Final Project

Analyzing Car Risk: Predictive Modeling and Insights in the Automotive Domain

Éloi Dallaire (260794674)

Desautels Faculty of Management, McGill University

MGSC 401: Statistical Foundations of Data Analytics

Professor Juan Camilo Serpa

December 10, 2023

1. Introduction

Embarking on the analysis of the Automobile dataset, our journey begins with a meticulous exploration of vehicle features and their impact on the elusive concept of car risk. Our initial focus was on fine-tuning and interpreting the Gradient Boosting Machine (GBM) Classification model. Despite encountering challenges, particularly due to the dataset's modest size, we delved into Exploratory Data Analysis (EDA) to explore patterns and potential risk indicators. As we steer through this report, we'll show the nuances of our model-building process and the complexities of predicting car risk. Join us on this analytical expedition, where each insight help guiding us towards a comprehensive understanding of car risk assessment and its potential applications in the automotive landscape.

2. Data Pre-processing

The initial assessment of the dataset revealed a relatively clean state, minimizing the need for extensive pre-processing. The first step was to handle potential duplicated entries. This step ensured the integrity of the dataset, preventing any duplicative influence on subsequent analyses. Next was the handling of missing values, which required careful consideration. At first glance, the dataset looked complete and free from any missing values. After analyzing the variable types and each of their list of unique values, we quickly figured that some recoding was necessary. The problem was caused by the missing entries entered as '?', hence not recognized properly as missing (NA). Reconverting all the numerical variables to a numeric type was the solution we were looking for. Now we could properly analyze and handle the missing data. Considering the importance that the missing values were taking in the whole dataset (close to 28%), we favorize an approach that would not delete any missing data. Hence, numerical features such as 'normalized losses,' 'bore,' 'horsepower,' 'stroke,' 'peak-rpm,' 'city.mpg,' and 'price' were imputed with their respective means. This choice was substantiated by the symmetric distribution of these variables, making the mean a suitable alternative. On the other hand, the remaining missing values were in 'number of doors' variable. With a quick look at the 'body-type' of those observations, we were easily able to deduct and input the missing rows. Concerning the data type of the target variable 'symboling', we decided to convert it in a factor. That enables us to stop treating it as a numerical, aligning with the classification task intended with our work. Lastly, the only variable that we decided to eliminate was 'engine location', exhibiting almost unilateral distribution (99%). Keeping variables with such distribution would simply not add any predictive power to our model.

3. Exploratory Data Analysis

Our dive into the Automobile dataset shifted gears into a detailed Exploratory Data Analysis (EDA). This part of the journey aimed to uncover patterns and insights using a mix of statistics and visuals. We kicked off by checking how numerical features relate to each other through a correlation heatmap (Appendix 1). We highlighted correlation coefficients greater than 0.80 (absolute value), relevant threshold for determining collinearity between predictors. This helped spot interesting connections like how 'horsepower' and 'normalized losses' interact or how price is heavily influenced by features related to the size or the weight of any parts of a vehicle. Creating a clear and color-coded visual helped us decide which features should be additionally explored. It also gave a first idea of what features we could expect to have great predictive power, easing the task of future feature selection. Moreover, looking at histograms and boxplots of all numerical variables (Appendix 2) gave us a closer look at the distribution of values and pointed potential outliers, notably in variables such as ‘compression-ratio’, ‘price’ and ‘engine size’. These outliers, representing data points that stand out, could have a great impact on the accuracy and robustness of our classification model. Moving to categorical features, pie charts showed the breakdown of categories in 'engine-type', 'fuel-system' and ‘number of cylinders’, adding more layers to our understanding (Appendix 6 to 8). For other categorical variables, we used bar plots to navigate the diverse landscape with a focus on our target variable ‘symboling’. For instance, we noted that specific car companies like ‘Volvo’ or ‘Subaru’ will tend to have a much better level of ‘symboling’ compared to other sports car companies such as ‘Porsche’ and ‘Alfa-Romeo’ (Appendix 3). This paves the way for intriguing insights on how the price and style of a vehicle would greatly influence the degree to which the auto is riskier than its price indicates. On another hand, charts in Appendix 4 gives a brief overview of the distribution of ‘symboling’ in relation with categorical variables with smaller cardinality (binary or ternary) such as ‘drive wheels’, ‘number of doors’, ‘aspiration’ and ‘fuel-type’. Concerning the price feature, Appendix 5 leads us to think that the most expensive cars are the ones ending up with both extreme of the symboling values (either very risky or very safe compared to its price).

In summary, the 'make' variable stood out as a key player, guiding our analytical ensemble with clarity. Categorical features played distinct roles, aligning with clustering dynamics and adding specific dimensions to our overall understanding. The insights from EDA became crucial, shaping our analytical strategy. EDA wasn't just about exploration; it laid the foundation for our predictive

journey. The insights gathered acted as a guide for subsequent tasks, from choosing features to tuning parameters for the Gradient Boosting Model. The transition from exploration to model preparation incorporated every observation from EDA, strengthening the groundwork for the next steps in our analysis.

4. Classification - Model Building & Methodology

After weighing the pros and cons, we decided towards a Gradient Boosting Machine (GBM) Classification model. This decision rested on its ability to handle complex relationships within the data while maintaining a good balance between interpretability and accuracy. Building the GBM model, we embarked on tuning hyperparameters. First of, we quickly noticed that the size of our dataset imposed a certain limit on certain hyperparameters. For instance, once 'train.fraction' was set to a standard value of 0.67, the maximum value of 'n.minobsinnode' that our model could take was 20. In other words, 141 observations would be used for the training stage, making impossible to require more than 20 observations as a minimum number of observations per terminal nodes. We decided to maximize this hyperparameter considering the high number of predictors of our initial model (24), and that a higher value would result in a reduced complexity. The small size of our dataset also had an impact on our method of cross-validation (CV). In the same sense, we could not use k-fold CV, as it would lead to having some subset of the test data with zero instances of certain categorical level. Consequently, we set the 'bag.fraction' hyperparameter to a usual 0.50 as an alternative to evaluate our model's performance. Choosing this out-of-sample method of cross-validation enabled us to not have to set aside an independent part of our data for testing purposes, which reduces the information available for learning the model structure.

Correspondingly, the most important part of the tuning process was to find the right balance between the number of iteration trees computed ('n.trees') and the learning rate of the training process ('shrinkage'). Understanding the relationship between those two hyperparameters is crucial and often the hardest concept to grasp from the tuning process. Notably, opting for smaller shrinkage values typically leads to enhanced predictive performance (but with decreased marginal improvement). However, this performance improvement comes at the cost of increased computational demands in terms of both storage and CPU time. For instance, a model with 'shrinkage=0.001' generally outperforms 'shrinkage=0.01' but requires ten times as many iterations, escalating storage and computation time. The rule of thumb is to set shrinkage as small as feasible while ensuring a reasonable model fit, often targeting 3,000 to 10,000 iterations with

shrinkage rates between 0.01 and 0.001 (Source). In our case, we set the shrinkage to 0.001. To determine the number of trees, our out-of-bag (OOB) method of cross-validation explained earlier enabled us to use the 'perf.gbm' function to properly estimate the optimal value (232). Moreover, we were able to verify the legitimacy of this value by activating the 'verbose' hyperparameter of our model. Doing so, we could note that 232 trees is the point where the model plateaus its improvement. This is observable in Appendix 7, where the red line, representing the validation deviance, starts increasing. At this point, monitoring potential performance discrepancies between training and validation deviance is very beneficial for future feature selection steps. This is explained by the fact that those discrepancies are often good signs of model overfitting. In summary, this fine-tuning process set the stage for feature selection, where the main objective remains to find the delicate balance between model complexity and generalization. Lastly, an integral phase involved feature selection, where we pruned the dataset for relevance. Leveraging the feature importance attribute of our model (Appendix 8), we carefully selected predictors. The goal here was to strike a balance, shedding unessential predictors without compromising the model's performance significantly.

5. Classification Results, Conclusions & Limitations

Despite our efforts, the classification model yielded an accuracy of only 29%, highlighting the intricacies involved in predicting car risk using the current dataset. Several factors contributed to these suboptimal results. The limited size of our dataset posed a significant challenge, making it difficult for the model to discern nuanced patterns and establish robust correlations. The dataset's inherent characteristics, including imbalanced classes and a scarcity of meaningful features, further complicated the predictive task. Imbalanced classes can lead the model to favor the majority class, compromising its ability to effectively predict the minority class, in our case, high-risk cars. The lack of diverse features limits the model's ability to capture the complexity of factors influencing car risk.

To enhance our model, a primary strategy involves expanding the dataset. A more extensive and diverse dataset provides the model with a broader range of examples, potentially improving its ability to generalize to unseen data. This expansion should focus on addressing class imbalances and incorporating additional relevant features, such as vehicle specifications, historical maintenance data, and owner characteristics. Considering alternative algorithms, such as Random

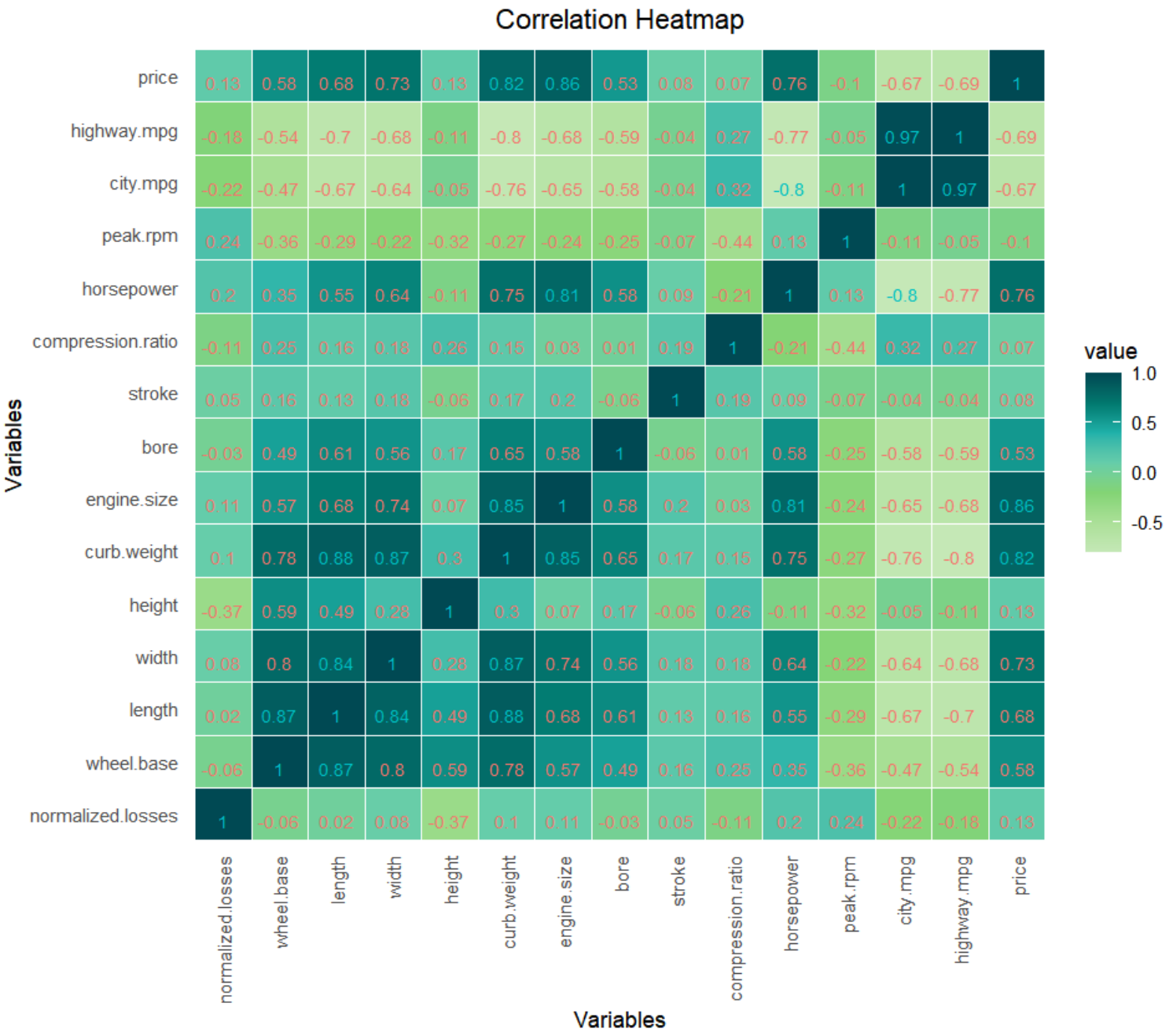
Forest, Classification Tree, and Logistic Regression, is crucial. A comparative analysis can reveal the strengths and weaknesses of each algorithm in the context of our specific prediction task.

Moreover, the dataset's modest size accentuates the need for nuanced analytical approaches. In tandem with expanding the dataset, incorporating a Clustering model adds depth to our analysis. Clustering, specifically K-Means, offers a holistic perspective by grouping similar instances together based on inherent patterns, unraveling latent structures within the data. By applying Principal Component Analysis (PCA) to reduce dimensionality and facilitate meaningful clustering, we gain insights into potential clusters of cars sharing common risk attributes. Exploring various algorithms, such as K-Prototypes and DBSCAN, ensures a comprehensive understanding of the dataset's structure. The Elbow and Silhouette methods guide us in determining the optimal number of clusters, aiding in the identification of distinct risk profiles. The clustering model supplements our classification efforts, shedding light on subtle relationships and interactions among features. For instance, it may reveal clusters of cars with similar risk profiles based on factors beyond those directly related to car specifications, offering a holistic view of risk determinants. While the clustering model may not directly predict risk levels, it complements the classification model by identifying nuanced associations and unveiling potential influencing factors. The combined insights from both models contribute to a comprehensive understanding of car risk, enriching the decision-making process for various stakeholders. As we move forward, refining the clustering model involves continuous iteration and validation. Exploring alternative clustering methods and incorporating additional features into the analysis can further enhance the model's precision. The ultimate aim is to create a robust framework that not only predicts risk categorization but also provides nuanced insights into the multifaceted nature of car risk, empowering users with comprehensive and actionable information.

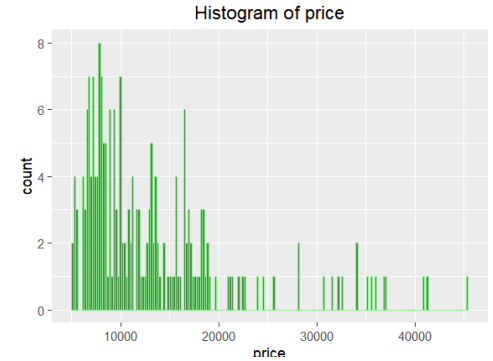
Despite the current limitations, our model holds potential business utility. Insurers could utilize it to refine risk assessment strategies, car dealerships to optimize inventory management, and individual buyers to make more informed purchasing decisions. The recommendation is to iteratively refine and expand the model, incorporating feedback from users and adapting to evolving trends in the automotive industry. In conclusion, while our initial classification model faced challenges, the outlined strategies pave the way for improvement. By addressing dataset limitations, exploring alternative algorithms, and considering business applications, we aim to develop a more robust model with broader implications in the domain of car risk assessment.

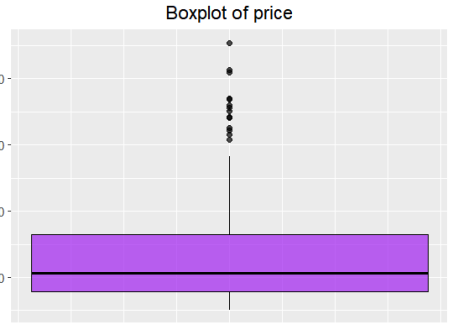
6. Appendices

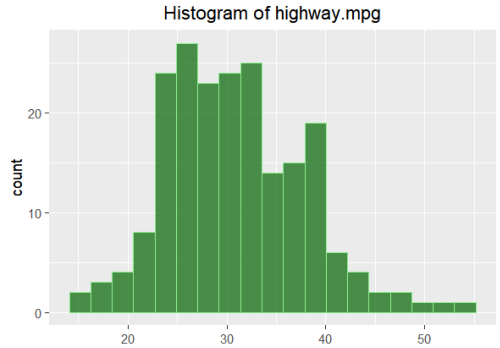
1. Correlation heatmap of all numerical variables

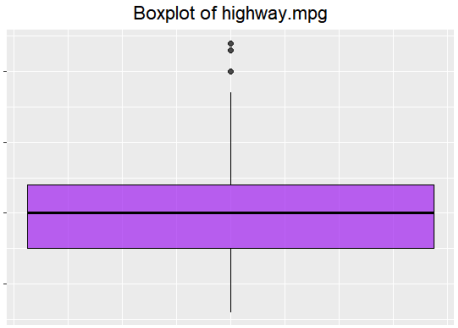


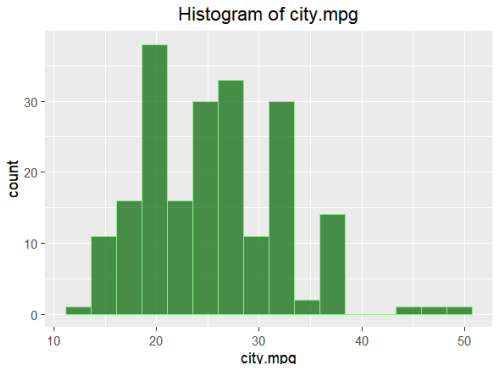
2. Numerical Variables Distributions

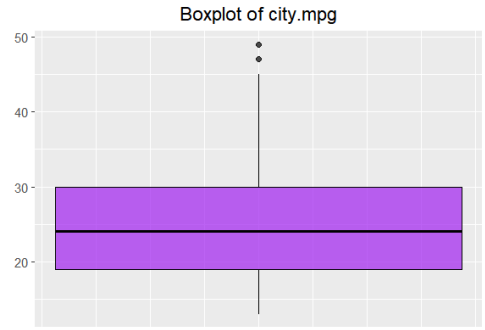


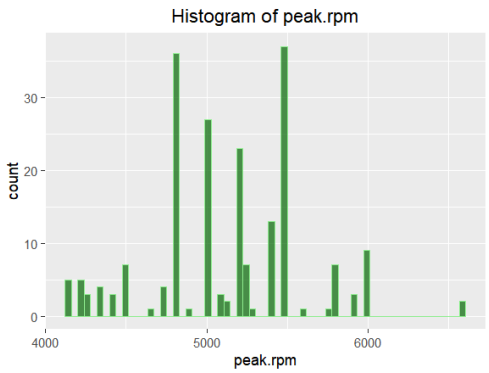


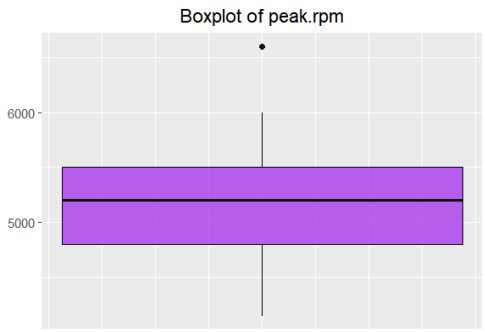


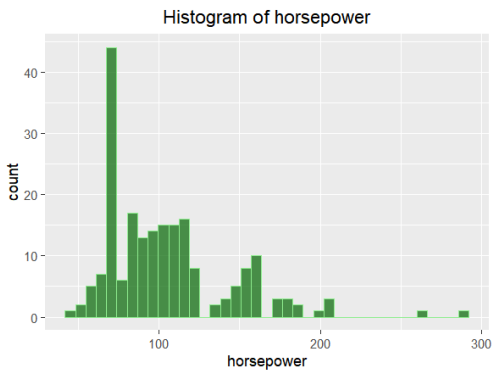


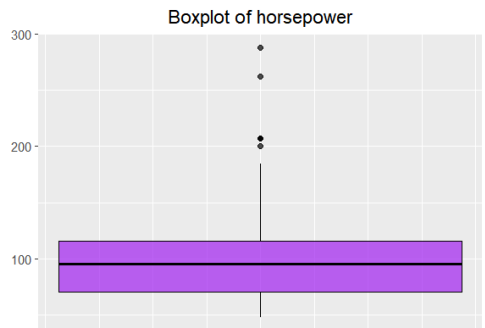


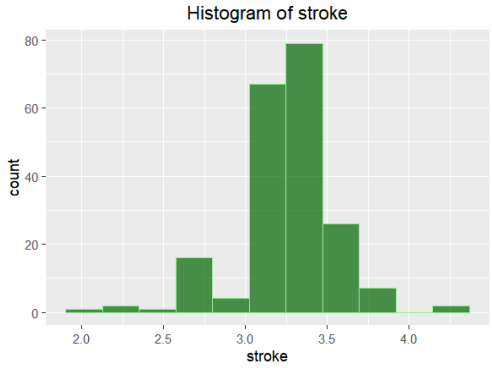


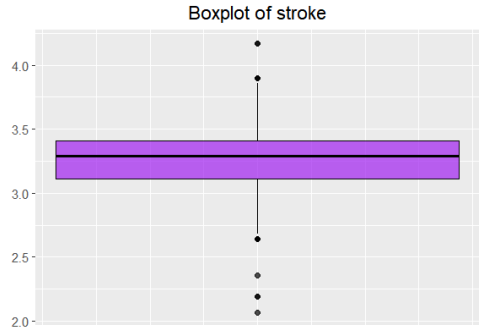


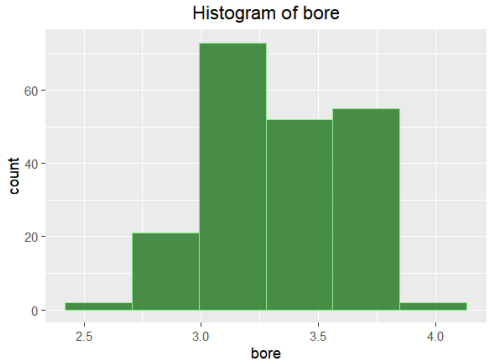


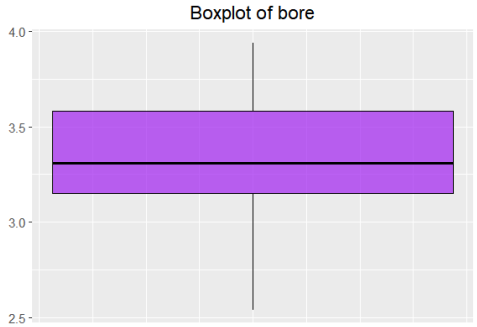


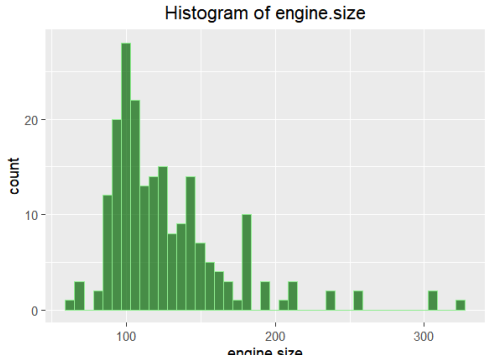


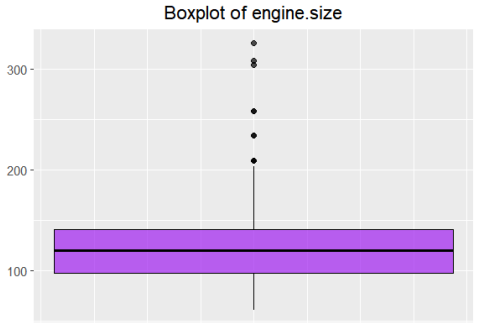

Histogram of stroke
count
stroke
0
20
40
60
80
2.0
2.5
3.0
3.5
4.0

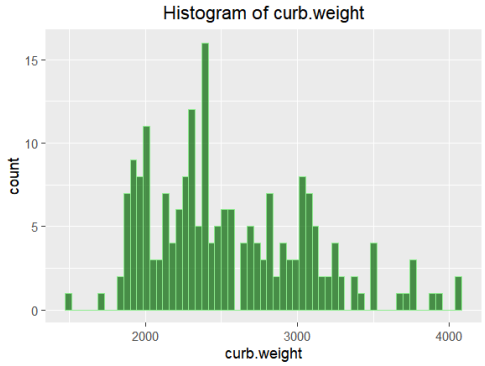

Boxplot of stroke
2.0
2.5
3.0
3.5
4.0

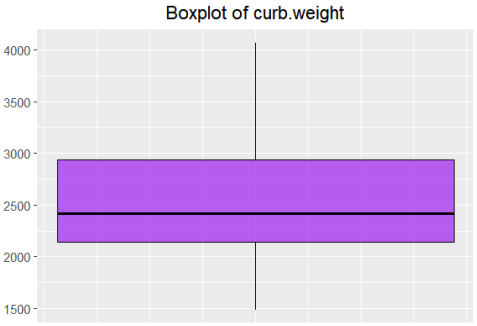

Histogram of bore
count
bore
0
20
40
60
2.5
3.0
3.5
4.0

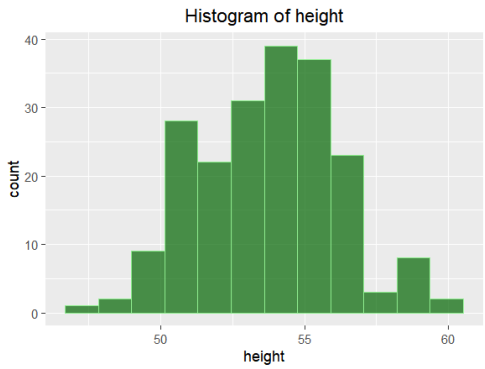

Boxplot of bore
2.5
3.0
3.5
4.0

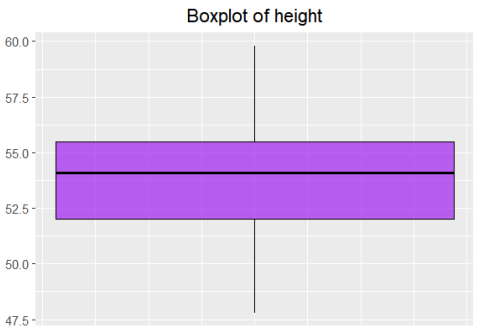

Histogram of engine.size
count
engine.size
0
10
20
100
200
300

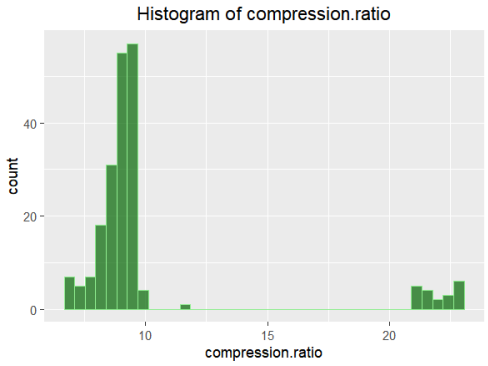

Boxplot of engine.size
100
200
300

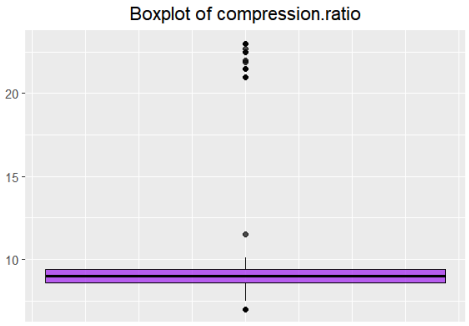

Histogram of curb.weight
count
curb.weight
0
5
10
15
2000
3000
4000

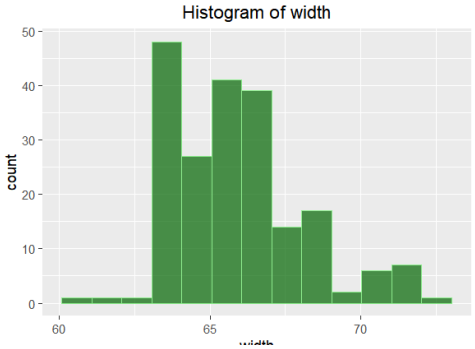

Boxplot of curb.weight
1500
2000
2500
3000
3500
4000

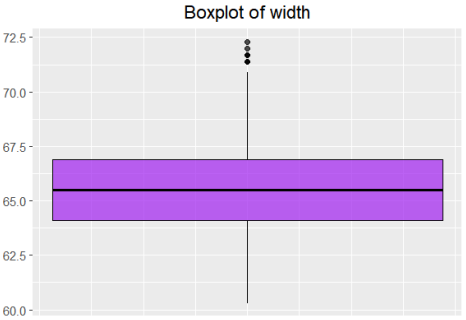

Histogram of height
count
height
0
10
20
30
40
50
55
60

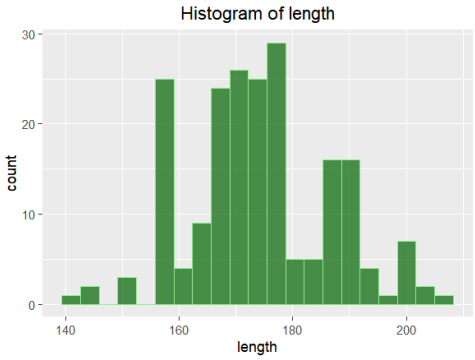

Boxplot of height
47.5
50.0
52.5
55.0
57.5
60.0

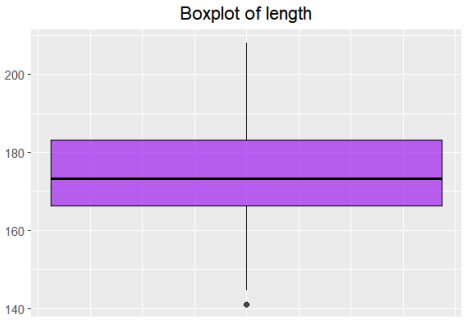

Histogram of compression.ratio
count
compression.ratio
0
20
40
10
15
20

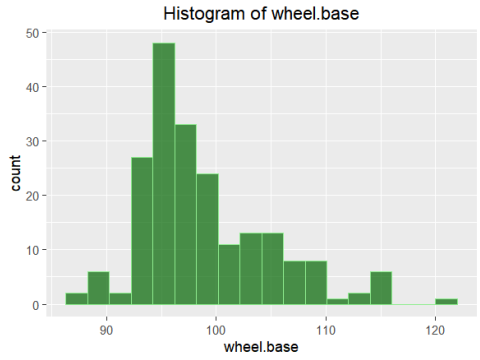

Boxplot of compression.ratio
10
15
20

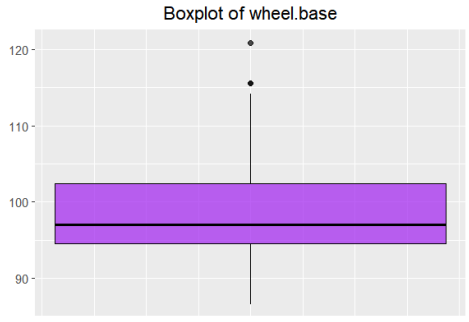

Histogram of width
count
0
10
20
30
40
50
60
65
70

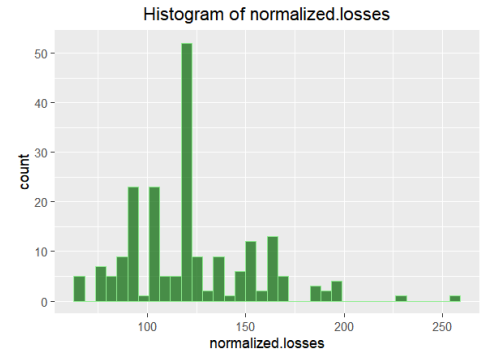

Boxplot of width
60.0
62.5
65.0
67.5
70.0
72.5

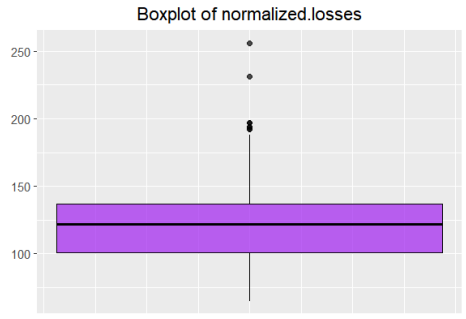

Histogram of length
count
length
0
10
20
30
140
160
180
200


Boxplot of length
140
160
180
200

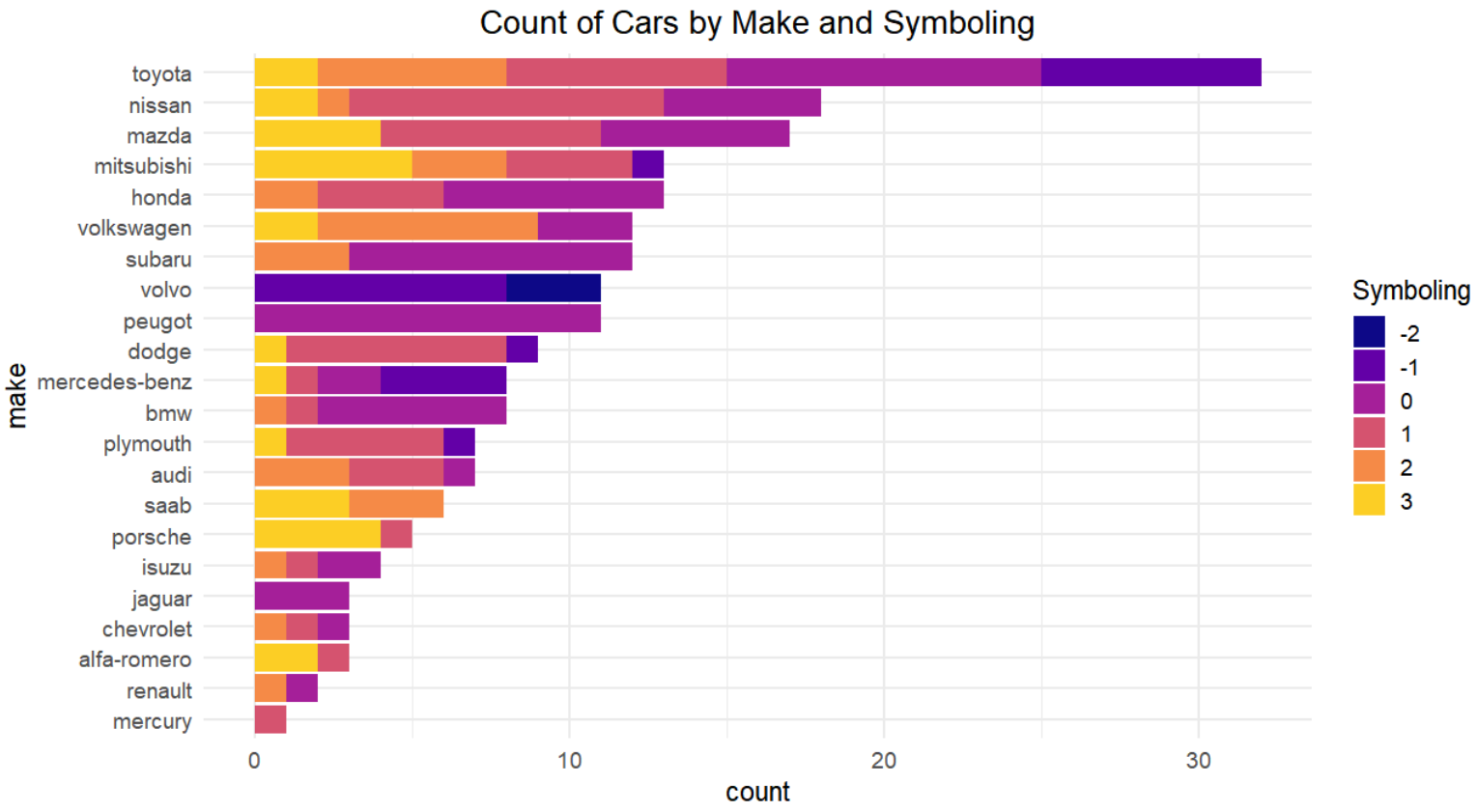

Histogram of wheel.base
count
wheel.base
0
10
20
30
40
50
90
100
110
120


Boxplot of wheel.base
90
100
110
120

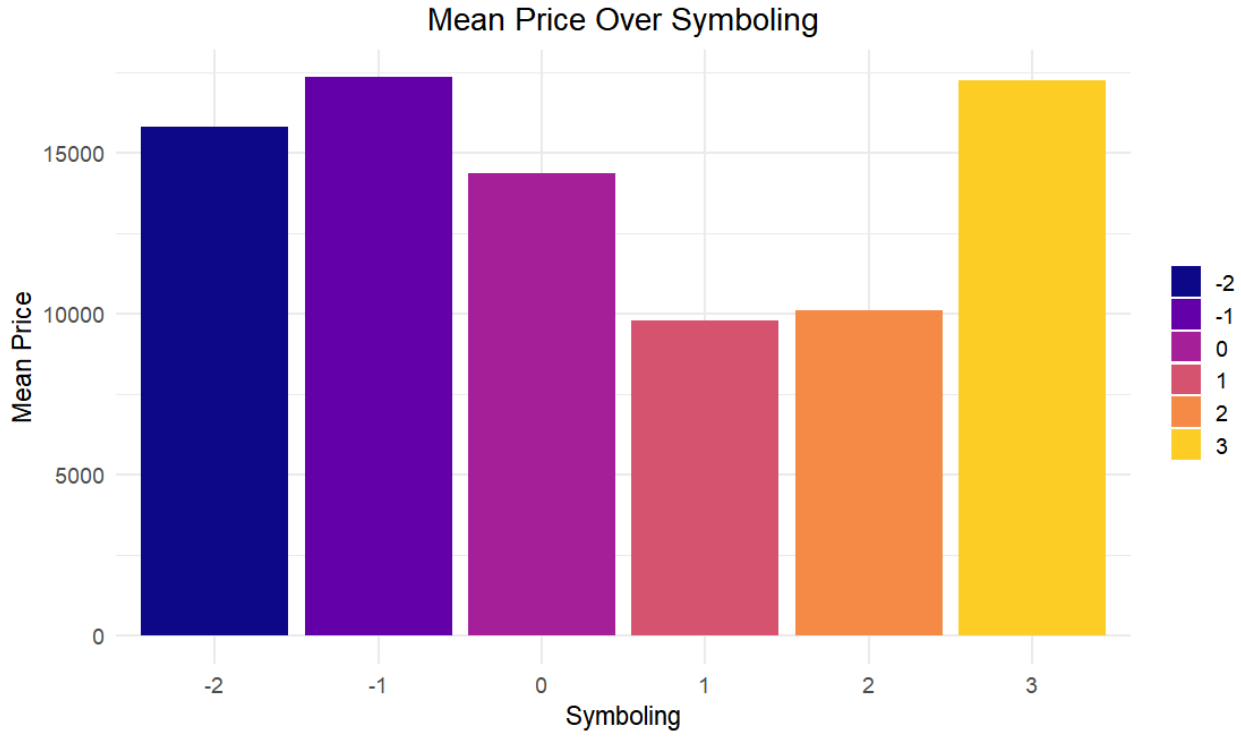

Histogram of normalized.losses
count
normalized.losses
0
10
20
30
40
50
100
150
200
250


Boxplot of normalized.losses
100
150
200
250

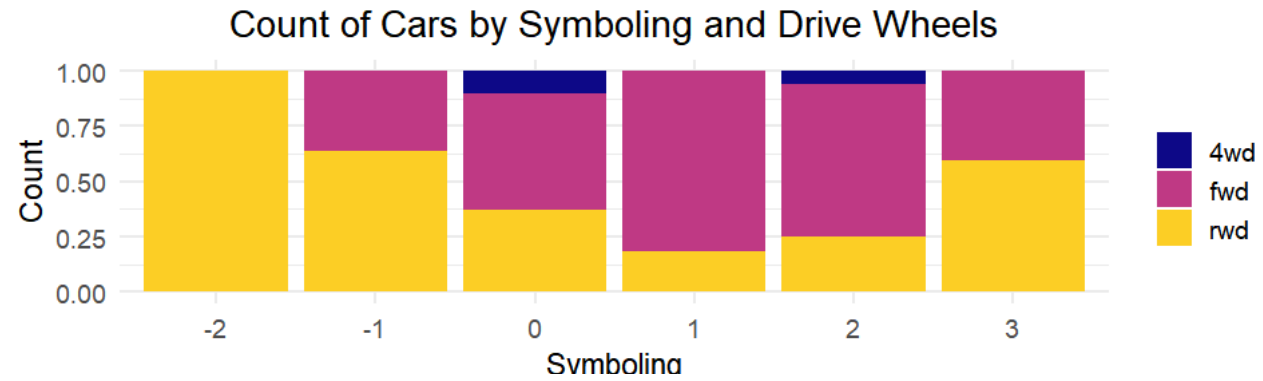
3. Distribution of Car Make by Symboling

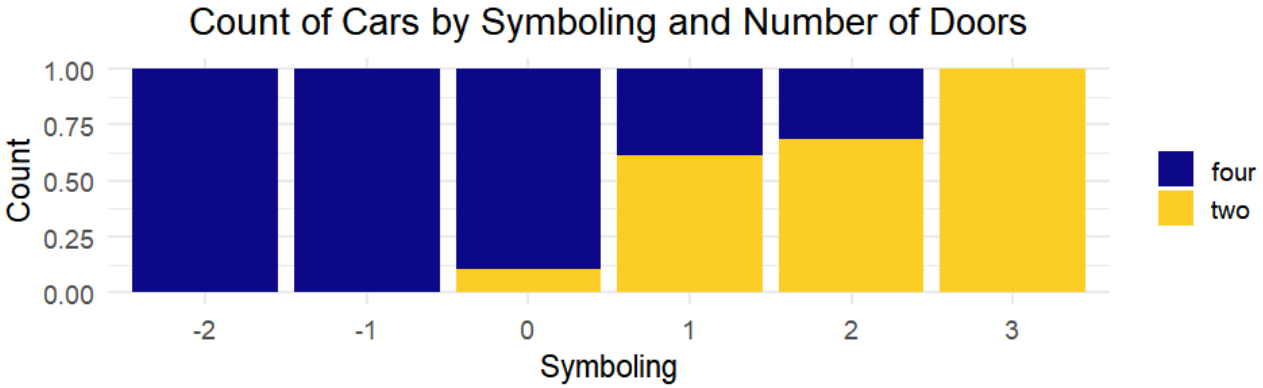


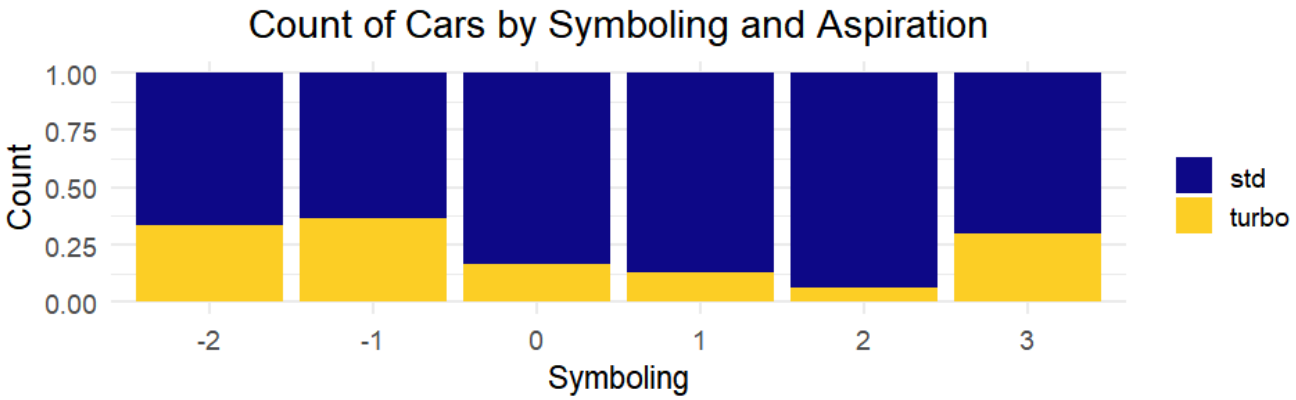
4. Distribution of Price by Symboling

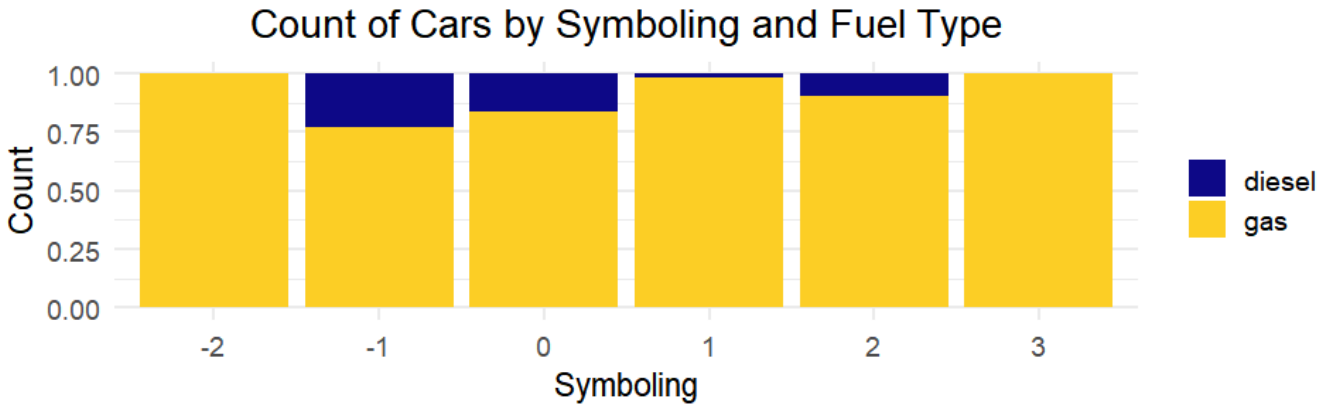


5. Distribution of Drive Wheels, Number of Doors, Aspiration & Fuel Type by Symboling

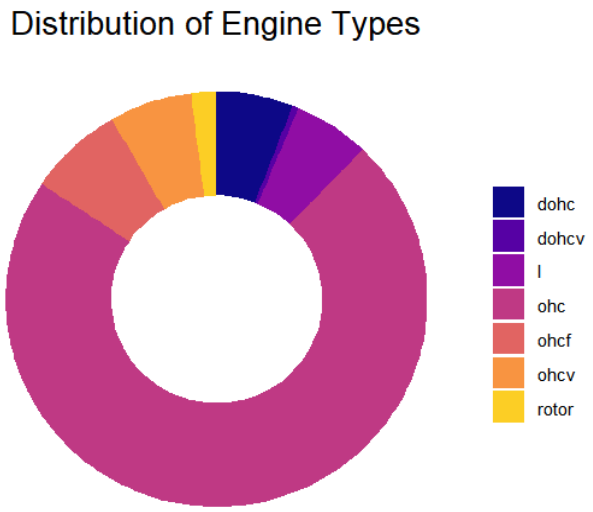


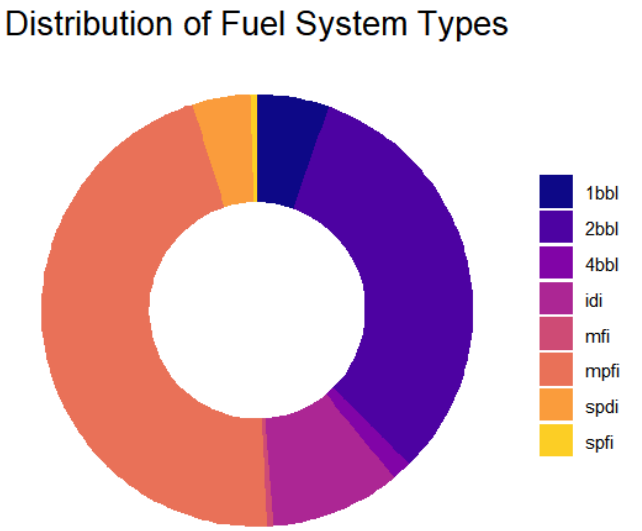


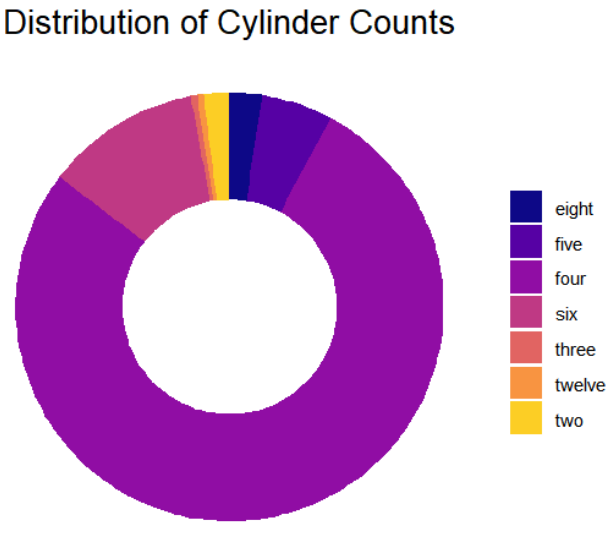




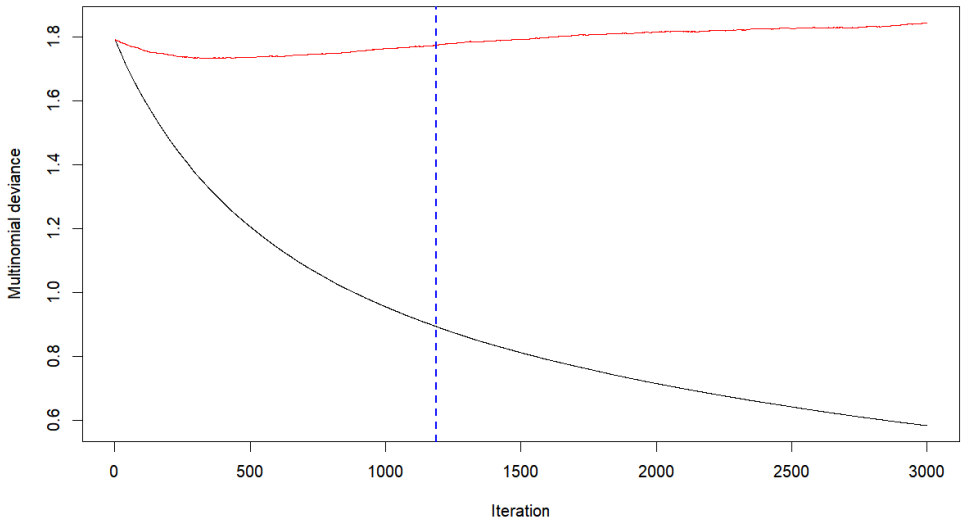
6. Distribution of Engine and Fuel System Types & Cylinder Counts







7. Evolution of GBM Performance over Trees Iterations



8. Initial Gradient Boosting Machine Feature Importance