

Proyecto1

Juan Luis Solórzano (carnet: 201598)

Micaela Yataz (carnet: 18960)

2025-01-20

```
##          id          budget          genres          homePage
## Min.      :      5    Min.      :      0    Length:10000    Length:10000
## 1st Qu.: 12286    1st Qu.:      0    Class :character    Class :character
## Median :152558    Median :   500000    Mode  :character    Mode  :character
## Mean    :249877    Mean    : 18551632
## 3rd Qu.:452022    3rd Qu.: 20000000
## Max.     :922260    Max.     :380000000
## productionCompany productionCompanyCountry productionCountry
## Length:10000      Length:10000      Length:10000
## Class :character  Class :character    Class :character
## Mode  :character  Mode  :character    Mode  :character
##
##
##
##          revenue          runtime          video          director
## Min.      :0.000e+00    Min.      : 0.0    Mode :logical    Length:10000
## 1st Qu.:0.000e+00    1st Qu.: 90.0    FALSE:9430      Class :character
## Median :1.631e+05     Median :100.0    TRUE :84        Mode  :character
## Mean    :5.674e+07     Mean    :100.3    NA's :486
## 3rd Qu.:4.480e+07     3rd Qu.:113.0
## Max.     :2.847e+09     Max.     :750.0
##          actors          actorsPopularity          actorsCharacter          originalTitle
## Length:10000      Length:10000      Length:10000      Length:10000
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##          title          originalLanguage          popularity          releaseDate
## Length:10000      Length:10000      Min.      :   4.258    Length:10000
## Class :character  Class :character  1st Qu.:   14.578    Class :character
## Mode  :character  Mode  :character  Median :   21.906    Mode  :character
##                                     Mean      :   51.394
##                                     3rd Qu.:   40.654
##                                     Max.      :11474.647
##          voteAvg          voteCount          genresAmount          productionCoAmount
## Min.      : 1.300    Min.      :    1    Min.      : 0.000    Min.      : 0.000
## 1st Qu.: 5.900    1st Qu.:  120    1st Qu.: 2.000    1st Qu.: 2.000
## Median : 6.500    Median :   415    Median : 3.000    Median : 3.000
## Mean    : 6.483    Mean    : 1342    Mean    : 2.596    Mean    : 3.171
## 3rd Qu.: 7.200    3rd Qu.: 1316    3rd Qu.: 3.000    3rd Qu.: 4.000
## Max.     :10.000    Max.     :30788    Max.     :16.000    Max.     :89.000
## productionCountriesAmount actorsAmount    castWomenAmount
```

```
## Min.      : 0.000           Min.      :    0   Length:10000
## 1st Qu.:  1.000           1st Qu.:   13   Class :character
## Median :  1.000           Median :   21   Mode  :character
## Mean    :  1.751           Mean    :  2148
## 3rd Qu.:  2.000           3rd Qu.:   36
## Max.    :155.000           Max.     :919590
## castMenAmount
## Length:10000
## Class :character
## Mode  :character
##
##
##
```

1. Clustering

1.1. Haga el preprocesamiento del dataset, explique qué variables no aportan información a la generación de grupos y por qué. Describa con qué variables calculará los grupos.

Como el algoritmo de k-medias y el clustering jerárquico necesitan de alguna medida de distancia, entre los datos, en una primera instancia vamos a tomar solo las variables numéricas y vamos a quitar el id por ser como el nombre de una película. Las variables que tomaremos en consideración son las siguientes:

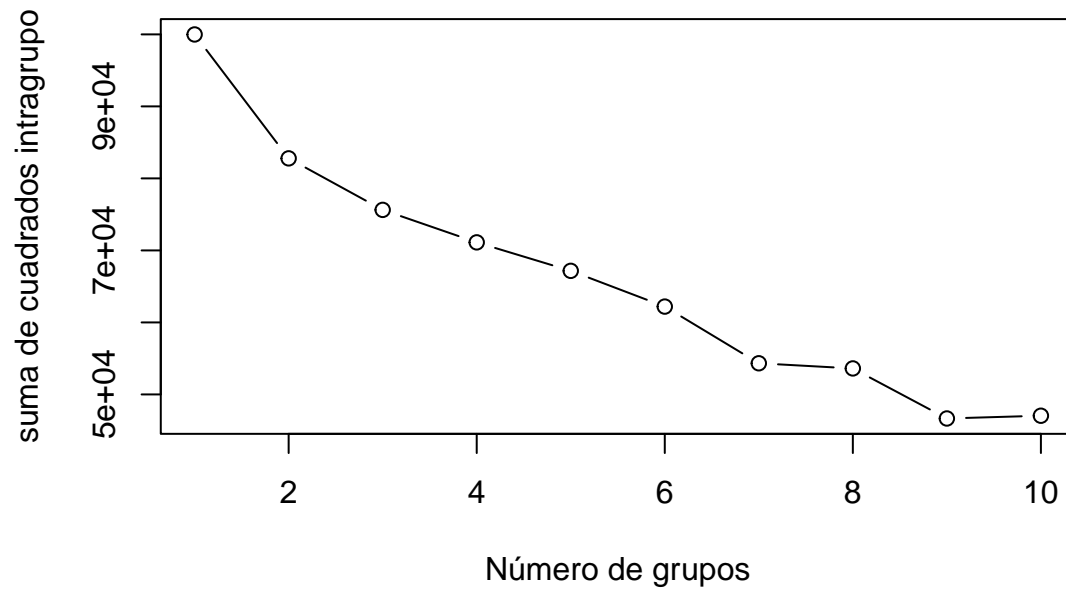
```
## 'data.frame': 10000 obs. of 10 variables:
## $ budget      : int  4000000 21000000 11000000 94000000 55000000 15000000 839727 128000000
## $ revenue     : num  4.26e+06 1.21e+07 7.75e+08 9.40e+08 6.77e+08 ...
## $ runtime     : int  98 110 121 100 142 122 119 141 126 149 ...
## $ popularity  : num  20.9 9.6 100 134.4 58.8 ...
## $ voteAvg     : num  5.7 6.5 8.2 7.8 8.5 8 8 7.9 7.5 8.2 ...
## $ voteCount   : int  2077 223 16598 15928 22045 9951 4253 1335 8726 1963 ...
## $ genresAmount : int  2 3 3 2 3 1 2 2 5 2 ...
## $ productionCoAmount : int  2 3 2 1 2 2 2 26 2 1 ...
## $ productionCountriesAmount: int  1 2 1 1 1 1 1 12 1 1 ...
## $ actorsAmount : int  25 15 105 24 76 40 152 29 117 24 ...
```

1.2. Analice la tendencia al agrupamiento usando el estadístico de Hopkins y la VAT (Visual Assessment of cluster Tendency). Esta última hágala si es posible, teniendo en cuenta las dimensiones del conjunto de datos. Discuta sus resultados e impresiones.

El estadístico de Hopkins es de 1 que es lejano a 0.5, entonces los datos no son aleatorios. Sin embargo no haremos un VAT por ser difícil de visualizar e interpretar con 10 variables.

1.3. Determine cuál es el número de grupos a formar más adecuado para los datos que está trabajando. Haga una gráfica de codo y explique la razón de la elección de la cantidad de clústeres con la que trabajará.

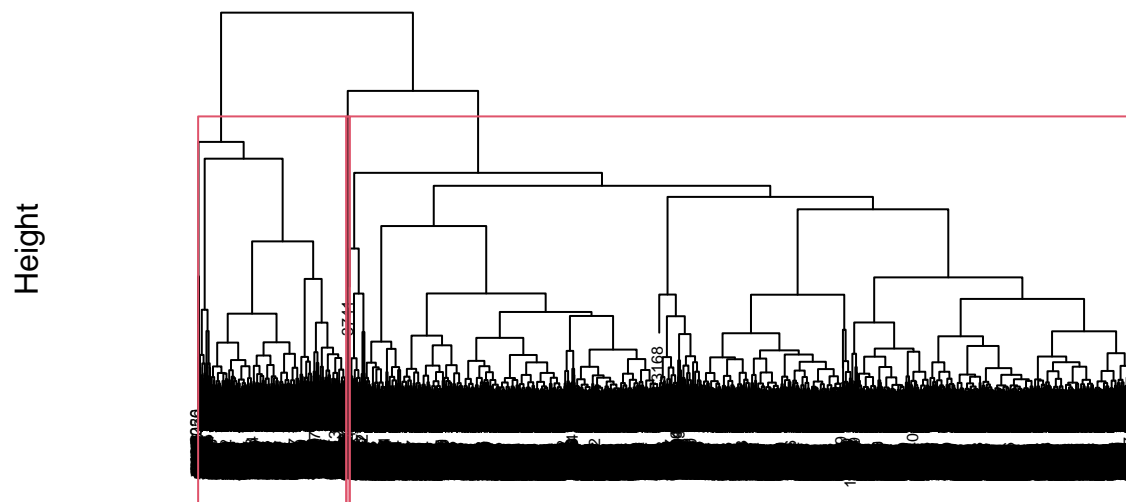
Para ello se usará el método de Codo



Como a partir de 3 grupos en adelante la suma de cuadrados intragrupo no disminuye tan rápido se elegirán 3 grupos.

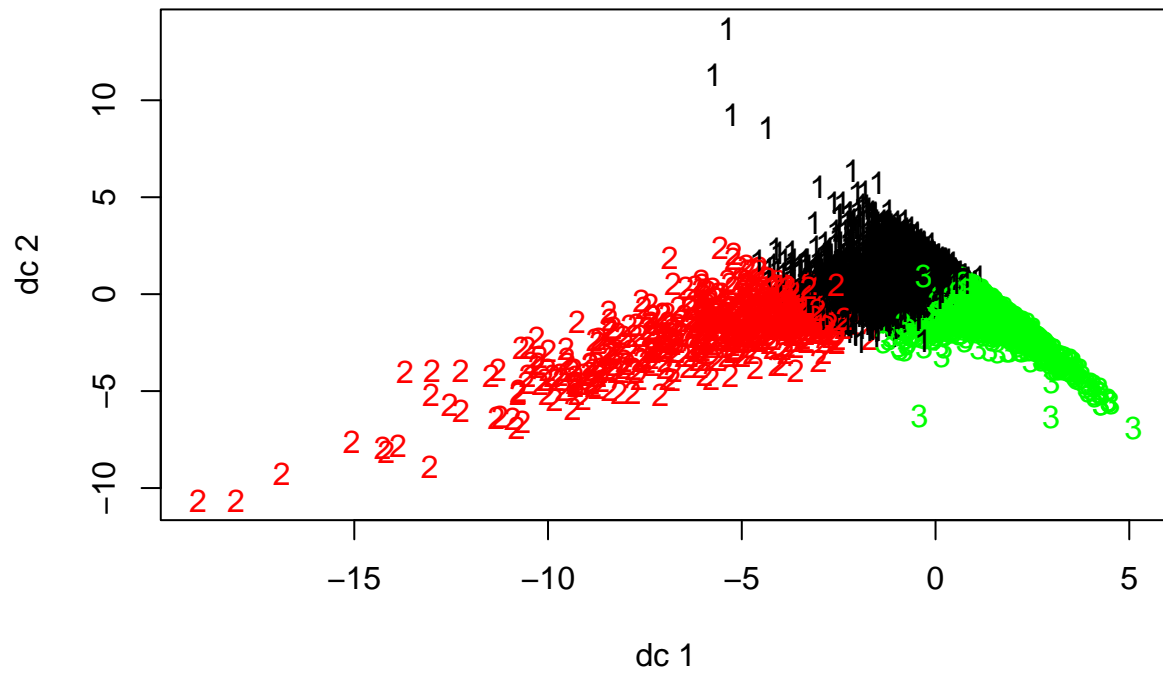
1.4. Utilice los algoritmos k-medias y clustering jerárquico para agrupar. Compare los resultados generados por cada uno.

Cluster Dendrogram

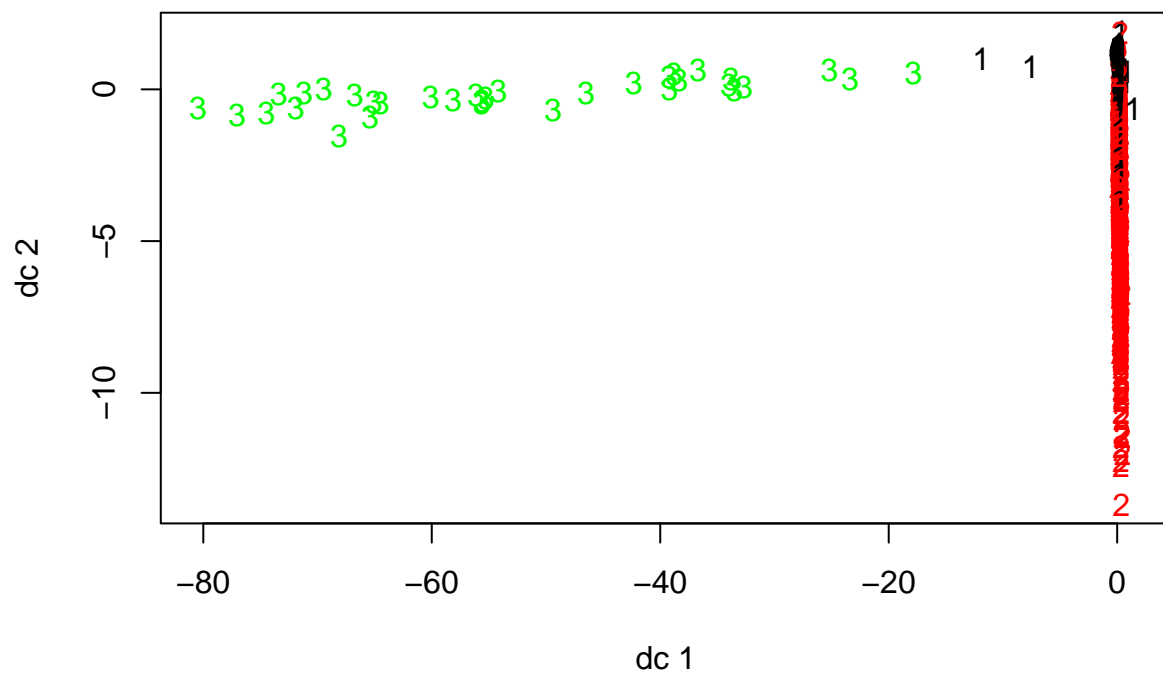


D
hclust (*, "ward.D2")

Clusters generados por K-means



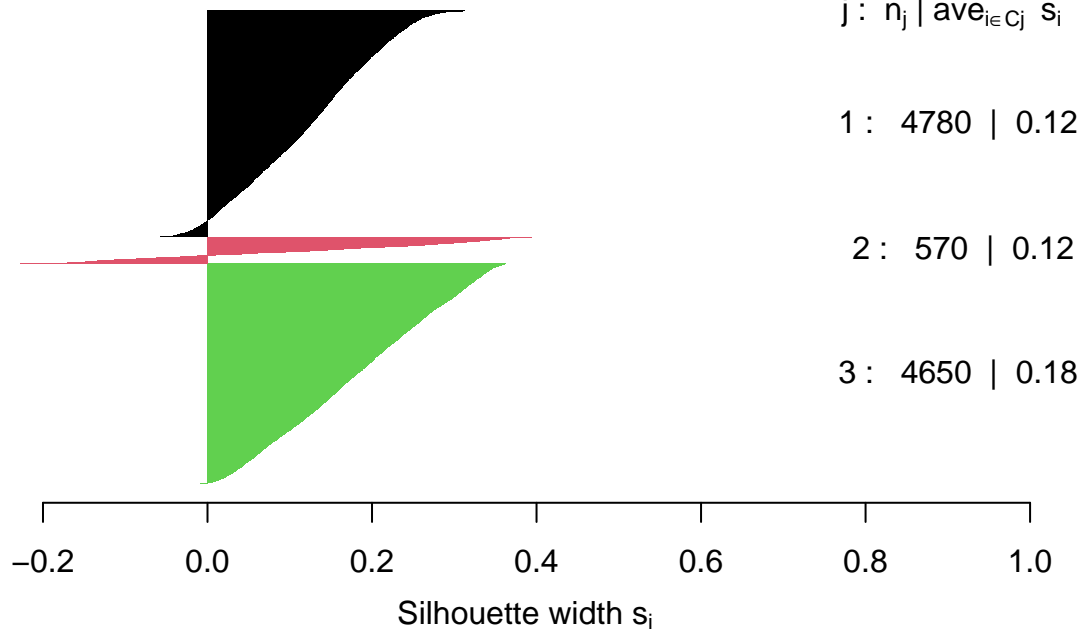
Clusters generados por Clustering Jerárquico



1.5. Determine la calidad del agrupamiento hecho por cada algoritmo con el método de la silueta. Discuta los resultados.

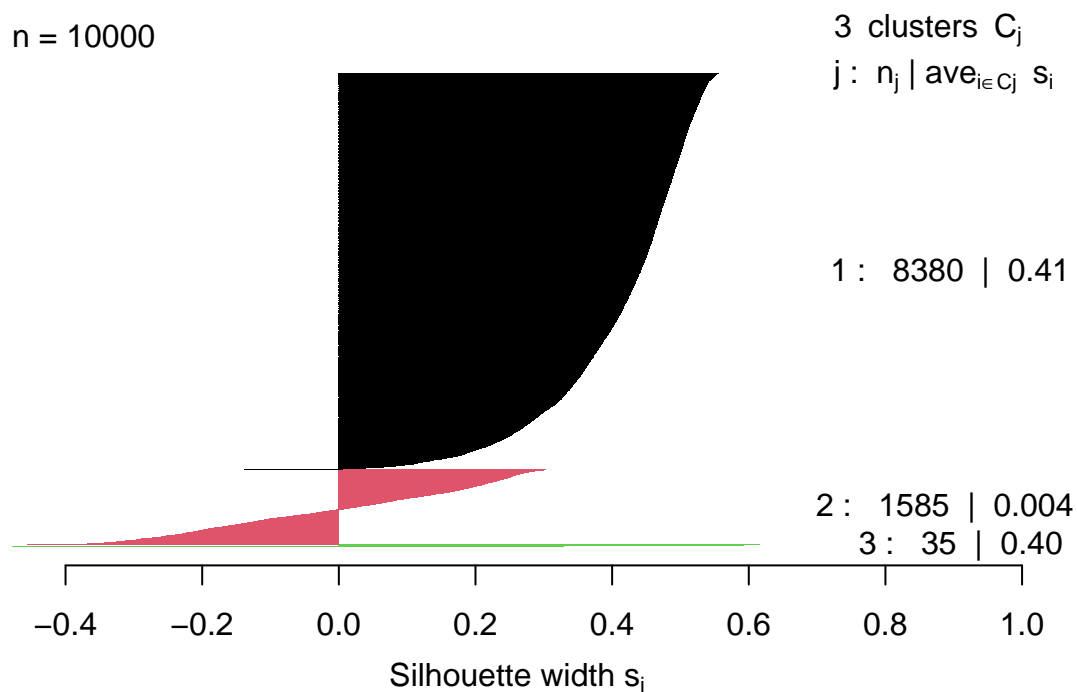
Silueta K-means

$n = 10000$



Silueta Clustering Jerárquico

$n = 10000$



Usando el método de la silueta el clustering jerárquico que tiene una silueta promedio de $0.35 > 0.15$ del clustering de Kmedias. Entonces para estos datos el clusterin jerárquico obtuvo un mejor resultado. Esto tiene sentido pues en la entrega anterior vimos que ninguna variable se comportaba de manera normal. En estos casos el k-medias no suele ser tan eficiente.

1.6. Interprete los grupos basado en el conocimiento que tiene de los datos. Recuerde investigar las medidas de tendencia central de las variables continuas y las tablas de frecuencia de las variables categóricas pertenecientes a cada grupo. Identifique hallazgos interesantes debido a las agrupaciones y describa para qué le podría servir.

```
## Tamaños de los dataframes:
```

```
## K-Means -> Cluster 1: 4780 | Cluster 2: 570 | Cluster 3: 4650
```

```
## Jerárquico -> Cluster 1: 8380 | Cluster 2: 1585 | Cluster 3: 35
```

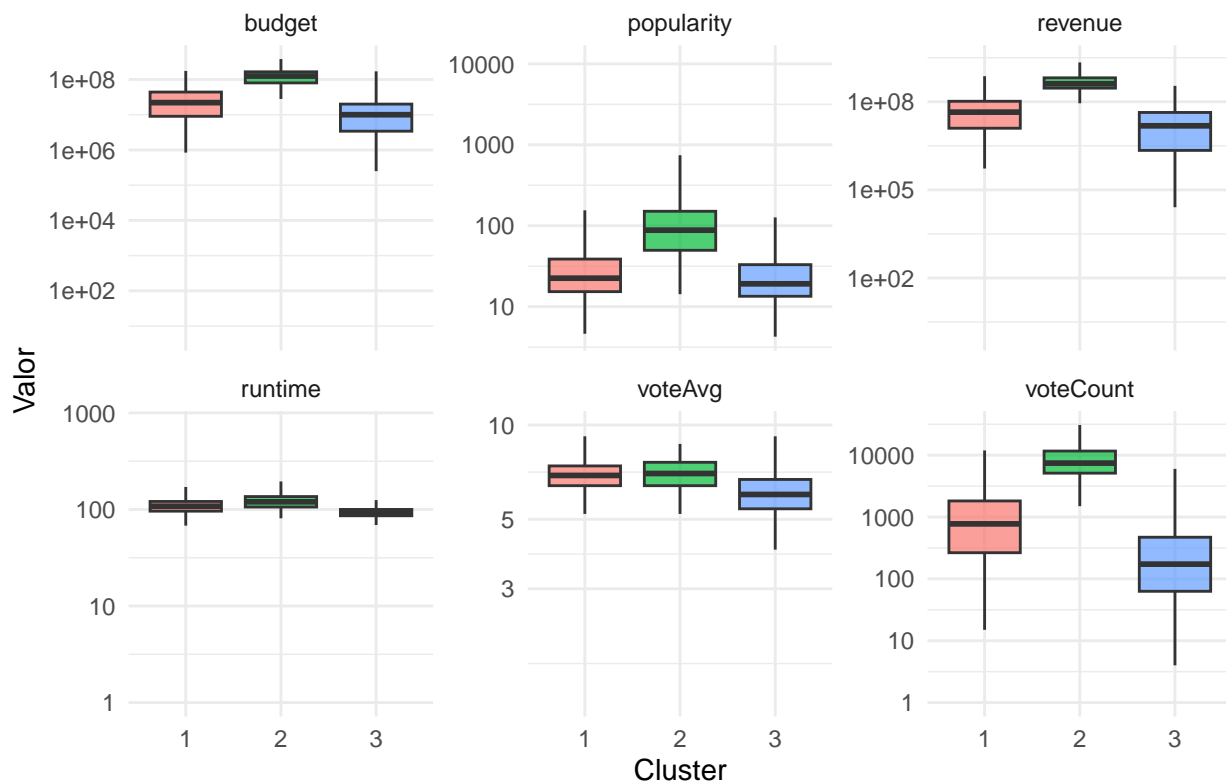
Como es difícil interpretar los números puros vamos a hacer unas gráficas de caja y bigote.

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
```

```
## Warning: Removed 9688 rows containing non-finite outside the scale range
```

```
## (`stat_boxplot()`).
```

Distribución de Variables por Clusters (K-Means)



Clusters generados por K-Means:

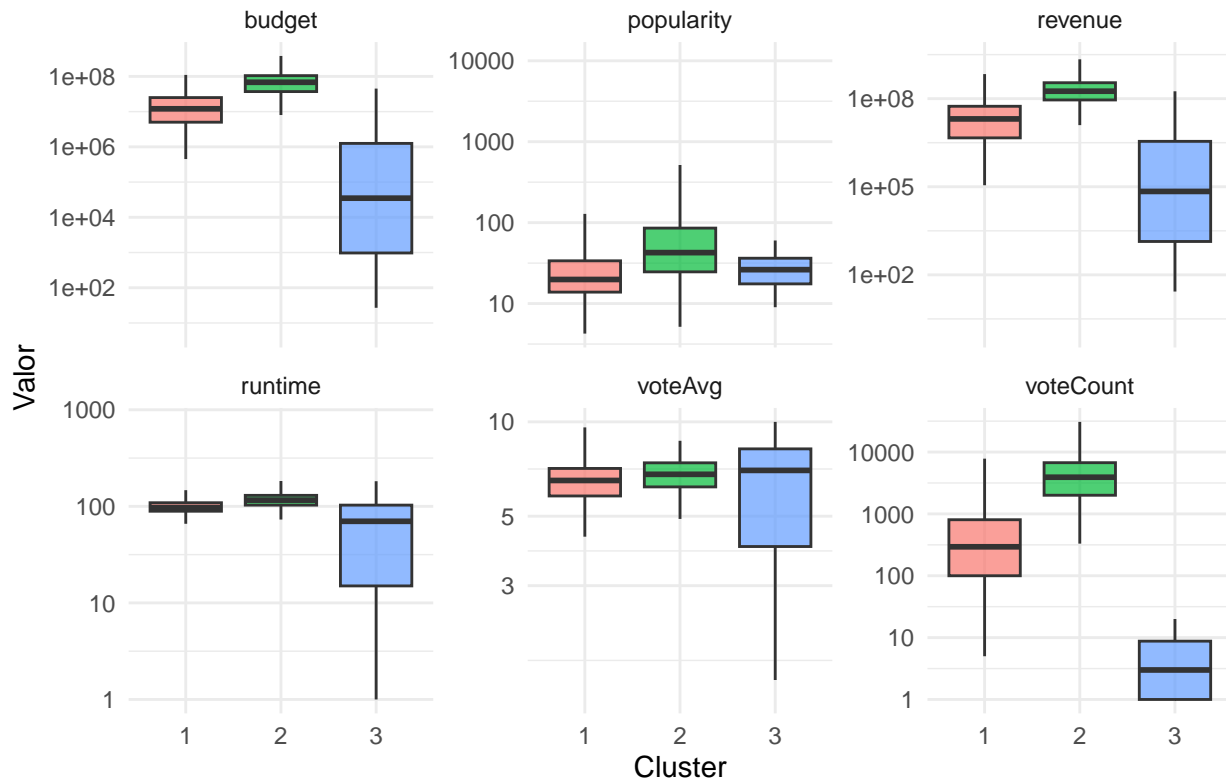
El grupo 1: Presenta valores intermedios en presupuesto, recaudación y popularidad. Probablemente incluye películas de presupuesto medio y recaudación media.

El grupo 2: Contiene películas con los presupuestos y recaudaciones más altos, alta popularidad y una gran cantidad de votos.

El grupo 3: Se caracteriza por películas con presupuestos y recaudaciones bajas, menor popularidad y poca cantidad de votos.

```
## Warning in scale_y_log10(): log-10 transformation introduced infinite values.
## Warning: Removed 9688 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

Distribución de Variables por Clusters (Jerárquico)



Clusters generados por Clustering Jerárquico:

El grupo 1: Incluye películas con recaudación y presupuesto moderado, popularidad mas bien baja.
 El grupo 2: Películas de alto presupuesto, altas recaudacion y popularidad
 El grupo 3: Agrupa peliculas de bajo presupuesto, con popularidades relativamente bajas y recaudaciones

2. Reglas de Asociación

2.1. Obtenga reglas de asociación interesantes del conjunto de datos usando el algoritmo “A priori”. Recuerde discretizar las variables numéricas. Genere reglas con diferentes niveles de confianza y soporte. Discuta los resultados. Si considera que debe eliminar variables

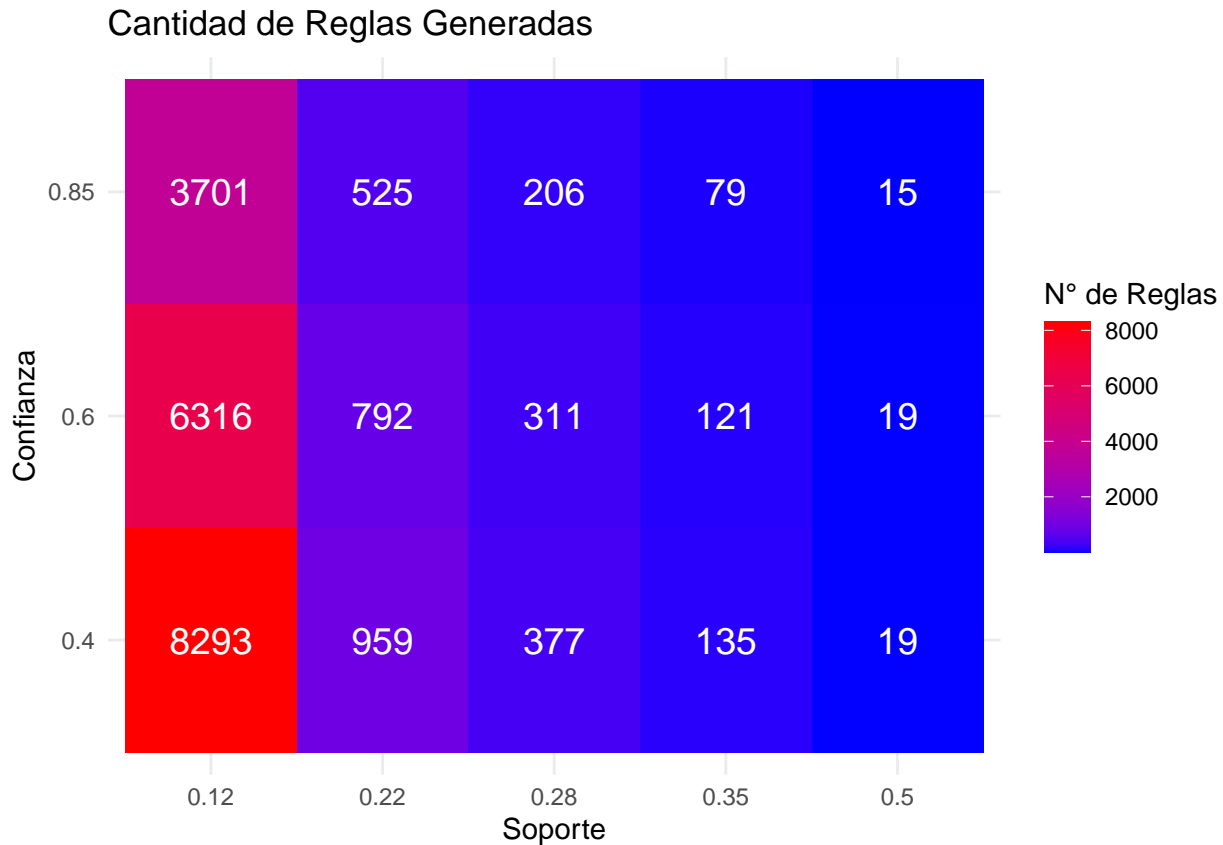
```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Attaching package: 'arules'
## The following object is masked from 'package:flexclust':
```

```
##
##      info
## The following object is masked from 'package:modeltools':
##
##      info
## The following object is masked from 'package:dplyr':
##
##      recode
## The following objects are masked from 'package:base':
##
##      abbreviate, write
## Warning: There were 3 warnings in `mutate()`.
## The first warning was:
## i In argument: `across(where(is.numeric), discretize)`.
## Caused by warning:
## ! The calculated breaks are: 0, 0, 1.2e+07, 3.8e+08
## Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.
## i Run `dplyr::last_dplyr_warnings()` to see the 2 remaining warnings.
## Warning: Column(s) 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18,
## 19, 20, 21, 22, 23, 24, 25, 26, 27 not logical or factor. Applying default
## discretization (see '? discretizeDF').
## Warning in discretize(x = c(4000000L, 21000000L, 11000000L, 94000000L, 55000000L, : The calculated breaks are:
## Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.
## Warning in discretize(x = c(4257354, 12136938, 775398007, 940335536, 677387716, : The calculated breaks are:
## Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.
## Warning in discretize(x = c(1L, 2L, 1L, 1L, 1L, 1L, 1L, 12L, 1L, 1L, 1L, : The calculated breaks are:
## Only unique breaks are used reducing the number of intervals. Look at ? discretize for details.
```

```
library(ggplot2)
```

```
# Crear el gráfico de calor
```

```
ggplot(resultados, aes(x = factor(Soporte), y = factor(Confianza), fill = Num_Reglas)) +
  geom_tile() +
  geom_text(aes(label = Num_Reglas), color = "white", size = 5) +
  scale_fill_gradient(low = "blue", high = "red") +
  labs(title = "Cantidad de Reglas Generadas",
       x = "Soporte",
       y = "Confianza",
       fill = "N° de Reglas") +
  theme_minimal()
```

Con el mapa de calor de el numero de reglas tenemos una idea de que soporte y confianza es buena idea elegir. Para discutir sobre las variables que mas importan para las reglas y las que menos vamos a elegir un soporte de 30% y una confianza de 60%

```
## Apriori
##
## Parameter specification:
## confidence minval smax arem aval originalSupport maxtime support minlen
##      0.85    0.1    1 none FALSE                TRUE     5     0.5     3
## maxlen target  ext
##      7  rules TRUE
##
## Algorithmic control:
## filter tree heap memopt load sort verbose
##    0.1 TRUE TRUE  FALSE TRUE     2    TRUE
##
## Absolute minimum support count: 5000
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[78214 item(s), 10000 transaction(s)] done [0.04s].
## sorting and recoding items ... [7 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [15 rule(s)] done [0.00s].
## creating S4 object ... done [0.00s].

##      lhs                                rhs
## [1] {budget=[0,1.2e+07),
```

support confidence

##	revenue=[0,2.03e+07)}	=> {Cluster_Jerarquico=1}	0.5903	0.9909350
## [2]	{budget=[0,1.2e+07),			
##	Cluster_Jerarquico=1}	=> {revenue=[0,2.03e+07)}	0.5903	0.9049517
## [3]	{revenue=[0,2.03e+07),			
##	Cluster_Jerarquico=1}	=> {budget=[0,1.2e+07)}	0.5903	0.9027374
## [4]	{budget=[0,1.2e+07),			
##	revenue=[0,2.03e+07)}	=> {productionCountriesAmount=[1,155]}	0.5870	0.9853953
## [5]	{budget=[0,1.2e+07),			
##	productionCountriesAmount=[1,155]}	=> {revenue=[0,2.03e+07)}	0.5870	0.8948171
## [6]	{revenue=[0,2.03e+07),			
##	productionCountriesAmount=[1,155]}	=> {budget=[0,1.2e+07)}	0.5870	0.8922329
## [7]	{budget=[0,1.2e+07),			
##	Cluster_Jerarquico=1}	=> {productionCountriesAmount=[1,155]}	0.6457	0.9898820
## [8]	{budget=[0,1.2e+07),			
##	productionCountriesAmount=[1,155]}	=> {Cluster_Jerarquico=1}	0.6457	0.9842988
## [9]	{revenue=[0,2.03e+07),			
##	Cluster_Jerarquico=1}	=> {productionCountriesAmount=[1,155]}	0.6473	0.9899067
## [10]	{revenue=[0,2.03e+07),			
##	productionCountriesAmount=[1,155]}	=> {Cluster_Jerarquico=1}	0.6473	0.9838881
## [11]	{originalLanguage=en,			
##	Cluster_Jerarquico=1}	=> {productionCountriesAmount=[1,155]}	0.6188	0.9934179
## [12]	{budget=[0,1.2e+07),			
##	revenue=[0,2.03e+07),			
##	Cluster_Jerarquico=1}	=> {productionCountriesAmount=[1,155]}	0.5837	0.9888192
## [13]	{budget=[0,1.2e+07),			
##	revenue=[0,2.03e+07),			
##	productionCountriesAmount=[1,155]}	=> {Cluster_Jerarquico=1}	0.5837	0.9943782
## [14]	{budget=[0,1.2e+07),			
##	productionCountriesAmount=[1,155],			
##	Cluster_Jerarquico=1}	=> {revenue=[0,2.03e+07)}	0.5837	0.9039802
## [15]	{revenue=[0,2.03e+07),			
##	productionCountriesAmount=[1,155],			
##	Cluster_Jerarquico=1}	=> {budget=[0,1.2e+07)}	0.5837	0.9017457

Sorprende, a la base de datos le agregue las etiquetas de los clusters. La mayoría de reglas tienen algo que ver con el cluster jerárquico 1. Pero las reglas tienen intervalos muy grandes, tanto que no aportan conocimiento relevante. Yo quitaría esa variable.

3 Análisis de componentes Principales

Estudia la matriz de correlación, la agrega y explica lo que observa en ella

En el análisis se realiza a partir de una muestra de datos de 10000 películas obtenidos de la plataforma The movie DB. Se evalúa la correlación entre variables. Se presentan las variables que se incluyen en el análisis:

Índice de popularidad de la película
Presupuesto de la película
Ingreso de la película
Duración de la película
Cantidad de géneros que representan la película
Cantidad de compañías productoras que participaron en la película
Cantidad de países que se llevó a cabo la película
Número de votos en la plataforma de la película
Promedio de votos en la plataforma de la película
Índice de popularidad del elenco de la película
Cantidad de personas que actúan en la película
Cantidad de actrices en el elenco de la película
Cantidad de actores en el elenco de la película.

```
datos$castWomenAmount <- as.numeric(datos$castWomenAmount)
```

```
## Warning: NAs introduced by coercion
```

[illegible]


```
##           popularity           budget           revenue
##           0.81           0.75           0.68
##           runtime           genresAmount           productionCoAmount
##           0.72           0.64           0.52
## productionCountriesAmount           voteCount           voteAvg
##           0.48           0.75           0.55
##           castWomenAmount           castMenAmount
##           0.46           0.51
```

El índice es de 0.67 lo cual es un valor regular, es suficiente pero no el ideal.

• Determina si vale la pena aplicar las componentes principales interpretando la prueba de esfericidad de Bartlett

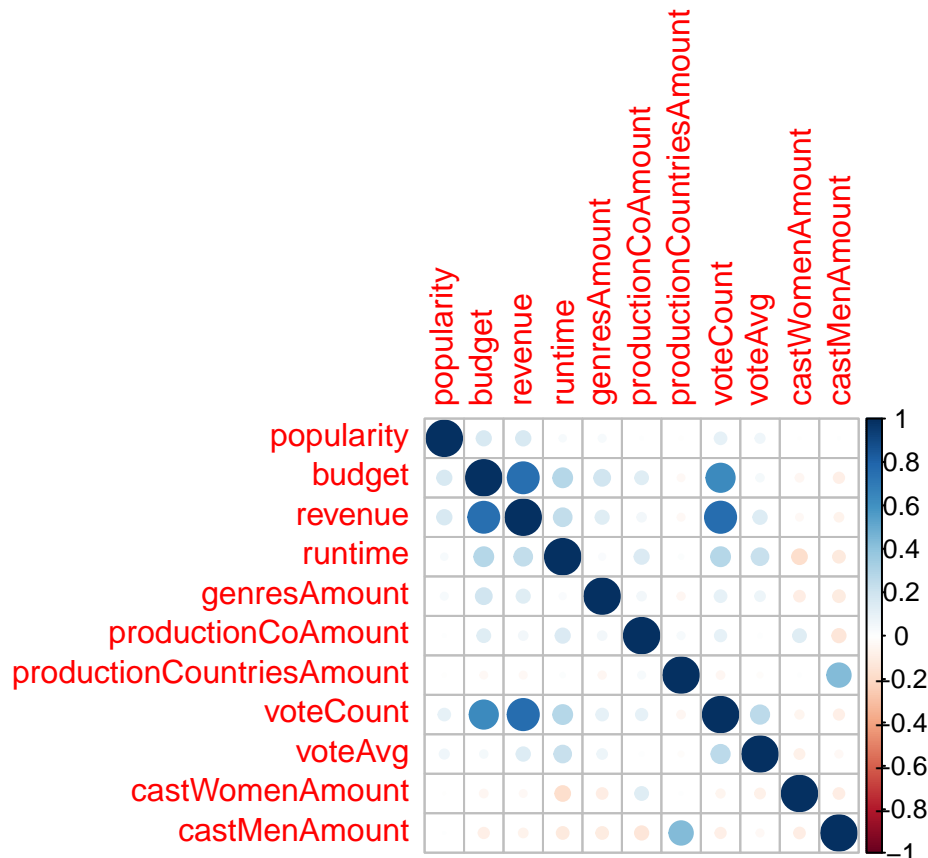
```
cortest.bartlett(sub_datos)
```

```
## R was not square, finding R from data
## $chisq
## [1] 24683.81
##
## $p.value
## [1] 0
##
## $df
## [1] 55
```

Como $P=0$ por lo que se rechaza la hipótesis nula, implicando que el análisis factorial es apropiado.

• Obtiene los componentes principales y explica cuántos seleccionará para explicar la mayor variabilidad posible.

```
matriz_cor <- cor(sub_datos, use = "pairwise.complete.obs" )
corrplot(matriz_cor)
```



Segun la tabla vemos que la

variable del presupuesto de la pelicula se correlaciona con ingreso de pelicula, asi mismo con el numero de votos en la plataforma. El numero de votos en la plataforma con el ingreso de la pelicula tambien estan correlacionados. El numero de actores en el elenco de las peliculas esta relacionado con la cantidad de paises en los que se rodo la pelicula.

#• Interpreta los coeficientes principales.

```
pca_result<-princomp(covmat=matriz_cor,use = "pairwise.complete.obs" )
```

```
## Warning: In princomp.default(covmat = matriz_cor, use = "pairwise.complete.obs") :  
## extra argument 'use' will be disregarded
```

```
#compPrinc<-prcomp(sub_datos, scale = T, use = "pairwise.complete.obs")
```

```
#compPrinc
```

```
summary(pca_result)
```

```
## Importance of components:
```

```
##              Comp.1    Comp.2    Comp.3    Comp.4    Comp.5  
## Standard deviation  1.6670613 1.2035413 1.0882368 1.0497615 0.99313055  
## Proportion of Variance 0.2526448 0.1316829 0.1076599 0.1001817 0.08966439  
## Cumulative Proportion 0.2526448 0.3843277 0.4919876 0.5921694 0.68183377  
##              Comp.6    Comp.7    Comp.8    Comp.9    Comp.10  
## Standard deviation  0.98388184 0.93595208 0.80386875 0.71186565 0.56377525  
## Proportion of Variance 0.08800213 0.07963694 0.05874591 0.04606843 0.02889478  
## Cumulative Proportion 0.76983590 0.84947284 0.90821875 0.95428717 0.98318195  
##              Comp.11  
## Standard deviation  0.43011460  
## Proportion of Variance 0.01681805  
## Cumulative Proportion 1.00000000
```

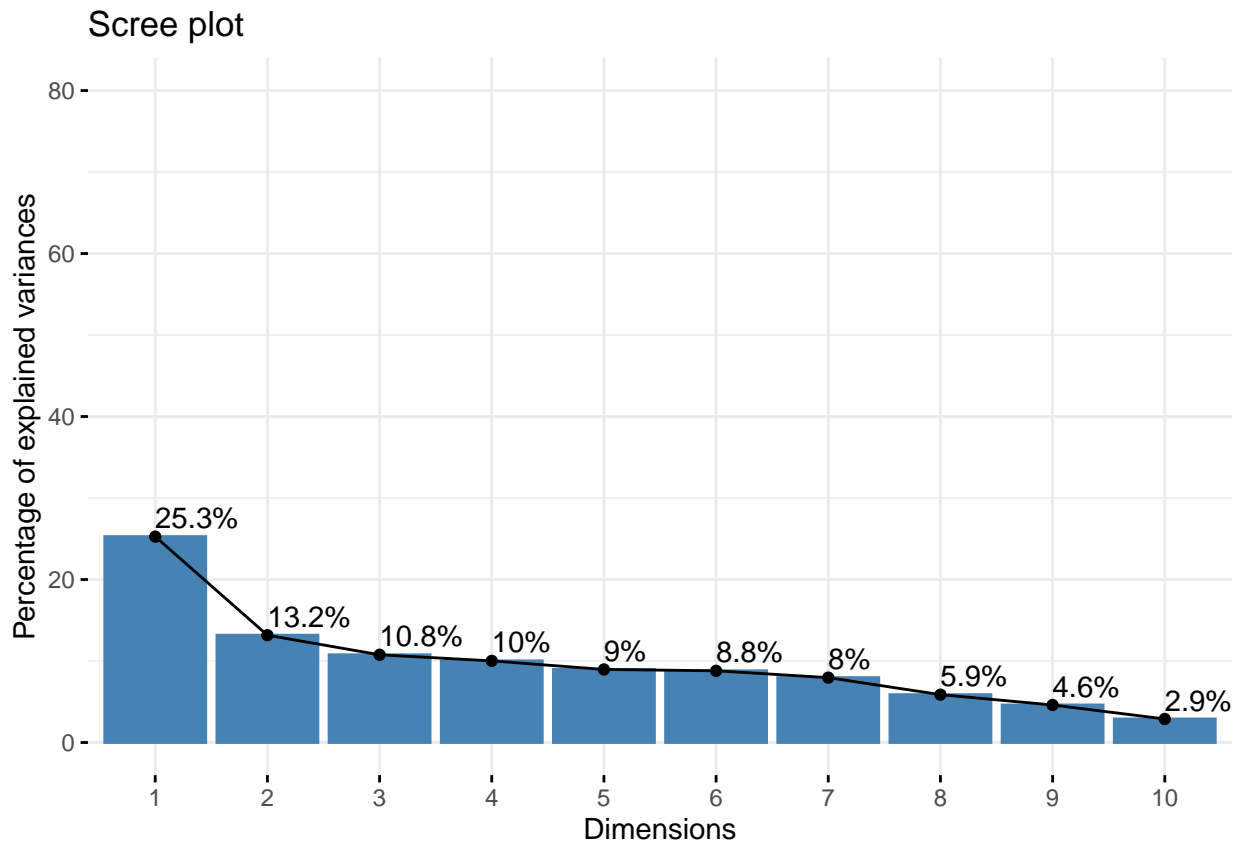
#regla de Kaiser

```
valores_propios<-pca_result$sdev^2  
valores_propios
```

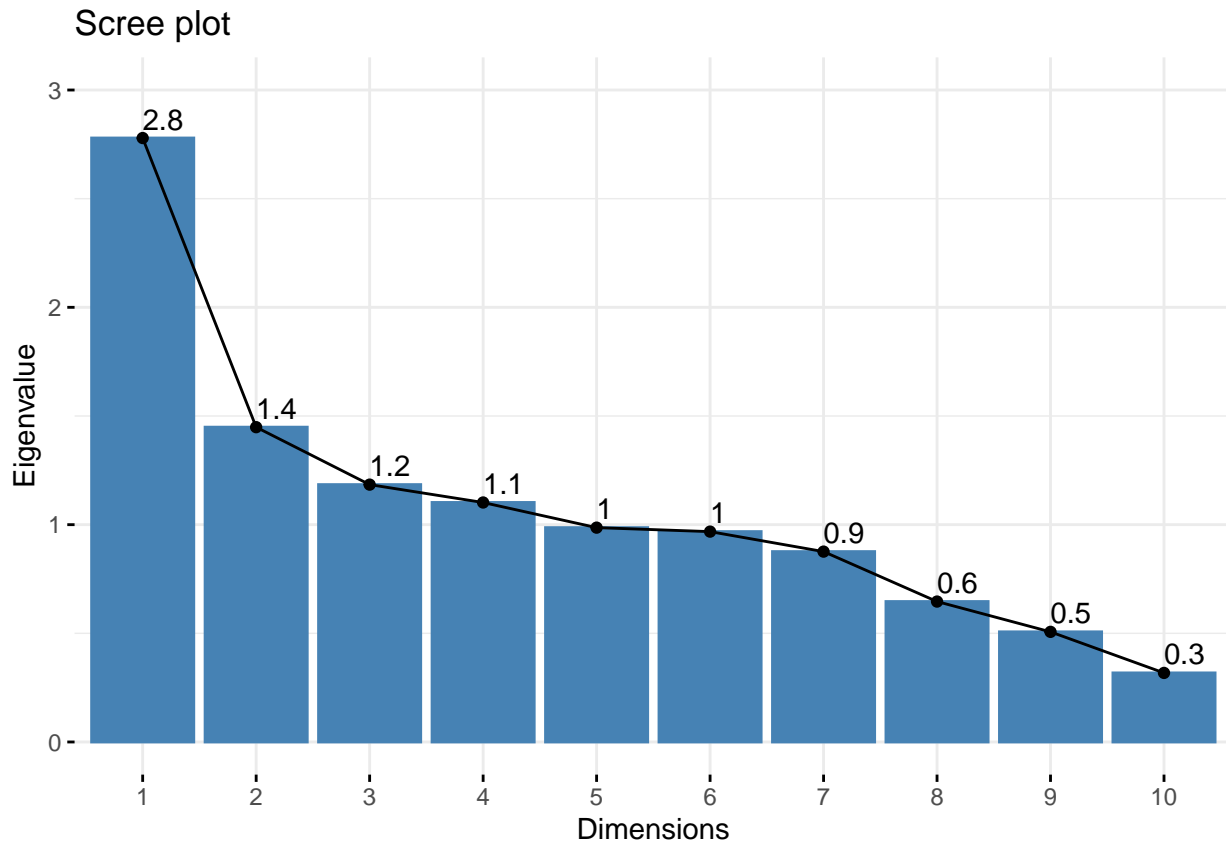
```
##      Comp.1      Comp.2      Comp.3      Comp.4      Comp.5      Comp.6      Comp.7      Comp.8  
## 2.7790932 1.4485116 1.1842592 1.1019991 0.9863083 0.9680235 0.8760063 0.6462050  
##      Comp.9      Comp.10      Comp.11  
## 0.5067527 0.3178425 0.1849986
```

Tomamos los 4 componentes principales.

```
fviz_eig(pca_result, addlabels = TRUE, ylim = c(0, 80))
```



```
fviz_eig(pca_result, addlabels = TRUE, choice = c("eigenvalue"), ylim = c(0, 3))
```



Interpretacion: EL componenete 1 se relaciona con el exito comercial y popularidad de la pelicula, las peliculas con alto presupuesto alto ingreso, y votaciones tienen valores altos en este componente.

El componente 2, puede indicar la cantidad de actores en el elenco pueden estar asociados con los países productores.

El componente 3, indica un mayor número de mujeres en el elenco tienden a tener menor puntuación.

El componente 4, Los valores altos podríamos relacionarlo con las películas independientes, es decir con menos productoras.