Explaorartory Data Analysis:

EDA Agenda:

business undestanding-->data Understanding-->data collection-->data preparation

The eda helps in undestanding the data and perform data analytics.

Data Analytics Process:
1)Data collection(Raw data)
2)Processing the data
3)Results in the clean data

The data processing and the Eda is an iterative process.
As we learn to explore the data we tend to understand the flaws of the data.
—>The Eda help in understanding the data and gain the confidence in the data so that the data can be given ML models
—>The Eda serve as baseline /dirty model where we further develop to for a perfect ML model
—>The Eda also helps in summarising the data and helps in gaining the visual understanding by visual analytics

1)Data sourcing
—>The first process in web scraping or collection the data from the internal company or external

2)data preprocessing or data cleaning:
After collecting the data we do the data cleaning two get rid of the unnecessary data.

** Sometimes we tend to go for Eda and in some other cases we directly start feeding the ML model

—>process in data cleaning:
1)handling the missing values
2)standardisation of the data
3)outlier treatment
4)Handle invalid values

The normalisation and standardisation is classified into the feature scaling

Doubt:The standardisation is both classified into both into data cleaning and the feature scaling what is the difference.

Handling the missing values:
1)deleting the rows or columns
—>The row can be deleted if the data is insignificant
—>The column can be removed If it has 75% of values are missing
**Deleting the data helps in giving clarity over the data without any confusion

**If there is data missing instead of replacing with the random value which would lead to faulty ml model we would try to replace the data by considering the total data mean/median/mode

After filling the test the model and obtain accuracy and if the accuracy is less then we would fill it out with median instead of mode.

Algorithm Imputation:
Using the algorithm to fill out the values.like KNN,Naive Bayes and random forest

**Diffrentiate the missing values as a test set and the values which are already there as the training set(This is a very big level work and accuracy testing is not possible without the data available)

Basic program for data cleaning:

```
Import padas as pd
Import numpy as np
Import Matplotlib.pyplot as plt


df=pd.read_csv("")#importing the dataset

Df.info()
```

By obtaining the info about the data we can find the number of null and non null values by which we can undestand whether to keep the column or
or remove column base on the percentage of the missing values

also we can find out the number of Null values by using this code directly
print(df.isnull.sum())

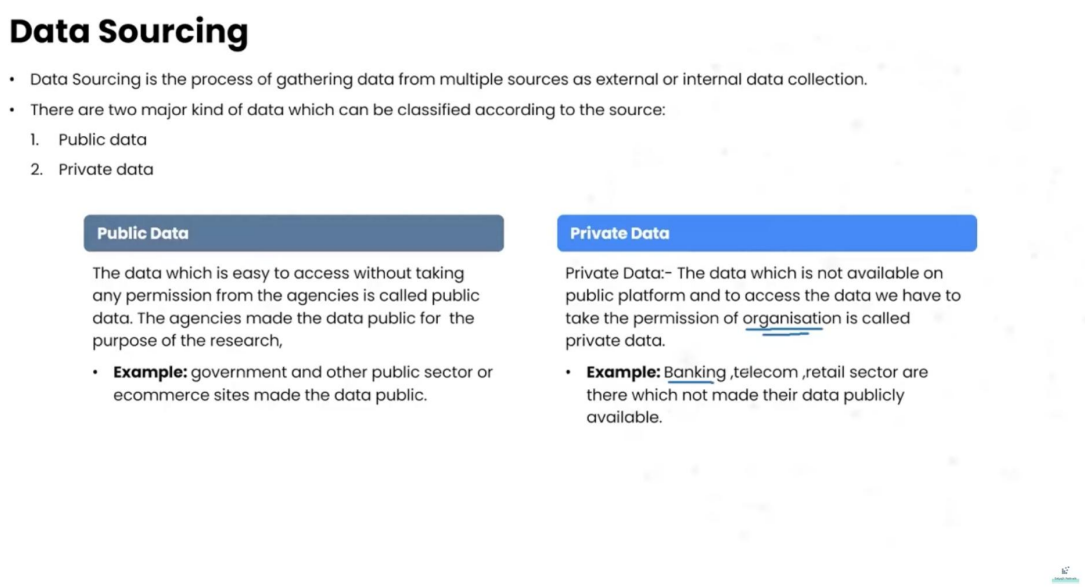
**deleting the column with missing data:

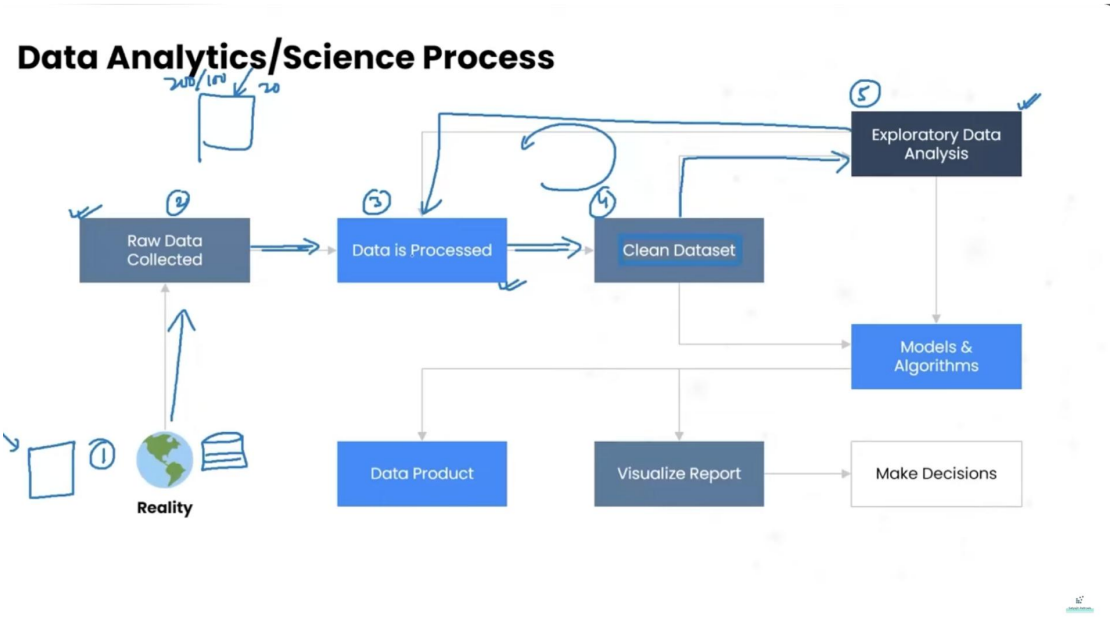
updated_df=df.dropna(axis=1)

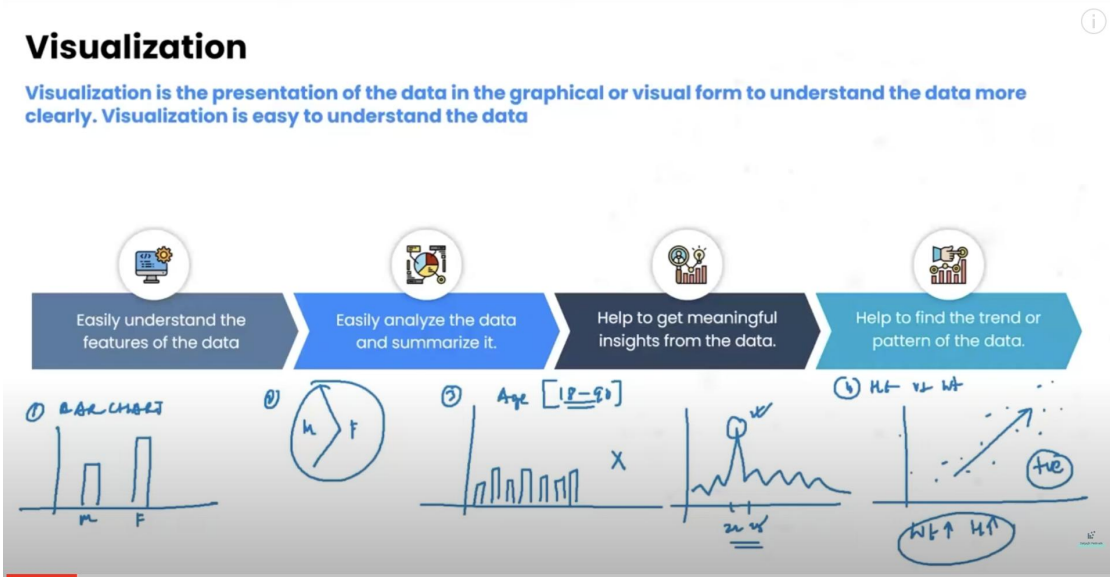
updated_df.info() #The updated data frame with remove columns

But the disadvantage if there is null values it completes delete it,which again leads to loosing the important data

In machine learning we cant predict whether deleting the columns or deleting the rows or imputing would work it wold al ways be like hit and try.







df['Age'].mean()

df["Age"].median()

updated_df['Age']=updated_df['age'].fillna(df['age'].mean())

updated_df['Age']=updated_df['age'].fillna(df['age'].median())

imputing the the values using the mean and median

for more outliers use mean and for less outlies use median.

There are also methods like
1) forward filling

updated_df[‘Age’]=updated_df[‘Age’].ffill()
2) and backward filling
updated_df[‘Age’]=updated_df[‘Age’].bfill()

There is no best techniques everything should be hit and try and which ever has the high accuray shoud be taken inti the account.

Time series Algorithm:
10
20
40
50
NA-->this vlaue is filled by using the before values and feeding it into time series algorithm and then result value would be filled but it is a tedious process.

Feature Scaling:
This technique is mostly used in predictive model building. But sometimes it is also used in eda.

**Feature scaling is the method to rescale the measurements the standarization and normalization comes under feature scaling.

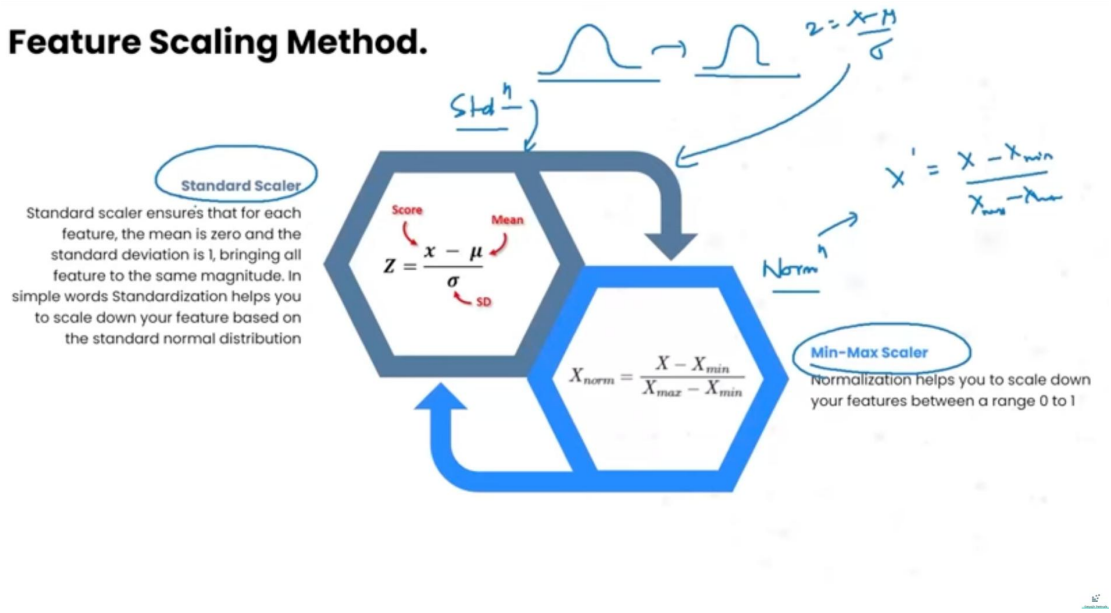
Height:(cm)
170
180
165
160
175

weight:
89
90
45
87
90

Consider the height and weight if it is given to the ml model then the model would be be by default giving more priority to the height since it is mor in numbber and less importance to the weight this should be normalizaed to make the model work and give correct output

To normalize it just divide the weight with the highest weight.so that the importance would be same and the scale would be only (0-1).

even perfrom the same for the height.These values called scale down values.



both normalization and standardization:
Standardization is making mean equal to zero and standard deviation equal to 1

normalization is making the values between the range 0-1

an Example to perform to Standardization:(follows the normal distribution)

z=(x-mean)/standard deviation.

converting the values to the z score and obtainging the values from the z-table.

-->sklearn library is used to perform all this standarization and the normalizaation.