Netflix Business Case-Study

1. Netflix – Data Visualization

1.1 Introduction

Netflix, Inc. is an American subscription streaming service and production company. It offers a library of films and television series through distribution deals as well as its own productions, known as Netflix Originals. As of March 31, 2023, with an estimated 232.5 million paid memberships in more than 190 countries, it is the most-subscribed video on demand streaming service. Founded by Reed Hastings and Marc Randolph in Scotts Valley, California, Netflix initially operated as a DVD sales and rental business. However, within a year, it shifted its focus exclusively to DVD rentals. In 2007, the company introduced streaming media and video on demand services, marking a significant step in its evolution.

1.1.1 Problem Statement

Analyzing the data and generating Insights that would help Netflix in deciding which type of Shows/Movies to produce more and how to grow business in different countries

1.2 About Data

The Dataset consists of data of range 2008-mid 2021 ,about 8807 tv shows and movies available , along with other details such as – cast, director, type ,ratings, release year ,duration etc. .The data is available in single csv file

1.3 Features of Dataset

Show_id: Unique ID for every Movie / Tv Show
Type: Identifier - A Movie or TV Show
Title: Title of the Movie / Tv Show
Director: Director of the Movie
Cast: Actors involved in the movie/show
Country: Country where the movie/show was produced
Date_added: Date it was added on Netflix
Release_year: Actual Release year of the movie/show
Rating: TV Rating of the movie/show
Duration: Total Duration - in minutes or number of seasons
Listed_in: Genre
Description: The summary description

Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary

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

Read File and show

```
df=pd.read_csv('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv'
)
df.head()
```

index	show_id	type	title	director	cast	country	date_added	release_year	rating	dura
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 m
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore Sithole, Cindy Mahlangu, Ryle De Morny, Greteli	South Africa	September 24, 2021	2021	TV-MA	2 Seas

index	show_id	type	title	director	cast	country	date_added	release_year	rating	dura
					Fincham, Sello Maake Ka-Ncube, Odwa Gwanya, Mekaila Mathys, Sandi Schultz, Duane Williams, Shamilla Miller, Patrick Mofokeng					
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffre, Salim Kechiouche, Noureddine Farihi, Geert Van Rampelberg, Bakary Diombera	NaN	September 24, 2021	2021	TV-MA	1 Seas
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Seas

```
df.info() #checking info
```

```
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)
```

```
df.dtypes
```

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

Shape

```
df.shape
```

(8807, 12)

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

index	show_id	type	title	director	cast	country	rating	duration	listed_in	descriptic
count	202065	202065	202065	202065	202065	202065	202065	202065	202065	202065
unique	8807	2	8807	5121	39297	197	14	220	73	8775
top	s7165	Movie	Kahlil Gibran's The Prophet	NotAvailable	NotAvailable	United States	TV-MA	1 Season	International Movies	A troubled young girl and her mothe with a subversive poet whose words imaginations.
freq	700	145917	700	50643	2149	61765	73985	35038	27141	700

```
df.isna().sum() #Check NAN Values
```

```
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
```

Preprocessing and unnesting

Filling NAN Space

```
df['director'] = df['director'].fillna('NotAvailable')
df['cast'] = df['cast'].fillna('NotAvailable')
df['country'] = df['country'].fillna(df['country'].mode()[0])
df['date_added'] = df['date_added'].fillna(df['date_added'].mode()[0])
df['duration'] = df['duration'].fillna(df['duration'].mode()[0])
df['rating'].fillna(df['rating'].mode()[0]
```

Splitting rows with multiple values

```
## Converting the columns to string tyoe before splitting
df['director'] = df['director'].astype(str)
df['cast'] = df['cast'].astype(str)
df['country'] = df['country'].astype(str)
df['listed_in'] = df['listed_in'].astype(str)
```

0s

```
df['cast'] = df['cast'].apply(lambda x: x.split(','))
df['director'] = df['director'].apply(lambda x: x.split(','))
df['country'] = df['country'].apply(lambda x: x.split(','))
df['listed_in'] = df['listed_in'].apply(lambda x: x.split(','))
```

0s

```
df = df.explode('cast')
df = df.explode('director')
df = df.explode('country')
df = df.explode('listed_in')
df.head()
```

	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	NotAvailable	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	NotAvailable	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows	After crossing paths at a party, a Cape Town t...
1	s2	TV Show	Blood & Water	NotAvailable	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	TV Dramas	After crossing paths at a party, a Cape Town t...

Converting data_added column to datetime

0s

```
df['date_added'] = pd.to_datetime(df['date_added'],format = 'mixed')
df['year'] = df['date_added'].dt.year
```

spliitting duration of movies and seasons

```
df['duration'] = df['duration'].astype(str)
df['movie_min'] = df[df['type']=='movie']['duration'].apply(lambda x: x.split(' ')[0])
df['seasons_no'] = df[df['type']=='Tv Show']['duration'].apply(lambda x: x.split(' ')[0])
df['rating'].unique()
```

```
array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',
    'TV-G', 'G', 'NC-17', '74 min', '84 min', '66 min', 'NR', nan,
    'TV-Y7-FV', 'UR'], dtype=object)
```

replacing rating values

```
df['rating'] = df['rating'].replace(['66 min', '74 min', '84 min'],np.nan)
def get_mode(series):
  return series.mode()[0] if not series.mode().empty else np.nan
df['rating'] = df.groupby('type')['rating'].transform(lambda x: x.fillna(get_mode(x)))
```

Analysis of Insights

From analysis it can be seen in total Netflix has 8807 total shows inkling Movies and Tv Shows both and out of which there are 6131-Movies and 676 Tv Shows.
Which shows the 70% percentage of movies and 30% of Tv shows

It is also noted that most common Genre is [International Movies , Dramas , Comedies ,Action,Documentaries].
Out of which International Movies hold the top position

It is also noted that Least common opted Genre is LGBTQ movies and sports movies

There are maximum movies with highest ratings of TV-MA

Maximum number of Movies and Tv Shows production has been noted in United States and India . United States holds the top position in to it

It is Seen Rajiv Chilaka holds the top position as Director and most for family and children entertainment

It is also observed addition of content has been increasing with years and maximum peak time was years 18,19,20,21
And 2020 has observed maximum number of production .The Same scenario has been seen for tv and movies both .
We have also observed we can see notable number of growth in Tv shows after 2018.

It is also observed that best time to launch movie is mostly month like July, Where as For Tv Shows December has been more preferred month.

Attributes

```
for i in df.columns:
  print(i,df[i].nunique())
  print('-'*20)
```

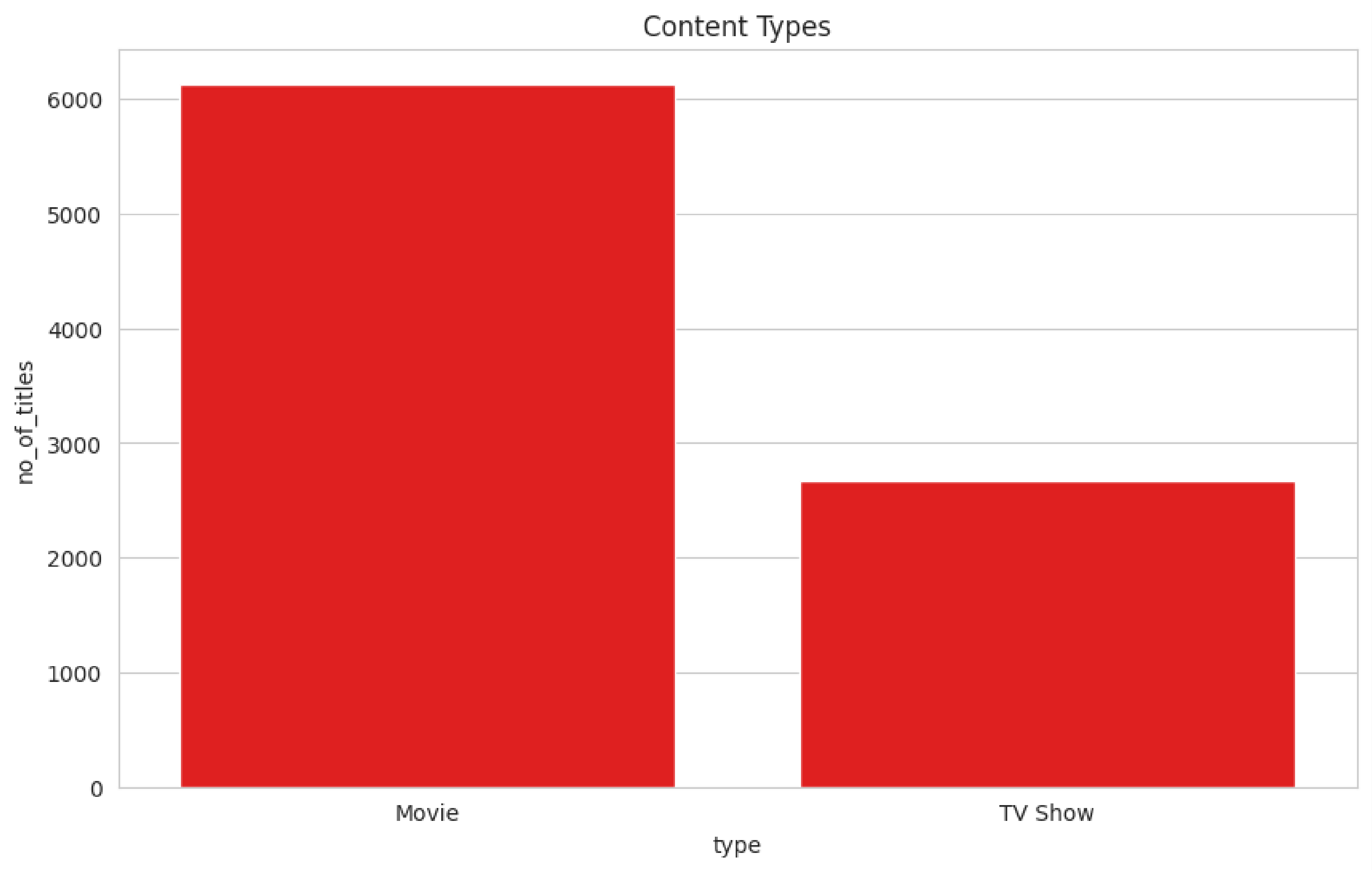
```
show_id 8807
--------------------
type 2
--------------------
title 8807
--------------------
director 5121
--------------------
cast 39297
--------------------
country 197
--------------------
date_added 1714
--------------------
release_year 74
--------------------
rating 14
--------------------
duration 220
--------------------
listed_in 73
--------------------
description 877
--------------------
year 14
--------------------
movie_min 1
--------------------
seasons_no 1
--------------------
month_added 12
--------------------
launch_time 1
```

Insights : showId is unique
Title is also unique
There are in total 73 genre

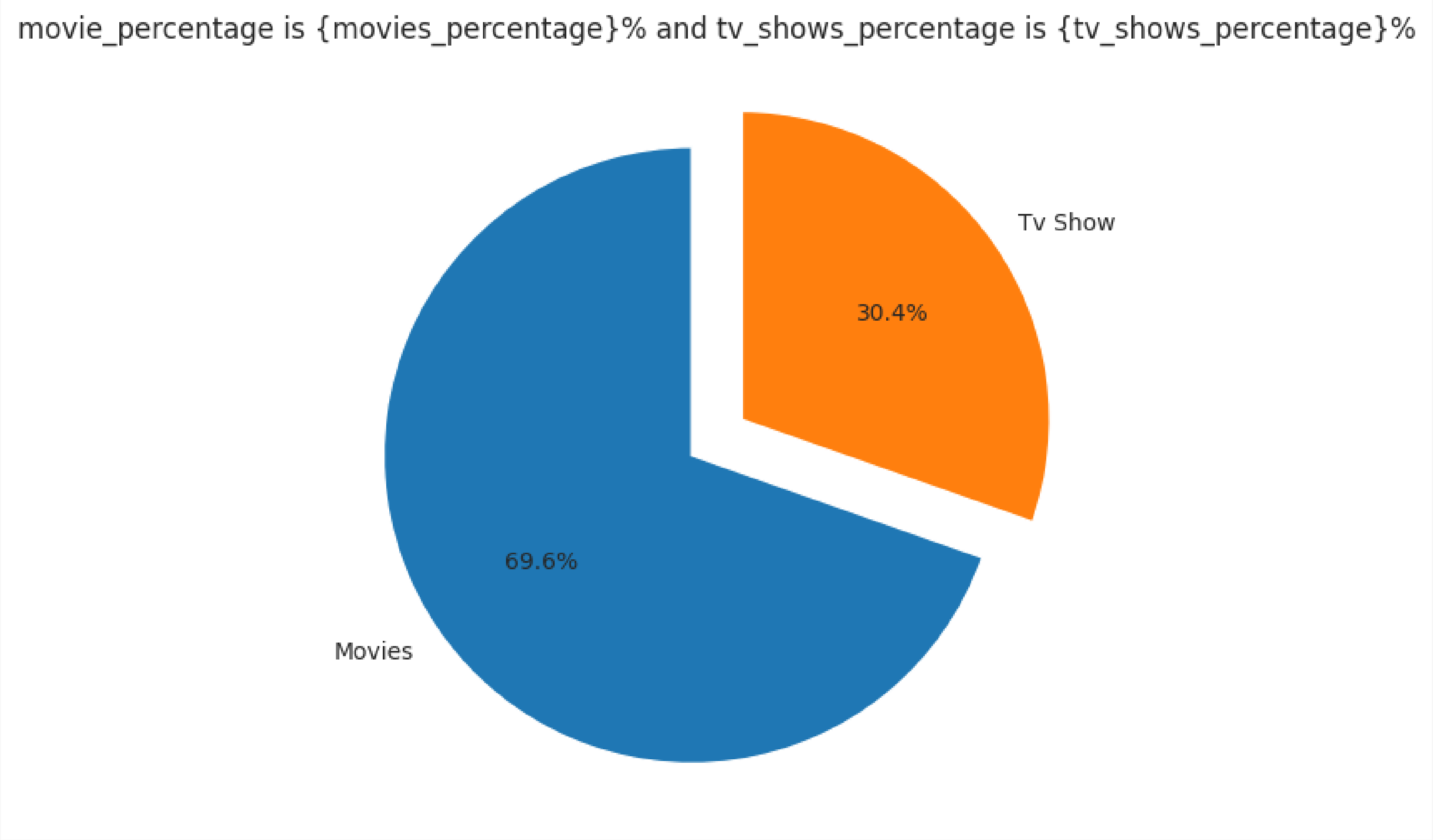
Content Types

0s

```
no_of_shows = pd.DataFrame(df.groupby('type')['show_id'].nunique()).reset_index()
no_of_shows.columns = ['type','no_of_titles']

sns.set_style("whitegrid")
plt.figure(figsize = (10,6))
sns.barplot(x = 'type',y = 'no_of_titles',data = no_of_shows,color='r')
plt.title('Content Types')
plt.show()
```

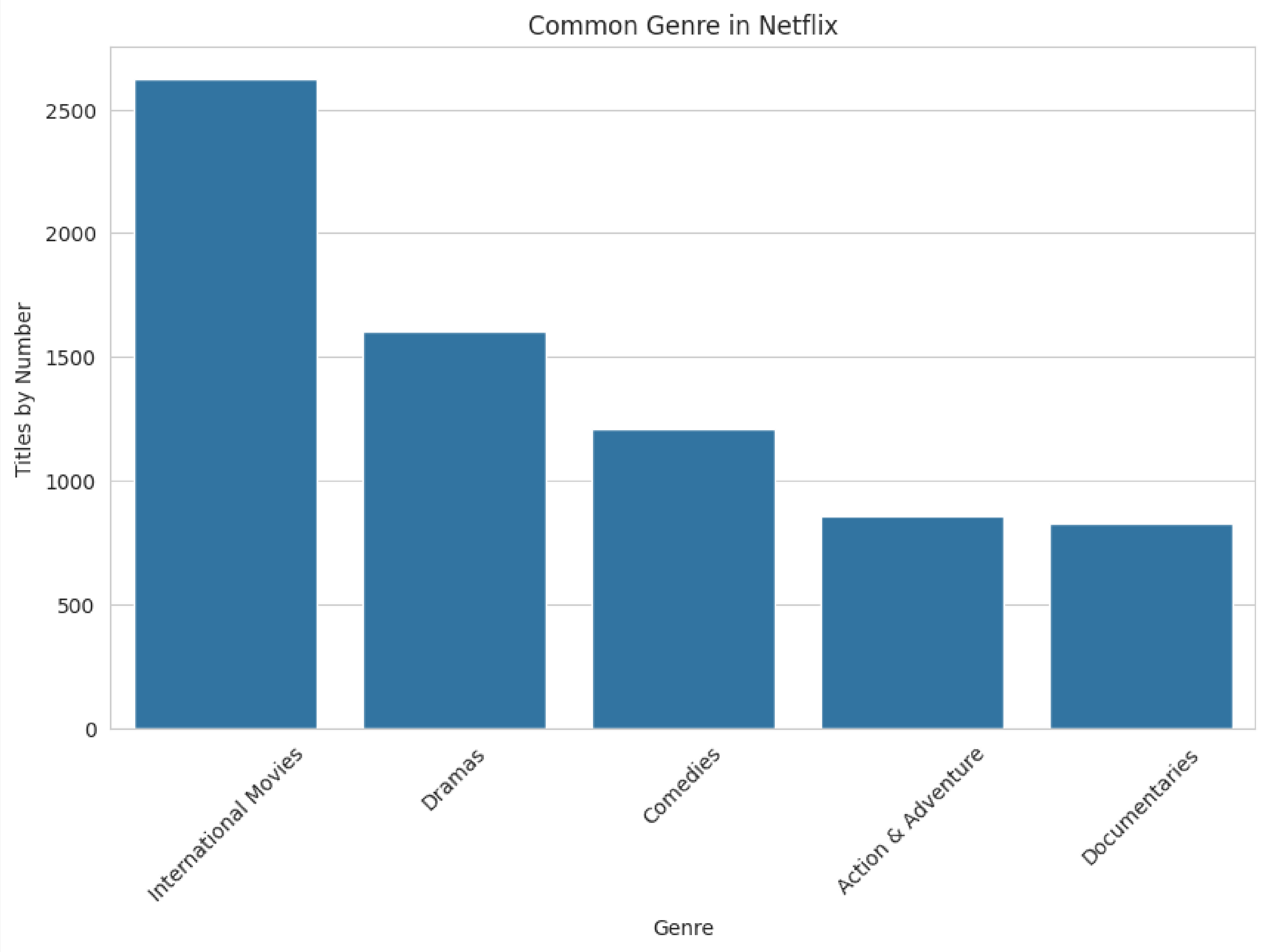


```
movies_percentage = round(df[df['type']=='Movie']['show_id'].nunique()/total_no_titles*100,2)
tv_shows_percentage = round(df[df['type']=='TV Show']['show_id'].nunique()/total_no_titles*100,2)
plt.figure(figsize = (10,6))
types = np.array([movies_percentage,tv_shows_percentage])
label = ['Movies','Tv Show']
plt.pie(types,labels = label,autopct='%1.1f%%',startangle=90,explode=(0.1,0.1))
plt.title('movie_percentage is {movies_percentage}% and tv_shows_percentage is {tv_shows_percentage}%')
plt.show()
```



Insights : Here we can there is vast difference in production of movies and tv shows . 70%Movies and 30% of TvShows

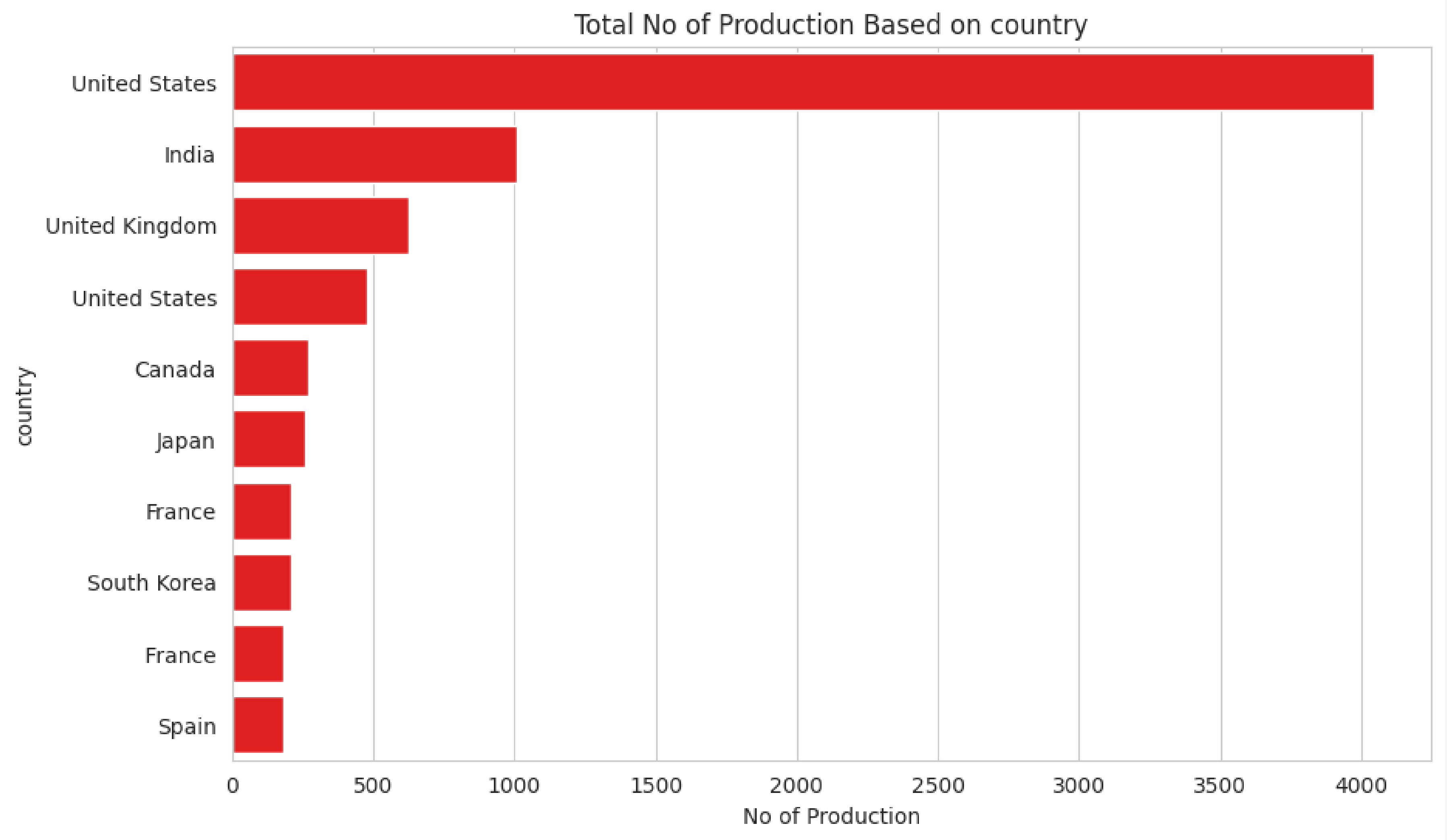
```
df_list_of_genres = pd.DataFrame(df.groupby('listed_in')['show_id'].nunique()).reset_index()
# df_listed_in
df_list_of_genres.columns = ['Genre','titles_number']

df_listed_in = df_list_of_genres.sort_values('titles_number',ascending = False).head(5)



sns.set_style("whitegrid")

plt.figure(figsize=(10, 6))
sns.barplot(data = df_listed_in, x = 'Genre',y = 'titles_number')
plt.xlabel('Genre')
plt.ylabel('Titles by Number')
plt.xticks(rotation = 45)
plt.title('Common Genre in Netflix')
plt.show()
```



Insights :: Most preferred Genre is International Movies followed by Drama

Which country has highest production of Movies and Tv shows

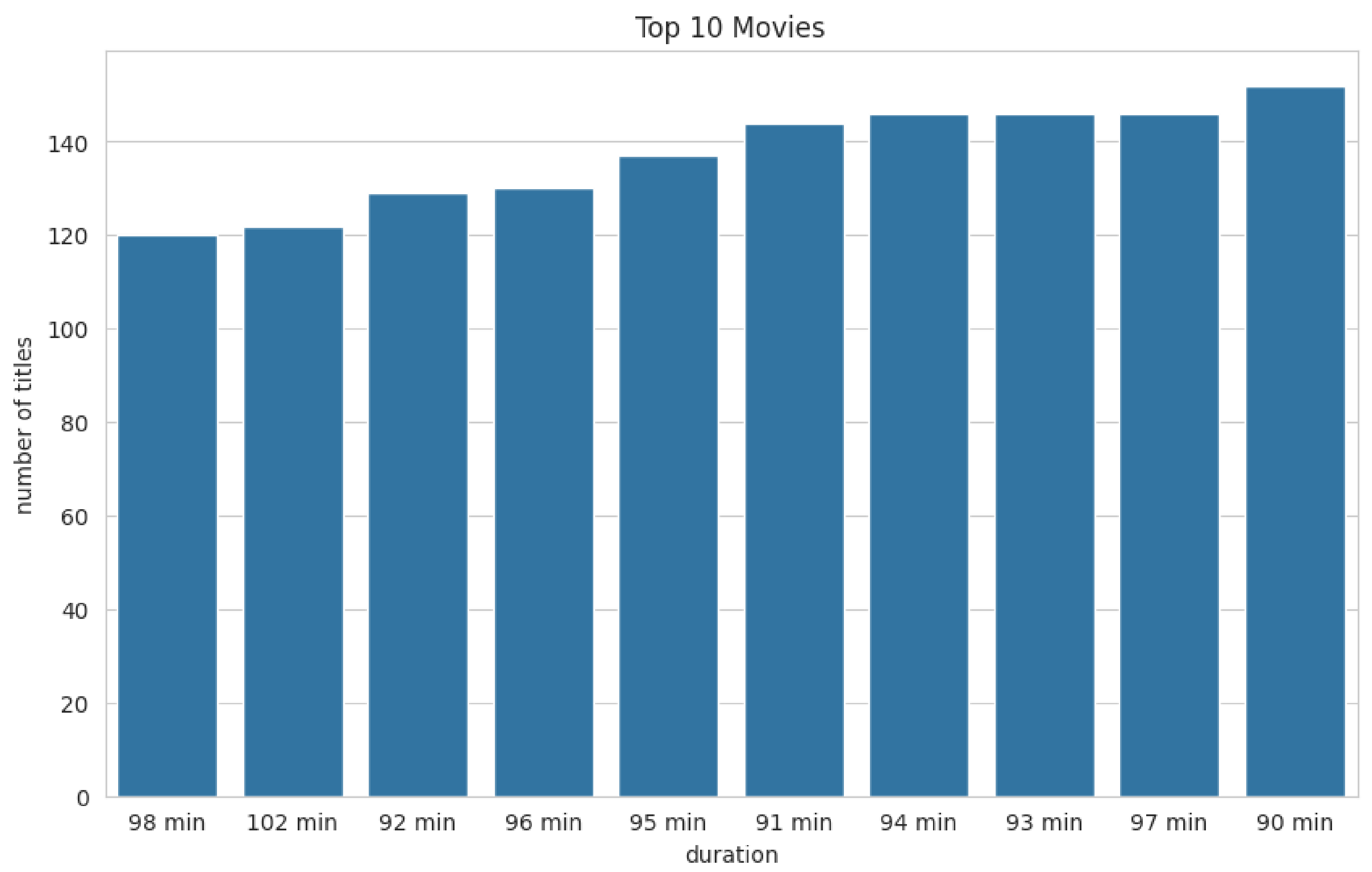
```
df_country = pd.DataFrame(df.groupby('country')['show_id'].nunique()).reset_index()
df_country.columns = ['country','No of Production']
df_country = df_country.sort_values('No of Production',ascending = False).head(10)

sns.set_style("whitegrid")

plt.figure(figsize=(10, 6))
sns.barplot(data = df_country, x = 'No of Production', y = 'country',color = 'r')
plt.title('Total No of Production Based on country')
plt.show()
```



Insights : We can see that United States holds the first position in production of shows and movies

Duration of top 10 Movies

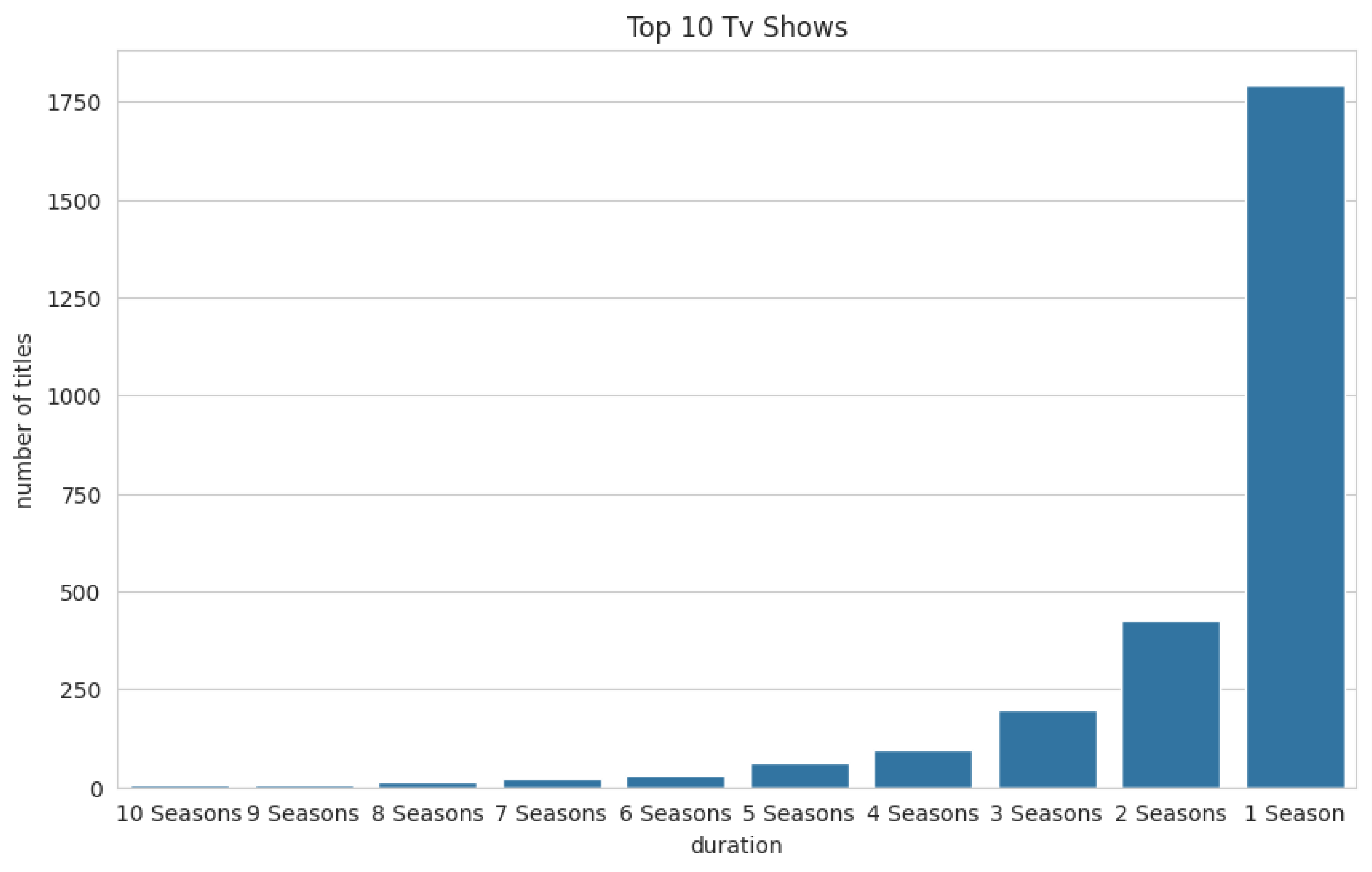
```
movies_data = df_by_duration[df_by_duration['type'] == 'Movie']
movies_data_sorted = movies_data.sort_values(by='number of titles', ascending= False).head(10)
top_10_movies_desc = movies_data_sorted.sort_values(by='number of titles', ascending=True)
plt.figure(figsize = (10,6))
sns.barplot(x = 'duration',y = 'number of titles',data = top_10_movies_desc)
plt.title('Top 10 Movies')
plt.show()
```



Insights : People like watching Movies with short duration .

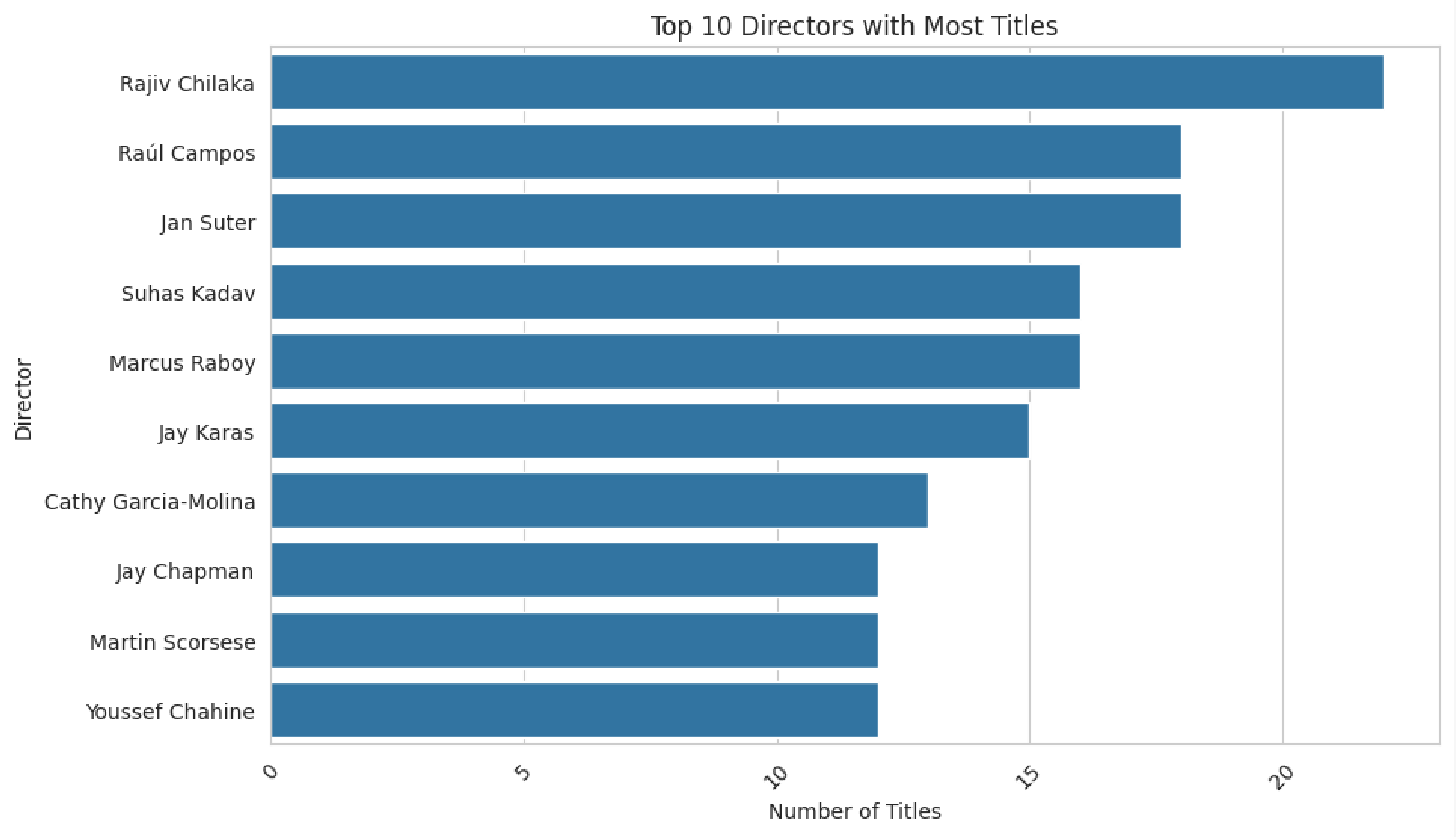
Duration of Top 10 Tv Shows

```
tv_Show_data = df_by_duration[df_by_duration['type'] == 'TV Show']
tv_Show_data_sorted = tv_Show_data.sort_values(by='number of titles', ascending= False).head(10)
top_10_tv_show_desc = tv_Show_data_sorted.sort_values(by='number of titles', ascending=True)
```



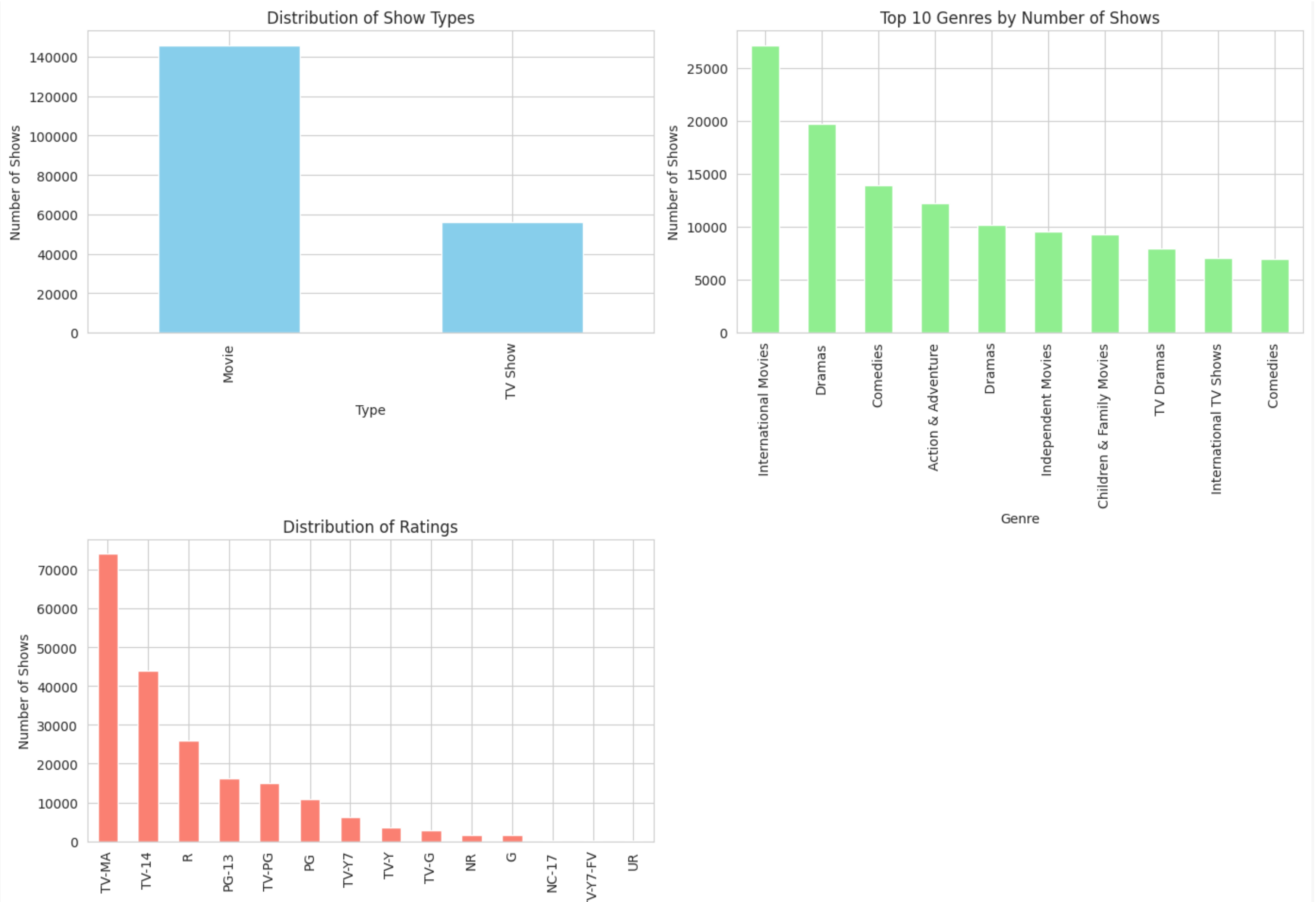
Insights: It can be noticed that people like watching shows with short durations and more engaging .

```
df_directors = pd.DataFrame(df.groupby('director')['title'].nunique()).reset_index()
df_directors = df_directors.sort_values('title',ascending = False).iloc[1:]
df_directors.head()

top_10_directors = df_directors.head(10)

plt.figure(figsize=(10, 6))
sns.barplot(x='title', y='director', data=top_10_directors)

# Add title and labels
plt.title('Top 10 Directors with Most Titles')
plt.xlabel('Number of Titles')
plt.ylabel('Director')

# Rotate x-axis labels for readability
plt.xticks(rotation=45)

# Show the plot
plt.show()
```



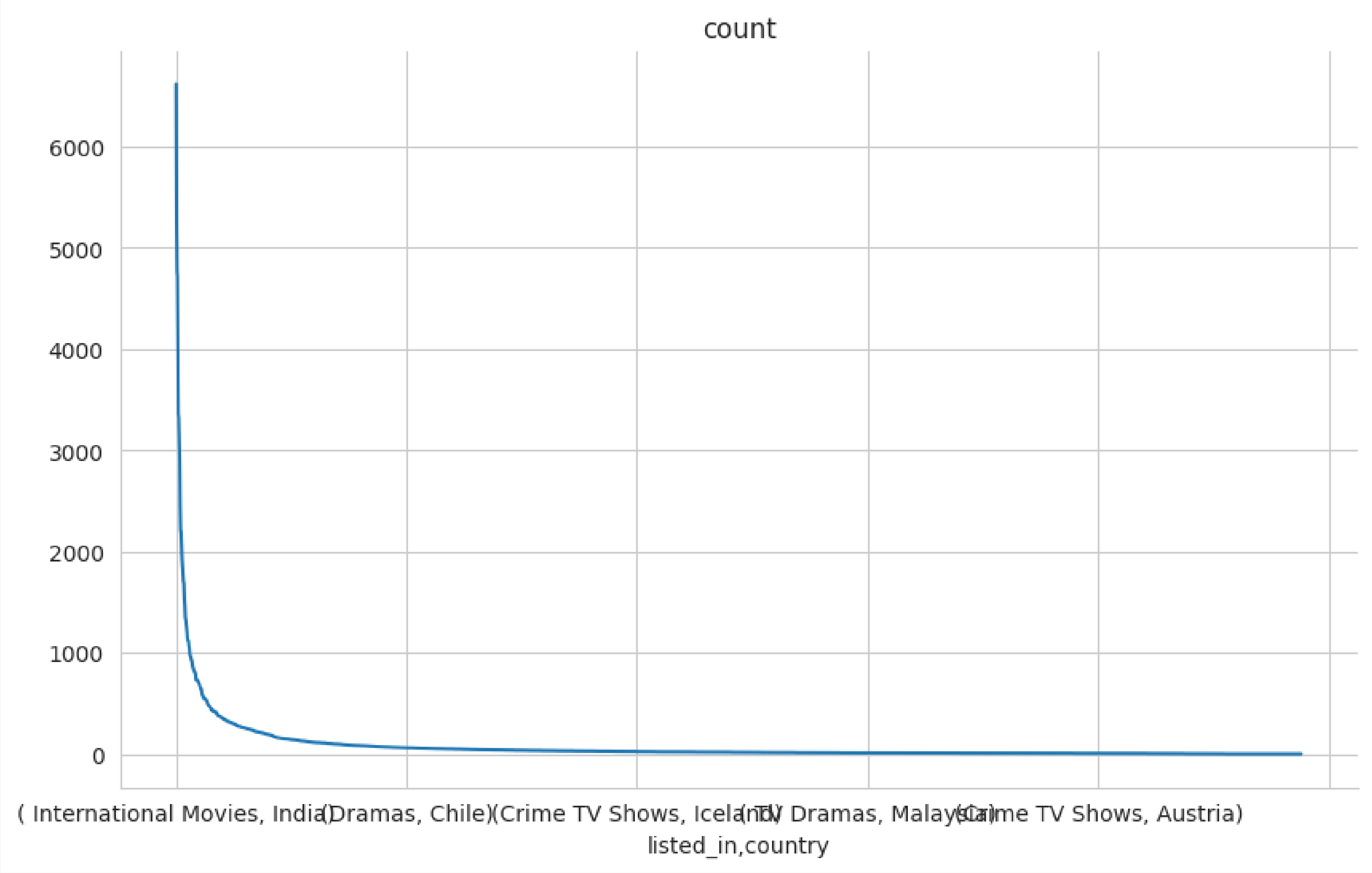
Insights: It can be seen Rajiv Chilaka is Top most director

```
Type of shows
type_counts = df['type'].value_counts()
type_counts

Genre Count
genre_counts = df['listed_in'].value_counts()
genre_counts

Rating count

rating_counts = df['rating'].value_counts()
rating_counts


plt.figure(figsize=(14, 10))

# Plot 1: Type of shows
plt.subplot(2, 2, 1)
type_counts.plot(kind='bar', color='skyblue')
plt.title('Distribution of Show Types')
plt.xlabel('Type')
plt.ylabel('Number of Shows')

# Plot 2: Genres
plt.subplot(2, 2, 2)
genre_counts.head(10).plot(kind='bar', color='lightgreen')
plt.title('Top 10 Genres by Number of Shows')
plt.xlabel('Genre')
plt.ylabel('Number of Shows')

# Plot 3: Ratings
plt.subplot(2, 2, 3)
rating_counts.plot(kind='bar', color='salmon')
plt.title('Distribution of Ratings')
plt.xlabel('Rating')
plt.ylabel('Number of Shows')

plt.tight_layout()
plt.show()
```



Genres popular accross different countries

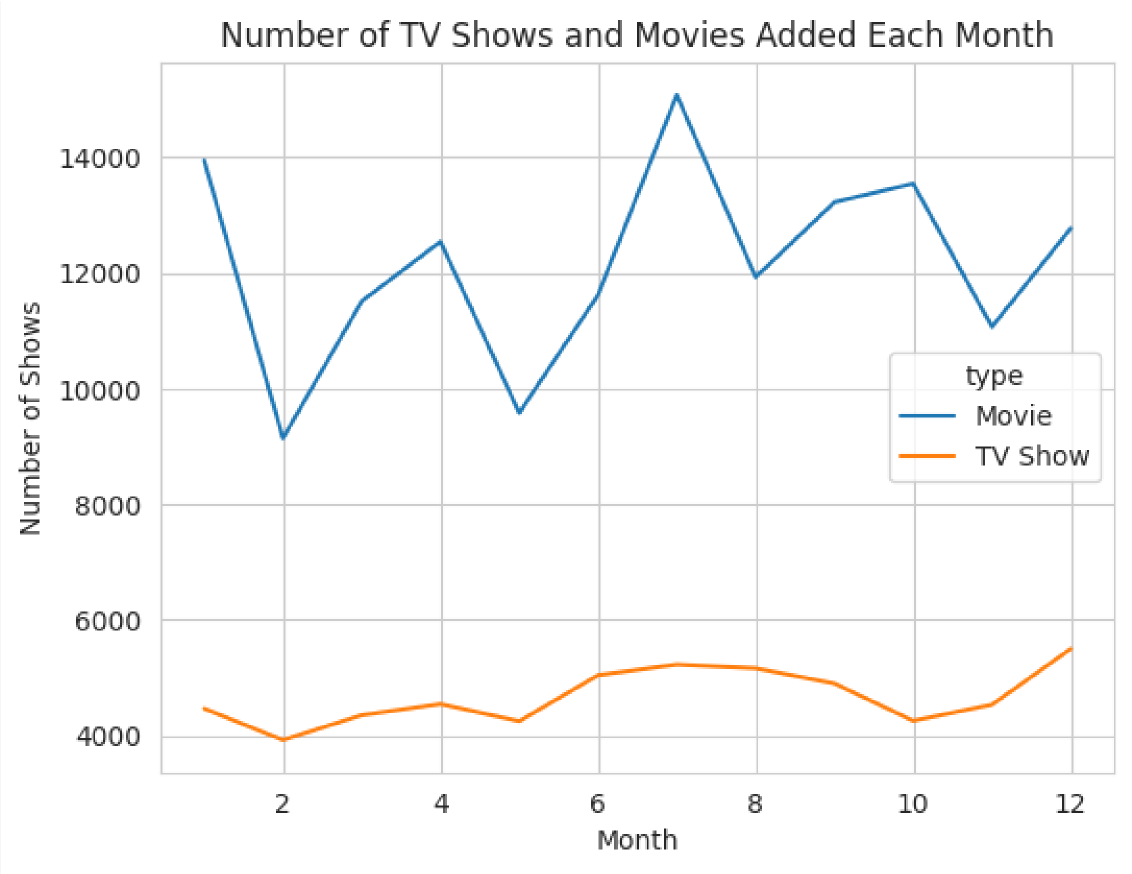
```
country_genre_counts = df.groupby(['country', 'listed_in']).size().reset_index(name='count')
#country_genre_counts = country_genre_counts.pivot(index='country', columns='listed_in', values='count')
country_genre_counts = country_genre_counts.sort_values('count', ascending=False)
country_genre_counts.head(10)



country_genre_counts = df.groupby(['country', 'listed_in']).size().reset_index(name='count')
count = country_genre_counts.groupby(['listed_in','country']).max().sort_values('count', ascending=False)
count
```



Insights : It can be seen maximum number of International Movies are seen in India and least number of crime shows in Austria

Number of Tv Shows and Movies added each month

```
df_rate = df.groupby(["month_added","type"]).agg({'type':'count'})
month = df_rate.rename(columns = {"type":"count"})
month.reset_index(inplace = True)
month.sort_values('count',ascending=False).head(20)

# Create a lineplot of the number of TV shows and movies added each month
sns.lineplot(data=month, x="month_added", y="count", hue="type")

# Set the title and axis labels
plt.title("Number of TV Shows and Movies Added Each Month")
plt.xlabel("Month")
plt.ylabel("Number of Shows")

# Show the plot
plt.show()
```



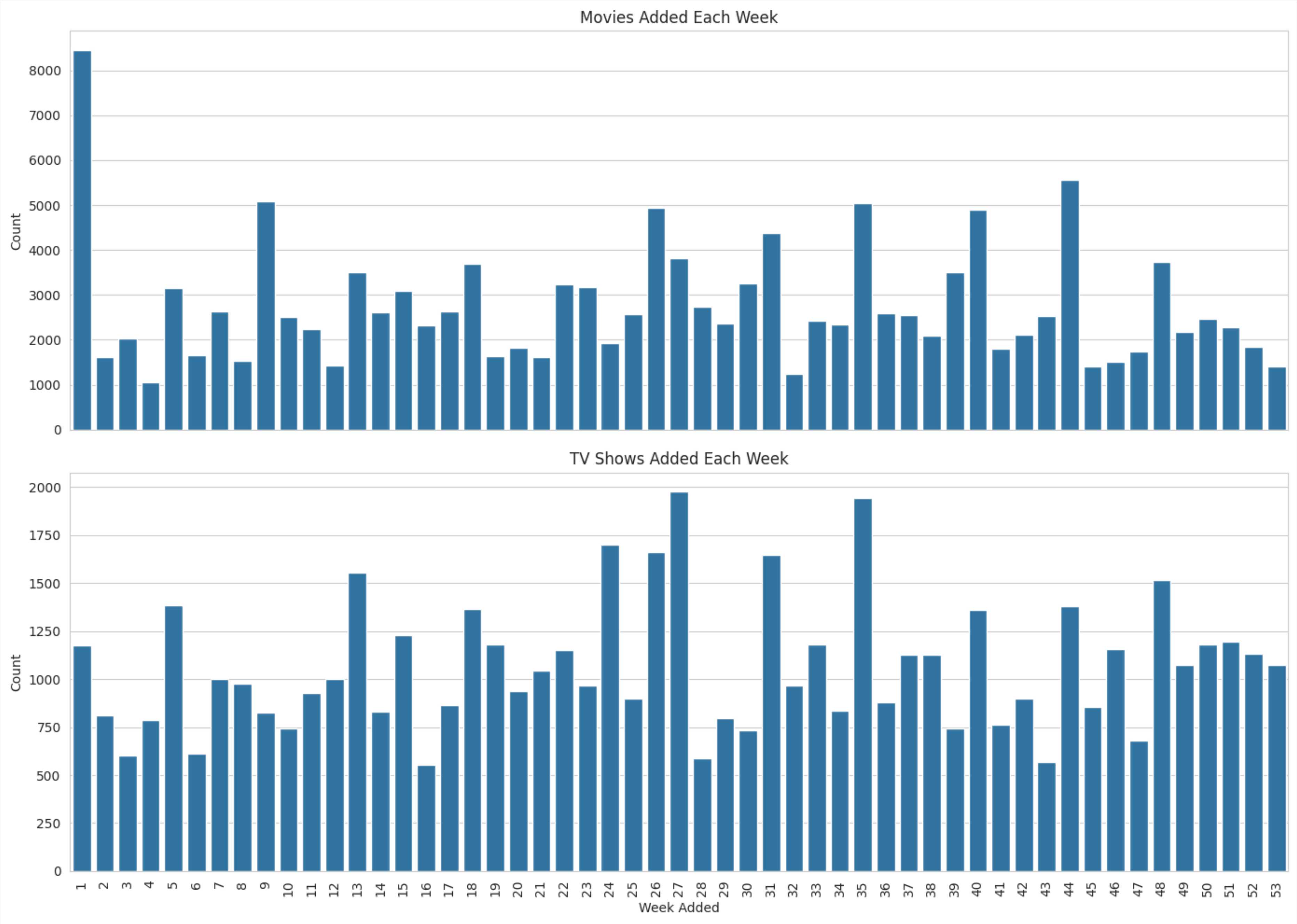
Insights: Here we can see number of Movies added is more in 7th month and least in Feb

Similarly for Tv Shows also we can notice that it has fluctuating addition pattern and more addition is seen in December month and least in Feb .

TV shows and Movies added each week

```
df['week_added'] = df['date_added'].dt.isocalendar().week
df_rate = df.groupby(["week_added","type"]).agg({'type':'count'})
week = df_rate.rename(columns = {"type":"count"})
week.reset_index(inplace = True)
week.sort_values('count',ascending=False).head(20)
movies = week[week['type'] == 'Movie']
tv_shows = week[week['type'] == 'TV Show']

# Create subplots
fig, axes = plt.subplots(2, 1, figsize=(14, 10), sharex=True)

# Plot Movies
sns.barplot(data=movies.sort_values('week_added'), x='week_added', y='count', ax=axes[0])
axes[0].set_title('Movies Added Each Week')
axes[0].set_xlabel('')
axes[0].set_ylabel('Count')
axes[0].tick_params(axis='x', rotation=90)

# Plot TV Shows
```

```
sns.barplot(data=tv_shows.sort_values('week_added'), x='week_added', y='count', ax=axes[1])
axes[1].set_title('TV Shows Added Each Week')
axes[1].set_xlabel('Week Added')
axes[1].set_ylabel('Count')
axes[1].tick_params(axis='x', rotation=90)

# Adjust layout
plt.tight_layout()
plt.show()
```



Insights : The plotting shows the count of movies and TV shows added each week throughout the years

Identifying peaks in specific weeks helps understand seasonal trends

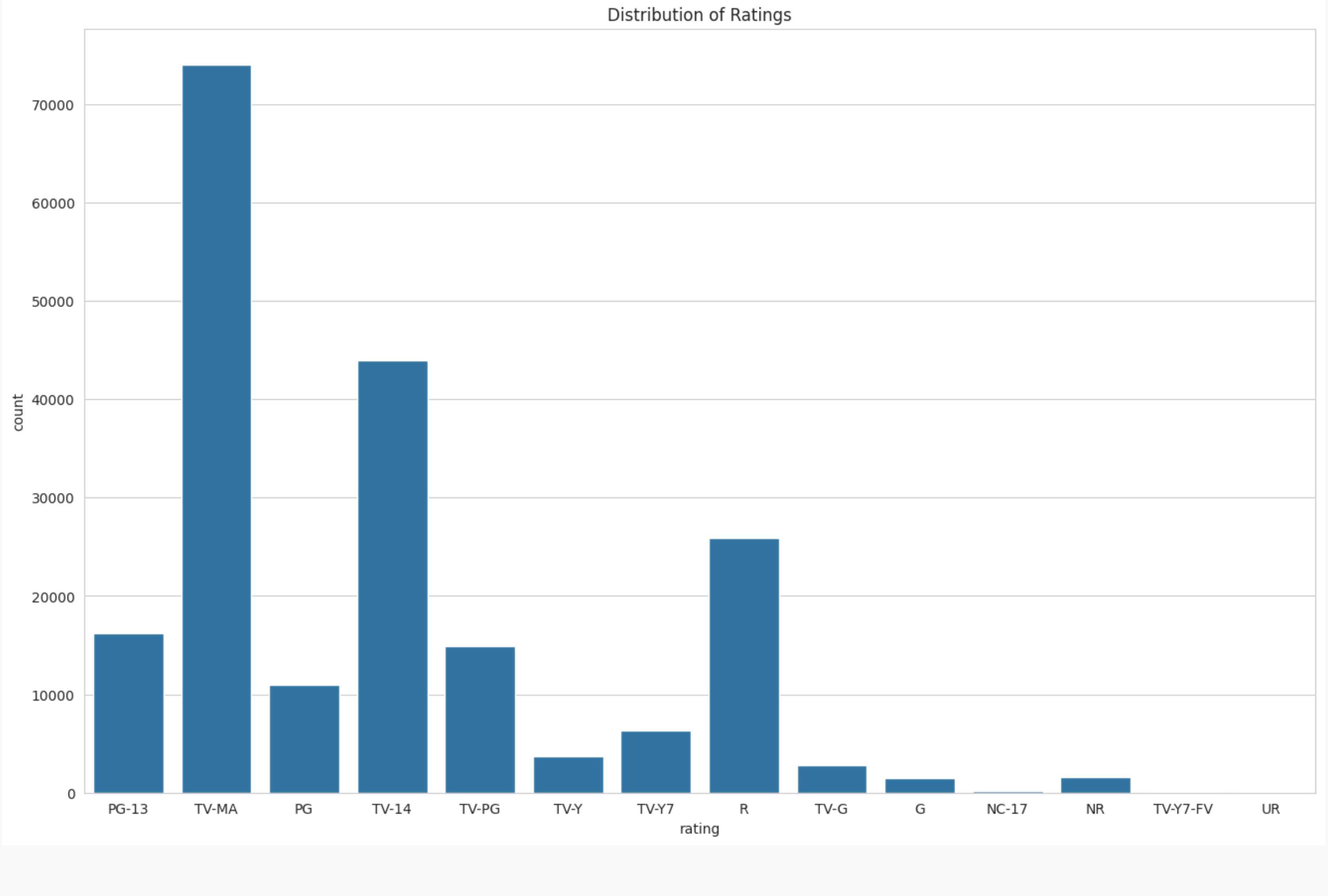
For Example high count of movies added in 27th week may be summer release strategy

Here by comparing two plots we can see that movies has almost similar release patterns where as TV

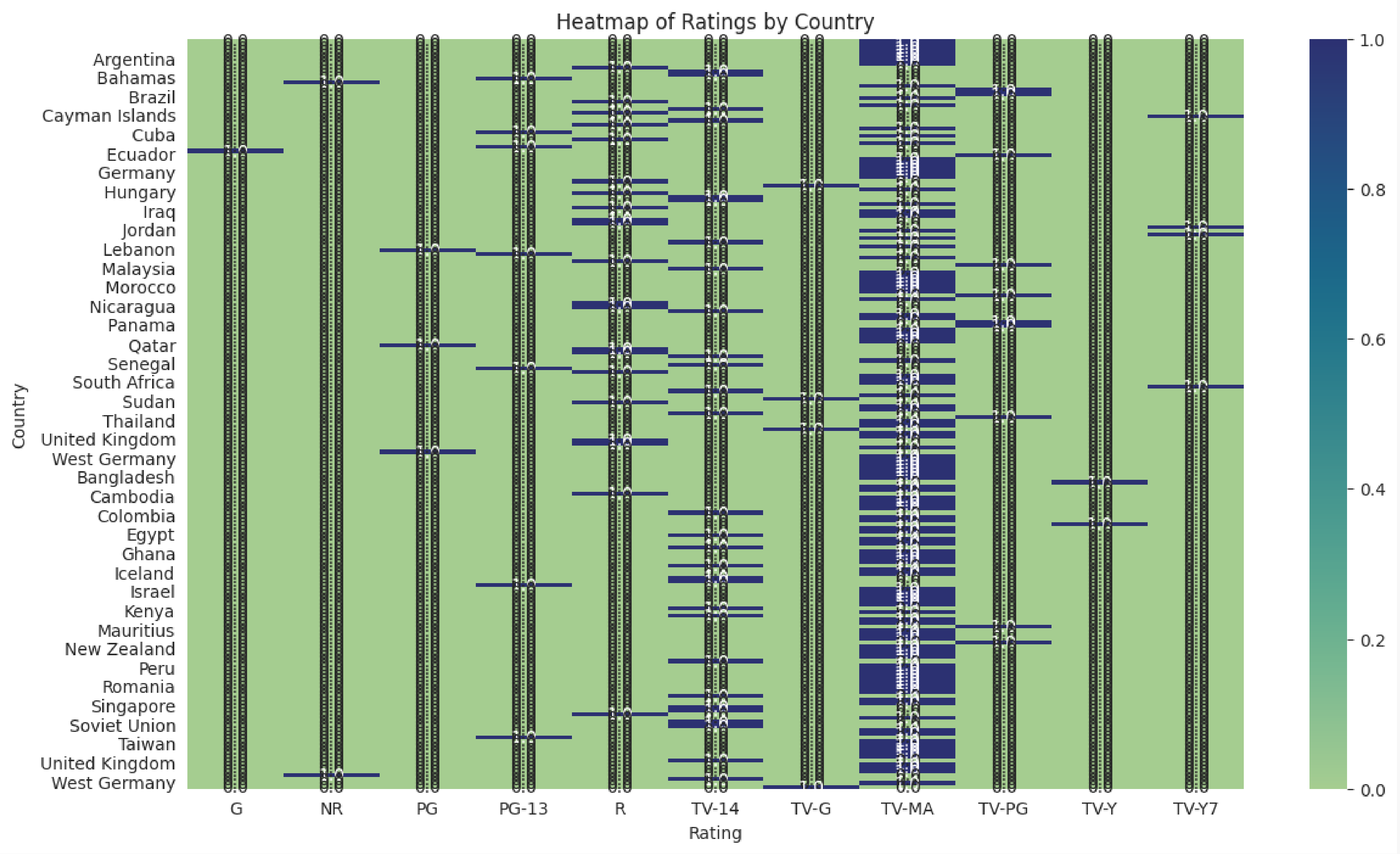
Tv Shows spike in particular weeks ,which might indicate preference for worthy release in this period.

Rating

```
plt.figure(figsize=(16, 10))
sns.countplot(x='rating', data=df)
plt.title('Distribution of Ratings')
plt.show()
```



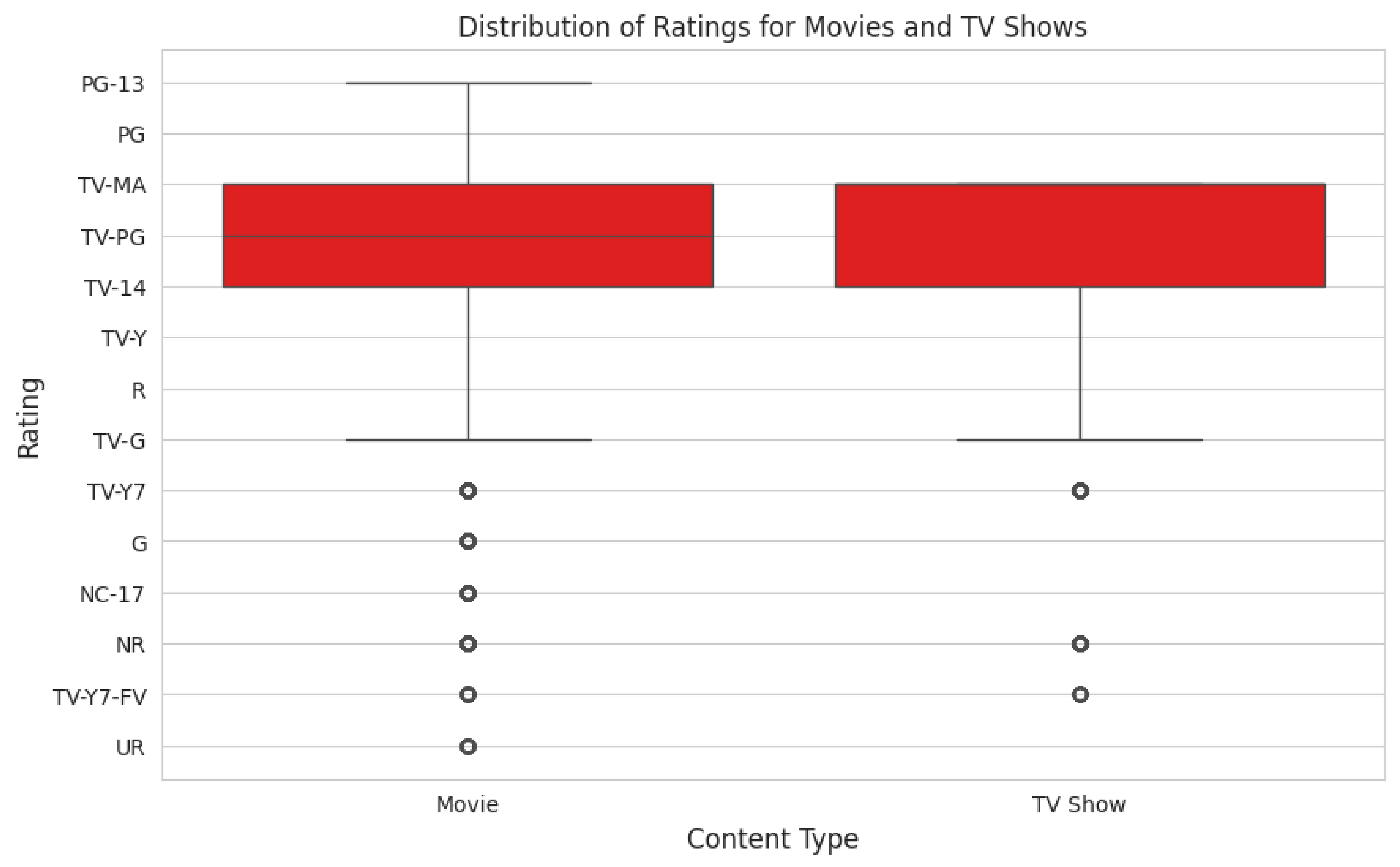
```
#Group by 'country' and 'rating' and count the number of shows in each combination
rating_counts = df.groupby(['country', 'rating']).size().reset_index(name='count')

# Find the most common rating for each country
most_common_ratings = rating_counts.loc[rating_counts.groupby('country')['count'].idxmax()]

pivot_table = most_common_ratings.pivot_table(index='country', columns='rating', aggfunc='size', fill_value=0)

# Plotting the heatmap
plt.figure(figsize=(14, 8))
sns.heatmap(pivot_table, annot=True,fmt='.1f',cmap="crest")
plt.title('Heatmap of Ratings by Country')
plt.xlabel('Rating')
plt.ylabel('Country')
plt.show()
```



Distribution of Ratings For Movies and TV Shows

0s

```
df_movies = df[df['type'] == 'Movie']
df_tv_shows = df[df['type'] == 'TV Show']
sns.set_style("whitegrid")
plt.figure(figsize=(10, 6))
sns.boxplot(x='type', y='rating', data=pd.concat([df_movies, df_tv_shows]), color = 'r')
plt.title('Distribution of Ratings for Movies and TV Shows')
plt.xlabel('Content Type',fontsize = 12)
plt.ylabel('Rating',fontsize = 12)
plt.show()
```



Insights on ratings :- Netflix uses various ratings to categories the nature of content and shows based on it

1. **PG-13**: Parents Strongly Cautioned. Some material may be inappropriate for children under 13.
2. **TV-MA**: Mature Audience Only. Specifically designed to be viewed by adults and may be unsuitable for children under 17.
3. **PG**: Parental Guidance Suggested. Some material may not be suitable for children.
4. **TV-14**: Parents Strongly Cautioned. Contains some material that many parents would find unsuitable for children under 14 years of age.
5. **TV-PG**: Parental Guidance Suggested. Contains material that parents may find unsuitable for younger children.
6. **TV-Y**: All Children. Suitable for all children.
7. **TV-Y7**: Directed to Older Children. Suitable for children age 7 and above.
8. **R**: Restricted. Contains some adult material. Parents are urged to learn more about the film before taking their children to see it.
9. **TV-G**: General Audience. Suitable for all ages.
10. **G**: General Audiences. All ages admitted.
11. **NC-17**: Adults Only. Clearly adult. Children are not admitted.
12. **TV-Y7-FV**: Directed to Older Children - Fantasy Violence. Suitable for children age 7 and older, with fantasy violence.
13. **UR**: Unrated. The content has not been rated by a recognized rating system.

    By identifying common rating in particular region we can understand regional content preferences .For example ,Here we can see TV-MA is most common rating in United states which is mature content .

    As here it is seen TV-MA,TV-14 ,TV-PG and R is mostly preferred ratings **.**

Business Insights

Netflix has majority of content released after 2018 .It is seen content for earlier years is less and hence could not engage senior citizens.IT can try and engage senior citizen by targeting senior citizen audience .

As we saw earlier, Maximum content is of TV-MA , TV-14 or PG and R . Which means 80% of content is either for adult or for children with parental control options.It could target on TV-G and for younger childrens who could be engaged in future .

Most of the Genre in Netflix is international movies and shows .We can increase audience engagement by more and more preferred genre in particular country .

Only top 10 countries contribute to the 70 % of Netflix content and rest comes for remaining countries hence Netflix could engage more and more countries to increase business and relatable audience

Even We could consider the duration of shows and Movies and work in accordance with it for upcoming shows and seasons .As we saw maximum viewers like watching movies having one seasons or with minimum number of time frame .

Consider what competitors are producing and identify gaps or opportunities where Netflix can differentiate itself from .

Netflix Should Focus more on producing movies along with Tv Shows according to the what we have seen from the given data

Recommendations

Very limited Genre has been Focused in other countries except United States .Hence every country area should try and add their cultural instinct to it and engage more audience through it .Determine the regional preferences for particular genre and type of content and particular target audience .

Collaborate with local content creator, Producers and distributors to strengthen the market

place

Try and release more and more original and something new story targeting on different audience groups and keep the waiting period short as now people keep searching new contents more and more.

Netflix Should Focus more on producing movies considering all kinds of ratings and delivering high quality content to audience.

Google Colab File Link

https://colab.research.google.com/drive/1zBBvD_QqER4KM9VAliTtNfZRRDNrp4Fe?usp=sharing