Case Study: Predicting Car Miles per Gallon (MPG) Using Linear Regression & Machine Learning

---

1. Introduction

The **Car-MPG** dataset contains various features of cars, such as the number of cylinders, engine displacement, horsepower, weight, acceleration, and model year, among others. The goal of this case study is to predict the **Miles Per Gallon (MPG)** of a car based on its features using linear regression.

---

2. Problem Statement

Predict the **MPG** of a car given its characteristics using a linear regression model. Linear regression will help us understand the relationship between the independent variables (e.g., weight, horsepower) and the dependent variable (MPG).

---

3. Dataset Overview

The dataset contains the following columns:

- mpg: The miles per gallon of the car (dependent variable).
- cylinders: The number of cylinders in the car.
- displacement: The engine displacement in cubic inches.
- horsepower: The horsepower of the car.
- weight: The weight of the car in pounds.
- acceleration: The car's acceleration (in seconds for 0 to 60 mph).
- model_year: The model year of the car.
- origin: The origin of the car (USA, Europe, or Japan).
- name: The name of the car (which we will not use for prediction).

---

4. Tentative Steps

Here's an outline of the steps we will follow in the case study:

1. **Data Loading and Preprocessing**
    - Load the dataset and check for missing values.
    - Handle missing values either by dropping or filling them.
    - Convert categorical variables (e.g., origin) into numeric representations.

2. **Exploratory Data Analysis (EDA)**
    - Analyze the relationships between the features and the target variable (MPG).
    - Visualize correlations and check for multicollinearity.
    - Create scatter plots to investigate how variables like horsepower, weight, and cylinders affect MPG.
3. **Data Splitting**
    - Split the data into **training** and **test** sets (e.g., 80% training, 20% testing).
4. **Feature Scaling**
    - Apply **standardization** or **normalization** (if needed) to scale the features for the linear regression model.
5. **Model Building**
    - Apply **linear regression** to train the model on the training dataset.
6. **Model Evaluation**
    - Evaluate the model's performance using metrics such as **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R-squared (R^2)**.
    - Visualize the **actual vs predicted MPG** using scatter plots.
7. **Residual Analysis**
    - Perform a residual analysis to check the model's assumptions (e.g., linearity, homoscedasticity).
8. **Model Refinement (Optional)**
    - Based on the model's performance, apply further steps for model improvement, such as feature engineering or regularization techniques like Ridge/Lasso regression if overfitting is observed.
9. **Conclusion**
    - Summarize the model's performance and suggest future improvements or potential applications of the model.

---

5. Data Preprocessing

Before applying the linear regression model, we need to preprocess the dataset.

- **Handle missing values**: We check for any missing data and decide whether to remove or fill in missing values.

- **Convert categorical variables**: The origin column is categorical and needs to be converted into numeric values using techniques like one-hot encoding or label encoding.
- **Feature scaling**: If necessary, we standardize or normalize continuous variables such as weight and horsepower.

---

6. Exploratory Data Analysis (EDA)

We will perform an exploratory data analysis to understand the relationships between features and the target variable (MPG). Key steps in this phase include:

- **Visualizing correlations**: We can use heatmaps to visualize how different features correlate with the MPG.
- **Scatter plots**: Scatter plots will help us see the relationship between the dependent variable (MPG) and independent variables like weight, horsepower, and displacement.

---

7. Model Building

We'll now build a regression model to predict the MPG (Linear Regression & Machine learning techniques)

1. **Split the data**: We split the data into training and testing sets. Typically, we use 70% for training and 30% for testing.
2. **Model training**: We apply linear regression using the training data.
3. **Model evaluation**: We evaluate the model using metrics such as:
    - **Mean Absolute Percentage Error (MAPE)**
    - **Root Mean Squared Error (RMSE)**
    - **Root Mean Squared Percentage Error (RMSPE)**
    - **R-squared (R^2)**: This tells us how well the independent variables explain the variability in the target variable (MPG).

---

8. Model Evaluation

- **Model performance**: We analyze how well the model performs on the testing dataset by comparing predicted MPG values to the actual values.
- **Residual analysis**: We check the residuals (differences between predicted and actual values) to ensure there is no pattern (indicating that the linear regression assumptions hold).

---

9. Interpretation of Results

- **Coefficients**: The coefficients of the linear regression model tell us how much each feature influences the predicted MPG. For example, if the coefficient for horsepower is negative, it implies that higher horsepower cars tend to have lower MPG.
- **Model significance**: We also analyze statistical significance (p-values) of the coefficients to ensure that the relationships we observe are meaningful and not due to random chance.

---

10. Model Improvement (if necessary)

- **Feature engineering**: We may try adding new features or transforming existing ones.
- **Regularization**: If the model shows overfitting, we may use techniques like **Ridge** , **Lasso** or **Elastic Net regression** to regularize the model.

---

11. Conclusion

Based on the results, we conclude whether the linear regression model is a good fit for predicting MPG and if the features we selected were appropriate. We can also suggest improvements for future models.