Hoja de Trabajo 2.
Clustering

INSTRUCCIONES:

Utilice el data set al que le hizo el análisis exploratorio en la hoja de trabajo anterior. Debe comparar los resultados generados por cada algoritmo de clustering. Genere un informe con el análisis del funcionamiento de los algoritmos y la interpretación de los clusters. Añada un apartado donde describa qué le pareció interesante de la información generada con el agrupamiento y de qué forma indagaría más en esa línea. Guarde el código que ha utilizado para hacer esta hoja de trabajo. Los lenguajes que tiene permitido usar son R o Python. **La calificación de cada ejercicio tomará en cuenta tanto lo escrito en el informe como el código.**

DESCRIPCIÓN DEL DATASET

El dataset contiene datos de 10866 películas obtenidos de la plataforma "The movie DB".

Variables:

- Id: Id de la película
- imdb_id: Id de la película en the movie data base (IMDB)
- popularity: Un índice de la popularidad de la película
- budget: El presupuesto para la película.
- Revenue: El ingreso de la película.
- original_title: El título original de la película.
- cast: Elenco de la película
- homepage: La página de inicio de la película
- director: Director de la película
- tagline: El eslogan de la película.
- keywords: Las palabras clave asociadas a la película.
- overview: Una breve trama de la película.
- runtime: La duración de la película.
- genres: El género de la película.
- production_companies: Las compañías productoras de la película.
- release_date: Fecha de lanzamiento de la película
- vote_count: El número de votos en la plataforma para la película.
- vote_average: El promedio de los votos en la plataforma para la película
- release_year: Año de lanzamiento de la película

ACTIVIDADES

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.
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.
3. Utilice 3 algoritmos existentes para agrupamiento. Compare los resultados generados por cada uno.
4. Determine la calidad del agrupamiento hecho por cada algoritmo con el método de la silueta. Discuta los resultados.
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.

EVALUACIÓN

- **(10 puntos)** Preprocesamiento: Elaboró el preprocesamiento de los datos necesario para utilizar los algoritmos de clustering.
- **(5 puntos)** Explicación del preprocesamiento: Explica por qué hace transformaciones en el conjunto de datos.
- **(15 puntos)** Determinación de la cantidad de grupos: Utiliza un procedimiento adecuado para determinar la cantidad de grupos que deberían formarse de acuerdo al conjunto de datos. Explica en que se basa para seleccionar el número de clústeres, interpretando los resultados del método usado. Se basa en gráficas para apoyar su decisión.
- **(15 puntos)** Clustering: Utiliza tres algoritmos de agrupamiento. Muestra el resultado generado por cada uno y lo compara.
- **(15 puntos)** Calidad del agrupamiento: Determina la calidad de los grupos arrojados por cada algoritmo. Discute los resultados y determina cuál va a usar y por qué para explorar e interpretar los grupos.
- **(30 puntos)** Interpretación de los grupos: Hace un análisis de los grupos generados. Explica los hallazgos interesantes que arrojaron. Muestra los elementos que utilizó para describir los grupos generados, medidas de tendencia central, tablas de frecuencia, etc. Explica como estos elementos ayudan a explicar los grupos.
- **(10 puntos)** Trabajo que sigue: Describe el trabajo que desarrollará a partir de la generación de grupos, las tendencias que investigará partiendo de lo que descubrió.

MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con la descripción del proceso de agrupamiento:
  - Descripción del preprocesamiento
  - Explicación de la selección del número adecuado de grupos
  - Comparación de los algoritmos de clustering, incluyendo la calidad del agrupamiento que hizo cada uno.
  - Interpretación de los grupos
  - Descripción del trabajo futuro.
- Script de R o de Python que utilizó debidamente organizado y comentado (Si utilizó jupyter notebooks o rmd debe añadir el html que se genera)
- Link de controlador de versiones utilizado.