```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd
path = "/content/drive/MyDrive/netflix data.csv"
df = pd.read_csv(path)

#dataset is now stored in a pandas dataframe

df.head(8808)
```

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021
...	...	...	...	...	...	...	...	...

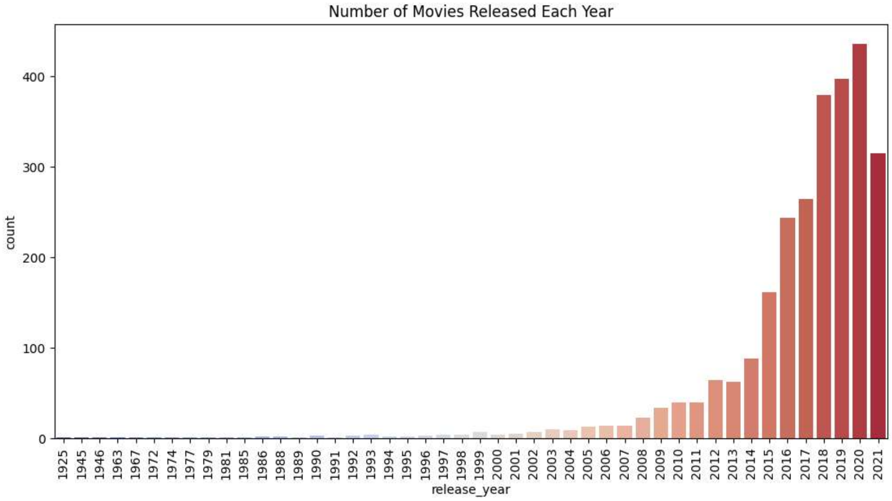
```
Q.1 # How has the number of movies released per year changed over the last 20-30 years?
```

```
# Filter for movies

movies_df = df[df['type'] == 'Movie']

current_year = pd.Timestamp.now().year

movies_last_20_30_years = movies_df[(movies_df['release_year'] >= current_year - 30) &
                                    (movies_df['release_year'] <= current_year - 20)]


# Group by 'release_year' and count the number of movies

movies_per_year = movies_last_20_30_years.groupby('release_year').size().reset_index(name= 'count')

print(movies_per_year)
```

```
    release_year  count
0           1994     20
1           1995     23
2           1996     21
3           1997     34
4           1998     32
5           1999     32
6           2000     33
7           2001     40
8           2002     44
9           2003     51
10          2004     55
```

```
# Countplot for release year

plt.figure(figsize=(12, 6))
sns.countplot(x='release_year', data=movies_df, palette='coolwarm')
plt.title('Number of Movies Released Each Year')
plt.xticks(rotation=90)
plt.show()
```

```
<ipython-input-22-b821235954ef>:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.

  sns.countplot(x='release_year', data=movies_df, palette='coolwarm')
```



INSIGHTS :

There has been a general upward trend in the number of movies released per year from the mid-1990s to the early 2000s. The counts consistently increase from around 20-30 movies per year in the mid-1990s to over 50 movies per year by the early 2000s.

Eventhough there is an increase in number of movies there is a fluctuation happened in the year 1998 to 2000

After 2000, the count of movies has stabilized there is no fluctuation from 2000 to 2004

RECOMMENDATIONS :

Check if there is any seasonal influence on the pattern of movie releases as it may explain the fluctuation from 1998 to 2004.

Perform a genre-specific analysis to understand if certain genres have driven the overall increase in movie releases or if there are shifts in popularity among genres over time.

Check on impact of technology in film-making and streaming platforms which has impacted the increased production of movies and TV shows.

```
Q.2 # Comparison of tv shows vs. movies.
```

```
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd

path = "/content/drive/MyDrive/netflix data.csv"

df = pd.read_csv(path)
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
movies_df = df[df['type'] == 'Movie']

tv_shows_df = df[df['type'] == 'TV Show']

#comparison on COUNT

num_movies_df = len(movies_df)
num_tv_shows_df = len(tv_shows_df)

print(f"Number of TV Shows: {num_tv_shows_df}")
print(f"Number of Movies: {num_movies_df}")
```

```
Number of TV Shows: 2676
Number of Movies: 6131
```

```
# Comparison on GENRE

from google.colab import drive
drive.mount('/content/drive')

import pandas as pd
path = "/content/drive/MyDrive/netflix data.csv"
df = pd.read_csv(path)
```

```
movies_df = df[df['type'] == 'Movie']

tv_shows_df = df[df['type'] == 'TV Show']

tv_genre_counts = tv_shows_df['listed_in'].value_counts()
movie_genre_counts = movies_df['listed_in'].value_counts()
print("Genre distribution in TV Shows:")
print(tv_genre_counts)
print("Genre distribution in Movies:")
print(movie_genre_counts)
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
Genre distribution in TV Shows:
```

```
listed_in
Kids' TV                                                 220
International TV Shows, TV Dramas                        121
Crime TV Shows, International TV Shows, TV Dramas        110
Kids' TV, TV Comedies                                     99
Reality TV                                                95
                                                        ... 
Kids' TV, TV Action & Adventure, TV Dramas                 1
British TV Shows, Kids' TV, TV Thrillers                   1
Reality TV, TV Horror, TV Thrillers                        1
TV Action & Adventure, TV Horror, TV Sci-Fi & Fantasy      1
Classic & Cult TV, Crime TV Shows, TV Dramas               1
Name: count, Length: 236, dtype: int64
Genre distribution in Movies:
listed_in
Dramas, International Movies                        362
Documentaries                                       359
Stand-Up Comedy                                     334
Comedies, Dramas, International Movies              274
Dramas, Independent Movies, International Movies    252
                                                   ... 
Sci-Fi & Fantasy                                      1
Sports Movies                                         1
Children & Family Movies, Comedies, Cult Movies       1
Cult Movies, Dramas, Music & Musicals                 1
Cult Movies, Dramas, Thrillers                        1
Name: count, Length: 278, dtype: int64
```

```
# Statistical summary

print("TV Shows DataFrame:")
print(tv_shows_df.head())
print()

print("Movies DataFrame:")
print(movies_df.head())
print()


print("TV Shows Columns:")
print(tv_shows_df.columns)
print()

print("Movies Columns:")
print(movies_df.columns)
print()
```

```
TV Shows DataFrame:
  show_id     type                  title         director  \
1      s2  TV Show          Blood & Water              NaN   
2      s3  TV Show              Ganglands  Julien Leclercq   
3      s4  TV Show  Jailbirds New Orleans              NaN   
4      s5  TV Show           Kota Factory              NaN   
5      s6  TV Show          Midnight Mass    Mike Flanagan   

                                                cast       country  \
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...  South Africa   
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...           NaN   
3                                                NaN           NaN   
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...         India   
5  Kate Siegel, Zach Gilford, Hamish Linklater, H...           NaN   

           date_added  release_year rating   duration  \
1  September 24, 2021          2021  TV-MA  2 Seasons   
2  September 24, 2021          2021  TV-MA   1 Season   
3  September 24, 2021          2021  TV-MA   1 Season   
4  September 24, 2021          2021  TV-MA  2 Seasons   
5  September 24, 2021          2021  TV-MA   1 Season   

                                           listed_in  \
1    International TV Shows, TV Dramas, TV Mysteries   
2  Crime TV Shows, International TV Shows, TV Act...   
3                             Docuseries, Reality TV   
4  International TV Shows, Romantic TV Shows, TV ...   
5                 TV Dramas, TV Horror, TV Mysteries   

                                         description  
1  After crossing paths at a party, a Cape Town t...  
2  To protect his family from a powerful drug lor...  
3  Feuds, flirtations and toilet talk go down amo...  
4  In a city of coaching centers known to train I...  
```

```
5  The arrival of a charismatic young priest brin...

Movies DataFrame:
   show_id   type                             title  \
0       s1  Movie              Dick Johnson Is Dead   
6       s7  Movie  My Little Pony: A New Generation   
7       s8  Movie                           Sankofa   
9      s10  Movie                       The Starling   
12     s13  Movie                       Je Suis Karl   

                         director  \
0                 Kirsten Johnson   
6   Robert Cullen, José Luis Ucha   
7                    Haile Gerima   
9                  Theodore Melfi   
12            Christian Schwochow   

                                                 cast  \
0                                                 NaN   
6   Vanessa Hudgens, Kimiko Glenn, James Marsden, ...   
7   Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...   
9   Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...   
12  Luna Wedler, Jannis Niewöhner, Milan Peschel, ...   
```

```
# Statistical summary for TV Shows
print("Statistical Summary for TV Shows:")
print(tv_shows_df.describe())
print()

# Statistical summary for Movies
print("Statistical Summary for Movies:")
print(movies_df.describe())
```

```
Statistical Summary for TV Shows:
       release_year
count   2676.000000
mean    2016.605755
std        5.740138
min     1925.000000
25%     2016.000000
50%     2018.000000
75%     2020.000000
max     2021.000000

Statistical Summary for Movies:
       release_year
count   6131.000000
mean    2013.121514
std        9.678169
min     1942.000000
25%     2012.000000
50%     2016.000000
75%     2018.000000
max     2021.000000
```

INSIGHTS :

TV Shows: The most prevalent genres are Kids' TV, International TV Shows, TV Dramas, and Crime TV Shows, International TV Shows, TV Dramas. Movies: The top genres include Dramas, International Movies, Documentaries, and Stand-Up Comedy.

The number of movies launch is 6131 and that of TV shows is 2676.

RECOMMENDATIONS :

TV Shows: Given the prominence of Kids' TV and International TV Shows, TV Dramas, consider expanding the production of content in these genres. Focus on developing engaging and culturally diverse series to cater to both younger audiences and international viewers.

Movies: With Dramas, International Movies and Documentaries leading in movies, continue to prioritize these genres. Additionally, explore opportunities to produce more Stand-Up Comedy specials, as they have shown popularity.

Tailor marketing campaigns and promotional efforts specifically for each genre. Use demographic data and viewer analytics to refine content recommendations and enhance viewer engagement.

Stay updated on evolving genre trends and audience preferences. Monitor shifts in viewing habits and genre popularity to adapt content strategies in real-time.

While focusing on popular genres, also encourage experimentation and innovation.

```
Q.3  What is the best time to launch TV Shows?

import pandas as pd

# Count the number of TV shows released in each year
release_year_counts = tv_shows_df['release_year'].value_counts()

# Find the year with the highest number of TV shows releases
max_release_count = release_year_counts.max()
best_release_years = release_year_counts[release_year_counts == max_release_count].index.tolist()

print(f"The best time(s) to launch a TV show based on release year:")
for year in best_release_years:
    print(f"- {year}: {max_release_count} TV shows released.")
```

```
Object `Shows` not found.
The best time(s) to launch a TV show based on release year:
- 2020: 436 TV shows released.
```

```
# Count the number of TV shows released in each year, grouped by genre
genre_counts = tv_shows_df.groupby(['listed_in', 'release_year']).size().reset_index(name='count')

# Find the genre with the highest number of TV show releases in each year
max_counts = genre_counts.groupby(['listed_in'])['count'].max()
best_years = genre_counts.merge(max_counts, on=['listed_in', 'count'])

print("Best time to launch a TV show based on genre:")
print(best_years)
```

```
Best time to launch a TV show based on genre:
                              listed_in  release_year  count
0                          Anime Series          2020      3
1                          Anime Series          2021      3
2          Anime Series, Crime TV Shows          2008      1
3          Anime Series, Crime TV Shows          2011      1
4          Anime Series, Crime TV Shows          2019      1
..                                  ...           ...    ...
363  TV Sci-Fi & Fantasy, TV Thrillers          2017      1
364                            TV Shows          2017      2
365                            TV Shows          2019      2
366                            TV Shows          2020      2
367                            TV Shows          2021      2

[368 rows x 3 columns]
```

***INSIGHTS : ***

The year 2020 saw the highest number of TV show releases, with 436 TV shows launched. This suggests that 2020 was a prolific year for TV show debuts, possibly indicating favorable conditions or trends in the industry at that time.

The dataset includes various genres associated with TV shows released across different years. Understanding the genre distribution can help identify which genres were popular in specific years and potentially predict audience preferences for future launches.

***RECOMMENDATIONS : ***

considering launching during peak years like 2020 could align with periods of higher industry activity and audience engagement.

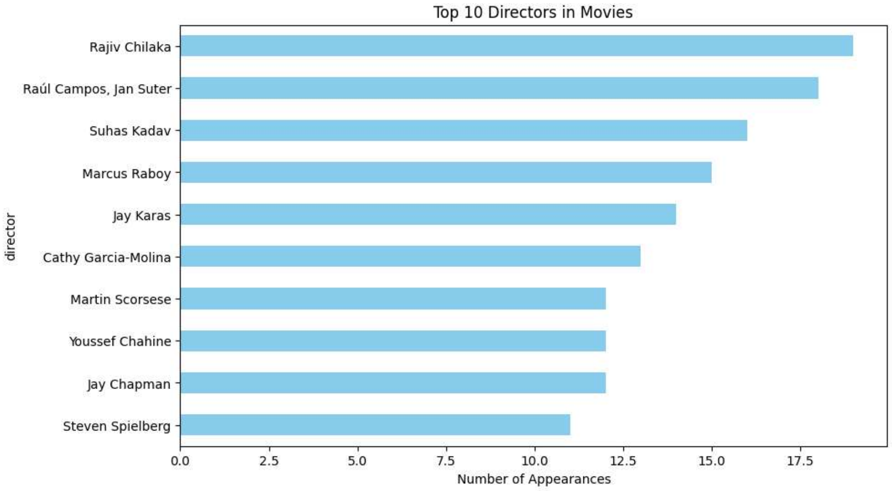
Continuously monitor industry trends and audience preferences to adapt launch strategies accordingly. Stay informed about emerging genres, content formats, and distribution platforms that could influence the success of TV show launches.

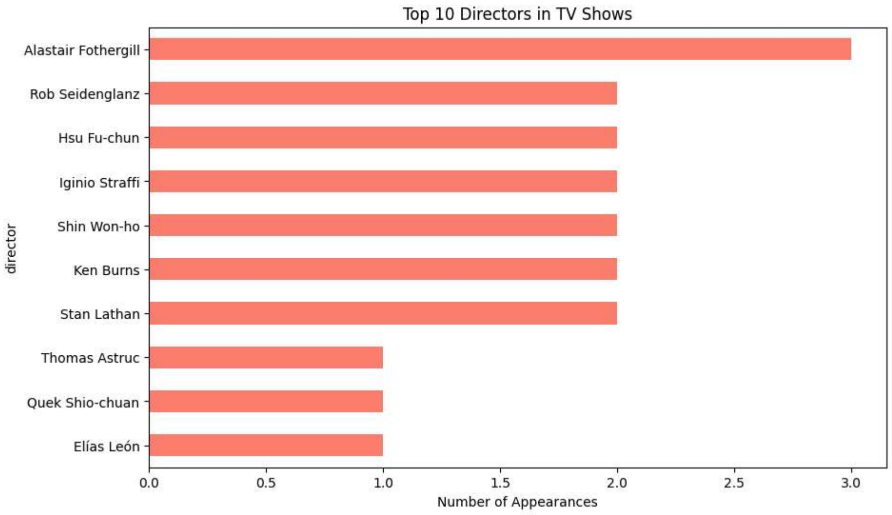
Avoid launching during periods of high competition or when audience attention may be diverted by major sporting events, holidays, or blockbuster movie releases.

```
Q.4 Analysis of actors/directors of different types of shows/movies?


# Counting occurrences of each director in movies and TV shows

director_counts_movies = movies_df['director'].value_counts()
director_counts_tv_shows = tv_shows_df['director'].value_counts()

# Display top directors by appearances

print("Top Directors in Movies:")
print(director_counts_movies.head(10))
print()

print("Top Directors in TV Shows:")
print(director_counts_tv_shows.head(10))
print()
```

```
Object `movies` not found.
Top Directors in Movies:
director
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Suhas Kadav               16
Marcus Raboy              15
Jay Karas                 14
Cathy Garcia-Molina       13
Martin Scorsese           12
Youssef Chahine           12
Jay Chapman               12
Steven Spielberg          11
Name: count, dtype: int64

Top Directors in TV Shows:
director
Alastair Fothergill    3
Rob Seidenglanz        2
Hsu Fu-chun            2
Iginio Straffi         2
Shin Won-ho            2
Ken Burns              2
Stan Lathan            2
Thomas Astruc          1
Quek Shio-chuan        1
Elías León             1
Name: count, dtype: int64
```

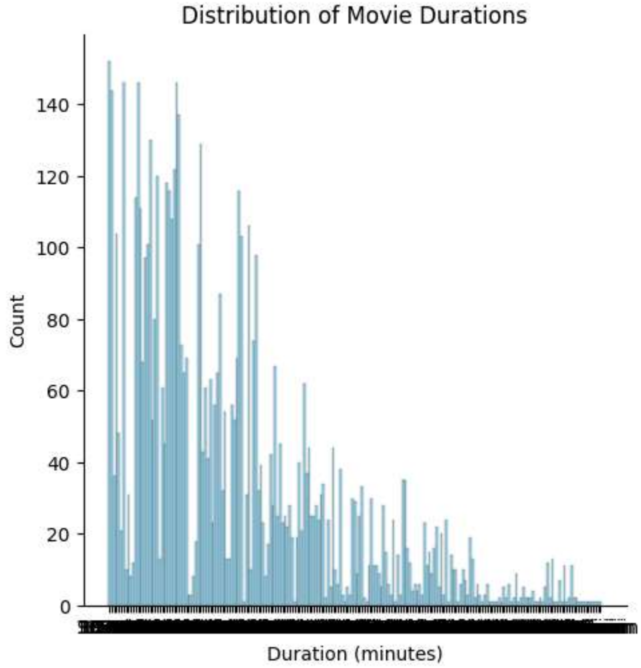
```
#Visualization

import matplotlib.pyplot as plt

# Plot top directors in movies

plt.figure(figsize=(10, 6))
director_counts_movies.head(10).plot(kind='barh', color='skyblue')
plt.title('Top 10 Directors in Movies')
plt.xlabel('Number of Appearances')
plt.gca().invert_yaxis()
plt.show()

# Plot top directors in TV shows

plt.figure(figsize=(10, 6))
director_counts_tv_shows.head(10).plot(kind='barh', color='salmon')
plt.title('Top 10 Directors in TV Shows')
plt.xlabel('Number of Appearances')
plt.gca().invert_yaxis()
plt.show()
```



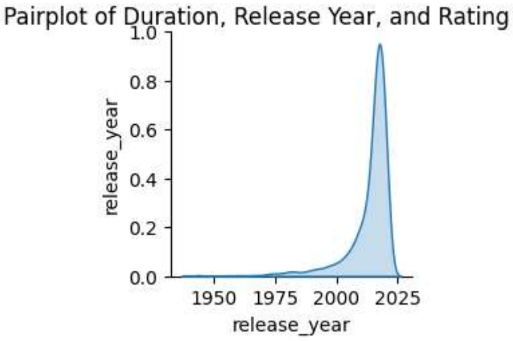


```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


# Displot

plt.figure(figsize=(10, 6))
sns.displot(movies_df['duration'], bins=20, kde=False, color='skyblue')
plt.title('Distribution of Movie Durations')
plt.xlabel('Duration (minutes)')
plt.ylabel('Count')
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



```
#Pairplot for selected variables

sns.pairplot(movies_df[['duration', 'release_year', 'rating']], diag_kind='kde')
plt.suptitle('Pairplot of Duration, Release Year, and Rating', y=1.02)
plt.show()
```



INSIGHTS :

The top directors in movies and TV shows have varying levels of involvement, with some directing a significant number of productions (example : Rajiv Chilaka with 19 movies) compared to others who have directed fewer.

Directors like Rajiv Chilaka, known for a large number of movies, may have a distinct style or preference for certain genres or formats. Understanding their strengths and preferences can help predict the type of content they are likely to produce in the future.

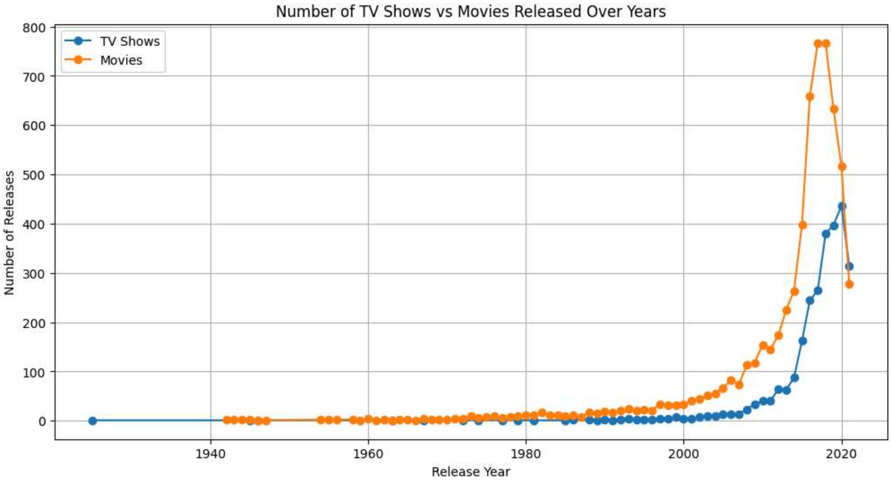
RECOMMENDATIONS :

Encourage collaboration between top directors and diverse talent pools to foster creativity and innovation in both movies and TV shows. Partnering with directors who have a strong track record can enhance the quality and appeal of productions.

Tailor content strategies based on the director's past successes and audience preferences. Directors with a proven track record in specific genres or styles can attract a loyal audience base that enjoys their particular brand of storytelling.

```
Q.5 Does Netflix has more focus on TV Shows than movies in recent years?

import pandas as pd
import matplotlib.pyplot as plt


# Counting releases by year for TV Shows and Movies
tv_shows_count = tv_shows_df['release_year'].value_counts().sort_index()
movies_count = movies_df['release_year'].value_counts().sort_index()

# Plotting the data
plt.figure(figsize=(12, 6))
plt.plot(tv_shows_count.index, tv_shows_count.values, label='TV Shows', marker='o')
plt.plot(movies_count.index, movies_count.values, label='Movies', marker='o')
plt.title('Number of TV Shows vs Movies Released Over Years')
plt.xlabel('Release Year')
plt.ylabel('Number of Releases')
plt.legend()
plt.grid(True)
plt.show()
```

```
Object `years` not found.
```



INSIGHTS :

It is clear that the number of movies is higher than number of TV shows in the recent years.

But number of TV shows has exponentially increased in the last 20 years.

RECOMMENDATIONS :

The increase in TV show production suggests a growing preference among viewers for television content in recent years. Therefore, it is advisable to prioritize the production of high-quality TV shows to align with this trend and meet audience demand effectively.

Although the number of movies remains higher than TV shows,maintaining a balanced approach where both movies and TV shows are produced concurrently is crucial.

Q.6 Understanding what content is available in different country?

```
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd

path = "/content/drive/MyDrive/netflix data.csv"

df = pd.read_csv(path)

movies_df = df[df['type'] == 'Movie']
```

```
Object `country` not found.
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

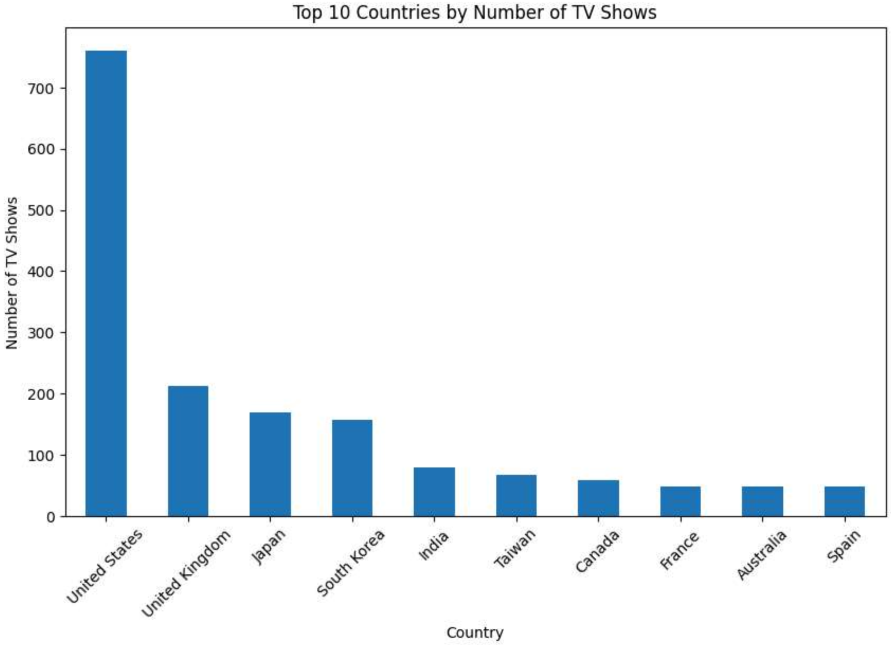
```
# Counting content by country

content_by_country = movies_df['country'].value_counts().head(10)

print(content_by_country)
```

```
country
United States     2058
India              893
United Kingdom     206
Canada             122
Spain               97
Egypt               92
Nigeria             86
Indonesia           77
Turkey              76
Japan               76
Name: count, dtype: int64
```

```
import matplotlib.pyplot as plt

movies_df = df[df['type'] == 'Movie']
content_by_country = movies_df['country'].value_counts().head(10)

plt.figure(figsize=(10, 6))
content_by_country.plot(kind='bar')
plt.title('Top 10 Countries by Number of Movies')
plt.xlabel('Country')
plt.ylabel('Number of Movies')
plt.xticks(rotation=45)
plt.show()
```

Top 10 Countries by Number of Movies
2000
1750
1500
1250
of Movies

```
import matplotlib.pyplot as plt

movies_df = df[df['type'] == 'TV Show']
content_by_country = movies_df['country'].value_counts().head(10)

plt.figure(figsize=(10, 6))
content_by_country.plot(kind='bar')
plt.title('Top 10 Countries by Number of TV Shows')
plt.xlabel('Country')
plt.ylabel('Number of TV Shows')
plt.xticks(rotation=45)
plt.show()
```


Top 10 Countries by Number of TV Shows
Number of TV Shows
United States
United Kingdom
Japan
South Korea
India
Taiwan
Canada
France
Australia
Spain
Country