PRACTICAL DATA SCIENCE

Assignment 2: DATA MODELLING

Analysing significance of current method of segmentation of cars in a showroom

26 May 2018

Akshay Sharma

RMIT University

Vishesh Jain

RMIT University

Author's Contact:
Akshay Sharma, s3669332@student.rmit.edu.au;
Vishesh Jain, s3666202@student.rmit.edu.au

Table of Contents

1. ABSTRACT

In this report, a hypothesis is tested, that the cars cannot be grouped based only on a single parameter such as 'number of cylinders' of a car.

K-Means, DBSCAN and K-modes, unsupervised machine learning clustering algorithms are used to find the relevant clusters in the data based on all the variables. The idea to use clustering algorithms is directly associated with the hypothesis, hence, number of groups/ clusters are found using these algorithms.

Based on the model evaluation, K-means is found to be the best fit model with the precision of 0.88 and F1-Score of 0.73. While the 'cylinder' has 5 levels, this model can identify four clusters in the data-set. Clear clusters can be seen for 6 & 8-cylinder cars, while there is confusion for 4-cylinder cars and clusters 0 & 2 both point towards it. As there are less number of data points for 3 & 5-cylinder cars, no cluster clearly indicate towards these cars.

Although, there was uneven distribution was found for 'cylinder' variable, through the pair-plot of K-Means model, even distribution in the clusters can be found.

This shows that the hypothesis is true, and the cars cannot be grouped based on a single parameter such as 'cylinder', and a better recommendation system shall be built considering all the available parameters.

2. INTRODUCTION

In today's world, the recommendation system has become an integral customer handling tool. With increase in design and manufacturing configurations of cars, it is becoming difficult to find similar cars using attributes such as engine displacement, origin, mpg, horsepower etc.

It is important to understand a buyer's needs and find cars according to a user's preference. When we visit a showroom, all cars are generally categorised based on one variable such as price or cylinder and this in turn acts as a filter rather than a grouping parameter.

In the field of machine learning, we have seen analysis of the data set using supervised machine learning algorithms and predict the miles-per-gallon of a car. In the given report, unsupervised machine learning or clustering algorithms have been used to identify the clusters using all the variables in the data set.

In the following report, we will test the ***hypothesis*** that the cars cannot be grouped based only on a single variable 'cylinder'.

To check this hypothesis, 'cylinder' is taken as the target variable and is kept of the data-set while doing model fitting. A confusion matrix and classification report are used to select the best model and present the results of the hypothesis test.

3. METHODOLGY

The dataset is taken from UCI Machine Learning Repository, 'Auto-Mpg Data' , which contains 398 instances, 5 continuous variables, 3 multi-valued discrete variables and 'car name' (unique for each instance).

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

	mpg	cylinders	displacement	horsepower	weight	acceleration	modelYear	origin	carName
0	18.0	8	307.0	130.0	3504.0	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165.0	3693.0	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150.0	3436.0	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150.0	3433.0	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140.0	3449.0	10.5	70	1	ford torino

Based on the hypothesis/ problem statement, 'cylinder' is taken as the target variable and taken out of the data set during model fitting. The values of this target variable will be later used to compare the actual groups and clustering groups predicted by the algorithm.

First, the data is explored using descriptive statistics and visualisations using:

- Histograms & Box Plots: To check the distribution of numerical variables in the data set
- Frequency & Percentage Plots: To check the distribution of categorical variables
- Pair Plot: To visualise the correlation of each numerical variable with each other

Next, in the data modelling stage, we used the following three unsupervised machine learning algorithms (Clustering):

- K-means: All numerical variables are used
- DBSCAN: All numerical variables are used
- K-modes: All numerical variables and categorical variables are used except 'horsepower', as during data exploration we saw that 'horsepower' has very little effect on the 'cylinder' variable

To analyse and evaluate the models above, following methods were used:

- Pair-Plot: To visualise clusters while observing the correlation of each numerical variable with each other
- Principal Component Analysis: The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. (Jolliffe, 2011)
- Confusion Matrix: It is a table that is used to visualise the performance of a model. Each row represents instance of the predicted class while each column represents the instance in an actual class (or vice-versa).
- Precision, Recall & F1-Score: Precision is accuracy of positive predictions; recall is fraction of positives that were correctly identified, and F1-Score is calculated by finding the harmonic mean of precision and recall

4. RESULTS

As per the data-set, we have following variables:

```
mpg             float64
cylinders         int64
displacement    float64
horsepower       object
weight          float64
acceleration    float64
modelYear         int64
origin            int64
carName          object
```

We performed data transformations to handle the mis-match of data types in import. Example. 'horsepower' has been imported as an object while this is a numerical variable. This mis-match can be due to missing values or character values in an integer column.

We also took 'modelYear', 'cylinder' and 'origin' and transformed them into categorical variables.

The new data set has following variables:

```
mpg             float64
cylinders        object
displacement    float64
horsepower      float64
weight          float64
acceleration    float64
modelYear        object
origin           object
carName          object
```

4.1. Descriptive Statistics

Using the describe() function in python, we obtain summarisation of numerical columns:

	mpg	displacement	horsepower	weight	acceleration
count	392.000000	392.000000	392.000000	392.000000	392.000000
mean	23.445918	194.411990	104.469388	2977.584184	15.541327
std	7.805007	104.644004	38.491160	849.402560	2.758864
min	9.000000	68.000000	46.000000	1613.000000	8.000000
25%	17.000000	105.000000	75.000000	2225.250000	13.775000
50%	22.750000	151.000000	93.500000	2803.500000	15.500000
75%	29.000000	275.750000	126.000000	3614.750000	17.025000
max	46.600000	455.000000	230.000000	5140.000000	24.800000

Observations:

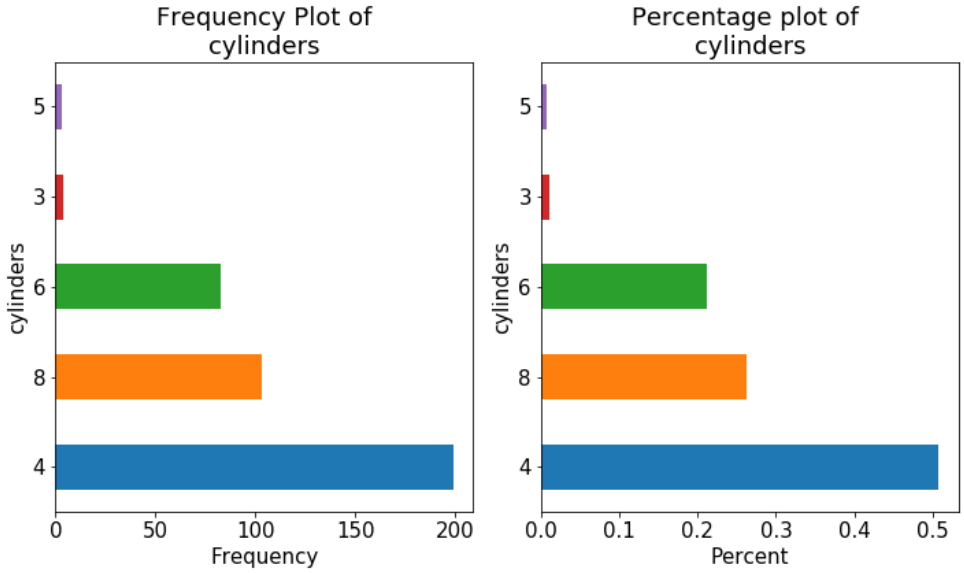
- There are no missing values as the count is same for all variables
- Seeing the interquartile range, there are possible outliers which will be explored using visualisations
- About Standard Deviation (std) in each variable, high variance can be seen. This will be handled through Box-Cox Transformation later in the Data Transformation

4.2. Data Exploration – Visualisation

First, we will look at the distribution of the target variable or the variable we have taken for building the confusion matrix

4.2.1. Univariate Analysis

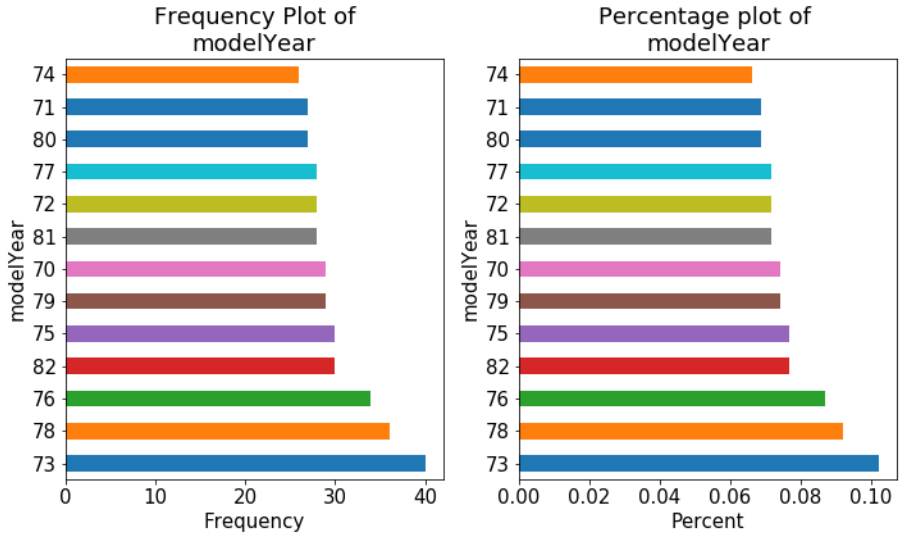
4.2.1.1. Cylinder – Target



Observations:

- Maximum number of instances belong to 4 cylinders
- Number of instances for 3 and 5 cylinders are very less as compared to others
- Number of instances for 6 and 8 cylinders are approximately equal
- The distribution is un-even based on the 'cylinder' of a car

4.2.1.2. Model Year

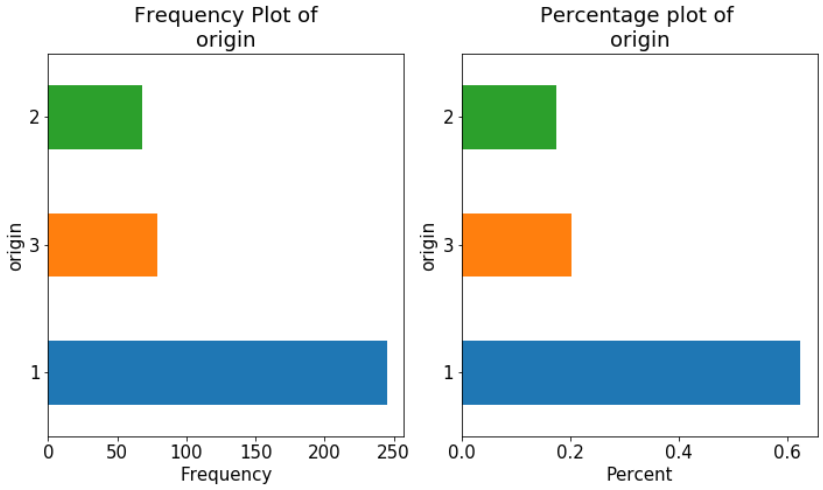


Observations:

- The instances belong to model year from 1974 to 1982
- An even distribution can be seen for all model years

- Maximum number of instances belong to year 1973
- Minimum number of instances belong to 1974

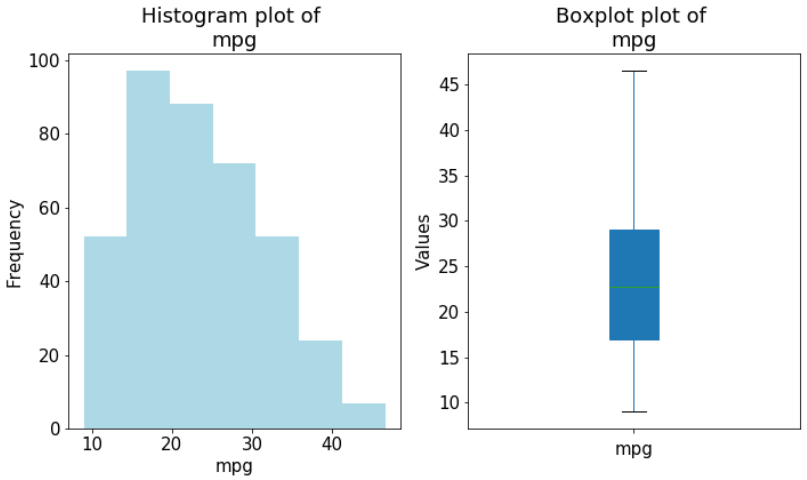
4.2.1.3. Origin



Observations:

- Maximum number of instances belong to 'Origin 1'
- The number of instances from 'Origin 2 & 3' are approximately equal, but are very less when compared to 'Origin 1'

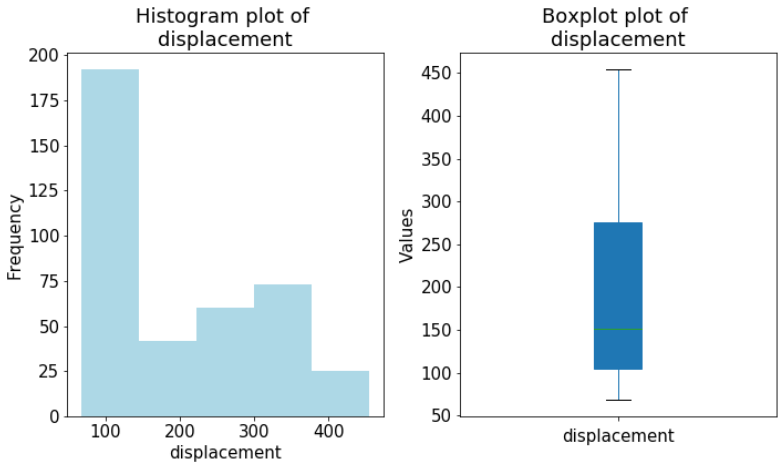
4.2.1.4. MPG – Miles Per Gallon



Observations:

- As per the histogram, the data is right-skewed
- As per Box plot, there are no outliers

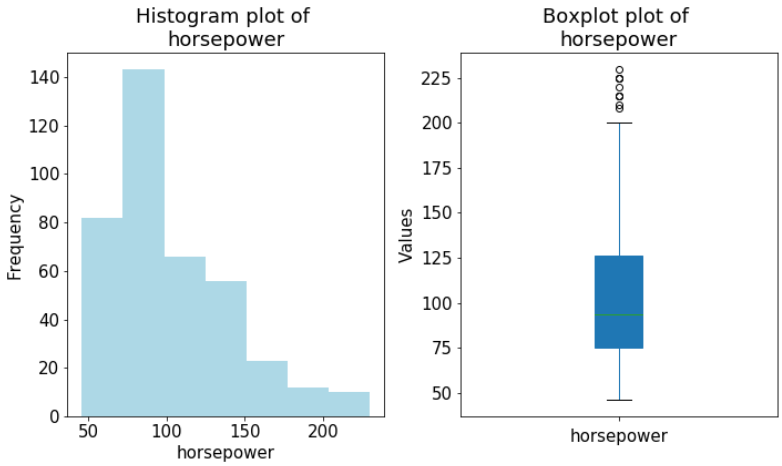
4.2.1.5. Displacement



Observations:

- As per the histogram, the data is clearly right-skewed
- Also, seeing the distribution, we can say that there might be an effect of other variables on displacement
- As per the Box Plot, there are no outliers, but the median is shifted towards a lower value than the mean

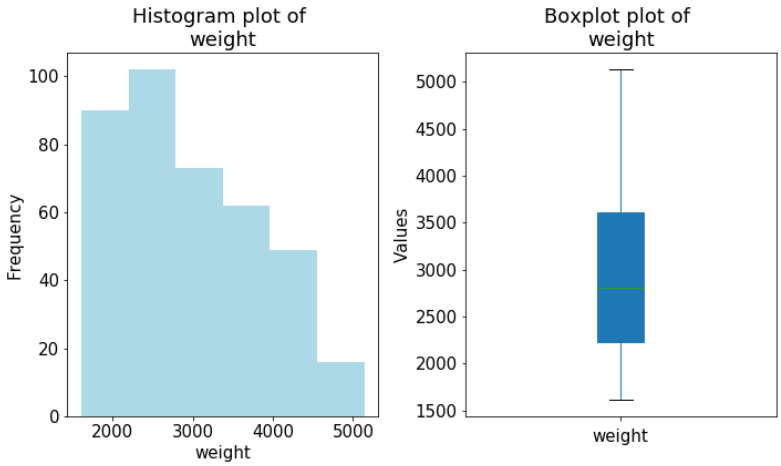
4.2.1.6. Horsepower



Observations:

- As per the histogram, the distribution is clearly right skewed
- As per the Box Plot, there are few outliers beyond horsepower of 200

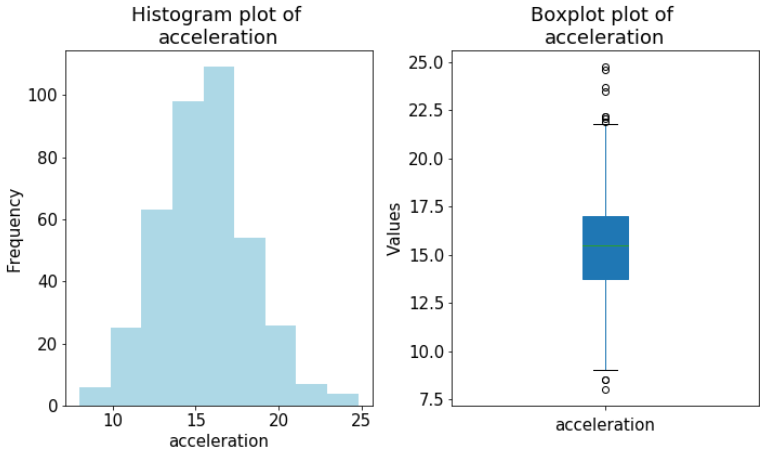
4.2.1.7. Weight



Observations:

- As per the histogram, the distribution of weight is also right-skewed
- As per the boxplot, there are no outliers

Acceleration



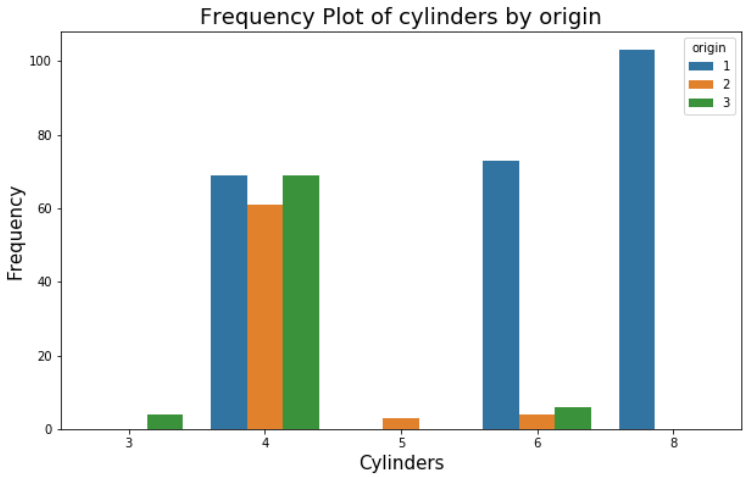
Observations:

- As per the histogram, the data for acceleration is normally distributed
- As per the boxplot, there are many outliers with values more than 22 or less than 9

4.2.2. Multi-Variate Analysis

Here, we will explore the relation between the ‘cylinder’ and other variables

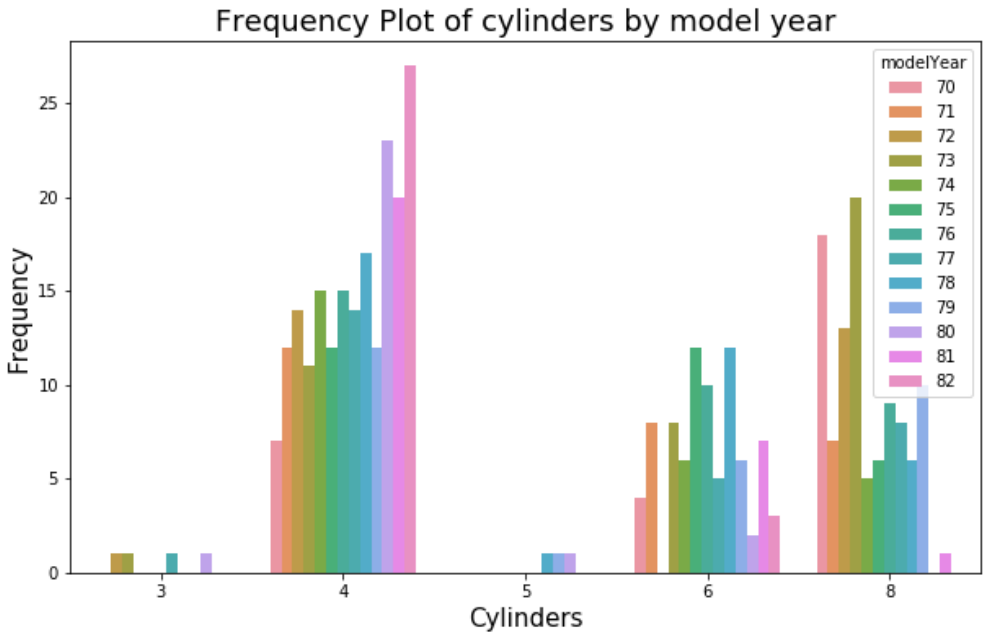
4.2.2.1. Cylinders and Origin



Observations:

- As seen in the univariate analysis, maximum number of instances are of 4-cylinder cars. Therefore, we can see an even distribution of instances from all three origins
- Instances of 8-cylinder cars are only from 'Origin 1'
- Similarly, instances of 3-cylinder and 5-cylinder cars are only from origin 3 and 2 respectively
- Instances of 6-cylinder cars have an uneven distribution. Maximum are from origin 1 and very few are from origin 2 & 3

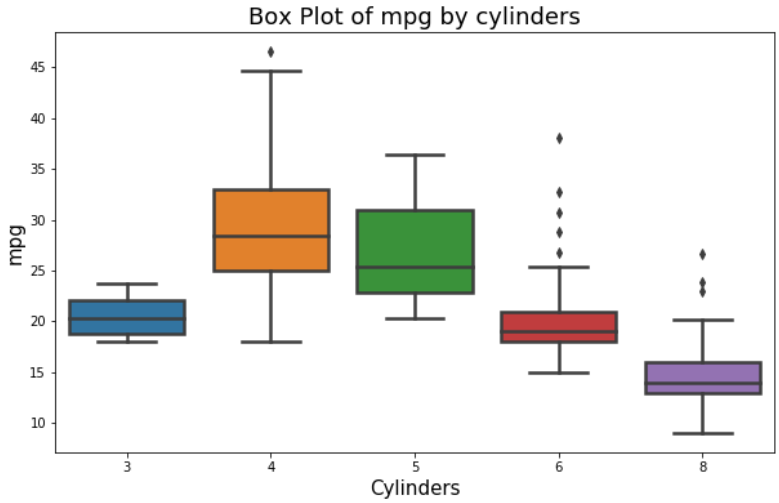
4.2.2.2. Cylinder & Model Year



Observations:

- Instances of 4-cylinder cars are from approximately all model year in the data set
- Instances of 6 & 8-cylinder cars can also be said to belong from all model years
- Instances of 3 & 5-cylinder cars are from very few years

4.2.2.3. Cylinder & MPG

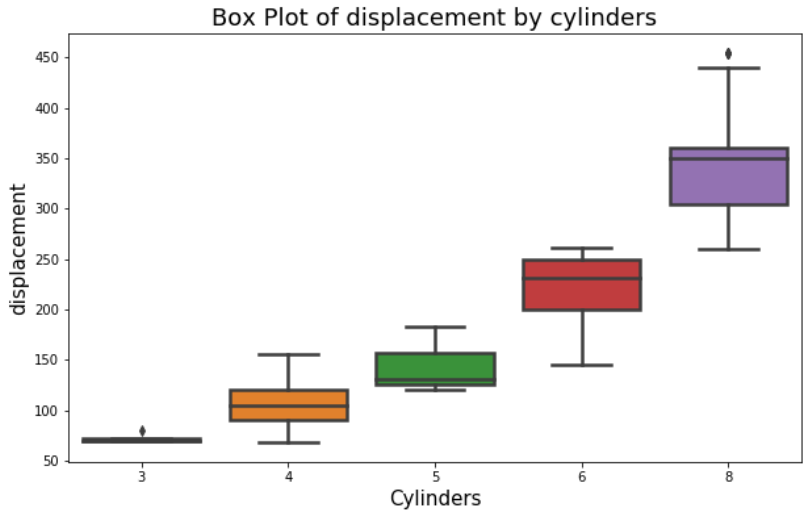


Observations:

- As per the boxplot, 4-cylinder cars have better MPG
- As expected, 8-cylinder cars have minimum MPG
- Although it can be said that with increase in number of cylinders, MPG decreases. But we can also see that 3-cylinder cars have lower MPG than 4-cylinder cars

MPG have a clear effect on the number of cylinders of a car.

4.2.2.4. Cylinder & Displacement

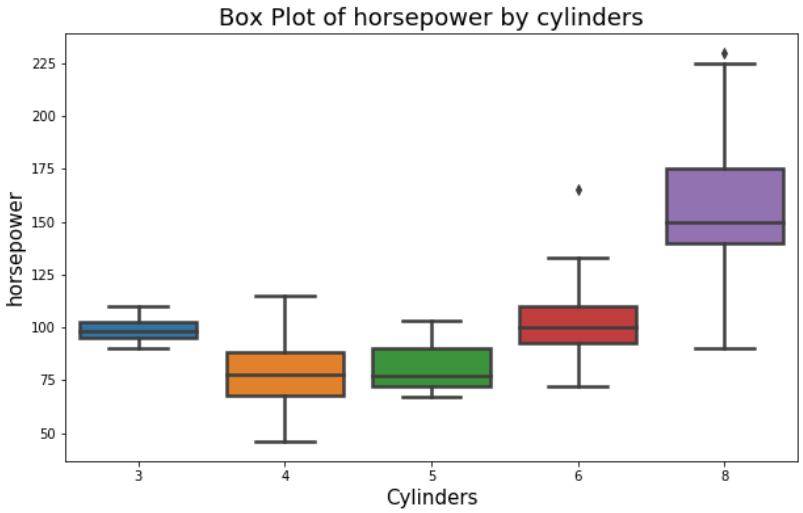


Observations:

- With increase in displacement, increase in number of cylinders can be clearly seen in the plot
- There is an outlier for 8-cylinder car and displacement more than 440

Displacement has a clear effect on the number of cylinders

4.2.2.5. Cylinder & Horsepower

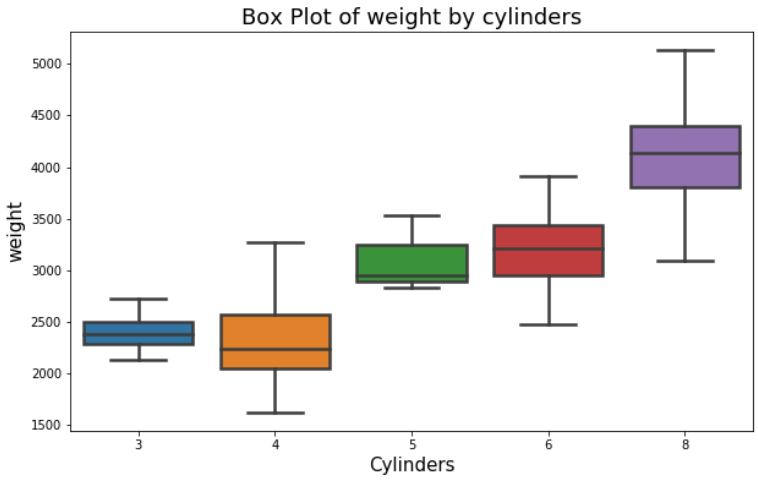


Observations:

- 4 & 5-cylinder cars have equal median values of horsepower
- Although there is a sudden increase in horsepower for 8-cylinder cars, it remains almost similar for other cylinder cars

Horsepower has very little effect on the number of cylinders of a car

4.2.2.6. Cylinder & Weight

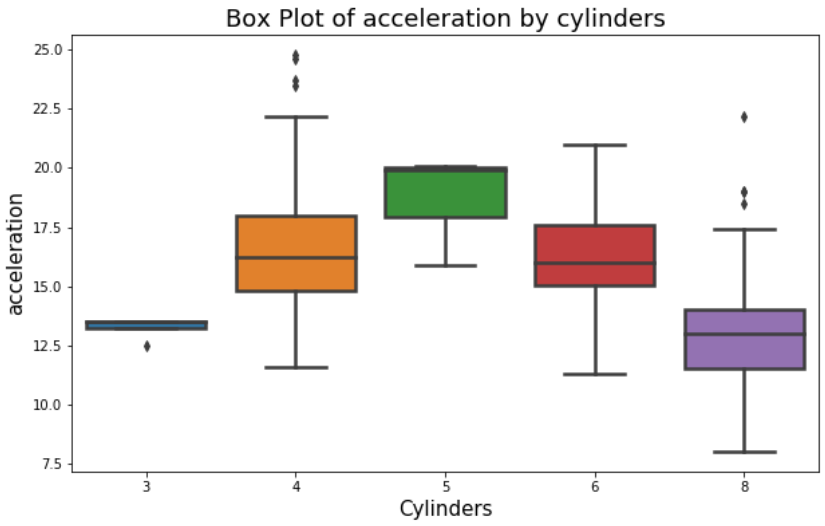


Observations:

- With increase in number of cylinders, the weight of the car also increases. But the same cannot be said for 3-cylinder cars as it does not have minimum median weight
- The distribution of weight in 5-cylinder cars is highly right-skewed

Weight also has an impact on the number of cylinders of a car

4.2.2.7. Cylinder & Acceleration

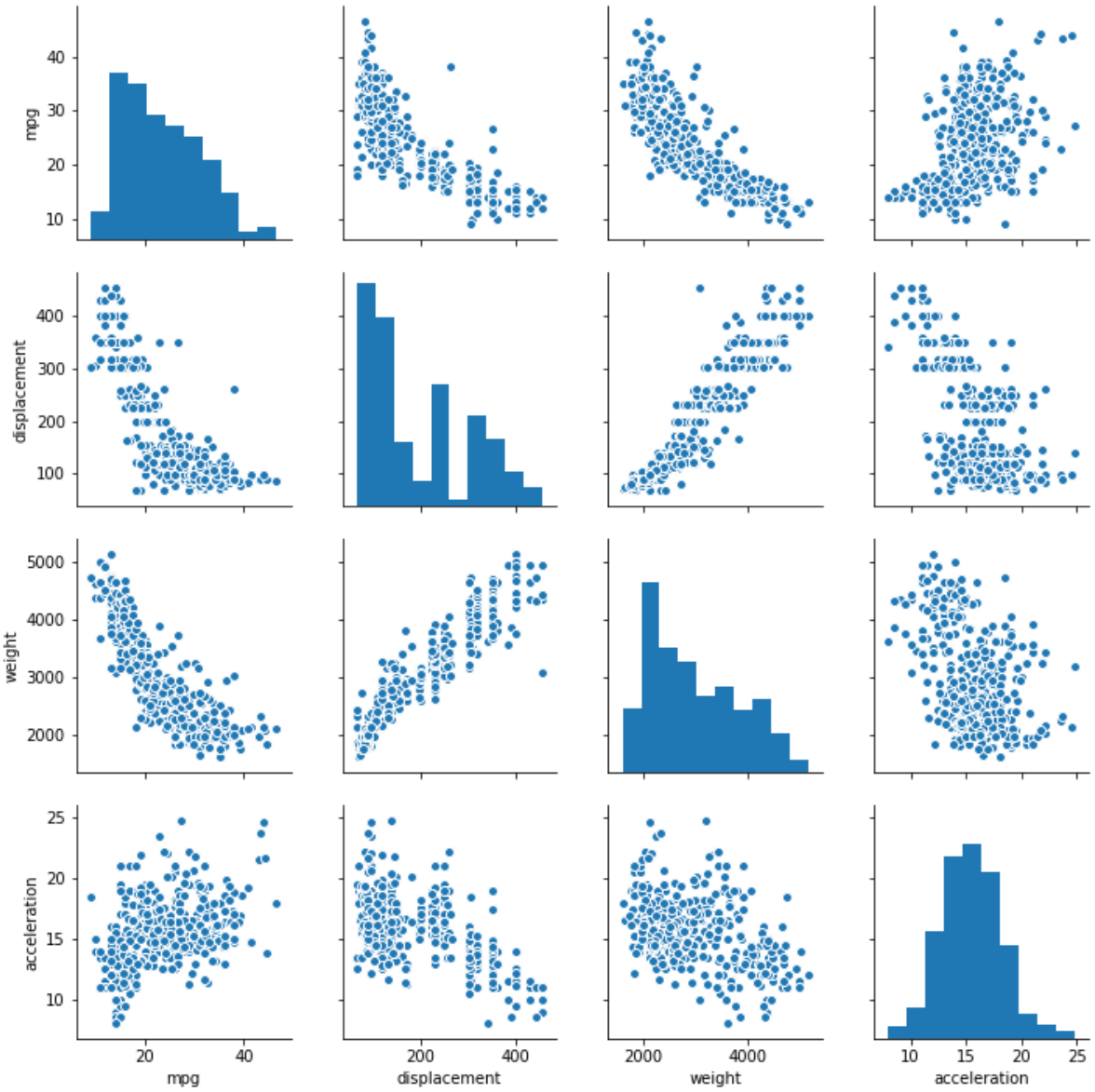


Observations:

- Many outliers can be seen for 4 & 8-cylinder cars
- The number of observation for acceleration in 3-cylinder cars is very less
- 5-cylinder cars have heavily skewed distribution for weight. Also, these cars have the maximum median acceleration
- The 4 & 6-cylinder cars have approximately equal acceleration

4.2.3. Pair-Plot

Using the pair plot, we will explore the correlation between all numerical variables.



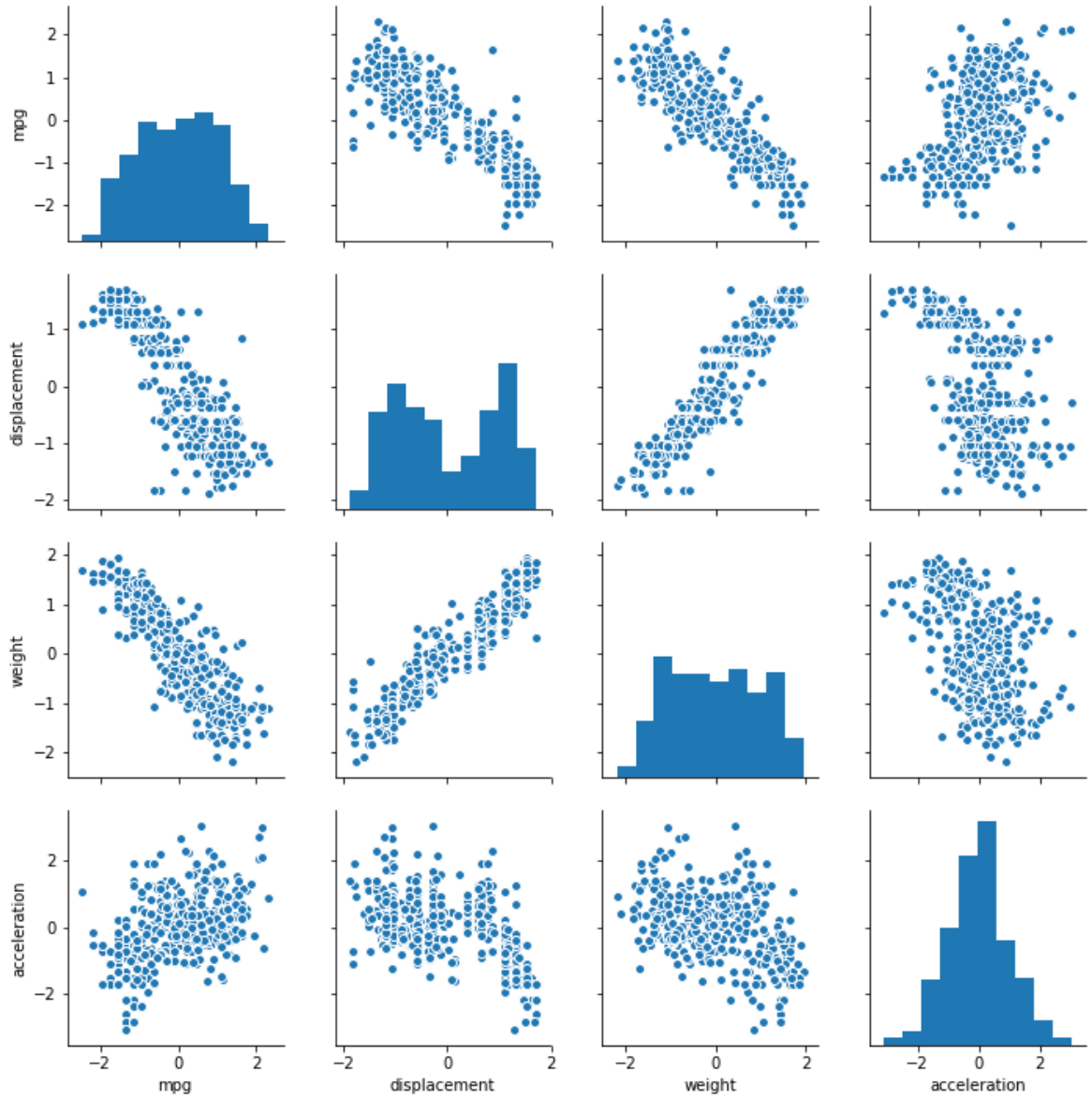
Observations:

- High positive correlation between weight and displacement
- Negative correlation between weight and MPG. Also, the negative correlation can be seen between displacement and MPG
- Acceleration has very weak correlation with other numerical variables

4.2.4. Data Transformation

As we saw earlier, we have skewed distribution for all numerical variables except acceleration. Therefore, we will apply Box-Cox Transformation to achieve approximately normal distribution and scaling is performed to transform data into equal scales, this can be seen below in pair-plot.

4.2.5. Transformed – Pair Plot



Observations:

- Symmetric or normal distribution can be seen for all numerical variables.
- Although displacement has a symmetric distribution, it is also bi-modal, which can be due to a categorical variable
- Now, correlations can be clearly seen between all variables except acceleration

Also, looking at the below table of correlation values:

	mpg	displacement	weight	acceleration
mpg	1.000000	-0.852084	-0.868520	0.452553
displacement	-0.852084	1.000000	0.942674	-0.494388
weight	-0.868520	0.942674	1.000000	-0.407256
acceleration	0.452553	-0.494388	-0.407256	1.000000

As observed in the pair-plot, acceleration has very weak correlations with other numerical variables. High positive correlation can be seen between weight and displacement and strong negative correlation of the two with MPG.

4.2.6. Data Exploration – Summary

Univariate and Multi-Variate analysis was used to observe that 'horsepower' had very little effect on the 'cylinder', therefore, it is not taken in further analysis.

The Box-Cox transformed series improved the distribution of each variable and presented a clear picture of correlation between each numerical variable. From this, it was found that 'acceleration' has very weak correlation with all other variables.

4.3. Data Modelling

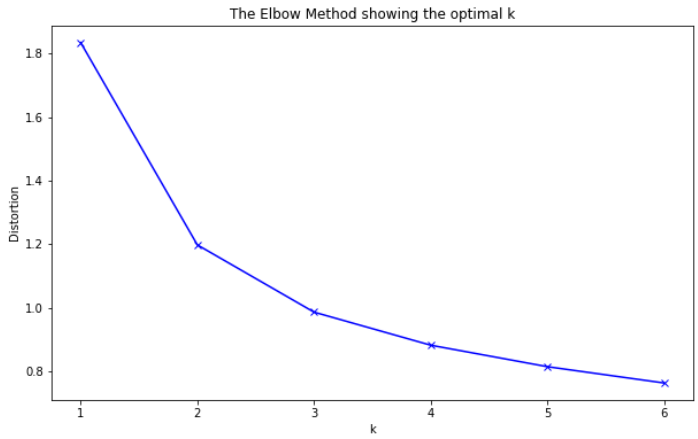
To explore the hypothesis, 'cylinder' variable is removed from the data-set. To obtain the clusters, following clustering models are explored:

4.3.1. K-Means

Based on this algorithm, clusters will be made based on the nearest mean. This model only uses the numerical variables of the dataset.

We will begin with finding the optimal number of clusters in the data-set using elbow & silhouette methods.

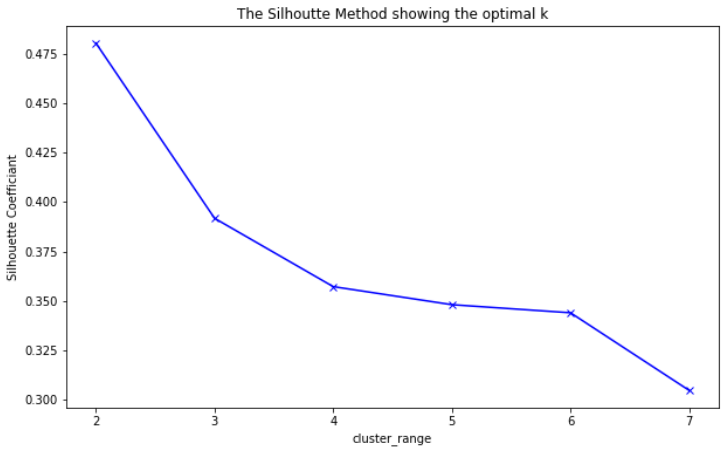
4.3.1.1. Elbow Method



The optimal value of *k* is taken where the fall becomes slower. From the above plot, the value of *k* can be 3 or 4.

Silhouette method is used to confirm the value of k.

4.3.1.2. Silhouette Method



Through this method, the optimal value of *k* is when there is an increase in the coefficient after the fall. From this, the value of *k* can be taken as 4.

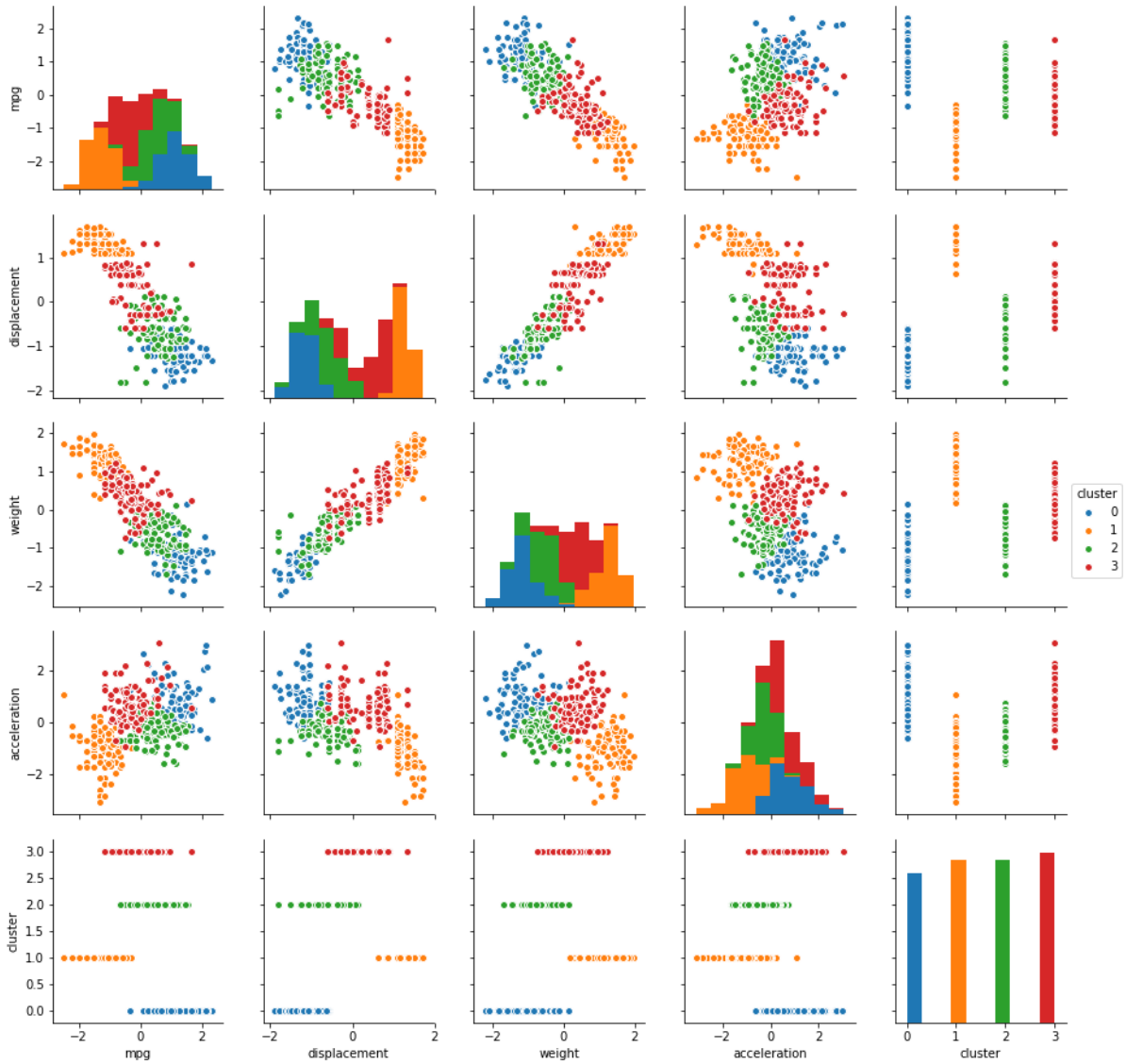
4.3.1.3. Model Fitting

As *k-value* is found 4 through both methods, this will be taken in the model fitting.

Pair plot and PCA will be used to visualise the clusters in transformed and untransformed data-set.

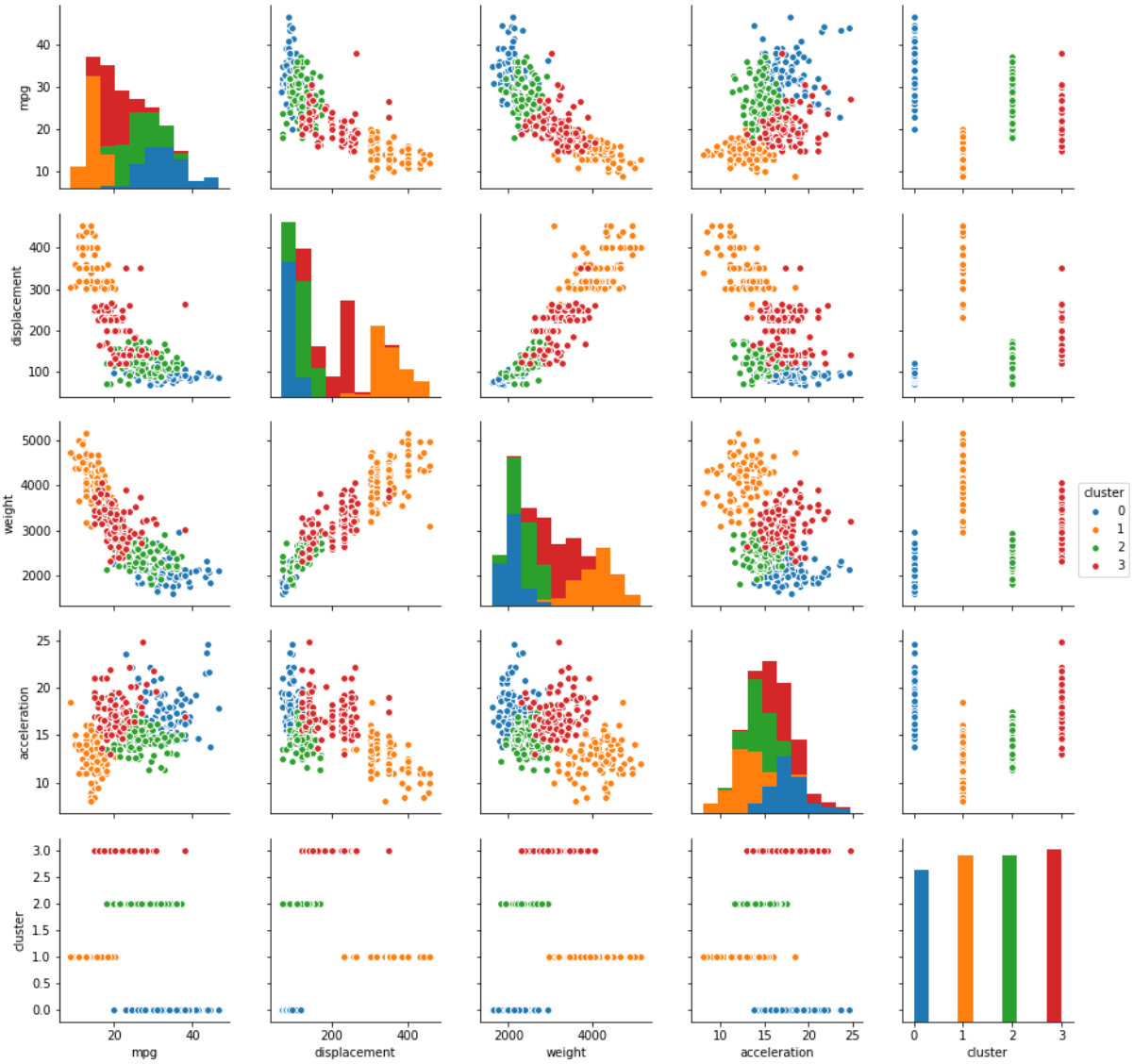
4.3.1.3.1. Pair-Plot:

Transformed Data



An even distribution between the 4-clusters can be observed.

Un-Transformed Data

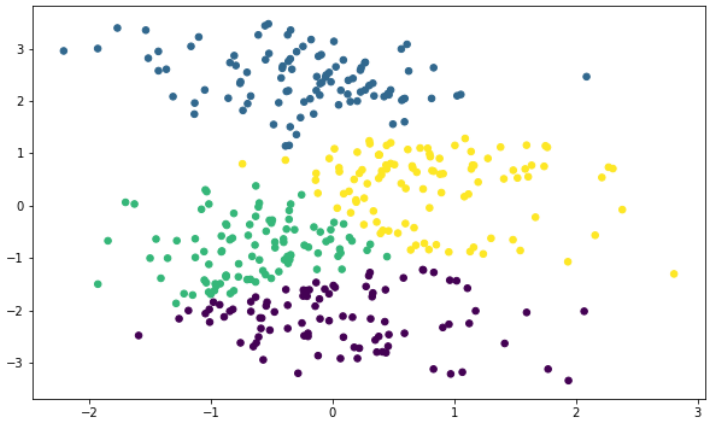


Same as seen in the transformed data pair-plot, although the data was not evenly distributed based on ‘cylinders’, the data in the clusters is approximately evenly distributed.

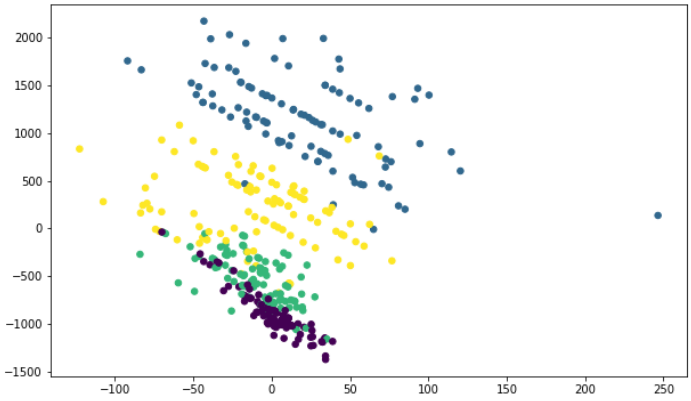
4.3.1.3.2. Principal Component Analysis

This is used to being n-dimensional data in two dimensions to better observe and visualise clusters in the model.

Transformed Data



Untransformed Data



While the clusters overlap in the untransformed data, clusters can be identified clearly in the transformed data.

4.3.2. DBSCAN (Density Based)

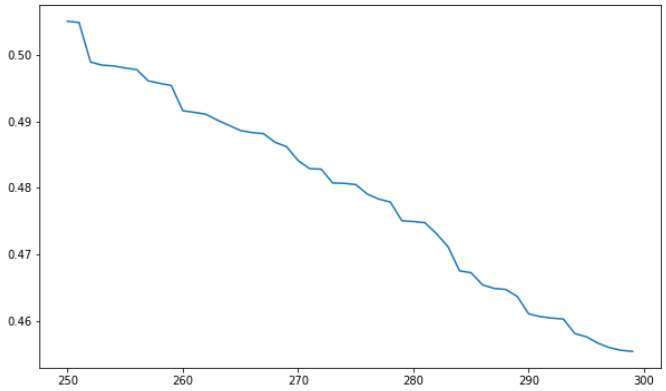
As the name suggests, the clusters are made based on the typical density of points that is considerably higher than outside of the cluster.

One of the important parameter of DBSCAN is *Epsilon* i.e. size of the neighbourhood. To identify an optimal value of the same, *k-nearest neighbour* (Classification Model) will be used to produce a k-distance graph.

4.3.2.1. K-Distance Graph

Good values of Epsilon are where the graph shows a strong bend. Also, DBSCAN is sensitive to this parameter, as choosing a too small value will result in large part of the data not clustered, while too high value will merge most objects in same cluster.

In general, small value are preferred.



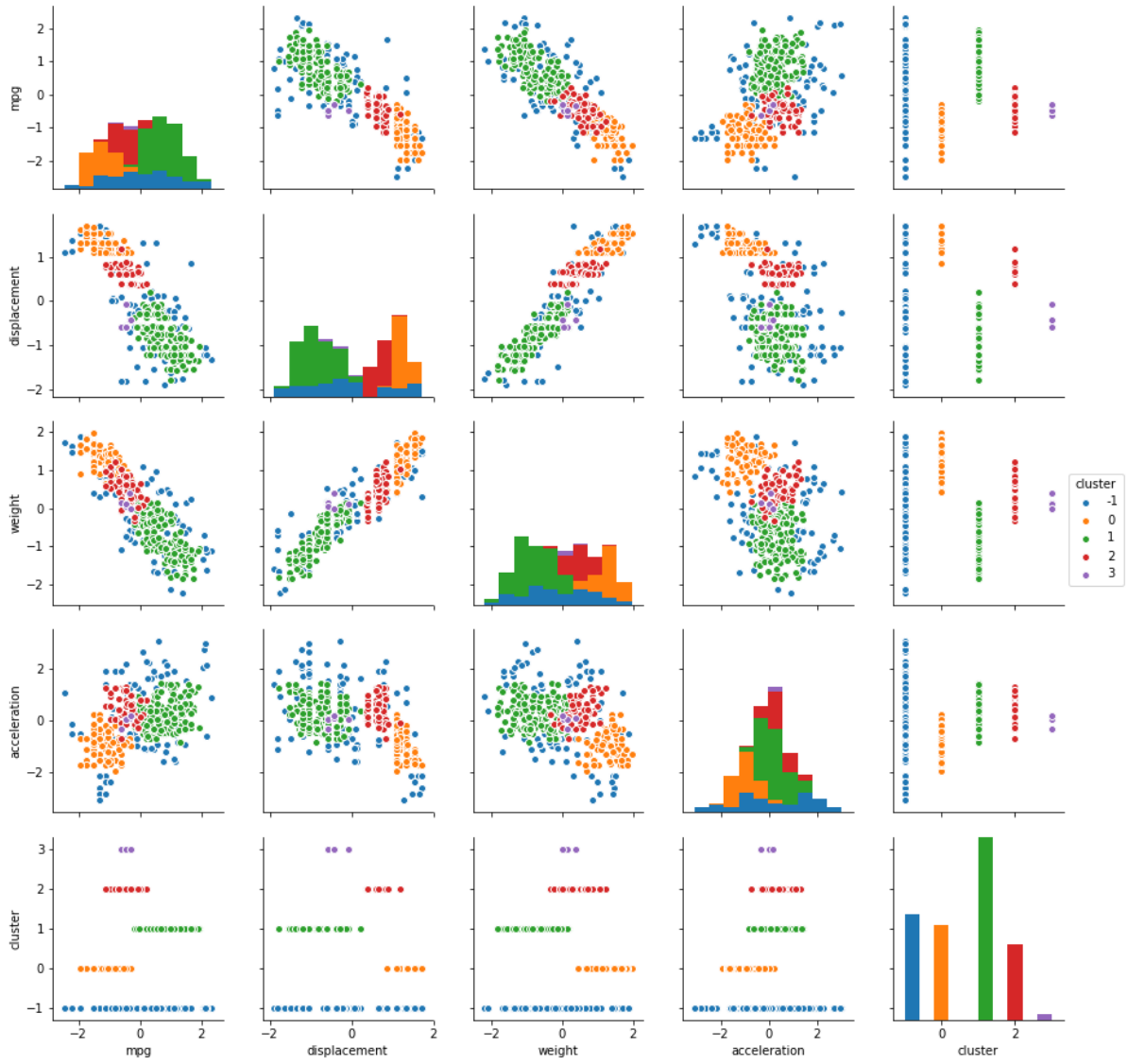
Although, a strong bend can't be clearly seen in the above *k-distance graph,* a considerably good bend can be seen at 0.468. Therefore, we will take *Eps = 0.468* in the model fitting.

4.3.2.2. Model Fitting

Same as K-means, pair plot and principal component analysis is used to visualise the clusters better.

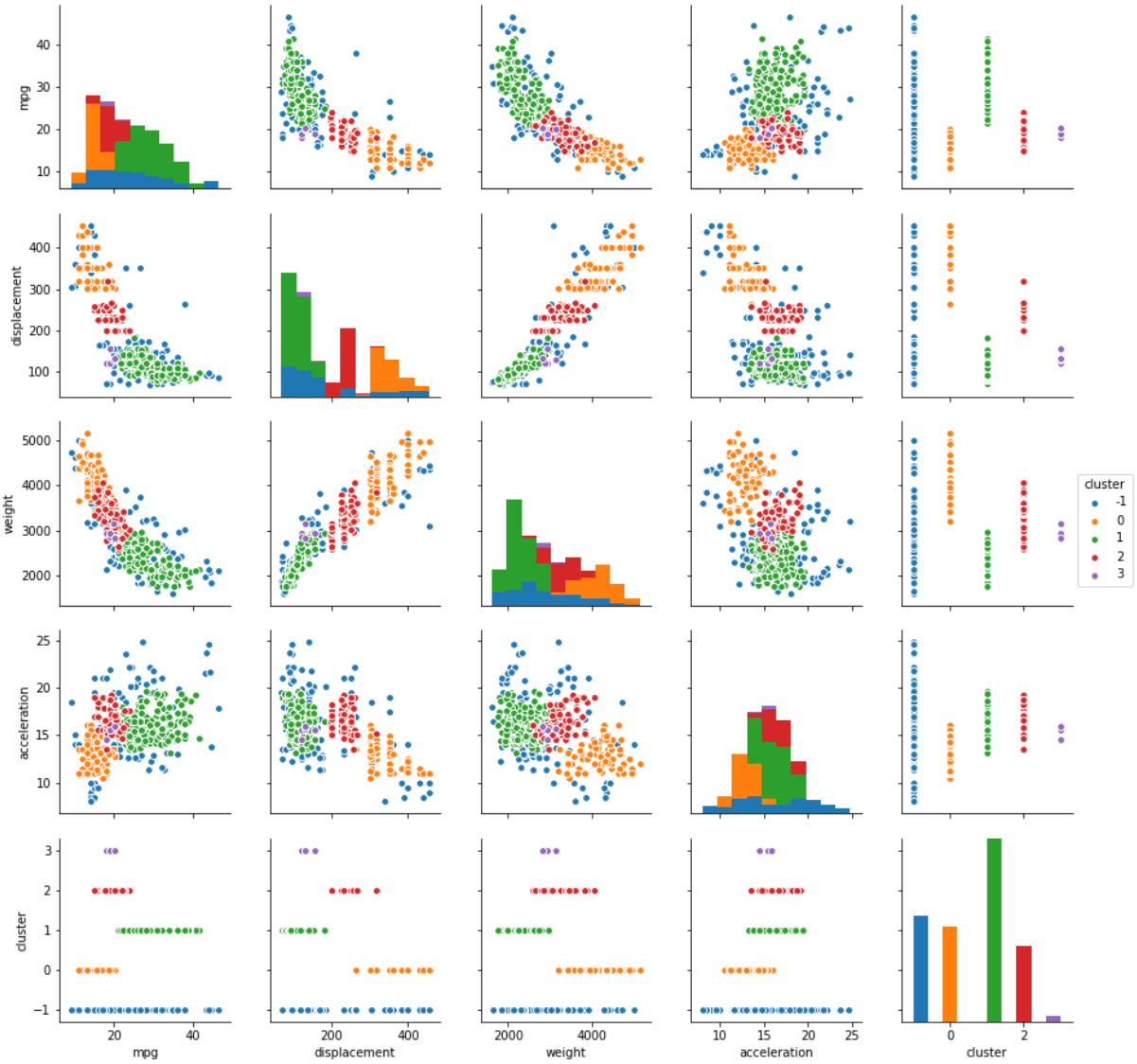
4.3.2.2.1. Pair-Plot

Transformed Data



As per DBSCAN of Transformed Data, five clusters are made, but, there is an uneven distribution in the clusters. There are very few points observed for 'Cluster – 3'.

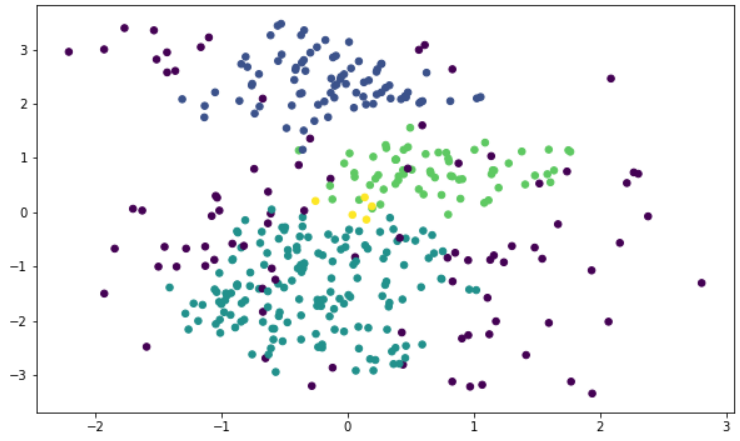
Un-Transformed Data



As seen in the transformed data, same can be observed in the pair-plot of the transformed data set.

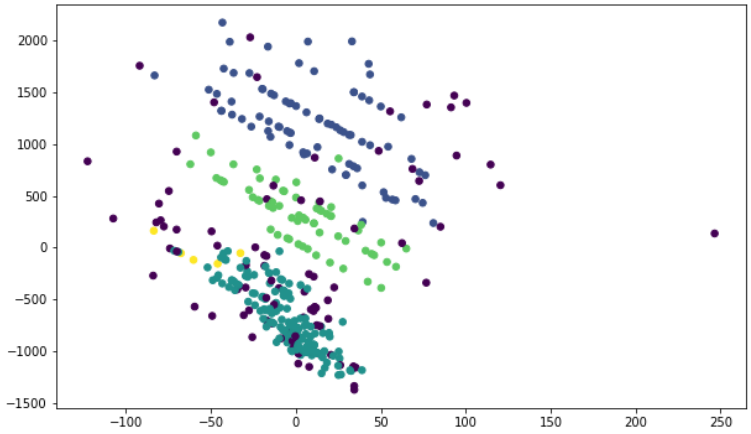
4.3.2.2.2. Principal Component Analysis

Transformed Data



Similar observation can be seen in PCA, the clusters are overlapping and not clear in the transformed data set.

Un-Transformed Data



PCA of the un-transformed data gives shows similar behaviour to PCA of the transformed data.

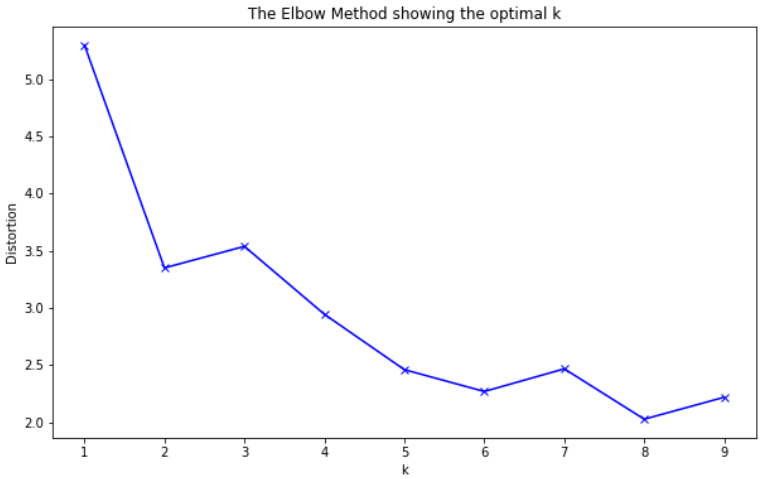
4.3.3. K-Modes

As stated earlier also, k-means and DBSCAN models use only numerical variables for fitting. In contrast, it is frequently needed to identify clusters based on the categorical or nominal scale data.

This procedure is analogous to MacQueen's (1967) K-means clustering procedure. A Drawback of the K-modes (as is true for most clustering procedures) is the lack of statistically valid/reliable indices for choosing the correctio number of clusters. (A.Chaturvedi, 2001).

Like k-means, elbow method will be used to determine the optimal *k.*

4.3.3.1. Elbow Method



In *K-modes,* the optimal value of 'k' is obtained where a strong rise is seen in the distortion value. As seen in the above graph, the same can be seen at 3. Therefore, *k=3* will be taken for model fitting.

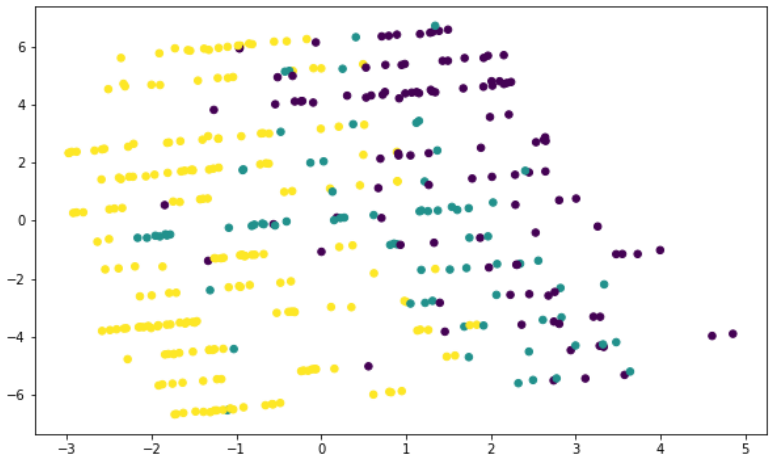
4.3.3.2. Model Fitting

4.3.3.2.1. Pair-Plot

Above pair-plot of the un-transformed data shows 3 clusters with uneven distribution in each. Maximum data points belong to ‘Cluster 2’ and minimum in ‘Cluster 1’.

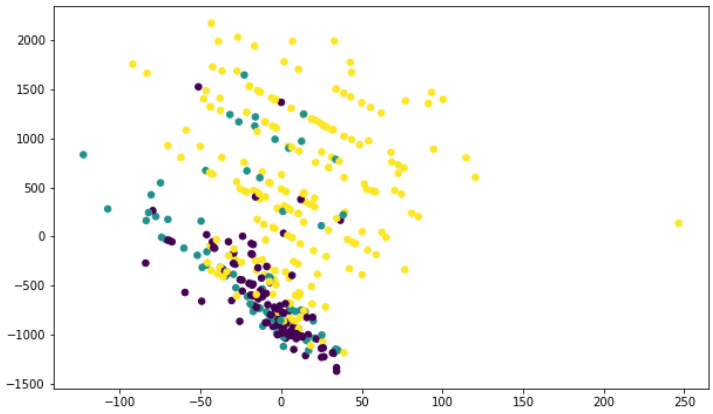
4.3.3.2.2. Principal Component Analysis

Transformed Data



As observed in the pair-plot also, the clusters are not evenly distributed and are overlapping.

Un-Transformed Data



As expected after observing the pair plot and PCA of transformed data, the clusters are overlapping and un-evenly distributed.

5. DISCUSSION

5.1. K-Means

- This model only uses numerical variables of the data-set and model clusters based on the distance between data points
- Effect size of categorical variables remain unknown which results in discrepancies in the cluster
- Maximum number of data points from different clusters are found in 4-cylinder cars, this is due to the availability of more data points in this category as compared to others. Therefore, there is bias present in the data.
- It is also known that when 'degree of freedom' increases, effect size of the variable also increases. This effect is more prevalent on 4-cylinder cars.
- These effects are not present in 8-cylinder cars and therefore, a clear cluster can be seen

Analysing different effects in the model, it can be said that the model suggests 4 clusters in a 5-level of cylinders data set.

5.2. DBSCAN

- This model also uses only numerical variables of the data set
- As it is density based, scaling of all the numerical variables through Box-Cox Transformation helped in making same weightage for all variables. After scaling, we are only dealing with variance of each variable
- Observing all factors together, the maximum relation is because of correlated factors. This could give us origin of a car as a better output of a car, which is in-line with the assumption that cars with same origin will be similar, as they would follow the same guidelines of a geographical region

DBSCAN suggests 5 clusters equal to the number of levels of cylinders in the dataset.

5.3. K-Modes

- This clustering model uses both numerical and categorical variables in model fitting
- The correlation and difference matrix of each categorical variable varies with other parameters
- Ideally, this model should give us better results as this is uses categorical variables as conditions and numerical variables as predictors.

K-modes suggests only 3 clusters, this can be due to less number of data points for 3 & 5-cylinder cars.

Further, we will explore and discuss the parameters used to evaluate the given models and then select the best fit clustering model.

5.4. Model Evaluation

Below table shows sample data set with the name of the clusters from all three candidate models.

	mpg	cylinders	displacement	horsepower	weight	acceleration	modelYear	origin	carName	kmeans	dbscan	kmodes
0	18.0	8	307.0	130.0	3504.0	12.0	70	1	chevrolet chevelle malibu	1	0	2
1	15.0	8	350.0	165.0	3693.0	11.5	70	1	buick skylark 320	1	0	2
2	18.0	8	318.0	150.0	3436.0	11.0	70	1	plymouth satellite	1	0	2
3	16.0	8	304.0	150.0	3433.0	12.0	70	1	amc rebel sst	1	0	2
4	17.0	8	302.0	140.0	3449.0	10.5	70	1	ford torino	1	0	2

As per the sample data table above, all three models can group the 8-cylinder cars under one cluster.

For evaluation, we will compare the cluster label with the 'cylinder' labels and create classification report and confusion matrix for all three models.

5.4.1. Classification Report

This report is used to check three parameters of a model:

- Precision: Accuracy of positive predictions
- Recall: Fraction of positives that were correctly identified
- F1-Score: It considers precision and recall. It is created by finding the harmonic mean of precision and recall

Parameter/Models	Precision	Recall	F1
K-Means	0.88	0.67	0.73
DBSCAN	0.80	0.50	0.58
K-Modes	0.59	0.40	0.46

Using the above table, we will select the model based on the F1-Score.

- K-Means and DBSCAN shows high precision of greater than 0.80, with precision of K-Means as 0.88
- Considering recall, only k-means shows a good fraction of positives
- As f1-score considers both precision and recall, k-means is chosen to be the best clustering model based on this, **F1-Score (K-Means) = 0.73**

5.4.2. Confusion Matrix

This matrix is used to evaluate or check the performance of each candidate model. It shows the number of points in each cluster as predicted by the model with actual labels.

cylinders / kmeans	3	4	5	6	8
0	0	90	1	0	0
1	0	0	0	2	97
2	4	86	0	9	0
3	0	23	2	72	6

cylinders / dbscan	3	4	5	6	8
-1	4	44	2	20	19
0	0	0	0	0	80
1	0	152	0	2	0
2	0	0	0	60	4
3	0	3	1	1	0

cylinders / kmodes	3	4	5	6	8
0	4	134	3	15	2
1	0	62	0	64	66
2	0	3	0	4	35

Considering K-Means as the best model based on F1-Score, following observations can be seen through its confusion matrix:

- Although we have 5 levels in the data set for 'cylinders', but K-Means suggests four clusters based on the optimal k-value through 'Elbow Method'. This indicates that at least two groups of cylinders belong to the same cluster
- K-means, although a good model, it is not able to cluster the 4-cylinder cars. It shows that 'Cluster 0 & 2' both have data points from the same cylinder category
- The behaviour of the model can be due to very less number of data points for 3 & 5-cylinder cars

With the above discussion, we can say that the hypothesis of grouping different cars on one parameter might not be the right approach and this grouping shall be based on other parameters as well.

6. CONCLUSION

Based on the results and discussion, we found that K-means is the best fit model. This model suggests 4 clusters for our data set. Based on the confusion matrix, it is observed that clusters of 6 & 8-cylinder cars are clearly identified by the model. Cluster 0 & 2 both suggested the grouping for 4-cylinder cars and there was no clear identification for 3 & 5-cylinder cars, which can be due to low number of data points in both categories.

Through unsupervised machine learning clustering algorithms, we can conclude that as per the popular practice, cars cannot be grouped based on only one variable, 'cylinder' being checked here. Although, as per 'section 4.2.1.1', we have uneven distribution based on the 'cylinder' of the car, we can see even distribution in the pair-plot of the K-Means clustering model, 'section 4.3.1.3.1'.

Through this analysis, the hypothesis tested is true and, it can be said that for a better identification of a customer's preference, cars should be grouped based on more than one parameter like 'cylinder' and through unsupervised machine learning, a recommendation system can be built for the customers.

This can be further extended to test this hypothesis on 'price' as well and if the same conclusion is drawn, the recommendation of vehicles shall be based on more than one parameter.

7. REFERENCES

A.Chaturvedi, P. G. J. C., 2001. K-modes Clustering. *Journal of Classification,* pp. 35-55.

Jolliffe, 2011. *Principal Component Analysis.* 2nd ed. New York: Springer, Berlin, Heidelberg.