Project 1 – 6372 Stats 2

Andre Mauldin, Ricco Ferraro, Ana Alfaro

**Introduction**

We will analyze the relationship between vehicle characteristics, MSRP and the relevance of the popularity score that is calculated across platforms.  We will explore what vehicle characteristics influence MSRP and whether the popularity rating has an influence in the price of a vehicle.

Data Clean Up

Upon closer inspection, it was determined that the data required some clean-up and pre-processing before fitting the models. The data in appendix sections 1.2 and 1.3 illustrates how these values confound the true vehicle price.

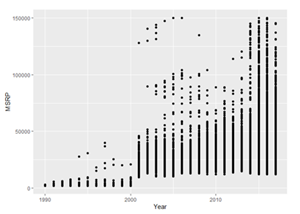
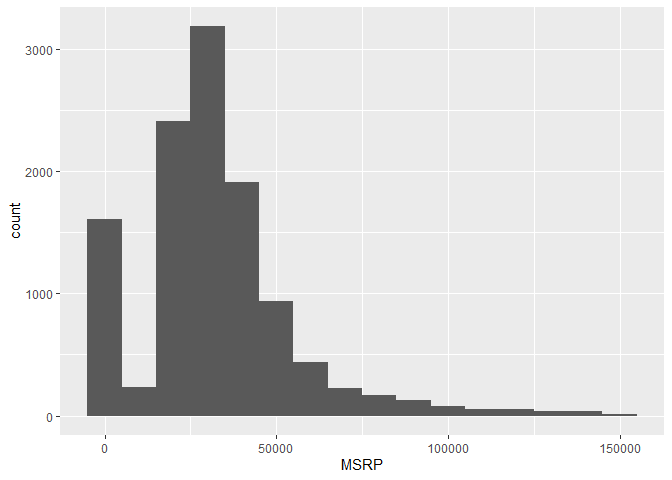
Categorical values

First, we examined the categorical attributes. Make and Model were excluded from the model due to the excess number of unique values.

Continuous values:

Plotting MSRP by year revealed a data quality issue, where vehicles manufactured prior to 1999 had a default value of $2000, which we believe would confound the true relationship of a vehicle’s price and the age of the car. We limited the data to cars manufactured after the year 2000.  A strong left skew was evident in the MSRP values, which is common in monetary values, as a result, we log transformed it for our analysis.

**MSRP by Year Distribution of MSRP Values**

Missing Data:

Engine HP had 69 missing values and Engine Cylinders had 30, we replaced them with the their median as determined by car size.

Engine Fuel Type and Transmission Type had blanks and ‘unknown’ values, we removed the 5 blank records, kept the category ‘unknown’ and excluded natural gas.

Outliers

There were two sets of outliers in the MSRP data. As seen in Appendix 1.6, we can observe there are car values for exceptionally expensive vehicles, we limited the vehicle data set to cars valued under $100,000

.

Data Types:

There were some inconsistencies in the numerical data types so we aligned them as doubles. Any attributes that were updated we appended the word Clean to the end of the column name.

Test / Train / Validation Data:

The data will be divided into Train, Test and Validation to ensure the models are not corrupted by the test results.

Popularity:

Evaluating the Popularity data in Appendix 1.7 it appears very bi-modal, we were not able to effectively delineate what drives the two distinct segments, there may be other attributes that are not considered in our dataset influencing this data, it could also be due to the collection method, source would be a good addition to this analysis.  Looking at more complex models, it was evident Popularity was influential as a interaction feature, this merits further exploration.

Evaluation after Data Prep:

Evaluating the data before and after the imputation and update of missing values, as seen in the summary statistics in Appendix 1.8, we can see that the population mean remained the about the same.

**Objective 1, Interpretative Model**

Aiming for interpretability, we modeled the linear relationship between various individual vehicle characteristics and MSRP.  To preserve interpretability of the model, we did not pursue interaction or quadratic terms.  We assessed the linear relationships, multicollinearity and verified the assumptions with MLR to determine which attributes were statistically significant.  We then applied a lasso regression to finalize the variable selection.  Appendix 2.5.

Collinearity

After preparing the data for modeling, we evaluate the relationship of the data by generating a correlation plot, We can observe that the correlations between Highway and City MPG, Engine HP and Number of Doors show signs of collinearity.



We verify our observations, with a first pass linear modeling of MSRP vs the cleaned subset of data, including the highly correlated attributes.

The pattern of the residuals on Appendix 2.1, appears random and the QQ plot looks normally distributed. After confirming our assumptions, we move on to look at the model results. The Variance Inflation Factor data for the regressor variables, , appendix 2.2, confirm the highly correlated attributes also carry a VIF greater than 2, so we will remove City MPG from the first pair and Number of Doors and reassess our model.

Lasso Regression

We ran a lasso regression, iterating to minimize lamda, and identified the variables that did not shrink to zero.  The highest coefficients were Cylinders, Fuel Type, Vehicle Style, Transmission, Driven Wheels and Vehicle Style, in that order.

Linear Model

Once we ran the lasso regression, we selected the attributes with coefficients that were not reduced to zero and applied them to our linear regression model for final assessment and interpretation. The resulting model was:



This model implies that the price of a vehicle depends on the Cylinders, the Engine Fuel Type the Vehicle Style and the Year a car was made.

*Median (MSRP | Engine Cylinders + Engine Fuel Type + Vehicle Style + Year) =*

*β\_0(Intercept)*

*+*

*β\_1(Engine.Cylinders.Clean4) + β\_2(Engine.Cylinders.Clean5) + β\_3(Engine.Cylinders.Clean6) + β\_4(Engine.Cylinders.Clean8 + β\_5(Engine.Cylinders.Clean10)+ β\_6(Engine.Cylinders.Clean12)*

 +

*β\_7(Engine.Fuel.Typeflex-fuel (premium unleaded recommended)+ β\_8(Engine.Fuel.Typeflex-fuel (premium unleaded required)+ β\_9(Engine.Fuel.Typeflex-fuel (unleaded/E85)) + β\_10(Engine.Fuel.Typeflex-fuel (unleaded/natural gas))+ β\_11(Engine.Fuel.Typepremium unleaded (recommended))+ β\_12(Engine.Fuel.Typepremium unleaded (required)) + β\_13(Engine.Fuel.Typeregular unleaded)*

+

*β\_14(Vehicle.Style2dr SUV) + β\_15(Vehicle.Style4dr Hatchback) + β\_16(Vehicle.Style4dr SUV) + β\_17(Vehicle.StyleCargo Minivan) + β\_18(Vehicle.StyleCargo Van) + β\_19(Vehicle.StyleConvertible) + β\_20(Vehicle.StyleConvertible SUV) + β\_21(Vehicle.StyleCoupe) + β\_22(Vehicle.StyleCrew Cab Pickup) + β\_23(Vehicle.StyleExtended Cab Pickup) + β\_24(Vehicle.StylePassenger Minivan) + β\_25(Vehicle.StylePassenger Van) + β\_26(Vehicle.StyleRegular Cab Pickup) + β\_27(Vehicle.StyleSedan) + β\_28(Vehicle.StyleWagon)*

+

*β\_29(Year)*

Interpretation 

**Engine Cylinders = 12**

With all other factors held constant, when a vehicle has 12 cylinders, with respect to our reference of 3 cylinders, the multiplicative effect on the Predicted Median MSRP = e^1.73 (with a 95% confidence interval between e^1.50 and e^1.97). This translates to a predicted median vehicle price increase of $5.64 with a 95% confidence interval of an increase of $4.48 and $7.17).

**Engine Cylinders = 6**

With all other factors held constant, when a vehicle has 6 cylinders, with respect to our reference of 3 cylinders, the multiplicative effect on the Predicted Median MSRP = e^.7598 (with a 95% confidence interval between e^.6582216 and e^0.86146459). This translates to a predicted median vehicle price increase of $2.14 with a 95% confidence interval of the increase being between $1.93 and $ 2.37).

For the rest of the parameter interpretations please refer to Appendix 2.6 for the list of

estimates and the 95% confidence intervals.

An Example of predicted log(MSRP) given a car with a 12 Cylinder Engine, Regular Unleaded, Convertible (which is a real car from our dataset) can be seen below.

*Predicted (log(MSRP) | 12 Cylinder Engine, Regular Unleaded, Convertible) =*

*1.73(12 Cylinder Engine) - .36 (Regular Unleaded) + .27 (Vehicle Style=Convertible) - 53*

**Objective 2: Multiple Linear Regression (Complex Model)**

With the goal of prediction and not interpretability, we fit a more complex multiple linear regression model to the car data mentioned above. For the complete list of variables included in the fully saturated model, please see Appendix section 3.2.

Data Prep, Variable selection Validation splits:

With regards to handling missing data and outliers (imputed or otherwise), we implemented the same approach as Objective 1 (described above). The same is true with how we split the training, test, and validation data (80%, 10%, 10% respectively).

Variable Selection (LASSO):

Log MSRP

Based on the analysis of the MSRP data in objective one, we continue to use log(MSRP) to conduct our analysis.

Fully Saturated Model

With the goal of prediction in mind, we implemented a constrained optimization technique called LASSO regression (L1 optimization). We did this against a fully saturated (order 2) multiple linear regression model. That is, we performed LASSO optimized regression on a model including:

* all squared terms, 2-way interactions, and independent variable estimates
* all levels of categorical explanatory variables, numerical explanatory variables

A complete list of all these features is in our “cleaned up” dataset and is described in Appendix 1.4.

Prior to running LASSO regression for our fully saturated model, we began with n=995 unique combinations of features. Again, this includes all levels of all variables in our cleaned-up dataset. The fully saturated model ends up looking something like:



Please see Appendix section 3.3 for the complete list of non-zero coefficients.

LASSO

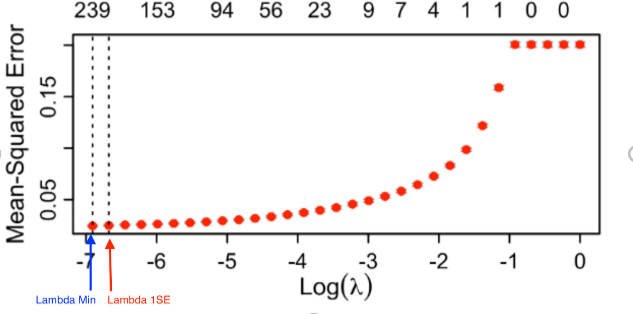
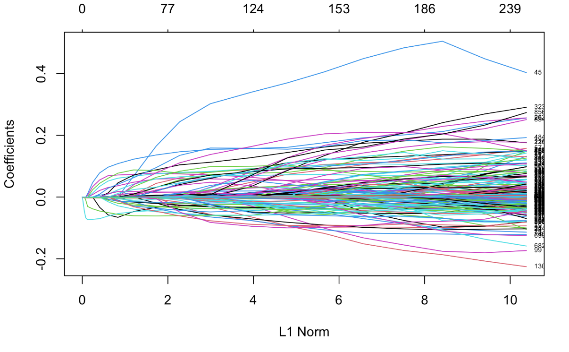
In order to parse out the relevant estimates (those which contribute the least to a L1-Penalty based error metric) from our fully saturated list of 995 features we performed LASSO L1-regularization on our full model. First, we scaled and centered our data set. Then we picked an optimal Lambda value. This is described below.

L1-Lambda

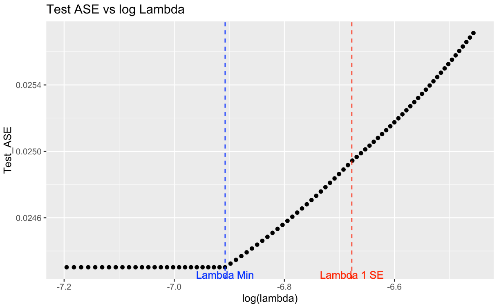
In order to determine the best nominal L1-weight to use for LASSO’s constraint multiplier (Lambda) we ran LASSO regression with a grid of potential lambdas against our training split of data. (See Appendix 3.5)

Note: “Lambda Min” represents the value of Lambda with the lowest associated MSE where  “Lambda 1SE” represents the highest value of Lambda which produces an MSE within 1 Standard Error of Lambda Min. In our case Lasso produces, Lambda Min = .001 and Lambda 1SE = .0012589.

The effect of increasing Lambda is that MSE increases for our training set.  Conversely, decreasing Lambda allows the coefficients of more features to diverge from zero and be considered significant by LASSO. The plots below illustrate these effects.

In order to check for overfitting, that is, whether it made the most sense to use Lambda Min vs Lamda 1SE with our LASSO model; we ran LASSO with a grid of lambdas against our test set. The results of this can be seen in the plot below.

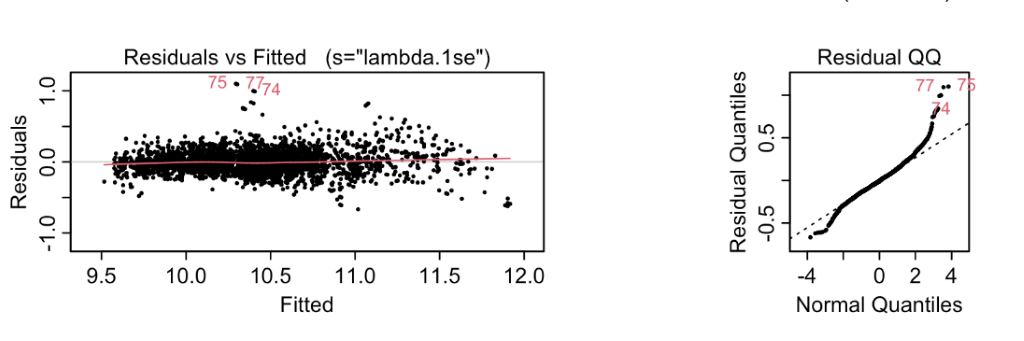


When validated against our test data, LASSO with Lambda Min continues to produce the lowest mean squared error.  This likely means that when we run LASSO using Lambda Min we are either, not overfitting badly, our test data is similar to our training data (due to large sample sizes and removal of outliers), or both. Further, the test MSE (f*or Log(MSRP)*) with Lambda min was **Test MSE = 0.02430239**. By comparison, our train MSE (*for Log(MSRP)*) with Lambda min was: **Train MSE = .023330728.**As such we will use Lambda Min for our LASSO regression.

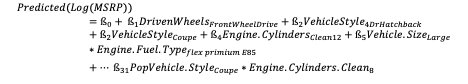
When this is done, our Validation MSE for Log(MSRP) is **Validation MSE = 0.02381303**. Because our validation MSE is roughly the same as the Test and Train MSE’s, we can cautiously suggest that our complex model is not overfitting badly.

When using LASSO penalization with our fully saturated model and Lambda Min = .001, we got a reduced list of influential regressor variables that is 241 elements long. Appendix 3.2.

Although we won’t dive into interpretation of our complex model’s residuals, For convenience they are visible in the plots below and in Appendix 3.1



Although we won’t dive into an interpretation of each of the 241 “significant” regressors in our complex Multiple Linear Regression model, it is worth noting that if such analysis were done, we would start with a reduced version of our 241 regressor model where we only use variables with an estimate from LASSO above 0.01. This filtering process yields a list of 31 regressors. Appendix 3.2 Such a model would look something like:



**Objective 3 Non-Parametric Model: KNN**

**Intuition**

K-Nearest Neighbor or K-NN is an algorithm that is used to classify or regress data.  For regression, the intuition behind KNN is that nearest neighbors are used to predict response variables such as Log(MSRP). The points are deemed "nearest” are those that are considered to have the closest Euclidean distance from the input we are attempting to regress. If k = 3, this would be the 3 nearest points.  Both K-NN and Linear Regression can be used to solve problems where the output needed is a continuous variable.

Scaling and centering is important since K-NN uses Euclidian distance to establish the closest points on the plot.  If the data is not on the same scale, the nearest neighbors could be closer or further away than they should.

Evaluation of a K-NN model depends on the problem.  K-NN models can easily overfit.

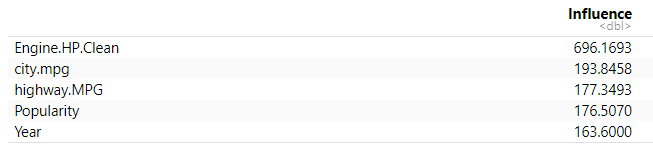
Variable Selection

Data Cleanup and removal of outliers the same way we did in Objective 1 (see above for more)

Further, we only used continuous explanatory variables in our KNN model. That is, no categorical variables were considered as part of any distance metric when performing KNN.

Next all continuous explanatory variables which were used were scaled and centered.

To select the best subset of variables for use in our KNN model we ran a random forest variable importance selection. The output of that selection was:

As can be seen, all continuous variables came back with relatively similar variable importance (besides Engine.HP.clean which is more important that the others).

Model:

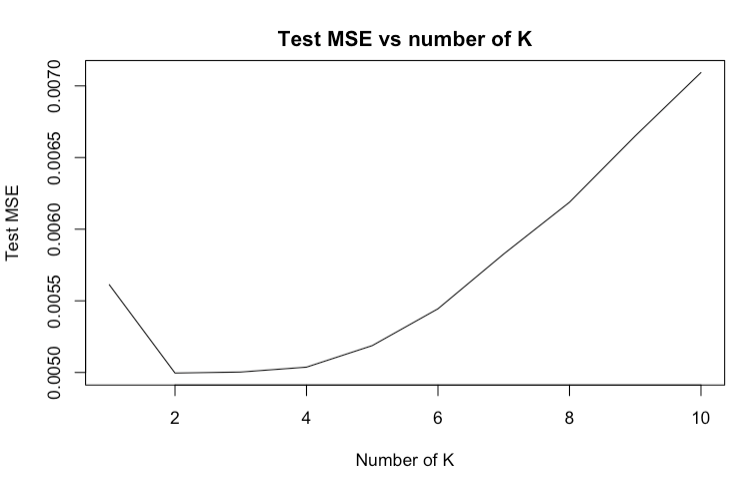
For the same reasons as noted in Objective 1, we are used Log(MSRP) for our response.

As such the Model we use in KNN ends up consisting of

|  |  |
| --- | --- |
| **Response** | **Explanatory Variables** |
| Log(MSRP) | Engine.HP.Clean, city.mpg, highway.MPG, Popularity, & Year |

Selection of Best K

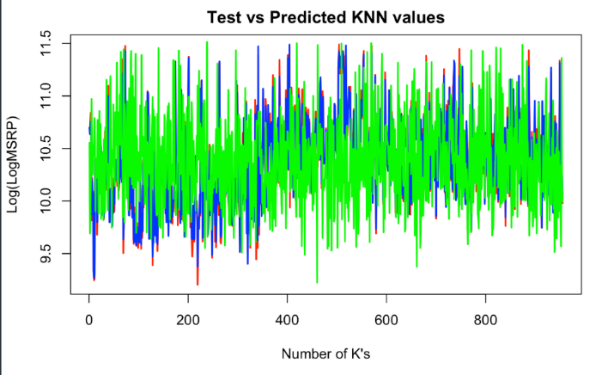
In order to select the best value of K to use in our KNN model, we ran an in iterative set of KNN fits, iterating on the value of K for our test dataset. The associated MSE values for each K are visible below for the Log(MSRP).



We ended up using K=2 because which had the lowest Test MSE of all values of K. Using K=2, we got a of **TestMSE = 0.005.**  The associated Test R-Squared values for the test set was 0.97055.  Validation MSE was 0.006.

Interpreting the output of KNN

Our KNN-2 model produced high R-Squared values and had small MSE for Log(MSRP) for both test and validation sets. This suggests that we have a good fit with KNN-2 and It is possible that we may be suffering from a small amount of overfitting as our validation MSE was very slightly higher.



Comparison of Model Results

|  |  |  |
| --- | --- | --- |
| **Model** | **Test MSE** | **Validation MSE** |
| **1 - Simple Multiple Linear Regression** | 0.0325 | 0.0329 |
| **2 - Complex Multiple Linear Regression - LASSO** | 0.023 | 0.024 |
| **3 - Non-Parametric: KNN-2** | 0.005 | 0.006 |

Our goal for model 1 (simple multiple linear regression) was interpretability. To make a comparison between model 1, model 2 (Complex MLR) and model 3 (Non-parametric KNN-2) which were strictly implemented and evaluated for the purpose of prediction, we must only look at Prediction metrics such as residuals, test and validation MSE metrics

When comparing the test and Validation MSE values for Model 1, Model 2 and Model 3, we see that the Non-parametric KNN-2 Model performs the best both on the test and validation split. It appears to outperform the parametric models with validation error at roughly 25% of our Complex model.

When comparing our complex model (model 2) to our Simple Multiple Linear Regression model (model 1) we see that the complex model performs with lower Test and Validation MSE.

If we had to choose a model, for the purposes of prediction, based on the results we have thus far, the KNN-2 model would be our choice.

It is possible that KNN-2 is mapping relationships between features in our data which are highly non-linear and that may be why it seems to perform the best.

**Summary**

Scope of Inference:

This is an observational study whose findings can be applied to the population of vehicles that were studied, we cannot extend this interpretation to the general population because we cannot verify the collection method and population subset from which this data was collected. As a result, we cannot make any causal inferences.

Objective 1:

Our interpretable model revealed that it was possible to predict vehicle price with a simple combination of the data attributes. Our model’s results feel intuitive, suggesting that a vehicle’s style and whether is powered by regular or unleaded fuel in addition to its number of cylinders impact the price of a car.

Objective 2:

Our Complex Linear Regression model highlights that for log transformed MSRP there is a linear relationship between MSRP (log transformed) and other explanatory variables in our model. It also highlights the fact that when we carefully include more variables from a fully saturated quadratic model, in final MLR model (by using LASSO regularization) we get improved prediction error and an improved overall prediction for Log-MSRP.

Our Non-Parametric model which uses a KNN-2 algorithm on Log-MSRP, we

Problems / Concerns:

There are some concerns with the quality of the data pre-2001, these values, if considered would confound the true price of the vehicle.

We observed in the fully saturated model that Popularity seems to be a valuable attribute when considered as an interaction term; however, the lack of domain knowledge regarding its source , scale caused by its collection method complicates this analysis. Ideally, this data would be collected along with the name of its source to better segment and interpret the values.

If we had more time, we would tidy the various marketing categories concatenated in varying sequence in a single field, to further understand this feature and its impact to both vehicle price and popularity.

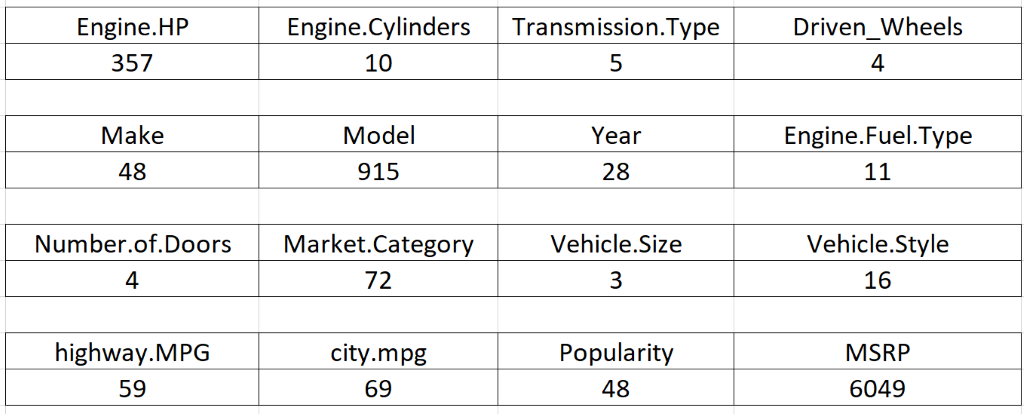
We would also take more care to assign a specific reference in our models, since the system auto-selected the first one, which in some cases were edge-case levels.

**Appendix**

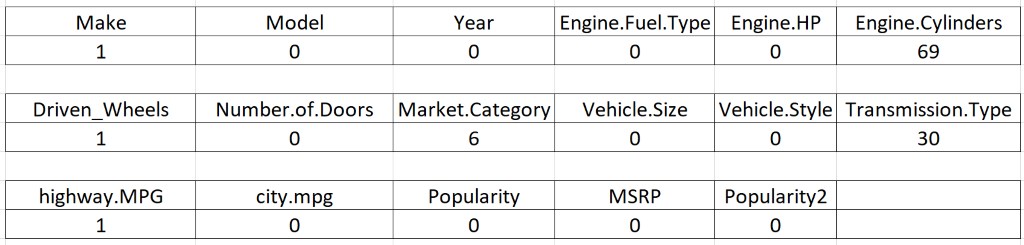
1. Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| MSRP | Numeric | The response variable |
| Car Make | Factor | The company that made the car. Ex: Honda, Toyota, etc. |
| Car Model | Factor | The model of the car. Ex: 4Runner, Accord, etc. |
| Year | Numeric | Year the car was produced |
| Engine Fuel Type | Factor | Type of fuel the car accepts.  Ex: Regular unleaded, Premium unleaded, Diesel |
| Engine HP | Numeric | Horsepower of the car’s engine. |
| Engine Cylinders | Numeric | Number of cylinders in the car’s engine. |
| Transmission Type | Factor | Type of transmission in the car. Usually manual or automatic, but there are a few specialty transmission types in the data. |
| Driven\_Wheels | Numeric | The wheels that are powered by the engine.  Ex: Front Wheel, Rear Wheel, Four Wheel Drive |
| Number of Doors | Numeric | The number of doors that the car has. Usually 2 or 4 |
| Market Category | Factor | Various special factors for each car.  Ex: Exotic, Luxury, High-Performance, Flex Fuel. Note: we created a new feature using Exotic/Not Exotic for our analysis |
| Vehicle Size | Factor | The size of the vehicle.  Ex: Midsize, Large, Compact |
| Vehicle Style | Factor | Body type of the vehicle. Ex: Coupe, Convertible, etc. |
| Highway MPG | Numeric | Fuel efficiency on the highway in MPG |
| City MPG | Numeric | Fuel efficiency in the city in MPG |
| Popularity | Numeric | A popularity score for each car.  The dataset does not detail how the popularity score is calculated. |

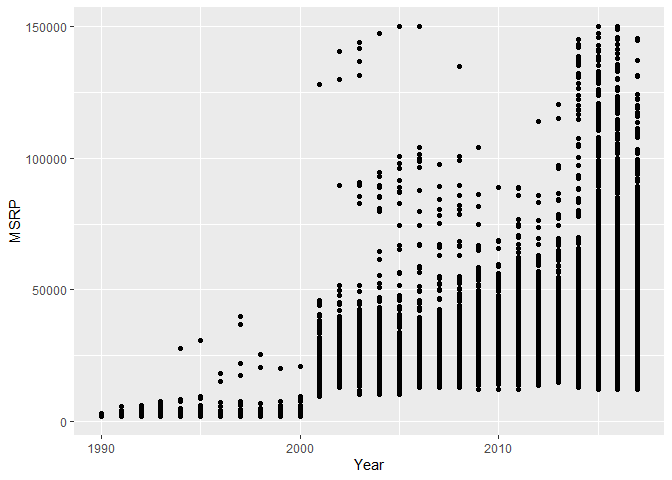
* 1. Unique values by attribute



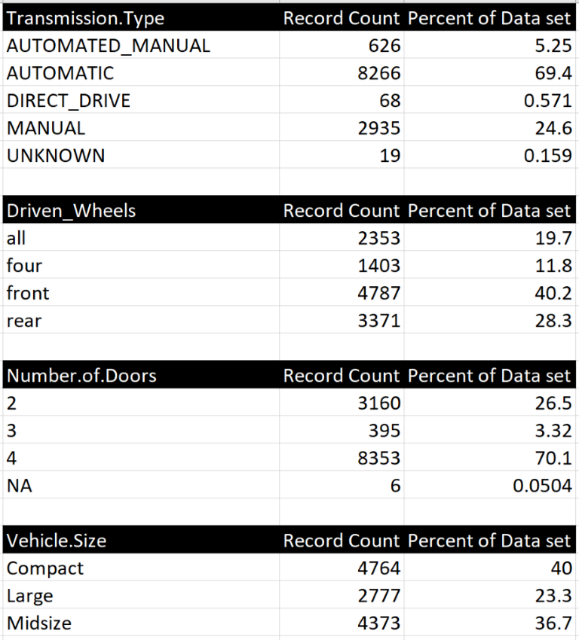
* 1. Count of N/As by column



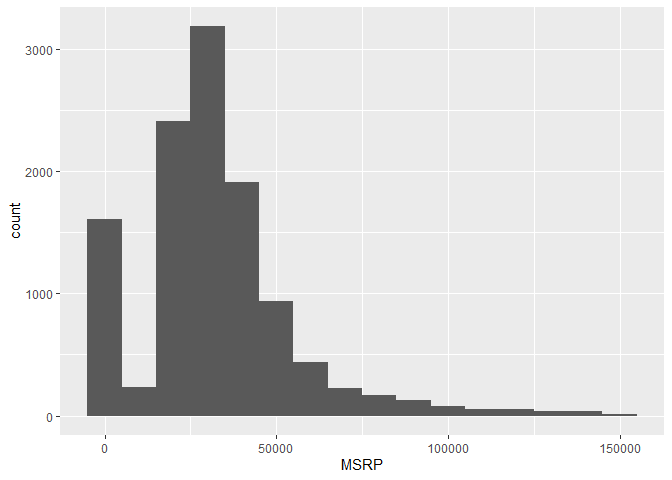
* 1. MSRP by Year



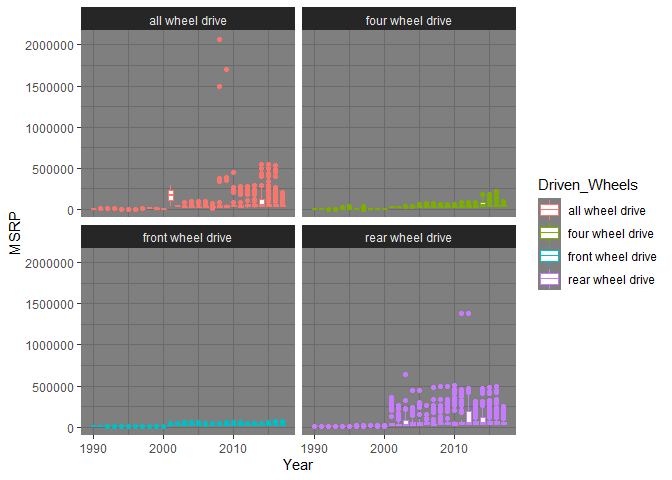
1.4 DISTRIBUTION OF DATA BY CATEGORY



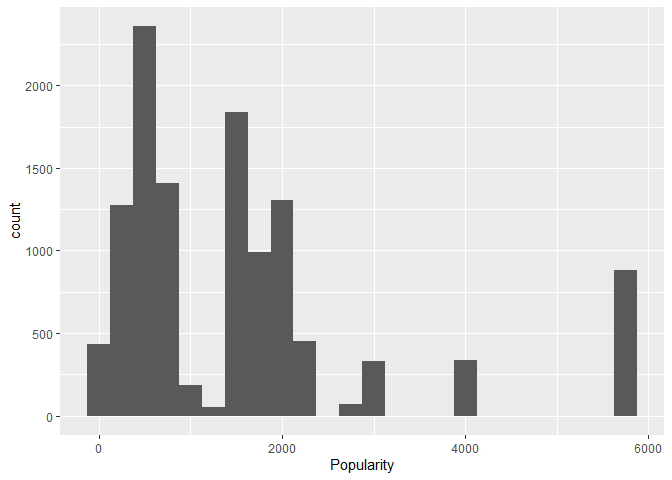
1.5 EDA: MSRP evaluation



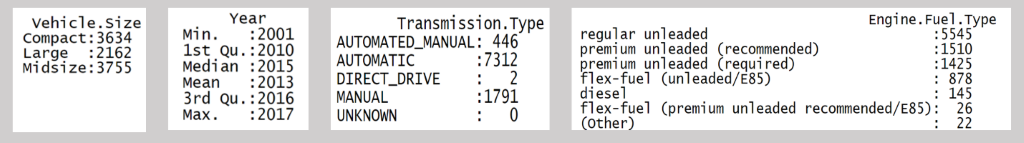
1.6 Outlier analysis

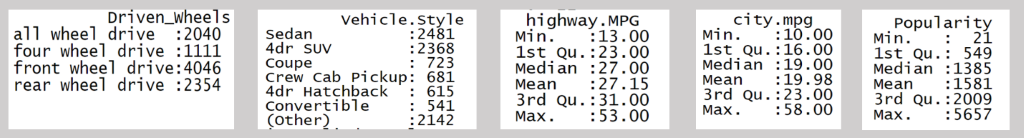


1.7 EDA: Popularity evaluation



1.8 Summary of Data Before and After Imputing (Data Prep Continued)

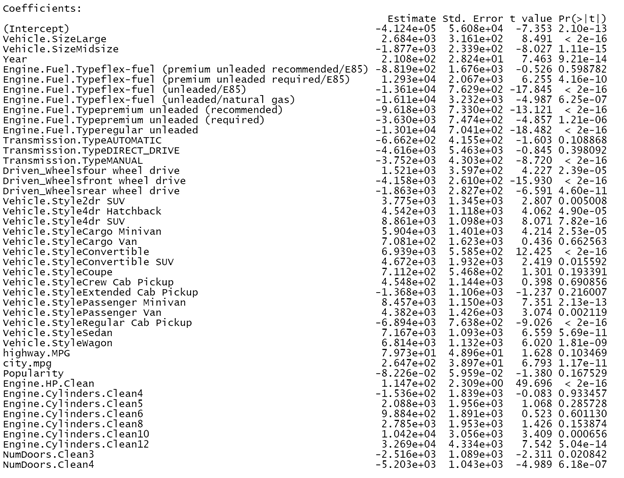




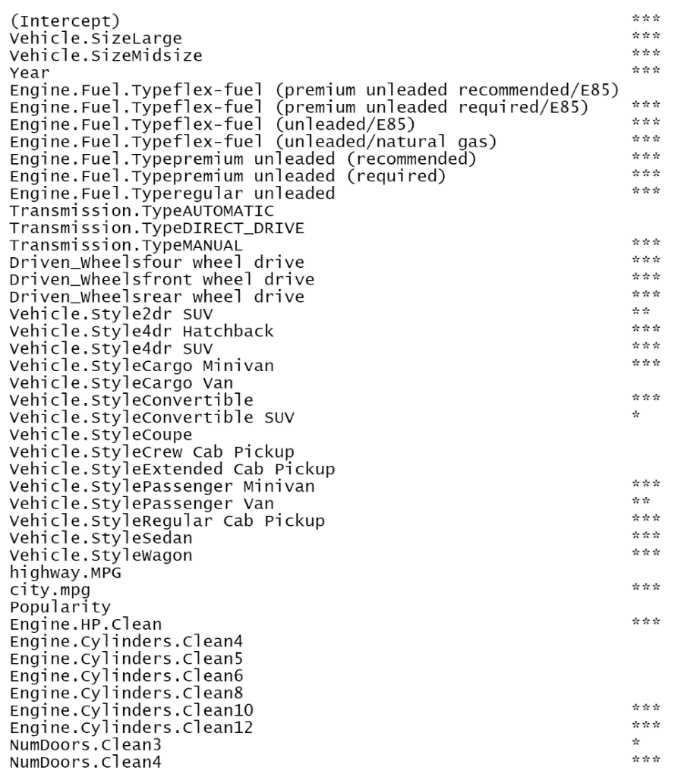


2.0 MSRP ~ ALL MODEL

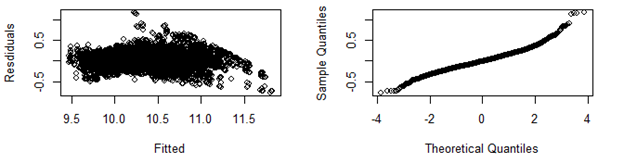
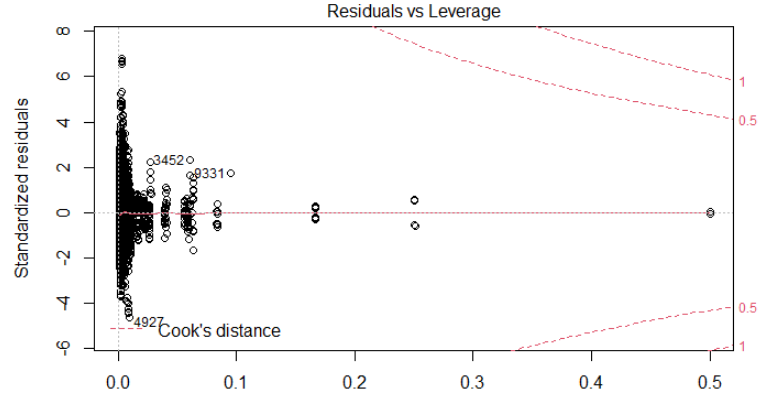




2.1 MSRP ~ ALL MODEL STATISTICALLY SIGNIFICANT



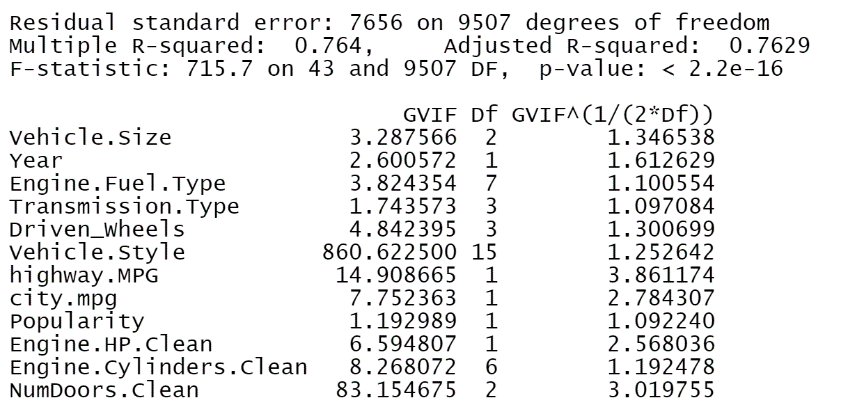
2.2 Verify assumptions

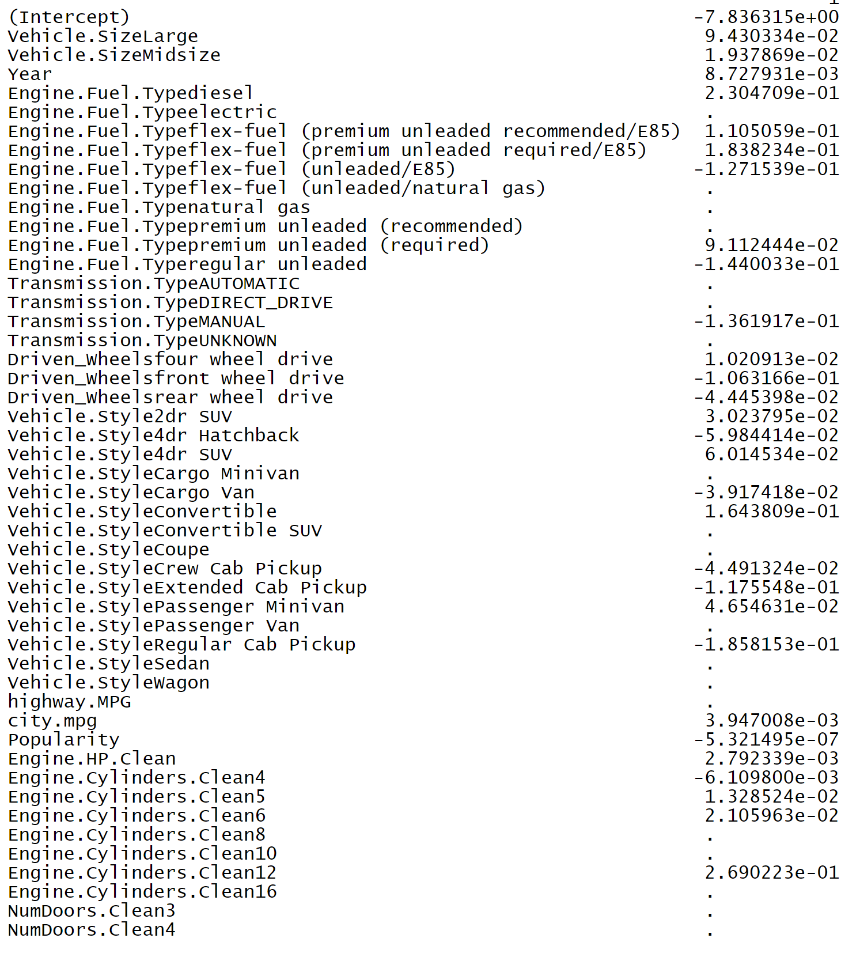
2.3 Correlation Plot



2.4 MSRP ~ . MODEL- VIF DATA



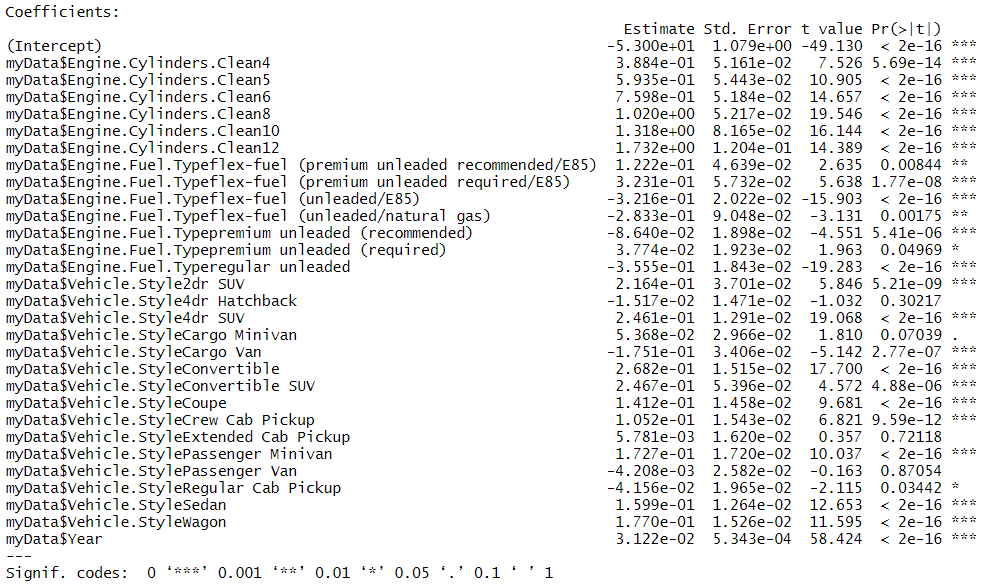
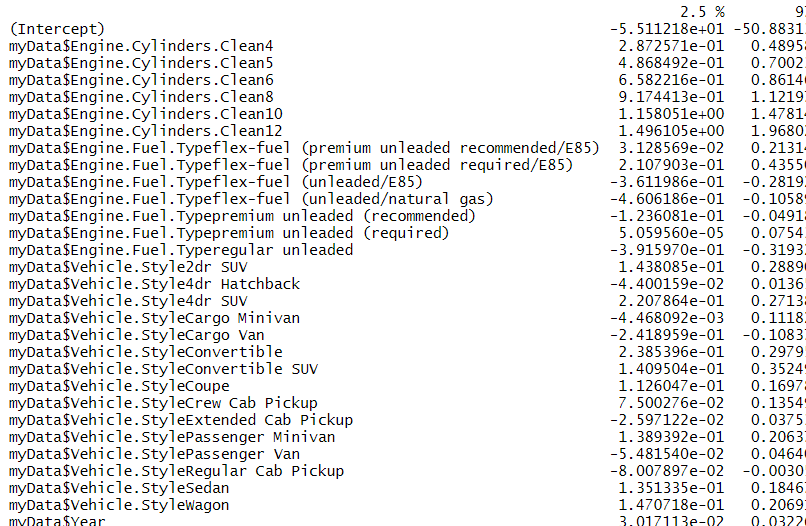
2.5 Lasso Regression All Data ~ Logged MSRP

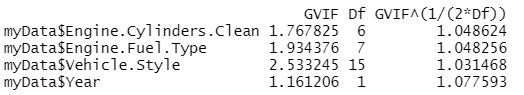


2.6 OBJECTIVE ONE ~ INTERPRETABLE MODEL

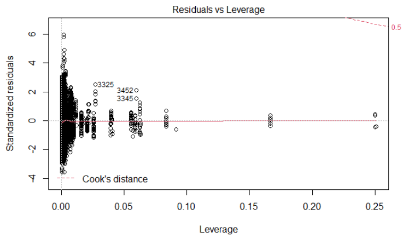
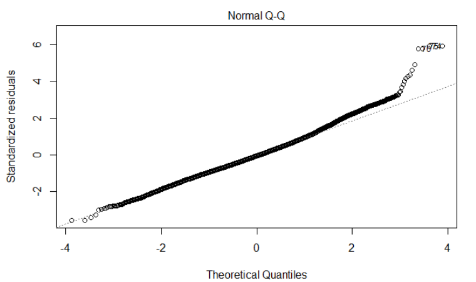
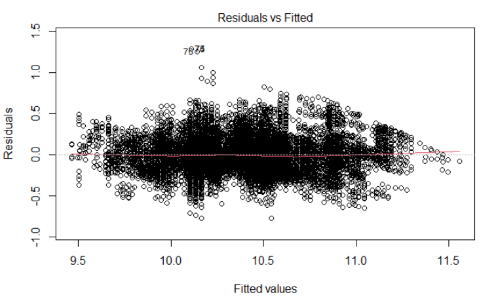




****

2.7 INTERPRETABLE MODEL - VIF

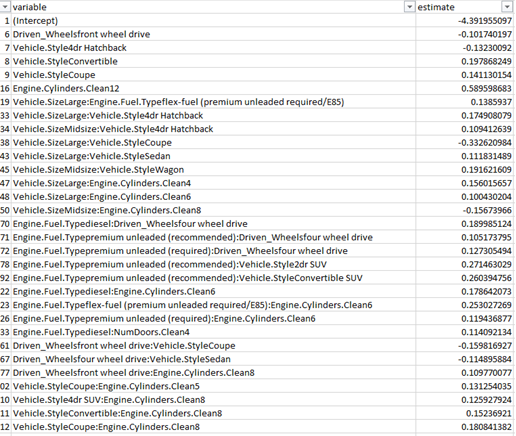
2.8 INTERPRETABLE MODEL – Diagnostic

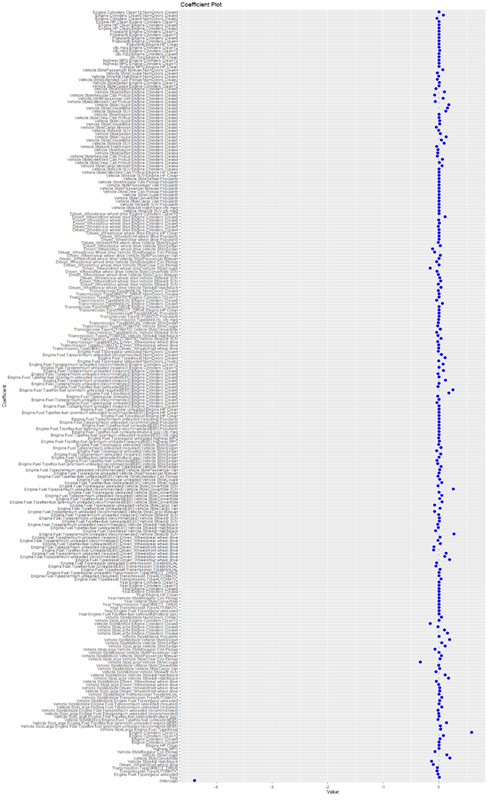


1. Fully Saturated Lasso Model top regressors

|  |
| --- |
| **Engine.Cylinders.Clean12** |
| Vehicle.Size\*Vehicle.Style |
| Engine.Fuel.Type\*Vehicle.Style |
| Engine.Fuel.Type\*Engine.Cylinders |
| **Vehicle.StyleConvertible** |
| Engine.Fuel.Type\*Driven\_Wheels |
| Vehicle.Style\*Engine.Cylinders |
| Driven\_Wheels\*Vehicle.Style |
| Vehicle.Size\*Engine.Cylinders |
| **Vehicle.StyleCoupe** |
| Vehicle.Size\*Engine.Fuel.Type |
| **Vehicle.Style4dr Hatchback** |
| Engine.Fuel.Type\*Num.Doors.Clean |
| Driven\_Wheels\*Engine.Cylinders |
| **Driven\_Wheelsfront wheel drive** |

3.1 Fully Saturated Lasso Model ~ coefficients > .01

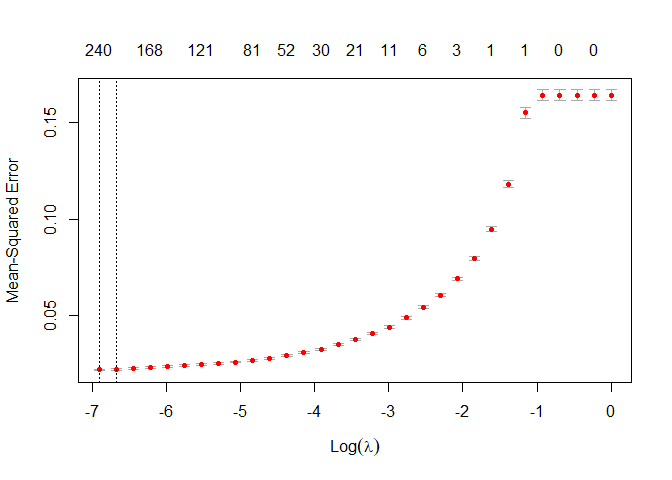


3.2 Fully Saturated Lasso Model ~ coefficient plot

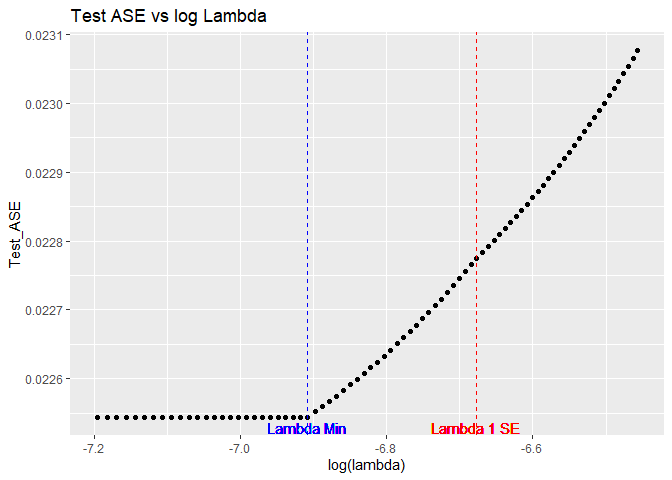
3.3 Fully Saturated Lasso Model ~ All non-zero coefficients

|  |  |
| --- | --- |
| variable | estimate |
| (Intercept) | -4.391955097 |
| Year | 0.007034003 |
| Engine.Fuel.Typeregular unleaded | -0.017248917 |
| Transmission.TypeAUTOMATIC | -0.03523848 |
| Transmission.TypeDIRECT\_DRIVE | -0.028531094 |
| Driven\_Wheelsfront wheel drive | -0.101740197 |
| Vehicle.Style4dr Hatchback | -0.13230092 |
| Vehicle.StyleConvertible | 0.197868249 |
| Vehicle.StyleCoupe | 0.141130154 |
| Vehicle.StyleRegular Cab Pickup | -0.074362395 |
| highway.MPG | -2.20E-05 |
| Engine.HP.Clean | 1.22E-06 |
| Engine.Cylinders.Clean4 | -0.00058933 |
| Engine.Cylinders.Clean8 | 0.019816785 |
| Engine.Cylinders.Clean10 | 0.082140387 |
| Engine.Cylinders.Clean12 | 0.589598683 |
| Vehicle.SizeLarge:Engine.Fuel.Typediesel | 0.031490198 |
| Vehicle.SizeLarge:Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85) | 0.028346107 |
| Vehicle.SizeLarge:Engine.Fuel.Typeflex-fuel (premium unleaded required/E85) | 0.1385937 |
| Vehicle.SizeMidsize:Engine.Fuel.Typeflex-fuel (unleaded/E85) | -0.017922758 |
| Vehicle.SizeLarge:Engine.Fuel.Typeflex-fuel (unleaded/natural gas) | -7.61E-05 |
| Vehicle.SizeLarge:Engine.Fuel.Typepremium unleaded (recommended) | -0.003444155 |
| Vehicle.SizeMidsize:Engine.Fuel.Typepremium unleaded (recommended) | -0.017937987 |
| Vehicle.SizeLarge:Engine.Fuel.Typepremium unleaded (required) | 0.048742274 |
| Vehicle.SizeMidsize:Engine.Fuel.Typepremium unleaded (required) | 0.003744238 |
| Vehicle.SizeMidsize:Engine.Fuel.Typeregular unleaded | 0.011715113 |
| Vehicle.SizeMidsize:Transmission.TypeAUTOMATIC | 0.052446078 |
| Vehicle.SizeMidsize:Transmission.TypeMANUAL | -0.000967092 |
| Vehicle.SizeLarge:Driven\_Wheelsfront wheel drive | 0.01137959 |
| Vehicle.SizeMidsize:Driven\_Wheelsfront wheel drive | 0.013494697 |
| Vehicle.SizeLarge:Driven\_Wheelsrear wheel drive | -0.00152986 |
| Vehicle.SizeMidsize:Driven\_Wheelsrear wheel drive | -0.046803993 |
| Vehicle.SizeLarge:Vehicle.Style4dr Hatchback | 0.174908079 |
| Vehicle.SizeMidsize:Vehicle.Style4dr Hatchback | 0.109412639 |
| Vehicle.SizeMidsize:Vehicle.Style4dr SUV | -0.05397705 |
| Vehicle.SizeMidsize:Vehicle.StyleCargo Van | 0.013103948 |
| Vehicle.SizeMidsize:Vehicle.StyleConvertible | -0.004648629 |
| Vehicle.SizeLarge:Vehicle.StyleCoupe | -0.332620984 |
| Vehicle.SizeLarge:Vehicle.StyleCrew Cab Pickup | 0.019977817 |
| Vehicle.SizeMidsize:Vehicle.StylePassenger Minivan | 0.044532795 |
| Vehicle.SizeMidsize:Vehicle.StylePassenger Van | 4.85E-05 |
| Vehicle.SizeLarge:Vehicle.StyleRegular Cab Pickup | -0.069566466 |
| Vehicle.SizeLarge:Vehicle.StyleSedan | 0.111831489 |
| Vehicle.SizeMidsize:Vehicle.StyleSedan | 0.051343185 |
| Vehicle.SizeMidsize:Vehicle.StyleWagon | 0.191621609 |
| Vehicle.SizeMidsize:Popularity | -2.60E-06 |
| Vehicle.SizeLarge:Engine.Cylinders.Clean4 | 0.156015657 |
| Vehicle.SizeLarge:Engine.Cylinders.Clean6 | 0.100430204 |
| Vehicle.SizeLarge:Engine.Cylinders.Clean8 | -0.015870005 |
| Vehicle.SizeMidsize:Engine.Cylinders.Clean8 | -0.15673966 |
| Vehicle.SizeLarge:Engine.Cylinders.Clean12 | 0.002637239 |
| Vehicle.SizeMidsize:NumDoors.Clean3 | 0.041972951 |
| Year:Engine.Fuel.Typeflex-fuel (unleaded/natural gas) | -1.20E-05 |
| Year:Engine.Fuel.Typeregular unleaded | -2.03E-05 |
| Year:Transmission.TypeAUTOMATIC | -1.44E-07 |
| Year:Transmission.TypeDIRECT\_DRIVE | -9.35E-08 |
| Year:Vehicle.StyleConvertible | 7.10E-09 |
| Year:Vehicle.StyleRegular Cab Pickup | -1.46E-07 |
| Year:Engine.HP.Clean | 7.19E-07 |
| Year:Engine.Cylinders.Clean4 | -4.55E-09 |
| Year:Engine.Cylinders.Clean8 | 4.56E-05 |
| Year:Engine.Cylinders.Clean10 | 2.60E-07 |
| Year:Engine.Cylinders.Clean12 | 7.20E-07 |
| Engine.Fuel.Typediesel:Transmission.TypeAUTOMATIC | 0.039828587 |
| Engine.Fuel.Typepremium unleaded (required):Transmission.TypeAUTOMATIC | -0.001259095 |
| Engine.Fuel.Typeregular unleaded:Transmission.TypeDIRECT\_DRIVE | -6.07E-05 |
| Engine.Fuel.Typediesel:Transmission.TypeMANUAL | 0.022437667 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Transmission.TypeMANUAL | -0.054985434 |
| Engine.Fuel.Typeregular unleaded:Transmission.TypeMANUAL | -0.038744363 |
| Engine.Fuel.Typediesel:Driven\_Wheelsfour wheel drive | 0.189985124 |
| Engine.Fuel.Typepremium unleaded (recommended):Driven\_Wheelsfour wheel drive | 0.105173795 |
| Engine.Fuel.Typepremium unleaded (required):Driven\_Wheelsfour wheel drive | 0.127305494 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Driven\_Wheelsfront wheel drive | -0.096557854 |
| Engine.Fuel.Typepremium unleaded (required):Driven\_Wheelsfront wheel drive | -0.089658617 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Driven\_Wheelsrear wheel drive | 0.024777979 |
| Engine.Fuel.Typepremium unleaded (recommended):Driven\_Wheelsrear wheel drive | -0.015403419 |
| Engine.Fuel.Typepremium unleaded (required):Driven\_Wheelsrear wheel drive | 0.021555128 |
| Engine.Fuel.Typepremium unleaded (recommended):Vehicle.Style2dr SUV | 0.271463029 |
| Engine.Fuel.Typediesel:Vehicle.Style4dr Hatchback | -0.059664164 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Vehicle.Style4dr Hatchback | -0.00854634 |
| Engine.Fuel.Typepremium unleaded (recommended):Vehicle.Style4dr Hatchback | -0.024783232 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Vehicle.Style4dr SUV | 0.005693584 |
| Engine.Fuel.Typepremium unleaded (recommended):Vehicle.Style4dr SUV | 0.006324302 |
| Engine.Fuel.Typepremium unleaded (required):Vehicle.Style4dr SUV | -0.00407387 |
| Engine.Fuel.Typepremium unleaded (recommended):Vehicle.StyleCargo Minivan | -0.080691453 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Vehicle.StyleCargo Van | 0.01747816 |
| Engine.Fuel.Typeregular unleaded:Vehicle.StyleCargo Van | -0.078228445 |
| Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85):Vehicle.StyleConvertible | 0.056495771 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Vehicle.StyleConvertible | 0.01410838 |
| Engine.Fuel.Typepremium unleaded (required):Vehicle.StyleConvertible | 0.04810462 |
| Engine.Fuel.Typeregular unleaded:Vehicle.StyleConvertible | 0.001892967 |
| Engine.Fuel.Typepremium unleaded (recommended):Vehicle.StyleConvertible SUV | 0.260394756 |
| Engine.Fuel.Typeregular unleaded:Vehicle.StyleConvertible SUV | -0.017341322 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Vehicle.StyleCoupe | 0.051392862 |
| Engine.Fuel.Typeregular unleaded:Vehicle.StyleCoupe | 0.013035414 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Vehicle.StyleExtended Cab Pickup | -0.019611536 |
| Engine.Fuel.Typeregular unleaded:Vehicle.StylePassenger Minivan | 0.004378092 |
| Engine.Fuel.Typepremium unleaded (recommended):Vehicle.StylePassenger Van | 0.016606774 |
| Engine.Fuel.Typediesel:Vehicle.StyleSedan | 0.039838189 |
| Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85):Vehicle.StyleSedan | -0.025273458 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Vehicle.StyleSedan | -0.037999824 |
| Engine.Fuel.Typeflex-fuel (unleaded/natural gas):Vehicle.StyleSedan | -6.79E-06 |
| Engine.Fuel.Typepremium unleaded (required):Vehicle.StyleSedan | 0.032229565 |
| Engine.Fuel.Typeregular unleaded:Vehicle.StyleSedan | -0.040765995 |
| Engine.Fuel.Typepremium unleaded (required):Vehicle.StyleWagon | 0.039258433 |
| Engine.Fuel.Typeregular unleaded:Vehicle.StyleWagon | -0.00294583 |
| Engine.Fuel.Typeflex-fuel (premium unleaded required/E85):highway.MPG | 0.004704078 |
| Engine.Fuel.Typeregular unleaded:highway.MPG | -0.00331805 |
| Engine.Fuel.Typeflex-fuel (premium unleaded required/E85):city.mpg | 9.93E-05 |
| Engine.Fuel.Typeflex-fuel (unleaded/natural gas):city.mpg | -5.65E-07 |
| Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85):Popularity | 7.43E-06 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Popularity | -1.93E-06 |
| Engine.Fuel.Typepremium unleaded (recommended):Popularity | -2.75E-06 |
| Engine.Fuel.Typepremium unleaded (required):Popularity | 2.13E-05 |
| Engine.Fuel.Typediesel:Engine.HP.Clean | 0.000196872 |
| Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85):Engine.HP.Clean | 0.000702246 |
| Engine.Fuel.Typeregular unleaded:Engine.HP.Clean | 0.000412489 |
| Engine.Fuel.Typepremium unleaded (required):Engine.Cylinders.Clean4 | -0.001395946 |
| Engine.Fuel.Typeregular unleaded:Engine.Cylinders.Clean4 | -0.032267514 |
| Engine.Fuel.Typepremium unleaded (recommended):Engine.Cylinders.Clean5 | 0.003588853 |
| Engine.Fuel.Typeregular unleaded:Engine.Cylinders.Clean5 | -0.021537774 |
| Engine.Fuel.Typediesel:Engine.Cylinders.Clean6 | 0.178642073 |
| Engine.Fuel.Typeflex-fuel (premium unleaded required/E85):Engine.Cylinders.Clean6 | 0.253027269 |
| Engine.Fuel.Typeflex-fuel (unleaded/E85):Engine.Cylinders.Clean6 | -0.014610055 |
| Engine.Fuel.Typepremium unleaded (recommended):Engine.Cylinders.Clean6 | 6.52E-05 |
| Engine.Fuel.Typepremium unleaded (required):Engine.Cylinders.Clean6 | 0.119436877 |
| Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85):Engine.Cylinders.Clean8 | 0.002297051 |
| Engine.Fuel.Typepremium unleaded (recommended):Engine.Cylinders.Clean8 | -0.017290585 |
| Engine.Fuel.Typepremium unleaded (required):Engine.Cylinders.Clean8 | 0.083114088 |
| Engine.Fuel.Typepremium unleaded (required):Engine.Cylinders.Clean10 | 0.00292684 |
| Engine.Fuel.Typepremium unleaded (recommended):Engine.Cylinders.Clean12 | 0.077320107 |
| Engine.Fuel.Typeregular unleaded:NumDoors.Clean3 | -0.002593543 |
| Engine.Fuel.Typediesel:NumDoors.Clean4 | 0.114092134 |
| Engine.Fuel.Typepremium unleaded (recommended):NumDoors.Clean4 | 0.042619385 |
| Engine.Fuel.Typeregular unleaded:NumDoors.Clean4 | -0.012719156 |
| Transmission.TypeDIRECT\_DRIVE:Driven\_Wheelsfront wheel drive | -0.001666998 |
| Transmission.TypeAUTOMATIC:Driven\_Wheelsrear wheel drive | -0.002198988 |
| Transmission.TypeMANUAL:Driven\_Wheelsrear wheel drive | -0.021070564 |
| Transmission.TypeAUTOMATIC:Vehicle.Style2dr SUV | 0.039999161 |
| Transmission.TypeAUTOMATIC:Vehicle.Style4dr Hatchback | -0.015881149 |
| Transmission.TypeMANUAL:Vehicle.Style4dr SUV | -0.039465504 |
| Transmission.TypeAUTOMATIC:Vehicle.StyleConvertible | 0.026283666 |
| Transmission.TypeAUTOMATIC:Vehicle.StyleCoupe | -0.024040832 |
| Transmission.TypeMANUAL:Vehicle.StyleSedan | -0.011850305 |
| Transmission.TypeMANUAL:city.mpg | -0.003901099 |
| Transmission.TypeAUTOMATIC:Popularity | -1.91E-06 |
| Transmission.TypeMANUAL:Popularity | -8.61E-06 |
| Transmission.TypeDIRECT\_DRIVE:Engine.HP.Clean | -7.66E-08 |
| Transmission.TypeDIRECT\_DRIVE:Engine.Cylinders.Clean4 | -2.89E-14 |
| Transmission.TypeMANUAL:Engine.Cylinders.Clean6 | 0.06456897 |
| Transmission.TypeMANUAL:Engine.Cylinders.Clean8 | -0.043923233 |
| Transmission.TypeAUTOMATIC:Engine.Cylinders.Clean12 | 0.000236228 |
| Transmission.TypeDIRECT\_DRIVE:NumDoors.Clean4 | -7.75E-05 |
| Transmission.TypeMANUAL:NumDoors.Clean4 | -0.008489677 |
| Driven\_Wheelsrear wheel drive:Vehicle.Style4dr Hatchback | 0.040523716 |
| Driven\_Wheelsfour wheel drive:Vehicle.Style4dr SUV | -0.011677578 |
| Driven\_Wheelsfront wheel drive:Vehicle.Style4dr SUV | 0.031662719 |
| Driven\_Wheelsrear wheel drive:Vehicle.Style4dr SUV | 0.013880531 |
| Driven\_Wheelsrear wheel drive:Vehicle.StyleCargo Minivan | -0.01348054 |
| Driven\_Wheelsrear wheel drive:Vehicle.StyleConvertible SUV | -0.03497651 |
| Driven\_Wheelsfront wheel drive:Vehicle.StyleCoupe | -0.159816927 |
| Driven\_Wheelsfour wheel drive:Vehicle.StyleCrew Cab Pickup | 0.003017614 |
| Driven\_Wheelsrear wheel drive:Vehicle.StyleExtended Cab Pickup | -0.040381992 |
| Driven\_Wheelsfront wheel drive:Vehicle.StylePassenger Minivan | -2.51E-05 |
| Driven\_Wheelsrear wheel drive:Vehicle.StylePassenger Van | 0.023883448 |
| Driven\_Wheelsrear wheel drive:Vehicle.StyleRegular Cab Pickup | -0.075421619 |
| Driven\_Wheelsfour wheel drive:Vehicle.StyleSedan | -0.114895884 |
| Driven\_Wheelsrear wheel drive:Vehicle.StyleSedan | 0.028627411 |
| Driven\_Wheelsfront wheel drive:Vehicle.StyleWagon | -0.011529504 |
| Driven\_Wheelsfour wheel drive:Popularity | 1.17E-06 |
| Driven\_Wheelsfront wheel drive:Popularity | -4.63E-06 |
| Driven\_Wheelsrear wheel drive:Engine.HP.Clean | -0.000255614 |
| Driven\_Wheelsrear wheel drive:Engine.Cylinders.Clean4 | 0.003948212 |
| Driven\_Wheelsrear wheel drive:Engine.Cylinders.Clean5 | -0.011922296 |
| Driven\_Wheelsfour wheel drive:Engine.Cylinders.Clean6 | 0.0092135 |
| Driven\_Wheelsfour wheel drive:Engine.Cylinders.Clean8 | -0.003578737 |
| Driven\_Wheelsfront wheel drive:Engine.Cylinders.Clean8 | 0.109770077 |
| Driven\_Wheelsrear wheel drive:Engine.Cylinders.Clean10 | 0.000171279 |
| Vehicle.Style2dr SUV:city.mpg | 0.001501678 |
| Vehicle.Style4dr Hatchback:city.mpg | 0.003583087 |
| Vehicle.Style4dr SUV:Popularity | -7.51E-06 |
| Vehicle.StyleCargo Van:Popularity | 7.33E-06 |
| Vehicle.StyleConvertible:Popularity | -1.03E-05 |
| Vehicle.StyleCoupe:Popularity | -2.04E-05 |
| Vehicle.StyleCrew Cab Pickup:Popularity | 3.79E-06 |
| Vehicle.StylePassenger Minivan:Popularity | 7.20E-06 |
| Vehicle.StylePassenger Van:Popularity | 2.63E-05 |
| Vehicle.StyleRegular Cab Pickup:Popularity | 8.86E-06 |
| Vehicle.StyleSedan:Popularity | -5.59E-07 |
| Vehicle.Style4dr SUV:Engine.HP.Clean | 0.000515478 |
| Vehicle.StyleExtended Cab Pickup:Engine.HP.Clean | -0.000117068 |
| Vehicle.Style2dr SUV:Engine.Cylinders.Clean4 | 0.000941664 |
| Vehicle.StyleCargo Minivan:Engine.Cylinders.Clean4 | 0.055933727 |
| Vehicle.StyleCrew Cab Pickup:Engine.Cylinders.Clean4 | 0.006886366 |
| Vehicle.StyleExtended Cab Pickup:Engine.Cylinders.Clean4 | 0.062608927 |
| Vehicle.StyleRegular Cab Pickup:Engine.Cylinders.Clean4 | -0.031348509 |
| Vehicle.StyleSedan:Engine.Cylinders.Clean4 | -0.02103876 |
| Vehicle.StyleWagon:Engine.Cylinders.Clean4 | -0.014954989 |
| Vehicle.Style4dr Hatchback:Engine.Cylinders.Clean5 | -0.038141278 |
| Vehicle.Style4dr SUV:Engine.Cylinders.Clean5 | 0.090581719 |
| Vehicle.StyleConvertible:Engine.Cylinders.Clean5 | 0.044086614 |
| Vehicle.StyleCoupe:Engine.Cylinders.Clean5 | 0.131254035 |
| Vehicle.StyleSedan:Engine.Cylinders.Clean5 | 0.027419564 |
| Vehicle.Style4dr SUV:Engine.Cylinders.Clean6 | -0.005421391 |
| Vehicle.StyleCargo Minivan:Engine.Cylinders.Clean6 | -0.04736555 |
| Vehicle.StyleConvertible:Engine.Cylinders.Clean6 | 0.014303536 |
| Vehicle.StyleCoupe:Engine.Cylinders.Clean6 | 0.012348008 |
| Vehicle.StyleCrew Cab Pickup:Engine.Cylinders.Clean6 | -0.001338878 |
| Vehicle.StyleWagon:Engine.Cylinders.Clean6 | 0.000403872 |
| Vehicle.Style4dr SUV:Engine.Cylinders.Clean8 | 0.125927924 |
| Vehicle.StyleConvertible:Engine.Cylinders.Clean8 | 0.15236921 |
| Vehicle.StyleCoupe:Engine.Cylinders.Clean8 | 0.180841382 |
| Vehicle.StyleExtended Cab Pickup:Engine.Cylinders.Clean8 | -0.068374916 |
| Vehicle.StylePassenger Van:Engine.Cylinders.Clean8 | 0.074301007 |
| Vehicle.StyleRegular Cab Pickup:Engine.Cylinders.Clean8 | -0.074643689 |
| Vehicle.StyleSedan:Engine.Cylinders.Clean8 | 0.096801315 |
| Vehicle.StyleWagon:Engine.Cylinders.Clean8 | -0.053073987 |
| Vehicle.StyleCoupe:Engine.Cylinders.Clean10 | 0.000175093 |
| Vehicle.StyleSedan:Engine.Cylinders.Clean12 | 0.007278907 |
| Vehicle.StyleExtended Cab Pickup:NumDoors.Clean3 | -0.028840371 |
| Vehicle.Style4dr Hatchback:NumDoors.Clean4 | -0.003934486 |
| Vehicle.StyleCoupe:NumDoors.Clean4 | -0.029064494 |
| Vehicle.StylePassenger Minivan:NumDoors.Clean4 | 0.066284267 |
| highway.MPG:Engine.HP.Clean | 6.74E-06 |
| highway.MPG:Engine.Cylinders.Clean10 | 2.14E-06 |
| highway.MPG:Engine.Cylinders.Clean12 | 3.56E-05 |
| city.mpg:Engine.HP.Clean | 5.39E-05 |
| city.mpg:Engine.Cylinders.Clean8 | 0.000196987 |
| city.mpg:Engine.Cylinders.Clean10 | 0.015688277 |
| city.mpg:Engine.Cylinders.Clean12 | 0.000405855 |
| Popularity:Engine.HP.Clean | -9.12E-09 |
| Popularity:Engine.Cylinders.Clean4 | 7.57E-06 |
| Popularity:Engine.Cylinders.Clean8 | -3.95E-06 |
| Popularity:Engine.Cylinders.Clean10 | 2.65E-06 |
| Popularity:Engine.Cylinders.Clean12 | 2.12E-06 |
| Engine.HP.Clean:Engine.Cylinders.Clean4 | 0.000120421 |
| Engine.HP.Clean:Engine.Cylinders.Clean8 | -6.03E-06 |
| Engine.HP.Clean:Engine.Cylinders.Clean10 | 4.34E-08 |
| Engine.Cylinders.Clean4:NumDoors.Clean3 | -0.033998496 |
| Engine.Cylinders.Clean5:NumDoors.Clean4 | 0.078560395 |
| Engine.Cylinders.Clean12:NumDoors.Clean4 | 0.000109866 |

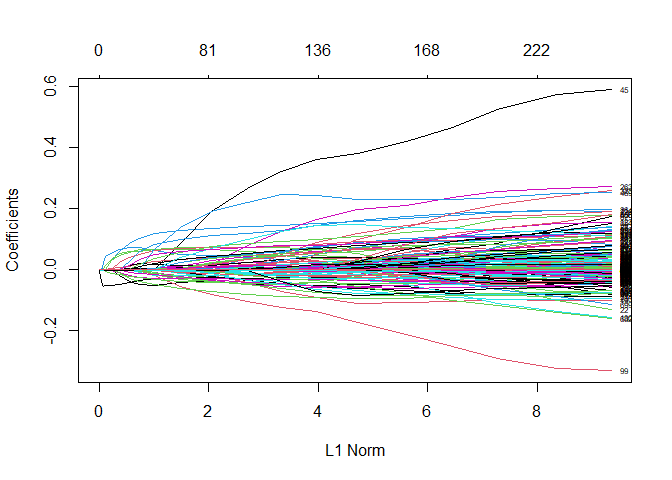
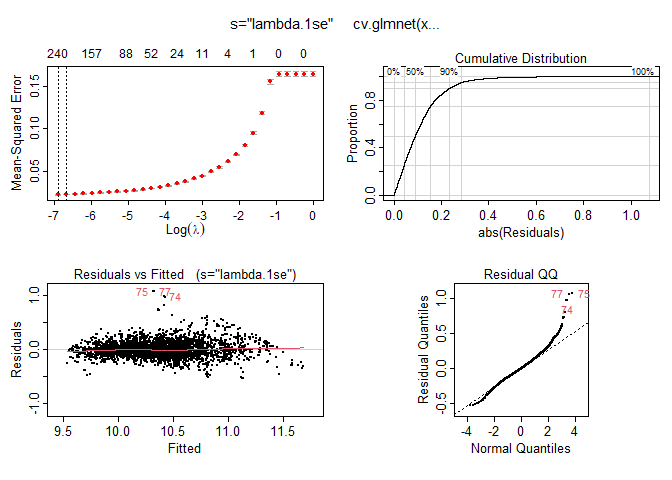
3.4 Fully Saturated Lasso Model ~ model diagnostics



3.5 Find best Lamda

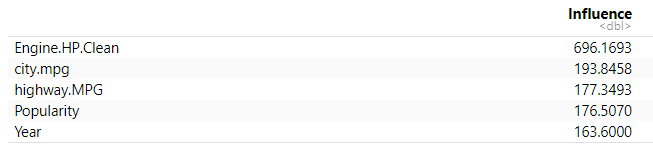


3.6 Model Statistics for Fully Saturated Lasso Model

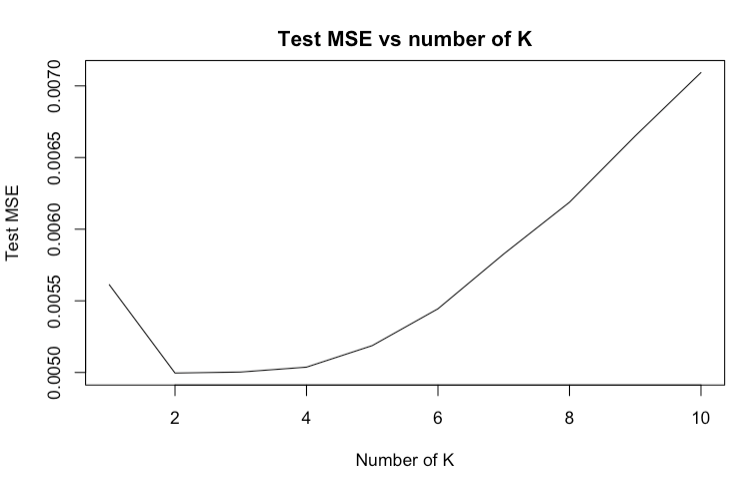


3.7 Random Forest Variable Importance

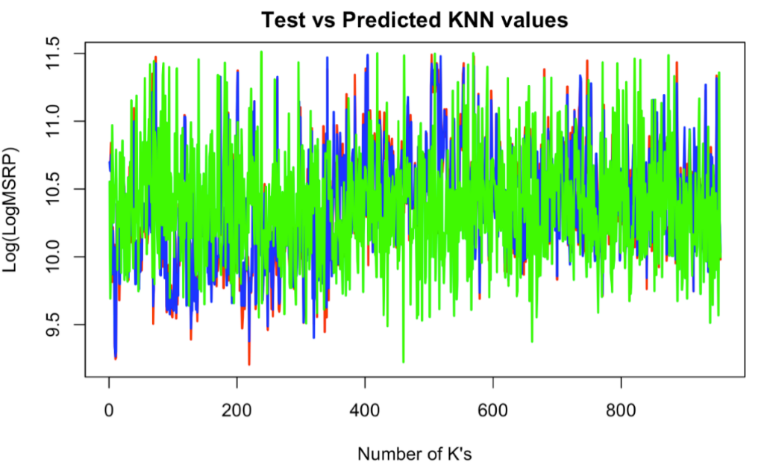
randomForest(log(MSRP)~., data=myDataKNN\_Scaled)



4.0 KNN Optimizer



4.1 KNN Plot



5.0 R-Code

