k\_means\_cluster.R

shiwam

2022-01-25

df=read.csv("Mall\_Customers.csv")  
head(df)

## CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.  
## 1 1 Male 19 15 39  
## 2 2 Male 21 15 81  
## 3 3 Female 20 16 6  
## 4 4 Female 23 16 77  
## 5 5 Female 31 17 40  
## 6 6 Female 22 17 76

summary(df)

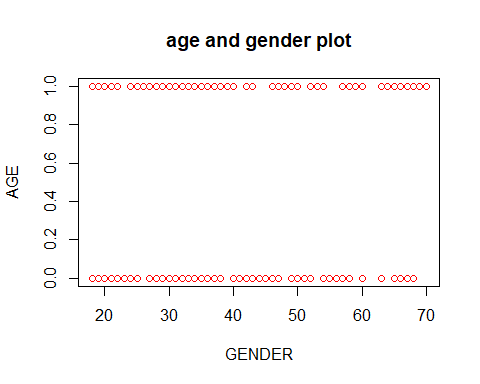
## CustomerID Gender Age Annual.Income..k..  
## Min. : 1.00 Length:200 Min. :18.00 Min. : 15.00   
## 1st Qu.: 50.75 Class :character 1st Qu.:28.75 1st Qu.: 41.50   
## Median :100.50 Mode :character Median :36.00 Median : 61.50   
## Mean :100.50 Mean :38.85 Mean : 60.56   
## 3rd Qu.:150.25 3rd Qu.:49.00 3rd Qu.: 78.00   
## Max. :200.00 Max. :70.00 Max. :137.00   
## Spending.Score..1.100.  
## Min. : 1.00   
## 1st Qu.:34.75   
## Median :50.00   
## Mean :50.20   
## 3rd Qu.:73.00   
## Max. :99.00

#CHANGING GENDER COLUMN AS BINARY VARIABLE  
df$Gender=as.factor(df$Gender)  
levels(df$Gender)=0:(length(df$Gender)-1)  
#here MALE=0, FEMALE=1  
df$Gender=as.numeric(as.character(df$Gender))  
#VISUALIZING AGE AND GENDER  
range(df$Age);range(df$Gender)

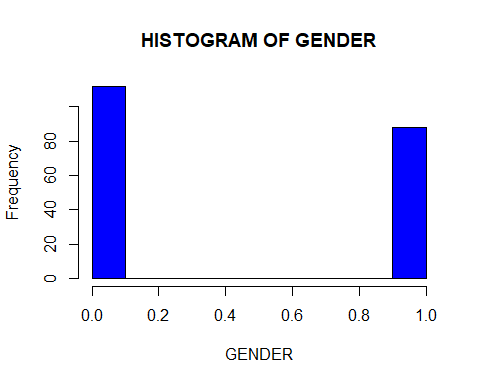
## [1] 18 70

## [1] 0 1

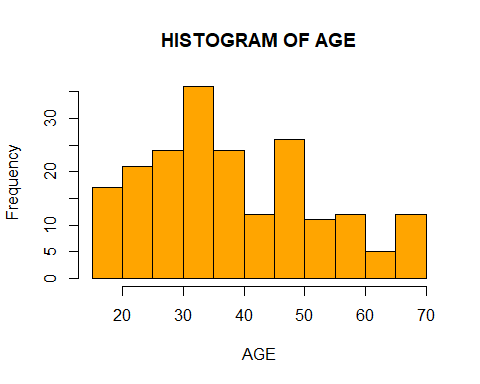
plot(df$Age,df$Gender,main = "age and gender plot",xlab = "GENDER",ylab = "AGE",col="red")



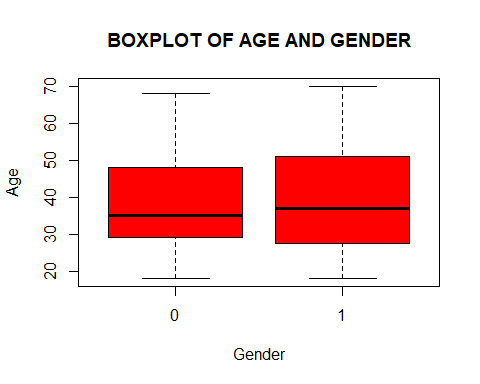
hist(df$Gender,main = "HISTOGRAM OF GENDER",xlab = "GENDER",col = "blue")



hist(df$Age,main = "HISTOGRAM OF AGE",xlab = "AGE",col = "orange")



boxplot(df$Age~df$Gender,ylab = "Age",xlab = "Gender",col="red",main="BOXPLOT OF AGE AND GENDER")



##########CLUSTER ANALYSIS########  
########K\_MEANS\_CLUSTERING(NON HIERARCHICAL APPROACH)######  
  
#AT FIRST FINDIING THE DISTANCE MATRIX OF THE DATASET  
#applying  
str(df)

## 'data.frame': 200 obs. of 5 variables:  
## $ CustomerID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Gender : num 1 1 0 0 0 0 0 0 1 0 ...  
## $ Age : int 19 21 20 23 31 22 35 23 64 30 ...  
## $ Annual.Income..k.. : int 15 15 16 16 17 17 18 18 19 19 ...  
## $ Spending.Score..1.100.: int 39 81 6 77 40 76 6 94 3 72 ...

mod=kmeans(df[,-c(1,2,3)],6)  
cc=mod$centers#cluster centroid  
cc

## Annual.Income..k.. Spending.Score..1.100.  
## 1 88.20000 17.114286  
## 2 86.53846 82.128205  
## 3 25.72727 79.363636  
## 4 31.53333 35.866667  
## 5 24.58333 9.583333  
## 6 56.05195 49.857143

rownames(cc)=c("centroid1","centroid2","centroid3","centroid4","centroid5","centroid6")  
cc

## Annual.Income..k.. Spending.Score..1.100.  
## centroid1 88.20000 17.114286  
## centroid2 86.53846 82.128205  
## centroid3 25.72727 79.363636  
## centroid4 31.53333 35.866667  
## centroid5 24.58333 9.583333  
## centroid6 56.05195 49.857143

#DISTANCE OF OBSERVATIONS FROM CLUSTER CENTROIDS   
#NOW MAKING A MATRIX OF SIZE= 200(OBSER.) x (6 clusters + 1 more column)  
#here 1 more column for clusterID(ON THIS COLUN WE SEE WHICH OBSERVATION IS ASSIGNED TO WHICH CLUSTER)  
  
  
#INITIALIZE (200 X 7) MATRIX  
DM= matrix(NA,200,7);head(DM)

## [,1] [,2] [,3] [,4] [,5] [,6] [,7]  
## [1,] NA NA NA NA NA NA NA  
## [2,] NA NA NA NA NA NA NA  
## [3,] NA NA NA NA NA NA NA  
## [4,] NA NA NA NA NA NA NA  
## [5,] NA NA NA NA NA NA NA  
## [6,] NA NA NA NA NA NA NA

rownames(DM)=c(1:200);head(DM)

## [,1] [,2] [,3] [,4] [,5] [,6] [,7]  
## 1 NA NA NA NA NA NA NA  
## 2 NA NA NA NA NA NA NA  
## 3 NA NA NA NA NA NA NA  
## 4 NA NA NA NA NA NA NA  
## 5 NA NA NA NA NA NA NA  
## 6 NA NA NA NA NA NA NA

colnames(DM)=c("clusterID","Dist\_clust1","Dist\_clust2","Dist\_clust3",  
 "Dist\_clust4","Dist\_clust5","Dist\_clust6")  
head(DM)

## clusterID Dist\_clust1 Dist\_clust2 Dist\_clust3 Dist\_clust4 Dist\_clust5  
## 1 NA NA NA NA NA NA  
## 2 NA NA NA NA NA NA  
## 3 NA NA NA NA NA NA  
## 4 NA NA NA NA NA NA  
## 5 NA NA NA NA NA NA  
## 6 NA NA NA NA NA NA  
## Dist\_clust6  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

#CLUSTER ASSIGNMENT  
DM[,1]=mod$cluster;head(DM)

## clusterID Dist\_clust1 Dist\_clust2 Dist\_clust3 Dist\_clust4 Dist\_clust5  
## 1 4 NA NA NA NA NA  
## 2 3 NA NA NA NA NA  
## 3 5 NA NA NA NA NA  
## 4 3 NA NA NA NA NA  
## 5 4 NA NA NA NA NA  
## 6 3 NA NA NA NA NA  
## Dist\_clust6  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

#COMPUTE DISTANCE (6 DISTANCE FOR 6 CENTROID)  
for (i in 1:200) {  
 DM[i,2]=dist(rbind(df[i,],cc[1,]),method = "euclidean")  
 DM[i,3]=dist(rbind(df[i,],cc[2,]),method = "euclidean")  
 DM[i,4]=dist(rbind(df[i,],cc[3,]),method = "euclidean")  
 DM[i,5]=dist(rbind(df[i,],cc[4,]),method = "euclidean")  
 DM[i,6]=dist(rbind(df[i,],cc[5,]),method = "euclidean")  
 DM[i,7]=dist(rbind(df[i,],cc[6,]),method = "euclidean")  
  
}  
head(DM)

## clusterID Dist\_clust1 Dist\_clust2 Dist\_clust3 Dist\_clust4 Dist\_clust5  
## 1 4 122.7895 158.8275 105.4343 52.87932 29.97001  
## 2 3 110.7351 150.2022 117.9998 71.28269 61.71468  
## 3 5 137.6999 170.3246 106.0756 57.27322 31.06769  
## 4 3 108.4447 148.5217 115.8531 67.66995 57.50392  
## 5 4 113.1823 151.2799 104.1583 49.17678 28.44647  
## 6 3 107.6152 147.4769 114.5348 66.04914 56.05832  
## Dist\_clust6  
## 1 91.08443  
## 2 91.51338  
## 3 101.24964  
## 4 88.72823  
## 5 84.00532  
## 6 87.33833

#ORDER BY CLUSTER ID  
head(DM[order(DM[,1]),])

## clusterID Dist\_clust1 Dist\_clust2 Dist\_clust3 Dist\_clust4 Dist\_clust5  
## 125 1 110.4523 125.3764 127.5127 106.1450 117.6755  
## 129 1 107.8910 122.3643 134.9075 114.5680 126.9549  
## 131 1 113.8651 127.5067 134.2590 114.4143 126.1568  
## 135 1 130.9450 141.7375 136.3243 118.8989 129.1964  
## 137 1 119.7691 132.3277 139.3029 120.6061 132.0500  
## 139 1 130.3080 140.7823 138.8990 121.8183 132.5074  
## Dist\_clust6  
## 125 97.31157  
## 129 100.96449  
## 131 103.66810  
## 135 114.28671  
## 137 110.11474  
## 139 115.51583

#DISTANCE BETWEEN CLUSTER CENTROID  
dist(cc,method = "euclidean",diag = T,upper = T)

## centroid1 centroid2 centroid3 centroid4 centroid5 centroid6  
## centroid1 0.00000 65.03515 88.19197 59.68888 64.06087 45.88673  
## centroid2 65.03515 0.00000 60.87400 71.87276 95.40019 44.39424  
## centroid3 88.19197 60.87400 0.00000 43.88276 69.78968 42.31098  
## centroid4 59.68888 71.87276 43.88276 0.00000 27.18669 28.22934  
## centroid5 64.06087 95.40019 69.78968 27.18669 0.00000 51.11021  
## centroid6 45.88673 44.39424 42.31098 28.22934 51.11021 0.00000

#SUMMARY OF CLUSTER DISTANCES  
DM2=matrix(NA,7,2);DM2

## [,1] [,2]  
## [1,] NA NA  
## [2,] NA NA  
## [3,] NA NA  
## [4,] NA NA  
## [5,] NA NA  
## [6,] NA NA  
## [7,] NA NA

rownames(DM2)=c("cluster1","cluster2","cluster3","cluster4","cluster5","cluster6","overall")  
colnames(DM2)=c("observation","Avg.dist\_in\_cluster");DM2#GIVES DISPERSION VALUE

## observation Avg.dist\_in\_cluster  
## cluster1 NA NA  
## cluster2 NA NA  
## cluster3 NA NA  
## cluster4 NA NA  
## cluster5 NA NA  
## cluster6 NA NA  
## overall NA NA

DM2[1:6,1]=mod$size;DM2

## observation Avg.dist\_in\_cluster  
## cluster1 35 NA  
## cluster2 39 NA  
## cluster3 22 NA  
## cluster4 15 NA  
## cluster5 12 NA  
## cluster6 77 NA  
## overall NA NA

DM2[7,1]=sum(DM2[1:6,1])  
for (i in 1:6) {  
 DM2[i,2]=mean(DM[which(DM[,1]==i),]);DM2  
 DM2[7,2]=mean(DM2[1:6,2]);DM2  
}  
  
head(DM2)

## observation Avg.dist\_in\_cluster  
## cluster1 35 127.06528  
## cluster2 39 123.82116  
## cluster3 22 77.50916  
## cluster4 15 67.82244  
## cluster5 12 77.32889  
## cluster6 77 76.30806

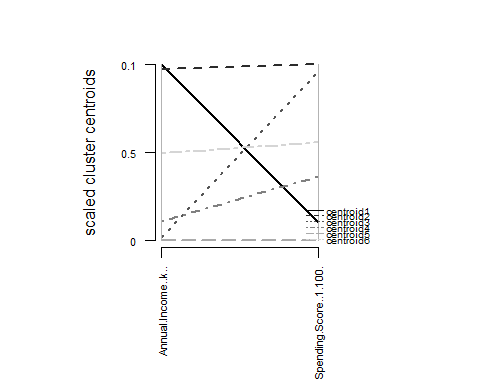
#LINE CHART OR PROFIT PLOT  
#PARALLEL COORDINATES PLOT(HELPS IN CHARACTERIZING THE CLUSTER)  
maxv=apply(cc,2, max)  
minv=apply(cc,2, min)  
cc.so1=as.data.frame(scale(cc,center =minv,scale = maxv-minv))  
cc.so1#here change the cluster as scaling

## Annual.Income..k.. Spending.Score..1.100.  
## centroid1 1.00000000 0.1038110  
## centroid2 0.97388203 1.0000000  
## centroid3 0.01798176 0.9618916  
## centroid4 0.10924810 0.3623045  
## centroid5 0.00000000 0.0000000  
## centroid6 0.49465991 0.5551572

library(MASS)

## Warning: package 'MASS' was built under R version 4.0.5

par(mar=c(0,0,0,0))  
par(mar=c(7.1,8.1,3,8.1),xpd=T,las=2,cex.axis=0.7)  
parcoord(cc.so1,col = gray(0:5/6),lty = c(1:6),lwd=2,las=3)  
axis(2,at=c(0,0.5,1),labels = c(0,0.5,0.1))  
title(ylab = "scaled cluster centroids")  
legend("bottomright",inset = c(-0.3,0)  
 ,c("centroid1","centroid2","centroid3","centroid4","centroid5","centroid6")  
 ,lty = c(1:6),cex=0.6,x.intersp = 0.4,y.intersp = 0.5,col = gray(0:5/6),bty = "n")



#plot cluster#2D REPRESENTATION OF CLUSTER  
library(cluster)

## Warning: package 'cluster' was built under R version 4.0.5

clusplot(df[,-c(1,2,3)],mod$cluster,main = "2D REPRESENTATION OF CLUSTER",color = T  
 ,labels = 2,lines = 0,shade = T)

