

# Introduction to R & it's some Application in Statistics



Dr. R S Rajput

Assistant Professor Computer Science

Department of Mathematics, Statistics, Computer Science

G.B. Pant University of Agriculture & Technology

Pantnagar, Uttarakhand, INDIA



# Introduction to R

- R is a powerful software environment for data management and data analysis.
- R is an Open source Software.
- Mostly R tools are working satisfactory on the low configured computer i.e. no advanced computing resources are required for all tools.
- Huge availability of references and Help.
- Well recognized in scientific computing.

# Resource of R

- Installation set up software, Installation guideline, software manual and help available following website.

<https://www.r-project.org/>

Download installation setup file

(e.g. \*.exe for windows based OS)

Simple to install: step-by-step wizard

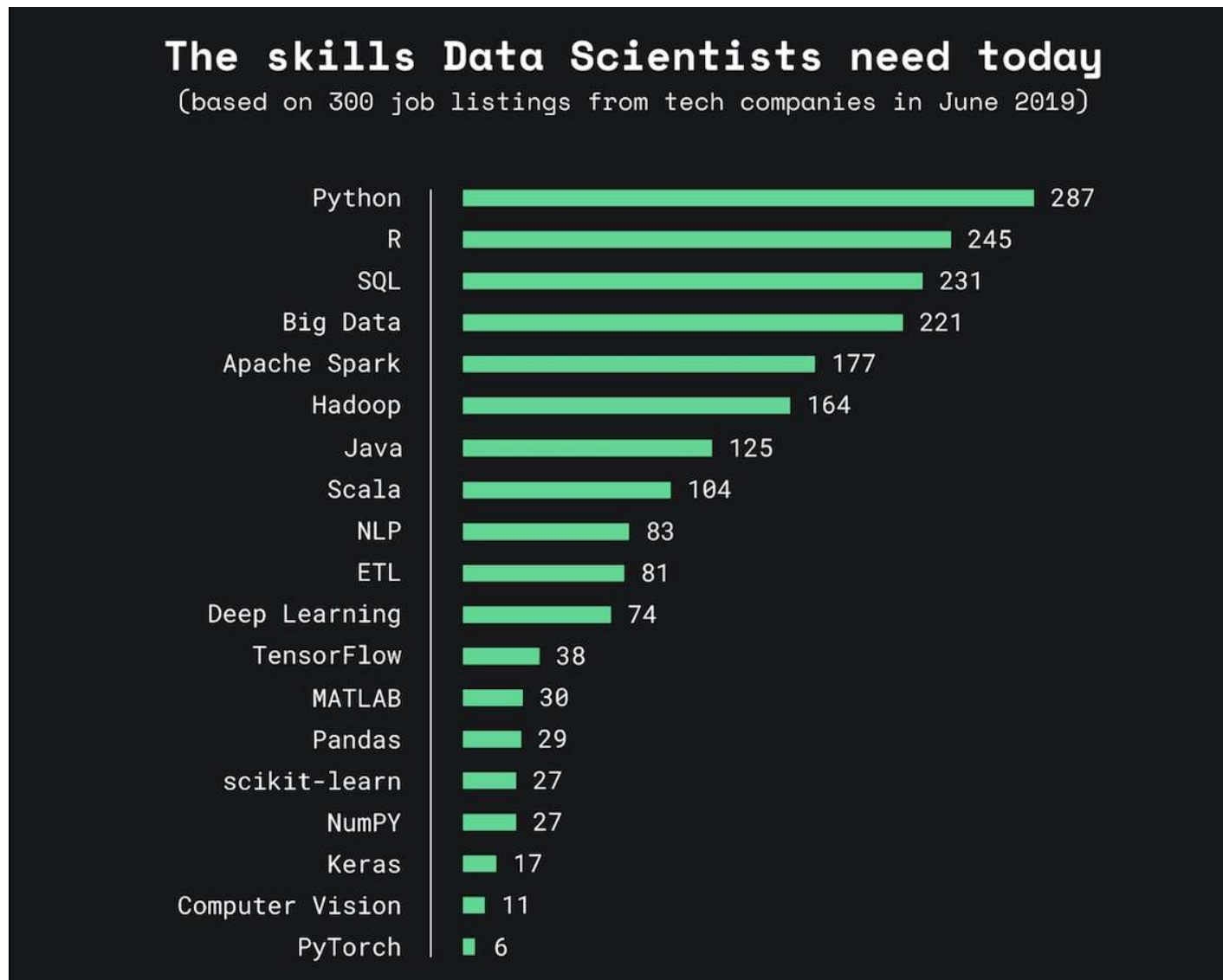
Less installation time: Approx. 10-15 minutes required

Not required for a expertise person for installation

# Why R



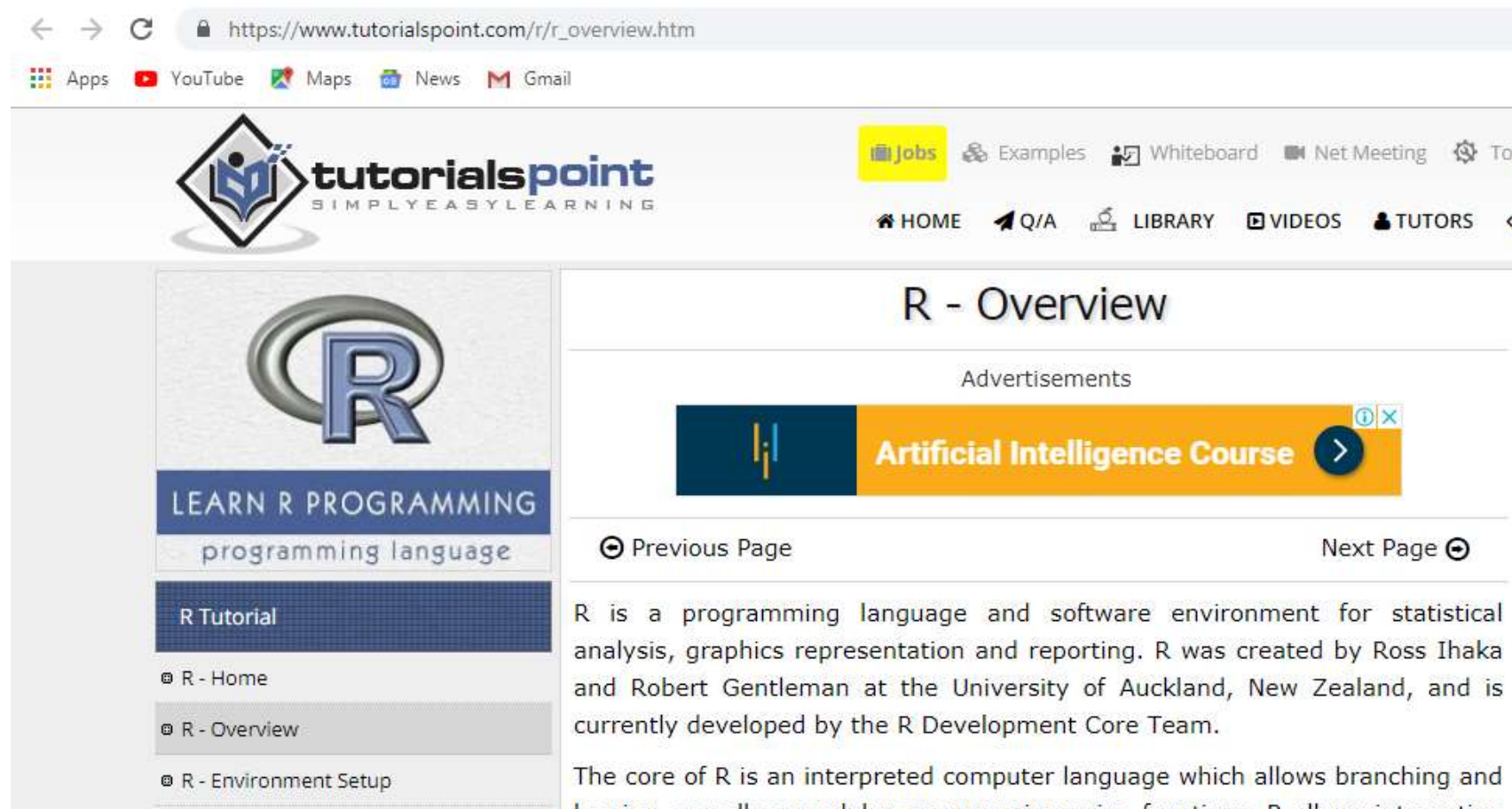
# Growth of R users



Rashi Desai (2020), The Most Popular Tools and Software for Data Science, <https://towardsdatascience.com/>

# Web Resource for Learning R

- [https://www.tutorialspoint.com/r/r\\_overview.htm](https://www.tutorialspoint.com/r/r_overview.htm)



# Web Resource for Learning R cont..

- <http://www.r-tutor.com/r-introduction>

www.r-tutor.com/r-introduction

News Gmail

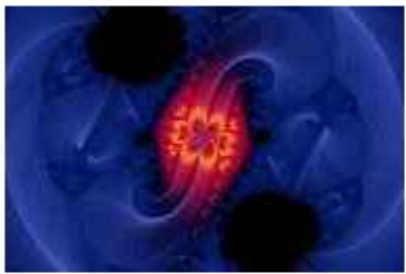
About | Contact | Resources | Terms of Use

## R Tutorial

An R Introduction to Statistics

HOME DOWNLOAD SALES EBOOK SITE MAP

### R Introduction




We offer here a couple of introductory tutorials on basic R concepts. It serves as background material for our main tutorial series *Elementary Statistics with R*.

The only hardware requirement for most of the R tutorials is a PC with the latest free open source R software installed. R has extensive documentation and active online community support. It is the perfect environment to get started in statistical computing.

Search this site:  Search

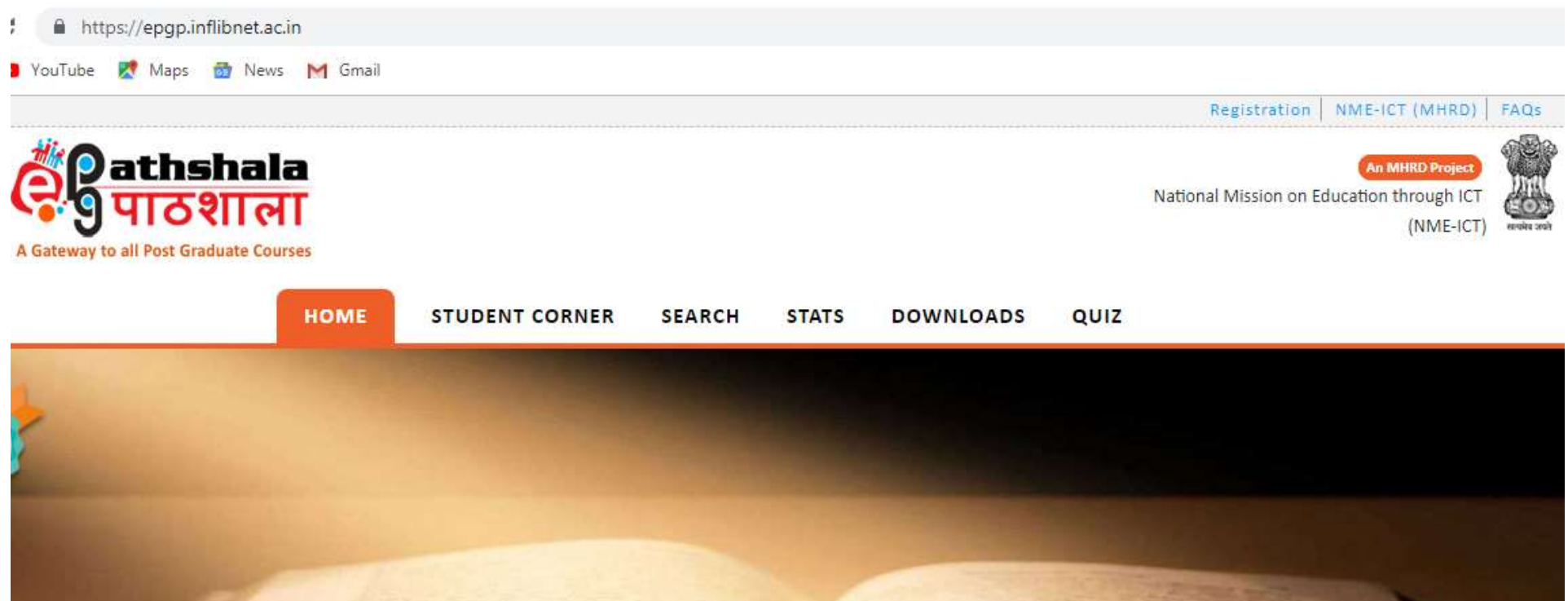
### R Tutorial eBook





# Web Resource for Learning R Cont..

- UGC (MHRD) s' e-PG Pathshala
- <http://www.ugc.ac.in> (go to e-PG Pathshala Link)
- <http://epgp.inflibnet.ac.in/>





# Starting R



- Start R by double clicking on R icon.
- A windows displayed, called R console with prompt ">".
- The > is called the R prompt.
- It is used to indicate where you are to type.

```
R : Copyright 2001, The R Development Core Team
Version 1.4.0 (2001-12-19)

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.

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

[Previously saved workspace restored]

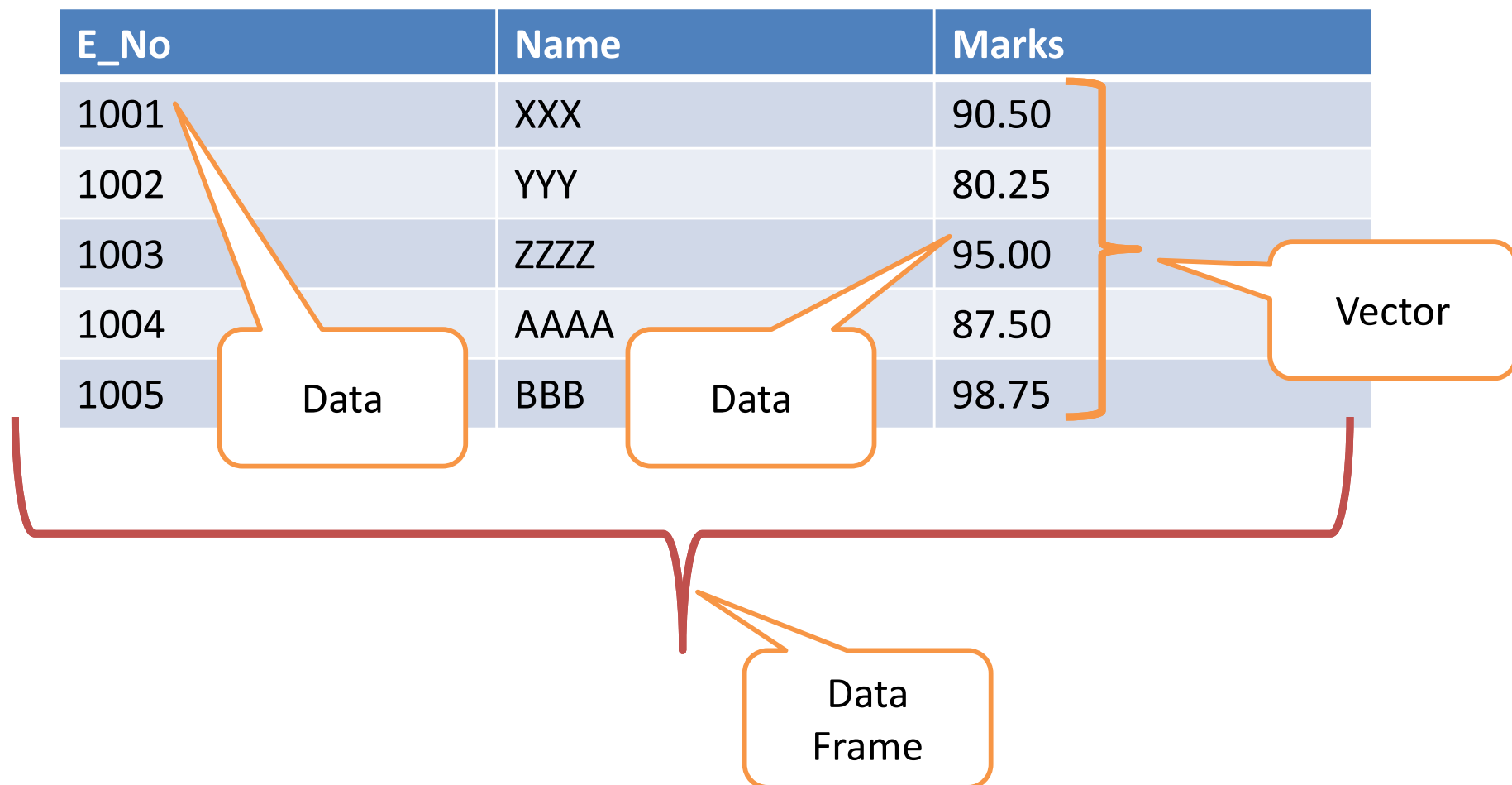
>
```

# Key Elements of R

- Data
- Vector
- Data Frame
- Operators
  - Mathematical, Relational, Logical, Assignment
- Function
  - Built in ( Library) function, User defined function
- Statements
  - Assignment, Conditional, Looping, Data Import & Export

# Data Management in R

- Data->Vector-> Data Frame



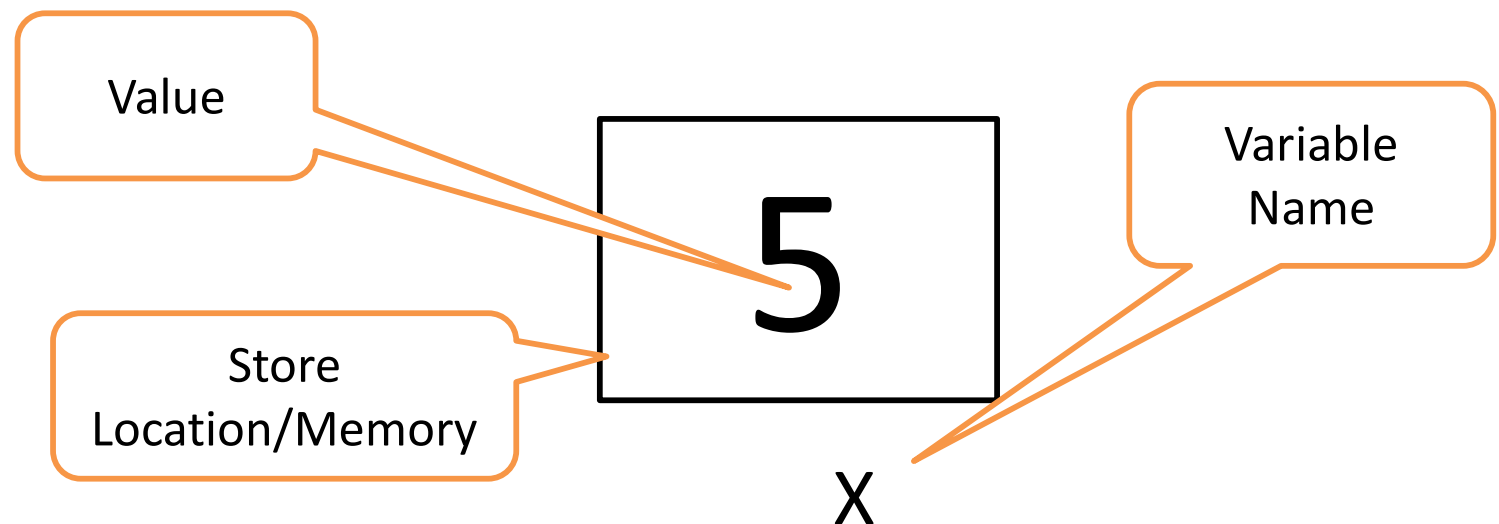
# Variable/ Variable Naming in R

- A variable name can contain letters (a-z, A-Z), digits (0-9), dot (.) and under score (\_)
- A variable name must be start with a letter or a dot.
- Example: if we want to store a value '5' into a variable X than R Code is.

>X<-5

OR

>X=5



# Data types in R

- Numeric
- Integer
- Complex
- Logical
- Character

# Basic Data Type

- Numeric Data
- Default Data type
- Decimal values are known as **numeric** in R
- Integer
- Numbers without decimal

```
>x=10.5
```

```
>x
```

```
[1] 10.5
```

```
>class(x)
```

```
[1]"numeric"
```

```
>y=as.integer(3)
```

```
>y
```

```
[1] 3
```

```
>class(y)
```

```
[1]"integer"
```

# Basic Data Type Cont..

- Complex Data
- A Complex value in R is defined via the pure imaginary value  $i$ .
- Logical Data
- A logical value is created via comparison between variables

$$i = \sqrt{-1}$$

```
>z=1+2i
```

```
>z
```

```
[1]1+2i
```

```
>class (z)
```

```
[1]"complex"
```

```
>x=1; y=2
```

```
>z=x>y
```

```
>z
```

```
[1]FALSE
```

```
>class (z)
```

```
[1]"logical"
```



# Basic Data Type Cont..

- Character Data
- A character object is used to represent string values in R.
- `class( )` function used to print class of data type.
- `#` is used to comment.

```
>x="Technology"
```

```
>class(x)
```

```
[1]"character"
```

# Vector

- A Vector is a container for data elements of the same basic data type. OR collection of similar data type.
- The container is known as **Vector**, and elements are known as components in vector.
- Generally a function 'c' is used to create vector.
- **'c' function**  
`>x=c (5, 6, 7, 3, 5)`
- Here 'x' is a vector and 5, 6, 7, 3, 5 are elements, x is numeric vector
- **Some vector functions**  
`>x, length(x), >x[5],>x[-2],>x[2:4]`

# Data Frame

- A data frame is used for storing data in as a tables.
- It is a list of related vectors having equal length.
- For example, the following variable df is a data frame containing three vectors n, s, b.

```
>n = c(2,3,5)
```

```
>s = c("aa","bb","cc")
```

```
>b = c(TRUE,FALSE,TRUE)
```

```
>df = data.frame(n,s,b)
```

```
>df
```

- Data frame define: Using function **data.frame**
- Inbuilt dataframe: mtcars, Orange, trees
- \$ Operator: retrieve vector from data frame e.g., df\$n

# Mathematical Operators

- (+) Addition
- (-) Subtraction
- (\*) Multiplication
- (/) Division
- (\*\*) or ^) Power
- (%%) Reminder

# Relational & Logical Operators

- $(==)$  Equal
- $(!=)$  Not equal
- $(<)$  Less than
- $(<=)$  Less than equal to
- $(>)$  Greater than
- $(>=)$  Greater than equal to
- $(!)$  Logical NOT
- $(&)$  Logical OR
- $(|)$  Logical AND

# Assignment Operator

- (= or <-)

Example(s)

```
> X=10
```

```
>V1=c (1,2,3,4,5,6,7,8,9)
```

# Date Import and Export from MS Excel

- Set working directory  
`>setwd (" C:\data")`
- Data import from Excel file ( as \*.csv file format)  
`>data =read.csv("input.csv")`
- Data Export to Excel file  
`>data = write.csv("output.csv")`



# Data

**Data means Observations or evidences.**

Singh, Y.K. (2008)

**Data known facts that can be recoded and that have implicit meaning.**

Elmasri R, Navathe S.B. (2008)

**Data is distinct pieces of information.**

(<https://www.webopedia.com>)

**Data is a set of values of subjects with respect to qualitative or quantitative variables.**

(<https://en.wikipedia.org/wiki/Data>)

**A collection of facts, such as numbers, words, measurements, observations or even just descriptions of things.**

(<https://www.mathsisfun.com/definitions/data.html>)

Example: weights, prices, costs, numbers of items sold, employee names, product names, addresses, tax codes, registration marks etc.

# Nature of Data

Data can be classified into two broad categories

- **Qualitative data**

For which numerical value can not be assigned, e.g. Division, Grades, Motivation, Confidence etc.

- Ordinal
- Nominal

- **Quantitative data**

For which numerical value can be assigned, e.g. Height, Weight, Speed etc.

Explore – ***mtcars*** dataset of R

# Variate

- The quantified variable is known as variate.
- Quantification is the process of assigning numeral values to variable.
- The Statistical analysis involves variate analysis: Uni-variate, Bi-variate, Multi-variate analysis.

# Variate Examples

## Univariate

Travel Time (minutes):

**15, 29, 8, 42, 35, 21, 18, 42, 26**

## Bivariate

### *Ice Cream Sales vs Temperature*

Temperature °C	Ice Cream Sales
14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408

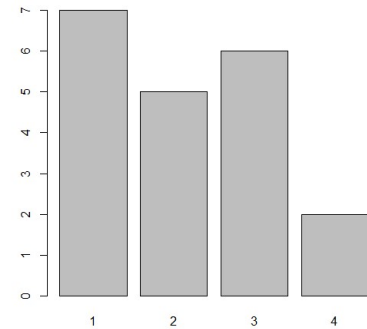
# Data Visualization

- **Uni-variate and Qualitative** (*ordinal*)
  - Pie diagram, Bar plot
- **Uni-variate and Quantitative**
  - Box plot, Histogram
- **Bi-variate (Qualitative** (*ordinal*), **Qualitative** (*ordinal*))
  - Mosaic Plot
- **Bi-variate (Qualitative** (*ordinal*), **Quantitative** )
  - Plot of Boxes, Bar plot
- **Bi-variate (Quantitative, Quantitative )**
  - Scatter plot, Line diagram

# Visualization Univariate Qualitative data

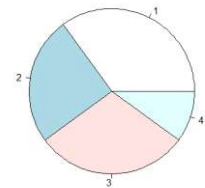
**Bar Chart:** Bar chart is a display frequency of a qualitative variable.

```
> grade=c(1,3,3,1,1,2,2,2,1,2,3,3,3,4,4,1,2,3,1,1)
> grade.freq=table(grade) # Frequency Vector
> grade.freq
> barplot(grade.freq)
```



**Pie chart:** Pie chart is a display percent frequency of a qualitative variable.

```
> pie(grade.freq)
```

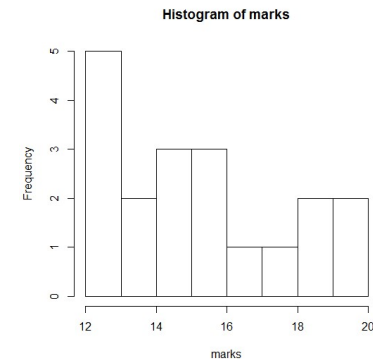
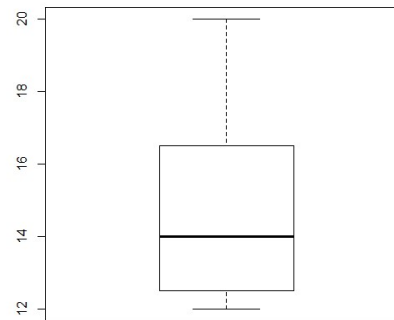


# Visualization Univariate Quantitative data

**Box Plot:** Box plot is used to summarize the distribution of a numeric variable.

```
>marks=c(12,12,14,15,13,14,16,17,12,15,18,18,19,20,12,13,12,14,15)
```

```
>boxplot(marks)
```



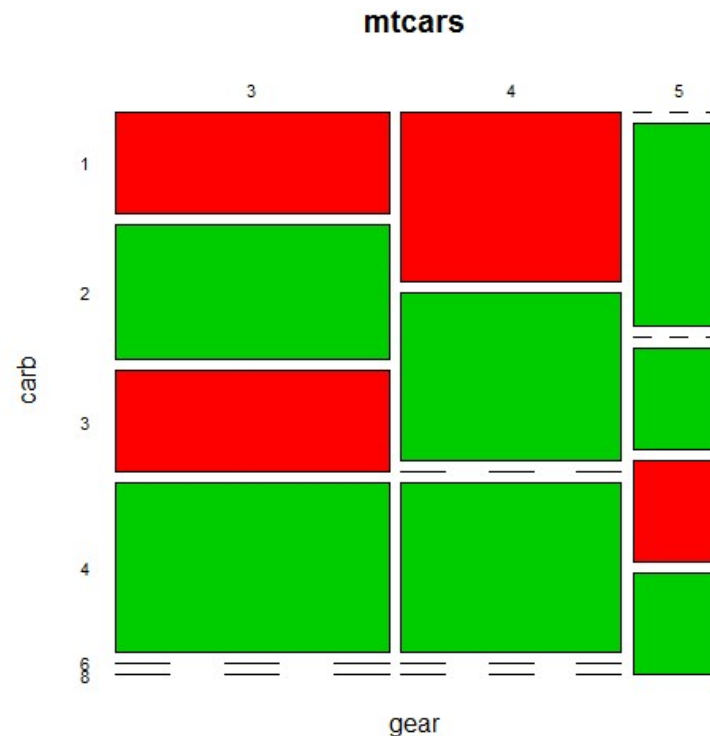
**Histograms:** Histogram is used to summarize the distribution of numeric variable.

```
>hist(marks, right=FALSE)
```



# Visualization Bivariate data (Qualitative (*ordinal*), Qualitative (*ordinal*))

- > mosaicplot(~ gear + carb, data = mtcars, color = 2:3, las = 1)

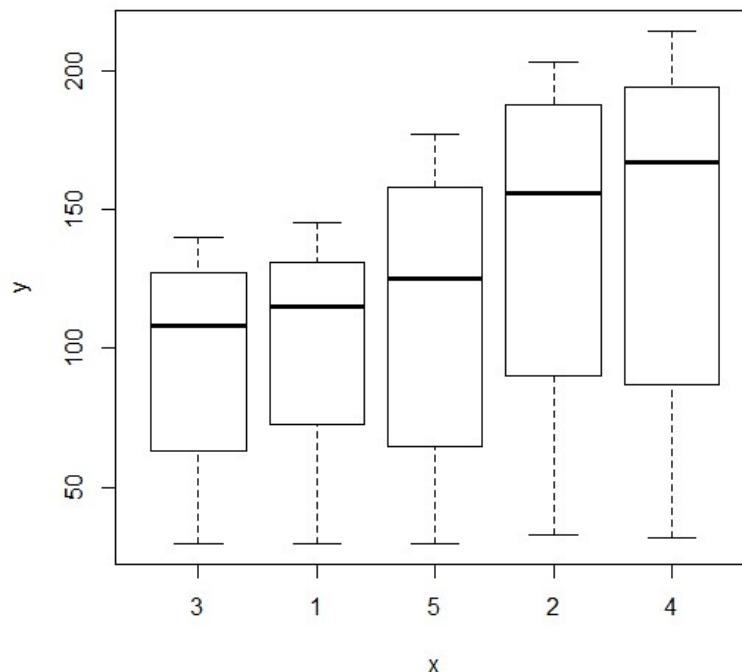


# Visualization Bivariate data (Qualitative (*ordinal*), Quantitative )

## Box Plot

```
>plot(Orange$Tree,Orange$circumference)
```

Tree is ordinal and, circumference is quantitative

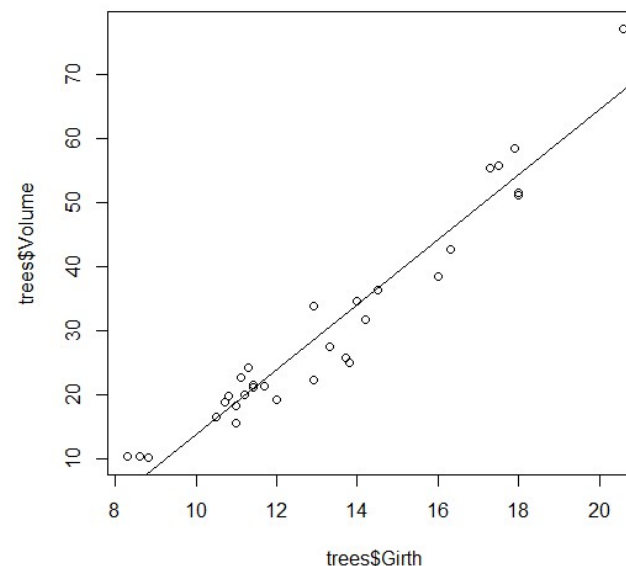
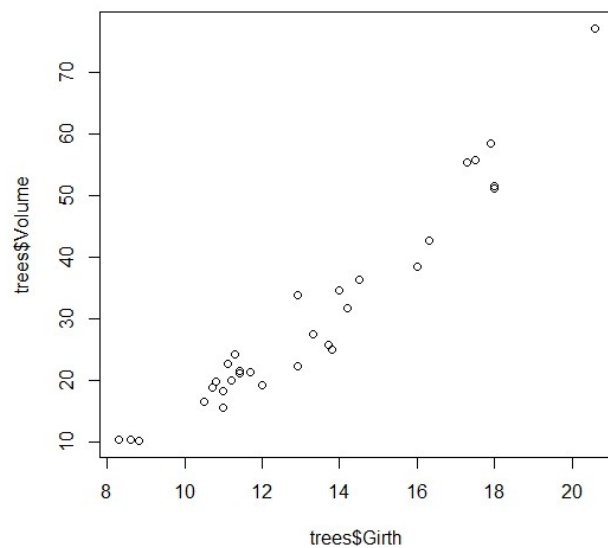


# Visualization Bivariate data (Quantitative, Quantitative )

**Scatter Plot** (Girth and Volume both quantitative)

```
>plot(trees$Girth, trees$Volume)
```

```
>abline(lm(trees$Volume~trees$Girth))
```



# Some Statistical Measures(Uni-variate)

- Mean
- Median
- Quartile
- Range
- Interquartile Range
- Variance
- Standard Deviation

# Some Statistical Measures Cont..

**Mean:** The mean of an observation variable is a numerical measure of the central location of the data values. It is the sum of its data values divided by data count.

```
>marks=c(12,12,14,15,13,14,16,17,12,15,18,18,19,20,12,13,12,14,15)  
>mean(marks)
```

**Median:** The median of an observation variable is the value at the middle when the data is sorted in ascending order. It is an ordinal measure of the central location of the data values.

```
>median(marks)
```

# Some Statistical Measures Cont..

**Quartile:** There are several quartiles of an observation variable. The first quartile, or lower quartile, is the value that cuts off the first 25% of the data when it is sorted in ascending order. The second quartile, or median, is the value that cuts off the first 50%. The third quartile, or upper quartile, is the value that cuts off the first 75%.

`>quantile(marks)`

**Range:** The range of an observation variable is the difference of its largest and smallest data values. It is a measure of how far apart the entire data spreads in value.

`>range(marks)`

# Some Statistical Measures Cont..

**Inter-quartile range:** The inter-quartile range of an observation variable is the difference of its upper and lower quartiles. It is a measure of how far apart the middle portion of data spreads in value

`>IQR(marks)`

**Variance:** The variance is a numerical measure of how the data values is dispersed around the mean.

`>var(marks)`

**Standard Deviation:** The standard deviation of an observation variable is the square root of its variance.

`>sd(marks)`



# t test

- Parametric Test: works on normally distributed scale data.
- Compares Two means
- There are different version for different design
  - One Sample i.e. One sample t-test
  - Dependent (related) Sample i.e. Paired t-test
  - Independent (unrelated) Sample i.e. Two sample t-test

# Hypothesis

- Null hypothesis: The sample come from the same population  $H_0$
- Alternative Hypothesis: The samples come from different populations  $H_1$

# t.test() Function

## General Form

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),  
mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

## Example

```
t.test (x, alt="less", mu=10)
```

```
t.test (x, y, alt="two.sided")
```

```
t.test(x, y, alt="less", paired = T)
```

# One Sample t-test

A sample of 24 people is taken. The length of time to prepare dinner is recorded in minutes, as given below

44.0, 51.9, 49.7, 40.0, 55.5, 43.4, 41.3, 45.2,  
40.7, 41.1, 49.1, 30.9, 45.2, 55.3, 52.1, 55.1,  
38.8, 43.1, 39.2, 58.6, 49.8, 43.2, 47.9, 46.6

Is there any evidence that the population mean time to prepare dinner is less than 48 minutes? Use a level of significance of 0.05.

# One Sample t-test

Step 1: One sample, Normally distributed ,one sample t test

Step 2

Null Hypothesis  $H_0: \mu = 48$

Alternative Hypothesis  $H_1: \mu < 48$

Step 3

>p=c (44.0, 51.9 .....47.9, 46.6)

>t.test(p, alt="less", mu=48 )

One Sample t-test

data: p

t = -1.7044, df = 24,

alternative hypothesis: true mean is less than 48

95 percent confidence interval:

-Inf 48.00909

sample estimates:

mean of x

45.628

# Interpretation one sample t test

The p-value is a number between 0 and 1 and interpreted in the following way

## **Right tailed or left tailed (95% confidence interval)**

- A small p-value (typically  $\leq 0.05$ ) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. ( $H_1$  ✓)
- A large p-value ( $> 0.05$ ) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis. ( $H_0$  ✓)

## **Two Tailed (95% confidence interval)**

- A small p-value (typically  $\leq 0.025$ ) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. ( $H_1$  ✓)
- A large p-value ( $> 0.025$ ) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis. ( $H_0$  ✓)

Since P value 0.0506 is greater than 0.05 so  $H_0$  is accepted

# Paired t-test

Compare the means of two conditions in which the same ( or closely matched) participants participated.

# Paired t-test

A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded. The mileage was recorded again for the same cars using the other kind of gasoline. Determine whether cars get significantly better mileage with premium gas.

Mileage

Regular 16,20,21,22,23,22,27,25,27,28

Premium 19,22,24,24,25,25,26,26,28,32



# Paired t-test

Step 1 Data Normally distributed, Paired

Step 2

H0 : Difference of mean mileage of premium gas with regular gas is equal

H1 : Difference of mean mileage of premium gas with regular gas is greater

Step 3

```
> reg=c(16,20,21,22,23,22,27,25,27,28)
```

```
> pre=c(19,22,24,24,25,25,26,26,28,32)
```

```
> t.test (pre, reg, alt="greater", paired=T)
```

Paired t-test

data: pre and reg

t = 4.4721, df = 9, p-value = 0.0007749

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

1.180207 Inf

sample estimates:

mean of the differences

2

# Two Sample t-test

- Compare the means of two groups of participants
- Groups are independent

# Two Sample t-test

6 subjects were given a drug (treatment group) and an additional 6 subjects a placebo (control group). Their reaction time to a stimulus was measured (in ms).

Control : 91, 87, 99, 77, 88, 91

Treat : 101, 110, 103, 93, 99, 104

To perform a two sample t-test for comparing the means of the treatment and control groups.

# Two Sample t-test

```
>Control = c(91, 87, 99, 77, 88, 91)
>Treat = c(101, 110, 103, 93, 99, 104)
> t.test(Control,Treat,alternative="less")
```

Two Sample t-test

data: Control and Treat

t = -3.4456, df = 10, p-value = 0.003136

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -6.082744

sample estimates:

mean of x mean of y

88.83333 101.66667

Since p value is 0.003136 is less than 0.05 so H1 is accepted.

# Chi-square tests

## **Goodness of fit**

A goodness of fit test, checks to see if the data came from some specified population.

## **Test for Independence**

Two random variables  $x$  and  $y$  are called independent if the probability distribution of one variable is not affected by the presence of another.

# Chi square Goodness of fit

## Step 1

Prepare frequency vector

Prepare corresponding Probability vector

## Step 2

$H_0$  : Fit is good

$H_1$  : Fit is not good

## Step 3

Use R Function

`>chisq.test( frequency vector, p=probability vector)`

## Step 4

Give result based on p-value

# Chi square Goodness of fit

If we toss a die 150 times and, find that we have the following distribution of rolls is the die fair?

Face	1	2	3	4	5	6
Number of rolls	22	21	22	27	22	36

Solution

H<sub>0</sub>: Die is fair, H<sub>1</sub>: Die is not fair

```
>freq=c(22,21,22,27,22,36)
```

```
>prob=c(1/6,1/6,1/6,1/6,1/6,1/6)
```

```
>chisq.test(freq,p=prob)
```

Chi-squared test for given probabilities

data: freq

X-squared = 6.72, df = 5, p-value = 0.2423

P-value is greater than 0.05 so that H<sub>0</sub> accepted.

# Chi square test for Independence

## Step 1

Prepare vectors

Prepare data frame

## Step 2

H0: Attributes are independents

H1: Attributes are not independents

## Step 3

Use R Function

```
>chisq.test(data_frame)
```

## Step 4

Give result based on p-value



# Chi square test for Independence

From a survey conducted in different regions (rural and urban)  
To know preference of persons to different television programs  
(educational and entertainment). Following data was obtained.

	Education	Entertainment
Rural	45	35
Urban	20	50

Test whether preference to a program depends on region.

# Chi square test for Independence

## Step 1

```
>rural =c(45,35)
>urban=c(20,50)
>survey1=data.frame(rural,urban)
```

## Step 2

H0: Performance of program independent on region

H1: Performance of program dependent on region

## Step3

Use R Function

```
>chisq.test(survey1)
```

Pearson's Chi-squared test

data: survey1

X-squared = 11.648, df = 1, p-value = 0.0006429

## Step4

Based on P ( $0.00064 < 0.05$ ) value Ho is rejected and H1 accepted, i.e. Performance of program dependent on region.

# Correlation

Correlation is used to test for a relationship between two numerical variables or two ranked (ordinal) variables.

Usually, in statistics, we measure three types of correlations:

1. Pearson correlation
2. Kendall rank correlation
3. Spearman correlation

Pearson  $r$  correlation-Parametric

Kendall Rank and Spearman correlation –Non Parametric

# Correlation

A simplified format is

**`cor(x, method= )`**

where

x: data frame

Method: Specifies the type of correlation.

Options are pearson (default), spearman or kendall.

# Correlation

## Exercise

Protein intake X and fat intake Y (in gm) for ten old women given as

X 56,47,33,39,42,38,46,47,38,32

Y 56,83,49,52,65,52,56,48,59,70

Calculate correlation Coefficient (Pearson) , draw scatter plot matrix and scatter plot

## Exercise

Find correlation coefficient (Pearson) between the sales and expenses from the data given below:

Firm: 1,2,3, 4,5,6,7,8,9,10

Sales (Rs Lakhs) 50,50,55,60,65,65,65,60,60,50

Expenses (Rs Lakhs): 11,13,14,16,16,15,15,,14,13,13

Draw scatter plot matrix, and scatter plot

# Correlation

## Exercise

Protein intake X and fat intake Y (in gm) for ten old women given as

X 56,47,33,39,42,38,46,47,38,32

Y 56,83,49,52,65,52,56,48,59,70

Calculate correlation Coefficient (Pearson) , draw scatter plot matrix and scatter plot

## Exercise

Find correlation coefficient (Pearson) between the sales and expenses from the data given below:

Firm: 1,2,3, 4,5,6,7,8,9,10

Sales (Rs Lakhs) 50,50,55,60,65,65,65,60,60,50

Expenses (Rs Lakhs): 11,13,14,16,16,15,15,,14,13,13

Draw scatter plot matrix, and scatter plot

# Simple Linear Regression

- ✓ A simple linear regression model describes the linear relationship between two variables, and the model equation can express as per equation,  $x$  is an independent variable, and  $y$  is the dependent variable. The constant numbers  $\alpha$  and  $\beta$  are called parameters, and  $\epsilon$  is the error term.
- ✓ Model equation:

$$Y = \alpha + \beta * x + e$$

(1)

# Significance test

- ✓ The significance test of linear regression focuses on the coefficient  $\beta$  of the regression model equation. Where  $\alpha$  is a constant and  $\beta$  is the slope.
- ✓ If the  $\beta$  is significantly different from zero then concluded, there is a significant relationship between the independent and dependent variables.
- ✓ Hypothesis for testing significance for linear regression as  
$$H_0: \beta = 0, H_a: \beta \neq 0.$$
- ✓ It is a point of consideration when the value of  $\beta$  in model equation is zero; then equation 1 represents constant value for all independent values.



# Residual

- ✓ Residual: Residual ( $e$ ) is a measure that obtained as a difference between the actual value of the dependent(target) variable ( $y$ ) and the predicted value ( $\hat{y}$ ) using model.
- ✓ Each data point has one residual.
- ✓ Residual = Observed value - Predicted value

# Coefficient of determination

- ✓ The coefficient of determination (denoted by  $R^2$  and ranges from 0 to 1) is a crucial measure of regression analysis.
- ✓  $R^2$  is described as the proportion of the variance in the dependent variable that is predictable from the independent variable.
- ✓  $R^2$  equal to 0, means that the dependent variable cannot be predicted and  $R^2$  equal to 1, indicate the dependent variable can be predicted from the independent variables without error.

# Example (Simple Linear Regression)

In the following table are recoded data showing the test score and their weekly sales

Table-1

## Sales and Test Score

Salesmen	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Test Score	40, 70, 50, 60, 80, 50, 90, 40, 60, 60
Sales in 000 Rs	2.5, 6.0, 4.0, 5.0, 4.0, 2.5, 5.5, 3.0, 4.5, 3.0

Develop a Simple Linear Regression Model based on above data.

# R Commands

```
> ts= c(40, 70, 50, 60, 80, 50, 90, 40, 60, 60)
```

```
> ws=c(2.5, 6.0, 4.0, 5.0, 4.0, 2.5, 5.5, 3.0, 4.5, 3.0)
```

```
> mode=lm(ws~ts)
```

```
> summary(mode)
```

```
Call:
```

```
lm(formula = ws ~ ts)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-1.08333	-0.82292	-0.02083	0.53125	1.45833

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.75000	1.17969	0.636	0.5427
ts	0.05417	0.01904	2.845	0.0216 *

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.9326 on 8 degrees of freedom
```

```
Multiple R-squared:  0.503,    Adjusted R-squared:  0.4408
```

```
F-statistic: 8.096 on 1 and 8 DF,  p-value: 0.02163
```

# Model

The model equation is as

$$\underline{\text{Weekly Sales} = 0.75 + 0.05417 * \text{Test Score} + \text{Error}}$$

The p-value of the score is 0.0216, which is less than 0.05, so that is evidence of alternation hypothesis. so that it is proof there is a significant relationship between sales and score.

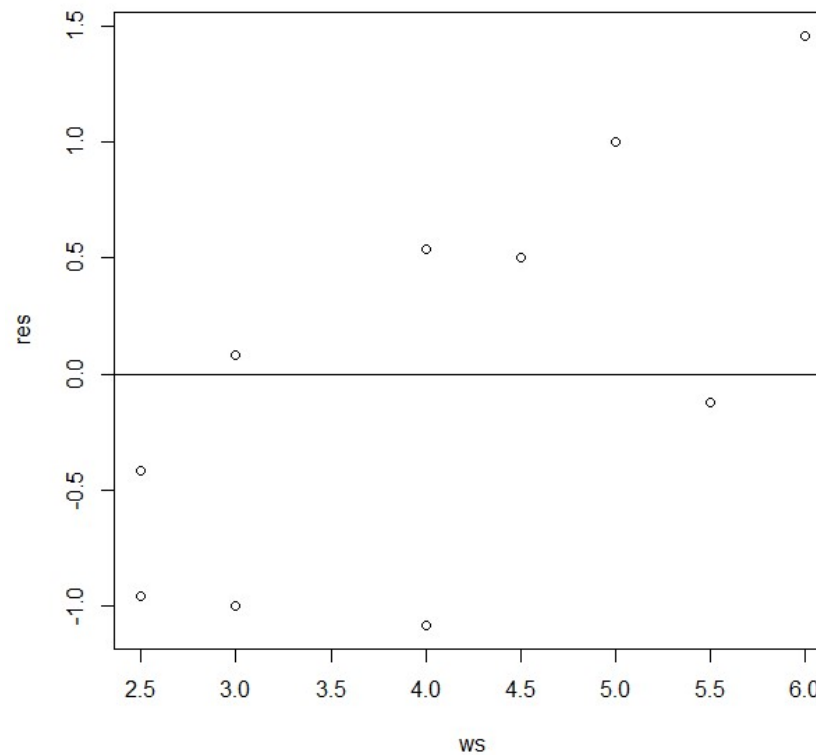
Coefficient of determination  $R^2$  is 0.503.

# Visualization of Simple Linear Regression

```
>res=resid(mode)
```

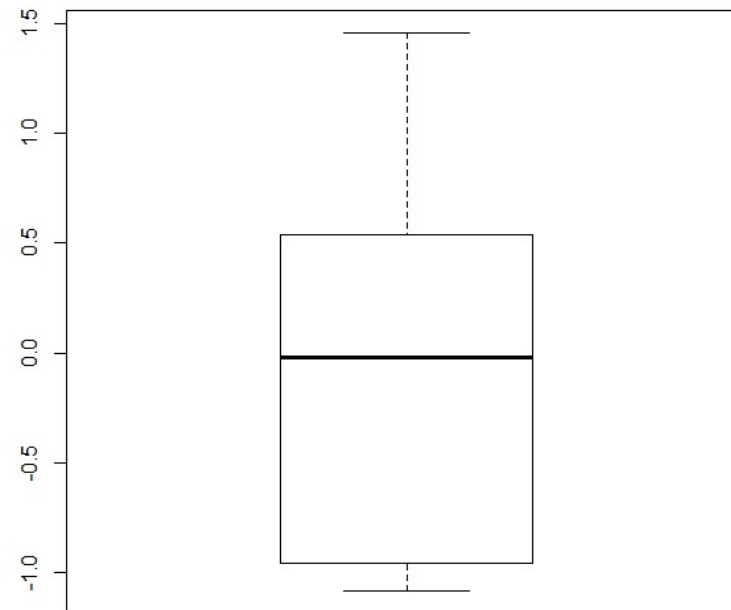
```
>plot(ws,res)
```

```
>abline(0,0)
```



# Visualization of Simple Linear Regression

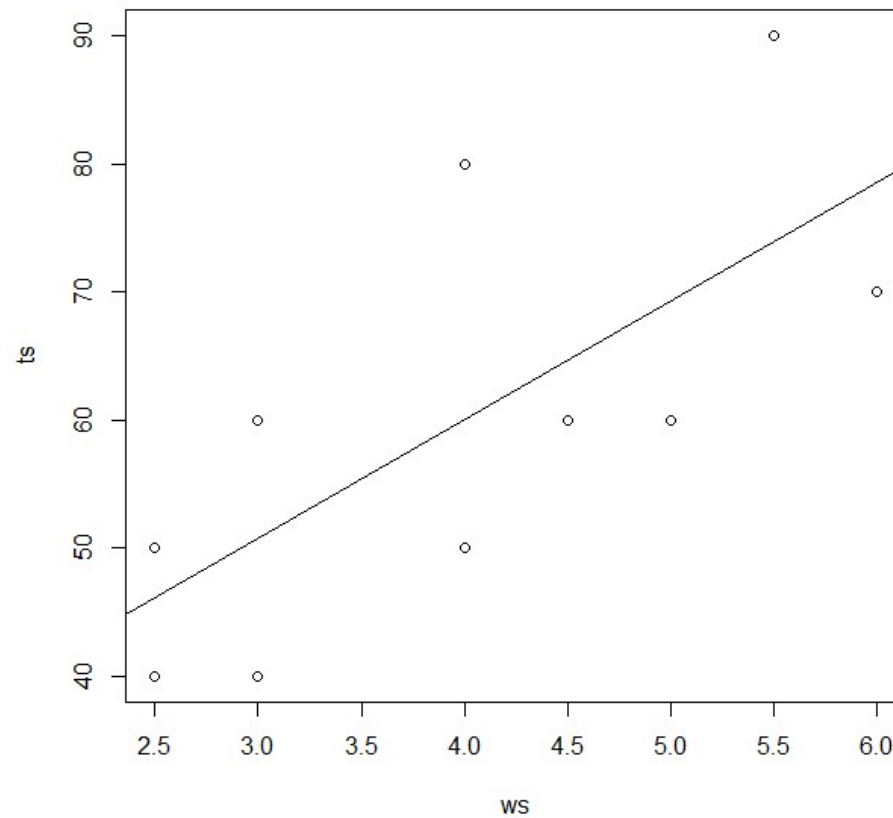
```
>boxplot(res)
```



# Visualization of Simple Linear Regression

```
> plot(ws,ts)
```

```
> abline(lm(ts~ws))
```





# Multiple Linear Regression

- ✓ A multiple linear regression model is advance regression model of simple linear regression model that describes a dependent variable  $y$  by many independent variables  $x_1, x_2, \dots, x_p$  ( $p > 1$ ) is expressed by the equation as follows.

$$Y = \alpha + \beta_1 * x_1 + \beta_2 * x_2 + \dots + e$$

(2)

# Example (Multiple Linear Regression)

An analyst was studying a chemical process expects the yield to be affected by the levels of two factors,  $x_1$  and  $x_2$ . Observations recorded for various levels of the two factors are shown in the following table 2. The analyst wants to fit a regression model to the data interaction between  $x_1$  and  $x_2$ .

•

**Table-2: Observations for Various Levels of The Two Factors  
And Yield**

Observation Number	Factor 1 ( $x_{i1}$ )	Factor 2 ( $x_{i2}$ )	Yield ( $y_i$ )
1	41.9	29.1	251.3
2	43.4	29.3	251.3
3	43.9	29.5	248.3
4	44.5	29.7	267.5
5	47.3	29.9	273.0
6	47.5	30.3	276.5
7	47.9	30.5	270.3
8	50.2	30.7	274.9
9	52.8	30.8	285.0
10	53.2	30.9	290.0
11	56.7	31.5	297.0
12	57.0	31.7	302.5
13	63.5	31.9	304.5
14	65.3	32.0	309.3
15	71.1	32.1	321.7
16	77.0	32.5	330.7
17	77.8	32.9	349.0

# Model (Multiple Linear Regression)

```
>x1=c(41.9,43.4,43.9,44.5,47.5, 47.3,47.9,50.2,52.8,53.2,56.7,57.0,63.5,65.3,71.1,77.0,77.8)
>x2=c(29.1, 29.3, 29.5, 29.7, 29.9, 30.3, 30.5, 30.7, 30.8, 30.9, 31.5, 31.5, 31.9, 32.0, 32.1,32.5, 32.9 )
>y=c(251.3,251.3,248.3,267.5,273.0,276.5,270.3,274.9,285.0,290.0,297.0,302.5,304.5,309.3,321.7,330.7,349.0
)
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min       1Q   Median       3Q      Max
-9.0145 -4.2791 -0.6348  4.5033  8.4946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -162.5937   109.1279  -1.490   0.1584
x1             1.2065     0.4243   2.844   0.0130 *
x2            12.4388     4.2554   2.923   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.607 on 14 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.962
F-statistic: 203.5 on 2 and 14 DF,  p-value: 4.492e-11
```

# Model (Multiple Linear Regression)

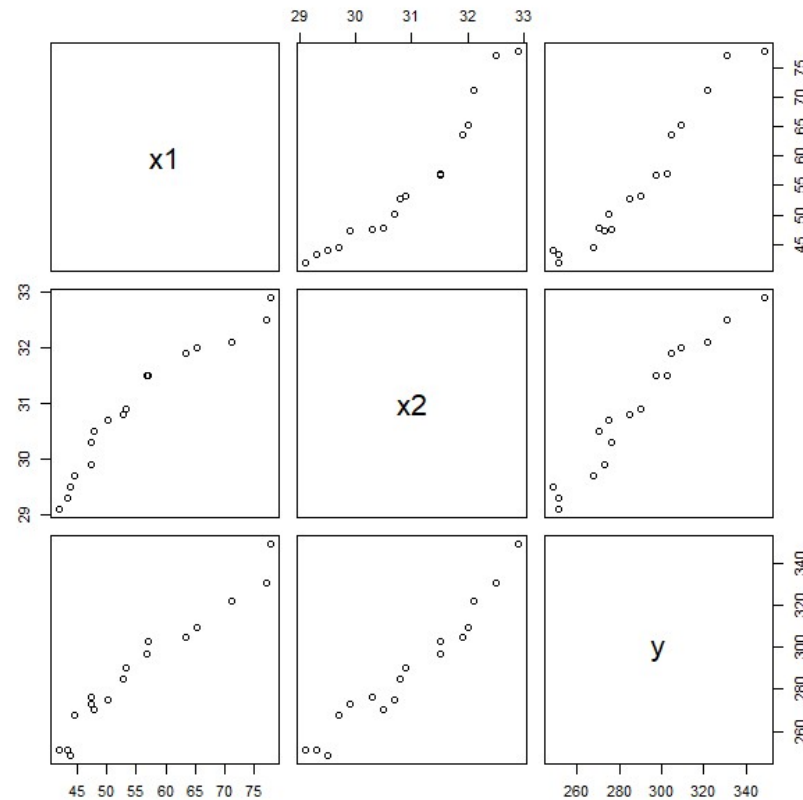
- ✓ The p-value of the  $x_1$  is 0.00724, which is less than 0.05, so there is evidence of alternation hypothesis. so that it is proof there is a significant relationship between  $y$  and  $x_1$ .
- ✓ The p-value of the  $x_2$  is 0.00827, which is less than 0.05, so there is evidence of alternation hypothesis. so that it is proof there is a significant relationship between  $y$  and  $x_2$ .
- ✓ Coefficient of determination  $R^2$  is 0.968
- ✓ The proposed multiple linear model equation is

$$y = -162.5937 + 1.2065 * x_1 + 12.065 * x_2 + \text{Error.}$$

# Visualization of Multiple Linear Regression

```
> df=data.frame(x1,x2,y)
```

```
> pairs(df)
```



# Question ?



“Thank You”

