Cars MSRP

DS6372: Project 1

By: Justin Ehly, Wang Renfeng, Allen Miller

Introduction

Manufacturer's Suggested Retail Price or MSRP is the price that a vehicle's manufacturer suggests that a specific vehicle should be sold for. Different features of a vehicle can contribute to what the MSRP is. This report will look at some of these features and explore the impact that they might have on the current retail price and predicting what suggested pricing should be in the future.

Data Description

The Autos dataset used was obtained from SMU's Applied Statistical Analysis project requirements. This included 11914 original observations for 48 distinct vehicle makes. 16 variables were originally explored to verify their significance on accurately predicting the MSRP. An explanation is provided in the EDA section about which variables were changed, removed, or added to the dataset to assist in performance, a full table with all variable names considered and their description is included in the appendix. A data description is available in Appendix Figure 1.1 and a summary of the original data is available below and in figure. Appendix Figure 1.2

```
 Vehicle.Size   highway.MPG       Popularity        MSRP        FactoryTuner Luxury     FlexFuel   Hatchback  Diesel
Compact:3856   Min.   : 12.00   Min.   :   2   Min.   : 10135   No :9525     No :7214   No :8907   No :9090   No :9886
Large  :2377   1st Qu.: 22.00   1st Qu.: 549   1st Qu.: 25095   Yes: 558     Yes:2869   Yes:1176   Yes: 993   Yes: 197
Midsize:3850   Median : 26.00   Median :1385   Median : 32800
               Mean   : 27.25   Mean   :1565   Mean   : 45914
               3rd Qu.: 31.00   3rd Qu.:2009   3rd Qu.: 45205
               Max.   :354.00   Max.   :5657   Max.   :643330

Hybrid     Crossover  Performance HighPerformance              Make_new       Year_new              Vehicle.Style_new
No :9736   No :8051   No :8233    No :8711        Insignificant Make:6954   2015   :2170   2dr SUV            :  33
Yes: 347   Yes:2032   Yes:1850    Yes:1372        Dodge              : 426   2016   :2157   4dr SUV            :2396
                                                  Cadillac           : 378   2017   :1668   Insignificant Style:5067
                                                  BMW                : 318   2014   : 589   Sedan              :2587
                                                  Audi               : 284   2012   : 386
                                                  Mercedes-Benz      : 283   2009   : 378
                                                  (Other)            :1440   (Other):2735

            Cylinders_new
12                 : 212
3                  :  18
Insignificant Style:9853


  Engine.HP      Transmission.Type           Driven_Wheels
 Min.   :  1.0   AUTOMATIC:7539    all wheel drive  :2282
 1st Qu.: 79.0   MANUAL   :2544    four wheel drive :1122
 Median :137.0                     front wheel drive:3999
 Mean   :139.8                     rear wheel drive :2680
 3rd Qu.:185.0
 Max.   :363.0
```

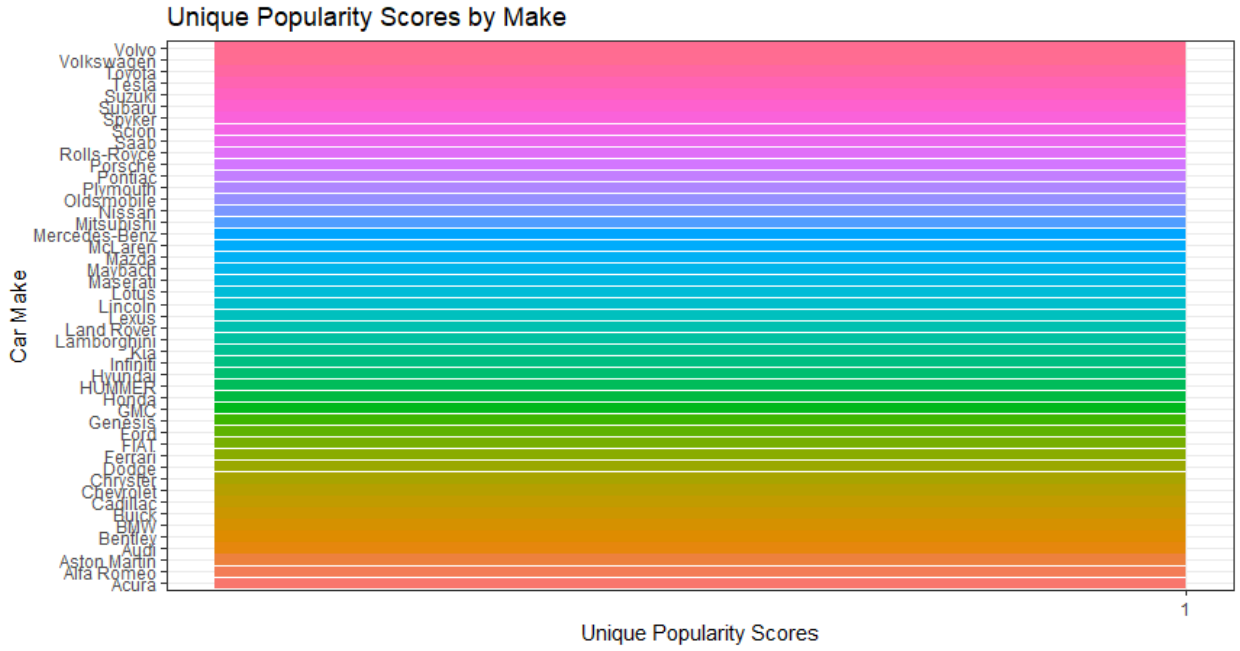
Exploratory Data Analysis

Starting with the summary data we noticed there was a maximum highway MPG of 354 and that was tied to an Audi A6, we looked that up and it should've been 34.

Missing Data:

We found the data contained over 3,700 missing values for the market category variable. After exploring the possibility of imputing the missing values we decided to remove this column due to its complexity and our inability to properly impute the missing data and based on the directions in the assignment we added in a column to determine if a car was exotic or not, the definer being cars with MSRP over $100,000. Other variables with missing data were horsepower, cylinders, transmission type, number of doors, city mpg and fuel type were all filled in using online research for each particular vehicle, this is charted in Appendix Figure 1.3 with additional details in Appendix Detail 1.1 and Detail 1.2.

Investigating the popularity of each vehicle against the MSRP we found that all models for each vehicle make had the same popularity score. This made the popularity variable essentially a unique identifier for each make of vehicle and provided mirrored results when estimating the MSRP using make or popularity as predictors. We decided to drop the popularity variable during variable selection as it would be redundant when make is included. The graph below shows that there was only a single unique popularity score per vehicle make, further testing confirmed no makes had a common popularity score.

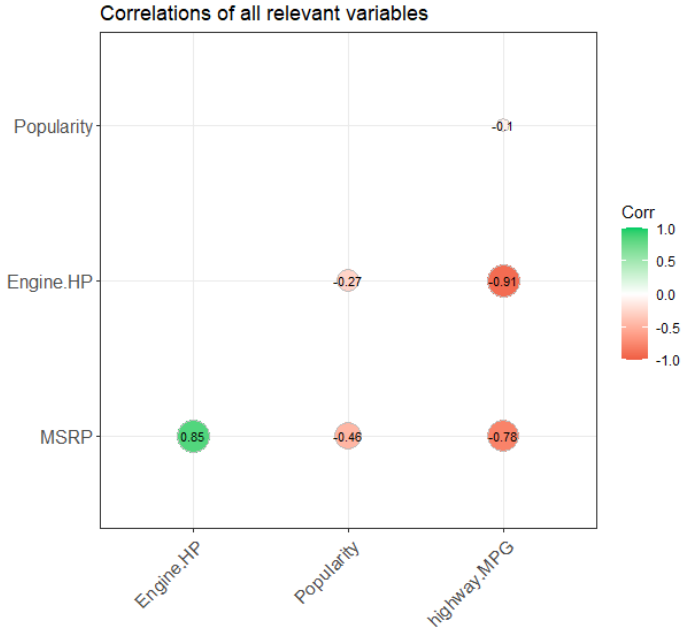


We also found that vehicles made produced prior to 2001 skewed the overall MSRP where the average price was $2,530 compared to average vehicle prices in 2001 that were $41,501. Appendix Figure 1.4 and Figure 1.5. Due to this we decided to remove any observations that were older than 2001 which helped normalize our MSRP data. Appendix Figure 1.6

Vehicles with MSRPs larger than $1,000,000 appeared to heavily skew the data Appendix Figure 1.7 so we chose to remove those as well and only keep vehicles with MSRPs below $1,000,000. Appendix Figure 1.8

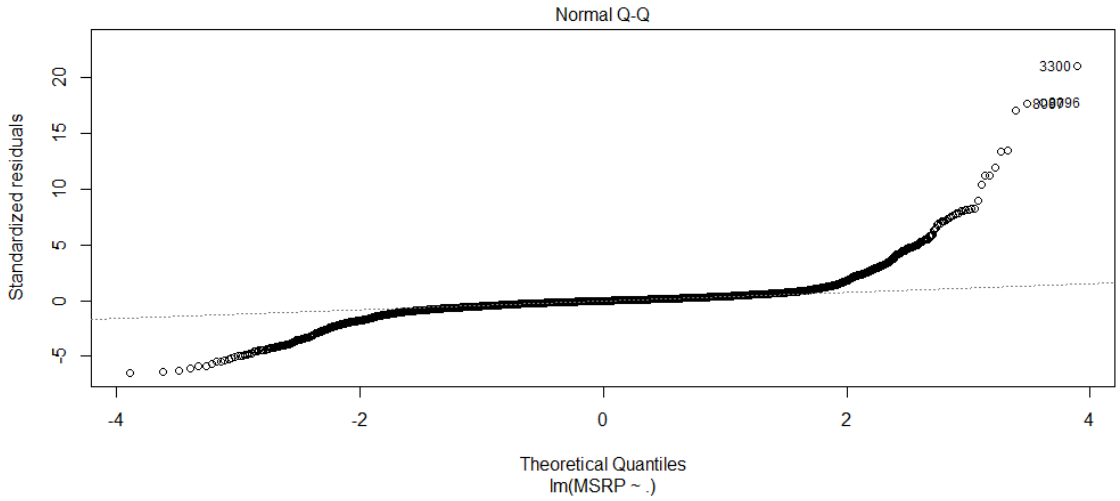
Correlation Between Continuous Variables.
Once the data was cleaned, we looked for visual indications of correlation between the continuous variables. With the Popularity variable determined to be a redundancy of the Make variable, we were left with 3 continuous variables in HP, Highway MPG and City MPG. Appendix Figure 1.9

We further confirmed our observations using a Correlation Graph (below) showing Engine horsepower and Highway MPG are highly correlated.

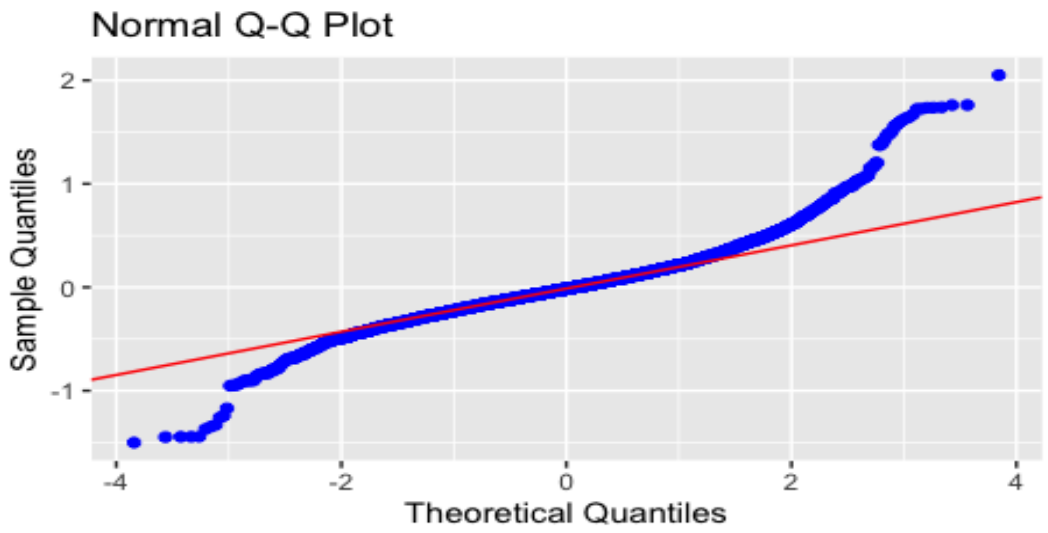


While assessing the residual plots for the MSRP models we found that we had an issue with linearity of the residuals Appendix Figure 1.10. To combat this, we decided to transform the MSRP variable by taking the log of the values. This transformation improved our linear distribution of the residuals. Appendix Figure 1.11 However, when interpreting the model, we will now need to use the median MSRP rather than the mean MSRP. A graphic depicting the logMSRP variable is available in the appendix section.

Before Log transformation.



After Log transformation



Objective 1

Problem

1. Build a highly interpretable model that identifies key relationships associated with vehicle MSRP

2. Build a highly predictive model for vehicle MSRP

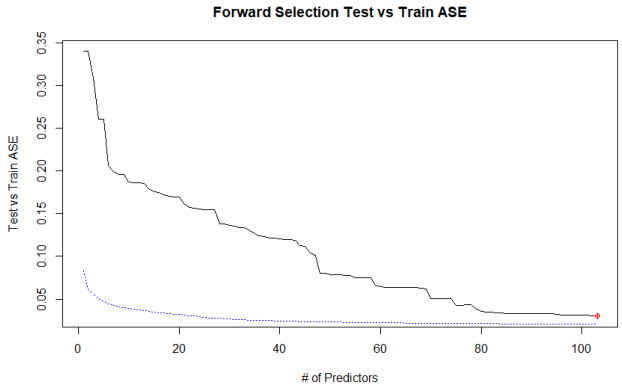
Approach

1. Export and clean the data
    a. Imputed, filled in missing values in all columns based on vehicle data online.
    b. Decomposed Market Category column into Exotic/ Not Exotic.
    c. Removed all the data before Year 2001 due to anomalies with the MSRPs.
    d. Removed a dataset whose highway MPG shows 354 as it is a data entry error.
2. Used ANOVA and p-values > 0.05 to remove variables that were insignificant Appendix Figure 1.12
3. With the remaining data, we used variable selection techniques to identify the best variables for the model. Techniques used: forward and backward selection and LASSO. (Stepwise with this many factorized variables seemed to lock up our computers)
4. Added quadratics for engine horsepower to the predictive model because a non-linear relationship was observed.
5. Added interactions for quadratics engine horsepower multiple cubic engine horsepower to add model complexity.
6. Hypothesis tests:
    a. Extra sum of squares: Check if interaction item is statistically significant.
    b. Checking the significance of the factorial variables. Engine fuel type was found to not be significant and was removed.
    c. Lack of fit: Lasso and Stepwise
7. Checked assumptions for full, interpretable, and predictive models.
8. Model diagnostics and comparisons
    a. ASE
    b. BIC
    c. R^2 / Adjusted R^2
    d. RMSE
9. Model selection based on fit test significance and model diagnostics metrics.

Model variable selection

Forward and Stepwise

Forward	Backward	Stepwise

		Because 116 predictors were in play. We cannot run Stepwise selection without completely bogging down our machines.
Predictors = 103	Predictors = 103	

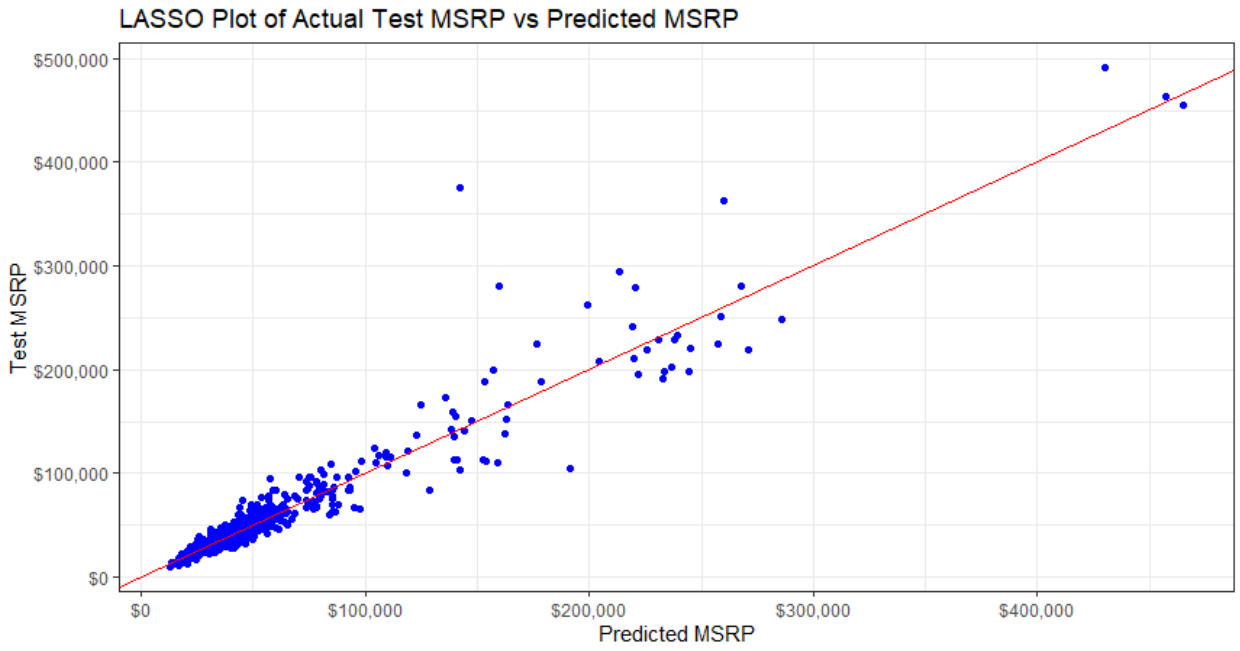
- Both forward selection and backward selection suggested the same amounts of predictors to be included in the model. These comparisons were done using ASE values.

	BIC	Adj R^2	MallowCP	Test ASE	ASE preds
Forward	84	93	93	0.03001098	103
Backward	70	91	88	0.02242308	103

LASSO

Ultimately, we decided to use the LASSO model as our simple model. It suggested our single continuous variable and 11 other categorical variables plus their specific significant levels to be included as predictors. This model was chosen due to its low RMSE value of 13,158 and high R^2 value of 0.9202 on our test and training sets.

The final RMSE and Adjusted R^2 values once run with our validation data split were 10,254 and 0.9342 respectively.



**Test MSRP vs Predicted MSRP LASSO*

Model 1 (Interpretable)

Assumptions

Normality
Due to the large sample size normality is not a concern as the central limit theorem should fix any non-normality.

Linear
Our original model contained a non-log transformed MSRP variable, with this there was a large amount of non-linear relationship when evaluating qq-plots. We decided that log transforming the response (MSRP) provided enough correction to the linearity issues that our model could still be used.

Independent Data
As we don't know much about the origin of the data set it's hard to establish that the data was truly collected independent of one another. However, we assumed the data was independent and proceeded cautiously.

Equal Variance
After we log transformed the MSRP variable we saw our residual plot go from patterned to a more random cloud. There still may be small evidence of unequal variance in the residual plot but it looks close enough that we felt comfortable moving forward with our analysis.

Parameter Interpretation

Predictor	Est. Value	Interpretation
Intercept	10.3019	When all other predictors are equal to zero it is estimated that the median MSRP for a vehicle is $29,789.16
MakeAlfa Romero	0.0351	When all other predictors are held constant it is estimated that when the Make is Alfa Romero there is a multiplicative change of $1.0357 in MSRP
Engine.Fuel.Typediesel	0.1727	When all other predictors are held constant it is estimated that when the Engine Fuel Type is Diesel there is a multiplicative change of $1.1885 in MSRP
Engine.HP	0.0024	When all other predictors are held constant it is estimated that a one unit increase in Engine Horsepower is associated with a multiplicative change of $1.0024 in MSRP
Engine.Cylinders8	- 0.0763	When all other predictors are held constant it is estimated that when the Engine Cylinders are 8 there is a multiplicative change of $ - 1.0793 in MSRP
Transmission.TypeAUTOMATIC	0.0099	When all other predictors are held constant it is estimated that when the Transmission Type is Automatic there is a multiplicative change of $1.0099 in MSRP
Driven_Wheelsfour wheel drive	0.0632	When all other predictors are held constant it is estimated that when the Vehicle is Four Wheel Drive there is a multiplicative change of $1.0652 in MSRP
Vehicle.SizeLarge	0.1069	When all other predictors are held constant it is estimated that when the Vehicle Size is Large there is a multiplicative change of $1.1128 in MSRP
Vehicle.StyleConvertible	0.2052	When all other predictors are held constant it is estimated that when the Vehicle Style is Convertible there is a multiplicative change of $1.2278 in MSRP
ExoticNot Exotic	- 0.3752	When all other predictors are held constant it is estimated that when the vehicle is Not Exotic there is a multiplicative change of $1.4553 in MSRP
Year2017	0.1188	When all other predictors are held constant it is estimated that when the year is 2017 there is a multiplicative change of $1.1261 in MSRP

**Only a partial list of predictors is included for interpretation. There were multiple levels for most variables included. We decided to interpret only one of the levels for each of the 11 plus the intercept.*

Objective 2

Problem

1. Add complexity to the simple model to see if we can improve performance and fit
2. Build a nonparametric model

Approach

1. Researched plots to see what transformation or interactions could be added to improve model
    a. Added a squared term to horsepower variable
    b. Decided against interaction due to the large number of categorical variables contained in the dataset
2. Hypothesis Testing
3. Compared the Simple and Complex models to see performance differences
4. Researched and Ran Random Tree algorithm to generate a nonparametric model
    a. Decided on Random Tree due to the large number of categorical variables
5. Compared all three models for best fit of data

Compare Competing Models

Too add complexity to our model and see if we could improve performance, we decided to add a square to our horsepower variable. This seemed to help with the overfitting that our simple model seemed to have issues with (R^2 of 0.7921), however, in doing so we did see a dramatic increase in our RMSE statistic to 31,435.

$$\log(MSRP) = Transmission.Type + Driven_{Wheels} + Vehicle.Size + Hybrid + Hatchback + (Engine.HP)^2 + (Engine.HP)^3$$

Our final RMSE and R^2 values were 24,961 and 0.7919 respectively. This showed that our model was able to account for about 79% of the variability in MSRP which was close to what our results were on our training and test sets; however, we did see a decrease of $6,000 in RMSE between our training/test sets and our validation sets. This means that there is still a potential for our predictions to be plus or minus $24,961 of what our true median value of MSRP at any given point.

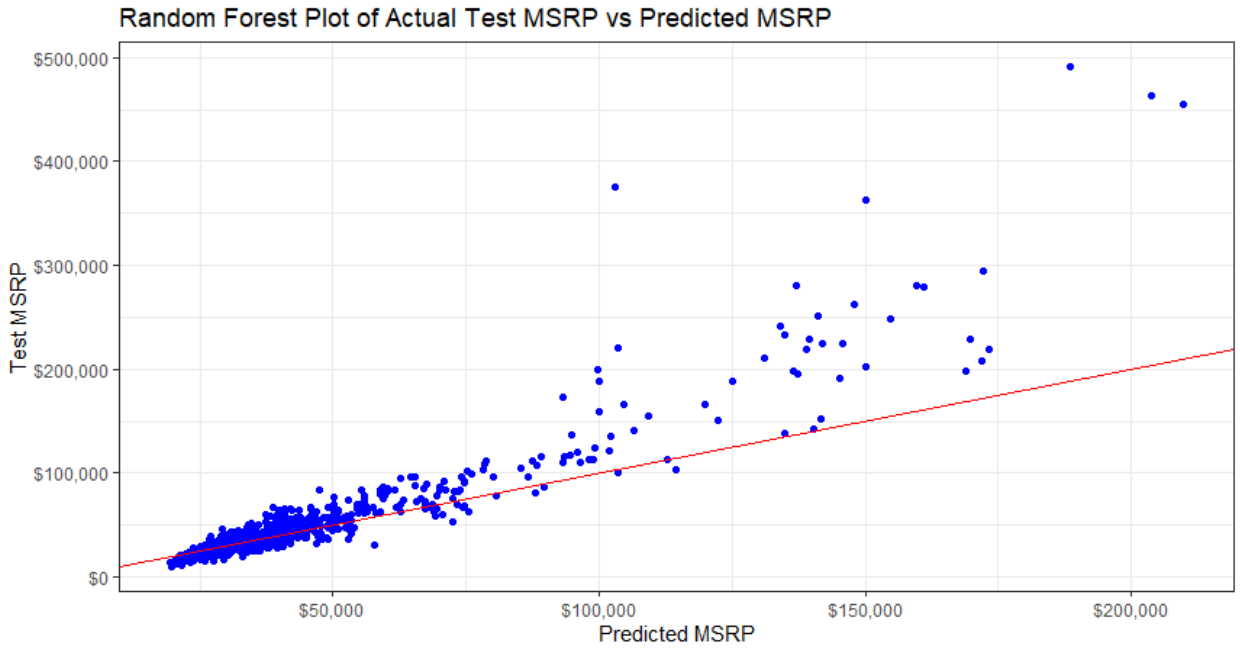
```
> Anova(model_afterlasso,type=3)
Anova Table (Type III tests)

Response: log1p(MSRP)
                             Sum Sq   Df    F value    Pr(>F)
(Intercept)                 19417.1    1 266828.339 < 2.2e-16 ***
Transmission.Type              46.9    3    214.824 < 2.2e-16 ***
Driven_Wheels                  21.7    3     99.233 < 2.2e-16 ***
Vehicle.Size                    3.0    2     20.565 1.233e-09 ***
I(Engine.HP^2)                 29.2    1    401.795 < 2.2e-16 ***
I(Engine.HP^3)                  1.2    1     16.208 5.727e-05 ***
I(Engine.HP^2):I(Engine.HP^3)   1.7    1     23.049 1.608e-06 ***
Residuals                     595.8 8188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> Anova(model_interaction,type=3)
Anova Table (Type III tests)

Response: log1p(MSRP)
                  Sum Sq   Df    F value    Pr(>F)
(Intercept)        33492    1 409605.137 < 2.2e-16 ***
I(Engine.HP^2)      1181    1  14449.044 < 2.2e-16 ***
Transmission.Type     46    3    186.279 < 2.2e-16 ***
Driven_Wheels         40    3    163.965 < 2.2e-16 ***
Vehicle.Size           4    2     26.594 3.073e-12 ***
Residuals            670 8190
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

With these results we find that the LASSO model provided a better overall fit for the data and a lower RMSE which will result in more accurate predictions for MSRP.

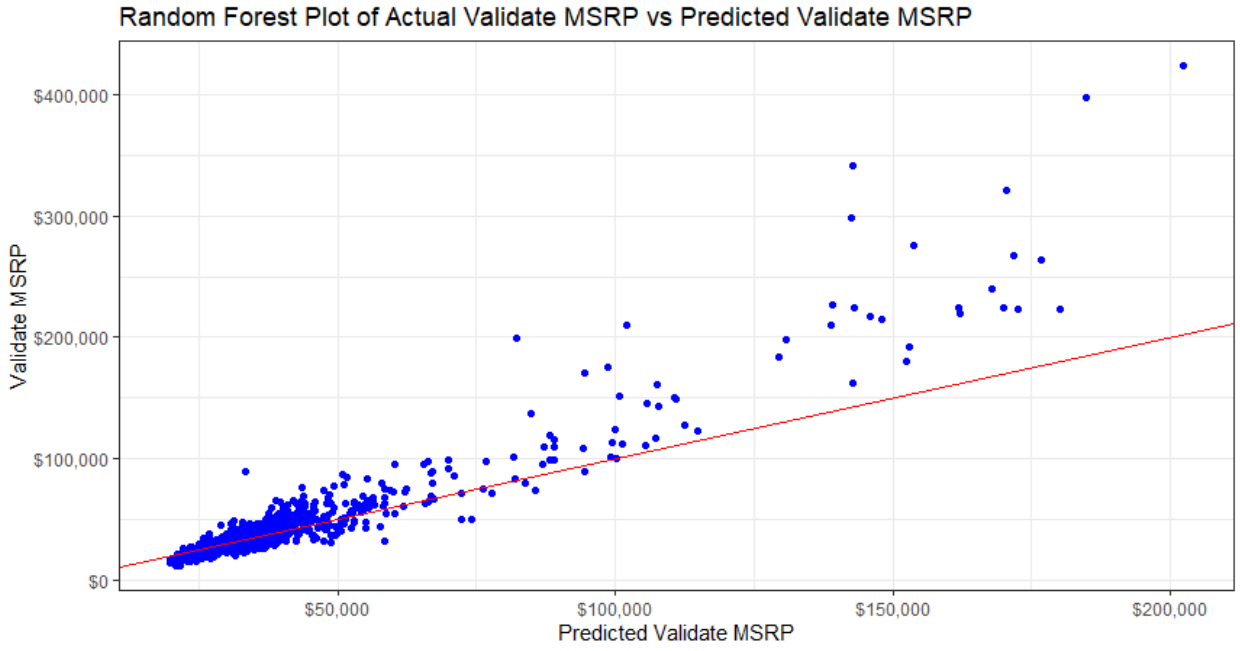
Nonparametric

Our nonparametric model was based off of a regression tree model. We felt this was the better fit for our data due to the high number of categorical predictors that we were working with. This model ended up being our highest performing model on the train/test set with an RMSE of 15,617 and an R^2 of 0.8874. This model did well at lowering our RMSE to a more tolerable level than our complex model while also lowering the R^2 statistic to a number that we felt like we weren't overfitting our data. Our final fit with the validation set provided us with an RMSE of 13,212 and an R^2 of 0.8966.

A Random Forest model was also generated which generated an RMSE of 24,879 and an R^2 of 0.8743 on the train/test set of data.



Our validation set gave us favorable results with an RMSE of 19,278 and an R^2 value of 0.8992.



Conclusion

Overall, we felt that our simple LASSO model, random tree model, and random forest model all predicted pretty well. We did decide though that for predicting MSRP on vehicles that contained predictors contained in our models the LASSO method performed best.

The LASSO model did well at keeping the deviation low ($10k) which in the scheme of vehicle sales, with such high MSRP numbers on some of the vehicles, is not very much. This model also didn't see a large drop in its R^2 value (0.92) when moving from the train/test data splits to the validate data split. This provides some reassurance that we are not over fitting the data too much and our model can be accurate when applied outside of this particular dataset.

Appendix

Data Description Figure 1.1

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.

***Above data shows the original variables in the dataset, the description discusses the transformations made to the data*

Original Data Statistics Figure 1.2

Summary Statistics Table		
Variable	**Statistics**	**Value**
Make	Length	11914
Make	Class	character
Make	Mode	character
Model	Length	11914

Model	Class	character
Model	Mode	character
Year	Min.	1990
Year	1st Qu.	2007
Year	Median	2015
Year	Mean	2010
Year	3rd Qu.	2016
Year	Max.	2017
Engine.Fuel.Type	Length	11914
Engine.Fuel.Type	Class	character
Engine.Fuel.Type	Mode	character
Engine.HP	Min.	55
Engine.HP	1st Qu.	170
Engine.HP	Median	227
Engine.HP	Mean	249.4
Engine.HP	3rd Qu.	300
Engine.HP	Max.	1001
Engine.HP	NA's	69
Engine.Cylinders	Min.	0
Engine.Cylinders	1st Qu.	4
Engine.Cylinders	Median	6
Engine.Cylinders	Mean	5.629
Engine.Cylinders	3rd Qu.	6
Engine.Cylinders	Max.	16
Engine.Cylinders	NA's	30
Transmission.Type	Length	11914
Transmission.Type	Class	character
Transmission.Type	Mode	character
Driven_Wheels	Length	11914
Driven_Wheels	Class	character
Driven_Wheels	Mode	character
Number.of.Doors	Min.	2
Number.of.Doors	1st Qu.	2
Number.of.Doors	Median	4
Number.of.Doors	Mean	3.436
Number.of.Doors	3rd Qu.	4
Number.of.Doors	Max.	4
Number.of.Doors	NA's	6
Market.Category	Length	11914
Market.Category	Class	character
Market.Category	Mode	character

Vehicle.Size	Length	11914
Vehicle.Size	Class	character
Vehicle.Size	Mode	character
Vehicle.Style	Length	11914
Vehicle.Style	Class	character
Vehicle.Style	Mode	character
highway.MPG	Min.	12
highway.MPG	1st Qu.	22
highway.MPG	Median	26
highway.MPG	Mean	26.64
highway.MPG	3rd Qu.	30
highway.MPG	Max.	354
city.mpg	Min.	7
city.mpg	1st Qu.	16
city.mpg	Median	18
city.mpg	Mean	19.73
city.mpg	3rd Qu.	22
city.mpg	Max.	137
Popularity	Min.	2
Popularity	1st Qu.	549
Popularity	Median	1385
Popularity	Mean	1555
Popularity	3rd Qu.	2009
Popularity	Max.	5657
MSRP	Min.	$2,000
MSRP	1st Qu.	$21,000
MSRP	Median	$29,995
MSRP	Mean	$40,595
MSRP	3rd Qu.	$42,231
MSRP	Max.	$2,065,902

Figure 1.2 Shows a cleaner set of summary statistics for each original variable.

Detail 1.1

Data types

8 of the 16 variables were originally cast as character type variables; we decided to update these types to factors to allow us to run better analysis on the levels of the variables. We felt that the year and number of doors variables made more sense as factors than continuous types. This allowed us to use them as categorical predictors that we could derive significant from while paired with other variables.

Detail 1.2

Missing and Wrong Values

The market category contains over three thousand missing values, after experimenting with how we could potentially use this variable and fill in missing values, we ultimately decided to remove the column as it would not be factored into any of our potential models due to its complexity.

We found that horsepower was missing 69 values (the graph shows 79, but that is due to our algorithm looking for values that are often used instead of NA such as 66 and 77), these values were relatively easy to find online and impute.

The cylinders variable contained 30 missing values, some of these were due to some models of vehicles being electric, or rotary engines and not having any cylinders. To fix this we changed the cylinders variable to a factor and added E as the value for electric vehicles and R for rotary vehicles.

The final missing values were the number of doors in for Tesla vehicles, these were added through searching for the correct values and manually adding them in.

Figure 1.3

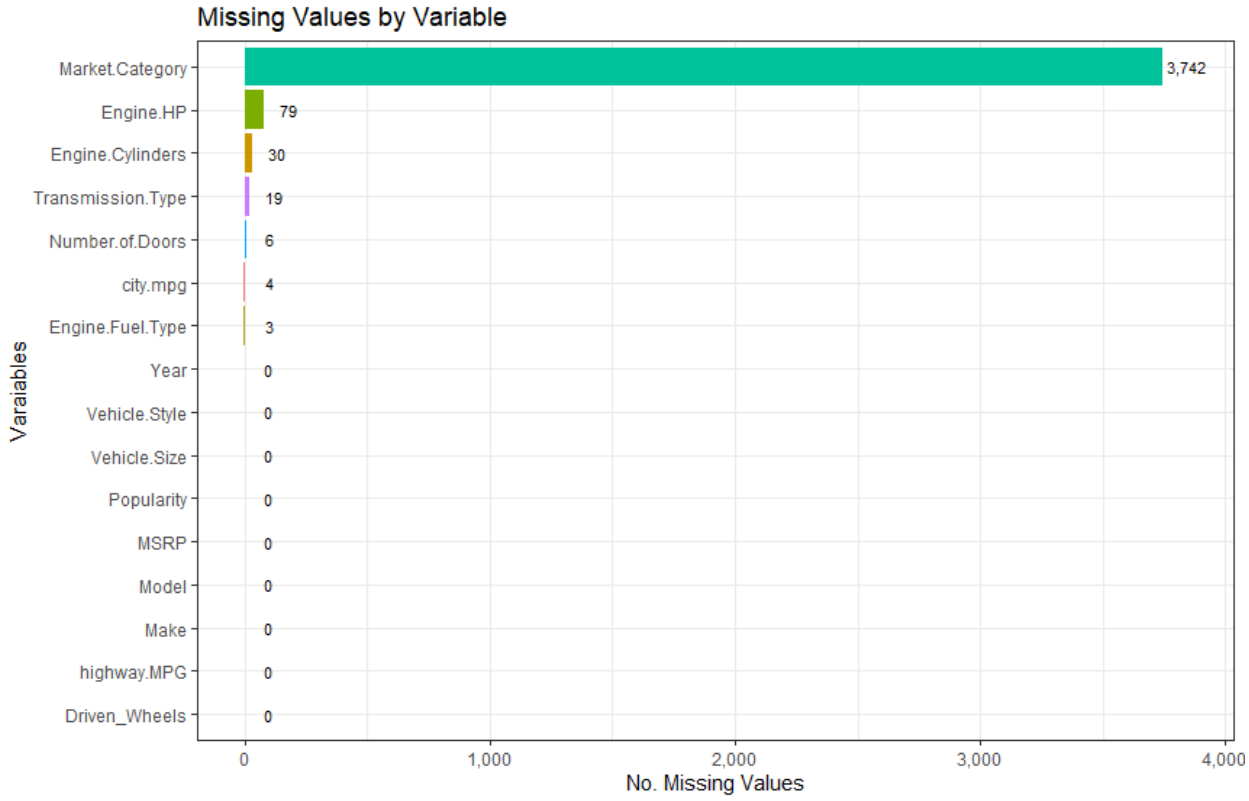


Figure 1.4

MSRP Min, Mean and Max Values			
Year	**Min(MSRP)**	**Mean(MSRP)**	**Max(MSRP)**
2000	$2,000	$2,530	$39,669
2001	$9,299	$41,501	$359,990
2002	$12,815	$33,744	$262,990
2003	$10,245	$39,774	$643,330
2004	$10,135	$36,107	$440,000
2005	$10,325	$36,747	$440,000
2006	$12,780	$32,056	$260,000
2007	$12,895	$34,974	$480,000
2008	$12,895	$44,175	$1,500,000
2009	$11,965	$45,458	$495,000
2010	$11,965	$50,642	$506,500
2011	$11,965	$57,548	$1,380,000
2012	$13,699	$59,516	$1,382,750
2013	$14,720	$48,699	$315,888
2014	$12,995	$63,226	$548,800
2015	$11,990	$46,794	$548,800
2016	$11,990	$47,221	$535,500
2017	$11,990	$42,192	$247,900

**MSRP statistics for model years newer than 1999*

Figure 1.5

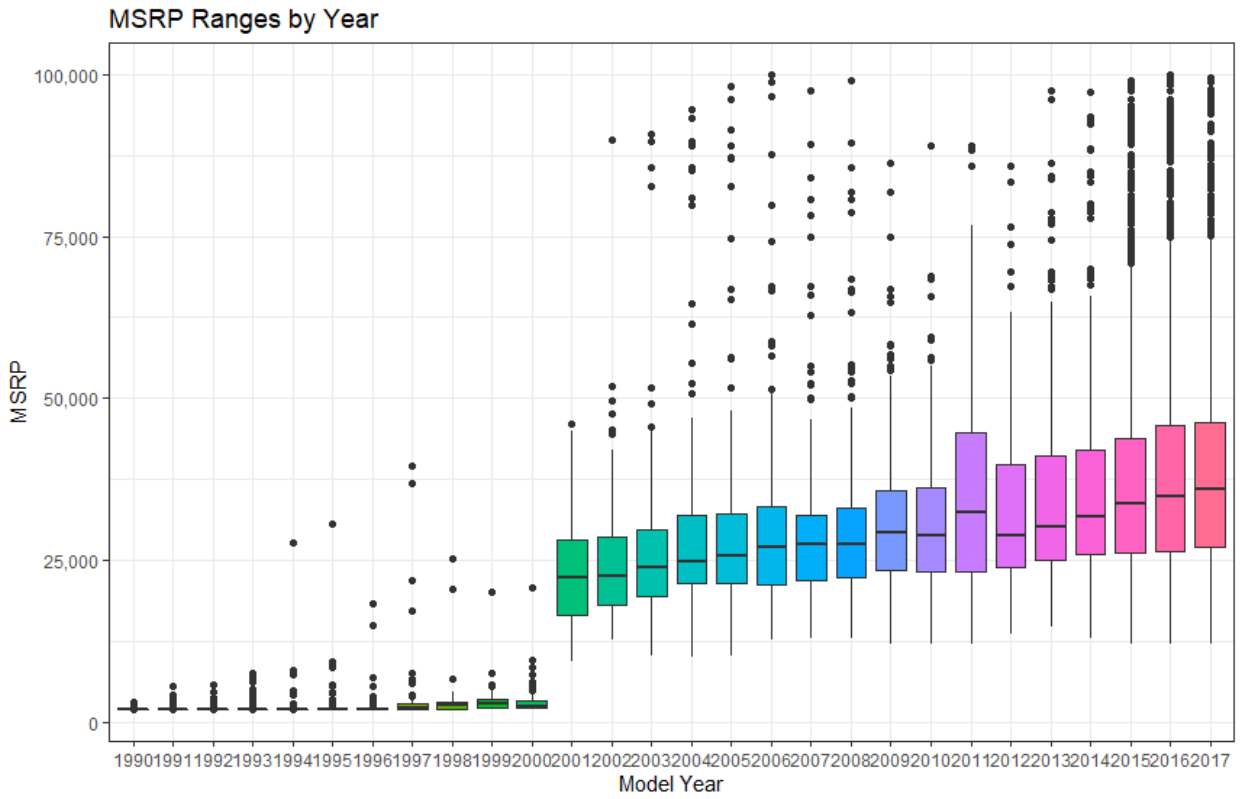


Figure 1.5 shows that model years older than 2001 increased the skewness of MSRP impacting the ability to provide an accurate prediciton.

Figure 1.6

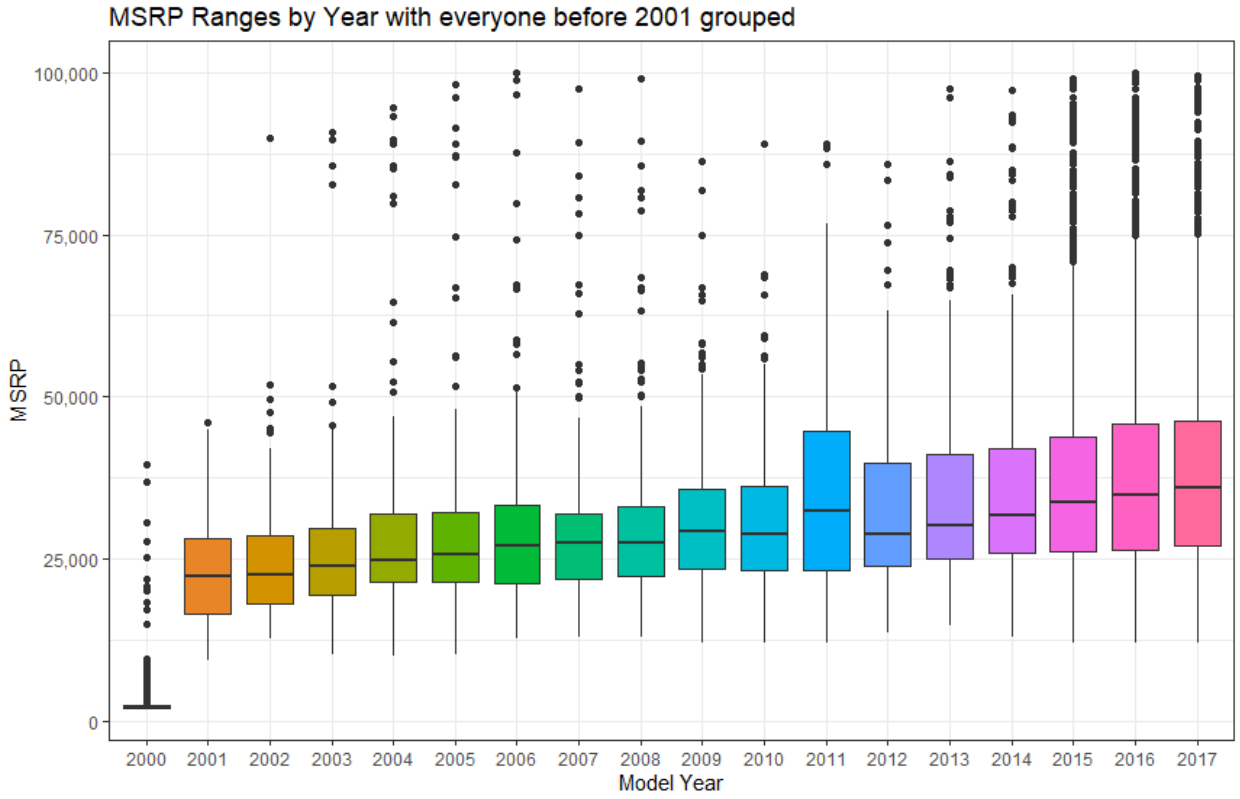


**Figure 1.6 shows all model years older than 2001 grouped into a single year. This decreases the skewness of MSRP, however those values were eventually dropped for analysis.*

Figure 1.7

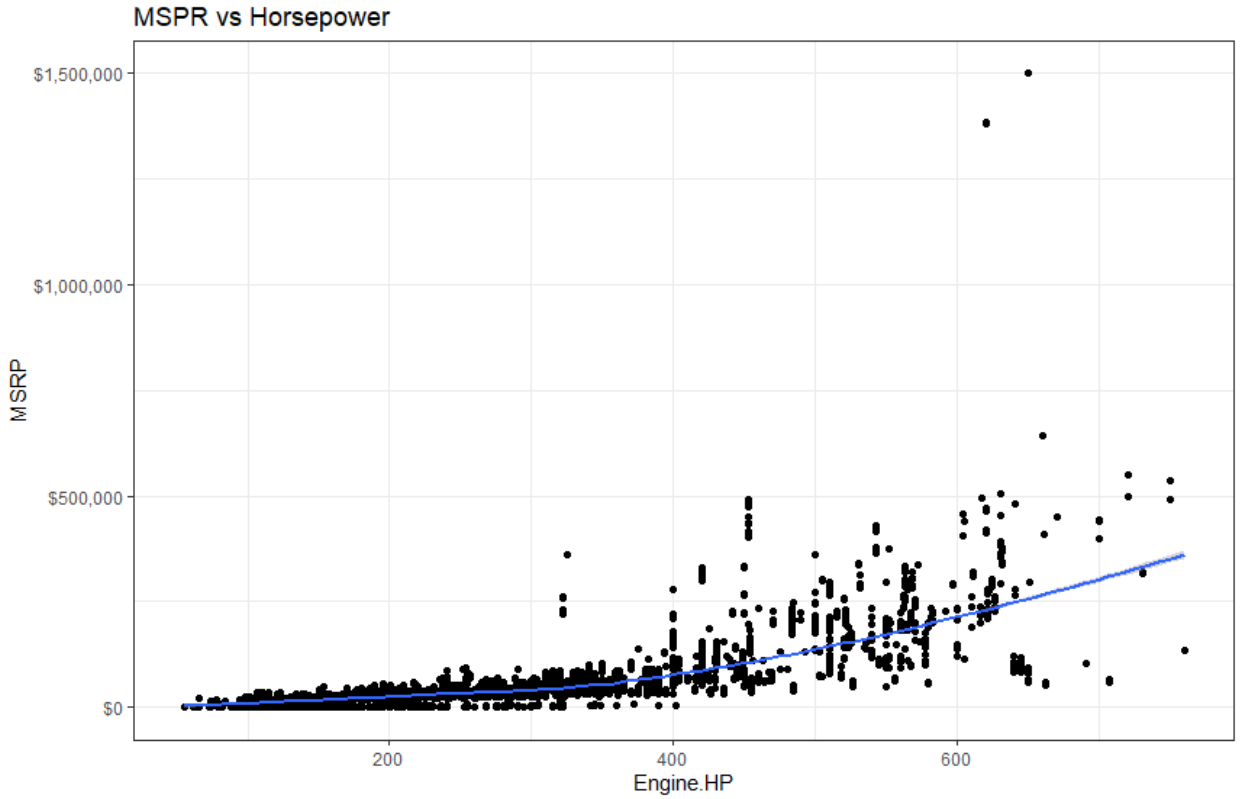


Figure 1.7 shows the correlation between MSRP and Engine.HP before limiting the maximum MSRP through data cleaning.

Figure 1.8

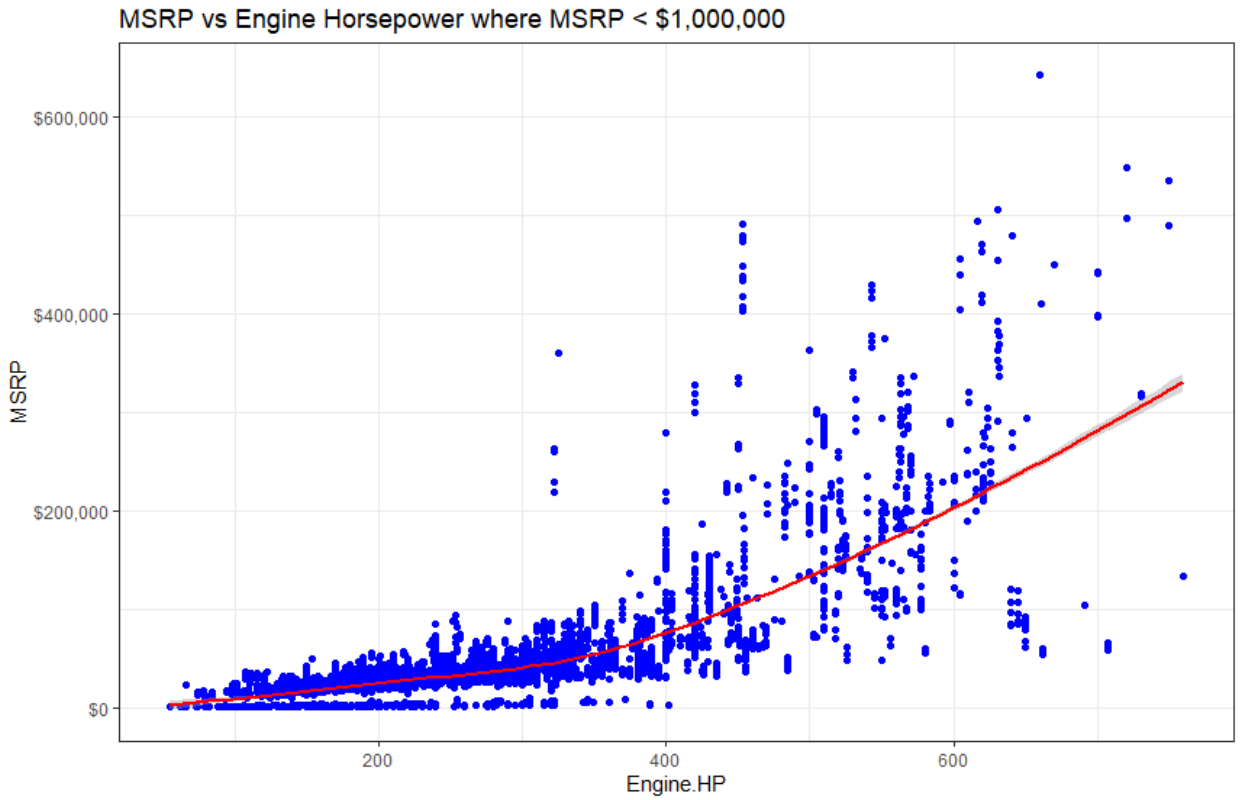


Figure 1.8 shows correlation between MSRP and Engine.HP For all MSRP values less than $1,000,000.

Figure 1.9

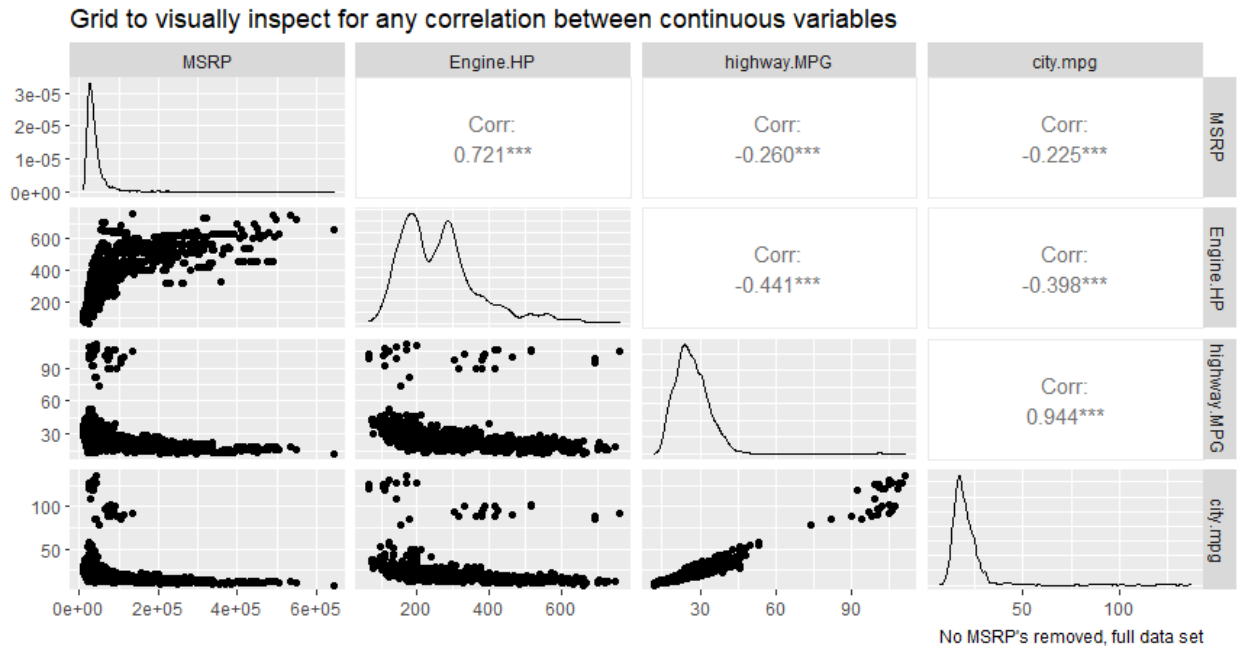


Figure 1.9 shows the realtionships between all the continuous variables in the data set after the data was cleaned.

Figure 1.10

**Before Log Transformation*

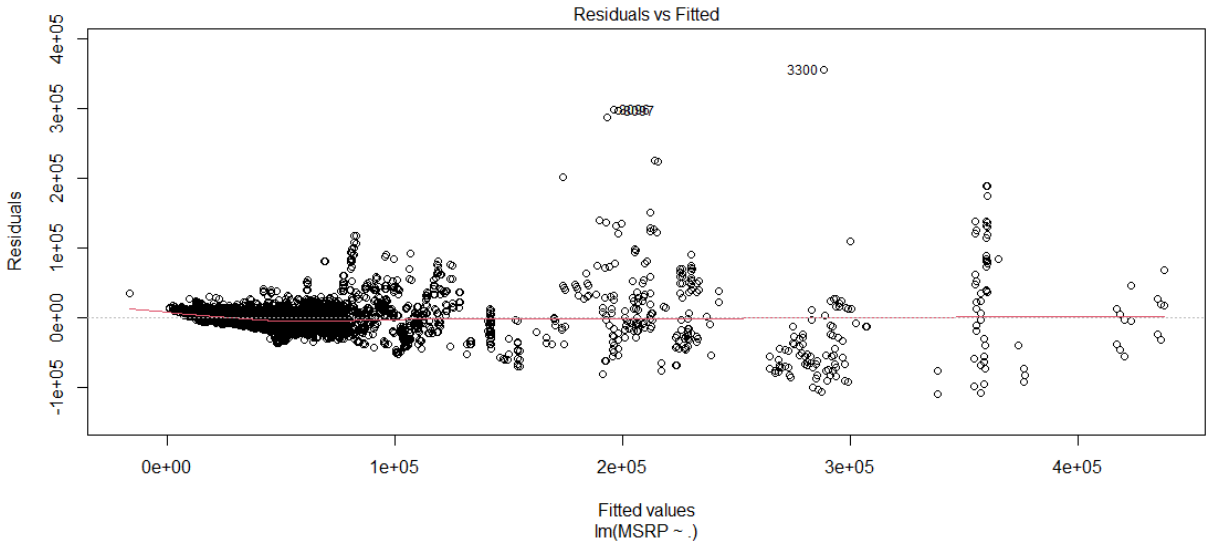
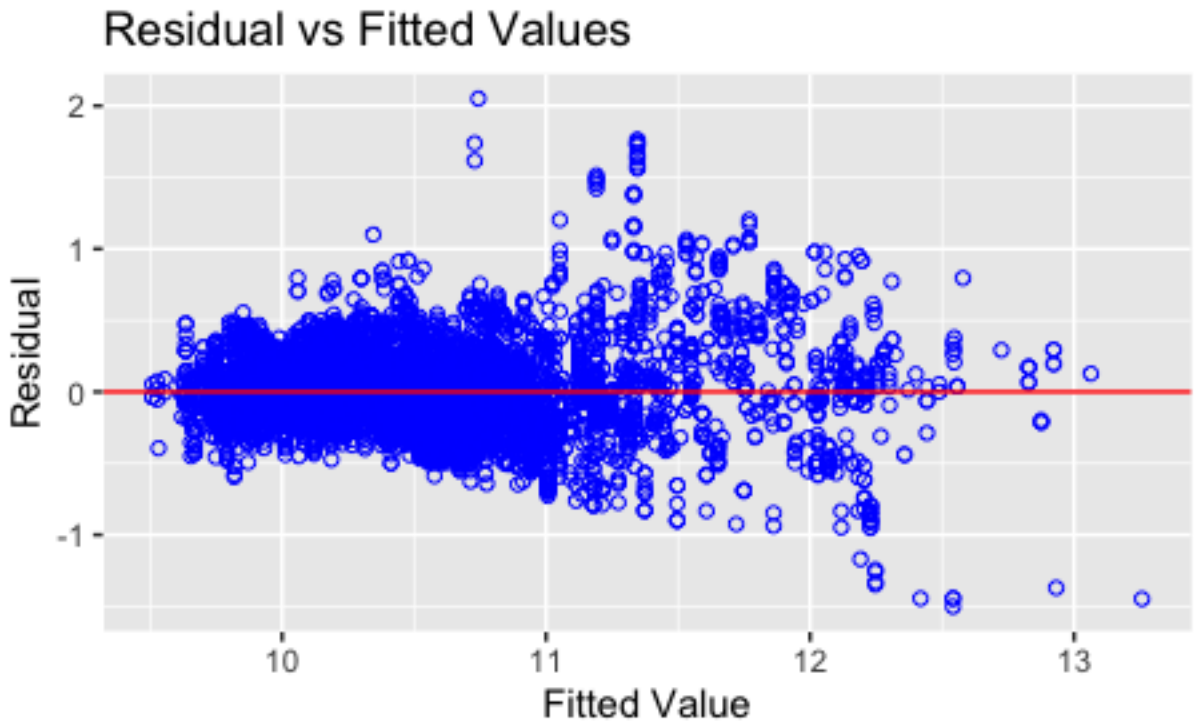


Figure 1.11



**Residuals after After Log Transformation of MSRP*

Link to Code and other useful Items (GitHub)
https://github.com/justinehly/6372-Auto-Pricing-Project

Figure 1.12

```
> aov.model<-aov(MSRP~.,data=autos)
> summary(aov.model)
                     Df    Sum Sq   Mean Sq   F value   Pr(>F)
Make                 46 2.004e+13 4.357e+11 13106.496  < 2e-16 ***
Model               695 4.926e+12 7.088e+09   213.225  < 2e-16 ***
Year                 16 5.848e+09 3.655e+08    10.996  < 2e-16 ***
Engine.Fuel.Type      9 1.204e+10 1.338e+09    40.243  < 2e-16 ***
Engine.HP             1 2.181e+11 2.181e+11  6561.340  < 2e-16 ***
Engine.Cylinders      6 5.661e+10 9.436e+09   283.863  < 2e-16 ***
Transmission.Type     3 7.163e+09 2.388e+09    71.830  < 2e-16 ***
Driven_Wheels         3 7.835e+09 2.612e+09    78.565  < 2e-16 ***
Number.of.Doors       2 3.917e+06 1.959e+06     0.059 0.942781
Vehicle.Size          2 4.843e+08 2.422e+08     7.285 0.000689 ***
Vehicle.Style        11 2.084e+10 1.894e+09    56.990  < 2e-16 ***
highway.MPG           1 1.071e+08 1.071e+08     3.221 0.072722 .
city.mpg              1 7.619e+06 7.619e+06     0.229 0.632125
Exotic                1 1.261e+10 1.261e+10   379.426  < 2e-16 ***
Residuals          9453 3.142e+11 3.324e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # doors, highway.MPG, city.mpg are all insig
> # model just has too many levels to work with.
```

**Doors, MPG’s all have insignificant p-values > 0.05.*