

▼ 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; sns.set()

# 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	Year	Genre
------	--------	-------------	---------	-------	------	-------

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).

▼ Análisis estadístico

1. Carga la tabla de datos y haz un análisis estadístico de las variables.

- Verifica la cantidad de datos que tienes, las variables que contiene cada vector de datos e identifica el tipo de variables.
- Analiza las variables para saber que representa cada una y en que rangos se encuentran. Si la descripción del problema no te lo indica, utiliza el máximo y el mínimo para encontrarlo.
- Basándote en la media, mediana y desviación estándar de cada variable, ¿qué conclusiones puedes entregar de los datos?
- Calcula la correlación de las variables que consideres relevantes.

```
# Escribe el código necesario para realizar el análisis estadístico descripto
# anteriormet.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550 entries, 0 to 549
Data columns (total 7 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Name             550 non-null   object
1   Author           550 non-null   object
2   User Rating      550 non-null   float64
3   Reviews          550 non-null   int64
4   Price            550 non-null   int64
5   Year             550 non-null   int64
6   Genre            550 non-null   object
dtypes: float64(1), int64(3), object(3)
memory usage: 30.2+ KB
```

```
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

```
df.corr()
```

	User Rating	Reviews	Price	Year
User Rating	1.000000	-0.001729	-0.133086	0.242383
Reviews	-0.001729	1.000000	-0.109182	0.263560
Price	-0.133086	-0.109182	1.000000	-0.153979
Year	0.242383	0.263560	-0.153979	1.000000

¿Cuáles son las variables relevantes e irrelevantes para el análisis?

Desde mi punto de vista las mas relevantes son: "user rating", "reviews" "price", ya que con ellas podriamos obtener una mayor idea de como evolucionó la percepcion del cine através de los años. A diferencia de "name", "author" y "genre", que solamente nos sirven para clasificar los datos.

▼ Análisis gráfico

Realiza el análisis de las variables usando diagramas de cajas y bigotes, histogramas y mapas de calor.

Responde las siguientes preguntas:

- ¿Hay alguna variable que no aporta información? Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?
- ¿Existen variables que tengan datos extraños?
- Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?
- ¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?

Haz un análisis estadístico de los datos antes de empezar con la segmentación. Debe contener al menos:

- 1 gráfico de caja (boxplot)
- 1 mapa de calor
- 1 gráfico de dispersión

Describe brevemente las conclusiones que se pueden obtener con las gráficas

```
fig, axs = plt.subplots(2, 3, figsize=(12, 8))
sns.boxplot(data=df, y = 'User Rating', x='Genre', ax=axs[0][0])
sns.boxplot(data=df, y = 'Reviews', ax=axs[0][1], x='Genre')
sns.boxplot(data=df, y = 'Price', ax=axs[0][2], x='Genre')
sns.boxplot(data=df, y = 'User Rating', x='Year', ax=axs[1][0])
sns.boxplot(data=df, y = 'Reviews', ax=axs[1][1], x='Year')
sns.boxplot(data=df, y = 'Price', ax=axs[1][2], x='Year')
plt.tight_layout()
plt.suptitle('Distribución de las variables numéricas por género y año', y=1.05)
```

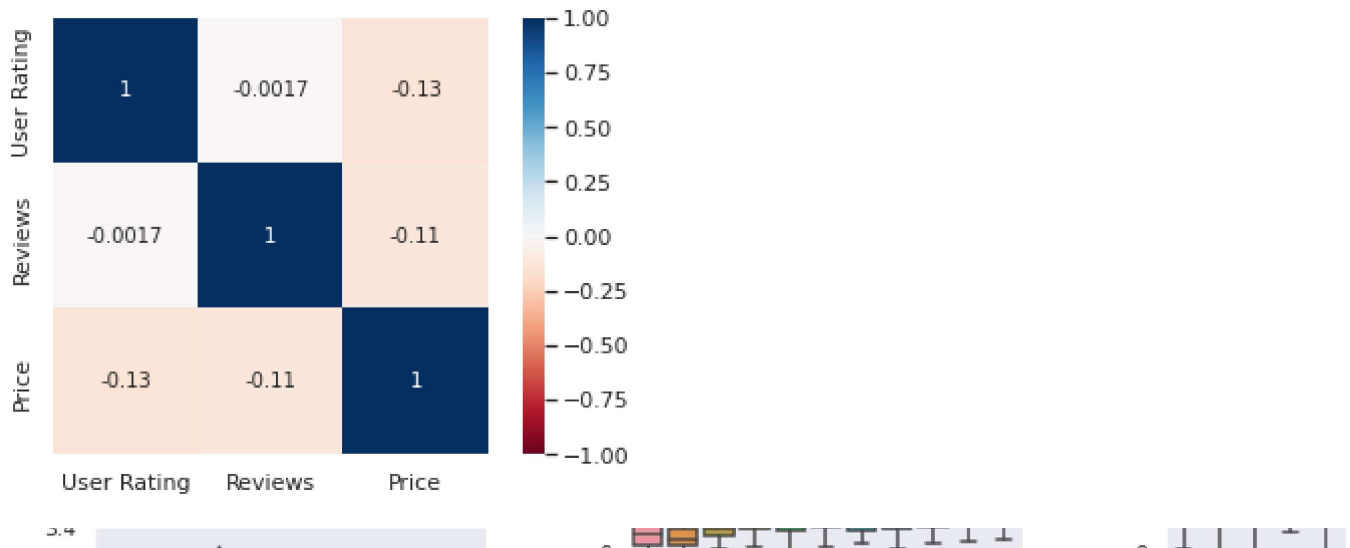
Text(0.5, 1.05, 'Distribución de las variables numéricas por género y año')

Distribución de las variables numéricas por género y año



```
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)
```

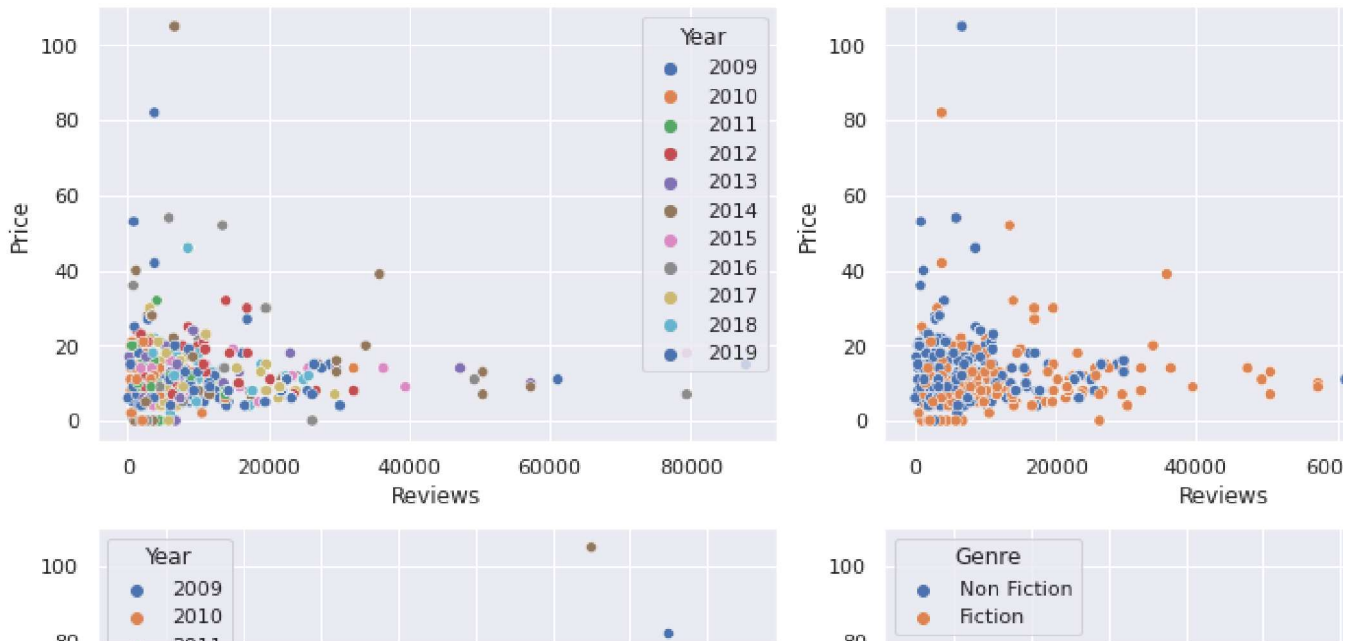
<matplotlib.axes._subplots.AxesSubplot at 0x7f89a4753410>



```
fig, axs = plt.subplots(2, 2, figsize=(12, 8))
sns.scatterplot(data=df, x='Reviews', y='Price', hue='Year', palette='deep', ax=axs[0][0])
sns.scatterplot(data=df, x='Reviews', y='Price', hue='Genre', palette='deep', ax=axs[0][1])
sns.scatterplot(data=df, x='User Rating', y='Price', hue='Year', palette='deep', ax=axs[1][0])
sns.scatterplot(data=df, x='User Rating', y='Price', hue='Genre', palette='deep', ax=axs[1][1])
plt.tight_layout()
plt.suptitle('Distribución del precio segun el numero de reviews por género y año', y=1.05)
```

```
Text(0.5, 1.05, 'Distribución del precio segun el numero de reviews por género y año')
```

Distribución del precio segun el numero de reviews por género y año



Hay alguna variable que no aporta información? Si tuvieras que eliminar variables, ¿cuáles quitarías y por qué?

En este analisis, la variable "year" denota que no es de gran relevancia, ya que a pesar de cambiar, los datos no se ven realmente afectados por ello.

¿Existen variables que tengan datos extraños?

La variable "price" tiene valores los cuales se desvian mucho de la media, esto provoca que algunos calculos sean afectados, por ellos normalmente en estos analisis los datos con mayor desviación son eliminados.

Si comparas las variables, ¿todas están en rangos similares? ¿Crees que esto afecte?

No todas tienen rangos similares, ya que se podria decir que son escalas, lo cual no afecta en los resultados del análisis

¿Puedes encontrar grupos que se parezcan? ¿Qué grupos son estos?

En las graficas de dispersión se puede notar una mayor relación entre la cantidad de reviews y el precio ya que en aquellas que hay mayor cantidad de reviews el precio es bajo, esto podria simbolizar que mientras mas gente pueda acceder a el mayor cantidad de reviews puede obtener, pero esto no cuadra con la grafica de calor ya que en este se nota que realmente no hay una relación significativa.

▼ Clústering


Una vez que hayas realizado un análisis preliminar, haz una segmentación utilizando el método de K-Means. Justifica el número de clusters que elegiste.

- Determina un valor de k
- Calcula los centros de los grupos resultantes del algoritmo k-means

Basado en los centros responde las siguientes preguntas

- ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?
- ¿Cómo obtuviste el valor de k a usar?
- ¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?
- ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?
- ¿Qué puedes decir de los datos basándose en los centros?

```
from sklearn.preprocessing import StandardScaler
numeric_cols = ['User Rating', 'Reviews', 'Price']
X = df2.loc[:, numeric_cols]
scaler = StandardScaler()
X_norm = scaler.fit_transform(X)
X_norm = pd.DataFrame(X_norm, columns=numeric_cols)
X_norm.head()
```

	User Rating	Reviews	Price	
0	0.359990	0.460453	-0.470810	
1	-0.080978	-0.844786	0.821609	
2	0.359990	0.599440	0.175400	
3	0.359990	0.808050	-0.655441	
4	0.800958	-0.365880	-0.101547	

```
# Implementa el algoritmo de kmeans y justifica la elección del número de
# clusters. Usa las variables numéricas.
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
kmax = 16
grupos = range(2, kmax)
wcss = []
sil_score = []
for k in grupos:
    model = KMeans(n_clusters=k)
    clusters = model.fit_predict(X_norm)
    wcss.append(model.inertia_)
    sil_score.append(silhouette_score(X_norm, clusters))
```

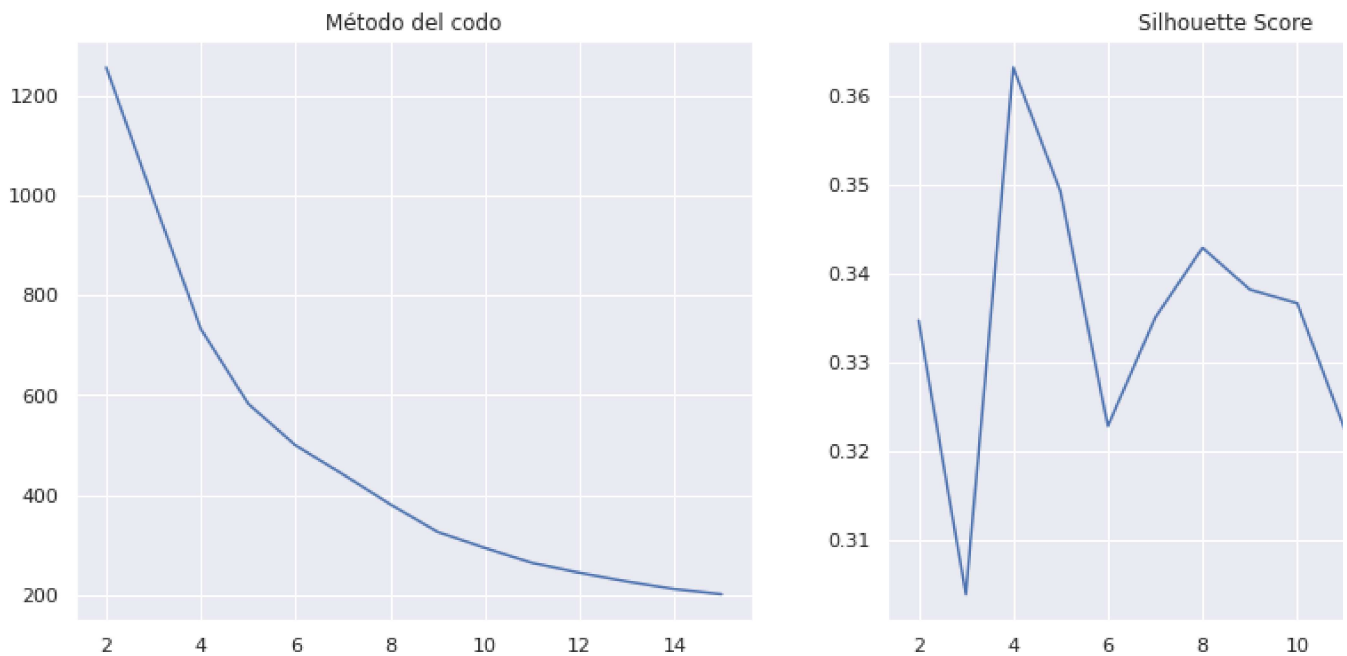
```
fig, axs = plt.subplots(1, 2, figsize=(15, 6))
```

```

fig, axs = plt.subplots(2, 1, figsize=(10, 10))
axs[0].plot(grupos, wcss)
axs[0].set_title('Método del codo')
axs[1].plot(grupos, sil_score)
axs[1].set_title('Silhouette Score')

```

Text(0.5, 1.0, 'Silhouette Score')



En 8 se comienza a estabilizar, por ello selecciono tomare k=8

```

model = KMeans(n_clusters=8)
clusters = model.fit_predict(X_norm)
df2['Grupo'] = clusters.astype('str')
df2.head(20)

```


	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
4	5,000 Awesome Facts (About Everything!) (Natio...	National Geographic Kids	4.8	7665	12
5	A Dance with Dragons (A Song of Ice and Fire)	George R. R. Martin	4.4	12643	11
6	A Game of Thrones / A Clash of Kings / A Storm...	George R. R. Martin	4.7	19735	30
-	A Game of Thrones / A Clash of Kings / A Storm...	George R. R. Martin	4.7	19735	30

- ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?

Si, debido a que con ello podemos ver tambien que datos no son relevantes para el analisis, por su separación del resto

- ¿Cómo obtuviste el valor de k a usar?

Observando la graficas en donde 8 es un numero en donde comienza la estabilización de la gráfica

- ¿Los centros serían más representativos si usaras un valor más alto? ¿Más bajo?

Mientras mas grupos haya se notara más cuales son los mas separados, pero tambien comenzara a ser mas difícl identificar cuales realmente estan separados del resto debido a que habra muchos intermedios. AL contraro de haber meno grupos esto hara que tal vez datos representativos no sean toamdos en cnet.

- ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?

Se necesitarian mas centros para llegara cubrir todos los outliers

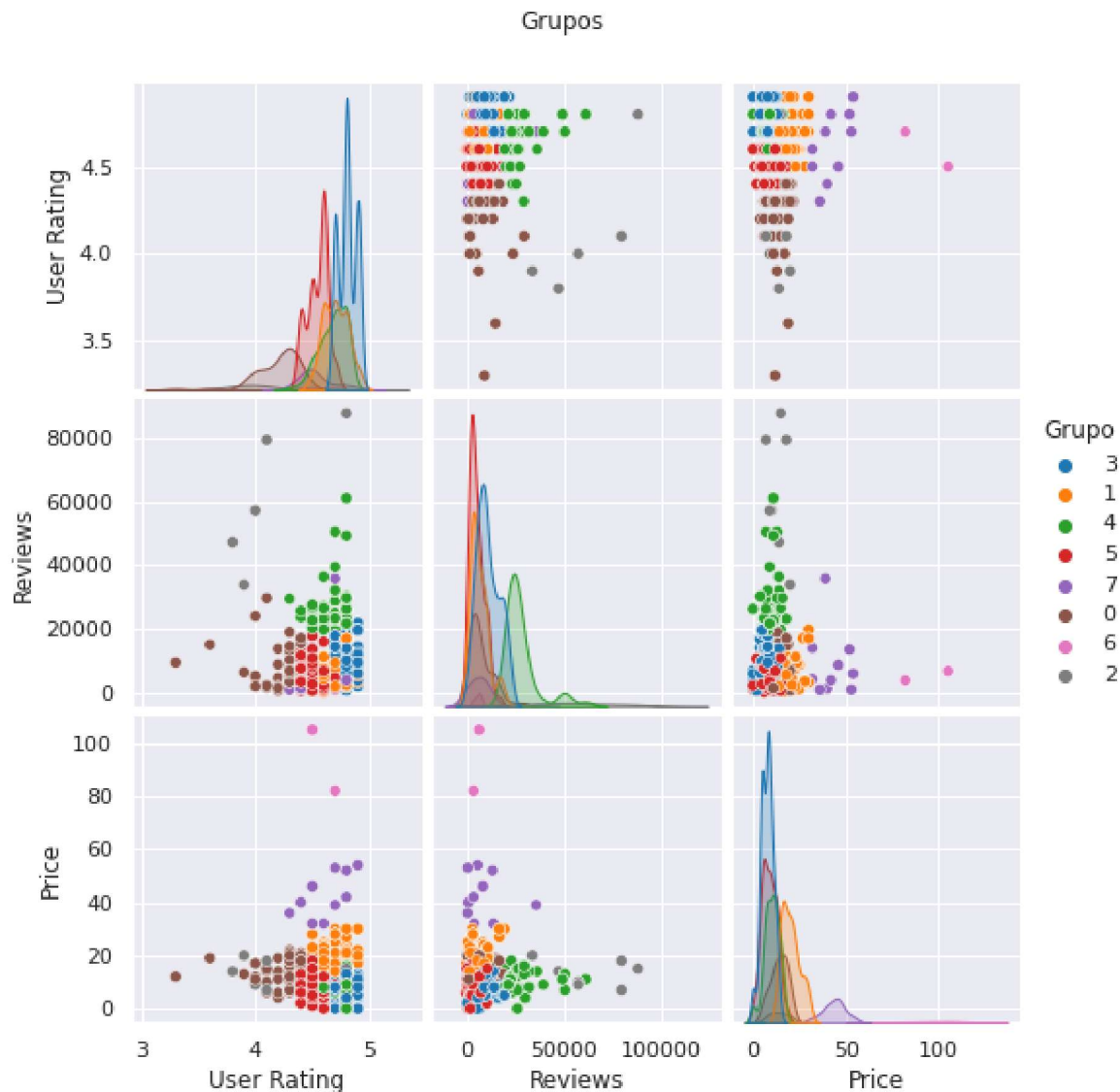
- ¿Qué puedes decir de los datos basándose en los centros? Que hay datos que afectan al analisis debido a que estan muy dispersos de los demás

Analiza las características de cada grupo. ¿Qué nombre le pondrías a cada segmento? Los grupos 2, 7 y 6, son los grupos no representativos Los 0 y 4, medianemte representativos Los 1, 3, 5 altamente representativos

```
# Haz un análisis por grupo para determinar las características que los hace
# únicos. Ten en cuenta todas las variables numéricas.
```

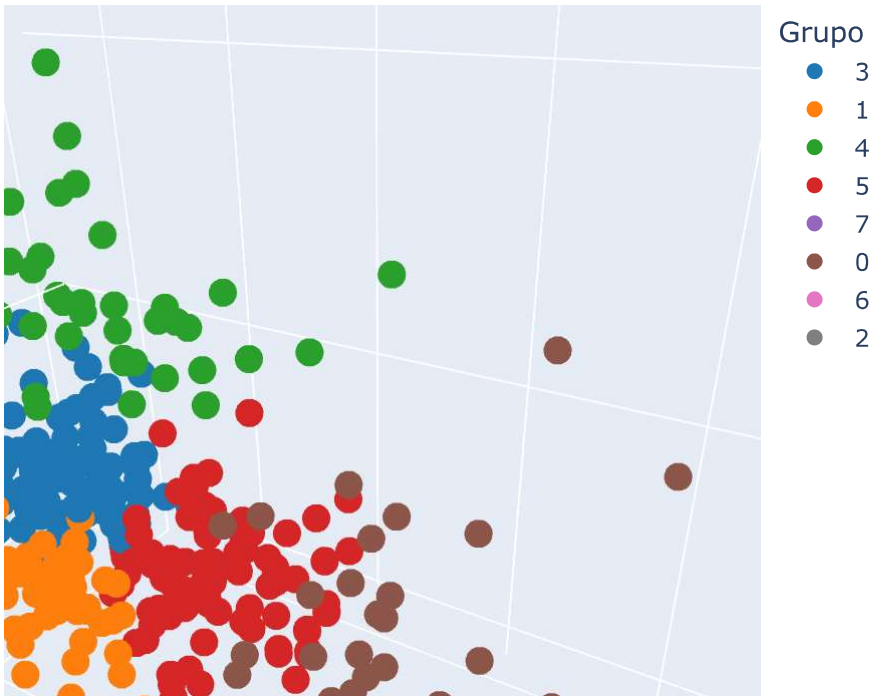
```
sns.pairplot(data=df2, hue='Grupo', palette='tab10')
plt.suptitle('Grupos', y=1.05)
```

```
Text(0.5, 1.05, 'Grupos')
```



```
# Grafica los grupos con un pairplot y con un scatterplot en 3D
# (si es necesario). Analiza las características de cada grupo.
# importar una librería más
import plotly.express as px
fig = px.scatter_3d(df2, x = 'User Rating', y = 'Price', z = 'Reviews',
                    title='8 grupos',
                    color='Grupo',
                    color_discrete_sequence=px.colors.qualitative.D3)
fig.show()
```

8 grupos



✓ 0 s se ejecutó 11:37

● ✕