Clustering project report

1. **Mahmoud Alaa Mahmoud : 20191615074**
2. **Khairy Mohamed Khairy : 20191480897**
3. **Ahmed Hossam El-din Mohamed : 20191613643**
4. **Ahmed Hossam El-din Mahrous : 20191566027**
5. **Yehia Osama Mohamed Fouad : 20191615496**

**Code:**

library(readxl)

data\_1 <- read\_excel("C:/Users/laphouse/Desktop/data\_1.csv")

View(data\_1)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

library(dplyr)

#select valid columns to cluster them, \*id & variance columns ignored

mydata = select(data\_1,c(2,3))

mydata

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#WSS plot to choose the optimum number of clusters

#WSS plot function

wssplot <- function(data, nc=15, seed=2345){

wss <- (nrow(data)-1)\*sum(apply(data,2,var))

for (i in 2:nc){

set.seed(seed)

wss[i] <- sum(kmeans(data, centers=i)$withinss)}

plot(1:nc, wss, type="b", xlab="Number of Clusters",

ylab="Within groups sum of squares")

wss

}

wssplot(mydata)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#kmeans clustering

Km = kmeans(mydata,4)

Km

#show centers of clusters

Km$centers

#Ploting the clusters

library(ggplot2)

autoplot(Km,mydata,frame=TRUE)

library(cluster)

clusplot(mydata, Km$cluster,color=T,shade=T)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#hierarchical clustering:

#Agglomerative method:

h\_data=(mydata)

h\_data

#compute the distance between data rows

distances=dist(h\_data, method ="euclidean")

distances

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#method 1

#Agglomerative method:

#complete with agnes

aglom\_h=agnes(h\_data,method="complete")

aglom\_h

#look at the agglomerative coeffecient

aglom\_h$ac

#the lower the agglomerative coeffecient, the better the clustering

pltree(aglom\_h)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#complete with hclust

hclusters=hclust(distances)

plot(hclusters)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#method2

#divisive method

divisive\_h=diana(h\_data)

divisive\_h$dc #Disisive coeffecient

#plot the dendogram

pltree(divisive\_h)

rect.hclust(aglom\_h, k = 2, border = 2:4)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#density-based clustering

d= select(data\_1,c(2,3))

library(fpc)

Dbscan\_cl<-dbscan(d, eps= 10, MinPts= 6)

Dbscan\_cl

Dbscan\_cl$cluster

plot(Dbscan\_cl,d, main = "DBScan")

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

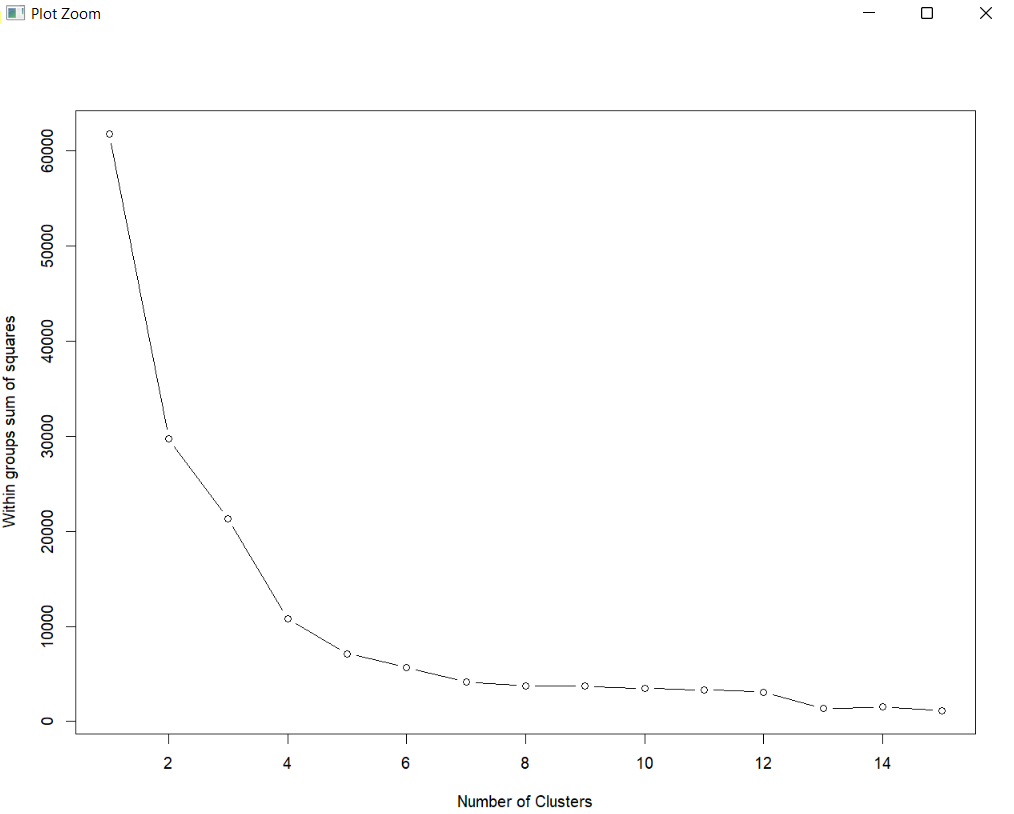
#Model-based clustering plots

library(mclust)

d.mclust<-Mclust(d)

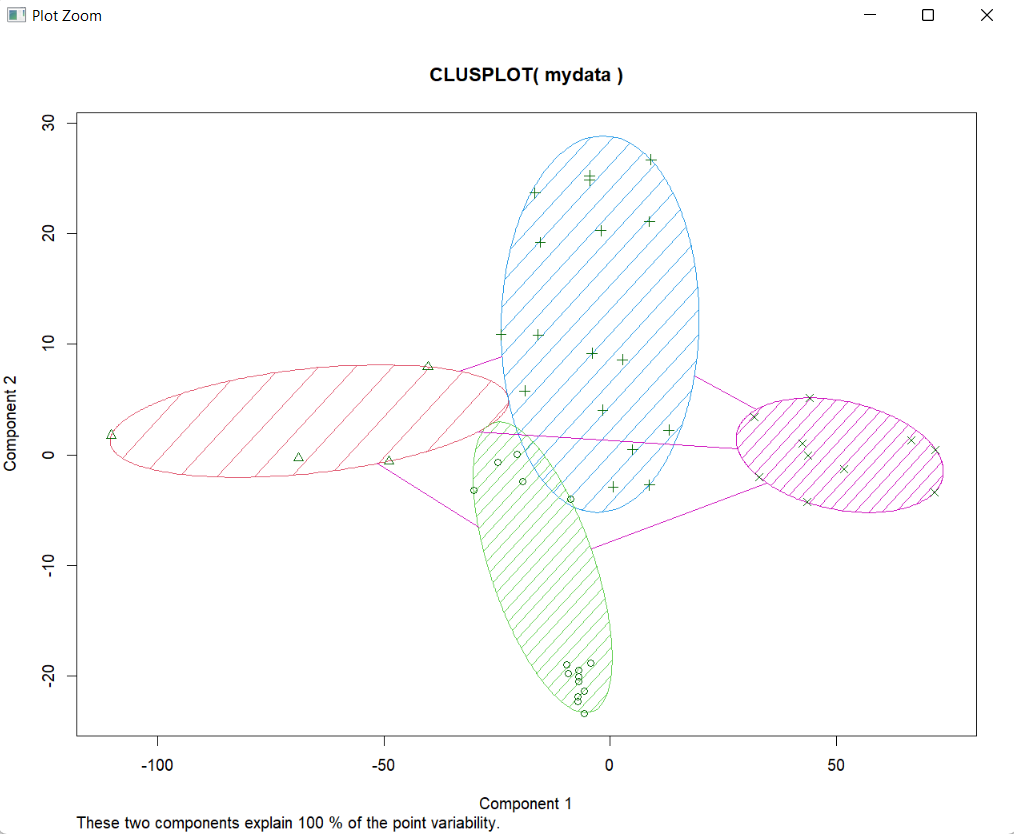
plot(d.mclust)

WSS method:



Elbow shows that the best no. Of clusters of the data will be at 4 clusters.

K-means clustering:

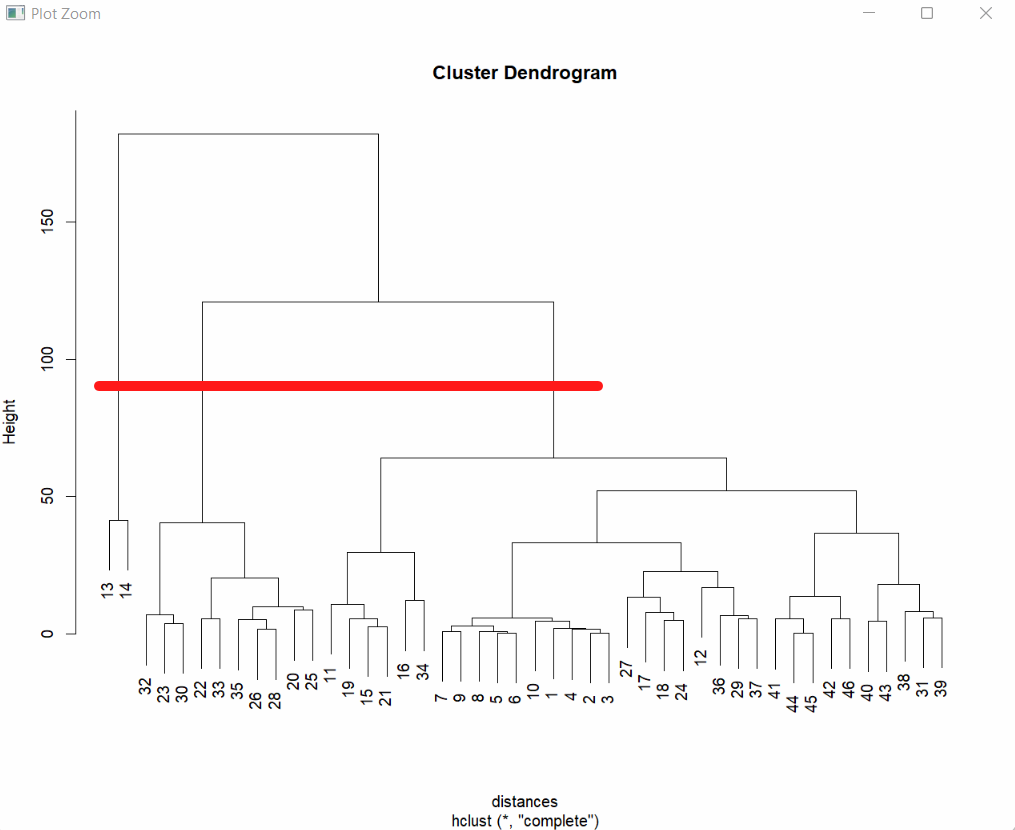


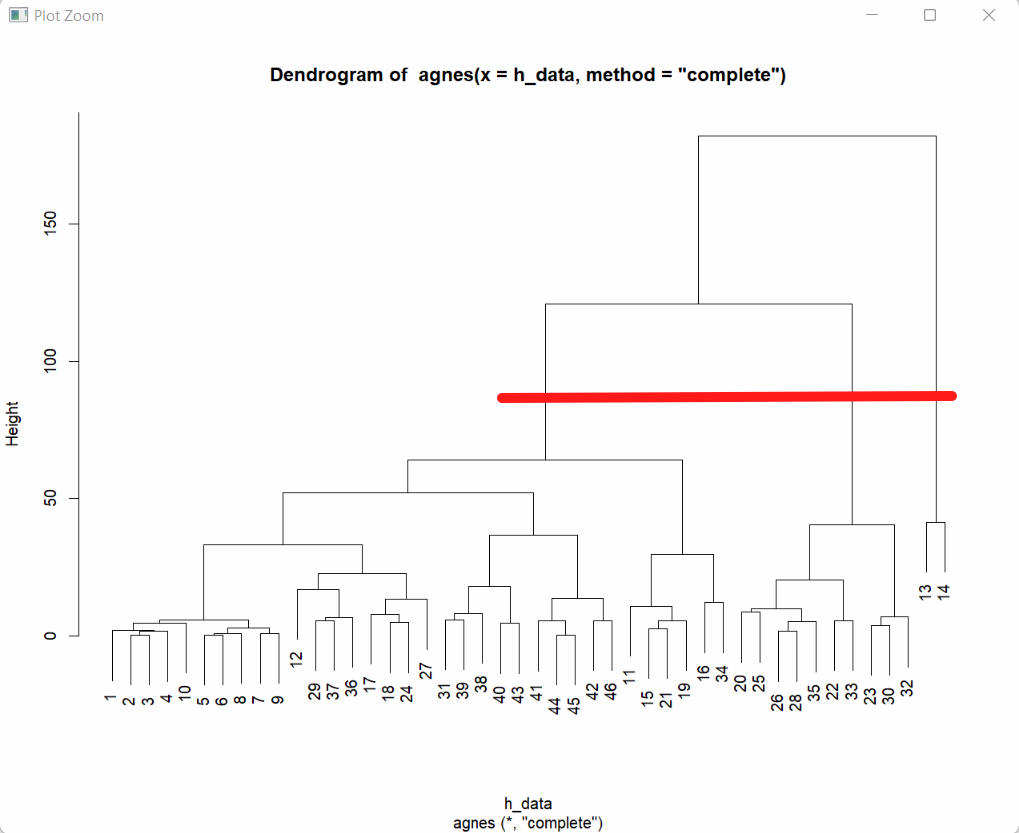
The figure of k-means above shows that the data has 4 clusters.

Hierarchical clustering:

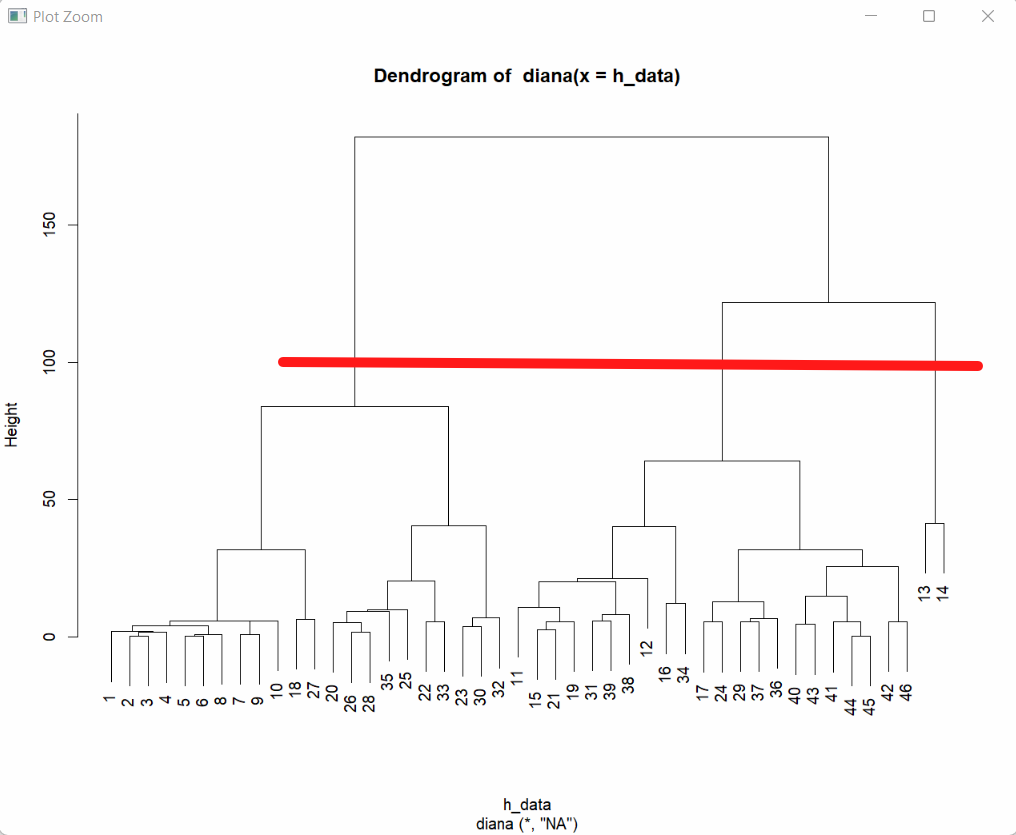
1)Agglomerative method:

1)complete linkage with hclust function



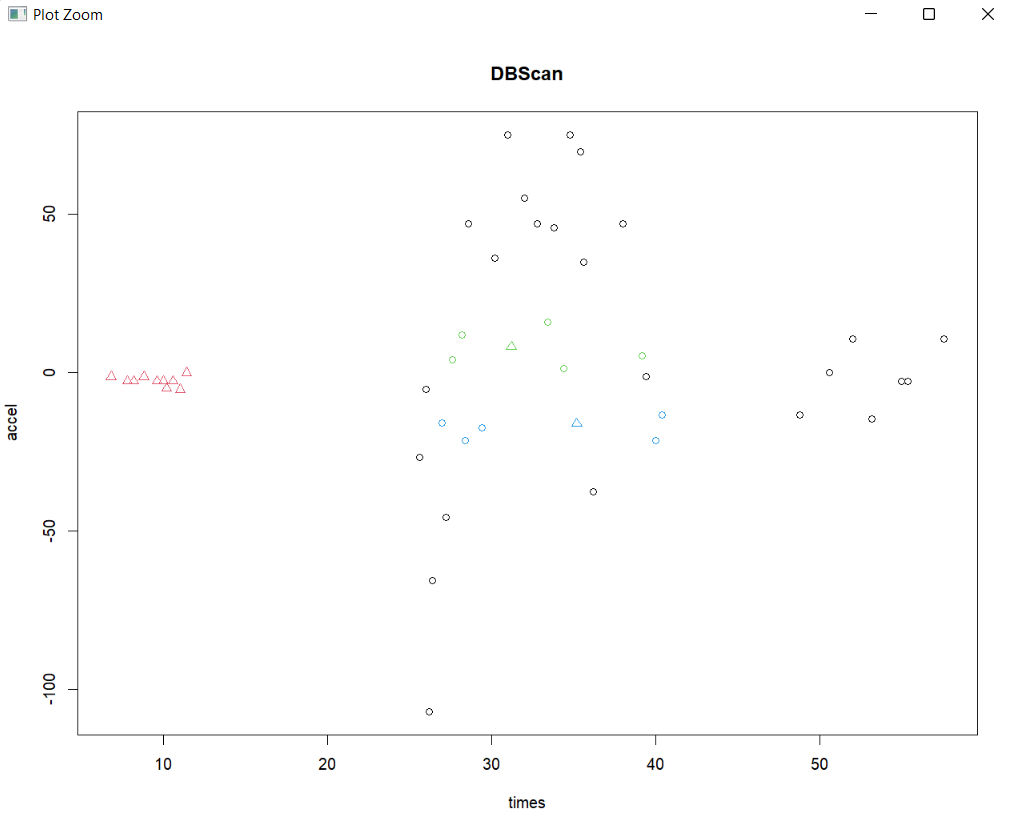
Complete method with Agnes method: 

* Agglomerative method with both hclust & agnes functions shows that there are 3 clusters.

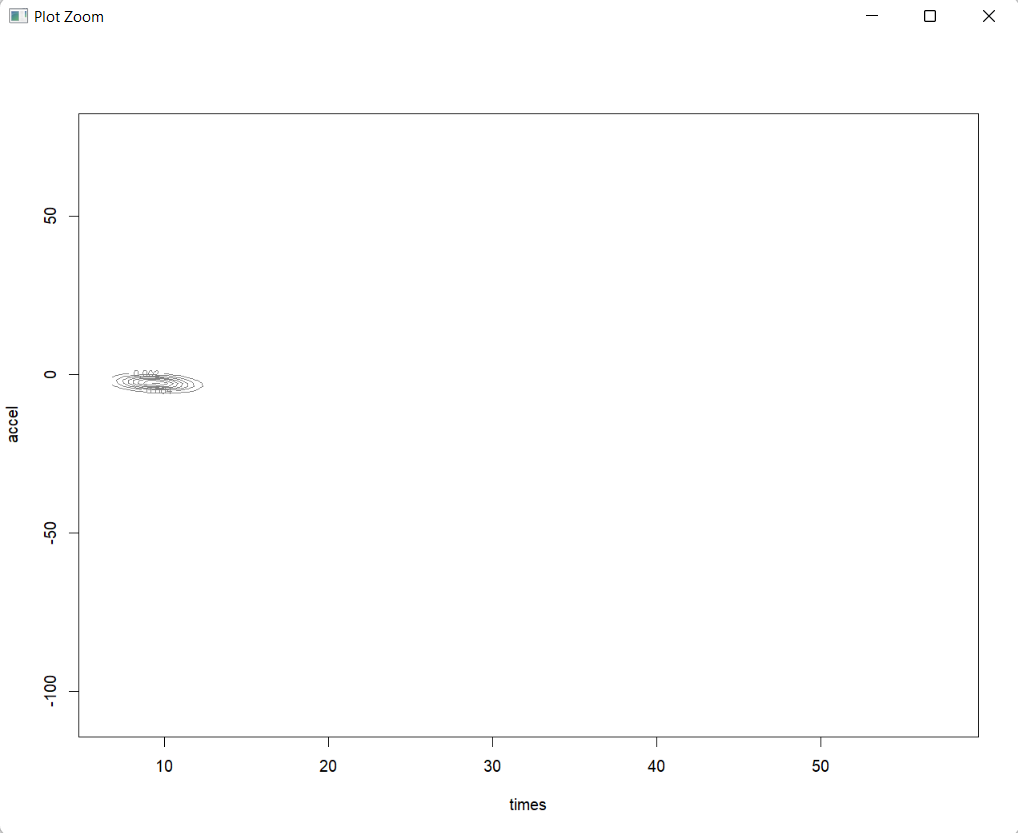
2) Divisive Method: 

Divisive method dendrogram shows that there are 3 clusters.

density-based clustering



Model-based clustering plots:



Advantages of K-Means :

1 – Simple to implement

2 – Easily adapts to new examples

3 – Can produce tighter clusters than hierarchal clustering

Disadvantages of K-Means:

1 – Choosing K means manually

2 – Dependent on initial values can produce different final clusters

3 – It is affected by outliers, we should consider removing them before clustering

Advantages of Hierarchal Clustering:

1. Simple to implement
2. Doesn’t need to pre specify the number of clusters
3. Dendrogram is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram

Disadvantages of Hierarchal Clustering:

1. Very sensitive to outliers
2. Once the instances have been assigned to a cluster they can’t no longer move around
3. High time complexity : Not suitable for large datasets

Advantages of Single linkage Method:

* Best of capturing clusters of different sizes
* Can deal with non-elliptical shapes

Disadvantages of Single linkage Method:

* Sensitive to outliers so it can’t group clusters properly

Advantages of Complete linkage Method:

* Does well in separating clusters if there is any noise between clusters

Disadvantages of Complete linkage Method:

* Tend to break large cluster
* Biased towards global clusters

Advantages of Average linkage Method:

* Does well in separating clusters if there is any noise between clusters

Disadvantages of Average linkage Method:

* Biased towards global clusters

Advantages of Density Based Algorithm:

* Doesn’t need to pre-specify the number of clusters
* Can identify noise data while cluster
* Sensitive to outliers

Disadvantages of Average linkage Method:

* Difficult to determine eps and min if the data isn’t well understood
* Cannot cluster data set with large differences in densities

Conclusion:

The data set has strata column which divide the data set into 3 categories so we already have information about the optimal number of clusters

From Hierarchal clustering we were able to identify the number of clusters which are 3 clusters from the dendogram with out pre-specification the number of clusters or finding the best parameters for the epsilon and minimum points in DBScan clustering algorithm