Actividad - Estadística básica

- **Nombre:** Manuel Villalpando Linares
- **Matrícula:** A01352033

Entregar: Archivo PDF de la actividad, así como el archivo .ipynb en tu repositorio. **Nota:** Recuerda habrá una penalización de **50** puntos si la actividad fue entregada fuera de la fecha límite.

Carga el conjunto de datos `bestsellers with categories.csv` (se encuentra en el repositorio de la clase) y realiza un análisis estadístico de las variables.

```
# 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.
df = pd.read_csv('bestsellers with categories.csv')
df.head(6)
```

	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	National	4.8	7665	12	2019	Non

El conjunto de datos es una tabla que contiene el top 50 de los libros más vendidos por Amazon por año desde 2009 hasta 2019. Cada libro está clasificado como Ficción o No ficción.

Las variables que contiene son:

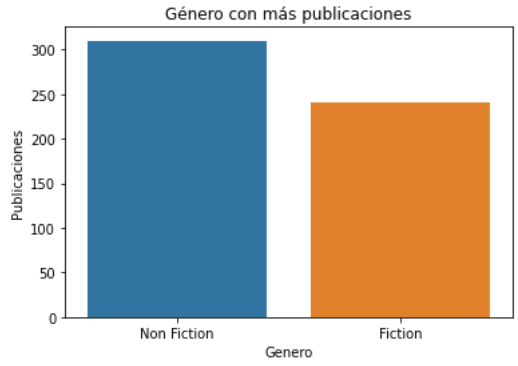
- **Name**: Nombre del libro.
- **Author**: Autor.
- **User Rating**: Calificación promedio que los usuarios asignaron al libro (1-5).
- **Reviews**: Número de reseñas.
- **Price**: Precio del libro.
- **Year**: Año de publicación.
- **Genre**: Género literario (ficción/no ficción).

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

```
        User Rating      Reviews      Price        Year
```

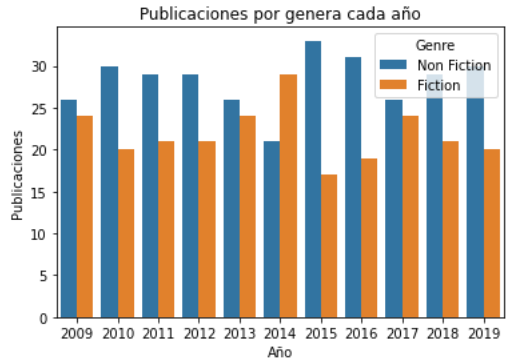
```
## ¿Cuál es el género con más publicaciones? Muéstralo en un gráfico.
fig = plt.figure(figsize=(6,4))

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

plt.title('Género con más publicaciones')
plt.xlabel('Genero')
plt.ylabel('Publicaciones')
```

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



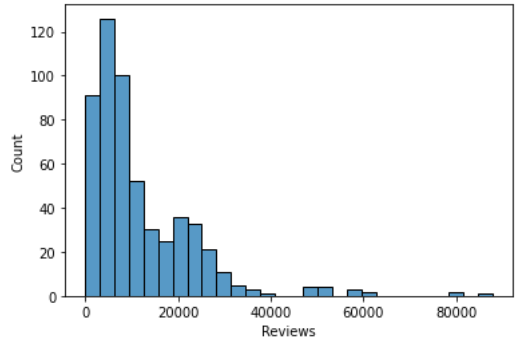

```
# ¿Cuántos libros del top 50 se publicaron por género en cada año? ¿Hay algún
# año donde hubo más libros de ficción en el top 50?. Muéstralo en un gráfico.
fig2 = plt.figure(figsize=(6,4))

sns.countplot(data=df, x = 'Year', hue='Genre')

plt.title('Publicaciones por genera cada año')
plt.xlabel('Año')
plt.ylabel('Publicaciones')
```

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



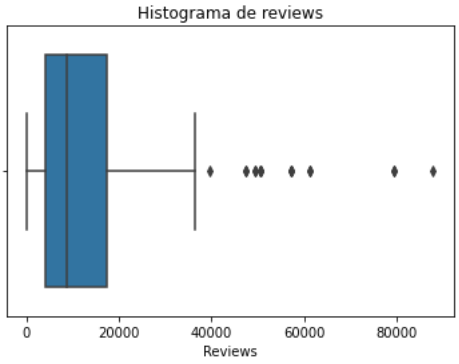

```
# ¿Cómo se distribuye la variable Review? Muéstra el histografa.
sns.histplot(data=df, x='Reviews')
```

```
<Axes: xlabel='Reviews', ylabel='Count'>
```



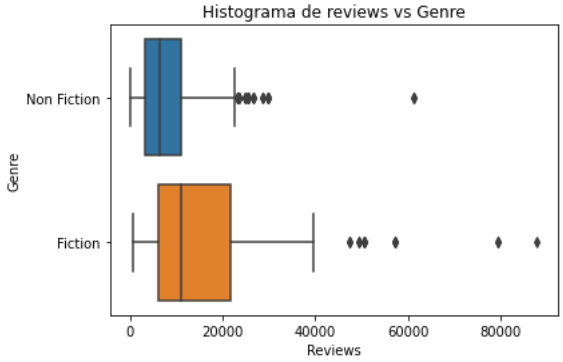
```
# Ahora muéstralo en un gráfico de caja y bigote.
fig3 = plt.figure(figsize=(6, 4))

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

plt.title('Histograma de reviews')
```

```
Text(0.5, 1.0, 'Histograma de reviews')
```



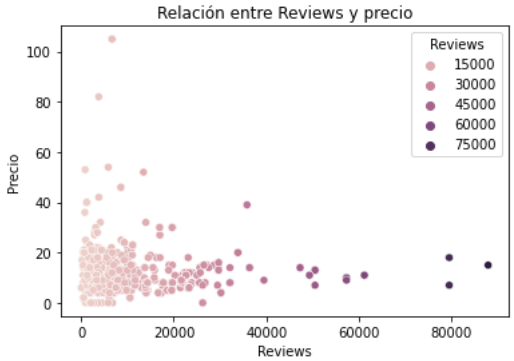

```
# ¿Cómo se compara la evaluación del libro por género? ¿Qué genero es mejor
# evaluado por los lectores? Muéstralo en un solo gráfico de caja y bigote.
fig4 = plt.figure(figsize=(6, 4))

sns.boxplot(data=df, x='Reviews', y="Genre")

plt.title('Histograma de reviews vs Genre')
```

```
Text(0.5, 1.0, 'Histograma de reviews vs Genre')
```



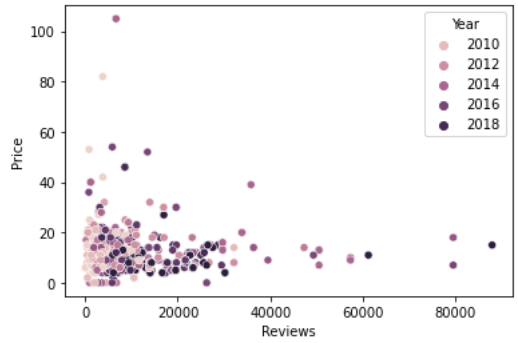

```
# ¿Cuál es la relación entre el número de reseñas y precios? Muéstralo en un
# gráfico de dispersión.
fig5 = plt.figure(figsize=(6, 4))

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

plt.title('Relación entre Reviews y precio')
plt.xlabel('Reviews')
plt.ylabel('Precio')
```

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



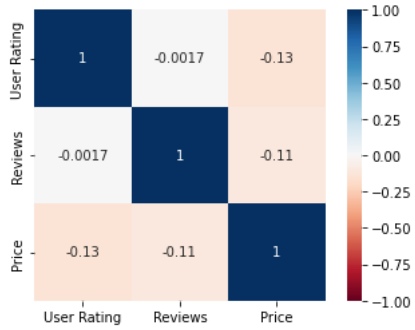
```
# De la pregunta anterior, ¿influye algo el año de publicación? ¿Cuál es la
# relación entre el número de reseñar, el precio y el año de publicación?
# IMPORTANTE: Selecciona una paleta de colores adeacuada.
fig6 = plt.figure(figsize=(6, 4))
sns.scatterplot(data=df, x='Reviews', y='Price', hue='Year')
```

```
<Axes: xlabel='Reviews', ylabel='Price'>
```




```
# ¿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.
df2 = pd.read_csv('bestsellers with categories.csv', usecols=['Price', 'Reviews', 'User Rating'])
dfcorr = df2.corr()

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

```
<Axes: >
```

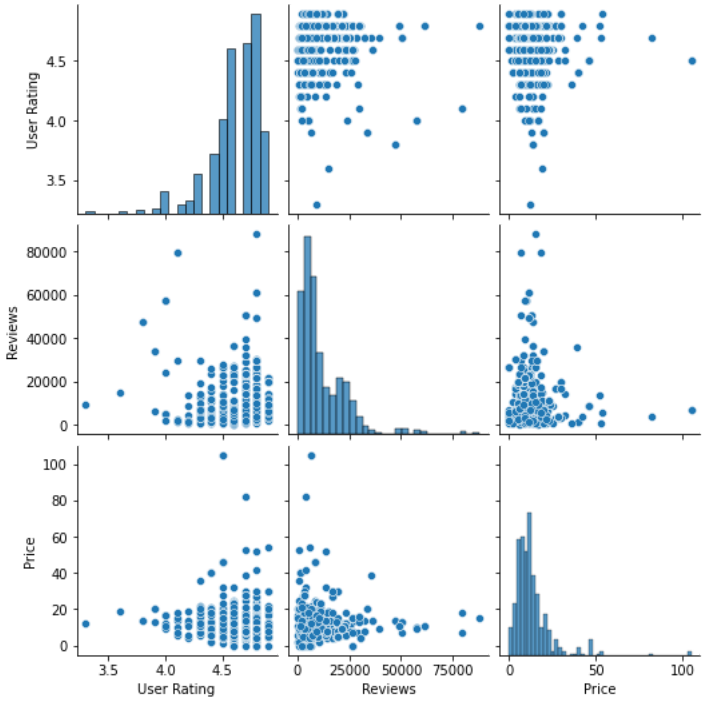


¿Cuáles variables tiene una fuerte relación positiva entre sí y cuáles tienen una fuerte relación negativa? (Esta pregunta no es de código) Responde la pregunta en la siguiente celda de texto.

Se puede apreciar que existe una ligera inclinacion a que las variables que tienen relacion negativa entre si son las de user rating y price, mientras que no se observa una fuerte relacion positiva como tal.

```
# 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.
sns.pairplot(data=df2)
```

```
<seaborn.axisgrid.PairGrid at 0x7efcfcff5fd0>
```



✓ 2 s completado a las 23:36