Exploratory Data Analysis

Vamshi Jaligama Saikrishna Reddy Kotha Abhishek Raj Sampath

ITC 510 Final Project

Exploratory data analysis is a process of identifying the missing , outliers in the data and cleaning the data. Finally, visualize the data and get the insights from the large data. We have done Exploratory data analysis(EDA) on dataset from Kaggle website called Netflix Movies and TV Shows .For this EDA we used Python as our primary language , libraries like pandas , NumPy , plotly , seaborn for the visualization of the data and for development environment we used the Jupyter notebook.

EDA Process:-

Firstly, we imported all the libraries that are required for the EDA like NumPy , pandas, matplotlib, plotly and seaborn and then read the csv file using the pandas as below :

In [3]:
```
#Importing the required libararies
import numpy as np
import pandas as pd
from matplotlib.pyplot import *
import matplotlib.pyplot as plt
import plotly.graph_objs as go
import plotly.express as px
from plotly.subplots import make_subplots
import seaborn as sns
from pandas_profiling import ProfileReport
```

In [4]:
```
#Reading the CSV File
file = pd.read_csv('netflix_titles.csv')
file.head(5)
```

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Next to get the idea about the data we use the info function and the descriptive statistics of the dataset

In [5]:
```
#Info about the Dataset
file.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
```

In [6]:
```
#Descriptive statistics of the dataset
file.describe()
```

Out[6]:

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

We need to look for the missing values in the dataset :

In [7]:
```
#Missing values in the dataset
file.isna().sum()
```

Out[7]:
```
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
```

Then we had replaced the missing country with the mode values. Removing the date_added rows. Replacing the missing rating with the mode values. Removing the missing duration as they are very few missing values. Also removing the missing value columns for the director , cast column's because those values are also not helpful for EDA :

In [8]:
```
#Replacing the missing country with mode values
country_mode = file['country'].mode().values[0]

file['country'] = file['country'].replace(np.nan,country_mode)
```

In [9]:
```
#Removing the date_added rows
file = file[file['date_added'].notna()]
```

In [10]:
```
#Replacing the missing country with mode values
rating_mode = file['rating'].mode().values[0]

file['rating'] = file['rating'].replace(np.nan,rating_mode)
```

In [11]:
```
#Removing the date_added rows
file = file[file['duration'].notna()]
```

Checking for the duplicate values in the dataset

Checking for Duplicate values in the Dataset

In [13]:
```
#Checking for duplicates in the dataset
duplicate_values = file.duplicated()
print("Number of duplicates in the dataset is :",duplicate_values.sum())
```

```
Number of duplicates in the dataset is : 0
```

Adding a column Movies ,TV Show added year to the Netflix form the existing date_added column:

In [14]:
```
#Adding a column movies/tvshow added year to the Netflix form the existing date_added column
file['available_year'] = file['date_added'].apply(lambda date : date.split(" ")[-1])
file.head(5)
```

Out[14]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	available_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	2021
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	2021

Only keeping the first country in the country column and removing the extra data :

```
In [17]:
# Only keeping the first country in the country column and removing the extra data
file['primary_country'] = file['country'].apply(lambda co : co.split(",")[0])
file.tail(10)
```

Out[17]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	available_year	pri
8797	s8798	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	September 13, 2018	2016	TV-Y7	3 Seasons	Kids' TV	Teen surfer Zak Storm is mysteriously transpor...	2018	
8798	s8799	Movie	Zed Plus	Chandra Prakash Dwivedi	Adil Hussain, Mona Singh, K.K. Raina, Sanjay M...	India	December 31, 2019	2014	TV-MA	131 min	Comedies, Dramas, International Movies	A philandering small-town mechanic's political...	2019	

To get the count of unique values of cast , director and titles we used the nunique function:

```
In [18]:
#unique director
file['director'].nunique()
```

Out[18]: 4527

```
In [19]:
#Unique title
file.title.nunique()
```

Out[19]: 8794

```
In [21]:
#Unique cast
file.cast.nunique()
```

Out[21]: 7681

To get the top 20 oldest shows in the netflix

```
In [28]:
#20 oldest TV shows on Netflix
old_tvshow = file.sort_values("release_year", ascending = True)
old_tvshow = old_tvshow[old_tvshow['type'] == "TV Show"]
old_tvshow[['title', "release_year", 'type',]][:20]
old_tvshow.rename(columns={'title': 'Title', 'release_year': 'Release Year', 'type':'Type'}, inplace=True)
old_tvshow = old_tvshow[['Release Year', "Title",'Type']].head(20)
old_tvshow.head(20).style.hide_index()
```

Out[28]:

Release Year	Title	Type
1925	Pioneers: First Women Filmmakers*	TV Show
1945	Five Came Back: The Reference Films	TV Show
1946	Pioneers of African-American Cinema	TV Show
1963	The Twilight Zone (Original Series)	TV Show
1967	The Andy Griffith Show	TV Show
1972	Monty Python's Fliegender Zirkus	TV Show
1974	Monty Python's Flying Circus	TV Show
1977	Dad's Army	TV Show
1979	El Chavo	TV Show
1981	Ninja Hattori	TV Show
1985	Robotech	TV Show

To get the Top 20 oldest Movies in the netflix

In [29]:
```
#20 oldest movies on netflix
old_movie = file.sort_values("release_year", ascending = True)
old_movie = old_movie[old_movie['type'] == "Movie"]
old_movie[['title', "release_year", 'type',]][:20]
old_movie.rename(columns={'title': 'Title', 'release_year': 'Release Year', 'type':'Type'}, inplace=True)
old_movie = old_movie[['Release Year', "Title",'Type']].head(20)
old_movie.head(20).style.hide_index()
```

Out[29]:

Release Year	Title	Type
1942	The Battle of Midway	Movie
1942	Prelude to War	Movie
1943	Undercover: How to Operate Behind Enemy Lines	Movie
1943	Why We Fight: The Battle of Russia	Movie
1943	WWII: Report from the Aleutians	Movie
1944	The Memphis Belle: A Story of a Flying Fortress	Movie
1944	The Negro Soldier	Movie
1944	Tunisian Victory	Movie
1945	Know Your Enemy - Japan	Movie

To get the data about the Movies and TV Shows that are added recently :

In [30]:
```
#Movies and TV Shows added recently
file['date_added'] = pd.to_datetime(file['date_added'])
newly_added = file.sort_values(by='date_added', ascending=False)
newly_added.rename(columns={'title': 'Title', 'date_added': 'Date Added', 'type':'Type', 'release_year':'Release Year', 'desc
newly_added = newly_added[['Date Added', "Title",'Type', 'Release Year', 'Description']][:15]
newly_added.head(20).style.hide_index()
```

Out[30]:

Date Added	Title	Type	Release Year	Description
2021-09-25 00:00:00	Dick Johnson Is Dead	Movie	2020	As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable.
2021-09-24 00:00:00	My Little Pony: A New Generation	Movie	2021	Equestria's divided. But a bright-eyed hero believes Earth Ponies, Pegasi and Unicorns should be pals — and, hoof to heart, she's determined to prove it.
2021-09-24 00:00:00	Vendetta: Truth, Lies and The Mafia	TV Show	2021	Sicily boasts a bold "Anti-Mafia" coalition. But what happens when those trying to bring down organized crime are accused of being criminals themselves?

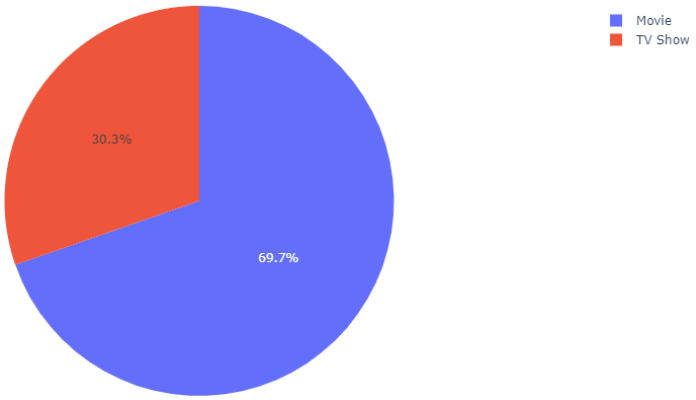
Popular actor in the Netflix based on the count :

Top 10 most casted Artist

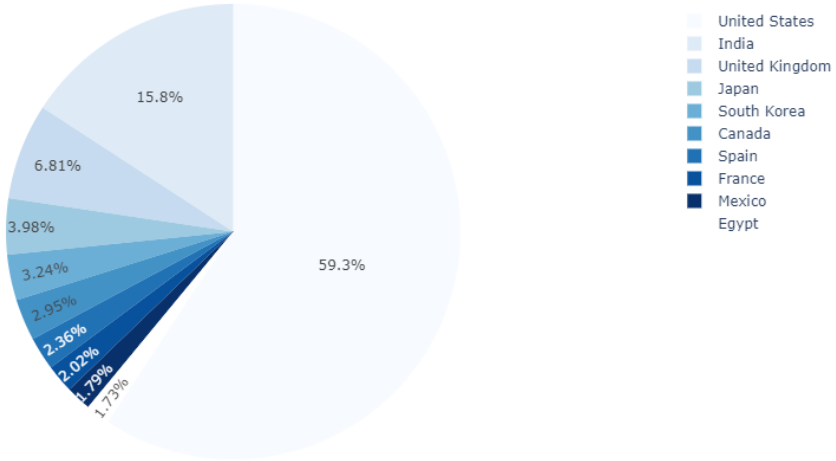
In [23]:
```
max_cast = file.copy()
max_cast = pd.concat([max_cast, file['cast'].str.split(",", expand=True)], axis=1)
max_cast = max_cast.melt(id_vars=["type","title"], value_vars=range(44), value_name="Cast_name")
max_cast = max_cast[max_cast["Cast_name"].notna()]
max_cast["Cast_name"] = max_cast["Cast_name"].str.strip()
max_cast.Cast_name.value_counts()[:10]
```

Out[23]:
```
Anupam Kher          43
Shah Rukh Khan       35
Julie Tejwani        33
Takahiro Sakurai     32
Naseeruddin Shah     32
Rupa Bhimani         31
Om Puri              30
Akshay Kumar         30
Yuki Kaji            29
Paresh Rawal         28
Name: Cast_name, dtype: int64
```

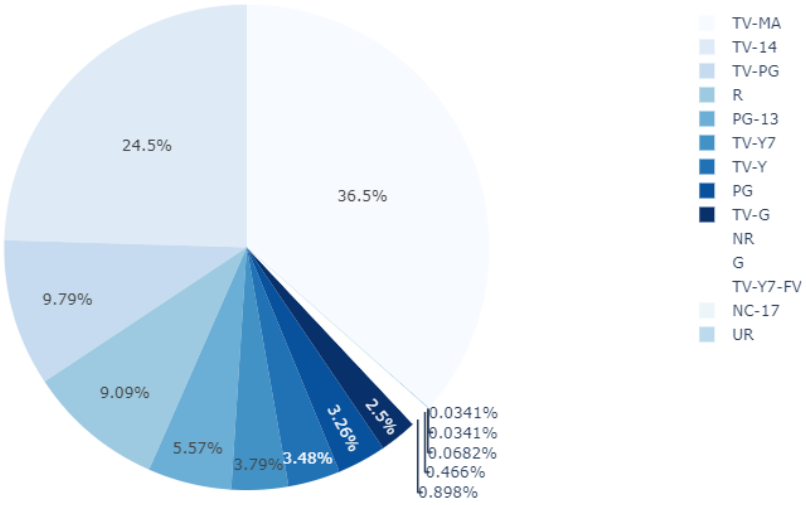
Pie chart about the count of Movies and TV Shows :



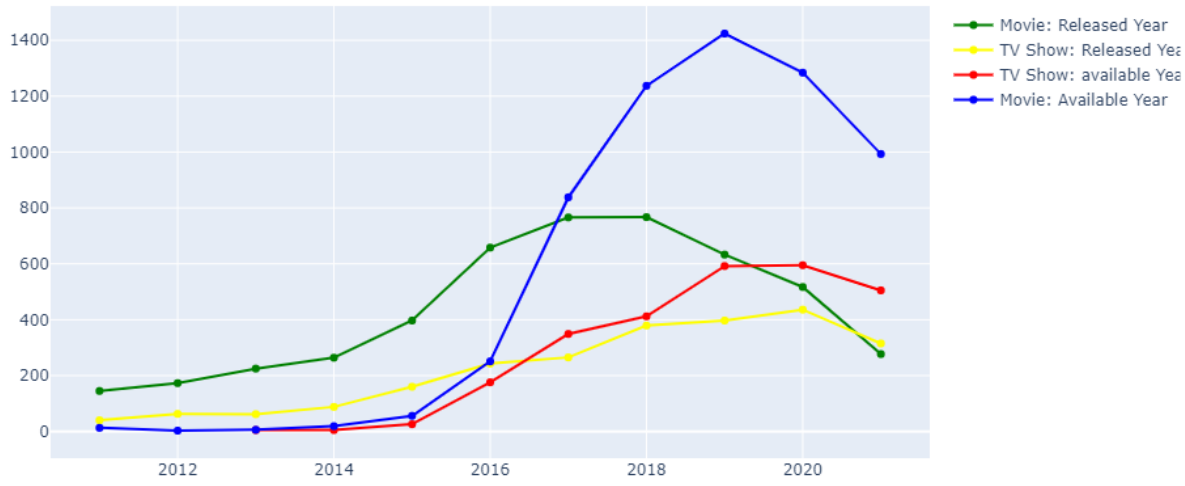
A pie chart about the country wise netflix content:



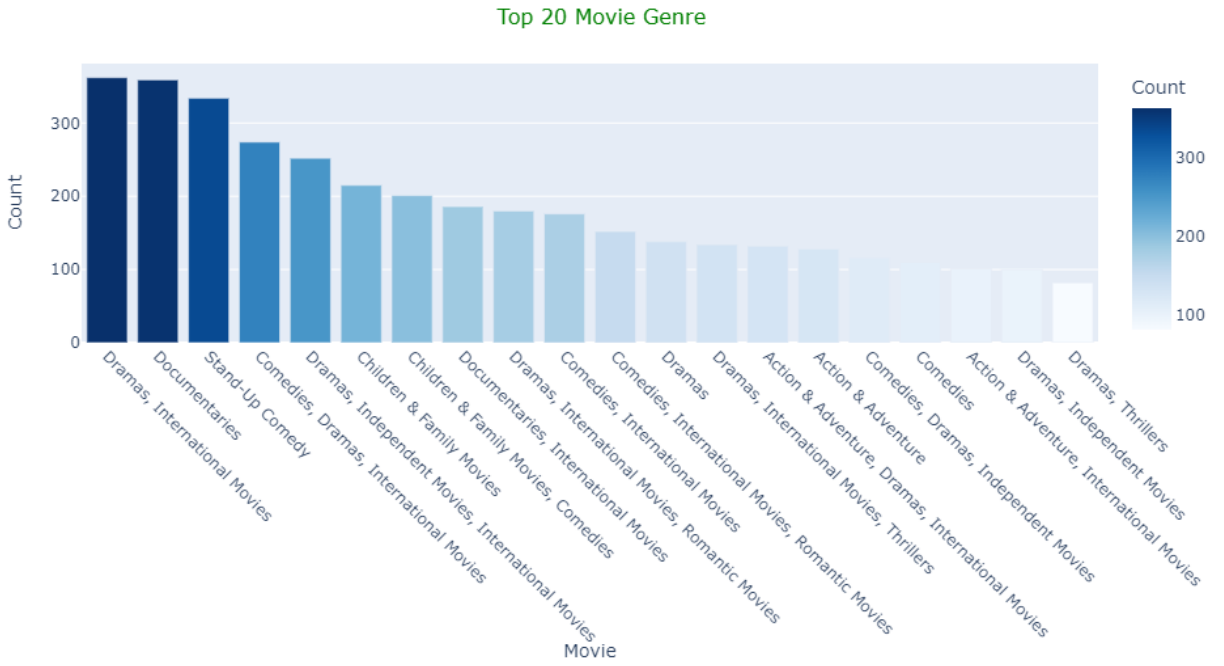
Piechart about the Movies / TV Show Ratings :



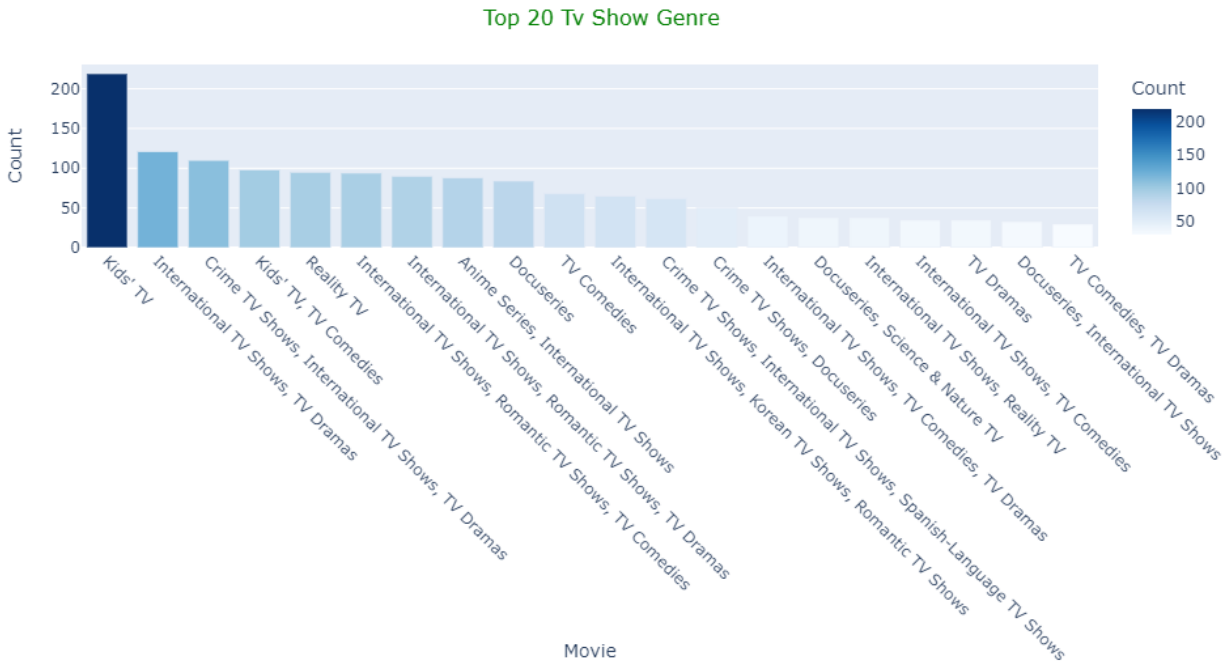
Scatter plot about the Movies / TV Show that are released year to the available year in the netflix :



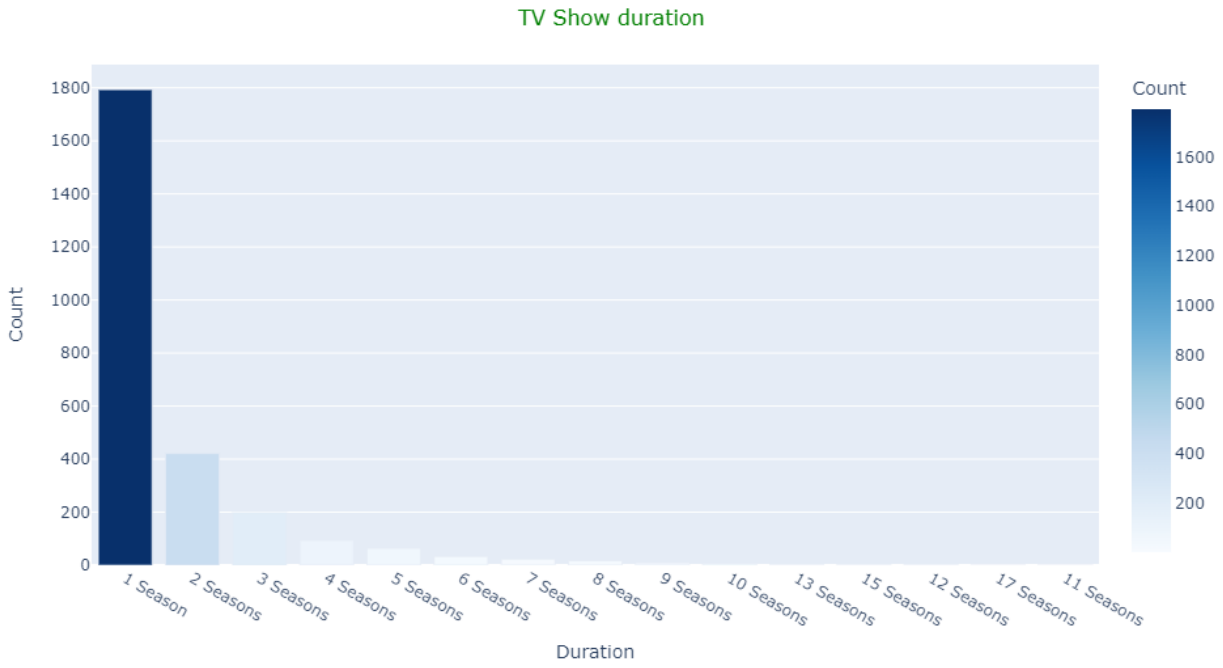
A Bar plot about the top 20 Movie Genre :



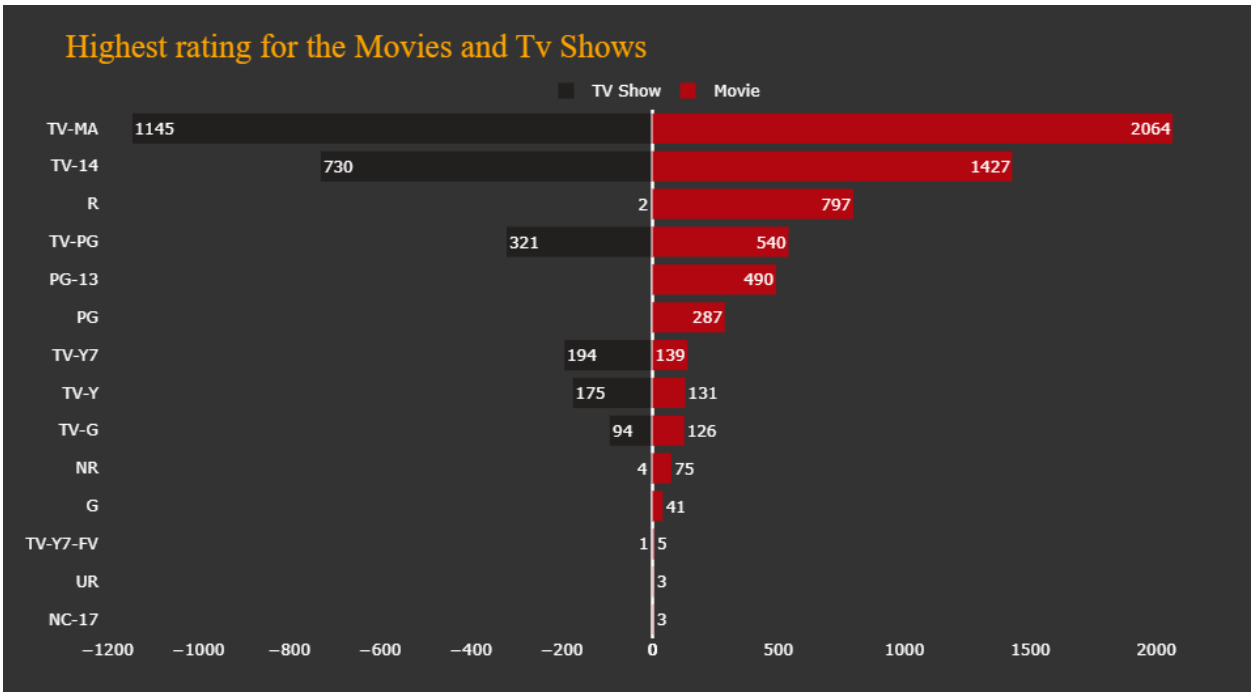
Bar chart about the TV Show Genre :



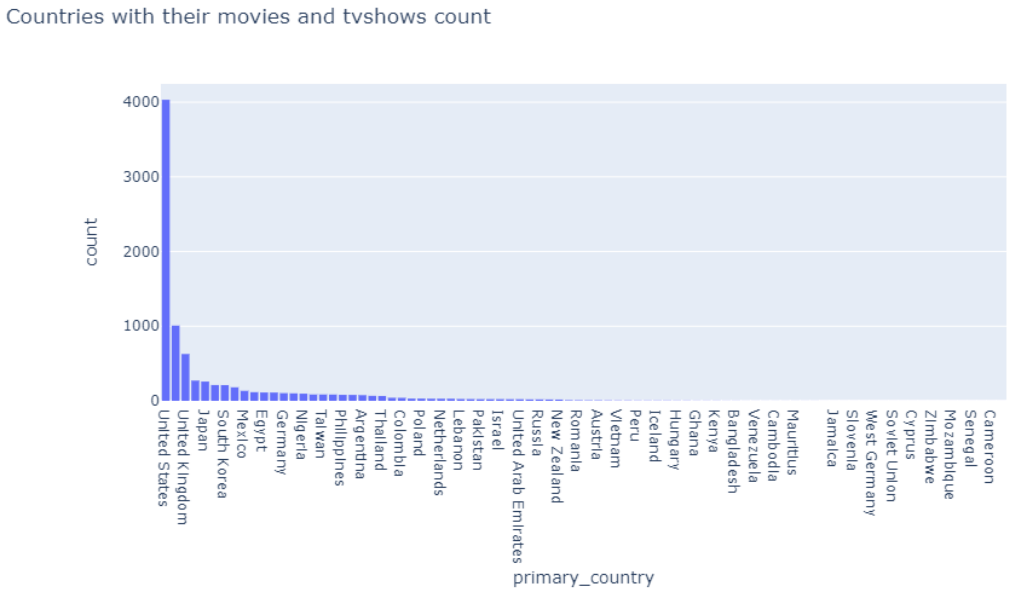
Bar chart about the distribution of the duration of TV Shows :



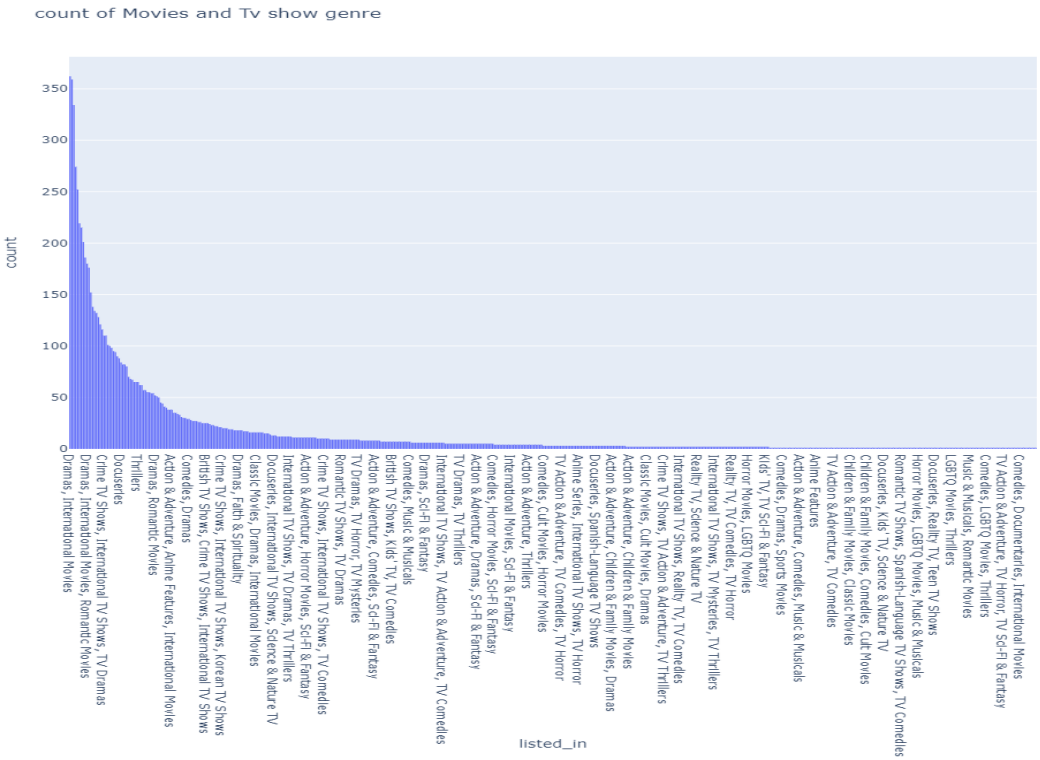
Bar chart about the highest rating of TV Shows and Movies :



Bar chart about the Count of TV Shows and Movies :



Bar chart about the count of TV Show and Movies about the Genre :



Bibliography

[1] A. Vidhya, "EDA in Python," [Online]. Available: https://www.analyticsvidhya.com/blog/2020/08/exploratory-data-analysiseda-from-scratch-in-python/.

[2] Kaggle, "Netflix Movies and TV Show," [Online]. Available: https://www.kaggle.com/shivamb/netflix-shows/code.