Assignment 1

Aim: Exploratory Data Analysis (EDA) on CarDekho Dataset

Objective: To perform Exploratory Data Analysis (EDA) on the CarDekho dataset and derive insights using Python.

Theory: Exploratory Data Analysis (EDA) is a crucial step in the machine learning process where data is visualized and analyzed to identify patterns and distributions. EDA helps in identifying outliers, handling missing values, and understanding relationships among various features. In this assignment, we conduct EDA on the CarDekho dataset to identify key factors influencing used car prices.

Importance of EDA in Machine Learning: EDA is an essential step in machine learning because it helps with:

- Understanding the structure and statistics of the dataset.
- Detecting missing values and outliers that may affect model accuracy.
- Identifying correlations between variables.
- Selecting the right preprocessing steps such as feature selection, normalization, and encoding categorical variables.
- Enhancing model interpretability by making data patterns and trends explicit.
- Ensuring high-quality input data, which improves model performance.

Dataset: The dataset used for this assignment is **CarDekho Dataset**. It contains the following features:

- **Car Name** - The name of the car.
- **Brand** - The manufacturer of the car.
- **Model** - The specific model of the car.
- **Vehicle Age** - The number of years since the car was manufactured.
- **Km Driven** - The total kilometers the car has been driven.
- **Fuel Type** - The type of fuel used (Petrol, Diesel, CNG, Electric, etc.).
- **Transmission Type** - Whether the car has an automatic or manual transmission.
- **Mileage** - The fuel efficiency of the car.
- **Engine** - The engine capacity in CC.
- **Max Power** - The maximum power output of the car.
- **Seats** - The number of seats in the car.
- **Seller Type** - The type of seller (Dealer, Individual, etc.).
- **Selling Price** - The price at which the car is being sold.

Steps of Implementation:

1. **Importing Libraries:**

- Analysis and visualization are performed using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn.

2. **Loading and Exploring the Dataset:**
   - The data is read using Pandas and explored using .head(), .tail(), and .info() to understand its structure.
3. **Checking for Missing Values and Duplicates:**
   - Missing values are detected and handled using .isnull().sum().
   - Duplicate records are removed to prevent redundancy.
4. **Descriptive Statistics:**
   - Summary statistics such as mean, median, and standard deviation are computed.
5. **Data Visualization:**
   - Histograms and distribution plots to understand the distribution of numerical attributes.
   - Boxplots to detect outliers in attributes like Selling Price, Km Driven, and Mileage.
   - Bar charts for categorical variables such as Brand, Model, and Fuel Type.
6. **Feature Analysis:**
   - Investigating how **vehicle age** affects selling price.
   - Analyzing the relationship between **fuel type** and car price trends.
   - Identifying correlations between **engine capacity, max power, mileage, and selling price** using heatmaps.

Conclusion:

- **Vehicle Age and Selling Price:** Older cars tend to have lower selling prices.
- **Impact of Fuel Type:** Certain fuel types (e.g., diesel cars) may have higher resale value depending on demand.
- **Car Specifications and Pricing:** Engine capacity, max power, and mileage significantly influence car prices.
- **Outlier Detection:** Boxplots reveal extreme values in selling prices and driven kilometers, which might need further analysis.

References:

- https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/
- GitHub Repository: https://github.com/Utkarsh-Rane43/ML-LAB---122B1F110