Perform clustering (Both hierarchical and K means clustering) for the airlines data to obtain optimum number of clusters.

Draw the inferences from the clusters obtained.

Hierarchical clustering

library(readxl)

## importing the data set

data <- read\_excel(choose.files())

View(data)

ncol(data)

##Removing the id coloum becase it does not effect our cluster results##

newdata <- data[,-1]

##As we performing clustering we should perform normalization of the Data set because the model depends on the euclineary distance##

nordata <- scale(newdata)

##Now we should find distance##

d <- dist(nordata,method = "euclidean")

str(d)

##Model##

Airline<- hclust(d,method = "complete")##H clustering model

plot(Airline,hang = -1)

rect.hclust(Airline,plot(Airline,hang = -1),k=5,border = "red")## rectangular cluster

groups <- cutree(Airline,k=5)##cutting the tree into 5 clusters

Air <- as.matrix(groups)##coverting the groups as the matrix form

Airfinal <- cbind(data,Air)

View(Airfinal)

table(Airfinal$Air)

write.csv(Airfinal, file="Airfinal.xls",row.names = F)

aggregate(data[,-1],by=list(Airfinal$Air),FUN=mean)

##I have divided the data set in 5 cluster because i have tryed some k values increase in in k does not effect the increase in number in the data points in the cluster so i took k value has 5

##The one has 3782 data point indicating that many of the data points near in one segement

##if want to target large number of customers we have to go with cluster one

##The least number of customer are in cluster 4 and 5

##here we can say most of data point are near in cluster one

##there are no dissimlarity in data points with distance not much far from each other

k-means

library(readxl)

## importing the data set

data <- read\_excel(choose.files())

View(data)

ncol(data)

##Removing the id coloum becase it does not effect our cluster results##

newdata <- data[,-1]

summary(newdata)

##As we performing clustering we should perform normalization of the Data set because the model depends on the euclineary distance##

nordata <- scale(newdata)

## to get accutarate k value

##creating a function with sum of the within cluster sum of squares

kmean\_withinss <- function(k) {

cluster <- kmeans(nordata, k)

return (cluster$tot.withinss)

}

##Testing the function with numerical 2 clusters

kmean\_withinss(2)

## Runing the algorithm n times

# Set maximum cluster

max\_k <-20

# Run algorithm over a range of k

wss <- sapply(2:max\_k, kmean\_withinss)

# Create a data frame to plot the graph

elbow <-data.frame(2:max\_k, wss)

library(ggplot2)

# Plot the graph with gglop

ggplot(elbow, aes(x = X2.max\_k, y = wss)) +

geom\_point() +

geom\_line() +

scale\_x\_continuous(breaks = seq(1, 20, by = 1))

## from this elbow plot we can say that k=5

model <- kmeans(nordata,5)

model$cluster## the data points belogs which cluster

model$centers##the cwentroid present in each cluster

model$withinss##with in sum of squre in it

##5024.9978 389.3969 7275.5673 7880.1659 4602.9266

model$tot.withinss

##25173.05

model$size

##156 15 1238 1104 1486

set.seed(2345)

library(animation)

kmeans.ani(nordata, 5)

* The cluster size in this k means spread then the hierarchical clustering
* So from the k means clustering you can more input compare to the hierarchical clustering
* Here smallest size cluster is 2
* So going with k-means clustering is better than the hierarchical clustering

