**CHAPTER – 1**

**EDA-Exploratory Data Analysis**

**Topics included in EDA**

1. What is EDA
2. Importance of EDA
3. Goals of EDA
4. Types of EDA
5. Feature Engineering.

**What is EDA?**

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. It involves analyzing and summarizing data to understand its main characteristics, often using visual methods. EDA helps to identify patterns, relationships, and anomalies in the data, which can guide further analysis and modeling decisions.

EDA is an iterative process and often leads to new questions and insights, guiding the direction of the data analysis. It is a critical step in any data science project as it helps in understanding the data better and making informed decisions.

**Importance of EDA**

Just like everything in this world, data has its imperfections. Raw data is usually skewed, may have outliers, or too many missing values. A model built on such data results in sub-optimal performance. In hurry to get to the [machine learning stage](https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/#:~:text=How%20to%20make%20EDA%20in,exploring%20and%20understanding%20the%20data.), some data professionals either entirely skip the exploratory data analysis (EDA) process or do a very mediocre job. This is a mistake with many implications, that includes generating inaccurate models, generating accurate models but on the wrong data, not creating the right types of variables in data preparation, and using resources inefficiently.

We will be using Pandas, Seaborn, numpy and Matplotlib libraries of Python to demonstrate various EDA techniques applied Dataset.

**Goals of EDA**

1. **Understand the data**: EDA helps you understand the structure, content, and quality of your data. It allows you to identify missing values, outliers, and other anomalies that could affect the analysis.
2. **Identify patterns and relationships**: EDA helps you identify patterns, trends, and relationships in the data. This can help you formulate hypotheses and guide further analysis.
3. **Guide data preprocessing**: EDA can help you decide how to preprocess your data, such as handling missing values, encoding categorical variables, and scaling features.
4. **Inform feature selection and engineering**: EDA can help you identify which features are most relevant to your analysis and whether you need to create new features through feature engineering.
5. **Detect outliers and anomalies**: EDA can help you identify outliers and anomalies in the data, which may need to be addressed before further analysis.
6. **Assess model assumptions**: EDA can help you assess whether the assumptions of your chosen model are met by the data.
7. **Communicate findings**: EDA can help you communicate your findings to stakeholders, providing insights into the data that may not be apparent from the raw data alone

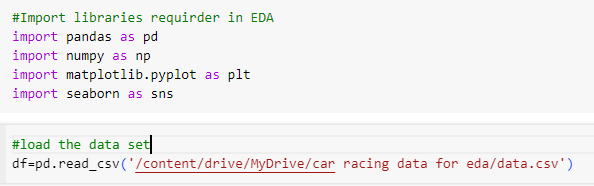
**Types of Exploratory Data Analysis**

Exploratory Data Analysis (EDA) can be broadly categorized into several types based on the techniques and methods used. Here are some common types of EDA:

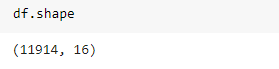
1. **Univariate Analysis**: This type of analysis focuses on one variable at a time. It involves examining the distribution of individual variables using summary statistics and visualizations like histograms, box plots, and bar charts.
2. **Bivariate Analysis**: Bivariate analysis looks at the relationship between two variables. It involves comparing the distribution of one variable across different levels of another variable using scatter plots, line graphs, or cross-tabulations.
3. **Multivariate Analysis**: Multivariate analysis involves analyzing more than two variables simultaneously. It can include techniques like principal component analysis (PCA), factor analysis, and cluster analysis to identify patterns and relationships among multiple variables.
4. **Statistical Analysis**: This type of analysis involves using statistical tests to make inferences about the data. It can include tests like t-tests, ANOVA, and correlation analysis to determine the significance of relationships and differences in the data.
5. **Visualization Techniques**: EDA often involves the use of various visualization techniques to explore the data. This can include plots like scatter plots, box plots, histograms, and heatmaps to visualize the distribution, relationships, and patterns in the data.
6. **Data Cleaning and Preprocessing**: EDA also includes the process of cleaning and preprocessing the data to prepare it for analysis. This can involve handling missing data, encoding categorical variables, and scaling or transforming features.
7. **Interactive EDA**: With the advent of interactive visualization tools and libraries, interactive EDA has become more common. This allows for more dynamic exploration of the data, enabling users to interactively manipulate and explore the data.

These types of EDA are not mutually exclusive, and often, a combination of these techniques is used to gain a comprehensive understanding of the data. The choice of techniques depends on the nature of the data and the goals of the analysis.

**Understanding dataset.**

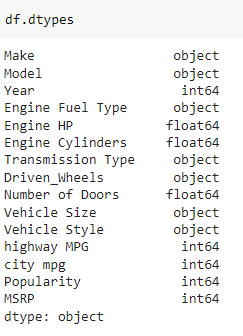
****

Its important to know the size of dataset



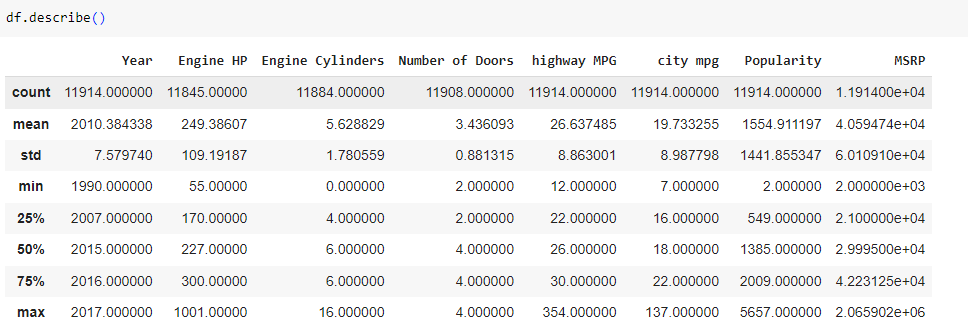
There are 11914 rows and 16 columns

To Know data type of each object

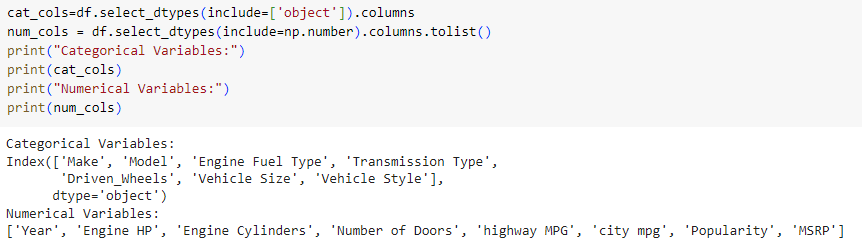


Based on data type we look for encoding and filling missing values .

**Describe data with some statistic info**

****

**Separate categorical and numerical columns**

****

**Types of encoding techniques**

1. **One-Hot Encoding**:
   * **Concept**: It creates binary columns for each category and marks the presence of a category with 1 and absence with 0.
   * **Example**: If you have a "Color" column with categories "Red," "Green," and "Blue," it will create three new columns: "Color\_Red," "Color\_Green," and "Color\_Blue."
   * **Use Cases**: Useful when there is no ordinal relationship between categories, and you want to avoid implying an order.
2. **Label Encoding**:
   * **Concept**: It assigns a unique integer to each category in a categorical variable.
   * **Example**: If you have a "Size" column with categories "Small," "Medium," and "Large," it might be encoded as (0, 1, 2) respectively.
   * **Use Cases**: Works well when there is an ordinal relationship between categories.
3. **Ordinal Encoding**:
   * **Concept**: Similar to label encoding but considers the order of the categories.
   * **Example**: If you have a "Temperature" column with categories "Low," "Medium," and "High," it might be encoded as (0, 1, 2) respectively, reflecting their order.
   * **Use Cases**: Suitable for categorical variables with clear ordinal relationships.
4. **Binary Encoding**:
   * **Concept**: Converts each integer to binary digits and uses those as separate features.
   * **Example**: If "Color" is label encoded as (0, 1, 2) and then binary encoded, "Color\_0" might become (0, 0), "Color\_1" might become (0, 1), and "Color\_2" might become (1, 0).
   * **Use Cases**: Useful for reducing the number of dimensions created by one-hot encoding.
5. **Target Encoding**:
   * **Concept**: It replaces each category with the mean (or other statistical measure) of the target variable for that category.
   * **Example**: If you have a "City" column and a target variable "Sales," target encoding for "City" calculates the mean sales for each city and replaces the city name with that mean.
   * **Use Cases**: Can be effective when there is a strong relationship between the categorical variable and the target variable
6. **Frequency Encoding**:
   * **Concept**: Replaces each category with the frequency of that category in the dataset.
   * **Example**: If "City" has three occurrences of "New York" and two occurrences of "Los Angeles," "New York" might be encoded as 3 and "Los Angeles" as 2.
   * **Use Cases**: Useful when the frequency of a category is relevant to the analysis

**To apply all these encoding technique we should have a knowledge of feature and field on which we are working {a field expert may require to know the relation }**

**Refrences**

1. [**analytics vidya (how eda works )**](https://www.analyticsvidhya.com/blog/2021/08/exploratory-data-analysis-and-visualization-techniques-in-data-science/)
2. [**gfg {eda and python libraries }**](https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/)
3. [**kaggle data set**](https://www.kaggle.com/code/imoore/intro-to-exploratory-data-analysis-eda-in-python)
4. [**categorical encoding**](https://towardsdatascience.com/eda-feature-engineering-and-encoding-categorical-data-75bed7af8692)
5. [**steps of EDA**](https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/)
6. [**google colab**](https://colab.research.google.com/drive/1MNSVjeoDQH3ZiHJiunCNDm6ii0uONAvY?usp=sharing)