Name: Justice Nii-Ayitey

Topic: Vehicle Fuel Economy Data

Stat517: Final Project Report

1. **Introduction**

We have different brands of vehicles such as Dodge, Subaru, Toyota, Volkswagen, Volvo, Audi, BMW and others. Each brand has different models. Toyota for instance, has different models such as Camry, RAV4, C-HR, Prius, Yaris, Land Cruiser, Tacoma, 4Runner, etc. Also, each vehicle has a fuel that it uses.

My motivation for this project was to determine if people buy cars based on the make and to know their fuel consumption and the cost involved per year.

Objectives

* Classifying vehicles based on their mpgData (i.e. if they have or do not have their average, minimum and maximum miles per gallon records)
* Predicting vehicles annual petroleum consumption in barrels
* Association based on vehicles make

Data Description

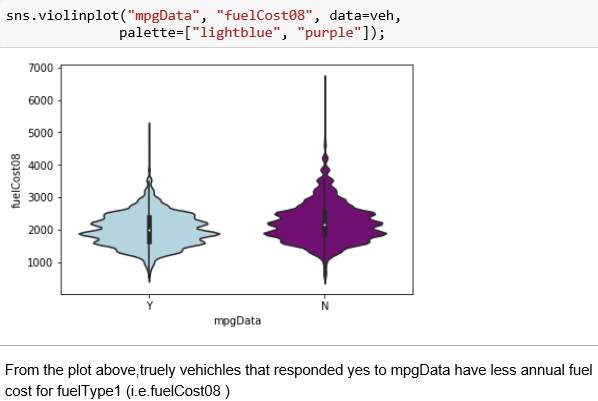
Had dataset from KAPSARC Data Portal (“Vehicle Fuel Economy Data — KAPSARC Data Portal” n.d.). I used 39588 total observations with 83 variables. Some variables of interest are below;

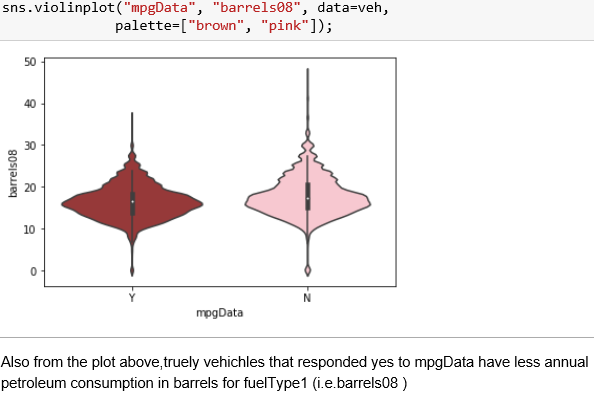
* *barrels08 - annual petroleum consumption in barrels for fuelType1*
* *fuelCost08 - annual fuel cost for fuelType1*
* *mpgData - has My MPG data*
* *make - manufacturer (division)*

1. **Methods**
2. Supervised Learning – Classification

The variable mpgdata was categorical (i.e. Yes or No) and was used as the response variable for the classification. Some models explored under the classification included K-Nearest Neighbor, Naïve Bayes, Logistic Regression, Decision Trees, Random Forest, Neural Networks and Support Vector machine.

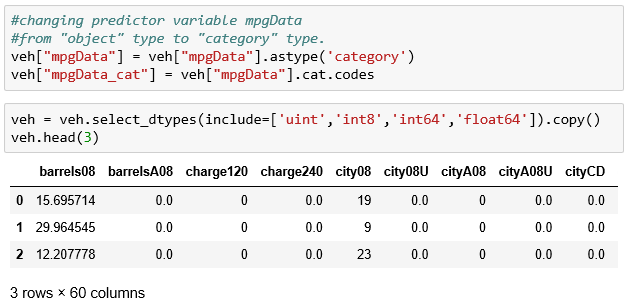
Below are some plots for visualization;

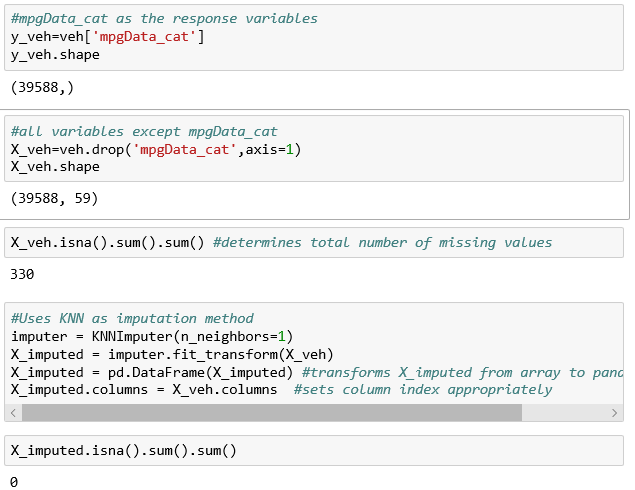




1. Preprocessing

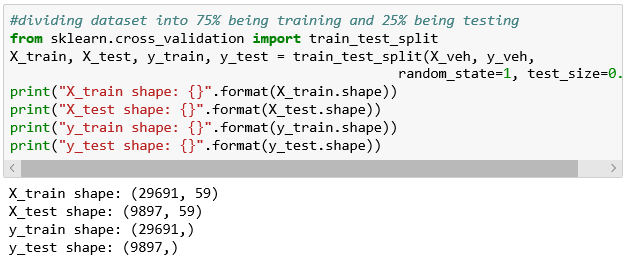
The first preprocessing made was to change the predictor variable mpgData from “object” type to “category” type. K-Nearest Neighbor was also used for imputation to replace missing values.





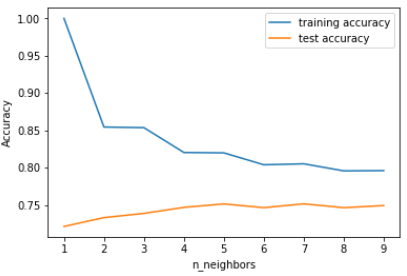
1. Partitioning

Dataset were being divided into training and testing which was used to explore various classification models



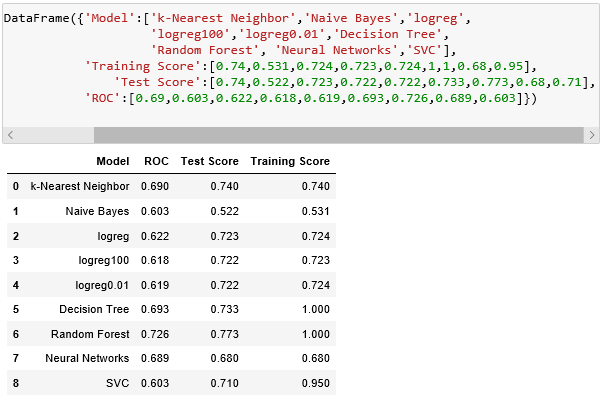
*Analyzing K-Neighbors Classifier*

With a range from 1 to 10, the test accuracy is high when n\_neighbors is at 7 from the plot below



1. Summary

The table below is a summary of ROC, Test Score and Training Score of the classification models.



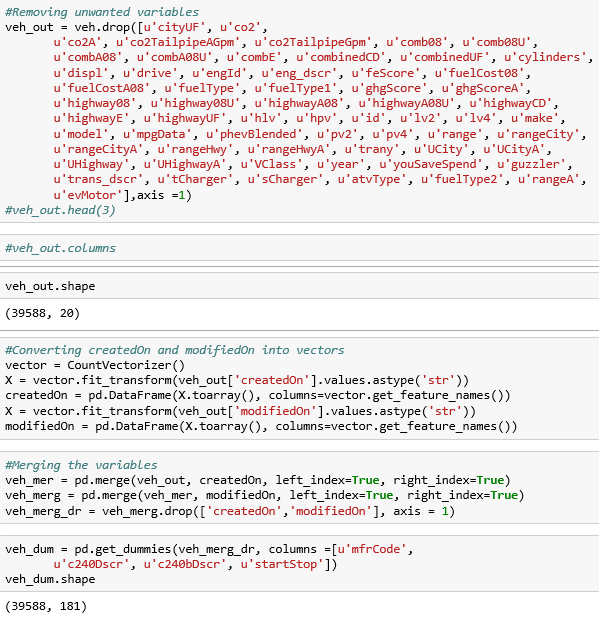
***Table 1****: Summary of Classification Models*

1. Supervised Learning-Regression

The variable barrels08 was continuous and hence it was used as the response variable for the regression analysis. Some models explored under regression are K-Neighbors Regressor, Linear Regression (aka Ordinary Least Squares), Ridge Regression, Lasso, Decision Tree Regressor and the Random Forest Regressor. Also, under the regression, R^2 was being performed for accuracy.

1. Preprocessing

The first preprocessing made under the regression analysis was to remove the unwanted variables. Afterwards, the variables (i.e. createdOn and modifiedOn) were being converted into into vectors. The variables were being merged and dummy coding was applied.



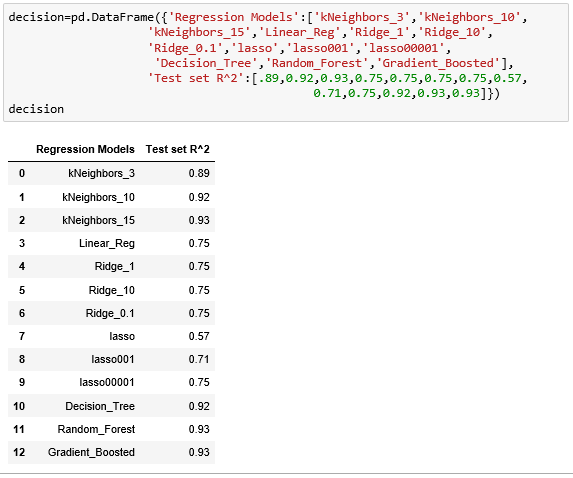
1. Partitioning

First, Principal Component Analysis (i.e. PCA) was being performed of which 150 PCA’s represents about 90% of the variability in the dataset. Afterwards, the dataset was being divided into training and testing and the models listed above were being explored for the regression analysis.



1. Summary

The table below is a summary of the Test set R^2 for the various regression models respectively.



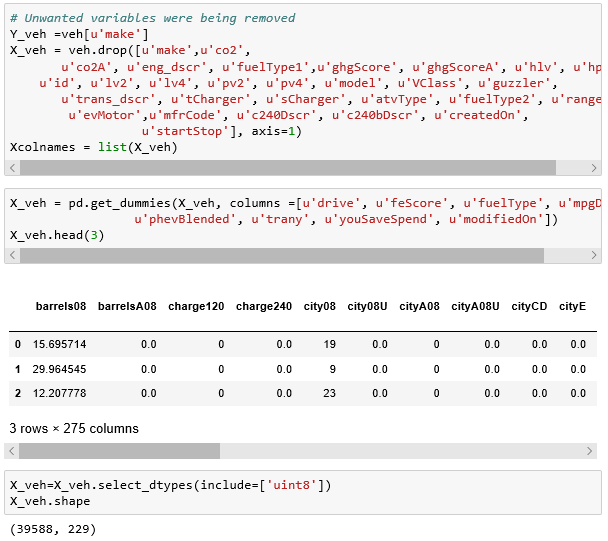
***Table 2****: Summary of Regression Models*

1. Unsupervised Learning-Clustering

The make (i.e. manufacturer) was considered here as the response variable and some clustering were being performed based on that.

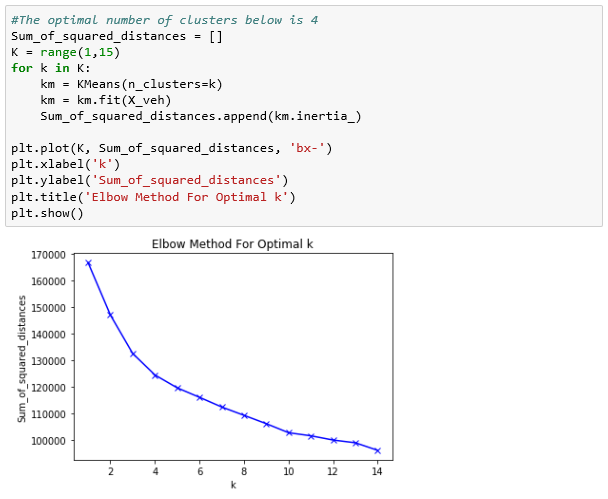
1. Preprocessing

The unwanted variables were being removed and some variables were being dummy coded.

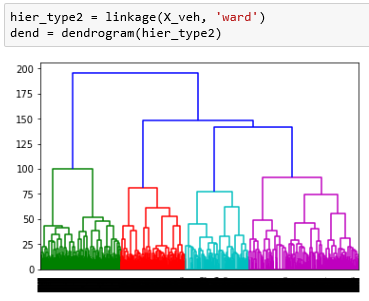


1. Silhouette and the Elbow Method were being used to determine the Optimal Number of Clusters.

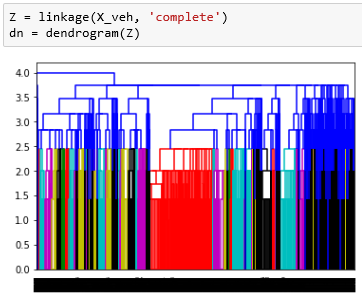




1. Hierarchical Clustering



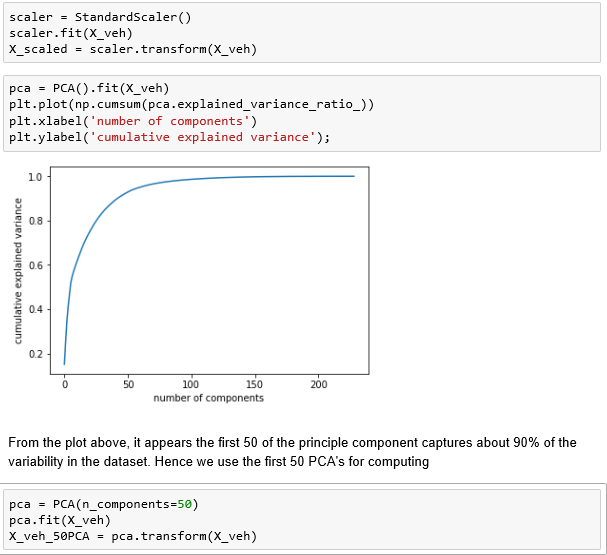
From the plot above, using “ward” for the linkage was able to produce 4 groups or clusters as the Elbow method suggested.



From this plot, “complete “was used for the linkage but didn’t produce good clusters.

1. Principal Component Analysis (PCA)

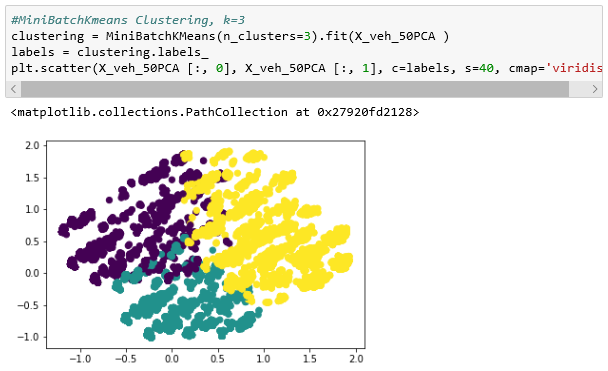
The PCA was used to reduce dimensions. Below is the code and the graph used;



1. K-means Clustering

MiniBatchKmeans was able to do good clustering as suggested by Silhouette when *k=2* with no overlapping. When *k=3 and k=4* for the MiniBatchKmeans, clustering was made but they were overlapping which do not make them good clusters. Below are graphs to illustrate that;







1. Unsupervised Learning-Association

Make as a variable was also used for the association to check factors or reasons for people going in for a make of vehicle.

1. Minor association performed in Python

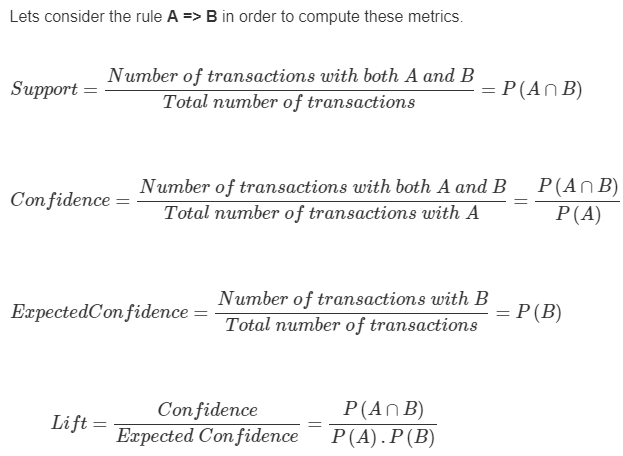


The first 20’s associations of the vehicles make can be seen in the plot above. The itemsets (phevBlended\_False) had the highest support (0.998257) while the itemsets (mpgData\_N, drive\_Rear-Wheel Drive) is with the lowest support (0.254067).

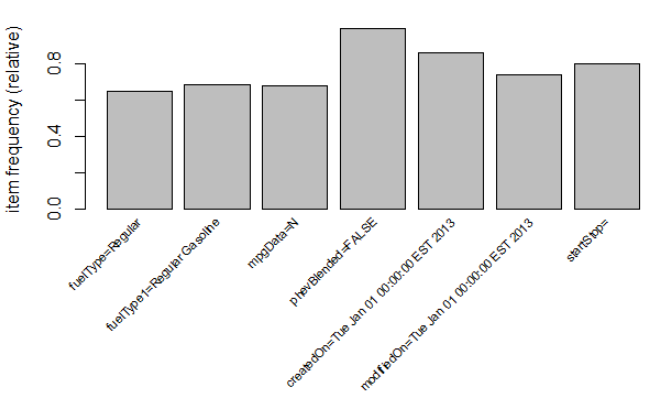
1. Major association performed in R using apriori ()

The apriori () generates the most relevant set of rules from a given transaction data. It also shows the support, confidence and lift of those rules. These three measures can be used to decide the relative strength of the rules.

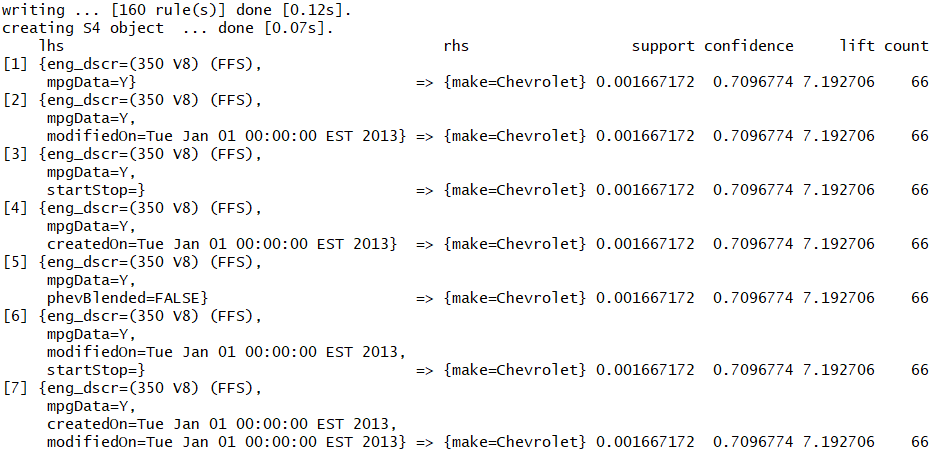
Lift is the factor by which, the co-occurrence of A and B exceeds the expected probability of A and B co-occurring, had they been independent. So the higher the lift, higher the chance of A and B occurring together(“Association Mining With R | arules” n.d.). Also, below is the mathematical representation of the three measure;

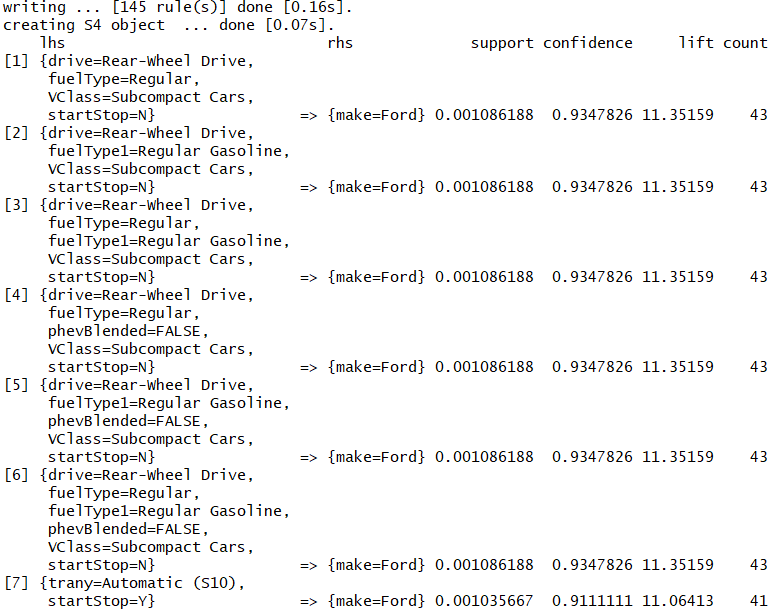


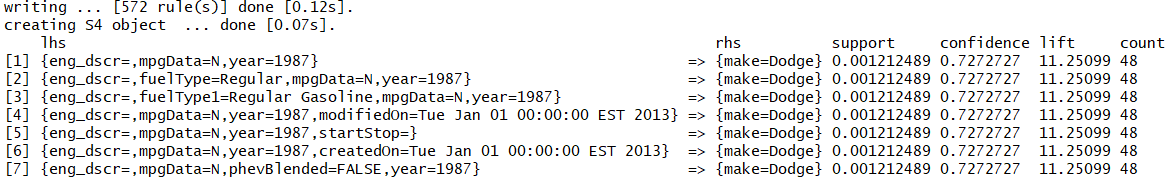
Descriptive Statistics after removing unwanted variables: the variable phevBlended is with the highest relative frequency (0.9) as seen below.

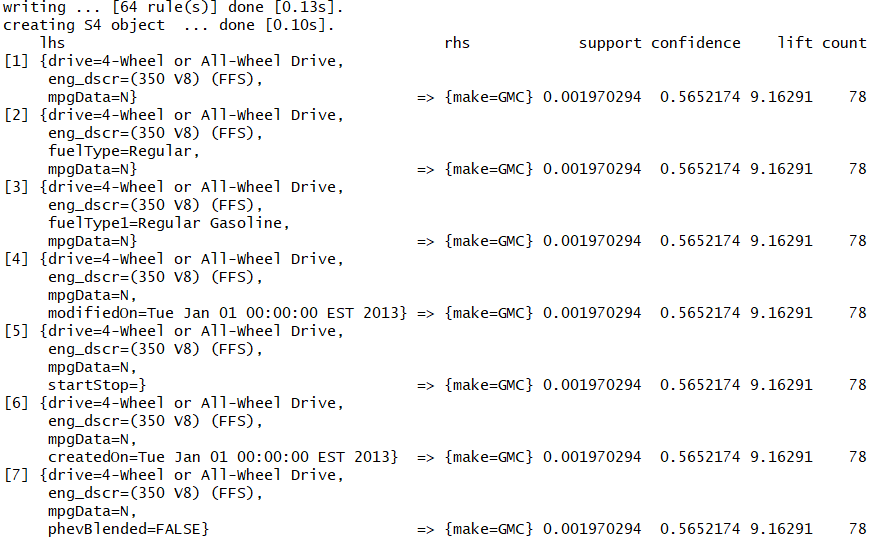


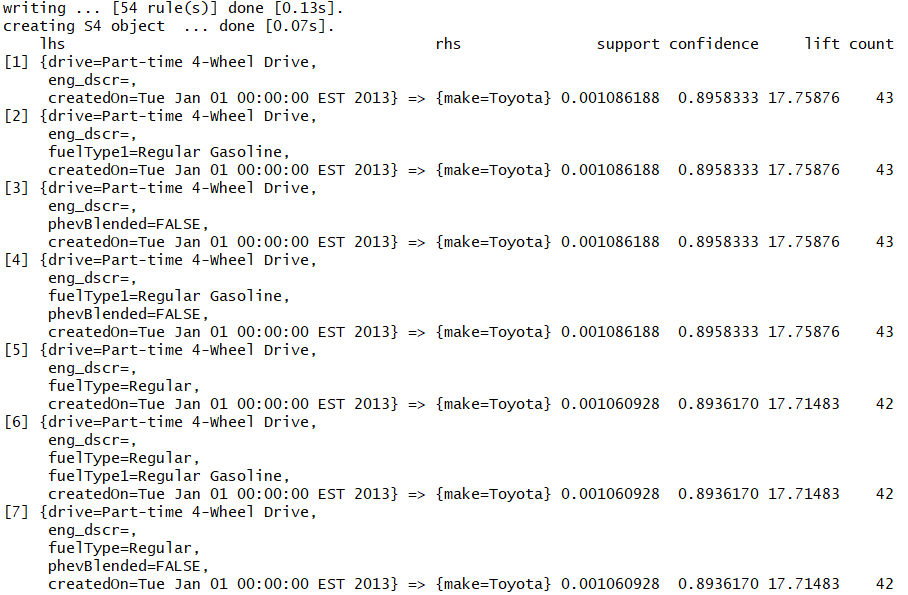
Chevrolet, Ford, Dodge, GMC and Toyota were the 5 types of make considered for the association.











From the output above, just the first 7 of each of the makes were inspected. Also, a fixed support of 0.001 with confidence of 0.55 was used for each make in the apriori for good comparison. From the outputs above, Dodge had the highest rules of 572 while Toyota is with the least rules of 54. Ford had the highest confidence of 0.93, followed by Toyota with a confidence of 0.9 while GMC is with the least confidence of 0.57. Toyota had the highest lift of 17.76 while Chevrolet is with lowest lift of 7.19. The highest count recorded was 78 which was from GMC and the lowest count was from Ford and Toyota

1. **Conclusion**

mpgData (vehichles that responded yes or no to mpgData) was categorical and it was considered as the reponse variable for the classification under supervised learning. From the **Table 1** above, Random Forest was the highest in both ROC, Test Score and Training Score. Hence it can be said as the best model for classifying mpgData respectivley.

barrels08 (annual petroleum consumption in barrels for fuelType1 (1)) was continuous and it was considered as the reponse variable for the regression analysis under supervised learning. From the **Table 2** above,lasso had the lowest test set R^2 of 0.57 while kNeighbors\_15, Random\_Forest, Gradient\_Boosted both had the highest test set R^2 (0.93), hence we can assume any of these as the best model for predicting barrels08 respectivley.

The optimal number of clusters from Silhouette was 2 and MinBatchMeans was able to cluster the dataset into 2 groups with no overlapping. Hence, 2 is the good number of clusters for the dataset.

In conclusion with the association, though Dodge had the highest rules of 572 and Ford had the highest confidence of 0.93 but Toyota had the highest lift of 17.76 and the higher the lift, higher the chance of A and B occurring together. Hence, I would say there is a strong association between the Left-hand side and Right-hand side of the Toyota make respectively based on the first 7 inspected.

1. **References**

“Association Mining With R | arules” (n.d.). Available athttp://r-statistics.co/Association-Mining-With-R.html.

“Vehicle Fuel Economy Data — KAPSARC Data Portal” (n.d.). Available athttps://datasource.kapsarc.org/explore/dataset/us-vehicle-fuel-economy-data-1984-2017/information/?disjunctive.make&disjunctive.model.