Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus, Querétaro

Herramientas computacionales: el arte de la analítica (Gpo 201)

Actividad: Visualización de datos

Estudiantes:

Karen Cebreros López A01704254

Profesor:
Pedro Pérez

Fecha de entrega:
Miércoles 11 de mayo del 2022

IMPORTAR LIBRERÍAS Y SUBIR EL ARCHIVO "bestsellers with categories.csv":

```
[1] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[2] from google.colab import files

    uploaded = files.upload()

    for fn in uploaded.keys():
      print('User uploaded file "{name}" with length {length} bytes'.format(
          name=fn, length=len(uploaded[fn])))
```

Choose Files bestsellers ...tegories.csv

- **bestsellers with categories.csv**(text/csv) - 51161 bytes, last modified: 5/8/2022 - 100% done

```
Saving bestsellers with categories.csv to bestsellers with categories.csv
User uploaded file "bestsellers with categories.csv" with length 51161 bytes
```

```
[3] df_bwc = pd.read_csv('bestsellers with categories.csv')
    df_bwc.head(5)
```

	Name	Author	User Rating	Reviews	Price	Year	Genre
0	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction
1	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction
2	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction
3	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction
4	5,000 Awesome Facts (About Everything!) (Natio...	National Geographic Kids	4.8	7665	12	2019	Non Fiction

---

Crea una tabla resumen con los estadísticas generales de las variables numéricas

```
[4] df_bwc.describe()
```

	User Rating	Reviews	Price	Year
count	550.000000	550.000000	550.000000	550.000000
mean	4.618364	11953.281818	13.100000	2014.000000
std	0.226980	11731.132017	10.842262	3.165156
min	3.300000	37.000000	0.000000	2009.000000
25%	4.500000	4058.000000	7.000000	2011.000000
50%	4.700000	8580.000000	11.000000	2014.000000
75%	4.800000	17253.250000	16.000000	2017.000000
max	4.900000	87841.000000	105.000000	2019.000000

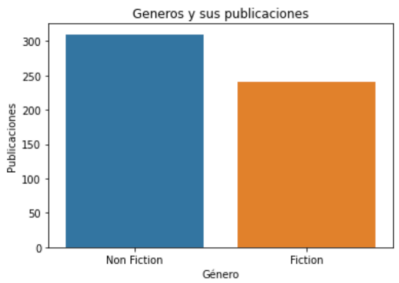
¿Cuál es el género con más publicaciones? Muéstralo en un gráfico

```
[5] fig = plt.figure(figsize=(6,4))

    sns.countplot(data=df_bwc, x = 'Genre')

    plt.title('Generos y sus publicaciones')
    plt.xlabel('Género')
    plt.ylabel('Publicaciones')
```

```
Text(0, 0.5, 'Publicaciones')
```



¿Cuántos libros del top 50 se publicaron por género en cada año? ¿Hay algún año donde hubo más libros de ficción en el top 50?. Muéstralo en un gráfico.

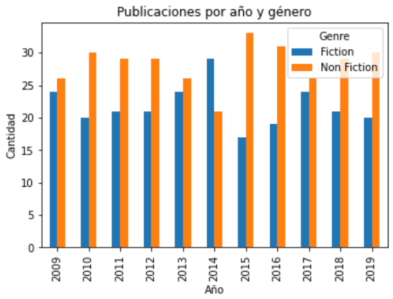
```
[10] df_yg = pd.crosstab(df_bwc['Year'], df_bwc['Genre'])
     df_yg
```

Genre	Fiction	Non Fiction
Year		
2009	24	26
2010	20	30
2011	21	29
2012	21	29
2013	24	26
2014	29	21
2015	17	33
2016	19	31
2017	24	26
2018	21	29
2019	20	30

```
[14] df_yg.plot(kind = 'bar')

     plt.title('Publicaciones por año y género')
     plt.xlabel('Año')
     plt.ylabel('Cantidad')
```

```
Text(0, 0.5, 'Cantidad')
```



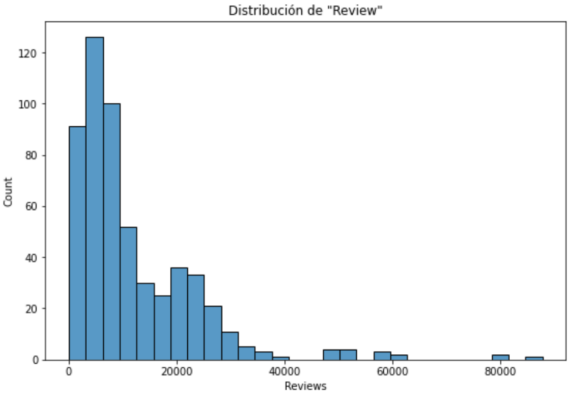
¿Cómo se distribuye la variable Review? Muéstra el histografa

```
[35] fig = plt.figure(figsize=(9, 6))

     sns.histplot(data=df_bwc, x='Reviews')

     plt.title('Distribución de "Review" - Histograma')
     plt.xlabel('Reviews')
```

```
Text(0.5, 0, 'Reviews')
```



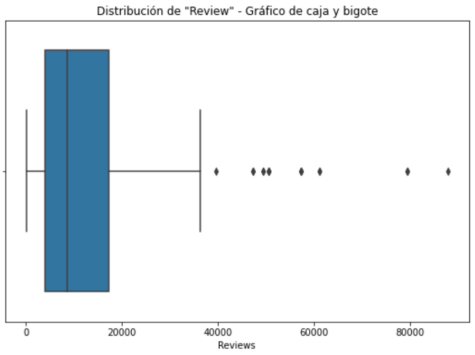
Ahora muéstralo en un gráfico de caja y bigote

```
[36] fig = plt.figure(figsize=(9, 6))

     sns.boxplot(data=df_bwc, x='Reviews')

     plt.title('Distribución de "Review" - Gráfico de caja y bigote')
     plt.xlabel('Reviews')
```

```
Text(0.5, 0, 'Reviews')
```



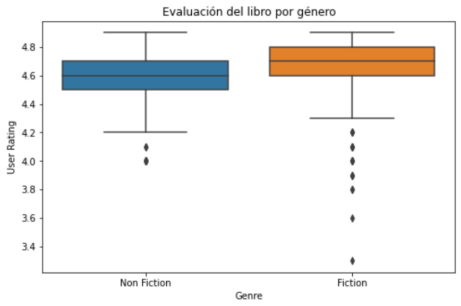
¿Cómo se compara la evaluación del libro por género? ¿Qué genero es mejor evaluado por los lectores? Muéstralo en un solo gráfico de caja y bigote

```
fig = plt.figure(figsize=(8,5))

sns.boxplot(data=df_bwc, x = 'Genre', y = 'User Rating')

plt.title('Evaluación del libro por género')
plt.xlabel('Genre')
plt.ylabel('User Rating')
```

```
Text(0, 0.5, 'User Rating')
```



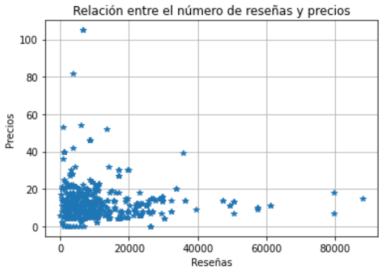
¿Cuál es la relación entre el número de reseñas y precios? Muéstralo en un gráfico de dispersión

```
[19] fig = plt.figure(figsize=(6,4))

     plt.plot(df_bwc['Reviews'], df_bwc['Price'], '*')

     plt.title('Relación entre el número de reseñas y precios')
     plt.xlabel('Reseñas')
     plt.ylabel('Precios')

     plt.grid(True)
```



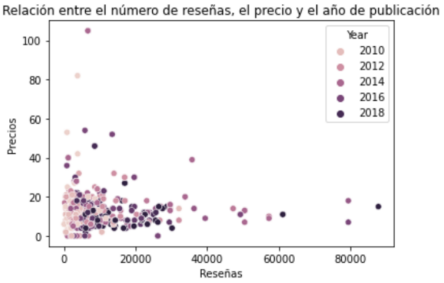
De la pregunta anterior, ¿influye algo el año de publicación? ¿Cuál es la relación entre el número de reseñas, el precio y el año de publicación?
IMPORTANTE: Selecciona una paleta de colores adeacuada

```
[22] fig = plt.figure(figsize=(6, 4))

     sns.scatterplot(data=df_bwc, x ='Reviews', y='Price', hue='Year')

     plt.title('Relación entre el número de reseñas, el precio y el año de publicación')
     plt.xlabel('Reseñas')
     plt.ylabel('Precios')
```

```
Text(0, 0.5, 'Precios')
```



¿Cuál es la correlación entre las variables numéricas? Muéstralo en un gráfico. La variable año, a pesar de ser numérica, la vamos a considerar como cualitativa, así que la eliminaremos del análisis

```
[29] df_temp = pd.DataFrame(df_bwc, columns = ['User Rating','Reviews','Price'])
     df_temp
```

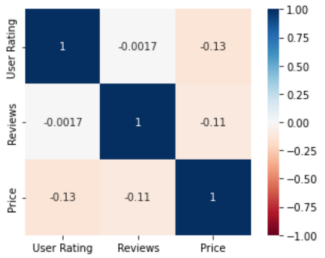
	User Rating	Reviews	Price
0	4.7	17350	8
1	4.6	2052	22
2	4.7	18979	15
3	4.7	21424	6
4	4.8	7665	12
...	...	...	...
545	4.9	9413	8
546	4.7	14331	8
547	4.7	14331	8
548	4.7	14331	8
549	4.7	14331	8

550 rows × 3 columns

```
[31] iris_corr = df_temp.corr()

     sns.heatmap(data=iris_corr, vmin=-1, vmax=1, cmap = 'RdBu', annot=True, square = True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7bb9829150>
```



¿Cuáles variables tiene una fuerte relación positiva entre sí y cuáles tienen una fuerte relación negativa? (Esta pregunta no es de código)
Responde la pregunta en la siguiente celda de texto:

- Las variables mismas entre sí, evidentemente tienen una correlación de 1 (positiva), mientras que entre "User Rating" y "Reviews", hay una correñación de 0. Y por otra parte, las correlaciones de "User Rating - Price" y "Reviews - Price" son menores a 0, por lo que se consideran como correlaciones negativas.

Haz una gráfica donde podemos comparar la relación entre las tres variables numéricas (User Rating, Reviews y Price) y que, además, podamos ver el efecto del libro. La variable año, a pesar de ser numérica, la vamos a considerar como cualitativa, así que la eliminaremos del análisis

```
[33] fig = plt.figure(figsize=(6, 4))

     sns.scatterplot(data = df_bwc, x = 'User Rating', y = 'Reviews', hue = 'Price')

     plt.title('Relación entre la calificación del usuario, el número de reseñas y el precio')
     plt.xlabel('User Rating')
     plt.ylabel('Reviews')
```

```
Text(0, 0.5, 'Reviews')
```