Analysis on Netflix Movies & TV Shows

Netflix is a popular service that people across the world use for entertainment. In this EDA, I will explore the netflix-shows dataset through visualizations and graphs using matplotlib and seaborn.

Package Install and Import

First, we wil install and import necessary packages

```
import jovian
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import matplotlib
```

```
netflix_titles_df = pd.read_csv('netflix_titles.csv')
netflix_titles_df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentar
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Internatio TV Shows, Dramas, Myste
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime Sho Internatio TV Shows, A
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuser Reality
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Internatio TV Sho Romantic Shows, T

Data Prepartion and Cleaning

```
netflix_titles_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   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      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

There are 8,807 entries and 12 columns to work in this EDA. Right off the bat there are a few columns that contain null values ('director','cast','country','rating').

```
netflix_titles_df.nunique()
```

```
show_id         8807
type               2
title           8807
director        4528
cast            7692
country          748
date_added      1767
release_year      74
rating            17
duration         220
listed_in        514
description     8775
dtype: int64
```

Handling Null Values

We can see that for each of the columns,there are a lot different unique values for some of them. It makes sense that show_id is large it is a unique key used to identify a movie/show. title, director, cast, country, date_added, listed_in, and description contain many unique values as well.

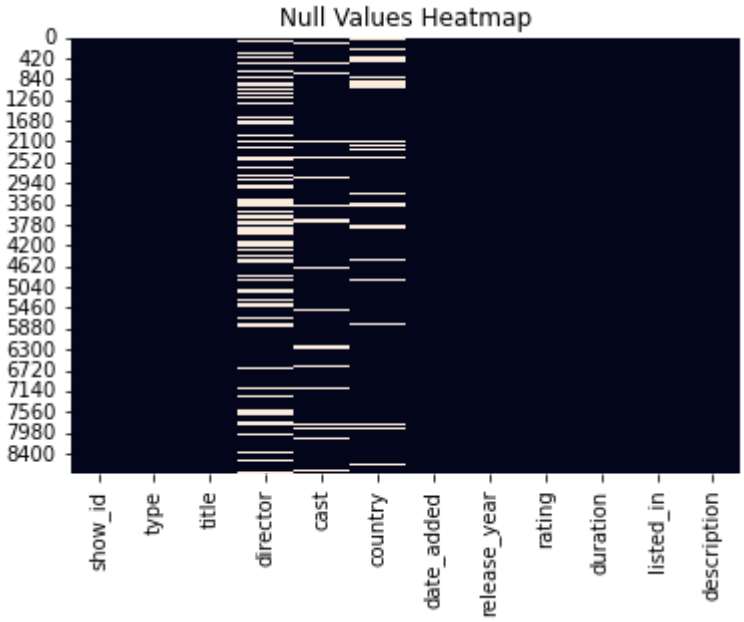
```
netflix_titles_df.isnull().values.any()
```

```
True
```

```
netflix_titles_df.isnull().sum().sum()
```

```
4307
```

```
sns.heatmap(netflix_titles_df.isnull(), cbar=False)
plt.title('Null Values Heatmap')
plt.show()
```




```
netflix_titles_df.isnull().sum()
```

```
show_id          0
type             0
title            0
director      2634
cast           825
country        831
date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description      0
dtype: int64
```

Above in the heatmap and table, we can see that there are quite a few null values in the dataset. There are a total of 4,307 null values across the entire dataset with 2,634 missing points under 'director', 825 under 'cast', 831 under

'country', 10 under 'date_added', 4 under 'rating', and 3 under 'duration'. We will have to handle all null data points before we can dive into EDA and modeling

```
netflix_titles_df['director'].fillna('No Director', inplace=True)
netflix_titles_df['cast'].fillna('No Cast', inplace=True)
netflix_titles_df['country'].fillna('Country Unavailable', inplace=True)
netflix_titles_df.dropna(subset=['date_added','rating','duration'],inplace=True)
```

```
netflix_titles_df.isnull().any()
```

```
show_id         False
type            False
title           False
director        False
cast            False
country         False
date_added      False
release_year    False
rating          False
duration        False
listed_in       False
description     False
dtype: bool
```

For null values, the easiest way to get rid of them would be to delete the rows with the missing data. However, this wouldn't be beneficial to our EDA since there is loss of information. Since 'director', 'cast', and 'country' contain the majority of null values, I will choose to treat each missing value as unavailable. The other two labels 'date_added' and 'rating' contains an insignificant portion of the data so I will drop them from the dataset. After, we can see that there are no more null values in the dataset.

Splitting the Dataset

Since the dataset can either contain movies or shows, it'd be nice to have datasets for both so we can take a deep dive into just Netflix movies or Netflix TV shows so we will create two new datasets. One for movies and the other one for shows.

```
netflix_movies_df = netflix_titles_df[netflix_titles_df['type']=='Movie'].copy()
netflix_movies_df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020	PG-13	90 min	D
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Country Unavailable	September 24, 2021	2021	PG	91 min	F

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	September 23, 2021	2021	TV-MA	127 min

```
netflix_movies_df = netflix_titles_df[netflix_titles_df['type']=='TV Show'].copy()
netflix_movies_df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	list
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Internat TV Sl TV Dra TV Mys
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Country Unavailable	September 24, 2021	2021	TV-MA	1 Season	Crin Sl Internat TV Sl TV
3	s4	TV Show	Jailbirds New Orleans	No Director	No Cast	Country Unavailable	September 24, 2021	2021	TV-MA	1 Season	Docus Reali
4	s5	TV Show	Kota Factory	No Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Internat TV Sl Rom TV Sl
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	Country Unavailable	September 24, 2021	2021	TV-MA	1 Season	TV Dra TV H TV Mys

Exploratory Analysis and Visualization

First we will begin analysis on the entire Netflix dataset consisting of both movies and shows. Revisiting the data, let us see how it looked like again.

```
netflix_titles_df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	l
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020	PG-13	90 min	Docume
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Interr TV Sh Drar M
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Country Unavailable	September 24, 2021	2021	TV-MA	1 Season	Cr Interr TV Sh
3	s4	TV Show	Jailbirds New Orleans	No Director	No Cast	Country Unavailable	September 24, 2021	2021	TV-MA	1 Season	Doc Re
4	s5	TV Show	Kota Factory	No Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Interr TV Roma Show

Netflix Film Types: Movie or TV Show

It'd be interesting to see the comparison between the total number of movies and shows in this dataset just to get an idea of which one is the majority.

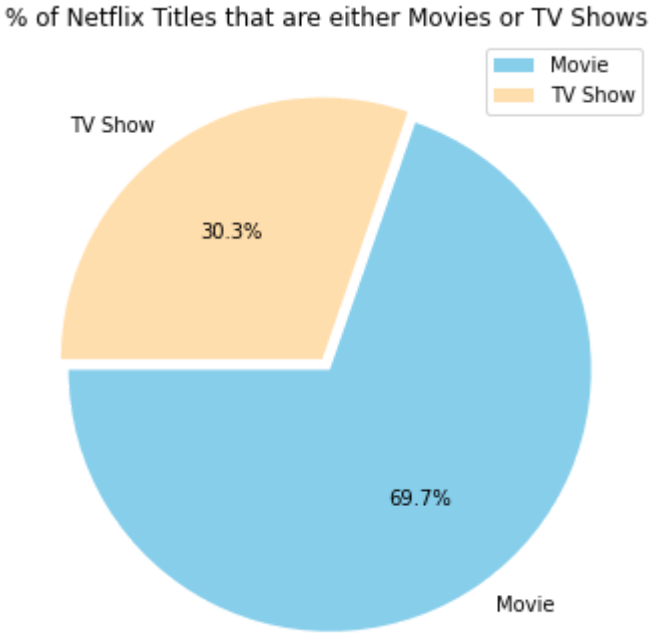
```
plt.figure(figsize=(7,5))
sns.barplot(x = "movies",data = netflix_titles_df.type, palette="pastel");
plt.title("Count of Movies and TV Shows")
plt.xlabel('Movie/TV Show')
plt.ylabel('Total Count')
plt.show()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Input In [35], in <cell line: 2>()
      1 plt.figure(figsize=(7,5))
----> 2 sns.barplot(x = "movies",data = netflix_titles_df.type, palette="pastel");
      3 plt.title("Count of Movies and TV Shows")
```

```
      4 plt.xlabel('Movie/TV Show')

File c:\program files\python38\lib\site-packages\seaborn\_decorators.py:46, in
_deprecate_positional_args.<locals>.inner_f(*args, **kwargs)
     36     warnings.warn(
     37         "Pass the following variable{} as {}keyword arg{}: {}. "
     38         "From version 0.12, the only valid positional argument "
   (...)
     43         FutureWarning
     44     )
     45 kwargs.update({k: arg for k, arg in zip(sig.parameters, args)})
---> 46 return f(**kwargs)

File c:\program files\python38\lib\site-packages\seaborn\categorical.py:3182, in
barplot(x, y, hue, data, order, hue_order, estimator, ci, n_boot, units, seed, orient,
color, palette, saturation, errcolor, errwidth, capsize, dodge, ax, **kwargs)
   3169 @_deprecate_positional_args
   3170 def barplot(
   3171     *,
   (...)
   3179     **kwargs,
   3180 ):
-> 3182     plotter = _BarPlotter(x, y, hue, data, order, hue_order,
   3183                           estimator, ci, n_boot, units, seed,
   3184                           orient, color, palette, saturation,
   3185                           errcolor, errwidth, capsize, dodge)
   3187     if ax is None:
   3188         ax = plt.gca()

File c:\program files\python38\lib\site-packages\seaborn\categorical.py:1584, in
_BarPlotter.__init__(self, x, y, hue, data, order, hue_order, estimator, ci, n_boot,
units, seed, orient, color, palette, saturation, errcolor, errwidth, capsize, dodge)
   1579 def __init__(self, x, y, hue, data, order, hue_order,
   1580              estimator, ci, n_boot, units, seed,
   1581              orient, color, palette, saturation, errcolor,
   1582              errwidth, capsize, dodge):
   1583     """Initialize the plotter."""
-> 1584     self.establish_variables(x, y, hue, data, orient,
   1585                              order, hue_order, units)
   1586     self.establish_colors(color, palette, saturation)
   1587     self.estimate_statistic(estimator, ci, n_boot, seed)

File c:\program files\python38\lib\site-packages\seaborn\categorical.py:153, in
_CategoricalPlotter.establish_variables(self, x, y, hue, data, orient, order, hue_order,
units)
    151     if isinstance(var, str):
    152         err = "Could not interpret input '{}'".format(var)
--> 153         raise ValueError(err)
    155 # Figure out the plotting orientation
```

```
    156 orient = infer_orient(
    157     x, y, orient, require_numeric=self.require_numeric
    158 )

ValueError: Could not interpret input 'movies'
<Figure size 504x360 with 0 Axes>
```

```
plt.figure(figsize=(12,6))
plt.title("% of Netflix Titles that are either Movies or TV Shows")
g = plt.pie(netflix_titles_df.type.value_counts(), explode=(0.025,0.025), labels=netfli
plt.legend()
plt.show()
```



So there are roughly 4,000+ movies and almost 2,000 shows with movies being the majority. This makes sense since shows are always an ongoing thing and have episodes. If we were to do a headcount of TV show episodes vs. movies, I am sure that TV shows would come out as the majority. However, in terms of title, there are far more movie titles (68.5%) than TV show titles (31.5%).

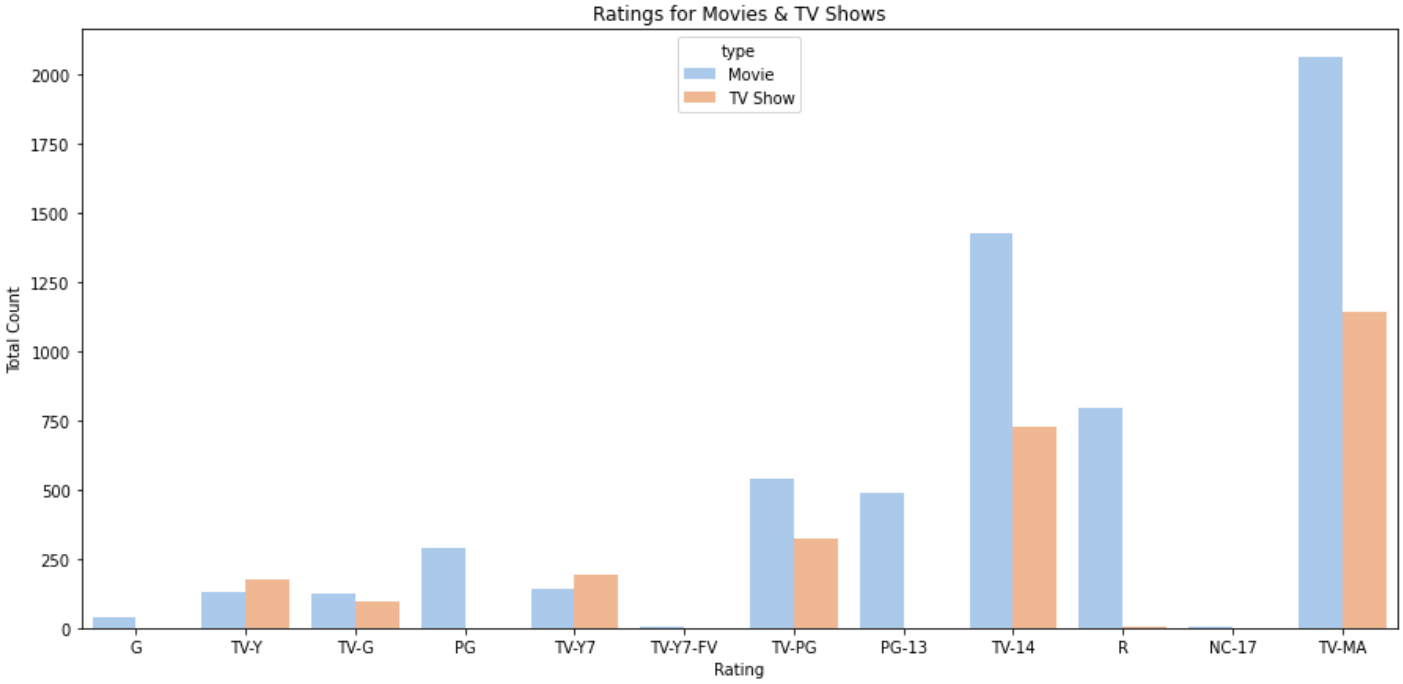
Netflix Film Ratings

Now, we will explore the ratings which are based on the film rating system. The ordering of the ratings will be based on the age of the respective audience from youngest to oldest. We will not include the ratings 'NR' and 'UR' in the visuals since they stand for unrated and non-rated content.

```
order =  ['G', 'TV-Y', 'TV-G', 'PG', 'TV-Y7', 'TV-Y7-FV', 'TV-PG', 'PG-13', 'TV-14', 'R
plt.figure(figsize=(15,7))
g = sns.countplot(netflix_titles_df.rating, hue=netflix_titles_df.type, order=order, pa
plt.title("Ratings for Movies & TV Shows")
plt.xlabel("Rating")
plt.ylabel("Total Count")
plt.show()
```

```
c:\program files\python38\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
```

```
Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```




```
import jovian
```

```
jovian.commit()
```

```
```