Clase11_pandas_pelis

May 22, 2023

1 Seminario de Lenguajes - Python

1.1 Cursada 2023

1.1.1 Clase 11: analizamos películas

El material de esta clase fue desarrollado en el marco del proyecto **"Ciencia de Datos en la escuela"**.

Se puede descargar desde https://ciencia_datos_escuela.gitlab.io/, donde también hay más ejemplos.

Vamos a mostrar otra herramienta para graficar: plotly

Vamos a trabajar con: - https://www.kaggle.com/datasets/shivamb/netflix-shows

1.2 Preparando el dataset de Netflix

[14]:
```
import pandas as pd
import plotly.express as px
```

[4]:
```
netflix_completo = pd.read_csv("netflix_titles.csv")
```

[5]:
```
netflix_completo.columns
```

[5]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

[6]:
```
netflix_completo.head()
```

[6]:
```
  show_id     type                  title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water              NaN
2      s3  TV Show              Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show           Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States

```

```
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban…   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi…            NaN
3                                                 NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K…          India

         date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                        listed_in  \
0                                   Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act…
3                             Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV …

                                         description
0  As her father nears the end of his life, filmm…
1  After crossing paths at a party, a Cape Town t…
2  To protect his family from a powerful drug lor…
3  Feuds, flirtations and toilet talk go down amo…
4  In a city of coaching centers known to train I…
```

Vemos que hay filas que tienen valores NaN, lo que significa que no hay un valor, para lo cual usamos el método dropna que elimina las filas que contienen este tipo de valores.

[7]:
```
netflix_completo_sin_na = netflix_completo.dropna()
```

[8]:
```
netflix_completo_sin_na.head()
```

[8]:
```
   show_id     type                          title             director  \
7       s8    Movie                         Sankofa         Haile Gerima
8       s9  TV Show  The Great British Baking Show      Andy Devonshire
9      s10    Movie                    The Starling       Theodore Melfi
12     s13    Movie                    Je Suis Karl  Christian Schwochow
24     s25    Movie                           Jeans           S. Shankar

                                                 cast  \
7   Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D…
8   Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho…
9   Melissa McCarthy, Chris O'Dowd, Kevin Kline, T…
12  Luna Wedler, Jannis Niewöhner, Milan Peschel, …
24  Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi…
```

```
                                                country          date_added  \
7   United States, Ghana, Burkina Faso, United Kin…  September 24, 2021
8                                      United Kingdom  September 24, 2021
9                                       United States  September 24, 2021
12                           Germany, Czech Republic  September 23, 2021
24                                             India  September 21, 2021

    release_year rating   duration  \
7           1993  TV-MA    125 min
8           2021  TV-14  9 Seasons
9           2021  PG-13    104 min
12          2021  TV-MA    127 min
24          1998  TV-14    166 min

                                          listed_in  \
7   Dramas, Independent Movies, International Movies
8                       British TV Shows, Reality TV
9                                   Comedies, Dramas
12                      Dramas, International Movies
24   Comedies, International Movies, Romantic Movies

                                          description
7   On a photo shoot in Ghana, an American model s…
8   A talented batch of amateur bakers face off in…
9   A woman adjusting to life after a loss contend…
12  After most of her family is murdered in a terr…
24  When the father of the man she loves insists t…
```

2 Desafío 1

¿Cuál es la cantidad de películas y/o series estrenadas por año, que contiene la plataforma Netflix?

[9]:
```
contenido_por_anio = netflix_completo[["show_id","type","release_year"]]
contenido_por_anio.head()
```

[9]:
```
  show_id     type  release_year
0      s1    Movie          2020
1      s2  TV Show          2021
2      s3  TV Show          2021
3      s4  TV Show          2021
4      s5  TV Show          2021
```

Podemos agrupar por año de estreno y tipo (si es Película o show de TV) y contar la cantidad de elementos en cada subgrupo con **count**.

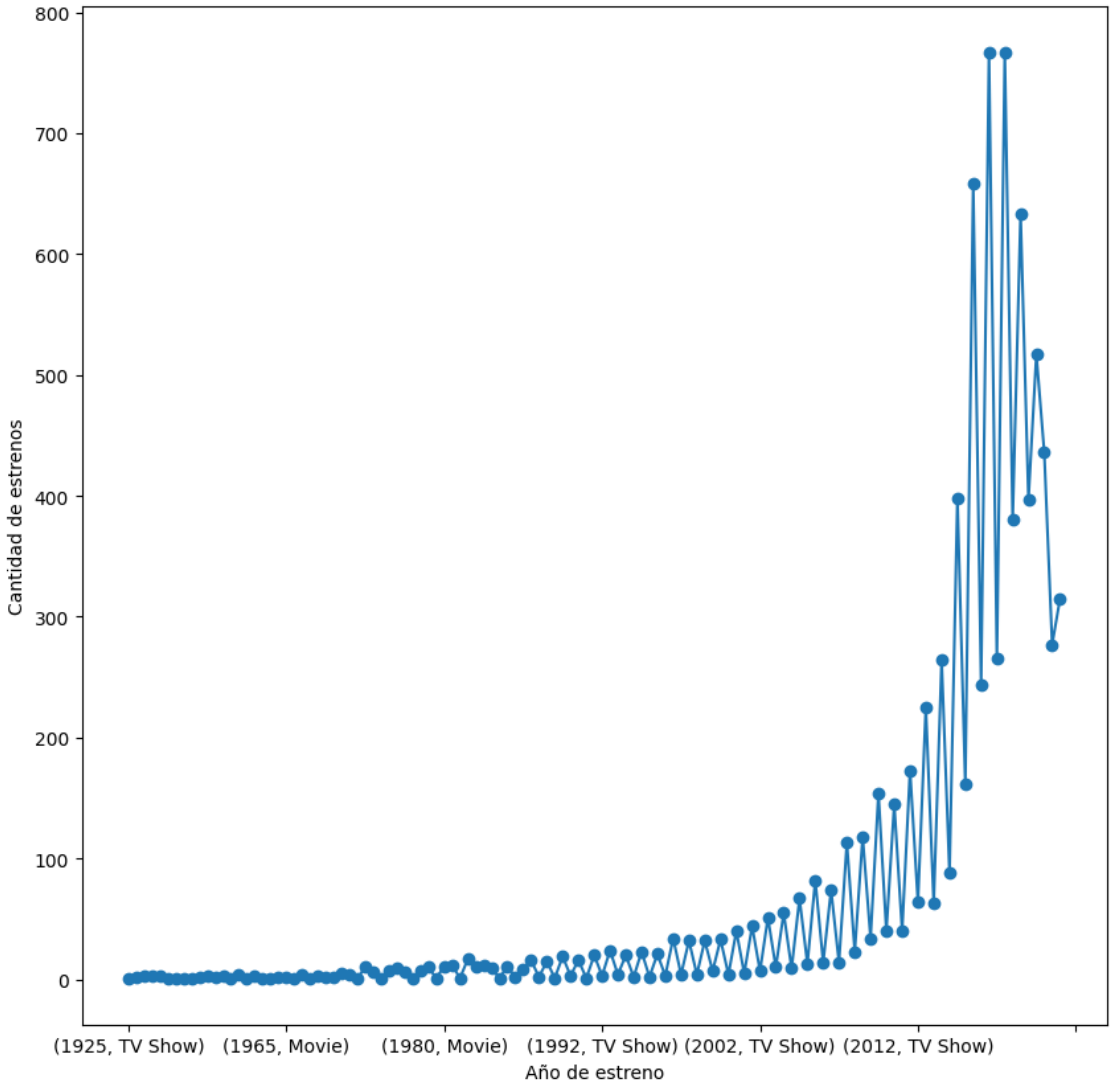
[10]:
```
contenido_por_anio = contenido_por_anio.
 ↪groupby(["release_year","type"])["show_id"].count()
contenido_por_anio.head(10)
```

[10]:
```
release_year  type
1925          TV Show    1
1942          Movie      2
1943          Movie      3
1944          Movie      3
1945          Movie      3
              TV Show    1
1946          Movie      1
              TV Show    1
1947          Movie      1
1954          Movie      2
Name: show_id, dtype: int64
```

3 Graficando con matplotlib

[12]:
```
contenido_por_anio.plot(figsize=(10,10), xlabel="Año de estreno",␣
 ↪ylabel="Cantidad de estrenos", marker="o")
```

[12]:
```
<Axes: xlabel='Año de estreno', ylabel='Cantidad de estrenos'>
```



4 Graficando con Plotly

En el caso de Plotly, usamos **plotly express**, una versión simplificada, muy parecida a matplotLib pero nos permite tener **gráficos interactivos**.

[15]:
```
fig= px.line(contenido_por_anio)
fig.show()
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[15], line 1
----> 1 fig= px.line(contenido_por_anio)
      2 fig.show()
```

```
File ~/ownCloud/Materias/Python/2023/entorno3.11/lib/python3.11/site-packages/
 ↪plotly/express/_chart_types.py:264, in line(data_frame, x, y, line_group,␣
 ↪color, line_dash, symbol, hover_name, hover_data, custom_data, text,␣
 ↪facet_row, facet_col, facet_col_wrap, facet_row_spacing, facet_col_spacing,␣
 ↪error_x, error_x_minus, error_y, error_y_minus, animation_frame,␣
 ↪animation_group, category_orders, labels, orientation,␣
 ↪color_discrete_sequence, color_discrete_map, line_dash_sequence,␣
 ↪line_dash_map, symbol_sequence, symbol_map, markers, log_x, log_y, range_x,␣
 ↪range_y, line_shape, render_mode, title, template, width, height)
    216 def line(
    217     data_frame=None,
    218     x=None,
   (…)
    258     height=None,
    259 ) -> go.Figure:
    260     """
    261     In a 2D line plot, each row of `data_frame` is represented as verte␣
 ↪of
    262     a polyline mark in 2D space.
    263     """
--> 264     return make_figure(args=locals(), constructor=go.Scatter)

File ~/ownCloud/Materias/Python/2023/entorno3.11/lib/python3.11/site-packages/
 ↪plotly/express/_core.py:1991, in make_figure(args, constructor, trace_patch,␣
 ↪layout_patch)
   1988 layout_patch = layout_patch or {}
   1989 apply_default_cascade(args)
-> 1991 args = build_dataframe(args, constructor)
   1992 if constructor in [go.Treemap, go.Sunburst, go.Icicle] and args["path"]␣
 ↪is not None:
   1993     args = process_dataframe_hierarchy(args)

File ~/ownCloud/Materias/Python/2023/entorno3.11/lib/python3.11/site-packages/
 ↪plotly/express/_core.py:1389, in build_dataframe(args, constructor)
   1387 if df_provided:
   1388     if isinstance(df_input.index, pd.MultiIndex):
-> 1389         raise TypeError(
   1390             "Data frame index is a pandas MultiIndex. "
   1391             "pandas MultiIndex is not supported by plotly express "
   1392             "at the moment."
   1393         )
   1394     args["wide_cross"] = df_input.index
   1395 else:

TypeError: Data frame index is a pandas MultiIndex. pandas MultiIndex is not␣
 ↪supported by plotly express at the moment.
```

5 ¿Qué pasó?

"**TypeError:** Data frame index is a pandas MultiIndex. **pandas MultiIndex is not supported** by plotly express at the moment."

La operación de agrupamiento genera un dataframe con más de un índice y esto no es soportado por plotly express.

Usamos el método unstack que retorna un nuevo dataframe con nuevos índice: ¿cuáles?

[16]:
```
contenido_por_anio = contenido_por_anio.unstack()
contenido_por_anio.head()
```

[16]:
```
type          Movie  TV Show
release_year
1925            NaN      1.0
1942            2.0      NaN
1943            3.0      NaN
1944            3.0      NaN
1945            3.0      1.0
```

6 Esto genera un nuevo problema

Hay valores NaN en aquellas interecciones que no faltan datos.

Por lo tando, podemos usar fillna con el valor 0, para indicar qué valores NaN o null deben ser llenados con 0.

[17]:
```
contenido_por_anio = contenido_por_anio.fillna(0)
contenido_por_anio.head()
```

[17]:
```
type          Movie  TV Show
release_year
1925            0.0      1.0
1942            2.0      0.0
1943            3.0      0.0
1944            3.0      0.0
1945            3.0      1.0
```

7 Ahora si podemos graficar

[19]:
```
fig= px.line(contenido_por_anio)
fig.show()
```

8 Seguimos la próxima ...