Netflix

En el presente trabajo se llevará a cabo un análisis de datos de Netflix utilizando un dataset obtenido de la plataforma de Kaggle. El conjunto de datos contiene información sobre películas y programas de televisión, con un total de 8791 registros. El objetivo principal es explorar diferentes aspectos del contenido (tipo de contenido, título, director, año, fecha que fue añadido, audiencia, duración, género y país de origen).

A lo largo del análisis, se crearán una serie de visualizaciones para examinar las distribuciones y relaciones entre estos datos.

Como primer paso, procederemos a cargar las librerias que necesitaremos a lo largo del trabajo

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotnine
from plotnine import *
import plotly.express as px
```

Lectura del dataset

Realizaremos la lectura de nuestro archivo que contiene el dataset de Netflix y comenzamos realizando una previsualización del contenido.

```
netflix=pd.read_csv ('C:\\Users\\celia\\OneDrive\\Escritorio\\DATA SCIENCE\\Visualizacion Avanzada\\netflix1.csv')

print(netflix.head())
```

```
                                     show_id   type  \
0                                         s1  Movie
1  s3,TV Show,Ganglands,Julien Leclercq,France,9/...    NaN
2  s6,TV Show,Midnight Mass,Mike Flanagan,United ...    NaN
3  s14,Movie,Confessions of an Invisible Girl,Bru...    NaN
4  s8,Movie,Sankofa,Haile Gerima,United States,9/...    NaN

                 title         director        country date_added  \
0  Dick Johnson Is Dead  Kirsten Johnson  United States  9/25/2021
1                   NaN              NaN            NaN        NaN
2                   NaN              NaN            NaN        NaN
3                   NaN              NaN            NaN        NaN
4                   NaN              NaN            NaN        NaN

   release_year rating duration      listed_in
0        2020.0  PG-13   90 min  Documentaries
1           NaN    NaN      NaN            NaN
2           NaN    NaN      NaN            NaN
3           NaN    NaN      NaN            NaN
4           NaN    NaN      NaN            NaN
```

Como podemos observar el conjunto de datos proporcionado contiene información sobre películas y programas de televisión disponibles en Netflix. Al examinar el dataset, identificamos un total de diez columnas, cada una con diferentes tipos de información. Algunas de estas columnas contienen datos completos y válidos mientras que otras presentan valores nulos (NaN)

Mostramos el tamaño de nuestro dataset. Consta de 8791 registros (filas) y 10 variables (columnas)

```
netflix.shape
```

```
(8791, 10)
```

A continuación mostramos un resumen estadístico y descriptivo de todas las columnas del dataset

```
netflix.describe(include = 'all')
```

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
count	8791	1760	1760	1760	1760	1760	1760.000000	1760	1760	1760
unique	8791	2	1760	920	37	875	NaN	11	146	26
top	s1	Movie	Dick Johnson Is Dead	Not Given	United States	1/1/2020	NaN	TV-MA	1 Season	Documentaries
freq	1	1246	1	570	1184	28	NaN	556	308	283
mean	NaN	NaN	NaN	NaN	NaN	NaN	2014.863636	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	7.548188	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	2014.000000	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	NaN	NaN	NaN

Resumimos con más detalle, para identificar la estructura con mayor claridad

```
]: print(netflix.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8791 entries, 0 to 8790
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   show_id       8791 non-null   object 
 1   type          1760 non-null   object 
 2   title         1760 non-null   object 
 3   director      1760 non-null   object 
 4   country       1760 non-null   object 
 5   date_added    1760 non-null   object 
 6   release_year  1760 non-null   float64
 7   rating        1760 non-null   object 
 8   duration      1760 non-null   object 
 9   listed_in     1760 non-null   object 
dtypes: float64(1), object(9)
memory usage: 686.9+ KB
None
```

Formateo de datos

A continuación realizaremos el formateo de los datos con el fin de mejorar su calidad y garantizar su precisión en el análisis. En primer lugar elimaremos las filas con valores nulos en columnas clave, asegurando que sólo se utilicen registros completos. Además eliminaremos espacios adicionales y reemplazaremos los valores inválidos en la columnas "show_id" que contienen comas. ¶

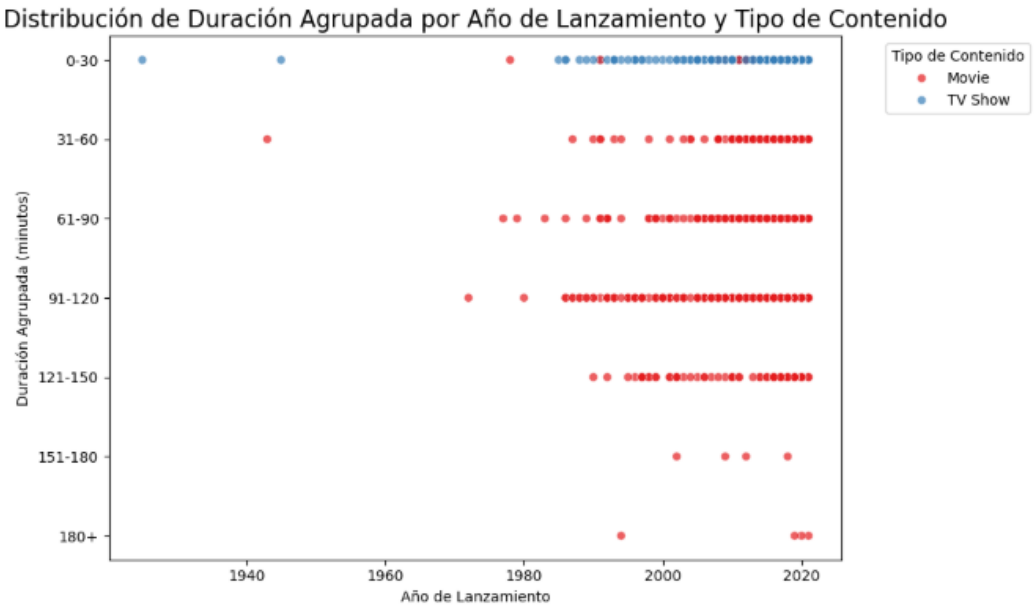
Este proceso da como resultado un conjunto de datos más limpio y consistente.

```
[9]: # Eliminamos filas con  valores NaN en las columnas citadas
netflix_cleaned = netflix.dropna(subset=['title', 'director', 'release_year', 'rating', 'type'])

# Limpiamos los espacios en las columnas especificadas
columns_to_strip = ['type', 'title', 'director', 'country']
for col in columns_to_strip:
    netflix_cleaned.loc[:, col] = netflix_cleaned[col].str.strip()

# Reemplazamos los registros con comas en 'show_id' por None
netflix.loc[netflix['show_id'].str.contains(',', na=False), 'show_id'] = None

# Rellenamos NaN en 'rating' con 'Unknown'
netflix_cleaned['rating'] = netflix_cleaned['rating'].fillna('Unknown')

# Limpiamos la columna 'duration' y convertimos a valores numéricos
netflix_cleaned['duration'] = netflix_cleaned['duration'].astype(str)
netflix_cleaned['duration'] = netflix_cleaned['duration'].str.extract('(\d+)').astype(float)

# Convertimos 'date_added' a datetime
netflix_cleaned['date_added'] = pd.to_datetime(netflix_cleaned['date_added'], errors='coerce')

# Convertimos 'release_year' a tipo entero
netflix_cleaned['release_year'] = netflix_cleaned['release_year'].astype('Int64')

# Creamos nuevas columnas 'release_month' y 'release_decade'
netflix_cleaned['release_month'] = netflix_cleaned['date_added'].dt.month
netflix_cleaned['release_decade'] = (netflix_cleaned['release_year'] // 10) * 10

# Visualizamos nuestro  dataframe limpio
print(netflix_cleaned.head())
```

```
   show_id     type                title          director        country  \
0       s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  United States
26   s7930    Movie       Samudri Lootere  Anirban Majumder      Not Given
30     s28    Movie            Grown Ups      Dennis Dugan  United States
32     s30    Movie             Paranoia    Robert Luketic  United States
34     s32  TV Show    Chicago Party Aunt         Not Given       Pakistan

   date_added  release_year rating  duration               listed_in  \
0  2021-09-25          2020  PG-13      90.0           Documentaries
26 2019-06-18          2018   TV-Y      65.0  Children & Family Movies
30 2021-09-20          2010  PG-13     103.0                Comedies
32 2021-09-19          2013  PG-13     106.0               Thrillers
34 2021-09-17          2021  TV-MA       1.0             TV Comedies

    release_month  release_decade
0               9            2020
26              6            2010
30              9            2010
32              9            2010
34              9            2020
```

Scatterplot duracion vs año lanzamiento

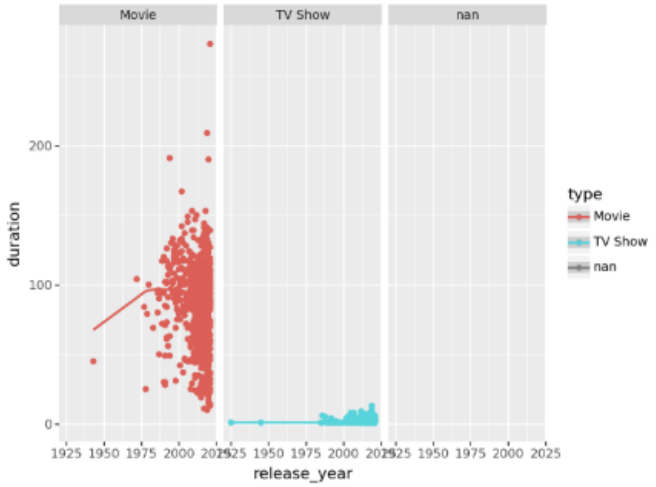
```
[6]: import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Nos aseguramos que la columna duration esta limpia
netflix['duration'] = netflix['duration'].str.extract('(\d+)').astype(float)

# Agrupamos la duración en intervalos de 30 minutos para una mejor visualizacion
bins = [0, 30, 60, 90, 120, 150, 180, 300]
labels = ['0-30', '31-60', '61-90', '91-120', '121-150', '151-180', '180+']
netflix['duration_grouped'] = pd.cut(netflix['duration'], bins=bins, labels=labels, right=False)

# Eliminamos filas con valores nulos en 'duration' o 'release_year'
netflix_cleaned = netflix.dropna(subset=['duration', 'release_year', 'type'])

# Creamos el gráfico scatterplot
plt.figure(figsize=(10, 6))
sns.scatterplot(data=netflix_cleaned,
                x='release_year',
                y='duration_grouped',
                hue='type',
                palette='Set1',
                alpha=0.7,
                edgecolor='w')


plt.title("Distribución de Duración Agrupada por Año de Lanzamiento y Tipo de Contenido", fontsize=16)
plt.xlabel("Año de Lanzamiento")
plt.ylabel("Duración Agrupada (minutos)")
plt.legend(title="Tipo de Contenido", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()

# Mostramos el gráfico
plt.show()
```



La gráfica muestra la distribución de los contenidos de Netflix en función del año de lanzamiento (eje X) y la duración (eje Y), diferenciados por tipo de contenido. Se observa que el período entre 2000 y 2020 es el de mayor presencia de contenido en la plataforma. En cuanto a la duración, los programas de televisión presentan un promedio cercano a los 30 minutos, lo cual se explica por la estructura de estos programas, que se dividen en episodios y no suelen tener una duración excesiva. Por otro lado, las películas de Netflix tienden a durar entre 1 y 2 horas, destacando especialmente el rango de 90 a 120 minutos.

```
from plotnine import ggplot, aes, geom_point, geom_smooth, facet_wrap
(
    ggplot(netflix)
    + aes(x = 'release_year',
          y = 'duration',
         color = 'type')
    + geom_point()
    + geom_smooth(method = 'lowess')
    + facet_wrap('~type')
)
```

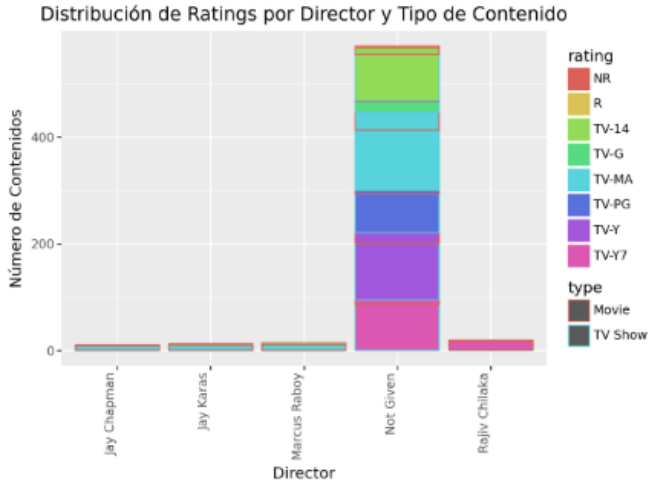
```
C:\Users\celia\anaconda3\Lib\site-packages\plotnine\stats\smoothers.py:347: PlotnineWarning: Confidence intervals are not yet implemented for lowess smo
othings.
C:\Users\celia\anaconda3\Lib\site-packages\plotnine\stats\smoothers.py:347: PlotnineWarning: Confidence intervals are not yet implemented for lowess smo
othings.
C:\Users\celia\anaconda3\Lib\site-packages\plotnine\layer.py:364: PlotnineWarning: geom_point : Removed 7031 rows containing missing values.
```



En esta gráfica de dispersión se analizan las diferencias en la duracion en minutos entre películas y programas de televisión, presentadas en dos gráficos separados para una mejor visualización. Se observa que las películas tienen una duración significativamente mayor.

Stacked Bar Chart

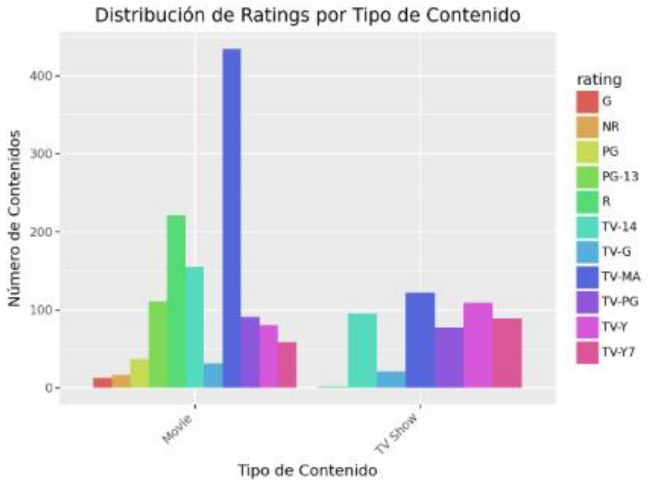
```
# Filtramos los 5 directores más frecuentes
top_directors = netflix['director'].value_counts().head(5).index
netflix_top_directors = netflix[netflix['director'].isin(top_directors)]

# Crear el gráfico con 'director', 'rating' y 'type'
ggplot(netflix_top_directors) + \
    aes(x='director', fill='rating', color='type') + \
    geom_bar(position='stack') + \
    theme(axis_text_x=element_text(rotation=90, hjust=1)) + \
    labs(title='Distribución de Ratings por Director y Tipo de Contenido',
         x='Director',
         y='Número de Contenidos')
```



En esta gráfica de barras apiladas se muestran los cinco directores más frecuentes. "Not Given" destaca por crear contenidos para todas las edades, incluyendo peliculas para adultos, programas para mayores de 14 años y para niños muy pequeños. Jay Chapman, Jay Karas y Marcus Rabboy han producido peliculas exclusivas para adultos, esto puede ser debido a que contengan contenido vulgar o escenas de violencia. Por último, Rajiv ha creado películas apropiadas para niños a partir de los 7 años.

Dodge Bar Chart

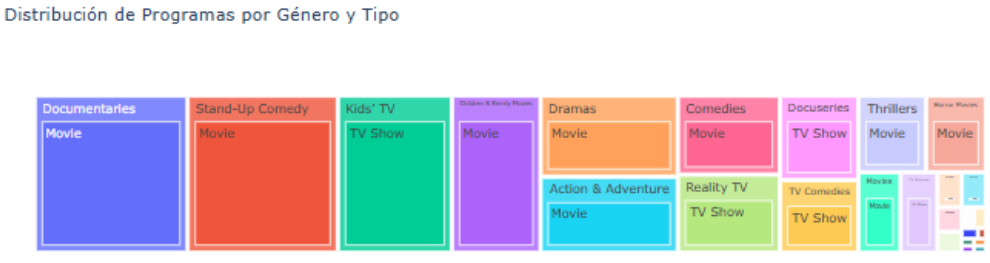
```
from plotnine import ggplot, aes, geom_bar, theme, element_text

# Filtramos los registros donde 'rating' y 'type' no sean NaN
netflix_cleaned = netflix.dropna(subset=['rating', 'type'])


ggplot(netflix_cleaned) + \
    aes(x='type', fill='rating') + \
    geom_bar(position='dodge') + \
    theme(axis_text_x=element_text(rotation=45, hjust=1)) + \
    labs(title='Distribución de Ratings por Tipo de Contenido',
         x='Tipo de Contenido',
         y='Número de Contenidos')
```



El gráfico muestra la distribución de los ratings por tipo de contenido en Netflix. Las películas predominan con clasificaciones para adultos y mayores de 14 años, mientras que los programas de televisión destacan por contenido para adultos y para niños. Esto refleja la diversidad en la oferta de Netflix, con un enfoque notable hacia audicencias adultas

Treemap distribucion de programas

```
import plotly.express as px

# Creamos una agrupación por 'listed_in' y 'type'
genre_counts = netflix.groupby(['listed_in', 'type']).size().reset_index(name='count')

fig = px.treemap(genre_counts,
                 path=['listed_in', 'type'],
                 values='count',
                 title="Distribución de Programas por Género y Tipo")
fig.show()
```

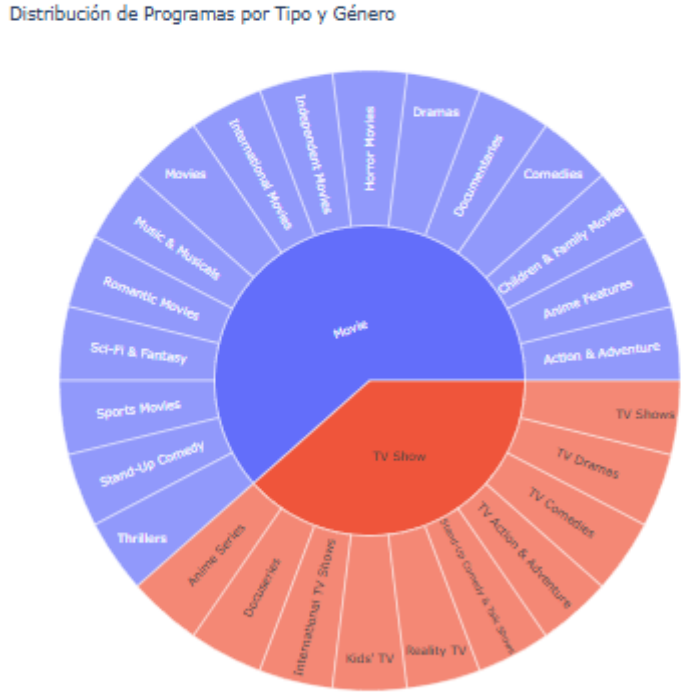


La gráfica muestra que los géneros con mayor cantidad de contenido en Netflix, son documentales, comedias y programas para niños y familias, los cuales destacan significativamente. En menor medida también se encuentran películas de drama, acción, aventura y reality shows. Los géneros con menor cantidad de contenido son peliculas de anime, deportes, fantasía y películas independientes, los cuales tienen una respresentación mucho más baja en comparación con los demás.

Esto refleja que netflix prioriza géneros de alta demanda como documentales, comedias y programas familiares, que tienen una audiencia amplia y constante. Además, producir estos géneros es más rentable y trascienden mejor a nivel global, en contraste con los géneros de nicho como anime o deportes.

Sunburst distribución por pais, tipo y género

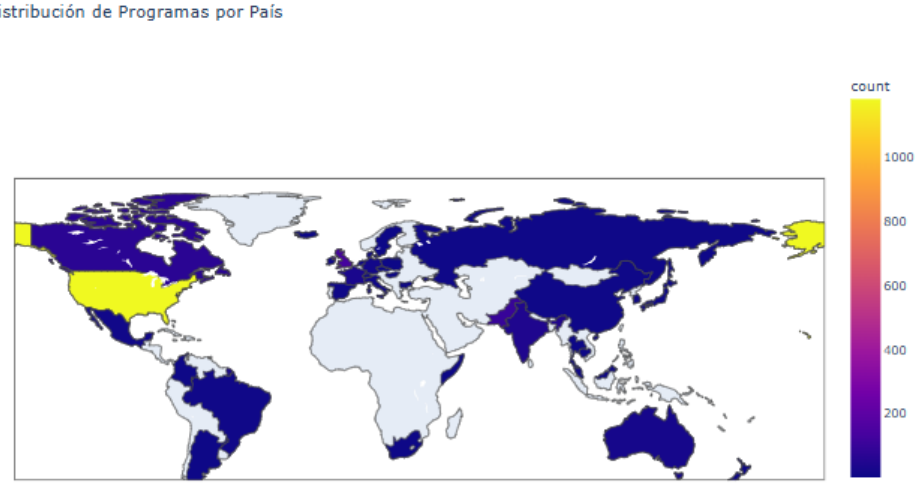
```
import plotly.express as px

# Limpiamos filas con valores nulos o duplicados en 'type' y 'listed_in'
netflix_clean = netflix.dropna(subset=['type', 'listed_in']).drop_duplicates(subset=['type', 'listed_in'])

# Creamos el grafico
fig = px.sunburst(netflix_clean, path=['type', 'listed_in'], title="Distribución de Programas por Tipo y Género")



# Ajustar el tamaño de la figura
fig.update_layout(
    width=800,   # Ancho de la figura
    height=800   # Alto de la figura
)

fig.show()
```



La distribución de contenidos en Netflix refleja una preponderancia de películas sobre programas de televisión. Dentro de las películas, se incluyen géneros como dramas, comedia y terror entre otros, mientras que en los programas de televisión se destacan categorías como anime, documentales,reality shows, docuseries, etc. Esta gráfica sugiere que la plataforma tiene una oferta variada, pero con un enfoque claro hacia el cine, especialmente en géneros populares. La menor proporción de contenido en la categoría de programas de televisión puede indicar una tendencia hacia una mayor producción cinematográfica.

Choropleth por país

```
0]: #Nos aseguramos tener la columna limpia
country_counts = netflix['country'].value_counts().reset_index()
country_counts.columns = ['country', 'count']

# Creamos el  mapa choropleth
fig = px.choropleth(country_counts,
                    locations='country',
                    locationmode='country names',
                    color='count',
                    hover_name='country',
                    title="Distribución de Programas por País")



# Ajustar el tamaño de la figura
fig.update_layout(
    width=1000,   # Ancho de la figura
    height=600    # Alto de la figura
)

fig.show()
```



Del mapa mundial observamos que el país con mayor cantidad de contenido en Netflix es Estados Unidos, destacando notablemente sobre el resto. En contraste, se identifican regiones como Africa, Portugal, Groenlandia, Finlandia, Ucrania y algunos paises de Asia donde no hay presencia de Netflix.

Esto evidencia una distribución desigual del contenido, con una clara concentración en ciertas áreas geográficas y ausencia en otras. Esto puede deberse a restricciones legales,licencias y la menor demanda de Netflix en esas zonas.

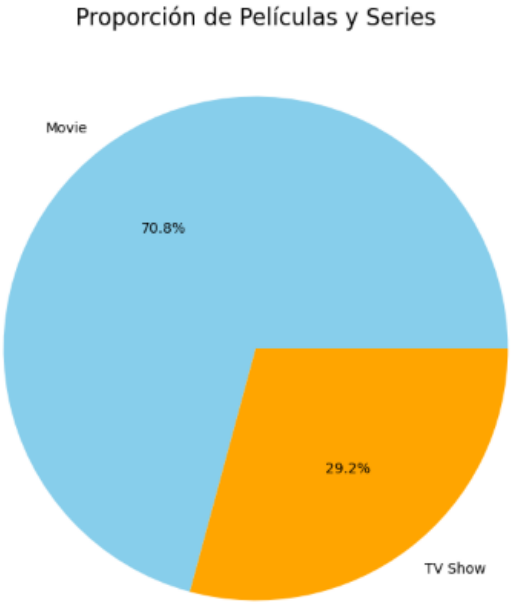
Gráfico de lineas con los programas que se añadieron por año

```
import pandas as pd
import matplotlib.pyplot as plt

# Convertimos la columna 'date_added' a formato datetime
netflix['date_added'] = pd.to_datetime(netflix['date_added'], errors='coerce')

# Extraemos el año de la columna 'date_added'
netflix['year_added'] = netflix['date_added'].dt.year

# Contamos la cantidad de programas añadidos por año
added_by_year = netflix['year_added'].value_counts().sort_index()

# Creamos el gráfico de línea
plt.figure(figsize=(10, 6))
plt.plot(added_by_year.index, added_by_year.values, marker='o', color='green', linestyle='-')
plt.title('Tendencia de Programas Añadidos por Año', fontsize=16)
plt.xlabel('Año')
plt.ylabel('Cantidad de Programas')
plt.grid(True)
plt.show()
```

La gráfica de linea muestra una tendencia de crecimiento notable en la cantidad de programas de Netflix desde el año 2015, alcanzando su punto máximo en 2019. A partir de ese año, se observa una ligera disminución, aunque no significativa. Sin embargo, la cantidad de programas se mantiene estable por encima de los 300, indicando una estabilidad en la producción de contenido tras el periodo de mayor crecimiento.

Pie Chart según el tipo de contenido ¶

```
netflix['type'].value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8, 8), colors=['skyblue', 'orange'])
plt.title('Proporción de Películas y Series', fontsize=16)
plt.ylabel('')
plt.show()
```

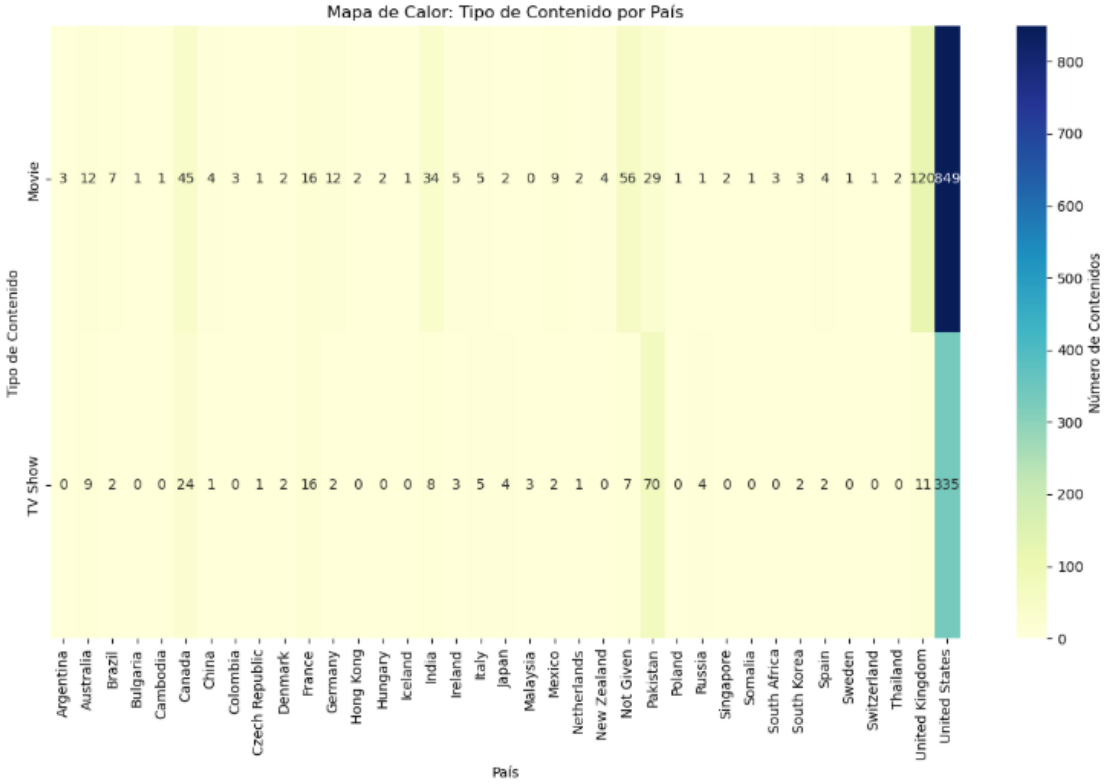


La gráfica Pie Chart muestra que el catálogo de Netflix está compuesto mayoritariamente por películas que representan un 70,8% del total, mientras que los programas de televisión constituyen un 29,2% restante.

Esto muestra que Netflix prioriza las películas en su catálogo, aunque mantiene una oferta significativa de programas de televisión para diversificar y atraer a distintas audiencias.

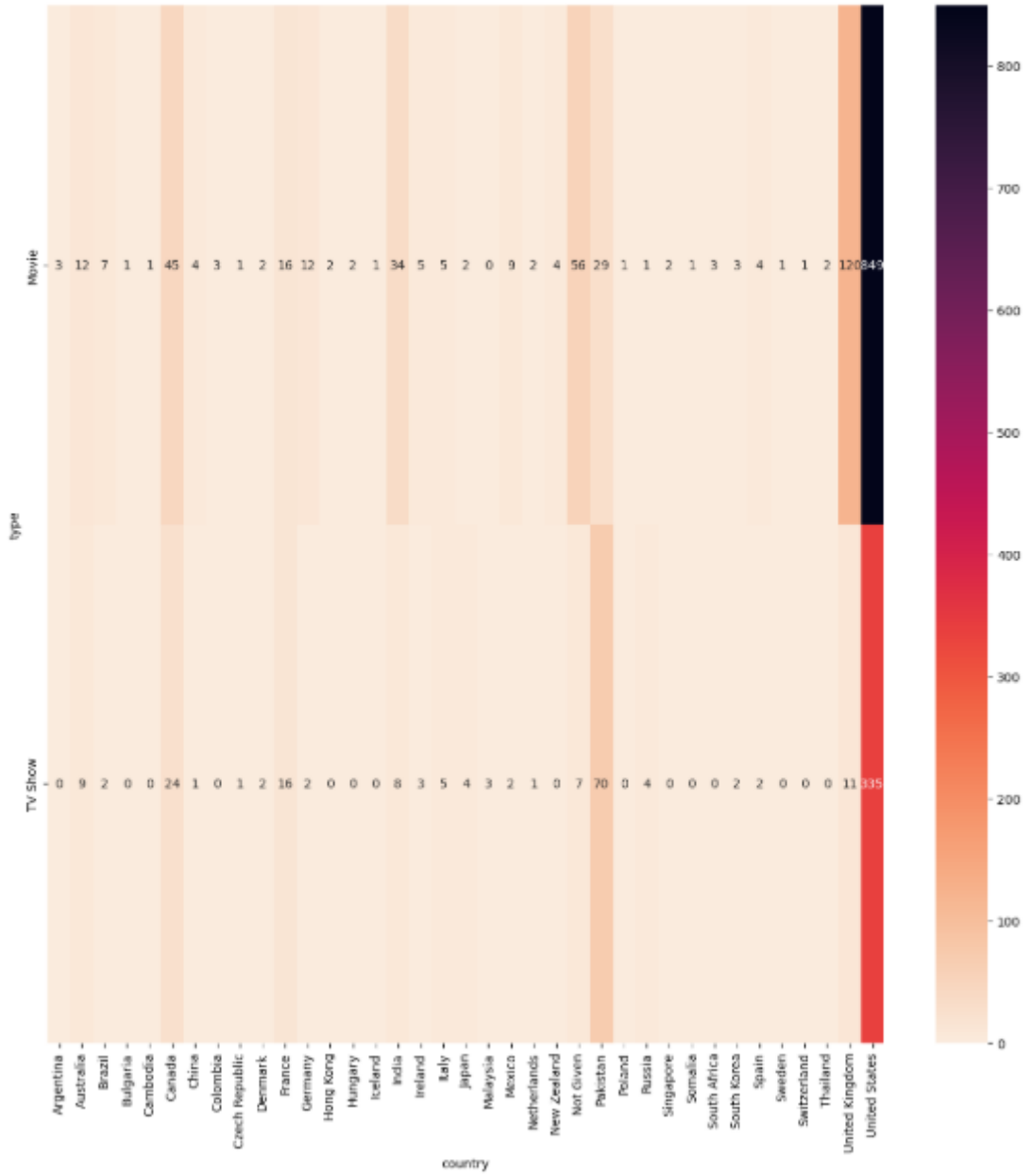
Mapa de Calor

```
# Creamos una tabla de contingencia entre 'type' y 'country'
heatmap_data = pd.crosstab(netflix['type'], netflix['country'])

# Creamos el heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(heatmap_data, annot=True, fmt='d', cmap='YlGnBu', cbar_kws={'label': 'Número de Contenidos'})

# Títulos y etiquetas
plt.title('Mapa de Calor: Tipo de Contenido por País')
plt.xlabel('País')
plt.ylabel('Tipo de Contenido')

# Mostramos el gráfico
plt.tight_layout()
plt.show()
```



La gráfica presenta un mapa de calor que muestra la distribución de contenido de Netflix según el tipo y el país. Se observa que Estados Unidos destaca significativamente en cuanto al número de contenidos, con más de 120,000 programas. En contraste, países como Bulgaria o Malaysia tienen una cantidad considerablemente menor. Esta distribución sugiere que Netflix tiene una mayor presencia en países como Estados Unidos y Reino Unido, mientras que otros países poseen una oferta más reducida, especialmente en cuanto a películas y series. En esta gráfica se muestra exactamente el número de contenido de cada tipo.

```
#Creamos otro mapa con dimension de 15x15
fig, ax = plt.subplots(figsize=(15, 15))
#Definimos nuestra paleta por colores
paleta2 = sns.color_palette("rocket_r", as_cmap=True)
#Creamos nuestro heatmap es base al tipo y país
ax = sns.heatmap(pd.crosstab(netflix['type'], netflix['country']),
                 cmap=paleta2, annot=True, fmt="g")
```



Conclusiones

El análisis de este dataset de Netflix revela una clara preferencia por las películas, que constituyen el 70,8% del catálogo, mientras que los programas de televisión representan un 29,2%. Los géneros más frecuentes son documentales, comedias y contenido familiar, lo que refleja la estrategia de ofrecer contenido de amplia demanda y accesible para diversas audiencias. Por otro lado, los géneros de nicho, como anime y deportes, tienen una representación menor.

Geográficamente, Estados Unidos destaca como país con mayor cantidad de contenido, mientras que algunas regiones como África y ciertas partes de Asia, muestran una presencia mínima o nula de Netflix. Este patrón podría deberse a restricciones de licencias o menor demanda de esas áreas.

En cuanto a la producción se observa un crecimiento significativo en el número de programas entre 2015 y 2019, seguido de una estabilización. En general, Netflix ha diversificado su oferta, priorizando películas y contenido para un público amplio, aunque con variaciones en la disponibilidad por regiones.