College of Professional Studies
Northeastern University San Jose

MPS Analytics

Course: ALY6020: Predictive Analytics

Assignment:

Module 2 Project - Building the Car of the Future

Submitted on:

Jan 25, 2024

Submitted to:	**Submitted by:**
Prof: BEHZAD AHMADI	NIKSHITA RANGANATHAN

INTRODUCTION

Understanding the dataset:

The cars dataset comprises 398 rows and 8 variables. The dependent variable is 'MPG' (Miles Per Gallon), which serves as the target attribute for analysis. The dataset under consideration provides a comprehensive collection of attributes related to various car models. Understanding and analyzing this dataset can offer valuable insights into the factors that influence the automotive industry and consumer choices.

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set. Each of the attributes contributes to the overall characterization of a car's performance, efficiency, and origin. The dataset's features are as follows:

1. MPG (Miles Per Gallon): This is a measure of fuel efficiency, indicating how far a car can travel per gallon of fuel.
2. Cylinders: Refers to the number of cylinders in the vehicle's engine, impacting power and fuel consumption.
3. Displacement: The total volume of all the cylinders in the engine, usually measured in cm cubic or liters. It's indicative of engine size and power.
4. Horsepower: A unit of measurement for engine power, denoting the vehicle's performance capabilities.
5. Weight: The total weight of the vehicle, influencing its acceleration, fuel efficiency, and handling.
6. Acceleration: A measure of how quickly the vehicle can increase its speed, typically from 0 to 60 mph.
7. Model Year: The year the vehicle model was released, indicating potential technological and design advancements.
8. US Made: A categorical variable indicating whether the car was manufactured in the US.

The dataset has values ranging from the 1970s to the early 1980s, a period known for significant shifts in automotive technology and fuel economy standards. A majority (over 62%) of the vehicles in the dataset are manufactured in the United States, reflecting their market dominance.

DATA CLEANING

- **Checking for the number of missing values in the dataset**

  The variables do not seem to have any null or NA values.

- **Handling abnormal values in the dataset**

  During the initial analysis, abnormal values in the form of "?" were identified in the Horsepower column. These values were handled by replacing them with the mean of the column.

- **Changing datatypes**

  The Horsepower columns is changed from object datatype to numeric for better analysis.

- **Removing duplicate rows**

  Duplicate rows in the dataset can lead to inconsistencies and affect the accuracy of data analysis. It is essential to find and remove any duplicate rows from the dataset before starting the analysis.
  There are no duplicate rows.

DATA EXPLORATION AND EDA

EDA was conducted to understand the relationships among the variables within the dataset and their impact on the target variable. This exploration was important for gaining valuable insights and identifying key patterns that would help in the development of a Linear Regression model.

Observations from the EDA:

Figure 1:

Here's an overview of the histograms for different variables -

- **MPG**: The distribution is somewhat right skewed, which indicates that while most cars have moderate fuel efficiency, a smaller number are highly fuel-efficient.
- **Cylinders**: The histogram shows discrete categories, which suggests that engines with 4 and 8 cylinders are more common in this dataset. There's a smaller number of cars with 6 cylinders.
- **Displacement**: The displacement histogram is also right skewed, similar to MPG, with more cars having lower engine displacements and fewer cars with high displacement.
- **Horsepower**: This histogram is also right-skewed, with a peak around the lower horsepower values.
- **Weight**: The weight distribution indicates that while there is a variety of car weights, there are comparatively fewer heavy cars.
- **Acceleration**: The acceleration histogram appears to be normally distributed, centering around a peak that suggests most cars have moderate acceleration capabilities.
- **Model Year**: There are several peaks, indicative of the years when more cars were manufactured or sold.
- **US Made**: This histogram is a binary categorical representation, showing a significant number of cars that are US-made and a smaller number that are not.

Figure 2:

The heatmap indicates that cars with more cylinders, higher horsepower and displacement, and greater weight tend to have lower MPG (Strong negative correlation).
Newer car models generally offer better fuel efficiency. Cars made in the US appear to be less fuel-efficient on average.

Figure 3:

From Figure 3, we can see that there is a clear downward trend (as the weight of a car increases, its MPG typically decreases). Most of the cars have 4 and 6-cylinder vehicles, suggesting these are common configurations. The 8-cylinder vehicles are predominantly heavier and have lower MPG, while the few 3-cylinder cars are lighter with higher MPG.

Figure 4:

There is a peak of around 75 HP where 4-cylinder cars dominate. As horsepower increases, particularly beyond 125 HP, the number of cars decreases, with 6 and 8-cylinder engines being more prevalent. 3 and 5-cylinder cars are less common across all horsepower ranges.

Figure 5:
4-cylinder cars are the most common among non-US made vehicles, whereas US made vehicles primarily feature 8-cylinder engines.

Figure 6:
It is observed that miles per gallon has generally improved from 1970 to 1982. The median MPG increased, and the range of MPG values widened, suggesting more variability in car fuel performance.

LINEAR REGRESSION MODEL

Initially, the dataset was divided into two segments: 80% dedicated to training the model, and the remaining 20% for testing purposes. A linear regression model is built using the statsmodels library to predict the "**MPG**" (Miles Per Gallon) based on predictor variables (**'Cylinders', 'Displacement', 'Horsepower', 'Weight', 'Acceleration', 'Model Year', 'US Made'**).

The R² score measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.

The model has an R² score of **0.8463**, indicating that it explains about 84.63% of the variability in the target variable. The most relevant variables for predicting MPG are Model Year, Weight, and US Made, with Model Year being the strongest predictor.

To enhance the performance of our regression model focused on estimating Miles Per Gallon (MPG), two feature selection techniques were used: backward elimination and forward selection.

Backward Elimination: This approach began with all potential predictors included in the model, with the least significant variable being removed in each iteration. This method identified that vehicle weight, the model year, and whether the car is US-made are key predictors of MPG.

Forward Selection: In contrast, this method started with no variables, adding the most statistically significant one at each step. In addition to the features identified by forward selection, backward elimination also highlighted 'Horsepower' as an important factor.

CONCLUSION

The study notes a significant improvement in MPG post-1973. This trend may be attributed to advancements in technology, stricter environmental regulations, and consumer demand for more fuel-efficient vehicles.
Key elements like vehicle weight, and model year play critical roles in determining a vehicle's fuel efficiency. US-made vehicles tend to have lower MPG compared to non-US made vehicles, possibly due to different manufacturing practices or market preferences.
While acceleration was initially considered, it appears to have a minimal impact on MPG compared to other factors.

By carefully considering these features, the team can develop vehicles that are not only more energy-efficient but also appealing to a global market.

REFERENCES

GeeksforGeeks. (2023a, September 6). *Sequential feature selection*. https://www.geeksforgeeks.org/sequential-feature-selection/

Zach. (2022, August 26). *How to perform OLS regression in Python (With example)*. Statology. https://www.statology.org/ols-regression-python/

Seaborn.lmplot() method. (n.d.). https://www.tutorialspoint.com/seaborn/seaborn_implot_method.html

APPENDIX

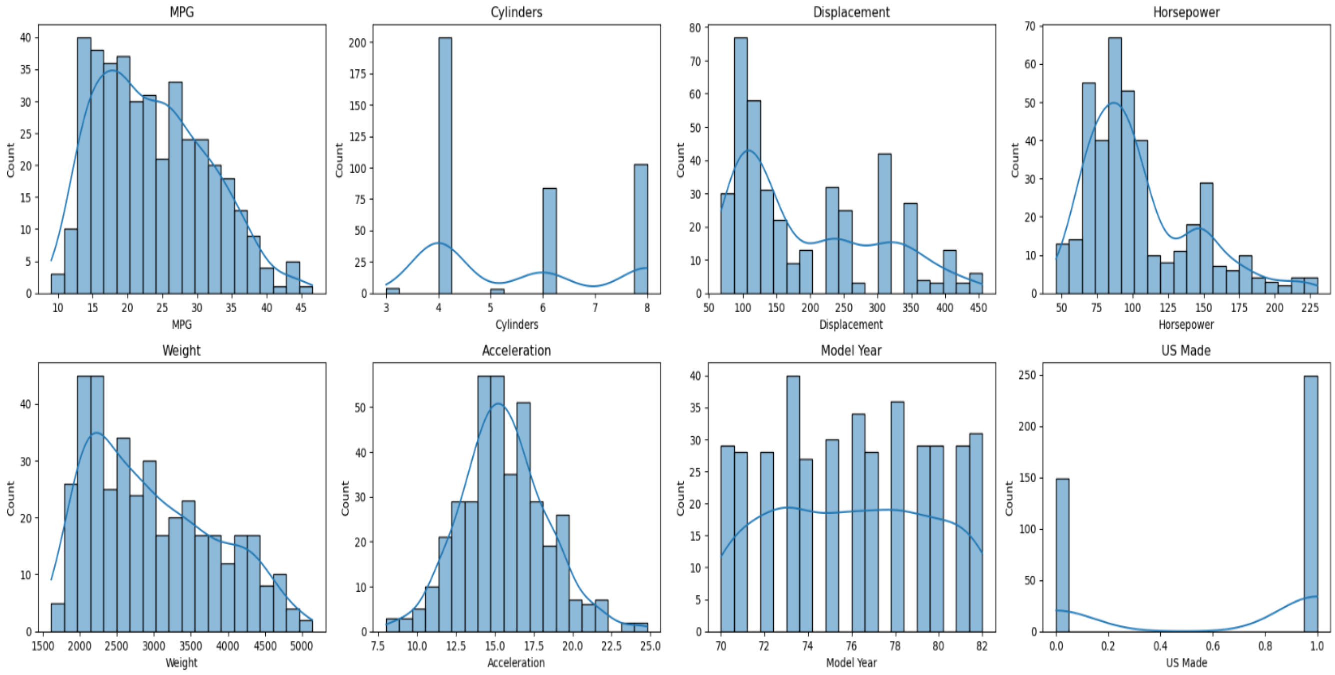
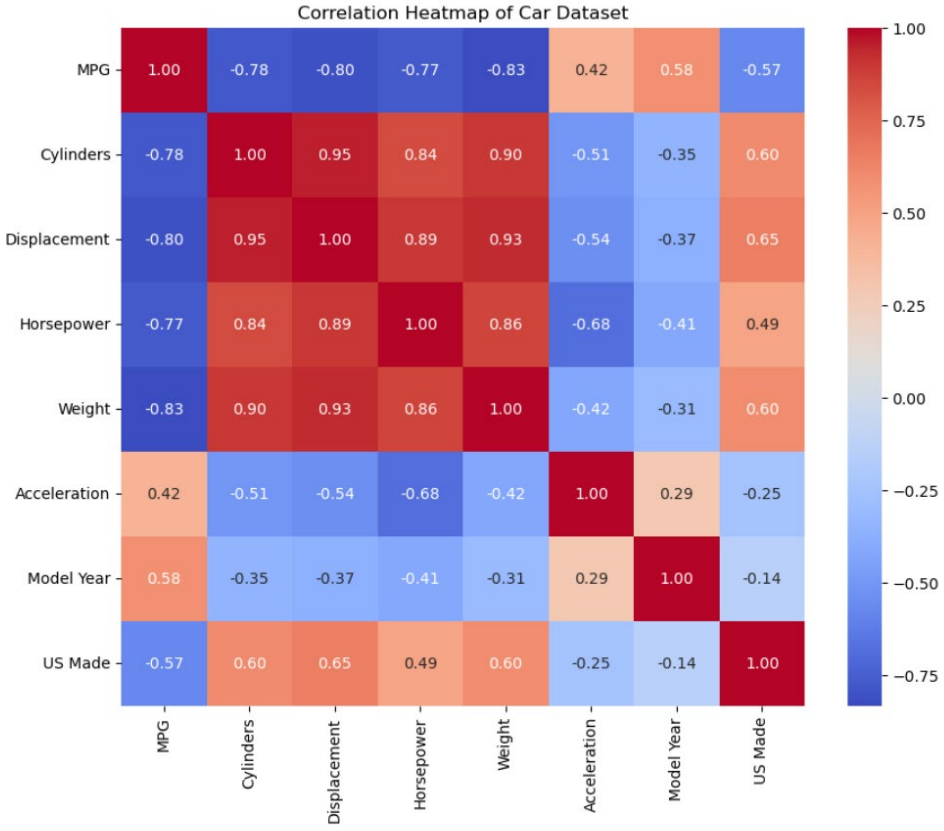


Figure 1 – Histograms for understanding distribution



	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	US Made
MPG	1.00	-0.78	-0.80	-0.77	-0.83	0.42	0.58	-0.57
Cylinders	-0.78	1.00	0.95	0.84	0.90	-0.51	-0.35	0.60
Displacement	-0.80	0.95	1.00	0.89	0.93	-0.54	-0.37	0.65
Horsepower	-0.77	0.84	0.89	1.00	0.86	-0.68	-0.41	0.49
Weight	-0.83	0.90	0.93	0.86	1.00	-0.42	-0.31	0.60
Acceleration	0.42	-0.51	-0.54	-0.68	-0.42	1.00	0.29	-0.25
Model Year	0.58	-0.35	-0.37	-0.41	-0.31	0.29	1.00	-0.14
US Made	-0.57	0.60	0.65	0.49	0.60	-0.25	-0.14	1.00

Figure 2 – Correlation matrix

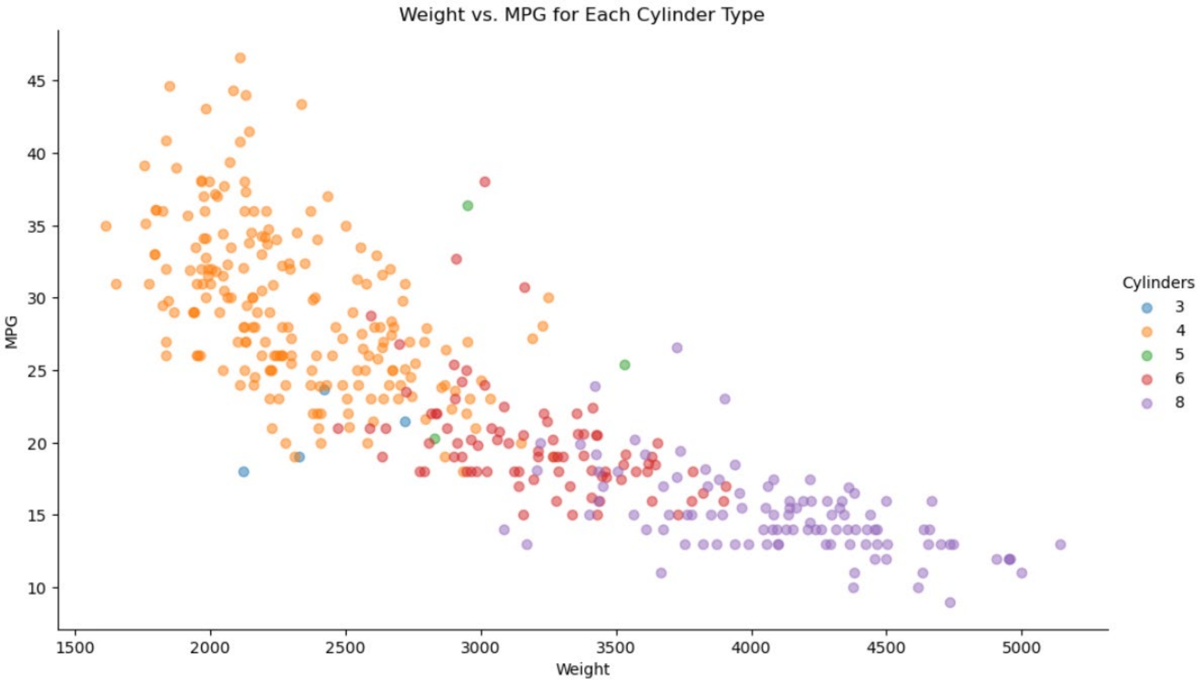


Figure 3 – MPG vs Weight

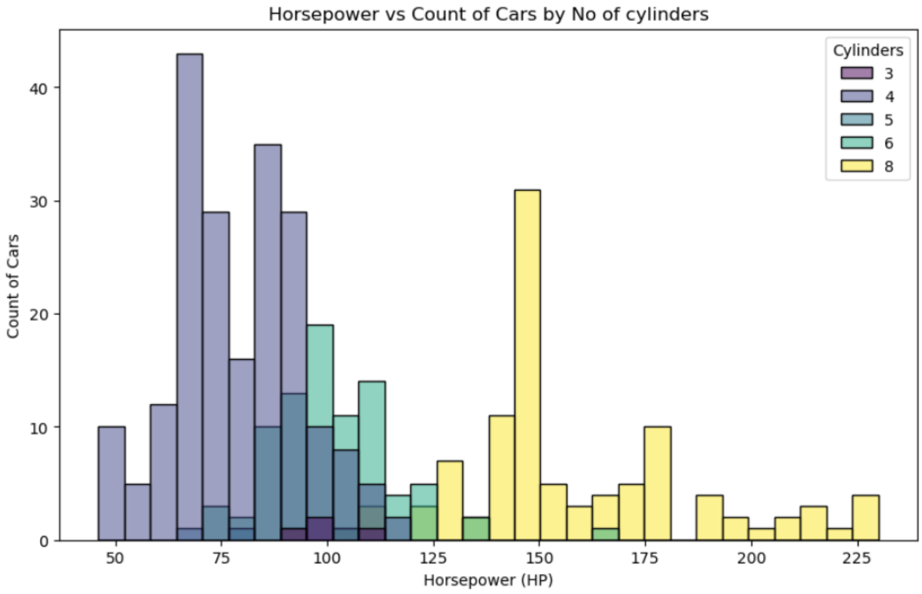


Figure 4 – Stacked bar graph (HP vs Count of cars by number of cylinders)

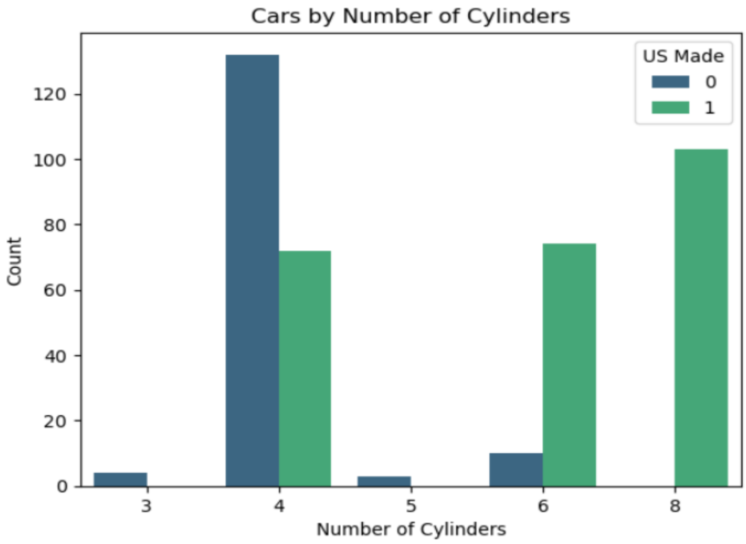


Figure 5 – Grouped bar chart (Count of different no of cylinders)

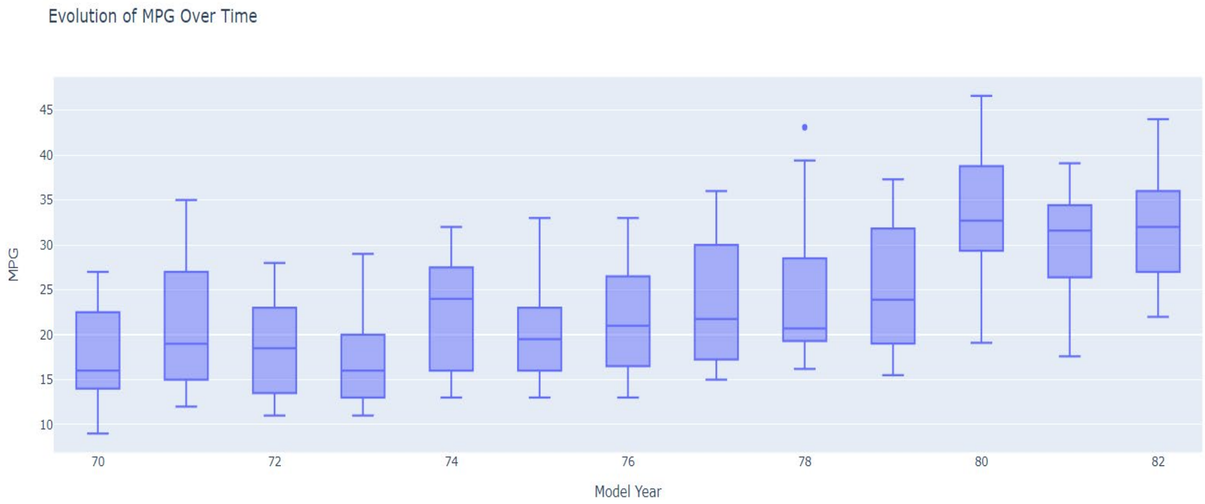
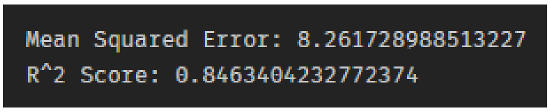
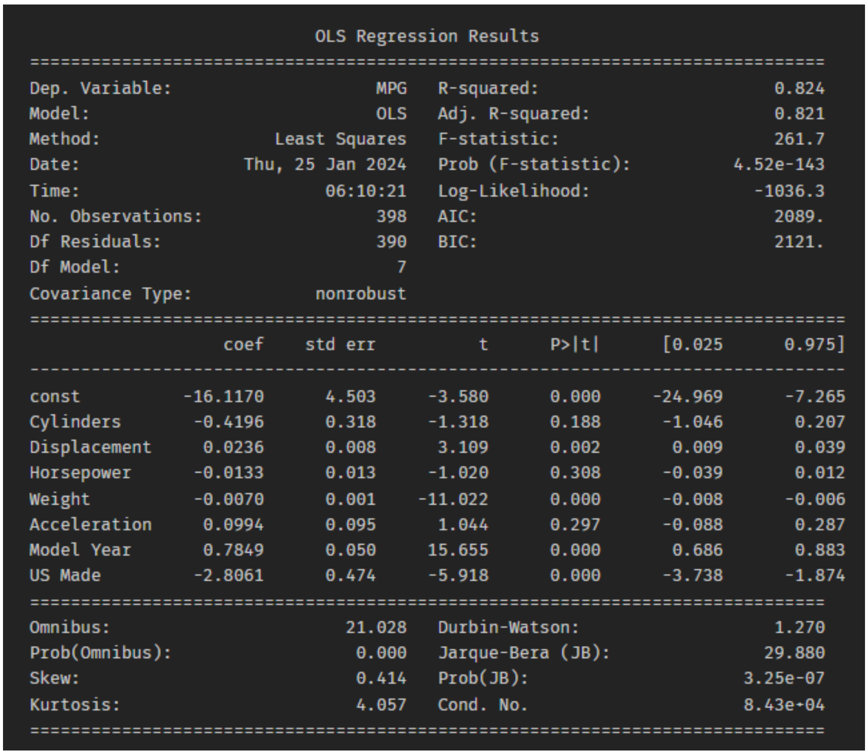


Figure 6 – MPG over time



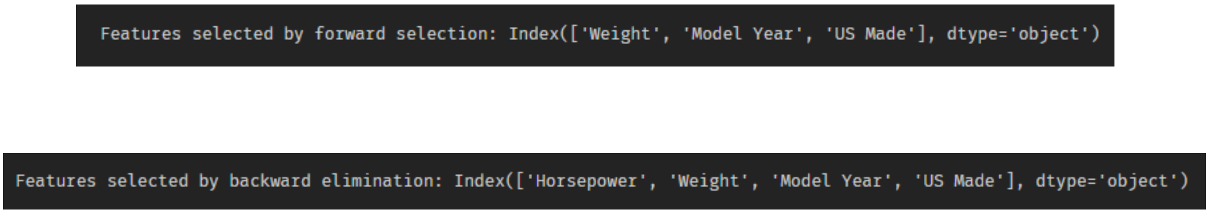
```
Mean Squared Error: 8.261728988513227
R^2 Score: 0.8463404232772374
```

Figure 7 – Linear regression model A



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    MPG   R-squared:                       0.824
Model:                            OLS   Adj. R-squared:                  0.821
Method:                 Least Squares   F-statistic:                     261.7
Date:                Thu, 25 Jan 2024   Prob (F-statistic):          4.52e-143
Time:                        06:10:21   Log-Likelihood:                -1036.3
No. Observations:                 398   AIC:                             2089.
Df Residuals:                     390   BIC:                             2121.
Df Model:                           7
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -16.1170      4.503     -3.580      0.000     -24.969      -7.265
Cylinders       -0.4196      0.318     -1.318      0.188      -1.046       0.207
Displacement     0.0236      0.008      3.109      0.002       0.009       0.039
Horsepower      -0.0133      0.013     -1.020      0.308      -0.039       0.012
Weight          -0.0070      0.001    -11.022      0.000      -0.008      -0.006
Acceleration     0.0994      0.095      1.044      0.297      -0.088       0.287
Model Year       0.7849      0.050     15.655      0.000       0.686       0.883
US Made         -2.8061      0.474     -5.918      0.000      -3.738      -1.874
==============================================================================
Omnibus:                       21.028   Durbin-Watson:                   1.270
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.880
Skew:                           0.414   Prob(JB):                     3.25e-07
Kurtosis:                       4.057   Cond. No.                     8.43e+04
==============================================================================
```

Figure 8 – Linear regression model B



```
Features selected by forward selection: Index(['Weight', 'Model Year', 'US Made'], dtype='object')
```

```
Features selected by backward elimination: Index(['Horsepower', 'Weight', 'Model Year', 'US Made'], dtype='object')
```

Figure 9 – Feature Selection