Universidad del Valle de Guatemala
Minería de datos
Juan Carlos Rosito
Sección 20

Mirka Monzón 18139
Andres Quinto 18288
Oscar De León 19298

Hoja de trabajo #2
Clustering

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.

```
##           variable
## 1               id
## 2   original_title
## 3 originalLanguage
## 4         homePage
## 5            video
## 6  actorsCharacter
```

Las variables listadas anteriormente, son las variables que considereamos no aportan informacion a la generacion de grupos ya que cada una de ellas tienen caracteristicas propias que no se relacionan con las demas y/o contienen informacion no usable.

2. Analice la tendencia al agrupamiento usando el estadístico de Hopkings y la VAT (Visual Assessment of cluster Tendency). Discuta sus resultados e impresiones.

Para ello necesitamos normalizar los datos de la db. Referencia para la funcion de hopkins
https://www.rdocumentation.org/packages/clustertend/versions/1.5/topics/hopkins

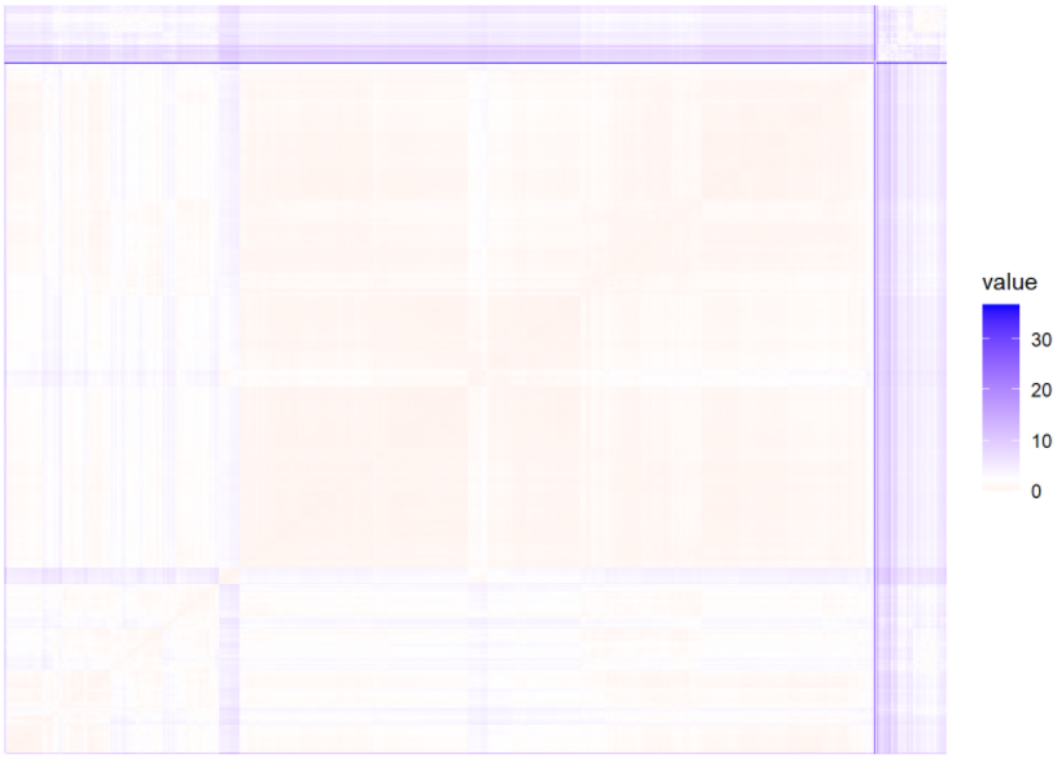
Code

```
## [1] 0.9953071
```

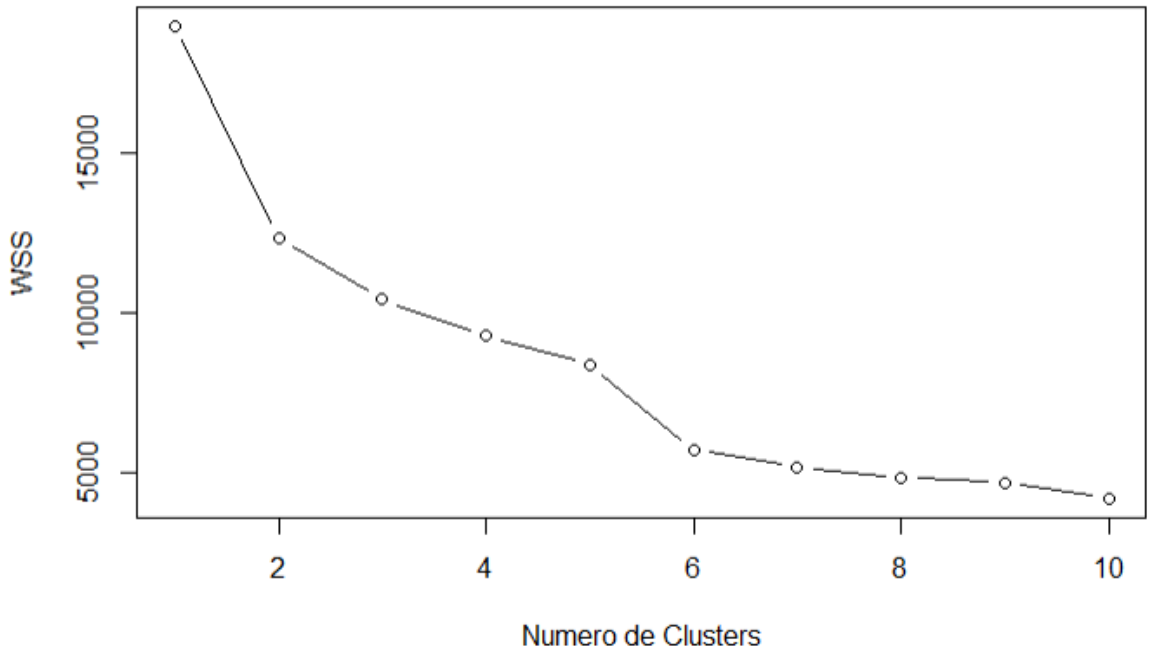
Code

Se puede apreciar que el valor que retorna la funcion de hopkins esta muy alejado de 0.5, es decir los datos recopilados no son aleatorios, dandonos a entender que el agrupamiento se puede facilitar. Para ello utilizaremos y analizaremos de forma grafica los datos.

Referencia a la libreria: https://cloud.r-project.org/web/packages/factoextra/factoextra.pdf



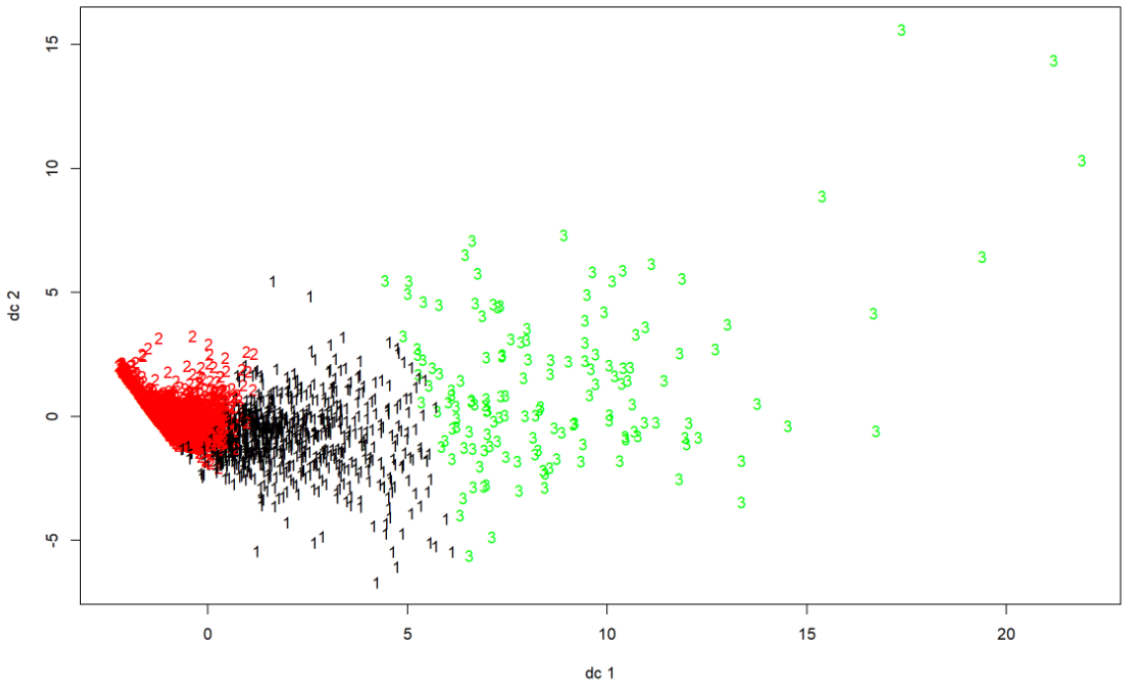
3. Determine cuál es el número de grupos a formar más adecuado para los datos que está trabajando. Haga una gráfica de codo y explique la razón de la elección de la cantidad de clústeres con la que trabajará.

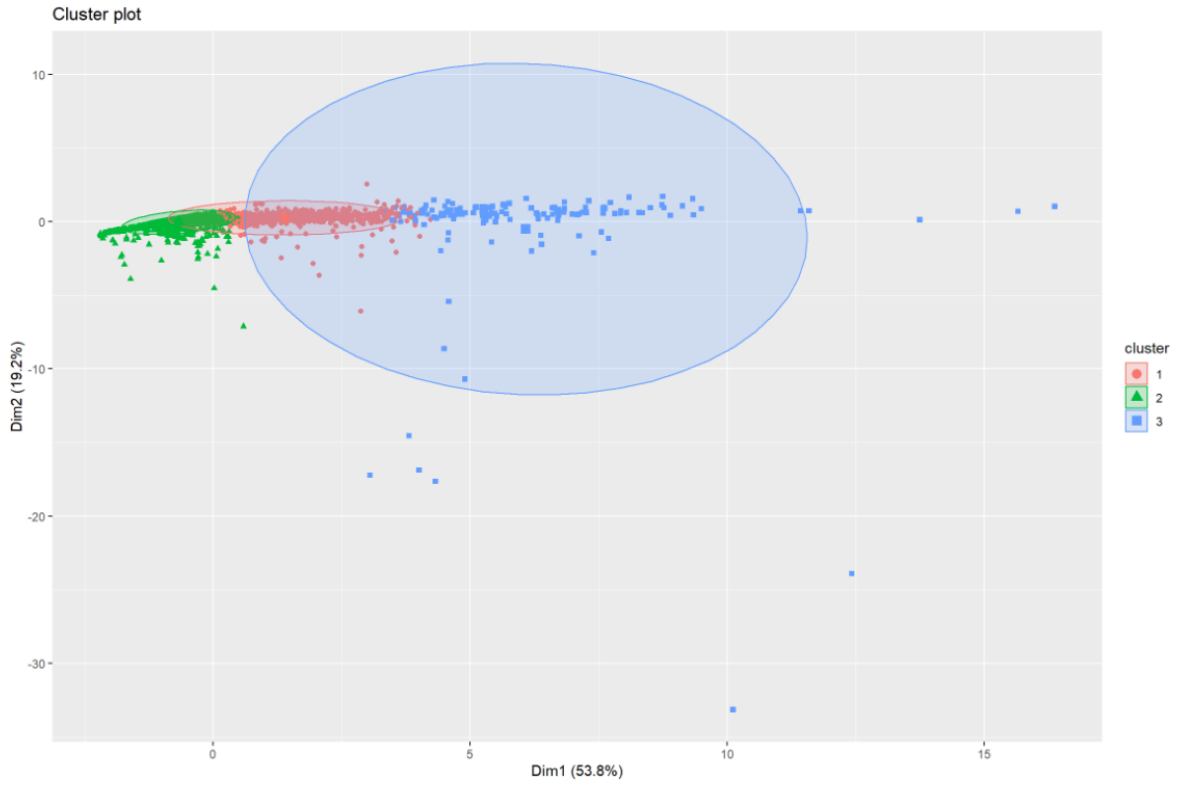


Basado en el resultado del método de codo, el número de clusters óptimo para analizar los datos es 6.

4. Utilice 3 algoritmos existentes para el agrupamiento. Compare los resultados generados por cada uno.

Agrupamiento por medio de k-means Referencia de la libreria para k-means; https://cran.r-project.org/web/packages/fpc/fpc.pdf

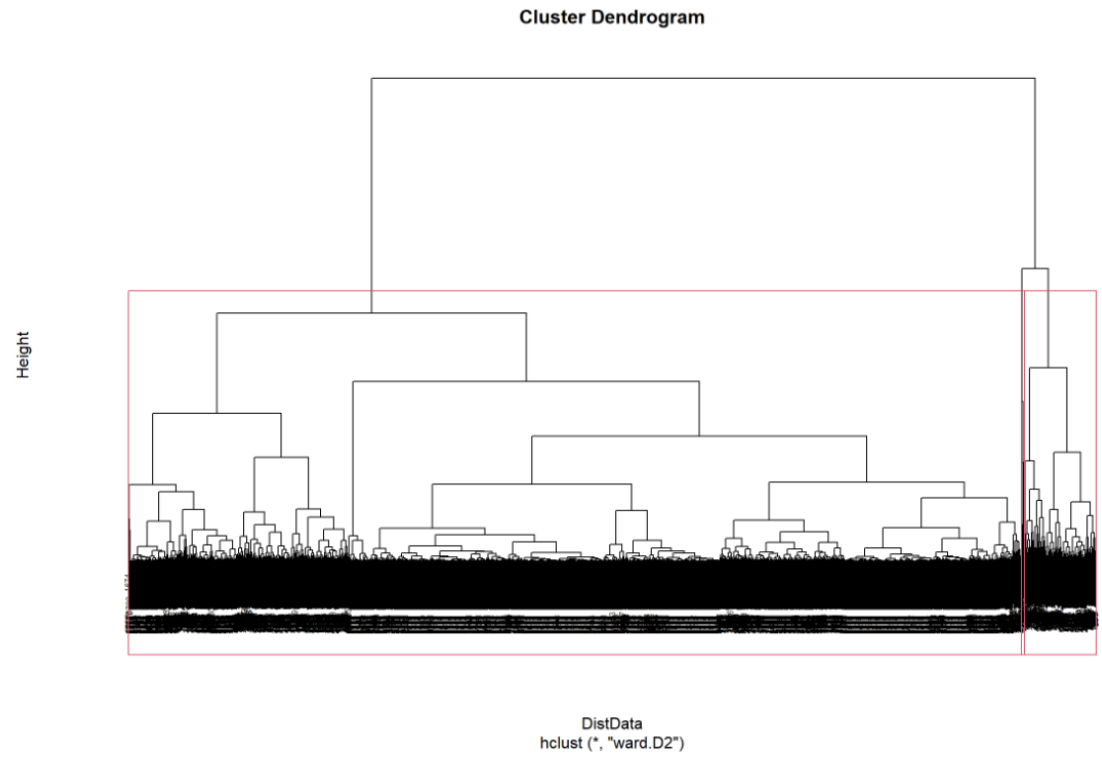




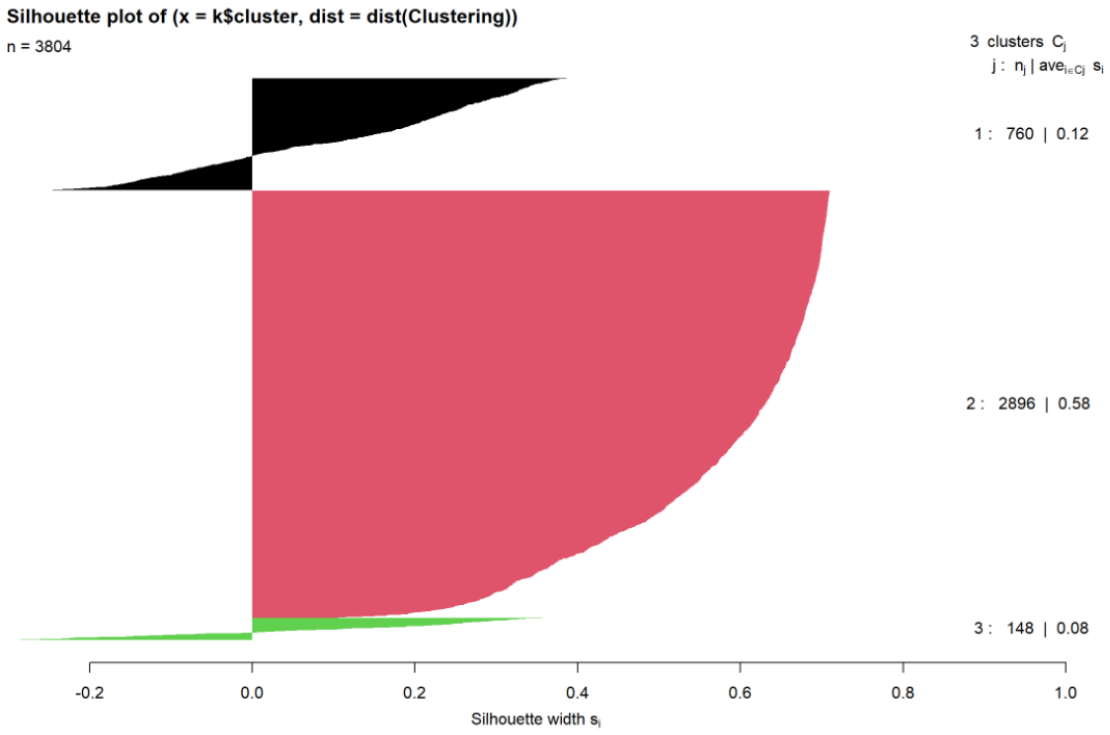

```
## [1] 0.4645744
```

Cluster jerarquico Para el cluster jerarquico se agrupan los datos basandose en la distancia que tienen entre si, de forma que los datos dentro de este cluster sean con mayor similitud entre si.

Code

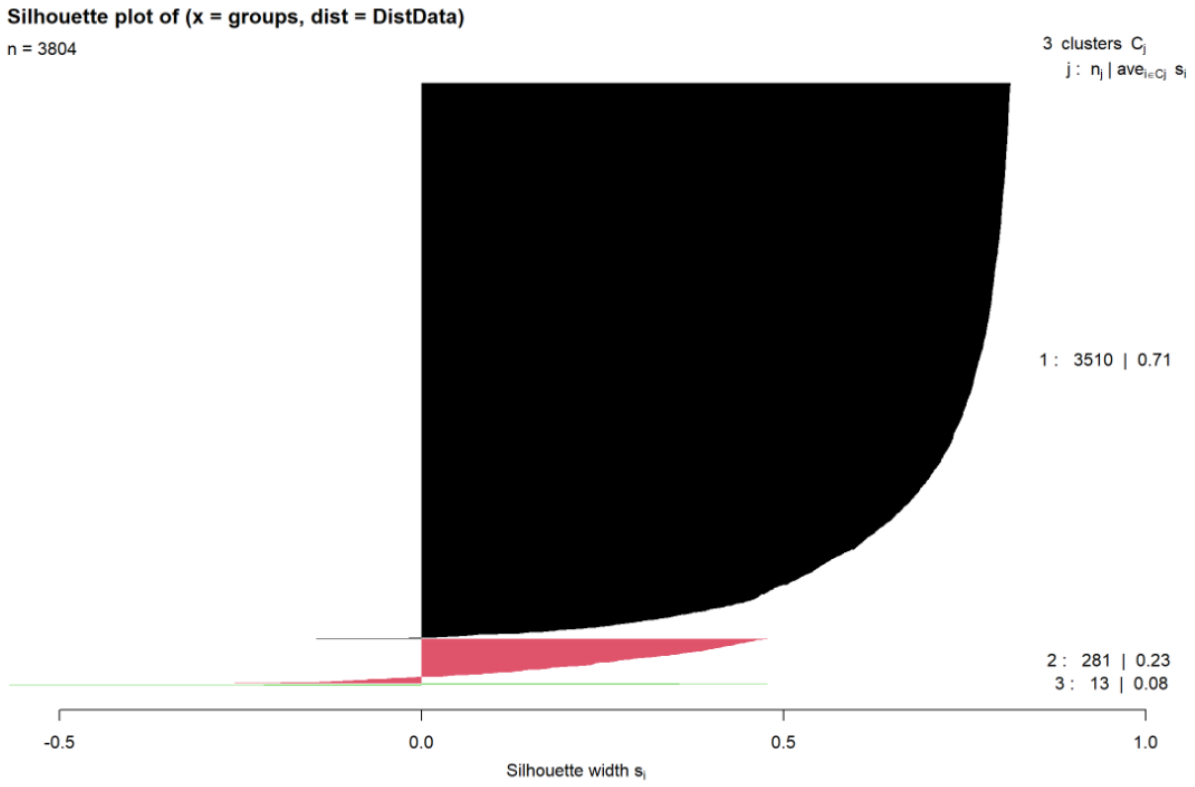


5. Determine la calidad del agrupamiento hecho por cada algoritmo con el método de la silueta. Discuta los resultados.

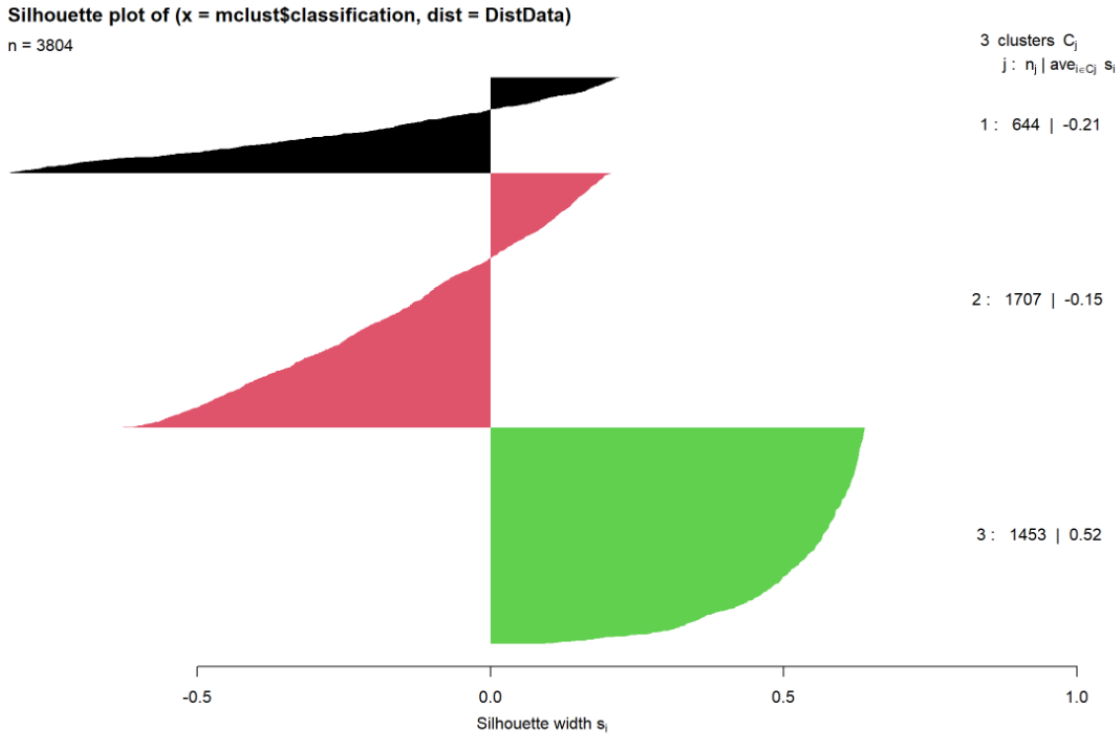


Average silhouette width : 0.46

El resultado es muy cercano a 1, el cual es un resultado deseable.



Average silhouette width : 0.67



Average silhouette width : 0.1

Podemos observar, que con este algoritmo obtuvimos 0.0998634 de promedio de la silueta, el valor es mas cercano a 0 que a 1, gracias a que posee demasiadas siluetas negativas.

El valor es mas cercano a 0 que a 1, gracias a que posee demasiadas siluetas negativas.

#####k-mean

Como se puede en la grafica realizada con el metodo de la silueta en el inciso anterior, la primera agrupacion se ve coherente, y en la segunda y tercera agrupacion se obtuvieron algunos valores atipicos, pero el coeficiente es bastante cercano a 1 lo que es adecuado.

#####Cluster jerarquico

Como se puede en la grafica realizada con el metodo de la silueta en el inciso anterior, se puede ver que los clusters involucrados son coherentes en la gran mayoria, solo con algunos datos atipicos.

#####Mezcla de gaussiano

Como se puede en la grafica realizada con el metodo de la silueta en el inciso anterior, esta grafica se obtuvo un mayor de datos atipicos, casi iguales a los datos 'normales', solo la terccera agrupacion es totalmete coherente y el coeficiente es mas cercano a 0.

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.

Medias

```{r Media popularity}
mean(x = NORMD$popularity, na.rm = TRUE)
```

```
[1] 68.1854
```

La media de la popularidad es 68.1854

```{r Media budget}
mean(x = NORMD$budget, na.rm = TRUE)
```

```
[1] 24335422
```

La media de presupuesto es 24335422

```{r Media revenue}
mean(x = NORMD$revenue, na.rm = TRUE)
```

```
[1] 78593938
```

```{r Media runtime}
mean(x = NORMD$runtime, na.rm = TRUE)
```

```
[1] 103.2592
```

La media de runtime es de 103.2592

```{r Media VoteCount}
round(mean(x = NORMD$voteCount, na.rm = TRUE))
```

```
[1] 1871
```

La media de votos es 1871

Modas

```{r Moda popularity}
tabla <- table(NORMD$popularity)
head(sort(tabla, decreasing = TRUE), n = 15)
```

```
15.804 39.372   8.54  9.336   9.34  9.363  9.608   9.83 10.171 10.192 10.243 10.303 10.393 10.472 10.767
     3      3      2      2      2      2      2      2      2      2      2      2      2      2      2
```

La moda de la popularidad es compartida entre 15.804 y 39.372 con 3 repeticiones cada una.

```{r Moda budget}
tabla <- table(NORMD$budget)
head(sort(tabla, decreasing = TRUE), n = 15)
```

```
       0  20000000  10000000  30000000  15000000  40000000  25000000  50000000   5000000  35000000  60000000
    1688        84        81        77        70        68        57        56        52        45        45
12000000 100000000 150000000   4000000
      42        34        33        32
```

La moda de presupuestos es de 20000000 con 84 repeticiones. El valor 0 ha sido ignorado debido a que se considera un NA o NULL.

```{r Moda revenue}
tabla <- table(NORMD$revenue)
head(sort(tabla, decreasing = TRUE), n = 15)
```

```
       0    7e+06     5000    2e+06  3600000  4100000  1.4e+07  1.9e+07 21200000  2.6e+07  2.7e+07    3e+07
    1558        3        2        2        2        2        2        2        2        2        2        2
34100000  4.3e+07 43300000
       2        2        2
```

La moda de los infresos es 7e+06 con 3 repeticiones. El valor 0 es ignorado debido a que se considera un NA o NULL.

```{r Moda runtime}
tabla <- table(NORMD$runtime)
head(sort(tabla, decreasing = TRUE), n = 15)
```

```
 90 100  97  93 102 110  92  95  94 107  96 108  98 109 104
143 115  98  93  90  88  86  86  84  84  83  83  82  82  80
```

La moda de runtime es 90 con 143 repeticiones.

```{r Moda VoteCount}
tabla <- table(NORMD$voteCount)
head(sort(tabla, decreasing = TRUE), n = 15)
```

```
  4  18 179   6  13  52  65   3  34  46  57  90 215   1   2
 15  13  12  11  11  11  11  10  10  10  10  10  10   9   9
```

La moda de votos por pelicula es de 4 con 15 repeticiones.