Exploración y análisis de datos sobre libros.

Descripción del proyecto

Los libros son una manera de pasar un buen rato por lo que cuando alguien le gusta un libro suele recomendarlo o posicionarlo dentro de sus favoritos, pero ocurre algo similar cuando un libro no nos gusta, decidimos darle una mala calificación o no recomendarlo y todos estos datos son utilizados para conocer qué libros les pueden interesar a otras personas.

El presente proyecto se encargará de analizar a partir de una serie de datos, cual es la relación de los libros más leídos y los libros con mejor puntuación, y comprobar si la audiencia prefiere los libros con mayor cantidad de páginas.

Fuente

El Dataset se obtuvo de Kaggle:
https://www.kaggle.com/datasets/dylanjcastillo/7k-books-with-metadata

También puedes encontrar el código en github: https://github.com/Salmailc/books_project.git

Propuesta

El desarrollo que se propone para este proyecto es, comenzar por la limpieza de los datos, analizar las variables, como es su correlación y el comportamiento que tienen, para posteriormente mostrar los resultados de lo analizado mediante gráficas.

Descripción de los datos

El Dataset está conformado por 6810 filas y 12 columnas (variables), las cuales se muestran a continuación:

Variable	Definición
isbn13	Número internacional normalizado del libro, es un número de 13 cifras
isbn10	Número internacional normalizado del libro, es un número de 10 cifras
title	Título del libro
subtitle	Subtítulo de los libro
authors	Autores del libro

categories	Categorías de libros
thumbnail	Enlace a una foto miniatura de la portada del libro
description	Descripción del libro
published_year	Año en que fue publicado el libro
average_rating	Es una medida de lo que una determinada población califica
num_pages	El número de páginas que tiene un libro
ratings_count	Es el recuento del número total de calificaciones

Las herramientas que se ocuparan

A continuación se muestran los programas que ocuparemos: numpy, pandas, seaborn, matplotlib, altair, sklearn, pandas_profiling.

EDA

Cargamos los datos de un documento en formato csv y que se encuentran previamente descargados en la pc, para esto ocuparemos pandas, *pd.read_csv()*
*Nota: Para poder cargar los datos es necesario que el ambiente de trabajo esté posicionado en la carpeta que contiene nuestro documento csv.

Para saber la cantidad de filas y columnas que tiene nuestro Dataset ocupamos *.shape.*

Utilizamos *.columns* para conocer con qué columnas cuenta nuestro Dataset.

Conoceremos la información del Dataset como lo son: el nombre de las columnas, si hay datos nulos, el tipo de variable (int, float, obj, etc.) por medio de *.info()*

Para conocer los datos estadísticos del Dataset ocupamos *.describe()* podremos observar la media, la desviación estándar, el valor mínimo, el valor máximo, el percentil 25, 50 y 75 de cada variable numérica.

En este caso se decidió cambiar los nombres de las columnas.

Se observó con anterioridad que teníamos algunos datos nulos (NaN), por lo que reemplazamos estos datos con el valor de la media en caso de los datos enteros, utilizando *.fillna()*

Para los datos que son nulos pero son de tipo objeto, se cambiaron por 0 ocupando *.fillna()*
También se cambiaron el tipo de variable en ***'Published year', 'Pages', 'Ratings count'***, ya que eran de tipo flotante y se cambiaron a entero, mediante *.astype()*

De *pandas_profiling* importamos *ProfileReport* es una herramienta que crea un formato html en donde se puede visualizar las variables y tablas de correlación.

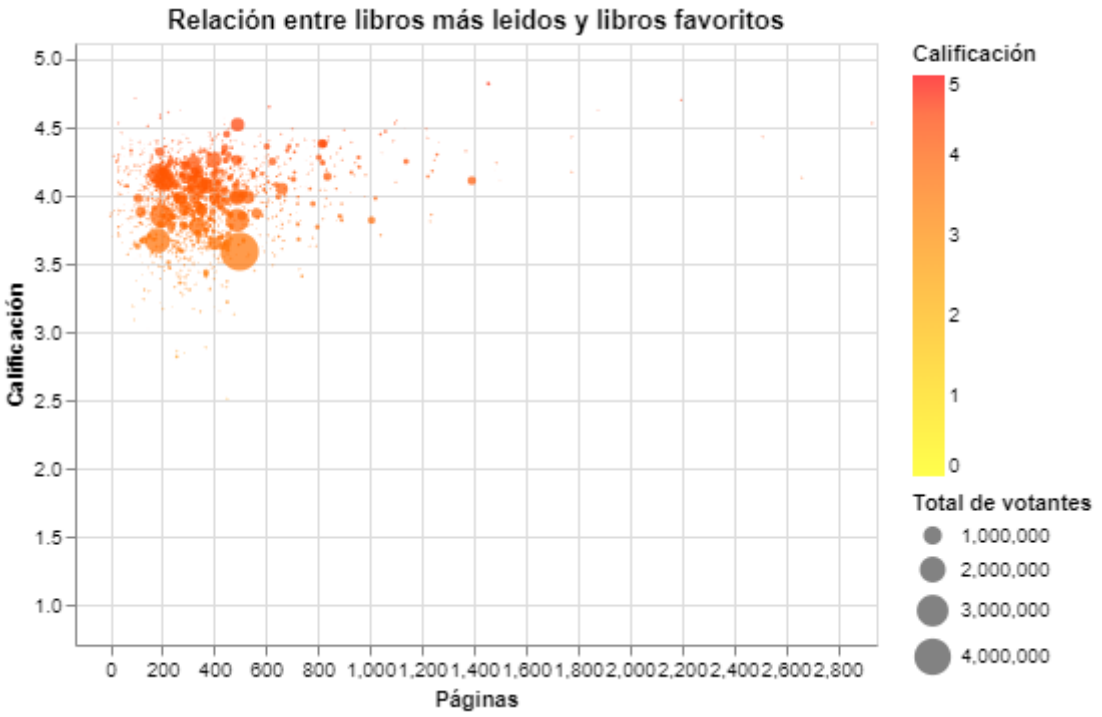
Otra manera de visualizar una tabla de correlación es utilizando seaborn *sns.pairplot()* en donde podemos colocar las variables que se quieren visualizar. Ya que se quiere saber si hay alguna influencia por parte de la cantidad de páginas, lo ocuparemos como la tonalidad.

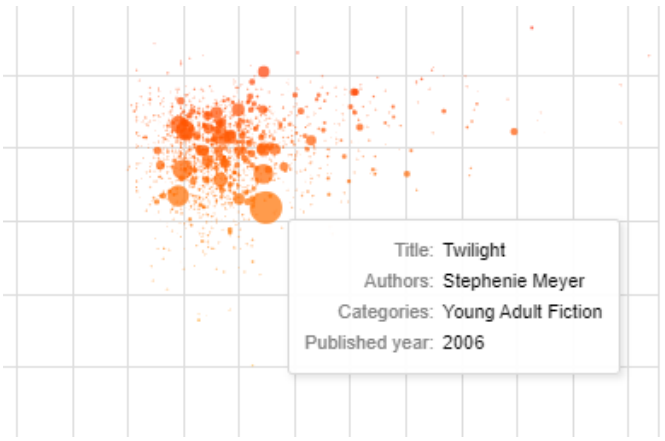
Separamos el Dataset en dos, uno para entrenamiento (train) y otro para prueba (test), utilizando *train_test_split,* en este decidí que la parte de prueba sea el 30% del total de datos.

Visualización

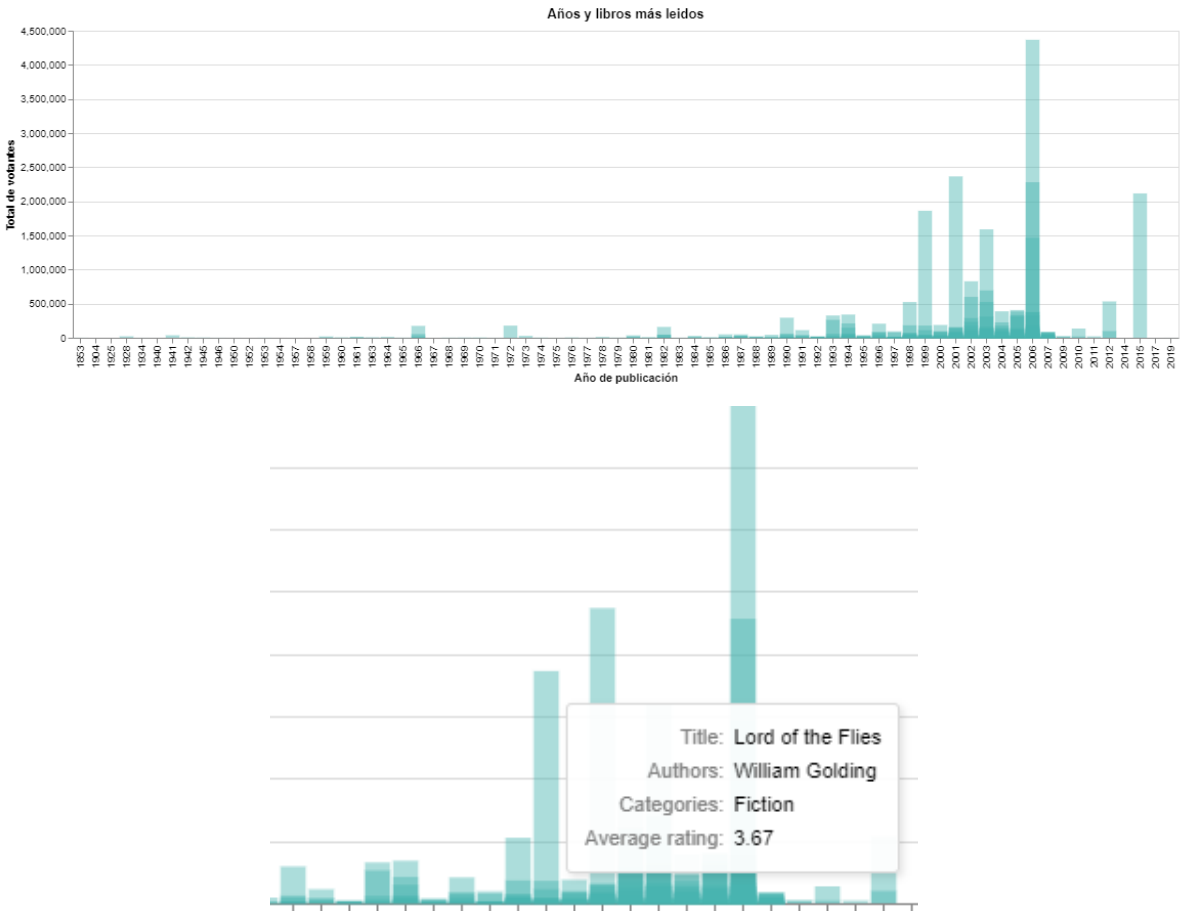
Utilizamos altair para crear una gráfica interactiva que nos pueda mostrar la relación que tiene la calificación que se le da a un libro (eje Y y escala de color), el número de páginas del libro (eje X) y el total de votantes de cada libro (tamaño de las esferas en la gráfica). Esto con la finalidad de comprender de manera visual si se tiene alguna interacción entre estas variables.
Al momento de posicionar el mouse en alguna de las esferas se mostrará una ventana con el título, el autor, la categoría y el año de publicación del libro, para que sea de conocimiento a cual libro pertenece dicha esfera.

*Nota: en el caso de la visualización ocuparemos los datos de prueba.





Utilizaremos otra gráfica de barras de altair para visualizar en qué años se leyeron los libros (eje X) y su cantidad de votantes (eje Y), ocupando la opacidad para diferenciar en las barras a los libros ya que muchos libros pertenecen al mismo año de publicación. Al igual que en la gráfica anterior al momento de posicionarse en una barra saldrá una ventana con el título, el autor, la categoría y la calificación que se les otorgó.



Modelado

Previamente ya habíamos dividido el Dataset en dos partes, un 70% para entrenamiento y un 30% para prueba.
Ahora observaremos la información de ambas partes con *.info()*, así como la información estadística con *.describe()*

Conclusiones

Se pudo analizar los datos con diferentes gráficas y se observó la relación que existe entre los libros más leídos y los libros con mejor puntuación, ya que los libros con mayor audiencia no siempre fueron los que tenían una mejor puntuación, por lo que podríamos concluir que aunque la gente elija leer libros no siempre terminan siendo de su agrado y en cambio hay libros muy buenos pero no son tan populares entre los lectores.

Por otra parte se pudo descubrir que la audiencia prefiere los libros con menor cantidad de páginas en comparación con otros libros con mayor número de páginas.

En la gráfica de barras titulada “Años y libros más leídos”, se noto que los libros con mayor cantidad de lectores comenzaron a partir del año 1996 al 2019, pero debemos considerar que el año máximo de nuestro Dataset es 2019.

*Nota:
Los planes que se tienen para el Dataset en un futuro es crear un modelo que ayude a recomendar libros, dependiendo de las preferencias literarias del usuario, es una opción para cuando no sabes qué otro libro puedes leer, o con qué libro empezar un buen pasatiempo o hábito.