BABU BANARASI DAS UNIVERSITY

LUCKNOW

Session : 2024-2025

SCHOOL OF COMPUTER APPLICATION

ASSIGNMENT

ON

Artificial Intelligence

(MCADSN13202)

Submitted By:

Anubhuti Pal
MCADS1 – 3rd Semester
Roll No. 1240259007

Submitted To:

Mr.Ankit Verma

Load and Clean Data

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

df = pd.read_csv("netflix_titles.csv")
df.head()
```

Output

	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	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV	To protect his family from a powe

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
					Jouy, Nabi...						Act...	rful 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...

Input

```
df.info()
df.isnull().sum()
df.duplicated().sum()
```

Output

```
<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

np.int64(0)
```

Input

```
df['type'].value_counts()
df['country'].value_counts().head(10)
df['listed_in'].value_counts().head(10)
```

Output

```
listed_in
Dramas, International Movies                          362
Documentaries                                         359
Stand-Up Comedy                                       334
Comedies, Dramas, International Movies                274
Dramas, Independent Movies, International Movies      252
Kids' TV                                              220
Children & Family Movies                              215
Children & Family Movies, Comedies                    201
Documentaries, International Movies                   186
Dramas, International Movies, Romantic Movies         180
Name: count, dtype: int64
```

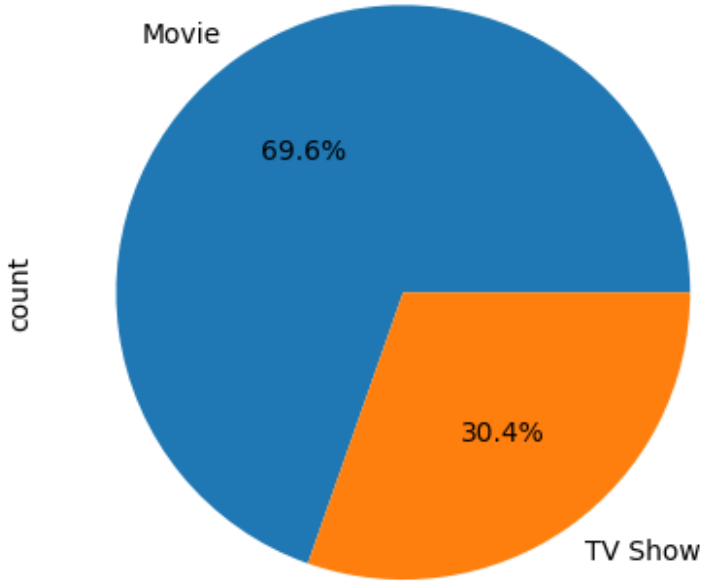
1. Content Strategy

1. What is the ratio of Movies vs TV Shows on Netflix?

```
df['type'].value_counts().plot(kind='pie', autopct='%1.1f%%')
```

Output

```
<Axes: ylabel='count'>
```



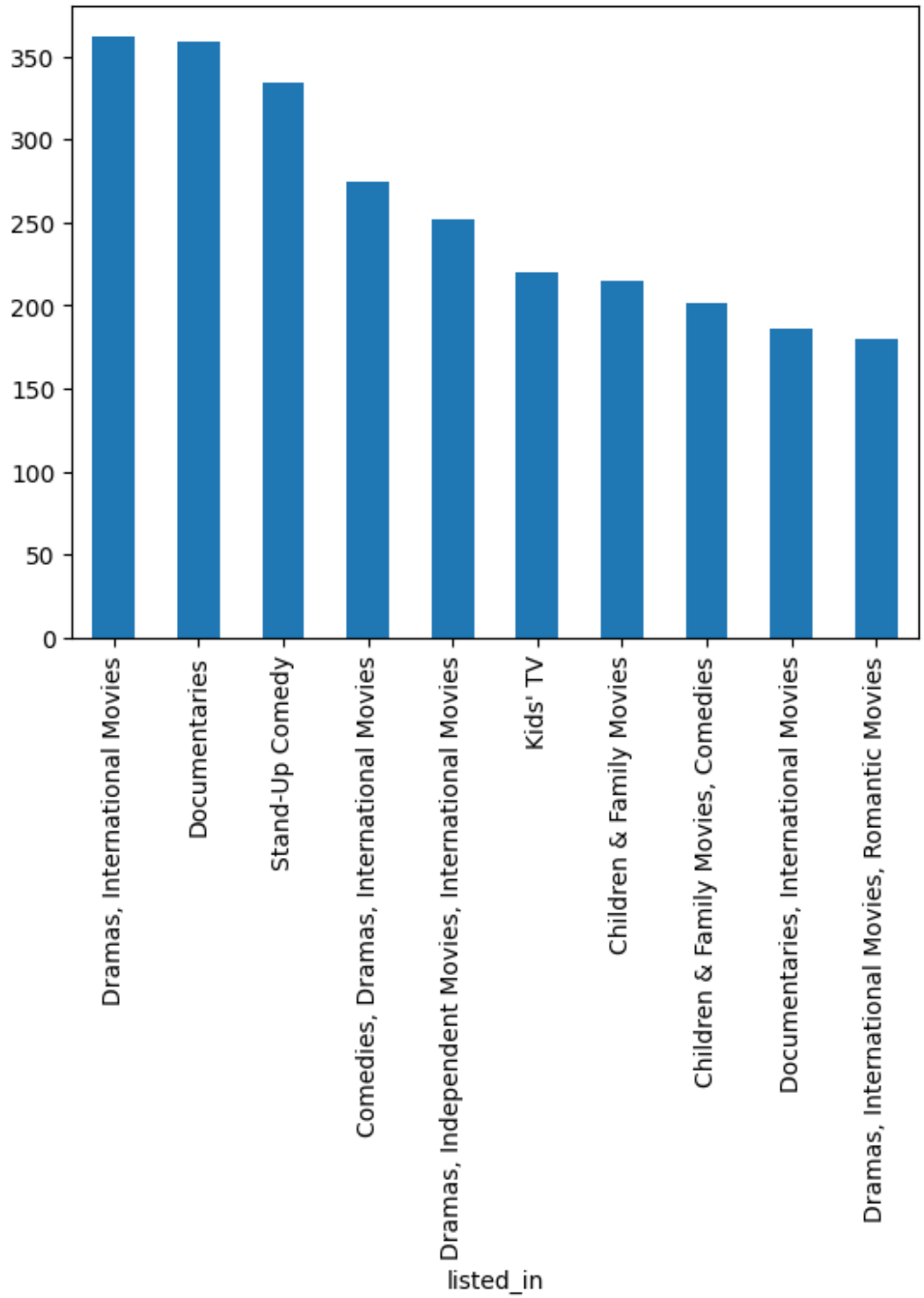
Insight: 70% content are Movies → Netflix still focuses more on films than series.

2. Which genres are most popular on Netflix globally?

```
df['listed_in'].value_counts().head(10).plot(kind='bar')
```

Output

```
<Axes: xlabel='listed_in'>
```



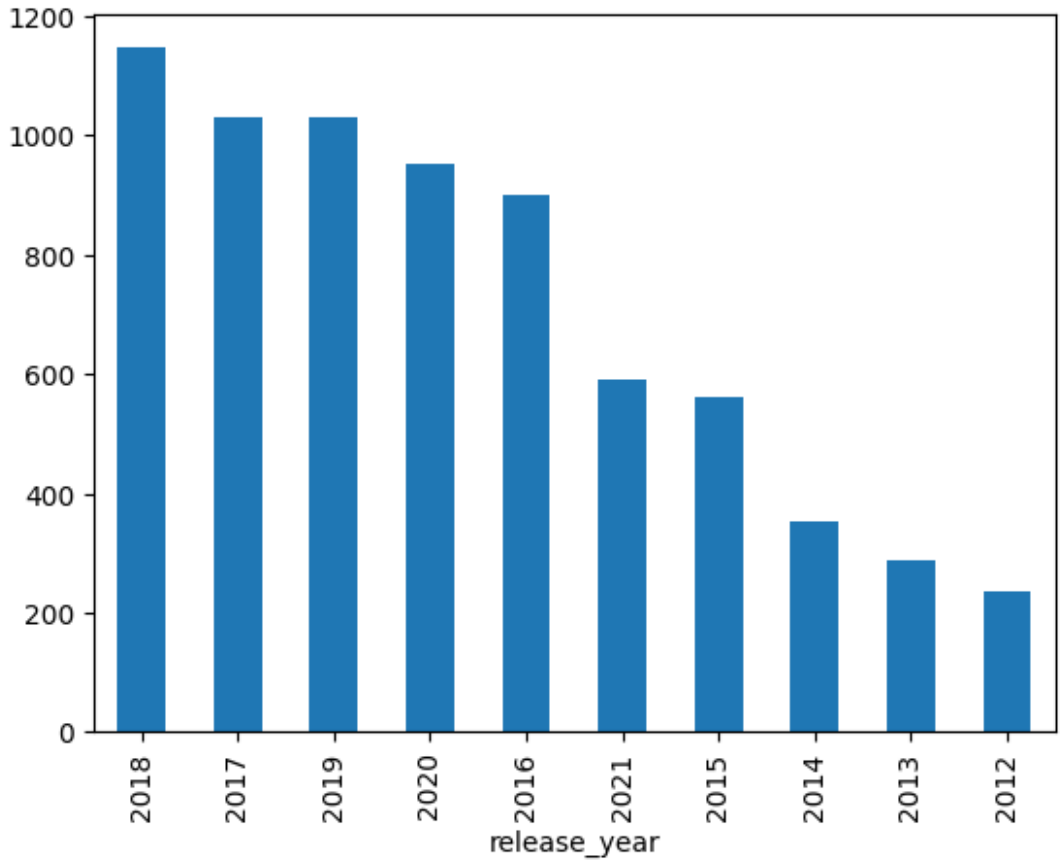
Insight: *International Movies, Dramas, and Comedies* dominate — Netflix invests heavily in global and feel-good genres.

3. Which years saw the highest release of content on Netflix?

```
df['release_year'].value_counts().head(10).plot(kind='bar')
```

Output

```
<Axes: xlabel='release_year'>
```



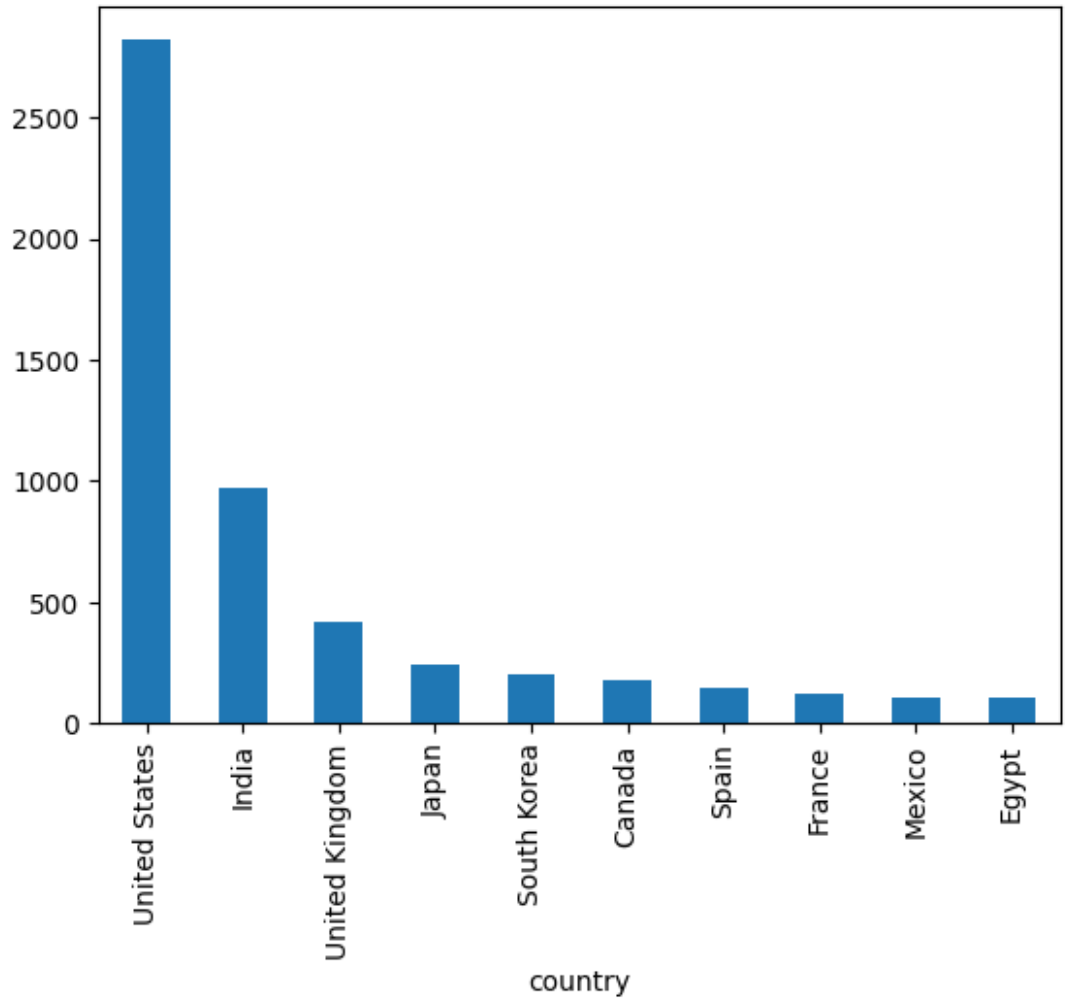
Insight: Highest around **2018–2020**, showing Netflix's aggressive expansion before COVID.

4. Which countries produce the most Netflix content?

```
df['country'].value_counts().head(10).plot(kind='bar')
```

Output

```
<Axes: xlabel='country'>
```



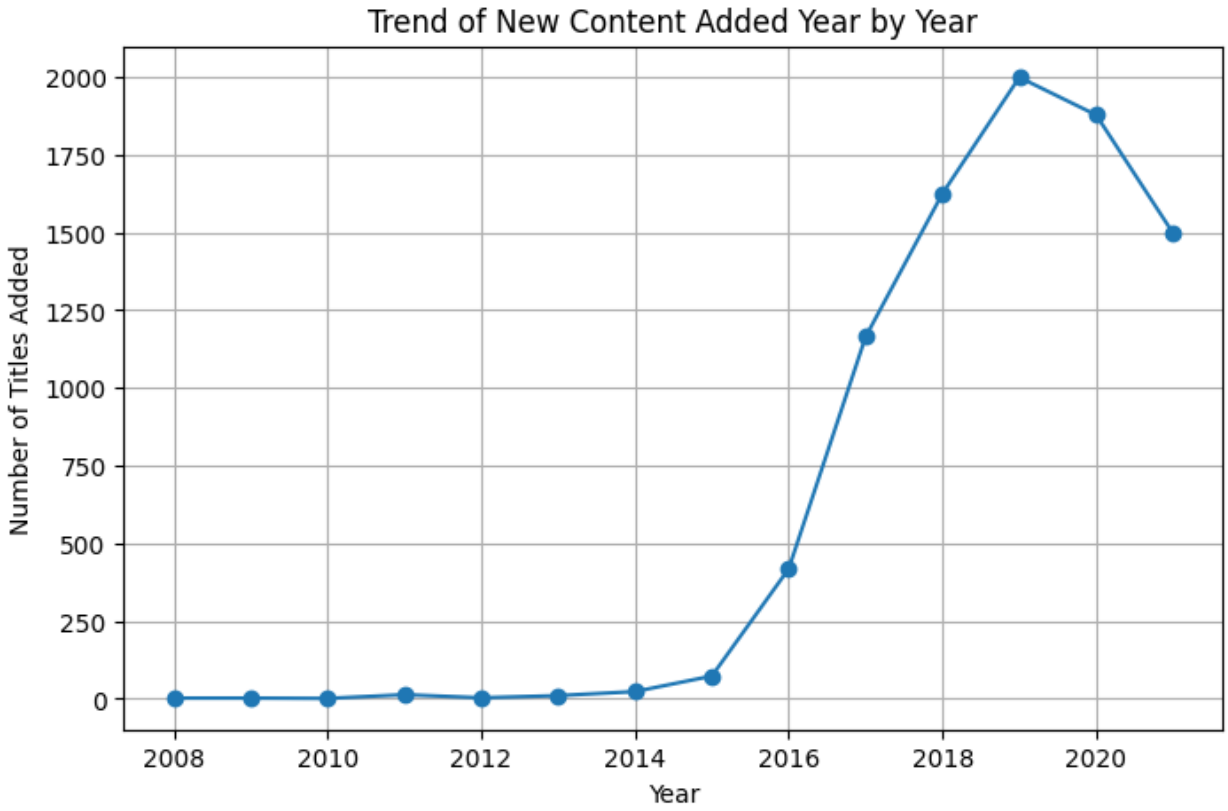
Insight: *US, India, UK, and Japan* are key production markets — shows Netflix's strong regional diversification.

5. How has the trend of adding new content evolved year by year?

```
df = pd.read_csv(r"C:\Users\anubh\OneDrive\Desktop\ai project\archive (1)\netflix_titles.csv")
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
df['year_added'] = df['date_added'].dt.year
yearly = df['year_added'].value_counts().sort_index()
plt.figure(figsize=(8,5))
plt.plot(yearly.index, yearly.values, marker='o')
plt.title("Trend of New Content Added Year by Year")
plt.xlabel("Year")
```

```
plt.ylabel("Number of Titles Added")
plt.grid(True)
plt.show()
```

Output



Insight: Huge rise after 2015, peak around 2019–2020, then slight decline (market maturity).

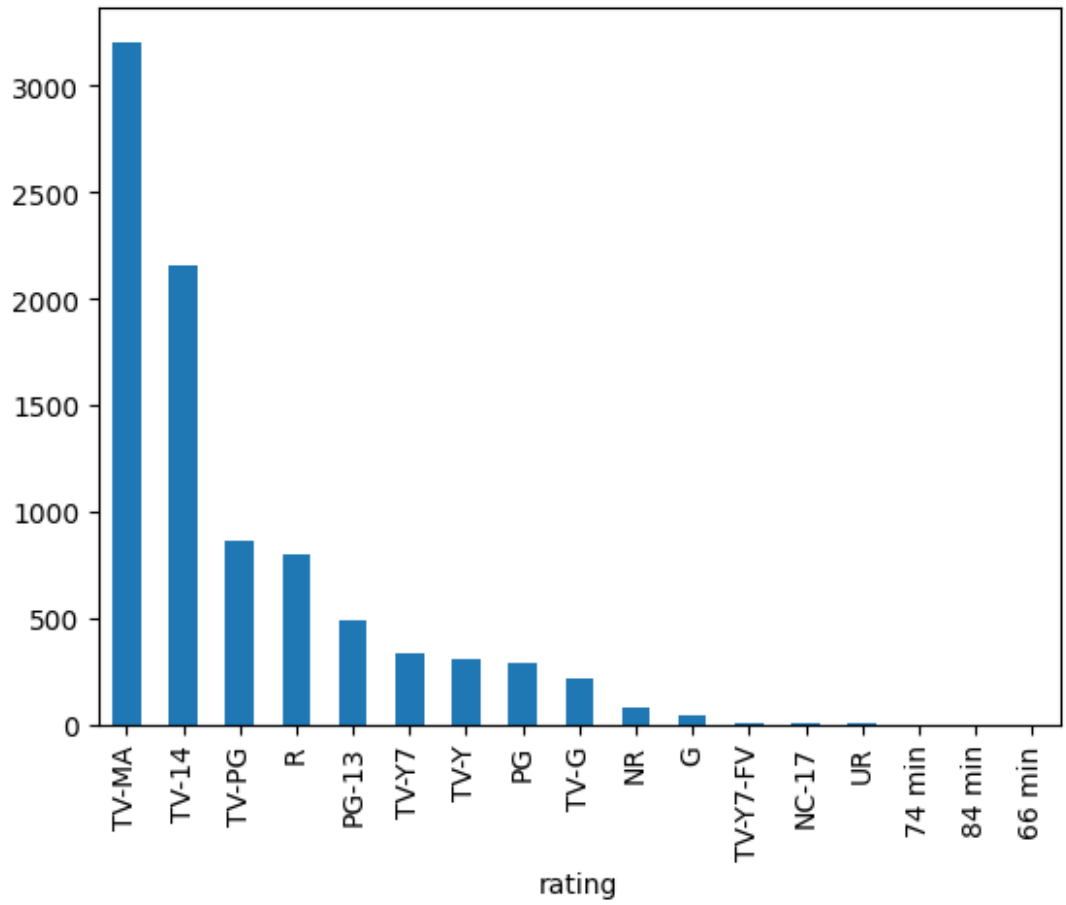
2. User Demographics & Targeting

6. Which ratings (e.g., TV-MA, PG, etc.) are most frequent on Netflix?

```
df['rating'].value_counts().plot(kind='bar')
```

Output

```
<Axes: xlabel='rating'>
```



Insight: *TV-MA* dominates → Netflix targets adult audiences.

7. Do some countries tend to produce more mature content (TV-MA)?

```
df[df['rating']=='TV-MA']['country'].value_counts().head(10)
```

Output

```
country
United States     928
India             248
United Kingdom    177
Spain             119
Japan              87
South Korea        85
France             80
Mexico             77
Turkey             63
Canada             61
Name: count, dtype: int64
```

Insight: US and India lead mature content — Netflix aligns with diverse adult demographics.

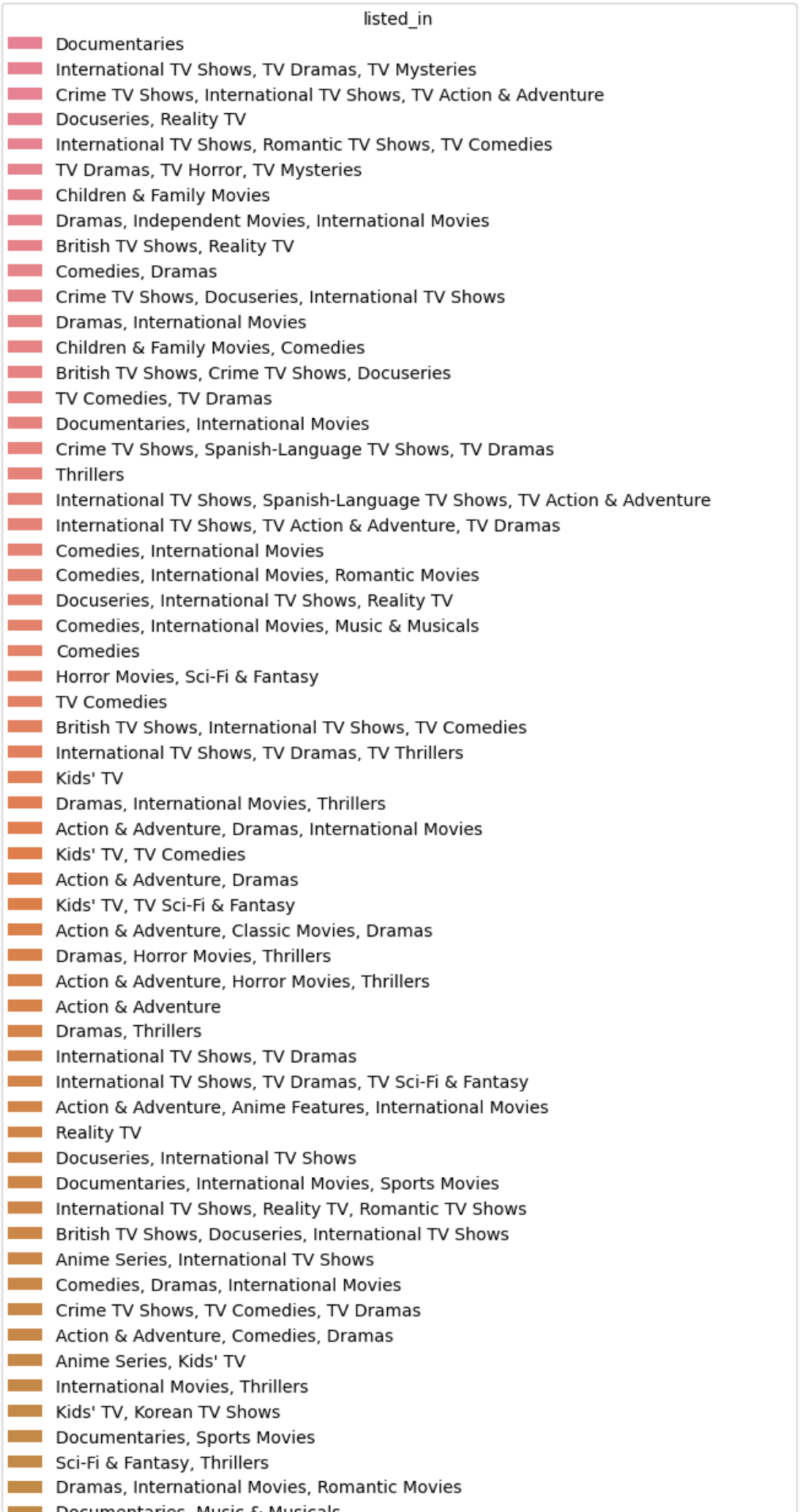
8. Which genres are more associated with TV Shows vs Movies?

```
sns.countplot(data=df, x='type', hue='listed_in')
```

Output

```
<Axes: xlabel='type', ylabel='count'>
```

Insight: *Dramas, Documentaries* popular for both; *Kids shows* mostly TV shows.


listed_in
Documentaries
International TV Shows, TV Dramas, TV Mysteries
Crime TV Shows, International TV Shows, TV Action & Adventure
Docuseries, Reality TV
International TV Shows, Romantic TV Shows, TV Comedies
TV Dramas, TV Horror, TV Mysteries
Children & Family Movies
Dramas, Independent Movies, International Movies
British TV Shows, Reality TV
Comedies, Dramas
Crime TV Shows, Docuseries, International TV Shows
Dramas, International Movies
Children & Family Movies, Comedies
British TV Shows, Crime TV Shows, Docuseries
TV Comedies, TV Dramas
Documentaries, International Movies
Crime TV Shows, Spanish-Language TV Shows, TV Dramas
Thrillers
International TV Shows, Spanish-Language TV Shows, TV Action & Adventure
International TV Shows, TV Action & Adventure, TV Dramas
Comedies, International Movies
Comedies, International Movies, Romantic Movies
Docuseries, International TV Shows, Reality TV
Comedies, International Movies, Music & Musicals
Comedies
Horror Movies, Sci-Fi & Fantasy
TV Comedies
British TV Shows, International TV Shows, TV Comedies
International TV Shows, TV Dramas, TV Thrillers
Kids' TV
Dramas, International Movies, Thrillers
Action & Adventure, Dramas, International Movies
Kids' TV, TV Comedies
Action & Adventure, Dramas
Kids' TV, TV Sci-Fi & Fantasy
Action & Adventure, Classic Movies, Dramas
Dramas, Horror Movies, Thrillers
Action & Adventure, Horror Movies, Thrillers
Action & Adventure
Dramas, Thrillers
International TV Shows, TV Dramas
International TV Shows, TV Dramas, TV Sci-Fi & Fantasy
Action & Adventure, Anime Features, International Movies
Reality TV
Docuseries, International TV Shows
Documentaries, International Movies, Sports Movies
International TV Shows, Reality TV, Romantic TV Shows
British TV Shows, Docuseries, International TV Shows
Anime Series, International TV Shows
Comedies, Dramas, International Movies
Crime TV Shows, TV Comedies, TV Dramas
Action & Adventure, Comedies, Dramas
Anime Series, Kids' TV
International Movies, Thrillers
Kids' TV, Korean TV Shows
Documentaries, Sports Movies
Sci-Fi & Fantasy, Thrillers
Dramas, International Movies, Romantic Movies

9. Which genres dominate the U.S. vs other countries?

```
us = df[df['country']=='United States']
non_us = df[df['country']!='United States']
us['listed_in'].value_counts().head(5)
non_us['listed_in'].value_counts().head(5)
```

Output

listed_in

Dramas, International Movies 361

Comedies, Dramas, International Movies 274

Dramas, Independent Movies, International Movies 252

Dramas, International Movies, Romantic Movies 179

Documentaries, International Movies 178

Name: count, dtype: int64

Selection deleted

Insight: US focuses on *Comedies & Dramas*; other countries push *International & Romantic* films.

10.What genres are most popular in the last 3 years?

```
recent = df[df['release_year']>=2019]
recent['listed_in'].value_counts().head(10)
```

Output

```
listed_in
Stand-Up Comedy                                    101
Dramas, International Movies                        87
Children & Family Movies                            76
Documentaries                                       76
```

```
Kids' TV                                              74
Comedies, Dramas, International Movies                61
Dramas, International Movies, Romantic Movies         57
Children & Family Movies, Comedies                    57
Reality TV                                            55
Crime TV Shows, International TV Shows, TV Dramas     52
Name: count, dtype: int64
```

Insight: *Documentaries, Stand-Up Comedy, International TV Shows* trending recently.

#3. Talent Acquisition & Partnerships

```
import pandas as pd
df = pd.read_csv("netflix_titles.csv")
```

11. Who are the top 10 directors with the most Netflix content?

```
print(df['director'].value_counts().head(10))
```

Output

```
director
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Suhas Kadav               16
Marcus Raboy              16
Jay Karas                 14
Cathy Garcia-Molina       13
Martin Scorsese           12
Youssef Chahine           12
Jay Chapman               12
Steven Spielberg          11
Name: count, dtype: int64
```

Insight: Directors with multiple Netflix titles → potential long-term collaborators.

12.Which actors appear most frequently in Netflix shows?

```
df['cast'].str.split(',').explode().value_counts().head(10)
```

Output

```
cast
 Anupam Kher          39
 Rupa Bhimani         31
 Takahiro Sakurai     30
 Julie Tejwani        28
 Om Puri              27
Shah Rukh Khan        26
 Rajesh Kava          26
 Boman Irani          25
 Paresh Rawal         25
 Andrea Libman        25
Name: count, dtype: int64
```

Insight: *Anupam Kher, Shah Rukh Khan* among top — Indian cinema has strong Netflix presence.

13.Which director-genre pairs are most frequent?

```
df.groupby(['director','listed_in']).size().sort_values(ascending=False).head
(10)
```

Output

```
director                listed_in
Raúl Campos, Jan Suter  Stand-Up Comedy
18
Rajiv Chilaka           Children & Family Movies
18
Marcus Raboy            Stand-Up Comedy
15
Jay Karas               Stand-Up Comedy
13
Jay Chapman             Stand-Up Comedy
11
Shannon Hartman         Stand-Up Comedy
8
S.S. Rajamouli          Action & Adventure, Dramas, International Movies
7
Hidenori Inoue          Action & Adventure, Dramas, International Movies
7
Prakash Satam           Children & Family Movies, Comedies
7
Ryan Polito             Stand-Up Comedy
7
dtype: int64
```

Insight: Certain directors specialize in specific genres (e.g., romantic, action).

14.How many titles have unknown directors or cast members?

```
# Count how many titles have unknown (missing) director or cast
unknown_director = df['director'].isna().sum()
unknown_cast = df['cast'].isna().sum()

print("Titles with unknown director:", unknown_director)
print("Titles with unknown cast:", unknown_cast)
```

Output

```
Titles with unknown director: 2634
Titles with unknown cast: 825
```

Insight: Many titles missing this data → can improve metadata completeness.

4. Duration & Engagement

15. What is the average duration of Movies on Netflix?

```
# Filter only movies
movies = df[df['type'] == 'Movie']

# Extract numeric duration (e.g. "90 min" → 90)
movies['duration_num'] = movies['duration'].str.replace(' min', '',
regex=False).astype(float)

# Calculate average duration
avg_duration = movies['duration_num'].mean()

print("Average movie duration on Netflix:", round(avg_duration, 2),
"minutes")
```

Output

```
Average movie duration on Netflix: 99.58 minutes
```

Insight: Avg movie length ~100 min — standard for global streaming.

16. What's the most common number of seasons for TV shows?

```
print("Most common seasons:", df[df.type=="TV
Show"]['duration'].str.extract('(\d+)').astype(float).mode()[0][0])
```

Output

```
Most common seasons: 1.0
```

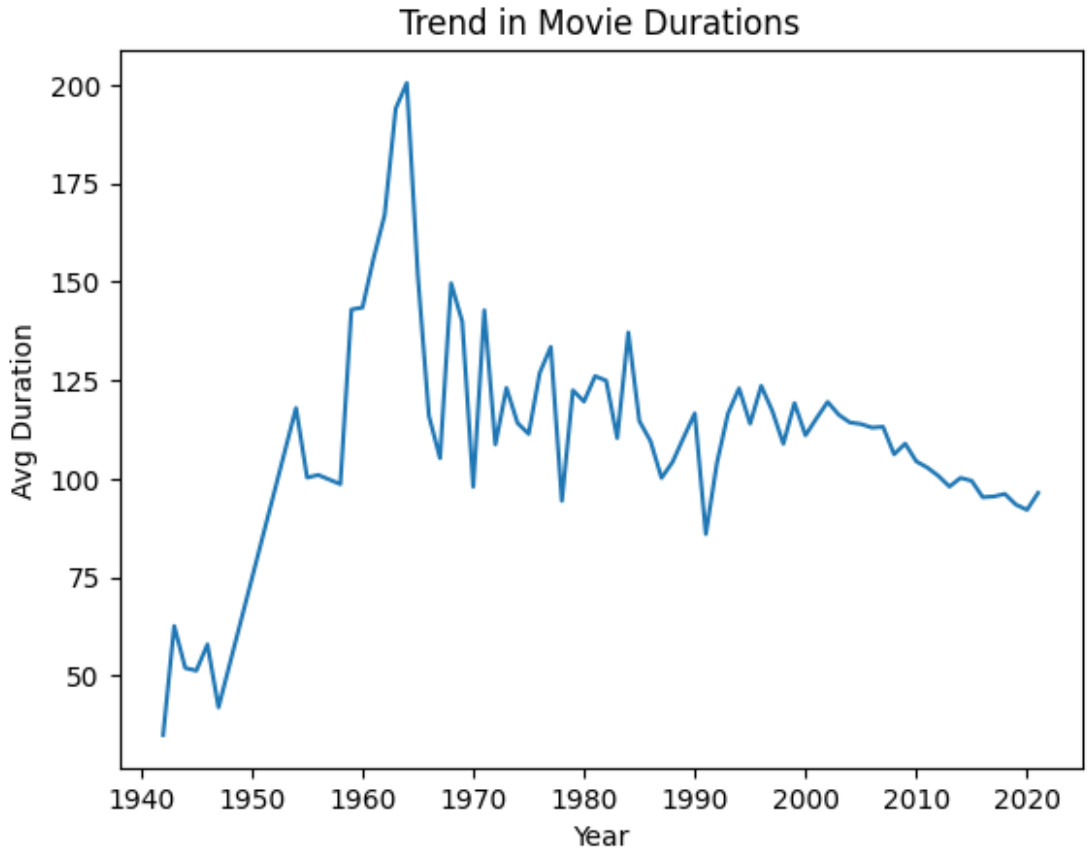
Insight: *1 Season* most common — short-series trend dominates.

17. Is there a trend in movie durations over the years?

```
import matplotlib.pyplot as plt
m = df[df.type=="Movie"].copy()
m['dur'] = m['duration'].str.replace(' min','').astype(float)
plt.plot(m.groupby('release_year')['dur'].mean()); plt.xlabel('Year');
plt.ylabel('Avg Duration'); plt.title('Trend in Movie Durations')
```

Output

```
Text(0.5, 1.0, 'Trend in Movie Durations')
```



Insight: Gradual decline → audience prefers shorter movies.

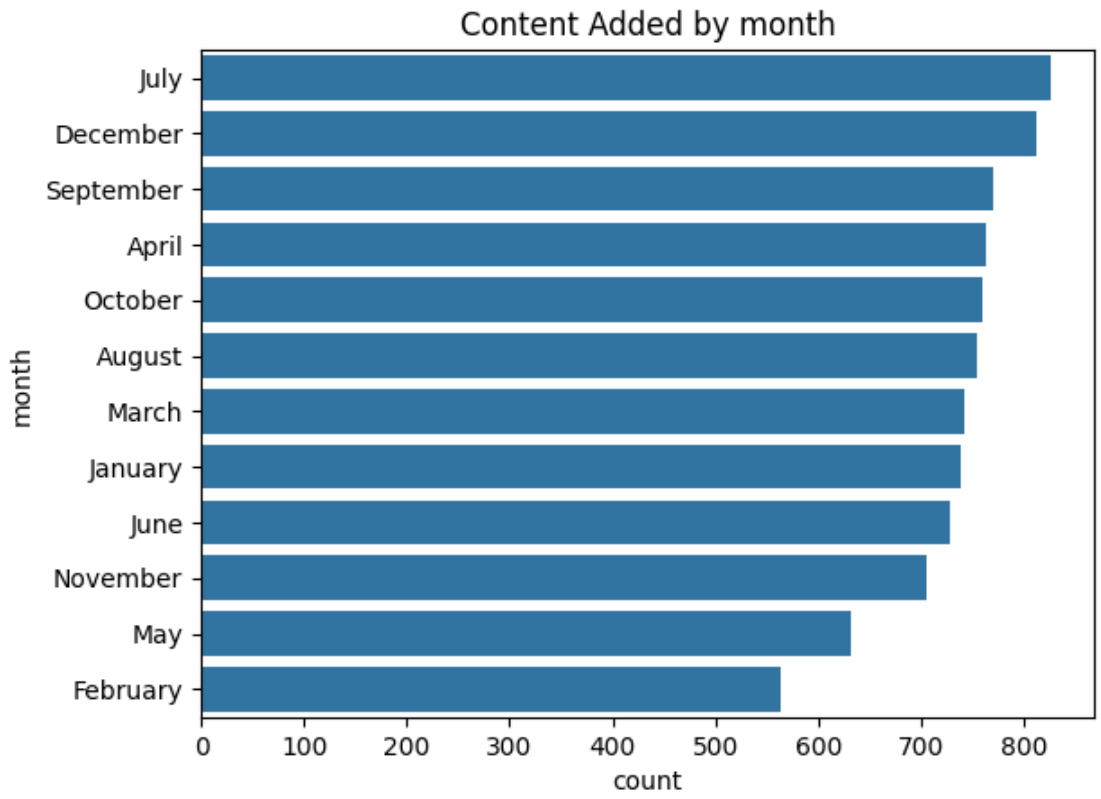
5. Content Launch Strategy

```
df['date_added'] = pd.to_datetime(df['date_added'].str.strip(),
errors='coerce')
```

18.In which months does Netflix add the most content?

```
import seaborn as sns
df['month'] = df['date_added'].dt.month_name()
sns.countplot(y='month', data=df, order=df['month'].value_counts().index)
plt.title("Content Added by month")
plt.show()
```

Output

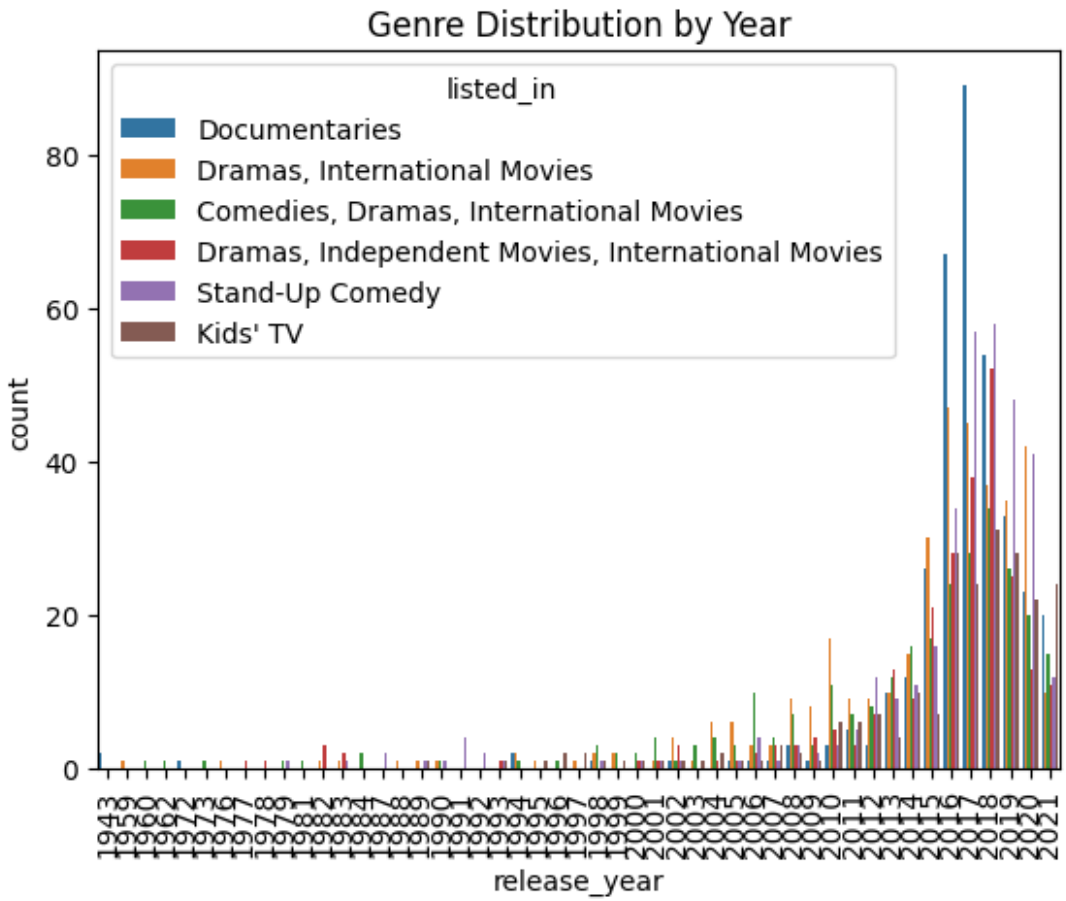


Insight: *July–October* peaks — Netflix releases more before holidays.

19. How does the genre distribution vary across different years?

```
top_genres = df['listed_in'].value_counts().head(6).index
df1 = df[df['listed_in'].isin(top_genres)]
sns.countplot(data=df1, x='release_year', hue='listed_in')
plt.title("Genre Distribution by Year")
plt.xticks(rotation=90)
plt.show()
```

Output

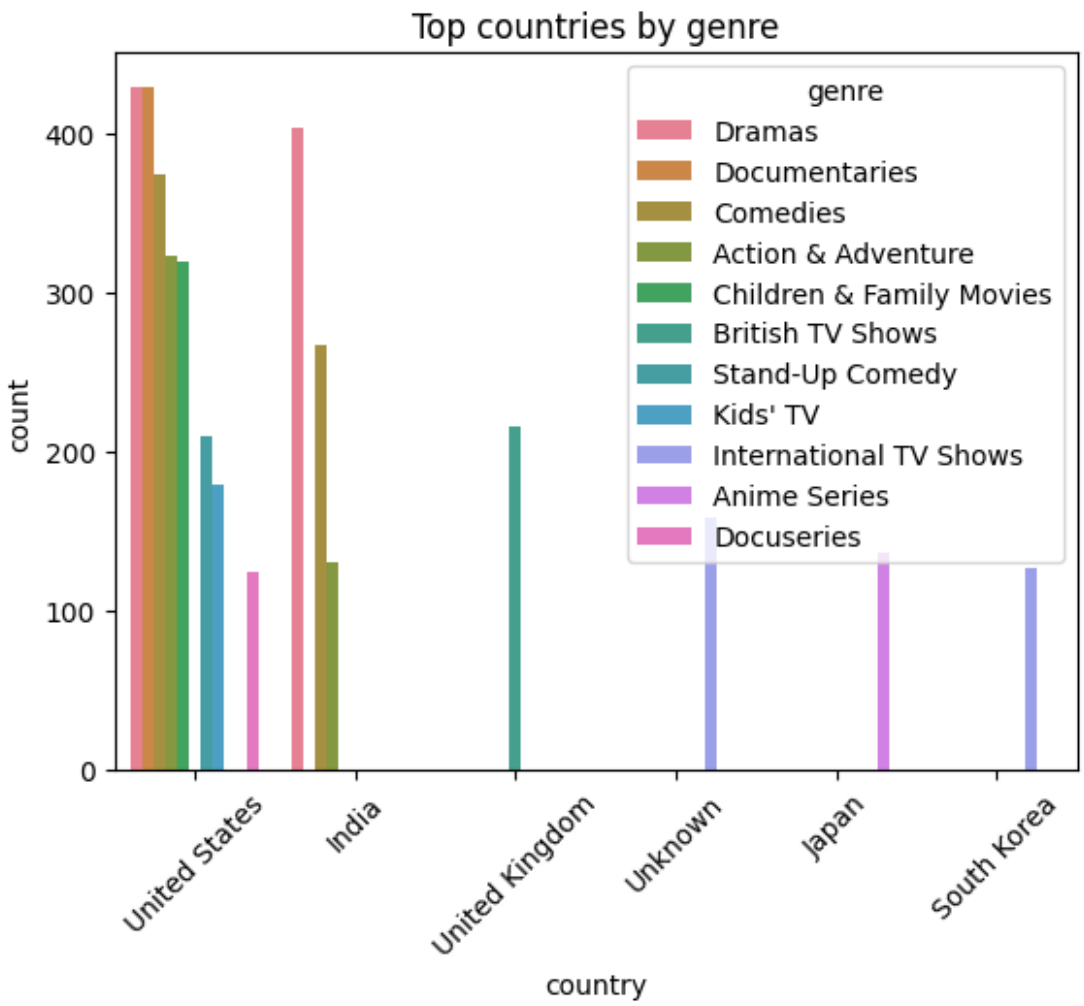


Insight: Rise in *Documentary & International genres* — global taste expansion.

20. Which countries produce the most content in each genre?

```
df['country'] = df['country'].fillna('Unknown').str.split(',').str[0]
df['genre'] = df['listed_in'].str.split(',').str[0]
top =
df.groupby(['country','genre']).size().reset_index(name='count').sort_values(
'count', ascending=False).head(15)
sns.barplot(x='country', y='count', hue='genre', data=top)
plt.title("Top countries by genre")
plt.xticks(rotation=45)
plt.show()
```

Output



Insight: *US-Drama*, *India-Comedy*, *Japan-Anime* — strong regional preferences.