SCALER NETFLIX EXPLORATORY DATA ANALYSIS

About NETFLIX

Netflix is one of the most popular media and video streaming platforms. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally. This tabular dataset 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.

Business Problem

Analyze the data and generate insights that could help Netflix ijn deciding which type of shows/movies to produce and how they can grow the business in different countries

PROBLEM STATEMENT

>

In the Netflix data Observe the attributes and data given.

In [ ]:
```
#importing required libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [ ]:
```
#loading the dataset
Ndf=pd.read_csv('/content/netflix_dataset.csv')
```

In [ ]:
```
Ndf.head()
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duratio
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 mi
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Seasor
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Seaso
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Seaso
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Seasor

In [ ]:
```
Ndf.describe
```

Out[ ]:
```
<bound method NDFrame.describe of      show_id     type                  title
director  \
0         s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show          Blood & Water              NaN
2         s3  TV Show              Ganglands  Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans              NaN
4         s5  TV Show           Kota Factory              NaN
...      ...      ...                    ...              ...
8802   s8803    Movie                 Zodiac    David Fincher
8803   s8804  TV Show            Zombie Dumb              NaN
8804   s8805    Movie             Zombieland  Ruben Fleischer
8805   s8806    Movie                   Zoom     Peter Hewitt
8806   s8807    Movie                Zubaan      Mozez Singh

                                                   cast        country  \
0                                                   NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                   NaN            NaN
4     Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India
...                                                 ...            ...
8802  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8803                                                NaN            NaN
8804  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8805  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8806  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India

              date_added  release_year rating   duration  \
0     September 25, 2021          2020  PG-13     90 min
1     September 24, 2021          2021  TV-MA  2 Seasons
2     September 24, 2021          2021  TV-MA   1 Season
3     September 24, 2021          2021  TV-MA   1 Season
4     September 24, 2021          2021  TV-MA  2 Seasons
...                  ...           ...    ...        ...
8802   November 20, 2019          2007      R    158 min
8803        July 1, 2019          2018  TV-Y7  2 Seasons
8804    November 1, 2019          2009      R     88 min
8805    January 11, 2020          2006     PG     88 min
8806       March 2, 2019          2015  TV-14    111 min

                                             listed_in  \
0                                       Documentaries
1         International TV Shows, TV Dramas, TV Mysteries
2     Crime TV Shows, International TV Shows, TV Act...
3                               Docuseries, Reality TV
4     International TV Shows, Romantic TV Shows, TV ...
...                                                 ...
8802                Cult Movies, Dramas, Thrillers
8803          Kids' TV, Korean TV Shows, TV Comedies
8804                        Comedies, Horror Movies
8805            Children & Family Movies, Comedies
8806    Dramas, International Movies, Music & Musicals

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

```
8802  A political cartoonist, a crime reporter and a...
8803  While living alone in a spooky town, a young g...
8804  Looking to survive in a world taken over by zo...
8805  Dragged from civilian life, a former superhero...
8806  A scrappy but poor boy worms his way into a ty...

[8807 rows x 12 columns]>
```

In [ ]:
```
Ndf.shape
```

Out[ ]:
```
(8807, 12)
```

In [ ]:
```
Ndf.dtypes
```

Out[ ]:
```
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 [ ]:
```
Ndf.isna().sum()
```

Out[ ]:
```
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
```

Director, cast, country, release_year has null values

In [ ]:
```
#Understanding unique values in each attribute

for col_id in Ndf:
  print(col_id ,':' , Ndf[col_id].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 [ ]:
```
#Understanding each unique value in each attribute
for col_id in Ndf:
  print(col_id)
  print(col_id ,':' , Ndf[col_id].value_counts())
  print('--'*50)
```

```
show_id
show_id : s1       1
s5875    1
s5869    1
s5870    1
s5871    1
        ..
s2931    1
s2930    1
s2929    1
s2928    1
s8807    1
Name: show_id, Length: 8807, dtype: int64
--------------------------------------------------------------------------------
---------------
type
type : Movie      6131
TV Show    2676
Name: type, dtype: int64
--------------------------------------------------------------------------------
---------------
title
title : Dick Johnson Is Dead                     1
Ip Man 2                                  1
Hannibal Buress: Comedy Camisado          1
Turbo FAST                                1
Masha's Tales                             1
                                         ..
Love for Sale 2                           1
ROAD TO ROMA                              1
Good Time                                 1
Captain Underpants Epic Choice-o-Rama     1
Zubaan                                    1
Name: title, Length: 8807, dtype: int64
--------------------------------------------------------------------------------
---------------
director
director : Rajiv Chilaka                      19
Raúl Campos, Jan Suter            18
Marcus Raboy                      16
Suhas Kadav                       16
Jay Karas                         14
                                  ..
Raymie Muzquiz, Stu Livingston     1
Joe Menendez                       1
Eric Bross                         1
Will Eisenberg                     1
Mozez Singh                        1
Name: director, Length: 4528, dtype: int64
--------------------------------------------------------------------------------
---------------
cast
cast : David Attenborough
19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapn
il
14
Samuel West
10
Jeff Dunham
```

```
7
David Spade, London Hughes, Fortune Feimster
6

..
Michael Peña, Diego Luna, Tenoch Huerta, Joaquin Cosio, José María Yazpik, Matt Letsc
her, Alyssa Diaz
1
Nick Lachey, Vanessa Lachey
1
Takeru Sato, Kasumi Arimura, Haru, Kentaro Sakaguchi, Takayuki Yamada, Kendo Kobayash
i, Ken Yasuda, Arata Furuta, Suzuki Matsuo, Koichi Yamadera, Arata Iura, Chikako Kak
u, Kotaro Yoshida     1
Toyin Abraham, Sambasa Nzeribe, Chioma Chukwuka Akpotha, Chioma Omeruah, Chiwetalu Ag
u, Dele Odule, Femi Adebayo, Bayray McNwizu, Biodun Stephen
1
Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malk
eet Rauni, Anita Shabdish, Chittaranjan Tripathy
1
Name: cast, Length: 7692, dtype: int64
--------------------------------------------------------------------------------
---------------
country
country : United States                              2818
India                                         972
United Kingdom                                419
Japan                                         245
South Korea                                   199
                                             ...
Romania, Bulgaria, Hungary                      1
Uruguay, Guatemala                              1
France, Senegal, Belgium                        1
Mexico, United States, Spain, Colombia          1
United Arab Emirates, Jordan                    1
Name: country, Length: 748, dtype: int64
--------------------------------------------------------------------------------
---------------
date_added
date_added : January 1, 2020      109
November 1, 2019      89
March 1, 2018         75
December 31, 2019     74
October 1, 2018       71
                     ...
December 4, 2016       1
November 21, 2016      1
November 19, 2016      1
November 17, 2016      1
January 11, 2020       1
Name: date_added, Length: 1767, dtype: int64
--------------------------------------------------------------------------------
---------------
release_year
release_year : 2018    1147
2017    1032
2019    1030
2020     953
2016     902
        ...
1959       1
```

```
1925       1
1961       1
1947       1
1966       1
Name: release_year, Length: 74, dtype: int64
--------------------------------------------------------------------------------
---------------
rating
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: rating, dtype: int64
--------------------------------------------------------------------------------
---------------
duration
duration : 1 Season     1793
2 Seasons     425
3 Seasons     199
90 min        152
94 min        146
              ...
16 min          1
186 min         1
193 min         1
189 min         1
191 min         1
Name: duration, Length: 220, dtype: int64
--------------------------------------------------------------------------------
---------------
listed_in
listed_in : Dramas, International Movies                          362
Documentaries                                        359
Stand-Up Comedy                                      334
Comedies, Dramas, International Movies               274
Dramas, Independent Movies, International Movies     252
                                                    ...
Kids' TV, TV Action & Adventure, TV Dramas             1
TV Comedies, TV Dramas, TV Horror                      1
Children & Family Movies, Comedies, LGBTQ Movies       1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows     1
Cult Movies, Dramas, Thrillers                         1
Name: listed_in, Length: 514, dtype: int64
--------------------------------------------------------------------------------
---------------
description
description : Paranormal activity at a lush, abandoned property alarms a group eager
```

```
to redevelop the site, but the eerie events may not be as unearthly as they think.
4
Challenged to compose 100 songs before he can marry the girl he loves, a tortured but
passionate singer-songwriter embarks on a poignant musical journey.    3
A surly septuagenarian gets another chance at her 20s after having her photo snapped
at a studio that magically takes 50 years off her life.                3
Multiple women report their husbands as missing but when it appears they are looking
for the same man, a police officer traces their cryptic connection.    3
Secrets bubble to the surface after a sensual encounter and an unforeseen crime entan
gle two friends and a woman caught between them.                       2

..
Sent away to evade an arranged marriage, a 14-year-old begins a harrowing journey of
sex work and poverty in the slums of Accra.                            1
When his partner in crime goes missing, a small-time crook’s life is transformed as h
e dedicates himself to raising the daughter his friend left behind.    1
During 1962’s Cuban missile crisis, a troubled math genius finds himself drafted to p
lay in a U.S.-Soviet chess match – and a deadly game of espionage.     1
A teen's discovery of a vintage Polaroid camera develops into a darker tale when she
finds that whoever takes their photo with it dies soon afterward.      1
A scrappy but poor boy worms his way into a tycoon's dysfunctional family, while faci
ng his fear of music and the truth about his past.                     1
Name: description, Length: 8775, dtype: int64
--------------------------------------------------------------------------------
---------------
```

In the given dataset we have show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in description attributes have to normalised

In the data, there are null values in country, directors, cast and release year, These will be preprocessed.

Preprocessing

In [ ]:
```
Ndf['director']
```

Out[ ]:
```
0       Kirsten Johnson
1                   NaN
2       Julien Leclercq
3                   NaN
4                   NaN
             ...
8802      David Fincher
8803                NaN
8804    Ruben Fleischer
8805       Peter Hewitt
8806        Mozez Singh
Name: director, Length: 8807, dtype: object
```

In [ ]:
```
#unnesting the directors by splitting them with , constraint

c1=Ndf['director'].apply(lambda x : str(x).split(', ')).tolist()
#print(c1)
Ndf1=pd.DataFrame(c1,index=Ndf['title'])
Ndf1=Ndf1.stack()
Ndf1=pd.DataFrame(Ndf1.reset_index())
Ndf1.rename(columns={0:'Director'},inplace=True)
```

```
Ndf1.drop(['level_1'],axis=1,inplace=True)
Ndf1.head()
```

Out[ ]:

	title	Director
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	nan
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	nan
4	Kota Factory	nan

In [ ]:

```
#unnesting the dataframe with cast by splitting them with , constraint
c2=Ndf['cast'].apply(lambda x : str(x).split(', ')).tolist()
#print(c2)
Ndf2=pd.DataFrame(c2,index=Ndf['title'])
Ndf2=Ndf2.stack()
Ndf2=pd.DataFrame(Ndf2.reset_index())
Ndf2.rename(columns={0:'Actors'},inplace=True)
Ndf2.drop(['level_1'],axis=1,inplace=True)
Ndf2.head()
```

Out[ ]:

	title	Actors
0	Dick Johnson Is Dead	nan
1	Blood & Water	Ama Qamata
2	Blood & Water	Khosi Ngema
3	Blood & Water	Gail Mabalane
4	Blood & Water	Thabang Molaba

In [ ]:

```
#unnesting the listed_in column by splitting them with , constarint

c3=Ndf['listed_in'].apply(lambda x : str(x).split(', ')).tolist()
#print(c3)
Ndf3=pd.DataFrame(c3,index=Ndf['title'])
Ndf3=Ndf3.stack()
Ndf3=pd.DataFrame(Ndf3.reset_index())
Ndf3.rename(columns={0:'Genres'},inplace=True)
Ndf3.drop(['level_1'],axis=1,inplace=True)
Ndf3.head()
```

Out[ ]:

	title	Genres
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Blood & Water	TV Dramas
3	Blood & Water	TV Mysteries
4	Ganglands	Crime TV Shows

In [ ]:
```
#unnesting the dataframe with country by splitting them with , constraint
c4=Ndf['country'].apply(lambda x : str(x).split(', ')).tolist()
#print(c4)
Ndf4=pd.DataFrame(c4,index=Ndf['title'])
Ndf4=Ndf4.stack()
Ndf4=pd.DataFrame(Ndf4.reset_index())
Ndf4.rename(columns={0:'country'},inplace=True)
Ndf4.drop(['level_1'],axis=1,inplace=True)
Ndf4.head()
```

Out[ ]:

	title	country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	nan
3	Jailbirds New Orleans	nan
4	Kota Factory	India

In [ ]:
```
# merging the Original datframe with these constraints on title as key
# If the actors, director are null they replaced with unknown Actor and Unknown Direct
# The null countries are were replace with np.nan

Ndf5=Ndf2.merge(Ndf1,on=['title'],how='inner')
Ndf6=Ndf5.merge(Ndf3,on=['title'],how='inner')
Ndf7=Ndf6.merge(Ndf4,on=['title'],how='inner')

Ndf7['Actors'].replace(['nan'],['unknown Actor'],inplace=True)
Ndf7['Director'].replace(['nan'],['Unknown Director'],inplace=True)
Ndf7['country'].replace(['nan'],[np.nan],inplace=True)

Ndf7.head()
```

Out[ ]:

	title	Actors	Director	Genres	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
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa

In [ ]:
```
Ndf7.shape
```

Out[ ]:
```
(201991, 5)
```

In [ ]:
```
Ndf7.isna().sum()
```

Out[ ]:
```
title           0
Actors          0
Director        0
Genres          0
country     11897
dtype: int64
```

In the Temperary dataframe we have 5 attributes in which country attribute have null values, this temporary dataframe will be merged with original dataframe in order to normalise.

In [ ]:
```
df=Ndf7.merge(Ndf[['show_id', 'type', 'title', 'date_added',
       'release_year', 'rating', 'duration',]],on=['title'],how='left')
df.head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
df.shape
```

Out[ ]:
```
(201991, 11)
```

In [ ]:
```
df.isna().sum()
```

Out[ ]:
```
title              0
Actors             0
Director           0
Genres             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 such cases the rating were replaced with NR ie No Rating.

In [ ]:
```
df.loc[df['duration'].isnull(),'duration']=df.loc[df['duration'].isnull(),'duration'].
df.loc[df['rating'].str.contains('min',na=False),'rating']='NR'
df.isnull().sum()
```

Out[ ]:
```
title              0
Actors             0
Director           0
Genres             0
country        11897
show_id            0
type               0
date_added       158
release_year       0
rating            67
duration           0
dtype: int64
```

In [ ]:
```
#Ratings can't be in min, so it has been made NR(i.e- Non Rated)
df.loc[df['rating'].str.contains('min', na=False),'rating']='NR'
df['rating'].fillna('NR',inplace=True)
pd.set_option('display.max_rows',None)
```

In [ ]:
```
df.isnull().sum()
```

Out[ ]:
```
title              0
Actors             0
Director           0
Genres             0
country        11897
show_id            0
type               0
date_added       158
release_year       0
rating             0
duration           0
dtype: int64
```

In [ ]:
```
df[df['date_added'].isnull()].head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year
136893	A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	British TV Shows	United Kingdom	s6067	TV Show	NaN	2013
136894	A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	TV Comedies	United Kingdom	s6067	TV Show	NaN	2013
136895	A Young Doctor's Notebook and Other Stories	Daniel Radcliffe	Unknown Director	TV Dramas	United Kingdom	s6067	TV Show	NaN	2013
136896	A Young Doctor's Notebook and Other Stories	Jon Hamm	Unknown Director	British TV Shows	United Kingdom	s6067	TV Show	NaN	2013
136897	A Young Doctor's Notebook and Other Stories	Jon Hamm	Unknown Director	TV Comedies	United Kingdom	s6067	TV Show	NaN	2013

In [ ]:
```
#date added is based on releases year , based on release yead we can add release for a

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

In [ ]:
```
df.dtypes
```

Out[ ]:
```
title           object
Actors          object
Director        object
Genres          object
country         object
show_id         object
type            object
date_added      object
release_year     int64
rating          object
duration        object
dtype: object
```

In [ ]:
```
#country of a movie is depicted by the director
# for a given country find the director and find the mode of countries in which direct
# impute the data

for i in df[df['country'].isnull()]['Director'].unique():
  if i in df[~df['country'].isnull()]['Director'].unique():
    coun=df[df['Director']==i]['country'].mode().values[0]
    df.loc[df['Director']==i,'country']=df.loc[df['Director']==i,'country'].fillna(cou
```

In [ ]:
```
df.isnull().sum()
```

Out[ ]:
```
title              0
Actors             0
Director           0
Genres             0
country         4276
show_id            0
type               0
date_added         0
release_year       0
rating             0
duration           0
dtype: int64
```

In [ ]:
```
#there are countries which are null in director too
# we can use actors mode to find the country name

for i in df[df['country'].isnull()]['Actors'].unique():
  if i in df[~df['country'].isnull()]['Actors'].unique():
    coun1=df[df['Actors']==i]['country'].mode().values[0]
    df.loc[df['Actors']==i,'country']=df.loc[df['Actors']==i,'country'].fillna(coun1)

df['country'].fillna('Unknown Country',inplace = True)
df.isnull().sum()
```

Out[ ]:
```
title           0
Actors          0
Director        0
Genres          0
country         0
show_id         0
type            0
date_added      0
release_year    0
rating          0
duration        0
dtype: int64
```

In [ ]:
```
df.shape
```

Out[ ]:
```
(201991, 11)
```

In [ ]:
```
#minutes can be removed from the data
df['duration']=df['duration'].str.replace("min",'')
df.head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
#we have season in duration we have to get new value for them
#we copy the duration attribute into another column and perform the operatiom

df['duration_c']=df['duration'].copy()
df1=df.copy()
```

In [ ]:
```
df1.loc[df1['duration_c'].str.contains('Season'),'duration_c']=0
df1['duration_c']=df1['duration_c'].astype('int')
df1.head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
df1['duration_c'].describe()
```

Out[ ]:
```
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_c, dtype: float64
```

In [ ]:
```
import seaborn as sns
sns.distplot(df1['duration_c'],hist=True,kde=True,bins=int(36),color='blue',
hist_kws={'edgecolor' : 'darkblue'}, kde_kws={'linewidth': 2})
plt.show()
```

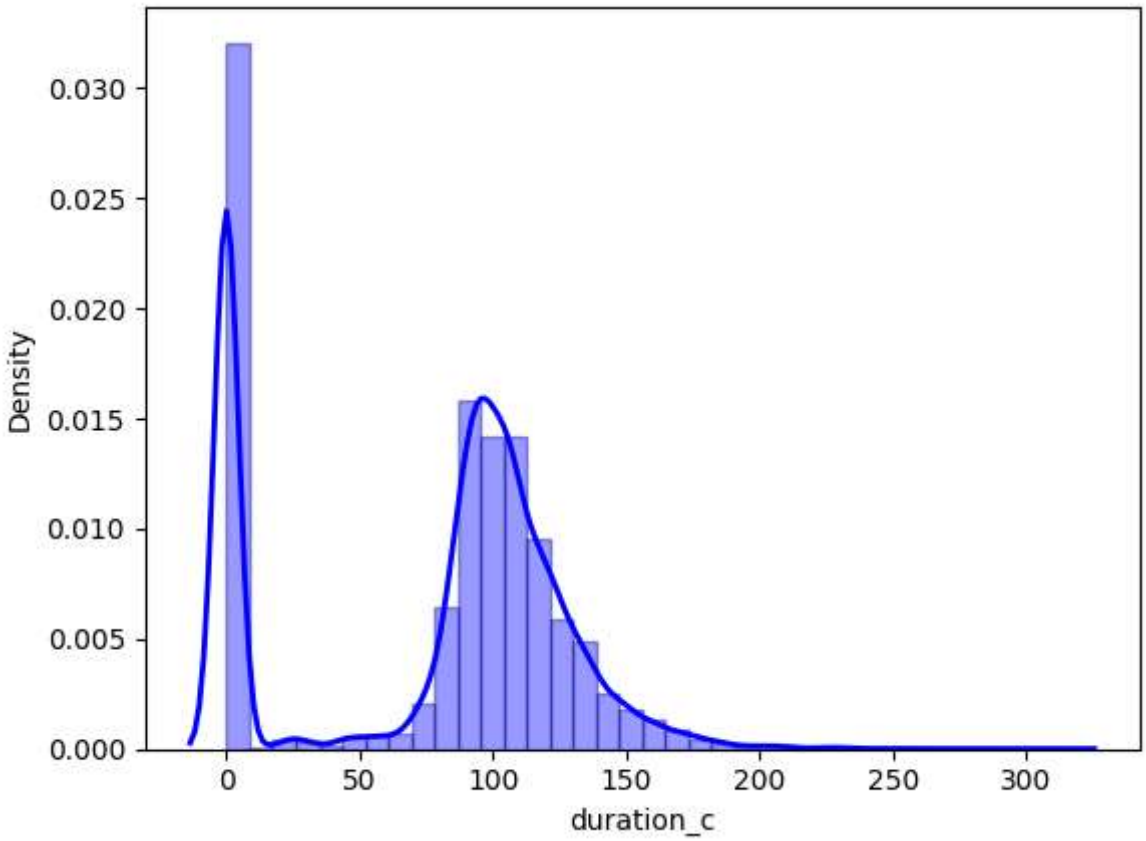
```
<ipython-input-39-1805e03ce925>:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df1['duration_c'],hist=True,kde=True,bins=int(36),color='blue',
```



In [ ]:
```
# for the duration copy label them according to the bins

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']
df1['duration_c']=pd.cut(df1['duration_c'],bins=bins1,labels=labels1)
df1.head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
df1.loc[~df1['duration'].str.contains('Season'),'duration']=df1.loc[~df1['duration'].s
df1.drop(['duration_c'],axis=1,inplace=True)
df1.head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
df1['duration'].value_counts()
```

Out[ ]:
```
80-100        52937
100-120       48724
1 Season      35035
120-150       26691
2 Seasons      9559
50-80          7700
150-200        6737
3 Seasons      5084
1-50           2530
4 Seasons      2134
5 Seasons      1698
7 Seasons       843
6 Seasons       633
200-315         524
8 Seasons       286
9 Seasons       257
10 Seasons      220
13 Seasons      132
12 Seasons      111
15 Seasons       96
17 Seasons       30
11 Seasons       30
Name: duration, dtype: int64
```

In [ ]:
```
#convert date added to the date time format so that we can get the data according to m

from datetime import datetime
from dateutil.parser import parse

arr=[]
for i in df1['date_added'].values:
  dt1=parse(i)
  arr.append(dt1.strftime('%Y-%m-%d'))
df1['Modified_added_date']=arr
df1['Modified_added_date']=pd.to_datetime(df1['Modified_added_date'])
df1['month_added']=df1['Modified_added_date'].dt.month
df1['week_added']=df1['Modified_added_date'].dt.week
df1['year_added']=df1['Modified_added_date'].dt.year
df1.head()
```

```
<ipython-input-43-5720fce22eae>:13: FutureWarning: Series.dt.weekofyear and Series.d
t.week have been deprecated. Please use Series.dt.isocalendar().week instead.
  df1['week_added']=df1['Modified_added_date'].dt.week
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
#few movies have same titles named in different languages like tamil hindi and so on
#so the presence of brackets and special chars in titles are to be removed

df1['title']=df1['title'].str.replace(r"\(.*\)",'')
df1.head()
```

```
<ipython-input-44-52ba41f0f41c>:4: FutureWarning: The default value of regex will cha
nge from True to False in a future version.
  df1['title']=df1['title'].str.replace(r"\(.*\)",'')
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
df1.shape
```

Out[ ]:
```
(201991, 15)
```

The data is completely preprocessed, we can use to visualise the data and analysis

Univariate Analysis

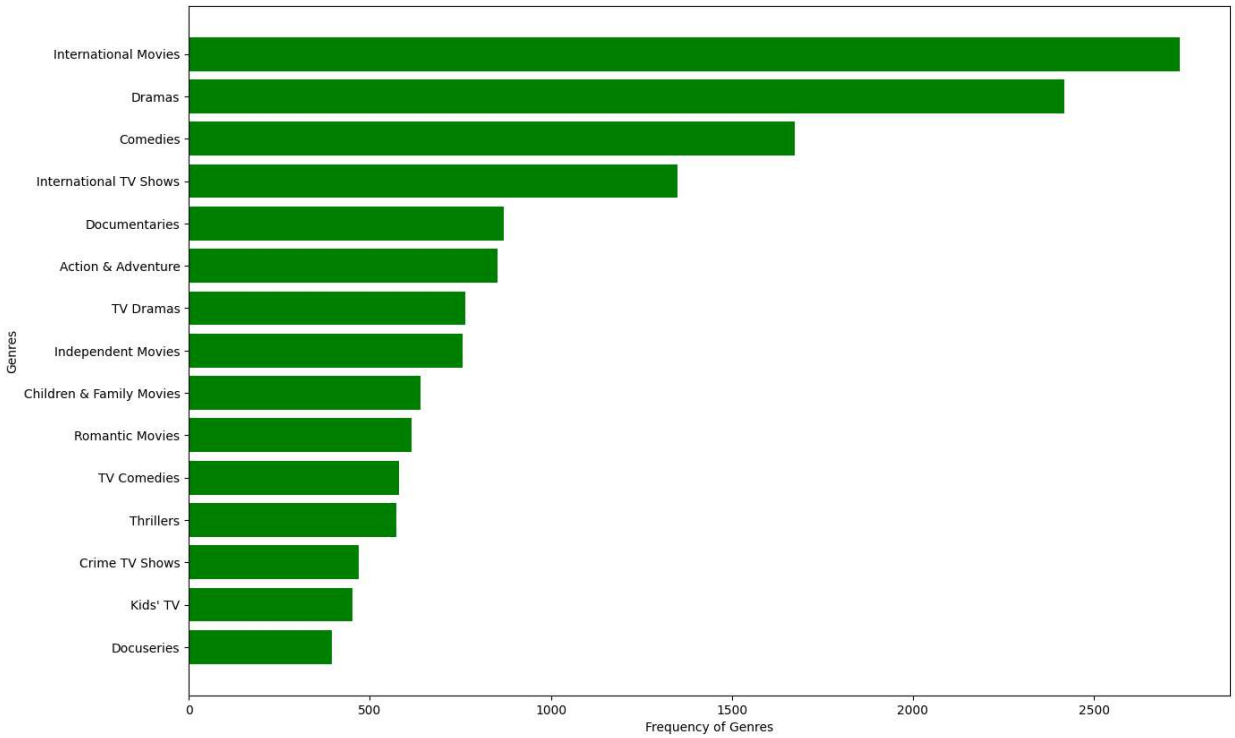
In [ ]:
```
#Number of distinct titles based in genres
df1.groupby(['Genres']).agg({'title':"nunique"})
```

Out[ ]:

Genres	title
Action & Adventure	854
Anime Features	71
Anime Series	176
British TV Shows	253
Children & Family Movies	639
Classic & Cult TV	28
Classic Movies	116
Comedies	1673
Crime TV Shows	470
Cult Movies	71
Documentaries	869
Docuseries	395
Dramas	2418
Faith & Spirituality	65
Horror Movies	353
Independent Movies	756
International Movies	2738
International TV Shows	1351
Kids' TV	451
Korean TV Shows	151
LGBTQ Movies	102
Movies	57
Music & Musicals	372
Reality TV	255
Romantic Movies	615
Romantic TV Shows	370
Sci-Fi & Fantasy	243
Science & Nature TV	92
Spanish-Language TV Shows	174
Sports Movies	219
Stand-Up Comedy	343
Stand-Up Comedy & Talk Shows	56
TV Action & Adventure	168

Genres	title
TV Comedies	581
TV Dramas	763
TV Horror	75
TV Mysteries	98
TV Sci-Fi & Fantasy	84
TV Shows	16
TV Thrillers	57
Teen TV Shows	69
Thrillers	573

In [ ]:
```
df_genre=df1.groupby(['Genres']).agg({'title':'nunique'}).reset_index().sort_values(by
plt.figure(figsize=(15,10))
plt.barh(df_genre[::-1]['Genres'],df_genre[::-1]['title'],color=['green'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```



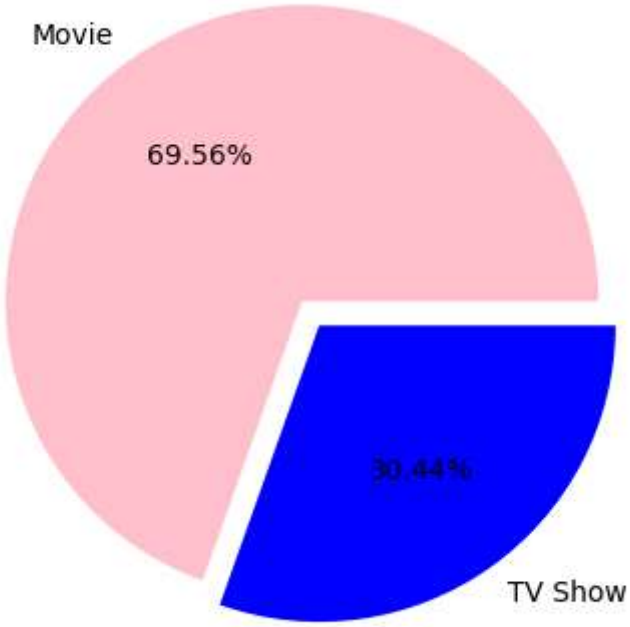
International Movies, Dramas, COmedies are most popular genres in Netflix content

In [ ]:
```
df1.groupby(['type']).agg({'title':'nunique'})
```

Out[ ]:

type	title
Movie	6115
TV Show	2676

In [ ]:
```
df_type=df1.groupby(['type']).agg({'title':'nunique'}).reset_index()
plt.pie(df_type['title'],explode=(0.05,0.05),labels=df_type['type'],colors=['pink','bl
plt.show()
```



In netflix content Movies play a major role than TV shie. we have 69% of movies where as 30 % of TV shows

In [ ]:
```
df_country=df1.groupby(['country']).agg({'title':'nunique'})
df_country
```

Out[ ]:

country	title
	3
Afghanistan	1
Albania	1
Algeria	3
Angola	2
Argentina	94
Armenia	1
Australia	162
Austria	12
Azerbaijan	1
Bahamas	1
Bangladesh	4
Belarus	1
Belgium	94
Bermuda	1
Botswana	1
Brazil	103
Bulgaria	10
Burkina Faso	1
Cambodia	5
Cambodia,	1
Cameroon	2
Canada	460
Cayman Islands	2
Chile	30
China	166
Colombia	54
Croatia	4
Cuba	2
Cyprus	1
Czech Republic	23
Denmark	50
Dominican Republic	1

country	title
East Germany	1
Ecuador	1
Egypt	134
Ethiopia	1
Finland	12
France	409
Georgia	2
Germany	231
Ghana	8
Greece	11
Guatemala	2
Hong Kong	110
Hungary	11
Iceland	11
India	1126
Indonesia	97
Iran	4
Iraq	2
Ireland	46
Israel	30
Italy	102
Jamaica	1
Japan	338
Jordan	10
Kazakhstan	1
Kenya	6
Kuwait	9
Latvia	1
Lebanon	33
Liechtenstein	1
Lithuania	1
Luxembourg	12
Malawi	1

country	title
Malaysia	26
Malta	3
Mauritius	3
Mexico	175
Mongolia	1
Montenegro	1
Morocco	6
Mozambique	1
Namibia	2
Nepal	2
Netherlands	50
New Zealand	33
Nicaragua	1
Nigeria	140
Norway	30
Pakistan	24
Palestine	1
Panama	1
Paraguay	1
Peru	11
Philippines	90
Poland	41
Poland,	1
Portugal	6
Puerto Rico	1
Qatar	10
Romania	14
Russia	27
Samoa	1
Saudi Arabia	14
Senegal	3
Serbia	7
Singapore	41

country	title
Slovakia	1
Slovenia	3
Somalia	1
South Africa	65
South Korea	235
Soviet Union	3
Spain	239
Sri Lanka	1
Sudan	1
Sweden	44
Switzerland	19
Syria	3
Taiwan	94
Thailand	74
Turkey	115
Uganda	1
Ukraine	3
United Arab Emirates	38
United Kingdom	829
United Kingdom,	2
United States	4245
United States,	1
Unknown Country	175
Uruguay	14
Vatican City	1
Venezuela	4
Vietnam	7
West Germany	5
Zimbabwe	3

country names cambodia and usa, uk have been represented more than once we can merge them

In [ ]:
```
df1['country']=df1['country'].str.replace(',','')
df1.head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

In [ ]:
```
df_country=df1.groupby(['country']).agg({'title':'nunique'})
df_country
```

Out[ ]:

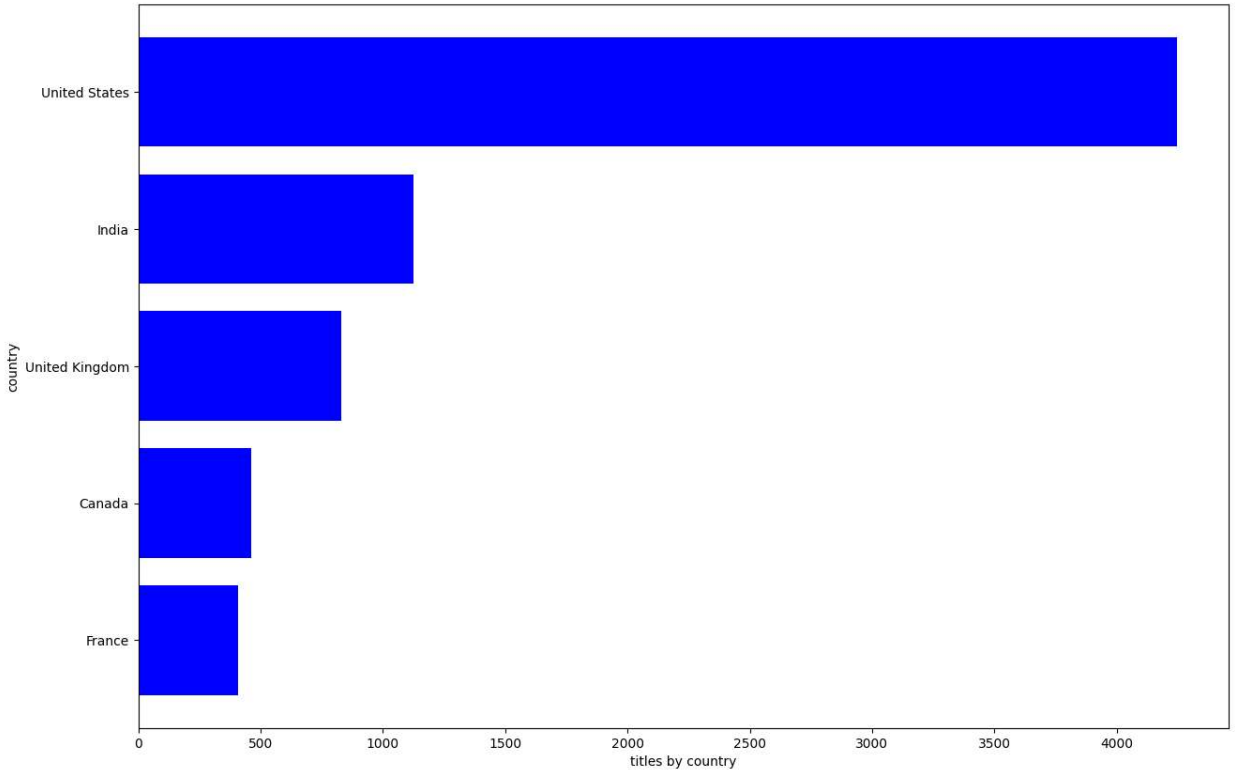
country	title
	3
Afghanistan	1
Albania	1
Algeria	3
Angola	2
Argentina	94
Armenia	1
Australia	162
Austria	12
Azerbaijan	1
Bahamas	1
Bangladesh	4
Belarus	1
Belgium	94
Bermuda	1
Botswana	1
Brazil	103
Bulgaria	10
Burkina Faso	1
Cambodia	6
Cameroon	2
Canada	460
Cayman Islands	2
Chile	30
China	166
Colombia	54
Croatia	4
Cuba	2
Cyprus	1
Czech Republic	23
Denmark	50
Dominican Republic	1
East Germany	1

country	title
Ecuador	1
Egypt	134
Ethiopia	1
Finland	12
France	409
Georgia	2
Germany	231
Ghana	8
Greece	11
Guatemala	2
Hong Kong	110
Hungary	11
Iceland	11
India	1126
Indonesia	97
Iran	4
Iraq	2
Ireland	46
Israel	30
Italy	102
Jamaica	1
Japan	338
Jordan	10
Kazakhstan	1
Kenya	6
Kuwait	9
Latvia	1
Lebanon	33
Liechtenstein	1
Lithuania	1
Luxembourg	12
Malawi	1
Malaysia	26

country	title
Malta	3
Mauritius	3
Mexico	175
Mongolia	1
Montenegro	1
Morocco	6
Mozambique	1
Namibia	2
Nepal	2
Netherlands	50
New Zealand	33
Nicaragua	1
Nigeria	140
Norway	30
Pakistan	24
Palestine	1
Panama	1
Paraguay	1
Peru	11
Philippines	90
Poland	42
Portugal	6
Puerto Rico	1
Qatar	10
Romania	14
Russia	27
Samoa	1
Saudi Arabia	14
Senegal	3
Serbia	7
Singapore	41
Slovakia	1
Slovenia	3

country	title
Somalia	1
South Africa	65
South Korea	235
Soviet Union	3
Spain	239
Sri Lanka	1
Sudan	1
Sweden	44
Switzerland	19
Syria	3
Taiwan	94
Thailand	74
Turkey	115
Uganda	1
Ukraine	3
United Arab Emirates	38
United Kingdom	831
United States	4246
Unknown Country	175
Uruguay	14
Vatican City	1
Venezuela	4
Vietnam	7
West Germany	5
Zimbabwe	3

In [ ]:
```
df_country=df1.groupby(['country']).agg({'title':'nunique'}).reset_index().sort_values
plt.figure(figsize=(15,10))
plt.barh(df_country[::-1]['country'],df_country[::-1]['title'],color='blue')
plt.xlabel('titles by country')
plt.ylabel('country')
plt.show()
```



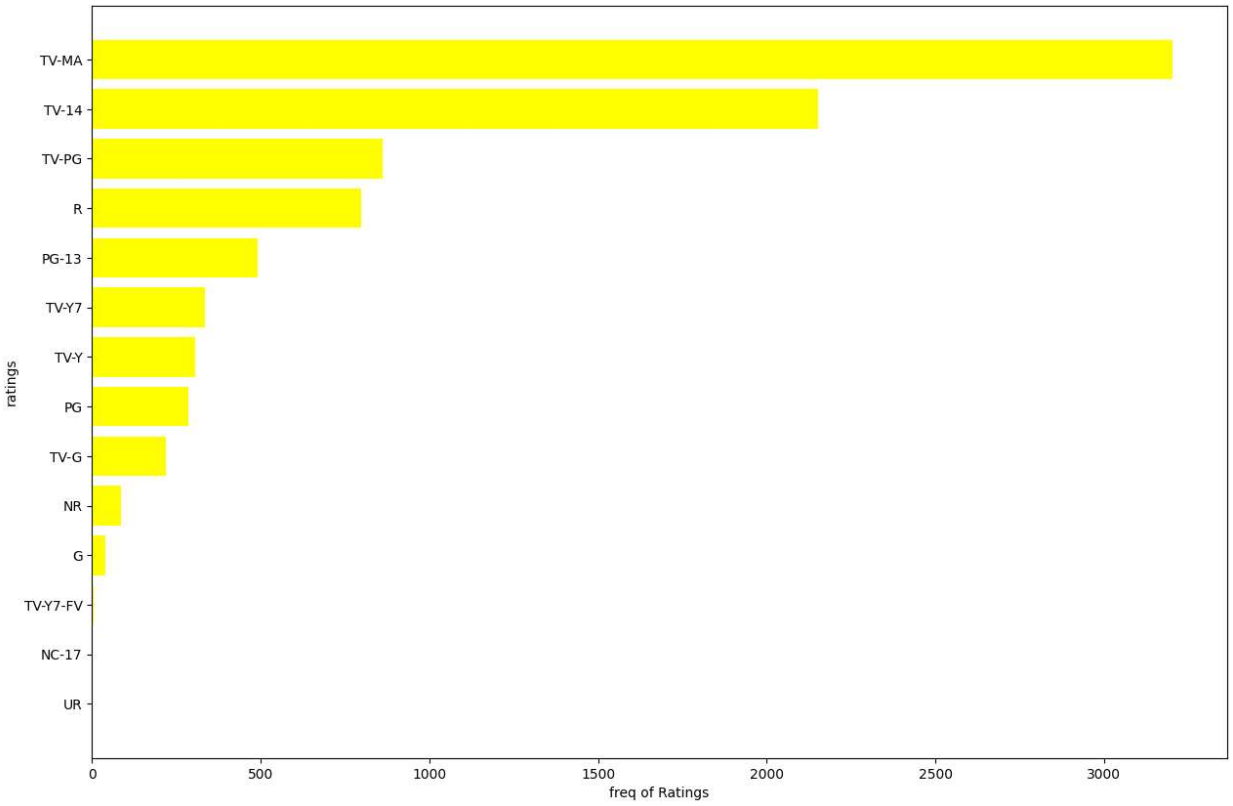
Us, India and UK are leading countries in making of shows in netflix

In [ ]:
```
df1.groupby(['rating']).agg({"title":'nunique'})
```

Out[ ]:

rating	title
G	41
NC-17	3
NR	87
PG	287
PG-13	490
R	799
TV-14	2151
TV-G	220
TV-MA	3204
TV-PG	863
TV-Y	305
TV-Y7	334
TV-Y7-FV	6
UR	3

In [ ]:
```
df_rating=df1.groupby(['rating']).agg({"title":'nunique'}).reset_index().sort_values(b
plt.figure(figsize=(15,10))
plt.barh(df_rating[::-1]['rating'],df_rating[::-1]['title'],color=['yellow'])
plt.xlabel('freq of Ratings')
plt.ylabel('ratings')
plt.show()
```



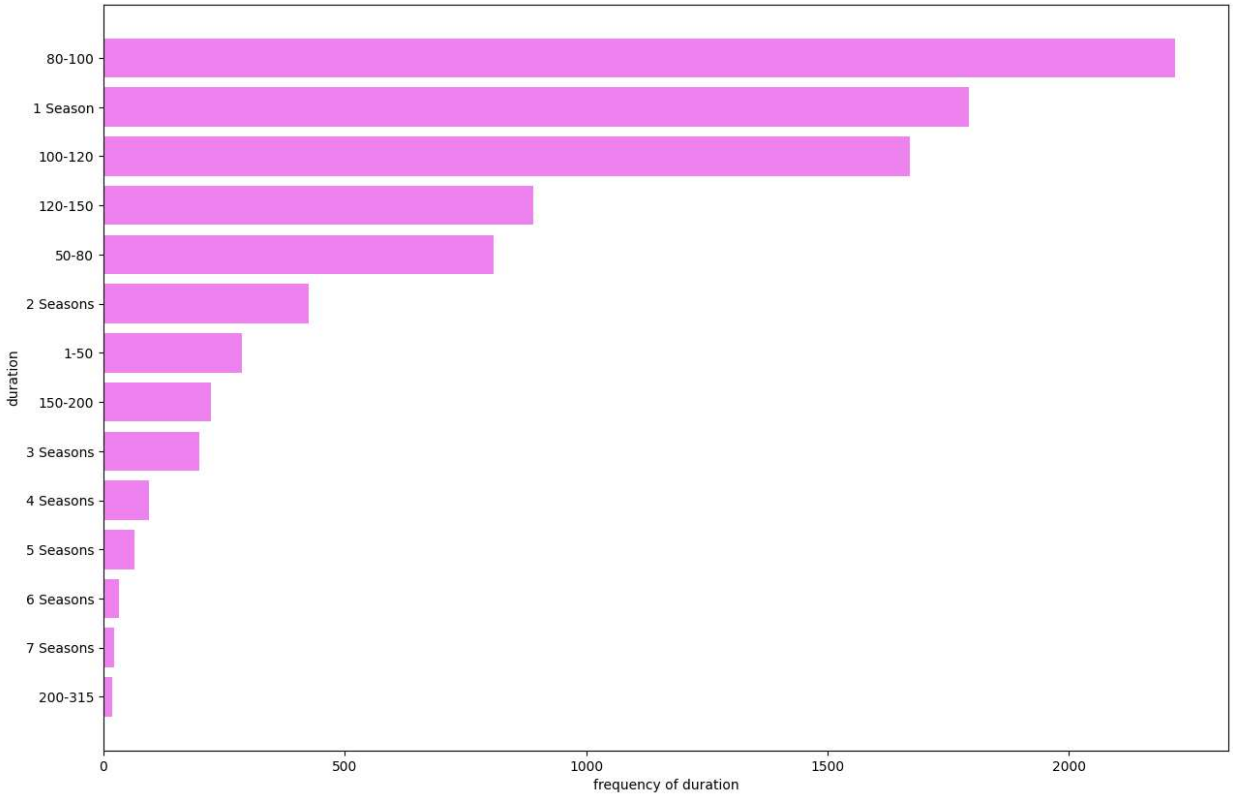
In netflix maturea audience(TV MA) and under 14 are high content

In [ ]:
```
df1.groupby(['duration']).agg({"title":"nunique"})
```

Out[ ]:

duration	title
1 Season	1793
1-50	287
10 Seasons	7
100-120	1671
11 Seasons	2
12 Seasons	2
120-150	891
13 Seasons	3
15 Seasons	2
150-200	222
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	2220
9 Seasons	9

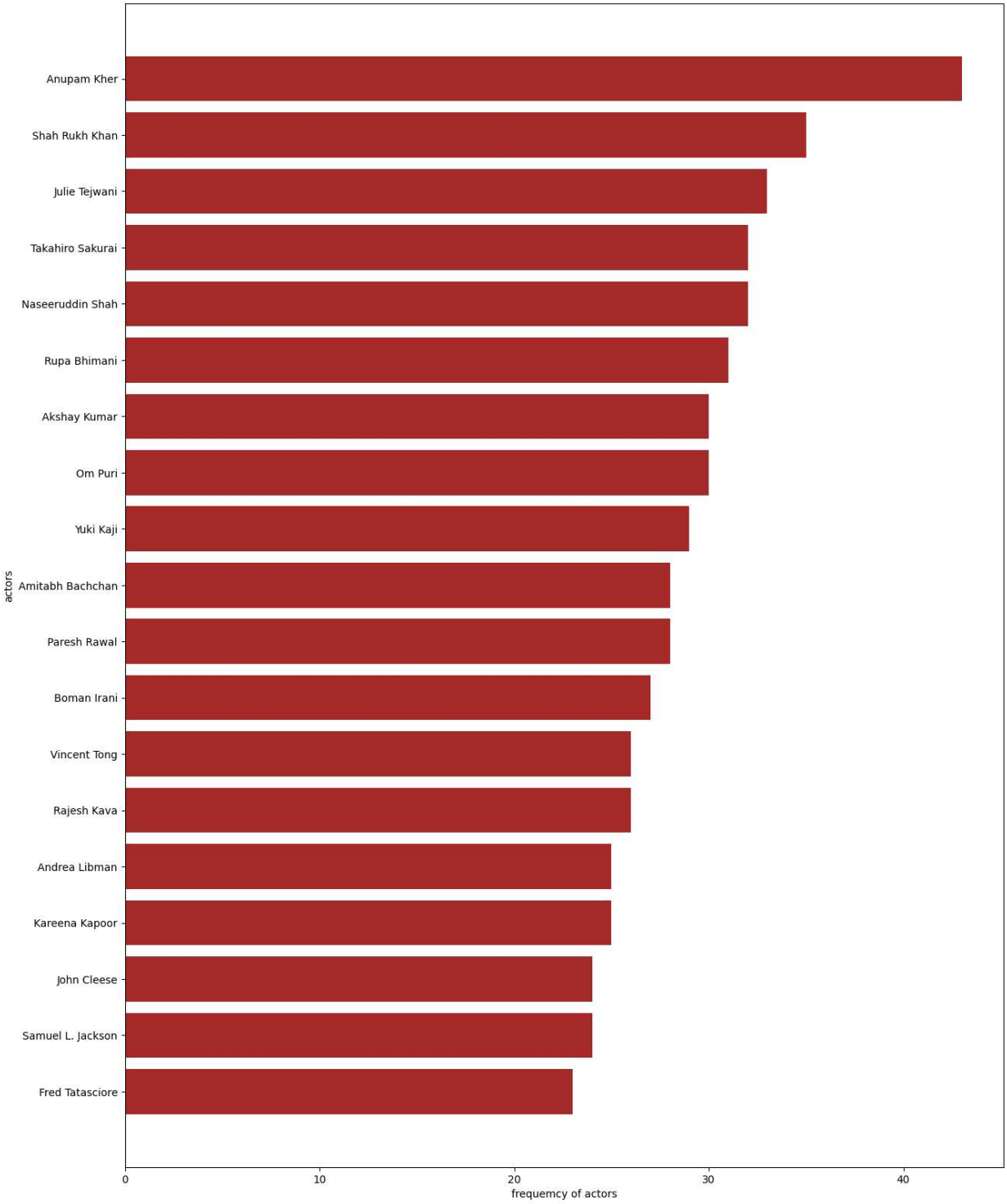
In [ ]:
```
df_duration=df1.groupby(['duration']).agg({'title':'nunique'}).reset_index().sort_valu
plt.figure(figsize=(15,10))
plt.barh(df_duration[::-1]['duration'],df_duration[::-1]['title'],color=['violet'])
plt.xlabel('frequency of duration')
plt.ylabel('duration')
plt.show()
```



> Most shows have duration time of 80 - 100 mins and 100-120 mins.
>
> Most Tv shows have 1 season

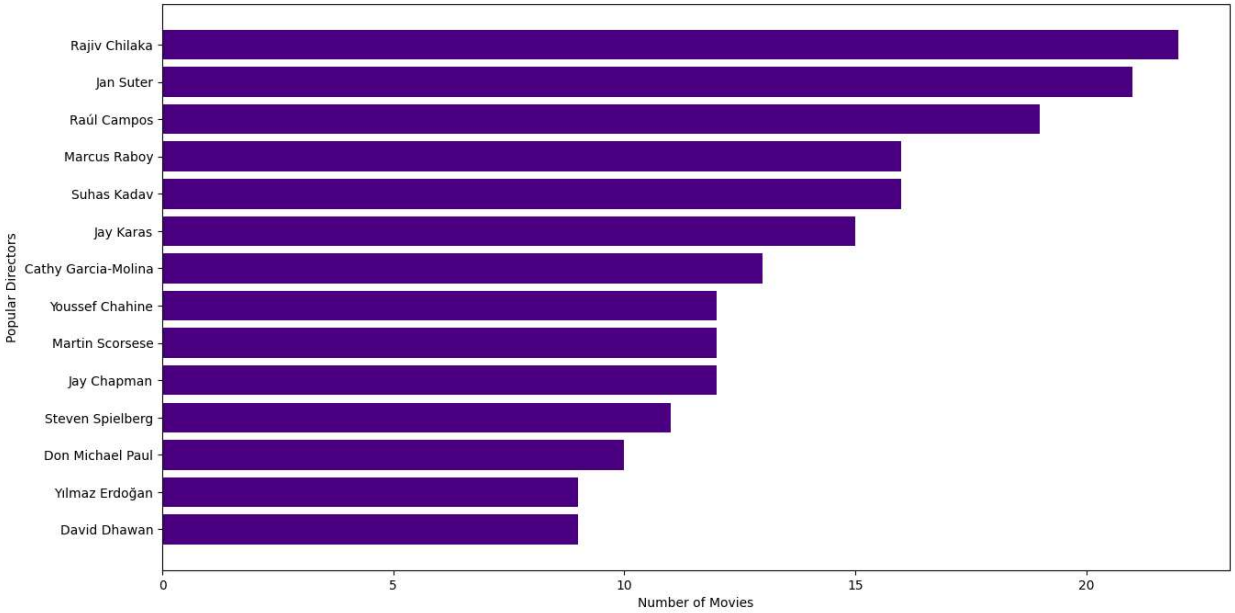
In [ ]:
```
df_actors=df1.groupby(['Actors']).agg({'title':"nunique"}).reset_index().sort_values(b
df_actors=df_actors[df_actors['Actors']!='unknown Actor']
plt.figure(figsize=(15,20))
plt.barh(df_actors[::-1]['Actors'],df_actors[::-1]['title'],color=['brown'])
plt.xlabel('frequemcy of actors')
plt.ylabel('actors')
plt.show
```

Out[ ]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



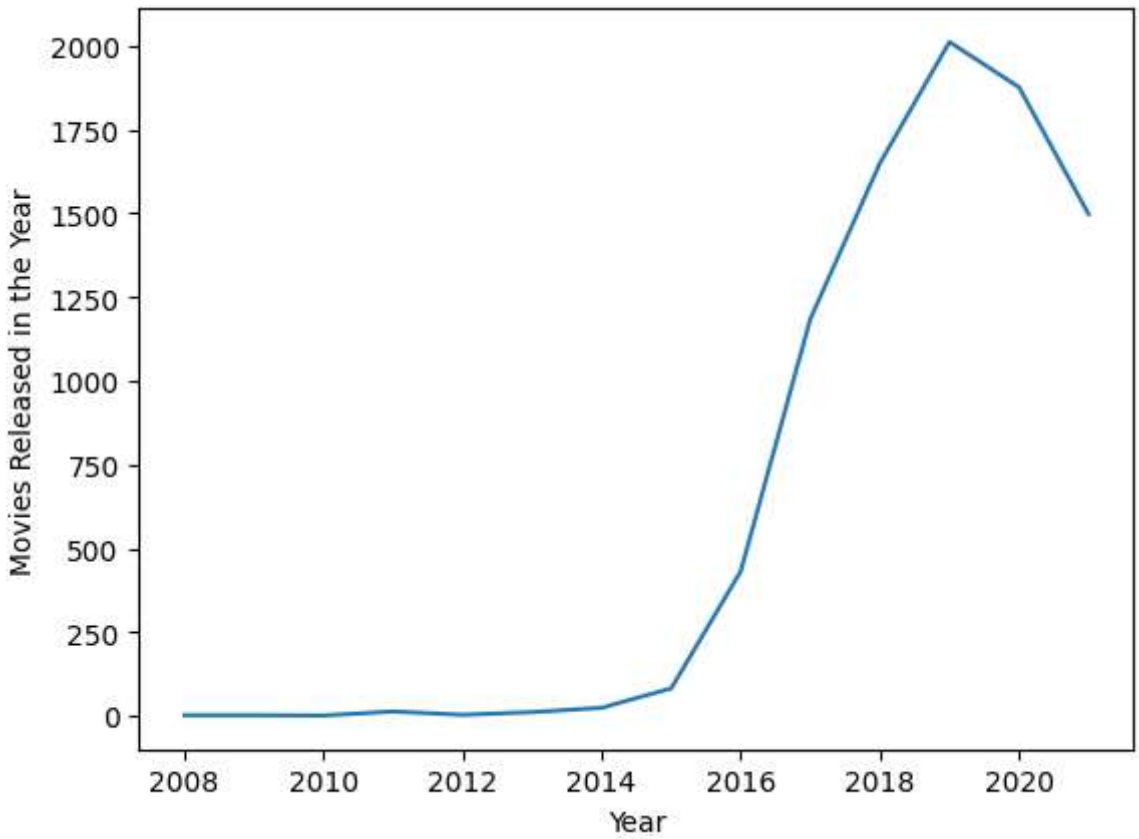
Anupam Kher, Shah Rukh Khan and Julie Tejwani are the most frequent actors in netflix content

In [ ]:
```
df_directors=df1.groupby(['Director']).agg({"title":"nunique"}).reset_index().sort_val
df_directors=df_directors[df_directors['Director']!='Unknown Director']
plt.figure(figsize=(15,8))
plt.barh(df_directors[::-1]['Director'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



Rajiv Chilaka, Jan Suter, Raul Campos are the most frequent director in Netflix content

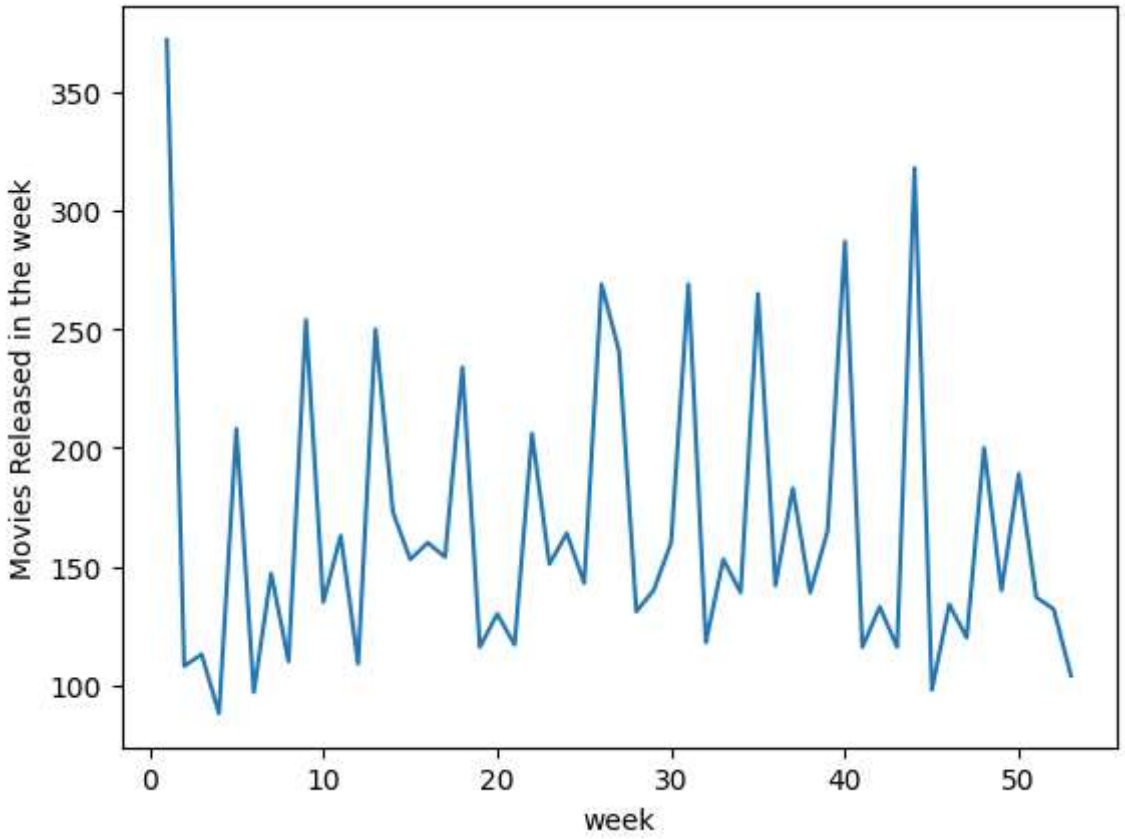
In [ ]:
```
df_year=df1.groupby(['year_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year_added', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



The amount of content added in the netflix has been rapidly increasing from the year 2016 to 2019. Then in started to reduce

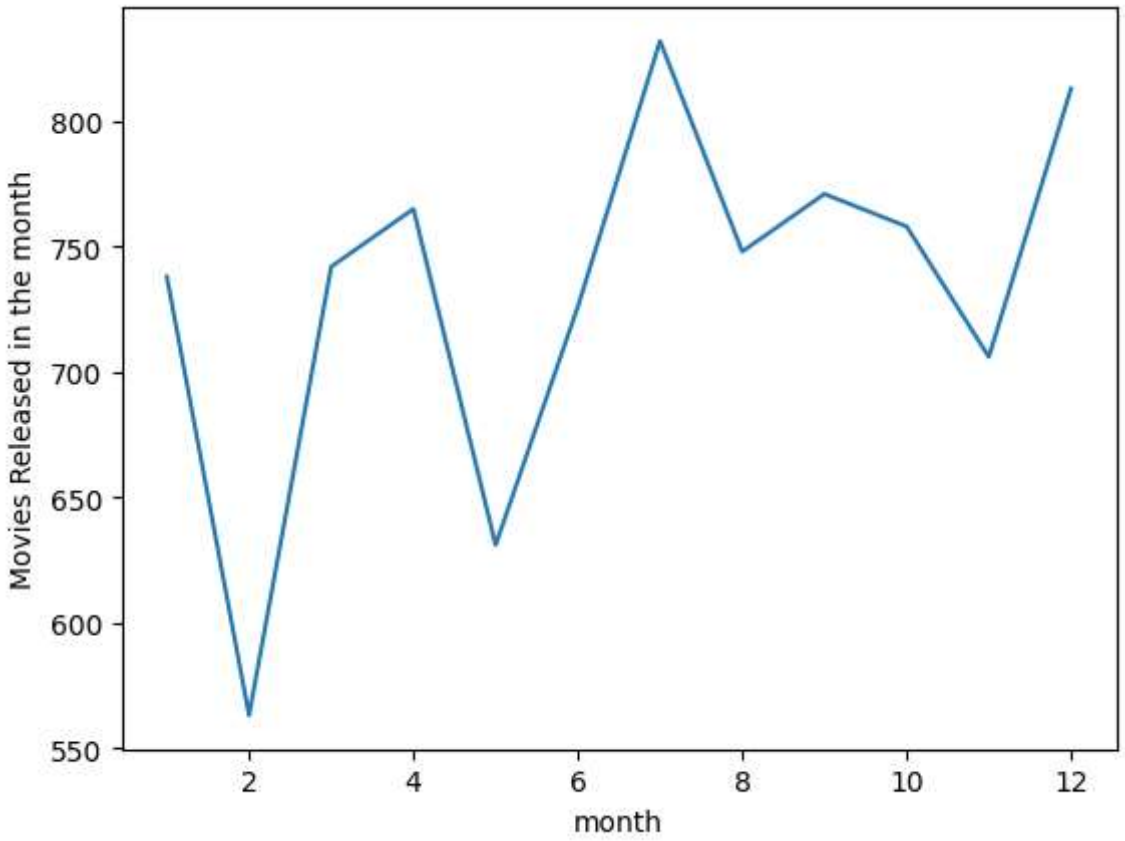
> It can be affected either by covid or external factors like other platforms that are available

In [ ]:
```
df_week=df1.groupby(['week_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_week, x='week_added', y='title')
plt.ylabel("Movies Released in the week")
plt.xlabel("week")
plt.show()
```



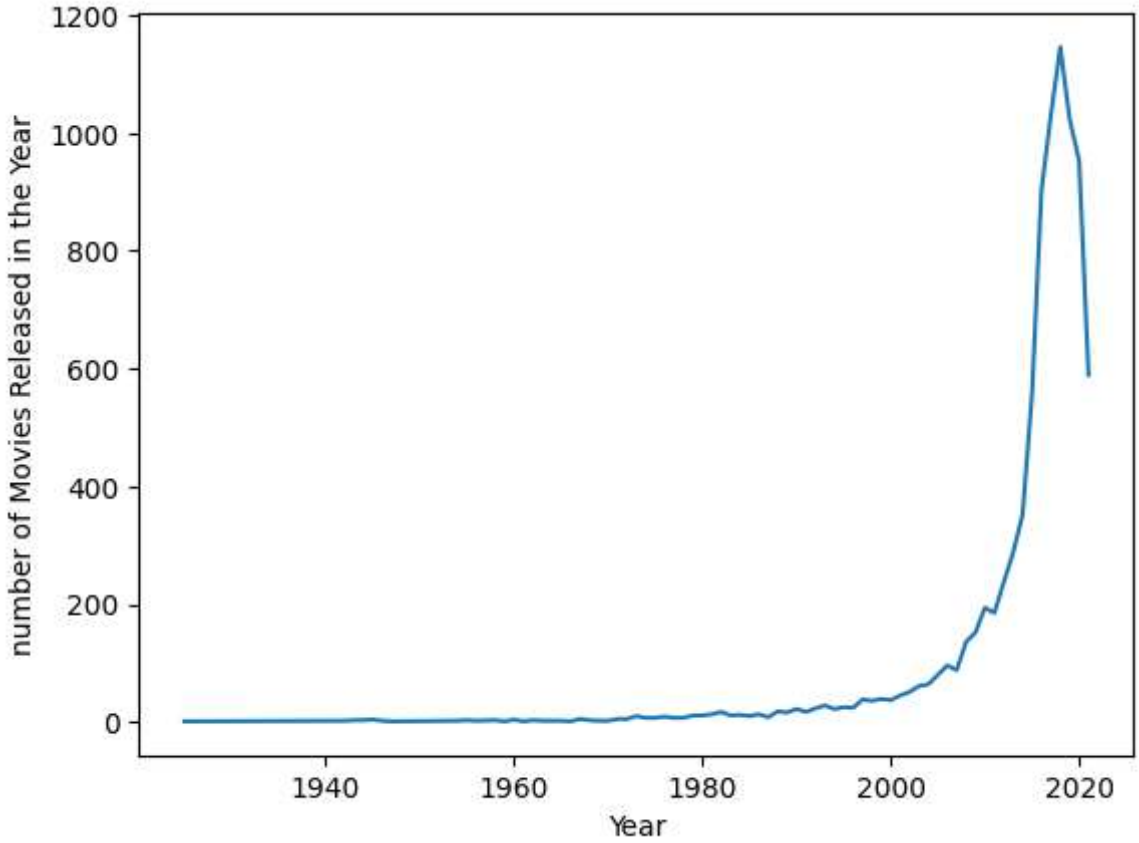
Most content is added in 0th, 40th and 27th week of the year may be because of holidays during that point of time

In [ ]:
```
df_month=df1.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()
```



The content releases is high in the start and end of the year

In [ ]:
```
df_year=df1.groupby(['release_year']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='release_year', y='title')
plt.ylabel("number of Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



Analysis based on movies and shows differently

In [ ]:
```
df_movies=df1[df1['type']=='Movie']
df_shows=df1[df1['type']=='TV Show']
```

In [ ]:
```
df_movies.head()
```

Out[ ]:

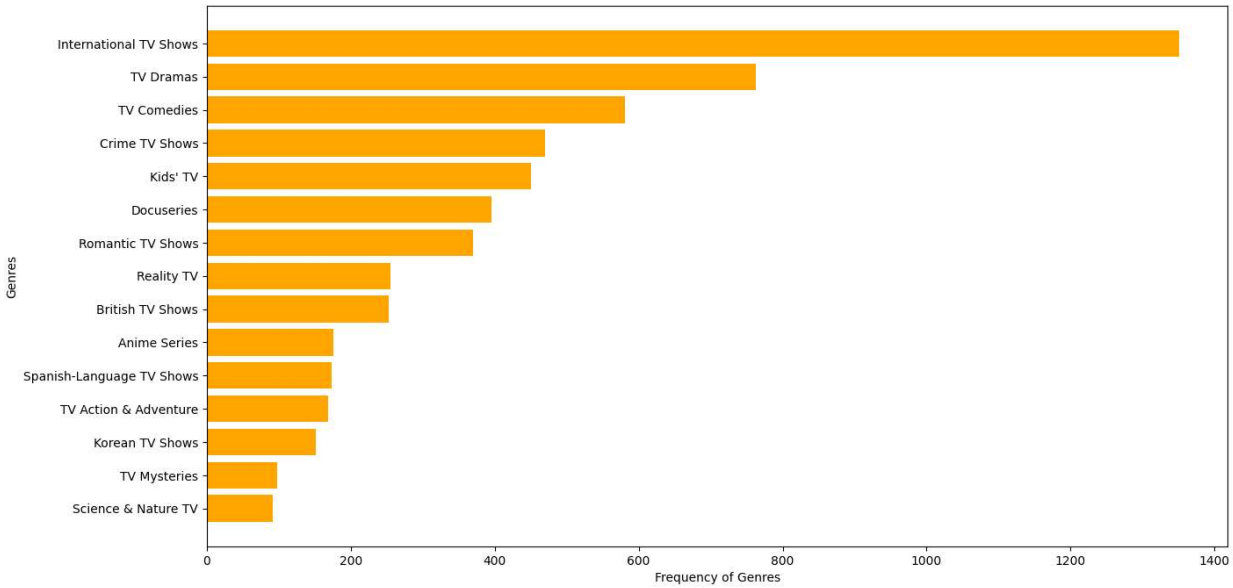
	title	Actors	Director	Genres	country	show_id	type	date_added	release_yea
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	202(
159	My Little Pony: A New Generation	Vanessa Hudgens	Robert Cullen	Children & Family Movies	United States	s7	Movie	September 24, 2021	202
160	My Little Pony: A New Generation	Vanessa Hudgens	José Luis Ucha	Children & Family Movies	United States	s7	Movie	September 24, 2021	202
161	My Little Pony: A New Generation	Kimiko Glenn	Robert Cullen	Children & Family Movies	United States	s7	Movie	September 24, 2021	202
162	My Little Pony: A New Generation	Kimiko Glenn	José Luis Ucha	Children & Family Movies	United States	s7	Movie	September 24, 2021	202

In [ ]:
```
df_shows.head()
```

Out[ ]:

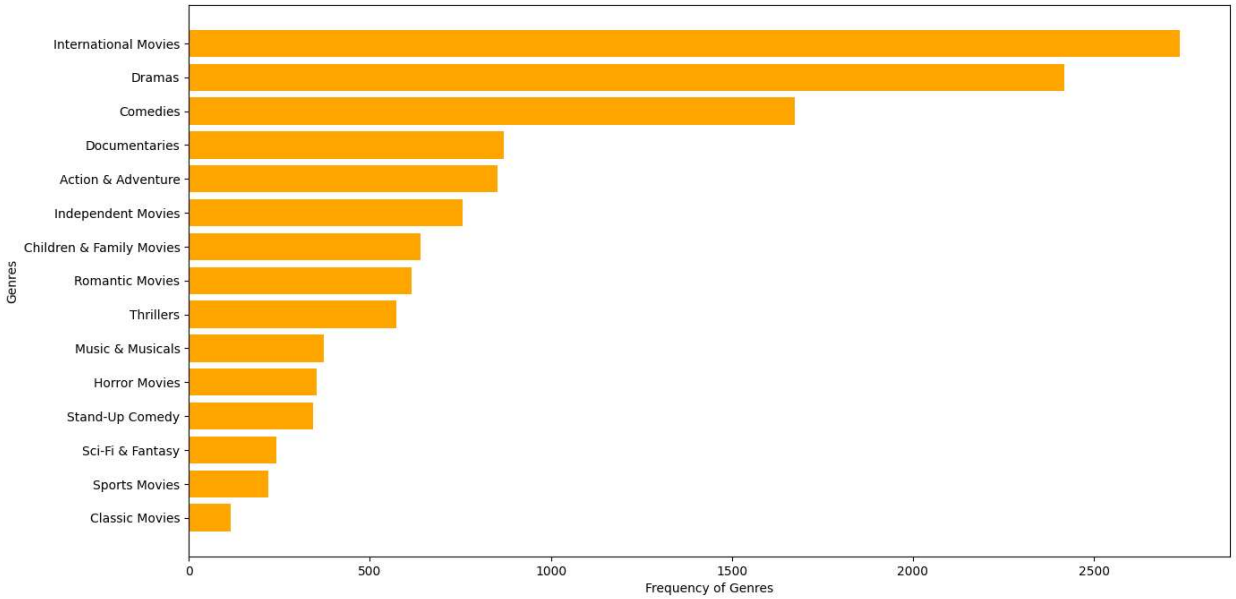
	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	rating
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA
5	Blood & Water	Khosi Ngema	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA

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



Most frequent shows are International drama, comedy and TV drama

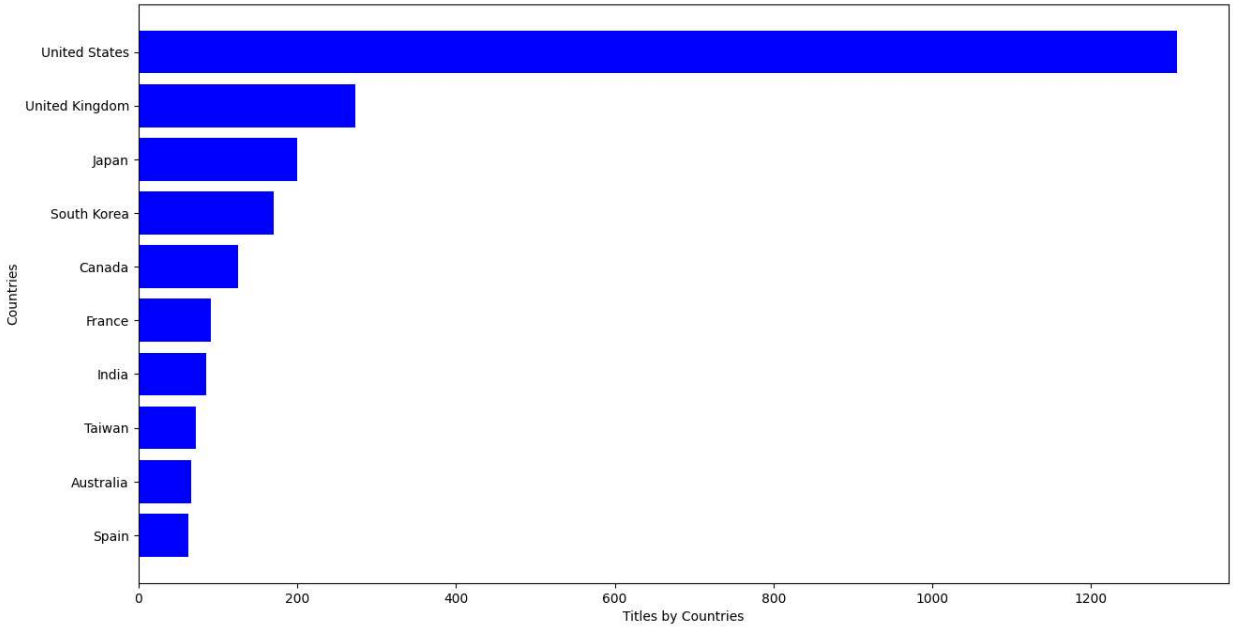
In [ ]:
```
df_genre=df_movies.groupby(['Genres']).agg({"title":"nunique"}).reset_index().sort_val
plt.figure(figsize=(15,8))
plt.barh(df_genre[::-1]['Genres'], df_genre[::-1]['title'],color=['orange'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```



Most frequent movies are International drama, comedy and TV drama

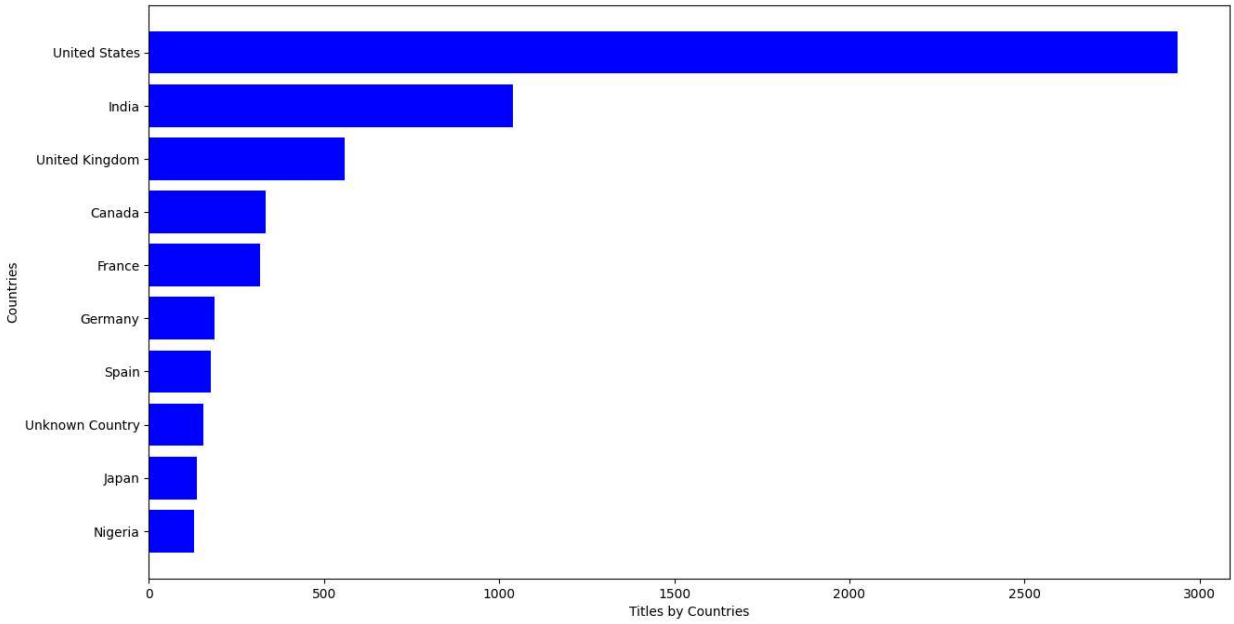
International, Drama and COmedies are most popular genres in both movies and TV shows

In [ ]:
```
df_country=df_shows.groupby(['country']).agg({"title":"nunique"}).reset_index().sort_v
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()
```



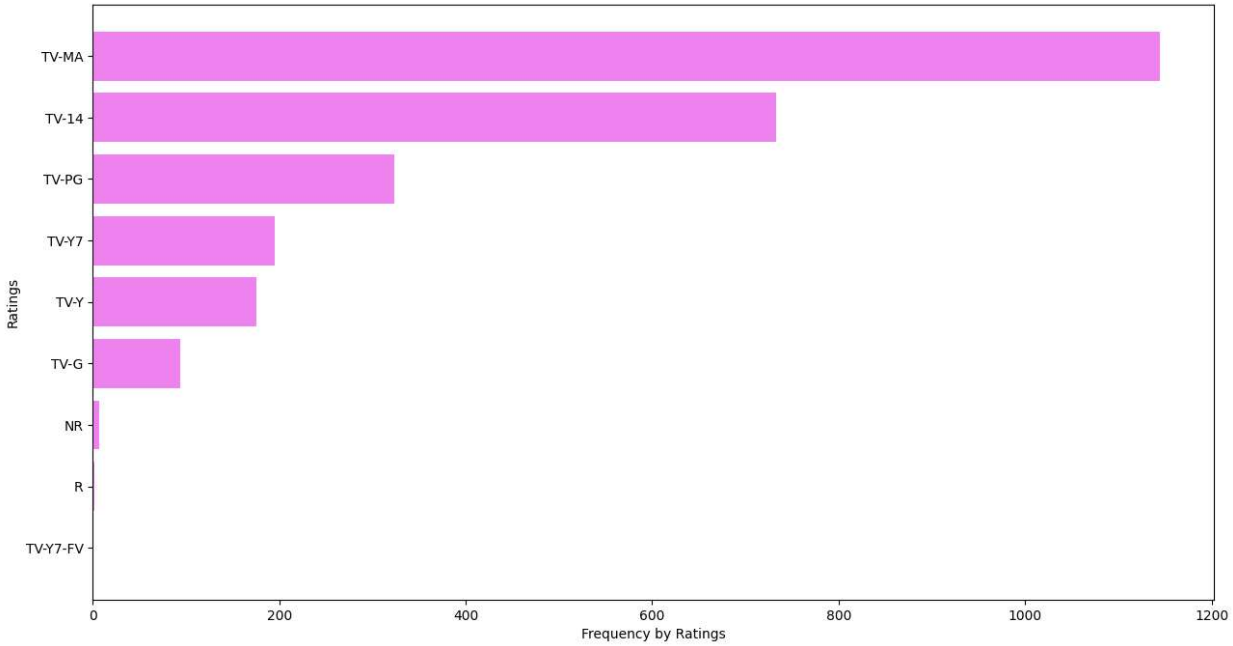
Most shows are from USA, UK and japan

In [ ]:
```
df_country=df_movies.groupby(['country']).agg({"title":"nunique"}).reset_index().sort_
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()
```



USA, India and UK are most movie makers in netflix

In [ ]:
```
df_rating=df_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_val
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()
```



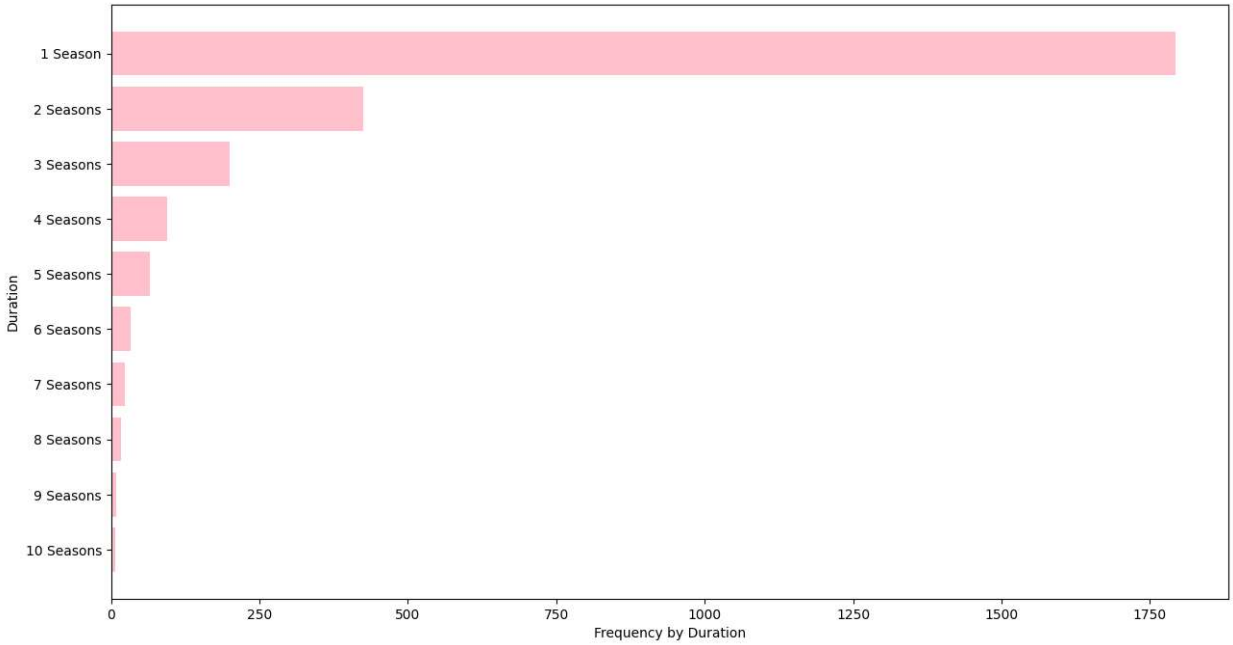
In [ ]:
```
df_rating=df_movies.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_va
plt.figure(figsize=(15,8))
plt.barh(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['violet'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```

In both Movies and TV shows TV-MA and TV-14 are top content

In movies R rated are next highest position

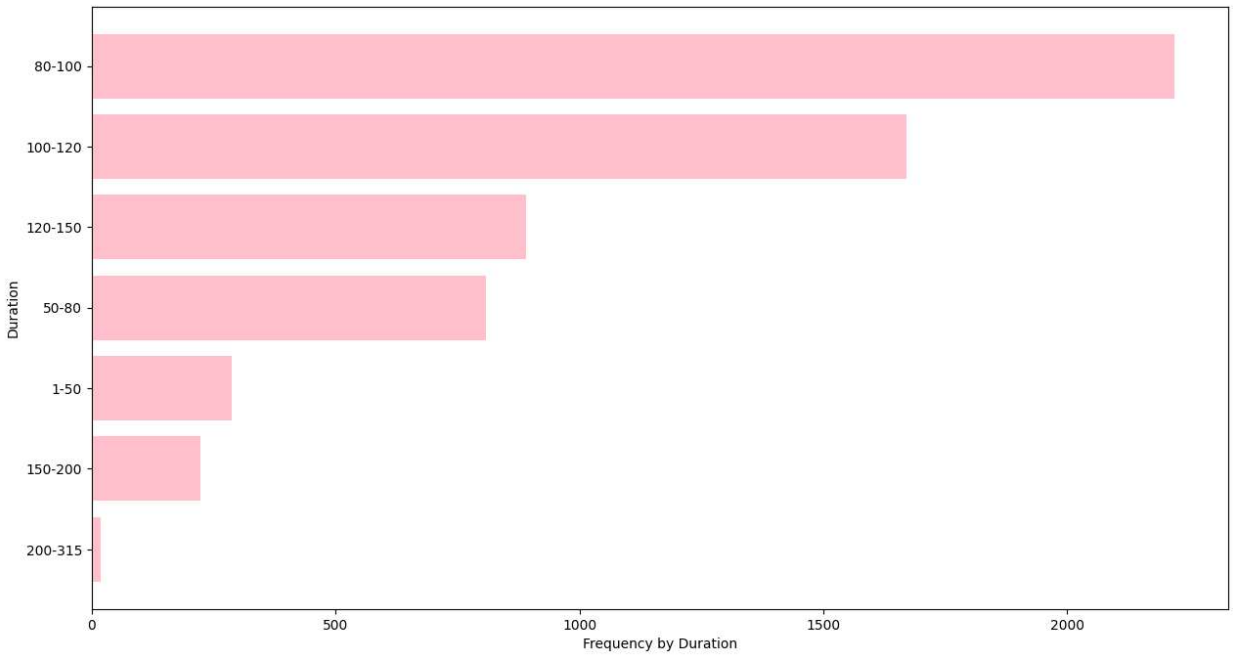
IN TV Shows TV-PG has next ranking

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



Most TV shows have 1,2 3 seasons in netflix

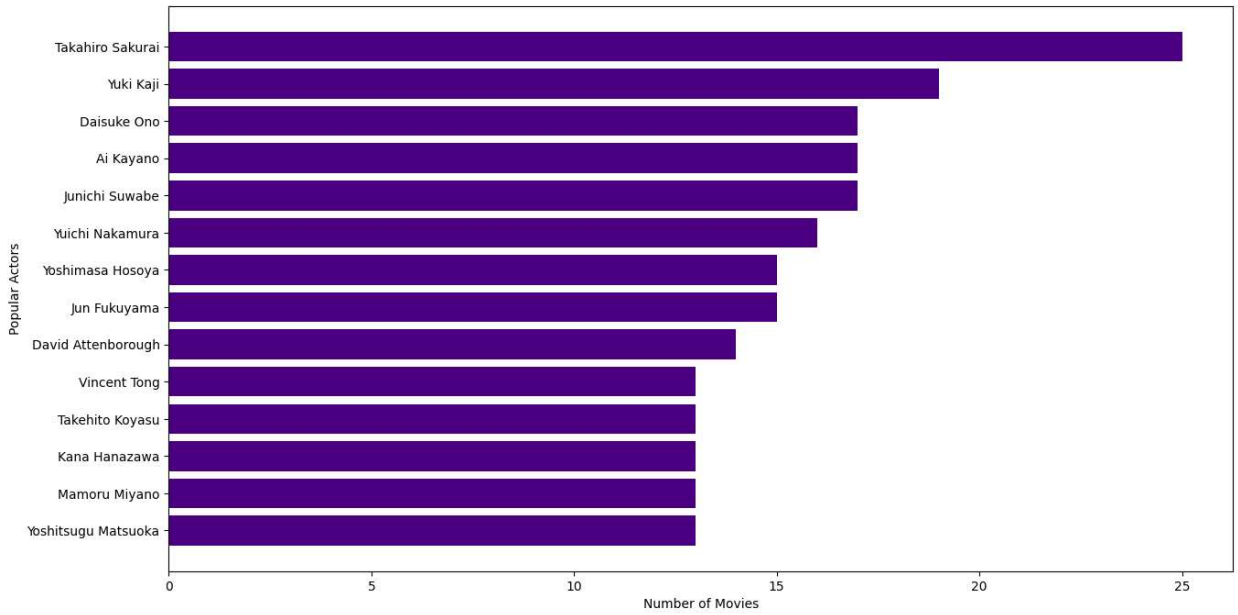
In [ ]:
```
df_duration=df_movies.groupby(['duration']).agg({"title":"nunique"}).reset_index().sor
plt.figure(figsize=(15,8))
plt.barh(df_duration[::-1]['duration'], df_duration[::-1]['title'],color=['pink'])
plt.xlabel('Frequency by Duration')
plt.ylabel('Duration')
plt.show()
```



Most movies have duration of 80-100 mins and 100-120 mins

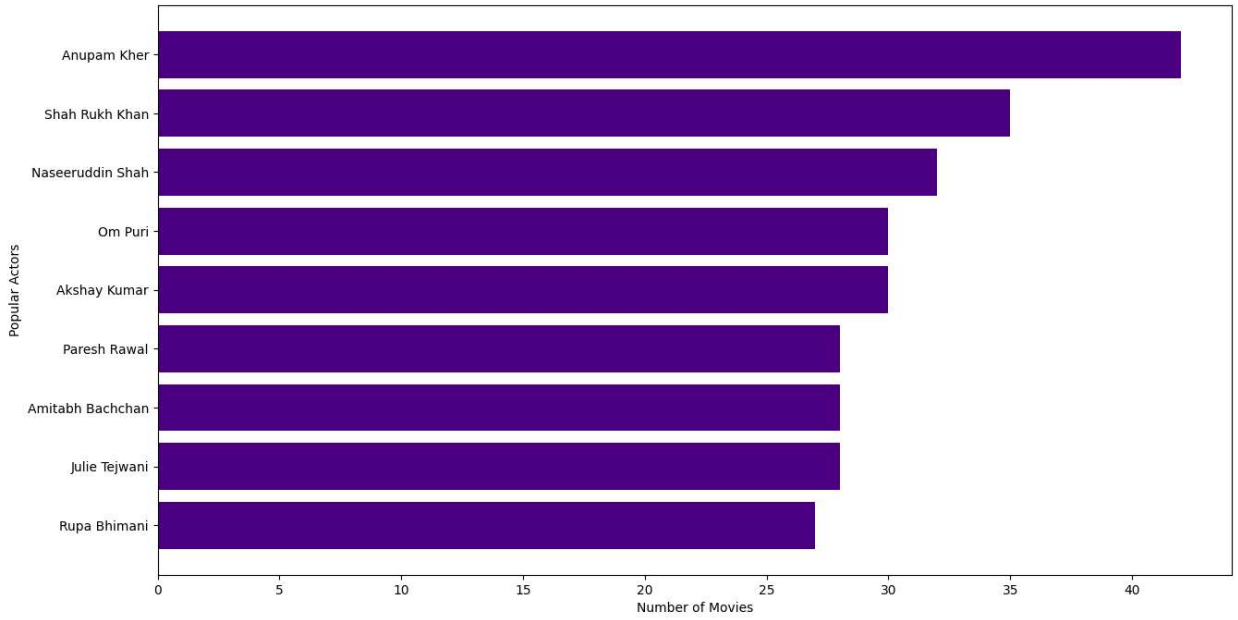
In [ ]:
```
df_actors=df_shows.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_val
df_actors=df_actors[df_actors['Actors']!='unknown Actor']
plt.figure(figsize=(15,8))
plt.barh(df_actors[::-1]['Actors'], df_actors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
```

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



N TV shows tha actors named as Takahiro sakurai, Yuki kaji and Daisuke Ono are the most frequent actors in netflix content

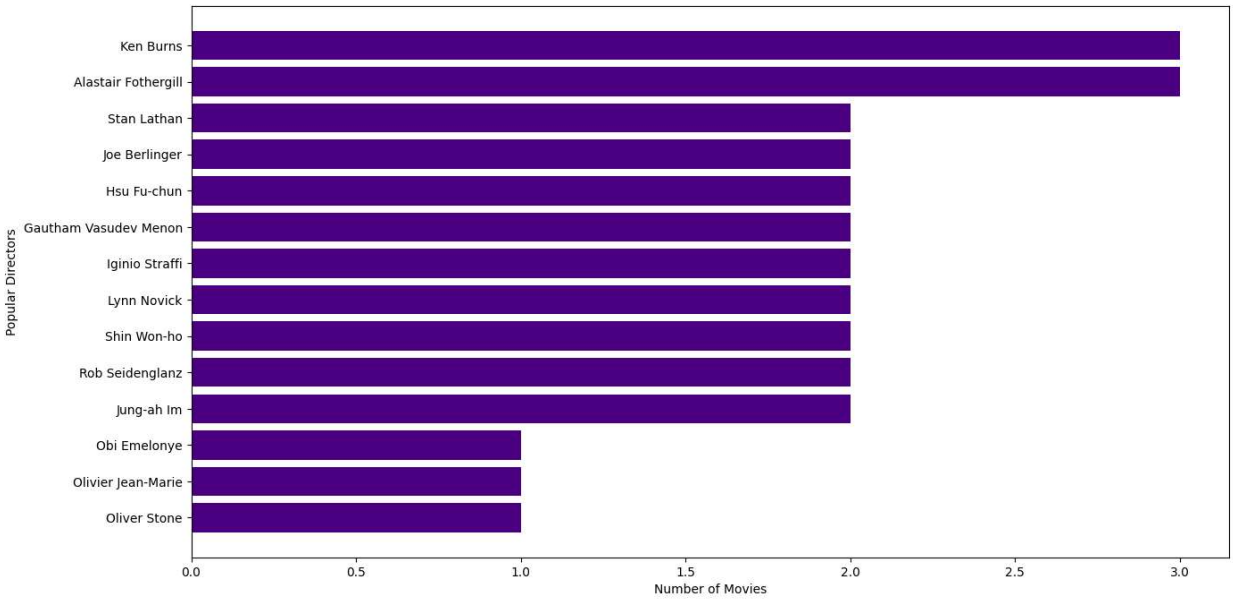
In [ ]:
```
df_actors=df_movies.groupby(['Actors']).agg({"title":"nunique"}).reset_index().sort_va
df_actors=df_actors[df_actors['Actors']!='unknown Actor']
plt.figure(figsize=(15,8))
plt.barh(df_actors[::-1]['Actors'], df_actors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular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



Anupam Kher, Sharukhan and Naseeruddin Shah are most frequent actors in movies of netflix

In [ ]:
```
df_directors=df_shows.groupby(['Director']).agg({"title":"nunique"}).reset_index().sor
df_directors=df_directors[df_directors['Director']!='Unknown Director']
```

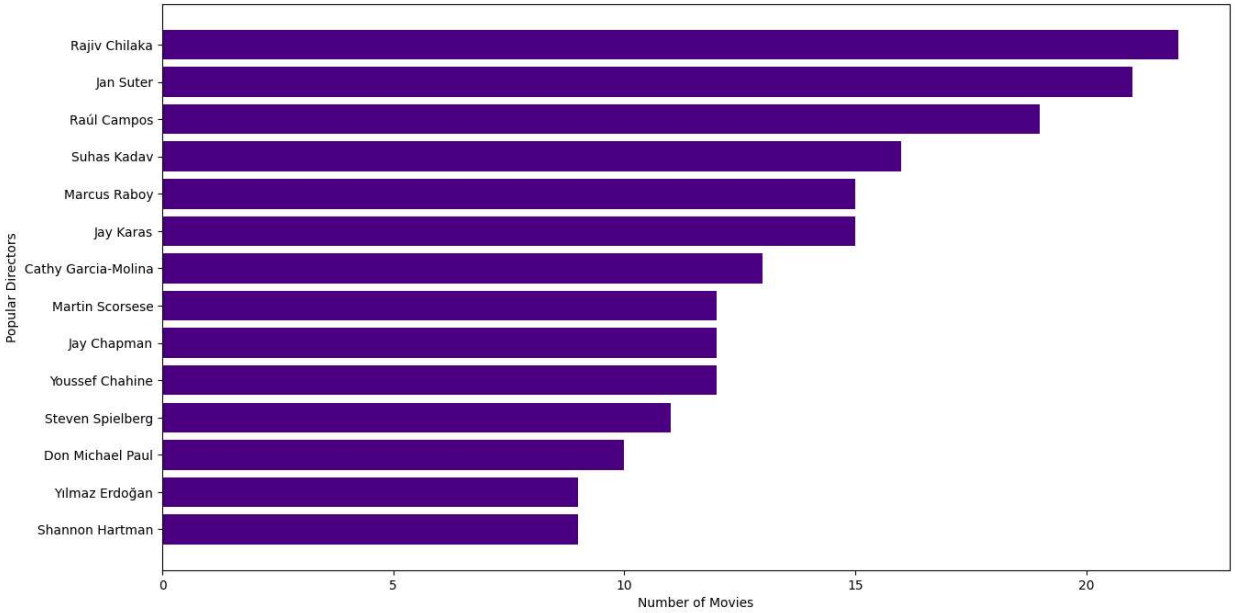
```
plt.figure(figsize=(15,8))
plt.barh(df_directors[::-1]['Director'], df_directors[::-1]['title'],color=['indigo'])
plt.xlabel('Number of Movies')
plt.ylabel('Popular Directors')
plt.show()
```



Top 10 directors in TV shows of netflix are

1. KenBurns
2. Alastair Fothergill
3. Stan Lathan
4. Joe Berlinger
5. Hsu Fu chun
6. Gautham Vasudev menon
7. Ignio straffi
8. Lynn Novic
9. Shin Won-ho
10. Rob Seidenglanz

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

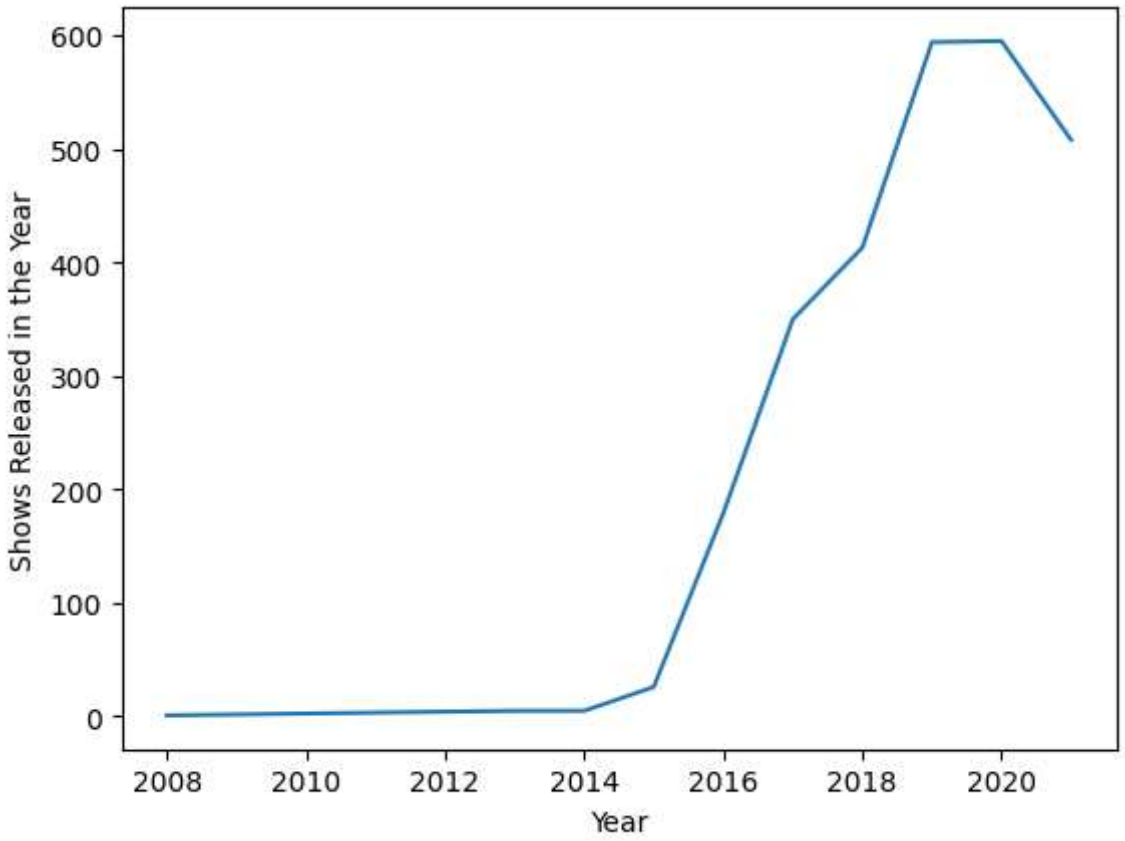


are most frequent directors in movies of netflix content

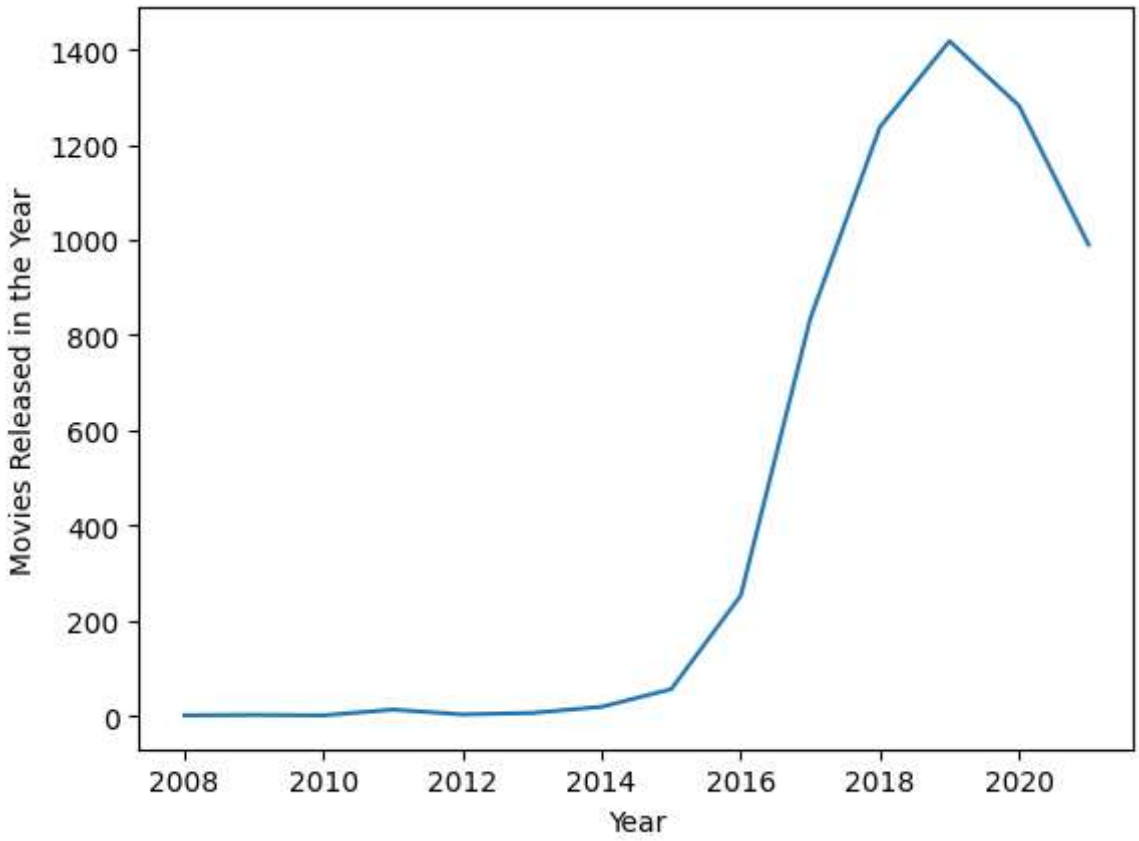
Top 10 directors in the movies of netflix are

1. Rajav Chilaka
2. Jan sutar
3. Raul Campos
4. Suhas Kadav
5. Marcus Raboy
6. Jay karas
7. Cathy Garcia-Molina
8. Martin Scorese
9. Jay chapman
10. Yousseff Chahine

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

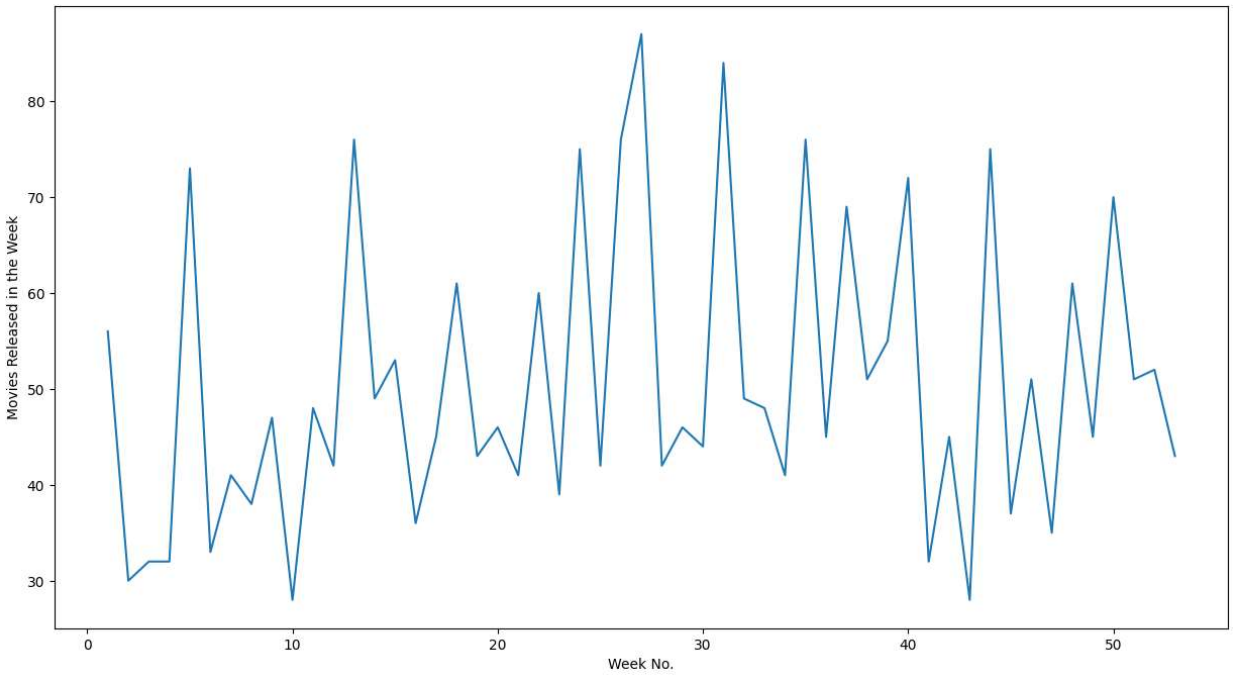


In [ ]:
```
df_year=df_movies.groupby(['year_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_year, x='year_added', y='title')
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



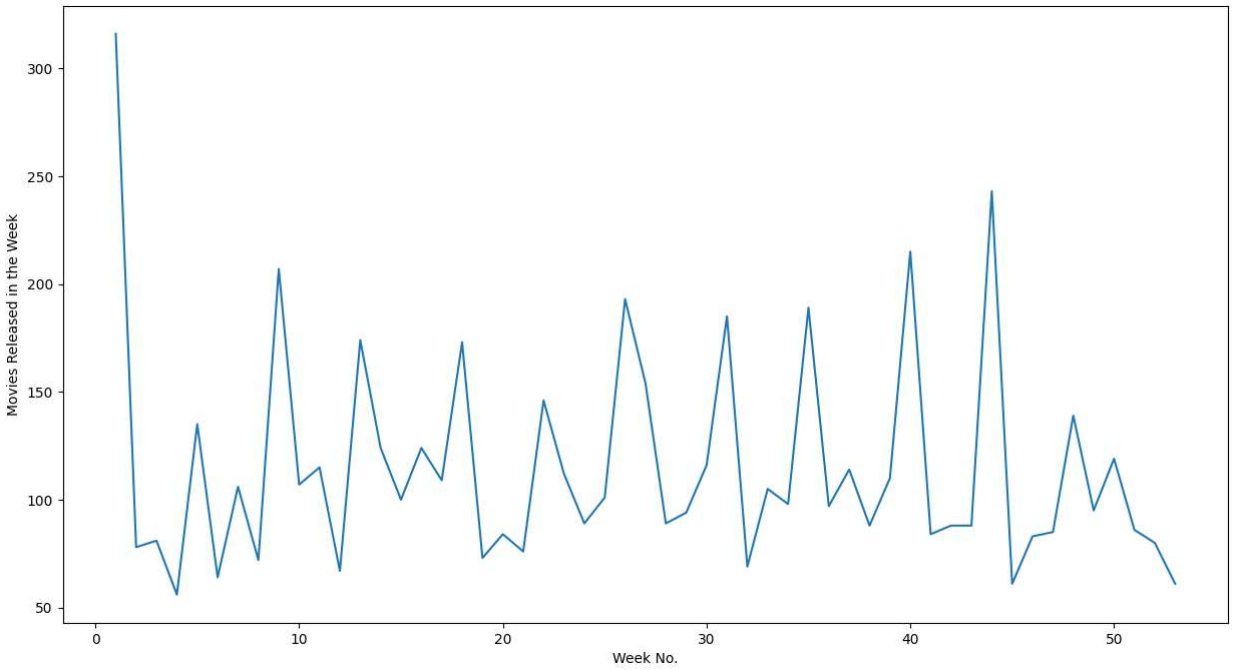
Most movues and TV shows are added from the year 2016

In [ ]:
```
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()
```



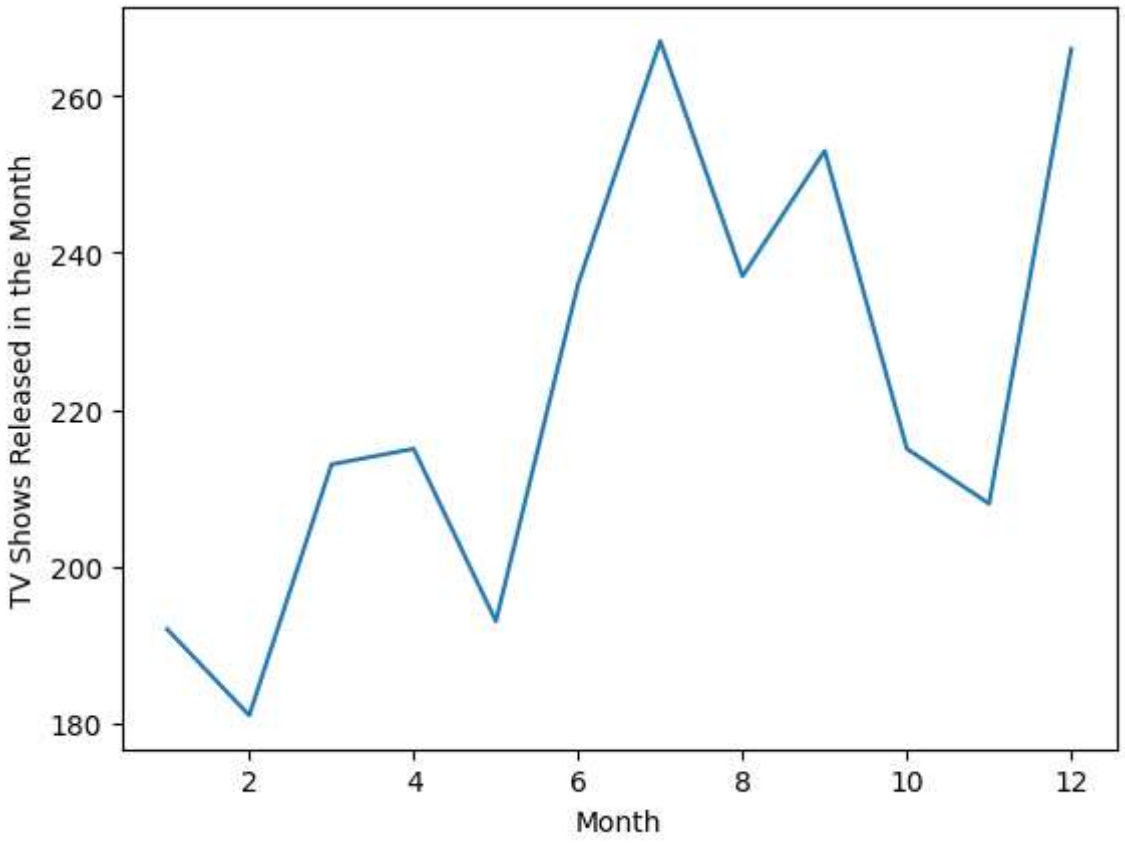
Most TV shows are added in the begining, mid and end of the year

In [ ]:
```
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()
```



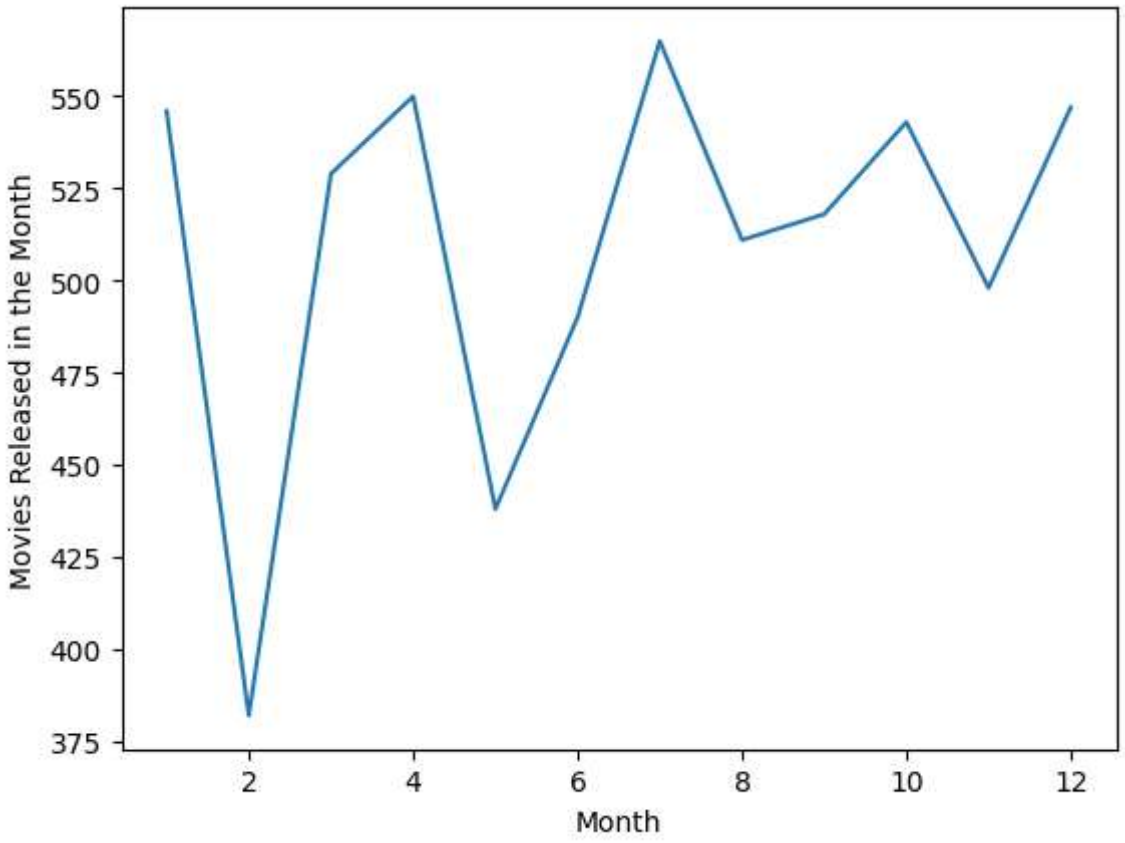
Most movies were added in the beginning and end of the year

In [ ]:
```
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()
```

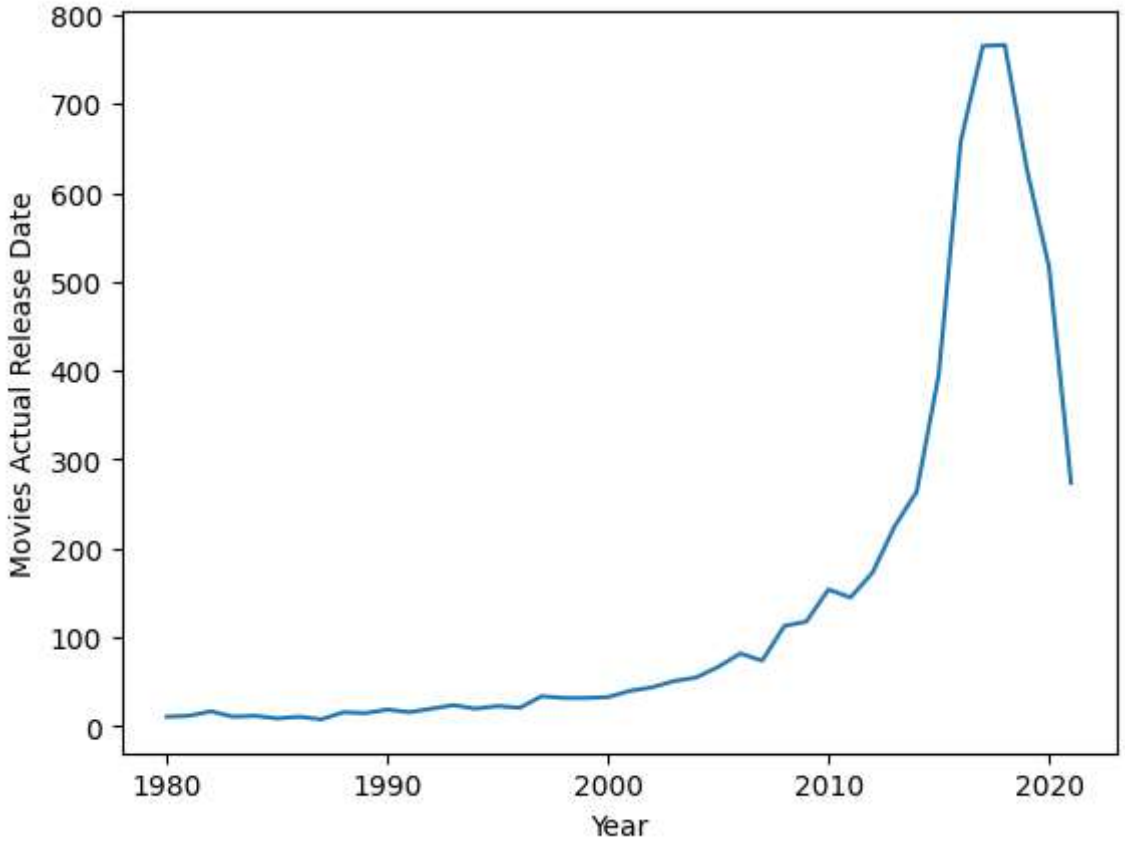


Most TV shows are added by the mid and end of the year

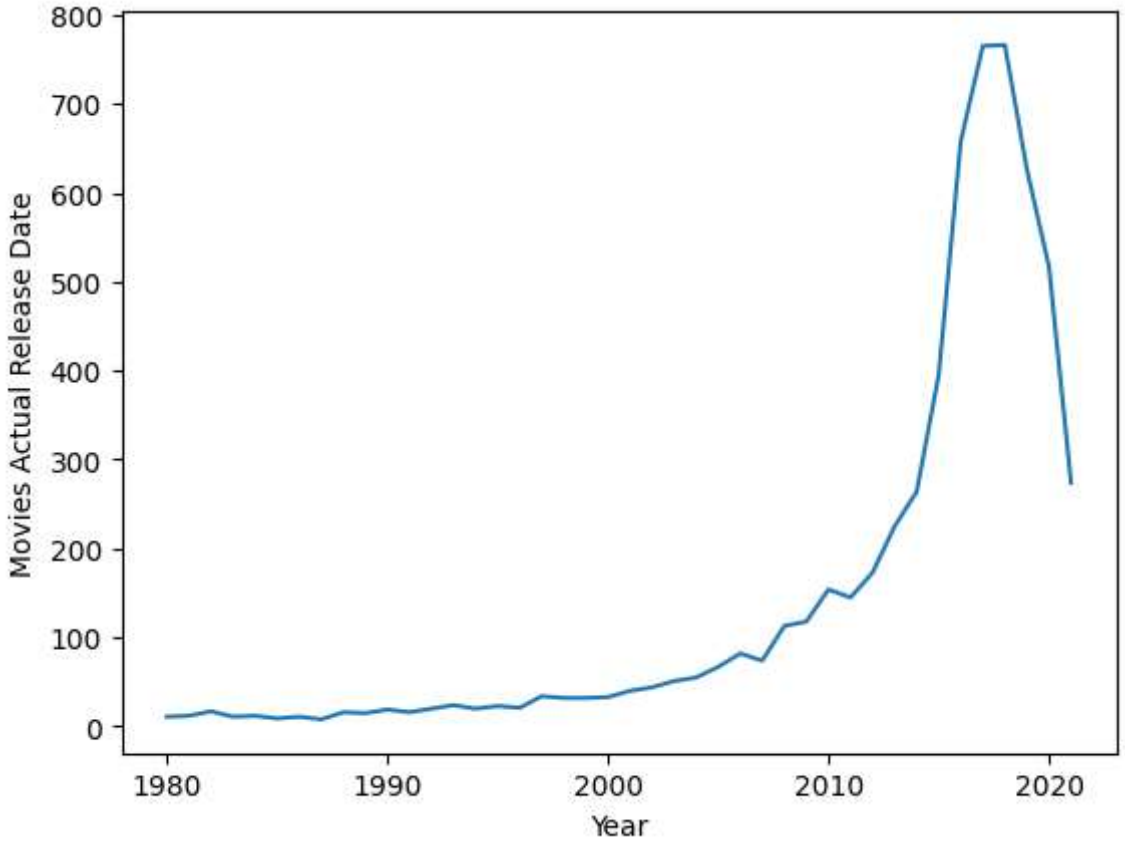
In [ ]:
```
df_month=df_movies.groupby(['month_added']).agg({"title":"nunique"}).reset_index()
sns.lineplot(data=df_month, x='month_added', y='title')
plt.ylabel("Movies Released in the Month")
plt.xlabel("Month")
plt.show()
```



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



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



Most movies and Tv shows are released from the year 2010

The dirrenece between the content released year are stated here

In [ ]:
```
yedf1['diff_year']=df1['year_added']-df1['release_year']
df1.head()
```

Out[ ]:

	title	Actors	Director	Genres	country	show_id	type	date_added	release_year	ra
0	Dick Johnson Is Dead	unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	P
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	

Most of the content has been released between 0 to 10 years

In [ ]:
```
df_movies=df1[df1['type']=='Movie']
df_shows=df1[df1['type']=='TV Show']
```

In [ ]:
```
import seaborn as sns
sns.distplot(df_movies['diff_year'],hist=True,kde=True,bins=int(32),color='blue',
hist_kws={'edgecolor' : 'darkblue'}, kde_kws={'linewidth': 2})
plt.show()
```

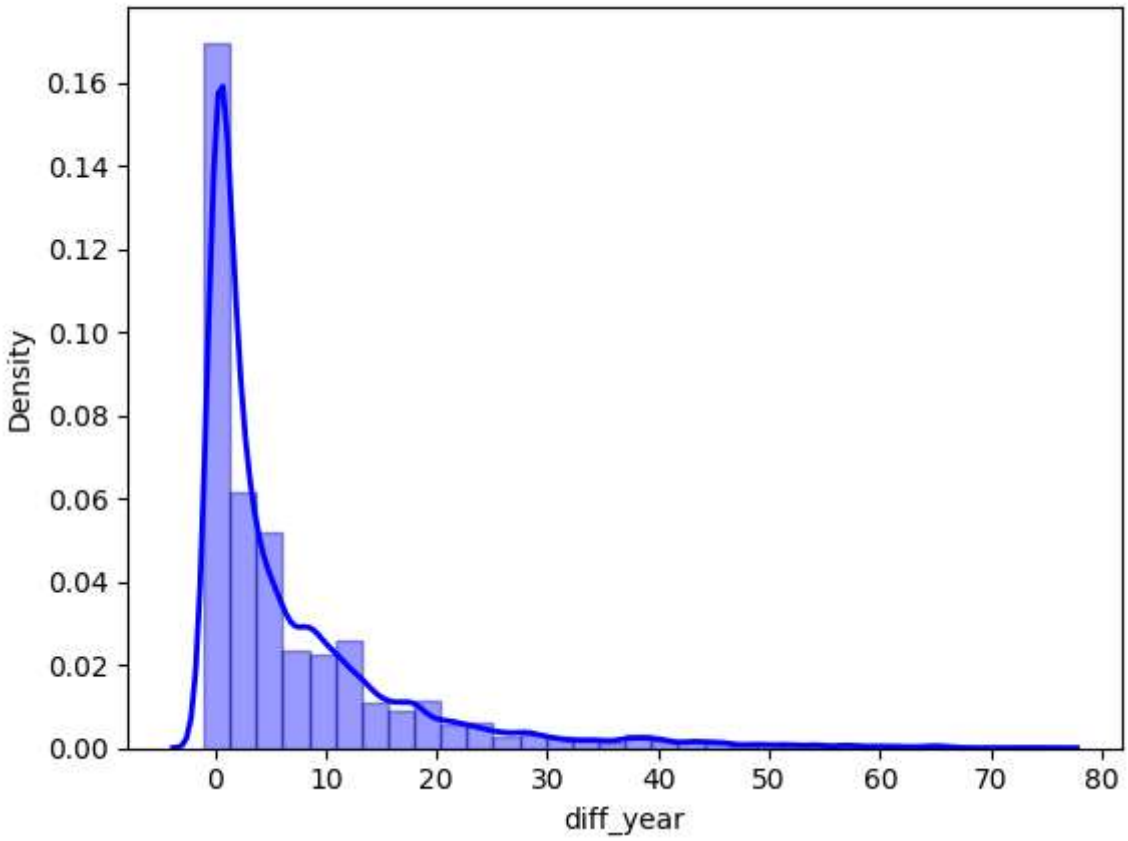
```
<ipython-input-104-8c34fd433f07>:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df_movies['diff_year'],hist=True,kde=True,bins=int(32),color='blue',
```



Most movies are added into netflix by 5 years after the release date

In [ ]:
```
import seaborn as sns
sns.distplot(df_shows['diff_year'],hist=True,kde=True,bins=int(32),color='blue',
hist_kws={'edgecolor' : 'darkblue'}, kde_kws={'linewidth': 2})
plt.show()
```

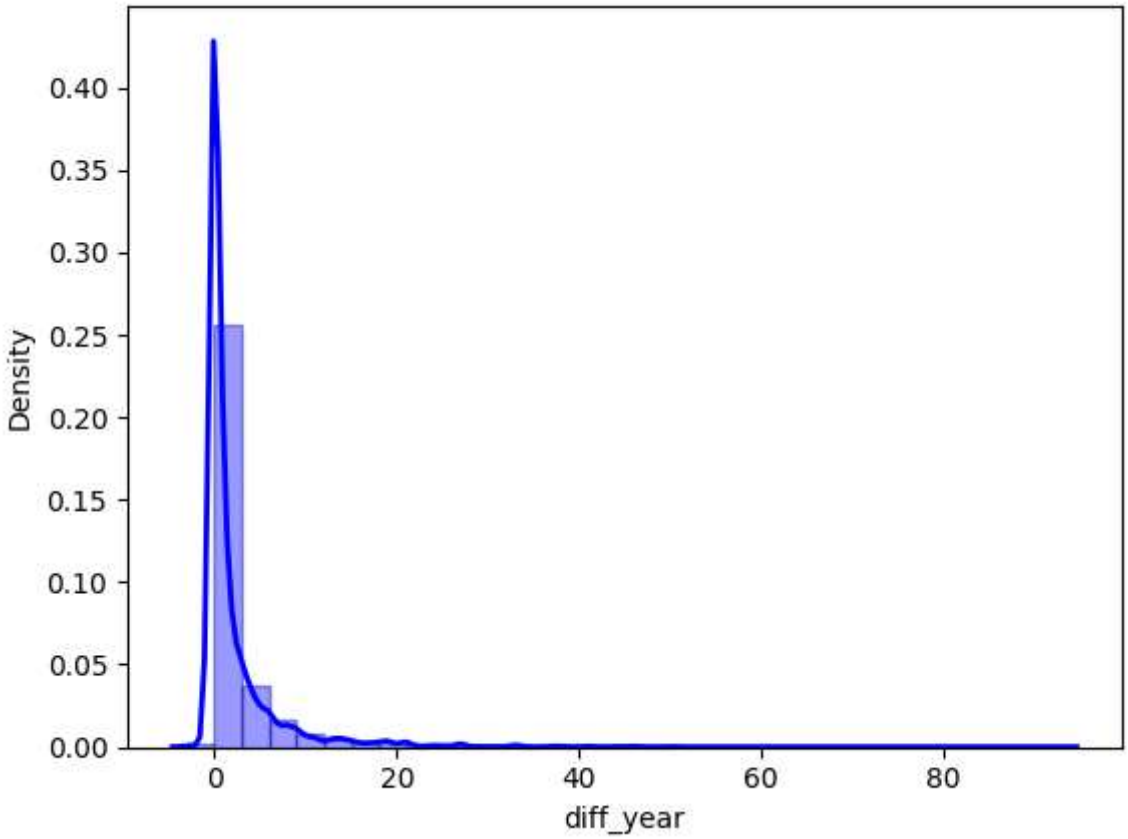
```
<ipython-input-105-7719ba5b6293>:2: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df_shows['diff_year'],hist=True,kde=True,bins=int(32),color='blue',
```



Most TV shows are added into netflix by 1 year after the release year

In [ ]:
```
from wordcloud import WordCloud
from wordcloud import ImageColorGenerator
from wordcloud import STOPWORDS
```

In [111…
```
text = " ".join(i for i in df1.Genres)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)
plt.figure( figsize=(10,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

In [112…
```
text = " ".join(i for i in df_movies.Genres)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)
plt.figure( figsize=(10,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Most watched Genres are International, Drama and comedy

In [113…
```
text = " ".join(i for i in df_shows.Genres)
stopwords = set(STOPWORDS)
wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)
plt.figure( figsize=(10,10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Most content in Tv shows are Internationa, Drama and Comedy

Recommendations

1. The most popular Genres are International, Drama and comedy in both TV shows and Movies
2. Add the movies and Tv shows into the netflix by 1 year after relase date.
3. Target audience are watching TV-MA and TV-14 movies and Tv shows so we can add more of these rating to the netflix.
4. People tend to watch 80 - 120 mins duration of the movies, It is recommended to add these durations
5. People tend to watch TV shows having 1 to 3 seasons, so add seasons with 1 to 3 duration
6. Add movies in the beginning and end of the year
7. Add Tv shows in the middle of the year
8. While adding the movies or TV shows consider the Popular directors and actors in the cast

In [113…

In [113…

In [ ]:

In [ ]:

In [ ]: