Punto 2: Documentación y análisis de módulos.

Jaime Santiago Almeida Salazar

Repositorio Github : https://github.com/santiagoalmeida08/BOOKS_PROYECT

Modulo 1: Carga y transformaciones

El módulo 1 fue creado con el propósito de cargar, limpiar y transformar los datos. Este proceso es crucial para preparar los datos para análisis posteriores y cualquier tipo de modelo. La limpieza y transformación de datos aseguran que los datos sean consistentes, completos y adecuados para el análisis, lo que mejora la calidad y precisión de los resultados obtenidos.

Hallazgos

- Se encontró que múltiples variables tenían una gran cantidad de valores nulos
- Se pensó en eliminar la variable “price” debido a su baja representatividad por su gran cantidad de valores nulos. Sin embargo, debido al gran volumen de datos y el bajo almacenamiento que se tiene en github se decidió eliminar sus nulos y conservar la variable.
- Se eliminaron las variables con referencia a links debido que eran innecesarias y no generarían aporte al modelo
- Se transformaron y se dio forma a las variables de review_time, categoría y autores
- Se reemplazaron los nulos de las variables user_id y profilename por “anónimos”, debido a que en estos sitios es común encontrar estos comentarios o las personas cierran su cuenta y la reseña pierde identidad.

Modulo 2: Análisis exploratorio

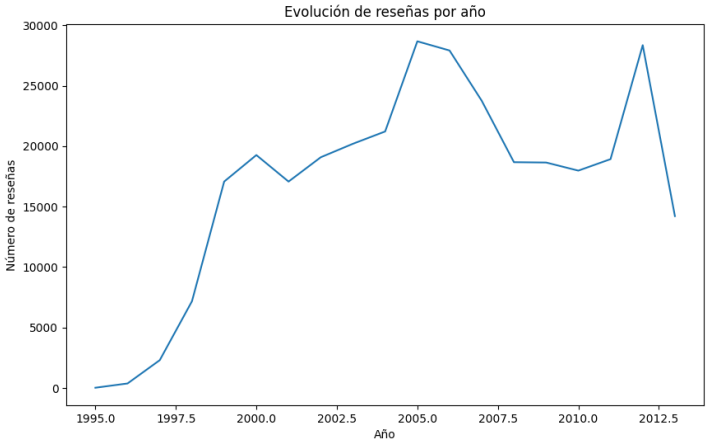
En este módulo se realiza un análisis exploratorio de las bases de datos con el fin de identificar tendencias, correlaciones o datos sesgados. Este análisis es crucial para comprender mejor los datos, identificar patrones y obtener información valiosa que puede guiar decisiones futuras en el proyecto

Hallazgos

- Al explorar mas a fondo las bases de datos se descubrió que la variable “publisheddate” contenía algunos registros con caracteres especiales, lo cual no se contemplo en el primero modulo.

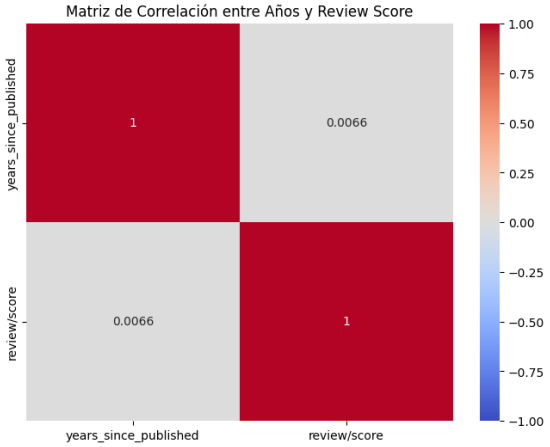
- El 75% de las personas ha escrito una reseña en amazon 10 años después de que el libro fuese publicado

- Se evidencia que el numero de reseñas ha disminuido a partir del año 2006



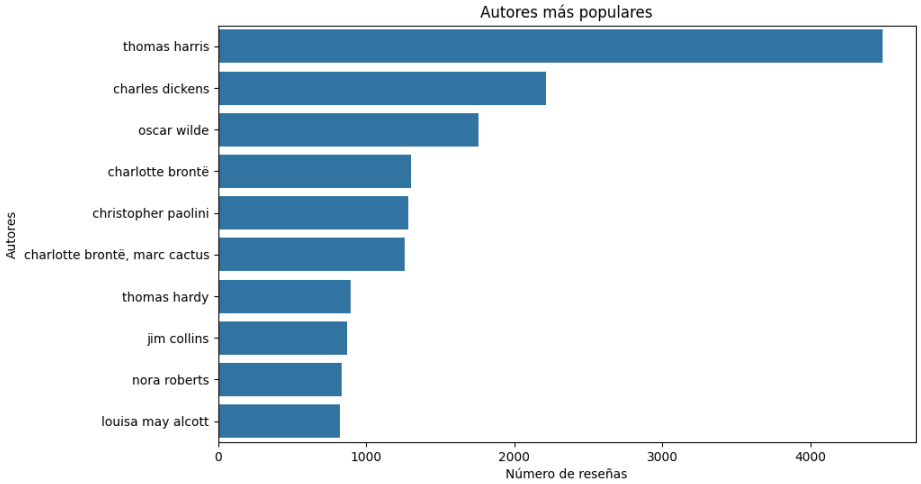
- El 50% de los libros tienen un costo menor a 10 dólares, sin embargo un 15% sobrepasa los 995 dólares

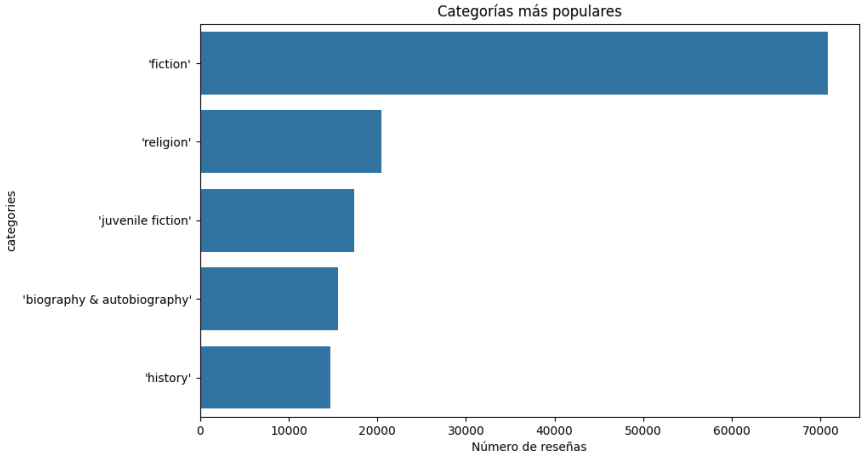
- El puntaje que los libros reciben son independientes al numero de años que pasen después de su publicación



- Se transformo la variable review/helpfulness , se separo en dos columnas con el fin de calcular el promedio de valoraciones y adaptar los datos, con ello obtendremos un puntaje en el cual, entre mas cercano sea a uno, mayor ayuda brindo la reseña los usuarios.

- Thomas Harris es el autor mas popular de la red y el genero de ficción es el mas popular ya que cuenta con mas de 70mil reseñas





- Se tuvieron 328.000 valoraciones totales

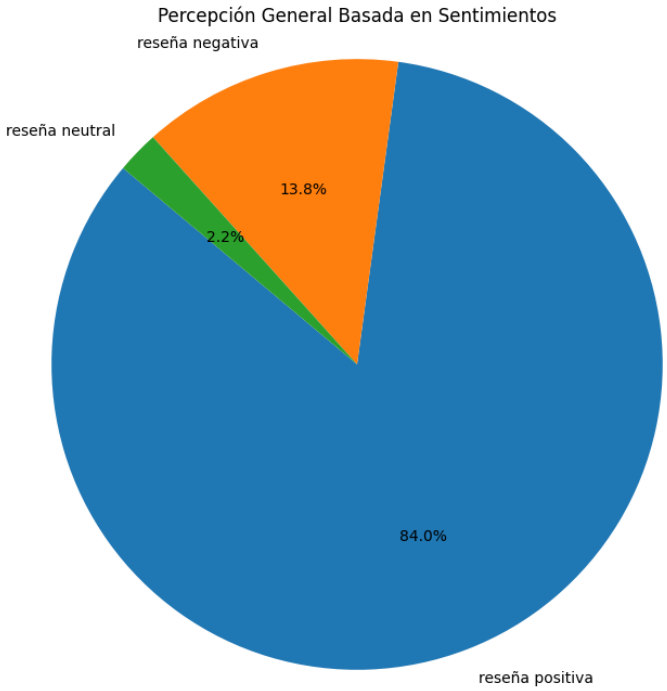
Modulo 3: Análisis de sentimientos

En este modulo se aborda un análisis de sentimientos a las reseñas que realizaron las personas para los diferentes libros. El objetivo principal del modulo es comprender de forma general, cual es la percepción que tienen los clientes obre los libros, autores y categorías. Para de esta forma mediante estas herramientas tener un acercamiento al cliente y plantear nuevos retos para mejorar la plataforma.

Hallazgos

- Al momento de abordar el análisis de sentimientos se consideraron las 3 tecnicas presentadas en la prueba (TextBlob, VADER o NLTK), se decidió trabajar con el algoritmo VADER debido a que se acoplaba perfectamente a la situación que se abordaba, pues este tiene un rendimiento optimo para textos no estructurados en inglés.

- La mayoría de las reseñas fueron positivas, seguidas por las negativas y, finalmente, las neutrales, este hallazgo sugiere que, en general, los lectores tienen una percepción positiva de los libros analizados, sin embargo, hay que buscar estrategias para disminuir el porcentaje de malas reseñas.



- Algunas categorías como judical error y hell in literature, mostraron un sentimiento promedio significativamente más alto que otras, lo que puede indicar una mayor satisfacción de los lectores con ciertos géneros

- El libro con mayor numero de reseñas positivas fue laodicean a story of today, el cual alcanzo un 0.9999 de sentimiento en la escala de compound

-

Modulo 4: Identificación de los mejores libros

Este módulo proporciona un sistema de recomendaciones para identificar los 10 mejores libros basándose en una combinación de tres métricas: número de reseñas, promedio de puntaje y sentimiento promedio. Para ello, se normalizaron las métricas y se combinaron en un puntaje único para facilitar la comparación y selección de los mejores libros. Este enfoque asegura que los libros recomendados sean populares, bien valorados y positivamente percibidos por los lectores.

Modulo 5: Automatización y ejecución