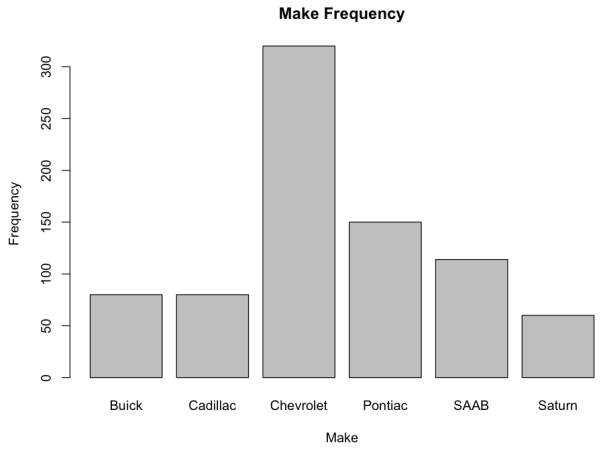
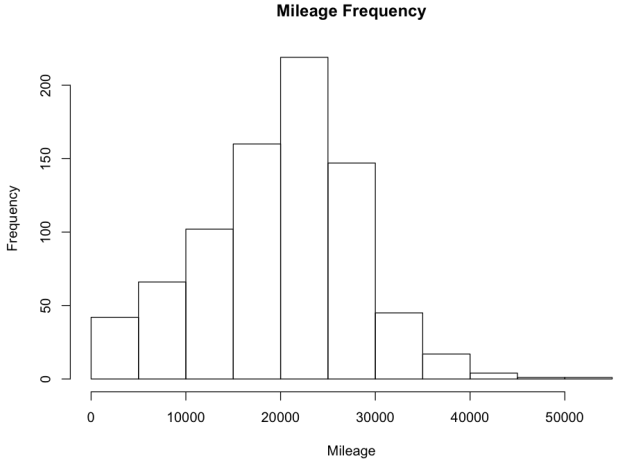
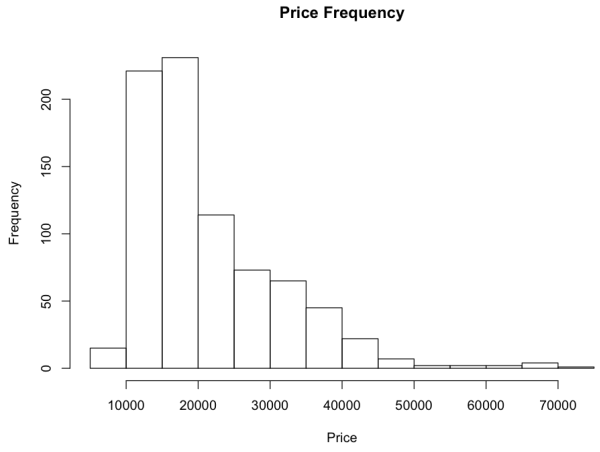
Stat 525 - IE Project Step 3 - Initial Exploration

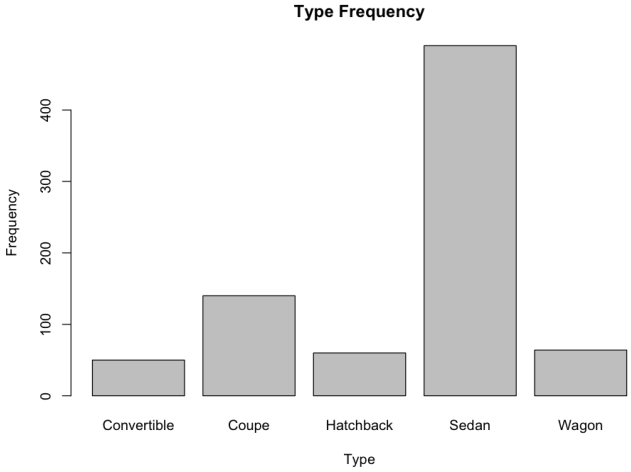
Connor, Ritwik, Evan, Jianjun

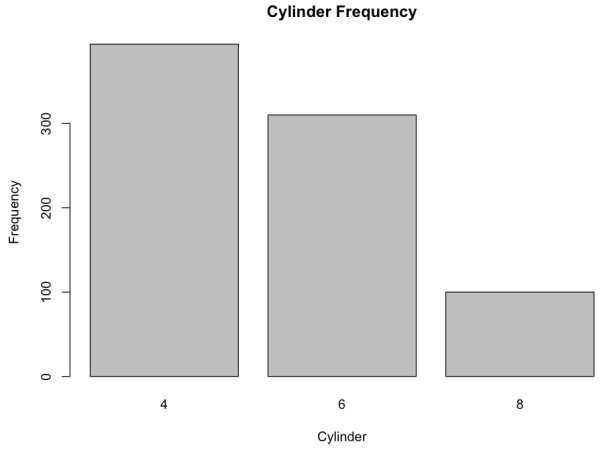
Data Frequency for X and Y

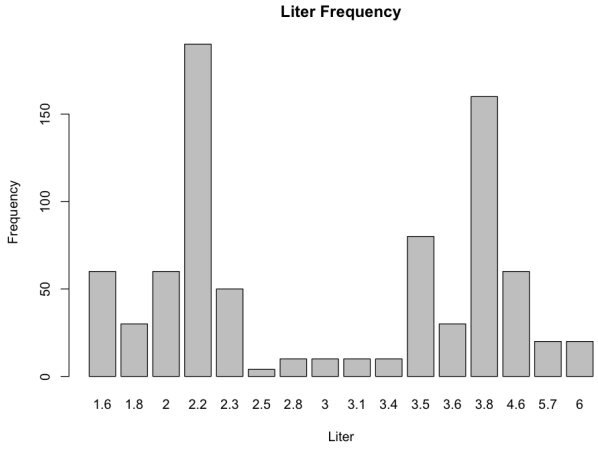


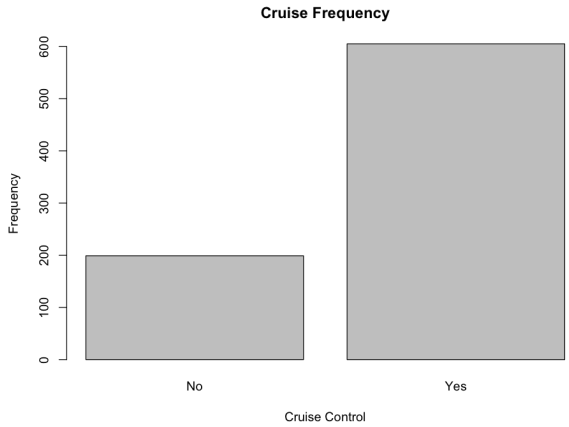


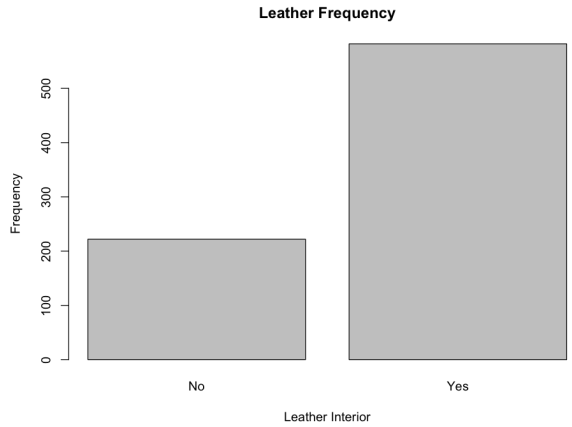


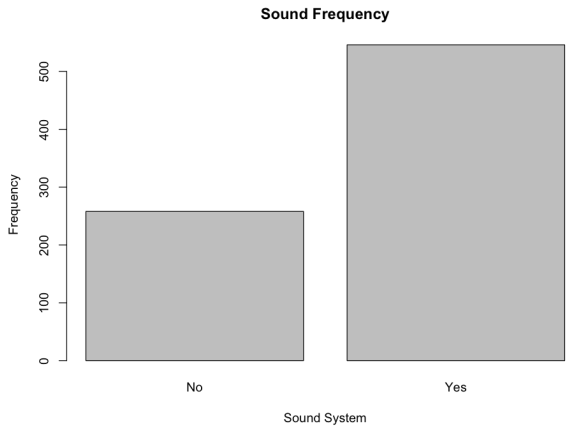


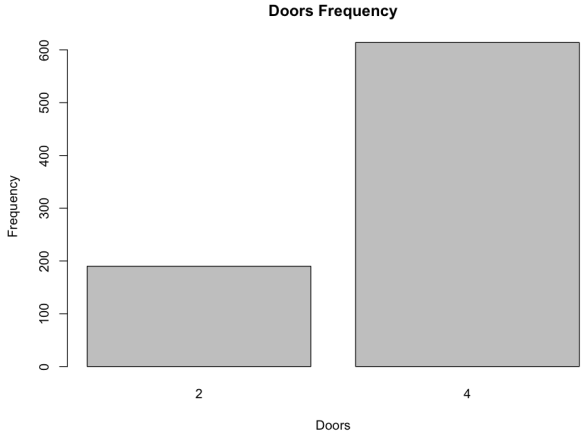


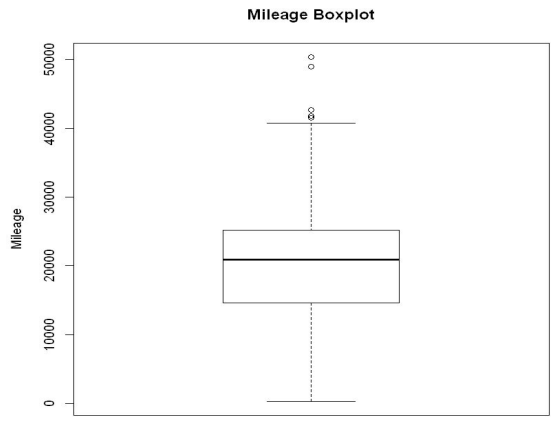


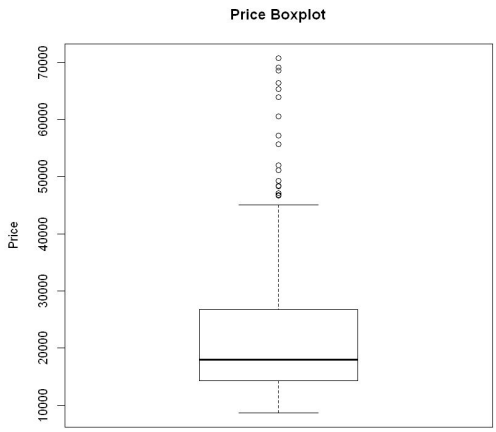






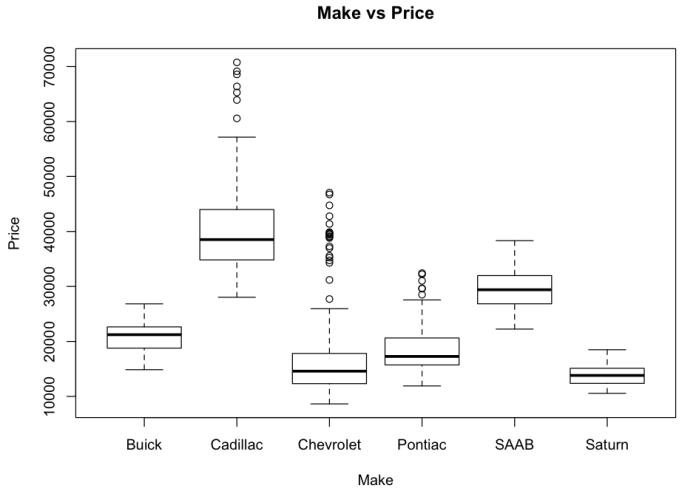


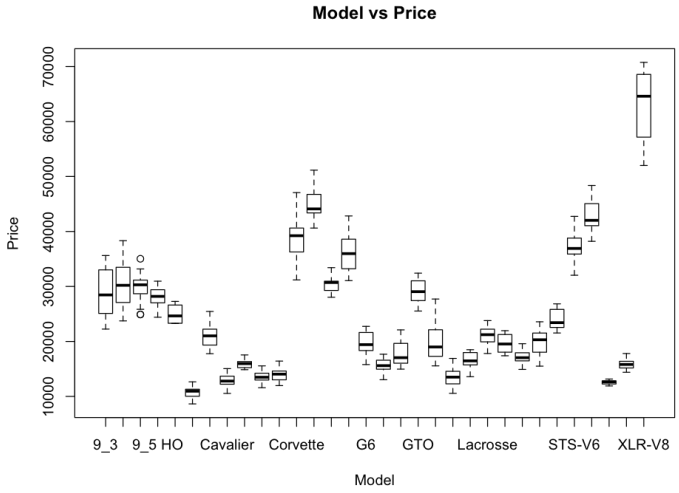


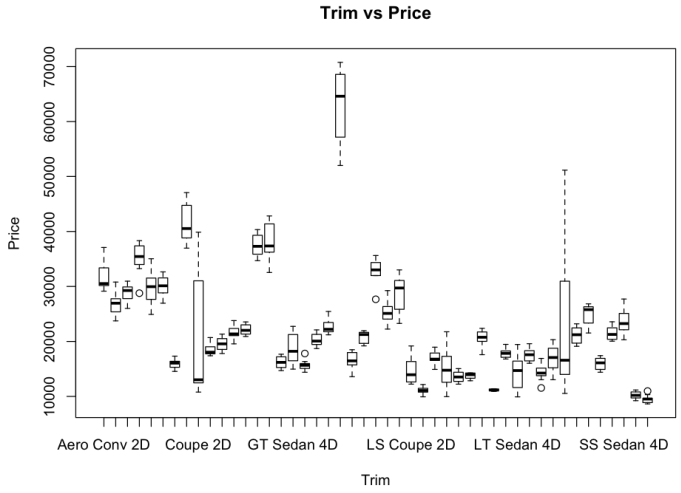


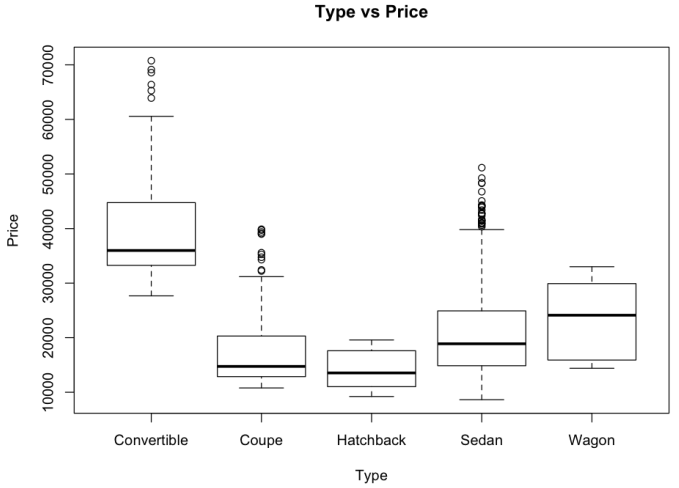
From these charts, we can see that there are some outliers in the data. For mileage, there are a few data points between 40,000 miles and 50,000 miles that are outliers. For price, there are quite a few outliers, ranging from ~$47,000 to $70,000. Upon investigating the dataset, many of the price outliers come from a specific make and model of vehicle (the Chevrolet Corvette) that is a more expensive and premium vehicle than the others in our data, so we will still include this data in our analysis. We excluded the Model and Trim charts as they are effectively unreadable due to the number of categorical options in both Model and Trim.

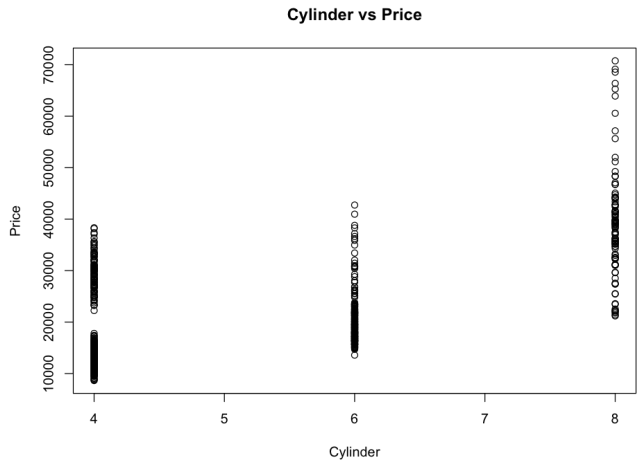
Scatter and Box Plots for Price vs. Potential Model Features

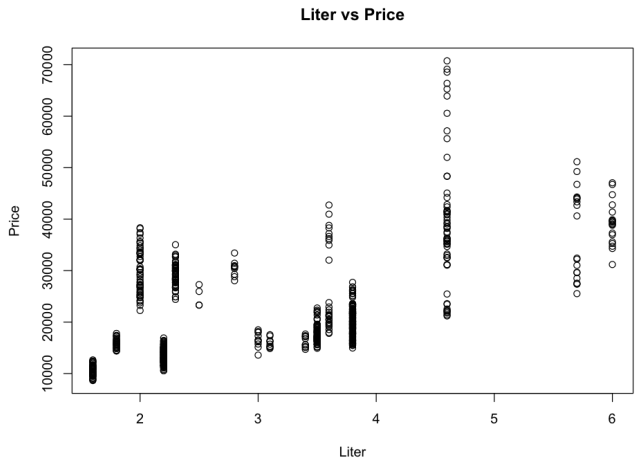


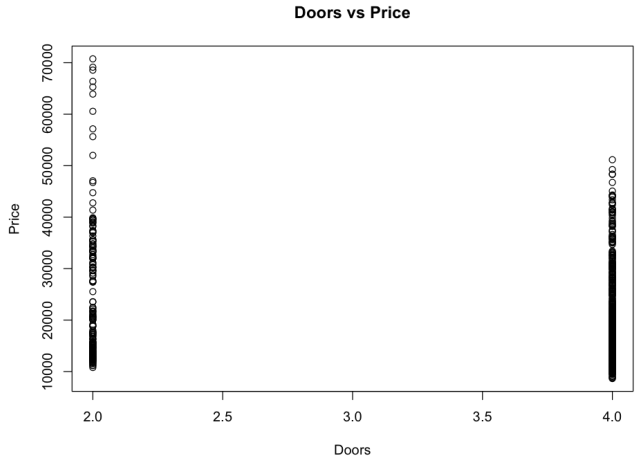


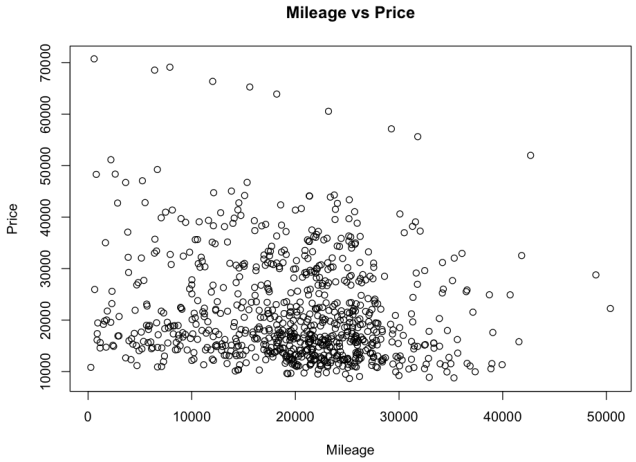


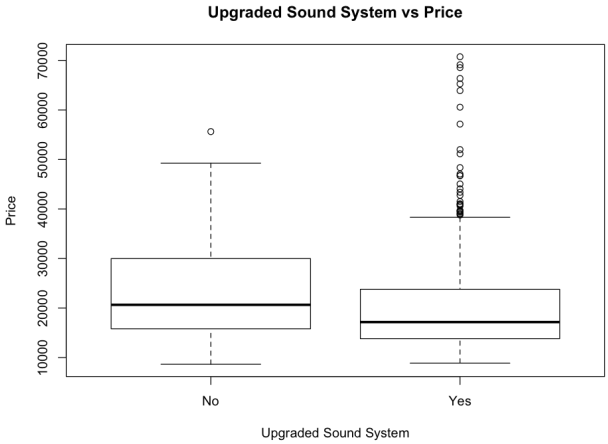


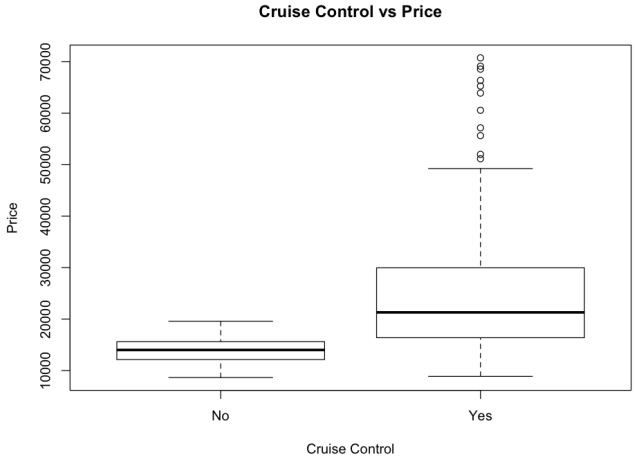


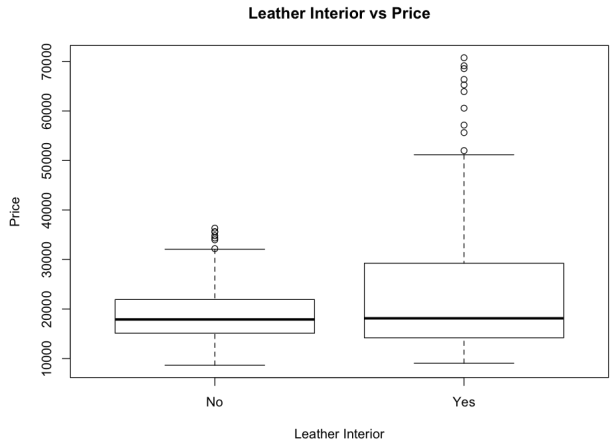










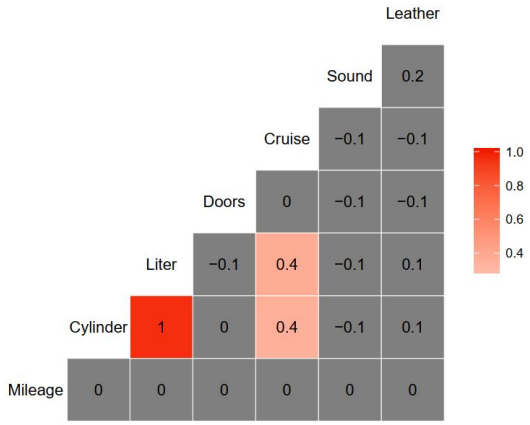


From looking at these charts, we can see some interesting patterns in our data. For example, in the “Upgraded Sound System” chart, we see that vehicles with an upgraded sound system have a lower price in the box and whiskers, but many outliers. We believe this is because the data does not indicate the quality of the sound system, but solely whether it was upgraded or not, and cheaper vehicles are more likely to have an upgraded sound system versus costlier vehicles that come with a better base sound system.

When it comes to linear relationships between quantitative X variables and Y, we can see a few variables that seem to exhibit a linear relationship. Mileage has a slight negative linear relation with Price, and Price has a slight positive relation with Cylinder and Liter.

For patterns between Y and categorical X variables, Cruise Control appears to have a positive linear relation with Price, as does having a Leather Interior. Make and Type both have a strong pattern with Y, where some Makes and some Types are clearly cheaper or more expensive than others.

Most of the X variables are not strongly correlated with each other - the one exception is Cylinder and Liter, as the two are strongly interlinked. The reason for this is that since our dataset only includes GM vehicles, only the GM family of engines is counted in our dataset, and for each liter number, there is only one engine that matches up, so that liter number will always have the same associated cylinder number. In other words, if a car has a 5.3 liter engine, it has a GM Vortec 5.3 V8, and so every 5.3 liter engine car will have an 8 cylinder engine. There are no engines with different cylinder numbers and the same displacement in our dataset.



During all of our preliminary analysis, we found that the Model and Trim variables are not useful for our model in their current state, as so many of them are specific to a particular model or make. In order to make the Trim and Model variables more useful, we plan to include them using interaction terms. If this still does not yield good performance in the model, we will transform the data by using the different parts of the trim and model (*e.g.* instead of using “Aero Sedan 4D” and “Aero Wagon 4D”, we can use “Sedan 4D” and “Wagon 4D”) so that the trim and model information can actually be effectively used in our model.

Descriptive Statistics

Price and Mileage

	Price	Mileage
Mean	21343.144	19831.93
SD	9884.853	8196.32
Median	18024.995	20913.50
IQR	12444.243	10589.50

Make

Buick	80
Cadillac	80
Chevrolet	320
Pontiac	150
SAAB	114
Saturn	60

Type

Convertible	50
Coupe	140
Hatchback	60
Sedan	490
Wagon	64

Doors

2	190
4	614

Cruise Control

No Cruise Control	199
Cruise Control	605

Leather Interior

No Leather Interior	222
Leather Interior	582

Sound System

Base Sound System	258
Upgraded Sound System	546

Cylinder

4	394
6	310
8	100

Liter

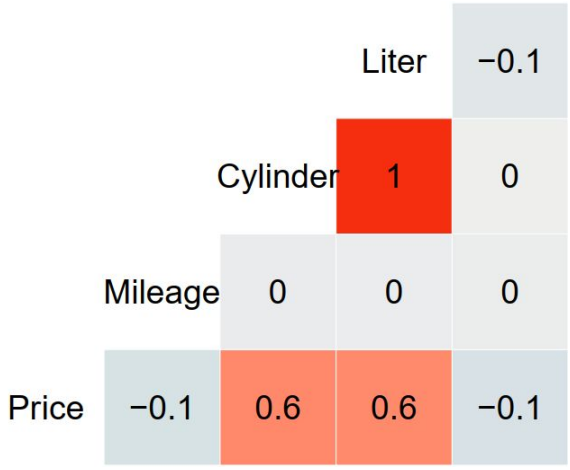
3.1	10
3.6	30
3.8	160
5.7	20
2.8	10
4.6	60
1.6	60
2.2	190
6	20
3.5	80
3.4	10
1.8	30
2	60
2.3	50
2.5	4
3	10

Model	
9-2X AWD	4
9_3	20
9_3 HO	40
9_5	30
9_5 HO	20
AVEO	60
Bonneville	30
Cavalier	60
Century	10
Classic	10

Trim	
Aero Conv 2D	10
Aero Sedan 4D	20
Aero Wagon 4D	10
Arc Conv 2D	10
Arc Sedan 4D	20
Arc Wagon 4D	10
AWD Sportwagon 4D	10
Conv 2D	10
Coupe 2D	50
Custom Sedan 4D	10

Here, a subset of the Model and Trim statistics are shown. Due to the number of categories in each, the full statistics of each are not shown.

Quantitative Variable Correlation



This is the correlation between quantitative variables in our dataset. While some other variables, such as doors, cruise control, sound system, and leather interior are also represented numerically (specifically as 0s and 1s), these variables are actually categorical (*e.g.* a car either has cruise control or doesn't) and are numerically represented as a formatting choice.

The correlation between liter and cylinder exists due the reasons explained above, but we also see a comparatively strong correlation between cylinder/liter and price.

Y Variable (Price) Patterns over Categorical X Variables

With regards to the **Make** simple linear model, the intercept of $20815.10 refers to the average cost of a Buick. If the make is a Cadillac or SAAB will increase this price, whereas if the make is a Chevrolet, Pontiac, or Saturn the expected price will fall.

The **Model** simple linear model intercept of $24960.90 refers to the expected cost of a "9-2X AWD". Specific makes such as 9-3, 9-3 HO, 9-5, 9-5 HO, Corvette, CST-V, CTS, Deville, GTO, STS-V6, STS-V8, and XLR-V8 will have higher expected costs. All other Models have a lower expected cost.

The **Trim** simple linear model has an intercept of $31764 which refers to the expected price of a "Aero Conv 2D" trim. Other trims such as Arc Conv 2D, Conv 2D, DHS Sedan 4D, DTS Sedan 4D, Hardtop Conv 2D, or Linear Conv 2D have higher expected costs, where every other trim has a lower expected price.

The **Type** simple linear model has an intercept of $40832 which refers to the expected price of Convertible type of car. Every other type has a lower expected price.

The **Cylinder** simple linear model has an intercept of -$17.06 and each additional cylinder adds an expected price 4054.20 dollars per cylinder.

The **Liter** simple linear model has an intercept of $6185.80 and each additional liter will add an expected 4990.40 dollars to the price.

The **Doors** simple linear model has an intercept of $27033.60 with each additional door decreasing the expected price by 1613.20 dollars.

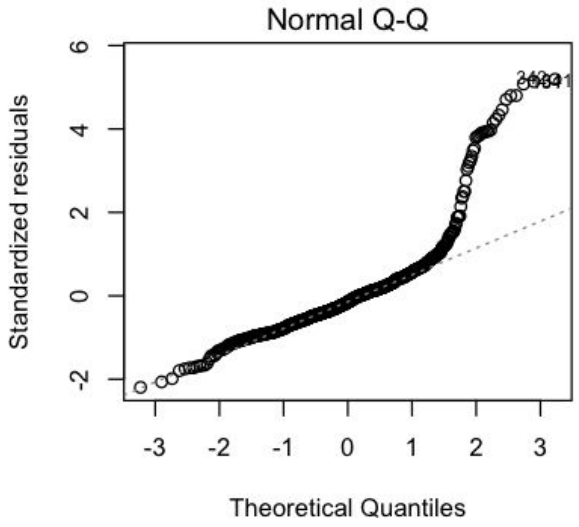
The **Cruise** simple linear model has an intercept of $13921.90 for the non-cruise control cars. Adding cruise control will increase the expected price by 9862.30 USD.

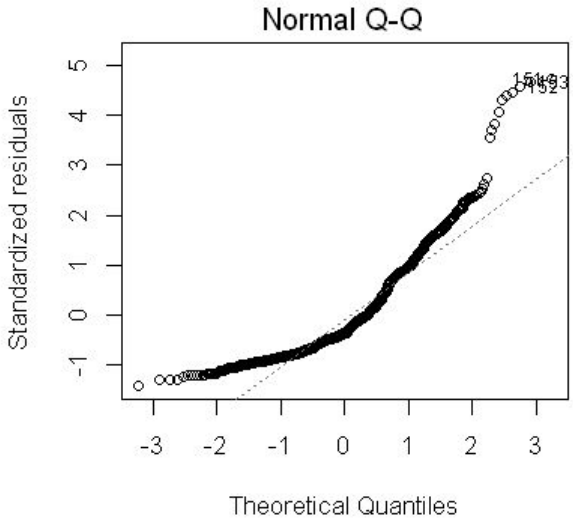
The **Sound** simple linear model has an intercept of $23130.10 for the non upgraded sound system cars. Upgrading the sound system will decrease the expected price of the car by 2631.40 USD.

The **Leather** simple linear model has an intercept of $18828.80 which is the expected price of a car without leather seats. Cars with leather seats have an increased value of 3473.50 US

Regression Analysis for Y vs. X variables

When looking at the regression models between our Y variable (Price) and every X variable we have, the only assumption violated in each case is that of residual normality. Here are some example Q-Q normality plots from our regression models:





Our first thought about how to address these issues was to remove the outlier data, but this would end up severely hampering our model's ability to handle a lot of the data in our dataset, so after some investigation we decided that a better approach would be using interaction terms to better fit our data, including all outliers. For example, we would only take "model" into consideration based on the "make", and we would only take "trim" into consideration based on the "model", and so on. By doing this, we will be able to create a model that better fits all of the data we are trying to represent. Since we have not determined how to use interaction terms yet, we will begin to research how to apply them in our model and include them in the model we will present in our next step.

Another issue that may be in play is that the depreciation curve of a vehicle is generally not linear, so while a linear relationship may exist between any X variable and our Y variable (and we did not find any evidence showing that our linearity assumption is violated for any model), we may use a log transformation on our Y variable in order to create a better relationship for our model. We will once again include this in the model that we will present in our next step.

Model Creation and Analysis - IE Project Step 4

```
library(knitr)
```

We plan to create our initial model using the following predictor variables with Price as our output variable:

- *Mileage*
- *Make*
- *Cruise*
- *Leather*
- *Make* × *Type* (interaction term)
- *Make* × *Cylinder* (interaction term)
- *Make* × *Doors* (interaction term)
- *Make* × *Sound* (interaction term)

We will first create our model with all of these variables, and then use the backwards elimination procedure to narrow down the model. Let us set our significance level $\alpha = 0.05$.

```
carsData <- read.csv("cars.csv")
attach(carsData)
model.all = lm(Price ~ Mileage + Make + Cruise + Leather + Make *
    Type + Make * Cylinder + Make * Doors + Make * Sound)
summary(model.all)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Make + Cruise + Leather + Make *
##     Type + Make * Cylinder + Make * Doors + Make * Sound)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7515.3 -1407.9    46.1  1335.0  8708.9
##
## Coefficients: (23 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.163e+04  2.817e+03   7.680 4.81e-14 ***
## Mileage                     -1.815e-01  1.058e-02 -17.157  < 2e-16 ***
## MakeCadillac                 2.056e+04  4.138e+03   4.968 8.32e-07 ***
## MakeChevrolet               -8.780e+03  3.087e+03  -2.844 0.004572 **
## MakePontiac                 -1.084e+04  2.951e+03  -3.672 0.000257 ***
## MakeSAAB                     8.716e+03  1.128e+03   7.726 3.42e-14 ***
## MakeSaturn                  -5.149e+03  9.561e+02  -5.385 9.60e-08 ***
## Cruise                       4.348e+02  2.563e+02   1.696 0.090239 .
## Leather                      8.563e+02  2.186e+02   3.917 9.76e-05 ***
## TypeCoupe                   -4.566e+03  9.078e+02  -5.030 6.09e-07 ***
## TypeHatchback               -1.537e+04  9.157e+02 -16.782  < 2e-16 ***
```

```
## TypeSedan                   -6.207e+03  5.730e+02 -10.833  < 2e-16 ***
## TypeWagon                   -4.641e+03  6.147e+02  -7.550 1.22e-13 ***
## Cylinder                     1.411e+03  4.550e+02   3.101 0.001995 **
## Doors                               NA         NA      NA       NA
## Sound                       -2.716e+02  5.754e+02  -0.472 0.637035
## MakeCadillac:TypeCoupe              NA         NA      NA       NA
## MakeChevrolet:TypeCoupe     -7.437e+03  1.270e+03  -5.856 6.98e-09 ***
## MakePontiac:TypeCoupe       -3.765e+03  1.214e+03  -3.102 0.001989 **
## MakeSAAB:TypeCoupe                  NA         NA      NA       NA
## MakeSaturn:TypeCoupe                NA         NA      NA       NA
## MakeCadillac:TypeHatchback          NA         NA      NA       NA
## MakeChevrolet:TypeHatchback         NA         NA      NA       NA
## MakePontiac:TypeHatchback           NA         NA      NA       NA
## MakeSAAB:TypeHatchback              NA         NA      NA       NA
## MakeSaturn:TypeHatchback            NA         NA      NA       NA
## MakeCadillac:TypeSedan      -1.729e+04  1.061e+03 -16.288  < 2e-16 ***
## MakeChevrolet:TypeSedan     -7.406e+03  1.063e+03  -6.970 6.79e-12 ***
## MakePontiac:TypeSedan       -2.138e+03  8.689e+02  -2.461 0.014081 *
## MakeSAAB:TypeSedan                  NA         NA      NA       NA
## MakeSaturn:TypeSedan                NA         NA      NA       NA
## MakeCadillac:TypeWagon              NA         NA      NA       NA
## MakeChevrolet:TypeWagon             NA         NA      NA       NA
## MakePontiac:TypeWagon               NA         NA      NA       NA
## MakeSAAB:TypeWagon                  NA         NA      NA       NA
## MakeSaturn:TypeWagon                NA         NA      NA       NA
## MakeCadillac:Cylinder        1.451e+03  5.674e+02   2.557 0.010741 *
## MakeChevrolet:Cylinder       2.454e+03  4.672e+02   5.253 1.93e-07 ***
## MakePontiac:Cylinder         1.780e+03  4.919e+02   3.619 0.000315 ***
## MakeSAAB:Cylinder                   NA         NA      NA       NA
## MakeSaturn:Cylinder                 NA         NA      NA       NA
## MakeCadillac:Doors                  NA         NA      NA       NA
## MakeChevrolet:Doors                 NA         NA      NA       NA
## MakePontiac:Doors                   NA         NA      NA       NA
## MakeSAAB:Doors                      NA         NA      NA       NA
## MakeSaturn:Doors                    NA         NA      NA       NA
## MakeCadillac:Sound           2.324e+02  8.284e+02   0.281 0.779120
## MakeChevrolet:Sound          1.827e+02  6.899e+02   0.265 0.791257
## MakePontiac:Sound            2.338e+02  7.297e+02   0.320 0.748743
## MakeSAAB:Sound               8.824e+02  7.444e+02   1.185 0.236228
## MakeSaturn:Sound             5.397e+02  8.552e+02   0.631 0.528176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2431 on 776 degrees of freedom
## Multiple R-squared:  0.9416, Adjusted R-squared:  0.9395
## F-statistic:   463 on 27 and 776 DF,  p-value: < 2.2e-16
```

The highest p-values are all from the $Make \times Sound$ variable. All of these p-values are higher than our significance level, 0.05, so we will eliminate it.

```
model.optimal = lm(Price ~ Mileage + Make + Cruise + Leather +
    Make * Type + Make * Cylinder + Make * Doors)
summary(model.optimal)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Make + Cruise + Leather + Make *
##     Type + Make * Cylinder + Make * Doors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7538.8 -1409.6    51.7  1354.5  8685.2
##
## Coefficients: (23 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.131e+04  2.779e+03   7.667 5.24e-14 ***
## Mileage                     -1.815e-01  1.053e-02 -17.236  < 2e-16 ***
## MakeCadillac                 2.081e+04  3.873e+03   5.372 1.03e-07 ***
## MakeChevrolet               -8.528e+03  3.026e+03  -2.818 0.004955 **
## MakePontiac                 -1.056e+04  2.890e+03  -3.653 0.000276 ***
## MakeSAAB                     9.329e+03  1.006e+03   9.269  < 2e-16 ***
## MakeSaturn                  -4.837e+03  8.148e+02  -5.936 4.40e-09 ***
## Cruise                       4.392e+02  2.554e+02   1.720 0.085853 .
## Leather                      8.425e+02  2.116e+02   3.981 7.50e-05 ***
## TypeCoupe                   -4.443e+03  8.979e+02  -4.948 9.18e-07 ***
## TypeHatchback               -1.537e+04  9.135e+02 -16.822  < 2e-16 ***
## TypeSedan                   -6.063e+03  5.608e+02 -10.811  < 2e-16 ***
## TypeWagon                   -4.615e+03  6.126e+02  -7.534 1.36e-13 ***
## Cylinder                     1.412e+03  4.536e+02   3.113 0.001920 **
## Doors                               NA         NA      NA       NA
## MakeCadillac:TypeCoupe              NA         NA      NA       NA
## MakeChevrolet:TypeCoupe     -7.566e+03  1.261e+03  -5.998 3.05e-09 ***
## MakePontiac:TypeCoupe       -3.883e+03  1.165e+03  -3.333 0.000900 ***
## MakeSAAB:TypeCoupe                  NA         NA      NA       NA
## MakeSaturn:TypeCoupe                NA         NA      NA       NA
## MakeCadillac:TypeHatchback          NA         NA      NA       NA
## MakeChevrolet:TypeHatchback         NA         NA      NA       NA
## MakePontiac:TypeHatchback           NA         NA      NA       NA
## MakeSAAB:TypeHatchback              NA         NA      NA       NA
## MakeSaturn:TypeHatchback            NA         NA      NA       NA
## MakeCadillac:TypeSedan      -1.741e+04  1.010e+03 -17.238  < 2e-16 ***
## MakeChevrolet:TypeSedan     -7.555e+03  1.054e+03  -7.169 1.76e-12 ***
## MakePontiac:TypeSedan       -2.272e+03  8.444e+02  -2.691 0.007283 **
## MakeSAAB:TypeSedan                  NA         NA      NA       NA
## MakeSaturn:TypeSedan                NA         NA      NA       NA
## MakeCadillac:TypeWagon              NA         NA      NA       NA
## MakeChevrolet:TypeWagon             NA         NA      NA       NA
## MakePontiac:TypeWagon               NA         NA      NA       NA
## MakeSAAB:TypeWagon                  NA         NA      NA       NA
## MakeSaturn:TypeWagon                NA         NA      NA       NA
## MakeCadillac:Cylinder        1.457e+03  5.561e+02   2.620 0.008967 **
## MakeChevrolet:Cylinder       2.456e+03  4.653e+02   5.279 1.68e-07 ***
## MakePontiac:Cylinder         1.783e+03  4.880e+02   3.653 0.000276 ***
## MakeSAAB:Cylinder                   NA         NA      NA       NA
## MakeSaturn:Cylinder                 NA         NA      NA       NA
## MakeCadillac:Doors                  NA         NA      NA       NA
## MakeChevrolet:Doors                 NA         NA      NA       NA
## MakePontiac:Doors                   NA         NA      NA       NA
```

```
## MakeSAAB:Doors                       NA         NA      NA       NA
## MakeSaturn:Doors                     NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2425 on 782 degrees of freedom
## Multiple R-squared:  0.9414, Adjusted R-squared:  0.9398
## F-statistic: 598.1 on 21 and 782 DF,  p-value: < 2.2e-16
```

The next highest p-value is 0.08, for the *Cruise* variable, and this is also higher than our significance level of 0.05, so we remove it.

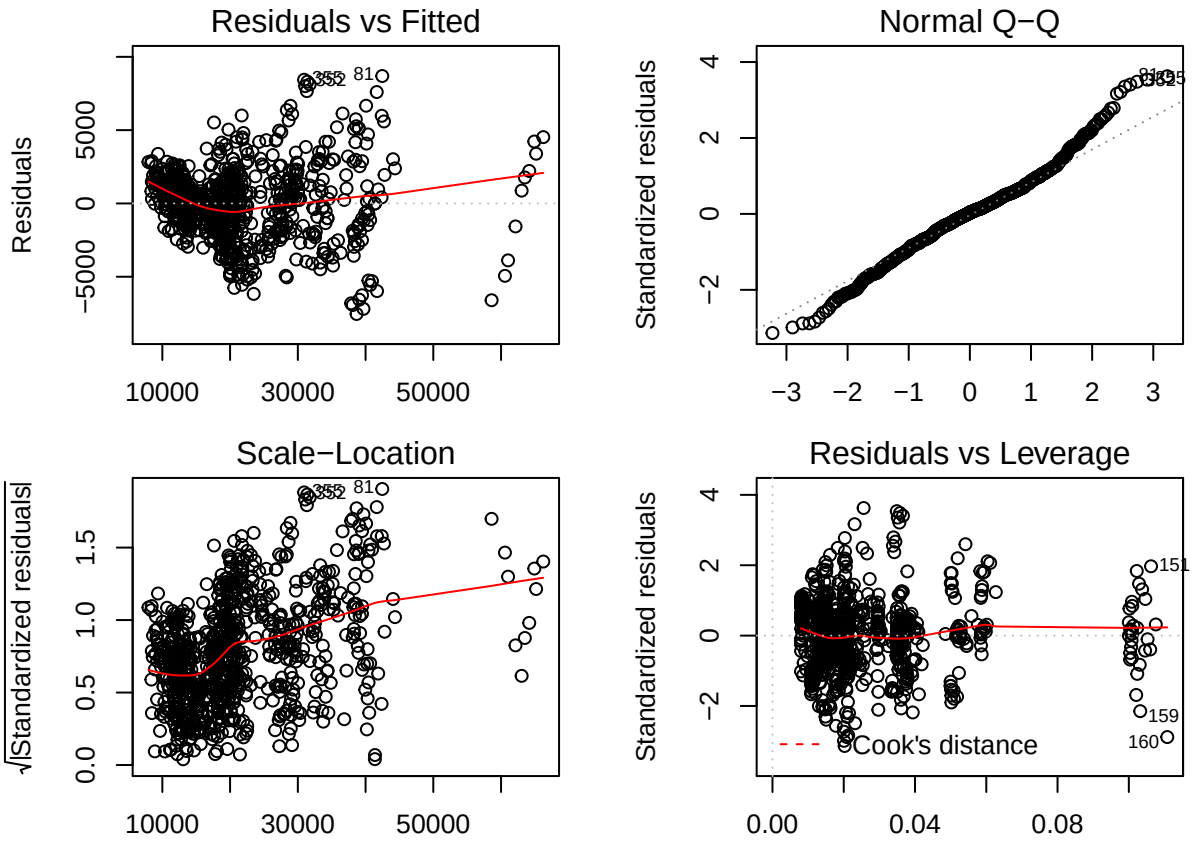
```
model.final = lm(Price ~ Mileage + Make + Leather + Make * Type +
    Make * Cylinder + Make * Doors)
summary(model.final)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Make + Leather + Make * Type +
##     Make * Cylinder + Make * Doors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7540.2 -1473.5    89.9  1316.0  8691.6
##
## Coefficients: (23 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.080e+04  2.767e+03   7.517 1.53e-13 ***
## Mileage                      -1.812e-01  1.054e-02 -17.183  < 2e-16 ***
## MakeCadillac                  2.178e+04  3.836e+03   5.677 1.93e-08 ***
## MakeChevrolet                -7.956e+03  3.012e+03  -2.642 0.008412 **
## MakePontiac                  -1.015e+04  2.884e+03  -3.519 0.000458 ***
## MakeSAAB                      9.652e+03  9.900e+02   9.750  < 2e-16 ***
## MakeSaturn                   -4.839e+03  8.159e+02  -5.931 4.51e-09 ***
## Leather                       8.144e+02  2.113e+02   3.855 0.000125 ***
## TypeCoupe                    -4.362e+03  8.978e+02  -4.858 1.43e-06 ***
## TypeHatchback                -1.551e+04  9.109e+02 -17.023  < 2e-16 ***
## TypeSedan                    -6.059e+03  5.615e+02 -10.791  < 2e-16 ***
## TypeWagon                    -4.605e+03  6.133e+02  -7.507 1.64e-13 ***
## Cylinder                      1.570e+03  4.448e+02   3.530 0.000439 ***
## Doors                                NA         NA      NA       NA
## MakeCadillac:TypeCoupe               NA         NA      NA       NA
## MakeChevrolet:TypeCoupe      -7.602e+03  1.263e+03  -6.021 2.67e-09 ***
## MakePontiac:TypeCoupe        -3.912e+03  1.167e+03  -3.353 0.000837 ***
## MakeSAAB:TypeCoupe                   NA         NA      NA       NA
## MakeSaturn:TypeCoupe                 NA         NA      NA       NA
## MakeCadillac:TypeHatchback           NA         NA      NA       NA
## MakeChevrolet:TypeHatchback          NA         NA      NA       NA
## MakePontiac:TypeHatchback            NA         NA      NA       NA
## MakeSAAB:TypeHatchback               NA         NA      NA       NA
## MakeSaturn:TypeHatchback             NA         NA      NA       NA
## MakeCadillac:TypeSedan       -1.742e+04  1.011e+03 -17.220  < 2e-16 ***
## MakeChevrolet:TypeSedan      -7.598e+03  1.055e+03  -7.203 1.39e-12 ***
```

```
## MakePontiac:TypeSedan       -2.123e+03  8.410e+02  -2.524 0.011785 *
## MakeSAAB:TypeSedan                  NA         NA      NA       NA
## MakeSaturn:TypeSedan                NA         NA      NA       NA
## MakeCadillac:TypeWagon              NA         NA      NA       NA
## MakeChevrolet:TypeWagon             NA         NA      NA       NA
## MakePontiac:TypeWagon               NA         NA      NA       NA
## MakeSAAB:TypeWagon                  NA         NA      NA       NA
## MakeSaturn:TypeWagon                NA         NA      NA       NA
## MakeCadillac:Cylinder        1.298e+03  5.491e+02   2.365 0.018286 *
## MakeChevrolet:Cylinder       2.348e+03  4.615e+02   5.087 4.57e-07 ***
## MakePontiac:Cylinder         1.685e+03  4.852e+02   3.472 0.000546 ***
## MakeSAAB:Cylinder                   NA         NA      NA       NA
## MakeSaturn:Cylinder                 NA         NA      NA       NA
## MakeCadillac:Doors                  NA         NA      NA       NA
## MakeChevrolet:Doors                 NA         NA      NA       NA
## MakePontiac:Doors                   NA         NA      NA       NA
## MakeSAAB:Doors                      NA         NA      NA       NA
## MakeSaturn:Doors                    NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2428 on 783 degrees of freedom
## Multiple R-squared:  0.9412, Adjusted R-squared:  0.9397
## F-statistic: 626.3 on 20 and 783 DF,  p-value: < 2.2e-16
```

The next highest p-value is 0.01 for the $Make \times Type$ variable, specifically in the interaction between $Make = Pontiac$ and $Type = Sedan$. This is lower than our significance level of 0.05, so we stop here and have our final model.

```
par(mfrow=c(2,2), mar=c(2, 4, 2, 2) + 0.1)
plot(model.final)
```

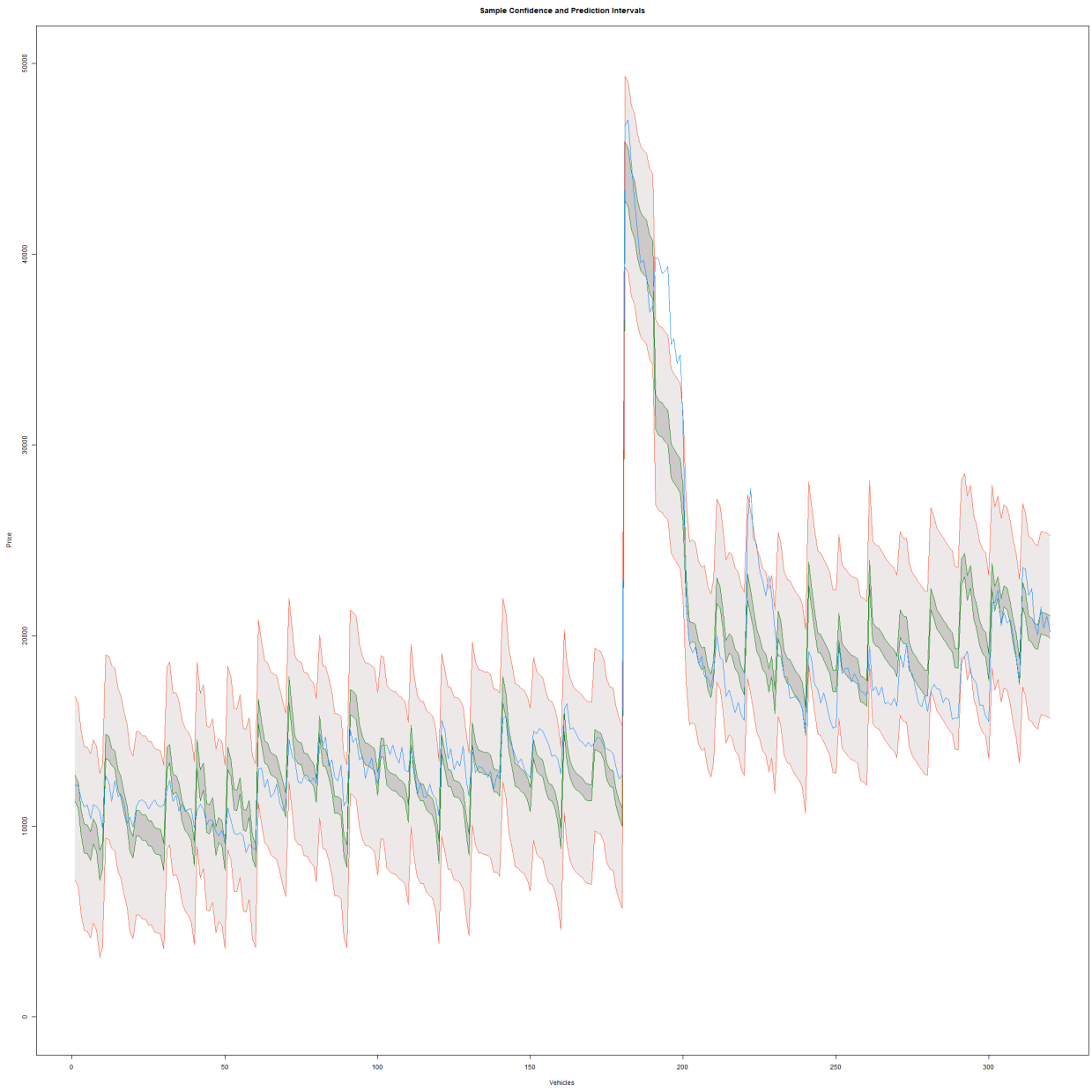


Here is a graph showing prediction and confidence intervals laid over the actual price data. The actual price data is the blue line, the confidence band is between the green lines, and the prediction band is between the red lines. This data is a subset of our total subset, using only vehicles with the make Chevrolet. Including the entire dataset makes the graph too complicated to read effectively.

```
# Mileage + Make + Leather + Make*Type + Make*Cylinder + Make*Doors
# interval.conf.make = predict(model.final, carsData[carsData$Make ==
# 'Chevrolet',], interval='confidence') interval.pred.make = predict(model.final,
# carsData[carsData$Make == 'Chevrolet',], interval='predict')

# png(filename='intervals.png', width=2048, height=2048, units='px')
# plot(1:length(carsData[carsData$Make == 'Chevrolet',]$Price),
# carsData[carsData$Make == 'Chevrolet',]$Price, col='dodgerblue', type='l',
# lwd=1, ylim=c(0, 50000), xlab='Vehicles', ylab='Price', main='Sample Confidence
# and Prediction Intervals') polygon(c(rev(1:length(carsData[carsData$Make ==
# 'Chevrolet',]$Price)), 1:length(carsData[carsData$Make ==
# 'Chevrolet',]$Price)), c(rev(interval.pred.make[ ,3]), interval.pred.make[
# ,2]), col = 'snow2', border = NA) polygon(c(rev(1:length(carsData[carsData$Make
# == 'Chevrolet',]$Price)), 1:length(carsData[carsData$Make ==
# 'Chevrolet',]$Price)), c(rev(interval.conf.make[ ,3]), interval.conf.make[
# ,2]), col = 'snow3', border = NA) lines(1:length(carsData[carsData$Make ==
# 'Chevrolet',]$Price), interval.conf.make[ ,3], lty = 'solid', lwd=1.5, col =
# 'forestgreen') lines(1:length(carsData[carsData$Make == 'Chevrolet',]$Price),
# interval.conf.make[ ,2], lty = 'solid', lwd=1.5, col = 'forestgreen')
# lines(1:length(carsData[carsData$Make == 'Chevrolet',]$Price),
# interval.pred.make[ ,3], lty = 'solid', lwd=1.5, col = 'tomato1')
```

```
# lines(1:length(carsData[carsData$Make == 'Chevrolet',]$Price),
# interval.pred.make[ ,2], lty = 'solid', lwd=1.5, col = 'tomato1')
# lines(1:length(carsData[carsData$Make == 'Chevrolet',]$Price),
# carsData[carsData$Make == 'Chevrolet',]$Price, col='dodgerblue', type='l',
# lwd=1.5)
```



It can be seen that the confidence interval does a decent job of capturing most of the data, and the prediction interval completely captures the real data.

One of the reasons that our model is effectively fit to our data is due to our interaction terms. In this case, specifically the $Make \times Cylinders$ interaction term and the $Make \times Doors$ allows us to capture some of the unique cases in our data (*e.g.* Corvettes and Silverado regular-cab trucks are both 2-door Chevrolet vehicles with 8-cylinder engines, but one is significantly more expensive).

```
random_vehicle <- carsData[242,]
random_vehicle
```

```
##        Price Mileage      Make    Model        Trim  Type Cylinder Liter Doors
## 242 14198.09   11322 Chevrolet Cavalier LS Sedan 4D Sedan        4   2.2     4
##     Cruise Sound Leather
## 242      1     1       1
```

```
pred.random = predict(model.final, random_vehicle, interval="predict")
```

```
## Warning in predict.lm(model.final, random_vehicle, interval = "predict"):
## prediction from a rank-deficient fit may be misleading
```

```
pred.random
```

```
##          fit     lwr      upr
## 242 13618.11 8829.47 18406.74
```

If we select a random data point, we see that our model effectively predicts the paradoxical relationship between leather/sound upgrades and price when it comes to low-trim, econobox vehicles.

```
summary(model.final)$r.squared
```

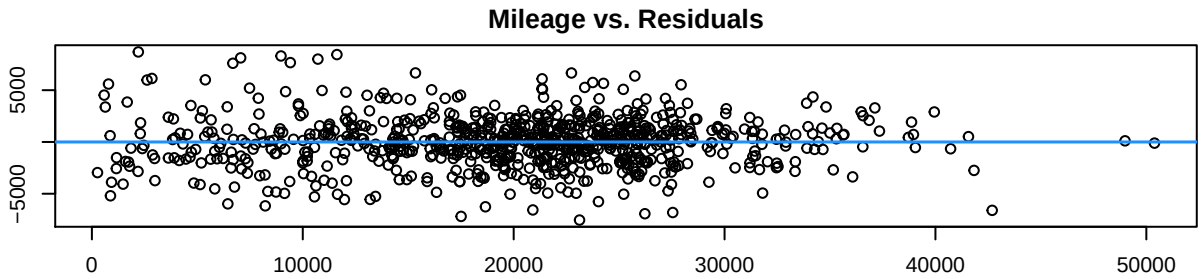
```
## [1] 0.9411677
```

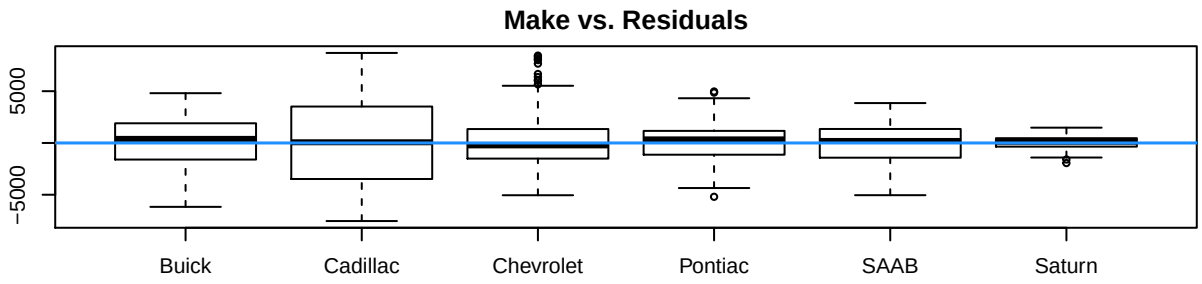
```
summary(model.final)$adj.r.squared
```

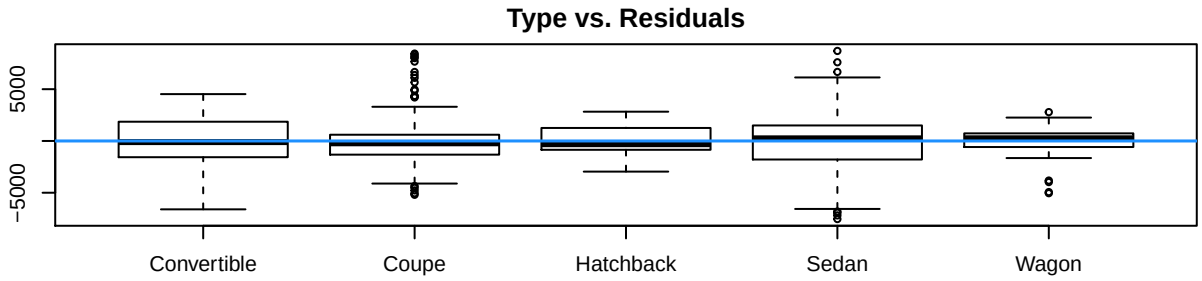
```
## [1] 0.939665
```

We see $R^2 = 0.941$ and $R^2_{adj} = 0.940$, indicating that we have a strong linear relationship between our predictors and our output variable.

```
par(mfrow=c(3,1), mar=c(2,2,2,2) + 0.1)
plot(Mileage, residuals(model.final), xlab="Mileage", ylab="Residuals", main="Mileage vs. Residuals")
abline(0, 0, col='dodgerblue', lwd=1.5)
plot(Make, residuals(model.final), xlab="Make", ylab="Residuals", main="Make vs. Residuals")
abline(0, 0, col='dodgerblue', lwd=1.5)
plot(Type, residuals(model.final), xlab="Type", ylab="Residuals", main="Type vs. Residuals")
abline(0, 0, col='dodgerblue', lwd=1.5)
```







Looking at our primary numerical and categorical variables, there does not appear to be any correlation between our variables and the residuals.

```
summary(model.final)$coefficients
```

```
##                                Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)                2.079716e+04 2.766651e+03   7.517087 1.533968e-13
## Mileage                   -1.811496e-01 1.054247e-02 -17.182839 2.151761e-56
## MakeCadillac               2.177765e+04 3.836102e+03   5.677026 1.930819e-08
## MakeChevrolet             -7.956433e+03 3.011771e+03  -2.641779 8.411745e-03
## MakePontiac               -1.014820e+04 2.883933e+03  -3.518876 4.583859e-04
## MakeSAAB                   9.652170e+03 9.899911e+02   9.749754 2.829416e-21
## MakeSaturn                -4.839148e+03 8.158628e+02  -5.931325 4.507353e-09
## Leather                    8.143812e+02 2.112648e+02   3.854790 1.253274e-04
## TypeCoupe                 -4.361674e+03 8.977551e+02  -4.858423 1.429422e-06
## TypeHatchback             -1.550670e+04 9.109454e+02 -17.022643 1.592194e-55
## TypeSedan                 -6.059224e+03 5.615267e+02 -10.790626 2.105139e-25
## TypeWagon                 -4.604616e+03 6.133480e+02  -7.507346 1.644252e-13
## Cylinder                   1.570238e+03 4.447704e+02   3.530447 4.390988e-04
## MakeChevrolet:TypeCoupe   -7.602030e+03 1.262686e+03  -6.020521 2.670847e-09
## MakePontiac:TypeCoupe     -3.911836e+03 1.166524e+03  -3.353412 8.365814e-04
## MakeCadillac:TypeSedan    -1.741775e+04 1.011457e+03 -17.220447 1.343427e-56
## MakeChevrolet:TypeSedan   -7.598327e+03 1.054881e+03  -7.203015 1.385000e-12
## MakePontiac:TypeSedan     -2.122942e+03 8.409526e+02  -2.524449 1.178455e-02
## MakeCadillac:Cylinder      1.298398e+03 5.490685e+02   2.364729 1.828629e-02
## MakeChevrolet:Cylinder     2.347643e+03 4.615391e+02   5.086553 4.567021e-07
```

```
## MakePontiac:Cylinder     1.684547e+03 4.852328e+02   3.471626 5.456696e-04
```

If we look at the coefficients, we can see the influence each predictor variable has on the vehicle's price. Higher mileage lowers price. Vehicles with the make of either Cadillac or SAAB have a higher price, whereas Chevrolet, Pontiac, and Saturn have a lower price. Having leather seats increases price. All body types reduce price, but by different amounts (being a hatchback reduces price by the least). We can also see the coefficients of our interaction terms and how these terms influence our model. For example, Chevrolet vehicles increase the most in price due to cylinder, as 8-cylinder Chevrolets tend to be Corvettes. On the other hand, Cadillac vehicles all tend to have more cylinders in their engines, and are more expensive due to a variety of other factors such as brand cachet, so Cylinder has less effect on the price of a Cadillac.

From analyzing these coefficients, we see that our model is doing a good job of accounting for all the predictor variables we used, especially with the interaction terms.

Car Values

Connor, Evan, Jianjun, Ritwik

Abstract

- Data collected from 2005 Kelley Blue Book for used GM cars stored in csv format
- Purpose
  - To determine car value based on variety of vehicular characteristics
- Population
  - Used 2005 GM cars, less than one year old, and in “excellent” condition
  - **Excellent condition** means that the vehicle **looks new**, is in **excellent mechanical condition** and needs **no reconditioning**. This vehicle has **never had any paint or body work** and is **free of rust**. The vehicle has a **clean Title History** and **will pass a smog and safety inspection**. The **engine compartment is clean**, with **no fluid leaks** and is **free of any wear or visible defects**. The vehicle also has **complete and verifiable service records**. Less than 5 percent of all used vehicles fall into this category.
- Dimensions
  - 804 observations with 12 variables each

Exploratory Data Analysis

Data Descriptions

Exploratory Data Analysis

Price
Suggested retail price (USD) of the used 2005 GM car in excellent condition.

Mileage
Odometer mileage.

Cylinder
Number of cylinders in the engine.

Liter
Engine displacement, in liters.

Doors
Number of doors.

Make
Manufacturer of the car such as Cadillac, Pontiac, and Chevrolet.

Model
Specific models for each vehicle such as Trans-Am, Silverado, or Regal.

Trim
Trim level for each vehicle, such as “Limited Sedan 4D”, “Coupe 2D”, or “LT MAXX Hback 4D”.

Type
Vehicle body type, such as sedan or coupe.

Cruise
Indicator variable representing whether the car has cruise control.

Sound
Indicator variable representing whether the car has *upgraded* speakers.

Leather
Indicator variable representing whether the car has leather seats.

Key

Outcome

Quantitative

Categorical

Factor

Descriptive Statistics

Exploratory Data Analysis

Cruise Control	
No Cruise Control	199
Cruise Control	605

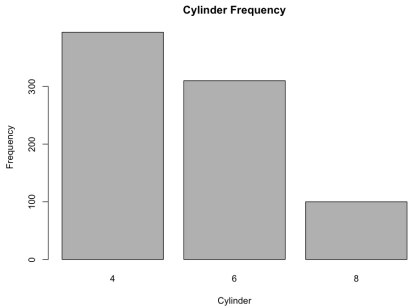
Doors	
2	190
4	614

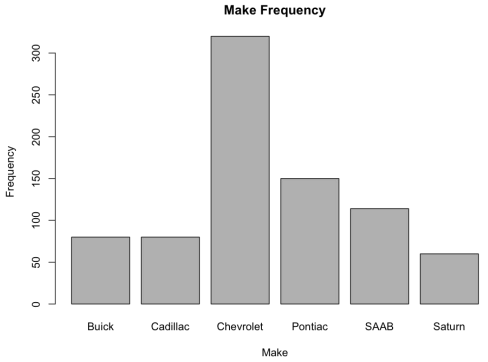
Sound System	
Base Sound System	258
Upgraded Sound System	546

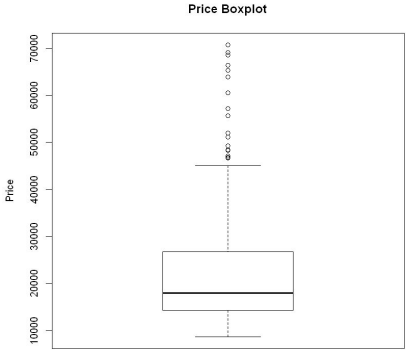
Cylinder	
4	394
6	310
8	100

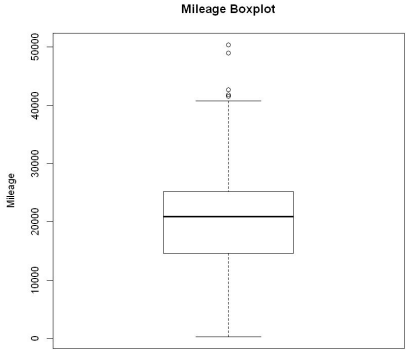
Leather Interior	
No Leather Interior	222
Leather Interior	582

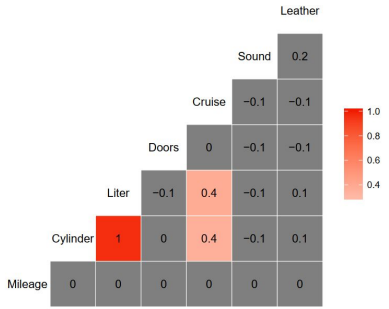
Type	
Convertible	50
Coupe	140
Hatchback	60
Sedan	490
Wagon	64











Multiple Linear Regression

Full Model Coefficients

Multiple Linear Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26408	514.2	51.36	1.418e-244
Mileage	-0.1854	0.004051	-45.75	1.303e-216
MakeCadillac	39454	572.7	68.89	1.215e-321
MakeChevrolet	-6518	484.1	-13.46	4.54e-37
MakePontiac	-7953	574.9	-13.83	8.116e-39
MakeSAAB	5254	879.9	5.97	3.693e-09
⋮	⋮	⋮	⋮	⋮
TrimSS Coupe 2D	5413	597	9.066	1.121e-18
TrimSS Sedan 4D	6183	510.7	12.11	7.053e-31
TrimSVM Hatchback 4D	-794.8	407.2	-1.952	0.05133
Cruise	69.4	101.5	0.6836	0.4944
Sound	211.3	79.76	2.649	0.008257
Leather	295.4	92.87	3.18	0.001533

Full Model Summary

Multiple Linear Regression

Table 2: Fitting linear model: Price ~ Mileage + Make + Model + Trim + Type + Cylinder + Liter + Doors + Cruise + Sound + Leather

Observations	Residual Std. Error	R^2	Adjusted R^2
804	903.3	0.9924	0.9916

Table 3: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mileage	1	1.606e+09	1.606e+09	1968	2.179e-209
Make	5	5.054e+10	1.011e+10	12389	0
Model	26	2.386e+10	917500599	1124	0
Trim	38	1.846e+09	48590983	59.55	4.738e-196
Cruise	1	413181	413181	0.5064	0.4769
Sound	1	5800306	5800306	7.108	0.007842
Leather	1	8253645	8253645	10.12	0.001533
Residuals	730	595658307	815970	NA	NA

Model Selection

Multiple Linear Regression

We had an intuitive idea that we would need to create a **reduced model** from our initial model that included every possible predictor.

We decided to use the **backwards elimination** procedure for eliminating predictor variables from our model.

Relying on **Adjusted R^2** as a criterion for model evaluation led us to an over-fitted model.

Overfitting made us **rely more on substantive reasons** for variables **rather than stepwise and other criterion-based analytical approaches**. We also wanted to **keep predictor count low** in order to make the model easier to handle.

BIC Model

Price ~ Mileage + Make + Type + Cylinder + Liter + Doors

Observations	**804**
Residual Std. Error	**2518**
R^2	**0.9361**
Adjusted R^2	**0.9351**

Final Predictor Variables

Multiple Linear Regression

Mileage
Good predictor for price and the only quantitative predictor.

Leather
Good factor predictor for price, that affects our data equally.

Make
Good categorical predictor for price.

Cylinder
Good factor predictor for price, and since it has an almost perfect correlation with **Liter**, we can use just **Cylinder** and simplify our model.

Type
Good categorical predictor for price, and since vehicle type is fundamentally related to the number of doors, we do not need to include **Doors** in our model, as its effect will be included in the effect of **Type**.

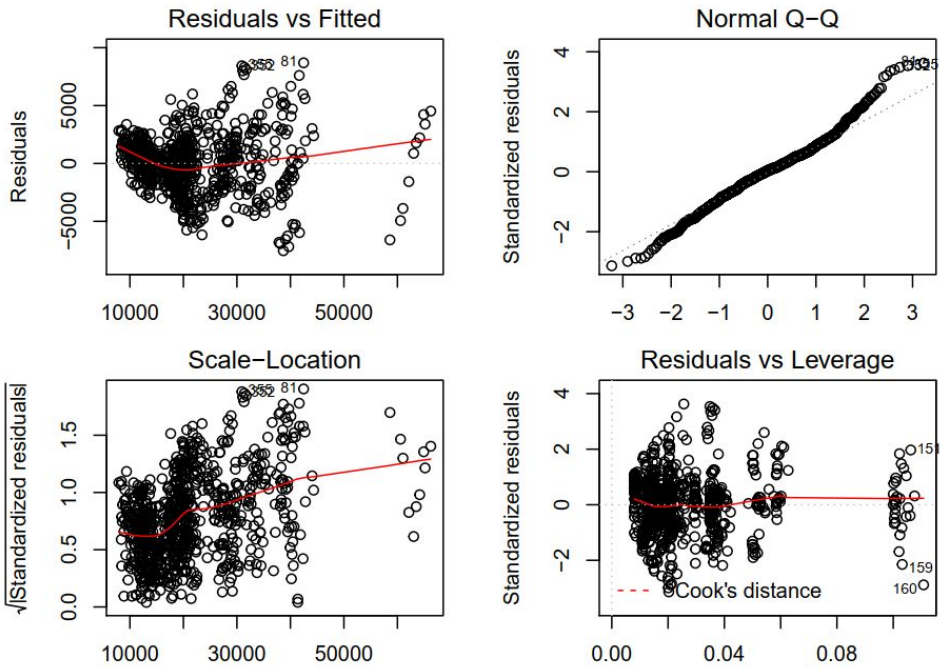
Make × Cylinder
Cylinder is a good predictor on its own, but some situations (for example, a 6-cylinder Chevrolet pickup is a low-price vehicle, whereas a 6-cylinder Chevrolet sedan is a medium-price vehicle) can only be addressed with an interaction term.

Make × Type
Type is a good predictor on its own, but some situations (such as a wagon being more expensive than a sedan for some makes) can only be addressed with an interaction term.

Key
- **Quantitative**
- **Categorical**
- **Factor**
- **Interaction Term**

Diagnostic Plots

Multiple Linear Regression



Looking at these diagnostic plots, we see slight heteroscedasticity for the residuals versus fitted values and a slightly heavy tail on the normality plot.

The heteroscedasticity of residuals against fitted values is explained by the lack of data for high-priced vehicles. Since most of our data is for low-to-medium priced vehicles, the model is not as confident about higher-priced vehicles, and thus we see the increase in residuals.

Since our model functions by using the higher-priced vehicles as the categorical standard, it tends to overestimate values for vehicles rather than underestimate values. This explains the slightly heavy tail in our Q-Q plot.

Results

Final Model Coefficients

Results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20797	2767	7.517	1.534e-13
Mileage	-0.1811	0.01054	-17.18	2.152e-56
MakeCadillac	21778	3836	5.677	1.931e-08
MakeChevrolet	-7956	3012	-2.642	0.008412
MakePontiac	-10148	2884	-3.519	0.0004584
MakeSAAB	9652	990	9.75	2.829e-21
MakeSaturn	-4839	815.9	-5.931	4.507e-09
Leather	814.4	211.3	3.855	0.0001253
TypeCoupe	-4362	897.8	-4.858	1.429e-06
TypeHatchback	-15507	910.9	-17.02	1.592e-55
TypeSedan	-6059	561.5	-10.79	2.105e-25
TypeWagon	-4605	613.3	-7.507	1.644e-13
Cylinder	1570	444.8	3.53	0.0004391
MakeChevrolet:TypeCoupe	-7602	1263	-6.021	2.671e-09
MakePontiac:TypeCoupe	-3912	1167	-3.353	0.0008366
MakeCadillac:TypeSedan	-17418	1011	-17.22	1.343e-56
MakeChevrolet:TypeSedan	-7598	1055	-7.203	1.385e-12
MakePontiac:TypeSedan	-2123	841	-2.524	0.01178
MakeCadillac:Cylinder	1298	549.1	2.365	0.01829
MakeChevrolet:Cylinder	2348	461.5	5.087	4.567e-07
MakePontiac:Cylinder	1685	485.2	3.472	0.0005457

Final Model Summary

Results

Table 4: Fitting linear model: Price ~ Mileage + Make + Leather + Make * Type + Make * Cylinder

Observations	Residual Std. Error	R^2	Adjusted R^2
804	2428	0.9412	0.9397

Table 5: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Mileage	1	1.606e+09	1.606e+09	272.3	9.902e-53
Make	5	5.054e+10	1.011e+10	1715	0
Leather	1	82501328	82501328	13.99	0.0001968
Type	4	9.202e+09	2.3e+09	390.2	9.942e-185
Cylinder	1	1.032e+10	1.032e+10	1751	7.098e-202
Make:Type	5	1.879e+09	375842848	63.75	7.835e-56
Make:Cylinder	3	209131230	69710410	11.82	1.399e-07
Residuals	783	4.616e+09	5895355	NA	NA

Conclusion

Results

- Traditional analytical model selection approaches don't alway work
  - Adjusted R^2 entirely fails to account for overfitting
- Domain knowledge becomes more important when overfitting issues arise
  - Ability to diagnose **Model** and **Trim** as responsible for overfitting was only possible due to domain knowledge about cars
- Many predictors rely on **Make**
  - **Make,** or "brand", is the most important variable when it comes to predicting price
  - **Mileage** is second most important
- Stay within scope of model
  - Rough estimate to ensure a fair price for similar set of cars
  - Used as more of a baseline

Issues and Next Steps

Results

- Overfitting in model
  - Use of cross validation to train and test model
- Too many levels in qualitative variables
  - Required to drop variables for exhaustive model searches
  - Cross validating becomes nearly impossible
- Old data and variables
  - Updated data set with 2020 variables
    - Fuel type (gasoline, diesel, electric, hydrogen)
    - Fuel efficiency
    - Drivetrain format (RWD, FWD, AWD, 4x4)
    - Seating capacity or layout (2+3, 2+2+2, 3+3, 2+3+3)
  - More data, especially for each category
  - Opening the population to more cars could resolve this problem