

Actividad - Estadística básica

- **Nombre:**Daniel Queijeiro Albo
- **Matrícula:** A01710441

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.
!curl -O https://raw.githubusercontent.com/Manchas2k4/tc1002S/main/datasets/bestsellers%20with%20categories.csv

df = pd.read_csv('bestsellers%20with%20categories.csv')
df.head(6)
```

% Total		% Received		% Xferd	Average	Speed	Time	Time	Time	Current					
100	51161	100	51161	0	0	103k	0	--:--:--	--:--:--	--:--:--	103k				
						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			
.		5,000 Awesome Facts (About Everything!)					National Geographic	4.8	7777	10	2010	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:

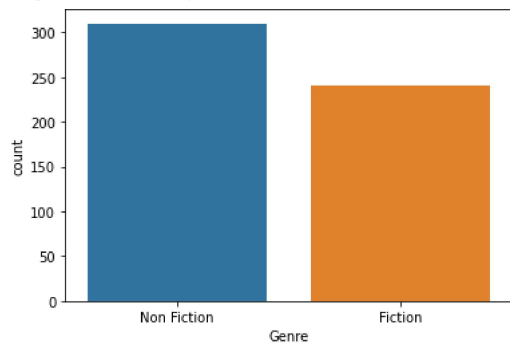
- **Name:** Nombre del libro.
- **Author:** Autor.
- **User Rating:** Calificación promedio que los usuarios asignaron al libro (1-5).
- **Reviews:** Número de reseñas.
- **Price:** Precio del libro.
- **Year:** Año de publicación.
- **Genre:** Género literario (ficción/no ficción).

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

	User Rating	Reviews	Price	Year
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

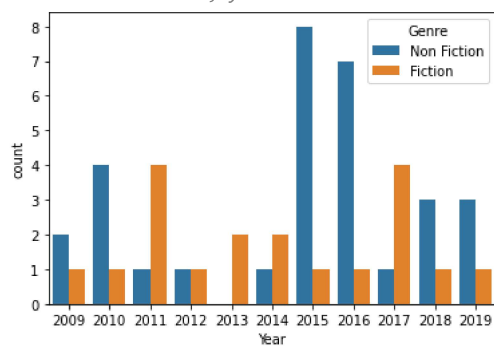
```
## ¿Cuál es el género con más publicaciones? Muéstralo en un gráfico.
sns.countplot(x='Genre', data=df)
print('El género con más publicaciones es el de non-fiction')
```

El género con más publicaciones es el de non-fiction



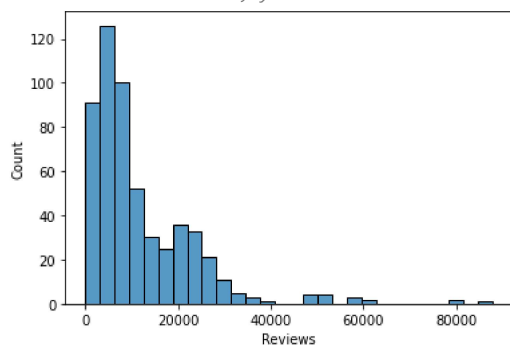
```
# ¿Cuántos libros del top 50 se publicaron por género en cada año? ¿Hay algún
# año donde hubo más libros de ficción en el top 50?. Muéstralo en un gráfico.
ndf = df.iloc[:50]
sns.countplot(x='Year', hue='Genre', data=ndf)
```

<Axes: xlabel='Year', ylabel='count'>



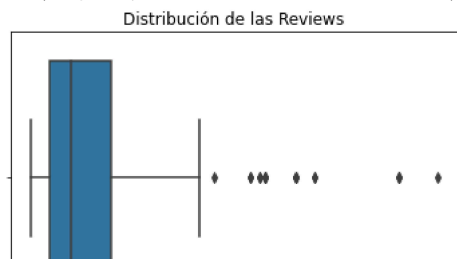
```
# ¿Cómo se distribuye la variable Review? Muéstra el histografa.
sns.histplot(data=df, x='Reviews')
```

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



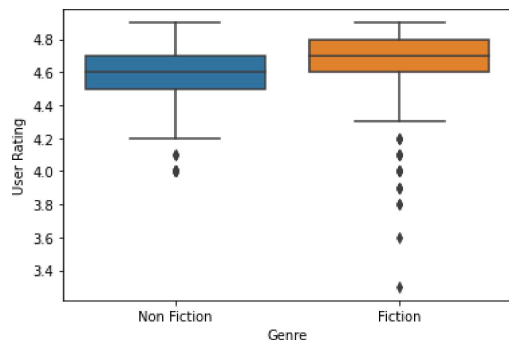
```
# Ahora muéstralo en un gráfico de caja y bigote.
sns.boxplot(data=df, x='Reviews')
plt.title('Distribución de las Reviews')
```

```
Text(0.5, 1.0, 'Distribución de las Reviews')
```



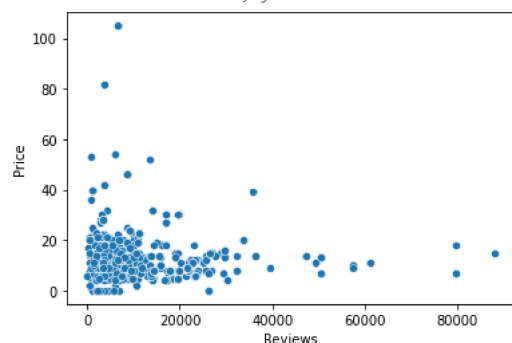
```
# ¿Cómo se compara la evaluación del libro por género? ¿Qué genero es mejor
# evaluado por los lectores? Muéstralo en un solo gráfico de caja y bigote.
sns.boxplot(x='Genre', y='User Rating', data=df)
print('El género con mejor rating es el de ficción')
```

El género con mejor rating es el de ficción



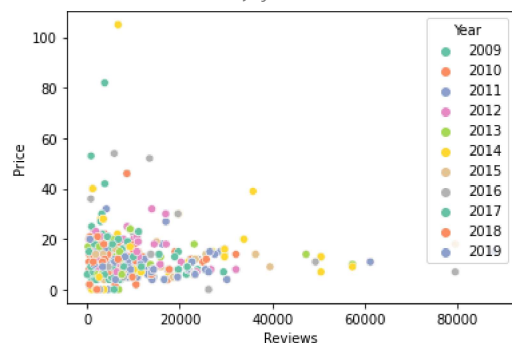
```
# ¿Cuál es la relación entre el número de reseñas y precios? Muéstralo en un
# gráfico de dispersión.
sns.scatterplot(x='Reviews', y='Price', data=df)
```

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



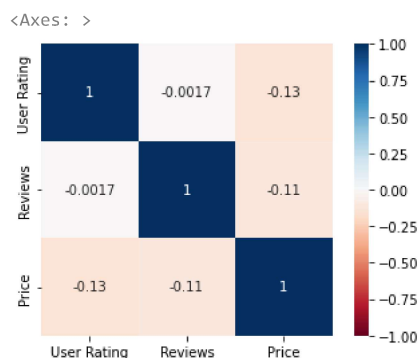
```
# De la pregunta anterior, ¿influye algo el año de publicación? ¿Cuál es la
# relación entre el número de reseñas, el precio y el año de publicación?
# IMPORTANTE: Selecciona una paleta de colores adecuada.
sns.scatterplot(x='Reviews', y='Price', data=df, palette=sns.color_palette('Set2'), hue='Year')
```

```
<ipython-input-10-e78836e178e7>:4: UserWarning:
The palette list has fewer values (8) than needed (11) and will cycle, which may produce an uninterpret
sns.scatterplot(x='Reviews', y='Price', data=df, palette=sns.color_palette('Set2'), 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.
ndf = df.drop('Year', axis=1)
ndf_corr = ndf.corr()

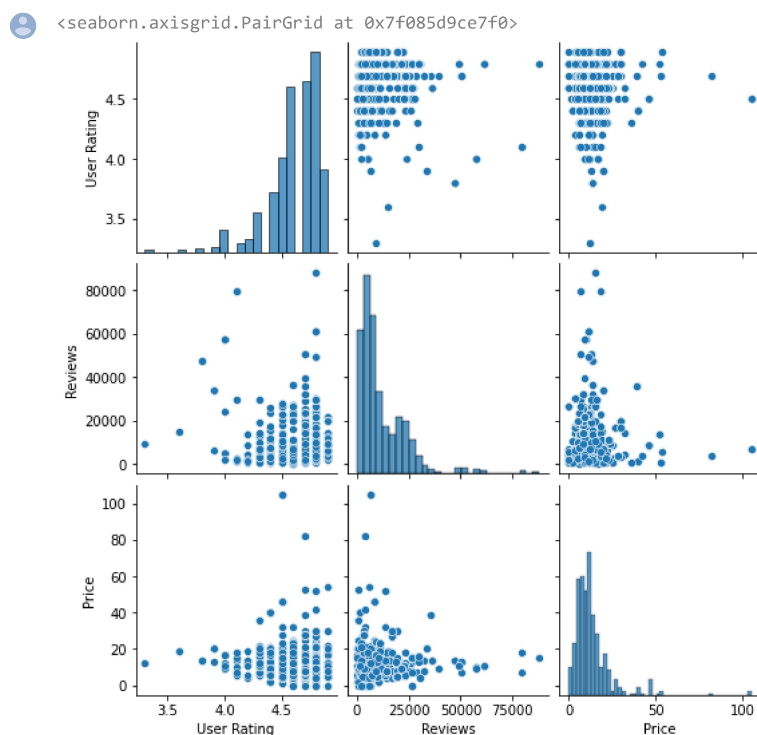
sns.heatmap(data=ndf_corr, vmin=-1, vmax=1, cmap = 'RdBu', annot=True, square = True)
```



¿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.

**** Escribe tu respuesta **** Las variables tienen una fuerte relación positiva con ellas mismas, y las más negativas se encuentran en las esquinas, entre las variables User Rating y 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. La variable año, a pesar de ser numérica, la vamos a considerar como
# cualitativa, así que la eliminaremos del análisis.
sns.pairplot(data=ndf)
```



✓ 0 s se ejecutó 22:43

