Ethan Colley

Project 1-Linear Analysis

For this project we were tasked with formatting a dataset of automobile information and produce a linear model that predicts fuel efficiency. To do this, we must first perform exploratory analysis on the dataset of automobiles. After this is done, then we can fit a linear regression model to the data and find a prediction for fuel efficiency.

Part 1:

To prepare our data for regression, we must first find all information about the dataset. After importing pandas and numpy, we get the shape of the dataset and find that there are 398 rows and 9 columns. By using the head function we can also see each columns specifically, those being mpg, cylinders, displacement, horsepower, weight, acceleration, model_year, origin, and car_name. After viewing the information in the dataset and trying to learn some information about the data, we see that there are multiple “?” in the horsepower column, and the column type is set to object. To perform our analysis, we must change this column type to a float so it can be graphed along with the other columns. I changed these “?” values to null values and then used the “.astype()” function to change the columns data type from an object to a float. From here I used the “.fillna” function to replace the null horsepower values in each column to the mean horsepower values from the rest of the columns. This made it so that each column had the same number of arguments, and there were no null values. I also decided to drop the “origin” and “car_name” columns from here, as I felt that the continent of origin and name of the car would have no real effect on the horsepower. Each continent and car brand make a multitude of different models with different efficiencies, so these columns are not going to have much to do with the fuel efficiency. With this data sorted and refined, we can now do the actual analysis of the data. The first plot that I made had to do with model year. I feel that fuel efficiency focus could change as the years progressed, and wanted to be sure there were many examples of vehicles from each year. I made a histogram and found there was no real skew to the data, meaning there was an even distribution of vehicles from 1970 to 1982. I also made a boxplot of this data for another visualization, which confirms that the data is extremely symmetrical. I also made a histogram of the cylinders, finding that most of the cars in this dataset have 4 cylinders, but there are still decent amounts of cases with 6 and 8 cylinders as well. The final univariate graph I made was a boxplot for the horsepower, where the mean horsepower seemed to be around 100 with outliers going up to 225, but the majority of the horsepowers remain around 75-125. Next I made a multivariate heatmap using all remaining columns in the dataset. This map shows that cylinders, displacement, horsepower, and weight all have a strong negative correlation with mpg(fuel efficiency). These columns also all have strong correlation with each other, and acceleration and model year do not seem to have a strong correlation with any column.

Part 2:

With analysis out of the way, I began on the linear regression model. The first step was creating a training and test set using the data. I chose to use the “train_test_split” model from sklearn, and set my test size to 0.5 so that the test and training sets would have equal shapes. I set X to take every column but mpg and y to be mpg, then used the linear regression model from sklearn. We see in our model that the test set has a score of 0.815 where a score of 1 is most accurate and a score of 0 is least accurate(I asked Chatgpt how to implement this score function). The training prediction had a score of 0.799 so the accuracy of the test prediction is higher. This would be the more preferable measurement for this problem since the test set was used after the model had been trained with the training data. I would express a fair confidence in the model since our test set was roughly 80% correct.