Exploratory Data Analysis of Auto-mpg Using Machine Learning Model:

Overview of the Dataset:

The dataset has 398 entries and 9 columns, including numerical features like mpg (miles per gallon), cylinders, displacement, weight, and acceleration, as well as categorical features like horsepower (which has been read as an object, indicating it may have non-numeric values) and car name.

Objective:

The aim of conducting EDA on the provided auto-mpg dataset is to identify, among other things, patterns, anomalies, tests of hypothesis
and verification of assumptions using summary statistics and graphical representations. EDA conducted on the auto-mpg dataset has the following aims:

a. **Identification of the Structure of the Data:** It includes identification of the data type, distribution, problems in data, such as missing and inconsistencies in data.
b. **Relationships among Features:** How Horsepower, Weight, or Cylinders affect mpg.
c. **Outlier Detection:** Find any outliers or anomalies that may influence model performance.
d. **Data Cleaning**: Clean issues such as missing values or incorrect data types. This will ensure the dataset is ready for analysis.
e. **Feature Importance:** Show the most important features in the prediction of MPG, which can be helpful in narrowing down the modeling effort.
f. **Initial Insights:** Create initial insights and hypotheses about data that can drive further analysis or modeling effort.

EDA is at the core of any data analysis workflow and basically sets the base for higher-end modeling techniques like regression analysis.

Steps for Building a Split Test Model with Regression Analysis:

a. **Data Cleaning:** Convert horsepower to numeric, handle missing values.
b. **Feature Selection:** Identify features for the regression model.
c. **Data Splitting:** Divide data into training and testing sets.
d. **Model Training:** Train the regression model on the training set.
e. **Model Evaluation:** Use metrics like Mean Squared Error (MSE) and R^2 to evaluate the model on the testing set.

Next Steps:

- **Feature Selection:** Use cylinders, displacement, horsepower, weight, acceleration, model year, and origin for the regression model.
- **Data Splitting:** Split the dataset into 80% training and 20% testing sets.
- **Model Training:** Train a linear regression model.

The linear regression model results:

MSE: 10.71

R^2 Score: 0.79

Interpretation:

MSE measures the average squared difference between actual and predicted values, with a lower value indicating better performance.

An R^2 Score of 0.79 means the model explains 79% of the variance in mpg, showing a good fit.