Netflix - Data Exploration and Visualisation

1. What type of content is available in different countries?
    - The **groupby()** function helps to get the data of content available in different countries.

```
#1. What type of content is available in different countries?
set1 = df.groupby("country")["type"].value_counts().sort_values(ascending = False).head()
set2 = df.groupby("country")["listed_in"].value_counts().sort_values(ascending = False).head()
```

    - Column using "type"

		count
country	type	
United States	Movie	2058
India	Movie	893
United States	TV Show	760
United Kingdom	TV Show	213
	Movie	206

    - Column using "**listed in**"

		count
country	listed_in	
United States	Documentaries	249
	Stand-Up Comedy	209
India	Comedies, Dramas, International Movies	120
	Dramas, International Movies	118
	Dramas, Independent Movies, International Movies	108

Insights:

United sates people watch more "**movies**" than "**TV shows**" and in movies they mostly watch "**Documentaries**" and "**stand-up comedy**" type of genres.

2. How has the number of movies released per year changed over the last 20-30 years?

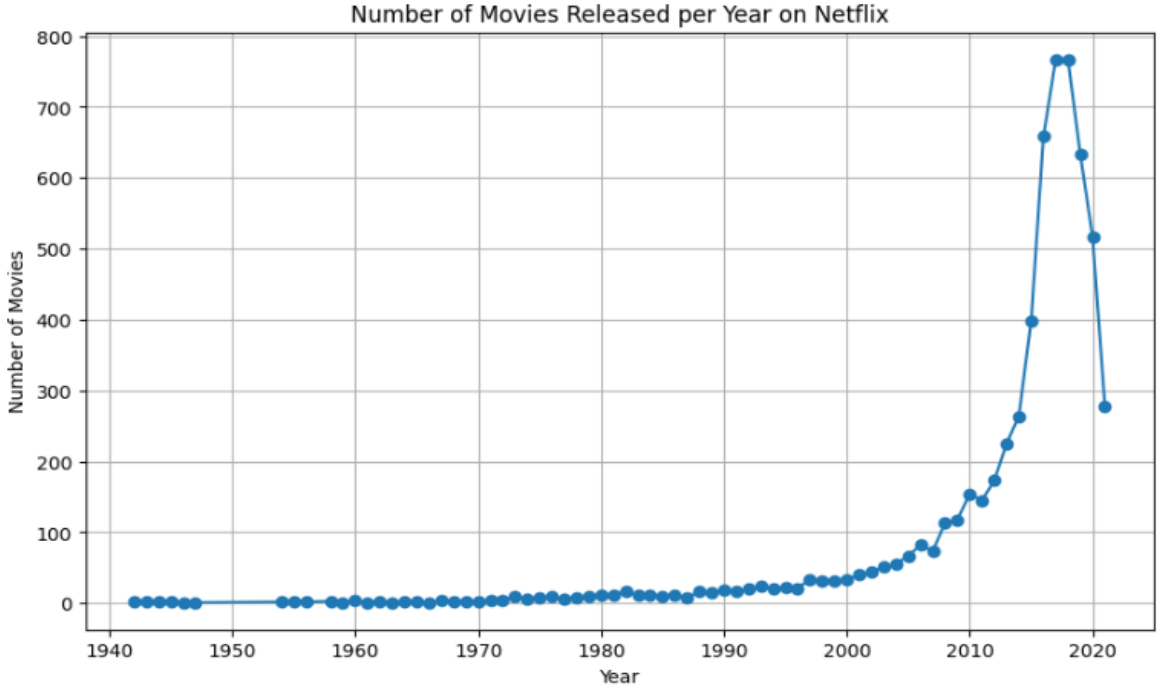
- We need to separate movies from the "**type**" column
- And using groupby function in "**Release_year**"

```
movies_per_year = df_movies.groupby('release_year').size()
movies_per_year
```

-

release_year	0
1942	2
1943	3
1944	3
1945	3
1946	1
...	...
2017	767
2018	767
2019	633
2020	517
2021	277

-



-

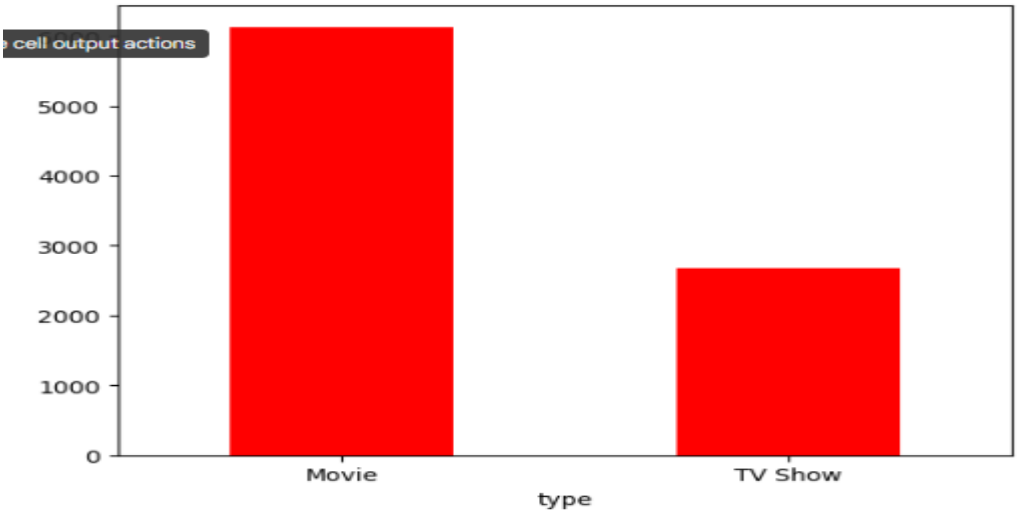
Insights:

There was steady growth in movie releases from the 1990s to the early 2010s, followed by a sharp rise, peaking between 2010 and 2018. After 2020, releases show signs of decline or stabilization as Netflix shifts focus from quantity to quality.

3. Comparison of TV shows vs. movies.
   - Using value_count() function, show that most people watch **movies** over **TV shows**, suggesting Netflix should invest in improving its TV show content

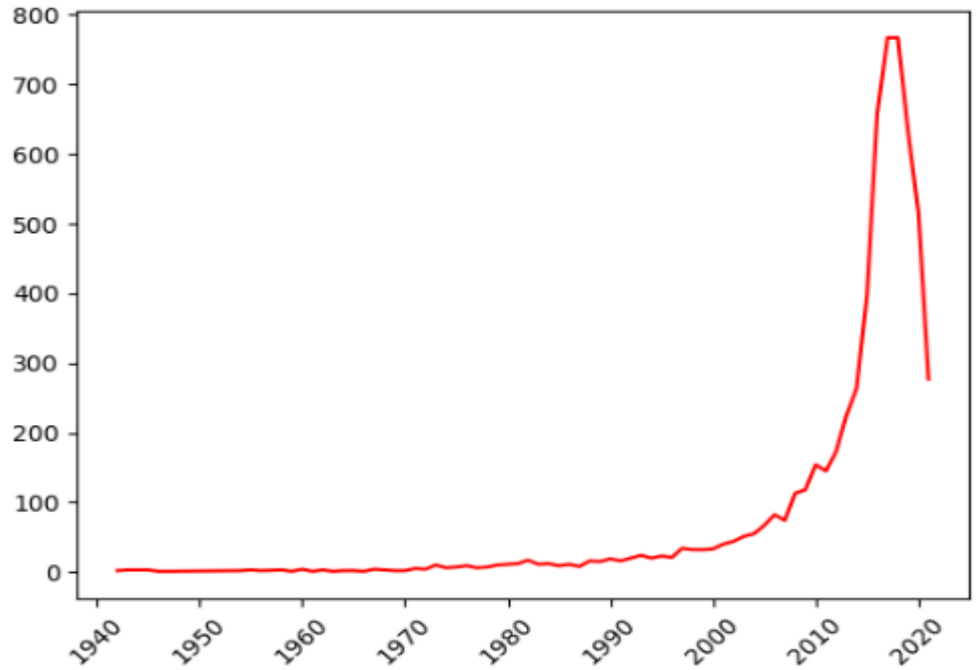
```
type_count = df["type"].value_counts()
type_count
```

type	count
Movie	6131
TV Show	2676

   - 

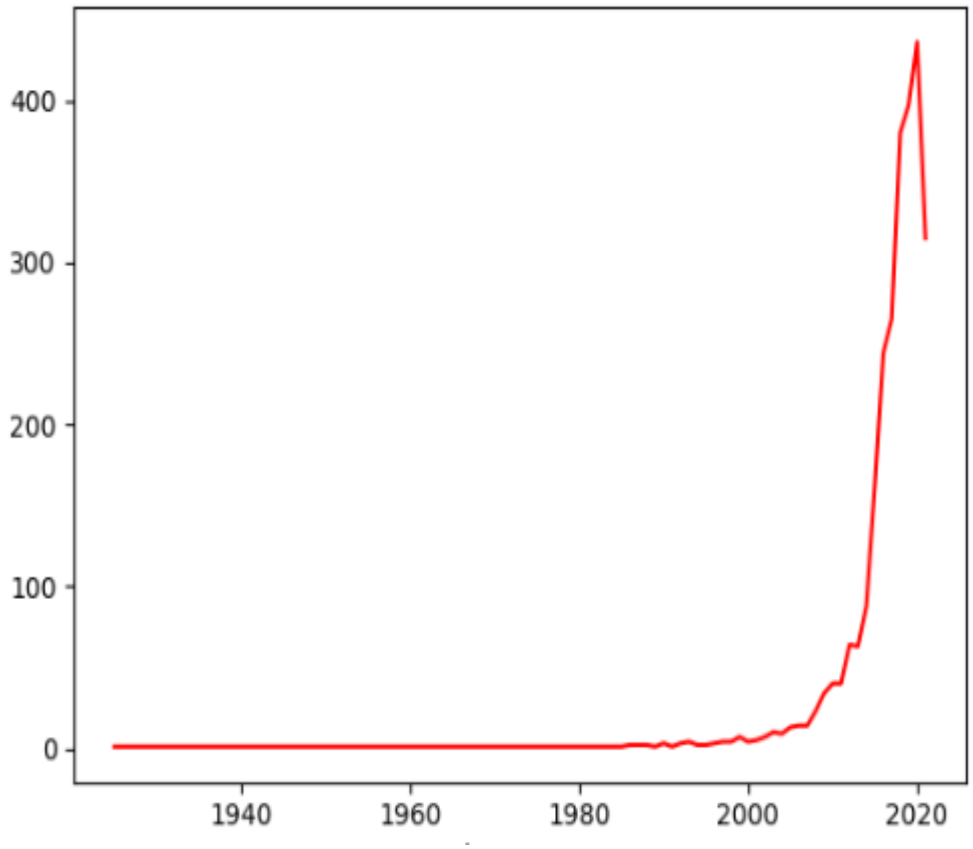
   - Analyse Movies trend per year by using groupby function

```
movies_per_year = df[df["type"]== "Movie"].groupby("release_year").size()
movies_per_year.plot(kind = "line", color = "red")
plt.xticks(rotation = 45)
plt.show()
```

   - 

- Analyse tv-shows trend per year by using groupby function

```
tv_shows_per_year = df[df["type"]== "TV Show"].groupby("release_year").size()
tv_shows_per_year.plot(kind = "line", color = "red")
plt.xticks(rotation = 360)
plt.show()
```



Insights:

Data shows that most people watch **movies** over **TV shows**, suggesting Netflix should invest in improving its TV show content.

4. What is the best time to launch a TV show
   - Separate the months from the “date added” column to get the months

```
df['month_added'] = df['date_added'].dt.month
df['month_added']
```

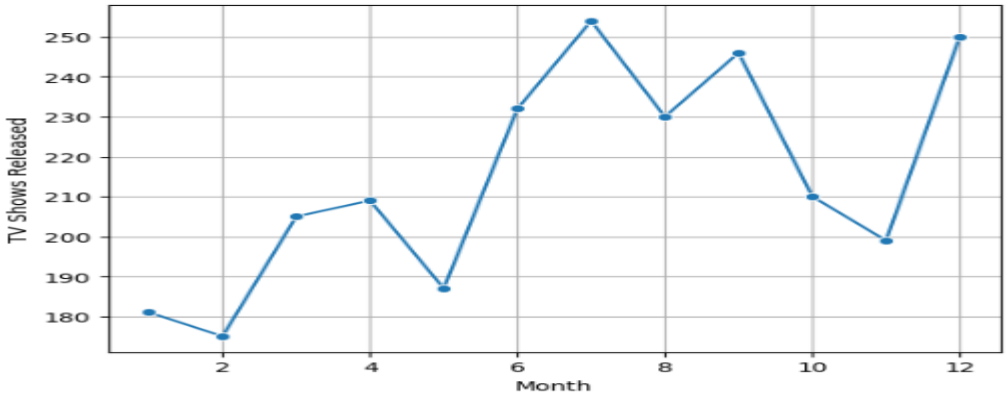
	month_added
0	9.0
1	9.0
2	9.0
3	9.0
4	9.0
...	...
8802	11.0
8803	7.0
8804	11.0
8805	1.0
8806	3.0

- And using masking filter “TV shows”
- Count TV Show releases by month

```
trend_month = df_tv_shows["month_added"].value_counts()
trend_month
```

month_added	count
7.0	254
12.0	250
9.0	246
6.0	232
8.0	230
10.0	210
4.0	209
3.0	205
11.0	199
5.0	187
1.0	181

```
sns.lineplot(data=trend_month, x='Month', y='TV Shows Released', marker='o')
plt.xticks(rotation = 360)
plt.grid(True)
plt.show()
```

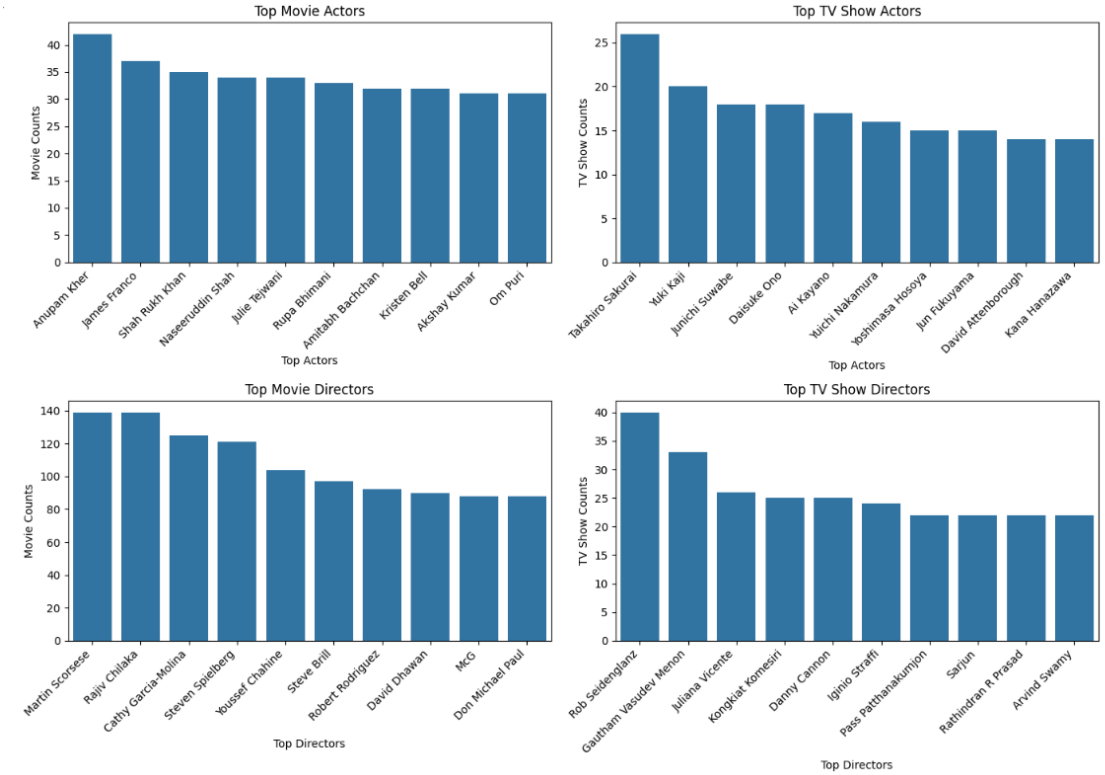


Insights:

TV show releases on Netflix peak in **December-January** and **summer months** like **June-July**, aligning with higher viewer engagement during holidays and vacations. These periods are ideal for launching new TV shows.

5. Analysis of actors/directors of different types of shows/movies.

- Analysing the data, we have missing and nested values in cast, listed in and director column.
- Use the fillna function to replace all null values with "Unknown", ensuring data consistency and preventing missing values from affecting the analysis.
- ```
  df[['cast', 'director']] = df[['cast', 'director']].fillna("unknown")
  ```
- To handle nested values, first use the `split` function to separate multiple entries within a column. Then, apply the `explode` function to transform these split values into individual rows, ensuring better data structure and analysis.
- ```
  df['cast']= df['cast'].apply(lambda x : x.split(', ')if isinstance(x, str)else x)
  df = df.explode('cast', ignore_index=True)
  ```
- Using the dataset, we identified the **top actors** in both **Movies and TV Shows**, as well as the **top directors** in each category, providing insights into their prominence across different formats.



- We identified the unique genres that actors and directors have worked in, highlighting their versatility and specialization across different categories.
- ```
  df.groupby("cast")['listed_in'].nunique().sort_values(ascending=False).head(10)
  ```

cast	listed_in
unknown	40
Ron Perlman	17
Kiernan Shipka	16
Gary Cole	16
Glenn Close	15
John Leguizamo	14
Anupam Kher	14
Jay Baruchel	14
Mae Whitman	14
Rajesh Sharma	14
  ```

Insights:

We analysed the most common actors and directors based on their appearances across titles. Some actors are more active in movies, while others dominate TV shows. Additionally, we found that while some actors and directors work across multiple genres, others specialize in specific ones.

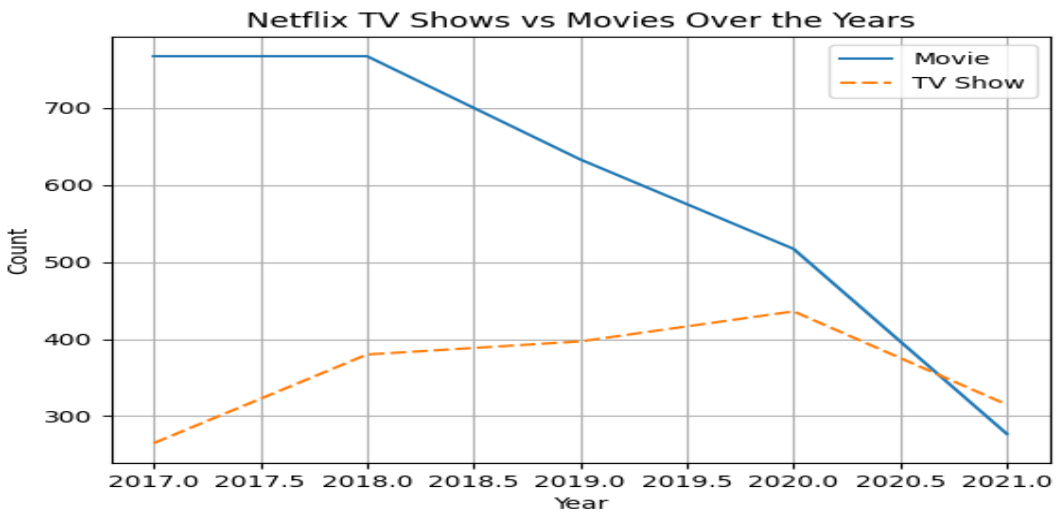
6. Does Netflix has more focus on TV Shows than movies in recent years

- Yes, Netflix has more focus on TV shows than movies
- Using crosstab function we get year wise count of movies and TV shows.

```
recent_yr = pd.crosstab(df['release_year'], df['type']).tail(5)
recent_yr
```

type	Movie	TV Show
release_year		
2017	767	265
2018	767	380
2019	633	397
2020	517	436
2021	277	315

- And using line plot we got a better visual that the Netflix focusing on TV shows more than movies



Insights:

Netflix has increasingly focused on TV shows, which are growing faster than movies, though movies still dominate in total numbers. This shift aligns with Netflix's strategy to boost engagement, as TV series retain viewers longer. In recent years, the gap between TV shows and movies has narrowed, highlighting a deliberate push toward serialized content while still expanding both categories.

7. Understanding what content is available in different countries

- Count the number of titles per country

```
country_counts = df['country'].value_counts().reset_index()
country_counts.columns = ['Country', 'Titles']
```

	Country	Titles
0	United States	59324
1	India	22814
2	United Kingdom	12945
3	unknown	11897
4	Japan	8679
5	France	8252
6	Canada	7915
7	Spain	5315
8	South Korea	5043
9	Germany	4383

- Count the number of Movies vs TV Shows per country

```
content_type = df.groupby(['country', 'type']).size().unstack(fill_value=0)
```

type	Movie	TV Show
country		
	24	8
Afghanistan	2	0
Albania	8	0
Algeria	77	0
Angola	32	0
Argentina	1325	455
Armenia	2	0
Australia	1551	1065
Austria	189	21
Azerbaijan	0	33

- Count genres per country

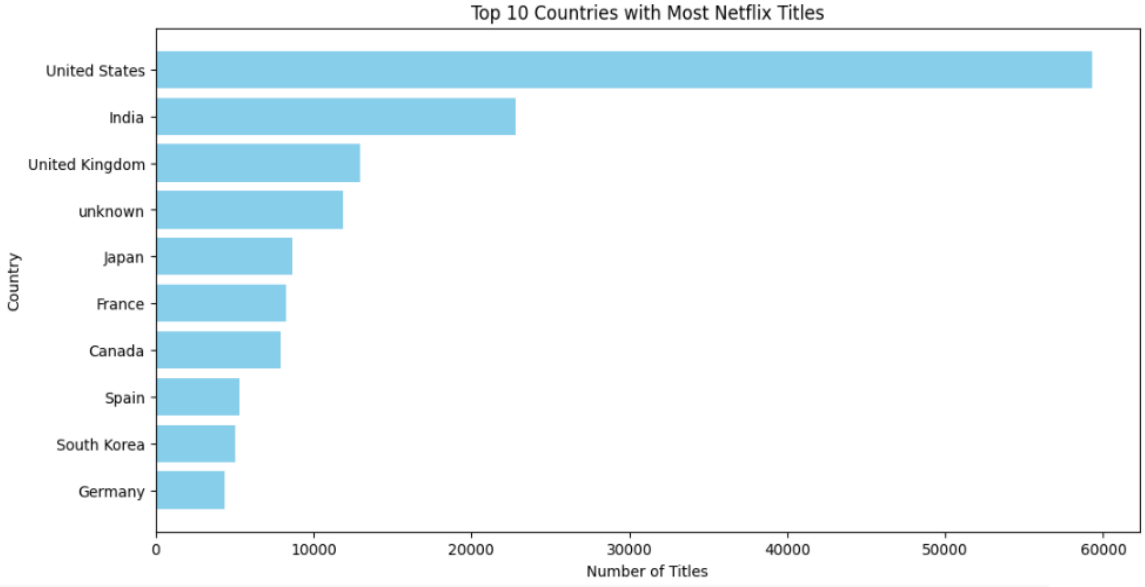
```
genre_counts = df_exploded.groupby(['country', 'listed_in']).size().reset_index(name='count')

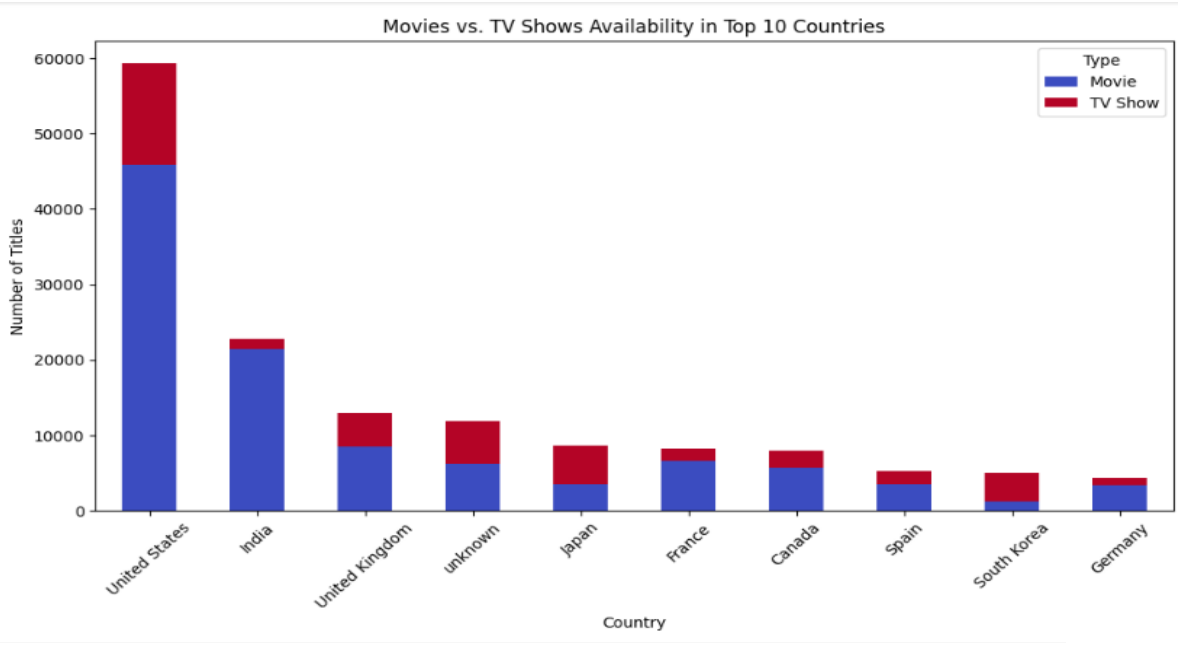
# Sort and display top genres per country
genre_sorted = genre_counts.sort_values(by=['country', 'count'], ascending=[True, False])
genre_sorted.head(20)
```

Afghanistan	International Movies	1
Albania	Dramas	4
Albania	International Movies	4
Algeria	Dramas	29
Algeria	International Movies	29
Algeria	Classic Movies	11
Algeria	Independent Movies	8
Angola	Action & Adventure	16
Angola	International Movies	16
Argentina	International Movies	455
Argentina	Dramas	306
Argentina	Spanish-Language TV Shows	144
Argentina	Comedies	139
Argentina	International TV Shows	134

```python
# Count how many times each title appears
title_counts = df['title'].value_counts()

# Get titles that appear only once (exclusive content)
exclusive_titles = df[df['title'].isin(title_counts[title_counts == 1].index)]

# Count exclusive titles per country
exclusive_counts = exclusive_titles.groupby('country').size().reset_index(name='Exclusive Count')

# Show top 10 countries with most exclusive content
exclusive_counts.sort_values(by='Exclusive Count', ascending=False).head(10)
```

	country	Exclusive Count
22	United States	466
24	unknown	90
21	United Kingdom	54
3	Canada	11
6	France	10
2	Brazil	7
9	India	7
13	Mexico	6
1	Australia	4
7	Germany	4





Insights:

Netflix's library varies by country due to licensing and regional preferences. The **largest collections** are found in the **US, India, and the UK**, offering diverse content, including Netflix Originals and local productions.

Movies vs. TV shows differ by region—**India and Brazil favourite movies**, while **Japan and South Korea lean towards TV shows**, particularly Anime and K-Dramas. In terms of **genres**, the **US prefers Drama and Comedy**, **India thrives on Action and Thriller**, and **Japan is dominated by Anime**.

Exclusive content is common in **India, South Korea, and Japan**, featuring region-specific productions unavailable elsewhere. Netflix's strategy ensures a **localized experience**, catering to each market's unique tastes.