Proyecto de Investigación

Melissa Ortega Galarza

2022-06-07

Capítulo I

Introducción.

Para este Proyecto de Investigación se ocupo el tema de un Dendrograma, para ello se ocupo una base de datos llamada "Top 50 Spotify - 2019" descargada de la pagina web Kaggle, cuenta con 50 observaciones y 6 variables, la cual trata de los mejores 50 Artistas en cuanto a sus canciones en Spotify dentro del año 2019 las variables son las siguientes:

- Artist Name : Nombre del Artista
- Popularity : Cuanto mayor sea el valor, más popular es el Artista.
- Acousticness : Cuanto mayor sea el valor, más fuerte es la canción
- Beats.Per.Minute : El tiempo de la canción del Artista
- Speechiness : Palabras dichas
- Energy : Cuanto mayor sea el valor mas energetico sera el Artista

El objetivo de esto es poder realizar un dendrograma a partir de nuestra base de datos y utilizando un codigo para ello, de tal manera que con esto podamos visualizar nuestros datos de manera clara y poder llegar a realizar conclusiones finales. Por otra parte se decidio ocupar la variable de ***Artist Name*** como nuestra variable principal para que después de esto podamos obtener nuestro análisis final.

Figura 1. Artistas que estan dentro del Top 50

Capítulo II

Tratamiento de la Matríz.

Cargamos la Libreria para leer la Base de Datos

```
library(readxl)
```

Cargamos la Base de Datos

```
spotify <- read_excel("spotify.xlsx")
```

Cambiamos el Nombre de la Matriz

```
M=spotify
```

Exploración de la Matriz

Dimensión

```
dim(M)
```

```
## [1] 50  6
```

Análisis:

Observamos que tenemos una matríz de 50 x 6

Nombre de las Variables

```
names(M)
```

```
## [1] "Artist.Name"      "Popularity"       "Acousticness.."   "Beats.Per.Minute"
## [5] "Speechiness."     "Energy"
```

Análisis:

Observamos el nombre de nuestras 6 variables

Tipo de Variable

```
str(M)
```

```
## tibble [50 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Artist.Name     : chr [1:50] "Shawn Mendes" "Anuel AA" "Ariana Grande" "Ed Sheeran" ...
##  $ Popularity      : num [1:50] 79 92 85 86 94 84 92 90 87 95 ...
##  $ Acousticness..  : num [1:50] 4 8 12 12 45 9 2 15 5 33 ...
##  $ Beats.Per.Minute: num [1:50] 117 105 190 93 150 102 180 111 136 135 ...
##  $ Speechiness.    : num [1:50] 3 9 46 19 7 4 29 9 10 38 ...
##  $ Energy          : num [1:50] 55 81 80 65 65 68 64 68 62 43 ...
```

Análisis:

Observamos que tenemos 1 variable de carácter y 5 de tipo númerico, por lo que se va a trabajar con la variable de carácter

Vemos si hay o no hay datos perdidos

```
anyNA(M)
```

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

Análisis:

Observamos que no tenemos ningún dato perdido.

Capítulo III

Metodología.

La metodología que se ocupo fue la de el hclust permite agregar rectángulos para indicar los grupos del dendrograma. Puedes seleccionar el número de grupos a ser mostrados con el argumento k . Ten en cuenta que puedes resaltar solo algunos rectángulos en base al número de grupos. Los llamados métodos jerarquicos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida desimilitud. Los métodos jerarquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorias presenta una gran diversidad de variantes.

1. *Calculamos la distancia de Mahalanobis de nuestra matríz.*

```
dist.M<-dist(M[,2:6])
```

2. *Convertimos los resultados del Cálculo de la distancia a una matríz de datos y que nos indique 3 digitos.*

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

```
##        1      2       3       4      5      6
## 1  0.000 32.265  88.899  32.326 55.776 21.095
## 2 32.265  0.000  93.059  23.495 60.481 16.371
## 3 88.899 93.059   0.000 101.804 67.201 98.295
## 4 32.326 23.495 101.804   0.000 67.424 18.111
## 5 55.776 60.481  67.201  67.424  0.000 60.975
## 6 21.095 16.371  98.295  18.111 60.975  0.000
```

3. *Cálculo del Dendrograma*

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

Instalamos las Paqueteria para el Dendrograma.

```
install.packages("dendextend")
install.packages("factoextra")
install.packages("ggplot2")
```

```
library(dendextend)
library(factoextra)
library(ggplot2)
```

Guardamos nuestras etiquetas en un objeto "L".

```
L=labels(dend.M)
labels(dend.M)=M$Artist.Name[L]
```

Capítulo IV

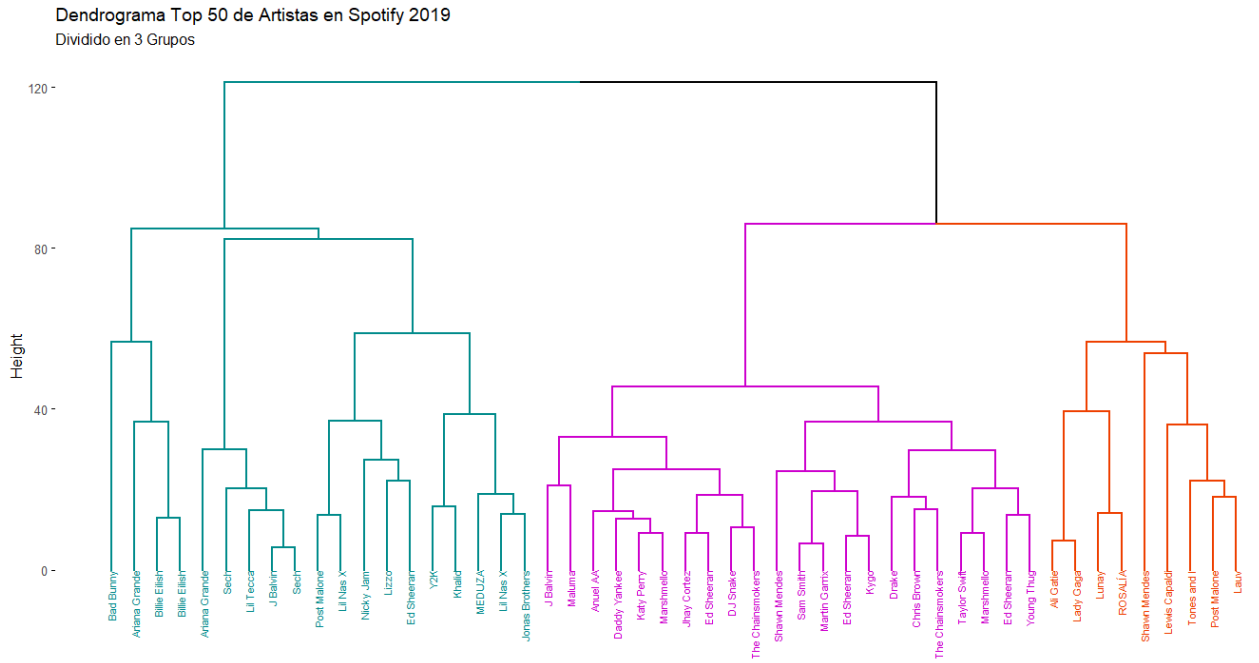
Resultados.

Dendrograma

El dendrograma es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus niveles de similitud. El nivel de similitud se mide en el eje vertical (alternativamente se puede mostrar el nivel de distancia) y las diferentes observaciones se especifican en el eje horizontal.

Creamos nuestro Dendrograma.

```
fviz_dend(dend.M,
          k = 3,cex=0.6,border=2:10,k_colors = c("darkcyan","magenta3","orangered2")) +
  labs(title = "Dendrograma Top 50 de Artistas en Spotify 2019",
       subtitle = "Dividido en 3 Grupos")
```



Capítulo V

Conclusiones.

Análisis del Gráfico del Dendrograma

Dentro del dendrograma observamos que esta dividido en 3 grupos esto quiere decir que los Artistas que estan dentro de cada grupo comparten mismas caracteristicas que es lo que hace que esten en el mismo grupo.

En el primer grupo que esta de color darkcyan se compone de 19 Artistas entre ellos Bad Bunny , Ariana Grande , J Balvin , Jonas Brothers, Billie Ellish, entre otros Artistas más , lo que hace que estos Artistas esten en este grupo es que comparten caracteristicas iguales o parecidas en cuanto a datos númericos y en esta caso es la popularidad , los beats por minuto , y la energía.

En el segundo grupo que esta de color magenta se compone de 22 Artistas entre ellos Maluma , Anuel , Katty Perry , Shawn Mendes , Daddy Yankee , entre otros Artistas más, lo que hace que estos Artistas esten en este grupo es que comparten caracteristicas iguales o parecidas en cuanto a datos numericos y en esta caso es la popularidad, los acusticos y la energía.

Finalmente en el tercer grupo nos encontramos con 9 Artistas entre ellos Lady Gaga , Lunay , Rosalia , Post Malone , Lauv , entre otros Artistas más, lo que hace que estos Artistas esten en este grupo es que comparten caracteristicas iguales o parecidas en cuanto a datos numericos y en esta caso es la popularidad, y la energía

Para concluir podemos observar que alguno de los Artistas se repite en algun grupo pero es debido que puede llegar a compartir datos numericos iguales o parecidos con algunso otros Artistas.

Esto de ver que comparten caracteristicas o datos númericos iguales o parecidos se obtuvo observando la base de datos y el gráfico de esta manera se tuvo que observar detalladamente porque se sepraban por grupos y que era lo que compartian esos Artistas. Por lo que podemos decir que un Dendrograma nos funciona para poder clasificar por grupos nuestras variables y es de gran utilidad un dendrograma.

Referencias

LEONARDO HENRIQUE (2019). Kaggle. Obtenido de: https://www.kaggle.com/datasets/leonardopena/top-spotify-songs-from-20102019-by-year

José Carlos Soage (2022) R CHARTS por R CODER https://r-charts.com/es/parte-todo/hclust/