Análisis K- medias

Antonio Hernández

2022-05-31

CAPITULO I

INTRODUCCIÓN

Un K-medias es un análisis mutivariante y mineria de datos que consiste en el agrupamiento de variables que presentan caracteristicas similares, estos estarán juntos en un mismo grupo y los que no comparte se separarán en otro grupo, para conocer si si los datos son similares o diferentes el algoritmo K-medias utiliza la distancia entre ellos, para tener una facil compresión del tema y de manera simple las observaciones que se parecen tendran una menor distancia entre sí, como medida tomaremos la distancia euclidia, aunque esta no es la unica, pero es la que mejor se deseempeña. La manera de trabajar este análisi es buscar patrones en los datos prediccion especifica ya que no hay una variable dependiente. La fragmentación de la población a partir de un número de K clusters, el algoritmo coloca los centroides (k puntos aleatorizados), se asigna en cada uno de los puntos la distancias pequeñas, las medias juegan un papel importante en este análisis ya que la media de las muestras más cercanas y esto genera nuevas asignación de muestras y esta esta más cerca de otro centroide, se repite iterativamente y los grupos se van ajustando.

Este análisis es útil cuando se tiene un gran número de casos.

MATRIZ DE DATOS

La matriz de datos para el análisis en cuestion es sustraida del paquete estadístico R, pero ¿Qué es lo que contiene esta base de datos?, bueno este dataset contiene como observaciones los estados de la unión americana, y como variables se encuentra los delitosn tales como, los asaltos, muertes,violaciones y la población del estado.

Exploración de la matriz

```
X<- USArrests
dim(X)
```

```
## [1] 50  4
```

```
anyNA(X)
```

```
## [1] FALSE
```

Como se puede observar la base elegida para este estudio tiene 4 variables cuantitativas que son las muertes, asaltos, población y violaciones, con 50 observaciones que son los estados de EUA, y no cuenta con datos faltantes.

Tratamiento de la matríz.

#Separacion de filas y columnas.

```
dim(X)
```

```
## [1] 50  4
```

```
n<-dim(X)[1]
p<-dim(X[2])
```

2.- Estandarizacion univariante.

```
X<-scale(X)
```

CAPITULO II

METODOLOGIA

Para dicho estudio se realizará un análisis multivariante de clostering o K medias, el cuál agrupará a los estados conforme a sus similitudes que existe entre ellos, para esto se debe realizar una exploración de la matriz para cerciorarse que no haya ningún dato faltante ya que si existen el análisis se vera afectado y no se podrá realizar, ya que se exploró prosigue estandarizar los datos para que ninguna de las variabes del dataset tenga un mayor peso con respecto a las demás, calcular las distancias euclidias de las observaciónes, hayar el número de closter recomedados por la función "silhouette", ya teniendo lo anterior se grafican los closters y se conoceran cuales son los estados que se alejan más de los demás estados.

RESULTADOS

1.- Cargar las librerias a utilizar.

```
library(tidyr)
library(cluster)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.7     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

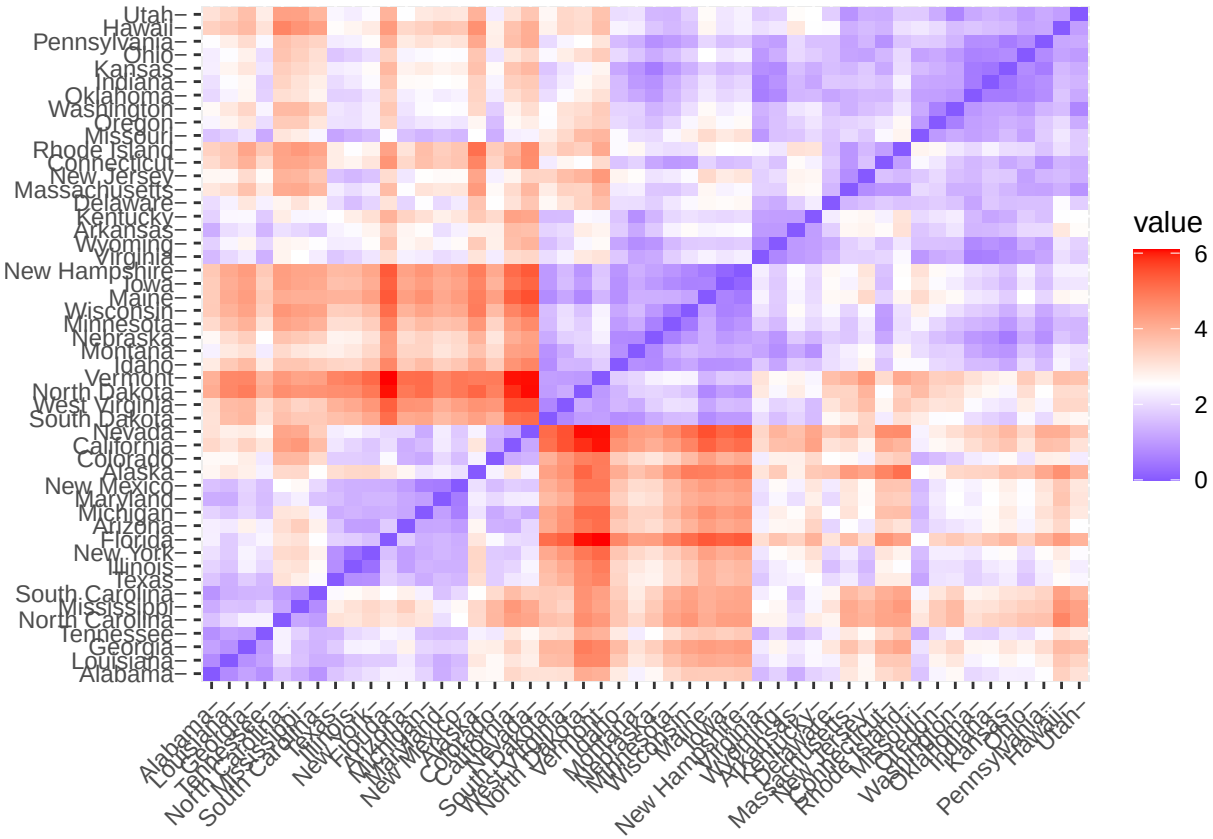
```
library(factoextra)
```

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

```
library(NbClust)
```

2.-Calcular la matriz de distacias.

```
m.distancia <- get_dist(X, method = "euclidean")
fviz_dist(m.distancia, gradient = list(low = "blue", mid = "white", high = "red"))
```



Como se puede observar hay variable con ciertas tendencias.

3.- Algoritmo k-medias (3 grupos)cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.

```
Kmeans.4<-kmeans(X, 2, nstart=25)
```

4.- centroides

```
Kmeans.4$centers
```

```
##      Murder    Assault   UrbanPop       Rape
## 1 -0.669956 -0.6758849 -0.1317235 -0.5646433
## 2  1.004934  1.0138274  0.1975853  0.8469650
```

5.- cluster de pertenencia.

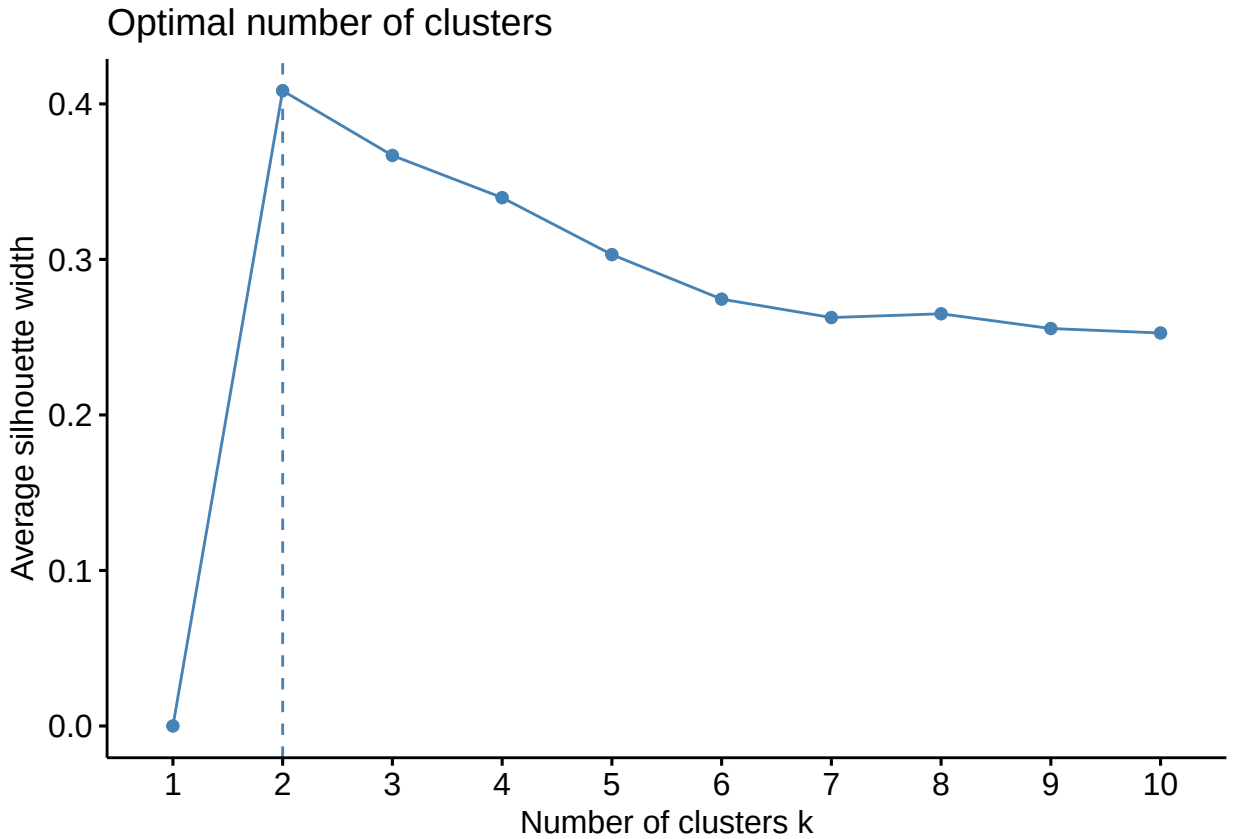
```
Kmeans.4$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              2              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              2              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              2              1              1
```

```
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              2              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              2              1              2              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              2              1              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              2              1              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              2              2              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              1              1              1
```

6.- Obterner el número de closter posibles con el metodo de la silueta.

```
fviz_nbclust(X, kmeans, method = "silhouette")
```



Con la grafica se nos da a conocer que el posible número de cluster es 2, ya que la linea punteada es la que nos indica el número.

7.- Calculo de la suma de los cuadrados generales.

```
SCDG<-sum(Kmeans.4$withinss)
SCDG
```

```
## [1] 102.8624
```

8.- Clusters

```
cl.kmeans<-Kmeans.4$cluster
cl.kmeans
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              2              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              2              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              2              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              2              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              2              1              2              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              2              1              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              2              1              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              2              2              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              1              1              1
```

9.-calculamos los dos clústers

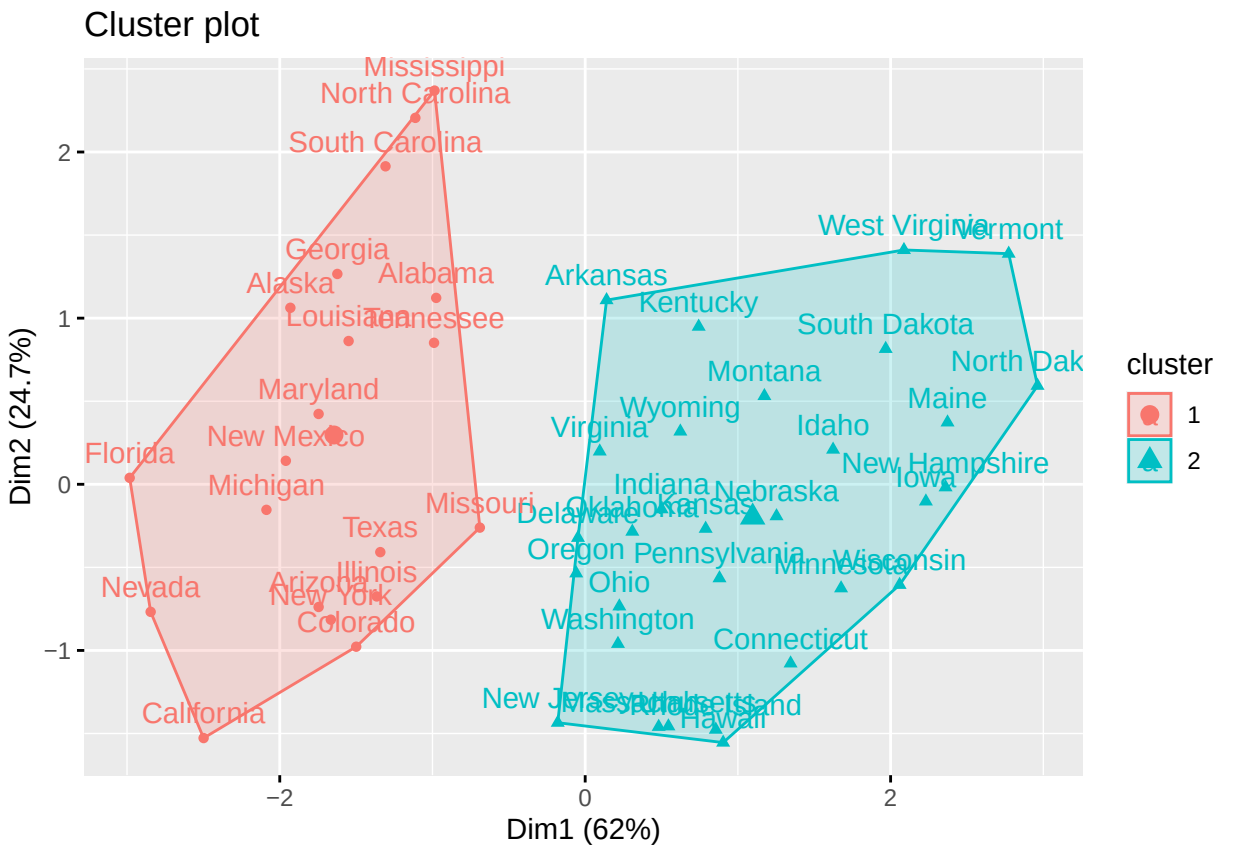
```
k2 <- kmeans(X, centers = 2, nstart = 25)
k2
```

```
## K-means clustering with 2 clusters of sizes 20, 30
##
## Cluster means:
##      Murder    Assault   UrbanPop       Rape
## 1  1.004934  1.0138274  0.1975853  0.8469650
## 2 -0.669956 -0.6758849 -0.1317235 -0.5646433
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              1              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
```

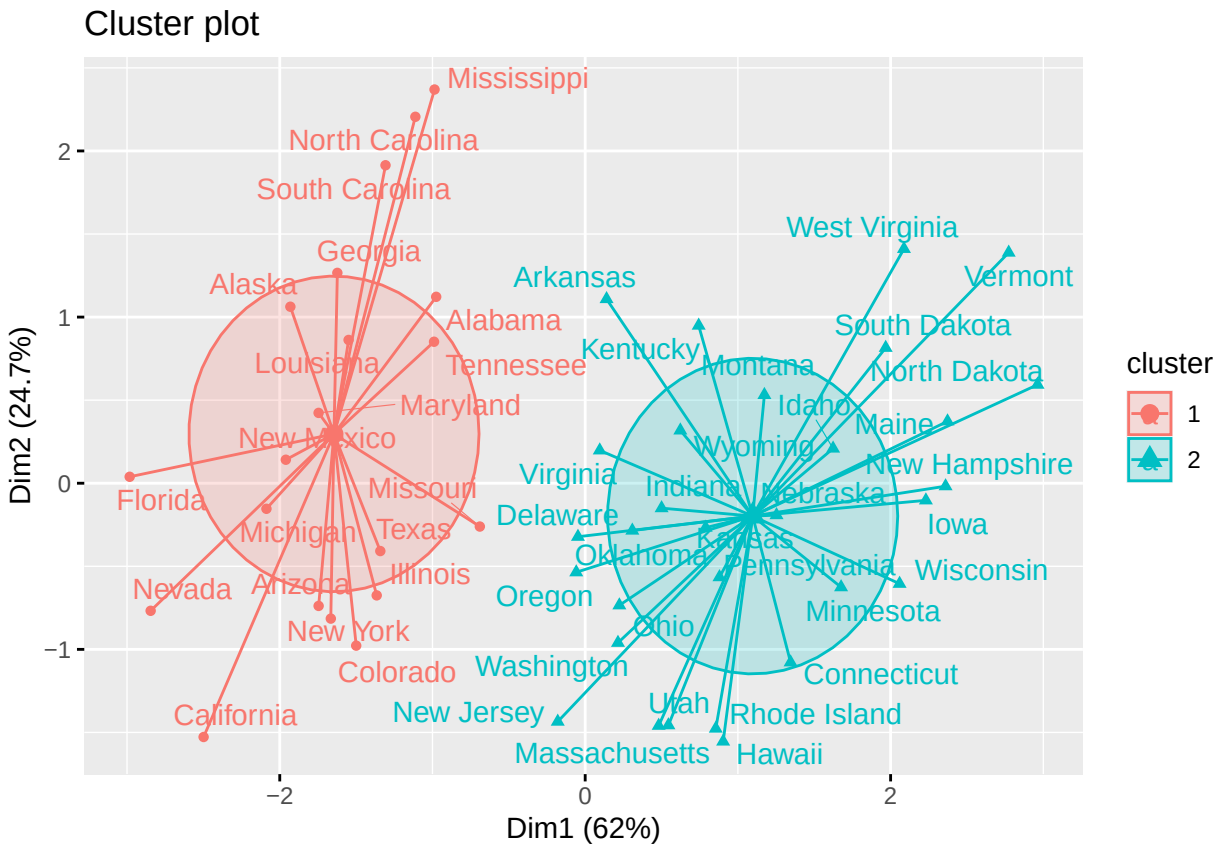
```
##              2             2             1             2             1
##  Massachusetts      Michigan     Minnesota   Mississippi      Missouri
##              2             1             2             1             1
##        Montana      Nebraska        Nevada New Hampshire    New Jersey
##              2             2             1             2             2
##     New Mexico      New York North Carolina  North Dakota          Ohio
##              1             1             1             2             2
##       Oklahoma        Oregon  Pennsylvania  Rhode Island South Carolina
##              2             2             2             2             1
##   South Dakota     Tennessee         Texas          Utah       Vermont
##              2             1             1             2             2
##       Virginia    Washington West Virginia     Wisconsin       Wyoming
##              2             2             2             2             2
##
## Within cluster sum of squares by cluster:
## [1] 46.74796 56.11445
##  (between_SS / total_SS =  47.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

10.- Graficar los cluster

```
fviz_cluster(k2, data = X)
```

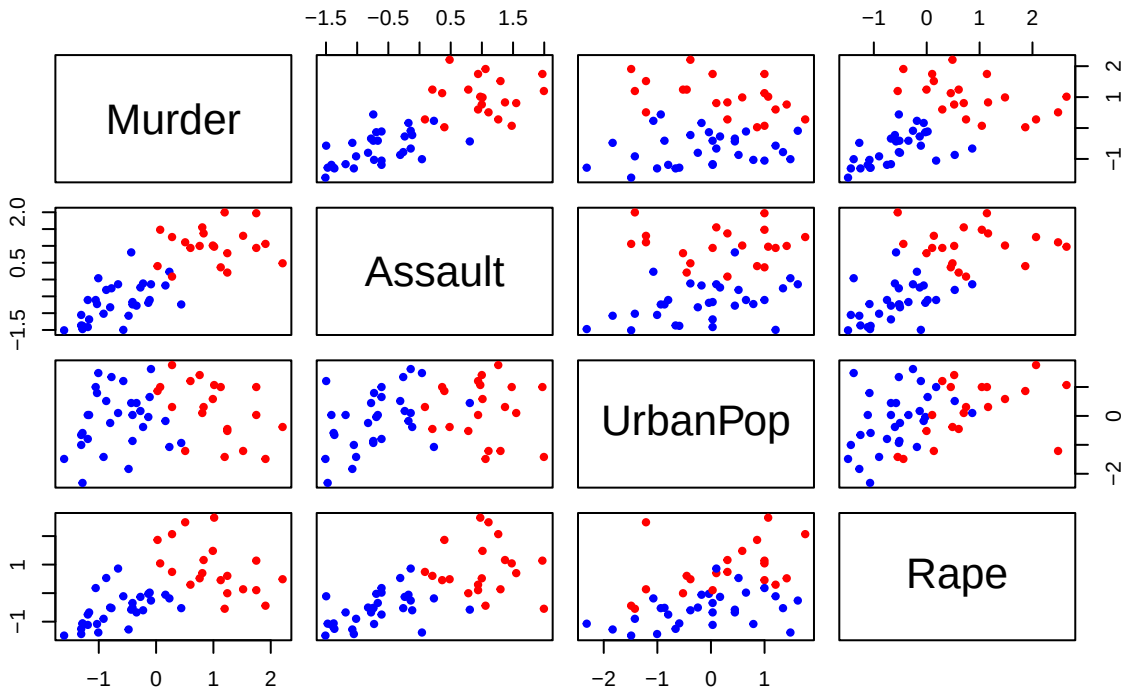


```
fviz_cluster(k2, data = X, ellipse.type = "euclid",repel = TRUE,star.plot = TRUE)
```



11.- Scatter plot con la division de grupos obtenidos (se utiliza la matriz de datos centrados).

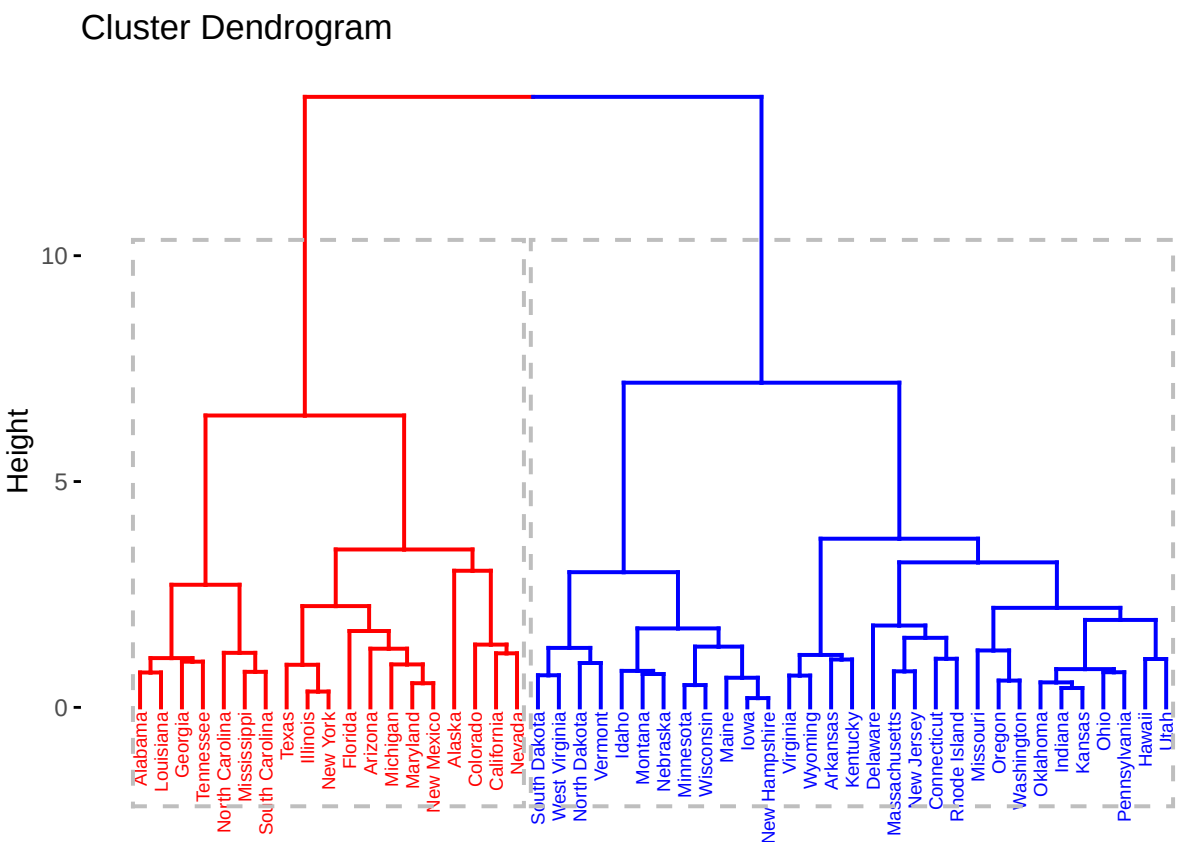
```
col.cluster<-c("blue", "red", "green")[cl.kmeans]
pairs(X, col=col.cluster, main="k-means", pch=20)
```



12.- Creación de un dendrograma para dicidir los dos closter y ver cuales son los estados que pertenecen a los dos diferentes closters.

```
res2 <- hcut(X, k = 2, stand = TRUE)
fviz_dend(res2, rect = TRUE, cex = 0.5,
          k_colors = c("red","blue"))
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



CAPITULO III

Concluciónes

Dentro del análisis se utilizó la distancia que existe entre los datos llamada “euclidia” la cual es resultado de la implementación del teorema de pitágoras, esto con fin de obtener la matriz de distancia para la elaboración de los closters más adelante, el calculo de los centroides por medio de la suma de los cuadrados generales en este caso es de 102.8624, el closter de pertenencia nos ayudara para porteriormente elaborar la compración con respecto a otro. Para obtener el número posible de closters es necesario ocupar la función silhouette o silueta en español, la cual recibe como datos principales la base de datos estandarizados, y se le especifica el cuál es el análisis que se esta realizando, “K- medias”, este es uno de los diferentes metodos para obtener el número recomendado de closter ya que existen wss y gap_stat que se encuentra en la paqueteria NbClust, 2 closters fue lo que indico la función.

Ya teniendo todos los calculos, es momento de generar los closter de manera grafica , lo cual nos dice que el closter rojo almacena 20 estados de los cuales se alejan de los demás estados con una distancia similares son Mississippi, California, le sigue Carolina del Norte, Carolina del Sur y Nevada, más sin embargo el closter azul tiene un mayor número de estados que se alejan de los datos con una menor distancia entre ellas, pero de manera drástica dichas entidades son Vermont, West Virgina, Dakota del Norte, Arkansas, Dakota del Sur, Massachusetts.

El scatter plot nos da de manera grafica los datos centrados, por ultimó se observa por medio de un dendrograma para conocer cuales son los estados que comparten similitudes entre ellos.

Referencias

JM Marin. (.).El procedimiento Conglomerados de K medias. En Análisis de Conglomerados(462-473) Estrategias de Trading. (2019). K-Means Clustering: Agrupamiento con Minería de datos. 6 junio 2022, de

Estrategias de Trading Sitio web: https://estrategiastrading.com/k-means/

La matriz de datos precargada en R.

Las paqueterias fueron: tidyr cluster tidyverse factoextra NbClust