Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus, Querétaro

Herramientas computacionales: el arte de la analítica (Gpo 201)

Actividad: K-means

Estudiantes:

Karen Cebreros López A01704254

Profesor:
Pedro Pérez

Fecha de entrega:
Jueves 12 de mayo del 2022

Análisis estadístico:

IMPORTAR LIBRERÍAS Y SUBIR EL ARCHIVO "bestsellers with categories.csv":

```
[1] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[2] 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])))
```

Choose Files bestsellers ...tegories.csv

- **bestsellers with categories.csv**(text/csv) - 51161 bytes, last modified: 5/8/2022 - 100% done

```
Saving bestsellers with categories.csv to bestsellers with categories.csv
User uploaded file "bestsellers with categories.csv" with length 51161 bytes
```

```
[8] df_bwc = pd.read_csv('bestsellers with categories.csv')
    df_bwc.head(5)
```

	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 Everything!) (Natio...	National Geographic Kids	4.8	7665	12	2019	Non Fiction

Tipos de datos

```
[5] df_bwc.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
```

Rango de los datos

```
[13] print("Rango de datos del 'User Rating': " + str(df_bwc['User Rating'].max() - df_bwc['User Rating'].min()))
     print("Rango de datos del 'Reviews': " + str(df_bwc['Reviews'].max() - df_bwc['User Rating'].min()))
     print("Rango de datos del 'Price': " + str(df_bwc['Price'].max() - df_bwc['User Rating'].min()))
     print("Rango de datos del 'Year': " + str(df_bwc['Year'].max() - df_bwc['User Rating'].min()))
```

```
Rango de datos del 'User Rating': 1.6000000000000005
Rango de datos del 'Reviews': 87837.7
Rango de datos del 'Price': 101.7
Rango de datos del 'Year': 2015.7
```

Crea una tabla resumen con los estadísticas generales de las variables numéricas

```
[4] df_bwc.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

¿Qué conclusiones puedes entregar de los datos?

- En general, podríamos decir que los libros cuentan con buenas calificaciones y reseñas por parte de los clientes, respecto a su contenido y precio. Pues en el caso

de las calificaciones, podemos ver que la media se encuentra entre 25% y el 50% de los datos, con un valor de 4.6.
Por otra parte, la media de las reseñas se encuentra entre el 50% y el 75% de los datos, al igual que la del precio.
La std del User Rating es de 0.22, lo que nos dice que los datos no se encuentran tan lejos de la media. (Para precio y reseñas esto no pasa tanto, por lo que nos dice que los datos están más dispersos).

```
[ ] df_temp = pd.DataFrame(df_bwc, columns = ['User Rating','Reviews','Price'])
    df_temp
```

	User Rating	Reviews	Price
0	4.7	17350	8
1	4.6	2052	22
2	4.7	18979	15
3	4.7	21424	6
4	4.8	7665	12
...	...	...	...
545	4.9	9413	8
546	4.7	14331	8
547	4.7	14331	8
548	4.7	14331	8
549	4.7	14331	8

550 rows × 3 columns

```
[ ] iris_corr = df_temp.corr()

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

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7bb9829150>
```

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

- Creo que las variables más relevantes para este análisis son aquellas que son cuantitativas y que sí se pueden correlacionar entre sí, como lo son la cantidad de reviews, el precio y las calificaciones de los usuarios.
  En este caso, yo creo que no es necesario tomar los años para el análisis porque es independiente de lo que ya se mencionó; por lo que la tomé como cualitativa y no relevante.

Análisis gráfico:

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

- Depende del gráfico que se requiera analizar, pero para las variables cuantitativas, yo quitaría el año porque no se relaciona con las calificaciones, reseñas y precios (es independiente el año en el que se publica el libro a todo esto).

Por otra parte, si quisiera analizar las calificaciones del usuario, dependiendo del género, quitaría el título y el autor.

Y por último, si quisiera analizar las reseñas en general por año, con el precio (para ver calidad-precio), quitaría las calificaciones individuales y las demás variables cualitativas.

¿Existen variables que tengan datos extraños?

- No.

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

- Las variables cuantitativas no tienen rangos similares, ya que se tratan cantidades muy diferentes. Por ejemplo… para el tema de las calificaciones de los usuarios, se trabaja con datos de 0 a 5, mientras que el número de reseñas, puede sobrepasar los miles dependiendo del libro.

  Yo creo que esto si afecta, ya que son valores muy diferentes y dispersos entre sí.

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

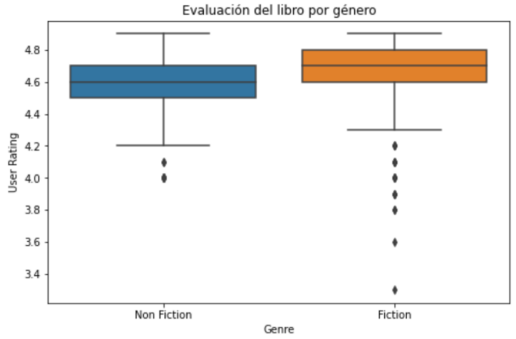
- Solamente el “User Rating” y el “Reviews” por la correlación representada en la parte de arriba (aunque ésta en sí no es muy grande y no cuenta en sí como positiva o negativa).

Gráficos para el análisis estadístico:

De caja y bigote:

```
[ ]  fig = plt.figure(figsize=(8,5))

     sns.boxplot(data=df_bwc, x = 'Genre', y = 'User Rating')

     plt.title('Evaluación del libro por género')
     plt.xlabel('Genre')
     plt.ylabel('User Rating')
```

```
Text(0, 0.5, 'User Rating')
```



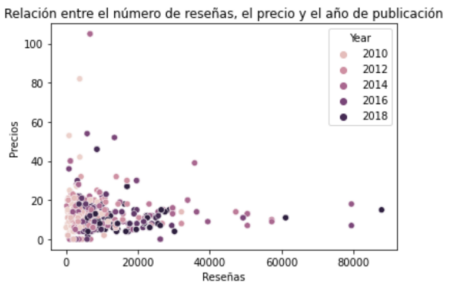
- En la gráfica anterior se puede ver la comparación entre las calificaciones de los usuarios y el género de los libros.

Podemos observar que el género de ficción, sobre pasa por bastante a los que no son de ficción.

De dispersión:

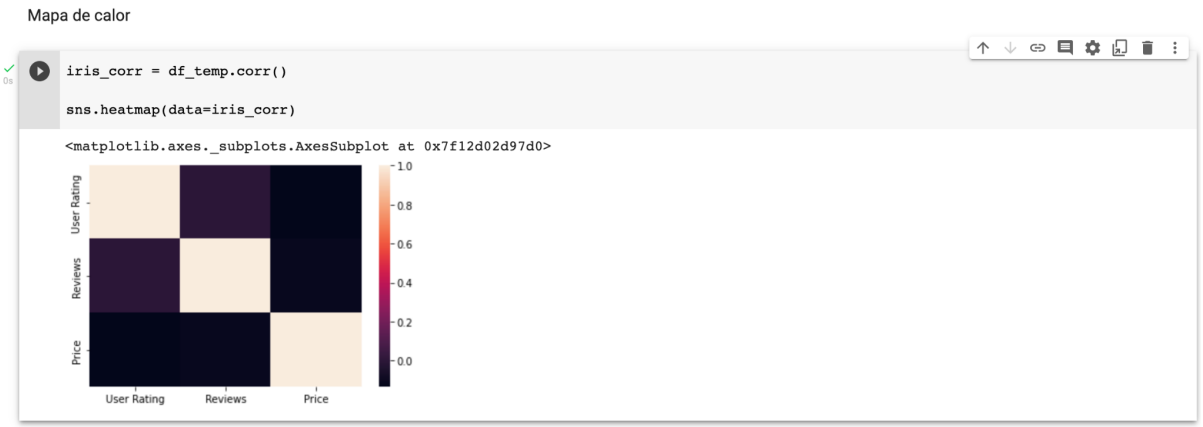
```
[ ]  fig = plt.figure(figsize=(6, 4))

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

     plt.title('Relación entre el número de reseñas, el precio y el año de publicación')
     plt.xlabel('Reseñas')
     plt.ylabel('Precios')
```

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



- En la gráfica de dispersión, comparé el número de reseñas con el precio (para ver calidad-precio) y clasificándose por el año de publicación; en donde podemos ver que en general, cada vez nos encontramos con datos un poco más dispersos.

De calor:



- Y por último, en el mapa de calor podemos ver la correlación entre las variables “user rating”, “reviews” y “price”.
  Como podemos observar, en realidad no hay gran correlación entre estas variables. Sin embargo, entre todas estas, las que muestran una correlación débil, pero no de 0, es la de “reviews” con “user rating”.

Conclusiones:

- Por las gráficas que yo analicé, quiero decir que el elegir bien las variables que se van a analizar entre sí es de vital importancia, ya que como se puede observar en la parte de arriba, muchas de ellas no presentan una relación muy fuerte y por ende podemos asumir que el analizarlas juntas no tiene mucho sentido.

  No obstante, usando los datos correctos y tipos de gráficas dependiendo a lo que se quiera analizar de estos, pude obtener cosas interesantes, como lo son la preferencia de género por las clasificaciones por el usuario.

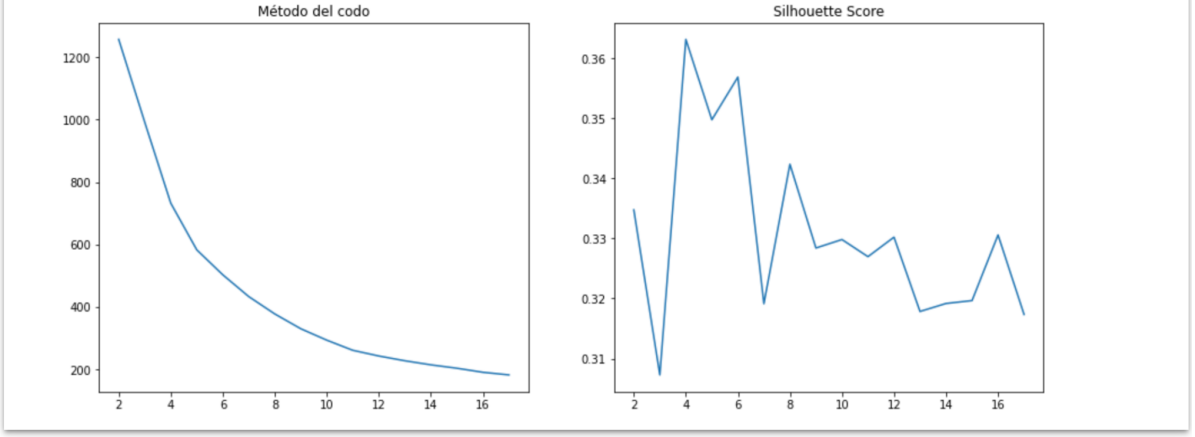
Clústering:

Algoritmo para diferentes valores de k

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

kmax   = 18
grupos = range(2, kmax)
wcss   = []
sil_score = []

# Ciclo para calcular K-Means para diferentes k
for k in grupos:
    # Clustering
    model = KMeans(n_clusters=k, random_state = 47)

    # Obtener las etiquetas
    clusters = model.fit_predict(X_norm)

    # Guardar WCSS
    wcss.append(model.inertia_)

    # Guardar Silhouette Score
    sil_score.append(silhouette_score(X_norm, clusters))
```

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

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')
```



- La kmax que yo utilicé fue de 18 porque al terminar de probar diferentes valores, vi que en ambos modelos coincidía con el pico y la caída (4) para determinar el número de grupos o clusters.

KMeans para los grupos

```
model = KMeans(n_clusters=4, random_state=47)
clusters = model.fit_predict(X_norm)

df_bwc['Grupo'] = clusters.astype('str')
df_bwc.head()
```

	Name	Author	User Rating	Reviews	Price	Year	Genre	Grupo
0	10-Day Green Smoothie Cleanse	JJ Smith	4.7	17350	8	2016	Non Fiction	0
1	11/22/63: A Novel	Stephen King	4.6	2052	22	2011	Fiction	2
2	12 Rules for Life: An Antidote to Chaos	Jordan B. Peterson	4.7	18979	15	2018	Non Fiction	0
3	1984 (Signet Classics)	George Orwell	4.7	21424	6	2017	Fiction	0
4	5,000 Awesome Facts (About Everything!) (Natio...	National Geographic Kids	4.8	7665	12	2019	Non Fiction	2

Centros de cada grupo

```
[44] df_bwc.groupby('Grupo').mean()
```

Grupo	User Rating	Reviews	Price	Year
0	4.727407	22135.555556	8.622222	2015.296296
1	4.616327	8457.918367	34.428571	2013.163265
2	4.243373	7951.469880	12.313253	2012.915663
3	4.566667	5719.666667	97.333333	2012.000000
4	4.693939	5903.750000	10.806818	2013.814394
5	4.412500	58490.375000	11.687500	2014.687500

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

- Yo creo que sí porque estamos hablando de un promedio de los valores y lo que aquí se conoce como la segmentación del comportamiento.

¿Cómo obtuviste el valor de k a usar?

- Al analizar la gráfica del “Silhouette Score”, vi que el pico más alto se presentaba en 4, por lo que sería el número de grupos más óptimo a utilizar.

  Utilizar una mayor cantidad de grupos podría entonces resultar contraproducente para el análisis y centros de los grupos.

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

- Si usamos una mayor cantidad de grupos no siendo necesario, el valor de los centros bajaría, haciendo que estos no fueran representativos de los datos. Y de la

misma manera, utilizar un número de grupos menor al óptimo presentado en el modelo, también generaría una representación errónea (más alta en centros), de esto.

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

- Tendríamos una mayor dispersión entre los grupos.
-

Análisis de las características de cada grupo:

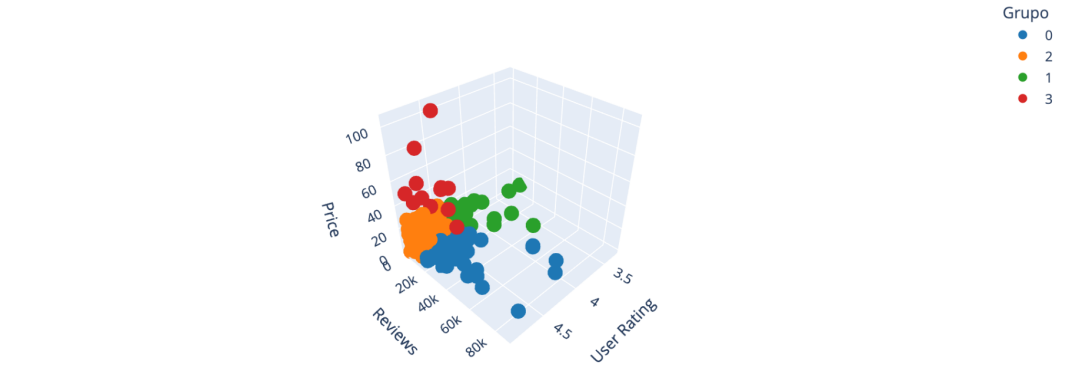
- El grupo 0 corresponde a las reseñas con precio bajo, con un rating medio y una cantidad de reseñas elevadas.
- El grupo 1 corresponde a las reseñas con precio bajo, con un rating medio y una cantidad de reseñas medio.
- El grupo 2 corresponde a las reseñas con precio alto, con un rating alto y una cantidad de reseñas baja.
- El grupo 3 corresponde a las reseñas con precio alto, con un rating bajo y una cantidad de reseñas bajo.

Gráfica de pairplot y scatterplot 3D

Pairplot:

Scatterplot 3D:

```
import plotly.express as px

fig = px.scatter_3d(df_bwc, x = 'User Rating', y = 'Reviews',
                    z = 'Price',
                    title='4 grupos de clientes',
                    color='Grupo',
                    color_discrete_sequence=px.colors.qualitative.D3)

fig.show()
```



- Podemos ver la división de los grupos entre las reseñas y las calificaciones de los usuarios.