Actividad - Estadística básica

- **Nombre:**Diego Arturo Padilla Domínguez
- **Matrícula:**A01552594

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.
from google.colab import drive
drive.mount('/gdrive')
%cd /gdrive/MyDrive/SemanaTec/Repositorio1/
```

```
Mounted at /gdrive
/gdrive/MyDrive/SemanaTec/Repositorio1
```

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

```
df.head(6)
```

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

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:

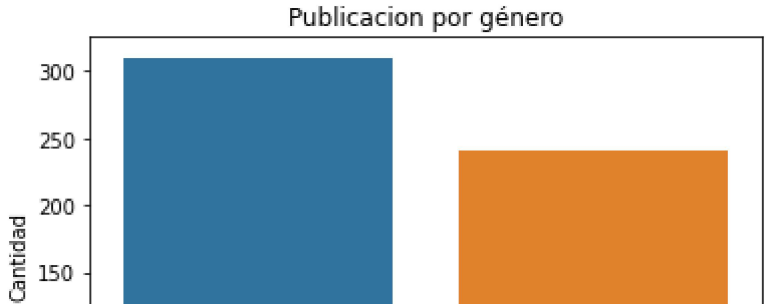
- **Name**: Nombre del libro.
- **Author**: Autor.
- **User Rating**: Calificación promedio que los usuarios asignaron al libro (1-5).
- **Reviews**: Número de reseñas.
- **Price**: Precio del libro.
- **Year**: Año de publicación.
- **Genre**: Género literario (ficción/no ficción).

```
# 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
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

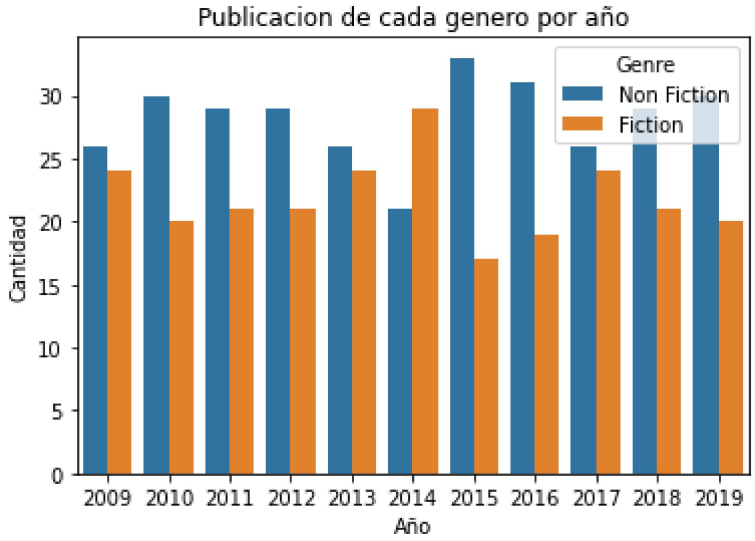
```
## ¿Cuál es el género con más publicaciones? Muéstralo en un gráfico.
fig = plt.figure(figsize=(6,4))
sns.countplot(data=df, x = 'Genre')
plt.title('Publicacion por género')
plt.xlabel('Género')
plt.ylabel('Cantidad')
```

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




```
# ¿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.
fig = plt.figure(figsize=(6,4))
sns.countplot(data=df, x="Year", hue="Genre")
plt.title('Publicación de cada género por año')
plt.xlabel('Año')
plt.ylabel('Cantidad')
```

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

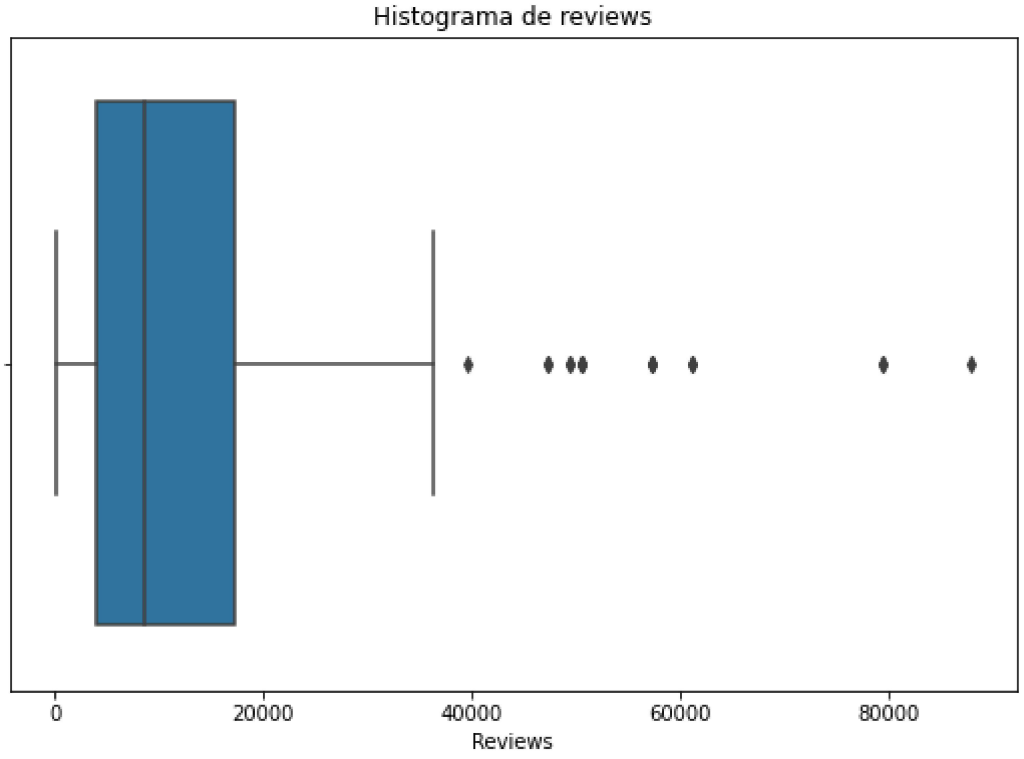



```
# ¿Cómo se distribuye la variable Review? Muéstra el histografa.
fig = plt.figure(figsize=(9,6))
sns.histplot(data=df, x='Reviews', kde=True)
plt.xlabel('Review')
plt.ylabel('Frecuencia')
plt.title('Histograma de reviews')
```

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

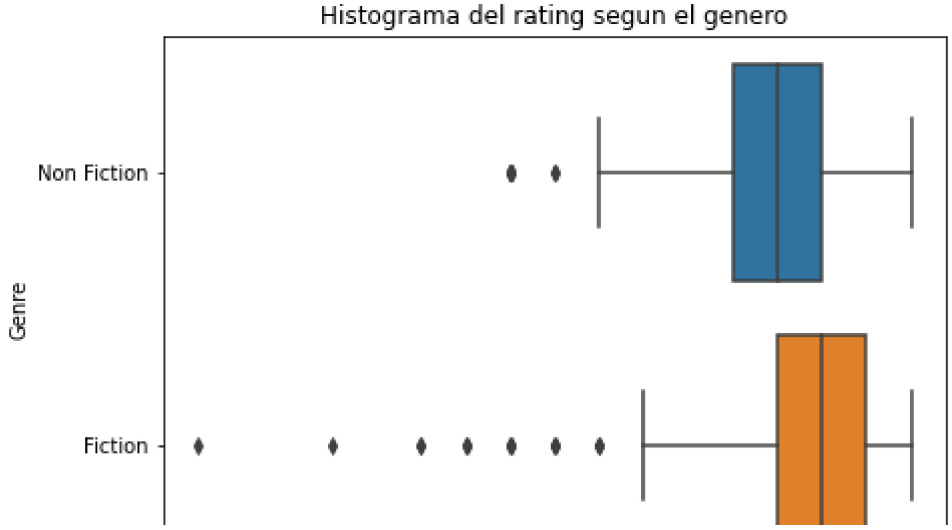

```
# Ahora muéstralo en un gráfico de caja y bigote.
fig = plt.figure(figsize=(9, 6))
sns.boxplot(data=df, x='Reviews')
plt.title('Histograma de reviews')
```

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



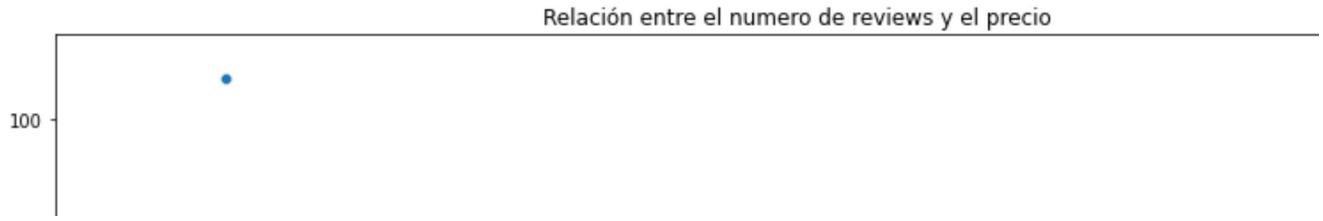

```
# ¿Cómo se compara la evaluación del libro por género? ¿Qué genero es mejor
# evaluado por los lectores? Muéstralo en un solo gráfico de caja y bigote.
fig = plt.figure(figsize=(7,5))
sns.boxplot(data=df, x='User Rating', y = 'Genre')
plt.title('Histograma del rating segun el genero')
```

```
Text(0.5, 1.0, 'Histograma del rating segun el genero')
```



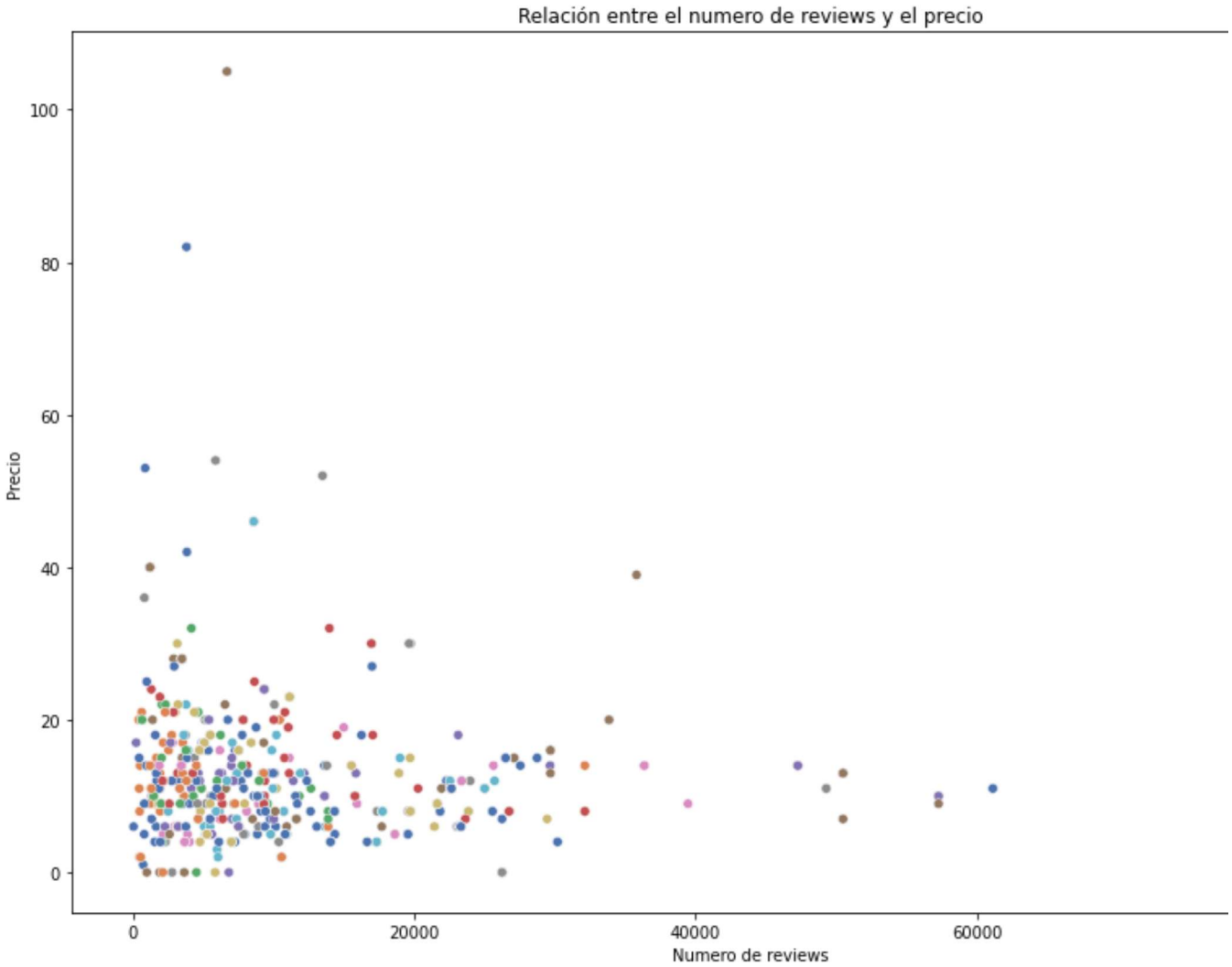

```
# ¿Cuál es la relación entre el número de reseñas y precios? Muéstralo en un
# gráfico de dispersión.
fig = plt.figure(figsize=(15, 10))
sns.scatterplot(data=df, x ='Reviews', y='Price')
plt.title('Relación entre el numero de reviews y el precio')
plt.xlabel('Numero de reviews')
plt.ylabel('Precio')
```

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



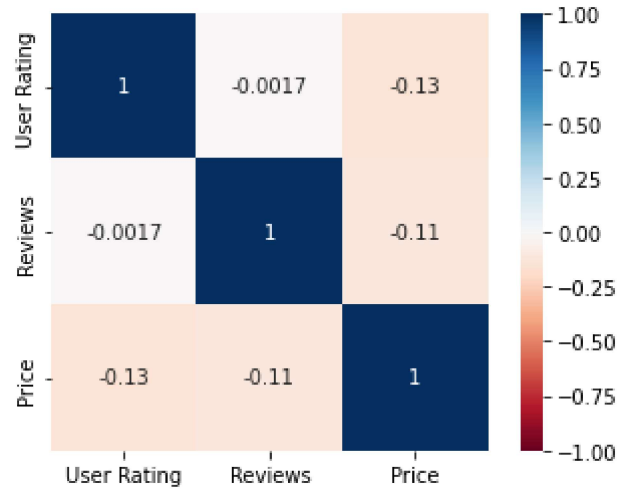

```
# De la pregunta anterior, ¿influye algo el año de publicación? ¿Cuál es la
# relación entre el número de reseñar, el precio y el año de publicación?
# IMPORTANTE: Selecciona una paleta de colores adeacuada.
fig = plt.figure(figsize=(15, 10))
sns.scatterplot(data=df, x ='Reviews', y='Price', hue = 'Year', palette = 'deep')
plt.title('Relación entre el numero de reviews y el precio')
plt.xlabel('Numero de reviews')
plt.ylabel('Precio')
```

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



```
# ¿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=[0, 1, 2, 3, 4, 6])
sns.heatmap(data=df2.corr(), vmin=-1, vmax=1, cmap = 'RdBu', annot=True, square = True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd320fd0fd0>
```

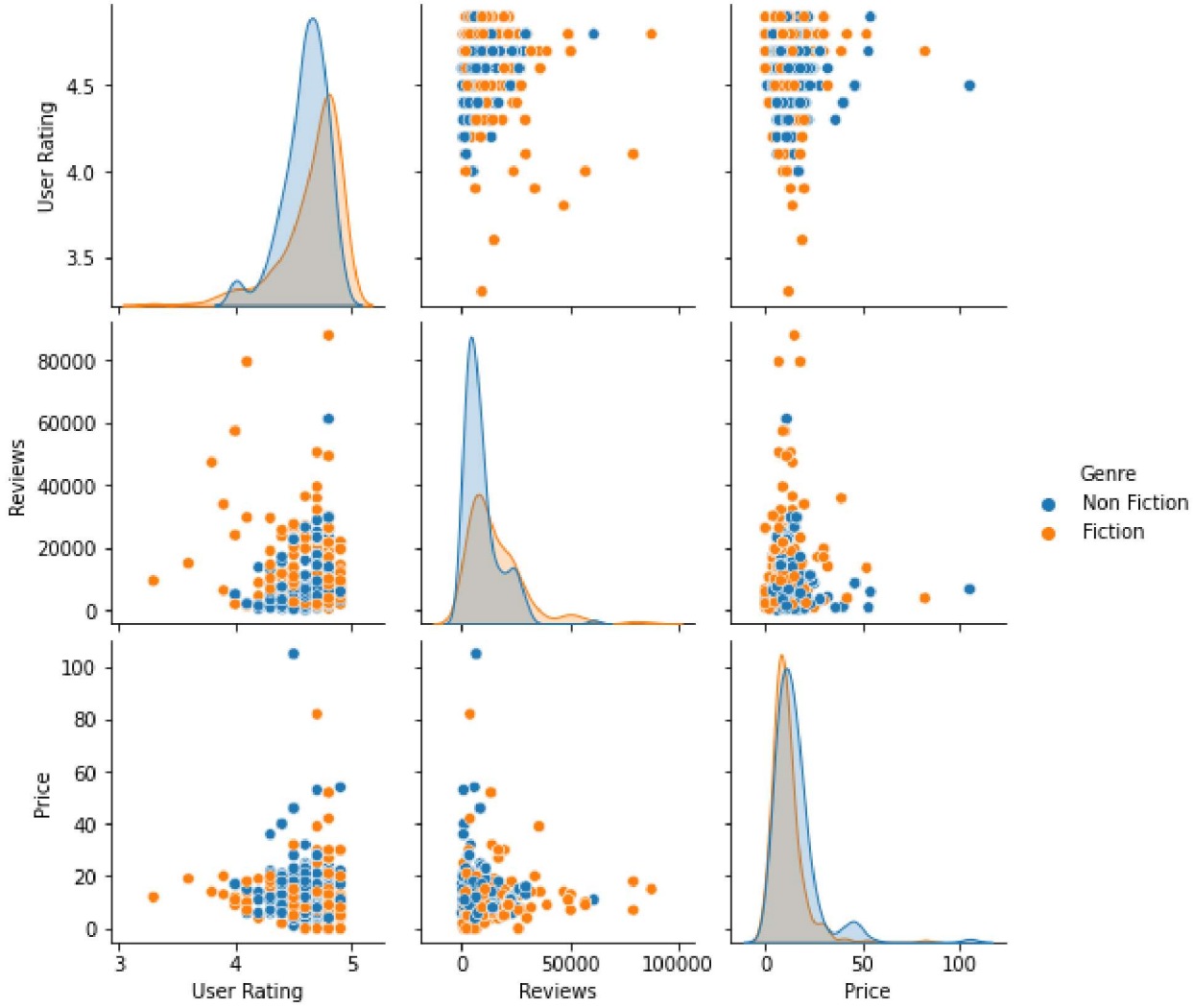


¿Cuáles variables tiene una fuerte relación positiva entre sí y cuáles tienen una fuerte relación negativa? (Esta pregunta no es de código) Responde la pregunta en la siguiente celda de texto.

```
# Las relaciones que hay, son demasiado débiles, esto nos da a entender que realmente no hay relación
```

```
# 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, hue = 'Genre')
```

```
<seaborn.axisgrid.PairGrid at 0x7fd31d9e77d0>
```



✓ 4 s se ejecutó 11:52