॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
INSTITUTE OF GEOINFORMATICS

Name: Alviya Ali

PRN: 24070243005

Analysis of Automobile MPG Dataset

INDEX

Objective

The objective of this report is to analyze key factors influencing a vehicle's miles per gallon (MPG) using statistical and machine learning methods. By visualizing data relationships and fitting predictive models, we aim to gain insights into which factors most impact fuel efficiency.

Introduction

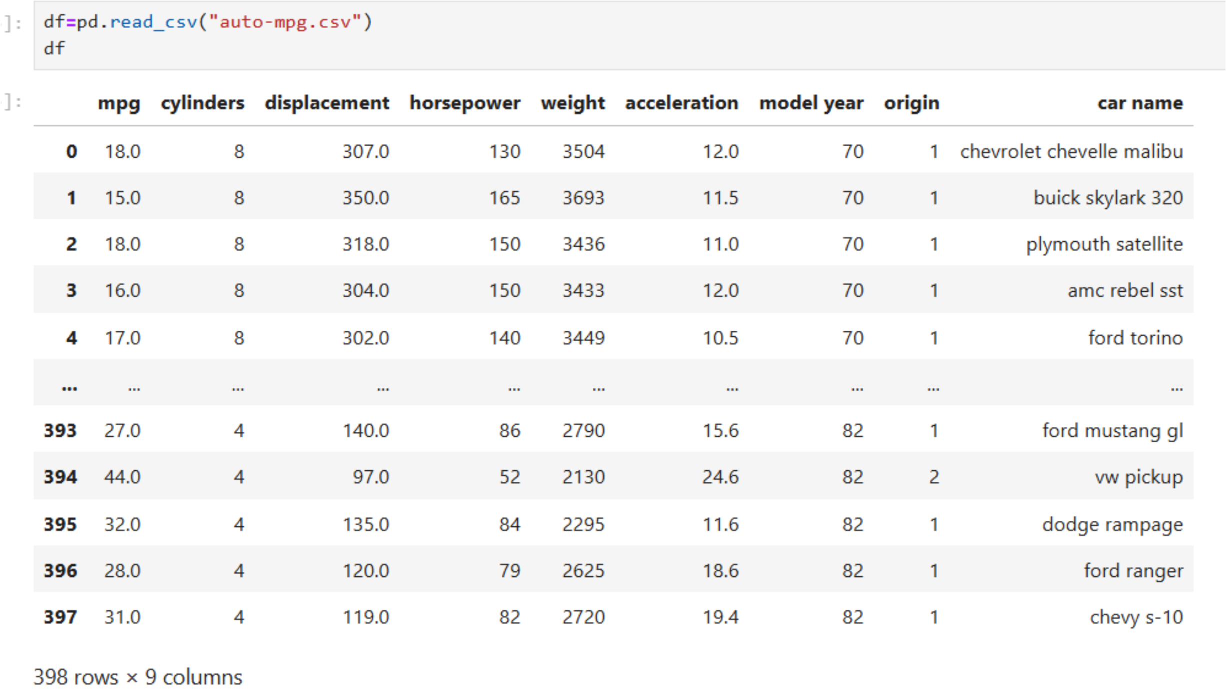
This report provides a comprehensive analysis and comparison of Multiple Regression Model. It contains information about various cars manufactured in the late 1970s and early 1980s, primarily focused on their fuel efficiency, measured in miles per gallon (MPG). The analysis includes a detailed examination of the dataset for regression analysis and to predict MPG based on a variety of features.

Key Attributes of the Dataset:

1. **Miles Per Gallon (MPG):** The target variable representing the car's fuel efficiency.
2. **Cylinders:** The number of cylinders in the car's engine, which influences performance and efficiency.
3. **Displacement:** The total volume of all the engine's cylinders, typically measured in cubic inches.
4. **Horsepower**: The power output of the car's engine.
5. **Weight:** The weight of the car in pounds.
6. **Acceleration:** The time it takes for the car to reach a certain speed from rest.
7. **Model Year:** The year the car model was manufactured, providing temporal context.
8. **Origin:** The geographical origin of the car (e.g., USA, Europe, Asia).
9. **Car Name:** A descriptive identifier of the car's make and model.

Data Collection and Sources

The dataset is collected from Kaggle and was available through the UCI Machine Learning Repository, a renowned repository for datasets commonly used in machine learning and data science research. It is often used for educational and research purposes, given its accessibility and historical significance. The UCI version of the dataset includes 398 entries with 9 attributes.



```
]: df=pd.read_csv("auto-mpg.csv")
   df
```

]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino
...	...	...	...	...	...	...	...	...	...
393	27.0	4	140.0	86	2790	15.6	82	1	ford mustang gl
394	44.0	4	97.0	52	2130	24.6	82	2	vw pickup
395	32.0	4	135.0	84	2295	11.6	82	1	dodge rampage
396	28.0	4	120.0	79	2625	18.6	82	1	ford ranger
397	31.0	4	119.0	82	2720	19.4	82	1	chevy s-10

398 rows × 9 columns

Statistical Insights

1. **Fuel Efficiency (MPG):**

- Mean: 23.5 MPG suggests the average vehicle in the dataset is moderately fuel-efficient.
- Range: MPG ranges from 9.0 to 46.0, indicating a diverse set of vehicles from low-efficiency to highly efficient ones.
- Dispersion: A standard deviation of 7.8 highlights moderate variability in MPG across the dataset.

2. **Weight:**

- The mean weight of vehicles is 2977.6 pounds, with a minimum of 1613 pounds and a maximum of 5140 pounds.
- Heavier vehicles tend to have lower MPG, which could be explored further in scatterplots or regression models.

3. **Horsepower:**

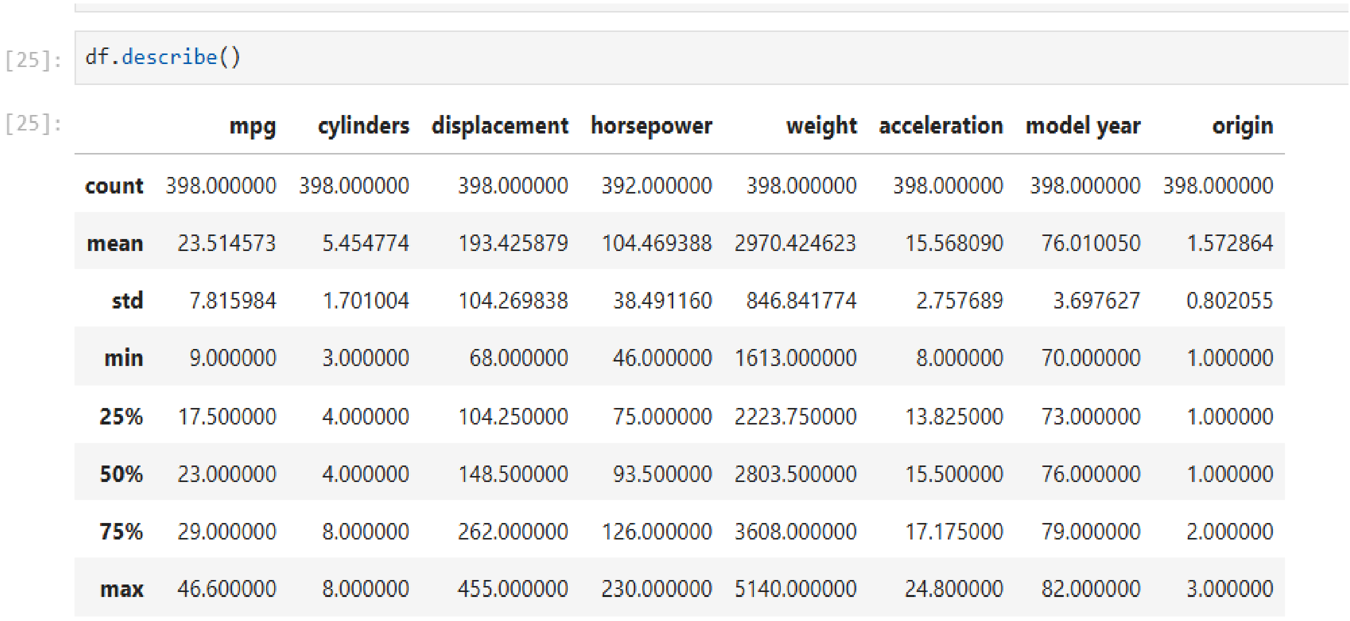
- The mean horsepower is 104.2, with a high variability (std = 30.5), reflecting a mix of low-performance and high-performance vehicles.

4. **Cylinders:**

- The 25th percentile at 4 cylinders and the 75th percentile at 8 cylinders confirm that most cars have 4 or 8 cylinders.

5. **Acceleration:**

- The dataset's acceleration values (mean: 15.6) indicate how quickly cars can reach a certain speed, varying between 8.0 and 24.8 seconds.



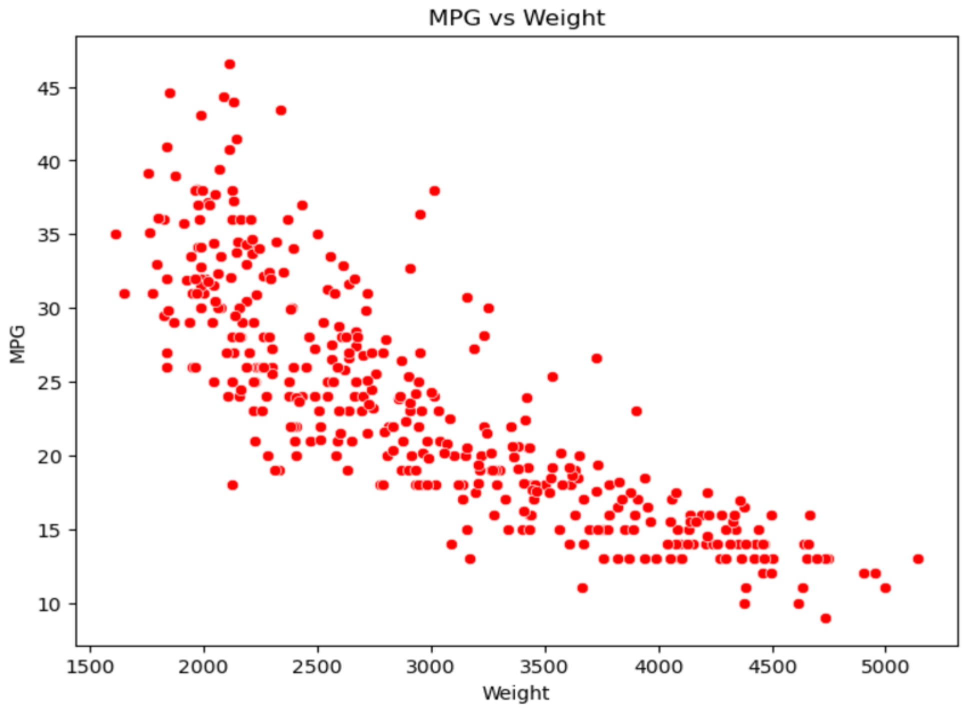
```
[25]: df.describe()
```

[25]:

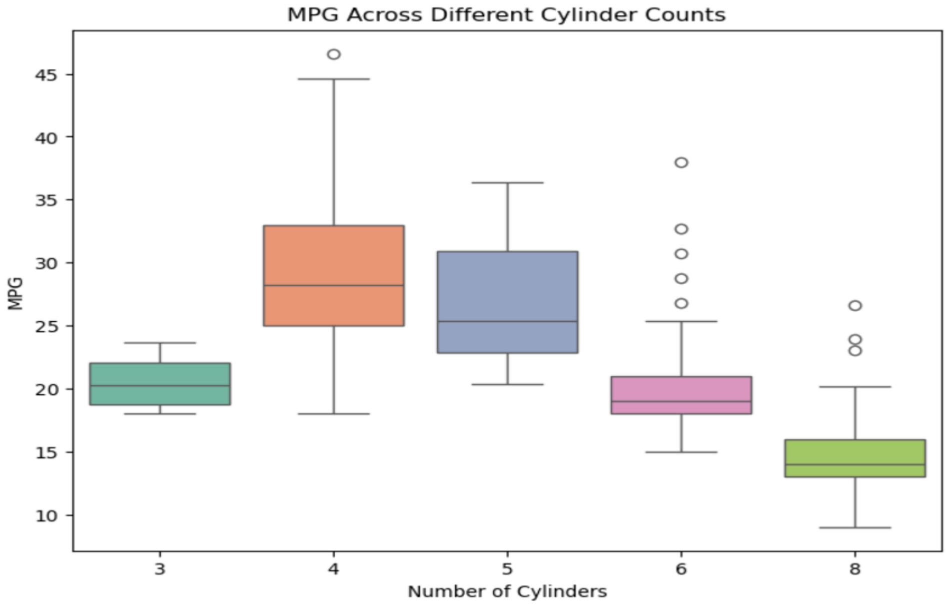
	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
count	398.000000	398.000000	398.000000	392.000000	398.000000	398.000000	398.000000	398.000000
mean	23.514573	5.454774	193.425879	104.469388	2970.424623	15.568090	76.010050	1.572864
std	7.815984	1.701004	104.269838	38.491160	846.841774	2.757689	3.697627	0.802055
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	70.000000	1.000000
25%	17.500000	4.000000	104.250000	75.000000	2223.750000	13.825000	73.000000	1.000000
50%	23.000000	4.000000	148.500000	93.500000	2803.500000	15.500000	76.000000	1.000000
75%	29.000000	8.000000	262.000000	126.000000	3608.000000	17.175000	79.000000	2.000000
max	46.600000	8.000000	455.000000	230.000000	5140.000000	24.800000	82.000000	3.000000

Visualization and different distribution

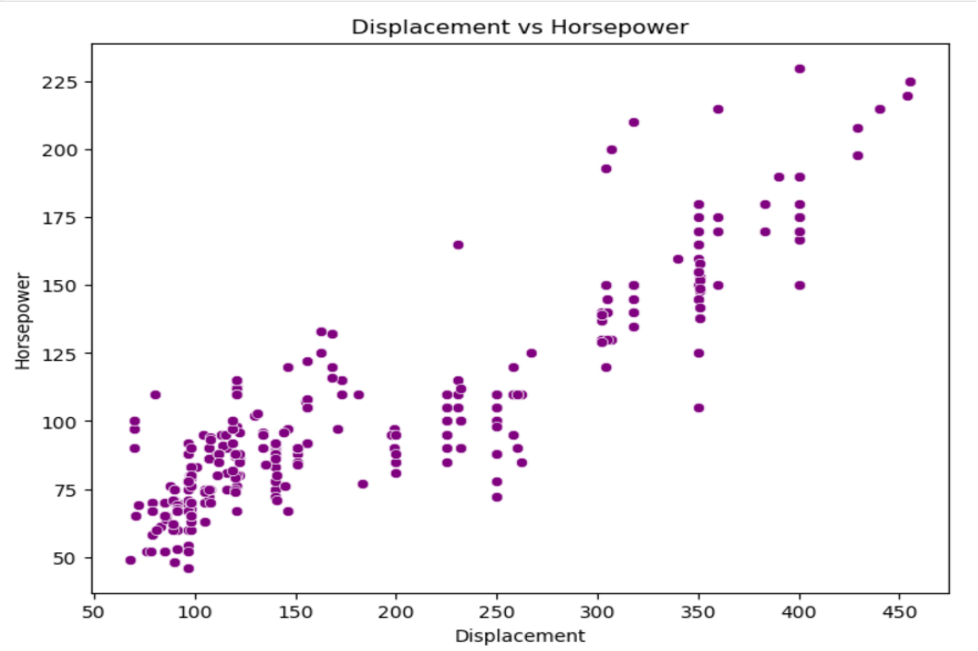
- Scatter plot between mpg and weight.



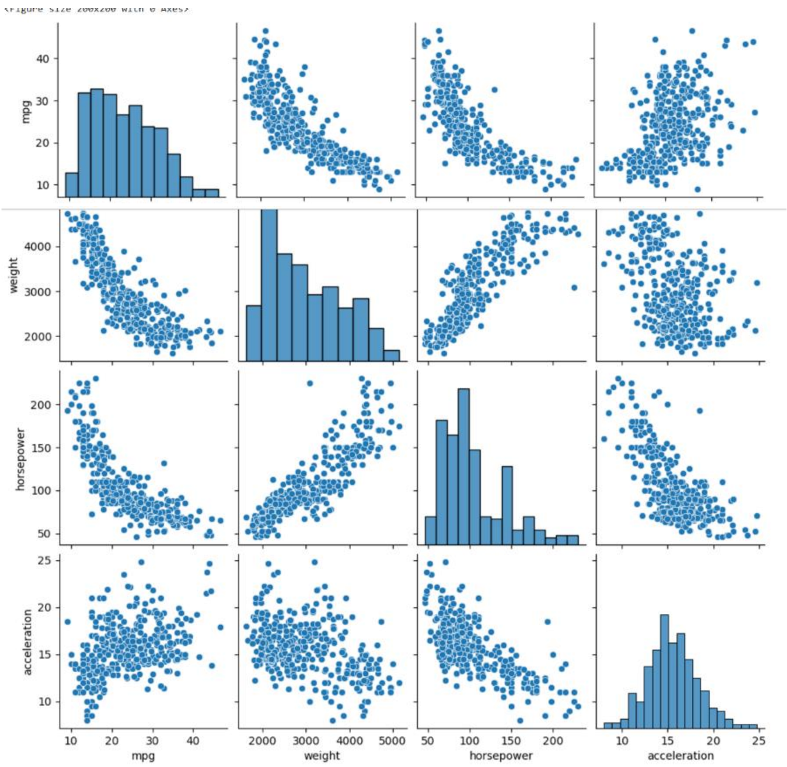
- Boxplot for mpg across cylinders.



- Scatter plot of displacement vs horsepower.

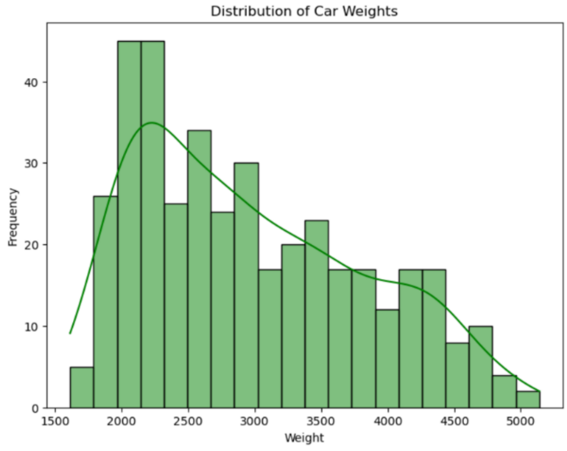


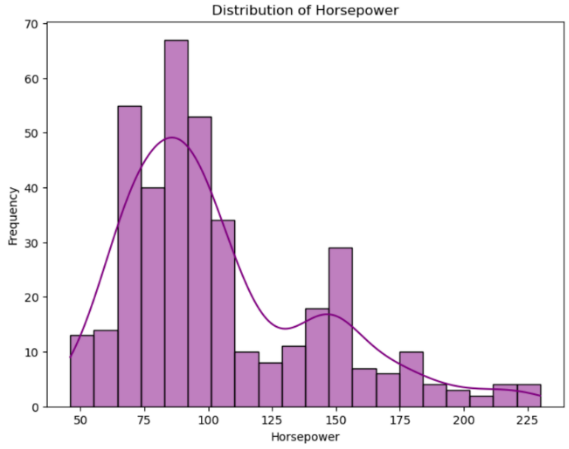
- Pairplot for mpg, weight, horsepower, and acceleration.



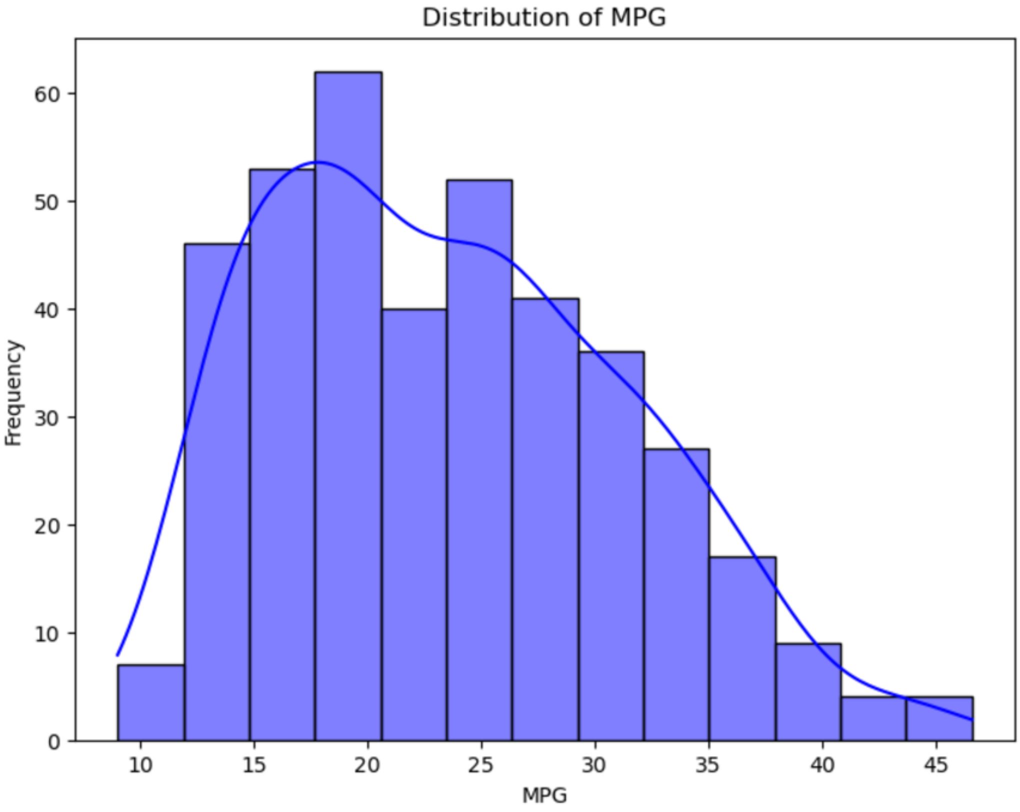
```
<Figure size 200x200 with 0 Axes>
```

- Histogram for car weight and Horsepower.





- Histogram for MPG.



Histogram Analysis: Distribution of Car Weight, Horsepower and MPG.

- **Distribution of Car Weight**

  The distribution of car weights is approximately right-skewed. Most cars cluster in the range of 2000–3000 pounds. Fewer cars are heavier than 4000 pounds, and very few exceed 5000 pounds. The peak of the distribution (mode) is around 2500 pounds, representing the most common car weight in the dataset.

1. Mean weight is likely higher than the mode due to the long tail on the heavier side.
2. Standard deviation is significant, reflecting the wide variety of car weights, from compact cars to heavy trucks.

- **Distribution of Horsepower**

  The distribution of horsepower is right-skewed. Most vehicles have horsepower in the range of 70–150. Few cars exceed 200 horsepower, likely representing performance-oriented models. A mode around 100 horsepower indicates the typical engine power of vehicles during this period.

1. The mean horsepower is pulled upward by the long tail of high-performance vehicles.
2. Standard deviation is moderate, reflecting some variability in engine power.

- **Distribution of MPG**

  The MPG distribution is left-skewed. The majority of vehicles achieve MPG in the range of 20–30. Fewer cars have low MPG (<15), typically heavier, high-horsepower vehicles. High-MPG outliers (>40 MPG) likely represent exceptionally efficient economy cars.

1. The median MPG is higher than the mean due to the skewness.
2. The wide range (9 to 46 MPG) highlights the diversity in vehicle efficiency.

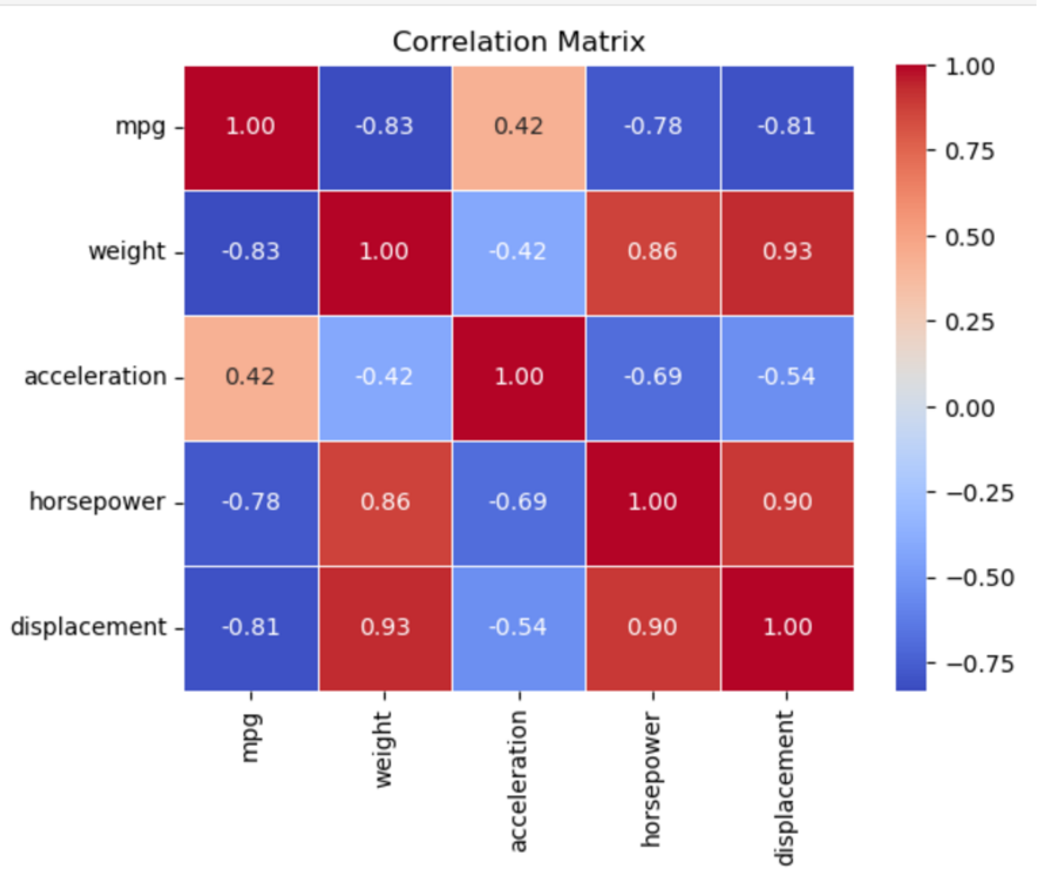
Correlation and Regression

Strongest Correlations:

- Weight and Displacement (0.93): Larger cars tend to have larger engines.
- MPG and Weight (-0.83): Lighter cars are more fuel-efficient.

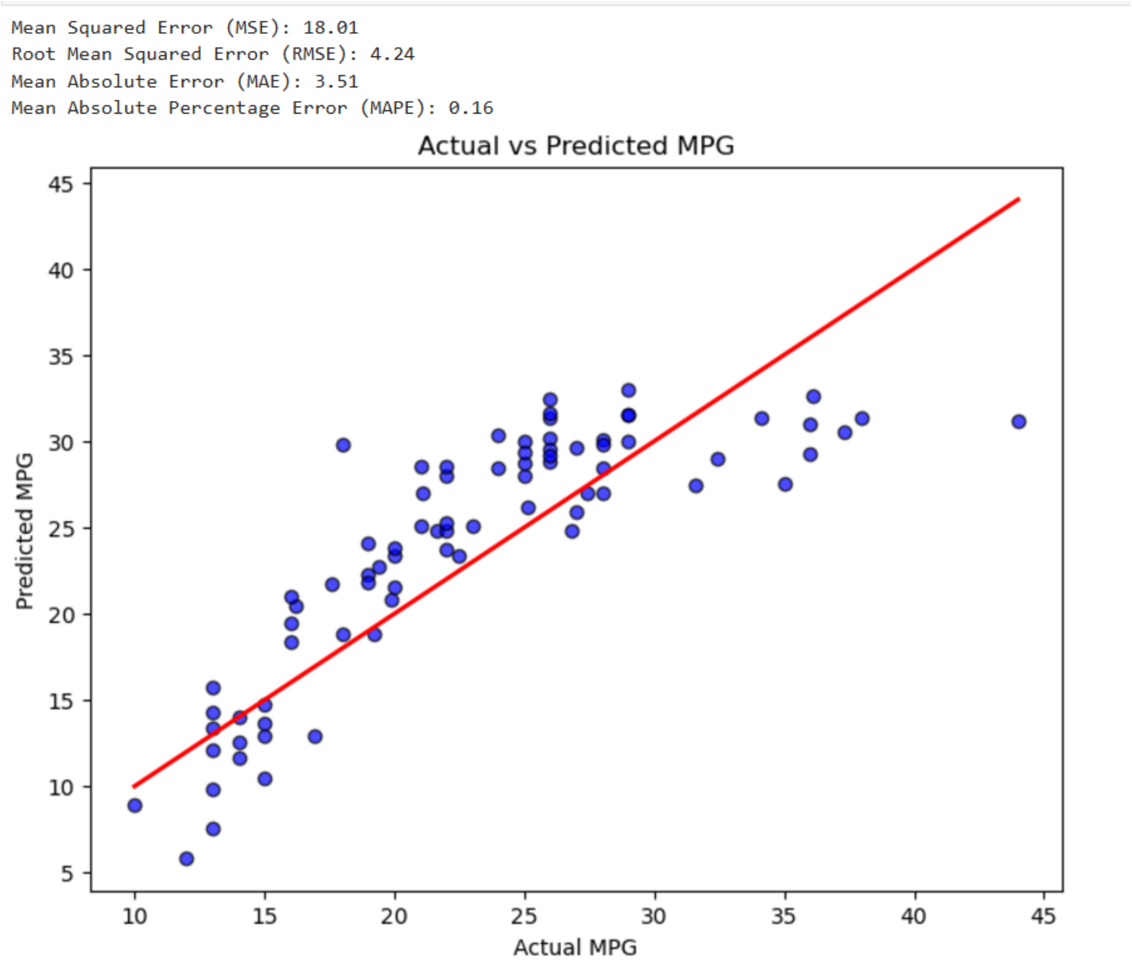
Weakest Correlations:

- Acceleration and Horsepower (-0.15): Indicates that acceleration times are not heavily influenced by horsepower in this dataset.



Regression model for independent variable (weight) and dependent variable (mpg).

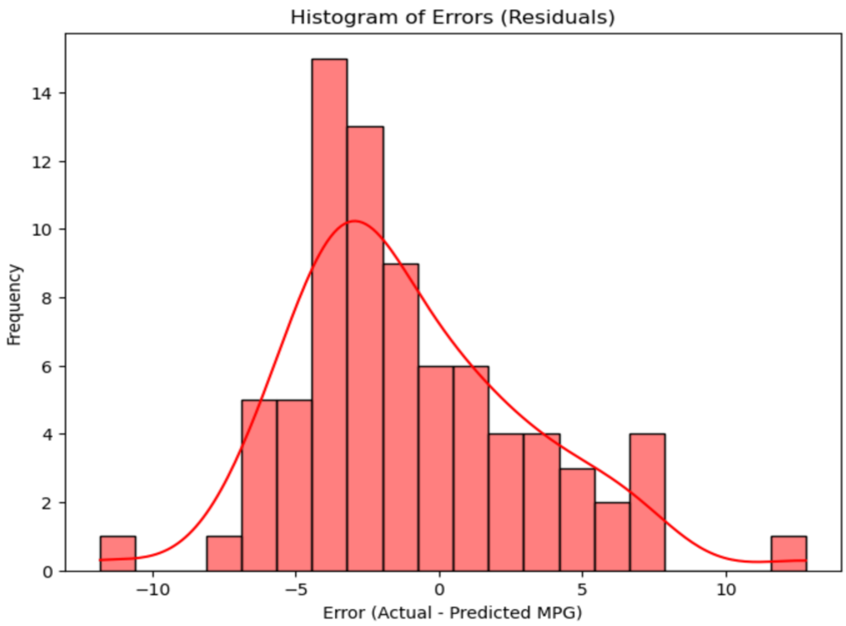
The linear regression model for MPG reveals that weight, horsepower, and displacement negatively impact fuel efficiency, while the model year positively correlates with better fuel economy. These insights underscore the trade-offs between vehicle performance and fuel efficiency. The model also highlights the importance of technological advancements in improving fuel efficiency over time. By examining the coefficients and evaluating the model’s performance through metrics like R², MSE, and RMSE, we’ve gained a deeper understanding of how vehicle characteristics influence fuel efficiency. This information is valuable not only for car manufacturers looking to optimize fuel efficiency but also for consumers who are interested in making environmentally and economically sound purchasing decisions.



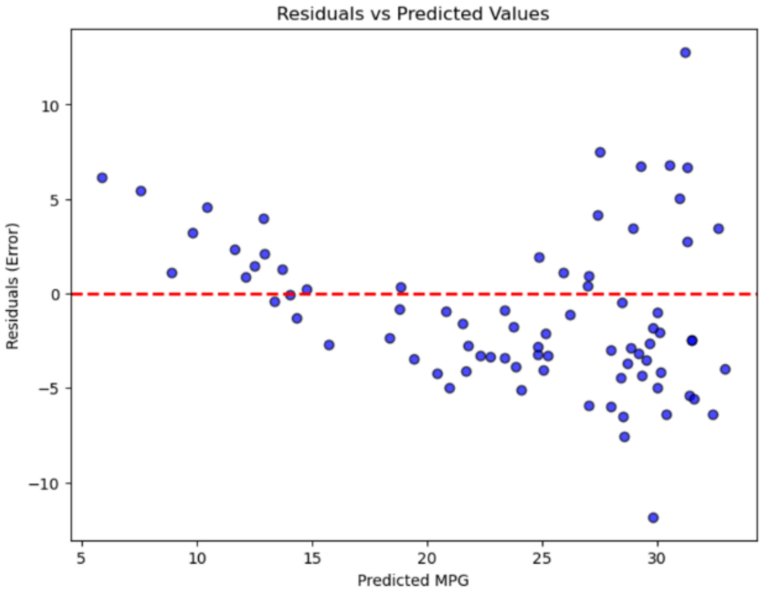
```
Mean Squared Error (MSE): 18.01
Root Mean Squared Error (RMSE): 4.24
Mean Absolute Error (MAE): 3.51
Mean Absolute Percentage Error (MAPE): 0.16
```

Error Analysis

A **bell-shaped** histogram centered around zero indicates that the residuals are approximately normally distributed, suggesting that the model is appropriate for the data and the assumptions of linear regression hold.

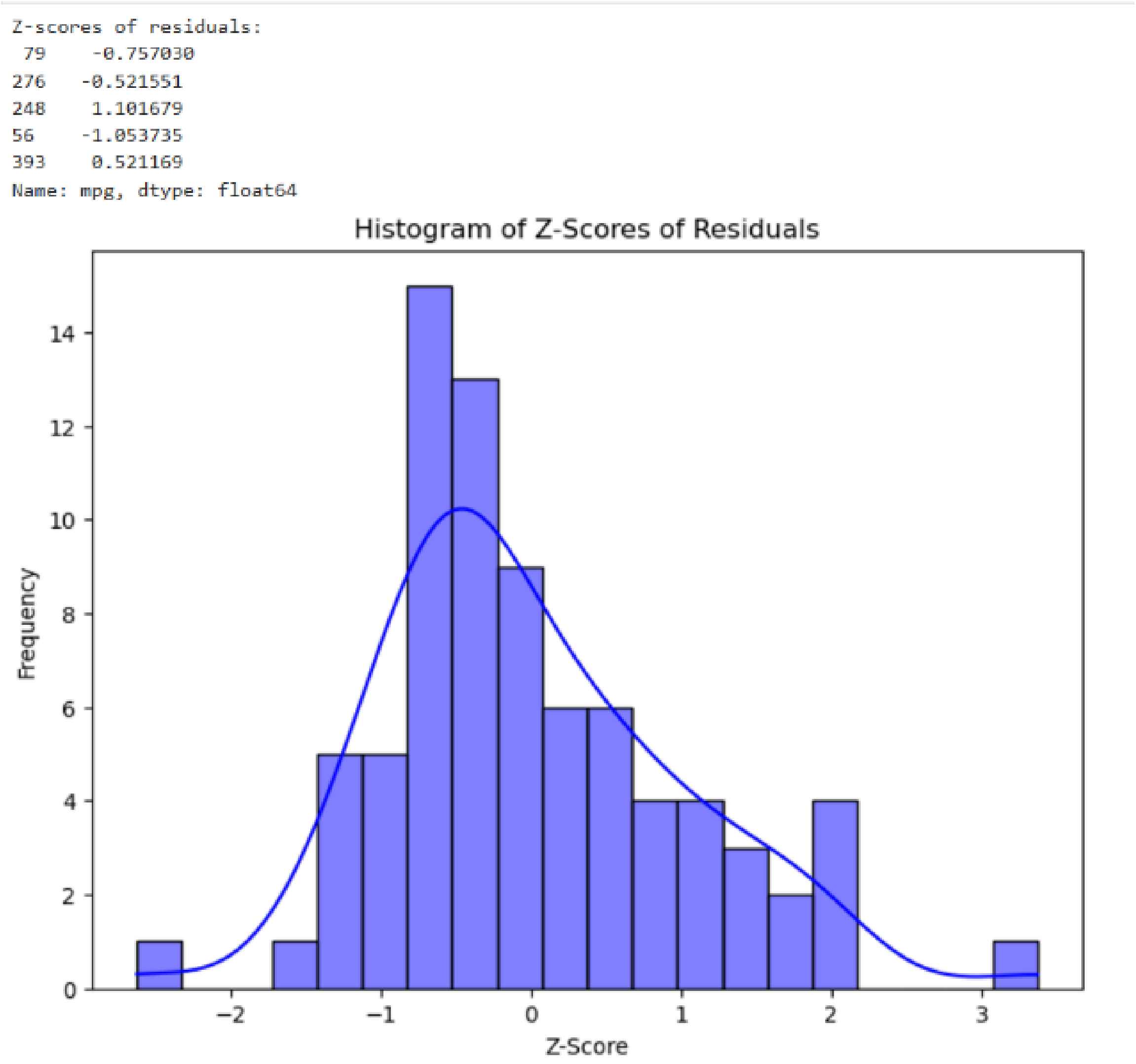


- Plot residuals vs predicted values to check for heteroscedasticity



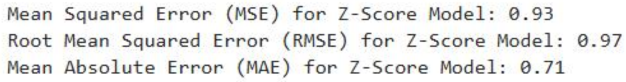
Standardizing the variable using Z-Score

Histogram looks like a bell curve, with most Z-scores near 0 and fewer as they move away from 0, this indicates that the residuals are normally distributed. This is a positive result, suggesting that the linear regression model fits the data well and the assumptions of normality are satisfied.

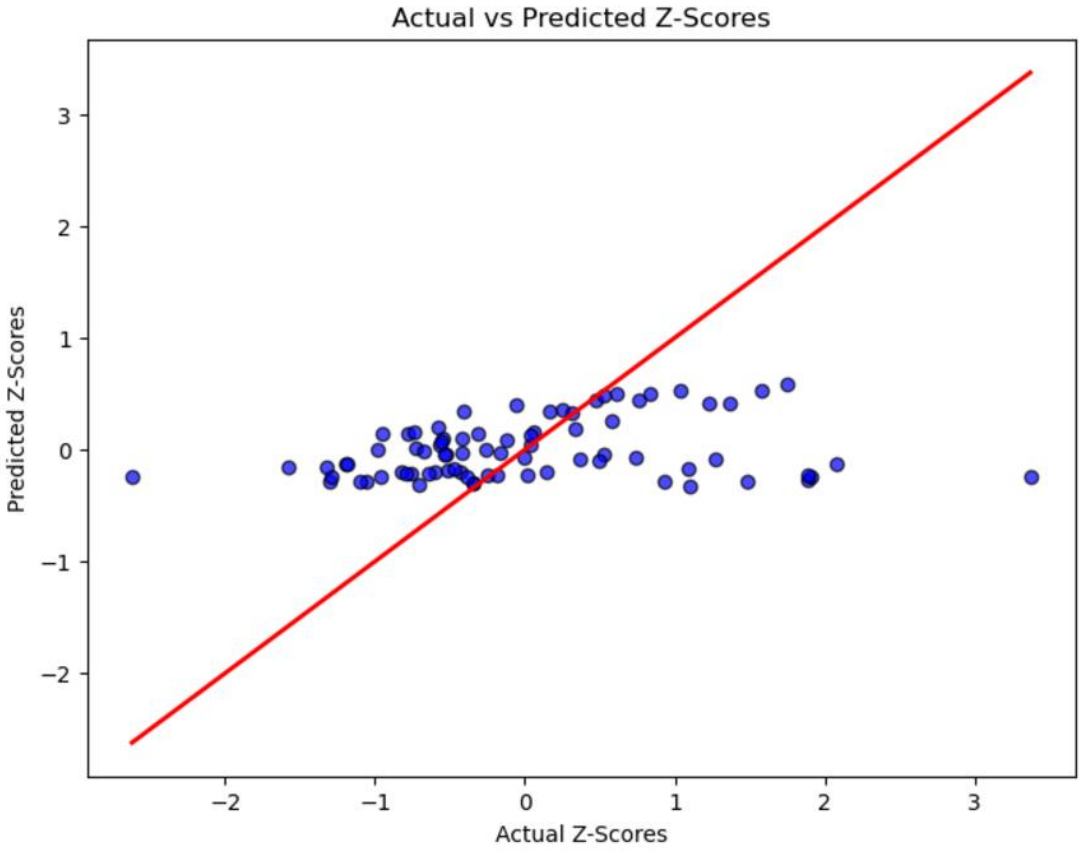


```
Z-scores of residuals:
 79    -0.757030
276   -0.521551
248    1.101679
56    -1.053735
393    0.521169
Name: mpg, dtype: float64
```

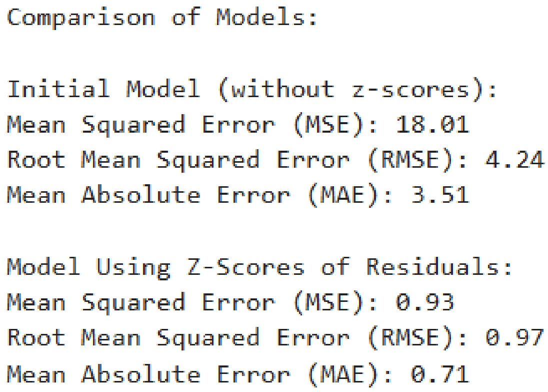
Z-scores as the target variable for the regression model.



```
Mean Squared Error (MSE) for Z-Score Model: 0.93
Root Mean Squared Error (RMSE) for Z-Score Model: 0.97
Mean Absolute Error (MAE) for Z-Score Model: 0.71
```



Comparison between initial model and Z-score model



```
Comparison of Models:

Initial Model (without z-scores):
Mean Squared Error (MSE): 18.01
Root Mean Squared Error (RMSE): 4.24
Mean Absolute Error (MAE): 3.51

Model Using Z-Scores of Residuals:
Mean Squared Error (MSE): 0.93
Root Mean Squared Error (RMSE): 0.97
Mean Absolute Error (MAE): 0.71
```