Univariate Analysis

Whenever we do single column analysis that can be called as Univariate Analysis.
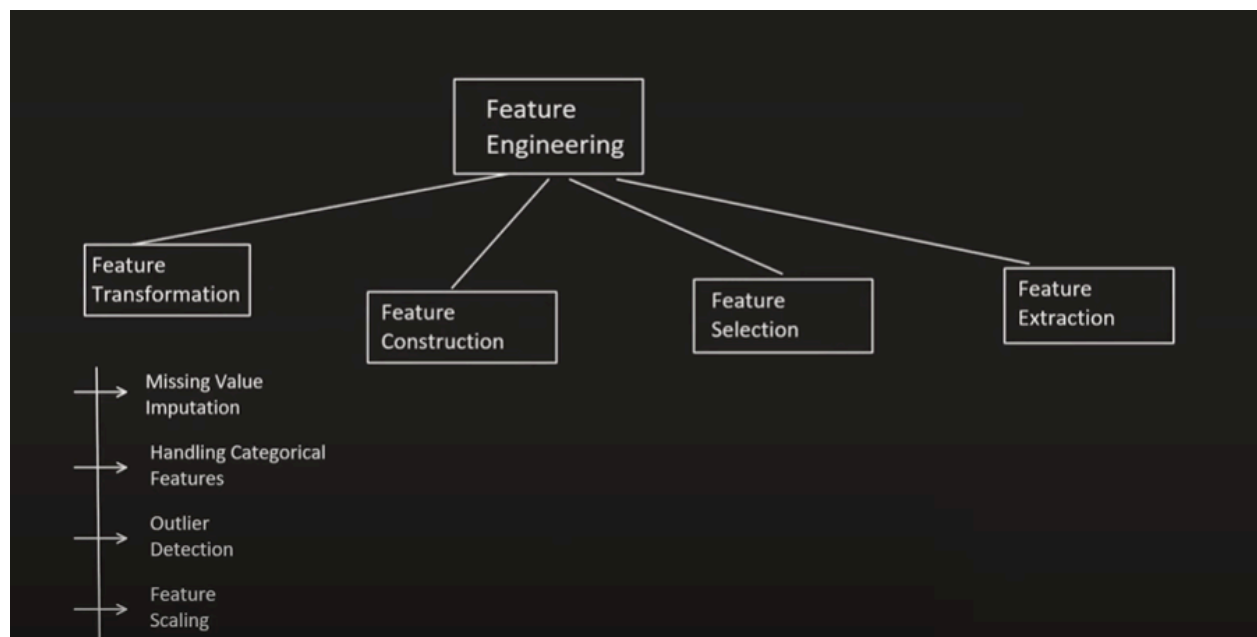
Bivariate Analysis

Whenever we do 2 column analysis, that can be called Bivariate Analysis.

Multivariate Analysis

If we are performing analysis on more than 2 columns that can be called as Multivariate Analysis.

Feature Engineering = Pre Processing of the Dataset

In Feature Construction: We Create columns on our own by examining the data.
In Feature Selection: We shortlist only those columns which are beneficial for our Model.
In Feature Extraction: Given features main se new features ko extract krna programmatically.

Feature Scaling:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

We do feature scaling in the end when we completely pre processed our data then we apply feature scaling.

Types of Feature Scaling:

- Standardization
- Normalization (Robust Scaler(tackle outliers), MinMax Scaler and etc)

Standardization is typically applied to numerical data, especially in machine learning and statistical analysis. It involves rescaling the features of a dataset so that they have a mean of 0 and a standard deviation of 1.

Normalization Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

Types of Normalization:

MinMax Scaling: This is the most basic form of normalization, where the values are scaled to a fixed range, typically between 0 and 1.

Mean normalization is a data preprocessing technique where the mean (average) of a feature is subtracted from each value in the feature, and then the result is divided by the range of the feature. It is being used in (when we want data in center)

MaxAbsScaler, also known as MaxAbsScaling, is a data preprocessing technique used to scale features by their maximum absolute value. MaxAbsScaling is applied to numerical data where the scale of features varies widely, and preserving the direction of the data is important. Examples of such data include:

- Text data represented using TF-IDF (Term Frequency-Inverse Document Frequency) vectors.
- Sparse matrices representing user-item interactions in recommendation systems.
- Data with outliers where Min-Max scaling might not be suitable.

Robust scaling, also known as robust normalization or feature scaling, is a data preprocessing technique used to scale features by removing the median and scaling them according to the interquartile range (IQR). This method is robust to outliers in the data because it uses percentile-based statistics rather than mean and standard deviation, which can be heavily influenced by outliers.

- Financial data with extreme values or anomalies.
- Sensor data where outliers may indicate malfunctioning equipment.
- Biological data with outliers caused by experimental error.