**WHAT IS EXPLORATORY DATA ANALYSIS?**

Exploratory Data Analysis (EDA) is a set of techniques that were developed by Tukey, John Wilder in 1970.This approach used to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques.

Today Data scientists and analysts spend most of their time in Data Wrangling and Exploratory Data Analysis also known as EDA.

**Why Exploratory Data Analysis or EDA** **is so important?**

Exploratory Data Analysis or EDA is used to take insights from the data. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques.

But, main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. It allows Analysts to understand the data better before making any assumptions. The outcomes of EDA help businesses to know their customers, expand their business and take decisions accordingly.

**Types of exploratory data analysis**

There are four primary types of EDA:

* **Univariate non-graphical.** This is simplest form of data analysis, where the data being analyzed consists of just one variable. Since it’s a single variable, it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
* **Univariate graphical.** Non-graphical methods don’t provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:
  + Stem-and-leaf plots, which show all data values and the shape of the distribution.
  + Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
  + Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
* **Multivariate nongraphical:**Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
* **Multivariate graphical:** Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

**Exploratory Data Analysis Tools**

**Pandas Profiling:** Pandas Profiling is an open source Python module which allows both non-technical users and data scientists to quickly perform EDA and present the information on a web-based interactive report. Using Pandas Profiling, you generate interactive graphs/charts and visualize the distribution of each variable in the dataset using just a few lines of code. Data scientists often use Pandas Profiling to save hours of time needed for the EDA process.

**Excel:** For many datasets, Excel is all that’s needed for data analysis. The advantages of Excel are that it’s easy to clean/manipulate the dataset using basic Excel functions, and it’s ultra-convenient to quickly create graphs/charts.  Although Excel is a paid program, Google Sheets is a free alternative that does exactly the same thing.

**Rattle (R Package):** R is complicated to learn with not so great documentation available, however, Rattle is the opposite. It is a graphical interface for R which allows in-depth data mining and requires no coding, no command line prompts - just clicks.

Rattle allows you to easily explore your data and create quick [visualizations](https://www.polymersearch.com/blog/data-visualization)(the process of turning datasets into charts, graphs, diagrams and other visuals. It can be used for analyzing data or presenting data.). You can also use it to clean & transform your data and build models. The tool is fast and ideal for handling big data for those who don’t know how to code.

**Trifacta:** allows you to prepare and explore any dataset on cloud data warehouse or cloud data lake houses through an interactive user interface. The tool uses in-built machine learning algorithms to guide you through the exploration of your data.

**KNIME**:

[**KNIME**](https://www.knime.com/) is a tool that allows you to dive deep into data processing without learning how to code. It is often used by data scientists, especially from the chem/biotech industry, for data processing and building production grade applications. It has plenty of features that’ll come in hand for exploratory data analysis including data cleansing and manipulation, merging datasets together, creating interactive visualizations and building models.

## DataPrep:

It is a tool on Python that saves countless hours of cleansing, preparing data and performing EDA. It works similarly to Pandas Profiling - that within a couple lines of code, you can plot a series of **interactive** graphs and distributions charts to get an overall sense of the data.

You can also find & analyze missing values and outliers within seconds using a few lines of code. This allows the user to be aware of data quality in each column and find possible reasons for these missing values or outliers.

Overall, DataPrep is a very powerful tool for cleansing data, analyzing missing variables, checking correlations and seeing the distribution of each variable.