Nombre: Aldo Tena García

Matrícula: A01275222

```
# Carga las librerías necesarias.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Carga el conjunto de datos al ambiente de Google Colab y muestra los primeros
# 6 renglones.
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])))
```

Elegir archivos bestsellers…egories.csv

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

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

```
df = pd.read_csv('bestsellers with categories.csv')
df.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

Name: Nombre del libro.

Author: Autor.

Se ha guardado correctamente

los usuarios asignaron al libro (1-5).

Reviews: Número de reseñas.

Price: Precio del libro.

Year: Año de publicación.

Genre: Género literario (ficción/no ficción).

```
# Crea una tabla resumen con los estadísticas generales de las variables
# numéricas.
df2 = df.describe()
df2
```

	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

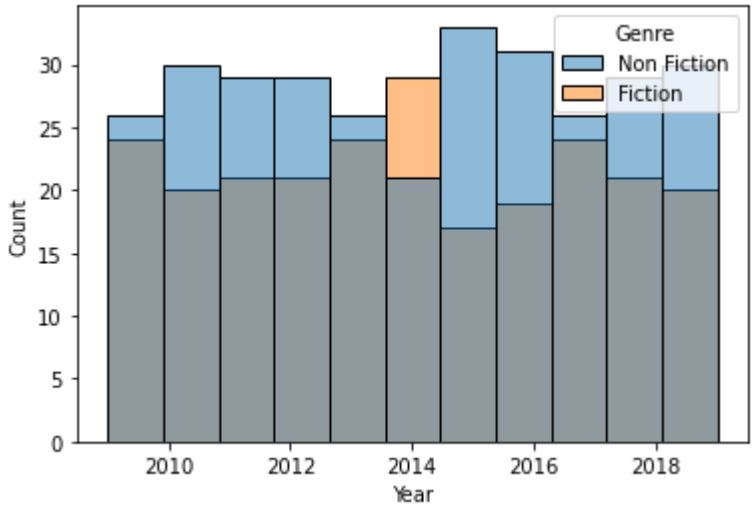
```
## ¿Cuál es el género con más publicaciones? Muéstralo en un gráfico.
sns.histplot(data=df, x='Genre')
```

Se ha guardado correctamente

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8a047fe8d0>
```

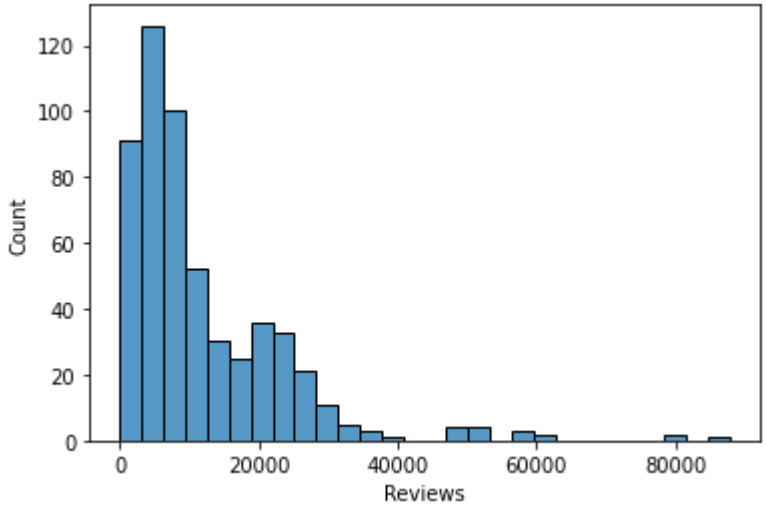
```
# ¿Cuántos libros del top 50 se publicaron por género en cada año? ¿Hay algún
# año donde hubo más libros de ficción en el top 50?. Muéstralo en un gráfico.
sns.histplot(data=df, x='Year', hue='Genre')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8a0474e750>
```




```
# ¿Cómo se distribuye la variable Review? Muéstra el histografa.
sns.histplot(data=df, x='Reviews')
```

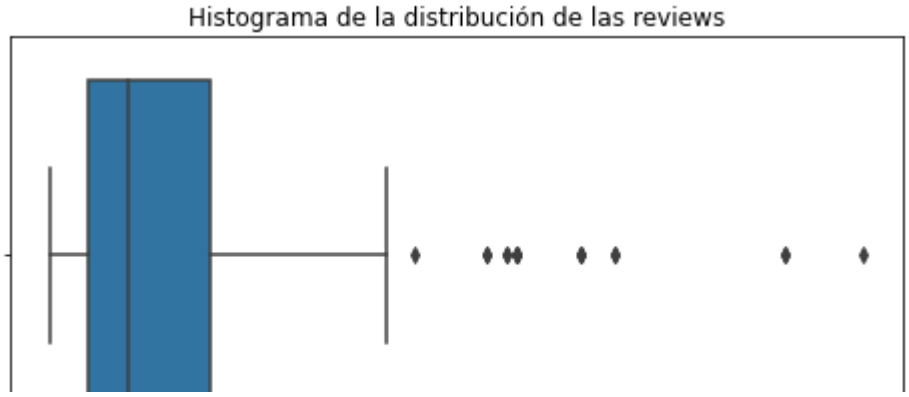
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8a0472b2d0>
```




```
# Ahora muéstralo en un gráfico de caja y bigote.
fig = plt.figure(figsize=(8, 4))
sns.boxplot(data=df, x='Reviews')
plt.title('Histograma de la distribución de las reviews')
```

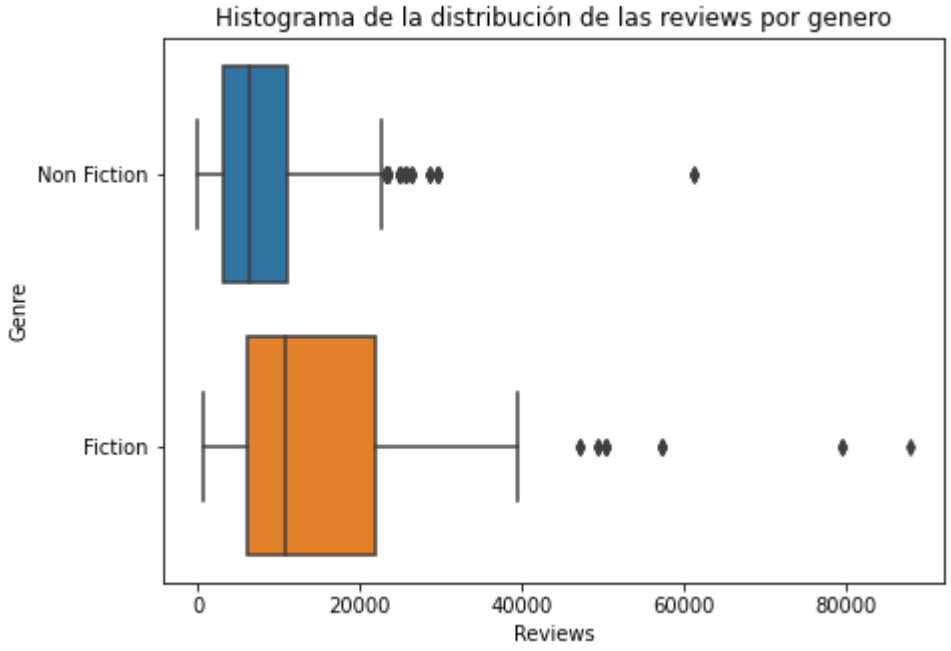
Se ha guardado correctamente

```
Text(0.5, 1.0, 'Histograma de la distribución de las 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=(7,5))
sns.boxplot(data=df, x='Reviews', y = 'Genre')
plt.title('Histograma de la distribución de las reviews por genero')
```

```
Text(0.5, 1.0, 'Histograma de la distribución de las reviews por genero')
```



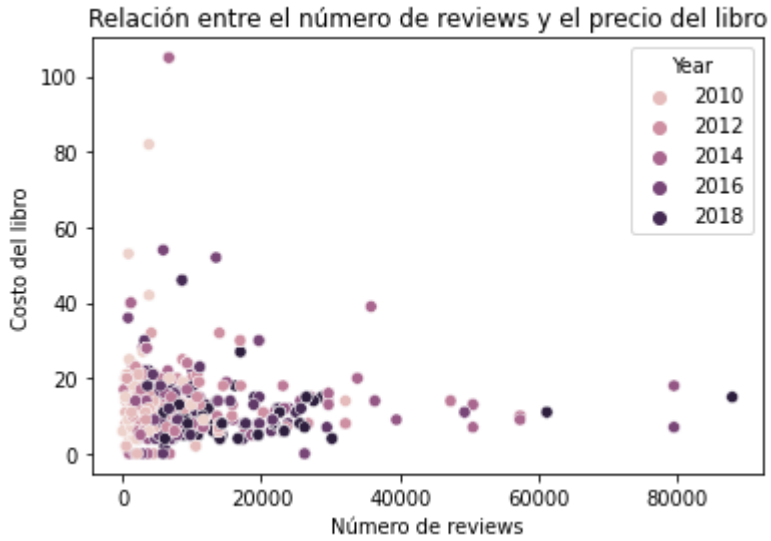

```
# ¿Cuál es la relación entre el número de reseñas y precios? Muéstralo en un
# gráfico de dispersión.
fig = plt.figure(figsize=(6, 4))
sns.scatterplot(data=df, x ='Reviews', y='Price')
plt.title('Relación entre el número de reviews y el precio del libro')
plt.xlabel('Número de reviews')
plt.ylabel('Costo del libro')
```

Se ha guardado correctamente

```
Text(0, 0.5, 'Costo del libro')
```


```
# 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.
fig = plt.figure(figsize=(6, 4))
sns.scatterplot(data=df, x ='Reviews', y='Price', hue='Year')
plt.title('Relación entre el número de reviews y el precio del libro')
plt.xlabel('Número de reviews')
plt.ylabel('Costo del libro')
```

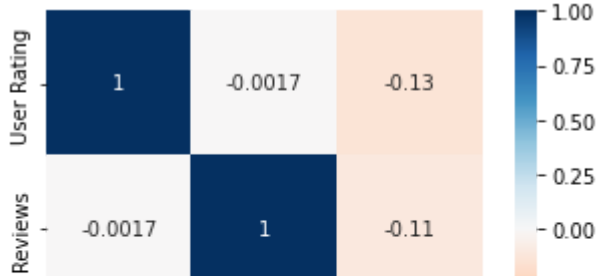
```
Text(0, 0.5, 'Costo del libro')
```




```
# ¿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.
dfwy = df.drop(columns='Year')
dfwyc = dfwy.corr()
sns.heatmap(data=dfwyc, vmin=-1, vmax=1, cmap = 'RdBu', annot=True, square = True)
```

Se ha guardado correctamente

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8a040aad10>
```

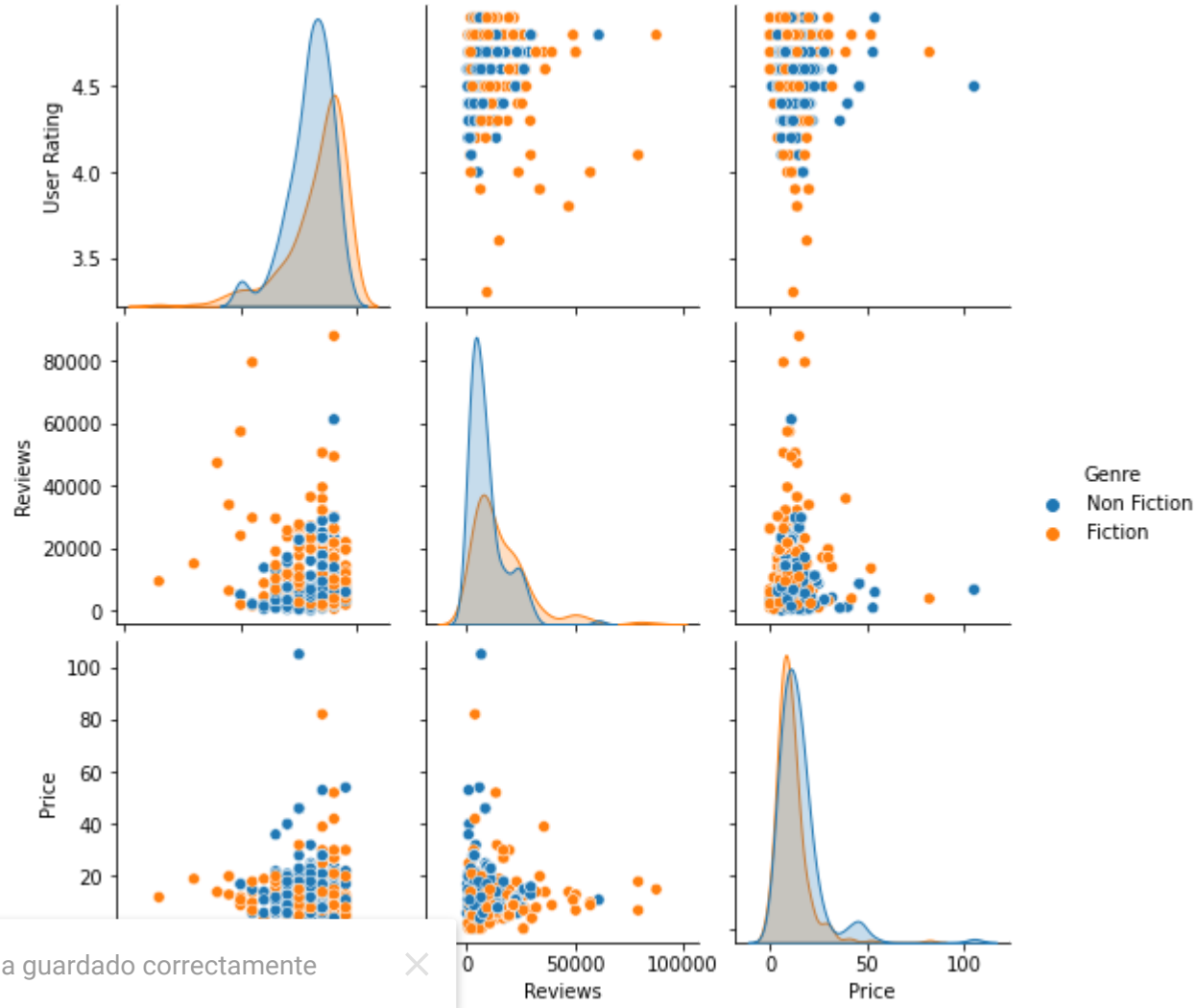


¿Cuáles variables tiene una fuerte relación positiva entre sí y cuáles tienen una fuerte relación negativa? No hay presentes variables númericas que tengan correlaciones fuertes entre sí dentro de estos datos, la relación mas fuerte que se observa es la negativa entre user rating y price.

User Rating Reviews Price

```
# 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(Genre). La variable año, a pesar de ser numérica, la vamos a considerar como
# cualitativa, así que la eliminaremos del análisis.
dfwy = df.drop(columns='Year')
sns.pairplot(data=dfwy, hue = 'Genre')
```

```
<seaborn.axisgrid.PairGrid at 0x7f8a0167c510>
```



Se ha guardado correctamente

✓ 4 s completado a las 21:09

Se ha guardado correctamente