MGSC 661 Final Project Report

Kelly Liu 260771986

1 Introduction

According to the most recent data from Transport Canada's National Collision Database (NCDB)[2], there were 1,768 cases of fatalities and 8,185 serious injuries caused by motor vehicle collisions in 2021. The Canadian government is working to improve road safety to have the safest roads in the world. The insurance company, which needs to pay for the car damage is also concerned about these road accidents. A risk factor called *symboling* was developed to represent the insurance risk rating of each car. Cars with a higher positive *symboling* value are considered riskier. A negative symboling value suggests a safer car. It would be interesting to see which features are commonly associated with lower risk. This analysis aims to analyze the relationship between *symboling* and other features like automobile specifications to provide insights to car buyer and owner, insurance companies, and the government on what factors might contribute to a car being considered risky.

2 Data Description and Exploratory Analysis

The dataset [1] includes specifications of automobiles, their assigned insurance risk rating, and normalized losses. It offers technical details like *engine size* and *horsepower*, design elements such as body style and dimensions, as well as insurance-related metrics indicated by *symboling* and *normalized losses*. With a mix of both categorical and numeric variables, the dataset offers a great opportunity to uncover underlying patterns, assess vehicle safety ratings, and evaluate factors influencing insurance risk.

I explored the data by checking the distributions of the categorical and numerical variables with histograms, boxplots, *table()*, and *skewness()* functions. Please find the histogram and boxplot diagrams in the Appendix - Key Variable Distributions Plots and Appendix - Other Variable Distributions Plots. For the Skewness, please refer to Appendix - Skewness Analysis.

2.1 Variable Distributions and Skewness

Among the categorical variables, *make* shows a diverse range of manufacturers, with Toyota being the most common. The *fuel type* is predominantly gas, while *aspiration* is majorly standard, barring a few turbo variants. The *number of doors* displays a balanced mix between two and four-door models. It was being converted to numeric data later on. The *number of cylinders* is a mix among eight, five, four, six, three. This variable was also converted to numeric data. There are 5 *engine types*: dohc (Dual OverHead Cam), l (L engine), ohc (OverHead Cam), ohcf (OverHead Cam and Valve F engine), ohcv (OverHead Cam and Valve).

For numeric variables, the *wheel base* ranges from 86.6 to 115.6, with a relatively symmetric distribution around the mean of 98.26. The *length*, *width*, and *height* exhibit varying dimensions of the cars with fairly symmetric distributions, suggesting a diverse set of designs and sizes. MPG stands for miles per gallon, which tells how many miles a car can go on a gallon of fuel. *City MPG* and *Highway MPG* are moderated skewed. *Curb weight* is the weight of an automobile without occupants or baggage, ranging significantly from 1488 to 4066, indicating a wide variance in car builds and potentially different classes from sports to utility vehicles. The *engine size* is highly skewed, with a broad range of values from 61 cc to 258 cc, indicating a diverse collection of vehicles from compact to larger, more powerful models. The *price* variable in the dataset is highly skewed, ranging from $5,118 to $35,056, with a median price of $9,233 and a mean of $11,446, illustrating a wide price spectrum with individual occurrences across various price points, indicative of a diverse array of vehicles from budget to luxury models. The *horsepower* variable is moderately skewed, exhibiting a moderate skewness with a value of 0.908, spans a range from 48 to 200, with the most frequent values clustered around the lower to mid-range, reflecting a dataset predominantly composed of the standard to moderately powered vehicles. The *compression ratio* is highly skewed, with values ranging from 7 to 23, with a significant concentration at common ratios like 9, indicating a mix of both standard and high-compression engines, typical in both conventional and performance-oriented vehicles. The *symboling*, the insurance risk rating, has values ranging from -2 to 3, suggesting 6 types of risk profiles among the cars. *Normalized losses* have a minimum of 65 and a maximum of 256, with a mean around 121.1, hinting at a moderate right-skewed distribution.

2.2 Correlations

According to the Correlations Table, the variable pairs that have correlation coefficient ≥ 0.8 are Wheel Base and Length, Wheel Base and Curb Weight, Length and Curb Weight, Length and Width, Width and Curb Weight, Width and Engine Size, Curb Weight and Engine Size, Engine Size and Horsepower, City MPG and Highway MPG. Please find the correlation matrix in the Appendix - Correlation.

3 Model Selection

To achieve the objective of finding the relationship between *symboling* and other features like automobile specifications, I planned to use PCA to find the influence of selected features on *symboling*. Followed by training a Random Forest Model to classify *symboling* based on the car specifications selected with the help of PCA.

3.1 Principal Component Analysis (PCA)

Based on the data exploration, I first ran a PCA analysis with all predictors, however, the initial PCA with all variables resulted in an overly complex plot. Therefore, by integrating the feature selection technique and domain knowledge, I came up with a more interpretable result and ends up with higher random forest classification accuracy rate.

I employed Recursive Feature Elimination (RFE), a wrapper method that combines the predictive power of Random Forest with cross-validation to ensure model robustness and generalizability. The RFE process iteratively creates models and removes the weakest features until the optimal set of features is determined. The outcome of RFE indicated a subset of features that yielded the best trade-off between model complexity and performance. The subset features are *make*, *wheel base*, *normalized loss*, *width*, *number of doors*. I added additional features based on domain knowledge, including *peak RPM*, *price*, *length*, *horsepower*, *city MPG*, *number of cylinders*, *compression ratio*, *engine size*, *curb weight*, were manually included for determining the influence on *symboling*.

After feature selection, Principal Component Analysis (PCA) was conducted on the numeric variables to reduce dimensionality while retaining the variance within the dataset. PCA helped to reduce the number of variables to a smaller set that still captured the most important information. The PCA scores were then visualized to observe the clustering of different *symboling* types, providing a visual representation of the underlying structure of the data. The coloring was scaled based on the *symboling*, from navy (low risk) to salmon (high risk), with white representing the median risk.

3.2 Random Forest Classification Model

Then a Random Forest model was implemented to classify *symboling*. The Random Forest's ensemble nature allowed for a nuanced understanding of the relationship between vehicle attributes and their associated risk ratings. The approach began by ensuring *symboling* as a categorical variable, which is fundamental since *symboling* represents discrete insurance risk categories. Tested the model performance with features selected with RFE, the feature selected by domain knowledge, and the reliable features from PCA. The accuracy scores are all the same, hence, features selected by RFE and domain knowledge were used to train the Random Forest model. *Normalized losses* were excluded from the predictors because it does not describe the car specification.

The dataset was split into training and test sets, with 80% of the data allocated for training. The train test data split allows models to learn from a substantial portion of the data while reserving a separate portion for unbiased evaluation. For hyperparameter tuning, the number of trees to grow was tested to have the best accuracy rate at 500.

The model's accuracy was evaluated using a confusion matrix, the table to describe the performance of a classification model. Additionally, an assessment of which features were most important in classifying *symboling* was done by examining the model's variable importance scores.

4 Results

4.1 Principal Component Analysis (PCA)

According to the Appendix - Principal Components Table, in the first principal component (pc1), the factors with higher weights are *wheel base*, *width*, *price*, *length*, *horsepower*, *city MPG*, *engine size*, and *curb weight*. This means that they are the most important features that capture the differences among observations. PC1 seems to be capturing a general car dimension and engine power-train factor and has the highest proportion of variance explained (PVE) of 53.3%, which can refer to the Appendix - PVE Plot. PVE is calculated for each principal component and represents the proportion (or percentage) of the total variance in the dataset that is explained by that component. The negative signs indicate that as the negative values increase, the value of PC1 decreases. It means that higher values of these variables might be associated with a lower *symboling* value, indicating a lower risk. Vice versa for the variables with positive signs.

The points in the Appendix - PCA Plot represent *symboling* rates, colored salmon for positive rates, and navy for negative rates. The plot shows which factors most strongly influence where the *symboling* end up. It helps to understand which factors are most important in distinguishing among automobiles with different risk levels and could guide automobile risk assessment and road safety regulations. The spread of points shows that automobiles with positive *symboling* (salmon dots) cluster together towards the top right, and the automobiles with negative *symboling* (navy dots) cluster together towards the bottom left. The automobiles in each group have similar factor profiles based on the principal components analyzed. The arrows on the plot represent the 13 factors. The direction of an arrow shows which way the factor 'pulls' the data, and the length shows how strong the factor's influence is. For instance, in the direction of where the most navy points are, we could infer that higher values of the bottom left 8 factors from *engine size* to *compression* *ratio* are more related to the negative *symboling*.

Generally speaking, variables that are the highly reliable indicators for automobiles are: *engine size*, *price*, *curb weight*, automobile size (*width & length*), *wheel base*, *number of doors*, *compression ratio*, *peak RPM*.

4.2 Random Forest Classification Model

The model's performance, measured by accuracy, achieved a score of 81.25%. This indicates that the model was able to correctly predict the insurance risk rating for a substantial majority of the cars in the test set, suggesting that the model is well-trained and capable of making reliable predictions.

Then I calculated the variable importance scores and visualized them with a bar chart referred to Appendix - Feature Importance Bar Chart. The MeanDecreaseGini metric, a measure of how each feature contributes to the homogeneity of the nodes and leaves in the model, revealed that *make*, *wheel base*, *width*, and *curb weight* were among the most influential features. This suggests that these variables are key drivers in determining a vehicle's insurance risk rating. For instance, the *make* of a car could be indicative of its overall build quality and safety features, which are critical in risk assessment. Similarly, *wheel base* and *width* might be proxies for the vehicle's size and stability, influencing its risk profile. Please find the key driver's distribution by *symboling* in the Appendix - Key Variable Distributions by Symboling Plots. The lower importance of features like *number of cylinders* and *city MPG* suggests that while they contribute to the model, their impact is less significant compared to other features. This insight can guide insurance companies in focusing on the more influential factors when assessing a vehicle's risk rating.

5 Recommendations and Conclusion

The *symboling* classification analysis, has led to insightful findings. The risk rating, pivotal for automobile owners, insurance companies, and regulatory bodies, correlates with various features of automobiles. This study involved exploring a dataset rich in automotive specifications and insurance risk metrics. By employing a mix of statistical techniques and machine learning models, particularly the PCA and Random Forest algorithm, the key insights that hold significant value for stakeholders are revealed.

5.1 Insights from the PCA and Random Forest Model

According to the Appendix - Principal Components Table and Appendix - PCA Plot, the major risk divers are smaller *engine size*, lower *price*, less *curb weight*, smaller automobile size (*width & length*), smaller *wheel base*, less *number of doors*, smaller *compression ratio*, higher *peak RPM*. The Random Forest model achieved at classifying *symboling* with a moderately high accuracy of 81.25%, which proves the model's capability to distinguish between different risk categories effectively. Key variables that emerged as influential in determining the symboling are *make*, *wheel base*, *width*, and *curb weight* (refer to Appendix - Feature Importance Bar Chart). The prominence of these features suggests that the physical attributes and the car manufacturer play a critical role in the risk assessment of a vehicle.

5.2 Recommendations for Automobile Buyers and Owners

The first advice to people when purchasing a vehicle, whether a new car or a second-hand car, is to consider not only the price and aesthetics but also the reliable car manufacturer, wheelbase, width, and curb weight. These factors significantly impact the car's safety rating and insurance costs. Vehicles with larger dimensions and greater weight, featuring lower peak RPM, and originating from manufacturers like Volvo, Jaguar, and Peugeot, known for their emphasis on safety, are likely to be more cost-effective over time. This knowledge can be crucial when budgeting for the total cost of owning a car, including insurance. Being informed about various factors, including the car's make and model, while choosing a vehicle with a lower risk rating can be a proactive step in managing potential risks associated with driving. Moreover, with an understanding of how factors affect insurance rates, people will be better positioned to negotiate with insurance providers.

In addition to the car's features, how car owners drive plays a vital role in road safety. Practicing safe driving habits, like not pushing the engine to its maximum capacity and maintaining the speed according to signs in the city and highways, not too slow and not too fast, can lower the chances of accidents. Careful driving behavior can also positively influence your car's insurance risk rating and reduce losses over time. Insurance companies often provide discounts to drivers who maintain good driving records.

5.3 Recommendations for Insurance Companies

The insights from the analysis can significantly aid insurance companies in enhancing their *symboling* classification process. Typically, vehicles are initially assigned a *symboling*, a risk factor symbol linked to their price. This symbol is then adjusted higher or lower depending on the perceived risk.

The use of PCA and Random Forest modeling offers insurance companies a pathway to make more informed decisions regarding *symboling* rankings. These companies should focus on the key features identified in our model, especially the make, dimensions, and attributes of a vehicle, to achieve a more accurate risk assessment. Also, incorporating these variables into the pricing models could lead to more precise premium calculations, accurately reflecting each vehicle's actual risk.

To maintain relevance and accuracy, I recommend that insurance companies periodically re-train and fine-tune the Random Forest model. This approach ensures that classifications stay current and reliable, adapting to changes in vehicle profiles and market trends.

5.4 Recommendations for Government Regulations

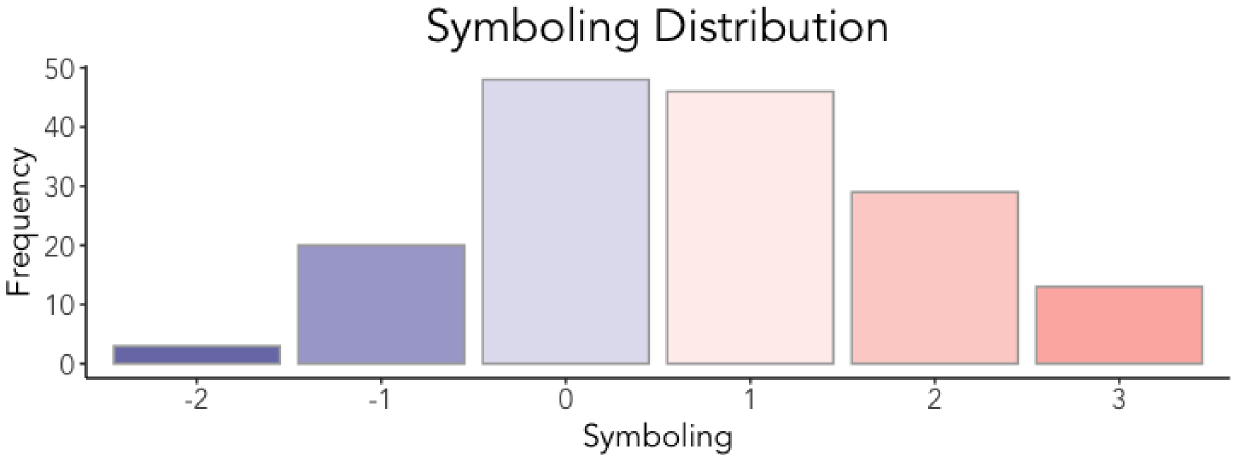
The government can enhance road safety by adjusting speed limits according to engine capacities and key features such as peak RPM, especially considering their correlation with risk levels. Additionally, regulating certain types of vehicles based on their features could be beneficial. For instance, as seen in the City MPG by Symboling Distribution in the Appendix - Key Variable Distributions by Symboling, vehicles with a city MPG lower than 20 are significantly riskier. This risk factor might be linked to specific vehicle characteristics, such as medium width and length combined with a lower height and just two doors after diving deep back into the dataset. These vehicles often possess high horsepower, resembling sports cars. Implementing targeted regulations for such vehicle categories could substantially improve road safety.

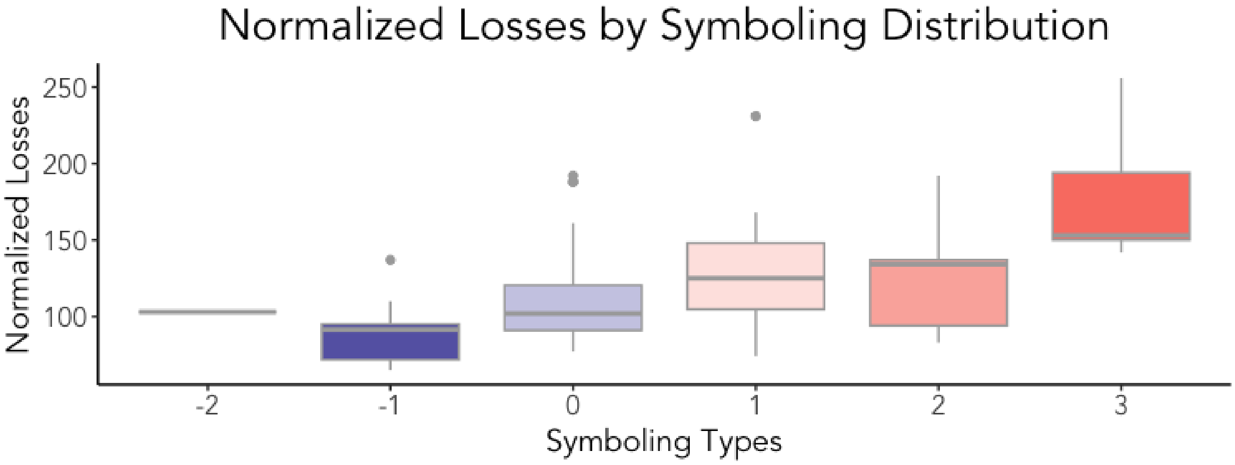
5.5 Conclusion

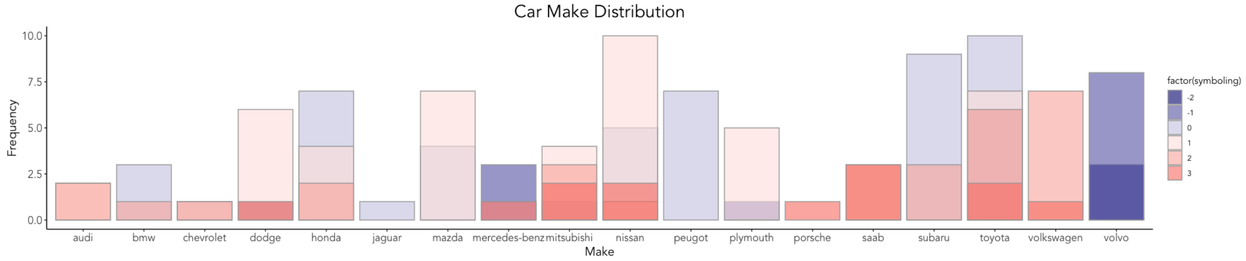
The study of the *symboling* attribute has shed light on the complex relationship between various car features and their associated insurance risks. These insights are invaluable for insurance companies, providing a basis for more refined risk assessment methods. For governmental authorities, this information supports a more informed approach to improving road safety regulations. Car buyers and owners can also benefit significantly, by gaining knowledge on selecting safer vehicles and adopting safer driving habits. Overall, this analysis not only enhances our understanding of the current situation but also lays the groundwork for future improvements in vehicle safety and insurance procedures. In terms of future opportunities, the accuracy of the model could potentially be further enhanced with additional data, such as the year of car manufacture and the age of the car owner...

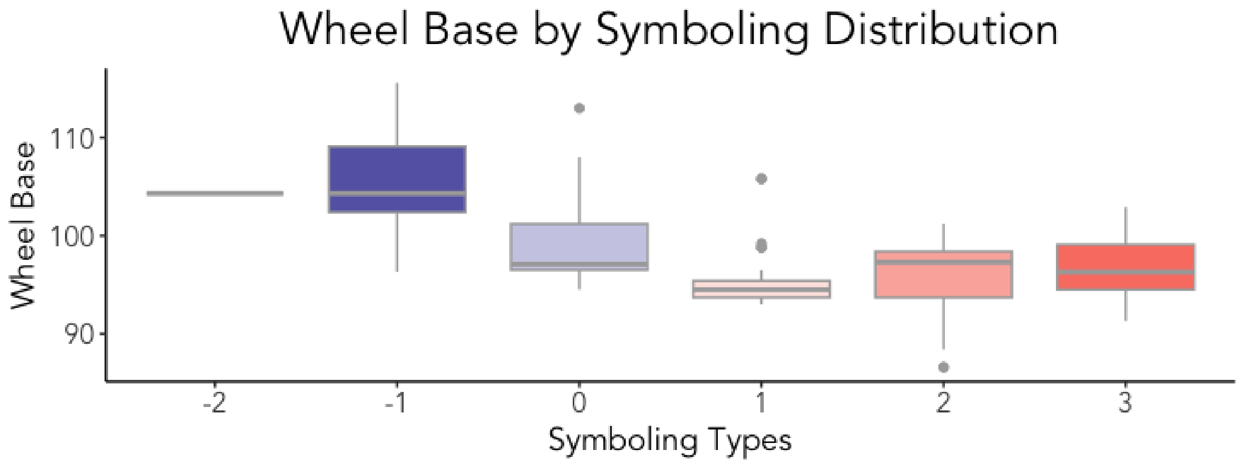
6 Appendices

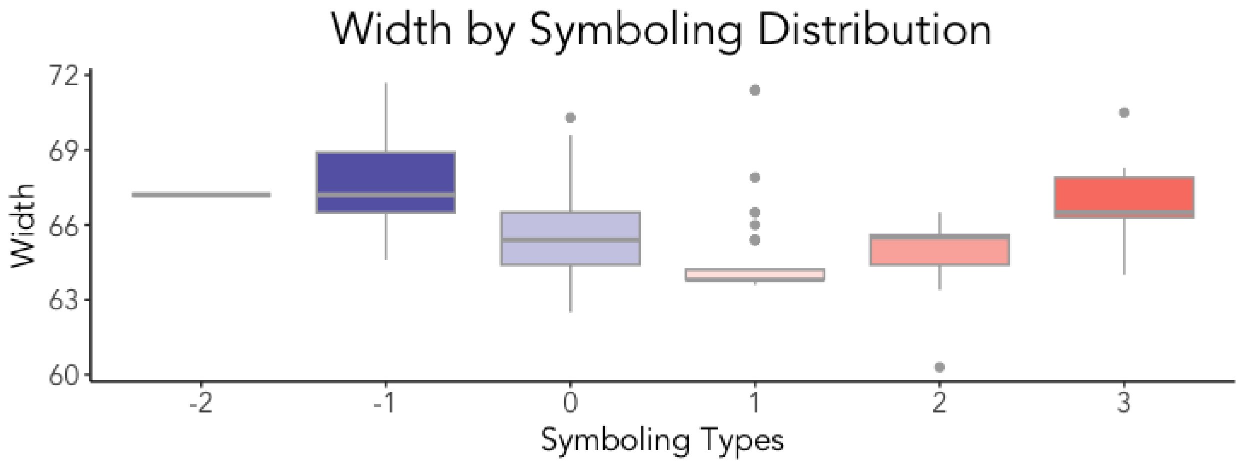
6.1 Appendix - Key Variable Distributions by Symboling

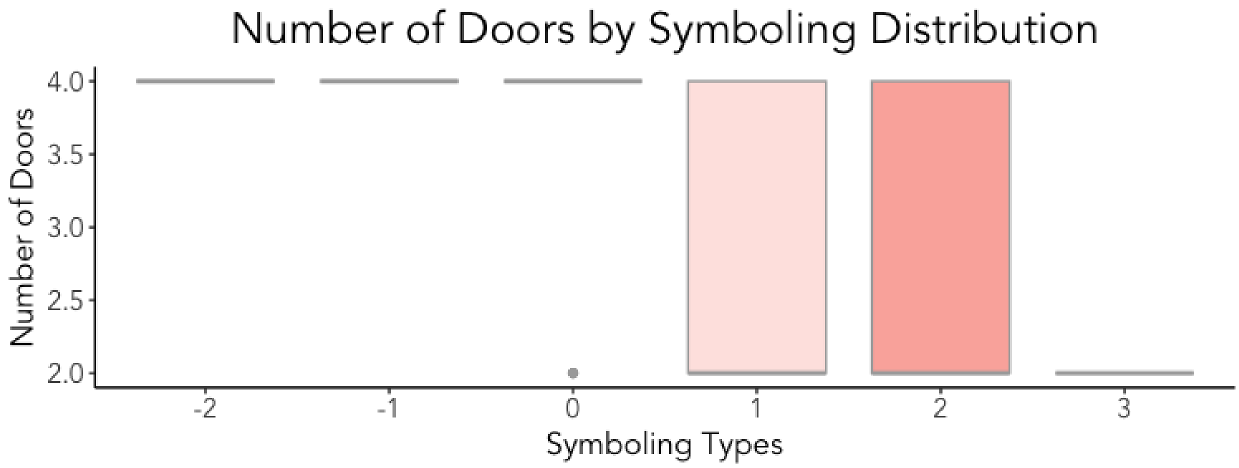


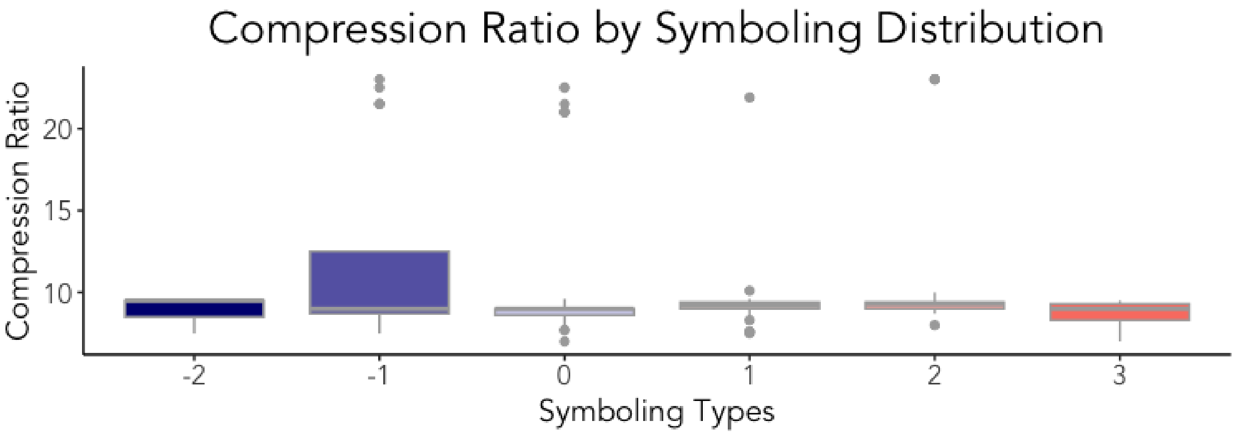


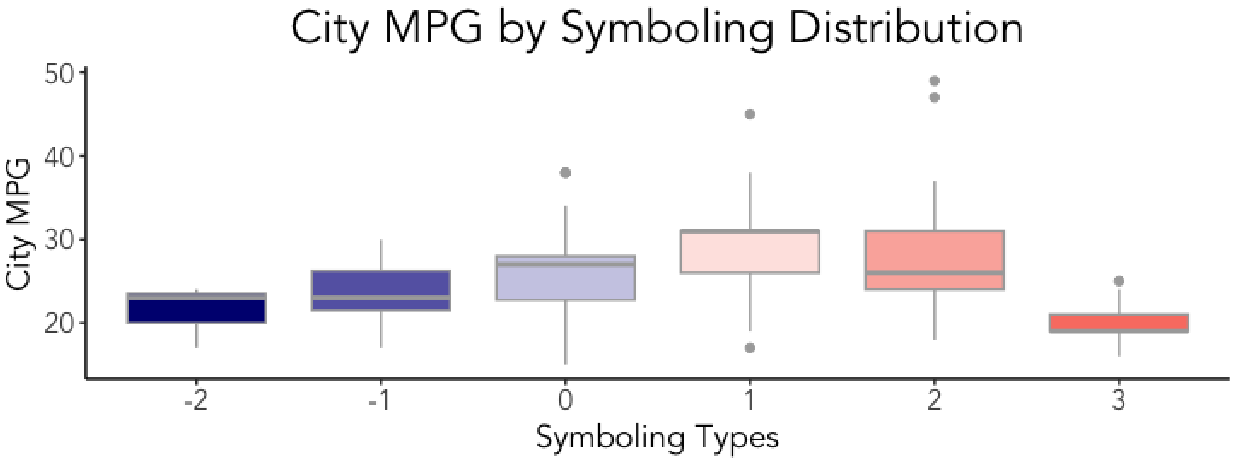


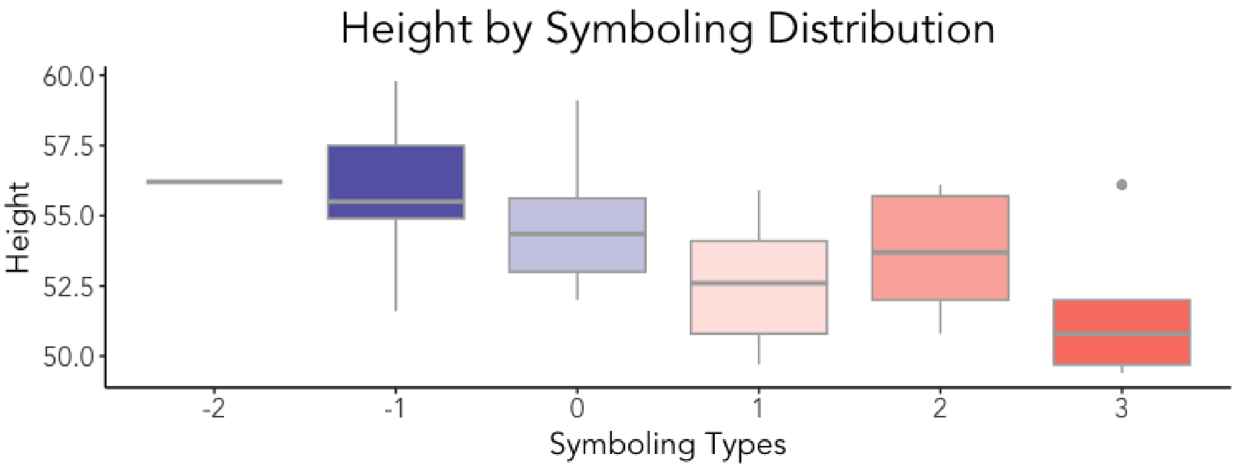

Wheel Base by Symboling Distribution
Wheel Base
110
100
90
-2
-1
0
1
2
3
Symboling Types

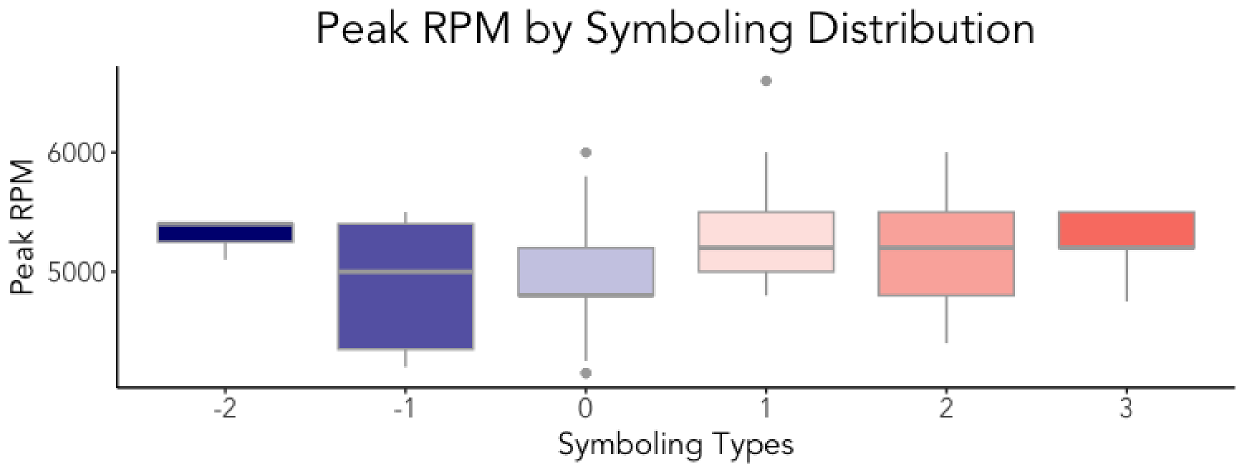

Width by Symboling Distribution
Width
72
69
66
63
60
-2
-1
0
1
2
3
Symboling Types


Number of Doors by Symboling Distribution
Number of Doors
4.0
3.5
3.0
2.5
2.0
-2
-1
0
1
2
3
Symboling Types

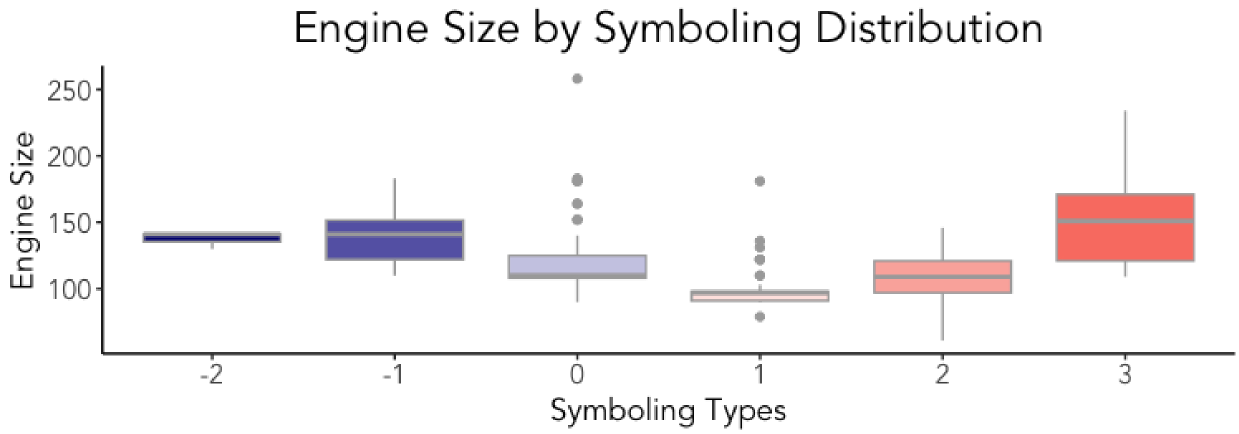

Compression Ratio by Symboling Distribution
Compression Ratio
20
15
10
-2
-1
0
1
2
3
Symboling Types


City MPG by Symboling Distribution
City MPG
50
40
30
20
-2
-1
0
1
2
3
Symboling Types

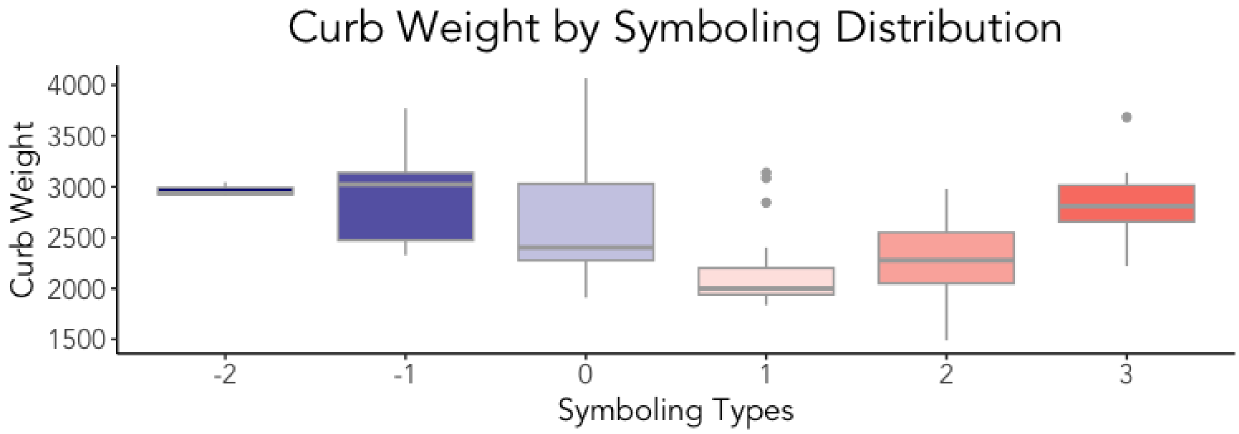

Height by Symboling Distribution
Height
60.0
57.5
55.0
52.5
50.0
-2
-1
0
1
2
3
Symboling Types



Peak RPM by Symboling Distribution

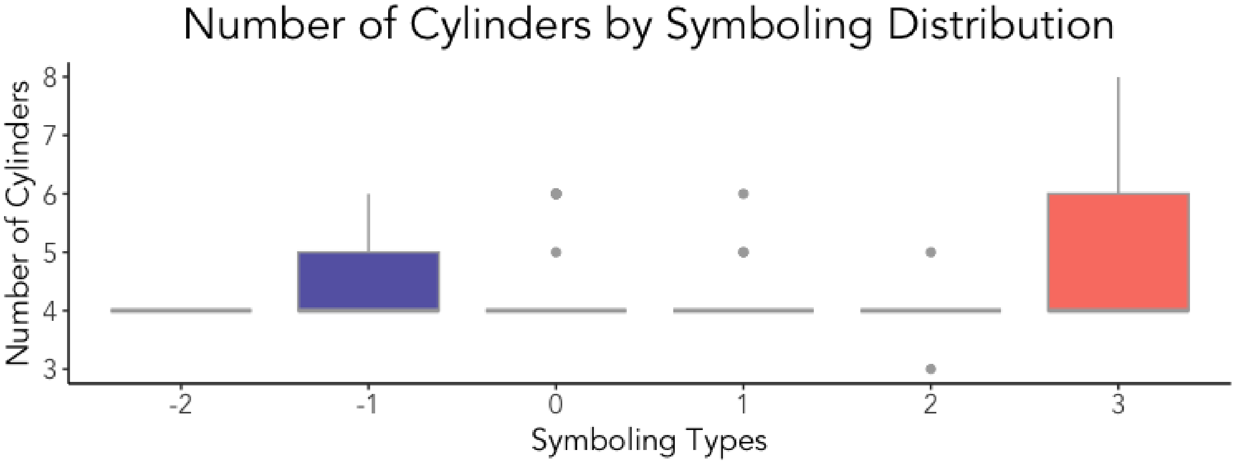


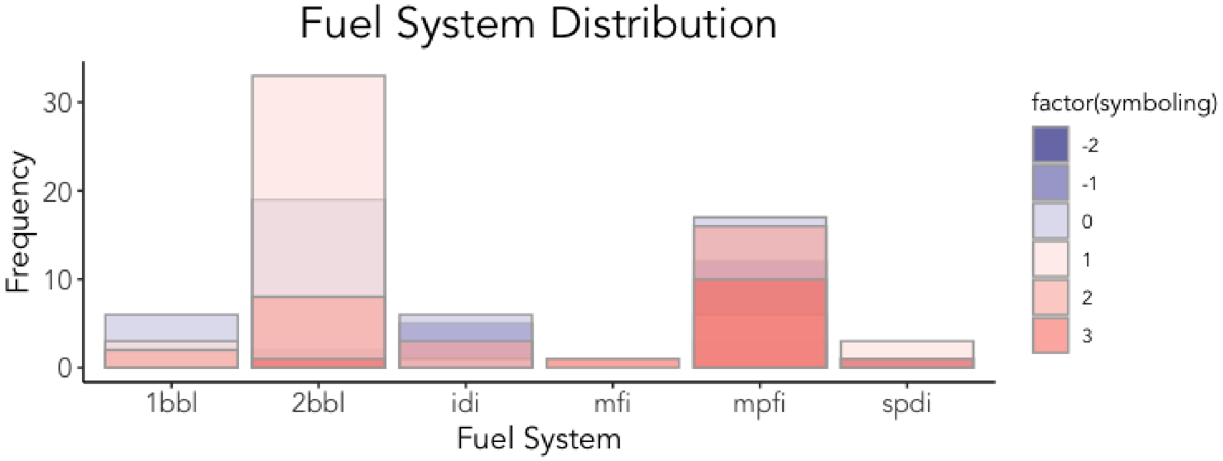
Engine Size by Symboling Distribution

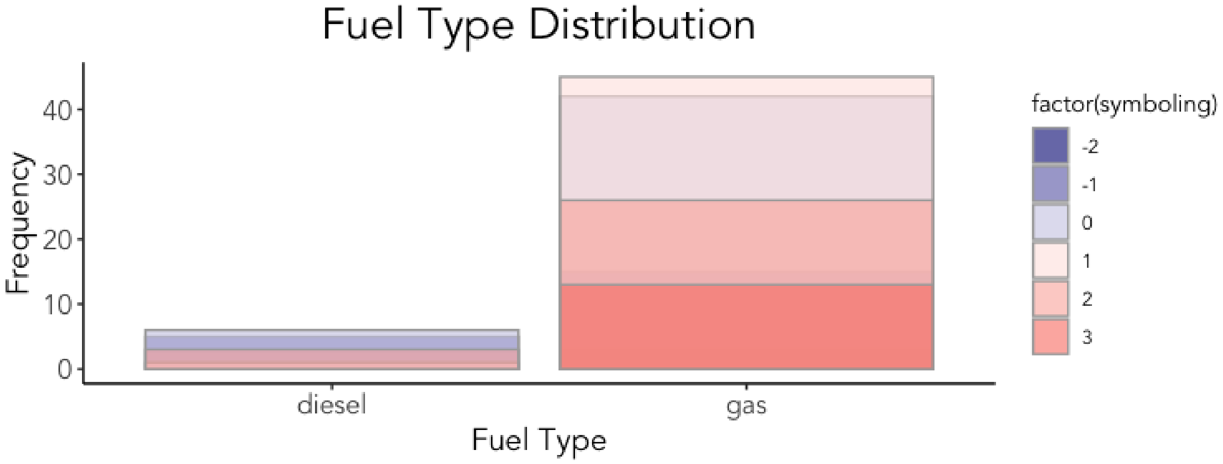


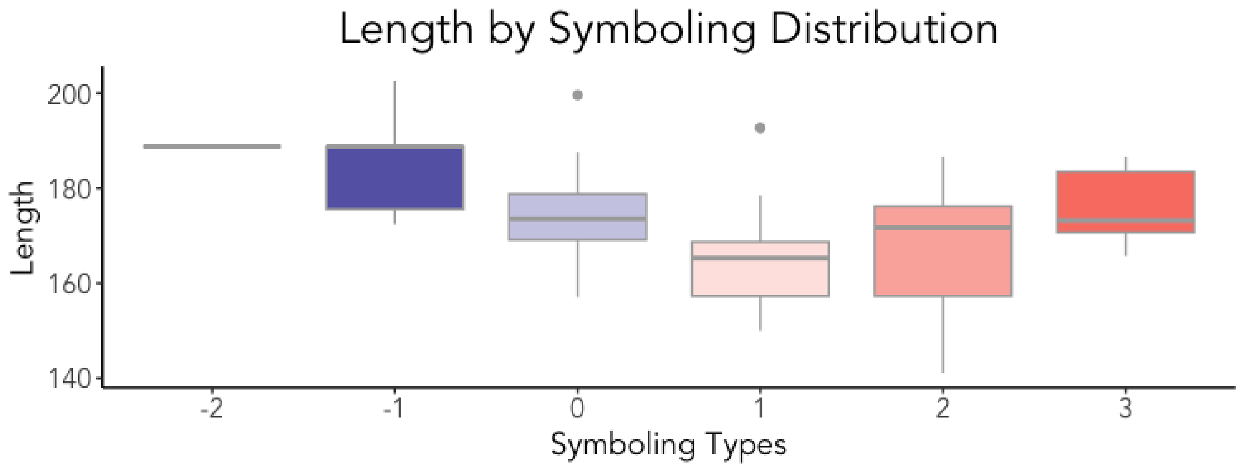
Curb Weight by Symboling Distribution

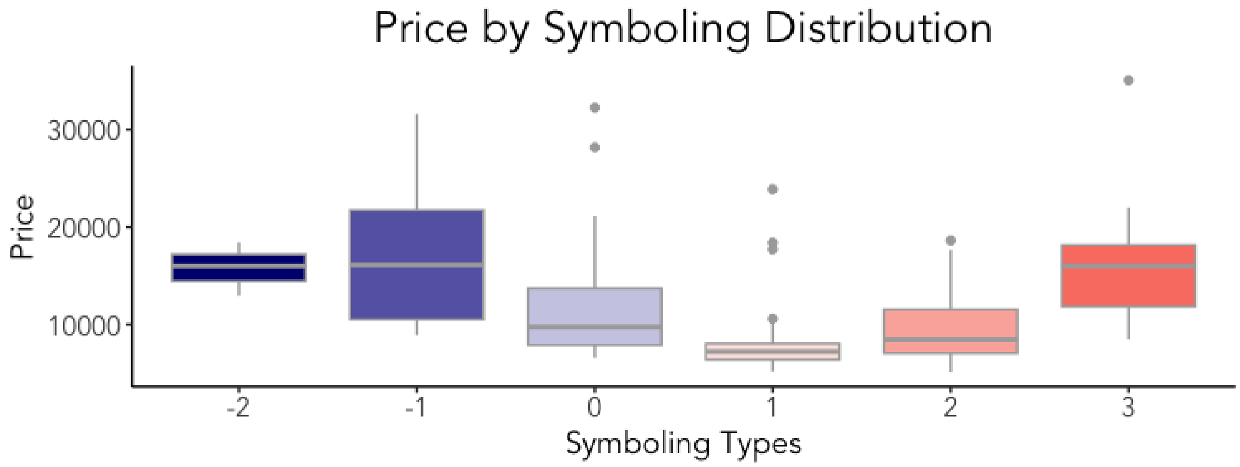
6.2 Appendix - Other Variable Distributions

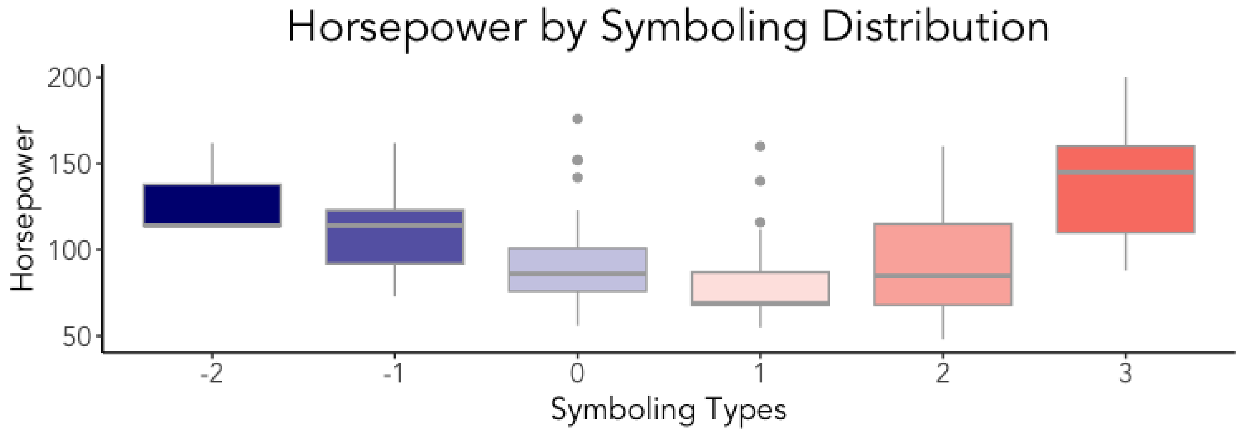




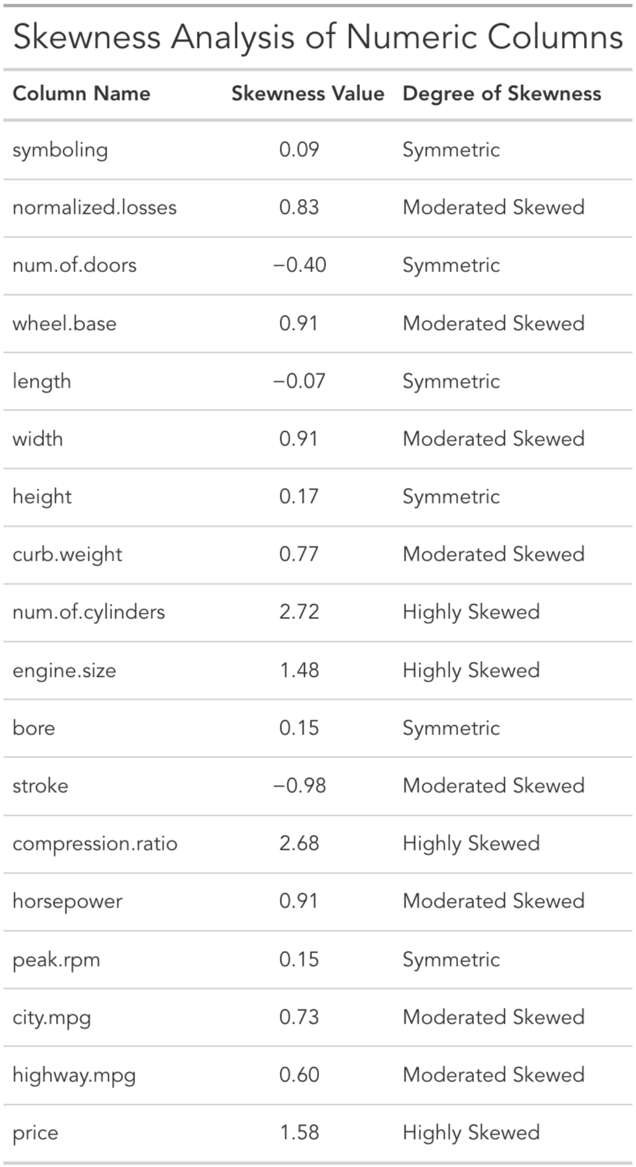



Length by Symboling Distribution
Length
200
180
160
140
-2
-1
0
1
2
3
Symboling Types


Price by Symboling Distribution
Price
30000
20000
10000
-2
-1
0
1
2
3
Symboling Types


Horsepower by Symboling Distribution
Horsepower
200
150
100
50
-2
-1
0
1
2
3
Symboling Types

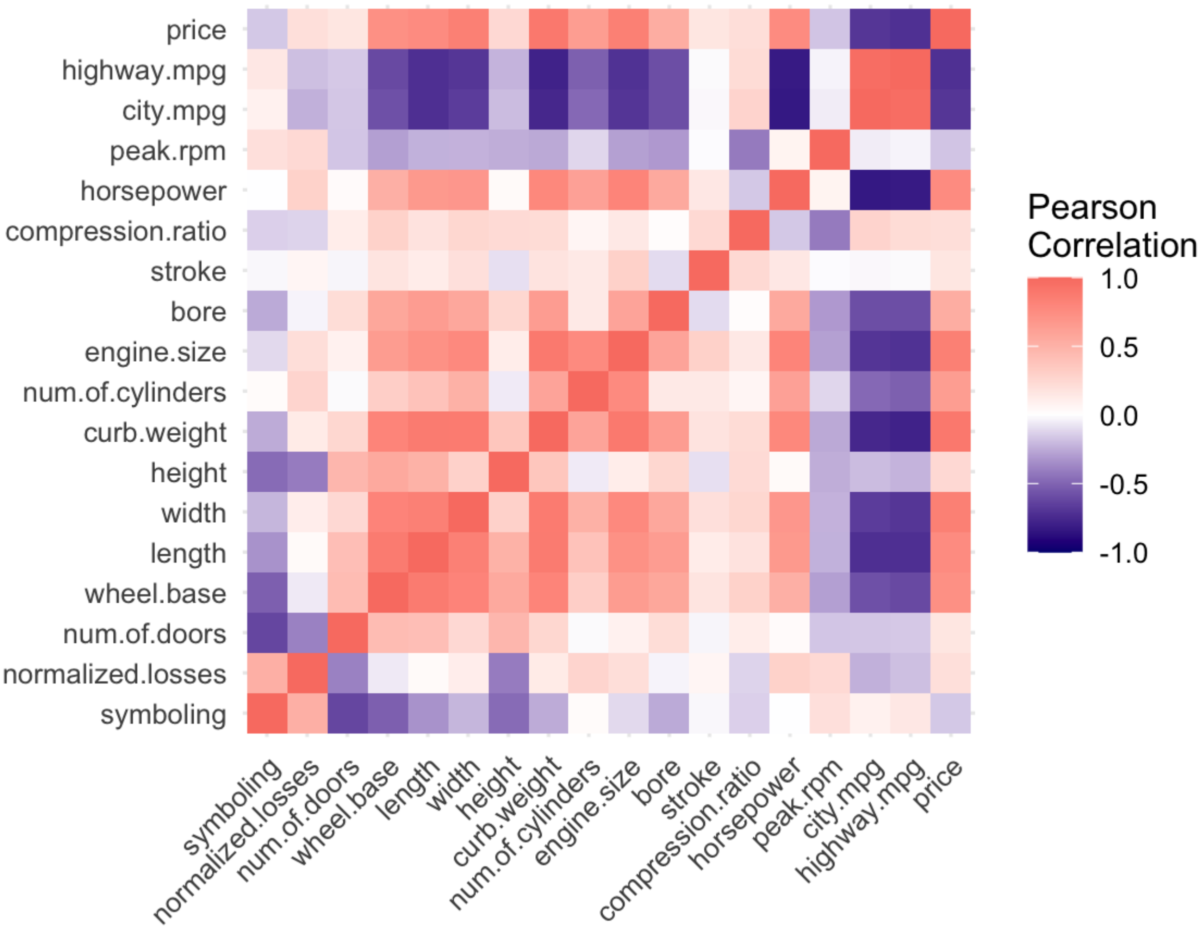
6.3 Appendix - Skewness Analysis



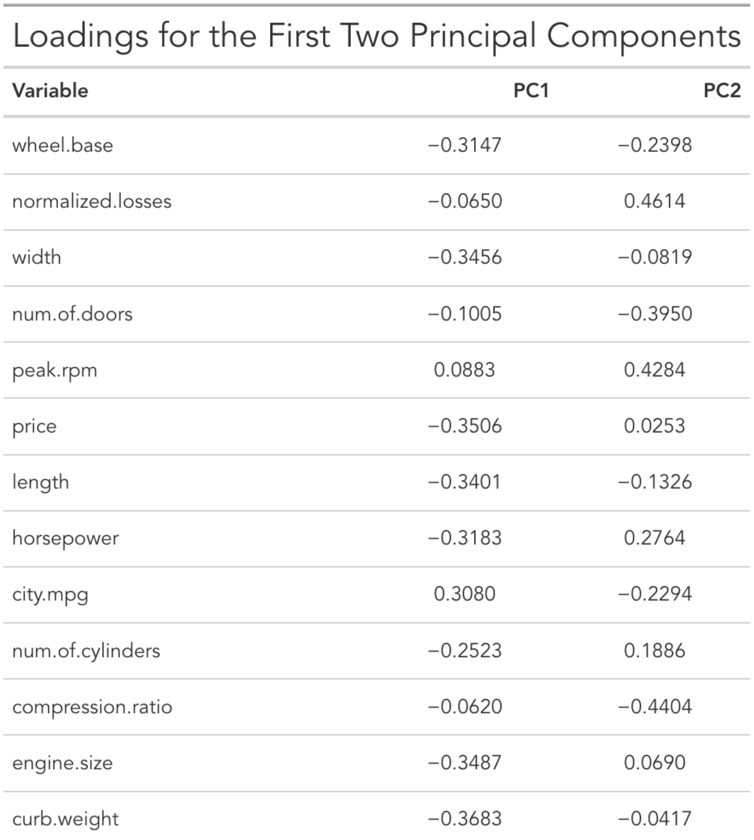
Skewness Analysis of Numeric Columns

Column Name	Skewness Value	Degree of Skewness
symboling	0.09	Symmetric
normalized.losses	0.83	Moderated Skewed
num.of.doors	−0.40	Symmetric
wheel.base	0.91	Moderated Skewed
length	−0.07	Symmetric
width	0.91	Moderated Skewed
height	0.17	Symmetric
curb.weight	0.77	Moderated Skewed
num.of.cylinders	2.72	Highly Skewed
engine.size	1.48	Highly Skewed
bore	0.15	Symmetric
stroke	−0.98	Moderated Skewed
compression.ratio	2.68	Highly Skewed
horsepower	0.91	Moderated Skewed
peak.rpm	0.15	Symmetric
city.mpg	0.73	Moderated Skewed
highway.mpg	0.60	Moderated Skewed
price	1.58	Highly Skewed

6.4 Appendix - Correlation



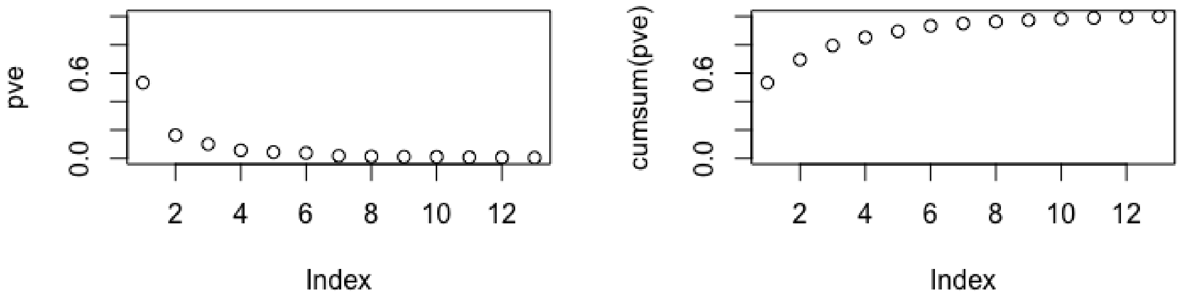
6.5 Appendix - Principal Components Table



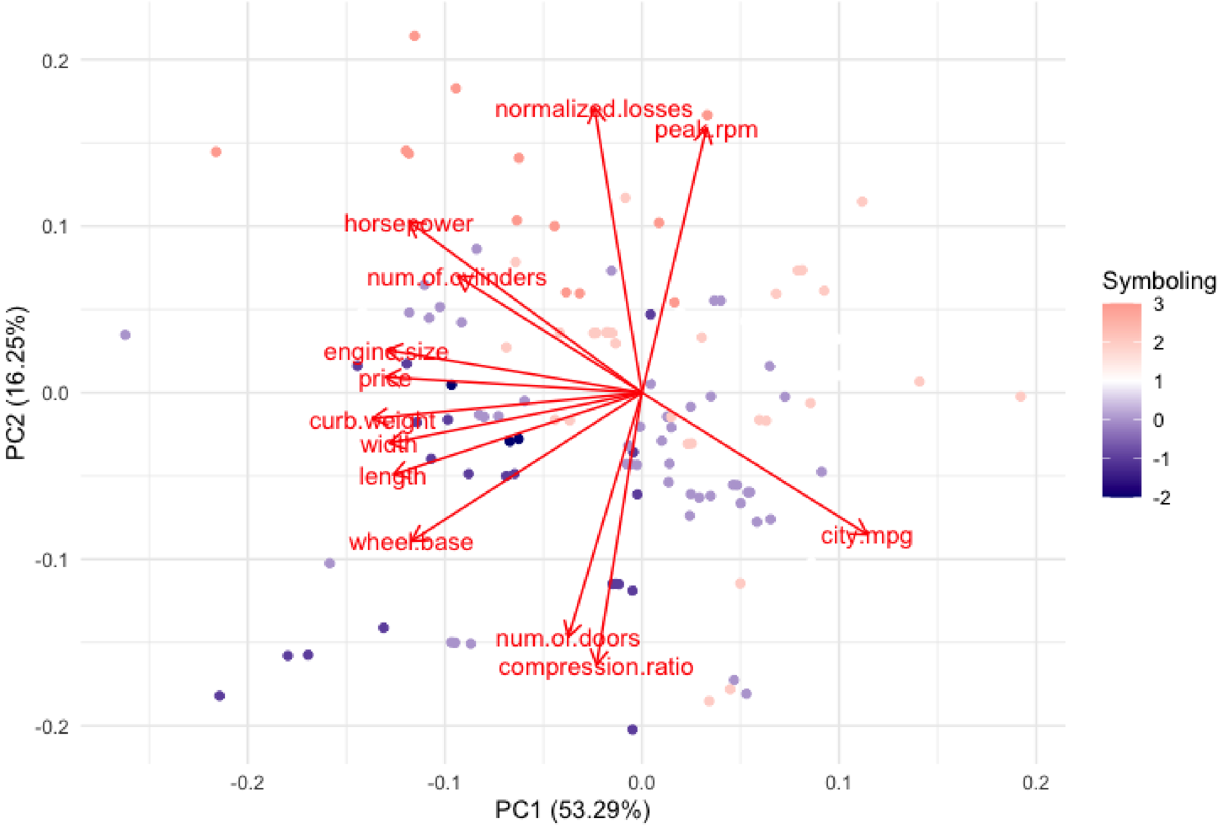
Loadings for the First Two Principal Components

Variable	PC1	PC2
wheel.base	−0.3147	−0.2398
normalized.losses	−0.0650	0.4614
width	−0.3456	−0.0819
num.of.doors	−0.1005	−0.3950
peak.rpm	0.0883	0.4284
price	−0.3506	0.0253
length	−0.3401	−0.1326
horsepower	−0.3183	0.2764
city.mpg	0.3080	−0.2294
num.of.cylinders	−0.2523	0.1886
compression.ratio	−0.0620	−0.4404
engine.size	−0.3487	0.0690
curb.weight	−0.3683	−0.0417

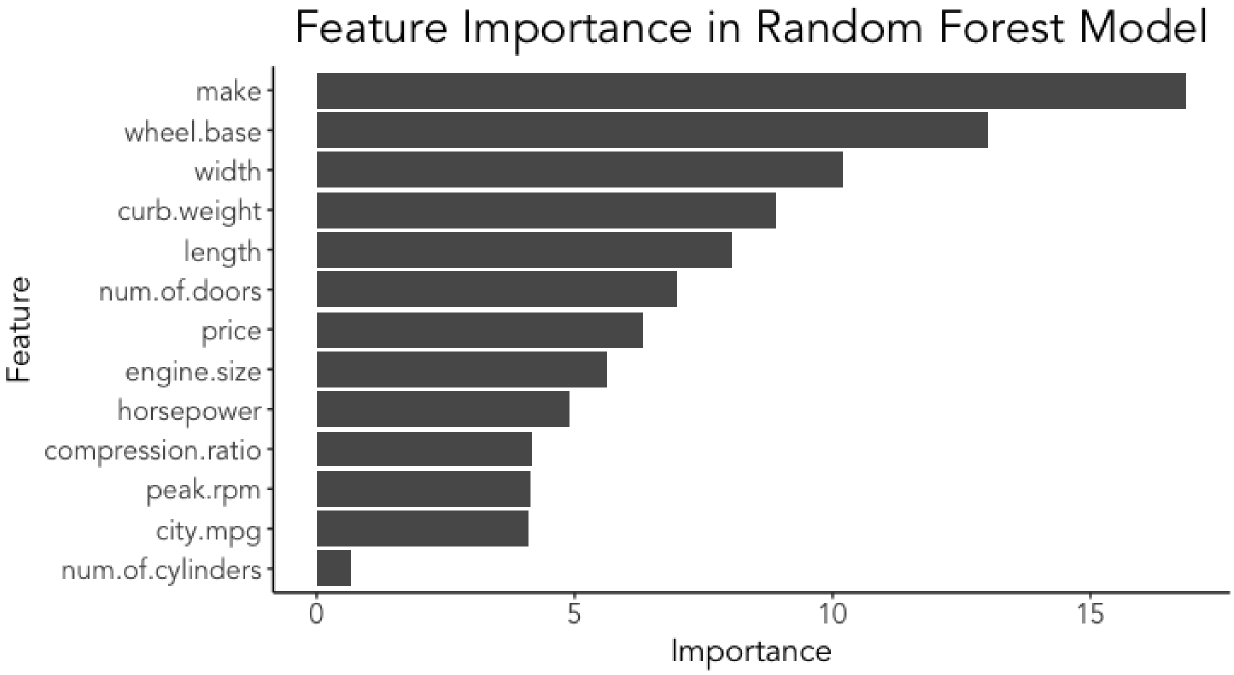
6.6 Appendix - PVE Plot



6.7 Appendix - PCA Plot



6.8 Appendix - Feature Importance Bar Chart



7 Code

The code is stored in the file called: ***Automobile.R***.

References

[1] Kaggle (2016). Automobile Dataset. https://www.kaggle.com/datasets/toramky/automobile-dataset/data

[2] Transport Canada (2021). Canadian Motor Vehicle Traffic Collision Statistics 2021. https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2021