Netflix Movies and TV Shows Analysis

Who doesn't like to relax and watch a movie or a TV show on Netflix? Well, I certainly do. For this project, I have decided to analyse a dataset containing information about Netflix Movies and TV Shows. The aim of this project is to showcase the skills that I have learned from the course "Data Analysis with Python: Zero to Pandas". I will use python libraries including Pandas, Matplotlib and Seaborn in this project.

Downloading the Dataset

I found this dataset on kaggle.com. There are many datasets on that website but I chose this one because I found it to be the most interesting. I used the openedatasets library to download the dataset directly from Kaggle to this python notebook.

```
!pip install jovian opendatasets --upgrade --quiet
```

Let's begin by downloading the data, and listing the files within the dataset.

```
# Change this
dataset_url = 'https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download'
```

```
import opendatasets as od
od.download(dataset_url)
```

```
Please provide your Kaggle credentials to download this dataset. Learn more:
http://bit.ly/kaggle-creds
Your Kaggle username: adhyannegi
Your Kaggle Key: ········
Downloading netflix-shows.zip to ./netflix-shows

100%|██████████| 1.34M/1.34M [00:00<00:00, 79.5MB/s]
```

The dataset has been downloaded and extracted.

```
# Change this
data_dir = './netflix-shows'
```

```
import os
os.listdir(data_dir)
```

```
['netflix_titles.csv']
```

```
project_name = "netflix-movies-and-tv-shows-analysis"
```

```
!pip install jovian --upgrade -q
```

Data Preparation and Cleaning

-> Imported the pandas library and loaded the dataset.

-> Printed out the dataset to make sure it was imported properly.

-> Displayed some basic information about the dataset.

-> Displayed number of unique values in each column of the dataset.

-> Found out that there were some null values in the dataset.

-> Wrote some commands to handle the null values carefully.

```
import pandas as pd
```

```
data_frame = pd.read_csv('./netflix-shows/netflix_titles.csv')
```

```
data_frame
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Docum
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Inter TV Sl Dra I
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	(Inte TV Sl
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Dc R
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Inte T' Rom Sho
...	...	...	...	...	...	...	...	...	...	...	

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cul
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	K Sl C
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	C Horrc
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	C Famil C
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Inte Movi &

8807 rows × 12 columns

This is the dataset.

```
data_frame.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
```

The info function gives us some basic information about the dataset. There are a total of 8807 entries and we can see that there are some null values in the dataset.

```
data_frame.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
```

This function gives us the count of unique values for each column.

```
data_frame.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
```

This function gives us the count of null values for each column. There are quite a bit of null values in this dataset, which may cause problems when we analyse the dataset, so this issue must be resolved.

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

Replacing null values with appropriate text messages.

```
data_frame.isnull().sum()
```

```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
dtype: int64
```

Now we can see that there are no null values in the dataset, we can move ahead with our analysis.

Exploratory Analysis and Visualization

In this section I analyse data using visuals. I use the matplotlib and the seaborn libraries to make bar graphs and pie charts to better analyse the data.

Let's begin by importing `matplotlib.pyplot and seaborn` .

```
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

```
data_frame.head()
```

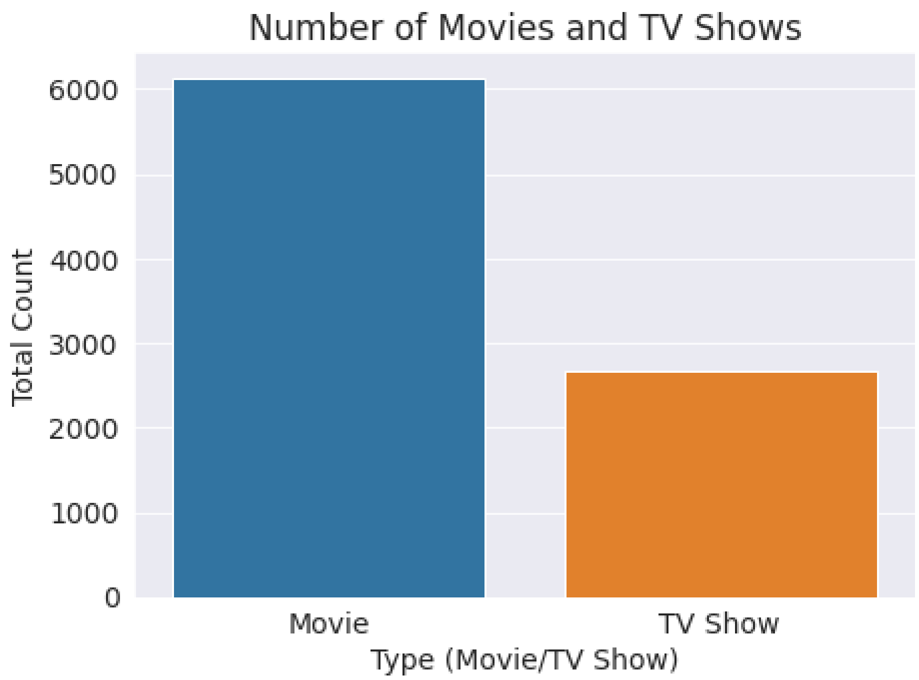
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	l
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020	PG-13	90 min	Docume
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Interr TV Sh Drar M

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	l
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

Revisiting the data set before we do some analysis.

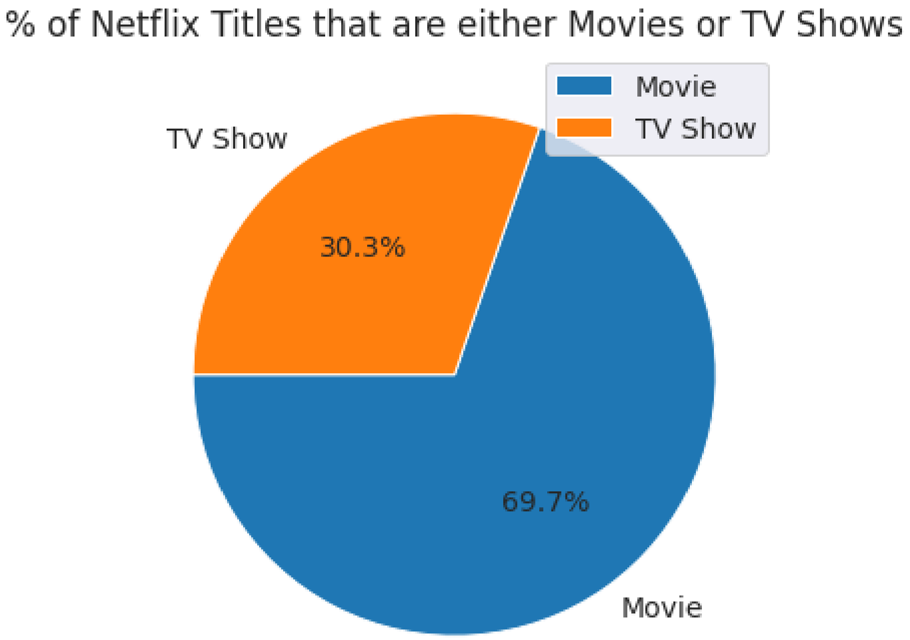
```
plt.figure(figsize=(7,5))
graph = sns.countplot(data_frame.type);
plt.title("Number of Movies and TV Shows")
plt.xlabel("Type (Movie/TV Show)")
plt.ylabel("Total Count")
plt.show()
```

```
/opt/conda/lib/python3.9/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(
```



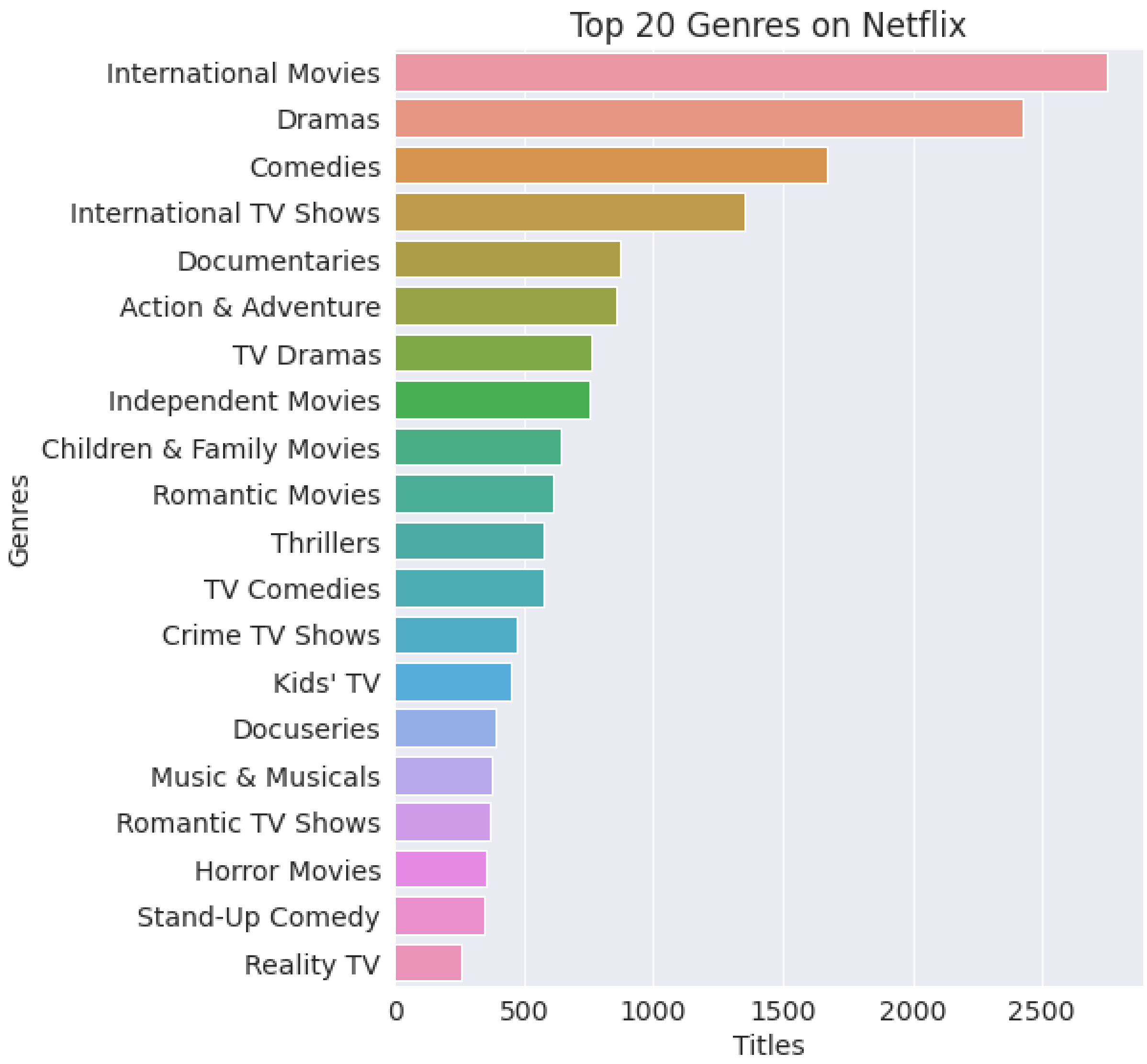
In this bar graph, we can see the number of Movies and the number of TV Shows in the dataset. From the visual, we can say that the number of movies is more than double the number of TV Shows in the dataset.

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



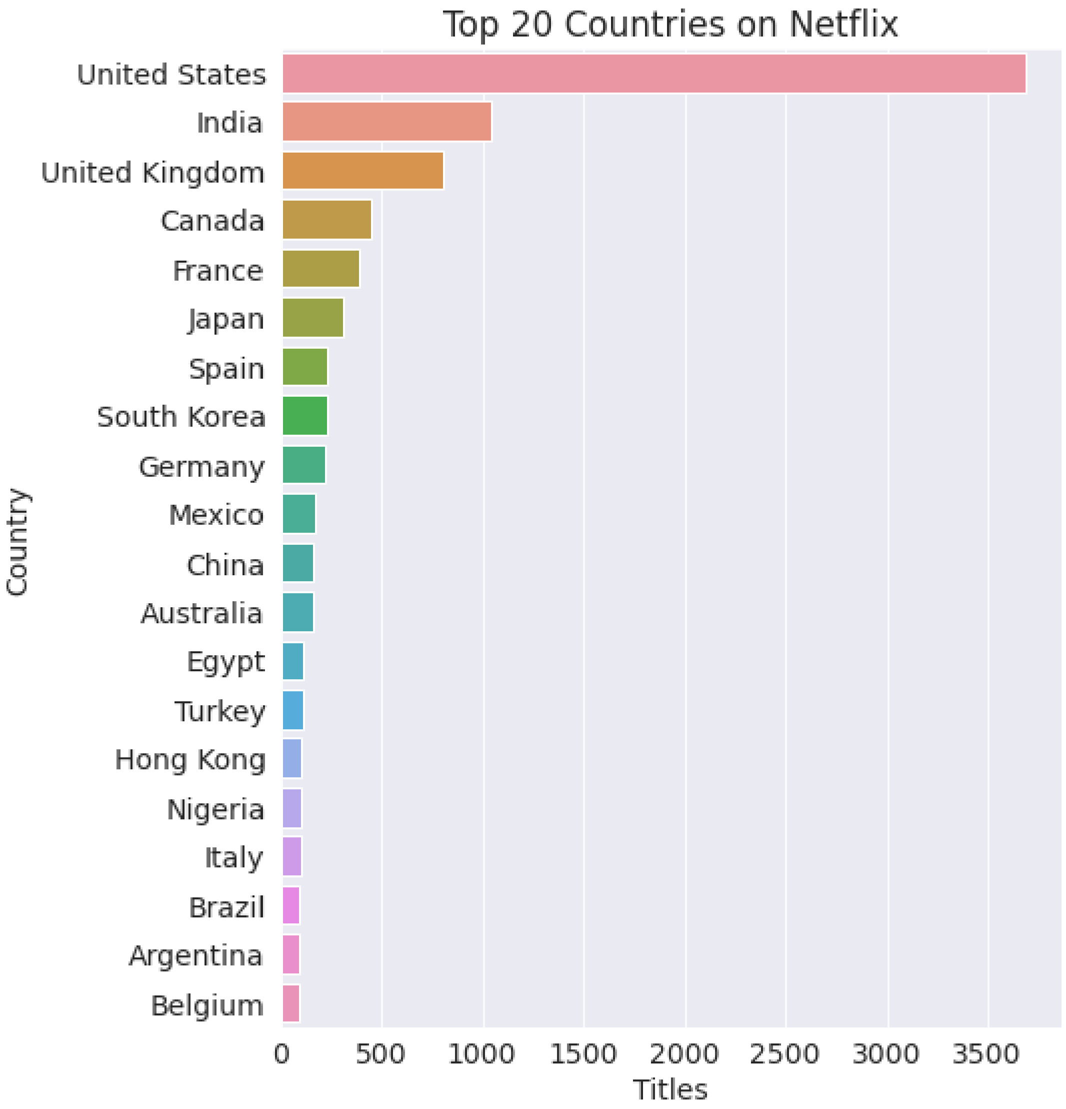
Using this pie chart, we can better analyse the data. We can now clearly say that 69.7% of the values are Movies and the other 30.3% are TV Shows.

```
filtered_genres = data_frame.set_index('title').listed_in.str.split(', ', expand=True).

plt.figure(figsize=(7,9))
g = sns.countplot(y = filtered_genres, order=filtered_genres.value_counts().index[:20])
plt.title('Top 20 Genres on Netflix')
plt.xlabel('Titles')
plt.ylabel('Genres')
plt.show()
```



This bar chart shows us the genre distribution in the dataset. The x-axis is the number of titles and the y-axis is the genre. "International Movies" is the most famous genre.

```
filtered_countries = data_frame.set_index('title').country.str.split(', ', expand=True)
filtered_countries = filtered_countries[filtered_countries != 'Country Unavailable']

plt.figure(figsize=(7,9))
g = sns.countplot(y = filtered_countries, order=filtered_countries.value_counts().index
plt.title('Top 20 Countries on Netflix')
plt.xlabel('Titles')
plt.ylabel('Country')
plt.show()
```



This bar chart shows us the top 20 countries with the most titles on Netlfix. United States has the maximum number of titles.

Asking and Answering Questions

In this section, I answer some of the questions that came to my mind when I looked at this dataset.

Q1: How many new Movies and TV Shows has Netflix added each year?

```
data_frame['year_added'] = pd.DatetimeIndex(data_frame['date_added']).year
each_year = data_frame['year_added'].value_counts().to_frame().reset_index().rename(col
each_year
```

	year	count
0	2019.0	2016
1	2020.0	1879
2	2018.0	1649
3	2021.0	1498
4	2017.0	1188
5	2016.0	429

	year	count
6	2015.0	82
7	2014.0	24
8	2011.0	13
9	2013.0	11
10	2012.0	3
11	2009.0	2
12	2008.0	2
13	2010.0	1

Here, we can see the number of Movies and TV Shows added each year on Netflix. 2019 was the year with the most additions and 2010 was the year with the least additions. I used a function from the pandas library to calculate how many TV Shows and Movies were added to Netflix in a particular year.

Q2: How many Movies and TV Shows did Adam Sandler star in?

```
count = 0
for actor in data_frame.cast:
    if 'Adam Sandler' in actor:
        count += 1

print("Adam Sandler starred in {} Movies and TV Shows.".format(count))
```

```
Adam Sandler starred in 20 Movies and TV Shows.
```

Q3: How many Movies and TV Shows were listed as comedies?

```
count = 0
for genre in data_frame.listed_in:
    if 'Comedies' in genre:
        count += 1

print("There are {} Movies and TV Shows listed as Comedies.".format(count))
```

```
There are 2247 Movies and TV Shows listed as Comedies.
```

Q4: How many movies and TV Shows were rated PG-13?

```
count = 0
for rate in data_frame.rating:
    if rate == "PG-13":
        count += 1

print("{} Movies and TV Shows were are rated PG-13.".format(count))
```

```
490 Movies and TV Shows were are rated PG-13.
```

Q5: How many Movies and TV Shows are from the United States?

```python
count = 0
for c in data_frame.country:
    if "United States" in c:
        count += 1

print("{} Movies and TV Shows are from the United States.".format(count))
```

```
3684 Movies and TV Shows are from the United States.
```

Inferences and Conclusion

It is clear from the analysis that Netflix has grown massively over the years. Netlix's original subscriber base was based solely in the United States, a large part of its success was due to the decision to expand to international markets. Through this dataset, we can see that a good amount of international movies and TV shows were added over the years as part of Netflix's global expansion.

References

https://www.kaggle.com/shivamb/netflix-shows

https://www.businessinsider.com/netflix-growth-comes-from-international-markets-2019-10

```python
import jovian
```

```python
jovian.commit()
```