**Exploratory Data Analysis**

Exploratory data analysis is a data analytics process to understand the data in depth, to analyse and investigate data sets to summarize their main characteristics, often with visual methods. It is used to see what data can reveal beyond the formal modelling/hypothesis testing task and provides a better understanding of data sets variables and the relationships between them. It makes it easier for data scientists to discover patterns, spot anomalies, test hypothesis or check assumption. EDA was originally developed by American mathematician John Tukey in the 1970s.

**Important of Exploratory Data Analysis**

EDA is very crucial in data science, below are few areas where EDA plays an important role

* It helps to look at data before making any assumptions and in choosing a better machine learning model.
* It helps to identify obvious errors as well as better understanding patterns within the data.
* It helps to detect outliers or anomalies events and to remove irregularities or unnecessary value.
* It helps to find interesting relations among the variables.

**Steps involved in EDA**

* Data Collection: This is an important part of exploratory data analysis. It is refer to as the process of finding and loading data into our system.
* Data Cleaning: This is refer to as the process of removing unwanted variables and values from our dataset and getting rid of any irregularities in it. Irregularities like; missing values, Incorrect formatting of data, incorrect headers, anomalies or outliers.

**What is an Outlier**

Outlier is a commonly used terminology by analysts and data scientists, it needs close attention because it might result in wildly wrong estimations. Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

The most commonly used method to detect outliers is visualization. We use various visualization methods, like Box-plot, Histogram, Scatter Plot and there are various thumb rules to detect outliers, which are: Inter-quartile Range, Capping Methods, Bi-variate and Multivariate, Outliers are typically measured using either an index of influence or leverage or distance. Popular indices such as Mahalanobis distance and cook’s are frequently used to detect outliers.

Outlier values can either be deleted, trim, transform, bin, or through imputing which can be done using mean, median or mode imputation methods.

**Level of Analysis**

There are three level of EDA in statistics which are;

* Uni-variate Analysis
* Bi-variate Analysis
* Multivariate Analysis

**Uni-variate Analysis:** This is the most basic form of statistical data data analysis technique. It is used where data being analyzed contains only one variable, it doesn’t deal with causes or effect relationships. The key objectives of Uni-variate analysis is to simply describe the data and to find patterns within the data, this can be done either with graphical means or non graphical means by finding certain mathematical values in the data. The non graphical Method can be done using Mean, Median, Mode, Dispersion, Variance, Range, Standard deviation e.t.c

Statistical techniques used in conducting Uni-variate analysis are mostly descriptive in nature which include;

Frequency distribution tables, Histogram, Frequency polygons, Pie and Bar charts.

**Bi-variate Analysis:** This is slightly more analytical than uni-variate analysis, it is used when the data contains two variables and researchers aim to undertake comparisons between the two data set. It is used to measure the correlation between two variables. It is done with scattered plot.

Statistical techniques used for conducting Bi-variate analysis are; Correlation coefficient, Regression analysis.

**Multivariate Analysis:** This is a more complex form of statistical analysis techniques and it is used when there are more than two variables in the data set.

Statistical techniques used are; Factors analysis, cluster analysis, variance Analysis, Discriminate Analysis, Multidimensional Scaling, Principal component Analysis, Redundancy analysis

**Feature Engineering**

Feature engineering is the science (and art) of extracting more information from existing data. i.e making data out of the existing dataframe.

Feature engineering can be divided in 2 steps:

* Variable transformation.
* Variable / Feature creation.

These two techniques are vital in data exploration and have a remarkable impact on the power of prediction. Let’s understand each of this step in more details.

**Variable Transformation**

It refers to the replacement of a variable by a function. In other words, transformation is a process that changes the distribution or relationship of a variable with others.

There are various methods used to transform variables.

* Logarithm: Log of a variable is a common transformation method used to change the shape of distribution of the variable on a distribution plot. It can’t be applied to zero or negative values as well.
* Square / Cube root: The square and cube root of a variable has a sound effect on variable distribution. However, it is not as significant as logarithmic transformation. Cube root has its own advantage. It can be applied to negative values including zero. Square root can be applied to positive values including zero.
* Binning: It is used to categorize variables. It is performed on original values, percentile or frequency. For numeric categorical variables, bins can be also created based on the quantiles

**Variable / Feature creation**

Feature / Variable creation is a process to generate new variables / features based on existing variable(s).

There are various techniques to create new features. Let’s look at the some of the commonly used methods:

Derived Variables: This refers to creating new variables from existing variable(s) using set of functions or different methods such as taking log of variables, binning variables and other methods of variable transformation can also be used to create new variables.

Creating Dummy Variables: The most common application of dummy variable is to convert categorical variable into numerical variables. Dummy variables are also called Indicator Variables. It is useful to take categorical variable as a predictor in statistical models.