MGSC 661 – Final Project

Tashfeen Ahmed – 261145667

Contents

1. Introduction 3
2. Data Description 3
3. Model Selection & Methodology 4
4. Clustering Model 6
5. Classification/Predictions and Conclusions 7
6. Appendix 9

1. Introduction

In the rapidly evolving field of data science, the ability to distill vast amounts of data into actionable insights is invaluable. This project, centered on a comprehensive analysis of automobile data, embodies this principle. It aims to delve deep into the intricacies of automotive data, utilizing a blend of predictive modeling and clustering techniques to extract meaningful insights.

The core of this project revolves around two primary analytical methods: Classification Model and clustering. In predictive modeling, I employed techniques like Random Forest and Decision Trees. These models predict the safety rating (symboling) of a car. This rating is an assessment of the likelihood that an insurer will have to pay out a claim and the potential cost of that claim.

Each car is given a risk factor symbol associated with its price. This symbol is an integer value that can range typically from -3 to +3. The value reflects the car's riskiness in terms of insurance:

- +3: Indicates a car is considered risky to insure.
- 0: Represents an average risk.
- -3: Suggests the car is deemed safer than average

The predictive modeling phase not only aims to predict but also to understand the factors that most significantly influence model metrics.

In clustering, the project uncovers natural groupings within the automobile data based on similarities across various features. I used the k-means clustering algorithm the key steps in this phase included determining the optimal number of clusters using the Elbow and Silhouette methods and analyzing the resulting clusters to gain insights about different types of automobiles in the dataset.

2. Data Description

The dataset I've used is sourced from Kaggle's called "Automobile Dataset". It contains a broad spectrum of variables, ranging from basic car attributes like make and body style to more technical aspects including engine size, horsepower, and price. The variables are a mix of categorical and numerical types, presenting an opportunity to explore diverse analytical techniques.

The automobile dataset presents a diverse array of variables, and the distribution and inter-variable relationships are depicted in a comprehensive pair plot (Appendix: Figure 2.1). Observations reveal a direct relationship between price and attributes like 'engine.size' and 'horsepower', highlighting a trend where increased performance attributes potentially command higher market prices. Furthermore, 'length' and 'width' appear to trend upwards with price, suggesting a consumer inclination towards more spacious vehicles being associated with a higher value. The plots also suggest a less apparent, yet notable, negative trend between 'city.mpg' and price, indicating that higher fuel economy does not necessarily correlate with a

higher price point, a counter-intuitive insight that may be of interest to business leaders strategizing in the competitive automobile market. The plot also shows variable distributions. For instance, the 'price' variable displays a right-skewed distribution, indicating that while most vehicles are clustered at a lower price range, there is a long tail of higher-priced automobiles, potentially luxury or performance models. 'Engine size' and 'horsepower' exhibit similar right skewness, hinting at a concentration of standard engine sizes with fewer high-performance outliers. In contrast, 'city.mpg' reveals a somewhat normal distribution, albeit with a slight skew towards higher efficiency, reflecting a market with a balanced mix of fuel-efficient and less economical vehicles. 'Width', 'Length' and height show a more symmetric distribution, suggesting that vehicle dimensions across the dataset do not vary as widely and are centered around a common design standard.

A correlation matrix was also built to remove highly correlated variables (Appendix: Figure 2.2). The correlation analysis guided the decision to streamline the dataset by excluding 'curb.weight', 'highway.mpg', and 'engine.size' due to their high correlation with other variables. The matrix revealed strong correlations such as 'curb.weight' and 'engine.size' (Corr: 0.8502), as well as 'city.mpg' and 'highway.mpg' (Corr: 0.971) and 'engine size' displayed a significant positive correlation with 'price' (Corr: 0.8888. While 'length' and 'width' and 'price' also showed some correlations, they were retained for their substantial business value, as they provide critical insights into vehicle design preferences and market trends.

Moreover, PCA biplot was also performed (Appendix: Figure 2.3). It captures both the magnitude and direction of variable correlations. The biplot also allows for the examination of categorical variables in conjunction with continuous ones, providing a nuanced perspective on how attributes like 'fuel.type' and 'body.style' might relate to other vehicle features. Although the PCA indicated no extreme correlations warranting variable removal, this visualization further aids in appreciating the complex interplay of features. Notably, vectors such as 'width', 'length', and 'wheel.base' point in a similar direction, indicating a high degree of correlation, consistent with the understanding that larger vehicles typically have longer wheelbases. The orthogonal positioning of 'city.mpg' relative to 'horsepower' underscores an inverse relationship, where fuel efficiency decreases as power increases. Moreover, as length, width and height of automobile increases city.mpg decreases (fuel efficiency decrease)

3. Model Selection & Methodology

3.1 Classification Models

The classification model is used to predict the insurance risk rating (symbolling). The 'symboling' attribute, originally an integer scale (range -3 to 3) representing the risk factor associated with a car's price, has been binary-encoded to simplify the model's interpretative capacity and because we have only 195 observations which would not be enough for 7 categories. This transformation is essential for ensuring that the model can generalize well, despite the limited sample size, and accurately predict whether a car is deemed riskier or safer. Cars with a 'symboling' above 0,

which are considered riskier than average, are now denoted by **1 (Risky)** , while those equal to or below 0, indicating average or less risk, are marked as **0 (Safe).** This binary transformation condenses the variable's complexity, allowing the model to focus on the dichotomy of risk, which is a critical consideration for insurers and customers alike.

Data Preprocessing

Initial exploration revealed missing values, notably within the 'normalized-losses' and 'num-of-doors' columns. 'normalized-losses' had to be dropped because of too many missing values (20%) while for the rest missing values rows were removed. Numeric and categorical columns that were necessary for prediction were selected and then highly correlated variables were removed using correlation matrix, PCA. List of columns used can be found in Appendix: Table 3.1. Considering the relatively small size of dataset outliers were not removed from numerical variables and using scatter plot within pair plots no extreme outliers were seen so there was no need for it. The ratio of the two classes (0 and 1) is close (approximately 1:1.19), which is not a severe imbalance. Often, severe imbalances are more like 1:10 or more extreme. So, this was deemed acceptable. However, to correct this, one could over-sample the minority classes by using Synthetic Minority Oversampling Technique SMOTE.

Modelling

For the classification model I assessed the performance of two tree-based model – Decision tree, Random Forest – to see which gave the best results I compared the accuracy metrics and recommended the best model as Random Forest

Random Forest

Random forests are intrinsically better than decision trees at dealing with overfitting as it uses bagging - randomly samples data and trains each tree on a random subset of data using a subset of the predictors. This allows us to estimate relative feature importance for each of our variables by comparing metrics like mean decrease in accuracy and Gini index for each set of predictors. The Random Forest algorithm was selected for its efficacy in handling the dataset's high-dimensional feature space, and its robustness against overfitting. An initial model was constructed to gauge the importance of each feature (Appendix: Figure 3.1), resulting in the removal of 'engine.location' due to its minimal impact on model accuracy, as indicated by its low mean decrease in accuracy and Gini importance. The Mean Decrease in Accuracy metric indicates how much each variable contributes to the model's predictive power; a higher value signifies greater importance. The Mean Decrease Gini, on the other hand, measures each variable's contribution to the homogeneity of the nodes and leaves in the model; a higher value indicates that the variable is better at splitting the data into pure subsets.

Subsequently, a k-fold cross-validation approach was employed to validate the model's performance, which is a robust method to assess a model's generalizability. In this context, accuracy refers to the proportion of correct predictions made by the model, while error denotes

the proportion of incorrect predictions. The Out-of-Bag (OOB) error estimate, an internal validation method unique to Random Forest, was calculated at 6.7%, providing an unbiased error rate as the model was not exposed to this data during training. The detailed metrics of the model's performance are systematically cataloged in Table 3.2 within the Appendix. Through this process, the optimal number of trees was determined to be 200 (Appendix: Figure 3.2), balancing computational efficiency with predictive performance.

Decision Tree

The decision tree methodology stands out for its interpretability, clearly delineating how decisions are made. The initial model used a complexity parameter (cp) of 0.01 to guard against overfitting, allowing for sufficient tree growth to capture the data's patterns without becoming overly complex. (Appendix: Figure 3.5). From the tree the most important criterion for num.of.doors is 4 and 65% observations fall down in left node. Most important criterion for wheel.base is 65 for bore is 3.2 and width is 65. In the left bottom leaf node 36% observations fall and most likely classification is 0 and 9% likelihood of being classified as 1 in this node

To find the optimal cp value a overfitted tree was trained and the out of sample performance was studied. (Appendix: Figure 3.3 and Figure 3.4). The best tree is also printed (Appendix: Figure 3.6) to understand the cuts as described for previous tree.

Subsequently, a k-fold cross-validation approach was employed to validate the model's performance.

Classification Results

Both models were tested using 5-fold cross validation. Model trains on 4 training folds and test on untrained test fold, this is repeated until all unseen fold are tested. Based on the results from Appendix: Table 3.2 - The Random Forest model, with its ensemble approach, demonstrated superior performance with an accuracy of 94.3% and an error rate of 5.6%, outshining the Decision Tree model in handling the complexity and variance within the data. Despite the Decision Tree's clear interpretability, it was Random Forest's robust predictive power and lower error rate that cemented its place as the preferred model for classifying the insurance risk rating of vehicles. This preference was further supported by the Random Forest's Out-of-Bag (OOB) error estimate, which reinforced the model's generalizability to unseen data. However, Decision trees are more interpretable (the decision split could be understood by humans) while the other model is a black box

4. Clustering Model

In the clustering phase of analysis, the focus shifted to discerning natural groupings within the automobile dataset based on shared characteristics to get the most insights. Implementing the k-means algorithm, I normalized all numerical variables to ensure uniformity in scale (0-1) range, missing values were also removed and numeric columns were selected. For selecting the optimum number of cluster silhouette and elbow method were used. The Silhouette method,

which measures how similar an object is within its cluster compared to other clusters, corroborated this finding by suggesting that three clusters provide clear and distinct groupings (Appendix: Figure 4.1). The Elbow method, which assesses the within-cluster sum of squares, suggested 3 clusters as the point where adding more clusters does not offer a significant improvement in variance explained (Appendix: Figure 4.2). The Elbow method's suggestion was favored as it offered a more significant drop in WSS at three clusters, signifying a substantial improvement in variance explained within clusters up to that point. In this case, the Elbow method presented a clear inflection at three clusters, suggesting that this was the most appropriate number to capture the dataset's inherent structure without overcomplicating the model. This point represents a balance between minimizing WSS and avoiding an excessive number of clusters, which could dilute the interpretability of the results.

4.1 Clustering Results

To see the results of the K-means clustering I plotted the cluster centroids for each variable in the cluster (Appendix: Figure 4.3). Here is a summary of each cluster:

Cluster 1 (Red): These vehicles are identified as the riskiest in terms of insurance. Despite their moderate price, they offer a balance in terms of size and engine power. They are not the most economical in terms of city mileage, which may suggest a trade-off between performance and fuel efficiency. This cluster may appeal to individuals who enjoy driving and are less sensitive to fuel costs and insurance premiums. They could be seen as performance-oriented cars that balance size and power without being the most expensive in the market.

Cluster 3 (Blue): These are the safest in terms of insurance risk, these vehicles are likely to attract lower insurance premiums, a key selling point for safety-conscious consumers. Their larger dimensions and engine power must be designed for comfort and safety. Despite their highest horsepower and engine size, they have lower risk symboling. These automobiles have the lowest city mileage and peak rpm which might be because of their big size. This cluster would be particularly appealing to those who value space, comfort, and a sense of security without compromising vehicle performance.

Cluster 2 (Green): Vehicles in this cluster, being moderate in risk, offer a middle ground between performance and safety. They are more economical (lowest price and highest city mileage), with lower horsepower and smaller engine sizes, which might indicate these vehicles are targeted towards individuals who need a reliable mode of transport without the heightened costs associated with high-performance cars. They have the smallest size (length, width, curb weight and height)

5. Classification/Predictions and Conclusions

In the realm of automobile insurance, understanding the risk is fundamental. As I delved into the predictive modeling of car insurance risk ratings, the insights gleaned from the Random Forest model, which exhibited a remarkable 94.3% accuracy, proved invaluable.

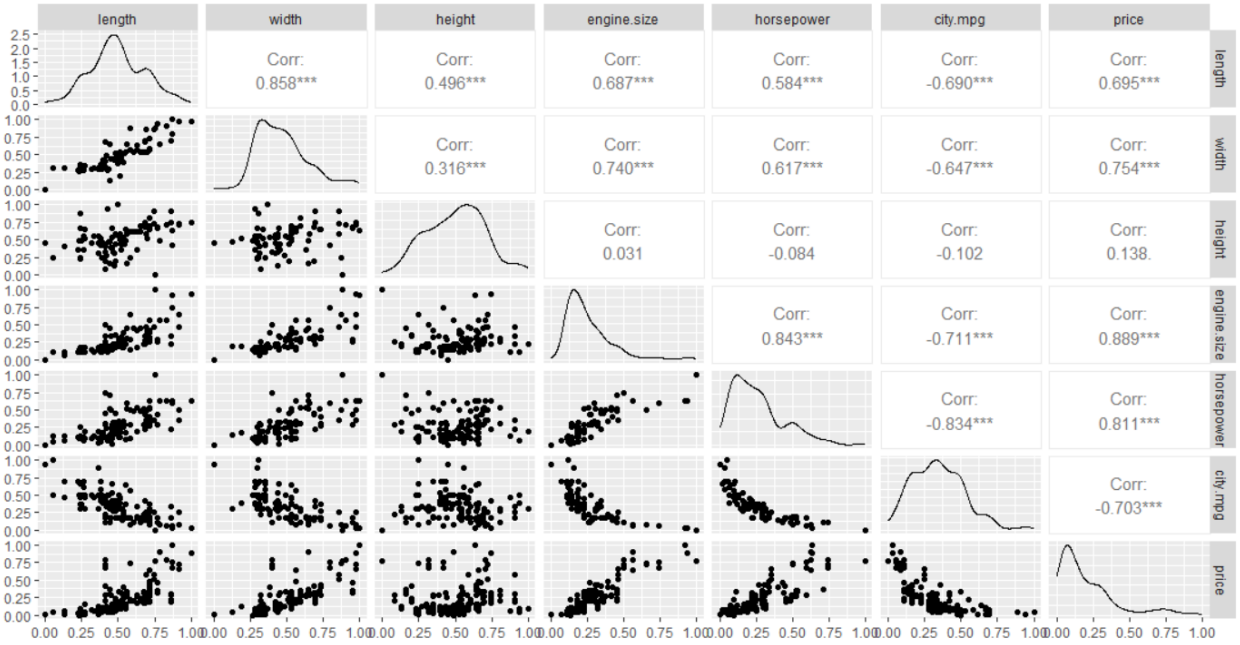
The classification model's ability to predict risk ratings with high accuracy is a beneficial for insurance managers. This predictive prowess facilitates better risk assessments, paving the way for more accurate premium calculations. The Random Forest model's feature importance graph highlights key attributes affecting vehicle insurance risk ratings. The 'number of doors' and 'wheelbase' emerge as top influencers. Knowing that certain attributes like the number of doors and wheelbase are top influencers in risk assessment, managers can price insurance premiums more accurately by building better models classification models. The PCA biplot also offers managers strategic insights. For example, horsepower and mileage are in the opposite directions showing more horsepower means less mileage and from clustering we know that lowest mileage automobiles with high horsepower are the safest.

The clustering model segments vehicles into distinct groups based on their characteristics. It is seen that the riskiest automobiles are those that have moderate dimensions, engine size, mileage, price, and horsepower. Insurers should design higher premiums and strict policies for such automobiles. Moreover, the safest cars are those with largest dimensions, engine size, bore, stroke, horsepower, highest price, and lowest mileage. Insurers can design lower premiums as there is less risk of claim. Looking at how dimensions, engine size, mileage, price, horsepower etc change in different clusters and how they affect risk insurance managers can better understand which features are related to higher risk. For example, an automobile with large dimensions, horsepower, higher price and low mileage would definitely be safer than other automobiles and a lower premium can be calculated for it.

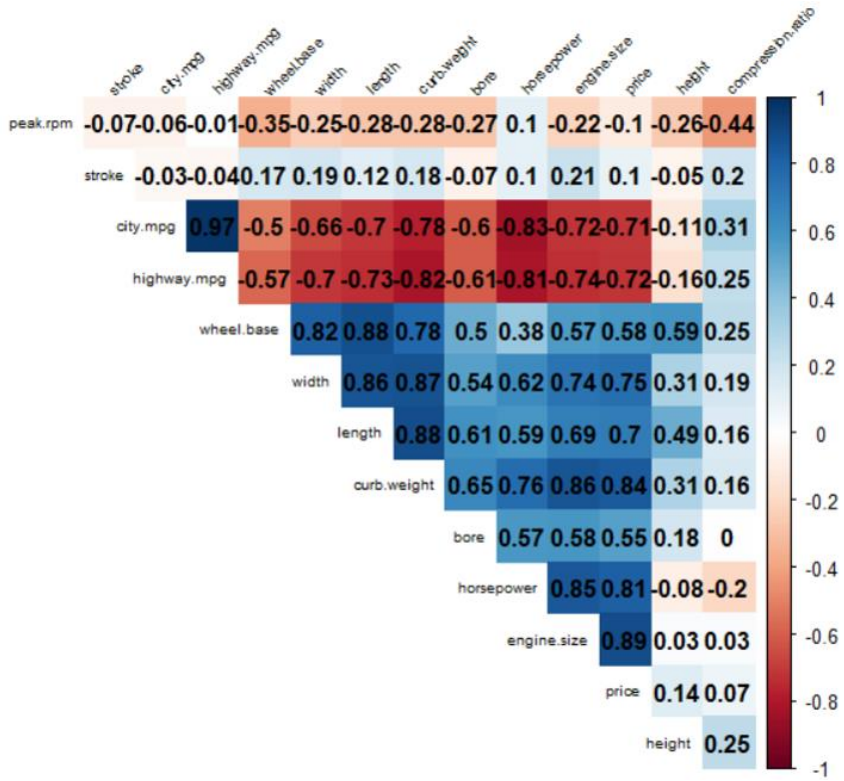
The fusion of predictive modeling and clustering provides a multidimensional view of the automotive market. For insurance managers, the practical application of these insights lies in their ability to inform decision-making, leading to enhanced risk management, customer satisfaction, and business growth. By combining the results from two techniques managers can classify risk and also understand what makes it risky!

6. Appendix

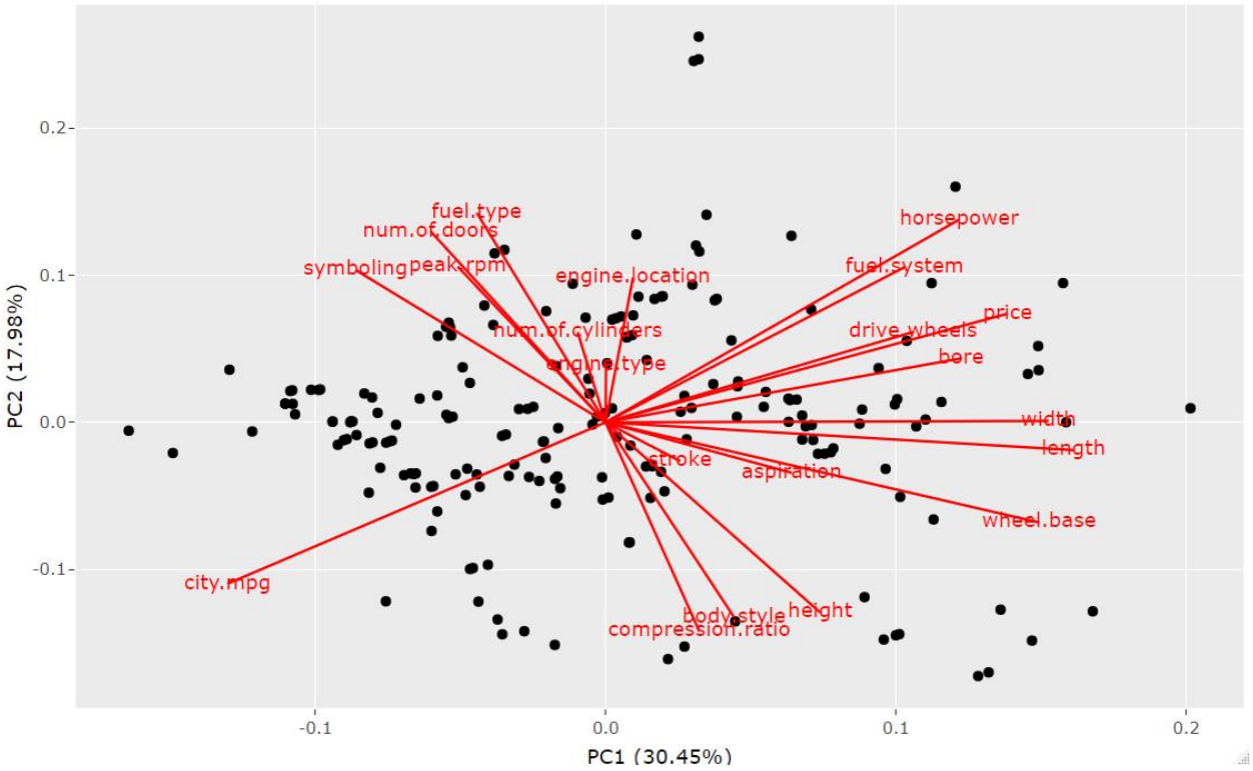
- **Figure 2.1**



- **Figure 2.2**



- **Figure 2.3**



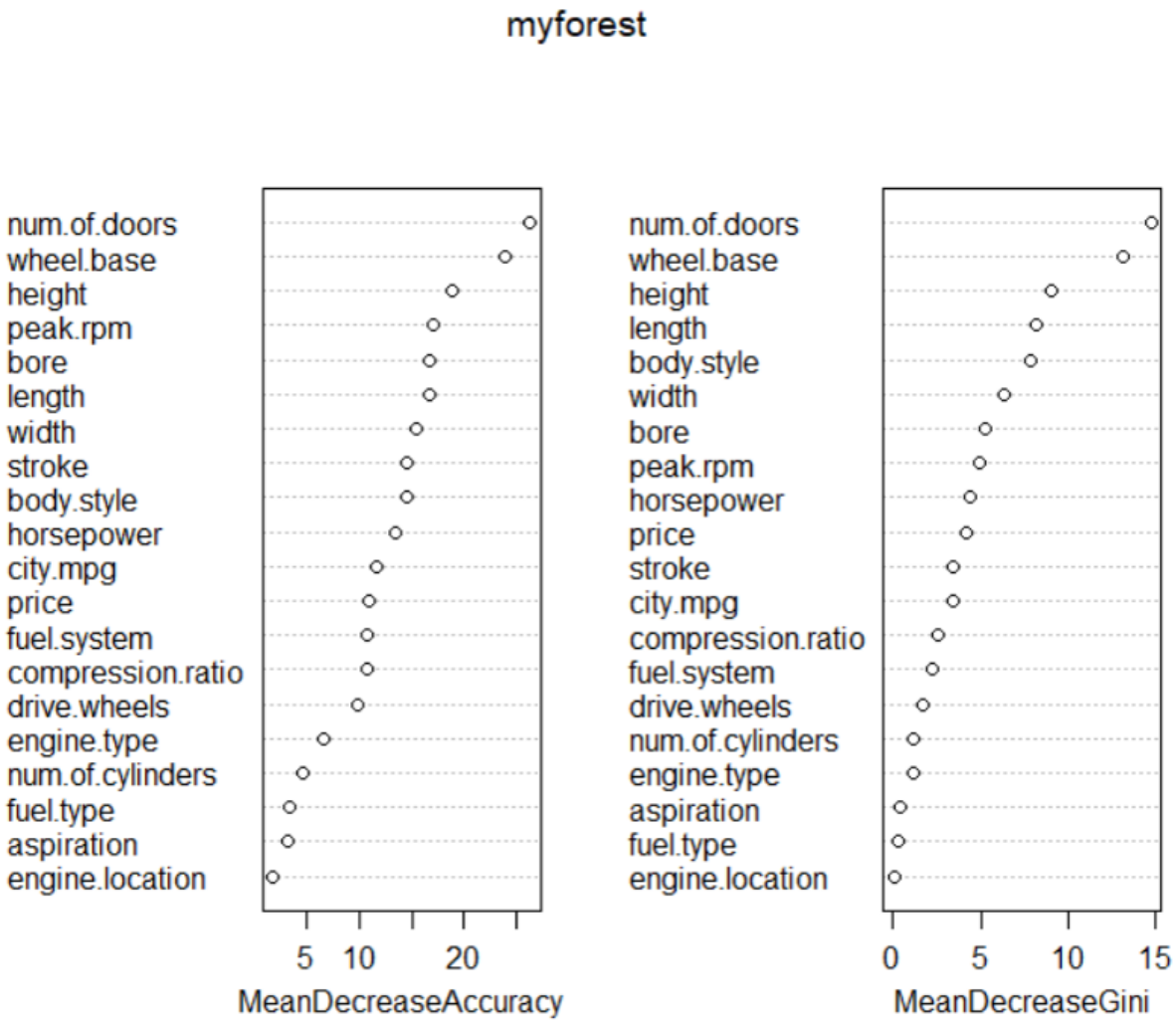
- **Table 3.1**

Predictor	**Type**
fuel. type	*Categorical*
aspiration	*Categorical*
num. of. doors	*Categorical*
body. style	*Categorical*
drive. wheels	*Categorical*
engine.type	*Categorical*
num.of.cylinders	*Categorical*
wheel.base	Numerical
length	Numerical
width	Numerical
height	Numerical
bore	Numerical
stroke	Numerical
compression.ratio	Numerical
Horsepower	Numerical
peak.rpm	Numerical
city.mpg	Numerical
price	Numerical

- **Table 3.2**

Model	Accuracy	Error Rate
Random Forest	94.3%	5.6%
Decision Tree	84.4%	15.5%

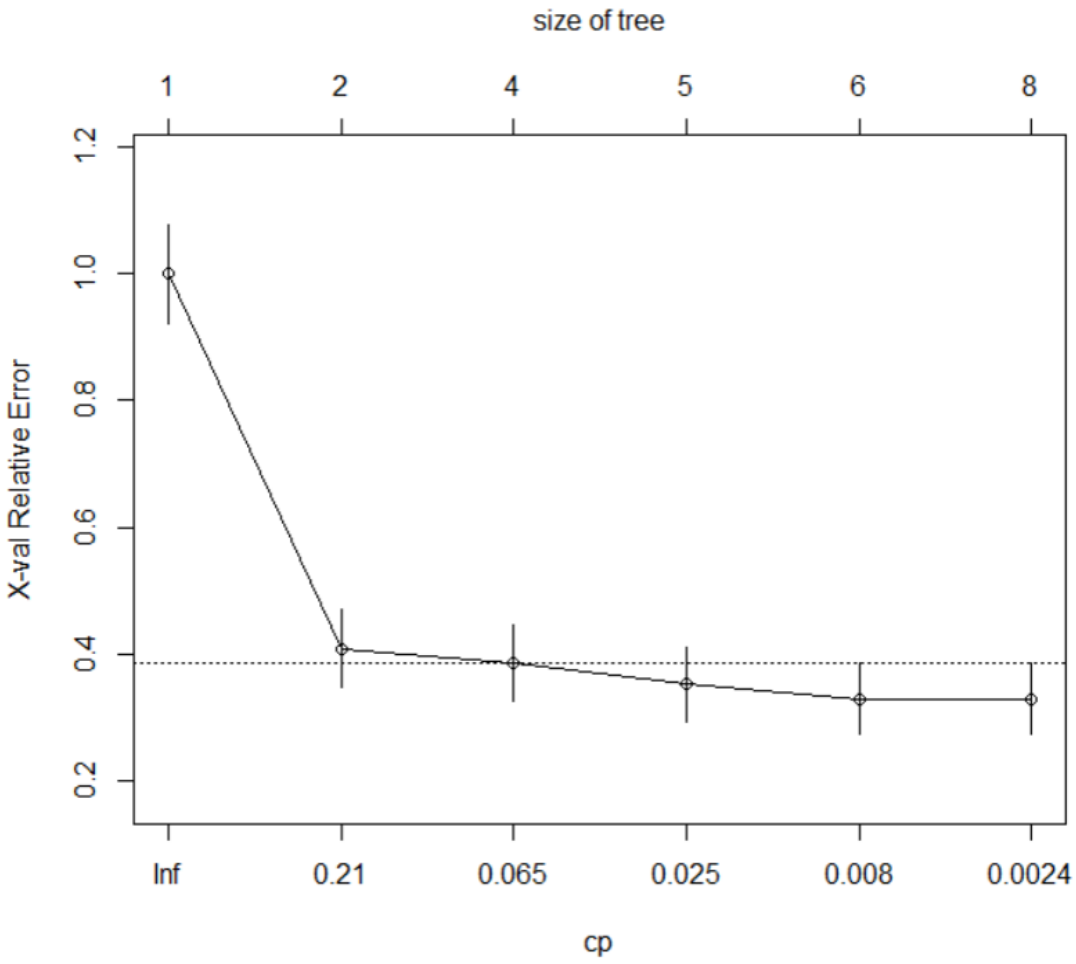
- **Figure 3.1**



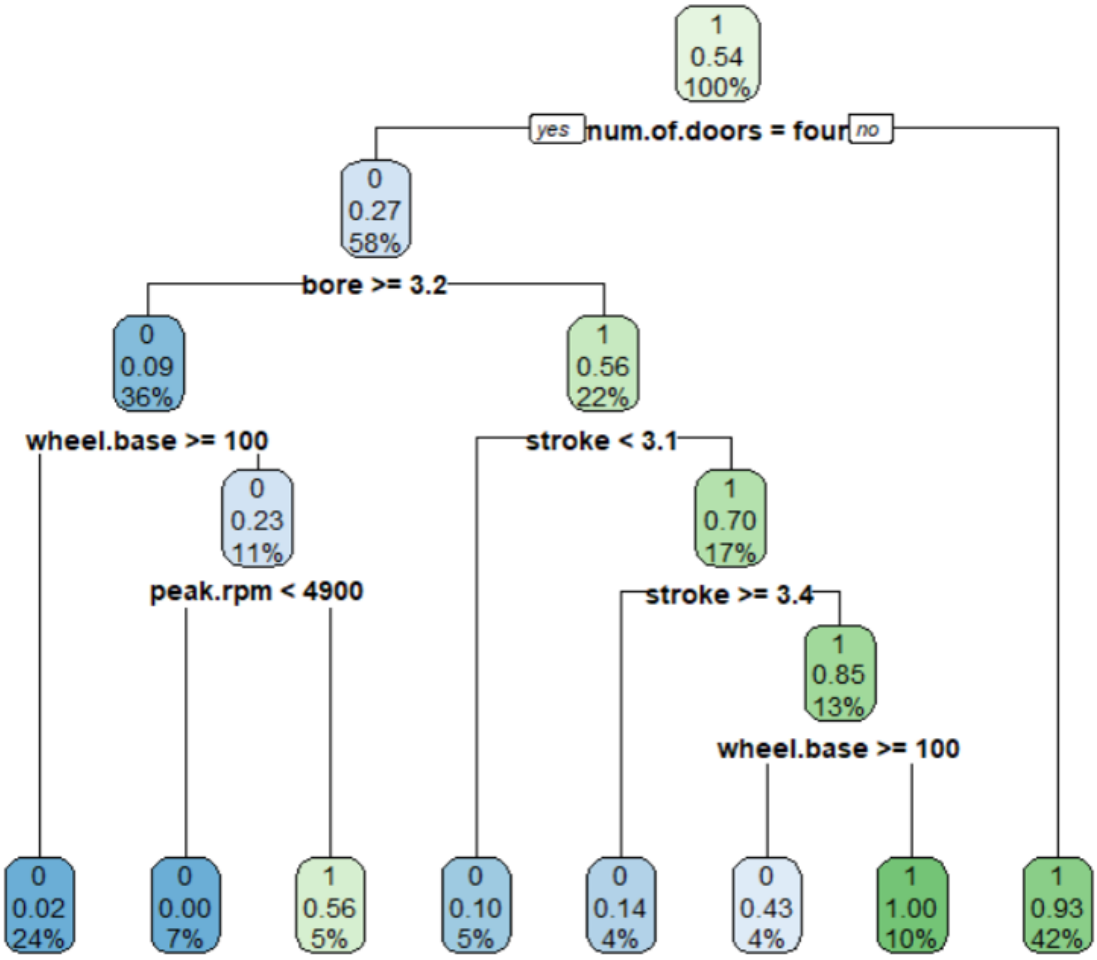
- **Figure 3.2**

ntree	OOB	1	2
50:	7.25%	9.09%	5.71%
100:	7.77%	9.09%	6.67%
150:	7.25%	9.09%	5.71%
200:	6.74%	9.09%	4.76%
250:	6.74%	9.09%	4.76%
300:	6.74%	9.09%	4.76%
350:	5.70%	6.82%	4.76%
400:	6.22%	7.95%	4.76%
450:	6.74%	7.95%	5.71%
500:	6.22%	6.82%	5.71%

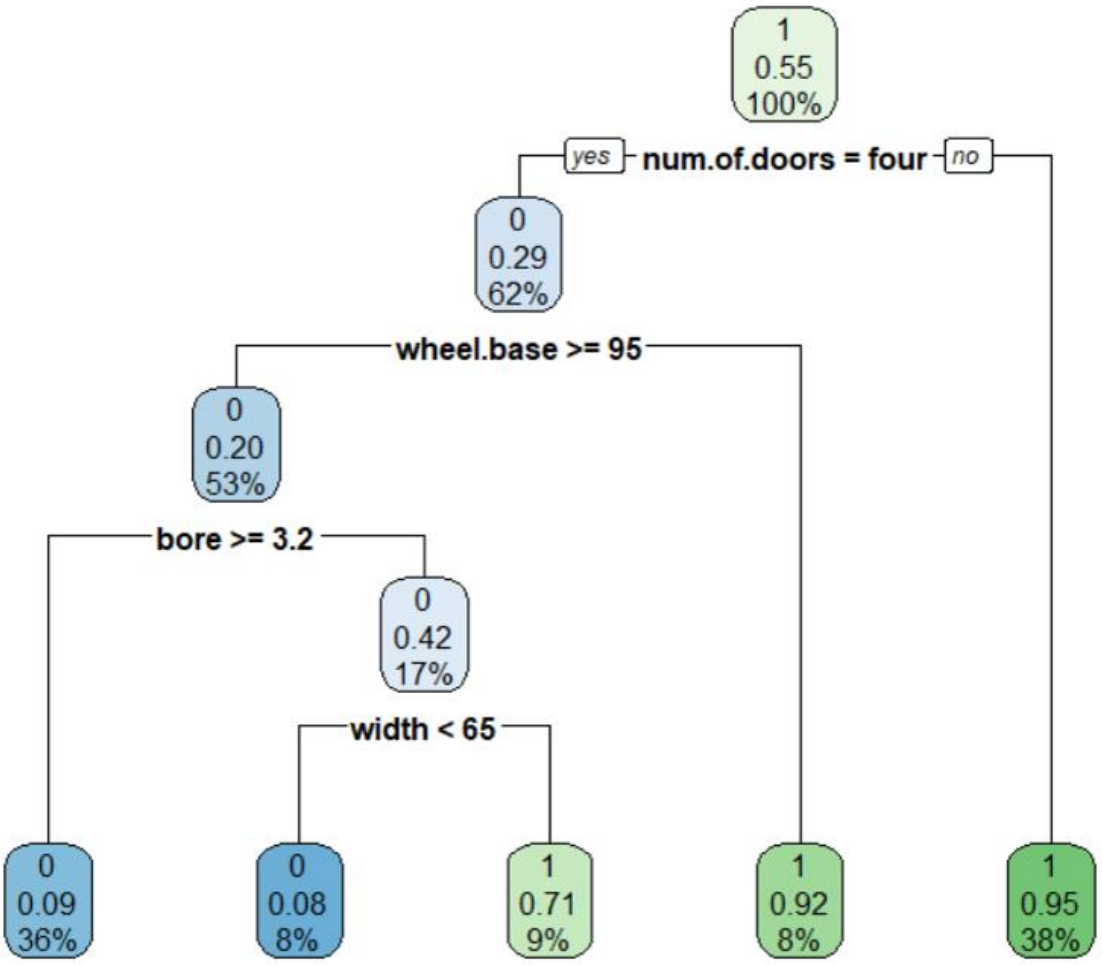
- **Figure 3.3**



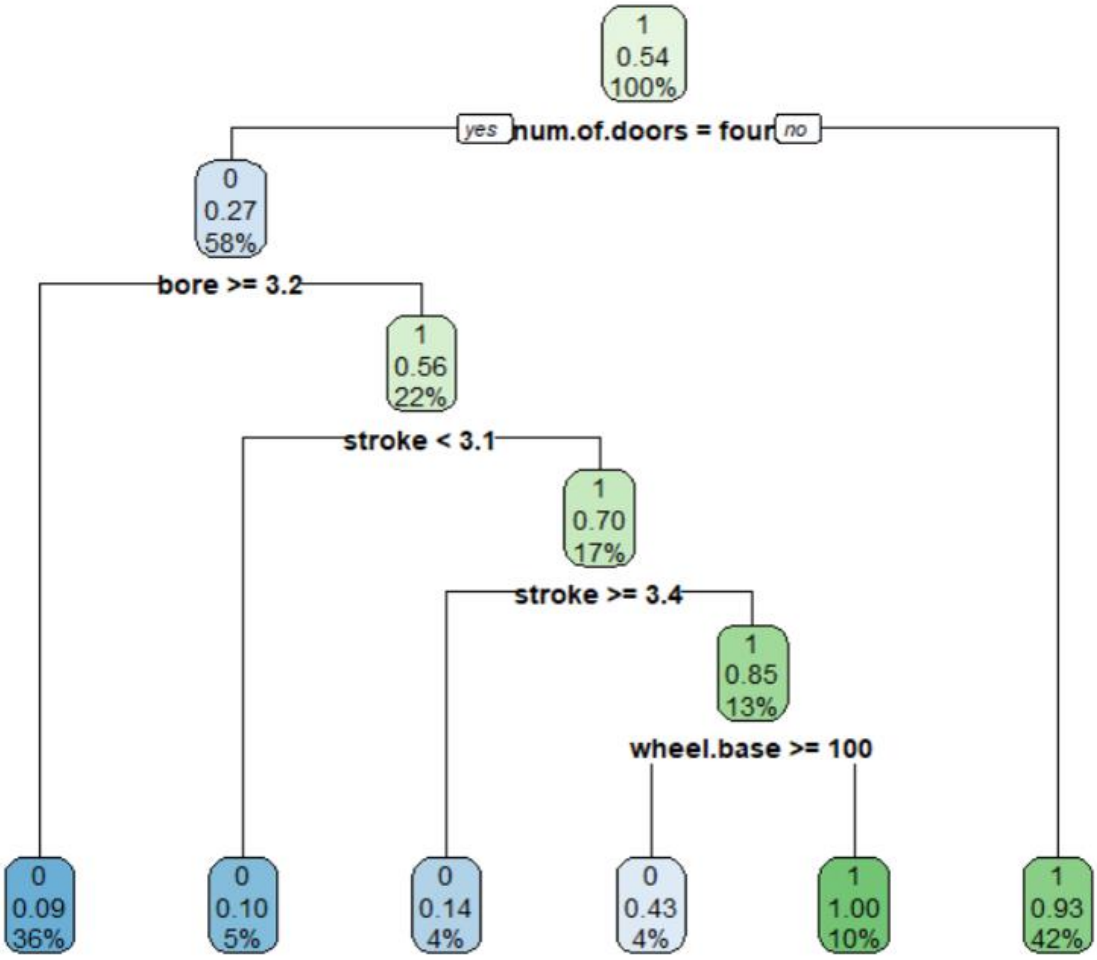
- **Figure 3.4**


1
0.54
100%
yes
num.of.doors = four
no
0
0.27
58%
bore >= 3.2
0
0.09
36%
1
0.56
22%
wheel.base >= 100
stroke < 3.1
0
0.23
11%
1
0.70
17%
peak.rpm < 4900
stroke >= 3.4
1
0.85
13%
wheel.base >= 100
0
0.02
24%
0
0.00
7%
1
0.56
5%
0
0.10
5%
0
0.14
4%
0
0.43
4%
1
1.00
10%
1
0.93
42%

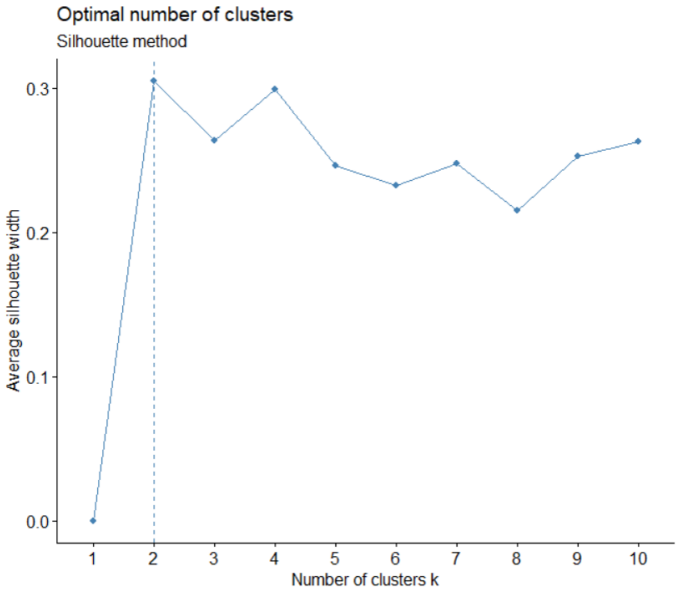
- **Figure 3.5**


1
0.55
100%
yes
num.of.doors = four
no
0
0.29
62%
wheel.base >= 95
0
0.20
53%
bore >= 3.2
0
0.42
17%
width < 65
0
0.09
36%
0
0.08
8%
1
0.71
9%
1
0.92
8%
1
0.95
38%

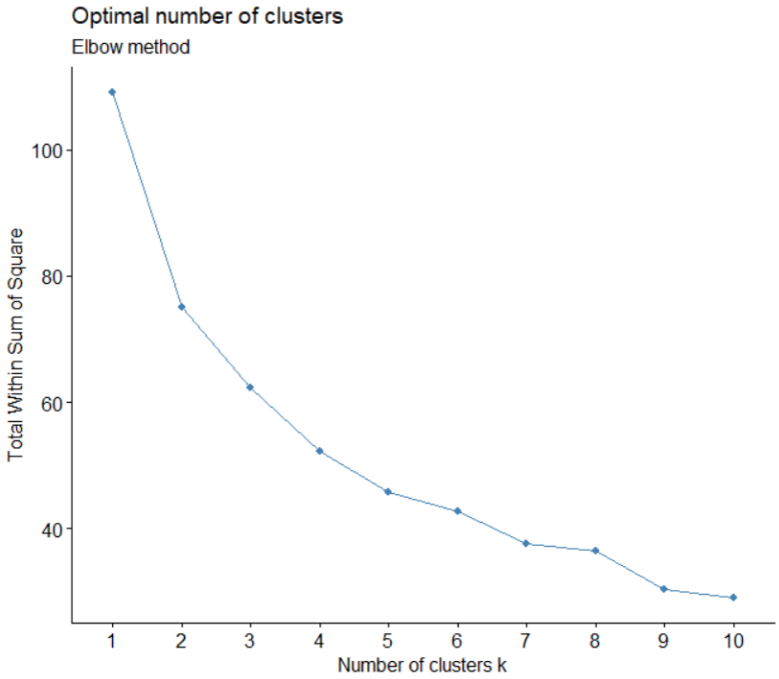
- **Figure 3.6**


1
0.54
100%
yes
num.of.doors = four
no
0
0.27
58%
bore >= 3.2
1
0.56
22%
stroke < 3.1
1
0.70
17%
stroke >= 3.4
1
0.85
13%
wheel.base >= 100
0
0.09
36%
0
0.10
5%
0
0.14
4%
0
0.43
4%
1
1.00
10%
1
0.93
42%

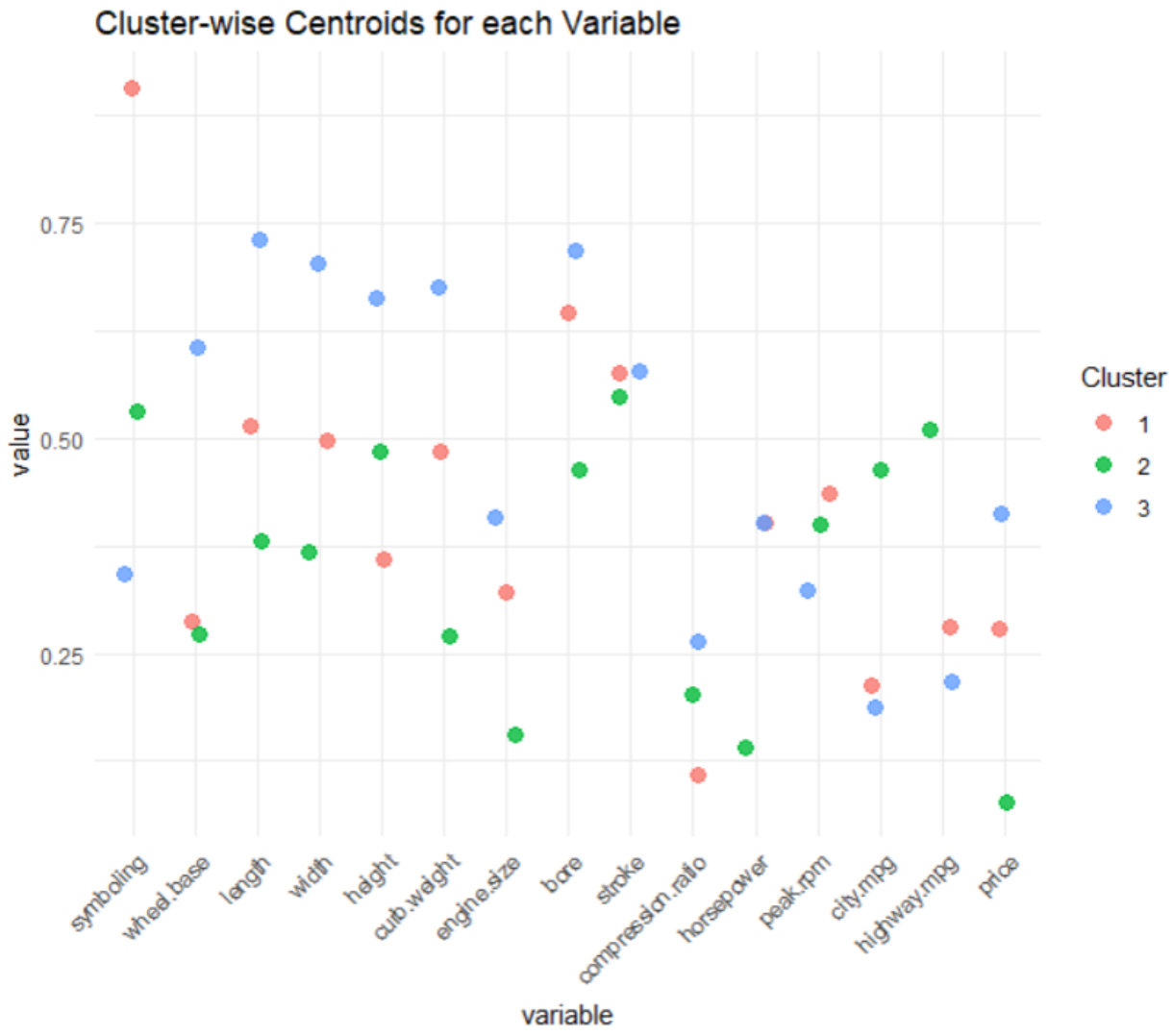
- **Figure 4.1**



- **Figure 4.2**



- **Figure 4.3**


Cluster-wise Centroids for each Variable
value
0.75
0.50
0.25
symboling
wheel.base
length
width
height
curb.weight
engine.size
bore
stroke
compression.ratio
horsepower
peak.rpm
city.mpg
highway.mpg
price
variable
Cluster
1
2
3