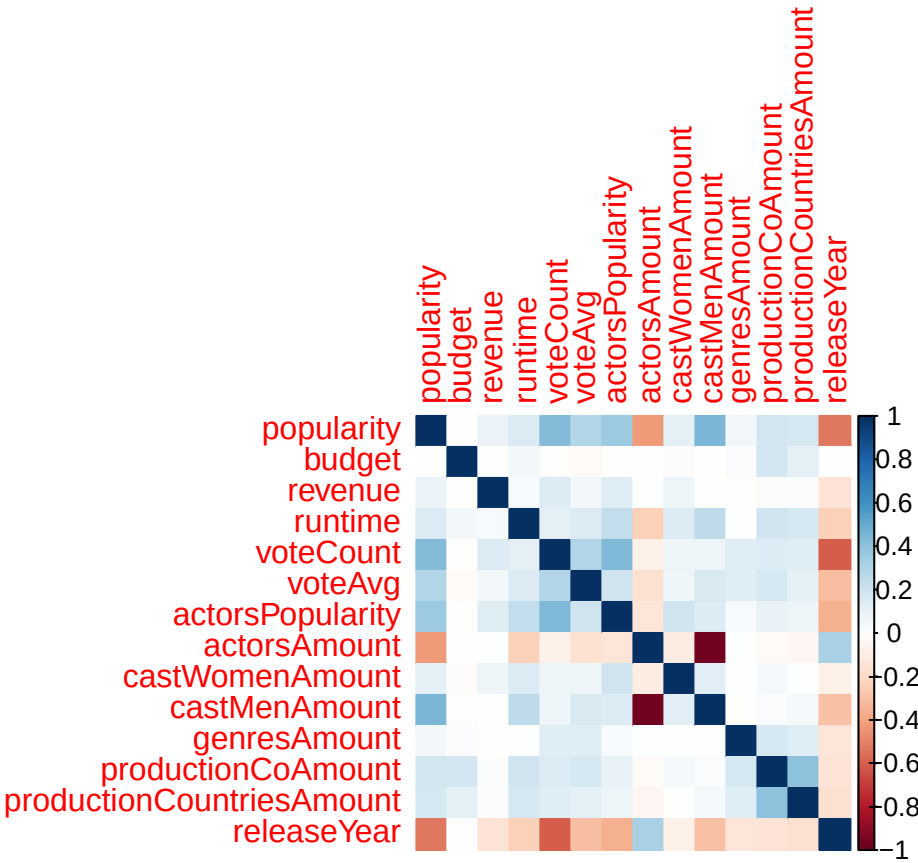
Laboratorio 2

Luis Pedro Lira 23669, Mario Rocha 23501, Juan Francisco Martínez 23617

2025-02-07

###Análisis de Componentes Principales (PCA)

###3.1 Selección de Variables y Preprocesamiento



###3.2 Matriz de Correlación

En la matriz de correlación se observan asociaciones importantes entre varias variables del conjunto de datos. Destaca una correlación positiva entre budget y revenue, lo cual es consistente con la lógica de la industria cinematográfica: producciones con mayor presupuesto tienden a generar mayores ingresos.

También se observa relación entre voteCount, popularity y actorsPopularity, lo que sugiere que películas con mayor visibilidad tienden a recibir más votos.

Asimismo, existe alta correlación entre actorsAmount, castMenAmount y castWomenAmount, lo cual es esperado debido a que representan componentes del tamaño total del elenco.

Estas correlaciones evidencian redundancia entre variables, lo que justifica la aplicación del Análisis de Componentes Principales para reducir la dimensionalidad del conjunto de datos.

###3.3 Medida de Adecuación Muestral (Índice KMO)

El índice KMO global obtenido fue de 0.69, lo cual indica una adecuación muestral aceptable para aplicar el Análisis de Componentes Principales.

Según la clasificación de Kaiser: • 0.90 → Excelente • 0.80 → Muy bueno • 0.70 → Bueno • 0.60 → Aceptable • <0.50 → Inadecuado

En este caso, el valor se encuentra cercano a 0.70, lo que sugiere que las correlaciones parciales entre variables no son excesivamente altas y que el análisis factorial es apropiado, aunque no óptimo.

###3.4 Prueba de Esfericidad de Bartlett

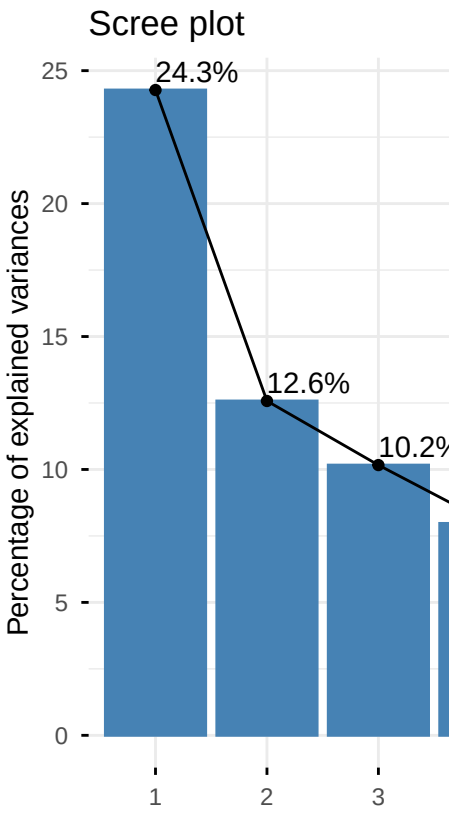
```
## $chisq
## [1] 5416.441
##
## $p.value
## [1] 0
##
## $df
## [1] 91
```

La prueba de esfericidad de Bartlett resultó altamente significativa ($\chi^2 = 5416.44$ p < 0.001), lo que permite rechazar la hipótesis nula de que la matriz de correlación es una matriz identidad.

Esto confirma que existe correlación suficiente entre las variables y que el Análisis de Componentes Principales es apropiado para este conjunto de datos.

###3.5 Aplicación del Análisis de Componentes Principales

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8434 1.3269 1.1929 1.05639 0.98343 0.95596 0.93911
## Proportion of Variance 0.2427 0.1258 0.1017 0.07971 0.06908 0.06528 0.06299
## Cumulative Proportion  0.2427 0.3685 0.4701 0.54983 0.61891 0.68419 0.74718
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.89506 0.87855 0.76897 0.75538 0.66624 0.57717 0.16590
## Proportion of Variance 0.05722 0.05513 0.04224 0.04076 0.03171 0.02379 0.00197
## Cumulative Proportion  0.80441 0.85954 0.90178 0.94253 0.97424 0.99803 1.00000
```



##3.6 Determinación del Número de Componentes (Scree Plot y Varianza Explicada)

```
##  [1] 3.39812748 1.76053146 1.42304682 1.11596354 0.96713359 0.91386270
##  [7] 0.88192312 0.80112853 0.77184452 0.59131760 0.57059723 0.44387318
## [13] 0.33312887 0.02752134
```

El primer componente principal explica el 24.3% de la variabilidad total, seguido por el segundo con 12.6% y el tercero con 10.2%.

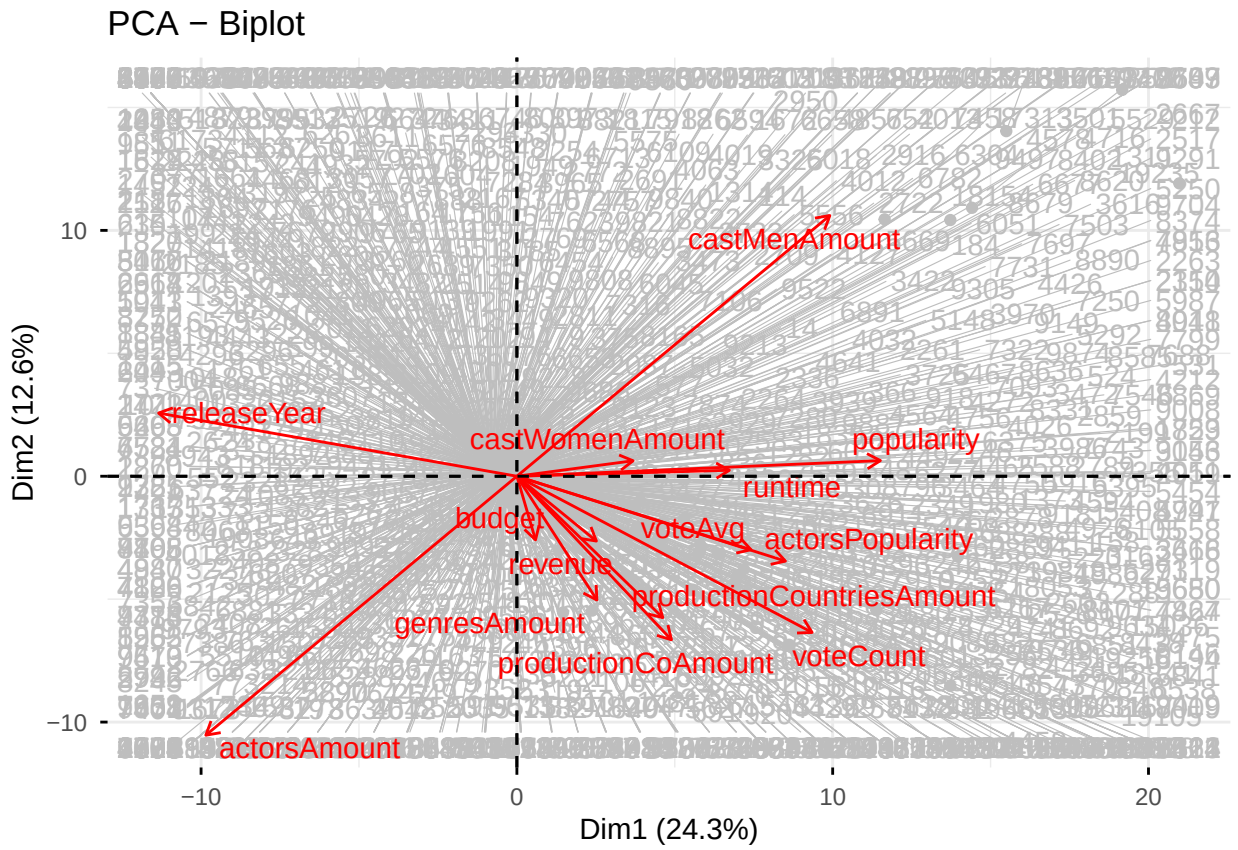
El gráfico de sedimentación muestra una disminución pronunciada después del tercer componente y una estabilización progresiva a partir del quinto componente.

Se decidió conservar los primeros 7 componentes, ya que explican aproximadamente el 74.7% de la variabilidad total, lo cual representa una reducción considerable de dimensionalidad manteniendo una proporción adecuada de información.

###3.7 Interpretación de los Componentes Principales

```
##                              PC1    PC2    PC3    PC4    PC5    PC6    PC7
## popularity                 0.410  0.031  0.076 -0.139  0.108  0.076  0.001
## budget                     0.021 -0.128 -0.384  0.184  0.542  0.052  0.671
## revenue                    0.090 -0.130  0.246  0.298  0.416 -0.747 -0.272
## runtime                    0.240  0.012 -0.207  0.410 -0.152  0.200 -0.231
## voteCount                  0.333 -0.314  0.318 -0.130  0.118  0.158  0.086
## voteAvg                    0.264 -0.148  0.049 -0.207 -0.221 -0.014 -0.029
## actorsPopularity           0.303 -0.171  0.282  0.272 -0.035  0.234  0.126
## actorsAmount              -0.351 -0.521  0.159  0.083 -0.045  0.124 -0.021
## castWomenAmount            0.131  0.031  0.088  0.600 -0.497 -0.180  0.315
## castMenAmount              0.353  0.525 -0.162 -0.060  0.031 -0.119  0.022
```

```
## genresAmount               0.090 -0.249 -0.150 -0.397 -0.404 -0.498  0.362
## productionCoAmount         0.174 -0.329 -0.482  0.059 -0.055 -0.021 -0.147
## productionCountriesAmount  0.165 -0.285 -0.461  0.023  0.016 -0.008 -0.380
## releaseYear               -0.404  0.127 -0.186  0.155 -0.137 -0.065 -0.013
##                              PC8    PC9   PC10   PC11   PC12   PC13   PC14
## popularity                 0.260 -0.130 -0.131  0.078  0.817 -0.101  0.049
## budget                    -0.160 -0.084  0.062 -0.131  0.044 -0.005  0.000
## revenue                   -0.125 -0.020 -0.062  0.015  0.024 -0.027  0.002
## runtime                   -0.557  0.436  0.201  0.051  0.225 -0.143  0.025
## voteCount                  0.110  0.076  0.237  0.086 -0.313 -0.669  0.008
## voteAvg                   -0.566 -0.660 -0.032 -0.233 -0.040  0.028 -0.004
## actorsPopularity           0.021  0.202 -0.703 -0.193 -0.184  0.216  0.003
## actorsAmount              -0.028 -0.014  0.046  0.007  0.249  0.034 -0.699
## castWomenAmount            0.310 -0.256  0.272 -0.012  0.014 -0.004  0.007
## castMenAmount              0.027  0.018 -0.083 -0.012 -0.174 -0.108 -0.710
## genresAmount              -0.083  0.425 -0.090 -0.056  0.092 -0.021  0.006
## productionCoAmount         0.105 -0.194 -0.227  0.685 -0.170  0.050 -0.007
## productionCountriesAmount  0.362  0.009  0.091 -0.624 -0.067 -0.012 -0.004
## releaseYear               -0.044 -0.139 -0.482 -0.112  0.104 -0.677  0.057
```



El primer componente parece representar una dimensión de impacto y visibilidad comercial, ya que agrupa variables relacionadas con popularidad, número de votos y popularidad del elenco.

La carga negativa en releaseYear sugiere que películas más antiguas tienden a tener mayor acumulación de popularidad y votos, lo cual puede deberse a mayor tiempo en el mercado.

###3.8 Conclusiones del PCA

El Análisis de Componentes Principales permitió reducir la dimensionalidad del conjunto de datos original compuesto por 14 variables numéricas, sintetizando la información en un menor número de componentes sin perder una proporción significativa de variabilidad.

Los primeros 7 componentes explican aproximadamente el 74.7% de la variabilidad total, lo cual representa una reducción considerable de la complejidad del dataset manteniendo la mayor parte de la información relevante.

El primer componente principal se asocia principalmente con variables relacionadas con popularidad, número de votos y popularidad del elenco, por lo que puede interpretarse como una dimensión de impacto y desempeño comercial.

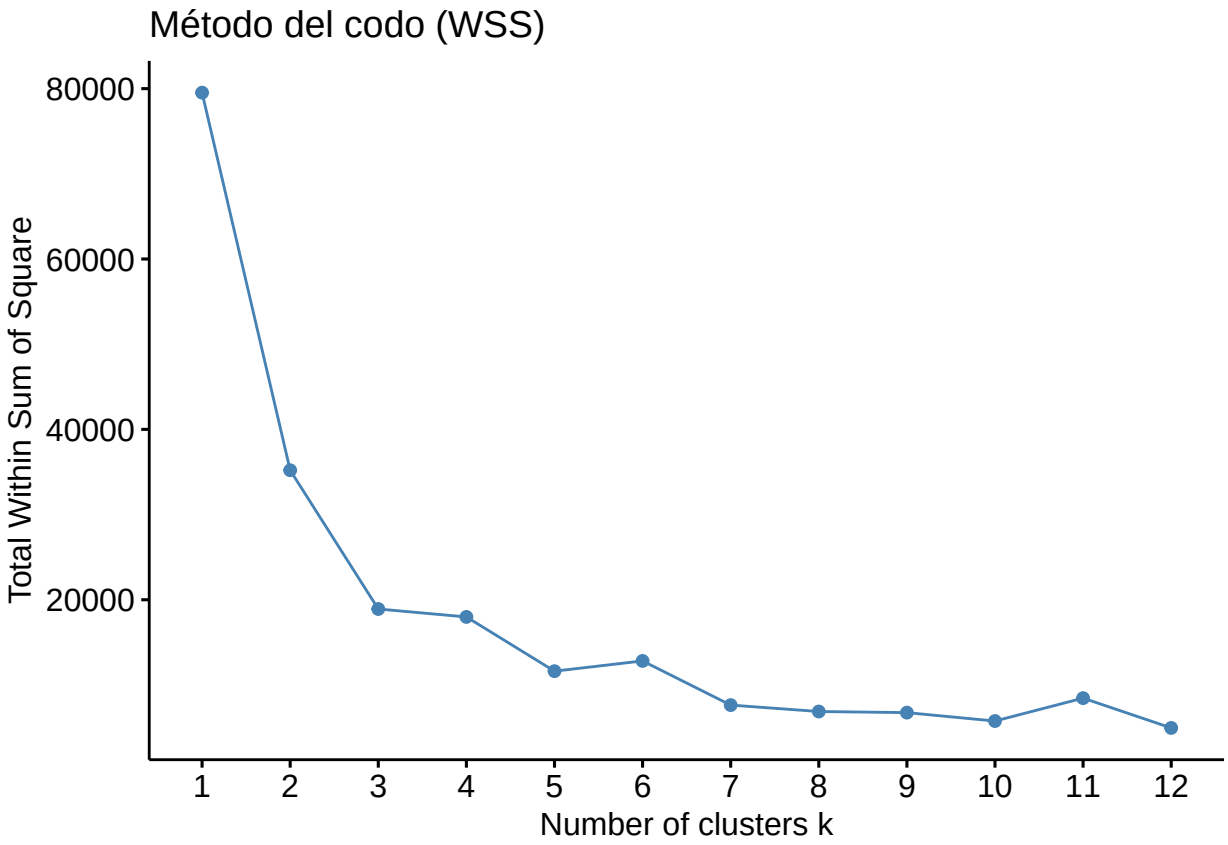
El segundo componente está vinculado con la composición del elenco, diferenciando películas según la cantidad y distribución de actores.

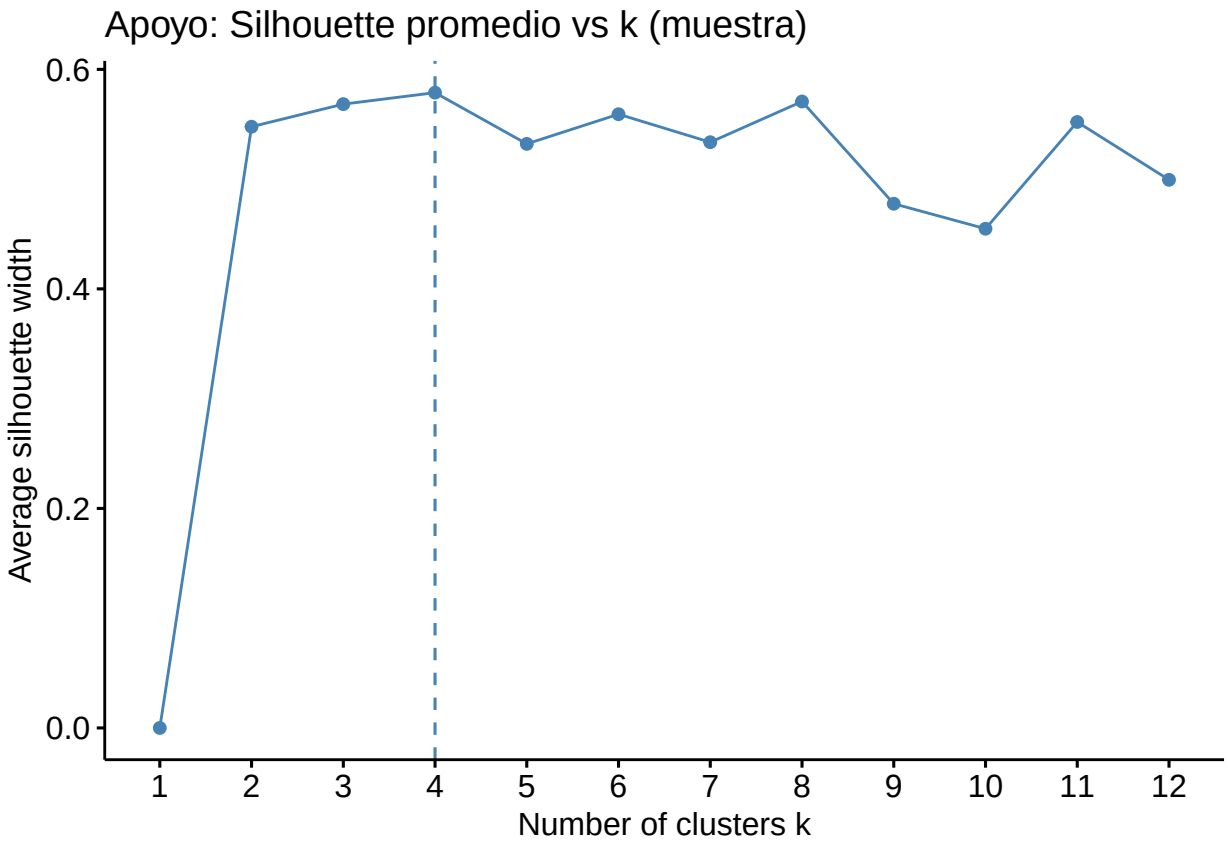
El tercer componente refleja características estructurales de producción, como el número de compañías productoras y países involucrados, lo cual sugiere distintos niveles de complejidad en las producciones cinematográficas.

En conjunto, el PCA revela que las películas pueden caracterizarse principalmente a través de tres grandes dimensiones: impacto comercial, estructura del elenco y complejidad de producción.

Estos resultados son útiles para CineVision Studios, ya que permiten comprender cuáles factores explican mayor variabilidad en el mercado cinematográfico y pueden servir como base para futuros modelos predictivos o segmentaciones estratégicas.

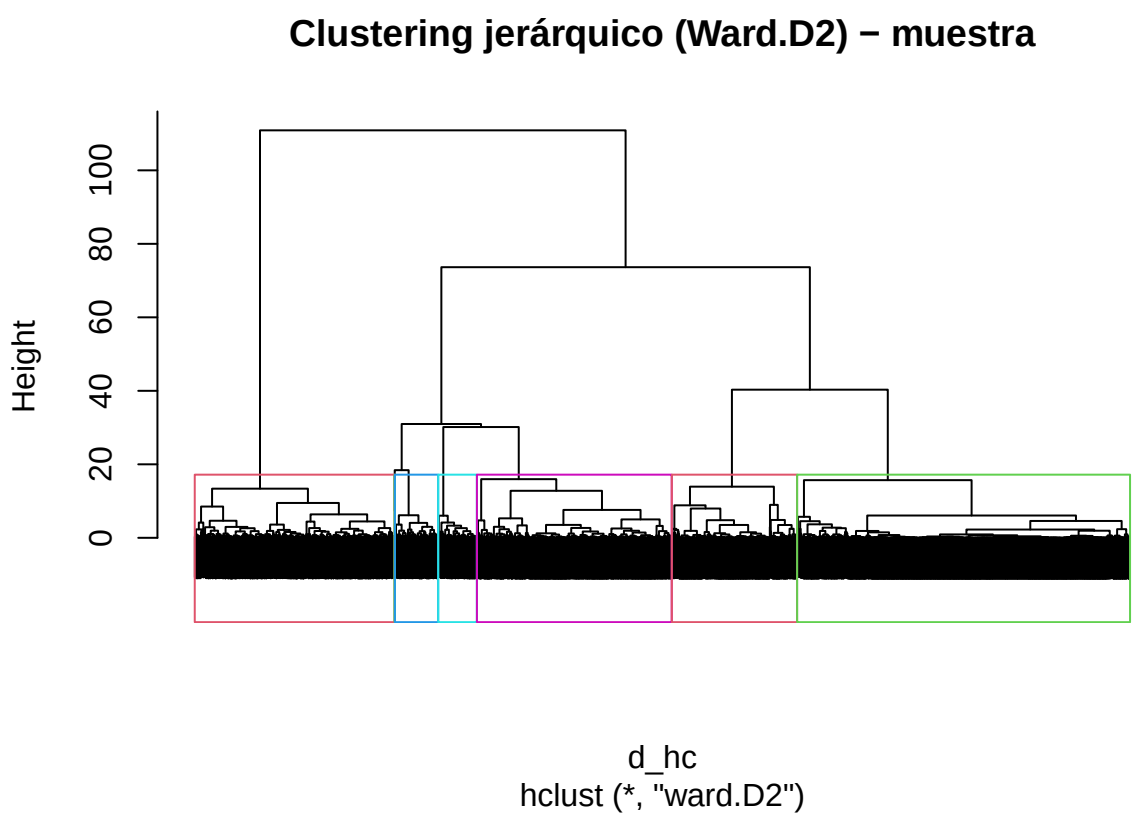
```
## Numéricas usadas para clustering:

## [1] "popularity" "budget"     "revenue"    "runtime"

##
## Numéricas NO encontradas (si esperabas verlas, revisa nombres):

##  [1] "vote_count"                  "vote_avg"
##  [3] "actors_popularity"           "genres_amount"
##  [5] "production_co_amount"        "production_countries_amount"
##  [7] "actors_amount"               "cast_women_amount"
##  [9] "cast_men_amount"             "release_year"

## Dimensiones de X (filas, columnas):

## [1] 19883     4

## $H
## [1] 0.05505128
```


Método del codo (WSS)
Total Within Sum of Square
80000
60000
40000
20000
1
2
3
4
5
6
7
8
9
10
11
12
Number of clusters k




```
## # A tibble: 10 x 2
##        k avg_silhouette
##    <int>          <dbl>
##  1     7          0.635
##  2     6          0.631
##  3     5          0.590
##  4     8          0.568
##  5     4          0.566
##  6     3          0.561
##  7     9          0.556
##  8    11          0.541
##  9    12          0.539
## 10    10          0.537
```

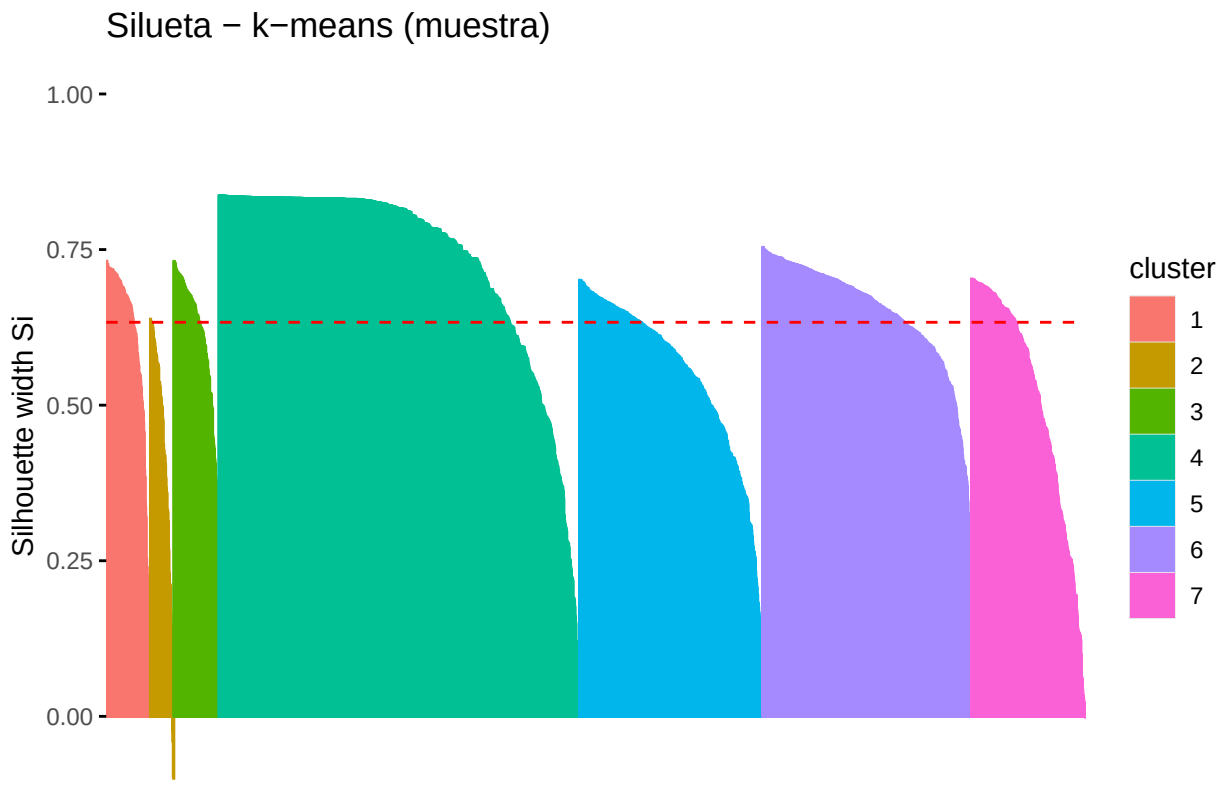
```
## k elegido (por silhouette en muestra): 7
```



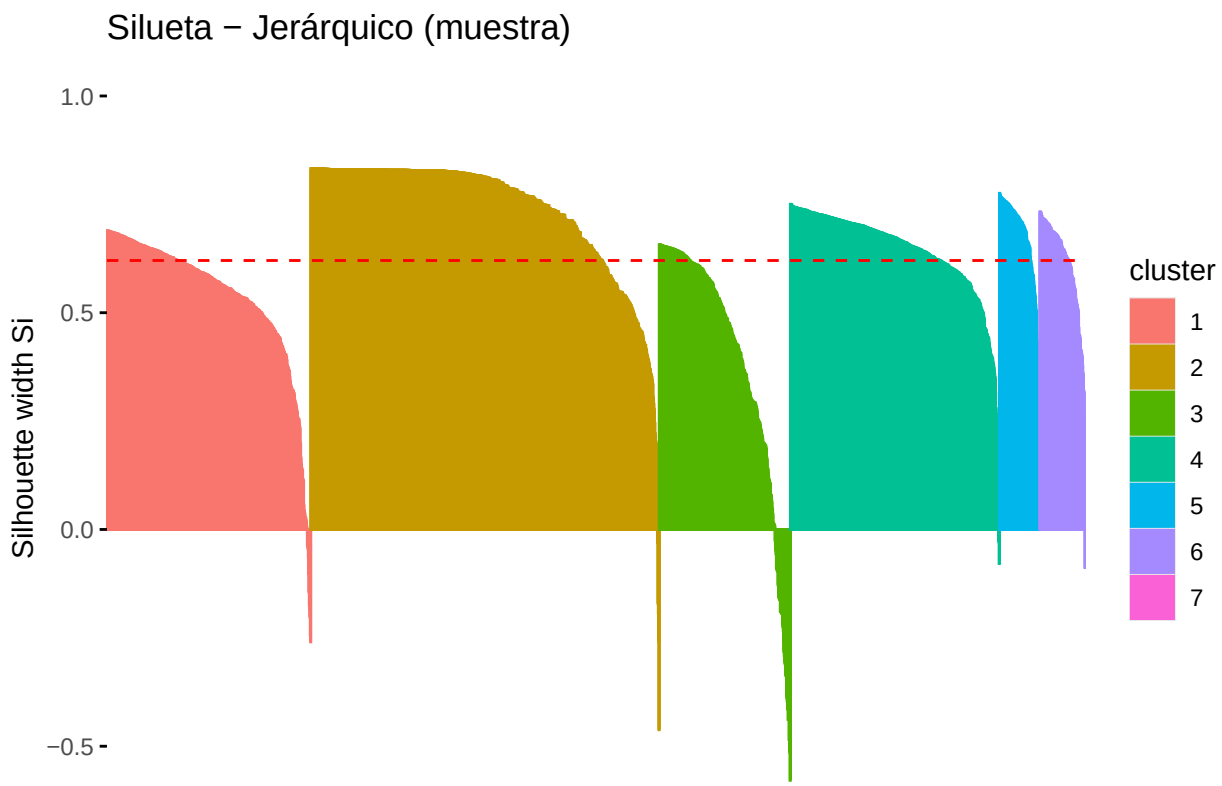

```
## Silhouette promedio (k-means, muestra): 0.6331
```

```
## Silhouette promedio (jerárquico, muestra): 0.6204
```

```
##   cluster size ave.sil.width
## 1       1  135          0.59
## 2       2   72          0.42
## 3       3  138          0.61
## 4       4 1105          0.73
## 5       5  561          0.55
## 6       6  640          0.65
## 7       7  349          0.50
```



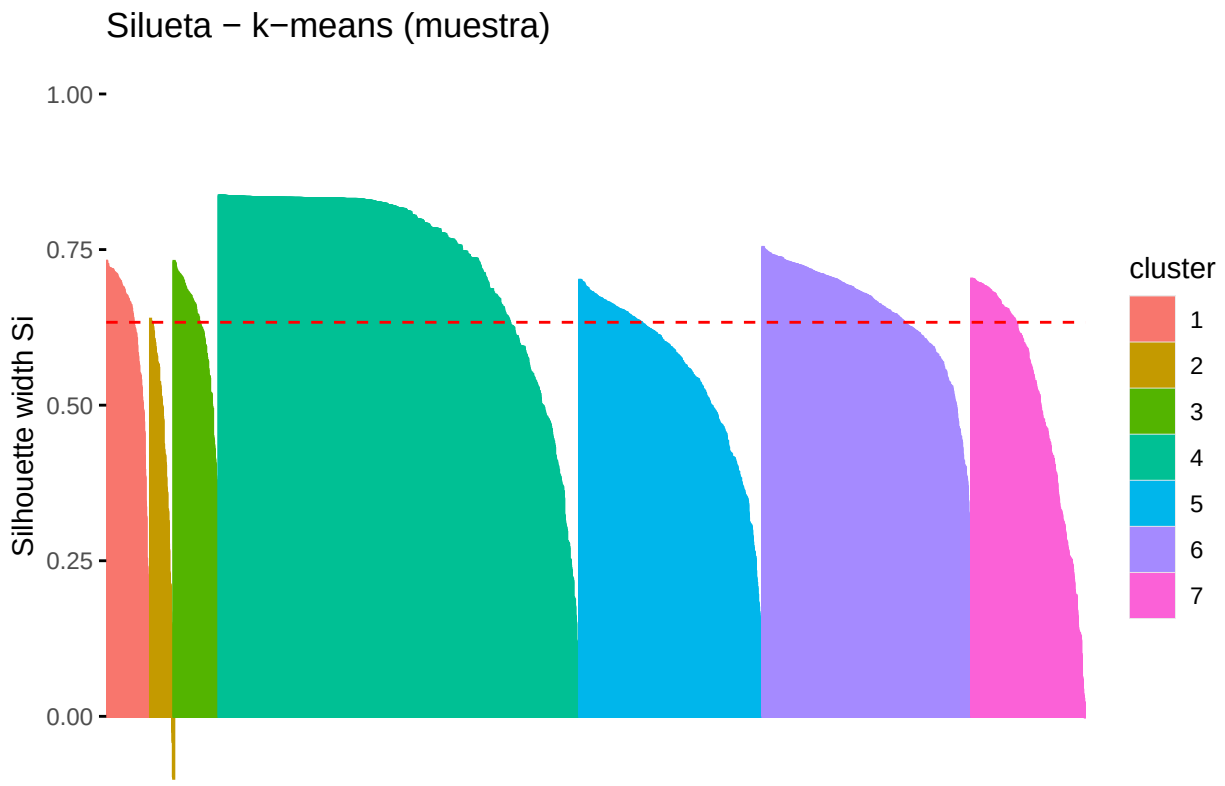

```
##   cluster size ave.sil.width
## 1       1  625          0.54
## 2       2 1068          0.74
## 3       3  402          0.39
## 4       4  641          0.64
## 5       5  124          0.67
## 6       6  139          0.61
## 7       7    1          0.00
```



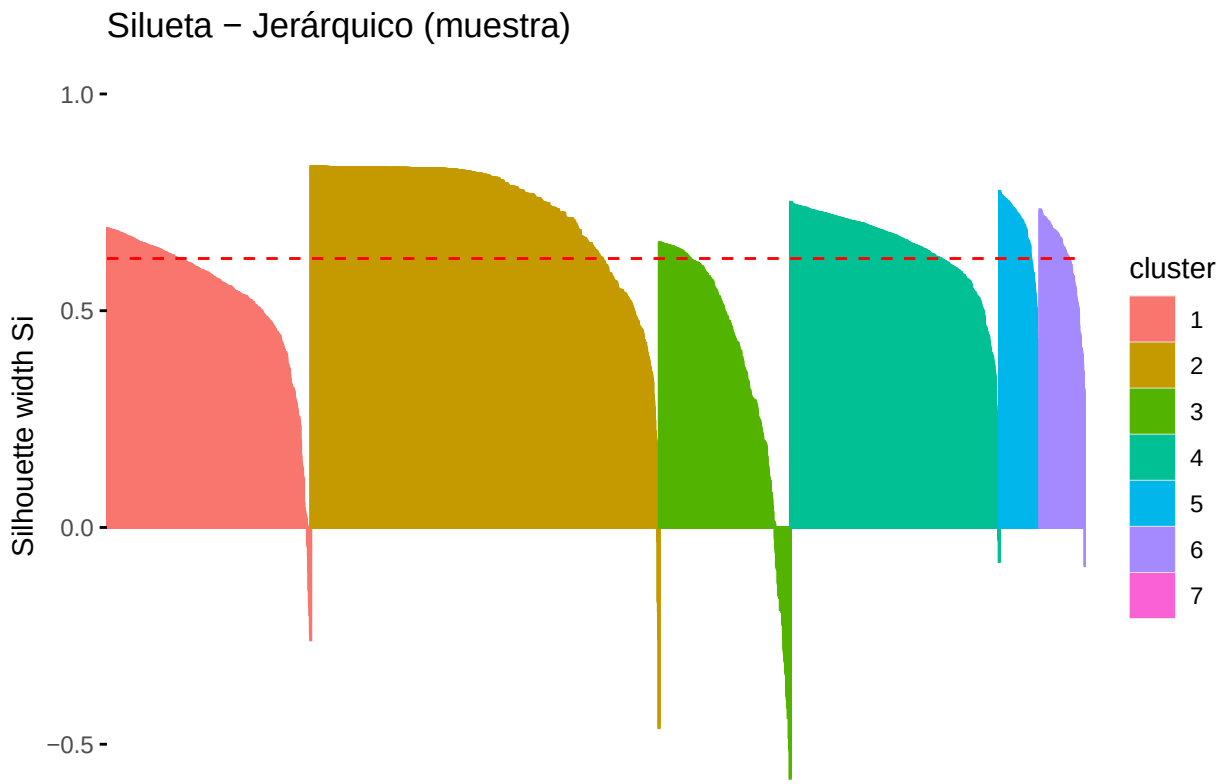

```
## Silhouette promedio (k-means, muestra): 0.6331
```

```
## Silhouette promedio (jerárquico, muestra): 0.6204
```

```
##   cluster size ave.sil.width
## 1       1  135          0.59
## 2       2   72          0.42
## 3       3  138          0.61
## 4       4 1105          0.73
## 5       5  561          0.55
## 6       6  640          0.65
## 7       7  349          0.50
```




```
##   cluster size ave.sil.width
## 1       1  625          0.54
## 2       2 1068          0.74
## 3       3  402          0.39
## 4       4  641          0.64
## 5       5  124          0.67
## 6       6  139          0.61
## 7       7    1          0.00
```


Silueta − Jerárquico (muestra)
Silhouette width Si
1.0
0.5
0.0
−0.5
cluster
1
2
3
4
5
6
7