Multiple Regression

How Much Is Your Car Worth?

Mina Akhlaghi

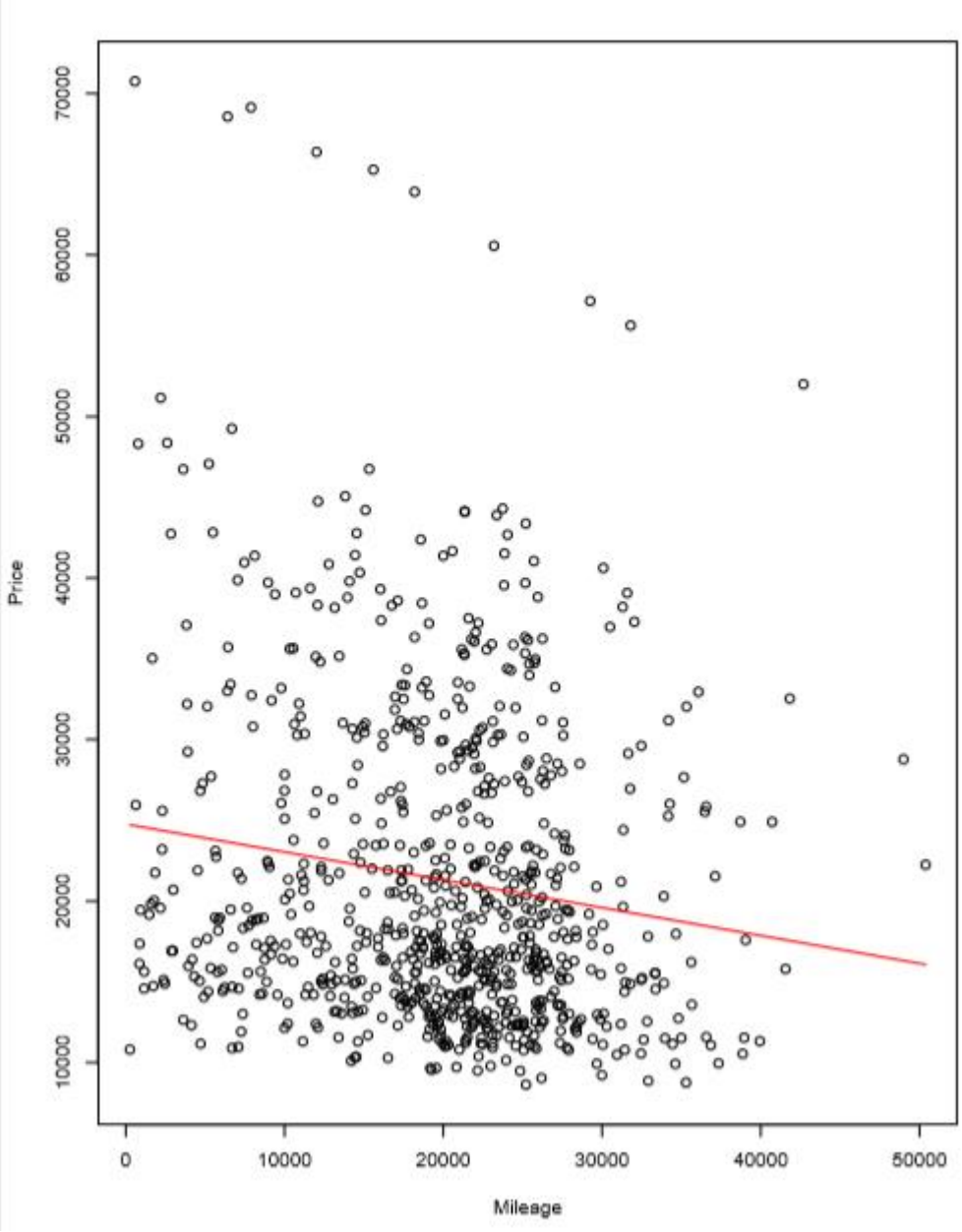
CONTENTS

OBJECTIVE

In the following report we will be displaying the predictive power of different variables in accordance with the price of a Car. We will display numerous techniques and offer an analysis of a number of different data points and various regression techniques in order to show the value of data analysis. We will demonstrate the predictive power of regression modelling when trying to predict a specific value.

SIMPLE LINEAR REGRESSION MODEL

Production of a Scatter Plot to display the relationship between Price and Mileage of a Car's predicted Price:



According to the scatter plot there is a strong relationship between price and mileage. The scatter plot shows us as mileage increases there is downward pressure on the price of cars. We also notice that the scatterplot shows a set of data points with a higher retail price that don't fall in the general cluster of data.

Least Square Regression:

```
Call:
lm(formula = Price ~ Mileage)

Residuals:
   Min     1Q Median     3Q    Max
-13905  -7254  -3520   5188  46091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.476e+04  9.044e+02  27.383  < 2e-16 ***
Mileage     -1.725e-01  4.215e-02  -4.093 4.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9789 on 802 degrees of freedom
Multiple R-squared:  0.02046,   Adjusted R-squared:  0.01924
F-statistic: 16.75 on 1 and 802 DF,  p-value: 4.685e-05
Analysis of Variance Table

Response: Price
           Df     Sum Sq    Mean Sq F value    Pr(>F)
Mileage     1 1.6056e+09 1605590375  16.755 4.685e-05 ***
Residuals 802 7.6856e+10   95830165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Price = 24723 – 0.17 Mileage
R Squared = 2.05
T - Statistic of B0= 27.83
T - Statistic of B1 = -4.09
P Value of B0 < 2e 16
P Value of B1 >4.68
While the p-value for b1 indicates that mileage is an important variable, the R-Sq value shows that the model does not account for much of the variation in prices and there is not a strong relationship between mileage and price.
Let's make the following Hypothesis:
H0: b1=0
H1: b1 is different than 0
T alpha = T 0.025 = 1,963 with DF=805 - 2=803
T stat = -4.093 |T stat| >T0.025;

We reject H0 There is evidence that Mileage is a strong indicator of price

Calculating the residual value of a Buick Century with 8221 miles:
We can calculate the value of the Buick by inputting the amount of Miles into the E(y) = B0+B1 equation.
The expected price of the Buick will be Price = 24723 - 0.17(8221)
Residual value = Actual price – Expected price = 17,314.103 – 23,342.33 = (6028.23)

COMPARING VARIABLE SELECTION TECHNIQUES

Single Variable Regression Model Technique:

Cylinder Regression Model

```
Call:
lm(formula = Price ~ Cylinder)

Residuals:
   Min     1Q Median     3Q    Max
-11216  -5230  -2749   2773  38339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -17.06    1126.94  -0.015    0.988
Cylinder     4054.20     206.85  19.600   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8133 on 802 degrees of freedom
Multiple R-squared:  0.3239,    Adjusted R-squared:  0.323
F-statistic: 384.1 on 1 and 802 DF,  p-value: < 2.2e-16
```

Liter Regression Model

```
Call:
lm(formula = Price ~ Liter)

Residuals:
   Min     1Q Median     3Q    Max
-10186  -5128  -3172   3032  41614

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6185.8      846.7   7.306 6.66e-13 ***
Liter         4990.4      262.0  19.050  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8207 on 802 degrees of freedom
Multiple R-squared:  0.3115,    Adjusted R-squared:  0.3107
F-statistic: 362.9 on 1 and 802 DF,  p-value: < 2.2e-16
```

Cruise Regression Model

```
Call:
lm(formula = Price ~ Cruise)

Residuals:
   Min     1Q Median     3Q    Max
-14913  -6020  -1454   3634  46971

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13921.9      632.7   22.00   <2e-16 ***
Cruise        9862.3      729.4   13.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8926 on 802 degrees of freedom
Multiple R-squared:  0.1856,    Adjusted R-squared:  0.1846
F-statistic: 182.8 on 1 and 802 DF,  p-value: < 2.2e-16
```

Doors Regression Model

```
Call:
lm(formula = Price ~ Doors)

Residuals:
   Min     1Q Median     3Q    Max
-13018  -7052  -2800   5420  46948

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23807.1      710.6  33.502  < 2e-16 ***
Doors4       -3226.5      813.2  -3.968 7.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9795 on 802 degrees of freedom
Multiple R-squared:  0.01925,   Adjusted R-squared:  0.01803
F-statistic: 15.74 on 1 and 802 DF,  p-value: 7.906e-05
```

Sound Regression Model

```
Call:
lm(formula = Price ~ Sound)

Residuals:
   Min     1Q Median     3Q    Max
-14491  -6874  -3184    5014  50257

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23130.1      611.0  37.856  < 2e-16 ***
Sound        -2631.4      741.4  -3.549 0.000409 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9814 on 802 degrees of freedom
Multiple R-squared:  0.01546,   Adjusted R-squared:  0.01423
F-statistic:  12.6 on 1 and 802 DF,  p-value: 0.0004092
```

Leather Regression Model

```
Call:
lm(formula = Price ~ Leather)

Residuals:
   Min     1Q Median     3Q    Max
-13260  -7435  -2691    5422  48453

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18828.8      655.6  28.720  < 2e-16 ***
Leather       3473.5      770.5   4.508 7.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9768 on 802 degrees of freedom
Multiple R-squared:  0.02471,   Adjusted R-squared:  0.02349
F-statistic: 20.32 on 1 and 802 DF,  p-value: 7.526e-06
```

Mileage Regression Model

```
Call:
lm(formula = Price ~ Mileage)

Residuals:
   Min     1Q Median     3Q    Max
-13905  -7254  -3520   5188  46091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.476e+04  9.044e+02  27.383  < 2e-16 ***
Mileage     -1.725e-01  4.215e-02  -4.093 4.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9789 on 802 degrees of freedom
Multiple R-squared:  0.02046,   Adjusted R-squared:  0.01924
F-statistic: 16.75 on 1 and 802 DF,  p-value: 4.685e-05
```

By analyzing the regression models with a single variable, it is apparent that the highest R-Sq for 1 variable regression is Cylinder with a 32.39% predictive power.

Use Of two Variables: one of which is fixed (Cylinder), to conduct 6 Regression Models:

Cylinder + Liter Regression Model

```
Call:
lm(formula = Price ~ Cylinder + Liter)

Residuals:
   Min     1Q Median     3Q    Max
-10479  -5182  -2944   3034  39076

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1372.4     1434.5   0.957    0.339
Cylinder      2976.4      719.8   4.135 3.92e-05 ***
Liter         1412.2      903.4   1.563    0.118
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8126 on 801 degrees of freedom
Multiple R-squared:  0.3259,    Adjusted R-squared:  0.3242
F-statistic: 193.6 on 2 and 801 DF,  p-value: < 2.2e-16
```

Cylinder + Door Regression Model

```
Call:
lm(formula = Price ~ Cylinder + Doors)

Residuals:
   Min     1Q Median     3Q    Max
-12093  -5565  -2888   3085  35847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5713.3     1614.9   3.538 0.000426 ***
Cylinder      4056.4      204.0  19.888  < 2e-16 ***
Doors        -1627.8      332.9  -4.890 1.22e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8019 on 801 degrees of freedom
Multiple R-squared:  0.3435,    Adjusted R-squared:  0.3418
F-statistic: 209.5 on 2 and 801 DF,  p-value: < 2.2e-16
```

Cylinder + Cruise Regression Model

```
Call:
lm(formula = Price ~ Cylinder + Cruise)

Residuals:
   Min     1Q Median     3Q    Max
-11724  -5695  -1961   3555  38661

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1046.4     1082.7  -0.967    0.334
Cylinder      3392.6      211.3  16.058   <2e-16 ***
Cruise        6000.4      678.8   8.839   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7768 on 801 degrees of freedom
Multiple R-squared:  0.3839,    Adjusted R-squared:  0.3824
F-statistic: 249.6 on 2 and 801 DF,  p-value: < 2.2e-16
```

Cylinder + Sound Regression Model

```
Call:
lm(formula = Price ~ Cylinder + Sound)

Residuals:
   Min     1Q Median     3Q    Max
-11946  -5429  -2607   2792  38970

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1293.7     1235.7   1.047   0.2955
Cylinder      4007.0      207.0  19.359   <2e-16 ***
Sound        -1563.7      614.8  -2.543   0.0112 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8106 on 801 degrees of freedom
Multiple R-squared:  0.3293,    Adjusted R-squared:  0.3276
F-statistic: 196.6 on 2 and 801 DF,  p-value: < 2.2e-16
```

Cylinder + Leather Regression Model

```
Call:
lm(formula = Price ~ Cylinder + Leather)

Residuals:
   Min     1Q Median     3Q    Max
-11748  -5318  -2838   3078  37807

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1528.9     1179.4  -1.296    0.195
Cylinder      3992.4      205.5  19.423  < 2e-16 ***
Leather       2538.3      637.5   3.981 7.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8059 on 801 degrees of freedom
Multiple R-squared:  0.337,     Adjusted R-squared:  0.3353
F-statistic: 203.6 on 2 and 801 DF,  p-value: < 2.2e-16
```

Cylinder + Mileage Regression Model

```
Call:
lm(formula = Price ~ Cylinder + Mileage)

Residuals:
   Min     1Q Median     3Q    Max
-10264  -5121  -2838   3102  35477

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3145.75032 1325.93436   2.372   0.0179 *
Cylinder    4027.67463  204.61180  19.684  < 2e-16 ***
Mileage       -0.15243    0.03464  -4.401 1.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8042 on 801 degrees of freedom
Multiple R-squared:  0.3398,    Adjusted R-squared:  0.3382
F-statistic: 206.2 on 2 and 801 DF,  p-value: < 2.2e-16
```

When conducting regression models with two variables we find that the best model with the highest R-SQ is Cylinder + Cruise Model with 38.39% predictive power and an increase of 6%.

Stepwise Regression Analysis:

```
Start:  AIC=14331.39
Price ~ Cylinder + Liter + Doors + Cruise + Sound + Leather +
    Mileage

           Df  Sum of Sq        RSS   AIC
- Liter     1   44992316 4.3492e+10 14330
<none>                  4.3447e+10 14331
- Sound     1  663674405 4.4111e+10 14342
- Doors     1 1265022645 4.4712e+10 14352
- Mileage   1 1548003060 4.4995e+10 14358
- Cylinder  1 1681894212 4.5129e+10 14360
- Leather   1 1714076275 4.5161e+10 14360
- Cruise    1 4986166639 4.8433e+10 14417

Step:  AIC=14330.22
Price ~ Cylinder + Doors + Cruise + Sound + Leather + Mileage

           Df  Sum of Sq        RSS   AIC
<none>                  4.3492e+10 14330
+ Liter     1 4.4992e+07 4.3447e+10 14331
- Sound     1 6.8659e+08 4.4178e+10 14341
- Doors     1 1.2297e+09 4.4722e+10 14351
- Mileage   1 1.5632e+09 4.5055e+10 14357
- Leather   1 1.6943e+09 4.5186e+10 14359
- Cruise    1 4.9514e+09 4.8443e+10 14415
- Cylinder  1 1.3563e+10 5.7055e+10 14546
```

The variables Sound, Doors, Mileage, Leather, Cruise, and Cylinder are all recommended by the stepwise regression. Only eliminate Liter as a possible variable to take out of the model. This may not be 100 percent accurate and we must analyze a bit further to determine if we should remove Liter from our model.

Subset Selection Model

```
Subset selection object
Call: regsubsets.formula(Price ~ Cylinder + Liter + Doors + Cruise +
    Sound + Leather + Mileage, data = cars, nbest = 1)
7 Variables  (and intercept)
         Forced in Forced out
Cylinder     FALSE      FALSE
Liter        FALSE      FALSE
Doors        FALSE      FALSE
Cruise       FALSE      FALSE
Sound        FALSE      FALSE
Leather      FALSE      FALSE
Mileage      FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         Cylinder Liter Doors Cruise Sound Leather Mileage
1  ( 1 ) "*"      " "   " "   " "    " "   " "     " "
2  ( 1 ) "*"      " "   " "   "*"    " "   " "     " "
3  ( 1 ) "*"      " "   " "   "*"    " "   "*"     " "
4  ( 1 ) "*"      " "   " "   "*"    " "   "*"     "*"
5  ( 1 ) "*"      " "   "*"   "*"    " "   "*"     "*"
6  ( 1 ) "*"      " "   "*"   "*"    "*"   "*"     "*"
7  ( 1 ) "*"      "*"   "*"   "*"    "*"   "*"     "*"
```

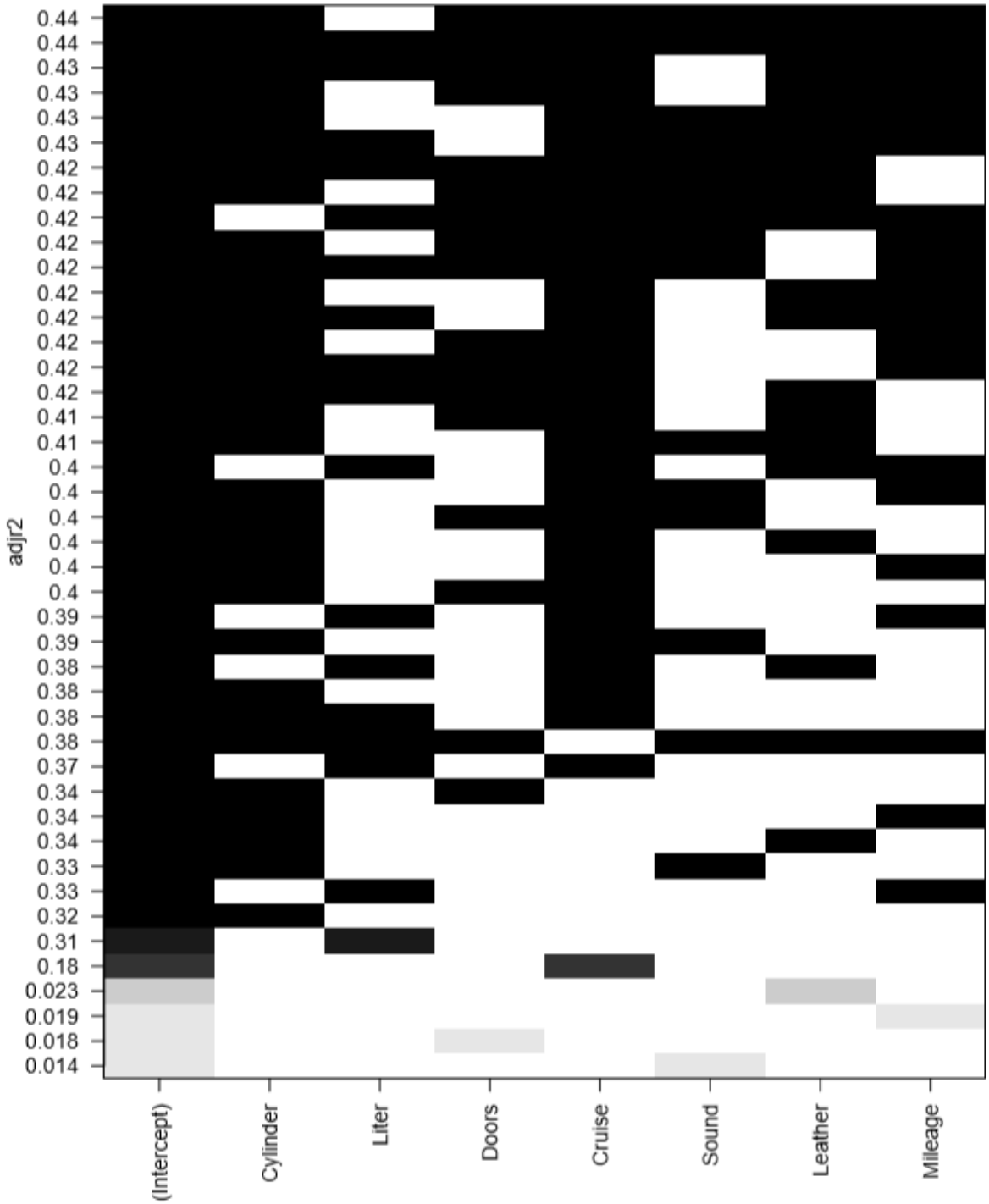
Based on the subset data Cylinder, Door, Cruise, Sound, Leather, Mileage should be used in the model. Liter can be eliminated from our optimal model based on its low adjusted R-SQ.

Analysis of Step Wise And Subset Techniques

According to the analyst given by each of the Stepwise Regression and the Subset Regression, each table gives us the same explanatory variables as important to the model. Each set of data tell us that Liter is not a significant determinant of Price.

The stepwise regression analysis tells us that the data variable "Liter" is not significant and can be safely removed from the model. The same can be said about the subset data. The adjusted R-Sq of Liter was lower than all other variables which tells us we can safely assume that Liter is not a significant variable for predicting price. The higher the adjusted R-SQ the better the data can be safely identified as significant which is why the subset data is a better indicator of Liters predictability of price.
We conducted one further test to determine which variable we should use in our regression model.

This was the AIC test model. Below the AIC table proves that Liter has the lowest Adjusted R-Sq and should be removed from our model:



Regression Equation:

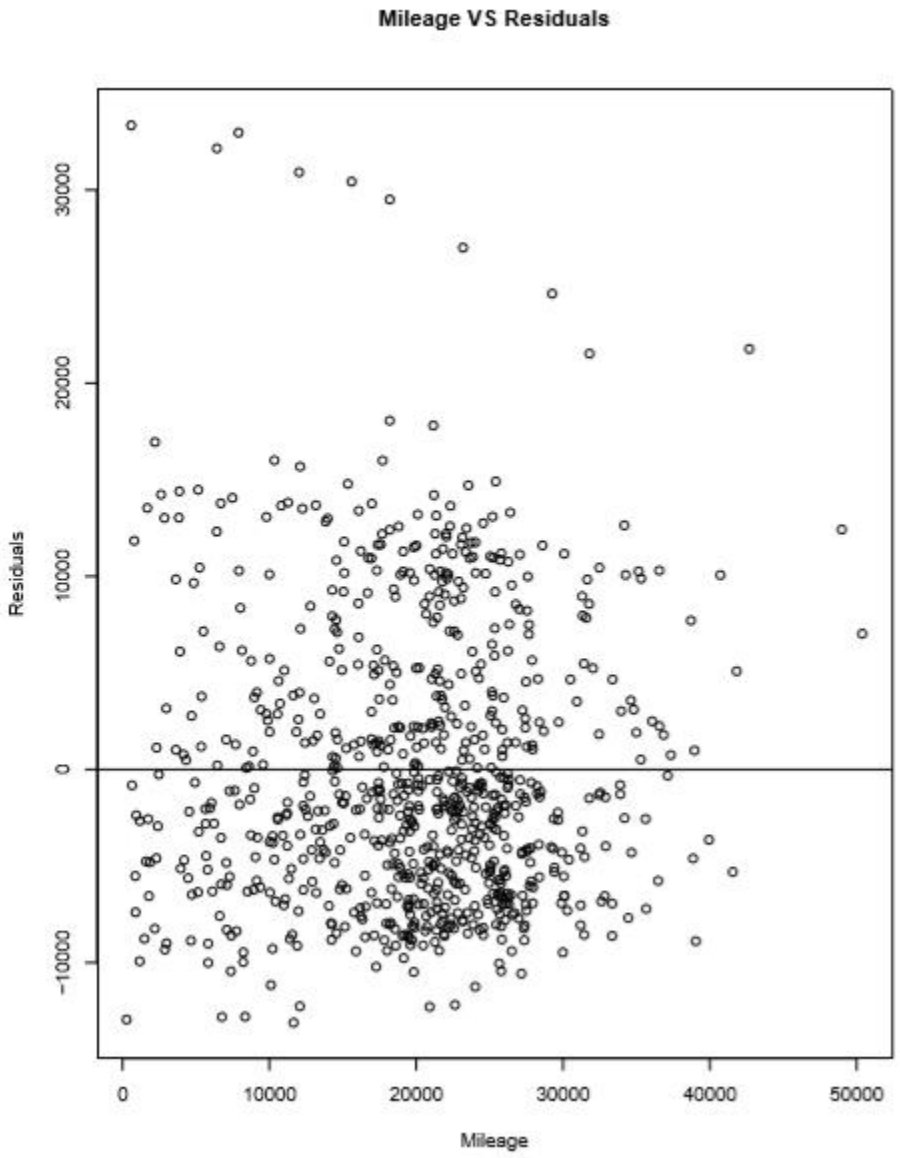
```
Call:
lm(formula = Price ~ Mileage + Cylinder + Cruise + Leather +
    Doors + Sound)

Residuals:
   Min     1Q Median     3Q    Max
-13104  -5566  -1544   3877  33349

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.396e+03  1.448e+03   3.035 0.002481 **
Mileage     -1.705e-01  3.186e-02  -5.352 1.14e-07 ***
Cylinder     3.200e+03  2.030e+02  15.765  < 2e-16 ***
Cruise1      6.206e+03  6.515e+02   9.525  < 2e-16 ***
Leather1     3.327e+03  5.971e+02   5.572 3.45e-08 ***
Doors4      -2.927e+03  6.165e+02  -4.747 2.45e-06 ***
Sound1      -2.024e+03  5.707e+02  -3.547 0.000412 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7387 on 797 degrees of freedom
Multiple R-squared:  0.4457,    Adjusted R-squared:  0.4415
F-statistic: 106.8 on 6 and 797 DF,  p-value: < 2.2e-16
```

This is our model based on best subset techniques, AIC, and stepwise regression. Best subset techniques and AIC proved to be our best predictor because we are focused on achieving the highest adjusted R-SQ and increasing our predictive power. Significance levels are important and need to be analyzed but when we did further testing we found the AIC to be our most significant predictor.

CHECKING THE MODEL ASSUMPTION

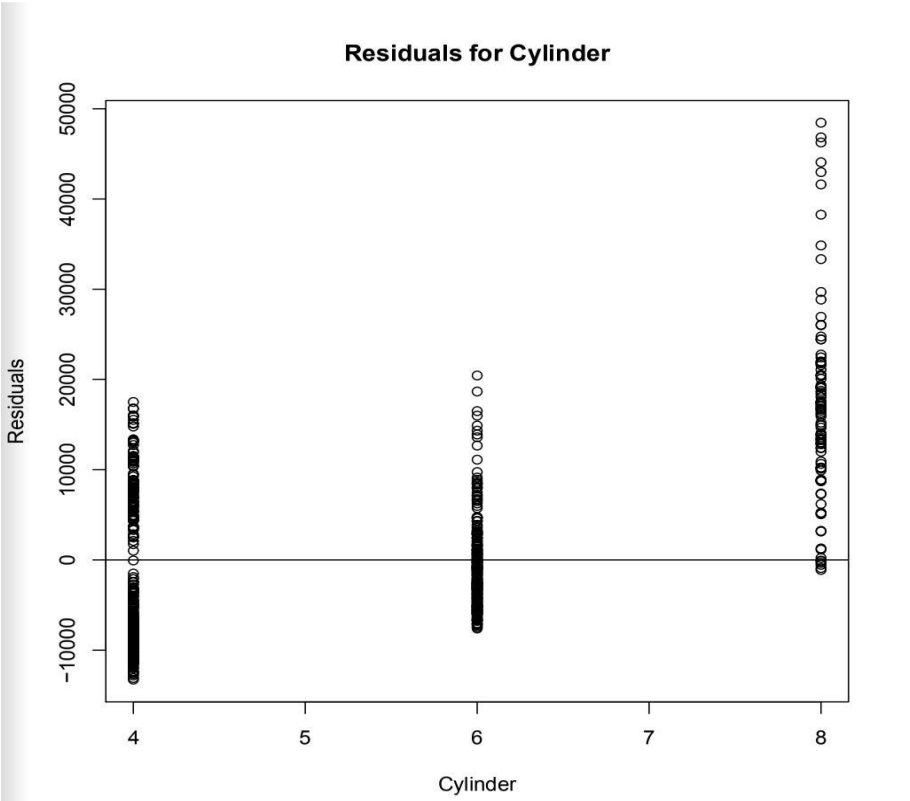
The following data plots will provide a visual distribution of Residuals vs each explanatory variable required in our model:

Residuals vs Mileage Plot

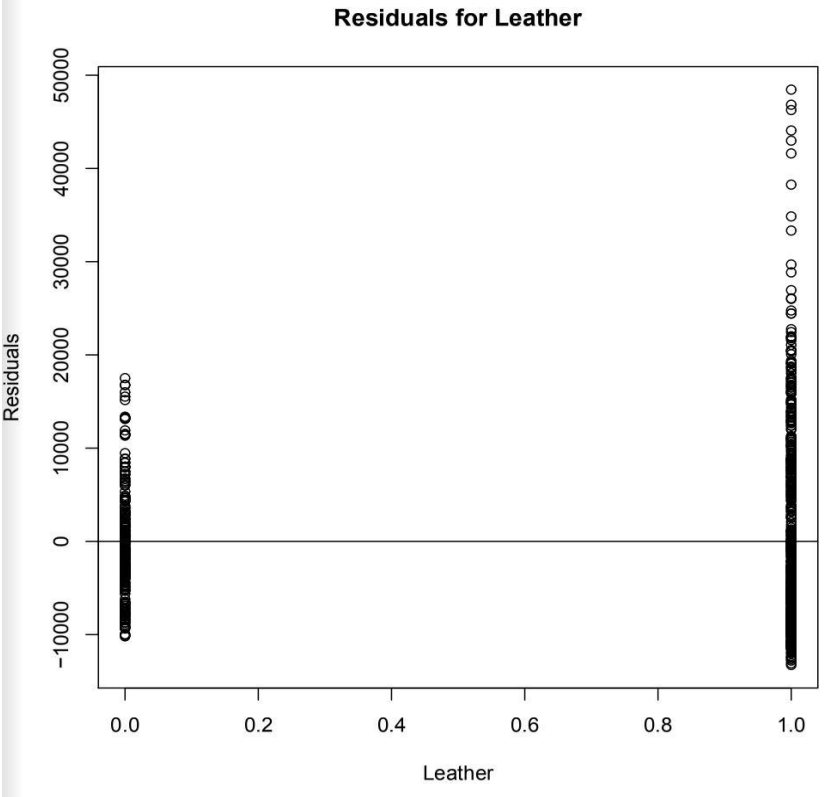


Analyzing this plot, we can see that as Mileage increases the size of residuals appears to increase with more residuals appearing with greater values. However, after drawing a vertical line corresponding to Mileage equal to 8000, we realized that the points in the plot of the residuals versus mileage are in fact balanced around the line Y=0, which indicates a right skewness pattern.

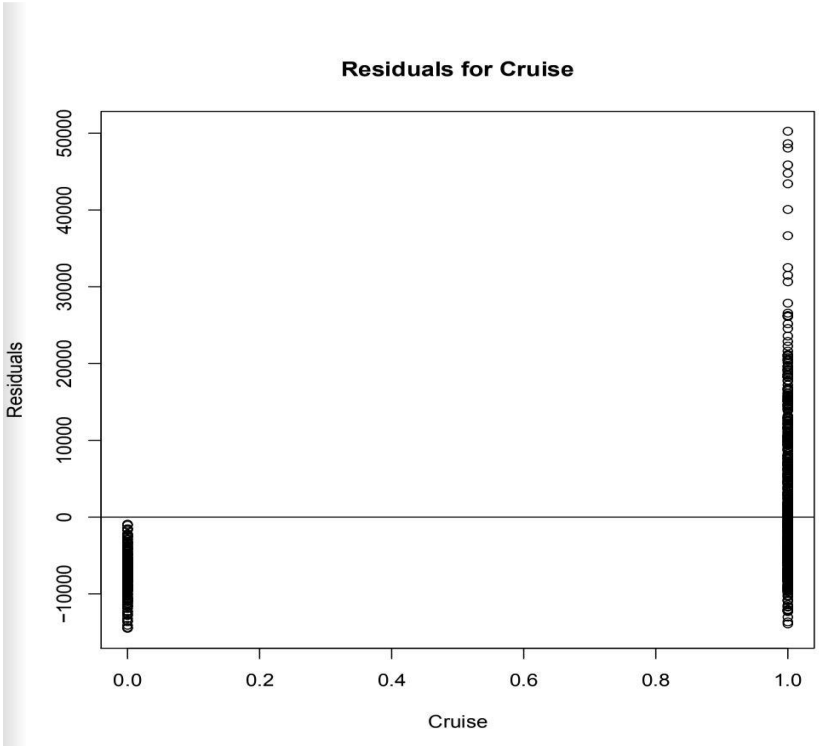
Residuals for Cylinder



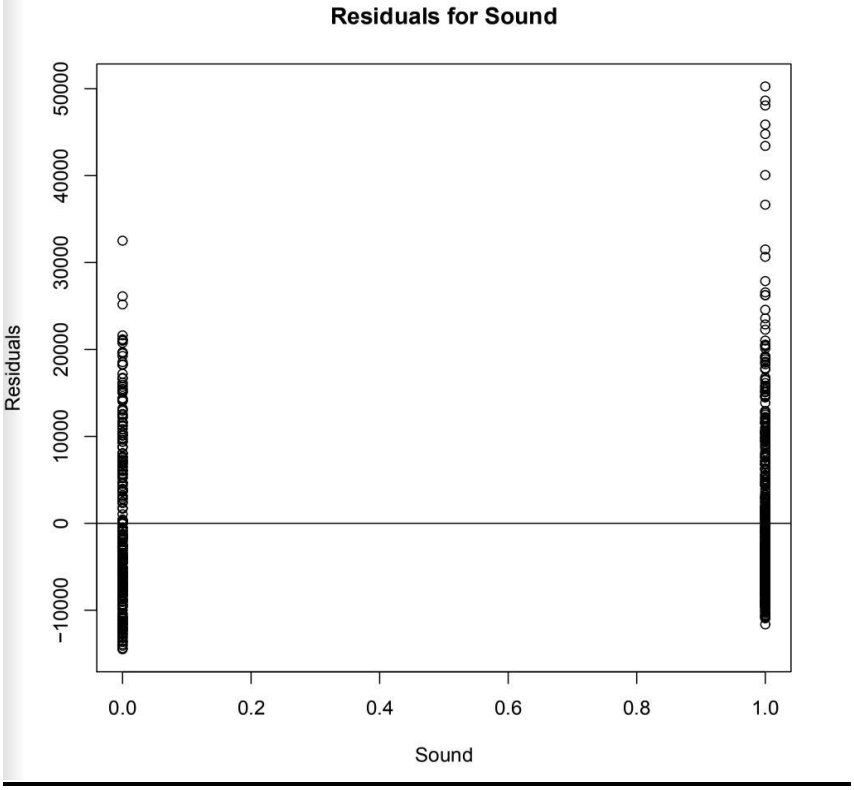
Residuals vs Leather Plot



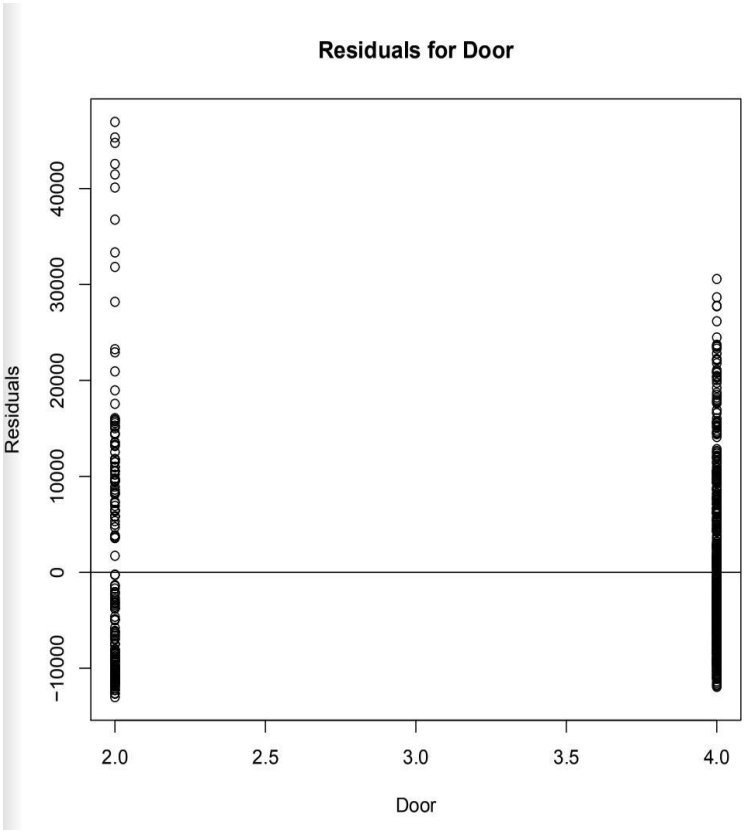
Residuals vs Cruise Plot



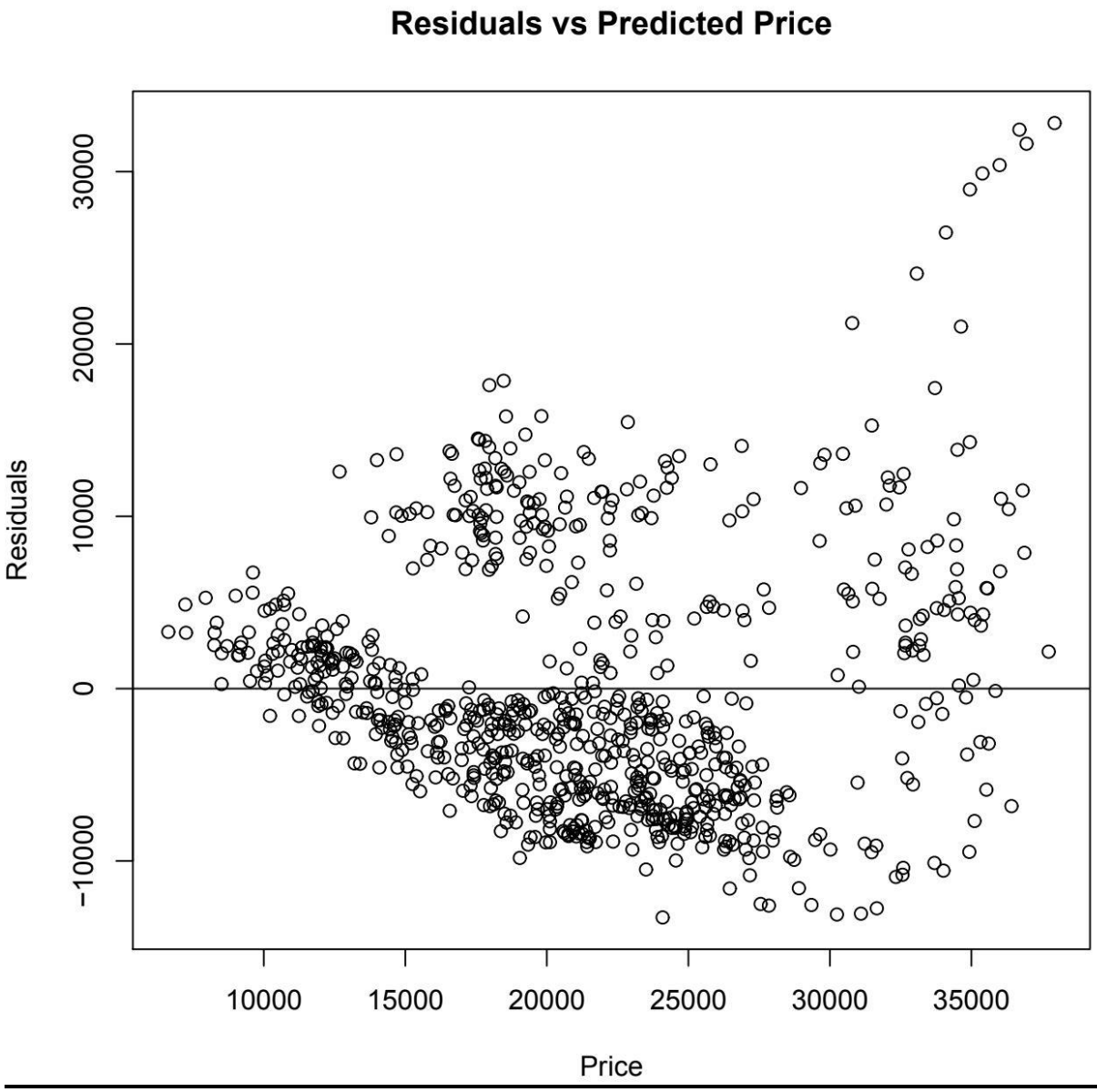
Residuals vs Sound Plot



Residuals for Door Plot



Residuals vs Predicted Price



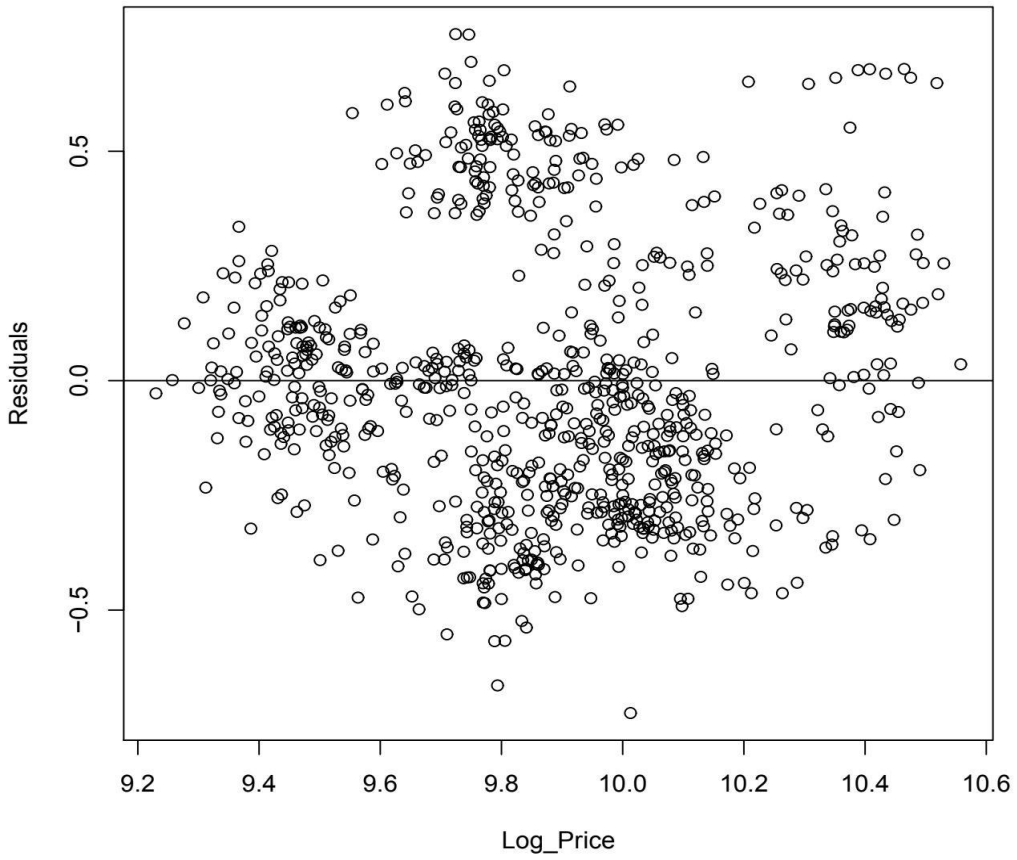
According to our observations in the plot of residuals versus fit plot, the size of the residuals seems to get larger as the prediction moves from small to large. There is a clear heteroskedasticity pattern that resembles a multiplicative error.

When Analyzing all the residual plot the patterns seen in the other graphs are those that represent homoscedasticity. Where the points outside of the outliers and the variance around the regression line is the same for all values of the predictor variable. The variance between the residuals and the explanatory variables is constant for all graphs.

Transformation Summary for Log Price function

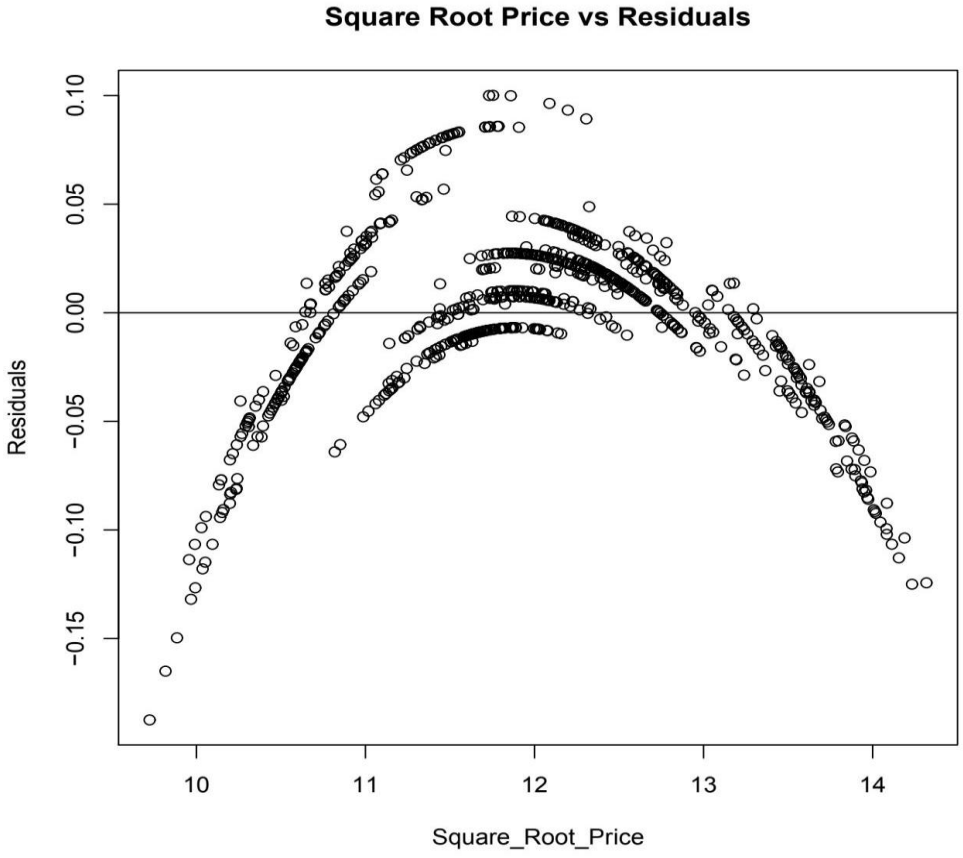
```
Call:
lm(formula = log(Price) ~ Mileage + Cylinder + Doors + Cruise +
    Sound + Leather)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72439 -0.23323 -0.03214  0.17048  0.75531

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.126e+00  5.800e-02 157.345  < 2e-16 ***
Mileage     -7.382e-06  1.276e-06  -5.786 1.03e-08 ***
Cylinder     1.302e-01  8.128e-03  16.018  < 2e-16 ***
Doors4      -7.424e-02  2.469e-02  -3.007 0.002723 **
Cruise1      3.208e-01  2.609e-02  12.298  < 2e-16 ***
Sound1      -8.720e-02  2.285e-02  -3.816 0.000146 ***
Leather1     1.214e-01  2.391e-02   5.078 4.75e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2958 on 797 degrees of freedom
Multiple R-squared:  0.4836,    Adjusted R-squared:  0.4797
F-statistic: 124.4 on 6 and 797 DF,  p-value: < 2.2e-16
```

Log Price vs Residuals



Transformation Summary for Square Root Price

```
Call:
lm(formula = sqrt(Price) ~ Mileage + Cylinder + Doors + Cruise +
    Sound + Leather)

Residuals:
   Min     1Q Median     3Q    Max
-47.82 -17.83  -3.76  13.94  74.75

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.739e+01  4.446e+00  19.659  < 2e-16 ***
Mileage     -5.449e-04  9.778e-05  -5.572 3.44e-08 ***
Cylinder     9.958e+00  6.230e-01  15.983  < 2e-16 ***
Doors4      -7.358e+00  1.892e+00  -3.888 0.000109 ***
Cruise1      2.215e+01  1.999e+00  11.080  < 2e-16 ***
Sound1      -6.660e+00  1.752e+00  -3.802 0.000154 ***
Leather1     9.934e+00  1.833e+00   5.421 7.87e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.67 on 797 degrees of freedom
Multiple R-squared:  0.4689,    Adjusted R-squared:  0.4649
F-statistic: 117.3 on 6 and 797 DF,  p-value: < 2.2e-16
```



Analysis of Transformation Summary

The Log Function of Price does the best job of reducing the heteroskedasticity and removing the residuals that most resemble a multiplicative error that the residuals vs price gave us. The R-Sq value of Log function is 48.36%. While the R-Sq value of the square root of Price is 46.89%.
In this case the best residual plot corresponds with the R-Sq value. In this case we can see that the Log function distributes the residual plots appropriately in a more random variety and the adjusted R-Sq offers us a higher predictive power at 48.36%.

Multicollinearity

Since we observed some outliers when looking at Price VS Mileage, we see that high end/sports cars have higher price. Even though the stepwise regression suggest that we eliminate Liter, we decided to check the multicollinearity by observing the Variance of Inflation Factors for the individual parameters. Looking at the figure below we can see the VIF for Cylinder and Liter are both greater than 10, which indicates multicollinearity.

```
> vif(fit)
 Cylinder     Liter     Doors    Cruise     Sound   Leather   Mileage
13.219830 13.518746  1.091984  1.187814  1.049451  1.051753  1.004130
```

We decided to look at the correlation between Liter and Cylinder, and we saw that they are highly correlated
(r =0.958).

```
> cor(Cylinder, Liter, method = "pearson", use = "complete.obs")
[1] 0.9578966
```

To be able to make a definitive decision of which variable to eliminate we decided to use three regression models: (1) Mileage and Liter, (2) Mileage and Cylinder, and (3) Mileage, Liter and Cylinder.

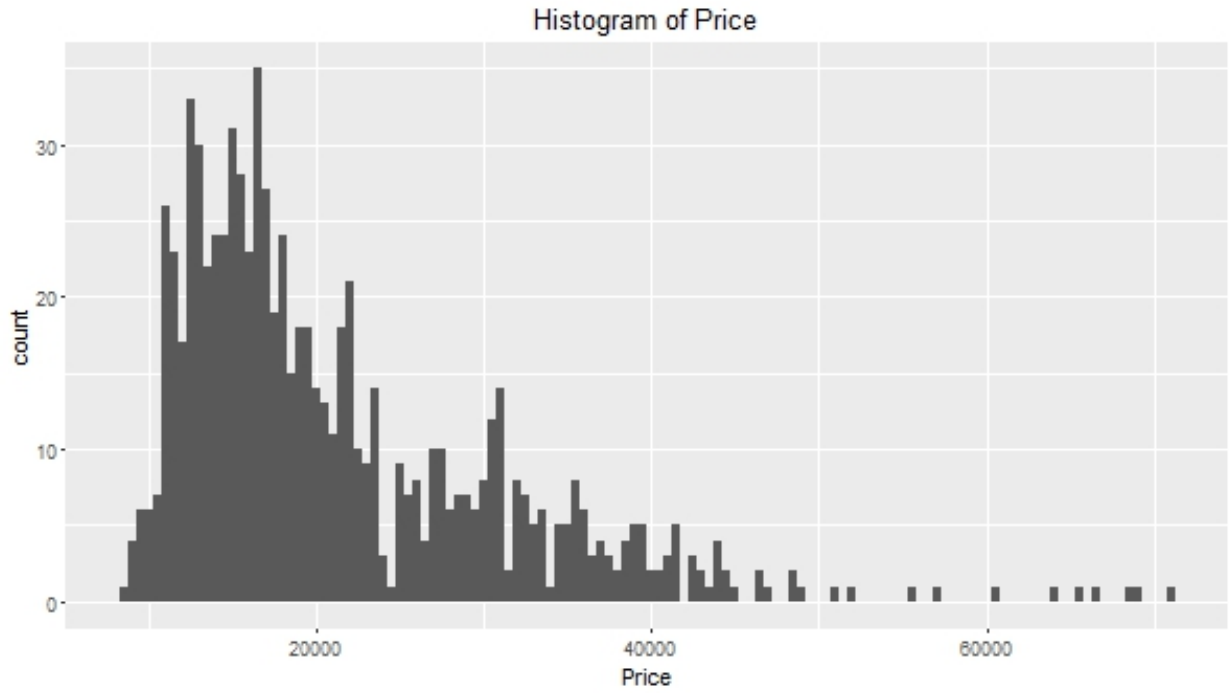
```
Call:
lm(formula = Price ~ Mileage + Liter)

Residuals:
   Min     1Q Median     3Q    Max
 -8817  -4990  -3337   3041  38568

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9426.60147 1095.07777   8.608  < 2e-16 ***
Mileage       -0.16003    0.03491  -4.584 5.28e-06 ***
Liter       4968.27812  258.80114  19.197  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8106 on 801 degrees of freedom
Multiple R-squared:  0.3291,   Adjusted R-squared:  0.3275
F-statistic: 196.5 on 2 and 801 DF,  p-value: < 2.2e-16


Call:
lm(formula = Price ~ Mileage + Cylinder)

Residuals:
   Min     1Q Median     3Q    Max
-10264  -5121  -2838   3102  35477

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3145.75032 1325.93436   2.372   0.0179 *
Mileage       -0.15243    0.03464  -4.401 1.22e-05 ***
Cylinder    4027.67463  204.61180  19.684  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8042 on 801 degrees of freedom
Multiple R-squared:  0.3398,   Adjusted R-squared:  0.3382
F-statistic: 206.2 on 2 and 801 DF,  p-value: < 2.2e-16


Call:
lm(formula = Price ~ Mileage + Liter + Cylinder)

Residuals:
   Min     1Q Median     3Q    Max
 -9552  -4905  -3026   2445  36246

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4707.61500 1602.86565   2.937  0.00341 **
Mileage       -0.15443    0.03461  -4.461 9.31e-06 ***
Liter       1545.25224  893.41064   1.730  0.08409 .
Cylinder    2847.93446  712.04020   4.000 6.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8032 on 800 degrees of freedom
Multiple R-squared:  0.3423,   Adjusted R-squared:  0.3398
F-statistic: 138.8 on 3 and 800 DF,  p-value: < 2.2e-16
```

```
> anova(fit15)
Analysis of Variance Table

Response: Price
           Df     Sum Sq    Mean Sq F value    Pr(>F)
Mileage     1 1.6056e+09 1.6056e+09  24.433 9.374e-07 ***
Liter       1 2.4218e+10 2.4218e+10 368.536 < 2.2e-16 ***
Residuals 801 5.2638e+10 6.5715e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(fit16)
Analysis of Variance Table

Response: Price
           Df     Sum Sq    Mean Sq F value    Pr(>F)
Mileage     1 1.6056e+09 1.6056e+09  24.828 7.681e-07 ***
Cylinder    1 2.5057e+10 2.5057e+10 387.478 < 2.2e-16 ***
Residuals 801 5.1799e+10 6.4667e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(fit17)
Analysis of Variance Table

Response: Price
           Df     Sum Sq    Mean Sq F value    Pr(>F)
Mileage     1 1.6056e+09 1.6056e+09  24.890 7.448e-07 ***
Liter       1 2.4218e+10 2.4218e+10 375.436 < 2.2e-16 ***
Cylinder    1 1.0319e+09 1.0319e+09  15.998 6.931e-05 ***
Residuals 800 5.1606e+10 6.4507e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

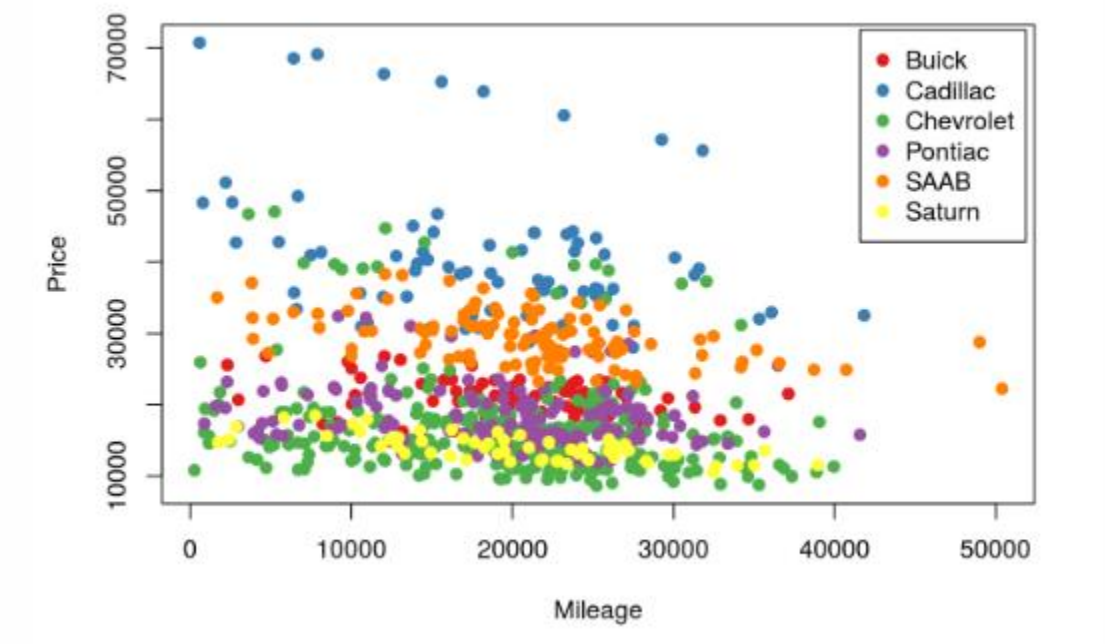
The R-Sq value and the t-tests for the regression coefficients show that Liter is significant in predicting Price in Model 1. Correspondingly, Cylinder is significant in Model 2. However, Model 3 only shows Liter as significant. We decide to look at both Cylinder and Liter when using additional variables into our equation.

Outliers and Influential Observations

Looking at the residual versus fit plot and residual versus mileage, it is quite clear that the outliers are the ones with a residual value of approximately 2.5 or higher
The Price variable histogram is right skewed and Price not normally distributed. We can say that, most of the cars has Price within the range of $10000 and $50000.

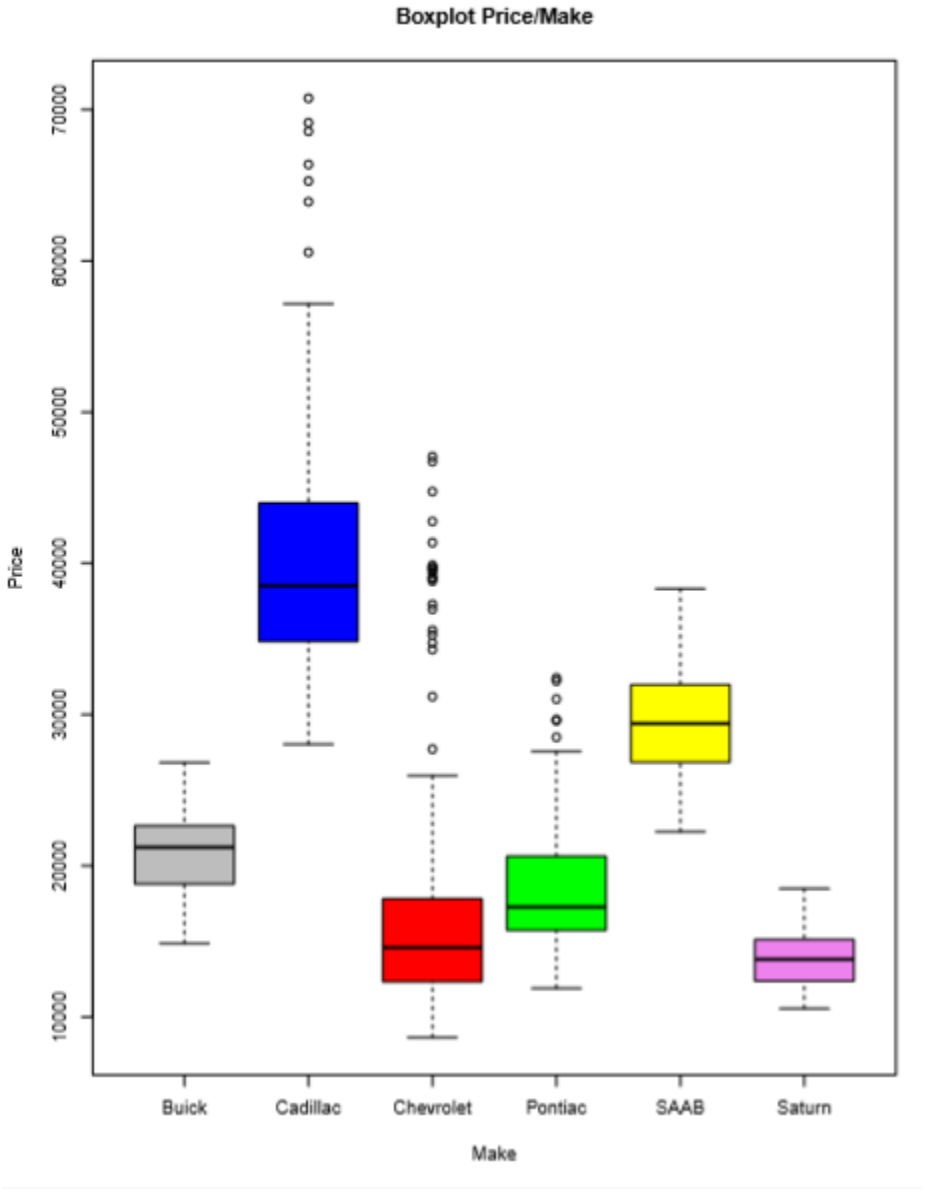


There is a lot of variation, including a set of high outliers. To explore that we start by looking at the Make of the cars. As shown below we used different colors to identify the cluster by the Make of the cars:



We can see now that there is a clear implication by Make. We can see Cadillac consistently near the top or Saturn consistently near the bottom. In this model we are confident that our residuals used earlier for the explanatory variables will not be helpful because in this case we are looking into the model of cars and ignoring the specifics of each model.
To explore further we look into the average price of the Make of the cars.

Group.1	x
Buick	20815.11
Cadillac	40936.34
Chevrolet	16427.60
Pontiac	18412.10
SAAB	29494.70
Saturn	13978.81



We can see that Cadillac has cars that are more expensive, and Saturn has cars that are less expensive.
We then find the Make of the cars that are more than $50,000 and take into account the Type of the car by looking at the price of cars that are more than $50,000 and make is Cadillac.

```
# A tibble: 11 x 1
   Make
   <chr>
 1 Cadillac
 2 Cadillac
 3 Cadillac
 4 Cadillac
 5 Cadillac
 6 Cadillac
 7 Cadillac
 8 Cadillac
 9 Cadillac
10 Cadillac
11 Cadillac
```

```
# A tibble: 11 x 1
   Type
   <chr>
 1 Sedan
 2 Convertible
 3 Convertible
 4 Convertible
 5 Convertible
 6 Convertible
 7 Convertible
 8 Convertible
 9 Convertible
10 Convertible
11 Convertible
```

We run the model with taking the variable “Make” into account

```
Call:
lm(formula = df$Price ~ Mileage + Make)

Residuals:
     Min       1Q   Median       3Q      Max
-11755.2  -3274.0   -701.8   1517.1  28174.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.431e+04  8.182e+02  29.705  < 2e-16 ***
Mileage       -1.709e-01  2.481e-02  -6.888 1.15e-11 ***
MakeCadillac   1.986e+04  9.093e+02  21.844  < 2e-16 ***
MakeChevrolet -4.520e+03  7.185e+02  -6.290 5.22e-10 ***
MakePontiac   -2.592e+03  7.959e+02  -3.257  0.00117 **
MakeSAAB       8.771e+03  8.381e+02  10.465  < 2e-16 ***
MakeSaturn    -6.852e+03  9.813e+02  -6.983 6.10e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5746 on 797 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6621
F-statistic: 263.3 on 6 and 797 DF,  p-value: < 2.2e-16
```

The R-Sq has improved substantially from 2% to 66%

Taking into account that both Liter and Cylinder are good prediction of price when measuring the engine size, we run the model with both Cylinder and Liter

```
Call:
lm(formula = df$Price ~ Mileage + Make + Cylinder + df$Door1 +
    df$Cruise1 + df$Sound1 + df$Leather1)

Residuals:
   Min     1Q Median     3Q    Max
-10430  -2089    -43   1767  20973

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.046e+03  1.079e+03   5.603 2.90e-08 ***
Mileage       -1.791e-01  1.594e-02 -11.234  < 2e-16 ***
MakeCadillac   1.345e+04  6.428e+02  20.921  < 2e-16 ***
MakeChevrolet -2.326e+03  5.153e+02  -4.514 7.34e-06 ***
MakePontiac   -2.061e+03  5.280e+02  -3.903 0.000103 ***
MakeSAAB       1.502e+04  6.266e+02  23.968  < 2e-16 ***
MakeSaturn    -2.073e+03  6.976e+02  -2.971 0.003057 **
Cylinder       3.741e+03  1.395e+02  26.822  < 2e-16 ***
df$Door1      -4.185e+03  3.211e+02 -13.030  < 2e-16 ***
df$Cruise1    -9.512e+01  3.668e+02  -0.259 0.795458
df$Sound1      7.339e+01  2.950e+02   0.249 0.803615
df$Leather1    4.904e+02  3.156e+02   1.554 0.120644
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3689 on 792 degrees of freedom
Multiple R-squared:  0.8627,    Adjusted R-squared:  0.8608
F-statistic: 452.3 on 11 and 792 DF,  p-value: < 2.2e-16
```

The R-Sq is 86% with Cylinder.

```
Call:
lm(formula = df$Price ~ Mileage + Make + Liter + df$Door1 + df$Cruise1 +
    df$Sound1 + df$Leather1)

Residuals:
    Min      1Q  Median      3Q     Max
-9559.8 -1907.3  -199.4  1348.1 22547.9

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.187e+04  8.665e+02  13.700  < 2e-16 ***
Mileage       -1.803e-01  1.496e-02 -12.051  < 2e-16 ***
MakeCadillac   1.610e+04  5.816e+02  27.689  < 2e-16 ***
MakeChevrolet -2.226e+03  4.832e+02  -4.606 4.79e-06 ***
MakePontiac   -1.787e+03  4.961e+02  -3.602 0.000335 ***
MakeSAAB       1.471e+04  5.760e+02  25.529  < 2e-16 ***
MakeSaturn    -2.271e+03  6.507e+02  -3.490 0.000510 ***
Liter          4.529e+03  1.490e+02  30.388  < 2e-16 ***
df$Door1      -3.452e+03  3.046e+02 -11.333  < 2e-16 ***
df$Cruise1    -5.053e+02  3.466e+02  -1.458 0.145255
df$Sound1     -3.386e+01  2.765e+02  -0.122 0.902559
df$Leather1    4.533e+01  2.973e+02   0.152 0.878852
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3462 on 792 degrees of freedom
Multiple R-squared:  0.879,     Adjusted R-squared:  0.8773
F-statistic:   523 on 11 and 792 DF,  p-value: < 2.2e-16
```

The R-Sq is at its highest with 88% with Liter

We then take Type into account and run the model again, this time without Cruise. We determined it is a potential outlier:

```
Call:
lm(formula = df$Price ~ Mileage + Make + Type + Liter + df$Door1 +
    df$Sound1 + df$Leather1)

Residuals:
    Min      1Q  Median      3Q     Max
-8599.5 -1364.9    44.5  1249.4 14589.5

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.013e+04  7.191e+02  27.993  < 2e-16 ***
Mileage       -1.847e-01  1.097e-02 -16.843  < 2e-16 ***
MakeCadillac   1.492e+04  4.284e+02  34.826  < 2e-16 ***
MakeChevrolet -1.616e+03  3.569e+02  -4.528 6.86e-06 ***
MakePontiac   -1.994e+03  3.680e+02  -5.419 7.98e-08 ***
MakeSAAB       1.097e+04  4.415e+02  24.855  < 2e-16 ***
MakeSaturn    -1.045e+03  4.702e+02  -2.223   0.0265 *
TypeCoupe     -1.172e+04  4.701e+02 -24.930  < 2e-16 ***
TypeHatchback -1.277e+04  5.463e+02 -23.386  < 2e-16 ***
TypeSedan     -1.227e+04  4.132e+02 -29.700  < 2e-16 ***
TypeWagon     -8.043e+03  5.132e+02 -15.671  < 2e-16 ***
Liter          4.463e+03  1.038e+02  42.975  < 2e-16 ***
df$Door1              NA         NA      NA       NA
df$Sound1      3.544e+02  2.043e+02   1.735   0.0832 .
df$Leather1    3.383e+02  2.177e+02   1.554   0.1206
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2537 on 790 degrees of freedom
Multiple R-squared:  0.9352,    Adjusted R-squared:  0.9341
F-statistic: 876.6 on 13 and 790 DF,  p-value: < 2.2e-16
```

R-Sq is at its highest, but we know that a good model is the simplest one.

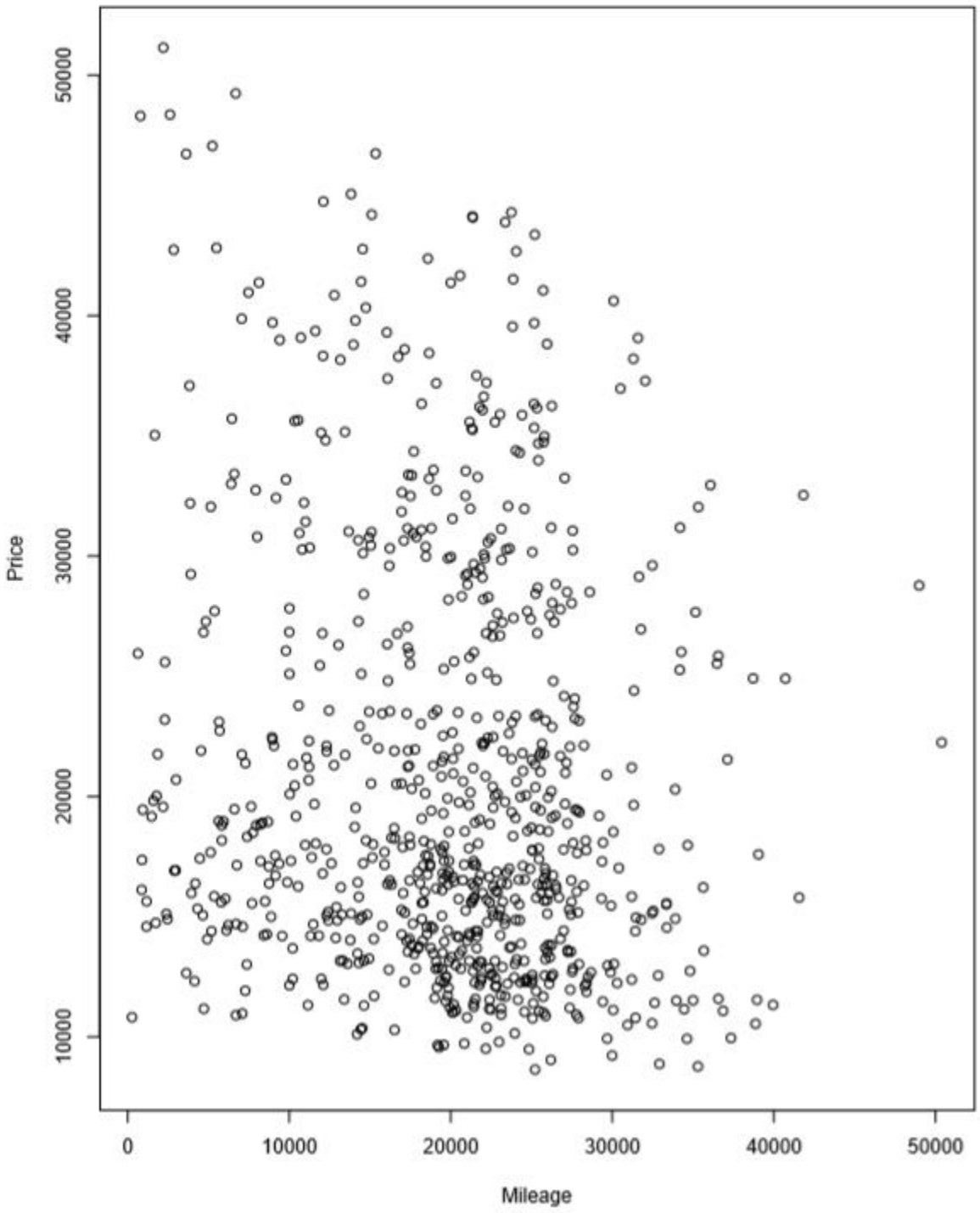
FINAL MODEL

After running the model with the individual Make and Types of the car and conducting a Step-Wise regression we arrive at the model below:

```
Call:
lm(formula = df$Price ~ Mileage + Liter + df$Door1 + df$Sound1 +
    df$Leather1 + df$Make_Buick + df$Make_Cadillac + df$Make_Chevrolet +
    df$Make_Pontiac + df$Make_SAAB + df$Type_Sedan + df$Type_Hatchback +
    df$Type_Convertible)

Residuals:
    Min      1Q  Median      3Q     Max
-8599.5 -1364.9    44.5  1249.4 14589.5

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.366e+03  5.263e+02  13.996  < 2e-16 ***
Mileage            -1.847e-01  1.097e-02 -16.843  < 2e-16 ***
Liter               4.463e+03  1.038e+02  42.975  < 2e-16 ***
df$Door1            3.677e+03  4.507e+02   8.159 1.33e-15 ***
df$Sound1           3.544e+02  2.043e+02   1.735   0.0832 .
df$Leather1         3.383e+02  2.177e+02   1.554   0.1206
df$Make_Buick       1.045e+03  4.702e+02   2.223   0.0265 *
df$Make_Cadillac    1.597e+04  5.057e+02  31.573  < 2e-16 ***
df$Make_Chevrolet  -5.709e+02  3.816e+02  -1.496   0.1350
df$Make_Pontiac    -9.490e+02  4.167e+02  -2.278   0.0230 *
df$Make_SAAB        1.202e+04  4.495e+02  26.738  < 2e-16 ***
df$Type_Sedan      -4.228e+03  3.862e+02 -10.947  < 2e-16 ***
df$Type_Hatchback  -4.732e+03  5.258e+02  -8.999  < 2e-16 ***
df$Type_Convertible 1.172e+04  4.701e+02  24.930  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2537 on 790 degrees of freedom
Multiple R-squared:  0.9352,    Adjusted R-squared:  0.9341
F-statistic: 876.6 on 13 and 790 DF,  p-value: < 2.2e-16
```

R-Sq is at 93.5%.

We remove the cluster of Cadillac Convertibles that we identified earlier and look at the relationship between Price and Mileage and run our model again:



```
Call:
lm(formula = Price ~ Mileage + Liter + Doors + Sound + Leather +
    df$Type_Sedan + df$Type_Hatchback + df$Type_Convertible +
    df$Make_Buick + df$Make_Chevrolet + df$Make_Cadillac + df$Make_Pontiac +
    df$Make_SAAB)

Residuals:
    Min      1Q  Median      3Q     Max
-7578.6 -1164.1    33.5  1179.1  7209.1

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.110e+03  4.332e+02  16.414  < 2e-16 ***
Mileage             -1.744e-01  9.136e-03 -19.093  < 2e-16 ***
Liter                4.515e+03  8.532e+01  52.916  < 2e-16 ***
Doors4               3.118e+03  3.712e+02   8.400  < 2e-16 ***
Sound1              -5.748e+01  1.694e+02  -0.339  0.73453
Leather1             3.438e+02  1.788e+02   1.923  0.05484 .
df$Type_Sedan       -3.507e+03  3.193e+02 -10.982  < 2e-16 ***
df$Type_Hatchback   -4.190e+03  4.327e+02  -9.684  < 2e-16 ***
df$Type_Convertible  8.346e+03  4.250e+02  19.638  < 2e-16 ***
df$Make_Buick        1.009e+03  3.861e+02   2.612  0.00916 **
df$Make_Chevrolet   -3.022e+02  3.137e+02  -0.963  0.33569
df$Make_Cadillac     1.439e+04  4.237e+02  33.952  < 2e-16 ***
df$Make_Pontiac     -8.026e+02  3.422e+02  -2.345  0.01927 *
df$Make_SAAB         1.317e+04  3.740e+02  35.201  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2084 on 780 degrees of freedom
Multiple R-squared:  0.9441,    Adjusted R-squared:  0.9432
F-statistic:  1013 on 13 and 780 DF,  p-value: < 2.2e-16
```

R-Sq value rose up to 94.32%

Running our model through AIC again we can see that we can remove Chevrolet, keeping R-Sq same.

```
Call:
lm(formula = Price ~ Mileage + Liter + Doors + Sound + Leather +
    df$Type_Sedan + df$Type_Hatchback + df$Type_Convertible +
    df$Make_Buick + df$Make_Cadillac + df$Make_Pontiac + df$Make_SAAB)

Residuals:
    Min      1Q  Median      3Q     Max
-7568.4 -1166.0    26.6  1192.2  7206.5

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.945e+03  3.981e+02  17.447  < 2e-16 ***
Mileage             -1.743e-01  9.134e-03 -19.080  < 2e-16 ***
Liter                4.501e+03  8.415e+01  53.490  < 2e-16 ***
Doors4               3.094e+03  3.704e+02   8.355 2.98e-16 ***
Sound1              -9.010e+01  1.660e+02  -0.543   0.5875
Leather1             3.206e+02  1.771e+02   1.810   0.0707 .
df$Type_Sedan       -3.488e+03  3.187e+02 -10.944  < 2e-16 ***
df$Type_Hatchback   -4.226e+03  4.310e+02  -9.805  < 2e-16 ***
df$Type_Convertible  8.324e+03  4.244e+02  19.616  < 2e-16 ***
df$Make_Buick        1.257e+03  2.880e+02   4.364 1.45e-05 ***
df$Make_Cadillac     1.465e+04  3.221e+02  45.484  < 2e-16 ***
df$Make_Pontiac     -5.540e+02  2.248e+02  -2.465   0.0139 *
df$Make_SAAB         1.341e+04  2.778e+02  48.257  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2083 on 781 degrees of freedom
Multiple R-squared:  0.944,     Adjusted R-squared:  0.9432
F-statistic:  1098 on 12 and 781 DF,  p-value: < 2.2e-16
```

Final model Log Price:

```
Call:
lm(formula = log(Price) ~ Mileage + Liter + Doors + Sound + Leather +
    df$Type_Sedan + df$Type_Hatchback + df$Type_Convertible +
    df$Make_Buick + df$Make_Chevrolet + df$Make_Cadillac + df$Make_Pontiac +
    df$Make_SAAB)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32892 -0.05887  0.00100  0.05980  0.28481

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.178e+00  1.871e-02 490.562  < 2e-16 ***
Mileage            -8.219e-06  3.946e-07 -20.829  < 2e-16 ***
Liter               2.237e-01  3.685e-03  60.705  < 2e-16 ***
Doors4              1.450e-01  1.603e-02   9.041  < 2e-16 ***
Sound1              7.448e-03  7.319e-03   1.018   0.3091
Leather1            1.290e-02  7.721e-03   1.670   0.0953 .
df$Type_Sedan      -1.535e-01  1.379e-02 -11.128  < 2e-16 ***
df$Type_Hatchback  -2.005e-01  1.869e-02 -10.726  < 2e-16 ***
df$Type_Convertible 2.781e-01  1.836e-02  15.150  < 2e-16 ***
df$Make_Buick       1.017e-01  1.668e-02   6.100 1.67e-09 ***
df$Make_Chevrolet  -2.529e-02  1.355e-02  -1.867   0.0623 .
df$Make_Cadillac    5.242e-01  1.830e-02  28.642  < 2e-16 ***
df$Make_Pontiac     4.087e-03  1.478e-02   0.276   0.7823
df$Make_SAAB        6.717e-01  1.615e-02  41.577  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08999 on 780 degrees of freedom
Multiple R-squared:  0.9479,    Adjusted R-squared:  0.947
F-statistic:  1091 on 13 and 780 DF,  p-value: < 2.2e-16
```

This is our final model and we have determined that in order to get the optimal model running log price with all other explanatory variables we determined as significant we can see that our R-Sq is 94.79%.

CONCLUSION

After performing all the regression analysis, we can conclude that; Mileage is not as significant as we expected to be to predict Price of the car and the most significant variable is Make of the car. When we look further we see that the Type of the car is also very significant, and among all Types, Cadillacs are the most significant but that falls under outliers. When we run our analysis with and without the cluster of outliers, we can see how it influence the coefficients in the regression line. By removing the cluster of outliers, we reach a more accurate model, and we use *log (Price)* transformation to get to our optimal model.