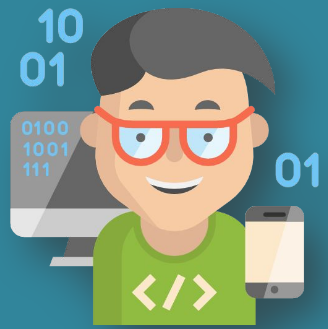


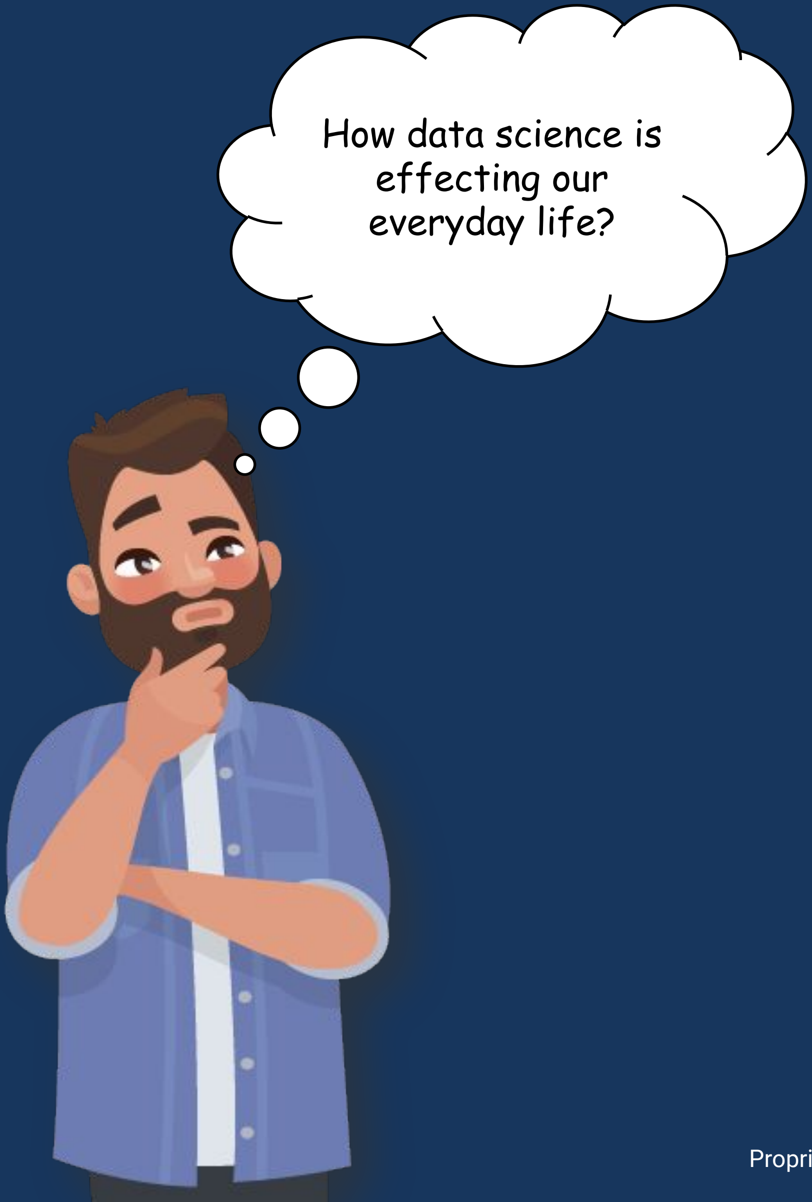
Top 10 Indian food analysis

Sampriti Chatterjee (Great Learning)

Agenda

1. Why do we need data science?
2. What is Data science?
3. Life cycle of Data science
4. Important statistics terms in data science
5. Install python
6. Demo: EDA on top 10 Indian food items

Why do we need Data Science?

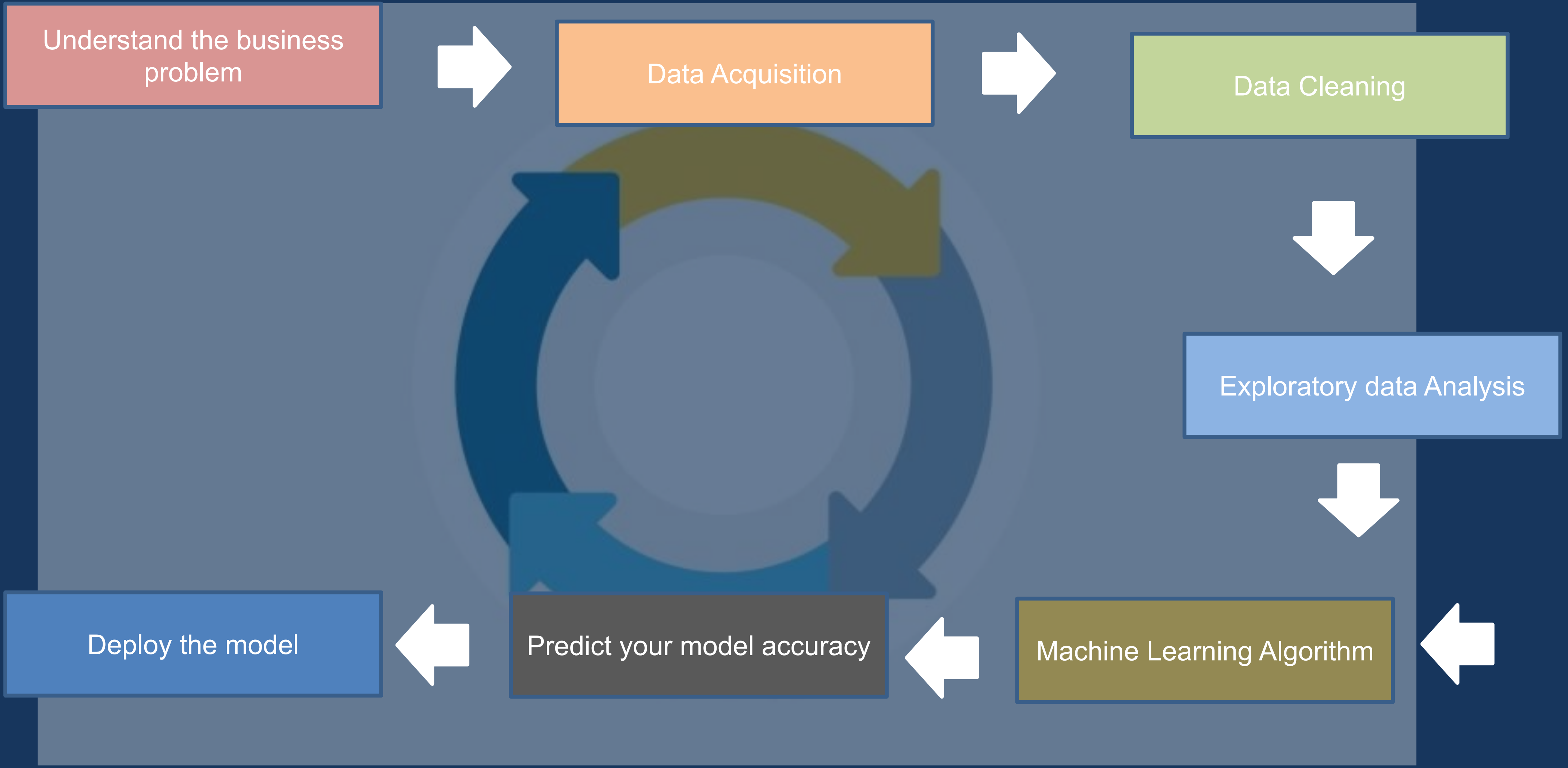


- In the past, we used to have data in a structured format but now as the volume of the data is increasing, so the number of structured data becomes very less, so to handle the massive amount of data we need data science techniques
- Those data can be used to get the proper business insights and the hidden trends from them.
- These insights helps the organization to predict the Future
- Using data science decision making can be faster and effective
- Helps to reduce the production cost
- Build model based on the data to give the ability to the machine to predicts on its own

What is Data Science?

Data science is a process to get some meaningful information from the massive amount of data. In simple terms, read and study the data to get proper intuitive insights. Data Science is a mixture of various tools, algorithms, and machine learning and deep learning concepts to discover hidden patterns from the raw and unstructured data

Life cycle of Data Science?



Understand the business problem

Data Acquisition

Data Cleaning

Exploratory data Analysis

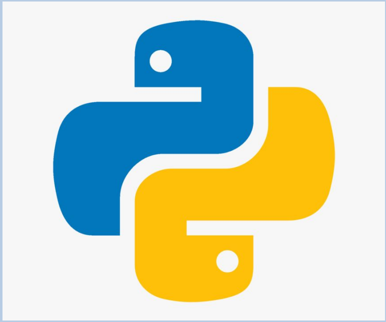
Machine Learning Algorithm

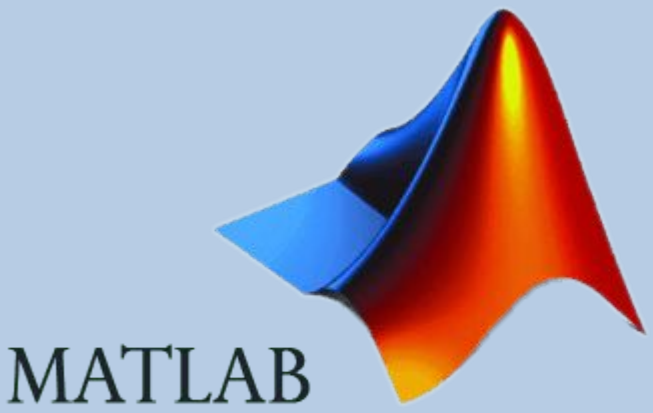
Predict your model accuracy

Deploy the model

Most Popular Programming Languages For Data Science?

Informatica™





Important statistics terms in data science

1 What is Statistics?

2 What is population?

3 What is parameter?

4 What is sample?

5 What is mean?

6 Types of analysis in statistics

7 What is Outlier?

8 What is Interquartile Range IQR?

9 What is upper and lower limits in interquartile range

10 What is null hypothesis?

11 What is p value?

What is Statistics?

Statistics is a part of integrated applied mathematics which deals with data

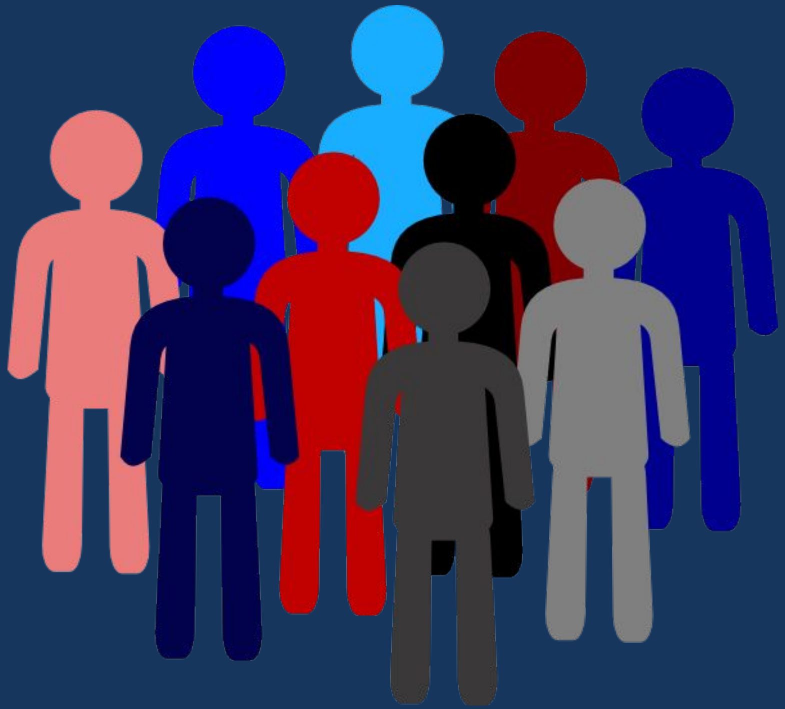
1. It helps to collect data and analyze them properly
2. With the help of statistics we can read the data and organize them in order to get the hidden information from them
3. In data science domain statistics concepts are used to process the complex data to get the insights from them using mathematical computations

What is Population?

Population terms in statistics use to refer the total set of observations

Example:

Suppose,
If we want to study a diabetes dataset to understand the symptoms and the other factors then the whole dataset is referred as population



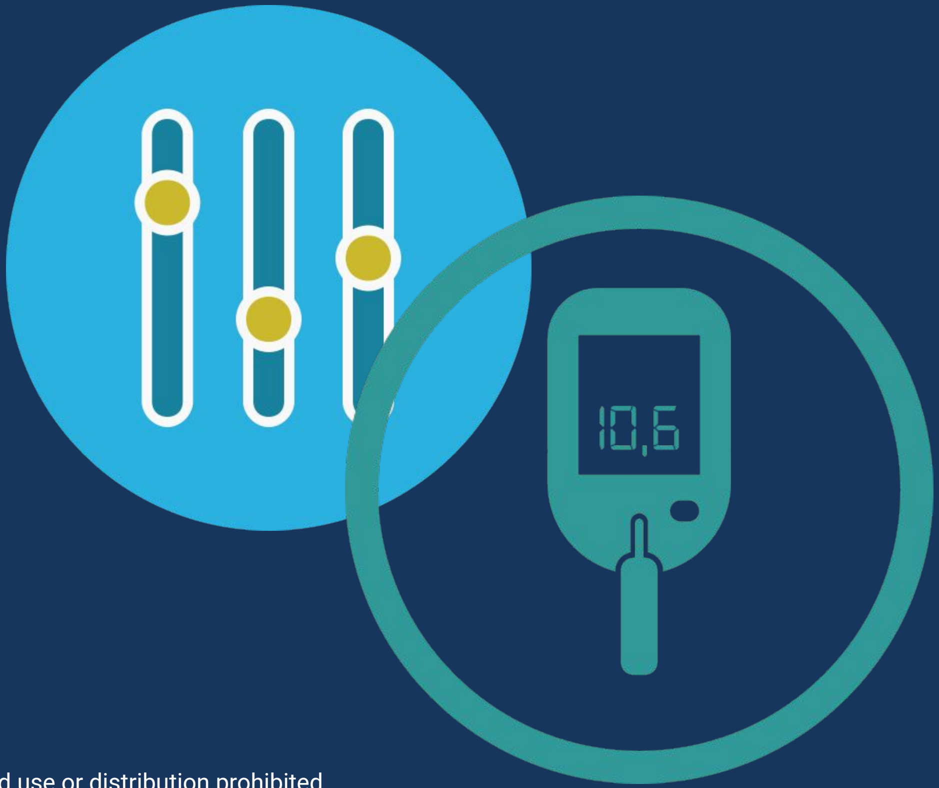
What is Parameter?

Parameters are referred to characteristics which describes the population

1 Parameters are like average or percentage which helps to describe the entire population

2 Mean and the standard deviation are two common parameters of population

3 Example: Average age for being diabetic is the parameter for whole diabetes dataset population

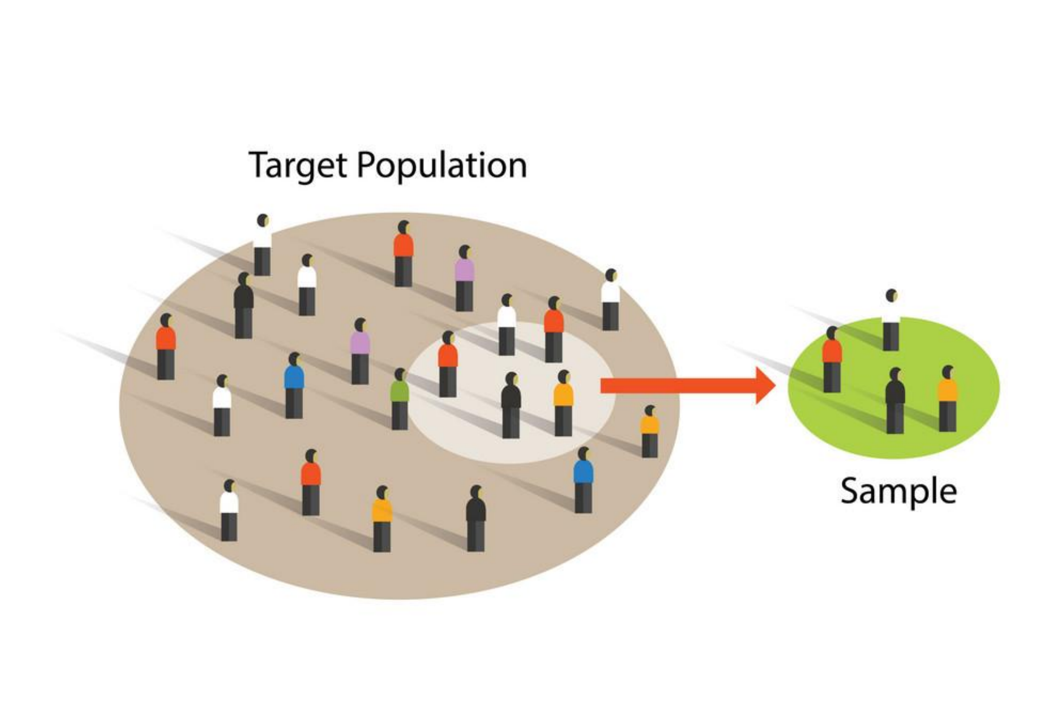


What is Sample?

Sample is basically a small part or portion of the large population

Example:

Suppose,
From the whole diabetes dataset you picked 100 rows of information to do the analysis, that 100 rows of information will be referred as **Sample**



Important statistics terms in data science

What is Mean?

Mean term referred as average value of the whole population

What is Median?

Median is the middle value of the data when your data is sorted in manner

What is Mode?

Mode stands for the most occurring element in the dataset

Types Of Analysis In Statistics

Descriptive statistics

Inferential Statistics

It helps to describe the data in mathematical or graphical way

Inferential statistics split the data into samples and applies probability to arrive to theconclusion

Important statistics terms in data science

Great Learning

What is Outlier?

Outliers in the dataset are referred as unusual value which can distort and violate statistical analysis

1 Outliers are basically experimental errors in the data

2 Some outliers are good for the dataset to detect anomaly like: detecting fraud transaction

3 It effects the mean and the standard deviation of the data and most of the machine learning technique does not perform good with outliers

What is Interquartile Range IQR?

Interquartile range divides the dataset into quartiles to measure the variability and the spread of the dataset

1. Splits the data into 4 equal part in sorted manner

2. Q1, Q2, Q3 are called first, second and third quartiles:
   - Q1 □ 25th percentile of the dataset
   - Q2□ 50th percentile of the dataset
   - Q3□ 75th percentile of the dataset

Formula: IQR□ Q3 – Q1

What is upper and lower limits in interquartile range

Lower and upper limit in the interquartile basically the range where data points lie

1 Formula to find the lower limit:

Lower_limit = Q1 - 1.5 IQR

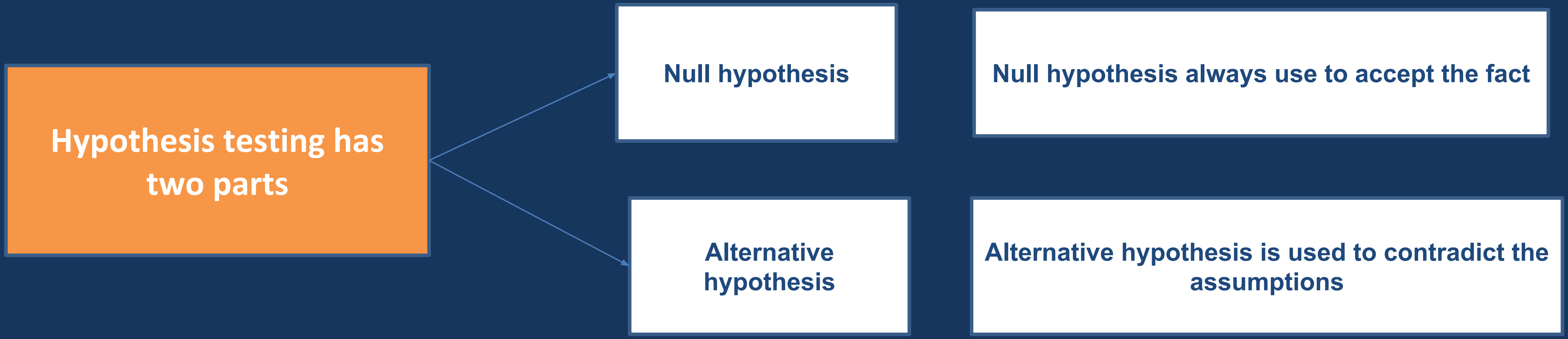
2 Formula to find the upper limit:

Upper_limit = Q3 + 1.5 * IQR

Important statistics terms in data science

What is Hypothesis testing?

Hypothesis testing is basically used to test the assumption which is taken based on observations and experiments



Hypothesis testing has two parts

Null hypothesis

Null hypothesis always use to accept the fact

Alternative hypothesis

Alternative hypothesis is used to contradict the assumptions

What is p value?

p value is used to support or reject the null hypothesis or the assumption

1. P value is basically the strong evidence to reject the null hypothesis

2. If p value is less than 0.05 then we accept the null hypothesis

EDA for Top 10 Indian food analysis

Thank You