Reporte de investigación_Dendrograma

Medel Colorado Yoselin Merari

2022-06-06

CAPÍTULO 1

Introducción

Netflix es una empresa comercial de entretenimiento que ofrece un servicio multimedia (películas y series de televisión) bajo demanda por Internet. Con este servicio, el usuario puede disfrutar de sus series y películas favoritas en cualquier dispositivo con conexión a internet, pagando una pequeña cantidad mensual o anual. Quizás no lo sepas, pero todo el contenido que Netflix muestra al usuario son recomendaciones muy precisas, de forma que coincidan casi al 100% con sus gustos. Más adelante se habla sobre cómo Netflix realiza recomendaciones. Además, estas recomendaciones pueden adaptarse a los hábitos del usuario. El Análisis de Conglomerados Jerárquico pretende identificar grupos homogéneos de variables en función de alguna característica. Una jerarquía indexada puede ser visualizada mediante un gráfico sencillo e intuitivo, llamado dendrograma. Para hacer el análisis de este trabajo se utilizó la metodología de un **dendrograma**; ya que es un grafo convexo, sin ciclos con un punto llamado **raíz** y **n** puntos extremos equidistantes de la raíz, que permite visualizar las agrupaciones en forma de árbol donde se van representando los datos por subcategorías. El objetivo de este trabajo es identificar las subcategorías de las variables de la matriz cuota de suscripción de Netflix en diferentes países, de acuerdo a sus características entre ellas.

CAPÍTULO 2

1.- Instalar paquetes para hacer el análisis

```
install.packages("readr")
library(readr)
install.packages("cluster.datasets")
library(cluster.datasets)
install.packages("dendextend")
library(dendextend)
install.packages("factoextra")
library(factoextra)
install.packages("ggplot2")
library(ggplot2)
install.packages("igraph")
library(igraph)
install.packages("stats")
library(stats)
install.packages("factoextra")
library(factoextra)
install.packages("scales")
library(scales)
```

```
install.packages("ggsci")
library(ggsci)
install.packages("cluster")
library(cluster)
install.packages("factoextra")
library(factoextra)
```

Tratamiento de la matriz

1.- Nombre de la matriz de datos: Cuota de suscripción de Netflix en diferentes países.

Se usa esta base de datos "Netflix_subscription_fee_Dec_2021 <- read_csv("Netflix subscription fee Dec-2021.csv")", contiene datos sobre los países que pagan el servicio de Netflix. La matriz de datos se descargó de la página web Kaggle, cuenta con 65 observaciones y 8 variables que son: Country_code, Country, Total Library Size, No. of TV Shows, No. of Movies, Cost Per Month - Basic $, Cost Per Month - Standard $ y Cost Per Month - Premium $.

1.1.- Matriz de datos

```
Netflix_subscription_fee_Dec_2021 <- read_csv("Netflix subscription fee Dec-2021.csv")
```

```
## Rows: 65 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Country_code, Country
## dbl (6): Total Library Size, No. of TV Shows, No. of Movies, Cost Per Month ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1.2.- Se renombra la matriz

```
NsfC2=Netflix_subscription_fee_Dec_2021
```

1.3.- Componentes de la matriz

```
head(NsfC2)
```

```
## # A tibble: 6 x 8
##   Country_code Country   `Total Library Size` `No. of TV Shows` `No. of Movies`
##   <chr>        <chr>                    <dbl>             <dbl>           <dbl>
## 1 ar           Argentina                 4760              3154            1606
## 2 au           Australia                 6114              4050            2064
## 3 at           Austria                   5640              3779            1861
## 4 be           Belgium                   4990              3374            1616
## 5 bo           Bolivia                   4991              3155            1836
## 6 br           Brazil                    4972              3162            1810
## # ... with 3 more variables: `Cost Per Month - Basic ($)` <dbl>,
```

```
## #   `Cost Per Month - Standard ($)` <dbl>, `Cost Per Month - Premium ($)` <dbl>
```

Así es como se visualizan los datos de la matriz. Donde:

- Country_code: código de país La abreviatura de los países, se encuentran ordenados alfabéticamente.
- Country: país Son 65 países, empezando con Argentina y finalmente Venezuela.
- Total Library Size: tamaño total de la libreria Con un mínimo de 2274 y un máximo de 7325.
- No. of TV Shows: no. de programas de televisión Con un mínimo de 1675 y un máximo de 5234.
- No. of Movies: no. de películas Con un mínimo de 373 y un máximo de 2387
- Cost Per Month - Basic $: costo por mes - básico $ Con un mínimo de $1.97 y un máximo de $12.88.
- Cost Per Month - Standard $: costo por mes - estándar $ Con un mínimo de $3 y un máximo de $20.46.
- Cost Per Month - Premium $: costo por mes - premium $ Con un mínimo de $4.02 y un máximo de $26.96.

1.4.- Exploración de la matriz

```
dim(NsfC2) # Dimensión
```

```
## [1] 65  8
```

```
str(NsfC2) # Tipo de variables
```

```
## spec_tbl_df [65 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Country_code                 : chr [1:65] "ar" "au" "at" "be" ...
##  $ Country                      : chr [1:65] "Argentina" "Australia" "Austria" "Belgium" ...
##  $ Total Library Size           : num [1:65] 4760 6114 5640 4990 4991 ...
##  $ No. of TV Shows              : num [1:65] 3154 4050 3779 3374 3155 ...
##  $ No. of Movies                : num [1:65] 1606 2064 1861 1616 1836 ...
##  $ Cost Per Month - Basic ($)   : num [1:65] 3.74 7.84 9.03 10.16 7.99 ...
##  $ Cost Per Month - Standard ($): num [1:65] 6.3 12.1 14.7 15.2 11 ...
##  $ Cost Per Month - Premium ($) : num [1:65] 9.26 16.39 20.32 20.32 13.99 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Country_code = col_character(),
##   ..   Country = col_character(),
##   ..   `Total Library Size` = col_double(),
##   ..   `No. of TV Shows` = col_double(),
##   ..   `No. of Movies` = col_double(),
##   ..   `Cost Per Month - Basic ($)` = col_double(),
##   ..   `Cost Per Month - Standard ($)` = col_double(),
##   ..   `Cost Per Month - Premium ($)` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
colnames(NsfC2) # Nombre de las variables
```

```
## [1] "Country_code"                 "Country"
## [3] "Total Library Size"           "No. of TV Shows"
## [5] "No. of Movies"                "Cost Per Month - Basic ($)"
## [7] "Cost Per Month - Standard ($)" "Cost Per Month - Premium ($)"
```

```
anyNA(NsfC2) # Presencia de NA
```

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

La dimensión de la matriz es de 65 observaciones y 8 variables, el tipo de variables que contiene son: 6 variables numéricas que tienen registrado el tamaño total de la biblioteca, no. de programas de TV, no. de películas, costo por mes - básico $, costo por mes - estándar $ y costo por mes - premium $ y 2 de carácter que tienen registrados los nombres de los países. No hay presencia de datos perdidos (NA).

CAPÍTULO 3

Metodología

hhclust

Para crear un dendrograma se necesita calcular las matrices de distancias de los datos con la función **dist** y luego el clúster jerárquico de la matriz de distancias con **hclust** para finalmente crear el dendrograma.

Sin embargo, en Hierarchical Clustering (hclust), los clusters se crean de manera que tengan un orden predeterminado, es decir, una jerarquía. Por ejemplo, para realizar este trabajo se utilizó una matriz de datos llamada: Cuota de suscripción de Netflix Dic-2021, donde se considera la jerarquía de que países pagan más y menos por este servicio.

CLARA

(Clustering Large Applications) es un método que combina la idea de K-medoids con el resampling para que pueda aplicarse a grandes volúmenes de datos. CLARA selecciona una muestra aleatoria de un tamaño determinado y le aplica el algoritmo de **PAM (K-medoids)** para encontrar los clusters óptimos acorde a esa muestra. Utilizando esos medoids se agrupan las observaciones de todo el set de datos. La calidad de los medoids resultantes se cuantifica con la suma total de las distancias entre cada observación del set de datos y su correspondiente medoid (suma total de distancias intra-clusters). CLARA repite este proceso un número predeterminado de veces con el objetivo de reducir el bias de muestreo. Por último, se seleccionan como clusters finales los obtenidos con aquellos medoids que han conseguido menor suma total de distancias. Se describen los pasos del algoritmo CLARA:

- Se divide aleatoriamente el set de datos en **n** partes de igual tamaño, donde **n** es un valor que determina el analista.

Para cada una de las **n** partes:

2.1 Aplicar el algoritmo **PAM** e identificar cuáles son los **k medoids**.

2.2 Utilizando los **medoids** del paso anterior agrupar todas las observaciones del set de datos.

2.3 Calcular la suma total de las distancias entre cada observación del set de datos y su correspondiente **medoid** (suma total de distancias intra-clusters).

Seleccionar como clustering final aquel que ha conseguido menor suma total de distancias intra-clusters en el paso.

CAPÍTULO 4

Resultados

1.- Cálculo de la matriz de distancia de Mahalonobis

```
dist.NsfC2<-dist(NsfC2[,3:6])
```

Se calcula la distancia de Mahalanobis de la variable 3 a la 6, ya que son variables númericas. Con este cálculo se obtiene la similitud y correlación que hay entre ellas.

1.1.- Convertir los resultados del cálculo de la distancia a una matriz de datos y me indique 3 dígitos

```
round(as.matrix(dist.NsfC2)[1:6, 1:6],3)
```

```
##          1        2        3        4        5        6
## 1    0.000 1686.983 1109.089  318.498  326.006  294.321
## 2 1686.983    0.000  582.518 1386.024 1454.008 1468.753
## 3 1109.089  582.518    0.000  804.084  900.668  910.787
## 4  318.498 1386.024  804.084    0.000  310.430  287.984
## 5  326.006 1454.008  900.668  310.430    0.000   33.127
## 6  294.321 1468.753  910.787  287.984   33.127    0.000
```

Se construye una matriz de la distancia de Mahalanobis y en el cálculo se hace un redondeo donde se utilizan los primeros 6 individuos que especifican la selección.

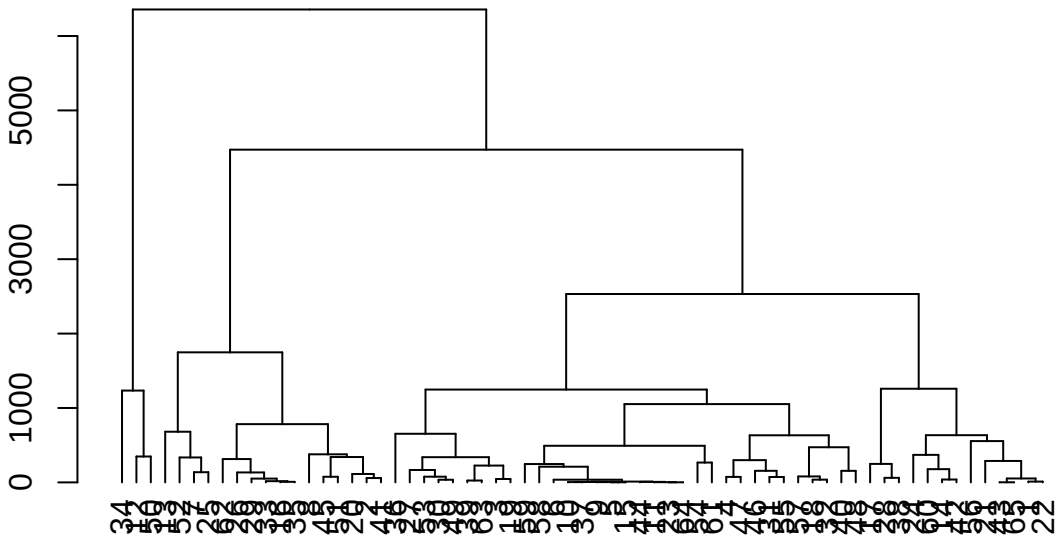
1.2.- Cálculo del dendrograma

```
dend.NsfC2<-as.dendrogram(hclust(dist.NsfC2))
```

Se calcula el dendrograma, utilizando el método de agrupación (hclust), ya que hace una agrupación jerárquica.

1.3.- Generación del dendrograma

```
plot(dend.NsfC2)
```



Se observan los agrupamientos, pero sin etiquetas asignadas.

1.4.- Agregar etiquetas al gráfico

```
NsfC2.country=NsfC2
NsfC2.country=NsfC2.country[,-1]
```

Se agregan etiquetas al gráfico para una mejor visualización.

2.- Modificar el dendrograma

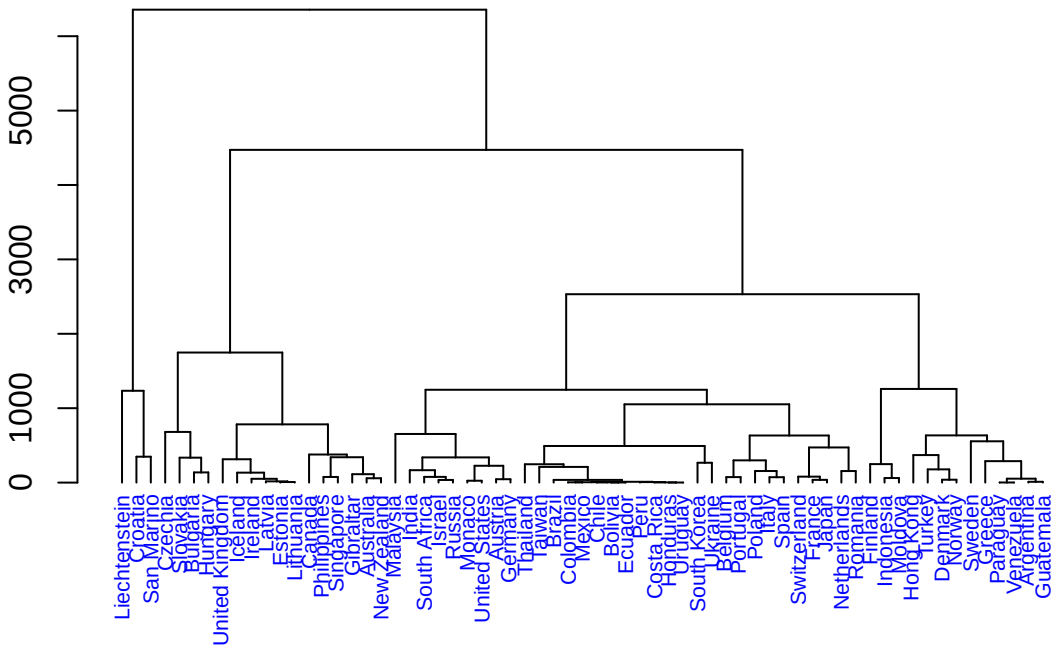
2.1.- Guardar las etiquetas en un objeto "L"

```
L=labels(dend.NsfC2)
labels(dend.NsfC2)=NsfC2$Country[L]
```

2.2.- Cambiar el tamaño de las etiquetas

```
dend.NsfC2 %>%
  set(what="labels_col", "blue") %>% #Colores etiqueta
  set(what="labels_cex", 0.7) %>%
  plot(main="Figura 1- Cuota de suscripción de Netflix en diferentes países ")
```

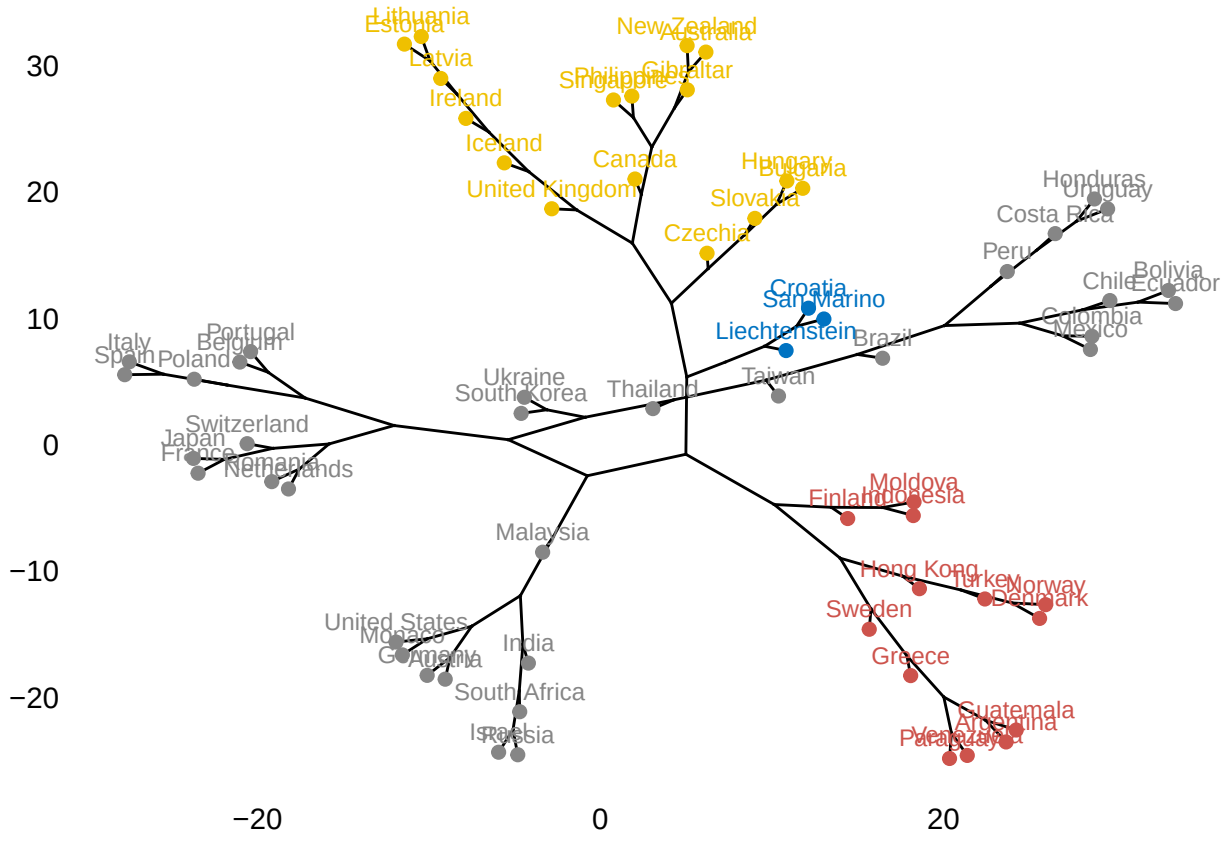
Figura 1− Cuota de suscripción de Netflix en diferentes países



Dendrograma en forma de árbol filogenético- Figura 2

```
Phylo = fviz_dend(dend.NsfC2, cex = 0.8, lwd = 0.8, k = 4,
rect = TRUE,
k_colors = "jco",
```

```
rect_border = "jco",
rect_fill = TRUE,
type = "phylogenic")
Phylo%>%
plot(main="Figura 2- Cuota de suscripción de Netflix en diferentes países ")
```



Clustering CLARA

Matriz con selección de variables

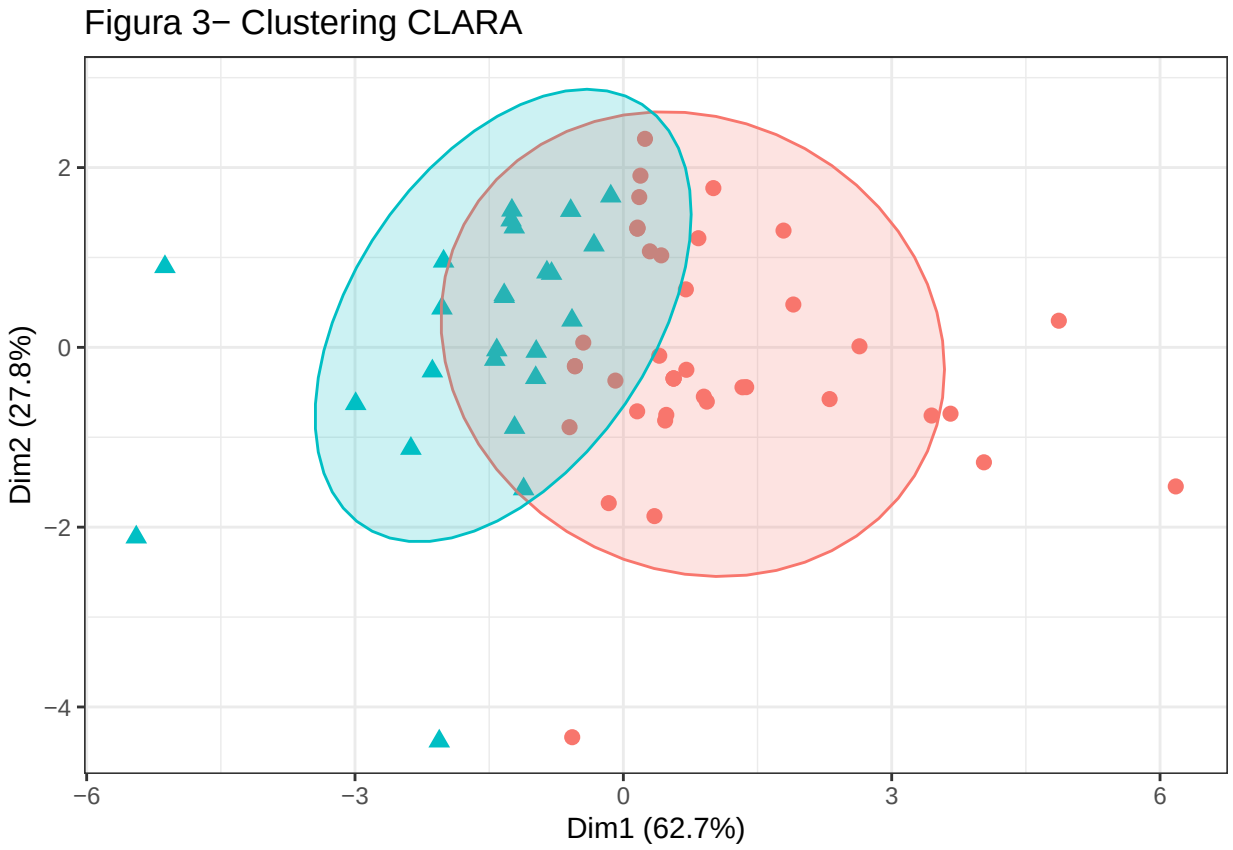
```
datosCLARA <- cbind(NsfC2$`No. of TV Shows`, NsfC2$`No. of Movies`, NsfC2$`Cost Per Month - Basic ($)`,
colnames(datosCLARA) <- c("A", "B", "C", "D","E","F")
head(datosCLARA)
```

```
##         A    B     C     D     E     F
## [1,] 3154 1606  3.74  3.74  6.30  9.26
## [2,] 4050 2064  7.84  7.84 12.12 16.39
## [3,] 3779 1861  9.03  9.03 14.67 20.32
## [4,] 3374 1616 10.16 10.16 15.24 20.32
## [5,] 3155 1836  7.99  7.99 10.99 13.99
## [6,] 3162 1810  4.61  4.61  7.11  9.96
```

```
clara_clusters <- clara(x = datosCLARA, k = 2, metric = "manhattan", stand = TRUE,
samples = 17, pamLike = TRUE)
clara_clusters
```

```
## Call:     clara(x = datosCLARA, k = 2, metric = "manhattan", stand = TRUE,      samples = 17, pamLik
## Medoids:
##         A    B    C    D     E     F
## [1,] 3155 1836 7.99 7.99 10.99 13.99
## [2,] 3779 1861 9.03 9.03 14.67 20.32
## Objective function:   4.574702
## Clustering vector:    int [1:65] 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 2 2 ...
## Cluster sizes:            40 25
## Best sample:
##  [1]  2  3  4  5  8  9 10 11 12 14 16 17 18 19 20 22 23 25 26 27 28 29 30 31 32
## [26] 34 35 37 38 39 41 43 44 45 46 47 49 50 51 54 59 60 62 65
##
## Available components:
##  [1] "sample"     "medoids"    "i.med"      "clustering" "objective"
##  [6] "clusinfo"   "diss"       "call"       "silinfo"    "data"
```

```
fviz_cluster(object = clara_clusters, ellipse.type = "t", geom = "point",
pointsize = 2.5) +
theme_bw() +
labs(title = "Figura 3- Clustering CLARA") +
theme(legend.position = "none")
```



CAPÍTULO 5

Conclusión

Finalmente, observando el dendrograma de la figura 1 cuota de suscripción de Netflix en diferentes países, hay claramente tres segmentos distintos.

En el primer subgrupo que es del lado izquierdo está conformado por 3 países (Liechtenstein, Croatia y San Marino).

El segundo subgrupo que es el intermedio está conformado por 16 países (Czechia, Slovakia, Bulgaria, Hungary, United Kingdom, Iceland, Ireland, Latvia, Estonia, Lithuania, Canada, Philippines, Singapore, Gibraltar, Australia y New Zealand).

En el tercer subgrupo que es del lado derecho está conformado por 45 países (Malaysia, India, South Africa, Israel, Russia, Monaco, United States, Austria, Germania, Thailand, Taiwan, Brazil, Colombia, Mexico, Chile, Bolivia, Ecuador, Peru, Costa Rica, Honduras, Uruguay, South Korea, Ukraine, Belgium, Portugal, Poland, Italy, Spain, Switzerland, France, Japan, Netherlands, Romania, Finland, Indonesia, Moldova, Hong Kong, Turkey, Denmark, Norway, Sweden, Greece, Paraguay, Venezuela, Argentina y Guatemala).

En la figura 2 que es un dendrograma en forma de árbol filogenético; se observa la división de clústeres de una forma organizada, es decir, se dividen por ramas que representan las características que comparten entre sí, entre más cerca estén comparten más similitudes. Es decir, cuando un grupo comparte características similares se genera una **rama**.

En la figura 3 de clustering CLARA hace una muestra aleatoria de las variables de la matriz de datos cuota de suscripción de Netflix en diferentes países donde genera agrupaciones de los datos que comparten características similares. Sin embargo, en el grupo de color azul se observan 3 triángulos fuera de la elipse lo que significa que no están bien clasificados, mientras que en el grupo color salmón se observan 5 puntos fuera de la elipse lo que significa que tampoco están bien clasificados.

REFERENCIAS

- Base de datos recuperada de: https://www.kaggle.com/datasets/prasertk/netflix-subscription-price-in-different-countries
- Agradecimientos Fuente de datos: https://www.comparitech.com/blog/vpn-privacy/countries-netflix-cost/ Crédito de la imagen de portada: https://www.pexels.com/photo/light-man-people-woman-5112410/
- Amat, J. (2017, septiembre). Clustering y heatmaps: aprendizaje no supervisado. RPubs. Recuperado de: https://rpubs.com/Joaquin_AR/310338
- Isaac, J. (2021, 22 abril). Cluster jerarquico en R. RPubs. Recuperado de: https://rpubs.com/jaimeisa acp/760355
- To cite R in publications use:

  R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- To cite package 'cluster.datasets' in publications use:

  Novomestky F (2013). *cluster.datasets: Cluster Analysis Data Sets*. R package version 1.0-1, https://CRAN.R-project.org/package=cluster.datasets.