Nombre: Jorge Guillermo Cañas Magaña

Actividad 1: Proyecto de Python

Dataset utilizado: *goodreadsbooks*

Introducción:

La selección del dataset se debe a mi interés por la lectura, actividad y hábito que considero de mucho beneficio. El archivo consta de 11123 filas y 12 columnas de información de libros, autores, ratings de lectores y campos identificadores.

Al realizar un análisis exploratorio de la data, se encontró que la mayoría de los libros están bien calificados por los lectores. Además, más del 70% de los libros han sido calificados por al menos 200 personas, es decir, se tiene la opinión de un buen número de lectores para intuir que la calidad de los libros en el data set es buena. Adicionalmente, el tipo de libro en general es de más de 200 páginas, es decir, son poco los libros que podríamos catalogar como cortos.

Null values:

Al revisar el archivo YData Profiling Report y, a través de varias revisiones efectuadas en el notebook, se comprobó que el dataset no tiene valores nulos, esto probablemente se deba a que el creador ya efectuó el trabajo de reemplazar y/o limpiar estos valores previo a cargar en kaggle.

Sin embargo, es de mencionar si en un dado caso me hubiera encontrado con columnas que contenían valores nulos, habría que evaluar tres puntos para proceder con la limpieza/ reemplazo:

1. Cantidad de valores nulos en la columna.
   Si son pocas filas, ente 0.1% y 10% de los datos, se podían eliminar las filas, considerando que el dataset cuenta con más de 11100 filas.
   Si, por otro lado, los datos nulos ascendían a más de 10% y hasta el 49% sería necesario evaluar el tipo de dato para proceder con alguna técnica de reemplazo.
2. Tipo de dato identificado con df.info():
   Para datos del tipo categórico es posible reemplazar los nulos con el valor más común dentro de la columna (moda).
   Ejemplo del código*: moda = df[average_rating].mode()[0]*
   df_books['average_rating'].replace(np.nan, moda)

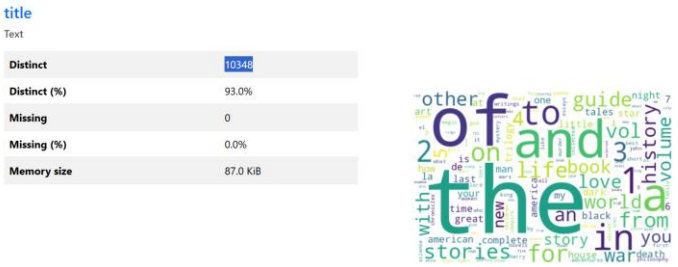
Para datos del tipo numérico, la media, mediana, moda o el valor cero son las opciones para reemplazar.

3. Definir para que se utilizará el data set:
   Si este se utilizará para algún algoritmo de recomendación, es necesario tener un dataset lo más completo posible, esto es crítico para las columnas como autores o casas editoriales, con el fin de evitar la recomendación de libros en base a parámetros erroneos errónea.

A continuación, se expone una breve revisión de algunos campos deldata set:

title:

Interesante notar que hay libros con nombres repetidos como podemos observar en esta imagen tomada de YData Profiling Report.



Considerando lo anterior, se ejecutó el siguiente segmento de código para identificar si esto se debe a libros con el mismo nombre y distinto autor o, definitivamente se tienen libros duplicados.

```
[65] #Se identifican combinaciones de título y autor repetidas dentro del data set
     df_books.value_counts(['title', 'authors'])[lambda x: x > 1]
```

El resultado es si, hay una pequeña proporción de registros de libros duplicados.

		count
title	authors	
Memoirs of a Geisha	Arthur Golden	5
One Hundred Years of Solitude	Gabriel GarcÃa MÃ¡rquez/Gregory Rabassa	5
'Salem's Lot	Stephen King	5
Sula	Toni Morrison	4
White Teeth	Zadie Smith	4
...	...	...
The Wonderful Story of Henry Sugar and Six More	Roald Dahl	2
Human Traces	Sebastian Faulks	2
Haruki Murakami and the Music of Words	Jay Rubin	2
I Don't Know How She Does It (Kate Reddy #1)	Allison Pearson	2
Bargaining for Advantage: Negotiation Strategies for Reasonable People	G. Richard Shell	2

263 rows × 1 columns

dtype: int64

Language_code:

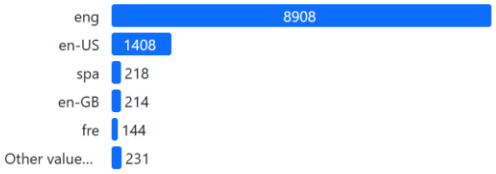
Revisando la columna language_code, se encuentra que más del 90% de los libros están escritos en idioma inglés, por lo tanto, es una base un tanto limitada para aquellos lectores que no sean conocedores del idioma.

language_code

Categorical

Imbalance

Distinct	27
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	87.0 KiB



Author:

Al revisar la columna autor se observa con claridad que hay muchos autores que tienen más de un libro en el data set.

authors

Text

Distinct	6639
Distinct (%)	59.7%
Missing	0
Missing (%)	0.0%
Memory size	87.0 KiB

Al ejecutar la siguiente línea de código se pudo obtener el top de autores con más número de libros en el data set:

```
[70] #Se identifican autores con más libros en la base
     df_books.value_counts('authors')[lambda x: x > 1][:5]
```

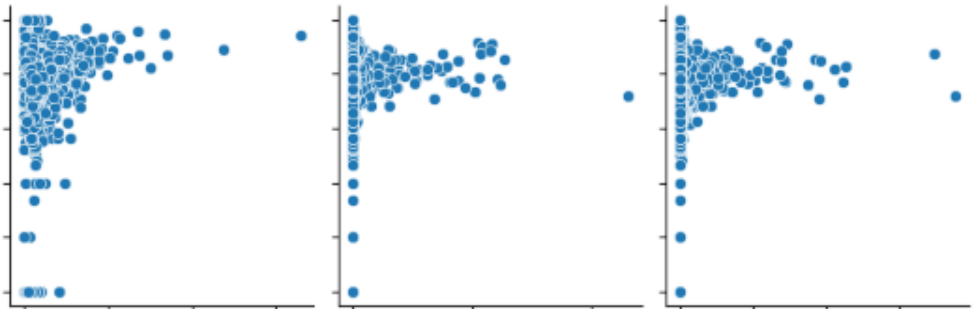
authors	count
P.G. Wodehouse	40
Stephen King	40
Rumiko Takahashi	39
Orson Scott Card	35
Agatha Christie	33

dtype: int64

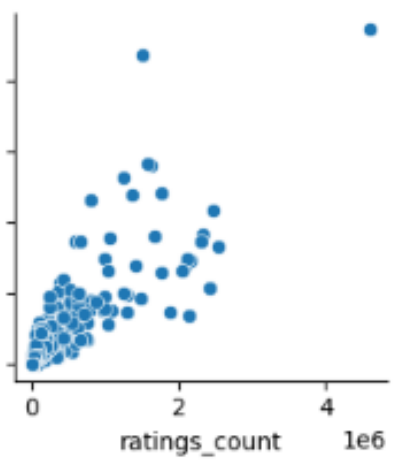
Por lo anterior, los fanáticos de estos autores estarán complacidos con los títulos que a partir de este data set se les pueda recomendar.

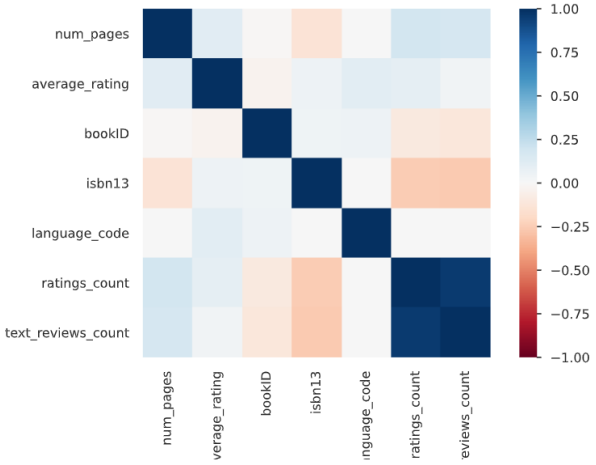
Conclusiones:

- Se cuenta con un data set de libros bien calificados y con muchos ratings de lectores que justifican las altas calificaciones.
- La disposición de los valores de los diagramas de dispersión de average_ratings vs num_pages, ratings_count y text_reviews_counts es bastante similar, lo que puede representar cierta correlación de variables.



- Las columnas rating_counts y reviews_counts se encuentra altamente correlacionadas positivamente, a mayor rating_counts mayores valores de reviews_counts. Para análisis más avanzados se recomienda descartar una de las columnas para evitar análisis redundantes.





- Sería interesante contar con una columna de genero/tipo de lectura, para proponer un sistema de recomendación estilo Netflix pero adaptado a libros, que no solo este basado en el autor o en la editorial.