Assignment 2

Aim: Implementation of Linear Regression on CarDekho Dataset

Objective: To implement and evaluate a Linear Regression model using Python to predict car prices based on different features in the dataset.

Theory: Linear Regression is a fundamental supervised learning algorithm used to predict numerical values based on input features. It assumes a linear relationship between the target variable (dependent variable) and the feature variables (independent variables). The objective of this assignment is to perform data preprocessing, train a Linear Regression model, validate its performance, and analyze the results based on the CarDekho dataset.

Importance of Linear Regression:

- **Predictive Analysis:** Helps estimate car prices based on various car attributes.
- **Feature Relationships:** Identifies relationships between vehicle age, mileage, engine power, and car price.
- **Efficiency:** Computationally less expensive and easy to implement.
- **Baseline Model:** Serves as a foundation for more complex regression models.

Dataset: The dataset used for this assignment is **CarDekho Dataset**. It contains various features that describe used cars, such as:

- **Vehicle Age:** Number of years since the car was manufactured.
- **Km Driven:** Total kilometers the car has been driven.
- **Fuel Type:** 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.
- **Selling Price:** Target variable representing the price at which the car is being sold.

Steps of Implementation:

1. **Importing Libraries:**
    - Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-Learn are used for data handling, visualization, and model training.
2. **Loading the Dataset:**
    - The dataset is loaded using Pandas, and an initial exploration is conducted using .head(), .info(), and .describe().
3. **Data Preprocessing:**
    - Handling missing values by filling categorical columns with mode and numerical columns with median.

- Encoding categorical variables using one-hot encoding.
- Defining the target variable (Selling Price) and feature set (X).
- Splitting the data into training and test sets (80% training, 20% testing).

4. **Training the Model:**
   - A Linear Regression model is trained on the dataset using Scikit-Learn.
5. **Making Predictions:**
   - The trained model is used to predict car prices on the test dataset.
6. **Model Evaluation:**
   - Performance measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared Score (R^2) are calculated.
7. **Visualization of Results:**
   - A scatter plot is used to compare actual and predicted car prices.

Conclusion:

- **Linear Regression Model Performance:** The model effectively predicts car prices based on key vehicle attributes.
- **Evaluation Metrics:**
  - **MAE:** Measures the average absolute difference between actual and predicted car prices.
  - **MSE & RMSE:** Indicate the spread of error.
  - **R^2 Score:** Determines how well independent variables explain price variation.
- **Visual Representation:** A scatter plot helps understand the correlation between predicted and actual car prices.