In [1]:

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

In [2]:

```
!gdown https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv
```

```
Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv (https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv)
To: C:\Users\anusha\Desktop\Numpy\netflix.csv

  0%|          | 0.00/3.40M [00:00<?, ?B/s]
 15%|#5        | 524k/3.40M [00:00<00:00, 3.05MB/s]
 31%|###       | 1.05M/3.40M [00:00<00:00, 2.80MB/s]
 46%|####6     | 1.57M/3.40M [00:00<00:00, 2.90MB/s]
 62%|######1   | 2.10M/3.40M [00:00<00:00, 2.96MB/s]
 77%|#######7  | 2.62M/3.40M [00:00<00:00, 2.99MB/s]
 93%|#########2| 3.15M/3.40M [00:01<00:00, 3.00MB/s]
100%|##########| 3.40M/3.40M [00:01<00:00, 2.99MB/s]
```

In [2]:

```
df = pd.read_csv("netflix.csv")
```

In [3]:

```
df
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

8807 rows × 12 columns

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

In [5]:

```
#converting date_added to datetime
df['date_added'] = pd.to_datetime(df['date_added'])
```

In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype         
---  ------        --------------  -----         
 0   show_id       8807 non-null   object        
 1   type          8807 non-null   object        
 2   title         8807 non-null   object        
 3   director      6173 non-null   object        
 4   cast          7982 non-null   object        
 5   country       7976 non-null   object        
 6   date_added    8797 non-null   datetime64[ns]
 7   release_year  8807 non-null   int64         
 8   rating        8803 non-null   object        
 9   duration      8804 non-null   object        
 10  listed_in     8807 non-null   object        
 11  description   8807 non-null   object        
dtypes: datetime64[ns](1), int64(1), object(10)
memory usage: 825.8+ KB
```

In [7]:

```
df
```

Out[7]:

	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	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

8807 rows × 12 columns

In [9]:

```
#1 defining the problem statement and Analysing basic metrics
```

In [ ]:

```
The netflix data is been provided where I need to analyse the data. Process it using Data exploration and visulisation technics
and generate valiable insights for the company which are data-driven . This insights would help Netflix to make correct decisions
that would help them to grow against they competitors. I need to concentatre more on which type of content has to been more added.
to the platform and also analyse the Netflix performance pre covid and post covid.
    The country anlyzation has to done to analyze individual countries content and performance . The Basic metrics provided are the
    data in most of the catagorical variable is nested. The date and time and release_date both are in object form . Duration has two
    different types measures  one in minutes and seasons.The missing values has to be filled.
```

In [ ]:

```
2.#Observations on the shape of data, data types of all the attributes,
#conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary
```

In [8]:

```
df.shape # shape of data
```

Out[8]:

```
(8807, 12)
```

In [9]:

```
df.info() #data types
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype         
---  ------        --------------  -----         
 0   show_id       8807 non-null   object        
 1   type          8807 non-null   object        
 2   title         8807 non-null   object        
 3   director      6173 non-null   object        
 4   cast          7982 non-null   object        
 5   country       7976 non-null   object        
 6   date_added    8797 non-null   datetime64[ns]
 7   release_year  8807 non-null   int64         
 8   rating        8803 non-null   object        
 9   duration      8804 non-null   object        
 10  listed_in     8807 non-null   object        
 11  description   8807 non-null   object        
dtypes: datetime64[ns](1), int64(1), object(10)
memory usage: 825.8+ KB
```

In [10]:

```
df.isna().sum() #missing values sum
```

Out[10]:

```
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 [11]:

```
df.describe() #statistical summary
```

Out[11]:

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

In [ ]:

```
#3 QUestion
#value counts and unique attributes
```

In [12]:

```
df.columns
```

Out[12]:

```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

In [13]:

```
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 : 1714
release_year : 74
rating : 17
duration : 220
listed_in : 514
description : 8775
```

In [14]:

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

Out[14]:

```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

In [12]:

```
df['title'].value_counts()
```

Out[12]:

```
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
```

In [18]:

```
df['director'].value_counts()
```

Out[18]:

```
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
```

In [20]:

```
df['cast'].value_counts()
```

Out[20]:

```
David Attenborough
19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil
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 Letscher, Alyssa Diaz
1
Nick Lachey, Vanessa Lachey
1
Takeru Sato, Kasumi Arimura, Haru, Kentaro Sakaguchi, Takayuki Yamada, Kendo Kobayashi, Ken Yasuda, Arata Furuta, Suzuki Ma
tsuo, Koichi Yamadera, Arata Iura, Chikako Kaku, Kotaro Yoshida     1
Toyin Abraham, Sambasa Nzeribe, Chioma Chukwuka Akpotha, Chioma Omeruah, Chiwetalu Agu, Dele Odule, Femi Adebayo, Bayray Mc
Nwizu, Biodun Stephen                                               1
Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malkeet Rauni, Anita Shabdish, Chittaranja
n Tripathy                                                          1
Name: cast, Length: 7692, dtype: int64
```

In [22]:

```
df['country'].value_counts()
```

Out[22]:

```
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
```

In [24]:

```
df['date_added'].value_counts()
```

Out[24]:

```
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
```

In [26]:

```
df['release_year'].value_counts()
```

Out[26]:

```
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
```

In [27]:

```
df['rating'].value_counts()
```

Out[27]:

```
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
```

In [28]:

```
df['duration'].value_counts()
```

Out[28]:

```
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
```

In [29]:

```
df['listed_in'].value_counts()
```

Out[29]:

```
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
```

In [30]:

```
df['description'].value_counts()
```

Out[30]:

```
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 ye
ars 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 tra
ces their cryptic connection.     3
Secrets bubble to the surface after a sensual encounter and an unforeseen crime entangle two friends and a woman caught bet
ween 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 A
ccra.                             1
When his partner in crime goes missing, a small-time crook’s life is transformed as he dedicates himself to raising the dau
ghter his friend left behind.     1
During 1962’s Cuban missile crisis, a troubled math genius finds himself drafted to play 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 w
ith it dies soon afterward.       1
A scrappy but poor boy worms his way into a tycoon's dysfunctional family, while facing his fear of music and the truth abo
ut his past.                      1
Name: description, Length: 8775, dtype: int64
```

In [ ]:

```
#4 unnesting the nested cloumns
```

In [15]:

```
#unnesting the country column
small_cast = df[['title','cast']]
small_cast['cast']=small_cast['cast'].apply(lambda x :str(x).split(','))
small_cast = small_cast.explode('cast')
small_cast.head()
```

```
C:\Users\anusha\AppData\Local\Temp\ipykernel_6112\1769347955.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vie
w-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  small_cast['cast']=small_cast['cast'].apply(lambda x :str(x).split(','))
```

Out[15]:

	title	cast
0	Dick Johnson Is Dead	nan
1	Blood & Water	Ama Qamata
1	Blood & Water	Khosi Ngema
1	Blood & Water	Gail Mabalane
1	Blood & Water	Thabang Molaba

In [16]:

```
#unnesting the country column
small_director = df[['title','director']]
small_director['director']=small_director['director'].apply(lambda x :str(x).split(','))
small_director = small_director.explode('director')
small_director.head()
```

```
C:\Users\anusha\AppData\Local\Temp\ipykernel_6112\2005900849.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vie
w-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  small_director['director']=small_director['director'].apply(lambda x :str(x).split(','))
```

Out[16]:

	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 [17]:

```
#unnesting the country column
small_country = df[['title','country']]
small_country['country']=small_country['country'].apply(lambda x :str(x).split(','))
small_country = small_country.explode('country')
small_country.head()
```

```
C:\Users\anusha\AppData\Local\Temp\ipykernel_6112\1352627097.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  small_country['country']=small_country['country'].apply(lambda x :str(x).split(','))
```

Out[17]:

	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 [18]:

```
#unnesting the listed_in column
small_listed_in = df[['title','listed_in']]
small_listed_in['listed_in']=small_listed_in['listed_in'].apply(lambda x :str(x).split(','))
small_listed_in = small_listed_in.explode('listed_in')
small_listed_in.head()
```

```
C:\Users\anusha\AppData\Local\Temp\ipykernel_6112\2629472403.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  small_listed_in['listed_in']=small_listed_in['listed_in'].apply(lambda x :str(x).split(','))
```

Out[18]:

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

In [ ]:

```
#merging
```

In [19]:

```
df5=small_director.merge(small_cast,on=['title'],how='inner')
df6=df5.merge(small_listed_in,on=['title'],how='inner')
df_new=df6.merge(small_country,on=['title'],how="inner")
df_new
```

Out[19]:

	title	director	cast	listed_in	country
0	Dick Johnson Is Dead	Kirsten Johnson	nan	Documentaries	United States
1	Blood & Water	nan	Ama Qamata	International TV Shows	South Africa
2	Blood & Water	nan	Ama Qamata	TV Dramas	South Africa
3	Blood & Water	nan	Ama Qamata	TV Mysteries	South Africa
4	Blood & Water	nan	Khosi Ngema	International TV Shows	South Africa
...	...	...	...	...	...
202060	Zubaan	Mozez Singh	Anita Shabdish	International Movies	India
202061	Zubaan	Mozez Singh	Anita Shabdish	Music & Musicals	India
202062	Zubaan	Mozez Singh	Chittaranjan Tripathy	Dramas	India
202063	Zubaan	Mozez Singh	Chittaranjan Tripathy	International Movies	India
202064	Zubaan	Mozez Singh	Chittaranjan Tripathy	Music & Musicals	India

202065 rows × 5 columns

In [20]:

```
df_new.rename(columns={'cast':'Actors','listed_in':'Genre'},inplace=True)
```

In [21]:

```
df_new
```

Out[21]:

	title	director	Actors	Genre	country
0	Dick Johnson Is Dead	Kirsten Johnson	nan	Documentaries	United States
1	Blood & Water	nan	Ama Qamata	International TV Shows	South Africa
2	Blood & Water	nan	Ama Qamata	TV Dramas	South Africa
3	Blood & Water	nan	Ama Qamata	TV Mysteries	South Africa
4	Blood & Water	nan	Khosi Ngema	International TV Shows	South Africa
...	...	...	...	...	...
202060	Zubaan	Mozez Singh	Anita Shabdish	International Movies	India
202061	Zubaan	Mozez Singh	Anita Shabdish	Music & Musicals	India
202062	Zubaan	Mozez Singh	Chittaranjan Tripathy	Dramas	India
202063	Zubaan	Mozez Singh	Chittaranjan Tripathy	International Movies	India
202064	Zubaan	Mozez Singh	Chittaranjan Tripathy	Music & Musicals	India

202065 rows × 5 columns

In [ ]:

```
#replacing the nan values in df_new
```

In [22]:

```
df_new['Actors'].replace(['nan'],['unknown Actor'],inplace = True)
df_new['director'].replace(['nan'],['unknown Director'],inplace=True)
df_new['country'].replace(['nan'],[np.nan],inplace = True)
df_new.head()
```

Out[22]:

	title	director	Actors	Genre	country
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa

In [ ]:

```
```

In [23]:

```
df.columns
```

Out[23]:

```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

In [24]:

```
df_final=df_new.merge(df[['show_id', 'type', 'title', 'date_added',
       'release_year', 'rating', 'duration']],on='title',how='left')
df_final.head()
```

Out[24]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90 min
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons

In [25]:

```
df_final.isna().sum() #we exploded the data so null values increased in country
```

Out[25]:

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

In [ ]:

```
```

In [24]:

```
#replacing the Null values in durations column with that of vlaues in rating column
```

In [26]:

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

In [27]:

```
df_final
```

Out[27]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90 min
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
...	...	...	...	...	...	...	...	...	...	...	...
202060	Zubaan	Mozez Singh	Anita Shabdish	International Movies	India	s8807	Movie	2019-03-02	2015	TV-14	111 min

In [28]:

```
#filling the null values in date_added by finding its mode using release date
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'].fillna(imp)
```

In [29]:

```
#filling the null values in country by finding its mode using director's column
for i in df_final[df_final['country'].isnull()]['director'].unique(): # all the places where corresponding country is missing for a direct
  if i in df_final[~df_final['country'].isnull()]['director'].unique():
    imp=df_final[df_final['director']==i]['country'].mode().values[0]
    df_final.loc[df_final['director']==i,'country']=df_final.loc[df_final['director']==i,'country'].fillna(imp)
```

In [30]:

```
df_final.isna().sum()
```

Out[30]:

```
title              0
director           0
Actors             0
Genre              0
country         4673
show_id            0
type               0
date_added         0
release_year       0
rating            67
duration           0
dtype: int64
```

In [33]:

```
#filling reamining the null values in country by finding its mode using Actor's column
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(imp)
#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[33]:

```
title           0
director        0
Actors          0
Genre           0
country         0
show_id         0
type            0
date_added      0
release_year    0
rating         67
duration        0
dtype: int64
```

In [34]:

```
##filling the null values in rating by finding
df_final.loc[df_final['rating'].str.contains('min', na=False),'rating']='NR'
df_final['rating'].fillna('NR',inplace=True)
```

In [35]:

```
df_final.isnull().sum() # final dataframe with no null values
```

Out[35]:

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

In [36]:

```
df_final
```

Out[36]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90 min
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
...	...	...	...	...	...	...	...	...	...	...	...
202060	Zubaan	Mozez Singh	Anita Shabdish	International Movies	India	s8807	Movie	2019-03-02	2015	TV-14	111 min
202061	Zubaan	Mozez Singh	Anita Shabdish	Music & Musicals	India	s8807	Movie	2019-03-02	2015	TV-14	111 min
202062	Zubaan	Mozez Singh	Chittaranjan Tripathy	Dramas	India	s8807	Movie	2019-03-02	2015	TV-14	111 min
202063	Zubaan	Mozez Singh	Chittaranjan Tripathy	International Movies	India	s8807	Movie	2019-03-02	2015	TV-14	111 min
202064	Zubaan	Mozez Singh	Chittaranjan Tripathy	Music & Musicals	India	s8807	Movie	2019-03-02	2015	TV-14	111 min

202065 rows × 11 columns

In [37]:

```
df_final['duration'].value_counts()
```

Out[37]:

```
1 Season     35035
2 Seasons     9559
3 Seasons     5084
94 min        4343
106 min       4040
             ...
3 min            4
5 min            3
11 min           2
8 min            2
9 min            2
Name: duration, Length: 220, dtype: int64
```

In [38]:

```
#removing mins from data
df_final['duration']=df_final['duration'].str.replace(" min","")
df_final.head()
```

Out[38]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons

In [39]:

```
df_final['duration'].unique()
```

Out[39]:

```
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 [40]:

```
df_final['duration_copy']=df_final['duration'].copy()
df_final1=df_final.copy()
```

In [41]:

```
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()
```

Out[41]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration	duration_copy
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90	90
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0

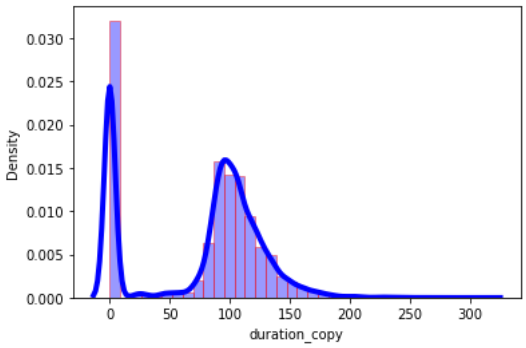
In [ ]:

```
# univariated displot for duration
```

In [42]:

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

```
C:\Users\anusha\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated functi
on and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with si
milar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



In [43]:

```
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()
```

Out[43]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration	duration_copy
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90	80-100
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1

In [44]:

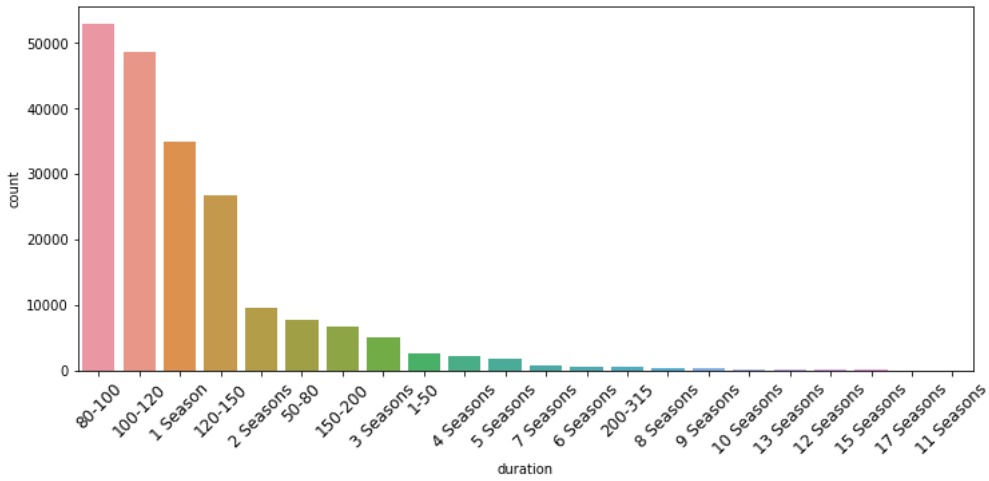
```
df_final1.loc[~df_final1['duration'].str.contains('Season'),'duration']=df_final1.loc[~df_final1['duration'].str.contains('Season'),'durat
df_final1.drop(['duration_copy'],axis=1,inplace=True)
df_final1.head()
```

Out[44]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons

In [45]:

```
plt.figure(figsize=(12,5))
sns.countplot(data=df_final1,
              x="duration",order=df_final1["duration"].value_counts().index)
plt.xticks(rotation=45,fontsize=12)
plt.show()
```



In [46]:

```
df_final1['title']=df_final1['title'].str.replace(r"\(.*\)","")
df_final1.head()
```

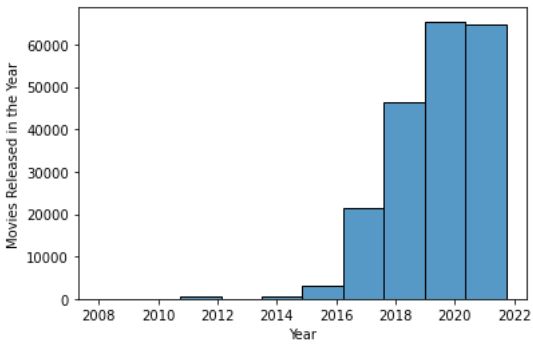
```
C:\Users\anusha\AppData\Local\Temp\ipykernel_6112\1805384656.py:1: FutureWarning: The default value of regex will change fr
om True to False in a future version.
  df_final1['title']=df_final1['title'].str.replace(r"\(.*\)","")
```

Out[46]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons

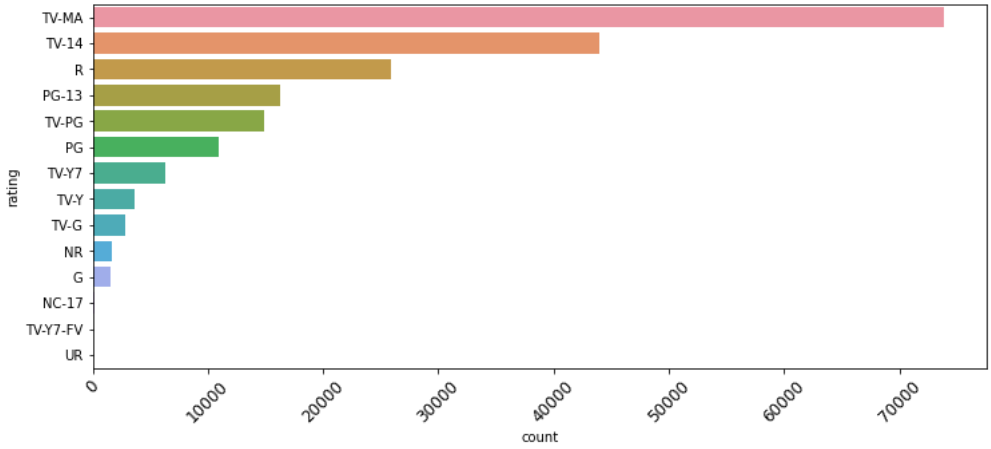
In [47]:

```
sns.histplot(df_final1['date_added'],bins=10)
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



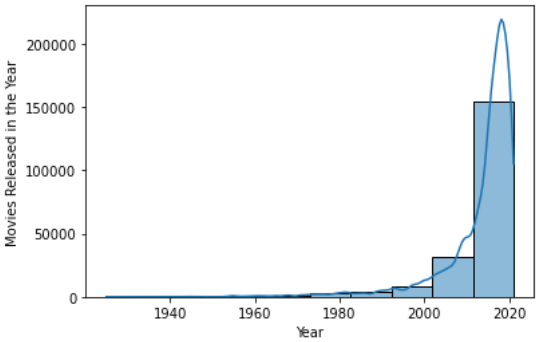
In [48]:

```
plt.figure(figsize=(12,5))
sns.countplot(y=df_final1["rating"],
              order=df_final1["rating"].value_counts().index)
plt.xticks(rotation=45,fontsize=12)
plt.show()
```



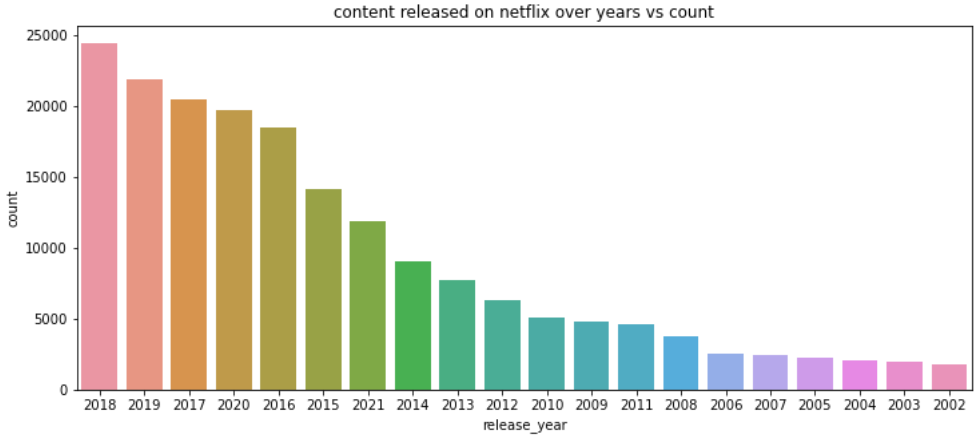
In [49]:

```
sns.histplot(df_final1['release_year'],bins=10,kde=True)
plt.ylabel("Movies Released in the Year")
plt.xlabel("Year")
plt.show()
```



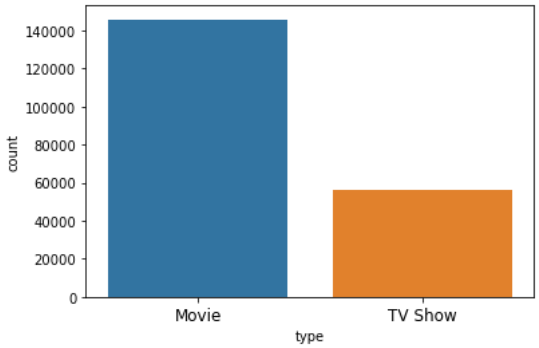
In [50]:

```
plt.figure(figsize=(12,5))
sns.countplot(data=df_final1,
              x="release_year",order=df_final1["release_year"].value_counts().index[0:20])
plt.title("content released on netflix over years vs count")
plt.show()
```



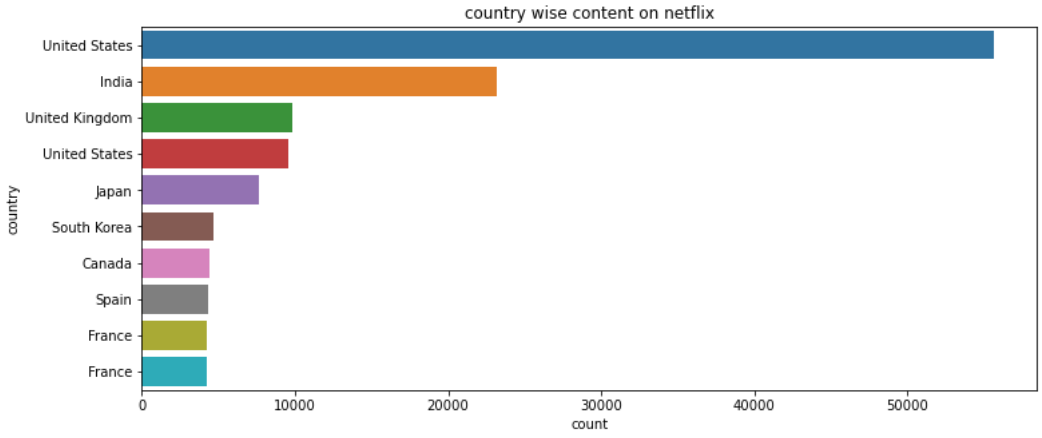
In [45]:

```
sns.countplot(data=df_final1,
              x="type",order=df_final1["type"].value_counts().index)
plt.xticks(fontsize=12)
plt.show()
```



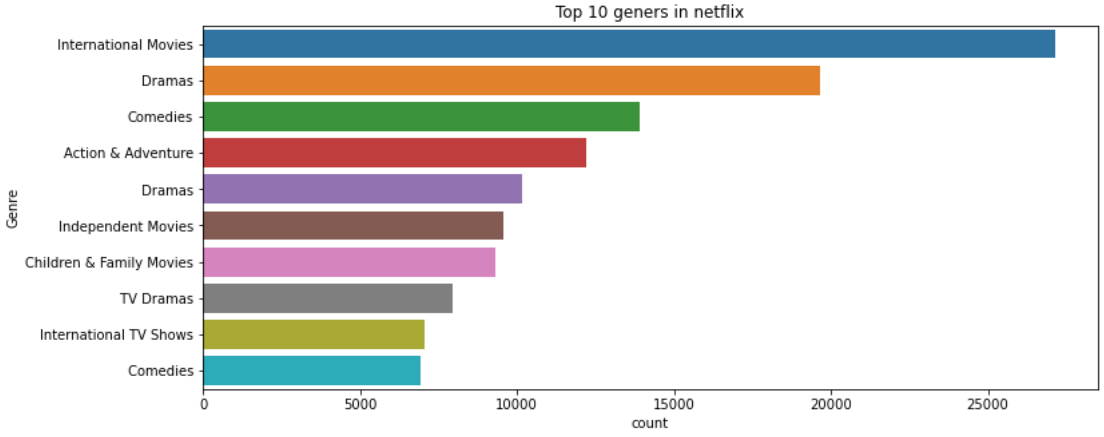
In [51]:

```
plt.figure(figsize=(12,5))
sns.countplot(data=df_final1,
              y="country",order=df_final1["country"].value_counts().index[0:10])
plt.title("country wise content on netflix")
plt.show()
```



In [52]:

```
plt.figure(figsize=(12,5))
sns.countplot(data=df_final1,
              y="Genre",order=df_final1["Genre"].value_counts().index[0:10])
plt.title("Top 10 geners in netflix")
plt.show()
```



In [ ]:

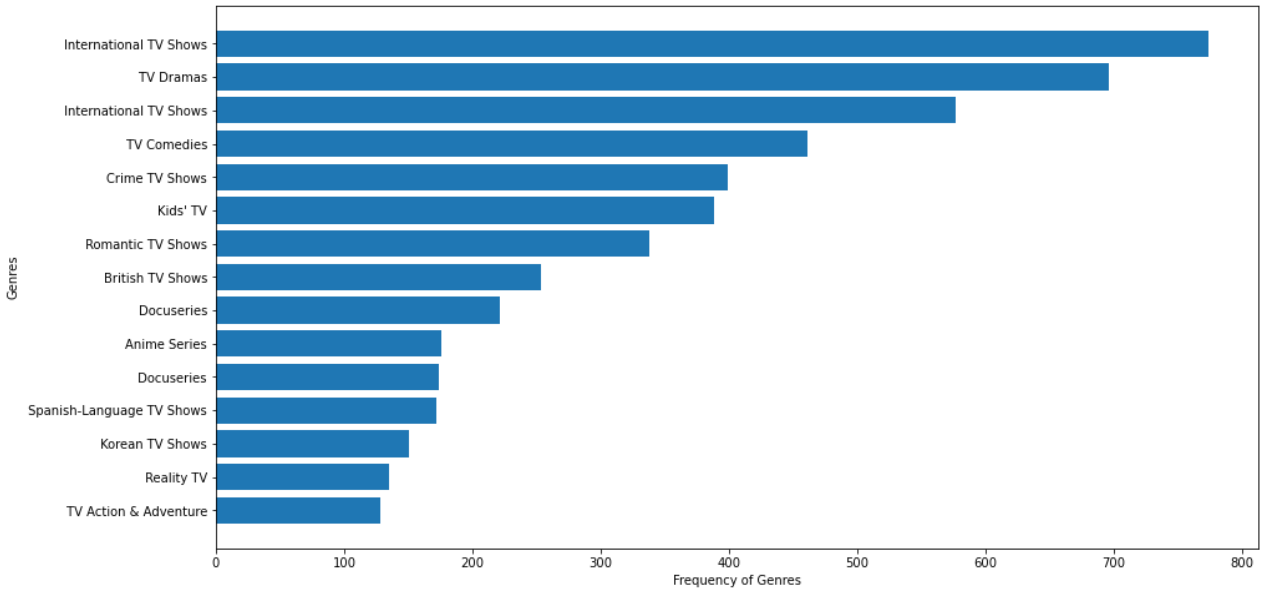
```
#Separate analysis for tvshows and movies
```

In [53]:

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

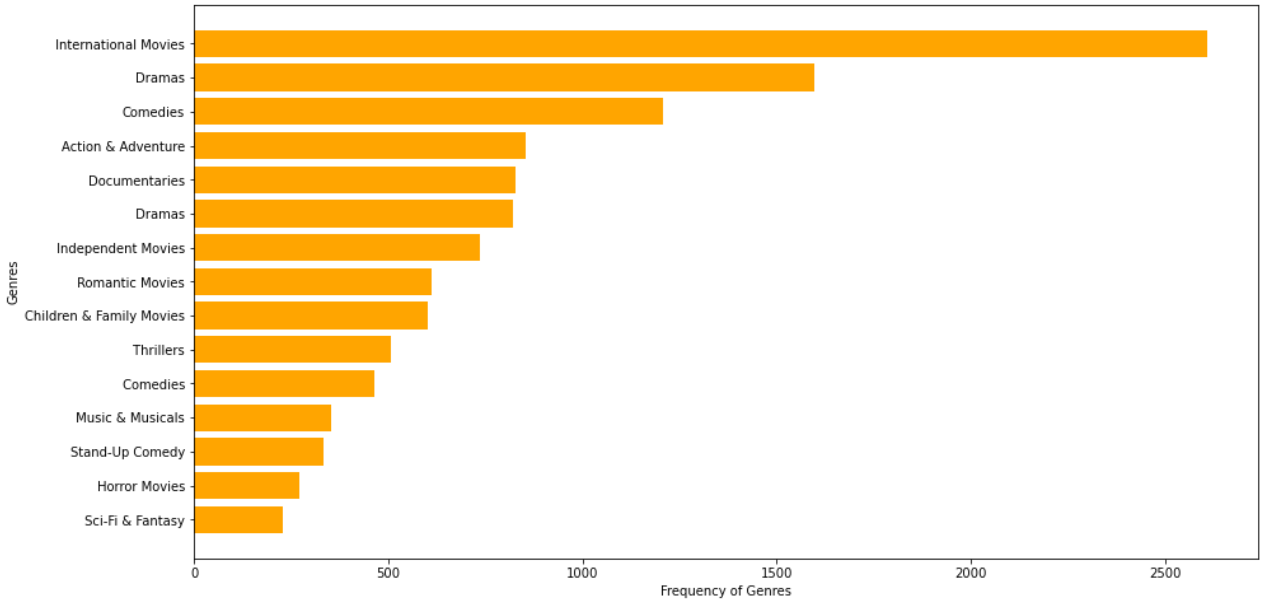
In [54]:

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



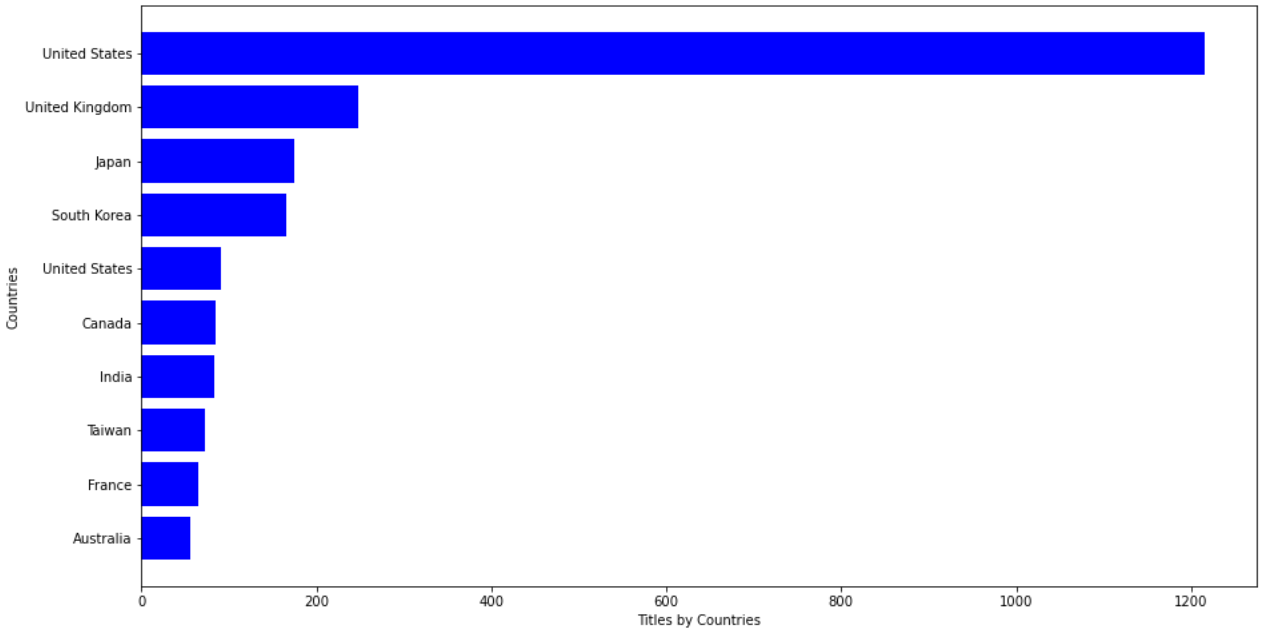
In [55]:

```
df_genre=df_movies.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



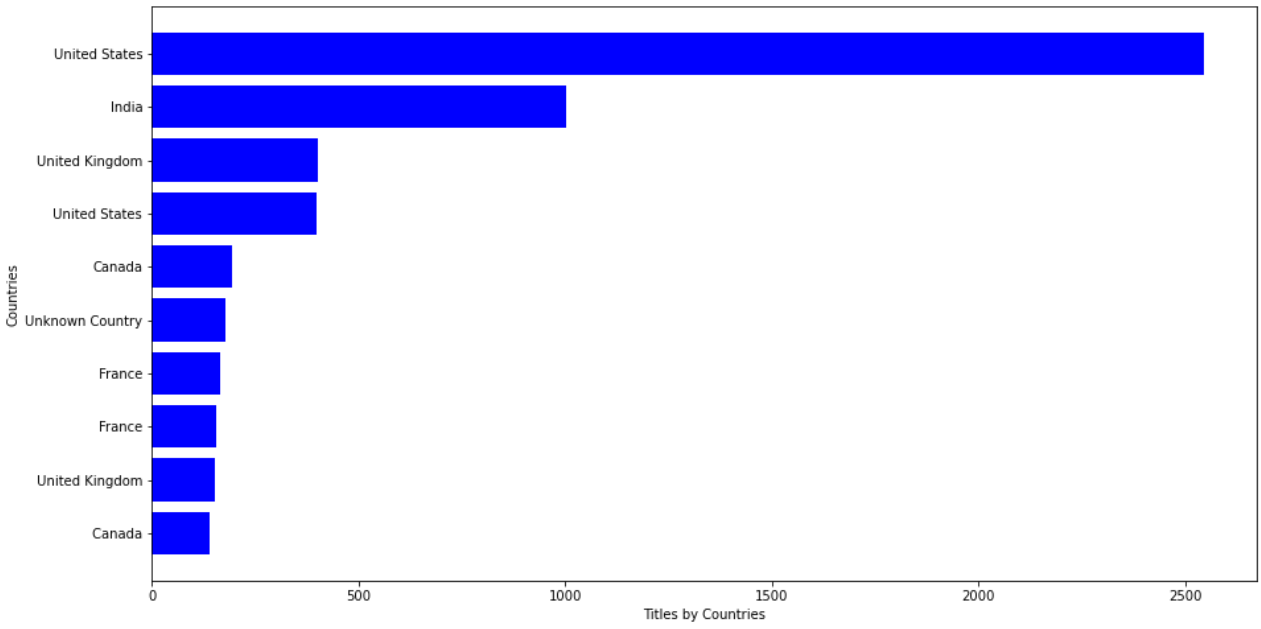
In [56]:

```
df_country=df_shows.groupby(['country']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:10]
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 [57]:

```
df_country=df_movies.groupby(['country']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:10]
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 [ ]:

```
#for categorial variable boxplot
```

In [58]:

```
df_final1['year'] = df_final1['date_added'].dt.year
df_final1.head(10)
```

Out[58]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration	year
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100	2021
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
4	Blood & Water	unknown Director	Khosi Ngema	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
5	Blood & Water	unknown Director	Khosi Ngema	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
6	Blood & Water	unknown Director	Khosi Ngema	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
7	Blood & Water	unknown Director	Gail Mabalane	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
8	Blood & Water	unknown Director	Gail Mabalane	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021
9	Blood & Water	unknown Director	Gail Mabalane	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021

In [59]:

```
df_final1['week_Added']=df_final1['date_added'].dt.week
df_final1['month_Added']=df_final1['date_added'].dt.month
df_final1.head(10)
```

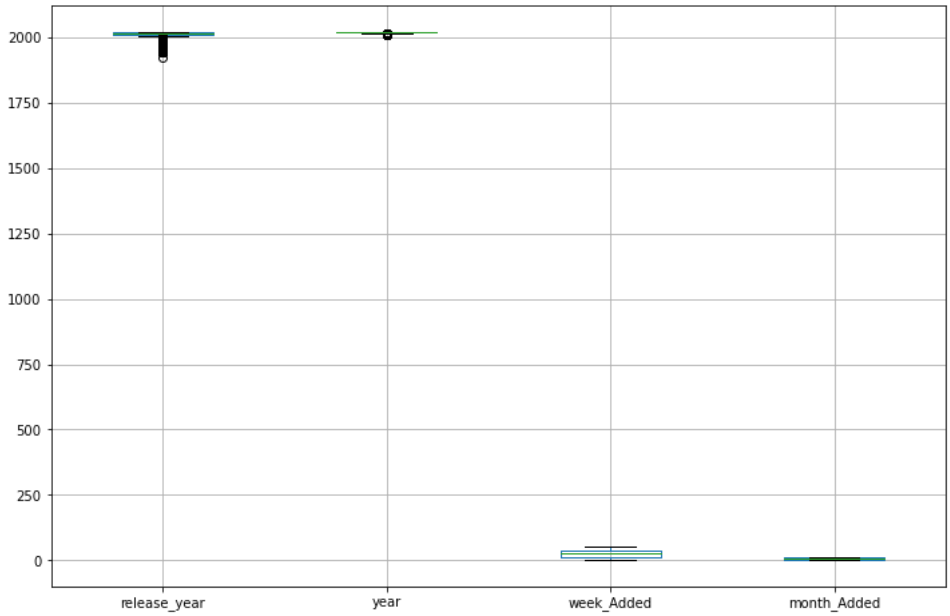
```
C:\Users\anusha\AppData\Local\Temp\ipykernel_6112\2352998475.py:1: FutureWarning: Series.dt.weekofyear and Series.dt.week
have been deprecated. Please use Series.dt.isocalendar().week instead.
  df_final1['week_Added']=df_final1['date_added'].dt.week
```

Out[59]:

	title	director	Actors	Genre	country	show_id	type	date_added	release_year	rating	duration	year	week_Added	month_Added
0	Dick Johnson Is Dead	Kirsten Johnson	unknown Actor	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100	2021	38	9
1	Blood & Water	unknown Director	Ama Qamata	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	38	9
2	Blood & Water	unknown Director	Ama Qamata	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	38	9
3	Blood & Water	unknown Director	Ama Qamata	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	38	9
4	Blood &	unknown	Khosi	International	South	s2	TV	2021-09-24	2021	TV-	2	2021	38	9

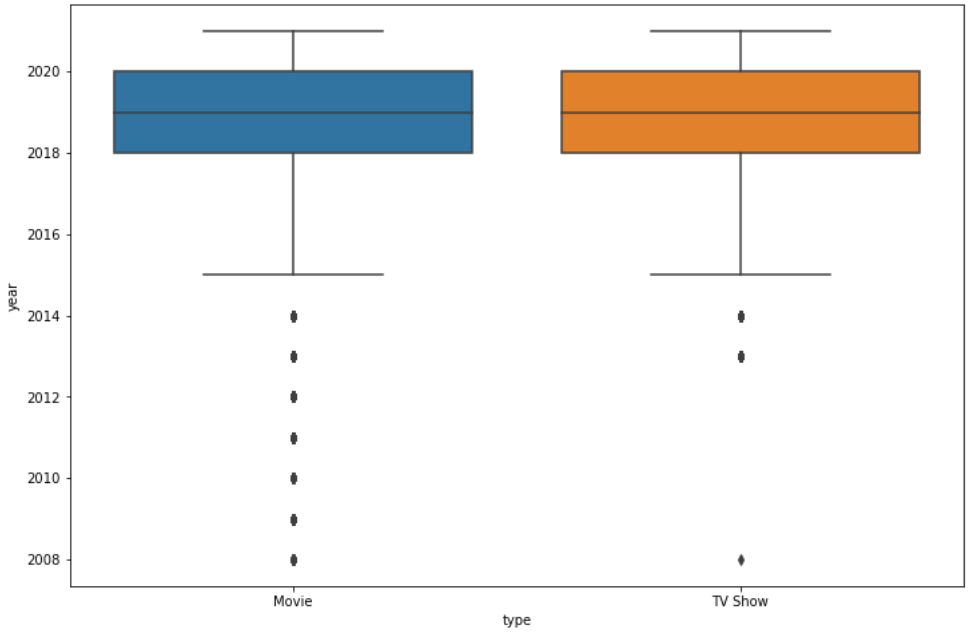
In [60]:

```
plt.figure(figsize=(12,8))
df_final1.boxplot()
plt.show()
```



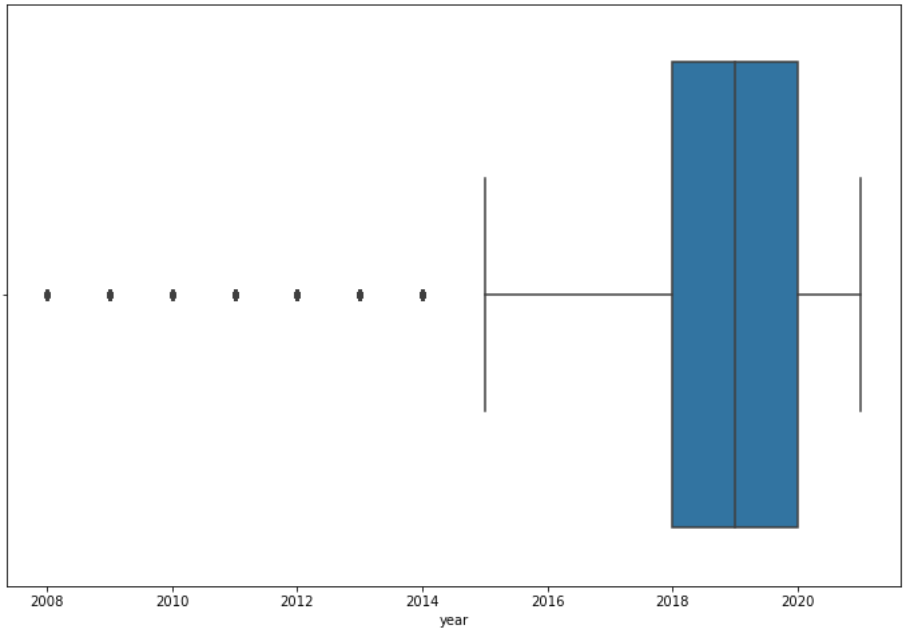
In [61]:

```
plt.figure(figsize=(12,8))
sns.boxplot(data = df_final1, x="type", y="year")
plt.show()
```



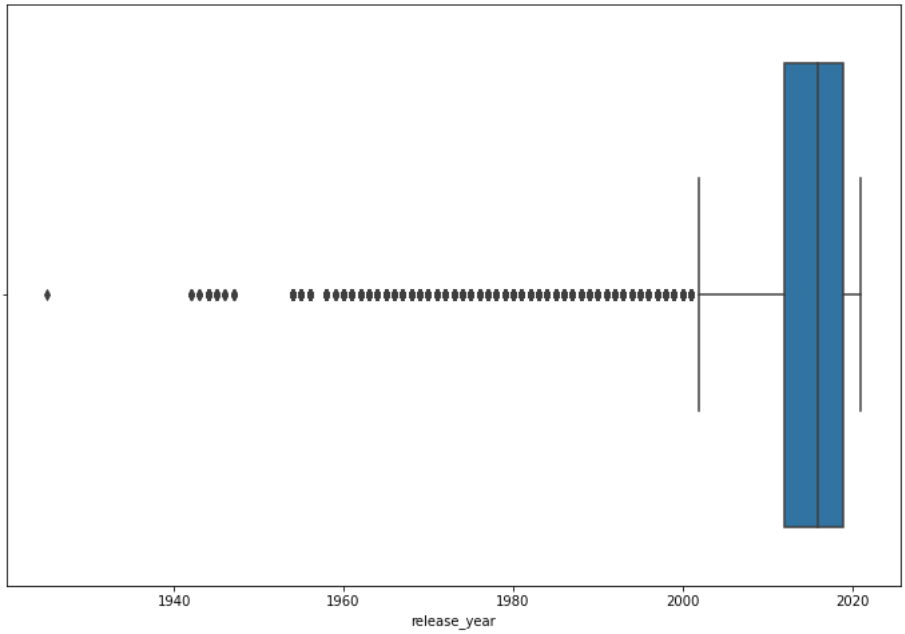
In [62]:

```
plt.figure(figsize=(12,8))
sns.boxplot(data = df_final1,  x="year")
plt.show()
```



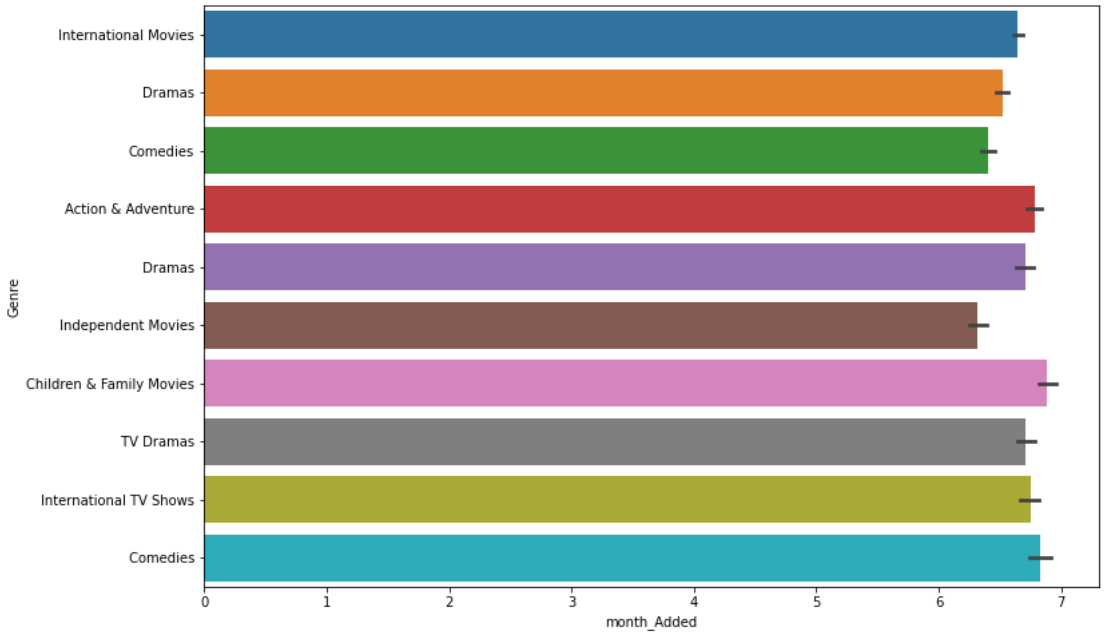
In [63]:

```
plt.figure(figsize=(12,8))
sns.boxplot(data = df_final1,  x="release_year")
plt.show()
```



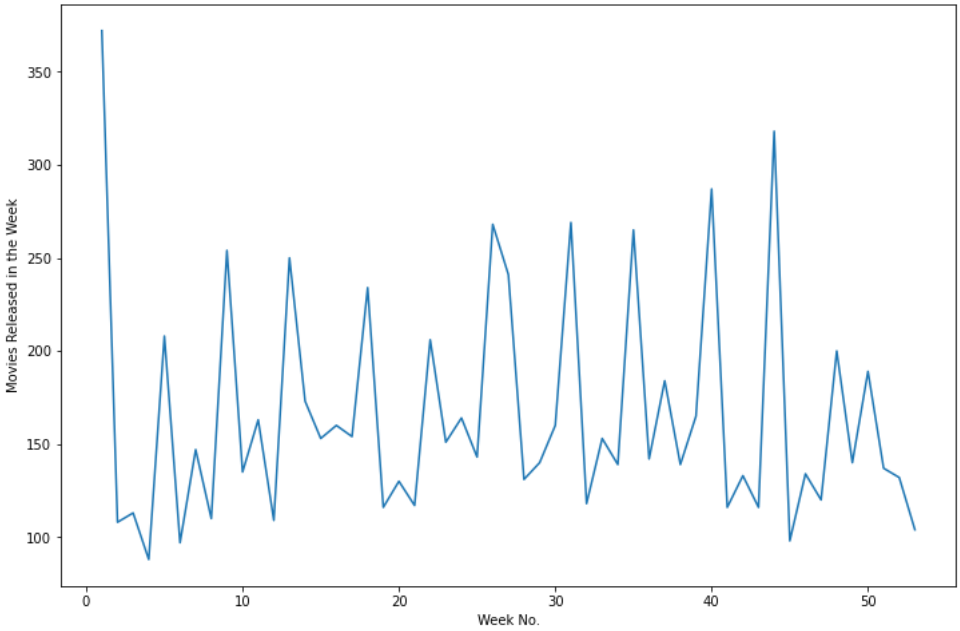
In [64]:

```
#bivariant analysis of month_added and genre
plt.figure(figsize=(12,8))
sns.barplot(data=df_final1,
            x="month_Added",
            y='Genre',order=df_final1["Genre"].value_counts().index[0:10])
plt.show()
```



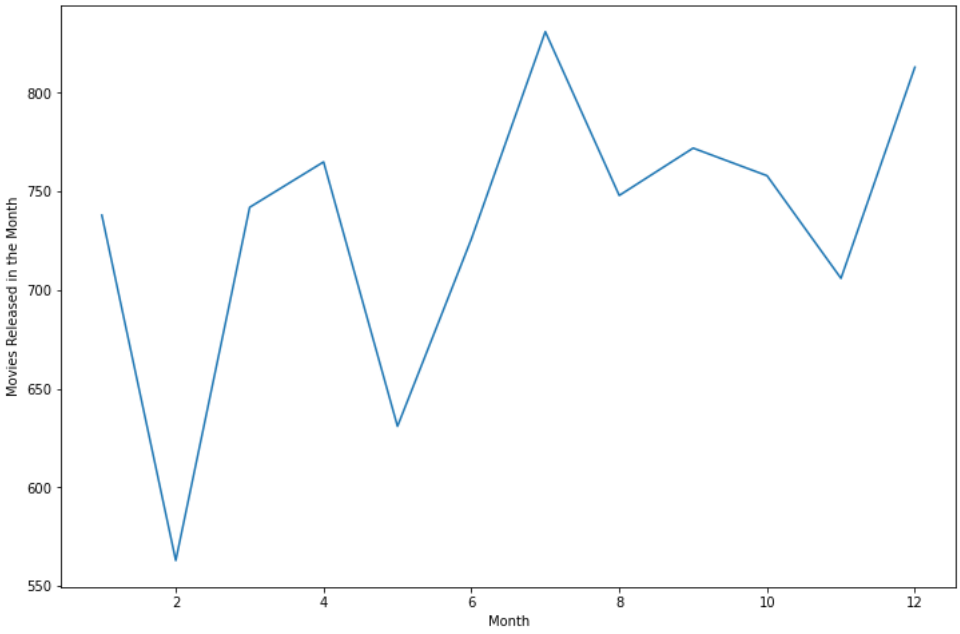
In [65]:

```
#univariate analysis of week_added
df_week=df_final1.groupby(['week_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(12,8))
sns.lineplot(data=df_week, x='week_Added', y='title')
plt.ylabel("Movies Released in the Week")
plt.xlabel("Week No.")
plt.show()
```



In [66]:

```
#univariate analysis of month_added
df_month=df_final1.groupby(['month_Added']).agg({"title":"nunique"}).reset_index()
plt.figure(figsize=(12,8))
sns.lineplot(data=df_month, x='month_Added', y='title')
plt.ylabel("Movies Released in the Month")
plt.xlabel("Month")
plt.show()
```



In [ ]:

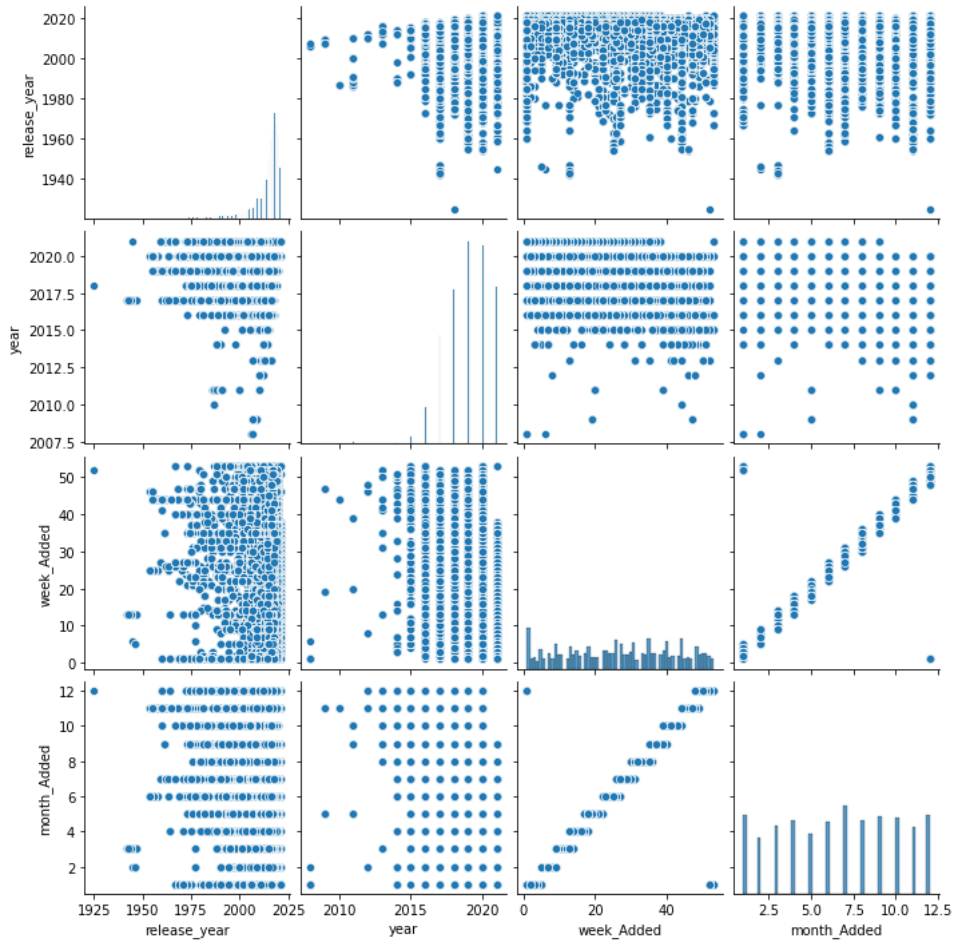
```
#For correlation: Heatmaps, Pairplots
```

In [67]:

```
sns.pairplot(data=df_final1)
```

Out[67]:

```
<seaborn.axisgrid.PairGrid at 0xe13a256b80>
```



In [68]:

```
#correlation
df_final1.corr()
```

Out[68]:

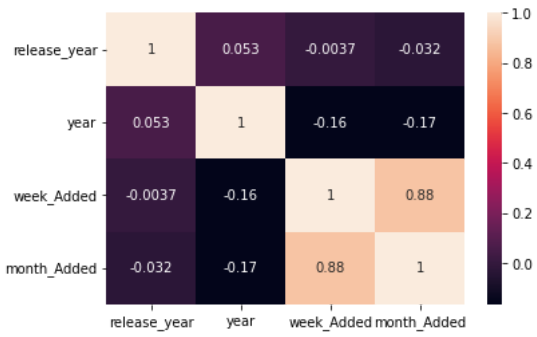
	release_year	year	week_Added	month_Added
release_year	1.000000	0.052768	-0.003676	-0.032396
year	0.052768	1.000000	-0.164870	-0.166904
week_Added	-0.003676	-0.164870	1.000000	0.879524
month_Added	-0.032396	-0.166904	0.879524	1.000000

In [69]:

```
sns.heatmap(df_final1.corr(),annot=True)
```

Out[69]:

```
<AxesSubplot:>
```



In [ ]:

```
#5 missing values and outliners check
```

In [ ]:

```
#Missing values
Have been filled by using different option for diffrent columns . First all the nested columns where exploded
becuase of which the missing values increased.
1. The directors,Actors,country where filled as 'unknown Directors', 'Unknown actor','unknown country'
2. The missing values in duration columns where filled by taking values in rating column
3. The rating column values where filled as'NR'
4.country column is imputed on the basis of director,suppose theres a null for country.when we have a director whose other
movies have a country given.Then we find the mode of country for the directorand imputes in place of nulls the
corresponding mode
5.date added column is imputed on the basis of release year,suppose theres a null for date_added.when release year was 2013.
Then we find the mode of date added for release year=2013 and imputes in place of nulls the corresponding mode.
#outiners check
They are negative outliners when I had done the boxplot for the type and year. No positive outliners . The negative outliners
in movies are more compared to tv shows.
```

In [ ]:

```

```

In [70]:

```
#range of attributes

df_final1.describe() # max year is 2021 and min year is 1925 and 2018 is mean year
```

Out[70]:

	release_year	year	week_Added	month_Added
count	202065.000000	202065.000000	202065.00000	202065.000000
mean	2013.448950	2018.964789	26.69835	6.636815
std	9.013616	1.551987	15.04647	3.441152
min	1925.000000	2008.000000	1.00000	1.000000
25%	2012.000000	2018.000000	14.00000	4.000000
50%	2016.000000	2019.000000	27.00000	7.000000
75%	2019.000000	2020.000000	39.00000	10.000000
max	2021.000000	2021.000000	53.00000	12.000000

In [ ]:

```
#Comments on the distribution of the variables and relationship between them
```

In [ ]:

```
#Comments for each univariate and bivariate plot
```

In [ ]:

```
#Univariate plots
--> In the duration copy plot we find that they are two distant types of duration count
--> In the countplot of duration I see that the 80-90 min bin has the highest count
-->In the histogram of date_added I can observe that more movies where relased duration 2020-2021 then there a slight drop
-->In the rating countplot the tv-ma has hightest count
--> The type plot shows more movies are produced by netflix
--> The release count plot shows 2018 has the highest year for movie release and kde plot it shows Content across
Netflix has increased from 2008 continuously till 2019
-->In the country plot the unites states top its followed by India
--> The Genre plot shows top 10 genres in which international dramas topped followed by Dramas and Comedies
--> According to the week plot Most of the Content across Netflix is added in the first week of the year and
it follows a bit of a cyclical pattern
--> The month plot shows Most of the content is added in the first and last months across Netflix

--> In the separate analysis of movies and Tv shows United States is leading across both TV Shows and Movies, UK also
provides great content across TV Showsand 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 [ ]:

```
#Business Insights
so far I have performed column wise univariate,bivariant and boxplot analysis . Following are the business insights by observing
the patterns in the data
--> Netflix produces more movies than tvshows
--> The united states topped in both international movies and tv shows. India lands in second place in production of movies
after US
--> Most content on the netflix is for mature audiences
-->In the year 2018 released more content compare to any other year
--> International movies and dramas are most popular on netflix
--> The duration two diffrent types of data one is seasons and other is in minutes
-->Net content release which are later uploaded to Netflix has increased since 1980 till 2020
though later reduced certainly due to COVID-19
--> The release content mostly during the first week  of the month. It concentrates more on first and last month's of the year
```

In [ ]:

```
#analysis for recommedations
```

In [ ]:

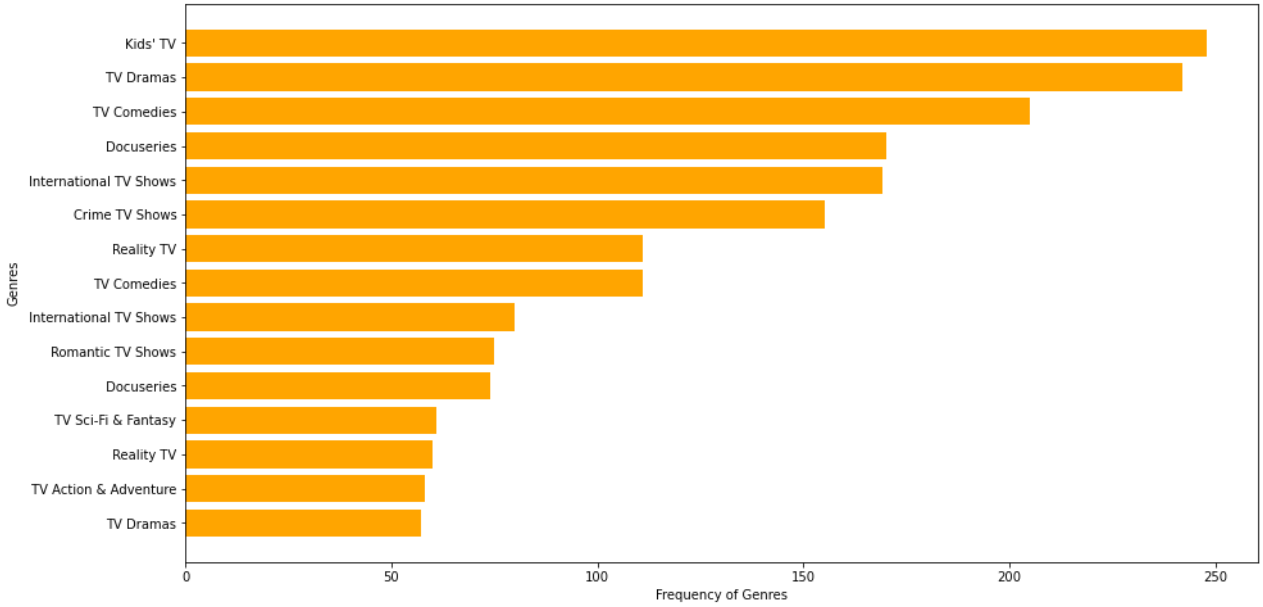
```
#Analysis of USA for movies and shows
```

In [71]:

```
#Analyzing USA for both shows and movies
df_usa_shows=df_final1[df_final1['country']=='United States'][df_final1[df_final1['country']=='United States']['type']=='TV Show']
df_usa_movies=df_final1[df_final1['country']=='United States'][df_final1[df_final1['country']=='United States']['type']=='Movie']
```

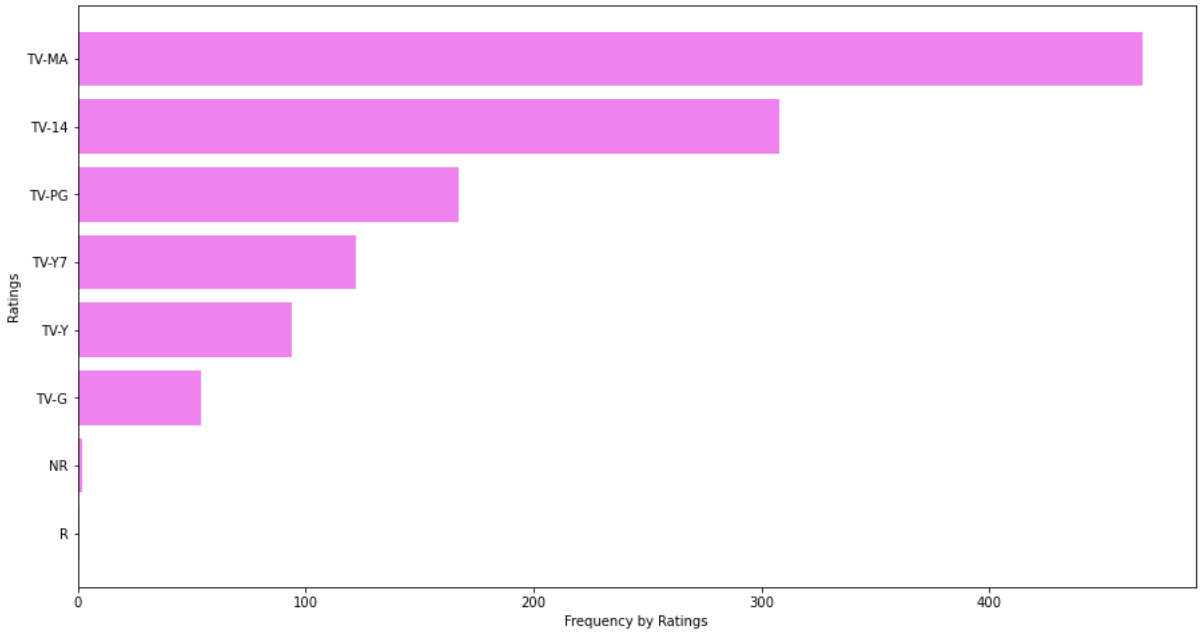
In [72]:

```
df_genre=df_usa_shows.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



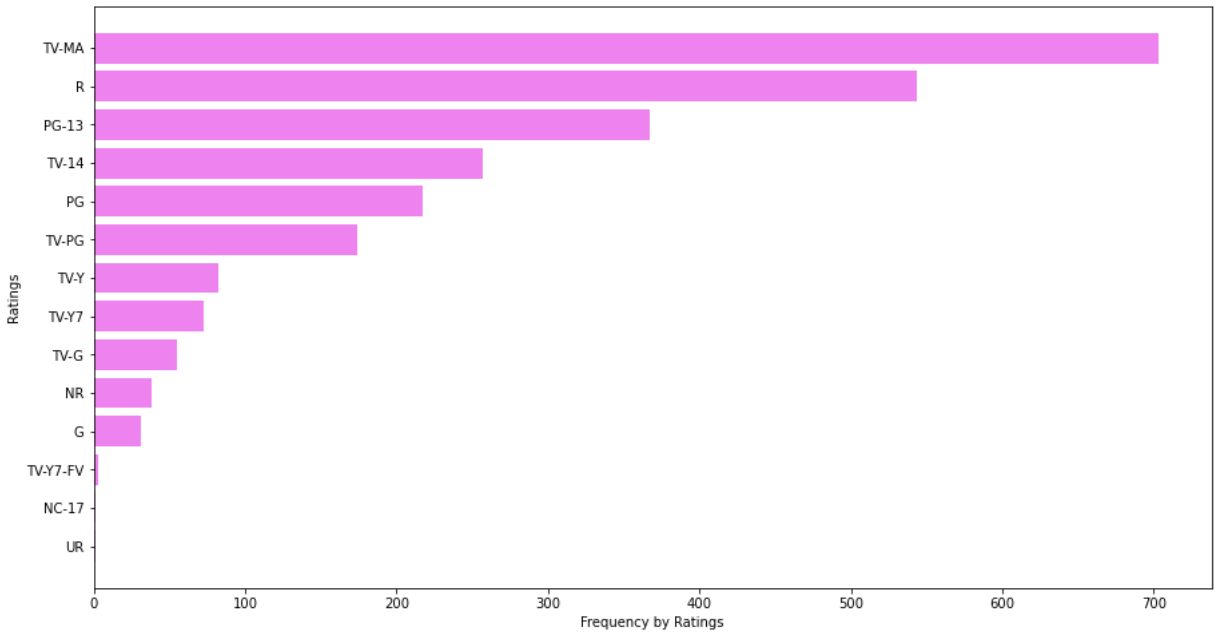
In [73]:

```
df_rating=df_usa_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



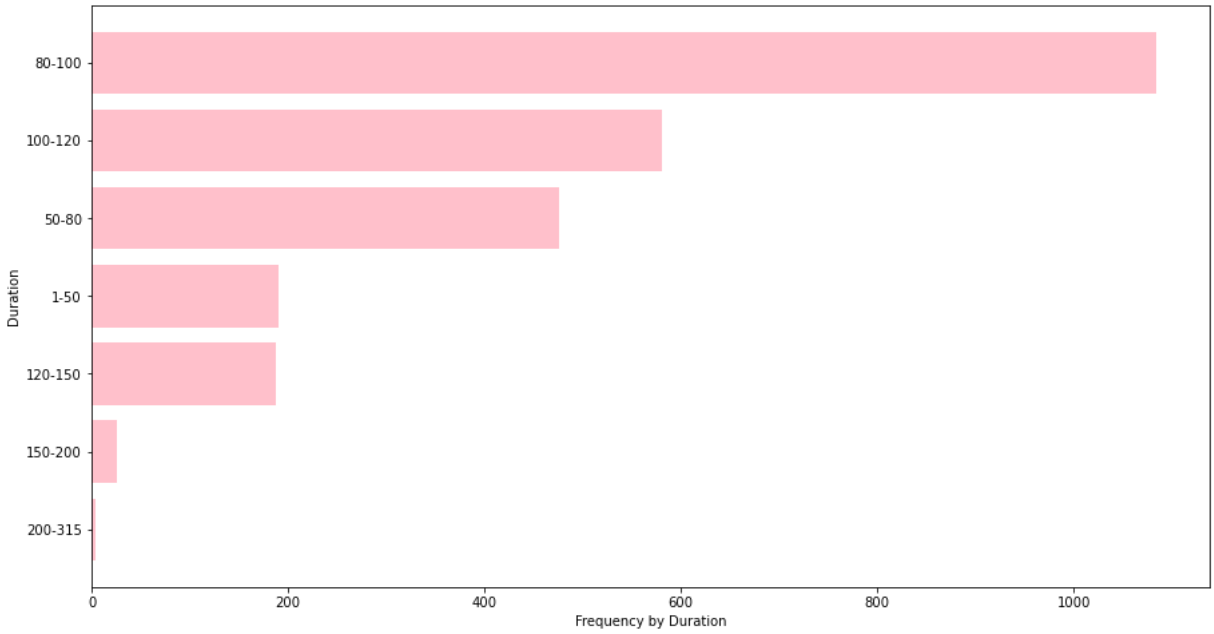
In [74]:

```
df_rating=df_usa_movies.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



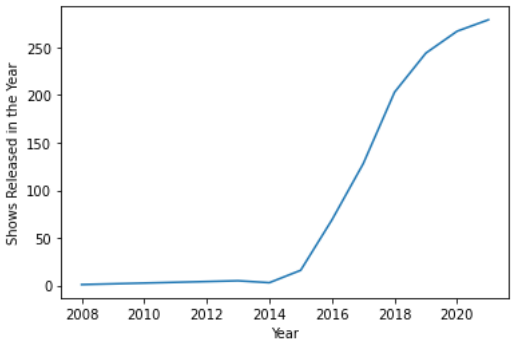
In [75]:

```
df_duration=df_usa_movies.groupby(['duration']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:10]
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()
```



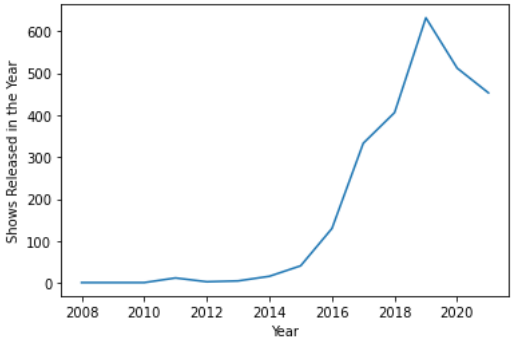
In [76]:

```
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()
```



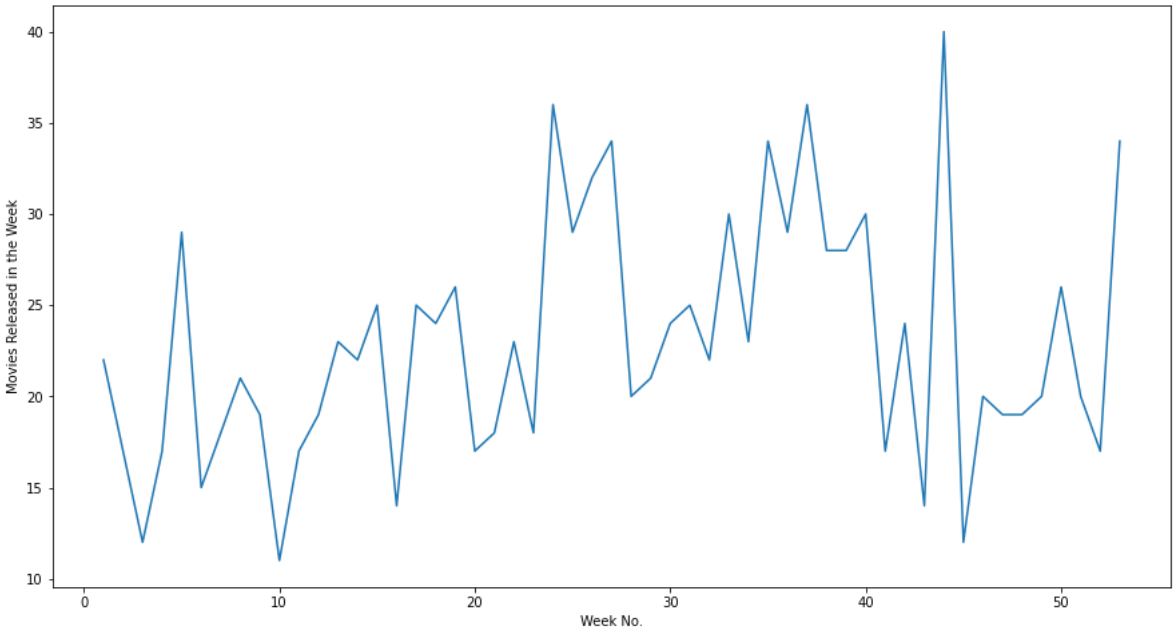
In [77]:

```
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()
```



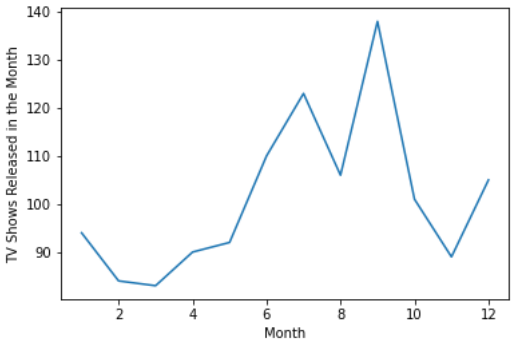
In [78]:

```
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()
```



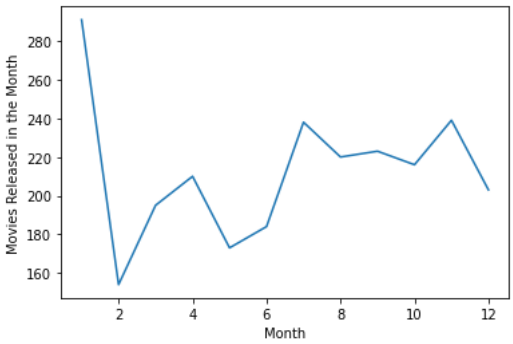
In [81]:

```
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()
```



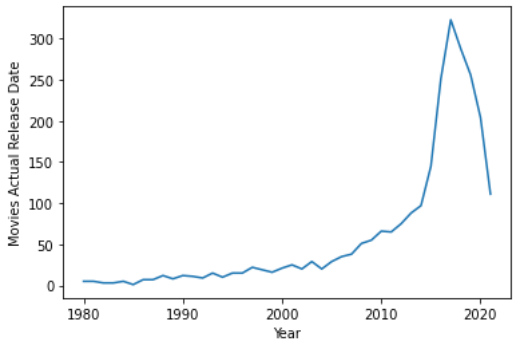
In [83]:

```
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()
```



In [84]:

```
df_release_year=df_usa_movies[df_usa_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 [ ]:

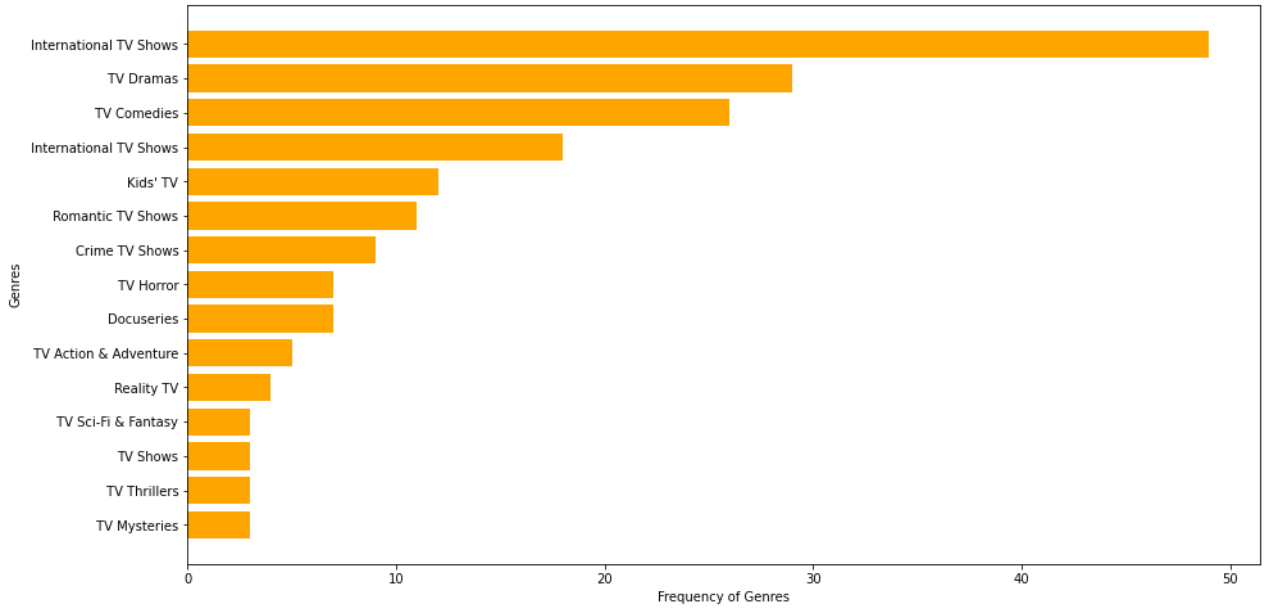
```
#Univariate Analysis separately for shows and movies in India
```

In [85]:

```
#Analyzing India for both shows and movies
df_india_shows=df_final1[df_final1['country']=='India'][df_final1[df_final1['country']=='India']['type']=='TV Show']
df_india_movies=df_final1[df_final1['country']=='India'][df_final1[df_final1['country']=='India']['type']=='Movie']
```

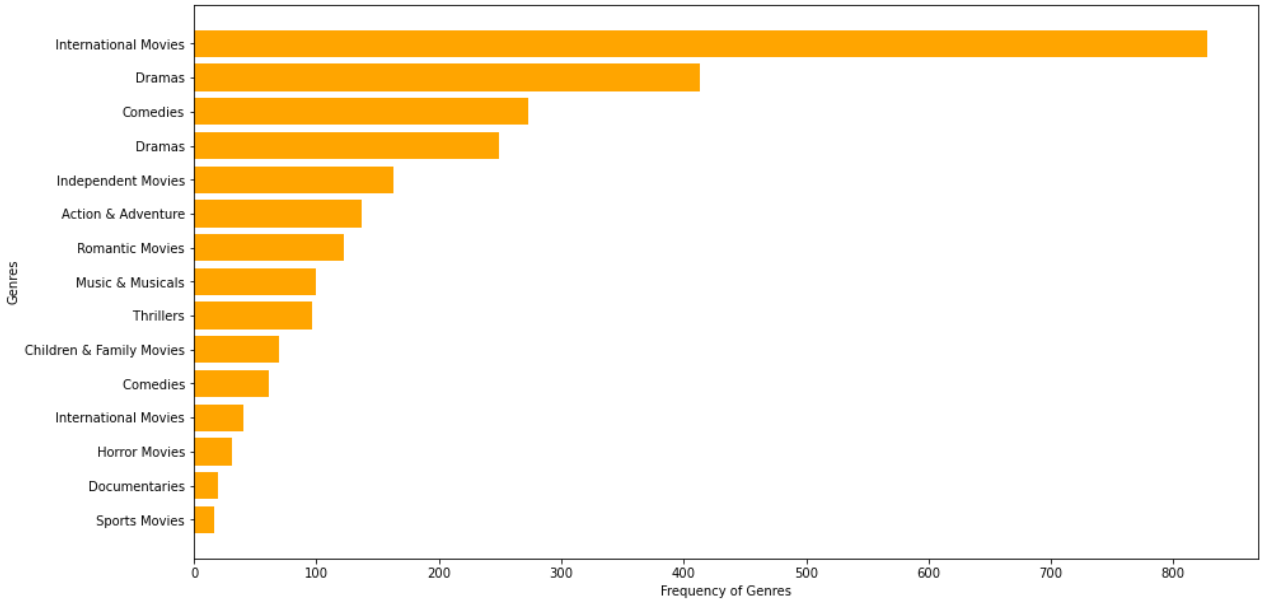
In [86]:

```
df_genre=df_india_shows.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



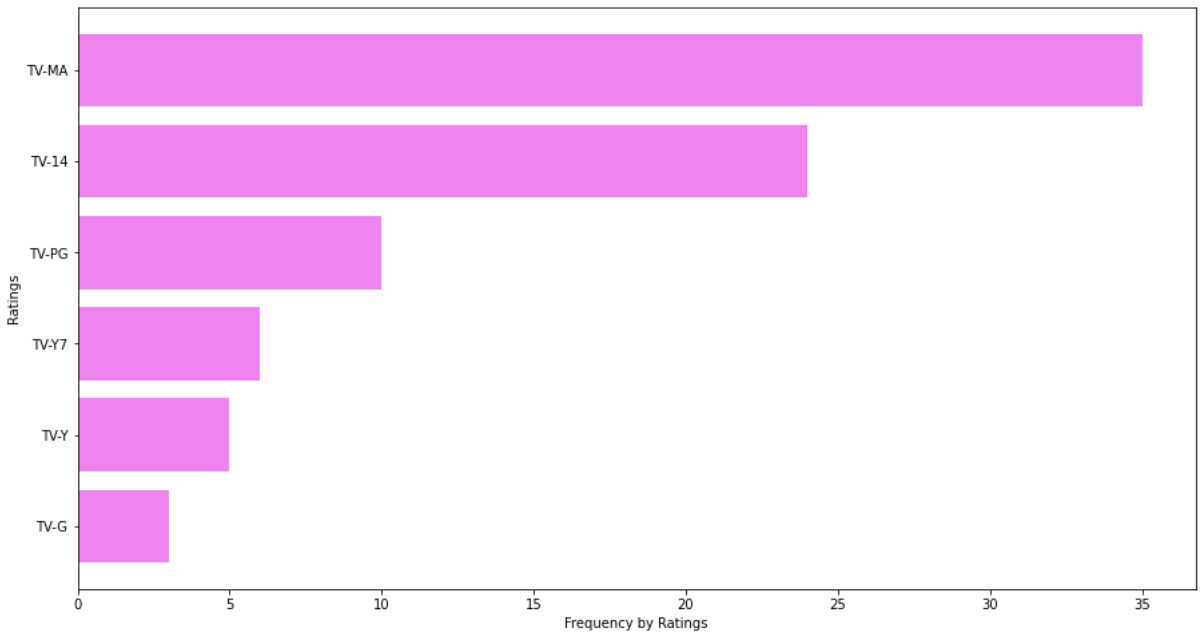
In [87]:

```
df_genre=df_india_movies.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



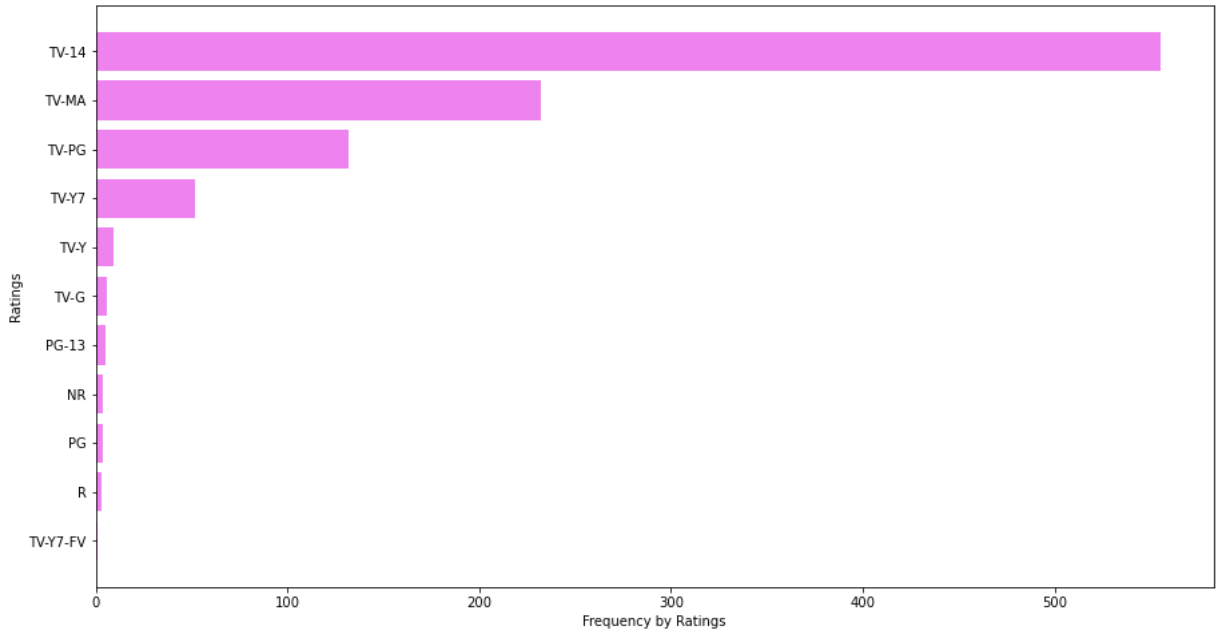
In [88]:

```
df_rating=df_india_shows.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



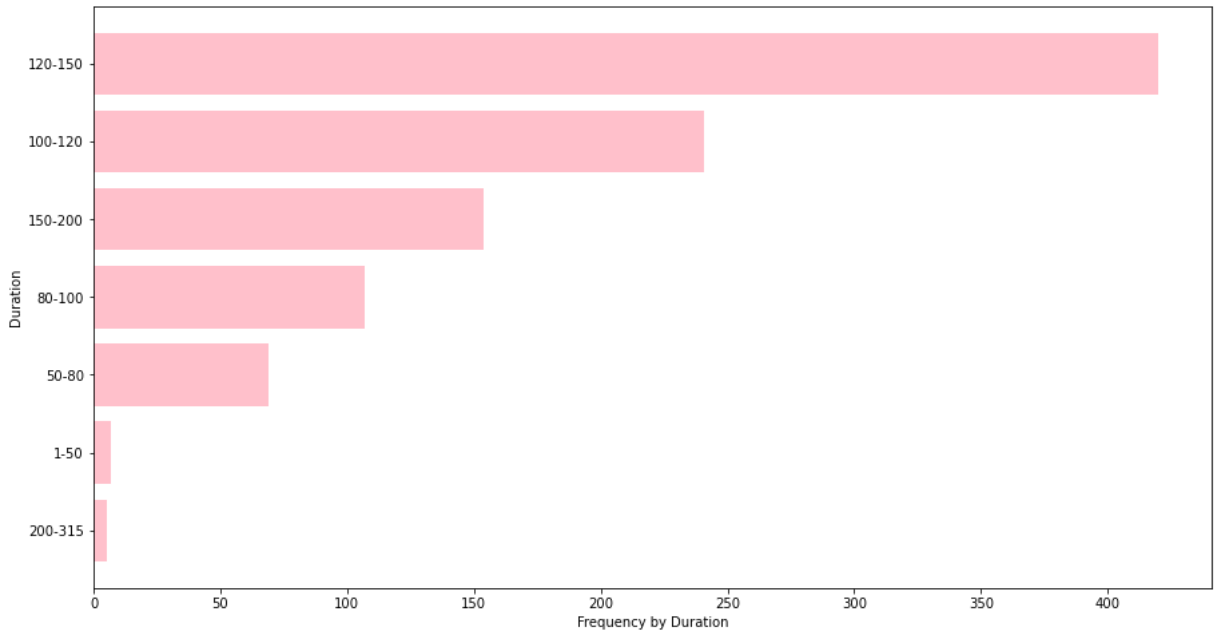
In [89]:

```
df_rating=df_india_movies.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
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()
```



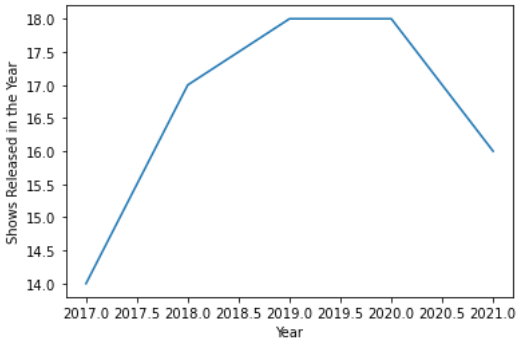
In [90]:

```
df_duration=df_india_movies.groupby(['duration']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:10]
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()
```



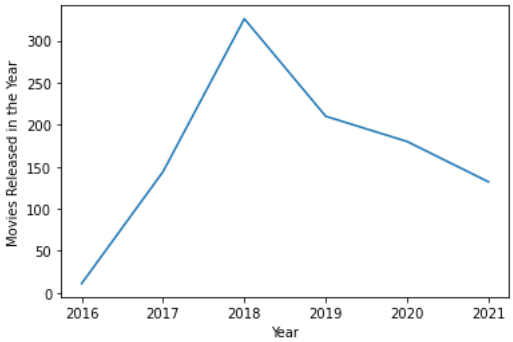
In [91]:

```
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()
```



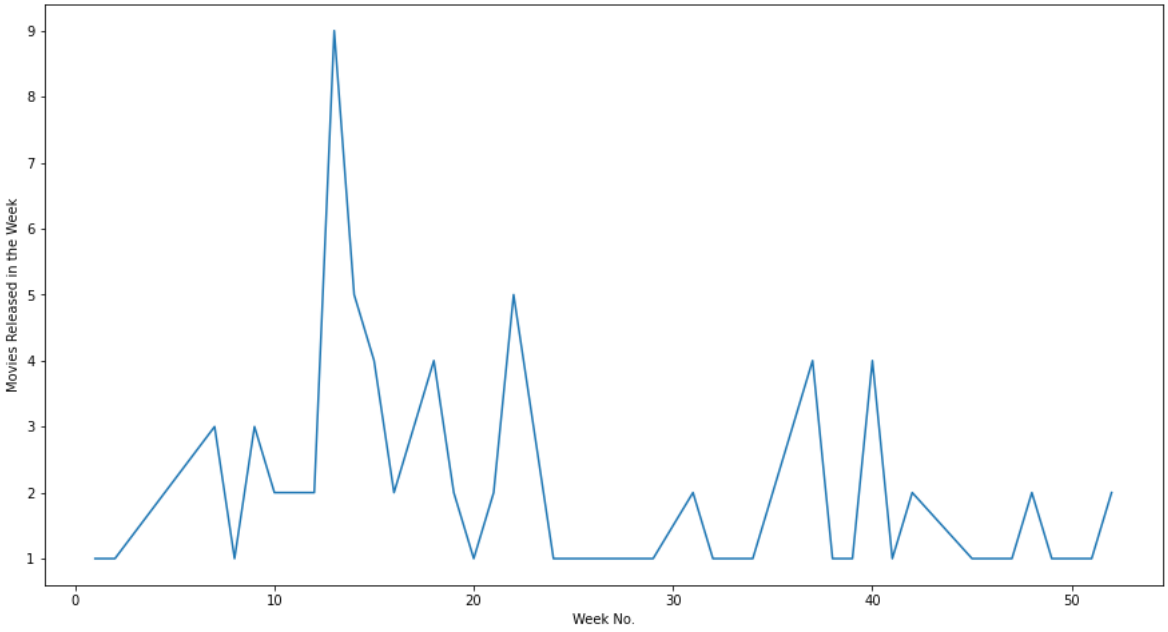
In [92]:

```
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()
```



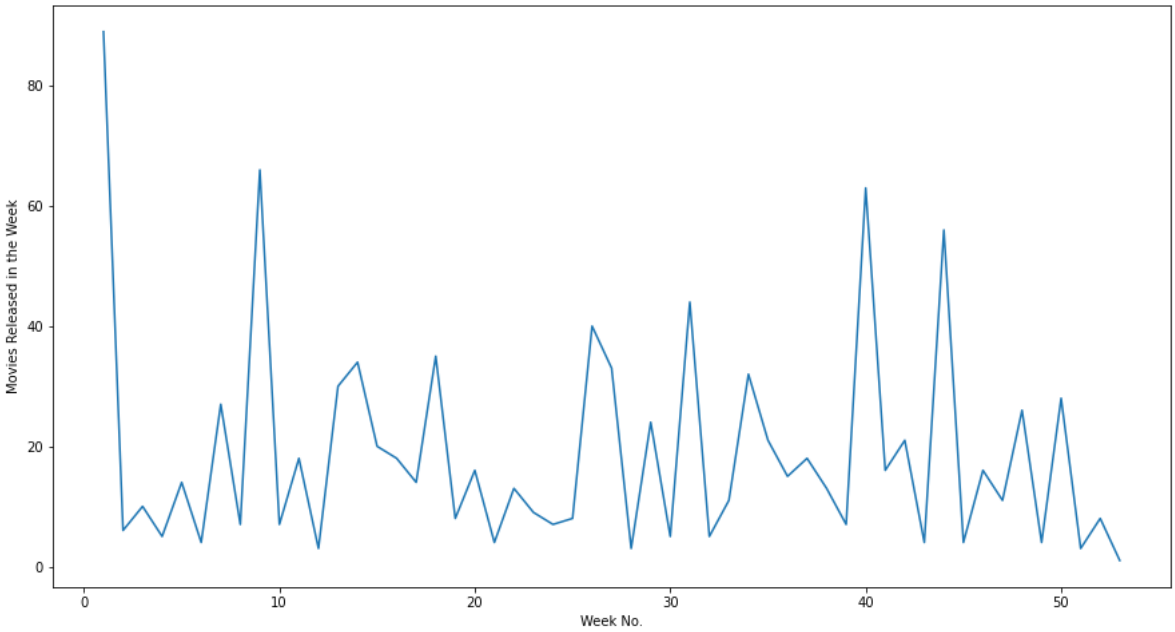
In [93]:

```
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()
```



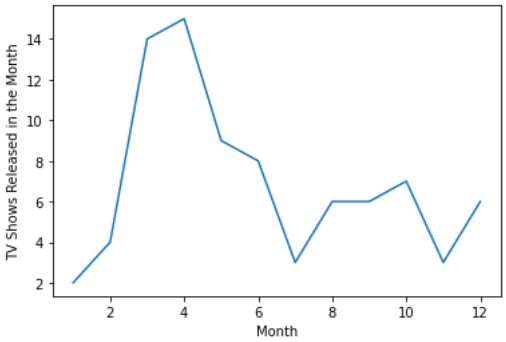
In [94]:

```
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()
```



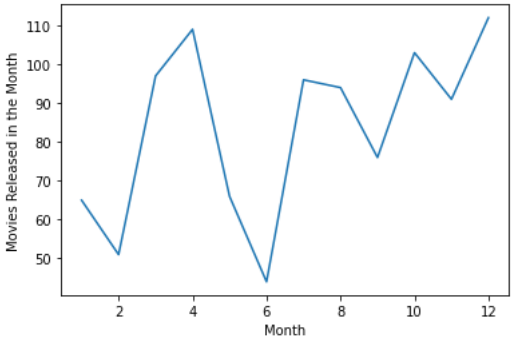
In [96]:

```
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()
```



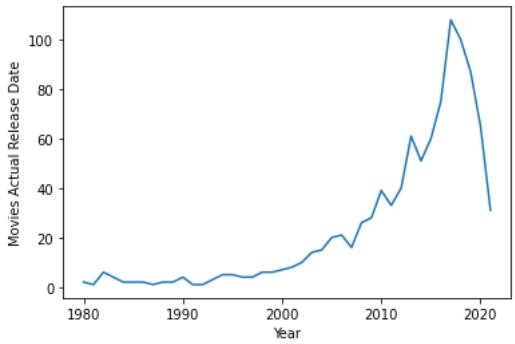
In [98]:

```
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()
```



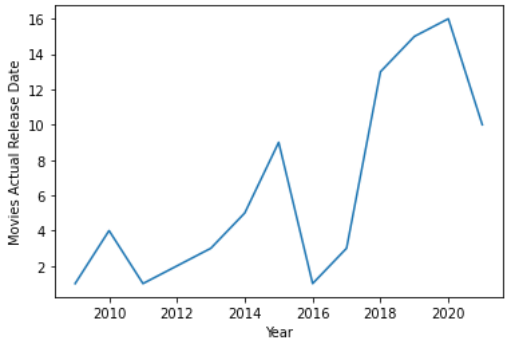
In [99]:

```
df_release_year=df_india_movies[df_india_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 [100]:

```
df_release_year=df_india_shows[df_india_shows['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 [ ]:

```
#from the above analysis recommendations are:

1) The most popular Genres across the countries and in both TV Shows and Movies are Drama, Comedy and International TV Shows/Movies, so c

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

6)Add movies for Indian Audience, it has been declining since 2018.
```