📄 Car Price Prediction Using Multiple Linear Regression

1. Project Overview

This project aims to build a **Multiple Linear Regression (MLR) model** to predict **car prices** based on various numerical and categorical features. The dataset used is sourced from Kaggle and contains information such as car year, engine size, mileage, brand, fuel type, transmission, and ownership details.

Objective:
To analyze the dataset, perform exploratory data analysis (EDA), and develop a regression model that accurately predicts car prices using multiple independent variables.

---

2. Dataset Description

Dataset Path:
`/kaggle/input/car-prepiction/car_price_dataset.csv`

Target Variable:

- `Price` – Selling price of the car

Independent Variables:

- Numeric: `Year`, `Engine_Size`, `Mileage`, `Doors`, `Owner_Count`
- Categorical: `Brand`, `Model`, `Fuel_Type`, `Transmission`

---

3. Exploratory Data Analysis (EDA)

EDA was performed to understand data distribution and relationships:

- **Univariate Analysis:**
  - Histograms and box plots identified distributions and outliers.
- **Bivariate Analysis:**
  - Scatter plots showed strong relationships:

        - `Year` vs `Price` → Strong positive correlation
        - `Mileage` vs `Price` → Strong negative correlation
- **Correlation Analysis:**
    - A heatmap confirmed that `Year` and `Mileage` are the most influential predictors.
- **Missing Values:**
    - Dataset was checked and handled appropriately.

---

4. Data Preprocessing

- Categorical variables were converted to numerical form using **One-Hot Encoding**.
- The dataset was split into:
    - **80% Training data**
    - **20% Testing data**

---

5. Model Building

A **Multiple Linear Regression** model was trained using `scikit-learn`.

Model Formula:
[
Price = b_0 + b_1(Year) + b_2(Engine_Size) + b_3(Mileage) + \dots
]

Where:

- (b_0) is the intercept
- (b_1, b_2, b_3) are coefficients learned from data

---

6. Model Evaluation

The model was evaluated using standard regression metrics:

- **MAE (Mean Absolute Error):** Measures average absolute error
- **MSE (Mean Squared Error):** Penalizes large errors
- **RMSE (Root Mean Squared Error):** Error in original price units
- **R^2 Score:** Explains variance in car prices

Residual plots and Actual vs Predicted plots were used to validate model assumptions.

---

7. Results & Insights

- Newer cars generally have higher prices.
- Higher mileage significantly reduces car price.
- Engine size showed weaker influence compared to year and mileage.
- The model performs reasonably well, but performance can improve with more data.

8. Conclusion & Future Work

The Multiple Linear Regression model successfully predicts car prices and aligns well with EDA insights.

Future Improvements:

- Apply feature scaling
- Use Ridge or Lasso Regression
- Perform cross-validation
- Try non-linear models (Random Forest, XGBoost)

---

9. Tools & Technologies

- **Python**
- **Pandas, NumPy**
- **Matplotlib, Seaborn**
- **Scikit-learn**
- **Kaggle Notebook Environment**