1 Introduction

This report is written from the data scientist's point of view working at Porsche. The anonymized name of this person is Kfool. The company is facing scrutiny from the media because a recent report showed that the company's products (Cars) have a very poor risk rating(symboling). The risk ratings are based on the price of the car and adjusted up or down depending on how risky the car is compared to its price. Kfool has been assigned a role to analyse the report and find what the current state of their company is compared to their competitors and how they can reduce this rating. Kfool is a proactive thinker and wants to develop a model that will help the CEO decide what this risk rating is going to be for the cars that they are going to build.

This report discusses how Principal Component Analysis (PCA) and K Means Clustering can be used to gain insights from data, as well as using Random Forest and Gradient Boosted Trees for classification. In addition, this report also covers how Correlation Plots and Boxplots can be used to find correlations between predictors and identify outliers. We will also provide practical examples of their applications. By the end of this report, readers would have an understanding of the various methods used to gain insights from data and the different types of classification methods.

PCA and K Means Clustering are two of the most widely used techniques for finding patterns and extracting meaningful insights from data. PCA is a dimensionality reduction technique that helps to reduce the number of variables in the dataset. K Means Clustering is an unsupervised learning algorithm that helps to group similar data points together into clusters.

Random Forest and Gradient Boosting Trees are two of the most popular algorithms used in machine learning for classification tasks. Correlation Plots and Boxplots are two useful data analysis tools which can be used to identify strong relationships between predictors and identify outliers in the data. These techniques can be used in combination to gain a better understanding of the data and to develop more accurate models.

2 Data Description

This data set includes details about three different aspects of a car: its characteristics, its assigned risk rating, and its normalized losses in comparison to other cars. The risk rating(Symboling) is determined by the car's initial price and any adjustments made to indicate whether it is more or less risky than the price suggests with a +3 indicating it is more risky and a -3 meaning it is probably quite safe. The normalized losses represent the average amount of money paid out for vehicles in a certain size classification per year. To understand more about the correlation and distribution of attributes, different plots (Table 1 and 2) were created.

2.1 Subset

We decided to generate a subset from the original dataset in order to maintain an unaffected reference before we started altering the dataset(referred to as the "final dataset" in the sections that follow). The label variable normalized-losses was not included in the subset because it was outside the scope of the study and wasn't necessary to achieve the study's goal. The characteristics in the dataset were improperly typecasted, which caused a significant disparity. For instance, "price" was displaying as a char attribute even though it should be an integer continuous attribute. To prevent mistakes during the subsequent modification, each attribute was typecasted to the required category.

2.2 Numerical Attribute Segregation

We started by separating the numerical predictors from the entire dataset. The primary motivation of doing this was to implement Principal component analysis and create correlation plots to understand the relation between these attributes. The columns of a categorical variable can be recognised by their use of alphabetical values. However, despite their numerical origin, attributes can still be categorised as

category on the basis of their characteristics. For instance, symboling, a numerical variable that exhibits various levels of risk, was regarded as a categorical predictor for this reason. Following the determination of the categories that each variable belongs to, the categorical columns in the final dataset were transformed into factors so that we could incorporate them into the models.

2.3 Removing Insignificant Predictors

Variables that are highly related to others were also removed. For instance, *price is* distinctly related to *engine.size*. Thus, the latter was dropped. The relations between these different attributes were identified through correlation plot (Refer Table 1) and Principal component analysis plot (Refer table 8).

2.4 Missing Values

We continued the pre-processing procedure with null value checks in our dataset on all the columns and found that none of the rows had null value. After further analyses we realised that missing values are labelled as '?'. We replaced all the '?' in out dataset with null values. Most missing values were in normalized-losses but it wasn't of concern as we had already dropped this column. Although the no. of missing values was very low in the remaining dataset, we decided not to drop the rows with the missing values as the volume of data was already very low . We did thorough analysis and realised that using simpler techniques such as using mean and median weren't going to be enough for all attributes since the missing values could be outliers and we don't want to create data leaks, making our model weak .So we decided to implement random forest to predict the missing values of attributes such as price, bore and stroke. We used median for num.of.doors, horsepower and peak.rpm but some grouping was done beforehand on the basis of correlation analysed before.

2.5 Derived Variables

We were concerned that we might not be able to find relation among different attributes when we create clusters, so we decided to bucket symboling into two categories, one covering negative value and one covering zero and positive values. The primary intention was to use this in the worst case scenario but we didn't had to use it as we made meaningful conclusions with the 6 different levels of symboling.

2.6 Final Adjustment

We were ware that Overfitting in random forest can happen when the model is too complex, meaning it has too many trees or too many levels in the trees. Once our training dataset was organized after performing all the pre-processing steps mentioned above, multiple models were built based on the filtered predictors and the best fits were identified. The final model was built on the training dataset and tested on testing dataset to check its accuracy. The split of training and testing from final data was 80/20.

3 Model Selection

3.1 Data Exploration Using PCA and Clustering

3.1.1 Problem statement and model formulation

The goal assigned was to find correlation between large number of continuous attributes in a dataset. It was easy to find relations between 2 attributes at a time but that would have only complicated the analysis as the total no. of attributes is very high.

Principal component analysis was used to reduce the number of dimensions in the dataset. It was used to identify patterns in data and to summarize the data in a more meaningful way while preserving as much of the information as possible. K means clustering was used on PCA to identify distinct clusters within the dimensionally reduced dataset, to group similar objects together, and to identify outliers within a dataset.

3.1.2 Issues

A major issue was that principal component analysis can only be used for numerical data and does not work with categorical variables. Therefore analysis of only numeral variables was done.
In PCA, Scaling=true data was used to make a data set more consistent by adjusting the range of values. It is also difficult to interpret the components generated by PCA in terms of their meaning. Therefore, Clustering was done on top on PCA to group similar instances together and do further analysis on them. Another issue was that The number of clusters needs to be specified before the algorithm is run, so we tried different clusters and got most meaningful result with 3 clusters.
We were also aware that K-means is prone to local optima, meaning it can find different clusters depending on its initial configuration

3.1.3 final model selection

2 PCAs were done, one with all the numerical attributes to find relation between those attributes and where the company's cars stand when compared to it's competitor's cars with respect to the characteristics of the car . This analysis is presented in plot 8. Clustering was done to segment different groups and find in which group the company belongs.
Another PCA was done by adding the symboling function as a numerical value with an objective to interpret its relation with the characteristics of the car. This helped us get a bigger picture of which characteristics of car were correlated to the sigmoid function. further analysis was done by creating clusters again. This is of immense importance as the goal was to find out how their car perform better on symboling function and which characteristics they should focus on to do so.

3.2 Classification using random forest

3.2.1 Problem statement and model formulation

The company wanted to predict the sigmoid function on the basis of the specification of the car so that they can ensure that they are not making a mistake and using right characteristics while building the car. Since we still hadn't found the significance of categorical attributes, Random Forest algorithm was used to do find the importance of different categorical variables and only most important ones were kept in the final model. Also, random forests are robust to outliers, since the decision trees making up the forest are not strongly influenced by a single outlier. The final model decided on random forest on basis of significance of each predictor was ran on gradient boosting because Gradient boosting typically gives more accurate results than random forests. No. of trees were decision by checking out of box error after every 50 trees generation and one with lowest error was kept.

3.2.2 Issues

Gradient boosted trees are prone to over fitting and therefore interaction depth was kept on the lower side. Also, to ensure that overfitting wasn't occurring. Final dataset was split into test and training and final predictions were compared with the test dataset. Random forest performs poorly when predictors are co-related so we tried to find the relation between attributes with PCA and correlation plots.

3.2.3 Final Model selection

After analysis the PCA report, only those numerical attributes were kept that were of immense importance and it was made sure that the numerical attributes aren't corelated to each other. Different combinations of categorical variables after removing those that were correlated were also used and only those of high significance were kept. The validation was done on test data to make sure overfitting is not happening.

4 Results

4.1 Principal Component Results

From the PCA of numerical variables as shown in table 8, we can see that there is high correlation between engine.size, price and bore. Highway.mpg and city.mpg are also highly co-related. Weight, width and length are also somewhat corelated. Plotting the cars made by Porsche on PCA in different color shows that their cars have characteristics of high horsepower and peak.rpm. low compression.raio and city.mpg. their price is also on the higher side.

From the PCA of numerical variables along with symboling, we can see that Porsche cars also have a high symboling no. which is a bad thing for them and something that they want to reduce.

4.2 K Means Clustering Results

Clustering results of PCA of numeric attributes with symboling are shown below. Violin plot with box plots were created to understand the nature of each attribute in different clusters. For the sake of simplicity and easy interpretation by higher manager authorities, we have mentioned values as (H,M,L) AS high, medium and low respectively as comparative values for clusters.Please refer to boxplot tables in appendix for more info.

Attribute	**Cluster 1**	**Cluster 2**	**Cluster 3**
symboling	H	M	L
price	M	L	H
stroke	M	M	M
bore	M	L	H
horsepower	M	L	H
peak.rpm	M	M	M
city.mpg	M	H	L
highway.mpg	M	H	L
wheel.base	M	L	H
width	M	L	H
height	L	M	H
engine.size	M	L	H
curb.weight	M	L	H

length	M	L	H

4.3 Random forest and Gradient Boosting result

model	OOB
myforest1=randomForest(symboling~length+curb.weight+wheel.base+price+curb.weight+horsepower+highway.mpg+make+fuel.type+aspiration+num.of.doors+body.style+drive.wheels+engine.location+engine.type+num.of.cylinders+fuel.system,ntree=1500, data=Automob, importance=TRUE, na.action = na.omit, do.trace=50)	14.6

myforest1=randomForest(symboling~length+curb.weight+wheel.base+price+curb.weight+horsepower+highway.mpg+make+aspiration+num.of.doors+body.style+drive.wheels+engine.type+num.of.cylinders+fuel.system, ntree=1500, data=Automob, importance=TRUE, na.action = na.omit, do.trace=50)	14.15
myforest1=randomForest(symboling~length+curb.weight+wheel.base+price+curb.weight+horsepower+highway.mpg+make+num.of.doors+body.style+drive.wheels+engine.type+num.of.cylinders+fuel.system, ntree=1500, data=Automob, importance=TRUE, na.action = na.omit, do.trace=50)	14.15

myforest1=randomForest(symboling~length+curb.weight+wheel.base+price+curb.weight+horsepower+highway.mpg+make+num.of.doors+body.style+drive.wheels+engine.type+fuel.system, ntree=1500, data=Automob, importance=TRUE, na.action = na.omit, do.trace=50)	13.66
boosted=gbm(symboling~length+curb.weight+wheel.base+price+curb.weight+horsepower+highway.mpg+make+num.of.doors+body.style+drive.wheels+engine.type+fuel.system, data=train_set,distribution= "gaussian",n.trees=10000, interaction.depth=4)	

Mean Error ^2 of final model on original data : 0.085
Mean Error ^2 of final model on test data : 2.83

5 Classification/predictions and conclusions

5.1 Business Insights

Analysing the PCA ,we can see the that cars made by Porsche (Table 9) are of high price , horsepower and peak.rpm and they also have a high value of symboling which is a bad thing.

Looking at table 9 , porsche's car should be where the dots are green .

Looking at significance of attributes for the final model created, Make of car helps a lot in deciding what the symboling is going to be ,unfortunately Porsche cant change that.

But it can change the no.of.doors which is the second most highest predictor.

More detail on the significance is mentioned in table 6.

Clustering shows how different groups have different characteristics of cars, Porsche should try to make cars with characteristics of clusters that has the most favourable Symboling value.

5.2 Future Recommendations and Improvements

The Mean^2 error in high on test dataset so there is a scope of improvement by reducing the overfitting as OOB error of model was very low.

More analysis of clusters can be done to help make better decision making, data given also had some missing values that were filled by us. If we had cleaner data than our insights would have been better.

We could have went in details with respect to K means clustering , we could have used different distance measures. Different distance measures can be used to determine the similarity between data points. Choosing the right distance measure can help improve the performance of the algorithm.

Maybe we could have analysed more clusters and went more in detail about the characteristics.

More analysis of categorical variables could have been done to interpret the correlation between them.

6 Appendices

Table 1

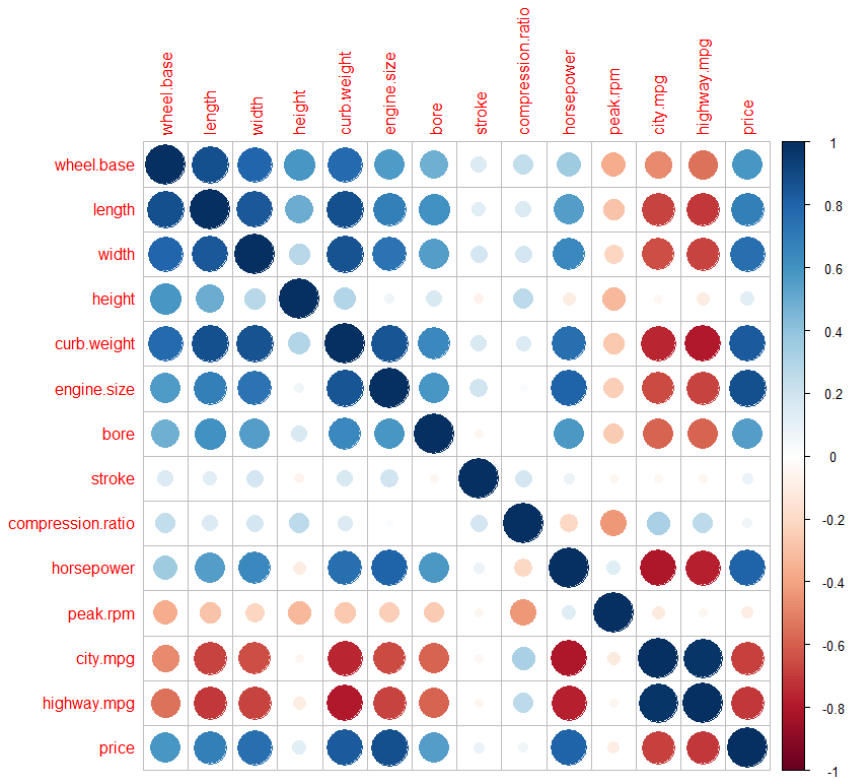


Table 2

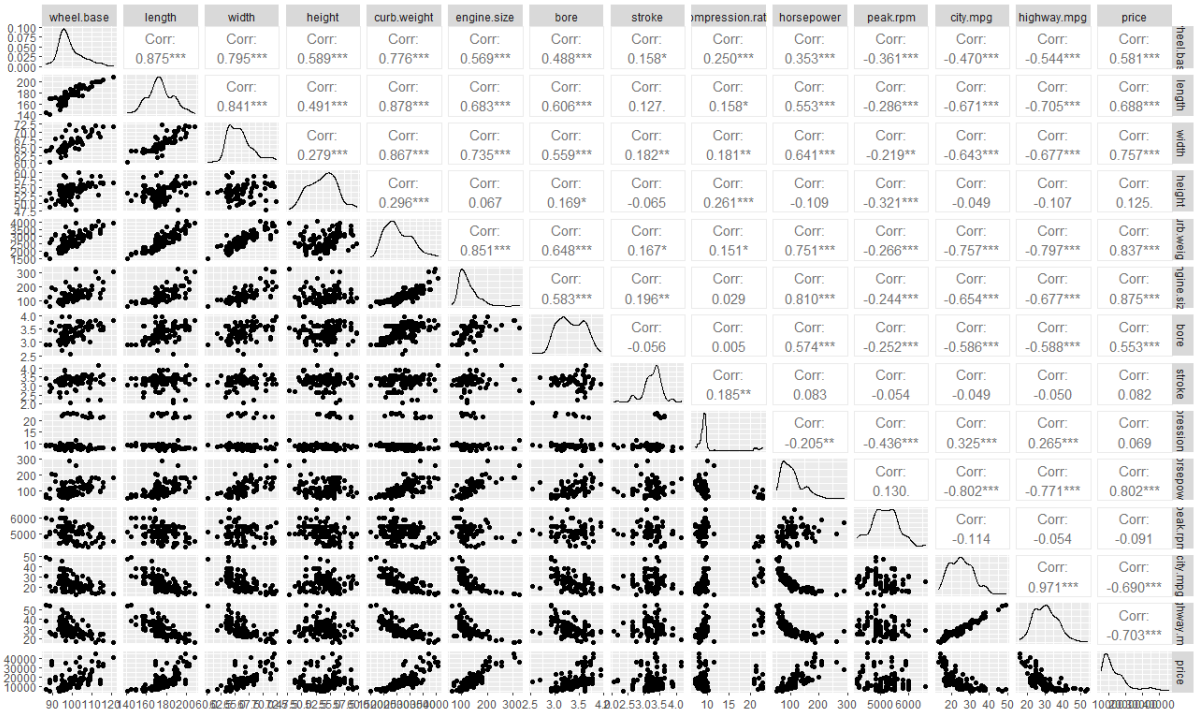


Table 3

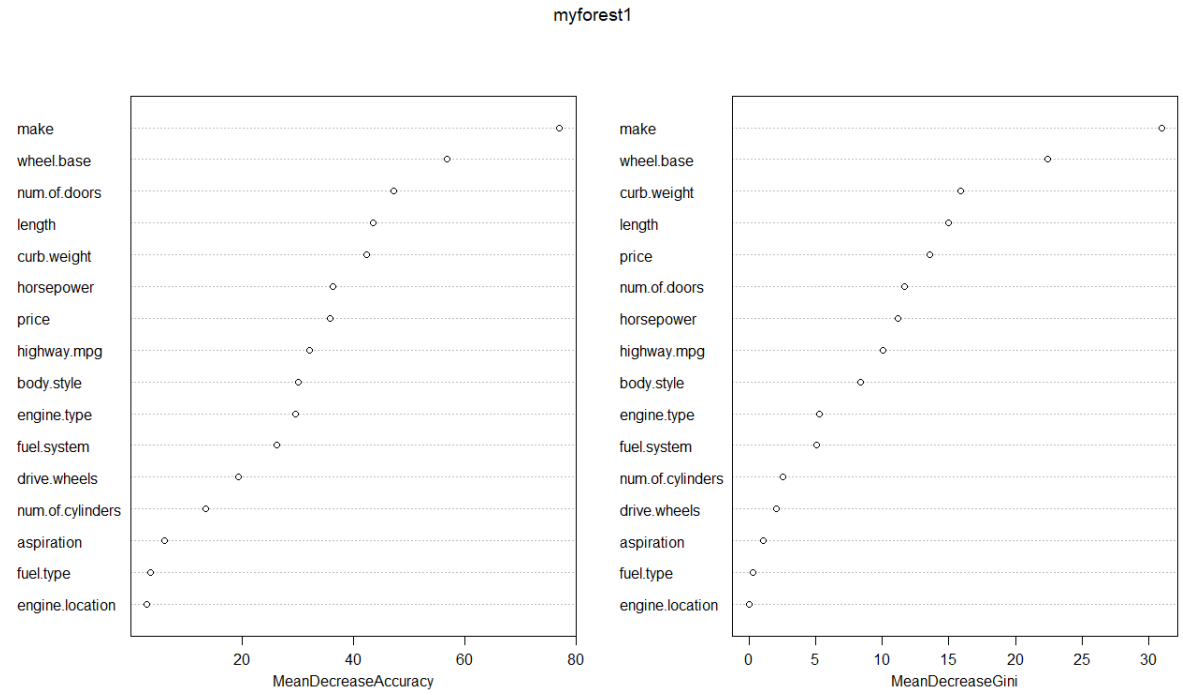


Table 4

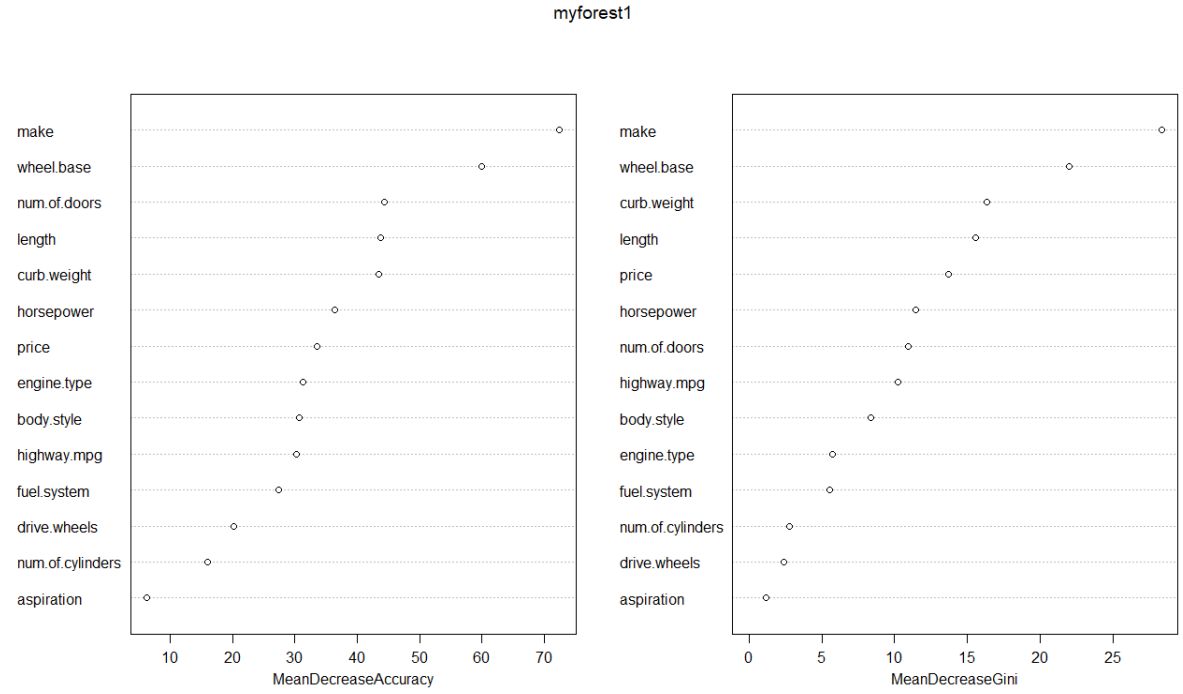


Table 5

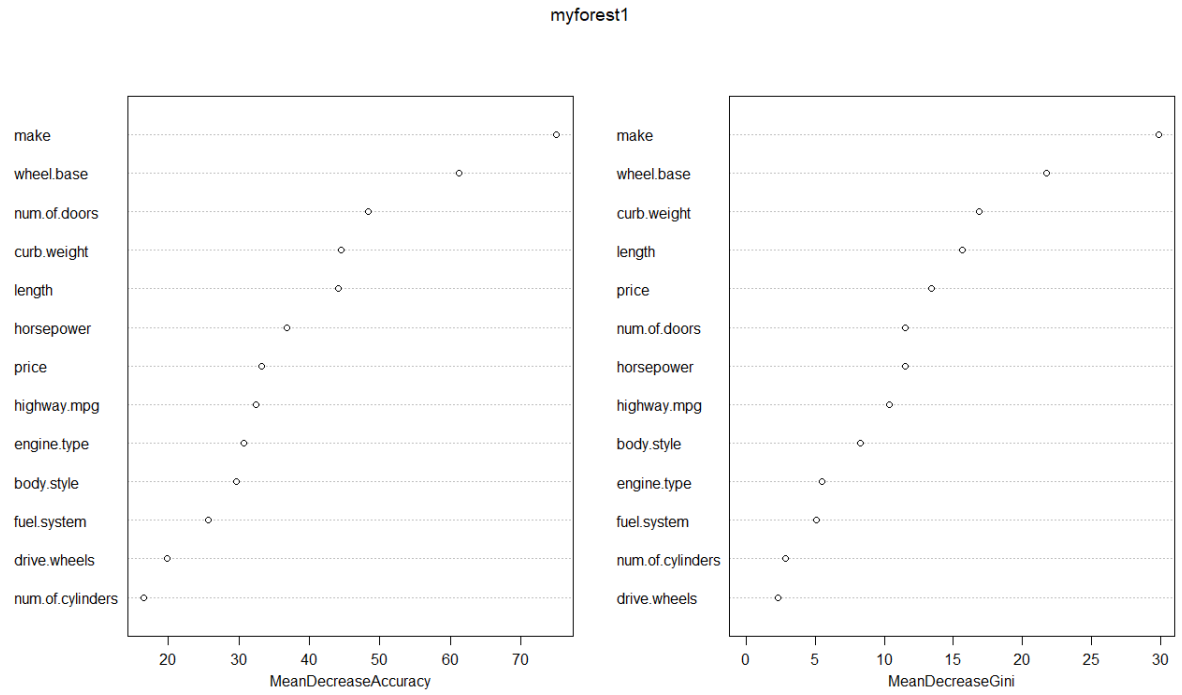


Table 6

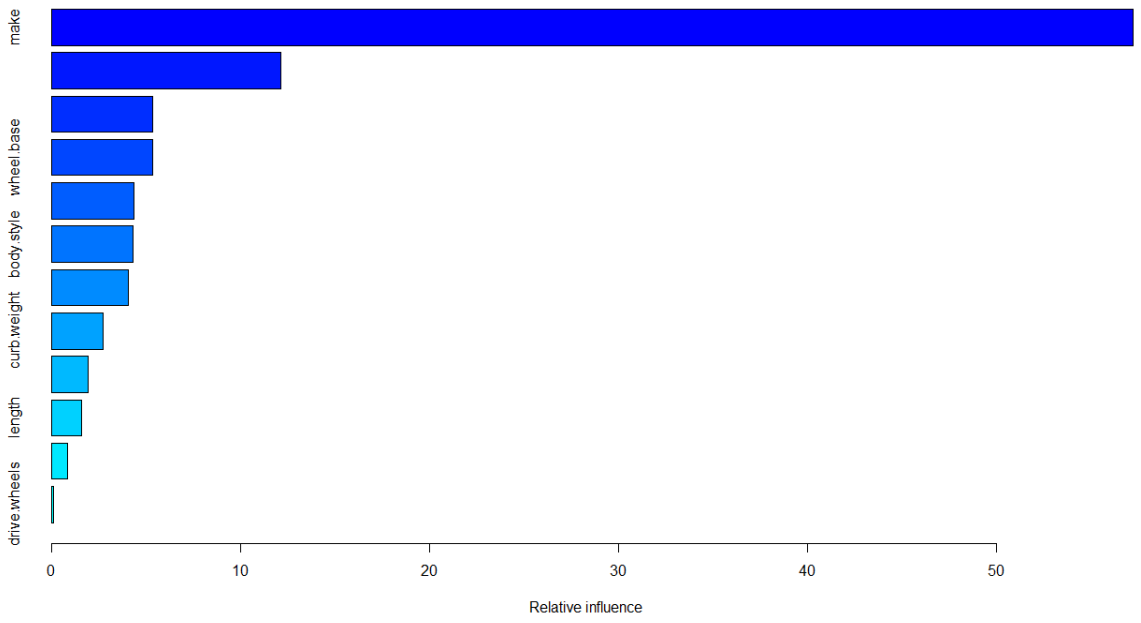


Table 7

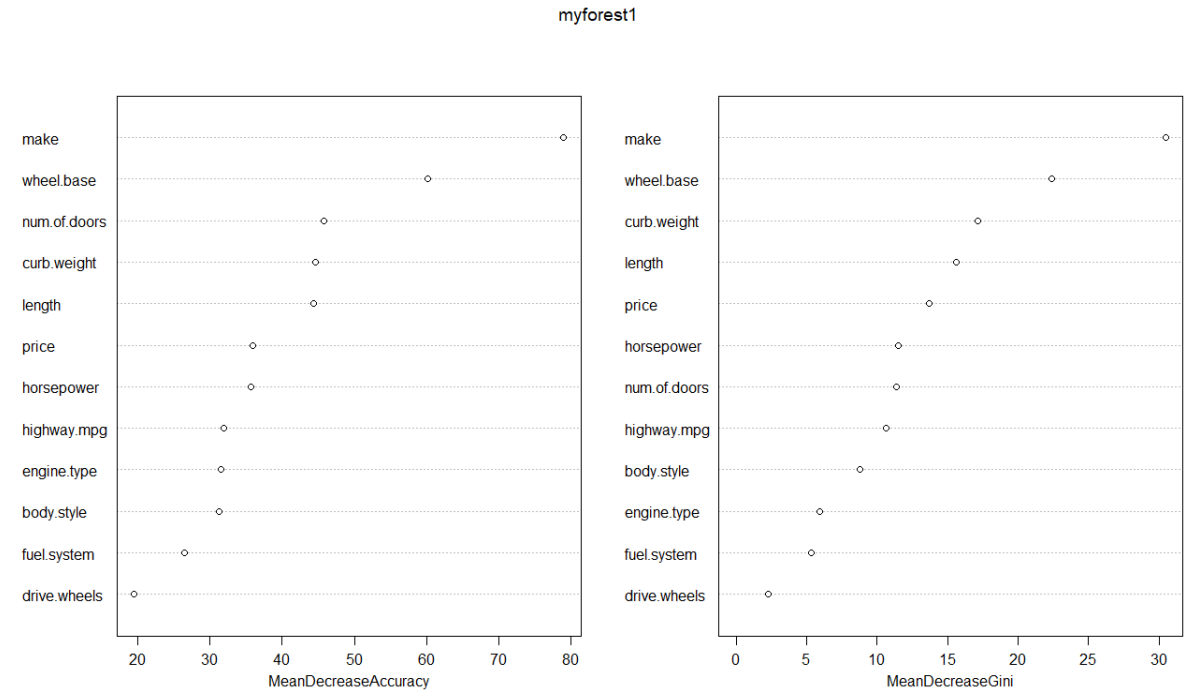


Table 8

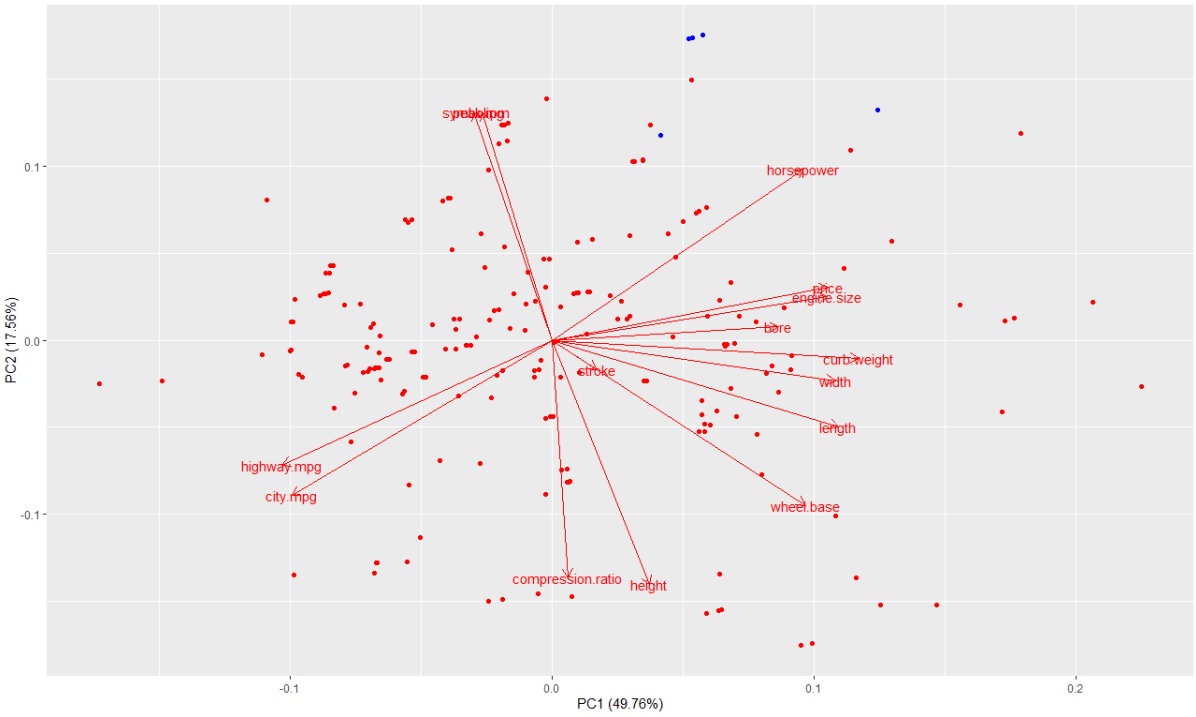
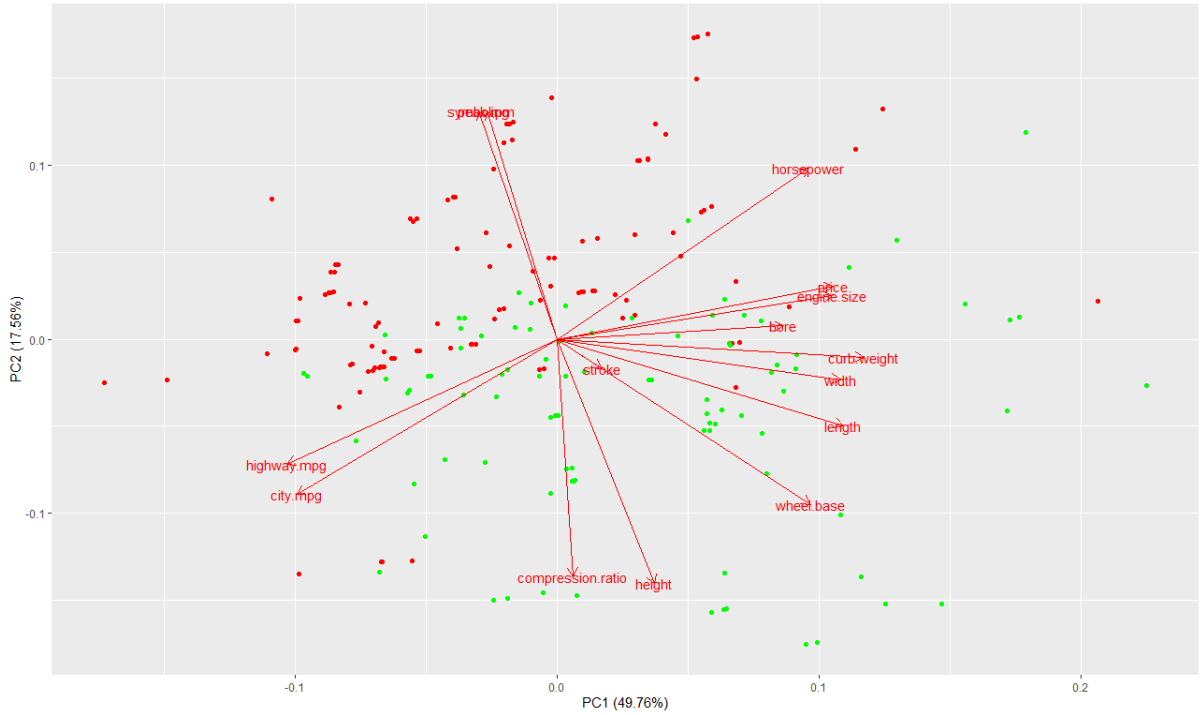
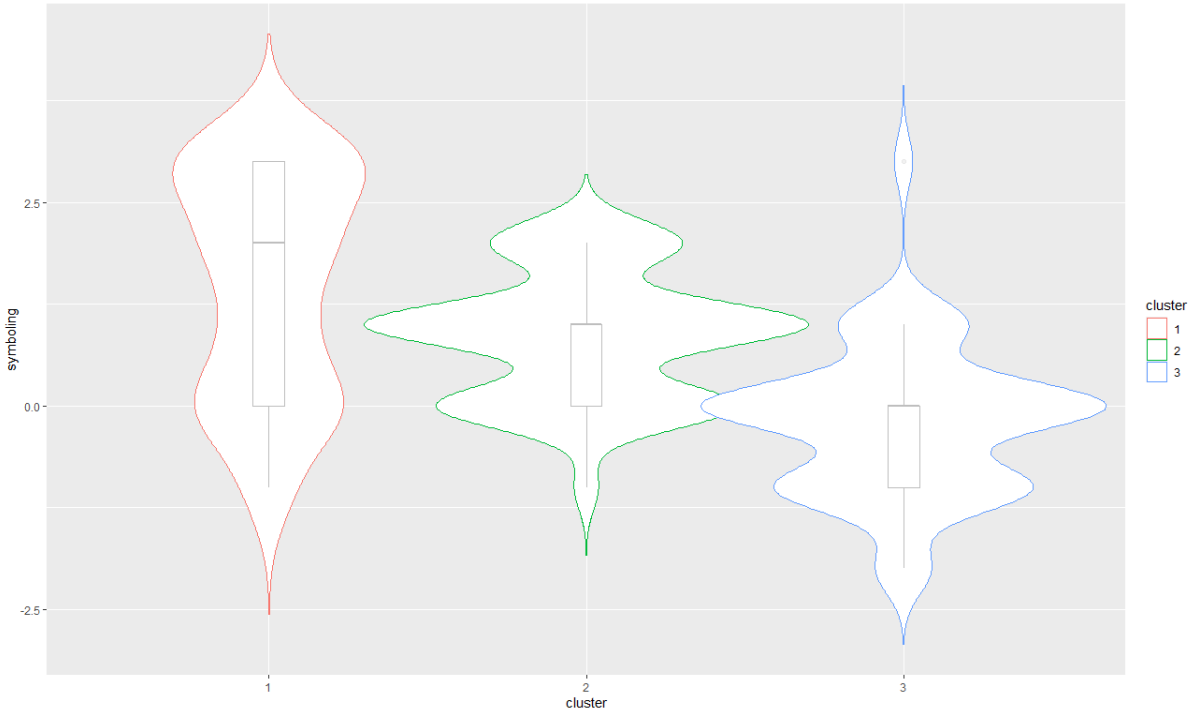
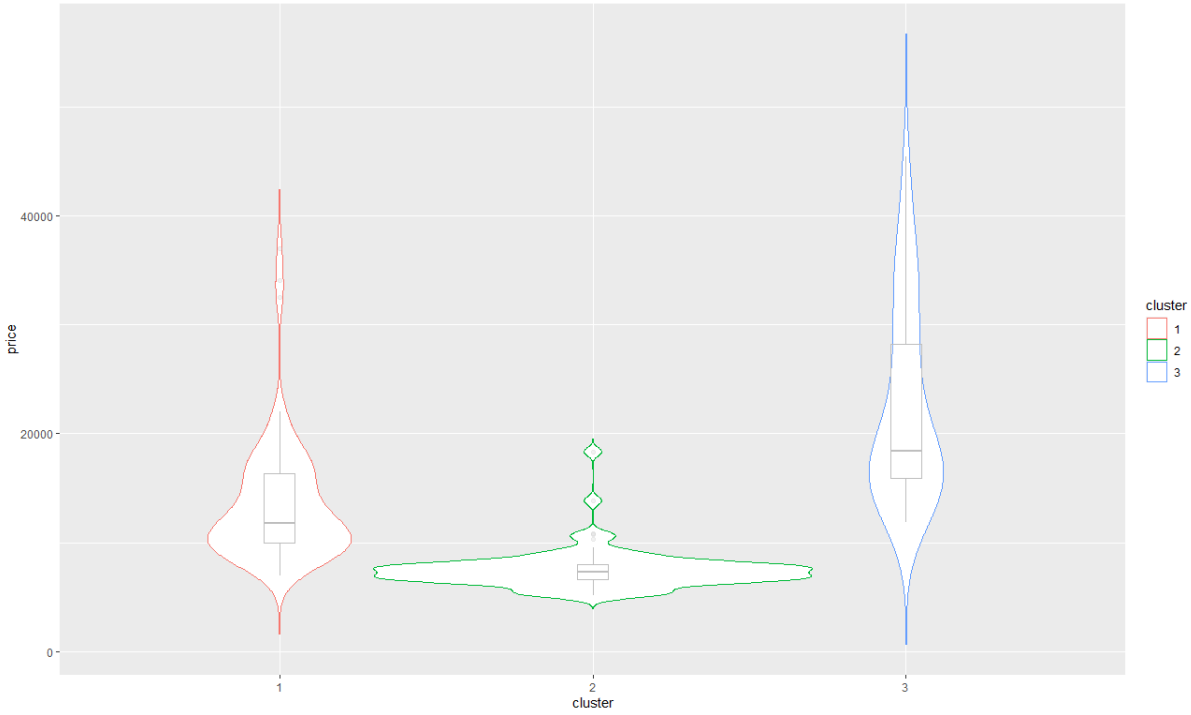


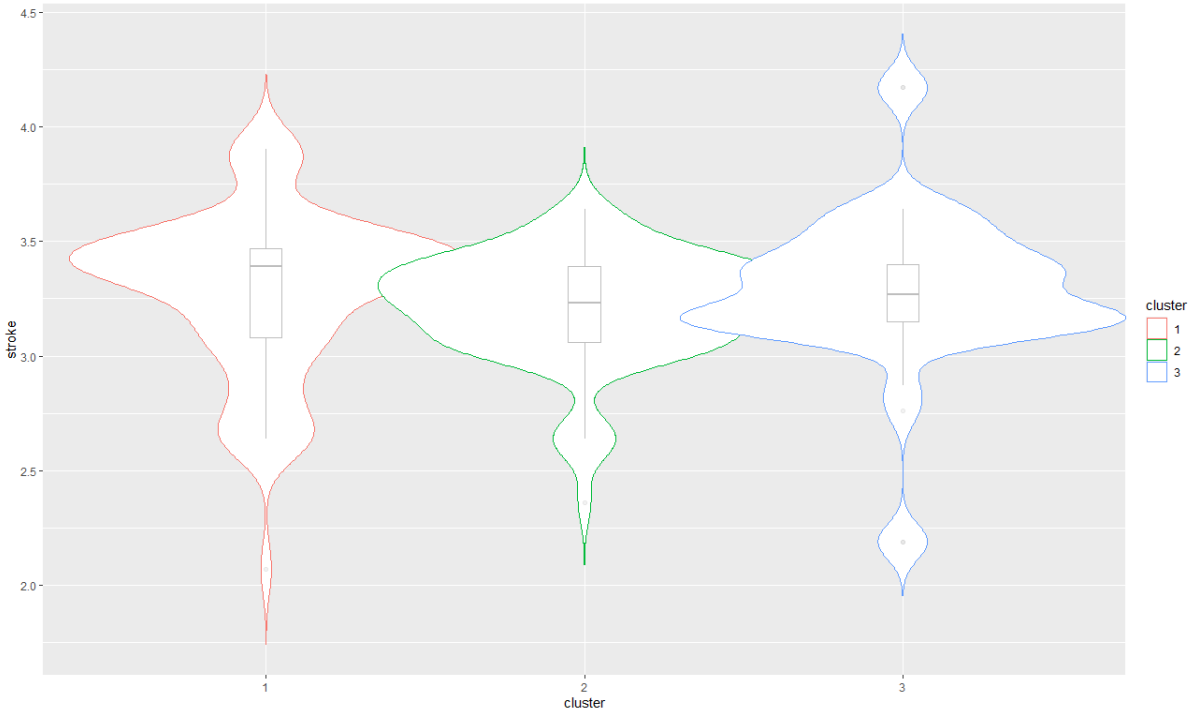
Table 9

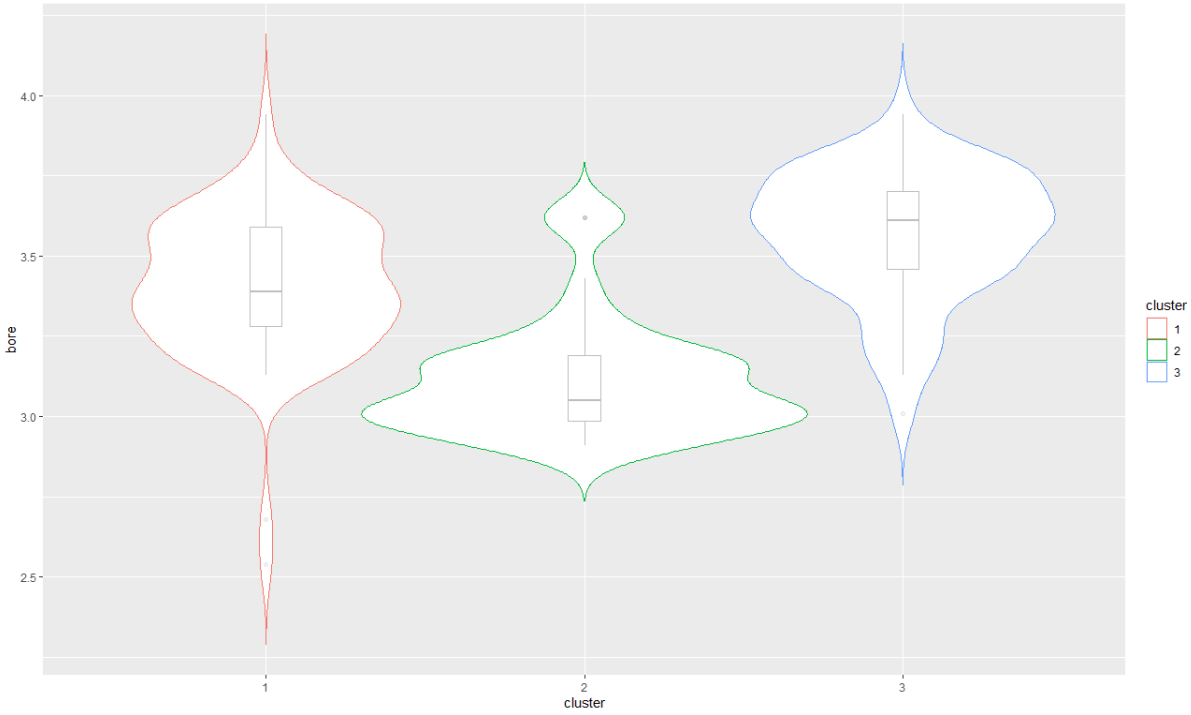

symboling
normalized.losses
horsepower
price
engine.size
bore
curb.weight
stroke
width
length
highway.mpg
city.mpg
wheel.base
compression.ratio
height
PC2 (17.56%)
PC1 (49.76%)
0.1
0.0
-0.1
-0.1
0.0
0.1
0.2

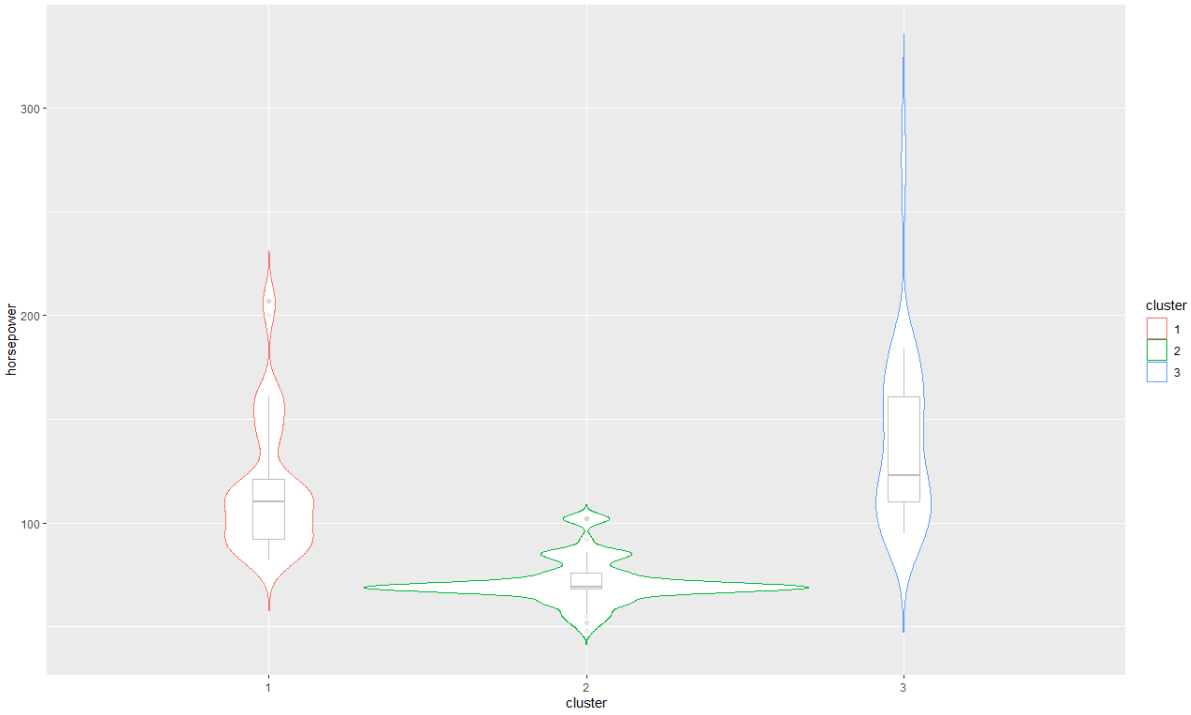
BOX PLOTS :

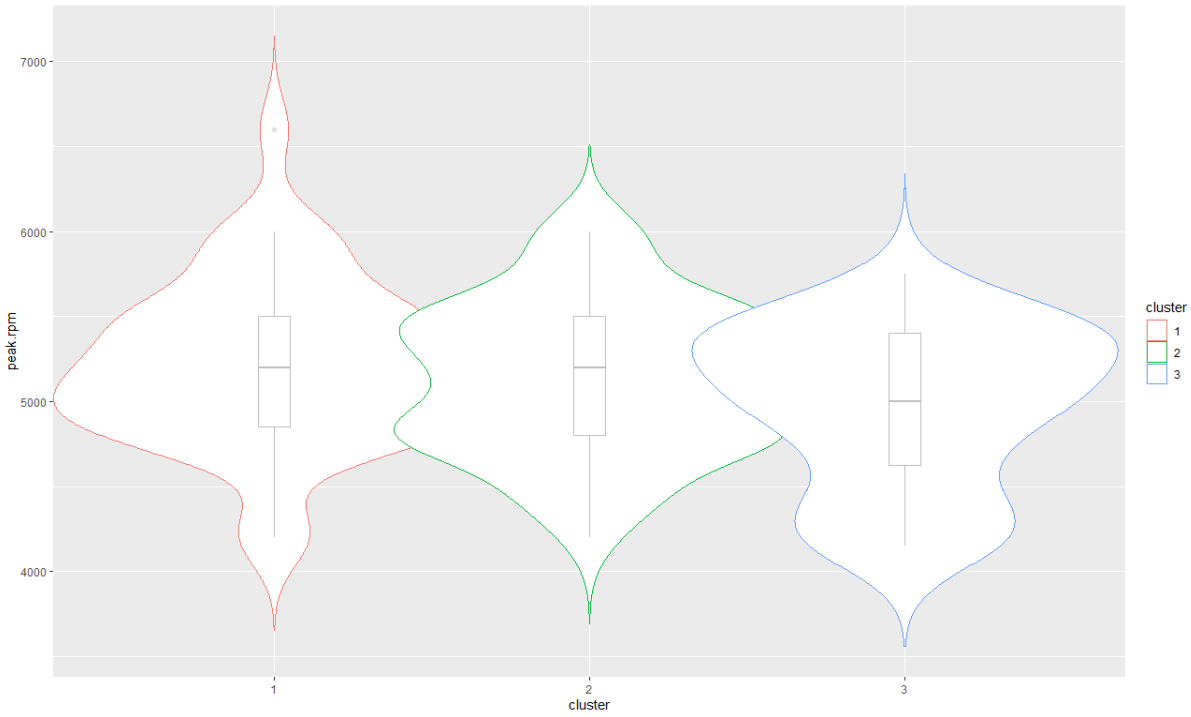

symboling
2.5
0.0
-2.5
1
2
3
cluster
cluster
1
2
3

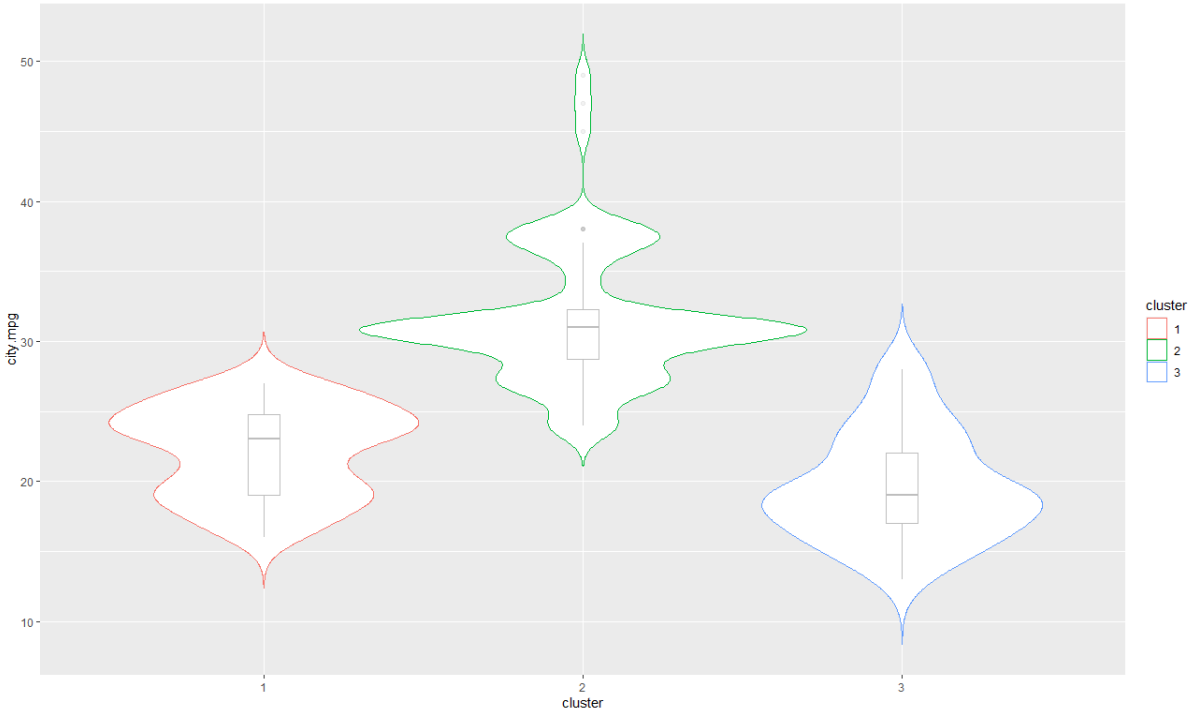

price
40000
20000
0
1
2
3
cluster
cluster
1
2
3

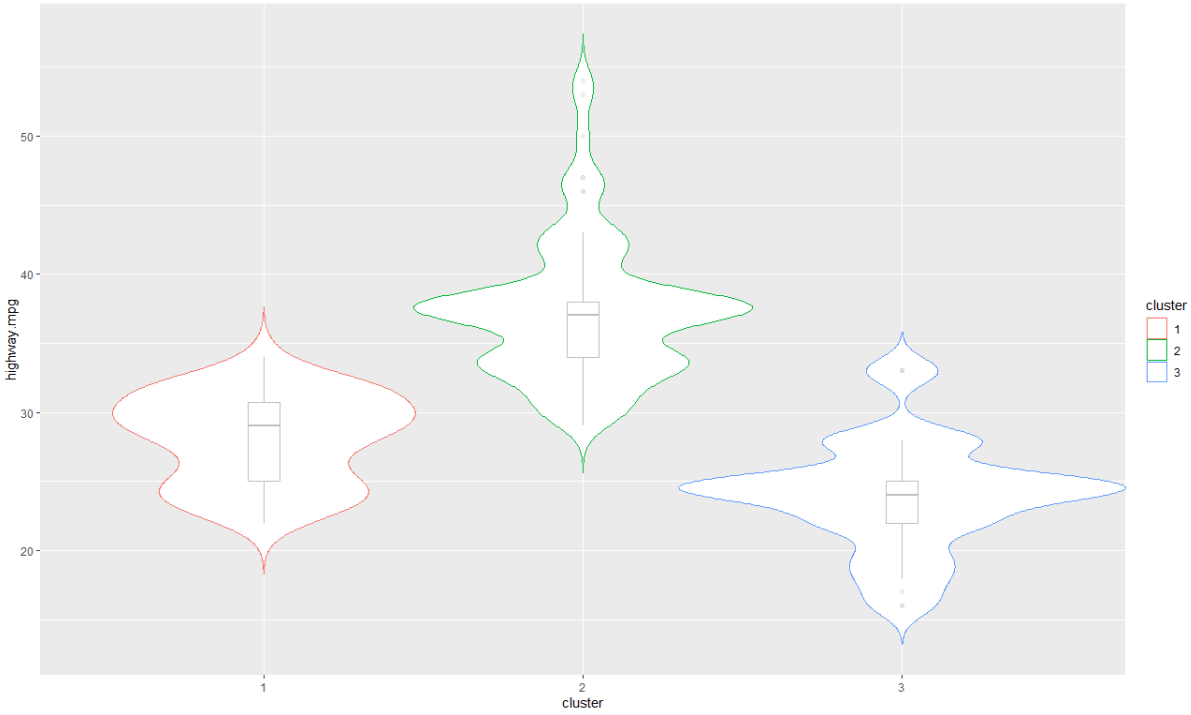

4.5
4.0
3.5
3.0
2.5
2.0
stroke
1
2
3
cluster
cluster
1
2
3

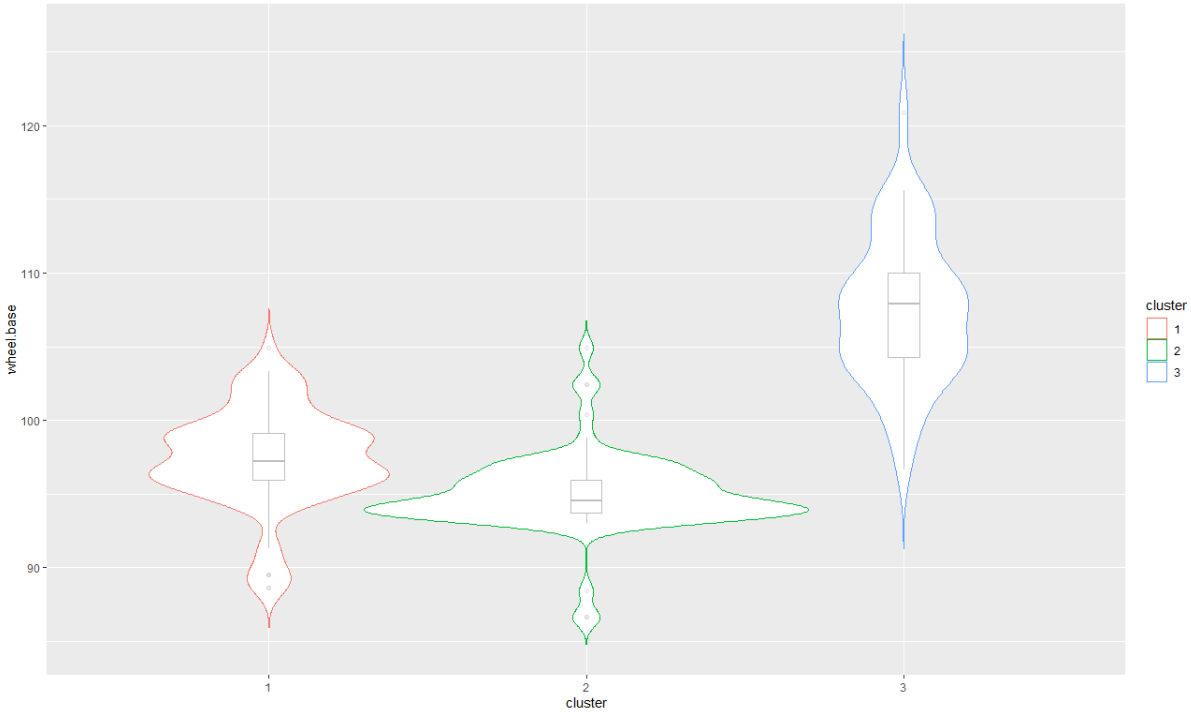

4.0
3.5
3.0
2.5
bore
1
2
3
cluster
cluster
1
2
3

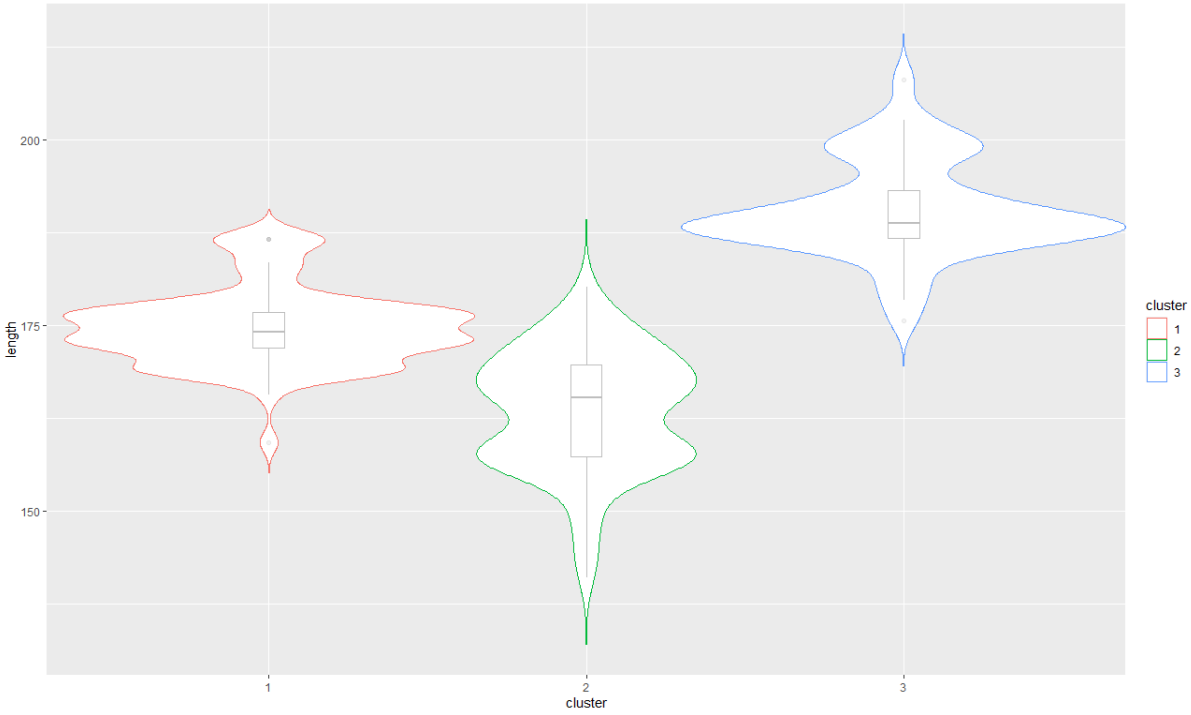

horsepower
300
200
100
1
2
3
cluster
cluster
1
2
3

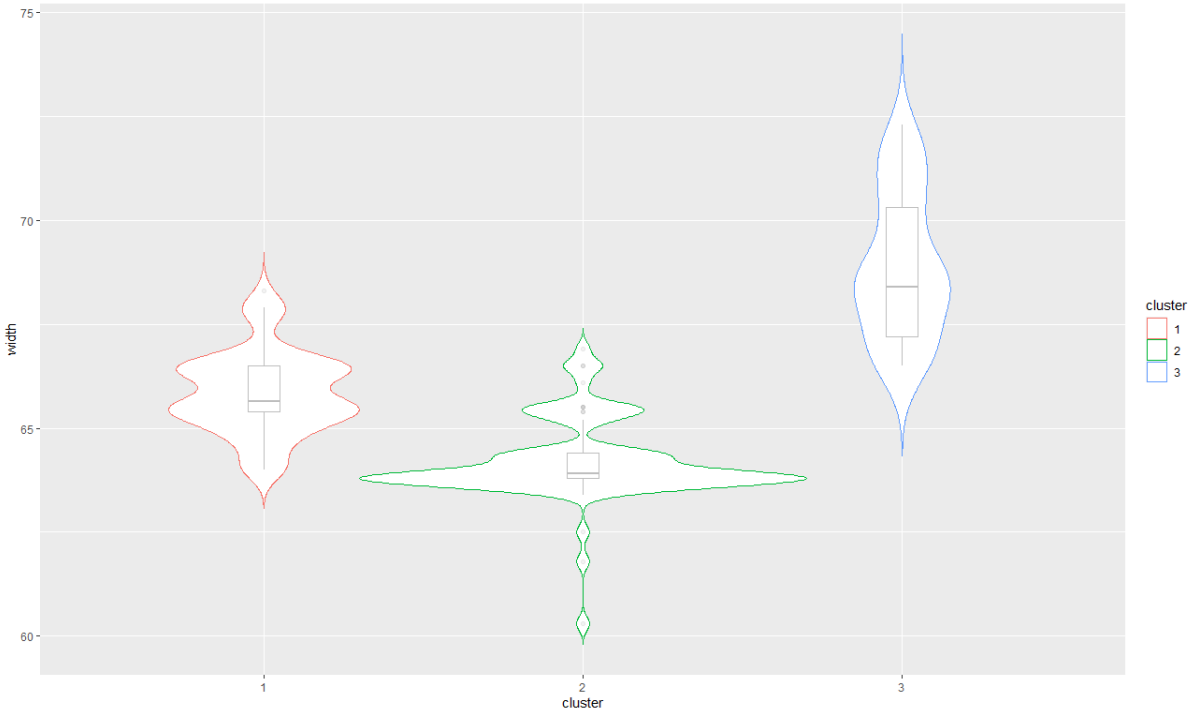

peak.rpm
7000
6000
5000
4000
1
2
3
cluster
cluster
1
2
3

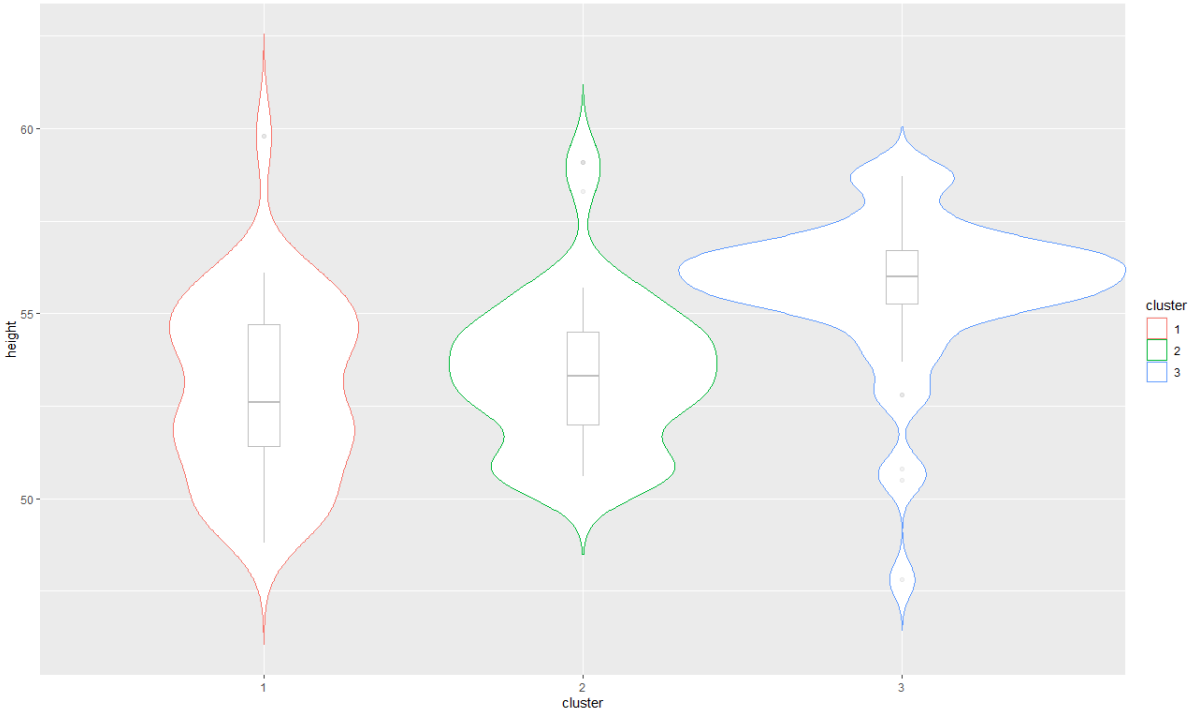

city.mpg
50
40
30
20
10
1
2
3
cluster
cluster
1
2
3

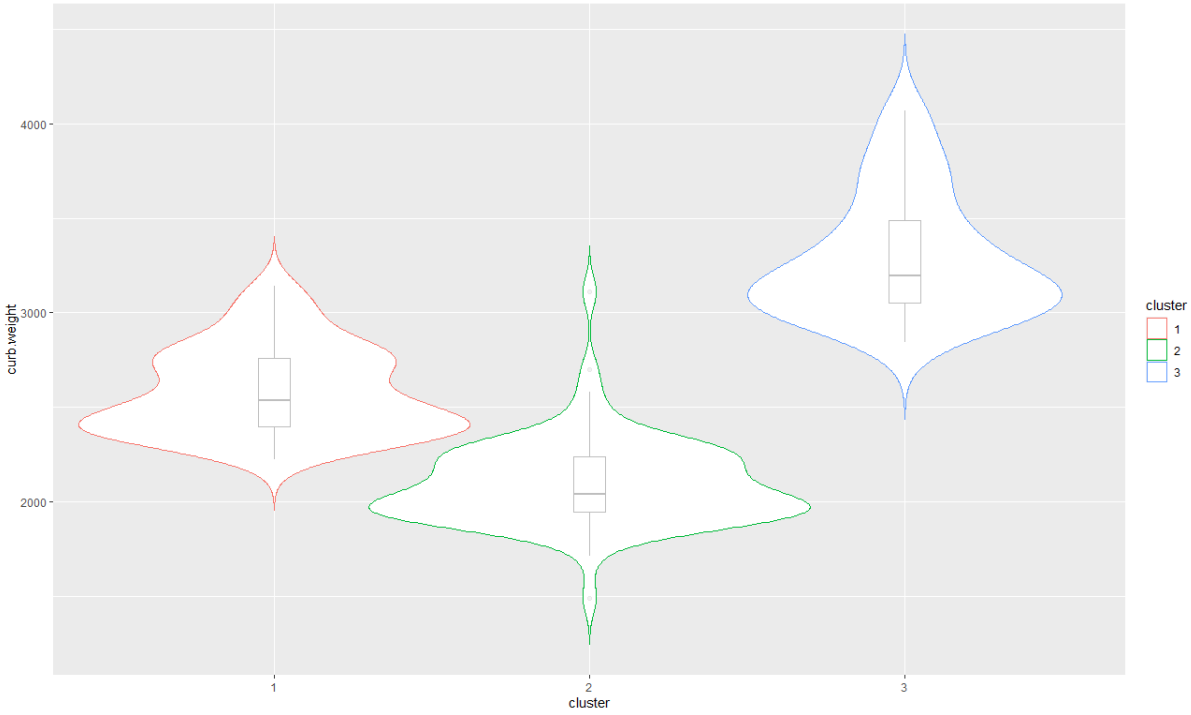

highway.mpg
50
40
30
20
1
2
3
cluster
cluster
1
2
3

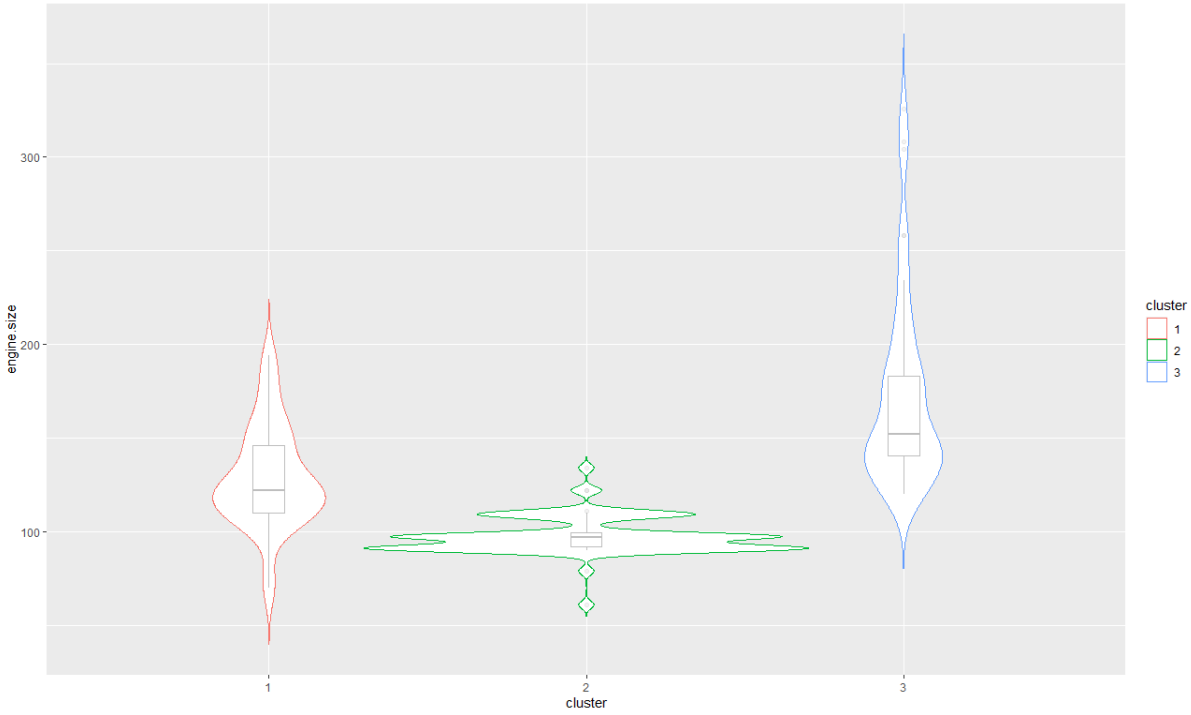

wheel.base
120
110
100
90
1
2
3
cluster
cluster
1
2
3


length
200
175
150
1
2
3
cluster
cluster
1
2
3


75
70
65
60
width
1
2
3
cluster
cluster
1
2
3


60
55
50
height
1
2
3
cluster
cluster
1
2
3





7 Code

The code file of our project is submitted with this report.