```
#Importing all the required libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv("/content/netflix.csv")
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descript
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her fa nears the of his film
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	A cros paths party, a C Tow
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 pro his fa fro powe drug l
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Fe flirtati and toilet go d an
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV	In a cit coacl cen know

```
df.shape
```

```
(8807, 12)
```

```
for i in df.columns:
  print(str(i)," : ",str(df[i].nunique()))
```

```
show_id  :  8807
type  :  2
title  :  8807
director  :  4528
cast  :  7692
country  :  748
date_added  :  1767
release_year  :  74
rating  :  17
duration  :  220
listed_in  :  514
description  :  8775
```

From the above we can observe all show_id's and titles are unique so need to check if any duplicated rows are there

```
df.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
```

If we observe all columns expect release_year are of object datatype

```
df.isnull().sum()
```

```
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
```

If we observe columns like director,cast,country has more null values

So we cant drop rows containing nulls in those columns

And also it is not appropriate to impute those columns with some metric(like most frequent category) so let us fill them with empty string

```
df['director']=df['director'].fillna("")
df['cast']=df['cast'].fillna("")
df['country']=df['country'].fillna("")
```

```
df.isnull().sum()
```

```
show_id          0
type             0
title            0
director         0
cast             0
country          0
date_added      10
release_year     0
rating           4
duration         3
listed_in        0
description      0
dtype: int64
```

Now we have very less number of rows with null values so we will drop them

```
df=df.dropna()
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descript
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her fa nears the of his film
1	s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	A cros paths party, a C Tow
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas,		September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International	To pro his fa fro

```
df.shape
```

```
(8790, 12)
```

```
date_format = '%B %d, %Y'

# Check if the date strings match the format
df['is_valid_date'] = pd.to_datetime(df['date_added'], format=date_format, errors='coerce').notna()
```

```
df['is_valid_date'].all()
```

```
False
```

So if we observe all date_added columns in the given format

```
df[~(df['is_valid_date'])]
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descr
6079	s6080	TV Show	Abnormal Summit	Jung-ah Im, Seung-uk Jo	Hyun-moo Jun, Si-kyung Sung, Se-yoon Yoo	South Korea	August 4, 2017	2017	TV-PG	2 Seasons	International TV Shows, Korean TV Shows, Stand...	Led b of c multina
6177	s6178	TV Show	忍者ハットリくん			Japan	December 23, 2018	2012	TV-Y7	2 Seasons	Anime Series, Kids' TV	Haili moun Iga
6213	s6214	TV Show	Bad Education		Jack Whitehall, Mathew Horne, Sarah Solemani, ...	United Kingdom	December 15, 2018	2014	TV-MA	3 Seasons	British TV Shows, TV Comedies	A teache posh Grov
6279	s6280	TV Show	Being Mary Jane: The Series		Gabrielle Union, Lisa Vidal, Margaret Avery, O...	United States	July 1, 2017	2016	TV-14	4 Seasons	Romantic TV Shows, TV Dramas	An si jo Ma
6304	s6305	TV Show	Big Dreams, Small Spaces		Monty Don	United Kingdom	July 26, 2019	2017	TV-G	3 Seasons	British TV Shows, International TV Shows, Real...	Wr pr Mo Eng
...	...	...	...	...	...	...	...	...	...	...	...	
8539	s8540	TV Show	The Tudors		Jonathan Rhys Meyers, Henry Cavill, James Frai...	Ireland, Canada, United States, United Kingdom	January 8, 2018	2010	TV-MA	4 Seasons	TV Dramas	splen sca Er
					Martin Sheen, Rob							This p

So there are 88 rows with date added column not matching required format.This can be due to trailing white spaces so we try to remove them and check

```
df['date_added']=df['date_added'].apply(lambda x:x.strip())
date_format = '%B %d, %Y'

# Check if the date strings match the format
df['is_valid_date'] = pd.to_datetime(df['date_added'], format=date_format, errors='coerce').notna()
df['is_valid_date'].all()
```

```
True
```

Now all values in date_added column are in required format so we convert into that required format

```
date_format = '%B %d, %Y'
df['date_added'] = pd.to_datetime(df['date_added'], format=date_format)
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descript
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her fa nears the of his film
1	s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	A cros paths party, a C Tow
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy,		2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To pro his fa fro powe drug

```
df.drop(['is_valid_date'],axis=1,inplace=True)
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descript
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her fa nears the of his film
1	s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	A cros paths party, a C Tow
					Sami Bouajila,						Crime TV	To pro

```
df.dtypes
```

```
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
description             object
dtype: object
```

So now let us drop columns show-id,title and description because they may not add any value to our analysis

```
df.drop(columns=['title','description','show_id'],inplace=True)
df.head()
```

	type	director	cast	country	date_added	release_year	rating	duration	listed_in
0	Movie	Kirsten Johnson		United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	TV Show		Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
2	TV Show	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...		2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...

If we observe have columns like cast have multiple values for each movie/show beacause one or more person can be acted in a movie/show.So we can convert that into list of names and also we can convert all names into lower case and remove trailing space

Similary we repeat this procedure for listed_in column

```
df['cast']=df['cast'].apply([lambda x:[i.strip().lower() for i in x.split(",") if i!=""]])
df['listed_in']=df['listed_in'].apply([lambda x:[i.strip().lower() for i in x.split(",") if i!=""]])
df.head()
```

	type	director	cast	country	date_added	release_year	rating	duration	listed_in
0	Movie	Kirsten Johnson	[]	United States	2021-09-25	2020	PG-13	90 min	[documentaries]
1	TV Show		[ama qamata, khosi ngema, gail mabalane, thaba...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	[international tv shows, tv dramas, tv mysteries]
2	TV Show	Julien Leclercq	[sami bouajila, tracy gotoas, samuel jouy, nab...		2021-09-24	2021	TV-MA	1 Season	[crime tv shows, international tv shows, tv ac...

There can be possibilty that that one movie can be directed by one more directors

```
df['director'].apply(lambda x:"," in x).any()
```

```
True
```

```
df.loc[df['director'].apply(lambda x:"," in x)]
```

	type	director	cast	country	date_added	release_year	rating	duration	listed_in
6	Movie	Robert Cullen, José Luis Ucha	[vanessa hudgens, kimiko glenn, james marsden,...		2021-09-24	2021	PG	91 min	[children & family movies]
16	Movie	Pedro de Echave García, Pablo Azorín Williams	[]		2021-09-22	2020	TV-MA	67 min	[documentaries, international movies]
23	Movie	Alex Woo, Stanley Moore	[maisie benson, paul killam, kerry gudjohnsen,...		2021-09-21	2021	TV-Y	61 min	[children & family movies]
30	Movie	Ashwiny Iyer Tiwari, Abhishek Chaubey, Saket C...	[abhishek banerjee, rinku rajguru, delzad hiwa...		2021-09-17	2021	TV-14	111 min	[dramas, independent movies, international mov...
68	Movie	Hanns-Bruno Kammertöns, Vanessa Nöcker, Michae...	[michael schumacher]		2021-09-15	2021	TV-14	113 min	[documentaries, international movies, sports m...
...	...	...	...	...	...	...	...	...	...
8727	Movie	Ritu Sarin, Tenzing Sonam	[]	United Kingdom, India, United States	2016-12-25	2013	NR	75 min	[documentaries, international movies]
8728	Movie	Heidi Brandenburg, Mathew Orzel	[]	Peru, United States, United Kingdom	2016-11-30	2016	TV-14	103 min	[documentaries, international movies]

So if we observe there are 614 rows that contain more than one director for movie/show

8737	Movie	Hansley, Alicky	[giles yeo]	Kingdom	2019-02-01	2016	TV-14	50 min	[documentaries]

```
df['director']=df['director'].apply([lambda x:[i.strip().lower() for i in x.split(",") if i!=""]])
df.head()
```

	type	director	cast	country	date_added	release_year	rating	duration	listed_in
0	Movie	[kirsten johnson]	[]	United States	2021-09-25	2020	PG-13	90 min	[documentaries]
1	TV Show	[]	[ama qamata, khosi ngema, gail mabalane, thaba...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	[international tv shows, tv dramas, tv mysteries]
2	TV Show	[julien leclercq]	[sami bouajila, tracy gotoas, samuel jouy, nab...		2021-09-24	2021	TV-MA	1 Season	[crime tv shows, international tv shows, tv ac...

Similarly some movies/shows are produced by multiple countries

```
df['country'].apply(lambda x:"," in x).any()
```

```
True
```

```
df['country']=df['country'].apply([lambda x:[i.strip().lower() for i in x.split(",") if i!=""]])
df.head()
```

	type	director	cast	country	date_added	release_year	rating	duration	listed_in
0	Movie	[kirsten johnson]	[]	[united states]	2021-09-25	2020	PG-13	90 min	[documentaries]
1	TV Show	[]	[ama qamata, khosi ngema, gail mabalane, thaba...	[south africa]	2021-09-24	2021	TV-MA	2 Seasons	[international tv shows, tv dramas, tv mysteries]
2	TV Show	[julien leclercq]	[sami bouajila, tracy gotoas, samuel jouy, nab...	[]	2021-09-24	2021	TV-MA	1 Season	[crime tv shows, international tv shows, tv ac...

```
def get_unique(df,col):
  x=df.loc[df[col].apply(lambda x:len(x)>0)][col].explode().reset_index(drop=True)
  return set(list(x)),len(set(list(x))),x.value_counts().reset_index()
```

Since some columns have list as values like cast ,country we cant directly use unique,nunique,value_counts function directly so we defined above function for that

```
df['type'].unique()
```

```
array(['Movie', 'TV Show'], dtype=object)
```

Now we split the dataset into two parts movies and shows

```
movies=df.loc[df['type']=="Movie"].reset_index(drop=True)
movies.drop(columns="type",inplace=True)
movies.head()
```

	director	cast	country	date_added	release_year	rating	duration	listed_in
0	[kirsten johnson]	[]	[united states]	2021-09-25	2020	PG-13	90 min	[documentaries]
1	[robert cullen, josé luis ucha]	[vanessa hudgens, kimiko glenn, james marsden,...	[]	2021-09-24	2021	PG	91 min	[children & family movies]
2	[haile gerima]	[kofi ghanaba, oyafunmike ogunlano, alexandra ...	[united states, ghana, burkina faso, united ki...	2021-09-24	1993	TV-MA	125 min	[dramas, independent movies, international mov...
3	[theodore melfi]	[melissa mccarthy, chris o'dowd, kevin kline, ...	[united states]	2021-09-24	2021	PG-13	104 min	[comedies, dramas]

```
shows=df.loc[df['type']=="TV Show"].reset_index(drop=True)
shows.drop(columns="type",inplace=True)
shows.head()
```

	director	cast	country	date_added	release_year	rating	duration	listed_in
0	[]	[ama qamata, khosi ngema, gail mabalane, thaba...	[south africa]	2021-09-24	2021	TV-MA	2 Seasons	[international tv shows, tv dramas, tv mysteries]
1	[julien leclercq]	[sami bouajila, tracy gotoas, samuel jouy, nab...	[]	2021-09-24	2021	TV-MA	1 Season	[crime tv shows, international tv shows, tv ac...
2	[]	[]	[]	2021-09-24	2021	TV-MA	1 Season	[docuseries, reality tv]
3	[]	[mayur more, jitendra kumar, ranjan raj, alam ...	[india]	2021-09-24	2021	TV-MA	2 Seasons	[international tv shows, romantic tv shows, tv...

Here we are converting duration of movie to integer

```
movies['duration_in_minutes']=movies['duration'].apply(lambda x:int(x.split(" ")[0]))
movies.drop(columns="duration",inplace=True)
movies.head()
```

	director	cast	country	date_added	release_year	rating	listed_in	duration_in_minutes
0	[kirsten johnson]	[]	[united states]	2021-09-25	2020	PG-13	[documentaries]	90
1	[robert cullen, josé luis ucha]	[vanessa hudgens, kimiko glenn, james marsden,...	[]	2021-09-24	2021	PG	[children & family movies]	91
2	[haile gerima]	[kofi ghanaba, oyafunmike ogunlano, alexandra ...	[united states, ghana, burkina faso, united ki...	2021-09-24	1993	TV-MA	[dramas, independent movies, international mov...	125
	[theodore	[melissa mccarthy, chris						

Here we are converting duration of show to integer

```
shows['Number of Seasons']=shows['duration'].apply(lambda x:int(x.split(" ")[0]))
shows.drop(columns="duration",inplace=True)
shows.head()
```

	director	cast	country	date_added	release_year	rating	listed_in	Number of Seasons
0	[]	[ama qamata, khosi ngema, gail mabalane, thaba...	[south africa]	2021-09-24	2021	TV-MA	[international tv shows, tv dramas, tv mysteries]	2
1	[julien leclercq]	[sami bouajila, tracy gotoas, samuel jouy, nab...	[]	2021-09-24	2021	TV-MA	[crime tv shows, international tv shows, tv ac...	1
2	[]	[]	[]	2021-09-24	2021	TV-MA	[docuseries, reality tv]	1
	[mayur more, jitendra kumar						[international tv shows,	

Now we will define some metrics

	[mike	[kate siegel, zach gilford,	[]	2021-09-24	2021	TV-MA	[tv dramas, tv horror, tv	1

```
average_duration_of_each_movie=movies['duration_in_minutes'].mean()
print("average_duration_of_each_movie : ",np.round(average_duration_of_each_movie,2)," minutes ")
```

```
average_duration_of_each_movie :  99.58  minutes
```

```
average_number_of_seasons_for_each_show=shows['Number of Seasons'].mean()
print("average_number_of_seasons_for_each_show : ",np.round(average_number_of_seasons_for_each_show,2)," sessons ")
```

```
average_number_of_seasons_for_each_show :  1.75  sessons
```

```
avg_number_of_cast=np.round(df.loc[df['cast'].apply(lambda x:len(x))!=0]['cast'].apply(lambda x:len(x)).mean(),2)
print("avg_number_of_cast for a movie/show : ",avg_number_of_cast)
```

```
avg_number_of_cast for a movie/show :  8.04
```

```
avg_number_of_cast=np.round(movies.loc[movies['cast'].apply(lambda x:len(x))!=0]['cast'].apply(lambda x:len(x)).mean(),2)
print("avg_number_of_cast for a movie : ",avg_number_of_cast)
```

```
avg_number_of_cast for a movie :  7.87
```

```
avg_number_of_cast=np.round(shows.loc[shows['cast'].apply(lambda x:len(x))!=0]['cast'].apply(lambda x:len(x)).mean(),2)
print("avg_number_of_cast for a show : ",avg_number_of_cast)
```

```
avg_number_of_cast for a show :  8.45
```

If we observe above result generally more cast members are present in TV Shows rather than movies

```
avg_number_of_listed_in=np.round(df.loc[df['listed_in'].apply(lambda x:len(x))!=0]['listed_in'].apply(lambda x:len(x)).mean(),2)
print("avg_number_of_listed_ins for a movie/show : ",avg_number_of_listed_in)
```

```
avg_number_of_listed_ins for a movie/show :  2.19
```

```
avg_number_of_listed_in=np.round(movies.loc[movies['listed_in'].apply(lambda x:len(x))!=0]['listed_in'].apply(lambda x:len(x)).m
print("avg_number_of_listed_ins for a movie : ",avg_number_of_listed_in)
```

```
avg_number_of_listed_ins for a movie :  2.15
```

```
avg_number_of_listed_in=np.round(shows.loc[shows['listed_in'].apply(lambda x:len(x))!=0]['listed_in'].apply(lambda x:len(x)).mea
print("avg_number_of_listed_ins for a show : ",avg_number_of_listed_in)
```

```
avg_number_of_listed_ins for a show :  2.29
```

```
director_unique=get_unique(df,"director")
cast_unique=get_unique(df,"cast")
country_unique=get_unique(df,"country")
listed_in_unique=get_unique(df,"listed_in")
```

```
print("Number of unique directors : ",director_unique[1])
print("Number of unique cast members : ",cast_unique[1])
print("Number of unique countries : ",country_unique[1])
print("Number of unique listed_ins : ",listed_in_unique[1])
```

```
Number of unique directors :  4987
Number of unique cast members :  36381
Number of unique countries :  122
Number of unique listed_ins :  42
```

If we observe here rajiv chilaka directed 22 movies or shows

```
director_unique[2].head(10) #Top 10 directors by number of movies/shows directed
```

	director	count
0	rajiv chilaka	22
1	jan suter	21
2	raúl campos	19
3	suhas kadav	16
4	marcus raboy	16
5	jay karas	15
6	cathy garcia-molina	13
7	martin scorsese	12
8	youssef chahine	12
9	jay chapman	12

```
cast_unique[2].head(10) #Top 10 actors by number of movies/shows acted
```

	cast	count
0	anupam kher	43
1	shah rukh khan	35
2	julie tejwani	33
3	naseeruddin shah	32
4	takahiro sakurai	32
5	rupa bhimani	31
6	akshay kumar	30
7	om puri	30
8	yuki kaji	29
9	paresh rawal	28

```
country_unique[2].head(10) #Top 10 countries by number of movies/shows produced
```

	country	count
0	united states	3681
1	india	1046
2	united kingdom	805
3	canada	445
4	france	393
5	japan	316
6	spain	232
7	south korea	231
8	germany	226
9	mexico	169

```
listed_in_unique[2].head(10) #Top 10 Generes in movies/shows
```

	listed_in	count
0	international movies	2752
1	dramas	2426
2	comedies	1674
3	international tv shows	1349
4	documentaries	869
5	action & adventure	859
6	tv dramas	762
7	independent movies	756
8	children & family movies	641
9	romantic movies	616

```
df['rating'].value_counts().reset_index() #Top 10 ratings in movies/shows
```

	rating	count
0	TV-MA	3205
1	TV-14	2157
2	TV-PG	861
3	R	799
4	PG-13	490
5	TV-Y7	333
6	TV-Y	306
7	PG	287
8	TV-G	220
9	NR	79
10	G	41
11	TV-Y7-FV	6
12	NC-17	3
13	UR	3

```
movie_director_unique=get_unique(movies,"director")
movie_cast_unique=get_unique(movies,"cast")
movie_country_unique=get_unique(movies,"country")
movie_listed_in_unique=get_unique(movies,"listed_in")


print("Number of unique movie directors : ",movie_director_unique[1])
print("Number of unique movie cast members : ",movie_cast_unique[1])
```

```
print("Number of unique  movie countries : ",movie_country_unique[1])
print("Number of unique  movie listed_ins : ",movie_listed_in_unique[1])
```

```
Number of unique movie directors :  4771
Number of unique movie cast members :  25936
Number of unique  movie countries :  117
Number of unique  movie listed_ins :  20
```

```
movie_director_unique[2].head(10) #Top 10 directors by number of movies directed
```

	director	count
0	rajiv chilaka	22
1	jan suter	21
2	raúl campos	19
3	suhas kadav	16
4	jay karas	15
5	marcus raboy	15
6	cathy garcia-molina	13
7	martin scorsese	12
8	youssef chahine	12
9	jay chapman	12

```
movie_cast_unique[2].head(10) #Top 10 actors by number of movies acted
```

	cast	count
0	anupam kher	42
1	shah rukh khan	35
2	naseeruddin shah	32
3	om puri	30
4	akshay kumar	30
5	julie tejwani	28
6	paresh rawal	28
7	amitabh bachchan	28
8	boman irani	27
9	rupa bhimani	27

```
movie_country_unique[2].head(10) #Top 10 countries by number of movies produced
```

	country	count
0	united states	2749
1	india	962
2	united kingdom	534
3	canada	319
4	france	303
5	germany	182
6	spain	171
7	japan	119
8	china	114
9	mexico	111

Insight and Reason

Netflix's strategy to acquire a large number of movies and TV shows from the United States is driven by Hollywood's global influence, the high production output and diversity of American content, strategic business decisions, and economic factors. These elements combined

make U.S. content a cornerstone of Netflix's library, catering to both local and international audiences and ensuring sustained subscriber growth and engagement. By analyzing the data, we can observe these trends and understand the rationale behind Netflix's content acquisition strategy.

Recommendation

Foster strategic partnerships with major Hollywood studios and production companies to secure exclusive streaming rights for blockbuster movies and popular TV series.

```
movie_listed_in_unique[2].head(10) #Top 10 geners in movies
```

	listed_in	count
0	international movies	2752
1	dramas	2426
2	comedies	1674
3	documentaries	869
4	action & adventure	859
5	independent movies	756
6	children & family movies	641
7	romantic movies	616
8	thrillers	577
9	music & musicals	375

```
movies['rating'].value_counts().reset_index() #Top 10 ratings in movies
```

	rating	count
0	TV-MA	2062
1	TV-14	1427
2	R	797
3	TV-PG	540
4	PG-13	490
5	PG	287
6	TV-Y7	139
7	TV-Y	131
8	TV-G	126
9	NR	75
10	G	41
11	TV-Y7-FV	5
12	NC-17	3
13	UR	3

```
show_director_unique=get_unique(shows,"director")
show_cast_unique=get_unique(shows,"cast")
show_country_unique=get_unique(shows,"country")
show_listed_in_unique=get_unique(shows,"listed_in")
```

```
print("Number of unique show directors : ",show_director_unique[1])
print("Number of unique show cast members : ",show_cast_unique[1])
print("Number of unique  show countries : ",show_country_unique[1])
print("Number of unique  show listed_ins : ",show_listed_in_unique[1])
```

```
Number of unique show directors :  299
Number of unique show cast members :  14800
Number of unique  show countries :  65
Number of unique  show listed_ins :  22
```

```
show_director_unique[2].head(10) #Top 10 directors by number of shows directed
```

	director	count
0	alastair fothergill	3
1	ken burns	3
2	jung-ah im	2
3	gautham vasudev menon	2
4	iginio straffi	2
5	hsu fu-chun	2
6	stan lathan	2
7	joe berlinger	2
8	shin won-ho	2
9	lynn novick	2

```
show_cast_unique[2].head(10) #Top 10 actorsby number of shows acted
```

	cast	count
0	takahiro sakurai	25
1	yuki kaji	19
2	daisuke ono	17
3	junichi suwabe	17
4	yuichi nakamura	16
5	ai kayano	16
6	jun fukuyama	15
7	yoshimasa hosoya	15
8	david attenborough	14
9	hiroshi kamiya	13

```
show_country_unique[2].head(10) #Top 10 countries by number of shows produced
```

	country	count
0	united states	932
1	united kingdom	271
2	japan	197
3	south korea	170
4	canada	126
5	france	90
6	india	84
7	taiwan	70
8	australia	64
9	spain	61

```
show_listed_in_unique[2].head(10)#Top 10 genres in shows
```

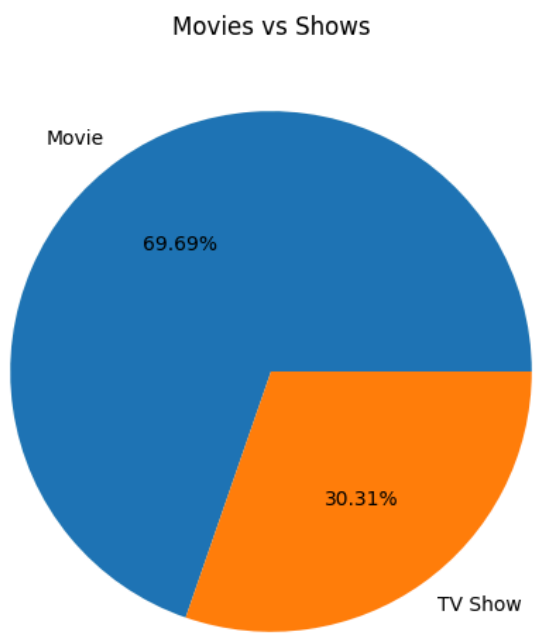
	listed_in	count
0	international tv shows	1349
1	tv dramas	762
2	tv comedies	573
3	crime tv shows	469
4	kids' tv	448
5	docuseries	394
6	romantic tv shows	370
7	reality tv	255
8	british tv shows	252
9	anime series	174

```
shows['rating'].value_counts() #Top 10 ratings in shows
```

```
rating
TV-MA       1143
TV-14        730
TV-PG        321
TV-Y7        194
TV-Y         175
TV-G          94
NR             4
R              2
TV-Y7-FV       1
Name: count, dtype: int64
```

Let us make pie chart on Number of Movies vs Tv Shows by netflix

```
plt.figure(figsize=(10,6))
plt.pie(df['type'].value_counts(),labels=df['type'].value_counts().index,autopct="%.2f%%")
plt.title("Movies vs Shows")
plt.show()
```



Insight and Reason

The observation that there are more movies than TV shows in a Netflix dataset can be attributed to several factors related to content acquisition, audience preferences, and platform strategy:

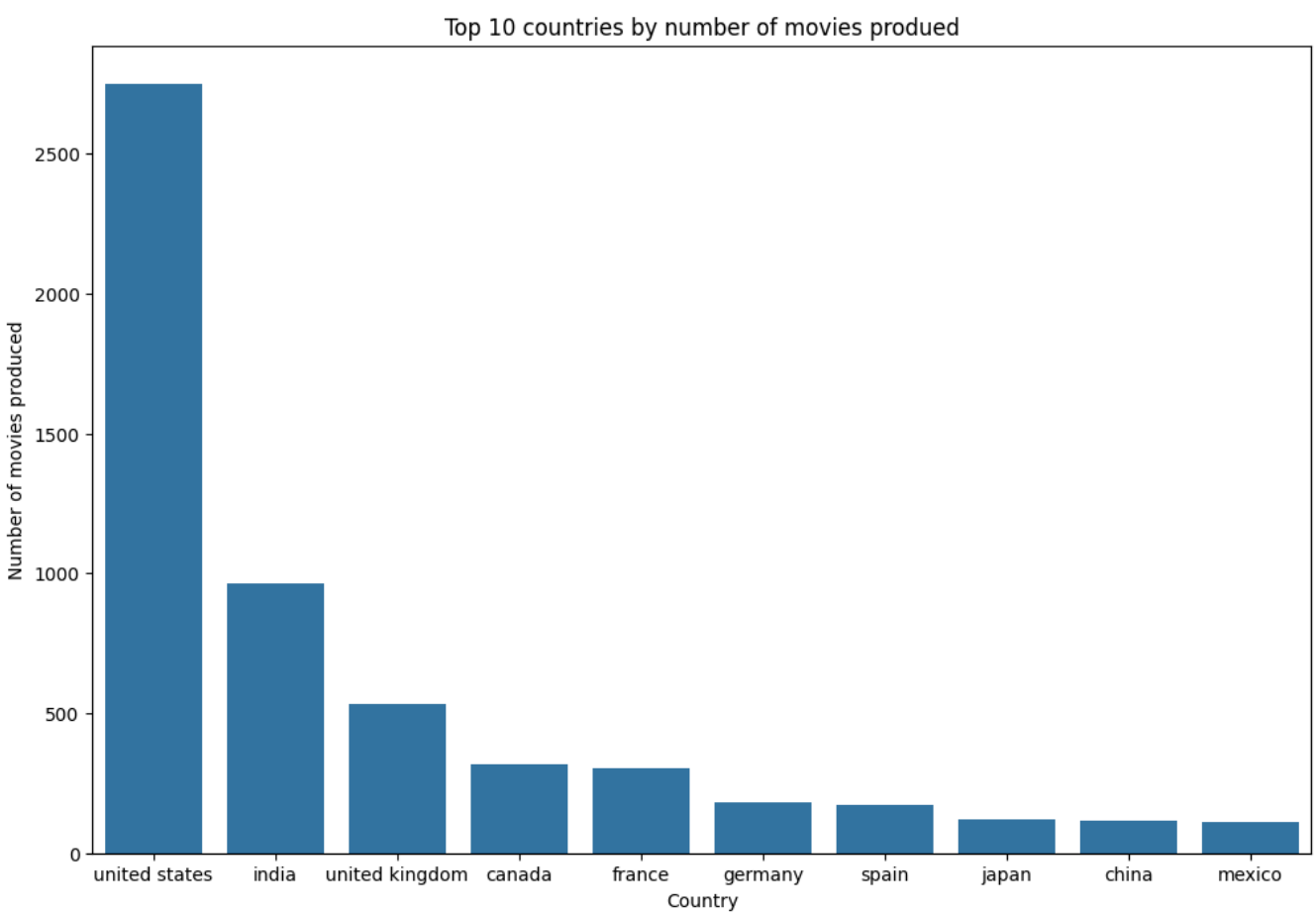
Variety and Global Appeal Broad Content Selection: Movies span various genres, languages, and cultural backgrounds, appealing to a wide audience globally. This diversity allows Netflix to cater to different viewer preferences and maximize its content library's appeal.

Global Licensing: Netflix acquires movies from around the world through licensing agreements with studios and distributors. This allows them to offer a broad range of content to their global subscriber base.

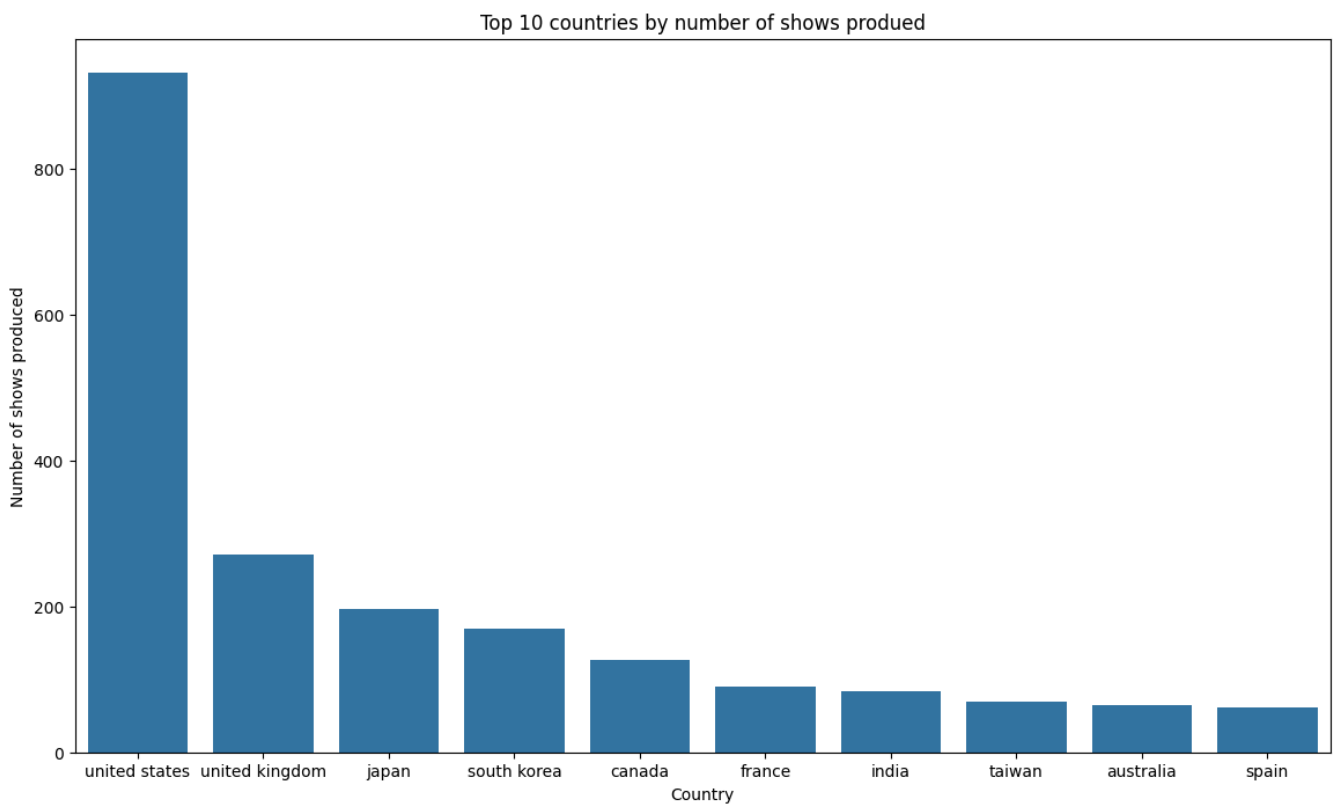
Recommendation

Predict how this balance might evolve over time based on emerging industry trends and invest accordingly

```
plt.figure(figsize=(12,8))
d=movie_country_unique[2].head(10)
sns.barplot(x="country",y="count",data=d)
plt.xlabel("Country")
plt.ylabel("Number of movies produced")
plt.title("Top 10 countries by number of movies produed")
plt.show()
```




```
plt.figure(figsize=(14,8))
d=show_country_unique[2].head(10)
sns.barplot(x="country",y="count",data=d)
plt.xlabel("Country")
plt.ylabel("Number of shows produced")
plt.title("Top 10 countries by number of shows produed")
plt.show()
```



Insight and Reason

If observe the above two plots we can observe netflix has aquired more movies from some countries but less number of shows from them

For example from india netflix has aquired second most number of movies among all countries but it aquired less number of movies compared to other countries.This can be due to following reasons:

Netflix's strategy to acquire more movies than TV shows from India is driven by historical and cultural preferences, economic efficiencies, market demand, and strategic business goals. This approach allows Netflix to tap into the vast and diverse Indian film industry, cater to local and global audiences, and strengthen its market position. By analyzing the data, we can observe these trends and validate the underlying reasons for Netflix's content acquisition strategy.

Recommendation

Forecast how Netflix’s acquisition of Indian movies might evolve in the future.

Will Netflix continue to invest in Indian cinema, and how might this impact its global market position?

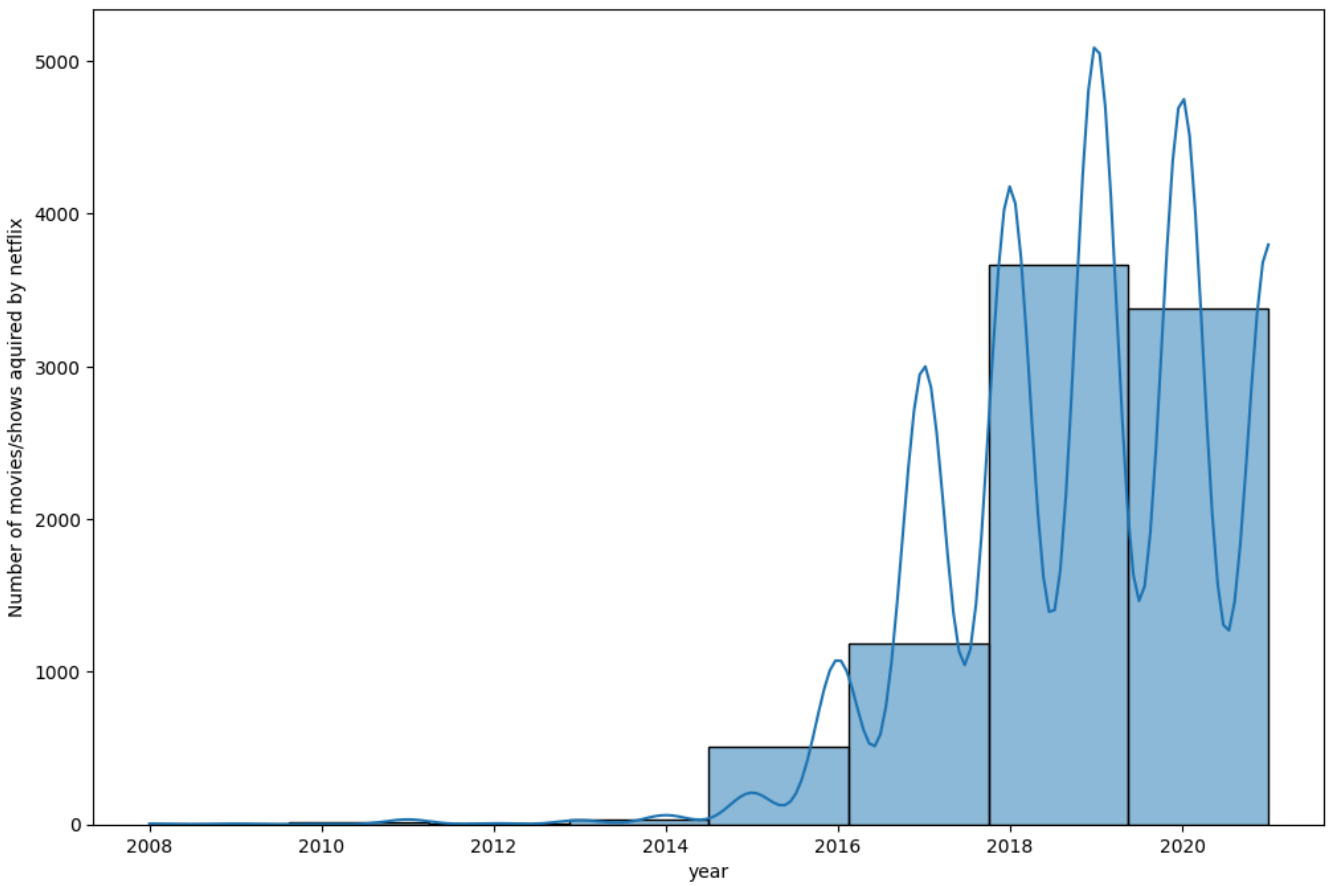
Anticipate trends in viewer preferences and consumption patterns in the streaming industry.

```
year=df['date_added'].dt.year
year.value_counts().reset_index().sort_values("date_added")
```

	date_added	count
12	2008	2
11	2009	2
13	2010	1
8	2011	13
10	2012	3
9	2013	11
7	2014	24
6	2015	82
5	2016	426
4	2017	1185
2	2018	1648
0	2019	2016
1	2020	1879
3	2021	1498

Double-click (or enter) to edit

```
year=df['date_added'].dt.year
plt.figure(figsize=(12,8))
plt.xlabel("year")
plt.ylabel("Number of movies/shows aquired by netflix")
sns.histplot(year,kde=True,bins=8)
plt.show()
```



Insight and Reason

If we observe number of shows/movies aquired by netflix have decreased from 2019 this can be due to

1.COVID-19 Pandemic:

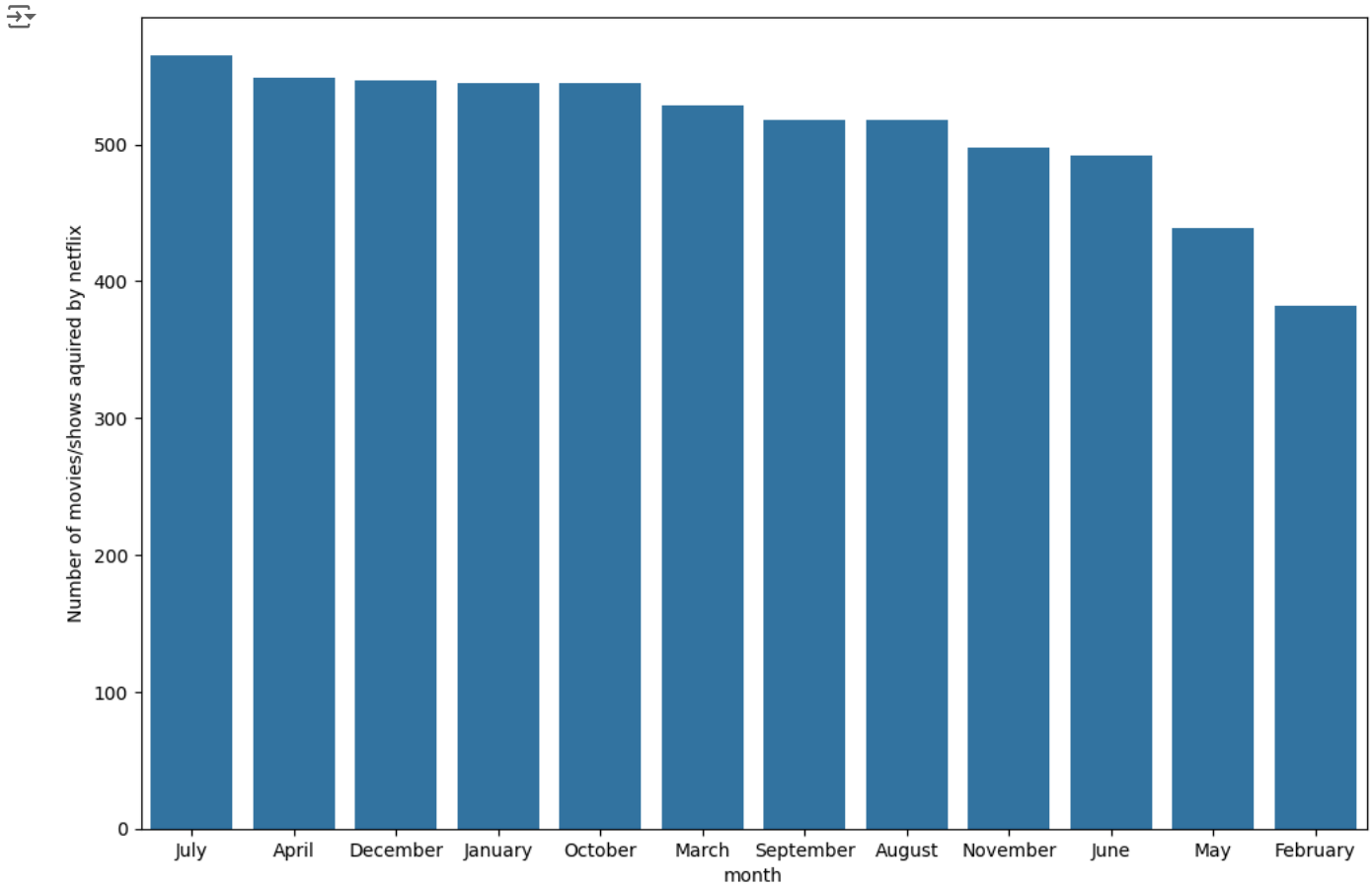
Global Impact on Production: The COVID-19 pandemic caused significant disruptions in film and TV show production worldwide, leading to delays and halts. Many projects that were slated for release or acquisition in 2020 and 2021 were postponed.

Safety Protocols: Even as productions resumed, new safety protocols and regulations slowed down the production process, reducing the overall output of new content.

2.Market Saturation and Competition

New Streaming Services: The emergence and growth of other streaming platforms like Disney+, HBO Max, Apple TV+, and Amazon Prime Video have intensified competition for acquiring content

```
month=movies['date_added'].dt.month_name()
plt.figure(figsize=(12,8))
plt.xlabel("month")
plt.ylabel("Number of movies/shows aquired by netflix")
sns.barplot(x=month.value_counts().index,y=month.value_counts())
plt.show()
```



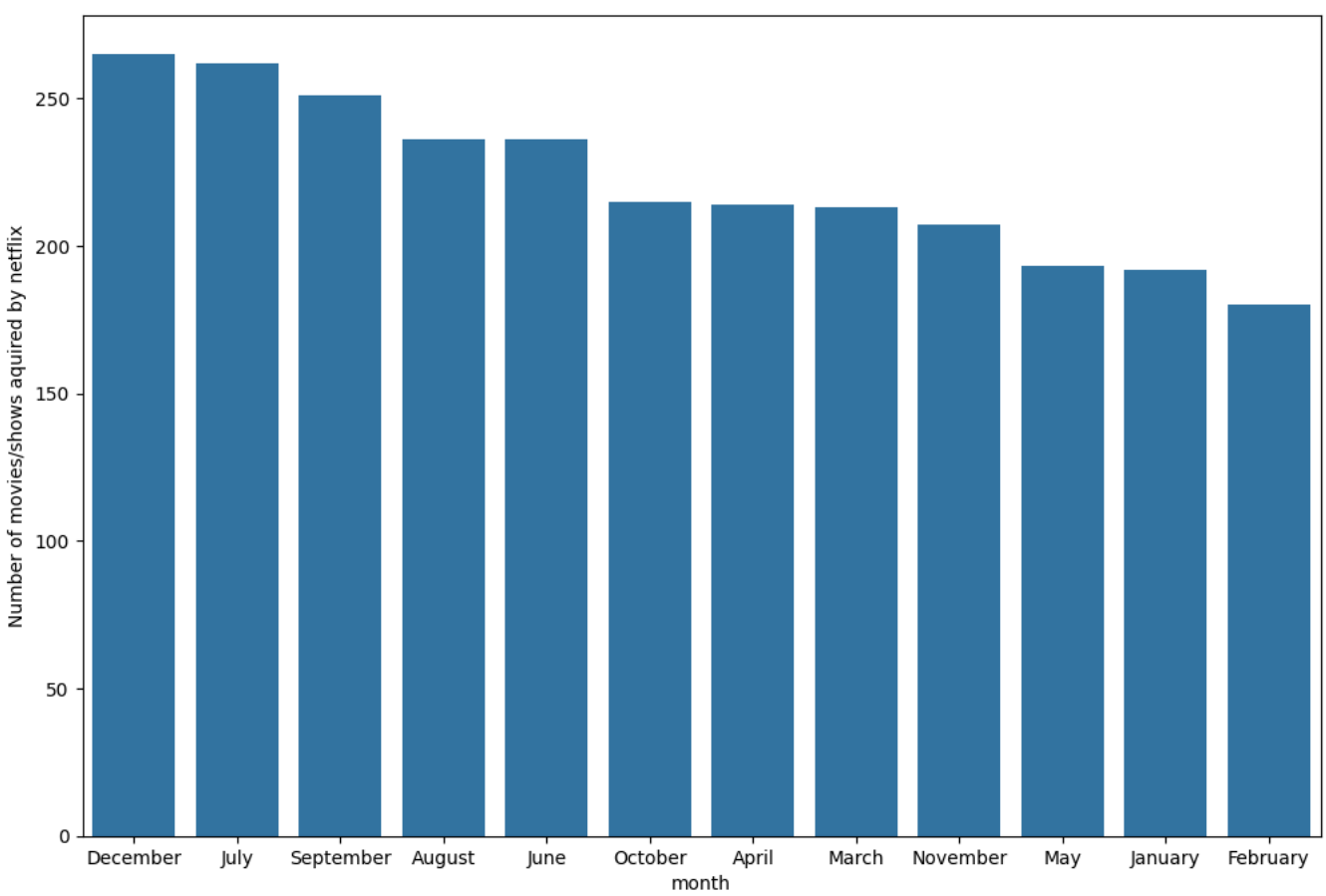
Insight and Reason

The observation that more movies are released by Netflix in July and April months in a Netflix dataset could be influenced by several factors related to content strategy, industry trends, and audience behavior. Here are some potential reasons:

Seasonal Trends and Holidays Summer Releases: July is often considered a peak period for movie releases in general, as it coincides with summer vacations in many countries. Netflix may strategically release more content during this time to capture increased viewership.

Spring Releases: April could also see increased releases due to spring breaks and holidays in various regions, leading to higher potential viewership.

```
month=shows['date_added'].dt.month_name()
plt.figure(figsize=(12,8))
plt.xlabel("month")
plt.ylabel("Number of movies/shows aquired by netflix")
sns.barplot(x=month.value_counts().index,y=month.value_counts())
plt.show()
```



Insight and Reason

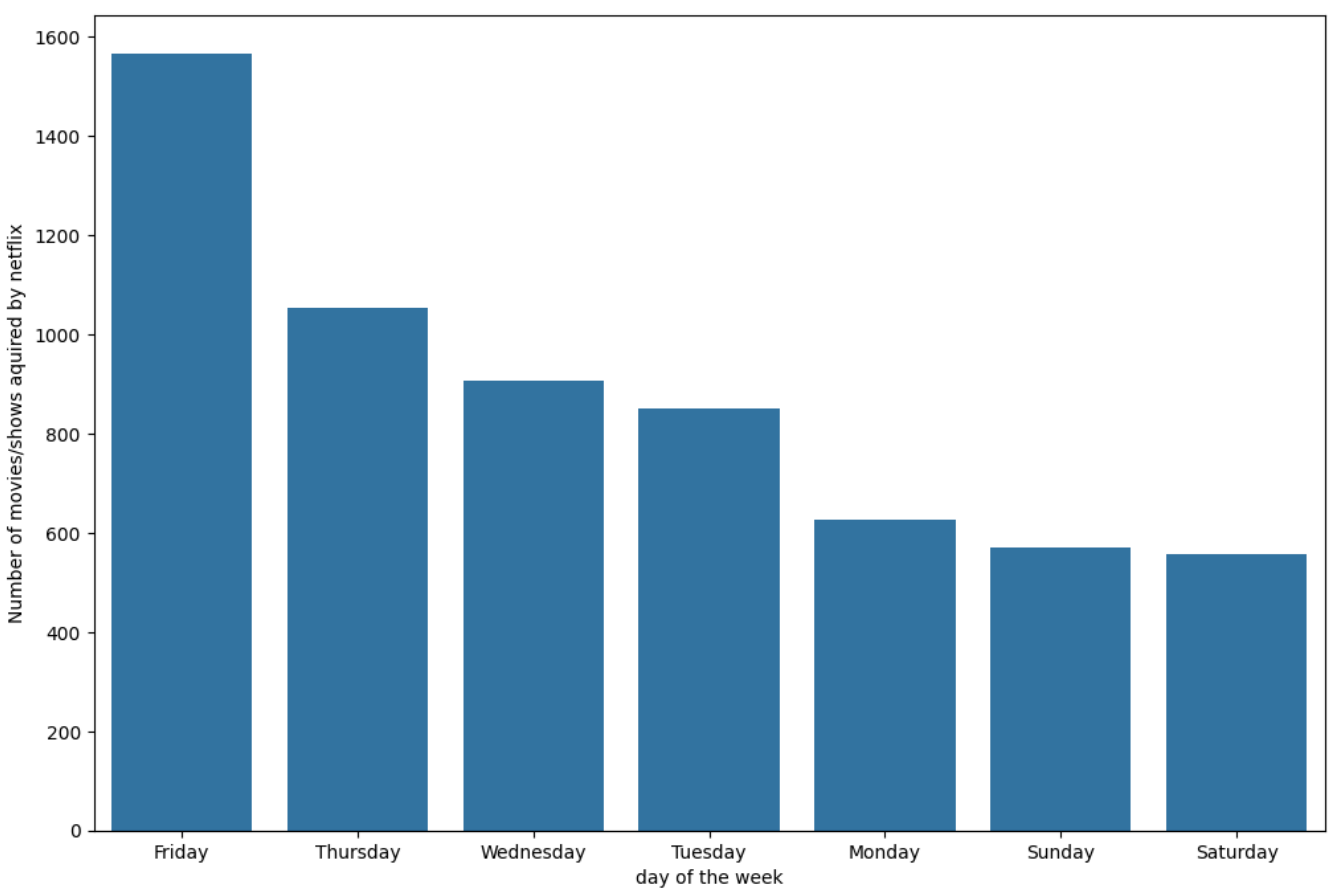
The trend of launching more shows in November and December by Netflix can be attributed to several strategic and practical considerations:

Holiday Season and Viewer Behavior:

Holiday Viewing: November and December coincide with major holidays such as Thanksgiving, Christmas, and New Year's Eve. During these festive periods, people often have more leisure time and are likely to engage in binge-watching. Netflix strategically releases new shows to capitalize on increased viewer engagement during these months.

Cultural and Regional Preferences: In many cultures, the holiday season is a time for family gatherings and indoor activities. Releasing new shows during this period aligns with cultural norms and viewer behavior patterns, maximizing viewership.

```
weekday=movies['date_added'].dt.day_name()
plt.figure(figsize=(12,8))
plt.xlabel("day of the week")
plt.ylabel("Number of movies/shows aquired by netflix")
sns.barplot(x=weekday.value_counts().index,y=weekday.value_counts())
plt.show()
```



Insight and Reason

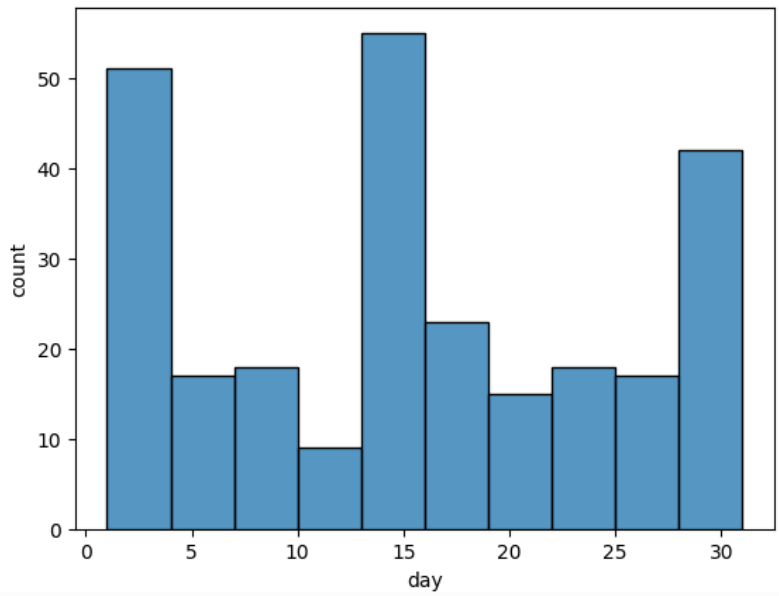
Netflix launches movies or shows on friday due to

Weekend Viewing: Fridays mark the beginning of the weekend for many people globally. Netflix capitalizes on this by releasing new content when viewers are more likely to have leisure time and engage in binge-watching.

Binge-Watching Culture: Many Netflix subscribers prefer to watch multiple episodes or movies in one sitting, especially over the weekend. Releasing new content on Fridays encourages binge-watching behavior and increases viewer retention.

```
x=shows.loc[shows["date_added"].dt.month==12]['date_added'].dt.day
plt.xlabel("day")
plt.ylabel("count")
sns.histplot(x)
```

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



Insight and Reason

The timing of show launches on Netflix, particularly why more shows are launched during December 10-15 compared to December 20-25, can be influenced by several strategic considerations:

1.Lead Time and Marketing Strategy Holiday Promotions: Shows launched during December 10-15 have more lead time before Christmas and New Year's Eve. This timing allows Netflix to build anticipation through marketing campaigns, generating buzz and maximizing visibility before the holiday season peaks.

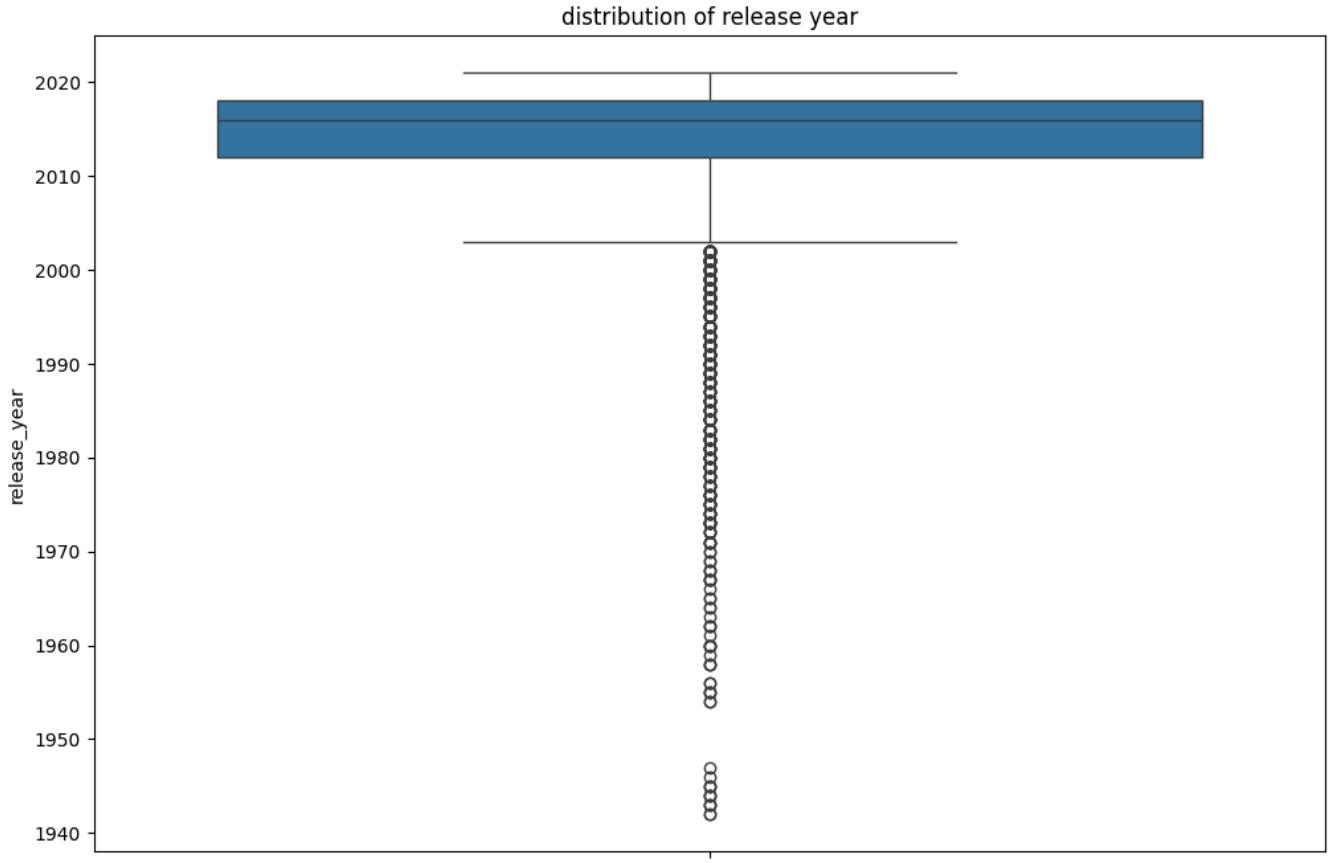
2.Avoiding Holiday Rush: Releasing shows earlier in December (around December 10-15) helps Netflix avoid the intense competition and saturation of content releases that typically occur closer to Christmas (December 20-25). By launching earlier, Netflix can capture viewer attention before they become overwhelmed with holiday activities and distractions.

Recommendation

If we have observed above insights Mid of december is best time to launch a TV Show

```
plt.figure(figsize=(12,8))
plt.title("distribution of release year")
sns.boxplot(movies['release_year'])
```

```
<Axes: title={'center': 'distribution of release year'}, ylabel='release_year'>
```



Insight and Reason

The observation that most movies released in a Netflix dataset are after the year 2000 can be attributed to several factors related to the evolution of the film industry, content acquisition strategies by streaming platforms like Netflix, and the availability and popularity of digital content. Here are some reasons why this might be the case:

Digital Transformation of Content Shift to Digital Distribution: Beginning in the late 1990s and early 2000s, there was a significant shift in the film industry towards digital distribution formats. This made it easier for newer movies to be distributed and streamed online, including through platforms like Netflix.

Availability of Digital Copies: Many older movies were originally distributed on physical formats like VHS tapes or DVDs, and they were gradually digitized and made available on streaming platforms post-2000.

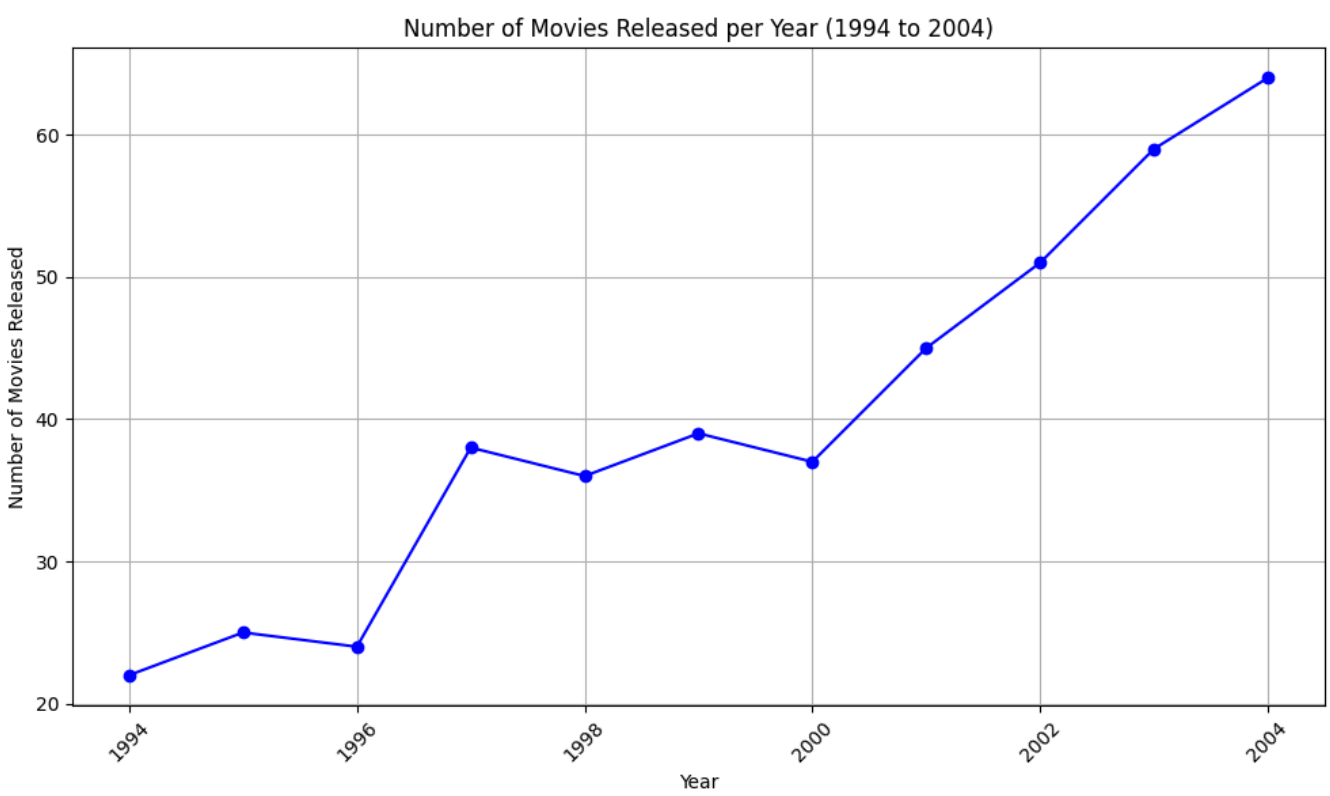
Recommendation

Consider the implications for user experience and satisfaction with the content library.

How does the availability of recent movies contribute to personalized recommendations and user retention?

Are there opportunities to further enhance the diversity or depth of recent movie offerings?

```
current_year = pd.Timestamp.now().year
start_year = current_year - 30  # 30 years ago
end_year = current_year - 20    # 20 years ago
movie_counts = df[df['release_year'].between(start_year, end_year)].groupby('release_year').size()
plt.figure(figsize=(10, 6))
plt.plot(movie_counts.index, movie_counts.values, marker='o', linestyle='-', color='b')
plt.title('Number of Movies Released per Year ({} to {})'.format(start_year, end_year))
plt.xlabel('Year')
plt.ylabel('Number of Movies Released')
plt.grid(True)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Insight and Reason

If we observe the above plot we can see that number of movies aquired from 1994 t0 2004 shown increasing trend.This can be due to

Growth of the Film Industry Expansion of Production Studios: Over the past few decades, there has been a proliferation of production studios globally, leading to an increase in the number of movies being produced each year.

Globalization: The film industry has become more globalized, with production companies from various countries contributing to the overall increase in movie releases.

Recommendation

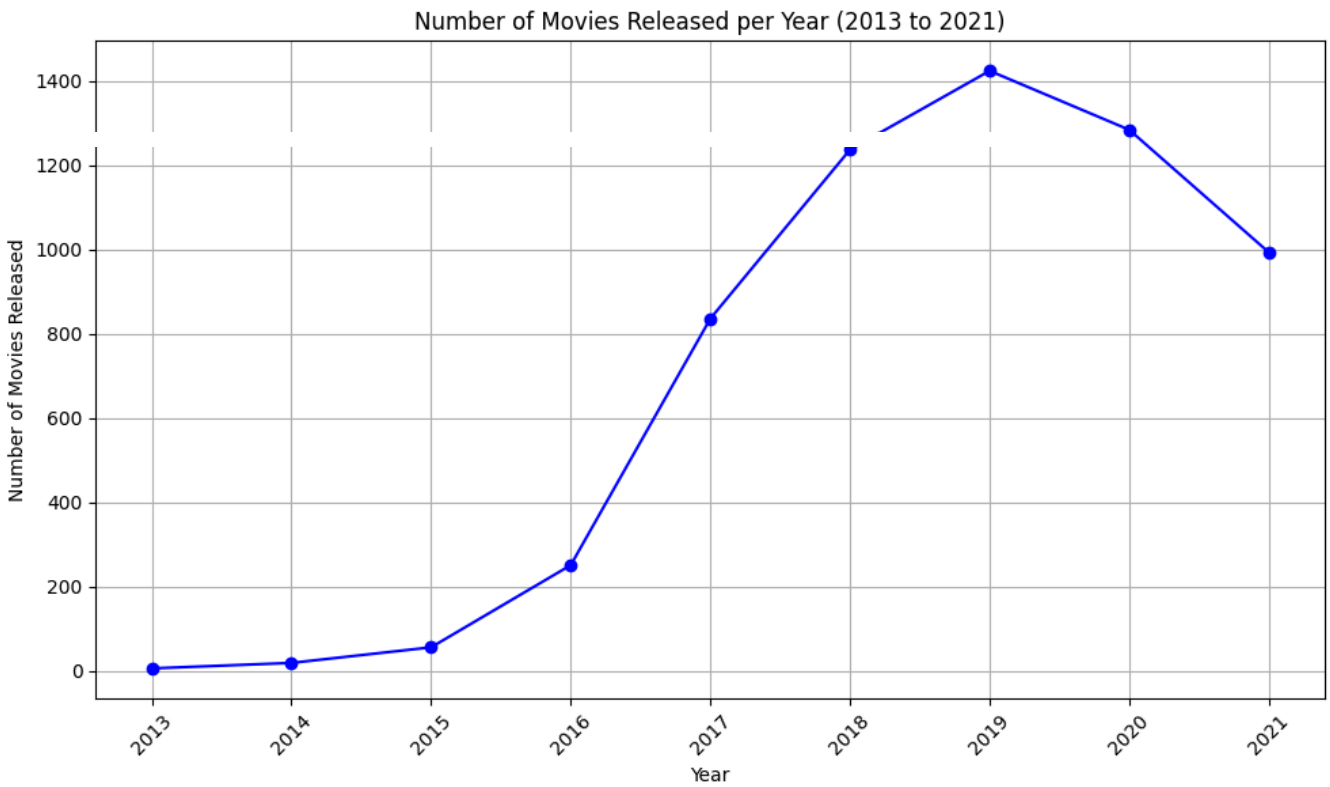
Analyze how this historical trend informs Netflix's current content strategy.

Consider how Netflix might leverage historical acquisition patterns to guide future content investments.

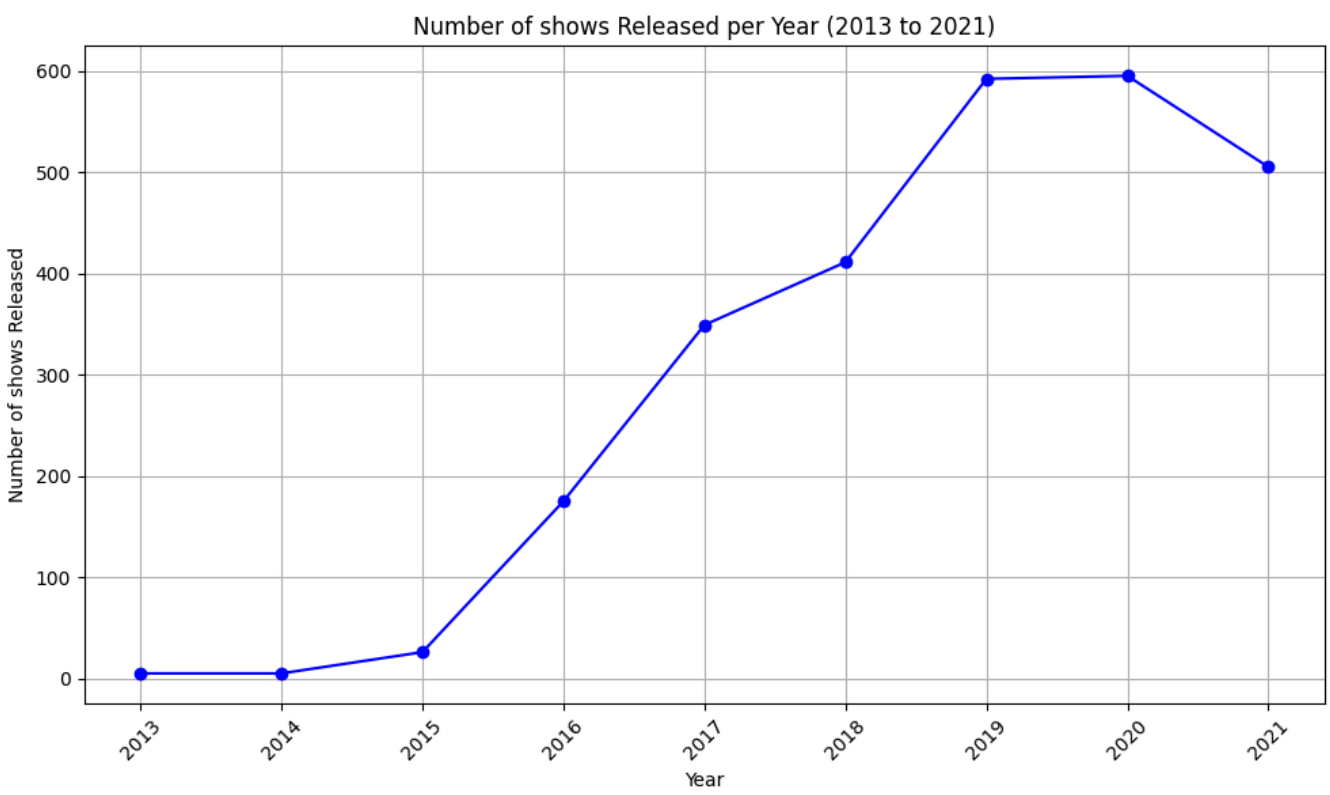
Discuss implications for subscriber retention, market positioning, and competitive advantage in the streaming landscape.

```
current_year = pd.Timestamp.now().year
start_year = movies['date_added'].dt.year.max() - 8
end_year = movies['date_added'].dt.year.max()
movie_counts =movies[movies['date_added'].dt.year.between(start_year, end_year)]['date_added'].dt.year.value_counts().reset_inde
plt.figure(figsize=(10, 6))
plt.plot(movie_counts['date_added'], movie_counts['count'] ,marker='o', linestyle='-', color='b')
```

```
plt.title('Number of Movies Released per Year ({} to {})'.format(start_year, end_year))
plt.xlabel('Year')
plt.ylabel('Number of Movies Released')
plt.grid(True)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```




```
current_year = pd.Timestamp.now().year
start_year = shows['date_added'].dt.year.max() - 8
end_year = shows['date_added'].dt.year.max()
show_counts =shows[shows['date_added'].dt.year.between(start_year, end_year)]['date_added'].dt.year.value_counts().reset_index()
plt.figure(figsize=(10, 6))
plt.plot(show_counts['date_added'], show_counts['count'] ,marker='o', linestyle='-', color='b')
plt.title('Number of shows Released per Year ({} to {})'.format(start_year, end_year))
plt.xlabel('Year')
plt.ylabel('Number of shows Released')
plt.grid(True)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Insight and Reason

Netflix's shift towards focusing more on TV shows than movies in recent years can be attributed to several strategic and market-driven factors:

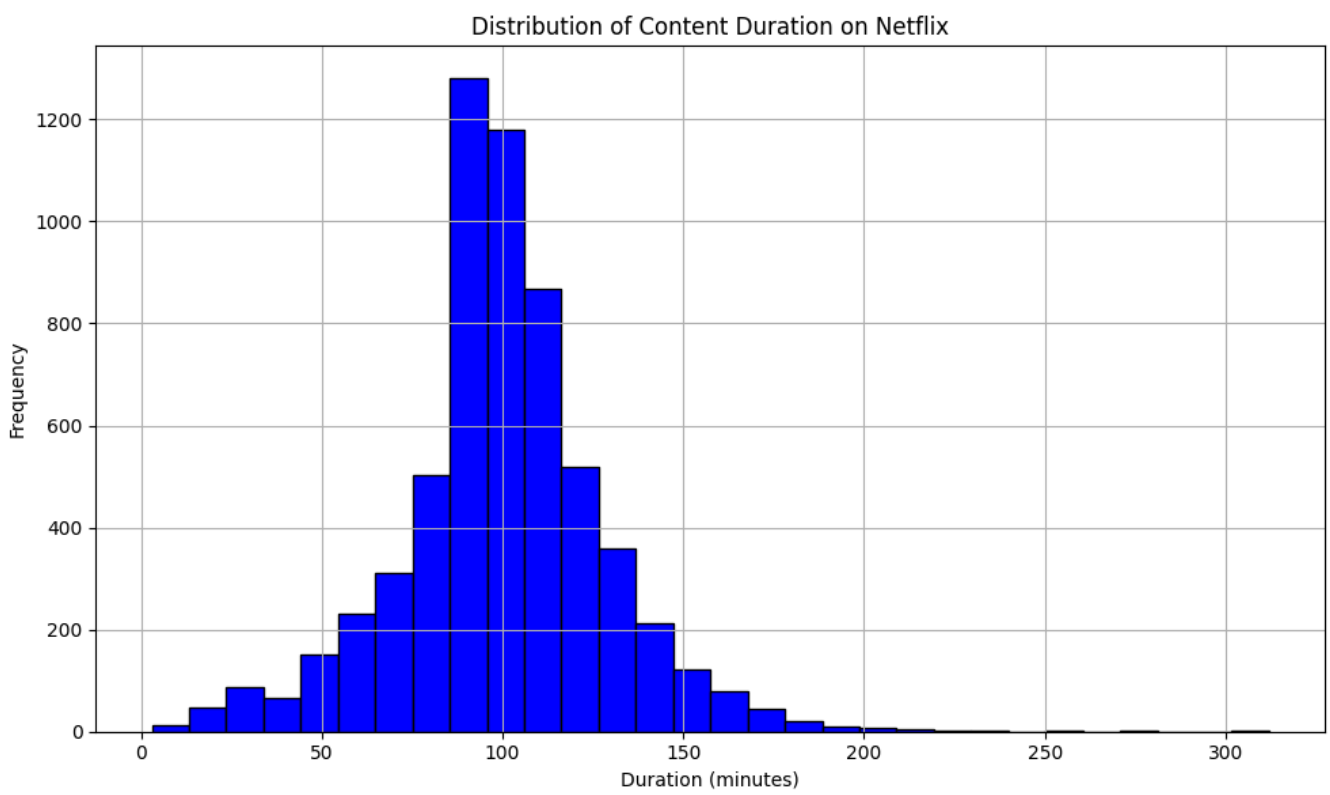
1. Audience Engagement and Viewing Behavior Changing Viewer Preferences: There has been a noticeable trend towards binge-watching TV series rather than watching movies in one sitting. TV shows provide longer engagement periods and can keep subscribers subscribed for longer durations.

Global Appeal: TV shows often have broader international appeal compared to movies, as they can cater to diverse cultural and linguistic preferences with serialized storytelling.

2. Content Production Economics Cost Efficiency: Producing TV shows can be more cost-effective than producing movies, especially when considering the duration of viewer engagement per dollar spent on content production.

Long-Term Content Strategy: TV shows allow for the development of a dedicated fan base and can generate ongoing interest over multiple seasons, enhancing long-term subscriber retention.

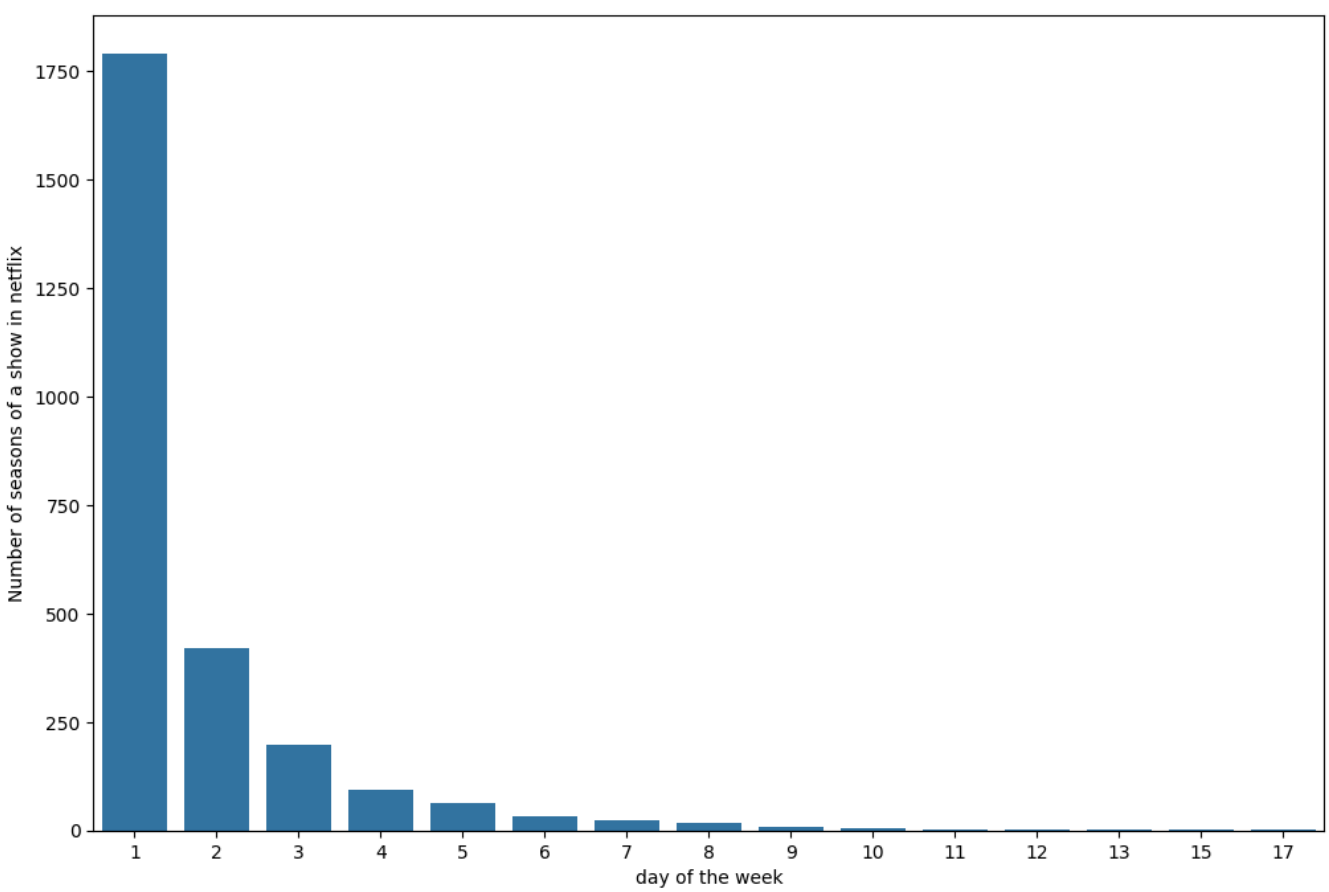
```
plt.figure(figsize=(10, 6))
plt.hist(movies['duration_in_minutes'], bins=30, color='blue', edgecolor='black')
plt.title('Distribution of Content Duration on Netflix')
plt.xlabel('Duration (minutes)')
plt.ylabel('Frequency')
plt.grid(True)
plt.tight_layout()
plt.show()
```



Insight and Reason

The prevalence of movies around 100 minutes in duration on Netflix likely reflects a combination of audience preferences, industry norms, production economics, and platform strategies aimed at optimizing viewer satisfaction and content diversity. While there are movies of various lengths available on Netflix, the popularity of movies around this duration indicates a balance between narrative completeness and viewer accessibility in the streaming era.

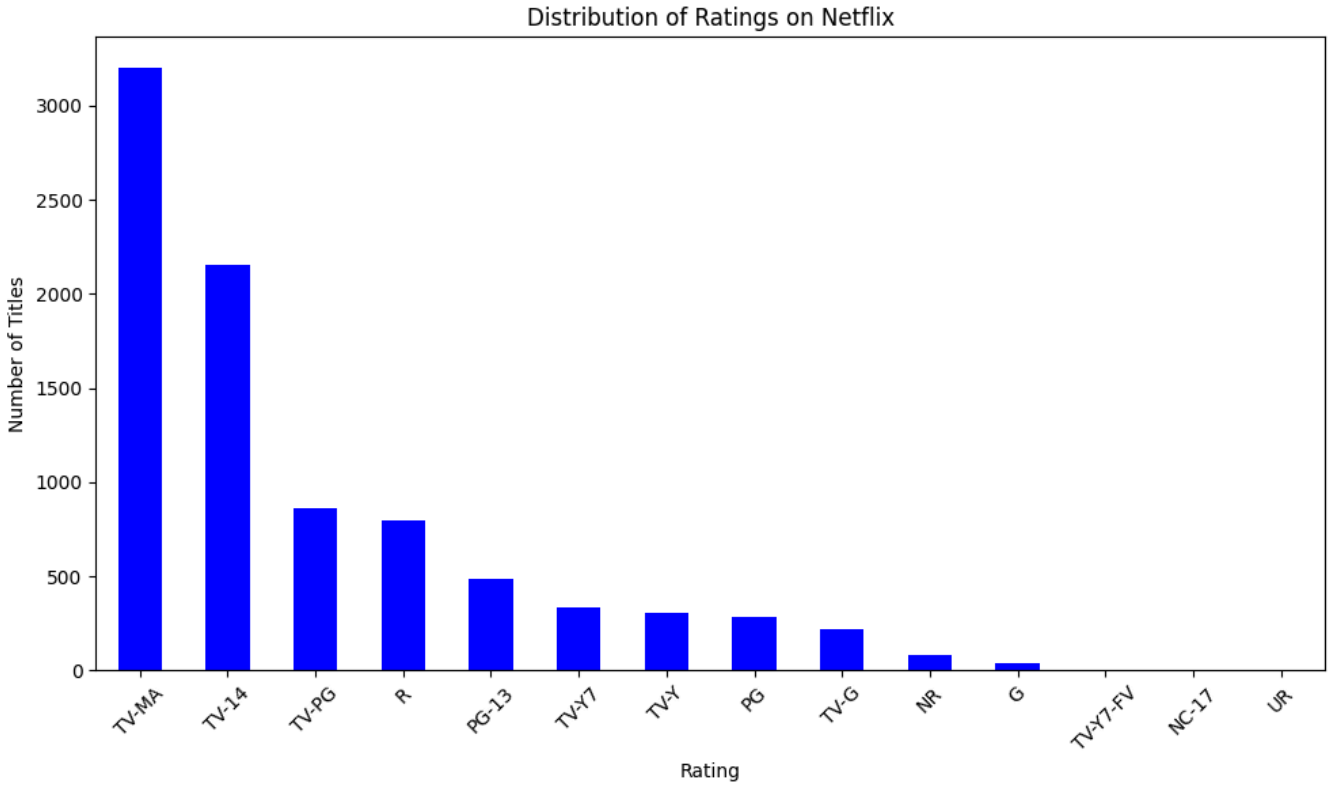
```
season=shows['Number of Seasons']
plt.figure(figsize=(12,8))
plt.xlabel("day of the week")
plt.ylabel("Number of seasons of a show in netflix")
sns.barplot(x=season.value_counts().index,y=season.value_counts())
plt.show()
```



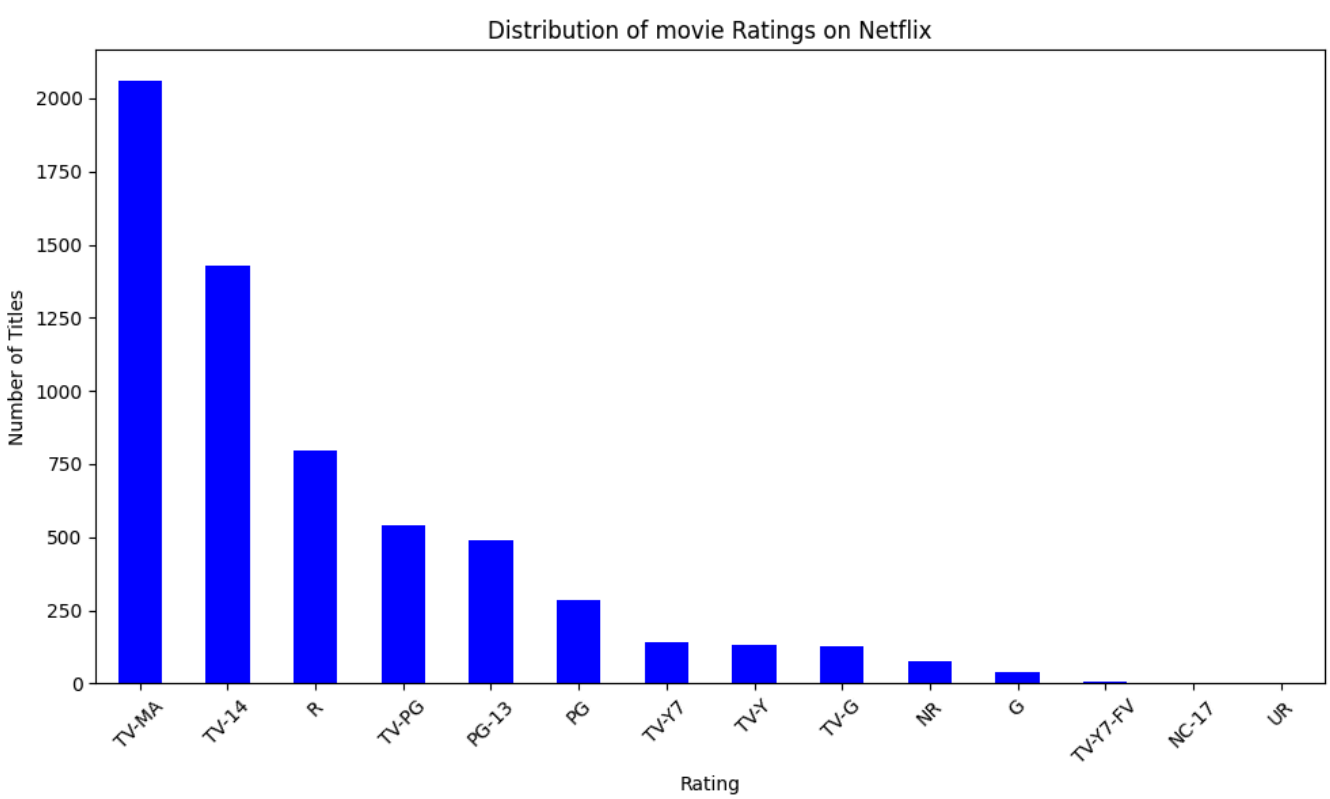
Insight and Reason

The predominance of TV shows with 1 or 2 seasons on Netflix reflects strategic decisions influenced by viewer behavior, production economics, licensing availability, and platform strategy. While there are exceptions with longer-running series and original productions, the focus on shorter series initially allows Netflix to manage risks, optimize viewer engagement, and maintain a dynamic content catalog that appeals to global audiences.

```
plt.figure(figsize=(10, 6))
df['rating'].value_counts().plot(kind='bar', color='blue')
plt.title('Distribution of Ratings on Netflix')
plt.xlabel('Rating')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

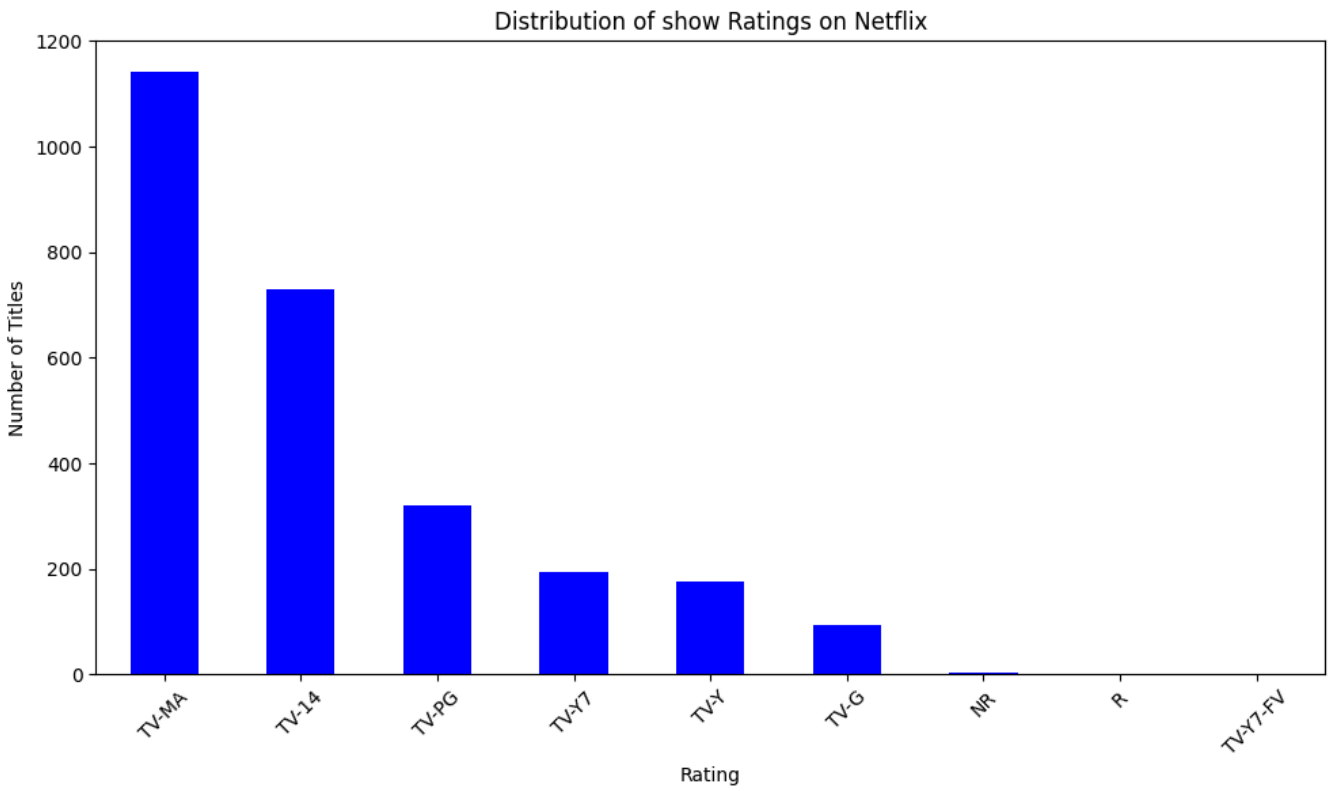



```
plt.figure(figsize=(10, 6))
movies['rating'].value_counts().plot(kind='bar', color='blue')
plt.title('Distribution of movie Ratings on Netflix')
plt.xlabel('Rating')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```




```
plt.figure(figsize=(10, 6))
shows['rating'].value_counts().plot(kind='bar', color='blue')
plt.title('Distribution of show Ratings on Netflix')
```

```
plt.xlabel('Rating')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Insight and Reason

The prevalence of TV-MA (Mature Audience) and TV-14 rating movies/shows on Netflix can be attributed to several strategic and audience-related factors:

Targeting Older Audiences: TV-MA and TV-14 ratings cater to older audiences who prefer content with mature themes, stronger language, and more intense or graphic content. Netflix's data analytics likely show a significant portion of their subscriber base falls into these demographics.

Diverse Content Preferences: Offering a variety of content ratings ensures Netflix can appeal to a broad spectrum of viewers, including those seeking more adult-oriented programming.

The prevalence of TV-PG ratings in TV shows compared to movies in a Netflix dataset can be influenced by several factors related to content categorization, viewer demographics, and platform strategy. Here are some possible reasons for this observation:

Target Audience: TV-PG ratings are typically designed for a broader audience, including children and families. TV shows often cater to diverse age groups, including younger viewers, where content is more likely to be rated as TV-PG to ensure broader accessibility.

Content Variety: Netflix may prioritize TV shows with TV-PG ratings to cater to family viewing preferences and ensure a balanced content library that appeals to a wide range of viewers.

Recommendations

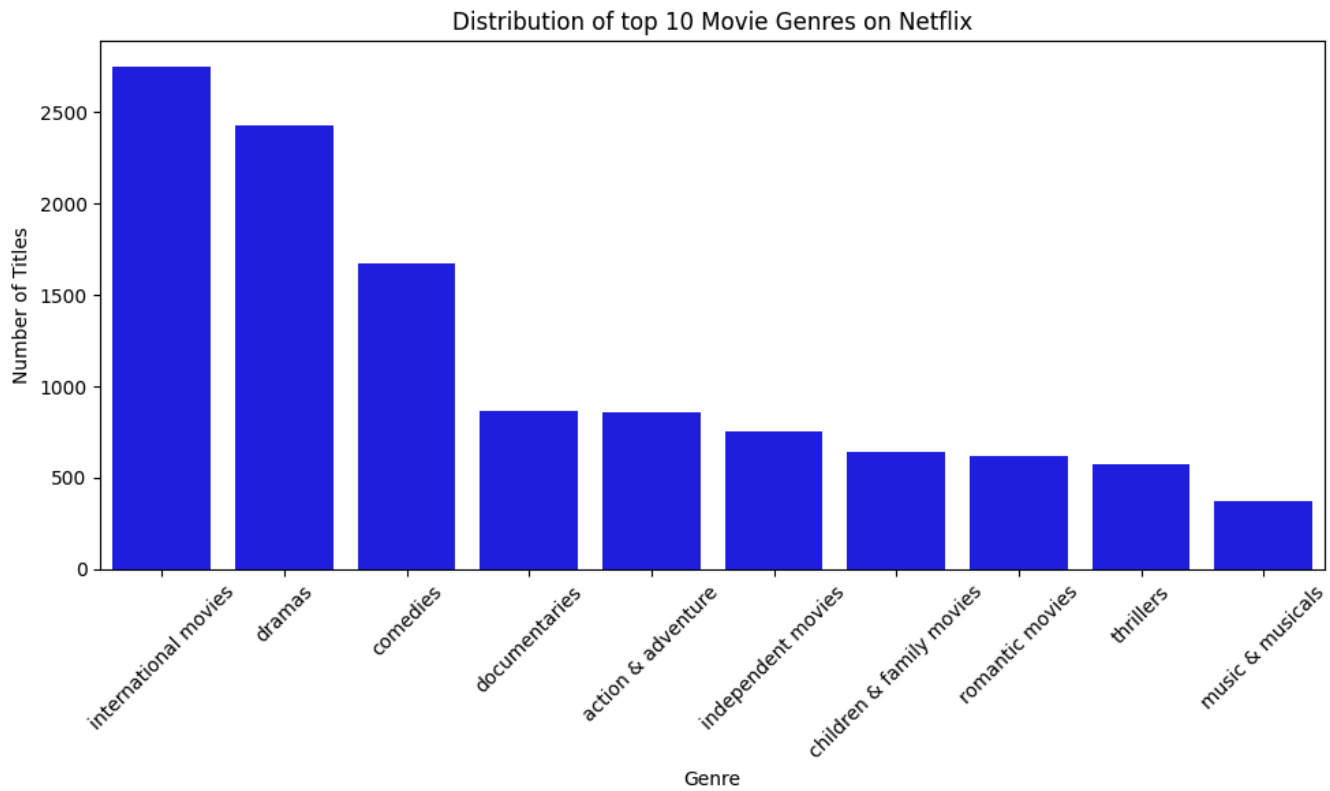
Regulatory and Cultural Factors: Consider potential regulatory implications and cultural sensitivities associated with mature content ratings.

Netflix's approach to content moderation and compliance with regional standards. Content Strategy Evolution: Predict how the prevalence of these ratings might evolve over time based on viewer demographics and market trends.

Shifts in audience preferences and global expansion strategies.

```
plt.figure(figsize=(10, 6))
sns.barplot(x=movie_listed_in_unique[2].head(10)['listed_in'],y=movie_listed_in_unique[2].head(10)['count'], color='blue')
plt.title('Distribution of top 10 Movie Genres on Netflix')
plt.xlabel('Genre')
plt.ylabel('Number of Titles')
```

```
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



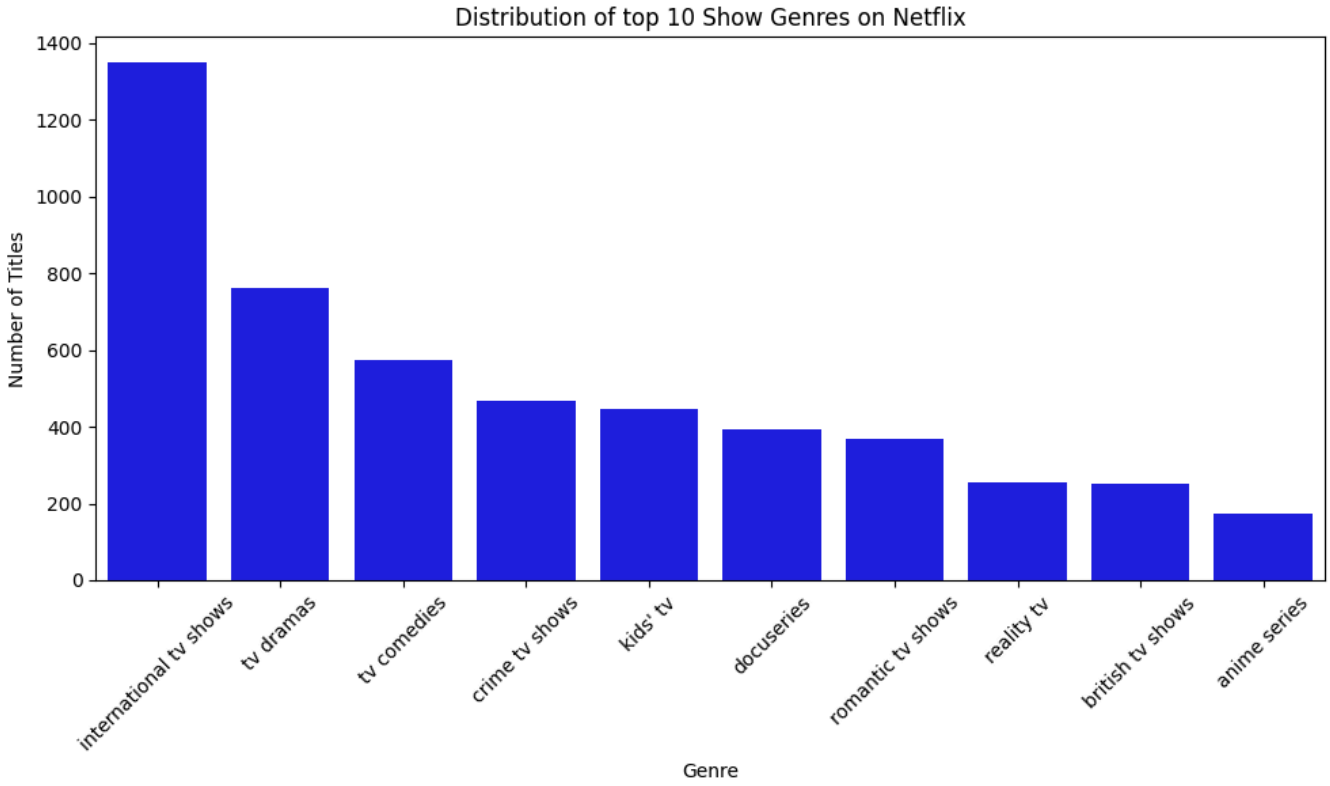
Here we can observe international movies,dramas,comedies are leading may be due to global appealing content

Recommendation

Continued Investment: Recommend continued investment in diverse international content to maintain and expand global audience engagement.

Enhanced Localization: Suggest enhancements in localization strategies (subtitles, dubbing, regional partnerships) to further enhance global appeal and market penetration.

```
plt.figure(figsize=(10, 6))
sns.barplot(x=show_listed_in_unique[2].head(10)['listed_in'],y=show_listed_in_unique[2].head(10)['count'], color='blue')
plt.title('Distribution of top 10 Show Genres on Netflix')
plt.xlabel('Genre')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



Here we can observe international tv shows,tv dramas,tv comedies are leading may be dy eto global appealing content

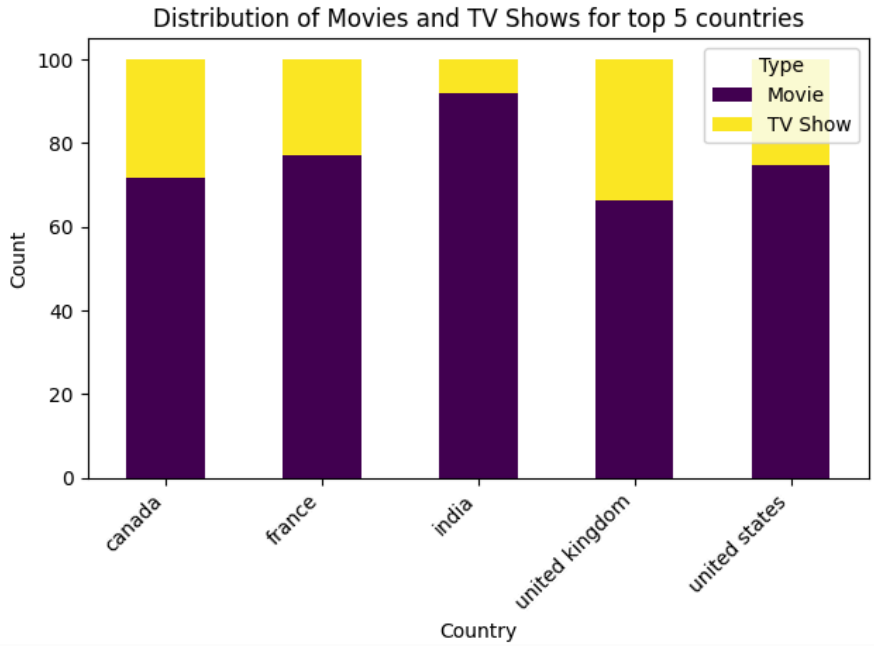
```
d=df.loc[df['country'].apply(lambda x:len(x)!=0)][['type','country']]
d=d.explode('country')
d=d.loc[d['country'].isin(country_unique[2].head()['country'])]
```

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

type	Movie	TV Show
country		
canada	319	126
france	303	90
india	962	84
united kingdom	534	271
united states	2749	932

```
plt.figure(figsize=(12, 8))
counts_percentage = counts.div(counts.sum(axis=1), axis=0) * 100
# Plot stacked bar chart
counts_percentage.plot(kind='bar', stacked=True,colormap='viridis')

plt.title('Distribution of Movies and TV Shows for top 5 countries')
plt.xlabel('Country')
plt.ylabel('Count')
plt.legend(title='Type', labels=['Movie', 'TV Show'])
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
<Figure size 1200x800 with 0 Axes>
```



Insight and Reason

If we observe the above plot there is good balance between shows and movies in countries like UK and US but in india it looks imbalanced

The structure of the Indian entertainment industry is such that TV shows are often produced for traditional television networks rather than streaming platforms. While this is changing with the rise of streaming services, the focus on movies remains strong.

Recommendation

Content Acquisition and Production:

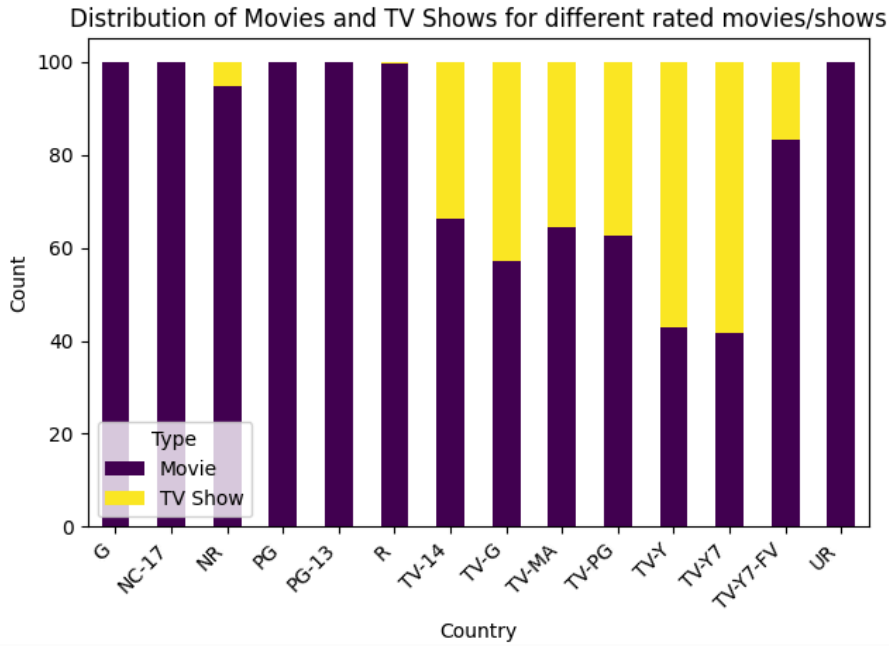
Recommend adjustments in Netflix's content strategy for India to achieve a more balanced offering. Suggest increasing investment in TV show production or acquiring more diverse TV content that resonates with Indian audiences.

User Experience and Engagement:

Propose enhancements to Netflix's recommendation algorithms to better cater to regional content preferences. Explore opportunities for promoting TV shows more effectively to balance the content mix in India.

```
d=df[['type','rating']]
counts = d.groupby(['rating', 'type']).size().unstack(fill_value=0)
plt.figure(figsize=(14, 8))
counts_percentage = counts.div(counts.sum(axis=1), axis=0) * 100
# Plot stacked bar chart
counts_percentage.plot(kind='bar', stacked=True,colormap='viridis')

plt.title('Distribution of Movies and TV Shows for different rated movies/shows')
plt.xlabel('Country')
plt.ylabel('Count')
plt.legend(title='Type', labels=['Movie', 'TV Show'])
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

```
<Figure size 1400x800 with 0 Axes>
```



Insight and Reason

If we observe some ratings like R,G and PG are completely dominated by movies while in some ratings like Tv-Y and TV-Y7 Tv shows are dominating

R-Rated Movies: Often include mature content, such as intense violence, strong language, and adult themes. Such content is more commonly explored in movies than in TV shows, given the shorter format and more significant impact that can be delivered in a film. G and PG Movies: These are family-friendly or child-friendly movies that are designed to be suitable for all or most audiences. Animated movies, family films, and children's films often fall into these categories, and movies have traditionally been a significant medium for delivering such content.

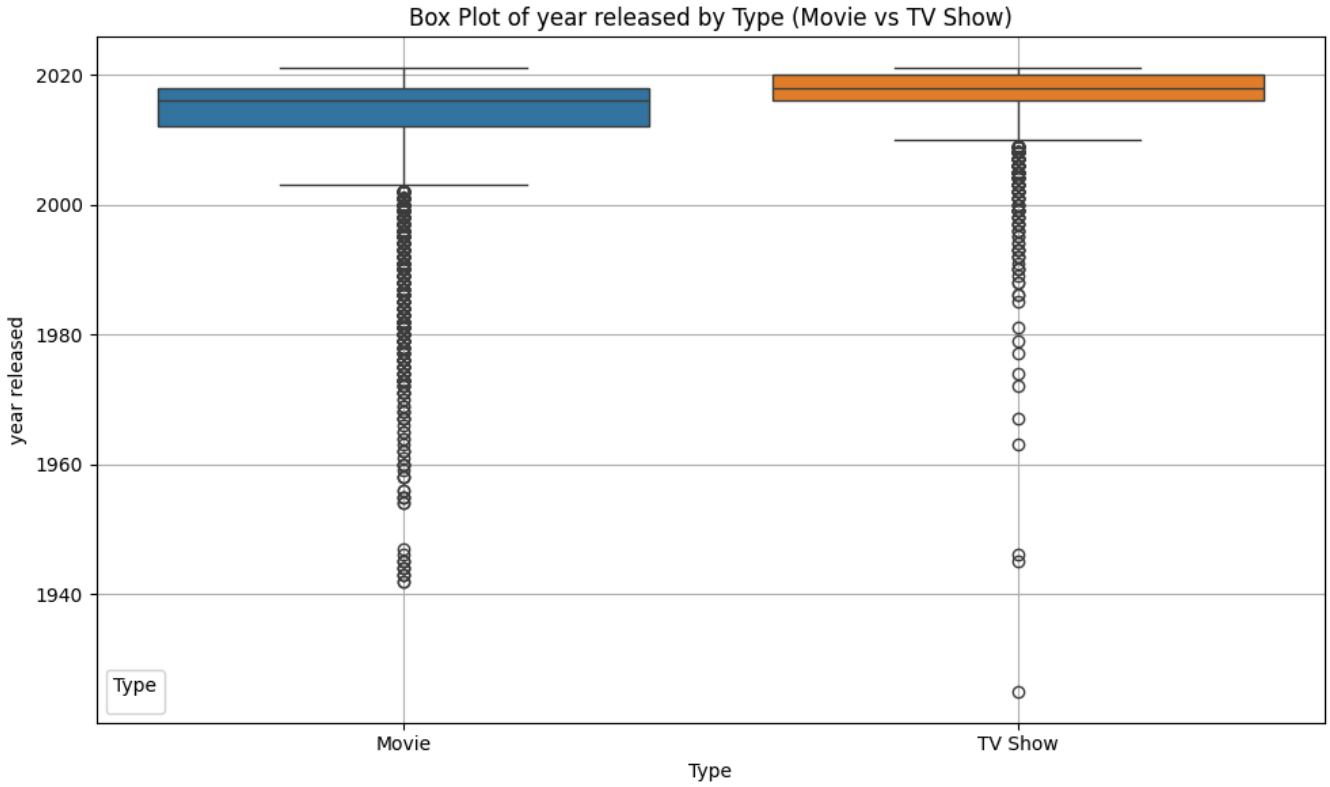
TV-Y (All Children) and TV-Y7 (Directed to Older Children) ratings are designed specifically for children. These ratings ensure that content is appropriate for young viewers, including educational shows, cartoons, and other children's programming.

Recommendation

Compare Netflix's distribution with other streaming platforms or traditional media. How do competitors like Disney+ or HBO Max handle content distribution across different ratings? What does this insight reveal about evolving viewer preferences and consumption habits in the streaming landscape?

```
d=df[['type','release_year']]

plt.figure(figsize=(10, 6))

# Create a box plot with hue using seaborn
sns.boxplot(x='type', y='release_year', data=d, hue='type')

# Formatting
plt.title('Box Plot of year released by Type (Movie vs TV Show)')
plt.xlabel('Type')
plt.ylabel('year released')
plt.legend(title='Type', loc='best')  # Adjust legend location if needed
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an unders
```



Insight and Reason If we observe the above plot average release year for TV show is more than movies.This can be due to

TV Licensing: Licensing agreements for TV shows often focus on recent seasons and series, as networks and producers aim to capitalize on current viewer interest. This results in a higher proportion of recent TV shows being available on streaming platforms.

Movie Licensing: Movie licensing deals can include a mix of recent blockbusters, older classics, and everything in between. This diversity in licensing agreements impacts the average release year for movies.

Recommendation

Predict how this trend might evolve in the future, considering shifts in content consumption habits and technological advancements:

Will the gap between average release years for movies and TV shows narrow or widen with the growth of streaming and on-demand viewing?

```
d=movies.explode('country')
d=d.loc[d['country'].isin(movie_country_unique[2].head(5)['country'])]
d=d.loc[d['rating'].isin(d['rating'].value_counts().head().index)]
```

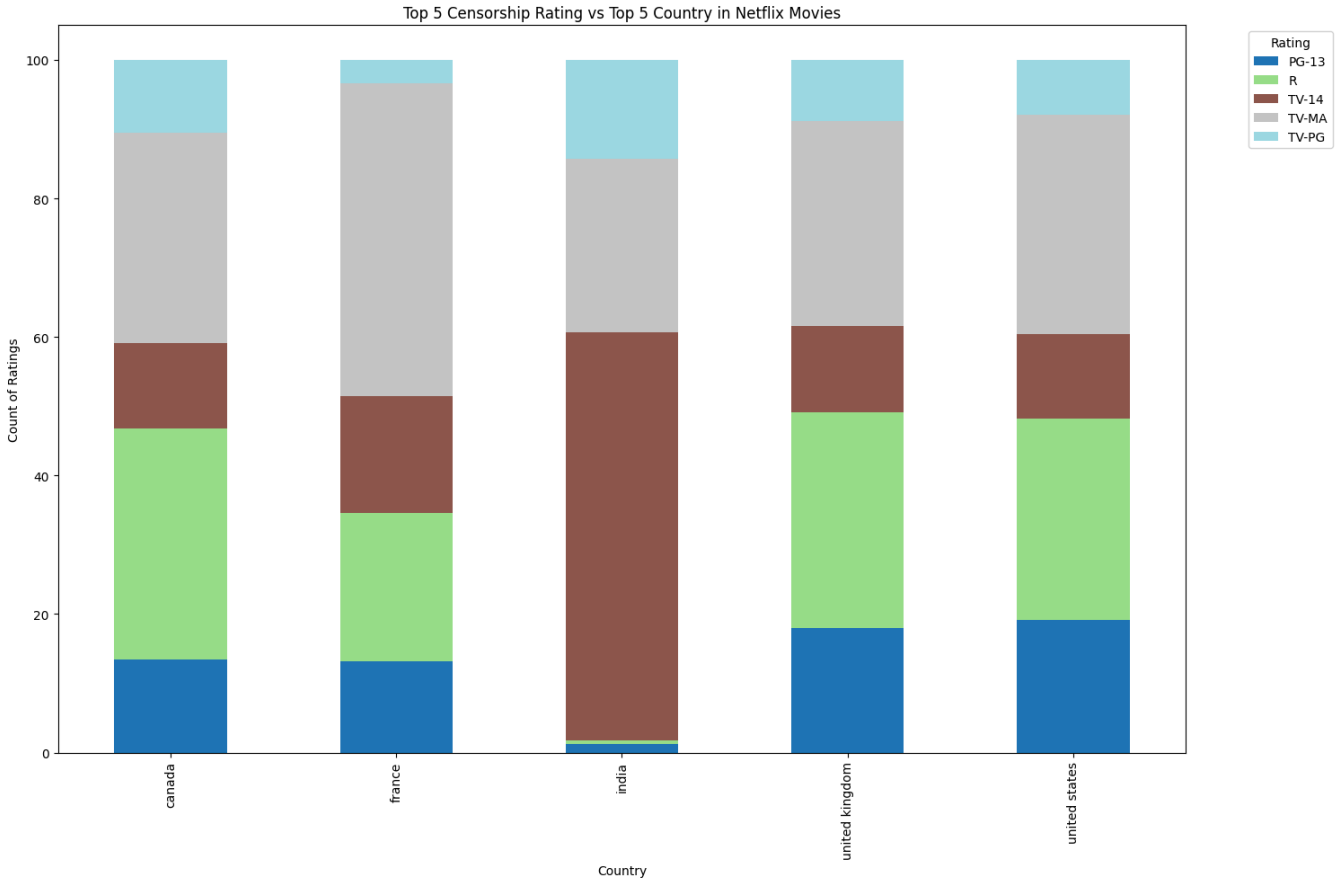
```
rating_counts = d.groupby(['country', 'rating']).size().unstack(fill_value=0)
rating_counts
```

rating	PG-13	R	TV-14	TV-MA	TV-PG
country					
canada	32	79	29	72	25
france	35	57	45	120	9
india	11	5	547	232	133
united kingdom	84	145	58	138	41
united states	433	660	276	719	180

```
plt.figure(figsize=(12, 8))
counts_percentage = (rating_counts.T.div(rating_counts.sum(axis=1))*100).T

counts_percentage.plot(kind='bar', stacked=True, colormap='tab20', figsize=(15, 10))
plt.title('Top 5 Censorship Rating vs Top 5 Country in Netflix Movies')
plt.xlabel('Country')
```

```
plt.ylabel('Count of Ratings')
plt.legend(title='Rating', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```

```
<Figure size 1200x800 with 0 Axes>
```



Insight and Reason

If we observe above plots most of movies produced by india are in TV-14 and TV-MA rated while countries US,UK,Canada had good proporation of R rated movies

Recommendations

Provide actionable recommendations based on the insights gained: Consider expanding content partnerships or production efforts in regions with high demand for specific ratings.

Enhance content diversity strategies to cater to varied audience preferences across different markets.

Continuously monitor and adapt content offerings based on evolving viewer behavior and regulatory landscapes.

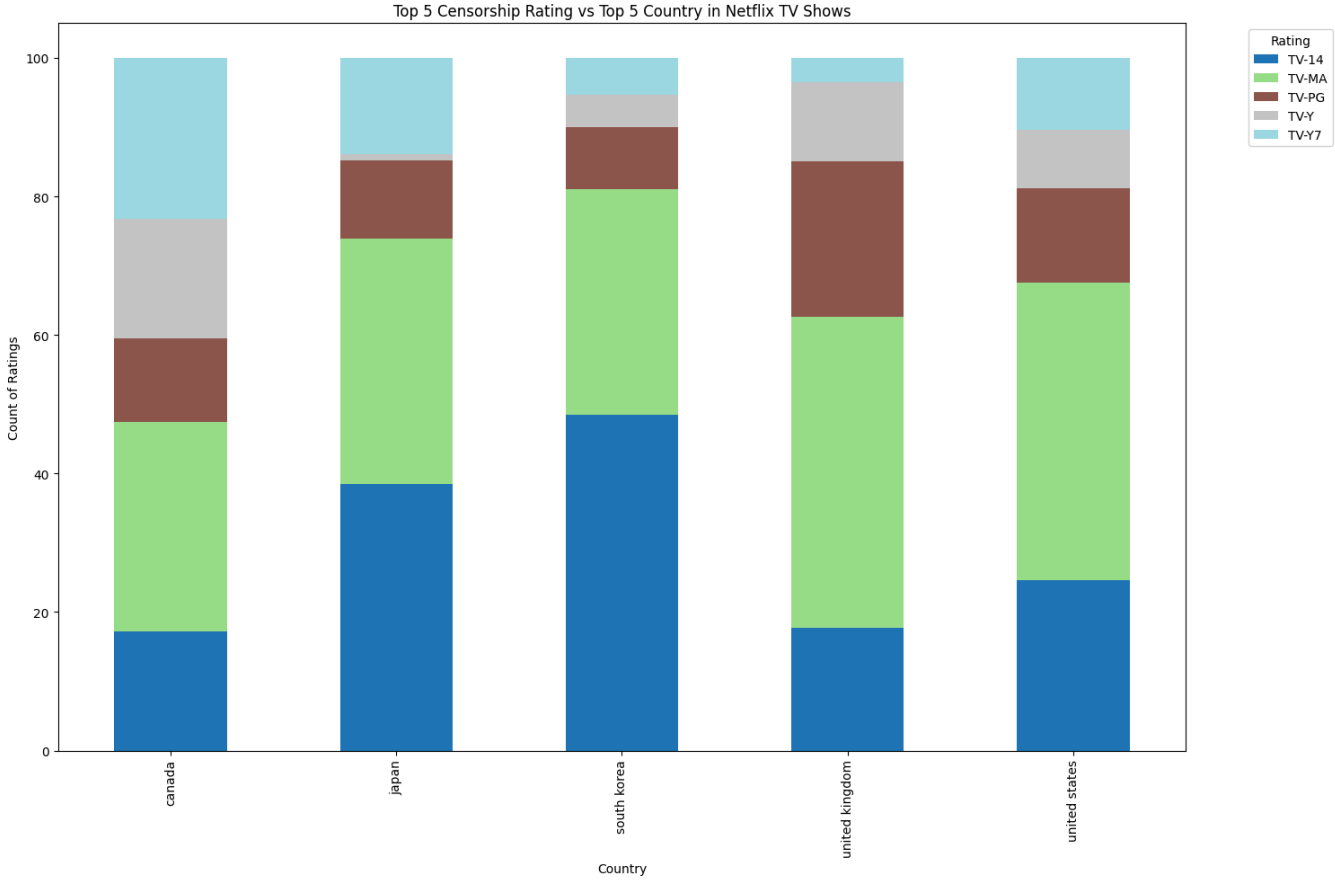
```
d=shows.explode('country')
d=d.loc[d['country'].isin(show_country_unique[2].head(5)['country'])]
d=d.loc[d['rating'].isin(d['rating'].value_counts().head().index)]
```

```
rating_counts = d.groupby(['country', 'rating']).size().unstack(fill_value=0)
rating_counts
```

rating	TV-14	TV-MA	TV-PG	TV-Y	TV-Y7
country					
canada	20	35	14	20	27
japan	75	69	22	2	27
south korea	82	55	15	8	9
united kingdom	45	114	57	29	9
united states	219	381	122	75	92

```
plt.figure(figsize=(12, 8))
counts_percentage = (rating_counts.T.div(rating_counts.sum(axis=1))*100).T

counts_percentage.plot(kind='bar', stacked=True, colormap='tab20', figsize=(15, 10))
plt.title('Top 5 Censorship Rating vs Top 5 Country in Netflix TV Shows')
plt.xlabel('Country')
plt.ylabel('Count of Ratings')
plt.legend(title='Rating', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```

```
<Figure size 1200x800 with 0 Axes>
```



Insight and Reason

If we observe above plots most of shows produced by south korea are in TV-14 and TV-MA rated while countries UKa had good proporation of PG rated shows

Recommendations

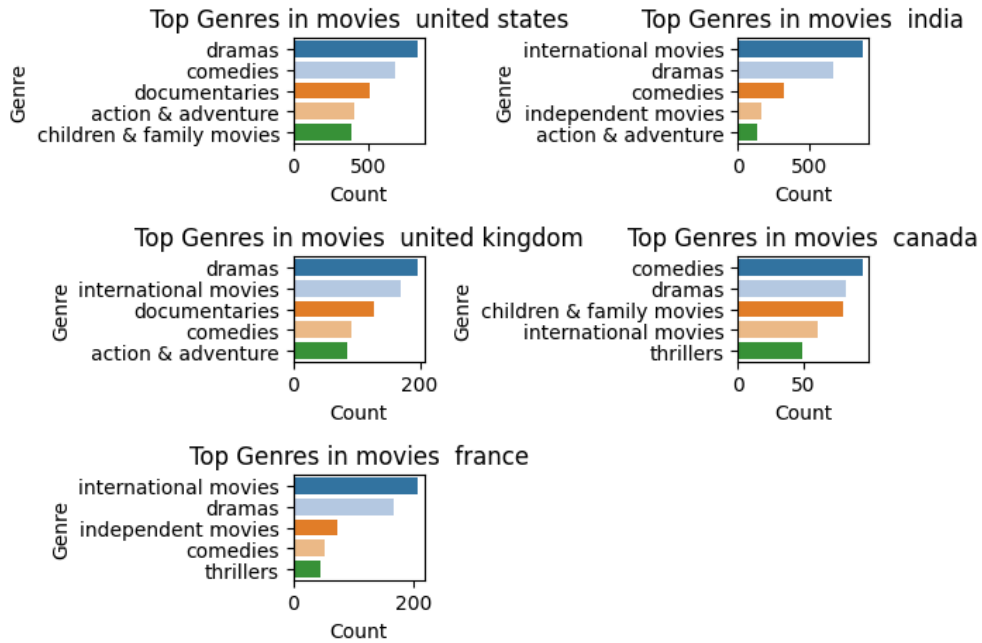
Predict how these rating distributions might evolve over time based on industry trends and changing viewer preferences.

Will South Korean productions expand into different rating categories as their global reach grows?

How might the UK's content strategy adapt to changing viewer expectations and regulatory standards?

```
d=movies.explode('country')
for i, country in enumerate(movie_country_unique[2]['country'].head()):
    plt.subplot(3, 2, i+1)
    a=d[d['country']==country]
    a=a.explode(['listed_in'])
    country_data = a['listed_in'].value_counts().head().reset_index()
    sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
    plt.title(f'Top Genres in movies  {country}')
    plt.xlabel('Count')
    plt.ylabel('Genre')

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

```
<ipython-input-92-ed16ca63eb6e>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-92-ed16ca63eb6e>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-92-ed16ca63eb6e>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-92-ed16ca63eb6e>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-92-ed16ca63eb6e>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
```



Insights

If we observe above plots dramas is dominating in US and UK,international movies dominating in India and France ,comedies in canada

Recommendations

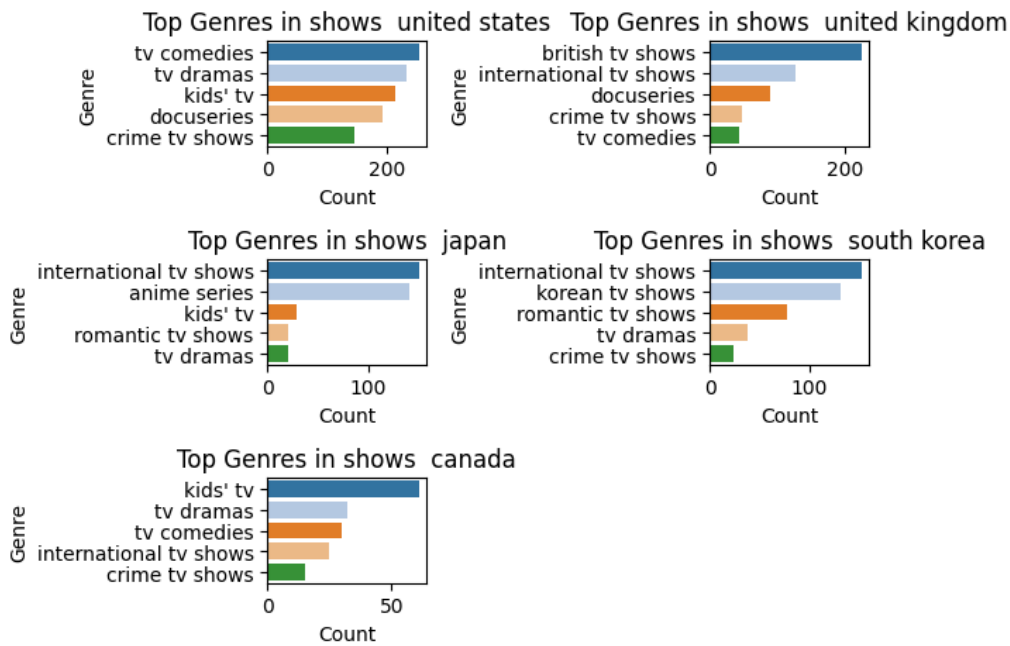
Predict how these genre preferences might evolve over time and how Netflix should adapt.

Will there be shifts in genre popularity based on global trends or socio-economic factors?

How can Netflix innovate to maintain or expand its audience base in each region?

```
d=shows.explode('country')
for i, country in enumerate(show_country_unique[2]['country'].head()):
    plt.subplot(3, 2, i+1)
    a=d[d['country']==country]
    a=a.explode(['listed_in'])
    country_data = a['listed_in'].value_counts().head().reset_index()
    sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
    plt.title(f'Top Genres in shows  {country}')
    plt.xlabel('Count')
    plt.ylabel('Genre')

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

```
<ipython-input-93-33c3a187e2a1>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-93-33c3a187e2a1>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-93-33c3a187e2a1>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-93-33c3a187e2a1>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
<ipython-input-93-33c3a187e2a1>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and

  sns.barplot(x='count', y='listed_in', data=country_data, palette='tab20')
```



Insights

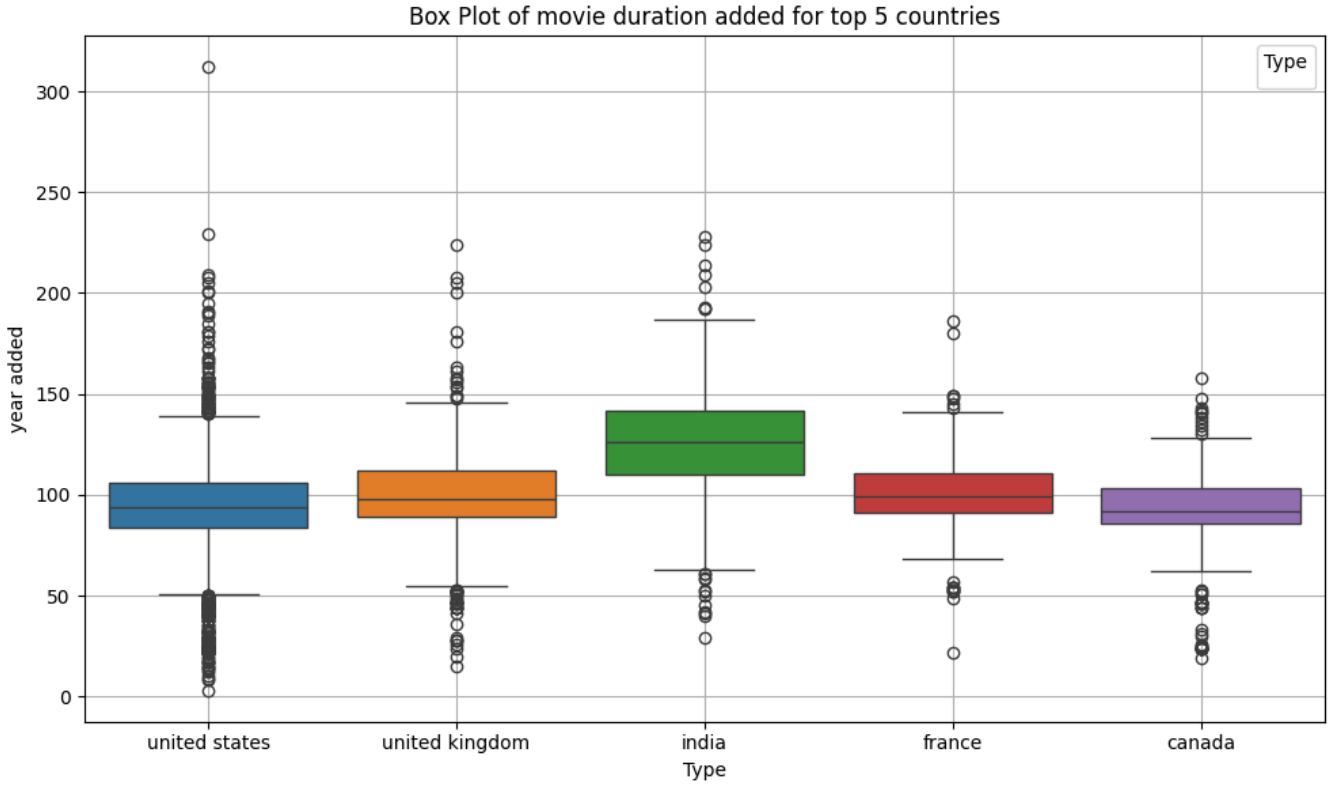
If we observe above plots interantional tv shows are dominating in japan and south korea,kids TV in canada,TV comedies in US,British TV Shows in UK

Recommendations

Competitive Landscape: Compare Netflix's dominance in specific genres across regions with competitors or traditional broadcasters. Identify strengths and potential areas for growth based on regional content preferences.

```
d=movies[['country','duration_in_minutes']]
d=d.explode('country')
d=d.loc[d['country'].isin(movie_country_unique[2].head()['country'])]
plt.figure(figsize=(10, 6))

# Create a box plot with hue using seaborn
sns.boxplot(x='country', y='duration_in_minutes', data=d, hue='country')

# Formatting
plt.title('Box Plot of movie duration added for top 5 countries')
plt.xlabel('Type')
plt.ylabel('year added')
plt.legend(title='Type', loc='best')  # Adjust legend location if needed
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an unders
```



Insight

If we observe the above plot average duration of movie is more in india compared to countries like US,UK,France,Canada

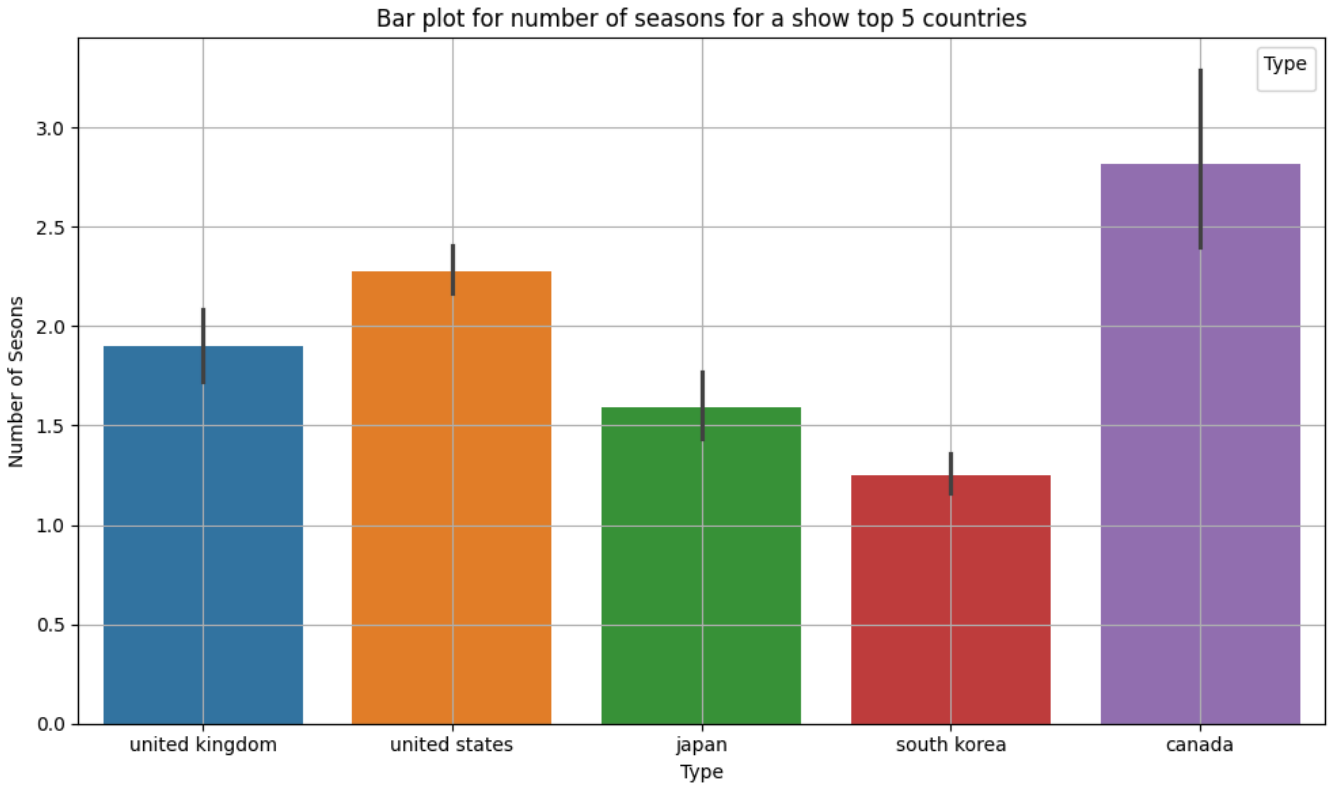
Outlier

If we observe the above plot some movies have very less duration less than 50 minutes which has to be considered as outliers

```
d=shows[['country','Number of Seasons']]
d=d.explode('country')
d=d.loc[d['country'].isin(show_country_unique[2].head()['country'])]
plt.figure(figsize=(10, 6))

# Create a box plot with hue using seaborn
sns.barplot(x='country', y='Number of Seasons', data=d, hue='country')
```

```
# Formatting
plt.title('Bar plot for number of seasons for a show top 5 countries')
plt.xlabel('Type')
plt.ylabel('Number of Sesons')
plt.legend(title='Type', loc='best')  # Adjust legend location if needed
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an unders
```

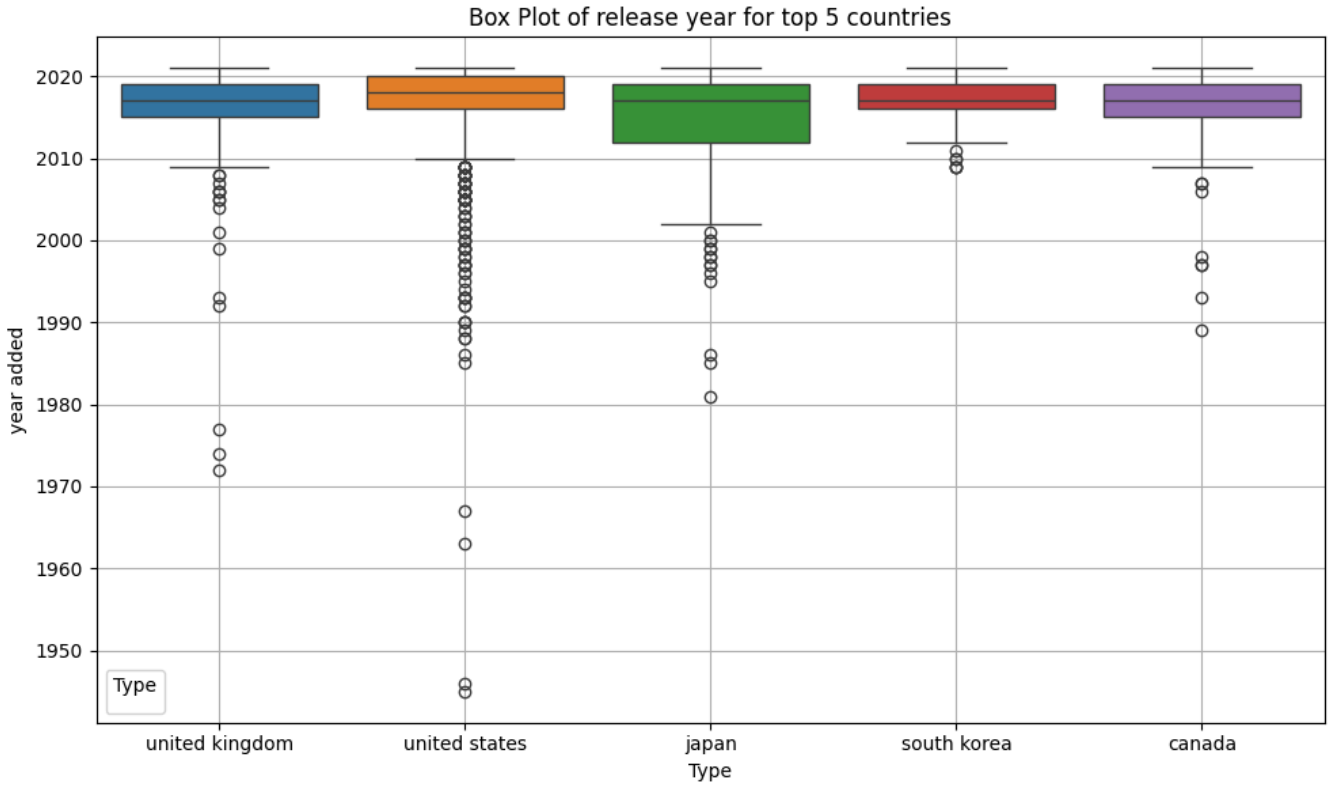


Insight

If we observe the above plot average number of seasons of TV Show is more in canada compared to countries like US,UK,japan,soth korea

```
d=shows[['country','release_year']]
d=d.explode('country')
d=d.loc[d['country'].isin(show_country_unique[2].head()['country'])]
plt.figure(figsize=(10, 6))

# Create a box plot with hue using seaborn
sns.boxplot(x='country', y='release_year', data=d, hue='country')

# Formatting
plt.title('Box Plot of release year for top 5 countries')
plt.xlabel('Type')
plt.ylabel('year added')
plt.legend(title='Type', loc='best')  # Adjust legend location if needed
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an unders
```



Insight

If we observe above box plot japan has more old TV shows compared to UK,US,south korea,canada

Recommendation

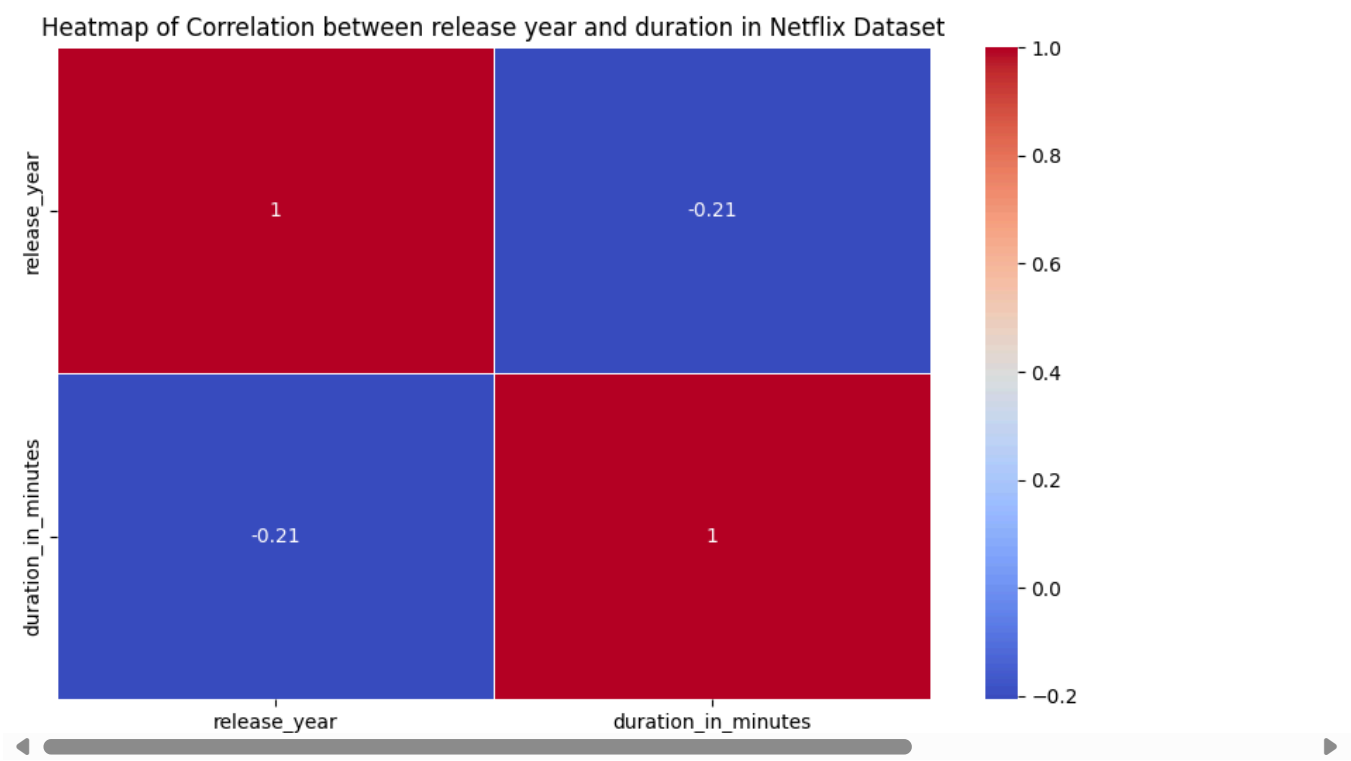
Discuss potential reasons for this disparity: Cultural Influence: Explore how cultural factors in Japan might prioritize preservation and accessibility of older TV shows.

Content Licensing and Acquisition: Analyze how Netflix's content acquisition strategies differ across regions, potentially affecting the availability of newer versus older content.

Outlier

If we observe above plot there are some movies which are released before 2000 are aquired by netflix can be outliers

```
correlation_matrix = movies[['release_year','duration_in_minutes']].corr()

# Create a heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=.5)
plt.title('Heatmap of Correlation between release year and duration in Netflix Dataset')
plt.show()
```



Insight

If we observe corrleation matrix it seems like there is negative corrleation between year released and duration of movie which can be intepreted as now movies are coming with shorter duration for better quality

```
d=movies[['rating','release_year']]
d=d.loc[d['rating'].isin(d['rating'].value_counts().head().index)]
plt.figure(figsize=(10, 6))

# Create a box plot with hue using seaborn
sns.boxplot(x='rating', y='release_year', data=d, hue='rating')

# Formatting
plt.title('Box Plot of release year for top 5 rated movies')
plt.xlabel('Type')
plt.ylabel('year added')
plt.legend(title='Type', loc='best')  # Adjust legend location if needed
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an unders
```

Box Plot of release year for top 5 rated movies