M2 Project Deliverable 2
Data Preparation & Information Modeling Report

Team
Snehita Moturu - G01388464
Pravallika Avula - G01388664
Sai Roopesh Diddi- G01353614

Prof. Charles Lynch

AIT664 - 001: Information: Representation, Processing & Visualization

George Mason University

08 October 2023

Abstract:

The automobile industry has had a huge impact on many countries since its introduction in the United States in 1895. There have been numerous improvements and diverse car models produced over the years. When consumers consider purchasing a car, they typically consider affordability, luxury, and safety.

Hypothesis:

We want to find patterns and relationships that can be used to anticipate a car's performance and fuel efficiency by analyzing these attributes collectively. The year, place of manufacture, and particular model name of an automobile all have an impact on its performance and fuel consumption, which are determined by variables like displacement, mpg, cylinders, horsepower, weight, and acceleration. We hypothesize that increased fuel economy, more horsepower, and less weight are related to newer model years, particular countries of origin, and particular car models.

Data Source:

This dataset consists of data by Ross Quinlan which covers from the year 1970 to 1982. This dataset was taken from UCI's Machine learning repository.

Attributes:

MPG: The fuel economy of an automobile is the relationship between the distance traveled and the amount of fuel consumed by the vehicle. MPG is continuous from 9 to 48 (Continuous).

Cylinders: A cylinder is the power unit of an engine; it's the chamber where the gasoline is burned and turned into power. Number of cylinder continuous from 3 to 8 (Multi-Valued Discrete)

Displacement: Engine displacement is the measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers. Displacement is continuous from 68 to 455 (Continuous).

Note: This displacement column(cubic inches) has been converted to cubic centimetres using a conversion factor as everyone is familiar with cubic centimeters(cc).

Horsepower: Horsepower is a unit of power used to measure the forcefulness of a vehicle's engine. Horsepower continuous from 46 to 230 (Continuous).

Weight: The weight of an object is related to the amount of force acting on the object, either due to gravity or to a reaction force that holds it in place. The weight of the car continues from 1613 to 5140 (Continuous).

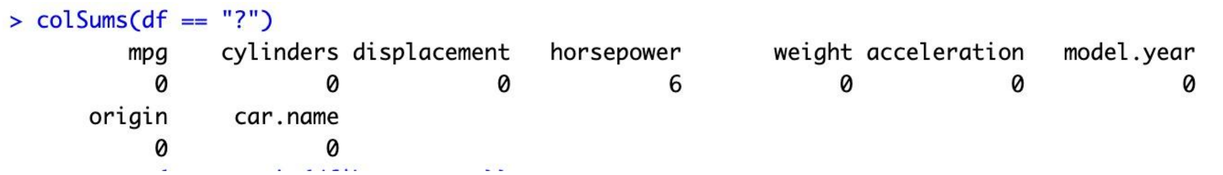
Acceleration: Is the rate of change of velocity of an object with respect to time. It continuous from 8 to 25 (Continuous).

Model Year: Is the year in which a product is manufactured. The year ranges from 1970 to 1982 (Multi-Valued Discrete)

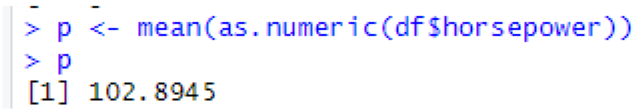
Origin: The Country which manufactures the automobile. (Multi-Valued Discrete)
1: USA
2: EUROPE
3: JAPAN

1. Data cleaning and preprocessing

#Checking for the missing values.

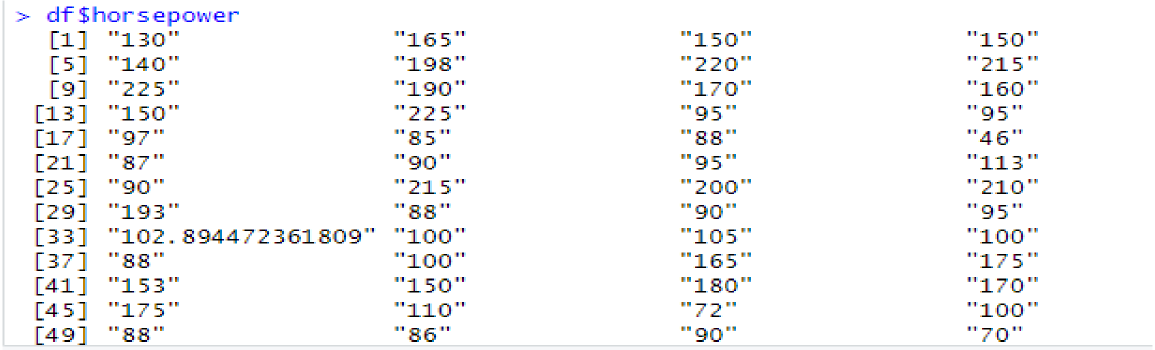


```
> colSums(df == "?")
         mpg    cylinders displacement   horsepower       weight acceleration   model.year
           0            0            0            6            0            0            0
      origin     car.name
           0            0
```



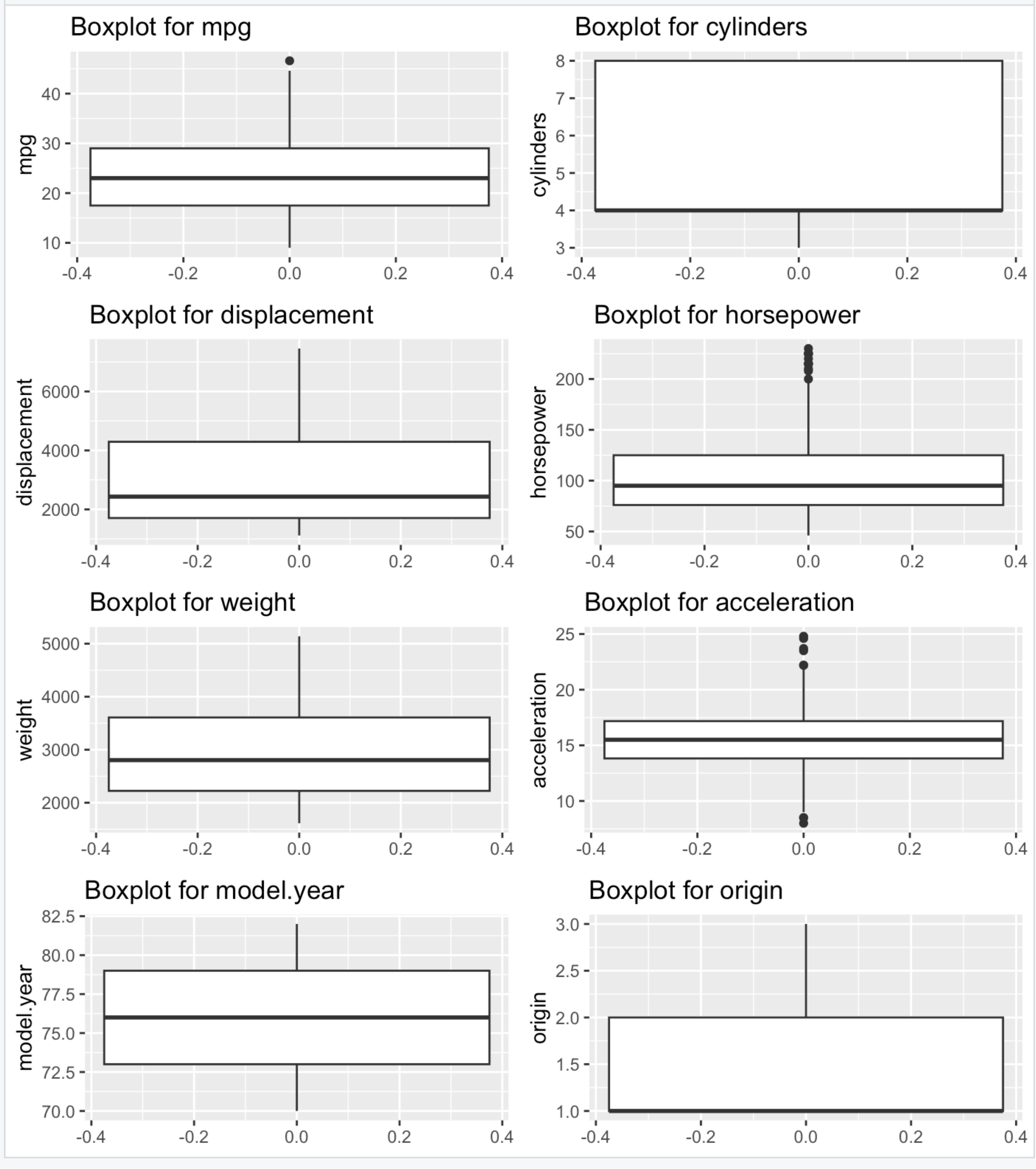
```
> p <- mean(as.numeric(df$horsepower))
> p
[1] 102.8945
```

The horsepower column has 6 missing values (?) and replaced the missing value with 0 to calculate the mean. Later, We replaced the null values with the mean.

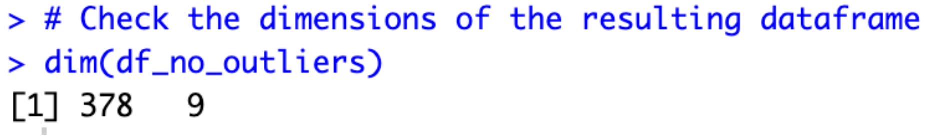


```
> df$horsepower
  [1] "130"               "165"               "150"               "150"
  [5] "140"               "198"               "220"               "215"
  [9] "225"               "190"               "170"               "160"
 [13] "150"               "225"               "95"                "95"
 [17] "97"                "85"                "88"                "46"
 [21] "87"                "90"                "95"                "113"
 [25] "90"                "215"               "200"               "210"
 [29] "193"               "88"                "90"                "95"
 [33] "102.894472361809"  "100"               "105"               "100"
 [37] "88"                "100"               "165"               "175"
 [41] "153"               "150"               "180"               "170"
 [45] "175"               "110"               "72"                "100"
 [49] "88"                "86"                "90"                "70"
```

The outliers are found by plotting boxplots.

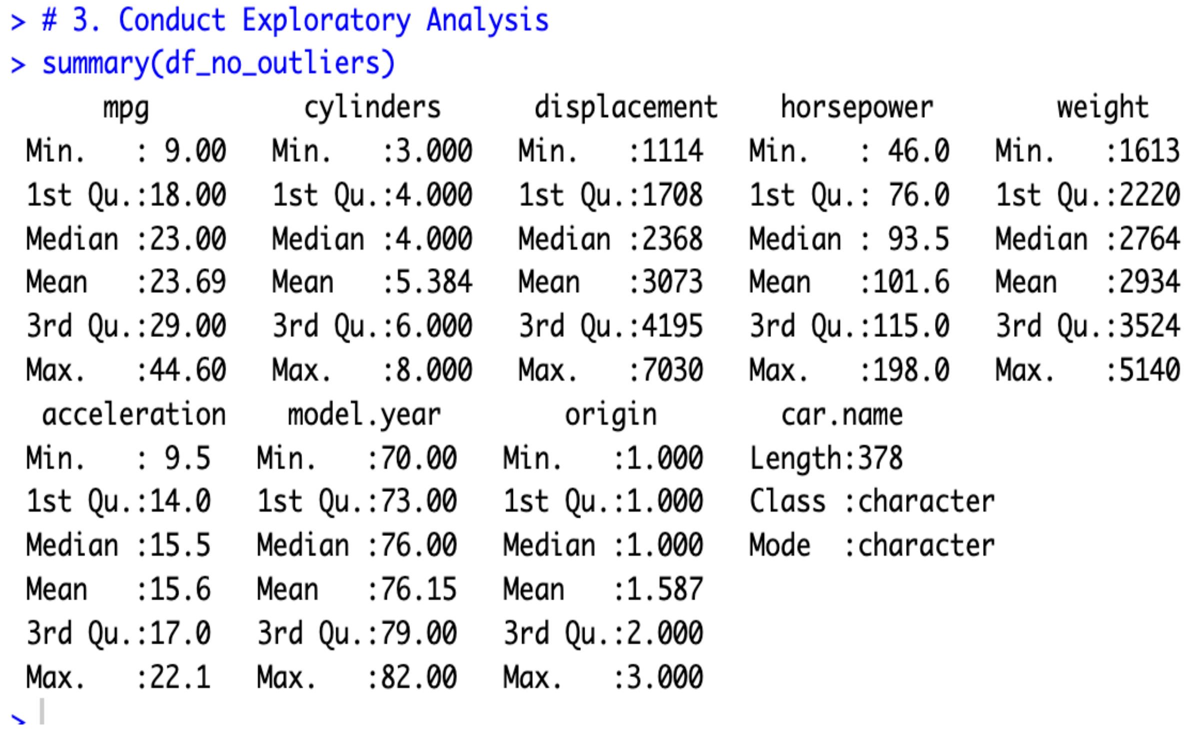


The outliers are removed by changing the first, third quantile, lower and upper limit and we can observe that the rows are reduced that means the outlier rows are removed.



```
> # Check the dimensions of the resulting dataframe
> dim(df_no_outliers)
[1] 378   9
```

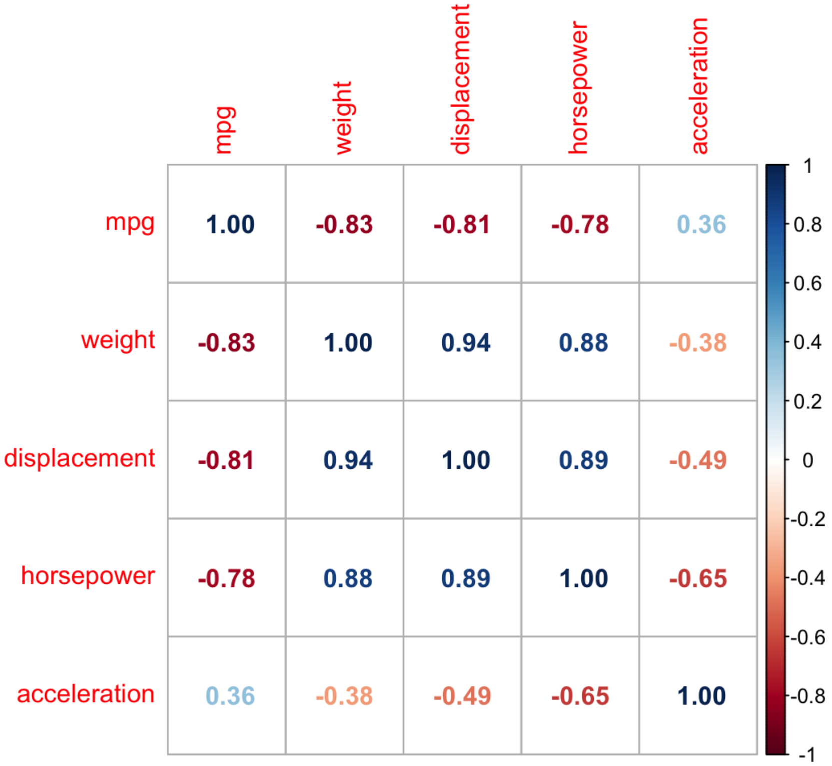
The below summarizes the attributes to calculate Minimum, Maximum values, Mean and Median. This gives insights into the data on the mean and median values how the data is varying.



```
> # 3. Conduct Exploratory Analysis
> summary(df_no_outliers)
      mpg          cylinders      displacement    horsepower        weight    
 Min.   : 9.00   Min.   :3.000   Min.   :1114   Min.   : 46.0   Min.   :1613  
 1st Qu.:18.00   1st Qu.:4.000   1st Qu.:1708   1st Qu.: 76.0   1st Qu.:2220  
 Median :23.00   Median :4.000   Median :2368   Median : 93.5   Median :2764  
 Mean   :23.69   Mean   :5.384   Mean   :3073   Mean   :101.6   Mean   :2934  
 3rd Qu.:29.00   3rd Qu.:6.000   3rd Qu.:4195   3rd Qu.:115.0   3rd Qu.:3524  
 Max.   :44.60   Max.   :8.000   Max.   :7030   Max.   :198.0   Max.   :5140  
  acceleration    model.year        origin        car.name        
 Min.   : 9.5   Min.   :70.00   Min.   :1.000   Length:378        
 1st Qu.:14.0   1st Qu.:73.00   1st Qu.:1.000   Class :character  
 Median :15.5   Median :76.00   Median :1.000   Mode  :character  
 Mean   :15.6   Mean   :76.15   Mean   :1.587                     
 3rd Qu.:17.0   3rd Qu.:79.00   3rd Qu.:2.000                     
 Max.   :22.1   Max.   :82.00   Max.   :3.000                     
>
```

2) Conduct Exploratory Analysis

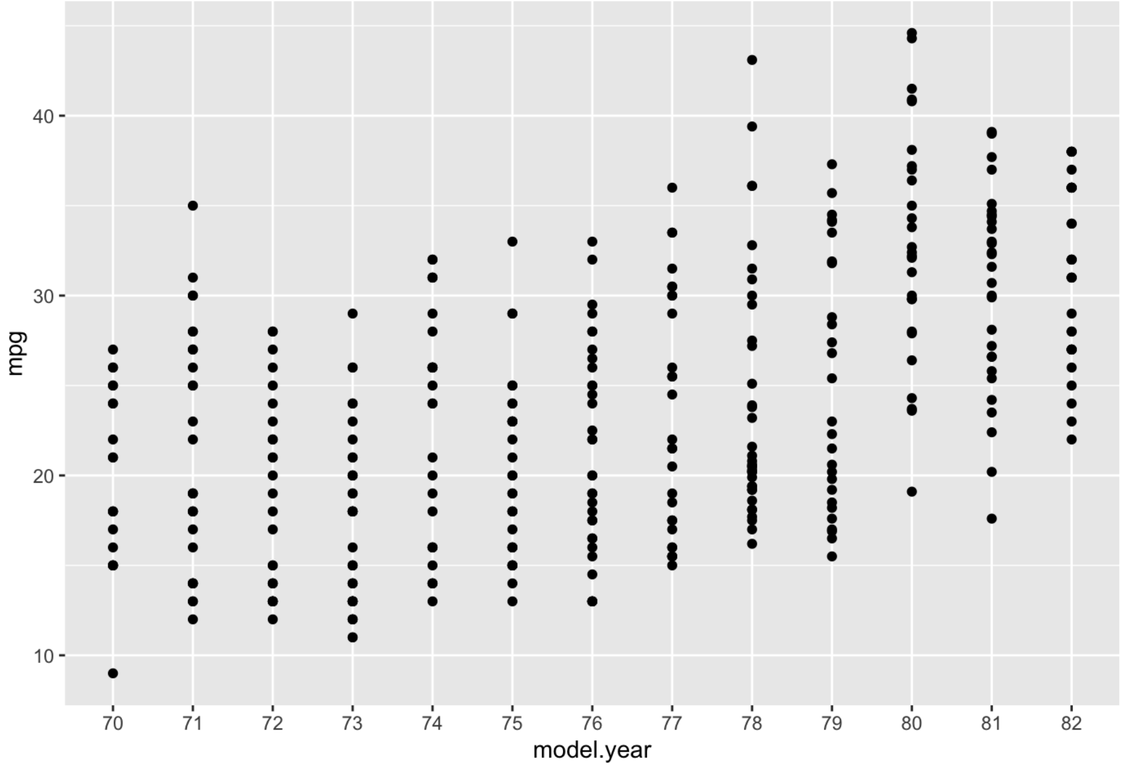
The Correlation plot is used to find the factors that are most correlated with Fuel Efficiency. From the scatter plot we can say that weight, displacement, horsepower are more correlated with mpg (fuel efficiency).



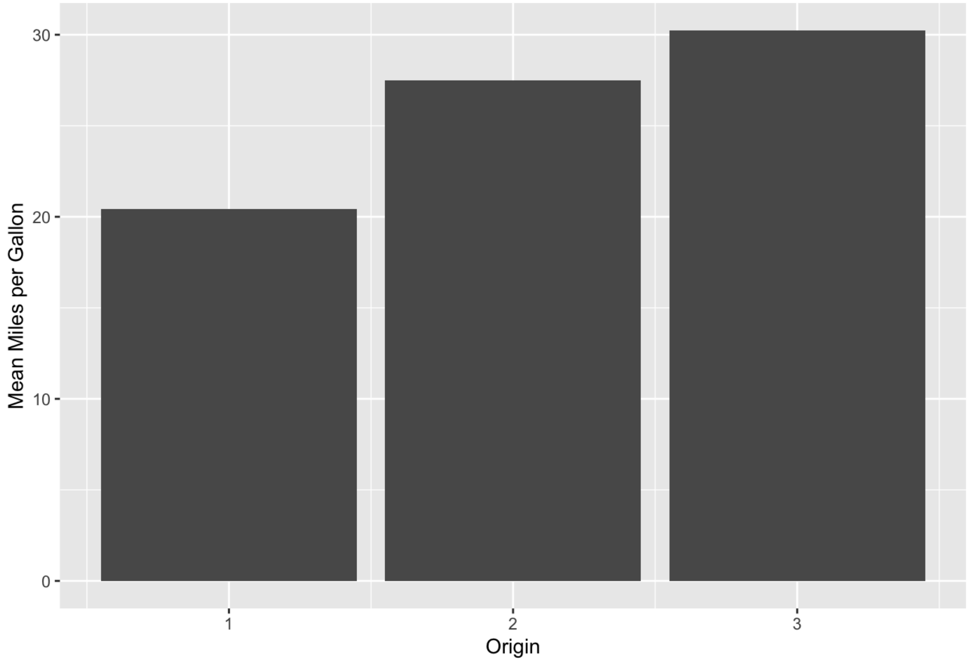
	mpg	weight	displacement	horsepower	acceleration
mpg	1.00	-0.83	-0.81	-0.78	0.36
weight	-0.83	1.00	0.94	0.88	-0.38
displacement	-0.81	0.94	1.00	0.89	-0.49
horsepower	-0.78	0.88	0.89	1.00	-0.65
acceleration	0.36	-0.38	-0.49	-0.65	1.00

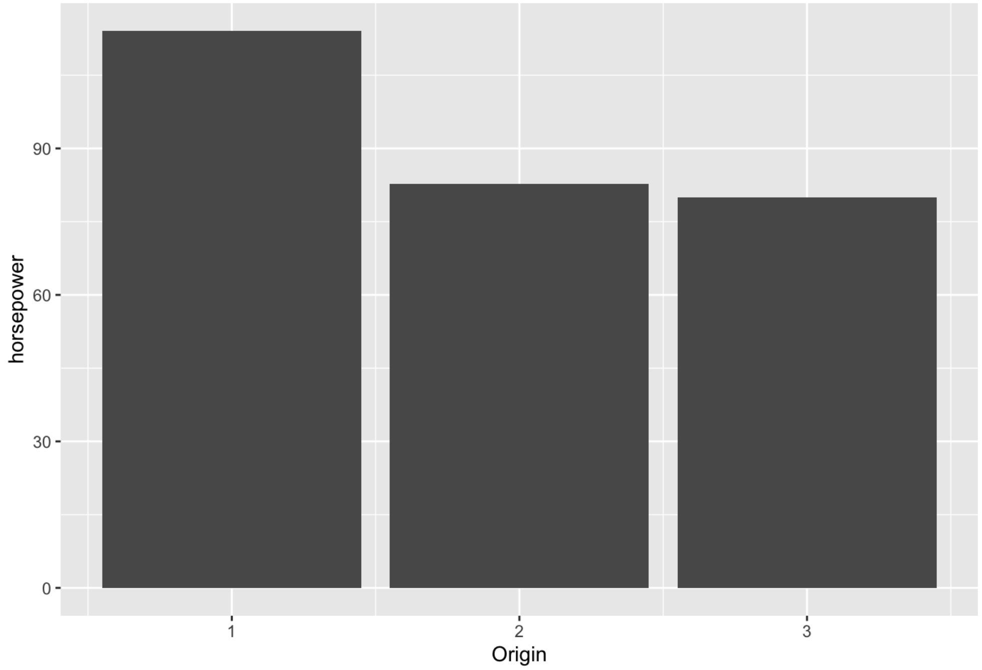
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Now, we determine how mpg is changing over time by plotting a scatterplot between mpg and model years. We can see that the mpg has increased from 1970 to 1982.

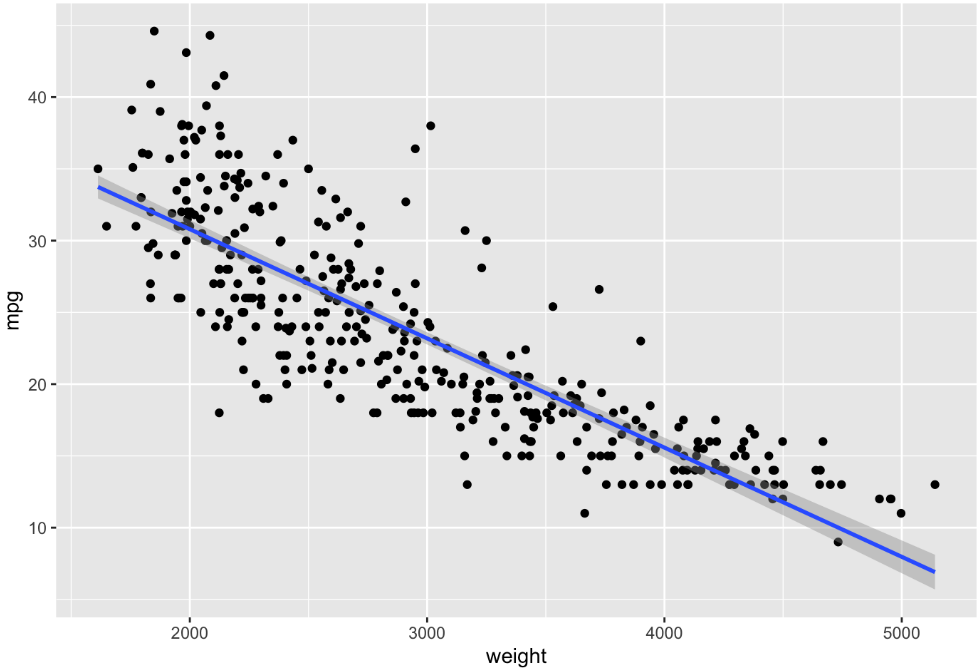


Now we determine which country has the highest mpg and horsepower. To know this we are plotting a bar graph between mpg and origin & horsepower and origin.





It is observed that the mpg is less and horsepower is more for USA originated cars. Whereas, mpg is higher and horsepower is lesser for Japanese originated cars.
Below is the scatterplot between weight and mpg to determine the relationship between the both. We can observe that the mpg is decreases as the weight increases.



Modeling the Fuel Efficiency of an Automobile :

To understand and predict the fuel efficiency (mpg) of an automobile, regression analysis is a valuable statistical tool. It helps answer critical questions such as: Which attributes are most influential in determining mpg? Which attributes can be disregarded? Additionally, it helps us assess our confidence in the relationships discovered.

Linear Regression:

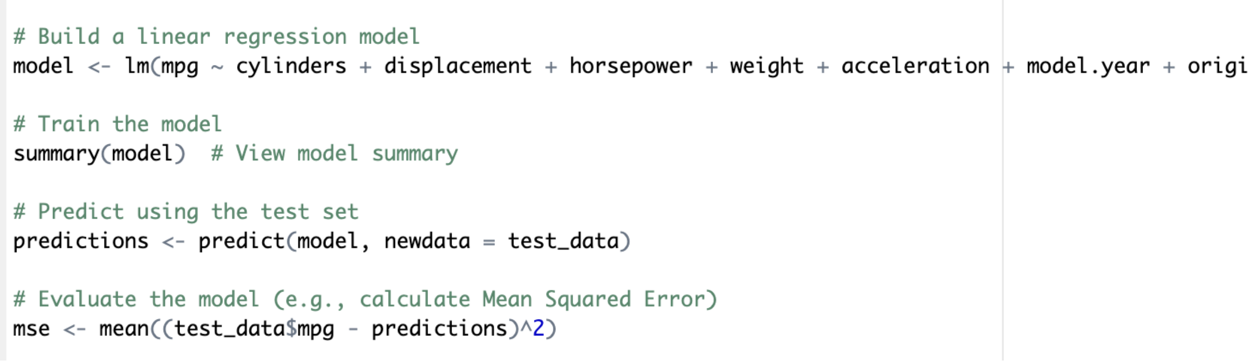
Linear regression is employed to model the relationship between variables, and in this case, it fits a linear equation to the observed data. In our context, the predictors (independent variables) include attributes like cylinders, displacement, horsepower, weight, acceleration, model year, and origin. The target variable (dependent variable) is the automobile's fuel efficiency, measured as mpg.

Splitting the Dataset:

Before building the regression model, we split the dataset into a training set and a test set it's crucial. This separation ensures that we can assess the model's performance on unseen data, helping us avoid overfitting.

Prediction on the Test Set:

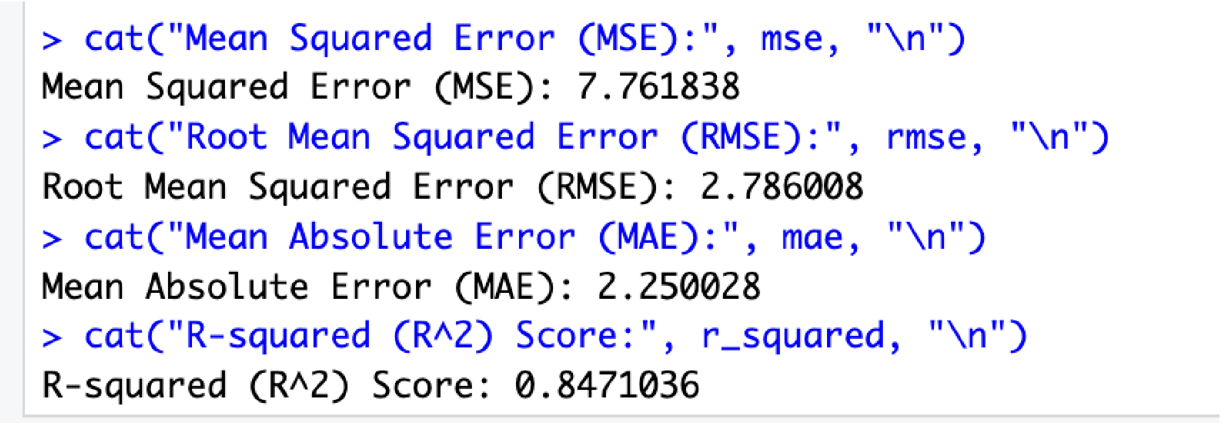
Once the model is trained, we use it to make predictions on the test dataset to estimate the fuel efficiency (mpg) of automobiles in the test set.



```
# Build a linear regression model
model <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + model.year + origi

# Train the model
summary(model)  # View model summary

# Predict using the test set
predictions <- predict(model, newdata = test_data)

# Evaluate the model (e.g., calculate Mean Squared Error)
mse <- mean((test_data$mpg - predictions)^2)
```

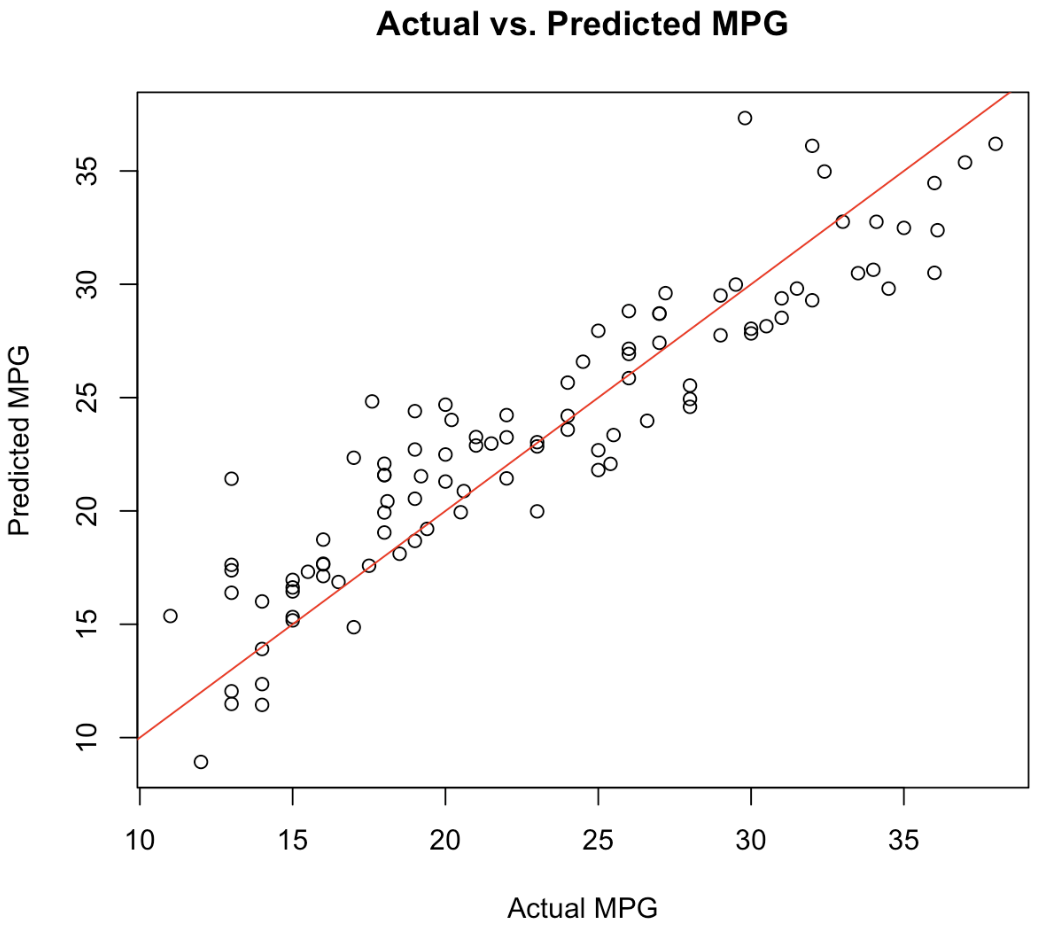
Evaluating the model:

To evaluate the model's performance, we computed various metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) coefficient of determination. These metrics collectively provide a comprehensive assessment of our linear regression model's performance. Our objective is to minimize MSE, RMSE, and MAE, indicating improved predictive accuracy, while striving for a higher R-squared score to signify a better fit of our model to the data. These metrics serve as valuable tools in quantifying the success of our model in predicting automobile fuel efficiency.



```
> cat("Mean Squared Error (MSE):", mse, "\n")
Mean Squared Error (MSE): 7.761838
> cat("Root Mean Squared Error (RMSE):", rmse, "\n")
Root Mean Squared Error (RMSE): 2.786008
> cat("Mean Absolute Error (MAE):", mae, "\n")
Mean Absolute Error (MAE): 2.250028
> cat("R-squared (R^2) Score:", r_squared, "\n")
R-squared (R^2) Score: 0.8471036
```

In general, the R-squared serves as a performance report card for our model. It gauges how effectively our model elucidates variations in automobile fuel efficiency (mpg). This score ranges between 0 and 1.0, with 0 implying no explanation of mpg variation and 1 indicating perfect explanation. With our R-squared score standing at 0.8415889, we find that our model accounts for approximately 84.16% of the mpg differences, primarily driven by attributes like cylinders, displacement, and horsepower. This underscores our model's strong fit with the data, affirming the effectiveness of our chosen attributes in both explaining and predicting fuel efficiency. This R-squared of 0.8415889 signifies a robust relationship between our predictors and fuel efficiency, solidifying the suitability of our linear regression model for mpg prediction based on the specified attributes.



The scatter plot shows the actual and predicted mpg values for a set of cars. The line of best fit shows that there is a positive relationship between the two variables, meaning that as the predicted mpg increases, the actual mpg also tends to increase.

References:

[1] Auto-mpg dataset. (2017, July 2). Kaggle. https://www.kaggle.com/datasets/uciml/autompg-dataset

[2] Quinlan,R.. (1993). Auto MPG. UCI Machine Learning Repository. (n.d.).https://doi.org/10.24432/C5859H.

[3] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

[4] R Core Team (2021). R: The R Project for Statistical Computing. (n.d.-b). https://www.r-project.org/