TrabajoFinalModulo5

María José Guzmán

2024-01-15

**Cargar paquetes**

library(openxlsx)  
library(stats)  
library(cluster)  
library("fpc")  
library("factoextra")

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(NbClust)  
library(htmltools)

## Cargar base

data <- read.xlsx("C:/Users/maria/Dropbox/Acádemico/Análisis de datos/Módulo 5/Base/parte 2/BOL\_BP\_MAY\_2017v2.xlsx")  
nombres <- data$BANCOS

*Modelo jerárquico*

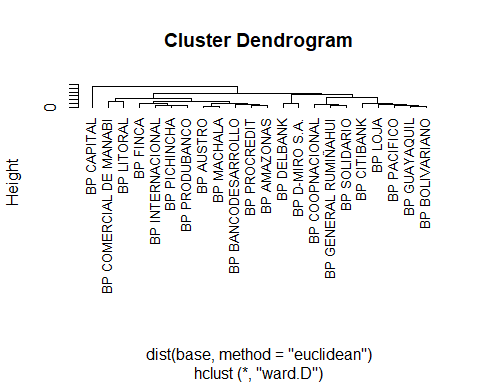
*Se escala la base para mantener a una sola escala todas las variables*

base <- as.data.frame(scale(data[,-1]))  
row.names(base) <- nombres

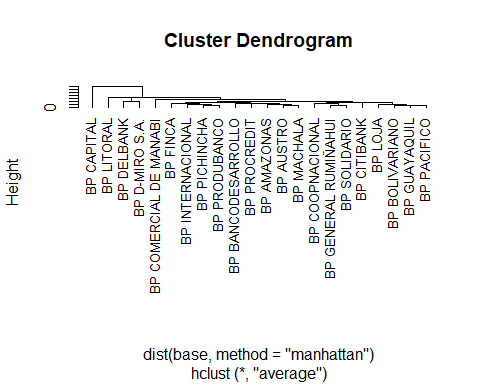
*Se especifica la distancia y el metodo de aglomeracion*

Donde para este caso se usa el método Euclidean con Ward,D mientras que para el otro se utiliza el método de Manhattan con el promedio

cluster <- hclust(dist(base, method = "euclidean"),  
 method = "ward.D")  
  
  
plot(cluster,hang = -0.01,cex=0.8)

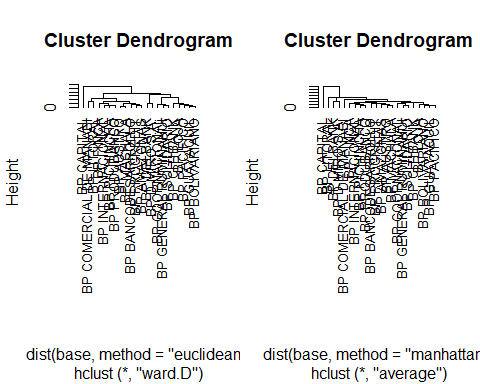


cluster2 <- hclust(dist(base, method = "manhattan"),  
 method = "average")  
  
  
plot(cluster2,hang = -0.01,cex=0.8)



**Compactar en 2 gráficos**

par(mfrow=c(1,2))  
plot(cluster,hang = -0.01,cex=0.8);plot(cluster2,hang = -0.01,cex=0.8)

 Realizando los gráficos con estas dos combinaciones de métodos y distancias se observa que la euclidea con el método de Ward.D es quien mejor realiza los cluster debido a que en el caso de promedios con Manhattan se observan que muchos bancos quedan agrupados juntos por lo que quiere decir que existen bancos muy similares y otros valores extremos que estan quedando solos.

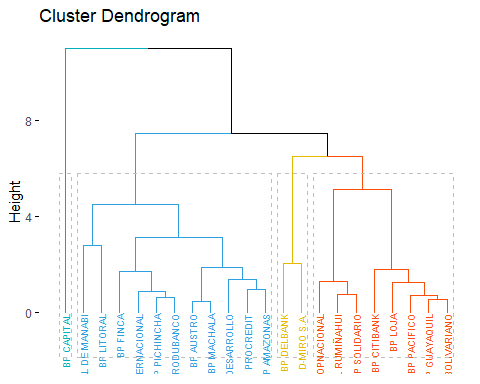
Tambien se realiza utilizando la función hcut, en este caso nos permite hacer otra combinación donde usamos entonces la que ya habíamos seleccionado de euclidean con Ward.D y la comparamos usando en h.metric una distancia basada en la correlación de pearson con el promedio.

res <- hcut(base, k = 4, stand = TRUE,   
 hc\_metric = "euclidean",hc\_method = "ward.D")  
res

##   
## Call:  
## stats::hclust(d = x, method = hc\_method)  
##   
## Cluster method : ward.D   
## Distance : euclidean   
## Number of objects: 22

fviz\_dend(res, rect = TRUE, cex = 0.5,  
 k\_colors = c("#00AFBB","#2E9FDF", "#E7B800", "#FC4E07"))

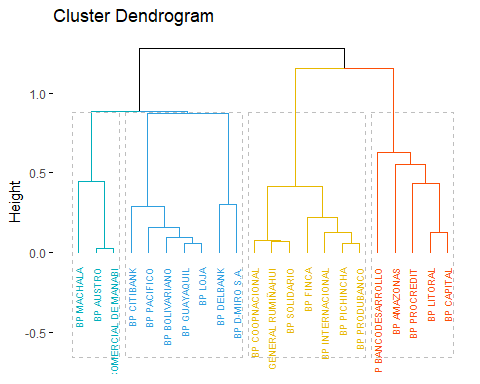
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as  
## of ggplot2 3.3.4.  
## ℹ The deprecated feature was likely used in the factoextra package.  
## Please report the issue at <https://github.com/kassambara/factoextra/issues>.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



res2 <- hcut(base, k = 4, stand = TRUE,   
 hc\_metric = "pearson",hc\_method = "average")  
res2

##   
## Call:  
## stats::hclust(d = x, method = hc\_method)  
##   
## Cluster method : average   
## Number of objects: 22

fviz\_dend(res2, rect = TRUE, cex = 0.5,  
 k\_colors = c("#00AFBB","#2E9FDF", "#E7B800", "#FC4E07"))



Al observar ambos gráficos se observa que realmente usando la correlacion como medida de distancia se agrupan mejor los bancos.

*Módelo no jerárquico*

**Media de cada cluster**

cnj <- kmeans(base,4)

aggregate(base,  
 by=list(cnj$cluster),  
 FUN=mean)

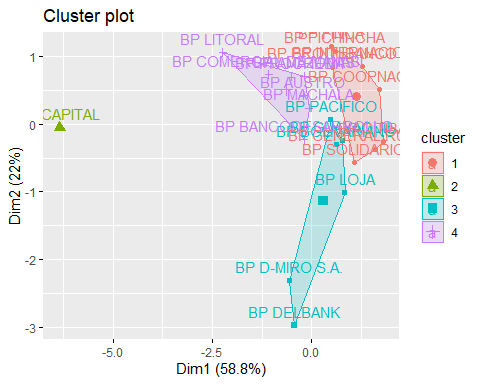
## Group.1 AP/TA MCT GO/MF RE/AP FD/TDCP  
## 1 1 0.8091827 -0.5882752 -0.4052557 0.6355750 -0.391010  
## 2 2 -1.8369098 3.5087289 4.0388074 -2.9704862 -0.379306  
## 3 3 -0.3375191 0.0679189 -0.3129738 0.3894900 1.128600  
## 4 4 -0.3730624 0.1128513 0.1544401 -0.6358648 -0.466316

cnj$centers

## AP/TA MCT GO/MF RE/AP FD/TDCP  
## 1 0.8091827 -0.5882752 -0.4052557 0.6355750 -0.391010  
## 2 -1.8369098 3.5087289 4.0388074 -2.9704862 -0.379306  
## 3 -0.3375191 0.0679189 -0.3129738 0.3894900 1.128600  
## 4 -0.3730624 0.1128513 0.1544401 -0.6358648 -0.466316

**Visualizando el cluster**

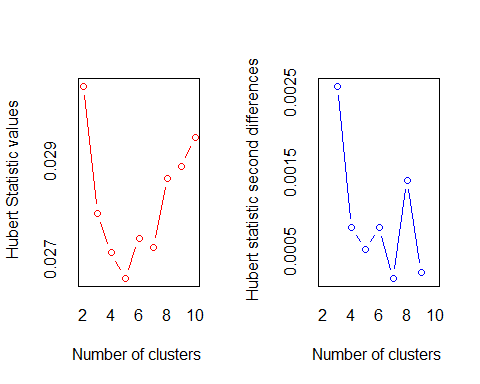
fviz\_cluster(cnj,data=base)



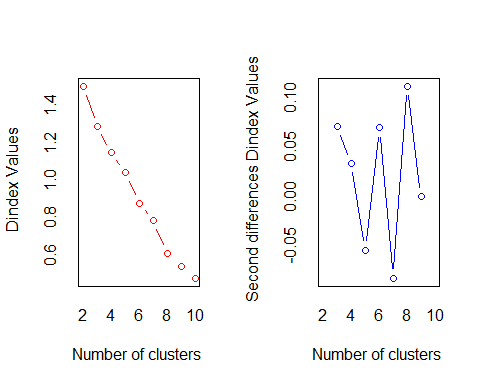
**Clusters óptimos**

clustersoptimos <- NbClust(base,  
 distance = "euclidean",  
 min.nc = 2,  
 max.nc = 10,  
 method="average",  
 index = "all")

## Warning in pf(beale, pp, df2): NaNs produced



## \*\*\* : The Hubert index is a graphical method of determining the number of clusters.  
## In the plot of Hubert index, we seek a significant knee that corresponds to a   
## significant increase of the value of the measure i.e the significant peak in Hubert  
## index second differences plot.   
##



## \*\*\* : The D index is a graphical method of determining the number of clusters.   
## In the plot of D index, we seek a significant knee (the significant peak in Dindex  
## second differences plot) that corresponds to a significant increase of the value of  
## the measure.   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## \* Among all indices:   
## \* 9 proposed 2 as the best number of clusters   
## \* 6 proposed 3 as the best number of clusters   
## \* 1 proposed 6 as the best number of clusters   
## \* 2 proposed 7 as the best number of clusters   
## \* 3 proposed 8 as the best number of clusters   
## \* 1 proposed 9 as the best number of clusters   
## \* 2 proposed 10 as the best number of clusters   
##   
## \*\*\*\*\* Conclusion \*\*\*\*\*   
##   
## \* According to the majority rule, the best number of clusters is 2   
##   
##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Evaluando la clusterizacion**

La silueta es una medida que indica como de bien estan agrupados o asignados lo cluster este comprendida entre -1 y 1 la idea es que esta cercano a 1. En el siguiente cuadro se puede observar que esta sobre 0.66 lo que indica que es cercano a uno para el cluster 1.

cnj <- kmeans(base,2)  
silueta <- silhouette(cnj$cluster,dist(base,method="euclidean"))  
   
fviz\_silhouette(silueta)

## cluster size ave.sil.width  
## 1 1 1 0.00  
## 2 2 21 0.66

