An Analysis of a Car's MSRP

Introduction

There are many factors that explain why a car's MSRP is priced the way it is. The purpose of this report is to explore the multitude of potential variables that best explains a car's worth. Two questions of interest will be explored in this study, first is the importance of a vehicle's popularity scores when it relates to MSRP. The second question that will be explored is predicting MSRP using the best possible model whether it is through linear regression or nonparametric models. This is an observational study so any conclusions that are made in this particular study are limited to the scope of the data collected and observed.

Data Description

The data used in this study is a car dataset that contains 16 variables such as Year, Make, Model, Popularity, MSRP and Vehicle Style along with a few others. This data is assumed to be privately collected and not available on a public source. There are 11,914 rows of information amongst the 16 different columns. A table including variable descriptions can be found in the appendix. Since this is an analysis of MSRP, this will be the response variable while the other 15 variables in this dataset are potential explanatory variables. Noticing working with the data there were multiple missing values that are not filled in or marked 'N/A' or filled in as 'Unknown.' Missing values are minimal since they are only present in 6 of the 16 different variables in the dataset. The market category variable is the only variable with a significant number of null values, 31%. (Figure 1.1). An exorbitant amount of tidying of the data is done before an explanatory analysis is done on the variables and determines any trends in the dataset.

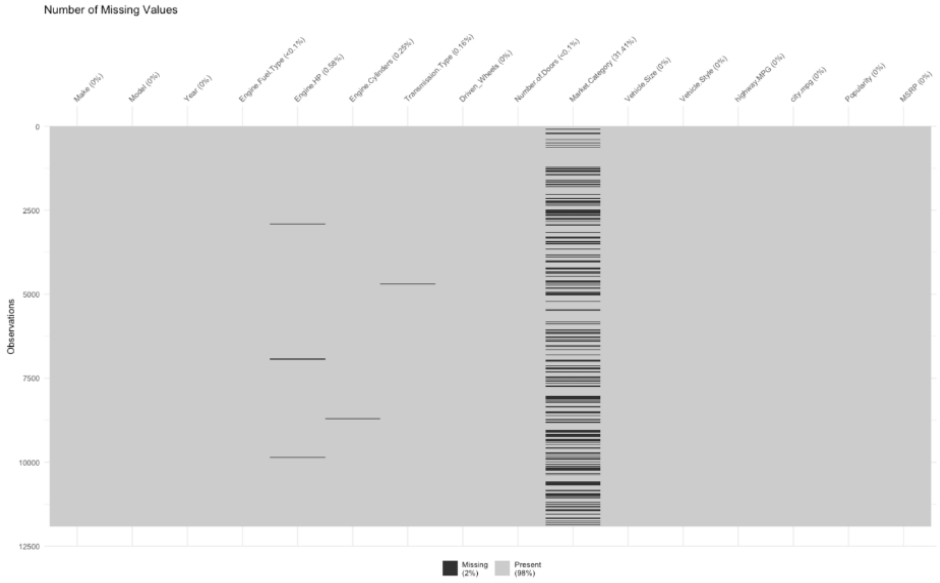


Figure 1.1

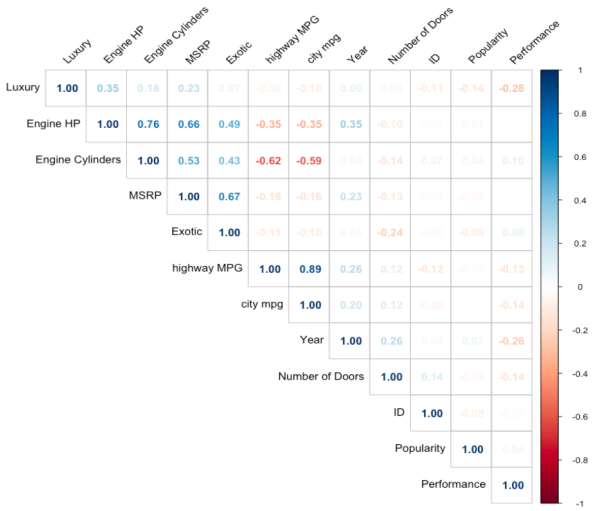
For many of the columns with null values (`Engine Fuel Type`, `Engine Cylinders`, etc.), imputation was successfully implemented using research into similar cars within the dataset. When there were no similar cars, quick internet research allowed us to find the appropriate values. However, `Market Category` was a uniquely difficult variable to deal with. Almost 4000

values were ‘N/A’. Because of the high variance within this variable, we decided to scrap it and create new variables, `Exotic` and `Luxury`, to replace it. Using a training and test split, we found that a simple Naive-Bayes model was successful at imputing `Exotic` for the unknown market categories (>94% accuracy). On the other hand, `Luxury` required a `Make`-specific approach to impute.

Exploratory Data Analysis (EDA)

The first step in our exploratory data analysis was to visualize the dataset by comparing each variable to MSRP in order to see the presence of patterns or relationships among variables.

We used a heat map to visualize the relationships between numerical variables and MSRP (Figure 1.2). The correlation heat map revealed the presence of multicollinearity in some independent variables. Multicollinearity was found between City MPG and Highway MPG as well as Engine Cylinders and Engine HP. Multicollinearity was solved by removing two variables, City MPG and Engine Cylinders.



Another interesting finding revealed by the correlation matrix above is the extremely low relationship between Popularity and any other variables in the model. A negative 0.048 correlation coefficient between Popularity and MSRP indicates a lack of relationship between them. After taking a closer look at popularity, something that stood up was the scale used to measure popularity. The lowest popularity scores are close to 0, while the highest scores are close to 5,000 (Figure 1.3). Additionally, the popularity scores jump from value to value, almost looking like a categorical variable instead of a continuous one.

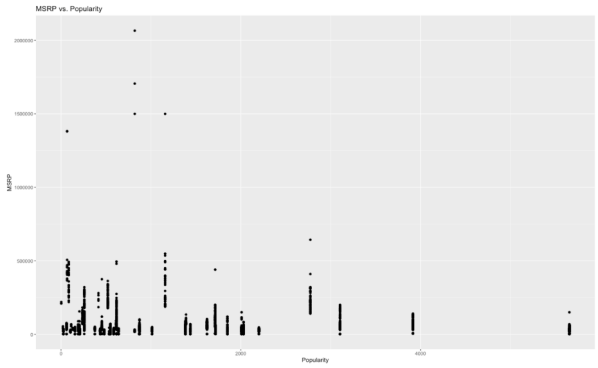
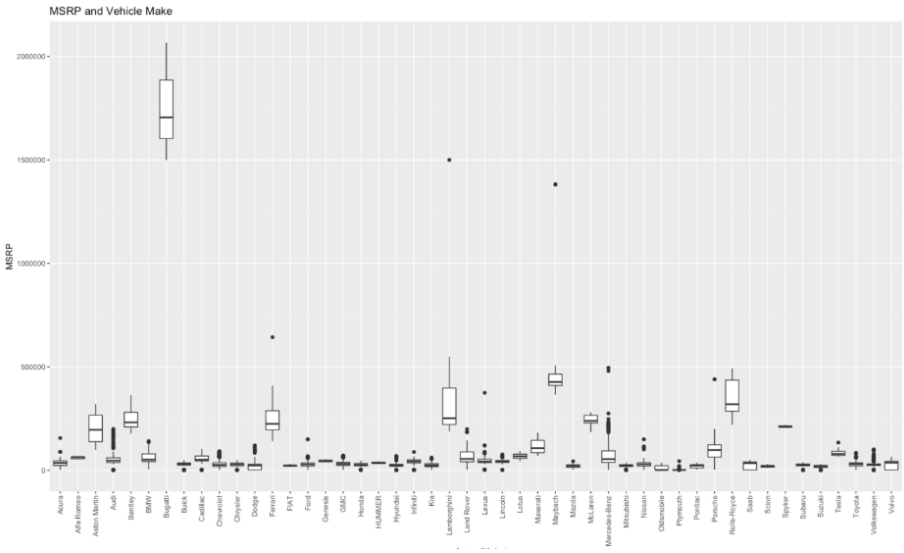


Figure 1.3

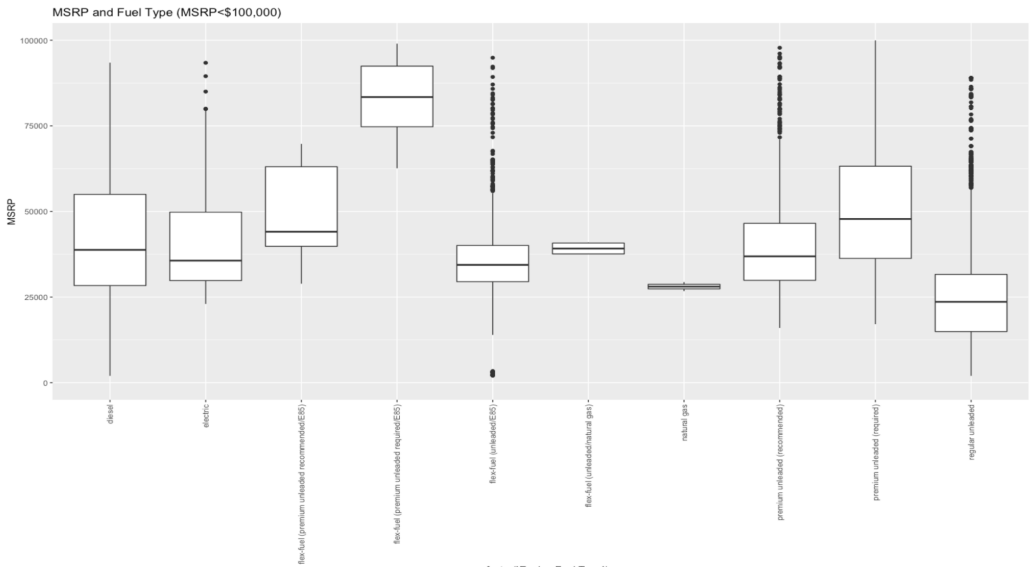
The next step in the exploratory data analysis was to visualize the relationship between the categorical variables in the dataset and MSRP. As seen on Figure 1.4, there are a few makes that are more influential on MSRP than others. Bugatti has a remarkably higher average MSRP than the rest of Makes in the dataset.

Figure 1.4



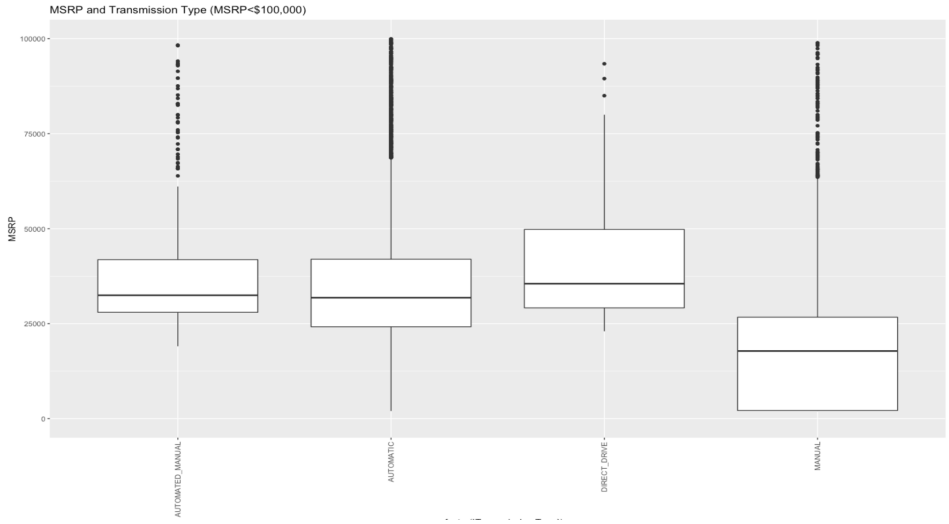
Different Fuel Types seem to have a similar average MSRP with the exception of two fuel types that have a significantly different average MSRP than the rest, “Flex-fuel (premium unleaded required)” with the highest average MSRP and “Regular Unleaded” with the lowest average MSRP. For scaling purposes, Figure 1.5, is only visualizing observations with a MSRP under $100,000 which make up around 95% of the dataset.

Figure 1.5



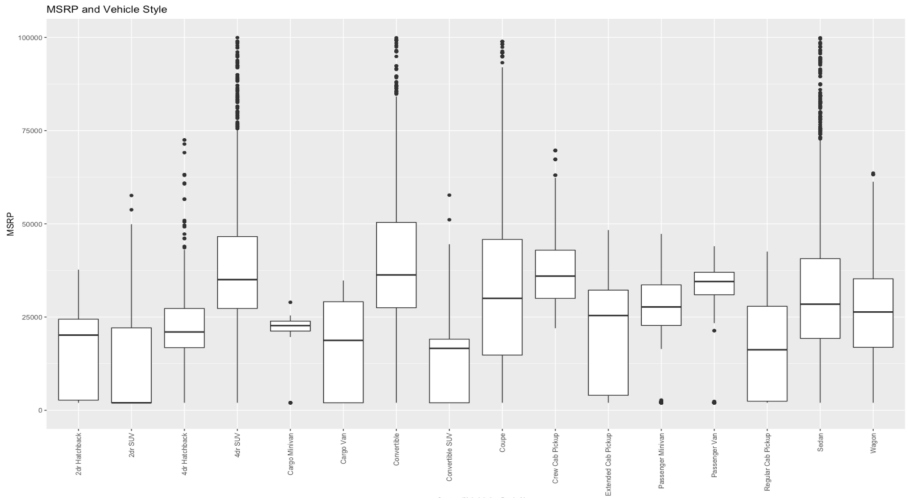
In the Transmission Type category cars under $100,000 that is labeled “Manual” has significantly lower average MSRP compared to the other transmission types.(Figure 1.6)

Figure 1.6



When it comes to vehicle styles, convertibles and coupe’s tend to have a higher MSRP than the other styles (Figure 1.7).

Figure 1.7



Questions of Interest:

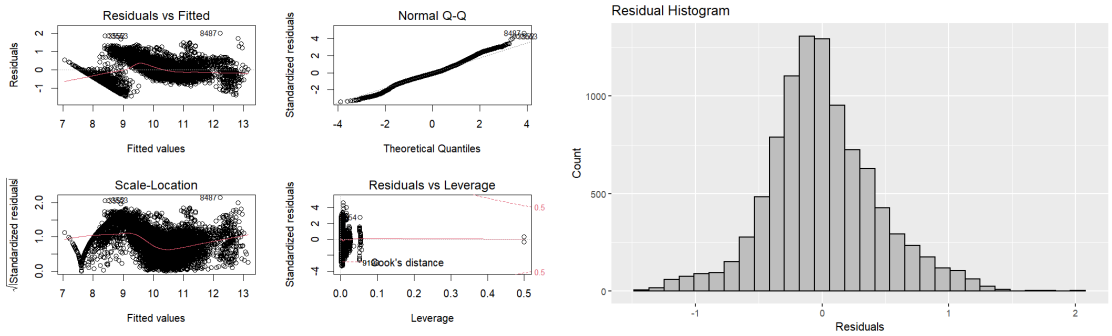
Objective 1

For the first question of interest, we were tasked with creating an interpretable model using the variables provided to predict MSRP, or retail price. Additionally, emphasis was placed on the importance of Popularity as it relates to MSRP.

Before building this model, data preprocessing and exploration had taken place. This allowed us to notice highly-correlated variables and alerted us to variables that would be difficult to work with. As a result, the variables ‘Engine Cylinders’ and ‘city mpg’ were removed from the dataset due to their high correlation with ‘Engine HP’ and ‘highway MPG’, respectively. Additionally, the variables ‘Make’, ‘Model’, and ‘Market Category’ were removed because of their high variance and dimensionality. Feeding them into the regression model would have required 100+ one-hot encoded variables or factors. Lastly, the ‘Year’ variable was turned into an ‘Age’ variable by simply calculating 2022-Year. After these changes, a regression model was fed with every independent variable aside from ID (an arbitrary identification column). One important note is that MSRP was log-transformed to meet the linear regression assumptions.

The next step was feature selection. We employed the stepAIC() function using stepwise selection to weed out insignificant predictors. Once a final model was output from this process, we observed the assumptions.

Figures 2.1 - 2.2 : Stepwise Selection Model Assumption Charts



As seen above, these assumptions look a bit risky, but passable. While there is some evidence of non-equal variance in the Residuals vs. Fitted chart, the Q-Q plot and residual histogram provided enough confidence to keep this model. Overall, there were few influential points, so the removal of them did not appear necessary.

Stepwise selection determined the most influential predictors of MSRP. Age, Exotic, and Engine HP are the top 3 dependent variables in the model (Appendix B). A 95% confidence interval was built to determine the relationship between each variable and MSRP (Appendix F). The confidence interval indicates that for each unit increase in Age, we are 95% confident that the MSRP decreases between 8.6% and 9.1% (log -0.09, log -0.0935). This makes sense based on the car industry as cars tend to lose value the older the car is. On the contrary, Luxury cars have a positive relationship with MSRP meaning that when a car is identified as Luxury (Luxury=1) we are 95% confident that the MSRP increases between 7.2% and 12.9% (log 0.0692, log .121). These estimates for the most influential predictors for MSRP are assuming all other variables in the model are held constant. From examining the data and domain knowledge about cars these four predictors are significant when determining MSRP. Whether it is considering the Year/Age the car is manufactured, the engine’s horsepower, or if the car is exotic or not can play an important role in what a particular individual is looking for. These

interpretations can be shared with the significant variables in the table below. More values for coefficient estimates and confidence intervals can be found in Appendix B and Appendix F.

Figure 2.3 - Regression Coefficient Estimates

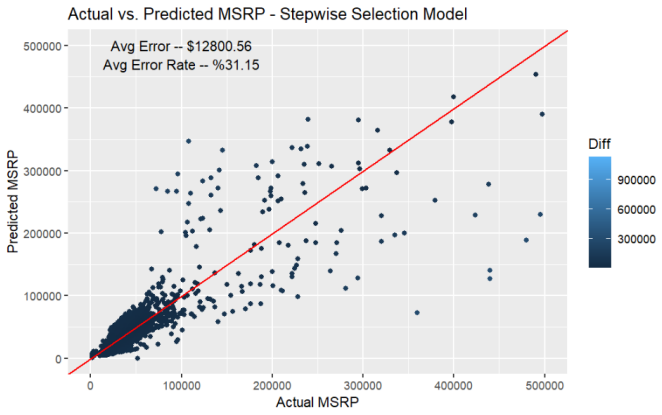
Variable	Estimate	Lower Conf. Interval	Upper Conf. Interval
Age	-0.0998	-0.0935	-0.09
Highway MPG	-.0133	-.0153	-.0112
Exotic	.9822	.9253	1.0391
Popularity	-.00001	-.00002	-.000005

Popularity's p-value is significantly larger compared to other variables in the model. While Popularity exists on a much larger scale than other variables such as `highway MPG` and `Exotic`, it's estimated coefficient is several magnitudes smaller than others within the model. Overall, Popularity tended to have statistical significance but no practical significance in the model with it being roughly zero. The popularity score given from social media did not really affect how a car's MSRP is determined. This conclusion is upheld by the essentially-zero correlation value between Popularity and MSRP.

Overall, the stepwise selection model did not perform to a satisfiable scale. With an average error rate exceeding 30% and an R-squared of 84.6%, more complicated models could perform much better. One cause of this lack of performance could be the proven difficulty of predicting high-MSRP cars. In the graph below, we can see that the average error (in USD) increases significantly as MSRP increases.

Figures 2.4 - 2.5 : Stepwise Model Accuracy Statistics and Error Graph

Test MSE (log scale)	.2409
Test MSE (USD scale)	2,306,091,868
Test R-squared	.8460



Objective 2

For the second part of this project, we were tasked with adding complexity to parametric and non-parametric regression models in the hopes of creating a more accurate MSRP prediction model. To accomplish this, our dataset was split into 80% training, 10% testing, and 10% validation.

Our first additional model utilizes Lasso to select the most important predictors within the dataset. A large part of this process is determining the most efficient parameter, known as lambda. Therefore, cross validation was used to hypertune the model.

Initially, a simple testing model that does not use cross validation was created. Below is a list of the variables that were selected using lasso, as well as a table of summary statistics from the lm() model utilizing these variables. A more detailed summary of the lm() model statistics is shown in the appendix.

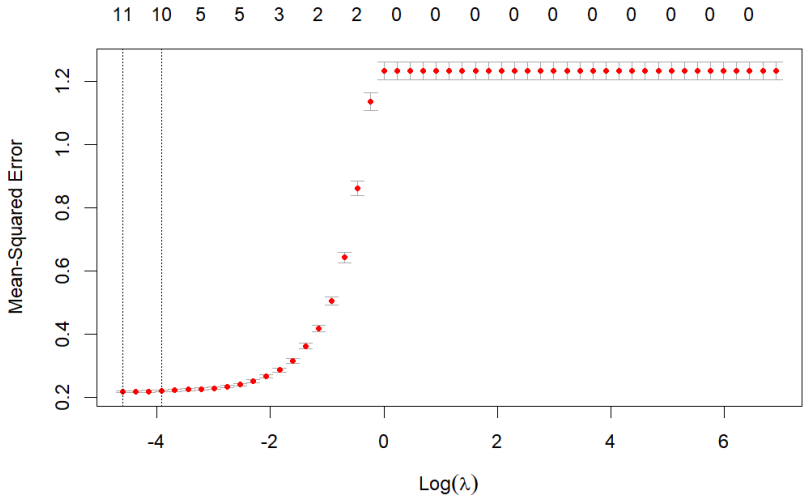
Figures 3.1 - 3.2 : Lasso Testing Model Coefficients and Accuracy Statistics

```
                              s1
(Intercept)          10.571986992559
Engine Fuel Type     -0.006612544511
Engine HP             0.003137233723
Transmission Type    -0.065938467409
Driven_Wheels        -0.006983906690
Number of Doors       .
Vehicle Size          .
Vehicle Style        -0.001538240953
highway MPG          -0.001733224076
Popularity           -0.000005652969
Exotic                1.046608911206
Luxury                0.169117438405
Age                  -0.088761899877
```

Test MSE (log scale)	.2188
Test MSE (US Dollar scale)	711,600,945
Adjusted R-squared	.8423

The next step was to utilize cross-validation to select the best-performing lambda. Below is a graph depicting the MSE as the lambda increases.

Figure 3.3 : Cross-Validation Lasso Lambda Chart



The two vertical lines highlight the placements of lambda.min and lambda.1se, the two best options for lambda. To test which option performed better, model statistics for both were calculated. The values of those statistics are included in the table below.

Figure 3.4 : Validated Lasso Model - Lambda Selection Criteria

	lambda.1se	lambda.min
MSE (log scale)	.2088	.2068
R-squared	.8206	.8227

Although both models performed similarly, lambda.min is marginally better. Therefore, this is the parameter we will use moving forward. Similar to the test model, below are the selected variables and lm() model statistics from the validated lambda.min model. A more detailed summary of the lm() model statistics is shown in the appendix.

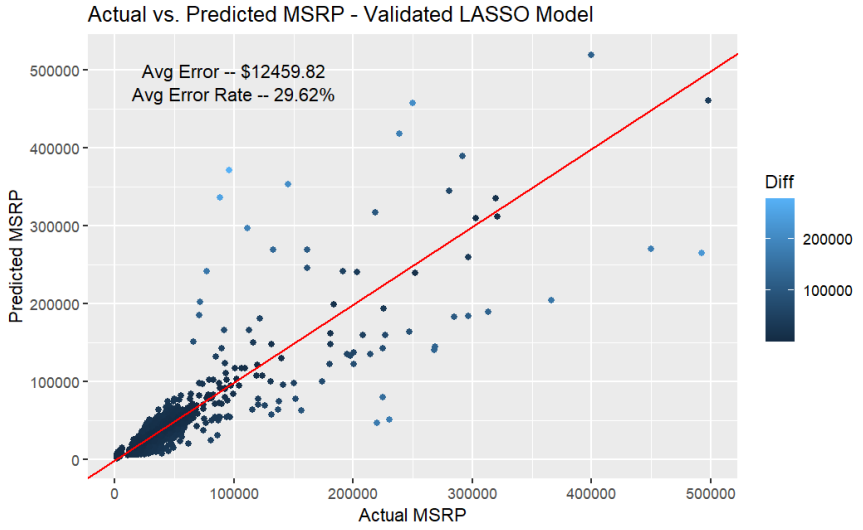
Figures 3.5 - 3.6 : Lasso Validated Model Coefficients and Accuracy Statistics

```
                              s1
(Intercept)       10.84303514202
Engine Fuel Type  -0.01533157809
Engine HP          0.00302128812
Transmission Type -0.07404173344
Driven_Wheels     -0.01493273978
Number of Doors    .
Vehicle Size      -0.01085341402
Vehicle Style     -0.00358593065
highway MPG       -0.00414968714
Popularity        -0.00001335189
Exotic             1.11386851407
Luxury             0.19302573901
Age               -0.09066836090
```

Validation MSE (log scale)	.2068
Validation MSE (US Dollar scale)	785,269,607
Adjusted R-squared	.8227

In comparison to the stepwise selection model, the validated lasso selection method is similar in Adj. R-squared, but the MSE is significantly better performing. Below is an actual vs. predicted MSRP graph for the predictions made from this model.

Figure 3.7 : Validated Lasso Accuracy Graph

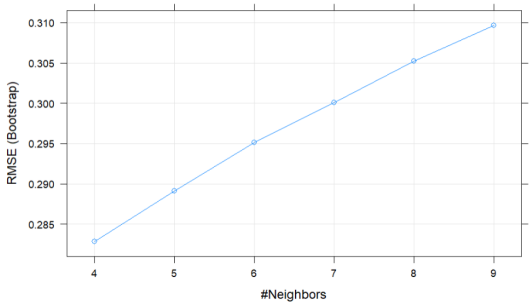


The farther an observation is from the red line, the less accurate the prediction is. It's extremely evident that low-MSRP cars are easier to predict, maybe because they are much more common. This model, on average, has an $12,459 error and a ~30% error rate. This is an improvement from the stepwise selection model graph.

Our next model utilizes Knn-regression. In short, this algorithm approximates the relationship between independent variables and the response value by combining observations within specific 'neighborhoods'. These neighborhoods have a size, K, that is chosen by the user. Similar to the Lasso process, we developed one 'test' model to hypertune this K value and then used a 'validate' dataset to determine the performance of the model with the hypertuned parameter.

After training a model over K's from 4 to 9, the testing predictions had the best performance at K=4. The performance of this testing model can be seen in the graph and table below.

Figures 3.9 - 3.10 : Testing Knn Regression Model Statistics



Test MSE (log scale)	.0799
Test MSE (US Dollar scale)	109,058,785
Adjusted R-squared	.9476

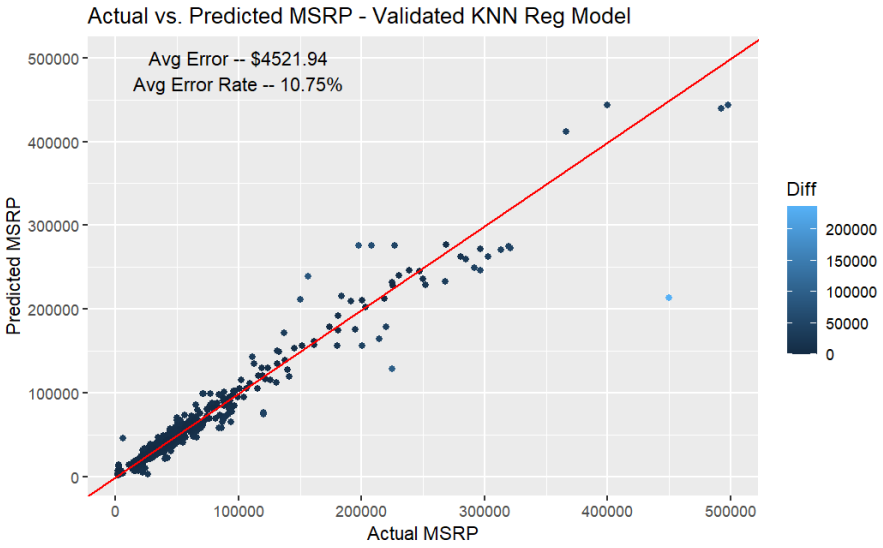
It is evident that this testing model is significantly better than any previous model within this project, having a Test MSE that is one seventh of the next lowest Test MSE (validation Lasso). Next, cross validation was used to hypertune the K parameter. Similar to the testing set, we found that K=4 was our best option. This validated model had similar results to the testing set, as both utilized the same parameters.

Figure 3.11 : Validated Knn Regression Model Statistics

Test MSE (log scale)	.0396
Test MSE (US Dollar scale)	133,047,991
Adjusted R-squared	.9476

Finally, an actual vs. predicted MSRP chart was created to visualize the accuracy of the predictions using this validated Knn regression model.

Figure 3.12 : Validated Knn Regression Model Accuracy Graph



In comparison with the previous models, the average error rate of ~11% is a massive improvement in performance. While this graph maintains that high-MSRP cars are the most difficult to predict, that effect seems much less significant here.

Conclusion

After testing and validating multiple MSRP-prediction models, the difference in performance metrics between models are significant. The table below lists each model's most important statistics.

Figure 4.1 : Model Summary Statistics

	Stepwise	Lasso	Knn
Test MSE (log scale)	.2409	.2188	.0799
Test MSE (USD scale)	2,306,091,868	711,600,945	109,058,785
Test R-squared	.8460	.8423	.9476
Validation MSE (log scale)	N/A	.2068	.0396
Validation MSE (USD scale)	N/A	785,269,607	133,047,991
Validation R-squared	N/A	.8227	.9476

Overall, each successive model improves within these performance metrics. While the stepwise & lasso models only explain ~84% of the variance in MSRP, the Knn regression model improves

to explaining almost 95% of the variance in MSRP. Additionally, the average difference in actual and predicted MSRP decreases by 63% between the stepwise model and the validated Knn model. We believe that the Knn regression model demonstrated the best results due to the fact the non-parametric method does not assume anything about the data it is using but instead tries to learn from the data overall. Non-parametric regression does require large sample sizes but the ability to adapt to the model structure and estimate to closely emulate the data.

Possible next steps could be exploring decision tree regression models, kernel regression, local regression and reexamining some of the independent variables. `Popularity' seems to have an arbitrary scale that lends itself to a distinct variable moreso than a continuous one. More detailed information on how Popularity is scored based on social media could create specific ranges to indicate most popular to least popular. Formula transparency would be a step in the right direction here, as there is no discernible relationship between `Popularity` and `MSRP`.

Appendix

The R code used in this project can be found in the project github repo. Project Github Repo Link: https://github.com/ericlaigaie/AppStat_Project1

Appendix A: Variable Types

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
Age	Numeric	2022 - Year
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.
Exotic	Numeric	A dummy variable (0 or 1) that describes if a vehicle is ‘Exotic’ or not.
Luxury	Numeric	A dummy variable (0 or 1) that describes if a vehicle is ‘Luxury’ or not.

Appendix B: Detailed summary of the stepwise test coefficients:

Knit on Save | Knit

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.45962 -0.25911 -0.03382  0.24775  1.98583

Coefficients:
                                                        Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                                         10.956880189  0.090103413 121.603 < 0.0000000000000002 ***
`Engine HP`                                          0.001672264  0.000088187  18.963 < 0.0000000000000002 ***
`Number of Doors`                                    0.105029498  0.015850974   6.626  0.00000000003635657 ***
`highway MPG`                                       -0.031716784  0.001576468 -20.119 < 0.0000000000000002 ***
Popularity                                          -0.000009010  0.000003292  -2.737              0.00621 **
Exotic                                               0.960522880  0.029055691  33.058 < 0.0000000000000002 ***
Luxury                                               0.095266114  0.013246940   7.192  0.00000000000068919 ***
Age                                                 -0.099811368  0.001026149 -97.268 < 0.0000000000000002 ***
`Engine Fuel Type_diesel`                            0.504823207  0.041614598  12.131 < 0.0000000000000002 ***
`Engine Fuel Type_electric`                          1.985169736  0.324122491   6.125  0.00000000094442978 ***
`Engine Fuel Type_flex-fuel (premium unleaded recommended/E85)` 0.343862717  0.100557107   3.420              0.00063 ***
`Engine Fuel Type_flex-fuel (premium unleaded required/E85)`    0.326168550  0.067681371   4.819  0.00000146394463129 ***
`Engine Fuel Type_flex-fuel (unleaded/E85)`          0.039692065  0.019514769   2.034              0.04198 *
`Engine Fuel Type_premium unleaded (recommended)`    0.126886555  0.016230262   7.818  0.00000000000000595 ***
`Engine Fuel Type_premium unleaded (required)`       0.413841225  0.018132257  22.823 < 0.0000000000000002 ***
`Transmission Type_AUTOMATED_MANUAL`                 0.167381177  0.022996254   7.279  0.00000000000036411 ***
`Transmission Type_AUTOMATIC`                        0.196560792  0.013371256  14.700 < 0.0000000000000002 ***
`Transmission Type_DIRECT_DRIVE`                     0.493393150  0.306351739   1.611              0.10731
`Driven_Wheels_all wheel drive`                      0.036573937  0.014976118   2.442              0.01462 *
`Driven_Wheels_four wheel drive`                     0.051439584  0.017509471   2.938              0.00331 **
`Driven_Wheels_front wheel drive`                    0.078397229  0.015947506   4.916  0.00000089835537476 ***
`Vehicle Size_Compact`                               0.049160500  0.012705694   3.869              0.00011 ***
`Vehicle Size_Large`                                 0.040947214  0.014047848   2.915              0.00357 **
`Vehicle Style_2dr Hatchback`                        0.166971207  0.038224601   4.368  0.00001266324956036 ***
`Vehicle Style_4dr Hatchback`                       -0.073891385  0.026912642  -2.746              0.00605 **
`Vehicle Style_4dr SUV`                             -0.052218367  0.021139894  -2.470              0.01352 *
`Vehicle Style_Cargo Van`                           -0.531792841  0.057669879  -9.221 < 0.0000000000000002 ***
`Vehicle Style_Convertible`                          0.342722753  0.036790830   9.315 < 0.0000000000000002 ***
`Vehicle Style_Convertible SUV`                      0.435722778  0.101604486   4.288  0.00001817351809067 ***
`Vehicle Style_Coupe`                                0.205584499  0.034984309   5.876  0.00000000433257464 ***
`Vehicle Style_Crew Cab Pickup`                     -0.118350121  0.028843573  -4.103  0.00004109448961814 ***
`Vehicle Style_Extended Cab Pickup`                 -0.218134904  0.028660139  -7.611  0.00000000000002978 ***
`Vehicle Style_Passenger Minivan`                    0.053019027  0.030120023   1.760              0.07840 .
`Vehicle Style_Passenger Van`                       -0.367130254  0.051309635  -7.155  0.00000000000089769 ***
`Vehicle Style_Regular Cab Pickup`                  -0.086467854  0.040284534  -2.146              0.03186 *
`Vehicle Style_Sedan`                               -0.046704130  0.021103597  -2.213              0.02692 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.432 on 9457 degrees of freedom
Multiple R-squared:  0.8474,    Adjusted R-squared:  0.8469
F-statistic:  1501 on 35 and 9457 DF,  p-value: < 0.00000000000000022
```

Some Model Stuff

Appendix C: Detailed summary of the lasso test coefficients:

```
Coefficients:
                                                      Estimate  Std. Error  t value            Pr(>|t|)
(Intercept)                                         11.06999023  0.07829625  141.386 < 0.0000000000000002 ***
`Engine Fuel Type`electric                           0.57683993  0.32252427    1.789             0.073718 .
`Engine Fuel Type`flex-fuel (premium unleaded recommended/E85) -0.09205310  0.09452363   -0.974             0.330144
`Engine Fuel Type`flex-fuel (premium unleaded required/E85)    -0.16002687  0.07267110   -2.202             0.027679 *
`Engine Fuel Type`flex-fuel (unleaded/E85)                     -0.38995295  0.04088501   -9.538 < 0.0000000000000002 ***
`Engine Fuel Type`flex-fuel (unleaded/natural gas)             -0.51175819  0.18418580   -2.778             0.005470 **
`Engine Fuel Type`natural gas                                  -0.12296038  0.31349212   -0.392             0.694897
`Engine Fuel Type`premium unleaded (recommended)               -0.27961685  0.03837528   -7.286 0.000000000000338551 ***
`Engine Fuel Type`premium unleaded (required)                   0.00308891  0.03917098    0.079             0.937148
`Engine Fuel Type`regular unleaded                             -0.42183582  0.03744236  -11.266 < 0.0000000000000002 ***
`Engine HP`                                                     0.00219890  0.00007388   29.762 < 0.0000000000000002 ***
`Transmission Type`AUTOMATIC                                    0.03388603  0.02056855    1.647             0.099488 .
`Transmission Type`DIRECT_DRIVE                                 0.16102838  0.31202769    0.516             0.605815
`Transmission Type`MANUAL                                      -0.17062667  0.02101379   -8.120 0.000000000000000513 ***
Driven_Wheelsfour wheel drive                                   0.05464757  0.01798994    3.038             0.002389 **
Driven_Wheelsfront wheel drive                                  0.02054822  0.01374078    1.495             0.134832
Driven_Wheelsrear wheel drive                                  -0.02954886  0.01366538   -2.162             0.030614 *
`Number of Doors`                                               0.12573046  0.01862392    6.751 0.000000000015364700 ***
`Vehicle Size`Large                                            -0.00277590  0.01488291   -0.187             0.852043
`Vehicle Size`Midsize                                          -0.04470224  0.01156184   -3.866             0.000111 ***
`Vehicle Style`2dr SUV                                         -0.10159525  0.04501066   -2.257             0.024018 *
`Vehicle Style`4dr Hatchback                                   -0.29862064  0.04385664   -6.809 0.000000000010300293 ***
`Vehicle Style`4dr SUV                                         -0.18507670  0.04244681   -4.360 0.000013103667671580 ***
`Vehicle Style`Cargo Minivan                                   -0.16624051  0.06389826   -2.602             0.009289 **
`Vehicle Style`Cargo Van                                       -0.55410356  0.05557632   -9.970 < 0.0000000000000002 ***
`Vehicle Style`Convertible                                      0.19359977  0.02665657    7.263 0.000000000000403021 ***
`Vehicle Style`Convertible SUV                                  0.23736300  0.08517646    2.787             0.005333 **
`Vehicle Style`Coupe                                            0.01181708  0.02508390    0.471             0.637577
`Vehicle Style`Crew Cab Pickup                                 -0.24051859  0.04586904   -5.244 0.000000160184661288 ***
`Vehicle Style`Extended Cab Pickup                             -0.30248464  0.03983645   -7.593 0.000000000000033555 ***
`Vehicle Style`Passenger Minivan                               -0.07421999  0.04442729   -1.671             0.094828 .
`Vehicle Style`Passenger Van                                   -0.35276054  0.05153833   -6.845 0.000000000008044975 ***
`Vehicle Style`Regular Cab Pickup                              -0.17166409  0.03361396   -5.107 0.000000332512768327 ***
`Vehicle Style`Sedan                                           -0.26138739  0.04214747   -6.202 0.000000000576993608 ***
`Vehicle Style`Wagon                                           -0.16393619  0.04503696   -3.640             0.000274 ***
`highway MPG`                                                  -0.01501610  0.00099670  -15.066 < 0.0000000000000002 ***
Popularity                                                     -0.00001135  0.00000302   -3.760             0.000171 ***
Exotic                                                          0.98122961  0.02624970   37.381 < 0.0000000000000002 ***
Luxury                                                          0.09102822  0.01200699    7.581 0.000000000000036763 ***
Age                                                            -0.09258641  0.00083145 -111.355 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4399 on 11874 degrees of freedom
Multiple R-squared:  0.8423,    Adjusted R-squared:  0.8417
F-statistic:  1626 on 39 and 11874 DF,  p-value: < 0.00000000000000022
```

Appendix D: Detailed summary of the Lasso validate coefficients:

```
Coefficients:
                                                    Estimate  Std. Error  t value             Pr(>|t|)
(Intercept)                                       11.06999023  0.07829625  141.386 < 0.0000000000000002 ***
`Engine Fuel Type`electric                         0.57683993  0.32252427    1.789             0.073718 .
`Engine Fuel Type`flex-fuel (premium unleaded recommended/E85) -0.09205310  0.09452363   -0.974             0.330144
`Engine Fuel Type`flex-fuel (premium unleaded required/E85)    -0.16002687  0.07267110   -2.202             0.027679 *
`Engine Fuel Type`flex-fuel (unleaded/E85)                     -0.38995295  0.04088501   -9.538 < 0.0000000000000002 ***
`Engine Fuel Type`flex-fuel (unleaded/natural gas)             -0.51175819  0.18418580   -2.778             0.005470 **
`Engine Fuel Type`natural gas                                  -0.12296038  0.31349212   -0.392             0.694897
`Engine Fuel Type`premium unleaded (recommended)               -0.27961685  0.03837528   -7.286 0.000000000000338551 ***
`Engine Fuel Type`premium unleaded (required)                   0.00308891  0.03917098    0.079             0.937148
`Engine Fuel Type`regular unleaded                             -0.42183582  0.03744236  -11.266 < 0.0000000000000002 ***
`Engine HP`                                                     0.00219890  0.00007388   29.762 < 0.0000000000000002 ***
`Transmission Type`AUTOMATIC                                    0.03388603  0.02056855    1.647             0.099488 .
`Transmission Type`DIRECT_DRIVE                                 0.16102838  0.31202769    0.516             0.605815
`Transmission Type`MANUAL                                      -0.17062667  0.02101379   -8.120 0.000000000000000513 ***
Driven_Wheelsfour wheel drive                                   0.05464757  0.01798994    3.038             0.002389 **
Driven_Wheelsfront wheel drive                                  0.02054822  0.01374078    1.495             0.134832
Driven_Wheelsrear wheel drive                                  -0.02954886  0.01366538   -2.162             0.030614 *
`Number of Doors`                                               0.12573046  0.01862392    6.751 0.000000000015364700 ***
`Vehicle Size`Large                                            -0.00277590  0.01488291   -0.187             0.852043
`Vehicle Size`Midsize                                          -0.04470224  0.01156184   -3.866             0.000111 ***
`Vehicle Style`2dr SUV                                         -0.10159525  0.04501066   -2.257             0.024018 *
`Vehicle Style`4dr Hatchback                                   -0.29862064  0.04385664   -6.809 0.000000000010300293 ***
`Vehicle Style`4dr SUV                                         -0.18507670  0.04244681   -4.360 0.000013103667671580 ***
`Vehicle Style`Cargo Minivan                                   -0.16624051  0.06389826   -2.602             0.009289 **
`Vehicle Style`Cargo Van                                       -0.55410356  0.05557632   -9.970 < 0.0000000000000002 ***
`Vehicle Style`Convertible                                      0.19359977  0.02665657    7.263 0.000000000000403021 ***
`Vehicle Style`Convertible SUV                                  0.23736300  0.08517646    2.787             0.005333 **
`Vehicle Style`Coupe                                            0.01181708  0.02508390    0.471             0.637577
`Vehicle Style`Crew Cab Pickup                                 -0.24051859  0.04586904   -5.244 0.000000160184661288 ***
`Vehicle Style`Extended Cab Pickup                             -0.30248464  0.03983645   -7.593 0.000000000000033555 ***
`Vehicle Style`Passenger Minivan                               -0.07421999  0.04442729   -1.671             0.094828 .
`Vehicle Style`Passenger Van                                   -0.35276054  0.05153833   -6.845 0.000000000008044975 ***
`Vehicle Style`Regular Cab Pickup                              -0.17166409  0.03361396   -5.107 0.000000332512768327 ***
`Vehicle Style`Sedan                                           -0.26138739  0.04214747   -6.202 0.000000000576993608 ***
`Vehicle Style`Wagon                                           -0.16393619  0.04503696   -3.640             0.000274 ***
`highway MPG`                                                  -0.01501610  0.00099670  -15.066 < 0.0000000000000002 ***
Popularity                                                     -0.00001135  0.00000302   -3.760             0.000171 ***
Exotic                                                          0.98122961  0.02624970   37.381 < 0.0000000000000002 ***
Luxury                                                          0.09102822  0.01200699    7.581 0.000000000000036763 ***
Age                                                            -0.09258641  0.00083145 -111.355 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4399 on 11874 degrees of freedom
Multiple R-squared:  0.8423,    Adjusted R-squared:  0.8417
F-statistic:  1626 on 39 and 11874 DF,  p-value: < 0.00000000000000022
```

Appendix E: Determining K in the KNN regression for the test and validate models:

```
[1] "TEST KNN MODEL RESULTS"
k-Nearest Neighbors

9527 samples
  12 predictor

Pre-processing: centered (12), scaled (12)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 9527, 9527, 9527, 9527, 9527, 9527, ...
Resampling results across tuning parameters:

  k  RMSE       Rsquared   MAE
  5  0.2928825  0.9301862  0.1409181
  7  0.3056601  0.9238755  0.1519943
  9  0.3149530  0.9191118  0.1607853

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 5.
```

```
[1] "VALIDATE KNN MODEL RESULTS"
k-Nearest Neighbors

9527 samples
  12 predictor

Pre-processing: centered (12), scaled (12)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 8573, 8574, 8575, 8575, 8576, 8572, ...
Resampling results across tuning parameters:

  k  RMSE       Rsquared   MAE
  4  0.2531730  0.9476476  0.1235803
  5  0.2609328  0.9445268  0.1308808
  6  0.2745656  0.9385849  0.1389301
  7  0.2837888  0.9344883  0.1453491
  8  0.2918668  0.9307513  0.1514015
  9  0.2972534  0.9281723  0.1562085

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 4.
```

Appendix F: Stepwise Coefficient - 95% Confidence Intervals:

	2.5 %	97.5 %
(Intercept)	10.06146342719	10.318801935469
`Engine HP`	0.00207604006	0.002393817109
`Number of Doors`	0.08868149353	0.137958904475
`highway MPG`	-0.01529508913	-0.011236791327
Popularity	-0.00001734428	-0.000004373835
Exotic	0.92532298828	1.039068283377
Luxury	0.06921048959	0.121381672352
Age	-0.09346075712	-0.089951194604
`Engine Fuel Type_diesel`	0.33670405259	0.495338596064
`Engine Fuel Type_electric`	1.04338483133	1.404437666687
`Engine Fuel Type_flex-fuel (premium unleaded recommended/E85)`	0.11717839614	0.508421514025
`Engine Fuel Type_flex-fuel (premium unleaded required/E85)`	0.08796357926	0.375718549823
`Engine Fuel Type_flex-fuel (unleaded/E85)`	-0.00721810504	0.068610632036
`Engine Fuel Type_premium unleaded (recommended)`	0.11409073877	0.178213144381
`Engine Fuel Type_premium unleaded (required)`	0.39294211490	0.464428992288
`Transmission Type_AUTOMATED_MANUAL`	0.13419271604	0.224626735074
`Transmission Type_AUTOMATIC`	0.18367925555	0.236399459093
`Driven_Wheels_all wheel drive`	0.00920744770	0.068697523518
`Driven_Wheels_four wheel drive`	0.06867886538	0.137909947892
`Driven_Wheels_front wheel drive`	0.02967683796	0.091623483572
`Vehicle Size_Compact`	0.01421591934	0.064386997855
`Vehicle Size_Large`	0.01168246777	0.067391317255
`Vehicle Style_2dr Hatchback`	0.09031881812	0.220483132064
`Vehicle Style_2dr SUV`	-0.01033028803	0.178562482450
`Vehicle Style_4dr Hatchback`	-0.16903231141	-0.079226727982
`Vehicle Style_Cargo Van`	-0.49563884986	-0.288624123656
`Vehicle Style_Convertible`	0.30160101105	0.423907200853
`Vehicle Style_Convertible SUV`	0.18156215697	0.569080838037
`Vehicle Style_Coupe`	0.12544578554	0.238943389473
`Vehicle Style_Crew Cab Pickup`	-0.10389422898	-0.011845347811
`Vehicle Style_Extended Cab Pickup`	-0.15972915141	-0.068426833847
`Vehicle Style_Passenger Minivan`	0.07184617822	0.175263248035
`Vehicle Style_Passenger Van`	-0.27062741060	-0.081324652300
`Vehicle Style_Sedan`	-0.11447490550	-0.058065335883