COE 379L Project 01: Used Cars

This project deals with a used car database, which provides mileage, horsepower, model year, and other specifications.

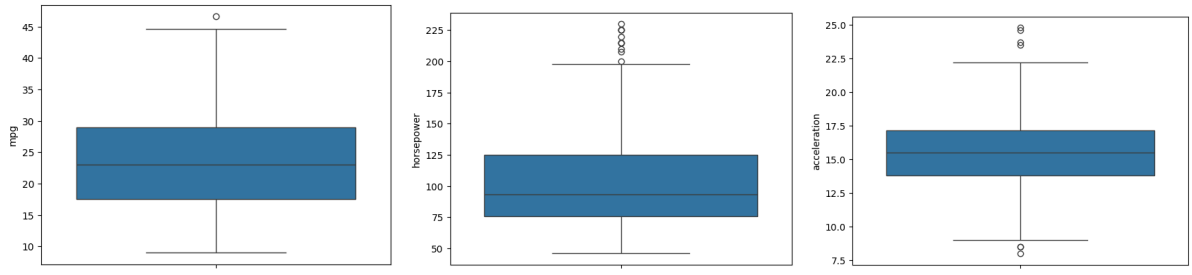
Preparing the Data

When looking at the data, the first thing I noticed was the horsepower for some vehicles was listed as "?". To fix this, these values were set as the median horsepower of the entire set. Now the column could be correctly saved as float values.

No null values or duplicate rows/columns were found, so we can now manipulate the categorical data to make it easier to find a correlation. I changed the origin column from a set of 1, 2, and 3 values into a pair of one-hot-encoded columns. Now we can analyze the data.

Data Insights

The first thing we can do to analyze the data is to find the data distributions for each column. A few columns had outliers, as shown in their box plots:



Figures 1, 2, and 3: MPG, Horsepower, and Acceleration

Similarly we can find correlation between different columns using a heatmap from Seaborn:

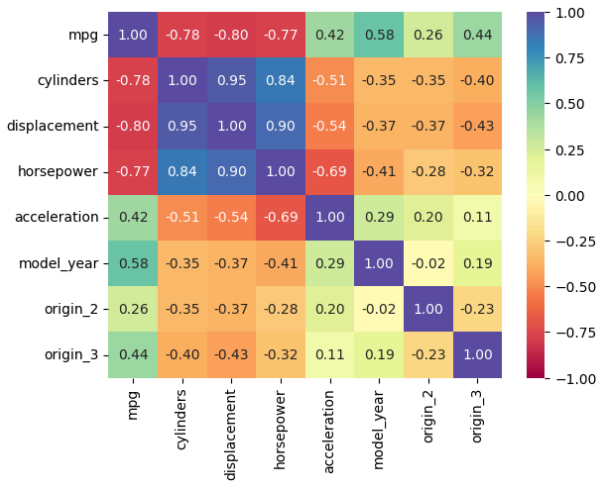
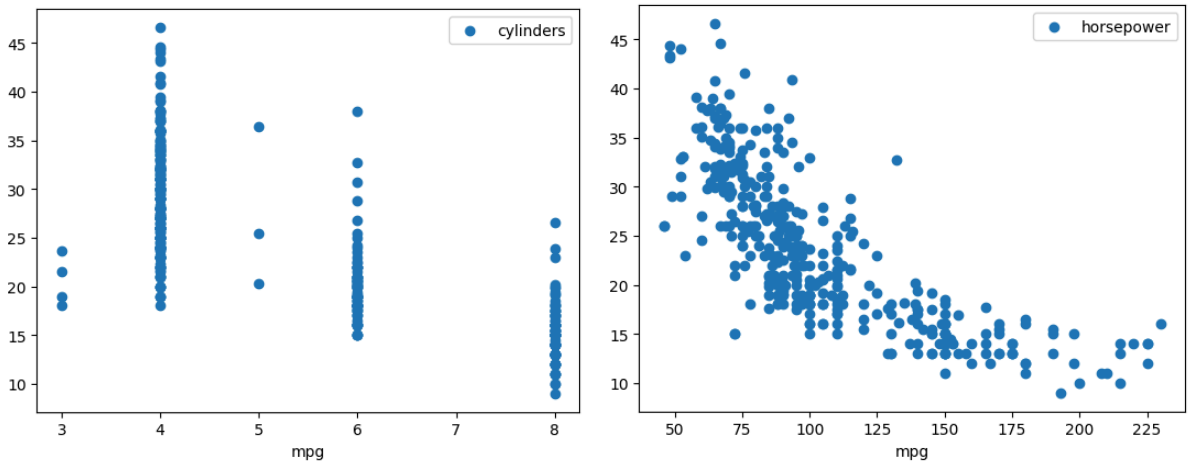


Figure 4: Correlation Heatmap

This gives us insight into which variables are most likely to affect the fuel efficiency. It seems that the cylinders, displacement, and horsepower are negatively correlated to MPG, while they are positively correlated to each other. All other correlations are too weak to mean anything.



Figures 5 and 6: Cylinders and Displacement vs MPG

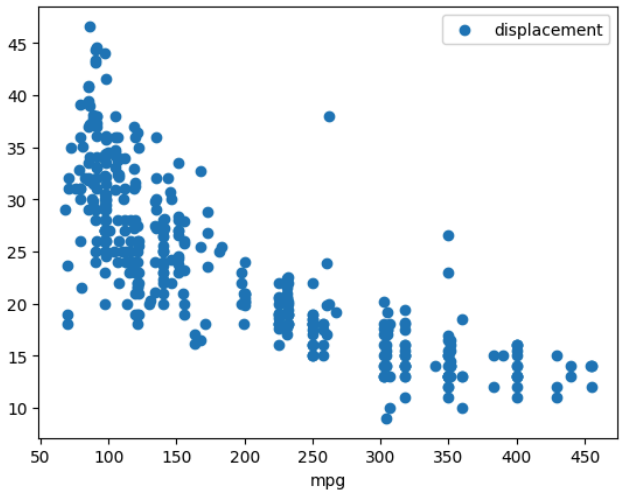


Figure 7: Displacement vs MPG

Training the Model

For this problem, we will be using a linear regression model to predict the MPG of each car. We must first split the data into a training set and test set at a ratio of 7:3. I decided to use the independent variables that are correlated to MPG from the previous section (cylinders, displacement, and horsepower). I then fit a linear regression model from Scikit Learn with these variables.

Model Accuracy

When testing this model on the training data, it gave an accuracy score of 0.647, while testing it on the test data, it gave an accuracy score of 0.700. This score comes from the built in score function of the linear regression model. When using the model to predict values from the data set, here are some results that I found:

Expected MPG	16.92	13.22	15.28	15.77	16.47	8.388	6.130	6.934	5.781	10.25
Predicted MPG	18	15	18	16	17	15	14	14	14	15

When looking at this example data, some of the predicted values appear to be close to the correct values. Meanwhile, other predicted values are quite far from the correct values. Because of this and the accuracy score, I would say I am not very confident in the model to predict MPG correctly.