

Exploratory Data Analysis



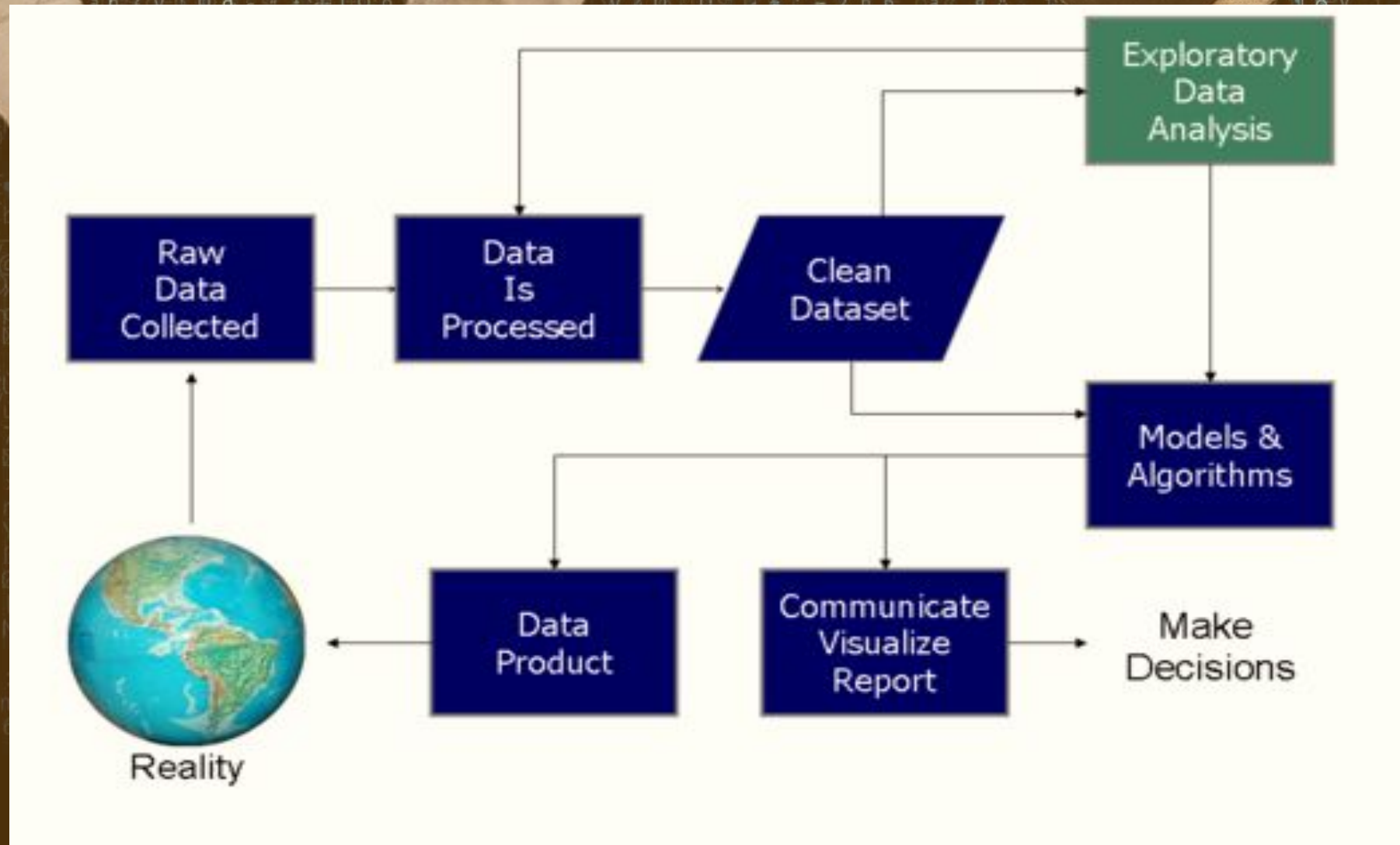
Beingdatum



Agenda

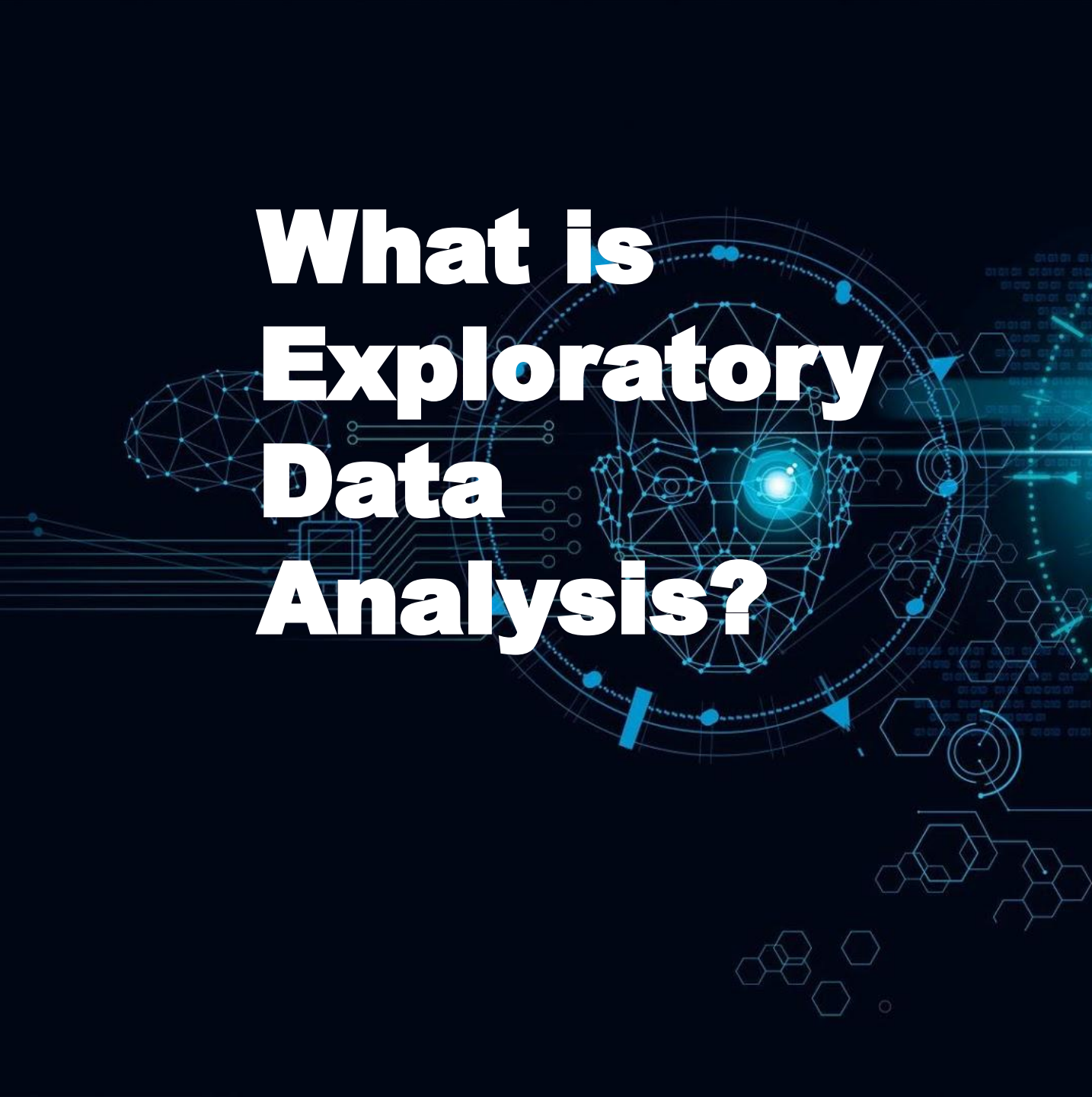
- 01 What is Exploratory Data Analysis?
- 02 Why EDA is important?
- 03 Visualisation
 - Important Charts for visualisation
- 04 Steps Involved in EDA:-
 - Data Sourcing
 - Data Cleaning
 - Univariate analysis with visualisation
 - Bivariate analysis with visualisation
 - Derived metrics
- 05 Uses Cases

Data Science Process



Exploratory Data Analysis

- ❑ Exploratory Data Analysis is an approach to analyse the datasets to summarize their main characteristics in form of visual methods.
- ❑ EDA is nothing but an data exploration technique to understand various aspects of the data.
- ❑ The main aim of EDA is to obtain confidence in a data to an extent where we are ready to engage a machine learning model.



What is Exploratory Data Analysis?

- ❖ EDA is important to analyse the data it's a first steps in data analysis process.
- ❖ EDA give a basic idea to understand the data and make sense of the data to figure out the question you need to ask and find out the best way to manipulate the dataset to get the answer of your question.
- ❖ Exploratory data analysis help us to finding the errors, discovering data, mapping out data structure, finding out anomalies.
- ❖ Exploratory data analysis is important for business process because we are preparing dataset for deep through analysis that will detect you business problem.
- ❖ EDA help to build a quick and dirty model, or a baseline model, which can serve as a comparison against later models that you will build.



Visualization

Visualisation is the presentation of the data in the graphical or visual form to understand the data more clearly. Visualisation is easy to understand the data.

Easily analyse the data and summarize it.

Easily understand the features of the data.

Help to find the trend or pattern of the data.

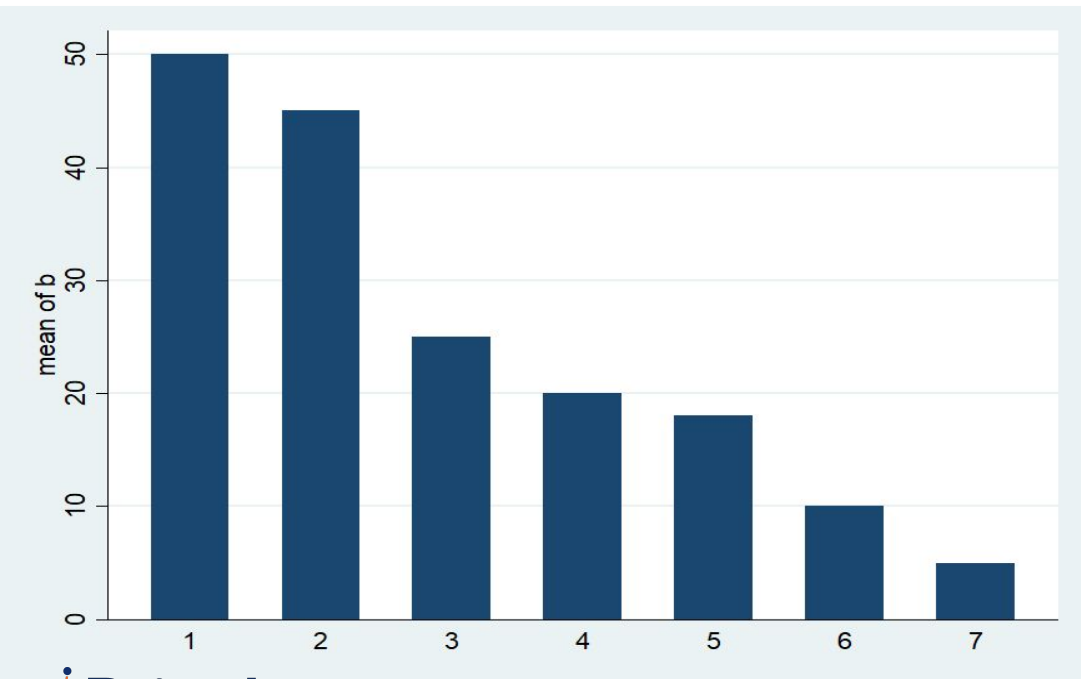
Help to get meaningful insights from the data.



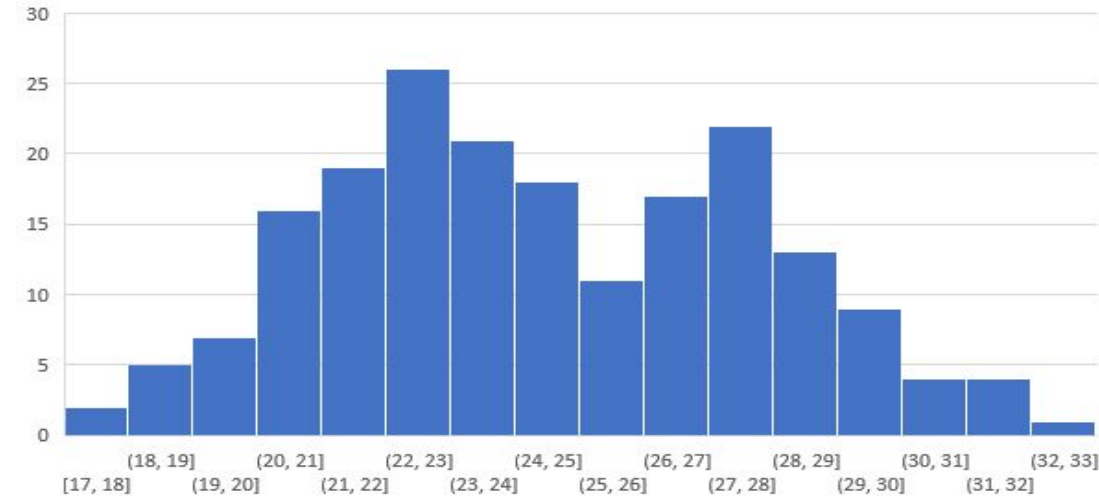
Important Charts for Visualisation

Histogram

Histogram represent the frequency distribution of the data



Age at first marriage - Females

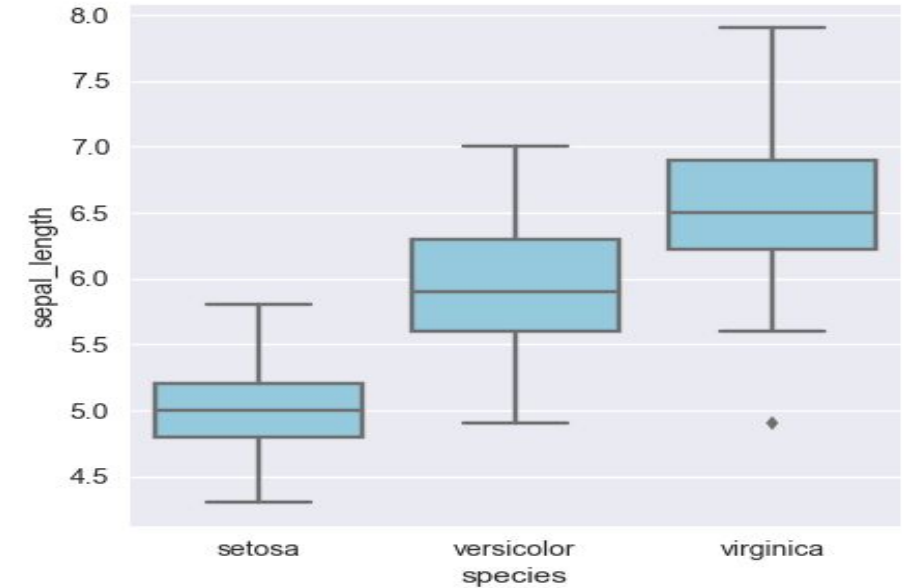


Bar Chart

Bar graph represent the total observation in the data for a particular category.

Box Plot

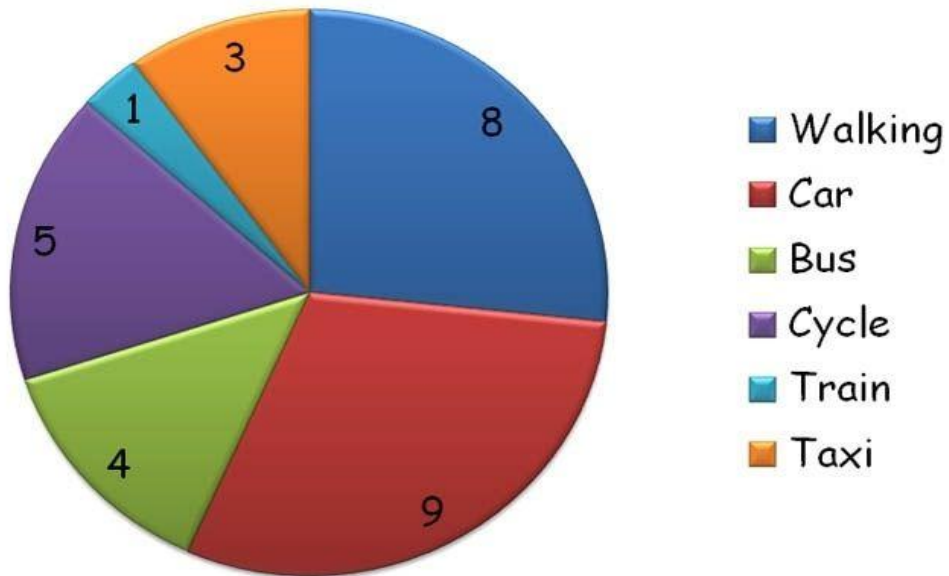
Boxplot display the distribution of the data based on five number summary(minimum, first quartile, median, third quartile, maximum)



Pie Chart

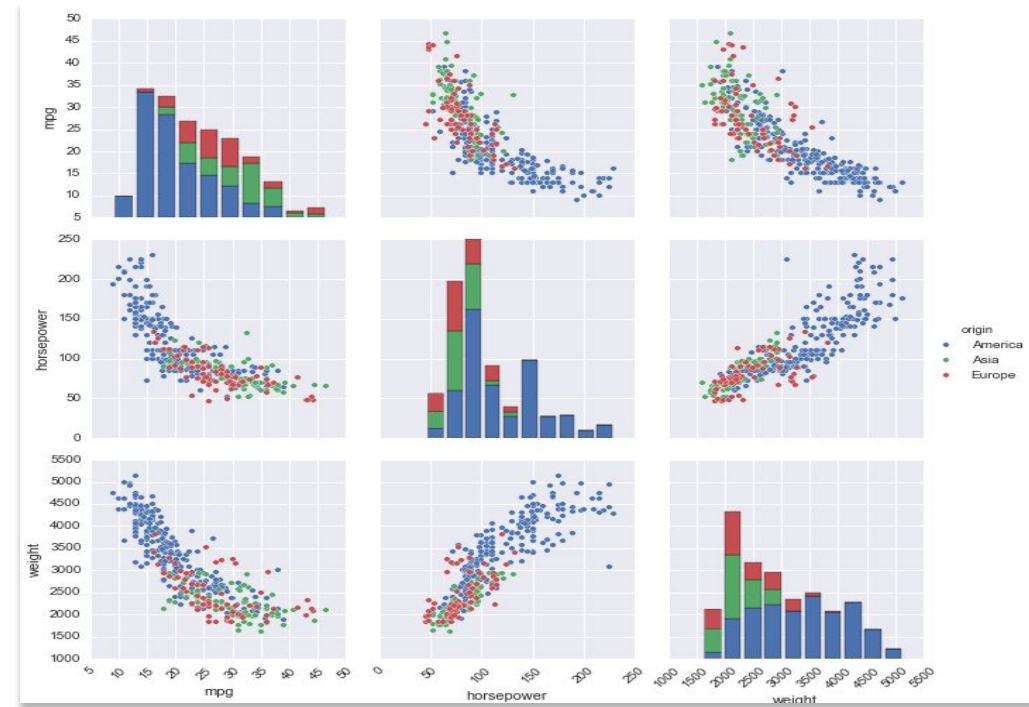
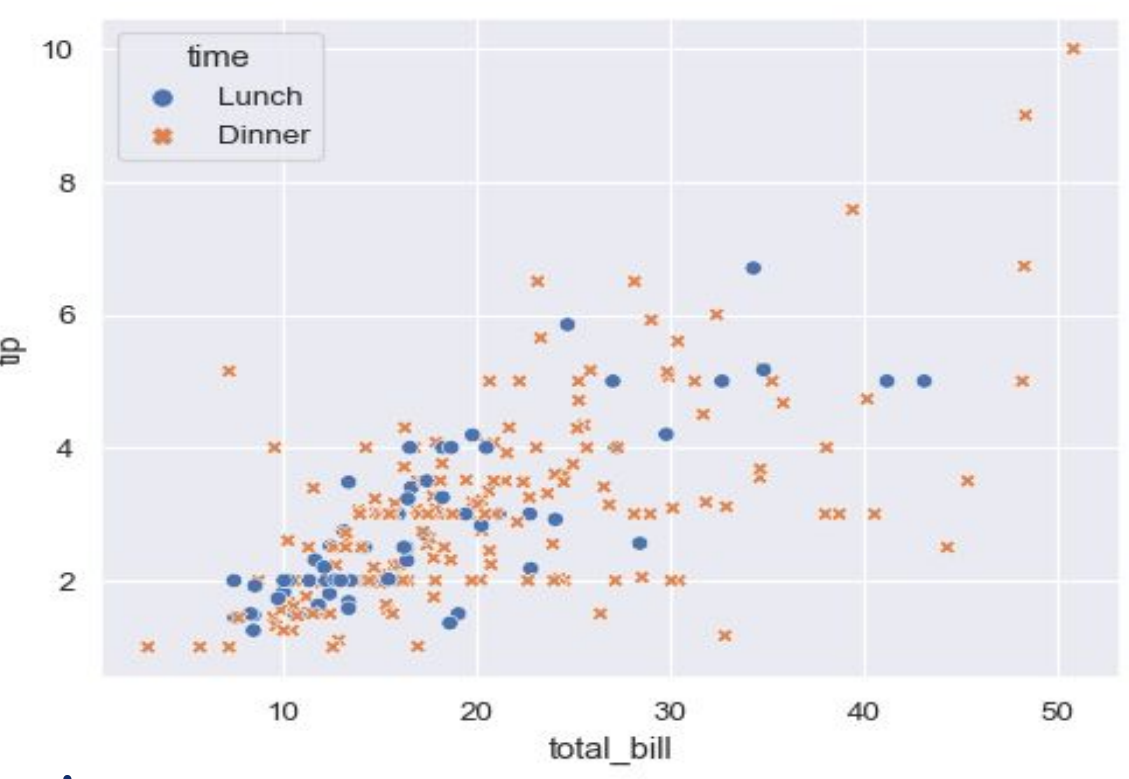
Pie chart represent the percentage of the data by each category.

Methods of Travelling to School



Pair-plot

Pair plot show the bivariate distribution of the datasets. It show the pairwise relationship between the variable

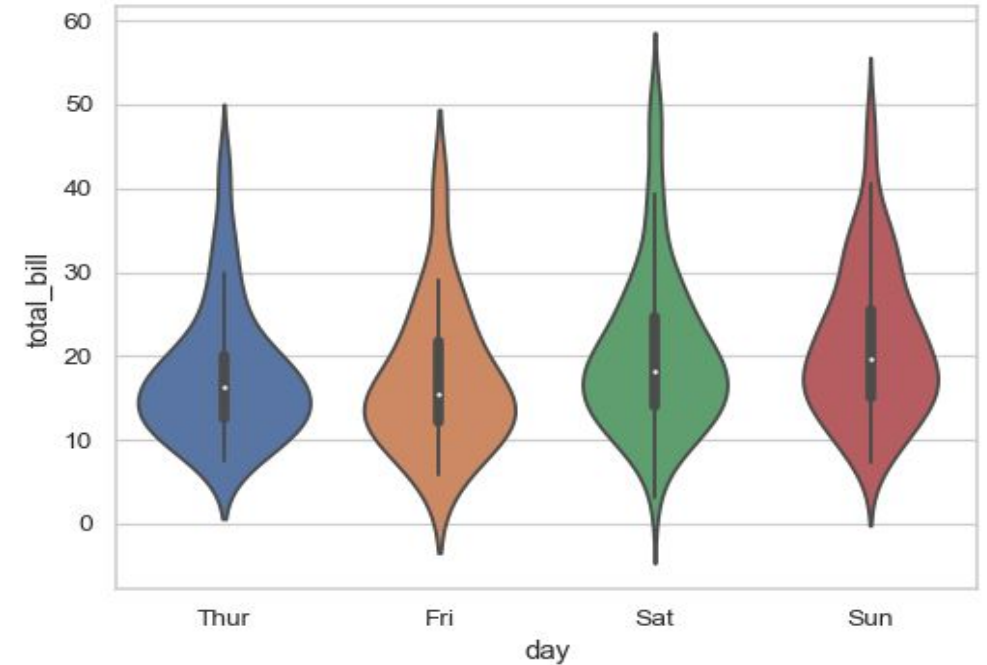


Scatter Plot

Scatter plot represent the relationship between two numerical variable. It show the correlation between two variable.

Violin Plot

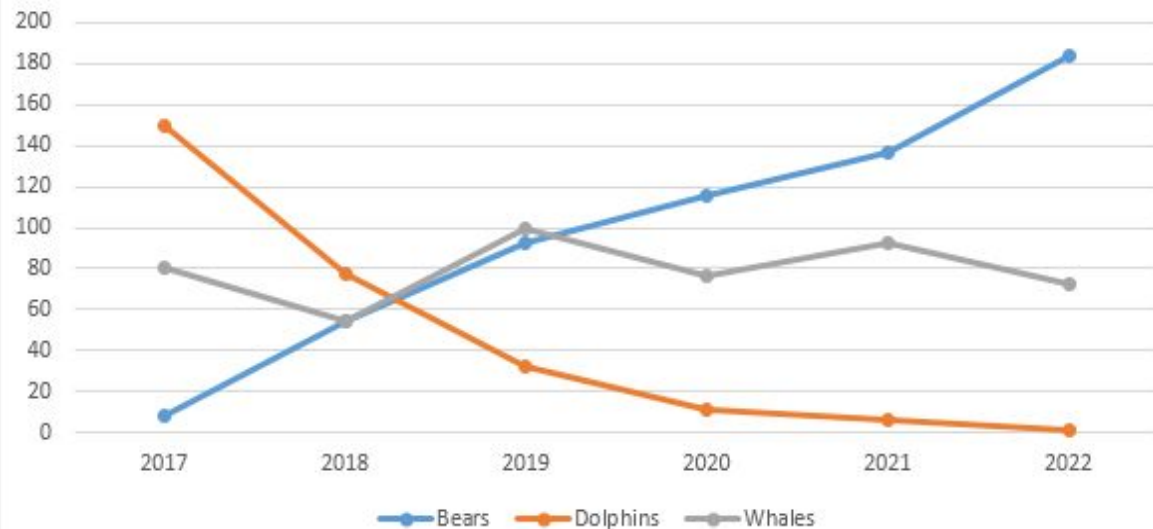
Violin chart are used to plot numeric data



Line Chart

Line chart are used to track change over line and short period of time. Line chart are used in time series data.

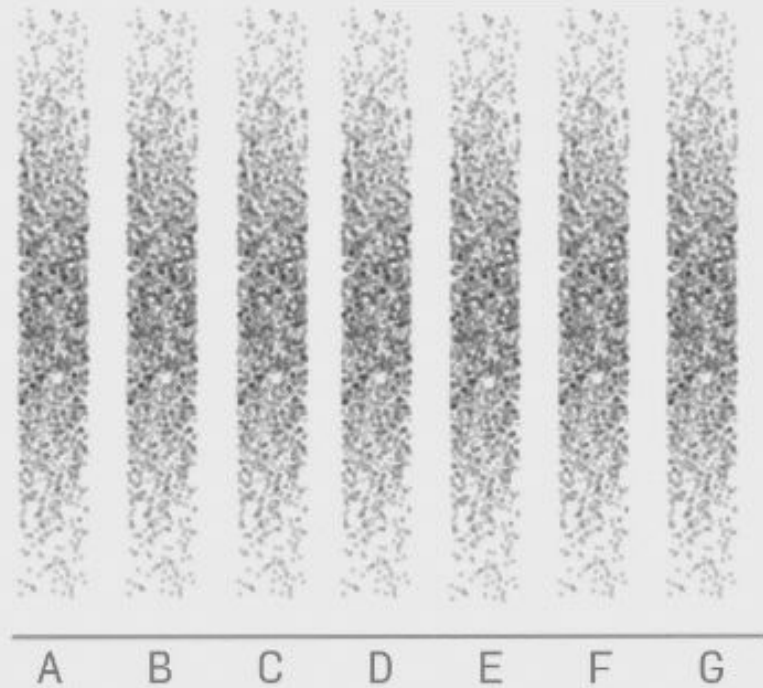
Wildlife Population



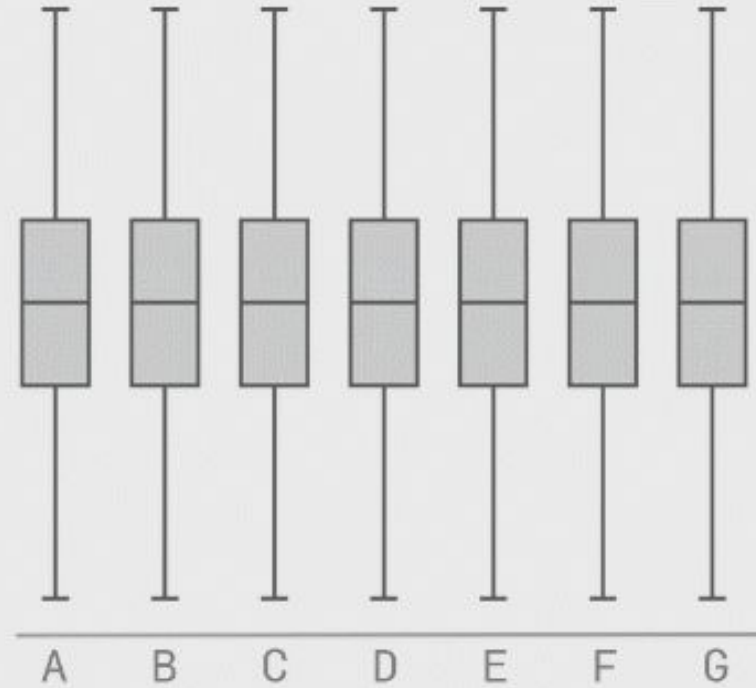
Box Plot vs ViolinPlot

- Violin Graphs are better than Box Plots
- Violin graph is visually intuitive & attractive.

Raw Data



Box-plot of the Data



Violin-plot of the Data

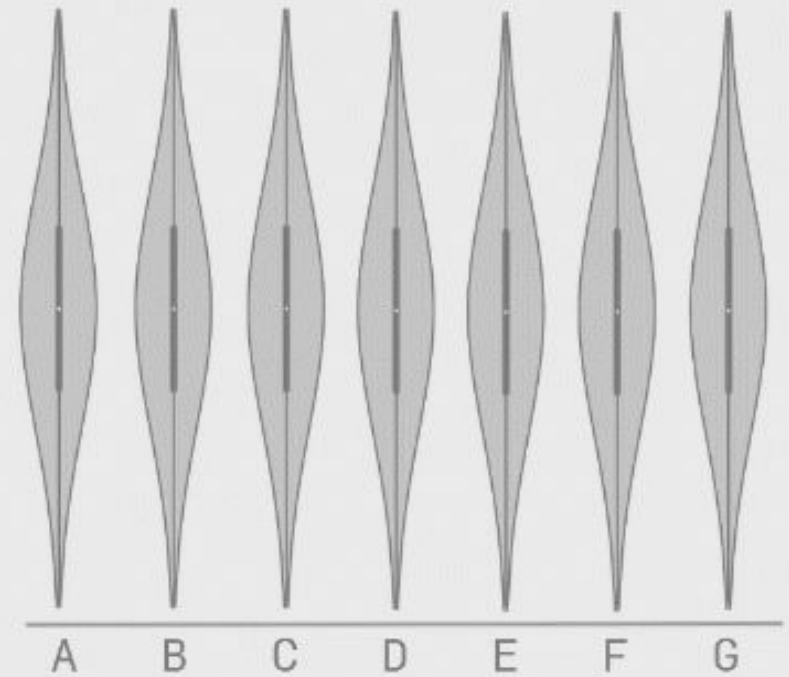
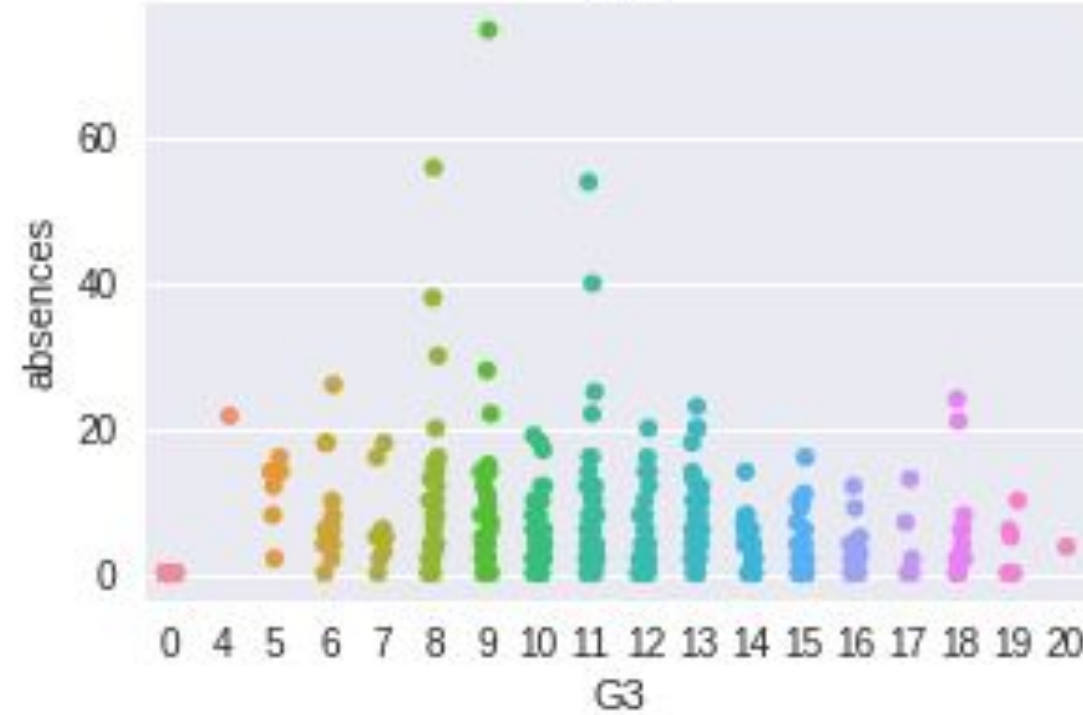


Image Courtesy: <https://blog.bioturing.com/>

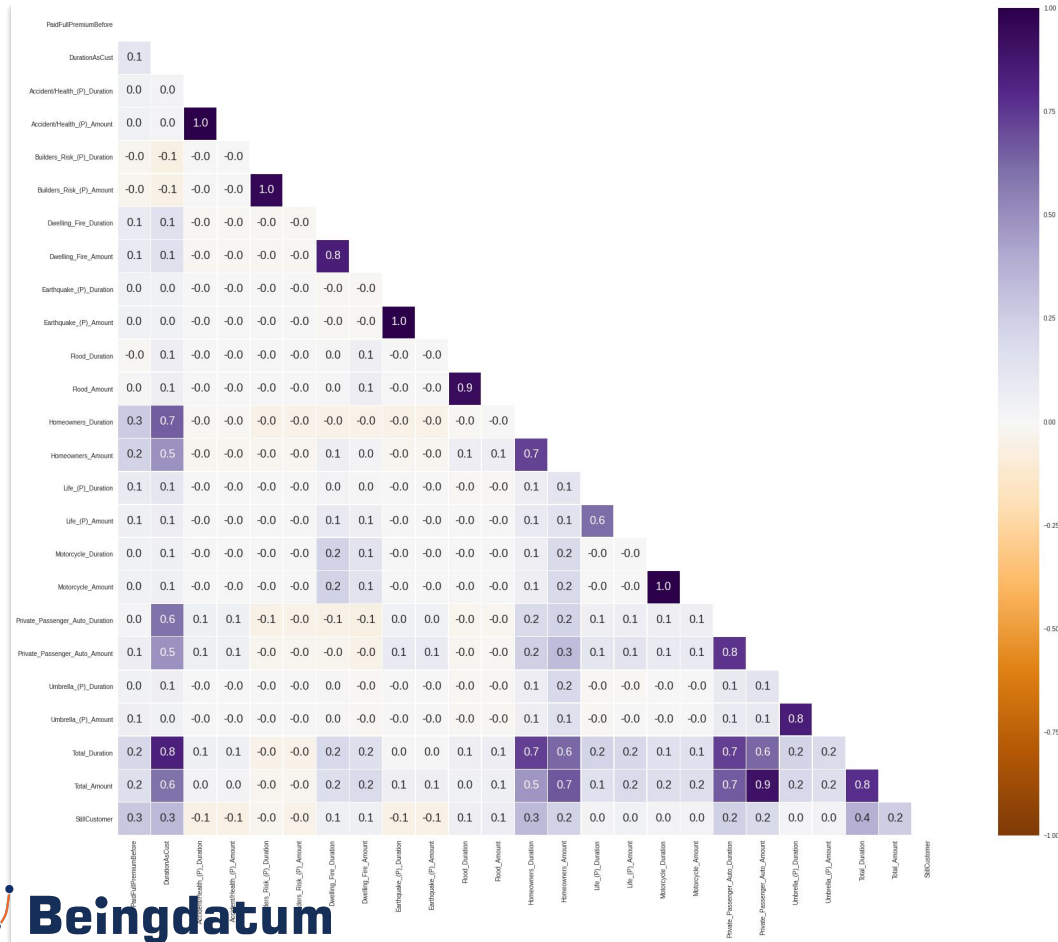
Strip Plot

Strip plots are used to draw a scatter plot for on the categorical basis



Heatmaps

A heat map is data analysis software that uses color the way a bar graph uses height and width.



Steps Involved in EDA

01

Data Sourcing

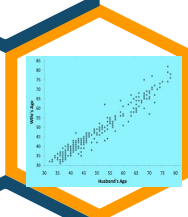


Data Cleaning

02

03

**Univariate Analysis
with Visualisation**



**Bivariate Analysis with
Visualisation**

04

05

Derived Metrics



Data Sourcing is the process of gathering data from multiple sources as external or internal data collection.

There are two major kind of data which can be classified according to the source:

1. Public data
2. Private data

Public Data:- The data which is easy to access without taking any permission from the agencies is called public data. The agencies made the data public for the purpose of the research. **Like** government and other public sector or ecommerce sites made there data public.

Private Data:- The data which is not available on public platform and to access the data we have to take the permission of organisation is called private data. **Like** Banking ,telecom ,retail sector are there which not made their data publicly available.



- After collecting the data , the next step is data cleaning. Data cleaning means that you get rid of any information that doesn't need to be there and clean up by mistake.
- Data Cleaning is the process of clean the data to improve the quality of the data for further data analysis and building a machine learning model.
- The benefit of data cleaning is that all the incorrect and irrelevant data is gone and we get the good quality of data which will help in improving the accuracy of our machine learning model.

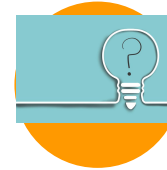
The following are some steps involve in Data Cleaning:

- Handle Missing Values
- Standardisation of the data
- Outlier Treatment
- Handle Invalid values



Some Questions you need to ask yourself about the data before cleaning the data

what you are reading, does it make sense?



Does this data make sense?

Does this data you're looking at match the column labels?

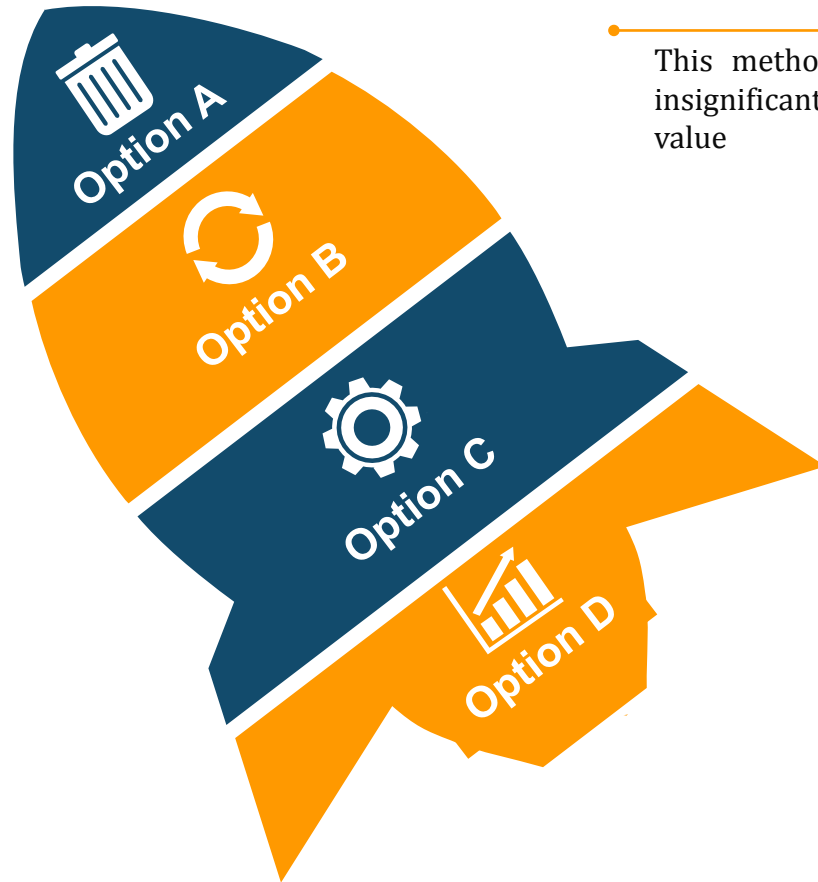


After compute the summarize statistics for numerical data, does it make sense?

Does the data follow the rules for this field?



Handle Missing Value



Delete Rows/Columns



This method we commonly used to handle missing values. Rows can be deleted if it has insignificant number of missing value Columns can be delete if it has more than 75% of missing value

Replacing with mean/median/mode



This method can be used on independent variable when it has numerical variables. On categorical feature we apply **mode** method to fill the missing value.

Algorithm Imputation



Some machine learning algorithm supports to handle missing value in the datasets. Like KNN, Naïve Bayes, Random forest.

Predicting the missing values



Prediction model is one of the advanced method to handle missing values. In this method dataset with no missing value become training set and dataset with missing value become the test set and the missing values is treated as target variable.

Example

For Numerical Data

Airlines	Ticket Price
Indigo	3887
Air Asia	7662
Jet Airways	-
Air India	5221
SpiceJet	4321

Suppose we have Airlines ticket price data in which there is missing value.
Steps to fill the numeric missing value:-

- 1) Compute the mean/median of the data
 $(3887+7662+5221+4321)/4 = \mathbf{5272.75}$
- 2) Substitute the Mean of the value in missing place.

For categorical Data

Airlines	Ticket price

Suppose we have Missing values in the categorical data:
Then we take the mode of the dataset a to fill the missing values:

Here :

Mode = Indigo

We substitute the Indigo in place of missing value in Airline column.

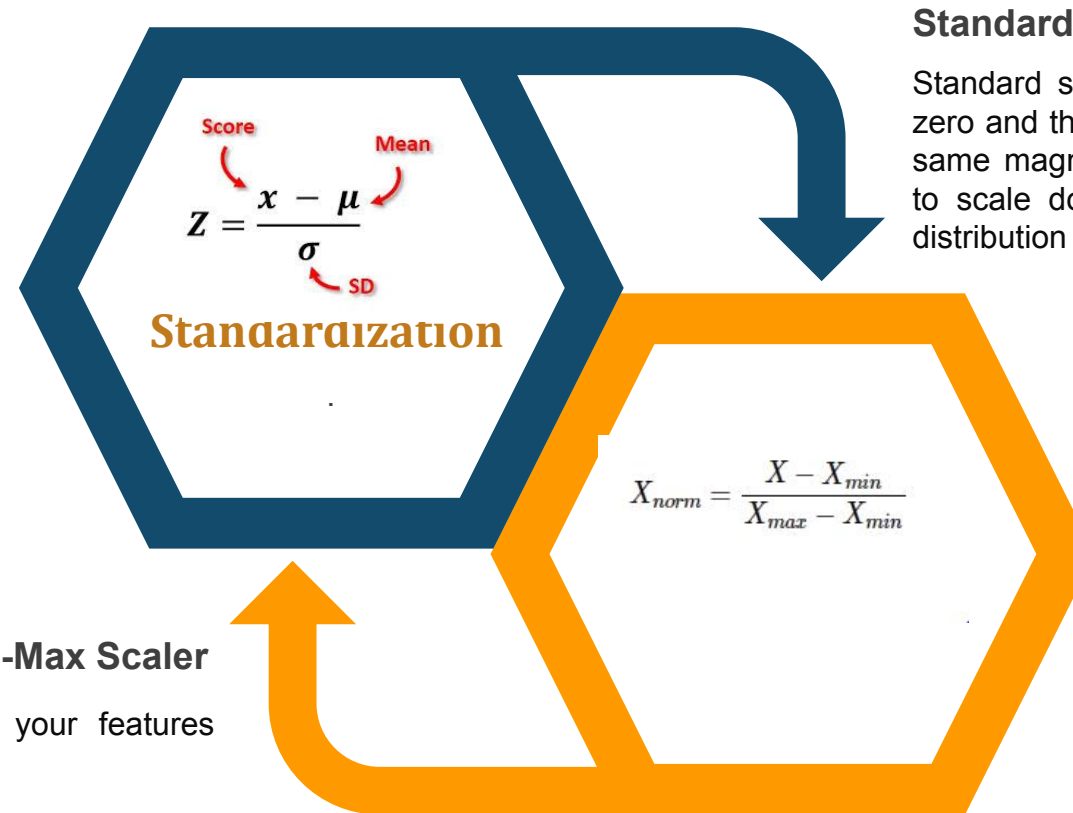
Standardization/ Feature Scaling

Feature scaling is the method to rescale the values present in the features. In feature scaling we convert the scale of different measurement into a single scale. It standardise the whole dataset in one range.

Importance of Feature Scaling:- When we are dealing with independent variable or features that differ from each other in terms of range of values or units of the features, then we have to normalise/standardise the data so that the difference in range of values doesn't affect the outcome of the data.



Feature Scaling Technique



Standard Scaler

Standard scaler ensures that for each feature, the mean is zero and the standard deviation is 1, bringing all feature to the same magnitude. In simple words Standardization helps you to scale down your feature based on the standard normal distribution

Min-Max Scaler

Normalization helps you to scale down your features between a range 0 to 1

Example

Normalisation

Age	Income (£)	New value
24	15000	$(15000 - 12000)/18000 = 0.16667$
30	12000	$(12000 - 12000)/18000 = 0$
28	30000	$(30000 - 12000)/18000 = 1$

Income Minimum = 12000

Income Maximum = 30000

$(\text{Max} - \text{min}) = (30000 - 12000) = 18000$

Hence, we have converted the income values between 0 and 1

Please note, the new values have

Minimum = 0

Maximum = 1

Average = $(15000 + 12000 + 30000)/3 = 19000$

Standard deviation = ??

Standardization

Age	Income (£)	New value of Income
24	15000	$(15000 - 19000)/\text{std} = ??$
30	12000	$(12000 - 19000)/\text{std} = ??$
28	30000	$(30000 - 19000)/\text{std} = ??$

Outlier Treatment

Outliers are the most extremes values in the data. It is a abnormal observations that deviate from the norm. Outliers do not fit in the normal behaviour of the data.

Detect Outliers using following methods:

- 1.Boxplot
2. Histogram
3. Scatter plot
- 4.Z-score
5. Interquartile range(values out of 1.5 time of IQR)

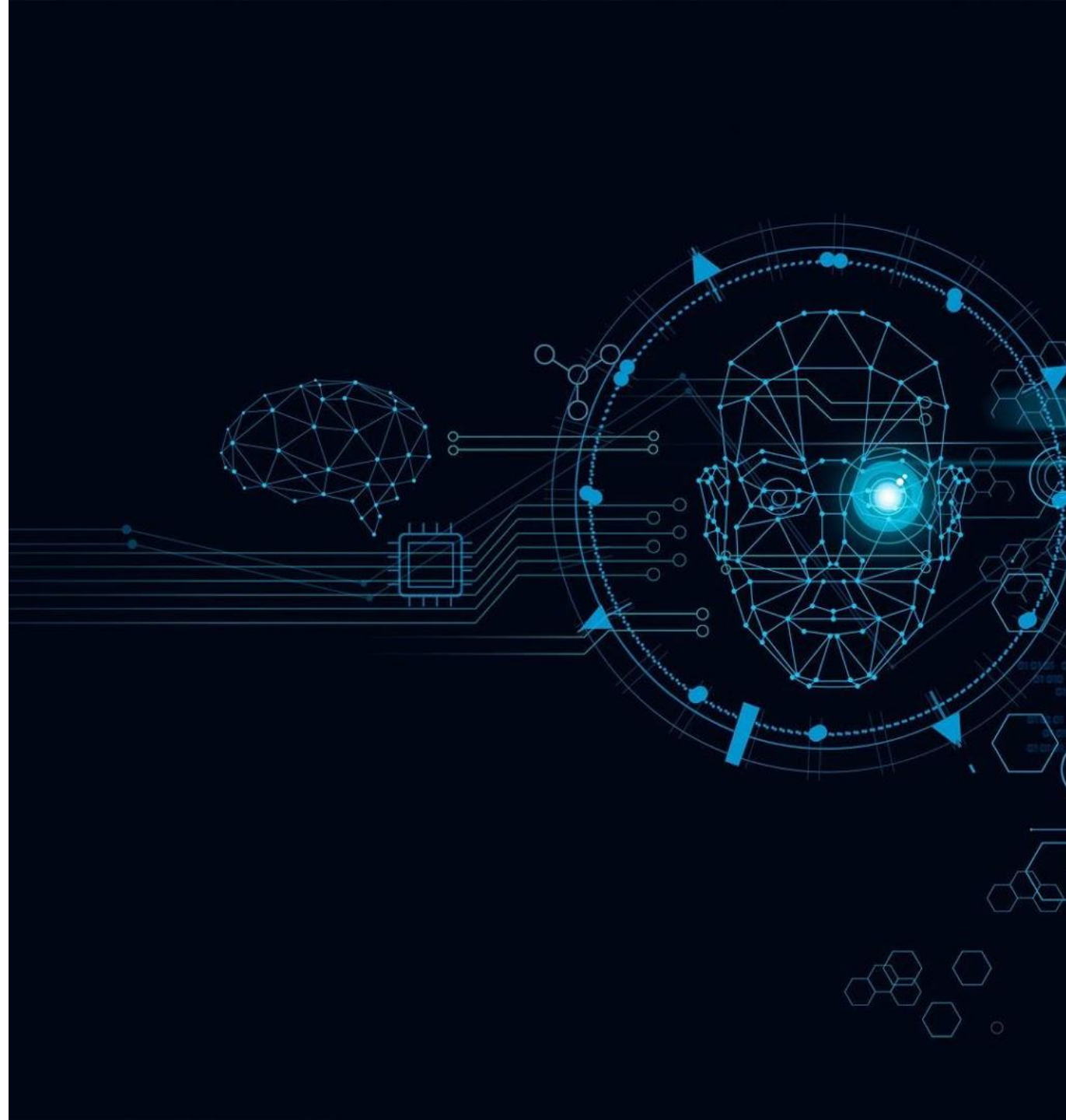
Handle Outlier using following methods:-

- 1.Remove the outliers.
- 2.Replace outlier with suitable values by using following methods:-
 - Quantile method
 - Interquartile range
- 3.Use that ML model which are not sensitive to outliers
Like:-KNN,Decision Tree,SVM,NaïveBayes,Ensemble methods



Handle Invalid Value

- **Encode Unicode properly:-** In case the data is being read as junk characters, try to change encoding, E.g. CP1252 instead of UTF-8.
- **Convert incorrect data types:-** Correct the incorrect data types to the correct data types for ease of analysis. E.g. if numeric values are stored as strings, it would not be possible to calculate metrics such as mean, median, etc. Some of the common data type corrections are — string to number: "12,300" to "12300"; string to date: "2013-Aug" to "2013/08"; number to string: "PIN Code 110001" to "110001"; etc.
- **Correct values that go beyond range:-** If some of the values are beyond logical range, e.g. temperature less than -273°C (0°K), you would need to correct them as required. A close look would help you check if there is scope for correction, or if the value needs to be removed.
- **Correct wrong structure:-** Values that don't follow a defined structure can be removed. E.g. In a data set containing pin codes of Indian cities, a pin code of 12 digits would be an invalid value and needs to be removed. Similarly, a phone number of 12 digits would be an invalid value



Univariate Analysis



Univariate analysis deal with analysing one feature at a time. Univariate analysis is important to understand the single variable at a time. The main purpose of univariate analysis is to make data easier to interpret. Mainly, Histogram is used to visualize univariate

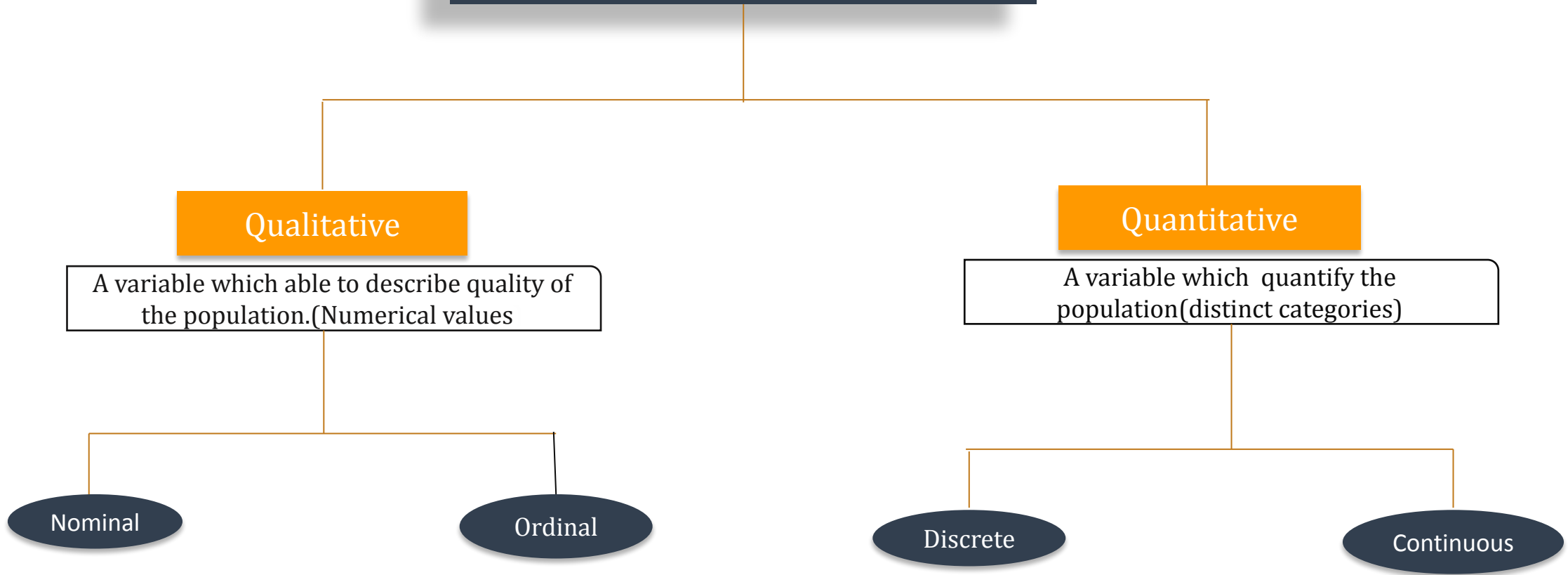
Univariate Types:

Frequency ➤ Measure of frequency include frequency table, where you tabulate how often a particular category or particular value appear in the dataset.

Central Tendency ➤ It is a typical value for a probability distribution for univariate measure of central tendency we have Mean, Median, Mode.

Dispersion ➤ It tries to capture how your data points jump around.

Types of Data



Types of Data



Discrete

It has a discrete value that means it takes only counted values not decimal values. Like count of students in class



Continuous

A number within a range of a value is usually measured, such as height.



Nominal

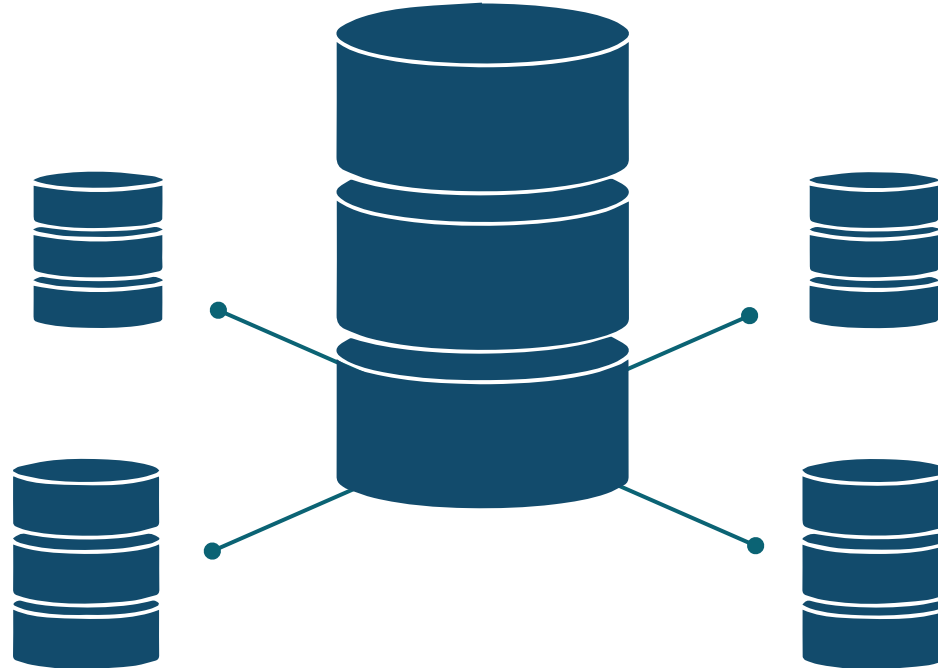
It represents qualitative information without order. Values represent discrete units.

Like Gender: Male/Female, Eye colour.



Ordinal

It represents qualitative information with order. It indicates the measurement classification are different and can be ranked. Let's say Economic status: high/medium/low which can be ordered as low, medium, high.



Segmented Univariate Analysis

Segmented Univariate Analysis allow you to compare subset of data it help us to understand how the relevant metric varies across the different segment.

The Standard process of segmented univariate analysis is as follow:

- Take a raw data
- Group by dimensions
- Summarise using a relevant metric like mean ,median.
- Compare the aggregate metric across the categories.



Bivariate Analysis

Data which has two variables ,you often want to measure the relationship that exists between these two variables.

Bi-variate Types:

Correlation ➤ Correlation measure the strength as well as the direction of the linear relationship between the two variables. Its range is from -1 to +1.

- If one increases as the other increases, the correlation is positive
- If one decreases as the other increases, the correlation is negative
- If one stays constant as the other varies, the correlation is zero

Covariance ➤ Covariance measure how much two random variable vary together. Its range is from $-\infty$ to $+\infty$.



Example

To see the distribution of two categorical variables. For example, if you want to compare the number of boys and girls who play games, you can make a 'cross table' as given below:

	Everyday	Never	Once a month	Once a week	Total
Boy	3474	154	150	780	4558
Girl	2776	175	200	1046	4197
Total	6250	329	350	1826	8755

To see the **distribution of two categorical variables** with one continuous variable. For example, you saw how a student's percentage in science is distributed based on the father's occupation (categorical variable 1) and the poverty level (categorical variable 2).

Derived Metrics

Derived metrics create a new variable from the existing variable to get a insightful information from the data by analysing the data.



Feature Binning

Feature binning converts or transform continuous/numeric variable to categorical variable. It can also be used to identify missing values or outliers.

Type of Binning:-

1. Unsupervised Binning
 - a) Equal width binning
 - b) Equal frequency binning
2. Supervised Binning
 - a) Entropy based binning



Un Supervised Binning

It transform continuous or numeric variable into categorical value without taking dependent variable into consideration

Equal Width

Equal width separate the continuous variable to several categories having same range of width.

Equal Frequency

Equal frequency separate the continuous variable into several categories having approximately same number of values.

Entropy Based

Entropy based binning separate the continuous or numeric variable majority of values in a category belong to same label of class.

Supervised Binning

It transform continuous or numeric variable into categorical value taking dependent variable into consideration.



Feature Encoding

Feature encoding help us to transform categorical data into numeric data.

Label encoding

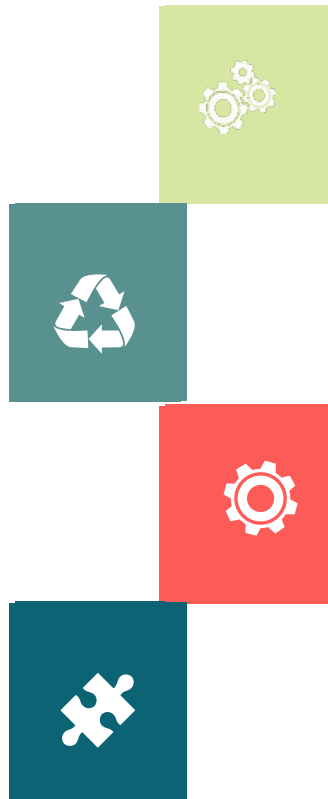


Label encoding is technique to transform categorical variables into numerical variables by assigning a numerical value to each of the categories.

Target Encoding



In target encoding, we calculate the average of the dependent variable for each category and replace the category variable with the mean value



One-Hot encoding



This technique is used when independent variables are nominal. It creates **k** different columns each for a category and replaces one column with 1 rest of the columns is 0.

Hash Encoder



TheHashencoder represents categorical independent variable using the new dimensions. Here, the user can fix the number of dimensions after transformation using component argument.

Uses Cases



Basically EDA is important in every business problem, it's a first crucial step in data analysis process. Some of the use cases where we use EDA is:-

- **Cancer Data Analysis :-** In this data set we have to predict who are suffering from cancer and who's not.
- **Fraud Data Analysis in E-commerce Transactions :-** In this dataset we have to detect the fraud in a E-commerce transaction.



THANK YOU

