Universidad del Valle de Guatemala
Minería de Datos

Eric Mendoza 15002
Samantha Duarte 16256

Hoja de Trabajo 2 - Clustering

Preprocesamiento

El resumen de la información es el siguiente:

```
                                               :10842
    runtime                   genres                      production_companies   release_date      vote_count       vote_average
 Min.   :  0.0   Comedy         : 712                                :1030    1/1/09  :    28   Min.   :  10.0   Min.   :1.500
 1st Qu.: 90.0   Drama          : 712   Paramount Pictures  : 156    1/1/08  :    21   1st Qu.:  17.0   1st Qu.:5.400
 Median : 99.0   Documentary    : 312   Universal Pictures  : 133    1/1/07  :    18   Median :  38.0   Median :6.000
 Mean   :102.1   Drama|Romance  : 289   Warner Bros.        :  84    1/1/05  :    16   Mean   : 217.4   Mean   :5.975
 3rd Qu.:111.0   Comedy|Drama   : 280   Walt Disney Pictures:  76    10/10/14:    15   3rd Qu.: 145.8   3rd Qu.:6.600
 Max.   :900.0   Comedy|Romance : 268   Columbia Pictures   :  72    1/1/03  :    13   Max.   :9767.0   Max.   :9.200
                 (Other)        :8293   (Other)             :9315    (Other) :10755
  release_year    budget_adj           revenue_adj
 Min.   :1960   Min.   :        0   Min.   :0.000e+00
 1st Qu.:1995   1st Qu.:        0   1st Qu.:0.000e+00
 Median :2006   Median :        0   Median :0.000e+00
 Mean   :2001   Mean   : 17551040   Mean   :5.136e+07
 3rd Qu.:2011   3rd Qu.: 20853251   3rd Qu.:3.370e+07
 Max.   :2015   Max.   :425000000   Max.   :2.827e+09
```

Nos interesan solamente las variables cuantitativas porque se pueden utilizar para realizar el clustering. Y de las variables numéricas, solamente nos interesan las variables que describen a la película significativamente. Por lo que se utilizarán las siguientes.

- Popularity
- Budget
- Revenue
- vote_average
- budget_adj
- revenue_adj

Ahora, se desea que se cuente con información completa y comparable, por tal razón, se procede a eliminar las casillas donde haya información faltante y luego a estandarizar los valores de todas las columnas para que sean comparables.

datos <- na.omit(datos)
datos <- scale(datos)

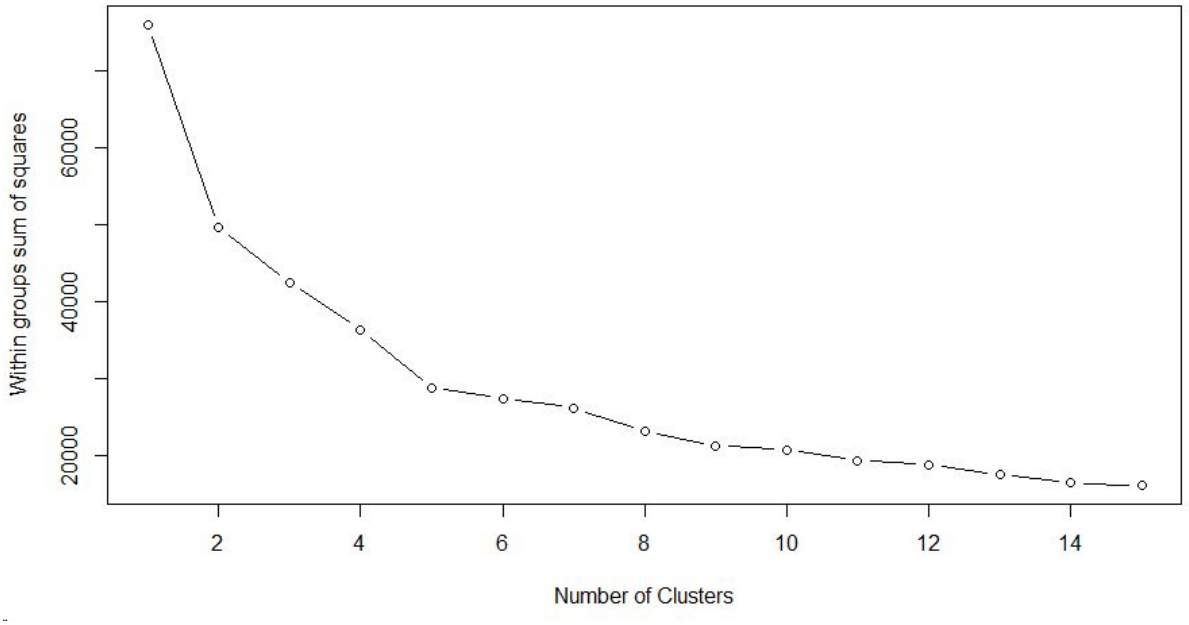
Selección de cantidad de grupos

Para seleccionar la cantidad de clusters adecuados para el conjunto de datos que se tiene, se procederá a utilizar 3 métodos diferentes con distintos valores de k (donde ‘k’ representa la cantidad de clusters). Luego se analizará la calidad de los clusters para cada método y cada k utilizado. Por último, se procederá a realizar una comparación de los tres métodos utilizados, buscando encontrar coincidencias entre los tres métodos para poder dar así un argumento suficientemente fundamentado de la elección de cantidad de clusters para el conjunto de datos.

Clustering

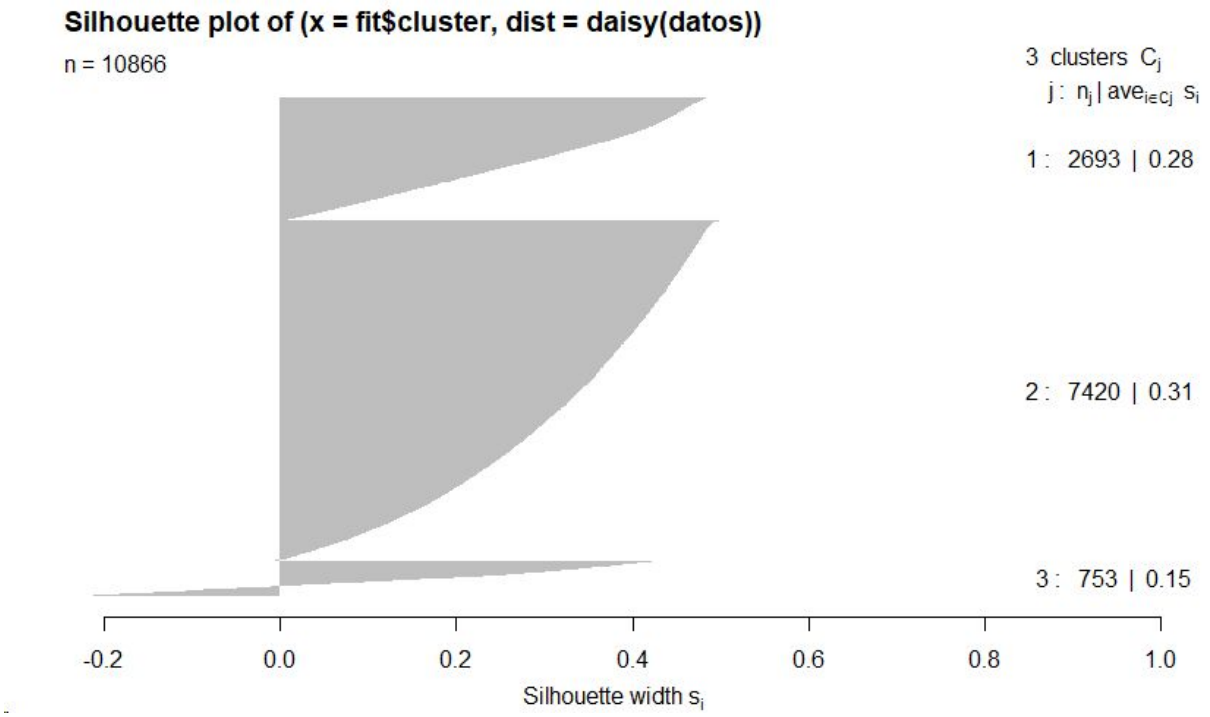
K-Means

Para calcular el número de clusters necesarios utilizando el método de k-means, se debe variar la variable K para obtener diferentes cantidades de clusters. Luego, para cada grupo de clusters, se encuentra la suma errores cuadrados y se grafica. La gráfica de codo para el conjunto de datos seleccionado es el siguiente:



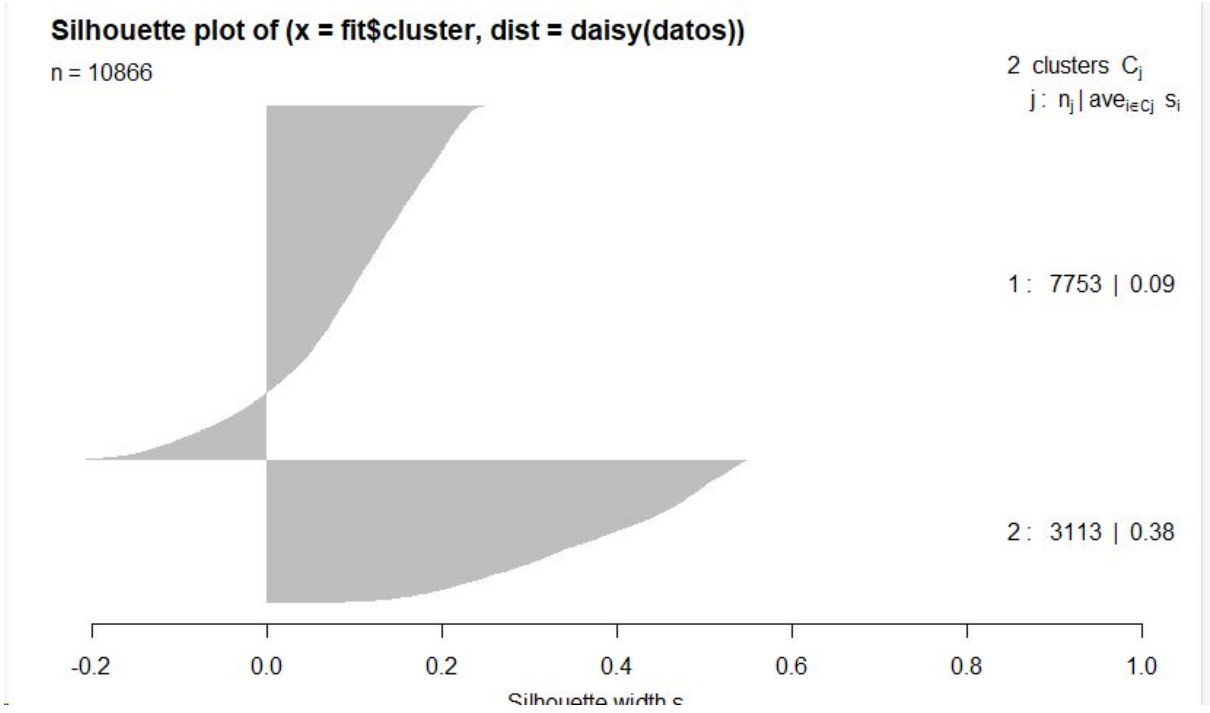
Gráfica 1: Gráfico de codo según la suma de errores cuadrados para el conjunto de datos

Se busca tener una cantidad de clusters que minimice el error, pero que sea significativo para el grupo. Como se observa en la Gráfica 1, existe una curva suave, por lo que no se puede determinar por este método cuál debería de ser la cantidad de clusters para este método. Por lo tanto, se procederá a realizar un test de Silhouette para determinar la calidad de los clusters. Se escoge K = 3 debido a que es el segundo después del codo de la gráfica, su gráfica de silueta es la siguiente.



Gráfica 2: Gráfico de silueta para k=3 utilizando k-medias

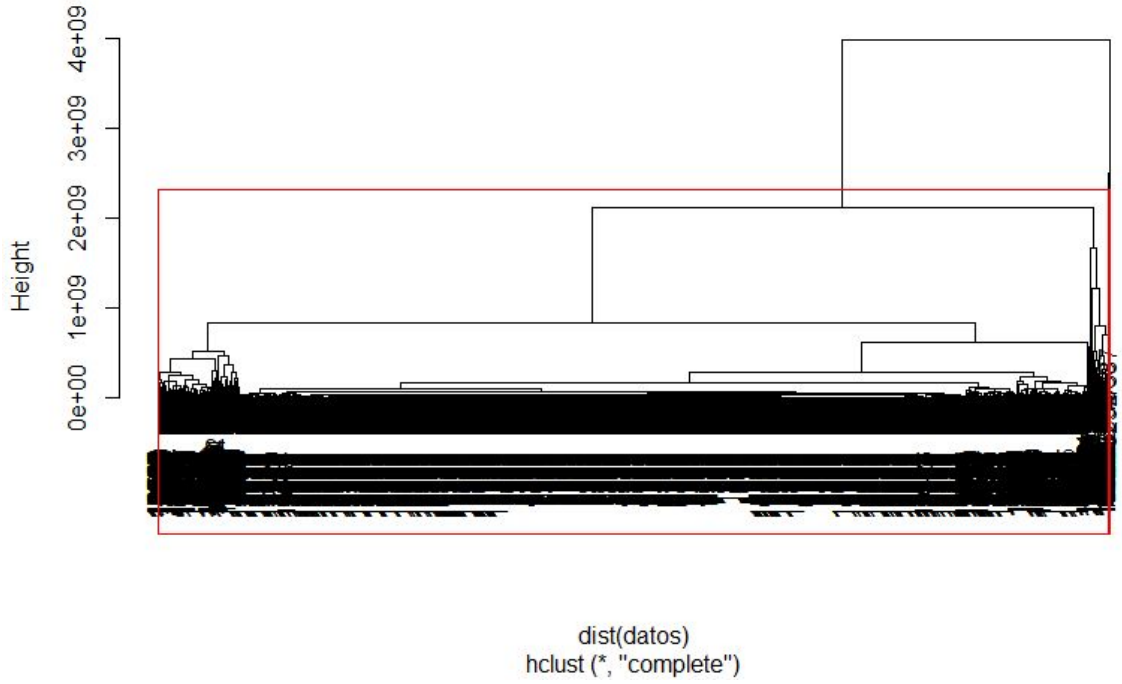
Como se observa en la gráfica, si se utilizan 3 clusters, se obtiene una silueta positiva, sin embargo, esta no es tan cercana a 1 como se quisiera. Por lo que se plantea que, para el conjunto de datos escogido, el método de k-medias no es el más apropiado. A continuación se muestra el gráfico de siluetas para 2 clusters con k medias.



Gráfica 3: Gráfico de silueta para k=2 utilizando k-medias

En este caso, valor de la silueta es aún más pequeño que en el caso anterior.

Clustering Jerárquico



Gráfica 4: Dendograma de clustering jerárquico según distancia máxima posible

Para utilizar el método de clustering jerárquico se coloca cada dato en su propio clúster. Luego se identifican los dos clusters más cercanos y se combinan en uno solo. Este proceso se repite hasta que todos los puntos están en un clúster. Puede observarse que existe una gran cantidad de datos y rangos equivalentes, por lo que no se observan diferencias con respecto al método anterior.

Se realizó un test de Silhouette con K=3. Así se determinó 0.9376693 para n, valor bastante cercano a 1. Esto indica la posibilidad de que el método de clustering jerárquico sea el método más óptimo para estos datos, idea que no se refleja tan bien en la gráfica 3.

Fuzzy means

En este método cada elemento tiene probabilidad de pertenecer a cada clúster, es decir, cada elemento tiene un grupo de coeficientes de pertenencia que pertenecen a un grado de membresía a un clúster específico. Es diferente al método de k-means, en el cual cada elemento es relacionado exactamente a un clúster. Se decidió generar 3 clústers. El método de la silueta para fuzzy means indicó un valor de 0.8027407 para n. Un valor cercano a uno, pero no tan cercano como el método de clustering jerárquico.

Calidad de agrupamiento

Cada método de clustering fue analizado con el método de Silhouette en sus respectivas secciones. Ahí se pueden encontrar sus respectivos análisis de calidad. Siempre se buscó que el valor de Silhouette fuera mayor a 0 y, además, que estuviera cercano a 1.

Análisis de los grupos generados

Al comparar los métodos de clustering utilizados, se llega a la conclusión que se deben escoger 3 clusters para el conjunto de datos escogido. Esto se debe a que en los tres métodos se obtuvo que dicho k logra caracterizar a las películas y es útil para realizar futuros análisis de la información. Además se obtuvo valores cercanos a 1 en las pruebas de silueta, lo cual indica que 3 es un número apropiado de clústers. Para utilizar los clusters con futuros análisis, se agrega una nueva columna al set de datos con el número del cluster al que pertenece la entrada.

datos <- data.frame(datos, fit$cluster)

```
   popularity          budget            revenue         vote_average      release_year       budget_adj
 Min.   :-0.64626   Min.   :-0.47312   Min.   :-0.3404   Min.   :-4.78529   Min.   :-3.2251   Min.   :-0.51160
 1st Qu.:-0.43878   1st Qu.:-0.47312   1st Qu.:-0.3404   1st Qu.:-0.61480   1st Qu.:-0.4935   1st Qu.:-0.51160
 Median :-0.26254   Median :-0.47312   Median :-0.3404   Median : 0.02682   Median : 0.3650   Median :-0.51160
 Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.: 0.06736   3rd Qu.: 0.01211   3rd Qu.:-0.1352   3rd Qu.: 0.66843   3rd Qu.: 0.7553   3rd Qu.: 0.09626
 Max.   :32.33334   Max.   :13.27505   Max.   :23.4325   Max.   : 3.44876   Max.   : 1.0675   Max.   :11.87685
  revenue_adj        fit.cluster
 Min.   :-0.3551   Min.   :1.000
 1st Qu.:-0.3551   1st Qu.:2.000
 Median :-0.3551   Median :2.000
 Mean   : 0.0000   Mean   :2.358
 3rd Qu.:-0.1222   3rd Qu.:3.000
 Max.   :19.1918   Max.   :3.000
```

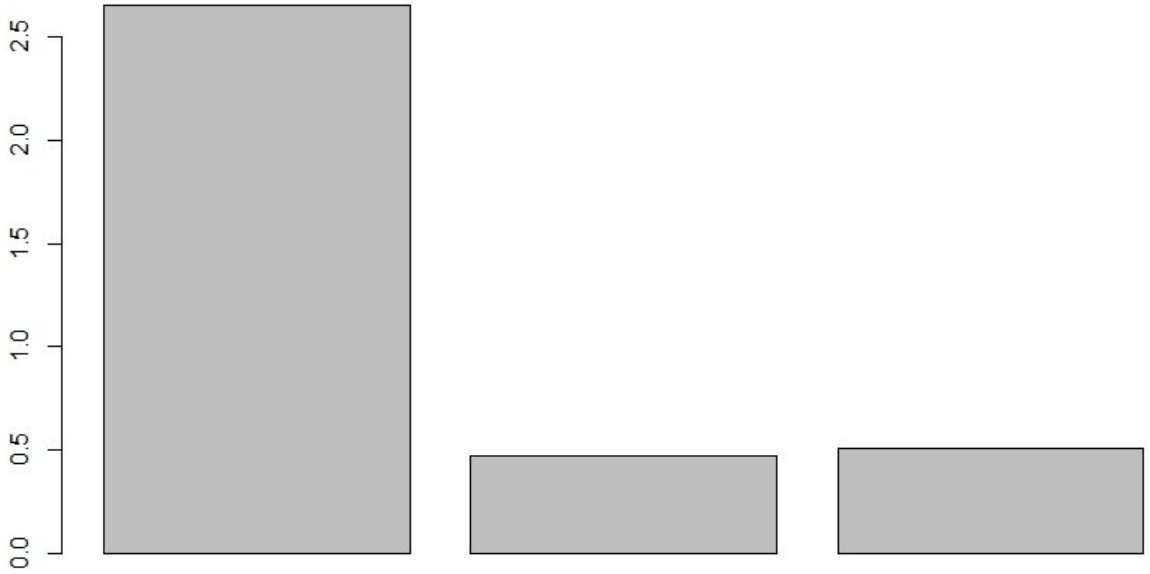
Cuadro 1: Resumen de la información ahora con columna cluster

Ahora se realizará un análisis para ver el presupuesto y las ganancias medias de las películas según su cluster.

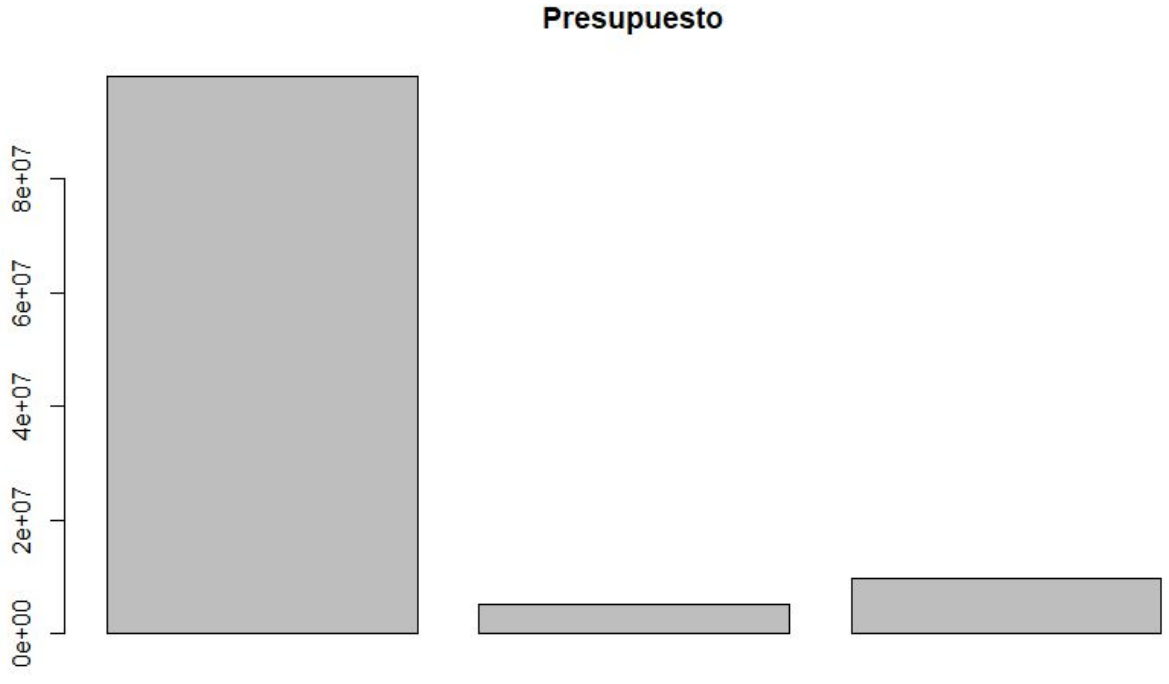
```
  Group.1       id imdb_id popularity   budget   revenue original_title cast homepage director tagline keywords overview
1       1 37093.07      NA  2.6530982 98084814 353946092             NA   NA       NA       NA      NA       NA       NA
2       2 17469.64      NA  0.4689349  5046043  18418956             NA   NA       NA       NA      NA       NA       NA
3       3 86641.04      NA  0.5072241  9632886  15713819             NA   NA       NA       NA      NA       NA       NA
   runtime genres production_companies release_date vote_count vote_average release_year budget_adj revenue_adj
1 119.40239     NA                  NA           NA 1560.60823     6.434529     2004.025  109059918   421233285
2 106.69625     NA                  NA           NA   96.03416     6.232826     1982.884   10666035    42216227
3  98.63329     NA                  NA           NA  125.12116     5.834677     2007.741   10763322    17149354
  fit.cluster
1           1
2           2
3           3
```

Cuadro 2: Media de todos los datos según su cluster

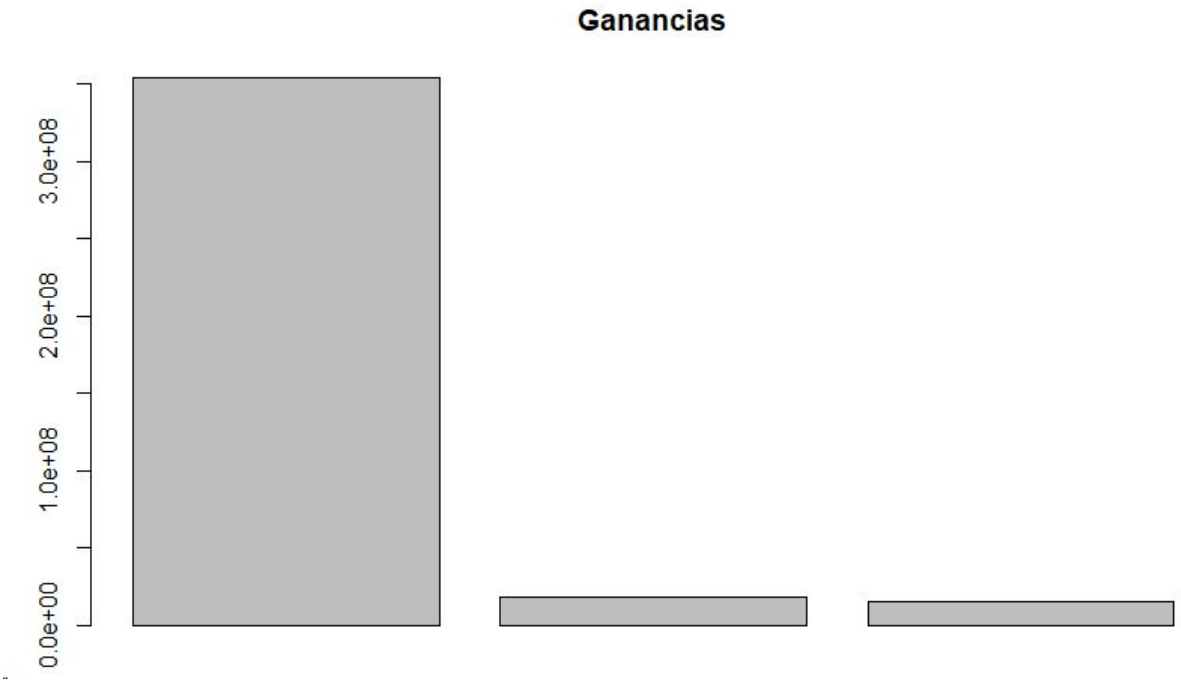
Como se observa en el Cuadro 2, la media de calificación para las películas en el cluster 1, 2 y 3, fueron 6.43, 6.23 y 5.8 respectivamente. También se observa que las películas del cluster 2 son más antiguas que las del cluster 1 y 3. Las películas que están en el cluster 1 tienen mejor popularidad que las película de los otros dos clusters, además, la media de presupuesto y de ingresos generados para este clúster es mucho mayor que para los otros dos. Por lo que se podría sacar la hipótesis que las películas con mejores presupuestos tienden a tener una mejor aceptación que las que tienen menor presupuesto.



Gráfica 6: Popularidad de clusters



Gráfica 7: Presupuesto según cluster



Gráfica 8: Ganancias según cluster.

Descripción del trabajo futuro

Ahora, lo que se debería proceder a realizar es el análisis de las variables no numéricas según el cluster al que pertenecen. Por ejemplo, podríamos encontrar cuales son las categorías de las películas más frecuentes en los clusters, o ver cuales directores son los más frecuentes en los clusters, entre otros más. También podría estudiarse la relación entre un cluster, de alguna variable específica, junto con una variable no número ca. Cómo la correlación popularidad-trama, para identificar qué tramas son las más populares.