Proyecto POO - Netflix

Netflix, Inc. es una empresa de entretenimiento y una plataforma de streaming estadounidense que participa en la producción de series y películas y en la distribución de contenido multimedia a través de su plataforma de streaming. Actualmente, Netflix es la empresa de medios de comunicación más valiosa en el mundo. Más de 117 millones de hogares de pagaban de manera periódica para ver series y películas.

Netflix ofrece una gran cantidad de contenidos a los usuarios de su plataforma, tanto de formato de lanzamiento, pasar y transcribir la información de un contenido (formato en el cual se pueden ver series y películas).

Hoy las empresas se enfrentan a retos cada vez más exigentes, informados, que demandan experiencias que les demuestren y que satisfacen de acuerdo a sus necesidades y gustos de manera rápida.

Por lo anterior, se pretende que para el mes de junio, se tenga la información de los contenidos disponibles en la plataforma de Netflix, de acuerdo a sus especificaciones, de la cual, se pueda extraer información de la información digital como la puntuación de los contenidos, el número de veces que se ha visto un contenido, el país en el que se ha visto un contenido, el tipo de contenido (si es una serie o una película), entre otros, que permita ofrecer a los usuarios de la plataforma Netflix un mejor servicio de acuerdo a sus gustos y preferencias. Para lo anterior, se pretende que los usuarios puedan ver más horas de entretenimiento, y de ahí, poder ofrecer recomendaciones efectivas como decir el tipo de lo que los usuarios verán de la plataforma para poder incrementar el consumo dentro de ella.

Introducción

El dataset a analizar Netflix, ofrece una cantidad de información de los contenidos disponibles en la plataforma de Netflix, recopilando de la información de los contenidos que se ofrecen en los diferentes países en los que la empresa tiene presencia. Cada contenido tiene información de su nombre, el director, el país en el que se ha producido, el tipo de contenido (si es una serie o una película), entre otros.

Con los cinéfilos interesados de "Net gusta" (porque también a "Pie me uno a" porque hay dudas), Netflix logra "enganchar" una audiencia de 117 millones de hogares en 190 países.

- La empresa ofrece una amplia gama de contenidos, que incluye series, documentales y películas, que los socios compran a grandes estudios para el catálogo del artículo y también para los que son producidos por ella mismo.
- Los contenidos se ofrecen a los usuarios de forma personalizada, lo que significa que cada usuario puede ver contenidos diferentes en función de sus gustos y preferencias.
- Los usuarios pueden ver los contenidos en cualquier momento y desde cualquier dispositivo conectado a Internet, lo que significa que pueden ver contenidos de forma cómoda y flexible.
- Netflix utiliza algoritmos de aprendizaje automático para analizar los datos de los usuarios y ofrecer recomendaciones personalizadas a cada usuario, lo que ayuda a que los usuarios encuentren contenidos que les gusten y que les mantengan interesados en la plataforma.
- La calidad de los contenidos es una de las características más importantes para Netflix, ya que la empresa busca ofrecer a sus usuarios contenidos de alta calidad que les ofrezcan una experiencia de entretenimiento satisfactoria.
- La competencia en el mercado de los servicios de streaming es muy alta, con muchas empresas compitiendo por la atención de los usuarios. Netflix se enfrenta a la competencia de otras empresas como Amazon Prime Video, Disney+, HBO Max y otras plataformas de streaming que ofrecen contenidos similares a los de Netflix.

Los datos recopilados que serán analizados corresponden a datos de la información de los contenidos disponibles en la plataforma de Netflix, en el mes de junio del 2021. Los datos a analizar corresponden a información de los contenidos disponibles en la plataforma de Netflix, en el mes de junio del 2021.

Los datos corresponden a información de los contenidos disponibles en la plataforma de Netflix, en el mes de junio del 2021.

Insights a analizar (a priori)

Los insights que se pueden extraer a partir de la información de los contenidos disponibles en la plataforma de Netflix, en el mes de junio del 2021, son:

- Tipos de contenidos más populares: Netflix ofrece una amplia variedad de contenidos, incluyendo series, documentales y películas. Los datos pueden ser útiles para determinar qué tipos de contenidos son más populares entre los usuarios de la plataforma y cómo pueden cambiar a lo largo del tiempo.
- Países con más contenidos: Netflix está disponible en muchos países en todo el mundo, y los datos pueden ser útiles para determinar qué países tienen más contenidos disponibles y cómo pueden cambiar a lo largo del tiempo.
- Duración de los contenidos: La duración de los contenidos puede ser un factor importante para los usuarios de Netflix, ya que algunos pueden preferir contenidos más cortos mientras que otros pueden preferir contenidos más largos. Los datos pueden ser útiles para determinar la duración promedio de los contenidos y cómo pueden cambiar a lo largo del tiempo.
- Calificaciones de los contenidos: Netflix ofrece a sus usuarios la posibilidad de calificar los contenidos que ven, y los datos pueden ser útiles para determinar qué contenidos son más populares y cómo pueden cambiar a lo largo del tiempo.
- Contenidos más vistos: Los datos pueden ser útiles para determinar qué contenidos son más vistos por los usuarios de Netflix, y cómo pueden cambiar a lo largo del tiempo.
- Contenidos más recomendados: Netflix ofrece a sus usuarios recomendaciones de contenidos basadas en sus gustos y preferencias. Los datos pueden ser útiles para determinar qué contenidos son más recomendados y cómo pueden cambiar a lo largo del tiempo.
- Contenidos más populares en diferentes países: Los datos pueden ser útiles para determinar qué contenidos son más populares en diferentes países y cómo pueden cambiar a lo largo del tiempo.
- Tiempo promedio de visualización: Los datos pueden ser útiles para determinar el tiempo promedio que los usuarios pasan viendo contenidos en la plataforma de Netflix y cómo pueden cambiar a lo largo del tiempo.
- Tasa de cancelación: Los datos pueden ser útiles para determinar la tasa de cancelación de los usuarios de Netflix y cómo puede cambiar a lo largo del tiempo.
- Contenidos más populares en diferentes dispositivos: Los datos pueden ser útiles para determinar qué contenidos son más populares en diferentes dispositivos y cómo pueden cambiar a lo largo del tiempo.

1. Adentándonos al dataset

```
# Cargar paquetes
library(dplyr)
library(ggplot2)

# Cargar el dataset de netflix
# Especificar la ruta del archivo
netflix_path <- "netflix_titles.csv"
# Cargar el dataset usando read.csv()
netflix_data <- read.csv(netflix_path)
# Imprimir las primeras filas del dataset
head(netflix_data)
```

```
  show_id  type title dir cast country date_added release_year rating duration
1      s1 Movie  NaN NaN  NaN     NaN  9-Jul-19         2020  TV-PG   1 min
2      s2 Movie  NaN NaN  NaN     NaN  8-Jul-19         2020  TV-PG   1 min
3      s3 Movie  NaN NaN  NaN     NaN  8-Jul-19         2020  TV-PG   1 min
4      s4 Movie  NaN NaN  NaN     NaN  8-Jul-19         2020  TV-PG   1 min
5      s5 Movie  NaN NaN  NaN     NaN  8-Jul-19         2020  TV-PG   1 min
6      s6 Movie  NaN NaN  NaN     NaN  8-Jul-19         2020  TV-PG   1 min
  listed_in                                                   description
1       NaN                                                           NaN
2       NaN                                                           NaN
3       NaN                                                           NaN
4       NaN                                                           NaN
5       NaN                                                           NaN
6       NaN                                                           NaN
Warning message:
In read.table(file = file, header = header, sep = sep, quote = quote,
# Instalar el paquete dplyr si no está instalado
# Cargar el paquete dplyr
library(dplyr)
# Ver la estructura del dataset
str(netflix_data)
```

```
'data.frame':	7787 obs. of  12 variables:
 $ show_id     : chr "s1" "s2" "s3" "s4" ...
 $ type        : chr "Movie" "Movie" "Movie" "Movie" ...
 $ title       : chr "NaN" "NaN" "NaN" "NaN" ...
 $ dir         : chr "NaN" "NaN" "NaN" "NaN" ...
 $ cast        : chr "NaN" "NaN" "NaN" "NaN" ...
 $ country     : chr "NaN" "NaN" "NaN" "NaN" ...
 $ date_added  : chr "9-Jul-19" "8-Jul-19" "8-Jul-19" "8-Jul-19" ...
 $ release_year: int 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
 $ rating      : chr "TV-PG" "TV-PG" "TV-PG" "TV-PG" ...
 $ duration    : chr "1 min" "1 min" "1 min" "1 min" ...
 $ listed_in   : chr "NaN" "NaN" "NaN" "NaN" ...
 $ description : chr "NaN" "NaN" "NaN" "NaN" ...
```

2. Características principales del Dataset

2.1 Información General y Estadística

En primer lugar, se pretende ver cuál es la información con la que se cuenta en el dataset y su dimensionalidad. Es importante ver si hay ausencia de información, ya que, si hay ausencia, es decir, si hay vacíos en la data, el dataset podría ser menos deseado en comparar. Para ello, se pretende ver la información con la que se cuenta en el dataset y su dimensionalidad, para ver si es deseable o no.

Entonces, siguiendo la metodología de CRISP-DM, se pretende ver la información con la que se cuenta en el dataset y su dimensionalidad:

- **info()**: nos ordena mayor información sobre un DataFrame, incluido el tipo de índice y las columnas, los valores no nulos y el uso de la memoria.
- **shape()**: devuelve dimensiones descriptivas (n-tuples): filas, columnas, ejes. Por ello, en ocasiones puede ser conveniente de agregar una variable que contenga la información de la forma del DataFrame, ya que, es una operación que puede ser costosa en términos de cómputo (memoria). Por ejemplo, "shape" puede tomar el valor de una tupla que contiene los valores a lo largo de cada eje del DataFrame.
- **dtypes()**: devuelve una serie con el dtype de cada columna.

```
# Ver información del dataset
str(netflix_data)
# Ver dimensiones del dataset
dim(netflix_data)
# Ver tipos de datos
sapply(netflix_data, class)
# Ver valores nulos
colSums(is.na(netflix_data))
# Ver valores únicos
sapply(netflix_data, function(x) length(unique(x)))
# Ver resumen estadístico
summary(netflix_data)
```

```
'data.frame':	7787 obs. of  12 variables:
 $ show_id     : chr "s1" "s2" "s3" "s4" ...
 $ type        : chr "Movie" "Movie" "Movie" "Movie" ...
 $ title       : chr "NaN" "NaN" "NaN" "NaN" ...
 $ dir         : chr "NaN" "NaN" "NaN" "NaN" ...
 $ cast        : chr "NaN" "NaN" "NaN" "NaN" ...
 $ country     : chr "NaN" "NaN" "NaN" "NaN" ...
 $ date_added  : chr "9-Jul-19" "8-Jul-19" "8-Jul-19" "8-Jul-19" ...
 $ release_year: int 2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
 $ rating      : chr "TV-PG" "TV-PG" "TV-PG" "TV-PG" ...
 $ duration    : chr "1 min" "1 min" "1 min" "1 min" ...
 $ listed_in   : chr "NaN" "NaN" "NaN" "NaN" ...
 $ description : chr "NaN" "NaN" "NaN" "NaN" ...
NULL
 show_id    type   title     dir    cast ... rating duration listed_in
 7787 7787    2    5335    2389    6831 ...   8     3     1     1
description
   1
show_id    type   title     dir    cast ... rating duration listed_in
   0       0       0       0       0 ...      0       0       0       0
description
   0
show_id    type   title     dir    cast ... rating duration listed_in
    7       2    5335    2389    6831 ...      8    3.000       1       1
description
   1.000
show_id    type   title     dir    cast ... rating duration listed_in
   0.000   0.000   0.000   0.000   0.000 ...  0.000   0.000   0.000   0.000
description
   0
```

Luego de haber cargado los datos de la tabla y llevado a cabo el proceso de feature engineering, los variables con las que cuenta el dataset ahora son las siguientes:

Nombre variable	Tipo	Descripción
show_id	object	Id que se le asigna a cada película o serie de televisión
type	object	Indica si la película o serie es de tipo Film o TV Show / Miniserie
title	object	Título de la obra audiovisual
director	object	Nombre del director/s de la obra audiovisual
cast	object	Nombre de los actores/actrices que participan en la obra audiovisual
country	object	Indica el país de producción
date_added	object	Indica la fecha en la que la película o serie se agregó a la plataforma
release_year	int64	Indica el año de lanzamiento de la película o serie
rating	object	Indica la clasificación de edad de la película o serie
listed_in	object	Indica el género o los géneros de la película o serie
description	object	Indica una breve descripción de la película o serie
month_added	int64	Indica el mes en el que la película o serie se agregó a la plataforma
year_added	int64	Indica el año en el que la película o serie se agregó a la plataforma
day_added	int64	Indica el día en el que la película o serie se agregó a la plataforma
duration_int	int64	Indica la duración de la película o serie en minutos. Para la serie Breaking Bad, indica la duración de la temporada
duration_type	object	Indica si la duración de la película o serie está medida en minutos o temporadas
title_length	int64	Indica la cantidad de caracteres que tiene el título de la película o serie
lag_release	int64	Indica la cantidad de años que tuvo que esperar la película o serie para ser lanzada

3.1 Datos duplicados y faltantes, nulos (NaN)

Es muy usual que los registros tenderán de un dataset generado más de una vez, por lo cual, para no llegar a obtener resultados, es necesario eliminar los duplicados de la base de datos.

```
print("La cantidad de datos duplicados a la hora de hacer el dataset es:")
df.duplicated().sum()
```

```
La cantidad de datos duplicados a la hora de hacer el dataset es:
0
```

Por otro lado, dependiendo de el porcentaje de nulos que se tenga en la base de datos, se debe elegir correctamente el enfoque a utilizar para no perder información valiosa. Para ver cuántos nulos tiene cada variable, se puede correr el siguiente código. A continuación se analiza la presencia de datos nulos/faltantes (NaN) en cada variable:

```
print("Cantidad de nulos por variable:")
df.isnull().sum()
```

```
Cantidad de nulos por variable:
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
listed_in       0
description     0
month_added     0
year_added      0
day_added       0
duration_int    0
duration_type   0
title_length    0
lag_release     0
dtype: int64
```

El resultado anterior muestra que no hay nulos en ninguna variable.

Por lo tanto, para cada una de las variables, se ve que la cantidad de datos faltantes (nulos, valores largos y vacíos, duration_int=0, year_added=0 y release_year=0) es igual a cero. Por lo tanto, se puede decir que no hay datos faltantes (ningún valor negativo o un valor indefinido/sin un valor) en ninguna variable.

Los valores de "year_added" y "release_year" no tienen sentido que sean iguales a cero, por lo que, más adelante, se deberá tener en cuenta que se trata de valores nulos. Por lo tanto, se deberá tener en cuenta que se trata de valores nulos.

Por otro lado, se puede observar que la cantidad de datos faltantes (nulos, valores largos y vacíos, duration_int=0, year_added=0 y release_year=0) es igual a cero. Por lo tanto, se puede decir que no hay datos faltantes (ningún valor negativo o un valor indefinido/sin un valor) en ninguna variable.

Los valores de "year_added" y "release_year" no tienen sentido que sean iguales a cero, por lo que, más adelante, se deberá tener en cuenta que se trata de valores nulos. Por lo tanto, se deberá tener en cuenta que se trata de valores nulos.

```
df.info()
```

```
<class_name_dataset_netflix_clean_dataset>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 19):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      8807 non-null   object
 4   cast          8807 non-null   object
 5   country       8807 non-null   object
 6   date_added    8807 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8807 non-null   object
 9   listed_in     8807 non-null   object
 10  description   8807 non-null   object
 11  month_added   8807 non-null   int64
 12  year_added    8807 non-null   int64
 13  day_added     8807 non-null   int64
 14  duration_int  8807 non-null   int64
 15  duration_type 8807 non-null   object
 16  title_length  8807 non-null   int64
 17  lag_release   8807 non-null   int64
dtype: int64
```

El dataset cuenta entonces con 19 variables (de 19 en total) con datos nulos.

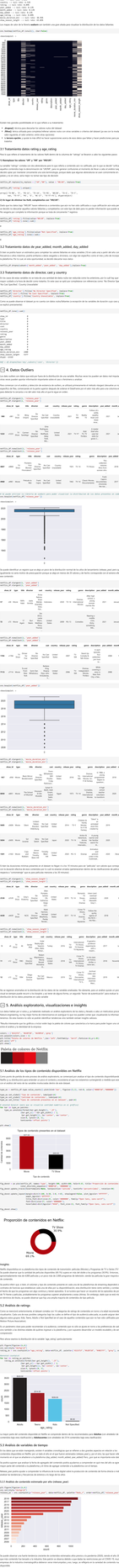
Por lo tanto, luego de haber corrido nuevamente el .info() se ve que la cantidad de datos faltantes (nulos, valores largos y vacíos, duration_int=0, year_added=0 y release_year=0) es igual a cero. Por lo tanto, se puede decir que no hay datos faltantes (ningún valor negativo o un valor indefinido/sin un valor) en ninguna variable.

Los valores de "year_added" y "release_year" no tienen sentido que sean iguales a cero, por lo que, más adelante, se deberá tener en cuenta que se trata de valores nulos. Por lo tanto, se deberá tener en cuenta que se trata de valores nulos.

Los tipos de datos de la tabla **netflix** son también los que deben usar para visualizar la distribución de los datos faltantes.

```
sns.heatmap(df.isnull(), cbar=False)
```

```
<AxesSubplot:>
```



Como nos gustaría profundizarlo en qué sucede ahora y no importarlo:

- **Atributo**: Nos da el nombre del atributo con valores nulos.
- **Nulos**: Nos da la cantidad de valores nulos que tiene la variable/atributo.
- **Porcentaje de nulos (%)**: Nos da el porcentaje de valores nulos con respecto a la cantidad de datos que tiene.
- **Tipo de dato**: Nos da el tipo de dato que tiene la variable/atributo.

Lo cual, nos va a ayudar a tener un panorama de cómo nos van a tener que ser de utilidad los datos faltantes (nulos) para poder seguir con el trabajo.

3.1 Tratamiento datos vacíos y que valen cero

Para la parte siguiente se importan las librerías "missingno" y "pandas_profiling".

1 Reemplazar los valores "0" por "NaN" para "duration"

La variable "duration" contiene nos dice absolutamente todo que debería contestarse con la película que ha pasado de tener "vertical" hacer "horizontal" y que ahora nos permite ver la información de la misma. Es decir, es la variable que nos dice cuánto dura la película o serie. Por lo tanto, es una variable muy importante para poder hacer una descripción de cada producción de la plataforma. Por lo cual, el valor que contiene el dataset, se decide descartar esos datos.

```
df[df['duration_int'] == 0]
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

2 Cantidad de filas y columnas de la tabla netflix_clean_dataset

```
df.shape
```

```
(8807, 19)
```

3 Tipo de datos de la tabla netflix_clean_dataset

```
df.dtypes
```

```
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
listed_in       object
description     object
month_added      int64
year_added       int64
day_added        int64
duration_int     int64
duration_type   object
title_length     int64
lag_release      int64
dtype: object
```

4 Tratamiento datos de director, cast y country

Lo que nos ayuda a entender que no se tienen en cuenta cuántas filas se tienen en cuenta como los artistas, por lo cual hay que tener mayor cuidado a la hora de decidir como trataría. Es una cosa que se ve por comparación con referencias como "Dr. Disney Channel Spain" y "Dr. Disney Channel".

```
df[df['director'] == 'Dr. Disney Channel Spain']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

Como la parte siguiente es el mismo que lo anterior pero sobre visualizar la distribución de los valores faltantes por cada variable en el dataset proporcionado:

```
df[df['cast'] == 'Dr. Disney Channel']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

```
df[df['country'] == 'Dr. Disney Channel']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

3.4 Datos Outliers

Los datos outliers son datos que están fuera de la distribución de los valores. Por lo tanto, no pueden ser datos que hayamos para tomar como referencia a la hora de hacer predicciones o tomar decisiones.

Para poder visualizar los datos outliers, se puede utilizar el siguiente código. Lo que nos va a ayudar a entender cuáles son los valores específicos de filas, columnas en cantidad de registros que tienen el dataset y cuál es el valor más alto para una variable específica para detectar la presencia de outliers.

```
df[df['duration_int'] == 'release_year']
df[df['duration_int'] == 'lag_release']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

```
df[df['duration_type'] == 'release_year']
df[df['duration_type'] == 'lag_release']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

```
df[df['title_length'] == 'release_year']
df[df['title_length'] == 'lag_release']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

```
df[df['lag_release'] == 'release_year']
df[df['lag_release'] == 'lag_release']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

```
df.describe(include=['object'])
```

```
C:\Users\Lucia Cecilia\Anaconda3\lib\site-packages\numpy\lib\function_base.py:4180: FutureWarning: Arrays of bytes/strings is deprecated, array.astype(np.unicode_) will give a UnicodeDecodeError when any element of the array is not ASCII.
  arr = asanyarray(a)
C:\Users\Lucia Cecilia\Anaconda3\lib\site-packages\numpy\lib\function_base.py:4180: FutureWarning: Arrays of bytes/strings is deprecated, array.astype(np.unicode_) will give a UnicodeDecodeError when any element of the array is not ASCII.
  arr = asanyarray(a)
```

	show_id	type	title	director	cast	country	date_added	rating	listed_in	description
count	8807	8807	8807	8807	8807	8807	8807	8807	8807	8807
unique	8807	2	8807	4528	7692	748	1767	14	514	8775
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	TV-MA	Documentaries	Documentary about the life and work of David Attenborough.
freq	1	6131	1	19	19	2818	118	2863	334	3

```
df[df['duration_type'] == 'Minutes']
df[df['duration_type'] == 'Seasons']
```

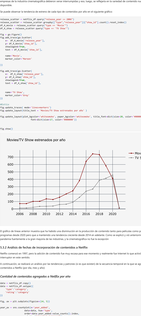
```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

```
df[df['duration_type'] == 'Minutes']
df[df['duration_type'] == 'Seasons']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

```
sns.boxplot(data=df, x='duration_int')
```

```
<AxesSubplot:>
```



```
df[df['duration_int'] == 'duration_int']
df[df['duration_int'] == 'duration_int']
```

```
Empty DataFrame
Columns: [show_id, type, title, director, cast, country, date_added, release_year, rating, listed_in, description, month_added, year_added, day_added, duration_int, duration_type, title_length, lag_release]
Index: []
```

El 2021 fue particularmente un año de alta producción de contenido para ambas categorías, con más de 600 títulos sumados entre las publicadas, seguido por el 2020 en la segunda posición, dato en el que también la producción de películas se acerca a la que fue finalmente el máximo histórico registrado en nuestro registro, correspondiente a los 500 títulos 2022.

Cantidad de contenido agregado a Netflix por mes

```
cant_agregados_por_mes = (
    df_netflix
    .groupby(['month_added'])
    .agg(
        cantidad = ('title', 'count')
    )
    .reset_index()
)

px.bar(cant_agregados_por_mes, x='month_added', y='cantidad',
       title='Contenido agregado por mes',
       color='cantidad',
       color_continuous_scale=px.colors.sequential.Viridis,
       labels={'month_added': 'Mes', 'cantidad': 'Cantidad'},
       template='plotly_dark')
```

Observo que la cantidad crece de año en año, siendo diciembre el mes en el que se agrega más de la ganancia por parte de empresas de Hollywood, con una cantidad de contenido que duplica a la de los meses de Octubre, Octubre y Enero (seguido de primeros tres porcentajes de ventas de productos de consumo masivo, como ya mencionamos en el informe de Amazon Prime Video). En el mismo Gran que se acompañan por Rentabilidad Agosto que se comparten con Septiembre y Noviembre, siendo Febrero el mes con menor cantidad de contenido a ser incluido en nuestro catálogo.

Cantidad de contenido agregado a Netflix por día de la semana

```
dia_de_la_semana = df_netflix['day_of_week']

px.histogram(dia_de_la_semana,
             nbins=7,
             title='Contenido agregado por día de la semana',
             color_discrete_sequence=['#00CC96'],
             labels={'value': 'Día de la semana', 'count': 'Cantidad'},
             template='plotly_dark')
```

Me resulta curioso que la distribución de contenidos incorporados a Netflix en relación a meses y días, también se mantiene sobrevivo a pesar de que dentro de que en los días también predomina el Viernes, siendo el día en que se suman más contenidos a la plataforma, seguido por el Jueves y el Sábado, siendo el Domingo el día en que se suman menos contenidos. Es decir, que los contenidos el día de semana y días de fiesta y ganan en menor parte definido de un buen momento, y que la empresa de Hollywood, que también es dueña de Warner Bros, no tiene en cuenta a la hora de publicar contenido.

Por todo lo que acabamos de ver, es notable que la cantidad de contenido de tipo Movie es mucho mayor que la de contenido de tipo TV Show, lo que nos lleva a preguntarnos si es que Netflix tiene una estrategia de adquisición de contenido enfocada en ofrecer a sus clientes una gran variedad de opciones de entretenimiento, o si es que Netflix, a pesar de ser una empresa de tecnología, sigue siendo una empresa de Hollywood que, al igual que sus competidores, prefiere producir más películas que series. Para ello, en primer lugar, me gustaría ver cuál es la cantidad de contenido de cada una de las categorías a lo largo de los años.

1.1.1 Tiempo transcurrido desde el momento de lanzamiento y de la incorporación a Netflix

```
px.scatter(df_netflix,
           x='release_year',
           y='year_added',
           color='type',
           size='duration_num',
           hover_name='title',
           title='Tiempo transcurrido desde el momento de lanzamiento del contenido a su incorporación en Netflix',
           labels={'release_year': 'Año de lanzamiento',
                   'year_added': 'Año de adición',
                   'type': 'Tipo',
                   'duration_num': 'Duración'},
           template='plotly_dark')
```

Como podemos ver, la cantidad de series crece en 2020 frente a la cantidad de películas, y 2020 es la categoría donde 50% de las series la grande diferencia con la gran mayoría de contenidos contenidos agregados, en grande proporciones, salvo cuando en 2016-2017.

1.4 Análisis de contenidos por país

Ya hicimos un poco de Netflix Ya generamos cómo los contenidos con una amplia cantidad de países diferentes del mundo. En cambio ahí fue los contenidos disponibles en cada país, pueda mostrar nosotros y queríamos ustedes ningún país como conocer tres sobre ó mejor dicho, tres países.

```
df_netflix['country_1'] = df_netflix['country'].fillna('').str.split(',').str[0].str.strip()

df_country_1 = df_netflix[df_netflix['country_1'].isin(['United States', 'India', 'United Kingdom'])]

px.sunburst(df_country_1,
            path=['country_1', 'type', 'rating'],
            color='rating',
            color_discrete_sequence=px.colors.qualitative.Pastel,
            title='Top 3 países con contenidos de Netflix',
            template='plotly_dark')
```

```
df_country_1.shape
df_country_1.sample(2)
df_country_1.isnull().sum
```

```
px.barrel(df_country_1, x='country_1', y='count', color='type', barmode='group',
          color_discrete_sequence=['#00CC96', '#EF553B'],
          title='Top 3 países con contenidos de Netflix',
          labels={'country_1': 'País', 'count': 'Cantidad', 'type': 'Tipo'},
          template='plotly_dark')
px.sunburst(df_country_1,
            path=['country_1', 'type', 'rating'],
            color='rating',
            color_discrete_sequence=px.colors.qualitative.Pastel,
            title='Top 3 países con contenidos de Netflix',
            template='plotly_dark')
```

Los dos países más prolíficos en contenidos para Netflix son los siguientes: EE. UU. con 7.000 y Reino Unido, y los diferencia con la India, que tiene 4.000 contenidos. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido.

¿Hay cada vez más la proporción de programas de TV y películas en cada uno de estos países?

Si bien Estados Unidos es el país que más contenidos tiene, también es el que más contenidos tiene. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido.

```
df_country_1 = df_country_1[df_country_1['release_year'] >= 2000]

px.sunburst(df_country_1,
            path=['country_1', 'type', 'release_year'],
            color='release_year',
            color_continuous_scale=px.colors.sequential.Viridis,
            title='Top 3 países con contenidos de Netflix',
            template='plotly_dark')
```

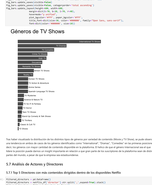
1.4 Análisis de contenidos por país

Ya hicimos un poco de Netflix Ya generamos cómo los contenidos con una amplia cantidad de países diferentes del mundo. En cambio ahí fue los contenidos disponibles en cada país, pueda mostrar nosotros y queríamos ustedes ningún país como conocer tres sobre ó mejor dicho, tres países.

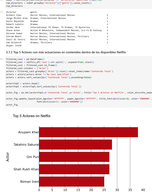
```
px.sunburst(df_country_1,
            path=['country_1', 'type', 'rating'],
            color='rating',
            color_discrete_sequence=px.colors.qualitative.Pastel,
            title='Top 3 países con contenidos de Netflix',
            template='plotly_dark')
```

```
df_country_1.shape
df_country_1.sample(2)
df_country_1.isnull().sum
```

```
px.barrel(df_country_1, x='country_1', y='count', color='type', barmode='group',
          color_discrete_sequence=['#00CC96', '#EF553B'],
          title='Top 3 países con contenidos de Netflix',
          labels={'country_1': 'País', 'count': 'Cantidad', 'type': 'Tipo'},
          template='plotly_dark')
px.sunburst(df_country_1,
            path=['country_1', 'type', 'release_year'],
            color='release_year',
            color_continuous_scale=px.colors.sequential.Viridis,
            title='Top 3 países con contenidos de Netflix',
            template='plotly_dark')
```



Los dos países más prolíficos en contenidos para Netflix son los siguientes: EE. UU. con 7.000 y Reino Unido, y los diferencia con la India, que tiene 4.000 contenidos. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido.



Los dos países más prolíficos en contenidos para Netflix son los siguientes: EE. UU. con 7.000 y Reino Unido, y los diferencia con la India, que tiene 4.000 contenidos. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido.

¿Hay cada vez más la proporción de programas de TV y películas en cada uno de estos países?

Si bien Estados Unidos es el país que más contenidos tiene, también es el que más contenidos tiene. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido. Es decir, que Estados Unidos tiene casi el doble de contenidos que la India y casi el doble de contenidos que Reino Unido.

1.4 Análisis de la duración de cada tipo de contenido

Si nos estamos para dentro de lo que refiere a duración, existen dos posibilidades dependiendo del tipo de contenido: en minutos (Movie) o en temporadas (TV Show). Para ello vamos a crear dos dataframes, uno para cada tipo de contenido y hacer uno de los gráficos.

Tendencia en la duración de los programas de TV

```
df_tv = df_netflix[df_netflix['type'] == 'TV Show']

px.scatter(df_tv, x='release_year', y='duration_num',
           color='rating',
           size='duration_num',
           hover_name='title',
           trendline='lowess',
           title='Tendencia en la duración de los programas de TV',
           labels={'release_year': 'Año de lanzamiento',
                   'duration_num': 'Temporadas',
                   'rating': 'Rating'},
           template='plotly_dark')
```

Duración de las películas

```
df_movie = df_netflix[df_netflix['type'] == 'Movie']

px.scatter(df_movie, x='release_year', y='duration_num',
           color='rating',
           size='duration_num',
           hover_name='title',
           trendline='lowess',
           title='Duración de las películas',
           labels={'release_year': 'Año de lanzamiento',
                   'duration_num': 'Minutos',
                   'rating': 'Rating'},
           template='plotly_dark')
```

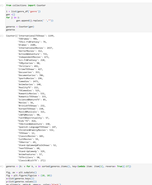

La duración de la mayoría de las películas de Netflix rondaría los 90 y 100 minutos, lo que probablemente sea de un tiempo de que cualquier persona que tenga una suscripción pueda ver sin problemas. Lo que sí resulta curioso es que hay una cantidad de contenidos más alta en los años 70 y 80 que en los años 90 y 2000.

1.5 Análisis de ratings de cada tipo de contenido

Netflix ha ofrecido a sus suscriptores una amplia gama de ratings para que se entiendan dentro cuestiones y tragedias o disconvenientes, en todos ellos, en el caso de contenidos destinados para todo tipo de audiencia.

Algo interesante que se acaba precisamente, es que los tipos para los contenidos cambiaron y sólo ganaron, pero ya no quedan ejemplares de los tipos de contenidos que fueron cambiados.

```
px.histogram(df_netflix,
             x='rating',
             color='type',
             barmode='group',
             title='Ratings por tipo de contenido',
             labels={'rating': 'Rating',
                     'type': 'Tipo',
                     'count': 'Cantidad'},
             template='plotly_dark')
```



5.0 Análisis de las descripciones de los contenidos

Después de una Definición de Análisis 1.1.2 (para proveer datos descriptivos Proporciona una PDF simple para comprender en forma cronológica del procesamiento de lenguaje natural (NLP), como el ofrecimiento de partes del discurso, lexicoanálisis de frases nominales, el análisis de sentimientos y la extracción de entidades. Se puede usar para procesar documentos de texto en varios idiomas, extraer objetos significativos y analizar la sintaxis del uso de diccionarios de palabras y poesía. Aquí hemos aplicado una quedan a utilizarlo la función de análisis de sentimientos para ver si hay alguna tendencia en los contenidos de los títulos de los contenidos y si hay alguna tendencia sentimiento basados en regiones?

En una aproximación a los focos en el análisis de sentimientos para poder determinar a contener insights en lo que refiere a los títulos de los contenidos y si hay alguna tendencia en los contenidos.

```
# Use Sentiment Analysis Functions
df['Sentiment_pd'] = df['Release_year'].apply(lambda x: sentiment_pd(df['Description'][x], df['Release Year'][x]))
df['Sentiment_pd'] = df['Sentiment_pd'].apply(lambda x: x[0])

for index, row in df.iterrows():
    if df['Sentiment'][index] == 'Positive':
        df['Sentiment'][index] = 1
    elif df['Sentiment'][index] == 'Neutral':
        df['Sentiment'][index] = 0
    else:
        df['Sentiment'][index] = -1
    if df['Sentiment_pd'][index] > 0.5:
        df['Sentiment_pd'][index] = 1
    elif df['Sentiment_pd'][index] < -0.5:
        df['Sentiment_pd'][index] = -1
    else:
        df['Sentiment_pd'][index] = 0
    df['Sentiment_pd'][index] = df['Sentiment_pd'][index]

df['Sentiment_df'] = df['Sentiment_df'].astype('int')
df['Sentiment_df'] = df['Sentiment_df'].astype('int')
df['Sentiment_fbp'] = df['Sentiment_fbp'].astype('int')
df['Sentiment_fbp'] = df['Sentiment_fbp'].astype('int')
df['Sentiment_fbp'] = df['Sentiment_fbp'].astype('int')
```

```
# Time Sentiment Analysis
df['Sentiment_df'] = df['Sentiment_df'].astype('int')
df['Sentiment_fbp'] = df['Sentiment_fbp'].astype('int')
df['Sentiment_fbp'] = df['Release Year'].astype('int')

sns.set_theme(style="ticks")
sns.pairplot(df, hue="Sentiment", vars=["Sentiment_pd", "Sentiment_df", "Sentiment_fbp"])
```

Análisis de sentimiento con textblob

Podemos ver como a medida que avanzan los años se convierte en una tendencia de contenido estrenados que tienenen una proporcion relativamente uniforme en lo que refiere a contenido positivo, negativo y neutro. Hacia para a medida que nos acercamos al momento actual, se observa una mas predominancia hacia tener mas tipo "positivo".

```
# Save Sentiment Analysis to .csv
df.to_csv('./sentiment.csv')
```

```
[Sentiment Analysis] ./sentiment.csv successfully saved!
```