IntelliPaat

Statistics and Probability

01 Intro to Statistics

02 Sampling

03 Central Tendencies

04 Variation

05 Correlation

06 Importance of Tables and charts

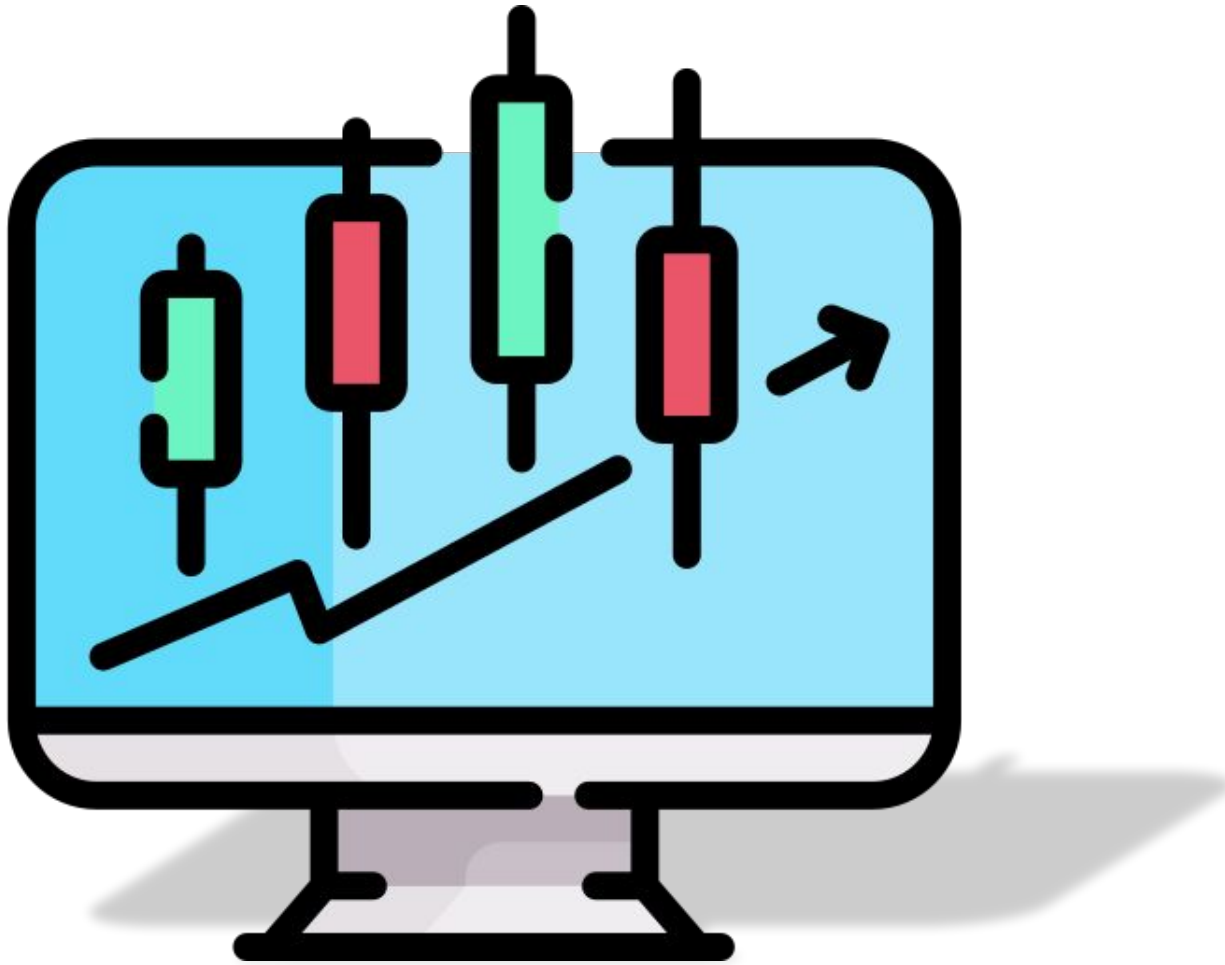
07 Probability

08 Probability in Business Analytics And Distribution

09 Hypothesis Testing , Estimation And Goodness of Fit

IntelliPaat

Introduction to Statistics



Fundamental of Data Scientist

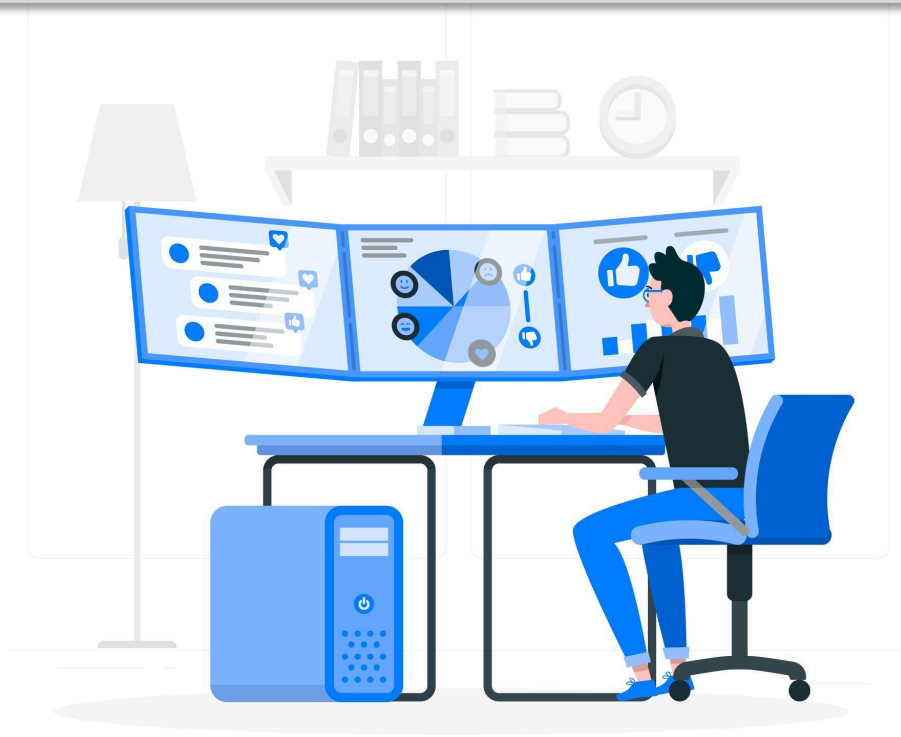
What is the kind of data we have ?

How is the data structured ?

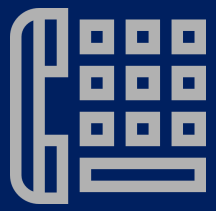
The concept of the ‘ Types of Data’ help us identify the right statistical and data science techniques that can be applied to the data.

Fundamental of Data Scientist

The fundamental job of a data scientist is to help stakeholders make informed business decisions.
Prior to deriving insights, it is imperative that we understand the data on hand



Fundamental of Data Scientist



One of the features of the data set may be the telephone numbers of participants. Since this is a numerical input one can technically calculate the mean of such numbers.
But in reality this is nominal data. It doesn't make statistical sense to calculate the mean of the phone numbers in the data set.

It is important to keep in mind that the software (Python in our case) will read whatever file is presented.
While the software has inbuilt test cases it cannot check for discrepancies in the data.

The GIGO (garbage in, garbage out) concept holds good.
So, for example, if you have categorical data and you run a multiple linear regression equation, the answer is unlikely to be correct.

Basic statistical concepts

These include things like the: Mean — the central value, commonly called the average. Median — the middle value if we ordered the data from low to high and divide it exactly in half. Mode- the value which occurs most often

Basic statistical concepts

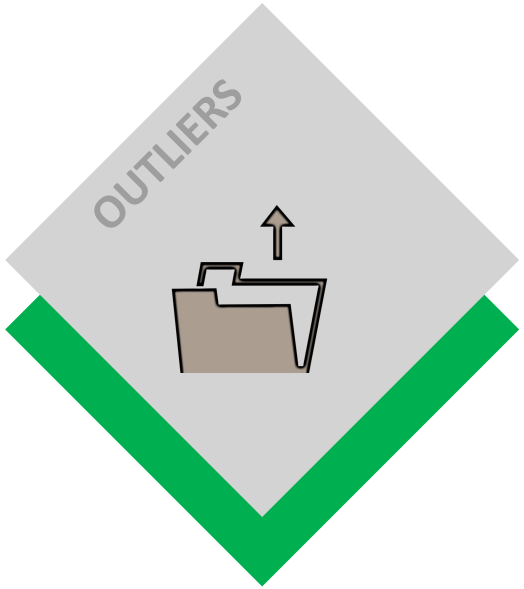
These include things like the: Mean — the central value, commonly called the average. Median — the middle value if we ordered the data from low to high and divide it exactly in half. Mode- the value which occurs most often

Variance helps us understand how spread the data is from the average value.

Basic statistical concepts

These include things like the: Mean — the central value, commonly called the average. Median — the middle value if we ordered the data from low to high and divide it exactly in half. Mode- the value which occurs most often

Variance helps us understand how spread the data is from the average value.

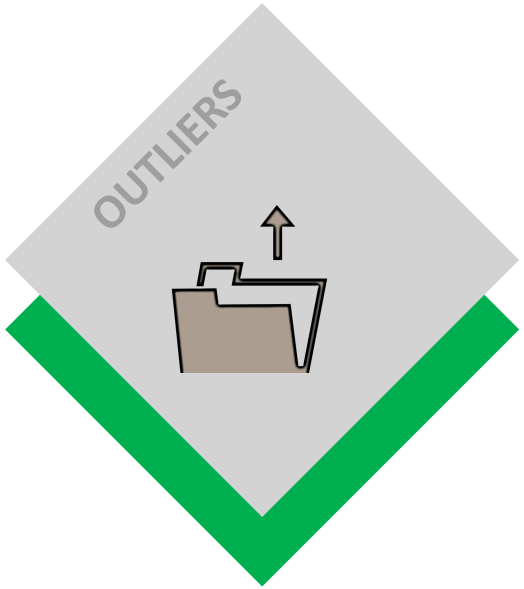


Outliers help us comprehend whether inclusion of such an event is essential to the study.

Basic statistical concepts

These include things like the: Mean — the central value, commonly called the average. Median — the middle value if we ordered the data from low to high and divide it exactly in half. Mode- the value which occurs most often

Variance helps us understand how spread the data is from the average value.



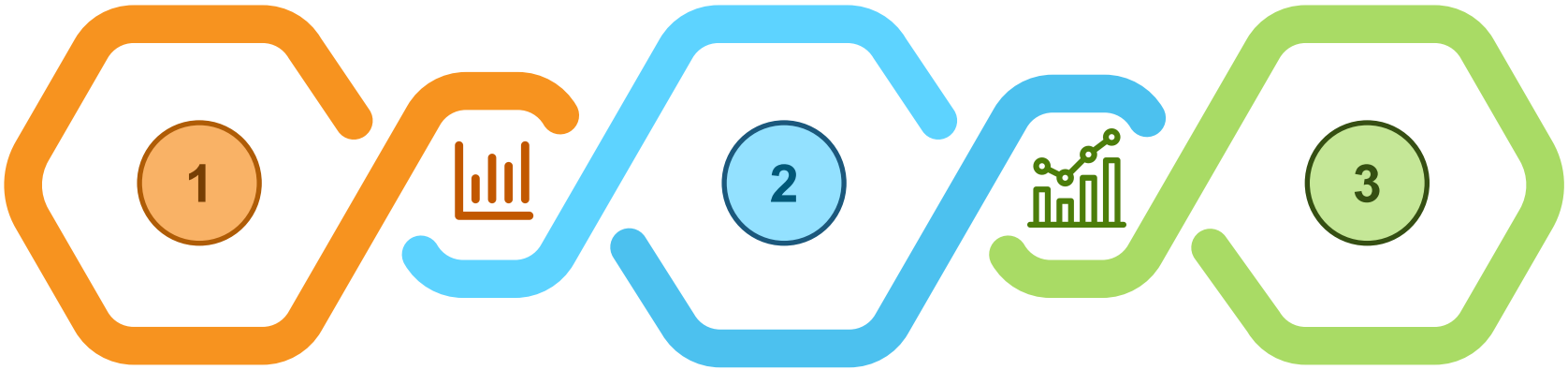
Outliers help us comprehend whether inclusion of such an event is essential to the study.

Determining which measure of central tendency needs to be opted for is also critical

Sampling

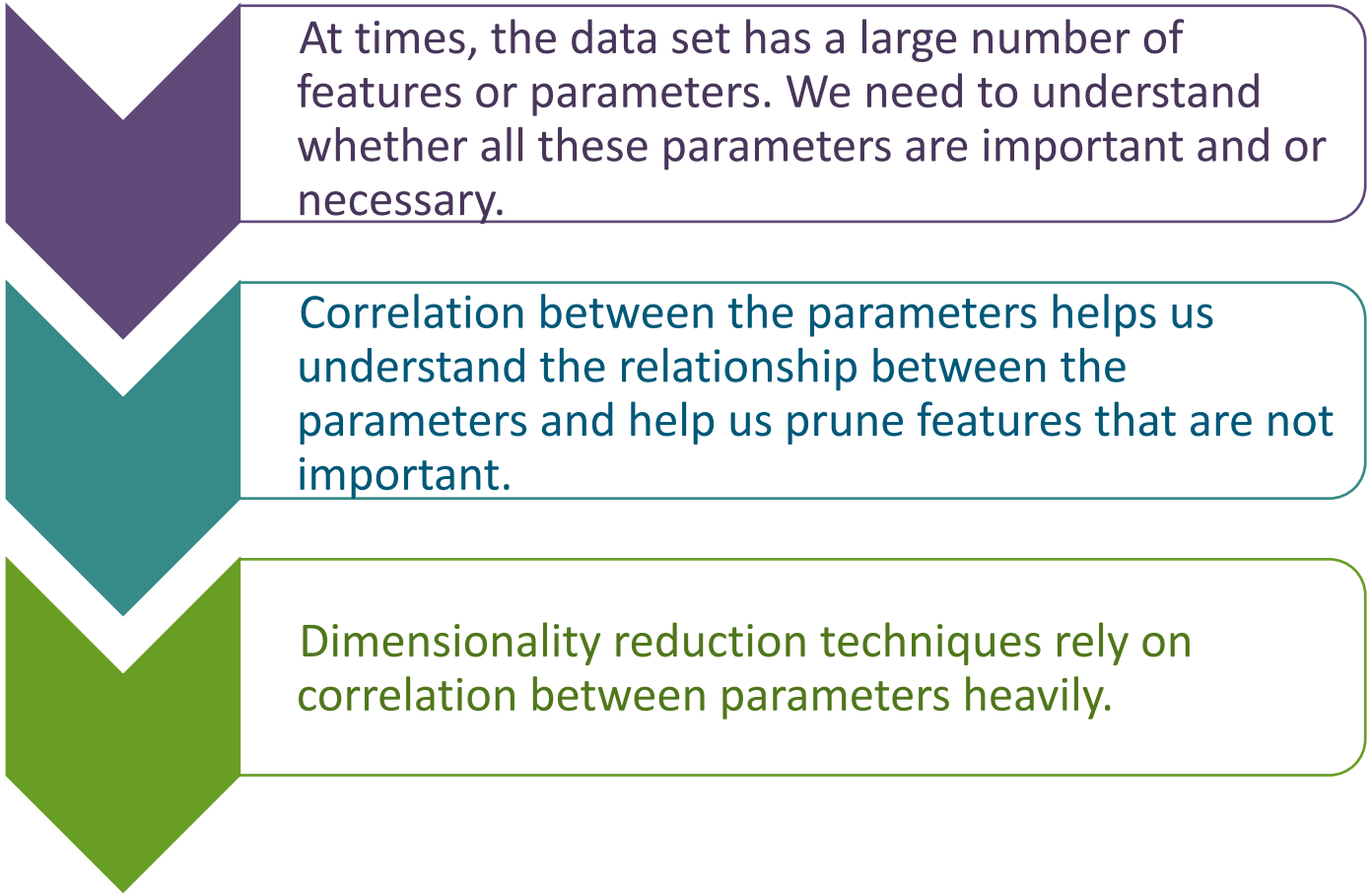
Predictive analytics involves analyzing historical data to predict future likely outcomes.

This knowledge is also very helpful to choosing the right technique that can be applied to derive insights



However, the data set we have to work, often is a sample of the population. Hence, it is important to understand if the sample is representative of the population.

The concept of correlation

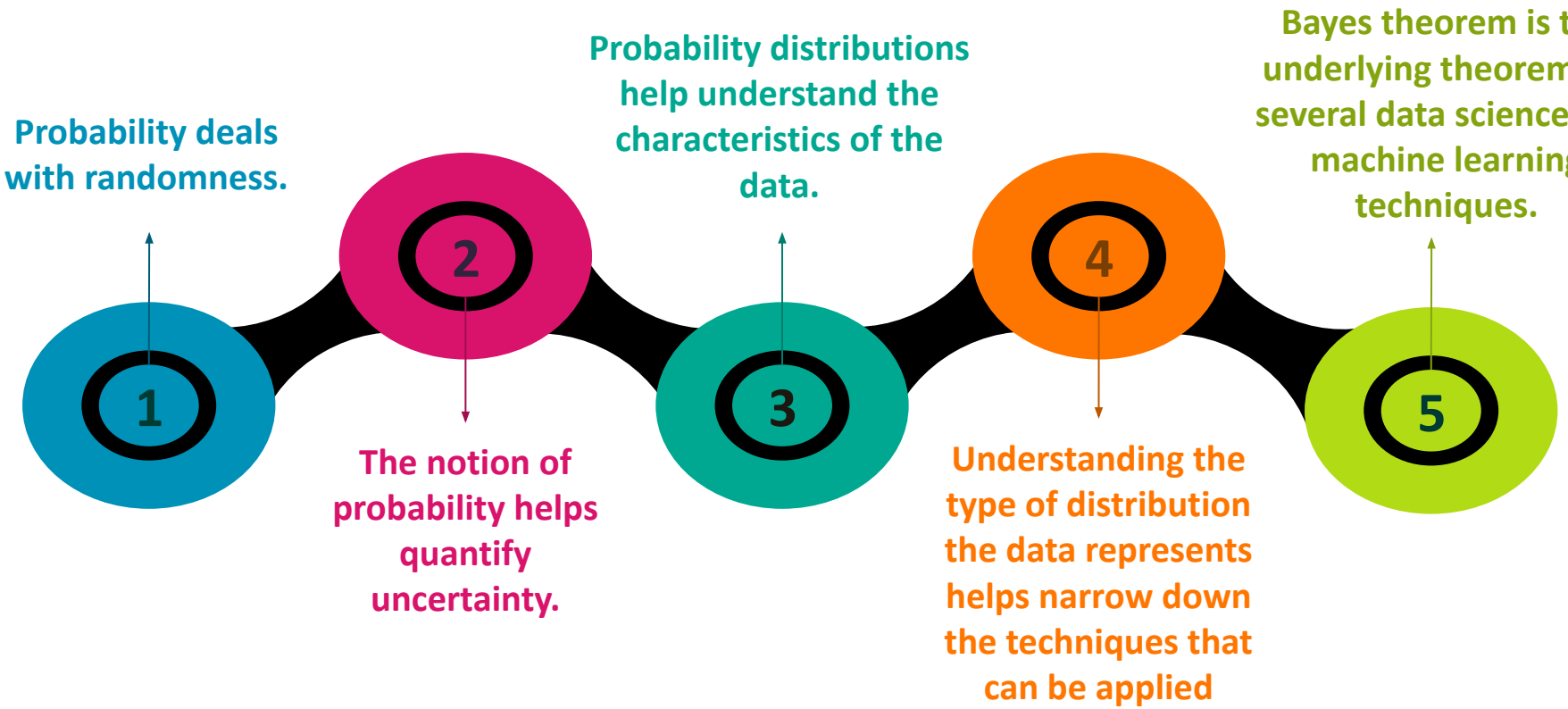


Probability and data science

Probability theory is the mathematical foundation of statistical inference, and it is required for analysing data affected by chance, making data scientists indispensable.

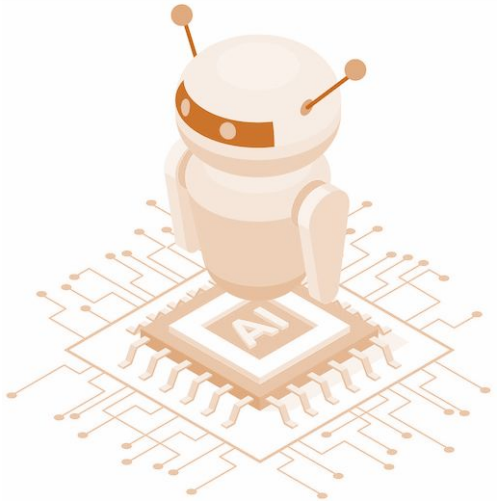
Probability and data science

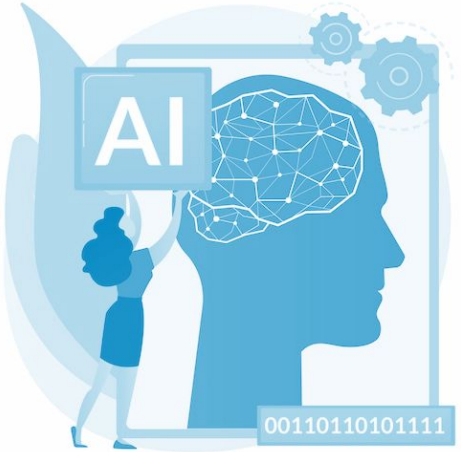
Probability theory is the mathematical foundation of statistical inference, and it is required for analysing data affected by chance, making data scientists indispensable.



Hypothesis Tests

Hypothesis tests are tests that are set up to evaluate whether statements about the population are supported by the data.





The output of several data science algorithms incorporate the results of the hypothesis test.

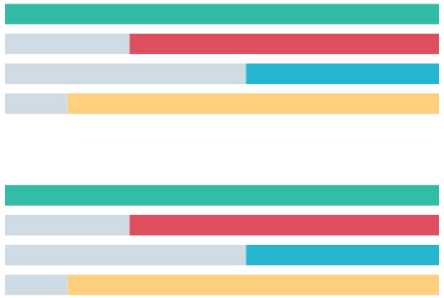
Understanding the concept of hypothesis testing is critical to interpreting the results of different algorithms

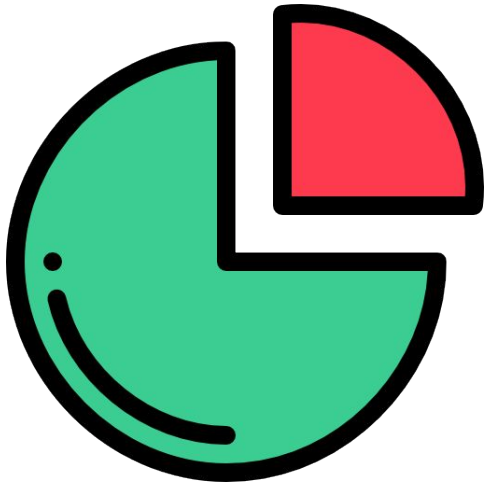
What is Statistics?

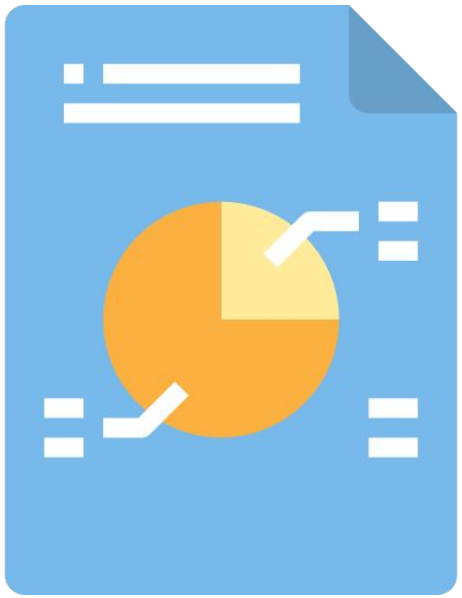
What is Statistics?

What?

Statistics is a branch of Mathematics that deals with collection, analyzing, and interpreting large amounts of data.



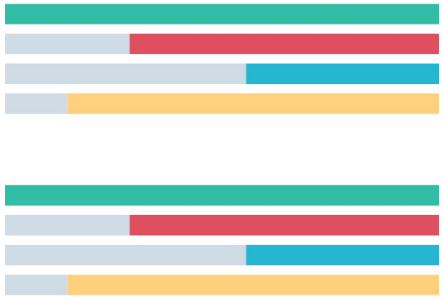


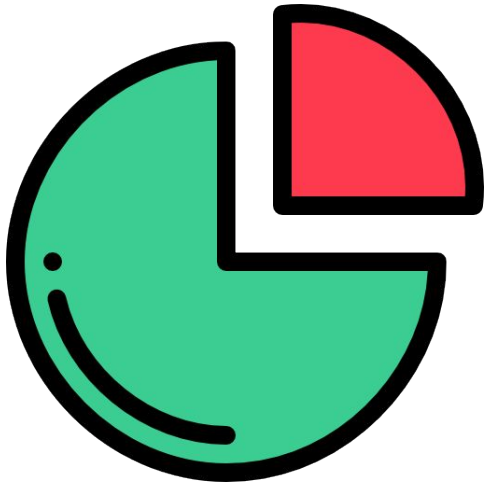


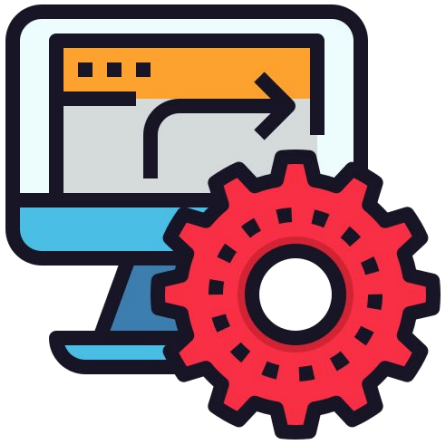
Why is Statistics important?

Why is Statistics important?

Statistics allows us to derive knowledge from large datasets and this knowledge can then be used to make predictions, decisions, classifications etc.



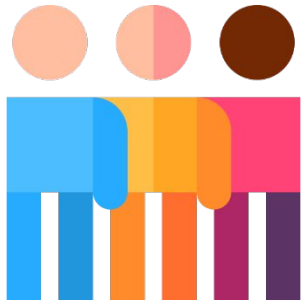




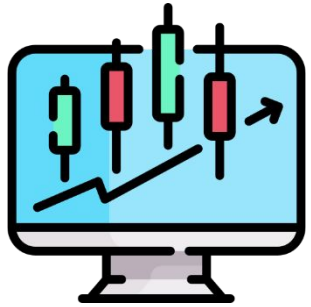
Where is Statistics used?

Where is Statistics used?

Statistics are used in various fields, some of them are:

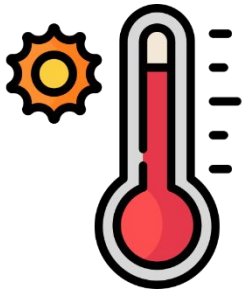


Medical Research



Stock Market

Sales Projection

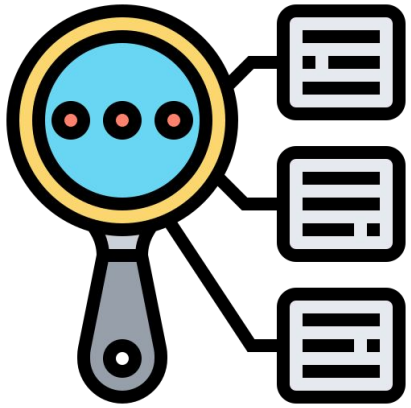


Weather Forecasting

Sampling

Sampling

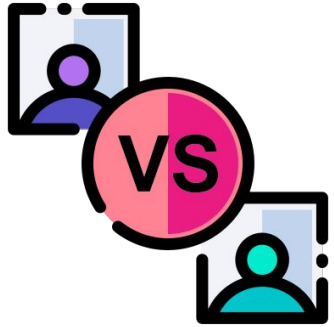
Sampling is the process of collecting data to perform analysis on

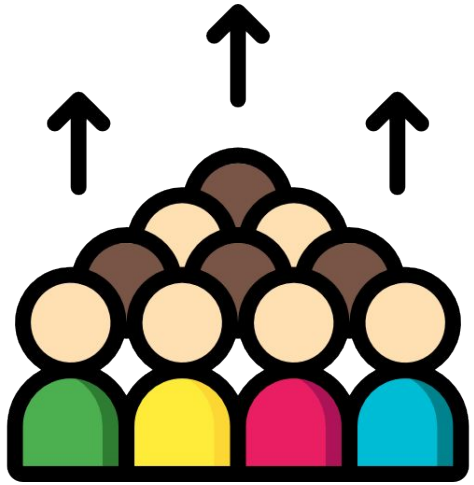


Sample vs Population

Sample vs Population

Population is the entire dataset such as the whole population of a country, **Sample** is subset of that population which is analyzed to make inferences

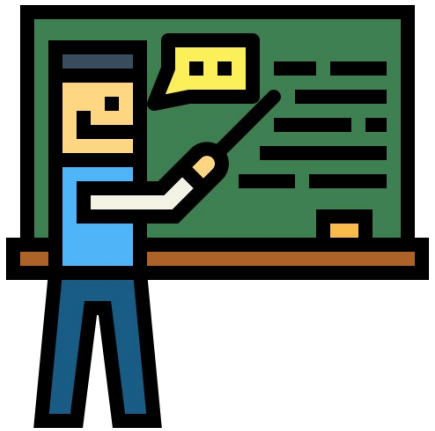


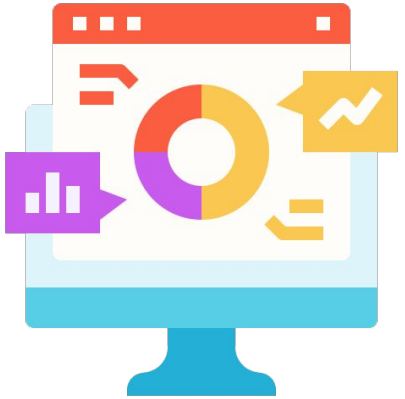


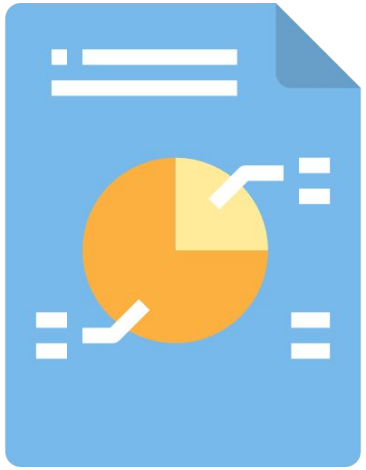
Random Sampling

Random Sampling

Random Sampling is the process of selecting a subset / sample from a population in such a way that every data point is equally likely to be included in the sample



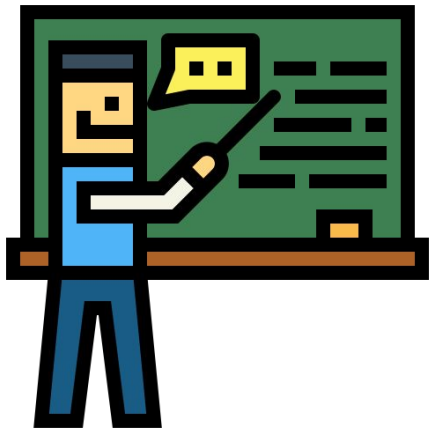


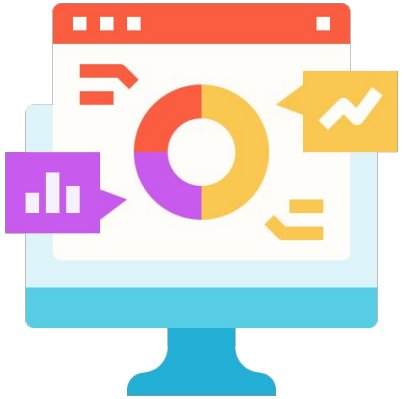


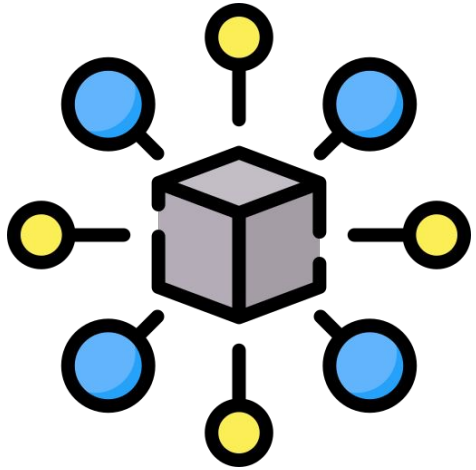
Stratified Sampling

Stratified Sampling

Stratified Sampling is the process of dividing your samples into layers or groups and then performing random sampling for each group







Central Tendencies

Central Tendencies

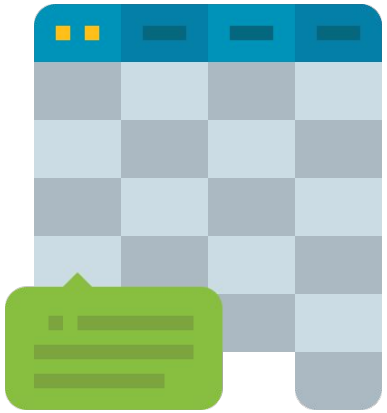
Central Tendency is used to indicate where does the middle or center of the distribution of our data lies

Mean

Mean

Mean is the average of the data. In simpler terms it's the sum of values divided by total number of values. It's represented by Greek letter Sigma

Mean



Mode

Mode

Mode is used to indicate the most frequent data point, in other words the one which occurs most number of times

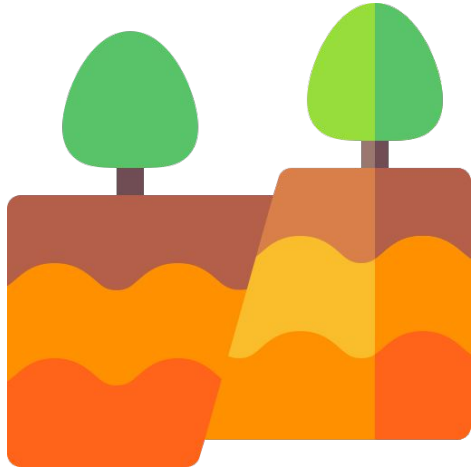
Mode

Median

Median

Median is the middle of the data. If the data is arranged in ascending order then the data element which occurs right at the center is the median

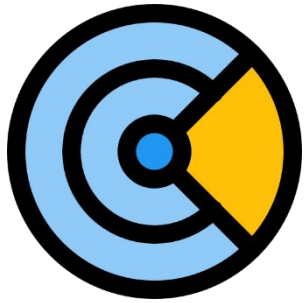
Median



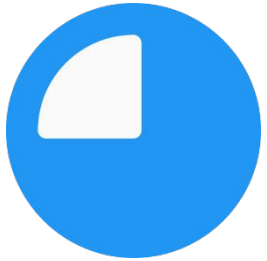
Variation

Variation

Variation in statistics is used to show how data is dispersed, or spread out. Several measures of variation are used in statistics.

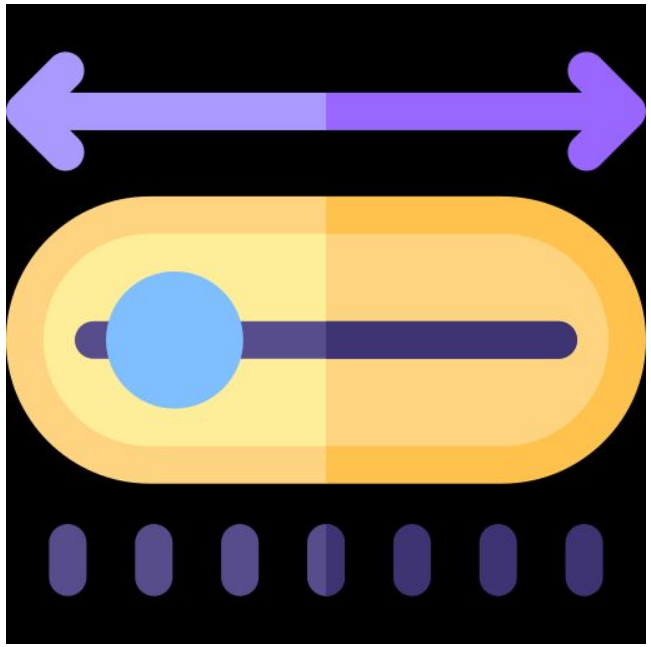


Range



Quartiles

Variance



Range

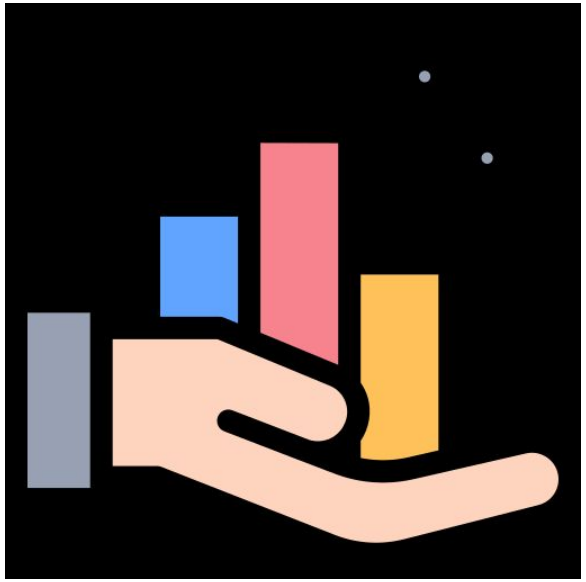
Range

Range is the difference between the highest and the lowest values in our dataset. Range tells us the distance between the lowest and highest values in our data

Percentiles

Percentiles

Percentiles are scores that are used to describe a value below which some Observations fall.
E.g.: If X is at 70th Percentile it mean 70% of other data points from our sample are below X



Quartiles

Quartiles

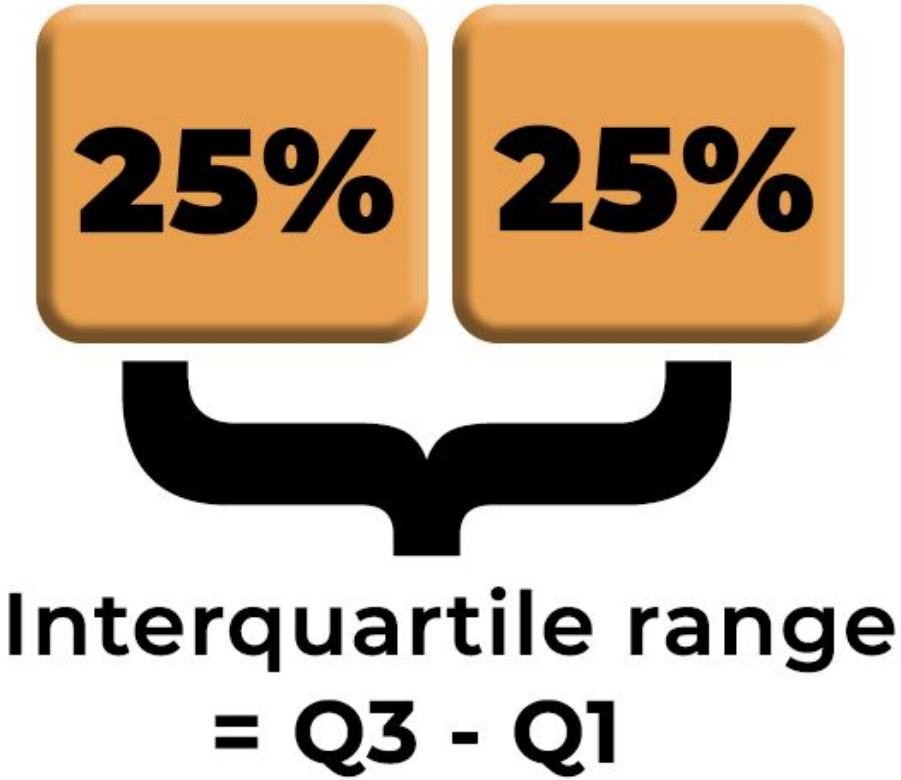
Quartiles are used to break the data into 4 parts so as to better find the spread of data in a way that is less influenced by outliers.

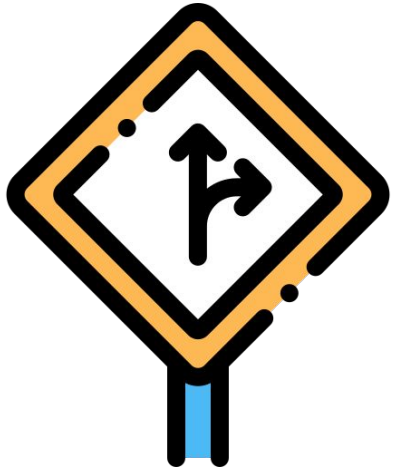
Quartiles are expressed in percentiles. 1st Quartile is 25th Percentile, 2nd Quartile is 50th Percentile (Median) and 3rd Quartile is 75th Percentile

Interquartile Range (IQR)

Interquartile Range (IQR)

Interquartile Range (IQR) is the difference between the lower and upper quartile. This gives us a better idea of the range of data.





Standard Variance and Standard Deviation

Standard Variance and Standard Deviation

Standard Variance measures how far a set of numbers are spread out from their average value.

Standard Deviation is used to express the magnitude by which the members of a group differ from the mean value for the group.

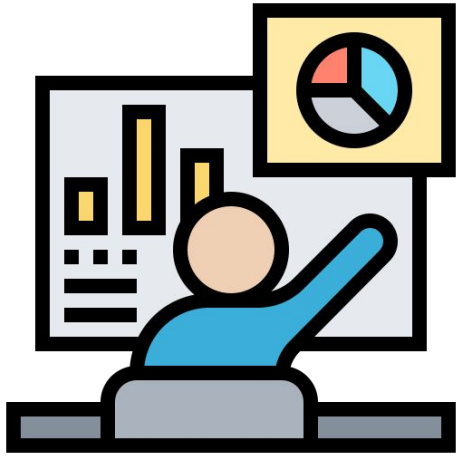
Standard Deviation is the square root of **Standard Variance.**

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$



Correlation

Correlation

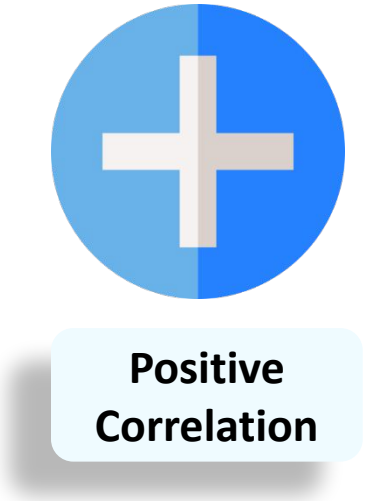
Correlation is a term that is a measure of the strength of a **linear relationship** between **two quantitative variables**

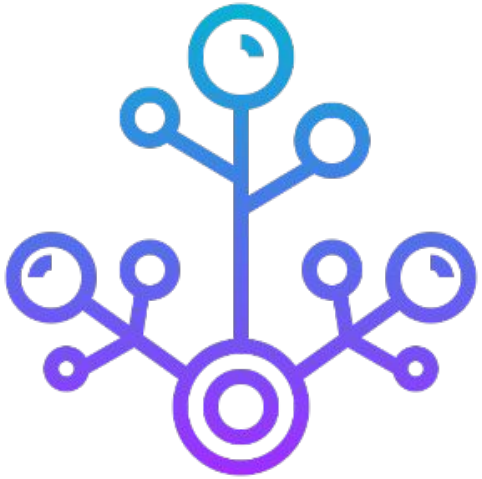
$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Positive Correlation

Positive Correlation

Positive Correlation is a term that is used to describe a **positive linear relationship** between **two quantitative variables**

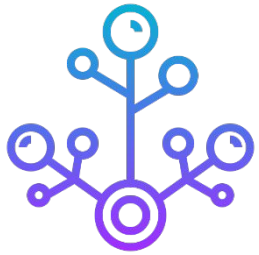




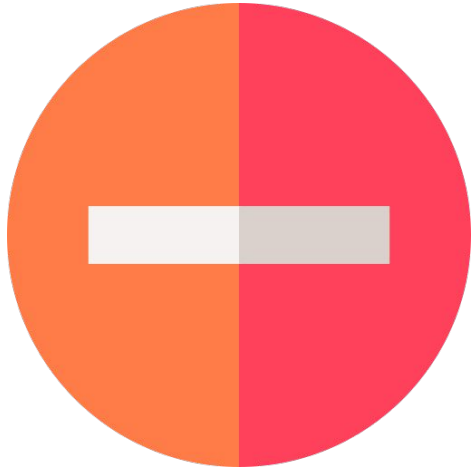
No Correlation

No Correlation

No Correlation is a term used to describe no **linear relationship** between **two quantitative variables**



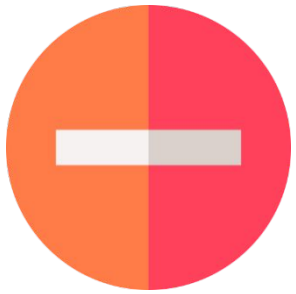
No Correlation



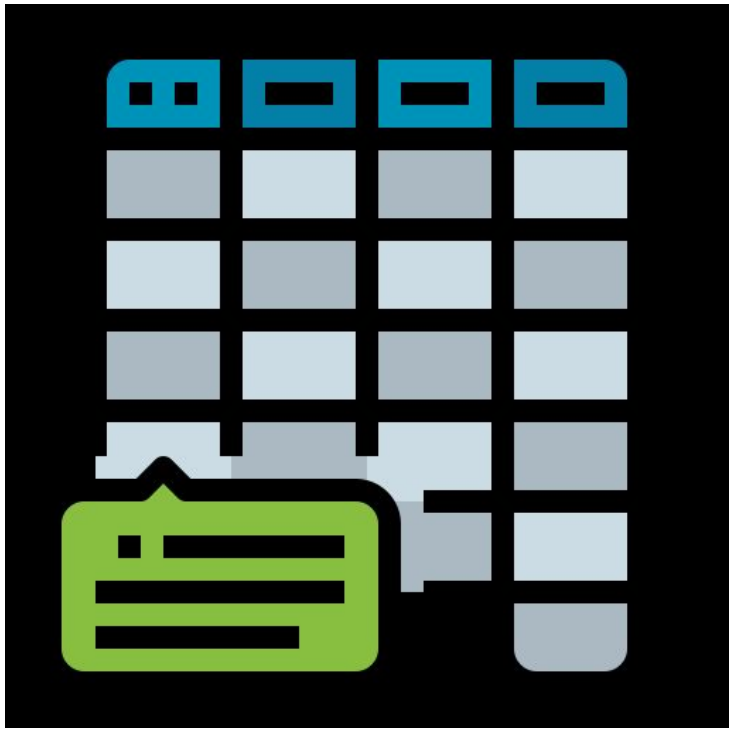
Negative Correlation

Negative Correlation

Negative Correlation is a term that is used to describe the strength of a **Negative linear relationship** between **two quantitative variables**



Negative Correlation



Tables

Tables

A way of presenting statistical data through a systematic arrangement of the numbers describing some mass phenomenon or process

A statistical table may be regarded as representing a subject and predicate. The meaning of each number is indicated by the headings of the corresponding row and column.

Charts

Charts

A statistical graph or chart is defined as the pictorial representation of statistical data in graphical form. The statistical graphs are used to represent a set of data to make it easier to understand and interpret statistical information.

Charts

A statistical graph or chart is defined as the pictorial representation of statistical data in graphical form. The statistical graphs are used to represent a set of data to make it easier to understand and interpret statistical information.

Lets list down the types of charts

Charts

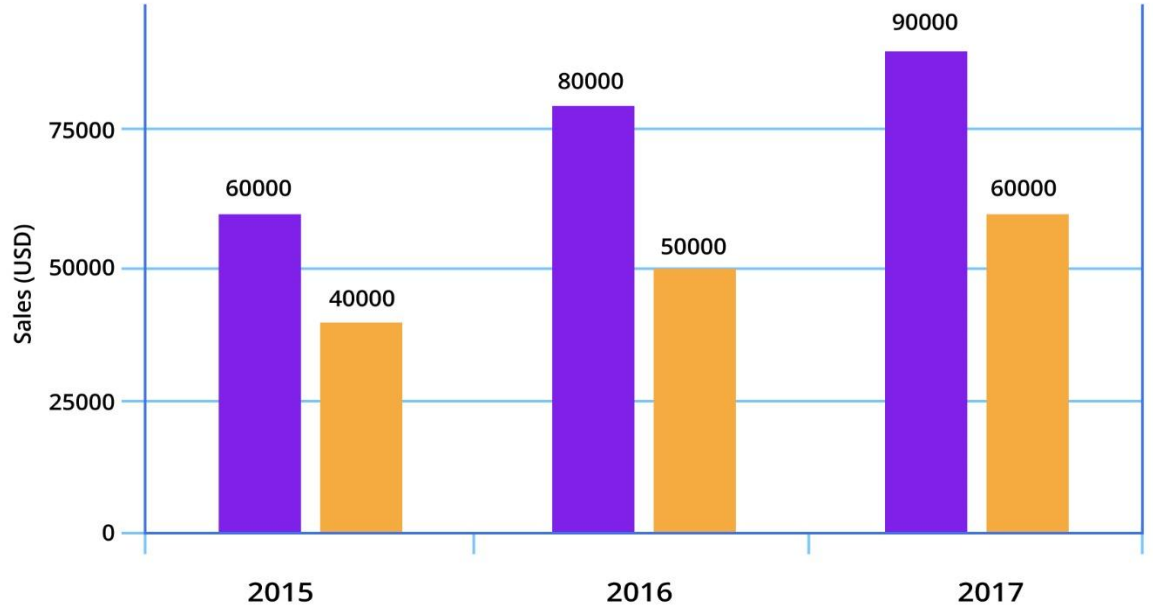
A statistical graph or chart is defined as the pictorial representation of statistical data in graphical form. The statistical graphs are used to represent a set of data to make it easier to understand and interpret statistical information.

Lets list down the types of charts

Types Of Charts
1.Bar chart
2.Histogram
3.Pie chart
4.Box chart
5.Line Graph
6.Area plot
7.Scatter plot

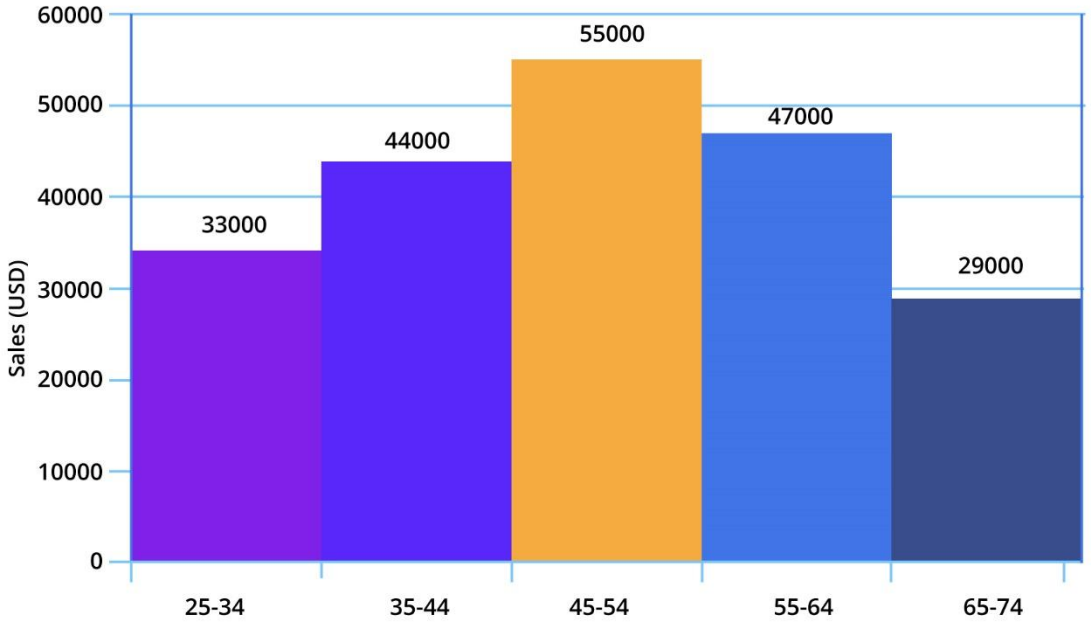
1. Bar chart

Bar charts are among the most frequently used chart types. As the name suggests a bar chart is composed of a series of bars illustrating a variable's development. Given that bar charts are such a common chart type, people are generally familiar with them and can understand them easily



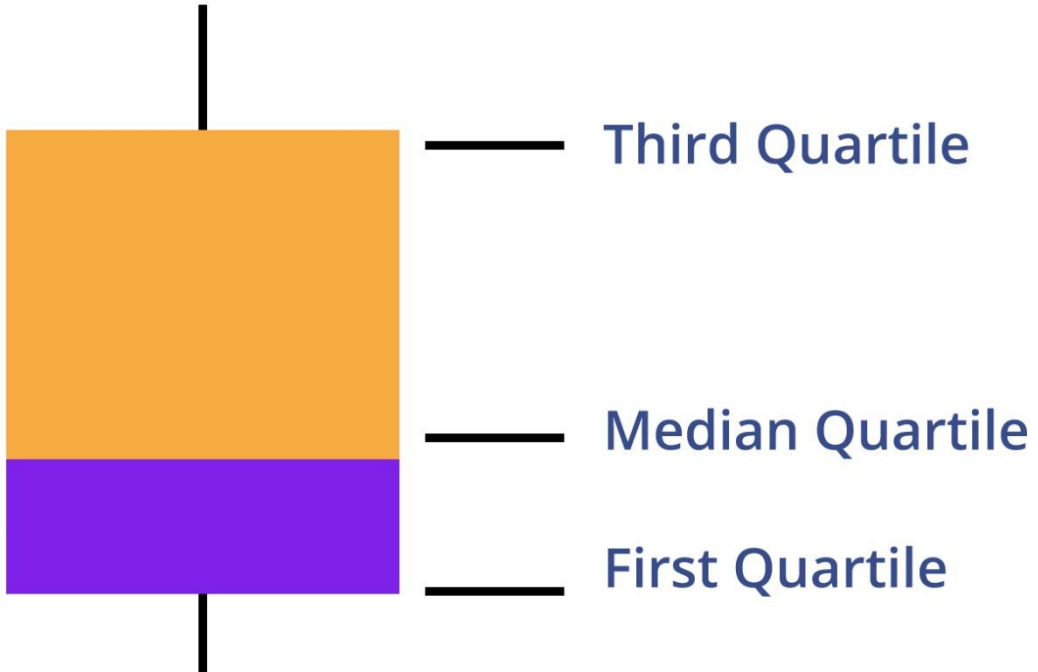
2. Histogram

A series of bins showing us the frequency of observations of a given variable. The definition of histogram charts is short and easy..



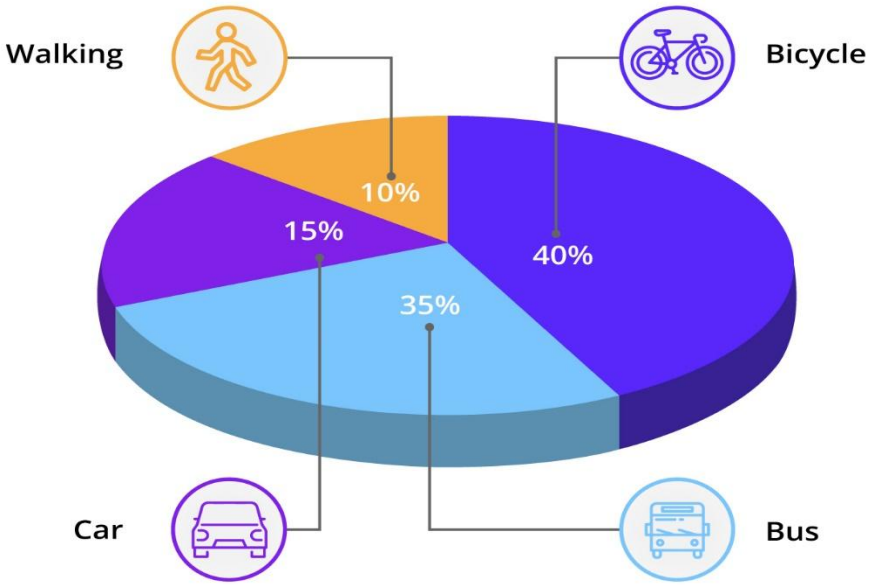
3. Box chart

Box plot, also called the box-and-whisker plot: a way to show the distribution of values based on the five-number summary: minimum, first quartile, median, third quartile, and maximum.



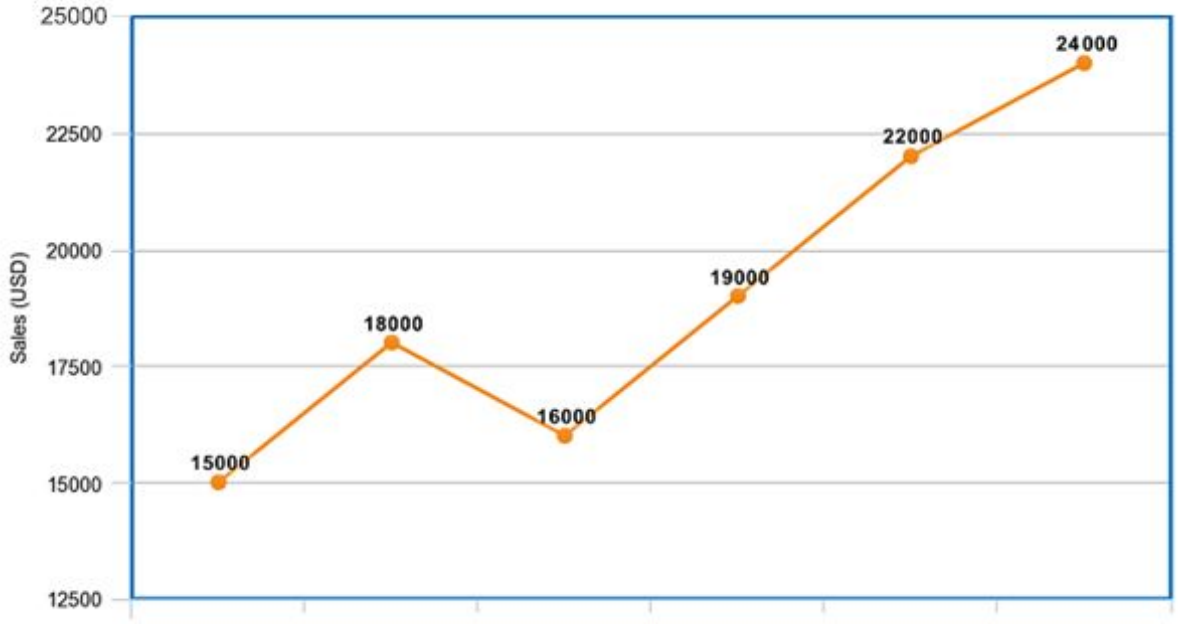
4. Pie chart

A pie chart is a circular graph divided into slices. The larger a slice is the bigger portion of the total quantity it represents.



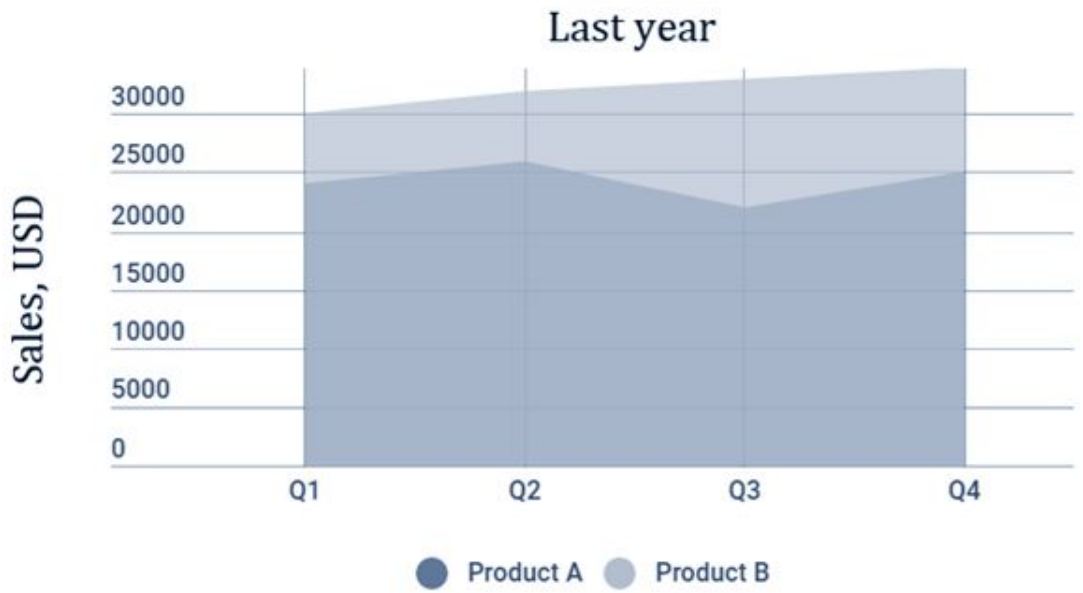
5. Line chart

A line chart is, as one can imagine, a line or multiple lines showing how single, or multiple variables develop over time. It is a great tool because we can easily highlight the magnitude of change of one or more variables over a period.



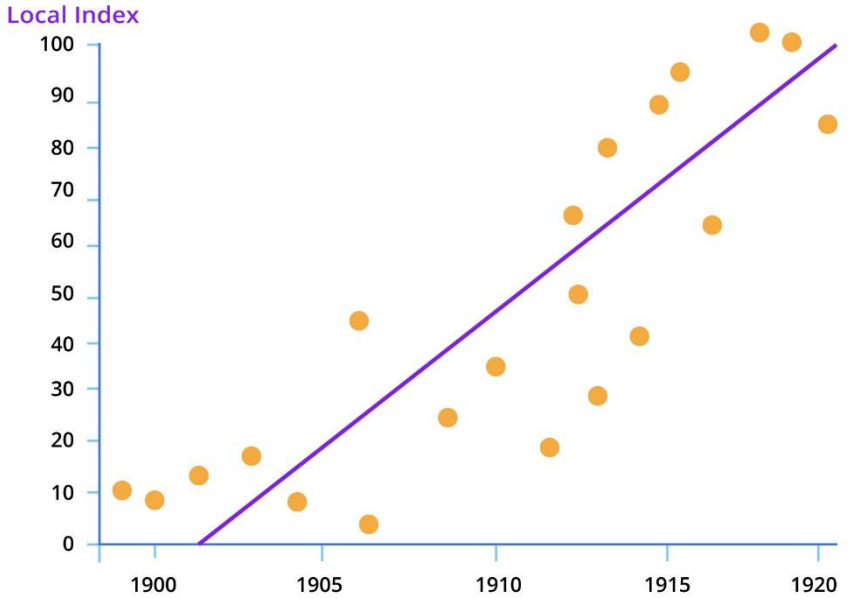
6. Area chart

Area charts are very similar to line charts. In fact, at first, I wanted to show them together. However, one major confusion could have arisen. So, please pay attention. The idea of an area chart is based on the line chart. Colored regions (areas) show us the development of each variable over time.



7. Scatter plot

A scatter plot is a type of chart that is often used in the fields of statistics and data science. It consists of multiple data points plotted across two axes. Each variable depicted in a scatter plot would have multiple observations. If a scatter plot includes more than two variables, then we would use different colors to signify that.



Probability

Introduction to Probability

Probability defines the likelihood of occurrence of an event

Introduction to Probability

Probability defines the likelihood of occurrence of an event

Probability can be defined as the ratio of the number of favorable outcomes to the total number of outcomes of an event.

Introduction to Probability

Probability defines the likelihood of occurrence of an event

Probability can be defined as the ratio of the number of favorable outcomes to the total number of outcomes of an event.

Probability can be defined as the ratio of the number of favorable outcomes to the total number of outcomes of an event.

Probability

For an experiment having 'n' number of outcomes, the number of favorable outcomes can be denoted by x. The formula to calculate the probability of an event is as follows.

$$\text{Probability(Event)} = \frac{\text{Favorable Outcomes}}{\text{Total Outcomes}}$$

Business Analytics

Probability In Business Analytics

One practical use for probability distributions and scenario analysis in business is to predict future levels of requirements that could boost the economy of the company.

Probability In Business Analytics

One practical use for probability distributions and scenario analysis in business is to predict future levels of requirements that could boost the economy of the company.

Using a scenario analysis based on a probability distribution can help a company frame its possible future values in terms of a likely sales level and a worst-case and best-case scenario.

Probability

Probability Distribution

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

Probability Distribution

A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on a number of factors.

Binomial Distribution

A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice)

Binomial Distribution

A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice)

For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

Poisson Distribution

A Poisson distribution is a tool that helps to predict the probability of certain events happening when you know how often the event has occurred. It gives us the probability of a given number of events happening in a fixed interval of time.

Poisson Distribution

A Poisson distribution is a tool that helps to predict the probability of certain events happening when you know how often the event has occurred. It gives us the probability of a given number of events happening in a fixed interval of time.

A textbook store rents an average of 200 books every Saturday night. Using this data, you can predict the probability that more books will sell (perhaps 300 or 400) on the following Saturday nights

Normal Distribution

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.
The area under the normal distribution curve represents probability and the total area under the curve sums to one.

Normal Distribution

The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.
The area under the normal distribution curve represents probability and the total area under the curve sums to one.

For example, if we randomly sampled 100 individuals we would expect to see a normal distribution frequency curve for many continuous variables, such as IQ, height, weight and blood pressure.

Hypothesis Testing and Estimation

Hypothesis testing refers to the process of making inferences or educated guesses about a particular parameter. This can either be done using statistics and sample data, or it can be done on the basis of an uncontrolled observational study.

Hypothesis Testing and Estimation

Hypothesis testing refers to the process of making inferences or educated guesses about a particular parameter. This can either be done using statistics and sample data, or it can be done on the basis of an uncontrolled observational study.

Estimation, in statistics, any of numerous procedures used to calculate the value of some property of a population from observations of a sample drawn from the population.

Goodness of Fit

The goodness-of-fit test is a statistical hypothesis test to see how well sample data fit a distribution from a population.
his test shows if your sample data represents the data you would expect to find in the actual population

Goodness of Fit

The goodness-of-fit test is a statistical hypothesis test to see how well sample data fit a distribution from a population.
his test shows if your sample data represents the data you would expect to find in the actual population

Goodness-of-fit establishes the discrepancy between the observed values and those that would be expected of the model in a normal distribution case.

India: +91-7847955955

US: 1-800-216-8930 (TOLL FREE)

sales@intellipaat.com

24/7 Chat with Our Course Advisor