# Proyecto final

Lino Oswaldo Sánchez Juárez

2/6/2022

# Capitulo 1

## Intoducción

Al usar un agrupamiento por K-means, podemos elegir el número de grupos que queremos trabajar con este algoritmo de agrupación. Siendo así un método que permite aplicar el criterio del investigador y la intuición de este pues el agrupamiento se realiza por la distancia que existen entre las observaciones.

En lo particular para mi el agrupamiento de K-means me pareció muy efectivo para poder clasificar las canciones de spotify del 2021 ya que al compartir las mismas variables y los géneros no son muy distintos entre ellos este clusterisación es efectiva para la naturaleza de estos datos; cabe señalar que este método es más utilizado en imágenes pues tienen muchas características diferentes es muy bueno probar la efectividad en un ramo diferente.

## Objetivo

Como objetivo principal es lograr con éxito el agrupamiento de las canciones con la mayor varaiabilidad explicada por los clusters.

# Capitulo 2

## Datos que se utilizan

Para este trabajo final se decidió usar una base de datos extraída de un portal de datos libres, la cual contiene información del top 50 canciones más escuchadas en el 2021; al extraer la base original, esta estaba conformada por 18 variables que recogían diferente clase de datos, es por eso que se hizo un filtrado de variables par solo trabajar con las de mayor interés para el investigador.

Cabe resaltar que la base fue revisada por datos faltantes errores de ortografía y demás desde la hoja de cálculo de Excel, para tener mayor comodidad de no hacerlo en el programa estadístico R studio y que el análisis fueran los más rápido posible y claros sin tener tantos pasos no necesarios.

Siendo así solo nos quedamos con las siguientes:

- **track name**: contiene los nombres de las canciones y es de naturaleza carácter.
- **popularity**: contiene valores numéricos sobre la popularidad, entre mas alto sea ese valor, más popular es la canción y de naturaleza numérica.

- **danceability**: contiene valores numéricos sobre que tan bailable es cuanto más alto sea el valor, más fácil será bailar esta canción de naturaleza numérica.
- **energy**: contiene valores numéricos sobre la energía de la canción, cuanto más alto sea el valor más energética sera la canción, de naturaleza numérica.
- **acousticness**: contiene valores numéricos de la acústica, entre más alto sea el valor, más acústica será la canción, de naturaleza numérica.
- **valence**: contiene valores numéricos sobre el estado de animo entre más alto sea el valor, más positivo será el estado de ánimo de la canción
- **tempo**: contiene los valores del ritmo de las canciones, de naturaleza numérica.

# Capitulo 3

## Metodología

El algoritmo K-means parte de un amuestra de n elementos con p variable. donde el objetivo es dividir la muestra en un número de grupos prefijado,(k).El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática. Este algoritmo esta compuesto por cuatro etapas:

1.- Selecciona K putos como centro de los grupos iniciales

1.1.- Asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos formados.

1.2.- Tomando como centro los k puntos más lejanos.

1.3.- Construyendo los grupos con información a priori o seleccionando los centros (a priori).

2.- Calcular las distancias euclídeas de cada elemento al centro de los k grupos y asignar cada elemento al grupo próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

3.- Definir un criterio de optimalidad y comprobar si reasignando uno a uno cada elemento de un grupo a otro mejora el criterio.

4.- Sí no es posible mejorar el criterio de optimalidad, terminar el proceso.

El algoritmo de K-means es un acercamiento al el maching lerning dándonos una introducción de como la clasificación automatizada basada en las distancia euclídeas es muy eficiente.

## comandos

### Metodo k-means

1.- Separacion de filas y columnas.

dim(basefreme) filas n<-dim(basefreme)[1] columnas p<-dim(basefreme)[2]

2.- Estandarizacion univariante. X.s<-scale(basefreme)

3.- Algoritmo k-medias (3 grupos) nstart: cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo. X.S: estadarizacion de los datos Kmeans.3<-kmeans(X.s, 2, nstart=25)

centroides Kmeans.3$centers

cluster de pertenencia Kmeans.3$cluster

4.- suma de cuadrados dentro de los grupos SCDG

SCDG<-sum(Kmeans.3$withinss) SCDG

5.- separa los Clusters cl.kmeans<-Kmeans.3$cluster cl.kmeans

6.- Scatter plot con la division de grupos obtenidos (se utiliza la matriz de datos centrados). col.cluster<-c("orange", "green")[cl.kmeans] pairs(X.s, col=col.cluster, main="k-meadias", pch=19)

Visualizacion con las dos componentes principales

library(cluster)

clusplot(X.s, cl.kmeans, main="Dos primeras componentes principales")

text(princomp(X.s)$score[,1:2], labels=rownames(X.s), pos=1, col="orange")

Silhouette

Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

1.- Generacion de los calculos dist.Euc<-dist(X.s, method = "euclidean") Sil.kmeans<-silhouette(cl.kmeans, dist.Euc)

2.- Generacion del grafico plot(Sil.kmeans, main="Silhouette for k-means", col="orange")

# Capitulo 4

## Resultados

### Base de datos

Librería necesaria

```
library(readxl)
```

```
base_1_1 <- read_excel("C:/Users/Usuario/Documents/Esatadistica multivariada/proyecto final dendograma/
attach(base_1_1)
```

### Convertir a data freme

```
basefreme<-data.frame(base_1_1)
basefreme[c("track_name")]<-NULL
```

### Etiquetas

```
rownames(basefreme)=paste(base_1_1$track_name)
colnames(basefreme)
```

```
## [1] "popularity"   "danceability" "energy"       "acousticness" "valence"
## [6] "tempo"
```

# Implemenctación del algoritmo K-means

1.- Separacion de filas y columnas.

filas

```
n<-dim(basefreme)[1]
```

columnas

```
p<-dim(basefreme)[2]
```

2.- Estandarizacion univariante.

```
X.s<-scale(basefreme)
```

Ya que las variables no tienen la misma unidad e medida estandarizamos escalando y creando un objeto que contenga estas escalaciones.

3.- Algoritmo k-medias (3 grupos) nstart: cantidad de subconjuntos aleatorios que se escogen para realizar los cálculos de algoritmo. X.S: estadarizacion de los datos

```
Kmeans.3<-kmeans(X.s, 2, nstart=25)
```

**centroides**

```
Kmeans.3$centers
```

```
##    popularity danceability     energy acousticness   valence      tempo
## 1 -0.06383451    0.2840984  0.4517998   -0.4872891  0.438541  0.1658140
## 2  0.14894718   -0.6628963 -1.0541996    1.1370080 -1.023262 -0.3868994
```

Extraemos los centroides de el objeto Kmeans.3 estos son los centroides del que se usaran para los clusters

**cluster de pertenencia**

```
Kmeans.3$cluster
```

```
##        drivers license                MONTERO                   STAY
##                      2                      1                      1
##               good 4 u             Levitating                Peaches
##                      1                      1                      1
##           Kiss Me More        Blinding Lights             Heat Waves
##                      1                      1                      2
##                Beggin'  Astronaut Inthe Ocean                 DÁKITI
##                      1                      1                      2
##          INDUSTRY BABY             Bad Habits        Save Your Tears
##                      1                      1                      1
```

```
##                  Butter      Leave The Door Op                deja vu
##                       1                      1                      2
##             Todo De Ti                   Mood           The Business
##                       1                      1                      2
##               Dynamite               Yonaguni       Watermelon Sugar
##                       1                      1                      1
##        Friday Dopamine              telepatía            WITHOUT YOU
##                       1                      1                      2
## Heartbreak Anniversary                traitor                  Pepas
##                       2                      2                      1
##              positions      Someone You Loved                Bandido
##                       1                      2                      1
##   IWANNA BE YOUR SLAVE                 RAPSTAR      LA NOCHE DE ANOCH
##                       1                      2                      1
##                Streets        Sweater Weather                   Fiel
##                       2                      1                      1
##           Need to Know        Don't Start Now               Lemonade
##                       2                      1                      1
##                  Woman                 Arcade              Good Days
##                       1                      2                      2
##          Qué Más Pues?           Head & Heart                  34+35
##                       1                      1                      1
##     you broke me first         Pareja Del Año
##                       2                      1
```

Visualizamos en que clusters están las observaciones, e identificamos para que en pasoso siguietes sepamos cules son los clusters de cada uno de ellos.

4.- suma de cuadrados dentro de los grupos SCDG

```
SCDG<-sum(Kmeans.3$withinss)
SCDG
```

```
## [1] 206.9465
```

El criterio de homogeneidad que se utiliza en el algoritmo de k-medias es la suma de cuadrados dentro de los grupos (SCDG) para todas las variables.

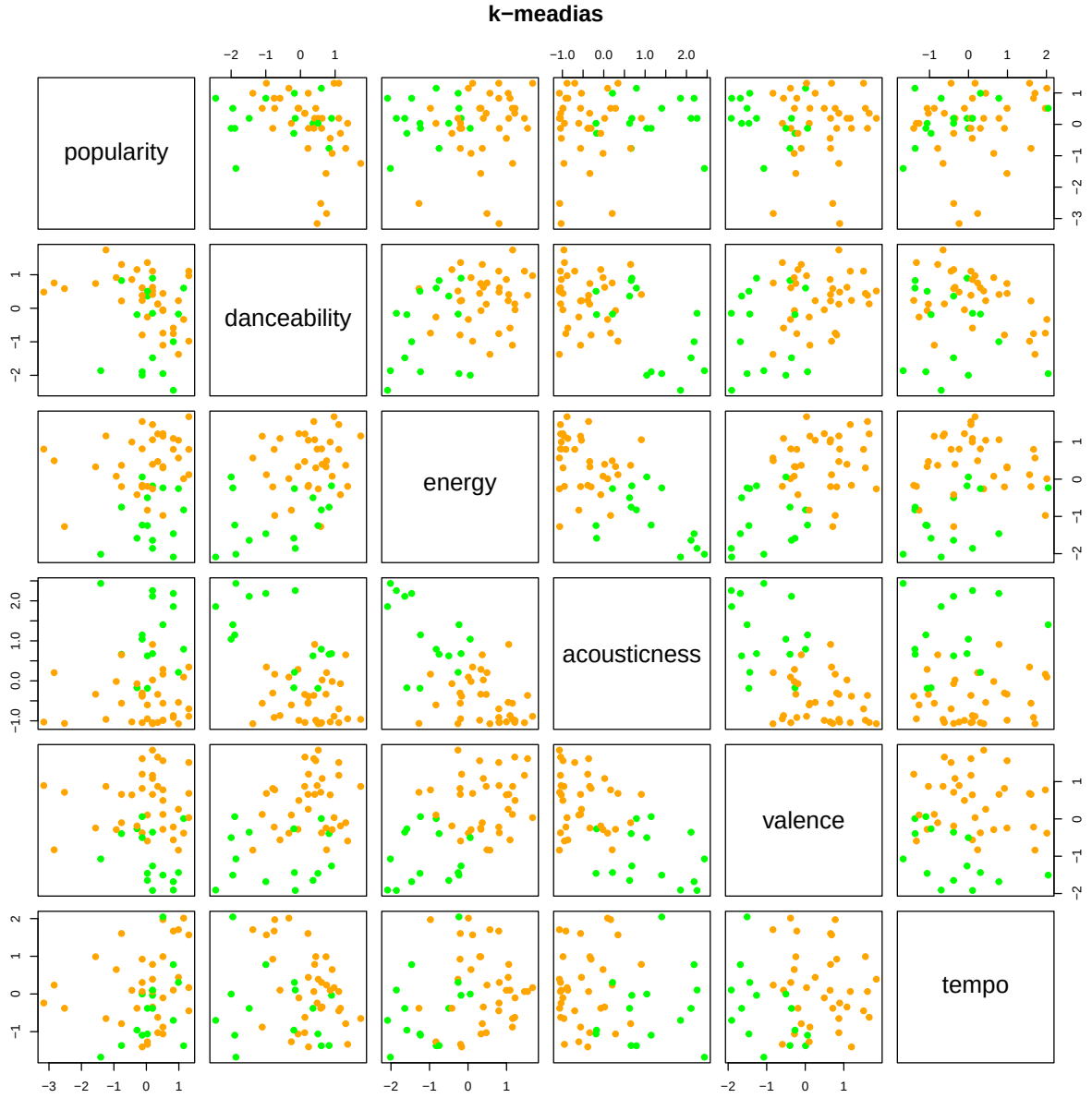
Esto equivale a la suma ponderada de las varianzas de las variables en los grupos; lo que se busca en concreto es que la suma de cuadrados se la menor posible puesto que entre más pequeña se la varianza los grupos serán más homogéneas.

5.- separa los Clusters Creamos un objeto y en donde estarán contenidos los clusters

```
cl.kmeans<-Kmeans.3$cluster
```

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

```
col.cluster<-c("orange", "green")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-meadias", pch=19)
```




Lo que podemos visualizar aquí es como están agrupados las observaciones y su correlación entre ellas, como se puede ver claramente las variables que se encuentran en la parte del centro son las que tiene una mejor correlación a comparación de las observadas en el exterior del gráfico, lo cierto es que no tenemos muchas correlaciones entre variables pero es un problema que no estén en una misma escala de mediada aunque ya escalamos la matriz para homogeneizar ese problema sale a relucir aquí aun así nuestro algoritmo es capaz de encontrar similitudes y agruparlas.

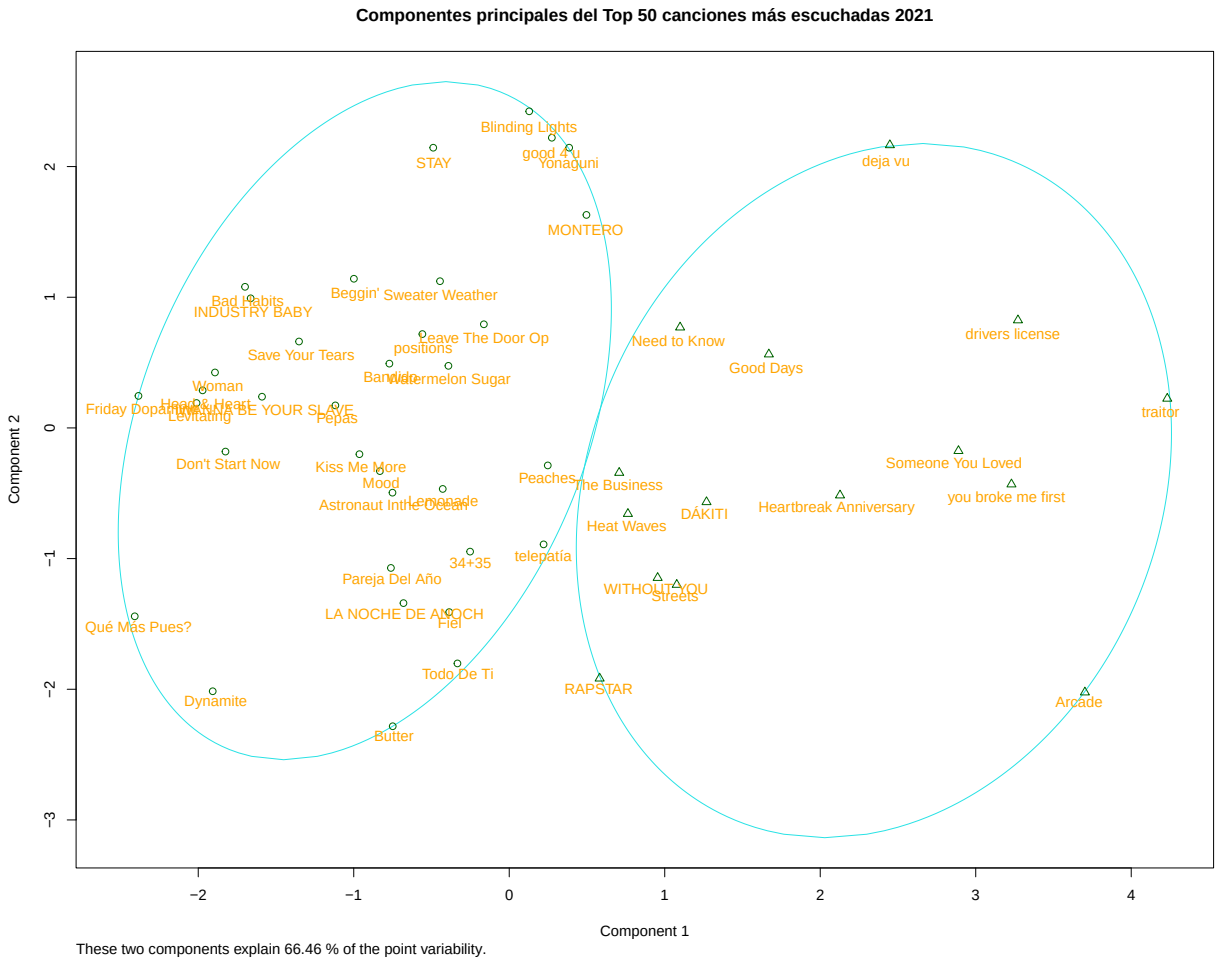
## Visualizacion con las dos componentes principales

librería necesaria

```
library(cluster)
```

```
clusplot(X.s, cl.kmeans,
         main="Componentes principales del Top 50 canciones más escuchadas 2021")

text(princomp(X.s)$score[,1:2],
     labels=rownames(X.s), pos=1, col="orange")
```




Este gráfico nos da una visualización de los cluster de una manera practica, dejando ver que el cluster 1 es el que tiene mas observaciones y el cluster 2 es un poco más disperso y con menos observaciones,pero es claro que los clusters están bien clasificadas nuestras observaciones, cluster 1 es en el que se encuentra la canción **Montero** y el cluster dos es en el que está la canción **deja vu**.

Ademas de que en estos dos componentes principales explicamos el 66.46 % de la variiabilidad que es un buen numero.

## Silhouette

Representación gráfica de la eficacia de clasificación de una observación dentro de un grupo.

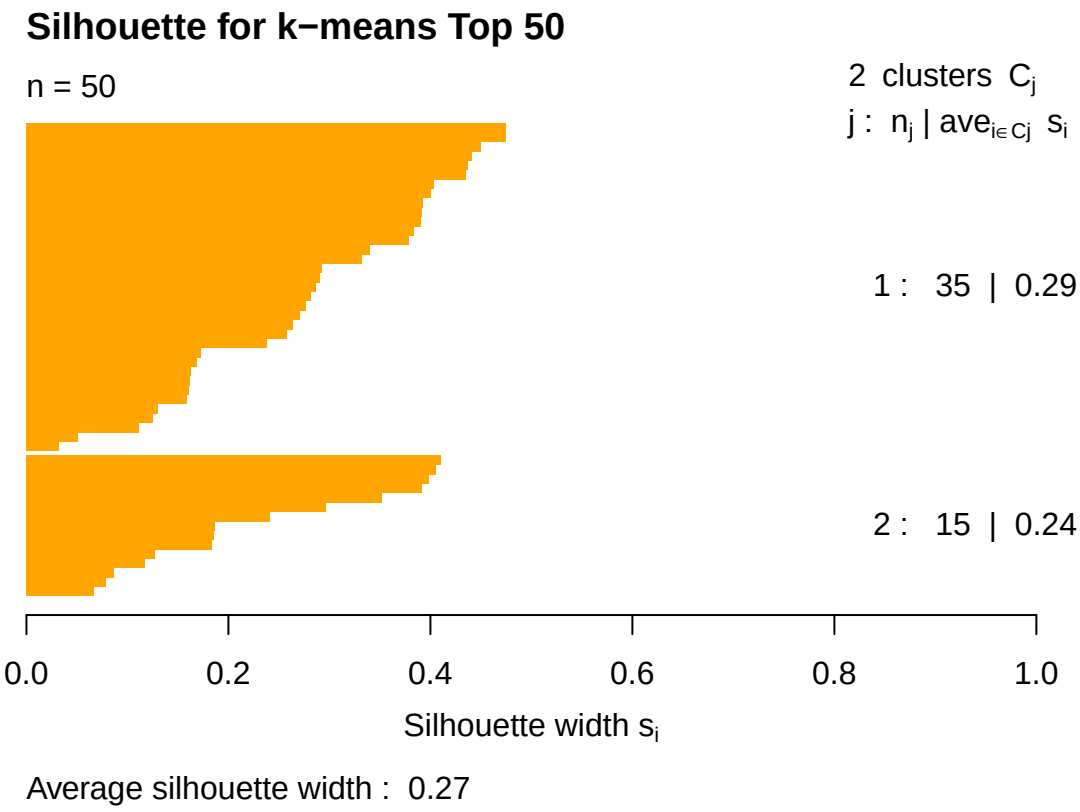
1.- Generación de los cálculos

Para realizar estos cálculos nos apoyamos de la distancia euclídea, de la matriz escalada y creamos un objeto para realizar el silhouette con la distancia euclídea calculada y los clusters.

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.kmeans<-silhouette(cl.kmeans, dist.Euc)
```

2.- Generación del gráfico

```
plot(Sil.kmeans, main="Silhouette for k-means Top 50 ",
     col="orange")
```




**Interpretación**

Observamos los anchos para cada silhouette por cluster es medianamente bueno en el 1 y 2 sin observaciones negativas, también como antes los mencionamos el cluster 1 su silhouette es el que más observaciones tiene aquí también se ve que tiene 35 observaciones y un ancho de *0.29*, en el cluster 2 su silhouette es mas pequeño y de *0.24* con 15 observaciones.

Adicionalmente el **average silhouette width**(ancho medio del silhouette)=0.27 no es muy alto debería ser mejor para estar seguros de que la cantidad de cluster que proponemos es la correcta, pero con simulaciones echas previas al reporte el tener dos grupos (clusters) para este análisis es el más optimo.

# Capitulo 5

## Concluciones

En conclusión hemos podido agrupar de manera correcta el **Top 50 canciones más escuchadas en Spotify 2021** obteniendo resultados satisfactorios para poder decir que la siguiente lista:

- Kiss Me More - Woman - Telepatía - Bandido - Dynamite - Leve the door open - Watermelon sugar
- Industry baby - Montero - Sweater Weather - Fiel - I wana be your slave - yonagui - Pepas
- Todo de ti - Blinding lights - Pareja del año - Good 4 u - Qué más pues? - Dont start now - La noche de anoche
- Friday Dopamine - Bad habits - Stay - beggin - Levitating - Head & heart - Lemonade
- Positions - Mood - Save your tears - Butter - astronaut in the ocean - Paches - 34 + 35

Que conforman el cluster 1 todas las canciones son de diferentes géneros,artistas,etc.Con nuestras variables de estudio comparten características que los agrupan dentro de el mismo.

Para el cluster 2 las siguientes canciones se agrupan en la siguiente lista:

- Drivers license - Heat weves - Need to konw
- Streets - The business - Traitor
- You broke me first - without you - Dákiti
- Some one you loved - Good days - Deja vu
- Arcade - Heart break Anniversary - Rap start

Son lo suficientemente parecidas para estar en este grupo y de igual forma son canciones que son de distintos géneros pero con las variables qué tenemos comparten características para estar agrupadas

Podemos decir que estamos conformes con la clasificación que nuestro estudio a arrogado; si queremos un desglose más puntual deberíamos hacer este análisis con ayuda de un experto en música y descubrir cosas que un estadístico desconoce.

# Referencias

La base de datos fue extraída de un portal de free data base: https://www.kaggle.com/datasets/equinxx/spotify-top-50-songs-in-2021

De la presentación de la profesora Lestrad Rodal Yendi: https://uvmx.sharepoint.com/:u:/s/EstadsticaMultivariada/EY_KQgBngppCiQ7_d_Z4W6cBV9WH5ap0Uw0uKjaGY6prdQ?e=754ygC

### NOTA IMPORTANTE

El libro en el que baso sus clases no lo encontré ni en Teams mucho menos en EMINUS; por eso no pude ponerlo en las referencias.