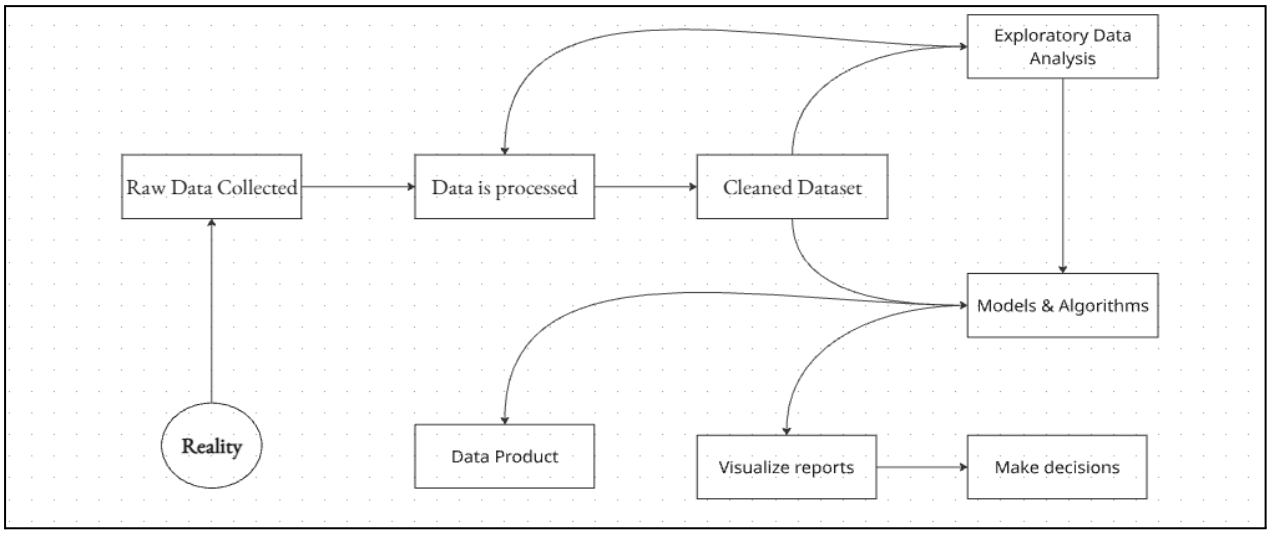
EXPLORATORY DATA ANALYSIS(EDA)



- It is an approach to analyze the dataset to summarise the main features in form of visual methods.
- It is a data exploration technique to understand various aspects.
- Data must be ready to use ML model.
- It help us finding errors,discovering data,mapping out data structures and finding out anomalies.
- Used for business aspect as considered for deep thorough analysis.
- It acts as a baseline model.

Data Sourcing

Public Data The data which is easy to access without taking permission from agencies. Eg:- Government and other public sector ecommerce sites made the data public.	**Private Data** The data which is not available on public platform and to access the data we need to seek permission from organisation. Eg:- Banking,telecom,retail sector are not made public.

Data Cleaning is avoiding the information that is not used and not needed. This is to improve the quality of data for further analysis. It also helps improve accuracy of our model. It involves:- handling missing and invalid values,outlier treatment and standardisation of the dataset.

Handle Missing Values

i. Del rows/columns:- insignificant missing value
ii. Replacing with mean/median when numerical features and mode when categorical feature.
iii. Algorithm imputation:- KNN,Naive Bayes and Random forest.
iv. Predicting the missing value:- missing value becomes the training set and dataset with missing value becomes the test set and missing values are treated as target variable.

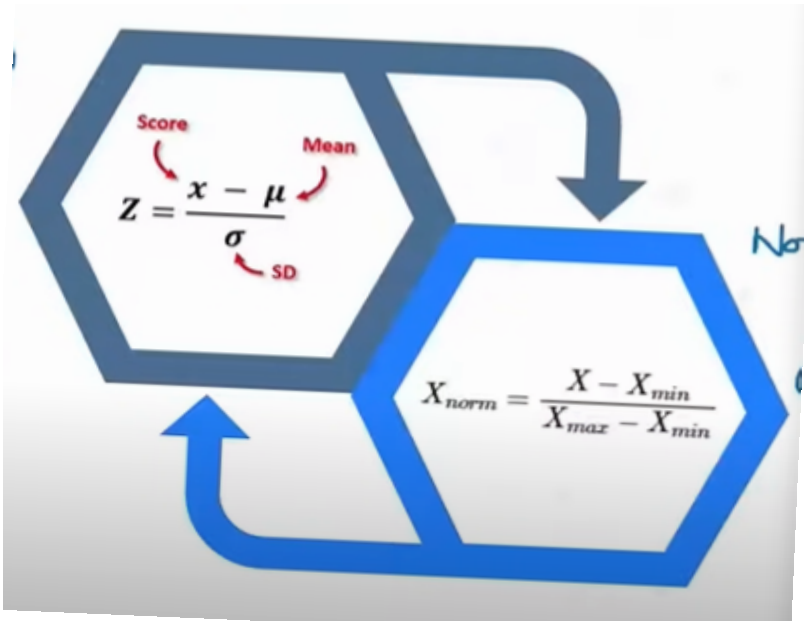
NOTE:- *While handling missing values,we can either try imputation or deleting rows/columns using hit and trial method.*

Feature Scaling

It is the method to rescale the values present in the features. In feature scaling,we convert the scale of different measurements into a single scale. It standardises the whole dataset in one range.
When we deal with independent variables that differ from each other in terms of range of values,then we have to standardize the data so that the difference in range of values doesn't affect the outcome of data.

Method used:-

1. Standard Scaler:- It ensures that for each feature,mean is zero and standard deviation is 1,bringing all features to the same magnitude. Standardization helps you to scale down your features based on the standard normal distribution



.

2. Min-Max Scaler:- Normalisation helps you to scale down your features between range 0 to 1.

Age	Income (£)	New value
24	15000	(15000 - 19000)/9643.65 = -0.4147
30	12000	(12000 - 19000)/9643.65 = -0.7258
28	30000	(30000-19000)/9643.65 = 1.1406

Average = (15000 + 12000 + 30000)/3 = 19000
Standard deviation = 9643.65

Hence, we have converted the income values to lower values using the z-score method.

x = c(-0.4147, -0.7258, 1.1406)
mean(x) = -0.000003 ~ 0
var(x) = 0.999 ~1

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
(Max – 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

Normalisation

It is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1.

Standardisation

It is another scaling technique where the values are centered around the mean with unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Outlier Treatment

Outliers are the extreme values in the data. It is an 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. Inter-quartile range(values out of 1.5 time of IQR)

Handling outliers using following methods:-

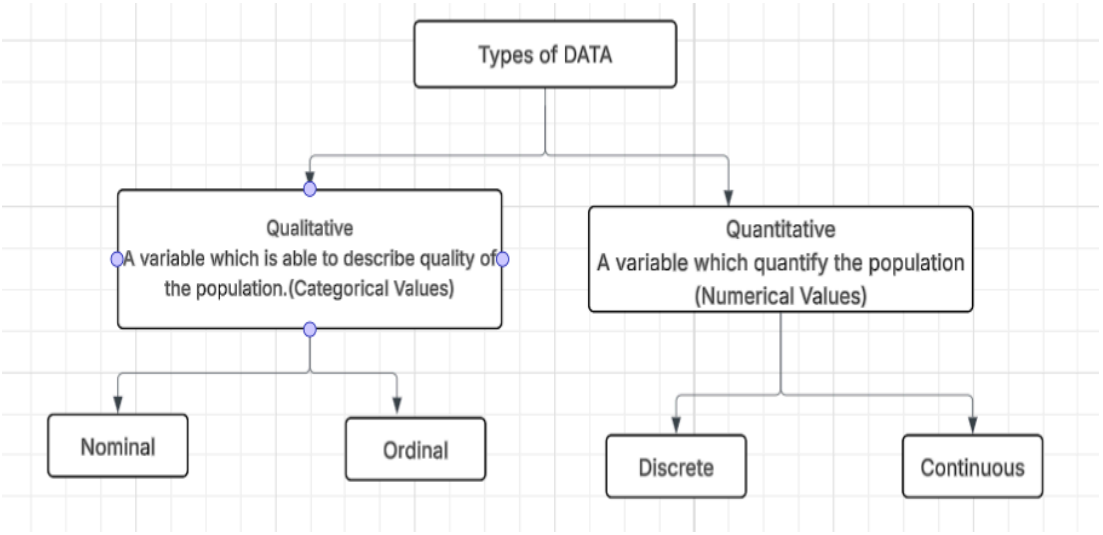
1. Remove the outliers.
2. Replace outlier with suitable value by Quantile method and Inter Quartile Range
3. Use the ML model not sensitive to outliers like KNN,Decision Tree, SVM, Naive Bayes,Ensemble methods

To check invalid data:-

1. Encode Unicode properly
   In case the data is being read as junk characters,try to change encoding. Eg: CP1252 instead of UTF-8
2. Convert incorrect data types

This is for ease of analysis. Eg:- If numeric values are stored as strings,it would not be possible to calculate metrics such as mean,median.
Some of common data type corrections are string to number and date.

3. Correct values that go beyond range
   If some of values are beyond the logical range, eg: temperature less than -273C 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,
4. Correct wrong structure
   Values that don't follow a defined structure can be removed.
   Eg:- in a dataset 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.



Types of Analysis:-

1. Univariate Analysis - data has only one variable
2. Bivariate Analysis- data has two variables. You often want to measure the relationship that exists

between these two categorical variables. It can also be performed with numerical values,or a combination of numerical and categorical values.

3. Multivariate Analysis:- Data which has more than two variables. You often want to measure the relationship that exists between these features.
4. Numerical Analysis:- Dealing with single numerical variable we might be interested in knowing their statistical information such as mean,median,25th percentile,75th percentile. While analysing multiple features, we might be interested in knowing their correlation with each other.

Derived Metrics

It creates a new variable from the existing variable to get a insightful information from the data by analysing the data.

1. Feature Binning
2. Feature Encoding
3. From domain knowledge
4. Calculated from Data

Feature Binning

It converts or transforms continuous/numeric variables to categorical variables.Also be used to identify missing values or outliers.

1. Unsupervised Binning - It converts or transforms continuous/numeric variables to categorical variables without taking dependent variables into consideration.
    a. Equal width binning - it separate the continuous variable to several categories having same range of width.

b. Equal frequency binning - it separate the continuous variable to several categories having approximately same number of values.

2. Supervised Binning - It converts or transforms continuous/numeric variables to categorical variables taking dependent variables into consideration.

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

Feature Encoding

1. Label Encoding is a technique to transform categorical variables into numerical variables by assigning value to each of the categories.
2. One -Hot Encoding is a technique 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. Here 0 represents the absence and 1 represents the presence.
3. Target Encoding is a technique that calculates the average of dependent variables for each category and replace the category variable with the mean value.
4. Hash Encoder represents categorical independent variable using the new dimensions. Here the user can fix the number of dimensions after transformation using component argument.

Customer Churn Analysis

Missing data-Initial Intuition

- For features with less missing values-can use regression to predict the missing values present depending on the feature.
- For features with very high number of missing values,it is better to drop those columns on analysis.
- As there's no criteria for deleting the columns with high number of missing values but you can delete the columns if you have more than 30-40% of missing values.