

Exploración y análisis de datos sobre libros.



Proyecto final

Bootcamp de Ciencia de Datos

Nombre:

Ing. Salma Itamar Lozada Carrasco



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:

- Cuál es la relación de los libros más leídos y los libros con mejor puntuación.
- Comprobar si la audiencia prefiere los libros con mayor cantidad de páginas.

Propuesta

El desarrollo que se propone para este proyecto es:



Limpieza de los datos.



Analizar las variables.



Como es la correlación y el comportamiento que tienen las variables.









Mostrar los resultados de lo analizado mediante gráficas



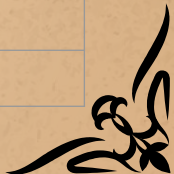
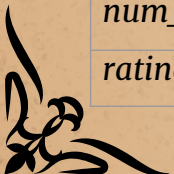
Descripción de los datos

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






Variable	Definición
<i>isbn13</i>	Número internacional normalizado del libro, es un número de 13 cifras
<i>isbn10</i>	Número internacional normalizado del libro, es un número de 10 cifras
<i>title</i>	Título del libro
<i>subtitle</i>	Subtítulo de los libro
<i>authors</i>	Autores del libro
<i>categories</i>	Categorías de libros
<i>thumbnail</i>	Enlace a una foto miniatura de la portada del libro
<i>description</i>	Descripción del libro
<i>published_year</i>	Año en que fue publicado el libro
<i>average_rating</i>	Es una medida de lo que una determinada población califica
<i>num_pages</i>	El número de páginas que tiene un libro
<i>ratings_count</i>	Es el recuento del número total de calificaciones






Herramientas



numpy
pandas
seaborn
matplotlib
altair
sklearn
pandas_profiling.



EDA

Se reemplazaron los datos nulos por el valor de la media, en el caso de los datos de tipo entero, para los de tipo objeto se cambiaron a "o". Se modificaron los nombres de las columnas y en el caso de 'Published year', 'Pages', 'Ratings count', ya que eran de tipo flotante y se cambiaron a entero

	ISBN_13	ISBN_10	Title	Subtitle	Authors	Categories	Thumbnail	Description	Published year	Average rating	Pages	Ratings count
0	9780002005883	0002005883	Gilead	0	Marilynne Robinson	Fiction	http://books.google.com/books/content?id=KQZCP...	A NOVEL THAT READERS and critics have been eag...	2004	3.85	247	361
1	9780002261982	0002261987	Spider's Web	A Novel	Charles Osborne; Agatha Christie	Detective and mystery stories	http://books.google.com/books/content?id=gA5GP...	A new 'Christie for Christmas' - a full-length...	2000	3.83	241	5164
2	9780006163831	0006163831	The One Tree	0	Stephen R. Donaldson	American fiction	http://books.google.com/books/content?id=OmQaw...	Volume Two of Stephen Donaldson's acclaimed se...	1982	3.97	479	172

Información estadística: podemos observar la media, la desviación estándar, el valor mínimo, el valor máximo, el percentil 25, 50 y 75 de cada variable numérica.

	ISBN_13	Published year	Average rating	Pages	Ratings count
count	6.810000e+03	6810.000000	6810.000000	6810.000000	6.810000e+03
mean	9.780677e+12	1998.629809	3.933284	348.179883	2.106910e+04
std	6.068911e+08	10.479654	0.330304	241.610246	1.371854e+05
min	9.780002e+12	1853.000000	0.000000	0.000000	0.000000e+00
25%	9.780330e+12	1996.000000	3.770000	208.000000	1.610000e+02
50%	9.780553e+12	2002.000000	3.950000	304.000000	1.041000e+03
75%	9.780810e+12	2005.000000	4.130000	418.000000	6.217250e+03
max	9.789042e+12	2019.000000	5.000000	3342.000000	5.629932e+06

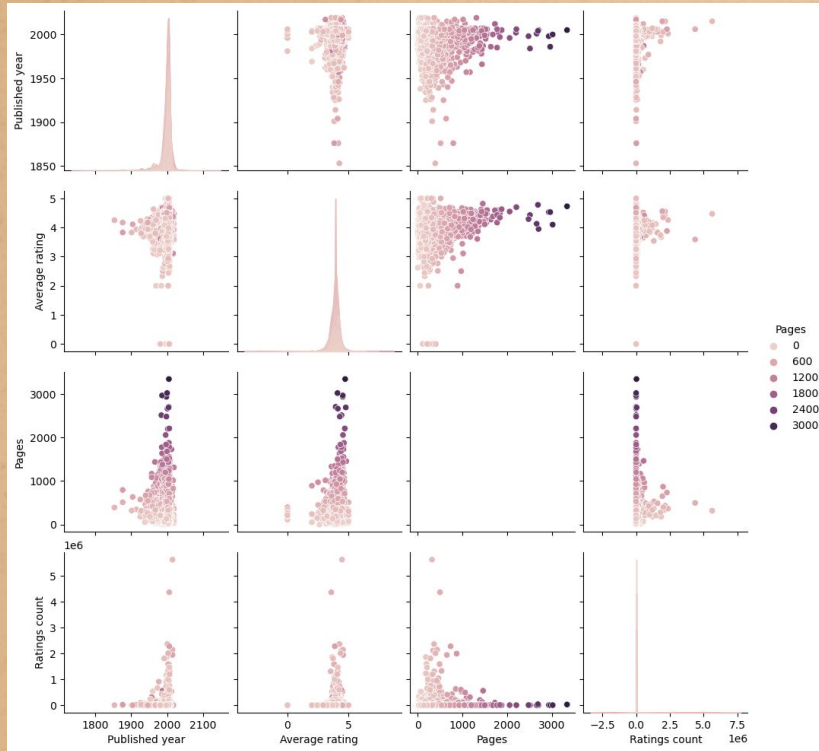
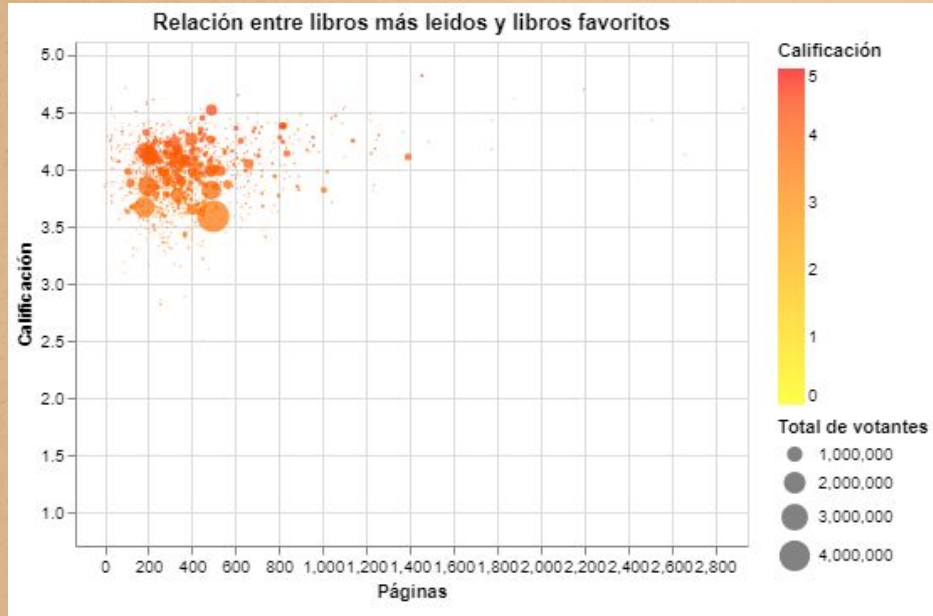


Tabla de correlación:
Utilizando `seaborn sns.pairplot()`,
en donde podemos visualizar.

Ya que se quiere saber si hay
alguna influencia por parte de la
cantidad de páginas, lo
ocuparemos como la tonalidad.

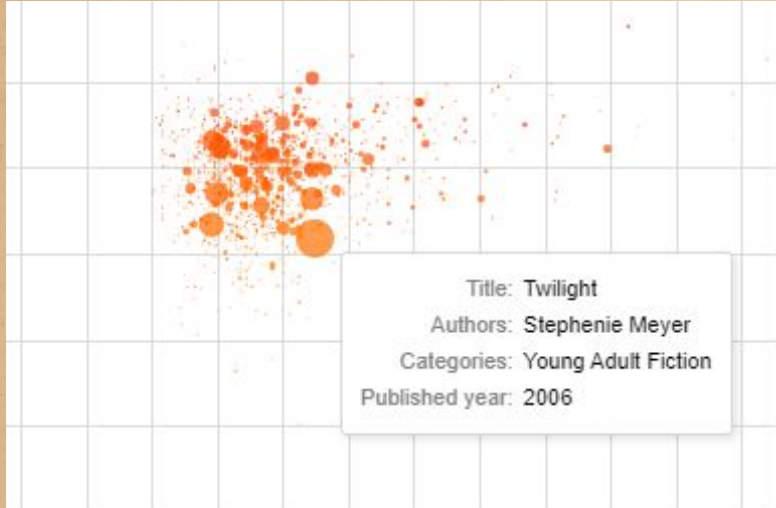
Lo que se puede observar es que
hay una relación más marcada
entre las calificaciones y el
número de páginas del libro.

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, el número de páginas del libro y el total de votantes de cada libro.

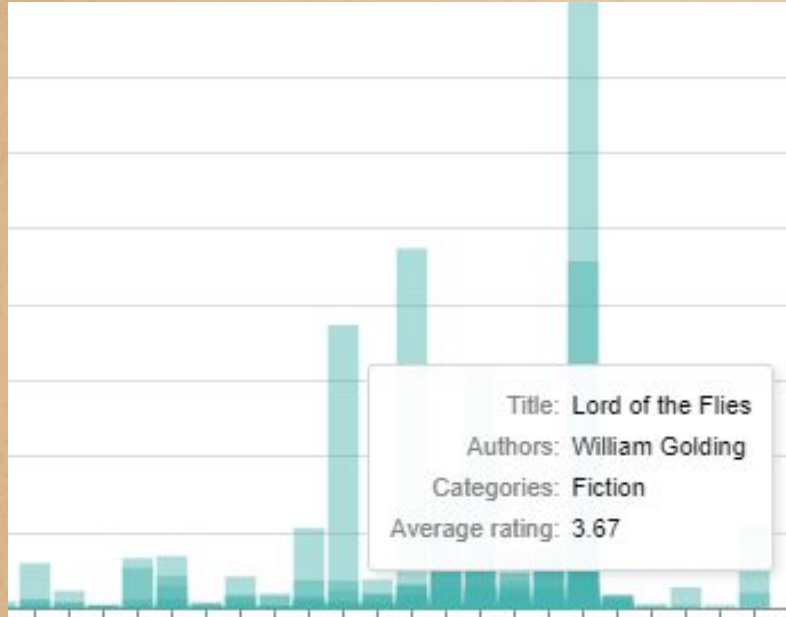
*Nota: en el caso de la visualización ocuparemos los datos de prueba.



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.

Gráfica de barras de altair para visualizar en qué años se leyeron los libros y su cantidad de votantes, 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.





Conclusiones

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.





Gracias

Fuente:

El Dataset se obtuvo de Kaggle:

[https://www.kaggle.com/datasets/dylanjcastillo/7k-books-with-metadata](https://www.kaggle.com/dylanjcastillo/7k-books-with-metadata)

También puedes encontrar el código en github:

https://github.com/Salmailc/books_project.git

