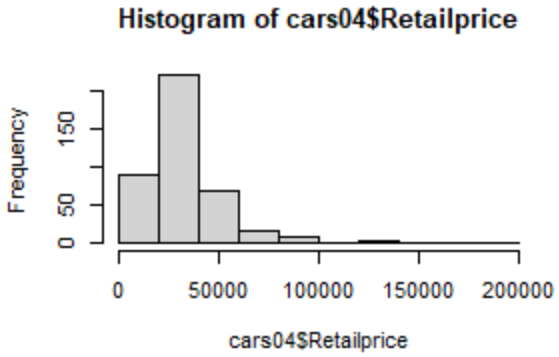
Midterm Project

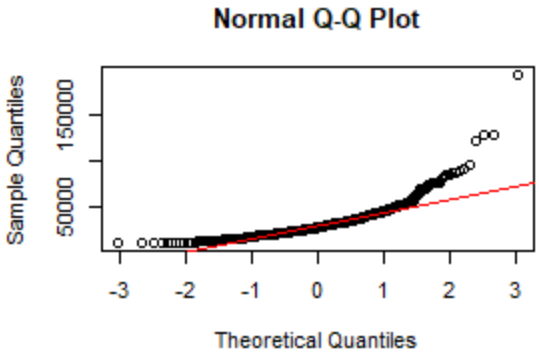
Jeff Watson

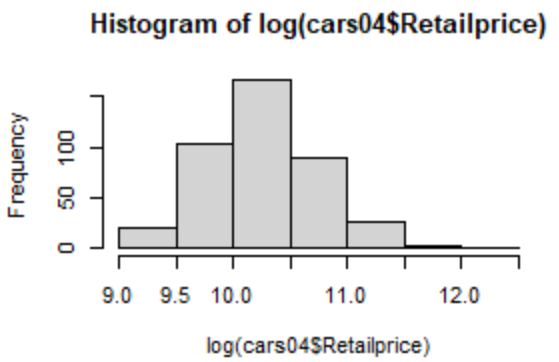
Determining the correct price for a product is one of the most important decisions a company can make. It is a key driver in the profitability of a business. To turn a profit, the retail price of a good must cover both cost and labor and be palatable to potential consumers. The price must also convey the perceived value of the product to the customer. If the price is too low, the consumer may think that the item is of poor quality. If the price is too high, the consumer may be off put and/or the good is out of their expected price range and not purchase it. Pricing is a delicate balance of showing value while remaining within the budget of the selected consumer slice. In this exercise, I will attempt to construct a model that will accurately predict the price of a new vehicle from 2004.

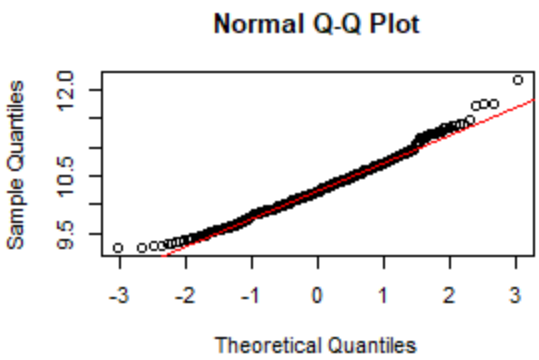
Step one is to load up the data and look at it. The first thing I notice is that `Pickup` has NAs for `Length` and `Height`. To preserve `Pickup` as a predictor variable, I will eliminate these variables. There are still a few NAs scattered about the data frame, but they don't appear to be part of a 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.

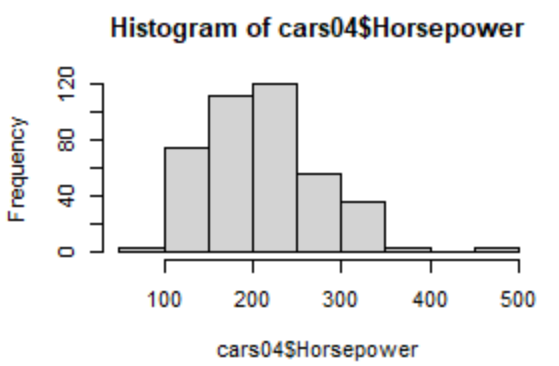
It's time to check normality and adjust for skewness where appropriate. I'll do this with histograms and Normal Q-Q plots. As you can see, there are several predictor variables that needed to be log-transformed due to their skewness, namely `Retailprice`, `Horsepower`, both MPG variables, and `Weight`.

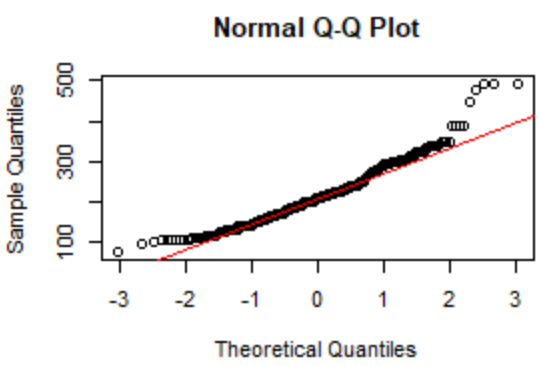

Histogram of cars04$Retailprice
Frequency
0
50
150
0
50000
100000
150000
200000
cars04$Retailprice

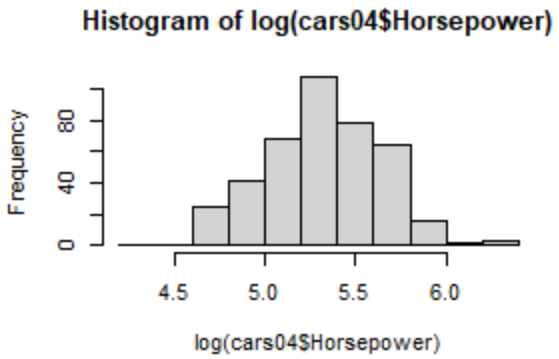

Normal Q-Q Plot
Sample Quantiles
50000
150000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

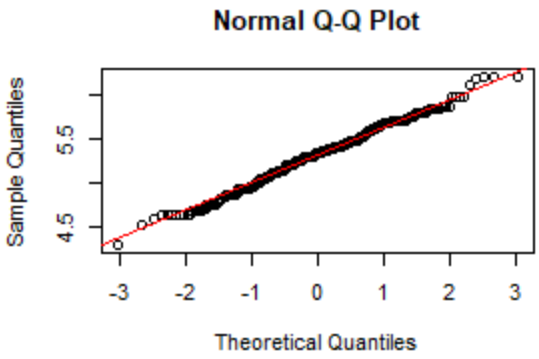

Histogram of log(cars04$Retailprice)
Frequency
0
50
100
9.0
9.5
10.0
11.0
12.0
log(cars04$Retailprice)

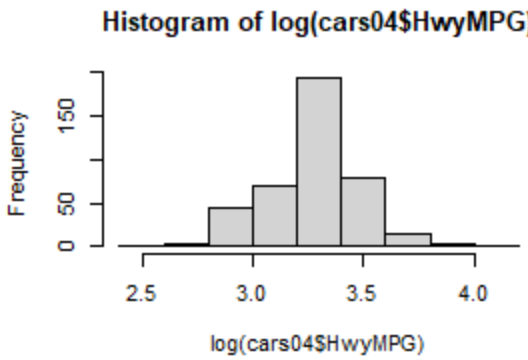

Normal Q-Q Plot
Sample Quantiles
9.5
10.5
12.0
-3
-2
-1
0
1
2
3
Theoretical Quantiles

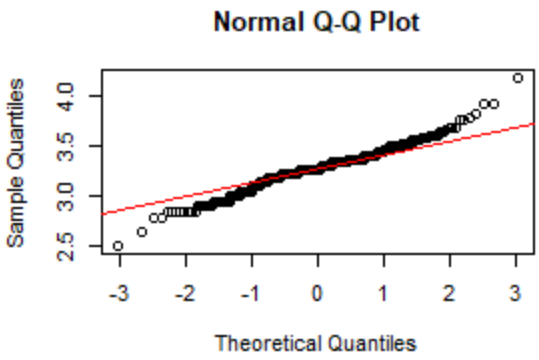

Histogram of cars04$Horsepower
Frequency
0
40
80
120
100
200
300
400
500
cars04$Horsepower

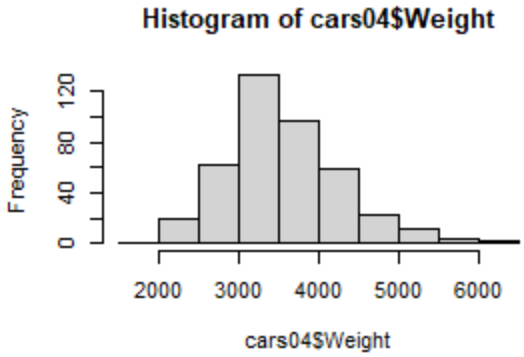

Normal Q-Q Plot
Sample Quantiles
100
300
500
-3
-2
-1
0
1
2
3
Theoretical Quantiles

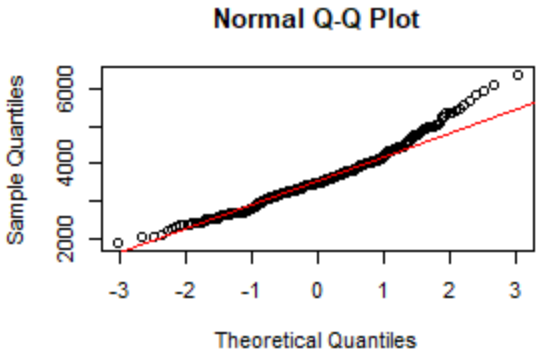

Histogram of log(cars04$Horsepower)
Frequency
0
40
80
4.5
5.0
5.5
6.0
log(cars04$Horsepower)

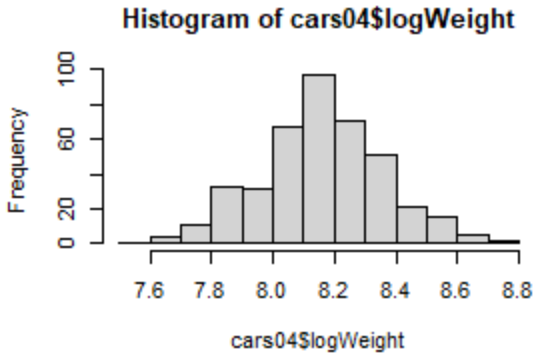

Normal Q-Q Plot
Sample Quantiles
4.5
5.5
-3
-2
-1
0
1
2
3
Theoretical Quantiles

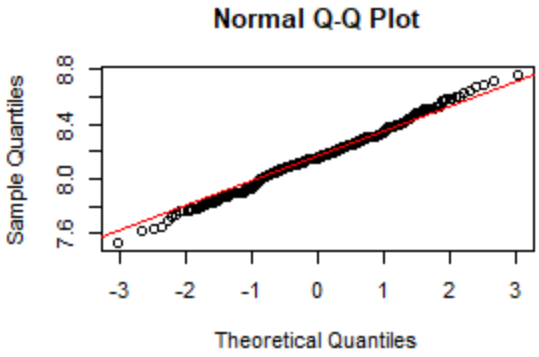

Histogram of log(cars04$HwyMPG)
Frequency
0
50
150
2.5
3.0
3.5
4.0
log(cars04$HwyMPG)

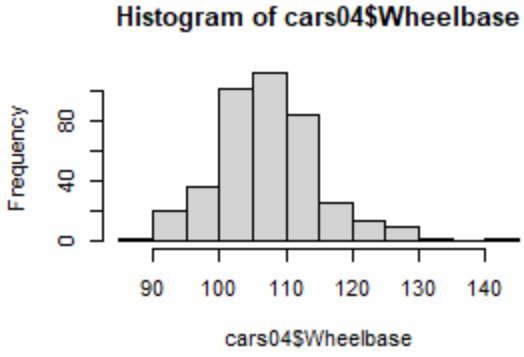

Normal Q-Q Plot
Sample Quantiles
2.5
3.0
3.5
4.0
-3
-2
-1
0
1
2
3
Theoretical Quantiles

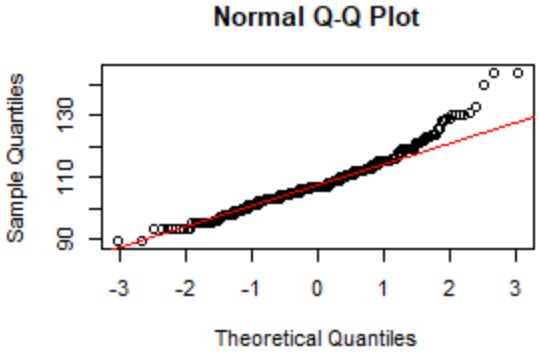

Histogram of cars04$Weight
Frequency
0
40
80
120
2000
3000
4000
5000
6000
cars04$Weight

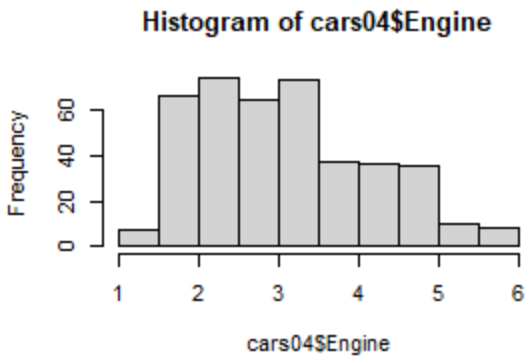

Normal Q-Q Plot
Sample Quantiles
2000
4000
6000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

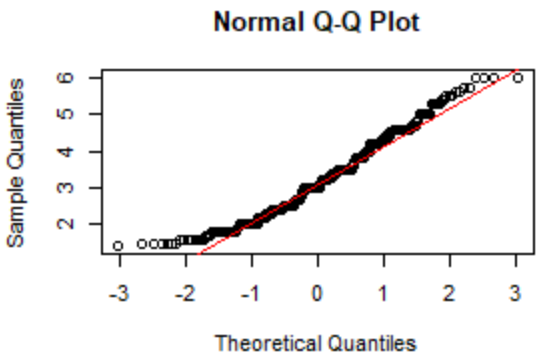

Histogram of cars04$logWeight
Frequency
0
20
60
100
7.6
7.8
8.0
8.2
8.4
8.6
8.8
cars04$logWeight

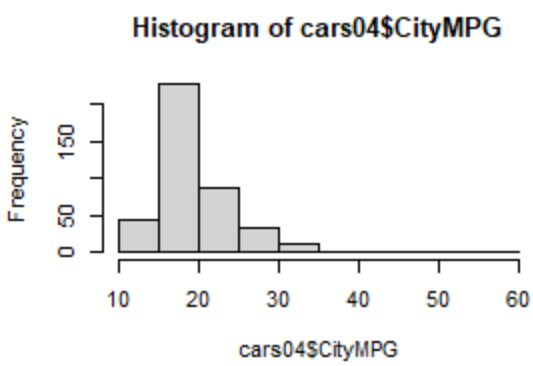

Normal Q-Q Plot
Sample Quantiles
7.6
8.0
8.4
8.8
-3
-2
-1
0
1
2
3
Theoretical Quantiles

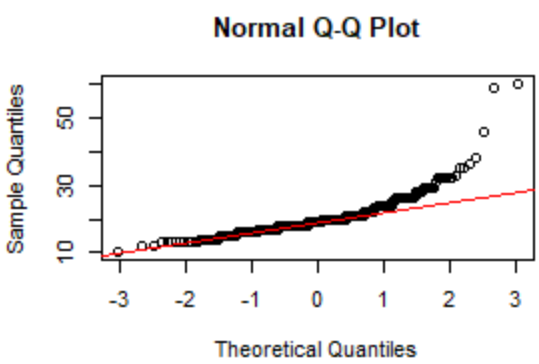

Histogram of cars04$Wheelbase
Frequency
0
40
80
90
100
110
120
130
140
cars04$Wheelbase

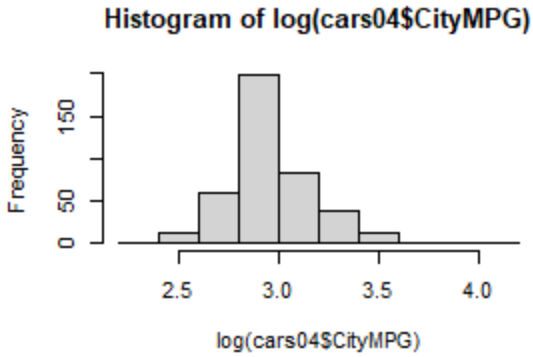

Normal Q-Q Plot
Sample Quantiles
90
110
130
-3
-2
-1
0
1
2
3
Theoretical Quantiles

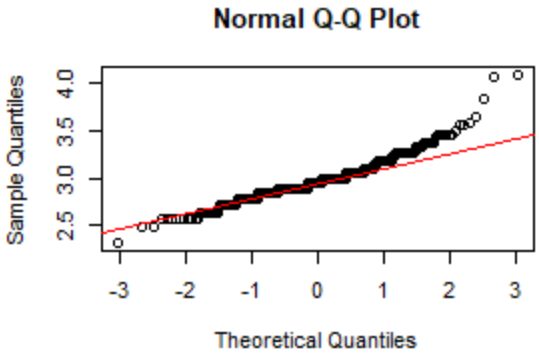

Histogram of cars04$Engine
Frequency
0
20
40
60
1
2
3
4
5
6
cars04$Engine

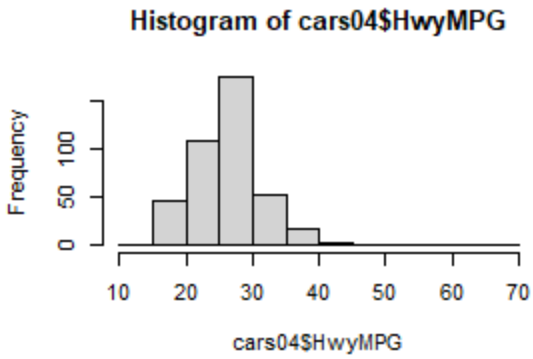

Normal Q-Q Plot
Sample Quantiles
2
3
4
5
6
-3
-2
-1
0
1
2
3
Theoretical Quantiles

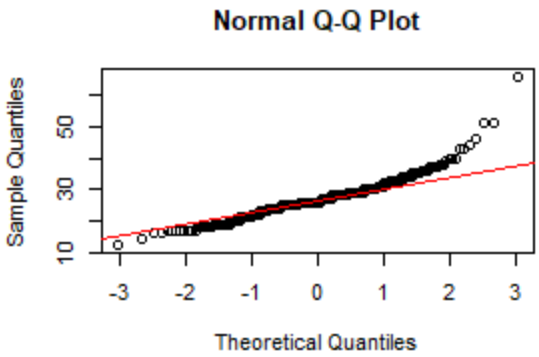

Histogram of cars04$CityMPG
Frequency
0
50
150
10
20
30
40
50
60
cars04$CityMPG


Normal Q-Q Plot
Sample Quantiles
10
30
50
-3
-2
-1
0
1
2
3
Theoretical Quantiles

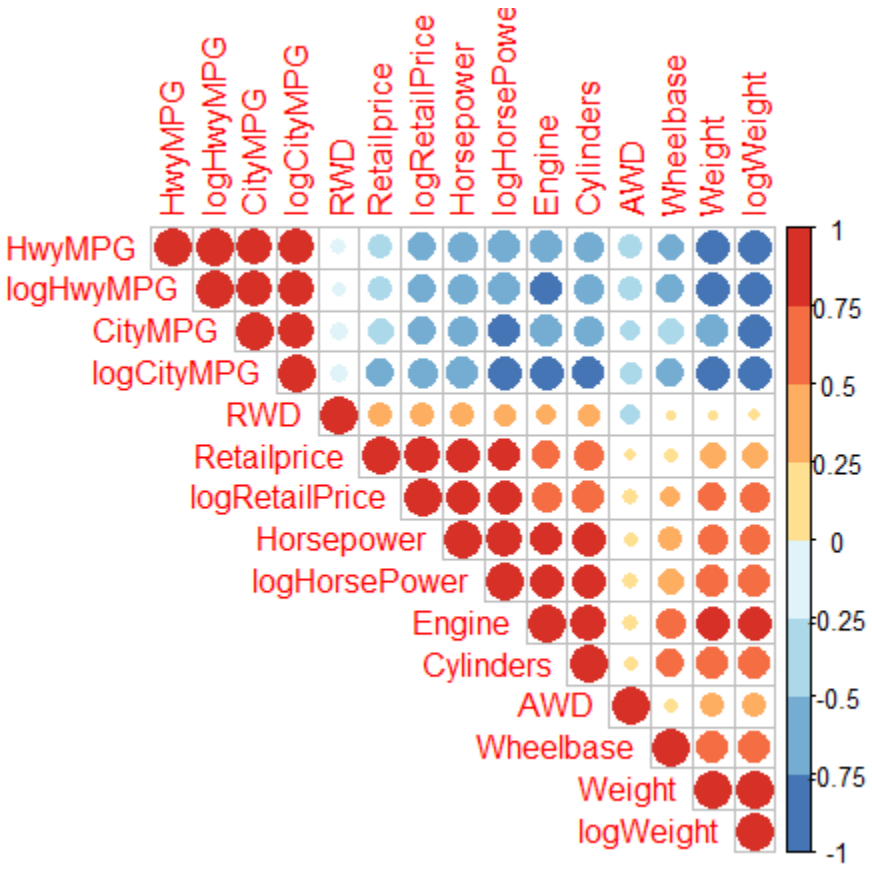

Histogram of log(cars04$CityMPG)
Frequency
0
50
150
2.5
3.0
3.5
4.0
log(cars04$CityMPG)


Normal Q-Q Plot
Sample Quantiles
2.5
3.0
3.5
4.0
-3
-2
-1
0
1
2
3
Theoretical Quantiles

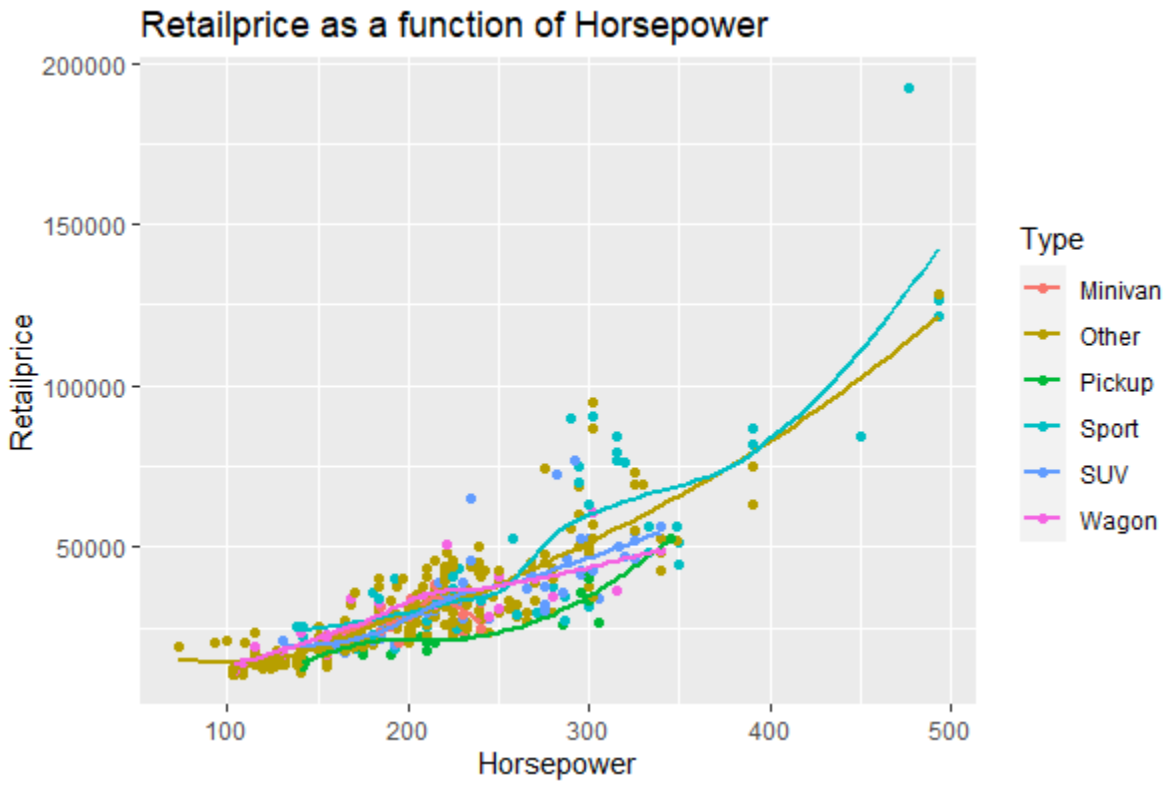

Histogram of cars04$HwyMPG
Frequency
0
50
100
10
20
30
40
50
60
70
cars04$HwyMPG

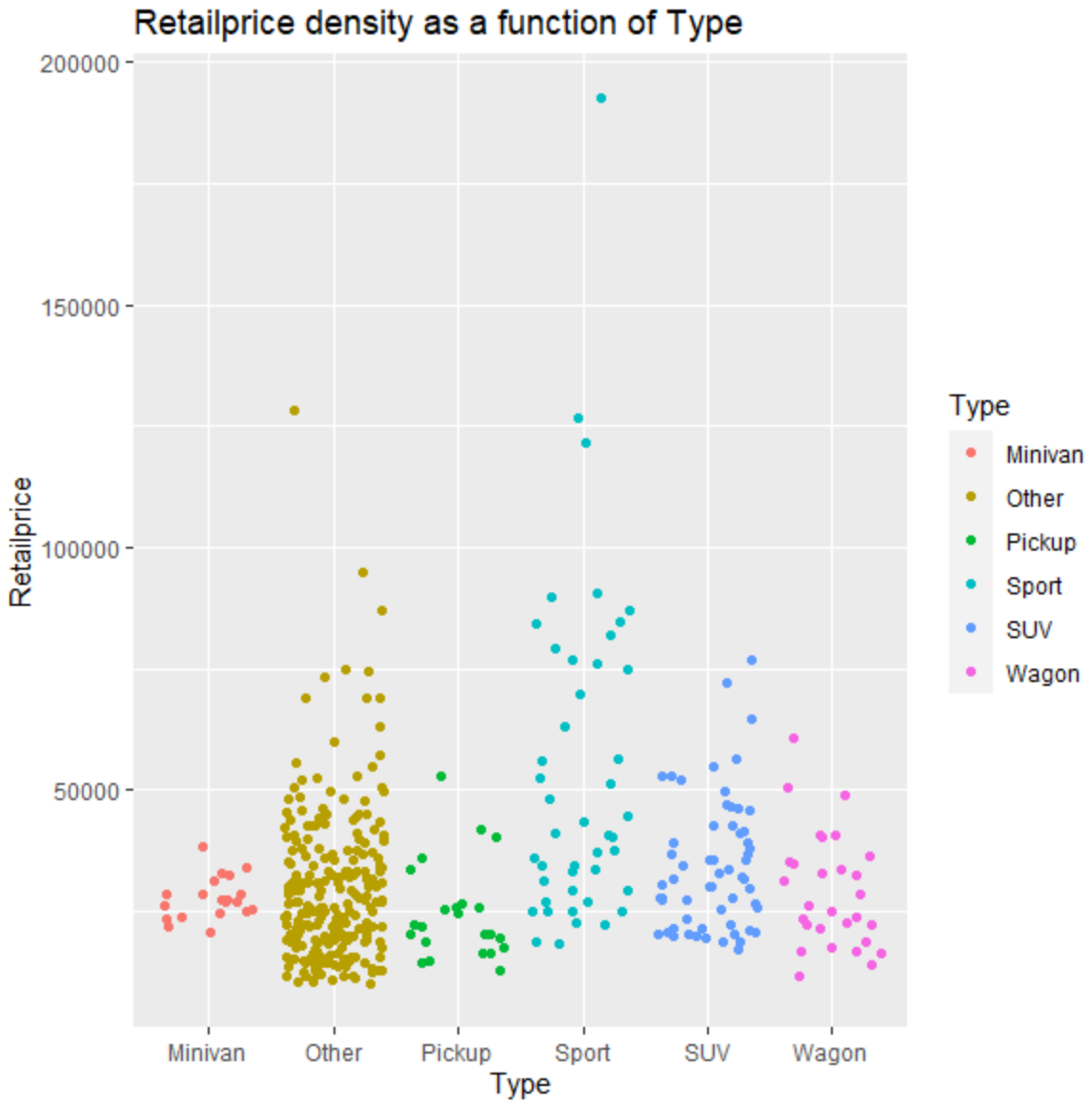

Normal Q-Q Plot
Sample Quantiles
10
30
50
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Next, I will check the data set for collinearity. I will employ a corrplot. There are 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, and heavier vehicles are required to move around big engines. 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 jitter plot helps demonstrate this interaction.



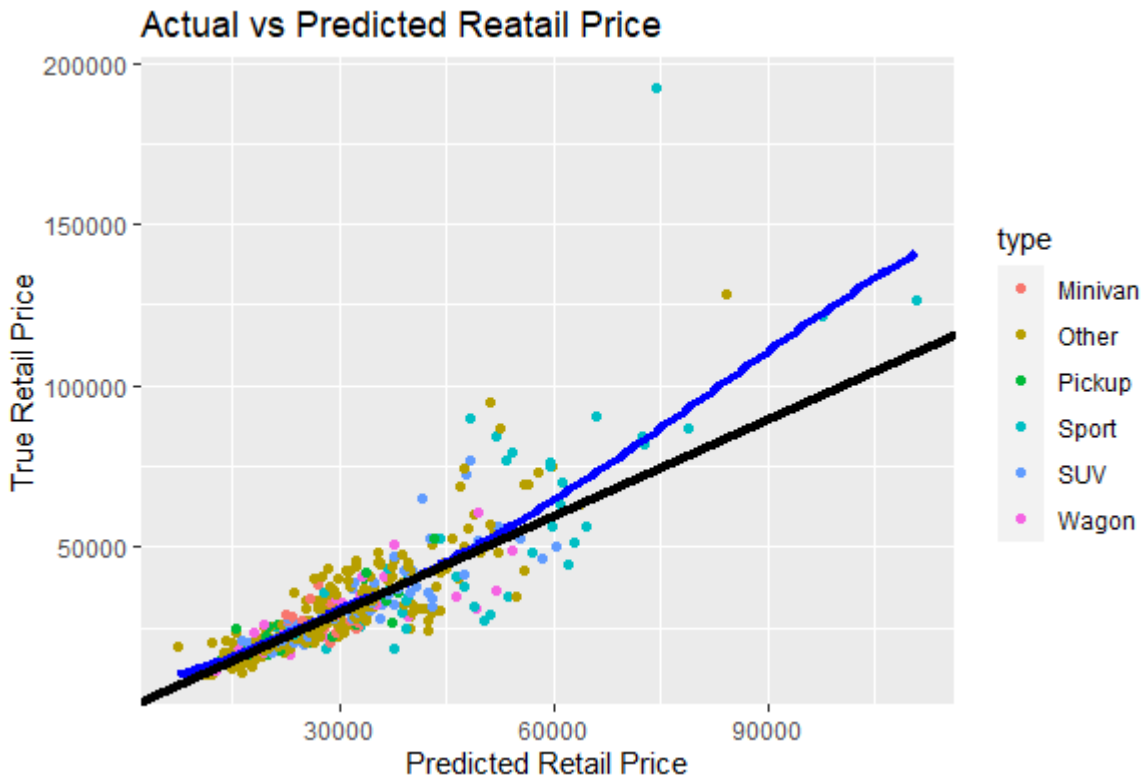


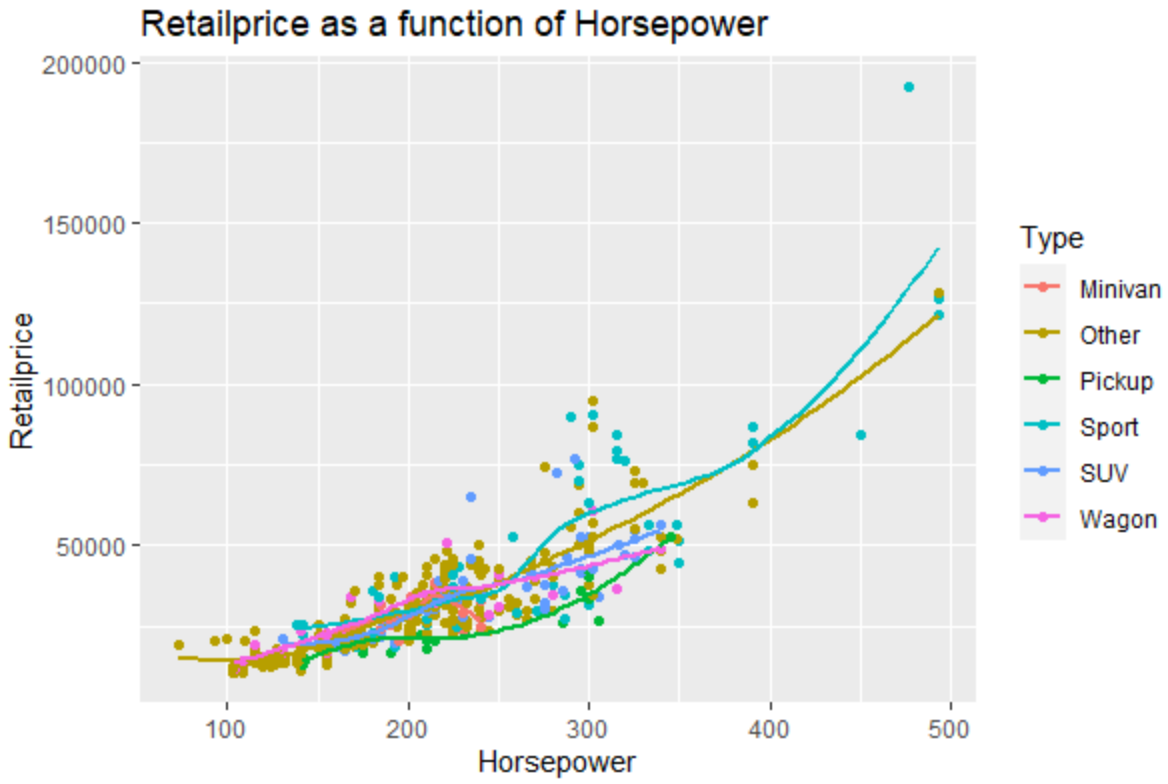
It's time to start eliminating predictor variables that don't add anything to the models we will use in our models to predict Retailprice. I will employ the elastic net method to select variables with zero impact on the model. I will run a glmnet method via the caret package to run multiple, successive, 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: Cylinders, logHorsePower, logWeight, Type, AWD, and RWD.

The model will be tested on both elastic net and robust regression models. I will perform an outer 5-fold cross-validation on an inner 10-fold cross-validation of both models. I will use a range of alpha values to control the balance between L1 and L2 regularization and find what value works best with the model. I will also use a range of lambda values to adjust the strength of the regularization penalty. When the models finish, I will output the best set of alpha and lambda values.

The results from the double cross-validation of the two models show that elastic net was the better model. The data had highly correlated predictor variables, so this makes sense. (I ran the model using a different seed while troubleshooting an error. The result was robust regression. There are a few extreme outliers in that data and that may speak to the fragility of the testing method.) The most important predictor variables, according to the model, were `logHorsePower` and `logWeight`. `HorsePower` makes sense as vehicles with higher horsepower tend to be more expensive. Economy cars tend to have lower horsepower. Strictly by itself, `Weight` is an odd statistic when it comes to predicting the price of a vehicle. If I were to guess, I'd say this may be an artifact created by collinearity. As stated previously, large engines are heavier than small engines. Large, heavy engines produce more horsepower than smaller, lighter engines. And large, heavy engines tend to be in more expensive cars. I would need to do more digging to sus out the reasoning behind this.

The best alpha and lambda values identified by the double-cross validation process were 0.01 and 0.0111. This shows that the model is minimally penalized by the regression coefficients and L1 and L2 penalty is also very small. The RMSE for the best model is 0.2081, the R^2 for the model is 0.7997, and the MAE is 0.1636. This shows that the model can accurately predict retail prices based on the statistics of vehicles.




Retailprice as a function of Horsepower
200000
150000
100000
50000
Retailprice
100
200
300
400
500
Horsepower
Type
Minivan
Other
Pickup
Sport
SUV
Wagon


Predicted Retail Price and Weight
100000
75000
50000
25000
Predicated Retailprice
2000
3000
4000
5000
6000
Weight
type
Minivan
Other
Pickup
Sport
SUV
Wagon

In conclusion, the model is relatively accurate at predicting the retail price of new vehicles. This demonstrates that ML models are appropriate tools in this sector and could help vehicle manufacturers make sound business decisions when determining how to price future models.