

Proyecto1

Gerardo Pineda 22880, Daniel Rayo 22933

2025-02-10

Clustering

Procesamiento del dataset

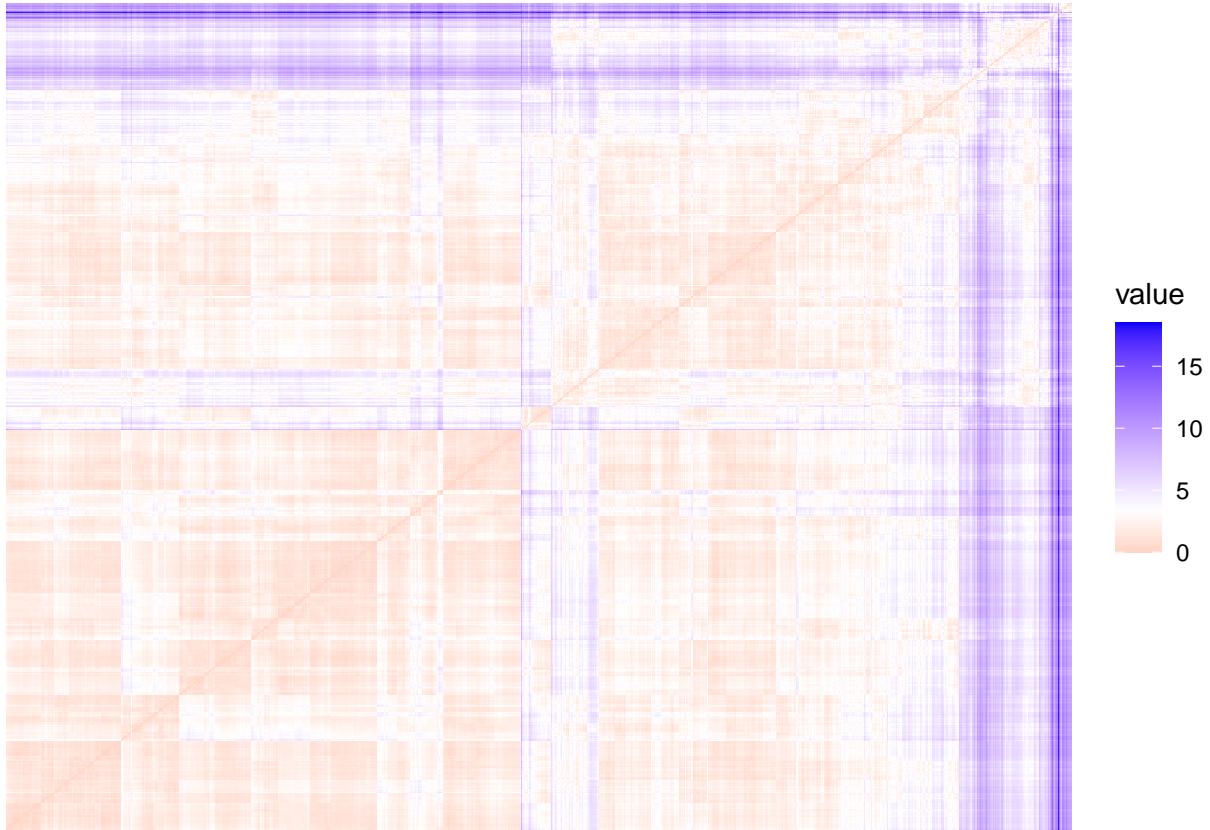
Para este apartado se decidió calcular los grupos en base a todas las variables numéricas, quitando todas las variables que no son de este tipo. Para posteriormente poder clasificar o ver la forma en la que se relacionan las variables por medio de estas que fueron descartadas en un inicio.

Las variables con las que se trabajaron los clusters son:

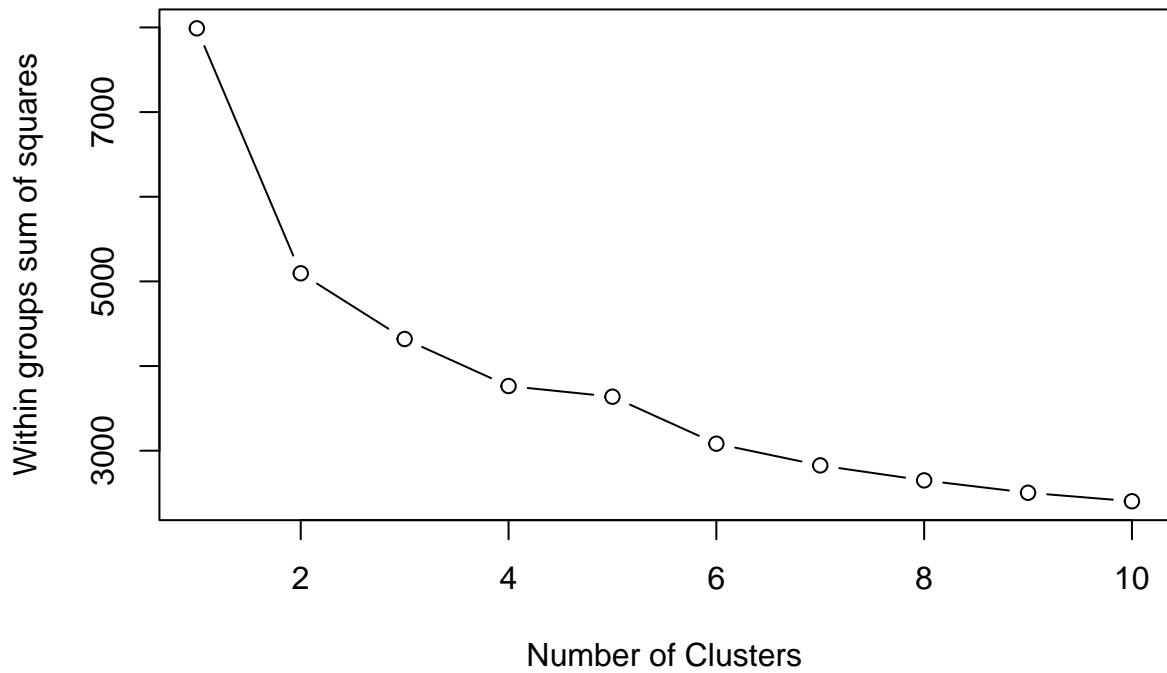
* *popularity*: Importante para agrupar películas por impacto mediático. * *Budget*: Ayuda a diferenciar producciones de alto y bajo costo. * *Revenue*: Relacionado con el éxito comercial. * *runtime*: Útil para agrupar películas cortas vs. largas. * *genresAmount*: Más géneros pueden significar una audiencia más diversa. * *voteCount*: Relacionado con la cantidad de personas que la vieron. * *voteAvg*: Permite separar películas mejor o peor valoradas. * *castMenAmount* y *castWomenAmount*: Para analizar el tamaño del reparto y su diversidad.

Primero necesitaremos verificar si vale la pena agrupar los datos. Usando el estadístico de Hopkins nos dio un resultado de 1, lo que indica que los resultados no son aleatorios y hay una alta posibilidad de que sea factible el agrupamiento.

Posteriormente se realizó un mapa de calor para verificar si realmente existen patrones.



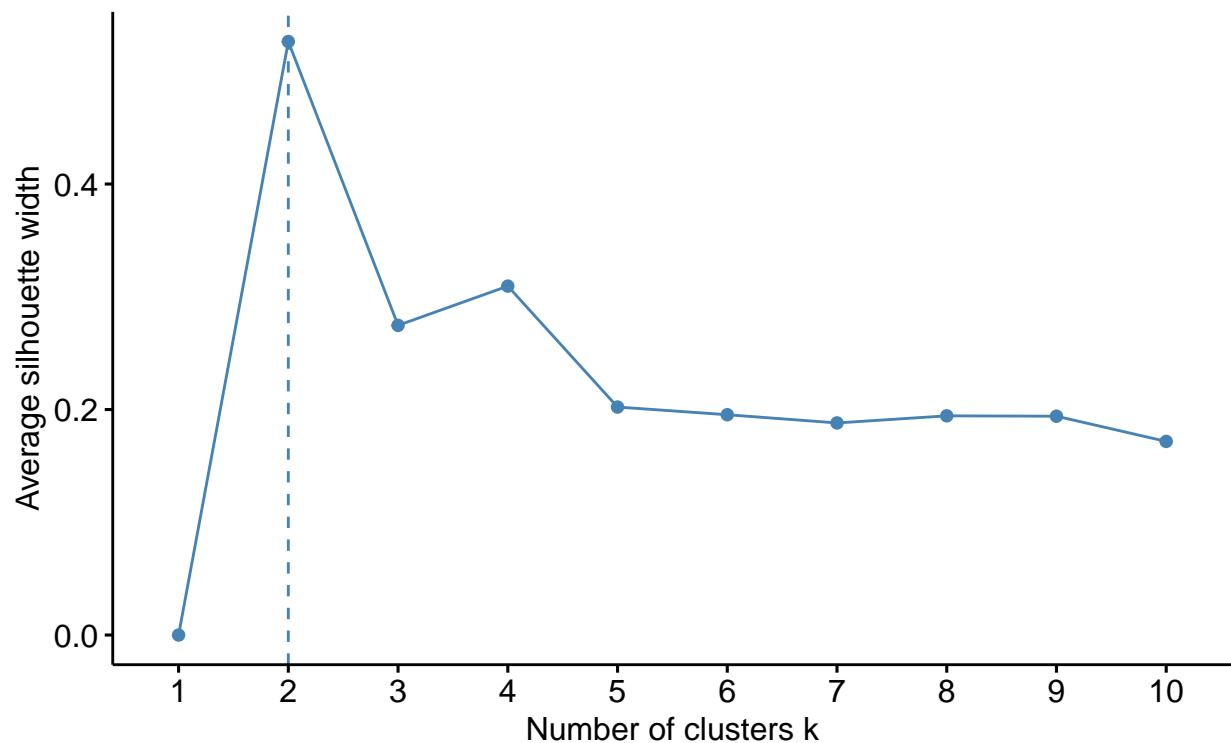
El mapa de calor revela que existen patrones claros en los datos, con al menos tres grupos de observaciones con similitudes internas. Esto refuerza la idea de que el dataset puede dividirse en tres clusters. La presencia de áreas con colores más intensos sugiere que algunas observaciones son más similares entre sí en ciertas características, mientras que otras tienen mayor variabilidad. El dataSet puede ser bueno para representar diferentes tipos de películas en términos de su presupuesto, popularidad y desempeño en taquilla.



El método del codo sugiere que un número óptimo de clusters es $k = 3$, ya que en este punto la reducción en la varianza dentro de los clusters deja de ser significativa.

Optimal number of clusters

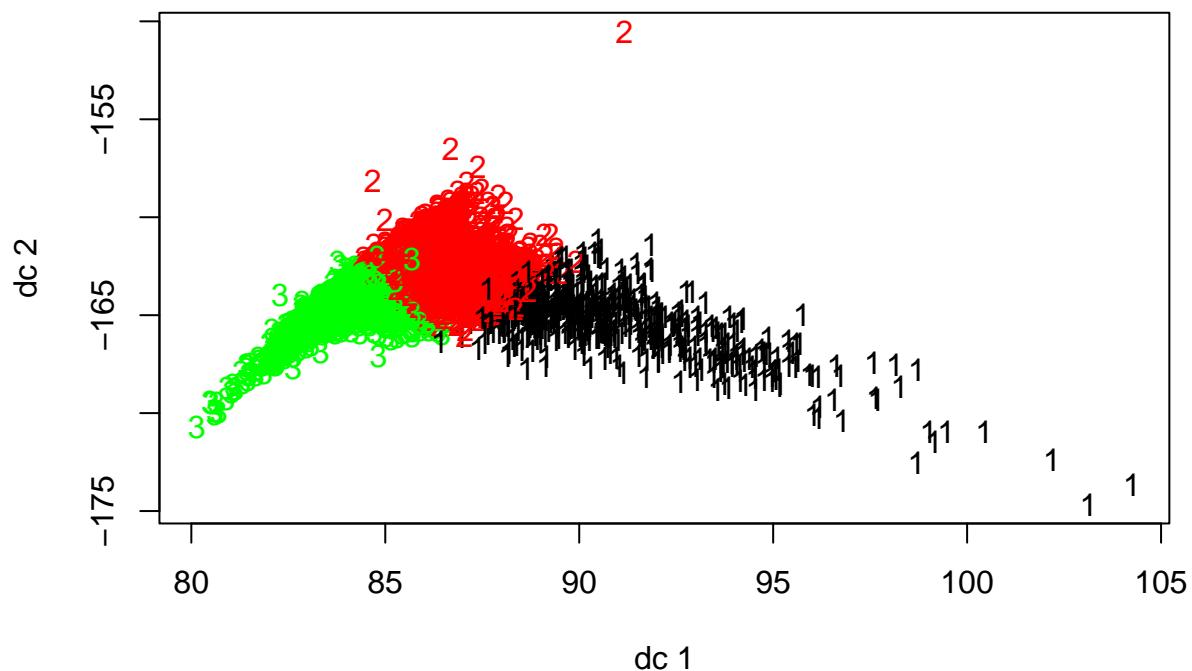
Silhouette method



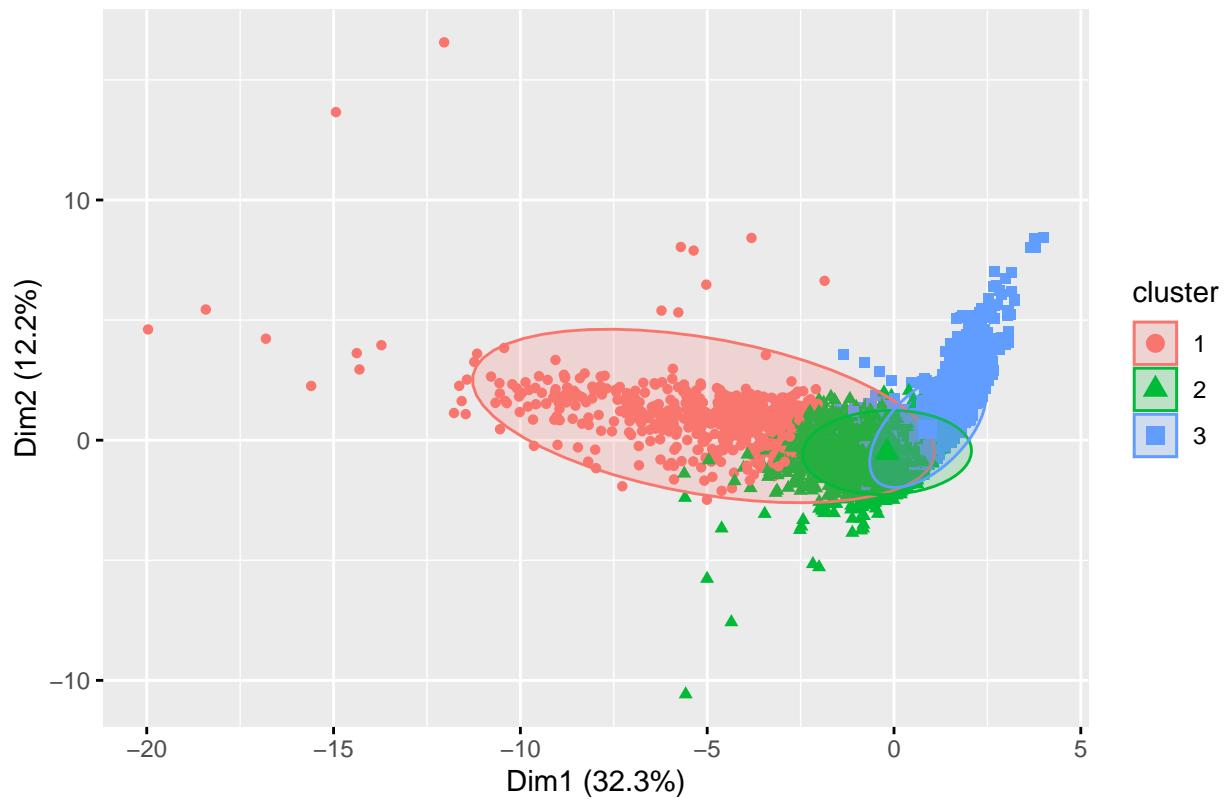
K means

Antes de empezar con el k mean comprobamos cual k era el mas adecuado para este clustering

$k = 3$



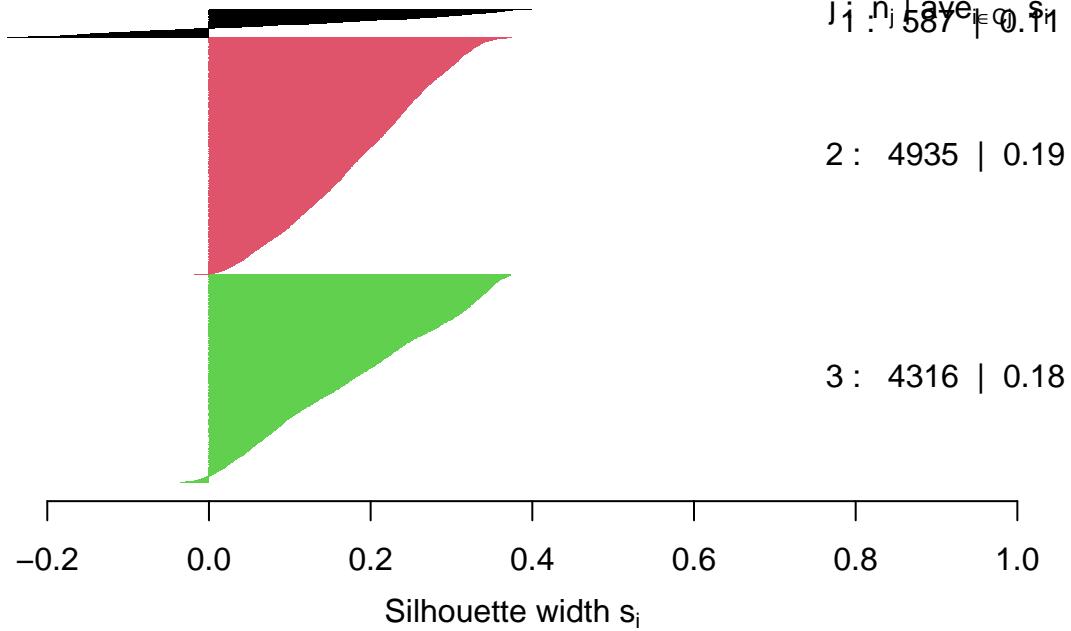
Cluster plot



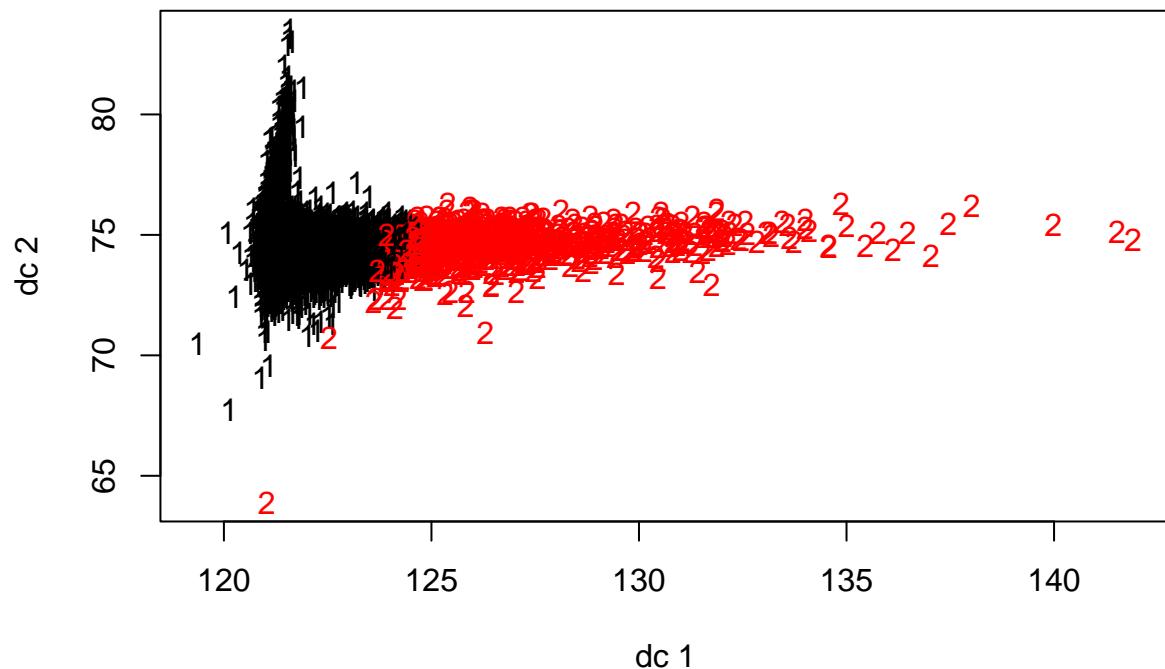
Silhouette plot of (x = km\$cluster, dist = dist(dataWithoutGrou

n = 9838

3 clusters C_j
 $j \in [1, n]$. Average silhouette width : 0.18



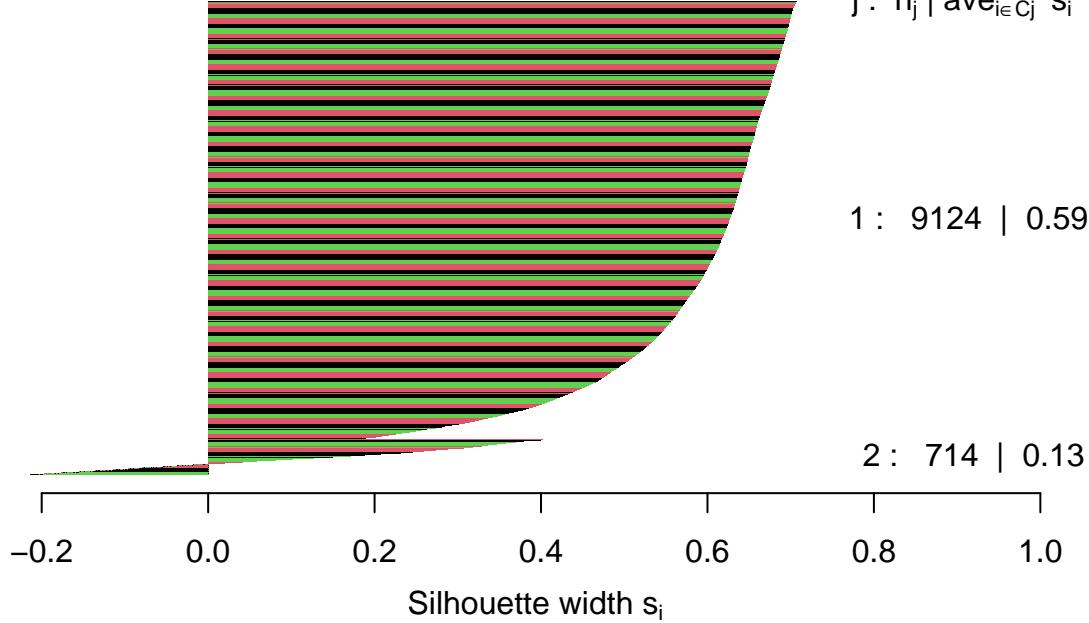
$K = 2$



Silhouette plot of (x = km\$cluster, dist = dist(dataWithoutGrou

n = 9838

2 clusters C_j
 $j : n_j | \text{ave}_{i \in C_j} s_i$

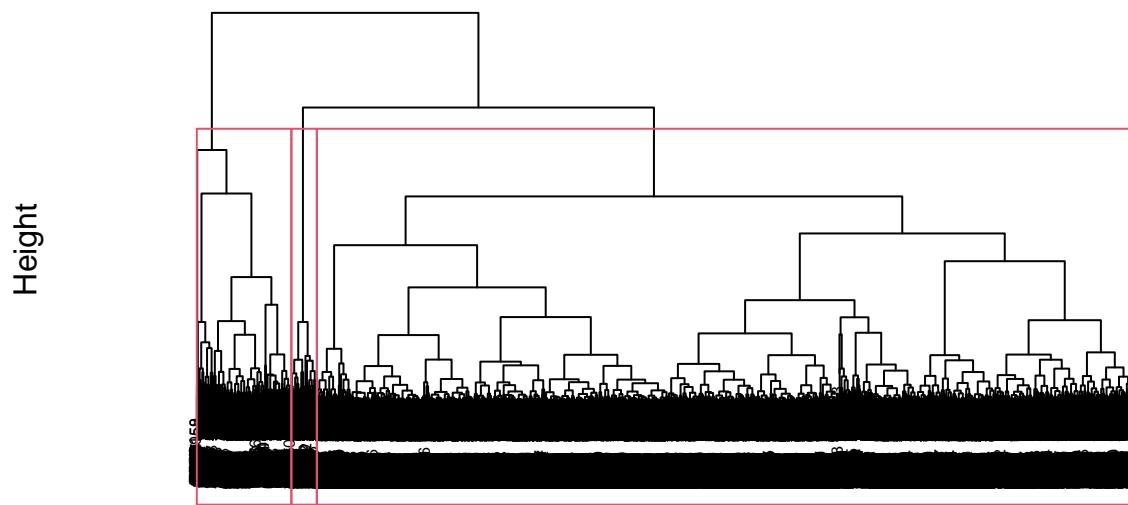


Average silhouette width : 0.55

Aunque el método del codo sugirió que k=3 podría ser un punto adecuado, al calcular el índice de silueta encontramos que para k=2 el valor es 0.55, significativamente mayor que el 0.18 obtenido con k=3. Esto indica que los clusters son más compactos y están mejor separados cuando se utilizan dos grupos en lugar de tres. Por esta razón, elegimos k=2, ya que proporciona una segmentación más clara y efectiva de los datos.

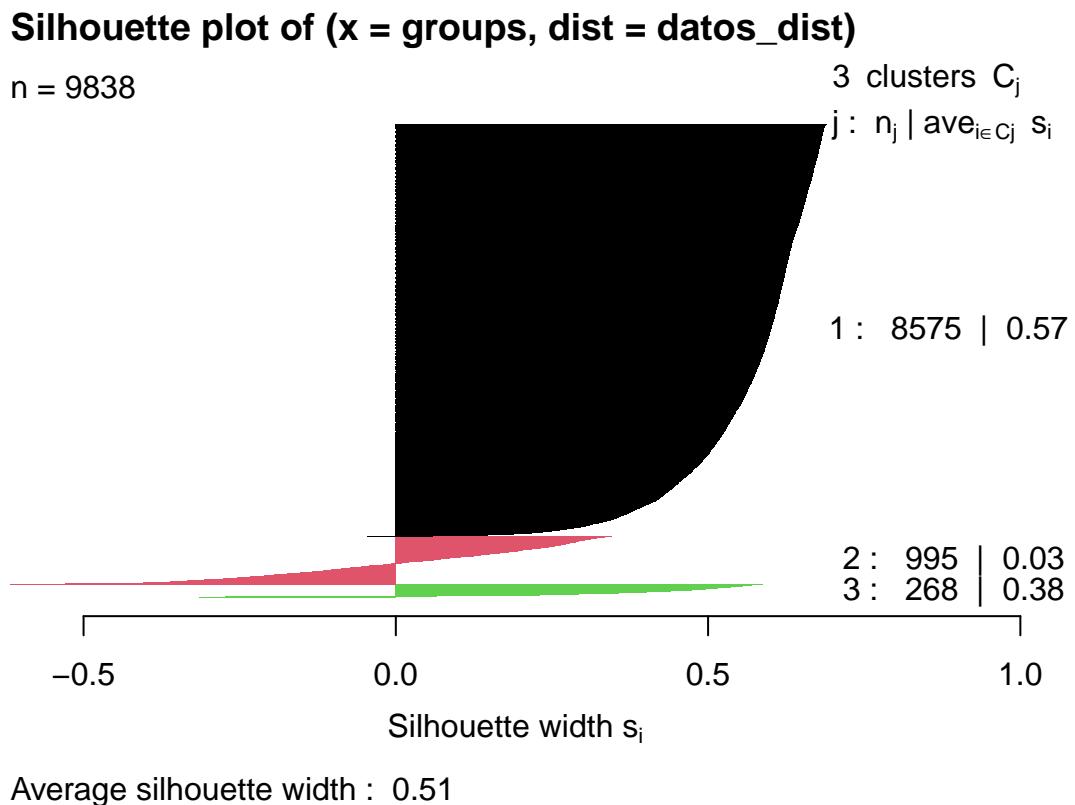
Cluster Jerarquico

Cluster Dendrogram



```
datos_dist  
hclust (*, "ward.D2")
```

Aunque con $k=2$ la silueta fue mejor en K-means, al analizar el dendrograma del clustering jerárquico, se observó que una división natural ocurre en tres grupos. Esto sugiere que, en lugar de forzar una separación en dos clusters, el modelo jerárquico es capaz de detectar subestructuras en los datos que K-means no capturó.



Interpretacion de los clusterings

Comparación de resultados:

Algoritmo	Número.de.clusters	Índice.de.silueta	Descripción
K-means	2	0.55	Mayor cohesión y separación clara
Jerárquico	3	0.51	El dendrograma sugiere 3 subgrupos bien diferenciados

Resumen de las variables con el metodo de k means

popularity_mean	budget_mean	revenue_mean	runtime_mean
48.78458	15966730	44958998	100.0764
57.73077	55790081	220130395	110.5000

Resumen de las variables con el metodo de cluster jerarquico

popularity_mean	budget_mean	revenue_mean	runtime_mean
47.50223	15864374	44216888	99.90915
58.84922	47740370	184325690	110.31055
76.28257	7372717	17966894	95.20149

Dado que cada algoritmo mostró un mejor desempeño con un número distinto de clusters, analizamos cómo se comportan los grupos generados en términos de las variables clave.

En general, los resultados de ambos métodos presentan distribuciones similares en algunas variables. Sin embargo, se observan diferencias significativas en presupuesto (budget) e ingresos (revenue). En el caso de k-means ($k=2$), los grupos reflejan una clara segmentación entre películas de mayor y menor presupuesto, con una diferencia notable en los ingresos generados.

Por otro lado, el clustering jerárquico ($k=3$) permite una segmentación más detallada, diferenciando un tercer grupo con películas de menor presupuesto y menor recaudación. Esta estructura ayuda a capturar mejor la diversidad de los datos, distribuyendo de manera más homogénea las películas en función de sus recursos financieros.

Comparando un poco más los dos métodos:

En cuanto al presupuesto e ingresos, el método jerárquico logra una mejor diferenciación, incorporando un grupo adicional de películas con un presupuesto bajo (7.3M) y una recaudación igualmente baja (17.9M), lo que no ocurre en k-means, donde la segmentación es más binaria.

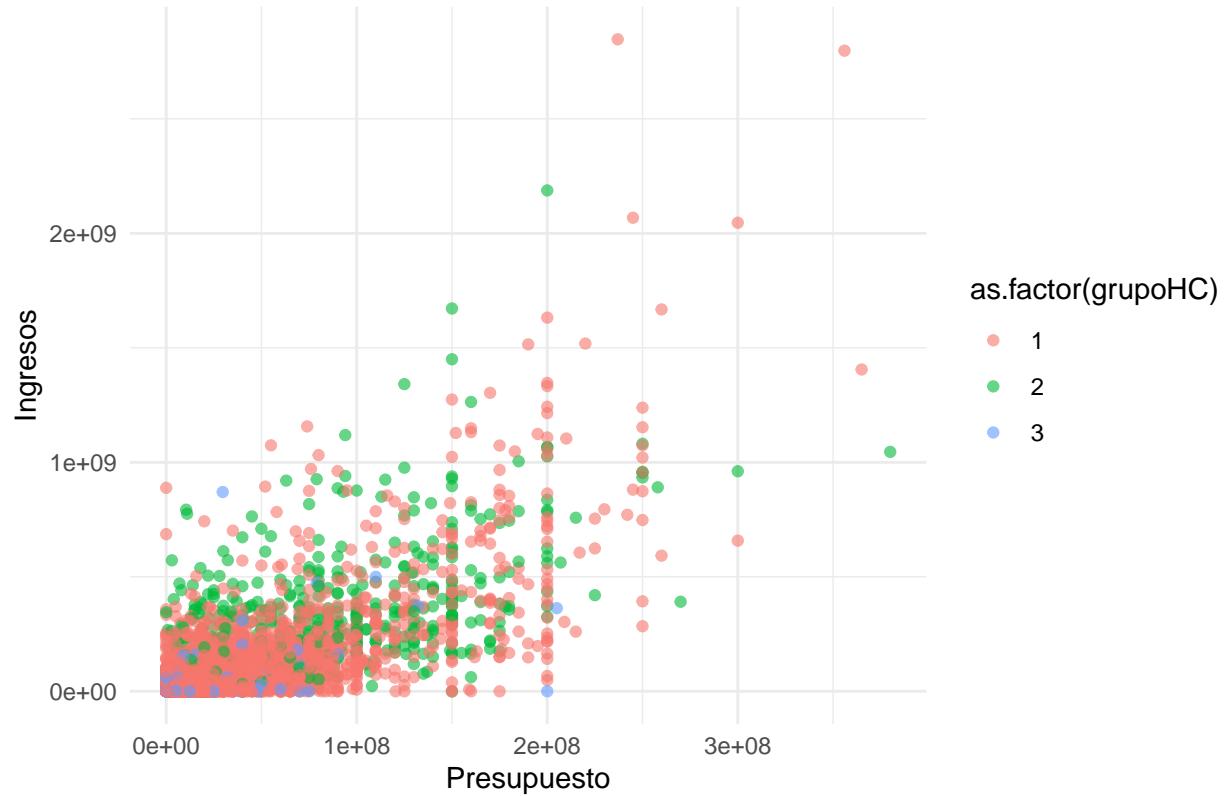
Además, la Duración de la película no parece tener diferencias significativas entre métodos, con un promedio cercano a los 100-110 minutos en la mayoría de los grupos.

También, en cuanto a la popularidad no se encuentran diferencias significativas.

elección de un metodo

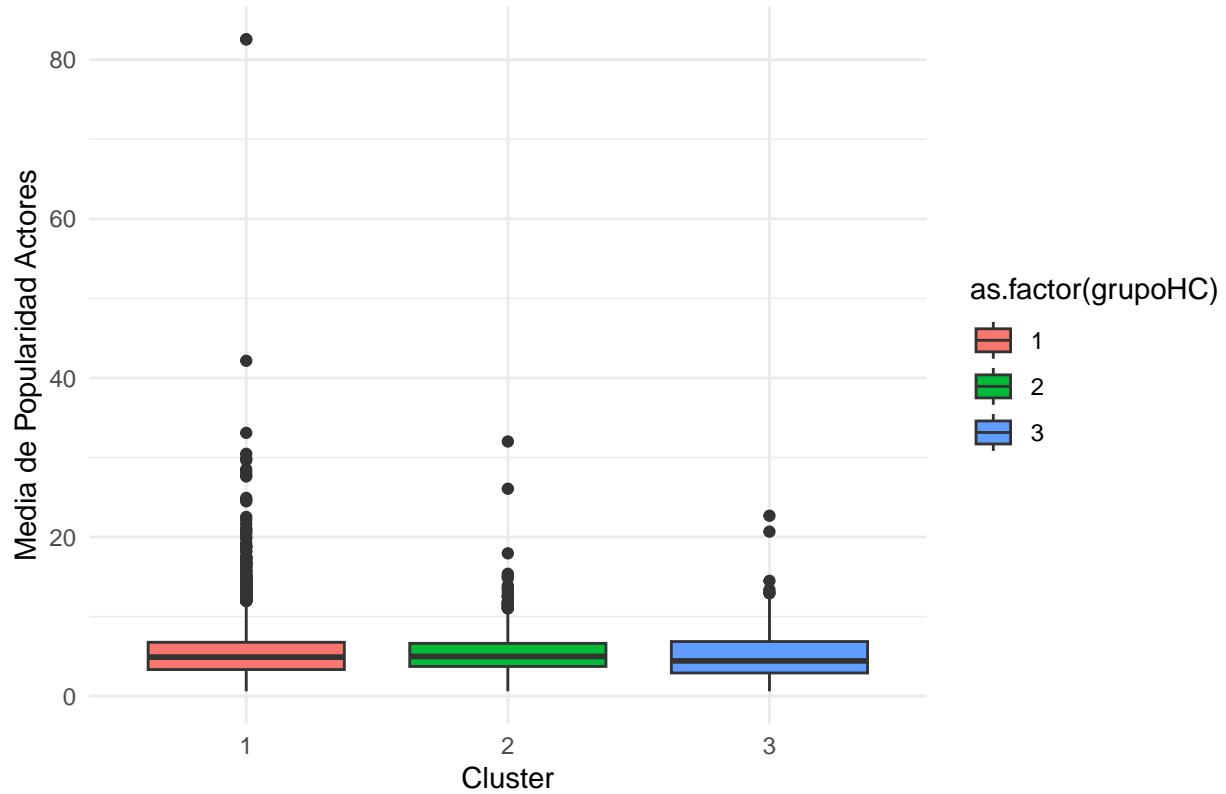
A pesar de que K-means con $k=2$ obtuvo un índice de silueta superior (0.55 frente a 0.51 del clustering jerárquico con $k=3$), se optó por utilizar el clustering jerárquico debido a su capacidad para capturar mayor variabilidad en los datos. El clustering jerárquico con $k=3$ permitió una clasificación más detallada, identificando un grupo intermedio que representa películas de presupuesto medio. Esta segmentación más rica facilita un análisis más preciso de las tendencias dentro de la industria cinematográfica, asegurando que las diferencias dentro de los datos sean mejor representadas. Por lo tanto, se eligió el clustering jerárquico para el análisis de tendencias centrales y frecuencias, ya que proporciona una interpretación más equilibrada sin perder información valiosa.

Relación entre Presupuesto e Ingresos por grupo



En la grafica se puede observar las películas en función de su presupuesto e ingresos. El grupo 1 tiene la mayor diversidad, mientras que los grupos 2 y 3 presentan características más específicas en términos de inversión y rentabilidad. Esta segmentación es útil para entender la relación entre el presupuesto y el éxito financiero.

Popularidad Promedio del Elenco por Cluster



Dado que la variable popularidad promedio del elenco no fue incluida en la formación de los clusters, es esperable que no haya una diferenciación clara entre los grupos con respecto a esta variable. Esto sugiere que la popularidad del elenco no estuvo directamente relacionada con los criterios de agrupamiento utilizados en el modelo.

Generos	Grupo 1	Grupo 2	Grupo 3
	38	4	5
Action	2220	329	64
Adventure	1467	266	33
Animation	1234	152	43
Comedy	2664	254	58
Crime	1060	125	28
Documentary	235	17	20
Drama	3282	386	87
Family	1197	163	32
Fantasy	1063	174	29
History	342	49	8
Horror	1356	99	38
Music	271	24	13
Mystery	701	78	15
Romance	1248	132	34
Science Fiction	1069	157	31
TV Movie	240	20	6
Thriller	2209	268	61
War	224	40	10
Western	109	6	3

Generos	Grupo 1	Grupo 2	Grupo 3
---------	---------	---------	---------

A partir de esta tabla podemos obtener que el Grupo 1 contiene la mayor diversidad y volumen de películas, lo que sugiere que es el cluster que abarca las producciones más comerciales y ampliamente distribuidas. El Grupo 2 es intermedio, con una presencia menor en todos los géneros, posiblemente representando películas con menor presupuesto o enfoque más específico. El Grupo 3 es el más reducido y especializado, con películas menos comerciales y potencialmente más orientadas a audiencias específicas, como documentales, cine independiente.

Grupo	Director	Conteo
1	Clint Eastwood	22
1	Kunihiro Yuyama	19
1	Robert Rodriguez	18
1	Sam Liu	18
2	Steven Spielberg	17
2	Tim Burton	11
2	Robert Zemeckis	10
3	Antoine Fuqua	2
3	Colin Trevorrow	2
3	Michael Dowse	2
3	Naoyoshi Shiotani	2

Se puede observar que en los distintos grupos se presentan las siguientes características: Cluster 1 agrupa a directores con una gran cantidad de películas, lo que podría significar que son responsables de sagas largas o producciones constantes. Cluster 2 contiene directores con menor volumen de películas pero con gran impacto cinematográfico. Cluster 3 tiene directores menos recurrentes en la base de datos, lo que sugiere películas más variadas en dirección.

Grupo	Calificación Promedio
2	6.746432
3	6.676119
1	6.454601

Si bien la diferencia no es tan notable, cuando se habla de calificaciones, esta mínima diferencia puede indicar factores. Se puede observar que las películas en el Grupo 2 pueden tener características que favorecen una mejor recepción por parte del público, como mayor presupuesto, actores reconocidos o mejor producción. Por otro lado, el Grupo 1, al tener la calificación más baja, podría estar compuesto por películas con menor impacto crítico.

Conclusiones

Basado en esta información se pudo detectar puntos clave:

- Grupo 1: son películas de bajo presupuesto con un alto rendimiento
 - Estas tienen un bajo presupuesto pero presentan un gran ingreso
 - Los géneros predominantes son: Horror, Thriller, Drama (películas de bajo costo suelen ser rentables en estos géneros).
 - Directores menos conocidos.

- Como se pudo observar en las calificaciones fue el menor impacto, porque estas películas no apuntan del todo a ser las mejores calificadas.
 - Si se quisiera hacer inversiones este grupo sería el mejor por su bajo presupuesto y gran recaudación.
 - Grupo 2: Blockbusters con Altos Presupuestos y Grandes Ingresos
 - Estas películas tienen alto presupuesto y gran ingreso
 - Géneros predominantes: Acción, Aventura, Ciencia Ficción.
 - Directores reconocidos: Spielberg, Zemeckis, Tim Burton.
 - Teniendo las películas mejores calificaciones.
 - Al ser Blockbusters aquí se podrían maximizar ingresos para tener franquicias o secuelas. Pero con riesgo financiero porque tienen un alto presupuesto y si fracasan sería una gran perdida.
 - Grupo 3: Producciones de Bajo Presupuesto y Bajo Rendimiento
 - Bajo presupuesto y bajo ingreso.
 - Géneros variados y películas independientes.
 - Este grupo de películas al ser menor pueden beneficiarse de modelos de distribución en streaming o festivales.
-

Analisis por Componentes principales

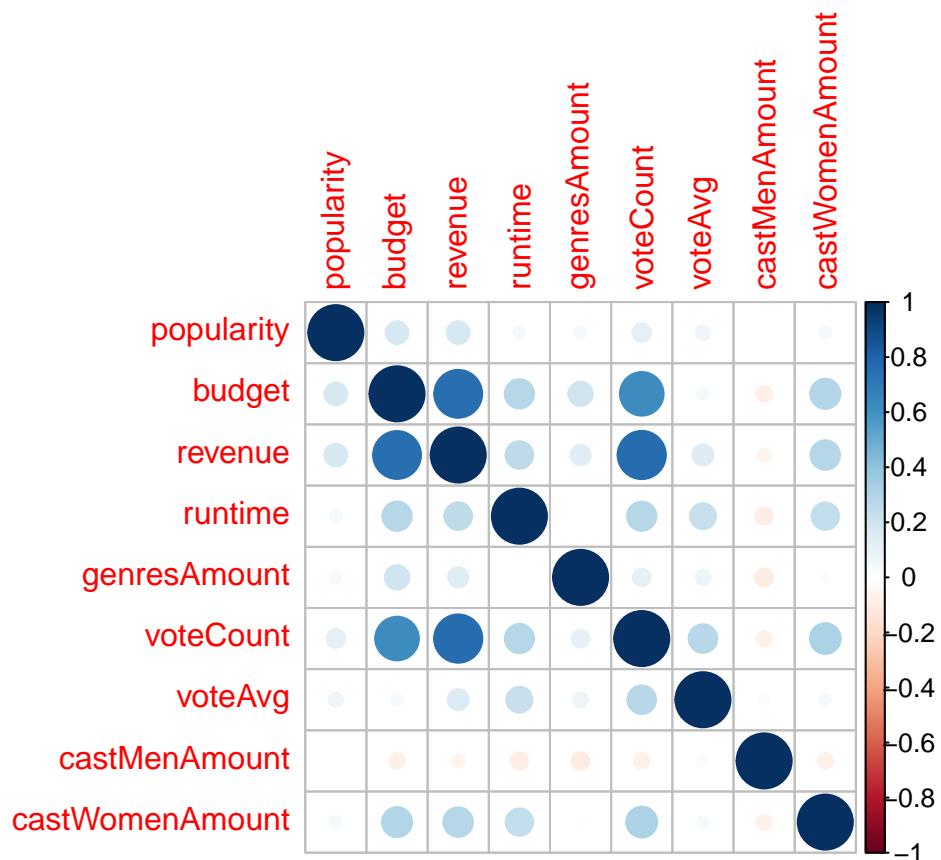
Para ello primero se requirió de la limpieza del dataset para contar solamente contar con un dataframe de variables cuantitativas. Este sería un ejemplo del Dataset con el que se trabajará el PCA.

Para ello se tomaron solo variables que eran inherentemente cuantitativas, ya que a pesar que las variables cualitativas aunque no se conviertan a un equivalente numérico, no tienen noción de orden y magnitud, y por lo tanto no aportando mucho valor al cálculo de componentes.

popularity	budget	revenue	runtime	genresAmount	voteCount	voteAvg	castMenAmount	castWomenAmount
20.880	4.0e+06	4257354	98	2	2077	5.7	9	15
9.596	2.1e+07	12136938	110	3	223	6.5	9	3
100.003	1.1e+07	775398007	121	3	16598	8.2	62	5
134.435	9.4e+07	940335536	100	2	15928	7.8	18	5
58.751	5.5e+07	677387716	142	3	22045	8.5	48	18
33.589	1.5e+07	356296601	122	1	9951	8.0	15	22

PCA es aplicable

Acto seguido, es importantes verificar si el PCA será una buena técnica para este conjunto de datos. Empezando por revisar la colinealidad de las variables:



Vemos un valor es 0.1038076, que es cercano a 0, y también podemos ver que hay una relación del estrecha (cercana al 60%) entre las ganancias y el presupuesto, así como número de votos y ganancias. Para reafirmar si el PCA será un método correcto, se corrieron tambien los test KMO y Test de esfericidad de Bartlet que arrojan los siguientes indicios:

KMO

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = as.matrix(PCA_raw_data))
## Overall MSA =  0.74
## MSA for each item =
##      popularity          budget         revenue        runtime   genresAmount
##            0.83           0.76           0.70           0.77           0.62
##      voteCount         voteAvg    castMenAmount castWomenAmount
##            0.76           0.51           0.69           0.86
```

Esfericidad de Bartlet

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 21952.59
##
## $p.value
## [1] 0
##
```

```
## $df
## [1] 36
```

Ambas estadísticas ofrecen resultados favorecedores al indicar que si hay relación entre las diferentes variables y datos con un MSA superior a 0.5 (MSA=0.71) y bartlet indicando que si hay interdependencia entre las variables ($p=0$). Es decir, podemos proceder a aplicar PCA.

Aplicación de PCA

Estas serían las componentes principales del subset de datos.

```
## [1] 0

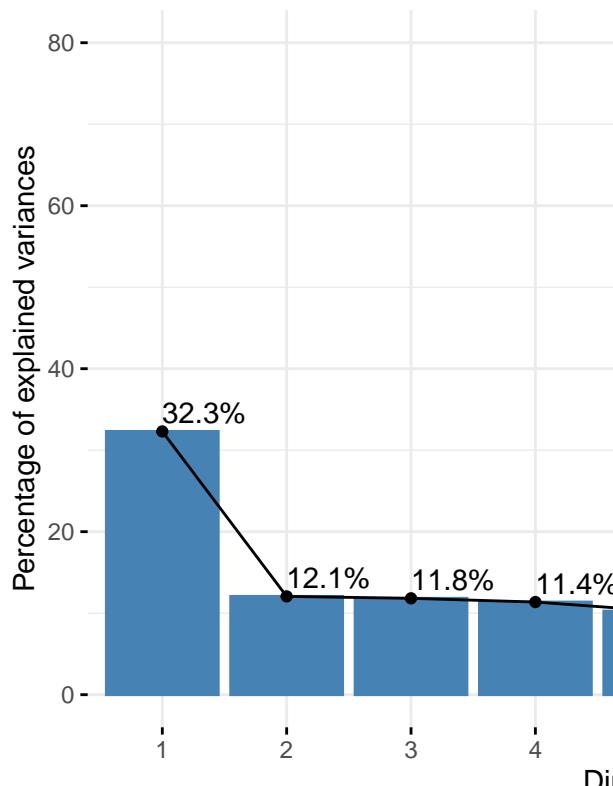
## Standard deviations (1, ..., p=9):
## [1] 1.7049488 1.0418268 1.0316224 1.0114821 0.9608926 0.8932974 0.8356471
## [8] 0.5598308 0.4328906
##
## Rotation (n x k) = (9 x 9):
##          PC1       PC2       PC3       PC4       PC5
## popularity 0.14442911 0.25163202 -0.046325371 -0.546055343 0.77245277
## budget    0.49165068 0.26462152 -0.065942086 0.091715572 -0.05615038
## revenue   0.51634229 0.19533683 -0.124308749 0.002249972 -0.11265397
## runtime   0.28181179 -0.55739929 0.005675743 0.061403136 0.12006918
## genresAmount 0.13279646 0.36462315 0.661119078 -0.168835944 -0.31140796
## voteCount  0.50291645 0.01211703 -0.091180674 -0.033431339 -0.16168200
## voteAvg   0.17315847 -0.57780254 0.203847257 -0.581365134 -0.20248834
## castMenAmount -0.09308855 0.13214542 -0.687088934 -0.385829772 -0.40577444
## castWomenAmount 0.28667649 -0.18273281 -0.137780490 0.416128762 0.21044219
##          PC6       PC7       PC8       PC9
## popularity 0.11602919 0.0051548668 -0.06925113 0.02303759
## budget    -0.07985002 0.1707947521 0.69240568 0.39669837
## revenue   -0.22250303 -0.0002660751 -0.03845208 -0.78491993
## runtime   0.26203349 0.7139916875 -0.10761065 -0.04053699
## genresAmount 0.51378232 0.0855564415 -0.12279137 -0.02205061
## voteCount  -0.24013787 -0.1401506521 -0.64205424 0.47097084
## voteAvg   -0.08592753 -0.3658471691 0.27172587 -0.02360236
## castMenAmount 0.42379450 0.0899690091 -0.01537798 0.01818567
## castWomenAmount 0.59444786 -0.5404791353 0.03656837 -0.03419912
```

Por la regla de Kaiser (las componentes con desviación estandar mayor a 1 son las que aportan mayor peso) se desvela que las primeras 77% componentes son las más importantes al explicar el 86% de la variación de los datos:

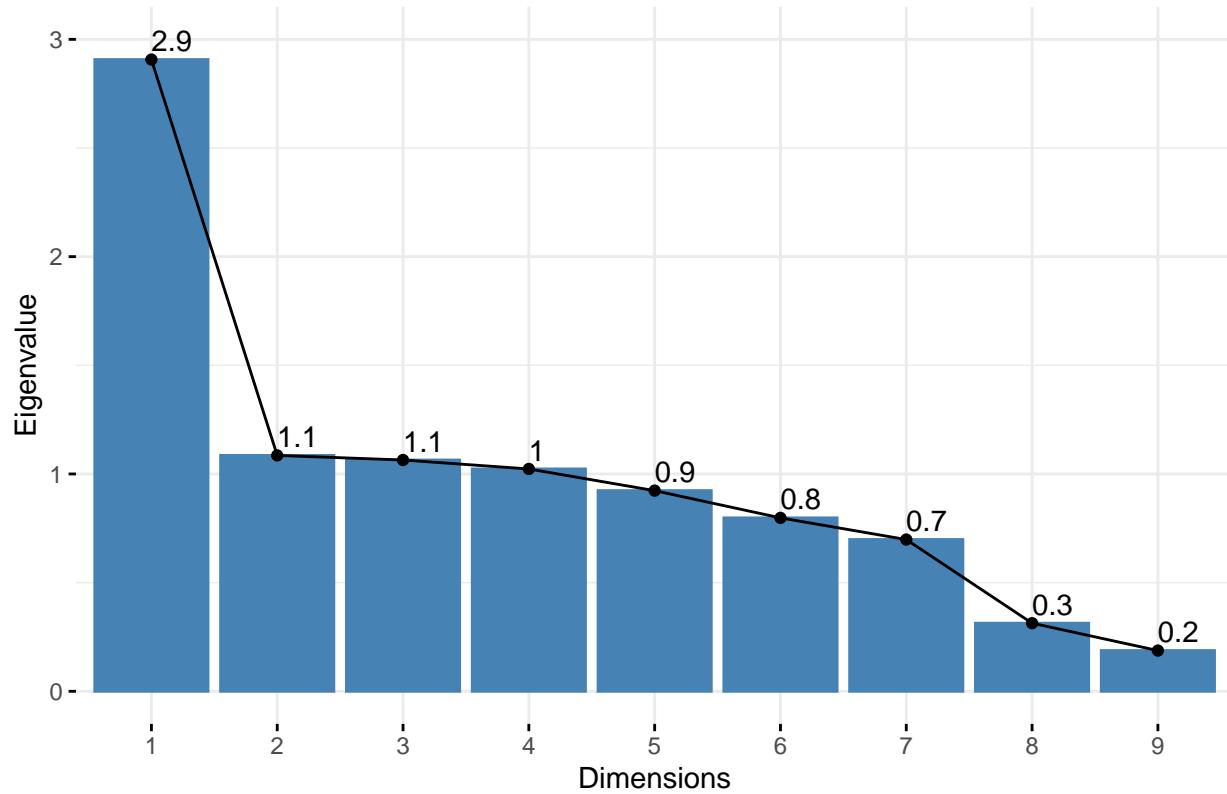
```
## Importance of components:
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Standard deviation 1.705 1.0418 1.0316 1.0115 0.9609 0.89330 0.83565
## Proportion of Variance 0.323 0.1206 0.1182 0.1137 0.1026 0.08866 0.07759
## Cumulative Proportion 0.323 0.4436 0.5618 0.6755 0.7781 0.86677 0.94436
##          PC8      PC9
## Standard deviation 0.55983 0.43289
## Proportion of Variance 0.03482 0.02082
## Cumulative Proportion 0.97918 1.00000
```

La gráfica recomienda el uso de 2 componentes, pero para encontrar, eso explicaría solo el 44% de la variabilidad de los datos, así que nos quedaremos en el análisis con componentes:

Scree plot



Scree plot

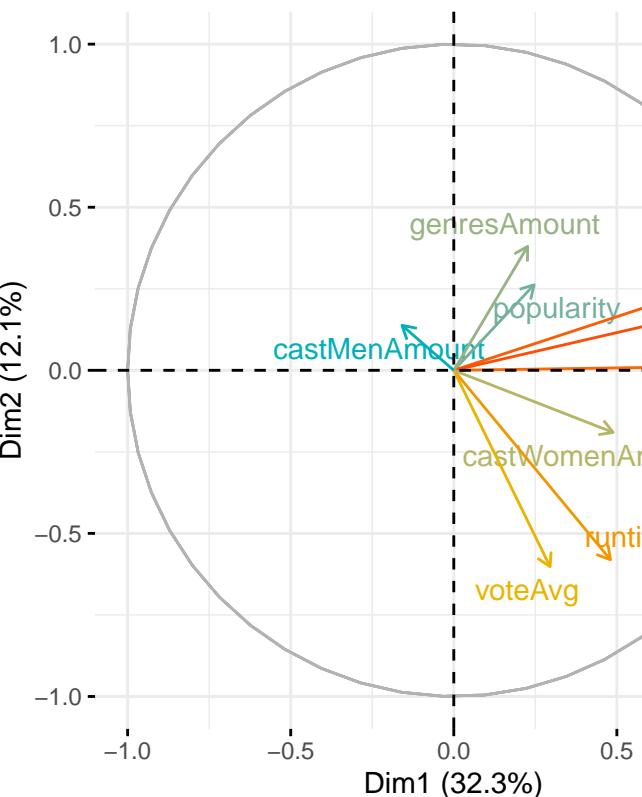


Interpretación de PSA

El siguiente gráfico muestra vectores proyectados sobre las 2 primeras componentes. Y se puede observar las relaciones descubiertas en pasos anteriores sobre la covarianza: Las componenetes de presupuesto, y número de votos y ganancias estan muy relacionadas positivamente entre, y bien representadas en la primera dimensión.

También se puede decir algo parecido del tiempo de expiración y el promedio de votos, solo que estos 2 ultimos

Variables – PCA



se encuentran menos representados en la 1 y 2 dimensión.