Exploratory Data Analysis

# Introduction:

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

# Importance:

* To identify the most **important variables/features** in your dataset.
* Testing a **hypothesis** or checking assumptions related to the dataset.
* To check the **quality of data** for further processing and cleaning.
* Deliver **data driven insights** to the business stakeholders.
* Verify expected **relationship** actually exist in data.
* To find the **unexpected structure** or insights in the data.

# Data science process:

* Business Objective.
* Data Requirement.
* Data Collection.
* Exploratory Data Analysis.
* Modeling.
* Evaluating.
* Deployment.
* Monitoring.

# Categories of data: (format)

* **Structured data type:** csv file, excel file, database file
* **Unstructured data type:** text, image, video, audio

# Types of Data:

* **Numerical:** Numerical data is commonly called quantitative data. Data that has numbers or can be counted are numerical data, whereas data that involves words or vocabulary are non-numerical data (data that cannot be manipulated mathematically using. standard arithmetic operators).

Eg: height of people ·stock prices ·Weight in pounds· Length in inches ·Distance in miles ·Number of days in a year · A heatmap of a web page.

* Discreate data- integer based; often counts some events. Like how many purchases did customer make in a year? How many times did I flip heads? When values in a data set are countable and can only take certain values, it is called discrete data. For example, number of students in a class, number of players required in a team, etc. We can easily count the variables in a discrete data.
* Continuous Data- it has infinite number of possible values like Daily wind speed Freezer temperature · Weight of newborn babies · Length of customer service calls.
* **Categorical:** Categorical variables represent types of data which may be divided into groups. Examples of categorical variables are race, sex, age group, educational level, cast, gender and political parties
* **Ordinal data:**it is classified into categories within a [variable](https://www.scribbr.com/methodology/types-of-variables/) that have a natural rank order. However, the distances between the categories are uneven or unknown. Eg: the level of education, the range of income, or the grades, rating, low, high, average.

# Python Packages for EDA:

* Pandas:
* NumPy:
* Matplotlib:
* Seaborn:

# Graphs:

* Bar chart: relation, gaps
* Pie chart: proportion
* Histogram: defines range, no gaps
* Scatter plot: plot in dots
* Heatmap: By observing how cell colors change across each axis, you can observe if there are any patterns in value for one or both variables.