Examen 1:Análisis 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 busca responder las preguntas propuestas utilizando la base de datos TMDB movies https://www.kaggle.com/juzershakir/tmdb-movies-dataset. La ubicación del código se encontrara en el siguiente link: 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. Preguntas

- Para el género comedia, ¿Qué películas gustan más según su popularidad?
- Las películas con mayores ingresos obtienen mayor rentabilidad? (rentabilidad = revenue - budget)

3. Pregunta 1

Para la primera pregunta, el criterio de 'gustar' esta definido por el campo 'voterating', esta variable es la calificación promedio de la película dada por los votantes. Para la palabra 'popularidad', existe un campo que indica este valor.

En el anterior trabajo, se supo que la DATASET contiene valores NAN, y mas en la columna de 'Genres', entonces es necesario una limpieza de DATASET. Posteriormente, se calculan los umbrales para 'popularity' y 'voteaverage', para 'popularity' se tomo la media de la popularidad y para 'voteaverage' se toma un umbral de 7.0 (este valor se calculo en base al numero de resultados que obtuvimos, para 6.5 todavía eran muchos resultados, es por eso que incrementamos a 7.0)

En la figura 1 vemos una lista de películas delimitadas por los umbrales. Hay que tener en cuenta que aquí se toma todas las películas que tengan Comedia, o solo una parte (ejemplo: [Avatar,ComedialSuspensolAcción])

Para la figura 2, solo se toma las películas que sean pura comedia, esto hace que el numero de películas escogidas sea menor y mas aun al delimitarlo por umbrales.

En los gráficos, se muestra en el eje X los resultados, es decir, el nombre de las películas.

La visualizacion de estos datos depende totalmente de los 2 umbrales, si se desea tener mejores resultados, es necesario estudiar estos umbrales, sacar un valor para 'popularidad' un valor mayor de la media y ser un poco mas ligero en la variable de 'votoaverage'.

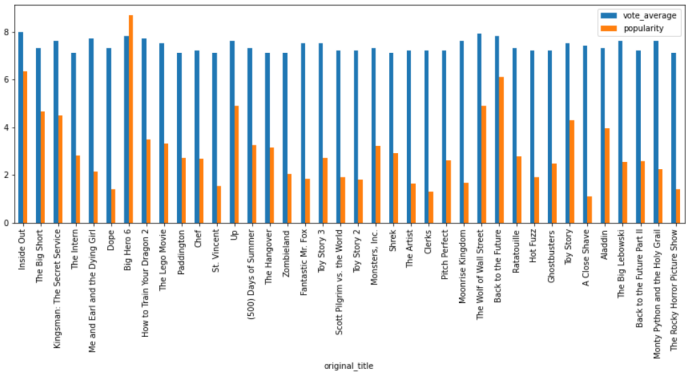

Figura 1: Title X Popularity+VoteAverage

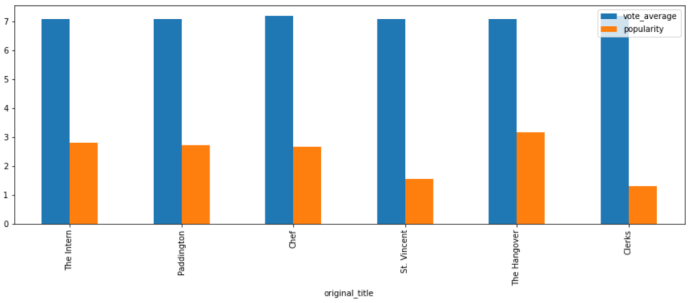

Figura 2: Title X Popularity+VoteAverage

Como tenemos pocos datos, se ha utilizado gráfico de barras para poder visualizar los datos y hacer una comparación entre estos 2 campos. Raramente, todas las películas con alta popularidad tienen una buena calificación, pues la popularidad depende del 'marketing' y lo que te enseñan por publicidad no siempre es lo que contiene la película.

4. Pregunta 2

Para la pregunta 2, se calcula una nueva columna, la rentabilidad, que es la resta de los ingresos con el presupuesto y como este se comporta.

En la figura 3, para visualizar mejor los datos, se han ordenado de mayor a menor y vemos que muy pocas películas tienen una alta rentabilidad y la mayoría de ellas es un valor constante, y también vemos muy pocas películas que poseen una rentabilidad negativa (esto significa que han fracasado en recuperar sus ingresos). La mayoría de películas, su rentabili-

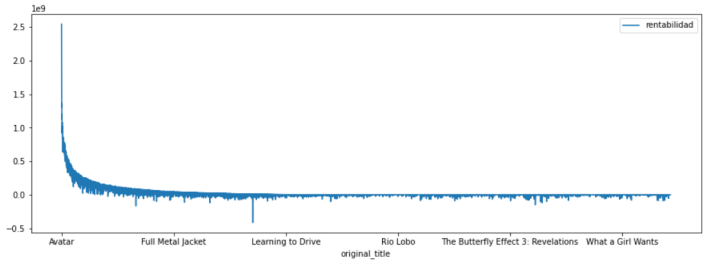

Figura 3: Title X Rentabilidad

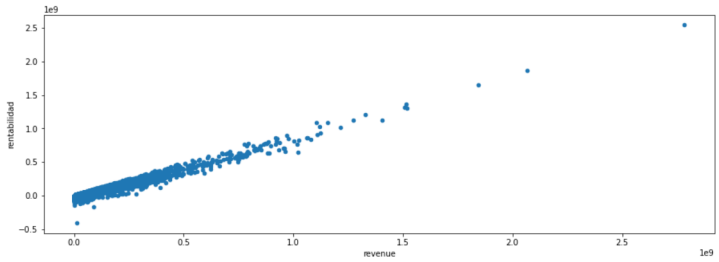

Figura 4: Ingresos X Rentabilidad

dad esta cercana a 0 y aproximadamente, solo un 5 % de ellas su rentabilidad es estable

Para esta figura, en el eje X esta los ingresos y en el eje Y la rentabilidad y vemos que si tienen una relación. Poco a poco esos puntos generar una recta casi perfecta, sin embargo, también vemos que muy pocas películas cuentan con altos ingresos y la mayoría de películas cuenta con pocos ingresos.

Como nuestros valores son números, usamos el gráfico SCATTER para su relación y ver el comportamiento de los datos.