**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.

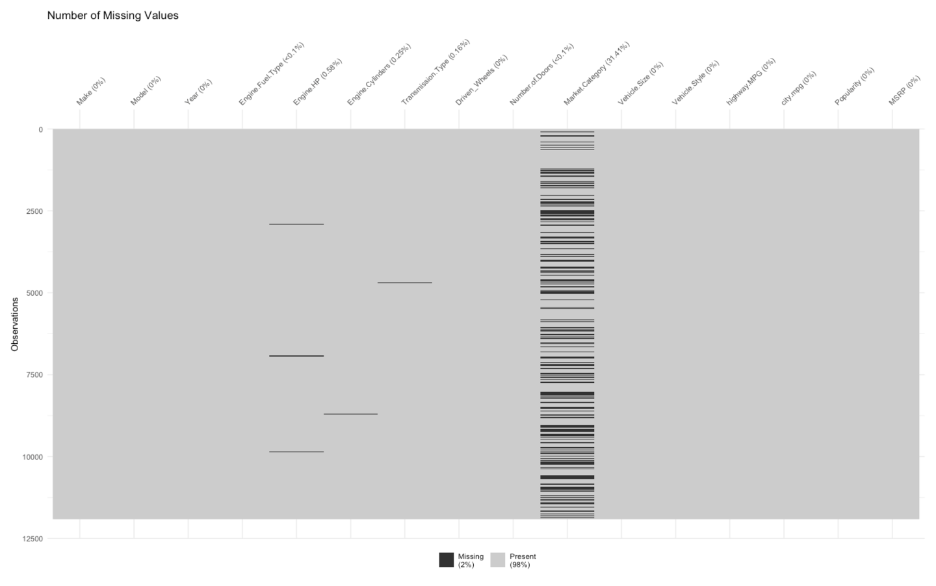
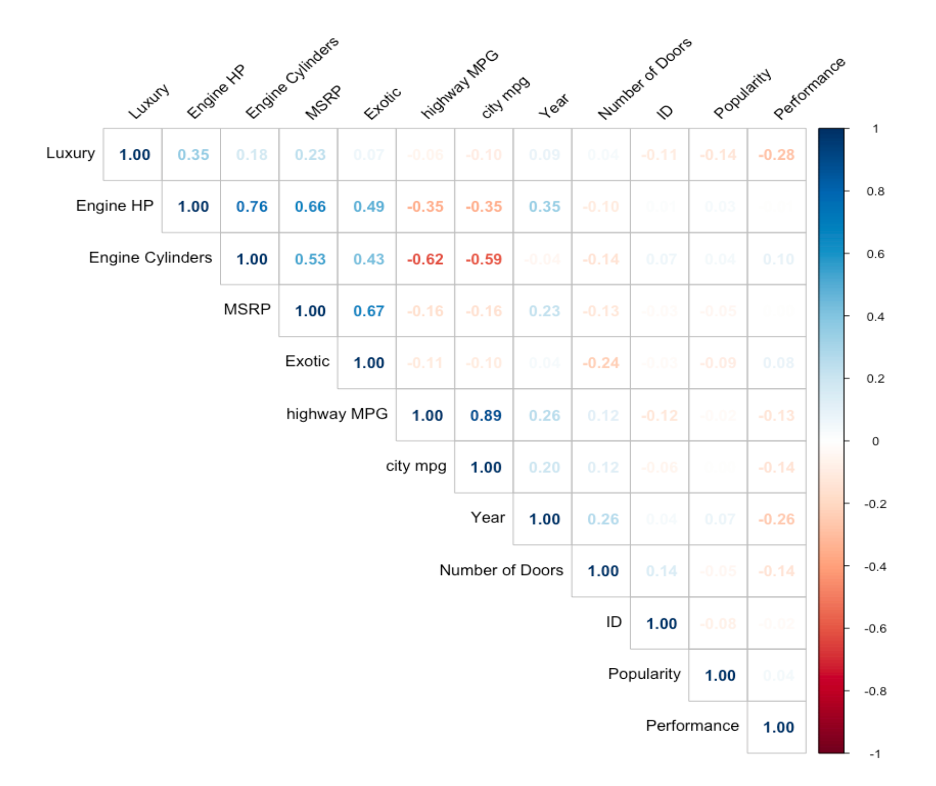


Figure 1.1

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.

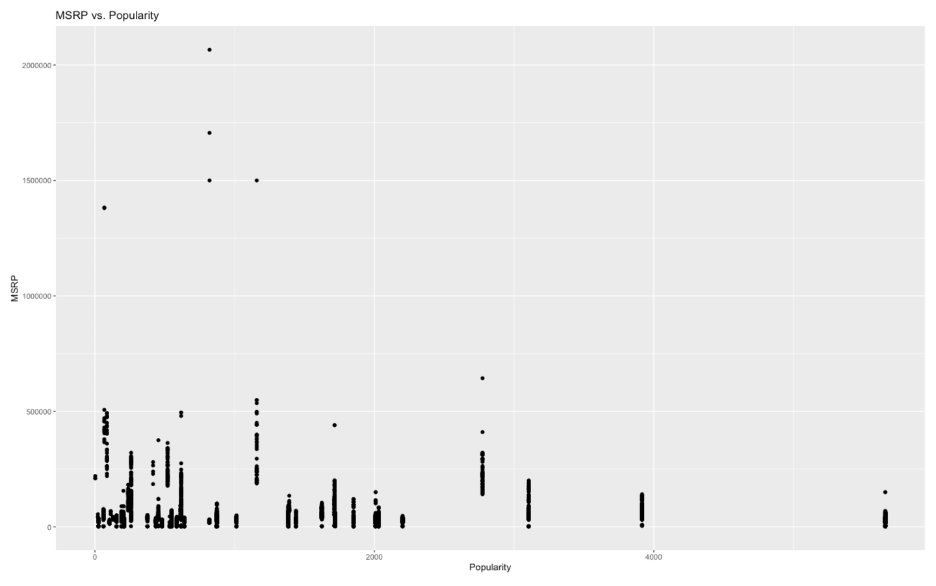
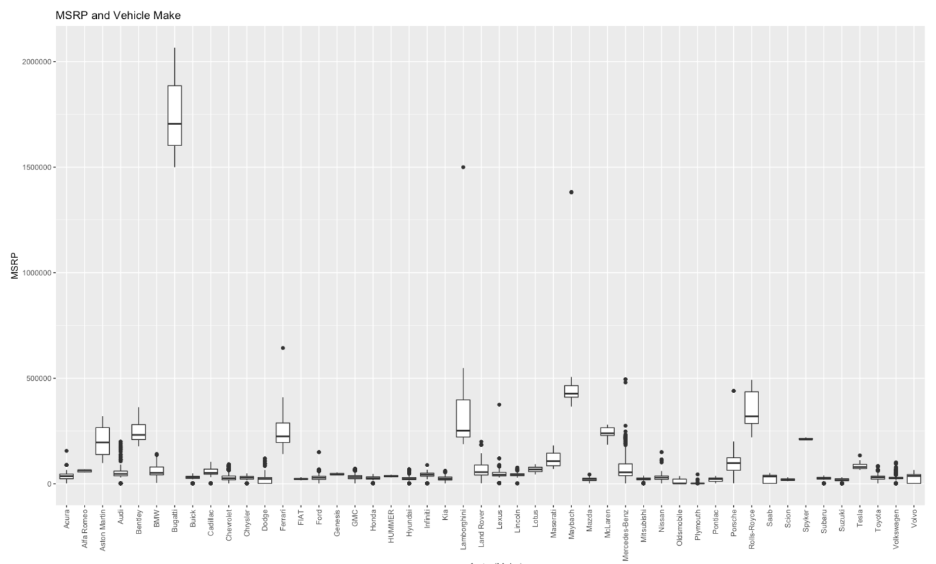


Figure 1.3

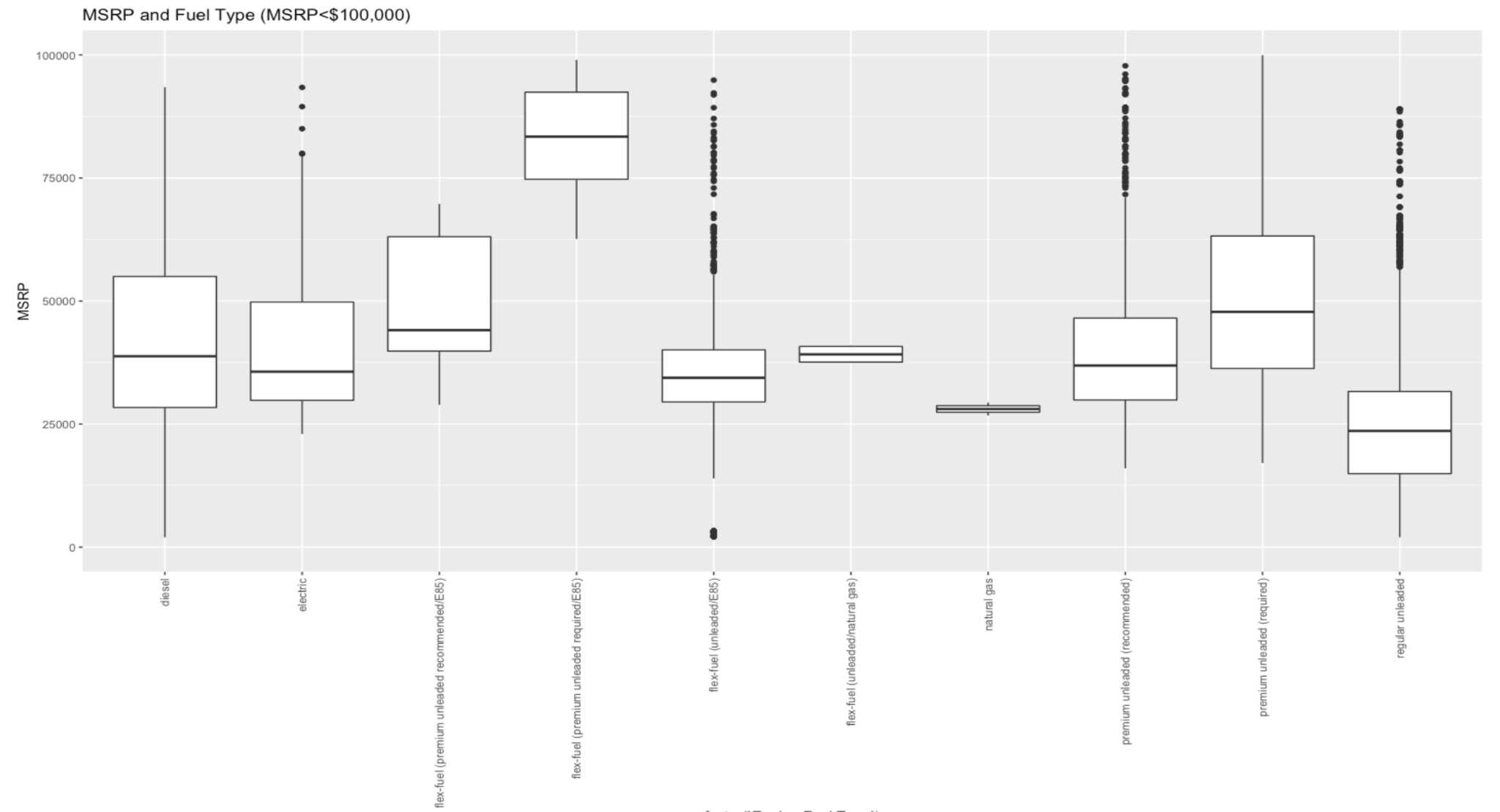
The next step in the exploratory data analysis was to visualize the relationship between the categorical variables in the dataset and MSRP. As seen on Figure 1.4, there are a few makes that are more influential on MSRP than others. Bugatti has a remarkably higher average MSRP than the rest of Makes in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

Figure 1.4



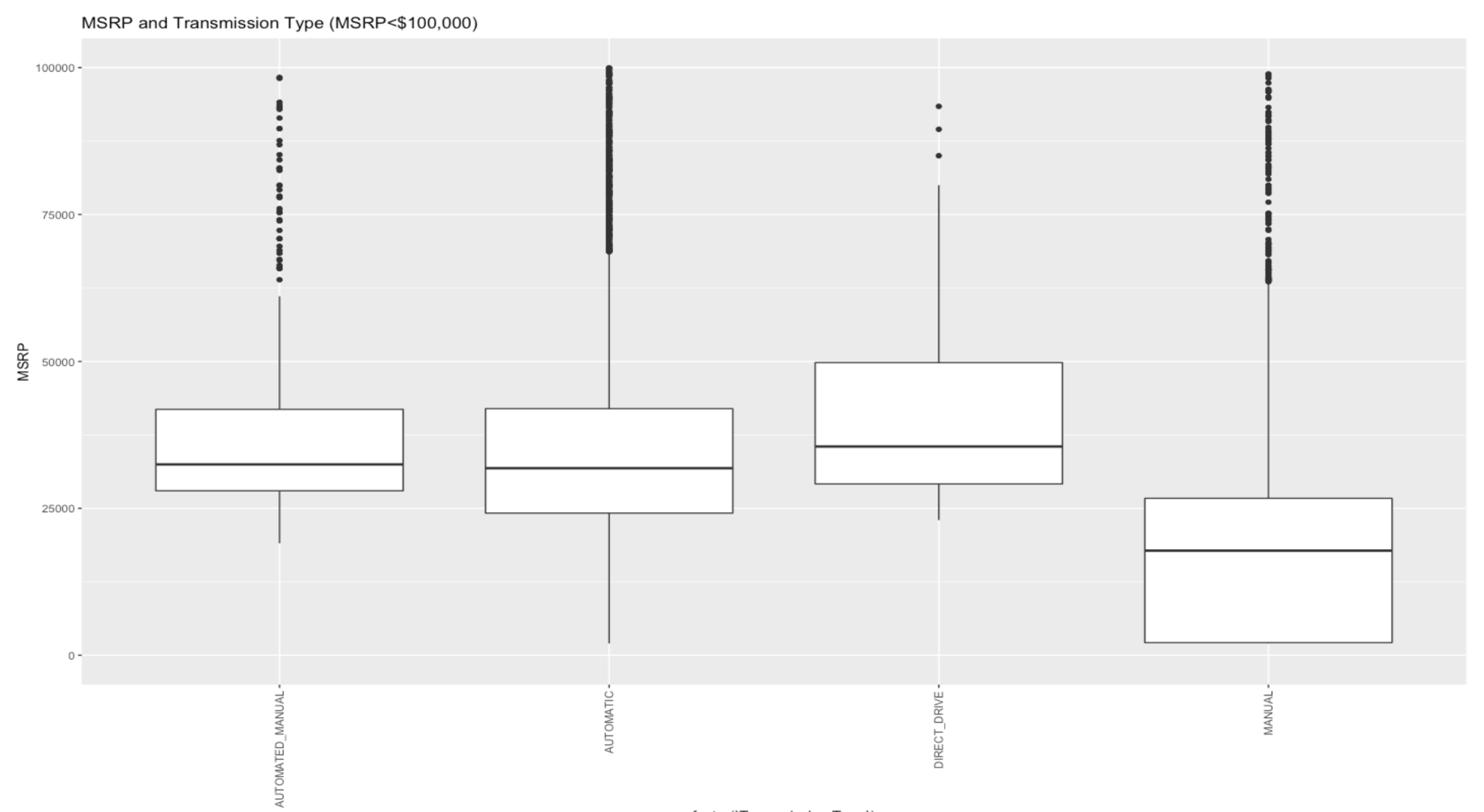
Different Fuel Types seem to have a similar average MSRP with the exception of two fuel types that have a significantly different average MSRP than the rest, “Flex-fuel (premium unleaded required)” with the highest average MSRP and “Regular Unleaded” with the lowest average MSRP. For scaling purposes,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

Figure 1.5



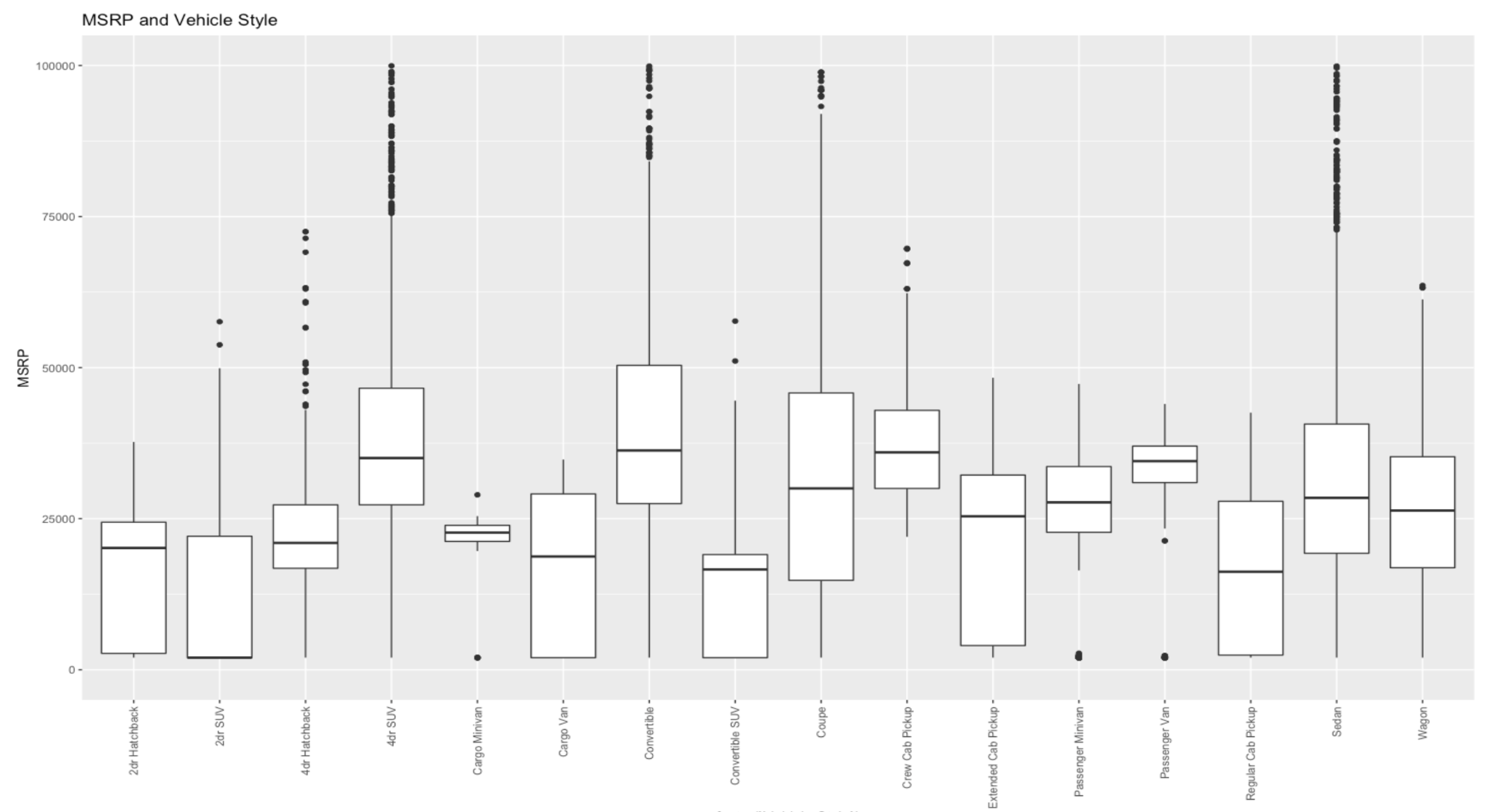
In the Transmission Type category cars under $100,000 that is labeled “Manual” has significantly lower average MSRP compared to the other transmission types.(Figure 1.6)

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:**

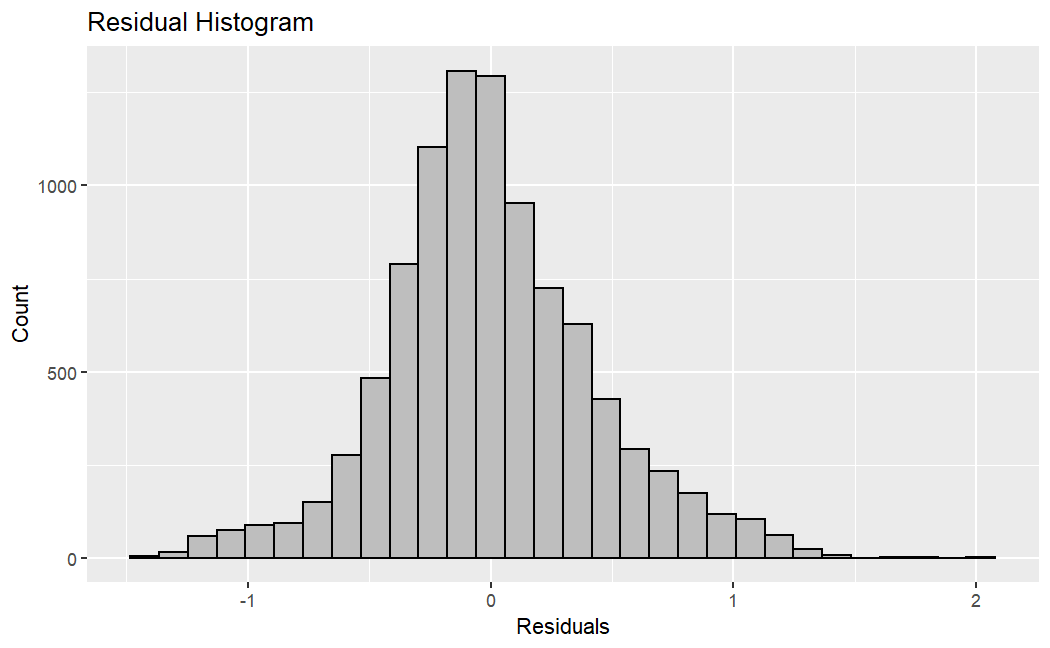
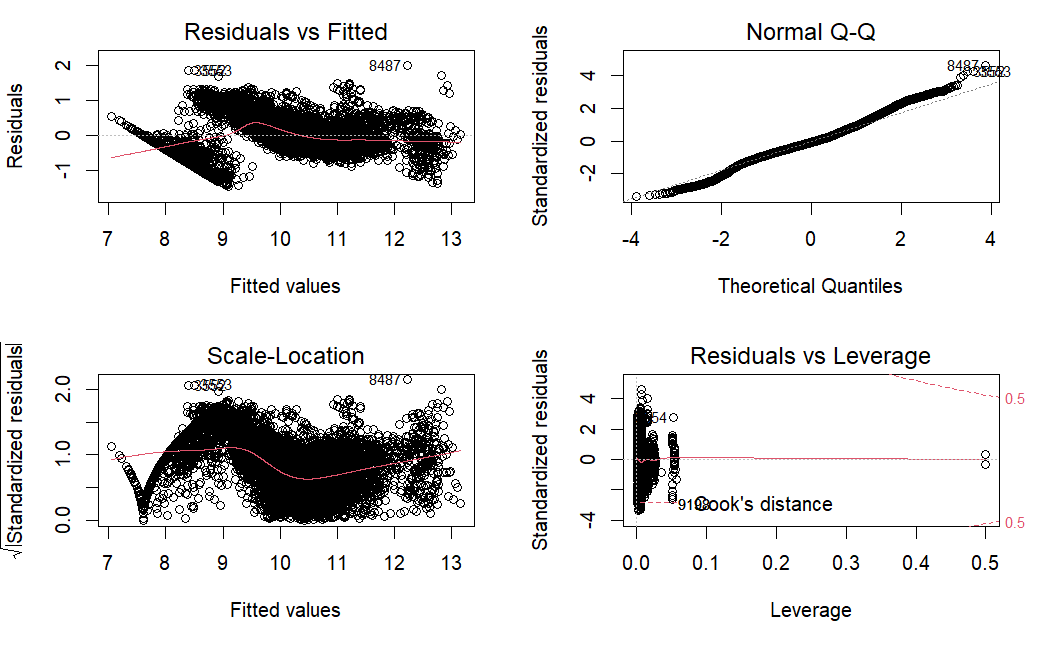
**Objective 1**

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*

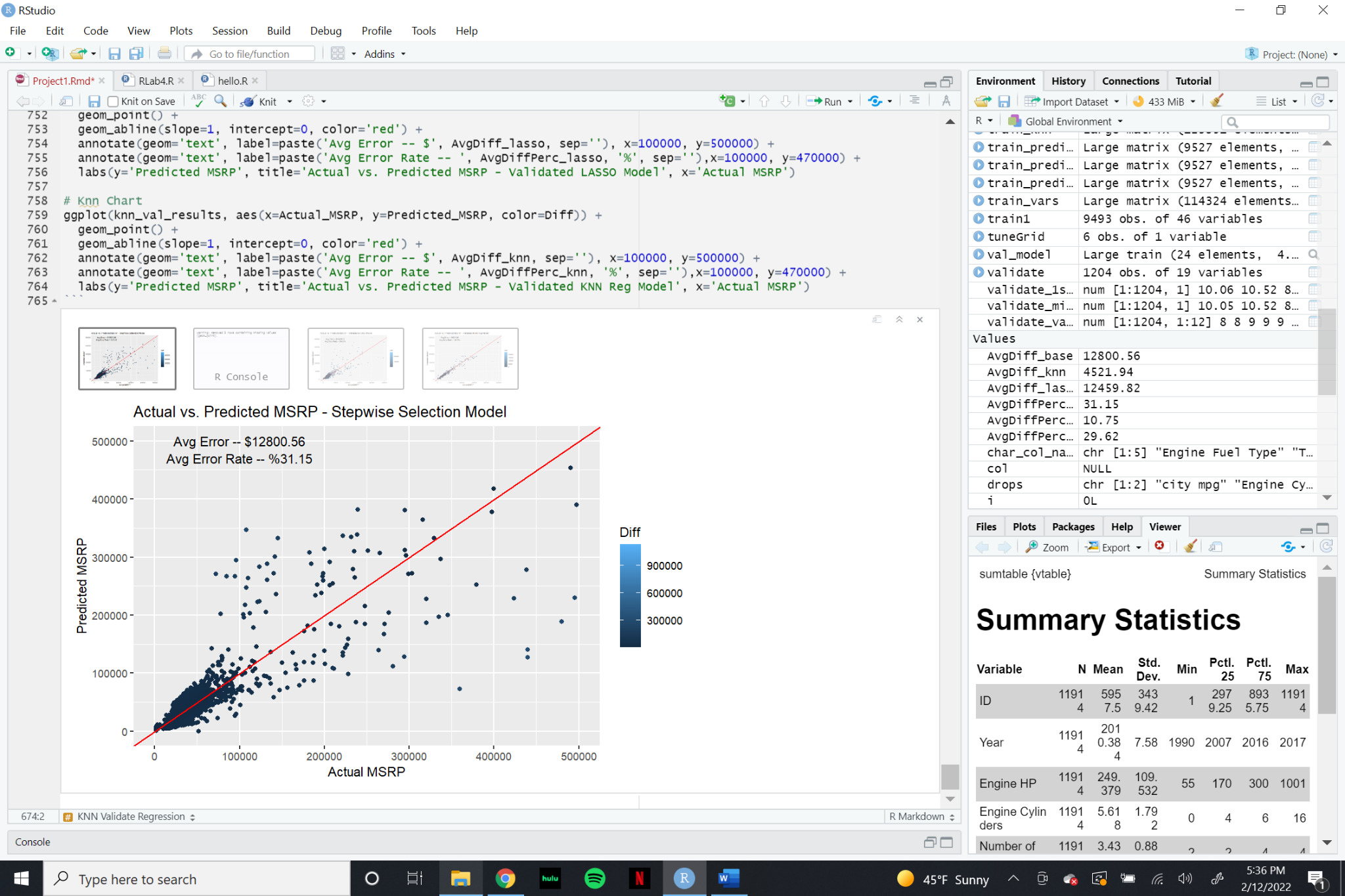
| Variable | Estimate | Lower Conf. Interval | Upper Conf. Interval |
| --- | --- | --- | --- |
| Age | -0.0998 | -0.0935 | -0.09 |
| Highway MPG | -.0133 | -.0153 | -.0112 |
| Exotic | .9822 | .9253 | 1.0391 |
| Popularity | -.00001 | -.00002 | -.000005 |

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 |

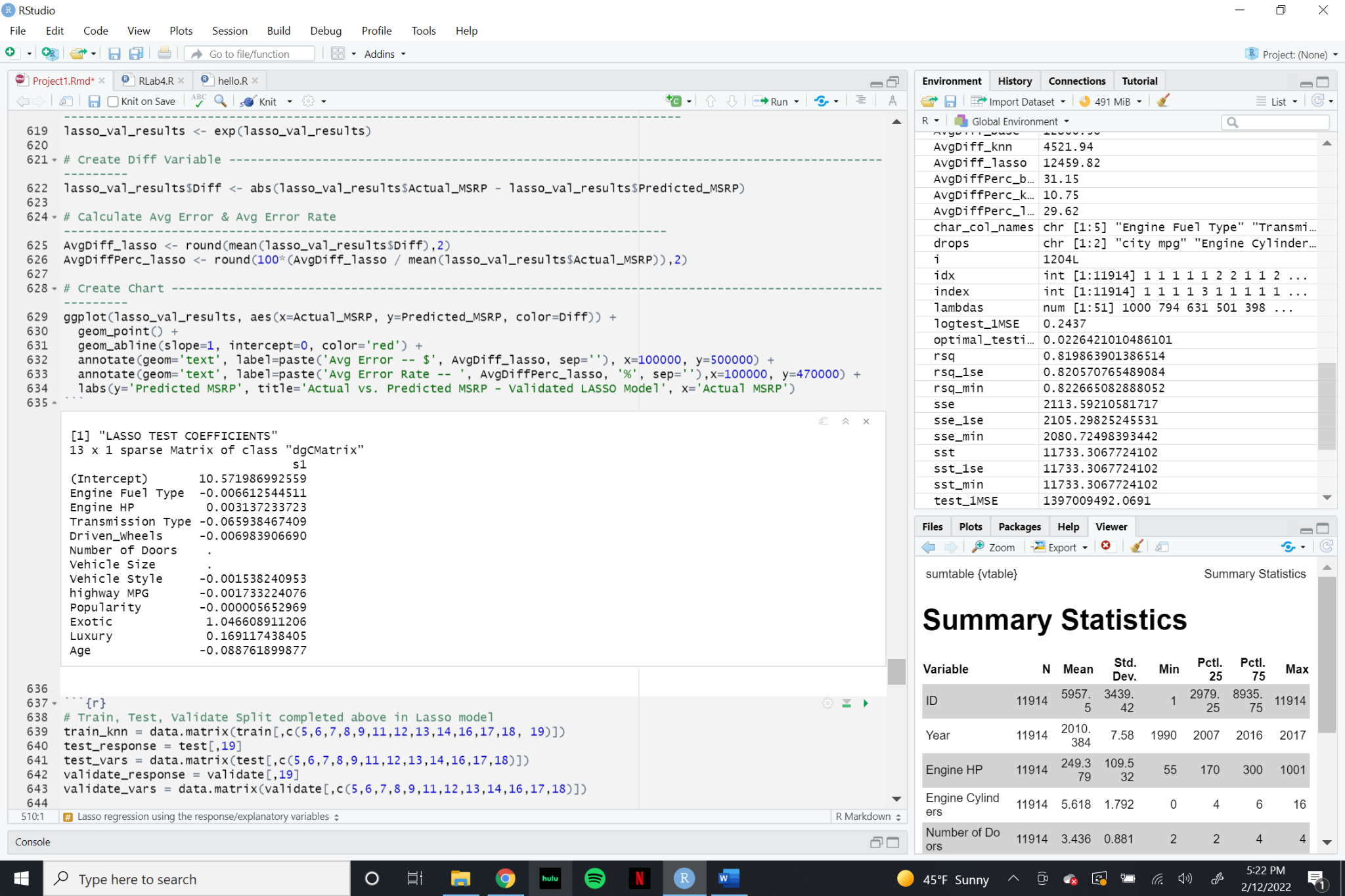


**Objective 2**

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.

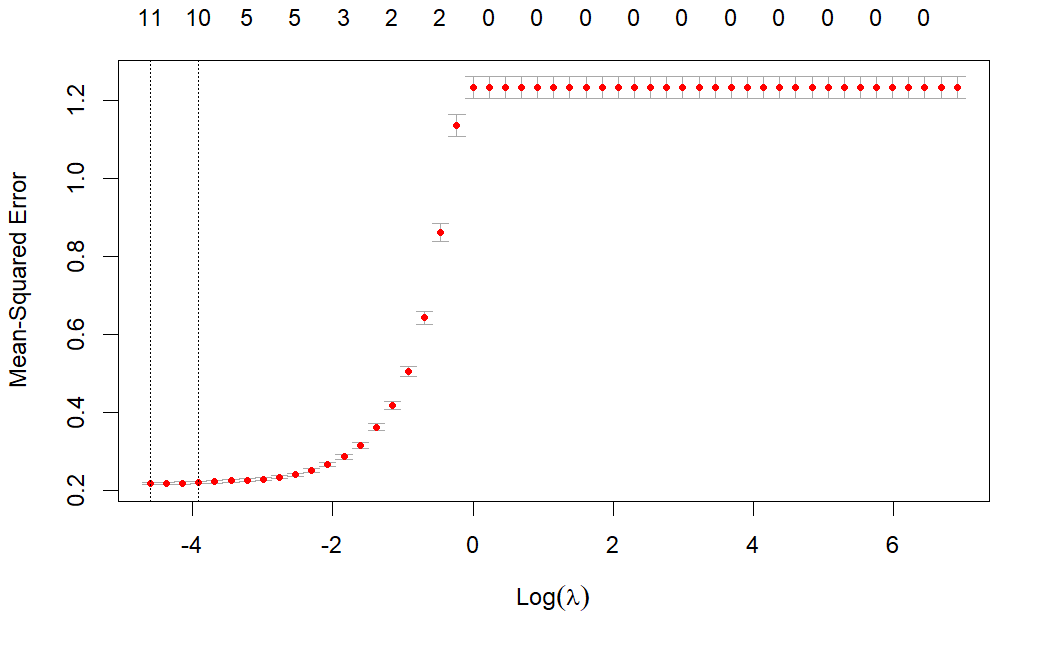
Initially, a simple testing model that does not use cross validation was created. Below is a list of the variables that were selected using lasso, as well as a table of summary statistics from the lm() model utilizing these variables. A more detailed summary of the lm() model statistics is shown in the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

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

| 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*

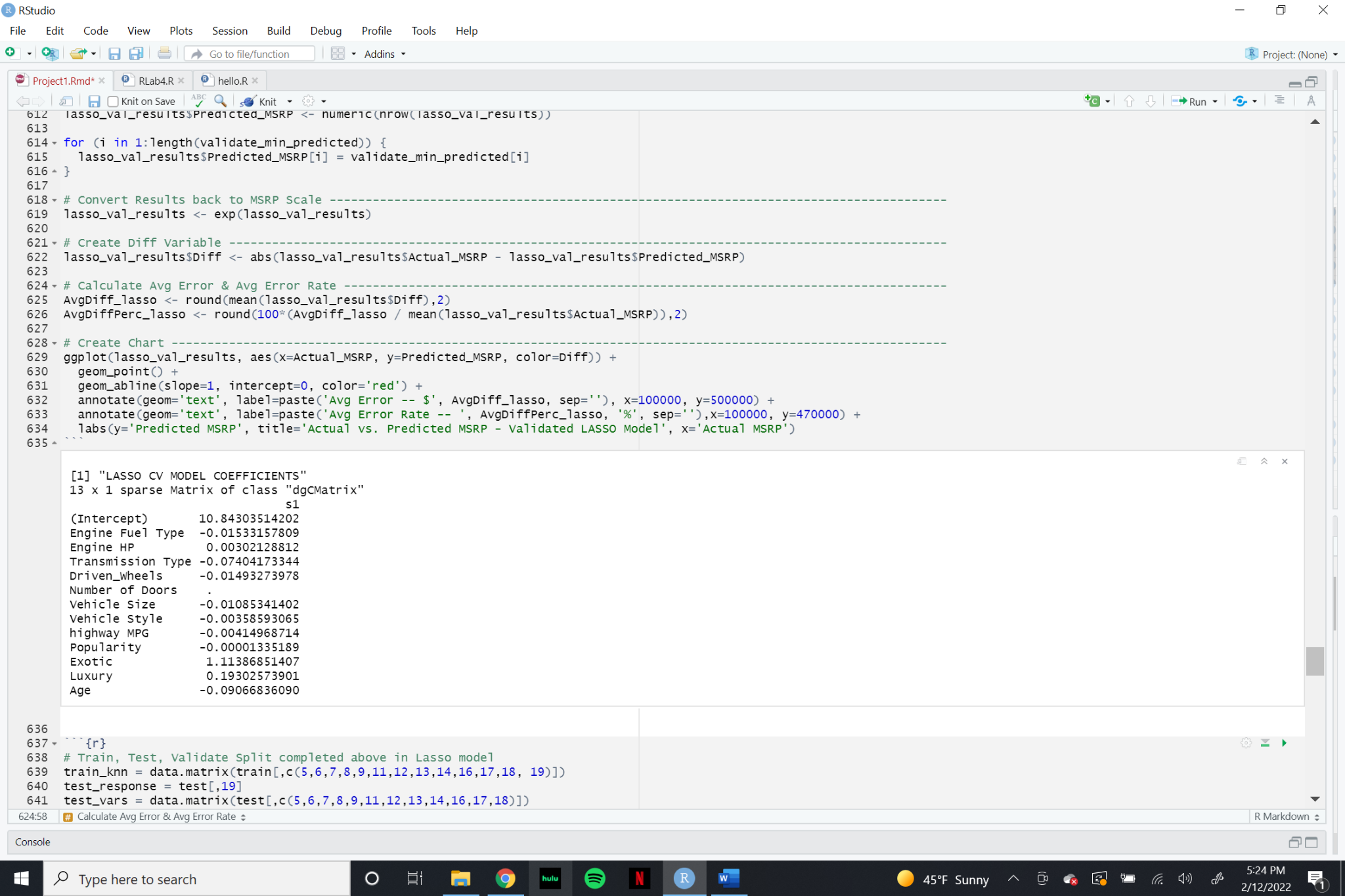


The two vertical lines highlight the placements of lambda.min and lambda.1se, the two best options for lambda. To test which option performed better, model statistics for both were calculated. The values of those statistics are included in the table below.

*Figure 3.4 : Validated Lasso Model - Lambda Selection Criteria*

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

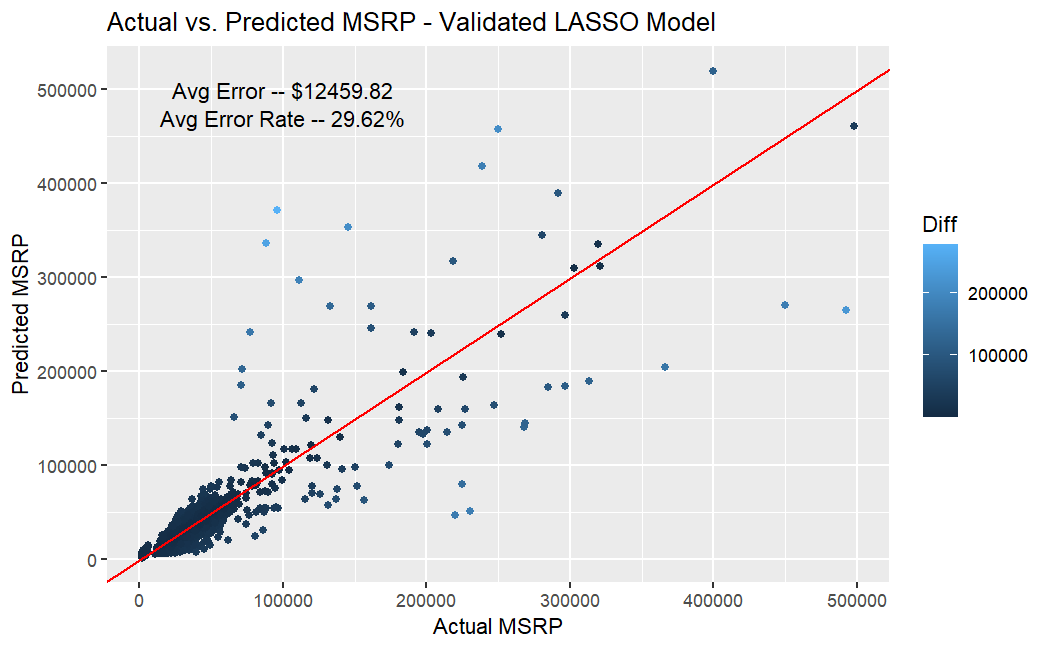
Although both models performed similarly, lambda.min is marginally better. Therefore, this is the parameter we will use moving forward. Similar to the test model, below are the selected variables and lm() model statistics from the validated lambda.min model. A more detailed summary of the lm() model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

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

| 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*

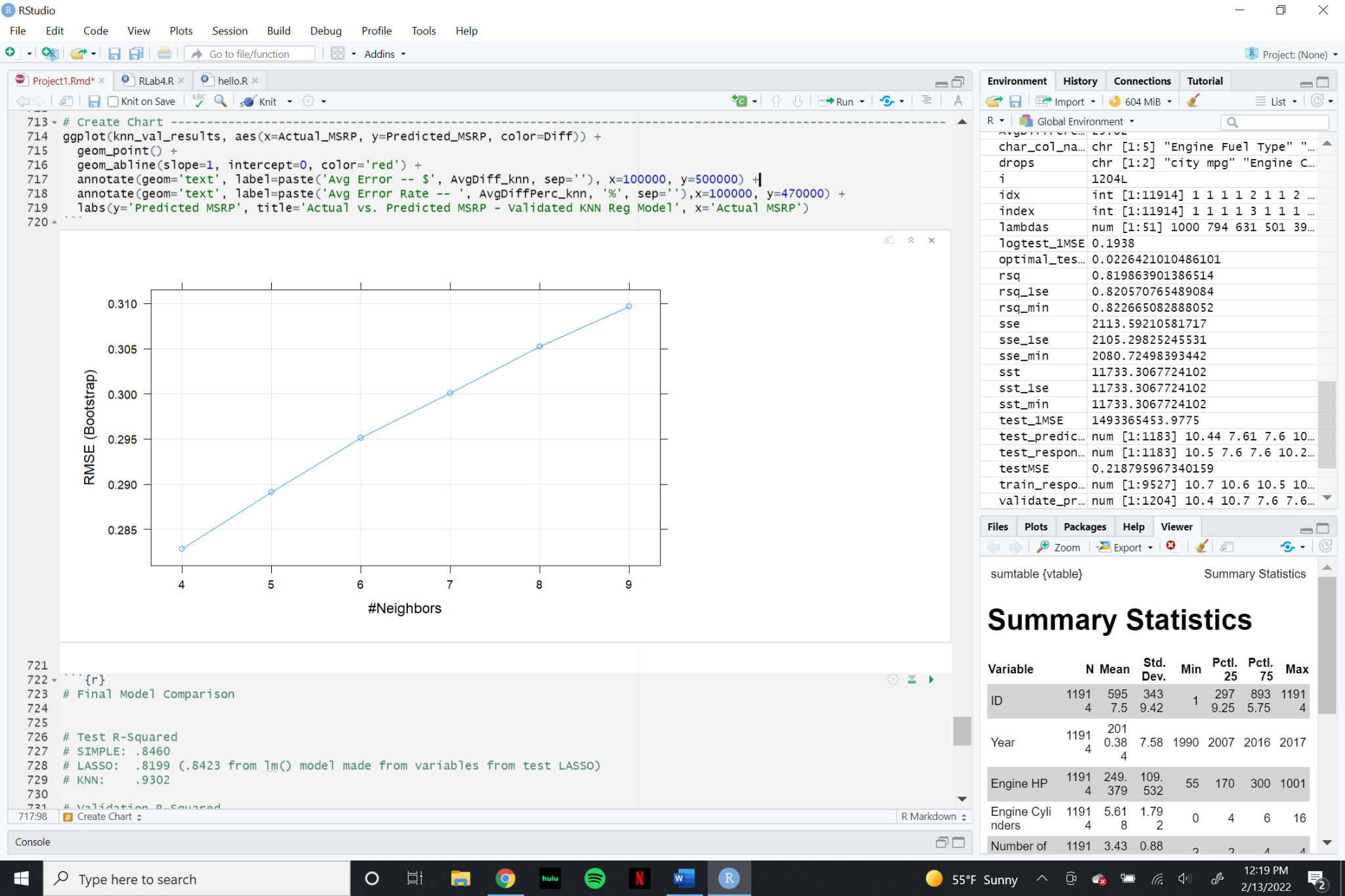


The farther an observation is from the red line, the less accurate the prediction is. It’s extremely evident that low-MSRP cars are easier to predict, maybe because they are much more common. This model, on average, has an $12,459 error and a ~30% error rate. This is an improvement from the stepwise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

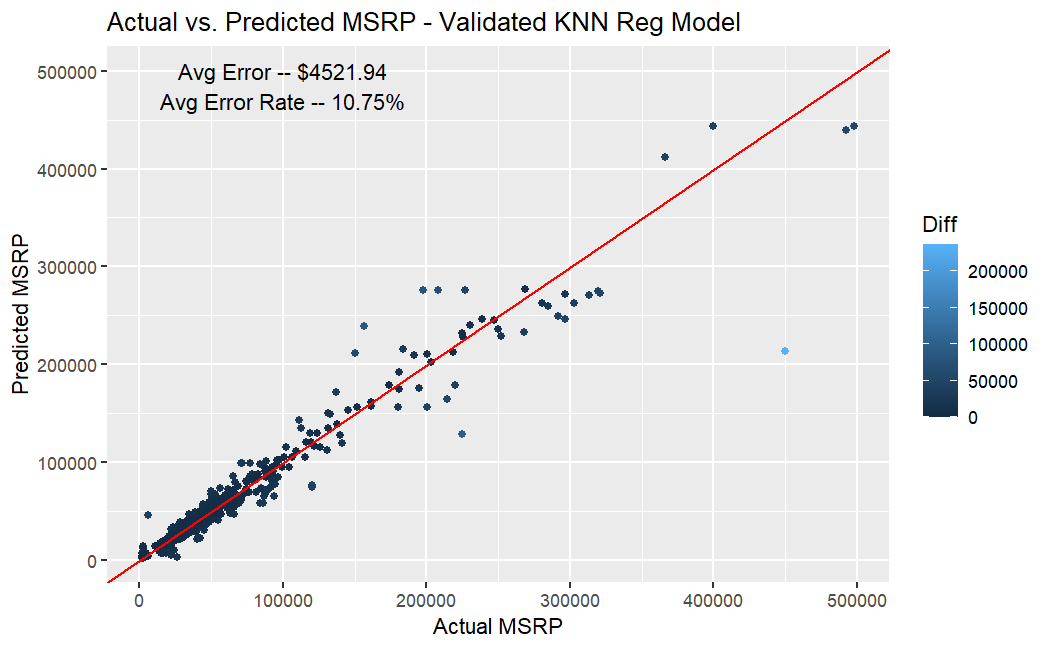
It is evident that this testing model is significantly better than any previous model within this project, having a Test MSE that is one seventh of the next lowest Test MSE (validation Lasso). Next, cross validation was used to hypertune the K parameter. Similar to the testing set, we found that K=4 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`.

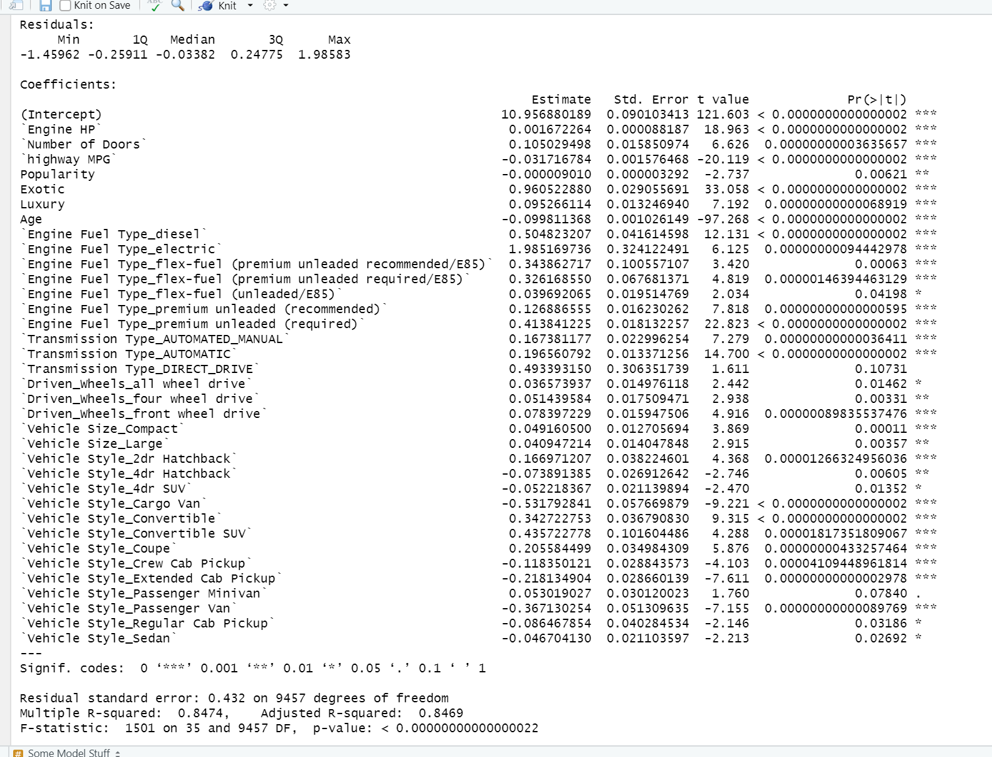
**Appendix**

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>

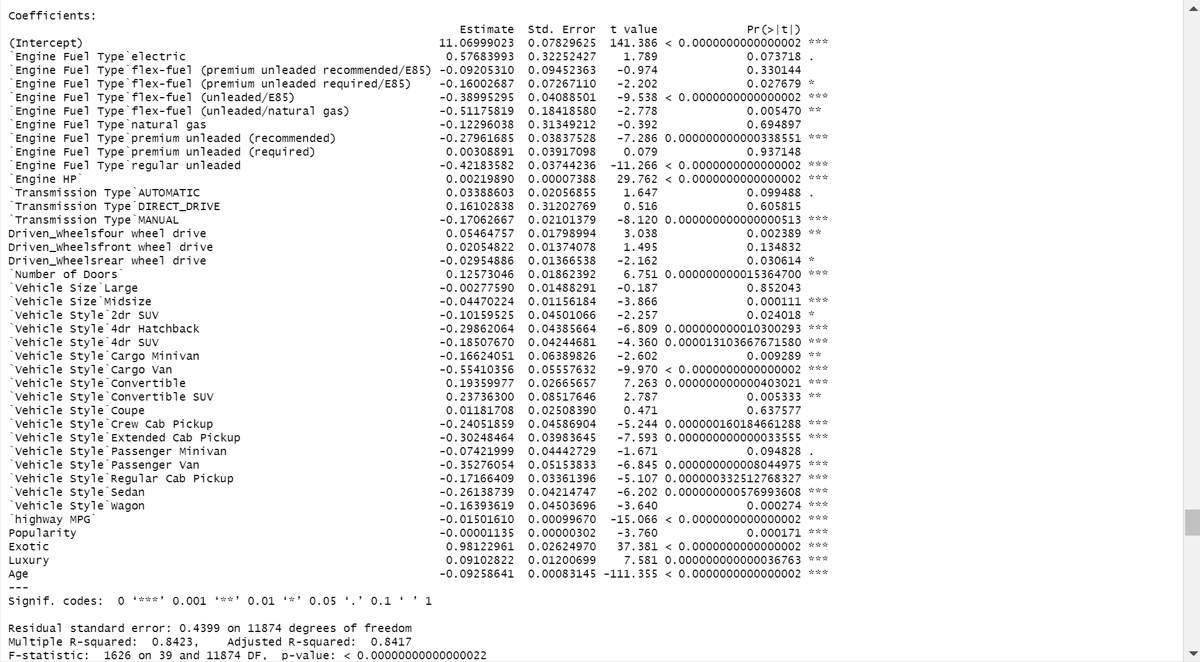
**Appendix A:**  Variable Types

| **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. |

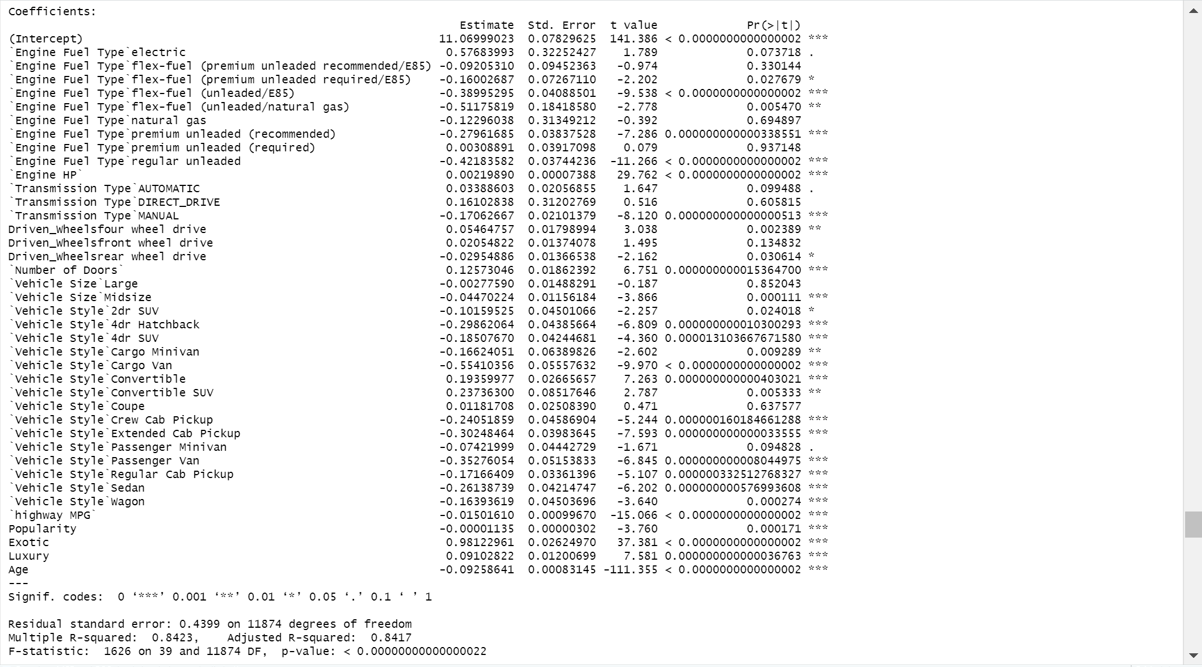
**Appendix B:** Detailed summary of the stepwise test coefficients:



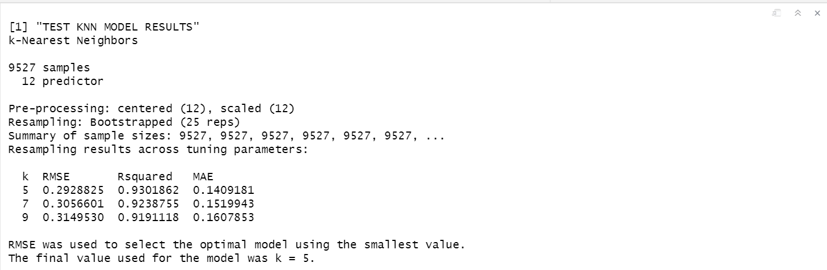
**Appendix C:** Detailed summary of the lasso test coefficients:

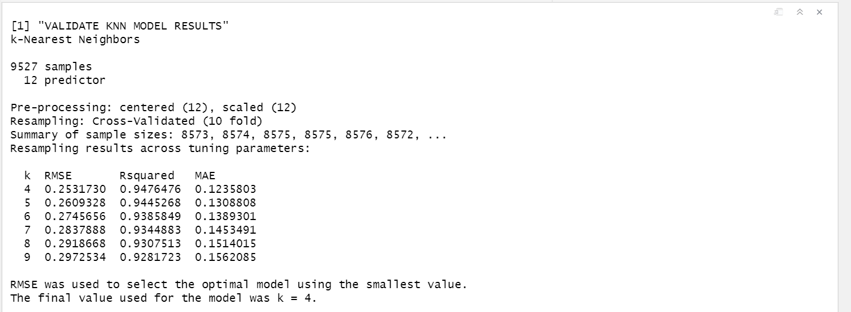


**Appendix D:** Detailed summary of the Lasso validate coefficients:



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





**Appendix F:** Stepwise Coefficient - 95% Confidence Intervals:

