Guillermo Santos 191517
Sara Paguaga 20634
Cristian Laynez 201281

Hoja de trabajo 2 - Clustering

1. Variables a eliminar
    a. id: Porque este es el identificador de la película, este no es relevante y no servirá tomarlo en cuenta.
    b. genres: Esta variable solo dice cuál es el género de cada película y no suele aportar información relevante y no son un valor numérico que aporte.
    c. homePage: no es una variable continua porque solo provee un enlace a una página web. Sería considerada una variable categórica nominal.
    d. productionCompany, productionCompanyCountry, productionCountry: Indican información sobre la producción de la película y son variables categóricas nominales.
    e. director: Indica el nombre de un director y es considerada una variable categórica nominal.
    f. video: Únicamente es una variable booleana
    g. actors: Menciona todos los actores que participaron en la película. No aporta en las métricas de dicha película.
    h. actorsCharacters
    i. originalTitle y title: No son un valor numérico y son valores únicos por película.
    j. originalLang: Es una variable categórica, a parte, son pocas opciones. Usualmente todas tienen el valor *en*.

    - Variables que serán tomadas en cuenta:
        - budget y revenue son de importancia dado que nos permiten saber el costo de hacer una película y las ganancias obtenidas, así como popularity nos indica el índice de popularidad de la película y esta variable se ve relacionada con voteAvg la cual indica el promedio de votos de una película. Es de importancia también poder saber el valor de cantidad de géneros que representan a la película con genresAmount. Además, se toman como variable de importancia releaseYear y releaseMonth para poder obtener otros datos de tiempo de la película. Finalmente, se tomó en cuenta actorsAmount la cual puede influir en la variable budget y averageActorPopularities que indica el indice de popularidad de un actor en la película´.

Guillermo Santos 191517
Sara Paguaga 20634
Cristian Laynez 201281

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

4. Utilice los algoritmos k-medias y clustering jerárquico para agrupar. Compare los resultados generados por cada uno.
   En Knit
5. Determine la calidad del agrupamiento hecho por cada algoritmo con el método de la silueta. Discuta los resultados.
   En Knit

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.
   En Knit

7. Trabajo que sigue: Describe el trabajo que desarrollará a partir de la generación de grupos, las tendencias que investigará partiendo de lo que descubrió.

   - El trabajo que se puede realizar a partir del clustering generado es poder identificar películas con factores relevantes en común, tal como: películas que hayan tenido bajo presupuesto pero altas calificaciones e ingreso monetario en taquilla o en el caso contrario alto presupuesto pero bajas calificaciones de y bajo ingreso monetario la crítica. Por otra parte se podría identificar la influencia de la popularidad de los actores y poder analizar películas con altas calificaciones independientemente si el presupuesto de la película es alto o bajo.
   - Las tendencias que se podrían investigar son:
     - Aumento o disminución del presupuesto e ingreso monetario/taquillero a lo largo del tiempo.
     - Índice popularidad según el género a lo largo del tiempo.
     - Características de una película exitosa.