Final Report

Justin Chan and Isaac Plotkin

3/12/2022

Research Question and Modeling Objective

The research question we have is how to best predict car prices using the various car features from our dataset. Therefore, our modeling objective is to create the best possible linear model using the most optimal set of features in our car dataset. Then we will be able to use this model to make predictions on car prices for new cars. We used *Predicting True Value of Used Car using Multiple Linear Regression Model* (D'Costa et al) and *Car Prediction using Machine Learning Techniques* (Gegic et al) as references for this paper.

Description of Data and Response Variable

Data

The observations (rows) in the dataset are cars with the columns being various features of the car. The dataset includes 205 rows and 26 columns. The first column is an observation index and the last column is **carprice** which is the response variable we are trying to predict. The rest of the columns are the 24 car features. There are 205 cars that we can use for our linear regression model.

The data was originally collected from various market surveys of different types of cars across the United States market around 1987 to learn how to price cars in China depending on the American market. There is an assumption that the cars in the dataset have been randomly chosen from the set of cars in the various market surveys.

The car dataset is from Kaggle: https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf

General Description of Variables

The following is the data dictionary of our dataset that gives a clear, general description of our variables/covariates that can be used in the model.

symboling: Its assigned insurance risk rating (Categorical)

carCompany: Name of car campany (Categorical)

fueltype: Car fuel type i.e gas or diesel (Categorical)

aspiration: Aspiration used in a car (Categorical)

doornumber: Number of doors in a car (Categorical)

carbody: Body of car (Categorical)

drivewheel: Type of drive wheel(Categorical)

enginelocation: Location of car engine (Categorical)

wheelbase: Wheelbase of car (Numeric)

carlength: Length of car (Numeric)

carwidth: Width of car (Numeric)

carheight: Height of car (Numeric)

curbweight: The weight of a car withoput occupants or baggage (Numeric)

enginetype: Type of engine (Categorical)

cylindernumber: Cylinder placed in car (Categorical)

enginesize: Size of car (Numeric)

fuelsystem: Fuel Sytem of car (Categorical)

boreratio: Boreration of car (Numeric)

stroke: Stroke or volume inside the engine (Numeric)

compressionratio: compression ratio of car (Numeric)

horsepower: Horsepower (Numeric)

peakrpm: car peak rpm (Numeric)

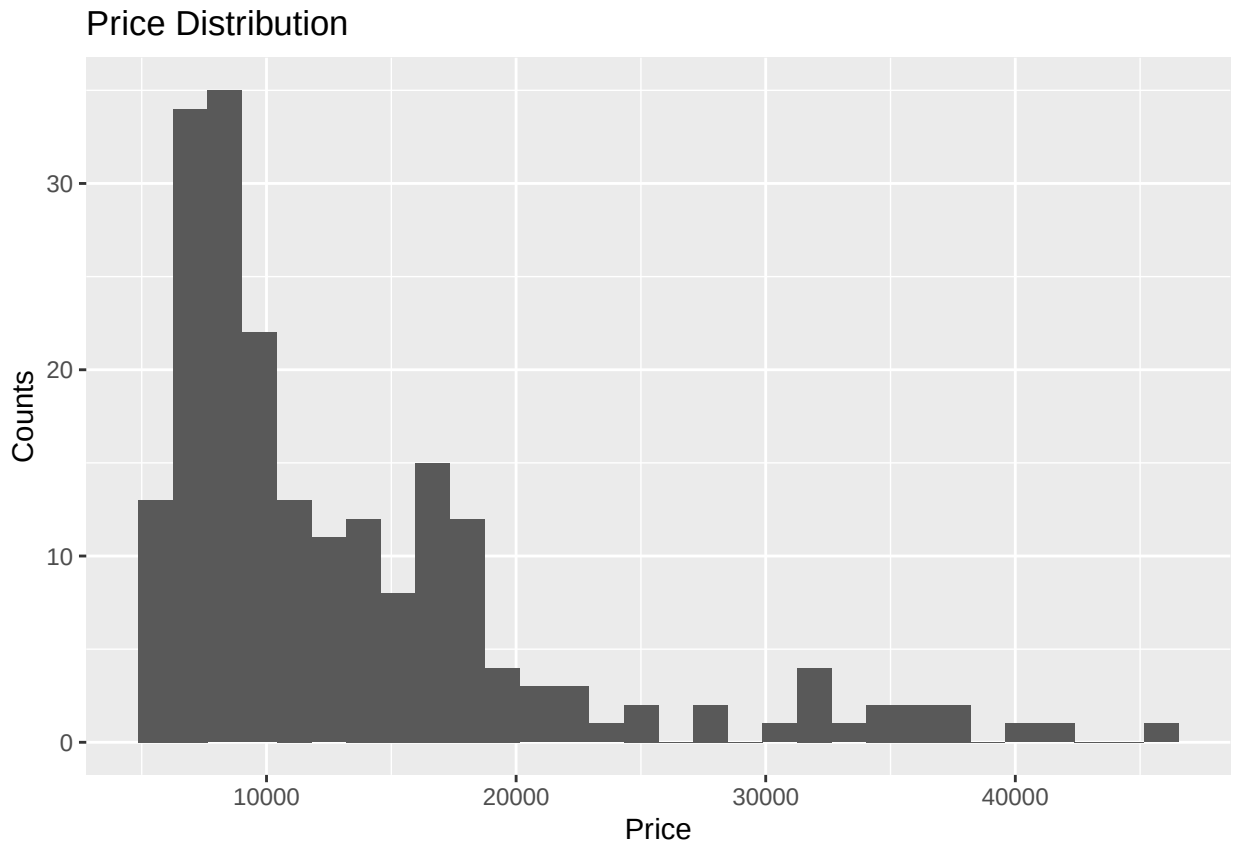
citympg: mileage in city (Numeric)

highwaympg: mileage on highway (Numeric)

price: price of car (Numeric)

Response Variable: Price

The response variable, **price**, is the price of the car in our dataset. In order to be able to predict **price**, we performed some initial univariate analysis of **price** to observe its spread in the dataset.



Price seems unimodal meaning that there is one peak. It also seems to be skewed to the right where there are many datapoints that have **price** around 5,000-10,000 but there are a few outliers that have **price** greater than $25,000. To follow up with our analysis, we also created summary statistics for **price** to see if the statistics reflected the graph we observed.

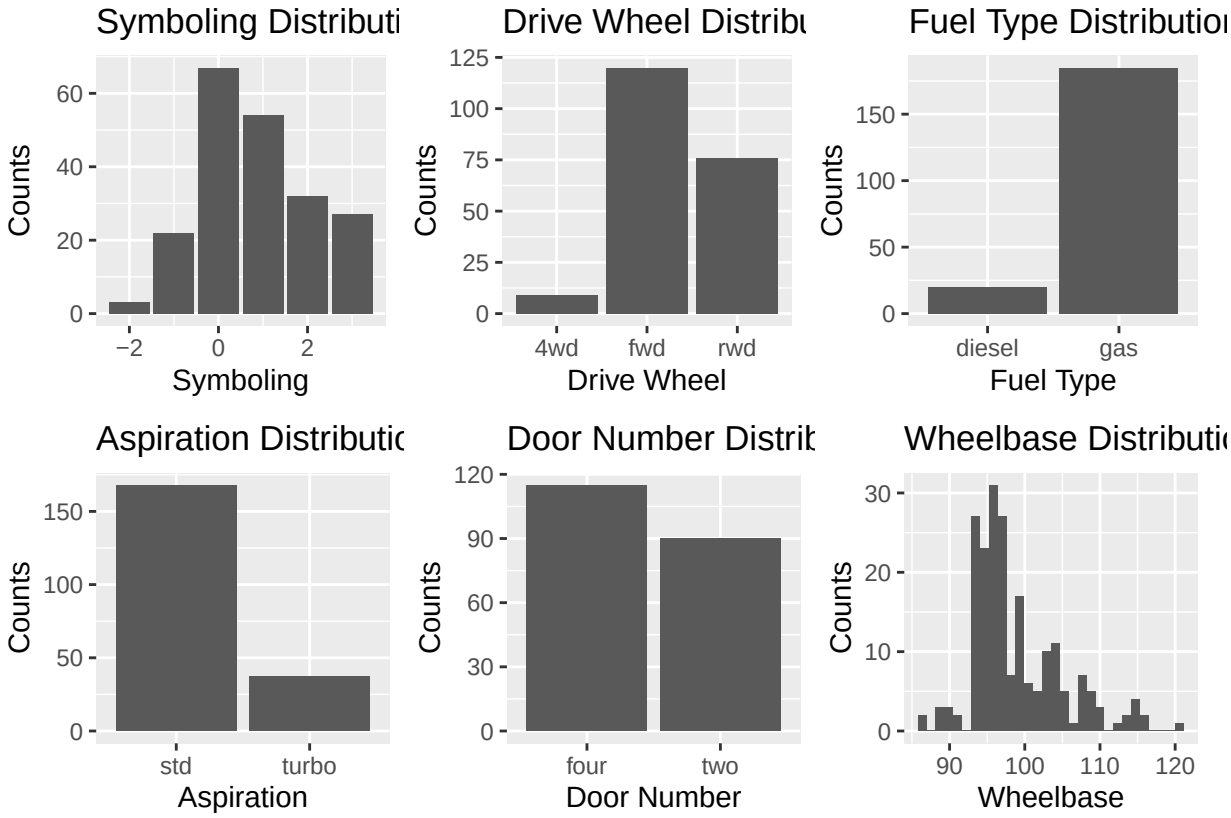
```
## # A tibble: 1 x 8
##     min    q1 median    q3   max   iqr   mean std_dev
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>
## 1  5118  7788  10295 16503 45400  8715 13277.   7989.
```

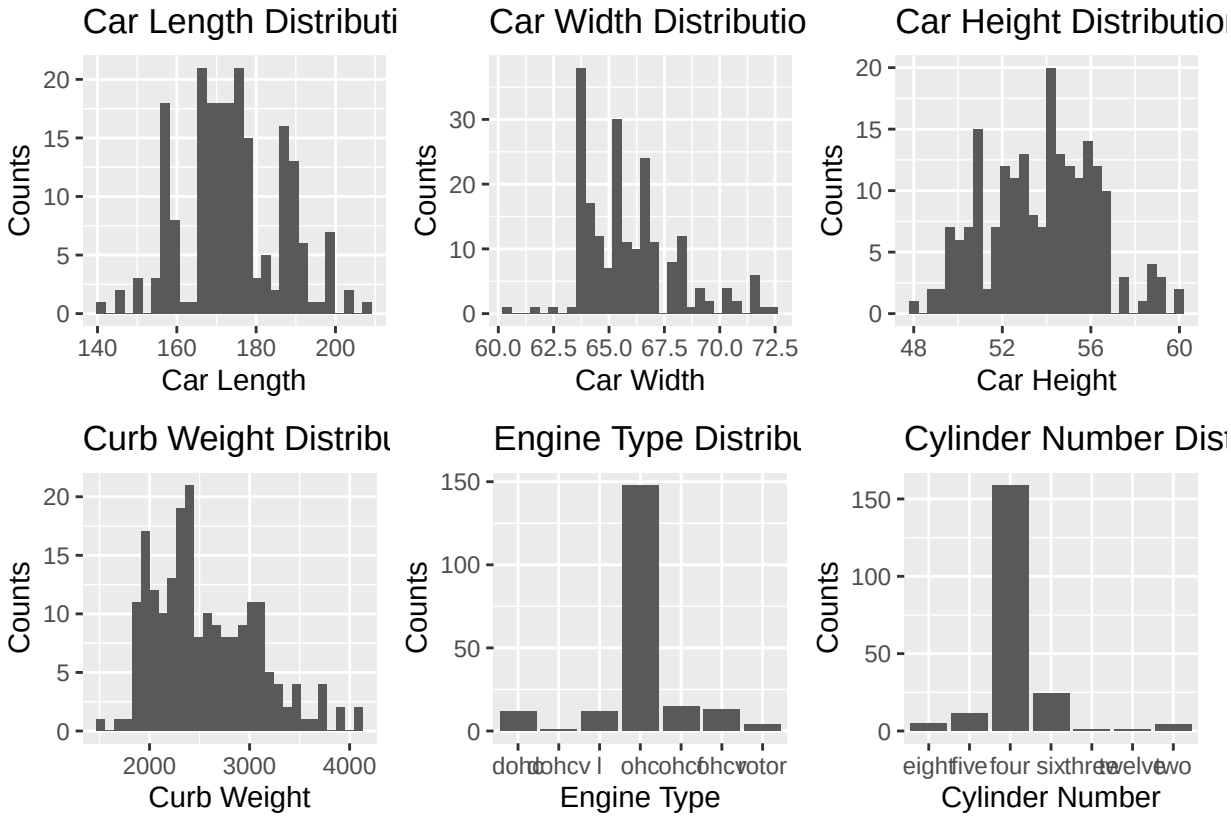
Our summary statistics further support the plot of **price** distribution. Since the mean ($13,276) is much larger than the median ($10,295) and $q1 - min$ ($2,670) is much less than $max - q3$ ($28,897), this proves that the distribution is right skewed and that there are outliers with high **price**.

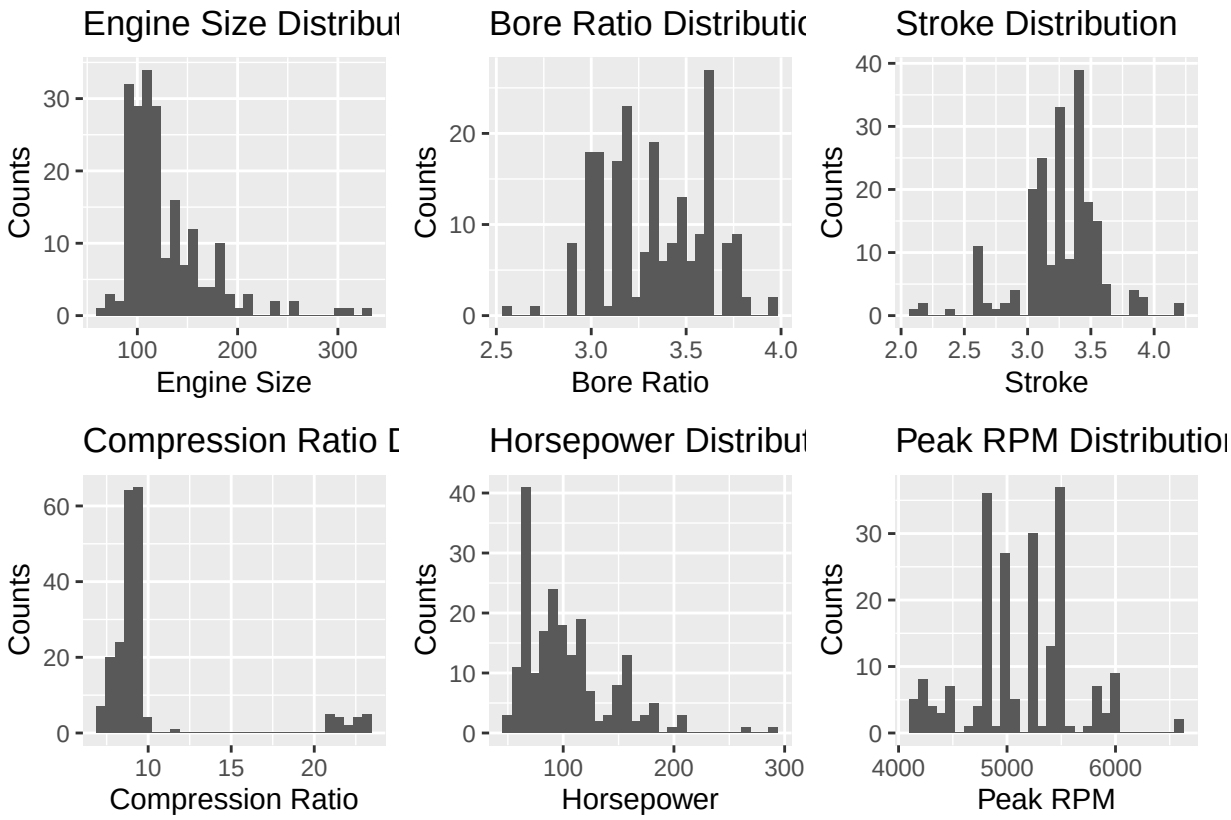
EDA

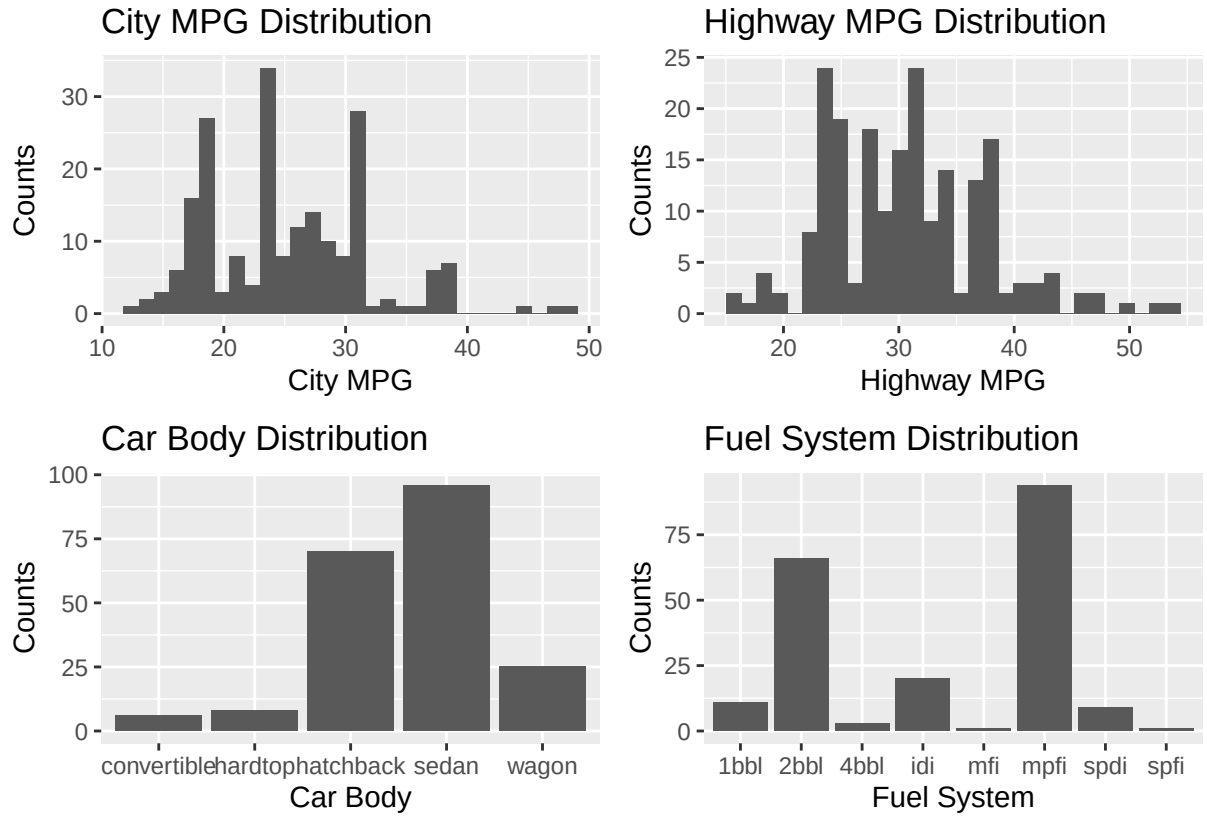
Univariate

In the following code block, we plotted the 23 covariates/possible predictor variables to do a simple univariate analysis. We used bar graphs for categorical variables and histograms for continuous variables. We formatted the graphs to be able to optimize for space on the pdf and still be able to see the visualization analysis for each variable.


Symboling Distributi
Counts
60
40
20
0
−2
0
2
Symboling
Drive Wheel Distribu
125
100
75
50
25
0
4wd
fwd
rwd
Drive Wheel
Fuel Type Distributio
150
100
50
0
diesel
gas
Fuel Type
Aspiration Distributio
150
100
50
0
std
turbo
Aspiration
Door Number Distrib
120
90
60
30
0
four
two
Door Number
Wheelbase Distributi
30
20
10
0
90
100
110
120
Wheelbase


Car Length Distributi
Counts
Car Length
Car Width Distributio
Counts
Car Width
Car Height Distributio
Counts
Car Height
Curb Weight Distribu
Counts
Curb Weight
Engine Type Distribu
Counts
Engine Type
Cylinder Number Dist
Counts
Cylinder Number


Engine Size Distribut
Counts
Engine Size
Bore Ratio Distributio
Counts
Bore Ratio
Stroke Distribution
Counts
Stroke
Compression Ratio D
Counts
Compression Ratio
Horsepower Distribut
Counts
Horsepower
Peak RPM Distributio
Counts
Peak RPM



From the first six graphs, most of the bar graphs have uneven distribution of observations across all categories. For the histograms, they all seemed to be unimodal, but some seemed a bit skewed to the right with outliers. Moving onto the next six graphs, the two bar graphs seems very skewed where one category in **enginetype** and **cylindernumber** have most of the observations. For the histograms, most of the them seem either unimodal or bimodal and are mostly skewed to the right. The next six graphs are all histograms where most of them have smaller peaks. A third of the histograms seem to have peak in the middle. Another third seem to have a peak on the left and are skewed right. The last third have a peak a bit to the right and are a little skewed to the left. For the next four graphs, the histograms seem to be trimodal where the middle peak is generally the highest and the bar graphs have two categories that have most of the observations. The next last graph has too many categories that have a count of 1 with a few of the categories having a count of 6.

After looking at all the graphs, we wanted to see the summary statistics of the univariate variables so we ran the summary method to see the individual statistics of each of our possible covariates.

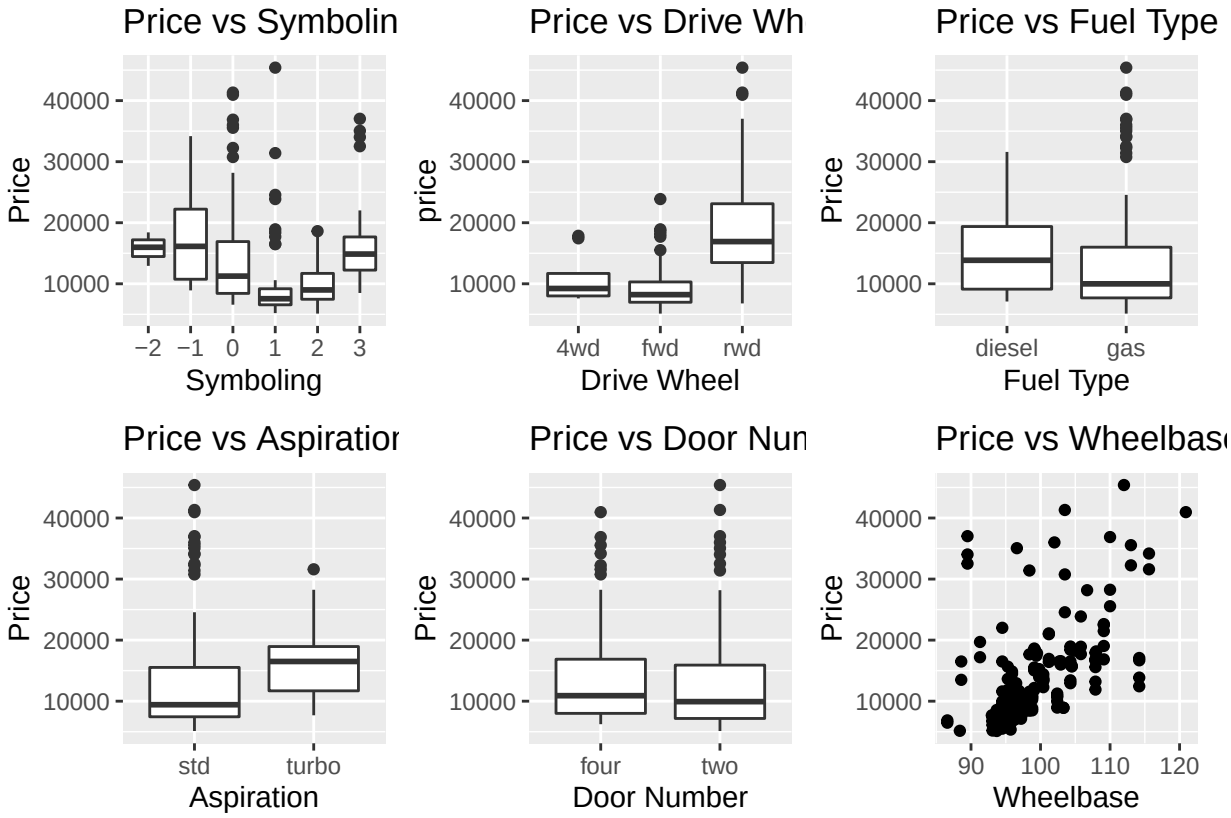
```
##      car_ID      symboling         CarName            fueltype
##  Min.   :  1   Min.   :-2.0000   Length:205         Length:205
##  1st Qu.: 52   1st Qu.: 0.0000   Class :character   Class :character
##  Median :103   Median : 1.0000   Mode  :character   Mode  :character
##  Mean   :103   Mean   : 0.8341
##  3rd Qu.:154   3rd Qu.: 2.0000
##  Max.   :205   Max.   : 3.0000
##   aspiration         doornumber          carbody           drivewheel
##  Length:205         Length:205         Length:205         Length:205
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

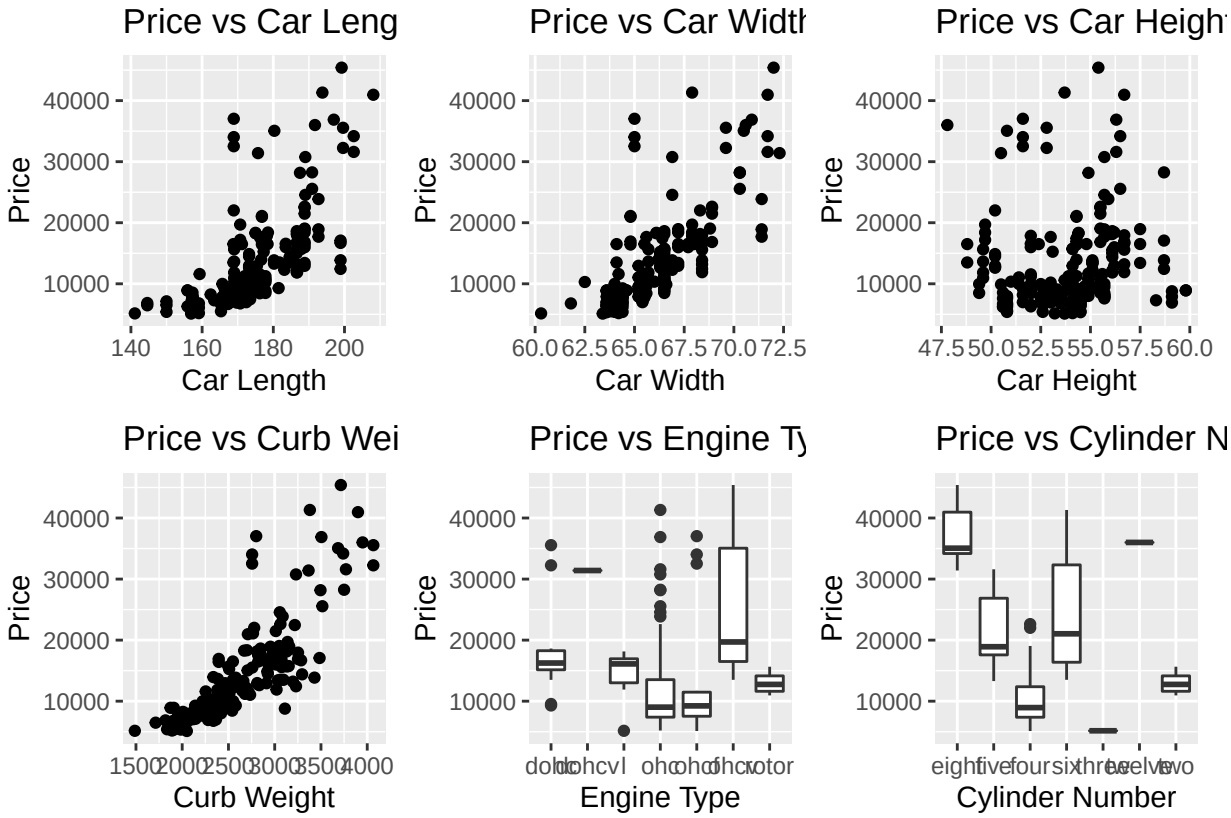
```
##
##  enginelocation       wheelbase        carlength        carwidth
##  Length:205         Min.   : 86.60   Min.   :141.1   Min.   :60.30
##  Class :character   1st Qu.: 94.50   1st Qu.:166.3   1st Qu.:64.10
##  Mode  :character   Median : 97.00   Median :173.2   Median :65.50
##                     Mean   : 98.76   Mean   :174.0   Mean   :65.91
##                     3rd Qu.:102.40   3rd Qu.:183.1   3rd Qu.:66.90
##                     Max.   :120.90   Max.   :208.1   Max.   :72.30
##    carheight       curbweight    enginetype        cylindernumber
##  Min.   :47.80   Min.   :1488   Length:205         Length:205
##  1st Qu.:52.00   1st Qu.:2145   Class :character   Class :character
##  Median :54.10   Median :2414   Mode  :character   Mode  :character
##  Mean   :53.72   Mean   :2556
##  3rd Qu.:55.50   3rd Qu.:2935
##  Max.   :59.80   Max.   :4066
##    enginesize     fuelsystem          boreratio        stroke
##  Min.   : 61.0   Length:205         Min.   :2.54   Min.   :2.070
##  1st Qu.: 97.0   Class :character   1st Qu.:3.15   1st Qu.:3.110
##  Median :120.0   Mode  :character   Median :3.31   Median :3.290
##  Mean   :126.9                      Mean   :3.33   Mean   :3.255
##  3rd Qu.:141.0                      3rd Qu.:3.58   3rd Qu.:3.410
##  Max.   :326.0                      Max.   :3.94   Max.   :4.170
##  compressionratio   horsepower       peakrpm        citympg
##  Min.   : 7.00    Min.   : 48.0   Min.   :4150   Min.   :13.00
##  1st Qu.: 8.60    1st Qu.: 70.0   1st Qu.:4800   1st Qu.:19.00
##  Median : 9.00    Median : 95.0   Median :5200   Median :24.00
##  Mean   :10.14    Mean   :104.1   Mean   :5125   Mean   :25.22
##  3rd Qu.: 9.40    3rd Qu.:116.0   3rd Qu.:5500   3rd Qu.:30.00
##  Max.   :23.00    Max.   :288.0   Max.   :6600   Max.   :49.00
##    highwaympg        price
##  Min.   :16.00   Min.   : 5118
##  1st Qu.:25.00   1st Qu.: 7788
##  Median :30.00   Median :10295
##  Mean   :30.75   Mean   :13277
##  3rd Qu.:34.00   3rd Qu.:16503
##  Max.   :54.00   Max.   :45400
```

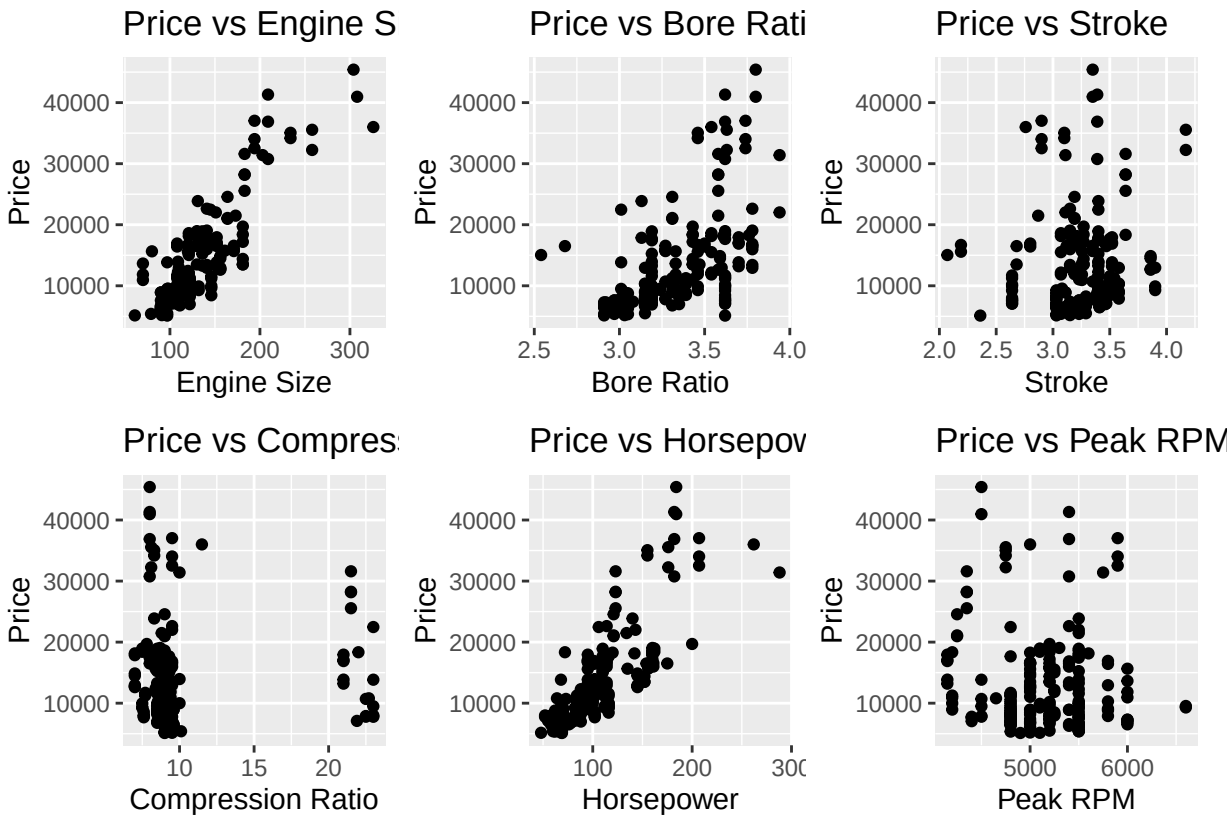
From these visualizations and statistics, we found the general distributions of each individual covariate which is always good to know before modeling. As we move onto bivariate analysis, we want to see how these distributions change when including price values to plot against them.

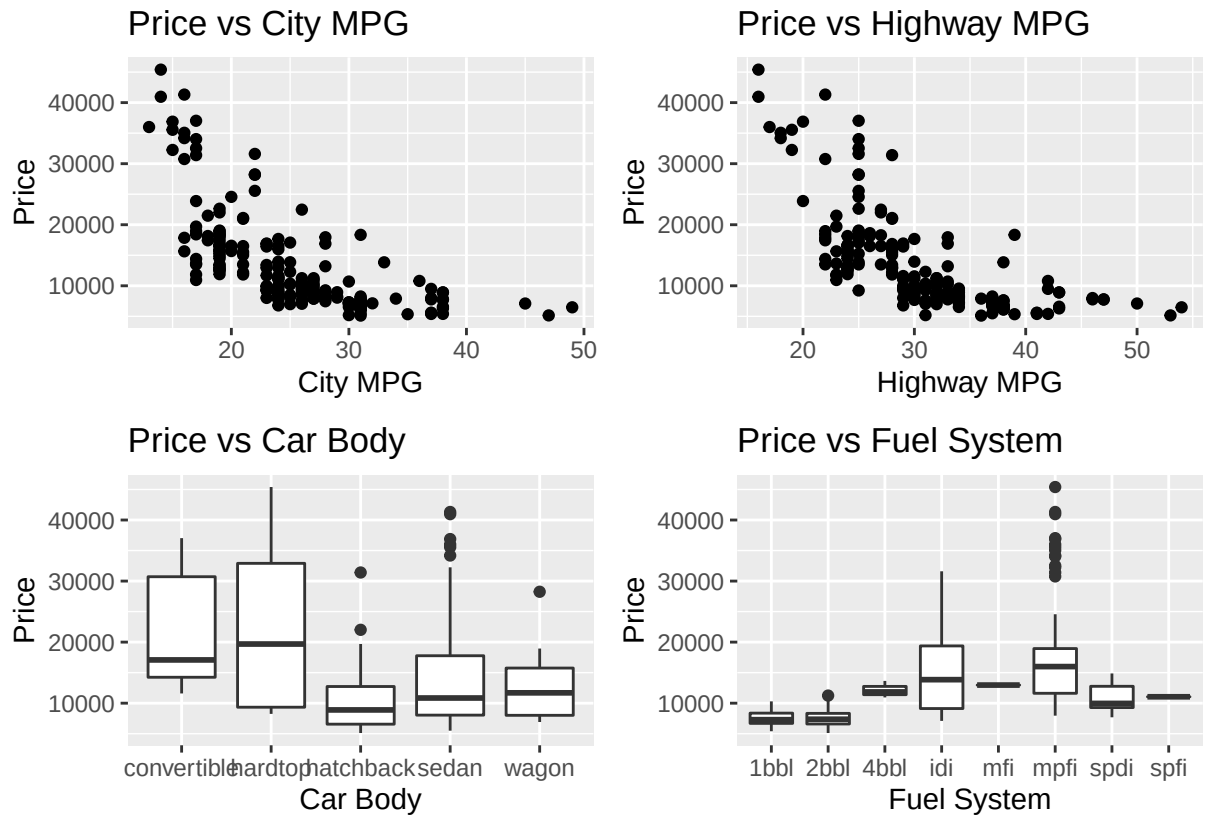
Bivariate

For bivariate analysis, we wanted to analyze each covariate vs **price** to see the relationship between each one. We want to be able to see first if a covariate could be used to distinguish price values for cars and see if there is a linear relationship between the predictor variable and our response variable. The following graphs show us **price** vs each individual covariate using box plots and scatter plots for categorical and continuous variables.


Price vs Symbolin
Price
40000
30000
20000
10000
−2
−1
0
1
2
3
Symboling
Price vs Drive Wh
price
40000
30000
20000
10000
4wd
fwd
rwd
Drive Wheel
Price vs Fuel Type
Price
40000
30000
20000
10000
diesel
gas
Fuel Type
Price vs Aspiratior
Price
40000
30000
20000
10000
std
turbo
Aspiration
Price vs Door Nur
Price
40000
30000
20000
10000
four
two
Door Number
Price vs Wheelbas
Price
40000
30000
20000
10000
90
100
110
120
Wheelbase


Price vs Car Leng
Price vs Car Width
Price vs Car Height
Price vs Curb Wei
Price vs Engine Ty
Price vs Cylinder N
Price
Car Length
Car Width
Car Height
Curb Weight
Engine Type
Cylinder Number
40000
30000
20000
10000
140
160
180
200
60.0
62.5
65.0
67.5
70.0
72.5
47.5
50.0
52.5
55.0
57.5
60.0
1500
2000
2500
3000
3500
4000
dohc
ohc
ohcv
rotor
eight
five
four
six
three
twelve
two

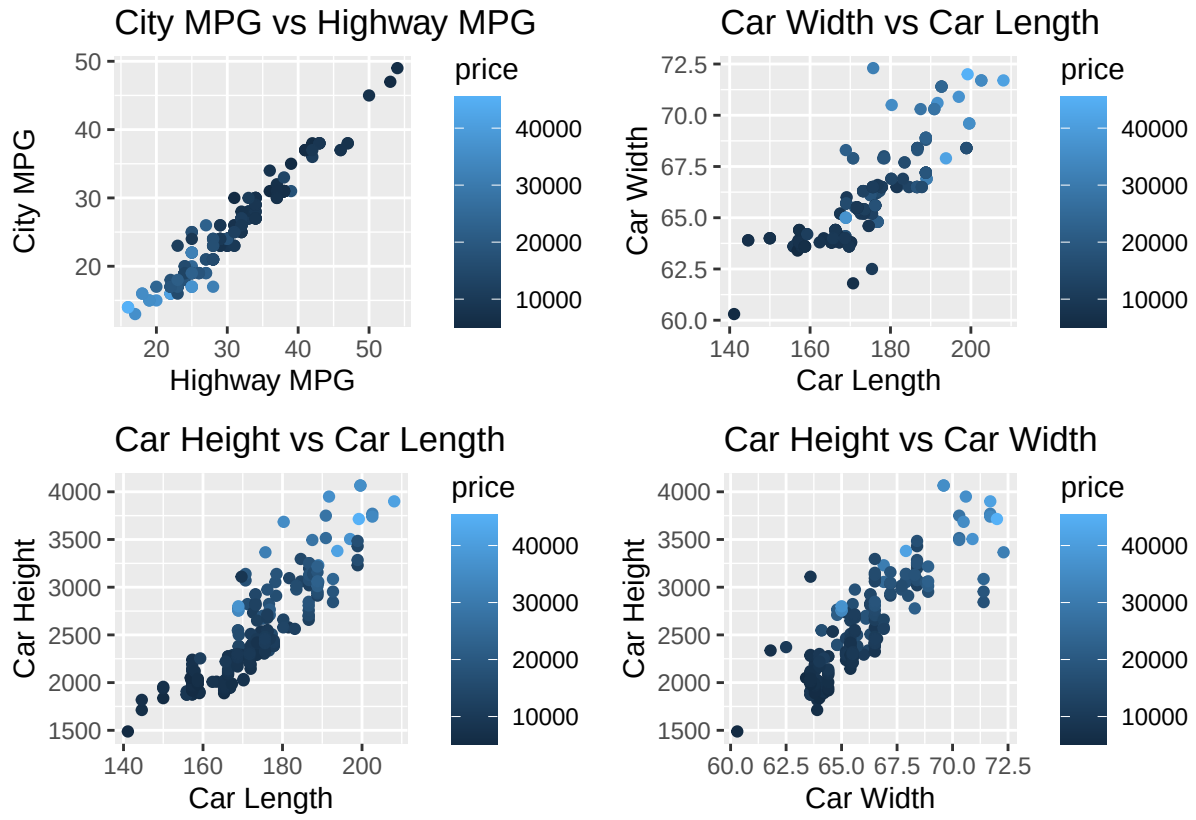

Price vs Engine S
Price vs Bore Rati
Price vs Stroke
Price vs Compres
Price vs Horsepov
Price vs Peak RPM
Price
Engine Size
Bore Ratio
Stroke
Compression Ratio
Horsepower
Peak RPM
40000
30000
20000
10000
100
200
300
2.5
3.0
3.5
4.0
2.0
10
15
20
5000
6000



When choosing initial variables to choose for our linear model, we wanted to find scatter plots that had a clear linear relationship and box plots where categories had different median values and price ranges. From our first six graphs, **symboling** and **drivewheel** had categories that had different distributions for **price**, so they are possible covariates. **Aspiration** and **wheelbase** seemed pretty reasonable to be covariates so they were also allowed to continue onto the next stage. **Fueltype** and **doornumber** were left out due to the minimal difference between their categories in the box plots. Out of the next six graphs analyzed, **carlength**, **carwidth**, and **curbweight** had the most clear linear relationship from the scatter plots. **Carheight** seemed to have a weak linear relationship and was eliminated. From the box plots, it seemed like the **cylindernumber** had each of its field to have their unique distributions for **price** which allowed it to continue as a possible covariate. **Enginetype** seemed to have much more outliers that caused more overlapping of price ranges across categories so we decided to leave it out of further analysis. For the next six graphs, the scatter plots that looked like they had a linear relationship were **enginesize**, **boreratio**, and **horsepower**. The other three predictor variables, **compressionratio**, **stroke**, and **peakrpm** did not seem to have any linear relationship with **price** so they were excluded as covariates. Of the next four graphs presented, **citympg** and **highwaympg** have a decreasing linear relationship between **price** so they are allowed to be included for further analysis. **Fuelsystem** also seemed that it could be possible to be used as a covariate based on a bit of the spread across the categories. On the other hand, **carbody** categories have quite a bit of overlapping across categories which caused us to eliminate it from further analysis. The last graph, **carName**, has both make and model of each car so there is only one car in each bin. Therefore, it seems to be useful, but there is not enough cars per category to accurately estimate price. Essentially, there were too many categories to be able to predict price so in order to avoid overfitting, **carname** was excluded from further analysis.

Multivariate

Mutlivariate analysis will be used to determine if there was interactions between particular covariates. Specifically the ones we believed that would be used in the model and seemed to be associated with each other. The multivariate analysis we did was against **highwaympg** and **citympg** since there were miles per gallon variables and would probably be associated with each other. In addition to these two covariates, it was believed that **carlength**, **carwidth**, and **curbweight** would also have interactions since the bigger the car, generally the bigger the width, length and weight of the car. Below are scatter plots of the covariates plotted against each other. **Price** is colored with lighter blue representing high car price and dark blue being low car prices.



Due to the clear relationships between **citympg**/**highwaympg**, **carheight**/**carlength**, and **carheight**/**carwidth**, these interactions may be something to consider when creating the linear model. **Carwidth** and **carlength** seemed to have the weakest relationship but, should still be considered when making the model.

Modeling Approach

The modeling approach we decided on was to select the final model covariates from the fourteen covariates that seemed to have a linear relationship with **price** in our bivariate analysis. A multiple linear regression model will be used because the outcome of car price is continuous and there are several covariates that have a linear relationship with **price**. We wanted to find the AIC and BIC for each combination of covariates to determine which selection of covariates would be best to use in our linear model. This helps us to prevent our model from overfitting. We wanted to search all possible models so we used the method dredge() from the MuMIn package to be able to create all the possible models from our covariates and order them by AIC and BIC values.

AIC

We want to choose the covariates for our model based on the lowest AIC value. In the block of code below, a model with the fourteen covariates is made. Then the model is tested with every possible combination of covariates and are ordered by AIC in ascending order. We selected the top 5 best models and recorded the covariates associated with them.

```
...
## Global model call: lm(formula = price ~ wheelbase + symboling + drivewheel + aspiration +
##     cylindernumber + curbweight * carlength * carwidth + enginesize +
##     horsepower + boreratio + fuelsystem + citympg * highwaympg,
##     data = car_data, na.action = "na.fail")
## ---
## Model selection table
##             (Int) asp        brr        crl        crw        cty        crb
## 19933   5.118e+05               -2.949e+03 -8.036e+03 -3.018e+02
## 28125   5.765e+05               -3.284e+03 -8.936e+03 -2.743e+02
## 151517  5.834e+05               -3.197e+03 -8.778e+03 -8.473e+02
## 28637   7.065e+05               -3.914e+03 -1.073e+04 -4.660e+02
## 28109   5.338e+05               -3.029e+03 -8.304e+03
...
```

BIC

As we mentioned above, we also wanted to evaluate based on BIC as a indicator of a good model. We repeated the same process by using the same model with all the covariates and doing a grid search using BIC which was ordered in ascending order. Similar to the AIC model, we took the 5 top models for BIC and recorded the variables associated with each model.

```
...
## Global model call: lm(formula = price ~ wheelbase + symboling + drivewheel + aspiration +
##     cylindernumber + curbweight * carlength * carwidth + enginesize +
##     horsepower + boreratio + fuelsystem + citympg * highwaympg,
##     data = car_data, na.action = "na.fail")
## ---
## Model selection table
##             (Int) asp        brr        crl        crw        cty        crb
## 18829   4.481e+05               -2.730e+03 -6.930e+03
## 27021   5.309e+05               -3.136e+03 -8.070e+03
## 22925   4.586e+05               -2.785e+03 -7.124e+03
## 18831   4.266e+05     -1.217e+03 -2.590e+03 -6.561e+03
## 19869   5.160e+05               -3.127e+03 -7.969e+03 -4.002e+02
...
```

After observing the top 5 multivariate linear models given by AIC and BIC, we looked to see the overlaps between the top models for each to determine the best model that has both a low AIC and BIC. The AIC top 5 models included much more features than the BIC models which is why we decided to select the BIC model that most closely resembled the AIC models. That model was number 5 on the BIC model which contained variables **carlength**, **curbweight**, **citympg**, **drivewheel**, **enginesize**, **highwaympg**, **horsepower citympg**, **drivewheel**, **enginesize**, and **horsepower**. The model also includes the interaction between **carlength** and **curbweight** which was present in every AIC and BIC model.

VIF

Next we ran VIF on the model to check for multicollinearity. **Carlength** had a VIF of 28.32, **curbweight** had a VIF of 253.92, and the **carlength**:**curbweight** interaction term had a VIF of 396.32. These are all greater than 10 so we removed **curbweight** from the model because it had a much larger VIF than **carlength**. **Citympg** and **highwaympg** also had a VIF greater than 10 which makes sense since these appear to be highly correlated features. Thus we decided to remove **citympg** from the model because it had a larger VIF than **highwaympg**. We then ran VIF again with the new selection of covariates and all of the VIF scores were below 10. See the second VIF table below to see the updated list of covariates that will be used for the final model.

	VIF
carlength	28.323388
curbweight	253.926364
citympg	28.102442
drivewheelfwd	7.527361
drivewheelrwd	7.038932
enginesize	6.041645
highwaympg	26.129437
horsepower	5.936810
carlength:curbweight	396.320168

	VIF
highwaympg	3.844498
carlength	2.705793
drivewheelfwd	6.585417
drivewheelrwd	6.759828
enginesize	3.894153
horsepower	4.511468

Output of Final Model

We created the final model Using the covariates we selected after evaluating AIC and BIC values of the model. The table below displays the coefficients for the final model.

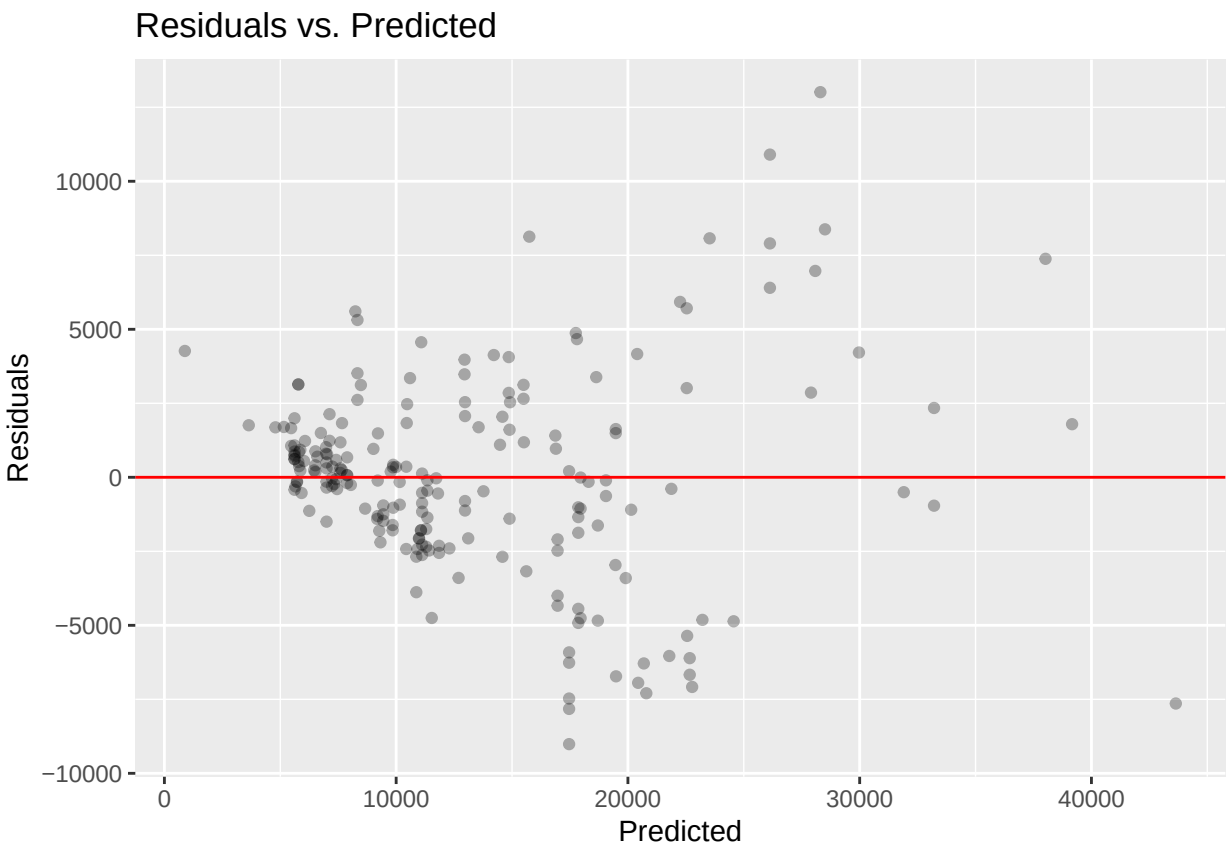
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-20177.318	6823.907	-2.957	0.003	-33634.181	-6720.454
highwaympg	35.703	68.491	0.521	0.603	-99.363	170.769
carlength	84.491	32.073	2.634	0.009	21.242	147.739
drivewheelfwd	-1262.343	1249.957	-1.010	0.314	-3727.280	1202.594
drivewheelrwd	1258.834	1291.724	0.975	0.331	-1288.468	3806.135
enginesize	98.617	11.399	8.651	0.000	76.137	121.097
horsepower	51.937	12.921	4.020	0.000	26.457	77.417

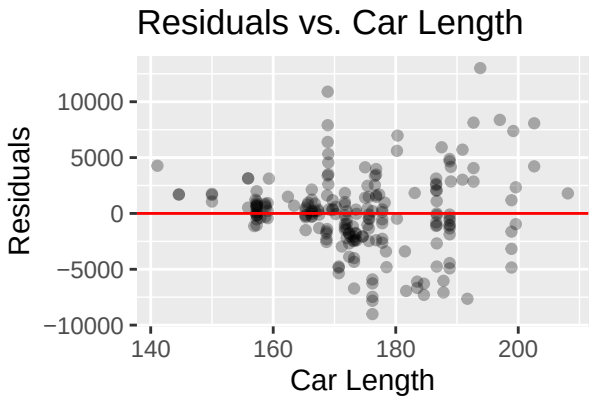
Assumptions

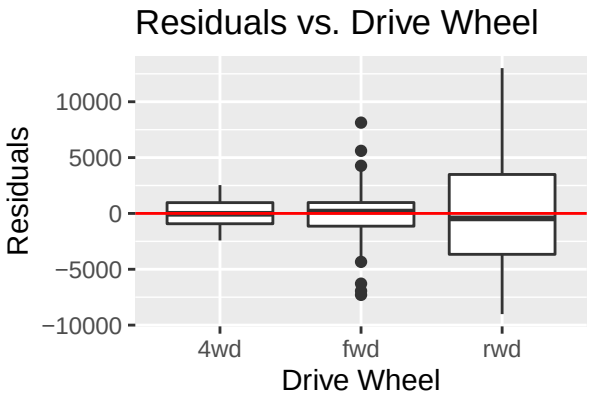
There are some assumptions that should be checked when using a multivariable linear model. These following assumptions should be checked:

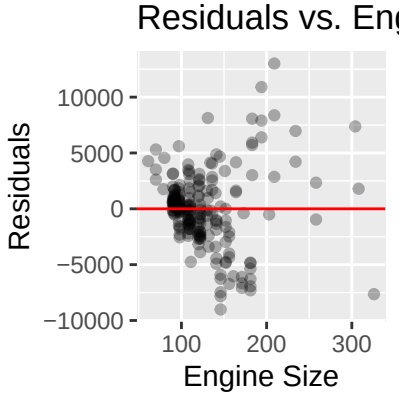
- Linearity: Response variable has a linear relationship with predictor variables. There should be no pattern in the plots unless in the case for interaction terms.
  - Residuals vs Predicted Values
  - Residuals vs Every Predictor Variables
- Constant Variance: The regression is the same for all predictor variables. The height cloud of points should be constant across the x-axis or across all predictor variable values.
  - Residuals vs Predicted Values
- Normality: Response variable follows a Normal distribution around its mean for every predictor variable. The histogram should be approximately unimodal and symmetric and the points on the QQ Plot should follow on the diagonal line.
  - Histogram of Residuals
  - Normal QQ-Plot of Residuals
- Independence: All observations are independent. There should not be pattern in residuals across the order of observations.
  - Residuals vs Observation Number

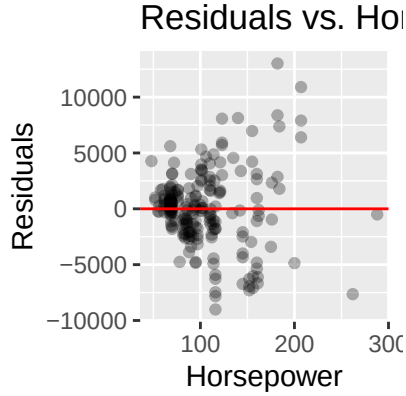
Linearity

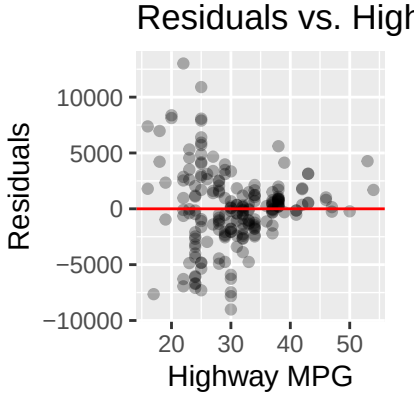

Residuals vs. Predicted
Residuals
Predicted
10000
5000
0
−5000
−10000
0
10000
20000
30000
40000


Residuals vs. Car Length
Residuals
Car Length
10000
5000
0
−5000
−10000
140
160
180
200


Residuals vs. Drive Wheel
Residuals
Drive Wheel
10000
5000
0
−5000
−10000
4wd
fwd
rwd

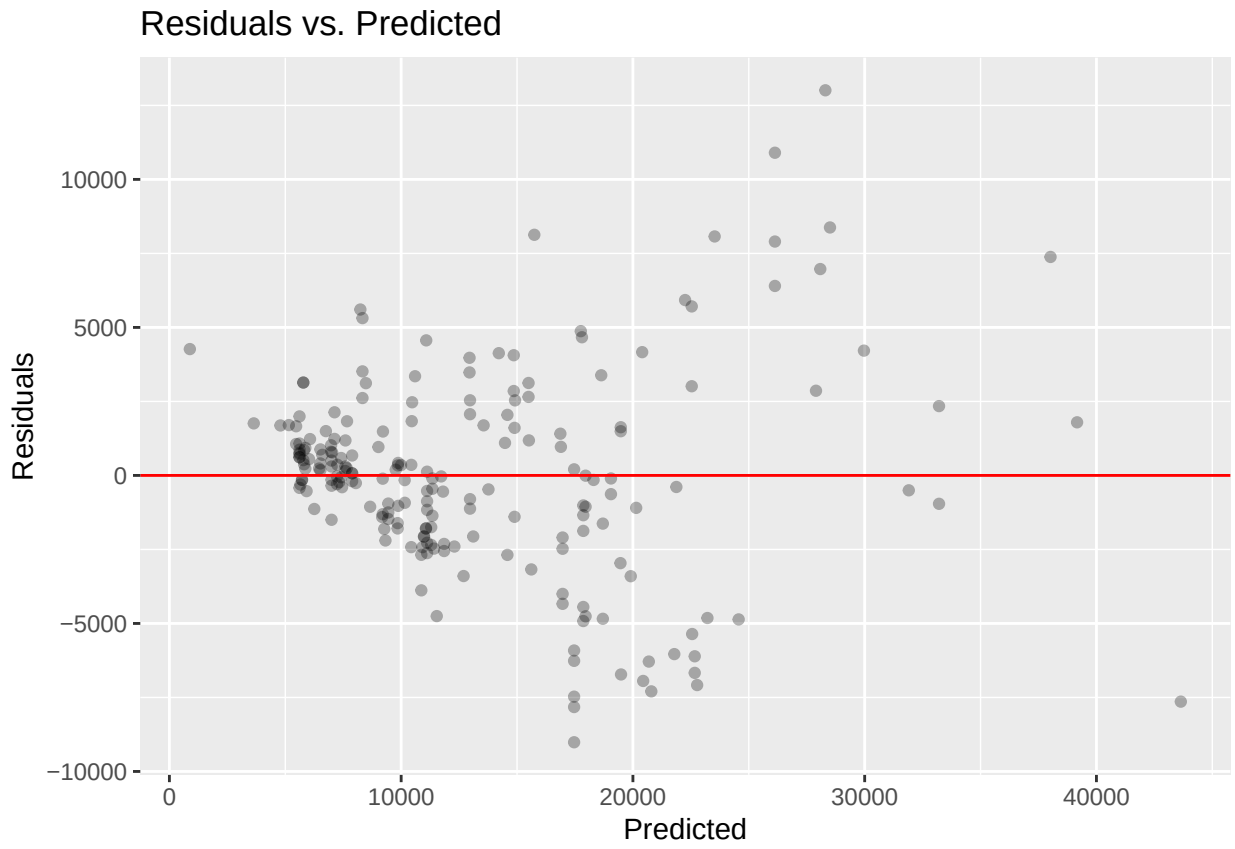

Residuals vs. Eng
Residuals
Engine Size
10000
5000
0
−5000
−10000
100
200
300


Residuals vs. Hor
Residuals
Horsepower
10000
5000
0
−5000
−10000
100
200
300


Residuals vs. High
Residuals
Highway MPG
10000
5000
0
−5000
−10000
20
30
40
50

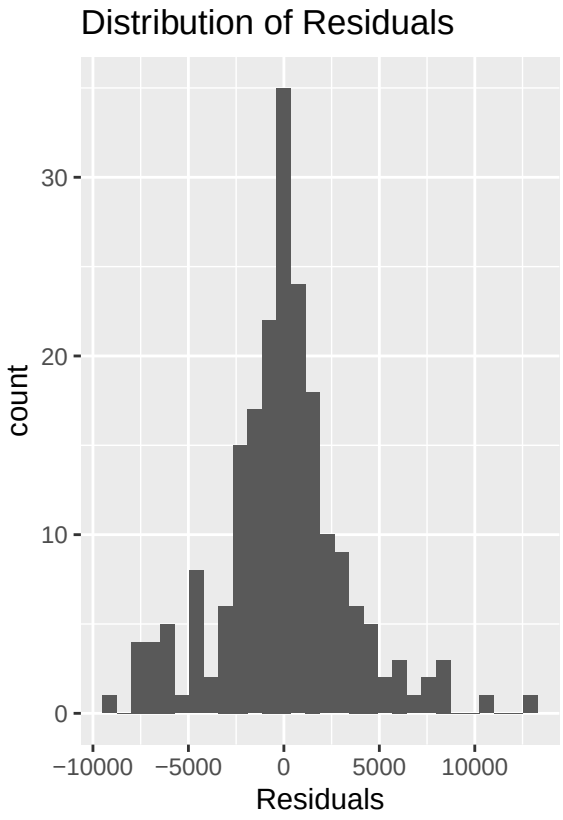
Based on the plots for checking the linearity assumption, it seems that the model does not fulfill the linearity assumption since there is a fan pattern in almost every predictor variable. Residuals tend to either increase or decrease as a predictor variable increases so it would seem the linearity assumption is not fulfilled. This also goes for the residuals versus predicted plots where there is also a fan pattern.

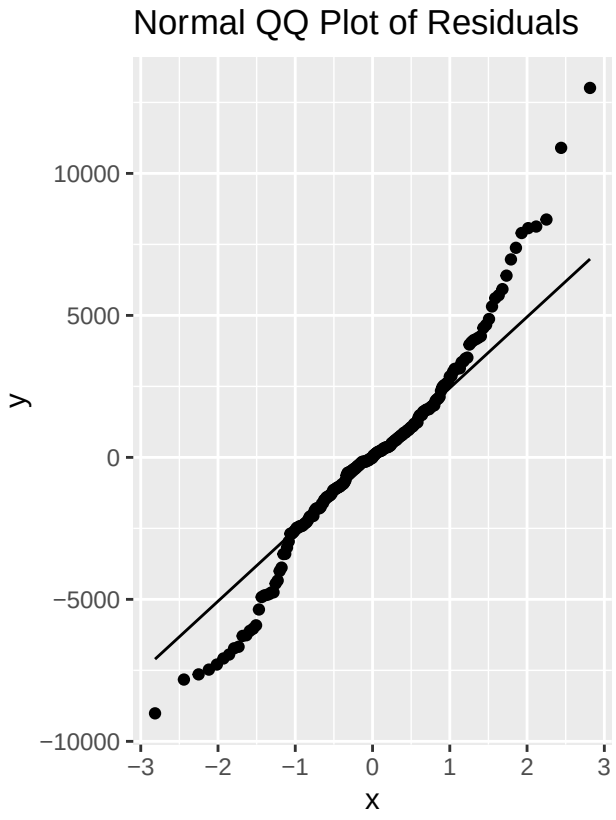
Constant Variance



The model does not meet the constant variance assumption since the graph does not follow a constant variance across all predicted variables. The variance of the residuals tend to increase as the predicted variables increase so it has a fanning pattern.

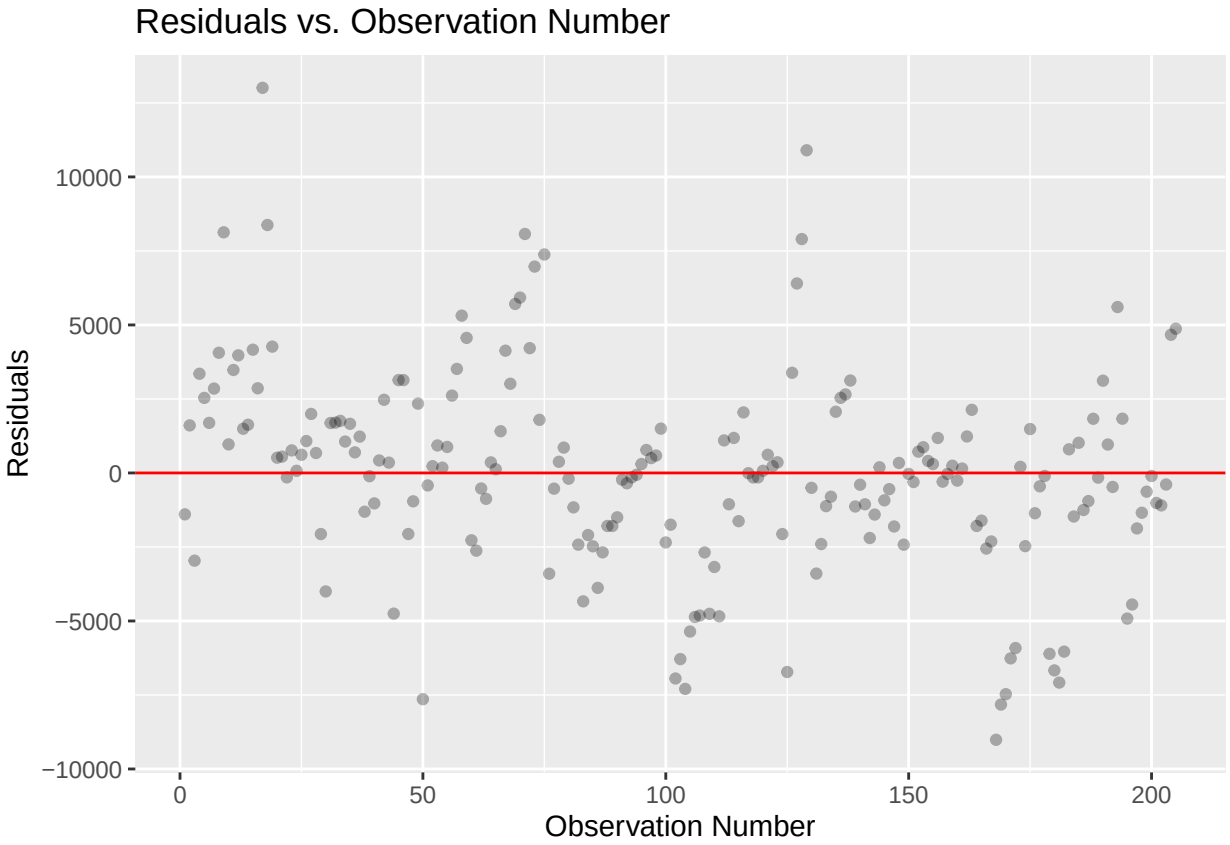
Normal Condition





The normal condition is not met since the QQ Plot is not following a completely diagonal line. The tails of the QQ plot are not on the diagonal line. Also the histogram of the residual is not normal.

Independence



The independence condition is met in our case since across the order of observations, there is constant variance. This means that the observations are most likely independent from one another.

Interpretations of Model Coefficients

The next part in our analysis is going to refer back to the coefficients in our model. Here is the table of coefficients.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-20177.318	6823.907	-2.957	0.003	-33634.181	-6720.454
highwaympg	35.703	68.491	0.521	0.603	-99.363	170.769
carlength	84.491	32.073	2.634	0.009	21.242	147.739
drivewheelfwd	-1262.343	1249.957	-1.010	0.314	-3727.280	1202.594
drivewheelrwd	1258.834	1291.724	0.975	0.331	-1288.468	3806.135
enginesize	98.617	11.399	8.651	0.000	76.137	121.097
horsepower	51.937	12.921	4.020	0.000	26.457	77.417

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$

$y = -20177.318 + 35.703 X_1 + 84.491 X_2 - 1262.343 X_3 + 1258.834 X_4 + 98.617 X_5 + 51.937 X_6 + \varepsilon$

Based on the table of the linear model above, the estimate intercept is telling us the price of a car with a **carlength** of 0, **enginesize** of 0, **highwaympg** of 0, **horsepower** of 0, and **4wd** which is not realistic at

all. The intercept estimate, -20177.318 does not tell us anything meaningful relative towards the data model. For the coefficients, if a car's **drivewheel** is **fwd** then the price prediction does down $1262.34 and if a its **drivewheel** is **rwd** then the price prediction does up $1262.34. The coefficient estimate 98.617 for engine size means that the for one unit increase in the **enginesize** there is a general increase of $98.62 in car price. The coefficient estimate 35.703 for **highwaympg** means that for every one unit increase **highwaympg**, the car price will increase by about $35.70. The coefficient estimate for **horsepower** is 51.937 which means that for every one unit increase in **horsepower**, there is an increase of about $51.94 in car price. The coefficient estimate 84.491 for **carlength** means that for one unit increase in **carlength**, there is a general increase of $84.49 in car price.

Limitations

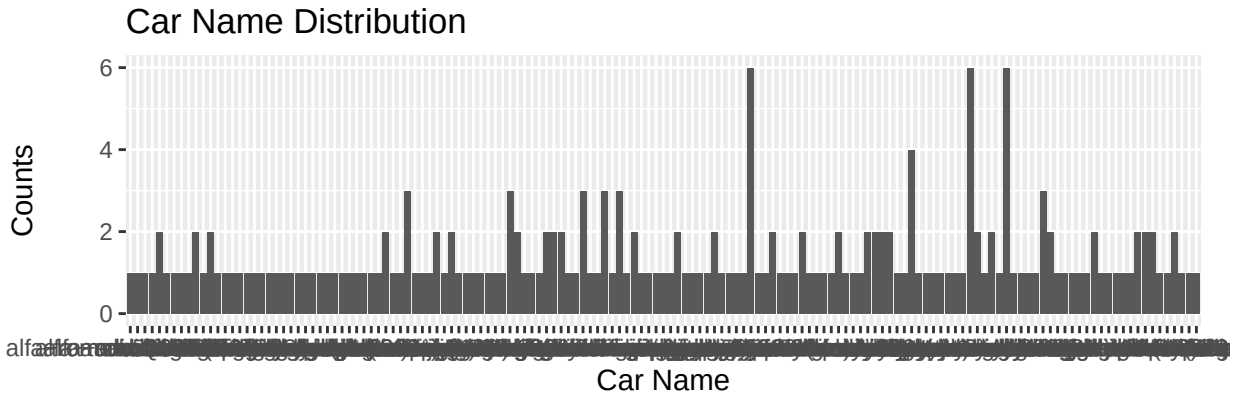
There are a few limitations of the model that should be acknowledged when trying to predict car price. To begin with, the model will not be able to accurately predict the price of a car that has covariate values outside the range of values that the model was trained on. For example, the final model was trained on the variable **carlength** which ranged from 140 to 200. Thus it would not be wise to predict the price of car with a **carlength** outside the range [140,200]. In other words, the model can interpolate but not extrapolate. Another limitation is that this model can only fit data linearly. Even though this dataset fits a linear model well, there might be non linear models that can predict **price** at a much higher accuracy.

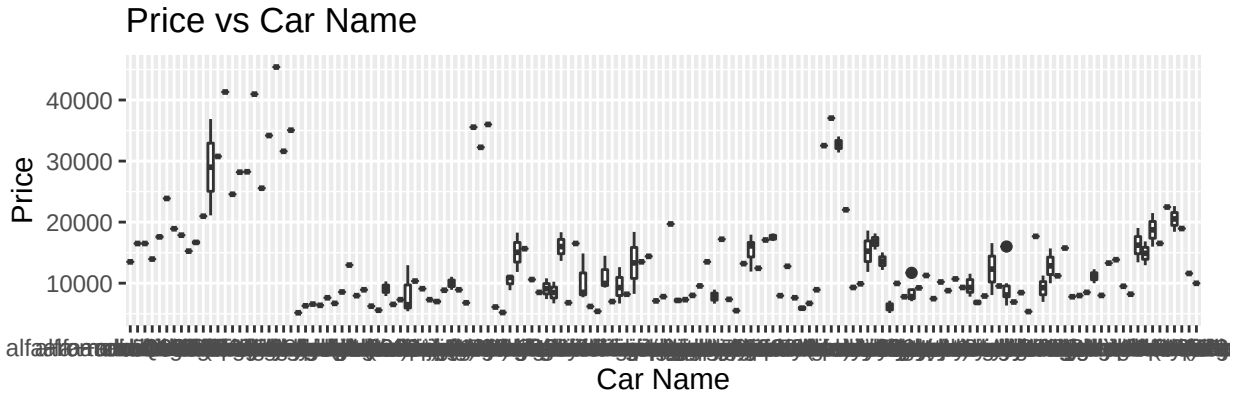
Conclusion

Multiple linear regression appears to be a good approach to accurately predicting car price. It is interesting to see how logical the final predictor variables are. In future work we might try to run a non linear machine learning model to see how its results compare to our linear model. The pro of using linear regression is that the results are easily interpretable, but if we used a black box machine learning algorithm, it would most likely predict car prices at a higher accuracy.

Additional Work

Univariate and bivariate plots for **carname**.


Car Name Distribution
Counts
6
4
2
0
Car Name


Price vs Car Name
Price
40000
30000
20000
10000
Car Name