Proyecto1

Juan Luis Solórzano (carnet: 201598) Micaela Yataz (carnet: 18960)

2025-01-20

```
##        id             budget             genres            homePage
##  Min.   :     5   Min.   :        0   Length:10000       Length:10000
##  1st Qu.: 12286   1st Qu.:        0   Class :character   Class :character
##  Median :152558   Median :   500000   Mode  :character   Mode  :character
##  Mean   :249877   Mean   : 18551632
##  3rd Qu.:452022   3rd Qu.: 20000000
##  Max.   :922260   Max.   :380000000
##  productionCompany  productionCompanyCountry productionCountry
##  Length:10000       Length:10000             Length:10000
##  Class :character   Class :character         Class :character
##  Mode  :character   Mode  :character         Mode  :character
##
##
##
##     revenue             runtime        video           director
##  Min.   :0.000e+00   Min.   :  0.0   Mode :logical   Length:10000
##  1st Qu.:0.000e+00   1st Qu.: 90.0   FALSE:9430      Class :character
##  Median :1.631e+05   Median :100.0   TRUE :84        Mode  :character
##  Mean   :5.674e+07   Mean   :100.3   NA's :486
##  3rd Qu.:4.480e+07   3rd Qu.:113.0
##  Max.   :2.847e+09   Max.   :750.0
##     actors          actorsPopularity   actorsCharacter    originalTitle
##  Length:10000       Length:10000       Length:10000       Length:10000
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     title           originalLanguage     popularity        releaseDate
##  Length:10000       Length:10000       Min.   :    4.258   Length:10000
##  Class :character   Class :character   1st Qu.:   14.578   Class :character
##  Mode  :character   Mode  :character   Median :   21.906   Mode  :character
##                                        Mean   :   51.394
##                                        3rd Qu.:   40.654
##                                        Max.   :11474.647
##     voteAvg         voteCount      genresAmount    productionCoAmount
##  Min.   : 1.300   Min.   :    1   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 5.900   1st Qu.:  120   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 6.500   Median :  415   Median : 3.000   Median : 3.000
##  Mean   : 6.483   Mean   : 1342   Mean   : 2.596   Mean   : 3.171
##  3rd Qu.: 7.200   3rd Qu.: 1316   3rd Qu.: 3.000   3rd Qu.: 4.000
##  Max.   :10.000   Max.   :30788   Max.   :16.000   Max.   :89.000
##  productionCountriesAmount  actorsAmount    castWomenAmount
```

```
##  Min.   :  0.000          Min.   :     0   Length:10000
##  1st Qu.:  1.000          1st Qu.:    13   Class :character
##  Median :  1.000          Median :    21   Mode  :character
##  Mean   :  1.751          Mean   :  2148
##  3rd Qu.:  2.000          3rd Qu.:    36
##  Max.   :155.000          Max.   :919590
##  castMenAmount
##  Length:10000
##  Class :character
##  Mode  :character
##
##
##
```

1. Clustering

1.1. 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.

Como el algoritmo de k-medias necesitan de alguna medida de distancia, entre los datos, en una primera instancia vamos a tomas solo las variables numéricas y vamos a quitar el id por ser como el nombre de una película. Las variables que tomaremos en consideración son las siguientes:

```
## 'data.frame':    10000 obs. of  10 variables:
##  $ budget                   : int  4000000 21000000 11000000 94000000 55000000 15000000 839727 12800
##  $ revenue                  : num  4.26e+06 1.21e+07 7.75e+08 9.40e+08 6.77e+08 ...
##  $ runtime                  : int  98 110 121 100 142 122 119 141 126 149 ...
##  $ popularity               : num  20.9 9.6 100 134.4 58.8 ...
##  $ voteAvg                  : num  5.7 6.5 8.2 7.8 8.5 8 8 7.9 7.5 8.2 ...
##  $ voteCount                : int  2077 223 16598 15928 22045 9951 4253 1335 8726 1963 ...
##  $ genresAmount             : int  2 3 3 2 3 1 2 2 5 2 ...
##  $ productionCoAmount       : int  2 3 2 1 2 2 2 26 2 1 ...
##  $ productionCountriesAmount: int  1 2 1 1 1 1 1 12 1 1 ...
##  $ actorsAmount             : int  25 15 105 24 76 40 152 29 117 24 ...
```

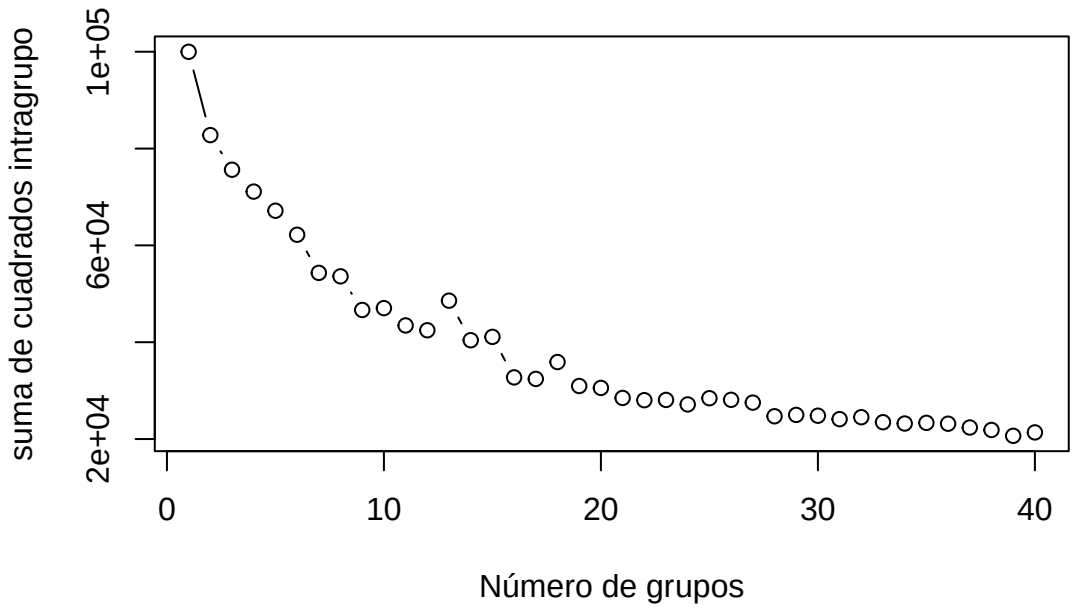
Luego los normalizamos y centramos para que todas las variables estén en una escala comparable.

1.2. Analice la tendencia al agrupamiento usando el estadístico de Hopkings y la VAT (Visual Assessment of cluster Tendency). Esta última hágala si es posible, teniendo en cuenta las dimensiones del conjunto de datos. Discuta sus resultados e impresiones.

El estadístico de Hopkings es de 1 que es lejano a 0.5, entonces los datos no son aleatorios. Sin embargo no haremos un VAT por ser difícil de visualizar e interpretar con 10 variables.

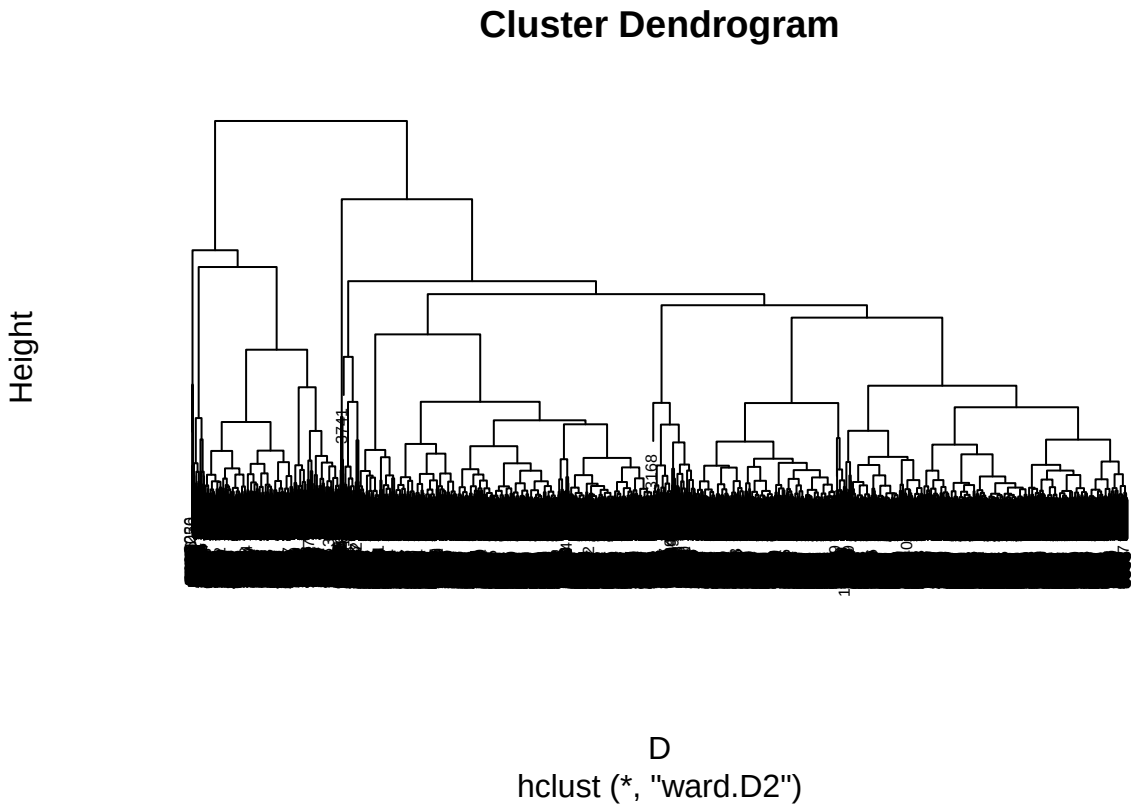
1.3. Determine cuál es el número de grupos a formar más adecuado para los datos que está trabajando. Haga una gráfica de codo y explique la razón de la elección de la cantidad de clústeres con la que trabajará.

Para ello se usará el metodo de Codo

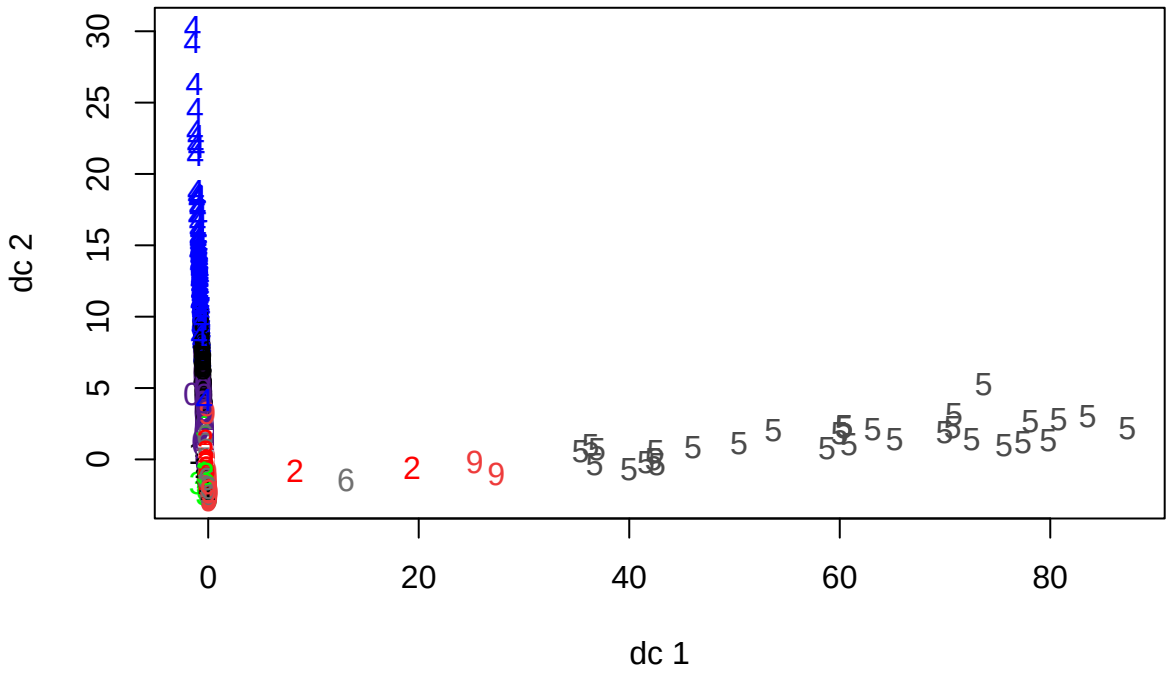


Como a partir de 10 grupos en adelante la suma de cuadrados intragrupo no disminuye significativamente se elegirán 10 grupos.

1.4. Utilice los algoritmos k-medias y clustering jerárquico para agrupar. Compare los resultados generados por cada uno.




```
plotcluster(datos_num,kmeans_result$cluster)
```



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

1.6. 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.