Universidad del Valle de Guatemala
Minería de Datos
Lynette García

Ivette Cardona -16020
Cristopher Recinos - 16005
Fecha: 27/02/2019

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.**

Variables descartadas	*Justificación*
• id • imdb_id • original_title • cast • homepage • director • tagline • keyword • overview • genres • production_companies • release_date • vote_average	Debido a que son variables categóricas y al realizar el clustering no se pueden realizar cálculos entre ellas, por ejemplo calcular la distancia, promedios, entre otros. En el caso de vote_average, se descarta debido a que ya se tiene una variable vote_count que cuenta con la misma información pero expandida.

Variables	*Justificación*
• popularity • budget • revenue • runtime • vote_count • release_year	Estas variables nos serán útiles para el clustering ya que es fácil comparar entre sus valores.

2. **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.**

Diagrama de codo (data)- método con for

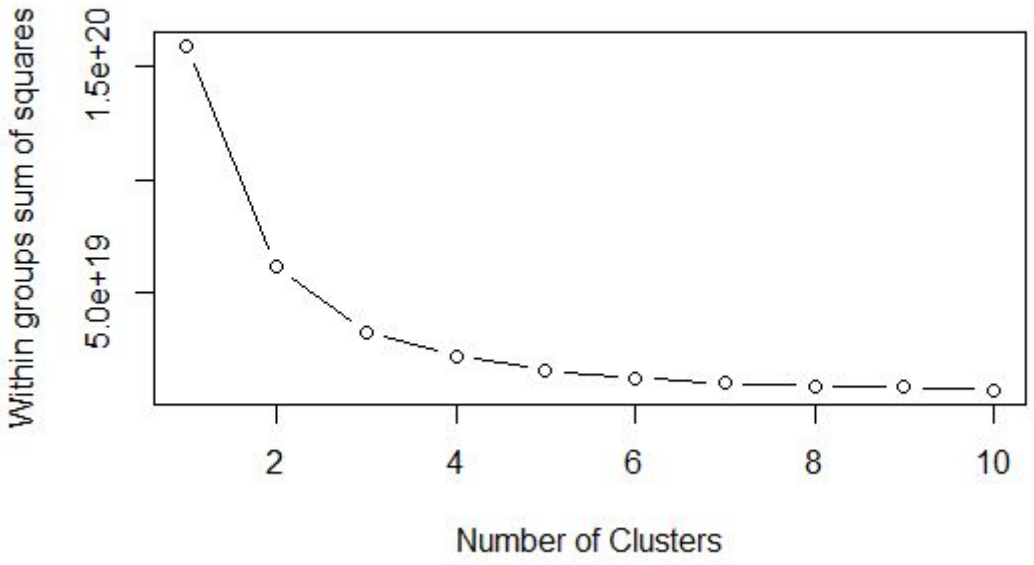


Figura 1: Diagrama de codo

Diagrama de codo (data escalada)- método con for

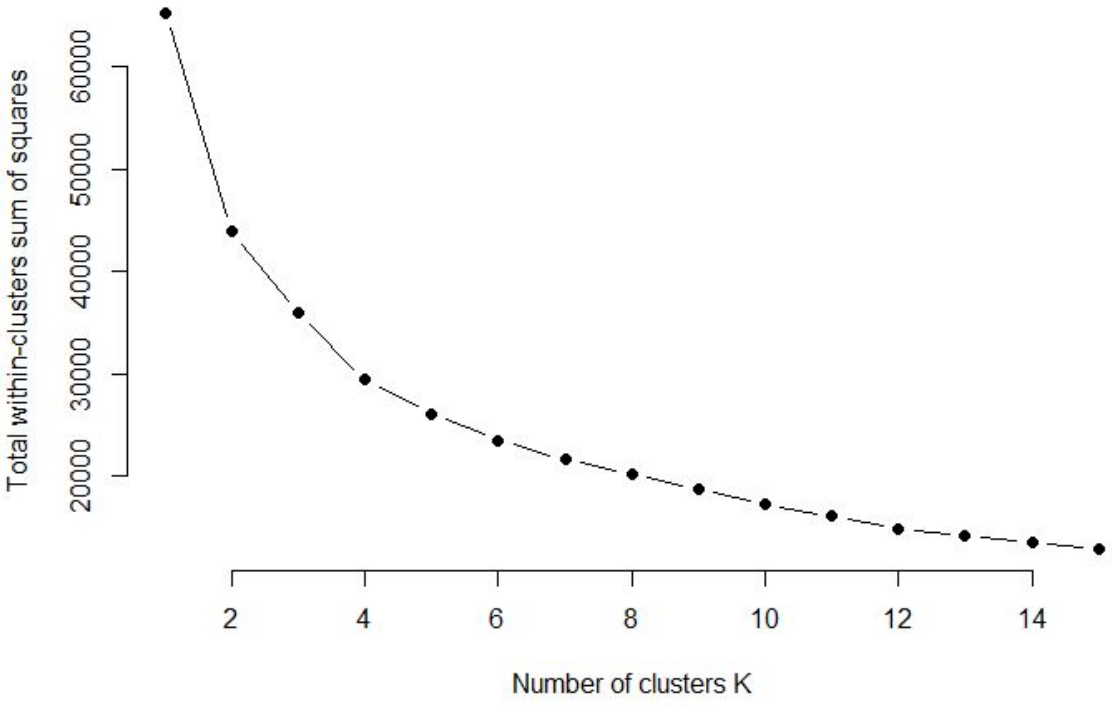


Figura 2: Diagrama de codo

Otros métodos

```
> clusGap(data, kmeans, 10, B = 100, verbose = interactive())
Clustering k = 1,2,..., K.max (= 10): .. done
Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
.................................................. 50
.................................................. 100
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = data, FUNcluster = kmeans, K.max = 10, B = 100, verbos
e = interactive())
B=100 simulated reference sets, k = 1..10; spaceH0="scaledPCA"
 --> Number of clusters (method 'firstSEmax', SE.factor=1): 1
          logW   E.logW      gap      SE.sim
 [1,] 26.06149 28.62571 2.564212 0.003753474
 [2,] 25.67651 28.05976 2.383252 0.003148747
 [3,] 25.36657 27.79246 2.425891 0.002952915
 [4,] 25.14584 27.63761 2.491767 0.003452918
 [5,] 25.08465 27.54707 2.462420 0.013845323
 [6,] 24.92615 27.47552 2.549370 0.003400851
 [7,] 24.81636 27.41004 2.593684 0.003747861
 [8,] 24.70129 27.34030 2.639008 0.003691755
 [9,] 24.63596 27.26997 2.634011 0.003703511
[10,] 24.59246 27.20288 2.610419 0.004776286
There were 12 warnings (use warnings() to see them)
```

Figura 3: Método clusGap

Kmeans

Al realizar el método kmean, se obtuvieron los siguientes resultados:

- centers 2:

```
Within cluster sum of squares by cluster:
[1] 14396.97 29478.60
 (between_SS / total_SS =  32.7 %)
```

Figura 4: kmeans 2

- centers 10:

```
Within cluster sum of squares by cluster:
 [1] 1142.1898  650.7891 1690.2739  585.3126  245.3316 2400.1289
 [7] 2239.2536 3499.3671 3317.4148 1555.1151
 (between_SS / total_SS =  73.4 %)
```

Figura 5: kmeans 10

- centers 12:

```
Within cluster sum of squares by cluster:
 [1] 1110.53261  645.31954 1529.16040  240.25668 2305.57970 1861.87454 1278.79099
 [8] 1488.68237 2103.59254  535.17357   87.69403 1738.07419
 (between_SS / total_SS =  77.1 %)
```

Figura 6: kmeans 12

- centers 20:

```
Within cluster sum of squares by cluster:
 [1]  363.41998  244.39788  728.90215  552.07626  443.68485
 [6]  875.66444  681.82899  677.68048   87.69403  639.80650
[11]  781.52076  181.32002  780.59120  622.13555  225.45362
[16]  902.20112  161.47529 1279.94116  336.16201  496.86904
 (between_SS / total_SS =  83.0 %)
```

Figura 7: kmeans 20

Bayesian Inference Criterion for k means

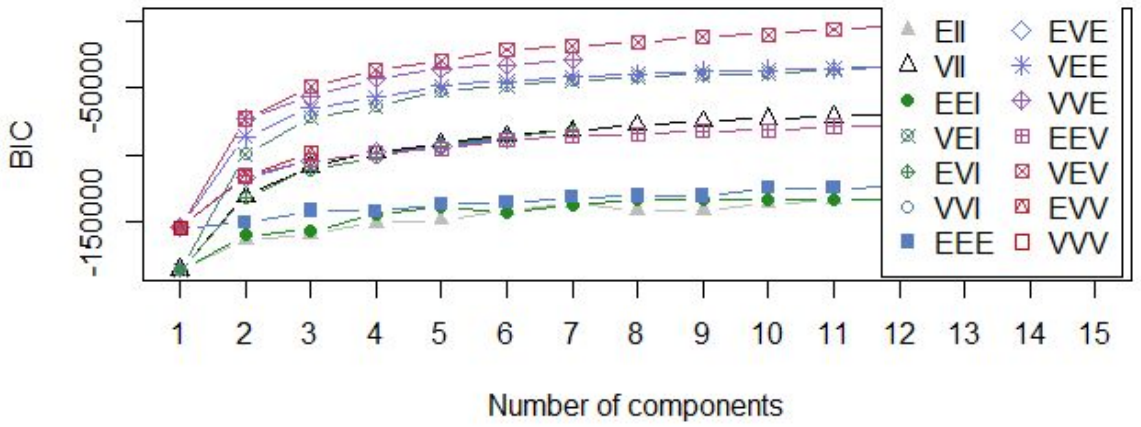


Figura 8: BIC

Resultado de clusGap

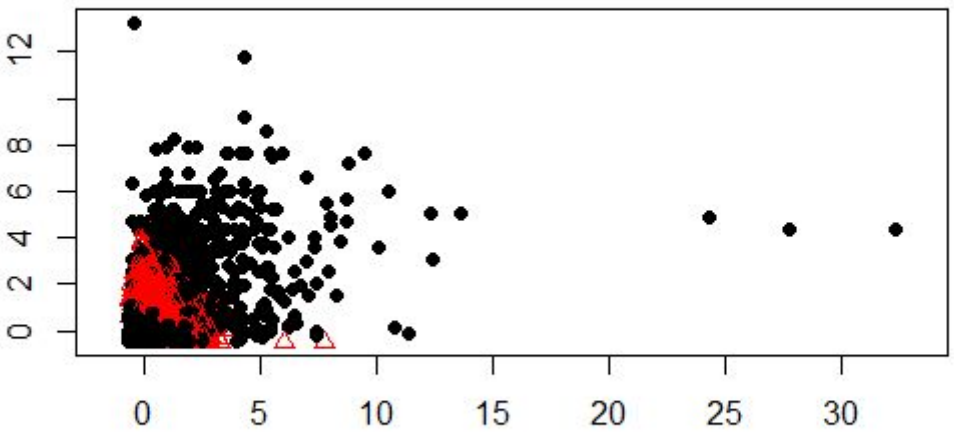


Figura 9: Gap stats

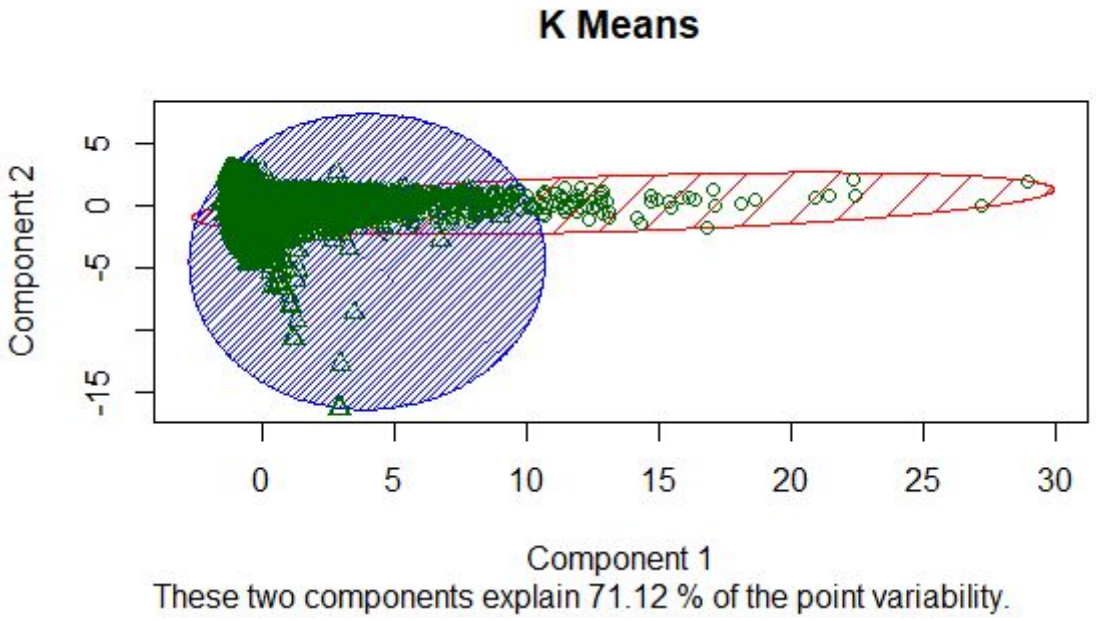
Análisis de los resultados:

Al comparar los resultados obtenidos, se puede observar claramente que el corte de los datos en los diagramas de codo obtenidos se realizan en el valor 2. Además, en la Figura 8, los datos se reúnen más en el punto 2, lo que indica que el resultado óptimo es 2. A su vez, en la Figura 9, los datos se ven divididos en 2.

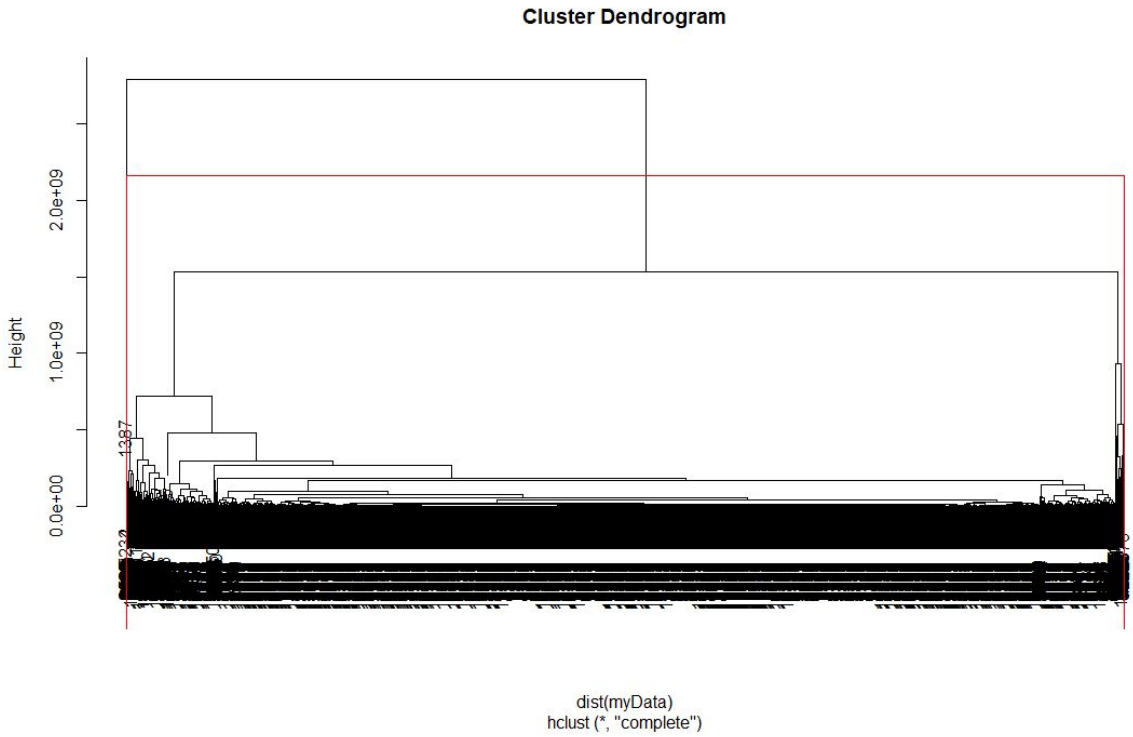
En conclusión, se realizaran 2 grupos de datos para los datos de las películas elegidos.

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

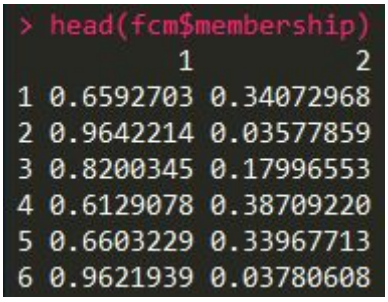
Kmeans

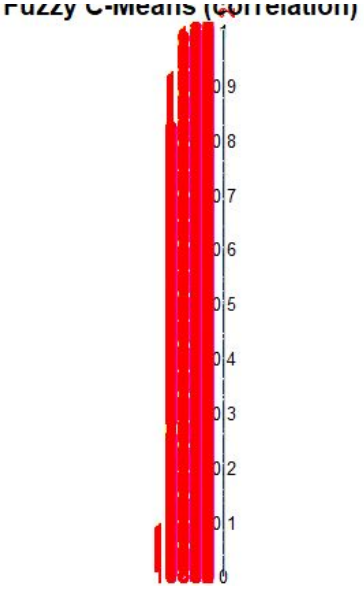


Clustering jerárquico



Fuzzy C-Means



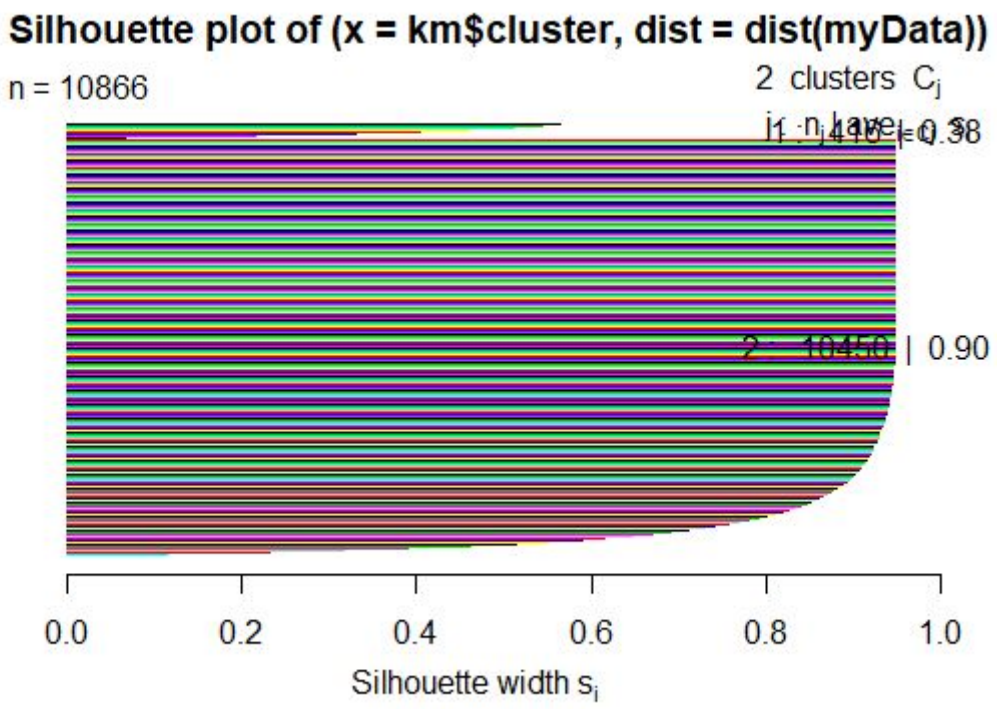


Análisis de los resultados:

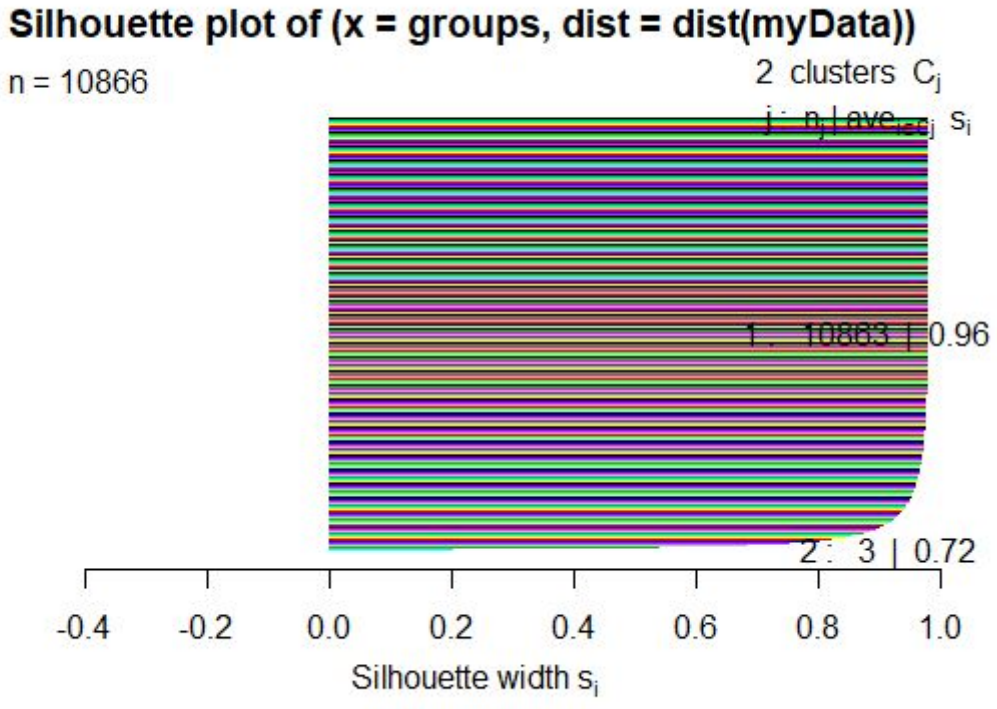
Debido a la gran cantidad de datos, se puede complicar la visualización de los datos por medio de las gráficas. Sin embargo, en los 3 casos se puede apreciar la distribución de los datos en dos grupos, como se estableció previamente. Los datos son bastante parecidos por lo que los grupos están casi encima del otro.

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

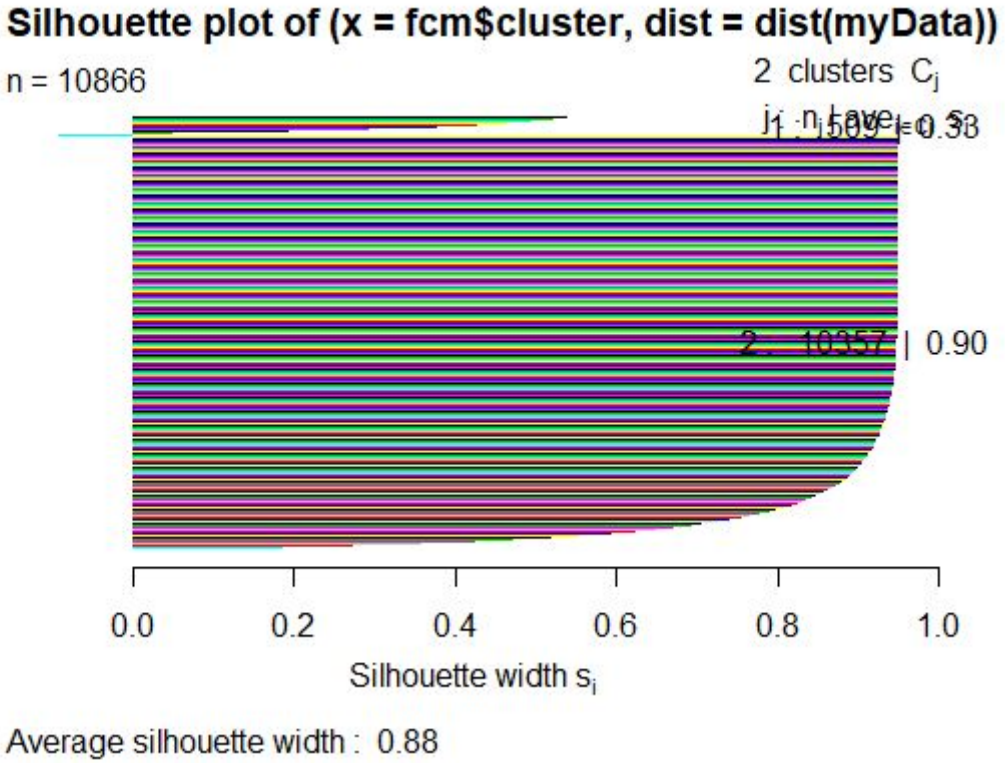
Silueta K-Means



Silueta en Clustering Jerárquico



Silueta en Fuzzy C-Means



Análisis de los resultados:

En los tres casos se tienen promedios que favorecen a la cantidad de grupos en los que se decidió dividir la data. Siendo 0.88, 0.96 y 0.88.

Los resultados cercanos a 1 nos muestran que la gran mayoría de datos están bien emparejados con sus respectivos clusters. En el que se obtuvo mejor resultado es el clustering jerárquico, por lo que nos parece apropiado utilizarlo.

Esto implica que la elección de 2 grupos de datos satisface a este caso en particular.

5. **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.**

Medidas de tendencia central

	Media	Mediana	Moda
popularity	0.646441	0.3838555	0.272995
budget	14625701	0	0
revenue	39823320	0	0
runtime	102.0709	99	90
vote_count	217.3897	38	10

Tabla de frecuencia release_year

1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
32	31	32	34	42	35	46	40	39	31	41	55	40	55	47	44
1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
47	57	65	57	78	82	81	80	105	109	121	125	145	137	132	133
1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
133	178	184	175	204	192	210	224	227	242	266	281	307	364	408	438
2008	2009	2010	2011	2012	2013	2014	2015								
496	533	490	540	588	659	700	629								

El runtime de las películas no varía mucho. Esto es algo esperado, ya que hay un estándar de duración del que la mayoría de películas no se sale. Se puede ver claramente en la moda de 90 minutos y en la media de no mucho más que 100 minutos. Este campo no nos dice mucho al agrupar las películas.

Por el contrario, popularity y vote_count son dos campos que si nos dicen más. Estos campos nos dicen que películas gustaron más al público.

Budget y revenue nos muestra que, en promedio, las ganancias de las películas superan a al presupuesto que tenían.

Por último podemos ver en la tabla de frecuencia de release_year, que cada año salen más películas que el anterior.