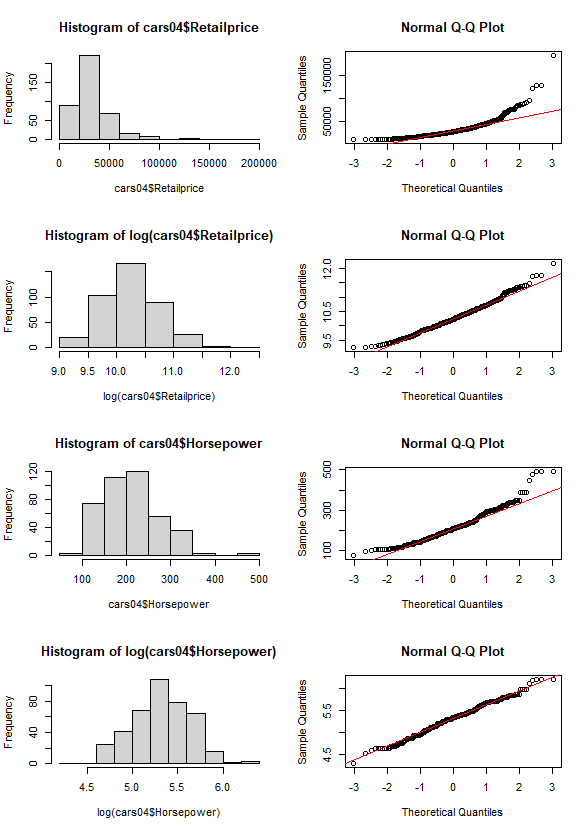
Midterm Project

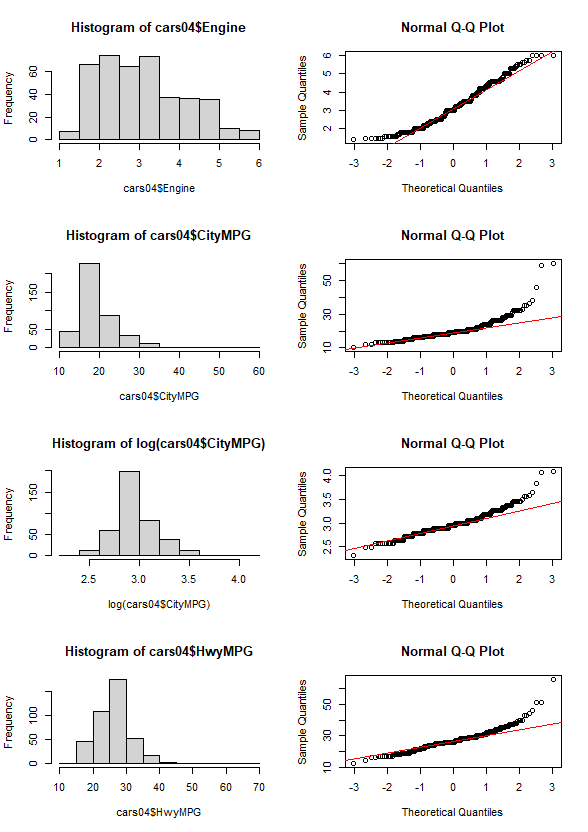
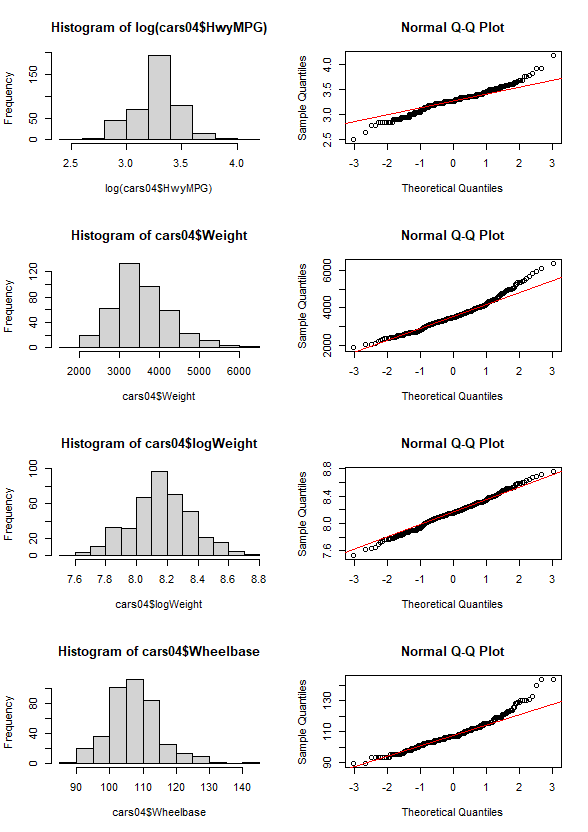
Jeff Watson

For the Midterm Project, I chose the 04cars data set. The data set contains statistics on new car and truck models from 2004. The response variable I’ll be checking is Retailprice. The potential predictor variables are: Engine (in terms of liters), Cylinders, Horsepower, CityMPG, HwyMPG, Length, Height, Weight, Wheelbase, AWD, RWD, and 5 categories of vehicle (Sport, SUV, Wagon, Minivan, Pickup). I will use both elastic net and robust regression models to makes my predictions.

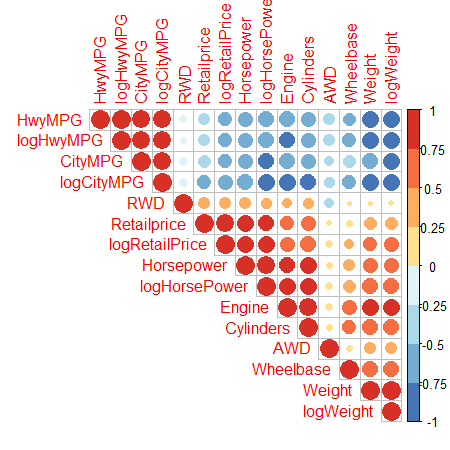
Step one is to load up the data and look at it. The first thing I notice is that `Pickup` has NAs for Length and Height. To preserve `Pickup` as a predictor variable, I will eliminate these variables. There are still a few NAs scattered about the data frame, but they don’t appear to be part of a pattern. Most of them are in the two MPG predictor variables. I will omit these rows from the data set. That brings the total rows down to 410 from 428. The next step is to factor the vehicle types and create a new variable `Type` and use it to store the 5 vehicle categories plus an `Other` category for when none of the model types are defined.

It's time to check normality and adjust for skewness where appropriate. I’ll do this with histograms and Normal Q-Q plots. As you can see, there are several predictor variable that needed to be log transformed due to their skewness, namely `Retailprice`, `Horsepower`, both MPG variables, and `Weight`.

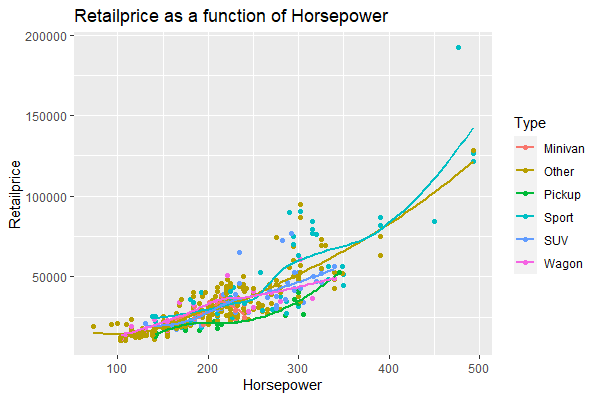


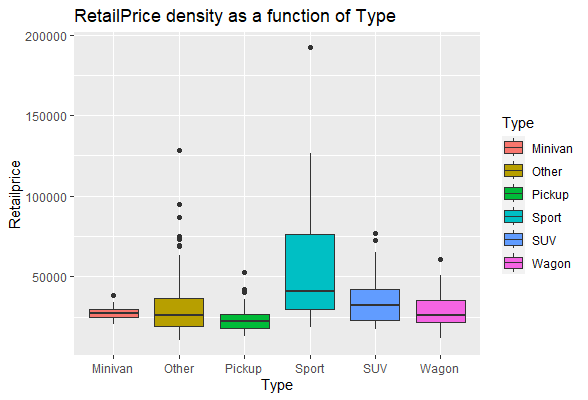


Next, I want to check the data set for collinearity. I will employ a corrplot. There is a lot of positive and negative correlations within the data. A lot of this is logical and can be reasoned out. For instance, `Engine` is measured in liters, and `Cylinders` are what fill up those liters. Larger engines require either larger cylinders or more cylinders. Larger cylinders tend to be the American solution. If there were a country of origin, there would probably be a higher correlation between 6-cylinder and 8-cylinder vehicles as opposed to 4-cylinder and >8-cylinder vehicles (the United States hasn’t had a production 12-cylinder engine since the 1940s). Weight and Engine have a high correlation, which makes sense as larger engines are heavier. Equally, Weight and the MPG variables have a negative correlation as it takes more fuel to move a heavy vehicle, and the positive correlation between Weight and Engine helps demonstrate that larger, heavier engines use more fuel to operate.



Let’s look at interactions between variables. First we’ll look at the interaction between `Horsepower` and `Retailprice`. As you can see below in the scatterplot, `Horsepower` shows a positive impact on `Retailprice`, especially with the `Sport` class. This makes sense as high-horsepower sportscars are the expensive dream cars found on posters on the walls of teenage boys. The boxplot helps demonstrate this interaction.





It's time to start eliminating predictor variables that don’t add anything to the models we will use in our models to predict Retailprice. I will employ the elastic net method to pick out variables with zero impact on the model. I will run a glmnet method via the caret package to run multiple, successive, 10-fold cross-validations beginning with all predictor variables and remove unnecessary predictors with each pass until I reach a stable model. For this data set, 3 cycles revealed the most appropriate predictors: logRetailPrice, Cylinders, logHorsePower, logWeight, Type, AWD, and RWD.