What is Data Science?

- Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.
- It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.
- Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems.
- It is the future of artificial intelligence.

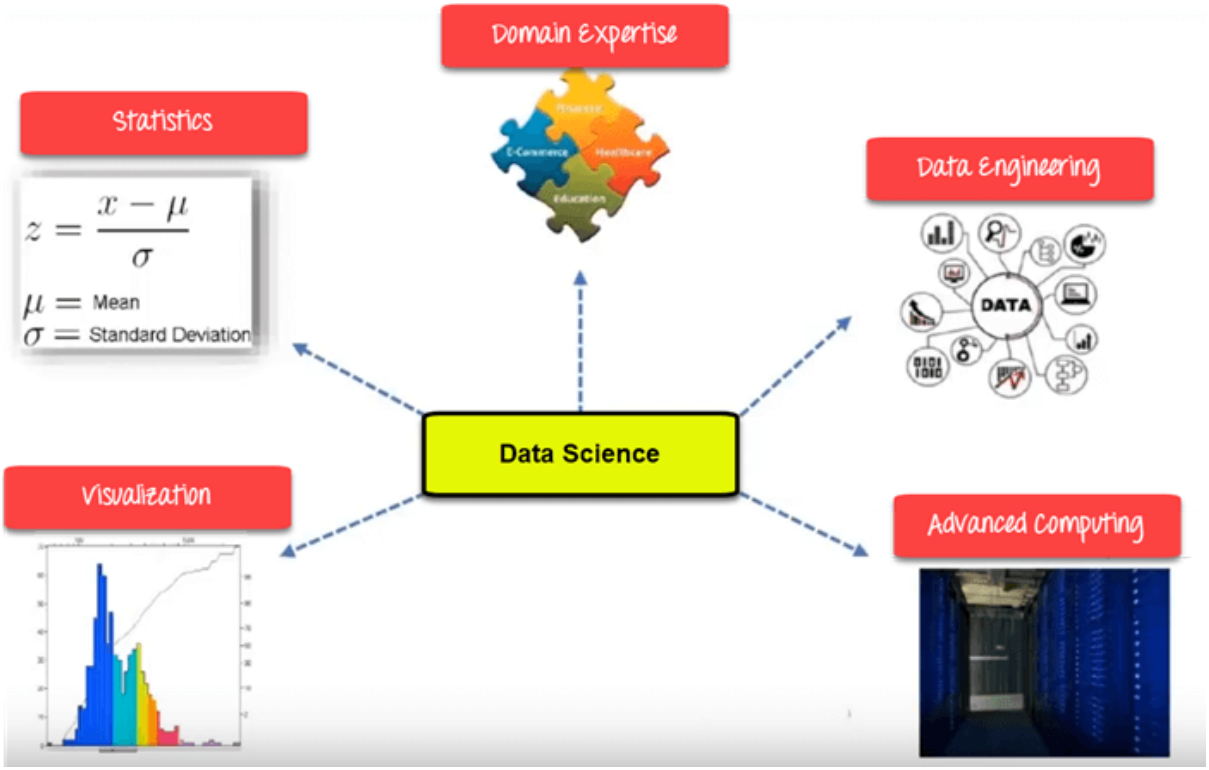
In short, we can say that data science is all about:

1. Asking the correct questions and analyzing the raw data.
2. Modeling the data using various complex and efficient algorithms.
3. Visualizing the data to get a better perspective.
4. Understanding the data to make better decisions and finding the final result.

Example:

- Let suppose we want to travel from station A to station B by car.
- Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective.
- All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Data Science Components



Statistics:

Statistics is the most critical unit of Data Science basics, and it is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.

Visualization:

Visualization technique helps you access huge amounts of data in easy to understand and digestible visuals.

Machine Learning:

Machine Learning explores the building and study of algorithms that learn to make predictions about unforeseen/future data.

Deep Learning:

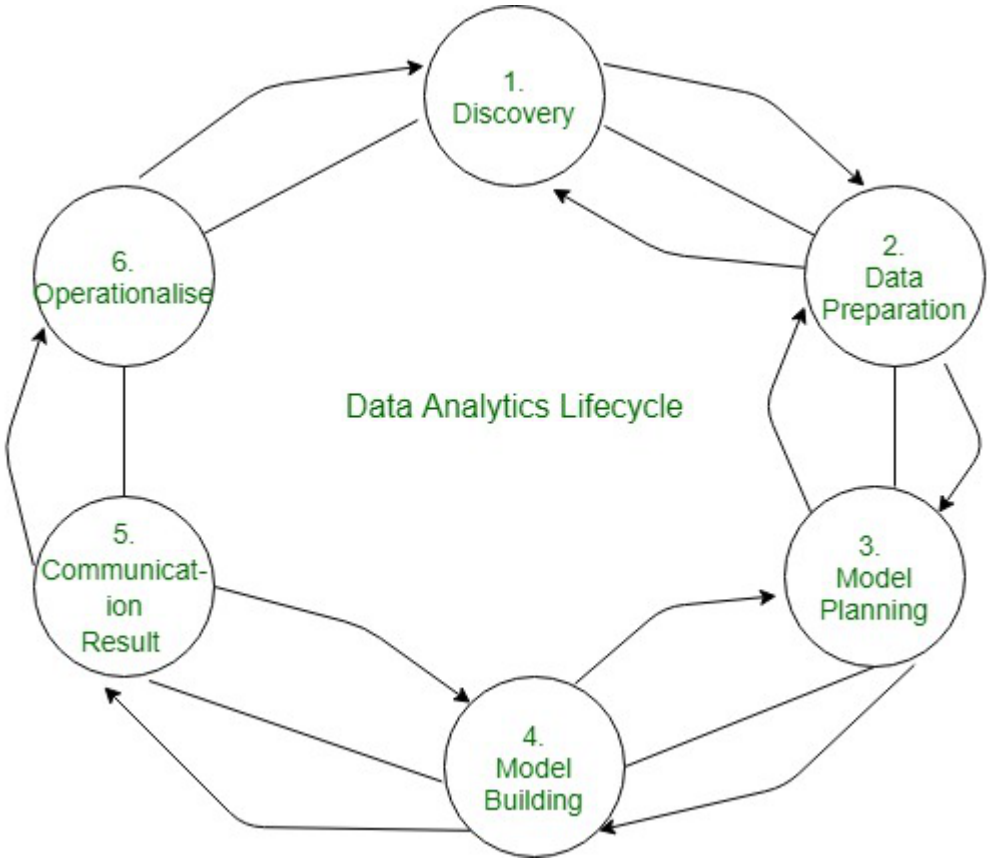
Deep Learning method is new machine learning research where the algorithm selects the analysis model to follow.

What is Data Analytics?

Big data analytics is the often complex process of examining big data to uncover information -- such as hidden patterns, correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

Data Analytics Lifecycle :

The Data analytic lifecycle is designed for Big Data problems and data science projects. The cycle is iterative to represent real project. To address the distinct requirements for performing analysis on Big Data, step – by – step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing, and repurposing data.



Phase 1: Discovery –

- The data science team learn and investigate the problem.
- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates initial hypothesis that can be later tested with data.

Phase 2: Data Preparation –

- Steps to explore, preprocess, and condition data prior to modeling and analysis.
- It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.
- Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning –

- Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- In this phase, data science team develop data sets for training, testing, and production purposes.
- Team builds and executes models based on the work done in the model planning phase.
- Several tools commonly used for this phase are – Matlab, STASTICA.

Phase 4: Model Building –

- Team develops datasets for testing, training, and production purposes.
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA.
- Commercial tools – Matlab , STASTICA.

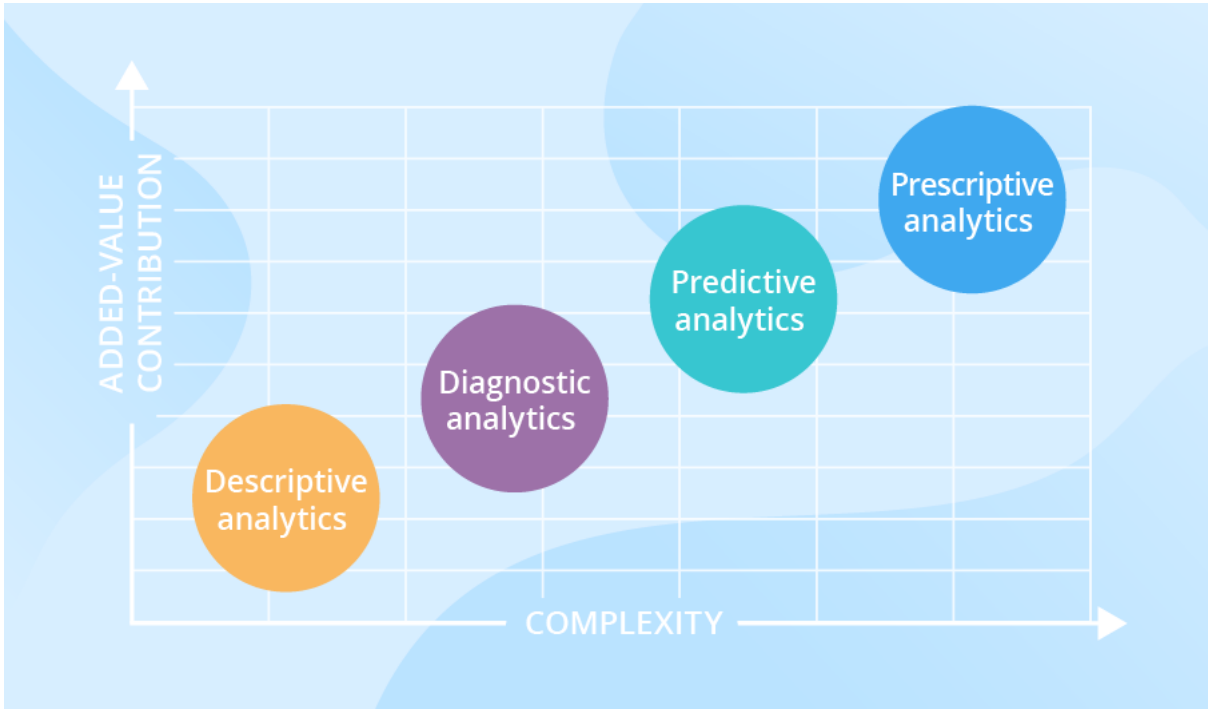
Phase 5: Communication Results –

- After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

Phase 6: Operationalize –

- The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes.
- Free or open source tools – Octave, WEKA, SQL, MADlib.

Types of Data Analytics



1)Descriptive analytics

- Descriptive analytics answers the question of ***what happened.***
- Descriptive analytics juggles raw data from multiple data sources to give valuable insights into the past.
- However, these findings simply signal that something is wrong or right, without explaining why.
- Ex-Performance of abc Shmpoo in year 2022.

2)Diagnostic analytics

- **Diagnostic** analytics tells ***why something happened.***
- Diagnostics analytics helps companies understand why a problem occurred.
- Big data technologies and tools allow users to mine and recover data that helps dissect an issue and prevent it from happening in the future.
- Example: An online retailer's sales have decreased even though customers continue to add items to their shopping carts. Diagnostics analytics helped to understand that the payment page was not working correctly for a few weeks

3)Predictive analytics

- **"What will happen in the future ?"(Forcasting);**
- Predictive analytics looks at past and present data to make predictions.
- With artificial intelligence (AI), machine learning, and data mining, users can analyse the data to predict market trends.
- Example Ex-Performance of abc Shmpoo in year 2023.

4)Prescriptive analytics

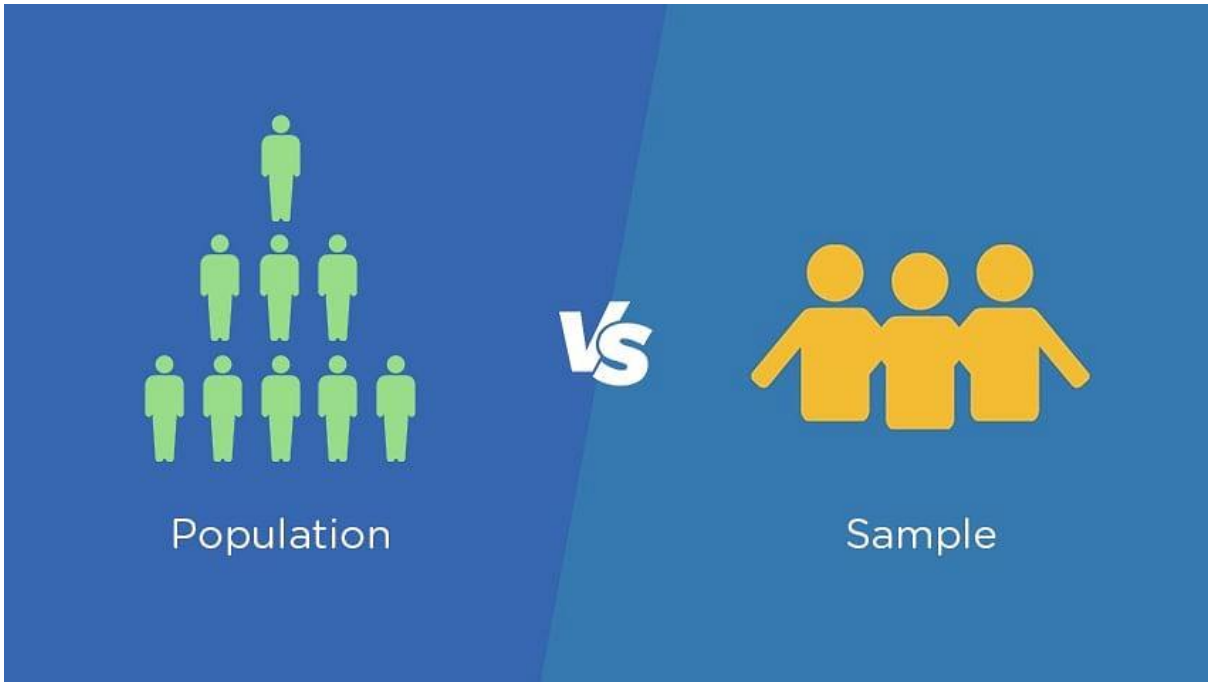
(Predictive+Prescriptive)

- The purpose of prescriptive analytics is to literally prescribe *what action to take* to eliminate a future problem or take full advantage of a promising trend.

- Suggest action and possible result.

- Prescriptive analytics uses advanced tools and technologies, like machine learning, business rules and algorithms, which makes it sophisticated to implement and manage.
- Ex. Self Driving Car

Populations and samples



Population:

- A complete collection of the objects or measurements is called the population or else everything in the group we want to learn about will be termed as population.
- In statistics population is the entire set of items from which data is drawn in the statistical study.
- It can be a group of individuals or a set of items.
- For example, All people living in India indicates the population of India.

There are different types of population. They are:

1. Finite Population
2. Infinite Population
3. Existent Population
4. Hypothetical Population

Let us discuss all the types one by one.

Finite Population

The finite population is also known as a countable population in which the population can be counted.

Examples of finite populations are employees of a company

Infinite Population

The infinite population is also known as an uncountable population in which the counting of units in the population is not possible.

Example of an infinite population is the number of germs in the patient's body is uncountable.

Existent Population

The population whose unit is available in solid form is known as existent population.

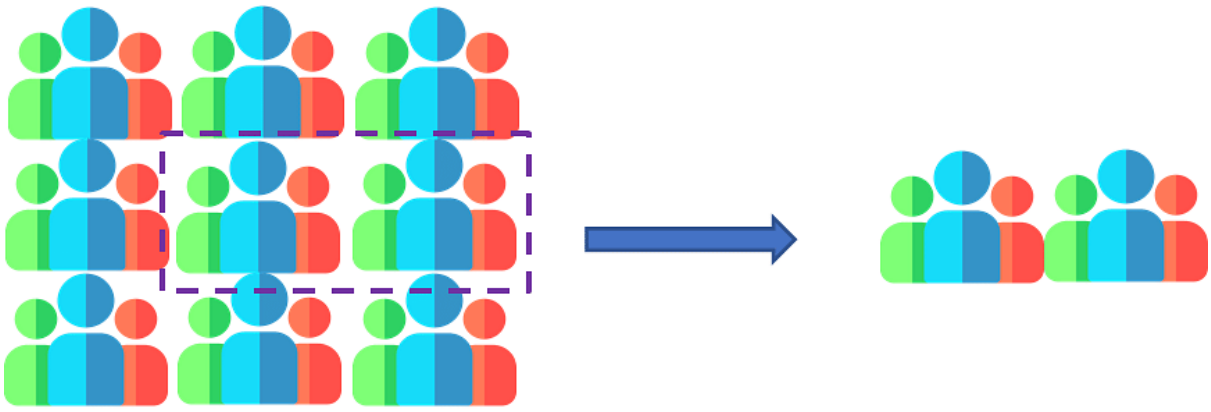
Examples are books, students etc.

Hypothetical Population

The population in which whose unit is not available in solid form is known as the hypothetical population.

Examples are an outcome of rolling the dice, the outcome of tossing a coin.

Sample:



- It includes one or more observations that are drawn from the population and the measurable characteristic of a sample is a statistic.
- Sampling is the process of selecting the sample from the population.
- For example, some people living in India is the sample of the population.

Basically, there are two types of sampling. They are:

1)Probability sampling

2)Non-probability sampling

1) Probability Sampling :

- In probability sampling, the population units cannot be selected at the discretion of the researcher.
- This can be dealt with following certain procedures which will ensure that every unit of the population consists of one fixed probability being included in the sample.
- Such a method is also called random sampling.

Some of the techniques used for probability sampling are:

- Simple random sampling
- Cluster sampling
- Stratified Sampling
- Disproportionate sampling
- Proportionate sampling
- Optimum allocation stratified sampling
- Multi-stage sampling

2) Non Probability Sampling :

- In non-probability sampling, the population units can be selected at the discretion of the researcher.
- Those samples will use the human judgements for selecting units and has no theoretical basis for estimating the characteristics of the population.
- Some of the techniques used for non-probability sampling are
- Quota sampling
- Judgement sampling
- Purposive sampling

Advantages of sampling

1. Low cost of sampling

- If data were to be collected for the entire population, the cost will be quite high. A sample is a small proportion of a population. So, the cost will be lower if data is collected for a sample of population which is a big advantage.

2. Less time consuming in sampling

- Use of sampling takes less time also. It consumes less time than census technique. Tabulation, analysis etc., take much less time in the case of a sample than in the case of a population.

3. Accuracy of data is high

- A sample represents the population from which its is drawn. It permits a high degree of accuracy due to a limited area of operations

4. Organization of convenience

- Organizational problems involved in sampling are very few. Since sample is of a small size, vast facilities are not required. Study of samples involves less space and equipment.

What is Statistical Modeling and How is it Used?

Statistical modeling is the process of applying statistical analysis to a dataset.

A **statistical model** is a mathematical representation (or mathematical model) of observed data.

Phases in statistical modeling :

1) Define and Design

- Write out research questions in theoretical and operational terms.
- Design the study
- Write an analysis plan
- Calculate sample size estimations.

2)Prepare and Explore

- Collect, code, enter and clean data
- Create new variables.

3)Refine the Model

- Refine predictors and check model fit.
- Test assumptions.

- Check for and resolve data issues

4)Answer the research Question

- Interpret result
- Write up result

1)Probability :

Probability means possibility. It is a branch of mathematics that deals with the occurrence of a random event. The value is expressed from zero to one. Probability has been introduced in Maths to predict how likely events are to happen.

Types of Probability :

1) **Classical:**

There are many possibilities which can equally happen without anything specifically expected to happen . "Whatever will be , will be".

There are 'n' number of event and we can find the probability of the happening of event by applying basic probability formulate.

Example: The probability of getting head in a single toss of a coin is ½ .This is Classical probability.

2**)Empirical** : It has expected theoretical outcome within a limited or controlled field. This type of probability is based on experiments.

3)Subjective : Subjective probability is an individual person measure of belief that an event will occur . It is vague and rarely accurate.

Applications of Probability

1. Weather Forecasting
2. Batting Average in Cricket
3. Politics
4. Flipping a coin or Dice

5. Lottery Tickets

6. Playing Cards

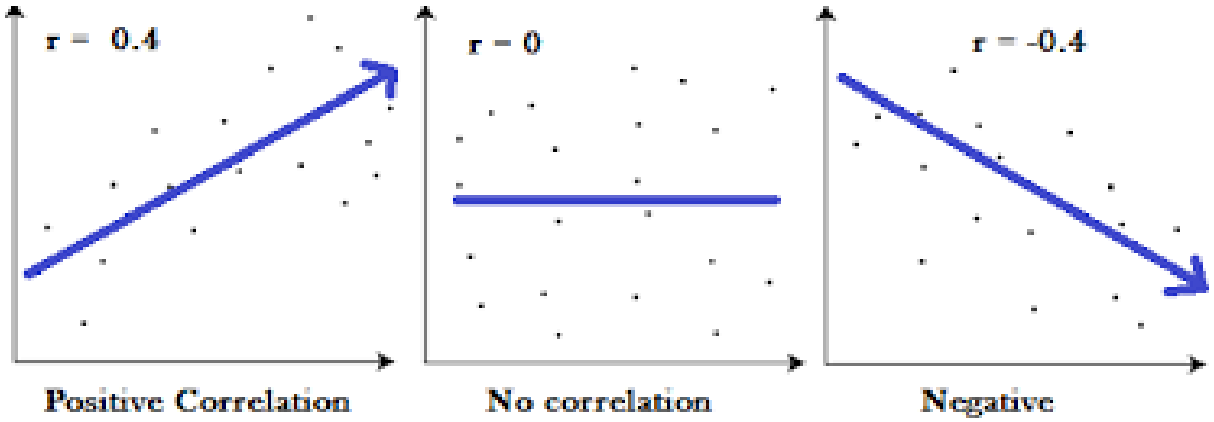
2)Correlation :

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related .

For example height and weight are related

Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –



1)Positive Correlation – when the values of the two variables move in the same direction so that an increase in the value of one variable is followed by an increase in the value of the other variable.

Examples of Positive Correlation:

1)The more money you make , the more taxes the government takes out of your check.

2)The more education you receive , the smarter you all be.

2)Negative Correlation – when the values of the two variables move in the opposite direction so that an decrease in the value of one variable is followed by decrease in the value of the other variable.

Examples of Negative Correlation:

1)The more you work in the office , the less time you'll spend at home.

3)No Correlation – when there is no linear dependence or no relation between the two variables.

Examples of No Correlation :

1)The smarter you are , the later you'll arrive at work.

2)The wealthier you are , the happier you'll be.

3) Regression

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.

It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

Types of Regression

1)Linear regression :

- It is used for predictive analysis. It is the simplest form of regression.
- It is a technique in which the dependent variable is continuous in nature.
- The relationship between the dependent variable and independent variable is assumed to be linear in nature.
- *The below-given equation is used to denote the linear regression model:*
- y=mx+c+e
- where m is the slope of the line, c is an intercept, and e represents the error in the model.

2)Logistic regression:

- Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete. Example: 0 or 1, true or false, etc.
- This means the target variable can have only two values.
- Logit function is used in Logistic Regression to measure the relationship between the target variable and independent variables.

- Below is the equation that denotes the logistic regression.
- logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3….+bkXk
- where p is the probability of occurrence of the feature.

3) Polynomial Regression :

- It is a type of Regression analysis that models the relationship of values of the Dependent variable "x" and Independent variables "y'' as non-linear.
- It is a special case of Multiple Linear Regression even though it fits a non-linear model to data.
- It is because data may be correlated but the relationship between two variables might not look linear.

4) Ridge regression :

- It is a technique for analyzing multiple regression data .when multicollinearity occurs , least squares estimates are unbiased

Application:

1) It may be used to compare a company's financial performance to that of a certain counterpart .

2) It can be used in the medical field to understand the relationships between drug dosage and blood pressure of the patients

3) Agriculture scientists frequently use Linear regression to see the impact of rainfall and fertilizer on the number

4)Distribution :

The distribution of a variable is a description of the relative number of times each possible outcome will occur in a number of trials .

The distribution of a statistical data set is a listing or function showing all the possible values of the data and how often occur.

Types of distribution :

1) **The Normal Distribution :**

It is often the case with medical data that histogram of a continuous variable obtained from a single measurement on different subject will have a characteristics 'bell-shaped' distribution known as a normal distribution

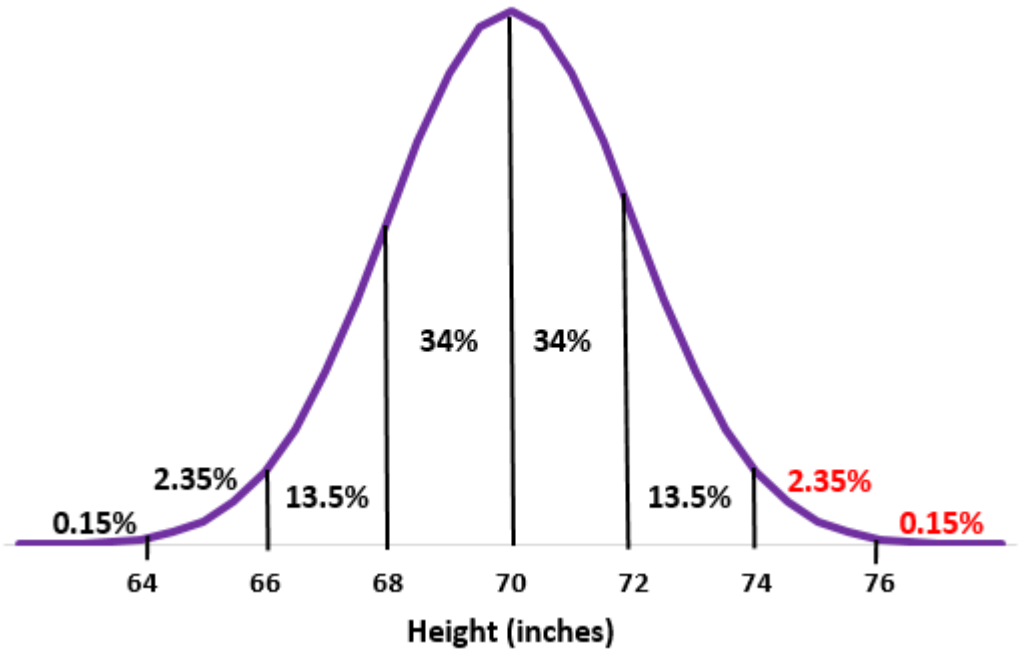
The **Normal Distribution** of random variable X is given by:

$$n(x; \mu, \sigma) = \{1/(\sqrt{2\pi})\sigma\}e(-1/2\sigma^2)(x-\mu)^2 \text{ for } -\infty<x<\infty$$

where

- μ *is mean*
- σ *is variance*

- **Examples**
- Salary of Working Class
- Heights of Male or Female
- The IQ Level of children



2)The Binomial Distribution :

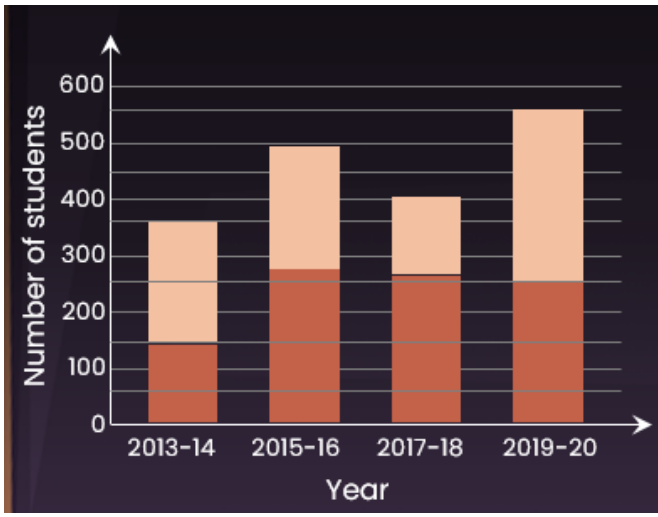
A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.

If Y is a Binomial Random Variable, where p is the probability of success in a given trial, q is the probability of failure, Let 'n' be the total number of trials, and 'x' be the number of successes, the Probability Function P(Y) for Binomial Distribution is given as:

$$P(Y) = nCx\ qn–xpx$$

where $x = 0,1,2…n$

- **Examples**
- To find the number of good and defective items produced by a factory.
- To find the number of girls and boys studying in a school.
- To find out the negative or positive feedback on something



Year	Number of girls	Number of boys	Total
2013-2014	150	200	350
2015-2016	280	220	500
2017-2018	260	140	400
2019-2020	250	300	550

3)The Poisson distribution :

In statistics a Poisson distribution is a statistical distribution that shows how many times an events is likely to occur within a specified period of time .

It is used for independent event which occur at a constant rate within a given interval of time .

Examples:

1. How many black colours are there in a random sample of 50 cars
2. No of cars arriving at a car wash during a 20 minute time interval

$$f(x; \lambda) = P(X=x) = (\lambda x e-\lambda)/x!$$

where,

- x *is number of times event occurred*

- $e = 2.718\ldots$
- λ *is mean value*

	A	B
1	**λ (mean number of successes)**	2.5
2		
3	**k (number of successes)**	
4	0	0.082085
5	1	0.205212
6	2	0.256516
7	3	0.213763
8	4	0.133602
9	5	0.066801
10	6	0.027834
11	7	0.009941
12	8	0.003106
13	9	0.000863
14	10	0.000216