Informe:Analisis Exploratorio de Datos

Jeferson Poma*
Departamento de Ciencia de la Computación
Universidad Nacional de San Agustín
Email: *jpomac@unsa.edu.pe

***Resumen*—El siguiente trabajo se analiza la base de datos TMDB movies del siguiente link https://www.kaggle.com/juzershakir/tmdb-movies-dataset. En la dataset se analiza como es su estructura, que valores tiene, si tiene valores faltantes, anómalos, y la relacion que tienen entre ellos. Link: https://github.com/st4rck19981/TOPDataScience/tree/main/Evaluacion1**

1. Herramientas

Para la resolución del informe, se uso la herramienta Jupyter Notebook, de Anaconda, y de el las siguientes librerias:

- PANDAS
- NUMPY
- MATPLOTLIB.PYPLOT

2. DATASET

Se pudo leer la DATASET gracias a la libreria PANDAS y su funcion .CSV. Lo primero es analizar con cuantos datos estamos trabajando.

- SHAPE: (10866, 21)

El segundo paso es identificar que columnas tenemos y con cuales vamos a trabajar (es importante entender el significado de cada columna para poder usarlas).

- id: Identificador
- imdbid: Otro identificador
- popularity: Popularidad de la película
- originaltitle: Nombre de la película
- homepage: URL de la película
- director: Director de la película
- keywords: Temas o objetos resaltantes que aparecen en la película
- genres: Uno o mas géneros de la película
- productioncompanies: Compañías que crearon la película
- releasedate: Fecha de lanzamiento
- votecount: Conteo de votos
- voteaverage: Promedio de votos
- releaseyear: Año de lanzamiento

En total tiene 21 columnas o etiquetas, de las cuales solo se utilizo 13 etiquetas. Las demás están en espera.

Después, se analizo si hay valores NAN en las columnas de las cuales se obtuvo lo siguiente: En la figura 1 vemos el conteo de valores NAN de cada columna, donde 0 significa que no hubo valores NAN. Si vemos en la columna de identificadores, a una le falta valores, y un identificador no puede tener valor NAN, por lo que habría que investigar que representa ese identificador y porque existen 10 valores NAN. Las demás columnas como 'Homepage', 'Director', 'Keywords', 'Genres', 'ProductionCompanies',no son tan necesarios, solo son necesarios si vamos a realizar consultas en base a estos campos, puesto que se descartarían las películas que no tengan estos campos.

```
In [85]: 1 data.isnull().sum()
Out[85]: id                        0
         imdb_id                  10
         popularity                0
         budget                    0
         revenue                   0
         original_title            0
         cast                     76
         homepage               7930
         director                 44
         tagline                2824
         keywords               1493
         overview                  4
         runtime                   0
         genres                   23
         production_companies   1030
         release_date              0
         vote_count                0
         vote_average              0
         release_year              0
         budget_adj                0
         revenue_adj               0
         dtype: int64
```

Figura 1: Conteo de valores NAN de cada Columna

3. Preguntas

- Que relación tiene la popularidad de las películas con respecto al tiempo?
- Que relación tiene el promedio de votos con respecto al tiempo?
- Que relación tiene los directores de las películas con respecto a su popularidad?
- Existe alguna relación del numero de votantes con respecto a la popularidad de la película?

4. Análisis

Para las posteriores análisis de datos, se gráfico cada conjunto de datos y se evaluó con respecto al tiempo y otros criterios.

4.1. Popularidad X Tiempo

En la figura 2 se hizo un gráfico de la popularidad, en el orden que estaba, pero esto no representa datos de importancia. Al momento de tomar el tiempo como nuestro eje X (columna='releasedate') nos salio que tenia películas del año 2060, y es porque al momento de convertir los datos a fechas, la fecha leída era "05/24/70ƶ PANDAS lo interpreta como si fuera del año 2070, y es un error a tener en cuenta porque estos errores pueden cambiar la interpretación que tenemos de los datos. **Rango de Popularidad: 6.500000000000001e-05 - 32.985763**

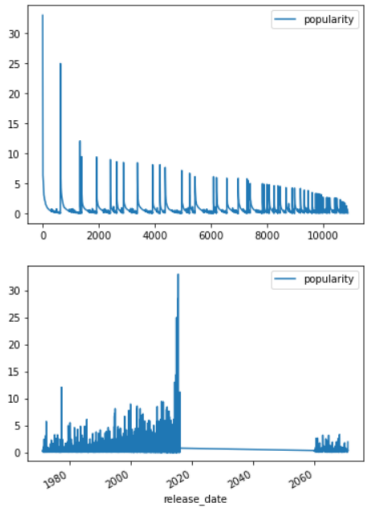

Figura 2: Popularidad X Tiempo

4.2. *Popularidad X Año*

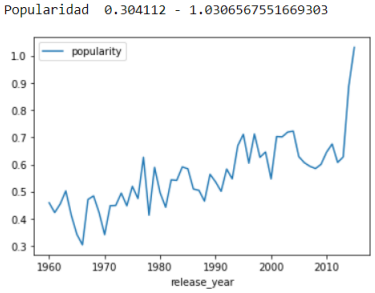

Figura 3: Popularidad X Año

En la figura 3 se agrupa los datos (las películas) por año y se hace un promedio de sus popularidades. El rango baja considerablemente (de los puntajes de 0 a 32), baja a un rango de (0.3 a 1), sin embargo, vemos el comportamiento de como la popularidad se incrementa a medida que pasa los años, y eso se debe a una era en donde las películas, son vistas por mas personas y por la calidad de ellas, la popularidad esta ligada a estos 2 conceptos.

4.3. *Promedio de Votos X Año*

En la figura **??** se agrupo por el año de lanzamiento de las películas, de las cuales se hizo un promedio del promedio de votos. Como vemos en la segunda gráfica, en anteriores años no había tanta diferencia en el promedio de votos de las películas, eso se debe a que no habían muchas personas que lo calificaran, pero ahora existe una mayor diferencia. En el primer gráfico vemos que el promedio del promedio de votos a estado bajando en estos años, uno se debe por la mayor cantidad de películas (buenas y malas) y mayores preferencias de los usuarios, pero esta mas ligara a la mayor cantidad de películas malas que existen ahora.

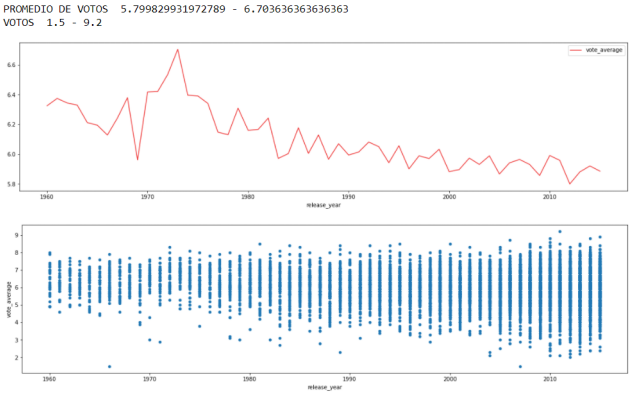

Figura 4: Promedio de Votos X Año

4.4. *Director X Popularidad*

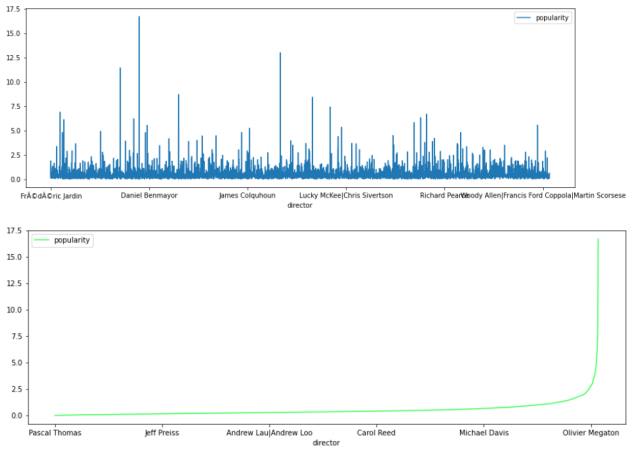

Figura 5: Director X Popularidad

Se agrupo las películas por el director que las produjo, una película no solo es hecha por 1 director, sino por varios, y eso no se toma en cuenta. Si la película esta hecha por 2 directores, este es un nuevo director. Como se ve en la figura 5, no muchas películas tienen una popularidad grande, de hecho, la mayoría su popularidad es menor que 2.5 y esto se ve mejor en el segundo gráfico. A partir de 2.5 es donde existe mayor popularidad de los directores.

4.5. *Director X Voto Promedio*

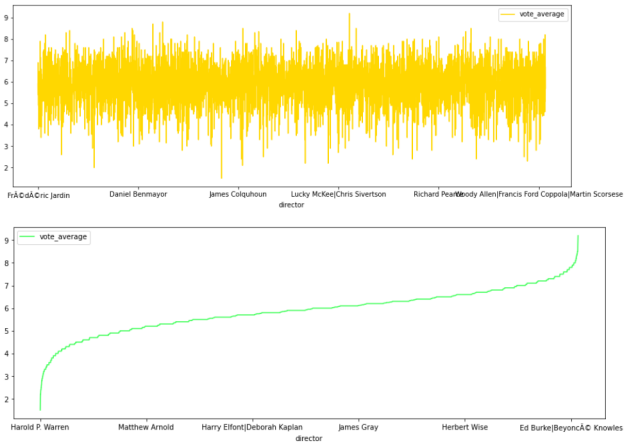

Figura 6: Director X Voto Promedio

Su gráfica es muy interesante, y es que hay mayor densidad en las partes mas extremas del gráfico. Pocos directores tiene un voto promedio malo y bueno. Y la mayoría de ellos su voto promedio esta por el valor de 5.

4.6. *Contador de Votantes X Popularidad*

Y si vemos como esta relacionado el numero de votantes con la popularidad de la película.

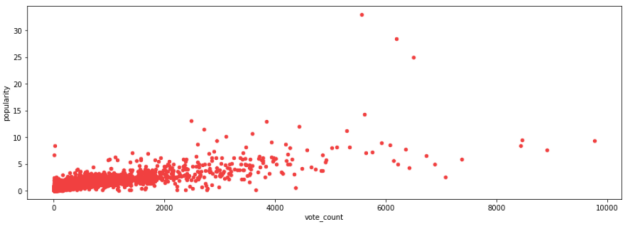

Figura 7: Contador de Votantes X Popularidad

Como vemos en la figura 7, vemos como una especie de recta, es decir, tiene una relación, y los valores que no están tan cerca a los grupos son valores raros que no se guían por esa recta. Por ejemplo, a una película que tiene una popularidad de 10, lo han calificado 10000 personas y una popularidad de 33, lo calificaron 6000 personas. Lo normal es que esta gráfica este guiada por una recta, pero vemos que no siempre es así.

5. Otras preguntas

De las preguntas que se plantearon al inicio, no se ha verificado con otros campos, seria interesante compararlos y ver que relación tienen con los otros campos. Al momento de analizar por Director, seria interesante separar a los directores y calcular en base a un solo director la popularidad que obtuvo con una de sus películas con respecto al tiempo, e igual su voto promedio con respecto al tiempo. E igual con contador de votantes X popularidad, analizar el numero de votantes con respecto al tiempo para ver si hubo un incremento de votantes ya sea por la mayor disponibilidad de internet o la popularidad de películas.