**INFORMATICS INSTITUTE OF TECHNOLOGY**

**IN COLLABORATION WITH**

**UNIVERSITY OF WESTMINSTER (UOW)**

**B.Eng. (Hons) Software Engineering**

**5DATA001C.2 – Machine Learning and Data Mining**

**Module Leader: Achala Aponso**

Coursework 01

2021

**UOW ID:** w1761265

**Student ID:** 2019281

**Student Full Name**: Mohammed Nazhim Kalam

**CLUSTERING PART**

1. **Discussion of the methodologies used in *reducing the dimensionality.***

* Since, this dataset “**vehicles.xlsx**” has a large number of features or columns (19 features), this is referred to as **a high dimensional dataset**.
* Having a large number of dimensions will learn to the **curse of dimensionality**.
* **Curse of dimensionality** means that’s as the number of dimensions (features/columns) of the dataset increases the points go further apart or data becomes extremely sparse leading the accuracy to decrease.
* To solve this problem only we have to perform **dimensionality reduction**.
* **PCA** (Principal Component Analysis) is the most popular technique which can be used for dimensionality reduction, PCA helps to identify all those high correlated variables, which are not related to the target variable at all and drop them out from further analysis.
* Therefore, using PCA we can convert **a high dimension data into low dimension** without loosing any of the important features which are correlated to the target class.

1. **What is PCA?**

* **PCA** or also known as Principal Component Analysis, is an unsupervised learning algorithm that is used specifically for dimensionality reduction in machine learning.
* When the PCA algorithm is applied it produces something called as the **Principal Components**.
* The aim of the PCA algorithm is to **lower-dimensions** from a higher dimension but still retain the quality of the data.

1. **What is Scaling and why do we need them?**

* Scaling or feature scaling is a technique which is used to standardize the data into a fixed range scale. In other words bring all the data which belonged to different scales into a single common unique scale.
* If we don’t standardize or scale our data, then we wont be getting much of a better result or a better accuracy, because all the features contributes in a different proportion not equally.

Eg:- Price of the house sale with time, we know in general price of the house increase with time but time and price of house are of different scale and the difference between the values of the price and time are quiet large hence data needs to be scaled down

1. **What are outliers?**

* Data points or data values which are completely out of the range from what its expected to be in are called outliers. In other words data collected due to some fault or error which makes it to fall way out of the range of data expected.
* Example, lets consider heights of student and there are 2 outliers present here which is 1ft and 8ft. (We know that in general there are no students with the height of 1ft or 8ft so these are outliers)

Heights: **1ft**, 5ft, 5.1ft, 4.9ft, 5.7ft, 5.9ft, **8ft**

1. **Why do we need to remove outliers from our dataset?**

* The presence of an outlier can affect the accuracy of the model we create. This is because the data is not clean. With clean data only we can a better accuracy model.
* Hence, we have to remove outliers when working with the data.

1. **How can we find out if there are outliers in our dataset?**

* By drawing up a box plot for each feature of your dataset, you will be able to find data points which go above the maximum range and data points going below the minimum range and these are the outliers and has to be removed.

1. **Explain how the Order of scaling and outlier removal is important?**

* We have to first remove the outliers from the dataset before scaling the dataset.
* If we don’t perform outlier removal before normalizing the data, then the dataset won’t be efficient enough for the modal to predict, this is because the resulting data won’t be properly standardized therefore you may end up getting different variables/columns having different standard deviations which is a problem.

1. **Briefly explain the meaning of:**
2. ***Accuracy***:

* Accuracy is one an evaluation metric type for classification models.
* Accuracy is also defined by the following formula:

**Accuracy** = Total Number of correct predictions / Total Number of predictions

**Accuracy = (True Positives + True Negatives) / Total Number of predictions**

In other words, accuracy represents how close a measurement comes to its true value.

1. ***Precision***:

* Precision is another evaluation metrics which is also referred to as the spread of the measured values.
* Precision is calculated by the following formula:

**Precision** = True Positives / (True Positives + False Positives)

* The result of this equation lies between 0 to 1, 0.0 indicating that there is no precision and 1.0 for full precision.

1. ***Recall***:

* Recall is another evaluation metrics that is used to quantify the number of correct positive predictions made out of all positive predictions that could have been made.
* Unlike precision, Recall provides an indication of missed positive predictions.
* Recall is calculated by the following formula:

Recall = True Positives / (True Positives + False Negatives)

1. **Results and Discussion using the Confusion matrix with respect to the calculation for the accuracy/recall and precision matrices.**

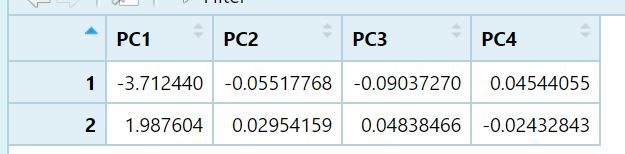
* When I used the automated tools inorder to calculate the number of clusters formed these are the following results which were obtained

**Elbow Method** (Automated) gave me 4 clusters as the best one.

Silhouette Method (Automated) gave me 2 clusters as the best one.

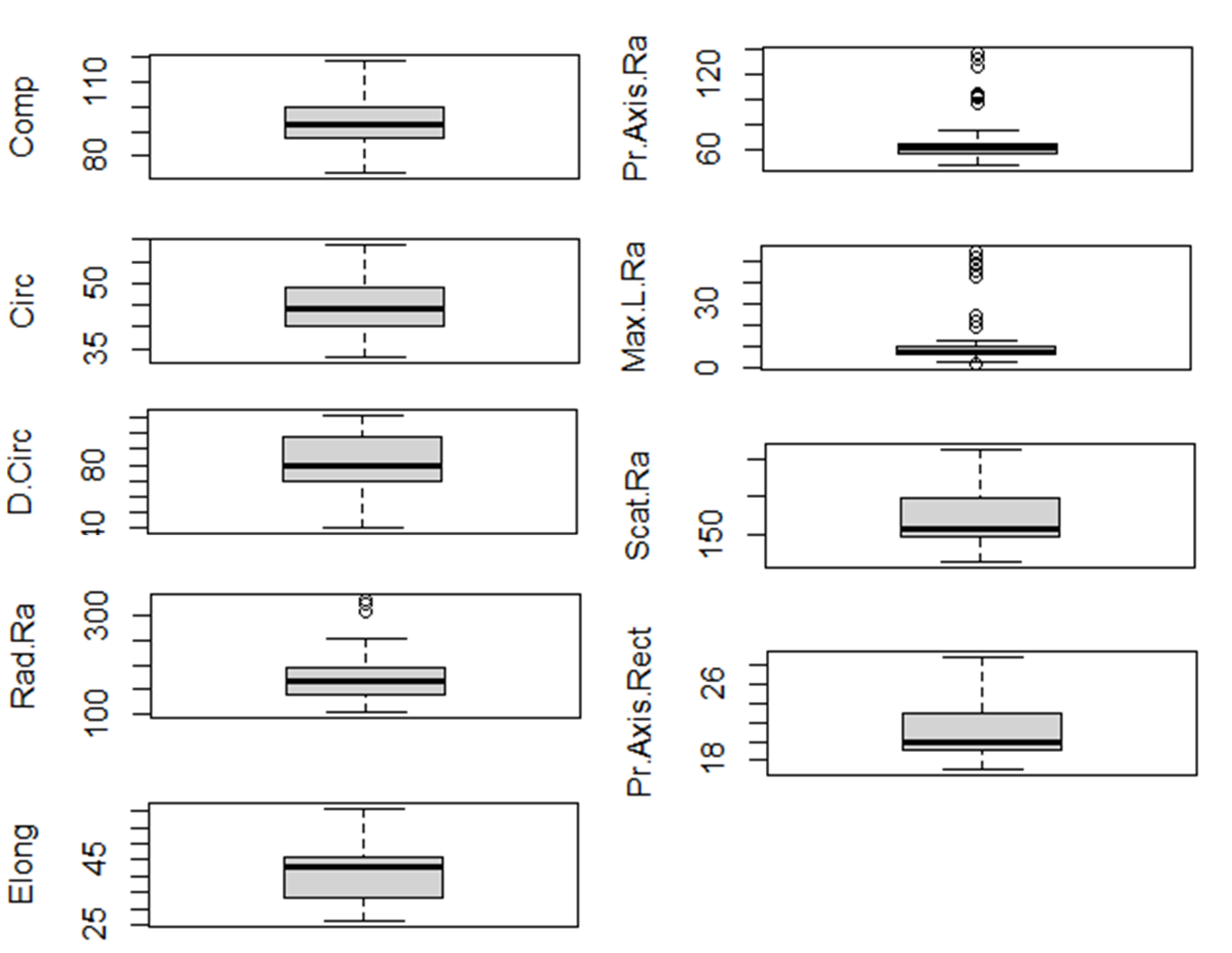
Gap Statistics Method (Automated) gave me 3 clusters as the best one.

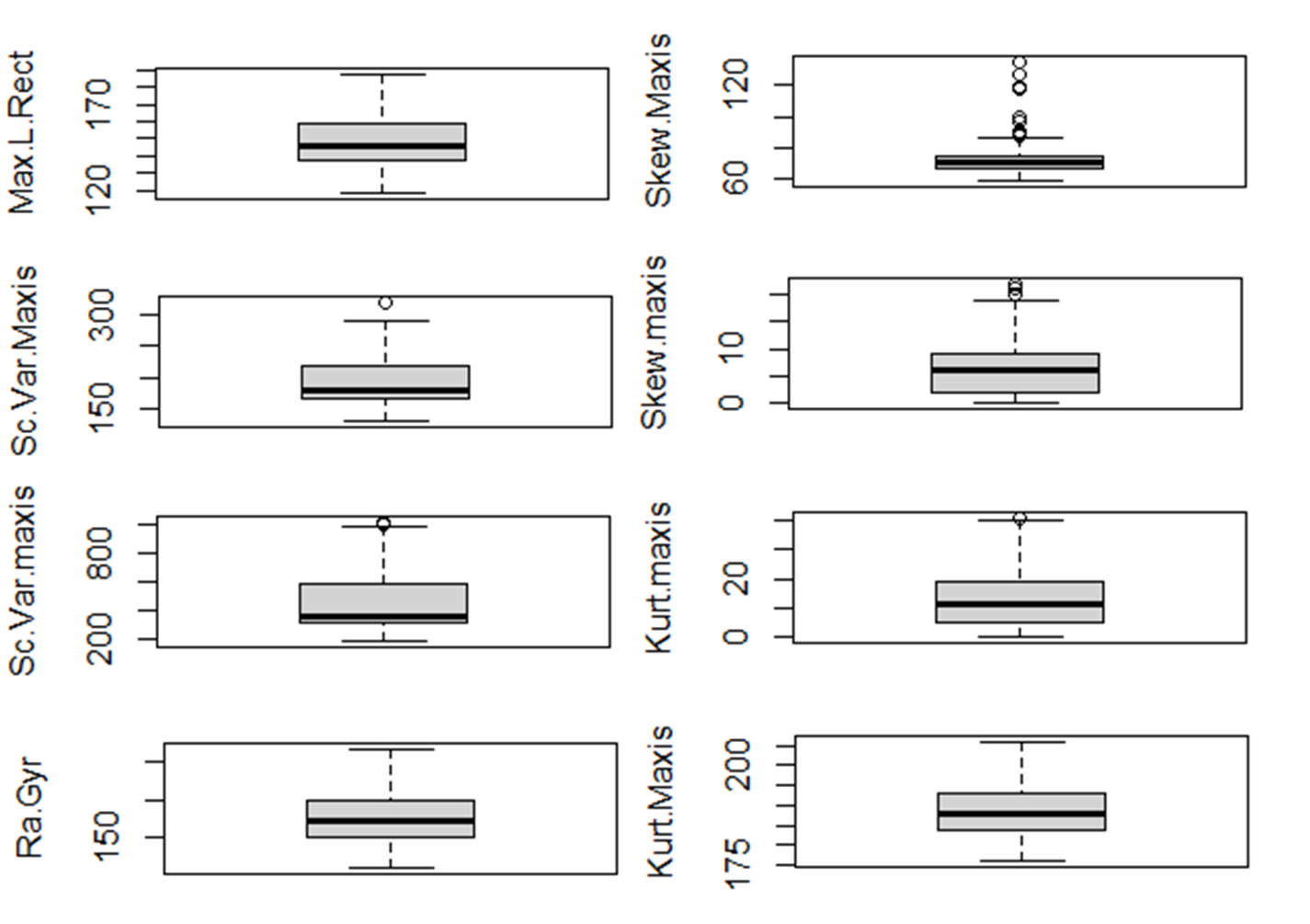
* I now manually created an elbow method to compute the best number of cluster by looping though 1 to 10 number of clusters and performing the KMeans clustering and calculating its accuracy using the confusion matrix. These are the follow results given for the clusters from 1 to 10 with its accuracy, precision and its recall value.
* 1 Cluster (Accuracy: 25.7%)
* 2 Cluster (Accuracy: 32.8%)
* 3 Cluster (Accuracy: 30.0%)
* 4 Cluster (Accuracy: 13.7%)
* 5 Cluster (Accuracy: 23.7%)
* 6 Cluster (Accuracy: 9.9%)
* 7 Cluster (Accuracy: 12.9%)
* 8 Cluster (Accuracy: 12.9%)
* 9 Cluster (Accuracy: 7.3%)
* 10 Cluster (Accuracy: 8.1%)
* Since the 2-Clusters have the highest accuracy I have taken 2 clusters as the optimal number of clusters and proceeded.
* These are the following coordinates of the centers of the clusters.



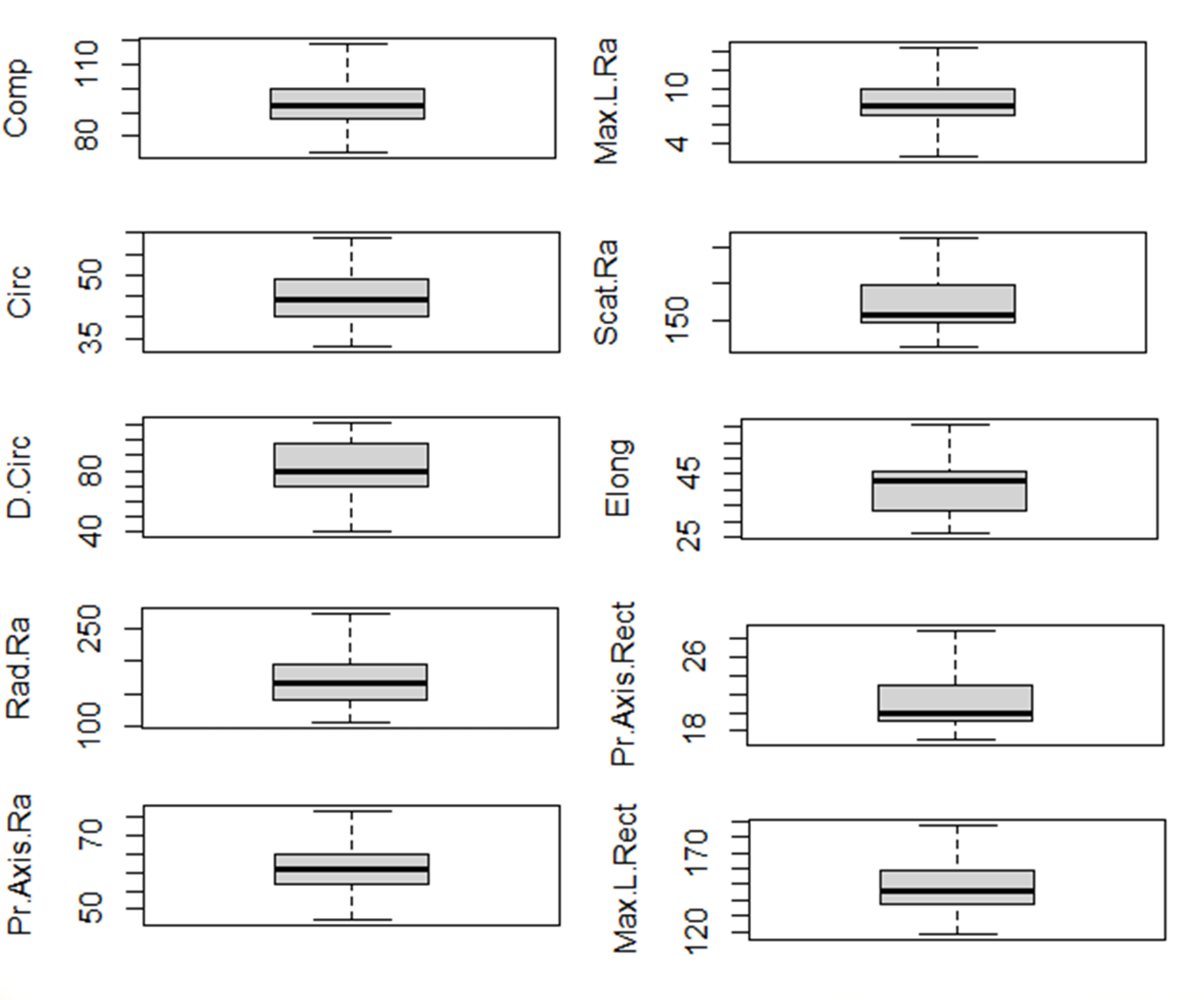
1. **The following are graphs obtained and its respective description about the graph**

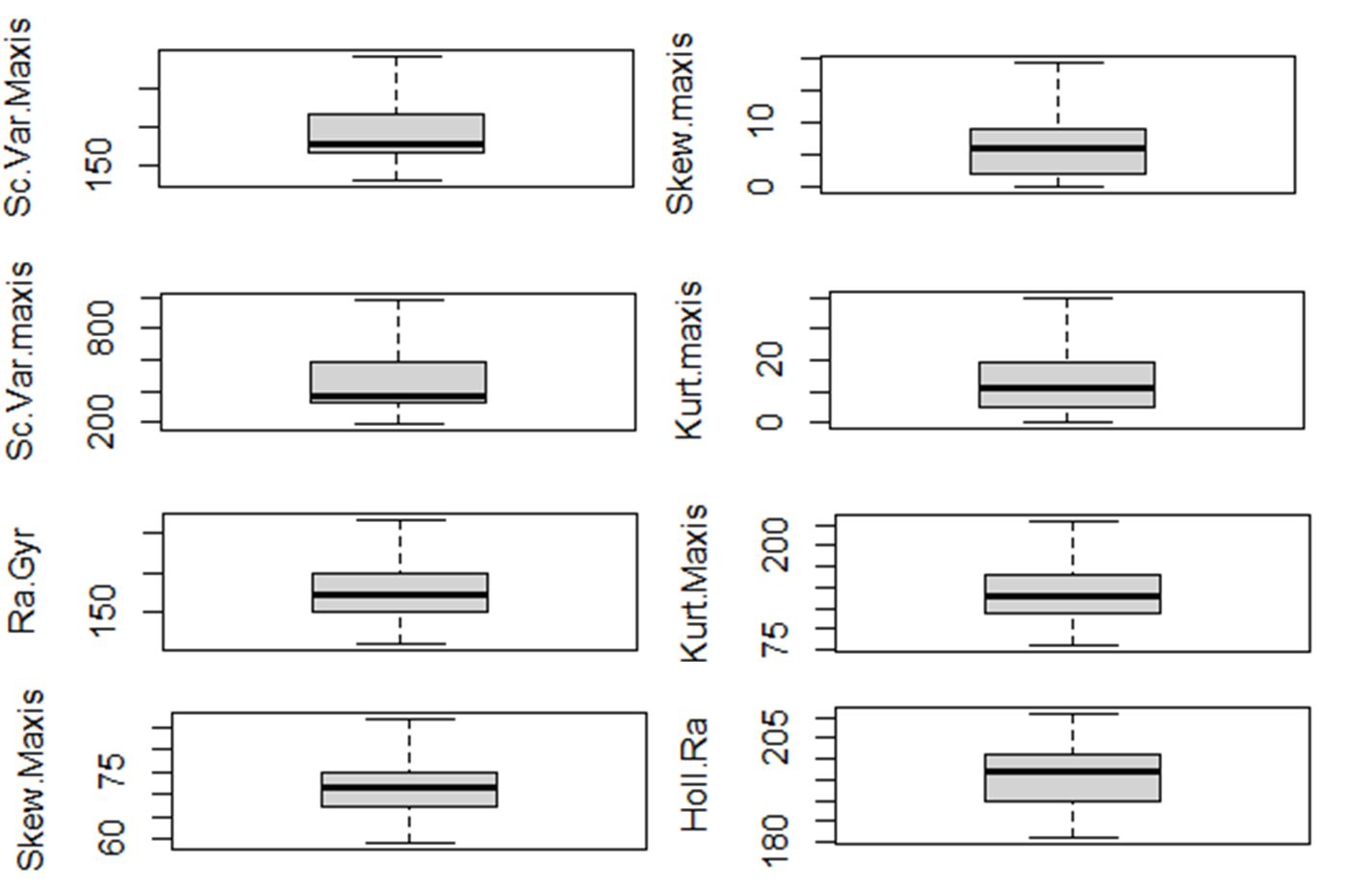
- The graphs below represent boxplots for all columns or features of the dataset which consist of outliers



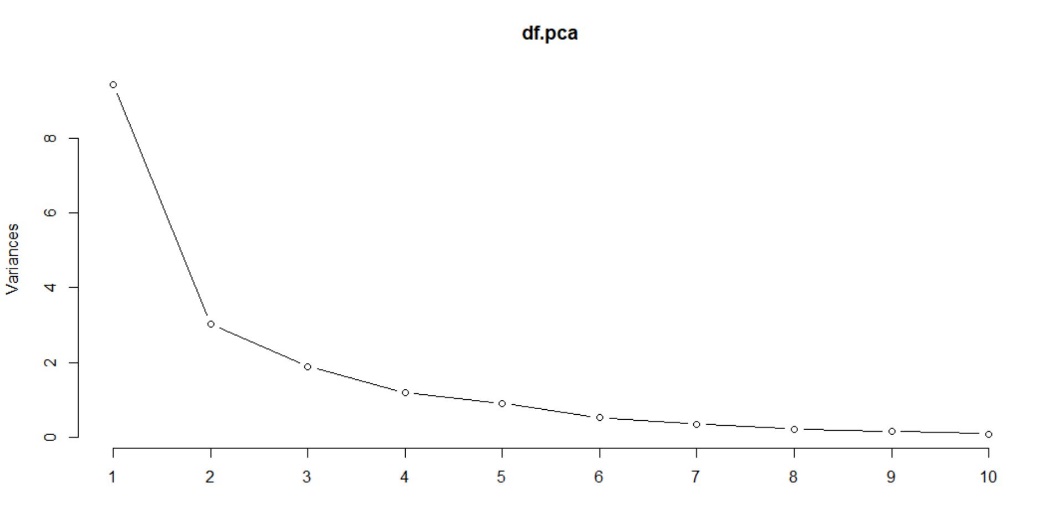


- The graphs below represent boxplots for all columns or features of the dataset where the outliers are removed.

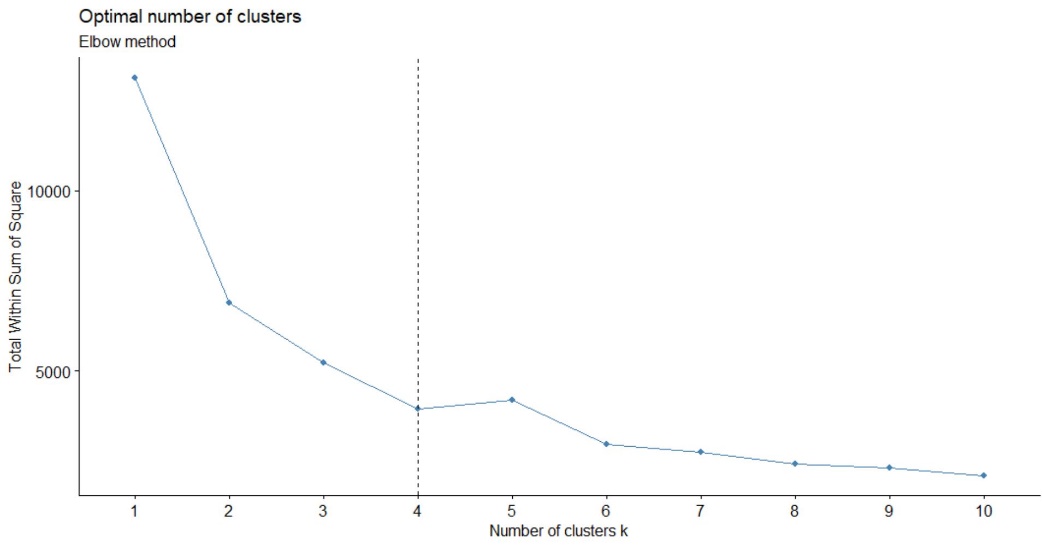




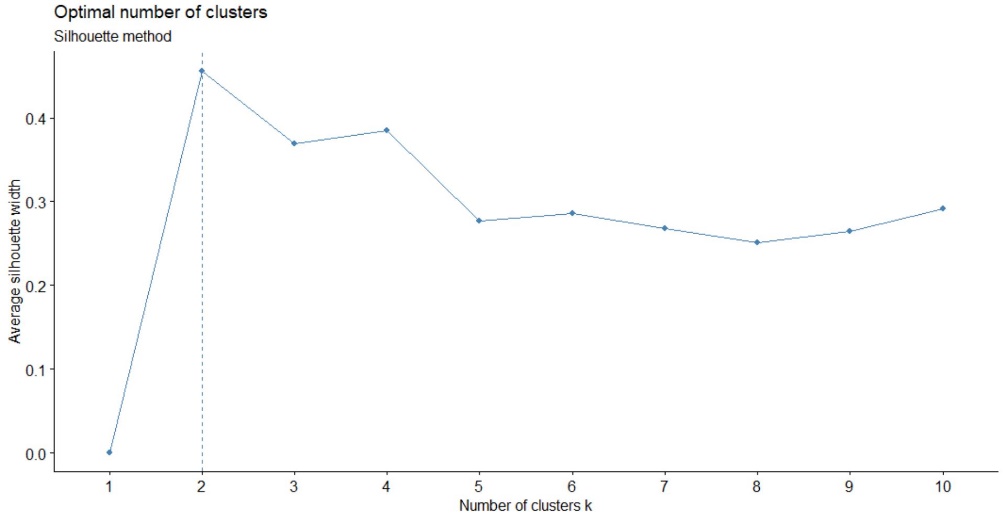
- Plotting the PCA data to find the best number of Principal Components to be taken. Using the elbow method of the plot above we can get the number of components which explains 85% or greater of the variation (BEST SET of components to take in this case the first 4 components are the best because it covers the greatest area of the graph)



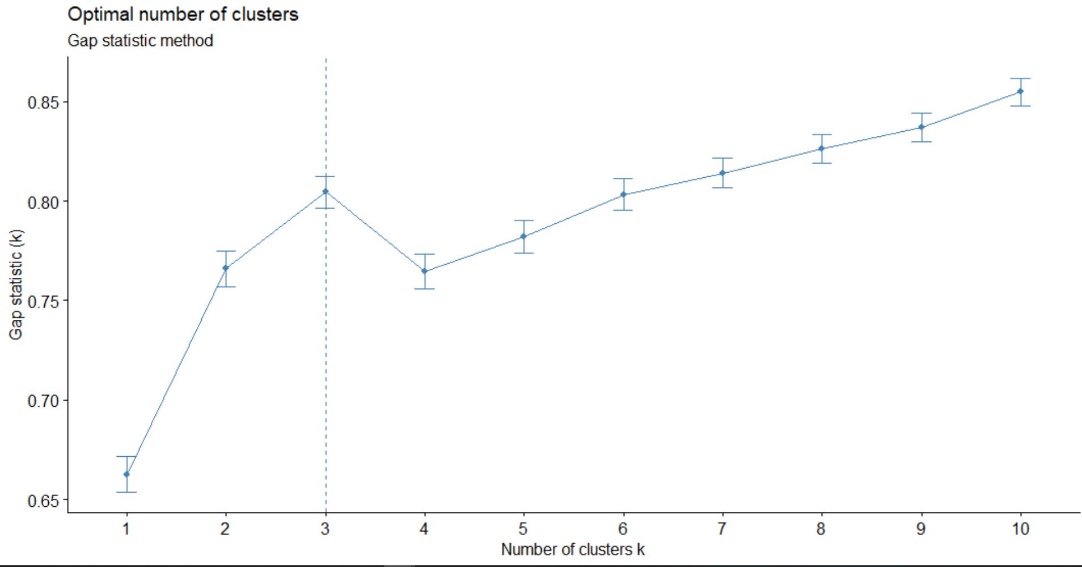
- Using the automated tools to find the best number of centroids (using the ELBOW method)



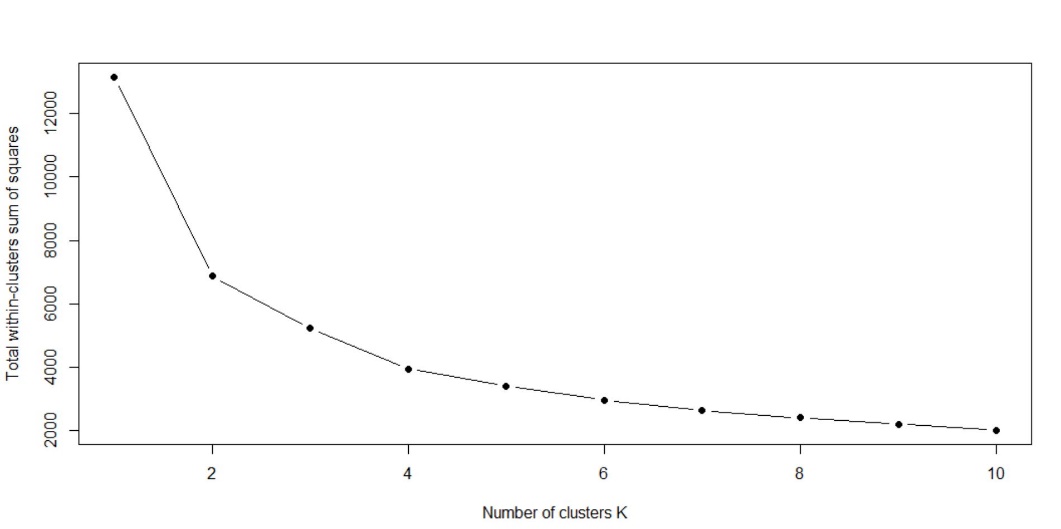
- Using the automated tools to find the best number of centroids (using the SILHOUETTE method)



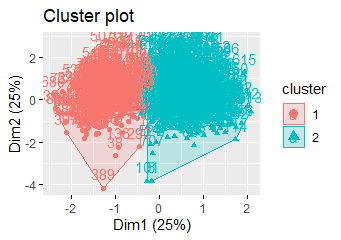
- Using the automated tools to find the best number of centroids (using the GAP STATISTIC method)



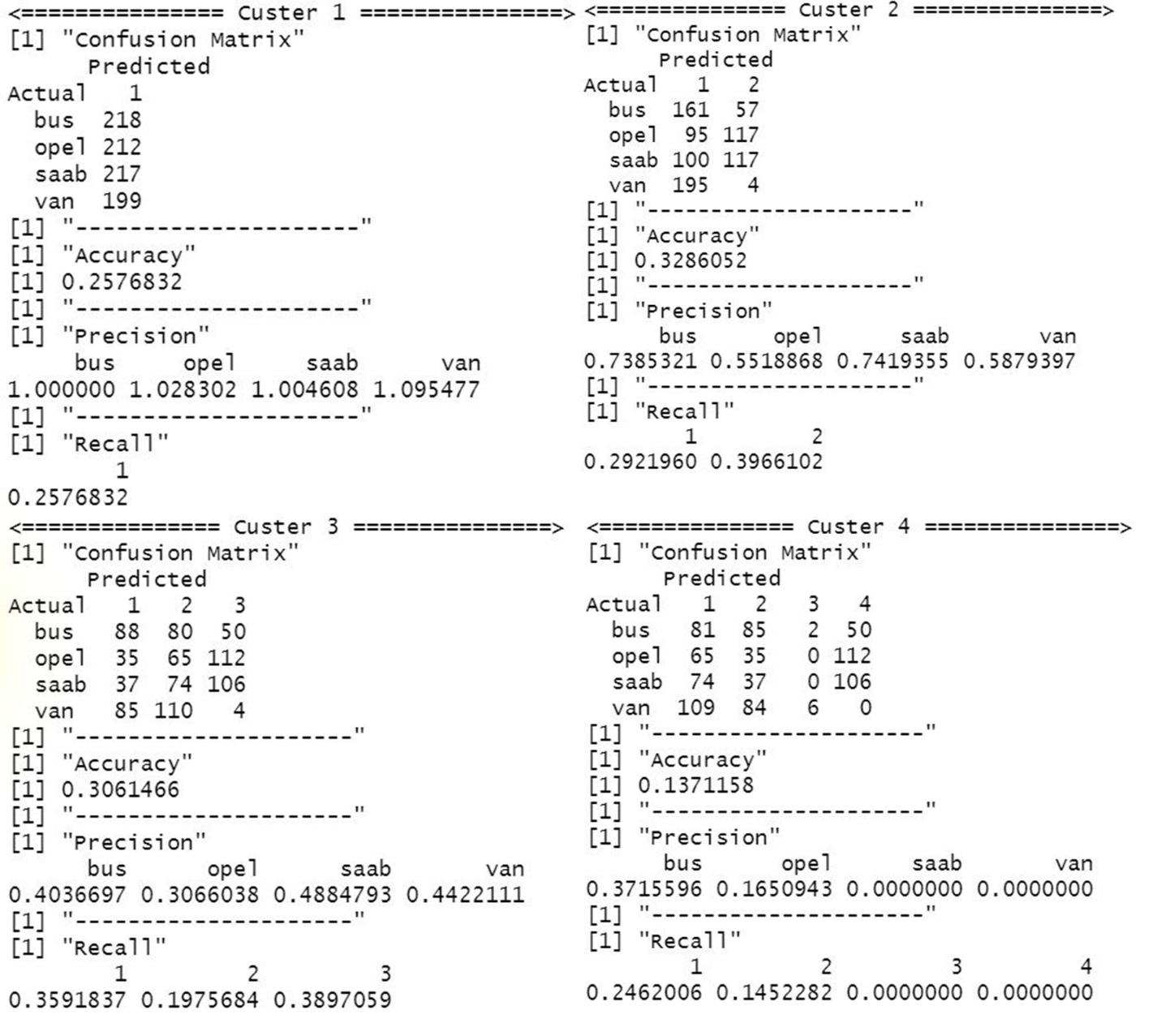
- Using the manual method to find the best number of centroids (using the ELBOW method)

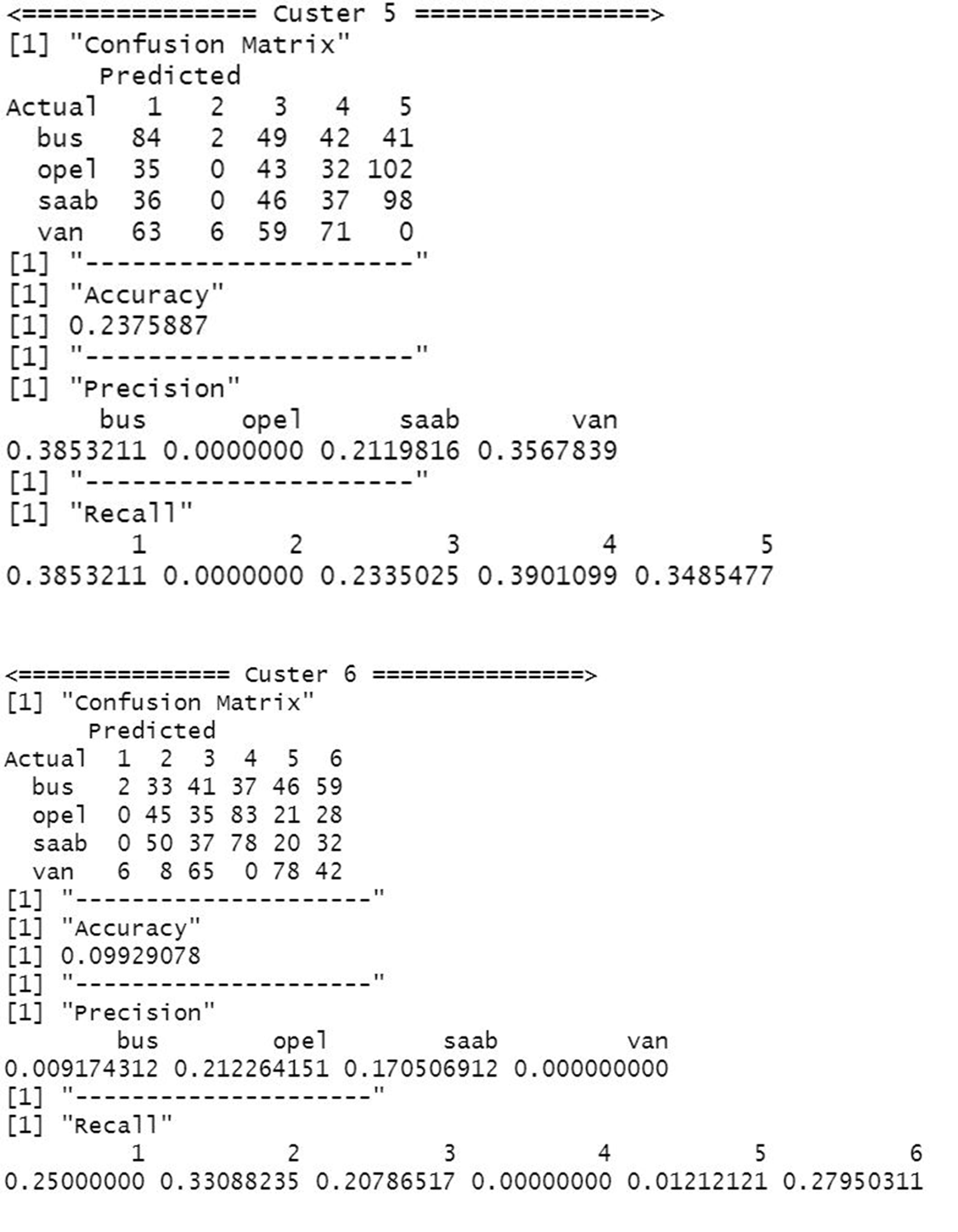


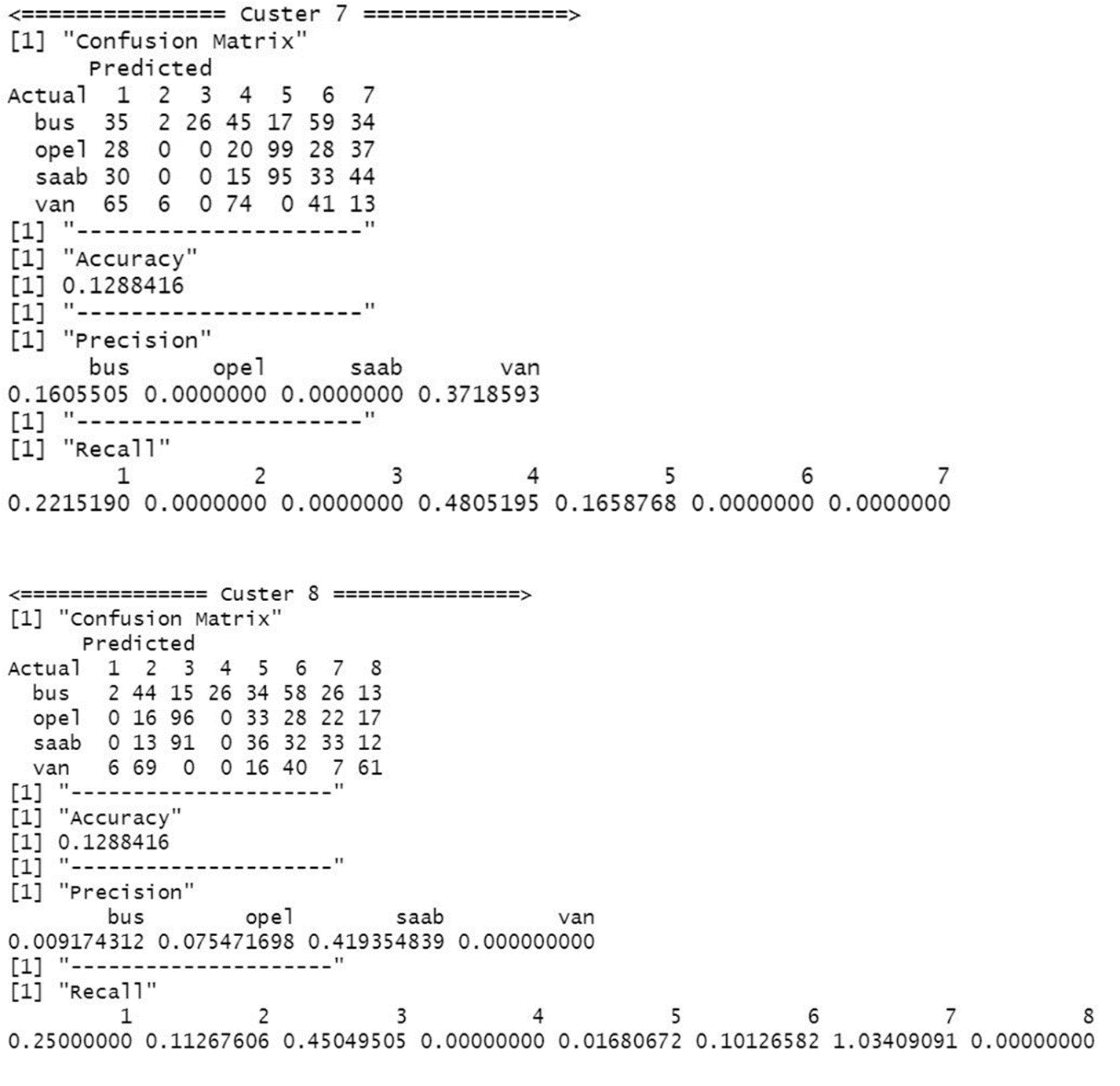
- Since I have got the best accuracy for 2 clusters I have included and image of the cluster separation.

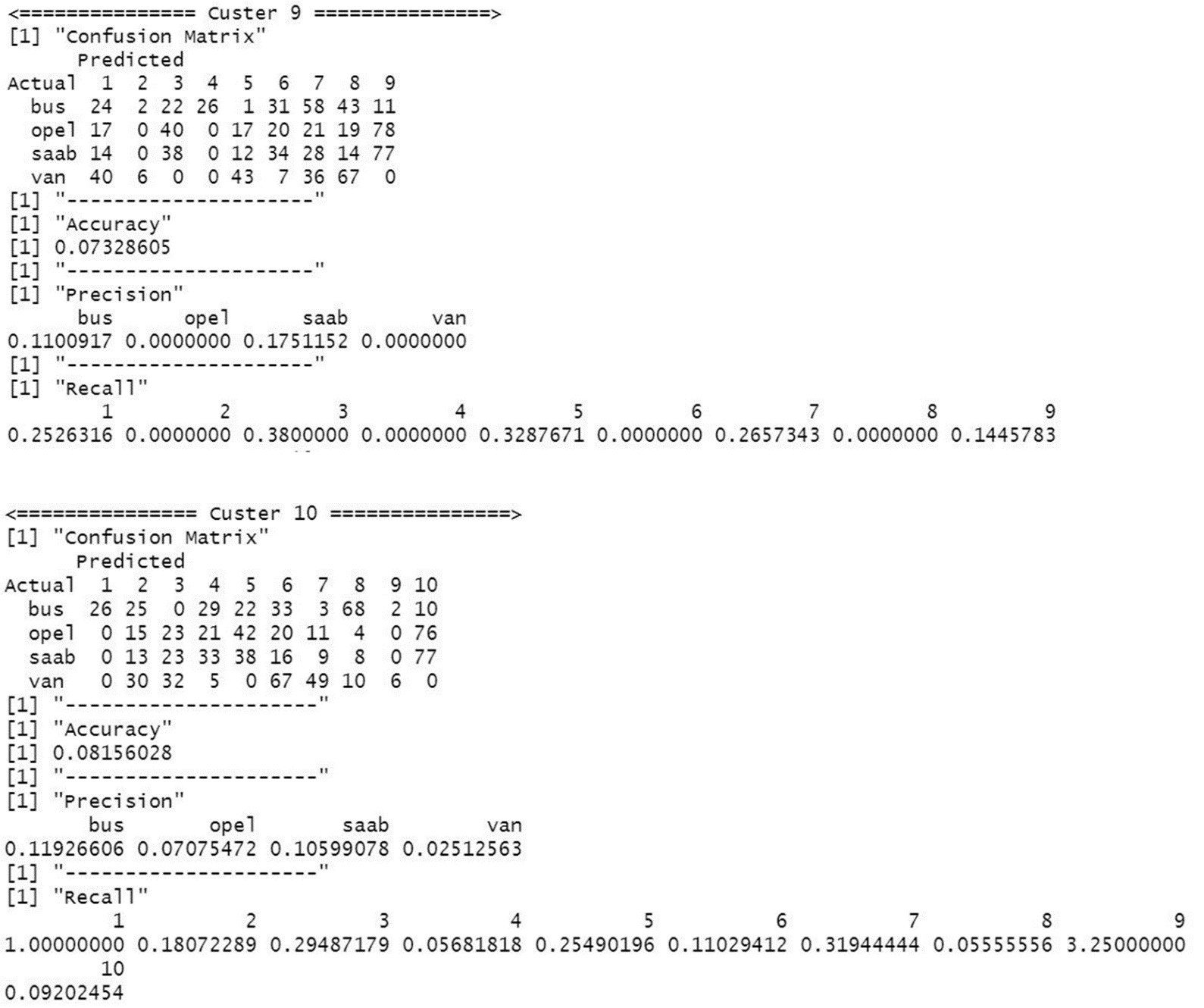


- These are the evaluation results obtained for the 10 clusters centroid I ran with to find the optimal cluster centroid number.









**References**

<https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/>

<https://discuss.analyticsvidhya.com/t/dimensionality-reduction-is-good-or-bad/2444#:~:text=The%20very%20purpose%20of%20applying,them%20out%20from%20further%20analysis.&text=So%20it%20is%20good%20to%20do%20the%20dimension%20reduction%20wherever%20possible>.

<https://www.javatpoint.com/principal-component-analysis#:~:text=Principal%20Component%20Analysis%20is%20an,dimensionality%20reduction%20in%20machine%20learning.&text=The%20PCA%20algorithm%20is%20based,Eigenvalues%20and%20Eigen%20factors>