Part 1: Exploratory Data Analysis

To prepare the data, I first opened it to see how many variables I was dealing with and what the values looked like. The shape and size of the dataset was 398 rows and 9 columns. Of these 9 variables, 2 of them were of the type “object”. The 2 that were type “object” were horsepower and car_name. When I looked at the values for horsepower, I noticed most of them were numeric values, however, there were a couple rows that contained a “?”. I checked which rows had a “?” entered as the value for horsepower and used those indices to drop those rows from the dataset. In total, I dropped 6 rows from the dataset. I then converted the values from strings to floats. In the end, the dataset had 392 rows and 9 columns, and no data was missing. Next, I performed one-hot encoding on the categorical data. I felt that the ‘origin’ variable was a great example of a categorical variable. The ‘origin’ variable had 3 categories: 1=American, 2=European, and 3=Japanese. Then I converted the categorical variable to a set of bit columns: ‘origin_2’ and ‘origin_3’.

I derived statistical information from the data and hypothesized that weight, horsepower and displacement were most likely to have outliers since they had the highest standard deviation of all the variables. When I did univariate and bivariate analysis plots, I made a boxplot of horsepower and found that it did in fact contain a couple of outliers. I made a histogram of fuel efficiency, since that was what this project is centered on. I found that fuel efficiency has a right skewed distribution, with most vehicles having a fuel efficiency of about 15-17 mpg. I hypothesized that the variable most likely to have the largest effect on fuel efficiency would be the vehicle’s weight, with heavier cars being less efficient than lighter cars. To visualize this, I created a scatter plot with vehicle weight on the x axis and fuel efficiency on the y axis. The scatter plot showed an overall negative trend. As vehicle weight increases, fuel efficiency

decreases. This relationship was further confirmed by the correlation matrix that I created. The correlation coefficient for fuel efficiency and weight was -0.83, meaning they are negatively correlated. The matrix also confirmed that weight had the greatest correlation with fuel efficiency. Horsepower and displacement also had correlation coefficients close to -1. Acceleration, on the other hand, proved to be a poor predictor of fuel efficiency, with a correlation coefficient of 0.42 .

Part 2: Linear Regression Fit

I first defined *x* to be all of the values except for fuel efficiency and car names so that I could use as much data as possible. I defined *y* to be all of the values of fuel efficiency, since that is what we are trying to predict. Then, I split the data into training and testing data, using 70% for training and 30% for testing. Next, I fit a linear regression model on the training data. To look at the accuracy of the model's ability to predict fuel efficiency for both the training and testing data, I used the score function on the linear regression model. For the testing data, the model had an accuracy of 0.824 . For the training data, the model had an accuracy of about 0.821 . Since both of these values are pretty close to 1, and the testing data has a higher accuracy than the training data, I am fairly confident in the model's ability to predict the fuel efficiency of a vehicle.