Unit I
Introduction to R, RStudio (GUI): R Windows Environment, introduction to various data types, Numeric, Character, date, data frame, array, matrix etc.,
Reading Datasets, Working with different file types .txt, .csv etc. Outliers, Combining Datasets, R Functions and loops.
Summary Statistics - Summarizing data with R, Probability, Expected, Random, Bivariate Random variables, Probability distribution. Central Limit Theorem etc.

Unit II
SQL using R & Correlation and Regression Analysis (NOS 2101):
Introduction to NoSQL, Connecting R to NoSQL databases. Excel and R integration with R connector. Regression Analysis, Assumptions of OLS Regression, Regression Modelling. Correlation, ANOVA, Forecasting, Heteroscedasticity, Autocorrelation, Introduction to Multiple Regression etc.

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. This programming language was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language.

The R environment
R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes:
an effective data handling and storage facility,
a suite of operators for calculations on arrays, in particular matrices,
a large, coherent, integrated collection of intermediate tools for data analysis,
graphical facilities for data analysis and display either on-screen or on hardcopy, and
a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

R is an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

Why R Programming Language?
R programming is used as a leading tool for machine learning, statistics, and data analysis. Objects, functions, and packages can easily be created by R.
It’s a platform-independent language. This means it can be applied to all operating system.

It’s an open-source free language. That means anyone can install it in any organization without purchasing a license.
R programming language is not only a statistic package but also allows us to integrate with other languages (C, C++). Thus, you can easily interact with many data sources and statistical packages.
The R programming language has a vast community of users and it’s growing day by day.
R is currently one of the most requested programming languages in the Data Science job market that makes it the hottest trend nowadays.

Features of R Programming Language

Statistical Features of R:

Basic Statistics: The most common basic statistics terms are the mean, mode, and median. These are all known as “Measures of Central Tendency.” So using the R language we can measure central tendency very easily.
Static graphics: R is rich with facilities for creating and developing interesting static graphics. R contains functionality for many plot types including graphic maps, mosaic plots, biplots, and the list goes on.
Probability distributions: Probability distributions play a vital role in statistics and by using R we can easily handle various types of probability distribution such as Binomial Distribution, Normal Distribution, Chi-squared Distribution and many more.
Data analysis: It provides a large, coherent and integrated collection of tools for data analysis.

Programming Features of R:

R Packages: One of the major features of R is it has a wide availability of libraries. R has CRAN(Comprehensive R Archive Network), which is a repository holding more than 10, 0000 packages.
Distributed Computing: Distributed computing is a model in which components of a software system are shared among multiple computers to improve efficiency and performance. Two new packages ddR and multidplyr used for distributed programming in R were released in November 2015.

Advantages of R:

R is the most comprehensive statistical analysis package. As new technology and concepts often appear first in R.
As R programming language is an open source. Thus, you can run R anywhere and at any time.
R programming language is suitable for GNU/Linux and Windows operating system.
R programming is cross-platform which runs on any operating system.
In R, everyone is welcome to provide new packages, bug fixes, and code enhancements.

Disadvantages of R:

In the R programming language, the standard of some packages is less than perfect.
Although, R commands give little pressure to memory management. So R programming language may consume all available memory.
In R basically, nobody to complain if something doesn’t work.
R programming language is much slower than other programming languages such as Python and MATLAB.

Applications of R:

We use R for Data Science. It gives us a broad variety of libraries related to statistics. It also provides the environment for statistical computing and design.
R is used by many quantitative analysts as its programming tool. Thus, it helps in data importing and cleaning.
R is the most prevalent language. So many data analysts and research programmers use it. Hence, it is used as a fundamental tool for finance.
Tech giants like Google, Facebook, bing, Twitter, Accenture, Wipro and many more using R nowadays.

Local Environment Setup

If you are still willing to set up your environment for R, you can follow the steps given below.

Windows Installation

You can download the Windows installer version of R from R-3.2.2 for Windows (32/64 bit) and save it in a local directory.

As it is a Windows installer (.exe) with a name "R-version-win.exe". You can just double click and run the installer accepting the default settings. If your Windows is 32-bit version, it installs the 32-bit version. But if your windows is 64-bit, then it installs both the 32-bit and 64-bit versions.

After installation you can locate the icon to run the Program in a directory structure "R\R3.2.2\bin\i386\Rgui.exe" under the Windows Program Files. Clicking this icon brings up the R-GUI which is the R console to do R Programming.

Linux Installation

R is available as a binary for many versions of Linux at the location R Binaries.

The instruction to install Linux varies from flavor to flavor. These steps are mentioned under each type of Linux version in the mentioned link. However, if you are in a hurry, then you can use yum command to install R as follows −

```
$ yum install R
```

Above command will install core functionality of R programming along with standard packages, still you need additional package, then you can launch R prompt as follows −

```
$ R
R version 3.2.0 (2015-04-16) -- "Full of  Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
```

Now you can use install command at R prompt to install the required package. For example, the following command will install plotrix package which is required for 3D charts.

```
> install.packages("plotrix")
```

R Command Prompt

Once you have R environment setup, then it's easy to start your R command prompt by just typing the following command at your command prompt −

```
$ R
```

This will launch R interpreter and you will get a prompt > where you can start typing your program as follows > myString <- "Hello, World!"
> print (myString)
[1] "Hello, World!"
Here first statement defines a string variable myString, where we assign a string "Hello, World!" and then next statement print() is being used to print the value stored in variable myString.

R Script File

Usually, you will do your programming by writing your programs in script files and then you execute those scripts at your command prompt with the help of R interpreter called Rscript. So let's start with writing following code in a text file called test.R as under −

```
# My first program in R Programming
myString <- "Hello, World!"
print ( myString)
```

Save the above code in a file test.R and execute it at Linux command prompt as given below. Even if you are using Windows or other system, syntax will remain same.

```
$ Rscript test.R
```

When we run the above program, it produces the following result.
[1] "Hello, World!"

Comments

Comments are like helping text in your R program and they are ignored by the interpreter while executing your actual program. Single comment is written using # in the beginning of the statement as follows −

```
# My first program in R Programming
```

R does not support multi-line comments but you can perform a trick which is something as follows −

```
if(FALSE) {
  "This is a demo for multi-line comments and it should be put inside either a
    single OR double quote"
}

myString <- "Hello, World!"
print ( myString)
[1] "Hello, World!"
```

Though above comments will be executed by R interpreter, they will not interfere with your actual program. You should put such comments inside, either single or double quote.

R - Data Types

In contrast to other programming languages like C and java in R, the variables are not declared as some data type. The variables are assigned with R-Objects and the data type of the R-object becomes the data type of the variable. There are many types of R-objects. The frequently used ones are −

Vectors
Lists
Matrices

Arrays
Factors
Data Frames
The simplest of these objects is the vector object and there are six data types of these atomic vectors, also termed as six classes of vectors. The other R-Objects are built upon the atomic vectors.

Data Type	Example	Verify
Logical	TRUE, FALSE	v <- TRUE print(class(v)) it produces the following result − [1] "logical"
Numeric	12.3, 5, 999	v <- 23.5 print(class(v)) it produces the following result − [1] "numeric"
Integer	2L, 34L, 0L	v <- 2L print(class(v)) it produces the following result − [1] "integer"
Complex	3 + 2i	v <- 2+5i print(class(v)) it produces the following result − [1] "complex"
Character	'a' , '"good", "TRUE", '23.4'	v <- "TRUE" print(class(v)) it produces the following result − [1] "character"
Raw	"Hello" is stored as 48 65 6c 6c 6f	v <- charToRaw("Hello") print(class(v)) it produces the following result − [1] "raw"

R - Variables
A variable provides us with named storage that our programs can manipulate. A variable in R can store an atomic vector, group of atomic vectors or a combination of many Robjects. A valid variable name consists of letters, numbers and the dot or underline characters. The variable name starts with a letter or the dot not followed by a number.

Variable Name	Validity	Reason
var_name2.	valid	Has letters, numbers, dot and underscore
var_name%	Invalid	Has the character '%'. Only dot(.) and underscore allowed.
2var_name	invalid	Starts with a number
.var_name, var.name	valid	Can start with a dot(.) but the dot(.)should not be followed by a number.
.2var_name	invalid	The starting dot is followed by a number making it invalid.

_var_name	invalid	Starts with _ which is not valid

Variable Assignment

The variables can be assigned values using leftward, rightward and equal to operator. The values of the variables can be printed using print() or cat() function. The cat() function combines multiple items into a continuous print output.

```
# Assignment using equal operator.
var.1 = c(0,1,2,3)

# Assignment using leftward operator.
var.2 <- c("learn","R")

# Assignment using rightward operator.
c(TRUE,1) -> var.3

print(var.1)
cat ("var.1 is ", var.1 ,"\n")
cat ("var.2 is ", var.2 ,"\n")
cat ("var.3 is ", var.3 ,"\n")
```

When we execute the above code, it produces the following result −

```
[1] 0 1 2 3
var.1 is  0 1 2 3
var.2 is  learn R
var.3 is  1 1
```

Note − The vector c(TRUE,1) has a mix of logical and numeric class. So logical class is coerced to numeric class making TRUE as 1.

Data Type of a Variable

In R, a variable itself is not declared of any data type, rather it gets the data type of the R - object assigned to it. So R is called a dynamically typed language, which means that we can change a variable's data type of the same variable again and again when using it in a program.

```
var_x <- "Hello"
cat("The class of var_x is ",class(var_x),"\n")

var_x <- 34.5
cat("  Now the class of var_x is ",class(var_x),"\n")

var_x <- 27L
cat("   Next the class of var_x becomes ",class(var_x),"\n")
```

When we execute the above code, it produces the following result −

```
The class of var_x is  character
   Now the class of var_x is  numeric
      Next the class of var_x becomes  integer
```

Finding Variables

To know all the variables currently available in the workspace we use the ls() function. Also the ls() function can use patterns to match the variable names.

```
print(ls())
```

When we execute the above code, it produces the following result −

```
[1] "my var"     "my_new_var" "my_var"     "var.1"
[5] "var.2"      "var.3"      "var.name"   "var_name2."
[9] "var_x"      "varname"
```

Note − It is a sample output depending on what variables are declared in your environment.

The ls() function can use patterns to match the variable names.

```
# List the variables starting with the pattern "var".
print(ls(pattern = "var"))
```

When we execute the above code, it produces the following result −

```
[1] "my var"     "my_new_var" "my_var"     "var.1"
[5] "var.2"      "var.3"      "var.name"   "var_name2."
[9] "var_x"      "varname"
```

The variables starting with dot(.) are hidden, they can be listed using "all.names = TRUE" argument to ls() function.

```
print(ls(all.name = TRUE))
```

When we execute the above code, it produces the following result −

```
[1] ".cars"        ".Random.seed" ".var_name"    ".varname"     ".varname2"
[6] "my var"       "my_new_var"   "my_var"       "var.1"        "var.2"
[11]"var.3"        "var.name"     "var_name2."   "var_x"
```

Deleting Variables

Variables can be deleted by using the rm() function. Below we delete the variable var.3. On printing the value of the variable error is thrown.

```
rm(var.3)
print(var.3)
```

When we execute the above code, it produces the following result −

```
[1] "var.3"
Error in print(var.3) : object 'var.3' not found
```

All the variables can be deleted by using the rm() and ls() function together.

```
rm(list = ls())
print(ls())
```

When we execute the above code, it produces the following result –

```
character(0)
```

R – Operators

An operator is a symbol that tells the compiler to perform specific mathematical or logical manipulations. R language is rich in built-in operators and provides following types of operators.

Types of Operators

We have the following types of operators in R programming −

Arithmetic Operators

Relational Operators

Logical Operators

Assignment Operators

Miscellaneous Operators

Arithmetic Operators

Following table shows the arithmetic operators supported by R language. The operators act on each element of the vector.

Operator	Description	Example
+	Adds two vectors	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v+t) it produces the following result − [1] 10.0 8.5 10.0
−	Subtracts second vector from the first	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v-t) it produces the following result − [1] -6.0 2.5 2.0
*	Multiplies both vectors	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v*t) it produces the following result − [1] 16.0 16.5 24.0
/	Divide the first vector with the second	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v/t) When we execute the above code, it produces the following result − [1] 0.250000 1.833333 1.500000
%%	Give the remainder of the first vector with the second	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v%%t) it produces the following result − [1] 2.0 2.5 2.0
%/%	The result of division of first vector with second (quotient)	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v%/%t) it produces the following result −

		[1] 0 1 1
^	The first vector raised to the exponent of second vector	v <- c(2,5.5,6) t <- c(8, 3, 4) print(v^t) it produces the following result − [1] 256.000 166.375 1296.000

<u>Relational Operators</u>

Following table shows the relational operators supported by R language. Each element of the first vector is compared with the corresponding element of the second vector. The result of comparison is a Boolean value.

Operator	**Description**	**Example**
>	Checks if each element of the first vector is greater than the corresponding element of the second vector.	v <- c(2,5.5,6,9) t <- c(8,2.5,14,9) print(v>t) it produces the following result − [1] FALSE TRUE FALSE FALSE
<	Checks if each element of the first vector is less than the corresponding element of the second vector.	v <- c(2,5.5,6,9) t <- c(8,2.5,14,9) print(v < t) it produces the following result − [1] TRUE FALSE TRUE FALSE
==	Checks if each element of the first vector is equal to the corresponding element of the second vector.	v <- c(2,5.5,6,9) t <- c(8,2.5,14,9) print(v == t) it produces the following result − [1] FALSE FALSE FALSE TRUE
<=	Checks if each element of the first vector is less than or equal to the corresponding element of the second vector.	v <- c(2,5.5,6,9) t <- c(8,2.5,14,9) print(v<=t) it produces the following result − [1] TRUE FALSE TRUE TRUE
>=	Checks if each element of the first vector is greater than or equal to the corresponding element of the second vector.	v <- c(2,5.5,6,9) t <- c(8,2.5,14,9) print(v>=t) it produces the following result − [1] FALSE TRUE FALSE TRUE
!=	Checks if each element of the first vector is unequal to the corresponding element of the second vector.	v <- c(2,5.5,6,9) t <- c(8,2.5,14,9) print(v!=t) it produces the following result − [1] TRUE TRUE TRUE FALSE

<u>Logical Operators</u>

Following table shows the logical operators supported by R language. It is applicable only to vectors of type logical, numeric or complex. All numbers greater than 1 are considered as logical value TRUE.

Each element of the first vector is compared with the corresponding element of the second vector. The result of comparison is a Boolean value.

<table>
<tr><th>Operator</th><th>Description</th><th>Example</th></tr>
<tr><td>&</td><td>It is called Element-wise Logical AND operator. It combines each element of the first vector with the corresponding element of the second vector and gives a output TRUE if both the elements are TRUE.</td><td>v <- c(3,1,TRUE,2+3i)
t <- c(4,1,FALSE,2+3i)
print(v&t)
it produces the following result −
[1] TRUE TRUE FALSE TRUE</td></tr>
<tr><td>|</td><td>It is called Element-wise Logical OR operator. It combines each element of the first vector with the corresponding element of the second vector and gives a output TRUE if one the elements is TRUE.</td><td>v <- c(3,0,TRUE,2+2i)
t <- c(4,0,FALSE,2+3i)
print(v|t)
it produces the following result −
[1] TRUE FALSE TRUE TRUE</td></tr>
<tr><td>!</td><td>It is called Logical NOT operator. Takes each element of the vector and gives the opposite logical value.</td><td>v <- c(3,0,TRUE,2+2i)
print(!v)
it produces the following result −
[1] FALSE TRUE FALSE FALSE</td></tr>
</table>

The logical operator && and || considers only the first element of the vectors and give a vector of single element as output.

<table>
<tr><th>Operator</th><th>Description</th><th>Example</th></tr>
<tr><td>&&</td><td>Called Logical AND operator. Takes first element of both the vectors and gives the TRUE only if both are TRUE.</td><td>v <- c(3,0,TRUE,2+2i)
t <- c(1,3,TRUE,2+3i)
print(v&&t)
it produces the following result −
[1] TRUE</td></tr>
<tr><td>||</td><td>Called Logical OR operator. Takes first element of both the vectors and gives the TRUE if one of them is TRUE.</td><td>v <- c(0,0,TRUE,2+2i)
t <- c(0,3,TRUE,2+3i)
print(v||t)
it produces the following result −
[1] FALSE</td></tr>
</table>

Assignment Operators

These operators are used to assign values to vectors.

<table>
<tr><th>Operator</th><th>Description</th><th>Example</th></tr>
<tr><td><−
or
=
or
<<−</td><td>Called Left Assignment</td><td>v1 <- c(3,1,TRUE,2+3i)
v2 <<- c(3,1,TRUE,2+3i)
v3 = c(3,1,TRUE,2+3i)
print(v1)
print(v2)</td></tr>
</table>

		print(v3) it produces the following result − [1] 3+0i 1+0i 1+0i 2+3i [1] 3+0i 1+0i 1+0i 2+3i [1] 3+0i 1+0i 1+0i 2+3i
-> or ->>	Called Right Assignment	c(3,1,TRUE,2+3i) -> v1 c(3,1,TRUE,2+3i) ->> v2 print(v1) print(v2) it produces the following result − [1] 3+0i 1+0i 1+0i 2+3i [1] 3+0i 1+0i 1+0i 2+3i

Miscellaneous Operators

These operators are used to for specific purpose and not general mathematical or logical computation.

Operator	Description	Example
:	Colon operator. It creates the series of numbers in sequence for a vector.	v <- 2:8 print(v) it produces the following result − [1] 2 3 4 5 6 7 8
%in%	This operator is used to identify if an element belongs to a vector.	v1 <- 8 v2 <- 12 t <- 1:10 print(v1 %in% t) print(v2 %in% t) it produces the following result − [1] TRUE [1] FALSE
%*%	This operator is used to multiply a matrix with its transpose.	M = matrix(c(2,6,5,1,10,4), nrow = 2,ncol = 3,byrow = TRUE) t = M %*% t(M) print(t) it produces the following result − [,1] [,2] [1,] 65 82 [2,] 82 117

R - Decision making

Decision making structures require the programmer to specify one or more conditions to be evaluated or tested by the program, along with a statement or statements to be executed if the condition is determined to be true, and optionally, other statements to be executed if the condition is determined to be false.

An if statement consists of a Boolean expression followed by one or more statements.

Syntax

The basic syntax for creating an if statement in R is −

```
if(boolean_expression) {
   // statement(s) will execute if the boolean expression is true.
}
```

If the Boolean expression evaluates to be true, then the block of code inside the if statement will be executed. If Boolean expression evaluates to be false, then the first set of code after the end of the if statement (after the closing curly brace) will be executed.

Eg.

```
x <- 30L
if(is.integer(x)) {
   print("X is an Integer")
}
```

When the above code is compiled and executed, it produces the following result −

```
[1] "X is an Integer"
```

An **if** statement can be followed by an optional **else** statement which executes when the boolean expression is false.

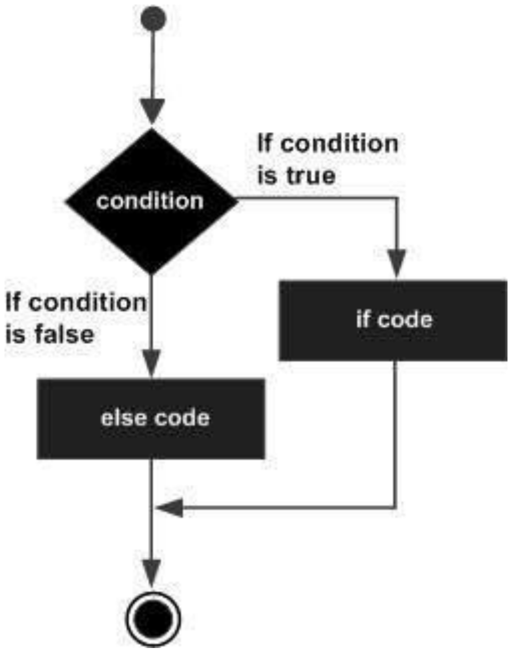
Syntax

The basic syntax for creating an **if...else** statement in R is −

```
if(boolean_expression) {
   // statement(s) will execute if the boolean expression is true.
} else {
   // statement(s) will execute if the boolean expression is false.
}
```

If the Boolean expression evaluates to be **true**, then the **if block** of code will be executed, otherwise **else block** of code will be executed.

Flow Diagram



Example

```
x <- c("what","is","truth")

if("Truth" %in% x) {
   print("Truth is found")
} else {
   print("Truth is not found")
}
```

When the above code is compiled and executed, it produces the following result −

```
[1] "Truth is not found"
```

Here "Truth" and "truth" are two different strings.

The if...else if...else Statement

An **if** statement can be followed by an optional **else if...else** statement, which is very useful to test various conditions using single if...else if statement.

When using **if**, **else if**, **else** statements there are few points to keep in mind.
An **if** can have zero or one **else** and it must come after any **else if**'s.
An **if** can have zero to many **else if's** and they must come before the else.
Once an **else if** succeeds, none of the remaining **else if**'s or **else**'s will be tested.

Syntax

The basic syntax for creating an **if...else if...else** statement in R is −

```
if(boolean_expression 1) {
   // Executes when the boolean expression 1 is true.
} else if( boolean_expression 2) {
   // Executes when the boolean expression 2 is true.
} else if( boolean_expression 3) {
   // Executes when the boolean expression 3 is true.
} else {
```

```
   // executes when none of the above condition is true.
}
```

Example

```
x <- c("what","is","truth")

if("Truth" %in% x) {
   print("Truth is found the first time")
} else if ("truth" %in% x) {
   print("truth is found the second time")
} else {
   print("No truth found")
}
```

When the above code is compiled and executed, it produces the following result −

```
[1] "truth is found the second time"
```

A **switch** statement allows a variable to be tested for equality against a list of values. Each value is called a case, and the variable being switched on is checked for each case.

Syntax

The basic syntax for creating a switch statement in R is −

```
switch(expression, case1, case2, case3....)
```

The following rules apply to a switch statement −

If the value of expression is not a character string it is coerced to integer.

You can have any number of case statements within a switch. Each case is followed by the value to be compared to and a colon.

If the value of the integer is between 1 and nargs()−1 (The max number of arguments)then the corresponding element of case condition is evaluated and the result returned.

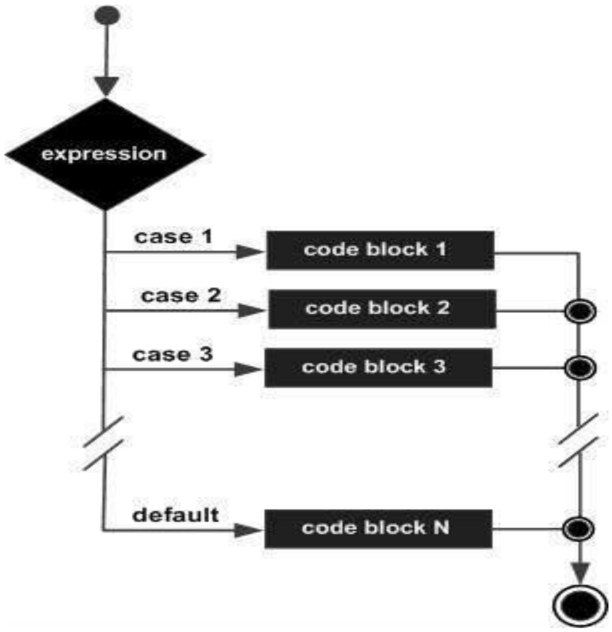
If expression evaluates to a character string then that string is matched (exactly) to the names of the elements.

If there is more than one match, the first matching element is returned.

No Default argument is available.

In the case of no match, if there is a unnamed element of ... its value is returned. (If there is more than one such argument an error is returned.)

Flow Diagram



Example

```
x <- switch(   3,   "first",   "second",   "third",   "fourth")
print(x)
```

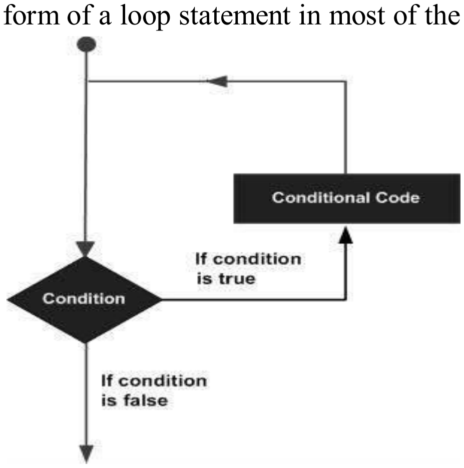
When the above code is compiled and executed, it produces the following result −

```
[1] "third"
```

R – Loops

Programming languages provide various control structures that allow for more complicated execution paths.

A loop statement allows us to execute a statement or group of statements multiple times and the following is the general form of a loop statement in most of the programming languages −

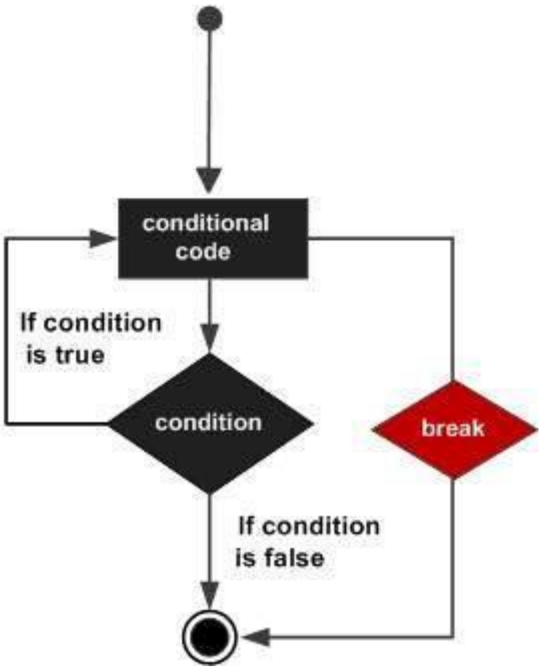


The **Repeat loop** executes the same code again and again until a stop condition is met.

Syntax

The basic syntax for creating a repeat loop in R is −

```
repeat {
   commands
   if(condition) {
      break
   }
}
```



Example

```
v <- c("Hello","loop")
cnt <- 2

repeat {
  print(v)
  cnt <- cnt+1

  if(cnt > 5) {
    break
  }
}
```

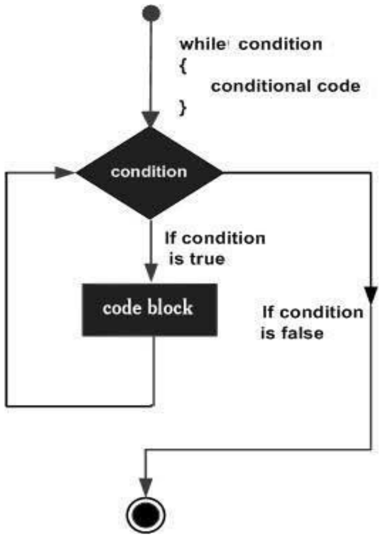
When the above code is compiled and executed, it produces the following result −

```
[1] "Hello" "loop"
[1] "Hello" "loop"
[1] "Hello" "loop"
[1] "Hello" "loop"
```

The **While loop** executes the same code again and again until a stop condition is met.

The basic syntax for creating a while loop in R is −

```
while (test_expression) {
  statement
}
```



Here key point of the **while** loop is that the loop might not ever run. When the condition is tested and the result is false, the loop body will be skipped and the first statement after the while loop will be executed.

Example

```
v <- c("Hello","while loop")
cnt <- 2

while (cnt < 7) {
   print(v)
   cnt = cnt + 1
}
```

When the above code is compiled and executed, it produces the following result −

```
[1] "Hello"  "while loop"
[1] "Hello"  "while loop"
[1] "Hello"  "while loop"
[1] "Hello"  "while loop"
[1] "Hello"  "while loop"
```

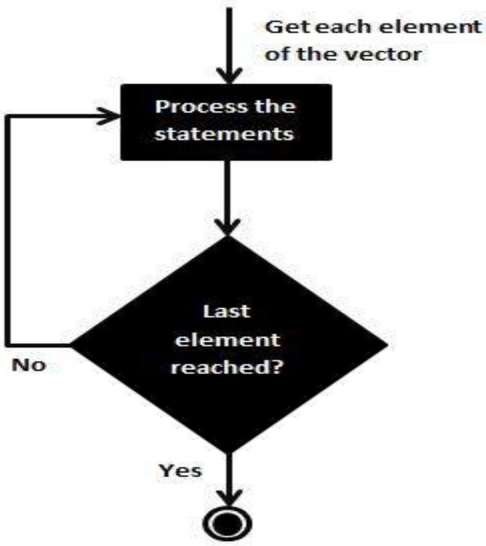
A **For loop** is a repetition control structure that allows you to efficiently write a loop that needs to execute a specific number of times.

Syntax

The basic syntax for creating a **for** loop statement in R is −

```
for (value in vector) {
   statements
}
```

Flow Diagram



R’s for loops are particularly flexible in that they are not limited to integers, or even numbers in the input. We can pass character vectors, logical vectors, lists or expressions.

Example
```
v <- LETTERS[1:4]
for ( i in v) {
   print(i)
}
```
When the above code is compiled and executed, it produces the following result −
```
[1] "A"
[1] "B"
[1] "C"
[1] "D"
```

Loop Control Statements

Loop control statements change execution from its normal sequence. When execution leaves a scope, all automatic objects that were created in that scope are destroyed.

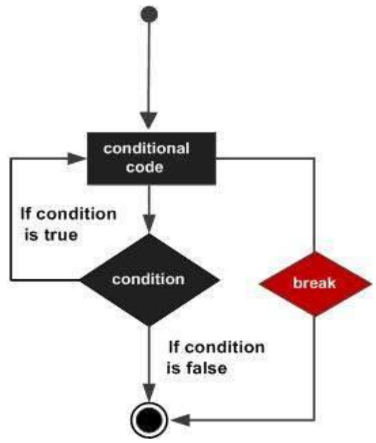
The **break statement** in R programming language has the following two usages −

When the break statement is encountered inside a loop, the loop is immediately terminated and program control resumes at the next statement following the loop.

It can be used to terminate a case in the switch statement (covered in the next chapter).

The basic syntax for creating a break statement in R is − break

Flow Diagram



Example

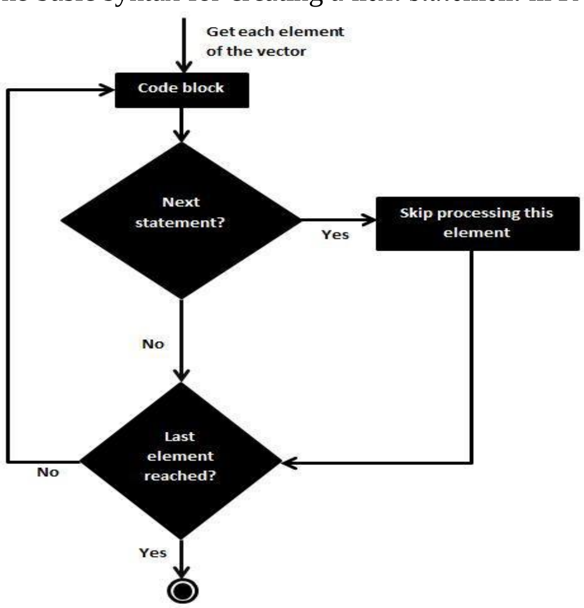
```
v <- c("Hello","loop")
cnt <- 2

repeat {
   print(v)
   cnt <- cnt + 1

   if(cnt > 5) {
      break
   }
}
```

When the above code is compiled and executed, it produces the following result −

```
[1] "Hello" "loop"
[1] "Hello" "loop"
[1] "Hello" "loop"
[1] "Hello" "loop"
```

The **next** statement in R programming language is useful when we want to skip the current iteration of a loop without terminating it. On encountering next, the R parser skips further evaluation and starts next iteration of the loop.

The basic syntax for creating a next statement in R is − next



Example

```
v <- LETTERS[1:6]
for ( i in v) {

  if (i == "D") {
    next
  }
  print(i)
}
```

When the above code is compiled and executed, it produces the following result −

```
[1] "A"
[1] "B"
[1] "C"
[1] "E"
[1] "F"
```

R – Functions

A function is a set of statements organized together to perform a specific task. R has a large number of in-built functions and the user can create their own functions.
In R, a function is an object so the R interpreter is able to pass control to the function, along with arguments that may be necessary for the function to accomplish the actions.
The function in turn performs its task and returns control to the interpreter as well as any result which may be stored in other objects.

Function Definition
An R function is created by using the keyword **function**. The basic syntax of an R function definition is as follows −

```
function_name <- function(arg_1, arg_2, ...) {
  Function body
}
```

Function Components
The different parts of a function are −
Function Name − This is the actual name of the function. It is stored in R environment as an object with this name.
Arguments − An argument is a placeholder. When a function is invoked, you pass a value to the argument. Arguments are optional; that is, a function may contain no arguments. Also arguments can have default values.
Function Body − The function body contains a collection of statements that defines what the function does.
Return Value − The return value of a function is the last expression in the function body to be evaluated.

R has many **in-built** functions which can be directly called in the program without defining them first. We can also create and use our own functions referred as **user defined** functions.

Built-in Function
Simple examples of in-built functions are **seq()**, **mean()**, **max()**, **sum(x)** and **paste(...)** etc. They are directly called by user written programs.

```
# Create a sequence of numbers from 32 to 44.
print(seq(32,44))
```

# Find mean of numbers from 25 to 82.
print(mean(25:82))

# Find sum of numbers frm 41 to 68.
print(sum(41:68))
When we execute the above code, it produces the following result −
[1] 32 33 34 35 36 37 38 39 40 41 42 43 44
[1] 53.5
[1] 1526

User-defined Function

We can create user-defined functions in R. They are specific to what a user wants and once created they can be used like the built-in functions. Below is an example of how a function is created and used.

```
# Create a function to print squares of numbers in sequence.
new.function <- function(a) {
  for(i in 1:a) {
    b <- i^2
    print(b)
  }
}
```

<u>Calling a Function</u>

```
# Create a function to print squares of numbers in sequence.
new.function <- function(a) {
  for(i in 1:a) {
    b <- i^2
    print(b)
  }
}

# Call the function new.function supplying 6 as an argument.
new.function(6)
```

When we execute the above code, it produces the following result −

```
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
[1] 36
```

<u>Calling a Function without an Argument</u>

```
# Create a function without an argument.
new.function <- function() {
  for(i in 1:5) {
    print(i^2)
  }
}

# Call the function without supplying an argument.
```

```
new.function()
```

When we execute the above code, it produces the following result −

```
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
```

Calling a Function with Argument Values (by position and by name)

The arguments to a function call can be supplied in the same sequence as defined in the function or they can be supplied in a different sequence but assigned to the names of the arguments.

```
# Create a function with arguments.
new.function <- function(a,b,c) {
  result <- a * b + c
  print(result)
}

# Call the function by position of arguments.
new.function(5,3,11)

# Call the function by names of the arguments.
new.function(a = 11, b = 5, c = 3)
```

When we execute the above code, it produces the following result −

```
[1] 26
[1] 58
```

Calling a Function with Default Argument

We can define the value of the arguments in the function definition and call the function without supplying any argument to get the default result. But we can also call such functions by supplying new values of the argument and get non default result.

```
# Create a function with arguments.
new.function <- function(a = 3, b = 6) {
  result <- a * b
  print(result)
}

# Call the function without giving any argument.
new.function()

# Call the function with giving new values of the argument.
new.function(9,5)
```

When we execute the above code, it produces the following result −

```
[1] 18
[1] 45
```

Lazy Evaluation of Function

Arguments to functions are evaluated lazily, which means so they are evaluated only when needed by the function body.

```
# Create a function with arguments.
new.function <- function(a, b) {
   print(a^2)
   print(a)
   print(b)
}

# Evaluate the function without supplying one of the arguments.
new.function(6)
```

When we execute the above code, it produces the following result −

```
[1] 36
[1] 6
Error in print(b) : argument "b" is missing, with no default
```

R - Strings

Any value written within a pair of single quote or double quotes in R is treated as a string. Internally R stores every string within double quotes, even when you create them with single quote.

Rules Applied in String Construction

The quotes at the beginning and end of a string should be both double quotes or both single quote. They cannot be mixed.

Double quotes can be inserted into a string starting and ending with single quote.

Single quote can be inserted into a string starting and ending with double quotes.

Double quotes cannot be inserted into a string starting and ending with double quotes.

Single quote cannot be inserted into a string starting and ending with single quote.

Examples of Valid Strings

Following examples clarify the rules about creating a string in R.

```
a <- 'Start and end with single quote'
print(a)

b <- "Start and end with double quotes"
print(b)

c <- "single quote ' in between double quotes"
print(c)

d <- 'Double quotes " in between single quote'
print(d)
```

When the above code is run we get the following output −

```
[1] "Start and end with single quote"
[1] "Start and end with double quotes"
[1] "single quote ' in between double quote"
[1] "Double quote \" in between single quote"
```

Examples of Invalid Strings

```
e <- 'Mixed quotes"
print(e)
```

f <- 'Single quote ' inside single quote'
print(f)

g <- "Double quotes " inside double quotes"
print(g)
When we run the script it fails giving below results.
Error: unexpected symbol in:
"print(e)
f <- 'Single"
Execution halted

String Manipulation

Concatenating Strings - paste() function

Many strings in R are combined using the **paste()** function. It can take any number of arguments to be combined together.
The basic syntax for paste function is −
paste(..., sep = " ", collapse = NULL)

Following is the description of the parameters used −
... represents any number of arguments to be combined.
sep represents any separator between the arguments. It is optional.
collapse is used to eliminate the space in between two strings. But not the space within two words of one string.

Example
a <- "Hello"
b <- 'How'
c <- "are you? "
print(paste(a,b,c))
print(paste(a,b,c, sep = "-"))
print(paste(a,b,c, sep = "", collapse = ""))

When we execute the above code, it produces the following result −
[1] "Hello How are you? "
[1] "Hello-How-are you? "
[1] "HelloHoware you? "

Formatting numbers & strings - format() function

Numbers and strings can be formatted to a specific style using **format()** function.

The basic syntax for format function is −
format(x, digits, nsmall, scientific, width, justify = c("left", "right", "centre", "none"))

Following is the description of the parameters used −
x is the vector input.
digits is the total number of digits displayed.
nsmall is the minimum number of digits to the right of the decimal point.
scientific is set to TRUE to display scientific notation.
width indicates the minimum width to be displayed by padding blanks in the beginning.
justify is the display of the string to left, right or center.

Example

```
# Total number of digits displayed. Last digit rounded off.
result <- format(23.123456789, digits = 9)
print(result)
# Display numbers in scientific notation.
result <- format(c(6, 13.14521), scientific = TRUE)
print(result)
# The minimum number of digits to the right of the decimal point.
result <- format(23.47, nsmall = 5)
print(result)
# Format treats everything as a string.
result <- format(6)
print(result)
# Numbers are padded with blank in the beginning for width.
result <- format(13.7, width = 6)
print(result)
# Left justify strings.
result <- format("Hello", width = 8, justify = "l")
print(result)
# Justfy string with center.
result <- format("Hello", width = 8, justify = "c")
print(result)
```

When we execute the above code, it produces the following result −

```
[1] "23.1234568"
[1] "6.000000e+00" "1.314521e+01"
[1] "23.47000"
[1] "6"
[1] "  13.7"
[1] "Hello   "
[1] " Hello  "
```

Counting number of characters in a string - nchar() function

This function counts the number of characters including spaces in a string.

The basic syntax for nchar() function is − nchar(x)

Following is the description of the parameters used −

x is the vector input.

Example

```
result <- nchar("Count the number of characters")
print(result)
```

When we execute the above code, it produces the following result −

```
[1] 30
```

Changing the case - toupper() & tolower() functions

These functions change the case of characters of a string.

The basic syntax for toupper() & tolower() function is −

```
toupper(x)
tolower(x)
```

Following is the description of the parameters used −

x is the vector input.

Example

```
# Changing to Upper case.
result <- toupper("Changing To Upper")
print(result)
# Changing to lower case.
result <- tolower("Changing To Lower")
print(result)
```

When we execute the above code, it produces the following result −

```
[1] "CHANGING TO UPPER"
[1] "changing to lower"
```

Extracting parts of a string - substring() function

This function extracts parts of a String.

The basic syntax for substring() function is −

```
substring(x,first,last)
```

Following is the description of the parameters used −

x is the character vector input.

first is the position of the first character to be extracted.

last is the position of the last character to be extracted.

Example

```
# Extract characters from 5th to 7th position.
result <- substring("Extract", 5, 7)
print(result)
```

When we execute the above code, it produces the following result −

```
[1] "act"
```

R – Vectors

Vectors are the most basic R data objects and there are six types of atomic vectors. They are logical, integer, double, complex, character and raw.

Vector Creation

Single Element Vector

Even when you write just one value in R, it becomes a vector of length 1 and belongs to one of the above vector types.

```
# Atomic vector of type character.
print("abc");
# Atomic vector of type double.
```

```
print(12.5)
# Atomic vector of type integer.
print(63L)
# Atomic vector of type logical.
print(TRUE)
# Atomic vector of type complex.
print(2+3i)
# Atomic vector of type raw.
print(charToRaw('hello'))
```

When we execute the above code, it produces the following result −

```
[1] "abc"
[1] 12.5
[1] 63
[1] TRUE
[1] 2+3i
[1] 68 65 6c 6c 6f
```

Multiple Elements Vector

Using colon operator with numeric data

```
# Creating a sequence from 5 to 13.
v <- 5:13
print(v)
# Creating a sequence from 6.6 to 12.6.
v <- 6.6:12.6
print(v)
# If the final element specified does not belong to the sequence then it is discarded.
v <- 3.8:11.4
print(v)
```

When we execute the above code, it produces the following result −

```
[1]  5  6  7  8  9 10 11 12 13
[1]  6.6  7.6  8.6  9.6 10.6 11.6 12.6
[1]  3.8  4.8  5.8  6.8  7.8  8.8  9.8 10.8
```

Using sequence (Seq.) operator

```
# Create vector with elements from 5 to 9 incrementing by 0.4.
print(seq(5, 9, by = 0.4))
```

When we execute the above code, it produces the following result −

```
[1] 5.0 5.4 5.8 6.2 6.6 7.0 7.4 7.8 8.2 8.6 9.0
```

Using the c() function

The non-character values are coerced to character type if one of the elements is a character.

```
# The logical and numeric values are converted to characters.
s <- c('apple','red',5,TRUE)
print(s)
```

When we execute the above code, it produces the following result −

```
[1] "apple" "red"   "5"     "TRUE"
```

Accessing Vector Elements

Elements of a Vector are accessed using indexing. The **[ ] brackets** are used for indexing. Indexing starts with position 1. Giving a negative value in the index drops that element from result.**TRUE**, **FALSE** or **0** and **1** can also be used for indexing.

```
# Accessing vector elements using position.
t <- c("Sun","Mon","Tue","Wed","Thurs","Fri","Sat")
u <- t[c(2,3,6)]
print(u)
# Accessing vector elements using logical indexing.
v <- t[c(TRUE,FALSE,FALSE,FALSE,FALSE,TRUE,FALSE)]
print(v)
# Accessing vector elements using negative indexing.
x <- t[c(-2,-5)]
print(x)
# Accessing vector elements using 0/1 indexing.
y <- t[c(0,0,0,0,0,0,1)]
print(y)
```

When we execute the above code, it produces the following result −

```
[1] "Mon" "Tue" "Fri"
[1] "Sun" "Fri"
[1] "Sun" "Tue" "Wed" "Fri" "Sat"
[1] "Sun"
```

Vector Manipulation

Vector arithmetic

Two vectors of same length can be added, subtracted, multiplied or divided giving the result as a vector output.

```
# Create two vectors.
v1 <- c(3,8,4,5,0,11)
v2 <- c(4,11,0,8,1,2)
# Vector addition.
add.result <- v1+v2
print(add.result)
# Vector subtraction.
sub.result <- v1-v2
print(sub.result)
# Vector multiplication.
multi.result <- v1*v2
print(multi.result)
# Vector division.
divi.result <- v1/v2
print(divi.result)
```

When we execute the above code, it produces the following result −

```
[1]  7 19  4 13  1 13
[1] -1 -3  4 -3 -1  9
[1] 12 88  0 40  0 22
[1] 0.7500000 0.7272727       Inf 0.6250000 0.0000000 5.5000000
```

Vector Element Recycling

If we apply arithmetic operations to two vectors of unequal length, then the elements of the shorter vector are recycled to complete the operations.

```
v1 <- c(3,8,4,5,0,11)
v2 <- c(4,11)
# V2 becomes c(4,11,4,11,4,11)
add.result <- v1+v2
print(add.result)
sub.result <- v1-v2
print(sub.result)
```

When we execute the above code, it produces the following result −

```
[1]  7 19  8 16  4 22
[1] -1 -3  0 -6 -4  0
```

Vector Element Sorting

Elements in a vector can be sorted using the **sort()** function.

```
v <- c(3,8,4,5,0,11, -9, 304)
# Sort the elements of the vector.
sort.result <- sort(v)
print(sort.result)
# Sort the elements in the reverse order.
revsort.result <- sort(v, decreasing = TRUE)
print(revsort.result)
# Sorting character vectors.
v <- c("Red","Blue","yellow","violet")
sort.result <- sort(v)
print(sort.result)
# Sorting character vectors in reverse order.
revsort.result <- sort(v, decreasing = TRUE)
print(revsort.result)
```

When we execute the above code, it produces the following result −

```
[1]  -9   0   3   4   5   8  11 304
[1] 304  11   8   5   4   3   0  -9
[1] "Blue"   "Red"    "violet" "yellow"
[1] "yellow" "violet" "Red"    "Blue"
```

R – Lists

Lists are the R objects which contain elements of different types like − numbers, strings, vectors and another list inside it. A list can also contain a matrix or a function as its elements. List is created using **list()** function.

Creating a List

Following is an example to create a list containing strings, numbers, vectors and a logical values.

```
# Create a list containing strings, numbers, vectors and a logical
# values.
list_data <- list("Red", "Green", c(21,32,11), TRUE, 51.23, 119.1)
print(list_data)
```

When we execute the above code, it produces the following result −

```
[[1]]
[1] "Red"
[[2]]
[1] "Green"
[[3]]
[1] 21 32 11
[[4]]
[1] TRUE
[[5]]
[1] 51.23
[[6]]
[1] 119.1
```

Naming List Elements

The list elements can be given names and they can be accessed using these names.

```
# Create a list containing a vector, a matrix and a list.
list_data <- list(c("Jan","Feb","Mar"), matrix(c(3,9,5,1,-2,8), nrow = 2),
   list("green",12.3))

# Give names to the elements in the list.
names(list_data) <- c("1st Quarter", "A_Matrix", "A Inner list")

# Show the list.
print(list_data)
```

When we execute the above code, it produces the following result −

```
$`1st_Quarter`
[1] "Jan" "Feb" "Mar"

$A_Matrix
     [,1] [,2] [,3]
[1,]    3    5   -2
[2,]    9    1    8

$A_Inner_list
$A_Inner_list[[1]]
[1] "green"

$A_Inner_list[[2]]
[1] 12.3
```

Accessing List Elements

Elements of the list can be accessed by the index of the element in the list. In case of named lists it can also be accessed using the names.

We continue to use the list in the above example −

```
# Create a list containing a vector, a matrix and a list.
list_data <- list(c("Jan","Feb","Mar"), matrix(c(3,9,5,1,-2,8), nrow = 2),
   list("green",12.3))

# Give names to the elements in the list.
names(list_data) <- c("1st Quarter", "A_Matrix", "A Inner list")

# Access the first element of the list.
print(list_data[1])

# Access the thrid element. As it is also a list, all its elements will be printed.
print(list_data[3])

# Access the list element using the name of the element.
print(list_data$A_Matrix)
```

When we execute the above code, it produces the following result −

```
$`1st_Quarter`
[1] "Jan" "Feb" "Mar"

$A_Inner_list
$A_Inner_list[[1]]
[1] "green"

$A_Inner_list[[2]]
[1] 12.3

     [,1] [,2] [,3]
[1,]    3    5   -2
[2,]    9    1    8
```

Manipulating List Elements

We can add, delete and update list elements as shown below. We can add and delete elements only at the end of a list. But we can update any element.

```
# Create a list containing a vector, a matrix and a list.
list_data <- list(c("Jan","Feb","Mar"), matrix(c(3,9,5,1,-2,8), nrow = 2), list("green",12.3))
# Give names to the elements in the list.
names(list_data) <- c("1st Quarter", "A_Matrix", "A Inner list")
# Add element at the end of the list.
list_data[4] <- "New element"
print(list_data[4])
# Remove the last element.
list_data[4] <- NULL
```

```
# Print the 4th Element.
print(list_data[4])
# Update the 3rd Element.
list_data[3] <- "updated element"
print(list_data[3])
```

When we execute the above code, it produces the following result −

```
[[1]]
[1] "New element"
$<NA>
NULL
$`A Inner list`
[1] "updated element"
```

Merging Lists

You can merge many lists into one list by placing all the lists inside one list() function.

```
# Create two lists.
list1 <- list(1,2,3)
list2 <- list("Sun","Mon","Tue")
# Merge the two lists.
merged.list <- c(list1,list2)
# Print the merged list.
print(merged.list)
```

When we execute the above code, it produces the following result −

```
[[1]]
[1] 1

[[2]]
[1] 2

[[3]]
[1] 3

[[4]]
[1] "Sun"

[[5]]
[1] "Mon"

[[6]]
[1] "Tue"
```

Converting List to Vector

A list can be converted to a vector so that the elements of the vector can be used for further manipulation. All the arithmetic operations on vectors can be applied after the list is converted into vectors. To do this conversion, we use the **unlist()** function. It takes the list as input and produces a vector.

```
# Create lists.
list1 <- list(1:5)
print(list1)
list2 <-list(10:14)
print(list2)

# Convert the lists to vectors.
v1 <- unlist(list1)
v2 <- unlist(list2)
print(v1)
print(v2)

# Now add the vectors
result <- v1+v2
print(result)
```

When we execute the above code, it produces the following result −

```
[[1]]
[1] 1 2 3 4 5

[[1]]
[1] 10 11 12 13 14

[1] 1 2 3 4 5
[1] 10 11 12 13 14
[1] 11 13 15 17 19
```

R – Matrices

Matrices are the R objects in which the elements are arranged in a two-dimensional rectangular layout. They contain elements of the same atomic types. Though we can create a matrix containing only characters or only logical values, they are not of much use. We use matrices containing numeric elements to be used in mathematical calculations.

A Matrix is created using the **matrix()** function.

The basic syntax for creating a matrix in R is −

```
matrix(data, nrow, ncol, byrow, dimnames)
```

Following is the description of the parameters used −

data is the input vector which becomes the data elements of the matrix.
nrow is the number of rows to be created.
ncol is the number of columns to be created.
byrow is a logical clue. If TRUE then the input vector elements are arranged by row.
dimname is the names assigned to the rows and columns.

Example
Create a matrix taking a vector of numbers as input.

```
# Elements are arranged sequentially by row.
M <- matrix(c(3:14), nrow = 4, byrow = TRUE)
```

```
print(M)

# Elements are arranged sequentially by column.
N <- matrix(c(3:14), nrow = 4, byrow = FALSE)
print(N)

# Define the column and row names.
rownames = c("row1", "row2", "row3", "row4")
colnames = c("col1", "col2", "col3")

P <- matrix(c(3:14), nrow = 4, byrow = TRUE, dimnames = list(rownames, colnames))
print(P)
```

When we execute the above code, it produces the following result −

```
     [,1] [,2] [,3]
[1,]    3    4    5
[2,]    6    7    8
[3,]    9   10   11
[4,]   12   13   14
     [,1] [,2] [,3]
[1,]    3    7   11
[2,]    4    8   12
[3,]    5    9   13
[4,]    6   10   14
     col1 col2 col3
row1    3    4    5
row2    6    7    8
row3    9   10   11
row4   12   13   14
```

Accessing Elements of a Matrix

Elements of a matrix can be accessed by using the column and row index of the element. We consider the matrix P above to find the specific elements below.

```
# Define the column and row names.
rownames = c("row1", "row2", "row3", "row4")
colnames = c("col1", "col2", "col3")

# Create the matrix.
P <- matrix(c(3:14), nrow = 4, byrow = TRUE, dimnames = list(rownames, colnames))

# Access the element at 3rd column and 1st row.
print(P[1,3])

# Access the element at 2nd column and 4th row.
print(P[4,2])

# Access only the  2nd row.
print(P[2,])
```

# Access only the 3rd column.
print(P[,3])
When we execute the above code, it produces the following result −

```
[1] 5
[1] 13
col1 col2 col3
   6    7    8
row1 row2 row3 row4
   5    8   11   14
```

Matrix Computations

Various mathematical operations are performed on the matrices using the R operators. The result of the operation is also a matrix.

The dimensions (number of rows and columns) should be same for the matrices involved in the operation.

<u>Matrix Addition & Subtraction</u>

```
# Create two 2x3 matrices.
matrix1 <- matrix(c(3, 9, -1, 4, 2, 6), nrow = 2)
print(matrix1)

matrix2 <- matrix(c(5, 2, 0, 9, 3, 4), nrow = 2)
print(matrix2)

# Add the matrices.
result <- matrix1 + matrix2
cat("Result of addition","\n")
print(result)

# Subtract the matrices
result <- matrix1 - matrix2
cat("Result of subtraction","\n")
print(result)
```

When we execute the above code, it produces the following result −

```
     [,1] [,2] [,3]
[1,]    3   -1    2
[2,]    9    4    6
     [,1] [,2] [,3]
[1,]    5    0    3
[2,]    2    9    4
Result of addition
     [,1] [,2] [,3]
[1,]    8   -1    5
[2,]   11   13   10
Result of subtraction
     [,1] [,2] [,3]
[1,]   -2   -1   -1
[2,]    7   -5    2
```

Matrix Multiplication & Division

```
# Create two 2x3 matrices.
matrix1 <- matrix(c(3, 9, -1, 4, 2, 6), nrow = 2)
print(matrix1)

matrix2 <- matrix(c(5, 2, 0, 9, 3, 4), nrow = 2)
print(matrix2)

# Multiply the matrices.
result <- matrix1 * matrix2
cat("Result of multiplication","\n")
print(result)

# Divide the matrices
result <- matrix1 / matrix2
cat("Result of division","\n")
print(result)
```

When we execute the above code, it produces the following result −

```
     [,1] [,2] [,3]
[1,]    3   -1    2
[2,]    9    4    6
     [,1] [,2] [,3]
[1,]    5    0    3
[2,]    2    9    4
Result of multiplication
     [,1] [,2] [,3]
[1,]   15    0    6
[2,]   18   36   24
Result of division
     [,1]      [,2]      [,3]
[1,]  0.6      -Inf 0.6666667
[2,]  4.5 0.4444444 1.5000000
```

R – Arrays

Arrays are the R data objects which can store data in more than two dimensions. For example − If we create an array of dimension (2, 3, 4) then it creates 4 rectangular matrices each with 2 rows and 3 columns. Arrays can store only data type.

An array is created using the **array()** function. It takes vectors as input and uses the values in the **dim** parameter to create an array.

Example

The following example creates an array of two 3x3 matrices each with 3 rows and 3 columns.

```
# Create two vectors of different lengths.
vector1 <- c(5,9,3)
vector2 <- c(10,11,12,13,14,15)
# Take these vectors as input to the array.
result <- array(c(vector1,vector2),dim = c(3,3,2))
```

```
print(result)
```

When we execute the above code, it produces the following result −

```
, , 1

     [,1] [,2] [,3]
[1,]    5   10   13
[2,]    9   11   14
[3,]    3   12   15

, , 2

     [,1] [,2] [,3]
[1,]    5   10   13
[2,]    9   11   14
[3,]    3   12   15
```

Naming Columns and Rows

We can give names to the rows, columns and matrices in the array by using the **dimnames** parameter.

```
# Create two vectors of different lengths.
vector1 <- c(5,9,3)
vector2 <- c(10,11,12,13,14,15)
column.names <- c("COL1","COL2","COL3")
row.names <- c("ROW1","ROW2","ROW3")
matrix.names <- c("Matrix1","Matrix2")

# Take these vectors as input to the array.
result <- array(c(vector1,vector2), dim = c(3,3,2), dimnames = list(row.names, column.names, matrix.names))
print(result)
```

When we execute the above code, it produces the following result −

```
, , Matrix1

     COL1 COL2 COL3
ROW1    5   10   13
ROW2    9   11   14
ROW3    3   12   15

, , Matrix2

     COL1 COL2 COL3
ROW1    5   10   13
ROW2    9   11   14
ROW3    3   12   15
```

Accessing Array Elements

```
# Create two vectors of different lengths.
vector1 <- c(5,9,3)
```

```
vector2 <- c(10,11,12,13,14,15)
column.names <- c("COL1","COL2","COL3")
row.names <- c("ROW1","ROW2","ROW3")
matrix.names <- c("Matrix1","Matrix2")
# Take these vectors as input to the array.
result <- array(c(vector1,vector2),dim = c(3,3,2),dimnames = list(row.names, column.names, matrix.names))
# Print the third row of the second matrix of the array.
print(result[3,,2])
# Print the element in the 1st row and 3rd column of the 1st matrix.
print(result[1,3,1])
# Print the 2nd Matrix.
print(result[,,2])
```

When we execute the above code, it produces the following result −

```
COL1 COL2 COL3
   3   12   15
[1] 13
     COL1 COL2 COL3
ROW1    5   10   13
ROW2    9   11   14
ROW3    3   12   15
```

Manipulating Array Elements

As array is made up matrices in multiple dimensions, the operations on elements of array are carried out by accessing elements of the matrices.

```
# Create two vectors of different lengths.
vector1 <- c(5,9,3)
vector2 <- c(10,11,12,13,14,15)
# Take these vectors as input to the array.
array1 <- array(c(vector1,vector2),dim = c(3,3,2))
# Create two vectors of different lengths.
vector3 <- c(9,1,0)
vector4 <- c(6,0,11,3,14,1,2,6,9)
array2 <- array(c(vector1,vector2),dim = c(3,3,2))
# create matrices from these arrays.
matrix1 <- array1[,,2]
matrix2 <- array2[,,2]
# Add the matrices.
result <- matrix1+matrix2
print(result)
```

When we execute the above code, it produces the following result −

```
     [,1] [,2] [,3]
[1,]   10   20   26
[2,]   18   22   28
[3,]    6   24   30
```

Calculations Across Array Elements

We can do calculations across the elements in an array using the **apply()** function.
Syntax
apply(x, margin, fun)

Following is the description of the parameters used −
x is an array.
margin is the name of the data set used.
fun is the function to be applied across the elements of the array.

Example
We use the apply() function below to calculate the sum of the elements in the rows of an array across all the matrices.

```
# Create two vectors of different lengths.
vector1 <- c(5,9,3)
vector2 <- c(10,11,12,13,14,15)
# Take these vectors as input to the array.
new.array <- array(c(vector1,vector2),dim = c(3,3,2))
print(new.array)

# Use apply to calculate the sum of the rows across all the matrices.
result <- apply(new.array, c(1), sum)
print(result)
```

When we execute the above code, it produces the following result −

```
, , 1

     [,1] [,2] [,3]
[1,]    5   10   13
[2,]    9   11   14
[3,]    3   12   15

, , 2

     [,1] [,2] [,3]
[1,]    5   10   13
[2,]    9   11   14
[3,]    3   12   15

[1] 56 68 60
```

R – Factors
Factors are the data objects which are used to categorize the data and store it as levels. They can store both strings and integers. They are useful in the columns which have a limited number of unique values. Like "Male, "Female" and True, False etc. They are useful in data analysis for statistical modeling.

Factors are created using the **factor ()** function by taking a vector as input.

Example
```
# Create a vector as input.
```

```
data <- c("East","West","East","North","North","East","West","West","West","East","North")

print(data)
print(is.factor(data))

# Apply the factor function.
factor_data <- factor(data)

print(factor_data)
print(is.factor(factor_data))
```

When we execute the above code, it produces the following result −

```
[1] "East"  "West"  "East"  "North" "North" "East"  "West"  "West"  "West"  "East" "North"
[1] FALSE
[1] East  West  East  North North East  West  West  West  East  North
Levels: East North West
[1] TRUE
```

Factors in Data Frame

On creating any data frame with a column of text data, R treats the text column as categorical data and creates factors on it.

```
# Create the vectors for data frame.
height <- c(132,151,162,139,166,147,122)
weight <- c(48,49,66,53,67,52,40)
gender <- c("male","male","female","female","male","female","male")
# Create the data frame.
input_data <- data.frame(height,weight,gender)
print(input_data)
# Test if the gender column is a factor.
print(is.factor(input_data$gender))
# Print the gender column so see the levels.
print(input_data$gender)
```

When we execute the above code, it produces the following result −

```
  height weight gender
1    132     48   male
2    151     49   male
3    162     66 female
4    139     53 female
5    166     67   male
6    147     52 female
7    122     40   male
[1] TRUE
[1] male   male   female female male   female male
Levels: female male
```

Changing the Order of Levels

The order of the levels in a factor can be changed by applying the factor function again with new order of the levels.

```
data <- c("East","West","East","North","North","East","West",
   "West","West","East","North")
# Create the factors
factor_data <- factor(data)
print(factor_data)
# Apply the factor function with required order of the level.
new_order_data <- factor(factor_data,levels = c("East","West","North"))
print(new_order_data)
```

When we execute the above code, it produces the following result −

```
 [1] East  West  East  North North East  West  West  West  East  North
Levels: East North West
 [1] East  West  East  North North East  West  West  West  East  North
Levels: East West North
```

Generating Factor Levels

We can generate factor levels by using the **gl()** function. It takes two integers as input which indicates how many levels and how many times each level.

Syntax

```
gl(n, k, labels)
```

Following is the description of the parameters used −

n is a integer giving the number of levels.

k is a integer giving the number of replications.

labels is a vector of labels for the resulting factor levels.

Example

```
v <- gl(3, 4, labels = c("Tampa", "Seattle","Boston"))
print(v)
```

When we execute the above code, it produces the following result −

```
Tampa   Tampa   Tampa   Tampa   Seattle Seattle Seattle Seattle Boston
[10] Boston  Boston  Boston
Levels: Tampa Seattle Boston
```

R - Data Frames

A data frame is a table or a two-dimensional array-like structure in which each column contains values of one variable and each row contains one set of values from each column.

Following are the characteristics of a data frame.
The column names should be non-empty.
The row names should be unique.
The data stored in a data frame can be of numeric, factor or character type.
Each column should contain same number of data items.

Create Data Frame

```
# Create the data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27")),
   stringsAsFactors = FALSE
)
# Print the data frame.
print(emp.data)
```

When we execute the above code, it produces the following result −

emp_	id	emp_name	salary	start_date
1	1	Rick	623.30	2012-01-01
2	2	Dan	515.20	2013-09-23
3	3	Michelle	611.00	2014-11-15
4	4	Ryan	729.00	2014-05-11
5	5	Gary	843.25	2015-03-27

<u>Get the Structure of the Data Frame</u>

The structure of the data frame can be seen by using **str()** function.

```
# Create the data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27")),
   stringsAsFactors = FALSE
)
# Get the structure of the data frame.
str(emp.data)
```

When we execute the above code, it produces the following result −

```
'data.frame':   5 obs. of  4 variables:
```

```
 $ emp_id    : int  1 2 3 4 5
 $ emp_name  : chr  "Rick" "Dan" "Michelle" "Ryan" ...
 $ salary    : num  623 515 611 729 843
 $ start_date: Date, format: "2012-01-01" "2013-09-23" "2014-11-15" "2014-05-11" ...
```

Summary of Data in Data Frame

The statistical summary and nature of the data can be obtained by applying **summary()** function.

```
# Create the data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27")),
   stringsAsFactors = FALSE
)
# Print the summary.
print(summary(emp.data))
```

When we execute the above code, it produces the following result −

```
     emp_id    emp_name             salary        start_date
 Min.   :1   Length:5          Min.   :515.2   Min.   :2012-01-01
 1st Qu.:2   Class :character   1st Qu.:611.0   1st Qu.:2013-09-23
 Median :3   Mode  :character   Median :623.3   Median :2014-05-11
 Mean   :3                      Mean   :664.4   Mean   :2014-01-14
 3rd Qu.:4                      3rd Qu.:729.0   3rd Qu.:2014-11-15
 Max.   :5                      Max.   :843.2   Max.   :2015-03-27
```

Extract Data from Data Frame

Extract specific column from a data frame using column name.

```
# Create the data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01","2013-09-23","2014-11-15","2014-05-11",
      "2015-03-27")),
   stringsAsFactors = FALSE
)
# Extract Specific columns.
result <- data.frame(emp.data$emp_name,emp.data$salary)
print(result)
```

When we execute the above code, it produces the following result −

```
  emp.data.emp_name emp.data.salary
1              Rick          623.30
2               Dan          515.20
```

```
3         Michelle         611.00
4            Ryan         729.00
5            Gary         843.25
```

Extract the first two rows and then all columns

```
# Create the data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27")),
   stringsAsFactors = FALSE
)
# Extract first two rows.
result <- emp.data[1:2,]
print(result)
```

When we execute the above code, it produces the following result −

```
  emp_id    emp_name   salary    start_date
1      1     Rick      623.3     2012-01-01
2      2     Dan       515.2     2013-09-23
```

Extract 3rd and 5th row with 2nd and 4th column

```
# Create the data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

         start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27")),
   stringsAsFactors = FALSE
)

# Extract 3rd and 5th row with 2nd and 4th column.
result <- emp.data[c(3,5),c(2,4)]
print(result)
```

When we execute the above code, it produces the following result −

```
  emp_name start_date
3 Michelle 2014-11-15
5     Gary 2015-03-27
```

Expand Data Frame

A data frame can be expanded by adding columns and rows.

Add Column

Just add the column vector using a new column name.

```
# Create the data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27")),
   stringsAsFactors = FALSE
)

# Add the "dept" coulmn.
emp.data$dept <- c("IT","Operations","IT","HR","Finance")
v <- emp.data
print(v)
```

When we execute the above code, it produces the following result −

```
  emp_id   emp_name    salary    start_date       dept
1     1    Rick        623.30    2012-01-01       IT
2     2    Dan         515.20    2013-09-23       Operations
3     3    Michelle    611.00    2014-11-15       IT
4     4    Ryan        729.00    2014-05-11       HR
5     5    Gary        843.25    2015-03-27       Finance
```

Add Row

To add more rows permanently to an existing data frame, we need to bring in the new rows in the same structure as the existing data frame and use the **rbind()** function.

In the example below we create a data frame with new rows and merge it with the existing data frame to create the final data frame.

```
# Create the first data frame.
emp.data <- data.frame(
   emp_id = c (1:5),
   emp_name = c("Rick","Dan","Michelle","Ryan","Gary"),
   salary = c(623.3,515.2,611.0,729.0,843.25),

   start_date = as.Date(c("2012-01-01", "2013-09-23", "2014-11-15", "2014-05-11",
      "2015-03-27")),
   dept = c("IT","Operations","IT","HR","Finance"),
   stringsAsFactors = FALSE
)

# Create the second data frame
emp.newdata <-            data.frame(
   emp_id = c (6:8),
   emp_name = c("Rasmi","Pranab","Tusar"),
   salary = c(578.0,722.5,632.8),
   start_date = as.Date(c("2013-05-21","2013-07-30","2014-06-17")),
```

```
   dept = c("IT","Operations","Fianance"),
   stringsAsFactors = FALSE
)

# Bind the two data frames.
emp.finaldata <- rbind(emp.data,emp.newdata)
print(emp.finaldata)
```

When we execute the above code, it produces the following result −

```
  emp_id    emp_name    salary     start_date       dept
1      1    Rick        623.30     2012-01-01       IT
2      2    Dan         515.20     2013-09-23       Operations
3      3    Michelle    611.00     2014-11-15       IT
4      4    Ryan        729.00     2014-05-11       HR
5      5    Gary        843.25     2015-03-27       Finance
6      6    Rasmi       578.00     2013-05-21       IT
7      7    Pranab      722.50     2013-07-30       Operations
8      8    Tusar       632.80     2014-06-17       Finance
```

R packages are a collection of R functions, complied code and sample data. They are stored under a directory called **"library"** in the R environment. By default, R installs a set of packages during installation. More packages are added later, when they are needed for some specific purpose. When we start the R console, only the default packages are available by default. Other packages which are already installed have to be loaded explicitly to be used by the R program that is going to use them.

All the packages available in R language are listed at `R Packages`.
Below is a list of commands to be used to check, verify and use the R packages.

Check Available R Packages

Get library locations containing R packages

```
.libPaths()
```

When we execute the above code, it produces the following result. It may vary depending on the local settings of your pc.

```
[2] "C:/Program Files/R/R-3.2.2/library"
```

Get the list of all the packages installed

```
library()
```

When we execute the above code, it produces the following result. It may vary depending on the local settings of your pc.

```
Packages in library ‘C:/Program Files/R/R-3.2.2/library’:

base                The R Base Package
boot                Bootstrap Functions (Originally by Angelo Canty
                    for S)
class               Functions for Classification
cluster             "Finding Groups in Data": Cluster Analysis
                    Extended Rousseeuw et al.
codetools           Code Analysis Tools for R
compiler            The R Compiler Package
datasets            The R Datasets Package
foreign             Read Data Stored by 'Minitab', 'S', 'SAS',
```

```
                'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...
graphics        The R Graphics Package
grDevices       The R Graphics Devices and Support for Colours
                and Fonts
grid            The Grid Graphics Package
KernSmooth      Functions for Kernel Smoothing Supporting Wand
                & Jones (1995)
lattice         Trellis Graphics for R
MASS            Support Functions and Datasets for Venables and
                Ripley's MASS
Matrix          Sparse and Dense Matrix Classes and Methods
methods         Formal Methods and Classes
mgcv            Mixed GAM Computation Vehicle with GCV/AIC/REML
                Smoothness Estimation
nlme            Linear and Nonlinear Mixed Effects Models
nnet            Feed-Forward Neural Networks and Multinomial
                Log-Linear Models
parallel        Support for Parallel computation in R
rpart           Recursive Partitioning and Regression Trees
spatial         Functions for Kriging and Point Pattern
                Analysis
splines         Regression Spline Functions and Classes
stats           The R Stats Package
stats4          Statistical Functions using S4 Classes
survival        Survival Analysis
tcltk           Tcl/Tk Interface
tools           Tools for Package Development
utils           The R Utils Package
```

Get all packages currently loaded in the R environment

```
search()
```

When we execute the above code, it produces the following result. It may vary depending on the local settings of your pc.

```
[1] ".GlobalEnv"        "package:stats"     "package:graphics"
[4] "package:grDevices" "package:utils"     "package:datasets"
[7] "package:methods"   "Autoloads"         "package:base"
```

Install a New Package

There are two ways to add new R packages. One is installing directly from the CRAN directory and another is downloading the package to your local system and installing it manually.

Install directly from CRAN

The following command gets the packages directly from CRAN webpage and installs the package in the R environment. You may be prompted to choose a nearest mirror. Choose the one appropriate to your location.

```
install.packages("Package Name")

# Install the package named "XML".
install.packages("XML")
```

Install package manually

Go to the link `R Packages` to download the package needed. Save the package as a **.zip** file in a suitable location in the local system.

Now you can run the following command to install this package in the R environment.

```
install.packages(file_name_with_path, repos = NULL, type = "source")

# Install the package named "XML"
install.packages("E:/XML_3.98-1.3.zip", repos = NULL, type = "source")
```

Load Package to Library

Before a package can be used in the code, it must be loaded to the current R environment. You also need to load a package that is already installed previously but not available in the current environment.

A package is loaded using the following command −

library("package Name", lib.loc = "path to library")

# Load the package named "XML"
install.packages("E:/XML_3.98-1.3.zip", repos = NULL, type = "source")

Data Reshaping

Data Reshaping in R is about changing the way data is organized into rows and columns. Most of the time data processing in R is done by taking the input data as a data frame. It is easy to extract data from the rows and columns of a data frame but there are situations when we need the data frame in a format that is different from format in which we received it. R has many functions to split, merge and change the rows to columns and vice-versa in a data frame.

Joining Columns and Rows in a Data Frame

We can join multiple vectors to create a data frame using the **cbind()** function. Also we can merge two data frames using **rbind()** function.

```
# Create vector objects.
city <- c("Tampa","Seattle","Hartford","Denver")
state <- c("FL","WA","CT","CO")
zipcode <- c(33602,98104,06161,80294)
# Combine above three vectors into one data frame.
addresses <- cbind(city,state,zipcode)
# Print a header.
cat("# # # # The First data frame\n")
# Print the data frame.
print(addresses)
# Create another data frame with similar columns
new.address <- data.frame(
   city = c("Lowry","Charlotte"),
   state = c("CO","FL"),
   zipcode = c("80230","33949"),
   stringsAsFactors = FALSE
)
# Print a header.
cat("# # # The Second data frame\n")
# Print the data frame.
print(new.address)
# Combine rows form both the data frames.
all.addresses <- rbind(addresses,new.address)
# Print a header.
cat("# # # The combined data frame\n")
# Print the result.
print(all.addresses)
```

When we execute the above code, it produces the following result −

```
# # # # The First data frame
     city       state zipcode
[1,] "Tampa"    "FL"  "33602"
[2,] "Seattle"  "WA"  "98104"
[3,] "Hartford" "CT"   "6161"
[4,] "Denver"   "CO"  "80294"

# # # The Second data frame
       city       state   zipcode
1      Lowry      CO      80230
```

```
2      Charlotte  FL      33949
```

```
# # # The combined data frame
       city      state zipcode
1      Tampa     FL    33602
2      Seattle   WA    98104
3      Hartford  CT     6161
4      Denver    CO    80294
5      Lowry     CO    80230
6     Charlotte  FL    33949
```

Merging Data Frames

We can merge two data frames by using the **merge()** function. The data frames must have same column names on which the merging happens.

In the example below, we consider the data sets about Diabetes in Pima Indian Women available in the library names "MASS". we merge the two data sets based on the values of blood pressure("bp") and body mass index("bmi"). On choosing these two columns for merging, the records where values of these two variables match in both data sets are combined together to form a single data frame.

```
library(MASS)
merged.Pima <- merge(x = Pima.te, y = Pima.tr,
   by.x = c("bp", "bmi"),
   by.y = c("bp", "bmi")
)
print(merged.Pima)
nrow(merged.Pima)
```

When we execute the above code, it produces the following result −

```
   bp  bmi npreg.x glu.x skin.x ped.x age.x type.x npreg.y glu.y skin.y ped.y
1  60 33.8       1   117     23 0.466    27     No       2   125     20 0.088
2  64 29.7       2    75     24 0.370    33     No       2   100     23 0.368
3  64 31.2       5   189     33 0.583    29    Yes       3   158     13 0.295
4  64 33.2       4   117     27 0.230    24     No       1    96     27 0.289
```

R - CSV Files

In R, we can read data from files stored outside the R environment. We can also write data into files which will be stored and accessed by the operating system. R can read and write into various file formats like csv, excel, xml etc.
In this chapter we will learn to read data from a csv file and then write data into a csv file. The file should be present in current working directory so that R can read it. Of course we can also set our own directory and read files from there.

Getting and Setting the Working Directory
You can check which directory the R workspace is pointing to using the **getwd()** function. You can also set a new working directory using **setwd()**function.

```
# Get and print current working directory.
print(getwd())
# Set current working directory.
setwd("/web/com")
# Get and print current working directory.
print(getwd())
```

When we execute the above code, it produces the following result −

```
[1] "/web/com/1441086124_2016"
[1] "/web/com"
```

This result depends on your OS and your current directory where you are working.

Input as CSV File
The csv file is a text file in which the values in the columns are separated by a comma.

Let's consider the following data present in the file named **input.csv**.

You can create this file using windows notepad by copying and pasting this data. Save the file as **input.csv** using the save As All files(*.*) option in notepad.

```
id,name,salary,start_date,dept
1,Rick,623.3,2012-01-01,IT
2,Dan,515.2,2013-09-23,Operations
3,Michelle,611,2014-11-15,IT
4,Ryan,729,2014-05-11,HR
5,Gary,843.25,2015-03-27,Finance
6,Nina,578,2013-05-21,IT
7,Simon,632.8,2013-07-30,Operations
8,Guru,722.5,2014-06-17,Finance
```

Reading a CSV File
Following is a simple example of **read.csv()** function to read a CSV file available in your current working directory −

```
data <- read.csv("input.csv")
print(data)
```

When we execute the above code, it produces the following result −

```
     id,  name,   salary,        start_date,   dept
1     1   Rick    623.30         2012-01-01     IT
```

2	2	Dan	515.20	2013-09-23	Operations
3	3	Michelle	611.00	2014-11-15	IT
4	4	Ryan	729.00	2014-05-11	HR
5	NA	Gary	843.25	2015-03-27	Finance
6	6	Nina	578.00	2013-05-21	IT
7	7	Simon	632.80	2013-07-30	Operations
8	8	Guru	722.50	2014-06-17	Finance

Analyzing the CSV File

By default the **read.csv()** function gives the output as a data frame. This can be easily checked as follows. Also we can check the number of columns and rows.

```
data <- read.csv("input.csv")
print(is.data.frame(data))
print(ncol(data))
print(nrow(data))
```

When we execute the above code, it produces the following result −

```
[1] TRUE
[1] 5
[1] 8
```

Once we read data in a data frame, we can apply all the functions applicable to data frames as explained in subsequent section.

Get the maximum salary

```
# Create a data frame.
data <- read.csv("input.csv")
# Get the max salary from data frame.
sal <- max(data$salary)
print(sal)
```

When we execute the above code, it produces the following result −

```
[1] 843.25
```

Get the details of the person with max salary

We can fetch rows meeting specific filter criteria similar to a SQL where clause.

```
# Create a data frame.
data <- read.csv("input.csv")
# Get the max salary from data frame.
sal <- max(data$salary)
# Get the person detail having max salary.
retval <- subset(data, salary == max(salary))
print(retval)
```

When we execute the above code, it produces the following result –

	id	name	salary	start_date	dept
5	NA	Gary	843.25	2015-03-27	Finance

Get all the people working in IT department

```
# Create a data frame.
data <- read.csv("input.csv")
retval <- subset( data, dept == "IT")
print(retval)
```

When we execute the above code, it produces the following result −

```
      id   name      salary             start_date   dept
1      1    Rick      623.3             2012-01-01   IT
3      3    Michelle  611.0             2014-11-15   IT
6      6    Nina      578.0             2013-05-21   IT
```

Get the persons in IT department whose salary is greater than 600

```
# Create a data frame.
data <- read.csv("input.csv")
info <- subset(data, salary > 600 & dept == "IT")
print(info)
```

When we execute the above code, it produces the following result −

```
      id   name      salary             start_date   dept
1      1    Rick      623.3             2012-01-01   IT
3      3    Michelle  611.0             2014-11-15   IT
```

Get the people who joined on or after 2014

```
# Create a data frame.
data <- read.csv("input.csv")

retval <- subset(data, as.Date(start_date) > as.Date("2014-01-01"))
print(retval)
```

When we execute the above code, it produces the following result −

```
      id   name     salary   start_date    dept
3      3    Michelle 611.00   2014-11-15    IT
4      4    Ryan     729.00   2014-05-11    HR
5     NA    Gary     843.25   2015-03-27    Finance
8      8    Guru     722.50   2014-06-17    Finance
```

Writing into a CSV File

R can create csv file form existing data frame. The **write.csv()** function is used to create the csv file. This file gets created in the working directory.

```
# Create a data frame.
data <- read.csv("input.csv")
retval <- subset(data, as.Date(start_date) > as.Date("2014-01-01"))

# Write filtered data into a new file.
write.csv(retval,"output.csv")
newdata <- read.csv("output.csv")
print(newdata)
```

When we execute the above code, it produces the following result −

```
  X     id   name      salary   start_date    dept
1 3      3   Michelle  611.00   2014-11-15    IT
2 4      4   Ryan      729.00   2014-05-11    HR
3 5     NA   Gary      843.25   2015-03-27    Finance
4 8      8   Guru      722.50   2014-06-17    Finance
```

Here the column X comes from the data set newper. This can be dropped using additional parameters while writing the file.

```
# Create a data frame.
data <- read.csv("input.csv")
retval <- subset(data, as.Date(start_date) > as.Date("2014-01-01"))

# Write filtered data into a new file.
write.csv(retval,"output.csv", row.names = FALSE)
newdata <- read.csv("output.csv")
print(newdata)
```

When we execute the above code, it produces the following result −

```
     id   name      salary   start_date    dept
1     3   Michelle  611.00   2014-11-15    IT
2     4   Ryan      729.00   2014-05-11    HR
3    NA   Gary      843.25   2015-03-27    Finance
4     8   Guru      722.50   2014-06-17    Finance
```

R - Excel File

Microsoft Excel is the most widely used spreadsheet program which stores data in the .xls or .xlsx format. R can read directly from these files using some excel specific packages. Few such packages are - XLConnect, xlsx, gdata etc. We will be using xlsx package. R can also write into excel file using this package.

Install xlsx Package

You can use the following command in the R console to install the "xlsx" package. It may ask to install some additional packages on which this package is dependent. Follow the same command with required package name to install the additional packages.

```
install.packages("xlsx")
```

Verify and Load the "xlsx" Package

Use the following command to verify and load the "xlsx" package.

```
# Verify the package is installed.
any(grepl("xlsx",installed.packages()))
# Load the library into R workspace.
library("xlsx")
```

When the script is run we get the following output.

```
[1] TRUE
Loading required package: rJava
Loading required package: methods
Loading required package: xlsxjars
```

Input as xlsx File

Open Microsoft excel. Copy and paste the following data in the work sheet named as sheet1.

id	name	salary	start_date	dept
1	Rick	623.3	1/1/2012	IT
2	Dan	515.2	9/23/2013	Operations
3	Michelle	611	11/15/2014	IT
4	Ryan	729	5/11/2014	HR
5	Gary	43.25	3/27/2015	Finance
6	Nina	578	5/21/2013	IT
7	Simon	632.8	7/30/2013	Operations
8	Guru	722.5	6/17/2014	Finance

Also copy and paste the following data to another worksheet and rename this worksheet to "city".

name	city
Rick	Seattle
Dan	Tampa
Michelle	Chicago
Ryan	Seattle
Gary	Houston
Nina	Boston
Simon	Mumbai
Guru	Dallas

Save the Excel file as "input.xlsx". You should save it in the current working directory of the R workspace.

Reading the Excel File

The input.xlsx is read by using the **read.xlsx()** function as shown below. The result is stored as a data frame in the R environment.

```
# Read the first worksheet in the file input.xlsx.
data <- read.xlsx("input.xlsx", sheetIndex = 1)
print(data)
```

When we execute the above code, it produces the following result −

```
     id,   name,     salary,   start_date,   dept
1      1    Rick      623.30    2012-01-01    IT
2      2    Dan       515.20    2013-09-23    Operations
3      3    Michelle  611.00    2014-11-15    IT
4      4    Ryan      729.00    2014-05-11    HR
5     NA    Gary      843.25    2015-03-27    Finance
6      6    Nina      578.00    2013-05-21    IT
7      7    Simon     632.80    2013-07-30    Operations
8      8    Guru      722.50    2014-06-17    Financ
```

R Data Visualization

In R, we can create visually appealing data visualizations by writing few lines of code. For this purpose, we use the diverse functionalities of R. Data visualization is an efficient technique for gaining insight about data through a visual medium. With the help of visualization techniques, a human can easily obtain information about hidden patterns in data that might be neglected.

By using the data visualization technique, we can work with large datasets to efficiently obtain key insights about it.

R Visualization Packages

R provides a series of packages for data visualization. These packages are as follows:

1) plotly

The plotly package provides online interactive and quality graphs. This package extends upon the JavaScript library ?plotly.js.

2) ggplot2

R allows us to create graphics declaratively. R provides the **ggplot** package for this purpose. This package is famous for its elegant and quality graphs, which sets it apart from other visualization packages.

3) tidyquant

The **tidyquant** is a financial package that is used for carrying out quantitative financial analysis. This package adds under tidyverse universe as a financial package that is used for importing, analyzing, and visualizing the data.

4) taucharts

Data plays an important role in taucharts. The library provides a declarative interface for rapid mapping of data fields to visual properties.

5) ggiraph

It is a tool that allows us to create dynamic ggplot graphs. This package allows us to add tooltips, JavaScript actions, and animations to the graphics.

6) geofacets

This package provides geofaceting functionality for 'ggplot2'. Geofaceting arranges a sequence of plots for different geographical entities into a grid that preserves some of the geographical orientation.

7) googleVis

googleVis provides an interface between R and Google's charts tools. With the help of this package, we can create web pages with interactive charts based on R data frames.

8) RColorBrewer

This package provides color schemes for maps and other graphics, which are designed by Cynthia Brewer.

9) dygraphs

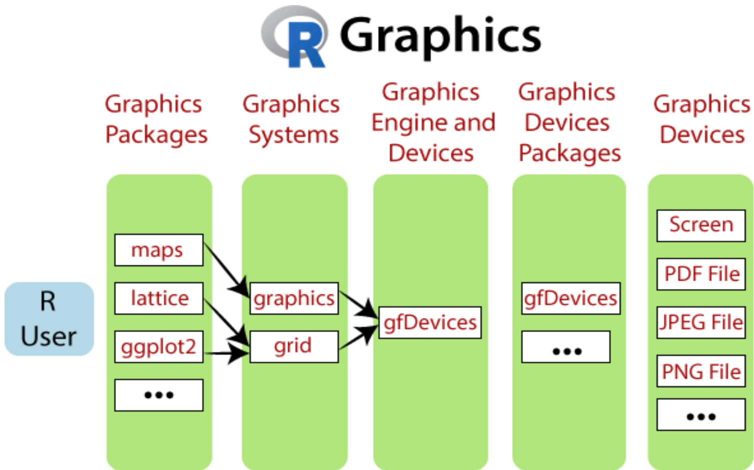
The dygraphs package is an R interface to the dygraphs JavaScript charting library. It provides rich features for charting time-series data in R.

10) shiny

R allows us to develop interactive and aesthetically pleasing web apps by providing a **shiny** package. This package provides various extensions with HTML widgets, CSS, and JavaScript.

R Graphics

Graphics play an important role in carrying out the important features of the data. Graphics are used to examine marginal distributions, relationships between variables, and summary of very large data. It is a very important complement for many statistical and computational techniques.



Standard Graphics

R standard graphics are available through package graphics, include several functions which provide statistical plots, like:
Scatterplots
Piecharts
Boxplots
Barplots etc.
We use the above graphs that are typically a single function call.

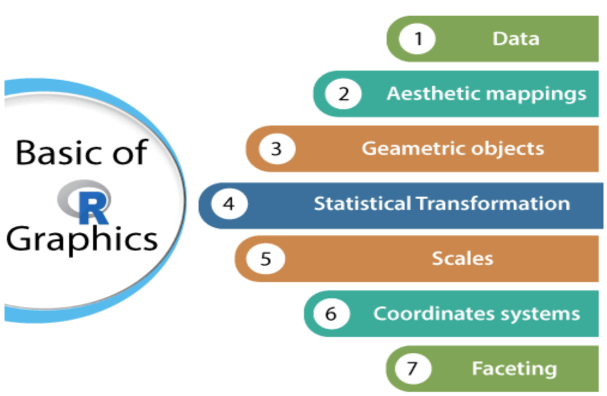
Graphics Devices

It is something where we can make a plot to appear. A graphics device is a window on your computer (screen device), a PDF file (file device), a Scalable Vector Graphics (SVG) file (file device), or a PNG or JPEG file (file device).

There are some of the following points which are essential to understand:
The functions of graphics devices produce output, which depends on the active graphics device.
A screen is the default and most frequently used device.
R graphical devices such as the PDF device, the JPEG device, etc. are used.
We just need to open the graphics output device which we want. Therefore, R takes care of producing the type of output which is required by the device.
For producing a certain plot on the screen or as a GIF R graphics file, the R code should exactly be the same. We only need to open the target output device before.
Several devices can be open at the same time, but there will be only one active device.

The basics of the grammar of graphics

There are some key elements of a statistical graphic. These elements are the basics of the grammar of graphics. Let's discuss each of the elements one by one to gain the basic knowledge of graphics.



1) Data

Data is the most crucial thing which is processed and generates an output.

2) Aesthetic Mappings

Aesthetic mappings are one of the most important elements of a statistical graphic. It controls the relation between graphics variables and data variables. In a scatter plot, it also helps to map the temperature variable of a data set into the X variable.

In graphics, it helps to map the species of a plant into the color of dots.

3) Geometric Objects

Geometric objects are used to express each observation by a point using the aesthetic mappings. It maps two variables in the data set into the x,y variables of the plot.

4) Statistical Transformations

Statistical transformations allow us to calculate the statistical analysis of the data in the plot. The statistical transformation uses the data and approximates it with the help of a regression line having x,y coordinates, and counts occurrences of certain values.

5) Scales

It is used to map the data values into values present in the coordinate system of the graphics device.

6) Coordinate system

The coordinate system plays an important role in the plotting of the data.

Cartesian

Plot

7) Faceting

Faceting is used to split the data into subgroups and draw sub-graphs for each group.

<u>Advantages of Data Visualization in R</u>

1. Understanding

It can be more attractive to look at the business. And, it is easier to understand through graphics and charts than a written document with text and numbers. Thus, it can attract a wider range of audiences. Also, it promotes the widespread use of business insights that come to make better decisions.

2. Efficiency

Its applications allow us to display a lot of information in a small space. Although, the decision-making process in business is inherently complex and multifunctional, displaying evaluation findings in a graph can allow companies to organize a lot of interrelated information in useful ways.

3. Location

Its app utilizing features such as Geographic Maps and GIS can be particularly relevant to wider business when the location is a very relevant factor. We will use maps to show business insights from various locations; also consider the seriousness of the issues, the reasons behind them, and working groups to address them.

<u>Disadvantages of Data Visualization in R</u>

1. Cost

R application development ranges a good amount of money. It may not be possible, especially for small companies, that many resources can be spent on purchasing them. To generate reports, many companies may employ professionals to create charts that can increase costs. Small enterprises are often operating in resource-limited settings, and are also receiving timely evaluation results that can often be of high importance.

2. Distraction

However, at times, data visualization apps create highly complex and fancy graphics-rich reports and charts, which may entice users to focus more on the form than the function. If we first add visual appeal, then the overall value of the graphic representation will be minimal. In resource-setting, it is required to understand how resources can be best used. And it is also not caught in the graphics trend without a clear purpose.

R Pie Charts

R programming language has several libraries for creating charts and graphs. A pie-chart is a representation of values in the form of slices of a circle with different colors. Slices are labeled with a description, and the numbers corresponding to each slice are also shown in the chart. However, pie charts are not recommended in the R documentation, and their characteristics are limited. The authors recommend a bar or dot plot on a pie chart because people are able to measure length more accurately than volume.

The Pie charts are created with the help of pie () function, which takes positive numbers as vector input. Additional parameters are used to control labels, colors, titles, etc.

There is the following syntax of the pie() function:

```
pie(X, Labels, Radius, Main, Col, Clockwise)
```

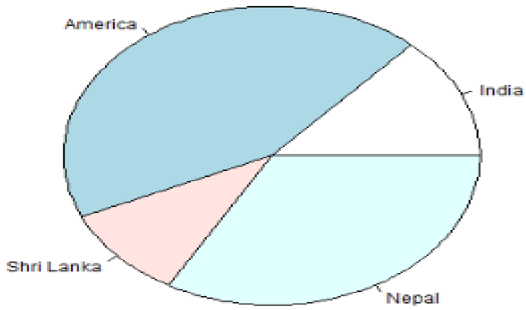
Here,

X is a vector that contains the numeric values used in the pie chart.
Labels are used to give the description to the slices.
Radius describes the radius of the pie chart.
Main describes the title of the chart.
Col defines the color palette.
Clockwise is a logical value that indicates the clockwise or anti-clockwise direction in which slices are drawn.

Example

```
# Creating data for the graph.
x <- c(20, 65, 15, 50)
labels <- c("India", "America", "Shri Lanka", "Nepal")
# Giving the chart file a name.
png(file = "Country.jpg")
# Plotting the chart.
pie(x,labels)
```

Output:



<u>Title and color</u>

A pie chart has several more features that we can use by adding more parameters to the pie() function. We can give a title to our pie chart by passing the main parameter. It tells the title of the pie chart to the pie() function. Apart from this, we can use a rainbow colour pallet while drawing the chart by passing the col parameter.

Note: The length of the pallet will be the same as the number of values that we have for the chart. So for that, we will use length() function.

Example

```
# Creating data for the graph.
```

```
x <- c(20, 65, 15, 50)
labels <- c("India", "America", "Shri Lanka", "Nepal")
# Giving the chart file a name.
png(file = "title_color.jpg")
# Plotting the chart.
pie(x,labels,main="Country Pie chart",col=rainbow(length(x)))
```

Output:

<u>Slice Percentage & Chart Legend</u>

There are two additional properties of the pie chart, i.e., slice percentage and chart legend. We can show the data in the form of percentage as well as we can add legends to plots in R by using the legend() function.

There is the following syntax of the legend() function.

legend(x,y=NULL,legend,fill,col,bg)

Here,

x and y are the coordinates to be used to position the legend.

legend is the text of legend

fill is the color to use for filling the boxes beside the legend text.

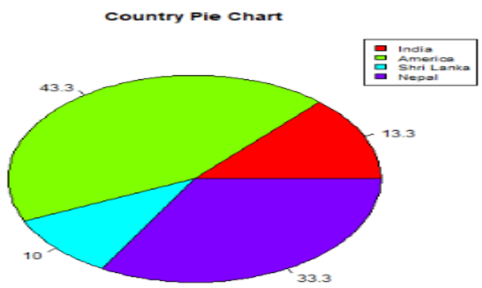
col defines the color of line and points besides the legend text.

bg is the background color for the legend box.

Example

```
# Creating data for the graph.
x <- c(20, 65, 15, 50)
labels <- c("India", "America", "Shri Lanka", "Nepal")
pie_percent<- round(100*x/sum(x), 1)
# Giving the chart file a name.
png(file = "per_pie.jpg")
# Plotting the chart.
pie(x, labels = pie_percent, main = "Country Pie Chart",col = rainbow(length(x)))
legend("topright", c("India", "America", "Shri Lanka", "Nepal"), cex = 0.8,
fill = rainbow(length(x)))
```

Output:



R Bar Charts

A bar chart is a pictorial representation in which numerical values of variables are represented by length or height of lines or rectangles of equal width. A bar chart is used for summarizing a set of categorical data. In bar chart, the data is shown through rectangular bars having the length of the bar proportional to the value of the variable.

In R, we can create a bar chart to visualize the data in an efficient manner. For this purpose, R provides the barplot() function, which has the following syntax:

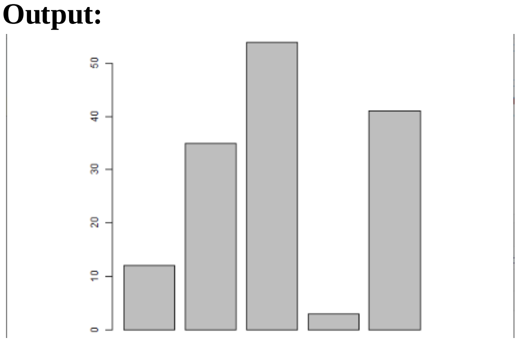
barplot(h, x, y, main, names.arg, col)

S.No	Parameter	Description
1.	H	A vector or matrix which contains numeric values used in the bar chart.
2.	xlab	A label for the x-axis.
3.	ylab	A label for the y-axis.
4.	main	A title of the bar chart.
5.	names.arg	A vector of names that appear under each bar.
6.	col	It is used to give colors to the bars in the graph.

Example

```
# Creating the data for Bar chart
H<- c(12,35,54,3,41)
# Giving the chart file a name
png(file = "bar_chart.png")
# Plotting the bar chart
barplot(H)
```

Output:



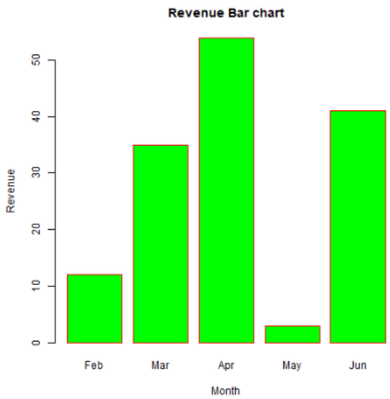
Labels, Title & Colors

Like pie charts, we can also add more functionalities in the bar chart by-passing more arguments in the barplot() functions. We can add a title in our bar chart or can add colors to the bar by adding the main and col parameters, respectively. We can add another parameter i.e., args.name, which is a vector that has the same number of values, which are fed as the input vector to describe the meaning of each bar.

Example

```
# Creating the data for Bar chart
H <- c(12,35,54,3,41)
M<- c("Feb","Mar","Apr","May","Jun")
# Giving the chart file a name
png(file = "bar_properties.png")
# Plotting the bar chart
barplot(H,names.arg=M,xlab="Month",ylab="Revenue",col="Green", main="Revenue Bar chart",border="red")
```

Output:



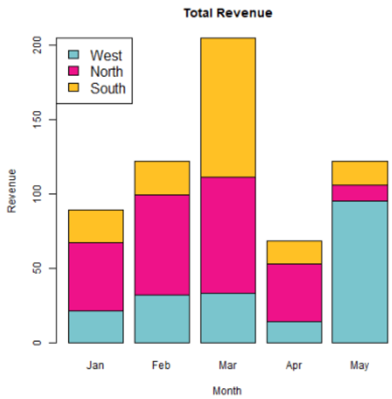
Group Bar Chart & Stacked Bar Chart

We can create bar charts with groups of bars and stacks using matrices as input values in each bar. One or more variables are represented as a matrix that is used to construct group bar charts and stacked bar charts.

Example

```
library(RColorBrewer)
months <- c("Jan","Feb","Mar","Apr","May")
regions <- c("West","North","South")
# Creating the matrix of the values.
Values <- matrix(c(21,32,33,14,95,46,67,78,39,11,22,23,94,15,16), nrow = 3, ncol = 5, byrow = TRUE)
# Giving the chart file a name
png(file = "stacked_chart.png")
# Creating the bar chart
barplot(Values, main = "Total Revenue", names.arg = months, xlab = "Month", ylab = "Revenue", ccol =c("cadetblue3","
deeppink2","goldenrod1"))
# Adding the legend to the chart
legend("topleft", regions, cex = 1.3, fill = c("cadetblue3","deeppink2","goldenrod1"))
```

Output:



R - Histograms

A histogram represents the frequencies of values of a variable bucketed into ranges. Histogram is similar to bar chat but the difference is it groups the values into continuous ranges. Each bar in histogram represents the height of the number of values present in that range.

R creates histogram using **hist()** function. This function takes a vector as an input and uses some more parameters to plot histograms.

The basic syntax for creating a histogram using R is −

```
hist(v,main,xlab,xlim,ylim,breaks,col,border)
```

Following is the description of the parameters used −

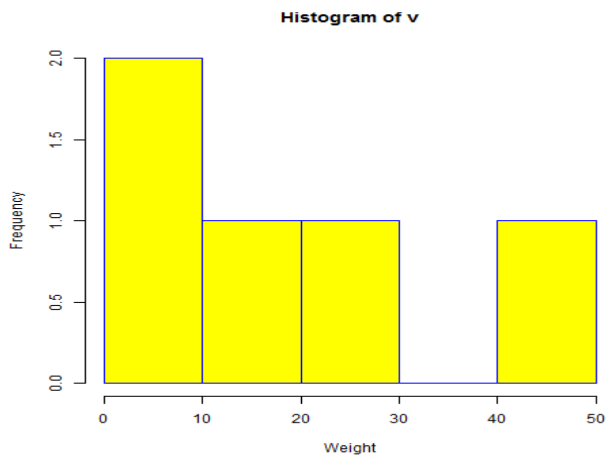
v is a vector containing numeric values used in histogram.
main indicates title of the chart.
col is used to set color of the bars.
border is used to set border color of each bar.
xlab is used to give description of x-axis.
xlim is used to specify the range of values on the x-axis.
ylim is used to specify the range of values on the y-axis.
breaks is used to mention the width of each bar.

Example

A simple histogram is created using input vector, label, col and border parameters.
The script given below will create and save the histogram in the current R working directory.

```
# Create data for the graph.
v <-  c(9,13,21,8,36,22,12,41,31,33,19)
# Give the chart file a name.
png(file = "histogram.png")
# Create the histogram.
hist(v,xlab = "Weight",col = "yellow",border = "blue")
```

When we execute the above code, it produces the following result −

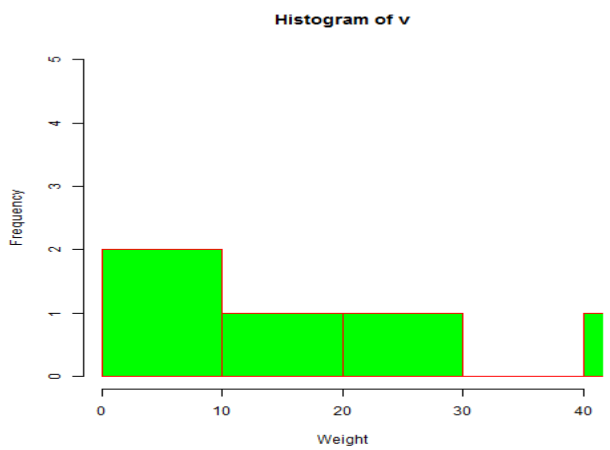


Range of X and Y values

To specify the range of values allowed in X axis and Y axis, we can use the xlim and ylim parameters. The width of each of the bar can be decided by using breaks.

```
# Create data for the graph.
v <- c(9,13,21,8,36,22,12,41,31,33,19)
# Give the chart file a name.
png(file = "histogram_lim_breaks.png")
# Create the histogram.
hist(v,xlab = "Weight",col = "green",border = "red", xlim = c(0,40), ylim = c(0,5),
   breaks = 5)
```

When we execute the above code, it produces the following result −



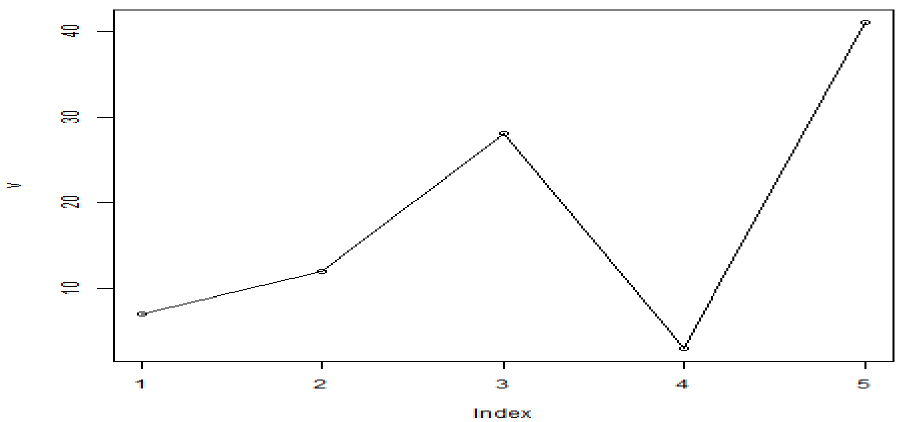
R - Line Graphs

A line chart is a graph that connects a series of points by drawing line segments between them. These points are ordered in one of their coordinate (usually the x-coordinate) value. Line charts are usually used in identifying the trends in data.
The **plot()** function in R is used to create the line graph.
Syntax
The basic syntax to create a line chart in R is −
plot(v,type,col,xlab,ylab)
Following is the description of the parameters used −
v is a vector containing the numeric values.
type takes the value "p" to draw only the points, "l" to draw only the lines and "o" to draw both points and lines.
xlab is the label for x axis.
ylab is the label for y axis.
main is the Title of the chart.
col is used to give colors to both the points and lines.

Example
A simple line chart is created using the input vector and the type parameter as "O". The below script will create and save a line chart in the current R working directory.

```
# Create the data for the chart.
v <- c(7,12,28,3,41)
# Give the chart file a name.
png(file = "line_chart.jpg")
# Plot the bar chart.
plot(v,type = "o")
```

When we execute the above code, it produces the following result −

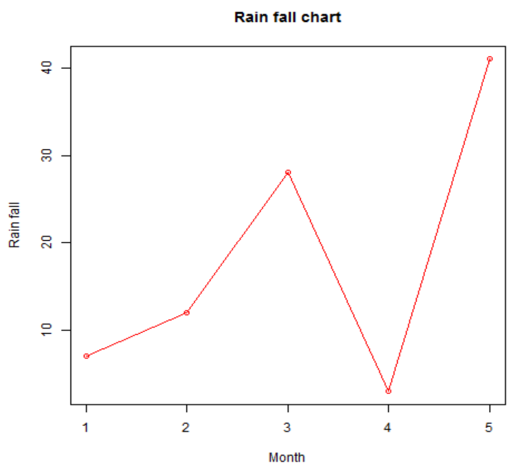


Line Chart Title, Color and Labels

The features of the line chart can be expanded by using additional parameters. We add color to the points and lines, give a title to the chart and add labels to the axes.

Example

```
# Create the data for the chart.
v <- c(7,12,28,3,41)
# Give the chart file a name.
png(file = "line_chart_label_colored.jpg")
# Plot the bar chart.
plot(v,type = "o", col = "red", xlab = "Month", ylab = "Rain fall", main = "Rain fall chart")
```

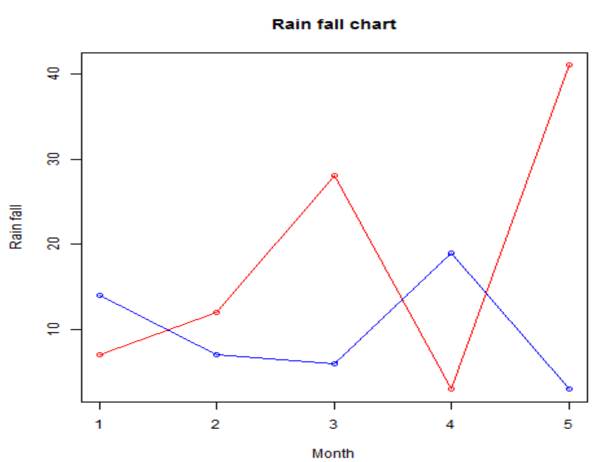


Multiple Lines in a Line Chart

More than one line can be drawn on the same chart by using the **lines()**function.

After the first line is plotted, the lines() function can use an additional vector as input to draw the second line in the chart,

```
# Create the data for the chart.
v <- c(7,12,28,3,41)
t <- c(14,7,6,19,3)
# Give the chart file a name.
png(file = "line_chart_2_lines.jpg")
# Plot the bar chart.
plot(v,type = "o",col = "red", xlab = "Month", ylab = "Rain fall", main = "Rain fall chart")
lines(t, type = "o", col = "blue")
```



R - Scatterplots

Scatterplots show many points plotted in the Cartesian plane. Each point represents the values of two variables. One variable is chosen in the horizontal axis and another in the vertical axis.

The simple scatterplot is created using the **plot()** function.
The basic syntax for creating scatterplot in R is −

```
plot(x, y, main, xlab, ylab, xlim, ylim, axes)
```

Following is the description of the parameters used −

x is the data set whose values are the horizontal coordinates.
y is the data set whose values are the vertical coordinates.
main is the tile of the graph.
xlab is the label in the horizontal axis.
ylab is the label in the vertical axis.
xlim is the limits of the values of x used for plotting.
ylim is the limits of the values of y used for plotting.
axes indicates whether both axes should be drawn on the plot.

Example

We use the data set **"mtcars"** available in the R environment to create a basic scatterplot. Let's use the columns "wt" and "mpg" in mtcars.

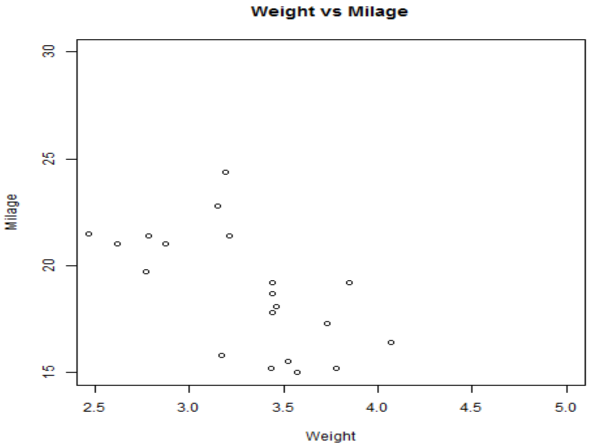
```
input <- mtcars[,c('wt','mpg')]
print(head(input))
```

	wt	mpg
Mazda RX4	2.620	21.0
Mazda RX4 Wag	2.875	21.0
Datsun 710	2.320	22.8
Hornet 4 Drive	3.215	21.4
Hornet Sportabout	3.440	18.7
Valiant	3.460	18.1

Creating the Scatterplot

The below script will create a scatterplot graph for the relation between wt(weight) and mpg(miles per gallon).

```
# Get the input values.
input <- mtcars[,c('wt','mpg')]
# Give the chart file a name.
png(file = "scatterplot.png")
# Plot the chart for cars with weight between 2.5 to 5 and mileage between 15 and 30.
plot(x = input$wt,y = input$mpg,
  xlab = "Weight",
  ylab = "Milage",
  xlim = c(2.5,5),
  ylim = c(15,30),
  main = "Weight vs Milage"
)
```

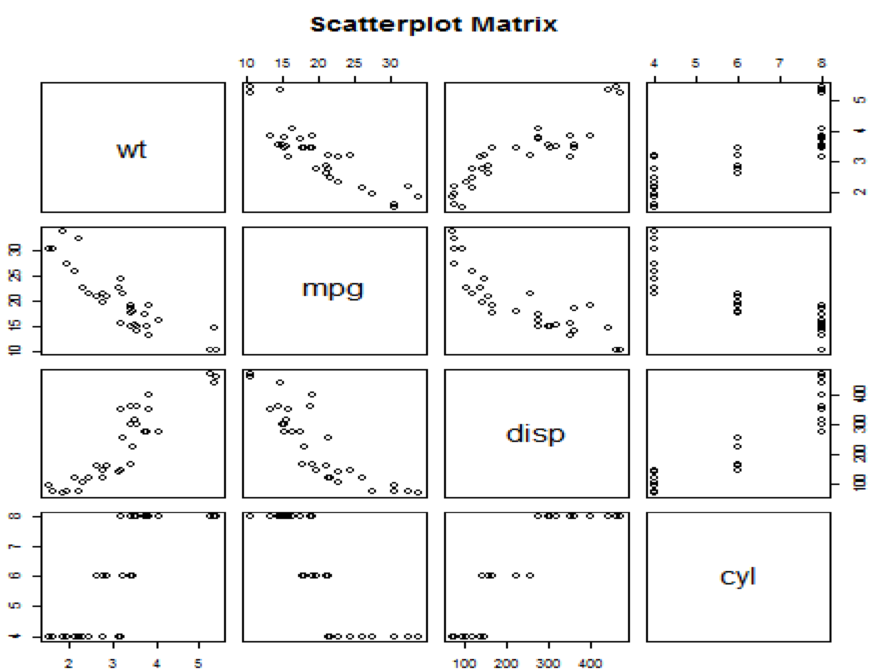


Scatterplot Matrices

When we have more than two variables and we want to find the correlation between one variable versus the remaining ones we use scatterplot matrix. We use **pairs()** function to create matrices of scatterplots.

The basic syntax for creating scatterplot matrices in R is −
pairs(formula, data)
Following is the description of the parameters used −
formula represents the series of variables used in pairs.
data represents the data set from which the variables will be taken.

Example
Each variable is paired up with each of the remaining variable. A scatterplot is plotted for each pair.
# Give the chart file a name.
png(file = "scatterplot_matrices.png")
# Plot the matrices between 4 variables giving 12 plots.
# One variable with 3 others and total 4 variables.
pairs(~wt+mpg+disp+cyl,data = mtcars, main = "Scatterplot Matrix")



Boxplots

Boxplots are a measure of how well distributed is the data in a data set. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data set. It is also useful in comparing the distribution of data across data sets by drawing boxplots for each of them.

Boxplots are created in R by using the **boxplot()** function.

Syntax

The basic syntax to create a boxplot in R is −

```
boxplot(x, data, notch, varwidth, names, main)
```

Following is the description of the parameters used −

x is a vector or a formula.

data is the data frame.

notch is a logical value. Set as TRUE to draw a notch.

varwidth is a logical value. Set as true to draw width of the box proportionate to the sample size.

names are the group labels which will be printed under each boxplot.

main is used to give a title to the graph.

Example

We use the data set "mtcars" available in the R environment to create a basic boxplot. Let's look at the columns "mpg" and "cyl" in mtcars.

```
input <- mtcars[,c('mpg','cyl')]
print(head(input))
```

When we execute above code, it produces following result −

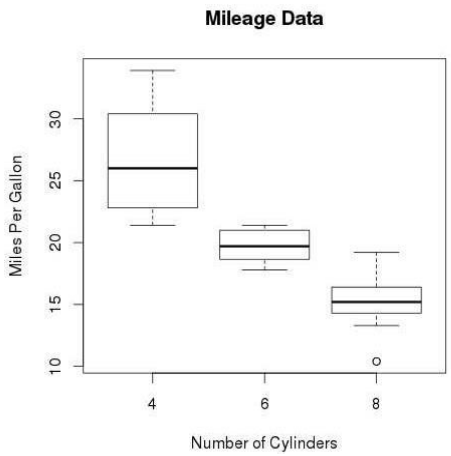
```
                  mpg  cyl
Mazda RX4         21.0   6
Mazda RX4 Wag     21.0   6
Datsun 710        22.8   4
Hornet 4 Drive    21.4   6
Hornet Sportabout 18.7   8
Valiant           18.1   6
```

Creating the Boxplot

The below script will create a boxplot graph for the relation between mpg (miles per gallon) and cyl (number of cylinders).

```
# Give the chart file a name.
png(file = "boxplot.png")

# Plot the chart.
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders",
   ylab = "Miles Per Gallon", main = "Mileage Data")
```

When we execute the above code, it produces the following result −

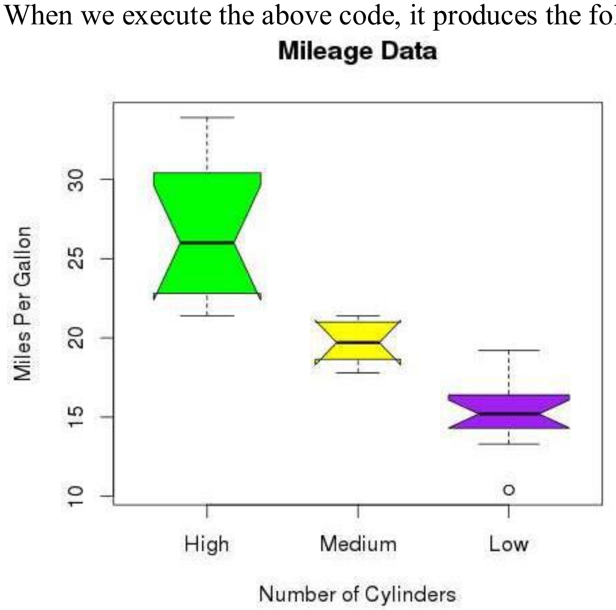


Boxplot with Notch

We can draw boxplot with notch to find out how the medians of different data groups match with each other.
The below script will create a boxplot graph with notch for each of the data group.

```
# Give the chart file a name.
png(file = "boxplot_with_notch.png")

# Plot the chart.
boxplot(mpg ~ cyl, data = mtcars,
   xlab = "Number of Cylinders",
   ylab = "Miles Per Gallon",
   main = "Mileage Data",
   notch = TRUE,
   varwidth = TRUE,
   col = c("green","yellow","purple"),
   names = c("High","Medium","Low")
)
```

When we execute the above code, it produces the following result −



Univariate Statistics

Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own.

Descriptive statistics describe and summarize data. Univariate descriptive statistics describe individual variables.

How to Analyze One Variable

1) Raw Data

Obtain a printout of the raw data for all the variables. Raw data resembles a matrix, with the variable names heading the columns, and the information for each case or record displayed across the rows.

Example: Raw data for a study of injuries among county workers (first 10 cases)

Injury Report No.	County Name	Cause of Injury	Severity of Injury
1	County A	Fall	3
2	County B	Auto	4
3	County C	Fall	6
4	County C	Fall	4
5	County B	Fall	5
6	County A	Violence	9
7	County A	Auto	3
8	County A	Violence	2
9	County A	Violence	9
10	County B	Auto	3

It is difficult to tell what is going on with each variable in this data set. Raw data is difficult to grasp, especially with large number of cases or records. Univariate descriptive statistics can summarize large quantities of numerical data and reveal patterns in the raw data. In order to present the information in a more organized format, start with univariate descriptive statistics for each variable.

For example, the variable Severity of Injury:

Severity of Injury
3
4
6
4
5
9
3
2
9
3

2) Frequency Distribution

Obtain a frequency distribution of the data for the variable. This is done by identifying the lowest and highest values of the variable, and then putting all the values of the variable in order from lowest to highest. Next, count the number of appearance of each value of the variable. This is a count of the frequency with which each value occurs in the data set.

For example, for the variable "Severity of Injury," the values range from 2 to 9.

Severity of Injury	Number of Injuries with this Severity
2	1
3	3
4	2
5	1
6	1
9	2
Total	10

3) Grouped Data

Decide on whether the data should be grouped into classes.

The severity of injury ratings can be collapsed into just a few categories or groups. Grouped data usually has from 3 to 7 groups. There should be no groups with a frequency of zero (for example, there are no injuries with a severity rating of 7 or 8).

One way to construct groups is to have equal class intervals (e.g., 1-3, 4-6, 7-9).

Another way to construct groups is to have about equal numbers of observations in each group.

(Remember that class intervals must be both mutually exclusive and exhaustive.)

Severity of Injury	Number of Injuries with this Severity
Mild (1-3)	4
Moderate (4-6)	4
Severe (6-9)	2
Total	10

4) Cumulative Distributions

Cumulative frequency distributions include a third column in the table (this can be done with either simple frequency distributions or with grouped data):

Severity of Injury	Number of Injuries	Cumulative frequency
2	1	1
3	3	4
4	2	6
5	1	7
6	1	8
9	2	10

A cumulative frequency distribution can answer questions such as, how many of the injuries were at level 5 or lower? Answer=7

5) Percentage Distributions

Frequencies can also be presented in the form of percentage distributions and cumulative percentages.

Severity of Injury	Percent of Injuries	Cumulative percentages
2	10	10
3	30	40
4	20	50
5	10	70
6	10	80
9	20	100

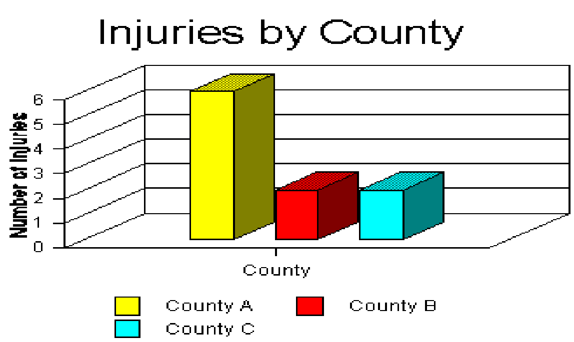
Why Graph? Graphing the Single Variable

Graphing is a way of visually presenting the data. Many people can grasp the information presented in a graph better than in a text format. The purpose of graphing is to:

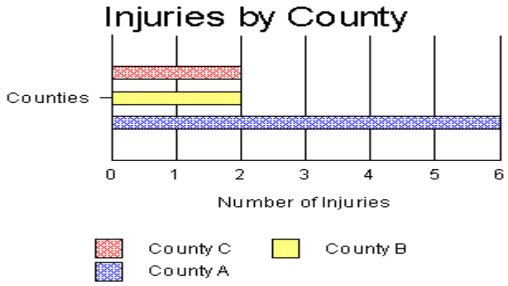
-present the data
-summarize the data
-enhance textual descriptions
-describe and explore the data
-make comparisons easy
-avoid distortion
-provoke thought about the data

Bar Graphs

Bar graphs are used to display the frequency distributions for variables measured at the nominal and ordinal levels. Bar graphs use the same width for all the bars on the graph, and there is space between the bars. Label the parts of the graph, including the title, the left (Y) or vertical axis, the right (X) or horizontal axis, and the bar labels.



Bar graphs can also be rotated so that the bars are parallel to the horizontal orientation of the page. For example,



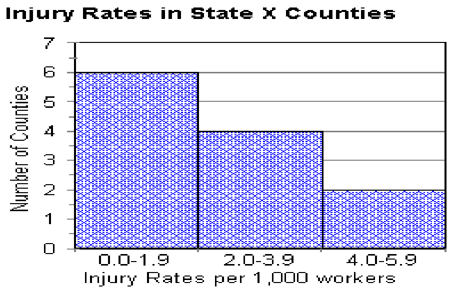
HISTOGRAM

A histogram is a chart that is similar to a bar chart, but it is used for interval and ratio level variables. With a histogram, the width of the bar is important, since it is the total area under the bar that represents the proportion of the phenomenon accounted for by each category. The bars convey the relationship of one group or class of the variable to the other(s).

For example, in the case of the counties and employee injuries, we might have information on the rate of injury according to the number of workers in each county in State X.

County Name	Rate of Injury per 1,000 workers
County A	5.5
County B	4.2
County C	3.8
County D	3.6
County E	3.4
County F	3.1
County G	1.8
County H	1.7
County I	1.6
County J	1.0
County K	0.9
County L	0.4

If we group the injury rates into three groups, then a low rate of injury would be 0.0-1.9 injuries per 1,000 workers; moderate would be 2.0-3.9; and high would be 4.0 and above (in this case, up to 5.9). This could be graphed as follows:

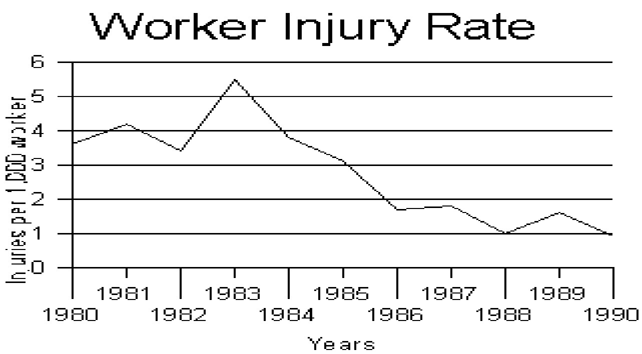


FREQUENCY POLYGON

A frequency polygon is another way of displaying information for an interval or ratio level variable. A frequency polygon displays the area under the curve that is represented by the values of the variable. This type of chart is also used to show time series graphs, or the changes in rates over time.

For example, the following table shows the average injury rate per 1,000 employes for counties in State X for the years 1980 to 1990.

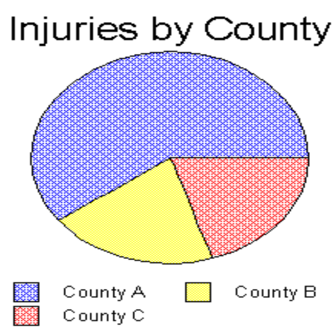
Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Rate	3.6	4.2	3.4	5.5	3.8	3.1	1.7	1.8	1.0	1.6	0.9



A cumulative frequency polygon is used to display the cumulative distribution of values for a variable.

PIE CHART

Another way to show the relationships between classes or categories of a variable is in a pie or circle chart. In a pie chart, each "slice" represents the proportion of the total phenomenon that is due to each of the classes or groups.



Rates and Ratios

Other ways to look at the sub-groups or classes within one variable is by the relation of each sub-group or class to the whole. This can be calculated with a proportion. A proportion is obtained by dividing the frequency of observations counted for one group or class (written as *f*) by the total number of observations counted for the variable (written as N).

This can be expressed as f / N

A percentage is the same as a proportion, multiplied by 100.

This can be expressed as f / N x 100

A rate is the relationship between two different numbers, for example, the number of injuries among county workers and the population of the county. This can be calculated as the first number (N_1, or injuries) divided by the second number (N_2, or population).

This can be expressed as N_1 / N_2

Many health statistics are expressed as rates, for example, the birth rate is the number of births per some population, such as number of births per 1,000 women.

Summary Statistics

Summary statistics include:

- **Minimum –** Get the Minimum element

Syntax:
min(data)

- **Maximum –** Get the Maximum element

Syntax:
max(data)

- **Mean –** Get the mean of the given elements

Syntax:
mean(data)

- **Median –** Get the median of the given elements

Syntax:
median(data)

- **Inter Quartile Range –** Get the IQR of the given elements

Syntax:
IQR(data)

- **Standard Deviation –** Get the standard deviation of the given elements

Syntax:
sd(data)

- **Range –** Get range from the elements

Syntax:
max(data)-min(data)

Example: R program to create a vector with 10 elements and display the Summary statistics.

```
# create a vector with 10 elements
data = c(1: 10)
# display
print(data)
# minimum
print(min(data))
# maximum
print(max(data))
# mean
print(mean(data))
# median
print(median(data))
# IQR
print(IQR(data))
# range
print(max(data)-min(data))
# standard deviation
print(sd(data))
```

Output:

```
[1]  1  2  3  4  5  6  7  8  9 10
[1] 1
[1] 10
[1] 5.5
```

[1] 5.5
[1] 4.5
[1] 9
[1] 3.02765

Frequency Table
We can display the frequency table using table() method, This will return the count of element occurrence.
Syntax:
table(data)
Example:
# create a vector with 10 elements
data = c(1: 10)
# display
print(data)
# display frequency table
print(table(data))

Output:

```
 [1]  1  2  3  4  5  6  7  8  9 10
data
 1  2  3  4  5  6  7  8  9 10
 1  1  1  1  1  1  1  1  1  1
```

Visualization
Here we can visualize the data using some plots

Boxplot
boxplot() function will result in a five-point summary(min, max, median, 1st quartile, 3rd quartile)
Syntax:
boxplot(data)
Example:
# create a vector with 10 elements
data = c(1: 10)
# display
print(data)
# display boxplot
print(boxplot(data))

Output:

```
[1]  1  2  3  4  5  6  7  8  9 10
$stats
     [,1]
[1,]  1.0
[2,]  3.0
[3,]  5.5
[4,]  8.0
[5,] 10.0
attr(,"class")
```

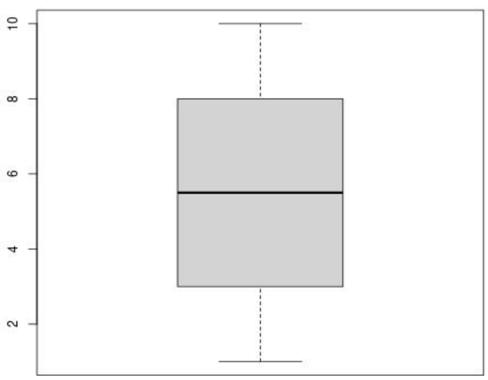
1
"integer"

$n
[1] 10

$conf
[,1]
[1,] 3.001801
[2,] 7.998199

$out
numeric(0)
$group
numeric(0)
$names
[1] "1"

Output:

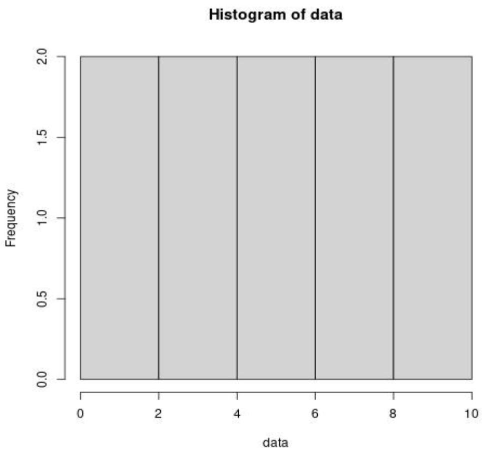


Histogram
This will return the histogram of the data and the function used is hist()
Syntax:
hist(data)
Example:
```
# create a vector with 10 elements
data = c(1: 10)
# display
print(data)
# display histogram
print(hist(data))
```

Output:
```
[1]  1  2  3  4  5  6  7  8  9 10
$breaks
[1]  0  2  4  6  8 10

$counts
[1] 2 2 2 2 2
```

$density
[1] 0.1 0.1 0.1 0.1 0.1
$mids
[1] 1 3 5 7 9
$xname
[1] "data"
$equidist
[1] TRUE
attr(,"class")
[1] "histogram"

Output:



Density plot

This will display the density plot . We have to use density() function along with plot() function.

Syntax:

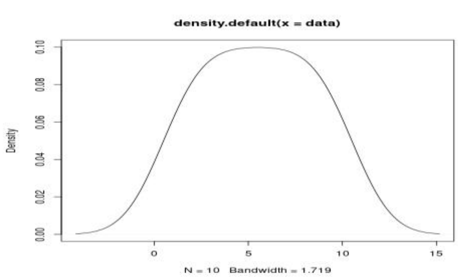
plot(density(data))

Example:

```
# create a vector with 10 elements
data = c(1: 10)
# display
print(data)
# display density plot
print(plot(density(data)))
```

Output:

[1] 1 2 3 4 5 6 7 8 9 10
NULL



The term **bivariate analysis** refers to the analysis of two variables.
The purpose of bivariate analysis is to understand the relationship between two variables

Handling bivariate categorical data
The table command will summarize bivariate data in a similar manner as it summarized univariate data.
Suppose a student survey is done to evaluate if students who smoke study less.
The data recorded is

Person	Smokes	amount of Studying
1	Y	less than 5 hours
2	N	5 - 10 hours
3	N	5 - 10 hours
4	Y	more than 10 hours
5	N	more than 10 hours
6	Y	less than 5 hours
7	Y	5 - 10 hours
8	Y	less than 5 hours
9	N	more than 5 hours
10	Y	5 - 10 hours

We can handle this in R by creating two vectors to hold our data, and then using the table command.

```
> smokes = c("Y","N","N","Y","N","Y","Y","Y","N","Y")
> amount = c(1,2,2,3,3,1,2,1,3,2)
> table(smokes,amount)
```

The command prop.table will compute this for us. It needs to be told the table to work on, and a number to indicate if you want the row proportions (a 1) or the column proportions (a 2) the default is to just find proportions.

```
> tmp=table(smokes,amount) # store the table
> old.digits = options("digits") # store the number of digits
> options(digits=3) # only print 3 decimal places
> prop.table(tmp,1) # the rows sum to 1 now
```

Plotting tabular data
You might wish to graphically represent the data summarized in a table.
For the smoking example, you could plot the amount variable for each of No or Yes, or the No and Yes variable for each level of smoking. In either case, you can use a barplot.

```
> barplot(table(smokes,amount))
> barplot(table(amount,smokes))
> smokes=factor(smokes) # for names
> barplot(table(smokes,amount),
+ beside=TRUE, # put beside not stacked
+ legend.text=T) # add legend
>
> barplot(table(amount,smokes),main="table(amount,smokes)",
+ beside=TRUE,
+ legend.text=c("less than 5","5-10","more than 10"))
```

There are three common ways to perform bivariate analysis:
1. Scatterplots

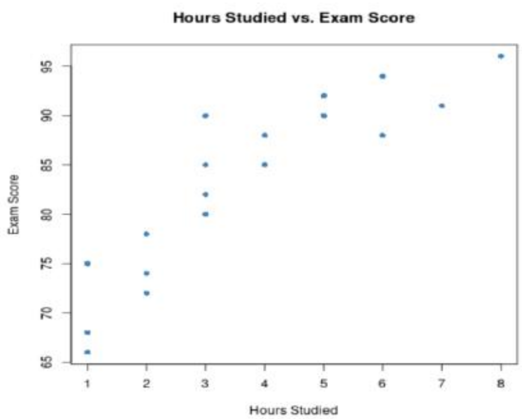
2. Correlation Coefficients
3. Simple Linear Regression

The following example shows how to perform each of these types of bivariate analysis using the following dataset that contains information about two variables: **(1)** Hours spent studying and **(2)** Exam score received by 20 different students:

```
#create data frame
df <- data.frame(hours=c(1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 5, 6, 6, 6, 7, 8),
                 score=c(75, 66, 68, 74, 78, 72, 85, 82, 90, 82,80, 88, 85, 90, 92, 94, 94, 88, 91, 96))

#view first six rows of data frame
head(df)

  hours score
1     1    75
2     1    66
3     1    68
4     2    74
5     2    78
6     2    72
```

1. Scatterplots
We can use the following syntax to create a scatterplot of hours studied vs. exam score in R:

```
#create scatterplot of hours studied vs. exam score
plot(df$hours, df$score, pch=16, col='steelblue',
     main='Hours Studied vs. Exam Score',
     xlab='Hours Studied', ylab='Exam Score')
```



The x-axis shows the hours studied and the y-axis shows the exam score received.
From the plot we can see that there is a positive relationship between the two variables: As hours studied increases, exam score tends to increase as well.

2. Correlation Coefficients
A Pearson Correlation Coefficient is a way to quantify the linear relationship between two variables.
We can use the **cor()** function in R to calculate the Pearson Correlation Coefficient between two variables:

```
#calculate correlation between hours studied and exam score received
cor(df$hours, df$score)

[1] 0.891306
```

The correlation coefficient turns out to be **0.891**.
This value is close to 1, which indicates a strong positive correlation between hours studied and exam score received.

3. Simple Linear Regression

Simple linear regression is a statistical method we can use to find the equation of the line that best "fits" a dataset, which we can then use to understand the exact relationship between two variables.

We can use the **lm()** function in R to fit a simple linear regression model for hours studied and exam score received:

```
#fit simple linear regression model
fit <- lm(score ~ hours, data=df)
#view summary of model
summary(fit)
```

```
Call:
lm(formula = score ~ hours, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-6.920 -3.927  1.309  1.903  9.385

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.0734     1.9651   35.15  < 2e-16 ***
hours         3.8471     0.4613    8.34 1.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.171 on 18 degrees of freedom
Multiple R-squared:  0.7944,	Adjusted R-squared:  0.783
F-statistic: 69.56 on 1 and 18 DF,  p-value: 1.347e-07
```

The fitted regression equation turns out to be:
Exam Score = 69.0734 + 3.8471*(hours studied)
This tells us that each additional hour studied is associated with an average increase of **3.8471** in exam score.
We can also use the fitted regression equation to predict the score that a student will receive based on their total hours studied.
For example, a student who studies for 3 hours is predicted to receive a score of **81.6147**:

- Exam Score = 69.0734 + 3.8471*(hours studied)
- Exam Score = 69.0734 + 3.8471*(3)
- Exam Score = 81.6147

Random Variable

A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. Random variables are often designated by letters and can be classified as discrete, which are variables that have specific values, or continuous, which are variables that can have any values within a continuous range.

Random variables are often used in econometric or regression analysis to determine statistical relationships among one another.

- A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.
- A random variable can be either discrete (having specific values) or continuous (any value in a continuous range).
- The use of random variables is most common in probability and statistics, where they are used to quantify outcomes of random occurrences.
- Risk analysts use random variables to estimate the probability of an adverse event occurring.

In probability and statistics, random variables are used to quantify outcomes of a random occurrence, and therefore, can take on many values. Random variables are required to be measurable and are typically real numbers.
For example, the letter X may be designated to represent the sum of the resulting numbers after three dice are rolled. In this case, X could be 3 (1 + 1+ 1), 18 (6 + 6 + 6), or somewhere between 3 and 18, since the highest number of a die is 6 and the lowest number is 1.

A random variable can be either discrete or continuous.

Discrete random variables take on a countable number of distinct values. Consider an experiment where a coin is tossed three times. If X represents the number of times that the coin comes up heads, then X is a discrete random variable that can only have the values 0, 1, 2, 3 (from no heads in three successive coin tosses to all heads). No other value is possible for X.

Continuous random variables can represent any value within a specified range or interval and can take on an infinite number of possible values. An example of a continuous random variable would be an experiment that involves measuring the amount of rainfall in a city over a year or the average height of a random group of 25 people.

Drawing on the latter, if Y represents the random variable for the average height of a random group of 25 people, you will find that the resulting outcome is a continuous figure since height may be 5 ft or 5.01 ft or 5.0001 ft. Clearly, there is an infinite number of possible values for height.

A random variable has a probability distribution that represents the likelihood that any of the possible values would occur. Let’s say that the random variable, Z, is the number on the top face of a die when it is rolled once. The possible values for Z will thus be 1, 2, 3, 4, 5, and 6. The probability of each of these values is 1/6 as they are all equally likely to be the value of Z.

For instance, the probability of getting a 3, or P (Z=3), when a die is thrown is 1/6, and so is the probability of having a 4 or a 2 or any other number on all six faces of a die. Note that the sum of all probabilities is 1.

Example of a Random Variable

A typical example of a random variable is the outcome of a coin toss. Consider a probability distribution in which the outcomes of a random event are not equally likely to happen. If the random variable Y is the number of heads we get from tossing two coins, then Y could be 0, 1, or 2. This means that we could have no heads, one head, or both heads on a two-coin toss.

However, the two coins land in four different ways: TT, HT, TH, and HH. Therefore, the P(Y=0) = 1/4 since we have one chance of getting no heads (i.e., two tails [TT] when the coins are tossed). Similarly, the probability of getting two heads (HH) is also 1/4. Notice that getting one head has a likelihood of occurring twice: in HT and TH. In this case, P (Y=1) = 2/4 = 1/2.

What Is a Discrete Random Variable?
A discrete random variable is a type of random variable that has a countable number of distinct values that can be assigned to it, such as in a coin toss.

What Is a Continuous Random Variable?
A continuous random variable stands for any amount within a specific range or set of points and can reflect an infinite number of potential values, such as the average rainfall in a region.

What Is a Mixed Random Variable?
A mixed random variable combines elements of both discrete and continuous random variables.

A random number is drawn from a population. A random variable is a variable for which we define a range of possible values and a probability distribution. The probability distribution specifies the probability that the variable assumes any value in its range.

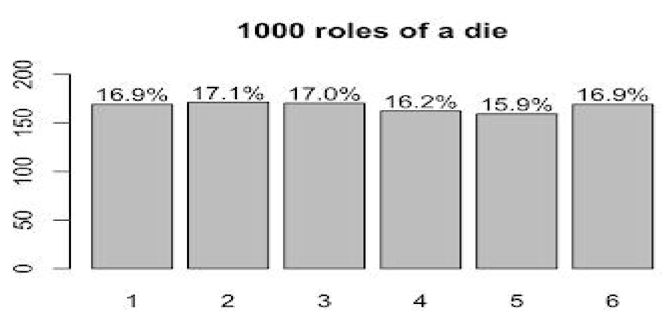
The range of a discrete random variable is a finite list of values. For any value k in the range, $0 \le P(X=k) \le 1$. The sum over all values k in the range is 1.

$$\sum_k P(X = k) = 1$$

R sample function implements these definitions.

For example, let’s ask R to create a six-sided die for us.
```
p.die <- rep(1/6,6)
sum(p.die)
```
Now, let's roll it 10 times.
```
die <- 1:6
sample(die, size=10, prob=p.die, replace=T)
```

Now, let's roll 1000 dice and plot the results.
```
s <- table(sample(die, size=1000, prob=p.die, replace=T))
lbls = sprintf("%0.1f%%", s/sum(s)*100)
barX <- barplot(s, ylim=c(0,200))
text(x=barX, y=s+10, label=lbls)
```



Expected value

Expected value (or population mean) of a discrete random variable *X* is the weighted average of the values in the range of *X*.

$$E(X) = \sum_{k} kP(X = k)$$

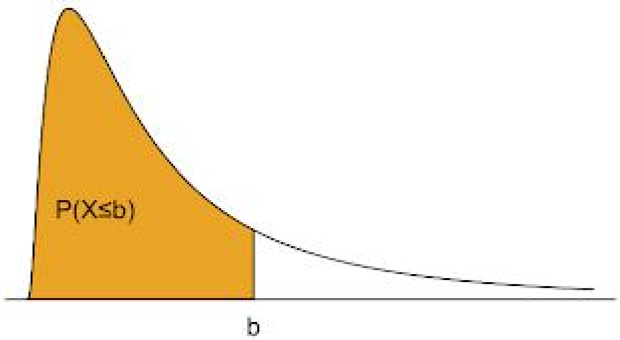
In the case of our six-sided die, the expected value is 3.5, computed like so:

```
sum(die*p.die)
```

A **continuous** random variable is described by a probability density function.

If *f(x)* is the probability density of a random variable *X*, $P(X \leq b)$ is the area under *f(x)* and to the left of *b*. The total area under $f(x) = 1$. $f(x) \geq 0$ for all possible values of *X*. $P(X>b) = 1 - P(X \leq b)$.

The expected value is the balance point where exactly half of the total area under *f(x)* is to the right and half is to the left.



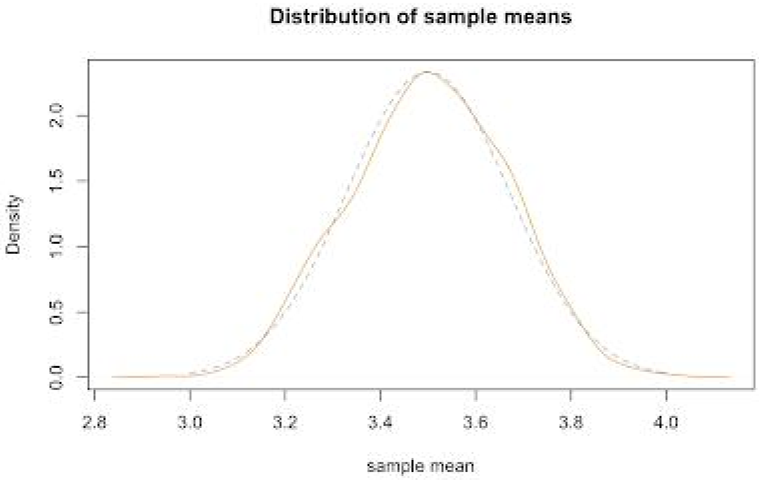
A **random sample** is a sequence of independent identically distributed random variables. A value derived from a random sample, such as sample mean, sample standard deviation, etc. is called a **statistic**. When we compute statistics of samples, our hope is that the sample statistic is not too far off from the equivalent measurement of the whole population.

The interesting thing is that derived statistics are also random variables. If we role our die several times, we have taken a random sample of size *n*. That sample can be summarized by computing its mean, denoted by *X bar*.

$$\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$$

The sample mean is, itself, a random variable with its own distribution. Define a function that generates a bunch of samples of size *n* and computes their means. It returns a vector of sample means.

```
generate.sample.means <- function(n) {
sample.means <- numeric()
for (i in 1:1000) {
sample.means <- append(sample.means, sum(sample(die, size=n, prob=p.die, replace=T))/n)
}
return (sample.means)
}
sample.means <- generate.sample.means(100)
plot(density(sample.means), main="Distribution of sample means",xlab="sample mean", col="orange")
```



It fits pretty closely to a normal distribution (dashed blue line). It's mean is about the same as the parent population, namely right around 3.5. The standard deviation of the sample means can be derived by dividing the standard deviation of the parent population by the square root of the sample size: $\sigma / \sqrt{n}$.

Let's compare the mean and *sd* of our sample with predicted values. We compute the standard deviation of our parent population of die rolls by squaring the deviation from the mean for each possible value and averaging that. Dividing that by the square root of our sample size gives the predicted *sd* of our sample means, about 0.17, which is equal with the actual *sd*.

```
> mean(sample.means)
[1] 3.49002
> sd(sample.means)
[1] 0.1704918
> sqrt(sum( (1:6-3.5)^2 ) / 6) / sqrt(100)
[1] 0.1707825
```

To overlay the normal distribution on the plot above, we used R's dnorm function like this:

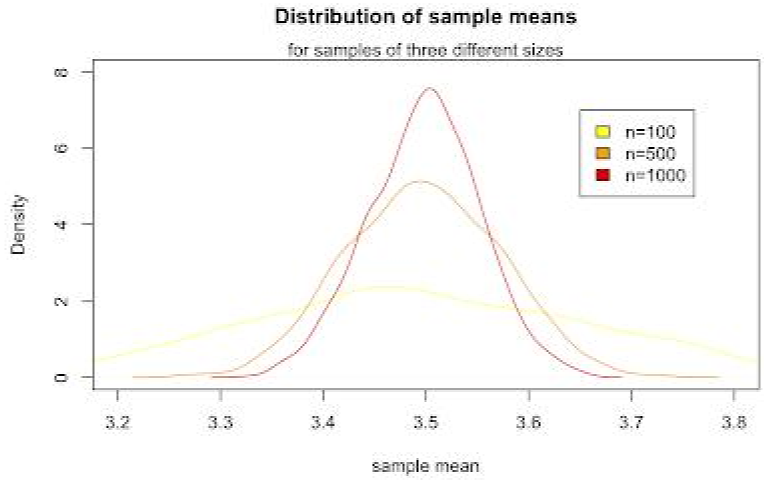
```
x = seq(3,4,0.01)
lines(x=x,y=dnorm(x,mean=3.5,sd=0.1707825), col=rgb(0x33,0x66,0xAA,0x90,maxColorValue=255), type="l", lty=2)
```

Inspection of the formula for the standard deviation of sample means supports our common sense intuition that a bigger sample will more likely reflect the whole population.

In particular, as the size of our sample goes up, our estimated mean is more likely to be closer to the parent population mean. This idea is known as the **law of large numbers**.

We can show that it works by creating similar plots with increasing *n*.

```
sample.means <- generate.sample.means(100)
plot(density(sample.means), main="Distribution of sample means", xlab="sample mean", col="yellow", xlim=c(3.2,3.8), ylim=c(0,8))
sample.means <- generate.sample.means(500)
lines(density(sample.means), col="orange")
sample.means <- generate.sample.means(1000)
lines(density(sample.means), col="red")
legend(3.6,7,c("n=100","n=500","n=1000"), fill=c("yellow", "orange", "red"))
```



What we've just discovered is the **central limit theorem**, which states that for any parent population with mean μ and standard deviation σ, the sampling distribution for large n is a normal distribution with mean μ and standard deviation $\sigma / \sqrt{n}$. Putting that in terms of the standard normal distribution Z gives:

$$P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le b\right) \approx P(Z \le b)$$

The normal distribution arises out of a means of samples from a discrete uniform distribution. Samples from *any* parent distribution give rise to the normal distribution in exactly the same way.

Notes

The graph above shaded in orange serving as an example probability density function is produced with the following R code:

```
plot(x=c(seq(0,5,0.01)), y=c(dlnorm(x=seq(0,5,0.01),meanlog=0,sdlog=1)), type="l", xlab='', ylab='', yaxt="n", xaxt="n", bty="n")
polygon(x=c(seq(0,2,0.01),2), y=c(dlnorm(x=seq(0,2,0.01),meanlog=0,sdlog=1),0), col="orange")
mtext('b', at=c(2), side=1)
text(0.6,0.2,"P(X≤b)")
abline(c(0,0),c(0,5))
```

Statistical analysis

Statistical analysis in R is performed by using many in-built functions. Most of these functions are part of the R base package. These functions take R vector as an input along with the arguments and give the result.

Mean

It is calculated by taking the sum of the values and dividing with the number of values in a data series.
The function **mean()** is used to calculate this in R.
The basic syntax for calculating mean in R is −

```
mean(x, trim = 0, na.rm = FALSE, ...)
```

Following is the description of the parameters used −
x is the input vector.
trim is used to drop some observations from both end of the sorted vector.
na.rm is used to remove the missing values from the input vector.

Example

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Find Mean.
result.mean <- mean(x)
print(result.mean)
```

When we execute the above code, it produces the following result −

```
[1] 8.22
```

Applying Trim Option

When trim parameter is supplied, the values in the vector get sorted and then the required numbers of observations are dropped from calculating the mean.
When trim = 0.3, 3 values from each end will be dropped from the calculations to find mean.
In this case the sorted vector is (−21, −5, 2, 3, 4.2, 7, 8, 12, 18, 54) and the values removed from the vector for calculating mean are (−21,−5,2) from left and (12,18,54) from right.

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Find Mean.
result.mean <-  mean(x,trim = 0.3)
print(result.mean)
```

When we execute the above code, it produces the following result −

```
[1] 5.55
```

Applying NA Option

If there are missing values, then the mean function returns NA.
To drop the missing values from the calculation use na.rm = TRUE which means remove the NA values.

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5,NA)
# Find mean.
result.mean <-  mean(x)
print(result.mean)
# Find mean dropping NA values.
```

```
result.mean <-  mean(x,na.rm = TRUE)
print(result.mean)
```
When we execute the above code, it produces the following result −
```
[1] NA
[1] 8.22
```

Median

The middle most value in a data series is called the median. The **median()** function is used in R to calculate this value.

The basic syntax for calculating median in R is −
```
median(x, na.rm = FALSE)
```

Following is the description of the parameters used −

x is the input vector.

na.rm is used to remove the missing values from the input vector.

Example
```
# Create the vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Find the median.
median.result <- median(x)
print(median.result)
```

When we execute the above code, it produces the following result −
```
[1] 5.6
```

Mode

The mode is the value that has highest number of occurrences in a set of data. Unlike mean and median, mode can have both numeric and character data.

R does not have a standard in-built function to calculate mode. So we create a user function to calculate mode of a data set in R. This function takes the vector as input and gives the mode value as output.

Example
```
# Create the function.
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
# Create the vector with numbers.
v <- c(2,1,2,3,1,2,3,4,1,5,5,3,2,3)
# Calculate the mode using the user function.
result <- getmode(v)
print(result)
# Create the vector with characters.
charv <- c("o","it","the","it","it")
# Calculate the mode using the user function.
result <- getmode(charv)
print(result)
```
When we execute the above code, it produces the following result −
```
[1] 2
[1] "it"
```

Regression analysis

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable.

Linear Regression

In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is 1. Mathematically a linear relationship represents a straight line when plotted as a graph. A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.

The general mathematical equation for a linear regression is −
y = ax + b

Following is the description of the parameters used −
y is the response variable.
x is the predictor variable.
a and **b** are constants which are called the coefficients.

Steps to Establish a Regression

A simple example of regression is predicting weight of a person when his height is known. To do this we need to have the relationship between height and weight of a person.

The steps to create the relationship are −

- Carry out the experiment of gathering a sample of observed values of height and corresponding weight.
- Create a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equation using these
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the weight of new persons, use the **predict()** function in R.

Input Data
Below is the sample data representing the observations −
Values of height
151, 174, 138, 186, 128, 136, 179, 163, 152, 131

Values of weight.
63, 81, 56, 91, 47, 57, 76, 72, 62, 48

lm() Function
This function creates the relationship model between the predictor and the response variable.

The basic syntax for **lm()** function in linear regression is −
lm(formula,data)

Following is the description of the parameters used −
formula is a symbol presenting the relation between x and y.
data is the vector on which the formula will be applied.

Create Relationship Model & get the Coefficients

x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)

```
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
# Apply the lm() function.
relation <- lm(y~x)
print(relation)
```

When we execute the above code, it produces the following result −

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)           x
   -38.4551      0.6746
```

Get the Summary of the Relationship

```
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
# Apply the lm() function.
relation <- lm(y~x)
print(summary(relation))
```

When we execute the above code, it produces the following result −

```
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q     Median      3Q     Max
-6.3002    -1.6629  0.0412    1.8944  3.9775
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.45509    8.04901  -4.778  0.00139 **
x             0.67461    0.05191  12.997 1.16e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.253 on 8 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06
```

predict() Function

The basic syntax for predict() in linear regression is −

```
predict(object, newdata)
```

Following is the description of the parameters used −

object is the formula which is already created using the lm() function.

newdata is the vector containing the new value for predictor variable.

Predict the weight of new persons

```
# The predictor vector.
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
```

# The resposne vector.
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
# Apply the lm() function.
relation <- lm(y~x)
# Find weight of a person with height 170.
a <- data.frame(x = 170)
result <- predict(relation,a)
print(result)

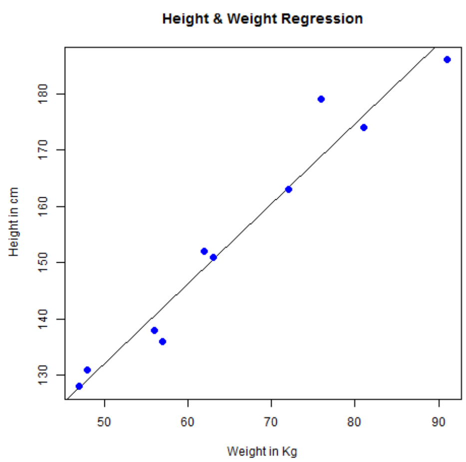
When we execute the above code, it produces the following result −

```
       1
76.22869
```

Visualize the Regression Graphically

```
# Create the predictor and response variable.
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
relation <- lm(y~x)
# Give the chart file a name.
png(file = "linearregression.png")
# Plot the chart.
plot(y,x,col = "blue",main = "Height & Weight Regression",
abline(lm(x~y)),cex = 1.3,pch = 16,xlab = "Weight in Kg",ylab = "Height in cm")
```

When we execute the above code, it produces the following result −



Multiple regression

Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor and one response variable, but in multiple regression we have more than one predictor variable and one response variable.

The general mathematical equation for multiple regression is −

y = a + b1x1 + b2x2 +...bnxn

Following is the description of the parameters used −

y is the response variable.

a, b1, b2...bn are the coefficients.

x1, x2, ...xn are the predictor variables.

We create the regression model using the **lm()** function in R. The model determines the value of the coefficients using the input data. Next we can predict the value of the response variable for a given set of predictor variables using these coefficients.

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for **lm()** function in multiple regression is −

lm(y ~ x1+x2+x3...,data)

Following is the description of the parameters used −

formula is a symbol presenting the relation between the response variable and predictor variables.

data is the vector on which the formula will be applied.

Example

Input Data

Consider the data set "mtcars" available in the R environment. It gives a comparison between different car models in terms of mileage per gallon (mpg), cylinder displacement("disp"), horse power("hp"), weight of the car("wt") and some more parameters.

The goal of the model is to establish the relationship between "mpg" as a response variable with "disp","hp" and "wt" as predictor variables. We create a subset of these variables from the mtcars data set for this purpose.

```
input <- mtcars[,c("mpg","disp","hp","wt")]
print(head(input))
```

When we execute the above code, it produces the following result −

	mpg	disp	hp	wt
Mazda RX4	21.0	160	110	2.620
Mazda RX4 Wag	21.0	160	110	2.875
Datsun 710	22.8	108	93	2.320
Hornet 4 Drive	21.4	258	110	3.215
Hornet Sportabout	18.7	360	175	3.440
Valiant	18.1	225	105	3.460

Create Relationship Model & get the Coefficients

```
input <- mtcars[,c("mpg","disp","hp","wt")]
# Create the relationship model.
```

```
model <- lm(mpg~disp+hp+wt, data = input)

# Show the model.
print(model)
# Get the Intercept and coefficients as vector elements.
cat("# # # # The Coefficient Values # # # ","\n")
a <- coef(model)[1]
print(a)
Xdisp <- coef(model)[2]
Xhp <- coef(model)[3]
Xwt <- coef(model)[4]
print(Xdisp)
print(Xhp)
print(Xwt)
```

When we execute the above code, it produces the following result −

```
Call:
lm(formula = mpg ~ disp + hp + wt, data = input)

Coefficients:
(Intercept)        disp          hp          wt
  37.105505     -0.000937       -0.031157    -3.800891

# # # # The Coefficient Values # # #
(Intercept)
   37.10551
         disp
-0.0009370091
         hp
-0.03115655
       wt
-3.800891
```

<u>Create Equation for Regression Model</u>

Based on the above intercept and coefficient values, we create the mathematical equation.

Y = a+Xdisp.x1+Xhp.x2+Xwt.x3

or

Y = 37.15+(-0.000937)*x1+(-0.0311)*x2+(-3.8008)*x3

<u>Apply Equation for predicting New Values</u>

We can use the regression equation created above to predict the mileage when a new set of values for displacement, horse power and weight is provided.

For a car with disp = 221, hp = 102 and wt = 2.91 the predicted mileage is −

Y = 37.15+(-0.000937)*221+(-0.0311)*102+(-3.8008)*2.91 = 22.7104

R - Logistic Regression

The Logistic Regression is a regression model in which the response variable (dependent variable) has categorical values such as True/False or 0/1. It actually measures the probability of a binary response as the value of response variable based on the mathematical equation relating it with the predictor variables.

The general mathematical equation for logistic regression is −

y = 1/(1+e^-(a+b1x1+b2x2+b3x3+...))

Following is the description of the parameters used −

y is the response variable.

x is the predictor variable.

a and **b** are the coefficients which are numeric constants.

The function used to create the regression model is the **glm()** function.

The basic syntax for **glm()** function in logistic regression is −

glm(formula,data,family)

Following is the description of the parameters used −

formula is the symbol presenting the relationship between the variables.

data is the data set giving the values of these variables.

family is R object to specify the details of the model. It's value is binomial for logistic regression.

Example

The in-built data set "mtcars" describes different models of a car with their various engine specifications. In "mtcars" data set, the transmission mode (automatic or manual) is described by the column am which is a binary value (0 or 1). We can create a logistic regression model between the columns "am" and 3 other columns - hp, wt and cyl.

```
# Select some columns form mtcars.
input <- mtcars[,c("am","cyl","hp","wt")]
print(head(input))
```

When we execute the above code, it produces the following result −

	am	cyl	hp	wt
Mazda RX4	1	6	110	2.620
Mazda RX4 Wag	1	6	110	2.875
Datsun 710	1	4	93	2.320
Hornet 4 Drive	0	6	110	3.215
Hornet Sportabout	0	8	175	3.440
Valiant	0	6	105	3.460

Create Regression Model

We use the **glm()** function to create the regression model and get its summary for analysis.

```
input <- mtcars[,c("am","cyl","hp","wt")]
am.data = glm(formula = am ~ cyl + hp + wt, data = input, family = binomial)
print(summary(am.data))
```

When we execute the above code, it produces the following result −

```
Call:
glm(formula = am ~ cyl + hp + wt, family = binomial, data = input)

Deviance Residuals:
     Min        1Q      Median       3Q       Max
-2.17272     -0.14907  -0.01464     0.14116   1.27641

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 19.70288    8.11637   2.428   0.0152 *
cyl          0.48760    1.07162   0.455   0.6491
hp           0.03259    0.01886   1.728   0.0840 .
wt          -9.14947    4.15332  -2.203   0.0276 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.2297  on 31  degrees of freedom
Residual deviance:  9.8415  on 28  degrees of freedom
AIC: 17.841

Number of Fisher Scoring iterations: 8
```

Conclusion

In the summary as the p-value in the last column is more than 0.05 for the variables "cyl" and "hp", we consider them to be insignificant in contributing to the value of the variable "am". Only weight (wt) impacts the "am" value in this regression model.

R - Normal Distribution

In a random collection of data from independent sources, it is generally observed that the distribution of data is normal. Which means, on plotting a graph with the value of the variable in the horizontal axis and the count of the values in the vertical axis we get a bell shape curve. The center of the curve represents the mean of the data set. In the graph, fifty percent of values lie to the left of the mean and the other fifty percent lie to the right of the graph. This is referred as normal distribution in statistics.

R has four in built functions to generate normal distribution. They are described below.

```
dnorm(x, mean, sd)
pnorm(x, mean, sd)
qnorm(p, mean, sd)
rnorm(n, mean, sd)
```

Following is the description of the parameters used in above functions −

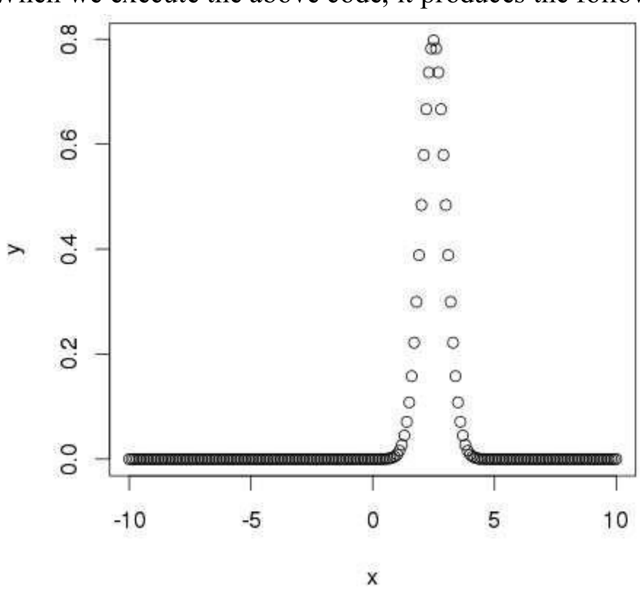
x is a vector of numbers.
p is a vector of probabilities.
n is number of observations(sample size).
mean is the mean value of the sample data. It's default value is zero.
sd is the standard deviation. It's default value is 1.

dnorm()

This function gives height of the probability distribution at each point for a given mean and standard deviation.

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.1.
x <- seq(-10, 10, by = .1)
# Choose the mean as 2.5 and standard deviation as 0.5.
y <- dnorm(x, mean = 2.5, sd = 0.5)
# Give the chart file a name.
png(file = "dnorm.png")
plot(x,y)
```

When we execute the above code, it produces the following result −

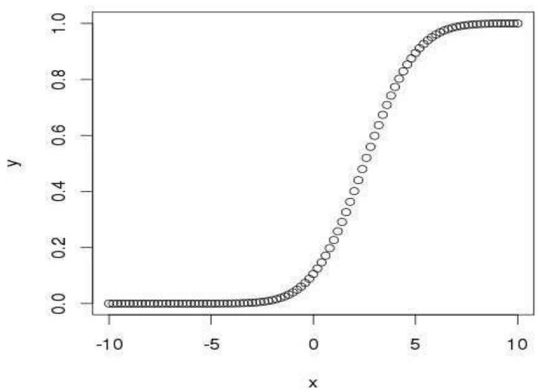


pnorm()

This function gives the probability of a normally distributed random number to be less that the value of a given number. It is also called "Cumulative Distribution Function".

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.2.
x <- seq(-10,10,by = .2)
# Choose the mean as 2.5 and standard deviation as 2.
y <- pnorm(x, mean = 2.5, sd = 2)
# Give the chart file a name.
png(file = "pnorm.png")
# Plot the graph.
plot(x,y)
```

When we execute the above code, it produces the following result −

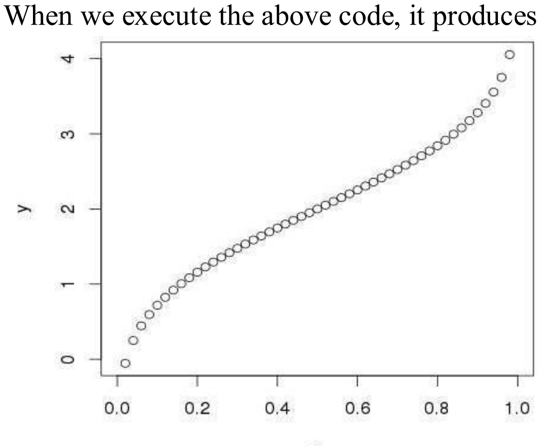


qnorm()

This function takes the probability value and gives a number whose cumulative value matches the probability value.

```
# Create a sequence of probability values incrementing by 0.02.
x <- seq(0, 1, by = 0.02)
# Choose the mean as 2 and standard deviation as 3.
y <- qnorm(x, mean = 2, sd = 1)
# Give the chart file a name.
png(file = "qnorm.png")
# Plot the graph.
plot(x,y)
```

When we execute the above code, it produces the following result −

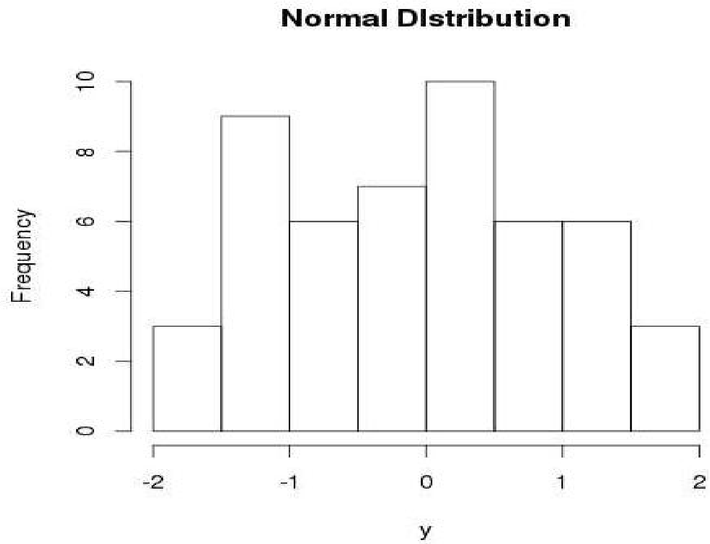


rnorm()

This function is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. We draw a histogram to show the distribution of the generated numbers.

```
# Create a sample of 50 numbers which are normally distributed.
y <- rnorm(50)
# Give the chart file a name.
png(file = "rnorm.png")
# Plot the histogram for this sample.
hist(y, main = "Normal DIstribution")
```

When we execute the above code, it produces the following result −



R - Binomial Distribution

The binomial distribution model deals with finding the probability of success of an event which has only two possible outcomes in a series of experiments. For example, tossing of a coin always gives a head or a tail. The probability of finding exactly 3 heads in tossing a coin repeatedly for 10 times is estimated during the binomial distribution.

R has four in-built functions to generate binomial distribution. They are described below.

```
dbinom(x, size, prob)
pbinom(x, size, prob)
qbinom(p, size, prob)
rbinom(n, size, prob)
```

Following is the description of the parameters used −

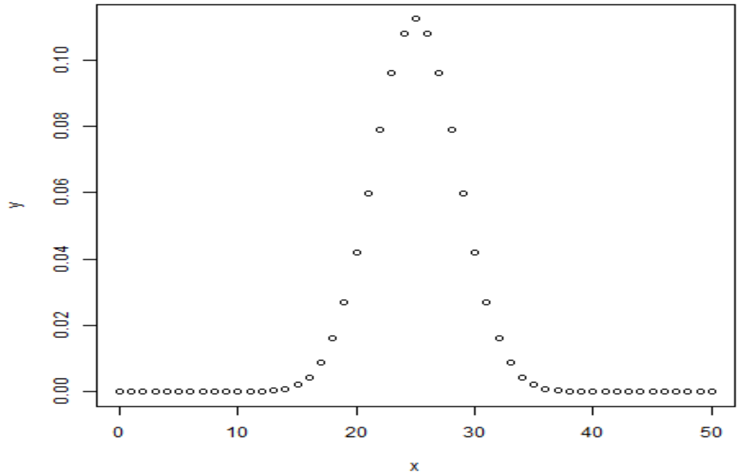
x is a vector of numbers.
p is a vector of probabilities.
n is number of observations.
size is the number of trials.
prob is the probability of success of each trial.

dbinom()

This function gives the probability density distribution at each point.

```
# Create a sample of 50 numbers which are incremented by 1.
x <- seq(0,50,by = 1)
# Create the binomial distribution.
y <- dbinom(x,50,0.5)
# Give the chart file a name.
png(file = "dbinom.png")
# Plot the graph for this sample.
plot(x,y)
```

When we execute the above code, it produces the following result −



pbinom()

This function gives the cumulative probability of an event. It is a single value representing the probability.

```
# Probability of getting 26 or less heads from a 51 tosses of a coin.
x <- pbinom(26,51,0.5)
```

```
print(x)
```

When we execute the above code, it produces the following result −

```
[1] 0.610116
```

qbinom()

This function takes the probability value and gives a number whose cumulative value matches the probability value.

```
# How many heads will have a probability of 0.25 will come out when a coin
# is tossed 51 times.
x <- qbinom(0.25,51,1/2)
print(x)
```

When we execute the above code, it produces the following result −

```
[1] 23
```

rbinom()

This function generates required number of random values of given probability from a given sample.

```
# Find 8 random values from a sample of 150 with probability of 0.4.
x <- rbinom(8,150,.4)
print(x)
```

When we execute the above code, it produces the following result −

```
[1] 58 61 59 66 55 60 61 67
```

R - Poisson Regression

Poisson Regression involves regression models in which the response variable is in the form of counts and not fractional numbers. For example, the count of number of births or number of wins in a football match series. Also the values of the response variables follow a Poisson distribution.

The general mathematical equation for Poisson regression is −

log(y) = a + b1x1 + b2x2 + bnxn.....

Following is the description of the parameters used −

y is the response variable.

a and **b** are the numeric coefficients.

x is the predictor variable.

The function used to create the Poisson regression model is the **glm()** function.

The basic syntax for **glm()** function in Poisson regression is −

glm(formula,data,family)

Following is the description of the parameters used in above functions −

formula is the symbol presenting the relationship between the variables.

data is the data set giving the values of these variables.

family is R object to specify the details of the model. It's value is 'Poisson' for Logistic Regression.

Example

We have the in-built data set "warpbreaks" which describes the effect of wool type (A or B) and tension (low, medium or high) on the number of warp breaks per loom. Let's consider "breaks" as the response variable which is a count of number of breaks. The wool "type" and "tension" are taken as predictor variables.

Input Data

```
input <- warpbreaks
print(head(input))
```

When we execute the above code, it produces the following result −

```
  breaks  wool  tension
1   26     A    L
2   30     A    L
3   54     A    L
4   25     A    L
5   70     A    L
6   52     A    L
```

Create Regression Model

```
output <-glm(formula = breaks ~ wool+tension, data = warpbreaks,    family = poisson)
print(summary(output))
```

When we execute the above code, it produces the following result −

```
Call:
glm(formula = breaks ~ wool + tension, family = poisson, data = warpbreaks)

Deviance Residuals:
    Min      1Q    Median      3Q     Max
```

```
 -3.6871  -1.6503  -0.4269    1.1902   4.2616

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.69196    0.04541  81.302  < 2e-16 ***
woolB       -0.20599    0.05157  -3.994 6.49e-05 ***
tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***
tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 297.37  on 53  degrees of freedom
Residual deviance: 210.39  on 50  degrees of freedom
AIC: 493.06

Number of Fisher Scoring iterations: 4
```

In the summary we look for the p-value in the last column to be less than 0.05 to consider an impact of the predictor variable on the response variable. As seen the wooltype B having tension type M and H have impact on the count of breaks.

Regression Modeling

Regression is a method to mathematically formulate relationship between variables that in due course can be used to estimate, interpolate and extrapolate.

Suppose we want to estimate the weight of individuals, which is influenced by height, diet, workout, etc. Here, *Weight* is the **predicted** variable. *Height*, *Diet*, *Workout* are **predictor** variables.

The predicted variable is a **dependant** variable in the sense that it depends on predictors. Predictors are also called as **independent** variables.

Regression reveals to what extent the predicted variable is affected by the predictors.
In other words, what amount of variation in predictors will result in variations of the predicted variable. The predicted variable is mathematically represented as Y. The predictor variables are represented as X1, X2, X3, etc. This mathematical relationship is often called the **regression model**.

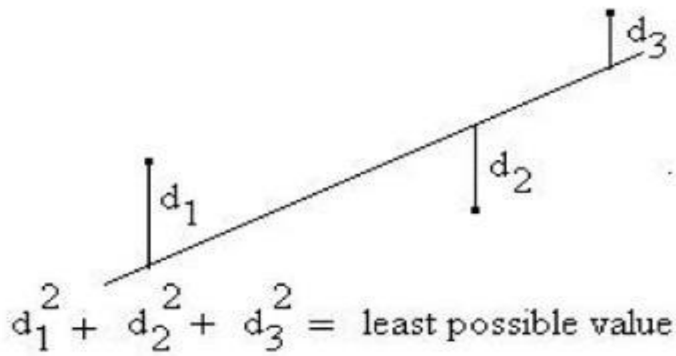
Regression is a branch of statistics. There are many types of regression. Regression is commonly used for prediction and forecasting.

What's a typical process for performing regression analysis?
First select a suitable predicted variable with acceptable measurement qualities such as reliability and validity. Likewise, select the predictors. When there's a single predictor, we call it **bivariate** analysis; anything more, we call it **multivariate** analysis.

Collect sufficient number of data points. Use a suitable estimation technique to arrive at the mathematical formula between predicted and predictor variables. No model is perfect. Hence, give error bounds.

Finally, assess the model's stability by applying it to different samples of the same population. When predictor variables are given for a new data point, estimate the predicted variable. If stable, the model's accuracy should not decrease. This process is called **model cross-validation**.

The **least-squares regression** line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.



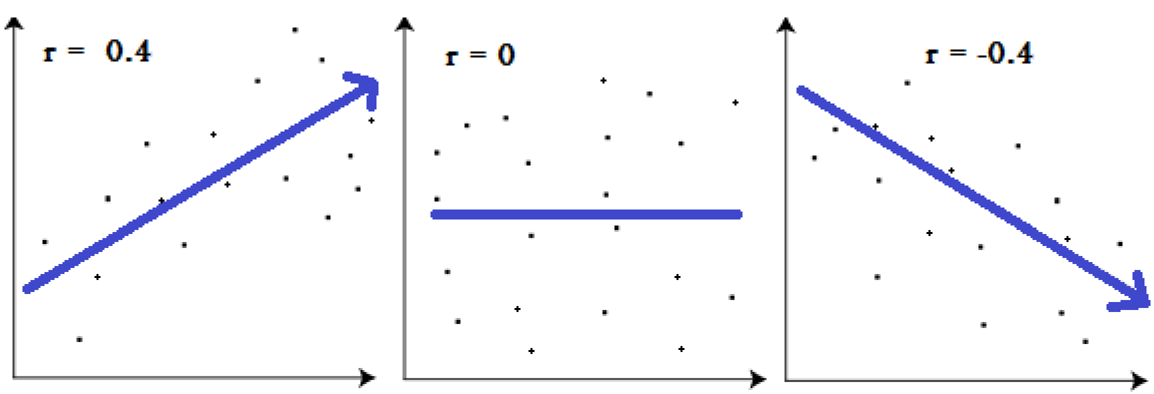
Least Squares is a term that signifies that the square of errors are at a minimum. The error is defined as the difference between observed value and predicted value. The objective of regression estimation is produce least squared errors as a result. When error approaches zero, we term it as *overfitting*.

Least Squares Method provides linear equations with unknowns that can be solved for any given data. The unknowns are regression parameters. The linear equations are called as *Normal Equations*. The normal equations are derived using calculus to minimize squared errors.

All other algorithms (*Artificial Neural Network (ANN), K-Nearest Neighbour (KNN)*, etc.) too attempt to minimize squared error unless the objective states otherwise.
Simply put, interpolation is estimation in familiar territory and extrapolation is estimation where not much of data is available due to various reasons—not collected or cannot be collected.

We can interpolate missing data points using regression. For instance, we want to estimate height given weight and data collection process missed out certain weights, we can use regression to interpolate. This missing data can estimated by other means too. The missing data estimation is called *imputation*.

The height and weight data is bound by nature and can be sourced. Say, we want to estimate future weight of an individual given historical weight variations of the individual. This is extrapolation. In regression, we call it *forecasting*. This is solved using a distinct set of techniques called as **Time Series Regression**.



Positive Correlation **No correlation** **Negative**

Correlation helps identify variables that can be applied for regression modelling. Correlations between each predictor and predicted variable are identified to decide on the predictors that need to be included in the model.

Correlation is defining the association between two variables. The effect of X (or X1, X2, X3...) on Y can be thus quantified:
Positive Correlation: Y goes up/down as X goes up/down. Correlation coefficient will be in the range [0,1].
Negative Correlation: Y goes up/down as X goes down/up. Correlation coefficient will be in the range [-1,0].
No Correlation: Y doesn't go up/down as X goes up/down. Correlation coefficient will be close to 0.

The equations below show the calculations sed to compute "r". However, you do not need to remember these equations. We will use R to do these calculations for us. Nevertheless, the equations give a sense of how "r" is computed.

$$r = \frac{Cov(X,Y)}{\sqrt{{s_x}^2 {s_y}^2}}$$

where Cov(X,Y) is the covariance, i.e., how far each observed (X,Y) pair is from the mean of X and the mean of Y, simultaneously, and and s_x^2 and s_y^2 are the sample variances for X and Y.
Cov (X,Y) is computed as:

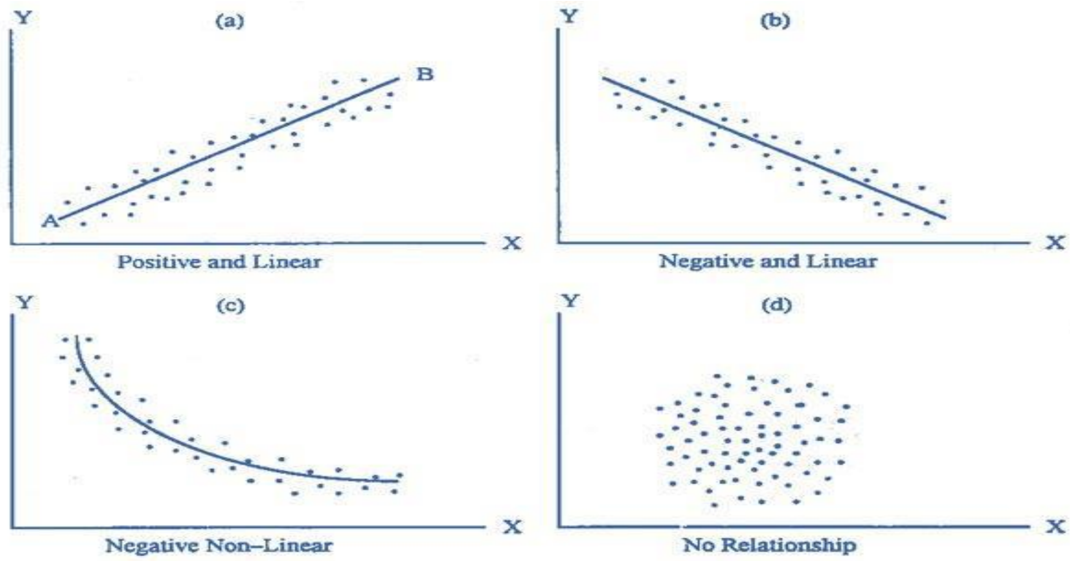
$$Cov\left(X, Y\right) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1}$$

You don't have to memorize or use these equations for hand calculations. Instead, we will use R to calculate correlation coefficients.
For example, we could use the following command to compute the correlation coefficient for AGE and TOTCHOL in a subset of the Framingham Heart Study as follows:
> cor(AGE,TOTCHOL)
[1] 0.2917043

What's the right interpretation of correlation coefficient?
The table below provides some guidelines for how to describe the strength of correlation coefficients, but these are just guidelines for description. Also, keep in mind that even weak correlations can be statistically significant.

Correlation Coefficient (r)	**Description (Rough Guideline)**
+1.0	Perfect positive + association
+0.8 to 1.0	Very strong + association
+0.6 to 0.8	Strong + association
+0.4 to 0.6	Moderate + association
+0.2 to 0.4	Weak + association
0.0 to +0.2	Very weak + or no association
0.0 to -0.2	Very weak - or no association
-0.2 to – 0.4	Weak - association
-0.4 to -0.6	Moderate - association
-0.6 to -0.8	Strong - association
-0.8 to -1.0	Very strong - association
-1.0	Perfect negative association

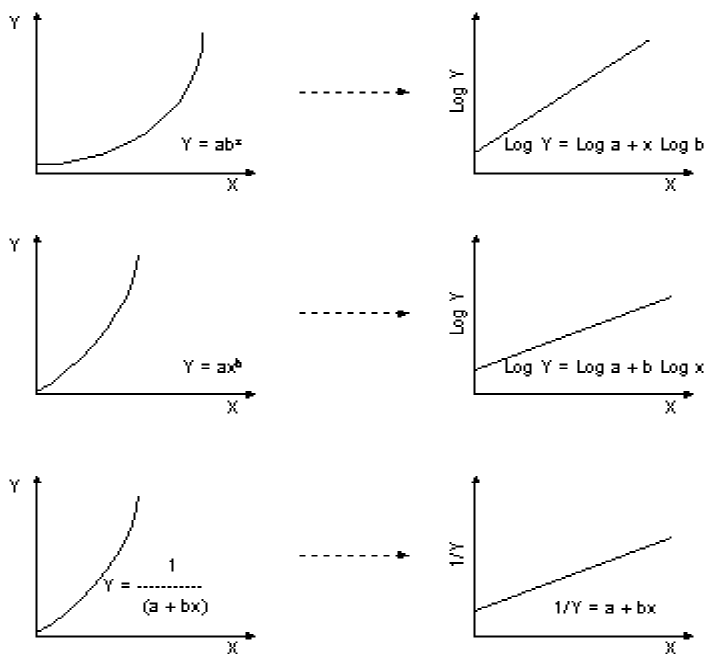
Examples of non-linear correlation



A non-linear correlation is where the relationship between the variables cannot be expressed by a straight line. We call this relationship *curvilinear*.

Non-linear relationship can exhibit monotonous positive, monotonous negative, or both patterns together.

How can we do data analysis when relationships are non-linear?



The correlation coefficient formula applies for only linear relationships. One common approach for non-linear correlations is to transform them into linear forms. If the relationship is curvilinear, we can apply transformations directly. Common transformations include logarithmic or inverse transformations.

If the relationship is non-linear but not curvilinear, we can split the data into distinct segments. Data within some segments may be linear. In other segments, if it's curvilinear, transformations can be applied to make them linear. Analysis is thus segment-wise, sometimes called **segmented regression**.

As an example, yield of mustard is not affected by soil salinity for low values. For salinity above a threshold, there's a negative linear relation. This dataset can be segmented at the threshold.

Causal relationship in regression

One study about college education, showed positive correlation between SAT scores of incoming students and their earnings when they graduate. Moreover, we can state that graduating from elite colleges (high SAT scores) had a role in higher salaries.

Causality or causation refers to the idea that variation in a predictor X *causes* variation in the predicted variable Y. This is distinct from regression, which is more about predicting Y based on its correlation with X. Regression does not claim that Y is caused by X.

Here are some possible examples of causality. High scores lead to higher earnings. Regular exercise results in better health. Current season influences power consumption. All pairs of variables that have causal relationship will exhibit significant correlation.

Sometimes correlations are purely coincidental. For example, non-commercial space launches and sociology doctorates awarded are completely unrelated but the image shows them to be strongly correlated. This is called a **Spurious Correlation**. This is a clear case where correlation does not imply causation.

Another example is when ice cream sales are positively correlated with violent crime. However, violent crime is not caused by ice cream sales. It so happens that there's a **confounding variable**, which in this case is weather. Hot weather influences both ice cream sales and violent crimes. It's therefore obvious that, correlation does not always imply causation

Correlation shouldn't be mistaken for causation. Look at the physical mechanism causing such a relationship. For example, is rain driving the sale of your product? Data may show a correlation. It need not be causal unless your product is an umbrella. However, proving causality is hard. At best, we can do randomized trials to establish causality.

Regression is a useful tool in either predictive or causal analysis. With the growth of Big Data, it's being used more often for predictive analysis.

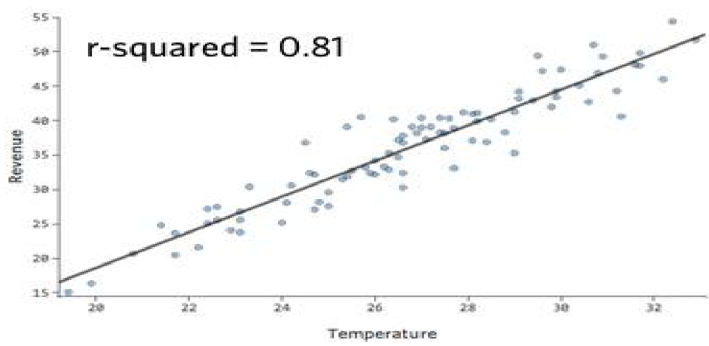
Regression model

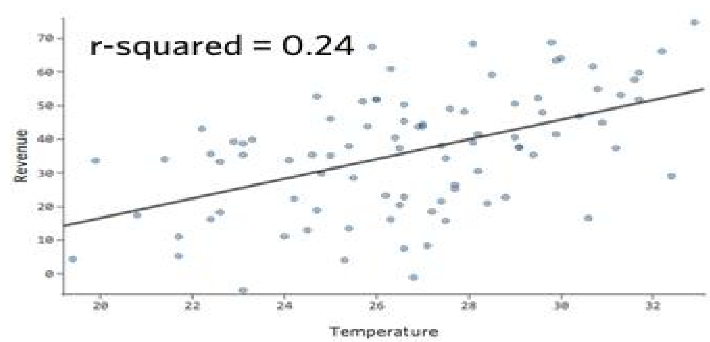
We call it a *model* when the relationship between variables is in a well-defined mathematical form: Y = f(X).

For instance, a linear relationship can be written as $f(X) = a + b_1X_1 + b_2X_2 + b_3X_3$, where 'a' is a constant and b1,b2,b3 are regression coefficients. a is constant effect while a unit change in X1, will result in b1 unit change in Y.

It's important to note that linearity is in terms of the coefficients, not in terms of predictor variables. For example, this model is still linear though it's quadratic in terms of X1: f(X)=a+b1X1+b2X2.

The accuracy of regression model is relative to base model. Called **R-Squared**, this measure is squared deviation from the expected value, which is mathematically defined below:





- For base model, the sum of squared deviation of actual value Y from mean value E(Y) is referred to as *Total Variance* or *SST (Total Sum of Squares)*. $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$

- For regression model, the sum of squared deviation of estimated value $\bar{y}$ from mean value E(Y) is referred to as *Explained Variance* or *SSR (Regression Sum of Squares)*. $SSR = \sum_{1=1}^{n}(\tilde{y_i} - \bar{y})^2$

- The accuracy of the model is called *R-Squared*. $R^2 = \frac{Explained\ Variance}{Total} = \frac{SSR}{SST}$

Higher the R^2, larger the explained variance and lower the unexplained. Hence, higher R^2 value is desired.
For example, if R^2=0.8, 80% of variation in data is explained by model.

Challenges with regression and how to overcome them
High multicollinearity is a challenge. It basically means one or more independent variables are highly linearly dependent on another independent variable. This makes it difficult to estimate the coefficients. One possible solution is to increase the sample size.

Another challenge is non-constant error variance, also called **heteroscedasticity**. An example of this when the observations "funnel out" as we move along the regression line. One solution is to use a Weighted Least Squares (WLS).

Regression assumes that errors from one observation are not related to other observations. This is often not true with time series data. **Autocorrelated errors** are therefore a challenge. One approach is to estimate the pattern in the errors and refine the regression model.

Another problem is **overfitting** that occurs when the model is "too well-trained". Such a model will not fit any other data. *Regularization* is the technique used to avoid overfitting. For parametric models, there are regression routines that address overfitting concerns. Lasso regression and ridge regression are a couple of such routines.

Heteroscedasticity in Regression Analysis

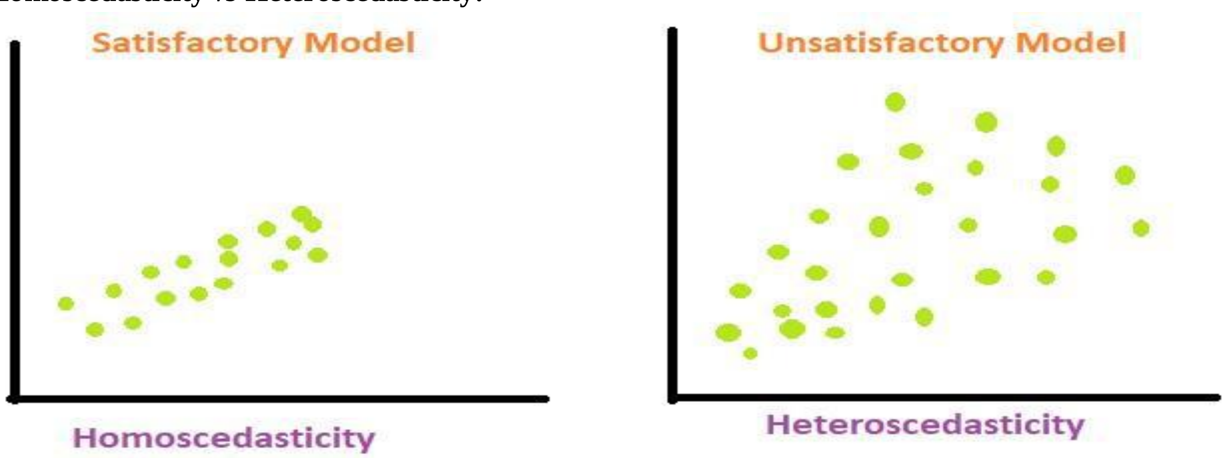
In Simple Linear Regression or Multiple Linear Regression we make some basic assumptions on the error term .

Assumptions:
1. Error has zero mean
2. Error has constant variance
3. Errors are uncorrelated
4. Errors are normally distributed

The *second assumption* is known as **Homoscedasticity** and therefore, the violation of this assumption is known as **Heteroscedasticity**.

When we perform regression, the data points are scattered around the fitted line. For a good regression model, the scattering should be as minimal as possible. When the scattering is uniform, the model is called homoscedastic. If not, the model is heteroscedastic.

Homoscedasticity *vs* Heteroscedasticity:



Therefore, in simple terms, we can define **heteroscedasticity** as the condition in which the variance of error term or the residual term in a regression model varies.
As you can see in the above diagram, in case of homoscedasticity, the data points are equally scattered while in case of heteroscedasticity the data points are not equally scattered.

What causes heteroscedasticity?
Most often the data itself is responsible for this kind of cone-shaped distribution. Sometimes it is very natural that the variance of the dependent variable varies and is not constant across the entire dataset.
For example, in the case of time-series data, the initial values of a dependent variable may be pretty much close together but during recent times, the scattering may be widened. The price of an apple in all the 50 states of the USA may be very much similar in the 1950s but it may vary wildly during recent times.
In other words, there may be greater variation among the recent data points than that of the earlier data.

Possible reasons of arising Heteroscedasticity:
1. Often occurs in those data sets which have a large range between the largest and the smallest observed values i.e. when there are outliers.
2. When model is not correctly specified.
3. If observations are mixed with different measures of scale.

4. When incorrect transformation of data is used to perform the regression.
5. Skewness in the distribution of a regressor, and may be some other sources.

Effects of Heteroscedasticity:

- As mentioned above that one of the assumption (assumption number 2) of linear regression is that there is no heteroscedasticity. Breaking this assumption means that *OLS (Ordinary Least Square)* estimators are not the *Best Linear Unbiased Estimator(BLUE)* and their variance is not the lowest of all other unbiased estimators.
- Estimators are no longer *best/efficient.*
- The tests of hypothesis (like t-test, F-test) are no longer valid due to the inconsistency in the co-variance matrix of the estimated regression coefficients.

Identifying Heteroscedasticity with residual plots:

As shown in the above figure, heteroscedasticity produces either outward opening funnel or outward closing funnel shape in residual plots.

Identifying Heteroscedasticity Through Statistical Tests:

The presence of heteroscedasticity can also be quantified using the algorithmic approach. There are some statistical tests or methods through which the presence or absence of heteroscedasticity can be established.

1. **The Breush – Pegan Test:** It tests whether the variance of the errors from regression is dependent on the values of the independent variables. In that case, heteroskedasticity is present.
2. **White test:** White test establishes whether the variance of the errors in a regression model is constant. To test for constant variance one undertakes an auxiliary regression analysis: this regresses the squared residuals from the original regression model onto a set of regressors that contain the original regressors along with their squares and cross-products.

Corrections for heteroscedasticity:

1. We can use different specification for the model.
2. Weighted Least Squares method is one of the common statistical method. This is the generalization of ordinary least square and linear regression in which the errors co-variance matrix is allowed to be different from an identity matrix.
3. **Use MINQUE:** The theory of ***Minimum Norm Quadratic Unbiased Estimation (MINQUE)*** involves three stages. First, defining a general class of potential estimators as quadratic functions of the observed data, where the estimators relate to a vector of model parameters. Secondly, specifying certain constraints on the desired properties of the estimators, such as unbiasedness and third, choosing the optimal estimator by minimizing a “norm” which measures the size of the covariance matrix of the estimators.

Heteroscedaticy makes a regression model less robust. Essentially, the predicted errors vary for different ranges of the independent variable, suggesting an imperfect model.

One should always check the residual plot after any regression analysis and ensure that the variability of the residuals doesn’t follow any pattern or in other words, the scattering of the residuals should be uniform across the regression line.

R - Analysis of Covariance

We use Regression analysis to create models which describe the effect of variation in predictor variables on the response variable. Sometimes, if we have a categorical variable with values like Yes/No or Male/Female etc. The simple regression analysis gives multiple results for each value of the categorical variable. In such scenario, we can study the effect of the categorical variable by using it along with the predictor variable and comparing the regression lines for each level of the categorical variable. Such an analysis is termed as **Analysis of Covariance** also called as **ANCOVA**.

Example

Consider the R built in data set mtcars. In it we observer that the field "am" represents the type of transmission (auto or manual). It is a categorical variable with values 0 and 1. The miles per gallon value(mpg) of a car can also depend on it besides the value of horse power("hp").

We study the effect of the value of "am" on the regression between "mpg" and "hp". It is done by using the **aov()** function followed by the **anova()** function to compare the multiple regressions.

Input Data

Create a data frame containing the fields "mpg", "hp" and "am" from the data set mtcars. Here we take "mpg" as the response variable, "hp" as the predictor variable and "am" as the categorical variable.

```
input <- mtcars[,c("am","mpg","hp")]
print(head(input))
```

When we execute the above code, it produces the following result −

	am	mpg	hp
Mazda RX4	1	21.0	110
Mazda RX4 Wag	1	21.0	110
Datsun 710	1	22.8	93
Hornet 4 Drive	0	21.4	110
Hornet Sportabout	0	18.7	175
Valiant	0	18.1	105

ANCOVA Analysis

We create a regression model taking "hp" as the predictor variable and "mpg" as the response variable taking into account the interaction between "am" and "hp".

Model with interaction between categorical variable and predictor variable

```
# Get the dataset.
input <- mtcars
# Create the regression model.
result <- aov(mpg~hp*am,data = input)
print(summary(result))
```

When we execute the above code, it produces the following result −

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
hp	1	678.4	678.4	77.391	1.50e-09 ***
am	1	202.2	202.2	23.072	4.75e-05 ***
hp:am	1	0.0	0.0	0.001	0.981
Residuals	28	245.4	8.8		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

This result shows that both horse power and transmission type has significant effect on miles per gallon as the p value in both cases is less than 0.05. But the interaction between these two variables is not significant as the p-value is more than 0.05.

Model without interaction between categorical variable and predictor variable

```
# Get the dataset.
input <- mtcars
# Create the regression model.
result <- aov(mpg~hp+am,data = input)
print(summary(result))
```

When we execute the above code, it produces the following result −

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
hp	1	678.4	678.4	80.15	7.63e-10 ***
am	1	202.2	202.2	23.89	3.46e-05 ***
Residuals	29	245.4	8.5		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

This result shows that both horse power and transmission type has significant effect on miles per gallon as the p value in both cases is less than 0.05.

Comparing Two Models

Now we can compare the two models to conclude if the interaction of the variables is truly in-significant. For this we use the **anova()** function.

```
# Get the dataset.
input <- mtcars
# Create the regression models.
result1 <- aov(mpg~hp*am,data = input)
result2 <- aov(mpg~hp+am,data = input)
# Compare the two models.
print(anova(result1,result2))
```

When we execute the above code, it produces the following result −

Model 1: mpg ~ hp * am
Model 2: mpg ~ hp + am

	Res.Df	RSS	Df	Sum of Sq	F Pr(>F)
1	28	245.43			
2	29	245.44	-1	-0.0052515	6e-04 0.9806

As the p-value is greater than 0.05 we conclude that the interaction between horse power and transmission type is not significant. So the mileage per gallon will depend in a similar manner on the horse power of the car in both auto and manual transmission mode.

R - Chi Square Test

Chi-Square test is a statistical method to determine if two categorical variables have a significant correlation between them. Both those variables should be from same population and they should be categorical like − Yes/No, Male/Female, Red/Green etc.

For example, we can build a data set with observations on people's ice-cream buying pattern and try to correlate the gender of a person with the flavor of the ice-cream they prefer. If a correlation is found we can plan for appropriate stock of flavors by knowing the number of gender of people visiting.

The function used for performing chi-Square test is chisq.test().

The basic syntax for creating a chi-square test in R is −

```
chisq.test(data)
```

Following is the description of the parameters used −

data is the data in form of a table containing the count value of the variables in the observation.

Example

We will take the Cars93 data in the "MASS" library which represents the sales of different models of car in the year 1993.

```
library("MASS")
print(str(Cars93))
```

When we execute the above code, it produces the following result −

```
'data.frame':   93 obs. of  27 variables:
 $ Manufacturer      : Factor w/ 32 levels "Acura","Audi",..: 1 1 2 2 3 4 4 4 4 5 ...
 $ Model             : Factor w/ 93 levels "100","190E","240",..: 49 56 9 1 6 24 54 74 73 35 ...
 $ Type              : Factor w/ 6 levels "Compact","Large",..: 4 3 1 3 3 3 2 2 3 2 ...
 $ Min.Price         : num  12.9 29.2 25.9 30.8 23.7 14.2 19.9 22.6 26.3 33 ...
 $ Price             : num  15.9 33.9 29.1 37.7 30 15.7 20.8 23.7 26.3 34.7 ...
 $ Max.Price         : num  18.8 38.7 32.3 44.6 36.2 17.3 21.7 24.9 26.3 36.3 ...
 $ MPG.city          : int  25 18 20 19 22 22 19 16 19 16 ...
 $ MPG.highway       : int  31 25 26 26 30 31 28 25 27 25 ...
 $ AirBags           : Factor w/ 3 levels "Driver & Passenger",..: 3 1 2 1 2 2 2 2 2 2 ...
 $ DriveTrain        : Factor w/ 3 levels "4WD","Front",..: 2 2 2 2 3 2 2 3 2 2 ...
 $ Cylinders         : Factor w/ 6 levels "3","4","5","6",..: 2 4 4 4 2 2 4 4 4 5 ...
 $ EngineSize        : num  1.8 3.2 2.8 2.8 3.5 2.2 3.8 5.7 3.8 4.9 ...
 $ Horsepower        : int  140 200 172 172 208 110 170 180 170 200 ...
 $ RPM               : int  6300 5500 5500 5500 5700 5200 4800 4000 4800 4100 ...
 $ Rev.per.mile      : int  2890 2335 2280 2535 2545 2565 1570 1320 1690 1510 ...
 $ Man.trans.avail   : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 1 1 1 ...
 $ Fuel.tank.capacity: num  13.2 18 16.9 21.1 21.1 16.4 18 23 18.8 18 ...
 $ Passengers        : int  5 5 5 6 4 6 6 6 5 6 ...
 $ Length            : int  177 195 180 193 186 189 200 216 198 206 ...
 $ Wheelbase         : int  102 115 102 106 109 105 111 116 108 114 ...
 $ Width             : int  68 71 67 70 69 69 74 78 73 73 ...
 $ Turn.circle       : int  37 38 37 37 39 41 42 45 41 43 ...
```

```
 $ Rear.seat.room    : num  26.5 30 28 31 27 28 30.5 30.5 26.5 35 ...
 $ Luggage.room      : int  11 15 14 17 13 16 17 21 14 18 ...
 $ Weight            : int  2705 3560 3375 3405 3640 2880 3470 4105 3495 3620 ...
 $ Origin            : Factor w/ 2 levels "USA","non-USA": 2 2 2 2 2 1 1 1 1 1 ...
 $ Make              : Factor w/ 93 levels "Acura Integra",..: 1 2 4 3 5 6 7 9 8 10 ...
```

The above result shows the dataset has many Factor variables which can be considered as categorical variables. For our model we will consider the variables "AirBags" and "Type". Here we aim to find out any significant correlation between the types of car sold and the type of Air bags it has. If correlation is observed we can estimate which types of cars can sell better with what types of air bags.

```
# Load the library.
library("MASS")
# Create a data frame from the main data set.
car.data <- data.frame(Cars93$AirBags, Cars93$Type)
# Create a table with the needed variables.
car.data = table(Cars93$AirBags, Cars93$Type)
print(car.data)
# Perform the Chi-Square test.
print(chisq.test(car.data))
```

When we execute the above code, it produces the following result −

```
                     Compact Large Midsize Small Sporty Van
  Driver & Passenger       2     4       7     0      3   0
  Driver only              9     7      11     5      8   3
  None                     5     0       4    16      3   6

        Pearson's Chi-squared test

data:  car.data
X-squared = 33.001, df = 10, p-value = 0.0002723

Warning message:
In chisq.test(car.data) : Chi-squared approximation may be incorrect
```

Conclusion

The result shows the p-value of less than 0.05 which indicates a string correlation.

T-Test Approach in R Programming

Suppose a businessman with two sweet shops in a town wants to check if the average number of sweets sold in a day in both the stores is the same or not.

So, the businessman takes the average number of sweets sold to 15 random people in the respective shops. He found out that the first shop sold 30 sweets on average whereas the second shop sold 40. So, from the owner's point of view, the second shop was doing better business than the former. But the thing to notice is that the data-set is based on a mere number of random people and they cannot represent all the customers. This is where T-testing comes into play it helps us to understand that the difference between the two means is real or simply by chance.

Mathematically, what the t-test does is, take a sample from both sets and establish the problem assuming a null hypothesis that the two means are the same.

Classification of T-tests

- One Sample T-test
- Two sample T-test
- Paired sample T-test

One Sample T-test

The One-Sample T-Test is used to test the statistical difference between a sample mean and a known or assumed/hypothesized value of the mean in the population.

So, for performing a one-sample t-test in R, we would use the syntax t.test(y, mu = 0) where x is the name of the variable of interest and mu is set equal to the mean specified by the null hypothesis.

```
set.seed(0)
sweetSold <- c(rnorm(50, mean = 140, sd = 5))
t.test(sweetSold, mu = 150) # Ho: mu = 150
```

Output:

```
        One Sample t-test

data:  sweetSold
t = -15.249, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 150
95 percent confidence interval:
 138.8176 141.4217
sample estimates:
mean of x
 140.1197
```

Two sample T-test

It is used to help us to understand that the difference between the two means is real or simply by chance. The general form of the test is t.test(y1, y2, paired=FALSE). By default, R assumes that the variances of y1 and y2 are unequal, thus defaulting to Welch's test. To toggle this, we use the flag var.equal=TRUE.

For Example:

```
set.seed(0)
shopOne <- rnorm(50, mean = 140, sd = 4.5)
shopTwo <- rnorm(50, mean = 150, sd = 4)
```

t.test(shopOne, shopTwo, var.equal = TRUE)

Output:

```
        Two Sample t-test

data:  shopOne and shopTwo
t = -12.883, df = 98, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.384388  -8.345285
sample estimates:
mean of x mean of y
 140.5360  150.4008
```

Paired Sample T-test

This is a statistical procedure that is used to determine whether the mean difference between two sets of observations is zero. In a paired sample t-test, each subject is measured two times, resulting in pairs of observations.
The test is run using the syntax t.test(y1, y2, paired=TRUE)
For Example:
set.seed(2820)

sweetOne <- c(rnorm(100, mean = 14, sd = 0.3))
sweetTwo <- c(rnorm(100, mean = 13, sd = 0.2))

t.test(sweetOne, sweetTwo, paired = TRUE)

Output:

```
        Paired t-test

data:  sweetOne and sweetTwo
t = 29.31, df = 99, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9892738 1.1329434
sample estimates:
mean of the differences
               1.061109
```

Fisher's F-Test in R Programming

Fisher's F test calculates the ratio between the larger variance and the smaller variance.
We use the F test when we want to check where means of three or more groups are different or not.
F-test is used to assess whether the variances of two populations (A and B) are equal.
The method is simple; it consists of taking the ratio between the larger variance and the smaller variance.
var.test() function in R Programming performs an F-test between 2 normal populations with the hypothesis that variances of the 2 populations are equal.

Formula for Fisher's F-Test

F = Larger Sample Variance / Smaller Sample Variance

Implementation in R

- To test the equality of variances between the two sample use **var.test(x, y)**
- To compare two variance use **var.test(x, y, alternative = "two.sided")**

Syntax:
var.test(x, y, alternative = "two.sided")
Parameters:
x, y: *numeric vectors*
alternative: *a character string specifying the alternative hypothesis.*

Example 1:

Let we have two samples x, y. The R function var.test() can be used to compare two variances as follow:

```
# Taking two samples
x <- rnorm(249, mean = 20)
y <- rnorm(79, mean = 30)
# var test in R
var.test(x, y, alternative = "two.sided")
```

Output:

```
F test to compare two variances
data:  x and y
F = 0.88707, num df = 248, denom df = 78, p-value = 0.4901
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.6071405 1.2521004
sample estimates:
ratio of variances
     0.8870677
```

It returns the following:

1. the value of the F test statistic.
2. the degrees of the freedom of the F distribution of the test statistic.
3. the p-value of the test 0.4901
4. a confidence interval for the ratio of the population variances.
5. the ratio of the sample variances 0.8870677

The p-value of F-test is p = 0.4901 which is greater than the alpha level 0.05. In conclusion, there is no difference between the two samples.

Example 2:
Let we have two random samples from two random population. Test whether two population have same variance.

```
# Taking two random samples
A = c(16, 17, 25, 26, 32,34, 38, 40, 42)
B = c(600, 590, 590, 630, 610, 630)
# var test in R
var.test(A, B, alternative = "two.sided")
```

Output:

```
        F test to compare two variances
data:  A and B
F = 0.27252, num df = 8, denom df = 5, p-value = 0.1012
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.04033118 1.31282683
sample estimates:
ratio of variances
        0.2725248
```

It returns the following:

1. the value of the F test statistic.
2. the degrees of the freedom of the F distribution of the test statistic.
3. the p-value of the test 0.1012
4. 95% confidence interval for the ratio of the population variances.
5. the ratio of the sample variances 0.2725248

The p-value of F-test is p = 0.1012 which is greater than the alpha level 0.05. In conclusion, there is no difference between the two samples.

Example 3:
Let we have two random sample.

```
# Taking two random samples
x = c(25, 29, 35, 46, 58, 66, 68)
y = c(14, 16, 24, 28, 32, 35,
                    37, 42, 43, 45, 47)
# var test in R
var.test(x, y)
```

Output:

```
F test to compare two variances
data:  x and y
F = 2.4081, num df = 6, denom df = 10, p-value = 0.2105
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
```

0.5913612 13.1514157
sample estimates:
ratio of variances
2.4081

It returns the following:

1. the value of the F test statistic.
2. the degrees of the freedom of the F distribution of the test statistic.
3. the p-value of the test 0.2105
4. 95% confidence interval for the ratio of the population variances.
5. the ratio of the sample variances 2.4081

The p-value of F-test is p = 0.2105 which is greater than the alpha level 0.05. In conclusion, there is no difference between the two samples.

R - Time Series Analysis

Time series is a series of data points in which each data point is associated with a timestamp. A simple example is the price of a stock in the stock market at different points of time on a given day. Another example is the amount of rainfall in a region at different months of the year. R language uses many functions to create, manipulate and plot the time series data. The data for the time series is stored in an R object called **time-series object**. It is also a R data object like a vector or data frame.

The time series object is created by using the **ts()** function.

The basic syntax for **ts()** function in time series analysis is −

```
timeseries.object.name <-  ts(data, start, end, frequency)
```

Following is the description of the parameters used −

data is a vector or matrix containing the values used in the time series.

start specifies the start time for the first observation in time series.

end specifies the end time for the last observation in time series.

frequency specifies the number of observations per unit time.

Except the parameter "data" all other parameters are optional.

Example

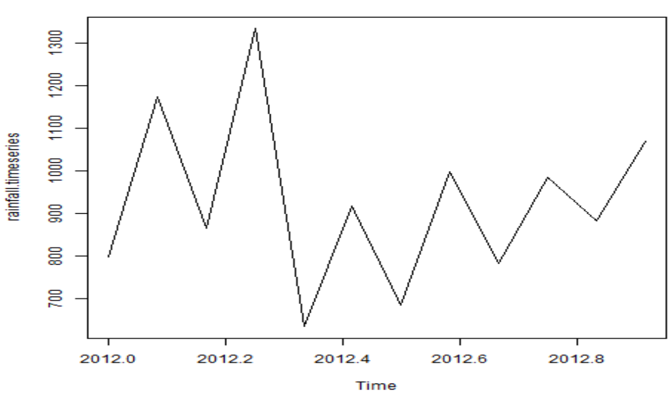
Consider the annual rainfall details at a place starting from January 2012. We create an R time series object for a period of 12 months and plot it.

```
# Get the data points in form of a R vector.
rainfall <- c(799,1174.8,865.1,1334.6,635.4,918.5,685.5,998.6,784.2,985,882.8,1071)
# Convert it to a time series object.
rainfall.timeseries <- ts(rainfall,start = c(2012,1),frequency = 12)
# Print the timeseries data.
print(rainfall.timeseries)
# Give the chart file a name.
png(file = "rainfall.png")
# Plot a graph of the time series.
plot(rainfall.timeseries)
```

When we execute the above code, it produces the following result and chart −

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2012	799.0	1174.8	865.1	1334.6	635.4	918.5	685.5	998.6	784.2	985.0	882.8	1071.0

The Time series chart −



Different Time Intervals

The value of the **frequency** parameter in the ts() function decides the time intervals at which the data points are measured. A value of 12 indicates that the time series is for 12 months. Other values and its meaning is as below −

frequency = 12 pegs the data points for every month of a year.
frequency = 4 pegs the data points for every quarter of a year.
frequency = 6 pegs the data points for every 10 minutes of an hour.
frequency = 24*6 pegs the data points for every 10 minutes of a day.

Multiple Time Series

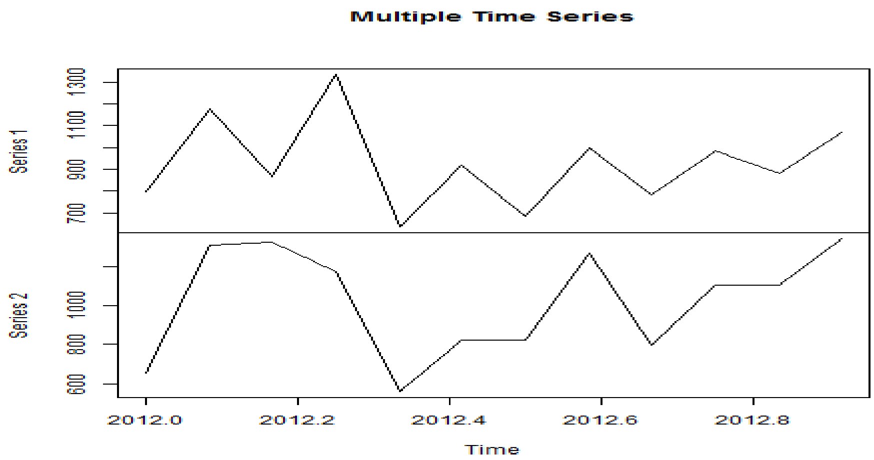
We can plot multiple time series in one chart by combining both the series into a matrix.

```
# Get the data points in form of a R vector.
rainfall1 <- c(799,1174.8,865.1,1334.6,635.4,918.5,685.5,998.6,784.2,985,882.8,1071)
rainfall2 <- c(655,1306.9,1323.4,1172.2,562.2,824,822.4,1265.5,799.6,1105.6,1106.7,1337.8)

# Convert them to a matrix.
combined.rainfall <-  matrix(c(rainfall1,rainfall2),nrow = 12)
# Convert it to a time series object.
rainfall.timeseries <- ts(combined.rainfall,start = c(2012,1),frequency = 12)
# Print the timeseries data.
print(rainfall.timeseries)
# Give the chart file a name.
png(file = "rainfall_combined.png")
# Plot a graph of the time series.
plot(rainfall.timeseries, main = "Multiple Time Series")
```

When we execute the above code, it produces the following result and chart −

	Series 1	Series 2
Jan 2012	799.0	655.0
Feb 2012	1174.8	1306.9
Mar 2012	865.1	1323.4
Apr 2012	1334.6	1172.2
May 2012	635.4	562.2
Jun 2012	918.5	824.0
Jul 2012	685.5	822.4
Aug 2012	998.6	1265.5
Sep 2012	784.2	799.6
Oct 2012	985.0	1105.6
Nov 2012	882.8	1106.7
Dec 2012	1071.0	1337.8

The Multiple Time series chart −



Forecasting is a technique that is popularly used in the field of machine learning for making business predictions. Companies use past time series forecasts and make business decisions for the future.

What is Time Series Forecasting?
Time series forecasting is the method of exploring and analyzing time-series data recorded or collected over a set period of time. This technique is used to forecast values and make future predictions. Not all data that have time values or date values as its features can be considered as a time series data. Any data fit for time series forecasting should consist of observations over a regular, continuous interval.

The **forecast** package in R (which is loaded automatically whenever you load the **fpp2** package) is used.

Many functions, including meanf(), naive(), snaive() and rwf(), produce output in the form of a forecast object (i.e., an object of class forecast). This allows other functions (such as autoplot()) to work consistently across a range of forecasting models.

Objects of class forecast contain information about the forecasting method, the data used, the point forecasts obtained, prediction intervals, residuals and fitted values. There are several functions designed to work with these objects including autoplot(), summary() and print().

The forecast() function works with many different types of inputs. It generally takes a time series or time series model as its main argument, and produces forecasts appropriately.

Eg:
```
forecast(ausbeer, h=4)
#>          Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
#> 2010 Q3          404.6 385.9 423.3 376.0 433.3
#> 2010 Q4          480.4 457.5 503.3 445.4 515.4
#> 2011 Q1          417.0 396.5 437.6 385.6 448.4
#> 2011 Q2          383.1 363.5 402.7 353.1 413.1
```

Time Series and Forecasting
R has extensive facilities for analyzing time series data. The steps are the creation of a time series, seasonal decomposition, modeling with exponential and ARIMA models, and forecasting with the forecast package.

<u>Creating a time series</u>

The **ts()** function will convert a numeric vector into an R time series object. The format is **ts(***vector*, **start=, end=, frequency=)** where start and end are the times of the first and last observation and frequency is the number of observations per unit time (1=annual, 4=quartly, 12=monthly, etc.).

```
# save a numeric vector containing 72 monthly observations
# from Jan 2009 to Dec 2014 as a time series object
myts <- ts(myvector, start=c(2009, 1), end=c(2014, 12), frequency=12)

# subset the time series (June 2014 to December 2014)
myts2 <- window(myts, start=c(2014, 6), end=c(2014, 12))

# plot series
plot(myts)
```

Seasonal Decomposition

A time series with additive trend, seasonal, and irregular components can be decomposed using the **stl()** function. Note that a series with multiplicative effects can often by transformed into series with additive effects through a log transformation (i.e., *newts* <- **log(***myts***)**).

```
# Seasonal decomposition
fit <- stl(myts, s.window="period")
plot(fit)

# additional plots
monthplot(myts)
library(forecast)
seasonplot(myts)
```

Exponential Models

Both the **HoltWinters()** function in the base installation, and the **ets()** function in the forecast package, can be used to fit exponential models.

```
# simple exponential - models level
fit <- HoltWinters(myts, beta=FALSE, gamma=FALSE)
# double exponential - models level and trend
fit <- HoltWinters(myts, gamma=FALSE)
# triple exponential - models level, trend, and seasonal components
fit <- HoltWinters(myts)

# predictive accuracy
library(forecast)
accuracy(fit)

# predict next three future values
library(forecast)
forecast(fit, 3)
plot(forecast(fit, 3))
```

ARIMA Models

The **arima()** function can be used to fit an autoregressive integrated moving averages model. Other useful functions include:

lag(*ts*, *k***)**	lagged version of time series, shifted back *k* observations
diff(*ts*, **differences=***d***)**	difference the time series *d* times
ndiffs(*ts***)**	Number of differences required to achieve stationarity (from the forecast package)
acf(*ts***)**	autocorrelation function
pacf(*ts***)**	partial autocorrelation function
adf.test(*ts***)**	Augemented Dickey-Fuller test. Rejecting the null hypothesis suggests that a time series is stationary (from the tseries package)
Box.test(*x*, **type="Ljung-Box")**	Pormanteau test that observations in vector or time series *x* are independent

Note that the forecast package has somewhat nicer versions of **acf()** and **pacf()** called **Acf()** and Pacf() respectively.

```
# fit an ARIMA model of order P, D, Q
fit <- arima(myts, order=c(p, d, q)

# predictive accuracy
library(forecast)
accuracy(fit)

# predict next 5 observations
library(forecast)
forecast(fit, 5)
plot(forecast(fit, 5))
```

<u>Automated Forecasting</u>

The forecast package provides functions for the automatic selection of exponential and ARIMA models. The **ets()** function supports both additive and multiplicative models. The auto.arima() function can handle both seasonal and nonseasonal ARIMA models. Models are chosen to maximize one of several fit criteria.

```
library(forecast)
# Automated forecasting using an exponential model
fit <- ets(myts)

# Automated forecasting using an ARIMA model
fit <- auto.arima(myts)
```

Outlier

Outlier is an unusual observation that is not consistent with the remaining observations in a sample dataset.

The outliers in a dataset can come from the following possible sources,

- contaminated data samples
- data points from different population
- incorrect sampling methods,
- underlying significant treatment response (e.g. biological variation of samples),
- error in data collection or analysis.

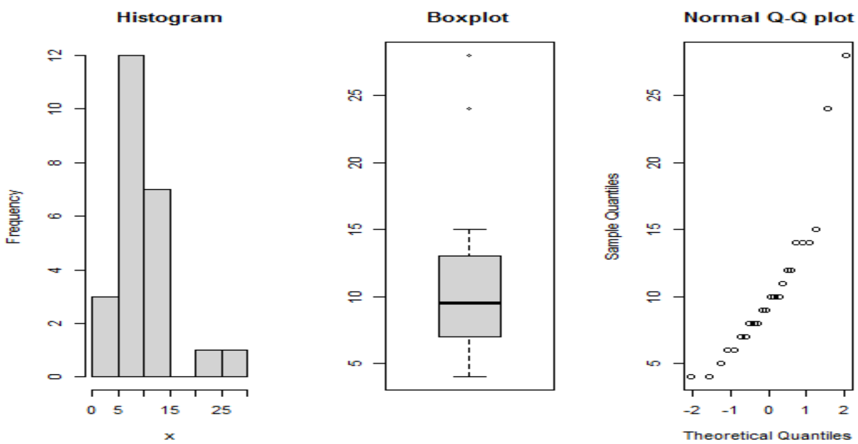
Outliers can largely influence the results of the statistical tests and hence it is necessary to find the outliers in the dataset. Most of the statistical tests and machine learning methods are sensitive to outliers and they must be removed before performing the analysis.

Statistical methods to find outliers

1. **Visual approaches** such as histogram, scatterplot (such as Q-Q plot), and boxplot are the easiest method to detect outliers.

Let's take an example of this univariate dataset [10,4,6,8,9,8,7,6,12,14,11,9,8,4,5,10,14,12,15,7,10,14,24,28] and identify outliers using visual approaches,

```
library(gridExtra)
x = c(10,4,6,8,9,8,7,6,12,14,11,9,8,4,5,10,14,12,15,7,10,14,24,28)
# histogram, Q-Q plot, and boxplot
par(mfrow = c(1, 3))
hist(x, main = "Histogram")
boxplot(x, main = "Boxplot")
qqnorm(x, main = "Normal Q-Q plot")
```



2. **Mean and Standard deviation (SD)**

Standard deviation is the positive square root of the variance. Standard deviation is one of the basic methods of statistical analysis. Standard deviation is commonly abbreviated as SD and denoted by 'σ' and it tells about the value that how much it has deviated from the mean value. If we get a low standard deviation then it means that the values tend to be close to the mean whereas a high standard deviation tells us that the values are far from the mean value.

Standard deviation is the degree of dispersion or the scatter of the data points relative to its mean, in descriptive statistics. It tells how the values are spread across the data sample and it is the measure of the variation of the data points from the mean. The standard deviation of a sample, statistical population, random variable, data set, or probability distribution is the square root of its variance.

```
x = c(10,4,6,8,9,8,7,6,12,14,11,9,8,4,5,10,14,12,15,7,10,14,24,28)

# get mean and Standard deviation
mean = mean(x)
std = sd(x)
# get threshold values for outliers
Tmin = mean-(3*std)
Tmax = mean+(3*std)
# find outlier
x[which(x < Tmin | x > Tmax)]
# remove outlier
x[which(x > Tmin & x < Tmax)]

[1] 28
[1] 10  4  6  8  9  8  7  6 12 14 11  9  8  4  5 10 14 12 15  7 10 14 24
```

The mean and Standard deviation (SD) method identified the value 28 as an outlier.

3. Median and Median Absolute Deviation (MAD)

The median of the dataset can be used in finding the outlier. Median is more robust to outliers as compared to mean. As opposed to mean, where the standard deviation is used for outlier detection, the median is used in Median Absolute Deviation (MAD) method for outlier detection.

MAD = b * median(|Y_i – median(Y_i|), where b is the scale factor and its value set as 1.4826 when data is normally distributed.

```
# example dataset
x = c(10,4,6,8,9,8,7,6,12,14,11,9,8,4,5,10,14,12,15,7,10,14,24,28)
# get median
med = median(x)
# subtract median from each value of x and get absolute deviation
abs_dev = abs(x-med)
# get MAD
mad = 1.4826 * median(abs_dev)
# get threshold values for outliers
Tmin = med-(3*mad)
Tmax = med+(3*mad)

# find outlier
x[which(x < Tmin | x > Tmax)]
[1] 24 28

# remove outlier
x[which(x > Tmin & x < Tmax)]
[1] 10  4  6  8  9  8  7  6 12 14 11  9  8  4  5 10 14 12 15  7 10 14
```

The median and median absolute deviation (MAD) method identified the values 24 and 28 as outliers.

4. Interquartile Range (IQR)

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

- Q1 represents the 25th percentile of the data.
- Q2 represents the 50th percentile of the data.
- Q3 represents the 75th percentile of the data.

If a dataset has *2n / 2n+1* data points, then
Q1 = median of the dataset.
Q2 = median of n smallest data points.
Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely Q1 and Q3: *IQR = Q3 – Q1*. The data points which fall below *Q1 – 1.5 IQR* or above *Q3 + 1.5 IQR* are outliers.

```
x = c(10,4,6,8,9,8,7,6,12,14,11,9,8,4,5,10,14,12,15,7,10,14,24,28)
# get values of Q1, Q3, and IQR
summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00    7.00    9.50   10.62   12.50   28.00

# get IQR
IQR(x)
[1] 5.5

# get threshold values for outliers
Tmin = 7-(1.5*5.5)
Tmax = 12.50+(1.5*5.5)

# find outlier
x[which(x < Tmin | x > Tmax)]
[1] 24 28

# remove outlier
x[which(x > Tmin & x < Tmax)]
[1] 10  4  6  8  9  8  7  6 12 14 11  9  8  4  5 10 14 12 15  7 10 14
```

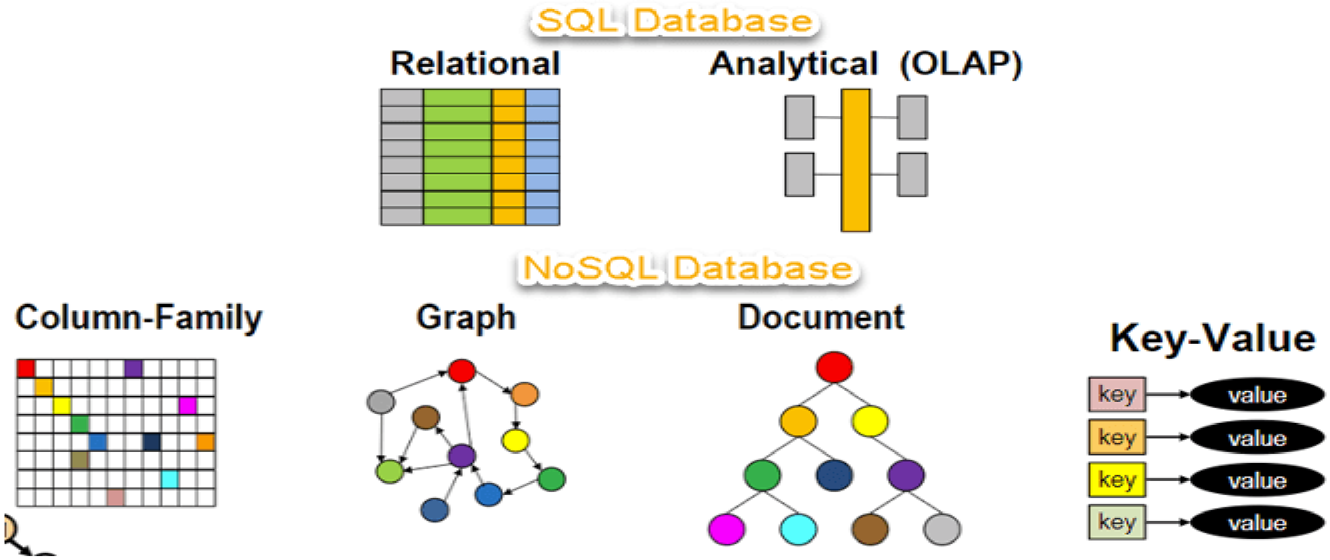
Based on IQR method, the values 24 and 28 are outliers in the dataset.

Introduction to NoSQL

NoSQL Database is a non-relational Data Management System that does not require a fixed schema. It avoids joins, and is easy to scale. The major purpose of using a NoSQL database is for distributed data stores with humongous data storage needs. NoSQL is used for Big data and real-time web apps. For example, companies like Twitter, Facebook and Google collect terabytes of user data every single day.

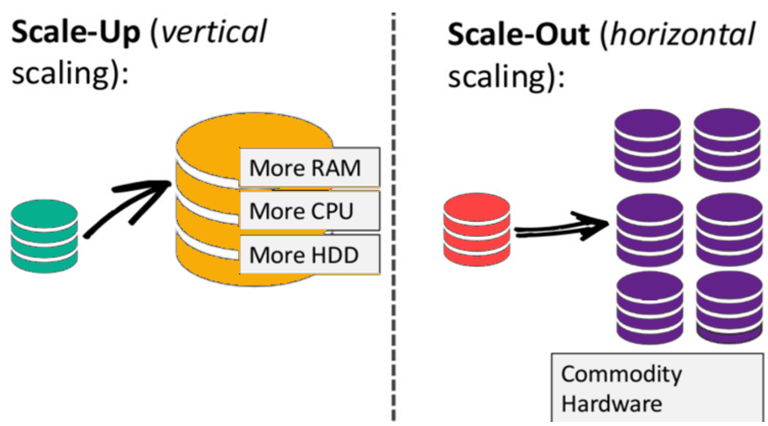
NoSQL database stands for "Not Only SQL" or "Not SQL." Though a better term would be "NoREL", NoSQL caught on. Carl Strozz introduced the NoSQL concept in 1998.

Traditional RDBMS uses SQL syntax to store and retrieve data for further insights. Instead, a NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data. Let's understand about NoSQL with a diagram in this NoSQL database tutorial:



Why NoSQL?

The concept of NoSQL databases became popular with Internet giants like Google, Facebook, Amazon, etc. who deal with huge volumes of data. The system response time becomes slow when you use RDBMS for massive volumes of data. To resolve this problem, we could "scale up" our systems by upgrading our existing hardware. This process is expensive. The alternative for this issue is to distribute database load on multiple hosts whenever the load increases. This method is known as "scaling out."



NoSQL database is non-relational, so it scales out better than relational databases as they are designed with web applications in mind.

Brief History of NoSQL Databases

- 1998- Carlo Strozzi use the term NoSQL for his lightweight, open-source relational database
- 2000- Graph database Neo4j is launched
- 2004- Google BigTable is launched
- 2005- CouchDB is launched
- 2007- The research paper on Amazon Dynamo is released
- 2008- Facebooks open sources the Cassandra project
- 2009- The term NoSQL was reintroduced

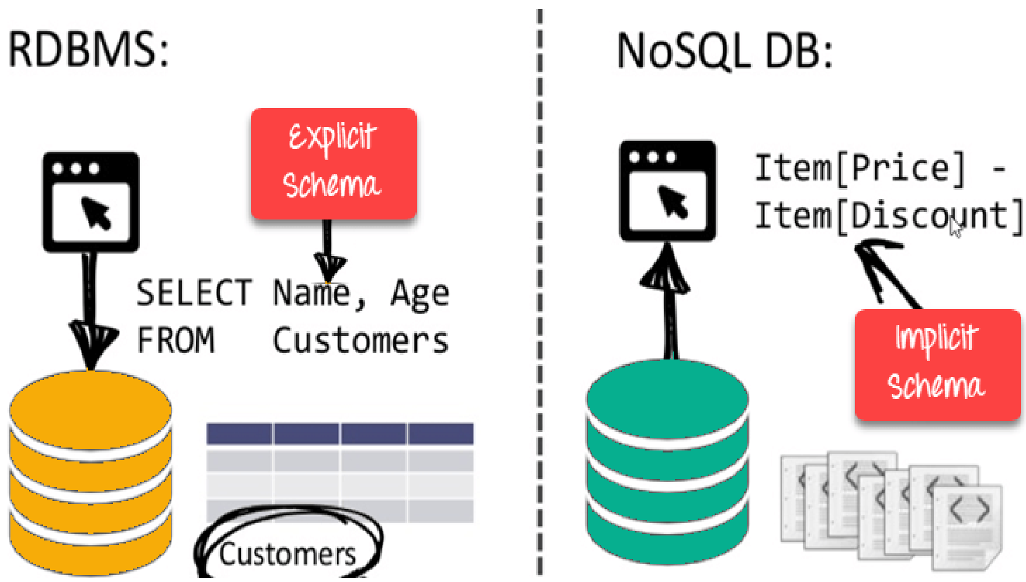
Features of NoSQL

Non-relational

- NoSQL databases never follow the relational model
- Never provide tables with flat fixed-column records
- Work with self-contained aggregates or BLOBs
- Doesn't require object-relational mapping and data normalization
- No complex features like query languages, query planners,referential integrity joins, ACID

Schema-free

- NoSQL databases are either schema-free or have relaxed schemas
- Do not require any sort of definition of the schema of the data
- Offers heterogeneous structures of data in the same domain
-



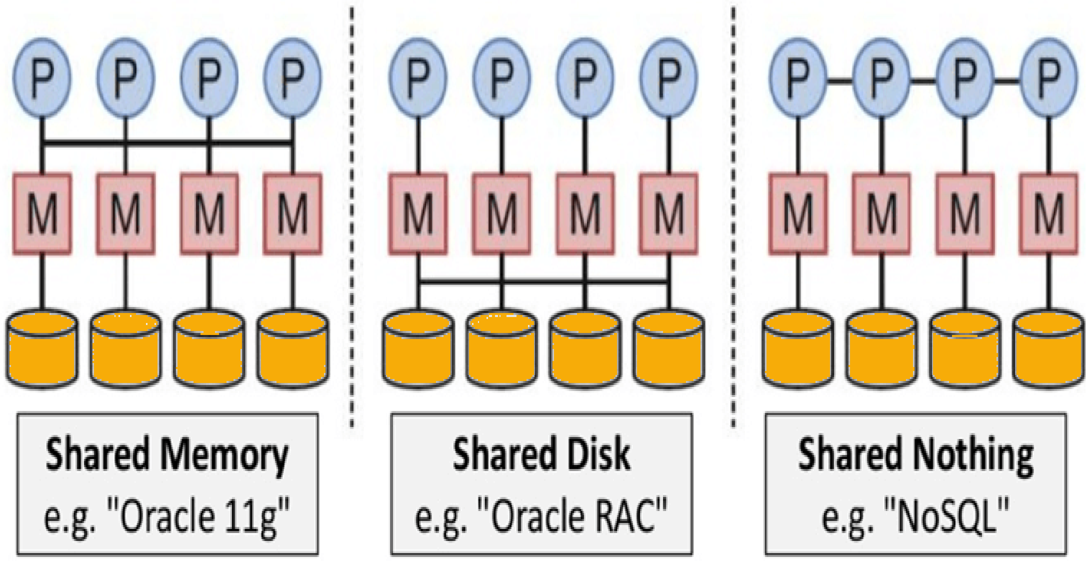
Simple API

- Offers easy to use interfaces for storage and querying data provided
- APIs allow low-level data manipulation & selection methods
- Text-based protocols mostly used with HTTP REST with JSON
- Mostly used no standard based NoSQL query language
- Web-enabled databases running as internet-facing services

Distributed

- Multiple NoSQL databases can be executed in a distributed fashion
- Offers auto-scaling and fail-over capabilities
- Often ACID concept can be sacrificed for scalability and throughput
- Mostly no synchronous replication between distributed nodes Asynchronous Multi-Master Replication, peer-to-peer, HDFS Replication
- Only providing eventual consistency

- Shared Nothing Architecture. This enables less coordination and higher distribution.

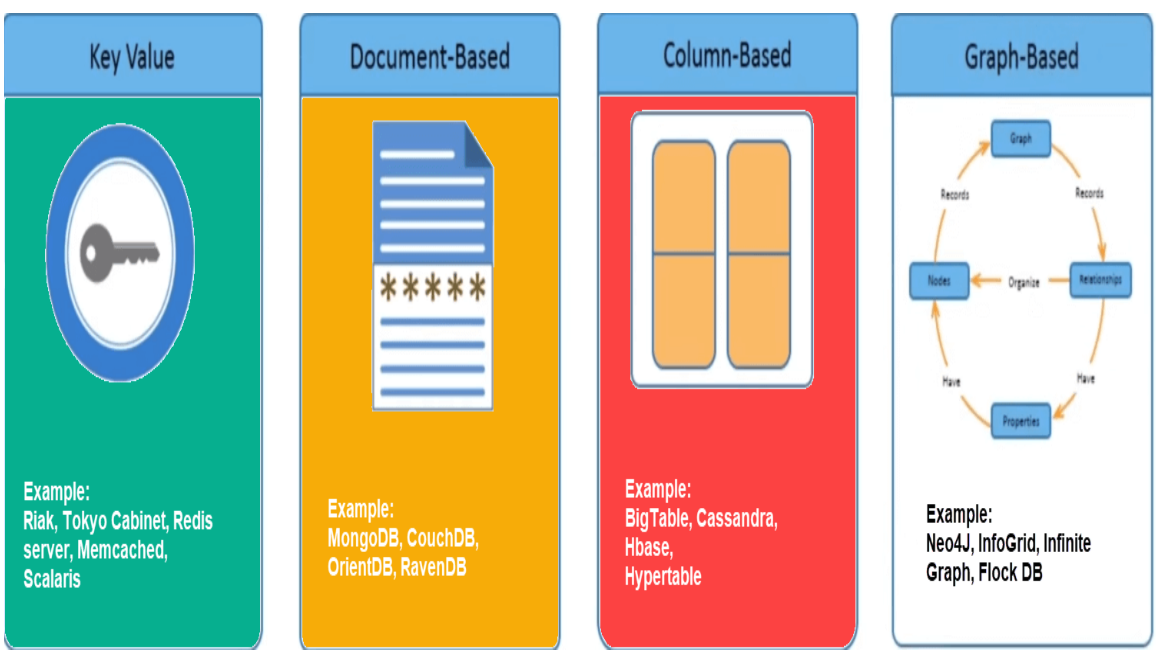


Types of NoSQL Databases

NoSQL Databases are mainly categorized into four types: Key-value pair, Column-oriented, Graph-based and Document-oriented. Every category has its unique attributes and limitations. None of the above-specified database is better to solve all the problems. Users should select the database based on their product needs.

Types of NoSQL Databases:

- Key-value Pair Based eg. Redis, Dynamo, Riak
- Column-oriented Graph (MariaDB, CrateDB, ClickHouse, Greenplum Database, Apache Hbase, Apache Kudu, Apache Parquet, Hypertable, MonetDB)
- Graphs based (Neo4j, RedisGraph, OrientDB)
- Document-oriented (MongoDB, DynamoDB and CosmosDB)



Key Value Pair Based

Data is stored in key/value pairs. It is designed in such a way to handle lots of data and heavy load.

Key-value pair storage databases store data as a hash table where each key is unique, and the value can be a JSON, BLOB(Binary Large Objects), string, etc.

For example, a key-value pair may contain a key like “Website” associated with a value like “Guru99”.

Key	Value
Name	Joe Bloggs
Age	42
Occupation	Stunt Double
Height	175cm
Weight	77kg

It is one of the most basic NoSQL database example. This kind of NoSQL database is used as a collection, dictionaries, associative arrays, etc. Key value stores help the developer to store schema-less data. They work best for shopping cart contents.

Redis, Dynamo, Riak are some NoSQL examples of key-value store DataBases. They are all based on Amazon’s Dynamo paper.

Column-based

Column-oriented databases work on columns and are based on BigTable paper by Google. Every column is treated separately. Values of single column databases are stored contiguously.

<table>
<tr><th colspan="4">ColumnFamily</th></tr>
<tr><td rowspan="6">Row Key</td><td colspan="3">Column Name</td></tr>
<tr><td>Key</td><td>Key</td><td>Key</td></tr>
<tr><td>Value</td><td>Value</td><td>Value</td></tr>
<tr><td colspan="3">Column Name</td></tr>
<tr><td>Key</td><td>Key</td><td>Key</td></tr>
<tr><td>Value</td><td>Value</td><td>Value</td></tr>
</table>

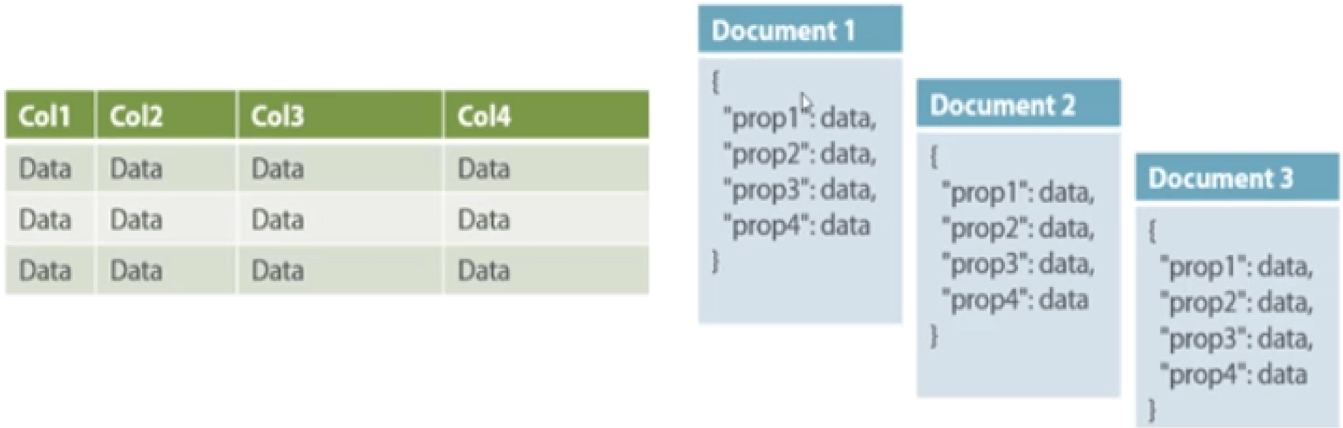
Column based NoSQL database

They deliver high performance on aggregation queries like SUM, COUNT, AVG, MIN etc. as the data is readily available in a column.

Column-based NoSQL databases are widely used to manage data warehouses, business intelligence, CRM, Library card catalogs, HBase, Cassandra, HBase, Hypertable are NoSQL query examples of column based database.

Document-Oriented:

Document-Oriented NoSQL DB stores and retrieves data as a key value pair but the value part is stored as a document. The document is stored in JSON or XML formats. The value is understood by the DB and can be queried.



Relational Vs. Document

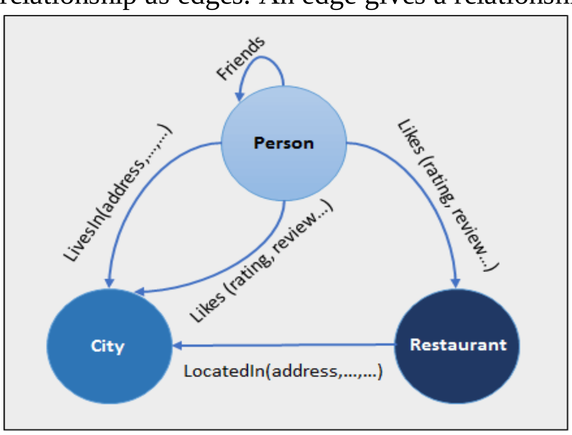
In this diagram on your left you can see we have rows and columns, and in the right, we have a document database which has a similar structure to JSON. Now for the relational database, you have to know what columns you have and so on. However, for a document database, you have data store like JSON object. You do not require to define which make it flexible.

The document type is mostly used for CMS systems, blogging platforms, real-time analytics & e-commerce applications. It should not use for complex transactions which require multiple operations or queries against varying aggregate structures.

Amazon SimpleDB, CouchDB, MongoDB, Riak, Lotus Notes, MongoDB, are popular Document originated DBMS systems.

Graph-Based

A graph type database stores entities as well the relations amongst those entities. The entity is stored as a node with the relationship as edges. An edge gives a relationship between nodes. Every node and edge has a unique identifier.



Compared to a relational database where tables are loosely connected, a Graph database is a multi-relational in nature. Traversing relationship is fast as they are already captured into the DB, and there is no need to calculate them.

Graph base database mostly used for social networks, logistics, spatial data.

Neo4J, Infinite Graph, OrientDB, FlockDB are some popular graph-based databases.

Query Mechanism tools for NoSQL

The most common data retrieval mechanism is the REST-based retrieval of a value based on its key/ID with GET resource

Document store Database offers more difficult queries as they understand the value in a key-value pair. For example, CouchDB allows defining views with MapReduce

What is the CAP Theorem?

CAP theorem is also called brewer's theorem. It states that is impossible for a distributed data store to offer more than two out of three guarantees

1. Consistency
2. Availability
3. Partition Tolerance

Consistency:

The data should remain consistent even after the execution of an operation. This means once data is written, any future read request should contain that data. For example, after updating the order status, all the clients should be able to see the same data.

Availability:

The database should always be available and responsive. It should not have any downtime.

Partition Tolerance:

Partition Tolerance means that the system should continue to function even if the communication among the servers is not stable. For example, the servers can be partitioned into multiple groups which may not communicate with each other. Here, if part of the database is unavailable, other parts are always unaffected.

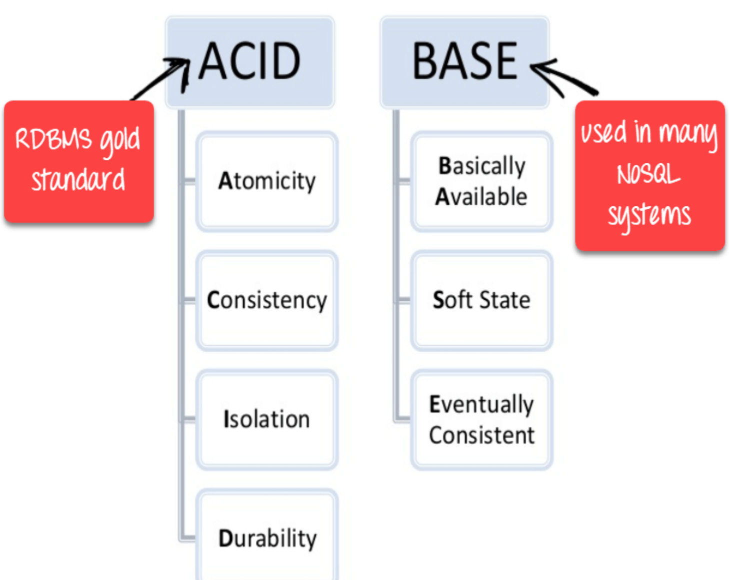
Eventual Consistency

The term "eventual consistency" means to have copies of data on multiple machines to get high availability and scalability. Thus, changes made to any data item on one machine has to be propagated to other replicas.

Data replication may not be instantaneous as some copies will be updated immediately while others in due course of time. These copies may be mutually, but in due course of time, they become consistent. Hence, the name eventual consistency.

BASE: **B**asically **A**vailable, **S**oft state, **E**ventual consistency

- Basically, available means DB is available all the time as per CAP theorem
- Soft state means even without an input; the system state may change
- Eventual consistency means that the system will become consistent over time



Advantages of NoSQL

- Can be used as Primary or Analytic Data Source
- Big Data Capability
- No Single Point of Failure
- Easy Replication
- No Need for Separate Caching Layer
- It provides fast performance and horizontal scalability.
- Can handle structured, semi-structured, and unstructured data with equal effect
- Object-oriented programming which is easy to use and flexible
- NoSQL databases don’t need a dedicated high-performance server
- Support Key Developer Languages and Platforms
- Simple to implement than using RDBMS
- It can serve as the primary data source for online applications.
- Handles big data which manages data velocity, variety, volume, and complexity
- Excels at distributed database and multi-data center operations
- Eliminates the need for a specific caching layer to store data
- Offers a flexible schema design which can easily be altered without downtime or service disruption

Disadvantages of NoSQL

- No standardization rules
- Limited query capabilities
- RDBMS databases and tools are comparatively mature
- It does not offer any traditional database capabilities, like consistency when multiple transactions are performed simultaneously.
- When the volume of data increases it is difficult to maintain unique values as keys become difficult
- Doesn’t work as well with relational data
- The learning curve is stiff for new developers
- Open source options so not so popular for enterprises.

Connecting R to NoSQL databases

Connect to MongoDB Database

MongoDB is a NoSQL database program using JSON type of documents with schemas. It's open source cross-platform database. MongoDB is the representative NoSQL database engine.

Inserts **web Crawling datasets including text data, image url, and so on** to NoSQL database.

Step 1. Install the Drivers

In between the mongolite and Rmongo, Lets assume you want RMongo. Install it.

```
install.packages("RMongo")
```

In R, type below code and install mongolite

```
library(mongolite)
```

Step 2. Create Connection

To establish the connection we have to first use the package. And then make the connection.

```
library(RMongo)
rmng <- mongoDbConnect('db')
```

(or)

```
url_path = 'mongodb+srv://<username here>:<password here>!@<cluster url>/admin'
#make connection object that specifies new database and collection (dataset)
mongo <- mongo(collection = "listingsAndReviews", db = "sample_airbnb", url = url_path, verbose = TRUE)
```

Sample url_path is provided. So, users must find your own url_path.

Next code is for connecting to database called sample_airbnb, which already uploaded freely as following instructions.

```
mongo_db <- mongo(collection = "listingsAndReviews", # Data Table
            db = "sample_airbnb", # DataBase
            url = url_path, verbose = TRUE)

print(mongo_db)
```

<u>Creation Database & Collection, insertion</u>

If users want to create database and collection, it's easy to create them with mongo() & collection=, db=. The sample code is below.

```
my_collection = mongo(collection = "name_of_collection", db = "name_of_database")
 # create connection, database and collection
my_collection$insert(name_of_collection)
```

Following the code, let's create insert iris data to mongoDB. Here, delete the current connection is important. rm(mongo_db) in this case.

```
rm(mongo_db) # disconnection
data("iris")
iris_collection <- mongo(collection = "iris", # Creating collection
            db = "sample_dataset_R", # Creating DataBase
            url = url_path, verbose = TRUE)
```

insert code
iris_collection$insert(iris)
##

Complete! Processed total of 150 rows.
List of 5
$ nInserted : num 150
$ nMatched : num 0
$ nRemoved : num 0
$ nUpserted : num 0
$ writeErrors: list()

Now, users go to mongodb cloud and please check database & collection. It's very successfully inserted to mongoDB cloud.

After inserting new data to mongoDB, then it's okay to disconnect iris_collection.
rm(iris_collection)

Connect to SQLite database
Embeds the SQLite database engine in R, providing a DBI-compliant interface.
SQLite is a public-domain, single-user, very light-weight database engine that implements a decent subset of the SQL 92 standard, including the core table creation, updating, insertion, and selection operations, plus transaction management.

You can install the latest released version of RSQLite from CRAN with:
install.packages("RSQLite")
Or install the latest development version from GitHub with:
 # install.packages("devtools")
devtools::install_github("rstats-db/RSQLite")

To establish the connection we have to first use the package. And then make the connection.
library(RSQLite)

setwd("/sqlitedatafolder")
sqlite <- dbDriver("SQLite")
conn <- dbConnect(sqlite,"rlang_db.sqlite3")

Basic Usage Code:
library(DBI)
 # Create an ephemeral in-memory RSQLite database
con <- dbConnect(RSQLite::SQLite(), ":memory:")

dbListTables(con) #will list tables

dbWriteTable(con, "mtcars", mtcars)
dbListTables(con)

dbListFields(con, "mtcars") #lists field names of mtcars dataset

dbReadTable(con, "mtcars") #displays the data stored in the dataset

```
# You can fetch all results:
res <- dbSendQuery(con, "SELECT * FROM mtcars WHERE cyl = 4")
dbFetch(res)

dbClearResult(res)

# Or a chunk at a time
res <- dbSendQuery(con, "SELECT * FROM mtcars WHERE cyl = 4")
while(!dbHasCompleted(res)){
  chunk <- dbFetch(res, n = 5)
  print(nrow(chunk))
}

# Clear the result
dbClearResult(res)

# Disconnect from the database
dbDisconnect(con)
```

Connect to CouchDB Database

To establish the connection we have to first use the package. And then make the connection.(outdated)

```
install.packages("R4CouchDB")
conn <- cdbIni()
conn$serverName
```

Database Connectivity with R Programming

To connect Database with R Programming we will be going to connect R script with MySQL Database.

To begin with the connection process, follow the steps given below:

Step 1: Create a database in MySQL with the following command

Step 2: To connect the database with R we can use R Studio.

Step 3: Use the following command to install the MySQL library in RStudio:

```
install.packages("RMySQL")
```

Now execute the following commands as RScript:

```
#To check whether the library is installed or not
library(RMySQL)

# Create a connection Object to MySQL database.
mysqlconnection = dbConnect(MySQL(), user = 'root', password = 'root', dbname = 'onlinetutorials', host = 'localhost')
typeof(mys)

# List the tables available in this database.
dbListTables(mysqlconnection)

# Query the "actor" tables to get all the rows.
a = dbSendQuery(mysqlconnection, "create table students(id int, name varchar(10))")
```

```
a = dbSendQuery(mysqlconnection, "insert into students values(101, 'amit')")
a = dbSendQuery(mysqlconnection, "insert into students values(102, 'aman')")
result = dbSendQuery(mysqlconnection, select * from students")

# Store the result in a R data frame object.
# n = 5 is used to fetch first 5 rows.
data.frame = fetch(result)
print(data.frame)
```

Unit III

Understand the Verticals - Engineering, Financial and others (NOS 2101):

Understanding systems viz. Engineering Design, Manufacturing, Smart Utilities, Production lines, Automotive, Technology etc. Understanding Business problems related to various businesses

What Are Business Verticals? (Definition)

Each business needs to discover a market that they believe they can serve in the most optimum manner possible. That is, finding your niche is of coveted importance.

When any business prizes itself at providing exclusively for a specific industry or demographic, they are called business verticals or vertical markets.

These "verticals" are businesses that aim at targeting specific audience needs and providing specialized services for the same.

- Business verticals don't traverse within various industries.
- They set out with a mission to amplify their position in a single, particular market.

Some broad examples of business verticals include insurance, banking, hospitals, retail, real estate, government, and more. Verticals can also be subcategorized into narrower niche markets.

Difference between Horizontal & Vertical Markets

A business vertical is all about scaling only within an already existing market whereas a horizontal market will allow a business to diversify itself and be omnipresent in multiple industries at a given time.

That means a business that sells in a varied number of industries is a part of horizontal markets.

Horizontal markets will try to maximize their outreach to each and every type of audience whereas vertical markets aim at a targeted consumer base.

Also, horizontal markets are the more basic of the two markets as they host a varying range of customers without further concerns about the industry they fall into while the vertical market is quite the opposite.

For example, A company like Whole Foods that only caters to its consumers' needs for organic supplies will be considered as a business vertical.

Whereas, a big box company like Walmart is an example of a horizontal market because it accommodates itself in a diverse range of business activities.

Benefits Of Business Verticals

Business verticals, when implemented in the right business and in the right manner can create long-standing and resounding results. Some benefits of business verticals are:

- It takes a village to understand the intricacies of business and by narrowing down a certain niche, business verticals become more able at assessing and serving the needs of their consumers.
- They enable better outreach from campaigns as they are directed at specifically targeted audiences.
- Developing strategies is easier as they have a narrower customer base and thus, they can leverage their position.
- With increased knowledge of an industry, a firm can be better prepared for external stimuli like changing economic issues, new market trends, variations in in-laws, etc.
- As verticals target specific industries, there is very little competition in newly-developing niches which helps to generate sales and attain higher market positions.
- From a consumer's perspective, a business vertical offers a simplified solution to the particular needs of the customer and strives to offer them the best resolution possible.

Disadvantages Of Business Verticals

Let's explore some of the drawbacks of business verticals:

- Distinct markets, although one of the biggest pros of vertical markets, also turns out to be a major con. Small markets bring in a smaller clientele, thus limiting exposure and loss of potential earnings.

- There's also plausibility to the idea that when a firm increases its pricing policies owing to its leading market position, it tends to shun out a major chunk of its prospective customers from its already compact audience.
- Along with all this comes the added risk factor, businesses tend to fall and downsize with changing trends, and other factors, this condenses an already concise and niche market.
- Vertical markets need to be assessed accurately enough to be able to understand if a business can survive in the specified niche or not, and a market as intricate as vertical won't make it any easier on you.
- Limited customer base. Focusing on a niche market may mean an organization will have a more difficult time finding or expanding its customer base.
- Limited revenue. If a vertical market doesn't have a large customer base, it may be difficult for an organization to generate the revenue and profit it needs to operate.
- Niche market shifts. If the needs of a niche customer base in a vertical market change suddenly, a company catering to that vertical market could lose a significant part of its customer base in a short period of time. The closing of restaurants and bars during the COVID-19 pandemic is an example of a niche market shift.

Niche marketing is a targeted marketing strategy aimed at small, specific and well-defined portions of the population...

To understand business verticals better, let's glance through some examples…

Some Examples of Business Verticals:

1. Many pet-related brands and stores like 'PetSmart' cater to niche markets of your pet's grooming and health needs. They exclusively provide pet-related products in their stores and are a great example of business verticals.
2. Up & coming makeup and beauty brands that have specifically started targeting skin types of people of color also fall under the category of vertical markets.
3. Consider a bank that specializes in home loans, it would also be a vertical market as only the potential homeowners who are looking for homes would revert to such a bank. That means the bank offers its services to a particular group of people in a specific industry; real estate.
4. Online institutions that offer educational courses also fall under this category of business verticals as they cover every course imaginable to a certain audience and fully make do of their niche; the educational industry.
5. A vegan restaurant serving only vegan food provides for only people who adopt a certain style of lifestyle namely veganism and thus places itself in a distinct market type.
6. Software customized and created specifically for a certain company's needs and its management also falls under the umbrella of vertical markets.

Businesses are not run blindly, you have to set your own set of goals that you want to achieve effectively to thrive in the market. However, setting goals and then working towards them without action plans bring nothing but ineffectiveness.

To ensure that the goals you have set for your business become achievable, you need to devise business strategies. A business strategy acts as a guide for the people working in your organization to make the right decisions at the right time.

Whether you are a businessman or an entrepreneur, you will need to develop business strategies so that your business can sustain itself in today's highly competitive markets.

When you set goals & objectives for your business, you need a roadmap that assists your organization in achieving them successfully. A business-level strategy behaves as that roadmap and also helps the business to gain a competitive edge and deliver better value to its customers.

In simpler words, *a business-level strategy is what gives direction to the efforts of a business organization.*

The most important thing you need to consider while devising a business-level strategy is that it should be in line with both the vision and the mission of your organization.

For instance, if you want to gain more market share, you need to focus more on making your products/services available at prices lower than the competitors. On the other hand, if your organization's aim is to deliver unique & top-quality products/services, you need to focus on innovation.

Create a Business Strategy

- Keep the Vision of Your Organization in View

Vision is what allows you to look into the future and see where your business is going to be in the upcoming years.

To develop an effective business strategy, you need to be well-aware of the vision of your business. After all, it's the vision of the business, which makes it much easier to define the long-term goals associated with organizational success.

- Set Crystal Clear Objectives and Milestones

It's difficult to achieve long-term goals without breaking them down into achievable objectives.

Once you have set a long-term goal for the business, the next thing you need to do is to figure out the answer to the question — how to know if you are making the right amount of progress?

This is where it is important while devising a business strategy, that you divide lengthy goals into small and actionable objectives. Set objectives that make it much easier to track the progress and achieve bigger goals in an effective and timely manner.

Moreover, you can also define various milestones and short-terms goals that can also help you to know from time to time that you are making progress in the right direction.

- Examine all the Factors that Have an Impact On your Business

It's a fact that each business has a set of factors that affects its performance.

Being a strategy maker, you need to take into all the environmental factors (internal and external) that have an impact on your business's operations. If you ignore such factors, it is quite possible that they will interfere with your business's operation and your strategy will become ineffective.

Common internal factors include plans & policies, financial resources, human resources, machinery & equipment, etc. Similarly, some of the most common external factors are, namely target customers, competitors, suppliers, economic & political policies of the region in which the business is operating, technology, etc.

- Take Inputs From Experienced Personnel

While the responsibility of developing a business strategy is limited to a few people, it's great if you take suggestions from others also; especially the senior members of your team.

This includes employees who have been working for your organization for a long time, or the experienced people working at the management level. It will be great to take inputs from different people as it will help you to address the issues at the ground level.

- Performance Assessment

A key point to keep in mind when you are developing a business strategy is that it should be actionable and measurable.

Being actionable means that you along with other people in your organization should be able to work practically on the objectives and milestones.

On the other hand, a business strategy is measurable when it is possible to track the outcome of the efforts made to achieve the objectives, and milestones defined within the strategy.

8 Best Strategies Practiced by Leading Businesses

1. Development of Innovative Products and Services

This strategy corresponds to the company's aim of focusing more on R&D work to develop advanced products and services.

In general, companies working in the field of technology, such as IT firms, automotive companies, space agencies, etc. follow this business strategy aggressively.

However, the strategy is also followed by many non-tech industries to develop new products that are not available in the market. Such companies make use of the latest technology in manufacturing and delivering innovative products & services to customers.

You need to invest a significant amount of money in the research & development of new products and services.

Being innovative in your business space makes it easier for your organization to get recognized easily by the target customers and thus generate a good amount of sales.

2. Expansion of the Product Line

Another strategy that is quite common among businesses is the expansion of their product line. This strategy is effective when you need to attract new customers by offering a wide variety of products.

Additionally, this strategy can also work for the existing customers as they are likely the ones who will be least reluctant in giving a try to your new products.

The strategy of adding new products will help to boost the sales of your business.

However, one particular risk associated with this strategy is that you need to give special attention to the product that you are willing to add in your product line otherwise, there can be higher chances of the failure of the product.

3. Tap Into a New Market

Within this strategy, your organization needs to enter a market that is new and has a significant amount of potential.

You can tap into new markets by selling your products or services in a new country or a geographical region. Moreover, targeting new customer segments also comes under this strategy.

The main motive of this strategy is to expand the customer base and get more sales by making products and services available to new customers.

However, giving life to a new market or entering a young market is challenging as it involves thorough research and higher risks.

4. Boost Customer Retention

Customer retention is important for any business to ensure that they have a fixed amount of market share.

The customer retention strategy focuses on improving the customer experience and making your active customers more loyal towards your products or services.

There are many ways of retaining customers such as delivering better customer service, giving special benefits to old customers, offering discounts to existing customers, etc. Customer retention is a common strategy to secure a certain portion of the market share that is enough for business to survive at critical times.

5. Marketing and Brand Awareness Campaigns

Marketing is an important and active part of any business. Running marketing campaigns from time to time helps a business to increase its sales and become visible to potential customers.

Additionally, brand awareness campaigns let people know about the business and what it has to offer to the customers.

In today's modern world, digital marketing has been teamed with conventional marketing to get a better reach. In addition, social media marketing has become the preferred choice for increasing brand awareness.

Marketing and brand awareness campaigns are ideal for attracting new customers and driving more sales.

6. Deployment of Latest and Relevant Technologies

This is a particular strategy that has become popular in recent years because of the emergence of technology.

Under this strategy, a business upgrades itself in terms of the technology that helps the business operations to execute smoothly and efficiently. The technological upgrades can be both in terms of hardware and software.

By deploying the latest technologies in your business, it is possible to increase the productivity of your employees and get more work done in less time.

For instance, a leading project management software can help you to manage all your projects, tasks, and teams from a single place. In addition, there are several features in a project management tool that makes it easy for teams to collaborate and stay on top of all their work.

7. Redefine Pricing Strategy

Redefining the pricing of the products is also a business strategy that organizations implement. Usually, there are two choices that you have while following this strategy.

You can either keep the prices of your products or services low enough to attract more customers.

However, you will need to increase the production capacity of your business to meet the huge demands, and also the profit margin is low.

The other choice is to keep the prices high enough so that your product/service is not within the reach of ordinary customers. The sales will be limited, but the profit margin will be quite high.

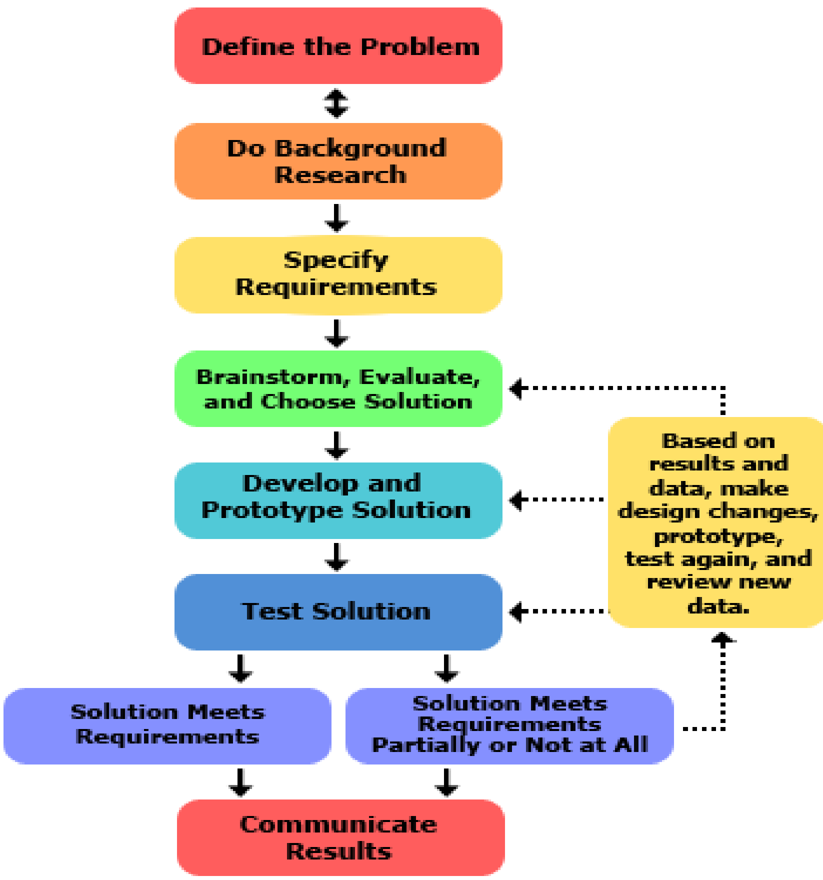
This type of pricing is ideal if you want to offer high-quality products to your customers and emerge as a premium brand.

8. Manage Resources Effectively

Resources are important for any business whether it's the raw materials used to create products or the workforce that manage the services offered by the firm.
By making the optimal use of your business's resources, it is possible to bring down the costs and increase profits.
The strategy for effective resource management involves taking necessary steps to avoid wastage of resources.
By managing the resources effectively, it is possible to increase the output at virtually the same or lower cost than before.
While this strategy sounds easy, implementing it can be difficult, especially in large businesses.

What is the Engineering Design Process?

The engineering design process is a series of steps that engineers follow to come up with a solution to a problem. Many times the solution involves designing a product (like a machine or computer code) that meets certain criteria and/or accomplishes a certain task. If your project involves making observations and doing experiments, you should probably follow the Scientific Method. If your project involves designing, building, and testing something, you should probably follow the Engineering Design Process. This diagram shows the steps of the engineering design process, and the table below describes each step in more detail:



Engineers do not always follow the engineering design process steps in order, one after another. It is very common to design something, test it, find a problem, and then go back to an earlier step to make a modification or change to your design. This way of working is called iteration.

1. Define the Problem

The engineering design process starts when you ask the following questions about problems that you observe:

What is the problem or need?
Who has the problem or need?
Why is it important to solve?

Finding an idea for your engineering project requires you to identify the needs of yourself, another person, or a group of people. The act of looking at the world around you to identify these needs is called need finding.

To help you find an idea for your engineering project:
Create a list of all the things that annoy or bother the people around you. Record this bug list in your Design Notebook.
Mind Map possible design problems, ideas, or areas of interest to you.

Once you have found an idea for your engineering project, describe the problem by writing a problem statement.

The format for writing a problem statement uses your answers to the questions and follows these guidelines:
Who need(s) what because why.
_____ need(s) _________ because ________.

Before moving forward with an idea for your engineering project, be sure to evaluate your problem.

One really great way to start the need-finding process is to make a "bug list." Think about all of the things that bug you or bug other people around you. Write them down. They may seem like small and silly problems, but they can spark ideas for a project or lead to larger problems that you may not have noticed otherwise.

Here are some examples of things you might find on someone's bug list:

- Uncomfortable airplane seats
- When one light on a string of Christmas lights goes out
- How quickly chewing gum loses flavor
- Moving (packing boxes, cleaning, unpacking, etc.)
- Public restrooms without toilet paper
- Long lines at amusement parks
- When food gets stuck in vending machines
- Dog or cat hair that gets stuck on clothing
- Sharing armrests with strangers at the movies
- Wasting water in the shower
- Losing one earring
- Draining tuna fish cans

There are two different types of potential project ideas that you have come up with on your bug list. First, there are the unsolved problems that don't currently have a solution. Second, there are poorly solved problems that have solutions, but the solutions are not entirely successful.

A design notebook is a way for a designer or engineer to keep a history of his or her design project from start to finish. It is a place to record research, observations, ideas, drawings, comments, and questions during the design process. At the end of your project, someone reviewing your design notebook should be able to understand fully how you got to your solution.

Everything goes in a design notebook. Your design notebook starts when you begin thinking about possible problems to solve. Write down everything you know about these problems and why you want to solve them. Then write down, draw, sketch, glue, or tape in every step of your process between this first step and your final solution.

Here are examples of what you might find in a design notebook:

- Notes on background research

- Interviews with users or experts
- Drawings and sketches
- Photos of competing products
- Lists of design requirements
- Questions/issues you face

Mind mapping refers to a technique that designers and engineers use to express and generate ideas. All that mind mapping really is, however, is a way to get all of the ideas in your head down onto paper. There is no right or wrong way to mind map. It is simply a visual representation of the thoughts in your head, and it often looks like organized chaos.

Mind mapping helps you to release all of the ideas in your head and gives you the opportunity to see those ideas visually. It is a fast and simple way to get your creative juices flowing, and the only tools you need are a pen or pencil and your design notebook.

To start a mind map, write down one, central idea or theme in the middle of a blank page. All mind maps have this common starting point. Then, stem off of the central idea by writing down anything that comes to your mind when thinking about the idea. You can include drawings, questions, comments, solutions, problems, etc. There are no limits. Simply write down everything that relates to the central theme or anything that enters your mind.

You can create a mind map at any stage in your design process and for absolutely any purpose. You can mind map at the very beginning before you have even decided what problem you are going to solve. You can also mind map to generate possible solutions to your problem or to identify different types of users for your project. Mind map whenever you feel the need to empty the thoughts in your head or whenever you feel stuck during the design process.

2. Do Background Research
Learn from the experiences of others — this can help you find out about existing solutions to similar problems, and avoid mistakes that were made in the past.

Background research is especially important for engineering design projects, because you can learn from the experience of others rather than blunder around and repeat their mistakes. To make a **background research plan**— a roadmap of the research questions you need to answer -- follow these steps:

1. Identify questions to ask about your **target user** or customer.
2. Identify questions to ask about the products that already exist to solve the problem you defined or a problem that is very similar.
3. Plan to research how your product will work and how to make it.
4. Network with other people with more experience than yourself: your mentors, parents, and teachers. Ask them: "What should I study to better understand my engineering project?" and "What area of science covers my project?" Better yet, ask even more specific questions.
5. Use this Background Research Plan Worksheet to help you develop your own plan.

The Focus of Your Background Research
For an engineering design project, you should do background research in two major areas:

- Users or customers
- Existing solutions

Users or Customers

- **Research your target user or customer.** Everything humans design is ultimately for the use of another human. (Think about it— even products designed for animals or plants are first purchased by another human!) Your choice of target user will sometimes have a big impact on your design requirements. For example, if you design something for a toddler, you need to make sure that there are no small parts that could be swallowed. Some customers are more sensitive to the cost than others, and so forth. You might describe your target user in any number of ways. Here are some examples:

- Age (old, young, infant)
- Gender
- Occupation
- Hobby interests
- Amateur or professional
- Whether users have disabilities and require accommodations
- Size
- First-time user or experienced user

Existing Solutions

- **Research the products that already exist to solve the problem you defined or a problem that is very similar.** No one wants to go to all the trouble of designing something they think is new, only to find that several people have already done it. That would be depressing! So, you want to investigate what's already out there. Only then can you be sure that you're making something that more effectively fills a need. And keep in mind that what is "better" depends on your requirements. You might want to build something that's been around for hundreds of years, but do it with recycled materials from around the house. The device might be old, but the construction materials new (or used!).
- **Research how your product will work and how to make it.** When it comes time to build their solution, savvy designers also want to use their research to help them find the best materials and way to do things, rather than starting from scratch. Background research is also important to help you understand the science or theory behind your solution. If you are entering a science fair, judges like to see that you understand why your product works the way it does and what causes it to perform better than other products.

How to Conduct the Research

Engineers are lucky, because there are three ways to do research regarding users and existing solutions:

- Observe users first-hand, either as they use a similar product or solution or in the environment in which they encounter the problem.
- Examine and analyze similar products and solutions. Looking at similar products is super important. Other engineers spent a lot of time designing them, so you might as well learn everything you can from their work. And it is fun! You might even want to take similar products apart! (Ask first!)
- Conduct library and Internet research.

Making a Background Research Plan: How to Know What Information to Look For

When you or your parents are driving a car, there are two ways to find your destination: drive around randomly until you finally stumble upon what you're looking for OR use a GPS or look at a map before you start. Finding information for your background research is similar. Since libraries and the Internet both contain millions of pages of information and facts, you might never find what you're looking for unless you start with a map! To avoid getting lost, you need a background research plan.

Target Users

To help clarify the definition of your target user, you'll want to ask questions like this:

- Who needs _________?
- Who wants _________?
- Who buys _________?
- What does my target user [a child, an elderly person, etc.] need or want in a _________?
- How much would my target user be willing to pay for a _________?

- What size should I make _________ for my target user?

Similar Products

Then, ask questions to help you understand products or programs that fill similar needs to the need you identified:

- What products fill a similar need?
- What are the strengths and weaknesses of products that fill a similar need?
- What are the key, must-have features of products that fill a similar need?
- Why did the engineers that built these products design them the way they did?
- How can I measure my design's improvement over existing designs?

How It Works and How to Make It

These are some example questions that will help you understand the science behind your design.

- Who invented _________?
- How does a __________ work?
- What are the different parts of a __________?
- What are the important characteristics of a __________?
- How is performance measured for a _________?
- Where does _________ get used?
- What is __________ made of?
- Why is __________ made from or using __________?
- What is the best material, component, or **algorithm** for building ________? (You may even ask this separately for the different parts of your device or program.)

Talk to People with More Experience: Networking

One of the most important things you can do while working on your project is talk to other people with more experience than yourself: your parents, teachers, and advisors. This process is called networking. Some advisors or mentors may have had classes or work experience related to the science involved in your project. Others may have used or even designed products like the one you are researching. Ask them, "What science concepts should I study to better understand my project?" Better yet, be as specific as you can when asking your questions.

And by the way, networking is something many adults don't expect students to be good at, so you can probably surprise them by doing a good job at it! The best networkers, of course, enjoy the spoils of victory. In other words, they get what they want more quickly, efficiently, and smoothly.

The reality is we have all networked at some point in our lives. Remember how you "networked" with your mom to buy you that cool water gun or "networked" with your grandpa to buy you that video game you always wanted? Well, now you are "networking" for knowledge. Train yourself to become a good networker, and you might just end up with a better project (and don't forget that you'll get a little smarter too in the process). So take our advice: work hard, but network harder.

Background Research Plan Worksheet

Prepare a Background Research Plan Worksheet for Engineering Design Projects to help you develop your own plan.

Background Research Plan Checklist

Answer the questions in the quick checklist below to evaluate your plan for background research.

What Makes a Good Background Research Plan?	For a Good Background Research Plan, You Should Answer "Yes" to Every Question
Have you identified questions to ask about your target user or customer?	Yes / No
Have you identified questions to ask about the products that already exist	Yes / No

What Makes a Good Background Research Plan?	For a Good Background Research Plan, You Should Answer "Yes" to Every Question
to solve the problem you defined or a problem that is very similar?	
Have you planned to research how your product will work and how to make it?	Yes / No

How to find information:

- Find and read the general information contained in an encyclopedia, dictionary, or textbook for each of your keywords.
- Use the bibliographies and sources in everything you read to find additional sources of information.
- Search periodical indexes at your local library.
- Search the Internet to get information from an organization, society or online database.
- Broaden your search by adding words to your search phrases in search engines. Narrow your search by subtracting words from or simplifying your search phrases.

A Checklist for Evaluating References

What Makes a Good Reference?	For a Good Reference, You Should Answer "Yes" to Every Question
Does your reference come from a credible source?	Yes / No
Is your reference current?	Yes / No
Is your reference objectively written, not biased towards one point of view?	Yes / No
Is your reference free of errors?	Yes / No
Does your reference properly cite its original sources?	Yes / No
Is the reference easy for other people to find or obtain?	Yes / No

Writing a Bibliography

A bibliography is a listing of the books, magazines, and Internet sources that you use in designing, carrying out, and understanding your science fair project. But, you develop a bibliography only after first preparing a background research plan — a road map of the research questions you need to answer. Before you compose your bibliography, you will need to develop your background research plan.

It consists of:

- the title page of a book, encyclopedia or dictionary
- the heading of an article
- the front, second, or editorial page of the newspaper
- the contents page of a journal or magazine
- the header (at the top) or footer (at the bottom) of a Web site

- the About or the Contact page of a Web site

Bibliography Checklist

What Makes a Good Bibliography?	For a Good Bibliography, You Should Answer "Yes" to Every Question
Have you included at least 3 sources of written information on your subject? (If you include Web pages, they should be in addition to the written sources.)	Yes / No
Have you included complete information to identify each of your sources (author's name, the title, the date, and where it was published)?	Yes / No
Have you used the proper format for each of your sources? Most teachers prefer the MLA or APA formats.	Yes / No
Is your Bibliography in alphabetical order, by author's last name?	Yes / No
Do you have sources of information to answer all of your research questions?	Yes / No

Writing a Research Paper for Your Project

As you do your research, follow your background research plan and take notes from your sources of information. These notes will help you write a better summary.

The purpose of your research paper is to give you the information to understand why your experiment turns out the way it does. The research paper should include:

- The history of similar experiments or inventions
- Definitions of all important words and concepts that describe your experiment
- Answers to all your background research plan questions
- Mathematical formulas, if any, that you will need to describe the results of your experiment

For every fact or picture in your research paper you should follow it with a citation telling the reader where you found the information. A citation is just the name of the author and the date of the publication placed in parentheses like this: (Author, date). This is called a reference citation when using APA format and parenthetical reference when using the MLA format. Its purpose is to document a source briefly, clearly, and accurately.

If you copy text from one of your sources, then place it in quotation marks in addition to following it with a citation. Be sure you understand and avoid plagiarism! Do not copy another person's work and call it your own. Always give credit where credit is due!

Most teachers want a research paper to have these sections, in order:

- Title page (with the title of your project, your name, and the date)
- Your report
- Bibliography
- Check with your teacher for additional requirements such as page numbers and a table of contents

3. Specify Requirements

Design requirements state the important characteristics that your solution must meet to succeed. One of the best ways to identify the design requirements for your solution is to analyze the concrete example of a similar, existing product, noting each of its key features.

Design requirements state the important characteristics that your design must meet in order to be successful.

- One of the best ways to identify the design requirements for your solution is to use the concrete example of a similar, existing product, noting each of its key features. Here is how to analyze:
    - A physical product
    - A software product or website
    - An environment
    - An experience
- To complete the requirements step of the design process, you should write a **design brief**; a document that holds all of the key information for solving your problem in one place.

Design Requirements

Design requirements state the important characteristics that your solution must meet to be successful.

For example, imagine that your problem statement relates to grocery store bags. You want to design a better grocery store bag--one that uses less expensive material than the paper and plastic bags that already exist. Your design requirements are the important characteristics that your bag must meet to be successful. Based on your problem statement, a successful bag would use less expensive material than existing bags and function properly as a grocery bag. Examples of some of your design requirements might be that the bag needs to:

- Have handles so that shoppers can carry multiple bags of groceries.
- Hold up to five pounds of food without breaking.
- Cost less than five cents to make.
- Collapse so that it can be stored in large quantities at grocery stores.

Effective design requirements are:

- *Needed* to solve your design problem. If it is not needed, leave it out. You'll have enough other things to work on!
- Feasible. A good design requirement is not just a wish. Ask if you have the time, money, materials, tools, and knowledge to make it happen.
- Subject to change as you do more research and design. Always ask yourself, is this requirement needed and feasible? If your answers to those questions change, it is OK to change the requirement.

Design requirements can fall into many different categories, such as size, cost, ease of use, and environmental impact, to name just a few.

One of the best ways to identify the design requirements for your project is to use the concrete example of a similar, existing product. Examine it in detail-take pictures, and take it apart if you have permission. Analyze how and why it works the way it does. Every single feature of the existing project represents a potential requirement for your design. (Of course, your design will have changes and improvements, so the requirements will not be identical.)

When you analyze an existing product, you build a mental library of techniques, mechanisms, and clever tricks. You acquire building blocks that you can use to construct your own designs. As you analyze more products, you can gain additional building blocks to use in your design. All designers do this!

How many design requirements should you have? For a project, three to five will often be a good number. For large, complex projects, there may be hundreds or even thousands of requirements.

First, as a group, your list of design requirements should provide a complete description of the key features that will make your design successful. Ask yourself, is anything missing?

Second, as a group, your list of design requirements should be feasible. Individually, your requirements might be feasible, but all together they might not be. For example, you might have time (or money or resources) to make one of them happen, but not all of them. Another potential problem might be that it is impossible to meet two or more of your requirements at the same time. For example, imagine that you are designing a toaster for a bagel shop. Two of your design requirements might be that the toaster needs to be large enough to toast ten bagels at a time, and it needs to fit on the bagel shop's counter. What if a toaster large enough to hold ten bagels at a time will not fit on the shop's counter? In cases like this, you must make a **trade-off**, a compromise or change in one or more requirements so that they can be met at the same time. In the toaster example, you would need to decide which is more important: toasting ten bagels at once or fitting the toaster on the counter? If the changes to your requirements make it impossible to solve your problem, you should look for a different problem to work on.

The Design Brief

To complete the requirements step of the design process, you should write a design brief. A design brief gathers all the key information for solving your problem in one place. It should contain:

- A description of your target user.
- A definition of the problem you intend to solve. [*Who*] need(s) [*what*] because [*why*].
- A description of how existing products are used and why they fail to address the problem.
- A list of all the requirements for your design.

For complex products with dozens of requirements or more, engineers supplement the design brief with detailed specifications such as a product requirements document (PRD) or a product design specification (PDS).

Design Brief Checklist

Answer the questions in the quick checklist below to grade your design brief.

What Makes a Good Design Brief?	For a Good Design Brief, You Should Answer "Yes" to Every Question
Does it define the problem you intend to solve? [*Who*] need(s) [*what*] because [*why*].	Yes / No
Does it describe how existing products are used and why they fail to address the problem?	Yes / No
Does it describe your target user?	Yes / No
Does it list all of the requirements for your design?	Yes / No
Is each design requirement *needed* to solve your problem? If it is not needed, leave it out. You'll have enough other things to work on!	Yes / No
Is each design requirement feasible? Ask if you have the time, money, materials, tools, and knowledge to make it happen. If you have conflicts between your requirements, have you investigated making trade-offs among them?	Yes / No

4. **Brainstorm Solutions**

There are always many good possibilities for solving design problems. If you focus on just one before looking at the alternatives, it is almost certain that you are overlooking a better solution. Good designers try to generate as many possible solutions as they can.

When solving a design problem, there are always many possible good solutions. If you focus on just one before looking at the alternatives, it is almost certain that you are overlooking a better solution. Good designers try to generate as many possible solutions as they can before choosing one that they feel is the best. Even "wild and crazy" design ideas that you end up rejecting might have some pieces that can make other designs better.

Ideation

Ideation, also known as idea generation, is the creative process of developing ideas. Start ideation after you have settled on a design problem that you want to solve and have done your background research, including the analysis of existing solutions. If you have not researched existing solutions, be sure to do so before starting ideation. Existing solutions are a great place to begin the ideation phase of your process because they give you a starting platform for ideas.

Generating lots of ideas is important to solving your design problem, so follow these key rules! One key rule for successful ideation is no limits. Start huge. Don't confine yourself to only one or two great ideas, and don't be afraid to think outside the box. No solutions are impossible during the ideation phase, so consider even the craziest of ideas. There will come a time later on when you will weigh your ideas against one another based on how easy they are to implement, but not yet. Ideation is the perfect time to put aside all judgment, and see how many design solutions you can come up with!

#1 Rule when Ideating: ***Don't settle for your first idea!***

If you think you have a great solution to your problem right from the beginning, you might be tempted to stick with that original idea. Even if it's the most perfect, without-a-doubt, best possible way to solve your problem -- don't stop here! Fixating on your first idea is a terrible mistake, because it stops your creative process before it even has the chance to get going. You never know what new ideas could branch off of your original idea or what new ideas might come to you over time, so you have to give the process (and yourself) a chance.

The lists are many creative techniques to help you come up with design ideas.

Existing Solutions

Existing solutions to your problem (or similar problems) are one of the best sources for creating design alternatives. Studying these designs will give you creative ideas of your own. Can the best features of existing solutions be combined in new ways? Can two entire solutions be combined to form one, better solution? Are there pieces missing from existing designs that if added, might make the designs more successful? Ask yourself these questions and see what new ideas you can come up with.

Analogies

By comparing your design problem to an entirely different situation, you may notice solutions that never would have come to mind otherwise. Try to create analogies between your design problem and random objects and people. For example, ask yourself:

- How is my design problem like [random object or problem]?
- How would I solve my problem using a [random object]?
- How would [random person, company, or group] solve my problem?

Choose random objects and people to create these analogies. Even though they may seem unrelated, the analogy will force your mind to come up with ideas to fit the specific cases of the random objects and people.

Example: Imagine you are designing a better lunchbox for students. Try these analogies to spark new and interesting design ideas...

Analogy: How is designing a lunchbox like designing a hotel?

Answer: When designing a hotel, you need to design for the people who will be staying in it. Think about the furniture, the decorations, the size of the rooms, etc. Try applying these to your lunchbox. What about the size of the lunchbox? Are there any components you could add to your lunchbox to serve as furniture-like features? Does the food in the lunchbox need furniture to sit on? You may never have considered these ideas without comparing a lunchbox to a hotel.

Analogy: How would I design a lunchbox using a skateboard?

Answer: You might create a lunchbox that has wheels, or a lunchbox that could be attached to a skateboard, or a skateboard that has a compartment to store food, or a lunchbox that could strap to the bottom of someone's feet. All of these are lunchbox designs that you might never have considered!

Analogy: How would Facebook design a lunchbox?

Answer: Facebook might design a lunchbox that you can take pictures with, or a lunchbox that has a computer screen on the inside. All of these are lunchbox designs you may never have thought about!

Brainstorming

Group brainstorming is a great way to generate lots and lots of ideas. Ask your friends, parents, and relatives if they would be willing to help you brainstorm ideas to your design problem. Gather a few of these people together for 30 minutes to an hour and tell them about your design problem. Then, leave the rest to discussion! Keep in mind:

- Fewer than five or six people per brainstorming session are best.
- No judgment! No ideas are bad ideas during ideation.
- Post-it notes are a great way for the people to show their ideas to the group.
- You should write down all of the ideas mentioned in your design notebook.

Sketching and Doodling

You can come up with great ideas by using all of the techniques listed, but ideation really isn't complete without sketching and doodling. Drawing is an ideal way to express your ideas and to visually connect multiple ideas to one another. Draw everything on your mind! Even if the idea is not fully developed, try to draw it and see what it looks like. Sketch all of the ideas that you have already come up with using other ideation techniques. By sketching, you will see new aspects of those ideas and be able to come up with even more.

"Sleep on It!"

Ideation isn't a one-day activity. In fact, it should be the longest phase of your entire design process. So don't feel like you need to come up with your perfect solution in one sitting. Ideate until you feel like you've run out of ideas. Then, sleep on it and return to ideation the next day or a few days after that. You will be surprised at how many more ideas you are able to come up with!

Create Multiple Solutions Checklist

Answer the questions in the quick checklist below to find out if you considered enough alternative (different) solutions.

What Makes Good Alternative Solutions?	For Good Alternative Solutions, You Should Answer "Yes" to Every Question
Did you use more than one of these ideation techniques to generate alternative solutions for your problem? • Examining existing solutions • Creating and using analogies • Conducting brainstorming sessions • Sketching and doodling	Yes / No
Did you come up with several possible solutions for your problem?	Yes / No

5. Choose the Best Solution

Look at whether each possible solution meets your design requirements. Some solutions probably meet more requirements than others. Reject solutions that do not meet the requirements.

Defining "Best"
Once you have created a number of possible solutions to your design problem, you need to choose which one is best.

Requirements
First, look at whether each possible solution met your design requirements. Consider solutions that did a much better job than others, and reject those that did not meet the requirements.

Nice to Haves, Desirables
In addition to your design requirements, you probably have some features that would be "nice to have" in your solution. These are things that are not quite as important as your design requirements; they are desirable, but not mandatory. Some of your possible solutions might include more of these nice-to-have features than others, and that is a possible reason they might be better.

Universal Design Criteria
Some criteria apply to virtually every design. Good designers consider them in every solution that they choose to implement.
Elegance. An elegant design solution is simple, clever, or ingenious. It might have fewer parts to wear out or fail. It might combine solutions from different areas in an inventive way not seen before. All good designers strive for elegance in their designs.
Robustness. A robust design is unlikely to fail, even when used in conditions more severe than it was designed for. It is sturdy or resilient, perhaps bending, but not breaking in hard use.
Aesthetics. If everything else is equal, people prefer a solution that is tasteful and pleasing to look at.
Cost. What will it cost? Can the target user afford the solution? Do you have enough money to build your prototype?
Resources. Do you have all the materials and equipment you need for your engineering project, or will you be able to obtain them quickly and at a very low cost?
Time. Do you have enough time to complete your design and make it before the due date? Allow time for doing additional research and fixing problems. It is very rare for everything to work correctly the first time.
Skill Required. Do you have the skills to build and implement your solution, or can you learn them in the time available?
Safety. Is your solution safe to build, use, store, and dispose of?

The Decision Matrix
As you compare potential solutions to your design brief and the universal criteria for a good design, it may be obvious which solution is the best. More often than not it helps to compare the solutions in a decision matrix. A decision matrix is a chart with your requirements and criteria on one axis and the different solutions on the other. Use a simple numeric evaluation scale to rate each solution against each of the criteria (2 = totally meets the criteria, 1 = somewhat meets the criteria, 0 = does not meet the criteria). Total up the columns to see which solution is best.

Alternatively, if you have colored stickers or pens, you can rate projects with a color scale (green = totally meets the criteria, yellow = somewhat meets the criteria, and red = does not meet the criteria). Using colors gives a highly visual indication of which solution is best (the more green the better!).

In our example, we lump together nice-to-have, desirable features and the universal design criteria into the "Other criteria" row of the decision matrix. That way these criteria serve as a tiebreaker, but they do not out-weigh "must-have" design requirements. You can make the design matrix with as many requirement rows and solution columns as you need, as shown in the examples.

Requirements and Criteria	Solution #1	Solution #2	Solution #3	Solution #4
Your requirement #1	1	2	2	1
Your requirement #2	1	1	2	1
Your requirement #3	2	2	2	2
Other criteria (nice-to-have and universal criteria)	1	1	2	1
Total Points	5	6	8	5

Pros and Cons
If your requirements and solutions are relatively simple, you can sometimes just list the pros and cons for each solution.

6. Develop the Solution
Development involves the refinement and improvement of a solution, and it continues throughout the design process, often even after a product ships to customers.

Development Work
Development involves the refinement and improvement of a solution.
Some development work might be needed very early in the design process in order to evaluate different potential solutions. For example, it might not be possible to compare the cost of early design concepts without doing some development work to estimate what the cost of an alternative would actually be.
Development work continues throughout the design process, often even after a product ships to customers.

Goals of Development Work

- Make It Work!

Of course, the primary goal of development work is to make a workable solution to your problem. This goal is true of all development efforts.

- Risk Reduction

In any complicated development effort, for example, building a robot for a competition, you will be uncertain about how well certain elements of your design solution will work. In a situation like this, it is very important to eliminate the uncertainty as soon as possible. This is called risk reduction. The longer you wait to eliminate the risk, the more likely that you will waste time on a solution that will fail. It is much better to find out that a potential solution will not work early in the design process before choosing your final solution for development. Fail fast, fail early!
To do risk reduction, use an appropriate method of development. For example, one strategy is to prototype just a small part of the potential solution, the risky part, to make sure that it will work.

- Optimization

Almost any design problem has multiple requirements. In many cases, requirements might conflict with each other, at least somewhat. For example, if you try to maximize almost any characteristic of a solution (speed, appearance, etc.), the cost will go up. Optimization is the process of finding the best trade-off between your different requirements, and it is an important part of almost every development effort.

Methods of Development

- Drawings

Designers use drawings to record ideas so that they are not forgotten, to communicate ideas to others, and to study how different parts of a design work together during development.

There are several types of drawings. Sketches are rough freehand drawings done very quickly and usually showing just the outlines of an object. Pictorial drawings portray a photo-like view of objects. Technical drawing is an accurate way of drawing that shows an object's true size and shape. It is often done with CAD (computer-aided design) software and is used in plans and blueprints that show how to construct an object. Technical drawings show in detail how the pieces of something relate to each other.

- Modeling

Models can be physical objects, such as a scale model of a solution that shows all the parts in correct proportion to each other. Generally, scale models do not actually work. Designers use them to visualize the solution and see how it looks.
A completely different kind of model is a mathematical or computer model. Designers use these models to predict how a solution will work. For complex problems, this can be extremely valuable, because the computer model is often much less expensive than building the solution itself, and it can assist the designer in making trade-offs among different requirements.

- Prototyping

A prototype is an operating version of a solution. Often a designer makes a prototype with different materials than the final version, and generally it is not as polished. Prototypes are a key step in the development of a final solution, allowing the designer to test how the solution will work and even show the solution to users for feedback.
Occasionally, designers will prototype pieces of the final solution very early in the design process. Sometimes designers will make several prototypes during the development of a solution.

- Storyboards

Storyboards are a series of graphic illustrations or images for the purpose of visualizing a video, website, software program, environment, user experience (like a theme park ride), or the like. Storyboards show how the solution appears as the user interacts with it over time, highlighting any problems in the flow of the experience.

- Analysis, Running the Numbers

Sometimes development work can be as simple as adding up the weight of all the components of a solution to see if the total weight meets the requirements. Similarly, you might add up the cost of all the parts to get a total cost or predict the speed of a vehicle by looking at the power of the engine. Analysis of this type is an important part of the development of many solutions, and it is often called running the numbers.

7. Build a Prototype

A prototype is an operating version of a solution. Often it is made with different materials than the final version, and generally it is not as polished.
Prototypes are a key step in the development of a final solution, allowing the designer to test how the solution will work and even show the solution to users for feedback. Occasionally, designers will prototype pieces of the final solution very early in the design process. Sometimes designers will make several prototypes during the development of a solution.
Prototypes can help you to develop the structure, function, and appearance of your solution.

8. Test and Redesign

The design process involves multiple iterations and redesigns of your final solution. You will likely test your solution, find new problems, make changes, and test new solutions before settling on a final design.
User Test

The first step in user testing is to get in contact with the users of your solution. Go back to your problem statement, and remember your potential users. You will want to test your solution on this group of people. For example, if you are designing a website for kindergarten students, you will want to test your final solution with kindergarteners.
When it comes to testing, there is no such thing as too many testers! The more people that you are able to test with the more you will find out. So try to find as many users as you can who are willing to help you. A good goal to reach for is three to five users for each round of testing.

The second step, after you have located three to five users, is to present your solution to these users while the users are in the **problem environment**. The problem environment is the situation or atmosphere in which the problem you are trying to solve happens.

- If your solution is a product, give the product to the users in the environment where they would use it. For example, if you designed a pair of sunglasses, you would give your sunglasses to the users outside while the weather is sunny. You wouldn't ask them to try the sunglasses at night or indoors, because those aren't situations where they would be using sunglasses.
- If your solution is a website or software product, ask users to test it on computers. You wouldn't want to just show them pictures or explain the website, because how they interact with the computer itself is also important.
- If your solution is an environment or experience, place your user in that environment or experience, and see how they react. For example, if you designed an after-school program for students, invite users to attend this program or a mock version of it. However, in many cases, having your users actually visit your designed environment or go through your designed experience will not be possible. An example is a student who designed a new school bus but obviously does not have an actual bus to use for testing. In a case like this, use a storyboard to present your solution to the users.

The third and final step of user testing is observing the interaction between the user and your solution. Record your observations in your design notebook. Watch closely as people use your product, navigate your website, or go through your designed environment or experience. Listen to what they say, but also watch what they do, see how they react, note where or when they get confused, and write down everything that happens during their interaction with the solution.

Below are three questions to ask during testing. The answers will be helpful when you move onto the redesign phase of test and redesign.

1. Are your users able to overcome the problem by using or interacting with your solution?
   - If yes, why are they successful?
   - If no, what problems do they encounter that prevent them from being successful?
2. Do the users ever need to ask you any questions when using or interacting with your solution?
   - If yes, what questions do they ask? During what part of their interaction do they ask these questions?
3. Do the users interact with your solution exactly the way that you intended for them to?
   - If no, what do they do differently?
4. If you have measurable targets for your solution, did you meet them?

Redesign

After you have tested your design, you will use your findings to complete a redesign of your solution. Use the findings from testing to:

- Fix any problems that occurred, and
- Further polish aspects of the design that were even more successful than you originally thought.

To make these changes, look at the answers to the four major questions you asked during testing:

1. Is your user able to overcome the problem by using or interacting with your solution?

If the answer is "yes," focus on why the user was successful. What specific aspects of your design helped the user to achieve success? Should those aspects become larger parts of your design? Should you make these features more

prominent or more obvious to the user? Consider emphasizing these aspects of your design. Then, in the next round of testing, see if the user is able to achieve success even more quickly and easily.
If the answer is "no," focus on the problems that users encountered during testing. What prevented them from achieving success? What changes to your design would eliminate these issues? Make these changes.

2. Does the user ever need to ask you any questions when using or interacting with your solution?

If the answer is "yes," focus on the questions that the users asked you. Why did they need to ask you a question? Were they confused? What part of the solution wasn't self-explanatory? You normally wouldn't be there to answer questions, so how can you make sure that the next users won't need to ask the same questions? Make changes that will eliminate these questions.

3. Does the user interact with your solution exactly the way that you intended for them to?

If the answer is "no," focus on what the users did that you hadn't intended to happen. Did their unexpected actions make your design more successful or less successful? If less successful, what changes could you make to your design to prevent these unexpected actions? What issues are causing the users to interact differently than intended, and how can you fix those issues? Make these changes.

4. If you have measurable targets for your solution, did you meet them?

If your design requirements call for your solution to be better, faster, or cheaper, you should measure the improvement that you made. If you met your targets, great! If not, how can you redesign your solution to improve its performance?

Once you have made changes to your design, go back and test again with your users. See if the improvements and changes you made negatively or positively affected your solution. Ask yourself the same four questions again, and then repeat the redesign again. Repeat this test and redesign process as many times as necessary to make your final solution as successful as possible. It may seem like you are doing the same thing over and over again, but with each test and redesign, you are greatly improving your project!

9. Communicate Results

To complete your project, communicate your results to others in a final report and/or a display board. Professional engineers always do the same, thoroughly documenting their solutions so that they can be manufactured and supported.

At this point, you are in the home stretch. Except for writing the abstract, preparing your project final report will just entail pulling together the information you have already collected into one large document.

Your final report will include these sections:
Title page.
Abstract. An abstract is an abbreviated version of your final report.
Table of contents.
Question, variables, and hypothesis.
Background research. This is the Research paper you wrote before you started your experiment.
Materials list.
Experimental procedure.
Data analysis and discussion. This section is a summary of what you found out in your experiment, focusing on your observations, data table, and graph(s), which should be included at this location in the report.
Conclusions.
Ideas for future research. Some science fairs want you to discuss what additional research you might want to do based on what you learned.
Acknowledgments. This is your opportunity to thank anyone who helped you with your science fair project, from a single individual to a company or government agency.
Writing a Bibliography (examples of APA & MLA styles).

Write the abstract section last, even though it will be one of the first sections of your final report.

Your final report will be several pages long, but don't be overwhelmed! Most of the sections are made up of information that you have already written. Gather up the information for each section and type it in a word processor if you haven't already.

Save your document often! You do not want to work hard getting something written the perfect way, only to have your computer crash and the information lost. Frequent file saving could save you a lot of trouble!

Remember to do a spelling and grammar check in your word processor. Also, have a few people proof read your final report. They may have some helpful comments!

Manufacturing

The term manufacturing refers to the processing of raw materials or parts into finished goods through the use of tools, human labor, machinery, and chemical processing.

Manufacturing allows businesses to sell finished products at a higher cost than the value of the raw materials used. Large-scale manufacturing allows for goods to be mass-produced using assembly line processes and advanced technologies as core assets. Efficient manufacturing techniques enable manufacturers to take advantage of economies of scale, producing more units at a lower cost.

Manufacturing is the process of turning raw materials or parts into finished goods through the use of tools, human labor, machinery, and chemical processing.
Manufacturing is integral to the economy.
Most products were handmade using human labor and basic tools before the Industrial Revolution.
The Industrial Revolution led to mass production, assembly line manufacturing, and the use of mechanization to manufacture larger quantities of goods at a lower cost.
Financial analysts study the ISM Manufacturing Report each month as a potential early indicator of the economy's health and where the stock market might be headed.

Understanding Manufacturing

Manufacturing is an integral and huge part of the economy. It involves the processing and refinement of raw materials, such as ore, wood, and foodstuffs, into finished products, such as metal goods, furniture, and processed foods.1

Converting these raw materials into something more useful adds value. This added value increases the price of finished products, making manufacturing a very profitable part of the business chain. Some people specialize in the skills required to manufacture goods, while others provide the funds that businesses need to purchase the tools and materials.

As noted above, efficiency in manufacturing can lead to higher productivity and cost savings. Manufacturers are able to accomplish this if they are able to:
Reduce redundancies
Improve the quality of work
Update equipment and procedures
Set realistic goals
Streamline intake, supply chain, and distribution channels2
Manufacturing is often reported on by the conference board and is well examined by economists.

Types of Manufacturing

How products are manufactured has changed over time. People have historically manufactured goods using raw materials. And in certain cases, they still do. Hand manufacturing involves the use of basic tools through more traditional processes. This form of manufacturing is often associated with decorative art, textile production, leatherwork, carpentry, and some metalwork.

Handmade goods are labor-intensive and require a lot of time. In some cases, they can command a high price, depending on the supplier and the type of goods. For instance, one-of-a-kind handmade fashion items can be sold at a higher price compared to something mass-produced. There are cases, though, where people who make goods using these techniques can be exploited, especially where labor laws are lax and demand for jobs is high.

Larger businesses use mechanization to mass-produce items on a much grander scale. This process involves the use of machines, which means that the manual manipulation of materials isn't necessarily required. Very little human capital is needed in the production process, although highly skilled individuals may be required to operate and ensure that machinery is running properly.

Manufacturing can fall into a few different categories, including:
Additive manufacturing: This type of manufacturing is commonly referred to as 3D printing. It involves the use of layers that are built up upon each other to create shapes and patterns in a three-dimensional process using a special piece of equipment, such as a 3D printer. 3
Advanced manufacturing: This method involves new forms of technology to improve the production process. Companies can add even more value to the raw materials they use to better serve their target markets. Newer technologies also help bring new products to market faster while increasing output. 4
Contract manufacturing: This is common in the manufacturing industry. Companies will enter into partnerships and business relationships with other firms to outsource certain manufacturing processes. For example, an automotive company may hire a third party to make parts that it will use in its assembly lines to make cars.5
3D printing has been around since the 1980s.6

Example of Manufacturing
Look around you and you'll find many examples of goods that have been manufactured. Clothing is just one example. Manufacturing clothing involves a number of steps and processes to go from raw materials to finished goods. The industry uses natural raw (cotton, silks, etc.) and synthetic materials (like rayon and polyester) to manufacture clothing. Companies may need both human labor and machinery to produce their goods through cutting, sewing, and other finishing techniques before they are sold to the public.

Business problems related to manufacturing :
1. Forecasting Demand for Products
Today, there are still many manufacturers who have difficulty forecasting future demand. The main problem is that they do not have advanced reporting tools that allow them to estimate how many items they should sell in the next few months or the following year. As a result, their products fail to meet the customer demand and they suffer lower sales.
The solutions:
In order to be able to forecast customer demand the right way, manufacturers should use accurate reporting tools that can make it easier for them to target sales and estimate how many and which items they should produce in the future.
In addition to utilizing software to make accurate forecasts, manufacturers also need to consider external events. For instance, currency exchange rates, rising fuel prices, current market trends, and so on. Make sure the sales and marketing teams are always up-to-date on these matters.

2. Controlling Inventory
Inventory management is still one of the main challenges in the manufacturing industry, but thanks to the help of automated solutions, it has become much simpler. Nevertheless, there are still many manufacturers, especially the small ones, who still manage their inventory manually.
Inventory tracking is a time-consuming process that can be streamlined with the help of software. Checking stock manually is very inefficient and prone to errors that can lead to inaccuracies, shortages and overstock, as well as unidentified damages.
The solutions:
To avoid unnecessary purchases of raw materials and equipment that lead to customer dissatisfaction, good inventory management strategies are highly needed. Routine inventory audits need to be carried out (both physical and cycle counts) to identify discrepancies between the numbers recorded in the company’s books and the actual number of items on hand. Barcode scanners can be used to speed up the tracking process.
Inventory managers need to ensure all the raw materials are sufficient for production and all the finished goods are ready to be delivered to end-users. This can be done very easily using reliable inventory management software. It allows

inventory managers to keep up with inventory levels at all times, receive alerts when inventory approaches the specified minimum amount, and reorder the right items at the right time.

Good inventory management software can also be integrated with barcode scanners to speed up the inventory tracking process and also with other software systems such as purchasing and accounting systems. Small-scale manufacturing businesses do not need to worry about having trouble investing in this software since there are affordable cloud solutions for them.

To be more precise, before deciding on getting the help of an inventory management software, it is better for business owners to learn about the software's pricing scheme calculation to avoid inefficient budgeting.

3. Improving Efficiency at Manufacturing Plants

Up until now, manufacturers have been looking for effective ways to reduce costs and improve efficiency at their plants. Many of them choose to sacrifice the quality of their products to reduce their production costs, but this will only reduce their profitability because dissatisfied customers will stop buying from them.

The solutions:

One of the most effective ways to optimize efficiency at manufacturing plants is to modernize the processes and systematize the workflows. Manufacturers need to minimize time-consuming and labor-intensive tasks, reduce material waste, optimize the use of equipment by minimizing damage, as well as simplify their supply chains. Enterprise Resource Planning (ERP) systems can facilitate all of them, thus enabling manufacturers to achieve optimal efficiency.

4. Increasing ROI

Any manufacturer would want to be able to increase their ROI. They would usually increase their sales or the price of their products. However, these aren't effective ways, especially when economic conditions are being erratic, reducing consumer purchasing power.

The solutions:

There are several ways to increase ROI such as:

i. Increase sales the right way

First, manufacturers should be able to clearly define their returns, since they can include higher sales, increased revenue, greater profits, reduced overhead or production costs, higher employee retention, and better customer satisfaction. Then, they need to set multiple benchmarks for their return goals. For example, instead of setting increased sales as a goal, they can increase sales during certain months in particular territories.

ii. Update marketing strategies

Manufacturers can also change their marketing strategies by utilizing digital marketing because the costs are cheaper than conventional methods (including placing advertisements on TV, newspapers, or billboards). This can be done by optimizing their website content, posting advertisements through search engines and social media, and personalizing email marketing campaigns. Furthermore, all these can now be automated by using advanced marketing automation software.

iii. Reduce costs

Manufacturers can reduce production costs by changing the design of packaging materials without having to sacrifice the quality of their products and negotiate with suppliers to get discounted prices.

With the help of reliable ERP software, manufacturers can not only reduce production costs, but also save labor costs, control the cost of procurement, and allocate costs appropriately.

5. Skilled Labor Shortage

Although automation and robotics can help fill the labor gap, human capabilities will still be needed to analyze and solve problems as well as manage outputs. With the baby boomer generation entering retirement, the manufacturing industry is facing a looming labor shortage. This is one of the biggest threats facing the manufacturing sector today.

The solutions:

To overcome the shortage of qualified workforce, manufacturers must be creative in recruiting employees by relying on various online platforms to post job openings. Producers must be truly selective when choosing prospective employees and ensure that they are able to work quickly, on target, under pressure, and do not object to shift work schedules.
In addition, manufacturers must be able to provide their new and existing employees with regular training to help increase their potential. Manufacturers need to consider using competency management software to bridge their workforce skill gaps.

6. Managing Sales Leads
Another challenge often faced by manufacturers is managing and prioritizing sales leads. Many of them treat their leads in the same way, but this isn’t the right method. Every sales leads have different characters, preferences, and needs. Therefore, manufactures should treat them specifically. Manufacturers also often find it difficult to identify potential leads so they often focus on unpromising opportunities and forget to follow up with high potential leads.
The solutions:
Manufacturers must truly understand their sales leads. In order to understand that, they need to dig deeper into their information. Manufacturers must have a system that makes it easy for them to keep, manage, and track prospect information.
The CRM system can help manufacturers understand their sales leads better through a centralized customer database, making it easier for them to identify qualified prospects so they can focus on the best opportunities, know the status of sales leads in real time without having to ask the sales team, as well as distribute leads to the right salespeople.

7. Coping with New Technological Advances
In reality, every year, there are always new technologies on horizon, which include the IoT, robotics, and manufacturing software. At the end, manufacturers left with a lot of decision that overwhelm them; Which ones do I should invest in? Will implementing this technology be a good decision for my manufacturing business?
The solutions:
Avoiding technology is certainly not a good move, because manufacturers do have to be able to adapt to any changes, including technology, as well as to be able to stay ahead in the highly competitive market. However, manufacturers should not be careless when deciding to implement automated solutions software.

Therefore, the best way that manufacturers can do is to discuss with all stakeholders and involve their employees. They have to figure out the difficulties they face on a daily basis. Manufacturers also need to consider the company’s budget to determine whether they are ready to invest in the desired technology. Should they invest in software like ERP or hardware like robotics.

When you start a small business or launch a startup, the one thing you can count on is the unexpected. No matter how thoroughly you plan, forecast, and test, problems are bound to arise. This is why as an entrepreneur, you need to know how to solve business problems effectively.

What is problem solving in business?
Problem solving in business relates to establishing processes that mitigate or remove obstacles currently preventing you from reaching strategic goals. These are typically complex issues that create a gap between actual results and your desired outcome. They may be present in a single team, operational process, or throughout your entire organization, typically without an immediate or obvious solution.

To approach problem solving successfully, you need to establish consistent processes that help you evaluate, explore solutions, prioritize execution, and measure success. In many ways, it should be similar to how you review business performance through a monthly plan review. You work through the same documentation, look for gaps, dig deeper to identify the root cause, and hash out options. Without this process, you simply cannot expect to solve problems efficiently or effectively.

Why problem solving is important for your business

While some would say problem-solving comes naturally, it's actually a skill you can grow and refine over time. Problem solving skills will help you and your team tackle critical issues and conflicts as they arise. It starts from the top. You as the business owner or CEO needing to display the type of level-headed problem solving that you expect to see from your employees.

Doing so will help you and your staff quickly deal with issues, establish and refine a problem solving process, turn challenges into opportunities, and generally keep a level head. Now, the best business leaders didn't just find a magic solution to solve their problems, they built processes and leveraged tools to find success. And you can do the same.

By following this 10-step process, you can develop your problem-solving skills and approach any issue that arises with confidence.

1. Define the problem

When a problem arises, it can be very easy to jump right into creating a solution. However, if you don't thoroughly examine what led to the problem in the first place, you may create a strategy that doesn't actually solve it. You may just be treating the symptoms.

For instance, if you realize that your sales from new customers are dropping, your first inclination might be to rush into putting together a marketing plan to increase exposure. But what if decreasing sales are just a symptom of the real problem?

When you define the problem, you want to be sure you're not missing the forest for the trees. If you have a large issue on your hands, you'll want to look at it from several different angles:

Competition

Is a competitor's promotion or pricing affecting your sales? Are there new entrants in your market? How are they marketing their product or business?

Business model

Is your business model sustainable? Is it realistic for how fast you want to grow? Should you explore different pricing or cost strategies?

Market factors

How are world events and the nation's economy affecting your customers and your sales?

Team

Are there any issues affecting your team? Do they have the tools and resources they need to succeed?

Goal alignment

Is everyone on your team working toward the same goal? Have you communicated your short-term and long-term business goals clearly and often?

There are a lot of ways to approach the issue when you're facing a serious business problem. The key is to make sure you're getting a full snapshot of what's going on so you don't waste money and resources on band-aid solutions.

Going back to our example, by looking at every facet of your business, you may discover that you're spending more on advertising than your competitors already. And instead, there's a communication gap within your team that's leading to the mishandling of new customers and therefore lost sales.

If you jumped into fixing the exposure of your brand, you would have been dumping more money into an area you're already winning. Potentially leading to greater losses as more and more new customers are dropped due to poor internal communication.

This is why it's so vital that you explore your blind spots and track the problem to its source.

2. Conduct a SWOT analysis

All good businesses solve some sort of problem for customers. What if your particular business problem is actually an opportunity, or even a strength if considered from a different angle? This is when you'd want to conduct a SWOT analysis to determine if that is in fact the case.

SWOT is a great tool for strategic planning and bringing multiple viewpoints to the table when you're looking at investing resources to solve a problem. This may even be incorporated in your attempts to identify the source of your problem, as it can quickly outline specific strengths and weaknesses of your business. And then by identifying any potential opportunities or threats, you can utilize your findings to kickstart a solution.

3. Identify multiple solutions with design thinking

As you approach solving your problem, you may want to consider using the design thinking approach. It's often used by organizations looking to solve big, community-based problems. One of its strengths is that it requires involving a wide range of people in the problem-solving process. Which leads to multiple perspectives and solutions arising.

This approach—applying your company's skills and expertise to a problem in the market—is the basis for design thinking.

It's not about finding the most complex problems to solve, but about finding common needs within the organization and in the real world and coming up with solutions that fit those needs. When you're solving business problems, this applies in the sense that you're looking for solutions that address underlying issues—you're looking at the big picture.

4. Conduct market research and customer outreach

Market research and customer outreach aren't the sorts of things small business owners and startups can do once and then cross off the list. When you're facing a roadblock, think back to the last time you did some solid market research or took a deep dive into understanding the competitive landscape.

Market research and the insights you get from customer outreach aren't a silver bullet. Many companies struggle with what they should do with conflicting data points. But it's worth struggling through and gathering information that can help you better understand your target market. Plus, your customers can be one of the best sources of criticism. It's actually a gift if you can avoid taking the negatives personally.

The worst thing you do when you're facing challenges is isolating yourself from your customers and ignore your competition. So survey your customers. Put together a competitive matrix.

5. Seek input from your team and your mentors

Don't do your SWOT analysis or design thinking work by yourself. The freedom to express concerns, opinions, and ideas will allow people in an organization to speak up. Their feedback is going to help you move faster and more efficiently. If you have a team in place, bring them into the discussion. You hired them to be experts in their area; use their expertise to navigate and dig deeper into underlying causes of problems and potential solutions.

If you're running your business solo, at least bring in a trusted mentor. SCORE offers a free business mentorship program if you don't already have one. It can also be helpful to connect with a strategic business advisor, especially if business financials aren't your strongest suit.

Quoting Stephen Covey, who said that "strength lies in differences, not in similarities," speaking to the importance of diversity when it comes to problem-solving in business. The more diverse a team is, the more often innovative solutions to the problems faced by the organization appear.

In fact, it has been found that groups that show greater diversity were better at solving problems than groups made up specifically of highly skilled problem solvers. So whoever you bring in to help you problem-solve, resist the urge to surround yourself with people who already agree with you about everything.

6. Apply lean planning for nimble execution

So you do your SWOT analysis and your design thinking exercise. You come up with a set of strong, data-driven ideas. But implementing them requires you to adjust your budget, or your strategic plan, or even your understanding of your target market.

Are you willing to change course? Can you quickly make adjustments? Well in order to grow, you can't be afraid to be nimble.

By adopting the lean business planning method—the process of revising your business strategy regularly—you'll be able to shift your strategies more fluidly. You don't want to change course every week, and you don't want to fall victim to shiny object thinking. But you can strike a balance that allows you to reduce your business's risk while keeping your team heading in the right direction.

Along the way, you'll make strategic decisions that don't pan out the way you hoped. The best thing you can do is test your ideas and iterate often so you're not wasting money and resources on things that don't work. That's Lean Planning.

7. Model different financial scenarios

When you're trying to solve a serious business problem, one of the best things you can do is build a few different financial forecasts so you can model different scenarios. You might find that the idea that seemed the strongest will take longer than you thought to reverse a negative financial trend. At the very least you'll have better insight into the financial impact of moving in a different direction.

The real benefit here is looking at different tactical approaches to the same problem. Maybe instead of increasing sales right now, you're better off in the long run if you adopt a strategy to reduce churn and retain your best customers. You won't know unless you model a few different scenarios. You can do this by using spreadsheets, and a tool like LivePlan can make it easier and quicker.

8. Watch your cash flow

While you're working to solve a challenging business problem, pay particular attention to your cash flow and your cash flow forecast. Understanding when your company is at risk of running out of cash in the bank can help you be proactive. It's a lot easier to get a line of credit while your financials still look good and healthy, than when you're one pay period away from ruin.

If you're dealing with a serious issue, it's easy to start to get tunnel vision. You'll benefit from maintaining a little breathing room for your business as you figure out what to do next.

9. Use a decision-making framework

Once you've gathered all the information you need, generated a number of ideas, and done some financial modeling, you might still feel uncertain. It's natural—you're not a fortune-teller. You're trying to make the best decision you can with the information you have. Do a matrix with the info and self-analyze.

10. Identify key metrics to track

How will you know your problem is solved? And not just the symptom—how will you know when you've addressed the underlying issues? Before you dive into enacting the solution, make sure you know what success looks like.

Decide on a few key performance indicators. Take a baseline measurement, and set a goal and a timeframe. You're essentially translating your solution into a plan, complete with milestones and goals. Without these, you've simply made a blind decision with no way to track success. You need those goals and milestones to make your plan real.

Problem solving skills to improve

1. Emotional intelligence

It can be very easy to make quick, emotional responses in a time of crisis or when discussing something you're passionate about. To avoid making assumptions and letting your emotions get the best of you, you need to focus on empathizing with others. This involves understanding your own emotional state, reactions and listening carefully to the responses of your team. The more you're able to listen carefully, the better you'll be at asking for and taking advice that actually leads to effective problem solving.

2. Research
Jumping right into a solution can immediately kill the possibility of solving your problem. Just like when you start a business, you need to do the research into what the problem you're solving actually is. Luckily, you can embed research into your problem solving by holding active reviews of financial performance and team processes. Simply asking "What? Where? When? How?" can lead to more in-depth explorations of potential issues.
The best thing you can do to grow your research abilities is to encourage and practice curiosity. Look at every problem as an opportunity. Something that may be trouble now, but is worth exploring and finding the right solution. You'll pick up best practices, useful tools and fine-tune your own research process the more you're willing to explore.

3. Brainstorming
Creatively brainstorming with your team is somewhat of an art form. There needs to be a willingness to throw everything at the wall and act as if nothing is a bad idea at the start. This style of collaboration encourages participation without fear of rejection. It also helps outline potential solutions outside of your current scope that you can refine and turn into realistic action.
Work on breaking down problems and try to give everyone in the room a voice. The more input you allow, the greater potential you have for finding the best solution.

4. Decisiveness
One thing that can drag out acting upon a potential solution, is being indecisive. If you aren't willing to state when the final cutoff for deliberation is, you simply won't take steps quickly enough. This is when having a process for problem solving comes in handy, as it purposefully outlines when you should start taking action.
Work on choosing decision-makers, identify necessary results and be prepared to analyze and adjust if necessary. You don't have to get it right every time, but taking action at the right time, even if it fails, is almost more vital than never taking a step.

5. Resilience (the capacity to recover quickly from difficulties; toughness)
Stemming off failure, you need to learn to be resilient. Again, no one gets it perfect every single time. There are so many factors in play to consider and sometimes even the most well-thought-out solution doesn't stick. Instead of being down on yourself or your team, look to separate yourself from the problem and continue to think of it as a puzzle worth solving. Every failure is a learning opportunity and it only helps you further refine and eliminate issues in your strategy.

Problem solving is a process
The key to effective problem-solving in business is the ability to adapt. You can waste a lot of resources on staying the wrong course for too long. So make a plan to reduce your risk now. Think about what you'd do if you were faced with a problem large enough to sink your business. Be as proactive as you can.

Business Analytics

Steps involved in BA process:

Step 1: Gather Background Info

This first step is where much of the ground work for a project is covered. Whether a project is brand new or existing, it's crucial for the business analyst to gather a significant amount of background information on the project

Step 2: Identify Stakeholders

The stakeholders on a project are the ones that give sign off on requirements, priorities, and make decisions. Therefore, identifying all of the stakeholders early on is important. A stakeholder wheel technique can be used which contains each stakeholder that impacts the project.

- Owners – shareholders, trustees or anyone who is sponsoring the project.
- Managers – senior or middle managers responsible for communication and monitoring the progress.
- Employees – developers, analysts, or testers responsible for delivering the project. Regulators – any regulators involved that monitor adherence to rules. For example, a regulator that monitors HIPAA compliance.
- Suppliers – any API provider or other supplier service needed that the project might interact with.
- Partners – individuals responsible for working alongside the project that provide complementary or supplementary products or services.
- Customers – the end users of the product.
- Competitors – a potential section of users of competitive products or inputs from the competitors themselves.

The document at this stage is the Stakeholder Matrix – a list of groups and stakeholders for each classification

Step 3: Discover Business Objectives

Establishing the business strategy and objectives and putting them on paper will help the business analyst and project managers stay focused on the vision and make course corrections along the way. It will also help during scope definition. Before getting too deep into the project, asking why the project is needed helps narrow down the business objectives.

A few techniques to assist in establishing business objectives for a project are:

- Benchmarking – understanding competitors and peers who work on the same level. An off's matrix can be used for this.
- SWOT Analysis – determine the strengths and weakness.
- Focus groups and brain storming.
- Documenting and conveying the business objectives to all of the stakeholders ensures a single vision.
- A business objective needs to be "SMART".
  - Specific –describes an outcome that is visible.
  - Measurable – outcome should be measurable.
  - Achievable – objective needs to be feasible to achieve.
  - Relevant – needs to be in alignment with the company's vision statement.
  - Time-bound – can be within a defined time frame.

The document at this stage is the: Business Objectives List

Step 4: Evaluate Options

To achieve the objective, it is important to determine the critical path among the various options available. Here are the steps involved in determining the best path:

- Identify options – brainstorming and focus group meetings help determine various options.
- Possible options can be:
- Customize/enhance an existing solution to achieve the business objective
- Purchase an existing service/system
- Build a product from scratch

- Integrate to other systems to achieve quick to market times.
- Shortlist options – narrowing down options from the big list to a short list of viable options.

A few factors to consider while narrowing this down are:
- Feasibility
- Budget/funds available
- Acceptable return on investment

Prepare a business case – based on the narrowed down options, a business analyst might need to prepare a business case using:
- Cost-benefit analysis – this looks at cost of pursuing an action and the benefits of that action.
- Impact analysis – identifying and presenting actions that affect the project or company that would have be factored into pursuing an action.
- Risk analysis – various risks that can be involved in pursuing an action
- Presentation of business case – based on the analysis of various options, the business analyst will present the options to the stakeholders to make a decision.

The document produced during this stage is a: Business Case Document

Step 5: Scope Definition
Based on the objective of the project and a team discussion, this step is when the scope is defined. A list of project development goals is detailed, along with a list of items that are not included in the project.
The scope definition document can include:
- Development modules in scope
- Development items out of scope
- Integrations in scope
- Integrations out of scope

The document produced in this stage is a: Scope Definition Document

Step 6: Business Analyst Delivery Plan
The business analyst and project owner will provide a detailed timeline for delivering the requirements to the development team.
A timeline for the requirements will be provided based on factors such as:
- Stakeholders and their availability
- Project scope
- Project methodology
- By dividing the requirements into deliverables and providing realistic dates for each of them, this will help plan resources and project timelines accordingly.

The document produced here is the: Business Analyst Delivery Plan

Step 7: Define Project Requirements
This step requires the business analyst to clarify the requirements to the business owner and get the ok to deliver them to the development team.
Requirements can be divided into functional and non-function. Non-functional requirements can be documented in a business requirement document which includes performance, scalability, and security.

Functional requirements are provided in terms of use cases, story boards, prototypes, and wireframes. A few techniques that help the requirement gathering phase are:
- Interviewing the stakeholders – asking when, how, where, and what is supposed to be achieved by the user helps arrive at requirements.

- Requirements defining techniques – use case templates, story boards, prototypes, or wireframes.

Based on the development method, requirements can be all delivered upfront e.g., the Waterfall technique. Most development companies shy away from Waterfall because it's difficult to accommodate changes along the way. For Agile projects, requirements can be delivered per sprint cycle. A business analyst will sequence deliverables to facilitate development plans.

Documents required in this stage are:

- Non-functional Requirements – business requirement document
- Functional Requirements – use cases

Step 8: Support Implementation through SDLC

A business analyst is involved through the technical implementation of requirements to ensure that everything aligns. These are a few steps during this phase:

- Reviewing the technical deliverables to align with requirements.
- Based on feedback from the development team, update or repackage requirements to facilitate implementation.
- Engage with quality analysts to ensure requirements are tested and requirements are understood.
- Manage changes from the business owner that are requested once the initial requirements are delivered and implemented.
- Facilitate user acceptance once the requirement implementation is done.

Step 9: Evaluate Value Added By Project

To maintain the business objective through the implementation, a constant evaluation on business outcomes needs to be maintained. Questions to ask can be – Are we on track? Is the solution delivering the value initially anticipated? A few key points are to:

- Evaluate the actual progress across the timeline and business objectives and provide stakeholders updates and answer questions.
- Based on the progress and feedback, suggest any modifications or initiatives required to realign the implementation phase with business objectives.

Understanding Business problems related to various businesses

Business Analytics:

The BA process can solve problems and identify opportunities to improve business performance. In this process, organizations may also determine strategies to guide operations and help to achieve competitive advantages. Typically, solving problems and identifying strategic opportunities to follow organization decision-making tasks. The latter, identifying opportunities can be viewed as a problem of strategy choice requiring a solution.

Business analytics begins with a *data set* (a simple collection of data or a data file) or commonly with a *database* (a collection of data files that contain information on people, locations, and so on). As databases grow, they need to be stored somewhere. Technologies such as *computer clouds* (hardware and software used for data remote storage, retrieval, and computational functions) and *data warehousing* (a collection of databases used for reporting and data analysis) store data. Database storage areas have become so large that a new term was devised to describe them. *Big data* describes the collection of data sets that are so large and complex that software systems are hardly able to process them (Isson and Harriott, 2013, pp. 57–61). Isson and Harriott (2013, p. 61) define *little data* as anything that is not big data. Little data describes the smaller data segments or files that help individual businesses keep track of customers. As a means of sorting through data to find useful information, the application of analytics has found new purpose.

Three terms in business literature are often related to one another: analytics, business analytics, and business intelligence. *Analytics* can be defined as a process that involves the use of statistical techniques (measures of central tendency, graphs, and so on), information system software (data mining, sorting routines), and operations research methodologies (linear programming) to explore, visualize, discover and communicate patterns or trends in data. Simply, analytics convert data into useful information. Analytics is an older term commonly applied to all disciplines, not just business. A typical example of the use of analytics is the weather measurements collected and converted into statistics, which in turn predict weather patterns.

There are many types of analytics, and there is a need to organize these types to understand their uses. We will adopt the three categories (*descriptive*, *predictive*, and *prescriptive*) that the *Institute of Operations Research and Management Sciences* (INFORMS) organization (www.informs.org) suggests for grouping the types of analytics (see Table 1.1). These types of analytics can be viewed independently. For example, some firms may only use descriptive analytics to provide information on decisions they face. Others may use a combination of analytic types to glean insightful information needed to plan and make decisions.

Types of Analytics

Type of Analytics	**Definition**
Descriptive	The application of simple statistical techniques that describes what is contained in a data set or database. Example: An age bar chart is used to depict retail shoppers for a department store that wants to target advertising to customers by age.
Predictive	An application of advanced statistical, information software, or operations research methods to identify predictive variables and build predictive models to identify trends and relationships not readily observed in a descriptive analysis. Example: Multiple regression is used to show the relationship (or lack of relationship) between age, weight, and exercise on diet food sales. Knowing that relationships exist helps explain why one set of independent variables influences dependent variables such as business performance.
Prescriptive	An application of decision science, management science, and operations research methodologies (applied mathematical techniques) to make best use of allocable resources. Example: A department store has a limited advertising budget to target customers. Linear programming models can be used to optimally allocate the budget to various advertising media.

The purposes and methodologies used for each of the three types of analytics differ. It is these differences that distinguish *analytics* from *business analytics*. Whereas analytics is focused on generating insightful information from data sources, business analytics goes the extra step to leverage analytics to create an improvement in measurable business performance. Whereas the process of analytics can involve any one of the three types of analytics, the major components of business analytics include all three used in combination to generate new, unique, and valuable information that can aid business organization decision-making. In addition, the three types of analytics are applied sequentially (descriptive, then predictive, then prescriptive). Therefore, *business analytics* (BA) can be defined as a process beginning with business-related data collection and consisting of sequential application of descriptive, predictive, and prescriptive major analytic components, the outcome of which supports and demonstrates business decision-making and organizational performance. Stubbs (2011, p. 11) believes that BA goes beyond plain analytics, requiring a clear relevancy to business, a resulting insight that will be implementable, and performance and value measurement to ensure a successful business result.

Analytic Purposes and Tools

Type of Analytics	**Purpose**	**Examples of Methodologies**

Type of Analytics	Purpose	Examples of Methodologies
Descriptive	To identify possible trends in large data sets or databases. The purpose is to get a rough picture of what generally the data looks like and what criteria might have potential for identifying trends or future business behavior.	Descriptive statistics, including measures of central tendency (mean, median, mode), measures of dispersion (standard deviation), charts, graphs, sorting methods, frequency distributions, probability distributions, and sampling methods.
Predictive	To build predictive models designed to identify and predict future trends.	Statistical methods like multiple regression and ANOVA. Information system methods like data mining and sorting. Operations research methods like forecasting models.
Prescriptive	To allocate resources optimally to take advantage of predicted trends or future opportunities.	Operations research methodologies like linear programming and decision theory.

Business intelligence (BI) can be defined as a set of processes and technologies that convert data into meaningful and useful information for business purposes. While some believe that BI is a broad subject that encompasses analytics, business analytics, and information systems (Bartlett, 2013, p.4), others believe it is mainly focused on collecting, storing, and exploring large database organizations for information useful to decision-making and planning (Negash, 2004). One function that is generally accepted as a major component of BI involves storing an organization's data in computer cloud storage or in data warehouses. Data warehousing is not an analytics or business analytics function, although the data can be used for analysis. In application, BI is focused on querying and reporting, but it can include reported information from a BA analysis. BI seeks to answer questions such as what is happening now and where, and also what business actions are needed based on prior experience. BA, on the other hand, can answer questions like why something is happening, what new trends may exist, what will happen next, and what is the best course for the future.

In summary, BA includes the same procedures as in plain analytics but has the additional requirement that the outcome of the analytic analysis must make a measurable impact on business performance. BA includes reporting results like BI but seeks to explain why the results occur based on the analysis rather than just reporting and storing the results, as is the case with BI.

.

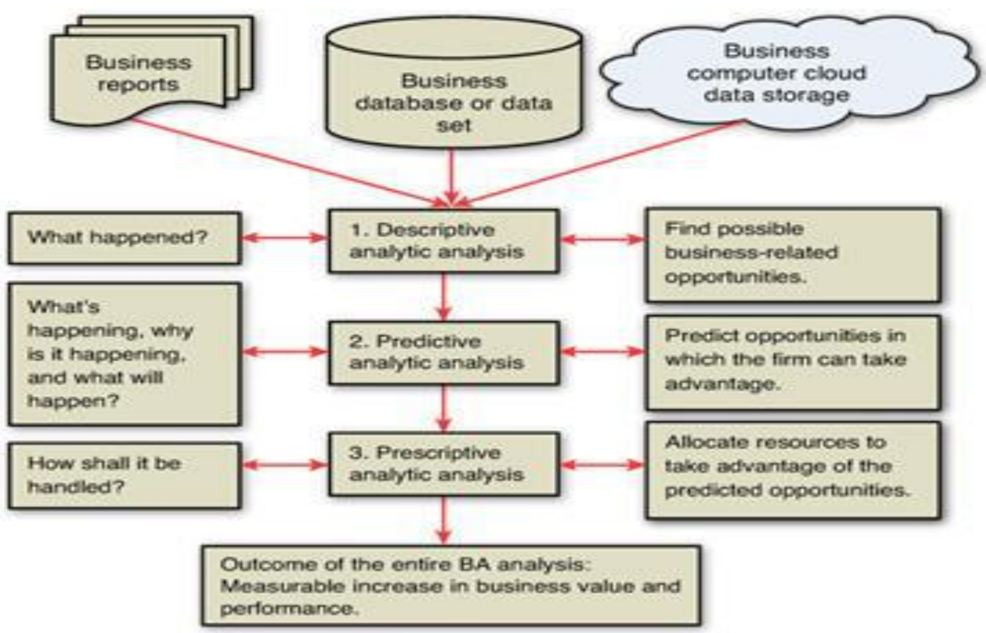
Characteristics of Analytics, Business Analytics, and Business Intelligence

Characteristics	Analytics	Business Analytics (BA)	Business Intelligence (BI)
Business performance planning role	What is happening, and what will be happening?	What is happening now, what will be happening, and what is the best strategy to deal with it?	What is happening now, and what have we done in the past to deal with it?
Use of descriptive analytics as a major component of analysis	Yes	Yes	Yes
Use of predictive analytics as a major component of analysis	Yes	Yes	No (only historically)

Characteristics	Analytics	Business Analytics (BA)	Business Intelligence (BI)
Use of prescriptive analytics as a major component of analysis	Yes	Yes	No (only historically)
Use of all three in combination	No	Yes	No
Business focus	Maybe	Yes	Yes
Focus of storing and maintaining data	No	No	Yes
Required focus of improving business value and performance	No		

Business Analytics Process

The complete *business analytic process* involves the three major component steps applied sequentially to a source of data. The outcome of the business analytic process must relate to business and seek to improve business performance in some way.



Business analytic process

The logic of the BA process is initially based on a question:
What valuable or problem-solving information is locked up in the sources of data that an organization has available?
At each of the three steps that make up the BA process, additional questions need to be answered.
Answering all these questions requires mining the information out of the data via the three steps of analysis that comprises the BA process. The analogy of digging in a mine is appropriate for the BA process because finding new, unique, and valuable information that can lead to a successful strategy is just as good as finding gold in a mine. SAS, a major analytic corporation (www.sas.com), actually has a step in its BA process, *Query Drilldown*, which refers to the mining effort of

questioning and finding answers to pull up useful information in the BA analysis. Many firms routinely undertake BA to solve specific problems, while other firms undertake BA to explore and discover new knowledge to guide organizational planning and decision-making to improve business performance.

The size of some data sources can be unmanageable, overly complex, and generally confusing. Sorting out data and trying to make sense of its informational value requires the application of descriptive analytics as a first step in the BA process. One might begin simply by sorting the data into groups using the four possible classification. Also, incorporating some of the data into spreadsheets like Excel and preparing cross tabulations and contingency tables are means of restricting the data into a more manageable data structure. Simple measures of central tendency and dispersion might be computed to try to capture possible opportunities for business improvement. Other descriptive analytic summarization methods, including charting, plotting, and graphing, can help decision makers visualize the data to better understand content opportunities.

Types of Data Measurement Classification Scales

Type of Data Measurement Scale	**Description**
Categorical Data	Data that is grouped by one or more characteristics. Categorical data usually involves cardinal numbers counted or expressed as percentages. Example 1: Product markets that can be characterized by categories of "high-end" products or "low-income" products, based on dollar sales. It is common to use this term to apply to data sets that contain items identified by categories as well as observations summarized in cross-tabulations or contingency tables.
Ordinal Data	Data that is ranked or ordered to show relational preference. Example 1: Football team rankings not based on points scored but on wins. Example 2: Ranking of business firms based on product quality.
Interval Data	Data that is arranged along a scale where each value is equally distant from others. It is ordinal data. Example 1: A temperature gauge. Example 2: A survey instrument using a Likert scale (that is, 1, 2, 3, 4, 5, 6, 7), where 1 to 2 is perceived as equidistant to the interval from 2 to 3, and so on. Note: In ordinal data, the ranking of firms might vary greatly from first place to second, but in interval data, they would have to be relationally proportional.
Ratio Data	Data expressed as a ratio on a continuous scale. Example 1: The ratio of firms with green manufacturing programs is twice that of firms without such a program.

From Step 1 in the *Descriptive Analytic analysis*, some patterns or variables of business behavior should be identified representing targets of business opportunities and possible (but not yet defined) future trend behavior. Additional effort (more mining) might be required, such as the generation of detailed statistical reports narrowly focused on the data related to targets of business opportunities to explain what is taking place in the data (what happened in the past). This is like a statistical search for predictive variables in data that may lead to patterns of behavior a firm might take advantage of if the patterns of behavior occur in the future. For example, a firm might find in its general sales information that during economic downtimes, certain products are sold to customers of a particular income level if certain advertising is undertaken. The sales, customers, and advertising variables may be in the form of any of the measurable scales for data, but they have to meet the three conditions of BA previously mentioned: clear relevancy to business, an implementable resulting insight, and performance and value measurement capabilities.

To determine whether observed trends and behavior found in the relationships of the descriptive analysis of Step 1 actually exist or hold true and can be used to forecast or predict the future, more advanced analysis is undertaken in Step 2, *Predictive Analytic analysis*, of the BA process.
There are many methods that can be used in this step of the BA process.

A commonly used methodology is multiple regression. This methodology is ideal for establishing whether a statistical relationship exists between the predictive variables found in the descriptive analysis. The relationship might be to show that a dependent variable is predictively associated with business value or performance of some kind.

For example, a firm might want to determine which of several promotion efforts (independent variables measured and represented in the model by dollars in TV ads, radio ads, personal selling, and/or magazine ads) is most efficient in generating customer sale dollars (the dependent variable and a measure of business performance). Care would have to be taken to ensure the multiple regression model was used in a valid and reliable way, which is why ANOVA and other statistical confirmatory analyses are used to support the model development.

Exploring a database using advanced statistical procedures to verify and confirm the best predictive variables is an important part of this step in the BA process. This answers the questions of what is currently happening and why it happened between the variables in the model.

A single or multiple regression model can often forecast a trend line into the future. When regression is not practical, other forecasting methods (exponential smoothing, smoothing averages) can be applied as predictive analytics to develop needed forecasts of business trends. The identification of future trends is the main output of Step 2 and the predictive analytics used to find them. This helps answer the question of what will happen.

If a firm knows where the future lies by forecasting trends as they would in Step 2 of the BA process, it can then take advantage of any possible opportunities predicted in that future state. In Step 3, *Prescriptive Analytics analysis*, operations research methodologies can be used to optimally allocate a firm's limited resources to take best advantage of the opportunities it found in the predicted future trends. Limits on human, technology, and financial resources prevent any firm from going after all opportunities they may have available at any one time. Using prescriptive analytics allows the firm to allocate limited resources to optimally achieve objectives as fully as possible. For example, *linear programming* (a constrained optimization methodology) has been used to maximize the profit in the design of supply chains (Paksoy et al., 2013). This third step in the BA process answers the question of how best to allocate and manage decision-making in the future.

Relationship of BA Process and Organization Decision-Making Process

The BA process can solve problems and identify opportunities to improve business performance. In the process, organizations may also determine strategies to guide operations and help achieve competitive advantages. Typically, solving problems and identifying strategic opportunities to follow are organization decision-making tasks. The latter, identifying opportunities can be viewed as a problem of strategy choice requiring a solution. It should come as no surprise that the BA process closely parallels classic organization decision-making processes. The business analytic process has an inherent relationship to the steps in typical organization decision-making processes.

Comparison of business analytics and organization decision-making processes

The *organization decision-making process* (ODMP) developed by Elbing (1970) and presented in above Figure is focused on decision making to solve problems but could also be applied to finding opportunities in data and deciding what is the best course of action to take advantage of them. The five-step ODMP begins with the perception of disequilibrium, or the awareness that a problem exists that needs a decision. Similarly, in the BA process, the first step is to recognize that databases may contain information that could both solve problems and find opportunities to improve business performance. Then in Step 2 of the ODMP, an exploration of the problem to determine its size, impact, and other factors is undertaken to diagnose what the problem is. Likewise, the BA descriptive analytic analysis explores factors that might prove useful in solving problems and offering opportunities. The ODMP problem statement step is similarly structured to the BA predictive analysis to find strategies, paths, or trends that clearly define a problem or opportunity for an organization to solve problems. Finally, the ODMP's last steps of strategy selection and implementation involve the same kinds of tasks that the BA process requires in the final prescriptive step (make an optimal selection of resource allocations that can be implemented for the betterment of the organization).

The decision-making foundation that has served ODMP for many decades parallels the BA process. The same logic serves both processes and supports organization decision-making skills and capacities.

Business intelligence (BI)

Business intelligence (BI) is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. BI technologies are capable of handling large amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities.

The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability. BI technologies provide historical, current and predictive views of business operations.

Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data, can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data.

Purpose of Business Intelligence:

Business intelligence can be applied to the following business purposes, in order to drive business value.

- Measurement: program that creates a hierarchy of performance metrics (see also Metrics Reference Model) and benchmarking that informs business leaders about progress towards business goals (business process management).
- Analytics: program that builds quantitative processes for a business to arrive at optimal decisions and to perform business knowledge discovery. Frequently involves: data mining, process mining, statistical analysis, predictive analytics, predictive modeling, business process modeling, data lineage, and complex event processing and prescriptive analytics.
- Reporting/enterprise reporting: program that builds infrastructure for strategic reporting to serve the strategic management of a business, not operational reporting. Frequently involves data visualization, executive information system and OLAP.
- Collaboration/collaboration platform: program that gets different areas (both inside and outside the business) to work together through data sharing and electronic data interchange.
- Knowledge management: program to make the company data-driven through strategies and practices to identify, create, represent, distribute, and enable adoption of insights and experiences that are true business knowledge.

Knowledge management leads to learning management and regulatory compliance.

Requirements gathering: Gather all the Data related to Business objective

There are many different approaches that can be used to gather information about a business. They include the following:

- Review business plans, existing models and other documentation
- Interview subject area experts
- Conduct fact-finding meetings
- Analyze application systems, forms, artifacts, reports, etc.

The business analyst should use one-on-one interviews early in the business analysis project to gauge the strengths and weaknesses of potential project participants and to obtain basic information about the business. Large meetings are not a good use of time for data gathering.

- Facilitated work sessions are a good mechanism for validating and refining “draft” requirements. They are also useful to prioritize final business requirements. Group dynamics can often generate even better ideas.
- Primary or local data is collected by the business owner and can be collected by survey, focus group or observation.
- Third party static data is purchased in bulk without a specific intent in mind. While easy to get (if you have the cash) this data is not specific to your business and can be tough to sort through as you often get quite a bit more data than you need to meet your objective.
- Dynamic data is collected through a third party process in near real-time from an event for a specific purpose (read into that VERY expensive).

Requirement gathering through surveys:

Surveys should follow a few basic and important rules:

a) Keep it VERY simple. It is recommended that to have one page with 3-4 questions maximum. Customers are visiting to purchase or to have an experience, not to fill out surveys.

b) Choose only one objective for the survey. Don’t try to answer too many questions; ultimately you won’t get much useful data that way because your customer will get confused and frustrated.
c) Don’t give the respondent any wiggle room. Open ended questions are tough to manage. Specific choices that are broad enough to capture real responses give you data that is much easier to use.
d) Always gather demographics. Why not? But rather than name and e-mail (leading to concerns with confidentiality and often less than truthful answers) gather gender, age and income; you might be surprised at who is actually buying what.

Interpret Data to make it useful for Business:
Business intelligence (BI) is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. BI technologies are capable of handling large amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities.

Goal of BI:
The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

Functions of BI:
BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

Uses:
BI can be used to support a wide range of business decisions ranging from operational to strategic.
- Basic operating decisions include product positioning or pricing.
- Strategic business decisions include priorities, goals and directions at the broadest level.

In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data.

Business intelligence is made up of an increasing number of components including:
- Multidimensional aggregation and allocation
- De-normalization, tagging and standardization
- Real time reporting with analytical alert
- A method of interfacing with unstructured data sources
- Group consolidation, budgeting and rolling forecasts
- Statistical inference and probabilistic simulation
- Key performance indicators optimization
- Version control and process management
- Open item management

SMART Utilities

S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology; often written as SMART) is a monitoring system included in computer hard disk drives (HDDs) and solid-state drives (SSDs) that detects and reports on various indicators of drive reliability, with the intent of enabling the anticipation of hardware failures.

When S.M.A.R.T. data indicates a possible imminent drive failure, software running on the host system may notify the user so stored data can be copied to another storage device, preventing data loss, and the failing drive can be replaced.

Smart utilities are companies in the electric, gas and water sectors that employ connected sensors across their grids in order to deliver services more efficiently and to analyze their operations. Most smart utilities are heavy users of the Internet of Things (IoT) concept of connected devices and incorporate the latest digital technologies in communications, software, computing and mapping to streamlines their businesses.

All utilities deal with an enormous amount of diverse data from multiple sources. Setting smart utilities apart from those slower to embrace digital technologies and change are the business benefits they frequently reap in diverse areas such as:

- Optimizing asset performance by identifying developing problems in real-time for field assets such as transformers and water mains in order to trigger repairs and avoid unplanned service disruptions.
- Using real-time business intelligence to improve operational efficiency, monitor remote sensors and meters, and make better, timelier business decisions.
- Leveraging the smart grid and intelligence from the Internet of Things (IoT) to provide key insights into customers' utility usage patterns.
- Preventing utility loss and provide early detection of supply leaks and thefts to optimize distribution.
- Modernizing the grid to more dynamically adjust and respond to multiple systems and providers, including alternative energy sources such as wind and solar.

IoT Analytics

Internet of Things (IoT) analytics is a data analysis tool that assesses the wide range of data collected from IoT devices. IoT analytics assesses vast quantities of data and produces useful information from it.

IoT analytics are usually discussed in tandem with Industrial IoT (IIoT). Data is collected from a wide range of sensors on manufacturing infrastructure, weather stations, smart meters, delivery trucks, and all forms of machinery. IoT analytics can be applied to managing data centers and applications handling retail and healthcare.

In many ways, IoT data is similar to big data. The key difference between the two is not just the quantity of data but the range of sources it is obtained from. All this data has to be processed into one comprehensible, single stream of data. Considering the several kinds of sources of information, data integration becomes quite difficult, and this is where IoT analytics makes a difference, though it can be tough to develop and implement.

Devices that power IoT Analytics

There are a wide range of IoT devices that collect data:

- Wearables

Dedicated trackers such as Fitbit or other smartwatches have gone beyond tracking steps. You can track your friends' fitness activities, compete with them, message, and even answer the phone by connecting your devices through the Internet. This information is tracked by fitness companies, enabling them to create customized packages if you sign up. This can include exercise routines, diet, goals, and more. The newest smart watches even monitor heart rates and rhythms and have accurately diagnosed heart problems in their wearers.

- Smart Home

Smart homes have security systems you can access and control when you are away from home, to appliances you can turn on and off with digital assistance. There is a wide range of devices that you can incorporate into your home and a wide range of data that can be collected to assess usage patterns, the efficacy of systems, and more.

- Healthcare

Healthcare has a wide range of IoT devices. Bluetooth technology creates hearing aids, records heart and blood pressure, and monitors pulse-based alarm systems that can call for help. This has helped enhance healthcare to a large extent. The data collected is invaluable in terms of creating newer and better technology.

- Voice-Activated Everything

Digital assistants are a form of IoT devices. Alexa, Siri, and Google take notes, find information, play music, order cabs, tell the weather, set alarms, and everything else. The internet regularly updates these digital assistants to improve functionality. Their data helps companies tailor their services for you, based on your everyday interaction with digital assistants.

How IoT Analytics Work and the Applications

With a wide range of devices, there is an endless stream of data in enormous quantities. IoT analytics helps analyze this data across all connected devices without hardware or infrastructure. As the needs of your organization change, computing power and data storage scales up or down accordingly, ensuring your IoT analysis has the requisite capacity.

1. The first step is to collect data aggregated from a variety of sources, in a range of formats, and at multiple frequencies.
2. This data is then processed with a wide range of external sources.
3. The information is then stored in a time-series for analysis.
4. The analysis can be done in multiple ways--with custom analysis systems, with standard SQL queries, or with machine learning analysis techniques. The results can be used to make a wide range of predictions.
5. With the information received, organizations can build several systems and applications to ease business processes.

Business Use Cases of IoT Analytics

- Smart agriculture: With IoT analytics, connected field machinery works based on information derived from IoT analysis. Analysis factors include time, geographical location, weather, altitude, and local environmental conditions. For example, irrigation systems can be optimized to deliver the exact amount of water as rainfall predictions.
- Regular restocking of supplies: Monitor inventories in real-time. A food vending company, with connected machines, can have their machines request restocking based on the depletion of products. This can be triggered when stocks in the machine reach a particular level.
- Predictive maintenance: Varying infrastructure needs regular maintenance. With IoT analysis, pre-set templates can help determine quality predictive maintenance models applied to specific needs. For example, in long-distance transport vehicles with heating and cooling systems--IoT analytics can determine when vehicles need an overhaul to ensure cargos are not damaged.
- Process efficiency scoring: Every company works with a range of processes in place. IoT analytics can measure the efficiency of these processes and make the necessary changes in them. Results from IoT analytics can identify bottlenecks--both current and potential--and can increase efficiencies.

Role of data analytics in the oil and gas industry

Big data and analytics may be new to some industries, but the oil and gas industry has long been dealing with large quantities of data to make technical decisions. In their quest to learn what lies below the surface and how to bring it out, energy companies have, for many years, invested in seismic software, visualization tools and other digital technologies.

The rise of pervasive computing devices, affordable sensors that collect and transmit data, new analytics tools and advanced storage capabilities is opening up more possibilities every year. Oil producers can capture more detailed data in real time at lower costs and from previously inaccessible areas to improve oilfield and plant performance. Oil and gas companies will need to improve their analytics capabilities in order to compete in an industry where decisions are being made faster and the stakes are growing higher.

Data-driven analytics is becoming an important point of competitive differentiation in the oil and gas industry. It is a set of tools and techniques to infer patterns and trends in data, and construct predictive models, which can assist in decision-making and optimization.

Oil and gas companies can leverage big data technologies, for instance, like the MapR Converged Data Platform which can be used for collecting, managing and rapidly analyzing seismic drilling and production data. Moreover, companies can leverage data to acquire new insights that can help boost drilling and production performance while addressing safety and environmental concerns. Additionally, MapR can help businesses capitalize on big data to optimize their operations, achieve lower costs and boost their competitive edge.

Business problems related to smart utilities:
Key challenges today include:

- National and international regulations driving the replacement of fossil fuel generation with cleaner, renewable methods of generating electricity to reduce Co2 emissions
- The distributed nature of renewable energy leading to a vast increase in the number of assets needing to be managed in real-time
- Secure monitoring of energy flows for management and financial settlement purposes as households and businesses become prosumers (a person who buys electronic goods that are of a standard between those aimed at consumers and professionals; a consumer who becomes involved with designing or customizing products for their own needs).
- Flat customer demand in developed economies and rising demand in developing economies
- The balancing of supply and demand to make electricity grids more efficient
- The rapid growth in grid scale storage for demand and frequency response applications
- The electrification of transport, resulting in millions of batteries being connected to the grid and time-shifting of peak demand

None of these challenges can be met without wholesale application of IoT technology.

Production Lines in Data Analytics

Data is everywhere on today's production lines. The challenge is that though manufacturers may be collecting large volumes of data, it often isn't being used proactively or efficiently to solve day-to-day problems. Data gathered from your manufacturing processes has the potential to improve your line efficiency and FTY (First Time Yield), boost product quality, and in some circumstances, even save your — but only if the systems are set up to allow for this kind of insight and action through data analysis.

A production line, also called an **assembly line** or factory line, refers to the organized path of assembly for a product. In most cases, the production line centers on a **conveyor** belt or other mechanical system that physically moves the product from one station to the next, and it is a common misnomer to call this conveyor a production line. At each station in the production process, a factory worker or machine adds a piece to the finished product, performs a **quality control** check or some other job that is essential to the completion of the project.

Consider the following best practices to start making better use of your production line data.
1. Consolidate your production data
Bring all of the data together from your plant floor systems and consolidate it into an accessible structure. A modern data management and analytics platform has the capability to ingest data from disparate sources – it is flexible and agnostic. This will eliminate the data silos common with using different machine vendors and allow you to scale down what it takes to manage the various types of software and hardware on the production floor. It will also enable you to do correlations from stations across the line, provide consistent reporting and a top-level view of manufacturing processes.

During this process, consider how you want to use the data to ensure you are collecting the right amount of useful information. Depending upon the size of your organization and the types of parts manufactured your needs may differ. If creating standard production reports is the main requirement, for example, less data is required than for test station optimization.

2. Standardize your production data reporting model

Having a standard data model for your plant floor systems greatly simplifies implementation and assures project success by levering standard off-the-shelf reporting. For instance, if your data model standard identifies traceability to a specific part by serial number (highly recommended), this will ensure all tooling suppliers program their logic controllers accordingly, or have the appropriate barcode readers/scanners or part-marking systems in place.

3. Look at parts, not just machines

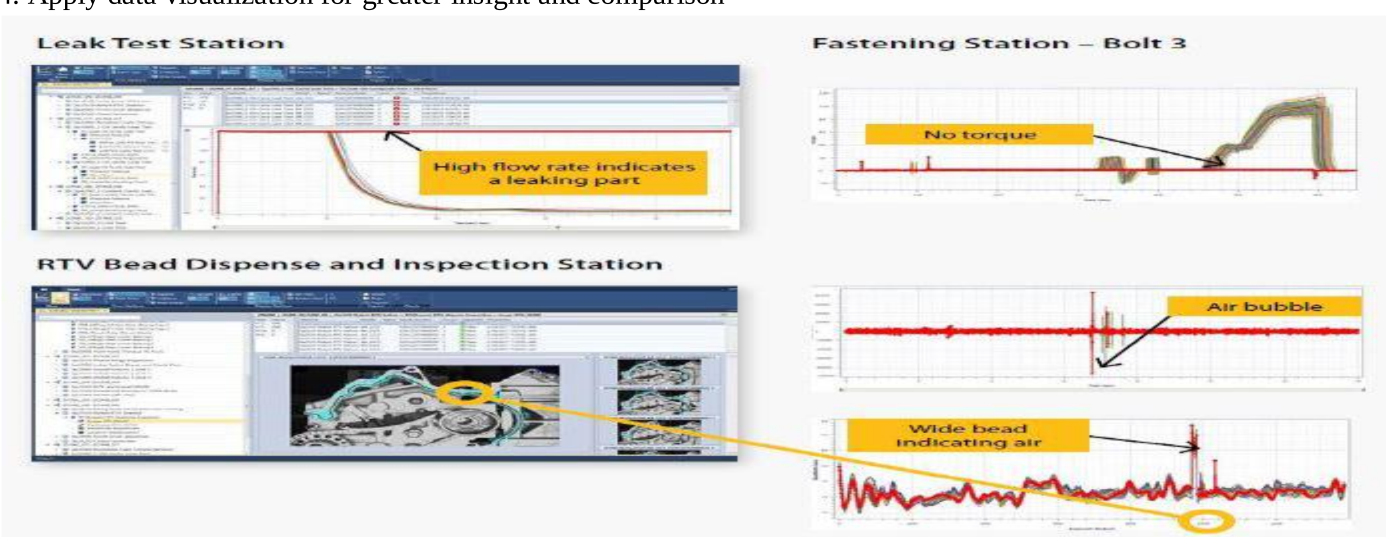
Monitoring performance of machines is important to ensure operational efficiency. However, focusing solely on this leaves a gap when trying to fix issues such as improving cycle time, tracing root cause of defects to avoid them in the future, etc.

Every cycle of every process and test station on the line generates data related to the part being produced or tested, including scalar data, digital process signatures and images. The more data you collect, the more insight you have into what has happened to a part at every assembly step.

Anything that happens upstream can have a bearing on a problem that arises further downstream. Take an engine assembly that fails a leak test, for example. The next step is that you'll need to identify what the cause of the failure is – a faulty gasket due to improper dispensing, an improperly installed gasket due to incorrect position, bolts that didn't tighten down correctly, poorly machined surfaces due to excessive vibration at a machining center? To get answers, you'll need to dig into your data.

Getting to the root cause of this flaw could require investigation of a dozen or more machining, dispensing, fitting and rundown operations, each with its own dataset. By having all the datasets relevant to a part collected and organized into a single birth history record, and indexed by part serial number, tracing root cause will be a much easier and effective process.

4. Apply data visualization for greater insight and comparison



In the above example, there is an issue caught at leak test. The signature data from the upstream processes – the RTV sealant dispense, dispense inspection and bolt fastening stations that went into creating the seal – provide visual information that can be interpreted to identify root cause. See how the bead dispense pressure, bead width and location and fastener torque angle waveforms clearly characterize the process and enable engineers and operators to quickly diagnose the root cause of the leak.

Using a tool that allows you to visualize your data from assembly stations makes the analysis of complex information easier. This in-depth analysis capability means that a part failure can easily be distinguished from a test malfunction. Detailed, signature-specific mathematical modeling can be used to find anomalies that require further investigation, pinpoint where problems occur during a process, and even optimize a test station by understanding how to shorten the test.

Automotives in Data Analytics

Though often overlooked, cars serve as a rich data source. Millions of transportation vehicles whizz past us on a regular basis, each of which generate swaths of useful information that automakers are now figuring out how to monetize. Some of the biggest passenger car automakers have more than 10 million vehicles' worth of data sitting in their data repositories. Failure to tap into these vast data stores amounts to lose value-added for customers, lost safety opportunities and lost revenue and business intelligence.

Creating business value from connected vehicle data is a non-trivial technical challenge, as any successful machine learning approach needs to contend with leveraging asynchronous and highly heterogeneous time-series data at scale in order to generate predictive insights.

As an example of the technical challenge described above, accurately estimating which vehicle on the road will experience a failure in the next month is in itself a highly unbalanced problem (Few failures are observed out of over 10 million and counting potential vehicle candidates), and requires processing the full historical set of data on all vehicles in the fleet, including hundreds of individual sensor streams, historical failure information (which includes unstructured text information from dealers and Service centers), as well as integrating a multitude of build and usage pattern information.

Indeed, intelligence gathering from automotive data is no walk in the park. However, we can leverage artificial intelligence technology in order to unlock automotive insights by way of machine learning approaches that have only recently been developed in academia. To paint the picture of how difficult this would be for Original Equipment Manufacturers (OEMs) to execute themselves, without partners from data science industry, they would have to build their own custom AI software for data extraction. This means that OEMs would have to go out of their way to find, recruit and hire world-class machine learning engineers—a challenge for even Silicon Valley startups. It is no wonder, then, that OEMs and players in the automotive industry are looking for smart solutions to the vehicle data analytics problem.

How Analytics is making the Automotive Companies from Good to Great

Several industries today are steering speedily towards the world of digitization; the Automotive Industry is a prominent field surging higher than ever. This industry has been making the right use of big data in order to enhance its current offerings and improve customer experience.

Imagine driving your car while all your calendar notes are being read out for you, your daily schedule is being created automatically, your car stopping on its own at your favorite breakfast spot and the map showing you the quickest way just when you need it. Sounds super dreamy, but as you wake up and smell the coffee you'll realize that this is also the new reality. Thanks to data analytics and the new age of technology, your daily journey has now been upgraded.
Being connected to the world with the help of analytics also reduces distraction, which in turn ensures customer safety. As more and more people invest in fully automated cars, the competitive automotive industry invests in big data analytics, cognitive IoT and predictive analysis to drive ahead of the competition.

The rise in the need for data analytics and new technology is clearly evident through the progress of the automotive industry. It is said that the use of artificial intelligence (AI) in the automotive industry is predicted to cross USD 12 billion by 2026. Soon it will begin to build its roots in every sector and modernize all the current products and services in a consumer friendly way.

Here are some of the ways of how analytics is making the automotive companies from good to great.

1. Enhancing automobile design and production

The role of data analytics in the Automobile Industry isn't just limited to self driving cars and Alexa connection. It's way more than that. Data Science, Machine Learning, and the use of AI can help automotive companies in research, development of new improved designs, enhance customer experience, cut down on costs and produce innovative models. Where manufacturing of automobiles is concerned, AI-based systems help the auto designers to create and monitor schedules more efficiently, provide advanced safety testing measures, and identify defects in the production process before going any further. With the access to predictive maintenance, manufacturing vehicles can be more efficient and inexpensive.

2. Improving customer experience with added safety and convenience

Data analytics not only help improve your travel experience by showing you the fastest routes, reminding you of calendar notes, and playing favorite songs whenever you want it to but they also help ensure your safety. Certain automobile tools recently designed by the automotive industry also indicate the driver about their speeding limits and sharp turns in the way.

3. Taking marketing and sales strategies to the next level

AI used in conned cars also helps offer localized information containing details about gas stations visited on the way to grocery stores. Later, the automotive marketers make use of this localized data to create custom deals based on customer choices. Predictive and prescriptive analytics also aid in providing them with useful data, which can then be turned into the most interesting offers for a delightful customer experience. An approach like this is helping the industry increase customer engagement and acceptance.

4. Helps companies in vehicle fleet management

Many companies these days are taking a data-driven approach to manage their fleet as it is one of the most effective ways to gain the exposure and visibility you require. Data analytics help highlight potential issues right before they pose to be huge threats in the future. It can also help you discover new opportunities. The most important of all is that it is fully customizable so you can track and note the metrics that matter the most to your business. Fuel is also an important and expensive aspect in fleet management. With the help of data analytics, you can track the fuel used by your employees, the cost, miles as well as the routes.

Technology in Data analytics

Data analytics (DA) is the process of examining data sets in order to find trends and draw conclusions about the information they contain. Increasingly, data analytics is done with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions. Scientists and researchers also use analytics tools to verify or disprove scientific models, theories and hypotheses.

As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics. In that sense, it's similar in nature to business analytics, another umbrella term for approaches to analyzing data. The difference is that the latter is oriented to business uses, while data analytics has a broader focus. The expansive view of the term isn't universal, though: In some cases, people use data analytics specifically to mean advanced analytics, treating BI as a separate category.

Data analytics initiatives can help businesses increase revenue, improve operational efficiency, optimize marketing campaigns and bolster customer service efforts. Analytics also enable organizations to respond quickly to emerging market trends and gain a competitive edge over business rivals. The ultimate goal of data analytics, however, is boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics. In addition, it can come from a mix of internal systems and external data sources.

Types of data analytics applications

At a high level, data analytics methodologies include exploratory data analysis (EDA) and confirmatory data analysis (CDA). EDA aims to find patterns and relationships in data, while CDA applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book *Exploratory Data Analysis*.

Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves the analysis of numerical data with quantifiable variables. These variables can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, as well as common phrases, themes and points of view.

At the application level, BI and reporting provide business executives and corporate workers with actionable information about key performance indicators, business operations, customers and more. In the past, data queries and reports typically were created for end users by BI developers who worked in IT. Now, more organizations use self-service BI tools that let executives, business analysts and operational workers run their own ad hoc queries and build reports themselves.

Advanced types of data analytics include data mining, which involves sorting through large data sets to identify trends, patterns and relationships. Another is predictive analytics, which seeks to predict customer behavior, equipment failures and other future business scenarios and events. Machine learning can also be used for data analytics, by running automated algorithms to churn through data sets more quickly than data scientists can do via conventional analytical modeling. Big data analytics applies data mining, predictive analytics and machine learning tools to data sets that can include a mix of structured, unstructured and semistructured data. Text mining provides a means of analyzing documents, emails and other text-based content.

Data analytics initiatives support a wide variety of business uses. For example, banks and credit card companies analyze withdrawal and spending patterns to prevent fraud and identity theft. E-commerce companies and marketing services providers use clickstream analysis to identify website visitors who are likely to buy a particular product or service -- based on navigation and page-viewing patterns. Healthcare organizations mine patient data to evaluate the effectiveness of treatments for cancer and other diseases.

Mobile network operators examine customer data to forecast churn; that enables them to take steps to prevent customers from defecting to rival vendors. To boost customer relationship management efforts, companies engage in CRM analytics to segment customers for marketing campaigns and equip call center workers with up-to-date information about callers.

Inside the data analytics process

Data analytics applications involve more than just analyzing data, particularly on advanced analytics projects. Much of the required work takes place upfront, in collecting, integrating and preparing data and then developing, testing and revising analytical models to ensure that they produce accurate results. In addition to data scientists and other data analysts, analytics teams often include data engineers, who create data pipelines and help prepare data sets for analysis.

The analytics process starts with data collection. Data scientists identify the information they need for a particular analytics application, and then work on their own or with data engineers and the IT staff to assemble it for use. Data from different source systems may need to be combined via data integration routines, transformed into a common format and loaded into an analytics system, such as a Hadoop cluster, NoSQL database or data warehouse.

In other cases, the collection process may consist of pulling a relevant subset out of a stream of data that flows into, for example, Hadoop. The data is then moved to a separate partition in the system so it can be analyzed without affecting the overall data set.

Once the data that's needed is in place, the next step is to find and fix data quality problems that could affect the accuracy of analytics applications. That includes running data profiling and data cleansing tasks to ensure the information in a data set is consistent and that errors and duplicate entries are eliminated. Additional data preparation work is done to manipulate and organize the data for the planned analytics use. Data governance policies are then applied to ensure that the data follows corporate standards and is being used properly.

From here, a data scientist builds an analytical model, using predictive modeling tools or other analytics software and programming languages such as Python, Scala, R and SQL. Typically, the model is initially run against a partial data set to test its accuracy; it's then revised and tested again as needed. This process is known as "training" the model until it functions as intended. Finally, the model is run in production mode against the full data set, something that can be done once to address a specific information need or on an ongoing basis as the data is updated.

In some cases, analytics applications can be set to automatically trigger business actions. An example is stock trades by a financial services firm. Otherwise, the last step in the data analytics process is communicating the results generated by analytical models to business executives and other end users. Charts and other infographics can be designed to make findings easier to understand. Data visualizations often are incorporated into BI dashboard applications that display data on a single screen and can be updated in real time as new information becomes available.

Data analytics vs. data science

As automation grows, data scientists will focus more on business needs, strategic oversight and deep learning. Data analysts who work in business intelligence will focus more on model creation and other routine tasks. In general, data scientists concentrate efforts on producing broad insights, while data analysts focus on answering specific questions. In terms of technical skills, future data scientists will need to focus more on the machine learning operations process, also called MLOps.

10 Most Common Technology Problems for Small Businesses

1. Data Sprawl and Integration Challenges

Software is an incredible thing. With the explosion of SaaS software, it’s relatively inexpensive for businesses to get the right tools to make their businesses better – and there are SO many. The problem is that there isn’t an all-in-one solution for most companies. As a result, work ends up being done in multiple systems, and if the tools aren’t integrated, you have a hard time finding anything anymore. We call this data sprawl.

Is it in SharePoint? Is it in OneDrive? Is it in the CRM? Is it in the project management tool?

Also, as new technologies are released, sometimes they don’t play nice with older systems or applications. For example, upgrading an operating system could cause applications or peripherals to stop working because they aren’t compatible.

Both data and integration challenges can result in the double-entry of data, having to look in multiple places for information, and other inefficiencies. Thorough planning before implementation can minimize these issues.

2. Backup and Disaster Recovery Challenges

If you’ve ever accidentally deleted a file, experienced a power outage, had a server crash, or had any other “data disaster,” then you know how important having a reliable backup and disaster recovery plan can be (or would have been!)

Having the right backup solution and disaster recovery procedures in place for when these events inevitably occur is critical to eliminating the downtime and expenses that accompany them.

So Much Data…What Needs to Be Backed Up?

We have the ability to create, collect, and store SO much data these days. The challenge is knowing what needs to be kept, for how long, and how quickly you might need it back if something bad happens.

Keep in mind that data backups have a cost. The amount of data you back up and the speed at which you'd need to get it back are two factors that impact the cost.

Data Backup and Remote Employees

Does your backup strategy include a way to backup the files your remote employees use?

Everything was a lot simpler when everyone worked under one roof – or a series of offices. But, with the recent prevalence of remote work, companies now have tens or hundreds of "home offices."

Ensuring all of the data they're working on makes it to a shared place where the IT department can see it and back it up is critical to maintaining collaboration and business continuity should a data disaster happen.

3. Root Causes Aren't Uncovered or Addressed

When IT departments or providers have a process for tracking issues, they're able to analyze the data and identify trends. When they dig into those trends, sometimes they find that an issue is a symptom of a more significant problem. Solving the root cause would make the problem go away for good and keep employees productive instead of on the phone with IT support.

4. Cyber Security Risks

Hackers are becoming more sophisticated, and small businesses are being targeted now more than ever before. Think of all of your trade secrets, confidential communication, customer information, and HR records that are stored on your company's devices.

What would be the impact if this data fell into the wrong hands? There is a cost-effective suite of IT security procedures and solutions worth considering. These pay dividends in the long run – an incident isn't worth the headache.

Data Security and Remote Employees

With a large portion of today's workforce working remotely or in a hybrid situation, a whole new set of business challenges have cropped up. One of those challenges is keeping your employees and their data safe from attackers while working remotely.

Providing workers with company-owned computers, using a VPN for remote access, and updating your data backup policy will help ensure that your remote worker's data is just as secure as your on-site employees.

5. Lack of Strategic IT Planning

Often, IT and business leaders are disconnected. The leaders see the bigger picture, but sometimes IT isn't informed to make sure that business goals can be accomplished without technology getting in the way.

How would business operations be different if the business and technology plans were aligned and the organization could scale without constraint?

For example, if you plan to grow significantly in the next three years and IT is in the loop, the infrastructure implemented for current challenges may be different so it can accommodate the long-term growth.

6. Frustrated Users (Employees)

Interacting with technology is a huge part of your employees' day. Using slow, outdated systems with frequent problems makes it much more difficult for them to be happy and productive, ultimately decreasing their output. How would it impact your business if you enabled your employees to get just 5 percent more accomplished every day simply by keeping your technology up to date?

Slow IT Response Times

When your IT team or managed IT services provider isn’t able to resolve issues promptly due to a lack of tools, expertise, or time, employees become frustrated, and productivity declines.

With the right IT support team (whether it’s outsourced, in-house, or a combination of both), users should be able to contact support and have their issues worked on swiftly. Some issues are more complex than others and may take a long time to be resolved if a vendor or manufacturer is involved. But you should expect that issues are being worked on as quickly as possible given each situation.
Evaluate your IT processes and procedures to ensure that issues are addressed and remedied in a timely manner.

Extra Step of Security
Security measures are a necessary evil given today’s cyber security landscape. They’re designed to help keep the wrong people out of the system, which means more time is spent verifying that you’re one of the right people.

Multi-factor authentication (MFA) is one of those steps that often frustrates users. It’s recommended for logging into accounts and accessing data as an essential layer in keeping business data safe.

Users are prompted to enter their password and a one-time code to verify that they are who they say they are. This second step of verification might seem like more of an annoyance to employees than a benefit, but implementing this procedure helps defend against an attacker accessing your data through an employee’s account.

Remote Work Experience
Working remotely is different than being in an office environment – for better or worse. In either case, employees need the right tools and access to systems and data regardless of where they’re working from.

Managing a team of remote workers is also a new and different experience for some. When everyone is in the office together, it’s easy to stop by people’s desks to connect, collaborate and mentor. Remote work doesn’t allow for the same kind of natural connections – it has to be intentional. With tools like Microsoft Teams, managers and teams can work together seamlessly. This helps employees get the support they need and stay connected.

With the right planning and technology, remote workers feel empowered and enabled to do their best work. As a result, their productivity will improve, and they will experience a better quality of work-life.

7. Employee Onboarding / Offboarding and Asset Management
Employee Onboarding
A new employee’s first impression of your company is typically based on your onboarding process, and you want it to go as smoothly as possible. This is especially true for remote workers who might never set foot in your physical office.
It can get rocky without a thorough new hire onboarding process. This ensures every new employee receives the right equipment, access, and training to hit the ground running and set them on the path towards success on their very first day.
Employee Offboarding
Offboarding employees is a security risk. After all, the departing employee had access to your systems and data. Therefore, no matter the reason for the separation (amicable or otherwise), ensuring any company-owned hardware is returned and their access is terminated successfully is vital to safeguarding your network and sensitive information.
A thoroughly documented process is critical to making sure nothing is missed.

Asset and Hardware Inventory Management
Not keeping track of your assets can be costly. For example, without a constantly updated inventory of both hardware and software (licensing), you could be paying for too many licenses, buying hardware you don’t need because it’s already available, or worse – a computer “walks away.”

Another issue that has become more common as businesses embrace remote work is hardware storage and inventory management. Without a physical office, where do you keep old hardware or new hardware that's waiting to be deployed when the time is right? And if it's stored offsite, how do you know what you have and if it's secure?
Many managed IT services providers now offer their clients secure storage and inventory management options. It makes sense since, in many cases, they are the ones preparing and deploying hardware as needed for your company.

8. Hardware Delays and Supply Chain Issues
The COVID-19 pandemic and its repercussions are being felt everywhere – supply chain issues are one of them.
Major shortages of everything from raw materials to computer chips is impacting our ability to get our hands on all kinds of goods.
Due to this shortage, companies are having to plan several months in advance, make purchase decisions quickly, and wait longer for the hardware to come in. It puts everyone in a frustrating spot.
For example, the timeline for getting a business-class computer currently is at least 90 days. That puts businesses in a tight spot if one machine breaks and they need a replacement, or if they hire new staff members and need new computers for them.
Also, since inventory is so limited, hardware quotes are being provided by manufacturers, and if the order isn't placed within 12-24 hours, the hardware is no longer available at that price – or no longer available at all.

9. Compliance Requirements and Cyber Insurance Applications
Accountability for cyber security is increasing as cyber security threats become more prevalent.
Many industries have regulatory compliance requirements such as HIPAA, ITAR, NIST, and CMMC. Navigating the technical and non-technical nuances of these requirements can be challenging.
Cyber security insurance has also grown in popularity given the losses companies have experienced related to cyber security incidents.
When an attack occurs, expenses and fines can mount up quickly. Cyber liability insurance is designed to help your business bounce back financially after a cyber attack. A cyber liability insurance policy covers the costs of ransom payments, data recovery, hardware replacement, regulatory fines, legal fees, and other expenses.
The difficulty arises when applying for the policy. The applications can be confusing and inconsistent.
You'll need to provide technical and non-technical information that addresses everything from your company's size to the exact details of your security and processes. This is because coverage and premiums are based on your security stature.
Getting the information needed for the more technical aspects of the application tends to be a roadblock for executives due to the complexity of cyber security. However, working closely with your IT manager or IT services provider to answer those questions can help you get the best coverage at the lowest prices and even reveal the holes in your security.

10. Facilitating Training
Every employee has a different comfort level with technology. And as technology has become more prevalent in daily business functions, training is even more important to keep employees happy, productive, and confident in their roles.
Training is also vital to the security of your business network and sensitive data.
You can quickly and easily find tutorials for all types of software applications, hardware, and even cyber security awareness training. However, accessing and facilitating those resources for your employees can be challenging, considering the abundance of readily available online training.
Your IT support team can help you determine the appropriate resources necessary to train your employees properly.

Unit IV
Manage your work to meet requirements (NOS 9001):
Understanding Learning objectives, Introduction to work & meeting requirements, Time Management, Work management & prioritization, Quality & Standards Adherence,

Understanding Learning Objectives

Definition: learning means acquisition of knowledge, skills, ability, expertise etc through study, experience and practice that result into long term change in one's behavior.
It is a process of acquiring knowledge through conscious and systematic efforts that affect one's subsequent behavior of a long time.
Learning is a key process in human behaviour. It refers to a spectrum of changes that take place as a result of one's experience. Learning may be defined as "any relatively permanent change in behaviour or behavioural potential produced by experience". One must remember that some behavioural changes occur due to the use of drugs, or fatigue. Such changes are temporary. They are not considered learning. Changes due to practice and experience, which are relatively permanent, are illustrative of learning.

Introduction:
A person's reaction pattern depends on the overall individual behavior and psychographic profile. The thinking and acting patterns are governed by cognitive construct such as perception, learning, attitudes, motivation and personality.
Understanding individual behavior is key to organizational behavior. Psychology and social psychology have made significant contribution towards understanding of individual behavior. These factors are referred as Dynamic of individual behavior.

Dynamic of individual behavior

The key dynamics that affect the individual behavior are:
1. Perception: it is important in understanding one's behavior and has direct application in organizational behavior and management.
2. Learning: learning is a dominant psychological process that determines one's behavior.
3. Attitudes and values: Attitude is cognitive aspect that always remains inside the person. We cannot see attitude but can see behavior that is triggered by it. Individual behavior is the outcome of learning experience. Learning experience or preferences are called as attitudes. Values are very closely related to attitudes. They are the basic conviction that society holds about all activities of life.
4. Personality: personality is one of the prime factors determining individual response to any stimulus. One's pattern of working depends on their personality.
5. Motivation: motivation is related to or based on motive. It is a process that stimulates people to put effort in specific direction for accomplishment of goals.
6. Positive organization behavior (POB): behavior is all about what people think. POB is the outcome of positive response of all participants in the organization.

Understanding Individual Behavior

There are millions of stars in the sky, but every star is different. Similarly, there are millions of people in the universe, but each person is different from the multiple perspectives. Organizations are composed of individuals. Each individual is different from each other on the basis of different psychological factors such as motives, aspirations, perceptions and abilities.

Individual behavior means some concrete actions by a person. For instance, how a teacher behaves in the class reflects his behavior. Organizational behavior has a great challenge as well as responsibility to deal with the difference in individual behavior in the context of organizations.

Psychology, a discipline contributing to organizational behavior, helps to examine how people are similar, and how they differ in their thinking, feeling, and behaviors. The behavior of an individual is influenced by various factors. Some of the factors lie within himself e.g. his instincts, personality traits, internal feelings etc. while some lie outside him comprising the external environment of which he is a part, e.g. weather conditions, events conveying some information, and other people's behavior that directly influences his behavior.

<u>Factors Influencing Individual Behavior</u>

Many characteristics of individual behavior (or human beings) are inherited and many others are formed because of various environmental factors. Individual behavior is influenced by personal factor, psychological factor, organizational factors and environmental factors.

1. Personal Factors

Every individual brings to the work place a variety of personal characteristics and attributes like age, sex, education, intelligence, marital status, religion and number of dependents etc.

i) Age

Age is an important factor that organizations are concern about. Age has an impact on the performance, turnover, absenteeism, productivity and satisfaction levels of employees. In the other simple words, performance depends on age. As age advances, performance is likely to decline. Similarly, aging has impact on turnover. The older people are less likely to quit the job. Age-absence relationship depends on whether absenteeism is avoidable or unavoidable. Generally, older employees have lower rate of avoidable absence than the younger employees. However, they have high rate of unavoidable absence. This probably because of poor health associated with old age. With regard to productivity, older age results in reduced productivity. This is because of the decline in individual's skills as he grows older in age. This is a positive association between age and satisfaction.

ii) Gender (Sex)

There are important differences between men and women which is because of their inherent qualities and family responsibilities that affect their job performance. Man is expected to be tough while a woman is expected to be gentle and highly emotional, are some of the stereotyped assumption that have no basis in genetic influences. These behaviors are developed due to differences in treatment that the boys and girls receive in the family environment.

Psychological studies have found that women are more willing to conform to authority and less aggressive. On the other hand, men are more aggressive, have more expectations, and are more ambitious than women. While the stability rate is higher among women employees, the absenteeism rate is also higher; this can be attributed to the fact that they are expected to shoulder more family responsibilities than men. Research studies on male and female employees say that there are no consistent male-female difference in problem-solving ability, analytical skills and competitive drive.

Even though some work roles are assumed to be the exclusive domain of modified to accommodate man in these positions. As far as the administration of the management process is concerned, women, in general do not differ from men in their operative behaviors.

iii. Emotional Intelligence

Highly intelligent employees learn job-related skills and other organizational practices quickly, and organizations have to spend less time for training them. People with high intelligence have a good decision-making ability because of their good analytical and reasoning skills. They are very productive and have a high achievement drive. This is the age of emotional intelligence, and today's recruiting managers make sure that candidates have it. The five components of emotional intelligence at work are:

- ***Self-awareness***: Self-confidence, realistic self-assessment and self-deprecating sense of humor.
- ***Self-regulation***: Trust worthiness and integrity, comfort with ambiguity and openness to change.
- ***Motivation***: A strong drive to achieve, even in the time of failure, and organizational commitment.
- ***Empathy:*** Expertise in building and retaining talent, cross-cultural sensitivity and service to clients and customers.
- ***Social Skills:*** Effectiveness in leading change, persuasiveness and expertise in building and leading teams.

iv) Marital Status

Marital status has influence on absenteeism, turn over and satisfaction. Married employees have fewer absences, undergo less turn over, and are more satisfied with their jobs than the unmarried ones. In fact, married employees are more conscious of their responsibilities. Satisfaction levels were found to be in equal ratio among both married and unmarried employees. Employees belonging to both categories were unsatisfied in terms of the salary they got from organizations. Marriage imposes additional responsibility hence the need for steady job and steady income.

v) Education

Education has its effect upon individual behavior largely through the level and type of education received. Increased levels of education serve to increase an individual's expectations about positive outcomes. These outcomes are generally perceived to be a more satisfying job, higher income level and greater alternative sources for occupational choice i.e. the good life.

vi) Religion

Religion and culture also determine attitude towards work and towards financial incentives. In other words, religion and religiously based cultures that play an important role in determining some aspects of individual behavior, especially those that concern morals, ethics and a code of conduct. Highly religious people have high moral standards and usually do not tell lies or talk ill of others. They are highly contented thus strive for the goal of achievement and self-fulfillment.

vii) Abilities

Ability refers to the actual skills and capabilities that a person possesses and is required for the effective performance of activities. Ability of an individual is made up of two sets of skills i.e. intellectual ability and physical ability. Railways need to ensure that its employees possess the necessary abilities to engage in the behavior required for effective performance. This is accomplished either by careful selection of people or by a combination of selection and training.

2. Psychological Factors

Psychological factors may include perception, values, attitudes, psychological needs, personality and motivation. All factors can be explained as follows:

i) Perception

Perception is an outcome of an object. It is the view point which one interprets a situation. For instance, a railway booking clerk facing a well-dressed person perceives him to be of high status and talks to him nicely, whereas he tends to ignore an ill-dressed person, or make him wait, though both the passengers want first class ticket. Perception is a major factor that contributes to individual behavior and differences, both in personal and organizational life.

ii) Values

Values are the foundation for behavioral difference among individuals, and they help in understanding the attributes and motivation. Values represent the basic conviction that a specific made of conduct is personally or socially preferable to an opposite mode of conduct. They have both context and intensity attributes. The context attribute says that a mode of

conduct is important and the intensity attribute gives importance to specific. Values are judgmental by nature, and every individual has a different conception of what is right, good or desirable and what is not right, not good or undesirable.
In the today's modern and competitive business environment, managers have to be capable of working with people from different cultures. Because values and attributes differ across cultures, and understanding of these differences will be helpful in explaining and predicting the behavior of employees form different cultures.

iii) Attitudes

The term 'Attitude' is very common and used in everyday life to describe people's behavior. It is a way of organizing a perception. In other words, it is more or less a stable tendency to feel, think, perceive and act in a certain manner towards an object or a situation. It is a tendency to act in a certain way, either favorably or unfavorably concerning objects, people or events. Attitude has three elements in it that lead to measurable outcomes. These are feelings, thoughts and behavior. Feelings and thoughts can be measured by simply asking individuals about their feelings and opinions. Behavior can be measured either by actual one's actions or simply by asking the person how he would act in a certain situation. By measuring and integrating these three elements, a person's attitude towards a given situation can be established.

In general, a person may have a positive attitude which means a good outlook of life, or negative attitude which means continuous complaining about problems in life. For example, if you showed a glass half-filled with water to A and B and asked them to describe it.

Features of Learning

The process of learning has certain distinctive characteristics. The first feature is that learning *always involves some kinds of experience.* We experience an event occurring in a certain sequence on a number of occasions. If an event happens then it may be followed by certain other events. For example, one learns that if the bell rings in the hostel after sunset, then dinner is ready to be served. Repeated experience of satisfaction after doing something in a specified manner leads to the formation of habit. Sometimes a single experience can lead to learning. A child strikes a matchstick on the side of a matchbox, and gets her/his fingers burnt. Such an experience makes the child learn to be careful in handling the matchbox in future.

Behavioural changes that occur due to learning are relatively permanent. They must be distinguished from the behavioural changes that are neither permanent nor learned.

For example, changes in behaviour often occur due to the effects of fatigue, habituation, and drugs. Suppose you are reading your textbook of psychology for sometime or you are trying to learn how to drive a motor car, a time comes when you will feel tired. You stop reading or driving. This is a behavioural change due to fatigue, and is temporary. It is not considered learning.
Let us take another case of change in one's behaviour. Suppose in the vicinity of your residence, a marriage is being performed. It generates a lot of noise, which continues till late night. In the beginning, the noise distracts you from whatever you are doing. You feel disturbed. While the noise continues, you make some orienting reflexes. These reflexes become weaker and weaker, and eventually become undetectable. This is also one kind of behavioural change. This change is due to continuous exposure to stimuli. It is called habituation. It is not due to learning.

You must have noticed that people who are on sedatives or drugs or alcohol, their behaviour changes as it affects physiological functions. Such changes are temporary in nature and disappear, as the effect wears out. Learning involves a sequence of psychological events. This will become clear if we were to describe a typical learning experiment.

Suppose psychologists are interested in understanding how a list of words is learned. They will go through the following sequence :
(i) do a pre-test to know how much the person knows before learning,
(ii) present the list of words to be remembered for a fixed time,

(iii) during this time the list of words is processed towards acquiring new knowledge,
(iv) after processing is complete, new knowledge is acquired (this is LEARNING), and
(v) after some time elapses, the processed information is recalled by the person.

By comparing the number of words which a person now knows as compared to what s/he knew in the pre-test, one infers that learning did take place. Thus, learning is an inferred process and is different from performance. Performance is a person's observed behaviour or response or action.

Let us understand what is meant by the term inference. Suppose you are asked by your teacher to memorise a poem. You read that poem a number of times. Then you say that you have learned the poem. You are asked to recite the poem and you are able to recite it. The recitation of the poem by you is your performance. On the basis of your performance, the teacher infers that you have learned the poem.

PARADIGM OF LEARNING

Learning takes place in many ways. There are some methods that are used in acquisition of simple responses while other methods are used in the acquisition of complex responses.

The simplest kind of learning is called conditioning. Two types of conditioning have been identified. The first one is called classical conditioning, and the second instrumental/operant conditioning. In addition, we have observational learning, cognitive learning, verbal learning, concept learning, and skill learning.

CLASSICAL CONDITIONING

Examples of classical conditioning abound in everyday life. Imagine you have just finished your lunch and you are feeling satisfied. Then you see some sweet dish served on the adjoining table. This signals its taste in your mouth, and triggers the secretion of saliva. You feel like eating it. This is a conditioned response (CR).

Let us take another example. In the early stages of childhood, one is naturally afraid of any loud noise. Suppose a small child catches an inflated balloon which bursts in her/his hands making a loud noise. The child becomes afraid. Now the next time s/he is made to hold a balloon, it becomes a signal or cue for noise and elicits fear response. This happens because of contiguous presentation of balloon as a conditioned stimulus (CS) and loud noise as an unconditioned stimulus (US).

OPERANT/INSTRUMENTAL CONDITIONING

This type of conditioning was first investigated by B.F. Skinner. Skinner studied occurrence of voluntary responses when an organism operates on the environment. He called them operants. *Operants are those behaviours or responses, which are emitted by animals and human beings voluntarily and are under their control.* The term operant is used because the organism operates on the environment. Conditioning of operant behaviour is called operant conditioning.

Examples of instrumental conditioning abound in our everyday life. Children who want to have some sweets in the absence of their mother learn to locate the jar in which mother hides the sweets for safekeeping and eat it. Children learn to be polite and say 'please' to get favours from their parents and others. One learns to operate mechanical gadgets such as radio, camera, T.V., etc. based on the principle of instrumental conditioning. As a matter of fact human beings learn short cuts to attain desired goals or ends through instrumental conditioning.

OBSERVATIONAL LEARNING

The next form of learning takes place by observing others. Earlier this form of learning was called imitation. Bandura and his colleagues in a series of experimental studies investigated observational learning in detail.

In this kind of learning, human beings learn social behaviours, therefore, it is sometimes called social learning. In many situations individuals do not know how to behave. They observe others and emulate their behaviour. This form of learning is called modeling.

Examples of observational learning abound in our social life. Fashion designers employ tall, pretty, and gracious young girls and tall, smart, and well-built young boys for popularizing clothes of different designs and fabrics. People observe them on televised fashion shows and advertisements in magazines and newspapers. They imitate these models. Observing superiors and likeable persons and then emulating their behaviour in a novel social situation is a common experience.

Cognitive learning

The word cognitive is derived from "cognition", which describes your ability to think, perceive, remember and solve problems. Cognitive learning is a learning process that results from the effective use and application of the brain. It focuses on helping you learn ways to maximize your brain's potential. When you master cognitive learning, you inculcate a habit of lifelong continuous learning.

Objectives or salient features of learning:

1. Psychological process: learning is an important psychological process that determines individual behavior.
2. Outcome of past experience: it is the result of past experience or practice.
3. Change in behavior: learning brings changes in the existing pattern of behavior. It leads to improvement or deterioration in behavior.
4. Conscious and deliberate: it is a conscious and deliberate process.
5. Need of reinforcement: it needs reinforcement i.e, any aspect that can strengthen learned things.
6. Permanent change: behavioral change resulting from learning is relatively permanent.
7. Experience plus practice: behavioral change must be based on experience plus practice.
8. Input in other cognitive processes: learning is an input to perception, personality, motivation and attitudes.

Organizational Behavior (OB)

Organizational behavior is the academic study of how people interact within groups. The principles of the study of organizational behavior are applied primarily in attempts to make businesses operate more effectively.

Understanding Organizational Behavior (OB)

The study of organizational behavior includes areas of research dedicated to improving job performance, increasing job satisfaction, promoting innovation, and encouraging leadership. Each has its own recommended actions, such as reorganizing groups, modifying compensation structures, or changing methods of performance evaluation.

Organizational Behavior (OB) can be defined as the understanding, prediction and management of human behavior both individually or in a group that occur within an organization.

Internal and external perspectives are the two theories of how organizational behavior can be viewed from an organization's point of view.

Importance of OB

While working in an organization, it is very important to understand others behavior as well as make others understand ours. In order to maintain a healthy working environment, we need to adapt to the environment and understand the goals we need to achieve. This can be done easily if we understand the importance of OB.

Following points bring out the importance of OB −

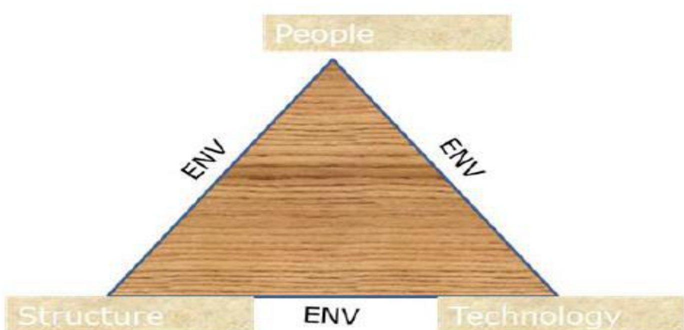
- It helps in explaining the interpersonal relationships employees share with each other as well as with their higher and lower subordinates.
- The prediction of individual behavior can be explained.
- It balances the cordial relationship in an enterprise by maintaining effective communication.
- It assists in marketing.
- It helps managers to encourage their sub-ordinates.
- Any change within the organization can be made easier.
- It helps in predicting human behavior & their application to achieve organizational goals.
- It helps in making the organization more effective.

Thus studying organizational behavior helps in recognizing the patterns of human behavior and in turn throws light on how these patterns profoundly influence the performance of an organization.

Organizational Behavior - Determinants

There are three major factors that affect OB. The working environment being the base for all three factors, they are also known as the determinants of OB. The three determinants are −

- People
- Structure
- Technology



People

An organization consists of people with different traits, personality, skills, qualities, interests, background, beliefs, values and intelligence. In order to maintain a healthy environment, all the employees should be treated equally and be judged according to their work and other aspects that affects the firm.

Example − A company offers campus placement to trainees from different states like Orissa, Haryana, Arunachal Pradesh and many more. However, during and after training, all trainees are examined only on the basis of their performance in the tasks assigned.

Organizational Structure

Structure is the layout design of an organization. It is the construction and arrangement of relationships, strategies according to the organizational goal.

Example − Organizational structure defines the relation of a manager with employees and co-workers.

Technology

Technology can be defined as the implementation of scientific knowledge for practical usage. It also provides the resources required by the people that affect their work and task performance in the right direction.

Example − Introduction of SAP, big data and other software in the market determines individual and organizational performance.

Environment

All companies function within a given internal and external environment. Internal environment can be defined as the conditions, factors, and elements within an enterprise that influences the activities, choices made by the firm, and especially the behavior of the employees. While external environment can be defined as outside factors that affect the company's ability to operate. Some of them can be manipulated by the company's marketing, while others require the company to make adjustments.

Some **examples of internal environment** include employee morale, culture changes, financial changes or issues, and some **examples of external environment** include political factors, changes to the economy and the company itself.

Organizational Behavior - Concepts

The concept of OB is based on two key elements namely −

- Nature of people
- Nature of the organization

Nature of People

In simple words, nature of people is the basic qualities of a person, or the character that personifies an individual they can be similar or unique. Talking at the organizational level, some major factors affecting the nature of people have been highlighted. They are −

- **Individual Difference** − It is the managerial approach towards each employee individually, that is one-on-one approach and not the statistical approach, that is, avoidance of single rule. **Example**− Manager should not be biased towards any particular employee rather should treat them equally and try not to judge anyone on any other factor apart from their work.
- **Perception** − It is a unique ability to observe, listen and conclude something. It believes in our senses. In short, the way we interpret things and have our point of view is our perception. **Example** − Aman thinks late night parties spoil youth while Anamika thinks late night parties are a way of making new friends. Here we see both Aman and Anamika have different perception about the same thing.
- **A whole person** − As we all know that a person's skill or brain cannot be employed we have to employee a whole person. Skill comes from background and knowledge. Our personal life cannot be totally separated from our work life, just like emotional conditions are not separable from physical conditions. So, people function is the functioning of a total human being not a specific feature of human being.

- **Motivated behavior** − It is the behavior implanted or caused by some motivation from some person, group or even a situation. In an organization, we can see two different types of motivated employees −
    - **Positive motivation** − Encouraging others to change their behavior or say complete a task by luring them with promotions or any other profits. Example − "If you complete this, you will gain this."
    - **Negative motivation** − Forcing or warning others to change their behavior else there can be serious consequences. Example − "If you don't complete this, you will be deprived from the office."
- **Value of person** − Employees want to be valued and appreciated for their skills and abilities followed by opportunities which help them develop themselves.

Nature of Organization

Nature of organization states the motive of the firm. It is the opportunities it provides in the global market. It also defines the employees' standard; in short, it defines the character of the company by acting as a mirror reflection of the company. We can understand the nature of any firm with its social system, the mutual interest it shares and the work ethics.

Let us take a quick look at all these factors −

- **Social system** − Every organization socializes with other firms, their customers, or simply the outer world, and all of its employees - their own social roles and status. Their behavior is mainly influenced by their group as well as individual drives. Social system are of two types namely −
    - **Formal** − Groups formed by people working together in a firm or people that belong to the same club is considered as **formal social system**. **Example** − A success party after getting a project.
    - **Informal** − A group of friends, people socializing with others freely, enjoying, partying or chilling. **Example** − Birthday party.
- **Mutual interest** − Every organization needs people and people need organizations to survive and prosper. Basically it's a mutual understanding between the organization and the employees that helps both reach their respective objectives. **Example** − We deposit our money in the bank, in return the bank gives us loan, interest, etc.
- **Ethics** − They are the moral principles of an individual, group, and organization. In order to attract and keep valuable employees, ethical treatment is necessary and some moral standards need to be set. In fact, companies are now establishing code of ethics training reward for notable ethical behavior.

Organizational Behavior - Scope

In a very broad sense, the **scope of OB** is the extent to which it can govern or influence the operations of an organization. The scope of OB integrates 3 concepts respectively −

Individual Behavior

It is the study of individual's personality, learning, attitudes, motivation, and job satisfaction. In this study, we interact with others in order to study about them and make our perception about them.

Example − The personal interview round is conducted to interact with candidates to check their skills, apart from those mentioned in the resume.

Inter-individual Behavior

It is the study conducted through communication between the employees among themselves as well as their subordinates, understanding people's leadership qualities, group dynamics, group conflicts, power and politics.

Example − A meeting to decide list of new board members.

Group Behavior

Group behavior studies the formation of organization, structure of organization and effectiveness of organization. The group efforts made towards the achievement of organization's goal is group behavior. In short, it is the way how a group behaves.

Example − Strike, rally etc.

Functions of a Manager

Functions of a manager are the various roles played by the manager in an organization. A manager is accountable for all the happenings in the firm and is answerable to the management. The seven major roles played by the manager are −

- Planning
- Organizing
- Staffing
- Directing/leading
- Coordinating
- Reporting
- Budgeting
- Controlling

Roles of a Manager

Now, let us see what exactly are these roles and their importance.

- **Planning** − The basic step required for any project, big or small, is the planning stage. The manager needs to plan the schedule and give the blueprint of how the task is to be done with all the necessary details, and also the manager should have a backup plan that if this doesn't work then what next. **Example** − There is a new project, how to start, human resource required, resources required, etc., everything should be planned.
- **Organizing** − Next comes the organizing part, where the manager needs to synchronize and have to make sure everything is going according to the plan. Everything should work as per the plan, and if not then the manager needs to look into the issue and make it work as planned. **Example** − A software tester is required, so organize the venue, date and time to interview those eligible for the post.
- **Staffing** − In simple words, staffing means grouping of people into different teams and allotting different tasks to them. If the team members have some disputes then the team member needs to report to the team leader who will forward it to the manager and the issue will be taken care of. **Example** − Assembling a new team for a new project.
- **Directing/Leading** − It is a manager's responsibility to guide the employees in all situations in order to avoid conflicts and delay in the task. Manager has to lead the employees so that they can get a clear idea about what is to be done and how to do it. **Example** − a team needs a team leader to look after each task that is accomplished, in-process, or aborted.
- **Coordinating** − It means bringing all the employees together by forming an efficient relationship and making them feel comfortable to share their views and issues freely. **Example** − Coordinating the schedule for a project.
- **Reporting** − The manager has to keep updated information about all the ongoing tasks, and it is the sole responsibility of the manager to report the updated status to the higher authorities; while all the employees are bound to report to the manager. **Example** − Keeping the respective directors informed about the progress on their respective projects.
- **Budgeting** − A task has to be completed within the given time frame as well as it should be cost efficient. The manager needs to be double sure that all the amount invested in the project doesn't exceed the budget given and in case of imbalance, the budgeting manager has to report to the management. **Example** − If budget allows to place three employees then five employees cannot be assigned for the task.
- **Controlling** − Last but of course not the least role played by the manager is having everything under control. Whether it is the budget, or resource allocation, everything should be in order. **Example** − All members of a team cannot be granted leave on the same day, as it affects work delivery.

Various Challenges of a Manager

We have seen the different roles a manager as to play in order to maintain the workflow balance in an organization. With all these responsibilities, there are some tough challenges a manager has to deal with while trying to balance everything. Following are some challenges a manager has to deal with −

- **Managing workforce diversity** − Manager shouldn't create or encourage discrimination among employees. Employees from different background, culture, and ethnicity should be treated as equal and rewards should be given only on the basis of work.
- **Improving quality and productivity** − It is the sole responsibility of the manager to increase the productivity without hampering the quality. It can be done in two ways −
    - **Totally quality management** − That is constant focus on customer satisfaction by improving organizational process.
    - **Process of engineering** − Focusing on the manufacturing of the product, so that the quality is not compromised.
- **Responding to labor storage** − If there is a labor shortage then the manager should quickly respond to solve this problem by arranging for the workforce required so that the product delivery is not delayed.
- **Eradication of labor shortage** − The manager needs to take quick action, if there is a labor shortage and should assure with backup plans so that there is no labor shortage in future.
- **Improving customer service** − Manager faces the challenge to constantly improve customer service to survive in an ever-competitive environment.
- **Improving ethical behavior** − Managers should make sure that the employees behave properly and maintain the decorum of the company. These are few major challenges a manager faces while trying to complete a project. To maintain work-life balance and for the betterment of the organization, the manager should try level best to resolve these challenges.

Organizational Behavior - Models

Organizational behavior reflects the behavior of the people and management all together, it is considered as field study not just a discipline. A discipline is an accepted science that is based upon theoretical foundation, whereas OB is an inter-disciplinary approach where knowledge from different disciplines like psychology, sociology, anthropology, etc. are included. It is used to solve organizational problems, especially those related to human beings.

There are four different types of models in OB. We will throw some light on each of these four models.

Autocratic Model

The root level of this model is power with a managerial orientation of authority. The employees in this model are oriented towards obedience and discipline. They are dependent on their boss. The employee requirement that is met is subsistence (survival). The performance result is less.

The major drawbacks of this model are people are easily frustrated, insecurity, dependency on the superiors, and minimum performance.

Custodial Model

The root level of this model is economic resources with a managerial orientation of money. The employees in this model are oriented towards security and benefits provided to them. They are dependent on the organization. The employee requirement that is met is security.

This model is adapted by firms having high resources as the name suggest. It is dependent on economic resources. This approach directs to depend on firm rather than on manager or boss. They give passive cooperation as they are satisfied but not strongly encouraged.

Supportive Model

The root level of this model is leadership with a managerial orientation of support. The employees in this model are oriented towards their job performance and participation. The employee requirement that is met is status and recognition. The performance result is awakened drives.

This model is dependent on leadership strive (attempt). It gives a climate to help employees grow and accomplish the job in the interest of the organization. Management job is to assist the employee's job performance. Employees feel a sense of participation.

Collegial Model

The root level of this model is partnership with a managerial orientation of teamwork. The employees in this model are oriented towards responsible behavior and self-discipline. The employee requirement that is met is self-actualization. The performance result is moderate zeal.

This is an extension of supportive model. The team work approach is adapted for this model. Self-discipline is maintained. Workers feel an obligation to uphold quality standard for the better image of the company. A sense of “accept” and “respect” is seen.

Organizational Behavior - Learning

Learning can be defined as the permanent change in behavior due to direct and indirect experience. It means change in behavior, attitude due to education and training, practice and experience. It is completed by acquisition of knowledge and skills, which are relatively permanent.

Nature of Learning

Nature of learning means the characteristic features of learning. Learning involves change; it may or may not guarantee improvement. It should be permanent in nature, that is learning is for lifelong.

The change in behavior is the result of experience, practice and training. Learning is reflected through behavior.

Factors Affecting Learning

Learning is based upon some key factors that decide what changes will be caused by this experience. The key elements or the major factors that affect learning are motivation, practice, environment, and mental group.

Coming back to these factors let us have a look on these factors −

- **Motivation** − The encouragement, the support one gets to complete a task, to achieve a goal is known as motivation. It is a very important aspect of learning as it acts gives us a positive energy to complete a task. **Example** − The coach motivated the players to win the match.
- **Practice** − We all know that ”Practice makes us perfect”. In order to be a perfectionist or at least complete the task, it is very important to practice what we have learnt. **Example** − We can be a programmer only when we execute the codes we have written.
- **Environment** − We learn from our surroundings, we learn from the people around us. They are of two types of environment – internal and external. **Example** − A child when at home learns from the family which is an internal environment, but when sent to school it is an external environment.
- **Mental group** − It describes our thinking by the group of people we chose to hang out with. In simple words, we make a group of those people with whom we connect. It can be for a social cause where people with the same mentality work in the same direction. **Example** − A group of readers, travelers, etc.

These are the main factors that influence what a person learns, these are the root level for our behavior and everything we do is connected to what we learn.

How Learning Occurs?

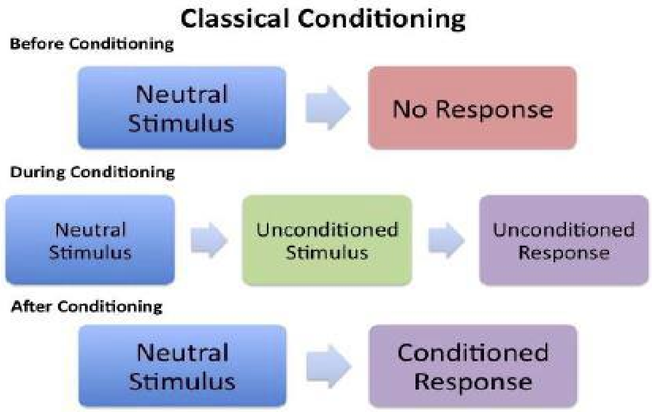
Learning can be understood clearly with the help of some theories that will explain our behavior. Some of the remarkable theories are −

- Classical Conditioning Theory
- Operant Conditioning Theory
- Social Learning Theory
- Cognitive Learning Theory

Classical Conditioning Theory

The classical conditioning occurs when a conditioned stimulus is coupled with an unconditioned stimulus. Usually, the conditioned stimulus (CS) is an impartial stimulus like the sound of a tuning fork, the unconditioned stimulus (US) is biologically effective like the taste of food and the unconditioned response (UR) to the unconditioned stimulus is an unlearned reflex response like salivation or sweating.

After this coupling process is repeated (for example, some learning may already occur after a single coupling), an individual shows a conditioned response (CR) to the conditioned stimulus, when the conditioned stimulus is presented alone. The conditioned response is mostly similar to the unconditioned response, but unlike the unconditioned response, it must be acquired through experience and is nearly impermanent.



Operant Conditioning Theory

Operant conditioning theory is also known as instrumental conditioning. This theory is a learning process in which behavior is sensitive to, or controlled by its outcomes.

Let's take an example of a child. A child may learn to open a box to get the candy inside, or learn to avoid touching a hot stove. In comparison, the classical conditioning develops a relationship between a stimulus and a behavior. The example can be further elaborated as the child may learn to salivate at the sight of candy, or to tremble at the sight of an angry parent.

In the 20th century, the study of animal learning was commanded by the analysis of these two sorts of learning, and they are still at the core of behavior analysis.

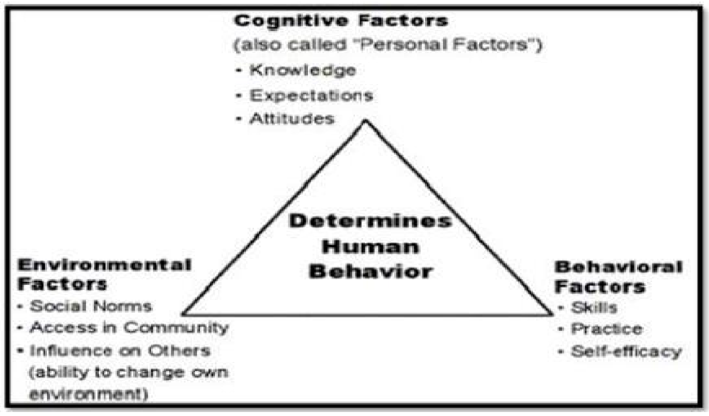
	Reinforcement Increase Behavior	Punishment Decrease Behavior
Positive Stimulus (something added)	Positive Reinforcement Add something to increase behavior	Positive Punishment Add something to decrease behavior
Negative Stimulus (something removed)	Negative Reinforcement Remove something to increase behavior	Negative Punishment Remove something to decrease behavior

Social Learning Theory

The key assumptions of social learning theory are as follows −

- Learning is not exactly behavioral; instead it is a cognitive process that takes place in a social context.
- Learning can occur by observing a behavior and by observing the outcomes of the behavior (known as vicarious reinforcement).
- Learning includes observation, extraction of information from those observations, and making decisions regarding the performance of the behavior (known as observational learning or modeling). Thus, learning can occur beyond an observable change in behavior.

- Reinforcement plays an important role in learning but is not completely responsible for learning.
- The learner is not a passive receiver of information. Understanding, environment, and behavior all mutually influence each other.



Cognitive Learning Theory

Cognition defines a person's ideas, thoughts, knowledge, interpretation, understanding about himself and environment.

This theory considers learning as the outcome of deliberate thinking on a problem or situation based upon known facts and responding in an objective and more oriented manner. It perceives that a person learns the meaning of various objects and events and also learns the response depending upon the meaning assigned to the stimuli.

This theory debates that the learner forms a cognitive structure in memory which stores organized information about the various events that occurs.

Learning & Organizational Behavior

An individual's behavior in an organization is directly or indirectly affected by learning.

Example − Employee skill, manager's attitude are all learned.

Behavior can be improved by following the listed tips −

- Reducing absenteeism by rewarding employees for their fair attendance.
- Improving employee discipline by dealing with employee's undesirable behavior, drinking at workplace, stealing, coming late, etc. by taking appropriate actions like oral reprimands, written warnings and suspension.
- Developing training programs more often so as to grab the trainees' attention, provide required motivational properties etc.

Organizational Behavior - Personality

The word personality is derived from a Greek word *"persona"* which means "to speak through." Personality is the combination of characteristics or qualities that forms a person's unique identity. It signifies the role which a person plays in public. Every individual has a unique, personal and major determinant of his behavior that defines his/her personality.

Personality trait is basically influenced by two major features −

- Inherited characteristics
- Learned characteristics

Inherited Characteristics

The features an individual acquires from their parents or forefathers, in other words the gifted features an individual possesses by birth is considered as inherited characteristics. It consists of the following features −

- Color of a person's eye
- Religion/Race of a person
- Shape of the nose
- Shape of earlobes

Learned Characteristics

Nobody learns everything by birth. First, our school is our home, then our society, followed by educational institutes. The characteristics an individual acquires by observing, practicing, and learning from others and the surroundings is known as learned characteristics.

Learned characteristics includes the following features −

- **Perception** − Result of different senses like feeling, hearing etc.
- **Values** − Influences perception of a situation, decision making process.
- **Personality** − Patterns of thinking, feeling, understanding and behaving.
- **Attitude** − Positive or negative attitude like expressing one's thought.

Traits of Personality

Personality traits are the enduring features that define an individual's behavior. A personality trait is a unique feature in an individual. Psychologists resolved that there are five major personality traits and every individual can be categorized into at least one of them. These five personality traits are −

- Extrovert
- Neurotic
- Open
- Agreeable
- Conscientious

Major Personality Attributes

Following are the five major personality attributes that influence OB −

Locus of Control

Locus of control is the center of control of an individual's code of conduct. People can be grouped into two categories i.e., internals and externals respectively.

People who consider themselves as the masters of their own fates are known as internals, while, those who affirm that their lives are controlled by outside forces known as externals.

Before making any decision, internals actively search for information, they are achievement driven, and want to command their environment. Thus, internals do well on jobs that craves complex information processing, taking initiative and independent action.

Externals, on the other hand, are more compliant, more willing to follow instructions, so, they do well in structured, routine jobs.

Machiavellianism (ma·kee·uh·**veh**·lee·uhn-ism)

Machiavellianism is being practical, emotionally distant, and believing that ends justify means.

Machiavellians always want to win and are great persuaders. Here are the significant features of high-mach individuals −

- High-Machs prefer precise interactions rather than beating about the bush.
- High-Machs tend to improvise; they do not necessarily abide by rules and regulations all the time.
- High-Machs get distracted by emotional details that are irrelevant to the outcome of a project.

Self-esteem

It is the extent up to which people either like or dislike themselves. Self-Esteem is directly related to the expectations of success and on-the-job satisfaction.

Individuals with high self-esteem think that they have what it takes to succeed. So, they take more challenges while selecting a job.

On the other hand, individuals with low self-esteem are more susceptible to external distractions. So, they are more likely to seek the approval of others and to adapt the beliefs and behaviors of those they respect.

Self-monitoring

Self-monitoring is the capability of regulating one's behavior according to social situations. Individuals with high self-monitoring skill easily adjust their behavior according to external, situational factors. Their impulsive talents allow them to present public personae which are completely different from their private personalities.
However, people with low self-monitoring skills cannot cover themselves. Regardless of any situation, they are always themselves. They have an attitude of, "what you see is what you get."

Risk taking
Generally, managers are reluctant on taking risks. However, individual risk-taking inclination affects the bulk of information required by the managers and how long it takes them to make decisions.
Thus, it is very important to recognize these differences and align risk-taking propensity with precise job demands that can make sense.

Theories of Personality
A theory is a simple model of reality that helps us understand, explain, predict and deal with reality. We have some theories that explain an individual's personality.

Sigmund Freud's Psychoanalytic Theory
This theory is based on the belief that man is encouraged more by unforeseen forces than the conscious and logical thought. Freud believed that most of the things in life are not present at the conscious level but they are present at an unconscious level.
The features of Freud's theory include three attributes − Id, Ego, and Superego.

- Id − It defines the innate component of personality. It is the impulsive and unconscious part of mind that seeks immediate satisfaction. Example − A hungry baby cries till he/she is fed.
- Ego − It is derived from Id and assists in dealing with the external world. It also helps in translating the inner needs into expressions. It deals with practical and rational thinking process. Example − We have a fight with our friend and expect the friend to talk first, even though both of us want to talk.
- Superego − It is different from ego and is partially unconscious. It includes the traditional values of society as interpreted by our parents. It also helps in the integral vision of punishment. Example − Ram came late today so he is grounded for a week.

Erikson's Theory
This theory states that personality is groomed throughout lifetime. He presents eight distinct stages each with two possible outcomes. Successful completion of each stage leads to a healthy personality. These stages are −

- **Infancy** − It is the period between 0-1 years of age. In this stage, children learn the ability to trust others depending on their caregivers. Unsuccessful completion in this stage results in anxiety and insecurity. **Example** − Children of this age are more comfortable with those faces they see more often and not with strangers.
- **Early Childhood** − It is the period between 1-3 years of age. In this stage, children learn to be independent. If given support, they become more confident else they become dependent over others. **Example** − Children in this age are taught how to walk, how to talk etc.
- **Play Age** − It is the period between 3-6 years of age. In this stage, children assert themselves frequently. The failure leads to development of a sense of guilt among them. **Example** − Children in this age group, need to be taught how to behave and should be taught to be focused.
- **School Age** − It is the period between 6 years of age till puberty. In this stage, children become more innovative. They feel confident and want to achieve their goals. If not encouraged they may feel inferior. **Example** − Teenagers should be protected and parents need to understand them and should handle them patiently.
- **Adolescence** − This stage is a transformation from childhood to adulthood. Here children find their own identity and should be guided and supported in order to help them choose the right direction. **Example** − Decision such as which stream to choose science or commerce etc. happens during this stage.

- **Young Childhood** − This stage is also known as young adulthood. Here, they begin to open up and become more intimate with others. **Example** − Making close friends.
- **Adulthood** − In this stage, they focus on establishing career and settling down with relationships that are important. **Example** − Applying for jobs.
- **Mature Adulthood** − In this stage, a person is old and thus in this stage the productivity slows down. **Example** − Taking care of the family.

Sheldon's Physiognomy Theory

This theory was proposed by William Sheldon. He presents personalities by classifying individuals into convenient categories based on their body shapes. They are −

- Endomorphs
- Mesomorphs
- Ectomorphs

Endomorphs

In this category, the body is soft and spherical. People with this kind of personality love comfort, eat a lot, like to be around people and desire affection. Some common endomorph features are large amount of fat accumulation, insatiable appetite, larger frame etc.

Some endomorph personalities are John Goodman, Jack Black etc.

Mesomorphs

In this category, the body is hard and rectangular physique. People with this kind of personality like to take risk, are courageous and have power. Some common mesomorph features are wide shoulders, small waist, low body fat.

Some mesomorph personalities are Jennifer Garner, Tina Turner etc.

Ectomorphs

In this category, the body is fragile, flat chest and delicate body. People with this kind of personality are anxious, ambitious and dedicated. Some common ectomorph features are narrow frame, low body fat, etc.

Some notable ectomorph personalities are Brad Pitt, Bruce Lee etc.

Organizational Behavior - Perception

Perception is an intellectual process of transforming sensory stimuli to meaningful information. It is the process of interpreting something that we see or hear in our mind and use it later to judge and give a verdict on a situation, person, group etc.

It can be divided into six types −

- **Of sound** − The ability to receive sound by identifying vibrations.
- **Of speech** − The competence of interpreting and understanding the sounds of language heard.
- **Touch** − Identifying objects through patterns of its surface by touching it.
- **Taste** − The ability to receive flavor of substances by tasting it through sensory organs known as taste buds.
- **Other senses** − They approve perception through body, like balance, acceleration, pain, time, sensation felt in throat and lungs etc.
- **Of the social world** − It permits people to understand other individuals and groups of their social world. **Example** − Priya goes to a restaurant and likes their customer service, so she will perceive that it is a good place to hang out and will recommend it to her friends, who may or may not like it. Priya's perception about the restaurant is good.

Perceptual Process

Perceptual processes are the different stages of perception we go through. The different stages are −

- Receiving

- Selecting
- Organizing
- Interpreting

Receiving

Receiving is the first and most important stage in the process of perception. It is the initial stage in which a person collects all information and receives the information through the sense organs.

Selecting

Selecting is the second stage in the process. Here a person doesn't receive the data randomly but selectively. A person selects some information out of all in accordance with his interest or needs. The selection of data is dominated by various external and internal factors.

- **External factors** − The factors that influence the perception of an individual externally are intensity, size, contrast, movement, repetition, familiarity, and novelty.
- **Internal factors** − The factors that influence the perception of an individual internally are psychological requirements, learning, background, experience, self-acceptance, and interest.

Organizing

Keeping things in order or say in a synchronized way is organizing. In order to make sense of the data received, it is important to organize them.

We can organize the data by −

- Grouping them on the basis of their similarity, proximity, closure, continuity.
- Establishing a figure ground is the basic process in perception. Here by figure we mean what is kept as main focus and by ground we mean background stimuli, which are not given attention.
- Perceptual constancy that is the tendency to stabilize perception so that contextual changes don't affect them.

Interpreting

Finally, we have the process of interpreting which means forming an idea about a particular object depending upon the need or interest. Interpretation means that the information we have sensed and organized, is finally given a meaning by turning it into something that can be categorized. It includes stereotyping, halo effect etc.

Importance of Perception in OB

We need to understand what the role of perception in an organization is. It is very important in establishing different role of perceptions like −

- Understanding the tasks to be performed.
- Understanding associated importance of tasks allotted.
- Understanding preferred behavior to complete respective tasks.
- Clarifying role perceptions.

For example, every member in a group has to be clear regarding the role allotted to them. Programmer writes the code, tester checks it, etc.

Organizational Behavior - Motivation

Motivation can be described as the internal force that impacts the direction, intensity, and endurance of a person's voluntary choice of behavior. It consists of −

- **Direction** − focused by goals.
- **Intensity** − bulk of effort allocated.
- **Persistence** − amount of time taken for the effort to be exerted.

Example − A team leader encourages team members to work efficiently.

Features of Motivation

Motivation is an internal feeling, that is, it defines the psychological state of a person. It is a continuous process and we should make sure that it is not disturbed. A person should be encouraged completely.
Motivation consists of three interacting and dependent elements −

- **Needs** − The requirements or deficiency which is created whenever there is physiological imbalance.
- **Drives** − The various camps or events organized to motivate the employees and give them new opportunities.
- **Incentives** − Employees need to be rewarded for their nice work in order to keep them encouraged.

Importance of Motivation
We need to motivate employees because of the following reasons −

- Motivated employee are more quality oriented.
- Highly motivated employees are more productive as compared to other employees.
- It helps in achieving three behavior dimension of human resource namely
    - Candidates must be attracted not only to join but also remain in the firm.
    - Employees must perform task in a dependable manner.
    - Employees should be creative, spontaneous and innovative at work.

Maslow's Hierarchy of Needs Theory
This theory was produced in order to answer the question "What motivates an individual". Every second need comes to force when the first need is satisfied completely. Maslow explained the hierarchy of needs by grouping them into two: deficiency needs and growth needs.

Physiological Needs
Every individual needs to take care of the basic requirements required to sustain. These requirements include food to eat, clothing to wear and shelter to live in. These necessities are relatively independent of each other but are finite.

Safety Needs
Everybody wants to stay in a protected environment with minimal danger so that they can have a peaceful life. Safety needs basically includes protection from physiological danger like accident and having economic security like bank accounts, health insurance
In an enterprise, it includes job security, salary increment, etc. The managerial practice to satisfy this involves offering pension scheme, provident fund, gratuity etc.

Social Needs
We have all heard that man is a social animal, we want to be there with those people where we are loved and we are accepted as we are; nobody wants to be judged. This is a common requirement every human desires.
This theory helps managers to think about encouraging their employees by identifying employee needs. In short, it presents motivation as constantly changing force, expressing itself to the constant need for fulfilment of new and higher levels of needs.

Esteem

Esteem means the typical human desire to be accepted and valued by others. People often involve in a profession or hobby to gain recognition, earn fame and respect. According to Maslow, the needs of humans have strict guidelines - the hierarchies rather than being sharply separated, are interrelated. This means that esteem and the consequent levels are not strictly separated but are closely related.

Self-Actualization

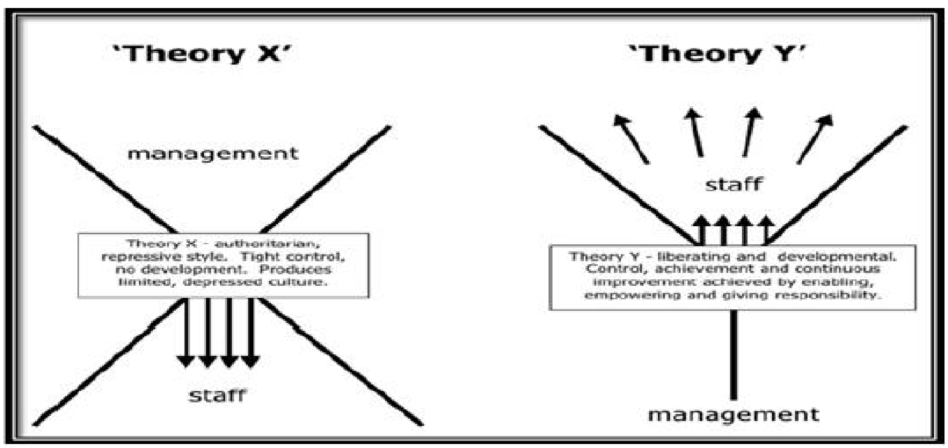
Self-actualization means realizing one's full potential. Maslow describes this as a desire to complete everything that one can, to become the most that one can be.

Theory X & Theory Y

Our management style is firmly influenced by our beliefs and assumptions about what encourages members of our team like: If we believe that our team members dislike work, then we tend towards an authoritarian style of management. However, if we assume that employees take pride in doing a good job, we tend to adopt a more participative style.

Douglas McGregor, the eminent social psychologist, divides management style into two contrasting theories −

- Theory X
- Theory Y



Theory X

This theory believes that employees are naturally unmotivated and dislike working, and this encourages an authoritarian style of management. According to this theory, management must firmly intervene to get things done. This style of management concludes that workers −

- Disfavor working.
- Abstain responsibility and the need to be directed.
- Need to be controlled, forced, and warned to deliver what's needed.
- Demand to be supervised at each and every step, with controls put in place.
- Require to be attracted to produce results, else they have no ambition or incentive to work.

McGregor observed that X-type workers are in fact mostly in minority, and yet in mass organizations, such as large scale production environment, X Theory management may be needed and can be unavoidable.

Theory Y

This theory explains a participative style of management, that is, distributive in nature. It concludes that employees are happy to work, are self-motivated and creative, and enjoy working with greater responsibility. It estimates that workers −

- Take responsibility willingly and are encouraged to fulfill the goals they are given.
- Explore and accept responsibility and do not need much guidance.
- Assume work as a natural part of life and solve work issues imaginatively.

In Y-type organizations, people at lower levels are engaged in decision making and have more responsibility.

Comparing Theory X & Theory Y
Let us now compare both the theories −

Motivation
Theory X considers that people dislike work; they want to avoid it and do not take responsibilities willingly. While, Theory Y considers that people are self-motivated, and sportingly take responsibilities.

Management Style and Control
In Theory X-type organization, management is authoritarian, and centralized control is maintained.
While in Theory Y-type organization, the management style is participative, employees are involved decision making, but the power retains to implement decisions.

Work Organization
Theory X employees are specialized and the same work cycle continues.
In Theory Y, the work tends to be coordinated around wider areas of skill or knowledge. Employees are also motivated to develop expertise, and make suggestions and improvements.

Rewards and Appraisals
Theory X-type organizations work on a 'carrot and stick' basis, and performance assessment is part of the overall mechanism of control and compensation.
Coming to Theory Y-type organizations, appraisal is also regular and crucial, but is usually a separate mechanism from organizational controls. Theory Y-type organizations provide employees frequent opportunities for promotion.

Application
Admitting the fact that Theory X management style is widely accepted as inferior to others, it has its place in large scale production procedure and unskilled production-line work.
Many of the principles of Theory Y are widely accepted by different types of organization that value and motivate active participation.
Theory Y-style management is appropriate for knowledge work and licensed services. Licensed service organizations naturally develop Theory Y-type practices by the nature of their work, even high structure knowledge framework, like call center operations, benefit from its principles to motivate knowledge sharing and continuous improvement.

Organizational Behavior - Groups
A group can be defined as two or more interacting and interdependent individuals who come together to achieve particular objectives. A group behavior can be stated as a course of action a group takes as a family. For example: Strike.

Types of Groups
There are two types of groups an individual forms. They are formal groups and informal groups. Let us know about these two groups.

Formal Groups

These are the type of work groups created by the organization and have designated work assignments and rooted tasks. The behavior of such groups is directed toward achieving organizational goals.

These can be further classified into two sub-groups −

- **Command group** − It is a group consisting of individuals who report directly to the manager.
- **Interest group** − It is a group formed by individuals working together to achieve a specific objective. **Example** − A group of workers working on a project and reporting to the same manager is considered as a command group. A group of friends chilling out together is considered as interest group or say members of a club.

Informal Groups

These groups are formed with friendships and common interests. These can be further classified into two sub-groups −

- **Task group** − Those working together to finish a job or task is known as task group.
- **Friendship group** − Those brought together because of their shared interests or common characteristics is known as friendship group.

Why Do People Join Groups

There is no particular reason answering why individuals join groups. Group helps individual to feel stronger, have fewer self-doubts, and be more contrary to threats.

The following points help us understand the need of joining a group by individuals −

- **Security mirrors strength in numbers**. Status pinpoints a prestige that comes from belonging to a specific group. Inclusion in a group is considered as important because it provides recognition and status.
- **Self-esteem transmits people's feelings of self-worth**. Membership can sometimes raise feelings of self-esteem like being accepted into a highly valued group.
- **Affiliation with groups can meet one's social needs**. Work groups significantly contribute to meet the need for friendships and social relations.
- **Groups represent power**. What mostly cannot be achieved individually becomes possible with group effort. Power might be aimed to protect themselves from unreasonable demands. Informal groups provide options for individuals to practice power.
- **People may join a group for goal achievement**. Sometimes it takes more than one person to accomplish a particular task.

Group Roles

The concept of roles is applicable to all employees within an organization as well as to their life outside the organization. A role is a set of expected behavior patterns attributed to the one who occupies the position demanded by the social unit. Individuals play multiple roles at the same time. Employees attempt to understand what kind of behavior is expected from them. An individual when presented by divergent role expectations experiences role conflict. Group roles are divided into three types −

- Task-oriented Roles
- Relationship-oriented Roles
- Individual Roles

Task-oriented Roles

Roles allotted to individuals according to their work and eligibility is known as task-oriented roles. Task-oriented roles can broadly divide individuals into six categories initiator, informer, clarifier, summarizer, reality tester and information seekers or providers respectively.

- **Initiator** − The one who proposes, suggests, defines.
- **Informer** − The one who offers facts, expresses feelings, gives opinions.
- **Clarifier** − The one who interprets, defines, clarifies everything.
- **Summarizer** − The one who links, restates, concludes, summarizes.

- **Reality Tester** − The one who provides critical analysis.
- **Information seekers or providers** − The one who gives information and data.

These roles present the work performed by different individuals according to their marked designation.

Relationship-oriented Roles

Roles that group individuals according to their efforts made to maintain healthy relationship in the group and achieve the goals are known as relationship-oriented roles. There are five categories of individuals in this category namely: harmonizer, gatekeeper, consensus tester, encourager, and compromiser.

- **Harmonizers** − The one who limits tension and reconciles disagreements.
- **Gatekeeper** − The one who ensures participation by all.
- **Consensus Tester** − The one who analyzes the decision-making process.
- **Encourager** − The one who is warm, responsive, active, shows acceptance.
- **Compromiser** − The one who admits error, limits conflict.

These roles depict the various roles an individual plays to maintain healthy self as well as group relationships.

Individual Roles

Roles that classify a person according to the measure of individual effort put in the project aimed is known as individual roles. Five types of individuals fall into these roles: aggressor, blocker, dominator, cavalier, and avoidance.

- **Aggressor** − The one who devalues others, attacks ideas.
- **Blocker** − The one who disagrees and rebels beyond reason.
- **Dominator** − The one who insists superiority to manipulate.
- **Cavalier** − The one who takes part in a group non-productively.
- **Avoidance** − The one who shows special interest to avoid task.

These are the various roles a person plays in an organization.

Well-Functioning Groups

We know what a group is, why it is important to form a group, and what the group-oriented roles are. Now we need to know how to mark a group as a well-functioning group, what features are necessary for a group to mark it as efficient.

A group is considered effective when it has the following characteristics.

- Relaxed, comfortable, friendly atmosphere.
- Task to be executed are well understood and accepted.
- Members listen well and actively participate in given assignments.
- Assignments are made clear and are accepted.
- Group is acquainted of its operation and function.
- People express their feelings and ideas openly.
- Consensus decision-making process is followed.
- Conflict and disagreement center regarding ideas or method.

Reference Groups

It is a group to which a person or another group is compared. Reference groups are used in order to examine and determine the nature of a person or other group's features and sociological attributes. It is the group to which a person relates or aspires to link himself or herself psychologically.

It is important for deciding a person's self-identity, attitudes, and social ties. It becomes the ground of reference in making comparisons or contrasts and in judging one's appearance and performance. These groups act as a benchmark and contrast needed for comparison and evaluation of group and personal characteristics.

An example of a reference group, would be the certainty of wealth. An individual in the US with an annual income of $70,000, may consider himself rich if he compares himself to those in the middle income strata, who earn roughly $22,000 a year. However, if the same person considers the relevant reference group to be those in the top 0.5% of households in the US, those making $1.8 million or more, then the individual's income of $70,000 would make him seem rather poor.

Group Decision Making

Group decision-making commonly known as collaborative decision-making is a situation faced when individuals collectively make a choice from the alternatives before them.

The decision is then no longer attributable to any individual group member as all the individuals and social group processes like social influence contribute to the decision outcome.

The decisions made by groups are mostly different from those made by individuals. For example, groups tend to make decisions that are more extreme than those made by individual members, as individuals tend to be biased.

Advantages of Group Decision Making

Group decision making has two advantages over individual decision making.

1. Synergy

It is the idea that the whole is greater than the aggregate of its parts. When a group makes a decision collectively, its judgment can be powerful than that of any of its members. Through discussing, questioning, and collaborative approach, group members can identify more complete and robust solutions and recommendations.

2. Sharing of information

Group decisions take into account a wider scope of information as each group member may contribute distinct information and expertise. Sharing information increases understanding, clarifies issues, and facilitates movement towards a collective decision.

Disadvantages of Group Decision Making

The major disadvantages of group decision making are as follows −

1. Diffusion of Responsibility

Group decision making results in distribution of responsibility that results in lack of accountability for outcomes. In this way, everyone is responsible for a decision, and no one really is. Moreover, group decisions can make it easier for members to refuse personal responsibilities and blame others for bad decisions.

2. Lower Efficiency

Group decisions can sometimes be less efficient than individual decisions. It takes additional time because there is a need of active participation, discussion, and coordination among group members. Without good facilitation and structure, meetings can get eliminated in trivial details that may matter a lot to one person but not to the others.

3. Groupthink

One of the biggest disadvantage of effective group decision making is groupthink. It is a psychological phenomenon that occurs within a group of people in which the wish for harmony or conformity results in an illogical or dysfunctional decision-making outcome.

By refraining themselves from outside influences and actively suppressing opposing viewpoints in the interest of minimizing conflict, group members reach a consensus decision without critical evaluation of substitute viewpoints.

Groupthink sometimes produces dehumanizing actions against the out-group.

Group Decision-Making Techniques

In order to eliminate group think and group shift from a group, we can use four different techniques that will help us make a collaborative decision that is best for the group. These techniques are −

- Brainstorming
- Nominal group thinking
- Didactic technique
- Delphi technique

Brainstorming

This technique includes a group of people, mostly between five and ten in number, sitting around a table, producing ideas in the form of free association. The main focus is on generation of ideas and not on evaluation of these ideas.

If more ideas can be originated, then it is likely that there will be a unique and creative idea among them. All these ideas are written on the blackboard with a piece of chalk so that all the team members can see every idea and try to improvise these ideas.

Brainstorming technique is very effective when the problem is comparatively precise and can be simply defined. A complex problem can be divided into parts and each part can be dealt with separately at a time.

Nominal Group Thinking

This technique is similar to brainstorming except that this approach is more structured. It motivates individual creativity. Members form the group for namesake and operate independently, originate ideas for solving the problem on their own, in silence and in writing. Members do not communicate well with each other so that strong personality domination is evaded.

The group coordinator either collects the written ideas or writes them on a large blackboard so that each member of the group can see what the ideas are. These ideas are further discussed one by one in turn and each participant is motivated to comment on these ideas in order to clarify and improve them. After all these ideas have been discussed, they are evaluated for their merits and drawbacks and each actively participating member is needed to vote on each idea and allot it a rank on the basis of priority of each alternative solution.

The idea with the highest cumulative ranking is selected as the final solution to the problem.

Didactic Interaction

This technique is applicable only in certain situations, but is an excellent method when a situation actually demands it. The type of problem should be such that it generates output in the form of yes, or no.

Say for example, a decision is to be made whether to buy or not to buy a product, to merge or not to merge, to expand or not to expand and so on. This type of decision requires an extensive and exhaustive discussion and investigation since a wrong decision can have serious consequences.

There are many advantages as well as disadvantages of this type of situation. The group that makes the decision is divided into two sub-groups, one in favor of the "go" decision and the opposing in favor of "no go" decision.

The first group enlists all the "pros" of the problem solution and the second group lists all the "cons". These groups meet and discuss their discoveries and their reasons.

After tiring discussions, the groups switch sides and try to find weaknesses in their own original standpoints. This interchange of ideas and understanding of various viewpoints results in mutual acceptance of the facts as they exist so that a solution can be put together around these facts and ultimately a final decision is reached.

Delphi Technique

This technique is the improvised version of the nominal group technique, except that it involves obtaining the opinions of experts physically distant from each other and unknown to each other.

This isolates group members from the undue influence of others. Basically, the types of problems sorted by this technique are not specific in nature or related to a particular situation at a given time.

For example, the technique could be used to explain the problems that could be created in the event of a war. The Delphi technique includes the following steps −

- The problem is first identified and a panel of experts is selected. These experts are asked to provide potential solutions through a series of thoughtfully designed questionnaires.
- Each expert concludes and returns the initial questionnaire.
- The results of the questionnaire are composed at a central location and the central coordinator prepares a second set of questionnaire based on the previous answers.
- Each member receives a copy of the results accompanied by the second questionnaire.
- Members are required to review the results and respond to the second questionnaire. The results typically trigger new solutions or motivate changes in the original ideas.
- The process is repeated until a general agreement is obtained.

Organizational Behavior - Leadership

Leadership can be defined as the ability of the management to make sound decisions and inspire others to perform well. It is the process of directing the behavior of others towards achieving a common goal. In short, leadership is getting things done through others.

Importance of Leadership

Leadership is very important in a firm as it leads to higher performance by the team members, it improves motivation and morale within the members, and helps to respond to change.

Leadership facilitates organizational success by creating responsibility and accountability among the members of the organization. In short, it increases value in an organization.

Leader Vs Manager

A leader is someone whom people follow or someone who guides or directs others. A manager is someone who is responsible for directing and controlling the work and staff in an organization, or of a department within it.

The main difference between the two is that a leader works by example, while a manager dictates expectations. If a manager goes against the rules, that will tarnish his position as a manager. If a leader goes against the example he or she is trying to set, that will be seen as a setback. Following are a few subtle differences between the two −

- A leader is an innovator and creator whereas a manager is a commander.
- A leader can’t be a manager but the opposite is possible, a manager is more than a leader.
- A leader does what is right, while the manager makes things right.
- A leader deals with change whereas a manager plans for a change.
- A leader gives direction to do something whereas the manager plans for everything that is to be done.
- A leader encourages people whereas the manager controls people.
- A leader handles communication, credibility, and empowerment whereas a manager deals with organizing and staffing.

Leadership Styles

Different leadership styles exist in work environments. The culture and goal of an organization determine which leadership style fits best. Some organizations offer different leadership styles within an organization, depending on the necessary tasks to complete and departmental needs.

We find five different leadership styles in the corporate world. They are as follows −

1. Laissez-Faire

A laissez-faire leader does not directly supervise employees and fails to provide regular updates to those under his supervision. Highly experienced and trained employees with minimal requirement of supervision fall under the laissez-faire leadership style.

But, not all employees possess these features. This leadership style blocks the production of employees needing supervision. The laissez-faire style implements no leadership or supervision efforts from managers, which can lead to poor production, lack of control and increasing costs.

2. Autocratic

The autocratic leadership style permits managers to make decisions alone without the input of others. Managers access total authority and impose their will on employees. No one opposes the decisions of autocratic leaders. Countries like Cuba and North Korea operate under the autocratic leadership style.

This leadership style benefits those who require direct supervision. Creative employees who participate in group functions detest this leadership style.

3. Participative

This is also known as the democratic leadership style. It values the input of team members and peers, but the responsibility of making the final decision rests with the participative leader. Participative leadership motivates employee morale because employees make contributions to the decision-making process. It accounts to a feeling that their opinions matter.

When an organization needs to make changes within itself, that is internally, the participative leadership style helps employees accept changes easily as they play a role in the process. This leadership style meets challenges when companies need to make a decision in a short period of time.

4. Transactional

Transactional leadership style is formed by the concept of reward and punishment. Transactional leaders believe that the employee's performance is completely dependent on these two factors. When there is an encouragement, the workers put in their best effort and the bonus is in monetary terms in most of the cases. In case they fail to achieve the set target they are given a negative appraisal.

Transactional leaders pay more attention to physical and security requirements of the employees.

5. Transformational

Transformational leadership has the ability to affect employee's perceptions through the returns that organization gets in the form of human capital benefits. These leaders have the ability to reap higher benefits by introducing knowledge management processes, encouraging interpersonal communication among employees and creating healthy organizational culture.

It helps in flourishing organizational innovation by creating a participative environment or culture. It promotes a culture where the employees have autonomy to speak about their experiences and share knowledge.

It has been seen that transformational leaders are more innovative than transactional and laisse-faire leaders.

Traditional Theory

Traditional theory is a theory based on different traits of a human beings. It assumes that leaders are born and not made. According to this theory, leadership behavior is the sum total of all traits that a leader possess.

Thus this theory gives the profile of a successful and complete leader. According to this theory, there are five human traits. They are −

- **Physical trait** − it includes energy, activity, appearance, and height.
- **Ability trait** − it includes judgement, knowledge, and fluency in speech.
- **Personal trait** − it includes self-confidence, creativity, and enthusiasm.
- **Work trait** − it includes organization and achievement.
- **Social trait** − it includes interpersonal skill, cooperativeness, popularity and prestige.

Drawbacks

Following are the major drawbacks of this theory −

- Traits are not arranged according to their importance.
- There is no quantitative tool to judge the human traits.
- This trait can’t be used universally.
- This trait can be achieved and developed.

- Situational factors are avoided.

Theories of Leadership

1. Behavioral Theory

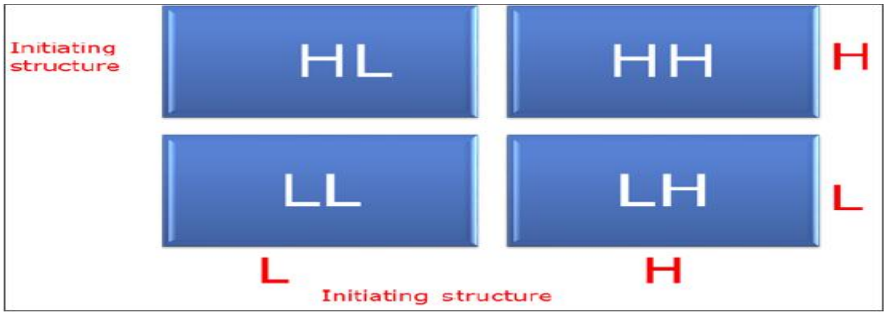
This theory explains the effectiveness of leadership. According to this theory, leadership has two qualities i.e., initiating structure and consideration. These qualities are tested with higher and lower levels with proper intersection of each other.

Initiating Structure

It is the level up to which a leader is task oriented and directs the employee towards achieving a goal. In this case, the leader gives instruction, makes plan and schedules work activities.

Consideration

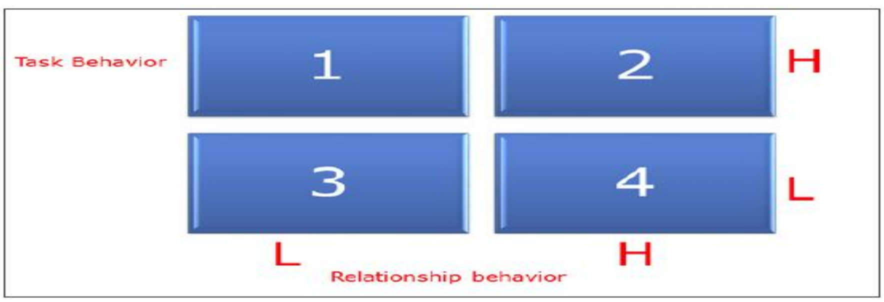
It is the level up to which a leader is concerned with the sub-ordinates, ideas and feelings. Considerate leaders are friendly, they show concern for sub-ordinates' well-being and satisfaction.



This type of leadership is achieved by performance and is found to be effective. But it is not the best way as situational factors are not taken into consideration.

2. Contingency Theory

According to this theory, propounded by Paul Hersey and Ken Blanchard, believes the effectiveness of a leader is dependent on the action or readiness of his followers. By readiness we mean the extent to which the followers are able and willing to achieve the goal. This theory is explained on the basis of four cases.



- **Case 1** − In case one, we have high relationship behavior and low task behavior. The leader motivates the followers and helps in decision making. Not much productivity can be seen in this case but the sense of togetherness is high.
- **Case 2** − In case two, we have high relationship behavior as well as high task behavior. In this combination, the leader explains the decision and helps to build confidence of the employees. In this case, productivity as well as loyalty towards the leader is more.
- **Case 3** − In case three, we have a combination of low relationship behavior and low task behavior. Here we see that the leader delegates the responsibility of decision making to the followers. In this case, there is poor communication as well as poor production.
- **Case 4** − Here we deal with a combination of low relationship behavior and high task behavior. The leader gives specific direction and supervises the performance. This theory is effective only if the leaders change their style irrespective of the readiness of the followers.

Conflict Management

Conflict can be defined as a mental struggle resulting from incompatible or opposing needs, drives, wishes, and external or internal demands. Where there are people, there is conflict.
They are usually taken in a negative association. However, this is inaccurate as conflicts are necessary for healthy relationships. It all depends on the approach we use to resolve the conflict.

Classification of Conflict

When we think of the different types of conflict, we might instantly think of the ones referred to in literature, especially in fiction. They can be applied to real life, of course. However, in contemporary times, types of conflict which are easily identifiable are classified into four different types −

- Intrapersonal
- Intragroup
- Interpersonal
- Intergroup

Intrapersonal Conflict

Intrapersonal conflict takes place within an individual. The person experiences it in his own mind. Thus, it is a type of conflict that is psychological involving the individual's thoughts, values, principles and emotions. Intrapersonal conflict may come in different forms, from the simple mundane ones like deciding whether or not to go vegan for lunch to ones that can affect major decisions such as choosing a career path.
However, this type of conflict can be quite difficult to handle, if you find it hard to decipher your inner struggles. It results in restlessness and uneasiness, or can even cause depression. On such occasions, it is advised to seek a way to let go of the anxiety by communicating with other people. Eventually, when the person finds himself/herself out of the situation, he/she can become more empowered as a person. Thus, the experience invokes a positive change which helps in personal growth.

Intragroup Conflict

Intragroup conflict occurs among individuals within a team. The incompatibilities and misunderstandings between team members lead to intragroup conflict. It starts from interpersonal disagreements like team members have different personalities which may lead to tension or differences in views and ideas. Say for example, during a presentation, members of the team might find the notions presented by the one presiding to be erroneous due to their differences in opinion.
Within a team, conflict can be helpful in coming up with decisions, which will eventually allow them to achieve their objectives as a team. But, if the degree of conflict disrupts harmony among the members, then some serious guidance from a different party will be needed for it to be settled.

Interpersonal Conflict

Interpersonal conflict means a conflict between two individuals. Basically, this occurs because of some differences in people. We have varied personalities which usually lead to incompatible choices and opinions. So, it is a natural occurrence which can eventually help in personal growth or developing our relationships with others.
In addition, adjustments are necessary for managing this type of conflict. However, when interpersonal conflict becomes too destructive, calling in a mediator helps so as to have the issue resolved.

Intergroup Conflict

Intergroup conflict occurs when a misunderstanding arises among different teams within an organization. For example, the marketing department of an organization can come in conflict with the customer support department. This is because of the varied sets of goals and interests of these different groups. In addition to this, competition also contributes to intergroup conflict. There are other factors which increase this type of conflict. Some of these factors may include a rivalry in resources or the boundaries set by a group to others which forms their own identity as a team.

Conflict should not always be perceived as a problem rather at times it is a chance for growth and can be an effective means of opening up among groups or individuals. However, when conflict begins to suppress or disrupt productivity and gives way to more conflicts, then conflict management is what is needed for problem resolution.

Conflict Resolution

Conflict resolution is a method by which two or more parties find a peaceful solution to a disagreement among them. The disagreement can be personal, financial, political, or emotional. When a disagreement arises, often the best course of action is negotiation to resolve the disagreement. We all know that when people gather for a discussion, it is not necessary that what one thinks is right the other thinks the same way, this difference in thinking or mentality leads to conflict.

"I'm doing my best at work and you expect me to do more! Why don't you ask the other team members?" This is the start of a conflict! Let us know about some of the conflict management techniques.

Conflict Management Techniques

We get into a conflict when the person opposite to us has a different mindset. It is very common in a workplace to get into differences of opinion. Sometimes there is a conflict between two or more employees, sometimes employees have a conflict with their managers and so on. Now the question is, how can we manage disagreements in ways that build personal and collegial relationships?

Here are five strategies from conflict management theory for managing stressful situations. None of them is a "one-size-fits-all" answer. Which one is the best in a given situation depends on variety of factors, including an appraisal of the levels of conflict.

- **Collaborating** − win/win
- **Compromising** − win some/lose some
- **Accommodating** − lose/win
- **Competing** − win/lose
- **Avoiding** − no winners/no losers

Collaborating

This technique follows the rule "I win, you win". Collaborating means working together by integrating ideas set out by multiple people. The objective here is to find a creative solution acceptable to everyone. It calls for a significant time commitment but is not appropriate for all conflicts.

This technique is used in situations where −

- There is a high level of trust
- We don't want to take complete responsibility
- We want others to also have "ownership" of solutions
- People involved are willing to change their thinking
- We need to work through animosity and hard feelings

However, this process takes a lot of time and energy and some may take advantage of other people's trust and openness.

Example − A businessman should work collaboratively with the manager to establish policies, but collaborative decision-making regarding office supplies wastes time better spent on other activities.

Compromising

This technique follows the rule "You bend, I bend". Compromising means adjusting with each other's opinions and ideas, and thinking of a solution where some points of both the parties can be entertained. Similarly, both the parties need to give up on some of their ideas and should agree with the other.

This technique can be used in situations where −

- People of equal levels are equally committed to goals
- Time can be saved by reaching intermediate settlements on individual parts of complex matters
- Goals are moderately important

Important values and long-term objectives can be derailed using this technique. This process may not work if initial demands are high and mainly if there's no commitment to honor the compromise solutions.
Example − Two friends had a fight and they decide to compromise with each other through mutual understanding.

Accommodating
This technique follows the rule "I lose, you win". Accommodating means giving up of ideas and thoughts so that the other party wins and the conflict ends. This technique can be used when −

- An issue is not that important to us as it is to the other person
- We realize we are wrong
- We are willing to let others learn by mistake
- We know we cannot win
- It is not the right time and we would prefer to simply build credit for the future
- Harmony is extremely important
- What the parties have in common is a good deal more important than their differences

However, using this technique, one's own ideas don't get attention and credibility, and influence can be lost.
Example − When we fight with someone we love we choose to let them win.

Competing
This technique follows the rule "I win, you lose". Competing means when there is a dispute a person or a group is not willing to collaborate or adjust but it simply wants the opposite party to lose. This technique can be used when −

- We know you are right.
- Time is short and a quick decision is to be made.
- A strong personality is trying to steamroll us and we don't want to be taken advantage of.
- We need to stand up for our rights.

This technique can further escalate conflict or losers may retaliate.
Example − When in a debate the party with more facts wins.

Avoiding
This technique follows the rule "No winners, no losers". Avoiding means the ideas suggested by both the parties are rejected and a third person is involved who takes a decision without favoring any of the parties. This technique can be used when −

- The conflict is small and relationships are at stake
- We are counting to ten to cool off
- More important issues are pressing and we feel we don't have time to deal with this particular one
- We have no power and we see no chance of getting our concerns met
- We are too emotionally involved and others around us can solve the conflict more successfully

Using this technique may lead to postponing the conflict, that may make matters worse.
Example − Rahul and Rohit had a fight, their mother came and punished both of them.

Organizational Behavior - Culture
Organizational culture can be defined as the group norms, values, beliefs and assumptions practiced in an organization. It brings stability and control within the firm. The organization is more stable and its objective can be understood more clearly.
Organizational culture helps the group members to resolve their differences, overcome the barriers and also helps them in tackling risks.
Elements of Organizational Culture
The two key elements seen in organizational culture are −

- **Visible elements** − These elements are seen by the outer world. Example, dress code, activities, setup, etc.

- **Invisible elements** − These inner elements of the group cannot be seen by people outside the group or firm. Example, values, norms, assumptions, etc. Now let us discuss some other elements of organizational culture. They are −
- **Stories** − Stories regarding the history of the firm, or founder.
- **Rituals** − Precise practices an organization follows as a habit.
- **Symbol** − The logo or signature or the style statement of a company.
- **Language** − A common language that can be followed by all, like English.
- **Practice** − Discipline, daily routine or say the tight schedule everyone follows without any failure.
- **Values and Norms** − The idea over which a company is based or the thought of the firm is considered as its value and the condition to adopt them are called norms.
- **Assumptions** − It means we consider something to be true without any facts. Assumptions can be used as the standard of working, means the employees prepare themselves to remain above standard.

Different Types of Organizational Culture

The culture a firm follows can be further classified into different types. They are −

- Mechanistic and Organic culture
- Authoritarian and Participative culture
- Subculture and Dominant culture
- Strong and Weak culture
- Entrepreneurial and Market culture

Mechanistic and Organic Culture

Mechanistic culture is formed by formal rule and standard operating procedures. Everything needs to be defined clearly to the employees like their task, responsibility and concerned authorities. Communication process is carried according to the direction given by the organization. Accountability is one of the key factors of mechanistic culture.

Organic culture is defined as the essence of social values in an organization. Thus there exists a high degree of sociability with very few formal rules and regulations in the company. It has a systematic hierarchy of authority that leads towards free flow of communication. Some key elements of organic culture include authority, responsibility, accountability and direct flow towards the employee.

Authoritarian and Participative Culture

Authoritarian culture means power of one. In this culture, power remains with the top level management. All the decisions are made by the top management with no employee involvement in the decision making as well as goal shaping process. The authority demands obedience from the employee and warns them for punishment in case of mistake or irregularity. This type of culture is followed by military organization.

In participative culture, employees actively participate in the decision making and goal shaping process. As the name suggests, it believes in collaborative decision making. In this type of culture, employees are perfectionist, active and professional. Along with group decision making, group problem solving process is also seen here.

Subculture and Dominant Culture

In subculture, some members of the organization make and follow a culture but not all members. It is a part of organizational culture, thus we can see many subcultures in an organization. Every department in a company have their own culture that gets converted to a subculture. So, the strength and adaptability of an organizational culture is dependent on the success of subculture.

In dominant culture, majority of subculture combine to become a dominant culture. The success of dominant culture is dependent on the homogeneity of the subculture, that is, the mixture of different cultures. At the same point of time, some cold war between a dominant culture and a minor culture can also be seen.

Strong and Weak Culture

In a strong culture, the employees are loyal and have a feeling of belongingness towards the organization. They are proud of their company as well as of the work they do and they slave towards their goal with proper coordination and control. Perception and commitment are two aspects that are seen within the employees. In this culture, there is less employee turnover and high productivity.
In a weak culture, the employees hardly praise their organization. There is no loyalty towards the company. Thus, employee dissatisfaction and high labor turnover are two aspects of this culture.

Entrepreneurial and Market Culture
Entrepreneurial culture is a flexible and risk-taking culture. Here the employees show their innovativeness in thinking and are experimental in practice. Individual initiations make the goal easy to achieve. Employees are given freedom in their activity. The organization rewards the employees for better performance.
Market culture is based on achievement of goal. It is a highly target-oriented and completely profit-oriented culture. Here the relationship between the employees and the organization is to achieve the goal. The social relation among the workers is not motivating.

How to Create an Organizational Culture
An organizational culture is created with the combination of certain criteria that are mentioned below −

- The founder of the organization may partly set a culture.
- The environment within which the organization standards may influence its activities to set a culture.
- Sometimes interchange of culture in between different organizations create different new cultures.
- The members of the organization may set a culture that is flexible to adapt.
- New cultures are also created in an organization due to demand of time and situation.

The culture of an organizational can change due to composition of workforce, merger and acquisition, planned organizational change, and influence of other organizational culture.

Organizational Behavior - Change
Organizational change can be defined as the alteration in structure, technology or people in an organization or behavior by an organization. Here we need to note that change in organizational culture is different from change in an organization. A new method or style or new rule is implemented here.
An organizational change occurs due to two major factors namely −

- **External factor** − External factors are those factors that are present outside the firm but force the firm to change or implement a new law, rule etc. For example, all banks are bound to follow the rules laid down by the RBI.
- **Internal factor** − Internal factors are those factors that are caused or introduced inside an organization that forces a change. For example, no smoking in the workplace.

Kurt Lewin’s Force Field Analysis
Kurt Lewin, is a noted organizational theorist, who proposed the force field analysis for organizational change. In this theory, he has prioritized two factors for change in an organization, namely −

- **Driving force** − Driving force can be defined as an organizational force that makes a change with respect to structure, people and technology. In short, it drives the organization from one culture to another.
- **Restoring force** − Restoring force is the force which changes the culture from the existing state to the old state. It indicates a backward motion while the driving force indicates a forward motion.

Importance of Organizational Change
There is a need of change in an organization because there is always a hope for further development, and in order to survive in a competitive market, the organization needs to be updated with changes. However, we have listed some reasons to explain why changes are deliberately made and carefully planned by the organization before implementation.

- It improves the means to satisfy the economic requirements of people.

- It enhances the profitability of organization.
- It promotes employee satisfaction and well-being.

Planned Change

We can define planned change as any kind of alteration or modification which is done in advance and differently for improvement.

The Need for Planned Change

Planned change takes places in an organization when there is a demand for change due to two types of forces. These forces are grouped into internal sources and external sources.

Internal forces that lead to a planned change in an organization include obsolescence of production and service, new market opportunities, new strategic direction, increasing workforce diversity, and shift in socio-cultural values.

External forces that lead to a planned change in an organization include regulators, competitors, market force, customers, and technology. Each of these forces can create pressing demand for change in small or big, public or private, business or non-business organizations.

Process of Planned Change

Once the management decides to implement some changes in the organization, it needs to be done carefully as it is a very sensitive issue. It is very important for all the employees to adapt to change. According to Kurt Lewin, the planned organizational change is implemented in three different stages. They are −

- **Unfreezing** − In this stage, the organization studies if the change is required or not, what and why is the change necessary. Considering the entire situation, the organization decides for appropriate change. Thus a plan and strategy is formulated as required.
- **Changing** − In this stage, the organization executes the plan and program for change. For this purpose, proper precautions are taken in order to maintain cooperation and coordination between the employees and management, avoiding miscommunication or disputes. Adequate supervision and control is arranged as needed.
- **Refreezing** − This is the final stage, in order to bring organizational change. By way of supervision, the organization tries to evaluate the effectiveness of change. Collecting all this information, the management interprets whether to continue or replace change by some other alternatives or to make further minor changes.

Types of Planned Change

On the basis of a company's requirement planned change is classified into three types. They are −

- Change in structure
- Change in technology
- Change in people

Change in Structure

We say that the planned change required is change in structure when development is required in these following areas −

- Change in management
- New management
- Change in position or location
- Change in objective, rules, regulations etc.
- Launching new branches

Change in Technology

We say that the planned change required is change in technology when development is required in these following areas −

- Need of office automation
- Installing new hardware and software
- Executing new working procedures
- New methods in production function

- Producing new products and devices
- New training, research and development program

Change in People

We say that the planned change required is change in people when development is required in these following areas −

- New candidate requirement
- Promotion or demotion
- Transfer to other location
- Suspension or dismissal
- Deputation
- Training and development

Organizational Behavior - Development

Organizational Development is a field of research, theory, as well as practice devoted to expanding the knowledge and effectiveness of how people accomplish successful organizational change and performance.

Organizational development is not an overnight transformation that can be done in an organization, rather it is a gradual process that has to be executed systematically and by taking care of the external environment.

Organization Development Techniques

Companies adopt organizational development technique to modify the behavior of people who are resisting change. It is a program to bring a change in the values, norms, attitudes, perception, and behavior of people and improve the quality of inter-personal relations. Some of the major organizational development techniques are −

- Sensitivity technique
- Survey feedback
- Process consultation
- Team building
- Intergroup development

Sensitivity Technique

Here sensitivity refers to the psychological aspect of human mind that has to be shaped to act as expected by the group. In this technique, one's own weakness is exposed and members understand how others react towards them. Stress is on group dynamics and tackling inter-relationship disputes.

The idea is to improve the behavior of people in order to maintain smooth inter-personal relationship without any power or influence. Members are motivated to have an open, heart-to-heart talk to develop mature relationship. Sensitivity training borders on psychotherapy where the emotions as well as body language are considered.

Survey Feedback

In this technique, the discrepancies among a group are weeded out using questionnaires, which identify the difference in perception amongst the same working family, group or department. The collected data is then tabulated and distributed for further deliberations. This acts as the basis for further discussions and the discrepancies if any can be sorted out by open discussions with all concerned, defending and opposing till a consensus is reached. This technique mainly focuses on ideas and not on persons who put up those ideas.

Process Consultation

In this technique, a firm may either seek the support of experts from within the firm or from outside. The firm must check that process consultation is done through an external expert with the needed support provided by the authorities from within the organization.

Team Building

In this technique, attempts are made at the group or inter-group level. The main objective is to improve co-ordination thereby improving the performance as a group. This can be done by goal setting, development of inter-personal relations, role analysis to identify roles and responsibilities and team process analysis.

Intergroup Development

Inter group development technique attempts to change the perceptions of groups about themselves or about other groups. This can be done by organizing independent group meeting, developing a list consisting of perception of itself, views about other departments and how others view them, trying to understand and resolve the actual cause of conflicts, or sub grouping the groups to remove difference in perceptions and impressions that groups have about each other.

Service-level agreements

A service-level agreement (SLA) defines the level of service you expect from a vendor, laying out the metrics by which service is measured, as well as remedies or penalties should agreed-on service levels not be achieved. It is a critical component of any technology vendor contract.

Service-level agreements, or SLAs, are business contracts between product or service providers and customers. SLAs document specific services to be provided, metrics that measure service performance and remedies for nonperformance.

SLAs are not limited to vendors and customers. In fact, they may be applied to any arrangement in which something -- such as an activity or process -- is to be performed by one party for another party to an agreed-upon level of performance.

For example, it is not uncommon for IT departments to establish SLAs with business units regarding the delivery of specific actions by IT. Most departments can execute an SLA with another department or departments for the provision or completion of a specific activity. In practice, SLAs with vendors are more common and are often the subject of discussion and dissention between participants based on satisfaction of an SLA's parameters.

Types of SLAs

IT product or service providers, cloud service providers, network service carriers and corporate IT organizations create SLAs with internal customers to establish expectations and evaluate vendor performance. SLAs come in three different forms:

1. **Service-based SLAs.** This form of SLA is for a service; it sets the same parameters for all customers using that service.
2. **Customer-based SLAs.** This SLA is based on an agreement between a provider and an individual customer and covers all services being provided to that customer.
3. **Multilevel SLAs.** This form of SLA focuses at the corporate level and is applied to all users in an organization. Multilevel SLAs are used to avoid duplicate or conflicting agreements across the organization.

Components of SLAs

SLAs typically include all or most of the following components; compliance with each of these components is essential:

- description of services to be provided;
- scope of the services;
- location(s) where services are to be provided;
- responsibilities and duties of the service provider;
- responsibilities and duties of the service recipient;
- description of acceptable performance levels;
- metrics to be used to evaluate performance;
- processes to monitor, track and evaluate performance;
- processes to resolve poor performance;
- remedies for failure to provide acceptable performance, time frames and escalation procedures;

- protection of intellectual property, as applicable;
- compliance with legislation, standards, regulation and acceptable practices; and
- termination of the agreement.

From a compliance perspective, parties to the SLA must agree on what is to be provided, the metrics to be satisfied, the method of monitoring and reporting service delivery, and remedies for failure to satisfy SLA requirements.
Each of the above components can be treated as an SLA compliance requirement. Users must decide which aspects of service provisioning are the most important from a business perspective, and these can be specified in compliance documents.

Team management

Team management is a manager or organization's ability to lead a group of people in accomplishing a task or common goal. Effective team management involves supporting, communicating with and uplifting team members so they perform to the best of their abilities and continue to grow as professionals.

Team management is important for a number of reasons within the workplace:
It promotes a unified approach to leadership within a company or team, especially when team building is implemented.
It makes it easier to solve problems through the implementation of negotiating and critical thinking.
It encourages open communication between managers and team members and emphasizes good communication skills and active listening.
It ensures managers and team members are working toward a common goal that has been clearly defined.
It helps managers clearly outline the roles and expectations for their team members.

Team Leader should practice the following:
Focus on serving rather than managing
Don't always assume you're right
Make transparency a priority
Set boundaries
Provide a positive workspace
Emphasize constant and effective communication within the workplace
Encourage and nurture your team's growth
Be open to change

Many common soft skills make individuals great team players. While soft skills are not as quickly learned as technical skills, they can be developed with time and practice.
Here are several qualities you can focus on to be a better team player:
1. You understand your role
As a team member, you understand your role within the team and work to achieve your duties to the best of your ability. Though you may offer help or solutions to other team members, you also should respect the boundaries of your position.
2. You welcome collaboration
Working with a team means there will be varying opinions and ideas. Even if you think your idea is best, you should listen to all ideas before pushing yours. Search for compromises, and remain respectful if your work is criticized.
3. You hold yourself accountable
Take responsibility for your mistakes and look for solutions. Understand how your actions impact the entire group. In doing so, you will learn from your errors and command more respect from your team.
4. You are flexible
You should readily accept any tasks your manager gives you. Flexibility in your role allows you to learn more and help your team. Look at every opportunity as a chance to learn.
5. You have a positive attitude

Maintaining a positive attitude even during stressful times helps the rest of your teamwork through that difficult time without getting upset. Your positive attitude will create a better atmosphere.
6. You commit to the team
You should be fully invested in the team. You will be a great team player if you show others that you believe in the group, the process, and the goals. This sort of positivity can radically increase morale and productivity.

How to be a better team player
Working well with others shows that you are committed to achieving both personal and organizational goals. Displaying consistent teamwork skills also indicates a strong work ethic, increasing your chances for raises, promotions, and other earnings. No matter your experience level or position, continuously focusing on becoming a better team player will succeed in your career.

Here are several ways you can focus on improving your teamwork skills:
1. Offer help. If you see a coworker who seems overwhelmed or is struggling to keep up with tasks, ask if you can help. Team players support each other during difficult times. Remember to ask for help, as well.
2. Actively listen. Active listening means hearing and thoughtfully responding to what your team member says. Ask questions about things you don't understand.
3. Communicate. Keep your team updated on your progress and what you need to be successful in your job. You should constantly communicate with your team to ensure that everyone is working toward the same goal and no one is repeating work.
4. Respect others. Recognize that other team members are also trying to fulfill their roles, and consider how you can support them. Take the time to get to know your team. Everyone has a role to play that is no less critical than your own.
5. Be a problem-solver. When you recognize a problem, could you take action to solve it? Brainstorm solutions to your concerns and ask for feedback.
6. Celebrate teammates' successes. If a member of your team succeeds in the workplace, so do you. It means you are one step closer to completing a goal. Celebrate their success. Also, stay updated on their personal lives and take the time to express interest and care.

If you're unsure about what areas you need to improve to be a better team player, ask a trusted friend or colleague for honest feedback about your teamwork skills. You might also consider asking someone you respect in your industry to be your mentor. Seeking out someone who has strong teamwork skills can help you improve your own.

Collaboration is a crucial part of working successfully, and learning how to be a positive force for your team is vital. When you aim to be a great team player, others will follow. In doing so, you can improve your workplace, grow personally and advance in your career.

Stakeholder management

Stakeholder management is the process of maintaining good relationships with the people who have most impact on your work. Communicating with each one in the right way can play a vital part in keeping them "on board."

1. Identify stakeholders
Before you can manage your stakeholders' expectations, you need to know who your stakeholders are.
Your stakeholders include anyone who may be impacted by or have an impact on your project or organization. This includes people inside your organization (employees, managers, executives, and shareholders) and external people, organizations, and community groups (customers, creditors, suppliers, resellers, schools, media, regulatory authorities, non-government organizations, community groups, and locals).

2. Manage stakeholders differently
Most projects have a range of different stakeholders. As such, they'll likely have a broad range of expectations.

You'll need to approach each group differently, providing information that's relevant to them. This is where stakeholder mapping can be invaluable. Understand your stakeholders in relation to three key criteria: Influence, Impact, Interest

3. Keep stakeholders in the loop
Provide regular updates to your stakeholders. Start by explaining the process, the timing and exactly what scope there is for stakeholders to influence decisions. Making sure that everyone understands the process and the role they can play in the process is a critical first step.

Put a plan or roadmap in place and share it. If your project is ongoing, your stakeholders will likely see it go through several stages and transitions. If you can clearly communicate this with a written plan or roadmap, they'll understand what's happening and when.
Encourage two-way communication. Whenever you communicate with stakeholders, make sure you're also ready to listen and receive feedback. And be ready to direct stakeholders towards factors they are able to influence so they can participate in the process.

4. Share evidence and research
Your stakeholders are likely informed by a range of perspectives and information sources outside of your control. So don't assume they have all the facts, and don't assume you have all the facts, either.
So, when you communicate, you may need to dispel myths by clearly stating the facts from a reliable source. And also be willing to listen to their perspectives – many stakeholders will make valid points you should take into consideration. Be prepared to adjust your plans, if needed.

5. Produce value regularly
Demonstrate that you're making progress or moving towards the overall outcome and communicate this regularly. Depending on the project timeline and stakeholders, you may need to provide progress updates anywhere from every 2 weeks to 2 months. Some ways you can update your stakeholders include:
Sending an email
Calling them on the phone
Hosting a community meeting
Posting an update on your website
Sharing social media posts

6. Track sentiment
Stakeholder sentiment is about how your stakeholders feel towards the project and any current issues that are coming up. When you know your stakeholder sentiment, you can understand your stakeholders better and see how to guide each group to appropriate expectations. One way to track sentiment is using your stakeholder management system. Darzin allows you to analyze your stakeholder communication and spot emerging issues by producing tag clouds. From there, you can adjust your message and/or project to better align stakeholder expectations with reality.

Time Management

Time management is the act or process of planning and exercising conscious control over the amount of time spent on specific activities, especially to increase effectiveness, efficiency or productivity.
It is an activity with the goal to maximize the overall benefit of a set of other activities within the boundary condition of a limited amount of time.

Time management may be aided by a range of skills, tools, and techniques used to manage time when accomplishing specific tasks, projects, and goals complying with a due date. Initially, time management referred to just business or work activities, but eventually the term broadened to include personal activities as well. A time management system is a designed combination of processes, tools, techniques, and methods. Time management is usually a necessity in any project development as it determines the project completion time and scope.

Time management has been considered to be a subset of different concepts such as:

- Project management: Time Management can be considered to be a project management subset and is more commonly known as project planning and project scheduling. Time Management has also been identified as one of the core functions identified in project management.
- Attention management: Attention Management relates to the management of cognitive resources, and in particular the time that humans allocate their mind (and organize the minds of their employees) to conduct some activities.

The major themes arising on time management include the following:

- Creating an environment conducive to effectiveness
- Setting of priorities
- Carrying out activity around those priorities
- The related process of reduction of time spent on non-priorities
- Incentives to modify behavior to ensure compliance with time-related deadlines.

It is rightly said "Time and Tide wait for none". An individual should understand the value of time for him to succeed in all aspects of life. People who waste time are the ones who fail to create an identity of their own.

What is Time Management?

- Time Management refers to managing time effectively so that the right time is allocated to the right activity.
- Effective time management allows individuals to assign specific time slots to activities as per their importance.
- Time Management refers to making the best use of time as time is always limited.

Time Management includes:

- **Effective Planning**

Plan your day well in advance. Prepare a To Do List or a "TASK PLAN". Jot down the important activities that need to be done in a single day against the time that should be allocated to each activity. High Priority work should come on top followed by those which do not need much of your importance at the moment. Complete pending tasks one by one. Do not begin fresh work unless you have finished your previous task. Tick the ones you have already completed. Ensure you finish the tasks within the stipulated time frame.

- **Setting Goals and Objectives**

Working without goals and targets in an organization would be similar to a situation where the captain of the ship loses his way in the sea. Yes, you would be lost. Set targets for yourself and make sure they are realistic ones and achievable.

- **Setting Deadlines**

Set deadlines for yourself and strive hard to complete tasks ahead of the deadlines. Do not wait for your superiors to ask you everytime. Learn to take ownership of work. One person who can best set the deadlines is you yourself. Ask yourself how much time needs to be devoted to a particular task and for how many days. Use a planner to mark the important dates against the set deadlines.

- **Delegation of Responsibilities**

Learn to say "NO" at workplace. Don't do everything on your own. There are other people as well. One should not accept something which he knows is difficult for him. The roles and responsibilities must be delegated as per interest and specialization of employees for them to finish tasks within deadlines. A person who does not have knowledge about something needs more time than someone who knows the work well.

- **Prioritizing Tasks**

Prioritize the tasks as per their importance and urgency. Know the difference between important and urgent work. Identify which tasks should be done within a day, which all should be done within a month and so on. Tasks which are most important should be done earlier.

- **Spending the right time on right activity**

Develop the habit of doing the right thing at the right time. Work done at the wrong time is not of much use. Don't waste a complete day on something which can be done in an hour or so. Also keep some time separate for your personal calls or checking updates on Facebook or Twitter. After all human being is not a machine.

For Effective Time Management one needs to be:
Organized - Avoid keeping stacks of file and heaps of paper at your workstation. Throw what all you don't need. Put important documents in folders. Keep the files in their respective drawers with labels on top of each file. It saves time which goes on unnecessary searching.
Don't misuse time - Do not kill time by loitering or gossiping around. Concentrate on your work and finish assignments on time. Remember your organization is not paying you for playing games on computer or peeping into other's cubicles. First complete your work and then do whatever you feel like doing. Don't wait till the last moment.
Be Focussed - One needs to be focused for effective time management.
Develop the habit of using planners, organizers, table top calendars for better time management. Set reminders on phones or your personal computers.

Time Management refers to making the best possible use of available time.
Managing time well enables an individual to do the right thing at the right time.
Time Management plays a pivotal role in one's personal as well as professional life.

Benefits of Time Management:

- **Time Management makes an individual punctual and disciplined**. One learns to work when it is actually required as a result of effective time management. To make the judicious use of time, individuals should prepare a "TASK PLAN" or a "TO DO" List at the start of the day to jot down activities which need to be done in a particular day as per their importance and urgency against the specific time slots assigned to each activity. A Task Plan gives individuals a sense of direction at the workplace. An individual knows how his day looks like and eventually works accordingly leading to an increased output.
- **One becomes more organized as a result of effective Time Management**. Keeping the things at their proper places minimizes the time which goes on unnecessary searching of documents, important files, folders, stationery items and so on. For better time management, individuals keep their workstations, study zones, cubicles, meeting areas clean and organized. People learn to manage things well as a result of Time Management.
- **Effective Time Management boosts an individual's morale and makes him confident**. As a result of Time Management, individuals accomplish tasks within the stipulated time frame, making them popular in their organization as well as amongst their peers. People who understand the value of time are the ones who manage to stand apart from the crowd. Individuals who finish off work on time are looked up to by others and are always the centre of attention everywhere.
- **Individuals who stick to a time plan are the ones who realize their goals and objectives within the shortest possible time span**. Managing time effectively helps employees to meet targets way ahead of deadlines and finish off task just when it is required.
- Effective Time Management helps an employee to reach the pinnacle of success quickly and stay firm at the top for a longer duration. An employee who works just for the sake of working fails to create an impression and is never taken seriously at work. Effective time management plays a pivotal role in increasing an individual's productivity. Output increases substantially when people manage their time well.
- **Better Time Management helps in better planning and eventually better forecasting**. Individuals learn to plan things well and know where exactly they stand five years from now.
- **Research says that individuals who accomplish tasks on time are less prone to stress and anxiety**. Remember there is no point in wasting time and cribbing later. Finish off pending work on time and then you would have ample time for your friends, relatives and family members.
- **Time Management enables an individual to prioritize tasks and activities at workplace**. It is foolish to stay overburdened. Do not accept anything and everything that comes your way.

- Time Management helps an individual to adopt a planned approach in life.

Time Management refers to assigning specific time slots to activities as per their importance and urgency in order to make the best possible use of time.
In a layman's language Time Management is nothing but to manage time well and doing things when they actually need to be done.
Every organization works on deadlines. An individual constantly needs to be on his toes to finish off assignments within stipulated time frame. It is essential for employees to understand the value of time for them to do well and make a mark of their own at the workplace.

How to practice effective time management in organization?

- **Know your targets well**. Do not hesitate to speak to your Boss if targets are unrealistic and unachievable within the allocated time slot. It is always better to discuss things at the initial stages than cutting a sorry figure later on. Accept tasks only when you are confident.
- **There is absolutely no harm in discussing work with your fellow workers**. You can't do almost everything on your own. Distribute work amongst your team members. It is foolish to over burden yourself. One must share his work load with others to finish assignments within the stipulated time frame. Know your capabilities.
- **Organize yourself**. Be very careful about your files, important documents, visiting cards, folders etc.Keep them at their proper places so that you do not waste half of your time in searching them.
- **Be loyal to your organization**. Do not work only when your superiors are around. Remember you are getting paid for your hard work. Concentrate on your own work rather than loitering and gossiping around. Do not waste time by playing games on computer or finding out what your fellow worker is up to.
- It is absolutely okay to call up family members or friends once in a while but make sure you do not end up in long phone calls while at work. Phone calls and messages are one of the biggest distractions at work.
- **Plan your things well in advance**. Do not work just for the sake of working. The first thing an employee should do in the morning is to jot down what all tasks he need to do in a single day against the time slot assigned to each task. Preparing a Task Plan right at the start of the day always helps and provides you a sense of direction at work. A "TO DO" List suggests you way forward. Tick off completed assignments. Make sure you finish tasks within the assigned deadlines.
- **Keep a notepad and pen handy**. Avoid writing on loose papers. You will never find them when you actually need something. Prefer using an organizer as it helps you plan your work better.
- **Eat only during lunch hours**. Eating while working not only makes you feel sleepy but also breaks continuity.
- **Be punctual**. Avoid taking frequent leaves from work unless it is an emergency. Make it a habit to reach office on time.
- **Do not keep things pending at your end**. Escalate matters immediately which need approval of higher authorities. Do not keep ignoring things. They would create problems for you sooner or later.

Types of Time Management Skills
Prioritizing
It might be impossible to do every single minute task expected of you. You also might want to do everything all at once. But you must prioritize so that you are able to complete the most important tasks in an order that makes sense. When assigning priority, consider such factors as when each task needs to be done, how long it might take, how important it might be to others in the organization, what could happen if a task is not done, and whether any task might be interrupted by bottlenecks in the process.

- Allocation
- Managing Expectations
- Waste Prevention
- Prioritizing requests and demands
- High-Value Activities (HVAs)

- Performance Reviews
- Goal Setting

Scheduling

Scheduling is important because some tasks have to be done at specific times. Scheduling affects your day, your week, your month, as well as other people's workflow. Most have specific times of the day when they are more or less productive as a result of energy levels and demands of the day. Schedules can be a good way to avoid procrastination, too.

- Scheduling Software
- Intentionality
- Punctuality
- Breaking broader goals into milestones
- Breaking up milestones into projects

Task Management

To-do lists (properly prioritized and integrated with your schedule) are a great way to avoid forgetting something important. They are also a great way to avoid spending all day thinking about everything you have to do. Remembering tasks takes energy and thinking about everything you have to do all week can be exhausting and overwhelming. Split all the necessary tasks up into a list for each day, and you won't have to worry about all of it all at once. Just take your tasks one day at a time.

- Proactive
- Batching
- Creating daily, weekly, and monthly to do lists
- Multitasking
- Thoroughness
- Organization
- Email Management

Workload Management

Pacing your work, even though it may seem an odd thing to call a skill, is an important time management concept. Although working long hours or skipping breaks can sometimes improve productivity in the short term, your exhaustion later will ensure that your overall productivity actually drops. Except for rare emergencies, it is important to resist the temptation to over work. Include necessary breaks, and a sensible quitting time, in your schedule.

Knowing and enforcing an optimum workload for yourself ensures consistency in your performance and avoids burnout. Employers want to be able to count on you for the long-term.

- Process Management
- Assertiveness
- Eliminating Waste
- Taking Breaks

Delegation

Depending on what type of work you do, you may be able to delegate some tasks. Knowing what and when to delegate is an important skill. Some people resist delegating, either because they want to maintain control or because they want to save money by not hiring assistants. Both approaches ultimately hurt productivity and raise costs.

Remember, however, that if you practice time management diligently and still can't get everything done, you may be trying to do too much. It is better to succeed at a few tasks than to attempt and fail at many.

- Seeking Expert Assistance
- Moderating Meetings
- Presentation
- Teamwork

- Leadership
- Collaboration
- Motivation

Timesheets are surprisingly important tools for project management in the following ways:

- By tracking time on a project, they help in qualifying an employee for a new project, role, or job based on performance.
- By analyzing time usage, you can determine the skill of someone in a particular area and utilize them in parts of a project for which they're better suited.
- They help in identifying areas in which a worker takes more time than allocated and needs additional training.
- They are helpful in measuring the overall efficiency of your project.
- They help in identifying the availability of workers assigned to a project, making it easy to distribute the workload among group members.
- They enable project managers to monitor progress and anticipate delays.
- They help project managers estimate expected costs and compare them to the real cost of the project during execution.

Project management

Project management skills are the competencies and traits a person needs in order to effectively coordinate a project from start to finish. A project manager leads a projects team using good communication, interpersonal skills, motivational skills, and organization. Project management is a simple term but it encompasses a wide range of skills and responsibilities. Project management includes planning projects, mapping out timelines, and executing each phase of the project.

What are project management skills?

Project management skills are a group of skills needed to initiate, plan, and execute a project. Project managers often have a team of people working on the project and everyone needs to work together to achieve a specific goal. Project management skills encompass a wide variety of hard skills, soft skills, and personality traits.

Examples of project management skills

A project manager is responsible for initiating and executing a project. They are also responsible for managing and motivating a team, managing the expectations of the stakeholders, and communicating with company executives and the stakeholders about the status of the project.

7 Essential Project Management Skills

1. Effective Communication

From project kick-offs to stakeholder meetings, project managers are constantly communicating. In fact, according to Cesar Abeid, host of the Project Management for the Masses Podcast, project managers spend a whopping 90 percent of their time communicating in some way. As such, project managers must have excellent communication skills in order to successfully lead projects to completion.

Poor communication, on the other hand, can lead to inefficiency and missed deadlines; according to a recent survey, 28 percent of employees report poor communication as the primary cause of failing to deliver a project on time.

In order to mitigate the risk of this, project managers should prioritize learning how to communicate effectively. This includes knowing how to approach people, create meaningful relationships with co-workers, and articulate a clearly established vision of what you wish to achieve. Taking the time to think about what you want to say—and how you want to say it—prior to communicating is a simple and productive first step toward honing this skill set.

Did You Know: Most companies regard communication skills as extremely important, regardless of department or industry. In fact, 77 percent of employers say that “soft skills” like this are just as important as technical skills.

2. Negotiation

From managing resources, to engaging suppliers, to dealing with team conflict, leading a project means constantly being involved in negotiations. An effective project manager is often a skillful negotiator with the ability to keep involved parties content and working toward a unified goal at all times.

Unavoidable discussions about budgets, scope creep, resources, and timelines can easily become adversarial if not handled tactfully, and savvy project managers instinctively know when—and how—to apply persuasive techniques that will encourage solutions and avoid damaging workplace relationships.

Depending on the situation at hand, project managers may choose to apply different negotiation styles, and it's important they have the ability to choose which will be the most effective in each given scenario. Where some may call for compromise (in which an agreement with the opposing party is reached), for example, others may call for collaboration (in which a win-win solution is reached), or competition (in which a project is controlling in a win-lose situation). No matter which style they choose, knowing how to effectively navigate these negotiation tactics is a vital tool for successful project managers.

3. Scheduling and Time Management

Naturally, scheduling is a core facet of the project management function. However, if you're adept at juggling multiple schedules and can anticipate roadblocks before they occur, you'll increase your chances of delivering successful projects.

The most common causes of project failure involve poor planning, and include changing priorities within an organization (40 percent), inaccurate requirements (38 percent), changes in project objectives (35 percent), and undefined project goals (30 percent). As this data shows, time management is clearly a valuable project management skill to attain.

The first step to effective time management for project managers starts with creating a well-defined project plan.

At the fundamental level, a project plan captures business requirements and project scope, while sequencing activities and resources—all of which will go a long way in saving time and money.

4. Leadership

When managing a team or project, it's crucial to have strong leadership skills. By effectively coaching, guiding, and motivating your co-workers, you can help move a project forward and deliver a positive outcome. Strong leaders also foster a productive work environment by communicating regularly with their teams and helping their employees develop important project management skills themselves.

Teams also work more efficiently when employees feel they are making a meaningful contribution to the project. You can help team members feel impactful by learning how to delegate tasks, provide constructive feedback, set goals, and evaluate individual and team performance. Be sure to recognize the achievements of your team members so they know you value their contributions, as well. Combining these elements with your own unique leadership style will help you to successfully manage projects while also becoming a more effective leader.

5. Technical Expertise

Successful project managers must not only possess the aforementioned soft skills but also have the technical know-how to move projects toward completion. PMs who can "speak the language" of their organization's subject matter experts will be able to communicate more effectively with their teams and have a better understanding of a project's inherent risks and potential roadblocks.

Additionally, a practical understanding of popular project management software can have a positive impact on a project manager's work. These types of programs are frequently used to plan, organize, and communicate with teammates while simultaneously managing resources, budgets, and schedules. In today's world, this means project managers must also constantly evolve, learning how to leverage the newest technologies available to successfully lead a project to completion.

Tools that may enhance a project manager's ability to communicate effectively with stakeholders range from shared drives (e.g., SharePoint, Microsoft Teams, Google Docs) to more technical programs. Microsoft Project and Jira, for example, are popular in IT settings, while Primavera is commonly used for scheduling in the government and construction industries.

6. Risk Management

Risk management, which involves identifying and planning for potential risks, is a critical competency for project managers. The most productive PMs are able to identify risks at the outset of a project and develop proper mitigation plans in the event that the risk does, in fact, occur.

Since risks are, by definition, uncertain events, it’s easy for project managers to ignore them and assume they will not happen. But by knowing—and acknowledging—what positive or negative impacts may happen to the project if the risk does occur, project managers can plan ahead to avoid major problems that might otherwise derail a project.

7. Critical Thinking and Problem Solving

While critical thinking is a skill all professionals could benefit from learning, it is particularly useful in the project management discipline. Rather than being reactive, the best project managers are proactive and use their critical thinking skills to navigate through tricky or ambiguous projects.

By remaining objective, analyzing the facts, and evaluating options without bias, project managers are able to solve complex problems for organizations while delivering results on time and within budget.

How to Develop Key Project Management Skills

1. Practice, Practice, Practice: Like any new ability, these skills take time to learn. If you’re currently in a project management role, be cognizant of utilizing these “tools” in your day-to-day work. If you’ve yet to break into the project management field, seek out opportunities for hands-on learning that will allow you to harness and refine them.
2. Attend Industry Events and Workshops: Attending local and online events will allow you to learn best practices and stay abreast of the latest project management trends.
3. Earn a Degree or Certificate in Project Management: One of the best ways to refine your project management skills is to earn a graduate certificate or degree. Programs like Northeastern’s Master of Science in Project Management provide students with the practical skills and technical expertise needed to lead complex projects to completion. Not only will you master the above skills (and more), but you’ll also have the chance to obtain hands-on experience through an experiential learning program.

Attention Management

Attention Management can be defined as a set of practices that increase people’s ability to concentrate on important things at work and in life. It also helps to arrange their priorities, so that they can deliver efficiently at organizational as well as personal levels.

Managers all over the world will tell you that a small but attentive workforce is much more productive than a large distracted workforce. An inattentive employee will make careless mistakes and end up being less productive. People from all fields have to be attentive to excel or at least retain their position.

Distractions include gossiping with co-workers, taking long breaks, handling personal matters during work hours, etc. These distractions generally take place due to boredom, working for long hours and mundane working conditions. One could prevent wasting time by focusing on specific goals.

Attention deficit can hamper the chances of a sincere employee too. If his teammates don’t pay attention to what he says, his work is less than effective. The ability of staying attentive helps professionals relate with their work on an emotional level. It also encourages them towards greater productivity to meet the company’s as well as their personal goals.

With increasing workload, people are easily stressed out and are easily distracted from work. Attention Management allows the managers and employees to focus more on their work with minimal distractions. This helps them to perform better and this in-turn leads them into getting a hike and a promotion.

We are not very far away from a future, where things like integrity, sincerity, etc., will make way for Attention as the most valuable commodity. On the other hand, work-related skills will make way for Focus as the most important skill in a workplace. To be relevant and successful in that future, we have to learn how to be focused and attentive from today itself.

What Causes Attention Deficit?

Social media has changed the way we people used to communicate, make new relationships, and develop them. It is also changing the way people used to handle multiple tasks at the same time. What was just a proposed theory years before, has now become an established fact that multitasking actually fails both the purposes, it was supposed to solve.

It is widely accepted now that multitasking actually takes more time to complete work and then, the quality is always poorer. In the age of technology and services, we have got used to instant gratification of curiosity and getting answers. In such situations, people have lost the patience to research for facts.

How Digital Media Changed Our Attention Spans?

Digital Media and technology have converged to produce more information now than ever. This means that people are encouraged to collect fleeting information, as compared to reading and analyzing the information. Then later, the facts are separated from fiction.

Attention has become a valuable asset today and the digital media uses it as currency. Gone are the days when online companies used to make money off direct sales. Nowadays, they have all entered into an indirect model of revenue generation. This is a model where they strive to become an established platform that is being visited by the internet-browsing crowd. It is all about **eyeballs, clicks, taglines, petitions**.

Attention Management - Deficit at Workplace

According to a poll released by Salary.com, 14% of employees of a company waste 3 hours daily, 22% waste 2 hours a day and 64% waste 1 hour a day. One of the biggest culprits was internet-surfing, which caused 48% of employees to waste their time on it.

A functional objective behind every company and subsequently every manager is to increase productivity. Every company wants the best from their workforce, but it would be overwhelming to use a small team to do multi-tasking and increase profits. This is where a manager comes into picture; a manager tries to extract more from less.

Managers need to be able to overcome their own distraction issues, both in their personal and professional lives, before they instruct their team-mates about it. A distracted manager cannot maintain an attentive team. So, the first thing one should consider is where they focus most of their attention.

CEOs Leads the Company

To determine the attention management of a company, one has to determine the attention of its CEO. A CEO should pay attention to the company and its employees to yield better productivity. It’s important that the CEO understands the different types of attention and various attention management techniques to encourage their employees.

Four Kinds of Attentive People

After studying the different attention spans of people from different fields and time management, experts have divided attention into four different areas. The skills of these people are taken from various spheres of life and age groups. These include −

- **Intentional** − People exhibiting intentional attention already know what they are doing and what they actually want to do. This is called working knowingly or “intentionally”. They plan strategically and note down the pros and cons of their decision and prioritize their activities accordingly.
- **Responsive** − People exhibiting responsive attention are not aware of the different situations they are going to face, but they respond according to the surroundings when they arrive at it. They don’t plan beforehand, but respond according to the current scenario. They even tend to spend lesser time in working intentionally.
- **Interrupted** − People exhibiting interrupted attention get easily distracted from their work. They don’t have a priority set to balance their life. In fact, they spend most of their time answering messages or handling situations, which are not at all related with their work or divert them from their work.
- **Unproductive** − People exhibiting unproductive attention are least bothered about their productivity. They spend most of their time letting their thoughts wander. They abuse office working time by taking breaks, chatting or any other unproductive works.

Attention Management - Types of Attention

Attention isn’t a constant phenomenon. A person can change his attention span and intensity of attention depending on different situations. There are three types of attention depending on the working type of an individual −

Focused Attention

Focused attention means “paying attention”. Under certain situations, people can channel their complete attention on a single task and everything else is considered as less important. This can be seen while studying for an exam or working on a project.

It is very difficult to sustain this type of attention for a very long time in normal cases, as it needs an engagement on a physiological level. Humans are bound to get tired from their work after certain period of time. For example, the same way a person is less likely to concentrate on studies, after studying vigorously for more than a couple of hours at a stretch.

Sustained Attention

Sustained Attention means concentrating on a certain time-consuming task. People employ such attention when they are watching a magic trick or watching an interesting movie. There are three stages of sustained attention −

- **Paying Attention** − Where you start focusing.
- **Keeping Attention** − Where you sustain your attention.
- **Ending attention** − When you finally stop paying attention.

This entire cycle is called one “Attention Span”. Once your attention ends, you will again need some time to focus again and remove the distractions. People get distracted from work which leaves the task incomplete, so one needs time to refocus and start after sometime.

Selective Attention

Selective Attention means focusing on a single stimulus in a complex setting. It’s like having a conversation in a crowded station, where it’s impossible to focus on everything going around. In this case, one has to really focus on the conversation and ignore everything else happening to get the message across.

One must have the ability to focus on a particular message or object by filtering all background noise. The negative part is that people tend to neglect what is going around (even if it’s important). In all this, the message they receive can easily be manipulated or misunderstood due to communication issues.

Alternating Attention

People exhibiting “alternating attention” have the ability of switching their attention to more than one task at the same time. Many of us have exhibited this attention by taking notes, while simultaneously listening and understanding a lecture. Here, the mind should be flexible and quick to understand and translate the information gathered.

Attentional Blink
According to a theory that was first proposed in the 90s, attention is just like vision. When we try to visualize two targets at the same time, one of them appears sharp while the other one gets blurred. Similarly, when people focus on two targets at the same time, they tend to miss the second one. When these targets are linked with strong emotions, it becomes easier to reduce attention blink.

Meditation Improves Attention
From the time you wake up till the time you sleep, your brain keeps sending and receiving messages through a network of nerve connections, which are called "brainwaves". Each brainwave is linked with different activities like sleep, attention, music, relaxation etc. These brainwaves are classified into Alpha, Beta, Theta, Delta and Gamma rays.

Visualization
In visualization techniques, a person is asked to use his imagination to create mental images that an instructor asks him to do. This increases brain activity and brings immense focus on work.
Every individual is different from the other, so they need different method or combination of methods to concentrate more. One has to understand the method best suited to him.

Activity to Increase Attentiveness
The following worksheet has a collection of activities that a person can engage in his daily life. These activities assure that he can manage to keep his attention sharp and focused. Most of these activities yield results on a long-term basis, so the practitioner needs to have patience while expecting results. However, he will start feeling an immediate change in his outlook in a week itself.

Work management and prioritization

Work management supports the commands and internal functions necessary to control system operation and the daily workload on the system. In addition, work management contains the functions that you need to distribute resources for your applications so that your system can handle your applications.
The purpose of your system is to perform work. Work management describes where work enters the system, where and with what resources work is processed, and where output from work goes.

Work Management functions
Work Management functions allow you to take any of the following courses of action on work lists
• View a summary of all work lists for your Organization.
• View a summary of all work lists proper to your user role and Organization Unit
• View a summary of all Immediate Action Work lists proper to your user role and Organization Unit.
• Select the work list that you want to view, or if you have update access to that list, on which you want to action the individual entries.
• View detailed statistics for the selected work list
• Select a revised presentation for the work list, which is in a different order, or filtered so that only those cases which have specific conditions present are shown
• Select a work list entry (work item) to view all available details on that work item

Six steps for expectation setting with the stakeholders
1. Describe the jobs in terms of major outcomes and link to the organization's need: The first step in expectation setting is to describe the job to the employees. They need to feel that there is a greater value to what they do. We need to specify that individual performance has an impact on the organization's mission.
2. Share expectations in terms of work style: it is not only important to talk about "what we do" but also on "how we expect to do it". The ground rules to be followed at the organization are
    • Always let the team know where the problems are

• Share concerns openly and look for solutions.
3. Maximize performance: identify what is required to complete the work. Set input as well as output expectations.
4. Establish priorities: use the time quadrant to establish priorities and also establish crisis plans.
5. Revalidate understanding: create documentation and communication plan to establish all discussions.
6. Establish progress check: schedule an early progress check to get things started the right way.

Work management

Organizations now have to constantly strive to achieve their annual goals and targets in the current competitive world. Fortunately, there's a way to do that without panicking and biting off your fingernails every other day.
And it starts with understanding work management.

Work management is a process that helps a business manages both its workflows and workloads simultaneously.
It's like a corporate superpower that can even leave the Avengers awestruck!

What Is Work Management?
Work management is a system that helps businesses create an efficient workflow structure that works for every level in the organization. It aims to streamline several business processes so that in-house and virtual teams can work efficiently to maximize profits.

And for this, you carefully assess current operational work to identify the shortcomings within the existing structure.
Work management also integrates procedures that help organizations align their work and execute it with perfection.
Ultimately, all of this contributes to *(drumroll, please…)*

- Boosting organizational performance
- Eliminating redundancies within the system
- Fulfilling client expectations *(and looking uber cool while doing it)*
- Utilizing resources efficiently
- Excellent team collaboration

Now, most people confuse work management with project management because both practices share some common ground.
However, they aren't the same.
So let's get our facts straight and quickly understand their differences.

Work management vs. project management

Sure, work management involves streamlining tasks, improving resource utilization, etc.
It's basically stuff that you perform with your go-to project management process too.

Comparing work management to project management	
Work Management	Project Management
Helps you organize and execute duties related to multiple projects and ad hoc tasks	Helps you plan and perform duties for an individual project
Lets you accomplish long-term goals focused on business efficiency and success	Lets you achieve short-term goals related to a project
It's a repetitive process with an evolving, flexible structure	It usually has a rigid structure with a set timescale and budget
Involves all the teams within an organization	Only involves the project team members associated with a particular project
Helps you achieve business goals	Helps you meet project expectations

The Work Management Process

The work management process consists of six steps that help you take a coherent and structured approach towards managing work:

- Work identification
- Work planning
- Work scheduling
- Work execution
- Work follow up
- Work analysis

The final objective of all these steps is to deliver quality deliverables that meet stakeholders‘ expectations and align with the approved budget.

1. Work identification

This step involves identifying the task at hand.

You define exactly what the work is, how to go about it, and when to complete it.

It’s also essential to document this information to ensure everyone’s on the same page.

2. Work planning

Work planning involves identifying the resources needed and creating time estimates to ensure that you get things done on time. Once you combine all the relevant data, you need to devise a plan of action.

One that is executable, unlike all your 2020 vacay plans…

Luckily, since work management apps have built-in Gantt charts, you won’t have to rely on spreadsheets for work planning!

3. Work scheduling

In this step, a team leader or project manager maps out the work across days, weeks, or even months.

And with these project management timelines in hand, you inform the team about their responsibilities in detail.

4. Work execution

With everything scheduled and mapped out, it’s time for some action.

Team members start working on the task by following the pre-defined plan of action discussed in the initial stages. You also prioritize any urgent work (and coffee refill breaks) to ensure that you meet your objectives on time.

5. Work follow-up

We all know how frustrating it can be when an important deadline is missed…

But plans don’t always go as planned, and that’s okay.

That’s why it’s essential to follow up on the work and check if all team members are working according to schedule.

If something doesn’t seem right, you can brainstorm together and create backup plans to help you inch closer to your deliverables.

6. Work analysis

What were the roadblocks?

What could you have done better?

In this stage, you refer to the collated data and documentation to assess the work quality. It’s useful for making a note of all the key takeaways to refine your processes.

Priority management

Priorities management is the practice of focusing time and resources towards work, projects, and tasks that impact high-value projects, accounts, and long-term goals.
Priority management is an essential part of time management and project management, where project managers adjust resources, schedules, and tasks to deliver projects on time and within scope.

Steps for effective priorities management

1. Understand top company objectives

For effective priorities management, it's essential to practice upward alignment before attempting downward alignment. You could be managing the most unified, productive team on the planet, but if the goals they're achieving aren't furthering the objectives most valued by stakeholders, what will that get you?
You don't want to be the team that spends time speeding a dozen new products to market when the executive team is more interested in reinvesting in the flagship product.
If you aren't clear about overall business goals, not to mention your boss's objectives for you and your team, it could be a failing on the part of upper management. But sitting back and waiting for clarity to float down from on high will hurt you more than it hurts them.
So speak up. Be bold. Proactively manage your boss until you have the tools and information you need to succeed.

2. Align team goals with company objectives

Armed with clarity around the company's intended destination, it's time to get your team members all rowing in the same direction. According to one Harvard Business study, 95 percent of a company's employees are unaware of or do not understand their company's strategic plan.

To combat this staggering statistic on your team, start with clear and frequent communication, especially about top company objectives. Set team and individual goals that align with company goals, and make sure you're measuring employees toward these objectives.

Next, address the unspoken assumptions about day-to-day prioritization.
We all come with our own personal biases about what matters most. Some will prioritize tasks based on who requested them—the higher the requestor on the org chart or more loudly they shout, the more important the project request.

Others depend more on when items are due—the earlier the due date, the more attention it gets. Still others will rely on their own pet whys—if it aligns strongly with their personal passions and preferences, it will get done first.
Instead, train your team to consider project value first. Projects that will deliver high ROI to the enterprise rank highest in importance, no matter who assigned them or how urgent they may seem.
But who determines what's valuable? That's where secret number three comes in.

3. Standardize and score work requests

Create standardized processes to initiate projects in the same format every time.
Work management software solutions have these capabilities built in, making sure employees aren't wasting valuable time combing through emails, voicemails, sticky notes, and meeting minutes to figure out what's expected of them. They'll know they have one work queue to consult, where essential details are readily accessible.

The Project Management Institute (PMI) reports that organizations with successful work performance measures (on time, on budget, and goals met) are almost three times more likely than organizations with poor work performance to use standardized practices throughout the organization, and have better outcomes as a result.

Standardized requests alone aren't quite enough, though, if they don't include some indication of importance, value, or priority. In new research findings from Business Improvements Architects, only 32 percent of respondents said they had a process for prioritizing projects. In the same study, 68 percent of organizations said they had no systematic approach in place to prioritize projects or link them to corporate and strategic goals.

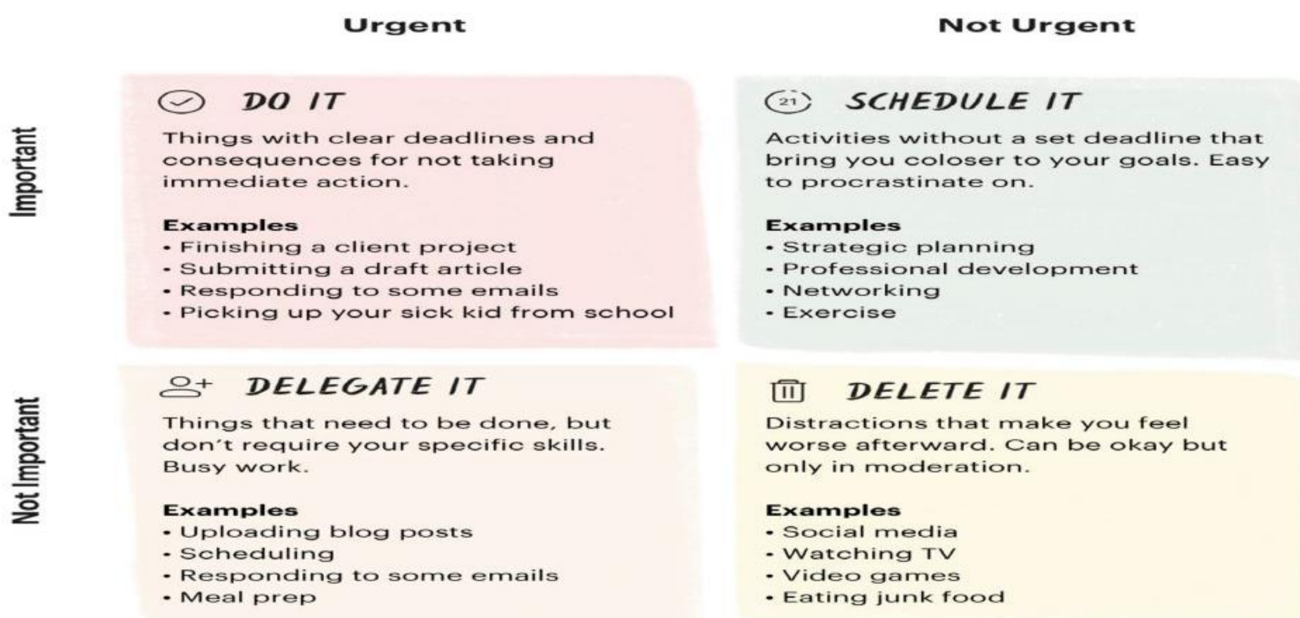
The solution? Implement a scorecard system that assigns strategic point values to all work, helping everyone easily determine which projects are essential and which are more flexible. Encourage an open dialogue as priorities shift and clash throughout the cycle of work.

4. Encourage team to make time for important but not urgent work

Steven Covey's seven famous habits have now been in circulation for 27 years—an entire lifetime for many of today's enterprise workers. But it never hurts to be reminded of tried-and-true principles.

Covey suggested dividing work into four quadrants:

1. Important and Urgent
2. Important and Not Urgent
3. Not Important and Urgent
4. Not Important and Not Urgent



It's easy to find yourself spending too much time hanging out in quadrant three (Not Important and Urgent), and without a scorecard system to help you define universal standards of "importance," many of your team members will be.

After all, "urgent" is one quality that's easy to identify, while other qualities can be more subjective. Make sure your priorities management system is designed to keep your team in quadrants one and two most of the time, with extremely rare forays into quadrant four.

5. Make course corrections

Once you've absorbed strategic company objectives, created a series of complementary goals for your own team, and figured out how to rank each incoming project in terms of value, your work has just begun. Now you need to make sure you have a good bird's eye view of what's going on with your team, so you can offer feedback and make adjustments along the way.

If you follow an Agile project management approach, your weekly stand-up meeting is a great place to do this, but it doesn't always require a meeting. Cloud-based work management solutions make it easy to offer feedback and course corrections minute by minute.

It's important to speak up when you see team members working on unimportant, not urgent work ahead of high-value projects, but it's arguably even more crucial to offer praise and positive reinforcement when you see:

- Team members prioritizing highest-value items first.
- Team members making time for important but not urgent work.
- Team members holding each other accountable to top priorities.
- Team members making suggestions for process improvements.

Priorities management is an ongoing effort

Like eating and organizing, priorities management is not one of those things that can be done "once and for all." It's an ongoing effort that requires constant vigilance.

But, you'll find that practicing alignment in directions, standardizing and scoring work requests, making time for important but not urgent work, and offering course corrections and positive feedback along the way is more than worth the effort.

After all, a boat full of people all rowing in the same direction will get much farther much faster than if everyone is focused on different destinations. It's your job to point them to the desired port, give them the tools they need, and then stay out of their way.

Core Competencies for Success

Performance criteria that are required to become competent in a job:

1. Adaptability and Change

   Adapts well to changes in work assignments and priorities.

   Adapts behavior or work methods in response to new information, changing conditions, or unexpected obstacles. Approaches change positively and adjusts behaviors accordingly.

   a. Understands changes in work tasks, situations, and environment as well as the logic or basis for change; actively seeks information about new work situations.
   b. Approaches change and new situations as positive opportunities for learning or growth; focuses on the beneficial aspects of change; speaks positively about the change to others.
   c. Modifies behavior to deal effectively with changes in the work environment; readily tries new approaches appropriate for new or changed situations; does not persist with ineffective behaviors.

2. Communication

   Clearly conveys information and ideas through a variety of media—written, verbal, and digital—to individuals and groups in a manner that engages the audience and helps them understand and retain the message.

   Demonstrates high level of professionalism.

   a. Demonstrates effective listening for accurate, empathetic understanding.
   b. Informs others involved in a project or effort about developments and plans.
   c. Shares important information from management with employees and others as appropriate.
   d. Informs manager about progress and problems.
   e. Initiates clear, continuous communication within area of responsibility.
   f. Gives and receives constructive feedback.

3. Collaboration and Teamwork

Develops and uses collaborative relationships to facilitate the accomplishment of work goals.

Identifies opportunities and takes action to build strategic relationships between one’s area and other areas, teams, departments, units, or organizations to help achieve business goals.

   a. Listens and responds appropriately to other team members' ideas.
   b. Offers support for others' ideas and proposals.
   c. Confers with other team members about their concerns.
   d. Expresses disagreement constructively (e.g., by emphasizing points of agreement, suggesting alternatives that may be acceptable to the group).

e. Reinforces and gives credit to team members for their contributions.
f. Gives honest, constructive feedback to other team members.
g. Provides assistance to others when needed.
h. Works toward solutions that all team members can support.

4. Continuous Learning

Demonstrates commitment to professional development; attends professional conferences, focuses on best practices, values cutting-edge practices and approaches.
Identifies new areas for learning, and takes advantage of a variety of learning activities and opportunities.
Introduces new knowledge and skills on the job.
Targets learning needs through informed feedback.

a. Identifies and seeks appropriate learning activities (e.g., courses, reading, self-study, coaching, and experiential learning) that help fulfill learning needs.
b. Maximizes learning and actively participates in learning activities (e.g., takes notes, asks questions, critically analyzes information, keeps on-the-job application in mind, does required tasks).
c. Applies new knowledge and skills to practical use on the job; furthers learning through trial and error.
d. Takes risks in learning, accepts unfamiliar or uncomfortable situation in order to learn, and asks questions. Accepts challenging and unfamiliar assignments.

5. Creativity and Innovation

Looks at situations from multiple perspectives. Creates solutions to problems using novel methods and processes.
Generates innovative solutions in work situations.
Experiments and takes risks with different and novel approaches to address work problems and opportunities, or to create something new.

a. Challenges conventional models and identifies assumptions about defining and solving problems.
b. Generates alternative approaches to problem solving.
c. Manages external constraints and approaches.
d. Draws upon multiple and diverse sources (individuals, disciplines, bodies of knowledge) for ideas and inspiration.
e. Thinks expansively and combines ideas in unique ways or makes connections between disparate ideas. Explores different lines of thought, views situations from multiple perspectives, and evaluates multiple solutions and approaches.
f. Ensures relevance by targeting important areas for innovation and developing solutions that address work priorities.

6. Critical Thinking and Problem Solving

Identifies, evaluates, and challenges assumptions that frame thinking and behavior.
Examines ideas—individual, organizational, social—from many perspectives.
Analyzes issues and solves problems with accuracy, clarity, depth, breadth, logic, open-mindedness, and fairness.
Takes informed action.

a. Undertakes a complex task by breaking it down into manageable parts in a systematic, detailed way.
b. Generates several possible explanations or alternatives, anticipates potential obstacles, and prepares contingency plans.
c. Identifies all information needed to solve a problem effectively.
d. Presents problem analysis and possible solutions rather than just identifying the problem.
e. Recognizes one's own limitations and errors, and takes steps for more accurate understanding and corrective action.

7. Organizational Awareness

Understands the structure, operations, culture, and context of the organization, including, relevant political, social, and economic forces.

a. Understands how organizational culture impacts how work gets done and takes this into account in planning and decision making.
b. Understands the goals and objectives of other departments and uses this information to build alliances and resolve issues.
c. Understands the interdependent nature of operations and the impact of various work units on workflow within the organization.
d. Understands how an individual’s decisions impact others across the organization and involves them appropriately.
e. Keeps up-to-date on what is happening across the organization.
f. Shares information with others across the organization based on an understanding of their priorities, goals, and objectives.

8. Self-Awareness

Understands and communicates, honestly, consistently, and effectively, concerning attitudes, beliefs, perspectives, values, interests, strengths, and feelings, all within a context of appropriate work behavior and performance, including periods of stress and adversity. Recognizes personal and professional strengths and limitations, areas for growth, and resources for improvement.

a. Understands how emotional responses to situations influence behavior and perception.
b. Maintains awareness of emotions and uses this information to guide one’s thinking and actions in a constructive manner.
c. Learns and uses strategies and techniques for managing stress and challenge in work situations in order to maintain momentum and diffuse conflicts.
d. Identifies one’s own and others’ strengths, and leverages skills to manage uncertain situations.
e. Recognizes gaps in one’s own skill set, and takes advantage of growth opportunities to enhance personal and professional effectiveness.
f. Acknowledges being “out of one’s depth” and seeks appropriate training or coaching.

9. Service and Quality Focus

Understands and maintains customer/client needs and relationships, internal and external, as the primary driver of work goals and activities. Values and earns customer trust and respect through meeting and exceeding customer expectations.

a. Solves customer problems quickly and effectively.
b. Communicates with customers to determine needs and level of satisfaction.
c. Asks well-formulated questions about needs and expectations.
d. Meets customer needs.
e. Takes personal responsibility for resolving service problems.
f. Measures and tracks student, patient, and customer satisfaction.
g. Presents in a positive manner with customers.

10. Valuing Diversity and Inclusion

Appreciates and incorporates the capabilities, perspectives, and contributions of all individuals. Works effectively with individuals of diverse style, ability, and motivation.

a. Seeks out and uses ideas, opinions, and talents from diverse and various sources and individuals to maximize on tasks or assignments.
b. Seeks understanding through building relationships with people from other cultures and backgrounds.
c. Champions the value of diversity through actions in the workplace (e.g., recruitment, professional development, promotion).
d. Confronts racist, sexist, and all other inappropriate behavior.
e. Challenges exclusionary organizational practices.

f. Examines own biases and behaviors to avoid stereotypical thoughts and actions. Plans and takes action that considers
g. diversity-related implications.

QUALITY STANDARDS ADHERENCE

The important thing is to make the Quality Objectives effective in addressing what needs to be improved. The objectives should be designed to be S.M.A.R.T (specific, measurable, achievable, realistic and time-based) and should have relevance at all levels of the company, meaning that each employee should understand how their job supports meeting the Quality Objectives.

To do this, the following should be addressed:
Specific: An objective needs to be clear and specific. Instead of saying "to improve non-conforming product," a specific Quality Objective would be "to reduce non-conformances on the third widget line," if the third widget production line is showing data as the most troublesome area for non-conforming product.

Measurable: If an objective can't be measured, how will you know if it has been obtained? In order to make a Quality Objective effective, it needs to be measurable, so this means that having an objective "to reduce non-conformances on the third widget line from 15% to 5%" is much more effective than saying "to improve quality of the products on the third widget line." You can measure the defects being made, and therefore make plans to reduce the number of defects, but a vague measure of "quality" is more ephemeral and very hard to plan improvements for.

Agreed: For an objective to be agreed it first needs to be created and approved by top level management. Once management agrees on the objective it needs to be communicated to each level of the organization that will be required to implement the plans to achieve the objective, and the people at these levels of the organization need to agree that the plan is achievable. Without this buy-in they may not fully work towards the goal and the plan may be doomed to failure.

Realistic: Being realistic with an objective will make selling it within your organization easier. If you tell your employees that you want to reduce defects from 50% to 2%, they will not be able to see how this is possible, especially if the plans around the object do not support the improvement. It is better to set realistic goals and overachieve than it is to set unrealistic goals and always fall short of the expectation.

Time-Based: To be truly effective, an objective needs to have a time associated with it. To say "reduce non-conformances on the third widget line from 15% to 5% in the next year" allows for better planning, since a plan needs to have dates in order to be properly tracked. Again, having the time associated will allow you to monitor how close you expect to be in achieving your goals.

Quality standards

Quality standards are defined as documents that provide requirements, specifications, guidelines, or characteristics that can be used consistently to ensure that materials, products, processes, and services are fit for their purpose.
Standards provide organizations with the shared vision, understanding, procedures, and vocabulary needed to meet the expectations of their stakeholders. Because standards present precise descriptions and terminology, they offer an objective and authoritative basis for organizations and consumers around the world to communicate and conduct business.

WHO USES QUALITY STANDARDS?

Organizations turn to standards for guidelines, definitions, and procedures that help them achieve objectives such as:

- Satisfying their customers' quality requirements
- Ensuring their products and services are safe
- Complying with regulations
- Meeting environmental objectives
- Protecting products against climatic or other adverse conditions
- Ensuring that internal processes are defined and controlled

Use of quality standards is voluntary, but may be expected by certain groups of stakeholders. Additionally, some organizations or government agencies may require suppliers and partners to use a specific standard as a condition of doing business.

WHY ARE STANDARDS IMPORTANT?

For businesses: Standards are important to the bottom line of every organization. Successful companies recognize standards as business tools that should be managed alongside quality, safety, intellectual property, and environmental policies. Standardization leads to lower costs by reducing redundancy, minimizing errors or recalls, and reducing time to market.

For the global economy: Businesses and organizations complying to quality standards helps products, services, and personnel cross borders and also ensures that products manufactured in one country can be sold and used in another.

For consumers: Many quality management standards provide safeguards for users of products and services, but standardization can also make consumers' lives simpler. A product or service based on an international standard will be compatible with more products or services worldwide, which increases the number of choices available across the globe.

Supporting an organization's goals and objectives, an ISO 9001 Quality Management documents the procedures, processes as well as responsibilities for attaining quality policies and objectives. ISO 9001 standard outlines the way an organization functions to meet the requirements of its stakeholders and customers based on eight quality management principles.

1. **Document the Process.** Document the business process and have stakeholders review and approve each step to meet quality expectations. A quality management system should hold all this process data – step by step.
2. **Ensure stakeholder agreement.** All stakeholders need to agree on the quality management process, with clear steps put in place and detailed instructions that meet the requirements for success.
3. **Define Key Performance Indicators (KPIs)**. Document the main drivers of the quality process and determine how they affect change. Critically examine the required inputs and expected results.

4. **Measure Results.** Measure the KPIs consistently for each business process and ensure any corrections and changes can be affected to optimize the process continuously.
5. **Assign Accountability**. Perform consistent review of the KPI's and hold the respective stakeholders accountable for the process.
6. **Automate the process.** As complexity can lead to errors, firstly simplify the quality management process before proceeding with automation.
7. **Evaluate Improvement Opportunities.** Develop a process to routinely evaluate quality standards and identify opportunities for improvement and innovation.
8. **Beware the pitfalls.** Embrace lessons learned from the organizations' team members or other companies to minimize catastrophic failures in the quality management process.

Employee work performance
A company's performance standards outline the expectations of its employees, including their roles and how they should act while in the workplace. Employers typically give new employees these standards while they're on boarding to explain common employee goals and requirements.

Performance standards are guidelines the employers give to their employees to outline what the company expects of them as a part of its team. They explain job duties and qualities and to what quality the employee should complete them.

Though employers use performance standards for all employees, they typically give these guidelines to new hires to help them understand their roles and responsibilities. Employers may also use performance standards to help evaluate team members. Performance standards should treat all employees equally and should be specific, measurable and easy to understand. There are three common levels of performance management:
Strategic: This level refers to goals on the organizational, or company, level. Strategic standards should align with the company's vision, objectives and values.
Operational: This level of performance management emphasizes how departmental activities work to achieve the company's goals. Employers may make performance standards that apply to the company's departments.
Individual: This level focuses on the individual employee to ensure that all employees are performing their tasks well. Individual standards evaluate employees' work and seek to improve the quality of their performances.

Performance standards and important because they provide employees with a framework of how the company expects them to work. This allows for open communication between the employer and the employee, which can help the employee understand their responsibilities. It also explains how employees can meet these goals. For example, communication might be a performance standard, and then the employer might explain what communication means to the company and how the employee can meet this standard. Performance standards are also important because employers can use them as evaluation tools.

Types of performance standards
Here are some common types of performance standards that you could use for your company:

Professionalism
This standard refers to how an employee acts within the workplace. Many companies value a respectful and serious workplace, which means that employees should act professionally. Professionalism in the workplace may include being calm and composed, polite and supportive toward other team members. It could also refer to other aspects of workplace culture, such as being punctual and dressing according to the company's dress code. Professionalism can be especially important in a company where employees work as a team or work with customers. For example, it's important for teachers to be professional around their students.

Example: The employee demonstrates a professional demeanor in the workplace, including both their actions and language. The employee also maintains respectful relationships with other employees, management and customers.

Teamwork

Employers may include performance standards regarding teamwork if their employees often work together. Teamwork is the ability to collaborate and work well with other people to complete certain tasks. In a work environment where employees may complete projects together, teamwork is a valuable performance standard to include. Companies may expect employers to help each other when necessary. For instance, employees have to be willing to share work-related information to their team when doing a project. Employers may emphasize that the workplace should be collaborative and employees are committed to helping each other.

Example: The employee actively helps other team members when necessary and appropriate. Employees may either ask for help or provide help to other employees.

Communication

This is the ability to convey messages or information clearly to other people. Communication refers to both written or oral conversation, such as writing an email or presenting on a topic. This requires clear language and strong writing skills. Employers may include this performance standard to promote open lines of communication between employees. In many situations, communication is a valuable skill for an employee to have because it can increase productivity and collaboration. When employees work together, they often communicate to complete a task. Communication can also refer to speaking with clients, which is essential to customer service positions.

Example: The employee clearly communicates with managers, team members and customers. The employee communicates accurate information when appropriate.

Problem-solving

Employers may add problem-solving as a performance standard as well. Problem-solving is the ability to assess an issue and provide possible solutions based on the situation. This is a common standard because most jobs require problem-solving skills regardless of the industry. For example, a dietitian may have to problem-solve to create a diet that works for a client's specific needs. A hiring manager may use problem-solving to create corresponding schedules for multiple employees. Problem-solving standards can evaluate how an employee reacts and resolves a challenge.

Example: The employee demonstrates strong problem-solving skills when faced with an issue. The employee can think of creative solutions and is flexible when problem-solving.

Accountability

This refers to taking responsibility for one's actions. Accountability is an important performance standard because it shows employees that they have ownership over their decisions. This performance standard also informs employees that the company expects them to work independently at times. Employers typically value when employees are determined and can complete tasks without direct management. Accountability can also mean that employees understand their roles and responsibilities as a team member. Employers can use accountability to evaluate an employee's independence and ability to take responsibility.

Example: The employee takes responsibility while at work, meaning they take ownership of their actions and obligations. The employee is able to work independently with limited supervision from management.

Time management

This is the ability to complete a set of given tasks by a deadline. Time management often requires organizational and prioritization skills. Employers may include time management as a performance standard to show that they expect employees to complete their duties by an indicated time. Time management is an important performance standard because many jobs require their employees to perform and complete multiple tasks during the workday or week. Employers can use this performance standard to evaluate how their employees spend their time while working.

Example: The employee finishes their duties by a given deadline. The employee uses prioritization to determine which tasks they should complete first in order to finish their projects on time.

Employee Performance Evaluation forms typically covers the following topics:

- Quality of work (accuracy, thoroughness, competence)
- Quantity of work (productivity level, time management, ability to meet deadlines)
- Job knowledge (skills and understanding of the work)
- Working relationships (ability to work with others, communication skills)
- Achievements

This unit covers the following:
Work requirements:
Activities (what you are required to do), deliverables (the outputs of your work), quantity (the volume of work you are expected to complete), standards (what is acceptable performance, including compliance with Service Level Agreements) and timing (when your work needs to be completed).

Appropriate people:
Line manager, the person requesting the work, members of the team/department, and members from other teams/departments

Resources: equipment, materials, information

Performance Criteria (PC) w.r.t. the Scope
To be competent on the job, you must be able to:
PC1. Establish and agree your work requirements with appropriate people
A good work relationship requires trust, respect, self-awareness, inclusion, and open communication.

- Trust: when you trust your team members, you can be open and honest in your thoughts and actions. And you don't have to waste time or energy "watching your back."
- Respect: teams working together with mutual respect value one another's input, and find solutions based on collective insight, wisdom, and creativity.
- Self-awareness: this means taking responsibility for your words and actions, and not letting your own negative emotions impact the people around you.
- Inclusion: don't just accept diverse people and opinions, but welcome them! For instance, when your colleagues offer different opinions from yours, factor their insights and perspective – or "cultural add " – into your decision-making.
- Open communication: all good relationships depend on open, honest communication. Whether you're sending emails or IMs, or meeting face-to-face or on video calls, the more effectively you communicate with those around you, the better you'll connect.

PC2. Keep your immediate work area clean and tidy
Office etiquette is the customary code of behavior that you’re expected to observe in the workplace. The rules of etiquette define what manners are appropriate and what conduct you should try to avoid.

Here are some actionable steps to help you present the proper office decorum:
Offer a polite greeting.
Make conversation.
Be mindful of others.
Silence your phone.
Give your undivided attention.
Keep the workplace clean.
Arrive on time.
Eat in the break room or outside the workplace.
Dress appropriately.
Meet in designated spaces.
Consider your body language.

PC3. Utilize your time effectively
Time management is the process of planning and exercising conscious control of the time spent on specific activities to work smarter than harder. It is a juggling act of various things that help you increase efficiency and strike a better work-life balance.

Improving your time management at work allows you to enhance your performance and achieve your desired goals with less effort and more effective strategies. However, failing to manage time or poor time management skills at work can result in:

- Missed deadlines and appointments
- Procrastination and lack of focus
- Lack of professionalism
- Inefficient workflow and low work quality
- Unwanted stress
- Poor professional reputation
- Strained workplace relationships
- Financial penalties
- Work and life imbalance

PC4. Use resources correctly and efficiently

Resource management is the process of pre-planning, scheduling, and allocating your resources to maximize efficiency.

A resource is anything that is needed to execute a task or project — this can be the skill sets of employees or the adoption of software. There are, therefore, many types of resource management, depending on the type of project you're working on. For example, if you're planning an event, a few resources include scheduling out staff for the event, planning what vendors to use for promotional materials, investing software that allows attendees to register, and budgeting for everything from giveaways to catering.

Some resource management techniques:

1. Resource Allocation

Resource allocation helps you get the most from your available resources. Based on team members' skills and capacity, resource allocation is the process of tackling projects using the resources you have at your disposal in the most efficient manner possible.

To get a clear view into allocation, project managers will often use resource allocation reports. These can give anywhere from a high-level view to a detailed run down of resource availability — helping you avoid schedule delays and going over budget. The better the reporting capabilities at your disposal, the more transparency and efficiency you will have over your projects.

2. Resource Leveling

What is resource leveling on a project and what are its objectives? This technique aims to discover underused or inefficiently used resources within the organization and work them to your advantage. An example of resource leveling is having a content writer who has experience in graphic design help out the design team by taking on small content tasks that require design work. If a team member can flex their design skills, the design team won't need to hire a freelancer if they suddenly get flooded with design requests

3. Resource Forecasting

Having a resource management plan is critical to optimizing people, materials, and budget efficiency. Resource forecasting allows you to predict your future resource requirements before a project begins. During the planning stages of a project, you could consider resource management software that forecasts the project's scope, possible constraints, unforeseen costs, and potential risks.

PC5. Treat confidential information correctly

Confidentiality means the state of keeping secret or not disclosing information. It comes from confide, meaning to trust someone or tell secrets to them.

Confidential information, therefore, is information that should be kept private or secret. Confidentiality is simply the act of keeping that information private.

Types of Confidential Information

1. Employee Information

In the course of the job, you will hear information about individuals within your organisation. Some of this will be unsubstantiated gossip, and some will be information that you have come across in the course of your work, especially if you are a line manager or you work in human resources.

If you come across private information in the course of your work, including any information that identifies an individual (name, address, maiden name and so on), then you should not disclose it to others. This also applies to information collected at interviews about ethnic background, disabilities and so on.

There are legal requirements about keeping that kind of information safe and secure in many countries. If you break those laws, you are likely to be personally liable and potentially face prosecution.

2. Managerial Information

Managerial information includes both information about individuals, such as disciplinary action, and also about broad management actions such as planned redundancies or employee relations issues.

Some types of managerial information will become public in due course—such as planned redundancies—and other information may remain confidential for much longer.

3. Organizational Information

Organizational information is also known as business information or ‘trade secrets’.

The definition covers anything not in the public domain that helps the organization do its work better or more efficiently. It would therefore include, for example, information about industrial processes, budgets, costs, forecasts, and even customer contact information.

WARNING!

Sometimes organizational information is covered by confidentiality agreements or contracts of employment. If you are tempted to disclose any information of this nature, you should check your contract first and, if necessary, get legal advice.

4. Customer or Contact Information

Customer and contact information is partially covered by ‘trade secrets’.

However, forthcoming changes to data protection law in Europe (and relating to any data held on a citizen of a European Union country) mean that it needs to be considered and held differently.

If you think this may affect you or your company, you are advised to seek legal advice.

5. Professional Information

Some professionals — including doctors, lawyers and accountants — come across information about individuals or organizations through their professional position.

For example:

Doctors know about the details of their patients’ conditions and treatments;

Lawyers know about details of wills and court cases, some of which may be protected by law; and

Accountants will know about their clients’ tax and income.

These professionals are often bound by professional codes of conduct as well as formal legal requirements.

PC6. Work in line with your organization’s policies and procedures

Policies and procedures go hand-in-hand but are not interchangeable.

A policy is a set of general guidelines that outline the organization’s plan for tackling an issue. Policies communicate the connection between the organization’s vision and values and its day-to-day operations.

A procedure explains a specific action plan for carrying out a policy. Procedures tells employees how to deal with a situation and when.

Using policies and procedures together gives employees a well-rounded view of their workplace. They know the type of culture that the organization is striving for, what behavior is expected of them and how to achieve both of these.

PC7. Work within the limits of your job role
Recognizing the importance of working within the limits of your own role and responsibilities maintains good relationships with colleagues who might otherwise think you're stepping on their toes.

PC8. Obtain guidance from appropriate people, where necessary
PC9. Ensure your work meets the agreed requirements

Knowledge and Understanding (K)
A. Organizational Context (Knowledge of the company/ organization and its processes)
You need to know and understand:
KA1. Your organization’s policies, procedures and priorities for your area of work and your role and responsibilities in carrying out your work
KA2. Limits of your responsibilities and when to involve others
KA3. Your specific work requirements and who these must be agreed with
KA4. The importance of having a tidy work area and how to do this
KA5. How to prioritize your workload according to urgency and importance and the Manage your work to meet requirements benefits of this
KA6. Your organization’s policies and procedures for dealing with confidential information and the importance of complying with these
KA7. The purpose of keeping others updated with the progress of your work
KA8. Who to obtain guidance from and the typical circumstances when this may be required
KA9. The purpose and value of being flexible and adapting work plans to reflect Change

B. Technical Knowledge
You need to know and understand:
KB1. The importance of completing work accurately and how to do this
KB2. Appropriate timescales for completing your work and the implications of not meeting these for you and the organization
KB3. Resources needed for your work and how to obtain and use these
A. Core Skills/ Generic Skills
Writing Skills
You need to know and understand how to:
SA1. Complete accurate work with attention to detail
Reading Skills
You need to know and understand how to:
SA2. Read instructions, guidelines, procedures, rules and service level agreements
Oral Communication (Listening and Speaking skills)
You need to know and understand how to:
SA3. Ask for clarification and advice from line managers
SA4. Communicate orally with colleagues

B. Professional Skills
Decision Making
You need to know and understand how to:
SB1. Make decisions on suitable courses Plan and Organize
Plan and Organize
You need to know and understand how to:
SB2. Plan and organize your work to achieve targets and deadlines
SB3. Agree objectives and work requirements

Customer Centricity
You need to know and understand how to:
SB4. Deliver consistent and reliable service to customers
SB5. Check your own work meets customer requirements
Problem Solving
You need to know and understand how to:
SB6. Refer anomalies to the line manager
SB7. Seek clarification on problems from others
Analytical Thinking
You need to know and understand how to:
SB8. Provide relevant information to others
SB9. Analyze needs, requirements and dependencies in order to meet your work requirements
Critical Thinking
You need to know and understand how to:
SB10. Apply judgments to different situations
Attention to Detail
You need to know and understand how to:
SB11. Check your work is complete and free from errors
SB12. Get your work checked by peers
Team Working
You need to know and understand how to:
SB13. Work effectively in a team environment

C. Technical Skill
You need to know and understand how to:
SC1. Use information technology effectively, to input and/or extract data accurately
SC2. Identify and refer anomalies in data
SC3. Store and retrieve information
SC4. Keep up to date with changes, procedures and practices in your role

SC1. Use information technology effectively, to input and/or extract data accurately
In order to make a decision, the managers need knowledge. In case of massive data amounts, issues may occur because of data analysis and necessary knowledge extract. Data is analyzed through an automated process, known as Knowledge Discovery in data mining techniques.
Data mining can be defined as a process of exploring and analysis for large amounts of data with a specific target on discovering significantly important patterns and rules. Data mining helps finding knowledge from raw, unprocessed data. Using data mining techniques allows extracting knowledge from the data mart, data warehouse and, in particular cases, even from operational databases.
In this context, data mining gets an important role in helping organizations to understand their customers and their behavior, keeping clients, stocks anticipation, sale policies optimization as well as other benefits which bring a considerable competitive advantage to the organization.
The main purpose of these techniques is to find patterns and hidden (but relevant) relations that might lead to revenue increase. The essential difference between data mining techniques and the conventional database operation techniques is that, for the second ones, the database becomes passive and is only being used for large amounts of data population, therefore helping in future finding of that specific data. Alternatively, the database is not passive anymore, being able to serve useful information regarding the business plans put in discussion.
Regarding data mining studies, two major types of them exists. One of them is represented by the hypothesis testing, which assumes exposing a theory regarding the relation between actions and their results. The second type of study is represented by the knowledge discovery. For this type of analysis, relations between data warehouse existing data are

tracked. This can be done by using data viewing tools or by using fundamental statistical analysis, such as correlation analysis.
Data mining techniques reside from classic statistical calculation, from database administration and from artificial intelligence. They are not a substitute for traditional statistical techniques, but an extension of graphical and statistical techniques.
Data mining uses a large variety of statistical algorithms, shape recognition, classification, fuzzy logic, machine learning, genetic algorithms, neural networks, data viewing etc., from which we can mention regression algorithms, decision algorithms, neural networks, clustering analysis.

SC2. Identify and refer anomalies in data
Anomaly detection, also called outlier detection, is the identification of unexpected events, observations, or items that differ significantly from the norm. Often applied to unlabeled data by data scientists in a process called unsupervised anomaly detection, any type of anomaly detection rests upon two basic assumptions:
Anomalies in data occur only very rarely
The features of data anomalies are significantly different from those of normal instances.
Typically, anomalous data is linked to some sort of problem or rare event such as hacking, bank fraud, malfunctioning equipment, structural defects / infrastructure failures, or textual errors. For this reason, identifying actual anomalies rather than false positives or data noise is essential from a business perspective.
There are three main classes of anomaly detection techniques: unsupervised, semi-supervised, and supervised. Essentially, the correct anomaly detection method depends on the available labels in the dataset.
Supervised anomaly detection techniques demand a data set with a complete set of “normal” and “abnormal” labels for a classification algorithm to work with. This kind of technique also involves training the classifier. This is similar to traditional pattern recognition, except that with outlier detection there is a naturally strong imbalance between the classes. Not all statistical classification algorithms are well-suited for the inherently unbalanced nature of anomaly detection.
Semi-supervised anomaly detection techniques use a normal, labeled training data set to construct a model representing normal behavior. They then use that model to detect anomalies by testing how likely the model is to generate any one instance encountered.
Unsupervised methods of anomaly detection detect anomalies in an unlabeled test set of data based solely on the intrinsic properties of that data. The working assumption is that, as in most cases, the large majority of the instances in the data set will be normal. The anomaly detection algorithm will then detect instances that appear to fit with the rest of the data set least congruently.

SC3. Store and retrieve information
Storing information is the process where information is deposited or stored in a storehouse (cabinets, HDD, memory stick, etc) and retrieving information is the process of obtaining the stored information resources relevant to the needs.
The main purpose of storing any information is for easy retrieval in the future when it is required. It is also part of a company’s business practice. Storing documents have to be done in proper procedure so that it is easier to find. Depending on the type of information, information can be stored in either fireproof cabinets for hard copies or hard disks or other electronic storage devices for soft copies.
Depending on the nature of the company or organisation or business, different types of information will have to be stored. It can be secrets related to the company, highly confidential files related to the company or employees or customers / clients. So files have to be stored following all the confidentiality procedures and protected according to data protection act as these are information necessary for providing quality service and also for the smooth running of the company.
The purpose of information retrieval is to provide quality service for the right person at the right time, with all the required information in hand. Only if data is stored in a procedural manner it can be easily retrieved. Information might be retrieved for marketing purposes, for communications, for monitoring purposes; for surveys and other

research that an organization or business might conduct. Information will also be retrieved in situations where information might have to be shared with partner companies and with the police or Inland Revenue.

SC4. Keep up to date with changes, procedures and practices in your role

The practice of keeping up to date with trends and changes in your industry is a necessity in our modern working world. It's an important ingredient to stay competitive and be successful.

- Subscribe to top industry magazines or/and online blogs
  An industry magazine or trade paper is the most traditional way to keep your knowledge up to date. Additionally, you can find blogs covering many industries and niche topics. You just need to do some digging. The process of searching for your go-to sources will be well worth the effort helping you to obtain a fresh perspective.
- Follow industry experts and influencers on social media
  This may be an obvious one. Social media is the best way to be instantly informed. It provides headlines and perspectives that you can then research more deeply if you find them relevant and interesting. Consider following your favorite blogs, trade papers, companies, and their prominent followers.
  As LinkedIn is a professional network, it is a fantastic place to find and follow knowledgeable people and influential companies. But don't discount Twitter, Instagram and Facebook.
- Attend conferences
  Attending a conference is an immersive way of keeping up to date. You will not only learn what is going on in your industry by going to a conference, but you can also really develop new ideas and speak to interesting people. They are also the best places to grow your professional network in your field. Just watch out, conferences can be expensive, so make sure you are selective.
- Network
  Find and maintain contacts in your industry and keep in touch with them. These may be people you have worked with, want to work with, or have met at a conference. There are also social media groups that you can join for every different industry imaginable, within which you can access a wealth of information and perspectives from a vast range of experience levels and locations.
  Whether you keep in touch with people by email, social media messages, or (the best way) by meeting in person – perhaps for a coffee or a working lunch; you can learn so much about what is happening in your industry from networking on a personal level like this.
- Take an online course
  Sometimes, keeping up to date with what you learn from reading and speaking to people is not enough, as the level of detail and understanding you require to stay relevant and progress is greater. The best way to improve your expertise in a situation like this is to take an online course. This could range from following a course to taking a full certifying program. As well as keeping up to date, further education always increases your value, be it directly or indirectly.

Understanding Learning objectives:
The benefits of this course include:
Efficient and Effective time management
Efficient – Meeting timelines
Effective – Meeting requirement for desired output
Awareness of the SSC (Sector Skill Council) environment and time zone understanding
Awareness of the SSC environment and importance of meeting timelines to handoffs
Review the course objectives listed above.
"To fulfil these objectives today, we'll be conducting a number of hands-on activities. Hopefully we can open up some good conversations and some of you can share your experiences so that we can make this session as interactive as possible. Your participation will be crucial to your learning experience and that of your peers here in the session today."
Question: Please share your thoughts on following?
A. Time is perishable – Cannot be created or recovered
B. Managing is only option – Prioritize
Importance of Time Management
The first part of this session discusses the following:
"Plan better avoid wastage"
Understanding the timelines of the deliverables.
Receiving the hand off from upstream teams at right time is critical to start self contribution and ensure passing the deliverables to downstream team. It is important to value others' time as well to ensure overall organizational timelines are met. Share the perspective of how important is time specifically in a global time zone mapping scenario
Suggested Responses:
Time management has to be looked at an organizational level and not just individual level
These Aspects teach us how to build the blocks of time management.
Time Management Aspects
Prompt participants to come up with some aspects and relate them back to here.

- □ Planning and goal setting
- □ Managing yourself
- □ Dealing with other people
- □ Your time
- □ Getting results

The first 4 Interconnect and Interact to give the 5th one – Results
Differentiate between Urgent and Important task
Assume importance as they demand immediate attention Important Task
May become urgent if left undone
Usually have a long term effect To judge importance vs. urgency, gauge tasks in terms of
Impact of doing them
Effect of not doing them

Main aim of prioritization is to avoid a crisis
We must Schedule our Priorities as opposed to Prioritizing our Schedule
1. Urgent and Important – Do Now
2. Not Urgent and Important – Schedule on your calendar
3. Urgent and Not Important – Delegate, Automate or Decline
4. Not Urgent Not Important – Delegate, Automate or Decline

Check Your Understanding

1. True or False? Time can be stored.
a. True
b. False

Suggested Responses:

False – Time once lost cannot be gotten back – hence important to plan time utilization properly

2. True or False? Time is perishable
a. True
b. False

Suggested Responses:

True – Time lost is lost for every – lost moments cannot be gotten back

3. True or False? Time management is required both at individual level and organizational level.
a. True
b. False

Suggested Responses:

True – plan for activities organizational level and also at individual level

4. True or False? Activities should be judged basis Urgency and Importance
c. True
d. False

Suggested Responses:

True – prioritization should be based on 2x2 matrix of urgency and importance

Team Exercise

Ask the participants to pick up the items listed below and place them in the Urgent/Important quadrant. Discuss the rationale of their thoughts and categorization. List the items and ask participants to classify them as per the quadrant.
Discuss the rationale of their thoughts and categorization.

Categorize the below items in the Time Management Quadrant

1. Wildly important goal
2. Last minute assignments from boss
3. Busy work
4. Personal health
5. Pressing problems
6. Crises
7. Planning
8. Time wasters
9. Professional development
10. Win-win performance agreement
11. Too many objectives
12. Vital customer call
13. Major Deadlines
14. Unimportant pre scheduled meetings
15. Meaningless management reports
16. Coaching and mentoring team
17. Low priority email
18. Other people's minor issues
19. Workplace gossip
20. Exercise
21. Needless interruptions
22. Defining contribution

23. Aimless Internet surfing
24. Irrelevant phone calls

Suggested Answers:
Depends on rationale shared
1. Wildly important goal – Q1
2. Last minute assignments from boss – Q1
3. Busy work – Q4 – Consumes time however not pressing
4. Personal health – Q4 – requires planning and care not pressing
5. Pressing problems – Q1 – has to be solved immediately
6. Crises – Q1 – have to tended to immediately
7. Planning – Q2 – Important but not urgent; should be done before crisis
8. Time wasters – Q4
9. Professional development – Q2
10. Win-win performance agreement – Q2 – Expectation setting part of planning
11. Too many objectives – Q3 – Prioritize further to establish which are important and pressing
12. Vital customer call – Q1 – Customer centricity
13. Major Deadlines – Q1
14. Unimportant pre scheduled meetings – Q3
15. Meaningless management reports – Q3 – Prioritize further to establish which are important and pressing
16. Coaching and mentoring team – Q2
17. Low priority email – Q3 – Prioritize further to establish which are important and pressing
18. Other people's minor issues – Q3 – May not be urgent but important for team building
19. Workplace gossip – Q4 – Non value add; occasionally creates negativity

Exercise – Q4 –
Important for health and personal wellbeing. To be done in spare and leisure time.
Cannot be ignored.

21. Needless interruptions – Q3
22. Defining contribution – Q2
23. Aimless Internet surfing – Q4
24. Irrelevant phone calls – Q4 – Reserve and avoid

Work Management and Prioritization
Preparing morning tea is a good example. Define time, no of family members, preparation required at night and then in the morning. Perfect execution to ensure good morning tea with family.
Gather responses.
Start the session by connecting the course content to the candidate responses.

Work Management
Six steps for expectation setting with the stakeholders
1. Describe the jobs in terms of major outcomes and link to the organization's need The first step in expectation setting is to describe the job to the employees. Employees need to feel there is a greater value to what they do.
We need to feel out individual performance has an impact on the organization's mission.
Answer this question: My work is the key to ensuring the organization's success because…
While completing the answer link it to
- Job Description
- Team and Organization's need
- Performance Criteria

2. Share expectations in terms of work style While setting expectation, it's not only important to talk about the "what we do" but also on "how we expect to do it". What are the ground rules for communication at the organization?
Sample ground rules
- Always let your team know where are the problems. Even if you have a solution, no one likes surprises.
- Share concerns openly and look for solutions
- If you see your colleagues doing something well, tell them. If you see them doing something poorly, tell them.

Sample work style questions
- Do you like to think about issues by discussing them in a meeting or having quite time alone?
- How do you prefer to plan your day?

3. Maximize Performance -
Identify what is required to complete the work:
Supervisor needs / Employee needs. Set input as well as output expectations.
In order to ensure employees are performing at their best, the supervisor needs to provide not only the resource (time, infrastructure, desk, recognition etc.) but also the right levels of direction (telling how to do the task) and support (engaging with employees about the task).
4. Establish priorities.
Establish thresh holds and crisis plan Use the time quadrant to establish priorities. Refer to earlier session.
5. Revalidate understanding.
Create documentation and communication plan to establish all discussion
When you are having a conversation about expectations with stakeholders, you're covering lot of details so you'll need to review to make sure you both have a common understanding of the commitments you have made.
6. Establish progress check
No matter how careful you have been in setting expectations, you'll want to follow up since there will be questions as work progresses.
Schedule an early progress check to get things started the right way, and agreed on scheduled/unscheduled further checks. Acknowledge good performance and point your ways to improve.

		Does this need to happen NOW?	
		Urgent	Not Urgent
Does this really matter?	Important	**I.** **Do it now** **Critical Activities**	**II.** **Do it next** **Important Goals**
	Not Important	**III.** **Delegate or Reject and Explain** **Interupptions**	**IV.** **Resist and Cease** **Distractions**

Importance

Urgency

Introduction to Analytics
(Associate Analytics – I)

Syllabus

Unit I: Introduction to Analytics and R programming (NOS 2101)

Introduction to R, R Studio (GUI): R Windows Environment, introduction to various data types, Numeric, Character, date, data frame, array, matrix etc., Reading Datasets, Working with different file types .txt,.csv etc. Outliers, Combining Datasets in R, Functions and loops.

Manage your work to meet requirements (NOS 9001)

Understanding Learning objectives, Introduction to work & meeting requirements, Time Management, Work management & prioritization, Quality & Standards Adherence,

Unit II: Summarizing Data & Revisiting Probability (NOS 2101)

Summary Statistics - Summarizing data with R, Probability, Expected, Random, Bivariate Random variables, Probability distribution. Central Limit Theorem etc.

Work effectively with Colleagues (NOS 9002)

Introduction to work effectively, Team Work, Professionalism, Effective Communication skills, etc.

Unit III: SQL using R

Introduction to NoSQL, Connecting R to NoSQL databases. Excel and R integration with R connector.

Unit IV: Correlation and Regression Analysis (NOS 9001)

Regression Analysis, Assumptions of OLS Regression, Regression Modelling. Correlation, ANOVA, Forecasting, Heteroscedasticity, Autocorrelation, Introduction to Multiple Regression etc.

Unit V: Understand the Verticals - Engineering, Financial and others (NOS 9002)

Understanding systems viz. Engineering Design, Manufacturing, Smart Utilities, Production lines, Automotive, Technology etc.

Understanding Business problems related to various businesses

Requirements Gathering: Gathering all the data related to Business objective

Reference Books:

1. **Introduction to Probability and Statistics Using R**, ISBN: 978-0-557-24979-4, is a textbook written for an undergraduate course in probability and statistics.
2. **An Introduction to R**, by Venables and Smith and the R Development Core Team. http://www.r-project.org/, see Manuals.

3. Montgomery, Douglas C., and George C. Runger, **Applied statistics and probability for engineers.** John Wiley & Sons, 2010.
4. The ***Basic Concepts of Time Series Analysis***. http://anson.ucdavis.edu/~azari/sta137/AuNotes.pdf
5. **Time Series Analysis and Mining with R,** Yanchang Zhao.

UNIT I

UNIT-I: Introduction to Analytics and R programming (NOS 2101)
Introduction to R, R Studio (GUI): R Windows Environment, introduction to various data types, Numeric, Character, date, data frame, array, matrix etc., Reading Datasets, Working with different file types .txt,.csv etc. Outliers, Combining Datasets in R, Functions and loops.

Manage your work to meet requirements (NOS 9001)

Understanding Learning objectives, Introduction to work & meeting requirements, Time Management, Work management & prioritization, Quality & Standards Adherence.

UNIT-I: Introduction to Analytics and R programming (NOS 2101)

Topic No.	Name of the Topic
1.1	Introduction to R
1.2	R Studio (GUI):
1.3	Introduction to various data types: Numeric, Character ,date, data frame, array, matrix etc.,
1.4	Reading Datasets and Working with different file types . txt,.csv etc.
1.5	Outliers
1.6	Combining Datasets in R
1.7	Functions and Loops
2.10	**Provide Data/Information in Standard formats (NOS 9004)**

1.1. Introduction to R

What is R?

R is a flexible and powerful open-source implementation of the language S (for statistics) developed by John Chambers and others at Bell Labs.

Why R?

Five reasons to learn and use R:

- ✓ R is open source and completely free. R community members regularly contribute packages to increase R's functionality.
- ✓ R is as good as commercially available statistical packages like SPSS, SAS, and Minitab.
- ✓ R has extensive statistical and graphing capabilities. R provides hundreds of built-in statistical functions as well as its own built-in programming language.
- ✓ R is used in teaching and performing computational statistics. It is the language of choice for many academics who teach computational statistics.
- ✓ Getting help from the R user community is easy. There are readily available online tutorials, data sets, and discussion forums about R.

R uses:

- ✓ R combines aspects of functional and object-oriented programming.
- ✓ R can use in interactive mode
- ✓ It is an interpreted language rather than a compiled one.
- ✓ Finding and fixing mistakes is typically much easier in R than in many other languages.

R Features:-

- ✓ Programming language for graphics and statistical computations
- ✓ Available freely under the GNU public license

- ✓ Used in data mining and statistical analysis
- ✓ Included time series analysis, linear and nonlinear modeling among others
- ✓ Very active community and package contributions
- ✓ Very little programming language knowledge necessary
- ✓ Can be downloaded from http://www.r-project.org/ **opensource**

What is CRAN?

CRAN abbreviates **Comprehensive R Archive Network** will provide binary files and follow the installation instructions and accepting all defaults. Download from *http://cran.r-project.org/* we can see the R Console window will be in the RGui (graphical user interface). Fig 1 is the sample

R GUI.

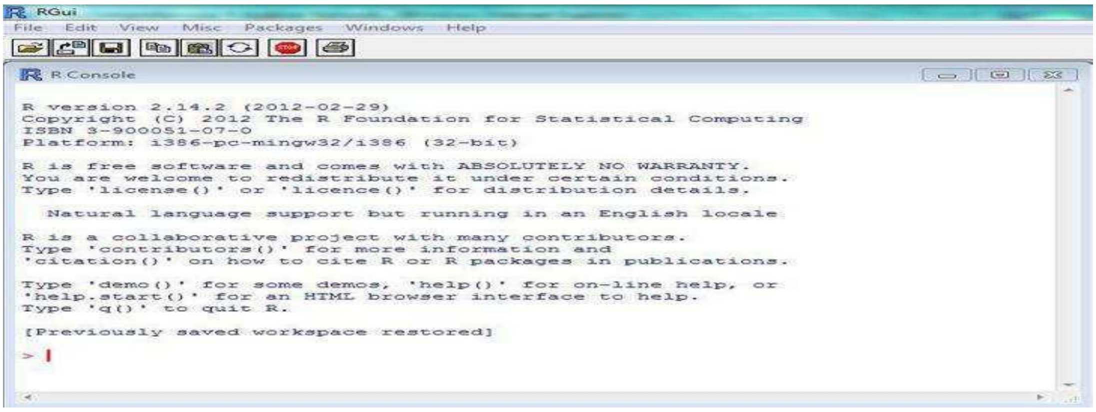


Figure 1. R console

R Studio: R Studio is an Integrated Development Environment (IDE) for R Language with advanced and more user-friendly GUI. R Studio allows the user to run R in a more user-friendly environment. It is open-source (i.e. free) and available at **http://www.rstudio.com/**.

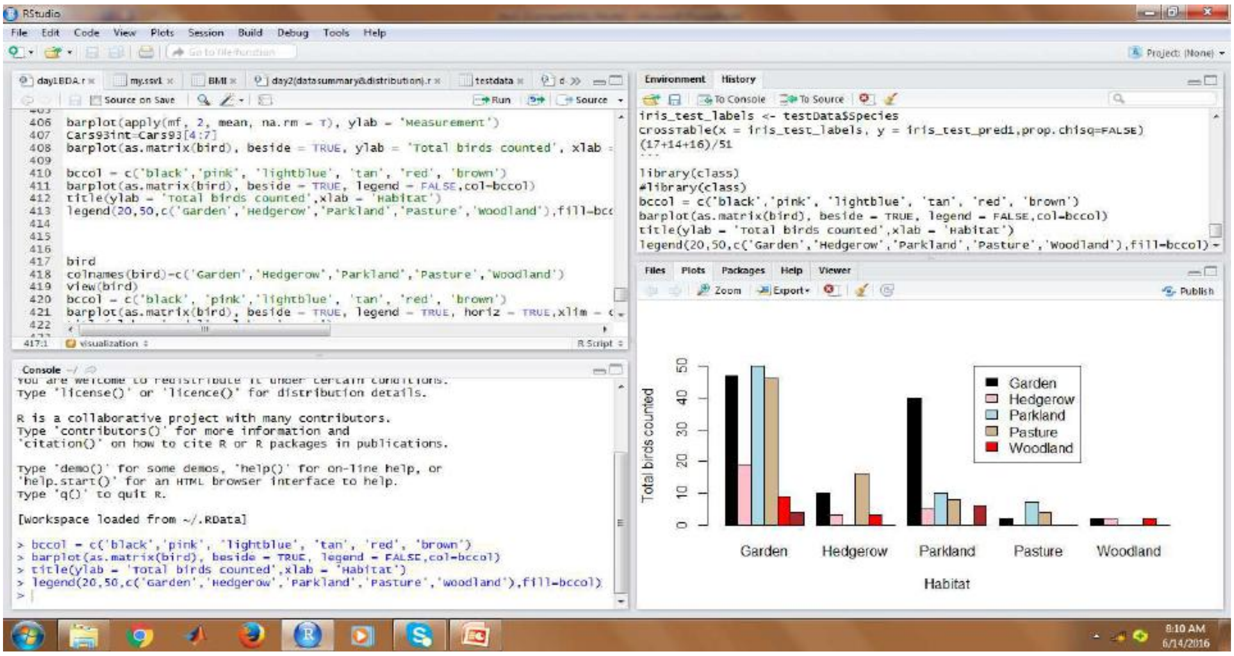


Figure 2. R studio GUI

Figure 2 shows the GUI of R Studio.The R Studio screen has four windows:
1. Console.
2. Workspace and history.
3. Files, plots, packages and help.

4. The R script(s) and data view.

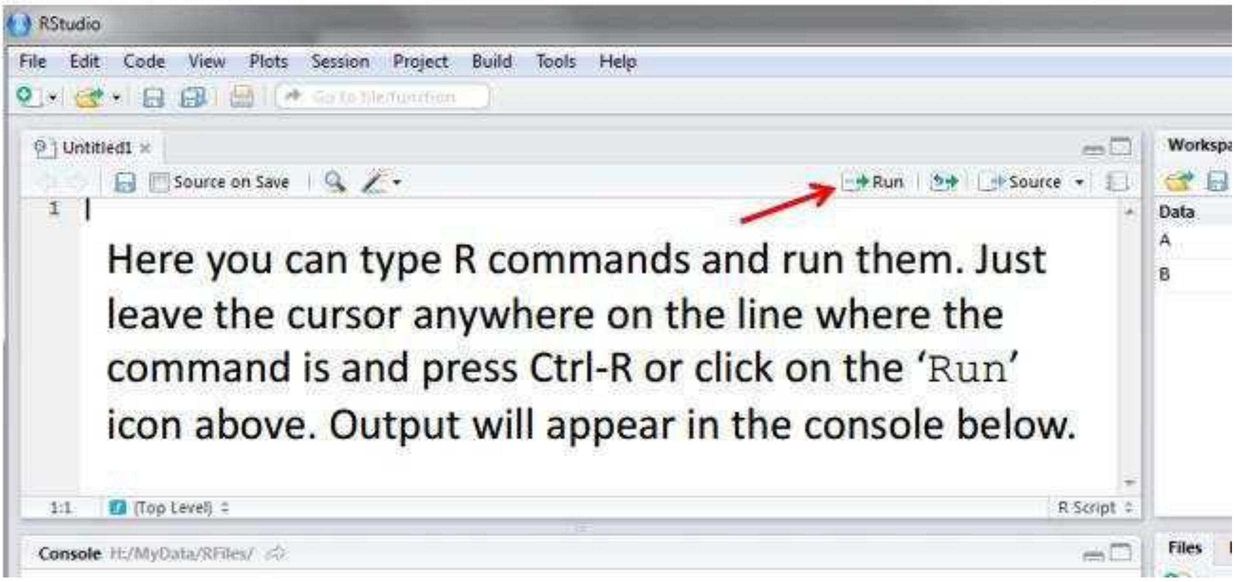
The R script is where you keep a record of your work.

Create a new R script file:

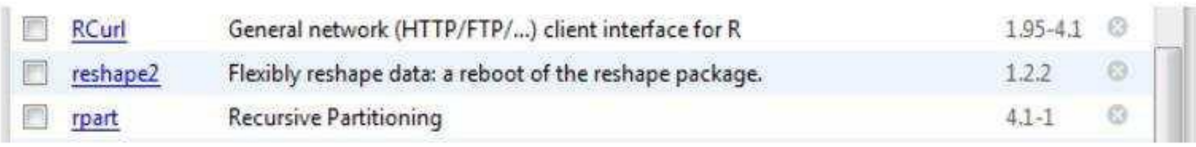
To create a new R script file:

1) File -> New -> R Script,
2) Click on the icon with the "+" sign and select "R Script"
3) Use shortcut as: Ctrl+Shift+N.

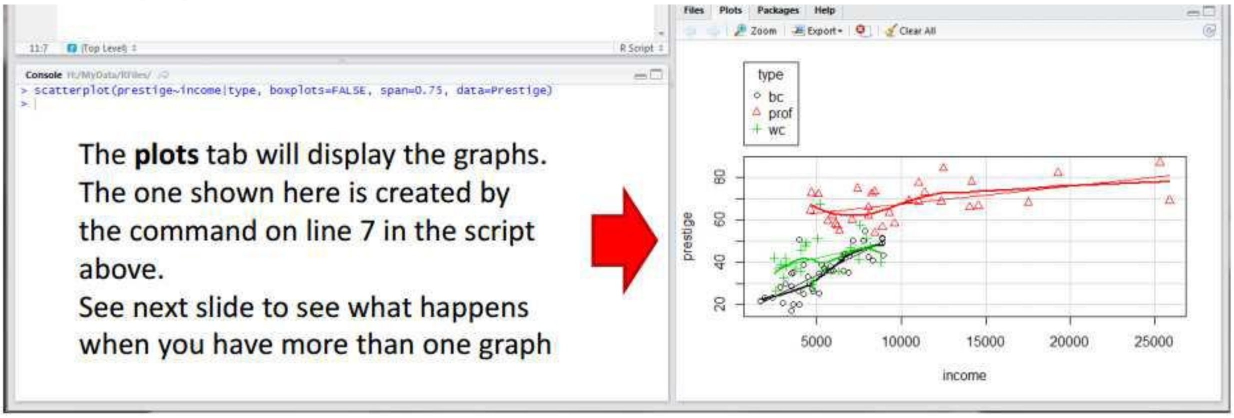
Running the R commands on R Script file:



Installing Packages:



Plots to display:

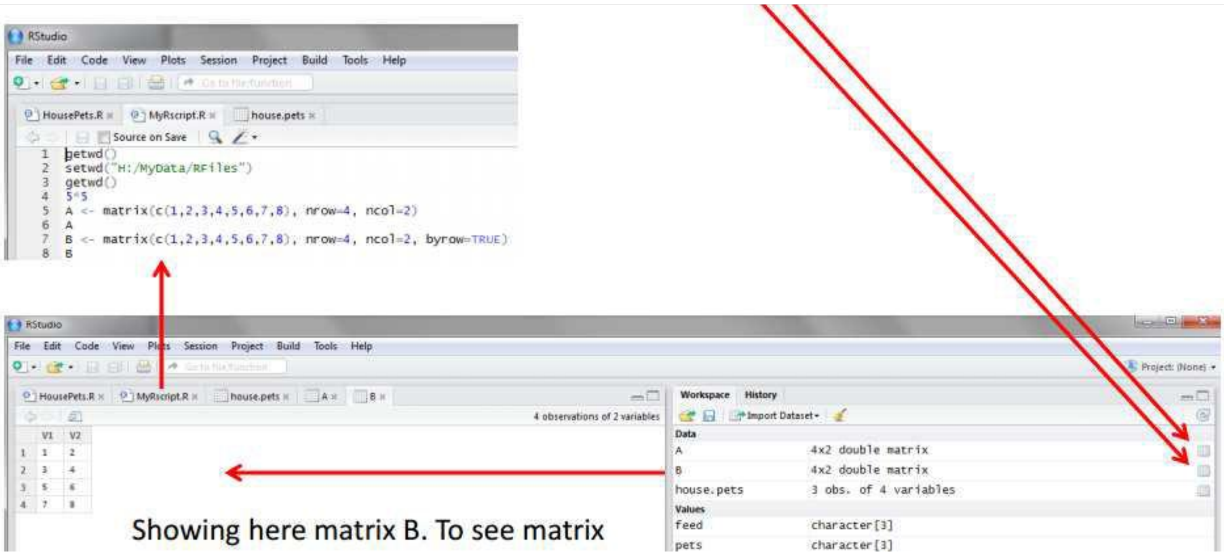


Console:

The console is where you can type commands and see output.

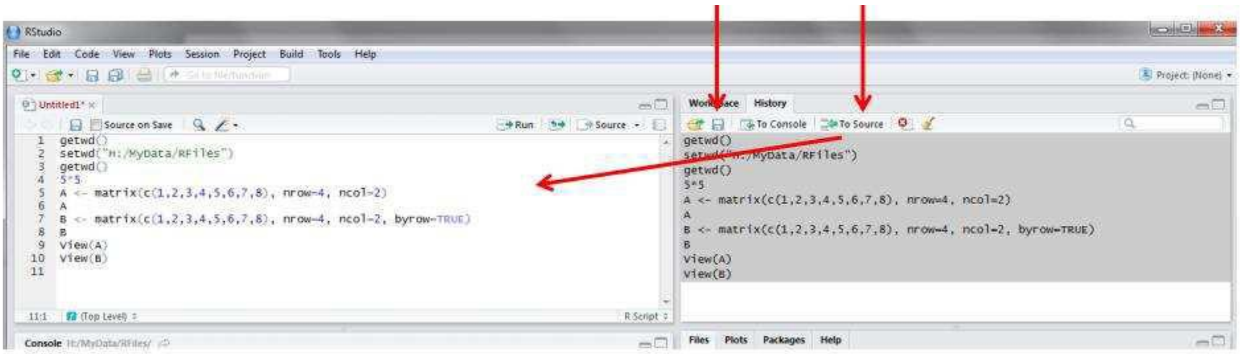
Workspace tab:

The workspace tab shows all the active objects (see next slide). The workspace tab stores any object, value, function or anything you create during your R session. In the example below, if you click on the dotted squares you can see the data on a screen to the left.



History tab:

The history tab shows a list of commands used so far.The history tab keeps a record of all previous commands. It helps when testing and running processes. Here you can either save the whole list or you can select the commands you want and send them to an R script to keep track of your work. In this example, we select all and click on the "To Source" icon, a window on the left will open with the list of commands. Make sure to save the 'untitled1' file as an *.R script.



Files Tab:

The files tab shows all the files and folders in your default workspace as if you were on a PC/Mac window. The plots tab will show all your graphs. The packages tab will list a series of packages or add-ons needed to run certain processes.

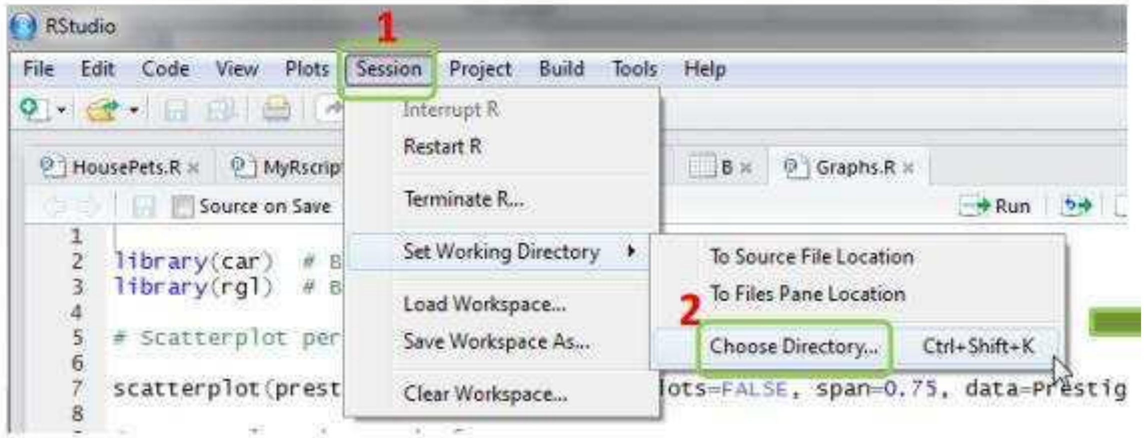
Changing the working directory:

To Show the present working directory (wd)

```
>getwd()
C:/mydocuments        #The default working directory is mydocuments
```

To change the working directory

```
>setwd("C:/myfolder/data")
```



First R program: Using R as calculator:

R commands can run in two ways:

1) Type at console and press enter to see the output. Output will get at console only in R studio.

2) Open new R Script file and write the command, keep the curser on the same line and press Ctrl+enter or click on Run. Then see the output at console along with command.

At console:

R as a calculator, typing commands directly into the R Console. Launch R and type the following code, pressing < Enter > after each command.

Type an *expression on console*.

R - Assignment Operators:

<- or = for assignment and == to test equality.

At the outer sider <- and = can be used similar. But we should be careful while using them in combined. In precise '<-' is prioritized than '=' . The operators <- and = assign into the environment in which they are evaluated. The operator <- can be used anywhere, whereas the operator = is only allowed at the top level (e.g., in the complete expression typed at the command prompt) or as one of the subexpressions in a braced list of expressions.

```
x <- y <- 5
x = y = 5
x = y <- 5
x <- y = 5
# Error in (x <- y) = 5 : could not find function "<-<-"
```

Example Exercise1:

```
> 2 * 2   ## Multiplication
[1] 4
> 2 / 2   ## Division
[1] 1
> 2 + 2   ## addition
[1] 4
> 2 – 2   ## subtraction
[1] 0
> 2 ^ 2   ## exponentiation
[1] 4
```

```
> q()      ## to quit
> y <- 3*exp(x)
> x <- 3*exp(x)
```

R expression:

At '>' (R-prompt) type the R expression and press enter

R output:

R labels each output value with a number in square brackets. As far as R is concerned, an individual number is a one-element vector. The [1] is simply the index of the first element of the vector.

Activity1: Calculate the following using R:
1. Log of 2
2.23 X 32
3. e3

Variable (Object) Names:

Certain variable names are reserved for particular purposes. Some reserved symbols are: **c q t C D F I T**
meaning of c q t C D F I T
? ## to see help document
?c ## c means Combine Values into a Vector or List
?q ## q means Terminate an R Session
?t ## t means Matrix Transpose
?C ## C means sets contrast for a factor
?D ## D means Symbolic and Algorithmic Derivatives of Simple Expressions
?F ## F means logical vector Character strings
c("T", "TRUE", "True", "true") are regarded as true, c("F", "FALSE", "False", "false") as false, and all others as NA.
>F ##[1] FALSE
?I ##Inhibit Interpretation/Conversion of Objects

Working on variables:

Operators in R:

Table - Arithmetic operators		**Table. Logical operators**	
Operator	**Description**	**Operator**	**Description**
+	Addition	**<**	less than
-	Subtraction	**<=**	less than or equal to
*****	Multiplication	**>**	greater than
/	Division	**>=**	greater than or equal to
^ or **	Exponentiation	**==**	exactly equal to
x %% y	modulus (x mod y) 5%%2 is 1	**!=**	not equal to
x %/% y	integer division 5%/%2 is 2	**!x**	Not x
		x | y	x OR y
		x & y	x AND y
		isTRUE(x)	test if X is TRUE

Getting Help in R:

>help(**function**), or use the

>?**keyword** checks in all packages--- function shortcut

>??**keyword** checks in content of all packages

Either of above all will open the R documentation.

Comment Notation in R:

is used to comment a line in R script.

List of Objects:

To see a listing of the objects in your workspace, you can use the **ls()** function. To get more detail, use **ls.str()**

> ls()

[1] "A" "acctdata" "address" "B" "b1"

[6] "balance" "c" "CareerSat" "chisquare" "colnames"

1.3. Introduction to various Data Types:

In Analytics the data is classified as Quantitative(numeric) and Qualitative(Character/Factor) on very broad level.

- Numeric Data: - It includes 0~9, "." and "- ve" sign.
- Character Data: - Everything except Numeric data type is Character.
  For Example, Names, Gender etc.

For Example, "1,2,3…" are Quantitative Data while "Good", "Bad" etc. are Qualitative Data. We can convert Qualitative Data into Quantitative Data using Ordinal Values. For Example, "Good" can be rated as 9 while "Average" can be rated as 5 and "Bad" can be rated as 0.

Table 1. List of Data types in R

Data Type		Verify

Logical	TRUE , FALSE	`v <- TRUE`
		`print(class(v))`
		`[1] "logical"`

Numeric	12.3, 5, 999	`v <- 23.5`
		`print(class(v))`
		`[1] "numeric"`

Integer	2L, 34L, 0L	`v <- 2L`
		`print(class(v))`
		`[1] "integer"`

Complex	3 + 2i	`v <- 2+5i`
		`print(class(v))`
		`[1] "complex"`

Character	'a' , '"good", "TRUE", '23.4'	`v <- "TRUE"`
		`print(class(v))`
		`[1] "character"`

Raw	"Hello" is stored as 48 65 6c 6c 6f	`v <- charToRaw("Hello")`
		`print(class(v))`
		`[1] "raw"`

mode() or Class():

These are used to know the type of data object type assigned.

Example:

Assign several different objects to x, and check the mode (storage class) of each object.

 # Declare variables of different types:

```
my_numeric <- 42
my_character <- "forty-two"
my_logical <- FALSE
```

 # Check which type these variables have:

```
>class(my_numeric)
[1] "numeric"
> class(my_character)
[1] "character"
> class(my_logical)
[1] "logical"
```

Mode vs Class:

'mode' is a mutually exclusive classification of objects according to their basic structure. The 'atomic' modes are numeric, complex, charcter and logical. Recursive objects have modes such as 'list' or 'function' or a few others. An object has one and only one mode.

'class' is a property assigned to an object that determines how generic functions operate with it. It is not a mutually exclusive classification. If an object has no specific class assigned to it, such as a simple numeric vector, it's class is usually the same as its mode, by convention.

Changing the mode of an object is often called 'coercion'. The mode of an object can change without necessarily changing the class.

e.g.

```
> x <- 1:16
> mode(x)
[1] "numeric"
> dim(x) <- c(4,4)
> mode(x)
[1] "numeric"
> class(x)
[1] "matrix"
> is.numeric(x)
[1] TRUE
> mode(x) <- "character"
> mode(x)
[1] "character"
> class(x)
[1] "matrix"
```

```
However:
```

```
> x <- factor(x)
> class(x)
[1] "factor"
> mode(x)
[1] "numeric"
```

 # Arithmetic operations on R objects

```
> x <- 2
> x
[1]  2
> x ^ x
```

```
[1]  4
> x ^ 2
[1]  4
> mode(x)    ## will returen the storage class of object
[1]  "numeric"
>  seq(1:10)   ## will create a vector of 1 to sequence numbers
[1] 1 2 3 4 5 6 7 8 9 10
>  x <- c(1:10)   ##vector of 1 to 10 digits
>  x
 [1]  1 2 3 4 5 6 7 8 9 10
> mode(x)
[1] "numeric"
> x <- c("Hello","world","!")
> mode(x)
[1] "character"
> x <- c(TRUE, TRUE, FALSE, FALSE, TRUE, FALSE, TRUE)
> mode(x)
[1] "logical"
> x <- list("R","12345",FALSE)
> x
[[1]]
[1] "R"
[[2]]
[1] "12345"
[[3]]
[1] FALSE
> mode(x)
[1] "list"
```

Create new variables using already available variables:
Example:
mydata$sum <- mydata$x1 + mydata$x2
New variable is created using two already available variables.
Modifying existing variable: Rename the existing variable by using *rename()* function.
For examples,
mydata<- rename(mydata, c(oldname="newname"))

1.3.1. Vectors:

Vector is the most common data structure in R. Vectors must be homogeneous i.e, the type of data in a given vector must all be the same. Vectors can be numeric, logical, or character. If a vector is mix data types then R forces (**coerces**, if you will) the data into one mode.
Creating a vector:
To create a vector , "concatenate" a list of numbers together to form a vector.
x < - c(1, 6, 4, 10, -2) ## c() to concatenate elements
my.vector<- 1:24 ## a numeric vector with 1 to 24 numbers
List of built-in functions to get useful summaries on vectors:
Example1:
> sum(x) ## sums the values in the vector
> length(x) ## produces the number of values in the vector, ie its length
> mean(x) ## the average (mean)
> var(x) ## the sample variance of the values in the vector
> sd(x) ## the sample standard deviation of the values in the vector (square root of the sample variance)
> max(x) ## the largest value in the vector
> min(x) ## the smallest number in the vector

```
> median(x) ## the sample median
> y < - sort(x) ## the values arranged in ascending order
```

Example2:

```
linkedin <- c(16, 9, 13, 5, 2, 17, 14)
> last <- tail(linkedin, 1)
> last
[1] 14
 > # Is last under 5 or above 10?
 > # Is last between 15 (exclusive) and 20 (inclusive)?
 > # Is last between 0 and 5 or between 10 and 15?
 > (last > 0 | last < 5)
[1] TRUE
> (last > 0 & last < 5)
[1] FALSE
> (last > 10 & last < 15)
[1] TRUE
```

Example3: Following are some other possibilities to create vectors

```
> x <- 1:10
> y <- seq(10) #Create a sequence
> z <- rep(1,10) #Create a repetitive pattern
> x
[1] 1 2 3 4 5 6 7 8 9 10
> y
[1] 1 2 3 4 5 6 7 8 9 10
> z
[1] 1 1 1 1 1 1 1 1 1 1
```

Adding elements to vector:

```
> x <- c(x, 11:15)
>x
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
```

Vector Arithmetic:

```
> x <- c(1:10)
> x
[1] 1 2 3 4 5 6 7 8 9 10
> y <- 10
> x + y
[1] 11 12 13 14 15 16 17 18 19 20
> 2 + 3 * x #Note the order of operations
[1] 5 8 11 14 17 20 23 26 29 32
> (2 + 3) * x #See the difference
[1] 5 10 15 20 25 30 35 40 45 50
> sqrt(x) #Square roots
[1] 1.000000 1.414214 1.732051 2.000000 2.236068 2.449490 2.645751 2.828427
[9] 3.000000 3.162278
> x %% 4 #This is the integer divide (modulo) operation
[1] 1 2 3 0 1 2 3 0 1 2
> y <- 3 + 2i #R does complex numbers
> re(y) #The real part of the complex number
[1] 3
> im(y) #The imaginary part of the complex number
[1] 2
> x * y
[1] 3+ 2i 6+ 4i 9+ 6i 12+ 8i 15 + 10i 18 + 12i 21 + 14i 24 + 16i 27 + 18i 30 + 20i
```

1.3.2. Matrices:

A matrix is a two-dimensional rectangular data set. Matrices must be heterogeneous i.e, the type of data in a given vector of all not in the same class.

Matrix can be created in three ways :

- matrix(): A vector input to the matrix function.
- Using rbind() and cbind() functions.
- Using dim() to the existing vector

Creating a matrix using matrix():

```
# Create a matrix.
M = matrix( c('a','a','b','c','b','a'), nrow=2,ncol=3,byrow = TRUE)
print(M)
     [,1] [,2] [,3]
[1,] "a"  "a"  "b"
[2,] "c"  "b"  "a"
```

Creating a matrix using rbind() or cbind():

First crate two vectors and then create a matrix using rbind() .It binds the two vectors data into two rows of matrix.

Example: To create a matrix havind the data as: 6,2 ,10 & 1, 3, -2

Step1: create two vectors as xr1,xr2

```
> xr1 <- c( 6, 2, 10)
> xr2 <- c(1, 3, -2)
> x <- rbind (xr1, xr2) ## binds the vectors into rows of a matrix(2X3)
> x
    [,1] [,2] [,3]
xr1    6    2   10
xr2    1    3   -2
```

First crate two vectors and then create a matrix using rbind() .It binds the two vectors data into two rows of matrix.

Example: To create a matrix havind the data as: 6,2 ,10 & 1, 3, -2

create two vectors as xr1,xr2

```
> y <- cbind(xr1, xr2) ## binds the same vectors into columns of a matrix(3X2)
> y
     xr1 xr2
[1,]   6   1
[2,]   2   3
[3,]  10  -2
```

Matrix operations:

```
> A <- matrix(c( 6, 1, 0, -3, -1, 2),3, 2, byrow = TRUE)
> B <- matrix(c( 4, 2, 0, 1, -5, -1),3, 2, byrow = TRUE)
>A (with output)
>B (with output)
> A + B
     [,1]  [,2]
[1,] 10    3
[2,] 0    -2
[3,] -6    1
> A - B
      [,1] [,2]
```

```
[1,]  2   -1
[2,]  0   -4
[3,]  4   3
> A * B # this is component-by-component multiplication, not matrix multiplication
     [,1] [,2]
[1,]  24  2
[2,]  0  -3
[3,]  5  -2
> t(A) ## Transpose of a matrix
[,1] [,2] [,3]
[1,] 6 0 -1
[2,] 1 -3 2
```

Alternative method to create a matrix using dim():

Create a vector and add the dimensions using the **dim ( )** function.It's especially useful if you have your data already in a vector.

Example: A vector with the numbers 1 through 24, like this:

>my.vector<- 1:24

You can easily convert that vector to an array exactly like my.array simply by assigning the dimensions, like this:

> dim(my.vector) <- c(3,4)

1.3.3. Array:

Arrays can be of any number of dimensions. The array function takes a **dim** attribute which creates the required number of dimension. In the below example we create an array with two elements which are 3x3 matrices each. Creating an array:

>my.array< - array(1:24, dim=c(3,4,2))

In the above example, "my.array" is the name of the array we have given. There are 24 units in this array mentioned as "1:24" and are divided in three dimensions "(3, 4, 2)".

Alternative: with existing vector and using dim()

> my.vector<- 1:24

To convert my.vector vector to an array exactly like my.array simply by assigning the dimensions, like this:

> dim(my.vector) <- c(3,4,2)

Activity 2:Create an Array with name "MySales" with 30 observations using following methods:

1. Defining the dimensions of the array as 3, 5 and 2.
2. By using Vector method.

1.3.4. Lists:

A list is a R object which can contain many different types of elements inside it like vectors, functions and even another list inside it.

>list1 <- list(c(2,5,3),21.3,sin) # Create a list.
>print(list1) # Print the list.

```
[[1]]
[1] 2 5 3
```

```
[[2]]
[1] 21.3
[[3]]
function (x)  .Primitive("sin")
```

1.3.5. Data Frames:

Data frames are tabular data objects. Unlike a matrix in data frame each column can contain different modes of data. The first column can be numeric while the second column can be character and third column can be logical. It is a list of vectors of equal length. Data Frames are created using the data.frame() function. It displays data along with header information.

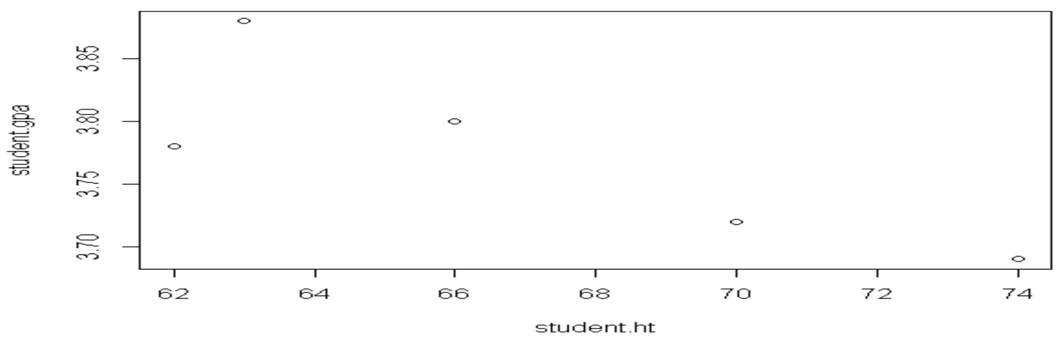
To retrieve data in a particular cell:

Enter its row and column coordinates in the single square bracket "[ ]" operator.

Example:

To retrieve the cell value from the first row, second column of mtcars.

```
>mtcars[1,2]
mtcars[row,column]
# Create the data frame.
>BMI <- data.frame(gender = c("Male", "Male","Female"), height = c(152, 171.5, 165), weight =
c(81,93, 78),Age =c(42,38,26))
>print(BMI)
    gender height weight Age
1   Male   152.0     81    42
2   Male   171.5     93    38
3  Female  165.0     78   26
Data1 :  Height GPA
 66        3.80
 62        3.78
 63        3.88
 70        3.72
 74        3.69
> student.ht <- c( 66, 62, 63, 70, 74)
> student.gpa < - c( 3.80, 3.78, 3.88, 3.72, 3.69)
> student.data1 < - data.frame(student.ht, student.gpa)
> student.data
    student.ht      student.gpa
1        66            3.80
2        62             3.78
3        63             3.88
4        70             3.72
5        74              3.69
> plot(student.ht, student.gpa)
```



1.3.6. Date:

Date() function is used to access the date and time in R

Sys.Date() : The current system date. This function returns a Date object.

class(Date)

A string in this format is treated as a character unless cast to a Date type.

>class("2010-06-16")

>class(as.Date("2010-06-16"))

You can also pass in dates in other formats and cast them as strings by specifying the format in use.

>as.Date('02/03/2004','%m/%d/%Y')

To format date information in a wide variety of string formats, use the strftime function.

>strftime(Sys.Date(),'%A: %B %d, %Y (Day %j of %Y)')

This returns the string "Tuesday: June 15, 2010 (Day 166 of 2010)"

Dates can be manipulated arithmetically.

Activity:

To return the next ten days…

seq(Sys.Date(),Sys.Date() + 10,1)

…or the last ten days…

seq(Sys.Date(),Sys.Date() – 10,-1)

1.4. Reading Datasets using R

We can import Datasets from various sources having various file types :

Example:

- **.csv or .txt format**
- **Big data tool – Impala**
- **CSV File**

The sample data can also be in comma separated values (CSV) format. Each cell inside such data file is separated by a special character, which usually is a comma, although other characters can be used as well. The first row of the data file should contain the column names instead of the actual data. Here is a sample of the expected format

```
Col1,Col2,Col3
100,a1,b1
200,a2,b2
300,a3,b3
```

After we copy and paste the data above in a file named "mydata.csv" with a text editor, we can read the data with the function read.csv.

In R data can read in two ways either from local disc or web.

From disc:

The data file location is known on local disc use: **read.csv() or read.table()** functions.

Path is not specific then use : **file.choose()**

> mydata = read.csv("mydata.csv") # read csv file

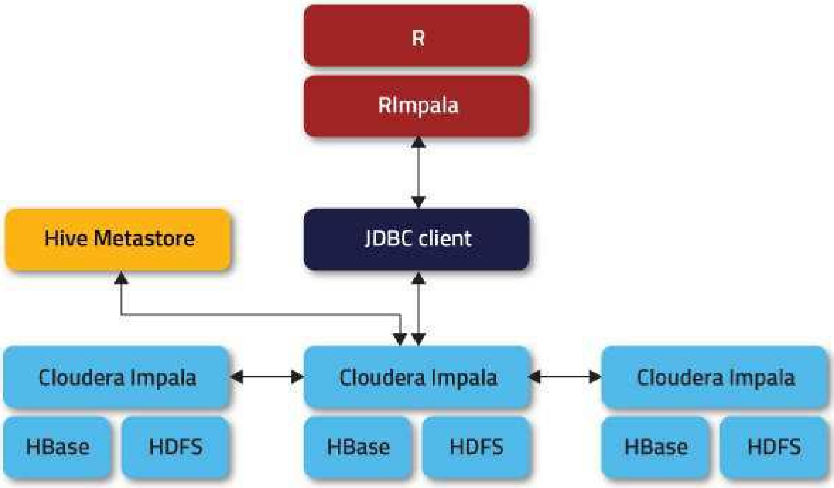
> mydata

From Web:

The URL of the data from web is pass to read.csv() or read.table() functions.

Big data tool – Impala:

- Cloudera 'Impala', which is a massively parallel processing (MPP) SQL query engine runs natively in Apache Hadoop.
- R package, RImpala, connects Impala to R.
- RImpala enables querying the data residing in HDFS and Apache HBase from R, which can be further processed as an R object using R functions.
- RImpala is now available for download from the Comprehensive R Archive Network (CRAN) under GNU General Public License (GPL3).
- This package is developed and maintained by MuSigma.



To install RImpala :

We use following code to install RImpala package.

>install. packages("RImpala")

Importing and Exporting CSV :

- Loading data – data(dataset_name)
- read and write functions
- getwd() and setwd(dir)
- read and write functions use full path name

Example:

read.csv ("C:/Rtutorials/Sampledata.csv").

Writing dataset :write () function.

getwd() means get the working directory (wd) and
setwd() is used to set the working directory.

Activity 3:

Import a CSV file in R and check the output.
Name, Age, Sex
Shaan, 21, M
Ritu, 24, F
Raj, 31, M

1.5. Outliers:

- Outlier is a point or an observation that deviates significantly from the other observations.

Reasons for outliers: Due to experimental errors or "special circumstances"

Outlier detection tests to check for outliers.

There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection.[7][8][9][10] Some are graphical such as normal probability plots. Others are model-based. Box plots are a hybrid.

Outlier treatments are three types:

– **Retention :**

There is no rigid mathematical definition of what constitutes an outlier; determining whether or not an observation is an outlier is ultimately a subjective exercise. There are various methods of outlier detection.[7][8][9][10] Some are graphical such as normal probability plots. Others are model-based. Box plots are a hybrid.

– **Exclusion :**

Deletion of outlier data is a controversial practice frowned upon by many scientists and science instructors; while mathematical criteria provide an objective and quantitative method for data rejection, they do not make the practice more scientifically or methodologically sound, especially in small sets or where a normal distribution cannot be assumed. Rejection of outliers is more acceptable in areas of practice where the underlying model of the process being measured and the usual distribution of measurement error are confidently known. An outlier resulting from an instrument reading error may be excluded but it is desirable that the reading is at least verified.

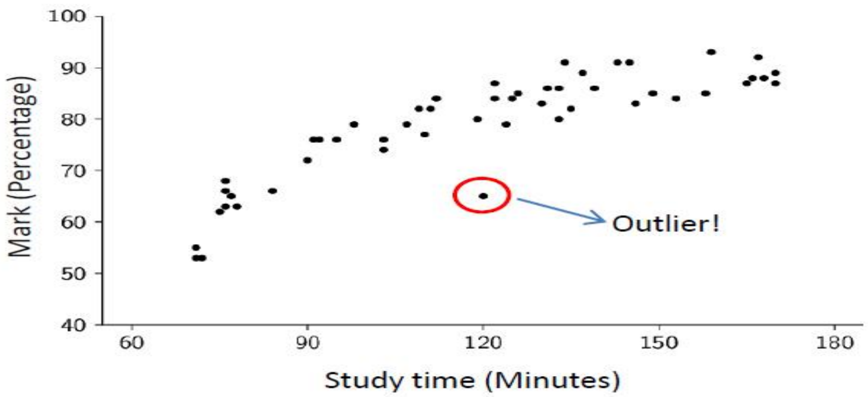
– **Other treatment methods**

OUTLIER package in R: to detect and treat outliers in Data.

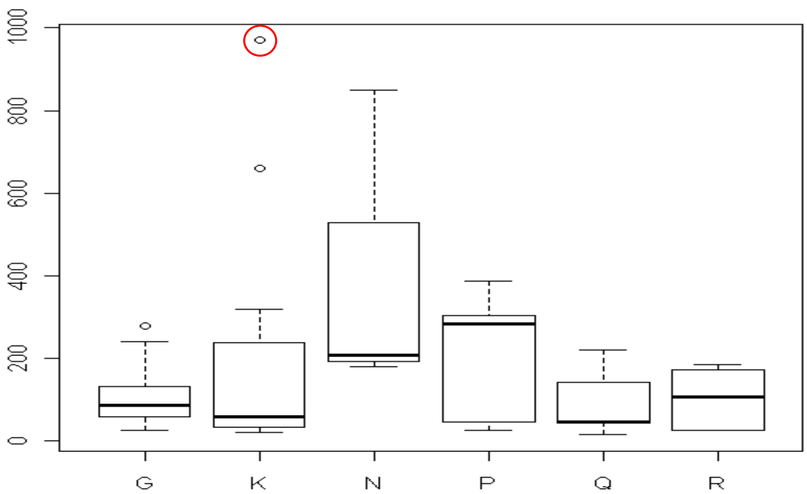
Outlier detection from graphical representation:

– Scatter plot and Box plot

Scatter Plot



Box Plot



The observations out of box are treated as outliers in data

Outliers and Missing Data treatment:

Missing Values

- In **R**, missing values are represented by the symbol **NA** (not available).
- Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number) and R outputs the result for dividing by zero as 'Inf'(Infinity).
- **R** uses the same symbol for character and numeric data.
- To Test missing values: is.na () function

Example:

```
y <- c(1,2,3,NA)
is.na(y)
[1] FALSE FALSE FALSE TRUE
mean(y)##Arithmetic functions on missing values
[1] NA
x <- c(1,2,NA,3)
  mean(x) # returns NA
  mean(x, na.rm=TRUE) # returns 2
```

- To remove missing values :**na.omit() function.**
  newdata<- na.omit(y)
- Alternative method using na.rm=TRUE
  mean(x, na.rm=TRUE)

PMM approach to treat missing values:

- **PMM**-> Predictive Mean Matching (PMM) is a semi-parametric imputation approach.
- It is similar to the regression method except that for each missing value, it fills in a value randomly from among the observed donor values from an observation
- whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model.
- **MICE Package -> Multiple Imputation by Chained Equations**
- MICE uses PMM to impute missing values in a dataset.

1.6. Combining Data sets in R

Merge(): To merge two data frames (datasets) horizontally. In most cases, we can join two data frames by one or more common key variables (i.e., an inner join).

To merge two data frames by ID:

    total <- merge(data frameA, data frameB, by="ID")

To merge on more than one criteria :To merge two data frames by ID and Country:

    total <- merge(data frameA,data frameB,by=c("ID","Country"))

To join two data frames (datasets) vertically : **rbind** function. The two data frames **must** have the same variables, but they do not have to be in the same order.

Example:

    total <- rbind(data frameA, data frameB)

Plyr package: Tools for Splitting, Applying and Combining Data. We use rbind.fill() in plyr package in R. It binds or combines a list of data frames filling missing columns with NA.

Example: rbind.fill(mtcars[c("mpg", "wt")], mtcars[c("wt", "cyl")])

In this all the missing value will be filled with NA.

1.7. Function and Loops:

In R functions and loop structure are created using for,while along with conditional statements if-else.

1.7.1. "FOR" Loop:

For loop is used to repeat an action for every value in a vector. "for" loop in R :

for(i in values){... do something ...}

This for loop consists of the following parts:

The keyword **for**, followed by parentheses.An identifier between the parentheses. In this example, we use i, but that can be any object name you like.The keyword **in**, which follows the identifier.

Syntax:

for (val in sequence) { statement}

```
Example :To count the number of even numbers in a vector.
#For Loop Example 1:
```

```
x<-c(34,5,8,1,2,67,98)
count<-0
for(y in x)
{
  if(y%%2==0)
  {
    count = count+1;
    print(paste("y=",y))
  }
}
print(paste("count=",count))

##The above can be written in simple as:
x<-c(34,5,8,1,2,67,98)
count<-0
for(y in x) if(y%%2==0){ count = count+1;print(paste("y=",y))}
print(paste("count=",count))
```

Output:

```
[1] "y= 34"
[1] "y= 8"
[1] "y= 2"
[1] "y= 98"
[1] "count= 4"
```

```
#For Loop Example 2:
for(i in -5:3) if(i>0) print(paste(i," is +ve ")) else
  if(i<0) print(paste(i," is -ve")) else print(paste(i," is Zero"))
```

output:

```
[1] "-5  is -ve"
[1] "-4  is -ve"
[1] "-3  is -ve"
[1] "-2  is -ve"
[1] "-1  is -ve"
[1] "0  is Zero"
[1] "1  is +ve "
[1] "2  is +ve "
[1] "3  is +ve "
```

```
#For Loop Example 3:
x<-c(2,NA,5)
for(n in x) print(n*2)
```

output:

```
[1] 4
[1] NA
[1] 10
```

1.7.2. While loop

while (test_expression) { statement }

```
Example1:
i<-1
while(i<=10) i=i+4
i
```

output:

```
 [1] 13
```

```
Example2:
i<-1
while(i<=10) {print(paste("i= ",i));i=i+4 }
i
```

output:

```
[1] "i=  1"
[1] "i=  5"
```

```
[1] "i=  9"
[1] 13
```

Example3:

```
i<-1
while(T) if(i>10) break else i=i+4
i
```

output:

```
[1] 13
```

1.7.4 repeat:

Syntax: repeat { statements }

```
#Example:
i<-1
repeat { i<-i+4 ; if(i>10) break }
i
```

output:

```
[1] 13
```

1.7.5. IF-ELSE Function

```
if ( test_expression1) {   statement1} else if ( test_expression2) {   statement2} else
if ( test_expression3) {   statement3} else   statement4
```

Example1:

```
x <- -5
y <- if(x > 0) 5 else 6
[1] 6
```

Example2:

```
x <- 0
if (x < 0)
{   print("Negative number")}
else if (x > 0)
{   print("Positive number")}
else
  print("Zero")
```

Output:

```
[1] Zero
```

Example3:

```
##ifelse is a function works like if-else
i<-1
ival<-ifelse(i>0,3,4)
ival
```

Output:

```
[1] 3
```

Example 4:

```
i<-0
if(i) i else 0
```

Output:

```
[1] 0
i<--4
if(i) i else 0
```

Output:

```
[1] -4
```

User defined functions:

To define functions in R using function()

One-Line Functions:

```
Example:
>log2 = function(x) log(x, base = 2)
> log2(64)
[1] 6
```

Using Default Values in Functions:

Functionname<-function(argument1,...){function body}

It returns by default the last statement return value.

Example1:

```
> sum<-function(i,j){ i+j }      ## function sum will take two integers and return the sum of integers
> sum(2,3)   ## calling of function
[1] 5
manning = function(radius, gradient, coef=0.1125) (radius^(2/3)*gradient^0.5/coef)
> manning(radius = 1, gradient = 1/500)
[1] 0.3975232
```

Manage your work to meet requirements (NOS 9001)
Understanding Learning objectives, Introduction to work & meeting requirements, Time Management, Work management & prioritization, Quality & Standards Adherence.

Understanding Learning objectives:
The benefits of this course include:

- Efficient and Effective time management
- Efficient – Meeting timelines
- Effective – Meeting requirement for desired output
- Awareness of the SSC (Sector Skill Council) environment and time zone understanding
- Awareness of the SSC environment and importance of meeting timelines to handoffs

Review the course objectives listed above.
"To fulfil these objectives today, we'll be conducting a number of hands-on activities. Hopefully we can open up some good conversations and some of you can share your experiences so that we can make this session as interactive as possible. Your participation will be crucial to your learning experience and that of your peers here in the session today."
Question: Please share your thoughts on following?
A. Time is perishable – Cannot be created or recovered
B. Managing is only option – Prioritize

Importance of Time Management

The first part of this session discusses the following:

- "Plan better avoid wastage"
- Understanding the timelines of the deliverables. Receiving the hand off from upstream teams
- at right time is critical to start self contribution and ensure passing the deliverables to
- downstream team.
- It is important to value others' time as well to ensure overall organizational timelines are met
- Share the perspective of how important is time specifically in a global time zone mapping scenario

"Yes, I have room in my schedule to attend a Time Management Seminar...the day after I retire!"

Suggested Responses:

- Time management has to be looked at an organizational level and not just individual level
- These Aspects teach us how to build the blocks of time management.

Time Management Aspects
Prompt participants to come up with some aspects and relate them back to here.

-
-
-
-
-

The first 4 Interconnect and Interact to give the 5th one – Results

Differentiate between Urgent and Important task

- Assume importance as they demand immediate attention Important Task
- May become urgent if left undone
- Usually have a long term effect To judge importance vs. urgency, gauge tasks in terms of
- Impact of doing them
- Effect of not doing them

Main aim of prioritization is to avoid a crisis
We must
Time Management quadrants
Schedule our Priorities as opposed to Prioritizing our Schedule
1. Urgent and Important – Do Now
2. Not Urgent and Important – Schedule on your calendar
3. Urgent and Not Important – Delegate, Automate or Decline
4. Not Urgent Not Important – Delegate, Automate or Decline

Check Your Understanding

1. True or False? Time can be stored.
   a. True
   b. False

Suggested Responses:
False – Time once lost cannot be gotten back – hence important to plan time utilization properly

2. True or False? Time is perishable
   a. True
   b. False

Suggested Responses:
True – Time lost is lost for every – lost moments cannot be gotten back

3. True or False? Time management is required both at individual level and organizational level.
   a. True
   b. False

Suggested Responses:
True – plan for activities organizational level and also at individual level

4. True or False? Activities should be judged basis Urgency and Importance
   c. True
   d. False

Suggested Responses:
True – prioritization should be based on 2x2 matrix of urgency and importance

Team Exercise

Ask the participants to pick up the items listed below and place them in the Urgent/Important quadrant. Discuss the rationale of their thoughts and categorization.

List the items and ask participants to classify them as per the quadrant.

Discuss the rationale of their thoughts and categorization.

Categorize the below items in the Time Management Quadrant

1. Wildly important goal
2. Last minute assignments from boss
3. Busy work
4. Personal health
5. Pressing problems
6. Crises
7. Planning
8. Time wasters
9. Professional development
10. Win-win performance agreement
11. Too many objectives
12. Vital customer call
13. Major Deadlines
14. Unimportant pre scheduled meetings
15. Meaningless management reports
16. Coaching and mentoring team
17. Low priority email
18. Other people's minor issues
19. Workplace gossip
20. Exercise
21. Needless interruptions
22. Defining contribution
23. Aimless Internet surfing
24. Irrelevant phone calls

Suggested Answers:

Depends on rationale shared

1. Wildly important goal – Q1
2. Last minute assignments from boss – Q1
3. Busy work – Q4 – Consumes time however not pressing
4. Personal health – Q4 – requires planning and care not pressing
5. Pressing problems – Q1 – has to be solved immediately
6. Crises – Q1 – have to tended to immediately
7. Planning – Q2 – Important but not urgent; should be done before crisis
8. Time wasters – Q4
9. Professional development – Q2
10. Win-win performance agreement – Q2 – Expectation setting part of planning
11. Too many objectives – Q3 – Prioritize further to establish which are important and pressing
12. Vital customer call – Q1 – Customer centricity
13. Major Deadlines – Q1
14. Unimportant pre scheduled meetings – Q3

15. Meaningless management reports – Q3 – Prioritize further to establish which are important and pressing
16. Coaching and mentoring team – Q2
17. Low priority email – Q3 – Prioritize further to establish which are important and pressing
18. Other people’s minor issues – Q3 – May not be urgent but important for team building
19. Workplace gossip – Q4 – Non value add; occasionally creates negativity

Exercise – Q4 –
Important for health and personal wellbeing. To be done in spare and leisure time.
Cannot be ignored.
21. Needless interruptions – Q3
22. Defining contribution – Q2
23. Aimless Internet surfing – Q4
24. Irrelevant phone calls – Q4 – Reserve and avoid

Summary

- It is important to manage time.
- To manage time one must:
- Prioritize
- Define Urgency
- Define Importance

Work Management and Prioritization

Preparing morning tea is a good example. Define time, no of family members, preparation required at night and then in the morning. Perfect execution to ensure good morning tea !!! with family.
Gather responses.
Start the session by connecting the course content to the candidate responses.

Work Management

Six steps for expectation setting with the stakeholders
1. Describe the jobs in terms of major outcomes and link to the organization's need The first step in expectation setting is to describe the job to the employees. Employees need to feel there is a greater value to what they do. We need to feel out individual performance has an impact on the organization's mission.
Answer this question: My work is the key to ensuring the organization's success because…
While completing the answer link it to
- Job Description
- Team and Organization's need
- Performance Criteria

2. Share expectations in terms of work style While setting expectation, it's not only important to talk about the "what we do" but also on "how we expect to do it". What are the ground rules for communication at the organization?

Sample ground rules
- Always let your tam know where are the problems. Even if you have a solution, no one likes surprises.
- Share concerns openly and look for solutions
- If you see your colleagues doing something well, tell them. If you see them doing something poorly, tell them.

Sample work style questions
- Do you like to think about issues by discussing them in a meeting or having quite time alone?
- How do you prefer to plan your day?

3. Maximize Performance -
Identify what is required to complete the work: Supervisor needs / Employee needs. Set input as well as output expectations.
In order to ensure employees are performing at their best, the supervisor needs to provide not only the resource (time, infrastructure, desk, recognition etc.) but also the right levels of direction (telling how to do the task) and support (engaging with employees about the task).

4. Establish priorities.
Establish thresh holds and crisis plan Use the time quadrant to establish priorities. Refer to earlier session.

5. Revalidate understanding.
Create documentation and communication plan to establish all discussion
When you are having a conversation about expectations with stakeholders, you're covering lot of details so you'll need to review to make sure you both have a common understanding of the commitments you have made.

6. Establish progress check
No matter how careful you have been in setting expectations, you'll want to follow up since there will be questions as work progresses.
Schedule an early progress check to get things started the right way, and agreed on scheduled/unscheduled further checks. Acknowledge good performance and point your ways to improve.

Does this really matter?	Does this need to happen NOW?	Urgent	Not Urgent
	Important	**I.** **Do it now** **Critical Activities**	**II.** **Do it next** **Important Goals**
	Not Important	**III.** **Delegate or Reject and Explain** **Interupptions**	**IV.** **Resist and Cease** **Distractions**

Importance

Urgency

***** End of Unit 1 *****

Introduction to Analytics
(Associate Analytics – I)
UNIT II

Summarizing Data & Revisiting Probability (NOS 2101)

Summary Statistics - Summarizing data with R, Probability, Expected, Random, Bivariate Random variables, Probability distribution. Central Limit Theorem etc.

Work effectively with Colleagues (NOS 9002)

Introduction to work effectively, Team Work, Professionalism, Effective Communication skills, etc.

S.No	Content
2.1	Summary Statistics - Summarizing data with R
2.2	Probability
2.3	Expected
2.4	Random Variables
2.5	Bivariate Random variables
2.6	Probability distribution
2.7	Central Limit Theorem etc.
2.8	**Work effectively with Colleagues (NOS 9002)**

2.1. Summary Statistics - Summarizing data with R:

Example1:

```
> grass
rich   graze
1   12   mow
2  15   mow
3   17   mow
4  11   mow
5  15   mow
6   8    unmow
7   9   unmow
8   7   unmow
9   9   unmow
```

a) summary():

It gives the summary statistics of data object in terms of min, max,1st Quartile and 3rd Quartile mean/median values.

```
> x<-c(1,2,3,4,5,6,7,8,9,10,11,12)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.75    6.50    6.50    9.25   12.00
> summary(grass)
Rich  graze
Min. : 7.00  mow :5  1st Qu : 9.00
 unmow:4 Median :11.00   Mean :11.44
3rd Qu.:15.00  Max. :17.00
> summary(graze)
Length     Class       Mode
   9        character    character
```

> summary(grass$graze)

mow unmow

5 4

b) str():

It gives the structure of data object in terms of class of object, No. of observations and each variable class and sample data.

Example2:

```
> str(mtcars)
'data.frame':  32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

> str(grass)

'data.frame': 9 obs. of 2 variables:

$ rich : int 12 15 17 11 15 8 9 7 9

$ graze: Factor w/ 2 levels "mow","unmow": 1 1 1 1 1 2 2

c) Tail():

It gives the last 6 observations of the given data object.

Example3: > tail(iris)

```
> tail(mtcars)
                mpg cyl  disp  hp drat    wt qsec vs am gear carb
Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.7  0  1    5    2
Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.5  0  1    5    4
Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.5  0  1    5    6
Maserati Bora  15.0   8 301.0 335 3.54 3.570 14.6  0  1    5    8
Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.6  1  1    4    2
> tail(HairEyeColor,2)
[1] 7 8
> tail(state.x77,2)
          Population Income Illiteracy Life Exp Murder HS Grad Frost  Area
Wisconsin       4589   4468        0.7    72.48    3.0    54.5   149 54464
Wyoming          376   4566        0.6    70.29    6.9    62.9   173 97203
```

> tail(grass)

rich graze

4 11 mow

5 15 mow

6 8 unmow

7 9 unmow

8 7 unmow

9 9 unmow

d) **Head():**

It displays the top 6 observations from dataset

Example:

```
> head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> head(iris,2)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
```

```
>head(grass)
   rich graze
1  12 mow
2  15 mow
3   17 mow
4   11 mow
 5   15 mow
 6   8 unmow
```

e) Names():

It returns the coloum names

```
> names(mtcars)
 [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear" "carb"
```

```
>names(grass)
graze  rich
```

f) nrow():

It returns the number of observations in the given dataset.

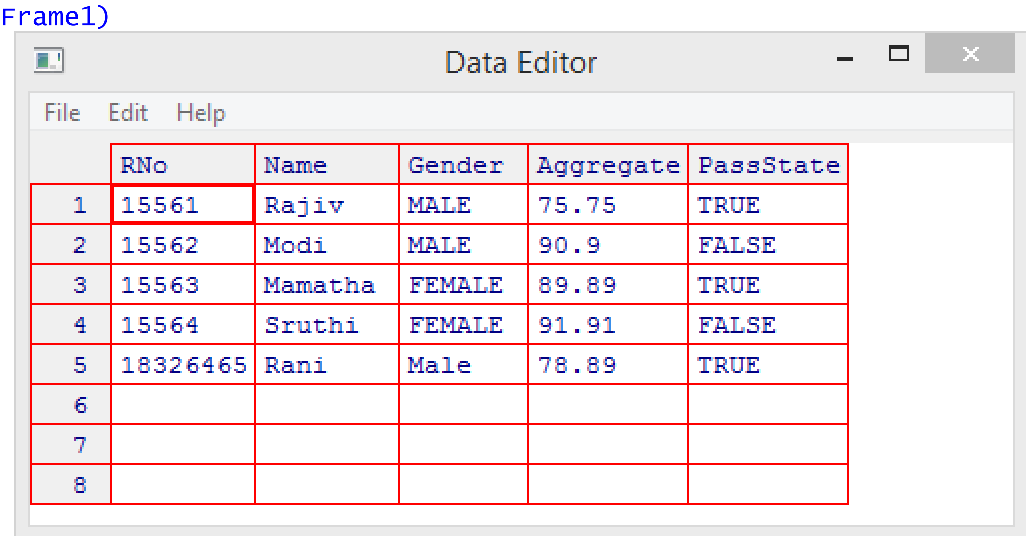
```
> dim(mtcars)
[1] 32 11
> nrow(mtcars)
[1] 32
> ncol(mtcars)
[1] 11
```

```
>nrow(iris)
9
```

g) fix(iris):

To fix the data in the given dataset.

```
> fix(mydFrame1)
```



Data Editor

File Edit Help

	RNo	Name	Gender	Aggregate	PassState
1	15561	Rajiv	MALE	75.75	TRUE
2	15562	Modi	MALE	90.9	FALSE
3	15563	Mamatha	FEMALE	89.89	TRUE
4	15564	Sruthi	FEMALE	91.91	FALSE
5	18326465	Rani	Male	78.89	TRUE
6					
7					
8					

h) With():

To replace $along with attribute names

i) Aggregate():

To get the summary statistic of specific column with respect to different levels in the class attribute.

aggregate(x ~ y, data, mean)

Here x is numeric and y is factor type

```
>aggregate(rich~graze, grass, mean)
   graze  rich
1  mow    14.00
2  unmow  8.25
```

j) Subset ():

To subset the data based on condition.

subset (data, x>7, select=c(x,y))

x is one of variable in data

select: to get the subset in specified order.

```
>subset(grass, rich>7, select=c(graze,rich))
  graze  rich
1  mow    12
2  mow    15
3  mow    17
4  mow    11
5  mow    15
6  unmow  8
7  unmow  9
9  unmow  9
```

Lab activity:

A researcher wants to understand the data collected by him about 3 species of flowers.

He wants the following:-

1. **The summary of 150 flower data including Sepal Length, Sepal Width, Petal Length and Petal Width. He also wants the summary of Sepal Length vs petal length.**

Solution:

To summarize data in R Studio we use majorly two functions Summary and Aggregate.

Using Summary command:

```
> summary_iris<- summary(iris)
> summary_iris
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species    ratio_s
epal_petal
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50   Min.
:1.050
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50   1st Qu.
:1.230
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50   Median
:1.411
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                   Mean
:2.018
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                   3rd Qu.
:3.176
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                   Max.
:4.833
```

We get Min, Max, 1st Quartile, 3rd Quartile, Median, Mean as an output of summary() command.

2. **He wants to understand the mean Petal Length of each species.**

Solution:

For getting detailed output of one or more functions we use aggregate() command.

Using Aggregate () command:

```
> aggregate(Sepal.Length~Species,iris,mean)
     Species Sepal.Length
1     setosa        5.006
2 versicolor        5.936
3  virginica        6.588
```

3. He wants to segregate the data of flowers having Sepal length greater than 7.

In the above example, we have calculated the mean sepal length of different species. Similarly we can calculate other functions also like frequency, median, summation etc.

For more details in terms of argument of Aggregate () command we use? Aggregate command to get help.

We also use subset () function to form subsets of data.

Using subset () command:

```
> sepalsub<- subset(iris,Sepal.Length>7)
> sepalsub
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species ratio_sepal_petal
103          7.1         3.0          5.9         2.1 virginica          1.203390
106          7.6         3.0          6.6         2.1 virginica          1.151515
108          7.3         2.9          6.3         1.8 virginica          1.158730
110          7.2         3.6          6.1         2.5 virginica          1.180328
118          7.7         3.8          6.7         2.2 virginica          1.149254
119          7.7         2.6          6.9         2.3 virginica          1.115942
123          7.7         2.8          6.7         2.0 virginica          1.149254
126          7.2         3.2          6.0         1.8 virginica          1.200000
130          7.2         3.0          5.8         1.6 virginica          1.241379
131          7.4         2.8          6.1         1.9 virginica          1.213115
132          7.9         3.8          6.4         2.0 virginica          1.234375
136          7.7         3.0          6.1         2.3 virginica          1.262295
```

4. He wants to segregate the data of flowers having Sepal length greater than 7 and Sepal width greater than 3 simultaneously.

Solution:

When we have to use more than 1 condition then we use & as shown below

```
> sepalsub<- subset(iris,Sepal.Length>7 & Sepal.Width>3)
> sepalsub
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species ratio_sepal_petal
110          7.2         3.6          6.1         2.5 virginica          1.180328
118          7.7         3.8          6.7         2.2 virginica          1.149254
126          7.2         3.2          6.0         1.8 virginica          1.200000
132          7.9         3.8          6.4         2.0 virginica          1.234375
>
```

5. He wants to view 1st 7 rows of data .

Solution:

For getting only few columns of requirement we use select () command in the argument:

```
> sepalsub<- subset(iris,Sepal.Length>7 & Sepal.Width>3,select=c(Sepal.Length,Sepal.Widt
h))
> sepalsub
    Sepal.Length Sepal.Width
110          7.2         3.6
118          7.7         3.8
126          7.2         3.2
132          7.9         3.8
```

For subsetting data without ant condition just based on rows and columns we use square brackets .

```
> iris11<- iris[1:7,]
> iris11
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species ratio_sepal_petal
1          5.1         3.5          1.4         0.2  setosa          3.642857
2          4.9         3.0          1.4         0.2  setosa          3.500000
3          4.7         3.2          1.3         0.2  setosa          3.615385
4          4.6         3.1          1.5         0.2  setosa          3.066667
5          5.0         3.6          1.4         0.2  setosa          3.571429
6          5.4         3.9          1.7         0.4  setosa          3.176471
7          4.6         3.4          1.4         0.3  setosa          3.285714
```

6. He wants to view 1st 3 rows and 1st 3 columns of data.

Solution:

```
> iris11<- iris[1:3,1:3]
> iris11
  Sepal.Length Sepal.Width Petal.Length
1          5.1         3.5          1.4
2          4.9         3.0          1.4
3          4.7         3.2          1.3
>
```

2.2 Basics of Probability

We shall introduce some of the basic concepts of probability theory by defining some terminology relating to *random experiments* (i.e., experiments whose outcomes are not predictable).

2.2.1. Terminology

Def. Outcome

The end result of an experiment. For example, if the experiment consists of throwing a die H or T, the outcome would be anyone of the six faces, F1,F2,F3,F4,F5,F6.

Def. Random experiment:

If an '**experiment**' is conducted for number of times, under essentially identical conditions, which has a set of all possible outcomes associated with it, if the result is not certain and is any one of the several possible outcomes is known as **Random Experiment.** In Simple an experiment whose outcomes are not known in advance.

Ex: Throwing a fair die, tossing a honest coin, measuring the noise voltage at the terminals of a resistor, etc.

Def. Sample space

The sample space of a random experiment is a mathematical abstraction that represent all possible outcomes of the experiment. We denote the sample space by S

Ex: In a random experiment of tossing 2 coins, S = {HH, HT, TH, TT}.

In the case a die, S = {1,2,3,4,5,6}

Each outcome of the experiment is represented by a point in S and is called a sample point. We use s (with or without a subscript), to denote a sample point. An event on the sample space is represented by an appropriate collection of sample point(s).

Def: Equally Likely Events:

Events are said to be equally likely when there is no reason to expect anyone of them rather than anyone of the others.

Def. Exhaustive Events

All possible events in any trial are known as Exhaustive events.

Ex:

In tossing a coin, there are two exhaustive elementary events, viz. Head and Tail.

In drawing 3 balls out 9 in a box, there are 9_{C_3} (9C3) exhaustive elementary events.

Def. Mutually exclusive (disjoint) events

Two events A and B are said to be mutually exclusive if they have no common elements (or outcomes).Hence if A and B are mutually exclusive, they cannot occur together. i.e., if the happening of any one of the events in a trial excludes the happening of any of others. (Two or more of the events can't happen simultaneously in the same trial.)

Ex: In tossing coin occurrence of the outcome 'Head' excludes the occurrence of 'Tail'.

Def. Classical Definition of Probability:

In a random experiment, let there be n mutually exclusive and equally likely elementary events. Let E be an event of the experiment. If m events are favorable to E, then the probability of E (Chance of occurrence of E) is defined as

$$p(E) = \frac{m}{n} = \frac{No.ofEventsFavourableToE}{TotalNo.ofEvents}$$

Note:

1. $0 \le \frac{m}{n} \le 1$
2. $0 \le p(E) \le 1$ and $0 \le p(\bar{E}) \le 1$

Random Variables – Distribution function:

A random variable, aleatory variable or stochastic variable is a variable whose value is subject to variations due to chance (i.e. randomness, in a mathematical sense).

Def: A ***random variable***, usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon

Let S be the sample of the random experiment. Random variable is a function whose domain is the set of outcomes w∈S and whose range is R, the set of real numbers. The random variable assigns a real value X(w) such that

1. The set {w/X(w)≤x} is an event for every x ∈ R, for which a probability is defined. This condition is called Measurability.
2. The probabilities of the events {w/X(w)=∞} and {w/X(w)=-∞} are equal to zero.
   i.e., p(X=∞) = p(X=-∞) = 0.
3. For A⊂S, there corresponds a set T⊂R called the image of A. Also for every T⊂R there exists in S the inverse image $X^{-1}(T) = \{w \in S \mid X(w) \in T\}$

In simple A random variable is a **real-valued function** defined on the points of a sample space.

Random variables are two broad categories:

- Random variable with discrete values
- Bivariate Random Variable

1. **Discrete Random Variable:** A random variable X which can take only a finite number of discrete values in an interval of domain is called a discrete random variable. In other words, if the random variable takes the values only on the set {0, 1,2,3,4,…n} is called Discrete Random variable.
   Ex: The printing mistakes in each page of a book, the number of tephone calls received by the receptionist are the examples of Discrete Random Variables.
   Thus to each outcome of S of a random experiment there corresponds a real number X(s) which is defined for each point of the sample S.
2. **Continuous Random Variable:** A Random variable X which can take values continuously i.e., which takes all possible values in a given interval is called a continuous random variable.
   Ex: the height, age and weight of individuals are the examples of continuous random variable
3. **Bivariate Random Variable:**

Bivariate Random Variables are those variables having only 2 possible outcomes.

Ex: Flip of coin(two outcomes: head/tail).

Probability distribution : Which describes how the values of a random variable are distributed. The probability distribution for a random variable X gives the possible values for X, and the probabilities associated with each possible value (i.e., the likelihood that the values will occur) The methods used to specify discrete prob. distributions are similar to (but slightly different from) those used to specify continuous prob. distributions.

Binomial distribution: The collection of all possible outcomes of a sequence of coin tossing

Normal distribution: The means of sufficiently large samples of a data population

Note: The characteristics of these theoretical distributions are well understood, they can be used to make Statistical inferences on the entire data population as a whole.

Example: Probability of ace of Diamond in a pack of 52 cards when 1 card is pulled out at random.

"At Random" means that there is no biased treatment

No. of Ace of Diamond in a pack = S = 1

Total no of possible outcomes = Total no. of cards in pack = 52

Probability of positive outcome = S/P = 1/52

That is we have 1.92% chance that we will get positive outcome.

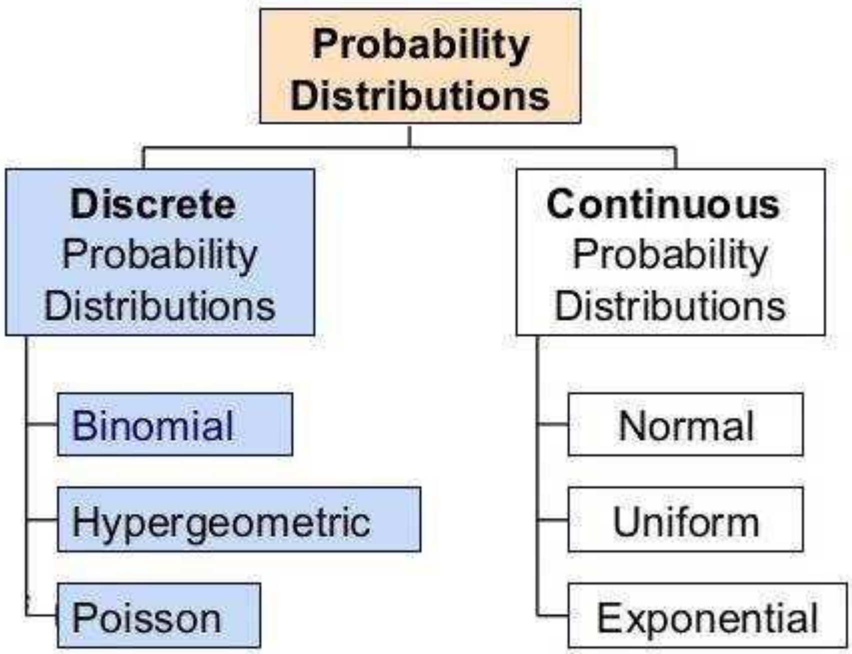
2.3.Expected value:

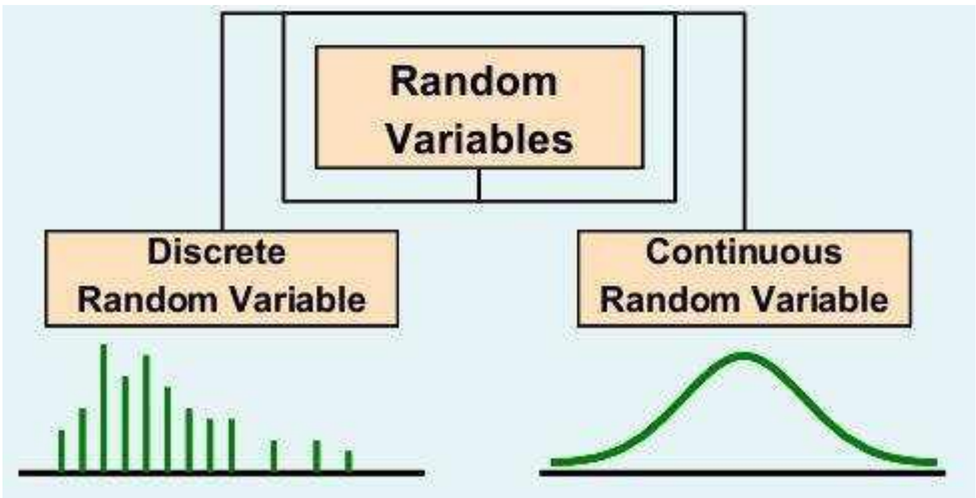
The expected value of a random variable is the long-run average value of repetitions of the experiment it represents.

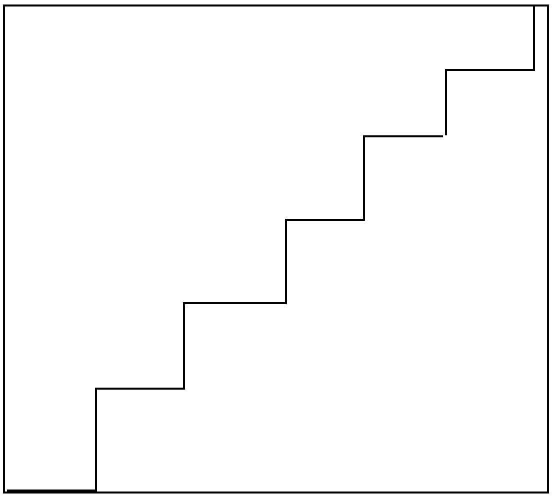
Example:

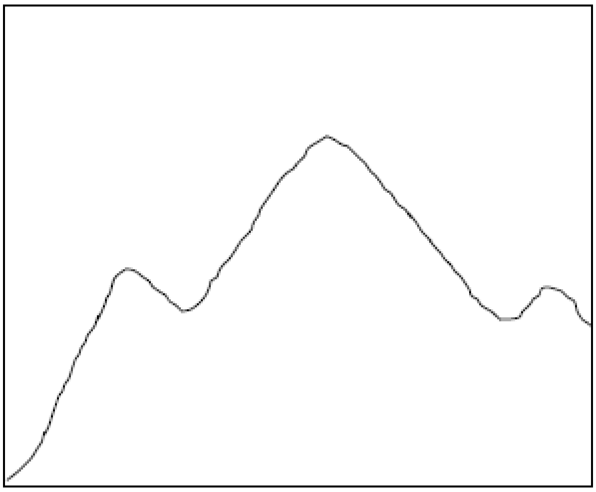
The expected value of a dice roll is 3.5 means the average of an extremely large number of dice rolls is practically always nearly equal to 3.5. Expected value is also known as the expectation, mathematical expectation, EV, mean, or first moment.

- **Expected value of a discrete random variable** is the probability-weighted average of all possible values
- **Continuous random variables are** the sum replaced by an integral and the probabilities by probability densities.









2.7.Probability Distribution Function (PDF):

It defines probability of outcomes based on certain conditions. Based on Conditions, there are majorly 5 types PDFs.

Types of Probability Distribution:

- Binomial Distribution
- Poisson Distribution
- Continuous Uniform Distribution
- Exponential Distribution
- Normal Distribution
- Chi-squared Distribution
- Student t Distribution
- F Distribution

Binomial Distribution

The **binomial distribution** is a discrete probability distribution. It describes the outcome of *n* independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure. If the probability of a successful trial is *p*, then the probability of having *x* successful outcomes in an experiment of *n* independent trials is as follows.

$$f(x) = (n_{c_x})\,p^x(1-p)^{(n-x)} \qquad \textit{Where } x = 0, 1, 2, \ . \ . \ . , n$$

Problem

Ex: Find the probability of getting 3 doublets when a pair of fair dice are thrown for 10 times.

Solution

n=no. of trials=10,
p=probability of success i.e., getting a doublet = 6/36=1/6
q=probability of failure=1-p=1-(1/6)=5/6
r=no. of successes expected=3

$$P(x=3) = (n_{c_x})\,p^x(1-p)^{(n-x)} = (n_{c_r})\,p^r(q)^{(n-r)}$$

$$= (10_{c_3})\,p^3(q)^{(10-3)}$$

$$= (10_{c_3})\left(\frac{1}{6}\right)^3\left(\frac{1}{6}\right)^{(10-3)} = 0.1550454$$

This can be computed in R as:

```
> choose(10,3)*((1/6)^3*(5/6)^7)  # Choose (10,3) is 10C3
[1] 0.1550454

This binomial distribution can be found using the formula in R as
> dbinom(3,size=10,prob=(1/6))
[1] 0.1550454
```

Problem: From the above problem find the probability of getting 3 or lesser doublets.

Solution:

```
>choose(10,0)*((1/6)^0*(5/6)^10  +
                 choose(10,1)*((1/6)^1*(5/6)^9) +
                                choose(10,2)*((1/6)^2*(5/6)^8) +
                                               choose(10,3)*((1/6)^3*(5/6)^7))
[1] 0.9302722
```

This can be obtained using cumulative binomial distribution function as

```
>pbinom(3,size=10,prob=(1/6),lower=T)
        [1] 0.9302722

>pbinom(3,size=10,prob=(1/6),lower=F) #probability of getting 4 or more doublets
        [1] 0.06972784

Note:> 0.9302722+ 0.06972784 = 1
```

Problem2

Suppose there are twelve multiple choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

Solution

Since only one out of five possible answers is correct, the probability of answering a question correctly by random is 1/5=0.2. We can find the probability of having exactly 4 correct answers by random attempts as follows.

dbinom(x,size,prob)

x: no. of successful outcomes (favourable)

size: n no.of independent trials

prob: probability of successful trial p

```
> dbinom(4, size=12, prob=0.2)
[1] 0.1329
```

To find the probability of having four or less correct answers by random attempts, we apply the function dbinom with $x = 0,\ldots,4$.

```
> dbinom (0, size=12, prob=0.2) +
      dbinom (1, size=12, prob=0.2) +
            dbinom(2, size=12, prob=0.2) +
                  dbinom(3, size=12, prob=0.2) +
                        dbinom(4, size=12, prob=0.2)
[1] 0.92744
```

Alternatively, we can use the cumulative probability function for binomial distribution pbinom.

```
> pbinom (4, size=12, prob=0.2)
[1] 0.92744
```

Answer:

The probability of four or less questions answered correctly by random in a twelve question multiple choice quiz is 92.7%.

Note: `dbinom` gives the density, `pbinom` gives the cumulative distribution function, `qbinom` gives the quantile function and `rbinom` generates random deviates. If `size` is not an integer, `NaN` is returned.

Poisson Distribution

The **Poisson distribution** is the probability distribution of independent event occurrences in an interval. If λ is the mean occurrence per interval, then the probability of having x occurrences within a given interval is:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \text{Where } x = 0, 1, 2, \ldots$$

Problem

If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

Solution

The probability of having *sixteen or less* cars crossing the bridge in a particular minute is given by the function ppois.

```
> ppois(16, lambda=12)   # lower tail
[1] 0.89871
```

Hence the probability of having seventeen or more cars crossing the bridge in a minute is in the *upper tail* of the probability density function.

```
> ppois(16, lambda=12, lower=FALSE)   # upper tail
[1] 0.10129
```

Similarly we can find the following:

```
> rpois(10, lambda=12)
 [1] 17 10  8 22  5 10 12 12  7 12
> dpois(16, lambda=12)
[1] 0.05429334
```

Answer

If there are twelve cars crossing a bridge per minute on average, the probability of having seventeen or more cars crossing the bridge in a particular minute is 10.1%.

Normal Distribution

The **normal distribution** is defined by the following probability density function, where μ is the population mean and σ^2 is the variance.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable X follows the normal distribution, then we write:

$X \sim N(\mu, \sigma^2)$

In particular, the normal distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard normal distribution*, and is denoted as $N(0,1)$. It can be graphed as follows.

Figure 1: Normal Distribution

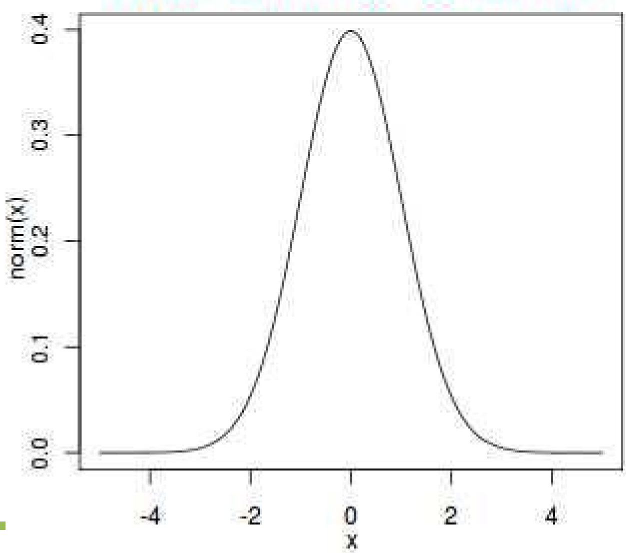


Figure 1 shows the normal distribution of sample data. The shape of a normal curve is highly dependent on the standard deviation.

Importance of Normal Distribution:

- Normal distribution is a continuous distribution that is "bell-shaped".
- Data are often assumed to be normal.
- Normal distributions can estimate probabilities over a ***continuous interval of data values.***

Properties:

The normal distribution f(x), with any mean μ and any positive deviation σ, has the following properties:

- It is symmetric around the point x = μ, which is at the same time the mode, the median and the mean of the distribution.
- It is unimodal: its first derivative is positive for x < μ, negative for x > μ, and zero only at x = μ.
- Its density has two inflection points (where the second derivative of is zero and changes sign), located one standard deviation away from the mean as x = μ − σ and x = μ + σ.
- Its density is log-concave.
- Its density is infinitely differentiable, indeed super smooth of order 2.

Its second derivative f′′(x) is equal to its derivative with respect to its variance σ2.

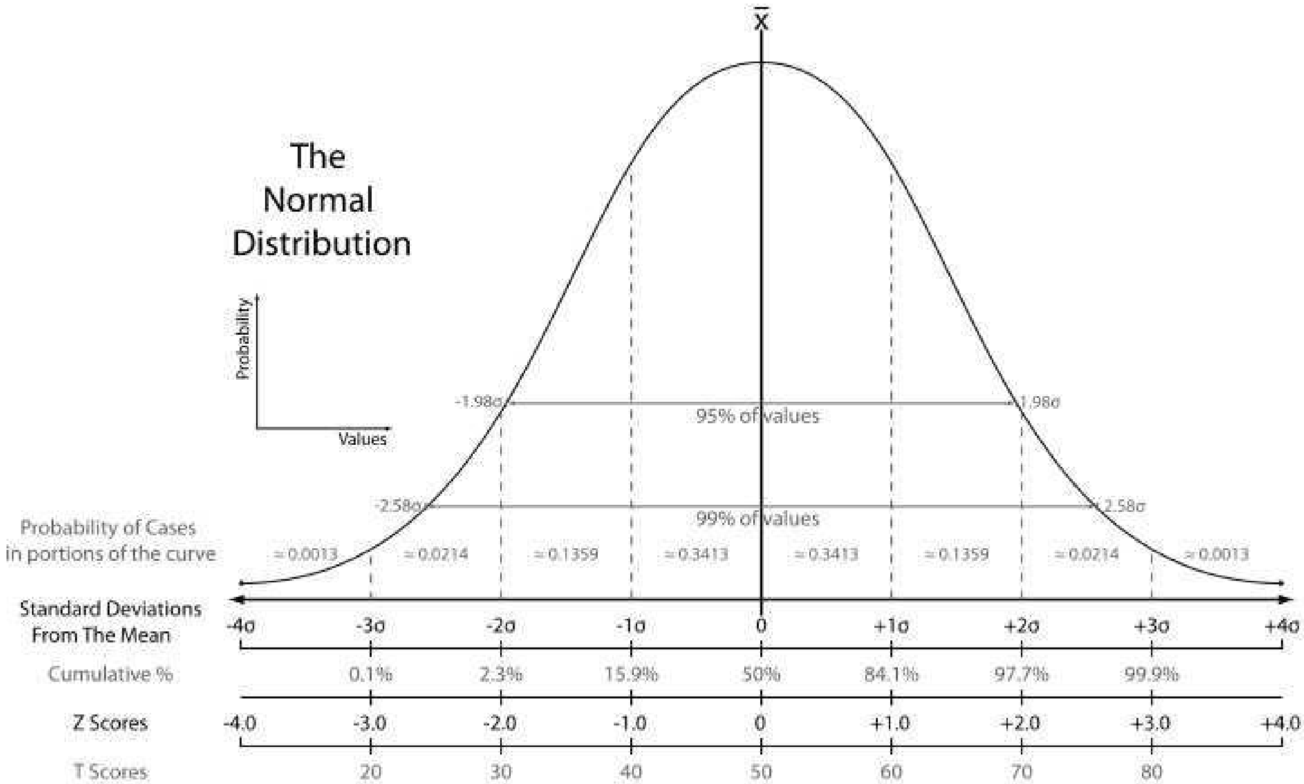


Figure 2: A normal distribution with Mean=0 and Standard deviation = 1

Normal Distribution in R:

Description:

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.

The normal distribution is important because of the **Central Limit Theorem**, which states that the population of all possible samples of size *n* from a population with mean μ and variance σ^2 approaches a normal distribution with mean μ and σ^2/n when *n* approaches infinity.

Problem exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

Solution

We apply the function pnorm of the normal distribution with mean 72 and standard deviation 15.2. Since we are looking for the percentage of students scoring higher than 84, we are interested in the *upper tail* of the normal distribution.

```
> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE)
[1] 0.21492
```

Answer

The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
 rnorm(n, mean = 0, sd = 1)
```

Arguments

- **x, q vector of quantiles.**
- **P vector of probabilities.**
- **N number of observations. If length(n) > 1, the length is taken to be the number required.**
- **Mean vector of means.**
- **Sd vector of standard deviations.**
- **log, log. P logical; if TRUE, probabilities p are given as log(p).**
- **lower.tail logical; if TRUE (default), probabilities are** $P[X \le x]$ **otherwise,** $P[X > x]$.
- **rnorm(n, mean = 0, sd = 1) as default**

The Central Limit Theorem

The central limit theorem and the law of large numbers are the two *fundamental theorems* of probability. Roughly, the central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution. The importance of the central limit theorem is hard to overstate; indeed it is the reason that many statistical procedures work.

The CLT says that if you take many repeated samples from a population, and calculate the *averages* or *sum* of each one, the collection of those averages will be normally distributed… and it doesn't matter what the shape of the source distribution is!

Lab Activity:

To generates 20 numbers with a mean of 5 and a standard deviation of 1:

```
> rnorm(20, mean = 5, sd = 1)
[1] 5.610090 5.042731 5.120978 4.582450 5.015839 3.577376 5.159308 6.496983
[9] 3.071729 6.187525 5.027074 3.517274 4.393562 3.866088 4.533490 6.021554
[17] 5.359491 5.265780 3.817124 5.855315
> pnorm(5, mean = 5, sd = 1)
[1] 0.5
> qnorm(0.5, 5, 1)
[1] 5
> dnorm(c(4,5,6), mean = 5, sd = 1)
[1] 0.2419707 0.3989423 0.2419707
```

Lab Activity 3: Probability Theories:

1. If you throw a dice 20 times then what is the probability that you get following results:

a. 3 sixes

Solution:

```
>
> dbinom(x=3,20,prob=1/6)
[1] 0.2378866
>
```

b. 6 sixes

```
> dbinom(x=6,20,prob=1/6)
[1] 0.06470515
>
```

c. 1,2 and 3 sixes

```
>
> pbinom(q=3,20,prob=1/6,lower.tail=T)
[1] 0.5665456
>
```

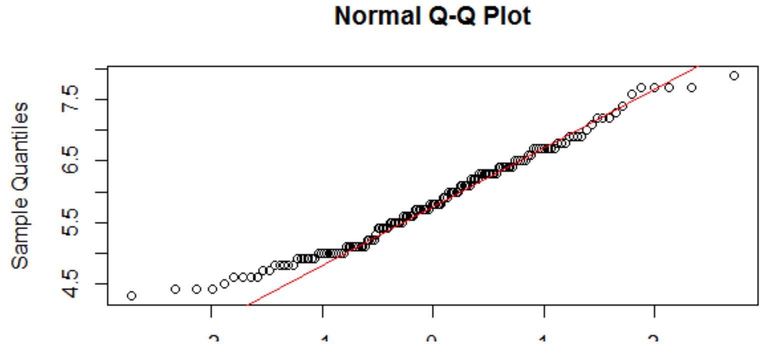
2. In Iris data set check whether Sepal Length is normally distributed or not.

Use : To find if the Sepal Length is normally distributed or not we use 2 commands- qqnorm() &qqline().

```
>
> qqnorm(iris$Sepal.Length)
> qqline(iris$Sepal.Length,col='red')
>
```

The qqnorm() shows the actual distribution of data while qqline() shows the line on which data would lie if the data is normally distributed. The deviation of plot from line shows that data is not normally distributed.

Figure3: Normal distribution of iris$ Sepal Length



3. Prove that population mean of Sepal length is different from mean of 1st 10 data significantly.

T-Test of sample subset of Iris data set.

```
> mean(iris$Sepal.Length)
[1] 5.843333
> iris.sub<- iris[1:10,1:1]
> t.test(iris.sub,alternative='less',mu=5.843)

        One Sample t-test

data:  iris.sub
t = -10.669, df = 9, p-value = 1.041e-06
alternative hypothesis: true mean is less than 5.843
95 percent confidence interval:
     -Inf 5.028894
sample estimates:
mean of x
     4.86
```

Here p-value is much less than 0.05. So we reject the null hypothesis and we accept the alternate hypothesis which says that mean of sample is less than the population mean.

$\mu_s < \mu_p$

Also sample mean is 4.86 and degree of freedom if 9 which is sample size -1.

Similarly we can do two sided test by writing alternative= “two sided”. And also paired sample t-test by using paired=TRUE as the part of argument.

Work effectively with Colleagues (NOS 9002)
Introduction to work effectively, Team Work, Professionalism, Effective Communication skills, etc.

Refer Students Hand Book and the ppt issued.

Introduction to work effectively

Team Work

1. Ways to Be More Effective at Work
   a) Trim Your Task List
   b) Swap Your To-Do List for a Schedule
   c) Stop While You're Still On a Roll
   d) Stay Organized
   e) Make Bad Habits More Difficult to Indulge(Spoil).
   f) Prioritize
   g) Tackle Your Most Important Tasks First
   h) Plan Tomorrow Tonight
   i) Use Idle Time to Knock Out Admin Tasks
   j) Schedule Meetings With Yourself
   k) Change Your Self-Talk
   l) Communicate and Clarify
   m) Find Ways to Do More of the Work You Enjoy

Team Work

1. What is team work?
2. How is it more advantageous?

A team comprises a group of people linked in a common purpose.

- Teams are especially appropriate for conducting tasks that are high in complexity and have many interdependent subtasks.
- Coming together is a beginning, keeping together is progress and working together is success.
- A team is a number of people associated together in work or activity.
- In a good team members create an environment that allows everyone to go beyond their limitation.

Why do we need teamwork –to make the organization profitable.

Team work vs. Individual work

- Team Work:
- Work Agree on goals/milestones
- Establish tasks to be completed
- Communicate / monitor progress
- Solve Problem
- Interpret Results
- Agree completion of projects

Individual work

- Work on tasks
- Work on new / revised tasks

Team Development

- Team building is any activity that builds and strengthens the team as a team.

Team building fundamentals

- Clear Expectations – Vision/Mission
- Context – Background – Why participation in Teams?
- Commitment – dedication – Service as valuable to Organization & Own
- Competence – Capability – Knowledge
- Charter – agreement – Assigned area of responsibility
- Control – Freedom & Limitations
- Collaboration – Team work
- Communication

- Creative Innovation
- Consequences – Accountable for rewards
- Coordination
- Cultural Change

Roles of team member

- Communicate
- Don't Blame Others
- Support Group Member's Ideas
- No Bragging(Arrogant) – No Full of yourself
- Listen Actively
- Get Involved
- Coach, Don't Demonstrate
- Provide Constructive Criticism
- Try To Be Positive
- Value Your Group's Ideas

Team Work: Pros and Cons

Summary:

- A team comprises a group of people linked in a common purpose.
- · Team work is essential to the success of every organization. In a good team, members create an environment that allows everyone to go beyond their limitation.
- · Some of the fundamentals on which a team is built are: Collaboration, Clear Expectations and Commitment

Professionalism

- Professionalism is the competence or set of skills that are expected from a professional.
- Professionalism determines how a person is perceived by his employer, co-workers, and casual contacts.
- How long does it take for someone to form an opinion about you?
- Studies have proved that it just takes six seconds for a person to form an opinion about another person.

How does someone form an opinion about you?

- Eye Contact – Maintaining eye contact with a person or the audience says that you are confident. It says that you are someone who can be trusted and hence can maintain contact with you.
- Handshake – Grasp the other person's hand firmly and shake it a few times. This shows that you are enthusiastic.
- Posture – Stand straight but not rigid, this will showcase that you are receptive and not very rigid in your thoughts.
- Clothing – Appropriate clothing says that you are a leader with a winning potential.

How to exhibit professionalism?

- Empathy (compassion)
- Positive Attitude
- Teamwork
- Professional Language
- Knowledge
- Punctual
- Confident
- Emotionally stable

Grooming

What are the colours that one can opt for work wear?

- A good rule of thumb is to have your pants, skirts and blazers in neutral colours. Neutrals are not only restricted to grey brown and off white - you can also take advantage of the beautiful navies, forest greens, burgundies, tans and caramel tones around.
- Pair these neutrals with blouses, scarves or other accessories in accent colours - ruby red, purple, teal blue, soft metallic and pinks are some examples.

Things to remember

- Wear neat clothes at work which are well ironed and do not stink.
- Ensure that the shoes are polished and the socks are clean
- Cut your nails on a regular basis and ensure that your hair is in place.
- Women should avoid wearing revealing clothes at work.
- Remember that the way one presents oneself plays a major role in the professional world

Effective Communication

- Effective communication is a mutual understanding of the message.
- Effective communication is essential to workplace effectiveness
- The purpose of building communication skills is to achieve greater understanding and meaning between people and to build a climate of trust, openness, and support.
- A big part of working well with other people is communicating effectively.

Things to remember

- Wear neat clothes at work which are well ironed and do not stink.
- Ensure that the shoes are polished and the socks are clean
- Cut your nails on a regular basis and ensure that your hair is in place.
- Women should avoid wearing revealing clothes at work.
- Remember that the way one presents oneself plays a major role in the professional world

Effective Communication

- Effective communication is a mutual understanding of the message.
- Effective communication is essential to workplace effectiveness
- The purpose of building communication skills is to achieve greater understanding and meaning between people and to build a climate of trust, openness, and support.
- A big part of working well with other people is communicating effectively.
- Sometimes we just don't realize how critical effective communication is to getting the job done.

What is Effective Communication?

- We cannot not communicate.
- **The question is: Are we communicating what we intend to communicate?**
- **Does the message we send match the message the other person receives?**
- Impression = Expression
- Real communication or understanding happens only when the receiver's impression matches what the sender intended through his or her expression.
- So the goal of effective communication is a mutual understanding of the message.

Forms of Communication

1. Verbal communication
2. Non verbal communication
3. Written communication

Verbal Communication :

- Verbal communication refers to the use of sounds and language to relay a message
- It serves as a vehicle for expressing desires, ideas and concepts and is vital to the processes of learning and teaching.
- verbal communication acts as the primary tool for expression between two or more people

Types of verbal communication

- Interpersonal communication and public speaking are the two basic types of verbal communication.
- Whereas public speaking involves one or more people delivering a message to a group
- interpersonal communication generally refers to a two-way exchange that involves both talking and listening.

Forms of non verbal communication

1. Ambulation is the way one walks
2. Touching is possibly the most powerful nonverbal communication form.
3. Eye contact is used to size up the trustworthiness of another.
4. 4.Posturing can constitute a set of potential signals that communicate how a person is experiencing the environment
5. Tics are involuntary nervous spasms that can be a key to indicate one is being threatened.
6. Sub-vocals are the non-words one says, such as "ugh" or "um." They are used when one is trying to find the right word..
7. Distancing is a person's psychological space. If this space is invaded, one can become somewhat tense, alert, or "jammed up.
8. Tics are involuntary nervous spasms that can be a key to indicate one is being threatened.
9. Sub-vocals are the non-words one says, such as "ugh" or "um." They are used when one is trying to find the right word..
10. Distancing is a person's psychological space. If this space is invaded, one can become somewhat tense, alert, or "jammed up.
11. Gesturing carries a great deal of meaning between people, but different gestures can mean different things to the sender and the receiver. This is especially true between cultures. Still, gestures are used to emphasize our words and to attempt to clarify our meaning.
12. Vocalism is the way a message is packaged and determines the signal that is given to another person. For example, the message, "I trust you," can have many meanings. "I trust you" could imply that someone else does not. "I trust you" could imply strong sincerity. "I trust you" could imply that the sender does not trust others.

Written Communication

- Written communication involves any type of message that makes use of the written word. Written communication is the most important and the most effective of any mode of business communication

- Examples of written communications generally used with clients or other businesses include email, Internet websites, letters, proposals, telegrams, faxes, postcards, contracts, advertisements, brochures, and news releases

Advantages and disadvantages of written communication:

Advantages

- Creates permanent record
- Allows to store information for future reference
- Easily distributed
- All recipients receive the same information
- Written communication helps in laying down apparent principles, policies and rules for running on an organization.
- It is a permanent means of communication. Thus, it is useful where record maintenance is required.
- Written communication is more precise and explicit
- Effective written communication develops and enhances organization's image
- It provides ready records and references
- Written communication is more precise and explicit.
- Effective written communication develops and enhances an organization's image
- Necessary for legal and binding documents

Disadvantages of Written Communication

- Written communication does not save upon the costs. It costs huge in terms of stationery and the manpower employed in writing/typing and delivering letters.
- Also, if the receivers of the written message are separated by distance and if they need to clear their doubts, the response is not spontaneous.
- - Written communication is time-consuming as the feedback is not immediate. The encoding and sending of message takes time.

2 – Ensuring Connectivity

- The content that comprises a piece of writing should reflect fluency and should be connected through a logical flow of thought, in order to prevent misinterpretation and catch the attention of the reader.
- Moreover, care should be taken to ensure that the flow is not brought about through a forced/deliberate use of connectives, as this make the piece extremely uninteresting and artificial.

3 – Steering Clear of Short Form

- People may not be aware of the meaning of various short forms and may thus find it difficult to interpret them.
- Moreover, short forms can at time be culture specific or even organization specific and may thus unnecessarily complicate the communication.

4 – Importance of Grammar, Spelling and Punctuation

- Improper grammar can at worst cause miscommunication and at least results in unwanted humour and should be thus avoided. So too, spellings can create the same effect or can even reflect a careless attitude on part of the sender.
- Finally, effective use of punctuations facilitates reading and interpretation and can in rare cases even prevent a completely different meaning, which can result in miscommunication

5 – Sensitivity to the Audience

- One needs to be aware of and sensitive to the emotions, need and nature of the audience in choosing the vocabulary, content, illustrations, formats and medium of communication, as a discomfort in the audience would hamper rather than facilitate communication.

6 – Importance of Creativity

- In order to hold the readers' attention one needs to be creative to break the tedium of writing and prevent monotony from creeping in.
- This is especially true in the case of all detailed writing that seeks to hold the readers' attention.

7 – Avoidance Excessive use of Jargons

- Excessive use of jargon(slang/terminology)

can put off a reader, who may not read further, as, unlike a captive audience, the choice of whether to participate in the communication rests considerably with the reader.

Go through the Facilitators Guide for Objective Questions/True or False Qns..They will be given for weekly tests.

***** End of Unit-2 *****

Unit III: SQL using R

3.1. Introduction to NoSQL:

Define Nosql Database:

NoSQL is originally referring to "non SQL" or "non-relational" and also called "Not only SQL" to emphasize that they may support SQL-like query languages. The RDBMS database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. NoSQL databases are increasingly used in big data and real-time web applications.

Benefits of NoSQL Database:

No SQL databases are more scalable and provide superior performance. The NoSQL data model addresses several issues that the relational model is not designed to address:

- Large volumes of structured, semi-structured, and unstructured data
- Agile sprints, quick iteration, and frequent code pushes
- Object-oriented programming that is easy to use and flexible
- Efficient, scale-out architecture instead of expensive, monolithic architecture

Classification of NoSQL databases based on data model: A basic classification based on data model, with examples:

- **Document**: Clusterpoint, Apache CouchDB, Couchbase, DocumentDB, HyperDex, Lotus Notes, MarkLogic, **MongoDB**, OrientDB, Qizx
- **Key-value**: CouchDB, Oracle NoSQL Database, Dynamo, FoundationDB, HyperDex, MemcacheDB, Redis, Riak, FairCom c-treeACE, Aerospike, OrientDB, MUMPS
- **Graph**: Allegro, Neo4J, InfiniteGraph, OrientDB, Virtuoso, Stardog
- **Multi-model**: OrientDB, FoundationDB, ArangoDB, Alchemy Database, CortexDB

Differences between SQL database and NoSQL Database:

		SQL Databases	NOSQL Databases
1.	**Types**	One type (SQL database) with minor variations	Many different types including key-value stores, document databases, wide-column stores, and graph databases
2.	**Development History**	Developed in 1970s to deal with first wave of data storage applications	Developed in 2000s to deal with limitations of SQL databases, particularly concerning scale, replication and unstructured data storage
3.	**Examples**	MySQL, Postgres, Oracle Database	MongoDB, Cassandra, HBase, Neo4j

4.	**Data Storage Model**	Individual records (e.g., "employees") are stored as rows in tables, with each column storing a specific piece of data about that record (e.g., "manager," "date hired," etc.), much like a spreadsheet. Separate data types are stored in separate tables, and then joined together when more complex queries are executed. For example, "offices" might be stored in one table, and "employees" in another. When a user wants to find the work address of an employee, the database engine joins the "employee" and "office" tables together to get all the information necessary.	Varies based on database type. For example: • key-value stores function similarly to SQL databases, but have only two columns ("key" and "value"), with more complex information sometimes stored within the "value" columns. • Document databases do away with the table-and-row model altogether, storing all relevant data together in single "document" in **JSON, XML**, or another format, which can nest values hierarchically.
5.	**Schemas**	Structure and data types are fixed in advance. To store information about a new data item, the entire database must be altered, during which time the database must be taken offline.	Typically dynamic. Records can add new information on the fly, and unlike SQL table rows, dissimilar data can be stored together as necessary. For some databases (e.g., wide-column stores), it is somewhat more challenging to add new fields dynamically.
6.	**Scaling**	Vertically, meaning a single server must be made increasingly powerful in order to deal with increased demand. It is possible to spread SQL databases over many servers, but significant additional engineering is generally required.	Horizontally, meaning that to add capacity, a database administrator can simply add more commodity servers or cloud instances. The database automatically spreads data across servers as necessary.
7.	**Development Model**	Mix of open-source (e.g., Postgres, MySQL) and closed source (e.g., Oracle Database)	Open-source
8.	**Supports Transactions**	Yes, updates can be configured to complete entirely or not at all	In certain circumstances and at certain levels (e.g., document level vs. database level)

9.	**Data Manipulation**	Specific language using Select, Insert, and Update statements, e.g. SELECT fields FROM table WHERE…	Through object-oriented APIs
10.	**Consistency**	Can be configured for strong consistency	Depends on product. Some provide strong consistency (e.g., MongoDB) whereas others offer eventual consistency (e.g., Cassandra)

Unit III: SQL using R

3.2. Connecting R to NoSQL databases

Lab Activity: No SQL Example: R script to access a XML file:

Step1: Install Packages plyr,XML

Step2: Take xml file url

Step3: create XML Internal Document type object in R using **xmlParse()**

Step4 :Convert xml object to list by using **xmlToList()**

Step5**:** convert list object to data frame by using **ldply(xl, data.frame)**

```
install.packages("XML")
install.packages("plyr")
> fileurl<-"http://www.w3schools.com/xml/simple.xml"
> doc<-xmlParse(fileurl,useInternalNodes=TRUE)
> class(doc)
[1] "XMLInternalDocument" "XMLAbstractDocument"
> doc
<?xml version="1.0" encoding="UTF-8"?>
<breakfast_menu>
  <food>
    <name>Belgian Waffles</name>
    <price>$5.95</price>
    <description>Two of our famous Belgian Waffles with plenty of real maple syrup</description>
    <calories>650</calories>
  </food>
  <food>
> xl<-xmlToList(doc)
> class(xl)
[1] "list"
> xl
$food
$food$name
[1] "Belgian Waffles"
$food$price
[1] "$5.95"
$food$description
```

```
[1] "Two of our famous Belgian Waffles with plenty of real maple syrup"
$food$calories
[1] "650"
$food
> data<-ldply(xl, data.frame)
> head(data)
   .id                        name price
1 food             Belgian Waffles $5.95
2 food   Strawberry Belgian Waffles $7.95
3 food Berry-Berry Belgian Waffles $8.95
4 food                French Toast $4.50
5 food         Homestyle Breakfast $6.95

Description:
1   Two of our famous Belgian Waffles with plenty of real maple syrup
2   Light Belgian waffles covered with strawberries and whipped cream
3   Light Belgian waffles covered with an assortment of fresh berries and whipped cream
4   Thick slices made from our homemade sourdough bread
5   Two eggs, bacon or sausage, toast, and our ever-popular hash browns
    calories
1     650
2     900
3     900
4     600
5     950
```

3.3. Excel and R integration with R connector

Different approaches in R to connect with Excel to perform read write and execute activities:

3.3.1. Read Excel spreadsheet in R:

Multiple Packages are available to access Excel sheet from R

1. **gdata:** This package requires you to install additional Perl libraries on Windows platforms but it's very powerful.

   ```
   require(gdata)
   myDf <- read.xls ("myfile.xlsx"), sheet = 1, header = TRUE)
   ```

2. **XLConnect:** It might be slow for large dataset but very powerful otherwise.

   ```
   require (XLConnect)
   ```

```
wb <- loadWorkbook("myfile.xlsx")
myDf <- readWorksheet(wb, sheet = "Sheet1", header = TRUE)
```

3. **xlsx:** This package requires JRM to install. This is suitable for java supported envirments.Prefer the read.xlsx2() over read.xlsx(), it's significantly faster for large dataset.

```
require(xlsx)
read.xlsx2("myfile.xlsx", sheetName = "Sheet1")
```

Lab activity : example R script:

```
install.packages("rjava")
install.packages("xlsx")
require(xlsx)
> read.xlsx2("myfile.xlsx", sheetName = "Sheet1")
```

Sno	Sname	Marks	Attendance	Contactno	Mailid
1	sri	45	45	988776655	s@mymail.com
2	vas	78	78	435465768	v@mymail.com
3	toni	34	46	-117845119	s@mymail.com
4	mac	90	89	-671156006	v@mymail.com
5	ros	25	23	-1224466893	s@mymail.com

xlsReadWrite: Available for Windows only. It's rather fast but doesn't support. **.xlsx** files which is a serious drawback. It has been removed from CRAN lately.

read.table("clipboard"): It allows to copy data from Excel and read it directly in R. This is the quick and dirty R/Excel interaction but it's very useful in some cases.

```
myDf<- read.table("clipboard")
```

3.3.2. Read R output in Excel:

First create a csv output from an R **data.frame** then read this file in Excel. There is one function that you need to know it's **write.table**. You might also want to consider: **write.csv** which uses "." for the decimal point and a comma for the separator and **write.csv2** which uses a comma for the decimal point and a semicolon for the separator.

```
x <- cbind(rnorm(20),runif(20))
colnames(x) <- c("A","B")
write.table(x,"your_path",sep=",",row.names=FALSE)
```

3.3.3. Execute R code in VBA:

RExcel is from my perspective the best suited tool but there is at least one alternative. You can run a batch file within the VBA code. If R.exe is in your PATH, the general syntax for the batch file (.bat)
R CMD BATCH [options] myRScript.R

Here's an example of how to integrate the batch file above within your VBA code.

3.3.4. Execute R code from an Excel spreadsheet

Rexcel is the only tool I know for the task. Generally speaking once you installed RExcel you insert the excel code within a cell and execute from RExcel spreadsheet menu. See the RExcel references below for an example.

a) Execute VBA code in R

This is something I came across but I never tested it myself. This is a two steps process. First write a VBscript wrapper that calls the VBA code. Second run the VBscript in R with the system or shell functions. The method is described in full details here.

b) Fully integrate R and Excel

RExcel is a project developped by Thomas Baier and Erich Neuwirth, "making R accessible from Excel and allowing to use Excel as a frontend to R". It allows communication in both directions: Excel to R and R to Excel and covers most of what is described above and more. I'm not going to put any example of RExcel use here as the topic is largely covered elsewhere but I will show you where to find the relevant information. There is a wiki for installing RExcel and an excellent tutorial available here. I also recommend the following two documents: RExcel – Using R from within Excel and High-Level Interface Between R and Excel. They both give an in-depth view of RExcel capabilities.

Introduction to Analytics (Associate Analytics – I)

UNIT IV: Correlation and Regression Analysis (NOS 9001)

4.1. Regression Analysis:

Regression modeling or analysis is a statistical process for estimating the relationships among variables. The main focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').The value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable.

In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is 1. Mathematically a linear relationship represents a straight line when plotted as a graph. A non-linear relationship where the exponent of any variable is not equal to 1 creates a curve.

The general mathematical equation for a linear regression is −

```
y = mx + c
```

Following is the description of the parameters used −

y is the response variable.

x is the predictor variable.

m(slope) and c(intercept) are constants which are called the coefficients.

In R , lm () function to do simple regression modeling.

The simple linear equation Y=mX+C , intercept "C" and the slope "m" . The below plot shows the linear regression.

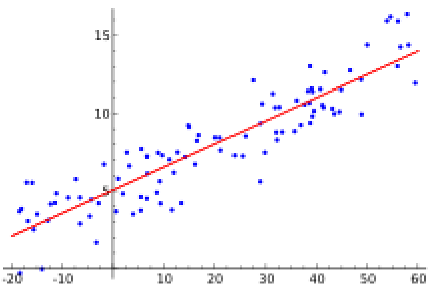


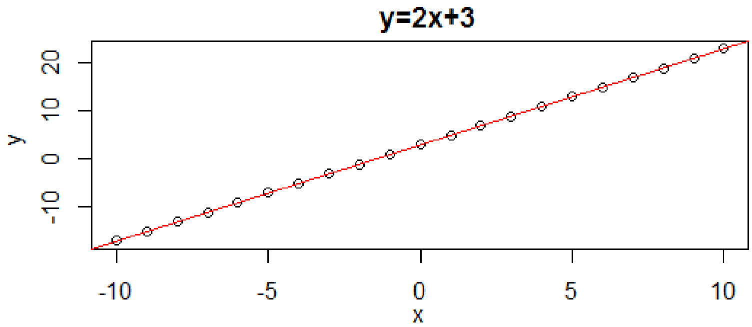
Figure Linear Regression plot

Example:

#Let us consider the equation y=mx+c with the sample values m=2, c=3

#Hence y=2x+3 will takes the following values

```
> x<-c(-10,-9,-8,-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7,8,9,10)
> y<-c(-17,-15,-13,-11,-9,-7,-5,-3,-1,1,3,5,7,9,11,13,15,17,19,21,23)
  # OR You can take y as #  > y<- 2*x+3
> relationxy<-lm(y~x)
> relationxy

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
          3            2
> plot(x,y)
> abline(relationxy,col='red')
> title(main="y=2x+3")
```



Example: (Case study #1)

Lab activity: Linear Regression for finding the relation between petal length and petal width in IRIS dataset:

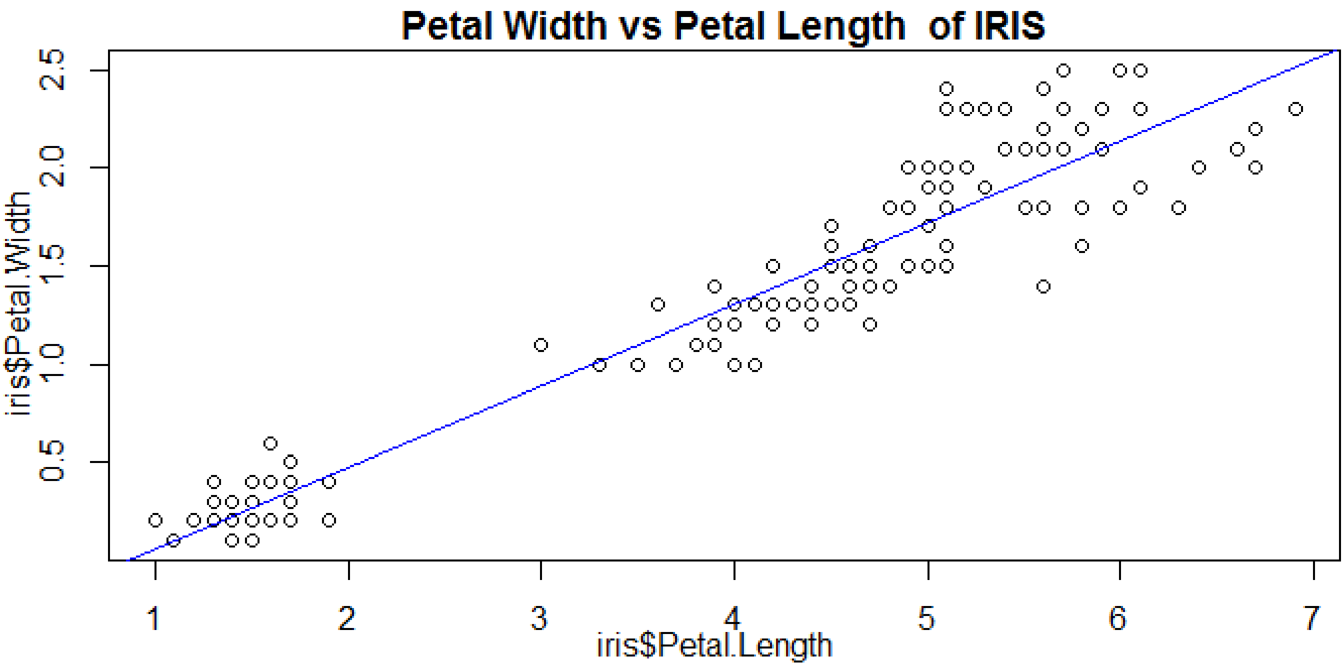
```
> fit <- lm(iris$Petal.Length ~ iris$Petal.Width)
> fit
Call:
lm(formula = iris$Petal.Length ~ iris$Petal.Width)
Coefficients:
     (Intercept)          iris$Petal.Width
           1.084                2.230
```

We get the intercept "C" and the slope "m" of the equation – Y=mX+C. Here m=2.230 and C=1.084 now we found the linear equation between petal length and petal width is

iris$Petal.Length=2.230* iris$Petal.Width+1.084

We can observe the plot and line below as

```
> plot(iris$Petal.Length,iris$Petal.Width)
> abline(lm(iris$Petal.Width~iris$Petal.Length ),col="blue")
> title(main="Petal Width vs Petal Length  of IRIS")
```



Example: (Case study #2)

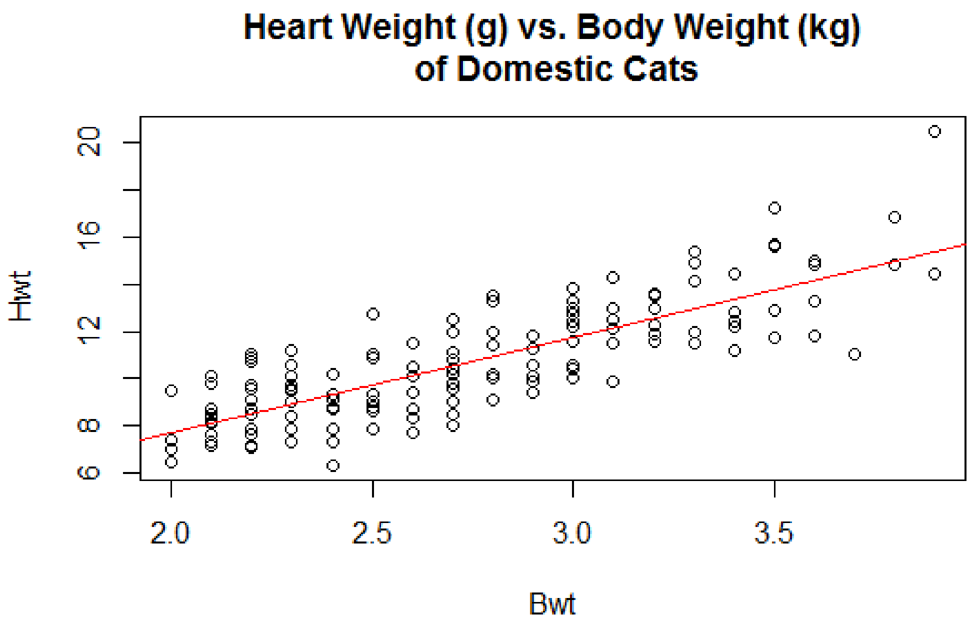
Relation between Heart weight and body weight of cats

- Generate a simple linear regression equation in two variables of ***cats*** dataset. The two variables are Heart Weight and Body Weight of the cats being examined in the research.
- Also find out if there is any relation between Heart Weight and Body Weight.
- Now check if Heart weight is affected by any other factor or variable.
- Find out how Heart Weight is affected by Body Weight and Sex together using Multiple Regression.

```
> library(MASS)   #  MASS Contains cats  #
> data(cats)   # data() was originally intended to allow users to load datasets from packages for use in their examples and as such it loaded the datasets into the work space. That need has been almost entirely superseded by lazy-loading of datasets.
> str(cats)
     'data.frame':    144 obs. of  3 variables:
      $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
      $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
      $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...

> summary(cats)
      Sex         Bwt              Hwt
      F:47    Min.   :2.000    Min.   : 6.30
      M:97    1st Qu.:2.300    1st Qu.: 8.95
              Median :2.700    Median :10.10
              Mean   :2.724    Mean   :10.63
              3rd Qu.:3.025    3rd Qu.:12.12
              Max.   :3.900    Max.   :20.50
```

"Bwt" is the body weight in kilograms, "Hwt" is the heart weight in grams, and "Sex" should be obvious. There are no missing values in any of the variables, so we are ready to begin by looking at a scatterplot.

```
> attach(cats)     # This works better rather using data(cats)
> plot(Bwt, Hwt)
> title(main="Heart Weight (g) vs. Body Weight (kg)\n of Domestic Cats")
> lmout<-lm(Hwt~Bwt)
> abline(lmout,col='red')
> lmout
Call:
lm(formula = Hwt ~ Bwt)
Coefficients:
(Intercept)          Bwt
    -0.3567       4.0341
```

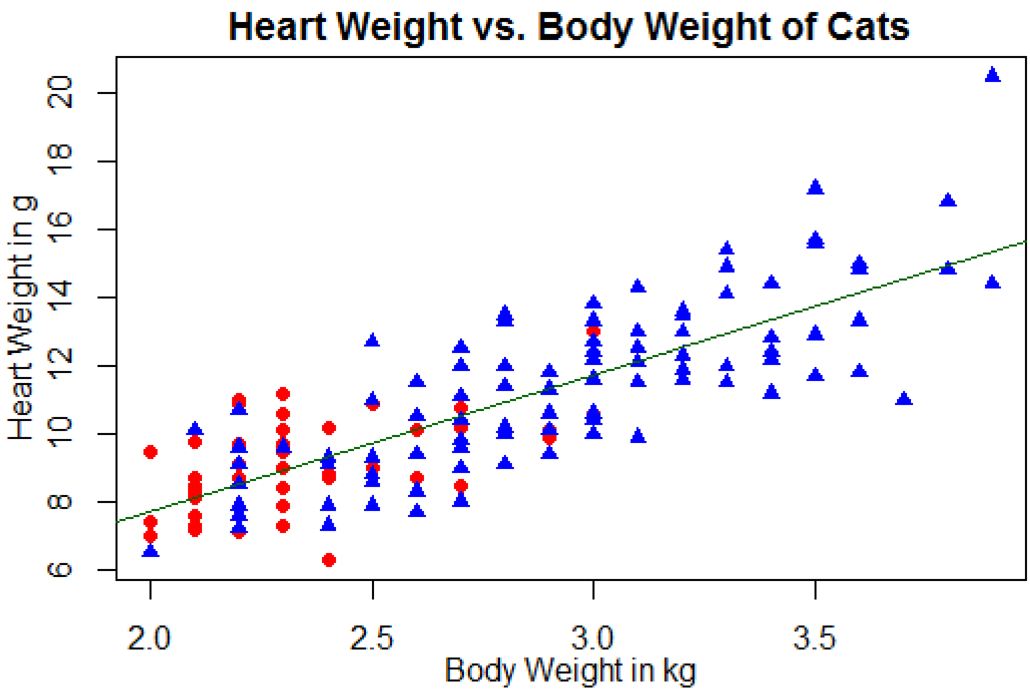


The plot() function gives a scatterplot whenever you feed it two numeric variables. The first variable listed will be plotted on the horizontal axis. A formula interface can also be used, in which case the response variable should come before the tilde and the variable to be plotted on the horizontal axis after. (Close the graphics window before doing this, because the output will look exactly the same.)

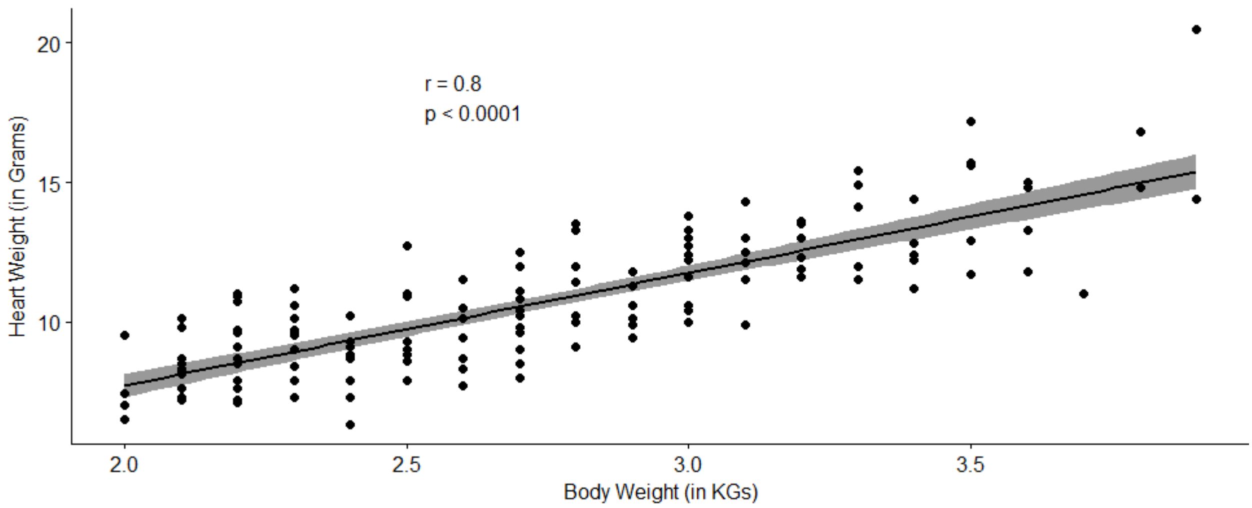
The scatterplot shows a fairly strong and reasonably linear relationship between the two variables. A Pearson product-moment correlation coefficient can be calculated using the cor() function (which will fail if there are missing values).

For a more revealing scatterplot, try this.

```
> with(cats, plot(Bwt, Hwt, type="n", las=.5, xlab="Body Weight in kg",
       ylab="Heart Weight in g",main="Heart Weight vs. Body Weight of Cats"))
> with(cats, points(Bwt[Sex=="F"], Hwt[Sex=="F"], pch=16, col="red"))
> with(cats, points(Bwt[Sex=="M"], Hwt[Sex=="M"], pch=17, col="blue"))
> abline(lm.out,col="dark green")
```



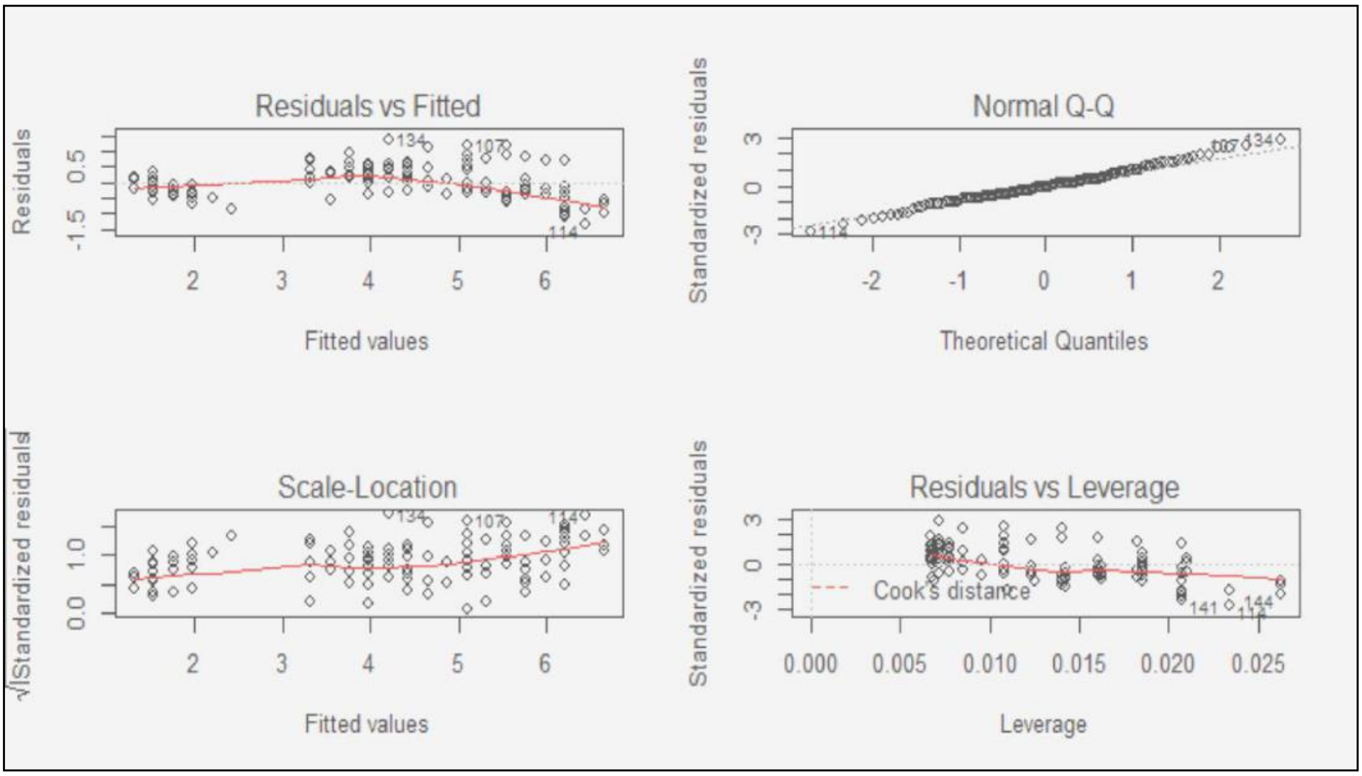

```
#Another Example with additional features
ggscatter(cats, x = "Bwt", y = "Hwt", add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson", xlab = "Body Weight (in KGs)",
          ylab = "Heart Weight (in Grams")
```



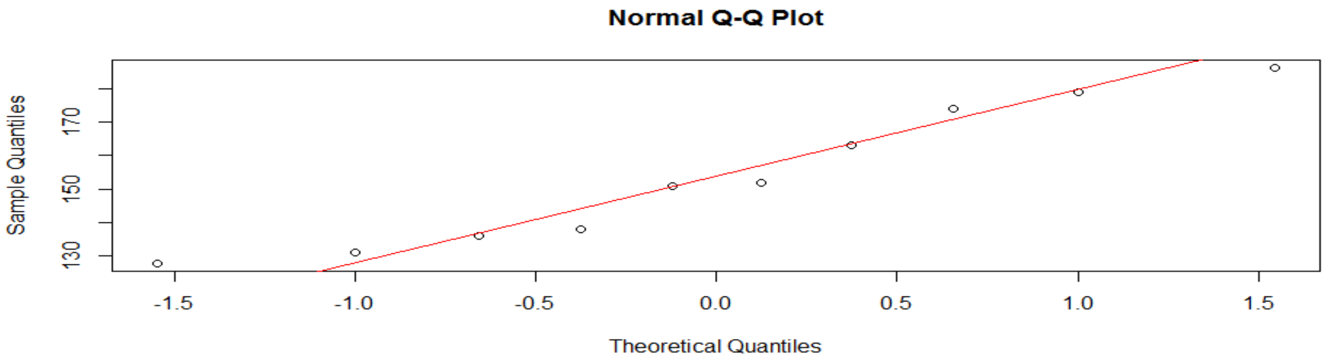
Visualization of fit data:

The fit information displays four charts: Residuals vs. Fitted, Normal Q-Q, Scale-Location, and Residuals vs. Leverage.

Below are the various graphs representing values of regression.




```
# The Normal Q-Q (quantile-quantile Normal Plot) for given vector of data items ca
n be drawn as/:
> x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131) # with single data set
> qqnorm(x)
> qqline(x,col='red')
```



4.2. Assumptions of OLS Regression

Ordinary least squares (OLS) Method:
Ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the differences between the observed responses and the predicted responses by the linear approximation of the data.
Assumptions of regression modeling:
For both simple linear and multiple regressions where the common assumptions are
a) The model is linear in the coefficients of the predictor with an additive random error term
b) The random error terms are

- Normally distributed with 0 mean and
- A variance that doesn't change as the values of the predictor covariates change.

4.3. Regression Modeling:

Regression modeling or analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

- **Understand influence of changes in dependent variable:**
  More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables, i.e the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution.
- **Estimation of continuous response variables:**
  Regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification. The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems.
- **Regression analysis uses:**
  It is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation.
- **Parametric and non-parametric regression:**
  Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data.
- **Nonparametric regression** refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.
- **Performance of regression analysis :**
  The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally.

4.3.1Regression residuals:

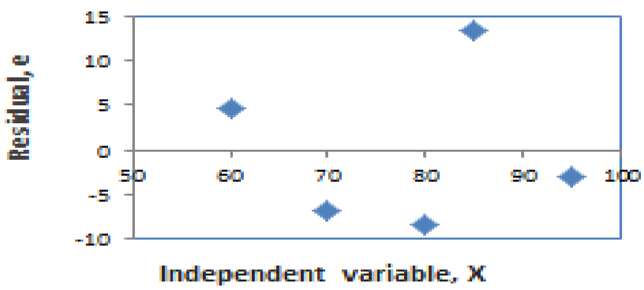
The residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest. Because a linear regression model is not always appropriate for the data, assess the appropriateness of the model by defining residuals and examining residual plots.

Residuals:

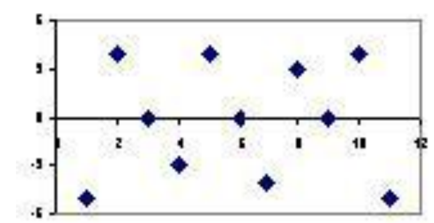
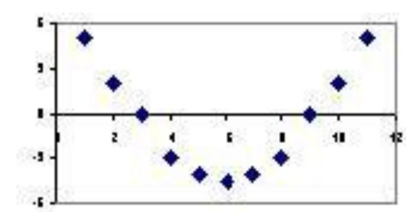
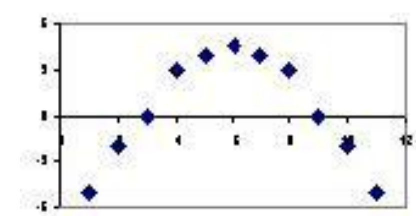
The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.
Residual = Observed value - Predicted value
$e = y - \hat{y}$

x	60	70	80	85	95
y	70	65	70	95	85
ŷ	65.411	71.849	78.288	81.507	87.945
e	4.589	-6.849	-8.288	13.493	-2.945

Both the sum and the mean of the residuals are equal to zero. That is, $\Sigma\, e = 0$ and e = 0. The above table shows inputs and outputs from a simple linear regression analysis.
Residual Plots:
A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.



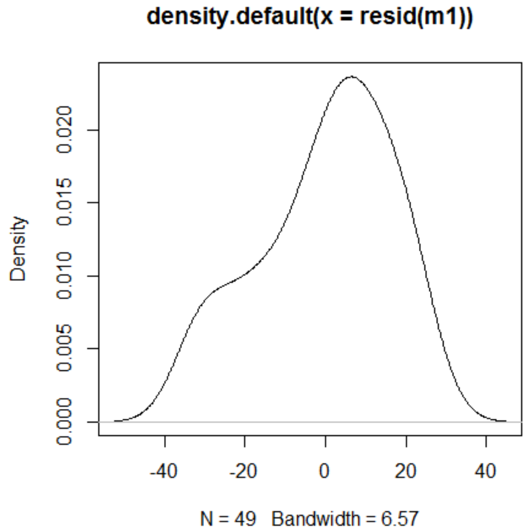
The chart on the right displays the residual (e) and independent variable (X) as a residual plot. The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data. Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

		
Random pattern	Non-random: U-shaped	Non-random: Inverted U

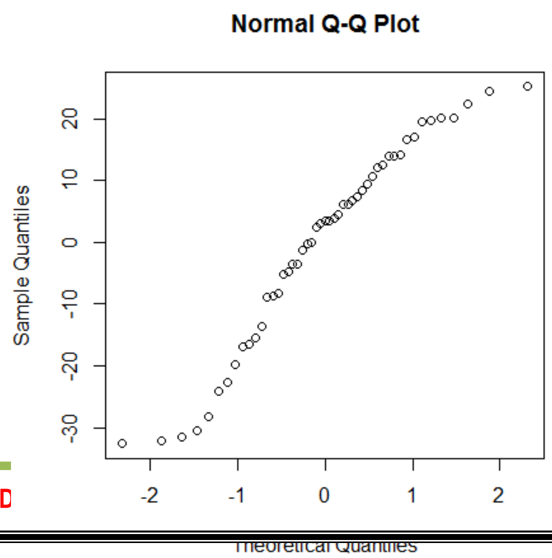
How to find Residuals and plot them?

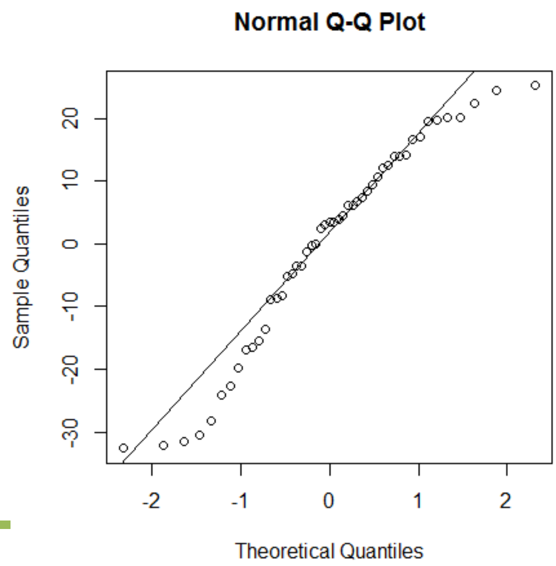
```
##Finding Residuals examples
x=c(21,34,6,47,10,49,23,32,12,16,29,49,28,8,57,9,31,10,21,26,31,52,21,8,18,5,18,26,
      27,26,32,2,59,58,19,14,16,9,23,28,34,70,69,54,39,9,21,54,26)
y = c(47,76,33,78,62,78,33,64,83,67,61,85,46,53,55,71,59,41,82,56,39,89,31,43,
      29,55, 81,82,82,85,59,74,80,88,29,58,71,60,86,91,72,89,80,84,54,71,75,84,79)

m1 <- lm(y~x)  #Create a linear model
resid(m1) #List of residuals
> resid(m1) #List of residuals
```

```
              1             2             3             4             5             6 
7             8 
-15.56228406   7.44941046 -22.65270081   3.46110498   4.50474366   2.53982722 -30.
48356182  -3.62931177 
              9            10            11            12            13            14 
15            16 
 24.58346589   6.74091036  -5.24739512   9.53982722 -19.78675624  -3.57397858 -24.
14528385  13.96538254 
             17            18            19            20            21            22 
23            24 
 -8.16867289 -16.49525634  19.43771594  -8.86547847 -28.16867289  12.15791057 -31.
56228406 -13.57397858 
             25            26            27            28            29            30 
31            32 
-32.18036741  -0.19206193  19.81963259  17.13452153  16.67388265  20.13452153  -8.
62931177  20.18985472 
             33            34            35            36            37            38 
39            40 
 -0.06656161   8.39407727 -32.64100629  -1.33781187  10.74091036   2.96538254  22.
51643818  25.21324376 
             41            42            43            44            45            46 
47            48 
  3.44941046   3.86641067  -4.67295044   6.23663280 -16.85378395  13.96538254  12.
43771594   6.23663280 
             49 
 14.13452153 
> plot(density(resid(m1))) #A density plot
```




```
> qqnorm(resid(m1)) # A quantile normal plot -
good for checking normality
> qqline(resid(m1))
```





4.4.Correlation:

The Correlation is a measure of association between two variables. Correlations are Positive and negative which are ranging between +1 and -1.

Positive correlation (0 to +1) example: Earning and expenditure

Negative correlation (-1 to 0) example : Speed and time

In R , correlation between x and y is by using cor(x,y) function.

Example:

> cor(iris$Petal.Length,iris$Petal.Width)

[1] 0.9628654

Means Petal Length and Petal Width are very strongly correlated.

Correlation Coefficient: correlation coefficient is indicated with r. The r value

- ✓ +1 : Perfectly positive
- ✓ -1 : Perfectly negative
- ✓ 0 – 0.2 : No or very weak association
- ✓ 0.2 – 0.4 : Weak association
- ✓ 0.4 – 0.6 : Moderate association
- ✓ 0.6 – 0.8 : Strong association
- ✓ 0.8 – 1.0 : Very strong to perfect association

```
> cor(Petal.Length, Petal.Width)
        Error in is.data.frame(y) : object 'Petal.Width' not found
> attach(iris)
> cor(Petal.Length, Petal.Width)
        [1] 0.9628654
> cor(Sepal.Length,Sepal.Width)
        [1] -0.1175698
> cor(Petal.Length,Sepal.Length)
        [1] 0.8717538
> cor(Petal.Length,Sepal.Width)
        [1] -0.4284401
> cor(Petal.Width,Sepal.Length)
        [1] 0.8179411

> iris1<-iris[1:4]
> iris1
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
6          5.4         3.9          1.7         0.4
7          4.6         3.4          1.4         0.3
.
.
.
```

```
[CONTD …]    # we have a total of 150 records in iris1, observe in iris 1 only fir
st four columns are considered, and species is left out to consider numeric data
> cor(iris1)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000


> cor.test(Petal.Length,Petal.Width)
  Pearson's product-moment correlation
data:  Petal.Length and Petal.Width
t = 43.387, df = 148, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9490525 0.9729853
sample estimates:
      cor
0.9628654    # which is the same value with cor() computation

> with(cats, cor.test(Bwt, Hwt))       # cor.test(~Bwt + Hwt, data=cats) also works
    Pearson's product-moment correlation   #^ Know More at the end of this unit..
  data:  Bwt and Hwt
  t = 16.1194, df = 142, p-value < 2.2e-16
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
   0.7375682 0.8552122
  sample estimates:
        cor
  0.8041274
```

Output Explanation:

The first line Pearson's Product-Moment Correlation (PPMC) (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Second Line the information about the data that we have considered for the correlation test.

On third line, you have the t-test for H0(Hypothesis Test HO & HA): As you can see, t is very large and p-value is very, very tiny, so you can reject the nil and be (almost) sure that the correlation on the population is not zero. P-value is just a yes/no kind of thing. Lower p-value does not mean stronger correlation. If p-value is <0.05 we can be reasonably certain that the two data columns are correlated, but a p-value of 0.01 does not mean the data is more strongly correlated than a set with p-value of 0.03. The t-statistic is used to calculate the p-value. The p value that's greater than 0.05 failed to reject the alternative hypothesis and the conclusion is that there is no significant correlation, namely accept the null hypothesis where correlation is equal to zero. Forth is the statement derived from alternative hypothesis test.

The fifth & sixth lines have very important information that the confidence interval at 95% for the correlation on the population. You can see that is between 0.73 and 0.86, so you can be (almost) sure that according to Cohen, you have a strong correlation on your population.

The last line is the ***correlation coefficient*** value 'r' obtained from the test.

```
> with(cats, cor.test(Bwt, Hwt, alternative="greater", conf.level=.8))

        Pearson's product-moment correlation

data:  Bwt and Hwt
t = 16.1194, df = 142, p-value < 2.2e-16
alternative hypothesis: true correlation is greater than 0
80 percent confidence interval:
 0.7776141 1.0000000
sample estimates:
```

```
        cor
  0.8041274
```

There is also a formula interface for cor.test(), but it's tricky. Both variables should be listed after the tilde. (Order of the variables doesn't matter in either the cor() or the cor.test() function.)

Covariance

The **covariance** of two variables *x* and *y* in a data set measures how the two are linearly related. A positive covariance would indicate a positive linear relationship between the variables, and a negative covariance would indicate the opposite.

The **sample covariance** is defined in terms of the sample means as:

$$\text{Covariance is } \sigma_{x,y} = \frac{\sum(x-\overline{x})(y-\overline{y})}{N}$$

$$\text{Correlation Coefficient is } \rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} \; where -1 \le \rho \le +1$$

In R, the covariance of x and y is by using cov(x, y) function.

Example: Covariance of petal length and petal width of iris dataset.

```
> cov(iris$Petal.Length,iris$Petal.Width)
[1] 1.295609
```

It means between Petal Length and Petal Width, the positive linear relationship existed.

Problem

Find the covariance of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the two variables.

Solution

We apply the cov function to compute the covariance of eruptions and waiting.

```
> duration = faithful$eruptions    # eruption durations
> waiting = faithful$waiting       # the waiting period
> cov(duration, waiting)           # apply the cov function
[1] 13.978
```

Answer

The covariance of eruption duration and waiting time is about 14. It indicates a positive linear relationship between the two variables.

The sample correlation coefficient is defined by the following formula, where sx and sy are the sample standard deviations, and cov(x,y) is the sample covariance.

$$r = \text{cov}(x, y)\frac{\text{cov}(x, y)}{\sqrt{S_x^2 S_y^2}}$$

4.5. ANOVA:

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as "variation" among and between groups).

In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups.

ANOVAs are useful for comparing (testing) three or more means (groups or variables) for statistical significance.

In R, to perform ANOVA test the built in function is anova()

Compute analysis of variance (or deviance) tables for one or more fitted model objects.

anova(object, ...)

Object : An object containing the results returned by a model fitting function (e.g., lm or glm).

Return value: an object of class anova includes:

Analysis-of-variance
Analysis-of-deviance tables.
Warning: The comparison between two or more models will only be valid if they are fitted to the same dataset. This may be a problem if there are missing values and R's default of na.action = na.omit is used.

```
> anova(fit)
Analysis of Variance Table

Response: iris$Petal.Length
                   Df Sum Sq Mean Sq F value    Pr(>F)
iris$Petal.Width    1 430.48  430.48  1882.5 < 2.2e-16 ***
Residuals         148  33.84    0.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> lm(Sepal.Length~Petal.Length+Petal.Width)

Call:
lm(formula = Sepal.Length ~ Petal.Length + Petal.Width)

Coefficients:
 (Intercept)  Petal.Length   Petal.Width
      4.1906        0.5418       -0.3196
```

4.6. Forecasting

Forecasting is the process of making predictions of the future based on past and present data and analysis of trends. A commonplace example might be estimation of some variable of interest at some specified future date.
Usage can differ between areas of application:
Example:
In hydrology, the terms "forecast" and "forecasting" are sometimes reserved for estimates of values at certain specific future times, The term "prediction" is used for more general estimates, such as the number of times floods will occur over a long period.

4.7. Heteroscedasticity:

A collection of random variables is heteroscedastic, if there are sub-populations that have different variability's from others. Here "variability" could be quantified by the variance or any other measure of statistical dispersion. Thus heteroscedasticity is the absence of homoscedasticity.

The existence of heteroscedasticity is a major concern in the application of regression analysis, including the analysis of variance, as it can invalidate statistical tests of significance that assume that the modeling errors are uncorrelated and uniform—hence that their variances do not vary with the effects being modeled. For instance, while the ordinary least squares estimator is still unbiased in the presence of heteroscedasticity, it is inefficient because the true variance and covariance are underestimated. Similarly, in testing for differences between sub-populations using a location test, some standard tests assume that variances within groups are equal.

4.8. Autocorrelation:

Autocorrelation, also known as serial correlation or cross-autocorrelation, is the cross-correlation of a signal with itself at different points in time. It is the similarity between observations as a function of the time lag between them. It is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals. In statistics, the autocorrelation of a random process describes the correlation between values of the process at different times, as a function of the two times or of the time lag.

Multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy.

4.9.Introduction to Multiple Regression:

The multiple regression is the relationship between several independent or predictor variables and a dependent or criterion variable.

For example: A real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating).

Lab Activity:

Multiple Regression model on Iris Data set

Step1: Subset the numeric data from iris dataset

Step2: Find the correlation among all variables

Step3: Find the formula based on highly correlated variables

Step4: call the glm()

```
>iris1<-iris[1:4]
> cor(iris1)
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Hence (sepal length, petal length),(petal width, sepal length) and (petal length ,petal width) are showing high correlation.

```
> glm(formula=iris$Petal.Length~iris$Petal.Width+iris$Sepal.Length)

Call:  glm(formula = iris$Petal.Length ~ iris$Petal.Width + iris$Sepal.Length)

Coefficients:
      (Intercept)   iris$Petal.Width  iris$Sepal.Length
         -1.5071            1.7481            0.5423

Degrees of Freedom: 149 Total (i.e. Null); 147 Residual
Null Deviance:          464.3
Residual Deviance: 23.9      AIC: 158.2
```

Hence the formula found from model is

iris$Petal.Length = 1.7481*iris$Petal.Width + 0.5423*iris$Sepal.Length-1.5071

study2:Multiple linear regression on MS application data:

```
# Multiple Linear Regression Example
fit <- lm(y ~ x1 + x2 + x3, data=mydata)
summary(fit) # show results
```

```
# Other useful functions
coefficients(fit) # model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit) # anova table
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnostics
```

```
>
> library(MASS)
> data(cats)
> attach(cats)
> lm.cats<- lm(Hwt~Bwt)
> summary(lm.cats)

Call:
lm(formula = Hwt ~ Bwt)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5694 -0.9634 -0.0921  1.0426  5.1238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3567     0.6923  -0.515    0.607
Bwt           4.0341     0.2503  16.119   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-squared:  0.6466,    Adjusted R-squared:  0.6441
F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

P Value is less than 0.05 which means we reject null hypothesis. Degree of Freedom is 142.
For other examples use the link: http://www.ats.ucla.edu/stat/r/dae/rreg.htm
Also refer to the book: - Practical Regression and Anova using R
Now to create a linear model of effect of Body Weight and Sex on Heart Weight we use multiple regression modeling.

Case study 3: Multiple linear regression on cats dataset:

```
> lm.cats<- lm(Hwt~Bwt+Sex)
> summary(lm.cats)

Call:
lm(formula = Hwt ~ Bwt + Sex)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5833 -0.9700 -0.0948  1.0432  5.1016 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.4149     0.7273  -0.571    0.569    
Bwt           4.0758     0.2948  13.826   <2e-16 ***
SexM         -0.0821     0.3040  -0.270    0.788    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.457 on 141 degrees of freedom
Multiple R-squared:  0.6468,    Adjusted R-squared:  0.6418 
F-statistic: 129.1 on 2 and 141 DF,  p-value: < 2.2e-16
```

So we can say that 65% variation in Heart Weight can be explained by the model.

The equation becomes $y=4.07x-0.08y-0.41$

Dummy Variables:

In regression analysis, a dummy variable (also known as an indicator variable, design variable, Boolean indicator, categorical variable, binary variable, or qualitative variable) is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. Dummy variables are used as devices to sort data into mutually exclusive categories (such as smoker/non-smoker, etc.).

In other words, Dummy variables are "proxy" variables or numeric stand-ins for qualitative facts in a regression model. In regression analysis, the dependent variables may be influenced not only by quantitative variables (income, output, prices, etc.), but also by qualitative variables (gender, religion, geographic region, etc.). A dummy independent variable (also called a dummy explanatory variable) which for some observation has a value of 0 will cause that variable's coefficient to have no role in influencing the dependent variable, while when the dummy takes on a value 1 its coefficient acts to alter the intercept.

Example:

Suppose Gender is one of the qualitative variables relevant to a regression. Then, female and male would be the categories included under the Gender variable. If female is arbitrarily assigned the value of 1, then male would get the value 0. Then the intercept (the value of the dependent variable if all other explanatory variables hypothetically took on the value zero) would be the constant term for males but would be the constant term plus the coefficient of the gender dummy in the case of females.

***** End of Unit-4 *****

Pearson's product-moment correlation

What is Pearson Correlation?

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation or PPMC. It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

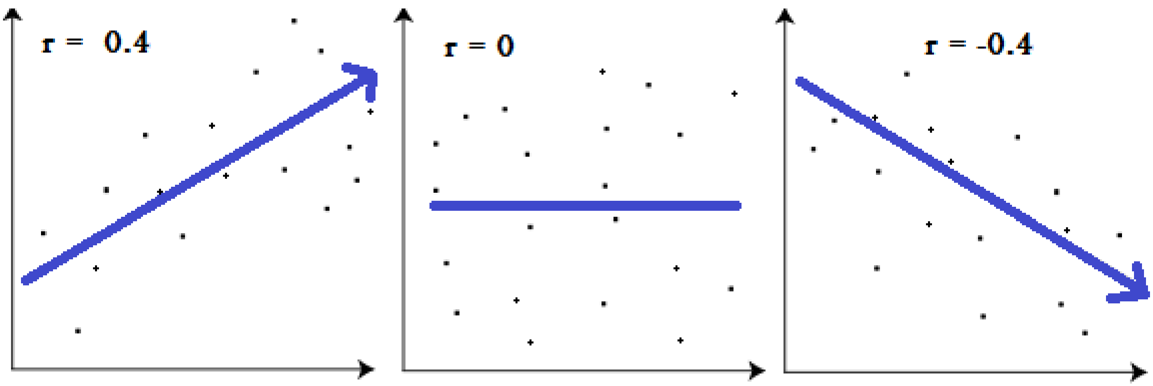
What are the Possible Values for the Pearson Correlation?

The results will be between -1 and 1. You will very rarely see 0, -1 or 1. You'll get a number somewhere in between those values. The closer the value of r gets to zero, the greater the variation the data points are around the line of best fit.

High correlation: .5 to 1.0 or -0.5 to 1.0.

Medium correlation: .3 to .5 or -0.3 to .5.

Low correlation: .1 to .3 or -0.1 to -0.3.



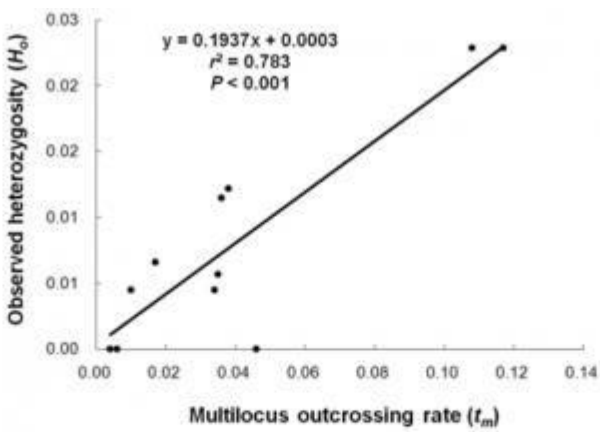
Positive Correlation   No correlation   Negative

Potential problems with Pearson correlation.

The PPMC is not able to tell the difference between dependent and independent variables. For example, if you are trying to find the correlation between a high calorie diet and diabetes, you might find a high correlation of .8. However, you could also work out the correlation coefficient formula with the variables switched around. In other words, you could say that diabetes causes a high calorie diet. That obviously makes no sense. Therefore, as a researcher you have to be aware of the data you are plugging in. In addition, the PPMC will not give you any information about the slope of the line; It only tells you whether there is a relationship.

Real Life Example

Pearson correlation is used in thousands of real life situations. For example, scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of the rice. Pearson's correlation between the two groups was analyzed. It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations. This figure is quite high, which suggested a fairly strong relationship.



If you're interested in seeing more examples of PPMC, you can find several studies on the National Institute of Health's Openi website, which shows result on studies as varied as breast cyst imaging to the role that carbohydrates play in weight loss.

Correlation Coefficient

The **correlation coefficient** of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related.

Formally, the **sample correlation coefficient** is defined by the following formula, where s_x and s_y are the sample standard deviations, and s_{xy} is the sample covariance.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Similarly, the **population correlation coefficient** is defined as follows, where σ_x and σ_y are the population standard deviations, and σ_{xy} is the population covariance.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

If the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

Problem

Find the correlation coefficient of eruption duration and waiting time in the data set faithful. Observe if there is any linear relationship between the variables.

Solution

We apply the cor function to compute the correlation coefficient of eruptions and waiting.

```
> duration = faithful$eruptions    # eruption durations
> waiting = faithful$waiting       # the waiting period
> cor(duration, waiting)           # apply the cor function
[1] 0.90081
```

Answer

The correlation coefficient of eruption duration and waiting time is 0.90081. Since it is rather close to 1, we can conclude that the variables are positively linearly related.

Introduction to Analytics (Associate Analytics – I)
UNIT V
Understand the Verticals - Engineering, Financial and others (NOS 9002)

Understand the Verticals _ Engineering, Financial and others (NOS 9002)
Understanding systems viz. Engineering Design, Manufacturing, Smart Utilities, Production lines, Automotive, Technology etc. Understanding Business problems related to various businesses.
Requirements Gathering
Gathering all the data related to Business objective

S. No	Content
5.1	Understanding systems viz. Engineering Design, Manufacturing, Smart Utilities, Production lines, Automotive, Technology etc
5.2	Understanding Business problems related to various businesses.
5.3	Requirements Gathering Gathering all the data related to Business objective

5.1.1 Engineering Design:
The engineering design process is a series of steps that engineers follow to come up with a solution to a problem. Many times the solution involves designing a product (like a machine or computer code) that meets certain criteria and/or accomplishes a certain task.
The engineering design process is a methodical series of steps that engineers use in creating functional products and processes. The process is highly iterative - parts of the process often need to be repeated many times before production phase can be entered - though the part(s) that get iterated and the number of such cycles in any given project can be highly variable.

Engineering Design Process describes the following stages:
1) Research
2) Conceptualization
3) Feasibility assessment
4) Establishing Design requirements
5) Preliminary design
6) Detailed design
7) Production planning and tool design, and
8) Production.

1. Research:
Research is a careful and detailed study into a specific problem, concern, or issue using the scientific method Research can be about anything, and we hear about all different types of research in the news. Cancer research has 'Breakthrough Cancer-Killing Treatment Has No Side Effects in Mice,' and 'Baby Born with HIV Cured.' Each of these began with an issue or a problem (such as cancer or HIV), and they had a question, like, 'Does medication X reduce cancerous tissue or HIV infections?'

But all I've said so far is what research has done (sort of like saying baking leads to apple pie; it doesn't really tell you anything other than the two are connected). To begin researching something, you have to have a problem, concern, or issue that has turned into a question. These can come from observing the world, prior research, professional literature, or from peers. Research really begins with the right question, because your question must be answerable. Questions like, 'How can I cure cancer?' aren't really answerable with a study. It's too vague and not testable.

3. Conceptualization:

Conceptualization is mental process of organizing one"s observations and experiences into meaningful and coherent wholes.

In research, conceptualization produces an agreed upon meaning for a concept for the purposes of research. Different meaning for a concept for the purposes of research. Different researchers may conceptualize a concept slightly differently. „ Conceptualization describes the indicators we'll use to measure Conceptualization describes the indicators we'll use to measure the concept and the different aspects of the concept.

3. Feasibility assessment:

Feasibility studies are almost always conducted where large sums are at stake. Also called feasibility analysis. In order to ensure the manufacturing facility to make a new item the engineers launched a feasibility study to determine the actual steps required to build the product.

4. Design Requirements :

The product/component to be analysed is characterised in terms of: functional requirements, objective of the materials selection process, constraints imposed by the requirements of the application, plus the free variable, which is usually one of the geometric dimensions of the product/component, such as thickness, which enables the constraints to be satisfied and the objective function to be maximised or minimised, depending on the application. Hence, the design requirement of the part/component is defined in terms of function, objective, constraints

5. Preliminary Design:

The preliminary design, or high-level design includes , often bridges a gap between design conception and detailed design, particularly in cases where the level of conceptualization achieved during ideation is not sufficient for full evaluation. So in this task, the overall system configuration is defined, and schematics, diagrams, and layouts of the project may provide early project configuration. This notably varies a lot by field, industry, and product. During detailed design and optimization, the parameters of the part being created will change, but the preliminary design focuses on creating the general framework to build the project on.

6. Detailed Design:

Detailed Design phase, which may consist of procurement of materials as well. This phase further elaborates each aspect of the project/product by complete description through solid modelling, drawings as well as specifications.

7. Production planning and tool design:

The production planning and tool design consists of planning how to mass-produce the product and which tools should be used in the manufacturing process. Tasks to complete in this step include selecting materials, selection of the production processes, determination of the sequence of operations, and selection of tools such as jigs, fixtures, metal cutting and metal or plastics forming tools. This task also involves additional prototype testing iterations to ensure the mass-produced version meets qualification testing standards.

8. Production:

Production is a process of workers combining various material inputs and immaterial inputs (plans, know-how) in order to make something for consumption (the output). It is the act of creating output, a good or service which has value and contributes to the utility of individuals.

5.1.2. Manufacturing:

Manufacturing is the production of goods for use or sale using labour and machines, tools, chemical and biological processing, or formulation. The term may refer to a range of human activity, from handicraft to high tech, but is most commonly applied to industrial production, in which raw materials are transformed into finished goods on a large scale. Such finished goods may be used for manufacturing other, more complex products, such as aircraft, household appliances or automobiles, or sold to wholesalers, who in turn sell them to retailers, who then sell them to end users – the "consumers".

Manufacturing takes turns under all types of economic systems. In a free market economy, manufacturing is usually directed toward the mass production of products for sale to consumers at a profit. In a collectivist economy, manufacturing is more frequently directed by the state to supply a centrally planned economy. In mixed market economies, manufacturing occurs under some degree of government regulation.

Modern manufacturing includes all intermediate processes required for the production and integration of a product's components. Some industries, such as semiconductor and steel manufacturers use the term fabrication instead.

The manufacturing sector is closely connected with engineering and industrial design. Examples of major manufacturers in North America include General Motors Corporation, General Electric, Procter & Gamble, General Dynamics, Boeing, Pfizer, and Precision Cast parts. Examples in Europe include Volkswagen Group, Siemens, and Michelin. Examples in Asia include Sony, Huawei, Lenovo, Toyota, Samsung, and Bridgestone.

5.1.3 Smart Utilities:

S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology; often written as SMART) is a monitoring system included in computer hard disk drives (HDDs) and solid-state drives (SSDs) that detects and reports on various indicators of drive reliability, with the intent of enabling the expectation of hardware failures.

When S.M.A.R.T. data indicates a possible forthcoming drive failure, software running on the host system may notify the user so stored data can be copied to another storage device, preventing data loss, and the failing drive can be replaced.

Smart Utility Systems is the leading provider of Software-as-a-Service (SaaS) solutions for Customer Engagement, Mobile Workforce, and Big Data Analytics to the Energy and Utility sector. We help utilities improve their operational efficiency and maximize revenue realization, through mobile and cloud technologies.

5.1.4 Production lines:

Production lines is an arrangement in a factory in which a thing being manufactured is passed through a set linear sequence of mechanical or manual operations. A production line is a set of sequential operations established in a factory whereby materials are put through a refining process to produce an end-product that is suitable for onward consumption; or components are assembled to make a finished article.

5.1.5 Automotive:

The automotive industry is a wide range of companies and organizations involved in the design, development, manufacturing, marketing, and selling of motor vehicles,[1] some of them are called automakers. It is one of the world's most important economic sectors by revenue. The automotive industry does not include industries will be dedicated to the maintenance of automobiles following delivery to the end-user (maybe), such as automobile repair shops and motor fuel filling stations.

5.1.6 Technology:

The application of scientific knowledge for practical purposes, especially in industry. Technology can be the knowledge of techniques, processes, and the like, or it can be embedded in machines which can be operated without detailed knowledge of their workings. The human species' use of technology began with the conversion of natural resources into simple tools. The prehistoric discovery of how to control fire and the later Neolithic Revolution increased the available sources of food and the invention of the wheel helped humans to travel in and control their environment. Developments in historic times, including the printing press, the telephone, and the Internet, have lessened physical barriers to communication and allowed humans to interact freely on a global scale. The steady progress of military technology has brought weapons of ever-increasing destructive power, from clubs to nuclear weapons.

Technology has many effects. It has helped develop more advanced economies (including today's global economy) and has allowed the rise of a leisure class. Many technological processes produce unwanted by-products known as pollution and deplete natural resources to the detriment of Earth's environment. Various implementations of technology influence the values of a society and new technology often raises new ethical questions. Examples include the rise of the notion of efficiency in terms of human productivity, and the challenges of bioethics.

Philosophical debates have arisen over the use of technology, with disagreements over whether technology improves the human condition or worsens it. Neo-Luddism, anarcho-primitivism, and similar reactionary movements criticise the pervasiveness of technology in the modern world, arguing that it harms the environment and alienates people; proponents of ideologies such as transhumanism and techno-progressivism view continued technological progress as beneficial to society and the human condition.

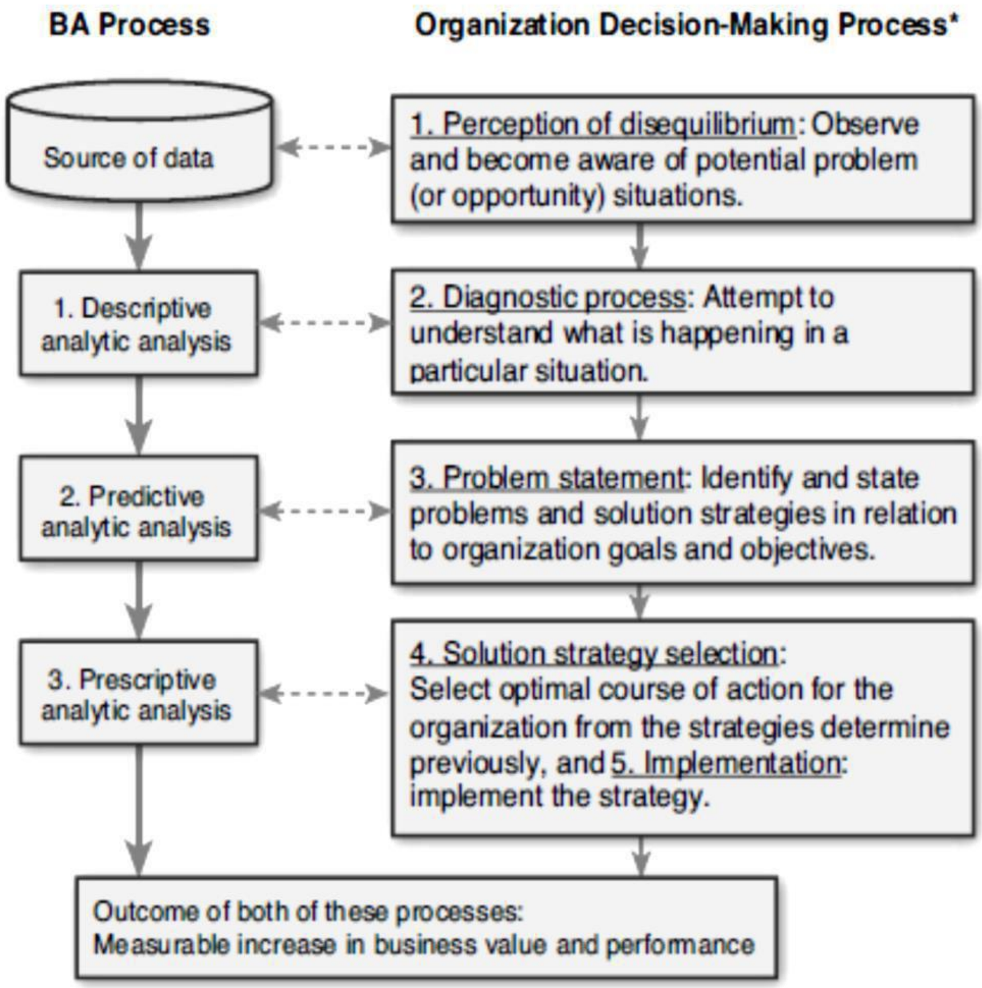
Until recently, it was believed that the development of technology was restricted only to human beings, but 21st century scientific studies indicate that other primates and certain dolphin communities have developed simple tools and passed their knowledge to other generations.

5.2 Understanding Business problems related to various businesses :

The BA process can solve problems and identify opportunities to improve business performance. In the process, organizations may also determine strategies to guide operations and help achieve competitive advantages. Typically, solving problems and identifying strategic opportunities to follow are organization decision-making tasks. The latter, identifying opportunities can be viewed as a problem of strategy choice requiring a solution.

A business analysis is the practice of identifying and clarifying problems or issues within a company and providing efficient solutions that satisfy the requirements of all stakeholders. ... Both information gathering techniques and communications with the various stakeholders are critical parts of the overall process.

5.3.1 Comparison of business analytics and organization decision-making processes:



5.3.2.Requirements Gathering: Gather all the Data related to Business objective

There are many different approaches that can be used to gather information about a business. They include the following:
1.Review business plans, existing models and other documentation
2.Interview subject area experts
3.Conduct fact-finding meetings
4.Analyze application systems, forms, artifacts, reports, etc.

The business analyst should use one-on-one interviews early in the business analysis project to gage the strengths and weaknesses of potential project participants and to obtain basic information about the business. Large meetings are not a good use of time for data gathering.
Facilitated work sessions are a good mechanism for validating and refining "draft" requirements. They are also useful to prioritize final business requirements. Group dynamics can often generate even better ideas.
Primary or local data is collected by the business owner and can be collected by survey, focus group or observation. Third party static data is purchased in bulk without a specific intent in mind. While easy to get (if you have the cash) this data is not specific to your business and can be tough to sort through as you often get quite a bit more data than you need to meet your objective. Dynamic data is collected through a third party process in near real-time from an event for a specific purpose (read into that VERY expensive).
Three key questions you need to ask before making a decision about the best method for your firm.
··What is the timeline required to accomplish your business objective?

-alone event or for part of a broader data collection effort?

How to interpret Data to make it useful for Business:
Business intelligence (BI) is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. BI technologies are capable of handling large amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities. The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.
BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.
BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data.

Business intelligence is made up of an increasing number of components including:
- Multidimensional aggregation and allocation
- Denormalization, tagging and standardization
- Realtime reporting with analytical alert
- A method of interfacing with unstructured data sources
- Group consolidation, budgeting and rolling forecasts
- Statistical inference and probabilistic simulation
- Key performance indicators optimization
- Version control and process management

- Open item management

Business intelligence can be applied to the following business purposes, in order to drive business value :

- Measurement – program that creates a hierarchy of performance metrics and benchmarking that informs business leaders about progress towards business goals (business process management).
- Analytics – program that builds quantitative processes for a business to arrive at optimal decisions and to perform business knowledge discovery. Frequently involves: data mining, process mining, statistical analysis, predictive analytics, predictive modeling, business process modeling, data lineage, complex event processing and prescriptive analytics.
- Reporting/enterprise reporting – program that builds infrastructure for strategic reporting to serve the strategic management of a business, not operational reporting. Frequently involves data visualization, executive information system and OLAP.
- Collaboration/collaboration platform – program that gets different areas (both inside and outside the business) to work together through data sharing and electronic data interchange.
- Knowledge management – program to make the company data-driven through strategies and practices to identify, create, represent, distribute, and enable adoption of insights and experiences that are true business knowledge. Knowledge management leads to learning management and regulatory compliance

In addition to the above, business intelligence can provide a pro-active approach, such as alert functionality that immediately notifies the end-user if certain conditions are met. For example, if some business metric exceeds a pre-defined threshold, the metric will be highlighted in standard reports, and the business analyst may be alerted via e-mail or another monitoring service. This end-to-end process requires data governance, which should be handled by the expert.

Data can be always gathered using surveys.
Your surveys should follow a few basic but important rules:
1. Keep it VERY simple. I recommend one page with 3-4 questions maximum. Customers are visiting to purchase or to have an experience, not to fill out surveys.
2. Choose only one objective for the survey. Don't try to answer too many questions, ultimately you won't get much useful data that way because your customer will get confused and frustrated.
3. Don't give the respondent any wiggle room. Open ended questions are tough to manage. Specific choices that are broad enough to capture real responses gives you data that is much easier to use.
4. Always gather demographics. Why not? But rather than name and e-mail (leading to concerns with confidentiality and often less than truthful answers) gather gender, age and income; you might be surprised at who is actually buying what.

***** End of Unit-5 *****