**What is Exploratory Data Analysis in Data Science?**

Exploratory Data Analysis (EDA) is one of the techniques used for extracting vital features and trends used by machine learning and deep learning models in Data Science.

**Importance of EDA in Data Science**

The Data Science field is now very important in the business world as it provides many opportunities to make vital business decisions by analyzing hugely gathered data. Understanding the data thoroughly needs its exploration from every aspect. The impactful features enable making meaningful and beneficial decisions; therefore, EDA occupies an invaluable place in Data science.

**Objective of Exploratory Data Analysis**

The overall objective of exploratory data analysis is to obtain vital insights and hence usually includes the following sub-objectives:

• Identifying and removing data outliers

• Identifying trends in time and space

• Uncover patterns related to the target

• Creating hypotheses and testing them through experiments

• Identifying new sources of data

**Role of EDA in Data Science**

The role of data exploration analysis is based on the use of objectives achieved as above.

After formatting the data, the performed analysis indicates patterns and trends that help to take the proper actions required to meet the expected goals of the business.

As we expect specific tasks to be done by any executive in a particular job position, it is expected that proper EDA will fully provide answers to queries related to a particular business decision.

As data science involves building models for prediction, they require optimum data features to be considered by the model.

Thus, EDA ensures that the correct ingredients in patterns and trends are made available for training the model to achieve the correct outcome, like a successful recipe.

Therefore, carrying out the right EDA with the correct tool based on befitting data will help achieve the expected goal.

**Steps Involved in Exploratory Data Analysis (EDA)**

 The key components in an EDA are the main steps undertaken to perform the EDA. These are as follows:

**1. Data Collection**

Nowadays, data is generated in huge volumes and various forms belonging to every sector of human life, like healthcare, sports, manufacturing, tourism, and so on. Every business knows the importance of using data beneficially by properly analyzing it. However, this depends on collecting the required data from various sources through surveys, social media, and customer reviews, to name a few. Without collecting sufficient and relevant data, further activities cannot begin.

**2. Finding all Variables and Understanding Them**

When the analysis process starts, the first focus is on the available data that gives a lot of information. This information contains changing values about various features or characteristics, which helps to understand and get valuable insights from them. It requires first identifying the important variables which affect the outcome and their possible impact. This step is crucial for the final result expected from any analysis.

**3. Cleaning the Dataset**

The next step is to clean the data set, which may contain null values and irrelevant information. These are to be removed so that data contains only those values that are relevant and important from the target point of view. This will not only reduce time but also reduces the computational power from an estimation point of view. Preprocessing takes care of all issues, such as identifying null values, outliers, anomaly detection, etc.

**4. Identify Correlated Variables**

Finding a correlation between variables helps to know how a particular variable is related to another. The correlation matrix method gives a clear picture of how different variables correlate, which further helps in understanding vital relationships among them.

**5. Choosing the Right Statistical Methods**

Depending on the data, categorical or numerical, the size, type of variables, and the purpose of analysis, different statistical tools are employed. Statistical formulae applied for numerical outputs give fair information, but graphical visuals are more appealing and easier to interpret.

**6. Visualizing and Analyzing Results**

Once the analysis is over, the findings are to be observed cautiously and carefully so that proper interpretation can be made. The trends in the spread of data and correlation between variables give good insights for making suitable changes in the data parameters. The data analyst should have the requisite capability to analyze and be well-versed in all analysis techniques. The results obtained will be appropriate to data of that particular domain and are suitable for use in retail, healthcare, and agriculture.

**Types of Exploratory Data Analysis**

There are three main types of EDA:

1. Univariate

2. Bivariate

3. Multivariate

**1. Univariate Non-Graphical**

It is the simplest of all types of data analysis used in practice. As the name suggests, uni means only one variable is considered whose data (referred to as population) is compiled and studied. **For example, data shows products produced each month for twelve months.** The main aim of univariate non-graphical EDA is to find out the details about the distribution of the population data and to know some specific parameters of statistics.

The significant parameters which are estimated from a distribution point of view are as follows:

• **Central Tendency:** This term refers to values located at the data's central position or middle zone. The three generally estimated parameters of central tendency are mean, median, and mode. Mean is the average of all values in data, while the mode is the value that occurs the maximum number of times. The Median is the middle value with equal observations to its left and right.

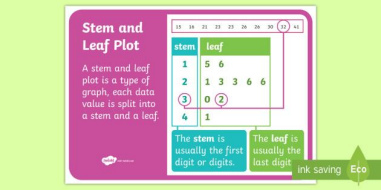
• **Range:** The range is the difference between the maximum and minimum value in the data, thus indicating how much the data is away from the central value on the higher and lower side.

• **Variance and Standard Deviation:** Two more useful parameters are standard deviation and variance. Variance is a measure of dispersion that indicates the spread of all data points in a data set. It is the measure of dispersion mostly used and is the mean squared difference between each data point and mean, while standard deviation is the square root value of it. The larger the value of standard deviation, the farther the spread of data, while a low value indicates more values clustering near the mean.

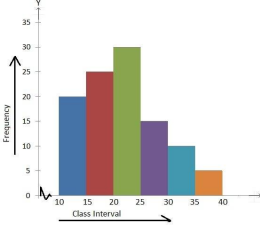
**2. Univariate Graphical**

The graphs in this section are based on **Auto MPG dataset** available on the UCI repository. Some common types of univariate graphics are:

• **Stem-and-leaf Plots:** This is a very simple but powerful EDA method used to display quantitative data but in a shortened format. It displays the values in the data set, keeping each observation intact but separating them as stem (the leading digits) and remaining or trailing digits as leaves. But histogram is mostly used in its place now.



• **Histograms (Bar Charts):** These plots are used to display both grouped or ungrouped data. On the x-axis, values of variables are plotted, while on the y-axis are the number of observations or frequencies. Histograms are very simple to quickly understand your data, which tell about values of data like central tendency, dispersion, outliers, etc. The simplest fundamental graph is a histogram, which is a bar plot with each bar representing the frequency, i.e., the count or proportion (the ratio of count to the total count of occurrences) for various values.

• 

**There are many types of histograms, a few of which are listed below:**

1. **Simple Bar Charts:** These are used to represent categorical variables with rectangular bars, where the different lengths correspond to the values of the variables.

2. **Multiple or Grouped charts:** Grouped bar charts are bar charts representing multiple sets of data items for comparison where a single color is used to denote one specific series in the dataset.

3. **Percentage Bar Charts:** These are bar graphs that depict the data in the form of percentages for each observation. The following image shows a percentage bar chart with dummy values.

4. **Box Plots:** These are used to display the distribution of quantitative value in the data. If the data set consists of categorical variables, the plots can show the comparison between them. Further, if outliers are present in the data, they can be easily identified. These graphs are very useful when comparisons are to be shown in percentages, like values in the 25 %, 50 %, and 75% range (quartiles).

**3. Multivariate Non-Graphical**

The multivariate non-graphical exploratory data analysis technique is usually used to show the connection between two or more variables with the help of either cross-tabulation or statistics.

• For categorical data, an extension of tabulation called cross-tabulation is extremely useful. For two variables, cross-tabulation is preferred by making a two-way table with column headings that match the amount of one variable and row headings that match the amount of the opposite two variables, then filling the counts with all subjects that share an equivalent pair of levels.

• For each categorical variable and one quantitative variable, we can generate statistical information for quantitative variables separately for every level of the specific variable. We then compare the statistics across the number of categorical variables.

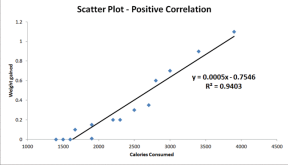
**4. Multivariate Graphical**

Graphics are used in multivariate graphical data to show the relationships between two or more variables. Here the outcome depends on more than two variables, while the change-causing variables can also be multiple.

Some common types of multivariate graphics include:

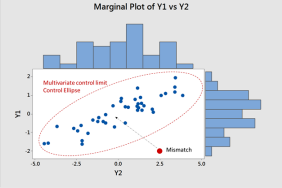
**A) Scatter Plot**

The essential graphical EDA technique for two quantitative variables is the scatter plot, so one variable appears on the x-axis and the other on the y-axis and, therefore, the point for every case in your dataset. This can be used for bivariate analysis.



**B) Multivariate Chart**

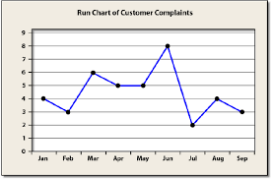
A Multivariate chart is a type of control chart used to monitor two or more interrelated process variables. This is beneficial in situations such as process control, where engineers are likely to benefit from using multivariate charts. These charts allow monitoring multiple parameters together in a single chart. A notable advantage of using multivariate charts is that they help minimize the total number of control charts for organizational processes. Pair plots generated using the Seaborn library are a good example of multivariate charts as they help visualize the relationships between all numerical variables in the entire dataset at once.



**C) Run Chart**

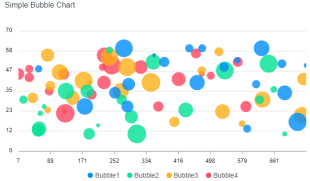
A **run chart** is a data line chart drawn over time. In other words, a run chart visually illustrates the process performance or data values in a time

sequence. Rather than summary statistics, seeing data across time yields a more accurate conclusion. A trend chart or time series plot is another name for a run chart. The plot below depicts dummy values of sales over a period of time.



**D) Bubble Chart**

Bubble charts scatter plots that display multiple circles (bubbles) in a two-dimensional plot. These are used to assess the relationships between three or more numeric variables. In a bubble chart, every single dot corresponds to one data point, and the values of the variables for each point are indicated by different positions such as horizontal, vertical, dot size, and dot colors.



**E) Heat Map**

A heat map is a colored graphical representation of multivariate data structured as a matrix of columns and rows. The heat map transforms the correlation matrix into color coding and represents these coefficients to visualize the strength of correlation among variables. It assists in finding the best features suitable for building accurate Machine Learning models.

