Universidad del Valle de Guatemala
Minería de Datos

Rene Ventura #19
Andres Paíz #19
Eduardo Ramírez #19946

Hoja de trabajo #2
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.

id: cuantitativa continua
Budget: cuantitativa continua
genres: cualitativa nominal
homePage: cualitativa nominal
productionCompany:cualitativa nominal
productionCompanyCountry: cualitativa nominal
productionCountry: cualitativa nominal
revenue: cuantitativa continua
Runtime: cuantitativa continua
video: cualitativa nominal
actors: cualitativa nominal
actorsPopularity: cuantitativa continua
actorsCharacter: cualitativa nominal
originalTitle: cualitativa nominal
title: cualitativa nominal
OriginalLanguage: cualitativa nominal
popularity: cuantitativa continua
releaseDate: cuantitativa discreta
voteAvg: cuantitativa continua
voteCount: cuantitativa discreta
genresAmount: cuantitativa discreta
productionCoAmount: cuantitativa discreta
productionCountriesAmount: cuantitativa discreta
actorsAmount: cuantitativa discreta
castWomenAmount: cualitativa
castMenAmount: culitativa

No se utilizaron las variables cualitativas, para la utilización de agrupamientos solamente tomamos en cuenta las que eran cuantitativas, como se puede observar en la imagen proporcionada.

Universidad del Valle de Guatemala
Minería de Datos

Rene Ventura #19
Andres Paíz #19
Eduardo Ramírez #19946

2. Analice la tendencia al agrupamiento usando el estadístico de Hopkings y la VAT (Visual Assessment of cluster Tendency). Discuta sus resultados e impresiones.

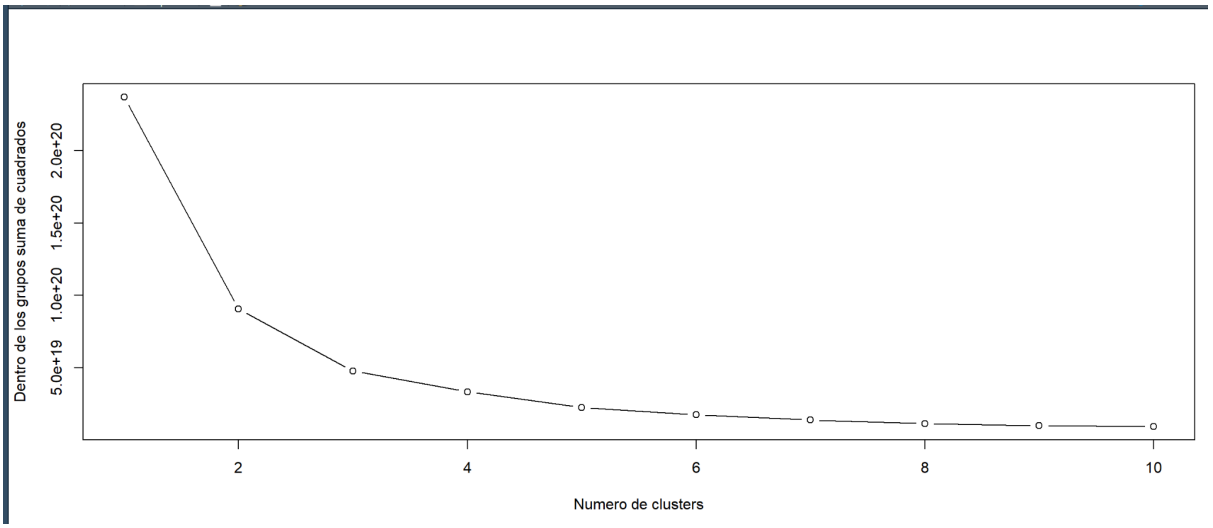
Al realizar el análisis de la tendencia al agrupamiento utilizando el estadístico de Hopkings da un valor de 0.9940868 y este valor está alejado de 0.5 por lo que los datos no son aleatorios hay altas posibilidades de que sea factible el agrupamiento.

```
library(cluster) #Para calcular la silueta
library(e1071)#para cmeans
library(mclust) #mixtures of gaussians
library(fpc) #para hacer el plotcluster
library(NbClust) #Para determinar el número de clusters óptimo
library(factoextra) #Para hacer gráficos bonitos de clustering
library(hopkins) #Para revisar si vale la pena hacer agrupamiento
library(GGally) #Para hacer el conjunto de graficos
library(FeatureImpCluster) #Para revisar la importancia de las variables
library(pheatmap) #Para hacer mapa de calor

#ejercicio2

datos<-read.csv("movies.csv")
datos<-datos[complete.cases(read.csv("movies.csv")),]
popular<-datos['popularity']
budget<-datos['budget']
revenue<-datos['revenue']
runtime<-datos['runtime']
votes<-datos['voteCount']
cl<-data.frame(popular,budget,revenue,runtime,votes)
clustering<-scale(cl)
hopkins(clustering)

datos<-dist(cl)
fviz_dist(datos, show_labels = F)
```

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.

Universidad del Valle de Guatemala
Minería de Datos

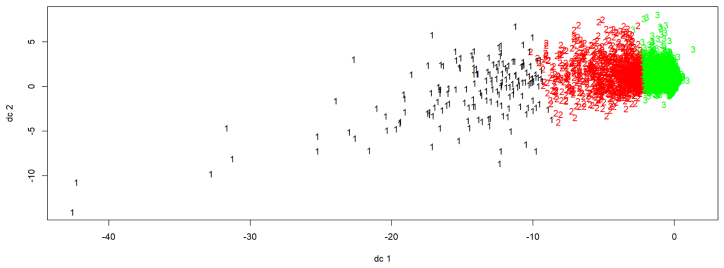
Rene Ventura #19
Andres Paíz #19
Eduardo Ramírez #19946



Podemos observar que el “codo” de la gráfica de codo está situado en 3, por lo cual ese número será la cantidad de clusters a utilizar.

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

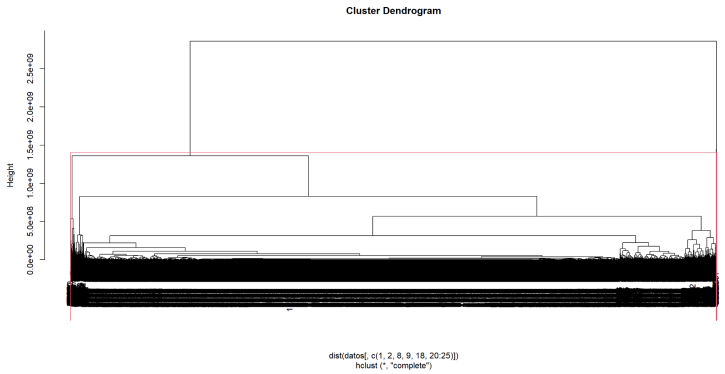
- K-means



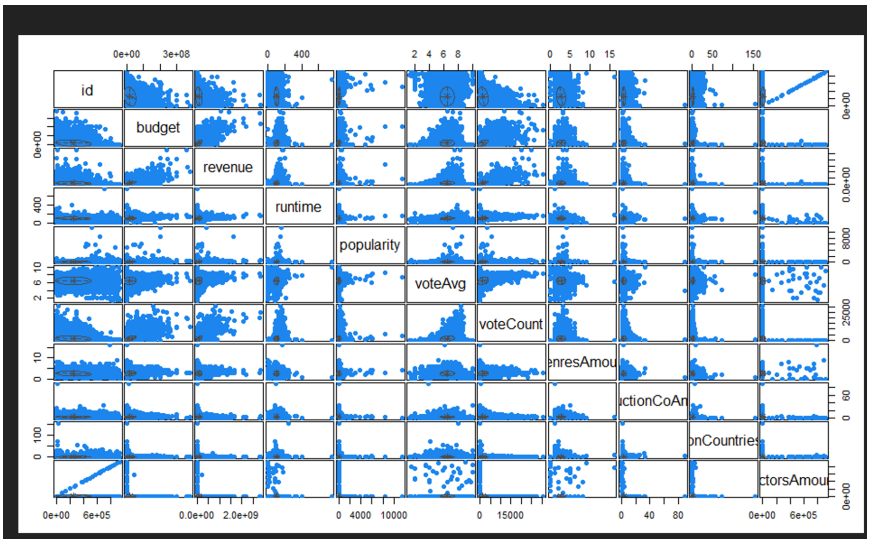
- Cluster Jerarquico

Universidad del Valle de Guatemala
Minería de Datos

Rene Ventura #19
Andres Paíz #19
Eduardo Ramírez #19946



- Gauss



El algoritmo de búsqueda k-means fue el más eficiente. A diferencia de la combinación gaussiana y el método jerárquico que se demoran minutos en obtener los resultados de clustering y luego mostrarlos gráficamente. Se puede observar que en el método jerárquico agrupó la mayoría de los los datos en un solo cluster, k-means se logró separar hasta en 3 grupos distintos y por último el método de gauss muestra los cluster probables cuando toma las dos variables del dataset que nos muestra la matriz.

Universidad del Valle de Guatemala
Minería de Datos

Rene Ventura #19
Andres Paíz #19
Eduardo Ramírez #19946

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

K means 0.8040449
Jerárquico 0.9314902
Gauss 0.3739027

```
> silkm<-silhouette(km$cluster,dist(datos[,c(1,2,8,9,18,20:25)]))
> mean(silkm[,3])
[1] 0.8040449
>
> silch<-silhouette(groups,dist(datos[,c(1,2,8,9,18,20:25)]))
> mean(silch[,3])
[1] 0.9314902
>
```

```
> mean(-silmg[,3])
[1] 0.3739027
>
```

El método de la silueta dio como resultado 0.8040449, el cual se puede decir que es un resultado cercano a 1, para el método de clustering de k means. Al ser cercano a 1 podemos afirmar que el fue eficaz. A pesar de esto, al observar el resultado del método Jerárquico podemos observar que es aún más cercano a 1 y por ende más eficaz. Finalmente el método de Gauss resulta en un número más cercano a 0 que a 1 por lo cual no es muy eficaz. Podemos afirmar que el método Jerárquico es el de mejor efectividad.

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.

Para realizar el análisis de Clusters se utilizó el método de k means, podemos observar que los datos pertenecientes a esos están basados en la magnitud de los valores de las variables numéricas. Observando el grupo 3

Universidad del Valle de Guatemala
Minería de Datos

Rene Ventura #19
Andres Paíz #19
Eduardo Ramírez #19946

podemos concluir que son los que tienen los valores más bajos. Los valores más bajos de la columna pertenecen al Cluster numero 3. Los valores están organizados de forma decreciente desde el Cluster 1 hasta el 3.