Análisis Clúster para estudiar los delitos en Estados Unidos

Daniel Sol Piedra

31/05/2022

Introduccón

En la cultura popular, incluso la más reciente, recogida en series de gran éxito como "The Wire", la sociedad estadounidense a menudo aparece plagada por la violencia y el crimen. Sin embargo, las estadísticas reflejan una mejora sustancial en la seguridad ciudadana durante la última década. De acuerdo con los datos del Departamento de Justicia, los crímenes violentos cayeron este año a su nivel más bajo desde 1973.

La tendencia a la baja en el número de homicidios y actos violentos registrados por la policía se ha mantenido durante los dos últimos años contradiciendo las previsiones de muchos expertos, que se temían un estallido del crimen a causa de la crisis económica que vive el país, la peor desde la Gran Depresión de los años treinta.

La reducción del crimen es especialmente espectacular en las tres mayores ciudades del país, y donde la sensación de inseguridad era mayor. En las dos últimas décadas, el número de homicidios cayó en Nueva York un 79%, pasando de las 2.245 muertes violentas del año 1990, a las 471 del 2009. En Chicago, la reducción es del 48%, pasando de 850 muertes a 458. Por su parte, en Los Ángeles, los homicidios cayeron un 68%, de 983 a 312.

Estos cambios han provocado que la seguridad ciudadana haya desaparecido en las encuestas como una de las principales prioridades de los votantes, siendo reemplazada por nuevas demandas de seguridad ante otro tipo de amenaza: el terrorismo. De hecho, en las pasadas elecciones legislativas, apenas si se oyó hablar de seguridad ciudadana en la campaña electoral.

Según los expertos, la caída del crimen se debe a varios factores. Por un lado, una mejor tecnología y metodología por parte de las autoridades policiales ha permitido un uso más eficiente de los recursos disponibles, y la aplicación de exitosas medidas de prevención. Además, se han llevado a cabo ambiciosas campañas contra algunas bandas de crimen organizado, a la vez que se apostaba por nuevos programas sociales destinados a los jóvenes de barrios marginales.

Sin embargo, algunos factores no responden directamente a una política de las autoridades policiales. Por ejemplo, el boom económico de los años 90 ofreció nuevas posibilidades de trabajo a muchos jóvenes que podrían haberse decantado por la delincuencia ante la falta de expectativas. Igualmente, las luchas intestinas de hoy en día entre las mafias de la droga nada tienen que ver con las auténticas guerras motivadas por el control del "crack" en los años 80.

La gran pregunta que se hacen muchos expertos es si estas mejoras son irreversibles, o en el futuro puede haber un repunte de la violencia. William Bratton, un especialista en seguridad que trabajó para los departamentos de policía de Nueva York y Los Angeles se muestra optimista sobre el futuro. “No veo [los homicidios y el crimen violento] volviendo a las cifras que vimos en el pasado”, sostiene Bratton, que considera que será clave evitar la llegada una nueva epidemia de droga como la del crack.

Descripción de la matriz de datos Este base de datos es del año de 1973 y contiene estadísticas en arrestos por cada 100,000 habitantes por agrasión, asesinatos y violación en cada uno de los 50 estados de EE.UU. Tambien se proporciona el porcentaje de la población que vive en esas áreas urbanas. Por último esta base de datos se obtuvo del software estadíatico R.

Exploración de la matriz de datos La base de datos a utilizar es de dimension 50 X 4 en donde se encuentran 50 observaciones de diversos estados de EE.UU y contiene 4 variables, como Asesinatos, asaltos, violación y porcentaje de población. Las cuatro variables son cuantitativas aunque de diferente escala, por lo que en el análisis se tienen que normalizar. No presenta Na´s o espacios vacíos.

Tratamiento de la matriz de datos Se tienen que normalizar los datos para hacer posible la construccion del análisis clúster.

Metodología de análisis

Para llevar a cabo el analisis clúster se toma la base de datos antes mencionada y se utilizan las librerias tales como tidyverse, cluster, factoextra, NbClust y tidr. El análisis clúster se llevo de la siguiente manera:

1. Se cargan las librerias antes mencionadas para poder extraer la base de datos.
2. Una vez cargadas las librerias lo que sigue es cargar la base de datos correspondiente.
3. Como ya se mencionó, se normalizan las puntuaciones, ya que las variables de estudio están en diferente escala, es decir las variables no miden la misma caracteristica.
4. Se crea un dendograma para poder apreciar visualmente cuantas agrupaciones es posble crear.
5. Una vez creado el dendograma lo que sigue es utilizar el método de distancias euclidianas,para clacular la matriz de distancias.
6. Mediante diversos métodos se logra crear el número de agrupaciones adecuada, ya que la mayoria de esos métodos proponen el mismo número adecuado.
7. Ya conociendo el numero de las agrupaciones lo qu sigue es calcularlas, mediante el método de K-medias.
8. Se realiza una representación gráfica de esas agrupaciones, en donde ya se puede apreciar las caracteríaticas de interés.

9. Como último paso, se añaden las 50 observaciones a la agrupacion que le corresponda.

Resultados

Análisis Clúster

Cargamos las librerías necesarias para poder ejecutar nuestro analisis Clúster.

```
ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}

packages <- c("tidyverse","cluster", "factoextra","NbClust","tidyr")
ipak(packages)

## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyvers
e 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conf
licts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: cluster

## Loading required package: factoextra

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

## Loading required package: NbClust

##  tidyverse    cluster factoextra    NbClust      tidyr
##       TRUE       TRUE       TRUE       TRUE       TRUE
```

Visualizamos la base de datos que trae R Project en donde se analizan delitos, numero de muertos, entre otros, en diversas partes de los Estados Unidos.

```
##                Murder Assault UrbanPop Rape
## Alabama          13.2     236       58 21.2
## Alaska           10.0     263       48 44.5
```

```
## Arizona            8.1     294       80 31.0
## Arkansas           8.8     190       50 19.5
## California         9.0     276       91 40.6
## Colorado           7.9     204       78 38.7
## Connecticut        3.3     110       77 11.1
## Delaware           5.9     238       72 15.8
## Florida           15.4     335       80 31.9
## Georgia           17.4     211       60 25.8
## Hawaii             5.3      46       83 20.2
## Idaho              2.6     120       54 14.2
## Illinois          10.4     249       83 24.0
## Indiana            7.2     113       65 21.0
## Iowa               2.2      56       57 11.3
## Kansas             6.0     115       66 18.0
## Kentucky           9.7     109       52 16.3
## Louisiana         15.4     249       66 22.2
## Maine              2.1      83       51  7.8
## Maryland          11.3     300       67 27.8
## Massachusetts      4.4     149       85 16.3
## Michigan          12.1     255       74 35.1
## Minnesota          2.7      72       66 14.9
## Mississippi       16.1     259       44 17.1
## Missouri           9.0     178       70 28.2
## Montana            6.0     109       53 16.4
## Nebraska           4.3     102       62 16.5
## Nevada            12.2     252       81 46.0
## New Hampshire      2.1      57       56  9.5
## New Jersey         7.4     159       89 18.8
## New Mexico        11.4     285       70 32.1
## New York          11.1     254       86 26.1
## North Carolina    13.0     337       45 16.1
## North Dakota       0.8      45       44  7.3
## Ohio               7.3     120       75 21.4
## Oklahoma           6.6     151       68 20.0
## Oregon             4.9     159       67 29.3
## Pennsylvania       6.3     106       72 14.9
## Rhode Island       3.4     174       87  8.3
## South Carolina    14.4     279       48 22.5
## South Dakota       3.8      86       45 12.8
## Tennessee         13.2     188       59 26.9
## Texas             12.7     201       80 25.5
## Utah               3.2     120       80 22.9
## Vermont            2.2      48       32 11.2
## Virginia           8.5     156       63 20.7
## Washington         4.0     145       73 26.2
## West Virginia      5.7      81       39  9.3
## Wisconsin          2.6      53       66 10.8
## Wyoming            6.8     161       60 15.6
```

Como podemos darnos cuenta estan en diferente escala nuestras variables, por lo que es necesario normalizarlas para poder hacer su respectivo análsisis.

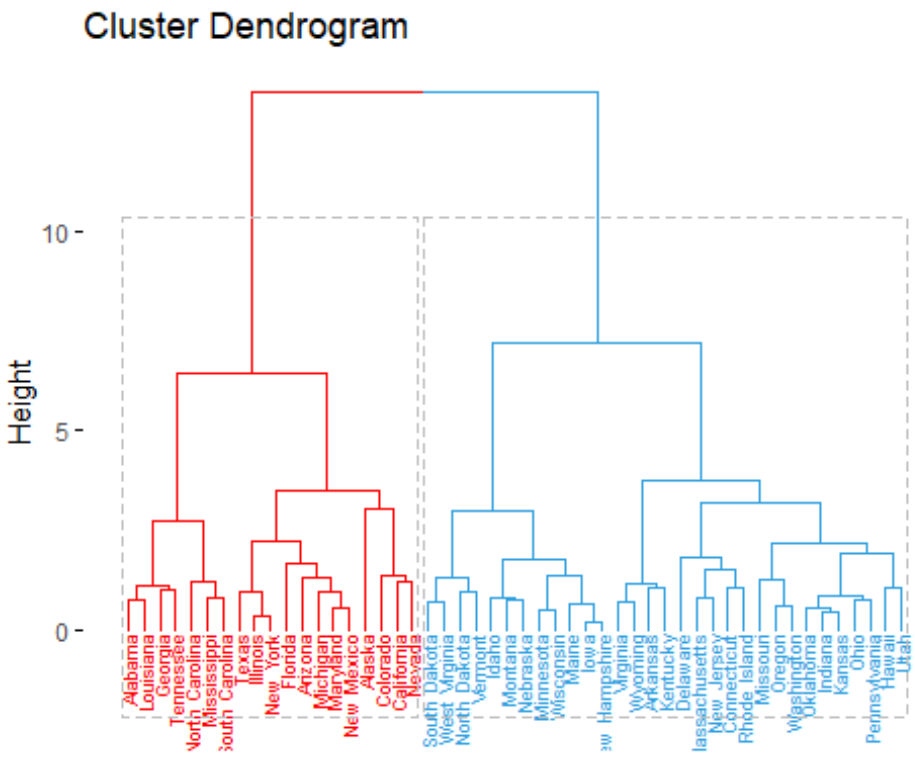
```
#normalizar las puntuaciones
df <- scale(df)
head(df)
```

```
##                Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
## California 0.27826823 1.2628144  1.7589234  2.067820292
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

Creamos el dendograma para poder ver las agrupaciones que se podrian formar

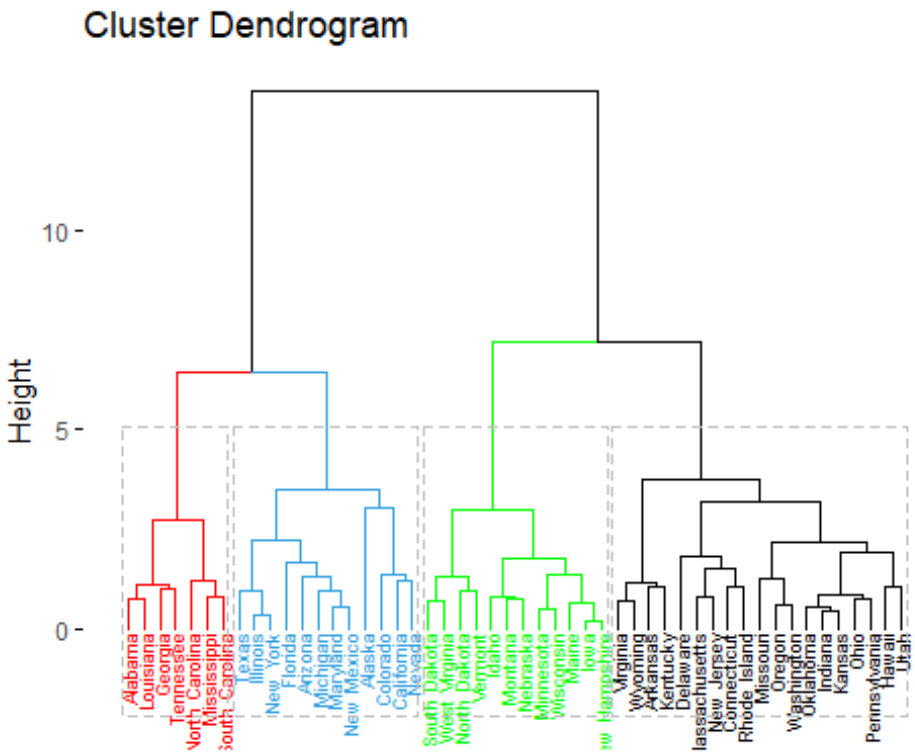
```
res2 <- hcut(df, k = 2, stand = TRUE)
fviz_dend(res2, rect = TRUE, cex = 0.5,
          k_colors = c("red","#2E9FDF"))
```

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




```
res4 <- hcut(df, k = 4, stand = TRUE)
fviz_dend(res4, rect = TRUE, cex = 0.5,
          k_colors = c("red","#2E9FDF","green","black"))
```

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

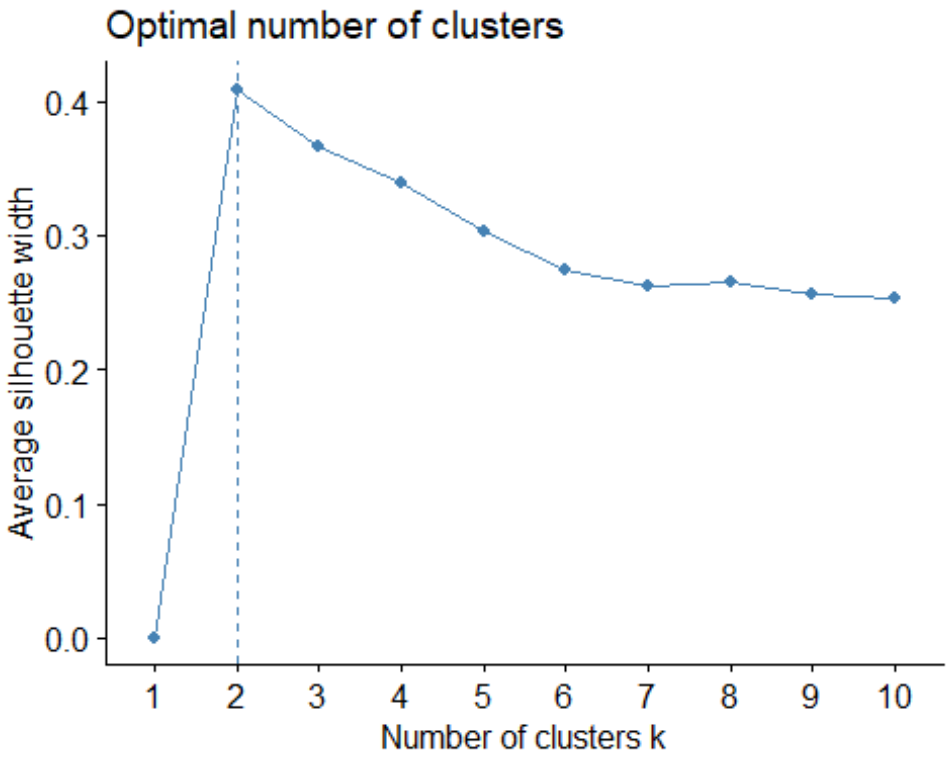


Es momento de calcular la matriz de diatancias en donde se usa el metodo de distancias euclidianas. Existen diversos métodos pero en este caso utilizaremos la antes mencionada

```
#calcular la matriz de distacias
m.distancia <- get_dist(df, method = "euclidean") #el método aceptado tam
bién puede ser: "maximum", "manhattan", "canberra", "binary", "minkowski"
, "pearson", "spearman" o "kendall"
```

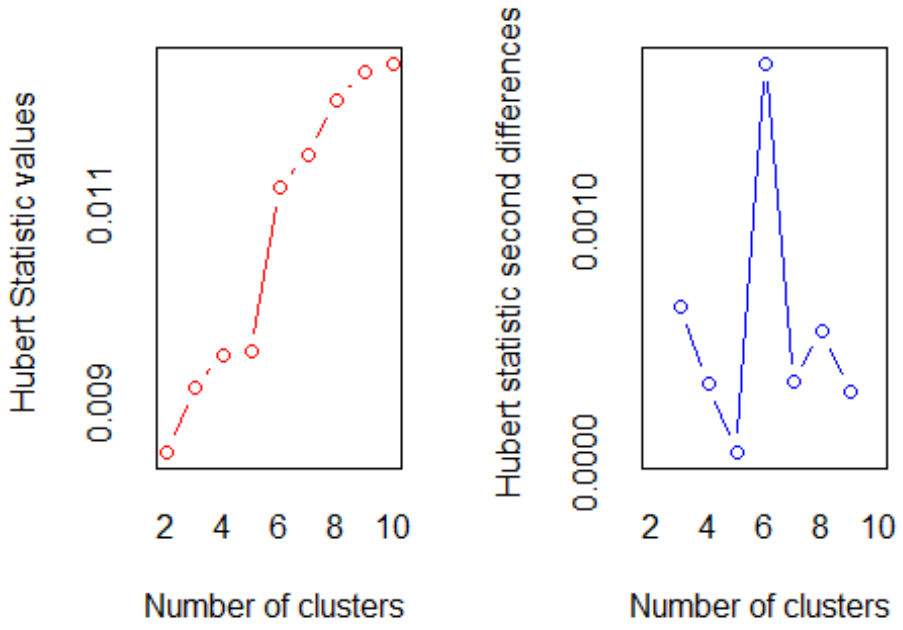
Se calcula el numero de clusters o agrupaciones adecuadas mediante el método slihouette, además de que existen diversos métodos.

```
#estimar el número de clústers
fviz_nbclust(df, kmeans, method = "silhouette")
```

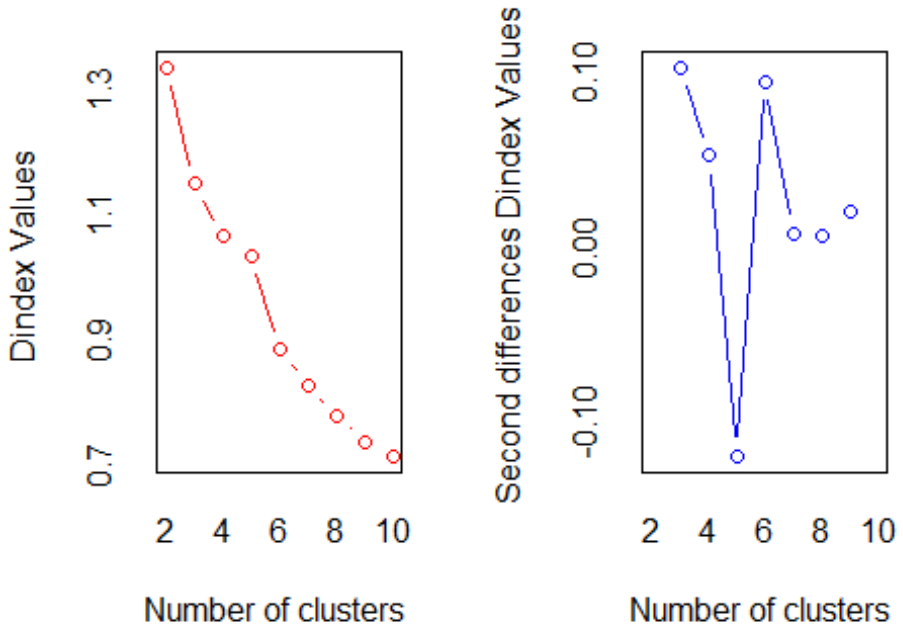


Con la siguiente función se calculan las agrupaciones adecuadas mediante diversos métodos, al final se nos imprime cual es el numero mas adecuado segun los métodos.

```
#con esta función se pueden calcular:
#the index to be calculated. This should be one of : "kl", "ch", "hartiga
n", "ccc", "scott",
#"marriot", "trcovw", "tracew", "friedman", "rubin", "cindex", "db", "sil
houette", "duda",
#"pseudot2", "beale", "ratkowsky", "ball", "ptbiserial", "gap", "frey", "
mcclain", "gamma",
#"gplus", "tau", "dunn", "hubert", "sdindex", "dindex", "sdbw", "all" (al
l indices except GAP,
#Gamma, Gplus and Tau), "alllong" (all indices with Gap, Gamma, Gplus and
Tau included).
resnumclust<-NbClust(df, distance = "euclidean", min.nc=2, max.nc=10, met
hod = "kmeans", index = "alllong")
```




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




```
## *** : The D index is a graphical method of determining the number of c
lusters.
##                 In the plot of D index, we seek a significant knee (th
e significant peak in Dindex
##                 second differences plot) that corresponds to a signifi
cant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 13 proposed 2 as the best number of clusters
## * 2 proposed 3 as the best number of clusters
## * 1 proposed 4 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 7 proposed 6 as the best number of clusters
## * 1 proposed 9 as the best number of clusters
## * 3 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************

fviz_nbclust(resnumclust)
```

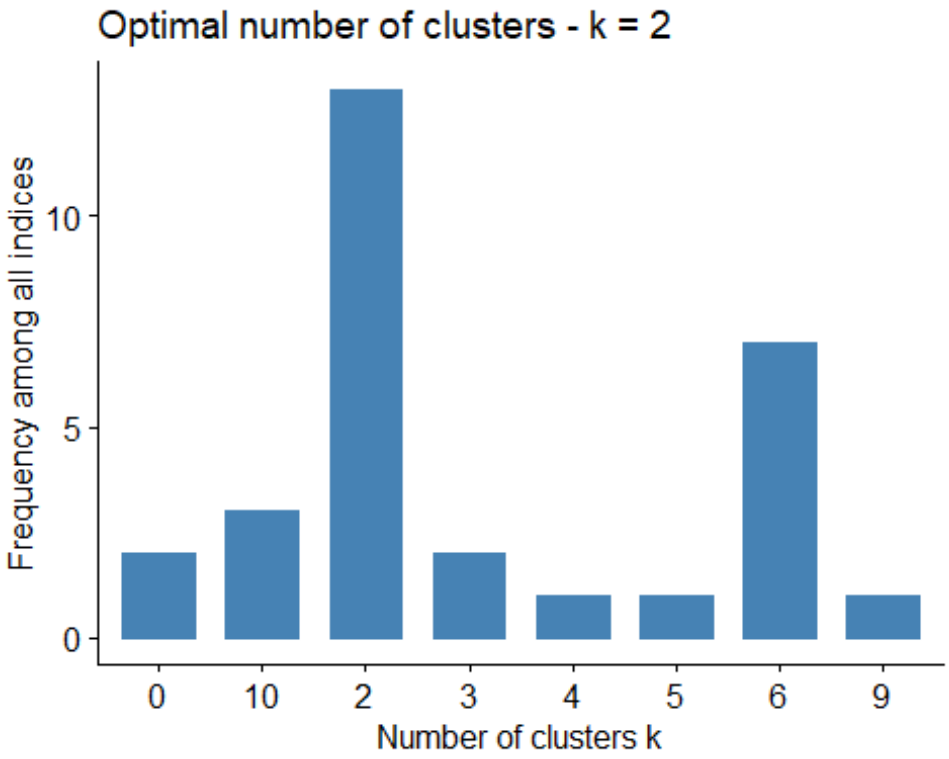
```
## Warning in if (class(best_nc) == "numeric") print(best_nc) else if
## (class(best_nc) == : la condición tiene longitud > 1 y sólo el primer
elemento
## será usado

## Warning in if (class(best_nc) == "matrix") .viz_NbClust(x, print.summa
ry, : la
## condición tiene longitud > 1 y sólo el primer elemento será usado

## Warning in if (class(best_nc) == "numeric") print(best_nc) else if
## (class(best_nc) == : la condición tiene longitud > 1 y sólo el primer
elemento
## será usado

## Warning in if (class(best_nc) == "matrix") {: la condición tiene longi
tud > 1 y
## sólo el primer elemento será usado

## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 13 proposed  2 as the best number of clusters
## * 2 proposed  3 as the best number of clusters
## * 1 proposed  4 as the best number of clusters
## * 1 proposed  5 as the best number of clusters
## * 7 proposed  6 as the best number of clusters
## * 1 proposed  9 as the best number of clusters
## * 3 proposed  10 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  2 .
```



Como se puede observar, el número de agrupaciones adecuadas es 2, ahora lo que sigue es calcular esas dos agrupaciones.

```
#calculamos los dos clústers
k2 <- kmeans(df, 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     Califo
rnia
##              1              1              1              2
1
##       Colorado    Connecticut       Delaware        Florida        Geo
rgia
##              1              2              2              1
1
##         Hawaii          Idaho       Illinois        Indiana
Iowa
##              2              2              1              2
2
```

```
##         Kansas       Kentucky      Louisiana          Maine       Mary
land
##              2              2              1              2
1
##  Massachusetts       Michigan      Minnesota    Mississippi       Miss
ouri
##              2              1              2              1
1
##        Montana       Nebraska         Nevada  New Hampshire     New Je
rsey
##              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 Caro
lina
##              2              2              2              2
1
##   South Dakota      Tennessee          Texas           Utah        Ver
mont
##              2              1              1              2
2
##       Virginia     Washington  West Virginia      Wisconsin        Wyo
ming
##              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.w
ithinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

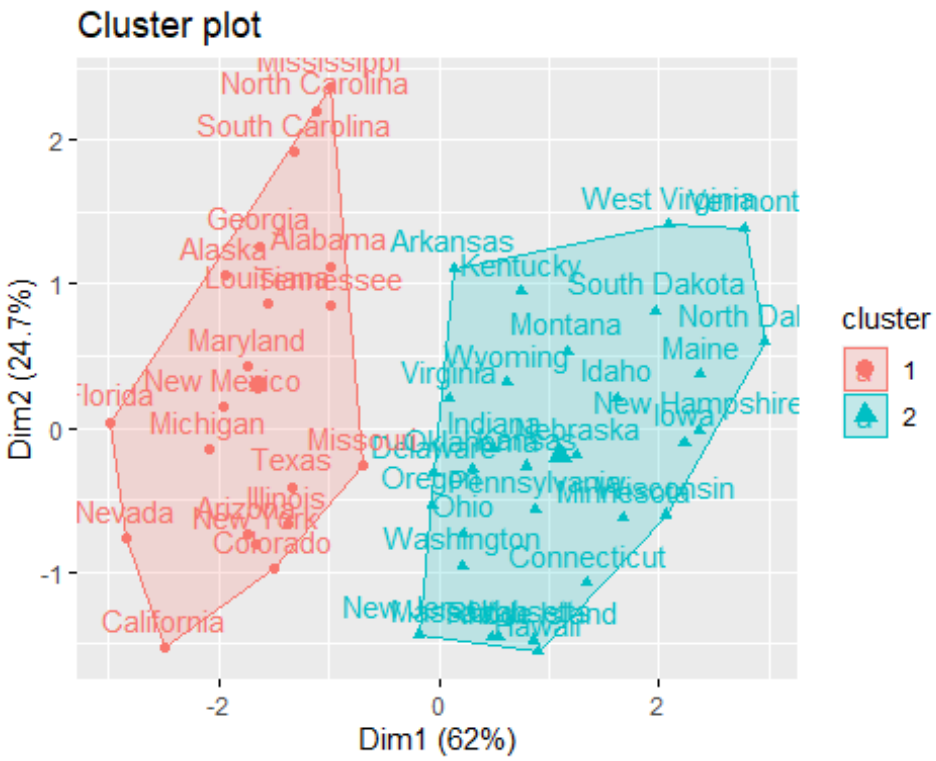
```
str(k2)
```

```
## List of 9
##  $ cluster     : Named int [1:50] 1 1 1 2 1 1 2 2 1 1 ...
##   ..- attr(*, "names")= chr [1:50] "Alabama" "Alaska" "Arizona" "Arkan
sas" ...
##  $ centers     : num [1:2, 1:4] 1.005 -0.67 1.014 -0.676 0.198 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ totss       : num 196
```

```
##  $ withinss    : num [1:2] 46.7 56.1
##  $ tot.withinss: num 103
##  $ betweenss   : num 93.1
##  $ size        : int [1:2] 20 30
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

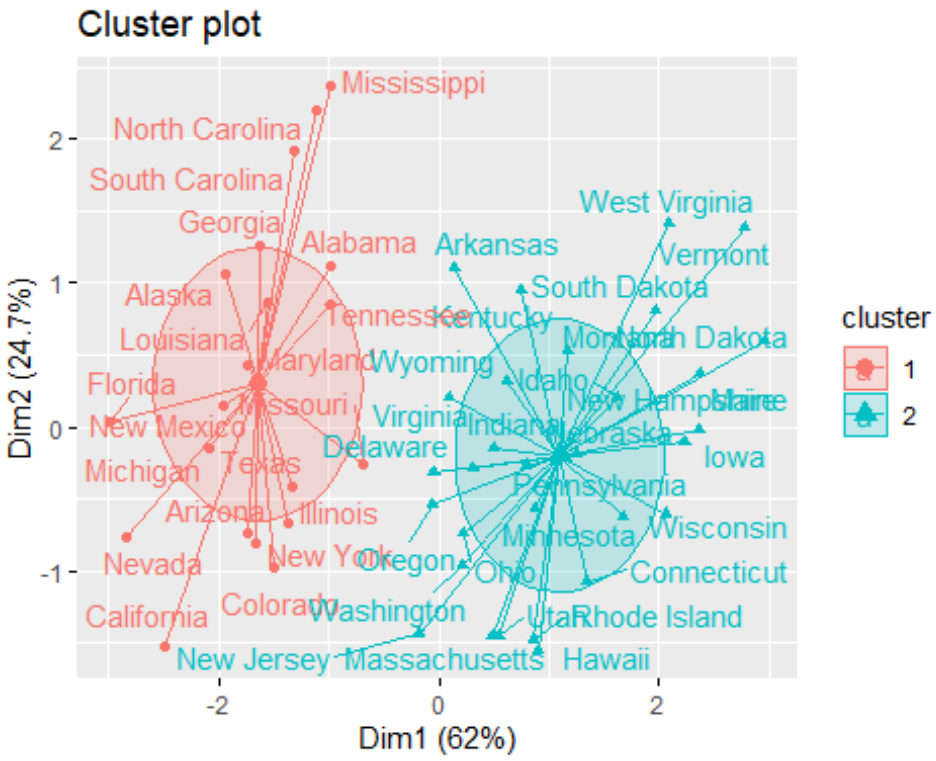
Una vez calculadas las dos agrupaciones, lo que sigue es representarlas en un grafico de K-means que es de utilidad para ver en que agrupacion estan distribuidos los estados de EUA.

```
#plotear los cluster
fviz_cluster(k2, data = df)
```




```
fviz_cluster(k2, data = df, ellipse.type = "euclid",repel = TRUE,star.plo
t = TRUE) #ellipse.type= "t", "norm", "euclid"
```

```
## Warning: ggrepel: 2 unlabeled data points (too many overlaps). Conside
r
## increasing max.overlaps
```



Podemos decir que en la agrupacion dos se encuentran los lugares mas peligrosos de acuerdo a sus numero de incidentes, etc. Por otro lado, en el cluster 1 se encuentran los estados menos peligrosos, como Arizona, Michigan, entre otros. Por ultimo se pasan las agrupaciones a la base inicial para sus posteriores análisis:

```
#pasar los cluster a mi df inicial para trabajar con ellos

USArrests %>%
  mutate(Cluster = k2$cluster) %>%
  group_by(Cluster) %>%
  summarise_all("mean")

## # A tibble: 2 x 5
##   Cluster Murder Assault UrbanPop  Rape
##     <int>  <dbl>   <dbl>    <dbl> <dbl>
## 1       1  12.2     255.     68.4  29.2
## 2       2   4.87    114.     63.6  15.9

df <- USArrests
df <- scale(df)
df<- as.data.frame(df)
df$clus<-as.factor(k2$cluster)
head(df)

##                Murder   Assault   UrbanPop         Rape clus
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473    1
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941    1
```

```
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388    1
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602    2
## California 0.27826823 1.2628144  1.7589234  2.067820292    1
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207    1
```

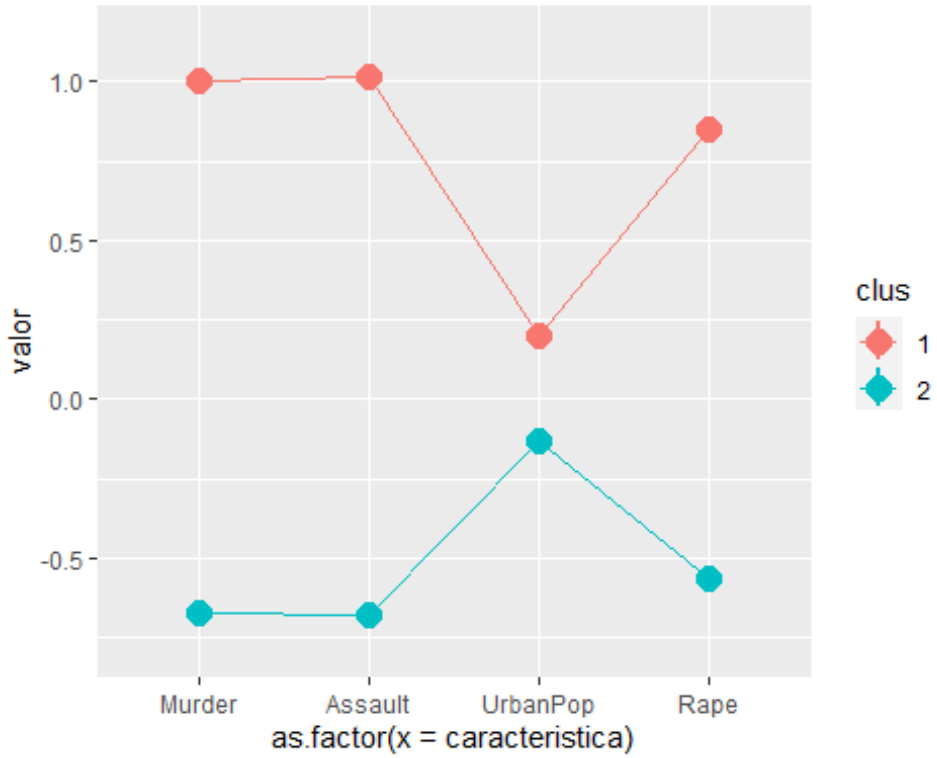
```
df$clus<-factor(df$clus)
data_long <- gather(df, caracteristica, valor, Murder:Rape, factor_key=TR
UE)
head(data_long)
```

```
##   clus caracteristica      valor
## 1    1         Murder 1.24256408
## 2    1         Murder 0.50786248
## 3    1         Murder 0.07163341
## 4    2         Murder 0.23234938
## 5    1         Murder 0.27826823
## 6    1         Murder 0.02571456
```

```
ggplot(data_long, aes(as.factor(x = caracteristica), y = valor,group=clus
, colour = clus)) +
  stat_summary(fun = mean, geom="pointrange", size = 1)+
  stat_summary(geom="line")
```

```
## No summary function supplied, defaulting to `mean_se()`
```

```
## Warning: Removed 8 rows containing missing values (geom_segment).
```



Conclusiones

Mediante el análisis clúster es posible agrupar a distintas poblaciones si se tienen variables cuantitativas, como puntuaciones, entre otros. Es de gran importancia para diversos investigadores ya que pueden agrupar sus poblaciones de estudio de acuerdo a sus propósitos y características de interés. Para este ejemplo práctico pudimos caracterizar de forma clara a cada uno de los estados de la base de datoa de acuerdoa a sus puntajes e cda una de sus variables, mediante el gráfico de K-medias podemos observar la agrupacion en donde se tienen a los estados con mayores índices delictivos y a los estadoa que más seguros son de acuerdo a nuestro análisis clúster.

Referencias

Unidad Editorial Internet. (2010, 29 diciembre). *Los crímenes violentos caen a su nivel más bajo en Estados Unidos desde 1973 | Estados Unidos | elmundo.es*. El mundo. Recuperado 28 de mayo de 2022, de https://www.elmundo.es/america/2010/12/29/estados_unidos/1293641776.html