

Netflix Data Analysis

Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally.

The study involves analysing the data and generating insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries.

- Understanding what content is available in different countries
- Network analysis of Actors / Directors and find interesting insights
- Does Netflix has more focus on TV Shows than movies in recent years. etc.

The dataset for this task consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

Data Source: <https://www.kaggle.com/datasets/shivamb/netflix-shows>

```
In [80]: #importing libraries for our purpose
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv('/kaggle/input/netflix/Netflix.csv')
df.head(2)
```

```
Out[80]:
```

	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, 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	After crossing paths at a party, a Cape Town t...

```
In [81]: #length of data
len(df)
```

```
Out[81]: 8807
```

```
In [82]: #checking datatypes
df.dtypes
```

```
Out[82]:
```

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

```
In [83]: #number of unique values in our data
for i in df.columns:
    print(i,':',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
```

```
In [84]: #checking null values in every column of our data
df.isnull().sum()
```

```
Out[84]: 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
```

```
In [85]: #checking the occurrences of each of the ratings
df['rating'].value_counts()
```

```
Out[85]: rating
TV-MA      3207
TV-14      2160
TV-PG      863
R          799
PG-13      490
TV-Y7      334
TV-Y       307
PG         287
TV-G       220
NR         80
G          41
TV-Y7-FV    6
NC-17      3
UR         3
74 min     1
84 min     1
66 min     1
Name: count, dtype: int64
```

```
In [86]: #unnesting the directors column, i.e- creating separate lines for each director in a movie
constraint1=df['director'].apply(lambda x: str(x).split(', ')).tolist()
df_new1=pd.DataFrame(constraint1,index=df['title'])
df_new1=df_new1.stack()
df_new1=pd.DataFrame(df_new1.reset_index())
df_new1.rename(columns={0:'Directors'},inplace=True)
df_new1.drop(['level_1'],axis=1,inplace=True)
df_new1.head(3)
```

```
Out[86]:      title    Directors
0 Dick Johnson Is Dead  Kirsten Johnson
1 Blood & Water        nan
2 Ganglands            Julien Leclercq
```

```
In [87]: #unnesting the cast column, i.e- creating separate lines for each cast member in a movie
constraint2=df['cast'].apply(lambda x: str(x).split(', ')).tolist()
df_new2=pd.DataFrame(constraint2,index=df['title'])
df_new2=df_new2.stack()
df_new2=pd.DataFrame(df_new2.reset_index())
df_new2.rename(columns={0:'Actors'},inplace=True)
df_new2.drop(['level_1'],axis=1,inplace=True)
df_new2.head(3)
```

```
Out[87]:      title    Actors
0 Dick Johnson Is Dead      nan
1 Blood & Water   Ama Qamata
2 Blood & Water   Khosi Ngema
```

```
In [88]: #unnesting the listed_in column, i.e- creating separate lines for each genre in a movie
constraint3=df['listed_in'].apply(lambda x: str(x).split(', ')).tolist()
df_new3=pd.DataFrame(constraint3,index=df['title'])
df_new3=df_new3.stack()
df_new3=pd.DataFrame(df_new3.reset_index())
df_new3.rename(columns={0:'Genre'},inplace=True)
df_new3.drop(['level_1'],axis=1,inplace=True)
df_new3.head(3)
```

```
Out[88]:      title    Genre
0 Dick Johnson Is Dead Documentaries
1 Blood & Water  International TV Shows
2 Blood & Water    TV Dramas
```

```
In [89]: #unnesting the country column, i.e- creating separate lines for each country in a movie
```

```

constraint4=df['country'].apply(lambda x: str(x).split(', ')).tolist()
df_new4=pd.DataFrame(constraint4,index=df['title'])
df_new4=df_new4.stack()
df_new4=pd.DataFrame(df_new4.reset_index())
df_new4.rename(columns={0:'country'},inplace=True)
df_new4.drop(['level_1'],axis=1,inplace=True)
df_new4.head(3)

```

Out[89]:

	title	country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	nan

In [90]:

```

#merging the unnested director data with unnested actors data
df_new5=df_new2.merge(df_new1,on=['title'],how='inner')
#merging the above merged data with unnested genre data
df_new6=df_new5.merge(df_new3,on=['title'],how='inner')
#merging the above merged data with unnested country data
df_new=df_new6.merge(df_new4,on=['title'],how='inner')
#replacing nan values of director and actor by Unknown Actor and Director
df_new['Actors'].replace(['nan'],['Unknown Actor'],inplace=True)
df_new['Directors'].replace(['nan'],['Unknown Director'],inplace=True)
df_new['country'].replace(['nan'],[np.nan],inplace=True)
df_new.head(3)

```

Out[90]:

	title	Actors	Directors	Genre	country
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa

In [91]:

```

#merging our unnested data with the original data
df_final=df_new.merge(df[['show_id', 'type', 'title', 'date_added',
                           'release_year', 'rating', 'duration']],on=['title'],how='left')
df_final.head(3)

```

Out[91]:

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

In [92]:

```
#now checking nulls
df_final.isnull().sum()
```

Out[92]:

title	0
Actors	0
Directors	0
Genre	0
country	11897
show_id	0
type	0
date_added	158
release_year	0
rating	67
duration	3
dtype: int64	

In duration column, it was observed that the nulls had values which were written in corresponding ratings column, i.e- you can't expect ratings to be in min. So the duration column nulls are replaced by corresponding values in ratings column

In [93]:

```

df_final.loc[df_final['duration'].isnull(),'duration']=df_final.loc[df_final['duration'].isnull(),'duration'].ffill()
df_final.loc[df_final['rating'].str.contains('min', na=False),'rating']='NR'
df_final.isnull().sum()

```

Out[93]:

title	0
Actors	0
Directors	0
Genre	0
country	11897
show_id	0
type	0
date_added	158
release_year	0
rating	67
duration	0
dtype: int64	

In [94]:

```
#Ratings can't be in min, so it has been made NR(i.e- Non Rated)
```

```
df_final.loc[df_final['rating'].str.contains('min', na=False), 'rating'] = 'NR'
df_final['rating'].fillna('NR', inplace=True)
pd.set_option('display.max_rows', None)
```

In [95]: #just an attempt to observe nulls in date_added column
df_final[df_final['date_added'].isnull()].head(3)

Out[95]:

		title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
136893		A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	British TV Shows	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons
136894		A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	TV Comedies	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons
136895		A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	TV Dramas	United Kingdom	s6067	TV Show	NaN	2013	TV-MA	2 Seasons

In [96]: #date added column is imputed on the basis of release year,i.e- suppose there's a null for date_added #when release year was 2013.So below piece of code just checks the mode of date added for release year=2013 # and imputes in place of nulls the corresponding mode

```
for i in df_final[df_final['date_added'].isnull()]['release_year'].unique():
    imp=df_final[df_final['release_year']==i]['date_added'].mode().values[0]
    df_final.loc[df_final['release_year']==i,'date_added']=df_final.loc[df_final['release_year']==i,'date_added']
```

In [97]: #country column is imputed on the basis of director,i.e- suppose there's a null for country #when we have a director whose other movies have a country given.So below piece of code just checks the mode of #country for the director # and imputes in place of nulls the corresponding mode

```
for i in df_final[df_final['country'].isnull()]['Directors'].unique():
    if i in df_final[~df_final['country'].isnull()]['Directors'].unique():
        imp=df_final[df_final['Directors']==i]['country'].mode().values[0]
        df_final.loc[df_final['Directors']==i,'country']=df_final.loc[df_final['Directors']==i,'country'].fillna(im)
```

So we imputed the country column on the basis of directors whose other movie titles had countries given. But there might be directors who have only one occurrence in our data. In that scenario, I have used Actors as a basis. i.e- for this Actor majorly acts in movies of which country? Imputation has been done on this basis. For remaining rows, country has been filled as Unknown Country

In [98]:

```
for i in df_final[df_final['country'].isnull()]['Actors'].unique():
    if i in df_final[~df_final['country'].isnull()]['Actors'].unique():
        imp=df_final[df_final['Actors']==i]['country'].mode().values[0]
        df_final.loc[df_final['Actors']==i,'country']=df_final.loc[df_final['Actors']==i,'country'].fillna(im)
#If there are still nulls, I just replace it by Unknown Country
df_final['country'].fillna('Unknown Country',inplace=True)
df_final.isnull().sum()
```

Out[98]:

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
	0	0	0	0	0	0	0	0	0	0	0
	Actors	0	0	0	0	0	0	0	0	0	0
	Directors	0	0	0	0	0	0	0	0	0	0
	Genre	0	0	0	0	0	0	0	0	0	0
	country	0	0	0	0	0	0	0	0	0	0
	show_id	0	0	0	0	0	0	0	0	0	0
	type	0	0	0	0	0	0	0	0	0	0
	date_added	0	0	0	0	0	0	0	0	0	0
	release_year	0	0	0	0	0	0	0	0	0	0
	rating	0	0	0	0	0	0	0	0	0	0
	duration	0	0	0	0	0	0	0	0	0	0
	dtype: int64										

In [99]: df_final.head(2)

Out[99]:

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

In [100]:

```
print(len(df_final['duration'].value_counts()))
df_final['duration'].value_counts()[:5]
```

Out[100]:

duration	count
1 Season	35035
2 Seasons	9559
3 Seasons	5084
94 min	4343
106 min	4040

Name: count, dtype: int64

In [101]: #removing mins from data

```
df_final['duration']=df_final['duration'].str.replace(" min","");
df_final.head(2)
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

```
In [102]: df_final['duration'].unique()
```

```
Out[102]: array(['90', '2 Seasons', '1 Season', '91', '125', '9 Seasons', '104', '127', '4 Seasons', '67', '94', '5 Seasons', '161', '61', '166', '147', '103', '97', '106', '111', '3 Seasons', '110', '105', '96', '124', '116', '98', '23', '115', '122', '99', '88', '100', '6 Seasons', '102', '93', '95', '85', '83', '113', '13', '182', '48', '145', '87', '92', '80', '117', '128', '119', '143', '114', '118', '108', '63', '121', '142', '154', '120', '82', '109', '101', '86', '229', '76', '89', '156', '112', '107', '129', '135', '136', '165', '150', '133', '70', '84', '140', '78', '7 Seasons', '64', '59', '139', '69', '148', '189', '141', '130', '138', '81', '132', '10 Seasons', '123', '65', '68', '66', '62', '74', '131', '39', '46', '38', '8 Seasons', '17 Seasons', '126', '155', '159', '137', '12', '273', '36', '34', '77', '60', '49', '58', '72', '204', '212', '25', '73', '29', '47', '32', '35', '71', '149', '33', '15', '54', '224', '162', '37', '75', '79', '55', '158', '164', '173', '181', '185', '21', '24', '51', '151', '42', '22', '134', '177', '13 Seasons', '52', '14', '53', '8', '57', '28', '50', '9', '26', '45', '171', '27', '44', '146', '20', '157', '17', '203', '41', '30', '194', '15 Seasons', '233', '237', '230', '195', '253', '152', '190', '160', '208', '180', '144', '5', '174', '170', '192', '209', '187', '172', '16', '186', '11', '193', '176', '56', '169', '40', '10', '3', '168', '312', '153', '214', '31', '163', '19', '12 Seasons', '179', '11 Seasons', '43', '200', '196', '167', '178', '228', '18', '205', '201', '191'], dtype=object)
```

```
In [103]: df_final['duration_copy']=df_final['duration'].copy()
df_final=df_final.copy()
```

```
In [104]: df_final1.loc[df_final1['duration_copy'].str.contains('Season'), 'duration_copy']=0
df_final1['duration_copy']=df_final1['duration_copy'].astype('int')
df_final1.head(2)
```

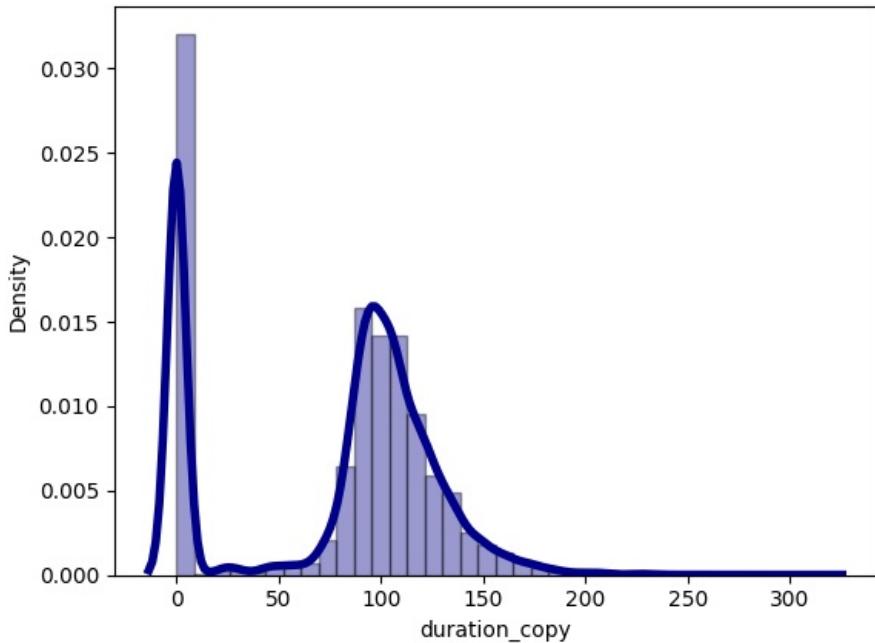
	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	duration_copy
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90	90
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	0

```
In [105]: df_final1['duration_copy'].describe()
```

```
Out[105]: count    201991.000000
mean      77.152789
std       52.269154
min       0.000000
25%      0.000000
50%      95.000000
75%     112.000000
max     312.000000
Name: duration_copy, dtype: float64
```

```
In [106]: import warnings
warnings.filterwarnings("ignore")
```

```
In [107]: import seaborn as sns
sns.distplot(df_final1['duration_copy'], hist=True, kde=True,
bins=int(36), color = 'darkblue',
hist_kws={'edgecolor':'black'},
kde_kws={'linewidth': 4})
plt.show()
```



```
In [108]: bins1 = [-1,1,50,80,100,120,150,200,315]
labels1 = ['<1','1-50','50-80','80-100','100-120','120-150','150-200','200-315']
df_final1['duration_copy'] = pd.cut(df_final1['duration_copy'],bins=bins1,labels=labels1)
df_final1.head(2)
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	duration_copy
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90	80-100
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	<1

```
In [109]: df_final1.loc[~df_final1['duration'].str.contains('Season')]['duration']=df_final1.loc[~df_final1['duration'].str.contains('Season')].drop(['duration_copy'],axis=1,inplace=True)
df_final1.head(2)
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	80-100
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

```
In [110]: print(len(df_final1['duration'].value_counts()))
df_final1['duration'].value_counts()[:5]
```

```
22
Out[110]: duration
80-100      52937
100-120     48724
1 Season    35035
120-150     26691
2 Seasons   9559
Name: count, dtype: int64
```

```
In [111]: from datetime import datetime
from dateutil.parser import parse
arr=[]
for i in df_final1['date_added'].values:
    dt1=parse(i)
    arr.append(dt1.strftime('%Y-%m-%d'))
df_final1['Modified_Added_date']=arr
df_final1['Modified_Added_date']=pd.to_datetime(df_final1['Modified_Added_date'])
df_final1['month_added']=df_final1['Modified_Added_date'].dt.month
df_final1['week_added']=df_final1['Modified_Added_date'].dt.strftime("%W")
df_final1['year']=df_final1['Modified_Added_date'].dt.year
df_final1.head(2)
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	80-100	2021-09-25
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	2021-09-24

```
In [112]: #Titles such as Bahubali(Hindi Version),Bahubali(Tamil Version) were there. Since it's only one movie in different presence of brackets and content between brackets is removed.
df_final['title']=df_final['title'].str.replace(r"\(.*)","",)
df_final.head(2)
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	80-100	2021-09-25
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	2021-09-24

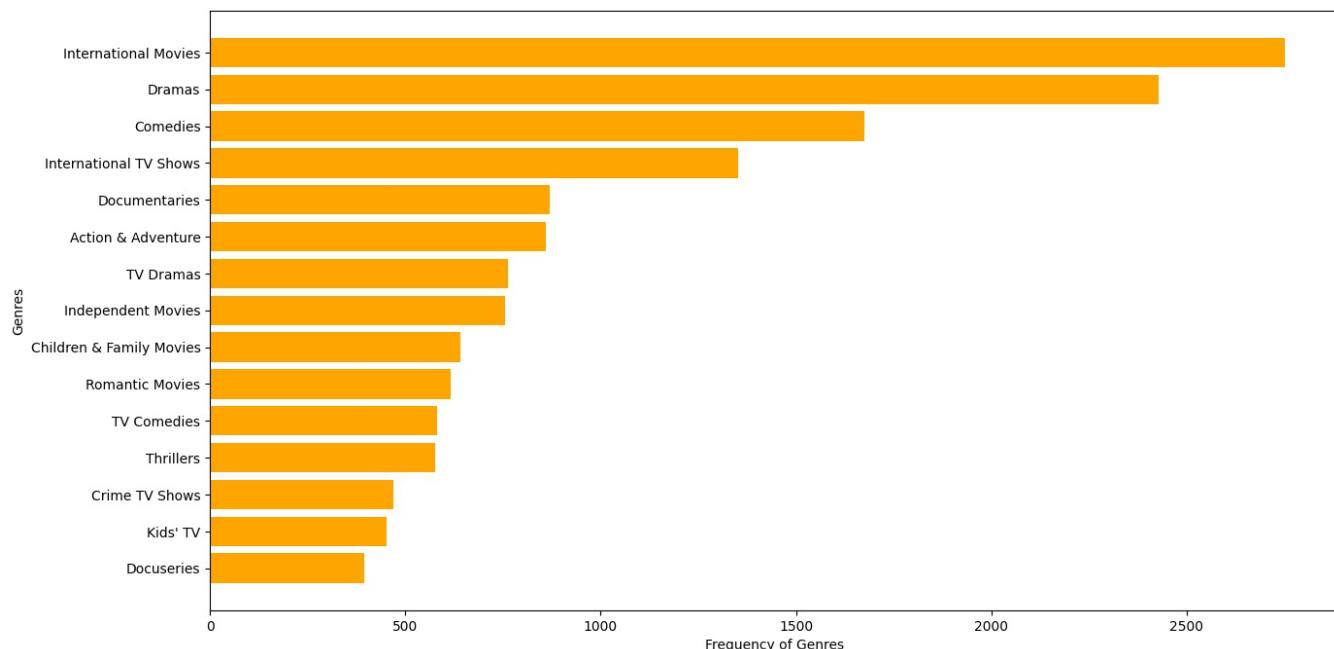
Univariate Analysis in terms of counts of each column

```
In [113]: #number of distinct titles on the basis of genre
print(len(df_final.groupby(['Genre']).agg({"title":"nunique"})))
df_final.groupby(['Genre']).agg({"title":"nunique"})[:5]
```

42

	title
Genre	
Action & Adventure	859
Anime Features	71
Anime Series	176
British TV Shows	253
Children & Family Movies	641

```
In [114]: df_genre=df_final.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barrh(df_genre[:::-1]['Genre'], df_genre[:::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```

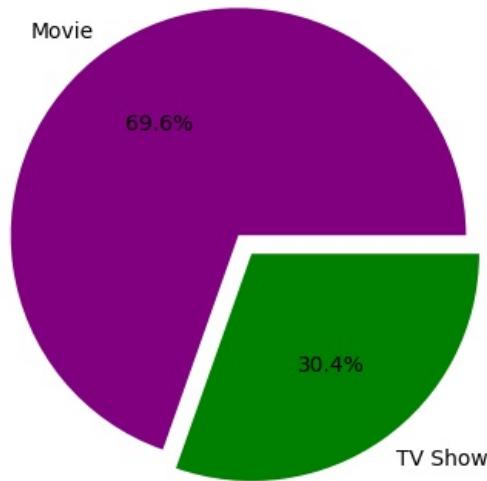


International Movies, Dramas and Comedies are the most popular .

```
In [115]: #number of distinct titles on the basis of type
df_final.groupby(['type']).agg({"title":"nunique"})
```

	title
type	
Movie	6131
TV Show	2676

```
In [116]: df_type=df_final.groupby(['type']).agg({"title":"nunique"}).reset_index()
plt.pie(df_type['title'],explode=(0.05,0.05), labels=df_type['type'],colors=['purple','green'], autopct='%.1f%%')
plt.show()
```



We have 70:30 ratio of Movies and TV Shows in our data

```
In [117]: #number of distinct titles on the basis of country
print(len(df_final1.groupby(['country']).agg({"title":"nunique"})))
df_final1.groupby(['country']).agg({"title":"nunique"})[:3]
```

128

```
Out[117]:      title
               country
-----|-----
               3
Afghanistan | 1
Albania     | 1
```

The above dataframe shows a flaw in which we are seeing countries, such as Cambodia and Cambodia, or United States and United States, are shown as different countries. They should have been same

```
In [118]: df_final1['country'] = df_final1['country'].str.replace(',', '')
df_final1.head(2)
```

```
Out[118]:   title  Actors  Directors  Genre  country  show_id  type  date_added  release_year  rating  duration  Modified_Added_date
0  Dick Johnson Is Dead  Unknown Actor  Kirsten Johnson  Documentaries  United States    s1  Movie  September 25, 2021  2020  PG-13  80-100  2021-09-25
1  Blood & Water  Ama Qamata  Unknown Director  International TV Shows  South Africa    s2  TV Show  September 24, 2021  2021  TV-MA  2 Seasons  2021-09-24
```

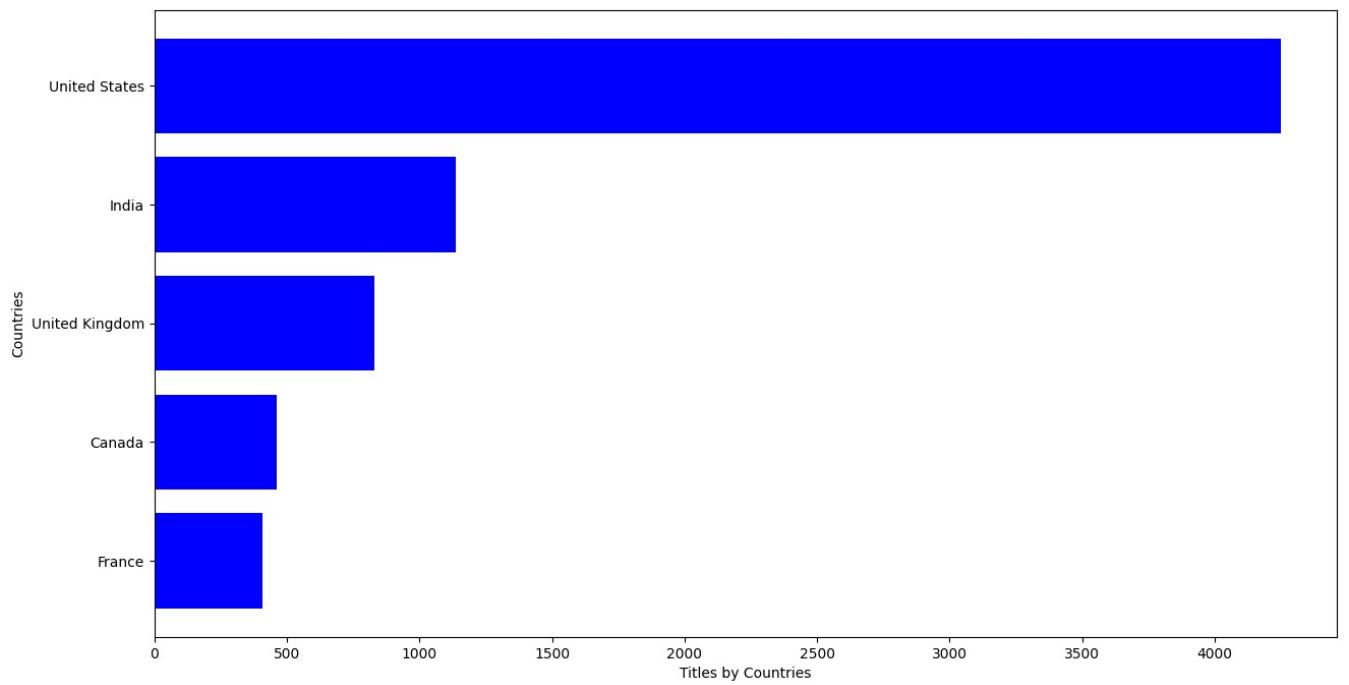
```
In [119]: #number of distinct titles on the basis of country
print(len(df_final1.groupby(['country']).agg({"title":"nunique"})))
df_final1.groupby(['country']).agg({"title":"nunique"})[:3]
```

124

```
Out[119]:      title
               country
-----|-----
               3
Afghanistan | 1
Albania     | 1
```

Now it is sorted.

```
In [120]: df_country=df_final1.groupby(['country']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_country[::-1]['country'], df_country[::-1]['title'], color=['blue'])
plt.xlabel('Titles by Countries')
plt.ylabel('Countries')
plt.show()
```

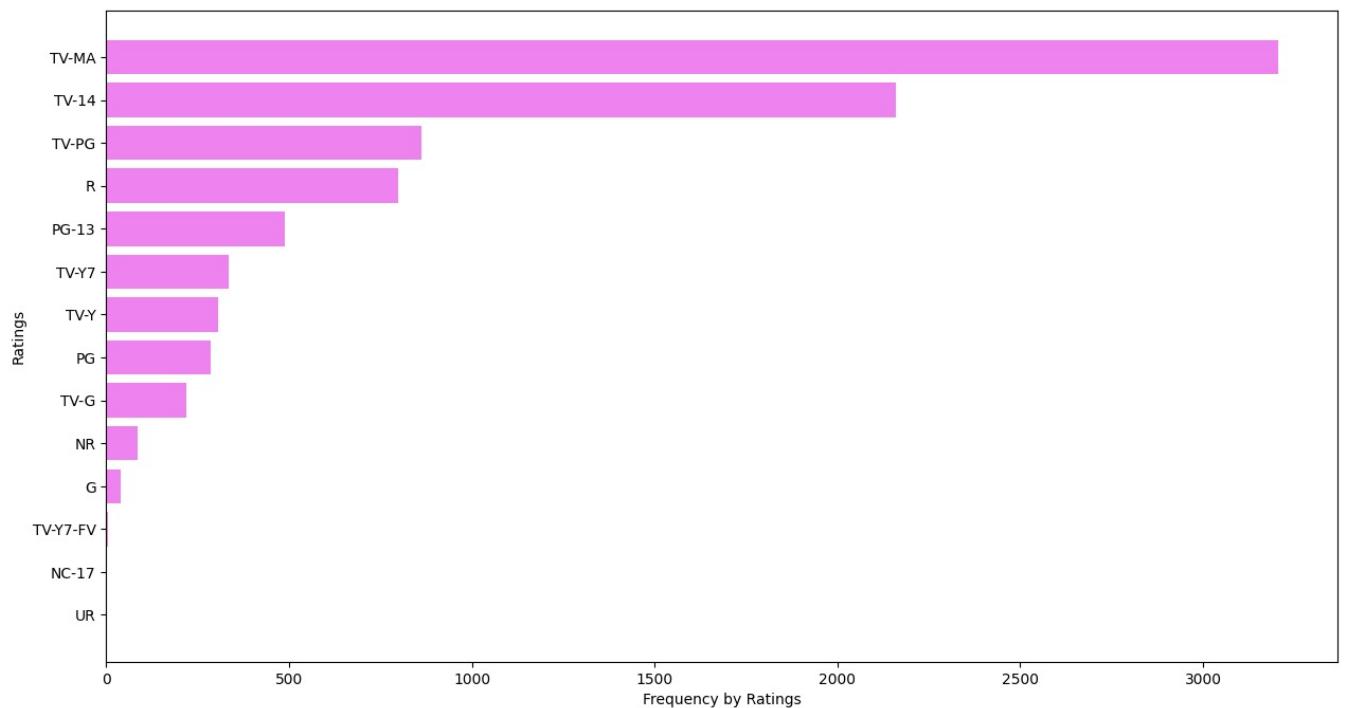


US, India, UK, Canada and France are leading countries in Content Creation on Netflix

```
In [121]: #number of distinct titles on the basis of rating
df_final1.groupby(['rating']).agg({"title":"nunique"})
```

```
Out[121]:      title
               rating
               G      41
               NC-17    3
               NR     87
               PG    287
               PG-13   490
               R     799
               TV-14  2160
               TV-G    220
               TV-MA  3207
               TV-PG   863
               TV-Y    307
               TV-Y7   334
               TV-Y7-FV    6
               UR      3
```

```
In [122]: df_rating=df_final1.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_rating[::-1]['rating'], df_rating[::-1]['title'], color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



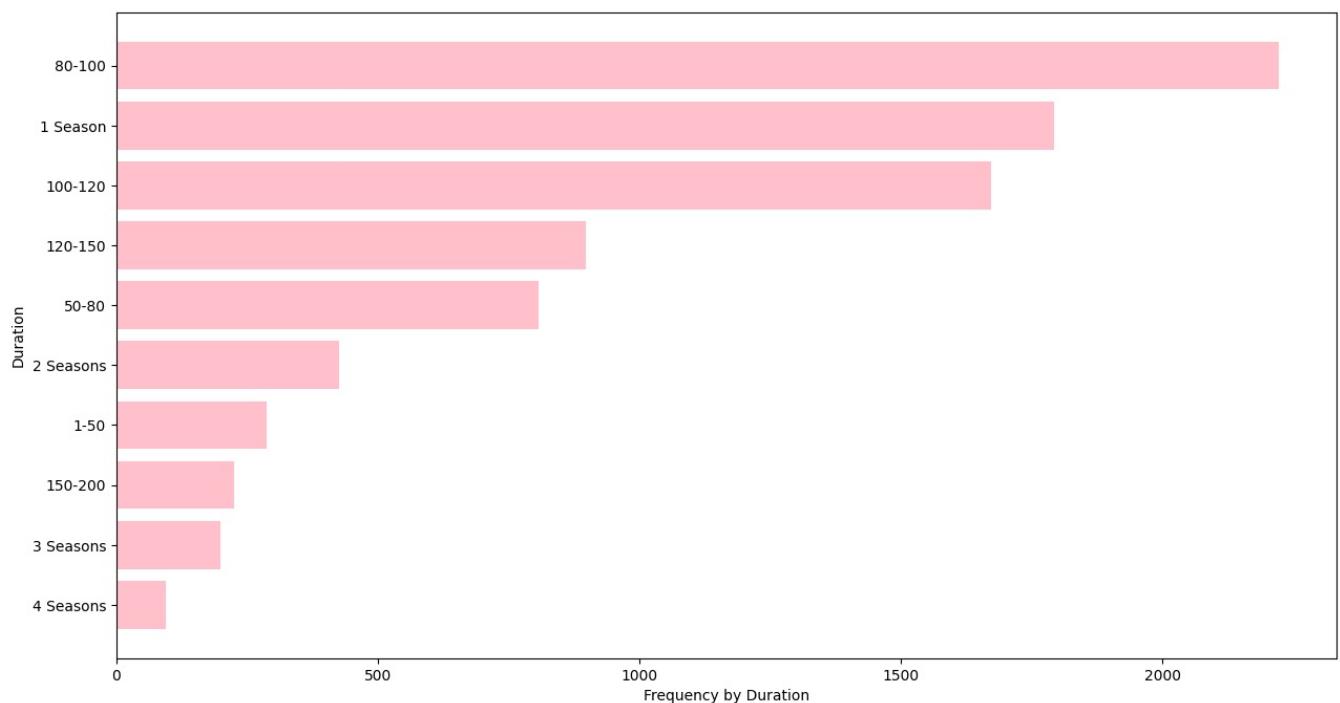
Most of the highly rated content on Netflix is intended for Mature Audiences, R Rated, content not intended for audience under 14 and those which require Parental Guidance

```
In [123]: #number of distinct titles on the basis of duration
df_final.groupby(['duration']).agg({'title':'nunique'})
```

Out[123]:

duration	title
1 Season	1793
1-50	287
10 Seasons	7
100-120	1672
11 Seasons	2
12 Seasons	2
120-150	897
13 Seasons	3
15 Seasons	2
150-200	226
17 Seasons	1
2 Seasons	425
200-315	19
3 Seasons	199
4 Seasons	95
5 Seasons	65
50-80	808
6 Seasons	33
7 Seasons	23
8 Seasons	17
80-100	2222
9 Seasons	9

```
In [124]: df_duration=df_final.groupby(['duration']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_duration[:10]['duration'], df_duration[:10]['title'],color=['pink'])
plt.xlabel('Frequency by Duration')
plt.ylabel('Duration')
plt.show()
```



The duration of Most Watched content in our whole data is 80-100 mins. These must be movies and Shows having only 1 Season.

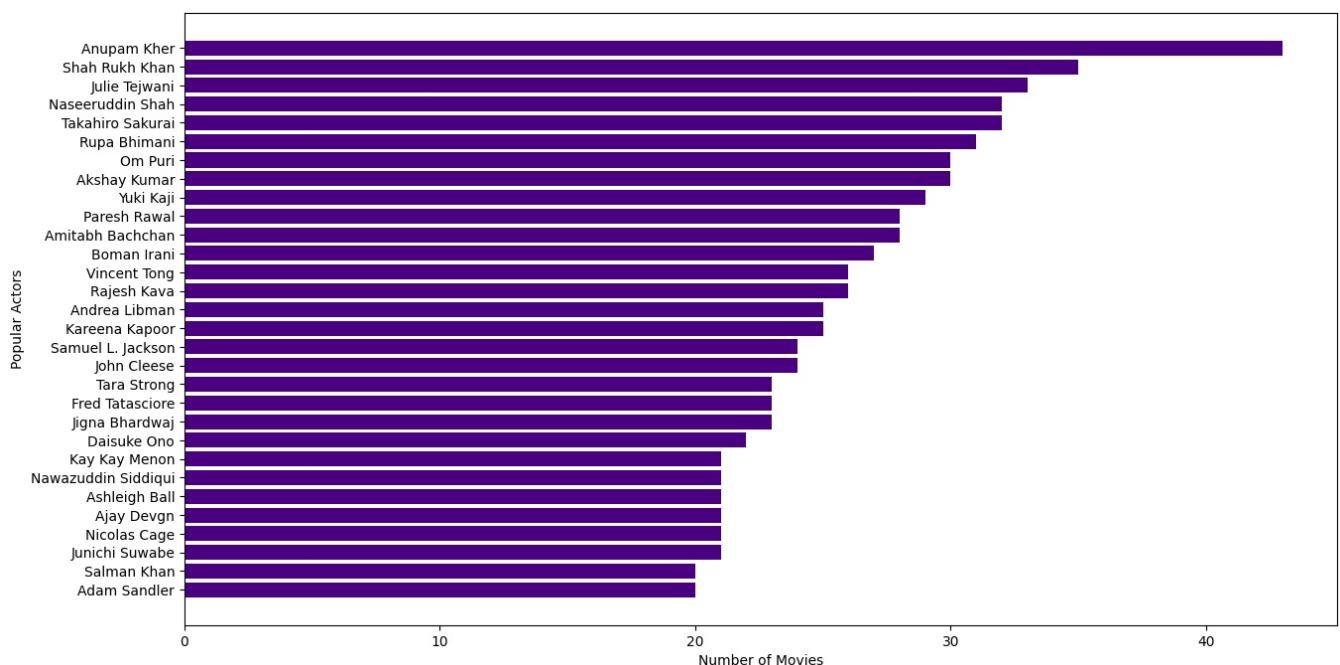
```
In [125]: #number of distinct titles on the basis of Actors
print(len(df_final1.groupby(['Actors']).agg({"title":"nunique"})))
df_final1.groupby(['Actors']).agg({"title":"nunique"})[:5]
```

36440

```
Out[125]:
```

Actors	
Jr.	2
"Riley" Lakdhar Dridi	1
'Najite Dede	2
2 Chainz	1
2Mex	1

```
In [126]: df_actors=df_final1.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barrh(df_actors[::-1]['Actors'], df_actors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actors')
plt.show()
```



Anupam Kher, SRK, Julie Tejwani, Naseeruddin Shah and Takahiro Sakurai occupy the top spot in Most Watched content.

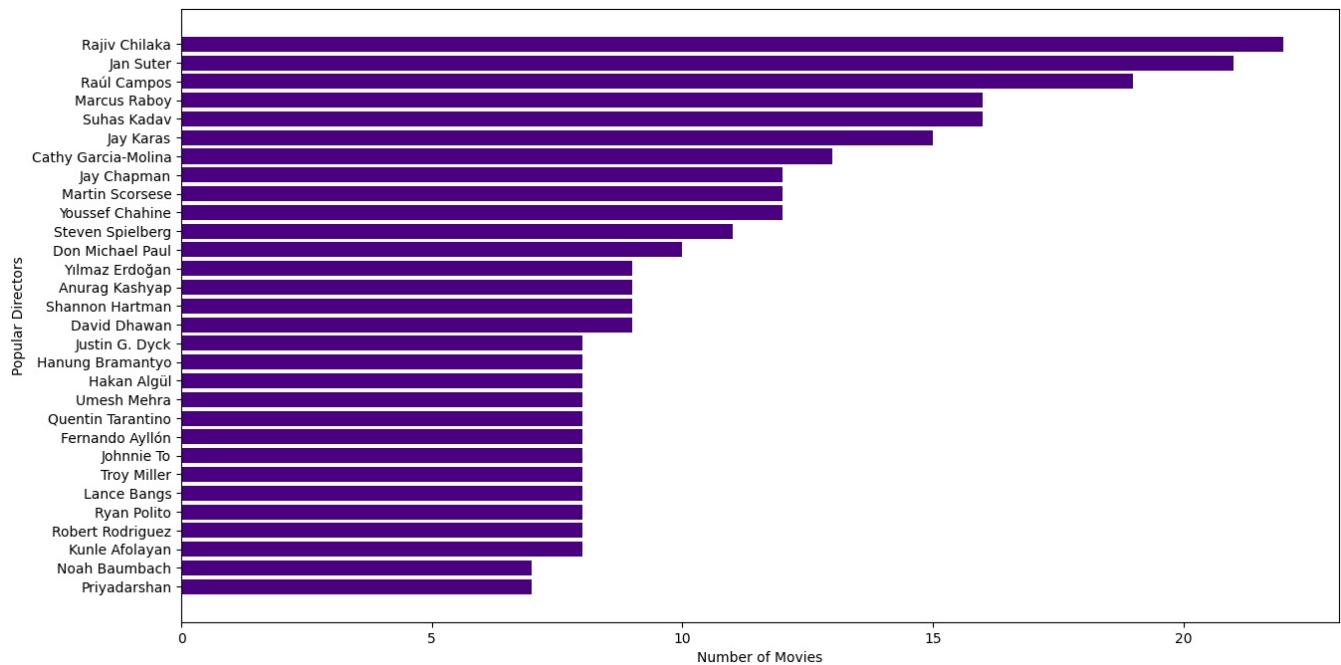
```
In [127]: #number of distinct titles on the basis of Actors
print(len(df_final1.groupby(['Directors']).agg({"title":"nunique"})))
df_final1.groupby(['Directors']).agg({"title":"nunique"})[:3]
```

4994

Out[127]:

Directors	title
A. L. Vijay	2
A. Raajdheep	1
A. Salaam	1

```
In [128]: df_directors=df_final1.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



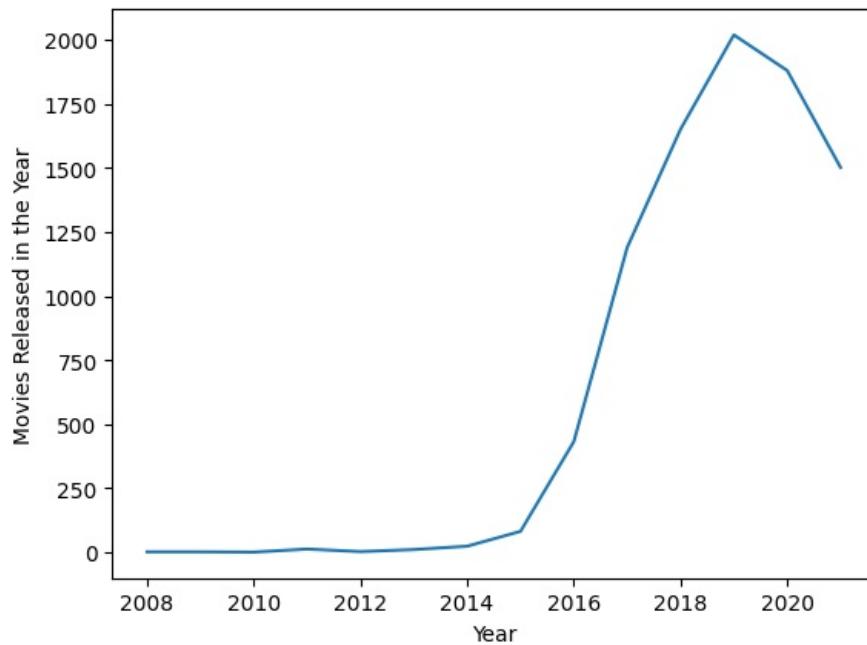
Rajiv Chilaka, Jan Suter and Raul Campos are the most popular directors across Netflix

```
In [129]: #number of distinct titles on the basis of year
df_final1.groupby(['year']).agg({"title":"nunique"})
```

Out[129]:

year	title
2008	2
2009	2
2010	1
2011	13
2012	3
2013	11
2014	24
2015	82
2016	432
2017	1189
2018	1650
2019	2018
2020	1879
2021	1501

```
In [130]: df_year=df_final1.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



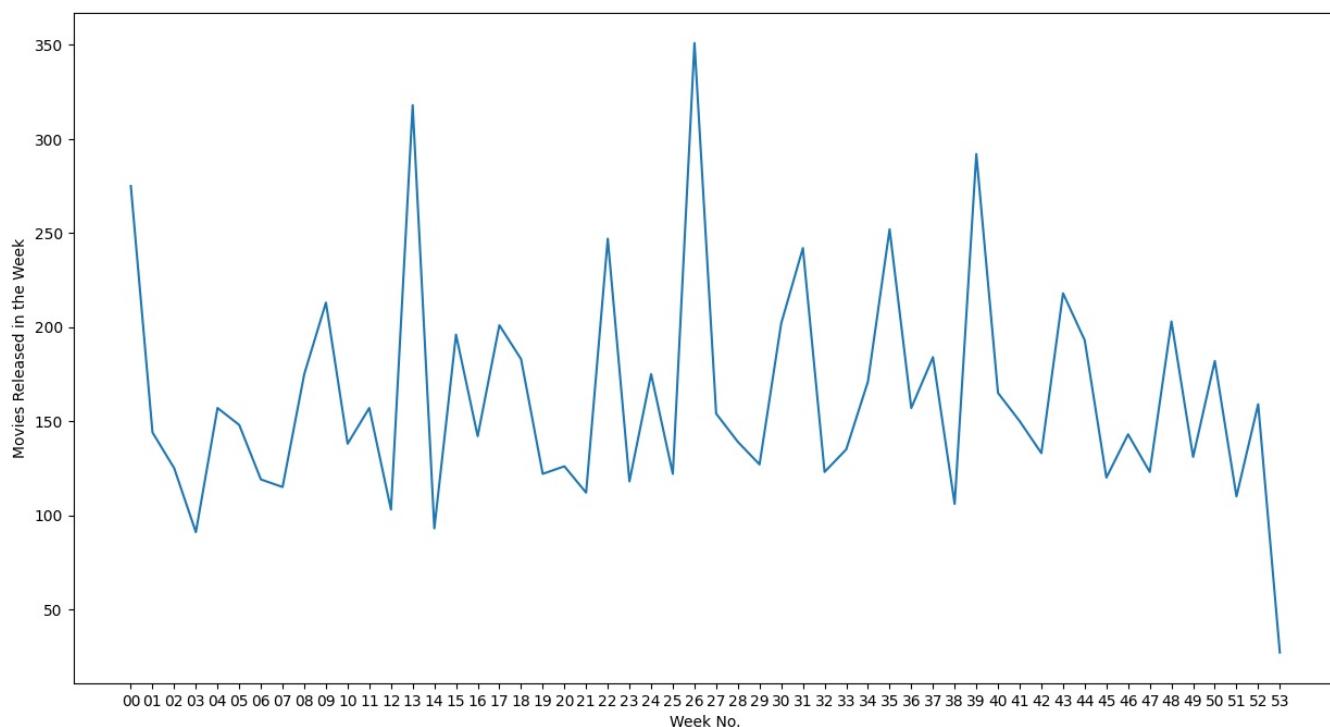
The Amount of Content across Netflix has increased from 2008 continuously till 2019. Then started decreasing from here(probably due to Covid)

```
In [131]: #number of distinct titles on the basis of week
print(len(df_final1.groupby(['week_Added']).agg({"title":"nunique"})))
df_final1.groupby(['week_Added']).agg({"title":"nunique"})[:3]
```

54

```
Out[131]: title
week_Added
00    275
01    144
02    125
```

```
In [132]: df_week=df_final1.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



Most of the Content across Netflix is added in the first week of the year and it follows a bit of a cyclical pattern

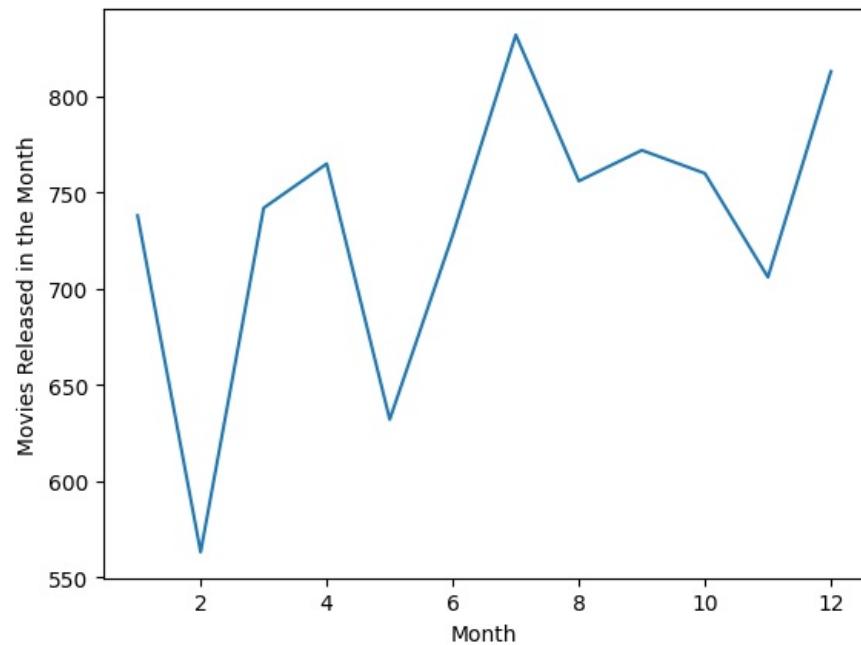
```
In [133]: #number of distinct titles on the basis of week
```

```
df_final1.groupby(['month_added']).agg({"title":"nunique"})
```

Out[133]:

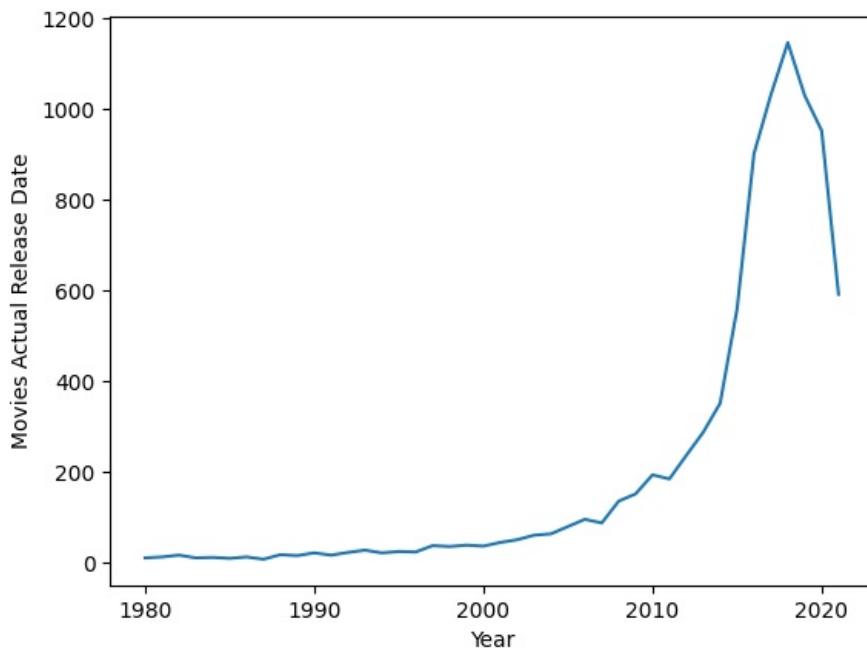
month_added	title
1	738
2	563
3	742
4	765
5	632
6	728
7	832
8	756
9	772
10	760
11	706
12	813

```
In [134]: df_month=df_final1.groupby(['month_added']).agg({"title":"nunique"}).reset_index()  
sns.lineplot(data=df_month, x='month_added', y='title')  
plt.ylabel("Movies Released in the Month")  
plt.xlabel("Month")  
plt.show()
```



Most of the content is added in the first and last months across Netflix(reinstating what we observed for first week in baove plot)

```
In [135]: df_release_year=df_final1[df_final1['release_year']>=1980].groupby(['release_year']).agg({"title":"nunique"}).reset_index()  
sns.lineplot(data=df_release_year, x='release_year', y='title')  
plt.ylabel("Movies Actual Release Date")  
plt.xlabel("Year")  
plt.show()
```

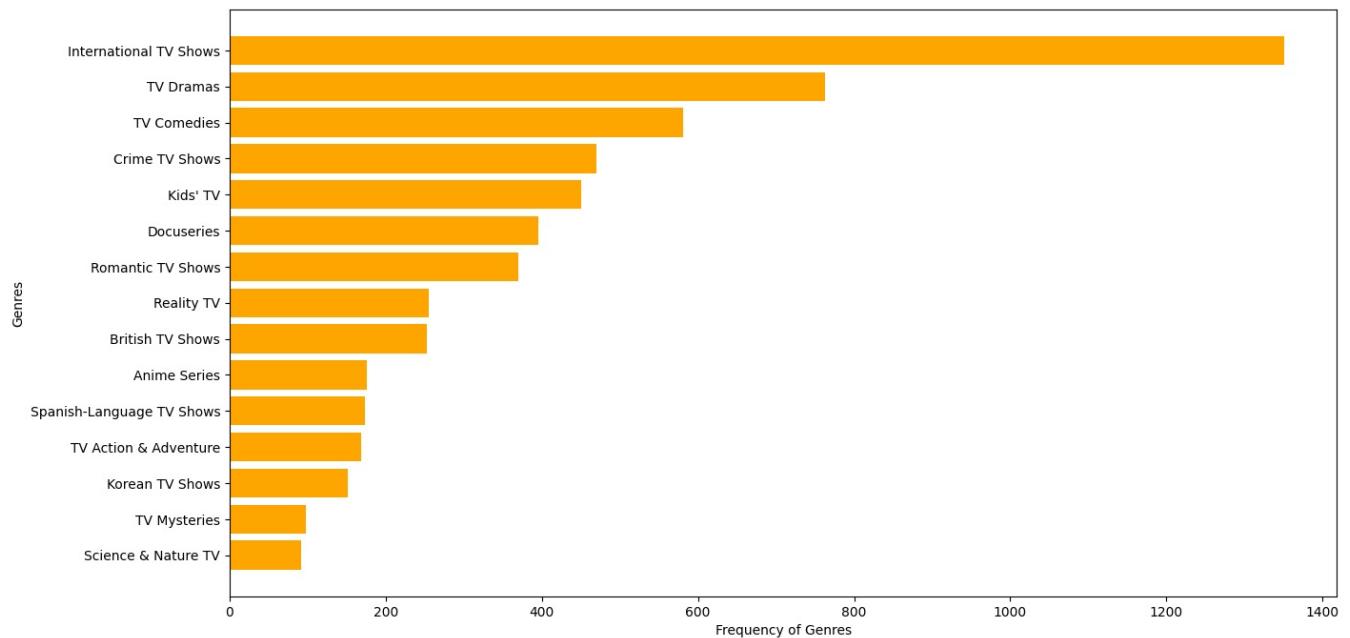


Net content release which are later uploaded to Netflix has increased since 1980 till 2020 though later reduced certainly due to COVID-19

Univariate Analysis separately for shows and movies

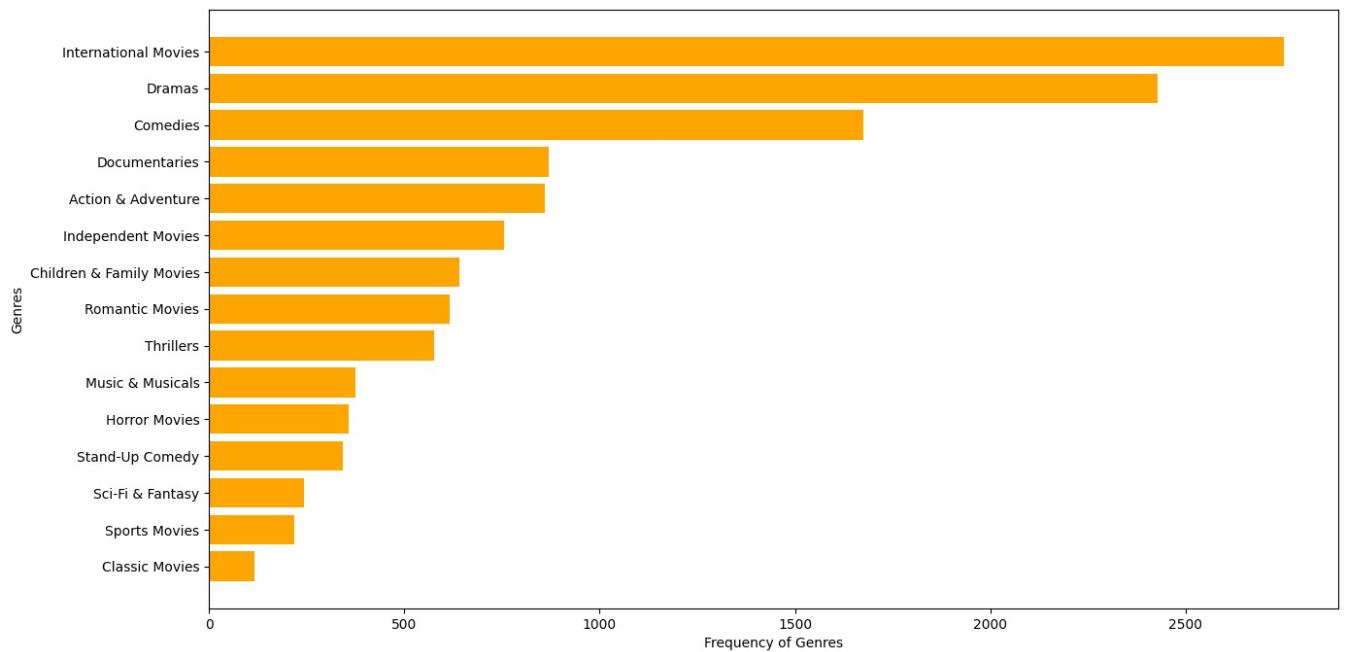
```
In [136]: df_shows=df_final1[df_final1['type']=='TV Show']
df_movies=df_final1[df_final1['type']=='Movie']
```

```
In [137]: df_genre=df_shows.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_genre[:::-1]['Genre'], df_genre[:::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```



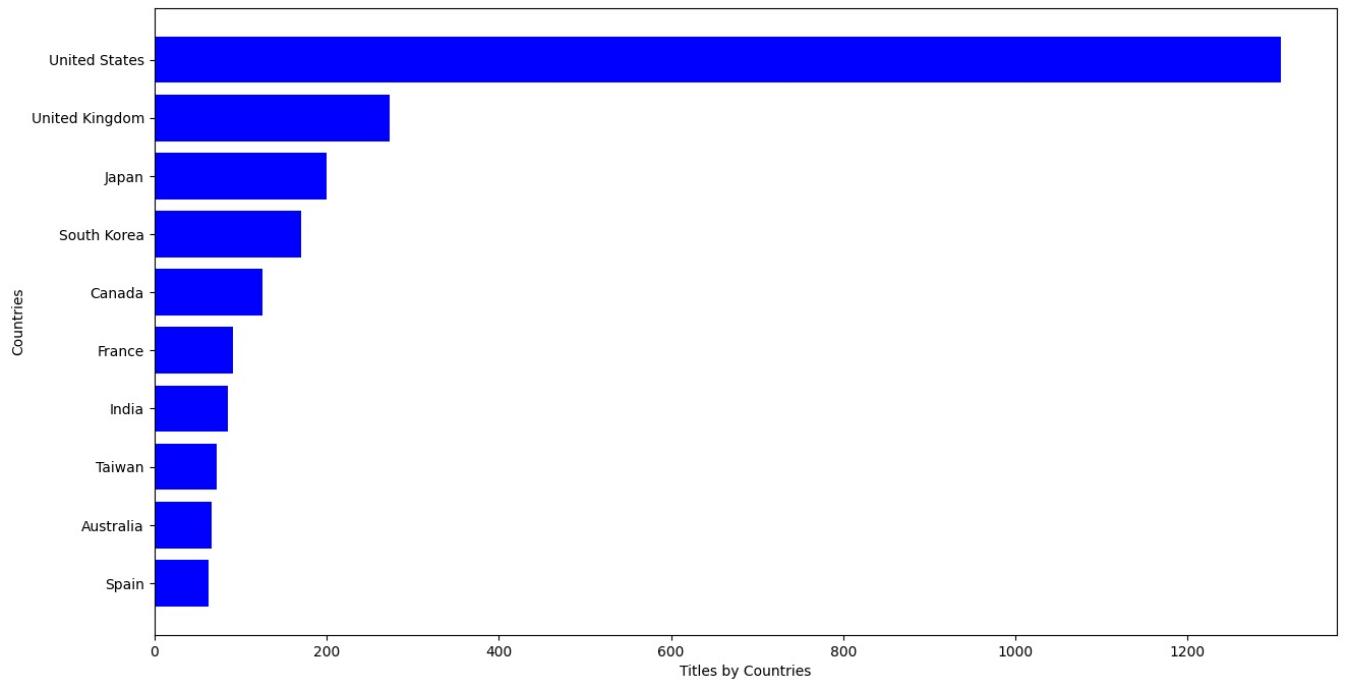
International TV Shows, Dramas and Comedy Genres are popular across TV Shows in Netflix

```
In [138]: df_genre=df_movies.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_genre[:::-1]['Genre'], df_genre[:::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```

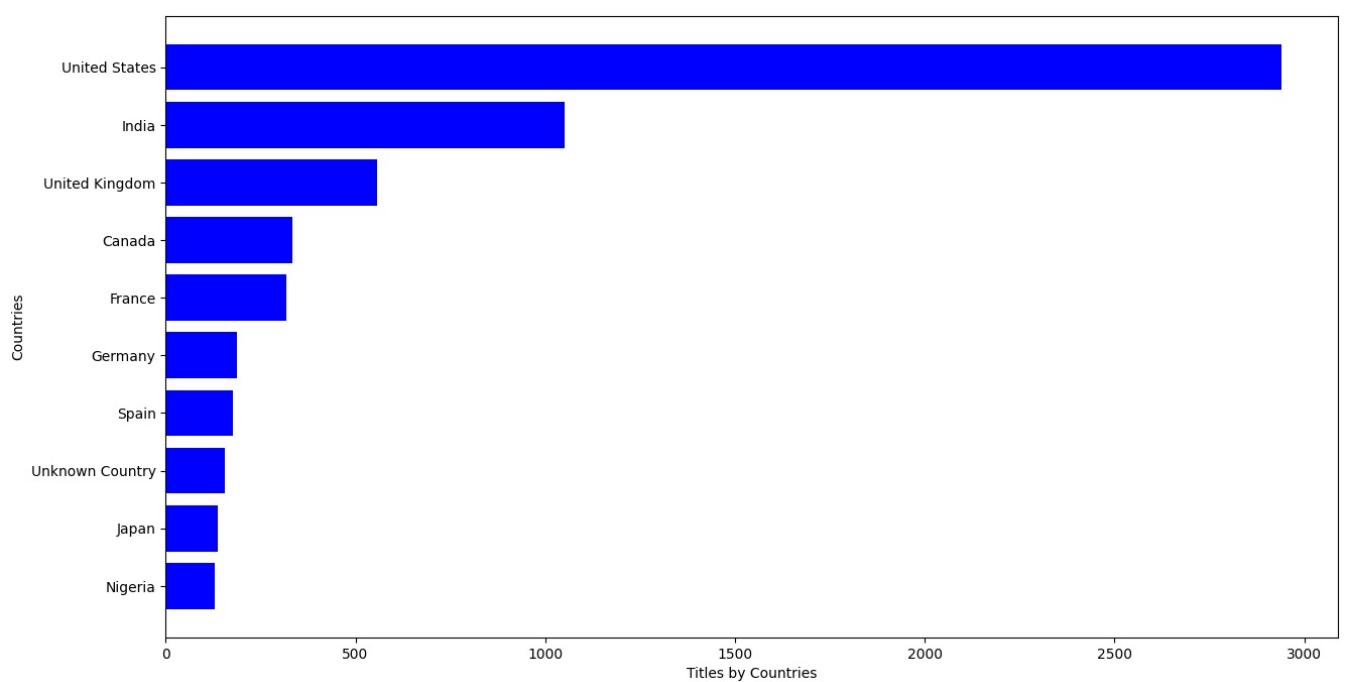


International Movies, Dramas and Comedy Genres are popular followed by Documentaries across Movies on Netflix

```
In [139]: df_country=df_shows.groupby(['country']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_country[::1]['country'], df_country[::1]['title'],color=['blue'])
plt.xlabel('Titles by Countries')
plt.ylabel('Countries')
plt.show()
```



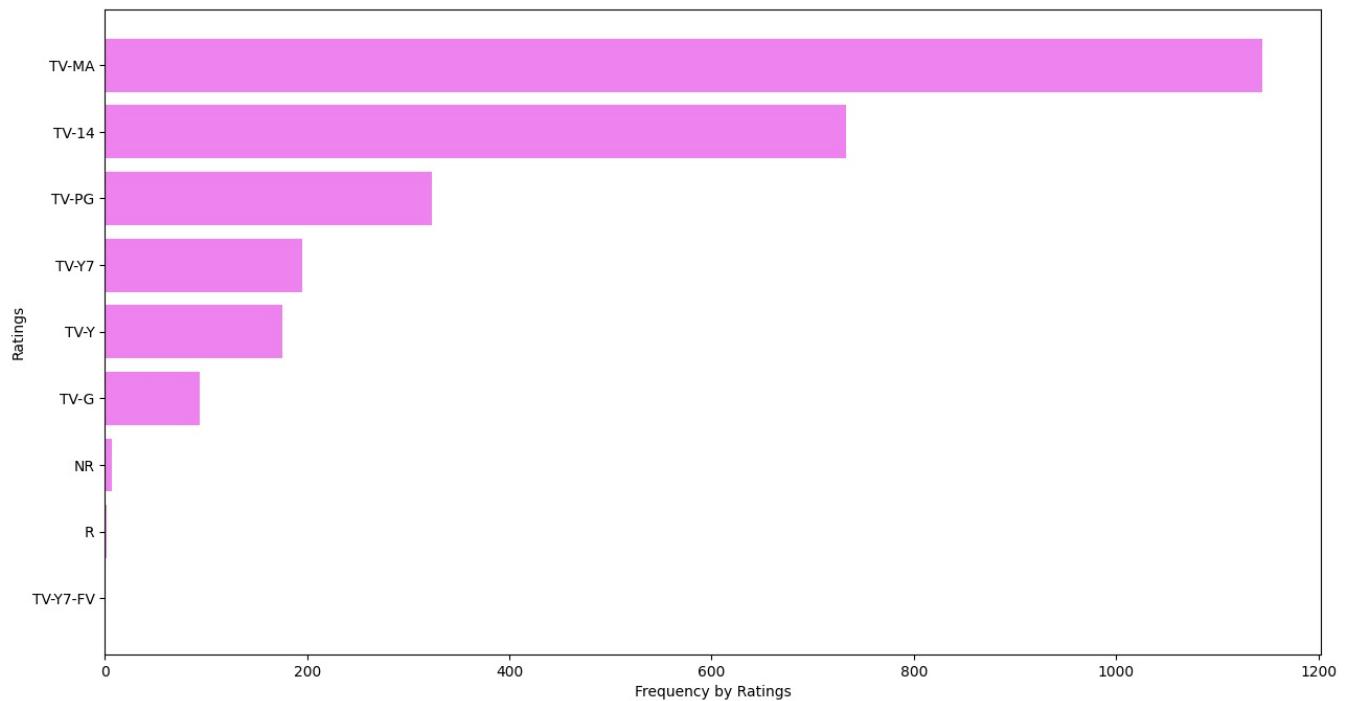
```
In [140]: df_country=df_movies.groupby(['country']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_country[::1]['country'], df_country[::1]['title'],color=['blue'])
plt.xlabel('Titles by Countries')
plt.ylabel('Countries')
plt.show()
```



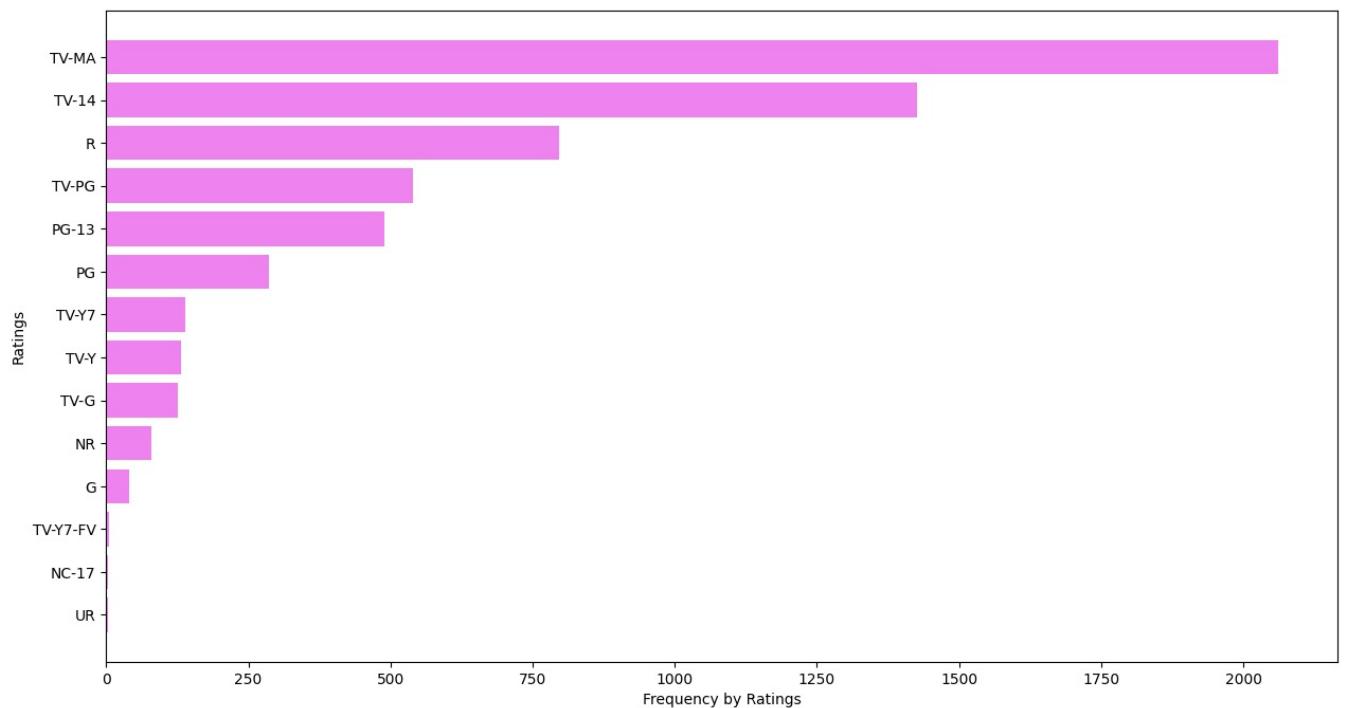
United States is leading across both TV Shows and Movies, UK also provides great content across TV Shows and Movies. Surprisingly India is much more prevalent in Movies as compared TV Shows.

Moreover the number of Movies created in India outweigh the sum of TV Shows and Movies across UK since India was rated as second in net sum of whole content across Netflix.

```
In [141]: df_rating=df_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



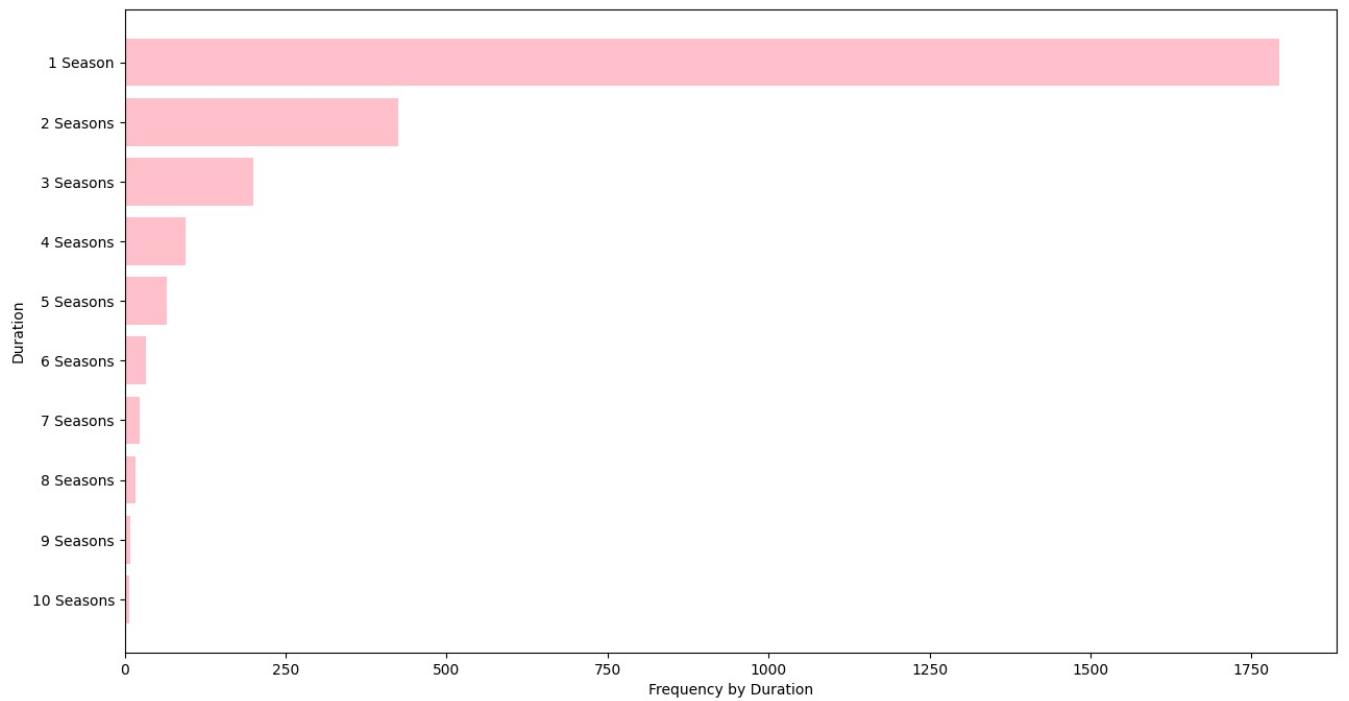
```
In [142]: df_rating=df_movies.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



So it seems plausible to conclude that the popular ratings across Netflix includes Mature Audiences and those appropriate for over 14/over 17 ages.

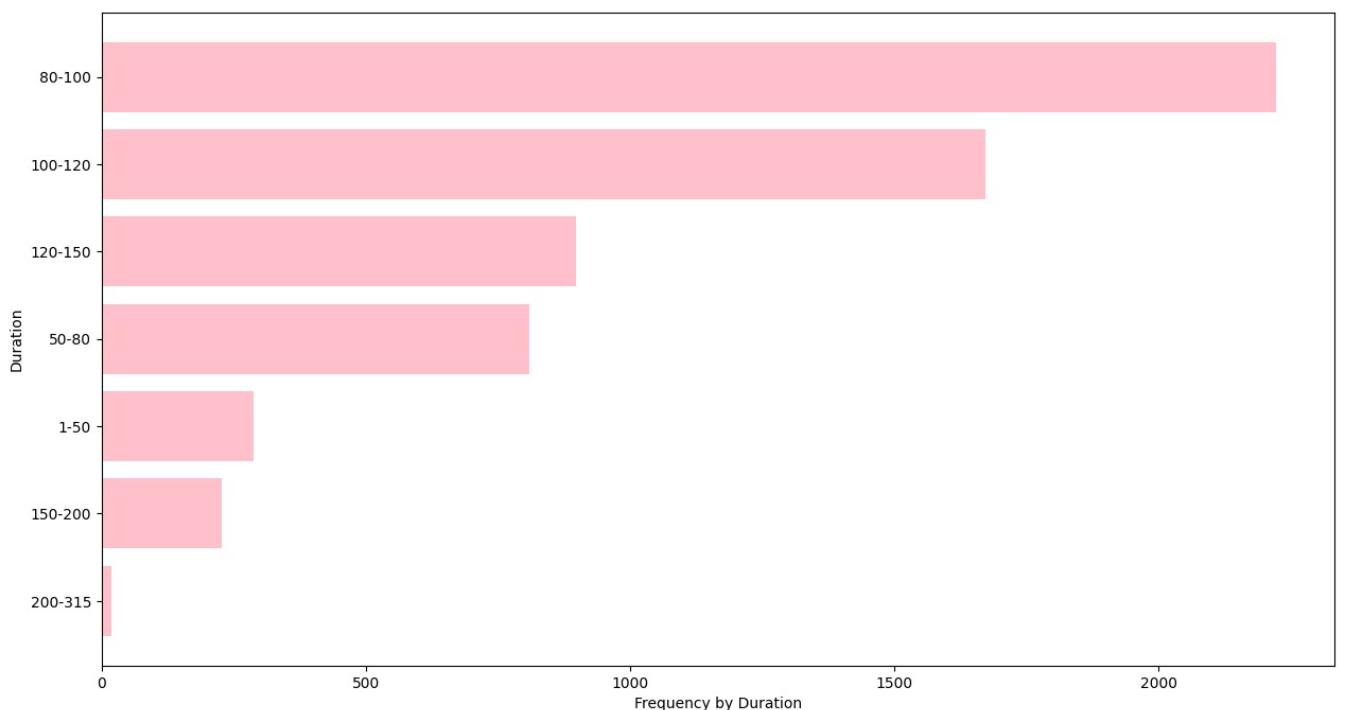
Moreover there are no TV Shows having a rating of R

```
In [143]: df_duration=df_shows.groupby(['duration']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_duration[::-1]['duration'], df_duration[::-1]['title'], color=['pink'])
plt.xlabel('Frequency by Duration')
plt.ylabel('Duration')
plt.show()
```



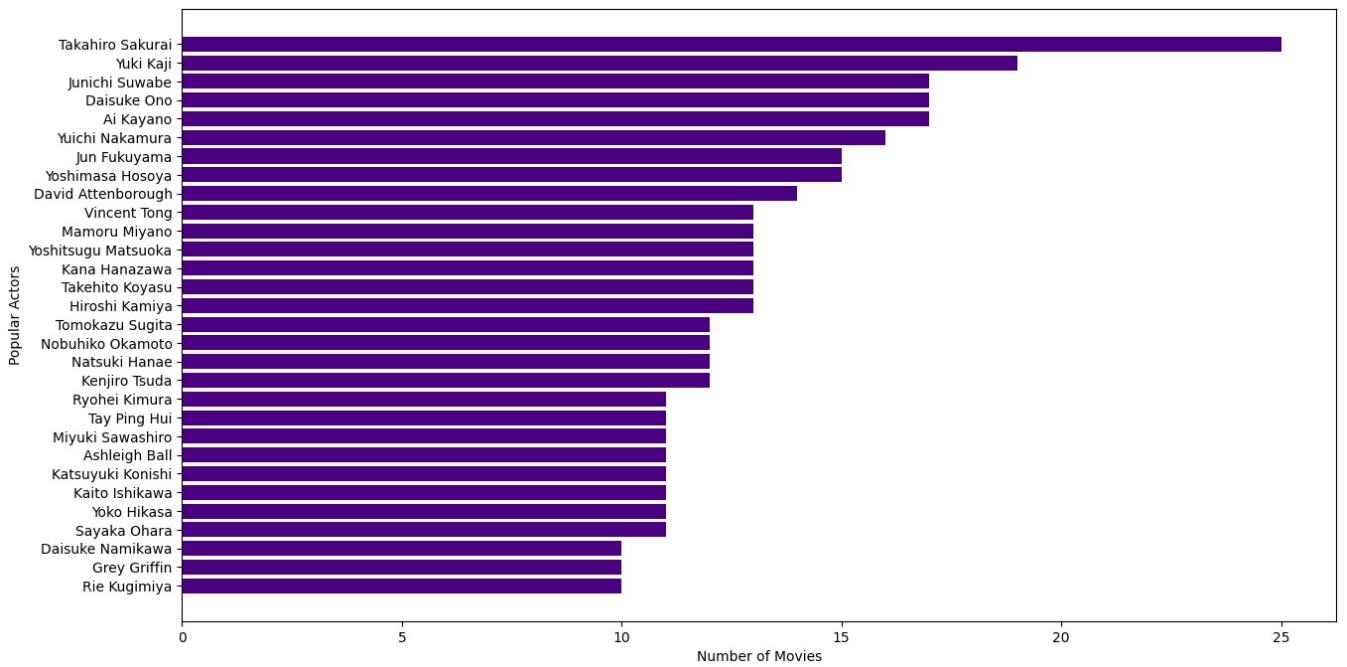
Across TV Shows, shows having only 1 Season are common as soon as the season length increases, the number of shows decrease and this definitely sounds as expected

```
In [144]: df_duration=df_movies.groupby(['duration']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_duration[::-1]['duration'], df_duration[::-1]['title'], color=['pink'])
plt.xlabel('Frequency by Duration')
plt.ylabel('Duration')
plt.show()
```



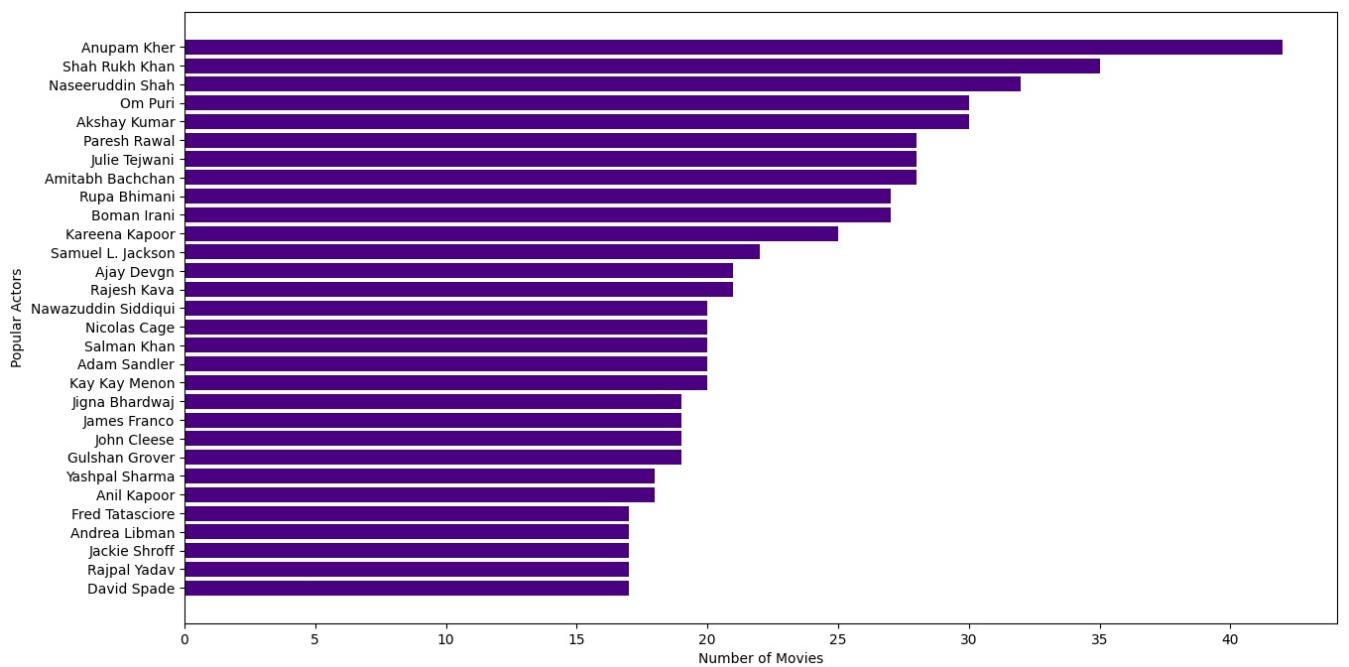
Across movies 80-100, 100-120 and 120-150 is the ranges of minutes for which most movies lie. So quite possibly 80-150 mins is the sweet spot we would be wanting for movies.

```
In [145]: df_actors=df_shows.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barh(df_actors[::-1]['Actors'], df_actors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actors')
plt.show()
```



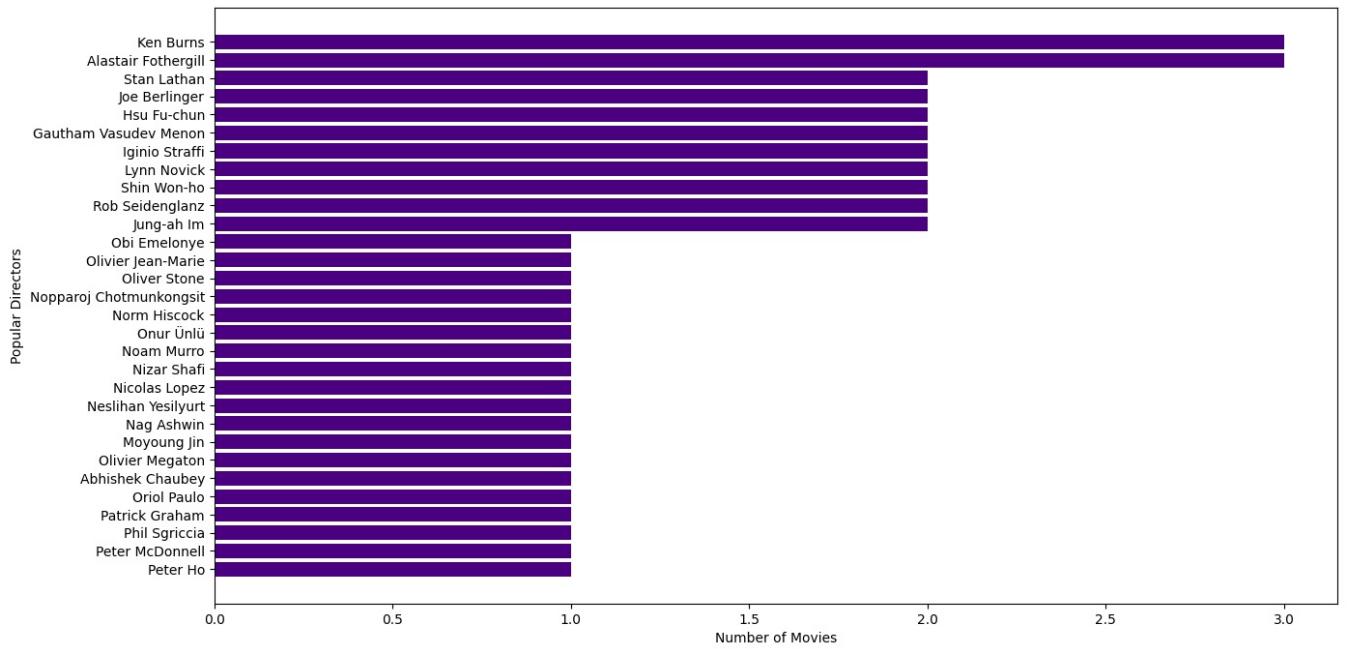
Takahiro Sakurai, Yuki Kaji and other South Korean/Japanese actors are the most popular actors across TV Shows

```
In [146]: df_actors=df_movies.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barh(df_actors[::-1]['Actors'], df_actors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actors')
plt.show()
```



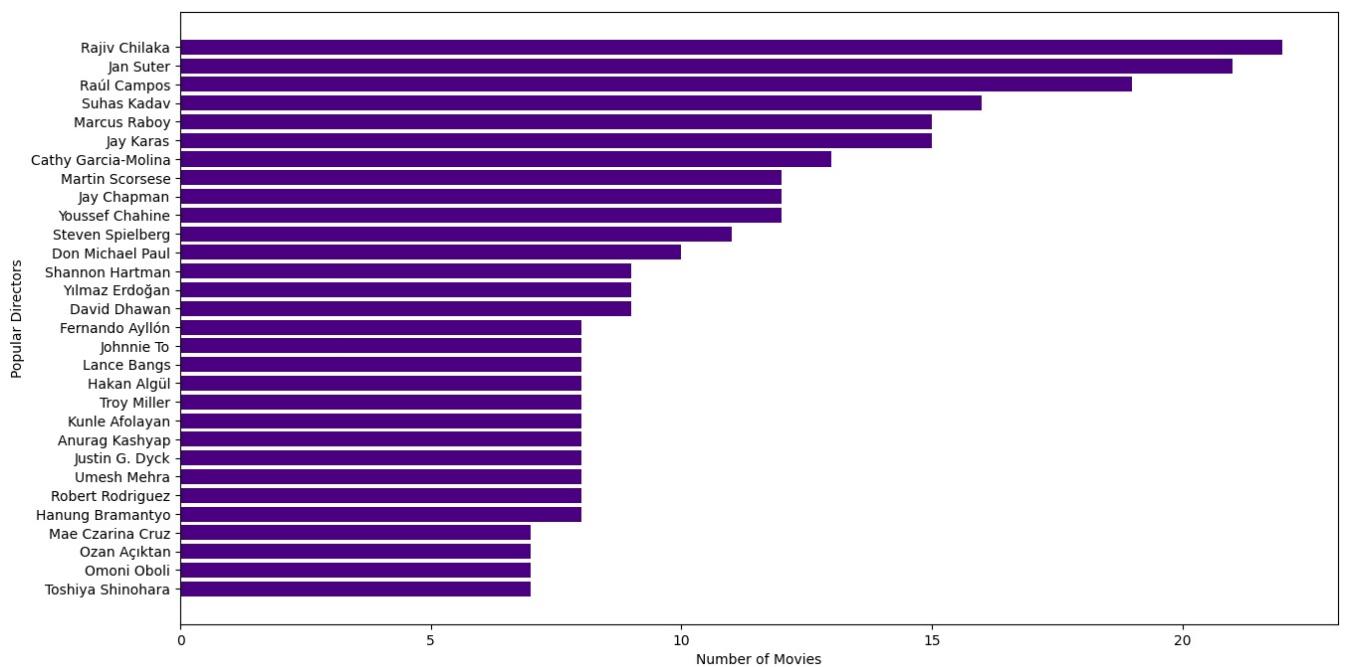
Our bollywood actors such as Anupam Kher, SRK, Naseeruddin Shah are very much popular across movies on Netflix

```
In [147]: df_directors=df_shows.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[:::-1]['Directors'], df_directors[:::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



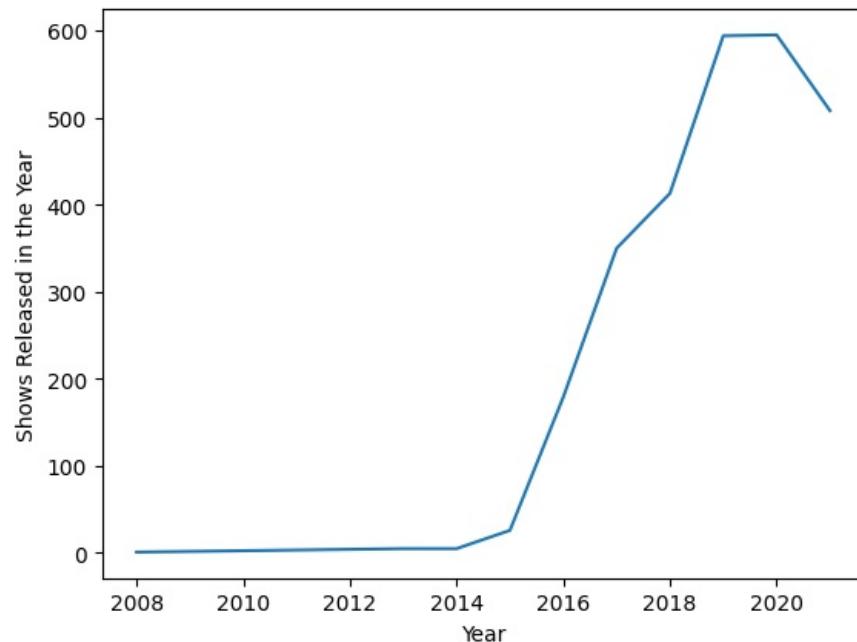
Ken Burns, Alastair Fothergill, Stan Lathan, Joe Barlinger are popular directors across TV Shows on Netflix

```
In [148]: df_directors=df_movies.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[:::-1]['Directors'], df_directors[:::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```

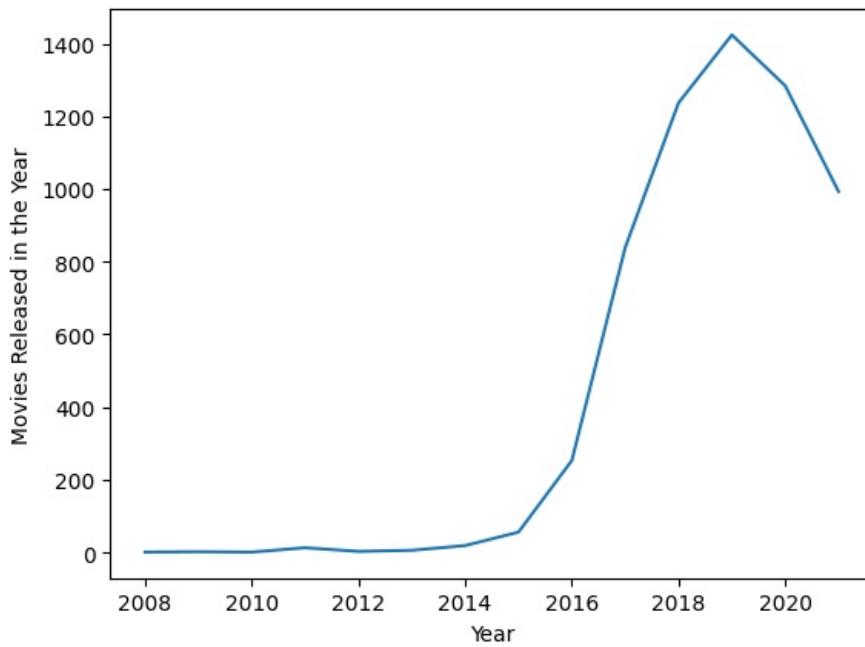


Rajiv Chilka, Jan Suter, Raul Campos, Suhas Kadav are popular directors across movies

```
In [149]: df_year=df_shows.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Shows Released in the Year")
plt.xlabel("Year")
plt.show()
```

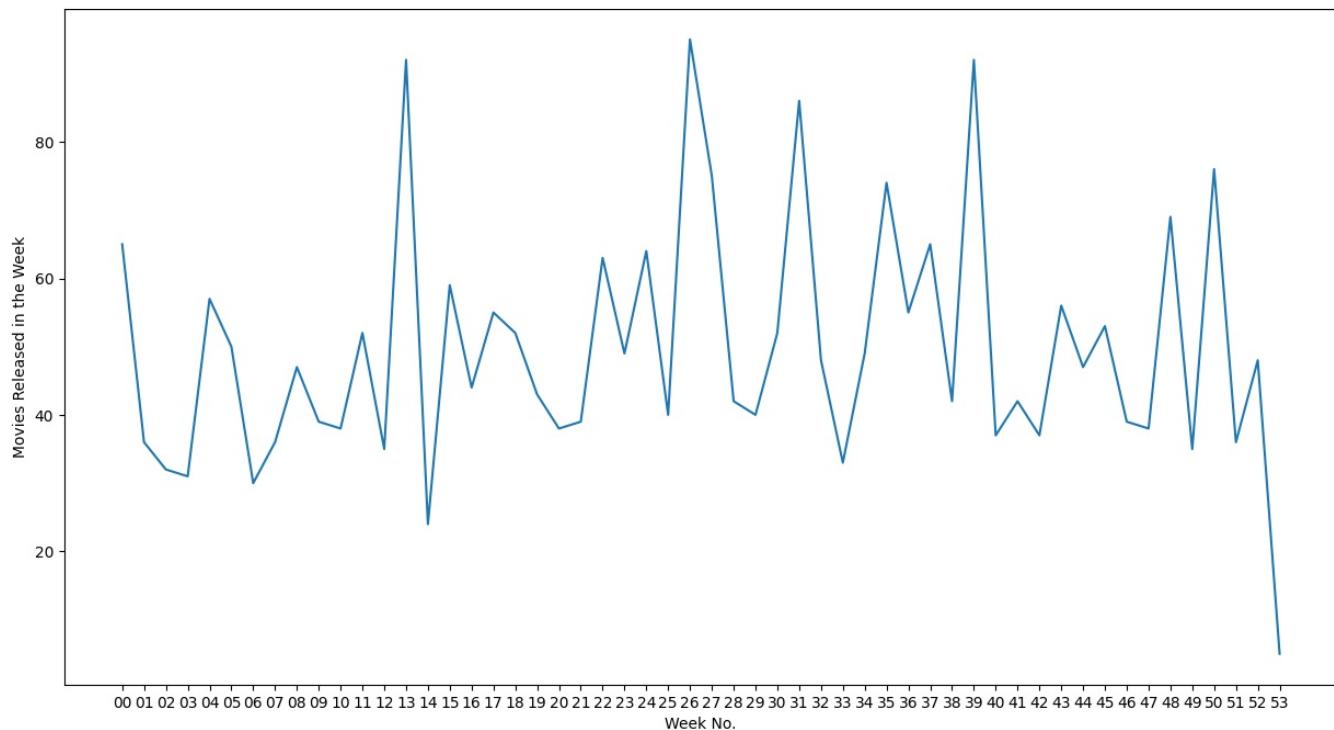


```
In [150]: df_year=df_movies.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```

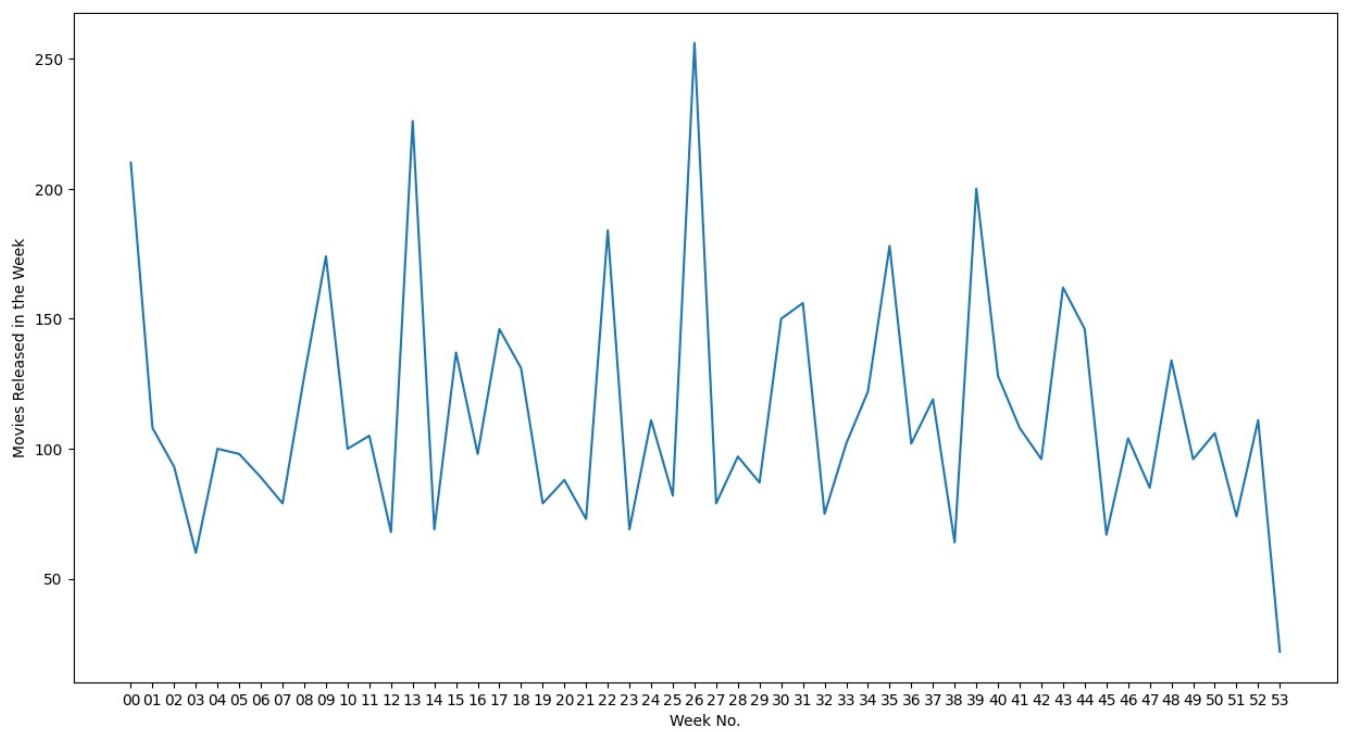


Till 2019, overall content across Netflix was increasing but due to Covid in 2020, though TV Shows didn't take a hit then Movies did take a hit. Well later in 2021, content across both was reduced significantly

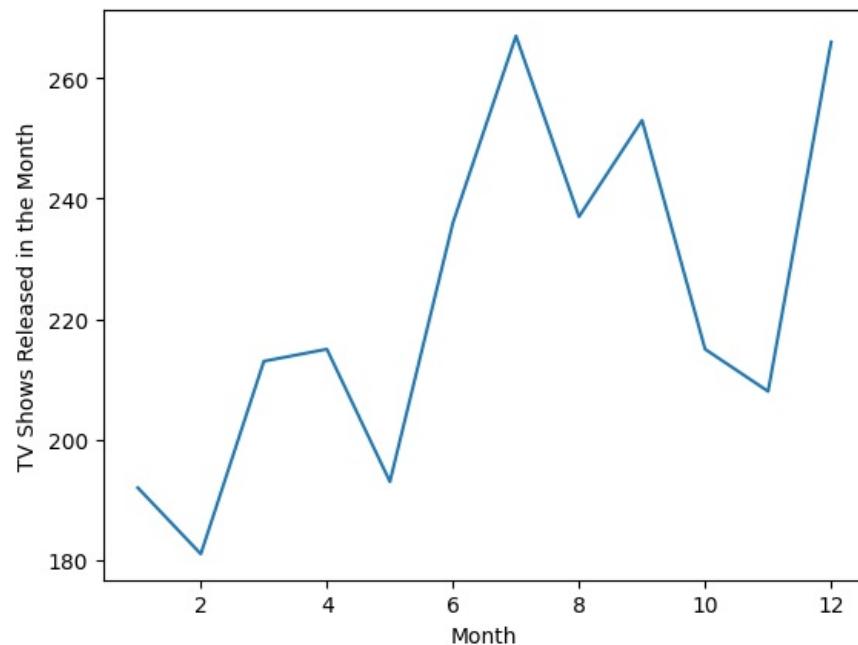
```
In [151]: df_week=df_shows.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



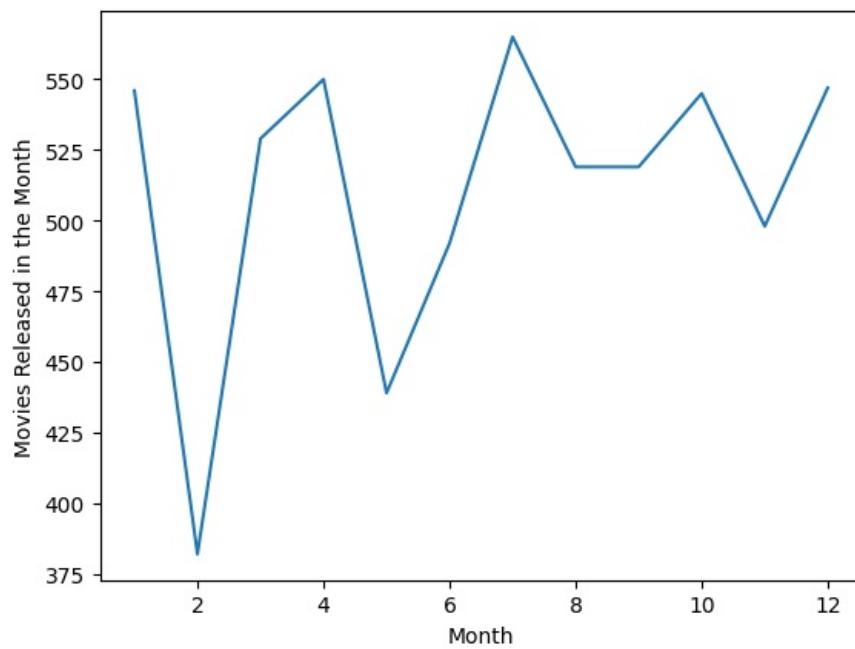
```
In [152]: df_week=df_movies.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



```
In [153]: df_month=df_shows.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("TV Shows Released in the Month")
plt.xlabel("Month")
plt.show()
```



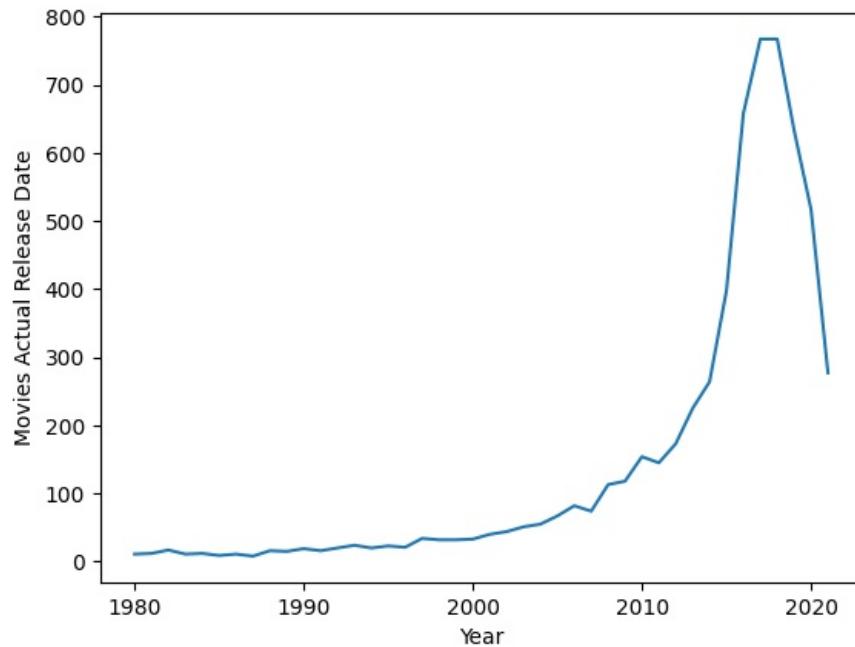
```
In [154]: df_month=df_movies.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("Movies Released in the Month")
plt.xlabel("Month")
plt.show()
```



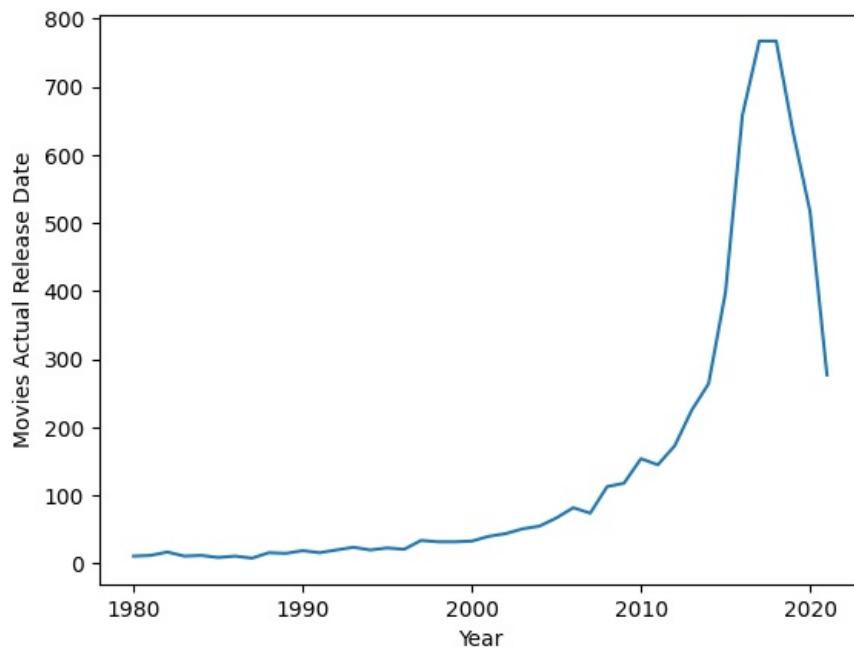
TV Shows are added in Netflix by a tremendous amount in mid weeks/months of the year, i.e- July

Movies are added in Netflix by a tremendous amount in first week/last month of current year and first month of next year

```
In [155]: df_release_year=df_movies[df_movies['release_year']>=1980].groupby(['release_year']).agg({'title':'nunique'}).reset_index()
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



```
In [156]: df_release_year=df_movies[df_movies['release_year']>=1980].groupby(['release_year']).agg({'title':'nunique'}).reset_index()
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



Actual Releases of both TV Shows and Movies have taken a hit after 2020

Questions to be Explored Now for Recommendations

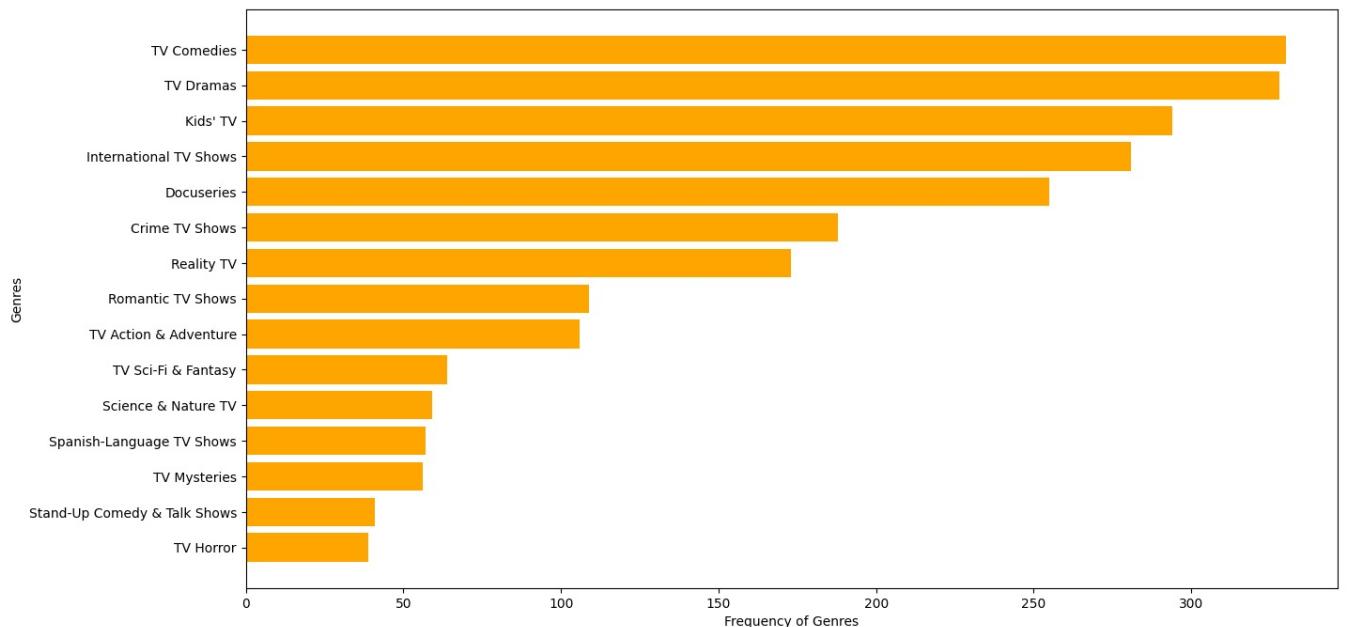
- 1) So this time, the granularity level is country and analysis of TV Shows/Movies the country brings. I am going to consider only the top countries individually for TV Shows and Movies. There are definitely some common countries too which bring out quality content in both TV Shows and Movies.
- 2) Which Genres do these countries offer and what are the intended audiences(Ratings) which are popular in Netflix?
- 3) In case of Movies, what is the duration/length of movies which makes them special and depicts attention span?
- 4) Who are the popular actors/directors across TV Shows and Movies in these countries?
- 5) In what time of the year, people tend to watch movies and shows in these countries?
- 6) Popular Actor and Director Combinations in these countries

```
In [157]: #below countries will be analyzed for both shows and movies
shows_and_movies=['United States','India','United Kingdom']
#below countries will be only analyzed on basis of shows
only_shows=['Japan','South Korea']
```

Univariate Analysis separately for shows and movies in USA

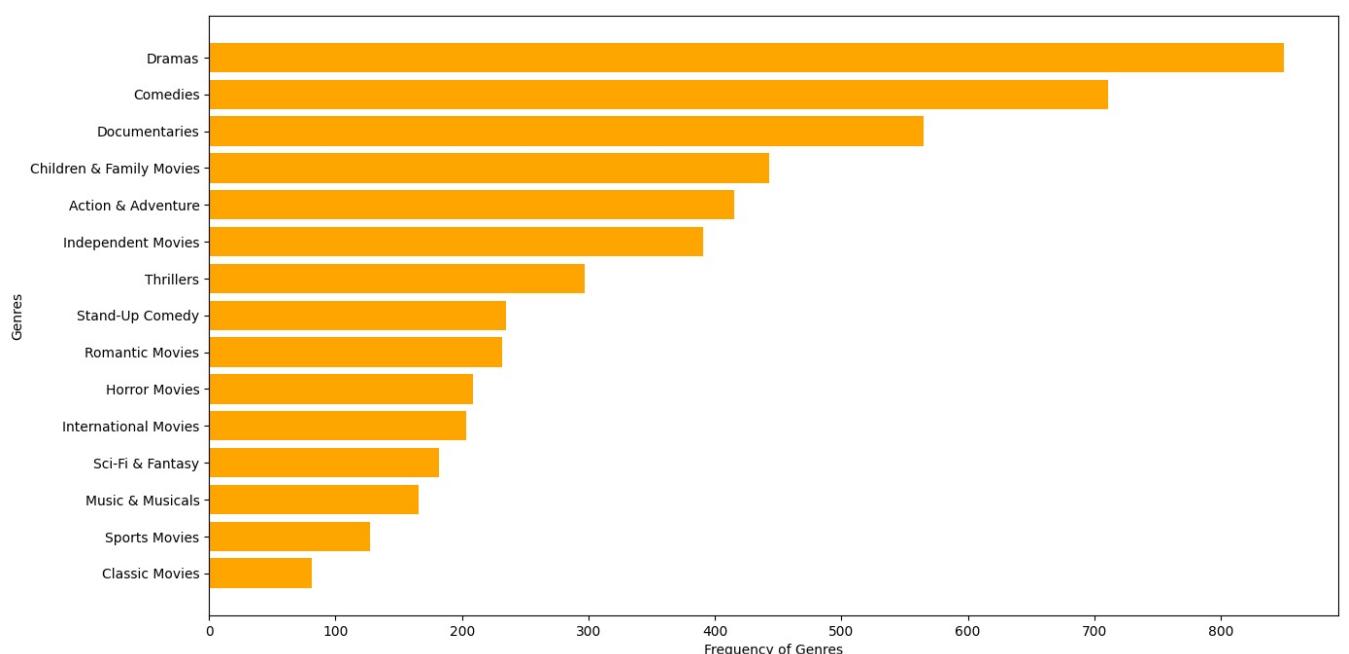
```
In [158]: #Analyzing USA for both shows and movies
df_usa_shows=df_final1[df_final1['country']=='United States'][df_final1[df_final1['country']=='United States']]
df_usa_movies=df_final1[df_final1['country']=='United States'][df_final1[df_final1['country']=='United States']]

In [159]: df_genre=df_usa_shows.groupby(['Genre']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_genre[:::-1]['Genre'], df_genre[:::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```



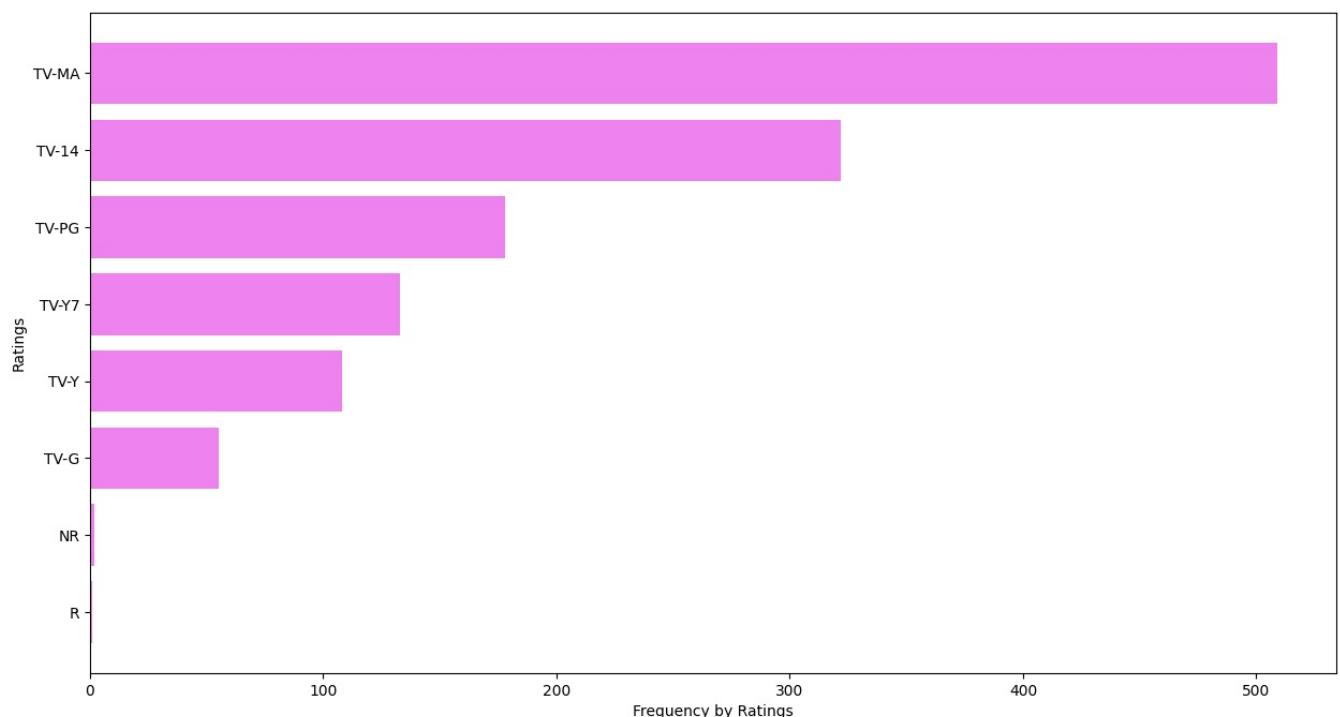
Dramas, Comedy, Kids' TV Shows, International TV Shows and Docuseries, Genres are popular in TV Series in USA

```
In [160]: df_genre=df_usa_movies.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_genre[::-1]['Genre'], df_genre[::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```

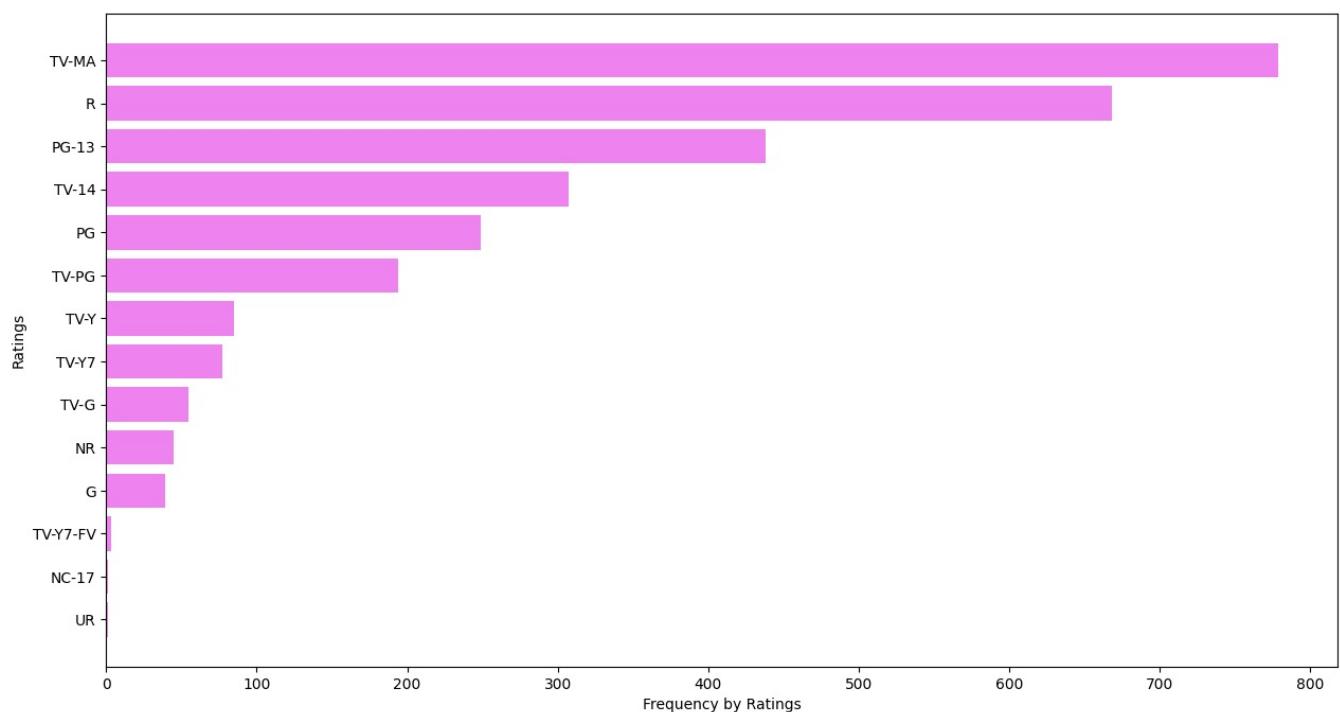


Dramas, Comedy, Documentaries, Family Movies and Action Genres in Movies are popular in USA

```
In [161]: df_rating=df_usa_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```

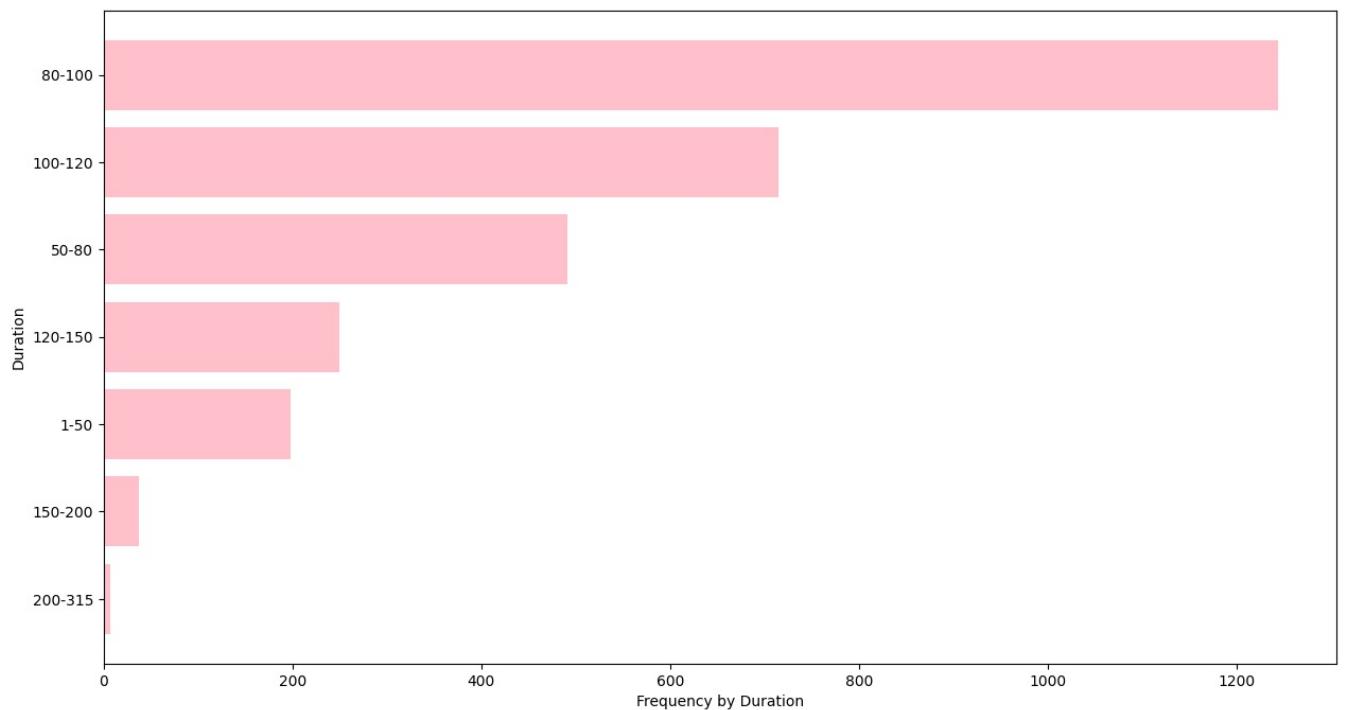


```
In [162]: df_rating=df_usa_movies.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=True)
plt.figure(figsize=(15,8))
plt.barrh(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



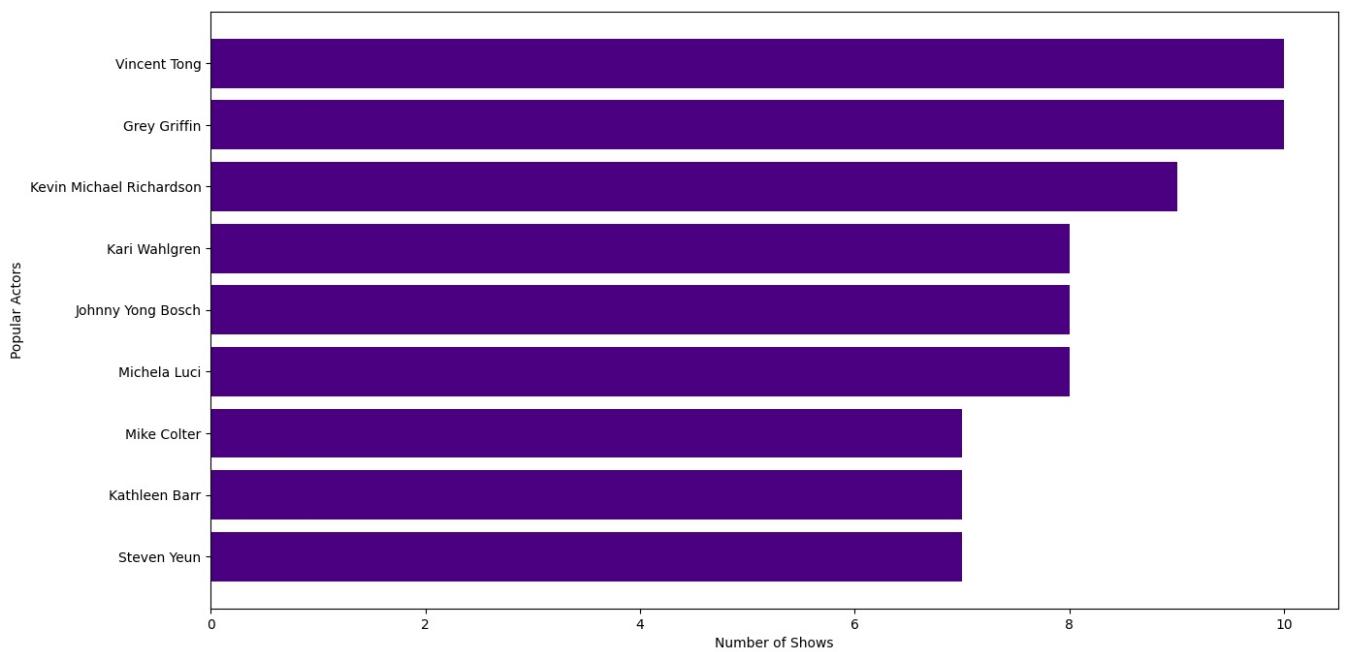
So it seems plausible to conclude that the popular ratings across Netflix includes Mature Audiences and those appropriate for over 14/over 17 ages in both Movies and TV Shows in USA

```
In [163]: df_duration=df_usa_movies.groupby(['duration']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=True)
plt.figure(figsize=(15,8))
plt.barrh(df_duration[::-1]['duration'], df_duration[::-1]['title'],color=['pink'])
plt.xlabel('Frequency by Duration')
plt.ylabel('Duration')
plt.show()
```



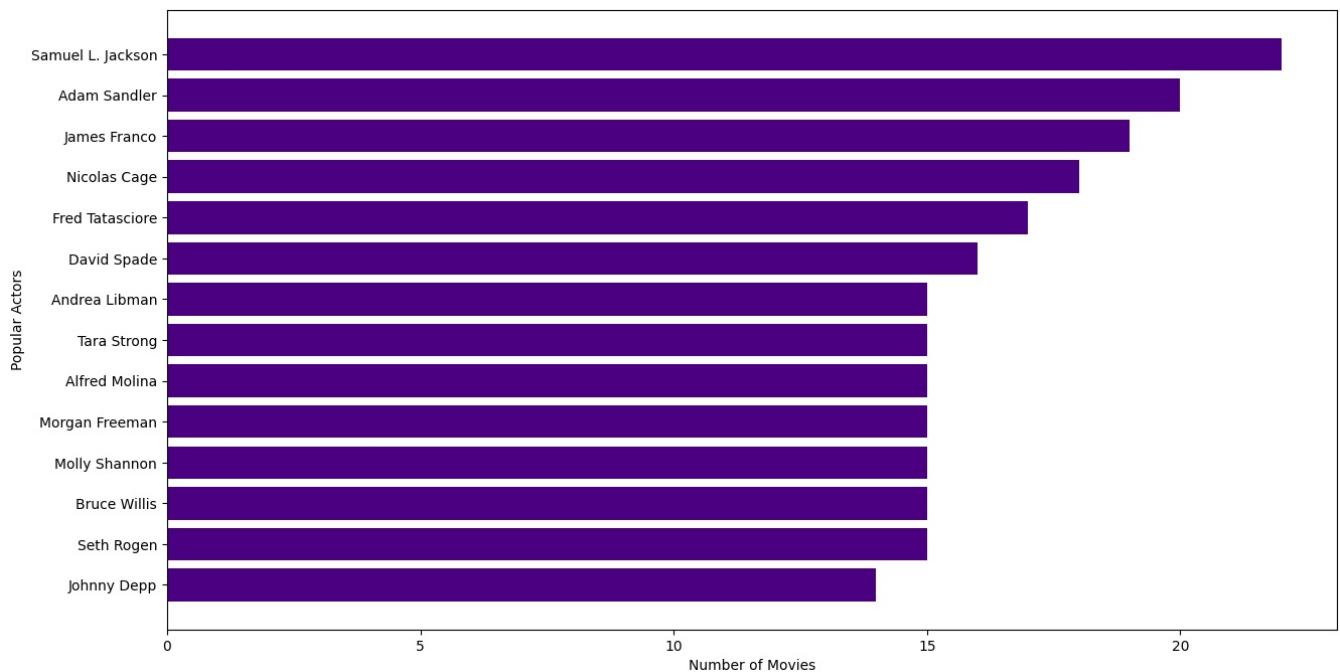
Across movies 80-100,100-120 is the ranges of minutes for which most movies lie. So quite possibly 80-120 mins is the sweet spot we would be wanting for movies in USA

```
In [164]: df_actors=df_usa_shows.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barrh(df_actors[:10]['Actors'], df_actors[:10]['title'], color=['indigo'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actors')
plt.show()
```



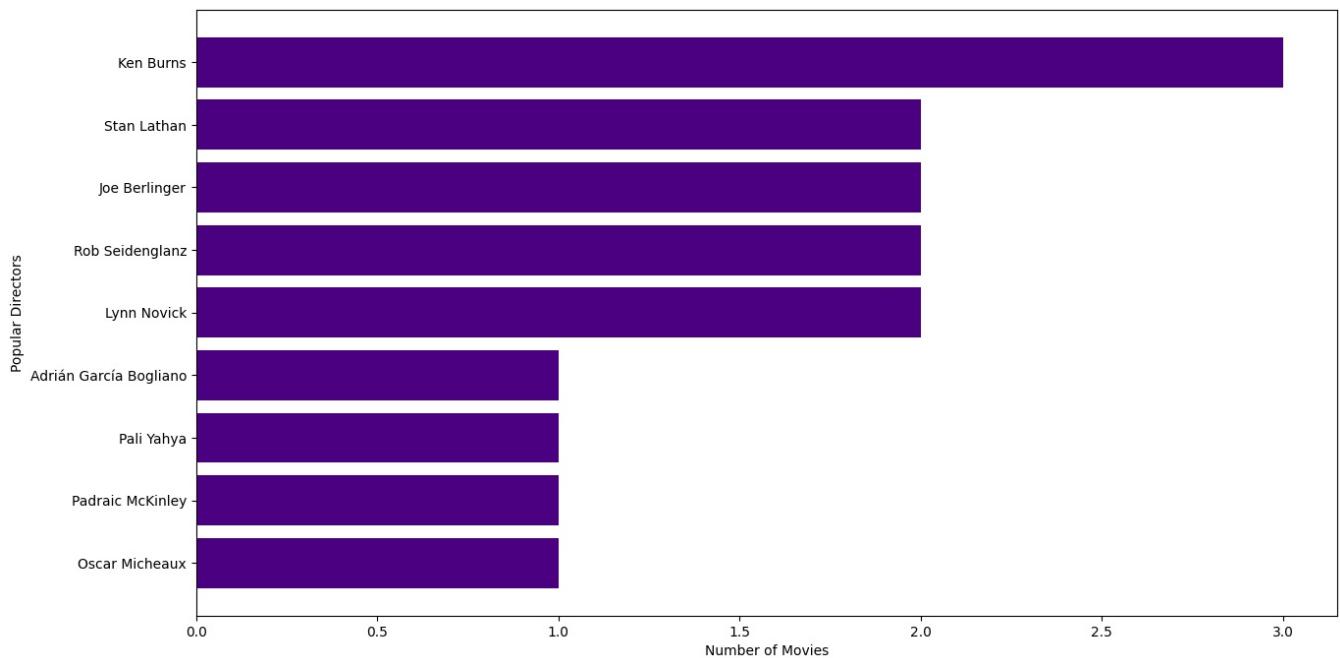
Vincent Tong, Grey Griffin and Kevin Richardson are the most popular actors across TV Shows in USA

```
In [165]: df_actors=df_usa_movies.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barrh(df_actors[:10]['Actors'], df_actors[:10]['title'], color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actors')
plt.show()
```



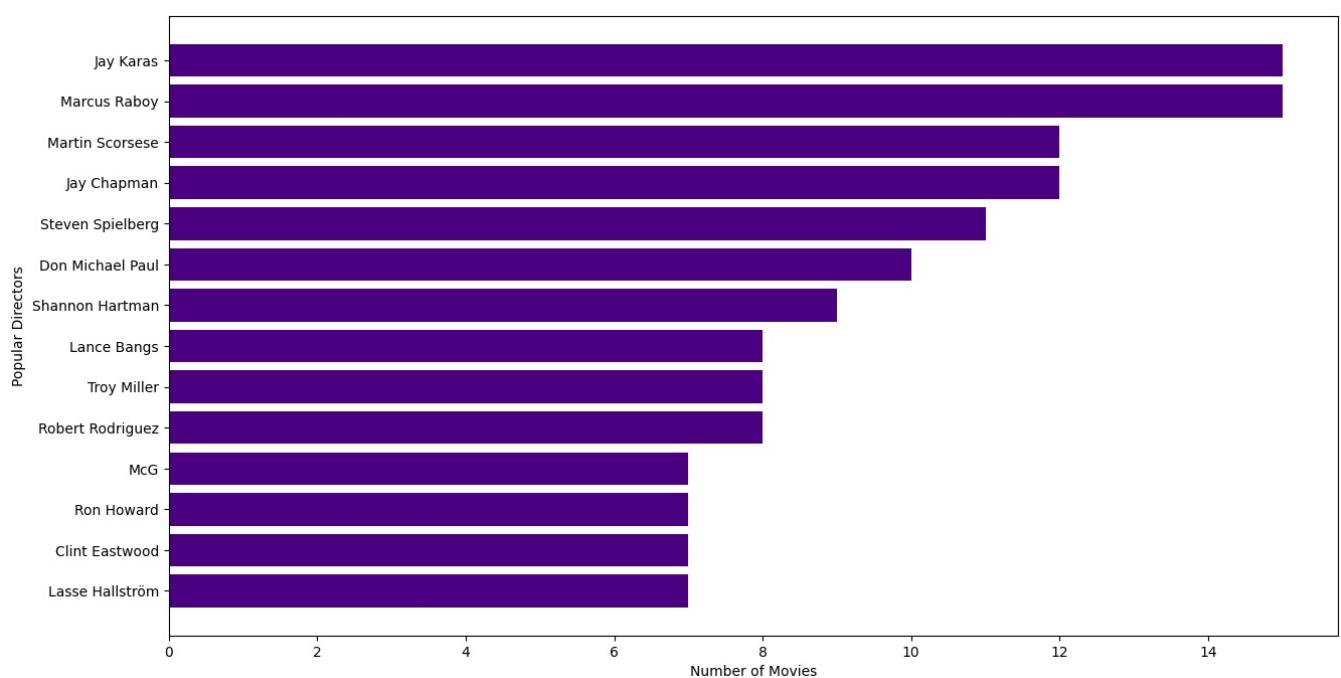
Samuel Jackson, Adam Sandler, James Franco and Nicolas Cage are very much popular across movies on Netflix in USA

```
In [166]: df_directors=df_usa_shows.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'])
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



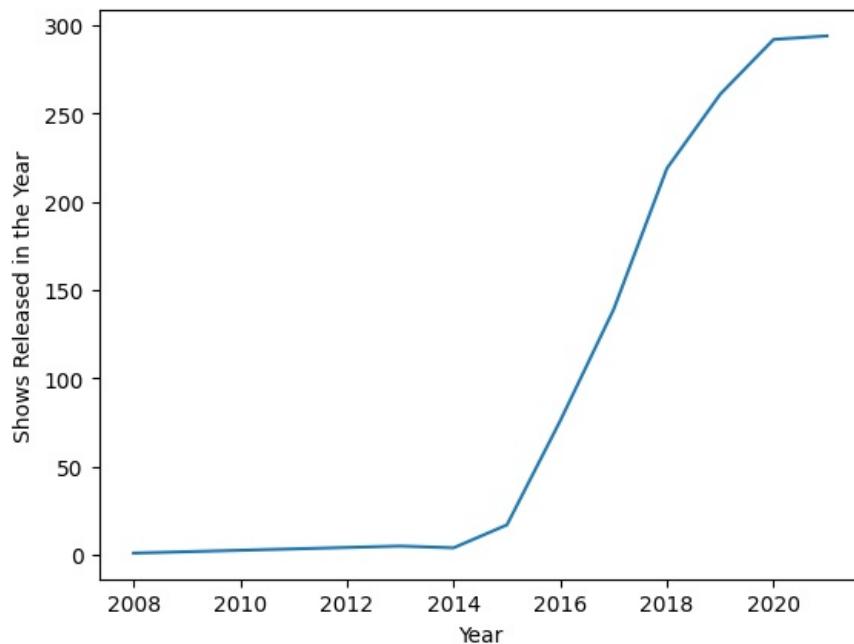
Ken Burns, Stan Lathan, Joe Barlinger are popular directors across TV Shows on Netflix in USA

```
In [167]: df_directors=df_usa_movies.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'])
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```

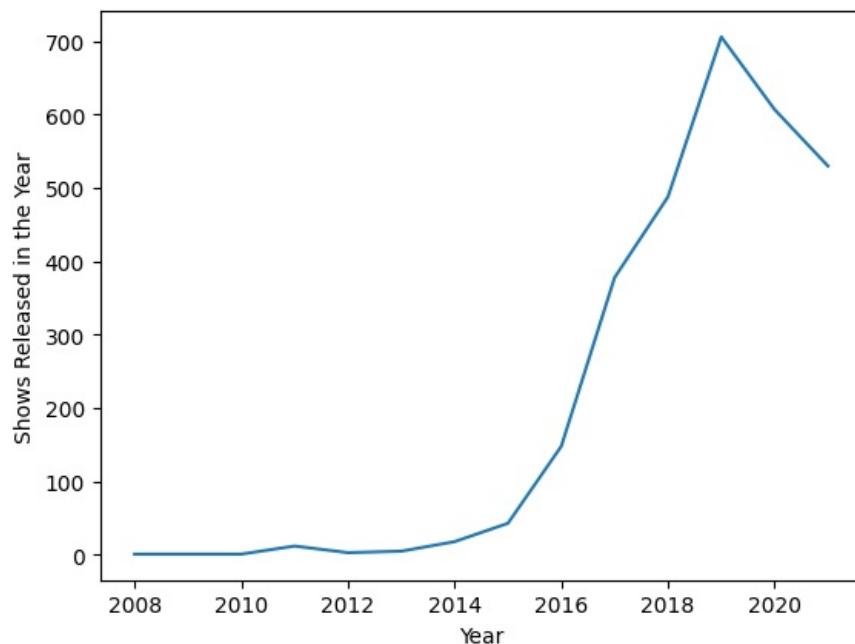


Jay Karas, Marcus Raboy, Martin Scorsese and Jay Chapman are popular directors across movies in USA

```
In [168]: df_year=df_usa_shows.groupby(['year']).agg({'title':'nunique'}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Shows Released in the Year")
plt.xlabel("Year")
plt.show()
```

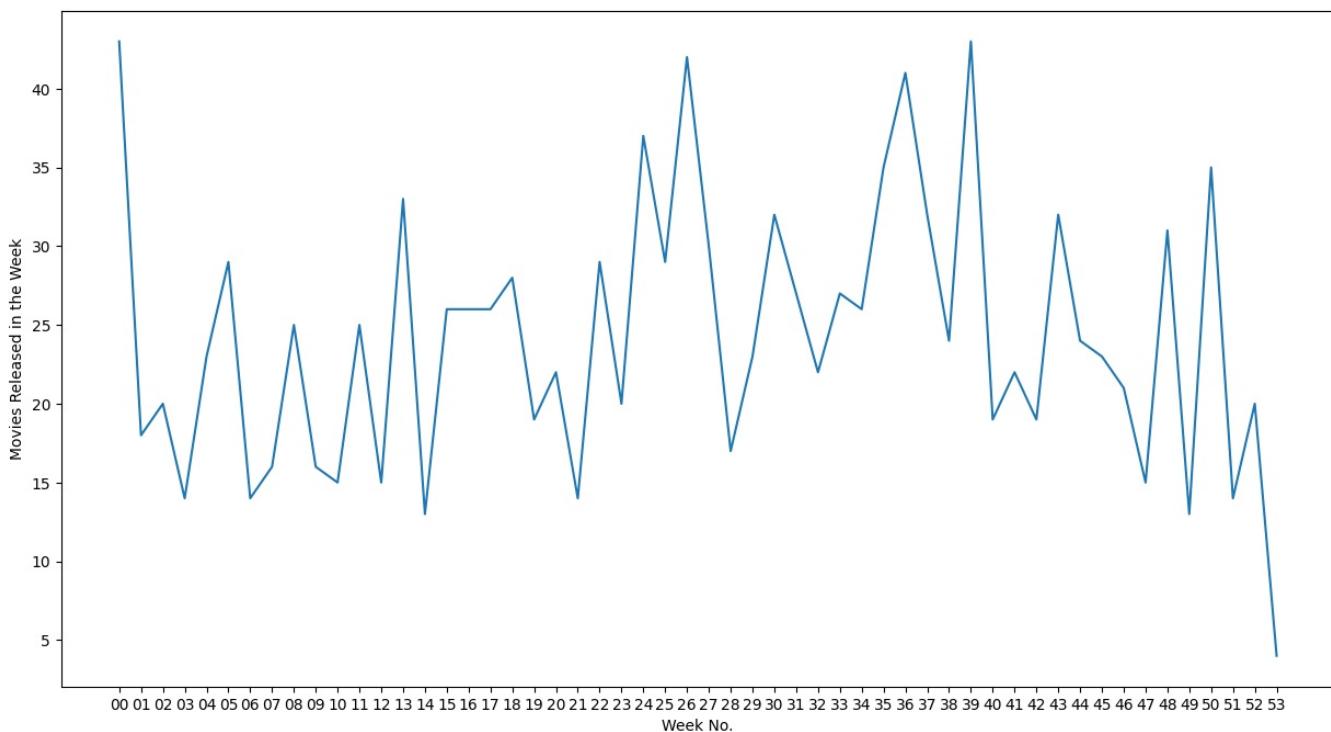


```
In [169]: df_year=df_usa_movies.groupby(['year']).agg({'title':'nunique'}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Shows Released in the Year")
plt.xlabel("Year")
plt.show()
```

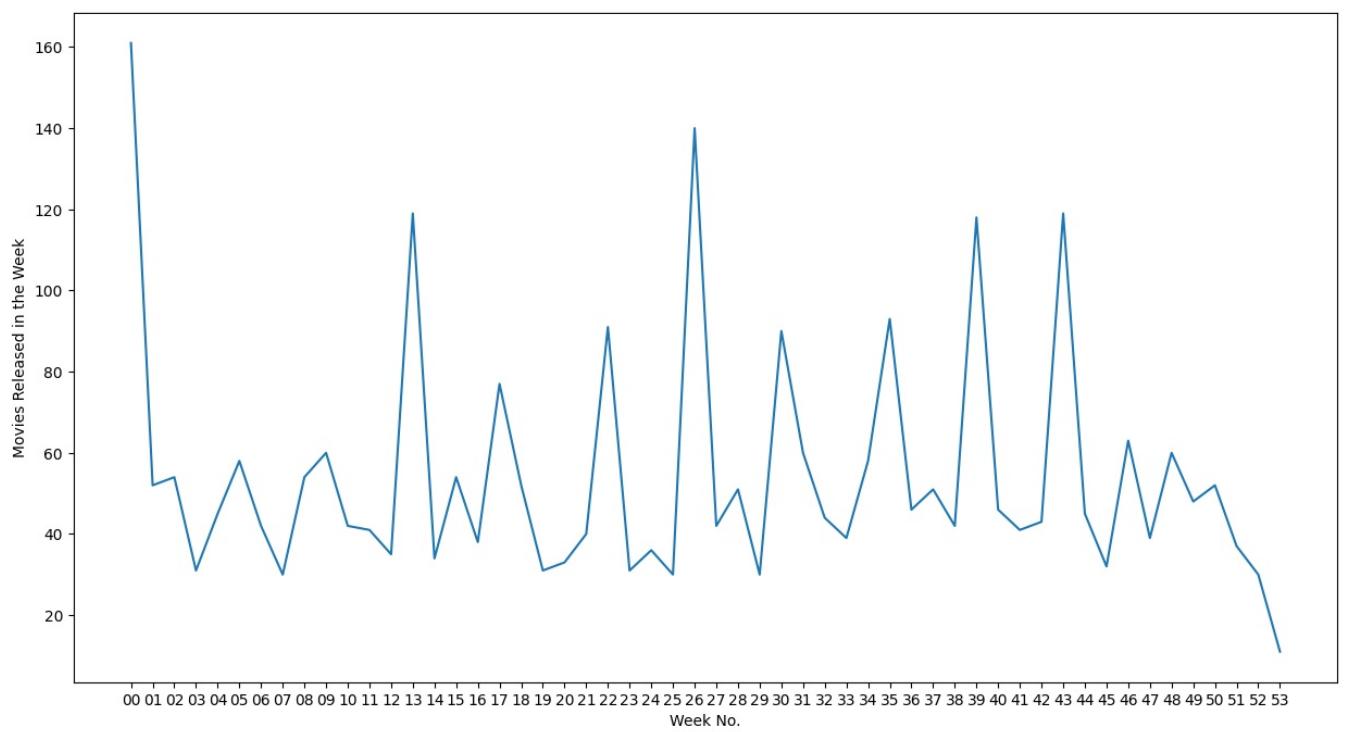


In USA, number of shows remained the same in 2021 as they were in 2020 while number of movies declined:

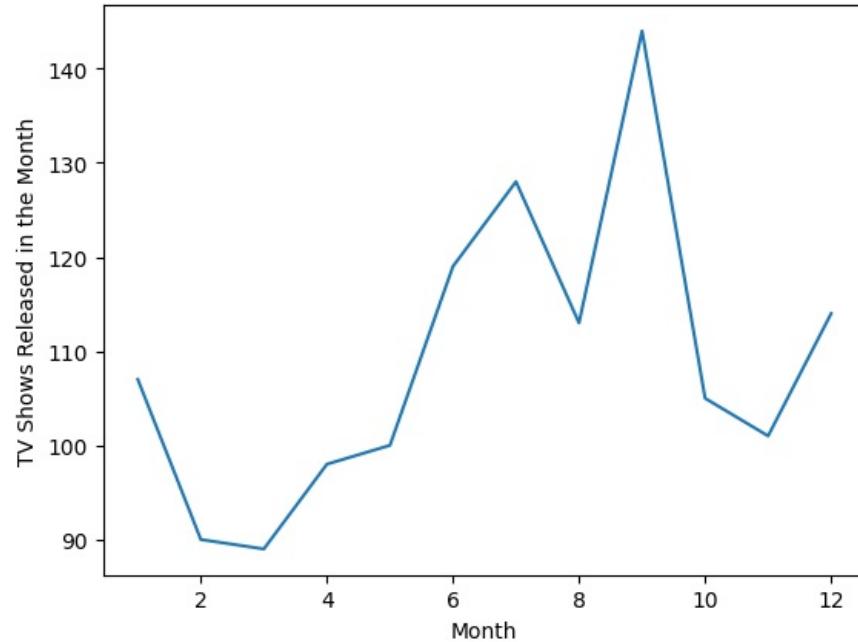
```
In [170]: df_week=df_usa_shows.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



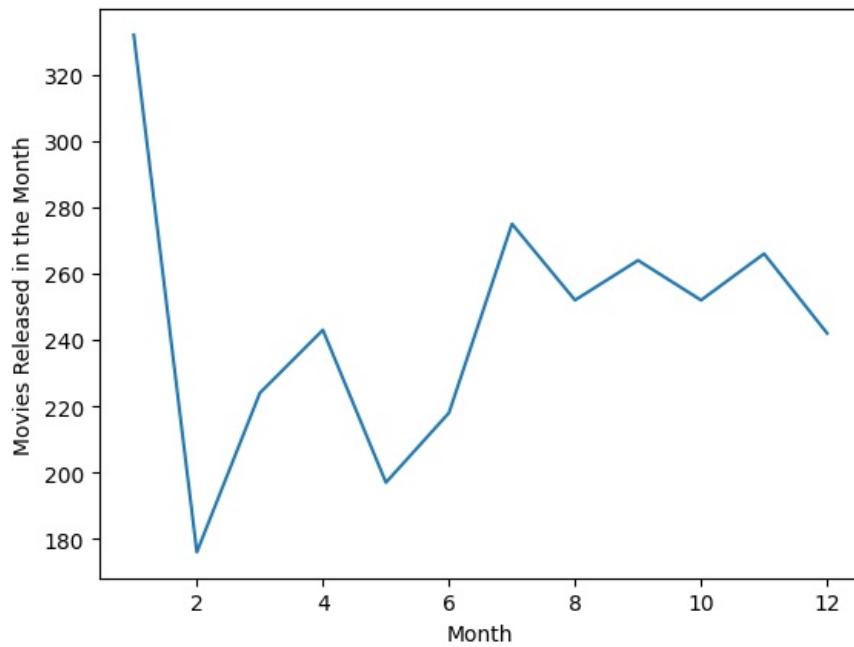
```
In [171]: df_week=df_usa_movies.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



```
In [172]: df_month=df_usa_shows.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("TV Shows Released in the Month")
plt.xlabel("Month")
plt.show()
```



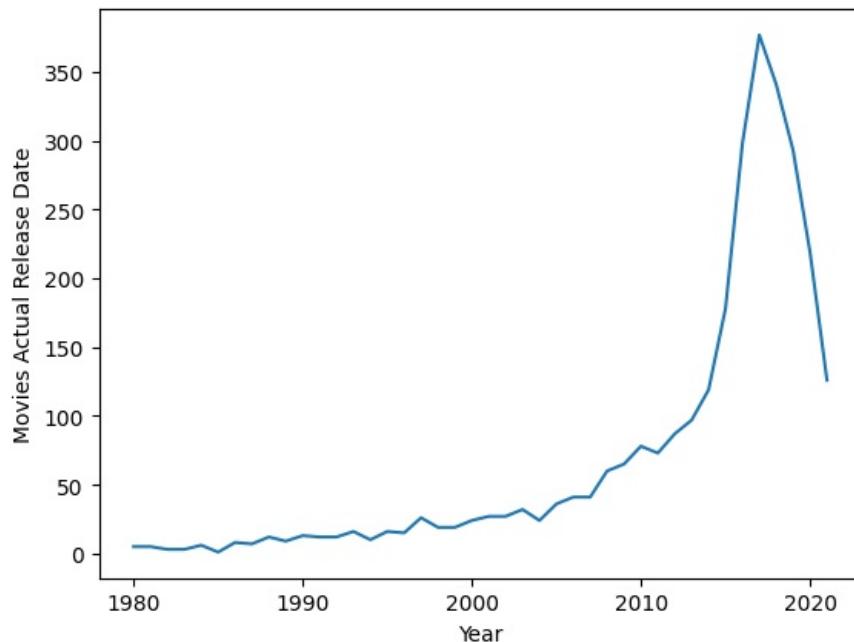
```
In [173]: df_month=df_usa_movies.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("Movies Released in the Month")
plt.xlabel("Month")
plt.show()
```



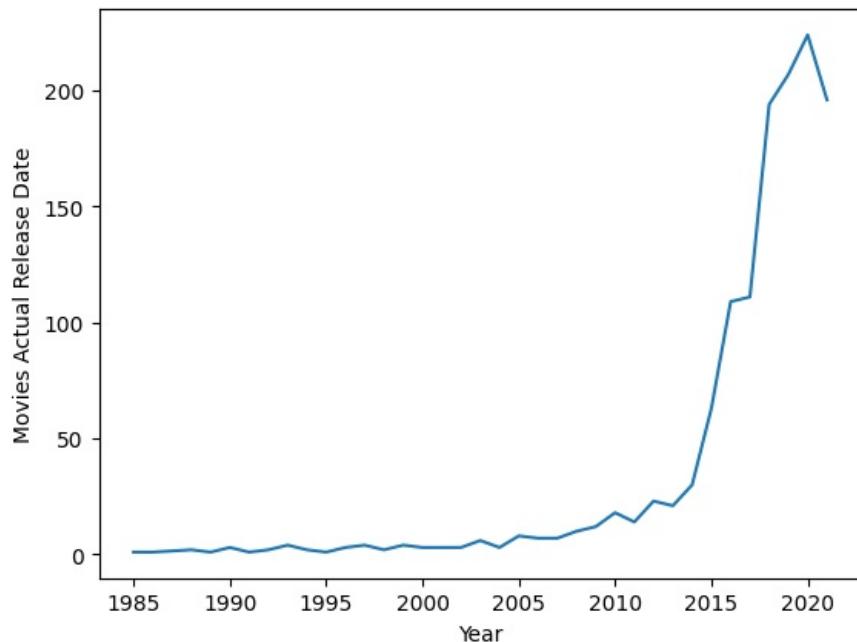
TV Shows are added in Netflix by a tremendous amount in July and September in USA

Movies are added in Netflix in USA by a tremendous amount in first week/last month of current year and first month of next year

```
In [174]: df_release_year=df_usa_movies[df_usa_movies['release_year']>=1980].groupby(['release_year']).agg({"title":"nunique"})
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



```
In [175]: df_release_year=df_usa_shows[df_usa_shows['release_year']>=1980].groupby(['release_year']).agg({"title":"nunique"})
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



In USA, though both Movies and Shows have reduced in 2021, the amount of decrease in number of TV Shows is small as compared to Movies

In [176]: df_usa_movies.head()

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	80-100	2021-09-
159	My Little Pony: A New Generation	Vanessa Hudgens	Robert Cullen	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-
160	My Little Pony: A New Generation	Vanessa Hudgens	José Luis Ucha	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-
161	My Little Pony: A New Generation	Kimiko Glenn	Robert Cullen	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-
162	My Little Pony: A New Generation	Kimiko Glenn	José Luis Ucha	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-

In [177]: #Analysing a combination of actors and directors

```
df_usa_movies['Actor_Director_Combination'] = df_usa_movies.Actors.str.cat(df_usa_movies.Directors, sep=' and ')
df_usa_movies_subset=df_usa_movies[df_usa_movies['Actors']!='Unknown Actor']
df_usa_movies_subset=df_usa_movies_subset[df_usa_movies_subset['Directors']!='Unknown Director']
df_usa_movies_subset.head()
```

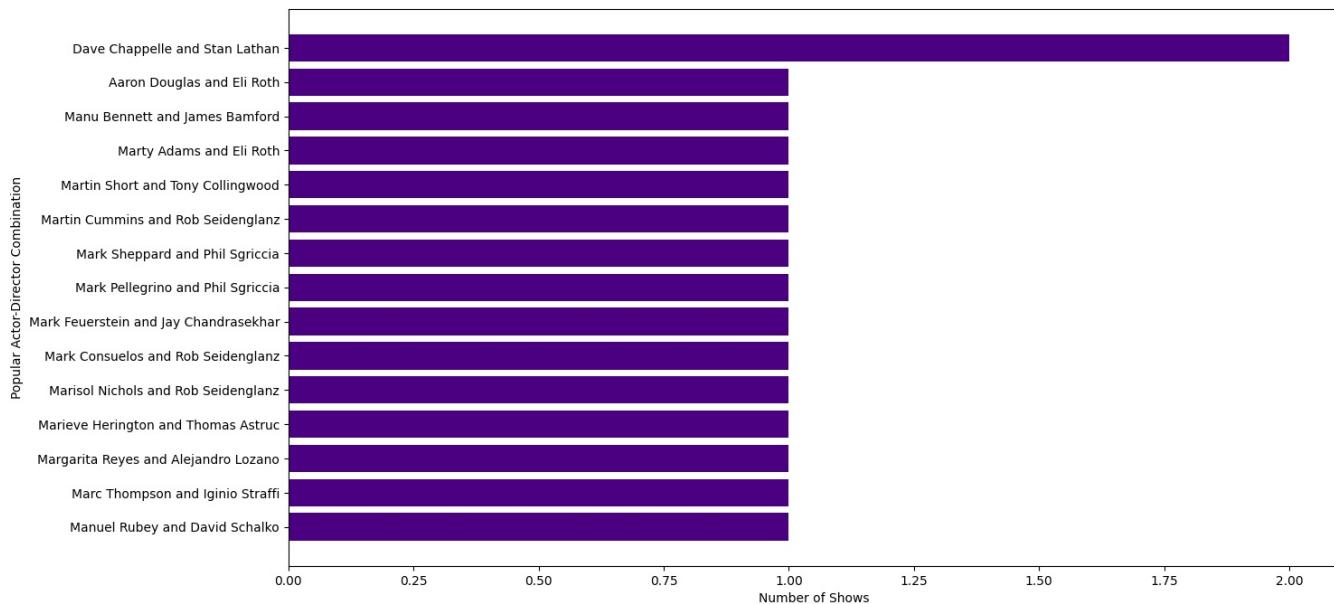
	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date	mc
159	My Little Pony: A New Generation	Vanessa Hudgens	Robert Cullen	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-24	
160	My Little Pony: A New Generation	Vanessa Hudgens	José Luis Ucha	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-24	
161	My Little Pony: A New Generation	Kimiko Glenn	Robert Cullen	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-24	
162	My Little Pony: A New Generation	Kimiko Glenn	José Luis Ucha	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-24	
163	My Little Pony: A New Generation	James Marsden	Robert Cullen	Children & Family Movies	United States	s7	Movie	September 24, 2021	2021	PG	80-100	2021-09-24	

```
In [178]: df_usa_shows['Actor_Director_Combination'] = df_usa_shows.Actors.str.cat(df_usa_shows.Directors, sep=' and ')
df_usa_shows_subset=df_usa_shows[df_usa_shows['Actors']!='Unknown Actor']
df_usa_shows_subset=df_usa_subset[df_usa_shows_subset['Directors']!='Unknown Director']
df_usa_shows_subset.head()
```

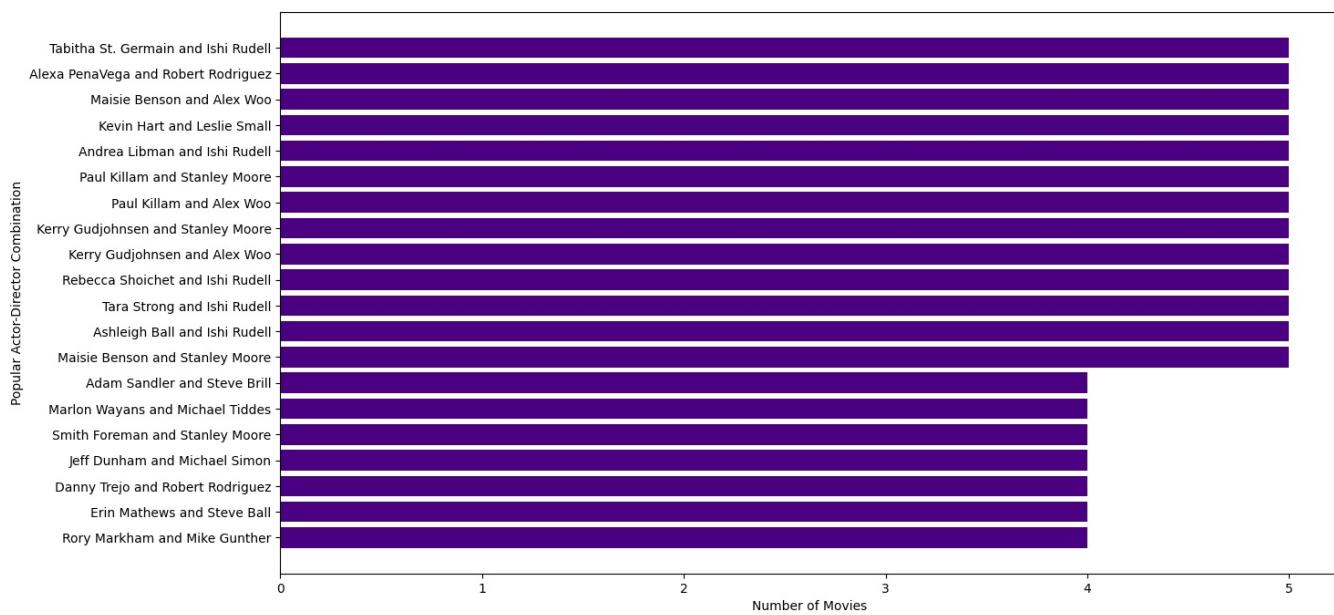
Out[178]:

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date	month
111	Midnight Mass	Kate Siegel	Mike Flanagan	TV Dramas	United States	s6	TV Show	September 24, 2021	2021	TV-MA	1 Season	2021-09-24	
112	Midnight Mass	Kate Siegel	Mike Flanagan	TV Horror	United States	s6	TV Show	September 24, 2021	2021	TV-MA	1 Season	2021-09-24	
113	Midnight Mass	Kate Siegel	Mike Flanagan	TV Mysteries	United States	s6	TV Show	September 24, 2021	2021	TV-MA	1 Season	2021-09-24	
114	Midnight Mass	Zach Gilford	Mike Flanagan	TV Dramas	United States	s6	TV Show	September 24, 2021	2021	TV-MA	1 Season	2021-09-24	
115	Midnight Mass	Zach Gilford	Mike Flanagan	TV Horror	United States	s6	TV Show	September 24, 2021	2021	TV-MA	1 Season	2021-09-24	

```
In [179]: df_actors_directors=df_usa_shows_subset.groupby(['Actor_Director_Combination']).agg({"title":"nunique"}).reset_
plt.figure(figsize=(15,8))
plt.barh(df_actors_directors[:::-1]['Actor_Director_Combination'], df_actors_directors[:::-1]['title'], color=['in
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actor-Director Combination')
plt.show()
```



```
In [180]: df_actors_directors=df_usa_movies_subset.groupby(['Actor_Director_Combination']).agg({"title":"nunique"}).reset_
plt.figure(figsize=(15,8))
plt.barh(df_actors_directors[:::-1]['Actor_Director_Combination'], df_actors_directors[:::-1]['title'], color=['in
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actor-Director Combination')
plt.show()
```



```
In [181]: df_actors_directors[:::-1]['Actor_Director_Combination'].values
```

```
Out[181]: array(['Rory Markham and Mike Gunther', 'Erin Mathews and Steve Ball',
   'Danny Trejo and Robert Rodriguez',
   'Jeff Dunham and Michael Simon', 'Smith Foreman and Stanley Moore',
   'Marlon Wayans and Michael Tiddes', 'Adam Sandler and Steve Brill',
   'Maisie Benson and Stanley Moore', 'Ashleigh Ball and Ishi Rudell',
   'Tara Strong and Ishi Rudell', 'Rebecca Shoichet and Ishi Rudell',
   'Kerry Gudjohnsen and Alex Woo',
   'Kerry Gudjohnsen and Stanley Moore', 'Paul Killam and Alex Woo',
   'Paul Killam and Stanley Moore', 'Andrea Libman and Ishi Rudell',
   'Kevin Hart and Leslie Small', 'Maisie Benson and Alex Woo',
   'Alexa PenaVega and Robert Rodriguez',
   'Tabitha St. Germain and Ishi Rudell'], dtype=object)
```

The Most Popular Actor Director Combination in Movies Across USA are:-

'Smith Foreman and Stanley Moore',
 'Marlon Wayans and Michael Tiddes',
 'Adam Sandler and Steve Brill',
 'Maisie Benson and Stanley Moore',
 'Ashleigh Ball and Ishi Rudell',
 'Tara Strong and Ishi Rudell',
 'Rebecca Shoichet and Ishi Rudell',
 'Kerry Gudjohnsen and Alex Woo',
 'Kerry Gudjohnsen and Stanley Moore',
 'Paul Killam and Alex Woo',
 'Paul Killam and Stanley Moore',
 'Andrea Libman and Ishi Rudell',
 'Kevin Hart and Leslie Small',
 'Maisie Benson and Alex Woo',
 'Alexa PenaVega and Robert Rodriguez',
 'Tabitha St. Germain and Ishi Rudell'

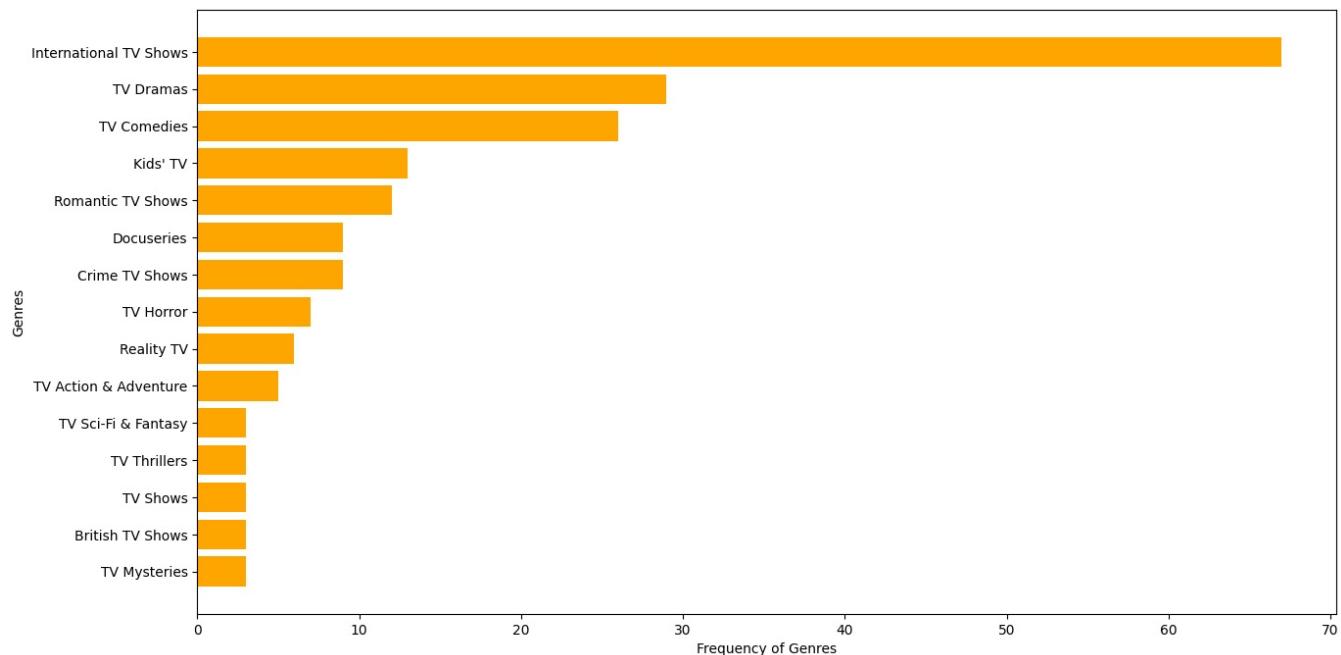
The Second Most Popular Actor Director Combination in Movies Across USA are:-

'Rory Markham and Mike Gunther',
 'Erin Mathews and Steve Ball',
 'Danny Trejo and Robert Rodriguez',
 'Jeff Dunham and Michael Simon'

Univariate Analysis separately for shows and movies in India

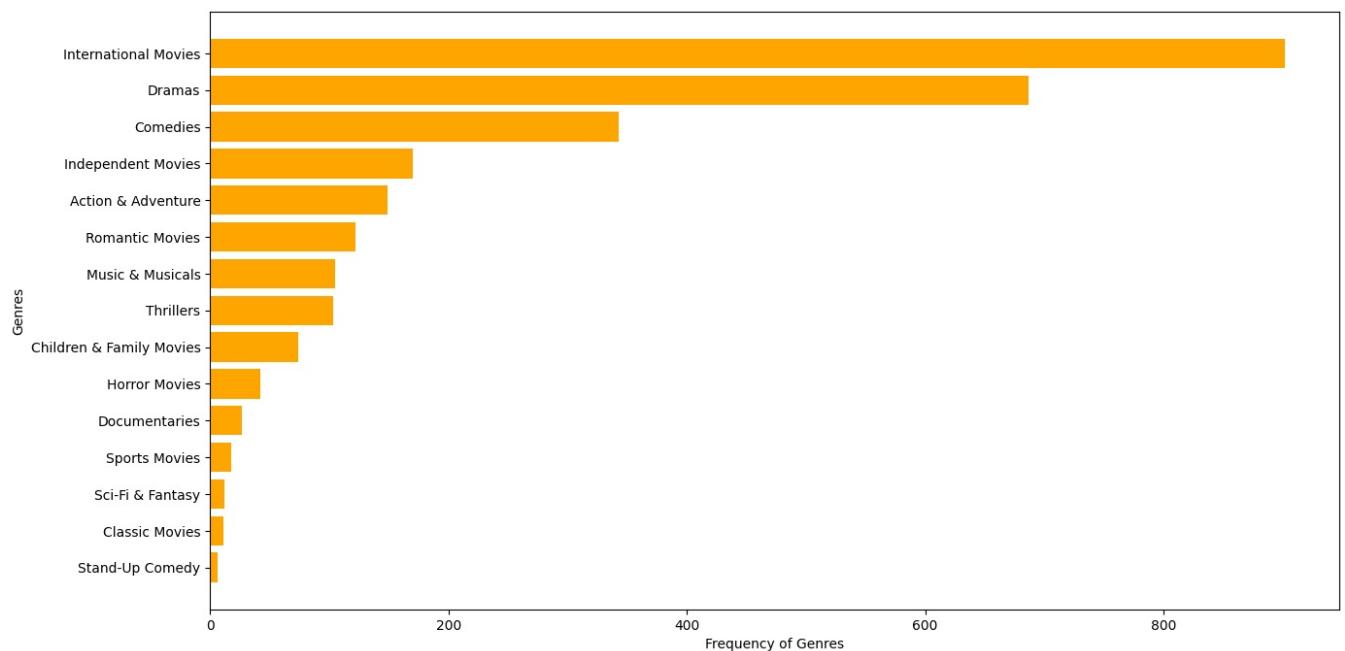
```
In [182... #Analyzing India for both shows and movies
df_india_shows=df_final1[df_final1['country']=='India'][df_final1[df_final1['country']=='India'][['type']=='TV Shows']]
df_india_movies=df_final1[df_final1['country']=='India'][df_final1[df_final1['country']=='India'][['type']=='Movies']]
```

```
In [183... df_genre=df_india_shows.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_genre[:::-1]['Genre'], df_genre[:::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```



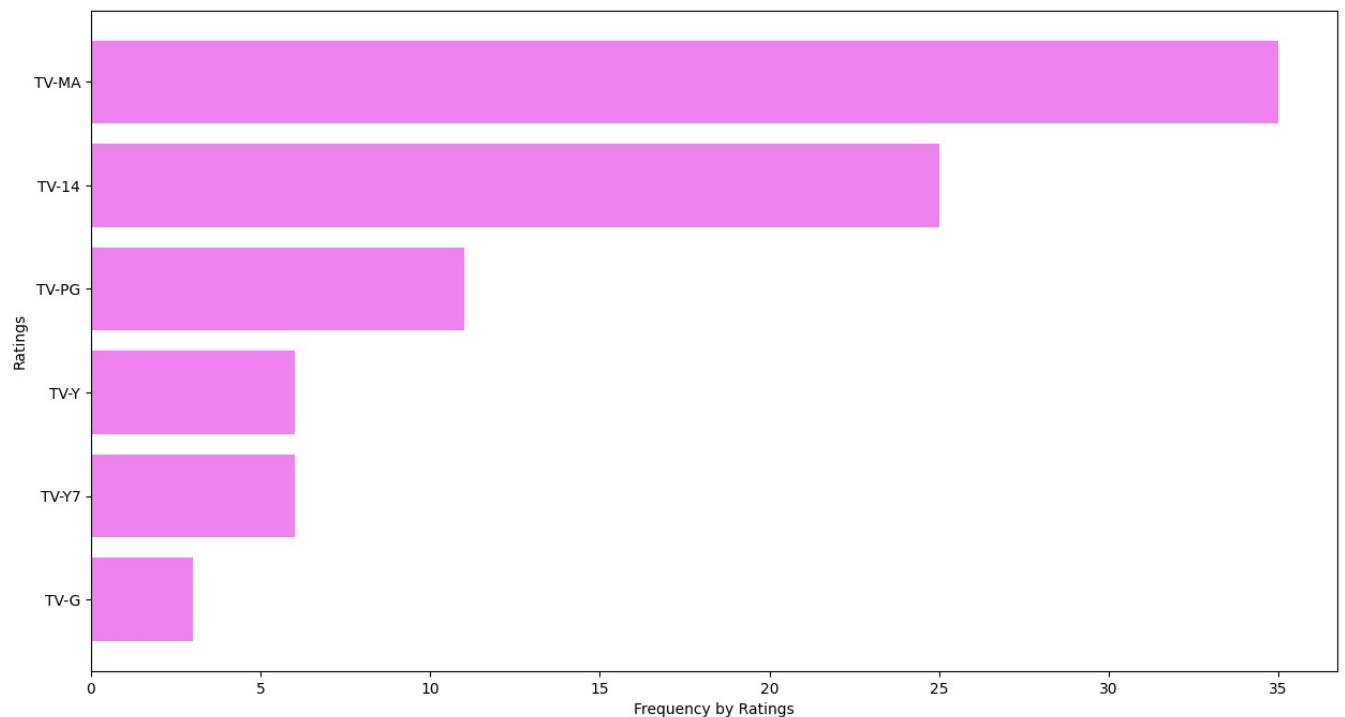
Dramas, Comedy, Kids 'TV Shows and International TV Shows Genres are popular in TV Series in India

```
In [184]: df_genre=df_india_movies.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_genre[::1]['Genre'], df_genre[::1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```

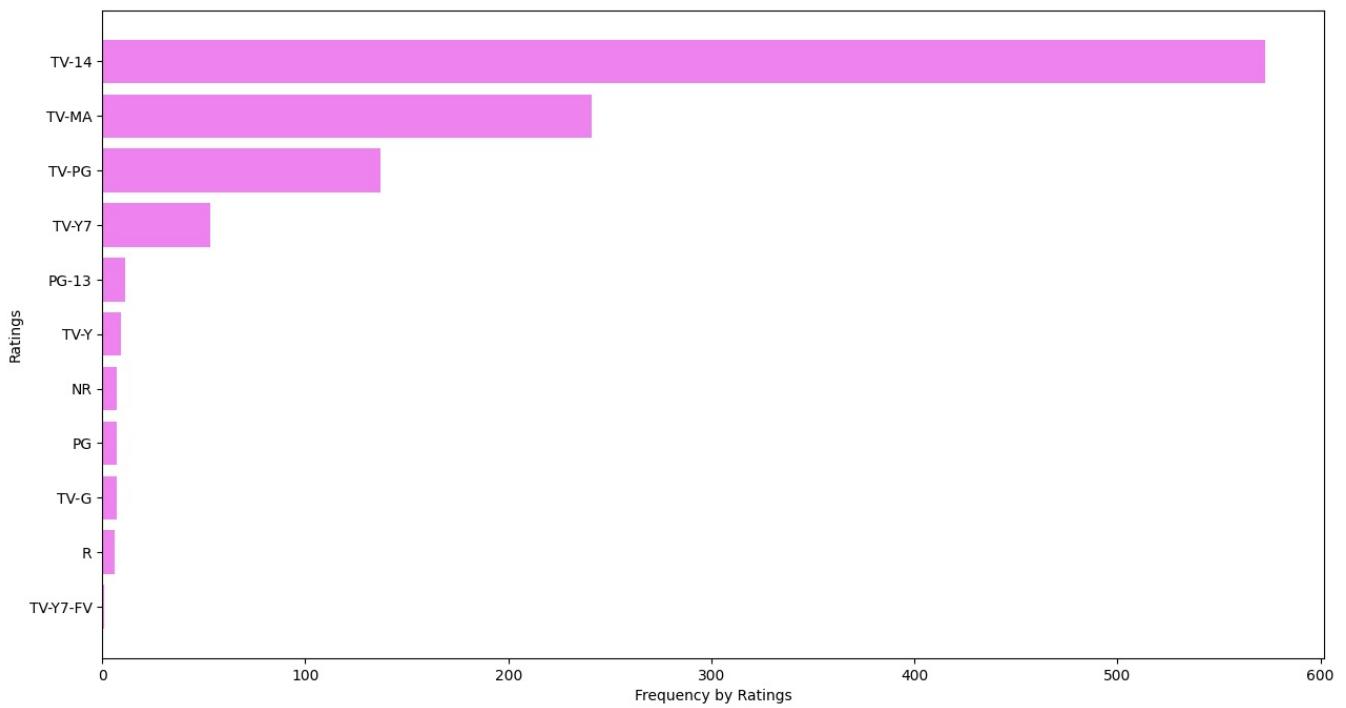


International Movies, Drama, Comedy, Independent Movies and Action, Romance Genres are prevalent in India

```
In [185]: df_rating=df_india_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_rating[::1]['rating'], df_rating[::1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



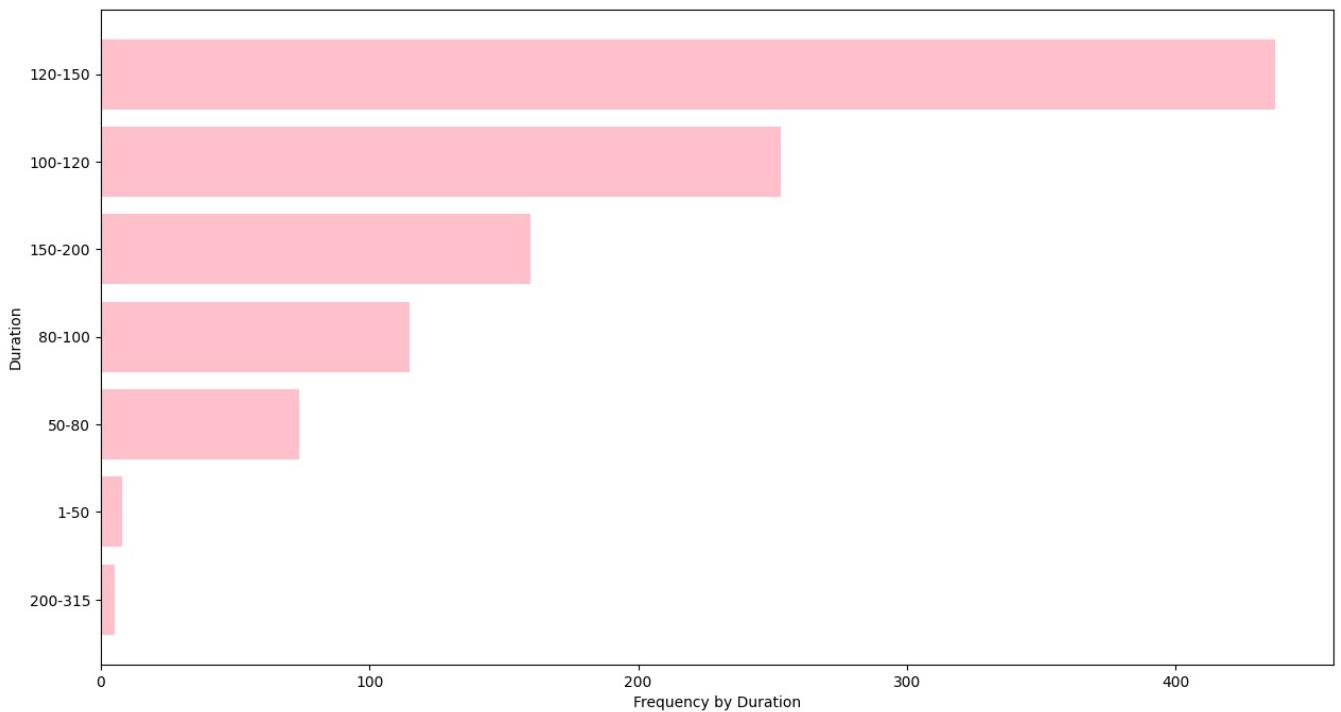
```
In [186]: df_rating=df_india_movies.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_rating[::1]['rating'], df_rating[::1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



So it seems plausible to conclude that the popular ratings across Netflix includes Mature Audiences in TV Shows and those appropriate for people over 14 in Movies in India.

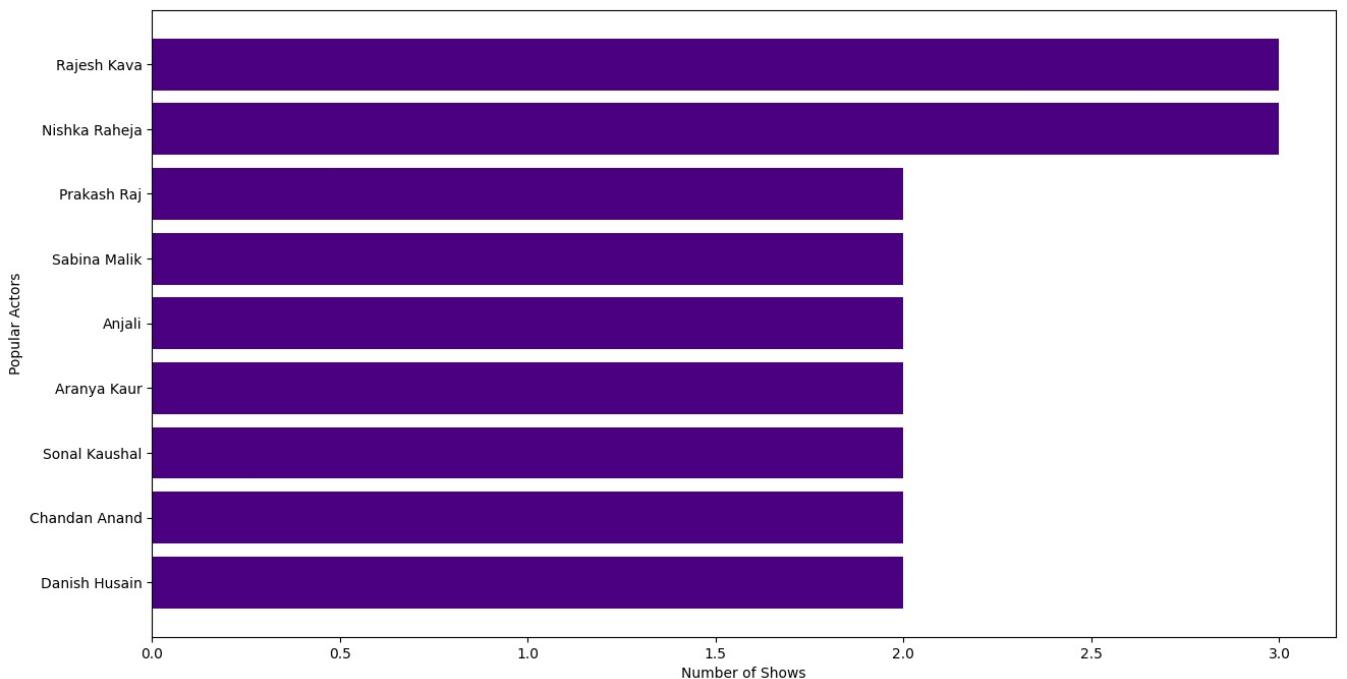
Now this indeed seems to be the case. Indian TV Shows in Netflix are without a shadow of doubt intended for Mature Audiences while Movies for over 14 years of age.

```
In [187]: df_duration=df_india_movies.groupby(['duration']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'])
plt.figure(figsize=(15,8))
plt.bahr(df_duration[::-1]['duration'], df_duration[::-1]['title'], color=['pink'])
plt.xlabel('Frequency by Duration')
plt.ylabel('Duration')
plt.show()
```



Across movies ranges of minutes in India are comparatively greater than USA with a sweet spot at 120-150 mins.

```
In [188]: df_actors=df_india_shows.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.bahr(df_actors[::-1]['Actors'], df_actors[::-1]['title'], color=['indigo'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actors')
plt.show()
```



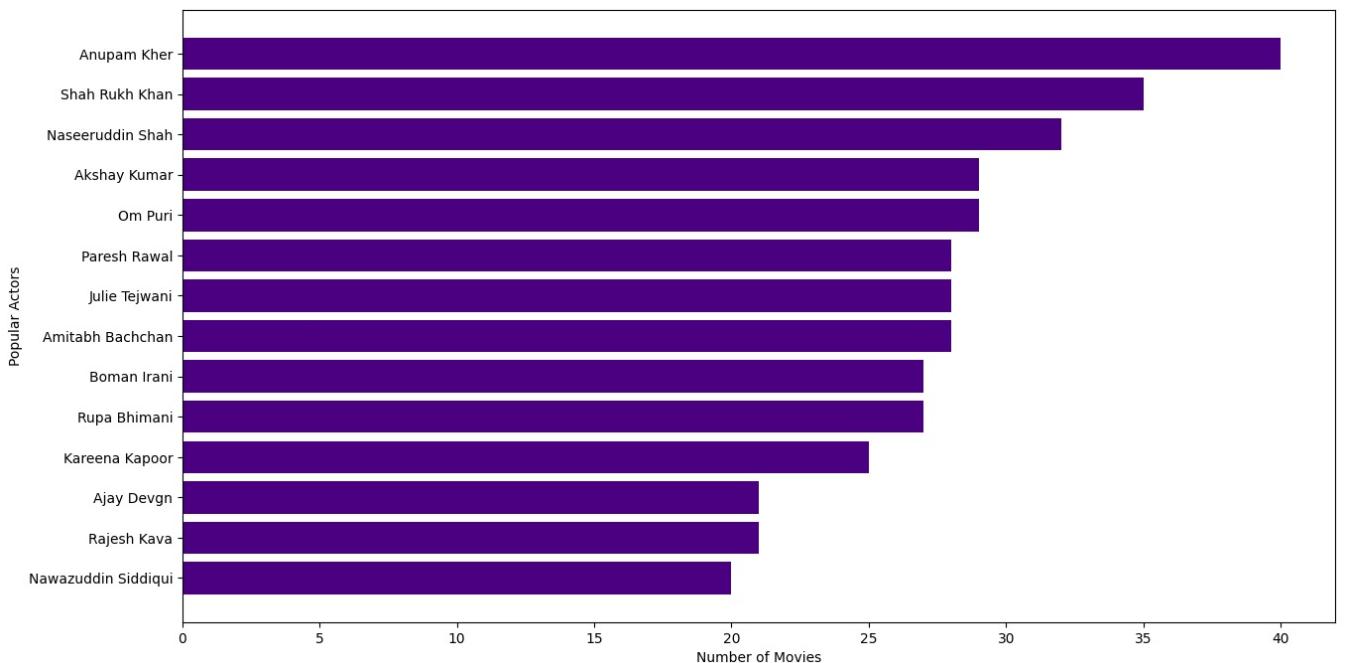
```
In [189]: df_actors['Actors'].values
```

```
Out[189]: array(['Rajesh Kava', 'Nishka Raheja', 'Prakash Raj', 'Sabina Malik',
   'Anjali', 'Aranya Kaur', 'Sonal Kaushal', 'Chandan Anand',
   'Danish Husain'], dtype=object)
```

Popular Actors in TV Shows in India are:-

- 'Rajesh Kava',
- 'Nishka Raheja',
- 'Prakash Raj',
- 'Sabina Malik',
- 'Anjali',
- 'Aranya Kaur',
- 'Sonal Kaushal',
- 'Chandan Anand',
- 'Danish Husain'

```
In [190]: df_actors=df_india_movies.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barrh(df_actors[:10]['Actors'], df_actors[:10]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actors')
plt.show()
```



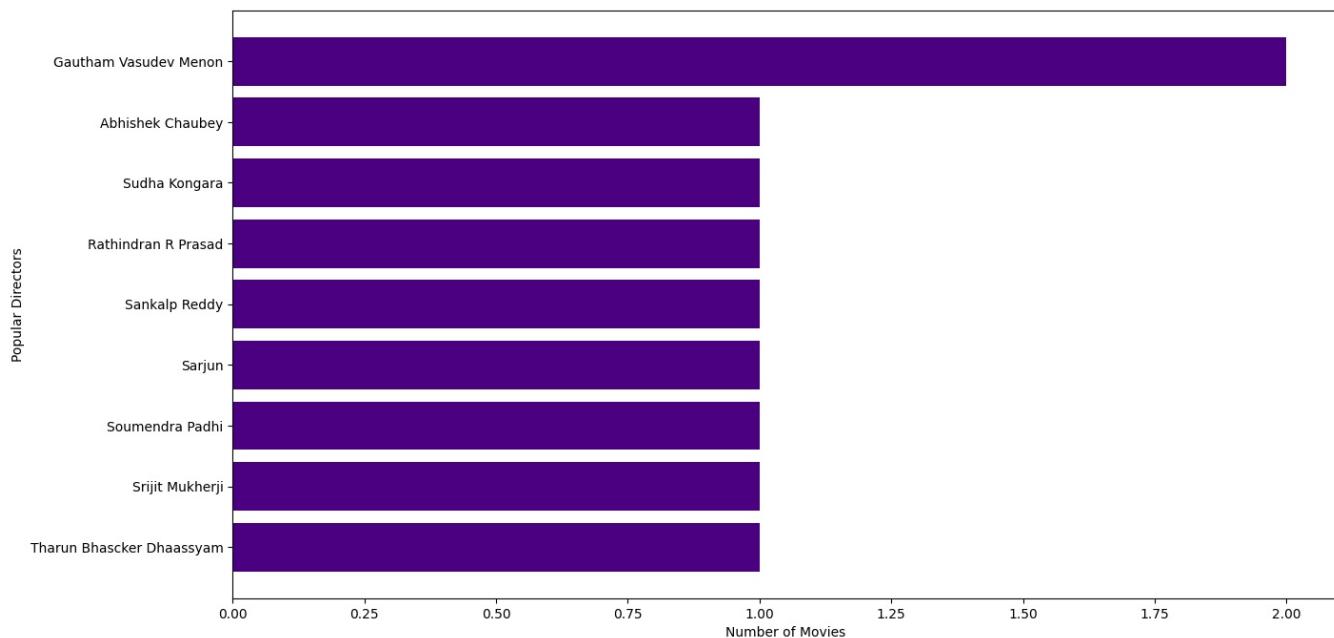
```
In [191]: df_actors['Actors'].values
```

```
Out[191]: array(['Anupam Kher', 'Shah Rukh Khan', 'Naseeruddin Shah',
   'Akshay Kumar', 'Om Puri', 'Paresh Rawal', 'Julie Tejwani',
   'Amitabh Bachchan', 'Boman Irani', 'Rupa Bhimani',
   'Kareena Kapoor', 'Ajay Devgn', 'Rajesh Kava',
   'Nawazuddin Siddiqui'], dtype=object)
```

Popular actors across Movies in India:-

```
'Anupam Kher',
'Shah Rukh Khan',
'Naseeruddin Shah',
'Akshay Kumar',
'Om Puri',
'Paresh Rawal',
'Julie Tejwani',
'Amitabh Bachchan',
'Boman Irani',
'Rupa Bhimani',
'Kareena Kapoor',
'Ajay Devgn',
'Rajesh Kava',
'Kay Kay Menon'
```

```
In [192... df_directors=df_india_shows.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['tit
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barrh(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



```
In [193... df_directors['Directors'].values
```

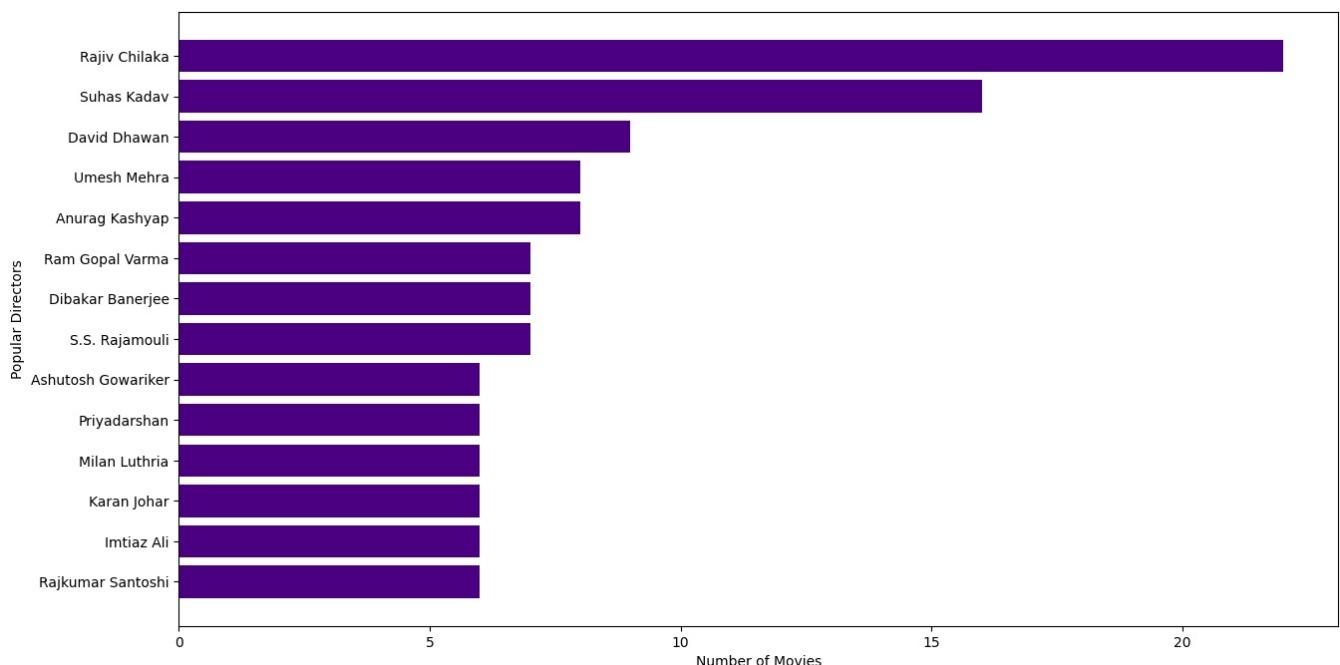
```
Out[193]: array(['Gautham Vasudev Menon', 'Abhishek Chaubey', 'Sudha Kongara',
   'Rathindran R Prasad', 'Sankalp Reddy', 'Sarjun',
   'Soumendra Padhi', 'Srijiit Mukherji', 'Tharun Bhascker Dhaassyam'],
  dtype=object)
```

Popular Directors Across Movies in India:-

```
'Gautham Vasudev Menon',
'Abhishek Chaubey',
'Sudha Kongara',
'Rathindran R Prasad',
'Sankalp Reddy',
'Sarjun',
'Soumendra Padhi',
'Srijiit Mukherji',
'Tharun Bhascker Dhaassyam'
```

```
In [194... df_directors=df_india_movies.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['tit
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barrh(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
```

```
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



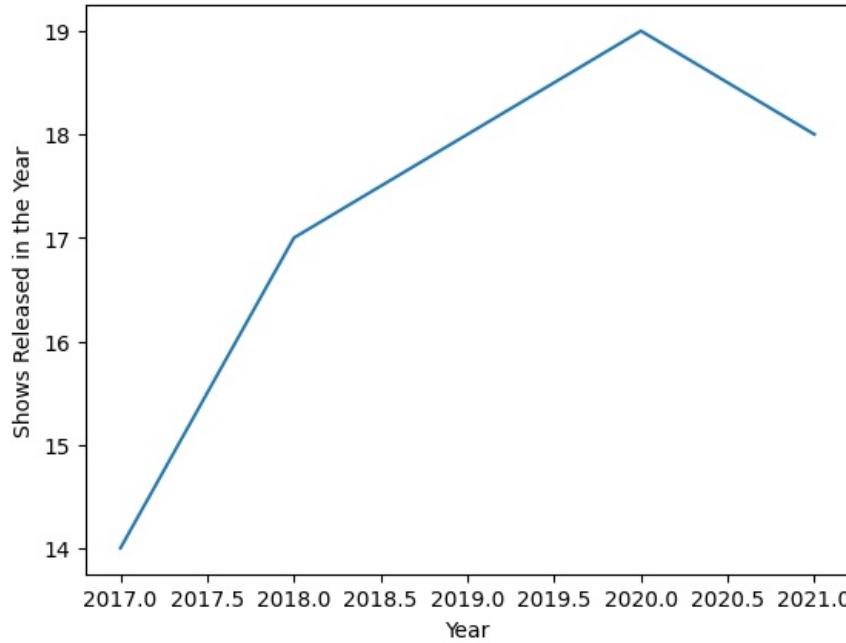
```
In [195]: df_directors['Directors'].values
```

```
Out[195]: array(['Rajiv Chilaka', 'Suhas Kadav', 'David Dhawan', 'Umesh Mehra',
 'Anurag Kashyap', 'Ram Gopal Varma', 'Dibakar Banerjee',
 'S.S. Rajamouli', 'Ashutosh Gowariker', 'Priyadarshan',
 'Milan Luthria', 'Karan Johar', 'Imtiaz Ali', 'Rajkumar Santoshi'],
 dtype=object)
```

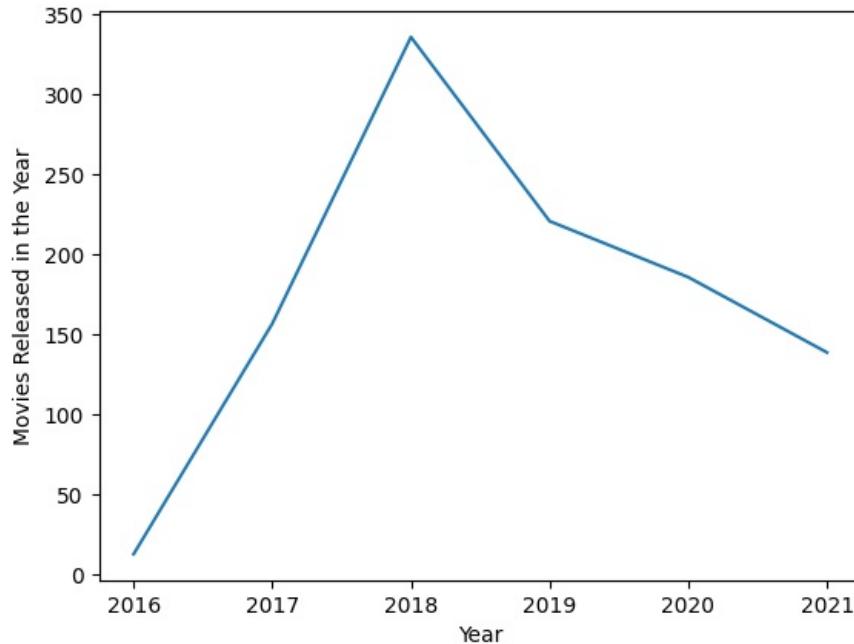
Popular directors across movies in India:-

'Rajiv Chilaka',
'Suhas Kadav',
'David Dhawan',
'Umesh Mehra',
'Anurag Kashyap',
'Ram Gopal Varma',
'Dibakar Banerjee',
'Zoya Akhtar',
'Tilak Shetty',
'Rajkumar Santoshi',
'Priyadarshan',
'Sooraj R. Barjatya',
'Ashutosh Gowariker',
'Milan Luthria'

```
In [196]: df_year=df_india_shows.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Shows Released in the Year")
plt.xlabel("Year")
plt.show()
```



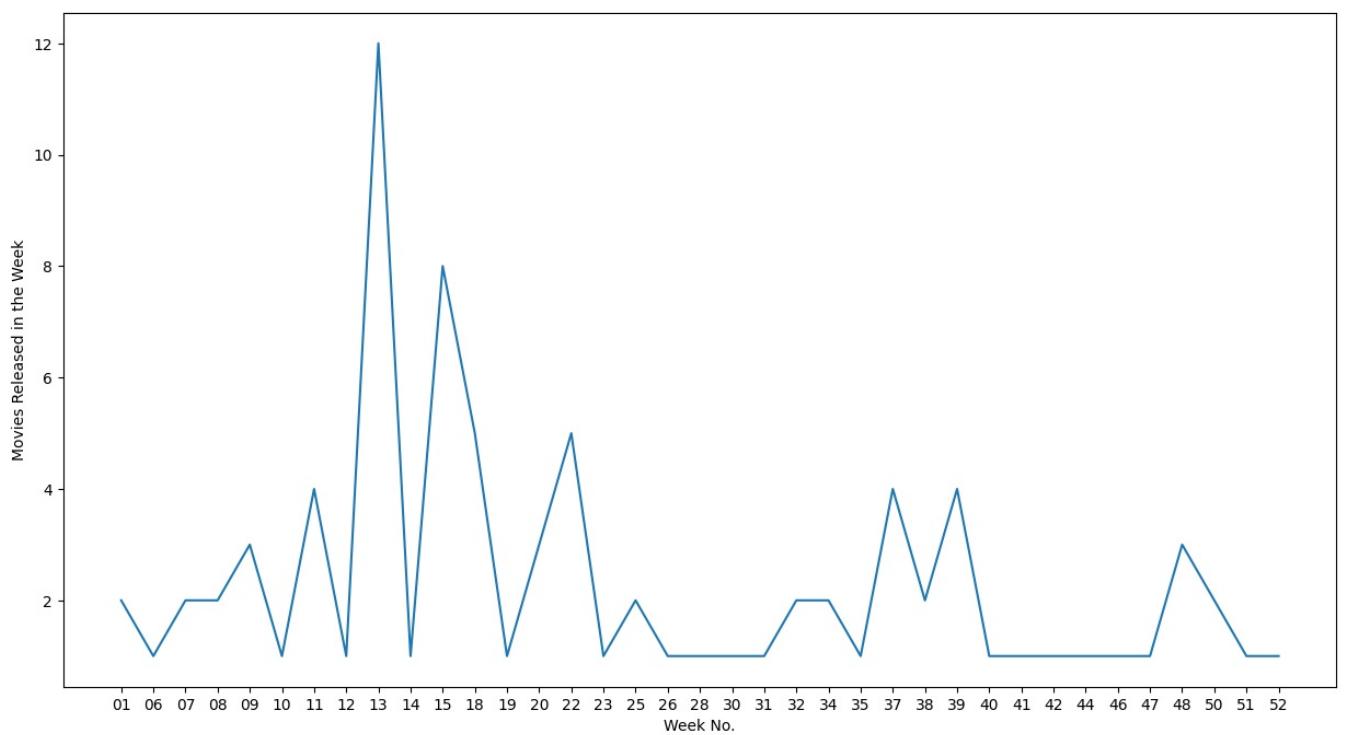
```
In [197]: df_year=df_india_movies.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



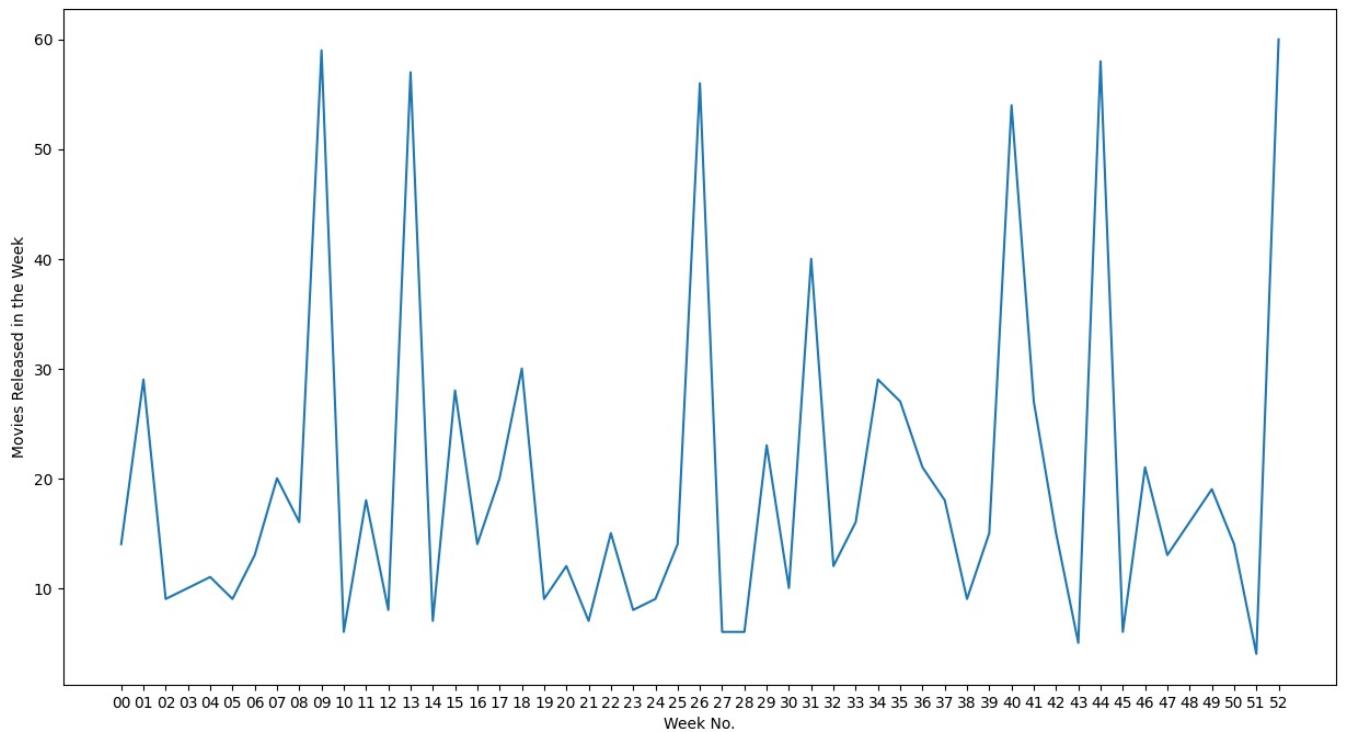
In India, TV Shows were increasingly being added till 2020, though the addition of shows reduced in 2021.

In India, Movies were increasingly added till 2018 but it has been a huge downhill since then. Now that's preposterous, since and something has to be recommended to the Netflix Team with regards to that.

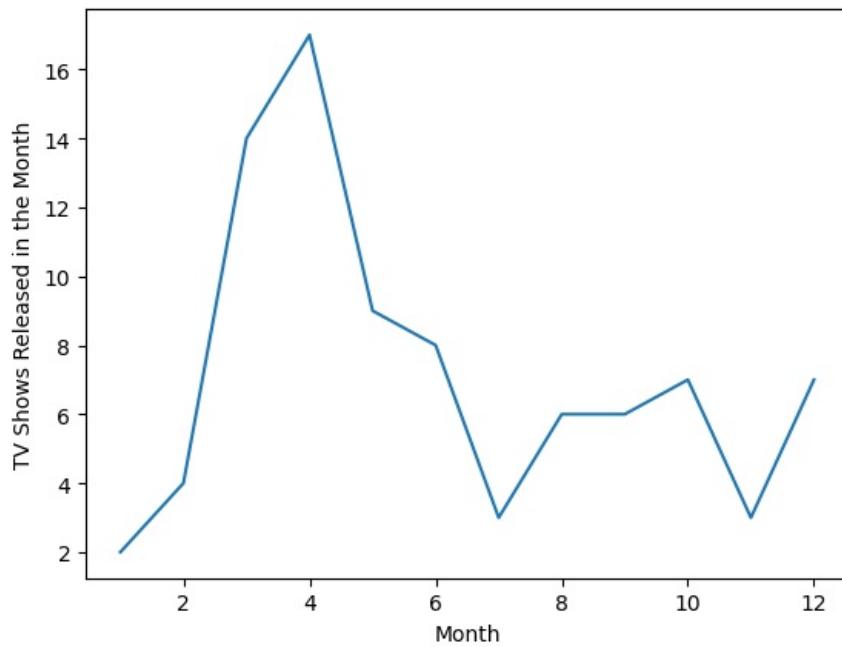
```
In [198]: df_week=df_india_shows.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



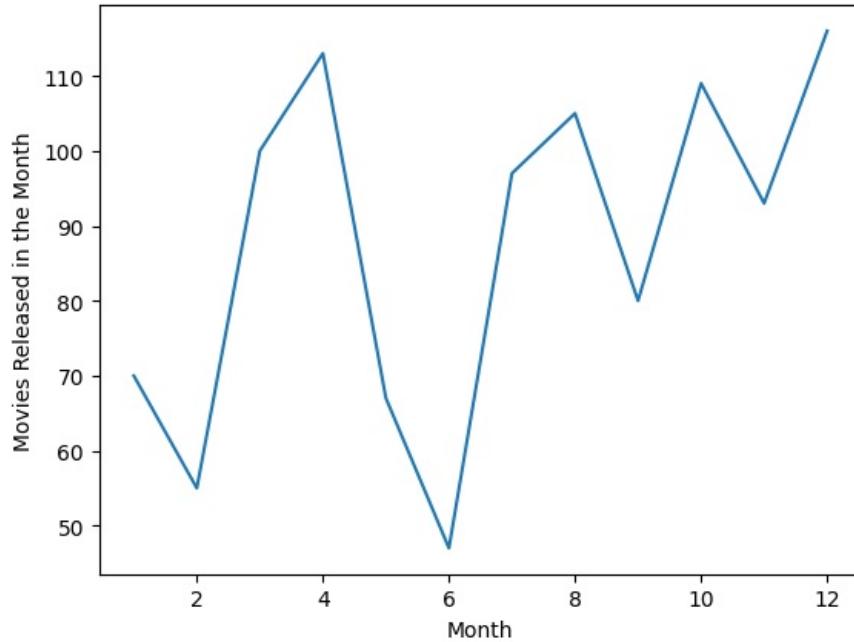
```
In [199]: df_week=df_india_movies.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



```
In [200]: df_month=df_india_shows.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("TV Shows Released in the Month")
plt.xlabel("Month")
plt.show()
```



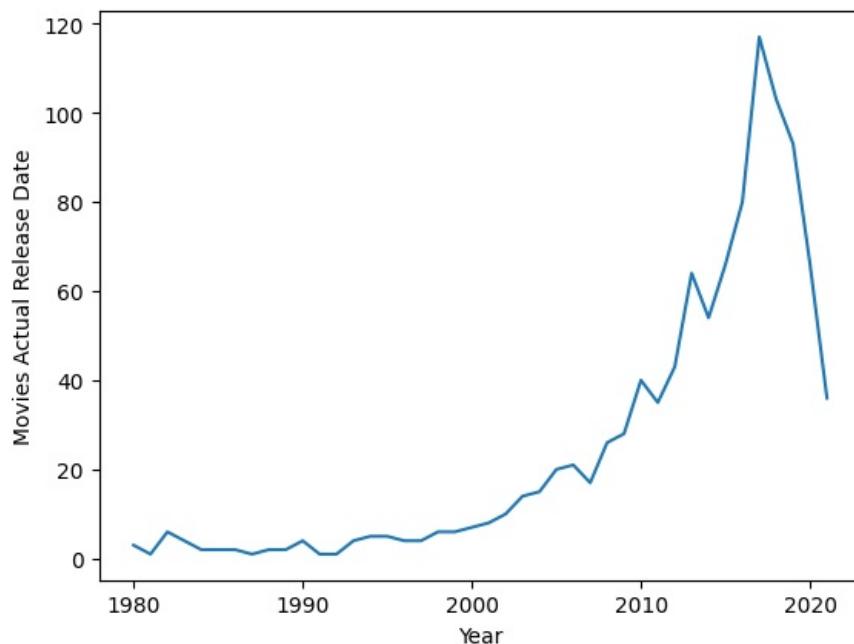
```
In [201]: df_month=df_india_movies.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("Movies Released in the Month")
plt.xlabel("Month")
plt.show()
```



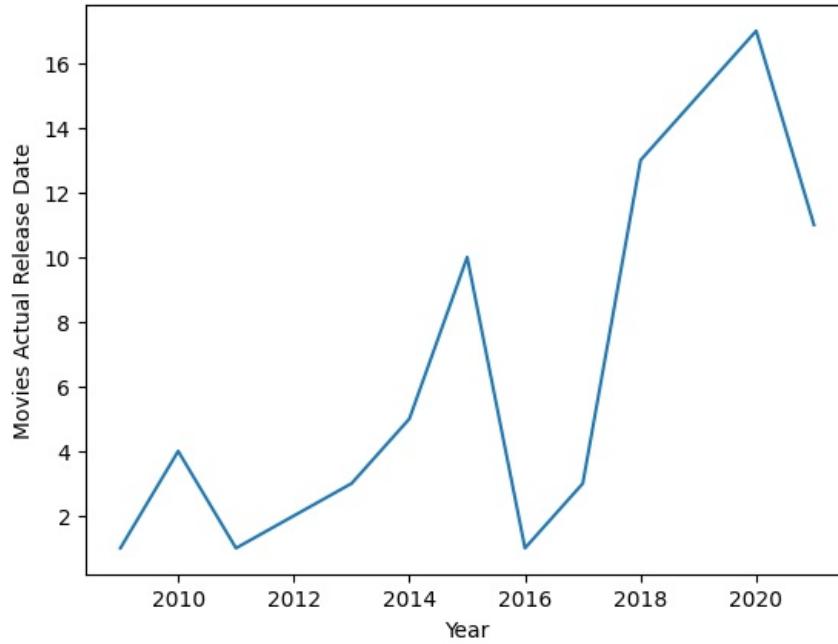
TV Shows are added in Netflix by a tremendous amount in April in India

Movies are added in Netflix in India by a tremendous amount in first week/last month of current year and first month of next year

```
In [202]: df_release_year=df_india_movies[df_india_movies['release_year']>=1980].groupby(['release_year']).agg({"title":"nunique"})
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



```
In [203]: df_release_year=df_india_shows[df_india_shows['release_year']>=1980].groupby(['release_year']).agg({"title": "count"})
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



The understandable trend amongs movies and TV Shows across India in Netflix is the reduction of movies after 2020

```
In [204]: #Analysing a combination of actors and directors
df_india_movies['Actor_Director_Combination'] = df_india_movies.Actors.str.cat(df_india_movies.Directors, sep=' and ')
df_india_movies_subset=df_india_movies[df_india_movies['Actors']!='Unknown Actor']
df_india_movies_subset=df_india_movies_subset[df_india_movies_subset['Directors']!='Unknown Director']
df_india_movies_subset.head()
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_A
621	Avvai Shanmughi	Kamal Hassan	K.S. Ravikumar	Comedies	India	s23	Movie	September 21, 2021	1996	TV-PG	150-200	
622	Avvai Shanmughi	Kamal Hassan	K.S. Ravikumar	International Movies	India	s23	Movie	September 21, 2021	1996	TV-PG	150-200	
629	Avvai Shanmughi	Nassar	K.S. Ravikumar	Comedies	India	s23	Movie	September 21, 2021	1996	TV-PG	150-200	
630	Avvai Shanmughi	Nassar	K.S. Ravikumar	International Movies	India	s23	Movie	September 21, 2021	1996	TV-PG	150-200	
631	Avvai Shanmughi	S.P. Balasubrahmanyam	K.S. Ravikumar	Comedies	India	s23	Movie	September 21, 2021	1996	TV-PG	150-200	

```
In [205]: df_india_shows['Actor_Director_Combination'] = df_india_shows.Actors.str.cat(df_india_shows.Directors, sep=' and ')
```

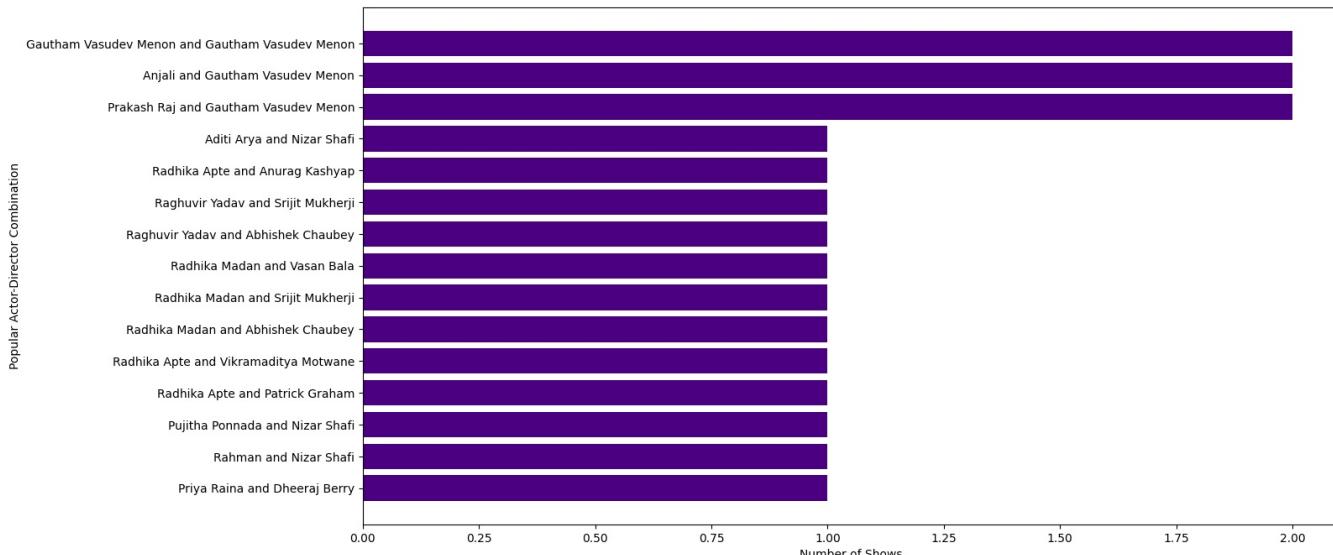
```
df_india_shows_subset=df_india_shows[df_india_shows['Actors']!='Unknown Actor']
df_india_shows_subset=df_india_shows_subset[df_india_shows_subset['Directors']!='Unknown Director']
df_india_shows_subset.head(2)
```

Out[205]:

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date	mo
7005	Navarasa	Suriya	Bejoy Nambiar	TV Shows	India	s298	TV Show	August 6, 2021	2021	TV-MA	1 Season	2021-08-06	
7006	Navarasa	Suriya	Priyadarshan	TV Shows	India	s298	TV Show	August 6, 2021	2021	TV-MA	1 Season	2021-08-06	

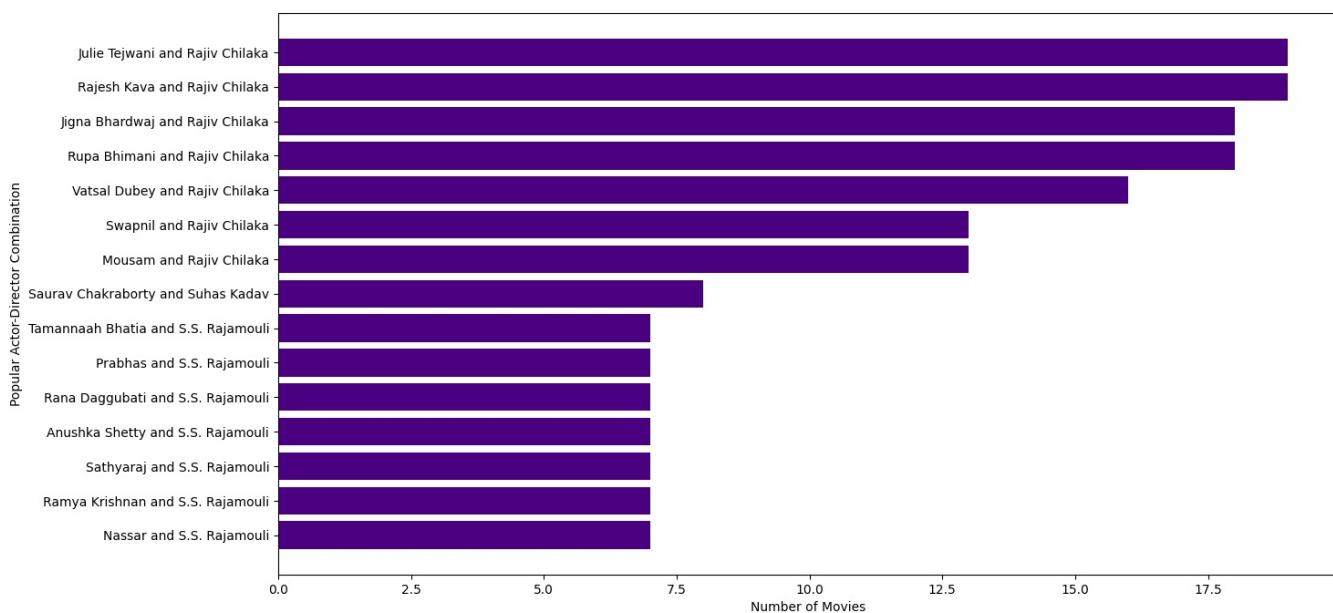
In [206]:

```
df_actors_directors=df_india_shows_subset.groupby(['Actor_Director_Combination']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
plt.barh(df_actors_directors[:::-1]['Actor_Director_Combination'], df_actors_directors[:::-1]['title'], color=['in'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actor-Director Combination')
plt.show()
```



In [207]:

```
df_actors_directors=df_india_movies_subset.groupby(['Actor_Director_Combination']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
plt.barh(df_actors_directors[:::-1]['Actor_Director_Combination'], df_actors_directors[:::-1]['title'], color=['in'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actor-Director Combination')
plt.show()
```



In [208]:

```
df_india_movies[df_india_movies['Directors']=='Rajiv Chilaka'].head(5)
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date	mc
10058	Chhota Bheem - Neeli Pahaadi	Vatsal Dubey	Rajiv Chilaka	Children & Family Movies	India	s407	Movie	July 22, 2021	2013	TV-Y7	50-80	2021-07-22	
10059	Chhota Bheem - Neeli Pahaadi	Julie Tejwani	Rajiv Chilaka	Children & Family Movies	India	s407	Movie	July 22, 2021	2013	TV-Y7	50-80	2021-07-22	
10060	Chhota Bheem - Neeli Pahaadi	Rupa Bhimani	Rajiv Chilaka	Children & Family Movies	India	s407	Movie	July 22, 2021	2013	TV-Y7	50-80	2021-07-22	
10061	Chhota Bheem - Neeli Pahaadi	Jigna Bhardwaj	Rajiv Chilaka	Children & Family Movies	India	s407	Movie	July 22, 2021	2013	TV-Y7	50-80	2021-07-22	
10062	Chhota Bheem - Neeli Pahaadi	Rajesh Kava	Rajiv Chilaka	Children & Family Movies	India	s407	Movie	July 22, 2021	2013	TV-Y7	50-80	2021-07-22	

It seems that Rajiv Chilaka has worked on Chota Bheem and has been able to create some good content in its movies. He can be relied on for more Chota Bheem stories

```
In [209]: df_actors_directors['Actor_Director_Combination'].values
```

```
Out[209]: array(['Julie Tejwani and Rajiv Chilaka', 'Rajesh Kava and Rajiv Chilaka',
 'Jigna Bhardwaj and Rajiv Chilaka',
 'Rupa Bhimani and Rajiv Chilaka', 'Vatsal Dubey and Rajiv Chilaka',
 'Swapnil and Rajiv Chilaka', 'Mousam and Rajiv Chilaka',
 'Saurav Chakraborty and Suhas Kadav',
 'Tamannaah Bhatia and S.S. Rajamouli',
 'Prabhias and S.S. Rajamouli', 'Rana Daggubati and S.S. Rajamouli',
 'Anushka Shetty and S.S. Rajamouli',
 'Sathyaraj and S.S. Rajamouli',
 'Ramya Krishnan and S.S. Rajamouli', 'Nassar and S.S. Rajamouli'],
 dtype=object)
```

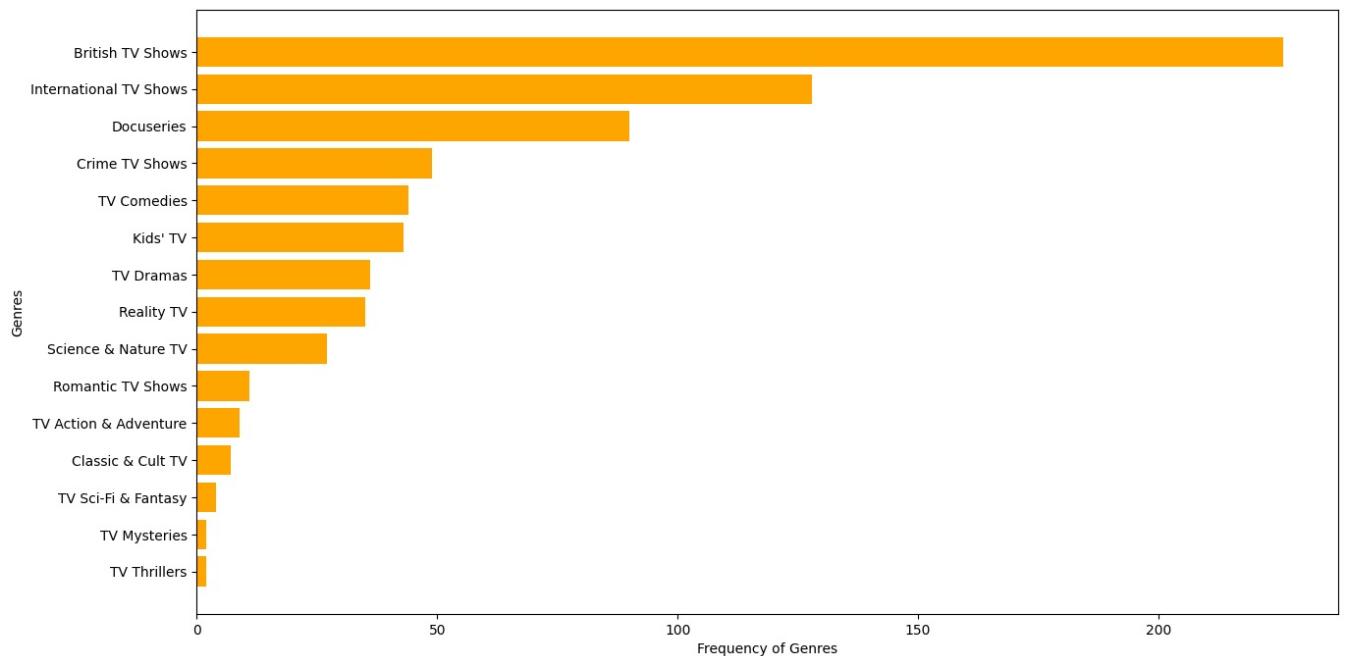
The Most Popular Actor Director Combination in Movies Across India are:-

- 'Rajesh Kava and Rajiv Chilaka',
- 'Julie Tejwani and Rajiv Chilaka',
- 'Rupa Bhimani and Rajiv Chilaka',
- 'Jigna Bhardwaj and Rajiv Chilaka',
- 'Vatsal Dubey and Rajiv Chilaka',
- 'Mousam and Rajiv Chilaka',
- 'Swapnil and Rajiv Chilaka',
- 'Saurav Chakraborty and Suhas Kadav',
- 'Smita Malhotra and Tilak Shetty',
- 'Anupam Kher and David Dhawan',
- 'Salman Khan and Sooraj R. Barjatya',

Univariate Analysis separately for shows and movies in United Kingdom

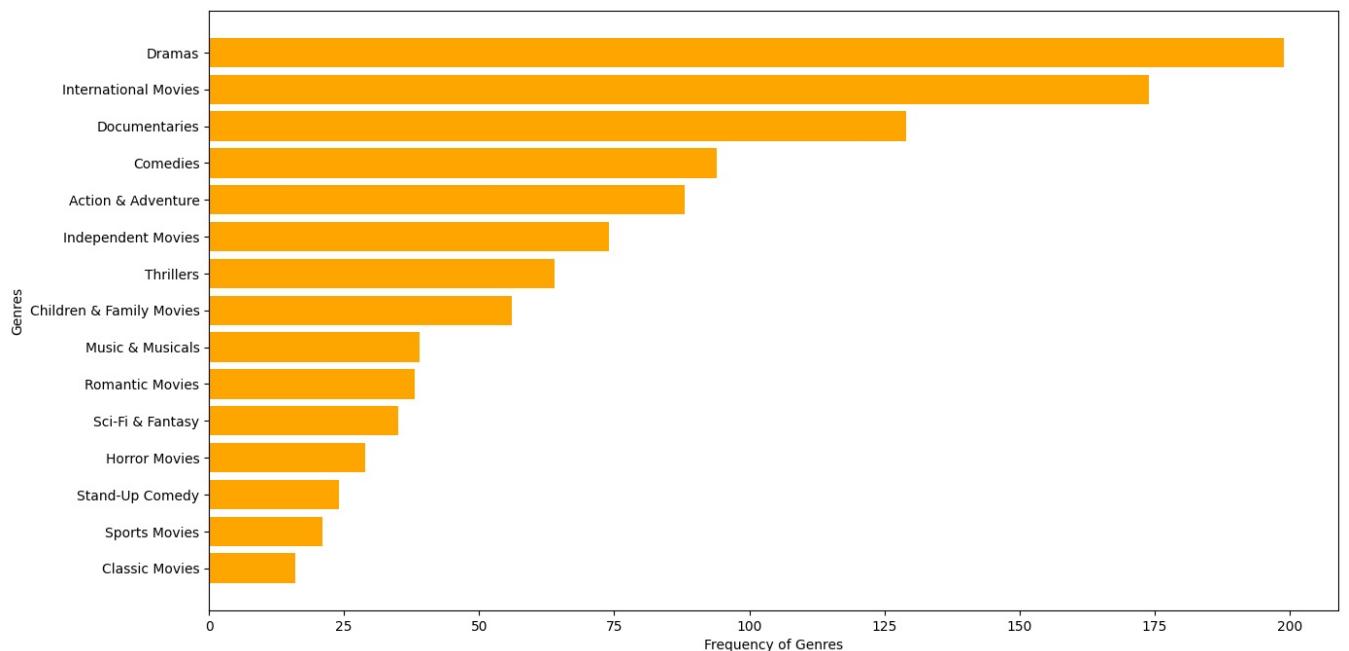
```
In [210]: #Analyzing India for both shows and movies
df_uk_shows=df_final[df_final['country']=='United Kingdom'][df_final[df_final['country']=='United Kingdom']]
df_uk_movies=df_final[df_final['country']=='United Kingdom'][df_final[df_final['country']=='United Kingdom']]
```

```
In [211]: df_genre=df_uk_shows.groupby(['Genre']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'], ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_genre[:10]['Genre'], df_genre[:10]['title'], color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```



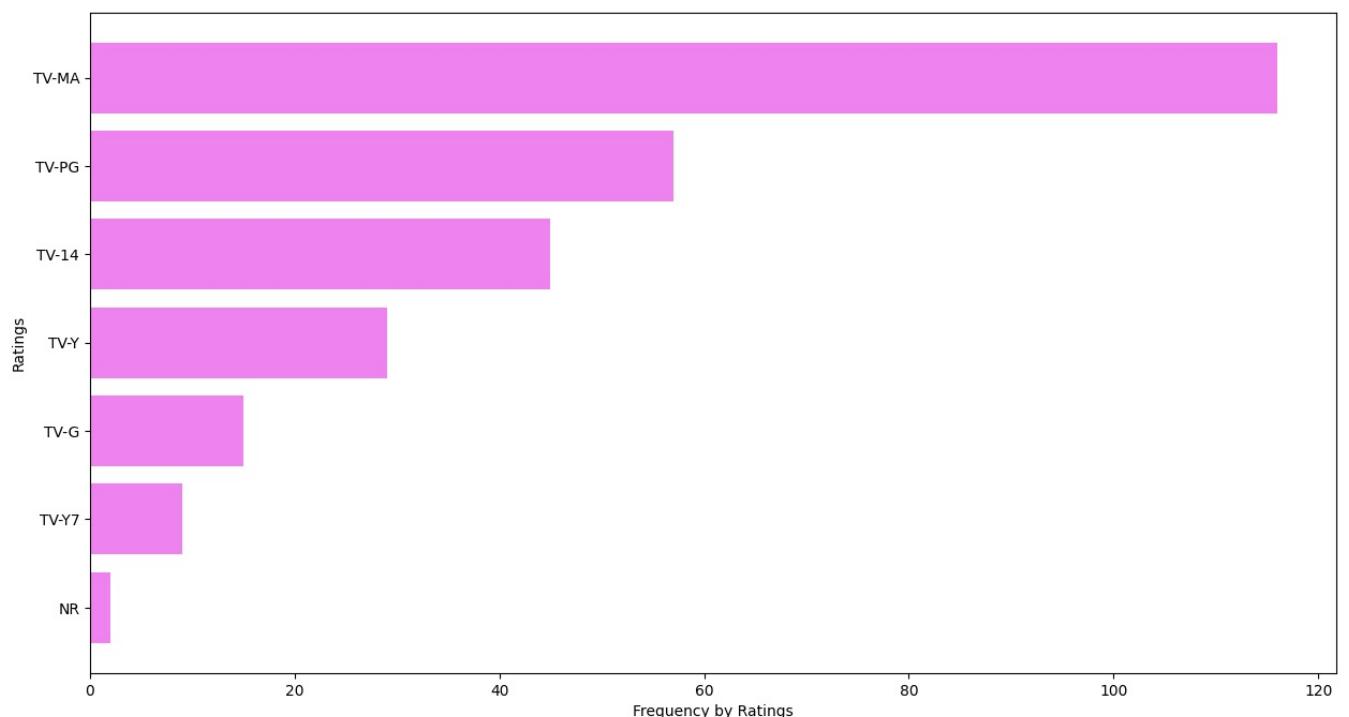
British TV Shows, International TV Shows, Docuseries, Crime, Comedy are widely watched Genres in TV Shows in UK

```
In [212]: df_genre=df_uk_movies.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_genre[::-1]['Genre'], df_genre[::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```

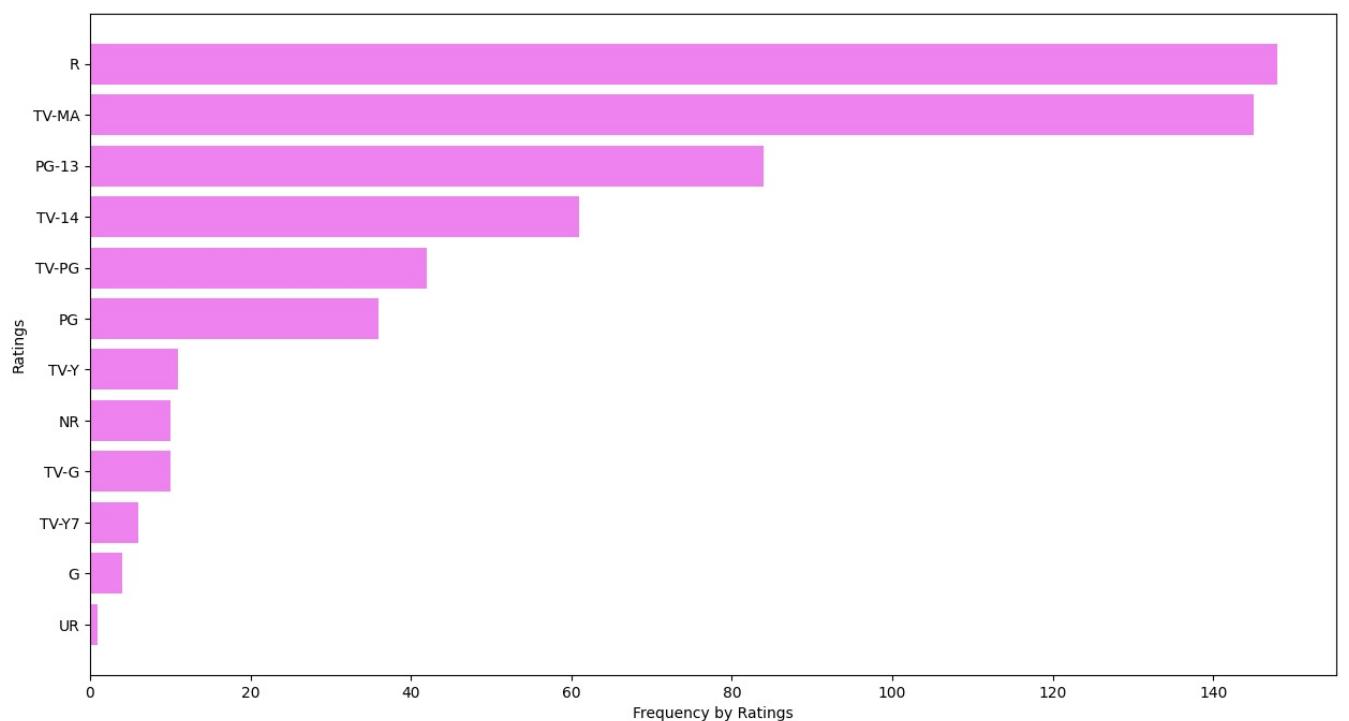


International Movies, Drama, Comedy, Independent Movies and Action, Romance Genres in Movies are prevalent in UK

```
In [213]: df_rating=df_uk_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.bart(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```

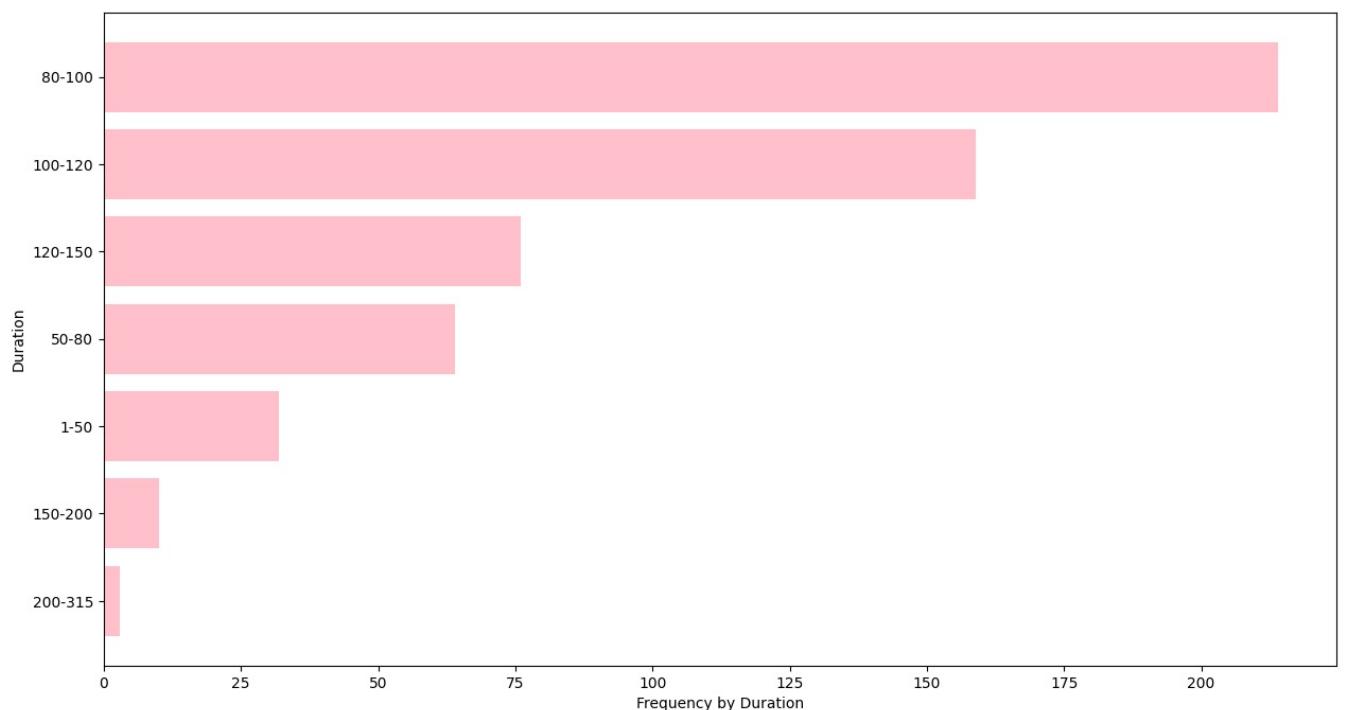


```
In [214]: df_rating=df_uk_movies.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
plt.figure(figsize=(15,8))
plt.barrh(df_rating[::-1]['rating'], df_rating[::-1]['title'], color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



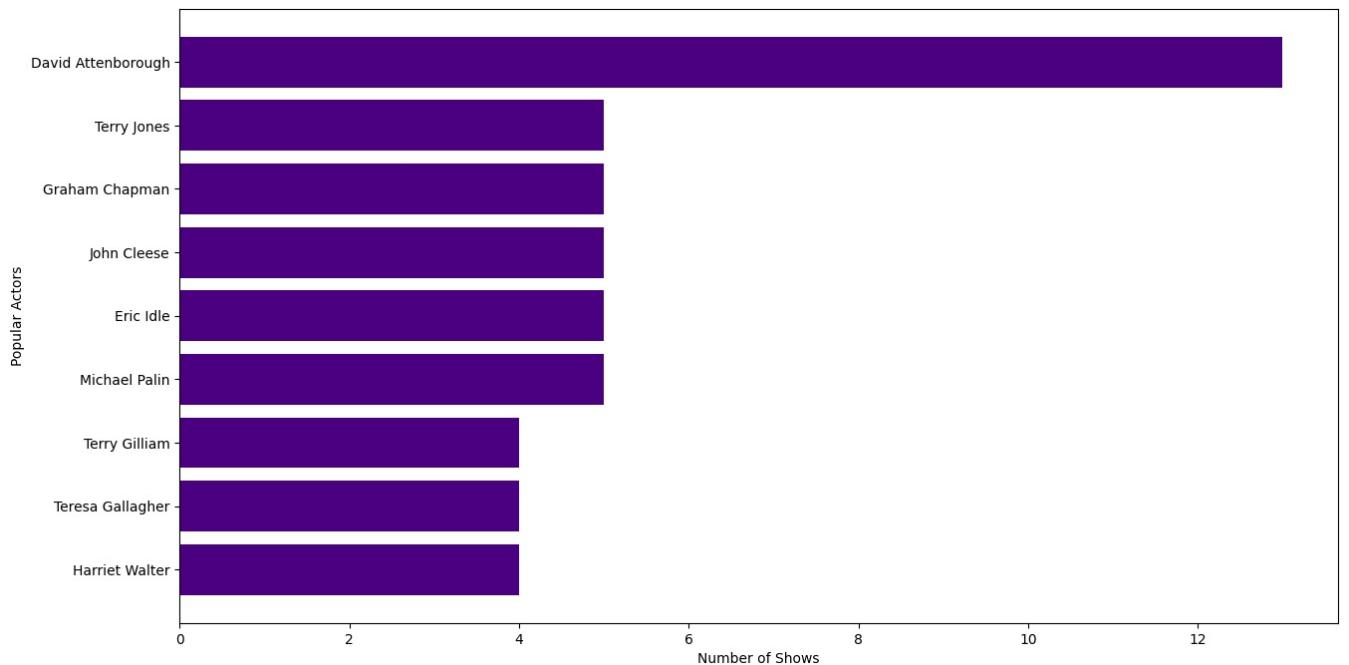
So it seems plausible to conclude that the popular ratings across Netflix includes Parental Guidance and Mature Audiences in TV Shows and R Rated+MA Rated in Movies in UK

```
In [215]: df_duration=df_uk_movies.groupby(['duration']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
plt.figure(figsize=(15,8))
plt.barrh(df_duration[::-1]['duration'], df_duration[::-1]['title'], color=['pink'])
plt.xlabel('Frequency by Duration')
plt.ylabel('Duration')
plt.show()
```



Across movies ranges of minutes in UK have a sweet spot at 80-120 mins.

```
In [216]: df_actors=df_uk_shows.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barrh(df_actors[:10]['Actors'], df_actors[:10]['title'],color=['indigo'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actors')
plt.show()
```



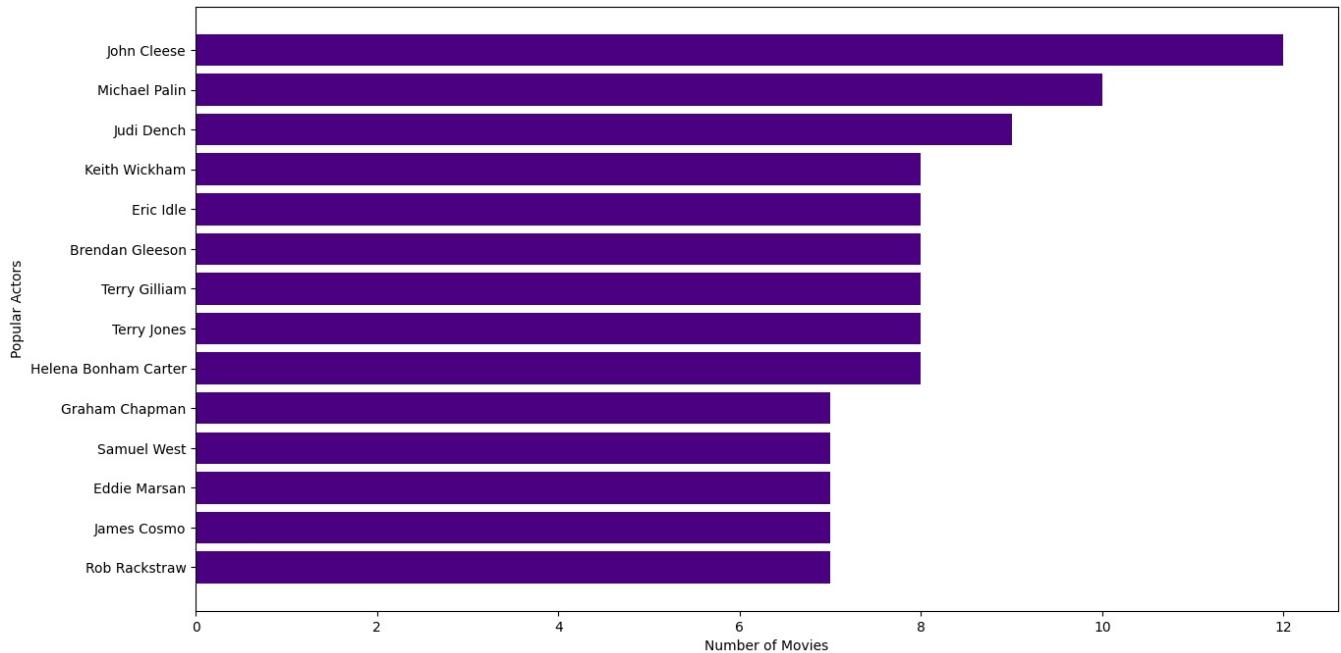
```
In [217]: df_actors['Actors'].values
```

```
Out[217]: array(['David Attenborough', 'Terry Jones', 'Graham Chapman',
       'John Cleese', 'Eric Idle', 'Michael Palin', 'Terry Gilliam',
       'Teresa Gallagher', 'Harriet Walter'], dtype=object)
```

Popular Actors in TV Shows in UK are:-

- 'David Attenborough',
- 'Terry Jones',
- 'Graham Chapman',
- 'John Cleese',
- 'Eric Idle',
- 'Michael Palin',
- 'Terry Gilliam',
- 'Teresa Gallagher',
- 'Harriet Walter'

```
In [218]: df_actors=df_uk_movies.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barh(df_actors[::-1]['Actors'], df_actors[::-1]['title'], color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actors')
plt.show()
```



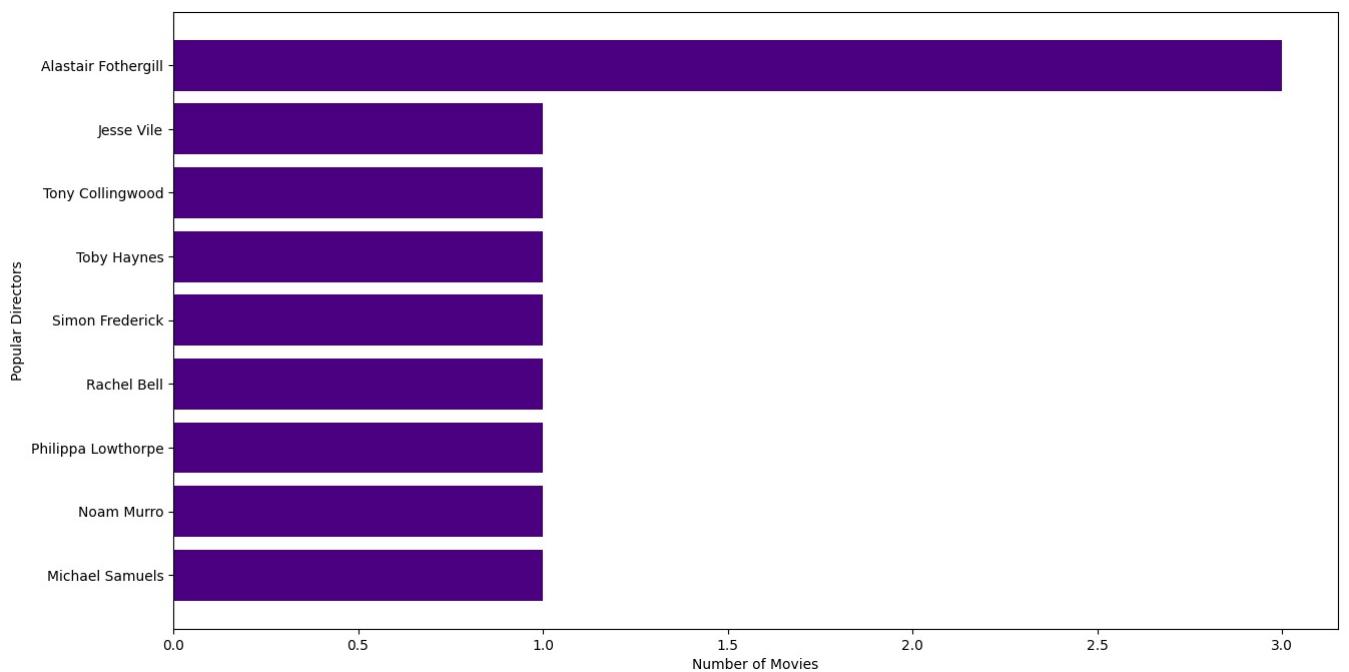
```
In [219]: df_actors['Actors'].values
```

```
Out[219]: array(['John Cleese', 'Michael Palin', 'Judi Dench', 'Keith Wickham',
       'Eric Idle', 'Brendan Gleeson', 'Terry Gilliam', 'Terry Jones',
       'Helena Bonham Carter', 'Graham Chapman', 'Samuel West',
       'Eddie Marsan', 'James Cosmo', 'Rob Rackstraw'], dtype=object)
```

Popular actors across Movies in UK:-

'John Cleese',
'Michael Palin',
'Judi Dench',
'Keith Wickham',
'Eric Idle',
'Brendan Gleeson',
'Terry Gilliam',
'Terry Jones',
'Helena Bonham Carter',
'Graham Chapman',
'Samuel West',
'Eddie Marsan',
'James Cosmo',
'Rob Rackstraw'

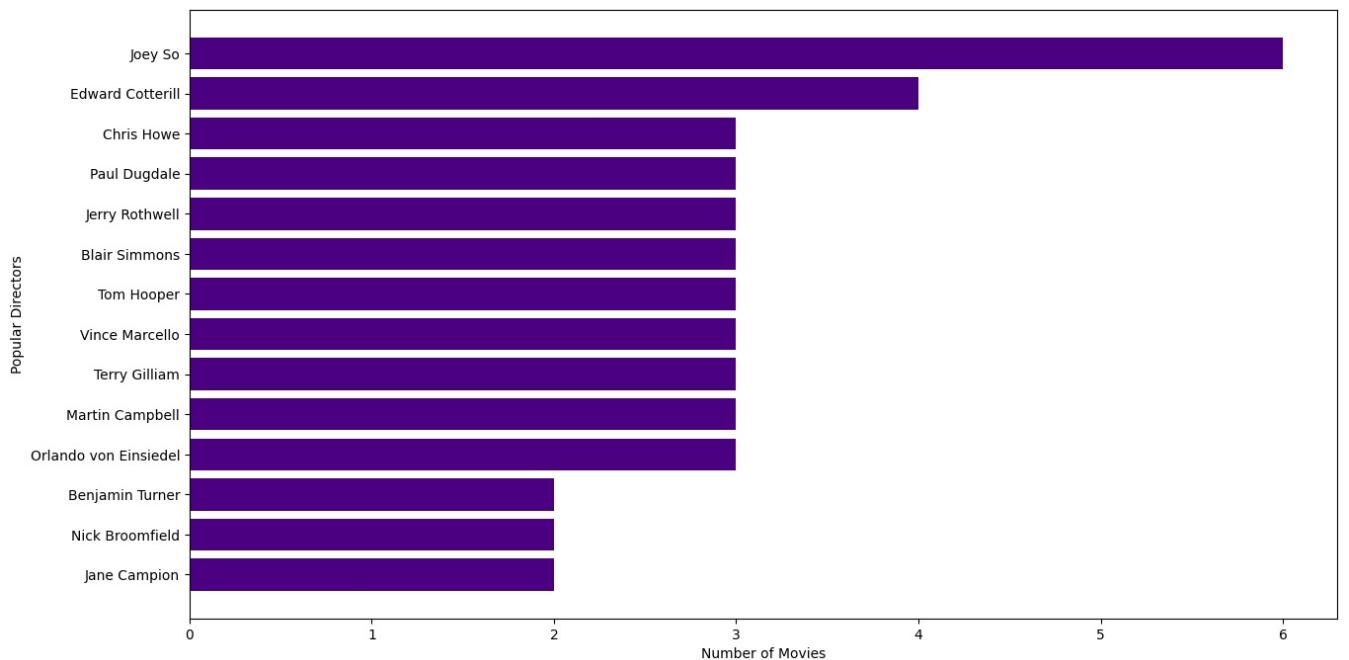
```
In [220]: df_directors=df_uk_shows.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'], ascending=False)
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[::-1]['Directors'], df_directors[::-1]['title'], color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



```
In [221]: df_directors['Directors'].values
```

```
Out[221]: array(['Alastair Fothergill', 'Jesse Vile', 'Tony Collingwood',
   'Toby Haynes', 'Simon Frederick', 'Rachel Bell',
   'Philippa Lowthorpe', 'Noam Murro', 'Michael Samuels'],
  dtype=object)
```

```
In [222]: df_directors=df_uk_movies.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'])
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barr(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



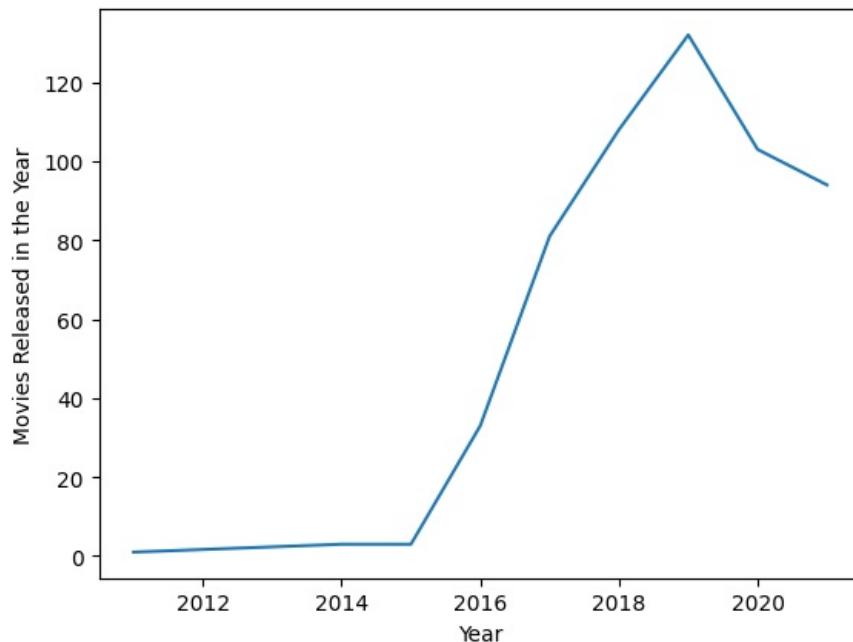
```
In [223]: df_directors['Directors'].values
```

```
Out[223]: array(['Joey So', 'Edward Cotterill', 'Chris Howe', 'Paul Dugdale',
   'Jerry Rothwell', 'Blair Simmons', 'Tom Hooper', 'Vince Marcello',
   'Terry Gilliam', 'Martin Campbell', 'Orlando von Einsiedel',
   'Benjamin Turner', 'Nick Broomfield', 'Jane Campion'], dtype=object)
```

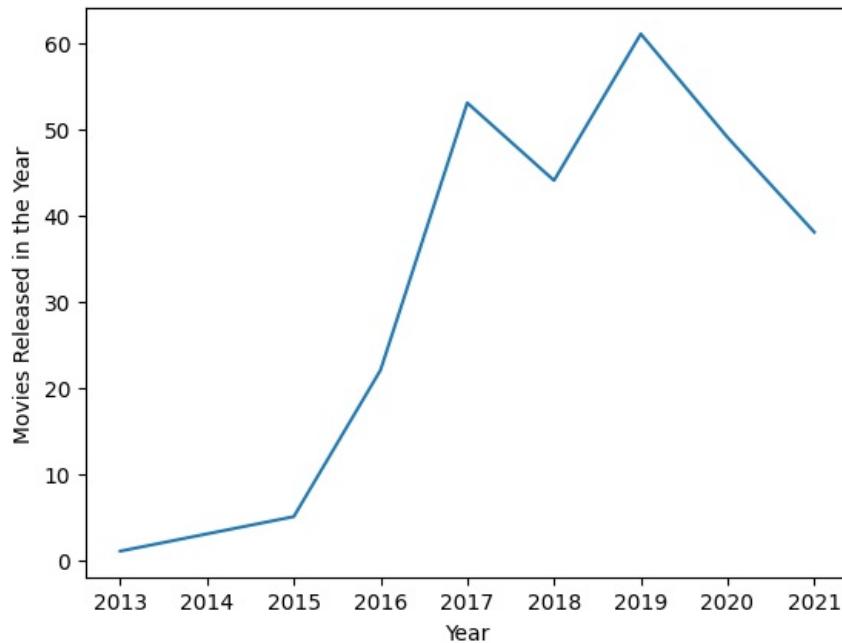
Popular directors across movies in UK:-

'Joey So',
'Edward Cotterill'

```
In [224]: df_year=df_uk_movies.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



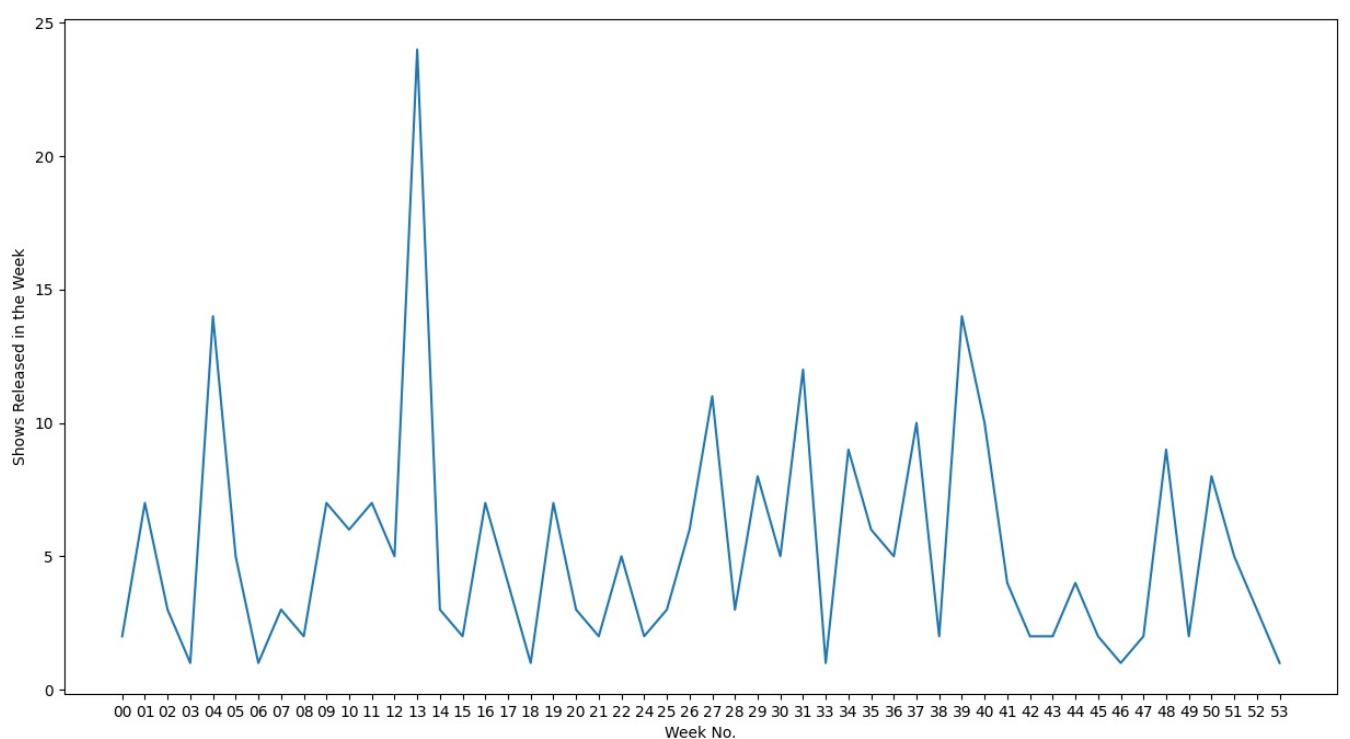
```
In [225]: df_year=df_uk_shows.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



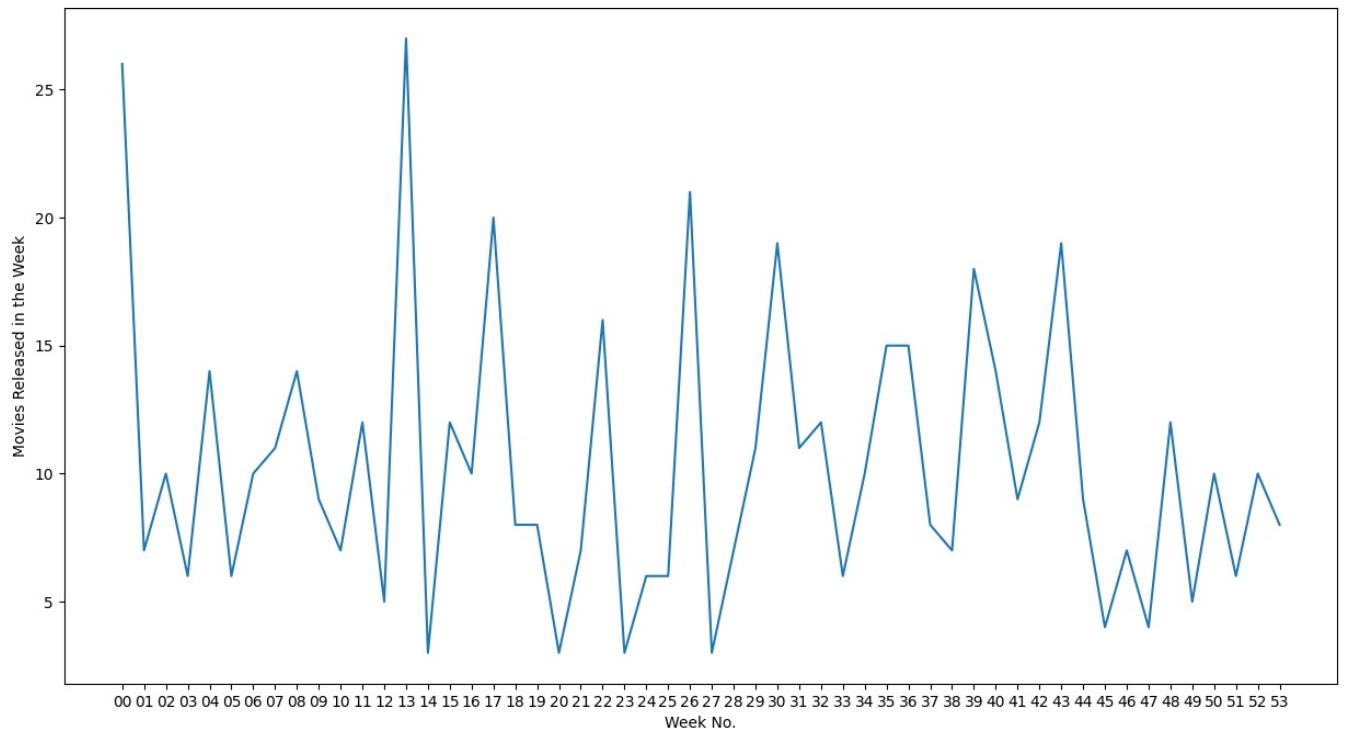
In terms of TV Shows, UK saw a downfall in 2018 from 2017, then a great increase in 2019 but has been reducing since then.

In terms of Movies, the number of popular movies in UK increased till 2019, since then it's decreasing.

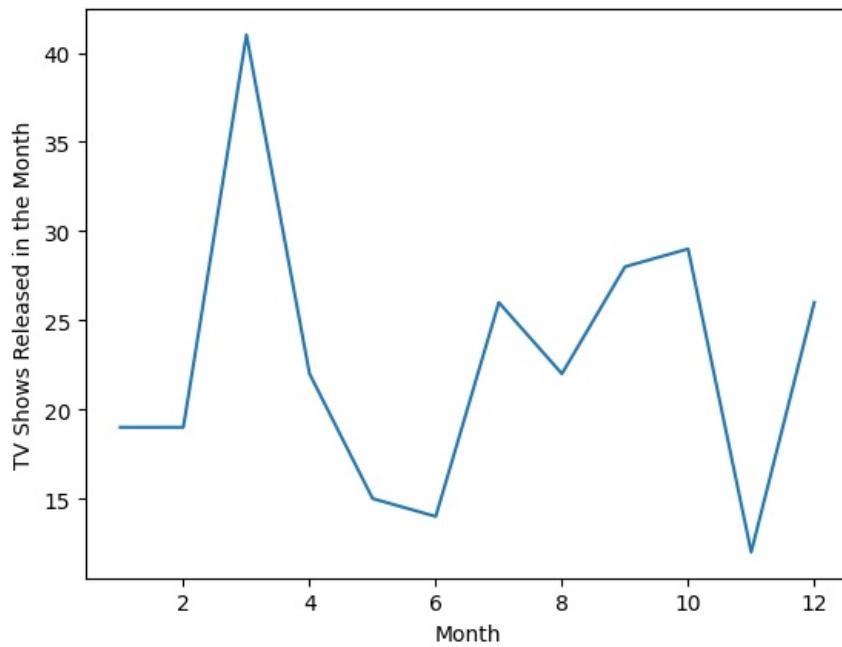
```
In [226]: df_week=df_uk_shows.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Shows Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



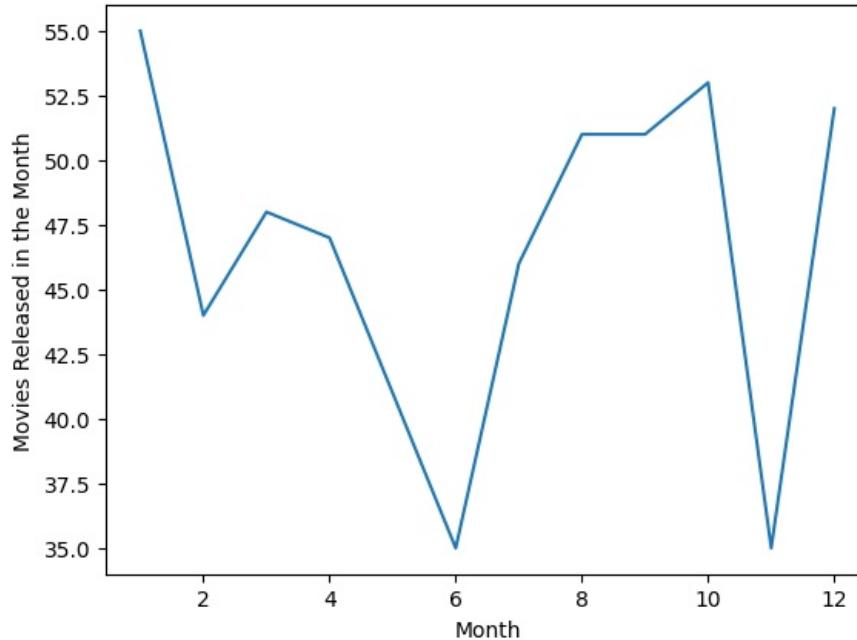
```
In [227]: df_week=df_uk_movies.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



```
In [228]: df_month=df_uk_shows.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("TV Shows Released in the Month")
plt.xlabel("Month")
plt.show()
```



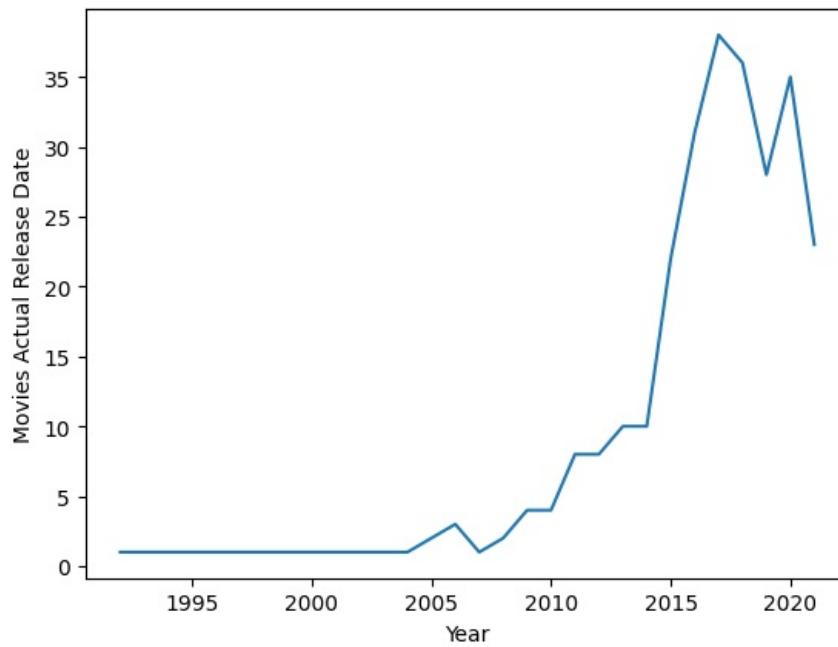
```
In [229]: df_month=df_uk_movies.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("Movies Released in the Month")
plt.xlabel("Month")
plt.show()
```



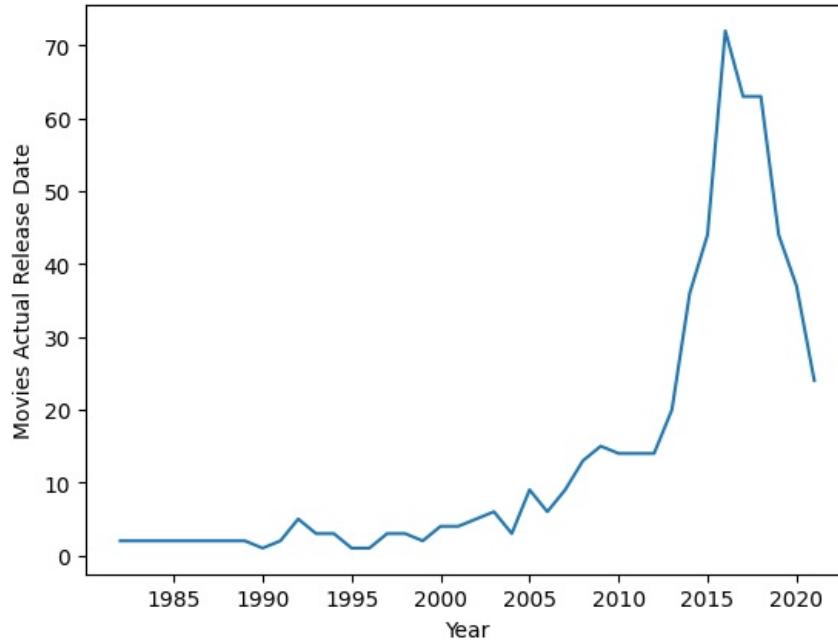
TV Shows are added in Netflix by a tremendous amount in March in UK

Movies are added in Netflix in India by a tremendous amount in first week/last month of current year and first month of next year

```
In [230]: df_release_year=df_uk_shows[df_uk_shows['release_year']>=1980].groupby(['release_year']).agg({"title":"nunique"})
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



```
In [231]: df_release_year=df_uk_movies[df_uk_movies['release_year']>=1980].groupby(['release_year']).agg({"title":"nunique"})
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



Same trend of reduction in movies and shows after 2020.

```
In [232]: #Analysing a combination of actors and directors
df_uk_movies['Actor_Director_Combination'] = df_uk_movies.Actors.str.cat(df_uk_movies.Directors, sep=' and ')
df_uk_movies_subset=df_uk_movies[df_uk_movies['Actors']!='Unknown Actor']
df_uk_movies_subset=df_uk_movies_subset[df_uk_movies_subset['Directors']!='Unknown Director']
df_uk_movies_subset.head(2)
```

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date
182	Sankofa	Kofi Ghanaba	Haile Gerima	Dramas	United Kingdom	s8	Movie	September 24, 2021	1993	TV-MA	120-150	2021-09-24
188	Sankofa	Kofi Ghanaba	Haile Gerima	Independent Movies	United Kingdom	s8	Movie	September 24, 2021	1993	TV-MA	120-150	2021-09-24

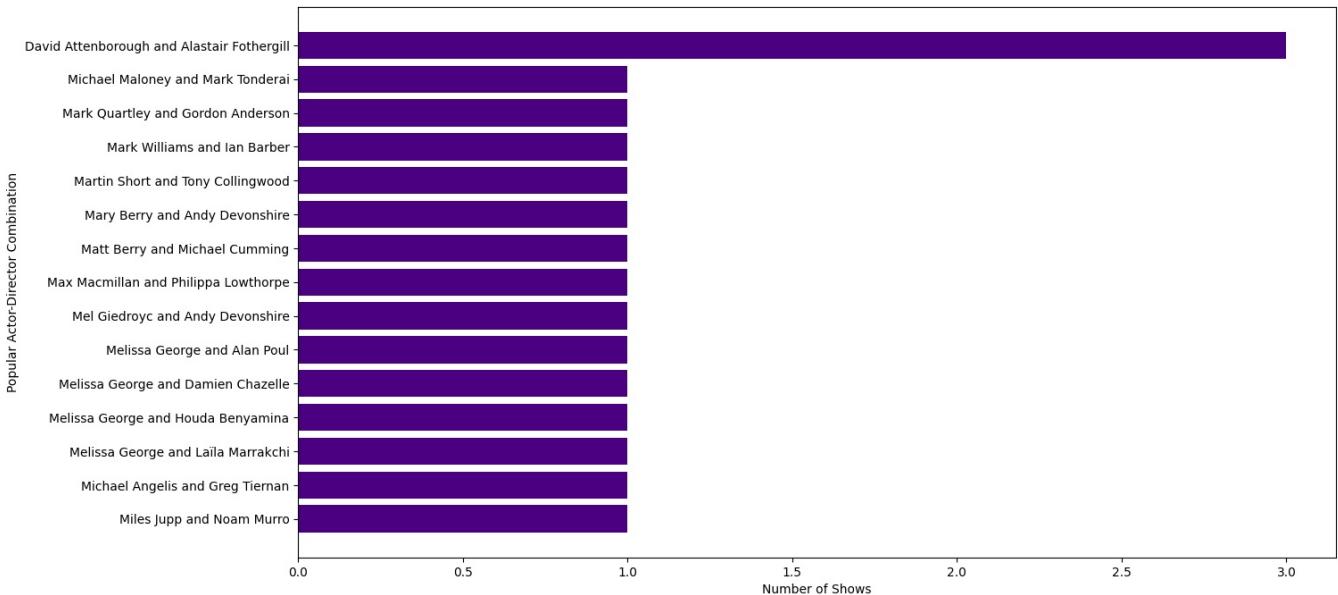
```
In [233]: df_uk_shows['Actor_Director_Combination'] = df_uk_shows.Actors.str.cat(df_uk_shows.Directors, sep=' and ')
df_uk_shows_subset=df_uk_shows[df_uk_shows['Actors']!='Unknown Actor']
df_uk_shows_subset=df_uk_shows_subset[df_uk_shows_subset['Directors']!='Unknown Director']
df_uk_shows_subset.head(2)
```

Out[233]:

	title	Actors	Directors	Genre	country	show_id	type	date_added	release_year	rating	duration	Modified_Added_date	month
323	The Great British Baking Show	Mel Giedroyc	Andy Devonshire	British TV Shows	United Kingdom	s9	TV Show	September 24, 2021	2021	TV-14	9 Seasons	2021-09-24	
324	The Great British Baking Show	Mel Giedroyc	Andy Devonshire	Reality TV	United Kingdom	s9	TV Show	September 24, 2021	2021	TV-14	9 Seasons	2021-09-24	

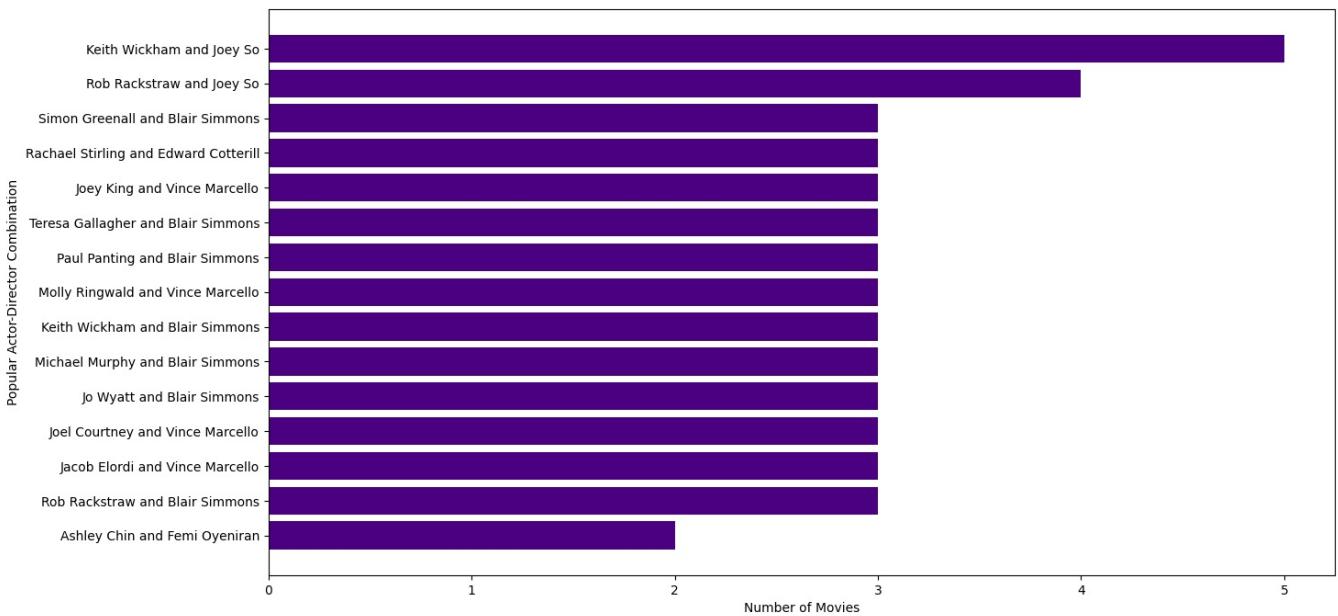
In [234]:

```
df_actors_directors=df_uk_shows_subset.groupby(['Actor_Director_Combination']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
plt.barh(df_actors_directors[:::-1]['Actor_Director_Combination'], df_actors_directors[:::-1]['title'], color=['in'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actor-Director Combination')
plt.show()
```



In [235]:

```
df_actors_directors=df_uk_movies_subset.groupby(['Actor_Director_Combination']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
plt.barh(df_actors_directors[:::-1]['Actor_Director_Combination'], df_actors_directors[:::-1]['title'], color=['in'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Actor-Director Combination')
plt.show()
```



In [236]:

```
df_actors_directors['Actor_Director_Combination'].values
```

```
Out[236]: array(['Keith Wickham and Joey So', 'Rob Rackstraw and Joey So',  
   'Simon Greenall and Blair Simmons',  
   'Rachael Stirling and Edward Cotterill',  
   'Joey King and Vince Marcello',  
   'Teresa Gallagher and Blair Simmons',  
   'Paul Panting and Blair Simmons',  
   'Molly Ringwald and Vince Marcello',  
   'Keith Wickham and Blair Simmons',  
   'Michael Murphy and Blair Simmons', 'Jo Wyatt and Blair Simmons',  
   'Joel Courtney and Vince Marcello',  
   'Jacob Elordi and Vince Marcello',  
   'Rob Rackstraw and Blair Simmons', 'Ashley Chin and Femi Oyeniran'],  
  dtype=object)
```

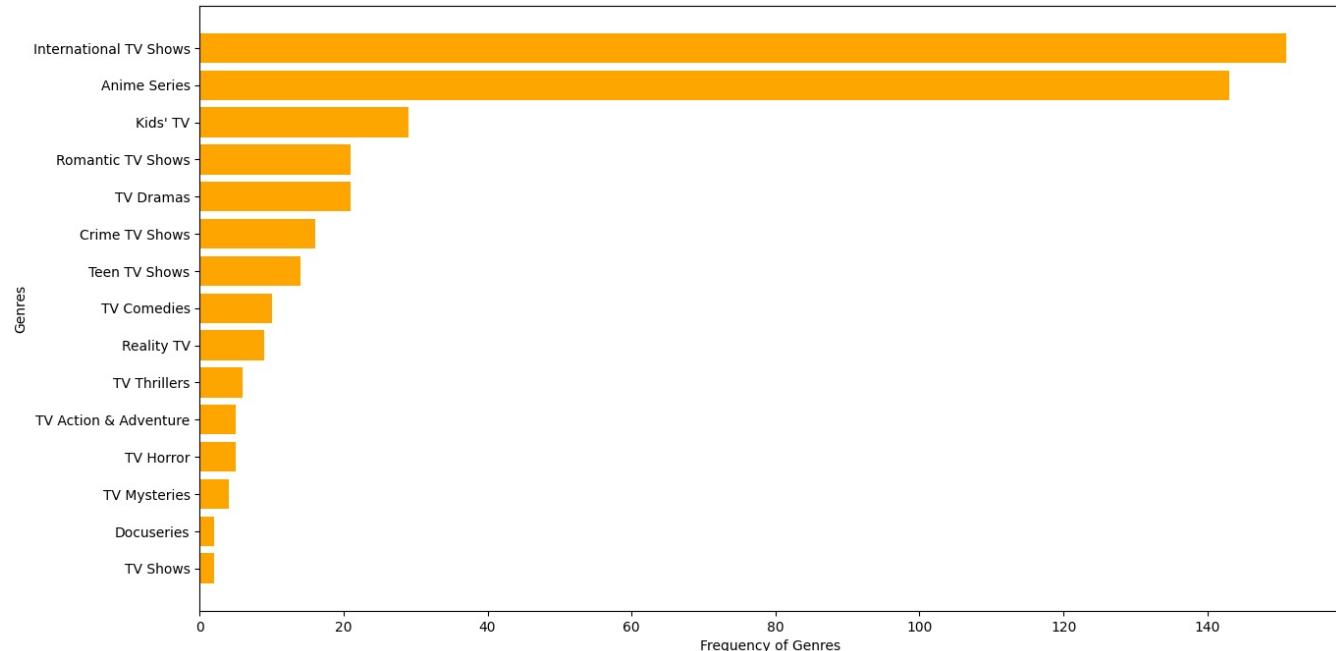
The Most Popular Actor Director Combination in Movies Across UK are:-

'Keith Wickham and Joey So',
'Rob Rackstraw and Joey So'

Univariate Analysis separately for shows in Japan

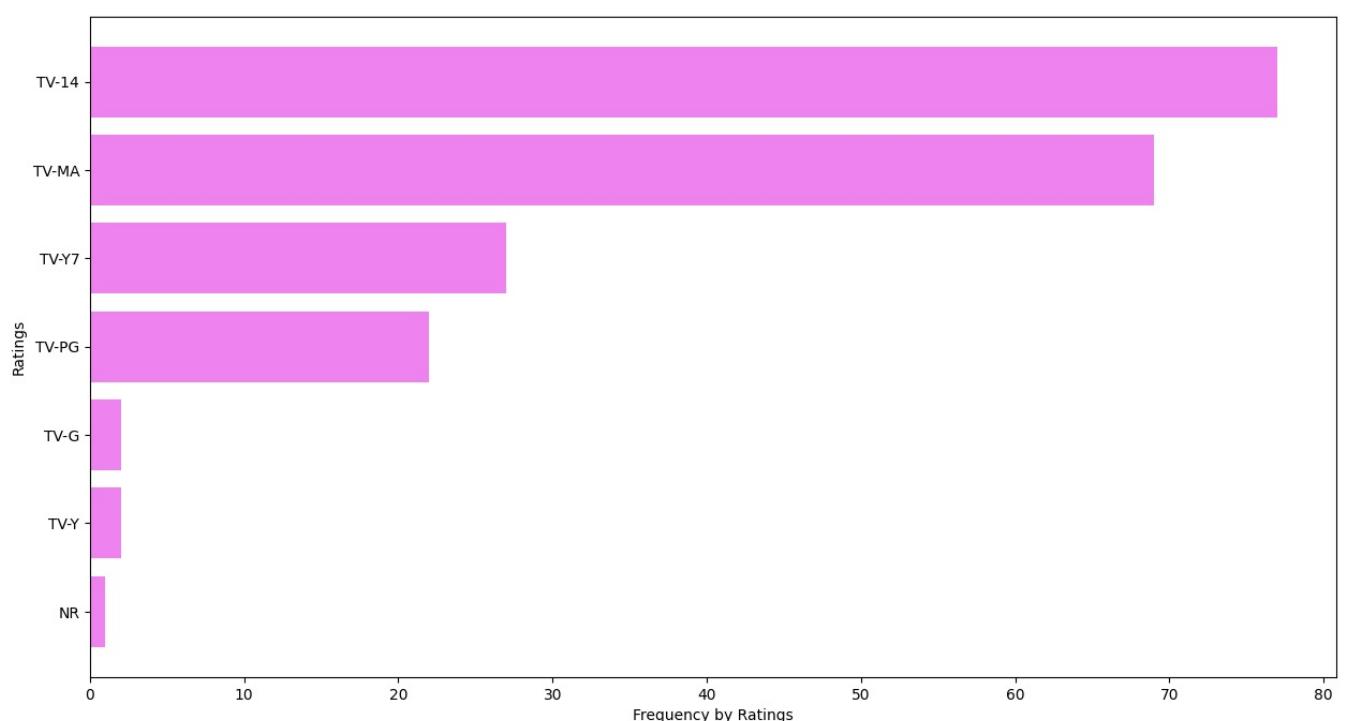
```
In [237]: #Analyzing India for both shows and movies  
df_japan_shows=df_final1[df_final1['country']=='Japan'][df_final1[df_final1['country']=='Japan']['type']=='TV S
```

```
In [238]: df_genre=df_japan_shows.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],asc  
plt.figure(figsize=(15,8))  
plt.bahr(df_genre[::-1]['Genre'], df_genre[::-1]['title'],color=['orange'])  
plt.xlabel('Frequency of Genres')  
plt.ylabel('Genres')  
plt.show()
```



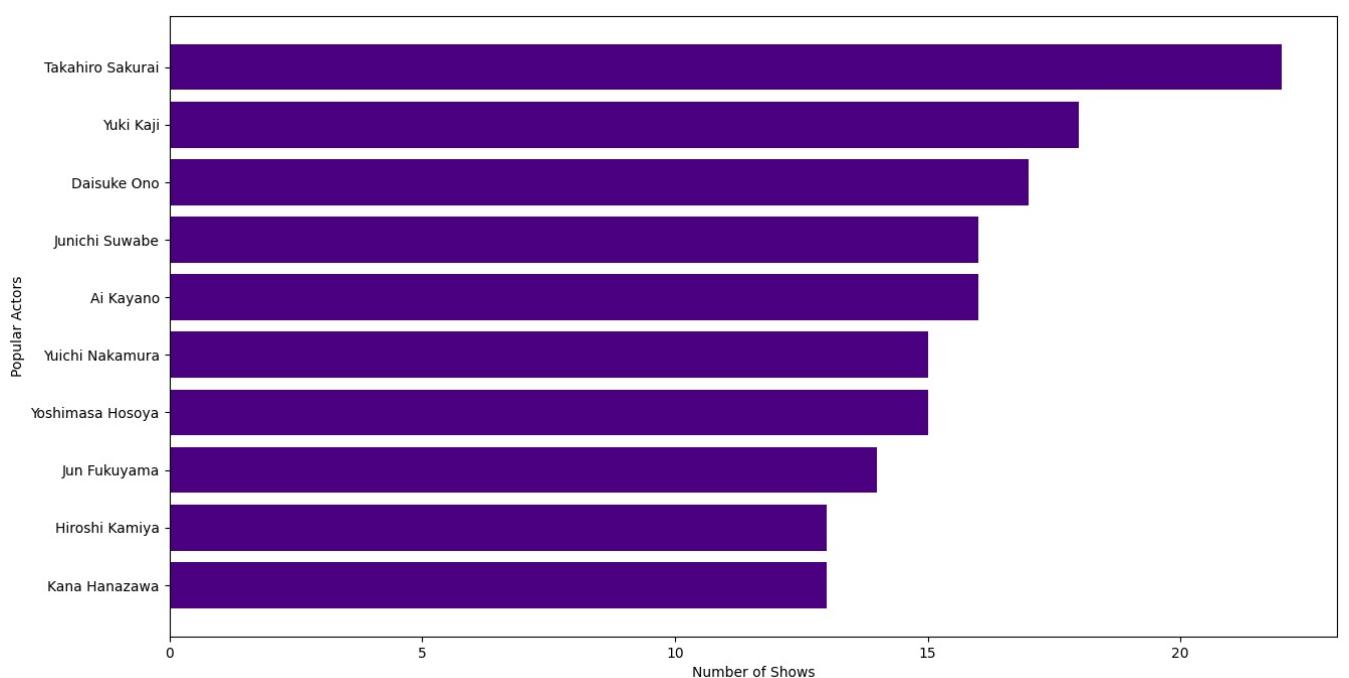
International TV Shows and Anime Genres are popular in TV Shows in Japan

```
In [239]: df_rating=df_japan_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],as  
plt.figure(figsize=(15,8))  
plt.bahr(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['violet'])  
plt.xlabel('Frequency by Ratings')  
plt.ylabel('Ratings')  
plt.show()
```



So it seems plausible to conclude that the popular ratings across Netflix includes TV-14 Mature Audiences in TV Shows

```
In [240]: df_actors=df_japan_shows.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barrh(df_actors[:10]['Actors'], df_actors[:10]['title'],color=['indigo'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actors')
plt.show()
```



```
In [241]: df_actors['Actors'].values
```

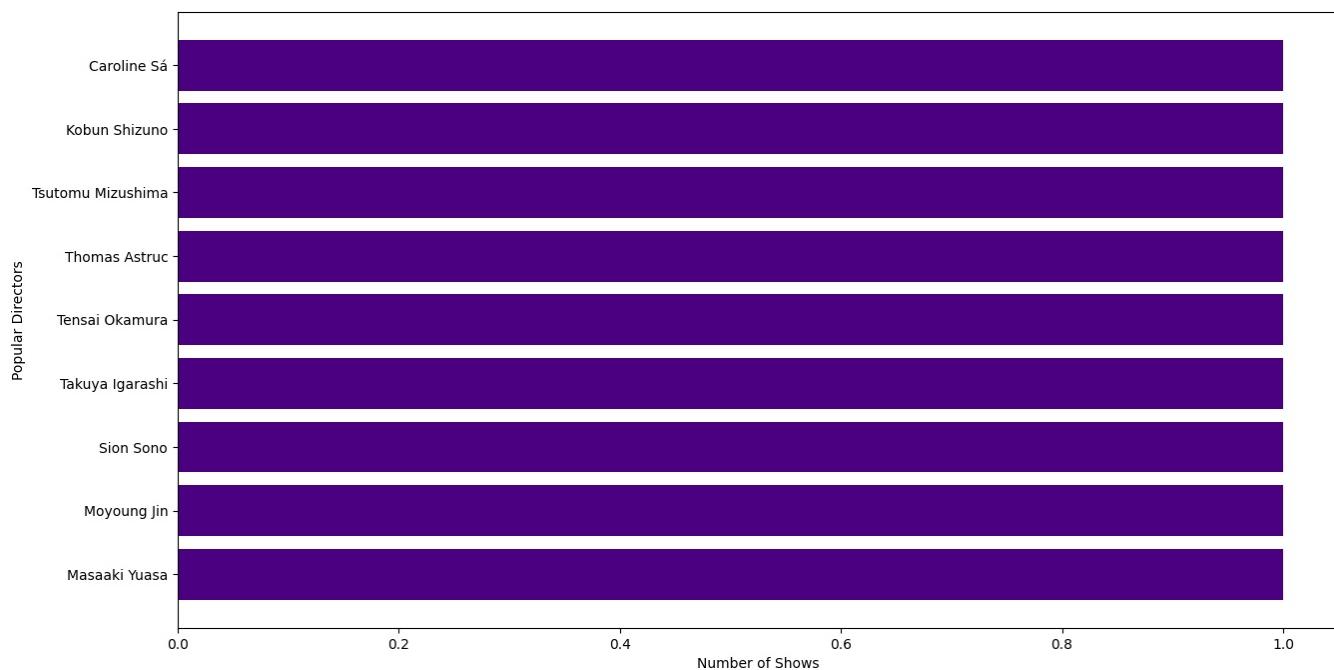
```
Out[241]: array(['Takahiro Sakurai', 'Yuki Kaji', 'Daisuke Ono', 'Junichi Suwabe',
       'Ai Kayano', 'Yuichi Nakamura', 'Yoshimasa Hosoya', 'Jun Fukuyama',
       'Hiroshi Kamiya', 'Kana Hanazawa'], dtype=object)
```

Popular Actors in TV Shows in Japan are:-

- 'Takahiro Sakurai',
- 'Yuki Kaji',
- 'Daisuke Ono',
- 'Junichi Suwabe',
- 'Ai Kayano',
- 'Yuichi Nakamura',
- 'Yoshimasa Hosoya',
- 'Jun Fukuyama',
- 'Hiroshi Kamiya',

'Kana Hanazawa'

```
In [242]: df_directors=df_japan_shows.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'])
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Directors')
plt.show()
```

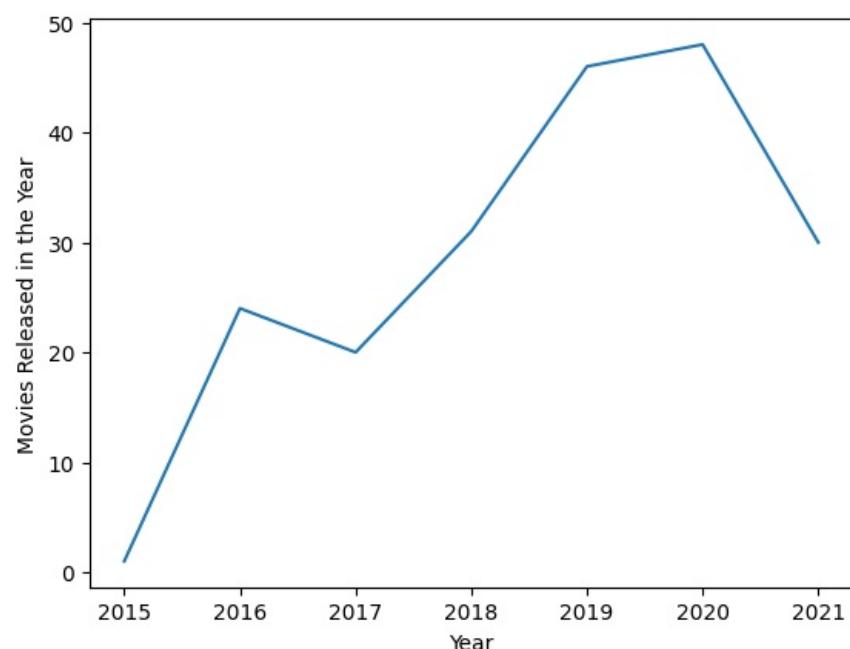


```
In [243]: df_directors['Directors'].values
```

```
Out[243]: array(['Caroline Sá', 'Kobun Shizuno', 'Tsutomu Mizushima',
   'Thomas Astruc', 'Tensai Okamura', 'Takuya Igarashi', 'Sion Sono',
   'Moyoung Jin', 'Masaaki Yuasa'], dtype=object)
```

All Directors are one time directors only

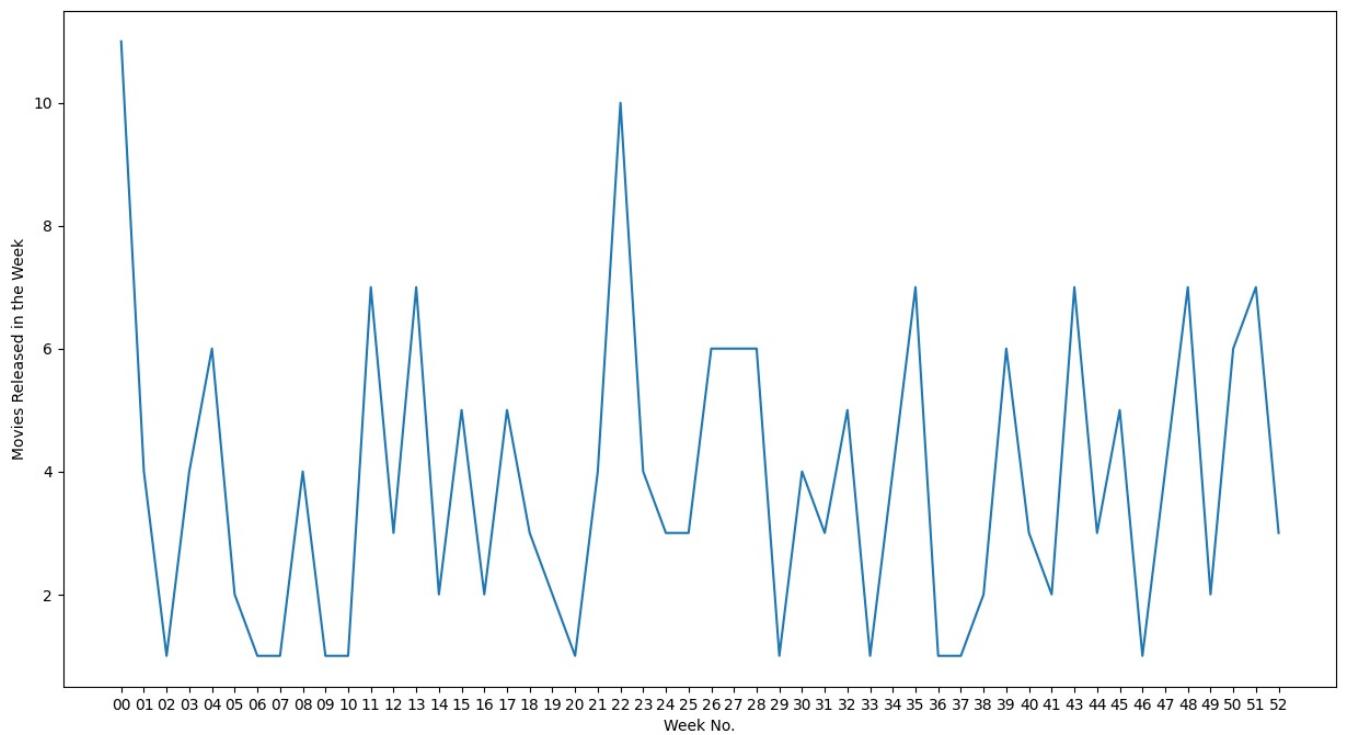
```
In [244]: df_year=df_japan_shows.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



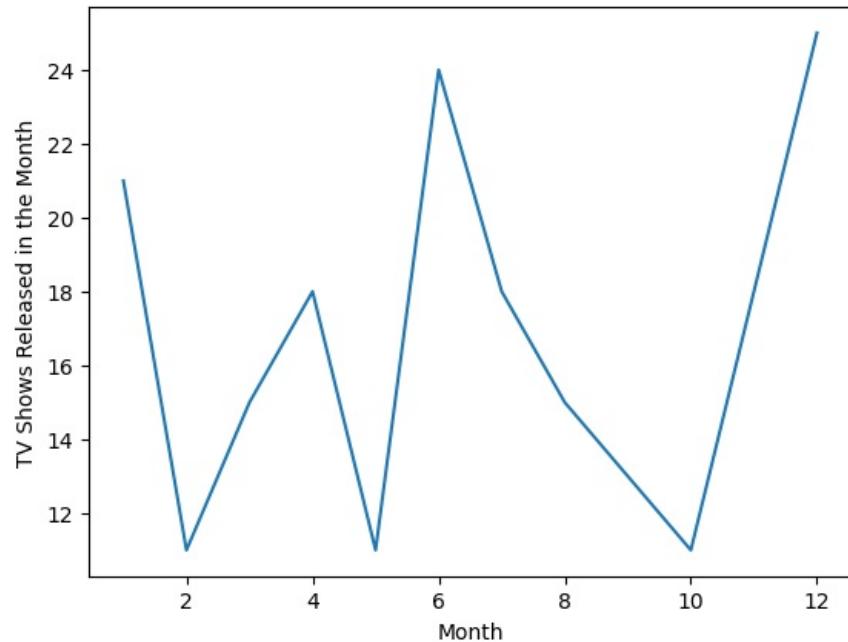
In Japan, TV Shows have diminished in 2017 from 2016 and then increased till 2020 after which it has reduced in 2021.

```
In [245]: df_week=df_japan_shows.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
```

```
plt.show()
```

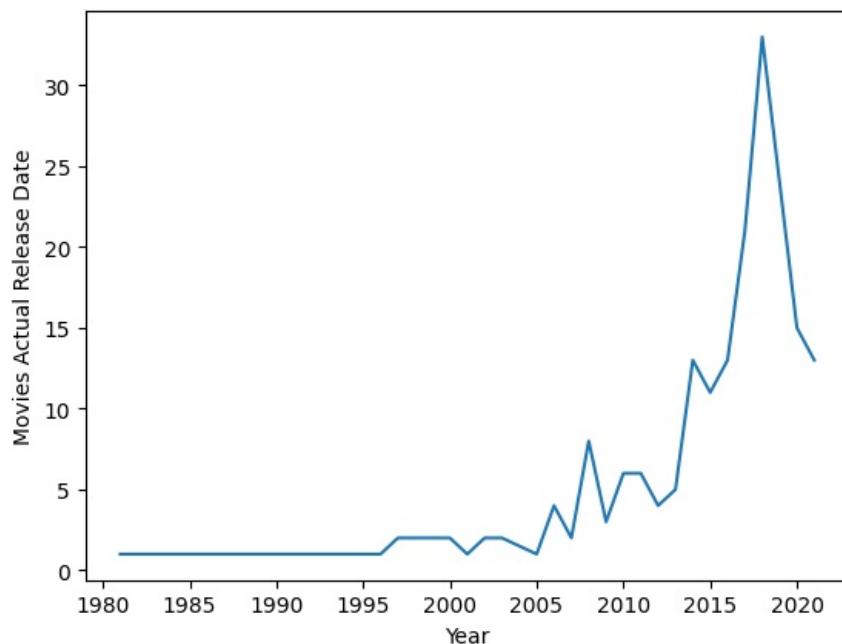


```
In [246]: df_month=df_japan_shows.groupby(['month_added']).agg({"title":"nunique"}).reset_index()  
sns.lineplot(data=df_month, x='month_added', y='title')  
plt.ylabel("TV Shows Released in the Month")  
plt.xlabel("Month")  
plt.show()
```



TV Shows are added in Netflix by significant numbers in April and January in Japan

```
In [247]: df_release_year=df_japan_shows[df_japan_shows['release_year']>=1980].groupby(['release_year']).agg({"title":"nu  
sns.lineplot(data=df_release_year, x='release_year', y='title')  
plt.ylabel("Movies Actual Release Date")  
plt.xlabel("Year")  
plt.show()
```

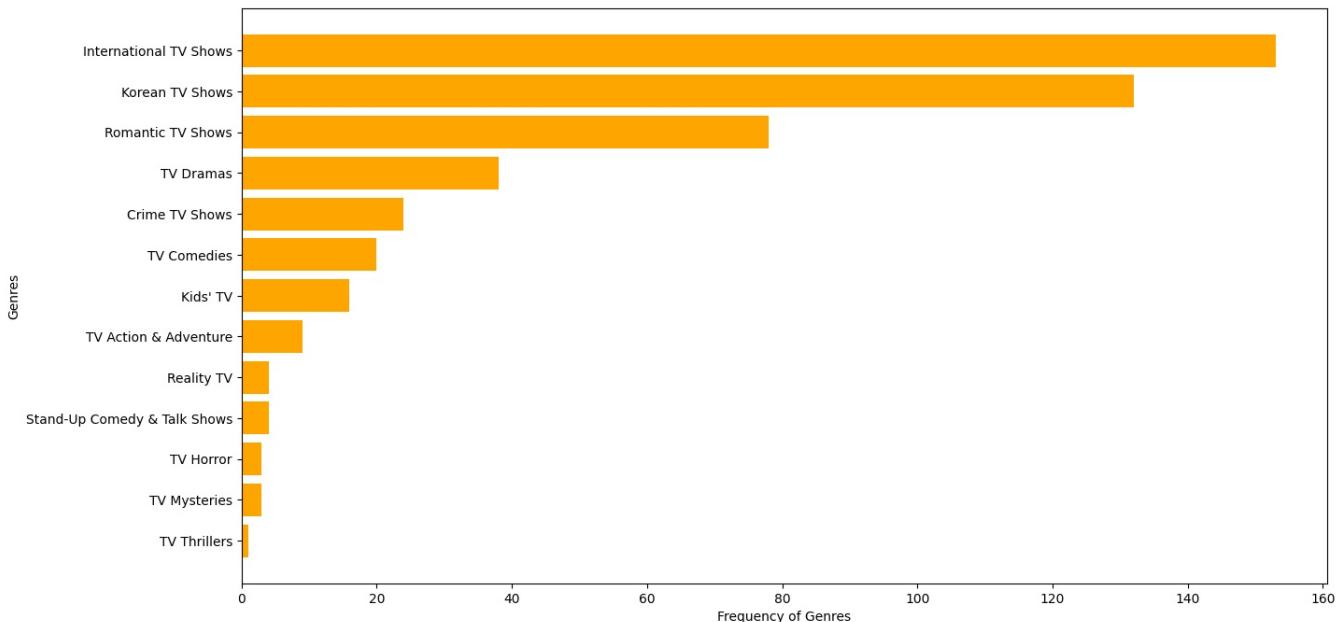


Reduction in TV Shows after 2019 in Japan

Univariate Analysis separately for shows in South Korea

```
In [248]: #Analyzing India for both shows and movies
df_sk_shows=df_final1[df_final1['country']=='South Korea'][df_final1['country']=='South Korea']['type']

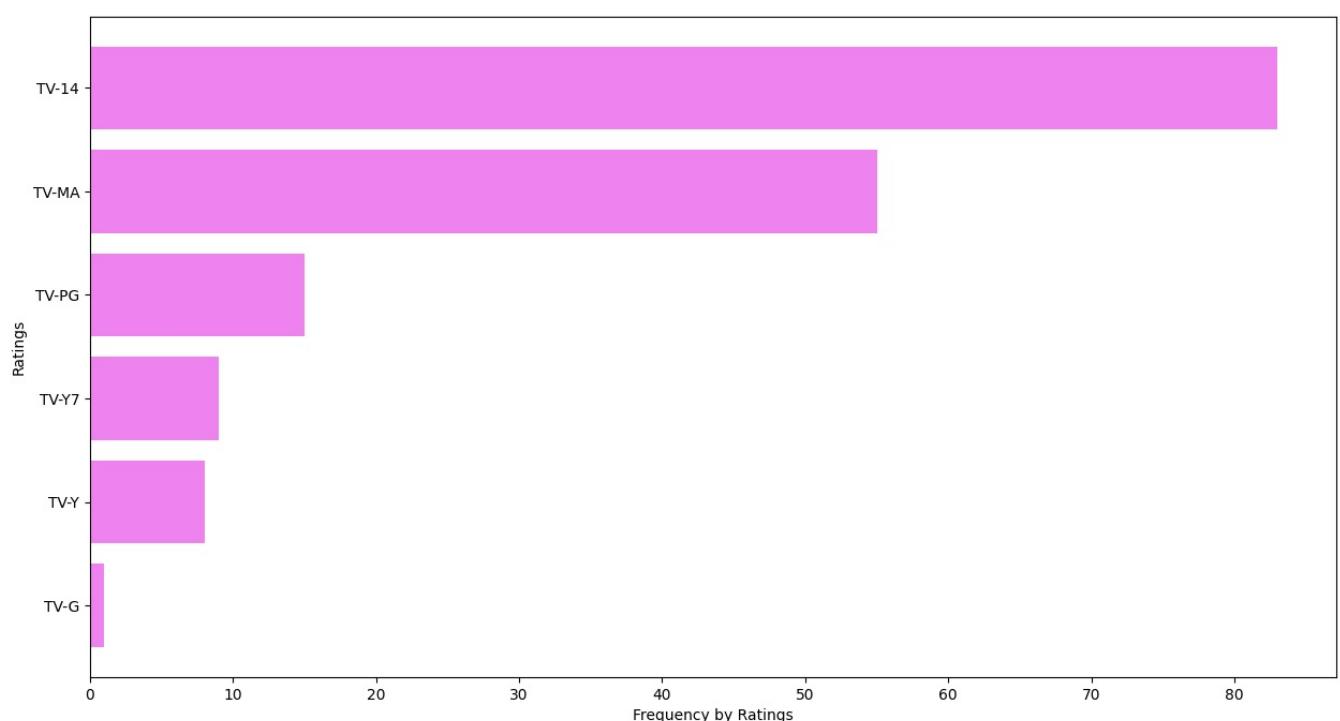
In [249]: df_genre=df_sk_shows.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_genre[::-1]['Genre'], df_genre[::-1]['title'], color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```



International TV Shows, Romantic TV Shows, Drama, Crime and Comedy Genres are popular in TV Shows in S.Korea.

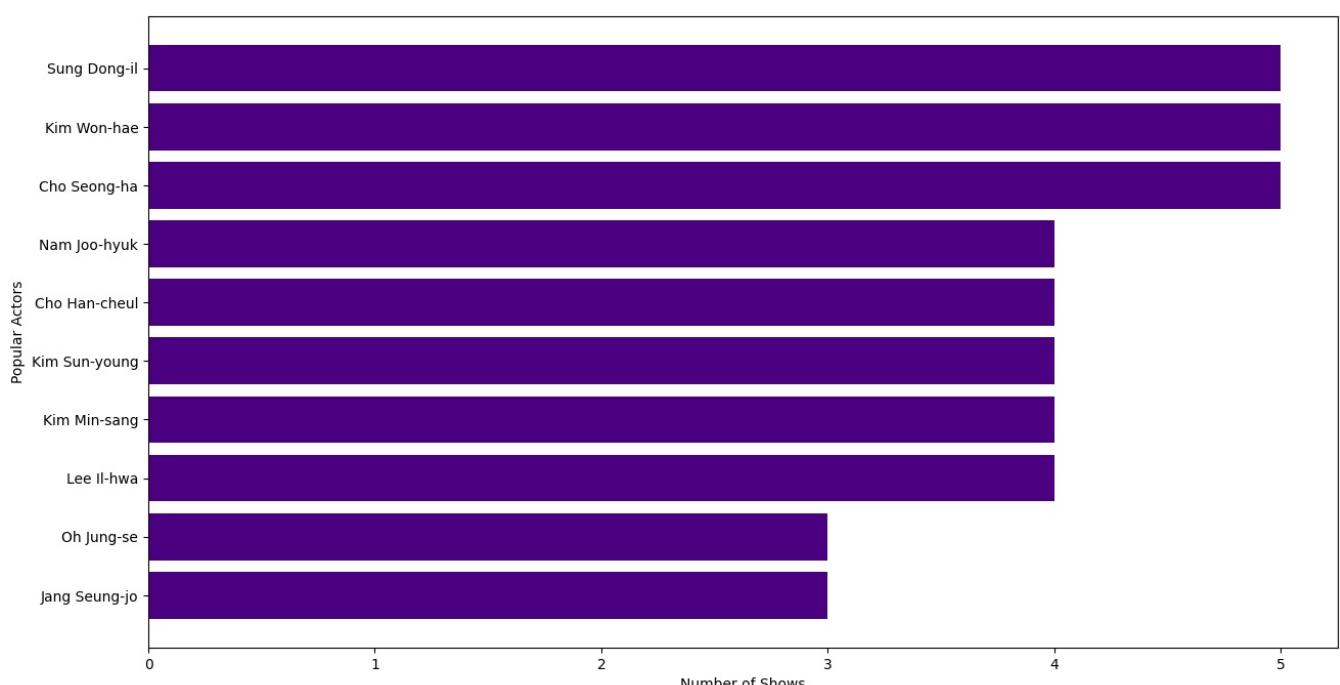
Only S.Korea has Romance as a top 3 favorable genre which depicts an inclination of their audience

```
In [250]: df_rating=df_sk_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
plt.figure(figsize=(15,8))
plt.barh(df_rating[::-1]['rating'], df_rating[::-1]['title'], color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```



So it seems plausible to conclude that the popular ratings across Netflix includes TV-14 and Mature Audiences in TV Shows

```
In [251]: df_actors=df_sk_shows.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_actors=df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,8))
plt.barrh(df_actors[::-1]['Actors'], df_actors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Shows')
plt.ylabel('Popular Actors')
plt.show()
```



```
In [252]: df_actors['Actors'].values
```

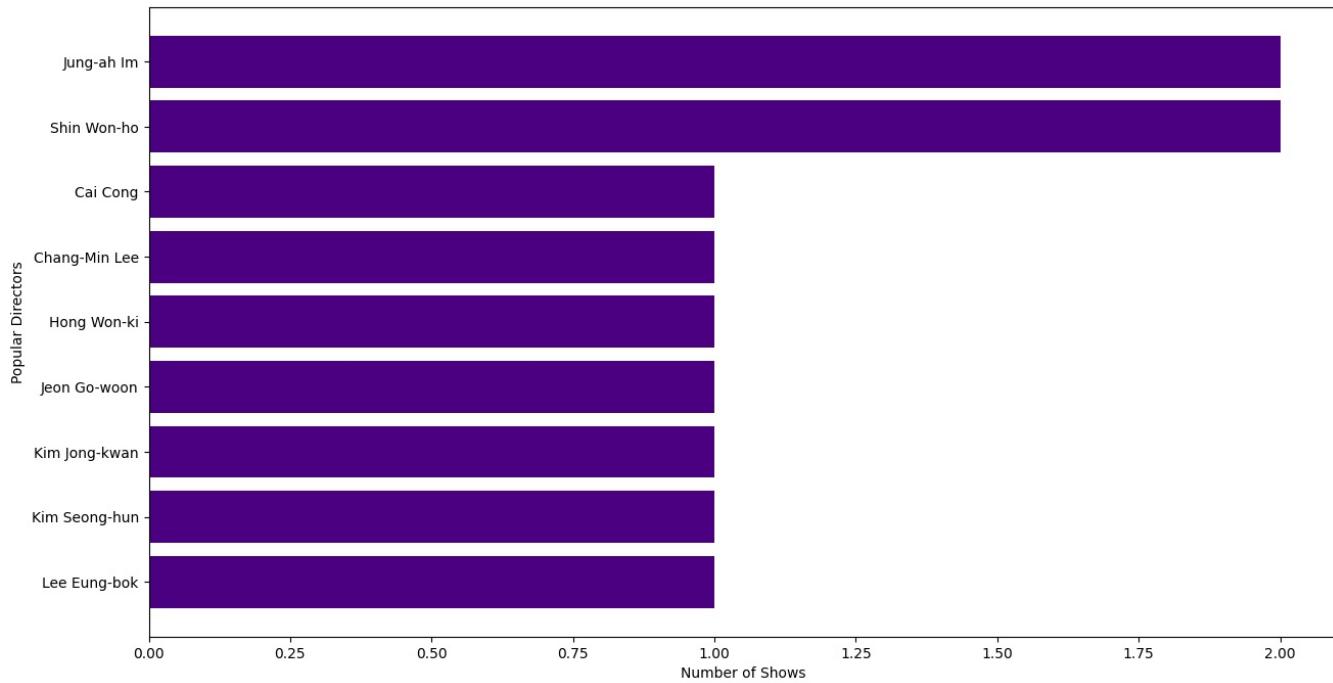
```
Out[252]: array(['Sung Dong-il', 'Kim Won-hae', 'Cho Seong-ha', 'Nam Joo-hyuk',
 'Cho Han-cheul', 'Kim Sun-young', 'Kim Min-sang', 'Lee Il-hwa',
 'Oh Jung-se', 'Jang Seung-jo'], dtype=object)
```

Popular Actors in TV Shows in South Korea are:-

- 'Sung Dong-il',
- 'Kim Won-hae',
- 'Cho Seong-ha',
- 'Nam Joo-hyuk'

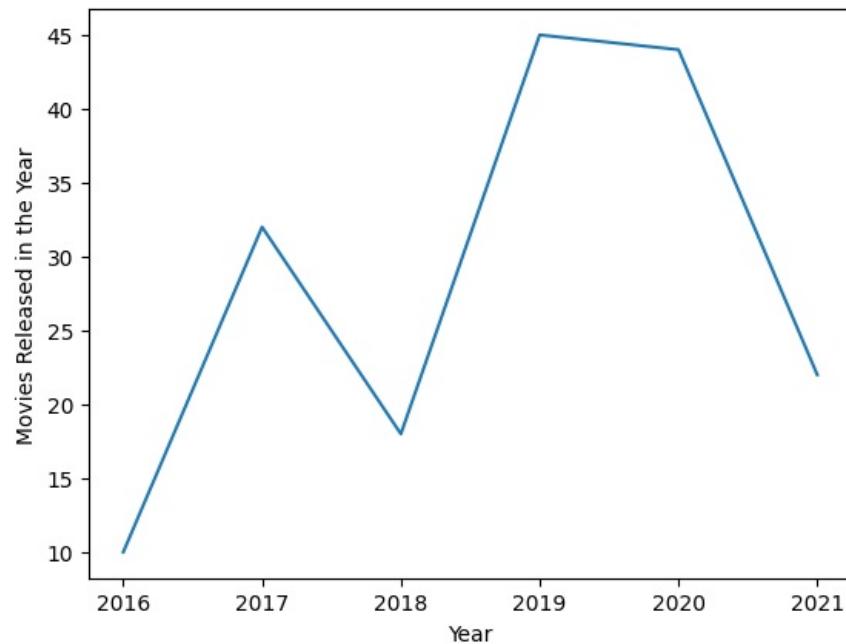
```
In [253]: df_directors=df_sk_shows.groupby(['Directors']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)
df_directors=df_directors[df_directors['Directors']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barrh(df_directors[::-1]['Directors'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Shows')
```

```
plt.ylabel('Popular Directors')
plt.show()
```



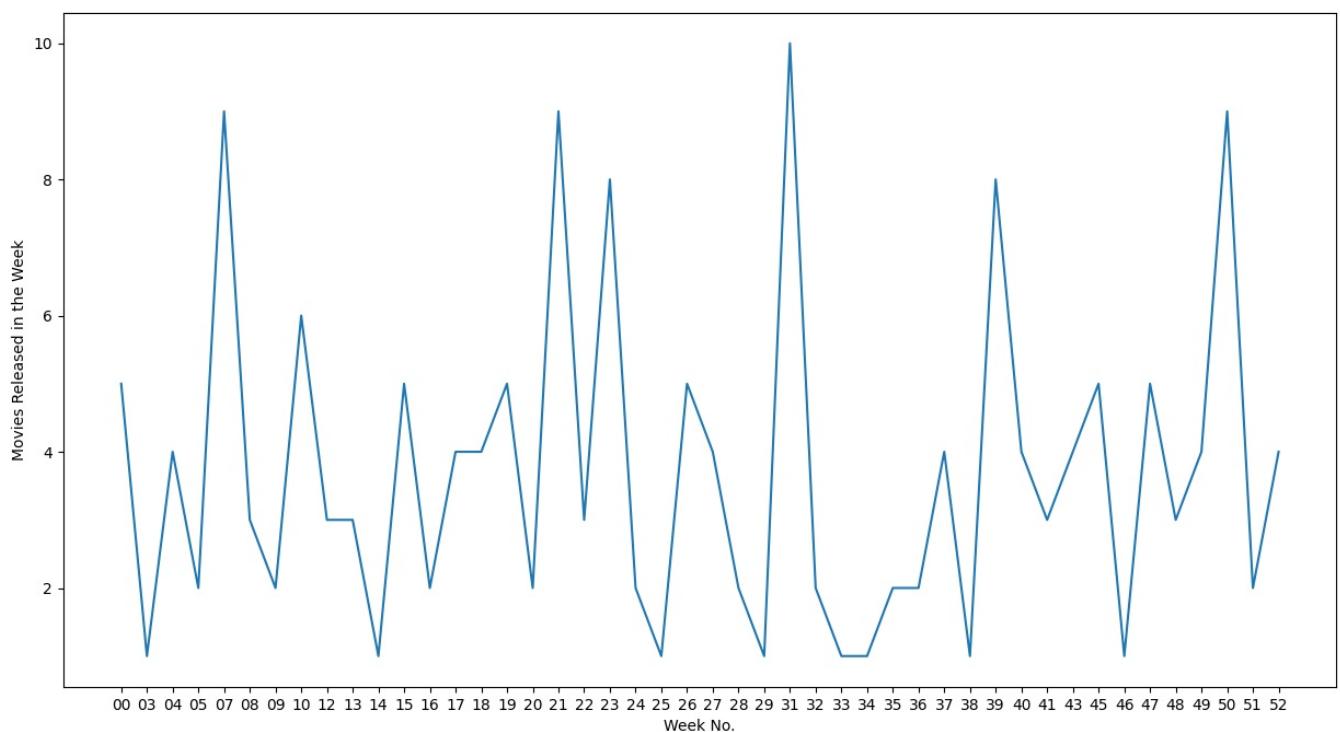
Two directors have directed 2 shows and rest all Directors are one time directors only

```
In [254]: df_year=df_sk_shows.groupby(['year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```

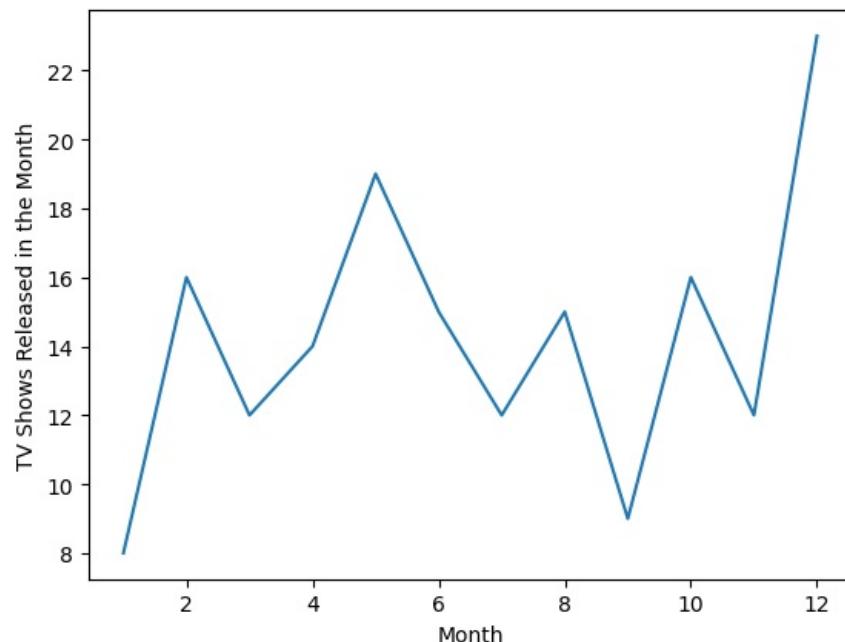


In South Korea, number of TV Shows reduced in 2018 from 2017, then increased till 2019 but have been on a heavy downfall since then

```
In [255]: df_week=df_sk_shows.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(15,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```

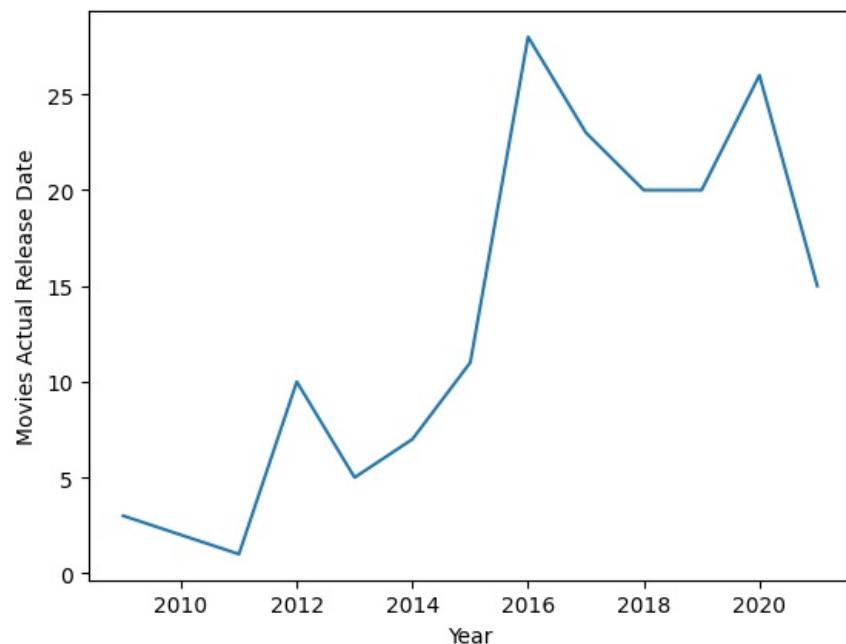


```
In [256]: df_month=df_sk_shows.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("TV Shows Released in the Month")
plt.xlabel("Month")
plt.show()
```



TV Shows are added in Netflix by significant numbers in May and January in South Korea

```
In [257]: df_release_year=df_sk_shows[df_sk_shows['release_year']>=1980].groupby(['release_year']).agg({"title":"nunique"})
sns.lineplot(data=df_release_year, x='release_year', y='title')
plt.ylabel("Movies Actual Release Date")
plt.xlabel("Year")
plt.show()
```



The number of TV Shows in S.Korea reached peak in 2016. It then reached a second peak in 2019. It has reduced in 2021 from 2020.

Recommendations

- 1) The most popular Genres across the countries and in both TV Shows and Movies are Drama, Comedy and International TV Shows/Movies, so content aligning to that is recommended.
 - 2) Add TV Shows in July/August and Movies in last week of the year/first month of the next year.
 - 3) For USA audience 80-120 mins is the recommended length for movies and Kids TV Shows are also popular along with the genres in first point, hence recommended.
 - 4) For UK audience, recommended length for movies is same as that of USA (80-120 mins)
 - 5) The target audience in USA and India is recommended to be 14+ and above ratings while for UK, its recommended to be completely Mature/R content .
 - 6) Add movies for Indian Audience, it has been declining since 2018.
 - 7) Anime Genre for Japan and Romantic Genre in TV Shows for South Korean audiences is recommended.
 - 8) While creating content, take into consideration the popular actors/directors for that country. Also take into account the director-actor combination which is highly recommended.