**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 won’t be getting much of a better result or a better accuracy, because all the features contribute in a different proportion not equally.

E.g.: - 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 quite 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, let’s 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.**

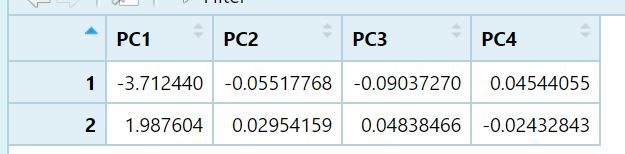
* When I used the automated tools inorder to calculate the number of clusters formed these are the following results which were obtained

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

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

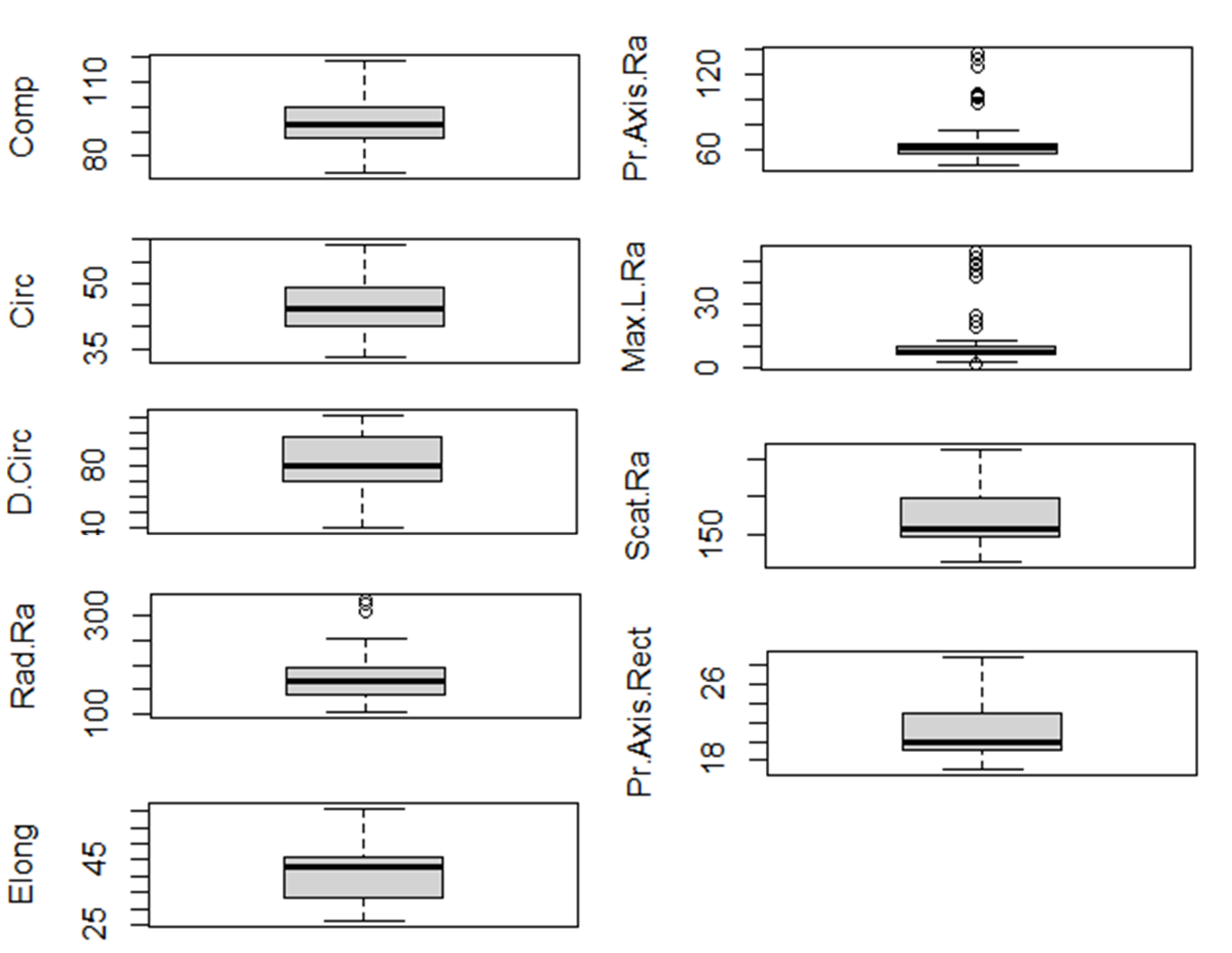
Gap Statistics Method (Automated) gave me 3 clusters as the best one.

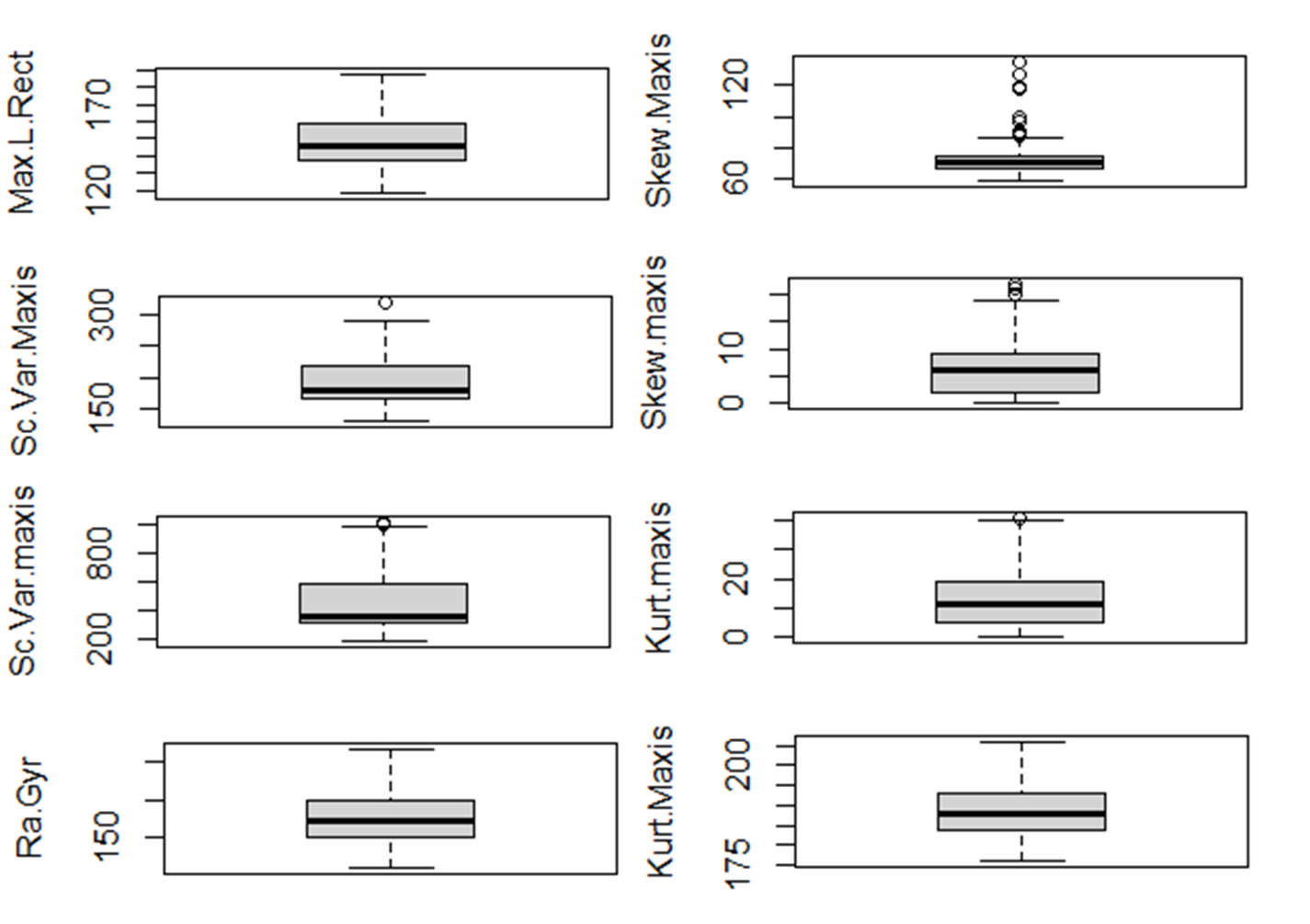
* I now manually created an elbow method to compute the best number of cluster by looping though 1 to 10 number of clusters and performing the KMeans clustering and calculating its accuracy using the confusion matrix. These are the follow results given for the clusters from 1 to 10 with its accuracy, 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.



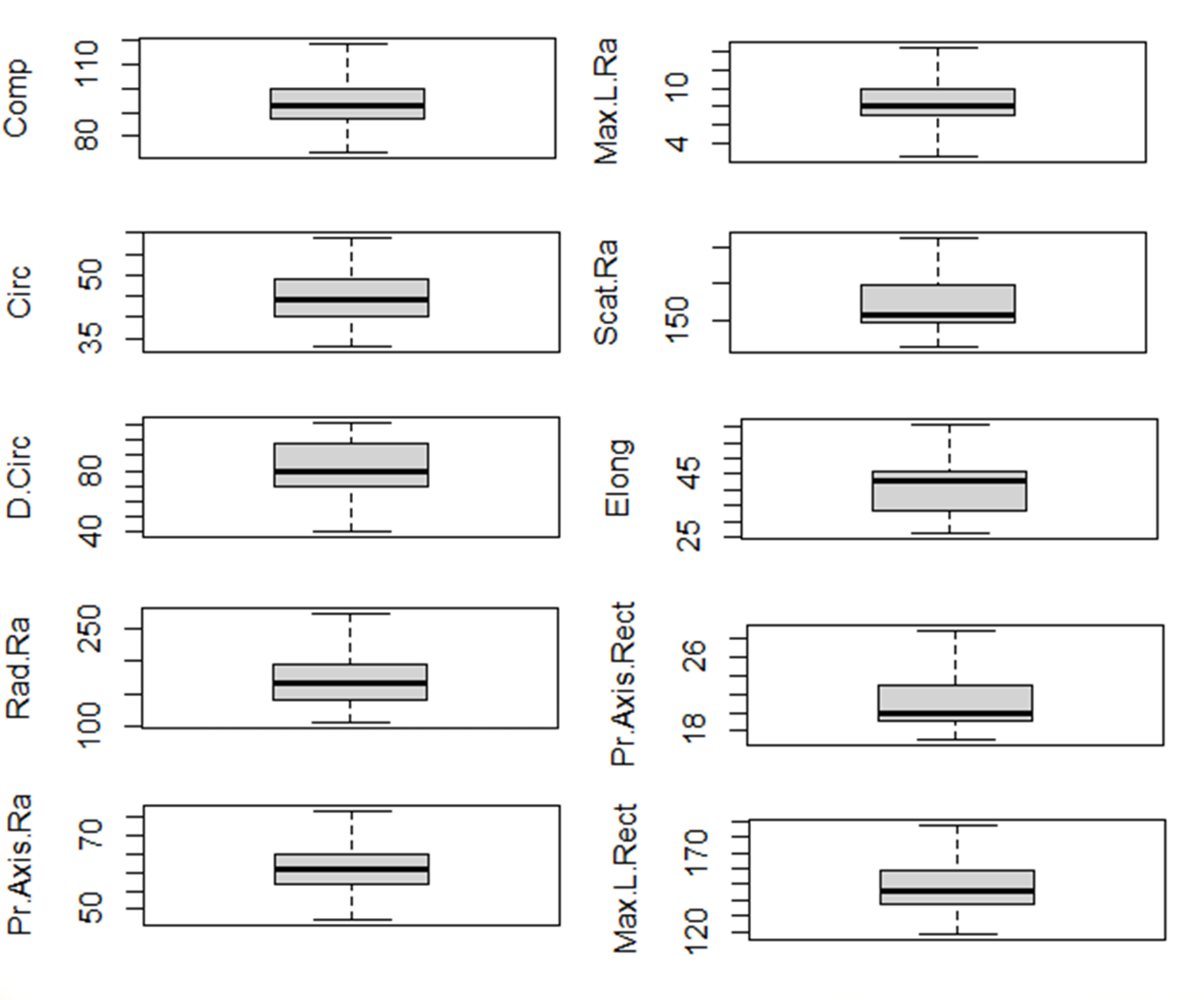
1. **The following are graphs obtained and its respective description about the graph**

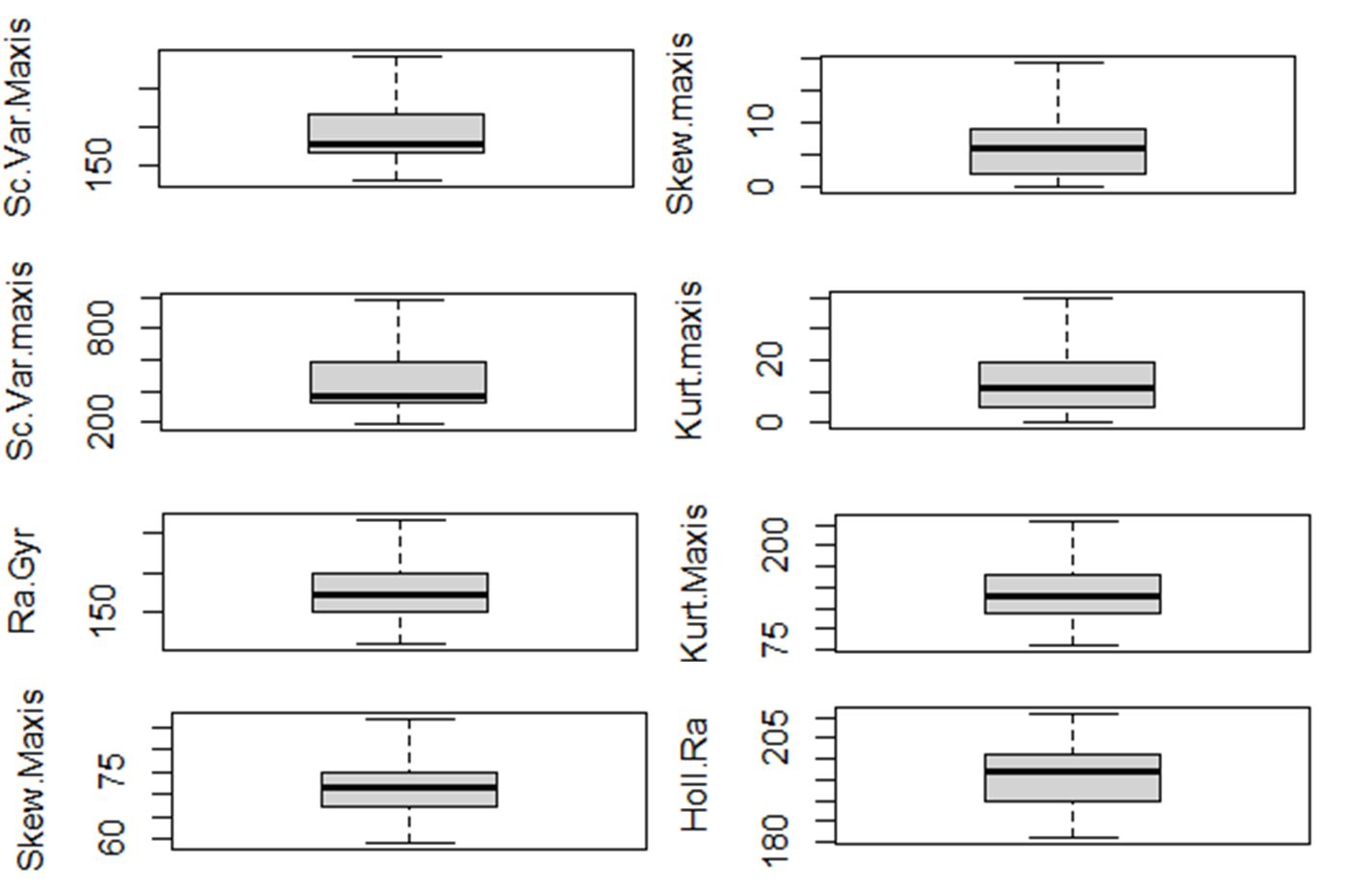
- The graphs below represent boxplots for all columns or features of the dataset which consist of outliers



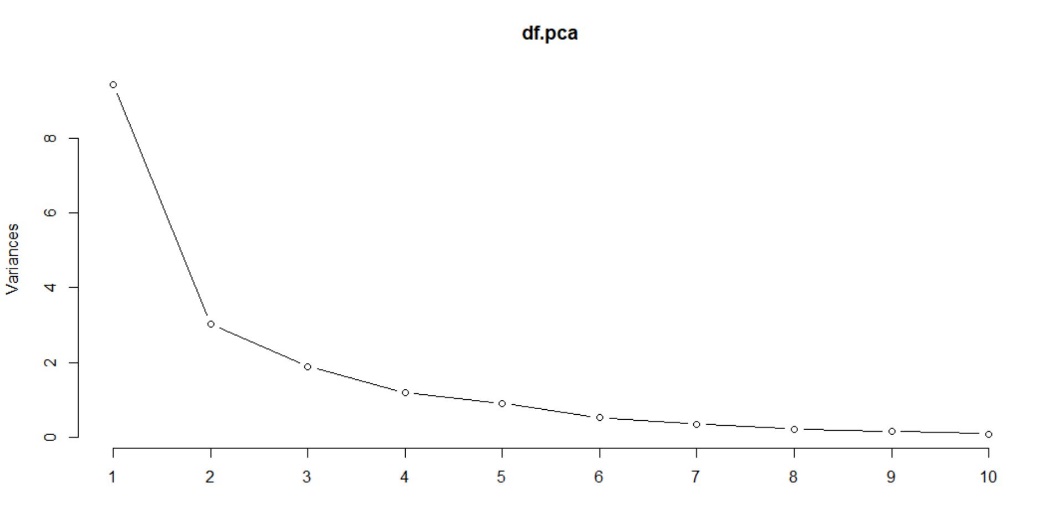


- The graphs below represent boxplots for all columns or features of the dataset where the outliers are removed.

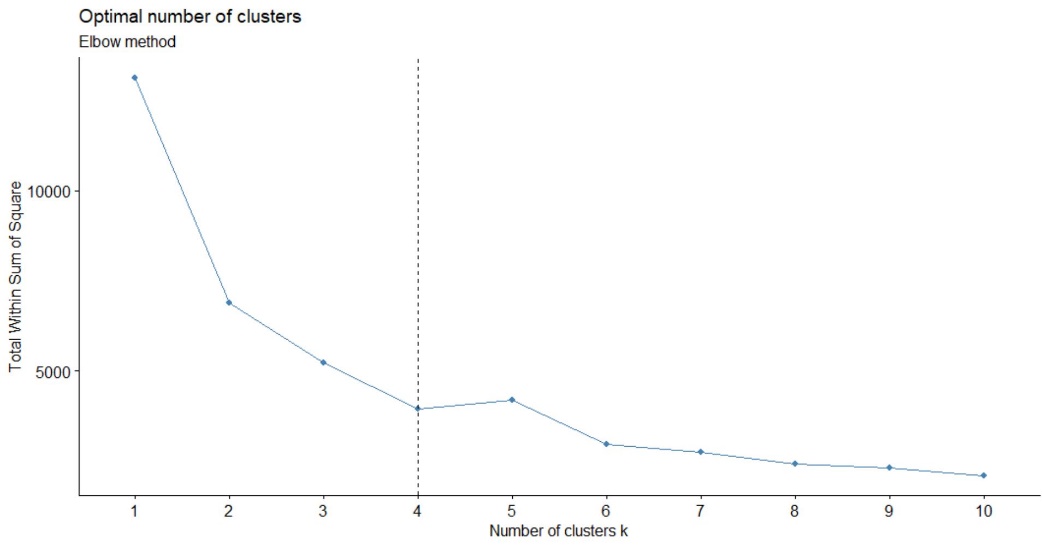




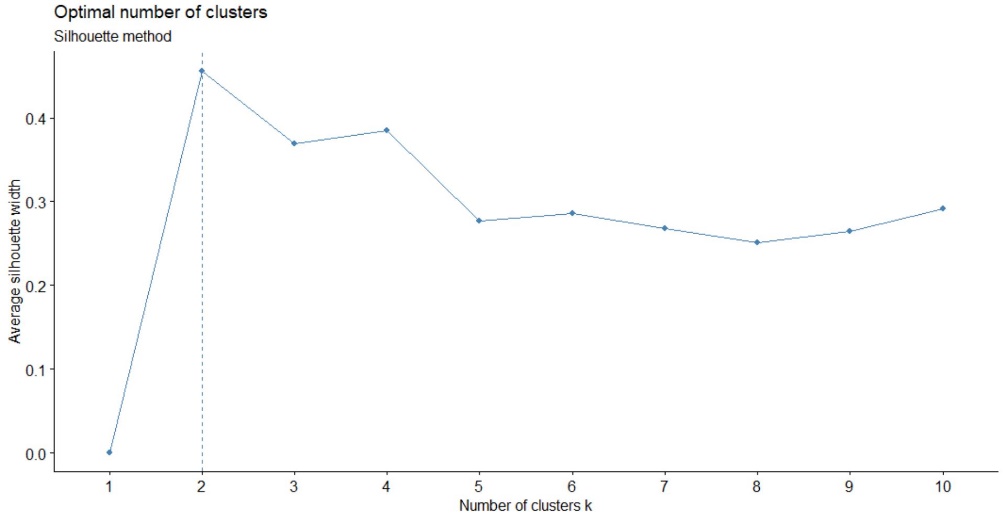
- Plotting the PCA data to find the best number of Principal Components to be taken. Using the elbow method of the plot above we can get the number of components which explains 85% or greater of the variation (BEST SET of components to take in this case the first 4 components are the best because it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



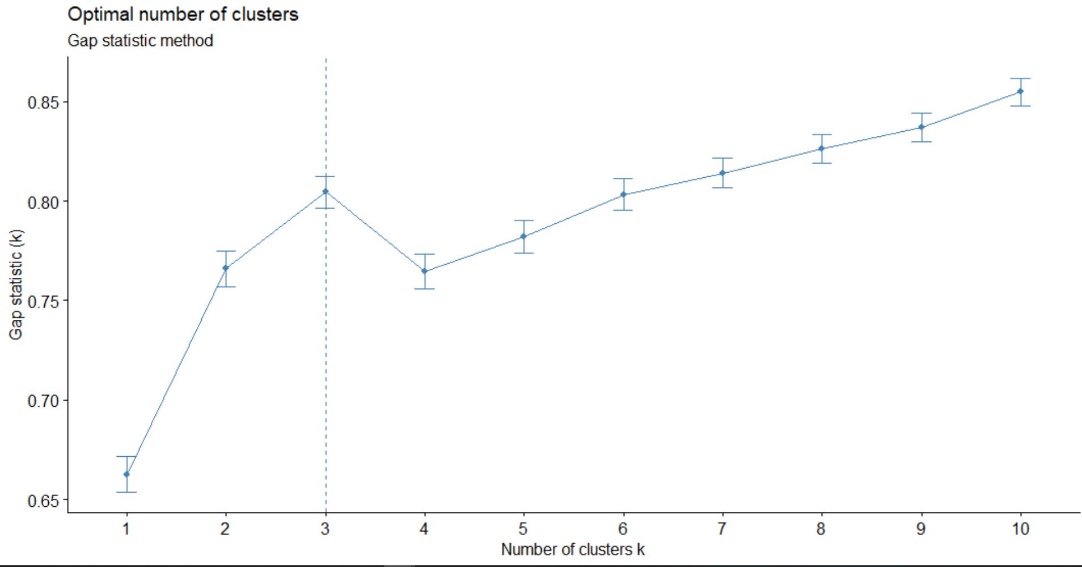
- Using the automated tools to find the best number of centroids (using the ELBOW method)



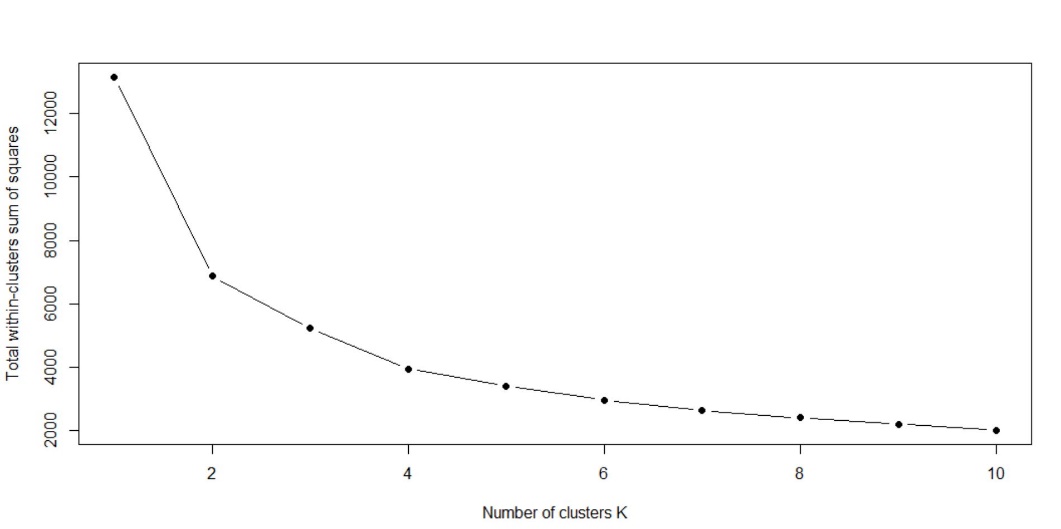
- Using the automated tools to find the best number of centroids (using the SILHOUETTE method)



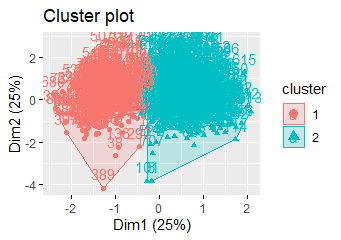
- Using the automated tools to find the best number of centroids (using the GAP STATISTIC method)



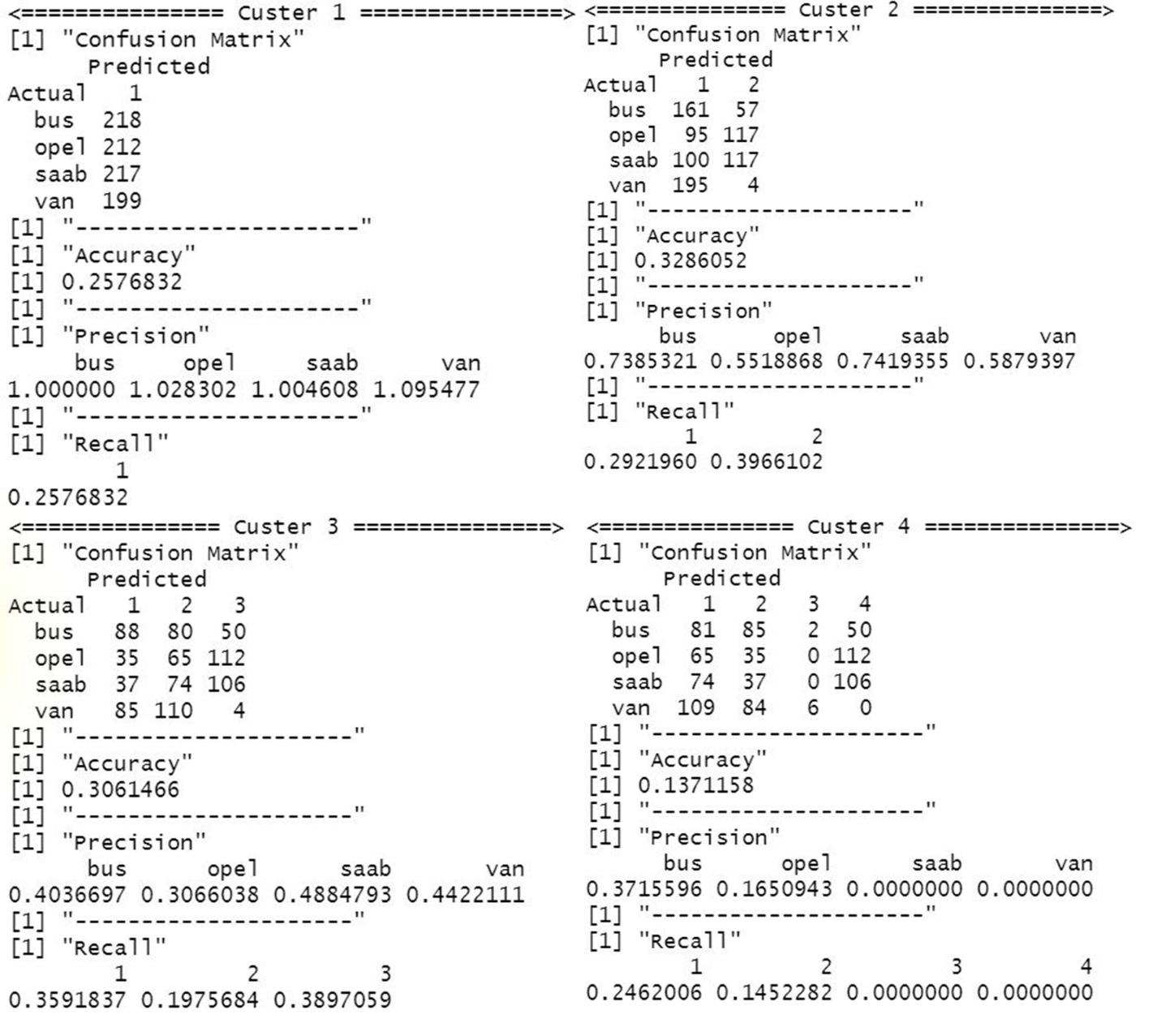
- Using the manual method to find the best number of centroids (using the ELBOW method)

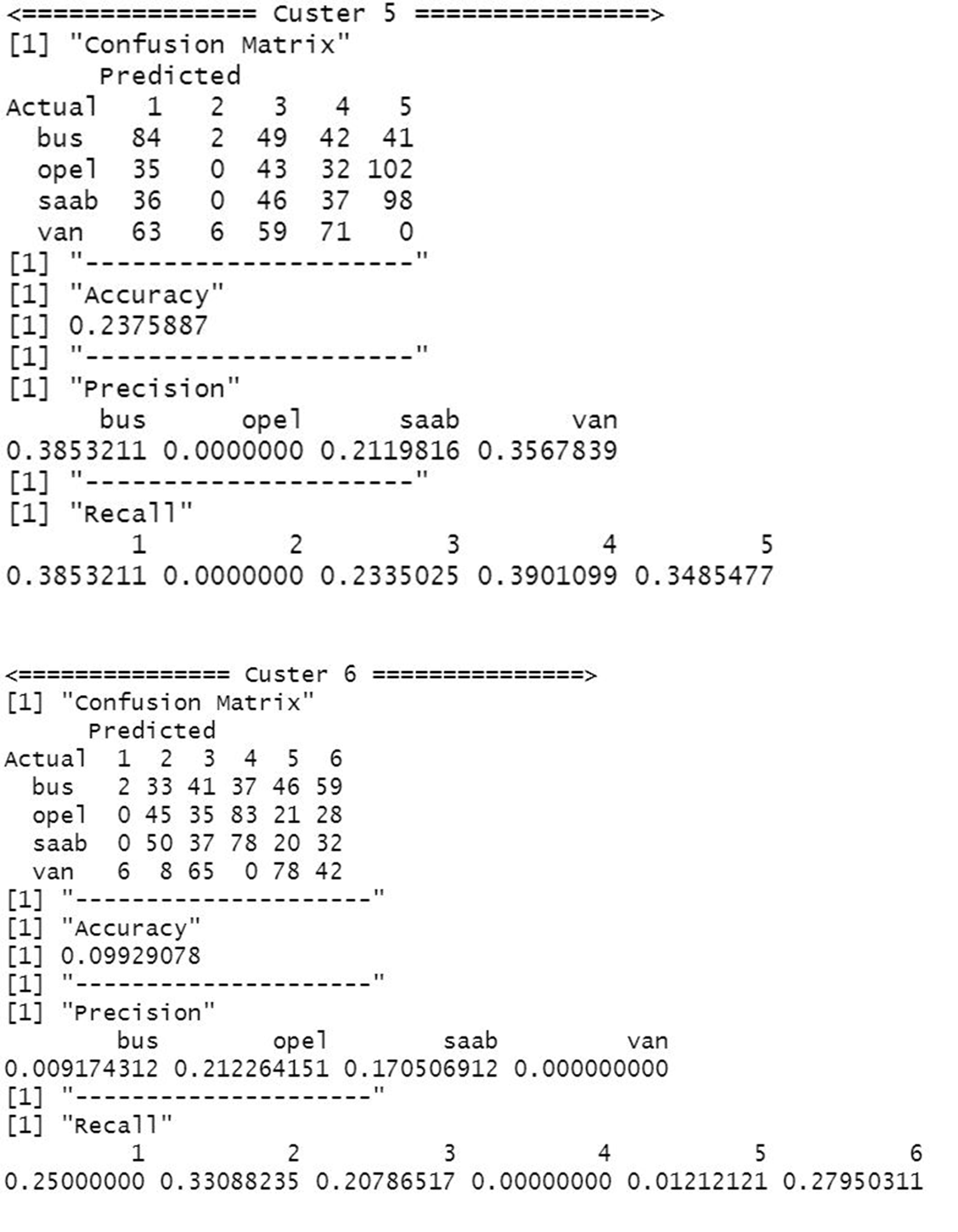


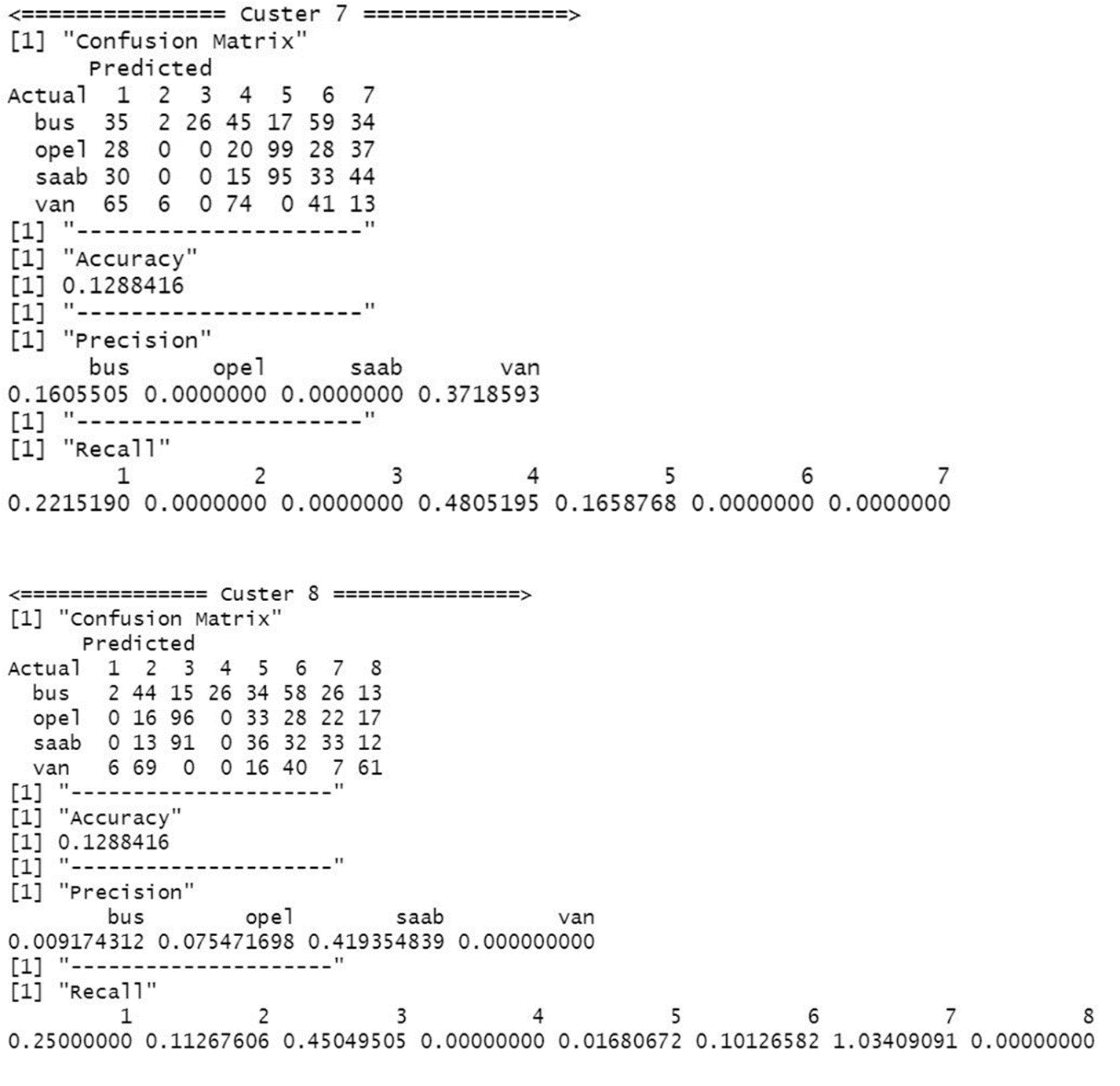
- Since I have got the best accuracy for 2 clusters I have included and image of the cluster separation.

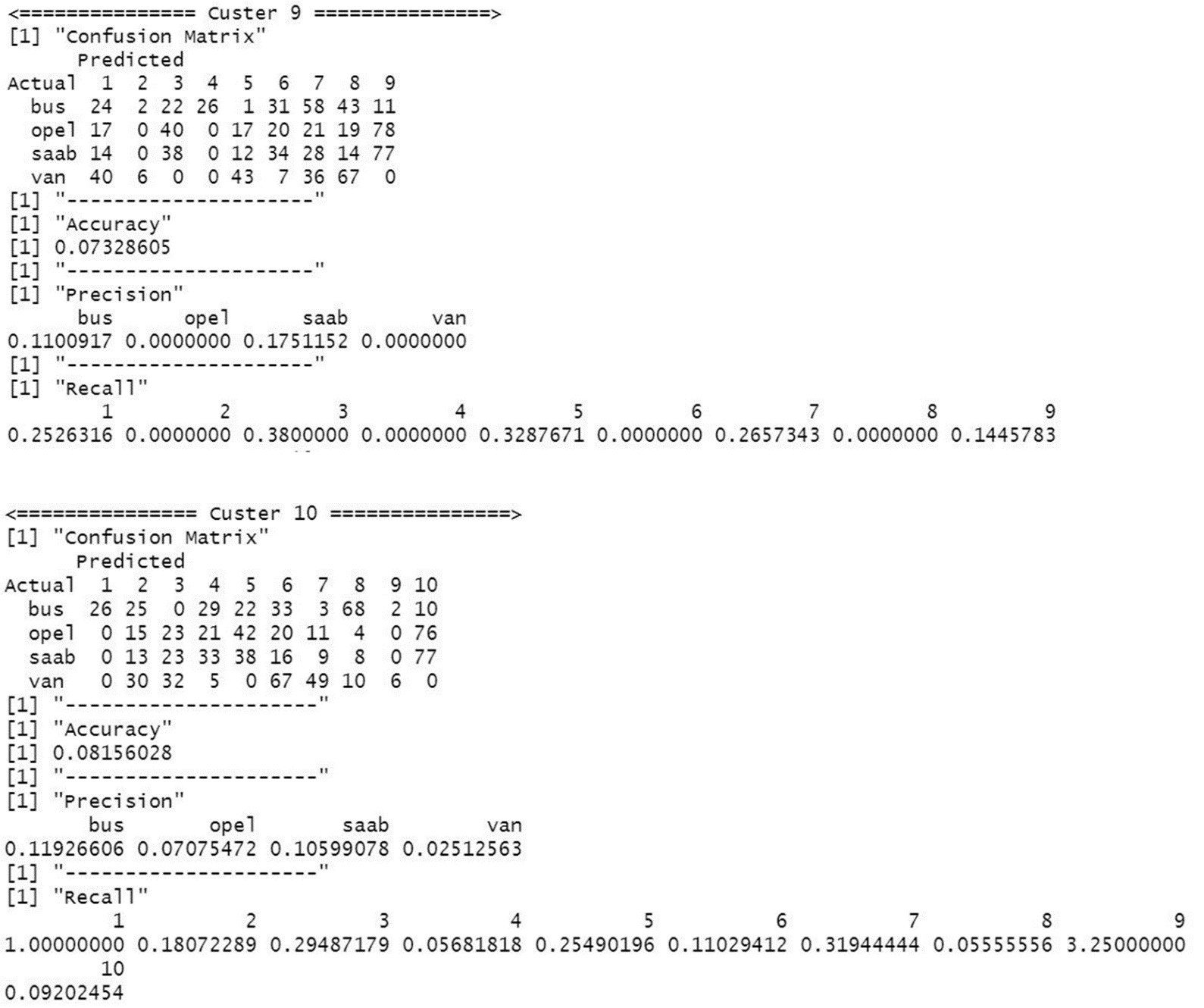


- These are the evaluation results obtained for the 10 clusters centroid I ran with to find the optimal cluster centroid number.









**FORECASTING PART**

1. **Discussion of the various schemes used to define the input vector.**
2. **Explain why normalization procedure is necessary for this specific type of NN.**

Since neural networks deals with a large number of layers and a large number of nodes then the time taken for the model to train properly will take a lot of time in general, but by normalizing the data input to the neural network it speeds up the learning process to a higher rate leading to faster or quicker convergence or increase the chances to reach the global minima quickly. Moreover, normalization is important give a consistence scale range for all the features there by making it more accurate for training and not making it bias to a specific feature due to its input size.

1. **Comparision table for the testing performance based on the statistical indices.**
2. **Brief explanation on the statistical indices (RMSE, MAE and MAPE)**

* **RMSE (Root Mean Square Error):** This is a statistical measure of the error of the model in predicting the respective data. Lower the RMSE value better the model is, better the prediction output is as well.
* **MAE (Mean Absolute Error):** This is basically the sum or total of average of the absolute difference between the predicted and actual values. So, what basically this means is that using the MAE we can get a clear picture as to how wrong the predictions are. Lower the MAE better the model
* **MAPE (Mean Absolute Percentage Error):** This is a statistical measure to define the accuracy of a model on a particular dataset or in other terms it is the percentage of average of absolute difference between predicted values and true values, divided by the true values. Lower the MAPE better the model again.

1. **Efficiency result from the best one hidden layer and two hidden layer network with their respective parameters.**
2. **Discussion about the best final MLP Network by checking with the evaluation stats and with predicted VS actual result graph**
3. **The Code for the forecasting part problem.**

**APPENDIX**

**CODE FOR PART 01 (CLUSTERING QUESTION)**

# --------------------------------------------------

# Name: Mohammed Nazhim Kalam

# Student ID: 2019281

# UoW ID: W1761265

# --------------------------------------------------

# CLUSTERING PART

# Installing package to read Excel Data-set

install.packages("readxl") # used to read excel data files

install.packages("factoextra") # used to determine the optimal number clusters

install.packages("NbClust") # used to compute about multiple methods at once,

# in order to find the optimal number of clusters.

install.packages("factoextra") # used to plot the clusters out

# Loading the package

library(readxl)

library(knitr)

library(tidymodels)

library(janitor)

library(flexclust)

library(dplyr)

library(factoextra)

library(NbClust)

library(haven)

library(factoextra)

# Reading the data-set "vehicles.xlsx"

df = read\_excel("./vehicles.xlsx")

View(df)

# converting the class column into factors because we are not able to get its count

df = mutate(df, Class = as\_factor(df$Class))

# Displaying the types of unique classes present in the data-set

summary(df)

# Removing the Sample index column and the class column from the data-set

df.filtered = subset(df, select = -c(Samples, Class))

# Viewing the filtered data-set

View(df.filtered)

# [PRE-PROCESSING DATA] PERFORMING SCALING AND OUTLIERS REMOVAL

# Checking for any null values present in the data-set (returned 0 so no null values present)

print(sum(is.na(df.filtered)))

# Checking the Summary of the data-set (We can see the stats of the data columns eg: mean, median etc.)

df.summary = summary(df.filtered)

# Plotting a box plot graph (Box plots are useful to detect potential outliers from the data-set)

display.boxplot = function(data, column.name){

boxplot(data, ylab = column.name)

}

# calling the display.boxplot function to display the boxplot data representation for each column.

display.boxplot(df.filtered$Comp, "Comp")

display.boxplot(df.filtered$Circ, "Circ")

display.boxplot(df.filtered$D.Circ, "D.Circ")

display.boxplot(df.filtered$Rad.Ra, "Rad.Ra")

display.boxplot(df.filtered$Pr.Axis.Ra, "Pr.Axis.Ra")

display.boxplot(df.filtered$Max.L.Ra, "Max.L.Ra")

display.boxplot(df.filtered$Scat.Ra, "Scat.Ra")

display.boxplot(df.filtered$Elong, "Elong")

display.boxplot(df.filtered$Pr.Axis.Rect, "Pr.Axis.Rect")

display.boxplot(df.filtered$Max.L.Rect, "Max.L.Rect")

display.boxplot(df.filtered$Sc.Var.Maxis, "Sc.Var.Maxis")

display.boxplot(df.filtered$Sc.Var.maxis, "Sc.Var.maxis")

display.boxplot(df.filtered$Ra.Gyr, "Ra.Gyr")

display.boxplot(df.filtered$Skew.Maxis, "Skew.Maxis")

display.boxplot(df.filtered$Skew.maxis, "Skew.maxis")

display.boxplot(df.filtered$Kurt.maxis, "Kurt.maxis")

display.boxplot(df.filtered$Kurt.Maxis, "Kurt.Maxis")

display.boxplot(df.filtered$Holl.Ra, "Holl.Ra")

# REMOVING THE OUTLIERS FROM THE DATASET

# Discarding the outliers from the data-set,

# any value greater than bench.mark value will be replace with the bench mark value

remove.outliers = function(data, column.name){

# filter with box plot and trimming out from

# "maximum": Q3 + 1.5\*IQR

# "minimum": Q1 -1.5\*IQR

# where interquartile range (IQR): 25th to the 75th percentile.

# Calculating the upper and lower limit for the data

bench.mark.upper = quantile(data, 0.75) + (1.5 \* IQR(data))

bench.mark.lower = quantile(data, 0.25) - (1.5 \* IQR(data))

# Replacing the outliers with the upper and lower limits

data[data > bench.mark.upper] = bench.mark.upper

data[data < bench.mark.lower] = bench.mark.lower

# Display the box-plot after removing the outlier

display.boxplot(data, column.name)

}

# calling the remove.outliers function to remove all the outliers from each column of the data

remove.outliers(df.filtered$Comp, "Comp")

remove.outliers(df.filtered$Circ, "Circ")

remove.outliers(df.filtered$D.Circ, "D.Circ")

remove.outliers(df.filtered$Rad.Ra, "Rad.Ra")

remove.outliers(df.filtered$Pr.Axis.Ra, "Pr.Axis.Ra")

remove.outliers(df.filtered$Max.L.Ra, "Max.L.Ra")

remove.outliers(df.filtered$Scat.Ra, "Scat.Ra")

remove.outliers(df.filtered$Elong, "Elong")

remove.outliers(df.filtered$Pr.Axis.Rect, "Pr.Axis.Rect")

remove.outliers(df.filtered$Max.L.Rect, "Max.L.Rect")

remove.outliers(df.filtered$Sc.Var.Maxis, "Sc.Var.Maxis")

remove.outliers(df.filtered$Sc.Var.maxis, "Sc.Var.maxis")

remove.outliers(df.filtered$Ra.Gyr, "Ra.Gyr")

remove.outliers(df.filtered$Skew.Maxis, "Skew.Maxis")

remove.outliers(df.filtered$Skew.maxis, "Skew.maxis")

remove.outliers(df.filtered$Kurt.maxis, "Kurt.maxis")

remove.outliers(df.filtered$Kurt.Maxis, "Kurt.Maxis")

remove.outliers(df.filtered$Holl.Ra, "Holl.Ra")

# NORMALIZING THE DATASET (BRINGING ALL THE DATA INTO A SINGLE UNQIUE SCALE)

# Performing normalization using the Z-Score Standardization

df.normalized = as.data.frame(scale(df.filtered))

View(df.normalized)

# PERFORMING PCA (PRINCIPAL COMPONENT ANALYSIS) / DIMENSIONALITY REDUCTION

df.pca = prcomp(df.normalized)

summary(df.pca)

# Plotting the PCA data to find the best number of Principal Components.

# Using the elbow method of the plot below we can get the number of components which

# explain 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 and has the sudden decrease after the 4th component

plot(df.pca)

plot(df.pca, type='l')

# comp.data contains the BEST PCA Component data extract

comp.data = data.frame(df.pca$x[,1:4])

View(comp.data)

# DETERMINE THE NUMBER OF CLUSTERS CENTERS (CENTROIDS) (via MANUAL and AUTOMATED TOOLS)

# AUTOMATED TOOLS TO FIND THE CENTROIDS

# USING ELBOW METHOD (Gave 4)

# The below method points out that 4 is the optimal number of centroids/clusters to be taken

fviz\_nbclust(comp.data, kmeans, method = "wss") +

geom\_vline(xintercept = 4, linetype = 2) +

labs(subtitle = "Elbow method")

# USING THE SILHOUETTE METHOD (Gave 2)

# The below method points out that 2 is the optimal number of centroids/clusters to be taken

fviz\_nbclust(comp.data, kmeans, method = "silhouette")+

labs(subtitle = "Silhouette method")

# USING GAP STATISTIC ( nboot = 50 to keep the function speedy

# recommended value: nboot= 500 for your analysis.

# Use verbose = FALSE to hide computing progression.)

# (Gave 3)

# The below method points out that 3 is the optimal number of centroids/clusters to be taken

set.seed(150)

fviz\_nbclust(comp.data, kmeans, nstart = 50, method = "gap\_stat", nboot = 50)+

labs(subtitle = "Gap statistic method")

# MANUALLY FIND THE CENTROIDS / CLUSTERS

# USING ELBOW METHOD

tot.withinss = vector(mode = "character", length = 10)

# Classification Report Function

classification\_report <- function(comparison\_table, dp = 2) {

#total counts

counts <- sum(comparison\_table)

#total sums for each column

column\_sums <- colSums(comparison\_table)

#total sums for each row

row\_sums <- rowSums(comparison\_table)

#true positive value

tp <- diag(comparison\_table)

#true negative

tn <- counts - (column\_sums + row\_sums - tp)

#false positive

fp <- row\_sums - tp

#false negative

fn <- column\_sums - tp

#precision

pr <- tp / (tp + fp)

#recall

re <- tp / (tp + fn)

#accuracy

ac <- sum(tp) / counts

# Displaying the Accuracy, Precision and Recall values

print("---------------------")

print("Accuracy")

print(ac)

print("---------------------")

print("Precision")

print(pr)

print("---------------------")

print("Recall")

print(re)

}

# Looping from 1 to the max optimal cluster to find its evaluation result

for (i in 1:10){

cat("<=============== ", "Custer ", i, " ===============>\n", sep = "")

set.seed(50)

# Performing Kmeans clustering

vehicleCluster = kmeans(comp.data, centers = i, nstart = 20)

# This is the confusion matrix

cm = as.matrix(table(Actual = df$Class, Predicted = vehicleCluster$cluster))

print("Confusion Matrix")

print(cm)

# Display Classification report

classification\_report(comparison\_table = cm)

# Total within-cluster sum of squares

tot.withinss[i] = vehicleCluster$tot.withinss

}

# plot to find the best number of clusters to be taken

plot(1:10,

tot.withinss,

type="b",

pch=19,

xlab = "Number of clusters K",

ylab = "Total within-clusters sum of squares")

# from the accuracy result we can see that we got the highest accuracy result for

# 2 clusters (33%)

set.seed(100)

vehicleCluster = kmeans(comp.data, centers = 2, nstart = 20)

fviz\_cluster(vehicleCluster, data = comp.data)

# Displaying the sizes(number of observations in each cluster) of each cluster

vehicleCluster$size

# Displaying the cluster distribution

vehicleCluster$cluster

# Getting the centers (A matrix of cluster centers).

# You get a 2 by 4 matrix representing the centers of each cluster (x,y,z,z')

# (4D coordinates) due to PCA we took 4D Data

vehicleCluster$centers

View(vehicleCluster$centers)

**CODE FOR PART 02 (FORECASTING QUESTION)**

**REFERENCES**

Brownlee, Jason. ‘How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification’. Machine Learning Mastery, 2 Jan. 2020, https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/.

Dimensionality Reduction Is Good or Bad’. Data Science, Analytics and Big Data Discussions, 24 July 2015, https://discuss.analyticsvidhya.com/t/dimensionality-reduction-is-good-or-bad/2444.

Principal Component Analysis - Javatpoint’. Www.Javatpoint.Com, https://www.javatpoint.com/principal-component-analysis. Accessed 12 Apr. 2021

Stöttner, Timo. ‘Why Data Should Be Normalized before Training a Neural Network’. Medium, 16 May 2019, <https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d>

Harrison, Myles. PCA and K-Means Clustering of Delta Aircraft | R-Bloggers. 23 June 2014, <https://www.r-bloggers.com/2014/06/pca-and-k-means-clustering-of-delta-aircraft/>

Determining The Optimal Number Of Clusters: 3 Must Know Methods’. Datanovia, https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/. Accessed 12 Apr. 2021

RPubs - k-Means Clustering in R for Beginners. [https://rpubs.com/Nitika/kmeans\_Iris. Accessed 12 Apr. 2021](https://rpubs.com/Nitika/kmeans_Iris.%20Accessed%2012%20Apr.%202021)

K-Means Clustering in R: Algorithm and Practical Examples’. Datanovia, https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/. Accessed 12 Apr. 2021

K-Means Clustering in R with Example. https://www.guru99.com/r-k-means-clustering.html. Accessed 12 Apr. 2021