Project I: MSRP Prediction Model

MSDS 6372

October 2nd , 2021

Randy Kim

Introduction

The vehicle buying process can and often is very stressful and uncomfortable for people. In the back of one's mind are often questions that linger which include "Am I getting this car for a reasonable price" and "What is behind the differences in prices between various cars?". As such, we've been asked to help with this topic by studying a multitude of variables given to us in a data set to understand and predict the MSRP (Manufacturer's Suggested Retail Price) of various vehicles. As a result of our analysis, we hope to help consumers become better informed as they embark on the vehicle purchasing process as well as automotive industry experts increase their understanding of drivers of MSRP. The target vehicle buying audience for this analysis is middle class consumers. One the one end of the price spectrum is very low-priced cars which we believe amenities and characteristics are of much less importance for such consumers as price is really all that matters for, they mainly want a cheap vehicle that runs. Then on the other end of the spectrum are very high priced and high-performance cars which are the opposite, for such those types of consumers care mostly about the amenities and performance but are less concerned with the price they pay. In between these two groups is our target audience which we believe represents a large portion of the vehicle buying population. As such, this analysis will target an MSRP range between $10,000 to $80,000.

Data Description

The given data set contains 11,914 observations of a variety of vehicles with 16 variables. There are 8 numeric variables and 8 categorical variables. The numeric variables contain Year, Engine Horsepower, Engine Cylinders, Number of Doors, highway and city MPG, Popularity, and MSRP, Manufacturer's Suggested Retail Price. The categorical variables contain 7177 different Makers, Model, Engine Fuel Type, Transmission Type, Driven Wheels, Market Category, Vehicle sizes, and Vehicle Style (Table I).

Exploratory Data Analysis

The first step we conducted with this analysis was to wrangle the data to clean, structure and enrich to a desired format. This process included the following adjustments to the data:

- Out of 11,914 observations, there are 105 missing values: 69 Engine Horsepower, 30 Engine Cylinders, and 6 Number of Doors (Table II). Even though these missing values are only 0.88 percent of the observations, certain models and makers make up most of the percentage.
    - We found multiple comparable individual vehicles and therefore calculated the average of such comps for horsepower, number of doors and number of cylinders to fill in many of these missing values.

- We found what appeared to be data errors with cars with a MSRP of $2,000 but these were filtered out by our selected price range.
- We fixed a type error with the MPG for an Audi A6 by replacing with average from similar model and year.
- Given our target audience of middle market consumers and the aforementioned MSRP range, we also made the following adjustments to the data in order to better align with this focus and in turn improve the statistical efficacy of our analysis and model:
    - Although there are a lot of options for affordable electric vehicles, there are only 66 observations of them in the data set. Given the underlying attributes between electric and non-electric cars often greatly differing, we decided that it would be best to focus our analysis on non-electric vehicles.
    - We removed vehicles with horsepower above 550 as above this level are deemed to be vehicles for which performance becomes a key driver thus price sensitivity is much less and therefore does not represent the population we are targeting with our analysis. For similar reasons, we removed vehicles with greater than 8 cylinders.
- Direct Drive was removed from transmission type because all but 2 of these observations represented electric cars. In fact, these two were hybrids which for this analysis will be treated as the electric cars.
- Three blank values were removed from Engine Fuel type.
- Although, including make and models in a prediction model would be increase the accuracy, it would not be efficient for the future predictions. For this analysis, we will exclude those two variables.

Upon finishing the wrangling process, we were left with 9,211 observations which we then segmented into training, test, and validation sets. The training set was used as part of our EDA and model fitting, while the test and validation sets were used to compare and then validate models. First, we assessed characteristics with the response variable, MSRP, which based on a histogram evidenced right skewness though the large sample size mitigates any such concern (Plot I). Then we explored the data for key insights and sought to efficiently look for potential explanatory power. We created a correlation matrix between the continuous variables and MSRP (Plot II). Engine horsepower showed strong correlation with MSRP as evidenced by an R squared of 0.78. Even after making the aforementioned adjustments to horsepower metric, the range was still somewhat wide from a low of 100 to a high of 526 thus evidencing those buyers have a range of power options to choose from and it appears you get what you pay for. The engine horsepower correlation with MSRP was further evidenced via a scatter plot which visually evidenced a high correlation as well (Plot III). Engine cylinders also had a correlation which was notable with an R squared of 0.51. Due to the high correlation between horsepower and cylinders as evidenced in the same correlation matrix and a scatter plot (Plot IV), we will likely only use one of these variables in our model due to the violation of independence. A

scatterplot of Year vs MSRP (Plot V) also evidenced an upward slope as it appears the MSRP has risen over time. Also notable is the range of vehicles by Year as the majority are from 2000 to the final year for the dataset of 2017. As evidenced by a boxplot segmented by engine fuel type (Plot VI), the difference in categories warrants attention and further analysis as a potential powerful variable. Similarly, vehicle size appears to show differences in the range of values per a boxplot by such categories (Plot VII). Though not overly evident, the driven wheel type also appears to be of interest based on differing ranges between such categories.

We were asked to address the potential importance of the "Popularity" variable. Initially, based on a correlation graph (Plot II) and a scatter plot with Popularity and MSRP (Plot VIII), we did not visually find strong evidence of explanatory power. However, after creating a simple linear regression model we found statistical evidence that this variable had explanatory power (p-value = 0.001) with the model assumptions being satisfied as well.

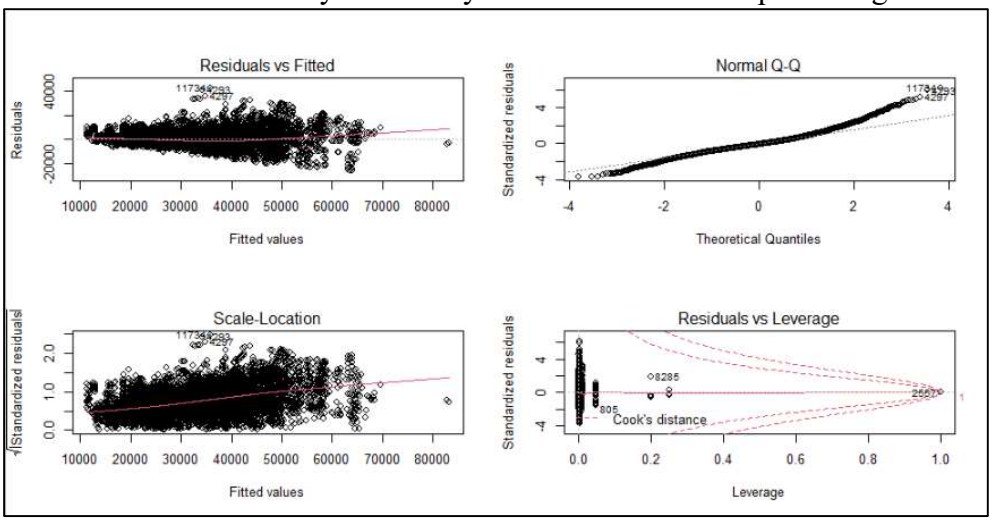
Objective 1

As part of our analysis, we were asked to build a highly interpretable regression model which would serve to help identify and explain in a clear manner key variable that had explanatory power of MSRP.

Checking Assumptions

The following details our assessment of the assumptions for our first model:

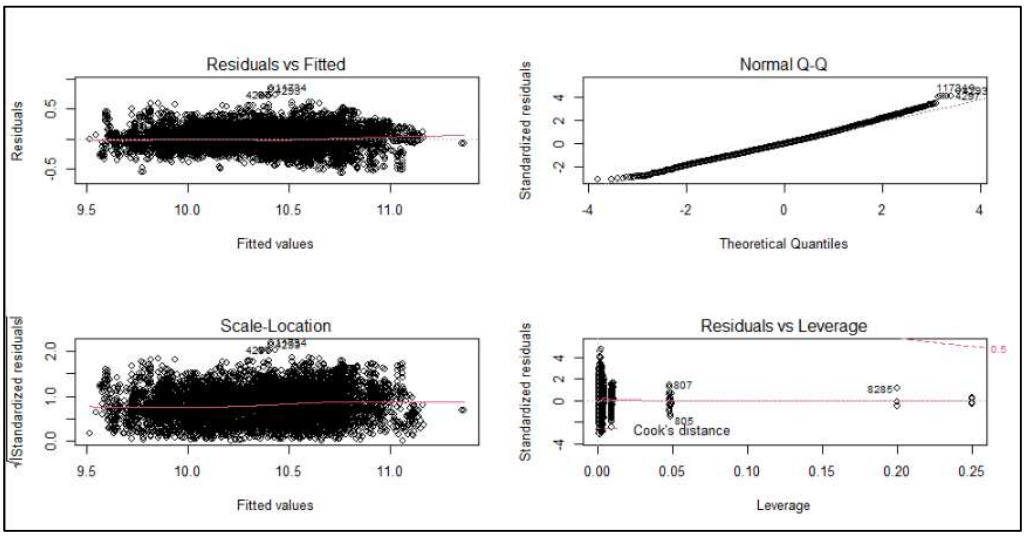
- Normality - Based on the QQ plot and residual plots below, there is evidence against normality, however, due to the large sample sizes, the sampling distribution of the difference of sample means will be normal due to the central limit theorem.
- Constant Variance - Based on the residual scatter plot below, there is evidence to suggest that the constant variance is violated with a cone shape.
- Independence - We chose to not include both Engine HP and Engine Cylinder due to evidence of multicollinearity. Similarly for the two variables pertaining to miles per



gallon (mpg). Besides these, we have no evidence against independence thus will assume this assumption is met.

The following modifications were made to the data considering the aforementioned findings:

- Log of the response variable, MSRP - based on (Plot III), we observed that there is evidence against constant variance and curviness in the scatter plot.
- Log of Engine HP - based on EDA scatterplot and correlation matrix, we decided to log this variable due to the appearance of non-constant variance in the scatterplot vs MSRP. After the transformation, the residuals appear to have variances that show somewhat of a random scatter cloud. We can observe some separation on the bottom right and that may be due to lack of data for that particular observation.
- After the transformation, the QQ plot and residual plots show that the normality and constant variance meet the assumptions. Leverage points have also improved from 1 to 0.25.



Variable Selection

Based on the results of the aforementioned EDA, we found the following to have explanatory power towards the response variable, MSRP and were also deemed to be interpretable: Engine HP, Year, Engine Fuel Type, Driven Wheels, Vehicle Size and Popularity. We found the model below to provide a good bias and variance balance, as well as provided interpretability.

Model 1 - Simple MLR Model

$$Predicted(log(MSRP) = \beta_0 + \beta_1 log(Engine.HP) + \beta_2 Year + \beta_3 Engine.Fuel.Type + \beta_4 Driven.Wheels + \beta_5 Vehicle.Size + \beta_6 Popularity$$

Summary statistics on this model is as follows:

- Overall F-test:

    - Null hypothesis: All of the regression coefficients are equal to 0. Alternative hypothesis: At least one is not equal to 0.
    - Using a significance threshold of 0.05, we reject the null hypothesis in favor of the alternative that there is at least one regression coefficient different from 0 (F-value 1525 with a p-value of < 2.2e-16).
- Adjusted R squared of 0.7679
- No notable variance inflation factors

Interpreting Coefficients

β_1Engine Horsepower:

- Holding all other variables constant, a doubling of Engine Horsepower is associated with a $2^{7.288e-01} = 1.66$ multiplicative change in the median of MSRP. A 95% confidence interval for this is between 1.63 and 1.68.

β_2Year:

- Holding all other variables constant, a one unit increase in Year is associated with a multiplicative change of $e^{1.214e-02} = 1.01$ in the median of MSRP. A 95% confidence interval for this is between 1.011 and 1.013.

β_3Engine Fuel Type:

- Adjustment of the intercept for (premium unleaded required/E85) with respect to a Diesel type. For MSRP of zero, the(premium unleaded required/E85) has an estimated median $e^{-2.637e-03} = 0.997$ units more than a Diesel type. A 95% confidence interval for this is between0.84 and 1.16.
- Adjustment of the intercept for flex-fuel (premium unleaded recommended/E85) with respect to a Diesel type. For MSRP of zero, the flex-fuel (premium unleaded recommended/E85) has an estimated median $e^{-1.337e-01} = 0.87$ units more than a Diesel type. A 95% confidence interval for this is between 0.80 and 0.95.

β_4Driven Wheel:

- Adjustment of the intercept for four-wheel drive with respect to all wheel drive. For MSRP of zero, the four-wheel drive has an estimated median $e^{-4.230e-02} = 0.96$ units more than all wheel drive. A 95% confidence interval for this is between 0.94 and 0.98.
- Adjustment of the intercept for front wheel drive with respect to all wheel drive. For MSRP of zero, the front wheel drive has an estimated median $e^{-9.673e-02} = 0.91$ units more than all wheel drive. A 95% confidence interval for this is between 0.897 and 0.92.

β_5Vehicle Size:

- Adjustment of the intercept for a large vehicle size with respect to compact vehicle size. For MSRP of zero, a large vehicle size has an estimated median $e^{9.970e-02} = 1.10$ units more than a compact vehicle size. A 95% confidence interval for this is between 1.09 and 1.12.
- Adjustment of the intercept for a medium vehicle size with respect to compact vehicle size. For MSRP of zero, a medium vehicle size has an estimated median $e^{5.633e-02} = 1.06$

units more than a compact vehicle size. A 95% confidence interval for this is between 1.05 and 1.07.

β_6 Popularity:

- Holding all other variables constant, a one unit increase in Popularity is associated with a multiplicative change of $e^{-5.826e-06} = 0.99$ in the median of MSRP. A 95% confidence interval for this is between $e^{-8.754e-06}$ and $e^{-2.899e-06}$.

Note, the rest of the variables are detailed in the Appendix (Plot IX).

Objective 2

We were also asked to develop and compare two other models which were premised on being more complex and evidencing more statistical predictive power, though at the expense of interpretability.

Complex Multiple Linear Regression Model

The first complex model we developed was leveraged based on the aforementioned simple multiple linear regression though we assessed interactions as well. We used our EDA as well as automated techniques (forward selection) to determine which interactions to include in this model. Such automated techniques involved using regsubsets from the leap package in R to aid in our variable selection. Due to the number of potential variables and leveraging our EDA, we did not allow this technique to choose all of the potential variables. As such, we decided to include the following variables in our selection process, all of which ended up being included in the final model:

- Year, Engine Fuel Type, Vehicle Size, Driven Wheels, Popularity - all these variables were deemed to be of interest based on our EDA which was previously discussed.
- Interaction variables: Vehicle Size & Driven Wheels, Driven Wheels & Engine HP, Vehicle Size & Engine HP - the first of these was confirmed via a 2-way ANOVA while the latter two via exploratory analysis.

Model 2 – Complex MLR:

$$Predicted(MSRP) = \beta_0 + \beta_1 log(Engine.HP) + \beta_2 Year + \beta_3 Engine.Fuel.Type + \beta_4 Drive.Wheel + \beta_5 Vehicle.Size + \beta_6 Popularity + \beta_7 (Driven.Wheels * Engine.HP) + \beta_8 (Vehicle.Size * Engine.HP)$$

Summary statistics and visual evidence of assumptions being met for this model are detailed in the Appendix (Plots X-XV).

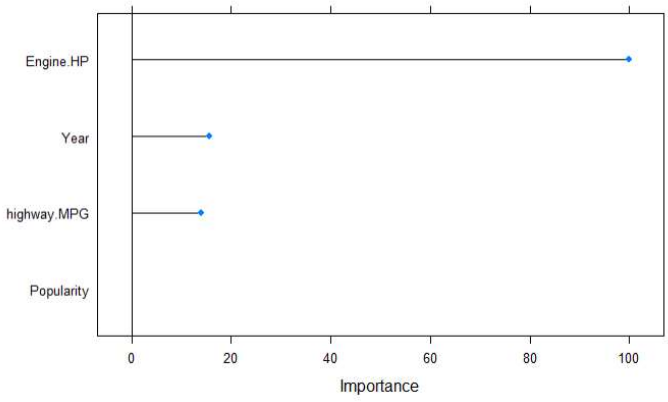
Nonparametric Regression Model

Additionally, we developed a nonparametric model. By way of background and insight, a parametric model assumes that the population can be adequately modeled by a probability distribution that has a fixed set of parameters, while a non-parametric model makes no assumptions about some probability distribution when modeling the data.

We decided to use K-Nearest Neighbors (KNN) for this model which is a supervised machine learning algorithm that can be used for classification and regression problems, the latter of which we used for this analysis. In this algorithm, the k is a constant defined by the modeler and in turn the nearest neighbors distances vector is quantified by it.

First, we narrowed down the variables to be fed into the model based on the aforementioned EDA and model creation as well as by limiting the variables to those which are non-categorical / numeric. We attempted creating numeric variables out of categorical variables (for example – Vehicle Size) but did not find this to be additive thus our model included the following variables: log(Engine HP), Year, Highway MPG and Popularity. We then used 10-fold cross validation, which means the data set used (in this case the training set used in the previous two models) was randomly partitioned into 10 equal size subsamples with each one used exactly once as part of validating of the model. Then a cross-validation process is repeated 10 times and averaged to produce a final estimation.

This 10-fold process concluded with k = 1 being the chosen value as evidenced by the best RMSE, Adjusted R-squared and MAE values (Plot XVI & XVII). More details are found in the Appendix (Plots XVI - XVIII). The graph below details the variable importance as determined by the KNN algorithm. Though Popularity was not deemed of importance it was left in the model though could justifiably be removed. Needless to say, we were pleasantly surprised at the simplicity but effectiveness of this KNN model.



Model Comparisons

Upon completion of building three different models, we compared them to each other using the validation data set. As evidenced in the table below, the KNN model was deemed to be the most powerful prediction model of the three based on the R-squared, RMSE and ASE calculations.

	Efficacy Metric		
Model	R-Squared	RMSE	ASE
Simple MLR	0.7882	0.1736	0.0392
Complex MLR	0.7951	0.1695	0.0316
KNN	0.8609	0.1394	0.0173

Conclusion

In conclusion, our analysis shows that given the provided data set we can with somewhat strong efficacy model out and predict MSRP. If one prefers a model that is highly interpretable and can be used in more practical settings and analysis, then the simple multiple linear regression is recommended. While if one is less concerned with interpretability and more so statistical significance, then using the non-parametric KNN model is recommended given its superior efficacy metrics.

Due to lack of insight into the source and collection process of the underlying data used in this analysis, details on causality and generalization are deemed inconclusive, thus we would assume neither to be true. Additionally, if we were given more time, data and resources we would find it quite interesting to test and calculate models on other vehicle data set to determine the efficacy of our models and if variable selection and explanatory power would be changed.

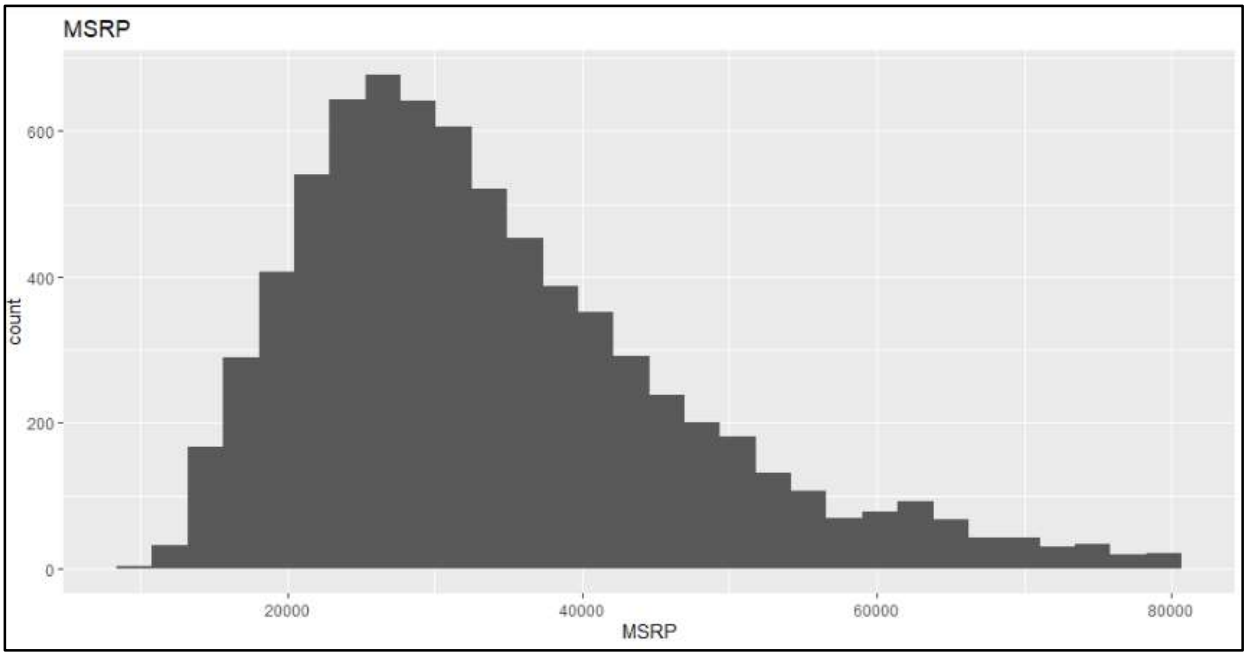
Appendix

Make	Factor w/ 48 levels	"Acura", "Alfa Romeo",..: 6 6 6 6 6 6 6 6 6
Model	Factor w/ 915 levels	"1 Series","1 Series M",..: 2 1 1 1 1 1 1 1 1
Year	Num	2011 2011 2011 2011 2011 ...
Engine Fuel Type	Factor w/ 11 levels	"","diesel", "electric",..: 10 10 10 10 10 10
Engine HP	Num	335 300 300 230 230 230 300 300 230 230 ...
Engine Cylinders	Num	6 6 6 6 6 6 6 6 6 6 ...
Transmission Type	Factor w/ 5 levels	"AUTOMATED_MANUAL",..: 4 4 4 4 4 4
Driven Wheels	Factor w/ 4 levels	"all wheel drive",..: 4 4 4 4 4 4
Number of Doors	Num	2 2 2 2 2 2 2 2 2 2 ...
Market Category	Factor w/ 72 levels	"Crossover","Crossover,Diesel",..: 39 68 65
Vehicle Size	Factor w/ 3 levels	"Compact", "Large",..: 1 1 1 1 1 1 1
Vehicle Style	Factor w/ 16 levels	"2dr Hatchback",..: 9 7 9 9 7 9 7 9 7 9 7
Highway.MPG	Num	26 28 28 28 28 28 26 28 28 27 ...
City.MPG	Num	9 19 20 18 18 18 17 20 18 18 ...
Popularity	Num	3916 3916 3916 3916 3916 ...
MSRP	Num	46135 40650 36350 29450 34500 ...

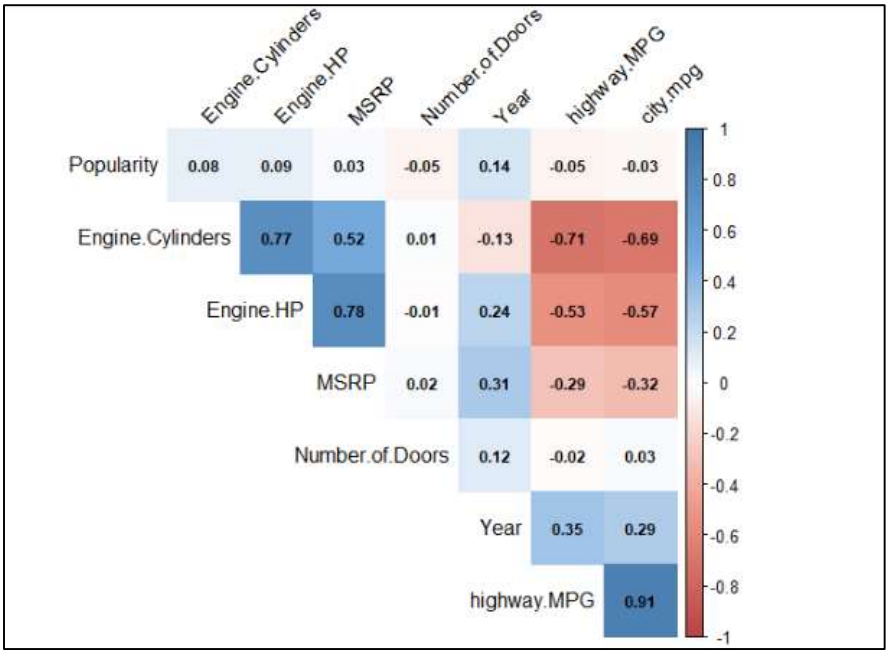
Table I. Dataset description and characteristics

Make	Model	Year	Engine Fuel Type
0	0	0	0
Engine HP	Engine Cylinders	Transmission Type	Driven Wheels
69	30	0	0
Number of Doors	Market Category	Vehicle Size	Vehicle Style
6	0	0	0
Highway MPG	City MPG	Popularity	MSRP
0	0	0	0

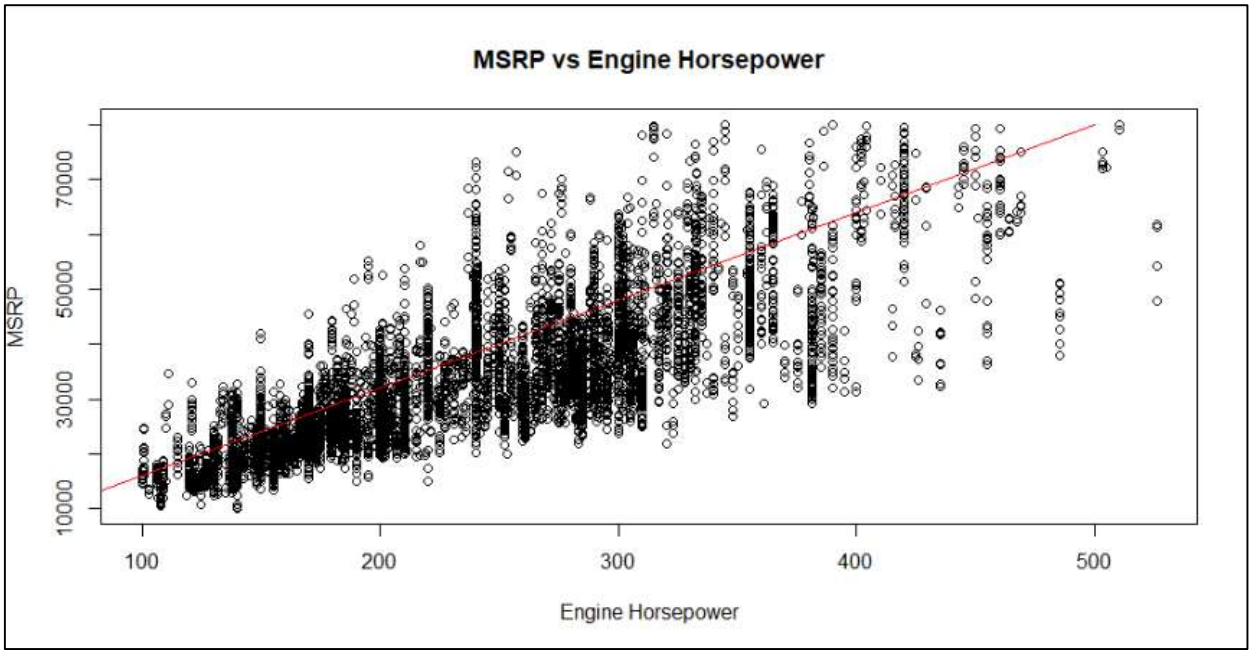
Table II. Missing values in each variable.



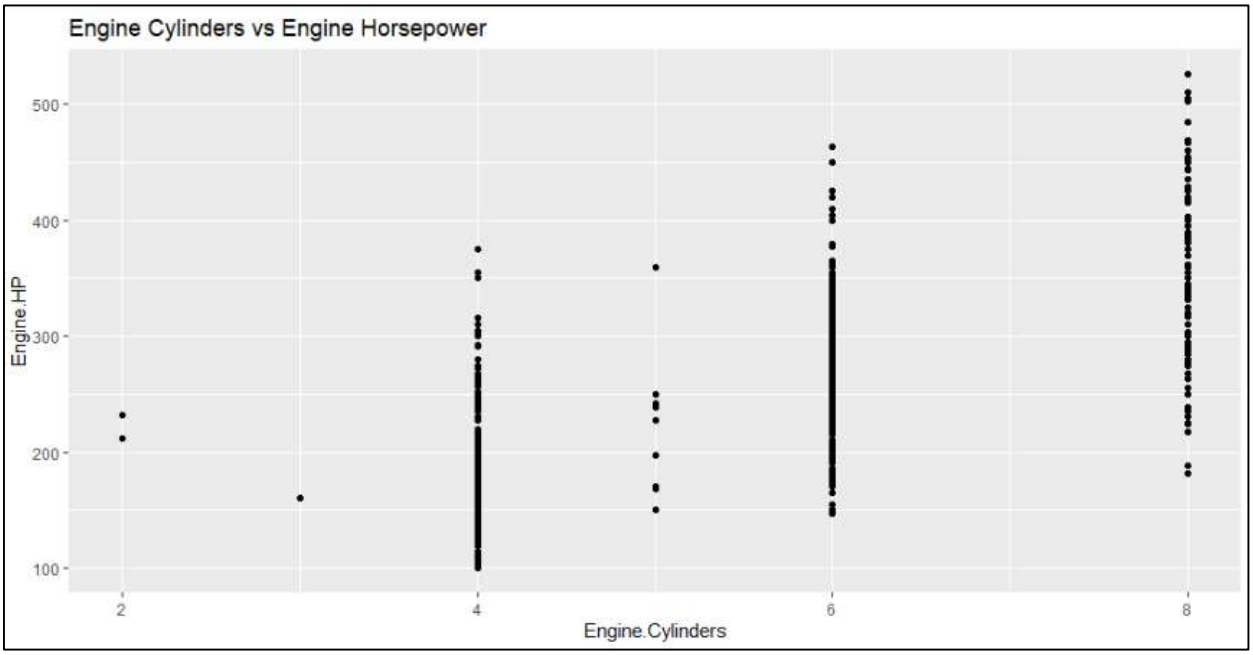
Plot I. MSRP histogram



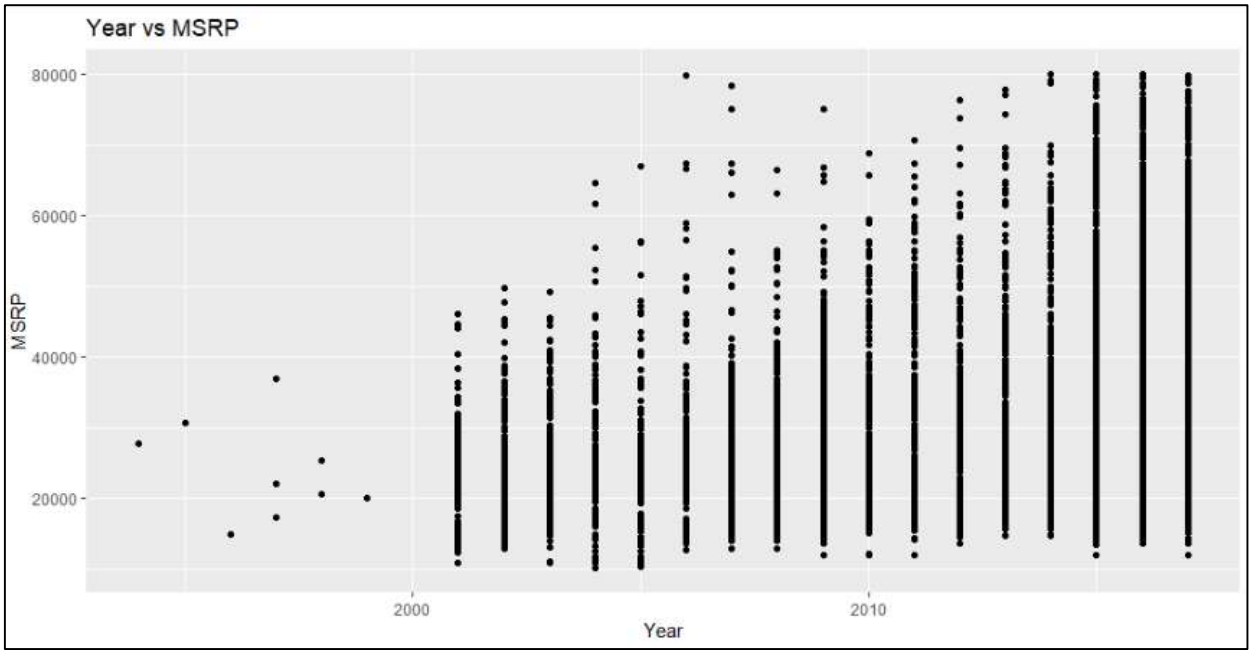
Plot II. MSRP vs Continuous variable: correlation matrix



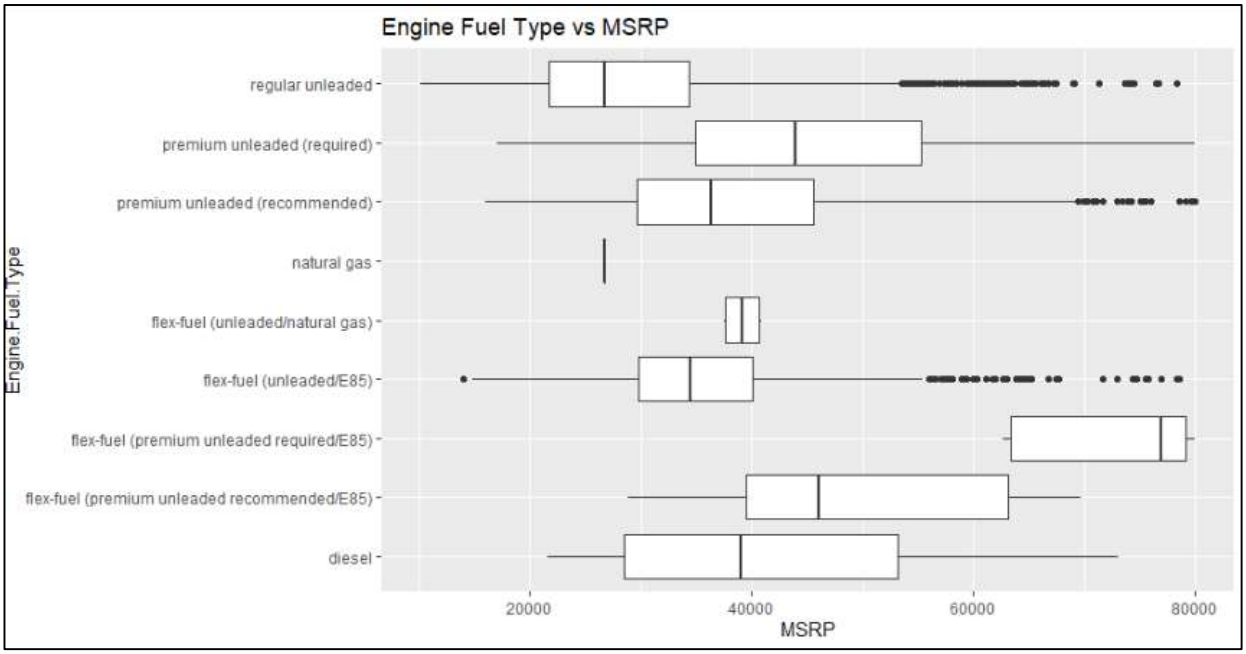
Plot III. MSRP vs Engine Horsepower



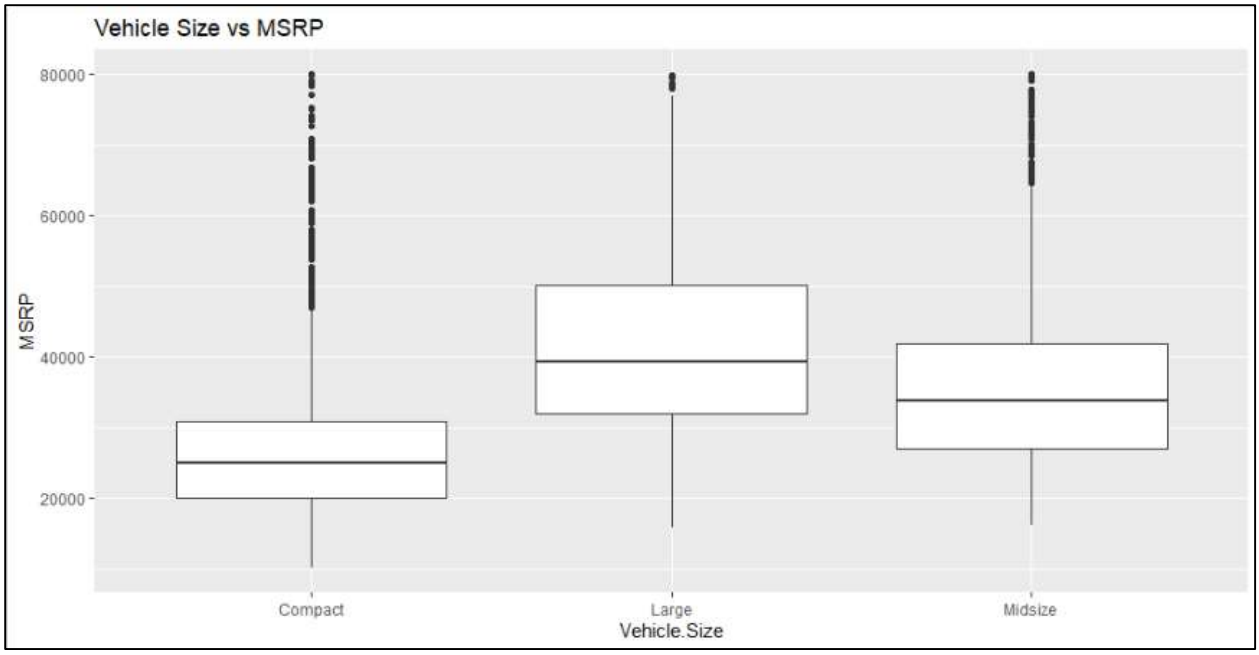
Plot IV. Engine Cylinders vs Engine Horsepower



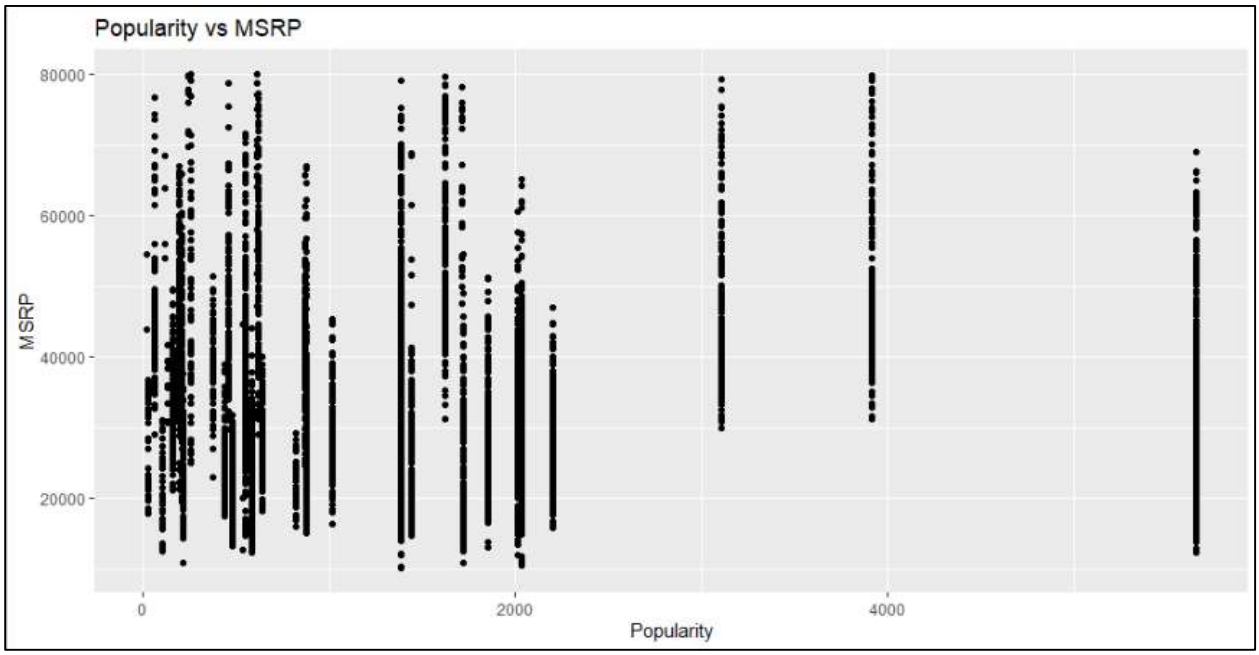
Plot V. Year vs MSRP



Plot VI. Engine Fuel Type vs MSRP



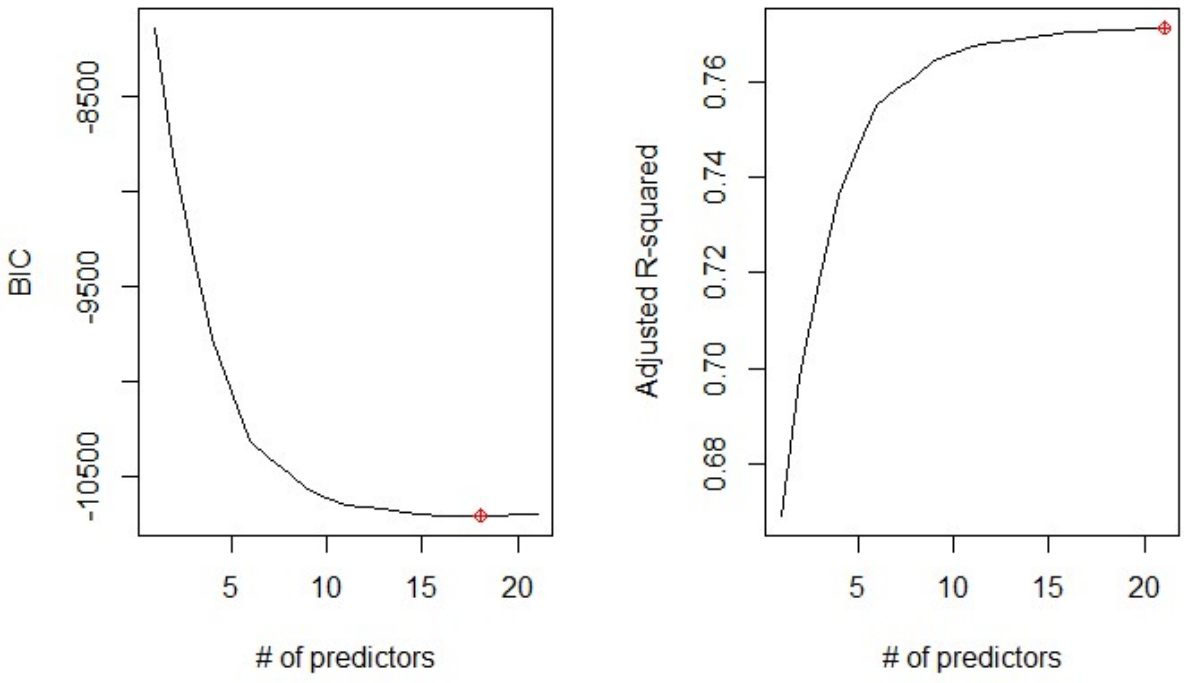
Plot VII. Vehicle Size vs MSRP



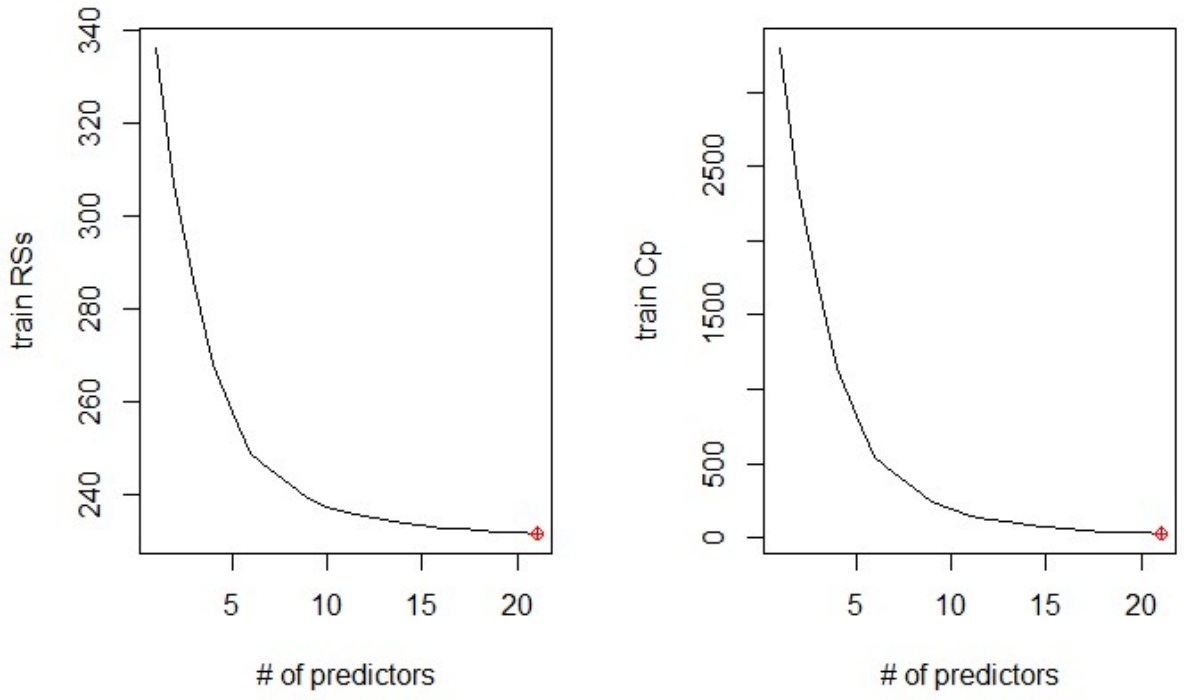
Plot VIII. Popularity vs MSRP

```
Call:
lm(formula = log(MSRP) ~ log(Engine.HP) + Year + Engine.Fuel.Type +
    Driven_Wheels + Vehicle.Size + Popularity, data = Autos_tr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56531 -0.11881 -0.00679  0.11067  0.85518

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           -1.763e+01  1.004e+00 -17.566  < 2e-16 ***
log(Engine.HP)                                         7.288e-01  1.037e-02  70.253  < 2e-16 ***
Year                                                   1.214e-02  5.049e-04  24.036  < 2e-16 ***
Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85) -1.353e-01  4.269e-02  -3.168  0.00154 **
Engine.Fuel.Typeflex-fuel (premium unleaded required/E85)    -1.058e-02  8.231e-02  -0.129  0.89770
Engine.Fuel.Typeflex-fuel (unleaded/E85)              -4.502e-01  1.843e-02 -24.424  < 2e-16 ***
Engine.Fuel.Typeflex-fuel (unleaded/natural gas)      -4.212e-01  9.128e-02  -4.615 4.00e-06 ***
Engine.Fuel.Typenatural gas                            5.800e-02  1.798e-01   0.323  0.74698
Engine.Fuel.Typepremium unleaded (recommended)        -2.870e-01  1.777e-02 -16.152  < 2e-16 ***
Engine.Fuel.Typepremium unleaded (required)           -1.991e-01  1.814e-02 -10.976  < 2e-16 ***
Engine.Fuel.Typeregular unleaded                      -4.130e-01  1.699e-02 -24.310  < 2e-16 ***
Driven_Wheelsfour wheel drive                         -4.178e-02  8.442e-03  -4.949 7.61e-07 ***
Driven_Wheelsfront wheel drive                        -9.714e-02  6.104e-03 -15.915  < 2e-16 ***
Driven_Wheelsrear wheel drive                         -9.098e-02  6.652e-03 -13.677  < 2e-16 ***
Vehicle.SizeLarge                                      1.007e-01  7.511e-03  13.407  < 2e-16 ***
Vehicle.SizeMidsize                                    5.606e-02  5.450e-03  10.286  < 2e-16 ***
Popularity                                            -5.827e-06  1.494e-06  -3.901 9.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1789 on 7352 degrees of freedom
Multiple R-squared:  0.7684,    Adjusted R-squared:  0.7679
F-statistic:  1525 on 16 and 7352 DF,  p-value: < 2.2e-16

                   GVIF Df GVIF^(1/(2*Df))
log(Engine.HP)  2.688325  1        1.639611
Year            1.200324  1        1.095593
Engine.Fuel.Type 1.820127  8        1.038141
Driven_Wheels   2.004039  3        1.122840
Vehicle.Size    2.110962  2        1.205369
Popularity      1.080841  1        1.039635
                                                        Estimate         2.5 %         97.5 %
(Intercept)                                           -1.763402e+01 -1.960187e+01 -1.566616e+01
log(Engine.HP)                                         7.287778e-01  7.084424e-01  7.491131e-01
Year                                                   1.213613e-02  1.114635e-02  1.312590e-02
Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85) -1.352538e-01 -2.189459e-01 -5.156172e-02
Engine.Fuel.Typeflex-fuel (premium unleaded required/E85)    -1.058287e-02 -1.719319e-01  1.507661e-01
Engine.Fuel.Typeflex-fuel (unleaded/E85)              -4.501822e-01 -4.863136e-01 -4.140507e-01
Engine.Fuel.Typeflex-fuel (unleaded/natural gas)      -4.212327e-01 -6.001594e-01 -2.423060e-01
Engine.Fuel.Typenatural gas                            5.799832e-02 -2.943890e-01  4.103857e-01
Engine.Fuel.Typepremium unleaded (recommended)        -2.870144e-01 -3.218478e-01 -2.521809e-01
Engine.Fuel.Typepremium unleaded (required)           -1.991010e-01 -2.346599e-01 -1.635421e-01
Engine.Fuel.Typeregular unleaded                      -4.130028e-01 -4.463060e-01 -3.796997e-01
Driven_Wheelsfour wheel drive                         -4.178354e-02 -5.833257e-02 -2.523450e-02
Driven_Wheelsfront wheel drive                        -9.713647e-02 -1.091014e-01 -8.517158e-02
Driven_Wheelsrear wheel drive                         -9.098086e-02 -1.040209e-01 -7.794086e-02
Vehicle.SizeLarge                                      1.006963e-01  8.597330e-02  1.154192e-01
Vehicle.SizeMidsize                                    5.605612e-02  4.537311e-02  6.673912e-02
Popularity                                            -5.826977e-06 -8.754764e-06 -2.899191e-06
```

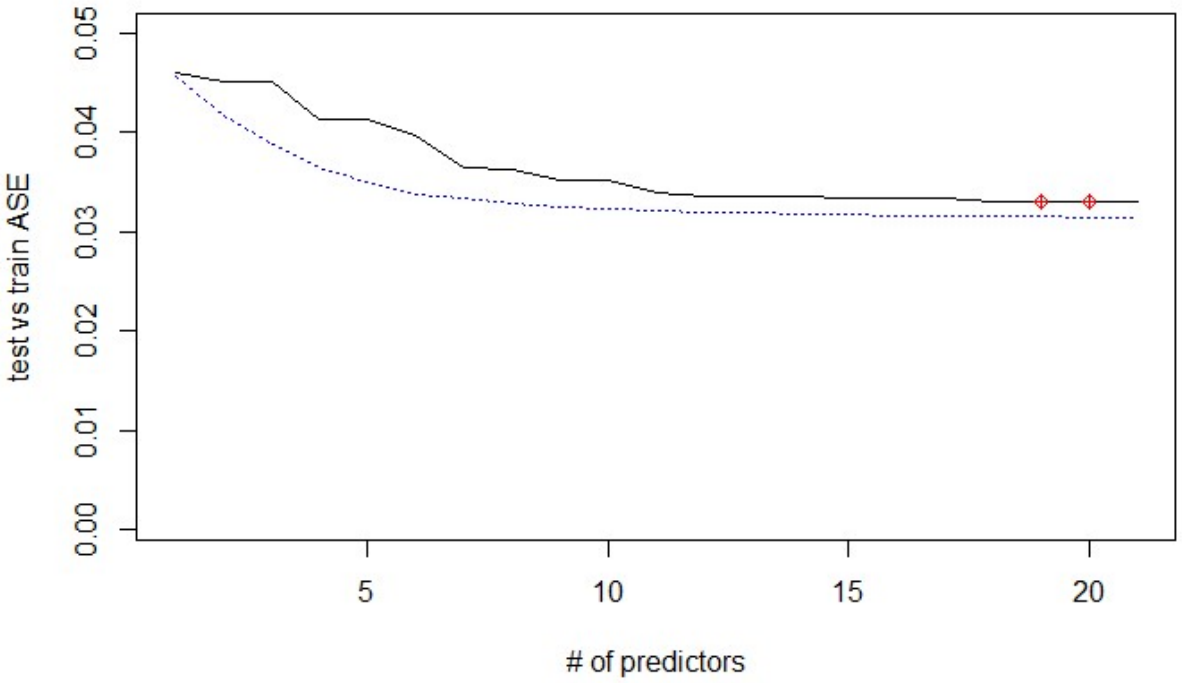
Plot IX. Simple MLR Model



Plot X. BIC & AdjR2 – forward selection



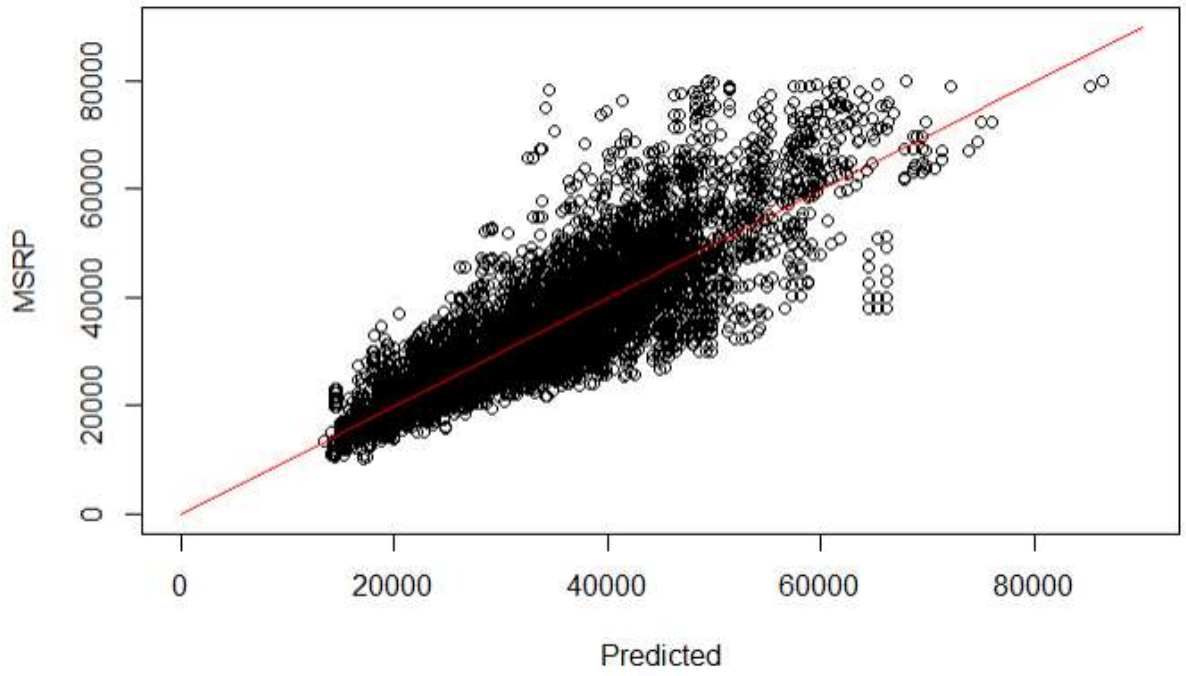
Plot XI. Train RSS & Train CP – forward selection



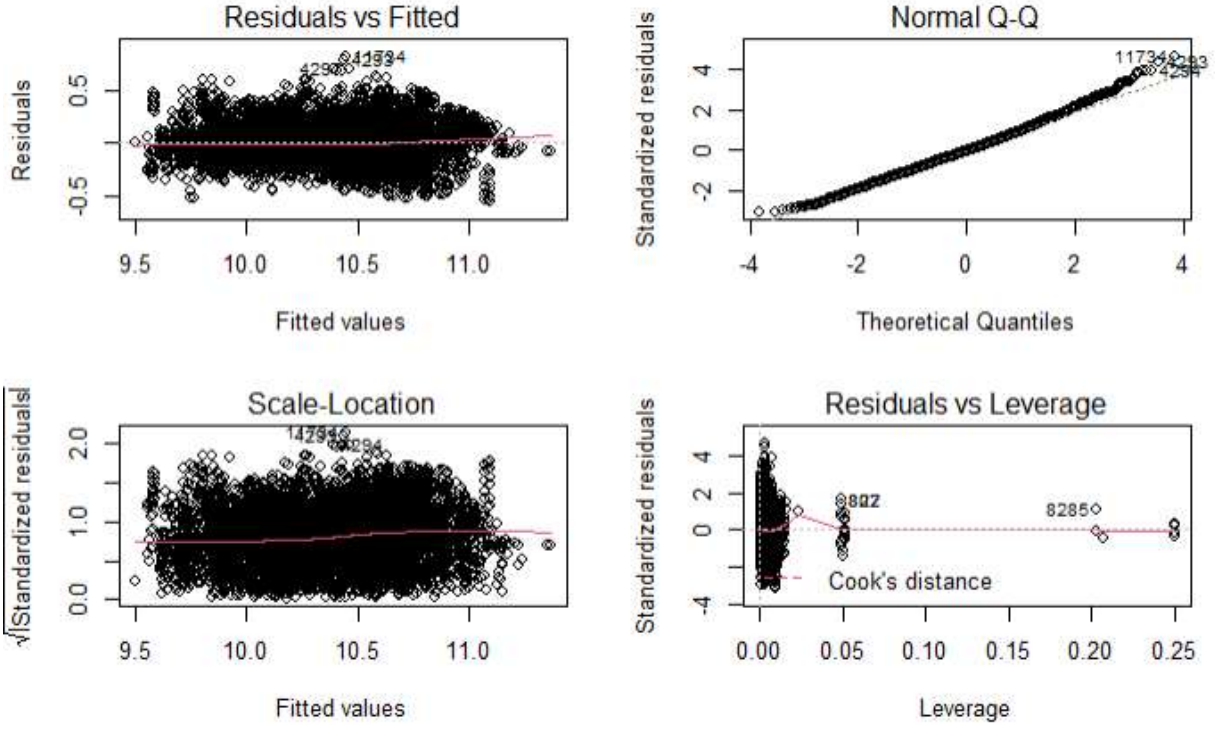
Plot XII. Test vs Train ASE

```
Call:
lm(formula = log(MSRP) ~ log(Engine.HP) + Year + Engine.Fuel.Type +
    Driven_Wheels + Vehicle.Size + Popularity + Engine.HP:Driven_Wheels +
    Vehicle.Size:Driven_Wheels + Engine.HP:Vehicle.Size, data = Autos_tr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.54967 -0.11920 -0.00613  0.11198  0.82053

Coefficients:
                                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                             -1.786e+01  1.053e+00 -16.969  < 2e-16 ***
log(Engine.HP)                                           5.788e-01  5.916e-02   9.784  < 2e-16 ***
Year                                                     1.257e-02  5.168e-04  24.333  < 2e-16 ***
Engine.Fuel.Typeflex-fuel (premium unleaded recommended/E85) -1.516e-01  4.277e-02  -3.546 0.000394 ***
Engine.Fuel.Typeflex-fuel (premium unleaded required/E85)    -3.870e-02  8.525e-02  -0.454 0.649838
Engine.Fuel.Typeflex-fuel (unleaded/E85)                 -4.414e-01  1.855e-02 -23.794  < 2e-16 ***
Engine.Fuel.Typeflex-fuel (unleaded/natural gas)         -4.936e-01  9.129e-02  -5.407 6.60e-08 ***
Engine.Fuel.Typenatural gas                               6.701e-02  1.784e-01   0.376 0.707239
Engine.Fuel.Typepremium unleaded (recommended)           -2.849e-01  1.775e-02 -16.051  < 2e-16 ***
Engine.Fuel.Typepremium unleaded (required)              -1.991e-01  1.819e-02 -10.944  < 2e-16 ***
Engine.Fuel.Typeregular unleaded                         -4.136e-01  1.700e-02 -24.331  < 2e-16 ***
Driven_Wheelsfour wheel drive                            -5.143e-02  3.620e-02  -1.421 0.155449
Driven_Wheelsfront wheel drive                           -1.637e-01  2.654e-02  -6.167 7.32e-10 ***
Driven_Wheelsrear wheel drive                            -3.123e-02  2.644e-02  -1.181 0.237528
Vehicle.SizeLarge                                         1.025e-01  3.679e-02   2.788 0.005323 **
Vehicle.SizeMidsize                                       1.703e-01  2.705e-02   6.297 3.21e-10 ***
Popularity                                               -5.795e-06  1.489e-06  -3.893 0.000100 ***
Driven_Wheelsall wheel drive:Engine.HP                    6.310e-04  2.892e-04   2.182 0.029123 *
Driven_Wheelsfour wheel drive:Engine.HP                   7.347e-04  2.927e-04   2.510 0.012085 *
Driven_Wheelsfront wheel drive:Engine.HP                  9.925e-04  3.490e-04   2.844 0.004469 **
Driven_Wheelsrear wheel drive:Engine.HP                   5.963e-04  2.481e-04   2.403 0.016274 *
Driven_Wheelsfour wheel drive:Vehicle.SizeLarge          -6.166e-02  2.542e-02  -2.426 0.015308 *
Driven_Wheelsfront wheel drive:Vehicle.SizeLarge          1.393e-02  2.219e-02   0.628 0.530302
Driven_Wheelsrear wheel drive:Vehicle.SizeLarge           -8.734e-02  2.049e-02  -4.264 2.04e-05 ***
Driven_Wheelsfour wheel drive:Vehicle.SizeMidsize         2.223e-02  2.249e-02   0.988 0.323111
Driven_Wheelsfront wheel drive:Vehicle.SizeMidsize        -2.513e-02  1.506e-02  -1.669 0.095212 .
Driven_Wheelsrear wheel drive:Vehicle.SizeMidsize         -7.250e-02  1.709e-02  -4.243 2.23e-05 ***
Vehicle.SizeLarge:Engine.HP                               8.087e-05  1.228e-04   0.659 0.510237
Vehicle.SizeMidsize:Engine.HP                            -3.750e-04  1.035e-04  -3.622 0.000294 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1775 on 7340 degrees of freedom
Multiple R-squared:  0.7725,    Adjusted R-squared:  0.7716
F-statistic: 890.1 on 28 and 7340 DF,  p-value: < 2.2e-16

                                GVIF Df GVIF^(1/(2*Df))
log(Engine.HP)             8.884128e+01  1        9.425565
Year                       1.277710e+00  1        1.130358
Engine.Fuel.Type           2.439389e+00  8        1.057317
Driven_Wheels              1.313251e+04  3        4.857259
Vehicle.Size               1.433666e+03  2        6.153356
Popularity                 1.091476e+00  1        1.044737
Driven_Wheels:Engine.HP    1.357218e+06  4        5.842264
Driven_Wheels:Vehicle.Size 2.241638e+03  6        1.902014
Vehicle.Size:Engine.HP     1.482857e+03  2        6.205471
```

Plot XIII. Complex MLR Model



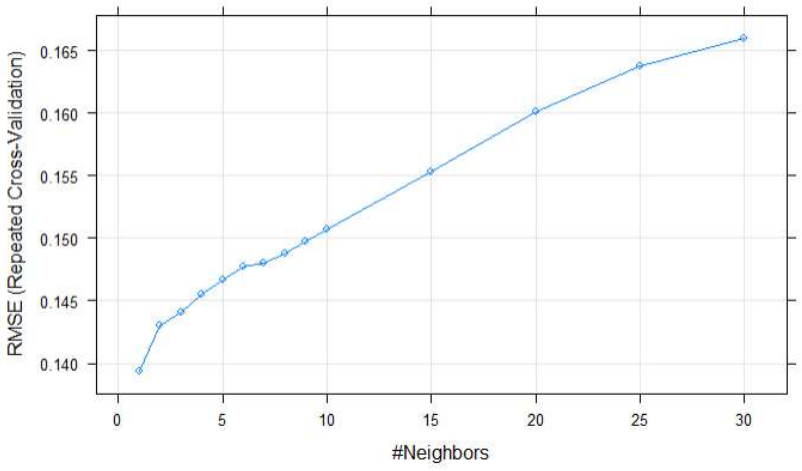
Plot XIV. Complex MLR Model



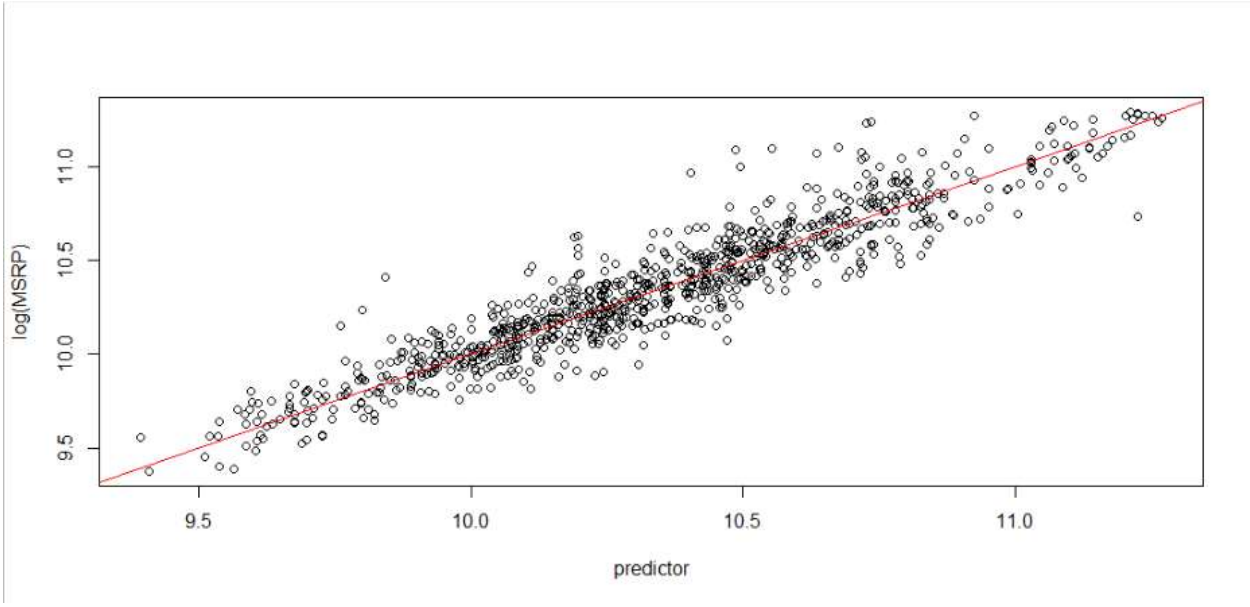
Plot XV. Complex MLR Model Residuals

```
k-Nearest Neighbors

7369 samples
   4 predictor

Pre-processing: centered (4), scaled (4)
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 6631, 6633, 6632, 6633, 6632, 6632, ...
Resampling results across tuning parameters:

  k   RMSE       Rsquared   MAE
   1  0.1394257  0.8609468  0.1050687
   2  0.1430540  0.8527229  0.1067321
   3  0.1440662  0.8502830  0.1081723
   4  0.1455265  0.8470323  0.1095051
   5  0.1466573  0.8444887  0.1107550
   6  0.1477453  0.8421557  0.1119507
   7  0.1480065  0.8414714  0.1125433
   8  0.1487910  0.8397916  0.1132322
   9  0.1497063  0.8378081  0.1140074
  10  0.1507431  0.8355900  0.1149344
  15  0.1552919  0.8257001  0.1194861
  20  0.1601185  0.8148161  0.1234509
  25  0.1637359  0.8066346  0.1265707
  30  0.1659425  0.8014953  0.1283422

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 1.
```

Plot XVI. KNN – 10 fold CV to choose k



Plot XVII. KNN – 10 fold CV to choose k



Plot XVIII. KNN validation