1. **Introduction:**

To accurately recognize the potential pricing risk within the competitive automobile market, many automotive companies will utilize advanced data analysis techniques to implement pricing strategies for their cars, trucks and SUV's. This analysis of pricing relies on differentiating characteristics of the vehicles, including several of the cars physical characteristics (i.e. engine, body style, etc.) as well as a symbolled value between -3 and +3 that represents the overall risk factor of that vehicles pricing structure (-3 representing the lowest risk up to +3 representing the highest risk).

The following analysis will evaluate twenty-two separate vehicle makes, with twenty-five explanatory variables that potentially affect each cars pricing model. One hundred and seventy-nine observations will be considered in this study. To more accurately predict the pricing of each vehicle, the number of explanatory variables within the dataset will be tested and reduced via variable selection strategies such as LASSO, forward, and step-wise selection, in order to identify the best fitting model. Once the best set of explanatory variables are identified, several questions will be addressed.

First, how does horsepower relate to a vehicles MPG (miles per gallon) in the city based on fuel type? Second, for vehicles with body styles categorized as a hardtop, hatchback, sedan or wagon, what is the mean miles per gallon in the city the car will achieve? With these answers, the analysis will then shift focus to predicting the pricing models of each vehicle, concluding with the top three risk adverse models being identified.

1. **Data Description:**

* Train data has 25 attributes and 179 observations.
* Test data has 24 attributes and 24 observations, with price being the attribute needs to be predicted.
* There are 19 observations total missing from the following attributes:
  + Make, Fuel Type, Number of Doors, Curb Weight, Bore, Stroke and Horse Power
* There are 14 continuous variables & 11 classification variables. Table 1.1 has the details of the attributes within this analysis.

Table 1.1

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes** | **Type** | **Example** | **Model** |
| symboling | Numeric | 1 | Class |
| make | Character | Audi | Class |
| fuel-type | Character | gas | Class |
| aspiration | Character | std | Class |
| num-of-doors | Character | two | Class |
| body-style | Character | sedan | Class |
| drive-wheels | Character | rwd | Class |
| engine-location | Character | front | Class |
| wheel-base | Numeric | 88.6 | Continuous |
| length | Numeric | 168.8 | Continuous |
| width | Numeric | 64.1 | Continuous |
| height | Numeric | 48.8 | Continuous |
| curb-weight | Numeric | 2548 | Continuous |
| engine-type | Character | dohc | Class |
| num-of-cylinders | Character | four | Class |
| engine-size | Numeric | 130 | Continuous |
| fuel-system | Character | mpfi | Class |
| bore | Numeric | 3.47 | Continuous |
| stroke | Numeric | 2.68 | Continuous |
| compression-ratio | Numeric | 9 | Continuous |
| horsepower | Numeric | 111 | Continuous |
| peak-rpm | Numeric | 5000 | Continuous |
| city-mpg | Numeric | 21 | Continuous |
| highway-mpg | Numeric | 27 | Continuous |
| price | Numeric | 13495 | Continuous |

For clarification, the symbolling attribute is derived from the ISO (International Organization of Standardization). “A Symbol is a code used in ISO's Vehicle Series Rating (VSR) program. The purpose of Vehicle Series Rating is to match premiums for each particular type of car to losses for that type of car.” ([*https://www.verisk.com/insurance/about/faq/iso-s-symbols-for-individual-makes-and-models-of-cars/*](https://www.verisk.com/insurance/about/faq/iso-s-symbols-for-individual-makes-and-models-of-cars/)).

1. **Clean the Data:**

As mentioned above, certain attributes within both the Train and Test data sets were found to be missing values. For the Train data, these attributes were identified as the fuel type, number of doors, curb weight, bore, stroke and horsepower variables. For the Test data, these attributes were identified as make and curb weight. During the initial data cleanup, the missing values which displayed as a “?” within the two datasets were replaced with a standard “NA” value. Additional cleanup included renaming columns for accessibility purposes and consistently casing all categorical variables (i.e. “Four” and “four” were transformed to all be lower case).

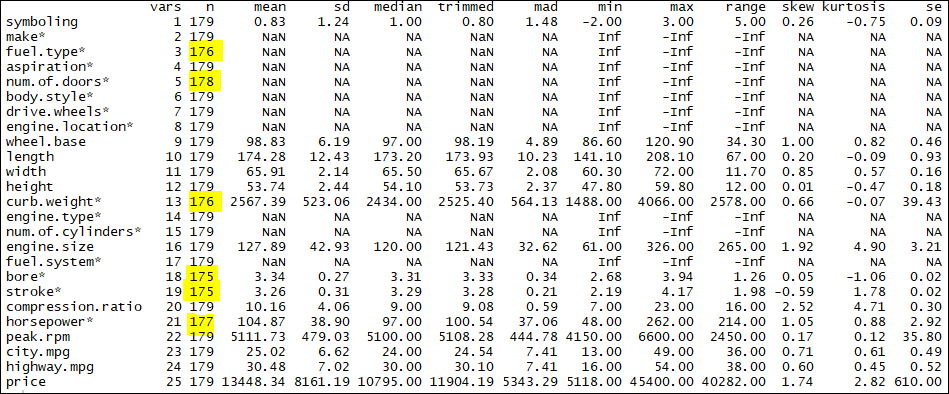
1. **Exploratory Analysis:**

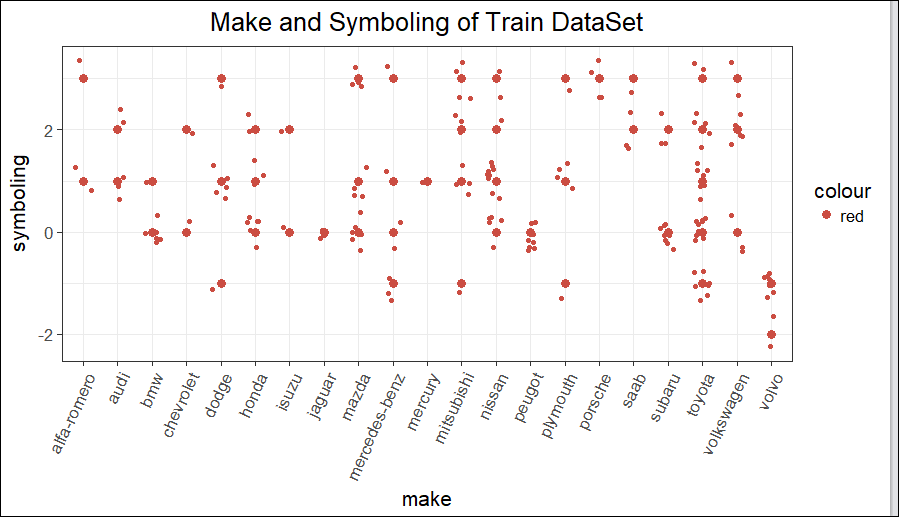
The primary data analysis within this study will be performed in both R and SAS. This will include reviewing the different explanatory variables applicable to all pertinent observations within the data set. Outliers, potential transformations and situations of multicollinearity will be reviewed. Specific questions pertaining to the vehicles and any price predictions will be answered and performed in SAS in sections VI and VII.

Preliminary Data Analysis in R:

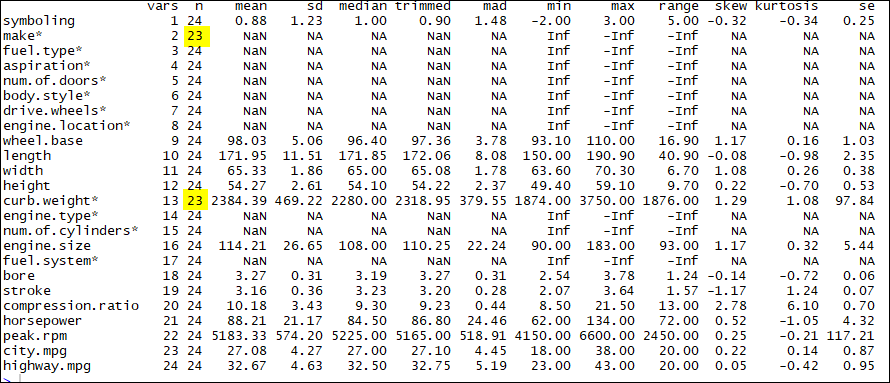
A statistical summary of the Train and Test data sets, as well as a scatterplot of the correlation between the make and symboling variables can be found below.

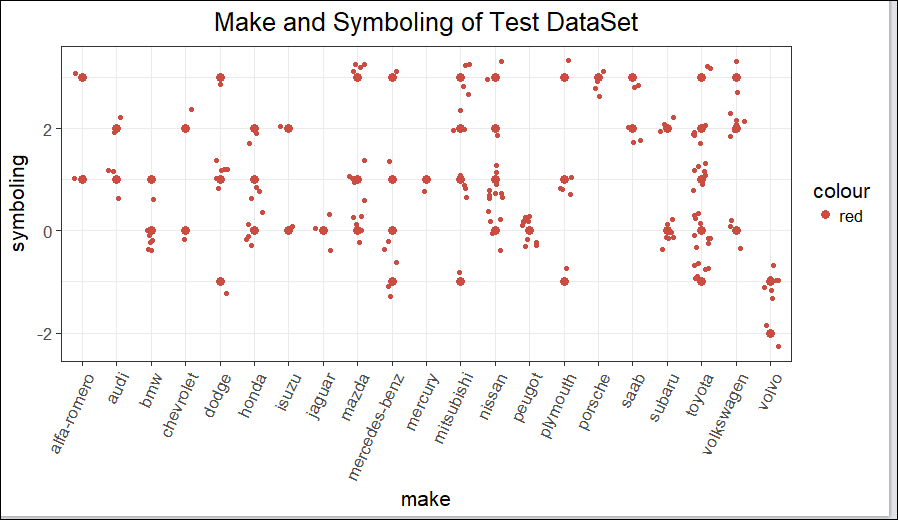
Train dataset with missing attributes columns highlighted. *(n = sample population of observational data)*





Test dataset with missing attribute columns highlighted. *(n = sample population of observational data)*



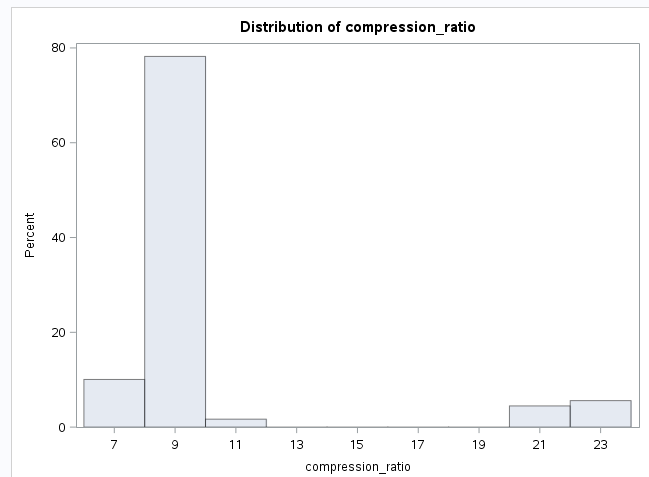


Preliminary Data Analysis in SAS:

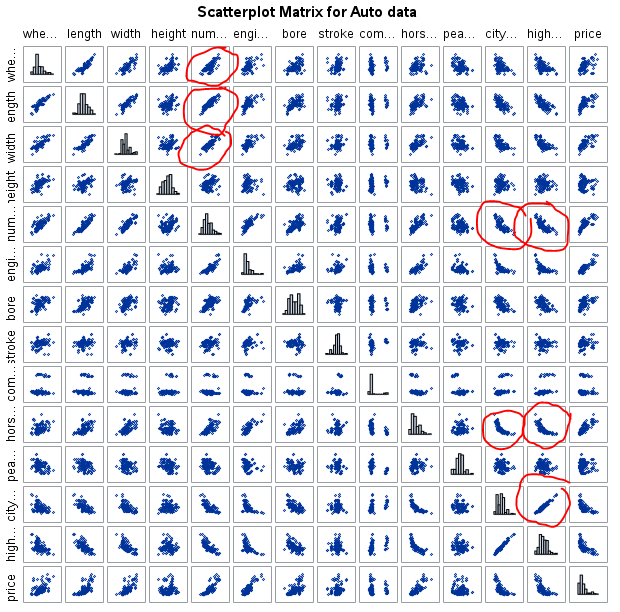
Judging from histograms within the preliminary SAS analysis, engine size, horsepower, price may benefit from some transformation due to skewness.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Compression ratio is a bimodal function and likely should not be treated as a simple continuous variable.



A simple scatter plot indicates that there are possible strong interactions between some of the variables. For example, curb weight and car dimensions are obviously correlated. Also, highway and city mileage are very strongly correlated. We should consider maybe dropping one the variables.



1. **Fit Models to Impute:**
2. **Toyota VP’s Question of Interest:**
3. **Predictive Model:**
4. **Conclusion:**
5. **Appendix:**

R code for Preliminary Analysis (section IV)



SAS code for Preliminary Analysis (section IV)

