# Index

- Types of Data
- Sample & Population
- Types of Sampling
- Descriptive Statistics
- Measure of Central Tendency ( Mean , Median, Mode )
- Measure of Variability ( Range, Variance, Standard Deviation)
- Measure of Position (Quartiles, Percentile, Z-Score)
- IQR, Skewness and Kurtosis

# Index

- Probability Theory (Sampling, Random Sampling, Stratified Sampling, Empirical Probability, Theoretical Probability, Joint Probability, Conditional Probability, Bayes Theorem)
- Distribution (Binomial Distribution and Poisson Distribution)
- Probability Distribution (Random Numbers, CDF, PDF)
- Empirical Rule ( 6 Sigma Rule )
- Central Limit Theorem
- Hypothesis Testing (z-test, t-test, p-value, ANOVA, correlation tests)

# Index

- Over and Under Sampling - SMOTE (Synthetic Minority Over-Sampling Technique)
- Covariance & Correlation
- Data Preprocessing Introduction
- Z Score Outlier Treatment
- IQR Outlier Treatment
- Scaling & Transformations
- Null Value imputation ( convert into bool, count , filling and dropping )

# Index

- Dealing with Categorical Data ( Label encoding & One Hot Encoder )
- Splitting of Data for Training and Testing

# EDA

Exploratory Data Analysis Is an approach to analyzing data.

# Purpose of EDA

- To check null values.
- TO check assumptions associated with any model fitting or hypothesis test.
- Create a list of outliers.
- Find parameter estimators

# Types of EDA

- Univariate Analysis (Graphical)
- Univariate Analysis (Non-Graphical)
- Multivariate Analysis (Graphical)
- Multivariate Analysis (Non-Graphical)

# Statistical Technique of EDA

- The Techniques are divided into Graphical and quantitative analysis.
- Basically EDA work on graphical technique.

# Graphical Technique of EDA

- Visualization by matplotlib or seaborn library.
- Matplotlib contain several plotting method like Histogram, Boxplot...

# STATISTICS

# Introduction To Statistic

Statistics is a branch of mathematics that deals with collecting, interpreting and organisation of data.

# Statistics Provides Methods For

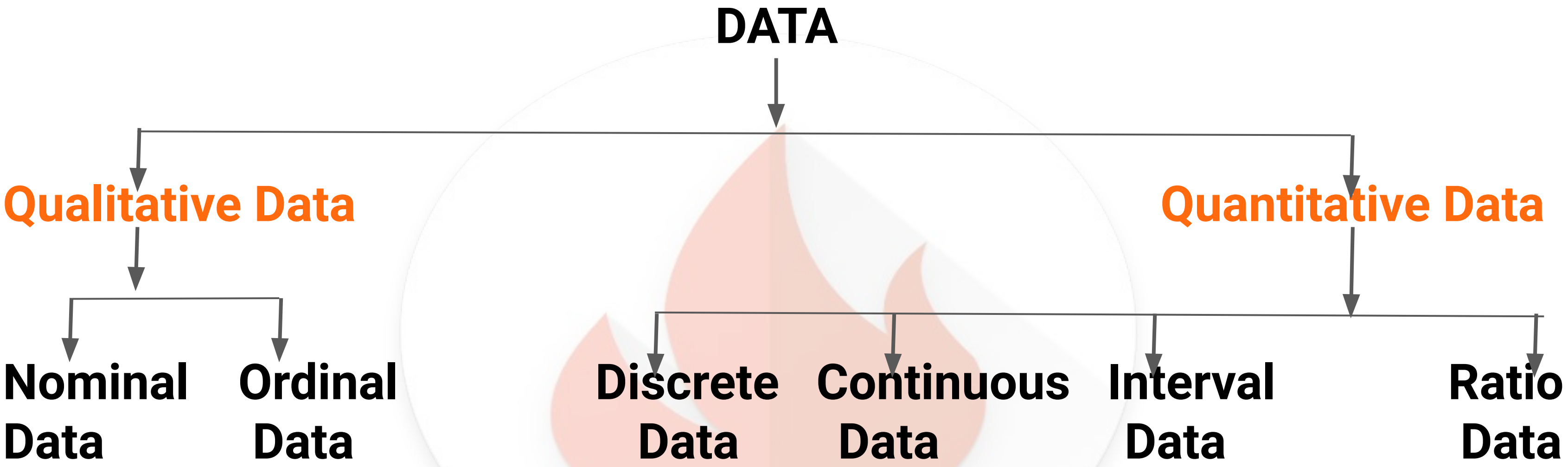
- **Design** :- Planning and carrying out research studies
- **Description** :- Summarising and exploring data.
- **Inference** :- Making predictions and generalising about phenomena represented by the data.

# Practical Problems

- **Agricultural problem:** Is new grain seed good or fertiliser is more productive?
- **Medical problem:** What is the right amount of dosage of drug to treatment?
- **Political science:** How accurate are the opinion polls?
- **Economics:** What will be the unemployment rate next year?
- **Technical problem:** How to improve quality of product?
- **Health problem:** What is the effectiveness of medical treatments?
- **Media problem:** What is the reaction of consumers to television advertising?
- **Social Problem:** What are the attitudes of young people toward sex and marriage?

# TYPES OF DATA

# Types of Data



DATA

Qualitative Data

Quantitative Data

Nominal Data

Ordinal Data

Discrete Data

Continuous Data

Interval Data

Ratio Data

# Types of Data

- **Quantitative Data:-**
- This types of data seem to be the easiest to explain.
  - Example:- How many, How much.
- It can be expressed as a number.
- These are easily open for statistical manipulation and can be represented by a wide variety of statistical types of graph.
- **There are two types**
  - **Discrete Data**
  - **Continuous Data**

# Types of Data

- **Qualitative Data:-**
- Can’t be expressed as a number.
- It mainly consists of words, pictures, symbols.
- Information can be sorted by category.
  - Example:- Why this has happened.
    - Colors
    - Popular Holiday Destination
- **There are two types**
  - **Nominal Data**
  - **Ordinal Data**

# TYPES OF STATISTICS

# Types Of Statistics

- Descriptive Statistics
- Inferential Statistics

# Descriptive Statistics

The branch of statistics devoted to the summarisation and description of data is called descriptive Statistics.

# Descriptive Statistics

1. It consist of methods for organising and summarising information.
2. Descriptive statistics is summarizing the data like mean, median, etc.
3. It includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.

# Descriptive Statistics

Descriptive statistics answer the following questions:

- What is the value that best describes the data set?
- How much a data set spreads from its average value?
- What is the smallest and largest number in a data set?

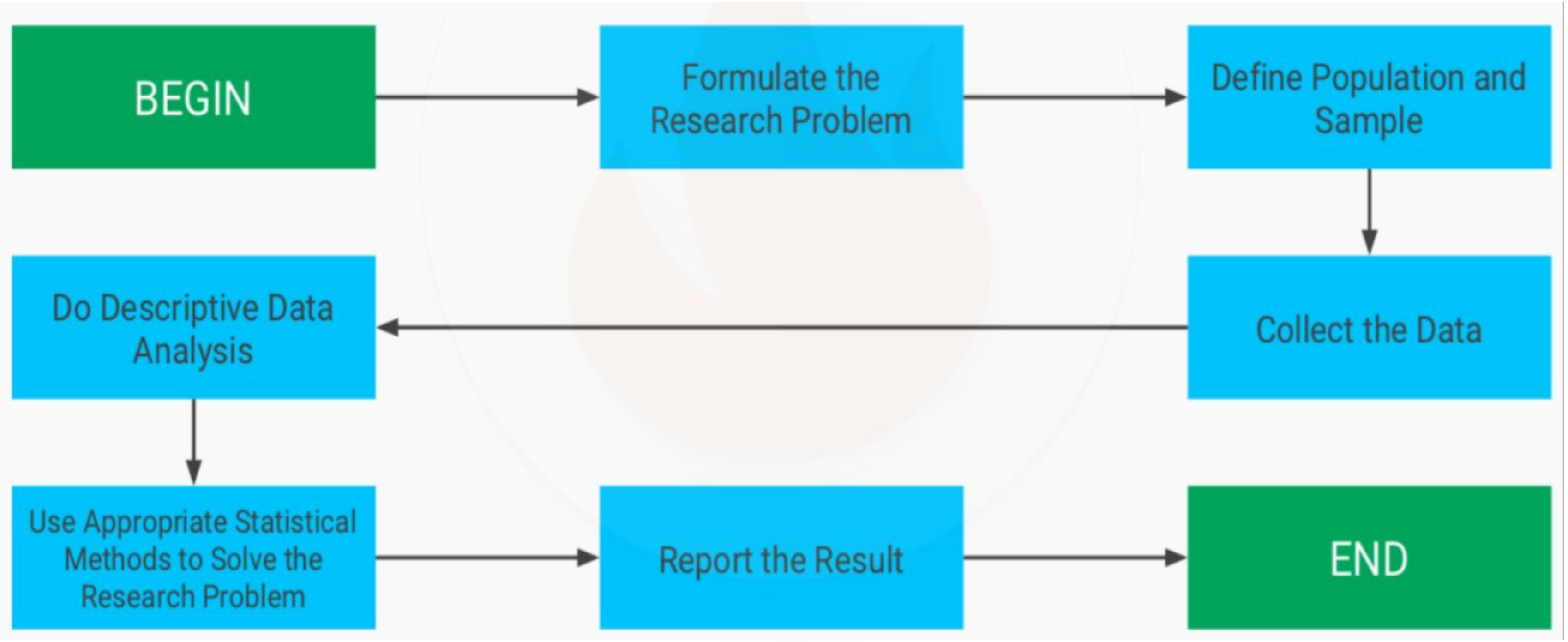
# Inferential Statistics

The branch of statistics concerned with using sample data to make an inference about a population of data is called inferential statistics.

# Inferential Statistics

1. Inferential statistics consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population.
2. It Includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.

# Statistical Data Analytics Process



BEGIN
Formulate the Research Problem
Define Population and Sample
Collect the Data
Do Descriptive Data Analysis
Use Appropriate Statistical Methods to Solve the Research Problem
Report the Result
END

# Population and Sample

# Population and Sample

- **Population** :- Population is the collection of all individuals or items under consideration in a statistical study. Population is all individuals, objects, or measurements whose properties are being studied.

- **Sample** :- Sample is that part of the population from which information is collected. Sample is a subset of the population studied.

# Population and Sample

- ➢ **Population**
- ➢ **Sample**

# Sampling

A sample should have the same characteristics as the population it is representing. There are several different methods of random sampling:

- Stratified Sampling
- Cluster Sampling
- Systematic Sampling
- Convenience Sampling

# Types Of Sampling

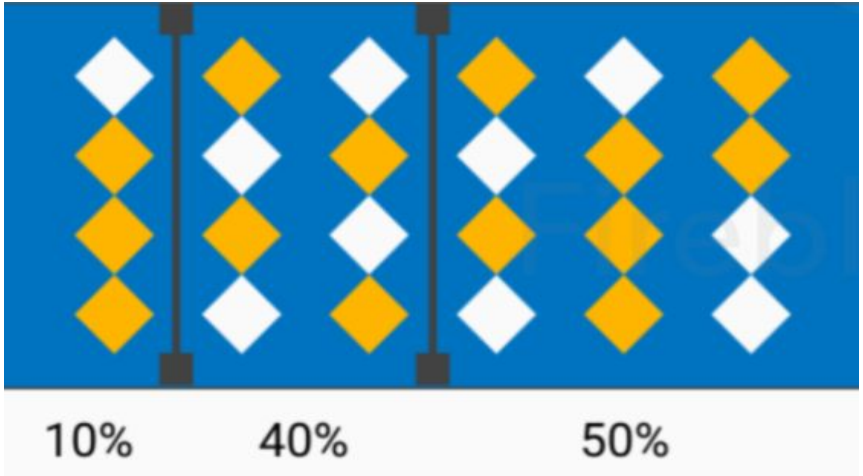
- **Random Sample** :- Pick Randomly From List

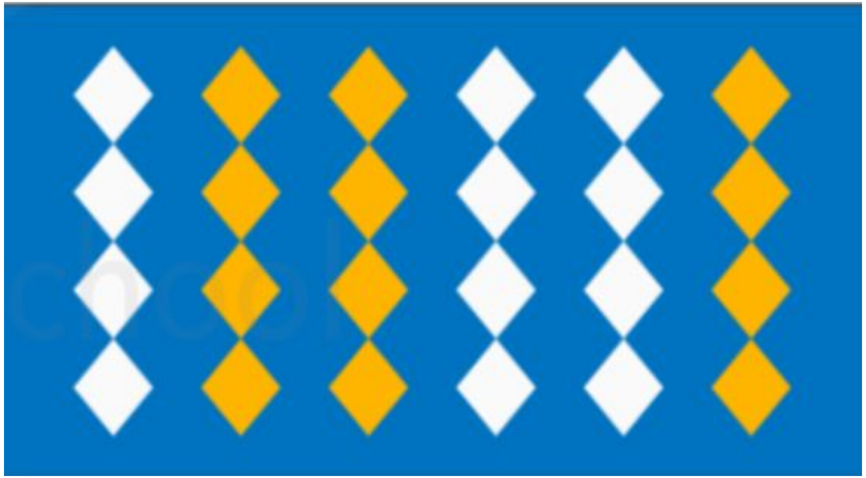
- **Systematic Sample** :- Select Every 3rd Group

# Types Of Sampling

**Stratified Sample** :- Choose Randomly, but in Ratio to Group Size

**Cluster Sample** :- Choose whole Group Randomly





# Sampling With Replacement

Sampling True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once.

Let’s say you had a population of 5 people, and you wanted to sample 2. Their names are: Mizan, Shaad , Rahul, Rucha, Smruti

- Rahul, Shaad
- Rucha, Rucha
- Rucha Smruti
- P(Rucha, Smruti) = (1/5) * (1/5) = 0.04

# Sampling Without Replacement

**Sampling without replacement** is a way to figure out probability without replacement. In other words, you don’t replace the first item you choose before you choose a second. This changes the odds of choosing sample items.

Let’s say you had a population of 5 people, and you wanted to sample 2. Their names are: Mizan, Shaad , Rahul, Rucha, Smruti

- Rahul, Shaad
- Rucha, Smruti
- Mizan, Rahul

P(Miza, Rahul) = (1/5) * (1/4) = .05

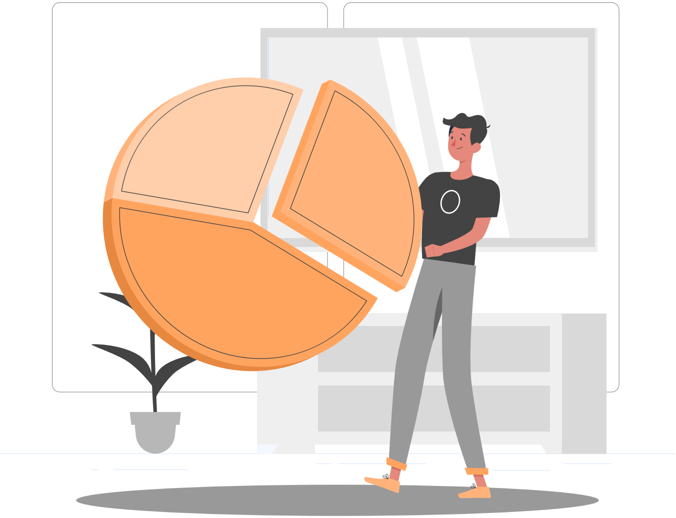
# Sampling

In Sampling **with replacement,** the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second.

**Covariance between the two is zero.**

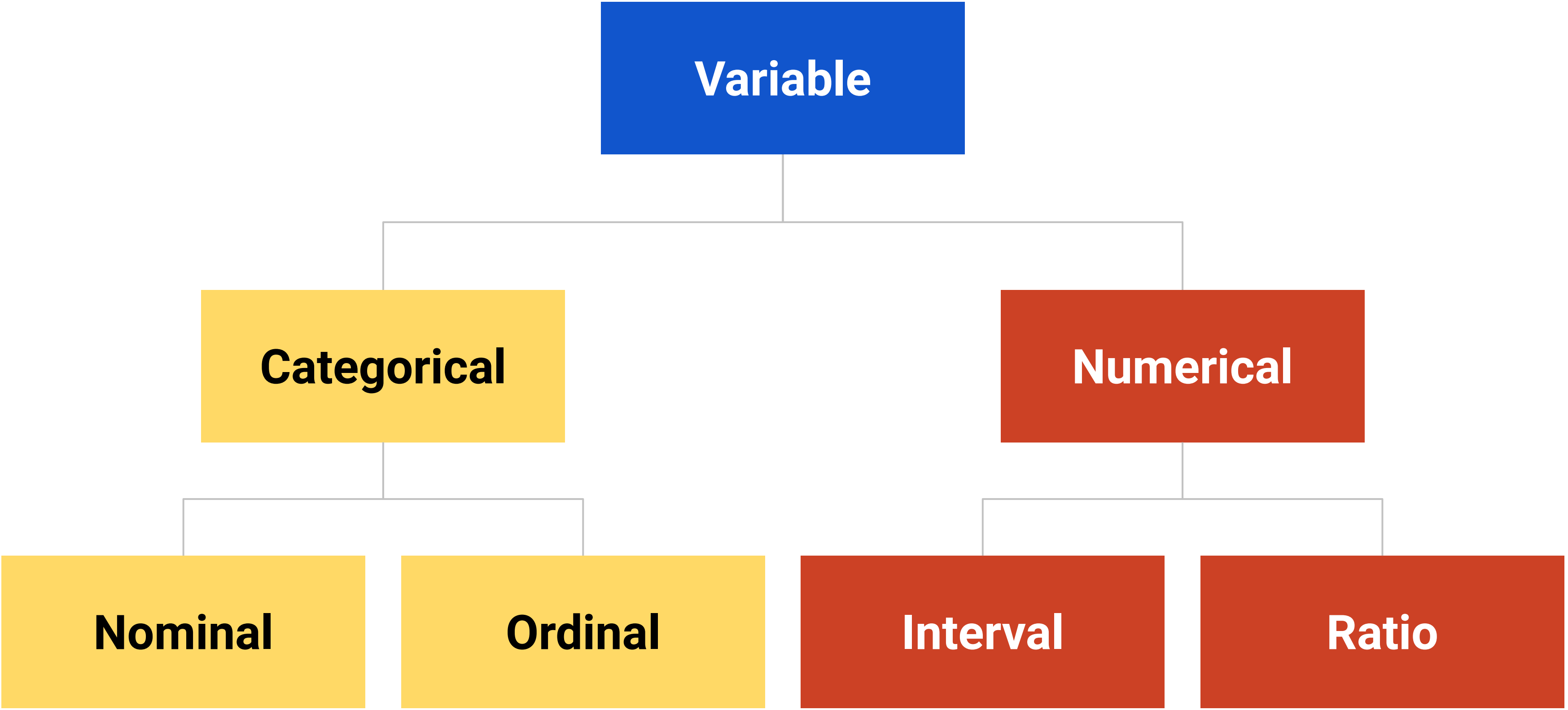
In Sampling **without replacement,** the two sample values aren't independent. Practically, this means that what we got on the for the first one affects what we can get for the second one.

**Covariance between the two isn't zero.**



# Types Of Variable

# Types Of Variable



# Categorical

Categorical Qualitative data are often termed categorical data. Variables that take on values that are names or labels are called as Categorical Variables.

# Nominal Variable (Unordered List)

A variable that has two or more categories, without any implied ordering.

Examples :

- Gender - Male, Female
- Marital Status - Unmarried, Married, Divorcee
- State - New Delhi, Haryana, Illinois, Michigan

# Ordinal Variable (Ordered List)

A variable that has two or more categories, with clear ordering.

Examples :

- Scale - Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree
- Rating - Very low, Low, Medium, Great, Very great

# Numerical

Numerical Quantitative data are often termed numerical data. Variables that take on values that are indicated by numbers are called as Numerical

# Interval

An interval variable is similar to an ordinal variable, except that the intervals between the values of the interval variable are equally spaced. In other words, it has order and equal intervals.

Examples :

Annual Income in Dollars - Three people who make $5,000, $10,000 and $15,000.

# Ratio

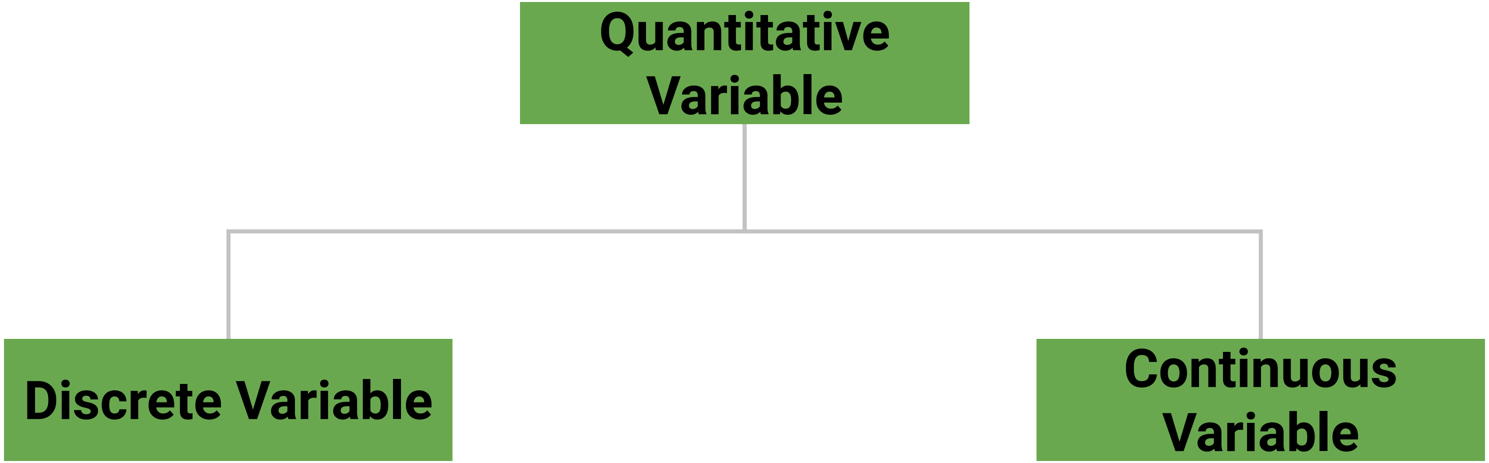
It is quantitative data which can be expressed in P/Q form where Q is NOT equal to 0.

Examples :

- Height
- Weight
- Length

# Quantitative (or Numerical) Variable

Quantitative (or Numerical) Variables Height, Weight, number of siblings of these variables yield numerical information (yield numerical measurements)



# Discrete Variable

All data that are the result of counting are called quantitative discrete data.

Example:

- The number of children in family
- The number of car accident on the certain road on different days
- The number of students taking ML Course.

# Continuous Variable

All data that are the result of measuring are quantitative continuous data assuming that we can measure accurately.

Example:

- Measuring angles in radians
- Measuring weight in kilograms
- Measuring height of the person in centimetres

# Qualitative (or Categorical)

Qualitative (or Categorical) Variables Sex, Marital Status, and Eye Color. yield non-numerical information (yield non-numerical measurements) and are examples of qualitative (or categorical) variables.

# Measure Of Central Tendency

# Measure Of Central Tendency

It describes a whole set of data with a single value that represents the centre of its distribution.

There are three main measures of central tendency: the mode, the median and the mean.

- **Mean : Average Value**
- **Median : Middle Value**
- **Mode : Most Frequent Value**

# MEAN

The mean can suggest a central value and series as a ‘Balance Point’ in a set of data.

The mean is the sum of all the values divided by the number of observations or sample size.

It is nothing but the average value. The mean of the values 5,6,6,8,9,9,9,9,10,10 is (5+6+6+8+9+9+9+9+10+10)/10 = 8.1

**Limitation:** It is affected by extreme values. Very large or very small numbers can distort the answer.

# MEDIAN

- Median is the point which divides the entire data into two equal halves.
- One-half of the data is less than the median & other half is greater than the same (median).
- The median not affected by extreme values.

# MEDIAN

The median is nothing more than the middle value of your observations when they are order from the smallest to the largest. It is the middle value. It splits the data in half. Half of the data are above the median; half of the data are below the median.

7, 8, 7, 6, 9, 8, 8 → 6, 7, 7, 8, 8, 8, 9 → 8 is the Median

7, 8, 7, 6, 9, 8, 8, 7 → 6, 7, 7, 7, 8, 8, 8, 9 → (7 + 8)/2 = 7.5 is the Median

**Advantage** : It is NOT affected by extreme values.

# MEDIAN

## Rules

1. If the dataset contains an **'ODD'** number of values, the median is the measurement associated with the **'middle ranked'** value.
2. If the dataset contains an **'EVEN'** number of values, the median is the measurement associated with the average of **'two ranked'** values.

# MODE

Mode: It is the value that occurs most frequently. In other words, mode is the most common outcome. Mode is the name of the category that occurs more often. There is a chance of having more than one

5, 6, 5, 7, 5, 8, 9, 5 → 5 is the Mode 5, 6, 6, 5, 7, 6, 5, 6, 8, 9, 5, 6 → 5 and 6 are mode.

# MODE

- If there is only **'One'** number that appears maximum number of times, the data has one mode called **'Uni-Modal'**.
- If there is only **'Two'** number that appears maximum number of times, the data has one mode called **'Bi-Modal'**.
- If there is only **'More than Two'** number that appears maximum number of times, the data has one mode called **'Multi-Modal'**.

# MODE

**Mode Advantage** : It can be used when the data is not numerical.

Disadvantage :

1. There may be no mode at all if none of the data is the same .

2. There may be more than one mode.

# WHEN TO USE WHAT MEASUREMENT OF CENTRAL TENDENCY?

# Central Tendency

**Mean** – When your data is not skewed i.e Symmetric/Normally Distributed. In other words, there are no extreme values present in the data set (Outliers).

**Median** – When your data is skewed or you are dealing with ordinal (ordered categories) data.

**Mode** - When dealing with nominal (unordered categories) data.

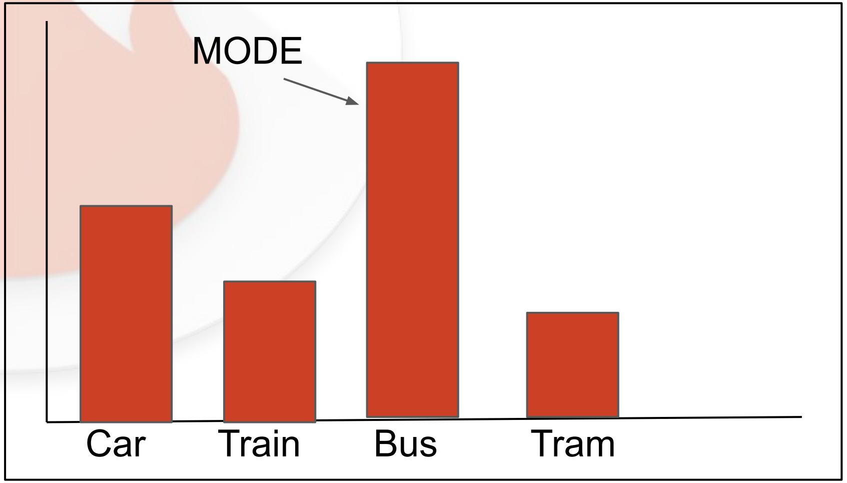
# When To Use Median Instead Of Mean

If your data is quantitative then go for mean or median. Basically, if your data is having some influential outliers or data is highly skewed the median is the best measurement for finding central tendency. Otherwise go for Mean.

Eg : Salary 15k 18k 16k 14k 15k 15k 12k 17k 90k 95k Mean is 30.1K whereas most workers have salaries in the $12k to 18k range. Hence Median is to be preferred.

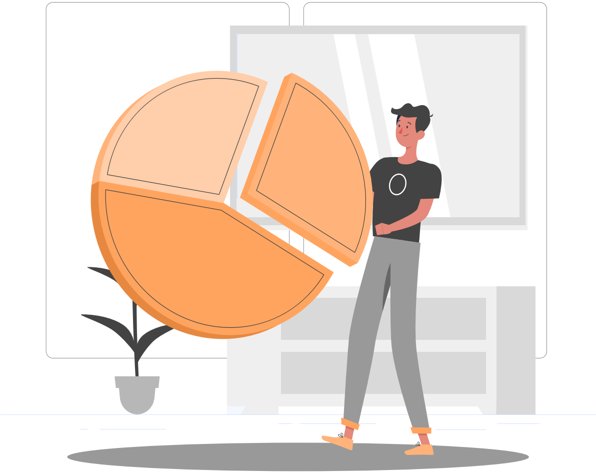
# When To Use Mode

If data is Categorical (Nominal or Ordinal) it is impossible to calculate mean or median. So, go for mode.Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:



# Summary

| Type Of Variable | Best Measure Of Central Tendency |
|---|---|
| Nominal | Mode |
| Ordinal | Median |
| Interval / Ratio(Not Skewed) | Mean |
| Interval / Ratio | Median |



# Measure Of Dispersion

# Mean Deviation

1. Find the mean of all values
2. Find the distance of each value from that mean (subtract the mean from each value, ignore minus signs)
3. Then find the mean of those distances

# Mean Deviation

The Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

**Step 1**:- Find the mean: Mean = (3 + 6 + 6 + 7 + 8 + 11 + 15 +16) / 8

= 72/8

= 9

**Step 2**:- Find the distance of each value from that mean:

# Mean Deviation

| Value | Distance From Mean = 9 |
|---|---|
| 3 | 6 |
| 6 | 3 |
| 6 | 3 |
| 7 | 2 |
| 8 | 1 |
| 11 | 2 |
| 15 | 6 |
| 16 | 7 |

# Mean Deviation

**Step 3** :- Find the mean of those distances: Mean Deviation = (6 + 3 + 3 + 2 + 1 + 2 + 6 + 7)/8 = 30/8 = 3.75

So, the **mean = 9**, and the **mean deviation = 3.75** It tells us how far, on average, all values are from the middle. In that example the values are, on average, 3.75 away from the middle.

**Note**: For deviation just think distance.

# Mean Deviation

**Formula :-**

Mean Deviation Formula: Mean Deviation: $(\Sigma \mid x - \mu \mid) / N$

- $\Sigma$ is Sigma, which means to sum up
- || (the vertical bars) mean Absolute Value, basically to ignore minus signs
- $x$ is each value
- $\mu$ is the mean
- $N$ is the number of values

# Absolute Deviation

Each distance we calculate is called an Absolute Deviation, because it is the Absolute Value of the deviation (how far from the mean).

Note: the mean deviation is sometimes called the **Mean Absolute Deviation (MAD)** because it is the mean of the absolute deviations.

Mean Deviation tells us how far, on average, all values are from the middle.

The deviations on one side of the mean should equal the deviations on the other side.

# Variance and Standard Deviation

- **Variance:** A simple measure of variation around the mean might take the difference between ‘each value’ & the ‘the mean’ & then sum these difference.

# Variance and Standard Deviation

Variance and Standard Deviation consider all the values of a variable to calculate the variability of the data. These are called as Measures of Spread or Dispersion.

- Standard Deviation: The Standard Deviation is a measure of how spread out numbers are.

- Variance: The average of the squared differences from the Mean. i.e the Square of the Standard Deviation.

# Variance and Standard Deviation

There are two types of variance and standard deviation in terms of Sample and Population.

## For Samples:

variance = s2 = Σ(x - μ)2 / n - 1

standard deviation s = √s2

## For Populations:

variance = σ2 = Σ(x - μ)2 / n

standard deviation = √ σ2

- x is individual one value
- n is size of population
- μ is the mean of population or sample

# Steps To Calculate Standard Deviation

1. Work out the Mean (the simple average of the numbers)

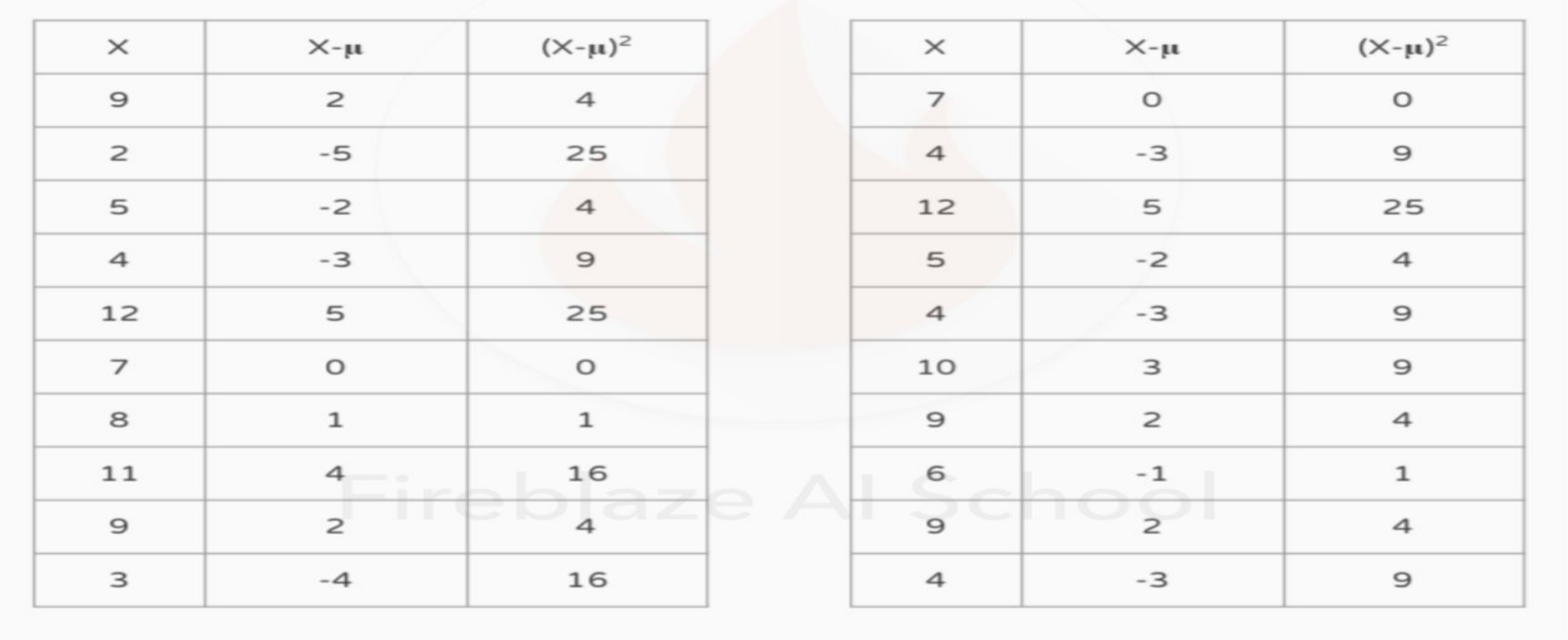
2. Then for each number: subtract the Mean and square the result

3. Then work out the mean of those squared differences.

4. Take the square root of that and we are done!

# Steps To Calculate Standard Deviation

Step 2. Then for each number: subtract the Mean and square the result.



| X | $X-\mu$ | $(X-\mu)^2$ |
|---|---|---|
| 9 | 2 | 4 |
| 2 | -5 | 25 |
| 5 | -2 | 4 |
| 4 | -3 | 9 |
| 12 | 5 | 25 |
| 7 | 0 | 0 |
| 8 | 1 | 1 |
| 11 | 4 | 16 |
| 9 | 2 | 4 |
| 3 | -4 | 16 |

| X | $X-\mu$ | $(X-\mu)^2$ |
|---|---|---|
| 7 | 0 | 0 |
| 4 | -3 | 9 |
| 12 | 5 | 25 |
| 5 | -2 | 4 |
| 4 | -3 | 9 |
| 10 | 3 | 9 |
| 9 | 2 | 4 |
| 6 | -1 | 1 |
| 9 | 2 | 4 |
| 4 | -3 | 9 |

# Steps To Calculate Standard Deviation

Step 3. Then work out the mean of those squared differences. To work out the mean, add up all the values then divide by how many. =Σ4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9 = 178

Here, n = 20

- Sample Variance ( s2 ) = 178/19 = 9.3
- Population Variance ( σ2 ) = 178/20 = 8.9
- Sample Standard Deviation ( s ) = √s2= √9.3 = 3
- Population Standard Deviation ( σ ) = √σ2 = √8.9 = 2.9

# Intuition

1. If variance is high, that means you have larger variability in your dataset. In the other way, we can say more values are spread out around your mean value.
2. Standard deviation represents the average distance of an observation from the mean.
3. The larger the standard deviation, larger the variability of the data.
4. A low standard deviation indicates that the data points tend to be close to the mean

# Standard Error

The Standard Error( SD ) is the standard deviation of the sampling distribution of a statistical mean

It is used to refer to an estimate of standard deviation, derived from a particular sample to compute the estimate

Sampling distribution is the probability distribution of a given statistic based on random sample

# Standard Error

Formula :-

$$SE = \sigma / \sqrt{n}$$

$\sigma$ is Standard Deviation

$n$ is Number Of Sample

# Frequency, Relative Frequency and CRF

- **Frequency** : A frequency is the number of times a value of the data occurs.Example: 20 students were asked how many hours they study per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3
- **Relative Frequency** : A relative frequency is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes.
- **Cumulative Relative Frequency** : Cumulative relative frequency is the accumulation of the previous relative frequencies.

# Frequency, Relative Frequency and CRF

| Data Value | Frequency | Relative Frequency | Cumulative Relative Frequency |
|---|---|---|---|
| 2 | 3 | 3 / 20 = 0.15 | 0.15 |
| 3 | 5 | 5 / 20 = 0.25 | 0.15 + 0.25 = 0.40 |
| 4 | 3 | 3 / 20 = 0.15 | 0.40 + 0.15 = 0.55 |
| 5 | 6 | 6 / 20 = 0.30 | 0.55 + 0.30 = 0.85 |
| 6 | 2 | 2 / 20 = 0.10 | 0.85 + 0.10 = 0.95 |
| 7 | 1 | 1 / 20 = 0.05 | 0.95 + 0.05 = 1 |

# Range

It is used to get good indication of how the values in a distribution are spread out i.e Measures of Variability.

- **Range**: Difference between Maximum and Minimum value in a distribution. Team1 → 2, 4, 10, 15, 24, 25, 40 → Range → 40 - 2 = 38 As ranges takes only the count of extreme values sometimes it may not give you a good impact on variability.
- **Disadvantage**: It is very sensitive to outliers and does not use all the observations in a data set.

# Measure of Position

# Percentile

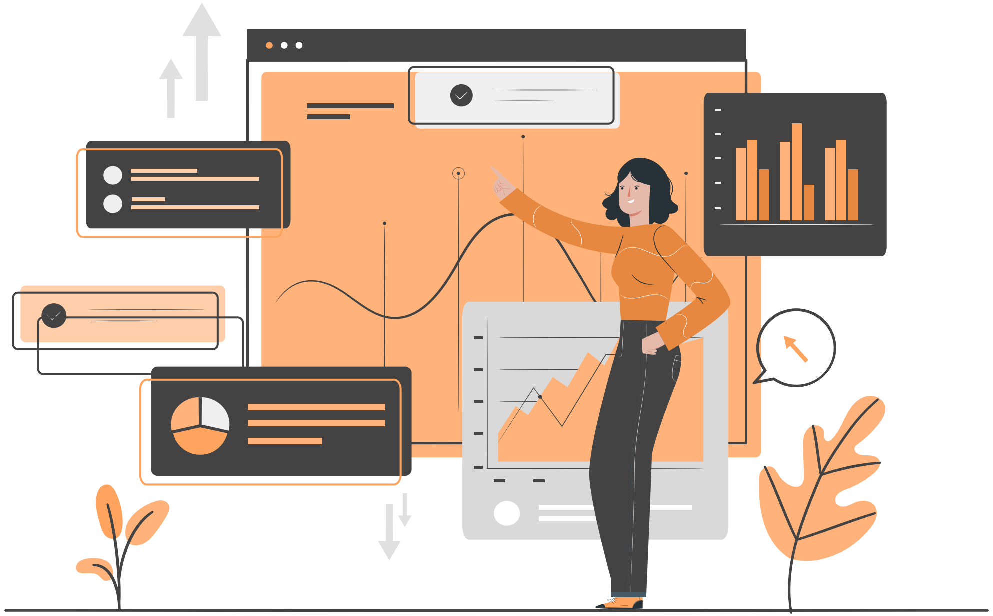
1. Related to quartiles are percentiles that split a variable into **100** equal parts.

# Interquartile Range

1. Quartiles split the values into four equal parts.
2. The **'First Quartile (Q1)'** divide the smallest 25% of the values from other 75%.
3. The **'Second Quartile (Q2)'** is the median. 50% of the values are smaller than or equal and 50% are larger than or equal to the median.
4. The **'Third Quartile (Q3)'** divide the smallest 75% of the values from larger 25%.

# Interquartile Range

1. Also called as **‘midspread’**.
2. Measure the difference in the center of a distribution between the third and first quartiles.
    a. $IQR = Q_3 - Q_1$



# Skewness & Kurtosis

# Skewness

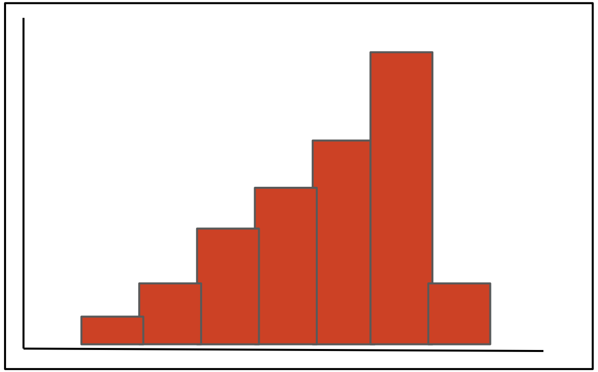
1. Skewness measure the extent to which the data values are not **'symmetrical'** around the mean.
2. There are three possibilities about skewness.
   a. mean < median (Negative Skewness / Left skewed)
   b. mean = median (symmetrical distribution)
   c. mean > median (Positive skewness/ Right skewed)

# Skewness

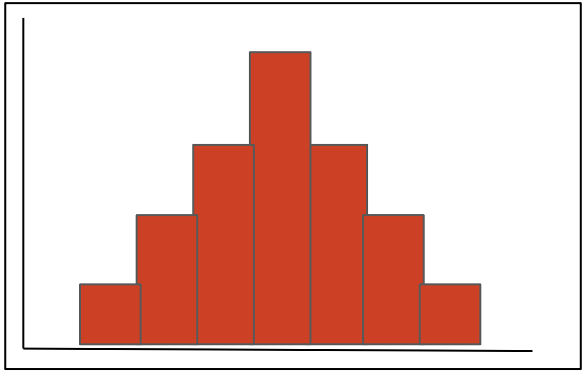
1. In a symmetrical distribution the values below the mean are distributed in exactly the same way as the values above the mean and skewness is **‘zero’**.
2. In a skewed distribution there is an imbalance of data values **‘below & above’** the mean and skewness is a **‘non-zero’**.
3. Less than **‘zero’** for a **left-skewed** distribution.
4. Greater than **‘zero’** for a **right-skewed** distribution.

# Skewness

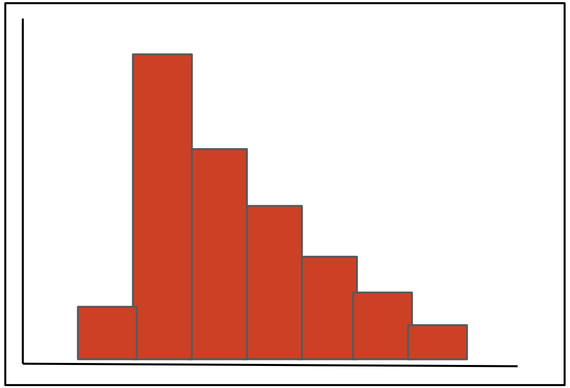
Left Skewed



Symmetric



Right Skewed



1. Middle one has the shape of a bell curve, has one peak, and is approximately symmetric.
2. Left one is left skewed and unimodal
3. Right one is right skewed and unimodal

# Skewness

## Left Skewed

1. In a left-skewed distribution most of the value are in the **upper portion** of the distribution.
2. Some extremely small values cause the long tail and distortion to the **'left'**.
3. **Mean** to be less than **Median**.

# Skewness

## Right Skewed

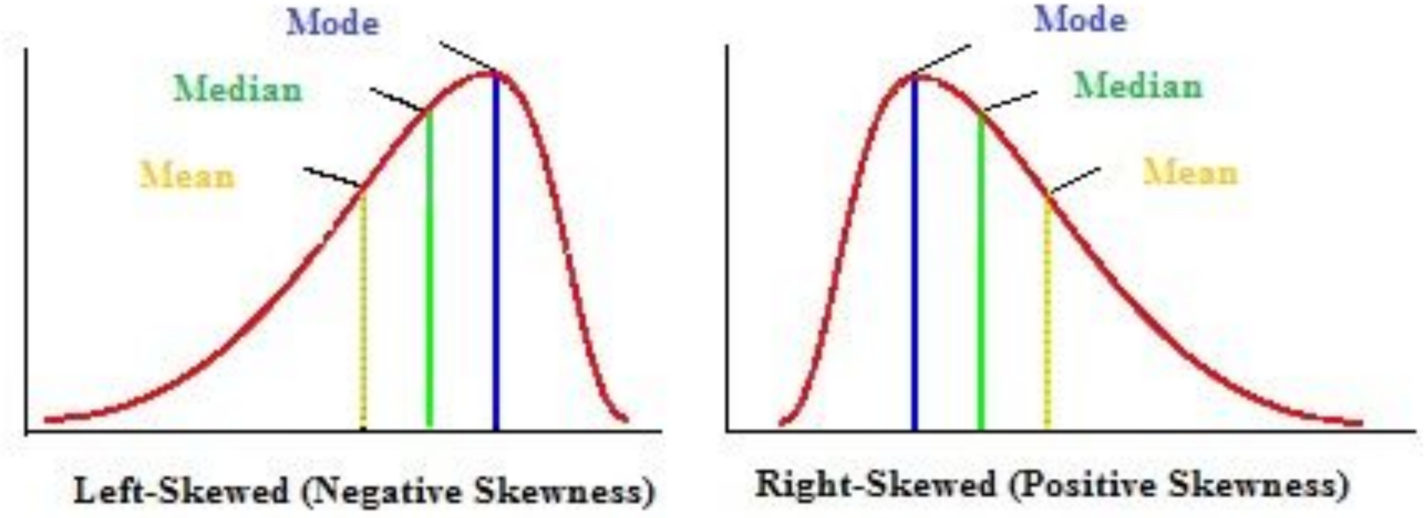
1. Most of the data values are in the lower portion of the distribution.

2. **Mean** to be more than **Median**.

## Symmetric

1. Values are equally distributed in the upper and lower portions.
2. This equality causes the portion of the curve below the mean to be a mirror image of the portion of the curve above the mean and makes the mean equal to the median.

# Skewness


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

# Kurtosis

1. In statistics, kurtosis is defined as the parameter of relative sharpness of the peak of the probability distribution curve.
2. Compares the shape of the peak to the shape of the peak of a **bell-shaped** normal distribution.
3. It is used to indicate the **Flatness or Peakedness** of the frequency distribution curve and measures the tails or outliers of the distribution.

# Types Of Kurtosis

**Mesokurtic :-** Mesokurtic is the distribution which has similar kurtosis as normal distribution kurtosis, which is zero.

**Leptokurtic :-** The distribution which has kurtosis greater than a Mesokurtic distribution. Tails of such distributions are thick and heavy.

- A distribution that has a shaper-rising center peak than the peak of normal distribution has **'+ve' kurtosis.**
- A value that is greater than 'zero' called as 'Lepokurtic'.

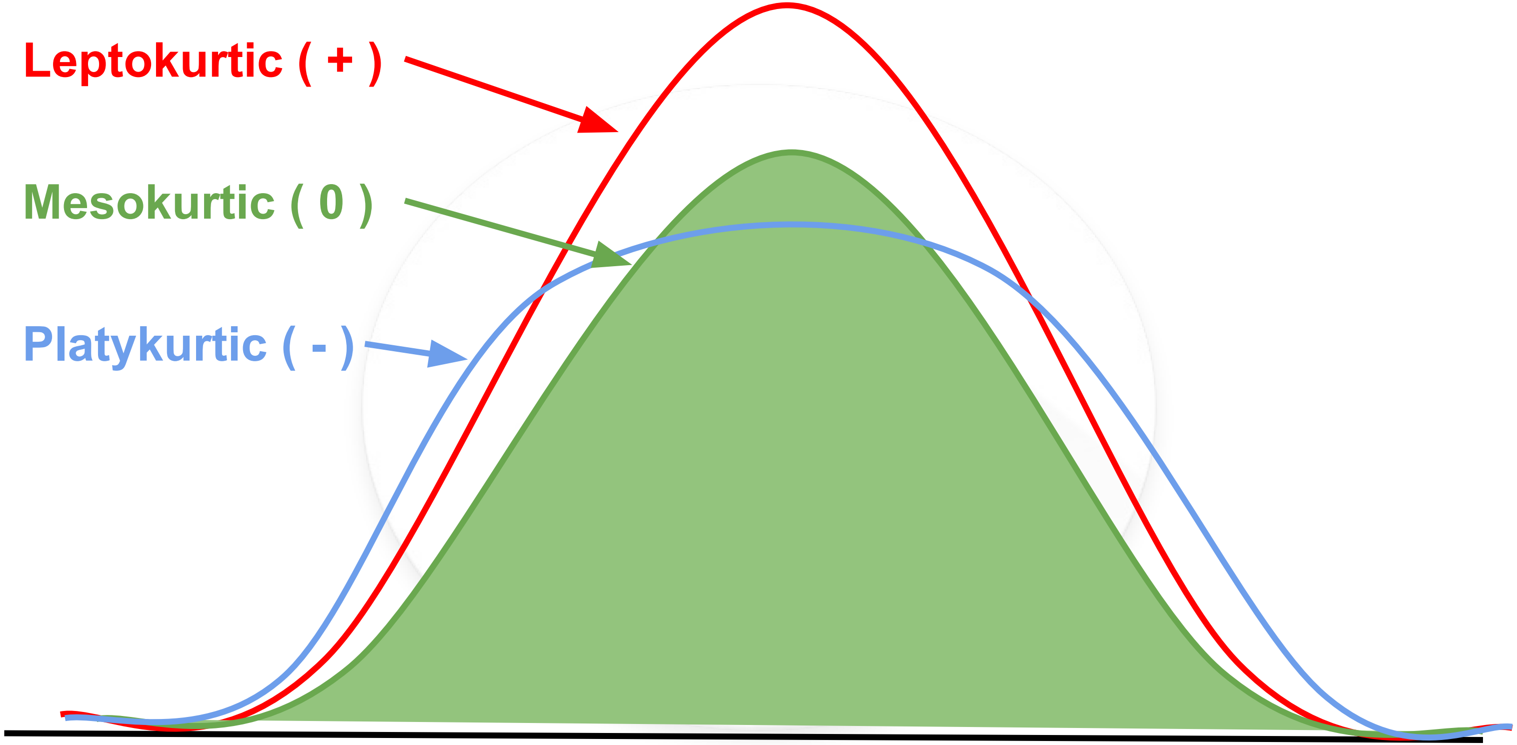
# Types Of Kurtosis

- A **‘Lepokurtic’** distribution has a higher concentration of values near to mean of distribution compared to normal distribution.

**Platykurtic** :- The distribution which has kurtosis lesser than a Mesokurtic distribution. Tails of such distributions thinner.

- A distribution that has a **‘slower-rising’** center peak than the peak of normal distribution has **‘-ve’** kurtosis.
- A values is less than zero called **‘Platykurtic’**.

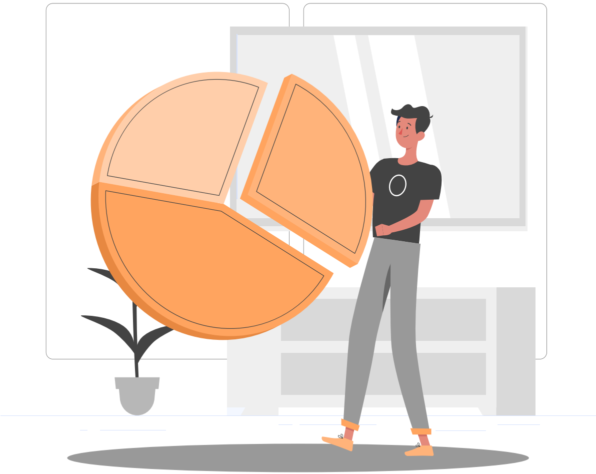
# Kurtosis



Leptokurtic ( + )

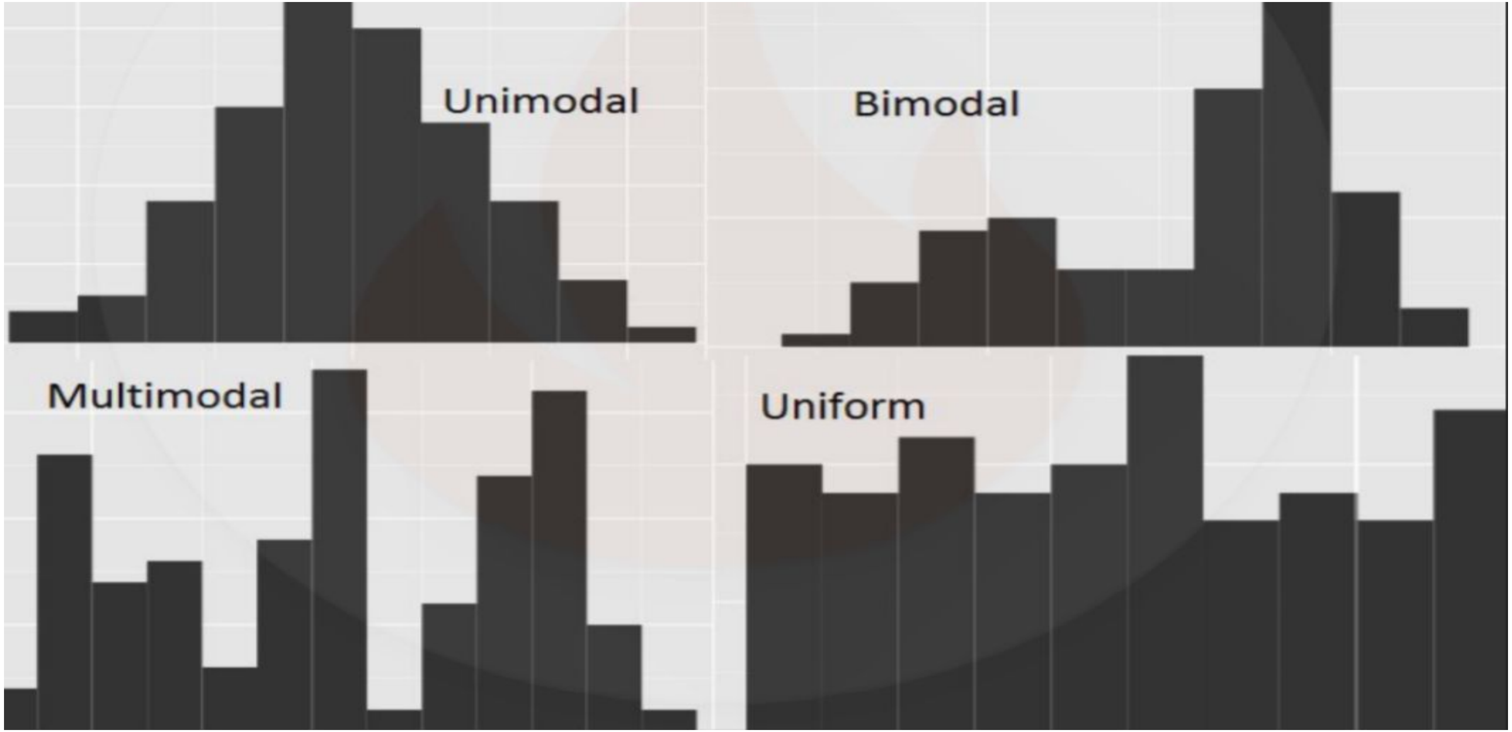
Mesokurtic ( 0 )

Platykurtic ( - )



# Modalities

# Four Kind Of Modalities



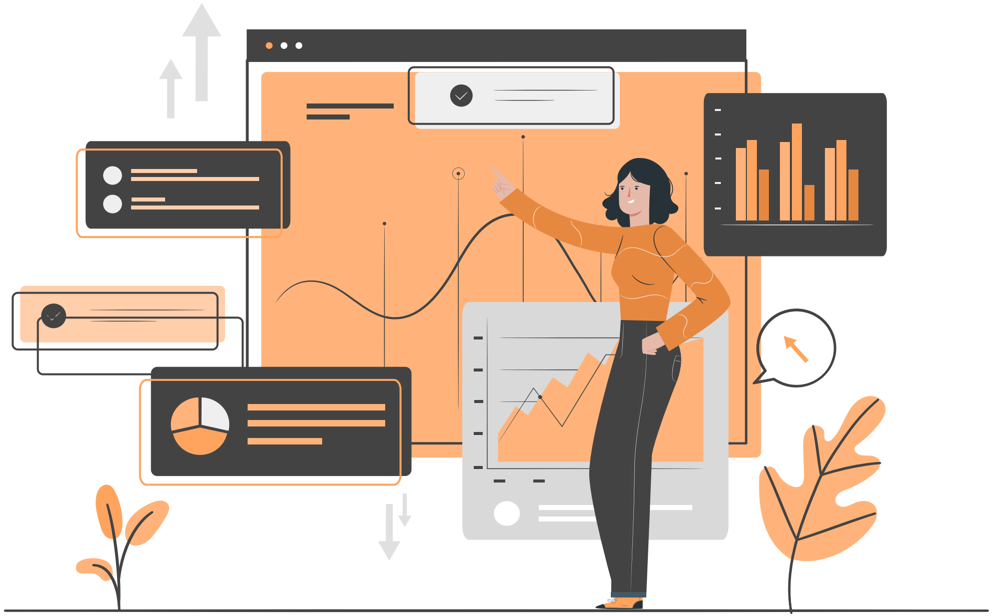
# Four Kind Of Modalities

**Unimodal** :- It has only one peak

**Bimodal** :- It has two peak

**Multimodal** :- It has many peak

**Uniform** :- All are distributed uniformly



# Interquartile Range (IQR)

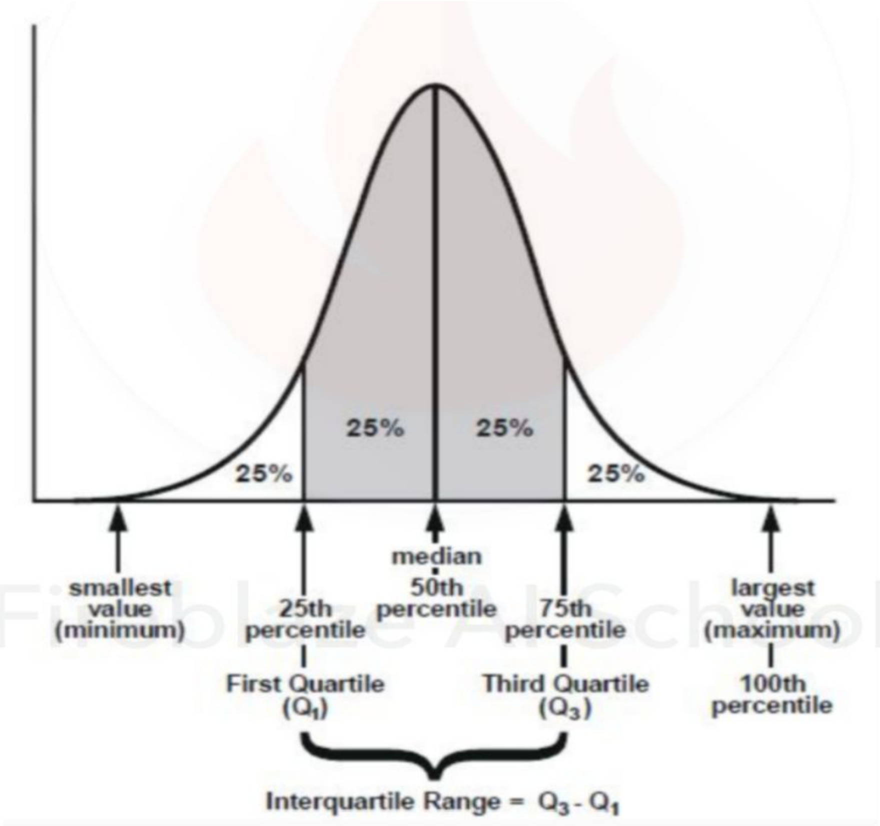
# Interquartile Range (IQR)

It equally divides the distribution into four equal parts called quartiles. First 25% is 1st quartile (Q1), last one is 3rd quartile (Q3) and middle one is 2nd quartile (Q2) and it leaves out the extreme values.

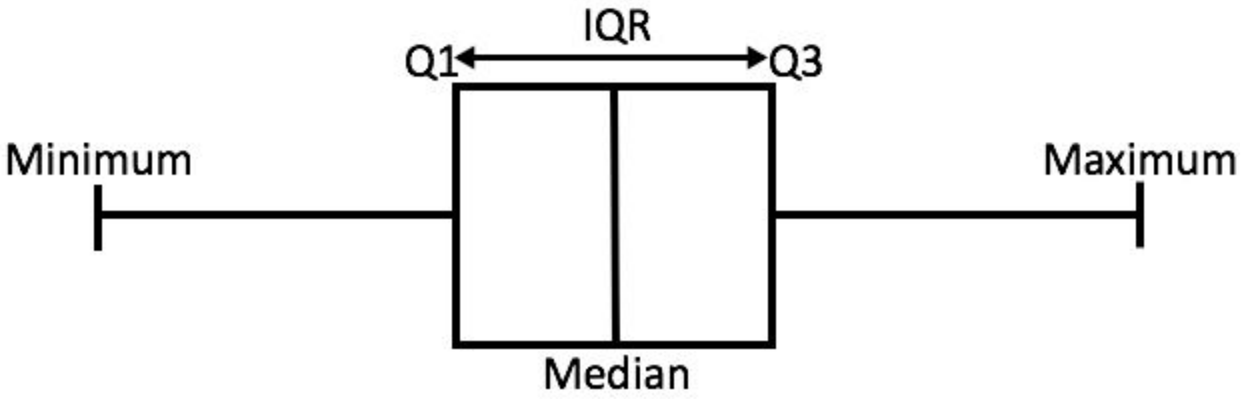
2nd quartile (Q2) divides the distribution into two equal parts of 50%. So, basically it is same as Median.

The interquartile range is the distance between the third and the first quartile, or, in other words, IQR equals Q3 minus Q1. **IQR = Q3- Q1**

# Interquartile Range (IQR)



# Interquartile Range (IQR)



# How To Calculate IQR

**Step 1 :-** Arrange data in ascending order from low to high.

**Step 2 :-** Find the median or in other words Q2.

**Step 3 :-** Then find Q1 by looking the median of the left side of Q2.

**Steps 4 :-** Similarly find Q3 by looking the median of the right of Q2.

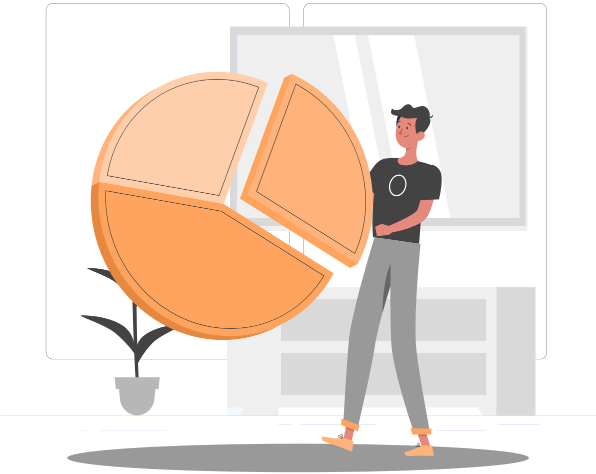
**Steps 5 :-** Now subtract Q1 from Q3 to get IQR.

# How To Calculate IQR

0, 2.3, 4.5, 10, 11, 12, 14.6, 17, 17, 19.7, 20, 23, 25

**14.6 is the Middle Value or Median or Q2**

Consider: 0, 2.3, 4.5, 10, 11, 12 → 4.5 + 10 = 14.5/2 = **7.25** → **Q1 Value** Consider: 17, 17, 19.7, 20, 23, 25 → 19.7 + 20 = 39.7/2 = **19.85** → **Q3** Value → Q3 Value IQR = Q3 - Q1 = 19.85 - 7.25 = **12.60** → **IQR Value**



# Z Score

# Z Score or Standard Score

Z-score is the number of standard deviations from the mean a data point is.

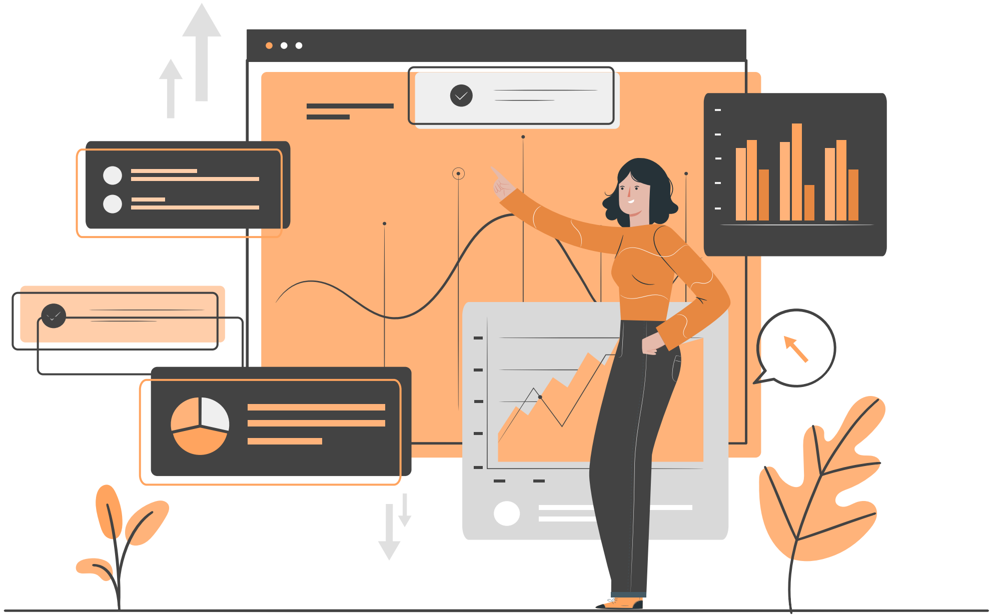
**Z Score = (x - μ) / σ**

- x : Value of the element
- μ : Population mean
- σ : Standard Deviation

A z-score of zero tells you the values is exactly average while a score of +3 tells you that the value is much higher than average.

# Z Score or Standard Score

- If a z-score is a +ve or -ve number, it indicates whether the value is above or below the mean and by how many standard deviations.
- Z-score help identify **'outliers'**.
- As general rule, a Z-score that is less than **-3.0** or greater than **+3.0** indicates as outlier value.



# Contingency Table

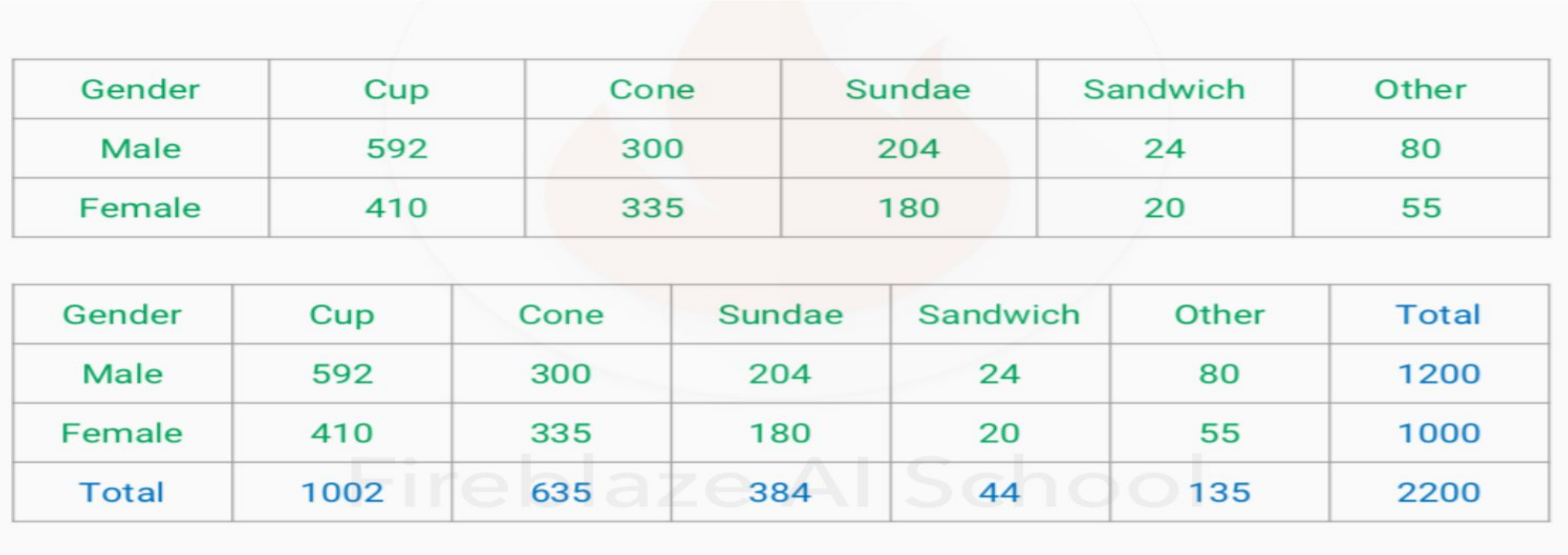
# Contingency Table

It is very similar to a frequency table. But the major difference is that a frequency table always concerns only one variable, whereas a contingency table concerns two variables.

To know the relationship between two ordinal or nominal variables then look for contingency table which displays this relationship.

# Contingency Table

Consider this sample, which shows gender and favourite way to eat ice cream.



| Gender | Cup | Cone | Sundae | Sandwich | Other |
|---|---|---|---|---|---|
| Male | 592 | 300 | 204 | 24 | 80 |
| Female | 410 | 335 | 180 | 20 | 55 |

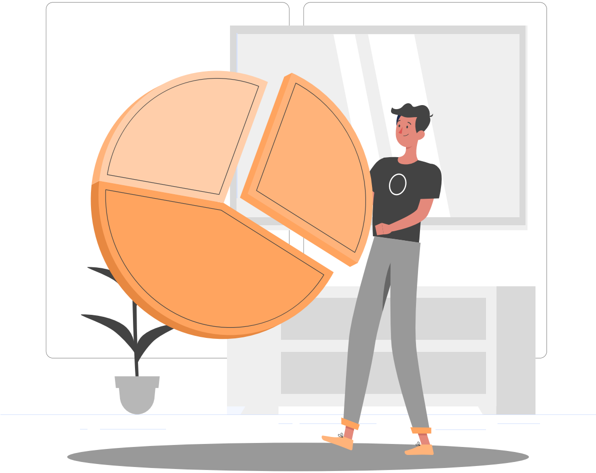
| Gender | Cup | Cone | Sundae | Sandwich | Other | Total |
|---|---|---|---|---|---|---|
| Male | 592 | 300 | 204 | 24 | 80 | 1200 |
| Female | 410 | 335 | 180 | 20 | 55 | 1000 |
| Total | 1002 | 635 | 384 | 44 | 135 | 2200 |

# Analysis

There is 1002/2200 = 45.54% probability that the person prefers his ice cream in a cup.

There is 24/1200 = 2% probability that a person prefers ice cream sandwich given that person is male.

These things are called Conditional proportion and Marginal proportion.



# Correlation Coefficient

# Correlation Coefficient

A contingency table is useful for nominal and ordinal variables, but not for quantitative variables. For quantitative variables, a scatterplot is more appropriate.

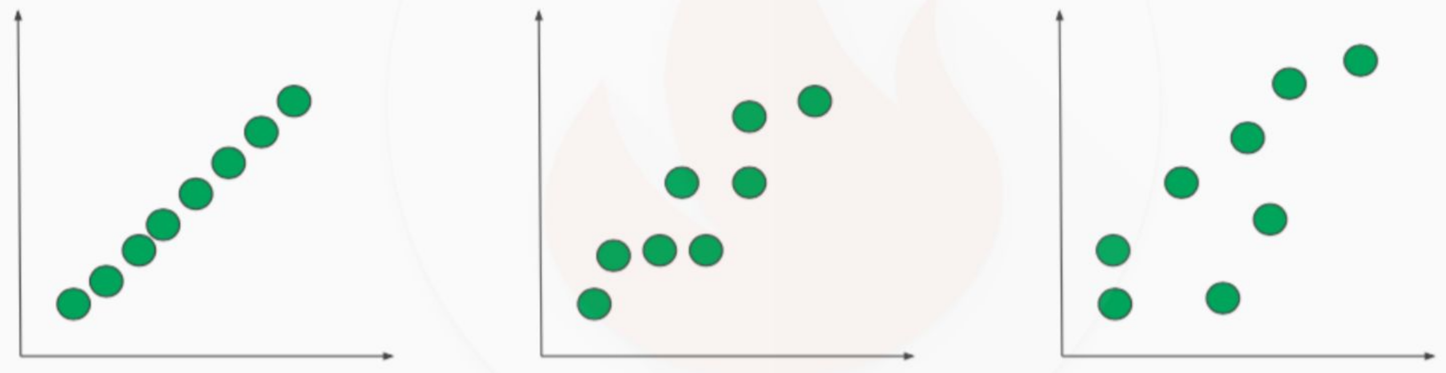
Scatter plot displays relation between two quantitative variables explanatory variable will be in X axis and Response variable will be in y axis.

**Pearson's r or Pearson Correlation:** When two sets of data are strongly linked together, they have a High Correlation.

# Correlation Coefficient

- The word Correlation is made of Co- (meaning "together"), and Relation
- Correlation is Positive when the values increase together, and
- Correlation is Negative when one value decreases as the other increases

# Scatter Plot

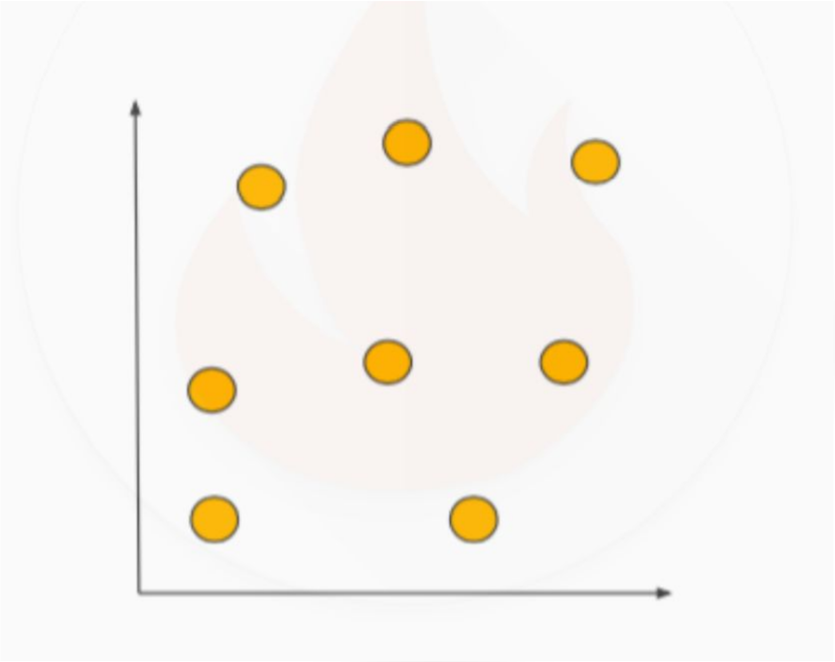


Perfect Positive Correlation Value = 1

High Positive Correlation Value = 0.9

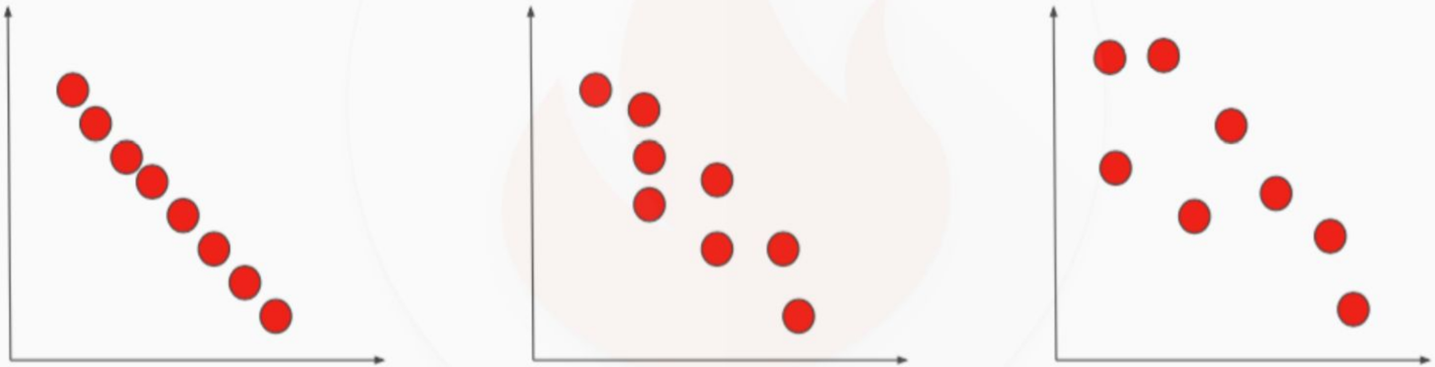
Low Positive Correlation Value = 0.5

# Scatter Plot



No Correlation Value = 0

# Scatter Plot



Perfect Negative Correlation
Value = -1

High Negative Correlation
Value = -0.9

Low Negative Correlation
Value = -0.5

# Correlation Coefficient

Correlation Coefficient can have a value:

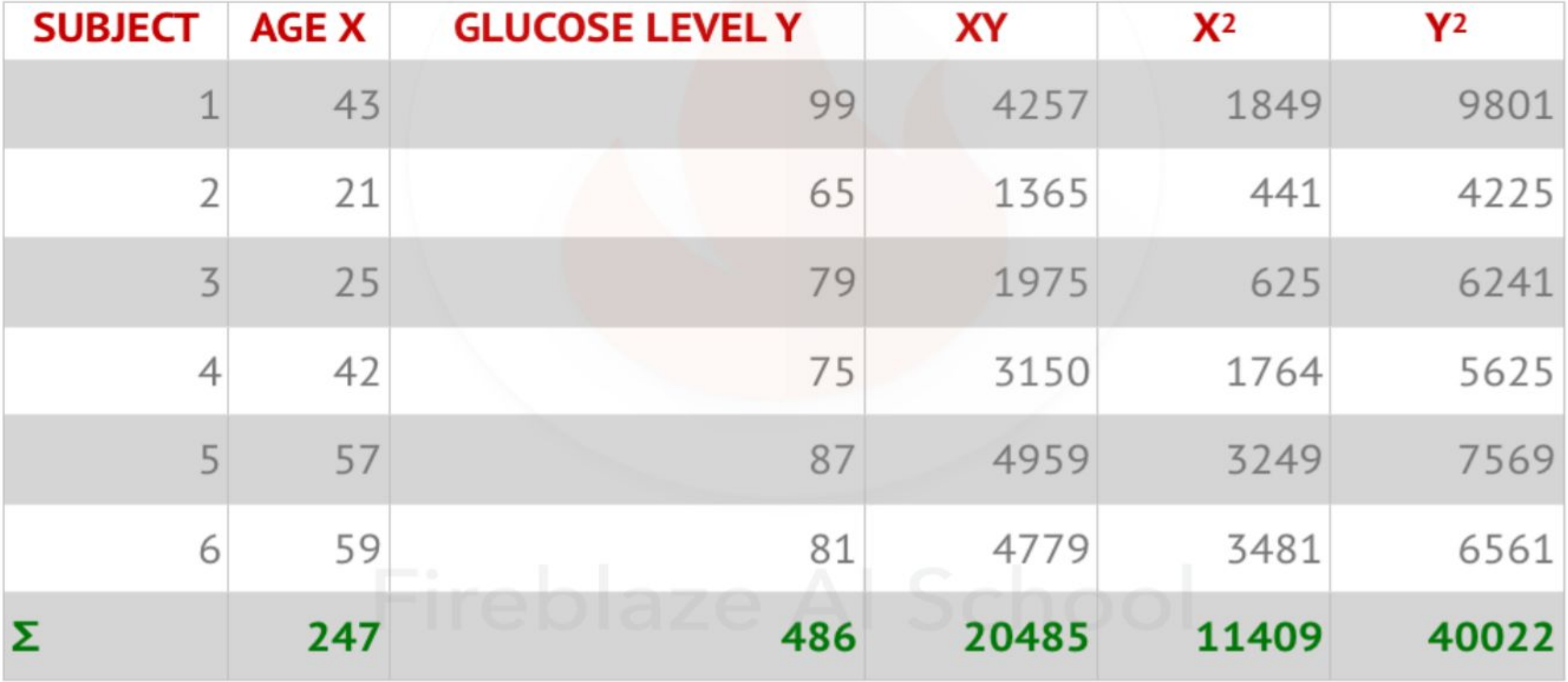
- 1 is a perfect positive correlation
- 0 is no correlation
- -1 is a perfect negative correlation
- The value shows how good the correlation is even if it is positive or negative. Note: Correlation is not Causation
- "Correlation Is Not Causation" ... which says that a correlation does not mean that one thing causes the other (there could be other reasons the data has a good correlation).

# Correlation Coefficient

Formula: Pearson’s

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

# Correlation Coefficient



| SUBJECT | AGE X | GLUCOSE LEVEL Y | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 43 | 99 | 4257 | 1849 | 9801 |
| 2 | 21 | 65 | 1365 | 441 | 4225 |
| 3 | 25 | 79 | 1975 | 625 | 6241 |
| 4 | 42 | 75 | 3150 | 1764 | 5625 |
| 5 | 57 | 87 | 4959 | 3249 | 7569 |
| 6 | 59 | 81 | 4779 | 3481 | 6561 |
| **Σ** | **247** | **486** | **20485** | **11409** | **40022** |

# Correlation Coefficient

From our table:

- $\Sigma x = 247$, $\Sigma y = 486$ , $\Sigma xy = 20,485$
- $\Sigma x2 = 11,409$, $\Sigma y2 = 40,022$

n is the sample size, in our case = 6
The correlation coefficient = 6(20,485) – (247 × 486) / [√[[6(11,409) – (247)2] × [6(40,022) – (486)2]] = 0.5298

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.

# Covariance

# Covariance

Covariance Formula:

For Population $\mathrm{Cov}(X,Y) = \Sigma( X_i - \bar{X} )(Y_j - \bar{Y}) / n$

For Sampling $\mathrm{Cov}(X,Y) = \Sigma( X_i - X )(Y_j - Y) / n - 1$

Where:

- $X_i$ – the values of the X-variable
- $Y_j$ – the values of the Y-variable
- $\bar{X}$ – the mean (average) of the X-variable
- $\bar{Y}$ – the mean (average) of the Y-variable
- n – the number of data points

# Covariance

**Covariance** measures the total variation of two random variables from their expected values.

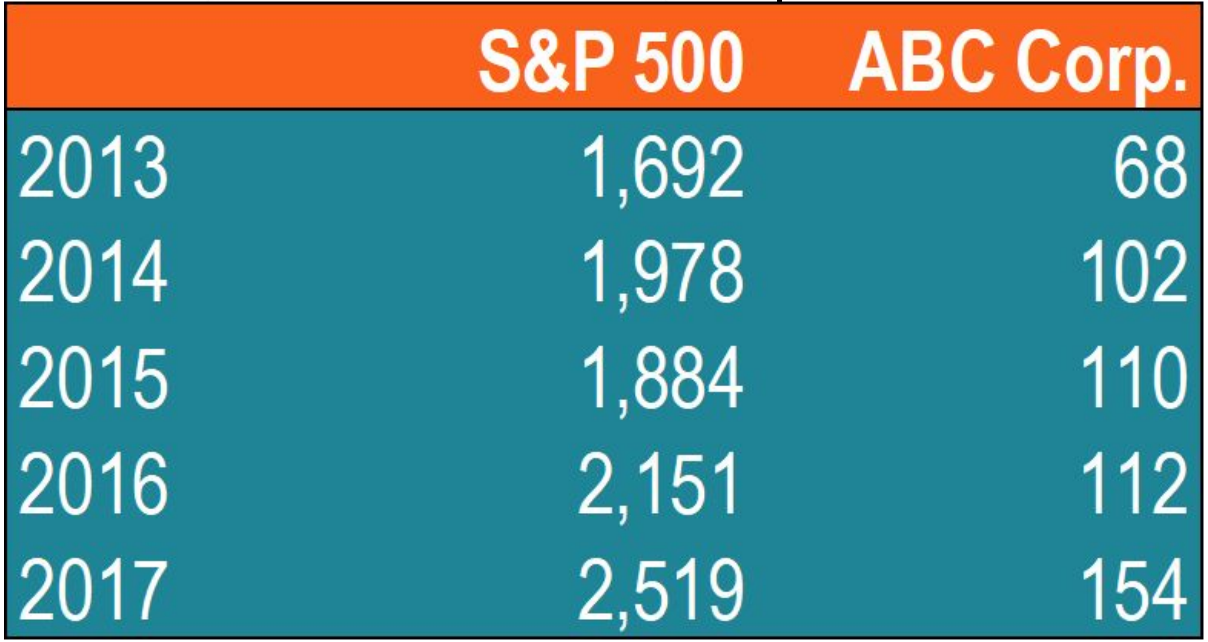
**“Covariance”** indicates the direction of the linear relationship between variables.

**“Correlation”** on the other hand measures both the strength and direction of the linear relationship between two variables

# Example of Covariance

Our objective is assess the directional relationship between the stock and the S&P 500.

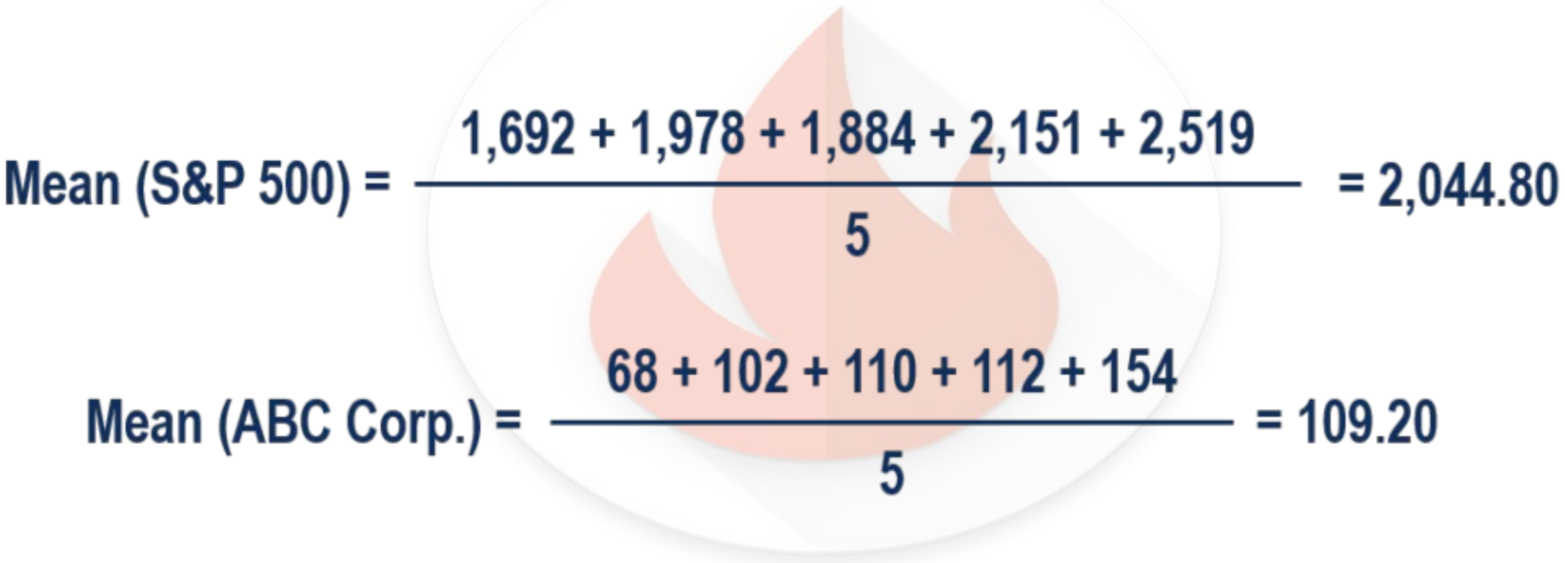
1. Obtain the data.



| | S&P 500 | ABC Corp. |
|---|---|---|
| 2013 | 1,692 | 68 |
| 2014 | 1,978 | 102 |
| 2015 | 1,884 | 110 |
| 2016 | 2,151 | 112 |
| 2017 | 2,519 | 154 |

# Example of Covariance

2. Calculate the mean (average) prices for each asset.

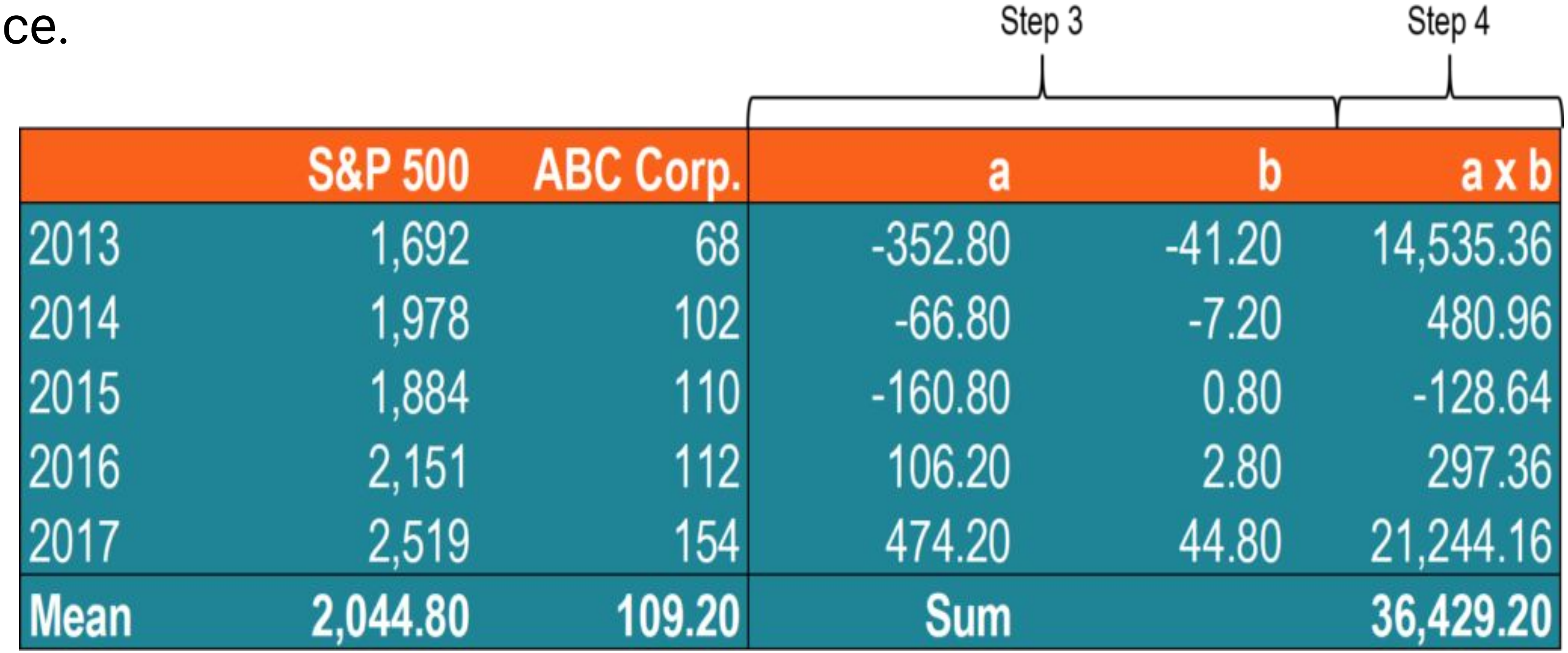


$$\text{Mean (S\&P 500)} = \frac{1{,}692 + 1{,}978 + 1{,}884 + 2{,}151 + 2{,}519}{5} = 2{,}044.80$$

$$\text{Mean (ABC Corp.)} = \frac{68 + 102 + 110 + 112 + 154}{5} = 109.20$$

# Example of Covariance

3. For each security, find the difference between each value and mean price.



Step 3

Step 4

| | S&P 500 | ABC Corp. | a | b | a x b |
|---|---|---|---|---|---|
| 2013 | 1,692 | 68 | -352.80 | -41.20 | 14,535.36 |
| 2014 | 1,978 | 102 | -66.80 | -7.20 | 480.96 |
| 2015 | 1,884 | 110 | -160.80 | 0.80 | -128.64 |
| 2016 | 2,151 | 112 | 106.20 | 2.80 | 297.36 |
| 2017 | 2,519 | 154 | 474.20 | 44.80 | 21,244.16 |
| **Mean** | **2,044.80** | **109.20** | **Sum** | | **36,429.20** |

# Example of Covariance

4. Multiply the results obtained in the previous step.

5. Using the number calculated in step 4, find the covariance.

$$\text{Cov(S\&P 500, ABC Corp.)} = \frac{36{,}429.20}{5 - 1} = 9{,}107.30$$

In such a case, the positive covariance indicates that the price of the stock and the S&P 500 tend to move in the same direction.

# Probability

A Probability is the numerical value representing the chance or possibility that a particular event will occur.

# Probability

1. The probability involved is a proportion whose value range between 0 & 1.
2. In probability theory, an **event** is a set of outcomes of an experiment to which a probability is assigned.

# Types of Probability

1. There are three types of probability.
    a. A Priori Probability
    b. Empirical Probability
    c. Subjective

- Probability of occurrences = X/T
    - X = number of ways in the event occurs
    - T = Total number of possible outcome.

# Types of Probability

**A Priori Probability:-** In a priori probability the probability of an occurrence is based on prior knowledge of the process involved.

i. A die has six faces. If you roll die, what is the probability that you will get a face with **‘five dots’**?

ii. The probability of getting a face with five dots is ⅙.

# Types of Probability

**Subjective Probability:-** In a subjective probability the probability differs from **‘person to person’**.

i. This probabilities to various outcomes usually based on a combination of an individual's past experience.

ii. Example:- The development team for a new product may design a probability of 0.60 to the chance of success for the product. While president of the company may be less optimisties an design probability of 0.30

# Types of Probability

**Empirical Probability:-** In a empirical probability the probability are based on **'observed data'**.

i. Survey are often used to generate empirical probabilities.

ii. Example:- Survey of population

# Types of Probability

**Joint Probability:-** Refer to the probability of an occurrence of involving two or more event.

i. P(A and B) is the probability that events A & B both occurs.

# Types of Probability

**Marginal Probability:-** The marginal probability of an event consists of a set of **'joint probability'**.

i. $P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \ldots\ldots.. + P(A \text{ and } B_k)$

Where, $B_1$, $B_2$, ……. $B_k$ are mutually exclusive.

# Types of Probability

**Conditional Probability:-** Conditional probability refers to the probability of event **‘A’** given information about the occurrence of another event **‘B’**.

Where, A and B are not independent it is often useful to compute the conditional **P(A|B)** which is the probability of **A** given that **B** occurred.

$$P(A|B) = P(A \text{ and } B) \mid P\ (B)$$

# Types of Probability

**Conditional Probability:-**

**“The chance of both things happening is the chance that the first one happens, and then the second one given the first happened”.**

# Distribution

- A distribution is simply a collection of data on a variable.
- The distribution describe the grouping or the density of the data called the "Probability Density Function".
- Also find out the likelihood of an observations having equal to or smaller than a given value.
- These relationships between is called as "Cumulative Density Function".

# Probability Density Function

- Is a mathematical expression that defines the distribution of the values for a **continuous** variable.
- There are three types of distribution
  - Normal Distribution
  - Uniform Distribution
  - Exponential Distribution

# Normal Distribution

## Normal Distribution

- The normal distribution like bell-shaped curve.
- The normal distribution is symmetrical, it means that most observed values tend to cluster around the mean.
- So, due to the distribution symmetrical shape, mean is equal to the median.
- Range of normal distribution from -ve infinity to +ve infinity.

# Normal Distribution

## Normal Distribution

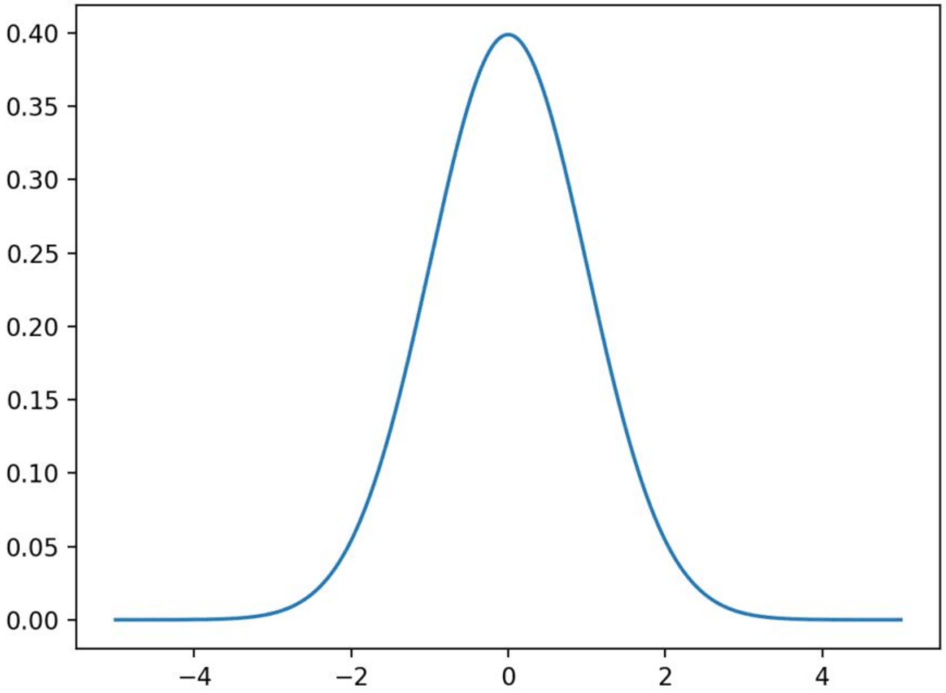
- The normal distribution also known as "**Gaussian Distribution**".
- The normal distribution provides the *statistical inference* because of its relationship to the "**Central Limit Theorem**".
- These distribution, calculate the probability that values occur within range.

# Normal Distribution Properties

## Normal Distribution

- It is symmetrical, and its "**mean and median** " are equal.
- Its interquartile range is equal to 1.33 *standard deviations.*
- It has an range (-ve infinity to +ve infinity).

# Normal Distribution



# Uniform Distribution

## Uniform Distribution

- Where the values are equally distributed in the range between the smallest value and the largest value.
- Also called as ‘rectangular distribution’.
- This distribution is symmetrical, therefore the mean equals the median.

# Exponential Distribution

## Exponential Distribution

- This distribution is skewed to the right, making the mean larger than the median.
- The range this distribution is **zero** to **+ve infinity**.

# Accuracy

Accuracy is how close a measured value is to the actual (true) value. aCcurate is Correct (a bullseye)

# Precision

Precision is how close the measured values are to each other.

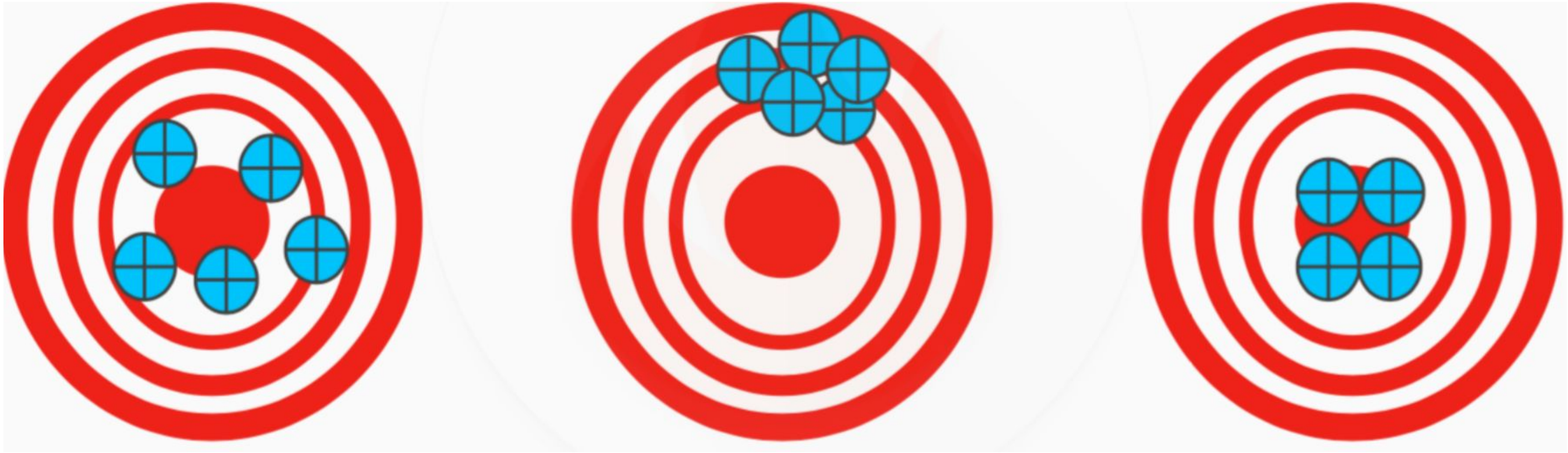
pRecise is Repeating (hitting the same spot, but maybe not the correct spot)

# Bias

When we measure something several times and all values are close, they may all be wrong if there is a "Bias"

Bias is a systematic (built-in) error which makes all measurements wrong by a certain amount.

# Accuracy and Precision



High Accuracy
Low Precision

Low Accuracy
High Precision

High Accuracy
High Precision

# Sampling Bias

A sampling bias is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

# Central Limit Theorem

# Central Limit Theorem

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution (also known as a “bell curve")

CLT is a statistical theory stating that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

# Central Limit Theorem

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger.
- Sample sizes equal to or greater than 30 are considered sufficient for the CLT to hold.
- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
- A sufficiently large sample size can predict the characteristics of a population accurately.

# Central Limit Theorem

Central limit theorem examples, you will be given:

- A population (i.e. 29-year-old males, seniors between 72 and 76, all registered vehicles, all cat owners)
- An average (i.e. 125 pounds, 24 hours, 15 years, $15.74)
- A standard deviation (i.e. 14.4lbs, 3 hours, 120 months, $196.42)
- A sample size (i.e. 15 males, 10 seniors, 79 cars, 100 households)

# Central Limit Theorem

1. General Steps

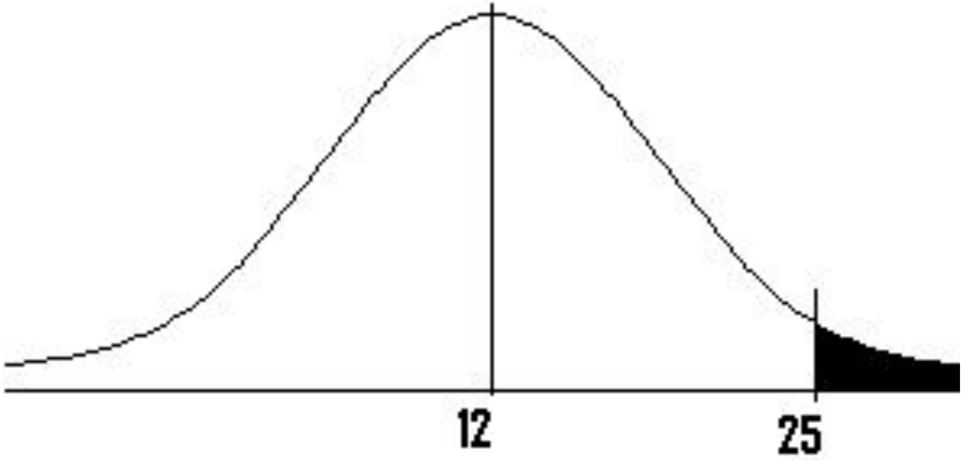
Step 1: Identify the parts of the problem. Your question should state:

- the mean (average or μ)
- the standard deviation (σ)
- population size
- sample size (n)

a number associated with “greater than” ( central limit theorem examples 2). Note: this is the sample mean. In other words, the problem is asking you “What is the probability that a sample mean of x items will be greater than this number?

# Central Limit Theorem

Step 2: Draw a graph. Label the center with the mean. Shade the area roughly above xbar (i.e. the “greater than” area).



# Central Limit Theorem

Step 3: Use the following formula to find the z-score. Plug in the numbers from step 1.

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

- Subtract the mean (μ in step 1) from the ‘greater than’ value (mean(x) step 1). Set this number aside for a moment.
- Divide the standard deviation (σ in step 1) by the square root of your sample (n in step 1). For example, if thirty six children are in your sample and your standard deviation is 3, then 3 / √36 = 0.5
- Divide your result from step 1 by your result from step 2 (i.e. step 1/step 2)

# Central Limit Theorem

Step 4: Subtract your z-score from 0.5. For example, if your score is 0.1554, then 0.5 – 0.1554 = 0.3446.

Step 5: Convert the decimal in Step 5 to a percentage. In our example, 0.3446 = 34.46%.

# Central Limit Theorem

Example:- A certain group of welfare recipients receives SNAP benefits of \$110 per week with a standard deviation of \$20. If a random sample of 25 people is taken, what is the probability their mean benefit will be greater than \$120 per week?

- Step 1: Insert the information into the z-formula: $= (120-110)/20 \sqrt{25} = 10/ (20/5) = 10/4 = 2.5$.

# Central Limit Theorem

- Step 2: Look up the z-score in a table (or calculate it using technology). A z-score of 2.5 has an area of roughly 49.38%. Adding 50% (for the left half of the curve), we get 99.38%.

# Central Limit Theorem

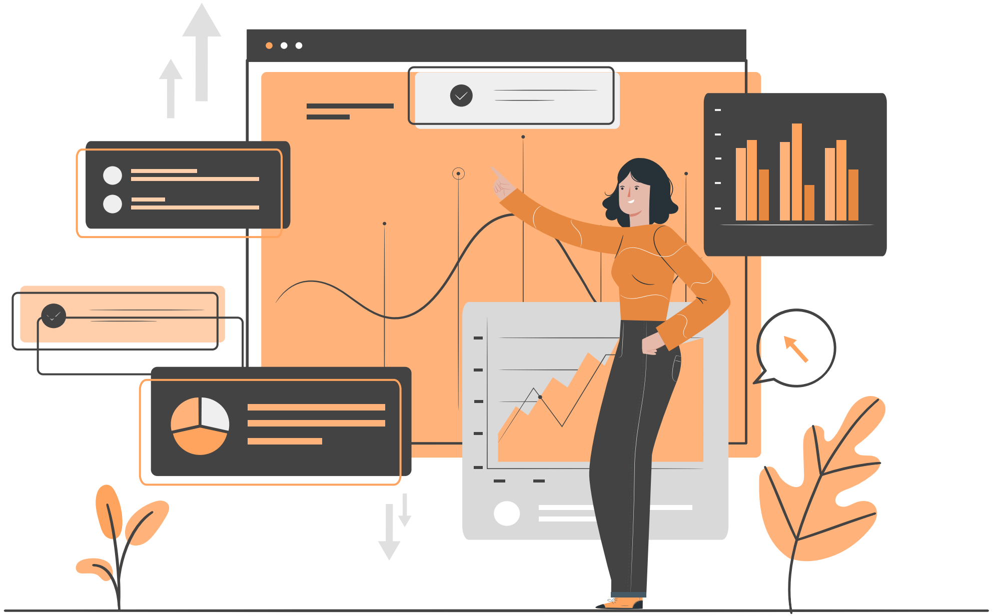
Example :- A population of 29 year-old males has a mean salary of $29,321 with a standard deviation of $2,120. If a sample of 100 men is taken, what is the probability their mean salaries will be less than $29,000?

- Step 1: Insert the values into the z-formula:= (29,000 – 29,321) / (2,120/√100) = -321/212 = -1.51.
- Step 2: Look up the z-score in the left-hand z-table (or use technology). -1.51 has an area of 93.45%.

# Central Limit Theorem

However, this is not the answer, as the question is asking for LESS THAN, and 93.45% is the area “greater than” so you need to subtract from 100%.

- 100% – 93.45% = 6.55% or about 0.07.



# Univariate & Bivariate Data

# Univariate Data

Univariate means “One Variable” (one type of data)

Example: Travel Time (in minutes): 15, 28, 9, 25, 36, 11, 45 The variable is Travel Time

Things to do with Univariate Data:

- Find a central value using mean, median and mode
- Find how spread out it is using range, quartiles and standard deviation
- Make plots like Bar Graphs, Pie Charts and Histograms

# Bivariate Data

Bivariate means “Two Variables” (two types of data)

- Example: Here are two variables Ice Cream Sales and Temperature: Univariate and Bivariate Data

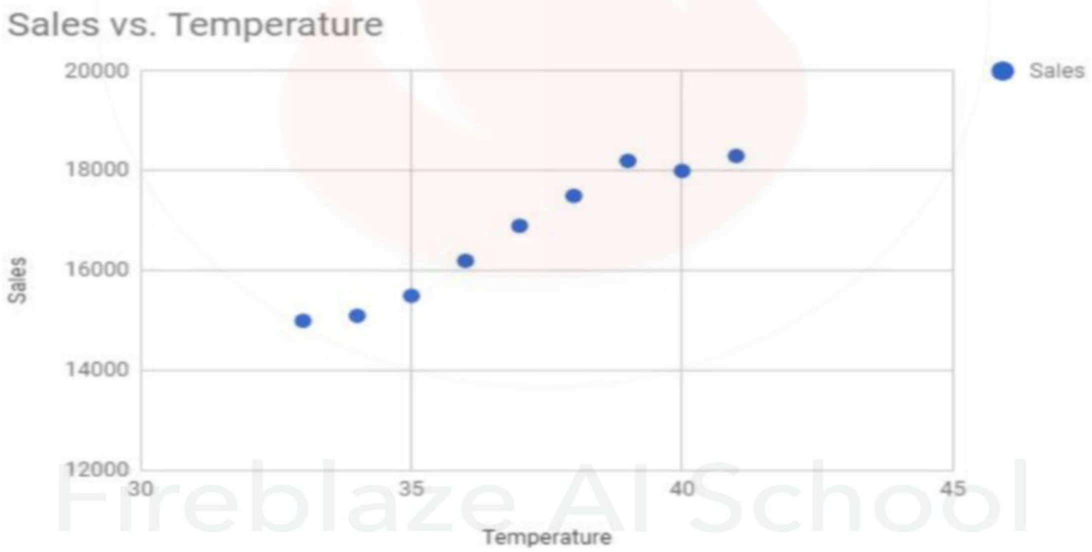
| Temperature | Ice Cream Sales (In Rupees) |
|---|---|
| 33 | 15000 |
| 37 | 16000 |
| 39 | 17500 |
| 42 | 22500 |

# Bivariate Data

With bivariate data: Comparing the two sets of data and finding any relationships. Use Tables, Scatter Plots, Correlation, Line of Best Fit to find out these relationships.

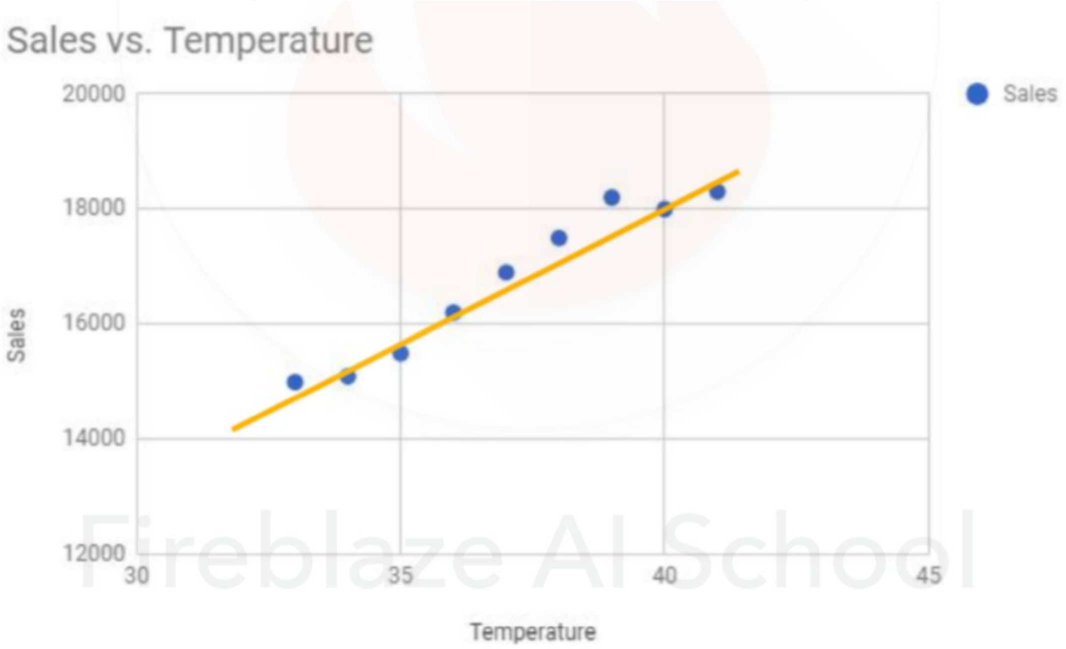
# Scatter Plot

A Scatter(XY) Plot has points that show the relationship between two sets of data



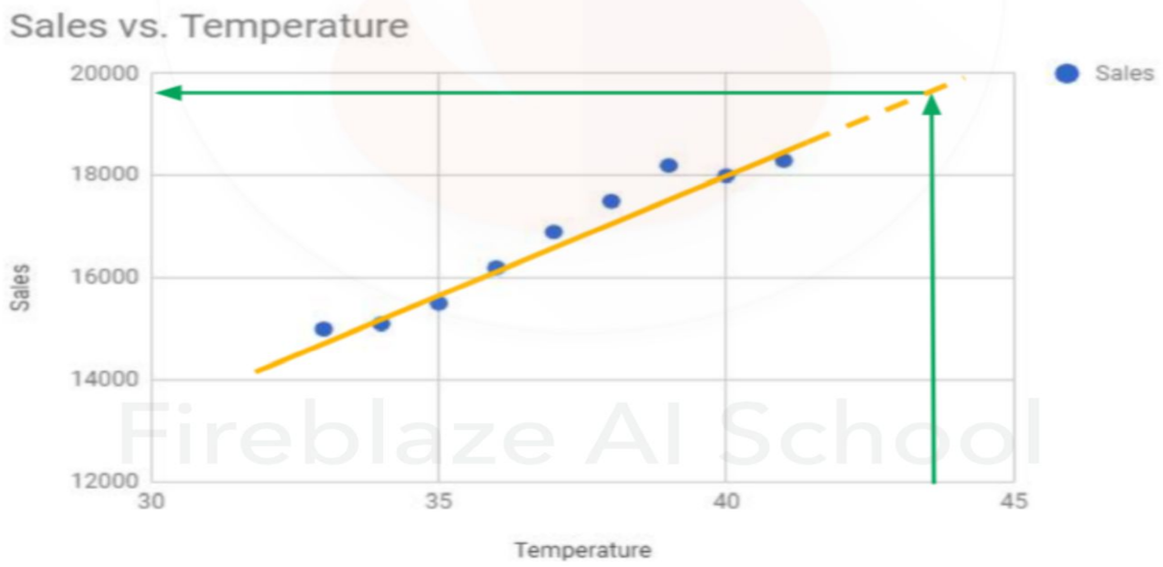
# Line Of Best Fit

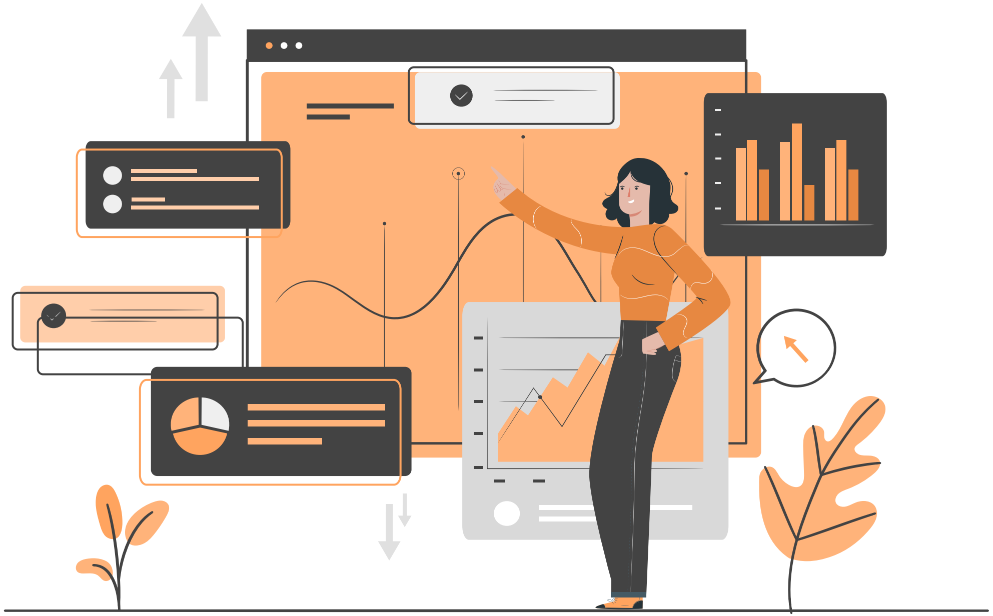
The line as close as possible to all points and as many points above the line as below.



# Extrapolation

Find a value outside set of data points. Using Linear Extrapolation, we can estimate Sales i.e. INR 19990 at Temperature 44°C





# Data Preprocessing

# Introduction To Data Preprocessing

Data Preprocessing is a technique that is used to convert the raw data into a clean data set.

Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn

# Methods of Data Preprocessing

- Outlier Treatment
- Null Value Imputation
- Dealing With Categorical data
- Scaling & Transformation
- Splitting the data-set

# Outliers

Outliers: "Outliers" are values that "lie outside" the other values. Example: Long Jump A new coach has been working with the Long Jump team this month, and the athletes' performance has changed.

- Augustus: +0.15m
- Tom: +0.11m
- June: +0.06m
- Carol: +0.06m
- Bob: + 0.12m
- Sam: -0.56m

So here, Sam is outlier

# Outliers

**“Outliers"** are values that "lie outside" the other values.

The mean is: Including “Sam” i.e. Outlier

Mean = (0.15+0.11+0.06+0.06+0.12-0.56) / 6 = -0.06 / 6 = -0.01m

So, on average the performance went **DOWN.**

The mean is: Excluding “Sam” i.e. Outlier

Mean = (0.15+0.11+0.06+0.06+0.12)/5 = 0.1 m .

So, on average the performance went **UP**

# Outliers

Outliers: "Outliers" are values that "lie outside" the other values. The median ("middle" value):

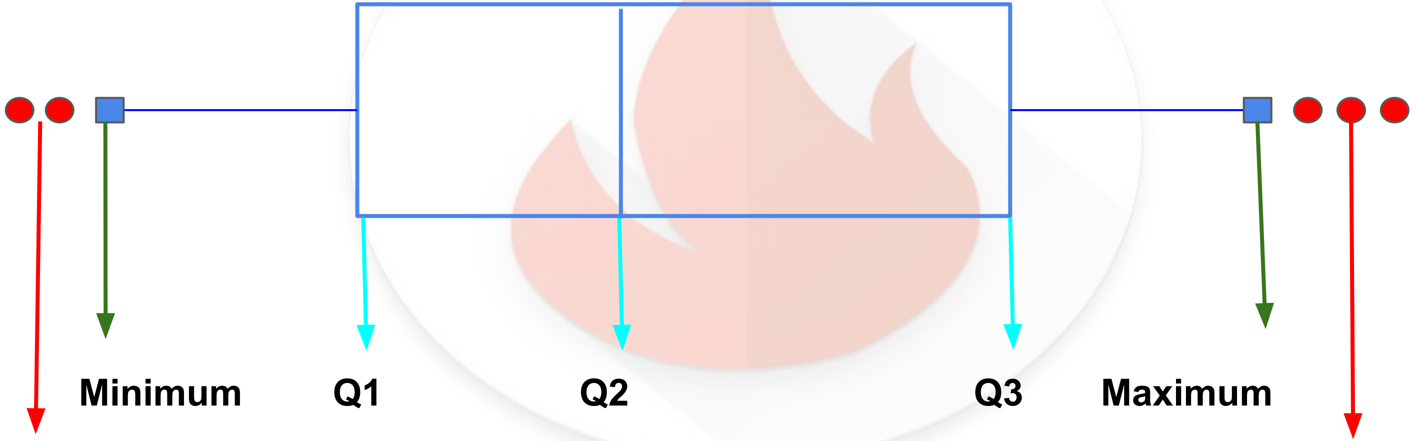
- including Sam is: 0.085
- without Sam is: 0.11 (went up a little)
- The mode (the most common value):
- including Sam is: 0.06
- without Sam is: 0.06 (stayed the same)
- The mode and median didn't change very much.

# Outlier Treatment Using IQR

- The main advantage of the IQR is that it is not affected by outliers because it doesn’t take into account observations below Q1 or above Q3.
- It might still be useful to look for possible outliers in your study.
- As a rule of thumb, observations can be qualified as outliers when they lie more than 1.5 IQR below the first quartile or 1.5 IQR above the third quartile. Outliers are values that "lie outside" the other values. Outliers = Q1 - 1.5 * IQR OR Outliers = Q3 + 1.5 * IQR

# Box Plot

There is one graph that is mainly used when you are describing centre and variability of your data. It is also useful for detecting outliers in the data.



Outlier = Value < Q1 - 1.5 * IQR

Outlier = Value > Q3 + 1.5 * IQR

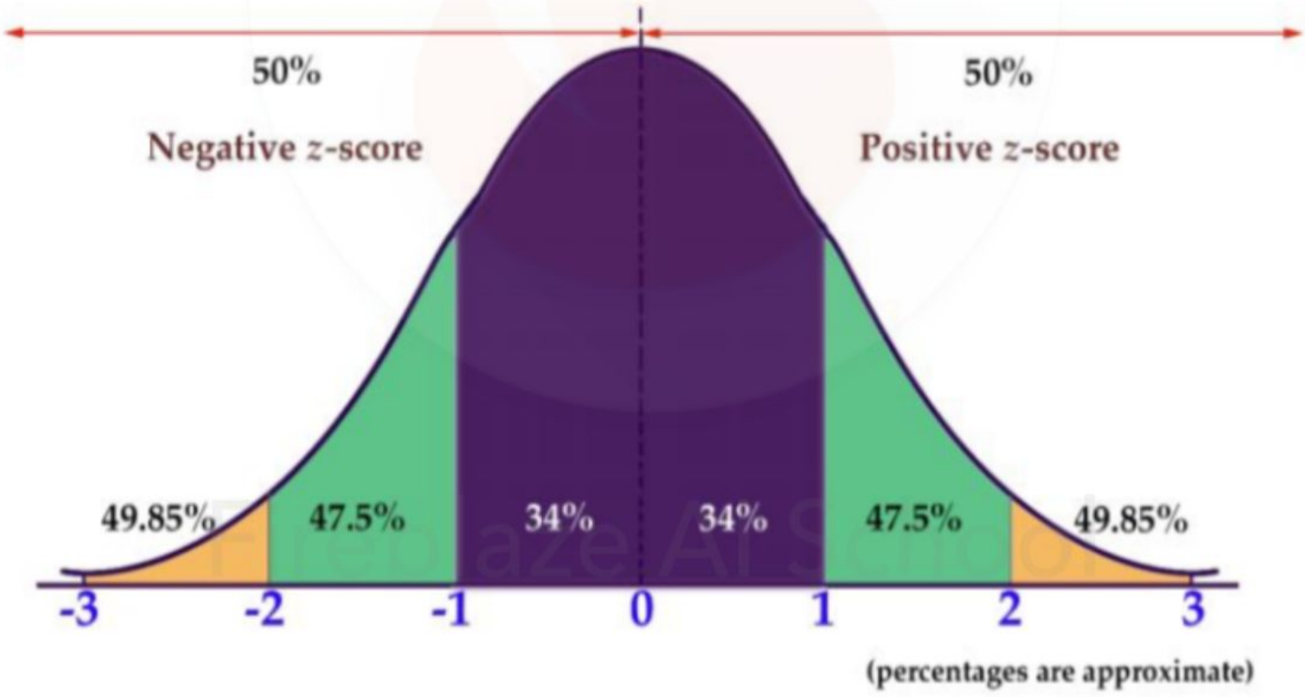
# Outlier Treatment Using Z Score

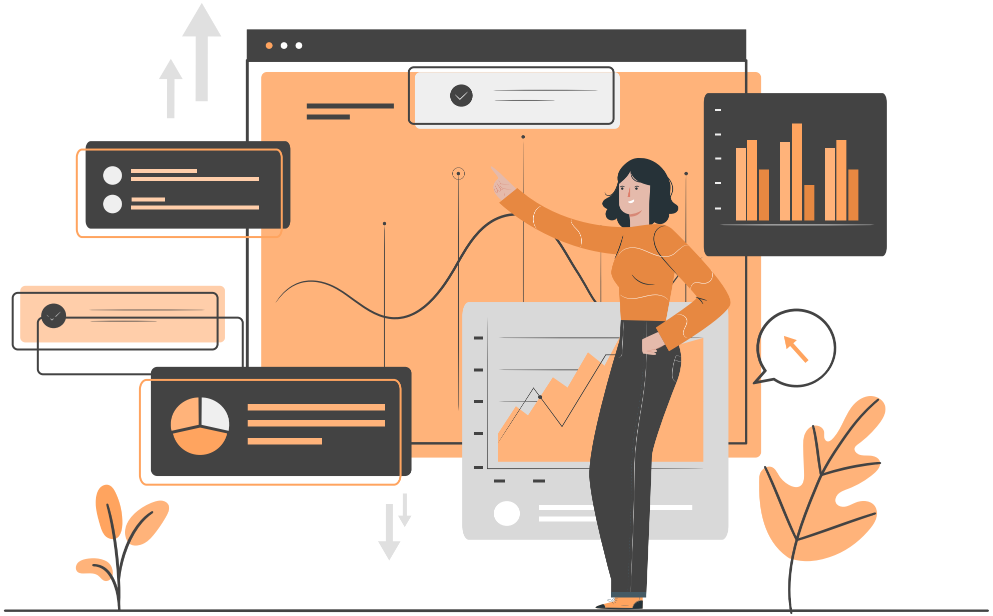
## Bell Shape Distribution and Empirical Rule:

If distribution is bell shape then it is assumed that about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; and about 99% have a z-score between -3 and 3.

According to the value of **z score is greater than 3 is consider as outlier**

# Bell Shaped Distribution





# Missing Values

# Missing Values

The missing of data occurs commonly because of the manual entry procedures and extraction methods.

Missing data can have a significant effect on the conclusions that can be drawn from the data.

Missing of data can be handled in two ways,

1. **Dropping**

2. **Imputing**

# Dropping

- Dropping Null Values
    - df.dropna()
    - axis – We can specify axis=0 if we want to remove the rows and axis=1 if we want to remove the columns.
    - how – If we specify how = ‘all’ then the rows and columns will only be dropped if all the values are NaN.By default how is set to ‘any’.
    - thresh – It determines the threshold value so if we specify thresh=5 then the rows having less than 5 real values will be dropped.

# Dropping

- subset —If we have 4 columns A,B,C and D then if we specify subset=[‘C’] then only the rows that have their C value as NaN will be removed.
- inplace — By default no changes will be made to your dataframe. So if you want these changes to reflect onto your dataframe then you need to use inplace = True.

# Imputation

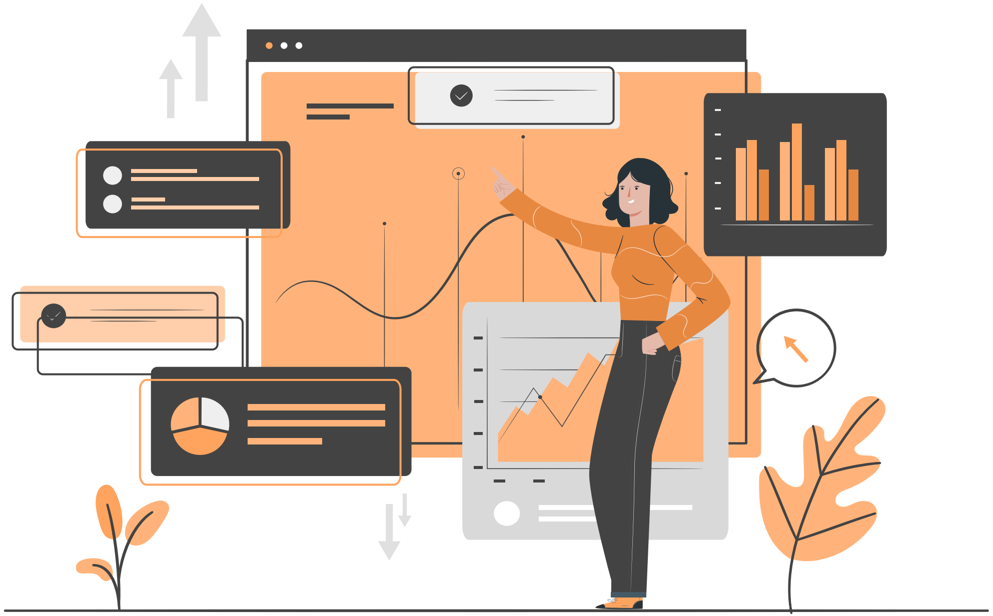
- Convert into boolean
    - df.isnull()
    - Returns a boolean matrix, if the value is NaN then True otherwise False
- Count Of Null Values
    - df.isnull().sum()
    - Returns the column names along with the number of NaN values in that particular column

# Imputation of Null Value

- Fill Null Values
  - df.fillna()
  - It replace all null values with given some input values ex. df.fillna(4) replace all null value with 4
- When your data is not skewed i.e Symmetric/Normally Distributed. In other words, there are no extreme values present in the data set then fill numerical null value using **Mean**.
- When your data is skewed or you are dealing with ordinal (ordered categories) data then fill numerical null value using **Median**.

# Imputation of Null Value

- When your data is skewed or you are dealing with ordinal (ordered categories) data then fill numerical null value using **Median**.
- When dealing with nominal (unordered categories) data the fill null value using **Mode**.



# Dealing With Categorical Data

# Dealing With Categorical Data

The machine learning models are based on mathematical equations and can operate only on numeric values.

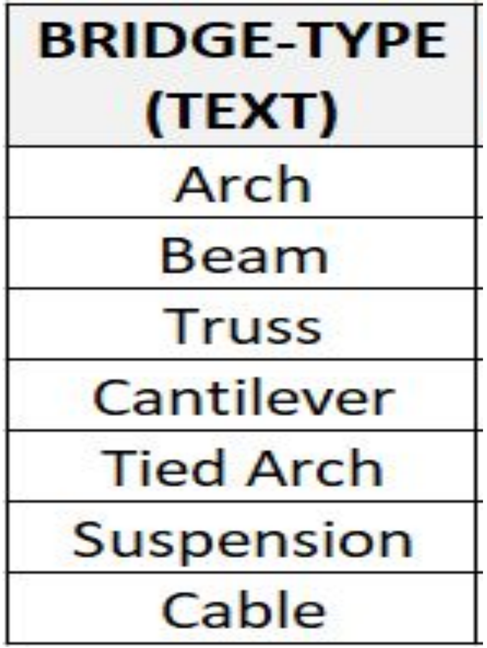
So, Need to transform these categorical data into numerical values.

Methods to transform these categorical data into numerical values.

1. **Label Encoding**
2. **One Hot Encoding**

# Label Encoding

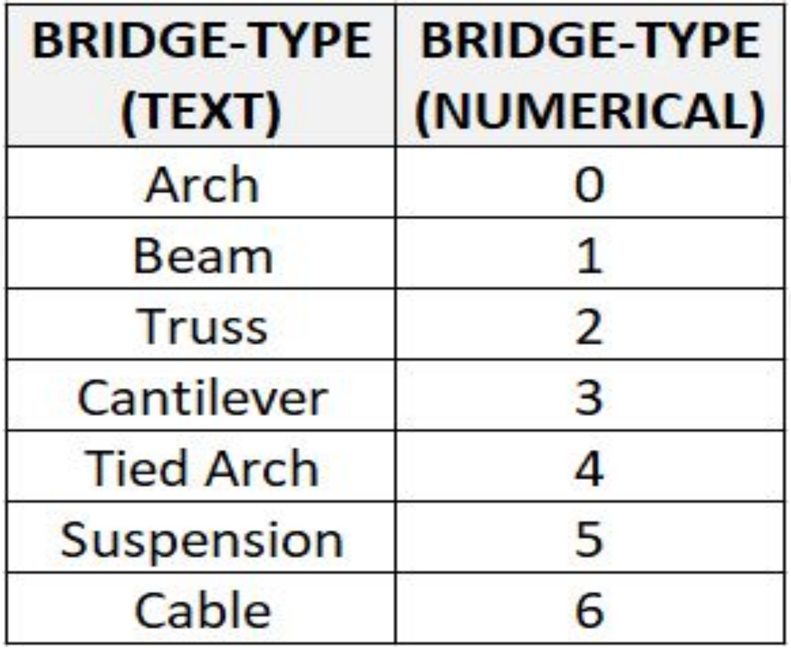
Consider a dataset of bridges having a column names bridge-types contain categorical values.



| BRIDGE-TYPE (TEXT) |
|---|
| Arch |
| Beam |
| Truss |
| Cantilever |
| Tied Arch |
| Suspension |
| Cable |

# Label Encoding

After applying label encoding categorical columns converted into numeric value



| BRIDGE-TYPE (TEXT) | BRIDGE-TYPE (NUMERICAL) |
|---|---|
| Arch | 0 |
| Beam | 1 |
| Truss | 2 |
| Cantilever | 3 |
| Tied Arch | 4 |
| Suspension | 5 |
| Cable | 6 |

# Label Encoding

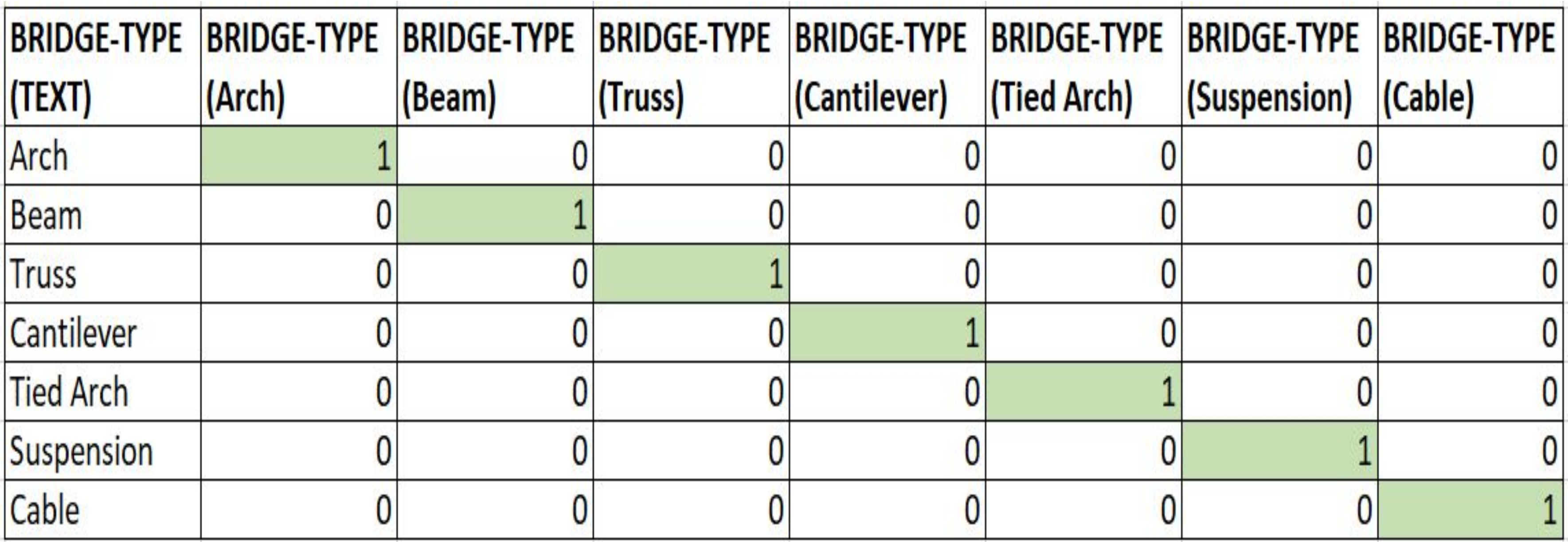
- Label encoding have  induces a problem with uses number sequencing.
- The problem using the number is that they introduce relation/comparison between them.
- The algorithm might misunderstand that data has some kind of hierarchy/order 0 < 1 < 2 … < 6 and might give 6X more weight to into calculation

# One Hot Encoding

In One Hot Encoding , each category value is converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column.

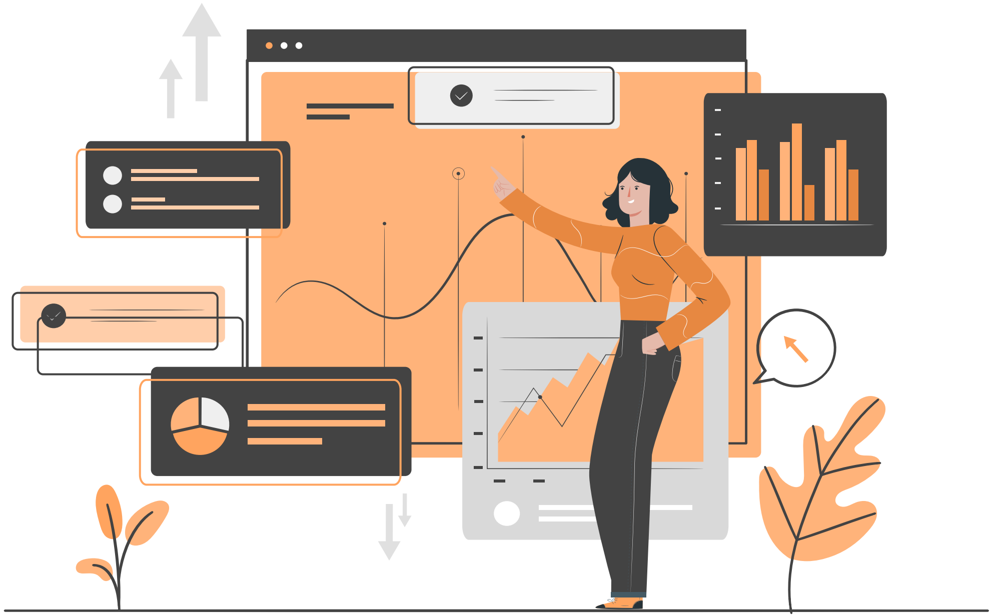
Let’s consider the previous example of bridge type with one-hot encoding.

# One Hot Encoding



| BRIDGE-TYPE (TEXT) | BRIDGE-TYPE (Arch) | BRIDGE-TYPE (Beam) | BRIDGE-TYPE (Truss) | BRIDGE-TYPE (Cantilever) | BRIDGE-TYPE (Tied Arch) | BRIDGE-TYPE (Suspension) | BRIDGE-TYPE (Cable) |
|---|---|---|---|---|---|---|---|
| Arch | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Beam | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Truss | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Cantilever | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Tied Arch | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Suspension | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Cable | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Though this approach eliminates the hierarchy/order issues but does have the downside of adding more columns to the data set.



# Scaling & Transformation

# Scaling And Transformation

Most machine learning algorithms take into account only the magnitude of the measurements, not the units of those measurements.

So that is expressed in a very high magnitude (number), may affect the prediction a lot more than an equally important feature.

Example :- you have two lengths, l1 = 250 cm and l2 = 2.5 m. We humans see that these two are identical lengths (l1 = l2), but most ML algorithms interpret this quite differently.

# Scaling And Transformation

You see, the algo is going to give a lot more weight to l1, just because it is expressed in a larger number , which in turn is going to have a much larger impact on the prediction than l2.

**Note:-** That not all algorithms behave this way. Certain types like Naive Bayes, Decision Trees, RF and XGB do not require feature scaling, because they work in a different manner.

# Scaling And Transformation

There are three types of Scaling

**MinMaxScaler** :- Use as your default

**StandardScaler** :- Use if you need normalized features

**Robust Scaler** :- Use if you have outliers and can handle a larger range

# Min Max Scalar

The Min-Max Scaler uses the following formula for calculating each feature:

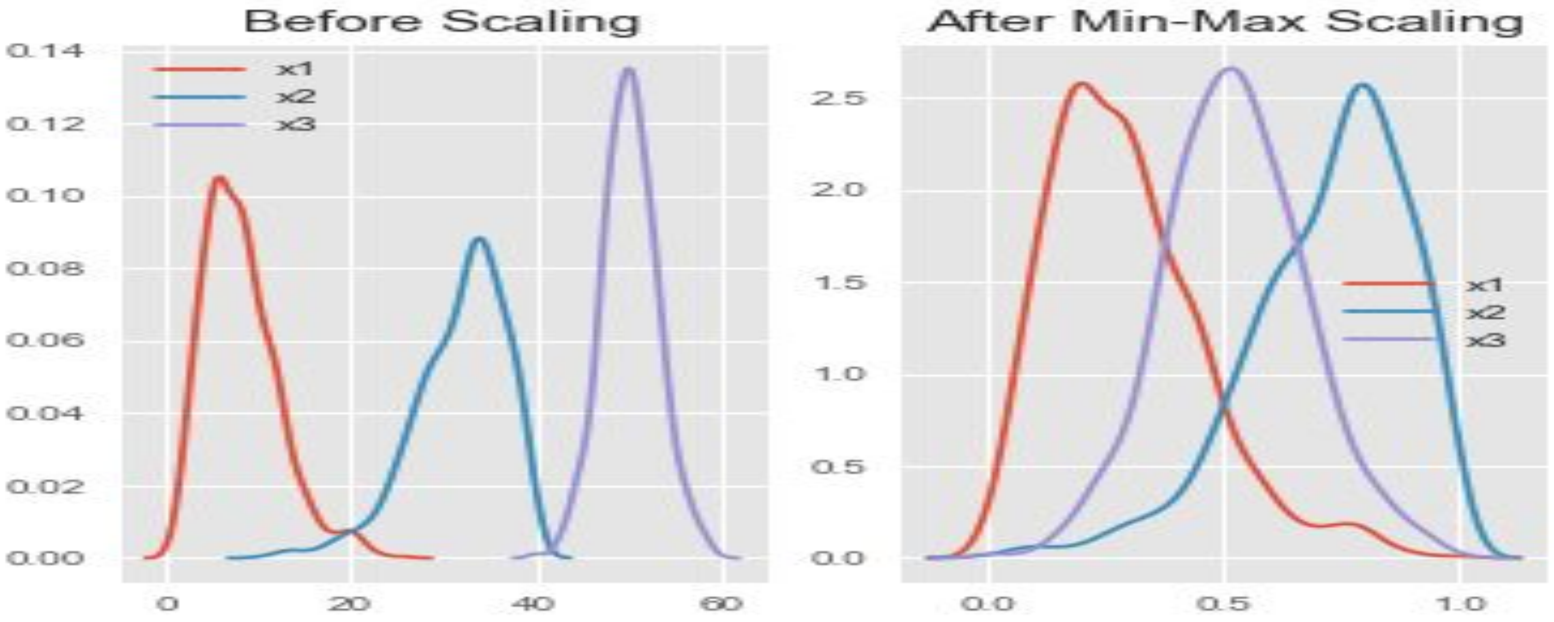
**Min - Max Scalar = (xi−min(x))/ (max(x)−min(x))**

It transforms the data so that it’s now in the range you specified. You specify the range by passing in a tuple to the Feature_Range parameter.

Note that, by default, it transforms the data into a range between 0 and 1 (-1 and 1, if there are negative values).

It can be used as an alternative to The Standard Scaler or when the data is not normally distributed.

# Min Max Scalar



It uses the min and max values, so it’s **very sensitive to outliers.**
Not to use it when your data has **noticeable outliers.**

# Standard Scalar

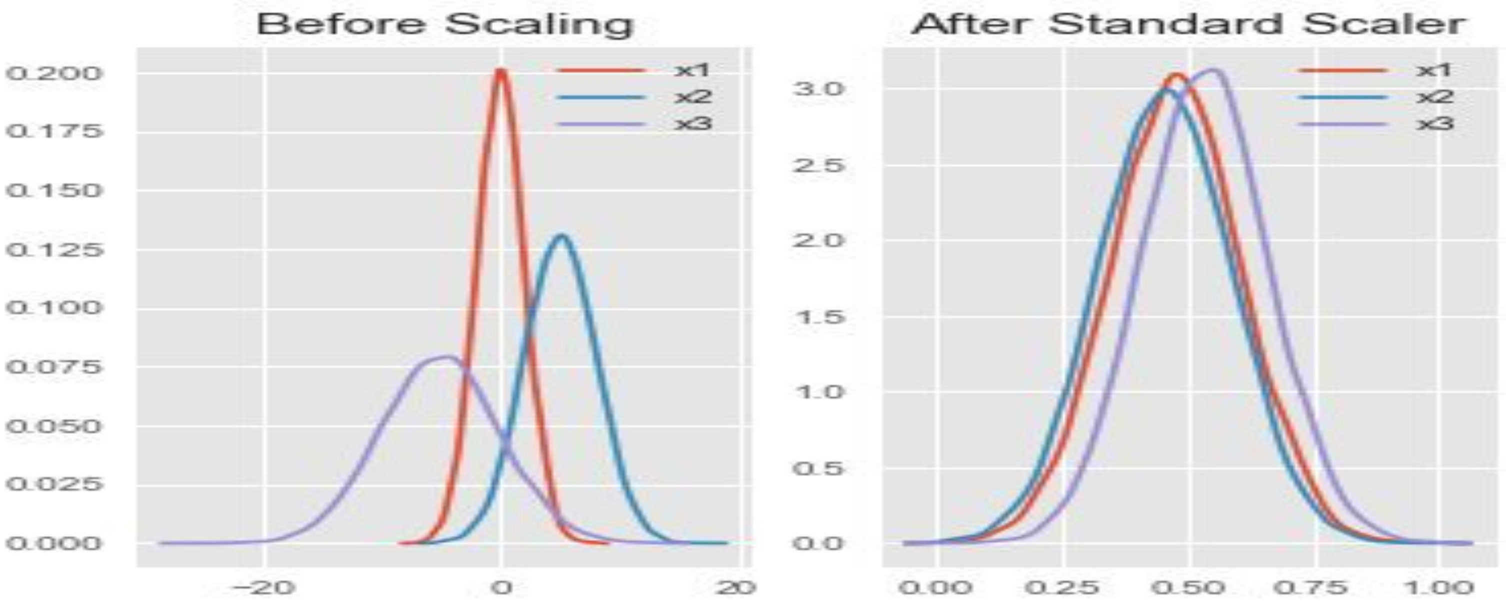
The Standard Scaler is one of the most widely used scaling algorithms out there. It assumes that your data follows a Gaussian Distribution (Gaussian distribution is the same thing as Normal distribution)

The Mean and the Standard Deviation are calculated for the feature and then the feature is scaled based on:

**SC= (xi–mean(x)) / stdev(x)**

The idea behind Standard Scaler is that it will transform your data, such that the distribution will have a mean value of 0 and a standard deviation of 1.

# Standard Scalar



If the data is not normally distributed, it’s not recommended to use the Standard Scaler.

# Robust Scalar

The Robust Scaler uses statistics that are robust to outliers:

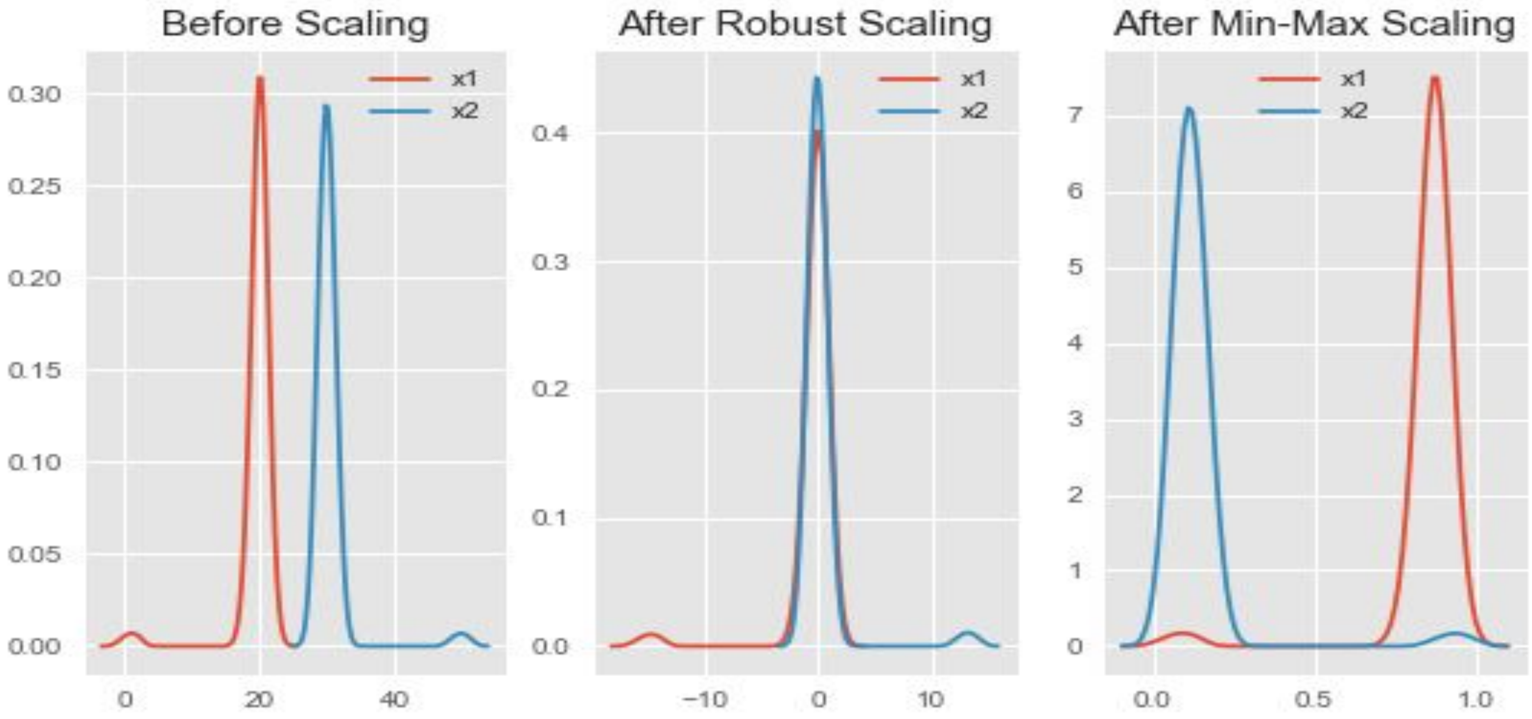
$$RS = (xi-Q1(x))/( Q3(x)-Q1(x))$$

It removes the median and uses the Interquartile Range. Q1 and Q3 represent quartiles.

The IQR is the range between the 1st quartile and the 3rd quartile.

This usage of interquartiles means that they focus on the parts where the bulk of the data is. This makes them very suitable for working with outliers.

# Robust Scalar



# Splitting Of Dataset

- Before applying machine learning models, we should split the data into two parts as the training set and test set.
- If we use 100% of the data gathered(full dataset) in training the model, we will be out of data for testing the accuracy of the model that we have built.
- So we generally split the dataset into 70:30 or 80:20 ratio (trainset:testset). Special care needs to be taken in splitting the data.

# Splitting Of Dataset

- **Training Data** :- The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.

- **Testing Data** :- After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.

# The End