

# Covariance



# Covariance

Covariance measures the direction of the relationship between two variables.

A positive covariance means that both variables tend to be high or low at the same time.

A negative covariance means that when one variable is high, the other tends to be low.

A covariance of zero indicates that there is no clear directional relationship between the variables being measured.

Population Covariance Formula

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

**Note :** The covariance can range from negative infinity to positive infinity. Thus, the value for a perfect linear relationship depends on the data.

# Correlation



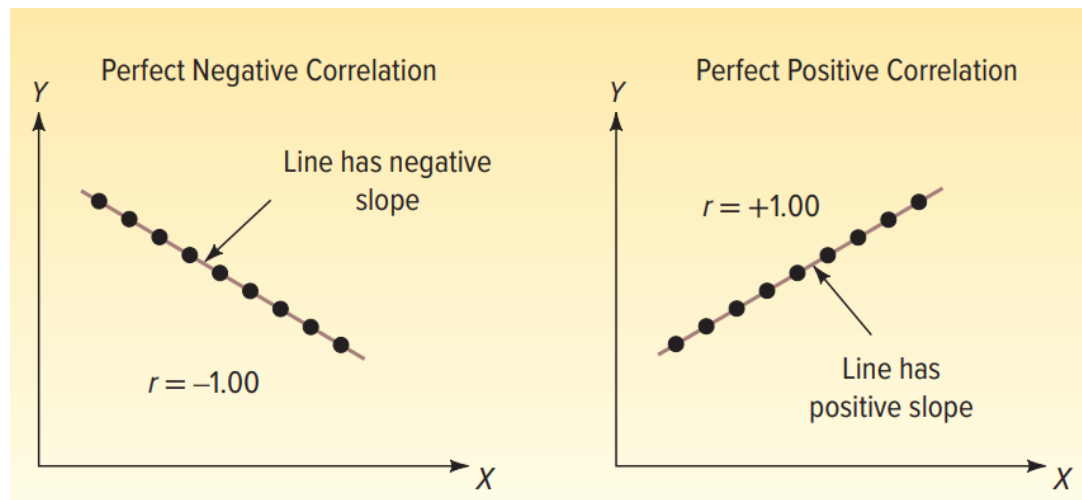
# Pearson's Correlation

Pearson's correlation coefficient can be calculated to measure the direction and strength of the relationship between two variables.

The correlation coefficient is computed as:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

The sample correlation coefficient, specified by  $r$ , ranges from  $-1.0$  to  $+1.0$ .



| Size of Correlation         | Interpretation                            |
|-----------------------------|---|
| .90 to 1.00 (–.90 to –1.00) | Very high positive (negative) correlation |
| .70 to .90 (–.70 to –.90)   | High positive (negative) correlation      |
| .50 to .70 (–.50 to –.70)   | Moderate positive (negative) correlation  |
| .30 to .50 (–.30 to –.50)   | Low positive (negative) correlation       |
| .00 to .30 (.00 to –.30)    | negligible correlation                    |

## Spearman's Rank Correlation

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables.

The formula for Spearman's rank coefficient is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

$\rho$  = Spearman's rank correlation coefficient

$d_i$  = Difference between the two ranks of each observation

$n$  = Number of observations

The Spearman Rank Correlation can take a value from +1 to -1 where,

- A value of +1 means a perfect association of rank
- A value of 0 means that there is no association between ranks
- A value of -1 means a perfect negative association of rank

## Example

Example : Pearson's correlation Coefficient

| ID       | Weight (X) | (X - $\bar{X}$ ) | (X - $\bar{X}$ ) <sup>2</sup> | Height (Y) | (Y - $\bar{Y}$ ) | (Y - $\bar{Y}$ ) <sup>2</sup> | (X - $\bar{X}$ )(Y - $\bar{Y}$ ) |
|----------|------------|------------------|-------------------------------|------------|------------------|-------------------------------|----------------------------------|
| 1        | 148        | -6.10            | 37.21                         | 64         | 1.00             | 1.00                          | -6.10                            |
| 2        | 172        | 17.90            | 320.41                        | 63         | 0.00             | 0.00                          | 0.00                             |
| 3        | 203        | 48.90            | 2391.21                       | 67         | 4.00             | 16.00                         | 195.60                           |
| 4        | 109        | -45.10           | 2034.01                       | 60         | -3.00            | 9.00                          | 135.30                           |
| 5        | 110        | -44.10           | 1944.81                       | 63         | 0.00             | 0.00                          | 0.00                             |
| 6        | 134        | -20.10           | 404.01                        | 62         | -1.00            | 1.00                          | 20.10                            |
| 7        | 195        | 40.90            | 1672.81                       | 59         | -4.00            | 16.00                         | -163.60                          |
| 8        | 147        | -7.10            | 50.41                         | 62         | -1.00            | 1.00                          | 7.10                             |
| 9        | 153        | -1.10            | 1.21                          | 66         | 3.00             | 9.00                          | -3.30                            |
| 10       | 170        | 15.90            | 252.81                        | 64         | 1.00             | 1.00                          | 15.90                            |
| $\Sigma$ | 1541       |                  | 9108.90                       | 630        |                  | 54.00                         | 201.00                           |

$$\bar{X} = 1541/10 = 154.10$$

$$\bar{Y} = 630/10 = 63.00$$

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$$r = \frac{201}{\sqrt{(9108.90)(54)}} \quad r = \frac{201.00}{701.34} = 0.29$$

Example : Spearman's Rank Correlation

| Students | Maths Rank | Science Rank | d  | d square |
|----------|------------|--------------|----|----------|
| A        | 35         | 24           | 11 | 121      |
| B        | 20         | 35           | 15 | 225      |
| C        | 49         | 39           | 10 | 100      |
| D        | 44         | 48           | 4  | 16       |
| E        | 30         | 45           | 15 | 225      |
|          |            |              |    | 14       |

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$= 1 - (6 * 14) / 5(25 - 1)$$

$$= 0.3$$

# Correlation vs Causation

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.

A correlation between variables, however, **does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.**

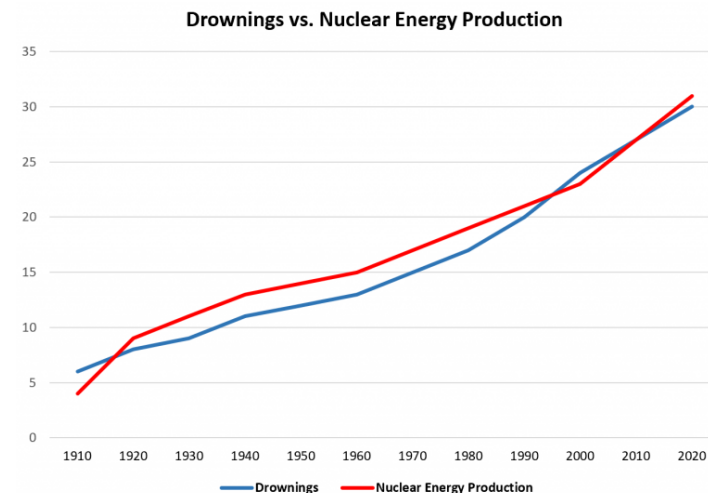
**Causation** indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events.

This is also referred to as cause and effect.

**Example :**

## **Pool Drownings vs. Nuclear Energy Production**

If we collect data for the total number of pool drownings each year and the total amount of energy produced by nuclear power plants each year, we would find that the two variables are highly correlated.



**Note : While causation and correlation can exist at the same time, correlation does not imply causation.**

# **Different Types of plots for Categorical Variables**





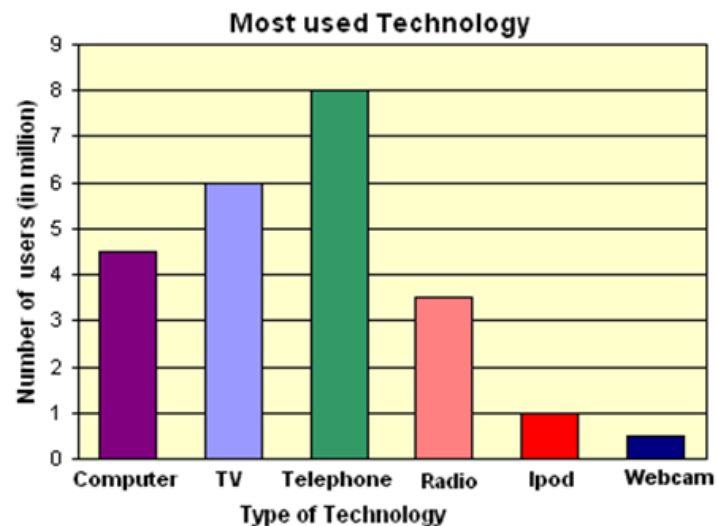
# Plots for Categorical Variables

## Bar Plot

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

The bars can be plotted vertically or horizontally.

Each categorical value claims one bar, and the length of each bar corresponds to the bar's value.



From the above Graph we can see that Telephone has most number of users followed by TV and Computer. Hence Bar plot can be used to compare categorical variables on the basis of counts.

# Pie Chart

- Pie charts show the size of items (called wedge) in one data series, proportional to the sum of the items.
- The data points in a pie chart are shown as a percentage of the whole pie.
- The total angle of  $360^\circ$  at the center of the circle is divided according to the values of the components.

A pie chart is used to compare frequencies of categorical variables.

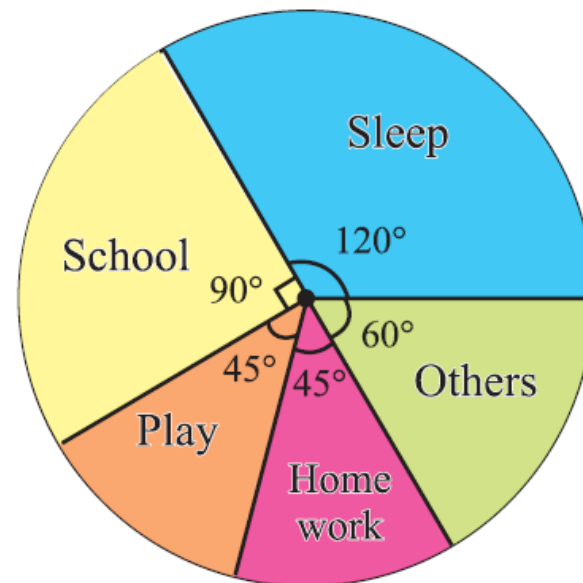
The central angle of a component is

$$= [\text{Value of the component} / \text{Total value}] \cdot 360^\circ$$


Example

| Activity        | Sleep | School | Play | Homework | Others |
|-----------------|-------|--------|------|----------|--------|
| Number of hours | 8     | 6      | 3    | 3        | 4      |

| Activity | Duration in hours | Central angle                               |
|----------|-------------------|---|
| Sleep    | 8                 | $\frac{8}{24} \times 360^\circ = 120^\circ$ |
| School   | 6                 | $\frac{6}{24} \times 360^\circ = 90^\circ$  |
| Play     | 3                 | $\frac{3}{24} \times 360^\circ = 45^\circ$  |
| Homework | 3                 | $\frac{3}{24} \times 360^\circ = 45^\circ$  |
| Others   | 4                 | $\frac{4}{24} \times 360^\circ = 60^\circ$  |
| Total    | 24                | $360^\circ$                                 |



# **Different Types of plots for Continuous Variables**

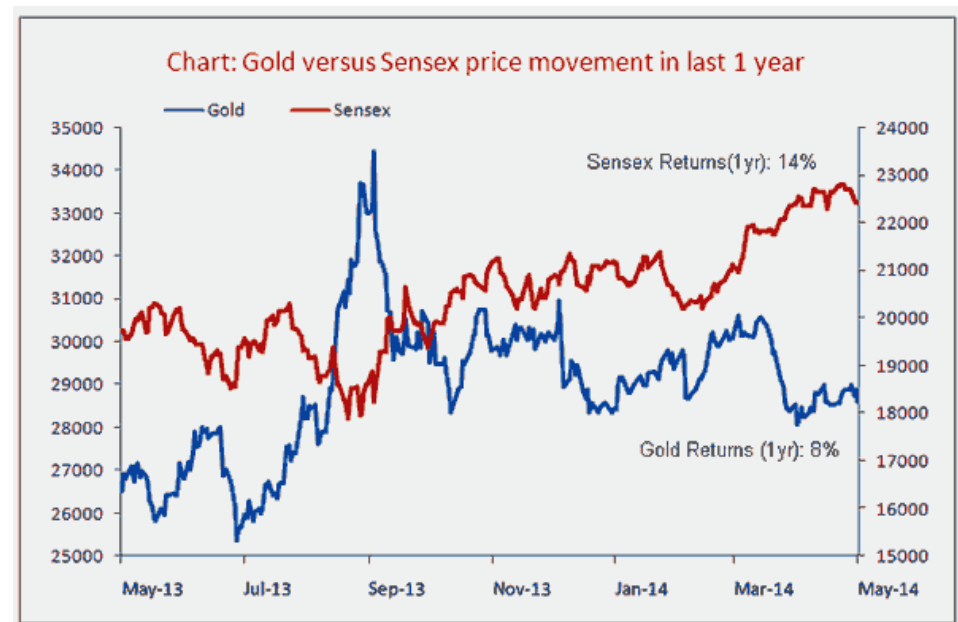
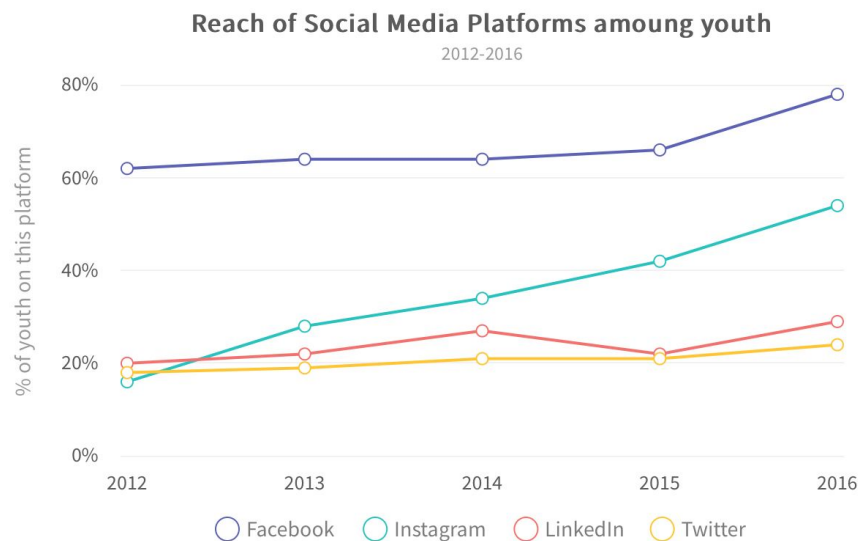


# Plots for Continuous Variables

## Line Plot

A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments.

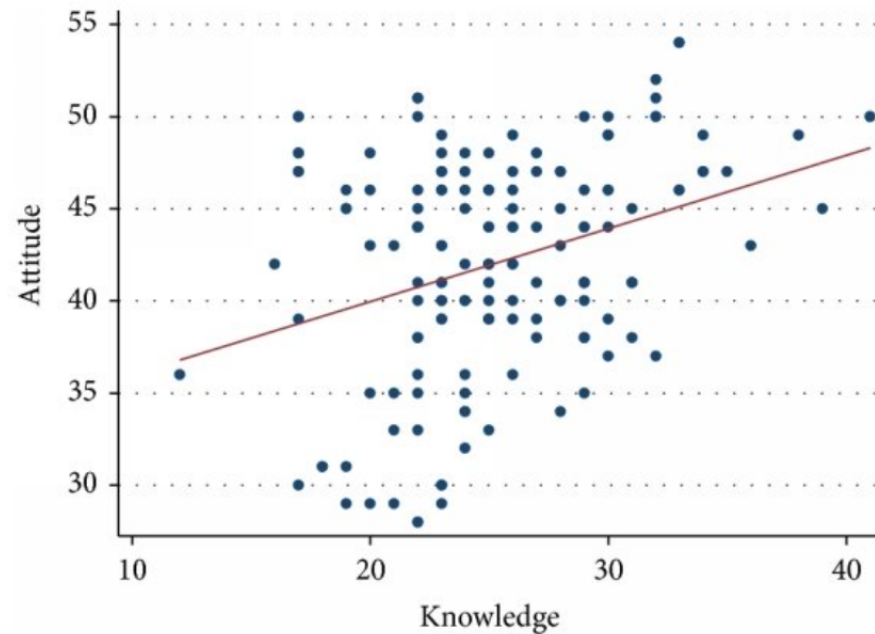
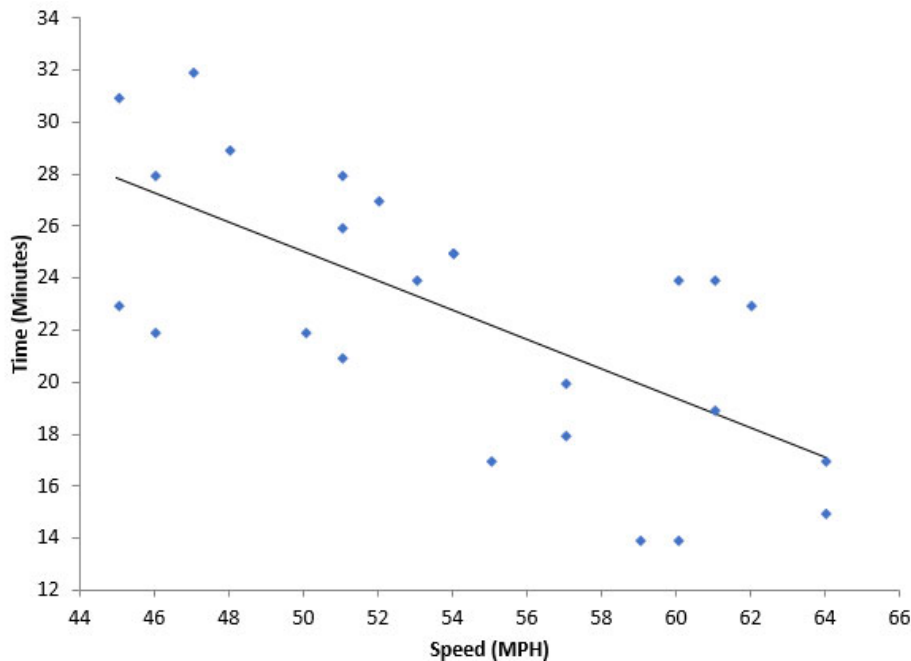
**Line Plots are widely used in time series to identify the pattern of the data.**



# Scatter Plot

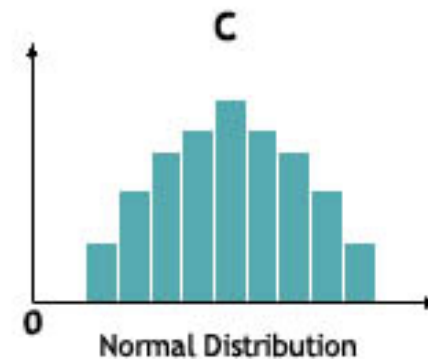
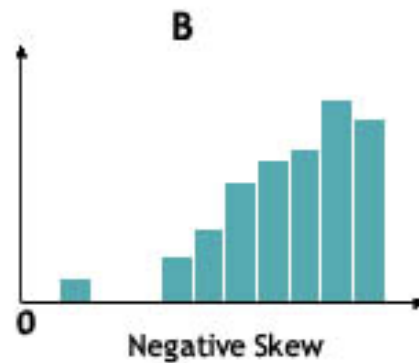
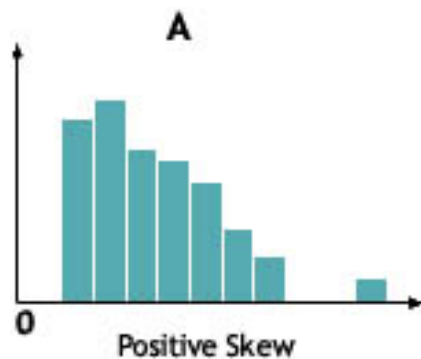
- Scatter plots are the graphs that present the relationship between two variables in a data-set.
- It represents data points on a two-dimensional plane or on a Cartesian system.
- In scatter plot data points on a horizontal and a vertical axis in the attempt to show how much one variable is affected by another.

**It is used in data analysis for checking correlation between two variables.**



# Histogram

- A histogram is an accurate representation of the distribution of numerical data. It is a kind of bar graph.
- Histogram is used when the data is numerical and You want to see the shape of the data's distribution, specially when determining whether the output of a process is distributed approximately normally.
- Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets".

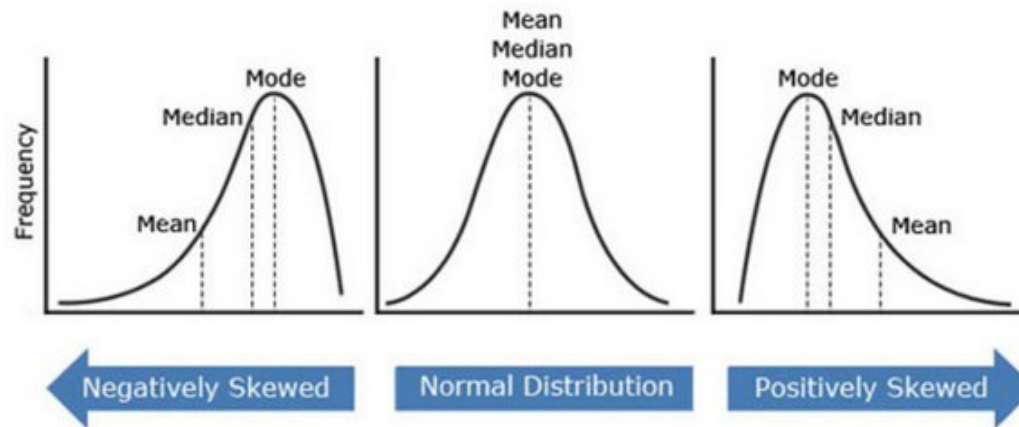


# Density Plot

**A density plot is a representation of the distribution of a numeric variable.**

It uses a kernel density estimate to show the probability density function of the variable (see more).  
It is a smoothed version of the histogram and is used in the same concept.

An advantage Density Plots have over Histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used (each bar used in a typical histogram).



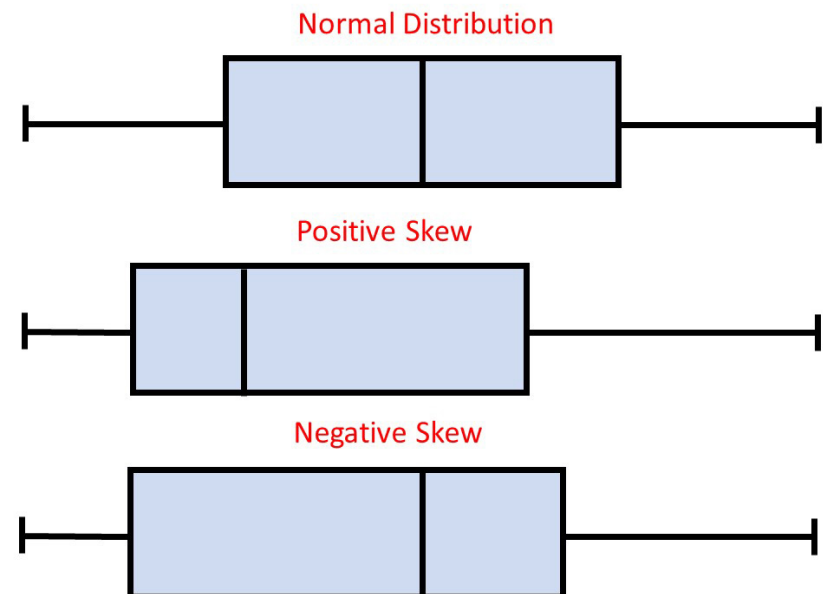
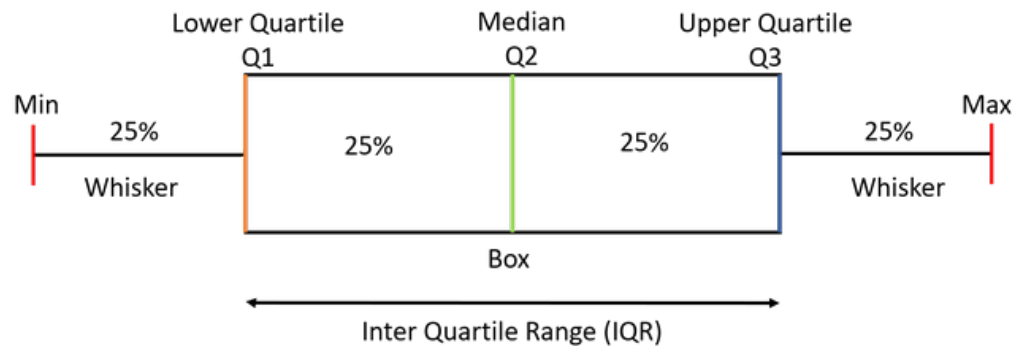
# Box Plot

A box plot which is also known as a whisker plot displays a summary of a set of data containing the

1. minimum
2. first quartile (Q1)
3. Median (Q2)
4. third quartile (Q3)
5. Maximum

Above five are referred to as five number summary in statistics.

Histogram is also used to check the distribution of a numerical variable by partitioning the data into four equal parts.





Thank you!

