Northeastern University

ALY6020: Predictive Analytics

Assignment 2: Building the car of the future

Abhigna Ramamurthy, 002982276

College of Professional Studies, Northeastern University

Professor: Justin Grosz

Jan 22, 2023

Introduction

This report includes comprehensive analysis on factors affecting mileage per gallon of car. The data is provided by a struggling car manufacturer who intends to understand what factors are contributing to better mileage among their existing cars and thus build a fuel-efficient car to compete at the highest level in the market. This is a noble cause which can help reduce emission, contributing to better environment consciousness. Various factors like number of cylinders, displacement (measurement of the cylinder volume swept by all the pistons), horsepower (power an engine produces), weight of the car, acceleration, model year, and finally an indicator to determine if the car was US made or not. Using these factors, the business purpose of this report is to determine the mileage per gallon (MPG) of the cars. The number of miles your car can travel on a single gallon of fuel is referred to as fuel efficiency. While "miles per gallon" is a numerical measure, "fuel economy" is a word that describes how well an automobile can use fuel. Thus, measuring the impact on mileage directly relates to fuel economy.

Data cleaning

Once the excel is imported in Jupyter Notebook, the dataset includes columns for MPG, cylinders, displacement, horsepower, weight, acceleration, model year and if the car was US made or not. It also contains 398 records and 8 features with 4 integer values, 3 floating data type values, and one object. The object data type value is ‘Horsepower’ which is intended to be numerical, hence further investigation on that revealed that there was ‘?’ value in 6 records. This anomaly made the numerical column to be automatically converted to object. This accounts to less than 2% of the total data, however, as we do not have lot of data to work with, I decided to keep the records. As the histogram of the attribute ‘Horsepower’ showed a right skewed long tail structure, I decided to replace the ‘?’ values with median value of 93.5 and converted the column into integer. Checking for missing values came clean as we did not have any missing values.

Another major column added to the dataset is “Age” of the car which was derived from “Model Year”. For example, if the car model is 70 (which refers to 1970) then age would be 53. While a car's age and the number of miles on its odometer are significant factors to consider, it's more crucial to look at how well the owner maintained the vehicle. Hence would be better to use age rather than just model year for analysis as it directly contributes to mileage over the years.

Finally, the check for outliers was performed using box plots as shown in Figure 1 Appendix 1. There are a few outliers in MPG with more than 40 miles per gallon, outliers in terms of horsepower with more than 200hp, and acceleration value being less than value of 10 or greater than 17. Upon further investigation, these are valid points and would add to the diversity of points explored in the analysis. Hence there was no removal of values at this stage.

Exploratory Data Analysis

In this section, exploratory approach to analyzing, summarizing, and identifying patterns, relationships, and anomalies are discussed for the car dataset. With approximately 63% of cars made in the US, we see some interesting patterns in the visual graphs. From Figure 2, Appendix 1, we can see that the histogram of cylinders versus the sum of MPG for each cylinder type has 4 cylinders as the most common and majority are US made cars. Area plot of displacement versus mileage in Figure 3, Appendix 1, shows that there is slight negative relationship between displacement and mileage. However, with all the noise in the histogram it is best analyzed in a machine learning model setting which can be seen in the next section.

An important aspect of a vehicle's fuel usage is horsepower. Histogram of horsepower versus mileage categorized by US made or not, in Figure 4, Appendix 1, can see that there 50-125 horsepower has the highest number of cars. Scatter plot with regression line for weight versus mileage in Figure 5, Appendix 1 shows that as the weight of the car increases, the mileage deceases. The attribute age has a similar impact on the mileage as seen in Figure 6 of Appendix 1.

The correlation matrix for the dataset shown in Figure 1 below shows positive relationship between acceleration and MPG indicating higher acceleration would lead to higher mileage. Weight is the attribute which shows high negative correlation with MPG which is in line with what was discussed before. Displacement also shows highly inverse proportionality towards MPG. Other attributes like US Made, Age, Horsepower, Cylinders show increasingly high negative correlation with MPG respectively.

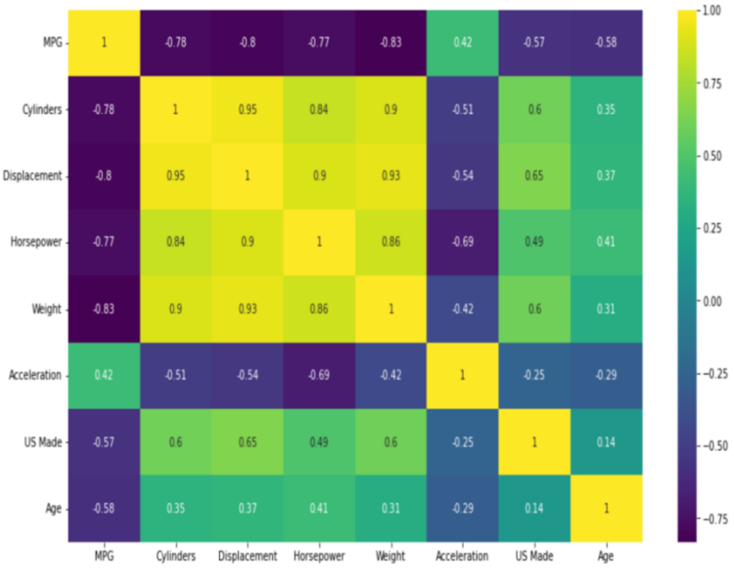


Figure 1: Correlation Matrix.

Analysis

Data Pre-processing

Upon data cleaning and exploratory data analysis, I did several steps to get the data ready for modelling which checking for all the assumptions for linear regression. First assumption check is to check for linearity. Scatter plot is used to check for linearity assumption where the relationship between the independent and dependent variables should be linear. We can see that not all the attributes are linearly distributed in Figure 6, Appendix 2. The second test is independence test where pair plots (Figure 7, Appendix 2) were used to determine that all the attributes are independent of each other. Homoscedasticity denotes that the variance of the residuals is equal or nearly equal along the regression line (Figure 8, Appendix 2). We may verify that there shouldn't be any pattern in the error terms by charting them against the expected terms. The next test was to investigate multicollinearity using variance inflation factor. At this stage I found that the values were highly correlated and removed the attributes “Cylinders” and “Weight” which had high VIG values to reduce the collinearity. The data was split into train and testing data with 80-20 ratio respectively. The values were also scaled using standard scalar method as the attributes had different ranges of values.

Data Modeling

A dependent variable (also known as the outcome variable or the response variable) and one or more independent variables are modeled using the statistical technique of linear regression (also known as predictor variables or explanatory variables). Finding the best-fitting straight line (or hyperplane in the case of several independent variables) that illustrates the relationship between the variables is the aim of linear regression. The equation $Y = a + bX$, where Y is the dependent variable, X is the independent variable, a is the y-intercept, and b is the slope of the line, is used to represent the line. Given the pretext of the report, the data will be run through a series of linear regression where the results will be evaluated to understand the highest impactful attributes towards MPG.

The base linear regression model was implemented using statsmodels.api for generating OLS regression results as shown in Figure 9, Appendix 3. The interpretation of the results is as follows: Let’s start Adjusted. R-squared reflects the fit of the model, higher value generally indicates better fit. The model has an adjusted R-squared value as 0.758 which is good and along with AIC value of 1770, can be used to compare the models after optimization. The p-value indicates the significance of the attribute towards the dependent variable MPG. A p-value of less than 0.05 is statistically significant, and in our results all the 4 attributes qualify to be significant as p-value is less than 0.05. The next values to consider are the coefficients. The constant coefficient value is 23.69 which refers to Y-intercept, that is, if every other attribute is 0, the MPG would be 23.69. Horsepower (co-efficient -2.8) and Age (co-efficient -2.6) have the highest co-efficient values with a negative value showcasing that higher the horsepower or age, lower the MPG. So, to have higher MPG and better fuel economy, we need optimal horsepower and newer cars. In general, more power results in more fuel being consumed. This means that by selecting a car with no more horsepower than you require, you may save money on fuel and lessen your influence on the environment. How old and how many miles the car has also run an important indicator for the condition of the car. A 10-year-old car with 100,000 miles may have received more value than a five-year-old model with 50,000 miles. Hence age can be considered as one of the major aspects to look at for the car manufacturer.

Optimization

Optimization is an essential part of modelling to enhance the performance of the model. In this section optimization of feature selection is done to evaluate if it improves the performance of the linear models, thereby, getting better attributes contributing towards MPG. Forward and Backward selection methods are done using SFS. The feature selection algorithm SFS (Sequential Forward Selection) is used in linear regression to choose the most useful characteristics that enhance the model's performance. The algorithm begins with a blank set of features and, until a stopping requirement is satisfied, adds the feature that leads to the greatest improvement in model performance at each step. This method is regarded as greedy since it always selects the feature that seems to be the best at that current stage, regardless of how it might affect the model's overall performance over time. This can also be used for backward by just having one of the parameters forward as false.

The results for forward selection method are shown in Figure 10, Appendix 3. Number of features selected for optimal output is 5. 'Cylinders', 'Displacement', 'Weight', 'US Made’, and 'Age' are the attributes considered. The model yielded Adjusted R-squared value of 0.815 which is higher than that of base model. AIC value decreased to 1686 which is an indicator of better model. The constant coefficient is 23.69, same as base model. The two most significant attributes are “Weight” (-5.99) and “Age” (-3.06) which implies lowering the weight and newer car (as also indicated in base model) results in higher MPG values.

The next step was to perform backward selection model. The results of backward selection are shown in Figure 11, Appendix 3. Four features resulted in best results for backward selection. The features used were 'Displacement', 'Weight', 'US Made', and 'Age'.

The model yielded Adjusted R-squared value of 0.815, same as forward selection optimization, but is higher than that of base model. AIC value decreased to 1685 which is an indicator of better model but is very close to 1686 value of forward selection method. The constant coefficient is 23.69, same as base model and forward selection model. The two most significant attributes are “Weight” (-6.02) and “Age” (-3.06) which implies lowering the weight and newer car (as also indicated in base model) results in higher MPG values. These significant values are like the ones found in forward selection.

Table 1: Benchmarking metrics used in the three methods

Method	***Adjusted R-Squared***	***AIC***	***Top 2 significant attributes***
Base model	0.758	1770	Horsepower, Age
Forward Selection	0.815	1686	Weight, Age
Backward Selection	0.815	1685	Weight, Age

Hence, we can conclude that optimization in any form, be it forward selection, or backward selection, has improved the model compared to the base model. The comparison of benchmarking metrics is shown above in Table 1. Adjusted R-Squared value has improved by more than 0.05 indicating better fit of the model given the data. AIC value can also be useful when comparing different models with different numbers of features. AIC penalizes models with more features, so if two models with different number of features have similar AIC values, that means that both models fit the data reasonably well, but the simpler one (model with fewer features) is probably better. Hence with lesser number of features on backward selection model, and higher Adjusted R-squared value and lower AIC values, is the best model for the given dataset.

Conclusion

Based on the above analysis, we can say that several factors affect the mileage per gallon of a car. Though good mileage refers to good fuel efficiency of the car, it might not always refer good fuel economy. Good maintenance of car over the years a big factor affecting fuel economy and thus age of the car is a big factor to consider. As shown in the optimized results above, age is one of the prominent factors affecting the MPG. Mileage has shown to decrease with older cars as compared to newer cars. Hence age is big factor while purchasing cars especially new cars. The car manufacturer needs to make sure the older cars are well maintained and offer maintenance packages to new customers so they can preserve the value of the car over the future years. Monitoring MPG values over time can go a long way in preserving the cars. Any drastic change in the numbers can be a sign of engine trouble.

Fuel economy is also affected by the size and weight of the car which in turn affects the aerodynamics of the car. Factors such as driving style, the type of tires, and the quality of the fuel can also affect fuel economy. Hence focusing on having a lighter model for the car with importance to main features along with a balance of luxury features is a good idea to invest in. More research on how to make cars lighter but at the same time be safe is a major topic across the industry. Keeping a look out for such research and development or having an in-house R&D team can help better the fuel economy and innovate the car of the future.

References

Backward Feature Elimination and its Implementation. (2021). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/backward-feature-elimination-and-its-implementation/

Forward Feature Selection | Implementation of Forward Feature Selection. (2021). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/04/forward-feature-selection-and-its-implementation/

Horsepower. (2023). Language selection - Natural Resources Canada / Sélection de la langue - Ressources naturelles Canada. https://www.nrcan.gc.ca/energy-efficiency/transportation-alternative-fuels/personal-vehicles/choosing-right-vehicle/tips-buying-fuel-efficient-vehicle/factors-affect-fuel-efficiency/horsepower/21028

How does fuel economy vary with engine displacement? (2018). Quora. https://www.quora.com/How-does-fuel-economy-vary-with-engine-displacement

Linear Regression in Python using Statsmodels – Data to Fish. (2022). Data to Fish – Data Science Tutorials. https://datatofish.com/statsmodels-linear-regression/

shrutimechlearn. (2020, June 15). Step by Step Assumptions - Linear Regression. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/code/shrutimechlearn/step-by-step-assumptions-linear-regression

Things You Should Know About Fuel Efficient Cars. (2023). Acura. https://www.acurasugarland.com/what-to-know-about-fuel-efficient-vehicles/#:~:text=Fuel%20economy%20refers%20to%20the,car%20can%20efficiently%20utilize%20fuel.

Appendix 1 Exploratory Data Analysis

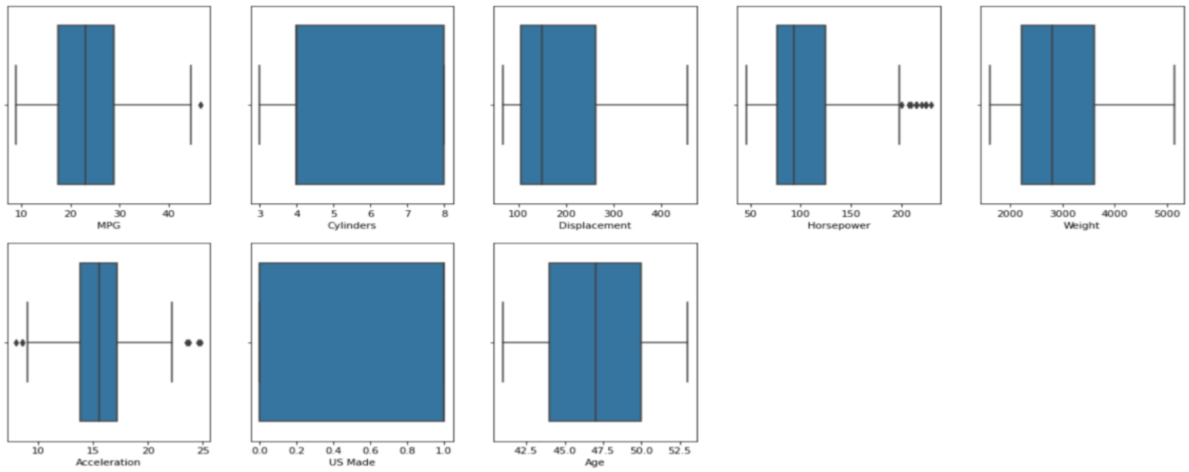


Figure 1: Box plot for outliers' check.

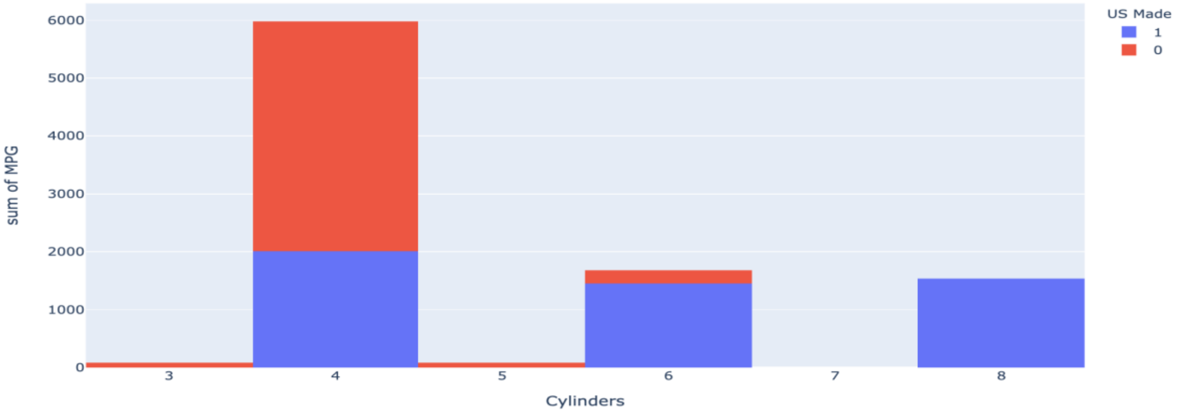


Figure 2: Histogram of cylinders categorized by US made or not.

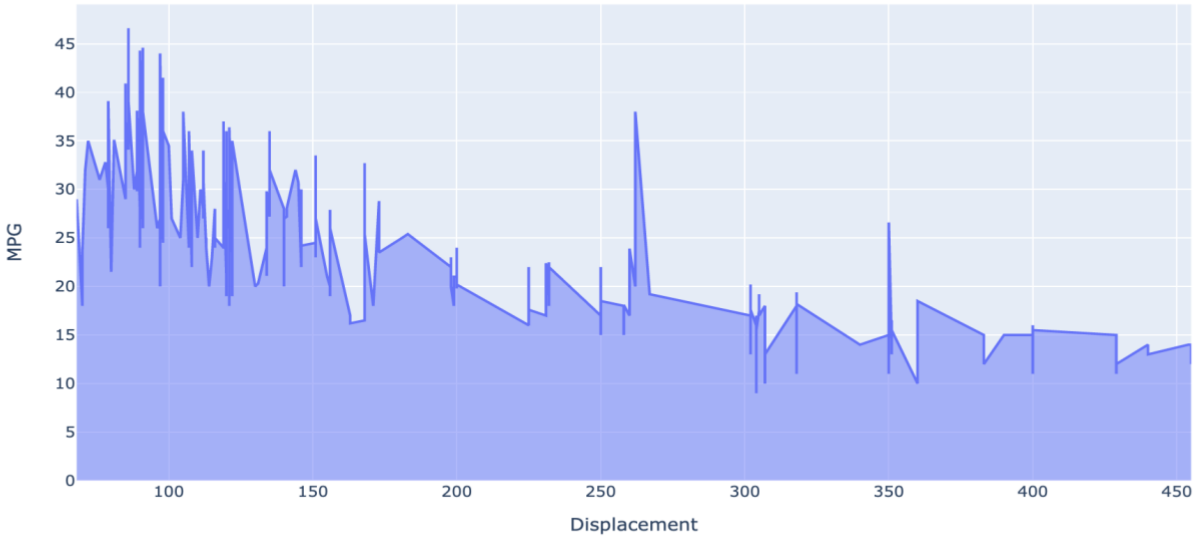


Figure 3: Area plot of displacement versus mileage.

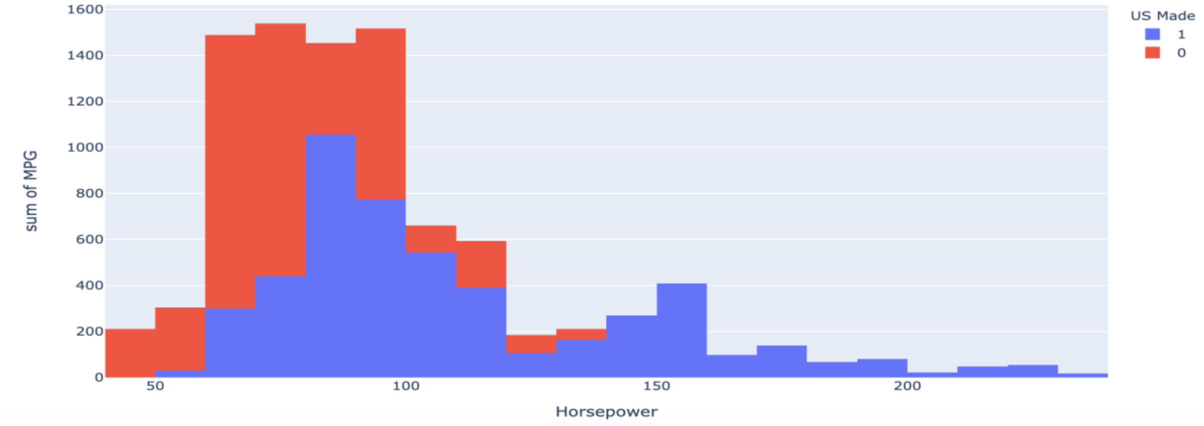


Figure 4: Horsepower versus mileage.

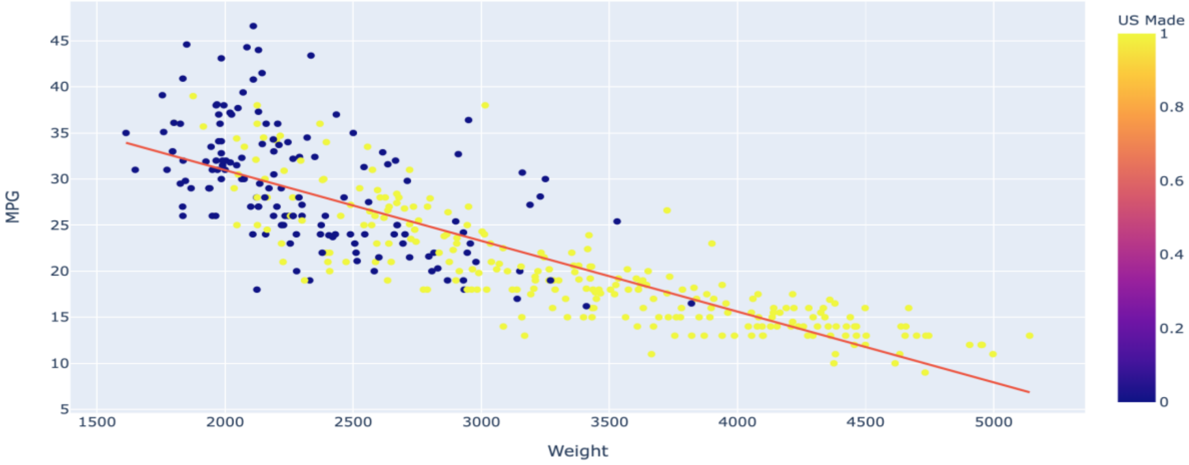


Figure 5: Scatter plot with regression line for weight versus mileage.

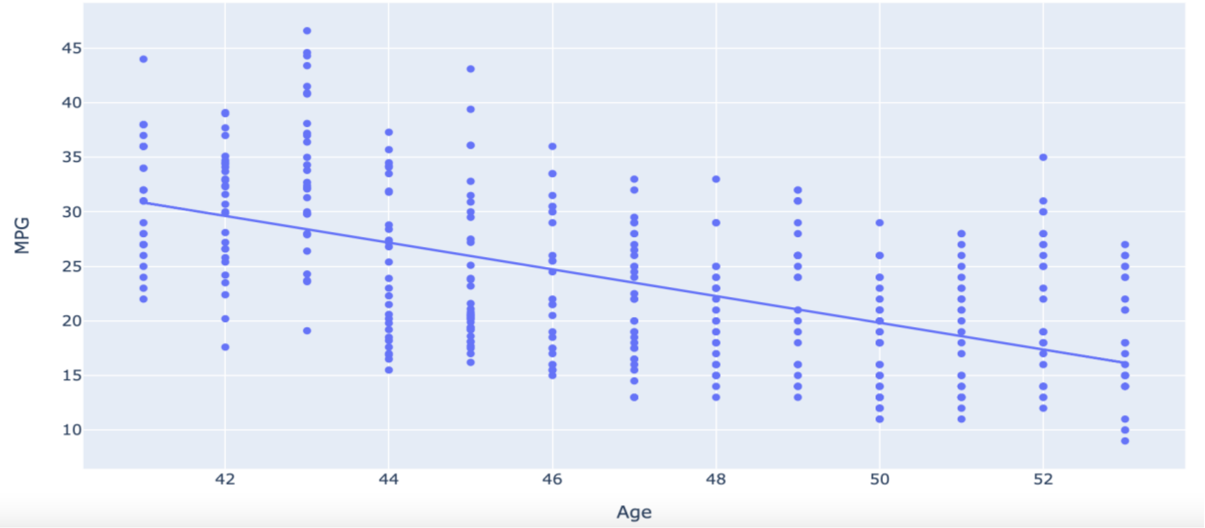


Figure 6: Scatter plot of age versus mileage.

Appendix 2 Linear Regression Assumptions

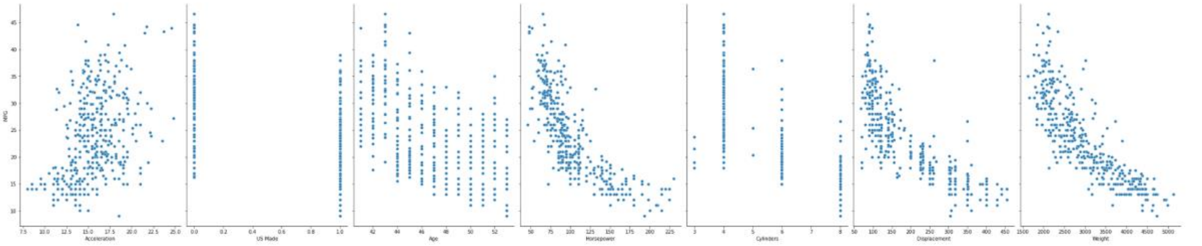


Figure 6: Linearity check.

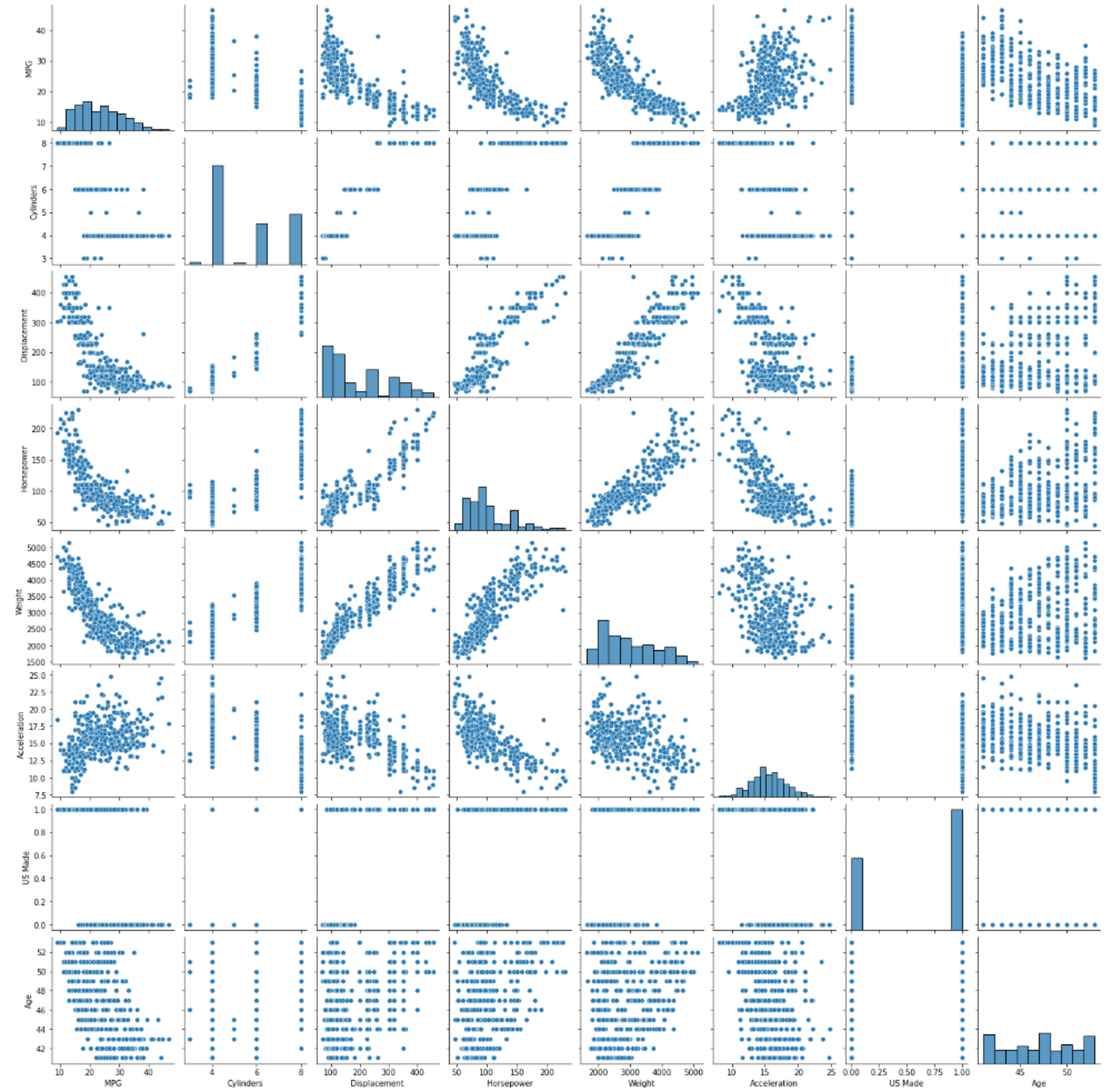
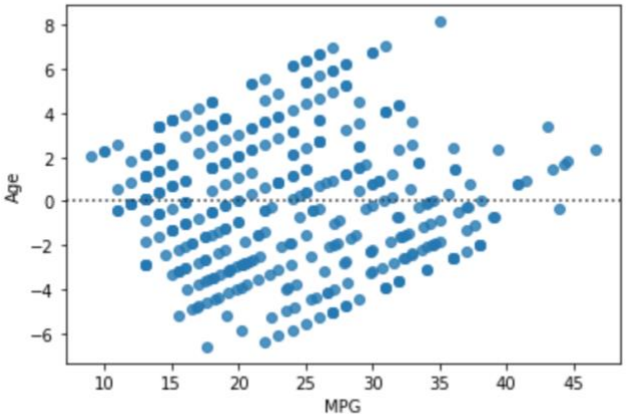


Figure 7: Independence test



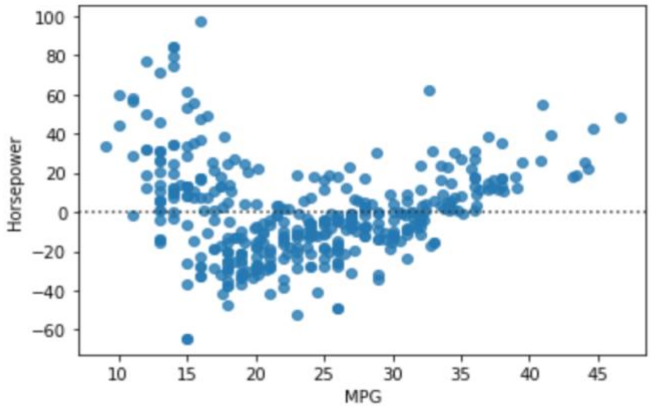
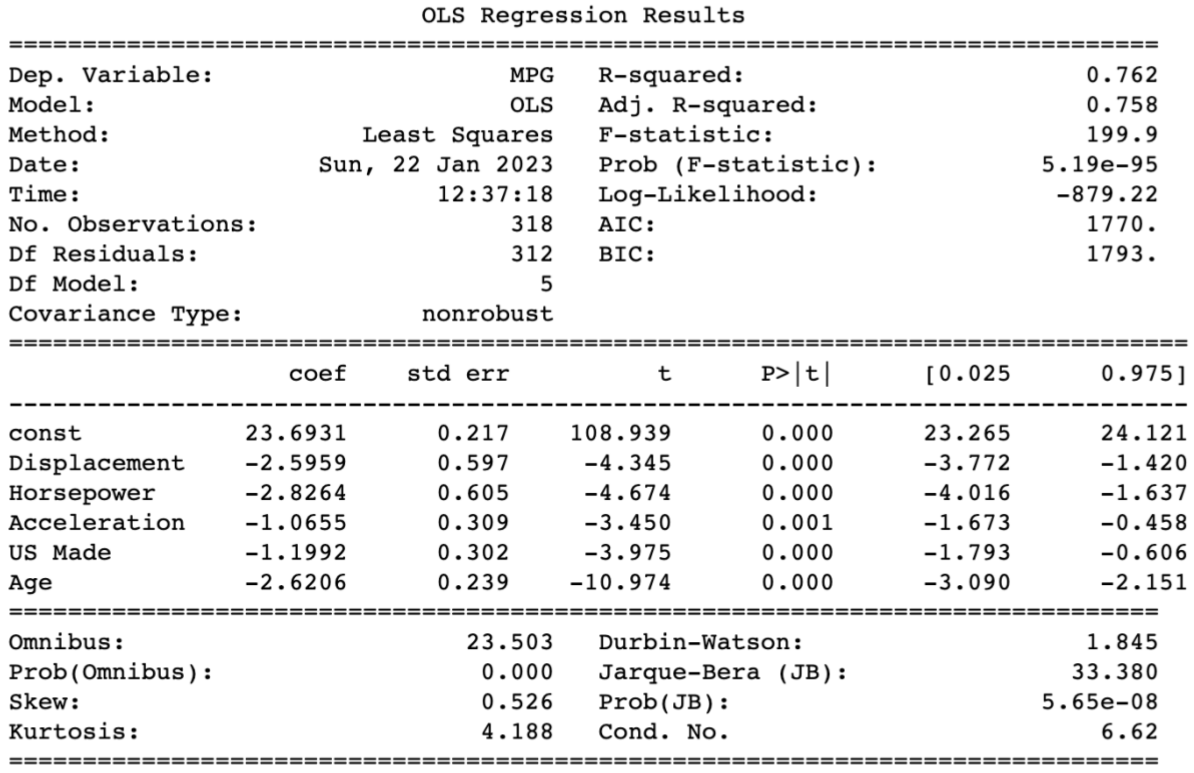


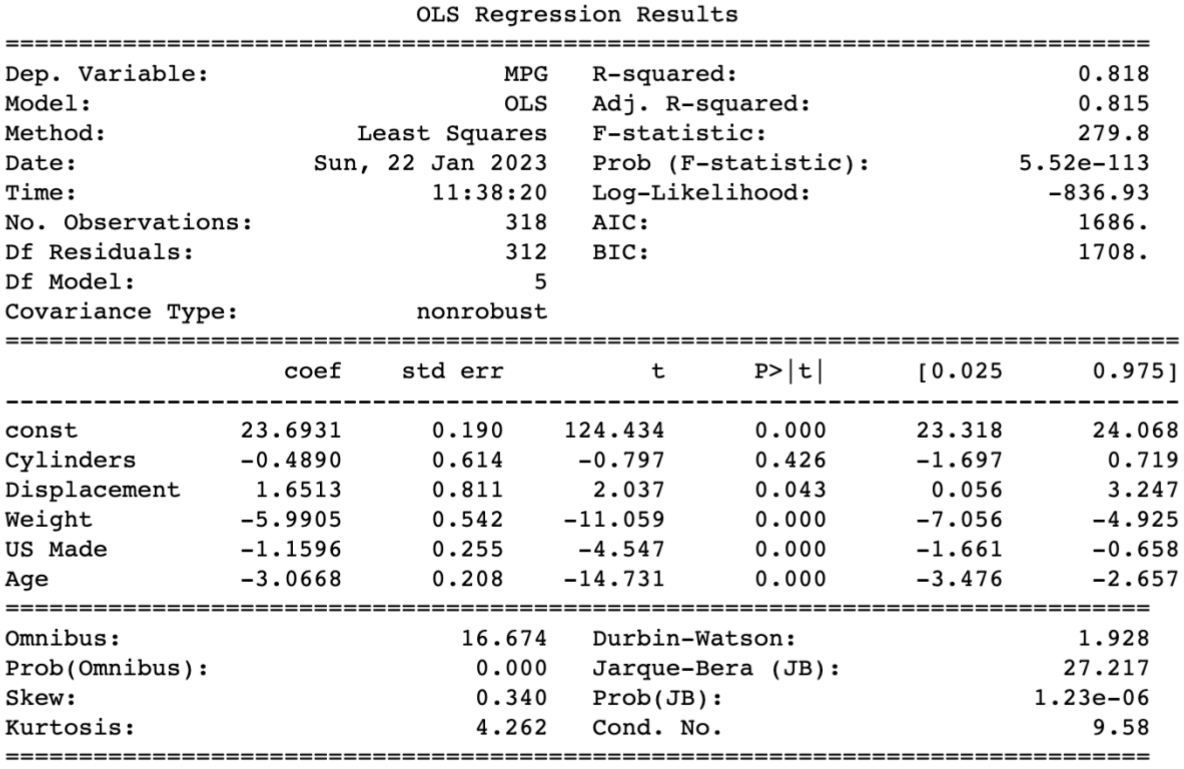
Figure 8: Test for Homoscedasticity

Appendix 3 Linear Regression Results



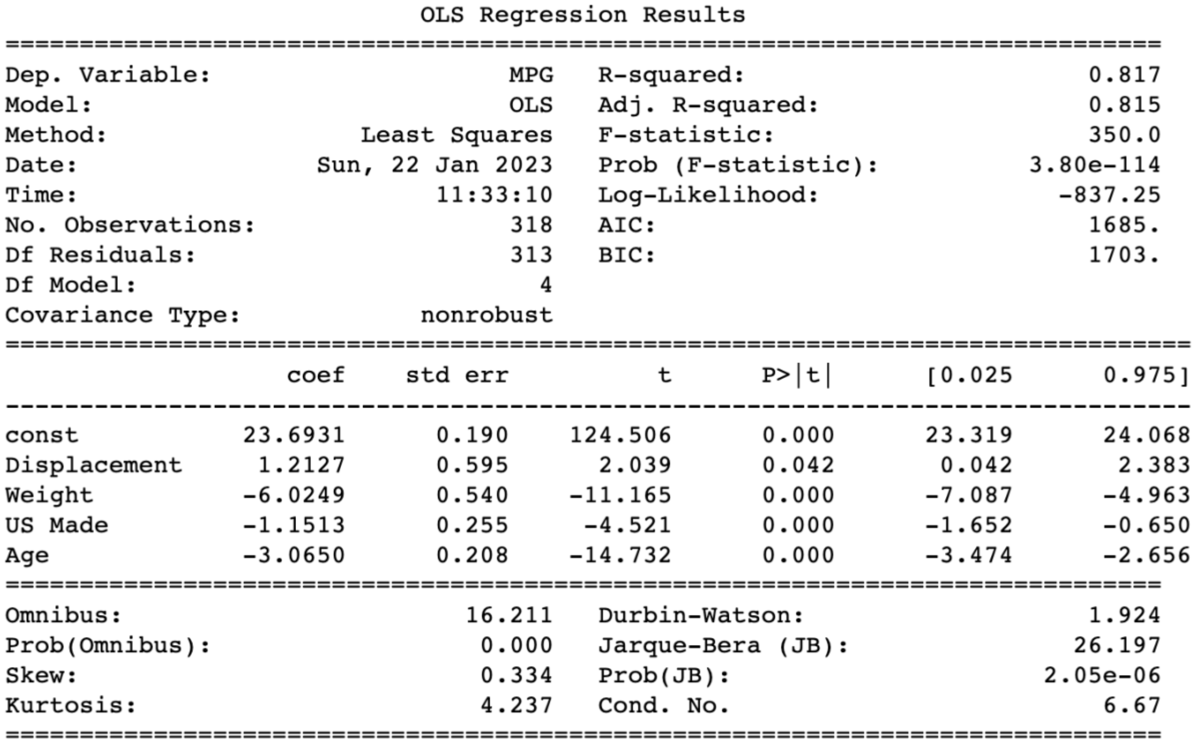
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    MPG   R-squared:                       0.762
Model:                            OLS   Adj. R-squared:                  0.758
Method:                 Least Squares   F-statistic:                     199.9
Date:                Sun, 22 Jan 2023   Prob (F-statistic):           5.19e-95
Time:                        12:37:18   Log-Likelihood:                -879.22
No. Observations:                 318   AIC:                             1770.
Df Residuals:                     312   BIC:                             1793.
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           23.6931      0.217    108.939      0.000      23.265      24.121
Displacement    -2.5959      0.597     -4.345      0.000      -3.772      -1.420
Horsepower      -2.8264      0.605     -4.674      0.000      -4.016      -1.637
Acceleration    -1.0655      0.309     -3.450      0.001      -1.673      -0.458
US Made         -1.1992      0.302     -3.975      0.000      -1.793      -0.606
Age             -2.6206      0.239    -10.974      0.000      -3.090      -2.151
==============================================================================
Omnibus:                       23.503   Durbin-Watson:                   1.845
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               33.380
Skew:                           0.526   Prob(JB):                     5.65e-08
Kurtosis:                       4.188   Cond. No.                         6.62
==============================================================================
```

Figure 9: Base model results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    MPG   R-squared:                       0.818
Model:                            OLS   Adj. R-squared:                  0.815
Method:                 Least Squares   F-statistic:                     279.8
Date:                Sun, 22 Jan 2023   Prob (F-statistic):          5.52e-113
Time:                        11:38:20   Log-Likelihood:                -836.93
No. Observations:                 318   AIC:                             1686.
Df Residuals:                     312   BIC:                             1708.
Df Model:                           5
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           23.6931      0.190    124.434      0.000      23.318      24.068
Cylinders       -0.4890      0.614     -0.797      0.426      -1.697       0.719
Displacement     1.6513      0.811      2.037      0.043       0.056       3.247
Weight          -5.9905      0.542    -11.059      0.000      -7.056      -4.925
US Made         -1.1596      0.255     -4.547      0.000      -1.661      -0.658
Age             -3.0668      0.208    -14.731      0.000      -3.476      -2.657
==============================================================================
Omnibus:                       16.674   Durbin-Watson:                   1.928
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.217
Skew:                           0.340   Prob(JB):                     1.23e-06
Kurtosis:                       4.262   Cond. No.                         9.58
==============================================================================
```

Figure 10: Forward selection optimization OLS results.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    MPG   R-squared:                       0.817
Model:                            OLS   Adj. R-squared:                  0.815
Method:                 Least Squares   F-statistic:                     350.0
Date:                Sun, 22 Jan 2023   Prob (F-statistic):          3.80e-114
Time:                        11:33:10   Log-Likelihood:                -837.25
No. Observations:                 318   AIC:                             1685.
Df Residuals:                     313   BIC:                             1703.
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           23.6931      0.190    124.506      0.000      23.319      24.068
Displacement     1.2127      0.595      2.039      0.042       0.042       2.383
Weight          -6.0249      0.540    -11.165      0.000      -7.087      -4.963
US Made         -1.1513      0.255     -4.521      0.000      -1.652      -0.650
Age             -3.0650      0.208    -14.732      0.000      -3.474      -2.656
==============================================================================
Omnibus:                       16.211   Durbin-Watson:                   1.924
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.197
Skew:                           0.334   Prob(JB):                     2.05e-06
Kurtosis:                       4.237   Cond. No.                         6.67
==============================================================================
```

Figure 11: Backward selection optimization OLS results.