# Actividad - Estadística básica

- **Nombre: Oscar Eduardo Nieto Espitia**
- **Matrícula: A01705090**

**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 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])))

df = pd.read_csv('bestsellers with categories.csv')
df1 = pd.read_csv('bestsellers with categories.csv')
df2 = pd.read_csv('bestsellers with categories.csv')
df1.head(6)
```

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:

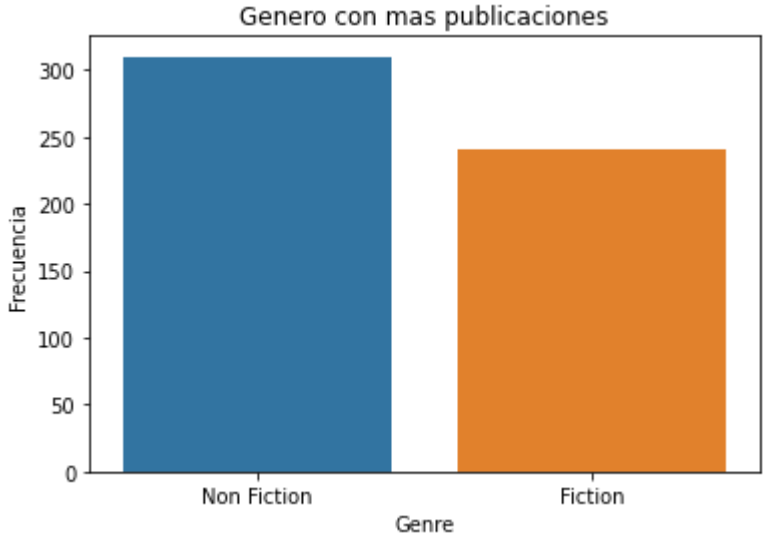
- **Name**: Nombre del libro.
- **Author**: Autor.
- **User Rating**: Calificación promedio que los usuarios asignaron al libro (1-5).
- **Reviews**: Número de reseñas.
- **Price**: Precio del libro.
- **Year**: Año de publicación.
- **Genre**: Género literario (ficción/no ficción).

```
# Crea una tabla resumen con los estadísticas generales de las variables
# numéricas.
df = df.drop('Name', axis=1)
df = df.drop('Author', axis=1)
df = df.drop('Genre', axis=1)
df.head(6)
```

| | User Rating | Reviews | Price | Year |
|---|---|---|---|---|
| **0** | 4.7 | 17350 | 8 | 2016 |
| **1** | 4.6 | 2052 | 22 | 2011 |
| **2** | 4.7 | 18979 | 15 | 2018 |
| **3** | 4.7 | 21424 | 6 | 2017 |
| **4** | 4.8 | 7665 | 12 | 2019 |
| **5** | 4.4 | 12643 | 11 | 2011 |

```
## ¿Cuál es el género con más publicaciones? Muéstralo en un gráfico.
fig = plt.figure(figsize=(6,4))
sns.countplot(data=df1, x = 'Genre')
plt.title('Genero con mas publicaciones')
plt.xlabel('Genre')
plt.ylabel('Frecuencia')
```

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




```
# ¿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.
# Tamaño de la imagen
fig = plt.figure(figsize=(9,6))

# Gráfico
sns.histplot(data=df1, x='Year', hue='Genre', bins=20, kde=True)

# Ejes y título
plt.xlabel('Anio')
plt.ylabel('Genero')
plt.title('Cantidad de libros por genero y anio')
```

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

```
# Ahora muéstralo en un gráfico de caja y bigote.
fig = plt.figure(figsize=(9, 6))

# Gráfico boxplot
sns.boxplot(data=df1, x='User Rating')

# Ejes y título
plt.title('Histograma del ancho de sépalo por especie')
```

```
# ¿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=(9,6))

# Gráfico
sns.histplot(data=df1, x='User Rating', hue='Genre', bins=20, kde=True)

# Ejes y título
plt.xlabel('Anio')
plt.ylabel('Genero')
plt.title('Cantidad de libros por genero y anio')
```

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

# Gráfico scatterplot.
sns.scatterplot(data=df, x ='Reviews', y='User Rating', hue='Price')
```

```
# Ejes y título. Colocamos la etiqueta correcta de acuerdo a la orientación.
plt.title("Relacion entre numero de resenias Rating por precio")
plt.xlabel('Reviews')
plt.ylabel('User Rating')
```

```
# 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=(6, 4))

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

# Ejes y título. Colocamos la etiqueta correcta de acuerdo a la orientación.
plt.title("Relacion entre numero de resenias Anio por precio")
plt.xlabel('Year')
plt.ylabel('User Rating')
```

```
# ¿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 = df
sns.pairplot(data=df2, hue='Year')
```

¿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 **

```
# 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
```

```
# del libro. La variable año, a pesar de ser numérica, la vamos a considerar como
# cualitativa, así que la eliminaremos del análisis.
df2 = df2.drop('Year', axis=1)
iris_corr = df2.corr()
sns.heatmap(data=iris_corr, vmin=-1, vmax=1, cmap = 'RdBu', annot=True, square =
```

Haz doble clic (o pulsa Intro) para editar

✓ 0 s completado a las 13:32