**Requirements (delete prior to final submission)**

* No more than 7 written pages (10 pages for appendix) (Max total = 17)
* single-spaced to doubled spaced
* 11-point font
* 1-inch margins
* page limitation includes graphics and tables
* Each graphic or table should be clearly labeled and discussed in the text.
* You may include more tables and figures in an appendix (no more than 10 pages)
* R code should also be included in the appendix
* Any table or figure that is included in the appendix must be referenced in the text.
* *build regression models*
* *identify key relationships and interpreting those relationships*
* *Address the importance of the “Popularity” variable*
* *How much general popularity can play a role in the retail price of a vehicle*
* *Provide detailed information on summary statistics, EDA, and your model building process*
* *Provide interpretation of the regression coefficients of your final model including hypothesis testing, interpretation of regression coefficients, and confidence intervals.*
* *Mention the Practical vs Statistical significance of the predictors.*
* *Answer any additional questions using your model that you deem are relevant.*
* *Use training data set for EDA and model fitting*
* *Use test set to compare models to make a final call*
* *Compare multiple models with the goal of developing a model that can predict the best and do well on future data*
* *Use training and test set to build at least one additional multiple linear regression model with more complexity than the interpretable model*
* *Use the ISLR text book (and internet) to identify one nonparametric technique to build a regression model*
* *Select from k-nearest neighbors’ regression or regression trees.*
* *Total you should have at least 3 models, 2 linear regression models and 1 nonparametric.*
* *For each of the three models, provide measures of fit for comparisons: test ASE and validation ASE are mandatory.*
* *You may also include additional metrics for completeness like R squared/Adjusted R squared, AIC, and BIC where applicable (only point where the validation set is being used).*
* *Providing additional insight as to why one model is better than the other, or why they are all the same is encourage.*
* After 2000 the MRSP has massive shift – explain it and use data, or split into 2 datasets, or drop/delete one of the two sides of data
* For each model interpret the MEANING of at least 1 numerical and 1 categorical variable. Show table of ALL variable results.
* Make sure to discuss the PROCESS and mindset of the EDA, not just the graphs
* 3-7 pages of WRITTEN text is sweet spot
* Start with “shotgun” graphs to look for correlation but then focus down into particular variables and display larger (more comprehensive) graphics

**Analysis of the Automotive Industry**

**Authors:** Nicole Assenza, Andrew Taylor and Ranjan Karki

In 2021, the worldwide automotive industry had an estimated value of almost 3 trillion dollars. More than 60 international manufacturers compete for their share of the market which sells roughly 76 million cars every year. As the amount of relevant data surrounding the industry continues to grow at an exponential rate, these companies are continually looking for reliable trends that yield even the slightest competitive advantage. This comprehensive report is intended to aid one such company in analysis and interpretation of the available data through multiple regression and nonparametric modeling.

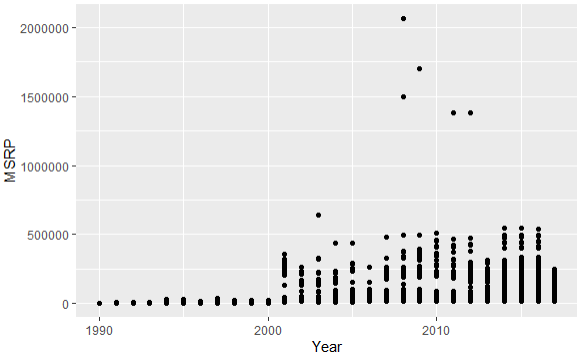
The data set of interest includes 11,914 records of discrete automobiles in production between 1990 and 2017. Each record hosts data points for 16 variables (9 numerical and 7 categorical) such as sale price (MSRP), Engine HP, Manufacturer, and “popularity score” to name a few. The full list of variables and their descriptions can be seen in **Table A-1** below.

**Table A-1**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| MSRP | Numeric | Sale Price (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, Diesel, etc. |
| Engine HP | Numeric | Horsepower of the car’s engine. |
| Engine Cylinders | Factor | Number of cylinders in the car’s engine. |
| Transmission Type | Factor | Type of transmission in the car. Ex: manual, automatic, etc. |
| Driven Wheels | Numeric | Wheels powered by engine. Ex: Front Wheel, Rear Wheel, etc. |
| Number of Doors | Numeric | The number of doors that the car has. Ex: 2, 4, etc. |
| Market Category | Factor | Various special factors for each car. Ex: Exotic, Luxury, etc. |
| Vehicle Size | Factor | The size of the vehicle. Ex: Midsize, Large, Compact, etc. |
| 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 numerical popularity score for each car. |

For the first objective of our analysis, we will assess the correlations of these variables to our response variable, the MSRP sale price. Then using various selection methods, we will build a simple regression model that can use the selected variables to predict a car’s MSRP with moderate accuracy, but more importantly, is easy to interpret. In order to have confidence in any future analysis or model building, the data must first be cleaned of errors, blanks, N/As, etc. For this particular dataset there were few enough of these null values that the records in question could be verified online at various car manufacturer websites to obtain and insert the correct values. In total we identified N/As in 69 records for “Engine HP”, 30 records for “Engine Cylinders”, and 6 records for “Number of Doors”. Furthermore, there were 3 records with missing “Engine Fuel Type” and 1 record with an incorrect value for “Highway MPG”. Once the missing data was resolved, the data was further cleaned by altering the data types of certain variables to fit our analysis strategy. By converting the values of “Engine Cylinders” for all electric vehicles to “0”, this variable is able to retain its numerical quality. Other variables were converted to factors to allow for analysis of particular levels including: Make, Engine Fuel Type, Transmission Type, Driven Wheels, Vehicle Size, Vehicle Style, and Market Category. The “Market Category” of the cars contains a combination of 11 descriptors such as “Luxury”, “Crossover”, “Hybrid”, etc. that offer good insight a particular car’s qualities and purpose as it pertains to the consumer. This information feels valuable in terms of representing certain features which are correlated to higher or lower sale prices. However, because of the way the current variable is displayed, it is not easy to comprehend or fit into regression analysis. To better utilize this portion of data, the Market Category variable was used to create new variables each representing whether or not a record contained those qualities. For example, a record with “Market Category” equal to “Luxury, Exotic, Performance” will have new variables “Luxury” “Exotic” and “Performance” equal to TRUE, while all other new variables equal to FALSE (“Hybrid” “Crossover” “Flex Fuel” “Hatchback” etc).

Now that the data is wrangled and cleaned, the exploratory data analysis can begin to better understand the effects each variable has on the response. After reviewing a matrix of correlation plots, the team was able to identify a few particulars that could be altered to assist in our final models. Once such example is the relationship between “Year” and “MSRP”. As seen below in **Plot B.1**, the MSRP values after the year 2000 appear drastically different than those before it. It appears there was a change after the year 2000 in how this set of data was captured which poses an issue if we treat each year the same. Therefore, the data was split into 2 subsets: cars with “Year” less than or equal to 2000 and those with “Year” greater than 2000. Due to the fact our customer of this analysis is most likely interested in using models to evaluate current and future cars, the “greater than 2000” dataset will be used for the remainder of the analysis while the other set will be ignored.



Correlations to MSRP from other variables are not as visually obvious as is. By taking the log transform of MSRP, much of the relationships become clearer. Taking the log of our response still maintains a level the of interpretability sought after, however, any further transforming of the explanatory variables will make it more difficult to understand practical explanations so we will avoid doing so while building the model for this first objective. Similarly, certain variables will also be too complex to interpret simply based on their data type and vast number of factor levels. Therefore, categorical variables such as “Make” “Model” and those related to “Market Category” will also be temporarily dropped from the analysis.

The remaining data variables can now be used to build various, simple models for comparison. In order to determine which of the remaining 12 variables should be included in the regression models, step-wise, forward, and backward selection techniques were used to evaluate each variable’s statistical significance as it relates to the impact it has on MSRP. Both the step-wise and forward selection methods advised on keeping the same 9 in the model whereas the backwards selection method directed the team towards dropping only one variable and keeping the remaining 11. To determine which of the 3 (2 since stepwise and forward are the same) models better represents MSRP, the summary statistics of each can be compared as seen below in **Table A.2**. The model using variables from the backwards selection method has a much lower AIC (1999.80) compared to the other two (7745.58).

Prior to building these models, the data was split into an 80% training set, 10% testing set, and 10% validation set. This proved useful as each of the models were built using the training set and can now be used to predict the values in the test set and compared for accuracy. In all three summary statistics used to determine which model performed better in their predictions, the forward/stepwise proved better than the backwards model as seen in the **Table A.3**.

Due to the fact the forward/stepwise model had a higher R values, lower Sum of Squared residuals, and less variables to ultimately interpret, these variables and coefficients will be used to build our first, simple model:

Log(MSRP) = 9.96 + 0.26(Engine.Cylinders) - 0.0908(Transmission.Type) + 0.014(City.MPG) – 0.13(Number.of.Doors) – 0.088(Driven\_Wheels) – 0.00001(Popularity) – 0.026(Engine.Fuel.Type)

Where:

Transmission Type of X = 1

Transmission Type of X = 1

Driven\_Wheels of X = 1

Driven\_Wheels of X = 1

Engine.Fuel.Type of X =

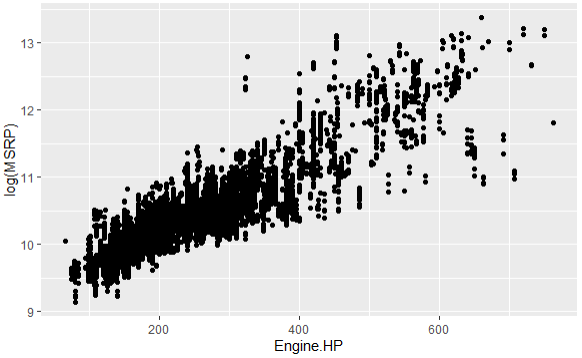
Engine.Fuel.Type of X =

\*interpretation of regression coeficcients, hypothesis testing, Confidence Intervals

One of the interesting variables this final model is “Popularity”. It was the “least significant” of the included variables. By viewing a plot Popularity vs MSRP (**Plot B.2**), we can see there is a slight positive correlation but overall but with such a random cloud of points along its range, it is obviously not a good representation of MSRP by itself.

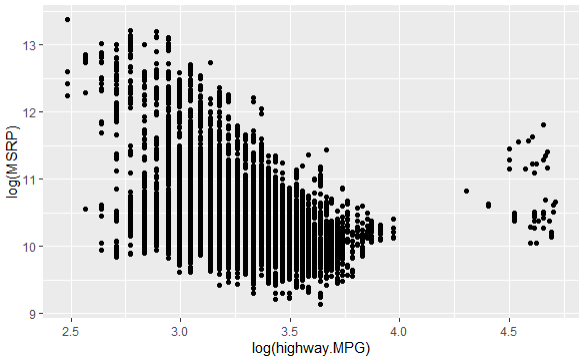
By looking at trends from the other included variables **(Plots B.3 – B.7**), we can make statements on their correlations to sale price.

**Plot B.3**

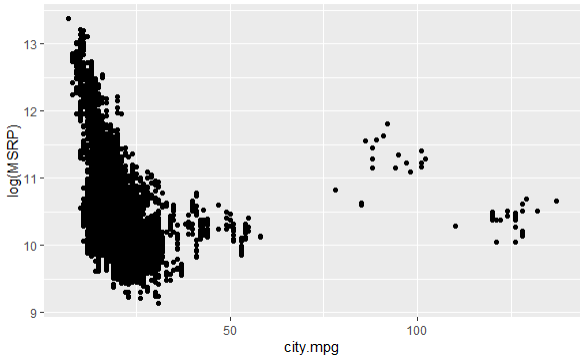


As the Engine Horsepower of a car increases, the log of MSRP increases

**Plot B.4**



**Plot B.5**



|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| MSRP | Numeric | Sale Price (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, Diesel, etc. |
| Engine HP | Numeric | Horsepower of the car’s engine. |
| Engine Cylinders | Factor | Number of cylinders in the car’s engine. |
| Transmission Type | Factor | Type of transmission in the car. Ex: manual, automatic, etc. |
| Driven Wheels | Numeric | Wheels powered by engine. Ex: Front Wheel, Rear Wheel, etc. |
| Number of Doors | Numeric | The number of doors that the car has. Ex: 2, 4, etc. |
| Market Category | Factor | Various special factors for each car. Ex: Exotic, Luxury, etc. |
| Vehicle Size | Factor | The size of the vehicle. Ex: Midsize, Large, Compact, etc. |
| 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 numerical popularity score for each car. |
| Crossover | Boolean | True if “Market Category” includes “Crossover”. False if not |
| Diesel | Boolean | True if “Market Category” includes “Diesel”. False if not |
| Exotic | Boolean | True if “Market Category” includes “Exotic”. False if not |
| Factory Tuner | Boolean | True if “Market Category” includes “Factory Tuner”. False if not |
| Flex Fuel | Boolean | True if “Market Category” includes “Flex Fuel”. False if not |
| Hatchback | Boolean | True if “Market Category” includes “Hatchback”. False if not |
| Hybrid | Boolean | True if “Market Category” includes “Hybrid”. False if not |
| Luxury | Boolean | True if “Market Category” includes “Luxury”. False if not |
| Performance | Boolean | True if “Market Category” includes “Performance”. False if not |

**Appendix**