Sub :

Day

Time : Date : / /

Introduction to Statistics

Definition: It is the science of collecting, analizing and organizing data.

Data: Facts or pieces of information

Example → height of students in a classroom
(178 cm, 180 cm, 195 cm)

Type of Statistics

1. Descriptive Statistics
2. Inferential Statistics

Descriptive Statistics:

It consists of organising and summarizing data.

It contains →
1. Measure of central tendancy (Mean, Median, Mode)
2. Measure of Dispersion (Variance, STD)
3. Different types of distribution of data

Sub :

Day

Time : Date : / /

② **Inferential statistics:** It consists of using data you have measured to form conclusion.

It contains → ① Z test, ② T test, ③ } Hypothesis testing: Ho, H1, P value, significance value

Understanding more about descriptive and inferencial statistics:

Suppose a school has 10 classes. From, 1 of the classes you are measuring the height of the students.

data = (130 cm, 150 cm, 165 cm, 165 cm, 170 cm, 172 cm)

Descriptive Statistics question:

What is the average height of the entire classroom?

Inferential statistics question:

Are the heights of the students in the classroom similar to what you expect in the entire college

Here, entire college = Population

classroom = Sample

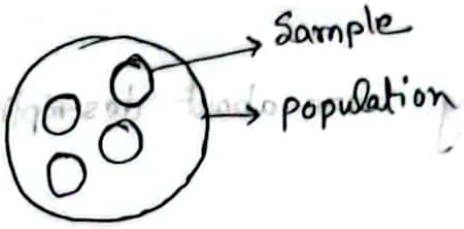
Sub :

Time : Date : / /

Population and Sample data:

Population: The group you are interested in studying

Sample: A subset of population.



- Data type
  - (Numbers) Quantitative
    - Discrete
      - Whole numbers
      - Examples,
        - Age,
        - No of Bank Acct.,
        - Num of family members
    - Continuous
      - Any values
      - Height, weight, Speed, Temp
  - Qualitative (Text) (Categorical)
    - Nominal
      - Exp →
        - Gender,
        - Blood Group,
        - Departments,
        - Pincode,
        - State
    - Ordinal (We assign Rank)
      - customer feedback
      - price category, (low, mid, high)

Sub :

Day

Time : Date : / /

Scale of data :

① Nominal scale data

→ Qualitative/Categorical data

→ Exp : Gender, colors,

→ Data which we can't rank (Text data)

→ Order does not matter

What analysis can we do from nominal scale data?

Suppose, there are 10 people. Each will say their favourite color.

From 10 people, Red → 5 people → 50% (5 out of 10 people)

Blue → 3 people → 30% (3 out of 10 people)

Orange → 2 people → 20% (2 out of 10 people)

Total = 10 people

Ordinal scale data :

→ Ranking is important

→ Order matter

→ Difference can't be measured

Sub :

Time : Date : / /

What analysis can we get from ordinal sale data?

Suppose, we have a customer review ~~fea~~ column.

→ 5 * (Rank 1) (Best)

→ 3 * (Rank 2) (Better)

→ 1 * (Rank 3) (Good)

Hence we can identify Rank 1 > Rank 2 > Rank 3. We can rank the feedbacks and draw conclusion about which one is better.

Interval scale of Data:

→ The order matter

→ Difference can be measured

→ Ratio can't be measured

→ No True "0" starting point

Let's take an example of temperature column.

~~Id~~ Temp
30°F
60°F
90°F
120°F

→ Here difference can be measured

60 - 30 = 30

Sub :

Day

Time : Date : / /

Temp
30°F
60°F
90°F
120°F

Ratio $\frac{30}{60} = \frac{1}{2} = 1:2$

Now, that doesn't mean 60°F heat is produce the double hit of 30°F though the value is double.

So, Ratio can't be measured.

Then, no True "0" starting point. In the case of temperature column, tempercatures also can be negative. Like -30°F, -60°F. So the values don't start from 0.

Ratio Scale Data:

→ The order matters

→ Differences are measurable

→ Ratios are also measurable

→ Contains a "0" starting point

Students mark → ~~30, 60~~

90
60
30
75
40
50

→ We can order them in ascending or descending

→ We can make desicions by ordering them so order matters.

Sub :

Day

Time : Date : / /

Marcks

90
60
30
40
75
50

$90 - 60 = 30$ → differences are measurable

$90/60 = \frac{3}{2} = 3:2$ Ratios are measurable and have meaning

→ Marcks can stant from 0. Marks can be -10

Some more Examples:

1) Nominal scale of data

→ Gender → Makital status → Eye color

→ Eye color → Blood type → Vehicle type

→ Departments → Pet ownership → Qualification

2) Ordinal scale data:

→ Social class → Food review → Performance Rating

→ Response scale → Pain Severity → Star Ratings

→ Order of finishing a race → Health condition saverity.

Sub :

Day

Time : Date : / /

Interval scale data:

→ Temperature in degree on F → Calendar dates → IQ scores

→ Time → SAT/GRE scores → pH scale → Musical pitch

→ Latitude and Longitude

Ratio Scale data:

→ Height → Weight → Age → Time in seconds → Amount of money

→ Distance → Number of items → Energy consumption

Scale of Measurement:

End of Introduction To Statistics

Sub :

Time : Date : / /

Measure of central tendency:

① Mean:

Population (N)

$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

$$\text{population mean } (\mu) = \sum_{i=1}^{N} \frac{X_i}{N}$$

$$= \left[\frac{1+1+2+2+3+3+4+5+5+6}{10}\right]$$

$$= \frac{32}{10} = 3.2$$

Sample (n)

$x = \{4, 5, 5, 6\}$

$$\text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$= \left[\frac{4+5+5+6}{4}\right]$$

$$= \frac{19}{4}$$

② ~~Mt~~ Median:

$X = \{4, 5, 2, 3, 2, 1\}$

→ Sort the variable ⇒ $X = \{1, 2, 2, 3, 4, 5\}$

→ Number of elements ⇒ count = 6

→ if count == even → $\{1, 2, (2, 3), 4, 5\}$

middle value

$\therefore$ median $= \frac{2+3}{2} = 2.5$

Sub :

Day

Time : Date : / /

→ if count = odd, suppose → $x = \{1, 1, 2, 3, 5, 8, 9\}$

↳ middle value

then,

↳ median = 3

(*) Why median instead of mean?

Outlier

$x = \{1, 2, 3, 4, 5\}$

(Mean) $\bar{x} = \frac{1+2+3+4+5}{5}$

$= 3$

$x = \{1, 2, 3, 4, 5, 100\}$

$\bar{x} = \frac{1+2+3+4+5+100}{6}$

$= \frac{115}{6} = 19$

For, having outliers, the mean value can be totally wrong. And the value can be shifted by a great amount.

So, if we ~~take~~ sort the values, then take the middle element that would be more accurate. That's why we need median, so the outliers won't effect the measures.

Mode: (Maximum Frequency)

$x = \{1, 3, 2, 2, 2, 2, 5, 7, 8, 8, 8, 9, 10\}$

Mode = 2 (which has the maximum frequency of 4)

Sub :

Day

Time : Date : / /

Measure of dispersion: (Spread of the data)

→ Variance → Standard Deviation

① Variance:

→ Population Variance (σ^2)

Formula → $$\sigma^2 = \sum_{j=1}^{N} \frac{(x_i - \mu)^2}{N}$$

x_i = data points

μ = population mean

N = Population size

→ Sample Variance

Formula $$s^2 = \sum_{j=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

x_i = data points

$\bar{x}$ = Sample mean

$n-1$ = sample size

⊛ Why we devide sample variance by n-1

The sample variance is devided by n-1 so that we can create an unbiased estimator of the population.

This whole thing is called: Basscle correction.

Sub :

Day

Time : Date : / /

Example of sample variance:

$x = \{1, 2, 3, 4, 5\}$

x	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
		Sum = 10

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$$= \frac{10}{5-1} = \frac{10}{4} = 2.5$$

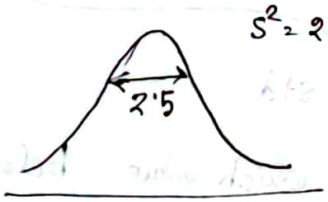
⊛ Why we calculate variance? What it shows?

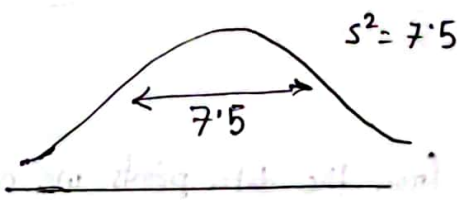
We ~~show~~ calculate variance (sample variance (S^2)) so that we can know the spread or dispersion of data.

Suppose for x_1 data points, variance $S^2 = 2.5$

for x_2 " " " $S^2 = 7.5$

If we plot them we can see the spread of data





Sub :

Time : Date : / /

② Standard Deviation:

→ Population SD → Sample SD

Population std:

$$\sigma = \sqrt{\text{variance } (\sigma^2)}$$

Sample std:

$$s = \sqrt{s^2} \text{ (sample variance)}$$

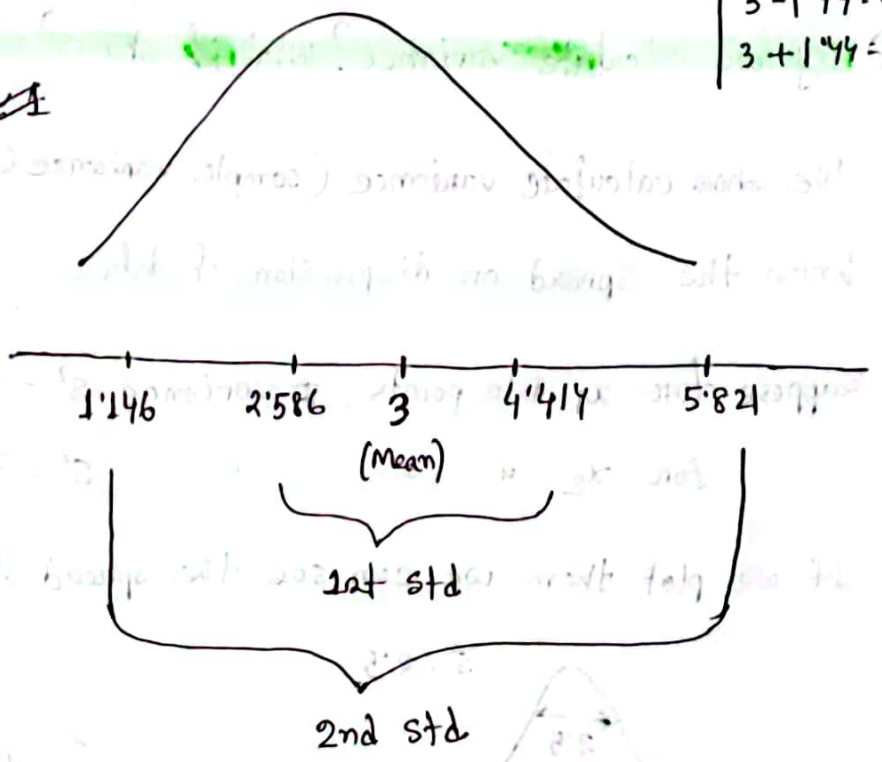
(*) What is the purpose of std?

$x = \{1, 2, 3, 4, 5\}$

mean, $\bar{x} = 3$

$\sigma = 1.44$ ~~≈ 1~~

$3 - 1.44 = 2.586$
$3 + 1.44 = 4.414$



from the data points we can decide which value belongs to which std range. Like $x = 4$, belongs to 1st std.

Sub :

Day

Time : Date : / /

Random Variable :

Random variable is the process of mapping the output of a random process or equipments to a number.

For example,

Tossing a coin $x = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if tail} \end{cases}$

So we are mapping the output of a toss (Head to 0)
(Tail to 1)

This process is called random variable.

Another example:

$$y = \{ \text{Sum of rolling of dice 7 times} \}$$

Here, y is also a random variable. Because we don't know what the outcome actually might be.

Sub :

Time : Date : / /

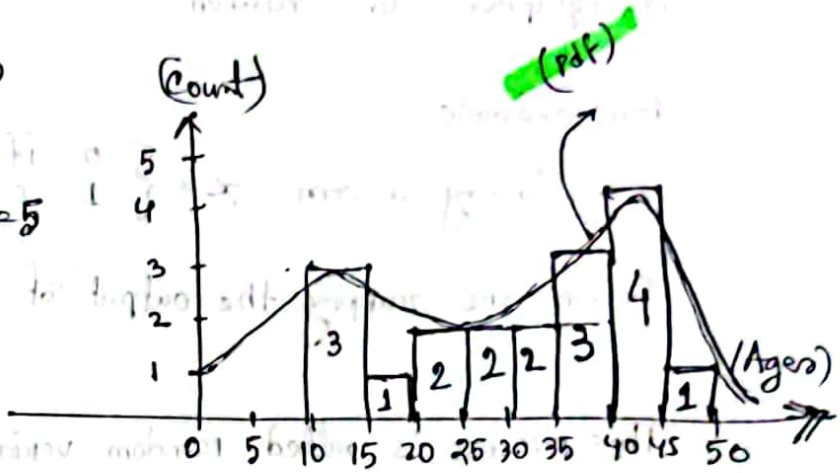
Histogram and skewness :

Ages = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 48, 50 }

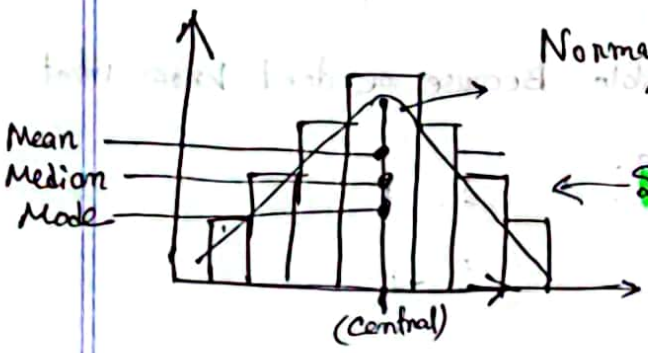
Range ~~0 - 60~~ 0 - 50

Bin number = 10

So, Bin size = $\frac{50}{10} = 5$



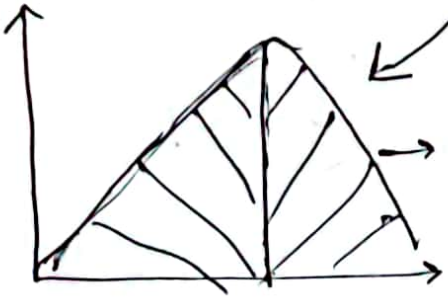
(Pdf = Probability ~~distro~~ distribution ~~distribution density~~ function)



Normal / Gaussian distribution

← Symmetrical Distribution

↳ There will be no skewness



→ Bell curve

In this distribution

[Mean = Median = Mode]

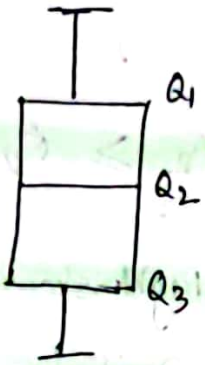
→ The mean, median and mode all are perfectly at the centre in this type of distribution.

Sub :

Day

Time : Date : / /

Box plot (For symmetrical distribution)

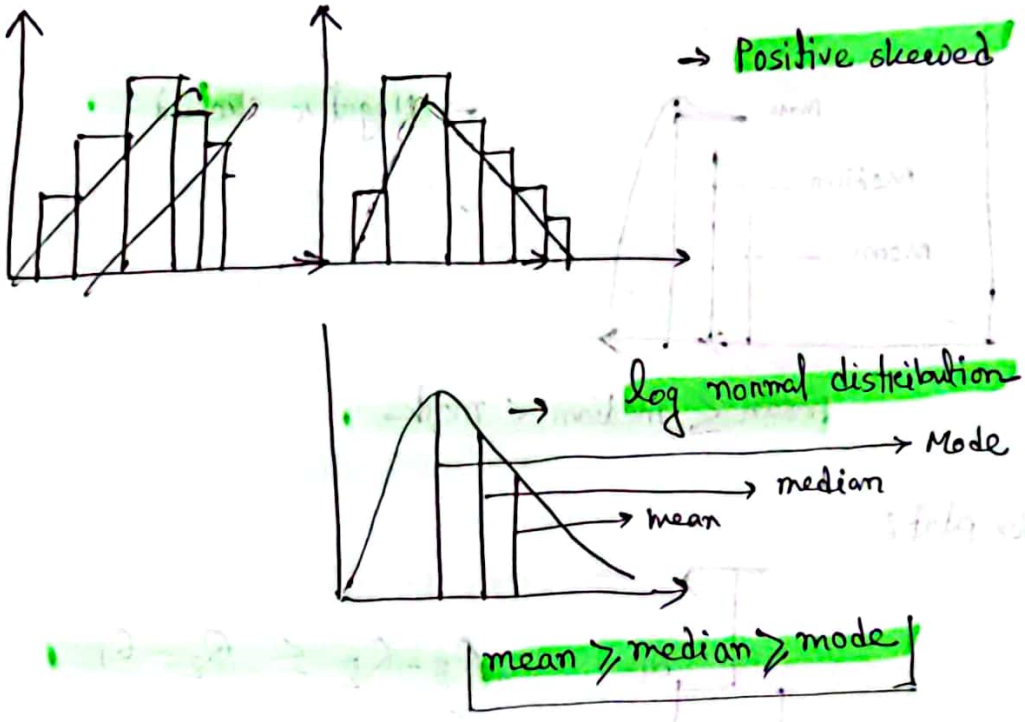


$Q_3 - Q_2 \approx Q_2 - Q_1$

This property also is for symmetric distribution.

→ No skewness

Right skewed distribution



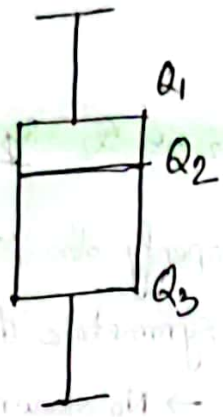
mean ≥ median ≥ mode

Sub :

Day

Time : Date : / /

Box plot: (log normal distribution) (~~Left~~ Right skewed)

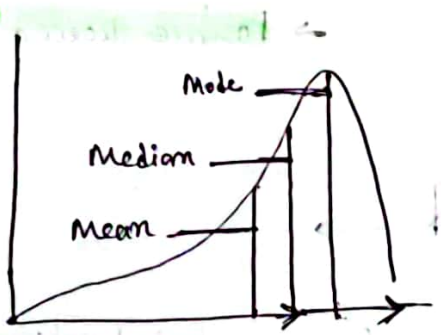


$Q_3 - Q_2 \geq Q_2 - Q_1$

↳ Right skewed

↳ Positively skewed.

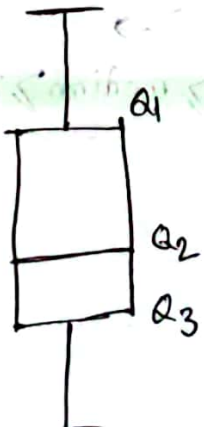
Left skewed Distribution



→ Negative skewed

mean ≤ median ≤ mode

Box plot:



$Q_3 - Q_2 \leq Q_2 - Q_1$

↳ Left skewed

↳ Negatively skewed.

Sub :

Day

Time : Date : / /

Covariance and Correlation:

x	y
2	3
4	5
6	7
8	2

[Relationship between x and y]

x ↑	y ↑
x ↓	y ↑
x ↑	y ↓
x ↓	y ↓

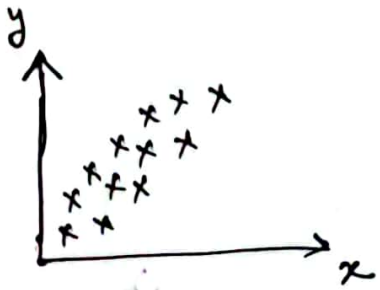
→ These relations are covariance and correlation

↑ → increasing

↓ → decreasing

x ↑	y ↑
x ↓	y ↓

Plot →



x ↑	y ↓
x ↓	y ↑

Plot →

What is the need of using it?

Suppose in a real life ml model for ea example the house price prediction model, there are two features/colums completely depend upon each othe

house size ↑ → house price ↑

house size ↓ → house price ↓

Sub :

Time : Date : / /

So, these relationships are really very necessary to predict the best outcome for the machine learning model.

Covariance:

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

x_i = x data prints

$\bar{x}$ = Sample mean (x)

y_i = y data prints

$\bar{y}$ = Sample mean (y)

=

We know,

$$\text{Variance } V = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$$\text{We can write} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

which is $= Cov(x,x)$

[Means x relation with x)
(which shows the spread of data)

So, the relation between variance and covariance is variance is a kind of covariance which shows the relation with it's ownself.

Sub :

Now **Positive Covariance** → $x\uparrow\ y\uparrow$, $x\downarrow\ y\downarrow$

Negative Covariance → $x\uparrow\ y\downarrow$, $x\downarrow\ y\uparrow$

Example of positive covariance:

x	y
2	3
4	5
6	7

$\bar{x} = 4,\ \bar{y} = 5$

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$= \frac{[(2-4)(3-4) + (4-4)(5-5) + (6-4)(7-5)]}{3-1}$$

$$= \frac{4+0+4}{2} = \frac{8}{2} = 4$$

which is a (+ve) covariance

Advantage

① Can find relationship between X and Y

Disadvantage

① Covariance does not have a specific limit value. For which sometimes it can't be measured how strongly a random variable (x) is dependent on another random variable (y)

Sub :

Day

Time : Date : / /

To tackle with the disadvantage of covariance, we use Pearson Correlation Coefficient.

The ~~out of~~ output of this method is always in the range [-1, 1]

The more the value towards +1, means the more positively correlated is x to y variable.

The more the value towards −1, means the more negatively correlated is x to y variable.

Formula, $\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \, \sigma_y}$

Another method: Spearman Rank Collection

Range [-1, 1]

R(x) : Rank of x

R(y) : Rank of y

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) \times \sigma(R(y))}$$

Sub :

Time : Date : / /

Example

x	y	R(x)	R(y)
1	2	5	5
3	4	4	4
5	6	3	3
7	8	2	1
0	7	6	2
8	1	1	6

[Here R(x) = 1, when x value is maximum]

[Here, R(y) = 1, when y value is maximum]

$$\text{Cov}(R(x), R(y)) = \sum_{i=1}^{n} \frac{(R(x_i) - R(\bar{x}))(R(y_i) - R(\bar{y}))}{n-1}$$

$$= \frac{(5-21)(5-21) + (4-21)(4-21) + (3-21)\times(3-21) + (2-21)(1-21) + (6-21)(2-21) + (1-21)(6-21)}{5}$$

$$= \frac{256 + 289 + 324 + 380 + 285 + 300}{5}$$

$$= \frac{1834}{5} = 366.8$$

$$\sigma(R(x)) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}} = \sqrt{\frac{256 + 289 + 324 + 361 + 225 + 400}{5}}$$

$$= 19.26$$

$$\sigma(R(y)) = 19.26$$

Sub :

Page :

Time : Date : / /

$$R_S = \frac{Cov(R_x, R_y)}{\sigma_{R_x}\, \sigma_{R_y}}$$

$$= \frac{366.8}{19.26 \times 19.26}$$

$= 0.989$ which is in rang $[-1, 1]$

and +ve correlation of x to y

Why we use this technique ? (Spearman Rank Collection)

During feature selection process, when we measure multiple column correlation with a single column, that time the column, whose R_S comes near 0, that means the column doesn't have a proper correlation with the main column which we are considering. So that time we can delete that particular column if needed.