Netflix Data Set

In [2]:

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

In [3]:

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

Netflix Data Set

In [58]:

```
df
```

Out[58]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021
...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009

	show_id	type	title	director	cast	country	date_added	release_year
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias	India	March 2, 2019	2015

In [5]:

```
df[["type","title","cast","rating"]]
```

Out[5]:

	type	title	cast	rating
0	Movie	Dick Johnson Is Dead	NaN	PG-13
1	TV Show	Blood & Water	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	TV-MA
2	TV Show	Ganglands	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	TV-MA
3	TV Show	Jailbirds New Orleans	NaN	TV-MA
4	TV Show	Kota Factory	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	TV-MA
...	...	...	...	...
8802	Movie	Zodiac	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	R
8803	TV Show	Zombie Dumb	NaN	TV-Y7
8804	Movie	Zombieland	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	R
8805	Movie	Zoom	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	PG
8806	Movie	Zubaan	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	TV-14

8807 rows × 4 columns

In [6]:

```python
df[df["release_year"]>2020][["title"]]
```

Out[6]:

	title
1	Blood & Water
2	Ganglands
3	Jailbirds New Orleans
4	Kota Factory
5	Midnight Mass
...	...
1468	What Happened to Mr. Cha?
1551	Hilda
1696	Polly Pocket
2920	Love Is Blind
8437	The Netflix Afterparty

592 rows × 1 columns

In [7]:

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

Out[7]:

```
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 [8]:

```python
df["director"].info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 8807 entries, 0 to 8806
Series name: director
Non-Null Count  Dtype
--------------  -----
6173 non-null   object
dtypes: object(1)
memory usage: 68.9+ KB
```

In [9]:

```python
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 [228]:

```python
newdf=df.copy(deep=True)
```

In [229]:

```
newdf
```

Out[229]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021
...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009

	show_id	type	title	director	cast	country	date_added	release_year
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015

8807 rows × 12 columns

In [12]:

```
# Check for null values
newdf.isnull().sum()
```

Out[12]:

```
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 [ ]:

```
# newdf1 = Unnested Cast
# newdf2 = unnested genres
# newdf3= unnested directors
#newdf4=unnested contries
```

In [13]:

```
# Basically unnesting the cast column
# constraint=newdf["cast"].str.split(",")      {Approach1}
constraint=newdf["cast"].apply(lambda x:str(x).split(',')).tolist()
```

In [14]:

```
constraint
```

```
  ' Thabang Molaba',
  ' Dillon Windvogel',
  ' Natasha Thahane',
  ' Arno Greeff',
  ' Xolile Tshabalala',
  ' Getmore Sithole',
  ' Cindy Mahlangu',
  ' Ryle De Morny',
  ' Greteli Fincham',
  ' Sello Maake Ka-Ncube',
  ' Odwa Gwanya',
  ' Mekaila Mathys',
  ' Sandi Schultz',
  ' Duane Williams',
  ' Shamilla Miller',
  ' Patrick Mofokeng'],
 ['Sami Bouajila',
  ' Tracy Gotoas',
  ' Samuel Jouy',
  ' Nabiha Akkari'
```

In [15]:

```
newdf1=pd.DataFrame(constraint,index=newdf["title"])
```

In [16]:

```
newdf1
```

Out[16]:

title	0	1	2	3	4	5	6	
Dick Johnson Is Dead	nan	None	None	None	None	None	None	
Blood & Water	Ama Qamata	Khosi Ngema	Gail Mabalane	Thabang Molaba	Dillon Windvogel	Natasha Thahane	Arno Greeff	Ts
Ganglands	Sami Bouajila	Tracy Gotoas	Samuel Jouy	Nabiha Akkari	Sofia Lesaffre	Salim Kechiouche	Noureddine Farihi	G Ran
Jailbirds New Orleans	nan	None	None	None	None	None	None	
Kota Factory	Mayur More	Jitendra Kumar	Ranjan Raj	Alam Khan	Ahsaas Channa	Revathi Pillai	Urvi Singh	Aru
...	...	...	...	...	...	...	...	
Zodiac	Mark Ruffalo	Jake Gyllenhaal	Robert Downey Jr.	Anthony Edwards	Brian Cox	Elias Koteas	Donal Logue	Joh
Zombie Dumb	nan	None	None	None	None	None	None	
Zombieland	Jesse Eisenberg	Woody Harrelson	Emma Stone	Abigail Breslin	Amber Heard	Bill Murray	Derek Graf	
Zoom	Tim Allen	Courteney Cox	Chevy Chase	Kate Mara	Ryan Newman	Michael Cassidy	Spencer Breslin	
Zubaan	Vicky Kaushal	Sarah-Jane Dias	Raaghav Chanana	Manish Chaudhary	Meghna Malik	Malkeet Rauni	Anita Shabdish	Chit

8807 rows × 50 columns

In [17]:

```
# we want the names on top down approach so we use stack
newdf1=newdf1.stack()
```

In [18]:

```
newdf1
```

Out[18]:

```
title
Dick Johnson Is Dead   0                       nan
Blood & Water          0               Ama Qamata
                       1              Khosi Ngema
                       2             Gail Mabalane
                       3           Thabang Molaba
                                             ...
Zubaan                 3          Manish Chaudhary
                       4             Meghna Malik
                       5            Malkeet Rauni
                       6            Anita Shabdish
                       7      Chittaranjan Tripathy
Length: 64951, dtype: object
```

In [19]:

```
newdf1=pd.DataFrame(newdf1.reset_index())
```

In [20]:

```
newdf1
```

Out[20]:

	title	level_1	0
0	Dick Johnson Is Dead	0	nan
1	Blood & Water	0	Ama Qamata
2	Blood & Water	1	Khosi Ngema
3	Blood & Water	2	Gail Mabalane
4	Blood & Water	3	Thabang Molaba
...	...	...	...
64946	Zubaan	3	Manish Chaudhary
64947	Zubaan	4	Meghna Malik
64948	Zubaan	5	Malkeet Rauni
64949	Zubaan	6	Anita Shabdish
64950	Zubaan	7	Chittaranjan Tripathy

64951 rows × 3 columns

In [21]:

```
# Renaming column '0' as 'Actors'
newdf1.rename(columns={0:"Actors"},inplace=True)
```

In [22]:

```
newdf1
```

Out[22]:

	title	level_1	Actors
0	Dick Johnson Is Dead	0	nan
1	Blood & Water	0	Ama Qamata
2	Blood & Water	1	Khosi Ngema
3	Blood & Water	2	Gail Mabalane
4	Blood & Water	3	Thabang Molaba
...	...	...	...
64946	Zubaan	3	Manish Chaudhary
64947	Zubaan	4	Meghna Malik
64948	Zubaan	5	Malkeet Rauni
64949	Zubaan	6	Anita Shabdish
64950	Zubaan	7	Chittaranjan Tripathy

64951 rows × 3 columns

In [23]:

```
newdf1.drop(["level_1"],axis=1,inplace=True)
```

In [24]:

```
newdf1
```

Out[24]:

	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
...	...	...
64946	Zubaan	Manish Chaudhary
64947	Zubaan	Meghna Malik
64948	Zubaan	Malkeet Rauni
64949	Zubaan	Anita Shabdish
64950	Zubaan	Chittaranjan Tripathy

64951 rows × 2 columns

In [25]:

```python
# Unnesting categories
genres=newdf["listed_in"].apply(lambda x:str(x).split(',')).tolist()
```

In [26]:

```python
genres
```

```
 ['TV Comedies'],
 ['British TV Shows', ' International TV Shows', ' TV Comedies'],
 ['International TV Shows', ' TV Dramas', ' TV Thrillers'],
 ["Kids' TV"],
 ['Dramas', ' International Movies', ' Thrillers'],
 ['Action & Adventure', ' Dramas', ' International Movies'],
 ["Kids' TV", ' TV Comedies'],
 ['Action & Adventure', ' Dramas'],
 ["Kids' TV"],
 ["Kids' TV", ' TV Sci-Fi & Fantasy'],
 ['Action & Adventure', ' Classic Movies', ' Dramas'],
 ['Dramas', ' Horror Movies', ' Thrillers'],
 ['Action & Adventure', ' Horror Movies', ' Thrillers'],
 ['Action & Adventure', ' Horror Movies', ' Thrillers'],
 ['Documentaries'],
 ['Action & Adventure'],

 ['International TV Shows', ' Romantic TV Shows', ' TV Comedies'],
 ['Dramas', ' Thrillers'],
 ['International TV Shows', ' TV Dramas'],
 ['International TV Shows', ' TV Dramas', ' TV Sci-Fi & Fantasy']
```

In [27]:

```python
newdf2=pd.DataFrame(genres,index=newdf["title"])
```

In [28]:

```python
newdf2.tail(10)
```

Out[28]:

title	0	1	2
Zak Storm	Kids' TV	None	None
Zed Plus	Comedies	Dramas	International Movies
Zenda	Dramas	International Movies	None
Zindagi Gulzar Hai	International TV Shows	Romantic TV Shows	TV Dramas
Zinzana	Dramas	International Movies	Thrillers
Zodiac	Cult Movies	Dramas	Thrillers
Zombie Dumb	Kids' TV	Korean TV Shows	TV Comedies
Zombieland	Comedies	Horror Movies	None
Zoom	Children & Family Movies	Comedies	None
Zubaan	Dramas	International Movies	Music & Musicals

In [145]:

```
newdf2
```

Out[145]:

title	0	1	2
Dick Johnson Is Dead	Documentaries	None	None
Blood & Water	International TV Shows	TV Dramas	TV Mysteries
Ganglands	Crime TV Shows	International TV Shows	TV Action & Adventure
Jailbirds New Orleans	Docuseries	Reality TV	None
Kota Factory	International TV Shows	Romantic TV Shows	TV Comedies
...	...	...	...
Zodiac	Cult Movies	Dramas	Thrillers
Zombie Dumb	Kids' TV	Korean TV Shows	TV Comedies
Zombieland	Comedies	Horror Movies	None
Zoom	Children & Family Movies	Comedies	None
Zubaan	Dramas	International Movies	Music & Musicals

8807 rows × 3 columns

In [146]:

```
newdf
```

Out[146]:

	show_id	type	title	director	cast	country	date_added	release_year	rat
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	P
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	.
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	.
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	.
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	.
...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	

	show_id	type	title	director	cast	country	date_added	release_year	rat
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV

8807 rows × 12 columns

In [29]:

```python
newdf["director"].value_counts()
```

Out[29]:

```
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 [31]:

```python
newdf2.rename(columns={0:"Category1",1:"Category2",2:"Category3"},inplace=True)
```

In [32]:

```python
dirs=newdf["director"].apply(lambda x:str(x).split(',')).tolist()
# genres=newdf["listed_in"].apply(lambda x:str(x).split(',')).tolist()
```

In [33]:

```
dirs
```

Out[33]:

```
[['Kirsten Johnson'],
 ['nan'],
 ['Julien Leclercq'],
 ['nan'],
 ['nan'],
 ['Mike Flanagan'],
 ['Robert Cullen', ' José Luis Ucha'],
 ['Haile Gerima'],
 ['Andy Devonshire'],
 ['Theodore Melfi'],
 ['nan'],
 ['Kongkiat Komesiri'],
 ['Christian Schwochow'],
 ['Bruno Garotti'],
 ['nan'],
 ['nan'],
 ['Pedro de Echave García', ' Pablo Azorín Williams'],
 ['nan'],
```

In [34]:

```
newdf3=pd.DataFrame(dirs,index=newdf["title"])
```

In [35]:

```
newdf3.value_counts()
```

Out[35]:

```
0                1                2                 3                  4
5                6                7                 8                  9
10               11               12
Chris Buck    Jennifer Lee    Patrick Osborne    Lauren MacMullan    John Kahrs
Nathan Greno    Byron Howard    Stevie Wermers    Dean Wellins    Kevin Deters
Roger Allers    Mike Gabriel    Mark Henn     1
dtype: int64
```

In [36]:

```
newdf3
```

Out[36]:

title	0	1	2	3	4	5	6	7	8	9	10	11
Dick Johnson Is Dead	Kirsten Johnson	None	None	None	None	None	None	None	None	None	None	None
Blood & Water	nan	None	None	None	None	None	None	None	None	None	None	None
Ganglands	Julien Leclercq	None	None	None	None	None	None	None	None	None	None	None
Jailbirds New Orleans	nan	None	None	None	None	None	None	None	None	None	None	None
Kota Factory	nan	None	None	None	None	None	None	None	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...	...	...
Zodiac	David Fincher	None	None	None	None	None	None	None	None	None	None	None
Zombie Dumb	nan	None	None	None	None	None	None	None	None	None	None	None
Zombieland	Ruben Fleischer	None	None	None	None	None	None	None	None	None	None	None
Zoom	Peter Hewitt	None	None	None	None	None	None	None	None	None	None	None
Zubaan	Mozez Singh	None	None	None	None	None	None	None	None	None	None	None

8807 rows × 13 columns

In [37]:

```
newdf[newdf["country"]=="India"]
```

Out[37]:

	show_id	type	title	director	cast	country	date_added	release_year
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021
24	s25	Movie	Jeans	S. Shankar	Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi...	India	September 21, 2021	1998
39	s40	TV Show	Chhota Bheem	NaN	Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jig...	India	September 16, 2021	2021
50	s51	TV Show	Dharmakshetra	NaN	Kashmira Irani, Chandan Anand, Dinesh Mehta, A...	India	September 15, 2021	2014
66	s67	TV Show	Raja Rasoi Aur Anya Kahaniyan	NaN	NaN	India	September 15, 2021	2014
...	...	...	...	...	...	...	...	...
8773	s8774	Movie	Yanda Kartavya Aahe	Kedar Shinde	Ankush Choudhary, Smita Shewale, Mohan Joshi, ...	India	January 1, 2018	2006
8775	s8776	TV Show	Yeh Meri Family	NaN	Vishesh Bansal, Mona Singh, Akarsh Khurana, Ah...	India	August 31, 2018	2018
8798	s8799	Movie	Zed Plus	Chandra Prakash Dwivedi	Adil Hussain, Mona Singh, K.K. Raina, Sanjay M...	India	December 31, 2019	2014
8799	s8800	Movie	Zenda	Avadhoot Gupte	Santosh Juvekar, Siddharth Chandekar, Sachit P...	India	February 15, 2018	2009

	show_id	type	title	director	cast	country	date_added	release_year
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015

972 rows × 12 columns

In [38]:

```
df_f=newdf1.merge(df[["show_id","type","title","director","country","listed_in","release_ye
```

In [39]:

```
df_f
```

Out[39]:

	title	Actors	show_id	type	director	country	listed_in	release_year	da
0	Dick Johnson Is Dead	nan	s1	Movie	Kirsten Johnson	United States	Documentaries	2020	S
1	Blood & Water	Ama Qamata	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	S
2	Blood & Water	Khosi Ngema	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	S
3	Blood & Water	Gail Mabalane	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	S
4	Blood & Water	Thabang Molaba	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	S
...	...	...	...	...	...	...	...	...	
64946	Zubaan	Manish Chaudhary	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64947	Zubaan	Meghna Malik	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64948	Zubaan	Malkeet Rauni	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64949	Zubaan	Anita Shabdish	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64950	Zubaan	Chittaranjan Tripathy	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	

64951 rows × 11 columns

In [45]:

```
df_f1=df_f.merge(newdf2,on="title",how="left")
```

In [85]:

```
# Final Unnnested Data Set
df_f1
```

Out[85]:

	title	Actors	show_id	type	director	country	listed_in	release_year	date
0	Dick Johnson Is Dead	nan	s1	Movie	Kirsten Johnson	United States	Documentaries	2020	Se 2
1	Blood & Water	Ama Qamata	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	Se 2
2	Blood & Water	Khosi Ngema	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	Se 2
3	Blood & Water	Gail Mabalane	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	Se 2
4	Blood & Water	Thabang Molaba	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	Se 2
...	...	...	...	...	...	...	...	...	
64946	Zubaan	Manish Chaudhary	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64947	Zubaan	Meghna Malik	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64948	Zubaan	Malkeet Rauni	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64949	Zubaan	Anita Shabdish	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	
64950	Zubaan	Chittaranjan Tripathy	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	

64951 rows × 14 columns

In [130]:

```python
# Copy of the original Dataset to perform analysis
d=newdf.copy()
d
```

Out[130]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021
...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009

	show_id	type	title	director	cast	country	date_added	release_year
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006
					Vicky Kaushal,			

In [131]:

```python
d["datetime"]=pd.to_datetime(d["date_added"])
```

In [132]:

```
d
```

Out[132]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021
...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009

	show_id	type	title	director	cast	country	date_added	release_year
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias,	India	March 2, 2019	2015

In [133]:

```
d.dtypes
```

Out[133]:

```
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
datetime     datetime64[ns]
dtype: object
```

In [134]:

```
d.isnull().sum()
```

Out[134]:

```
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
datetime        10
dtype: int64
```

In [60]:

```
# Filling missing values
```

In [137]:

```python
# Director
d.director.fillna(value="Missing",inplace=True)
# Cast
d.cast.fillna(value="Missing",inplace=True)
# Country
d.country.fillna(value="Missing",inplace=True)
# Date_Added
d.date_added.fillna(value="Missing",inplace=True)
# DateTime
d.datetime.fillna(value="Missing",inplace=True)
```

In [138]:

```python
d.isnull().sum()
```

Out[138]:

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

In [ ]:

```python
"""""As we are examining the data on the basis of duration and rating so we can't fill them
so we drop them"""
```

In [94]:

```python
d.dropna(inplace=True)
```

In [95]:

```python
d.type.value_counts()
```

Out[95]:

```
Movie      6126
TV Show    2674
Name: type, dtype: int64
```

In [96]:

```python
# Taking the above into consideration we plot a pie chart for a clear understanding
```

In [97]:

```python
# graph1=showing tv shows vs movies in the dataset
graph1=d.type.value_counts()
plt.pie(graph1,labels=graph1.index,autopct='%.1f')
plt.legend()
plt.show()
```


In [ ]:
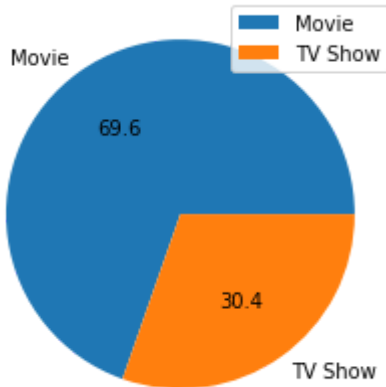
```python
"""As per the pie chart we can say that number of movies are more than tv shows"""
```

In [68]:

```python
# TV Show and Movie releases with respect to year of last 20-30 years
```

In [98]:

```python
years=d[["type","release_year"]]
```

In [99]:

```python
years=years[years["release_year"]>1990]
```

In [100]:

```python
years.release_year.value_counts().sort_index()
```

Out[100]:

```
1991      17
1992      23
1993      28
1994      22
1995      25
1996      24
1997      38
1998      36
1999      39
2000      37
2001      45
2002      51
2003      61
2004      64
2005      80
2006      96
2007      88
2008     136
2009     152
2010     193
2011     185
2012     237
2013     287
2014     352
2015     557
2016     902
2017    1030
2018    1147
2019    1030
2020     953
2021     592
Name: release_year, dtype: int64
```

In [101]:

```python
plt.figure(figsize=(10,5))
count_total=sns.countplot(x="release_year",data=years)
count_total.set_xticklabels(years["release_year"].value_counts().sort_index().index,rotatio
# years["release_year"].value_counts().sort_index().plot(kind="bar",figsize=(10,5))
plt.show()
```
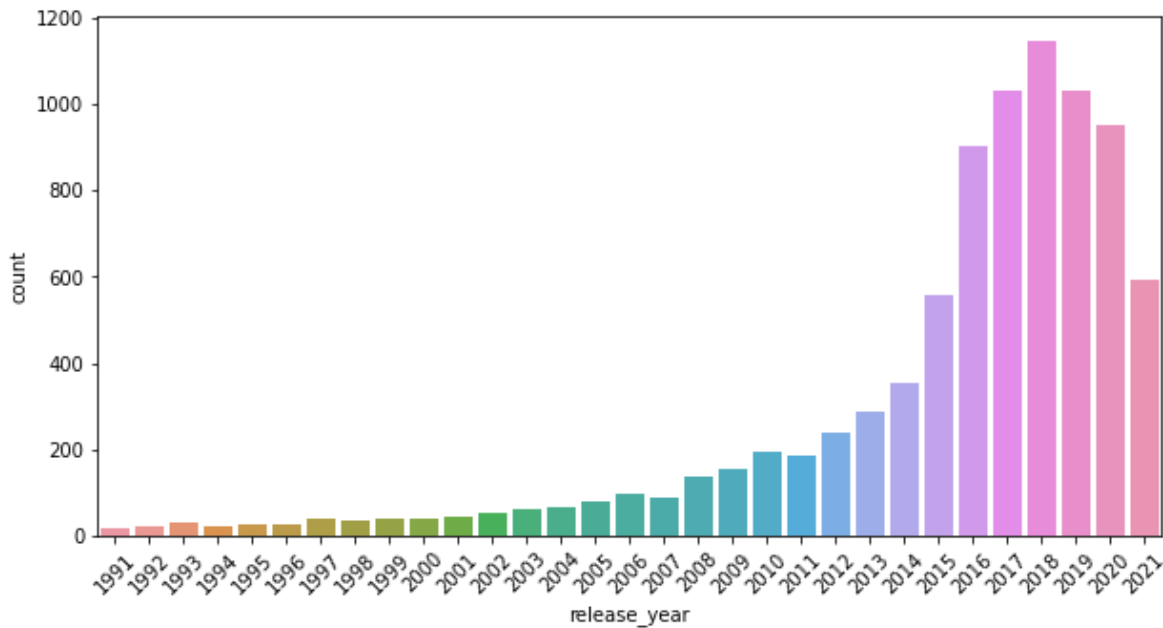

In [102]:

```python
# According to the graph the number of releases has increased exponentially as compared to
# Check both above and below for a proper movies vs tv shows comparision
```

In [108]:

```python
# graph 2 : Count of Movies and TV show side by side
plt.figure(figsize=(11,6))
graph2=sns.countplot(x="release_year",data=years,hue="type")
graph2.set_xticklabels(years["release_year"].value_counts().sort_index().index,rotation=45)
for i in graph2.containers:
    graph2.bar_label(i)
plt.show()
```


In [ ]:
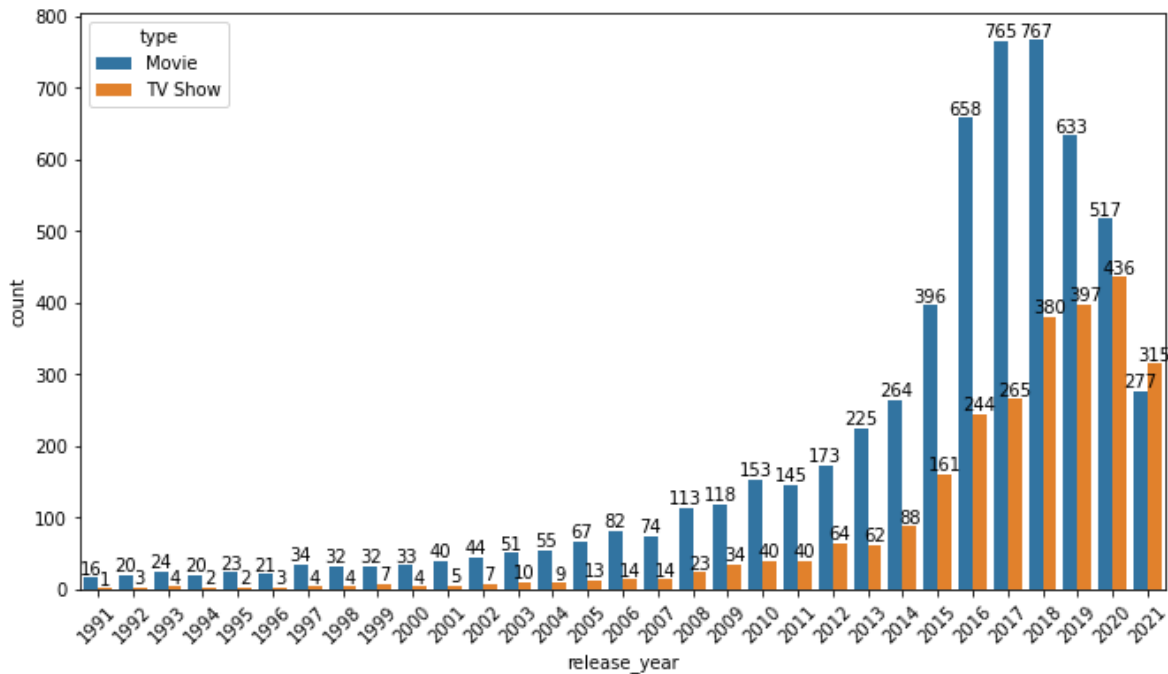
```python
# Count of TV shows and Movies acc to the year and release
"""No of movies and shows releasing as compared to 20-30 years , movies has increased by 7
1)In year 2021 more focus was into TV shows than Movies where as the avg is complete vice v
```

In [106]:

```python
# Top 18 countries producing the most content for netflix
```

In [107]:

```python
d.country.value_counts()[:18]
```

Out[107]:

```
United States     2815
India              972
Missing            830
United Kingdom     419
Japan              244
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Egypt              106
Turkey             105
Nigeria             95
Australia           86
Taiwan              81
Indonesia           79
Brazil              77
Philippines         75
Name: country, dtype: int64
```

In [113]:

```python
# graph3 : country wise distribution of netflix shows(production)
plt.figure(figsize=(10,5))
graph3=sns.barplot(x=d.country.value_counts()[:18].index,y=d.country.value_counts()[:18].va
graph3.set_xticklabels(d.country.value_counts()[:18].index,rotation=45)
# show value above each graph
for i in graph3.containers:
    graph3.bar_label(i)
plt.show()
```
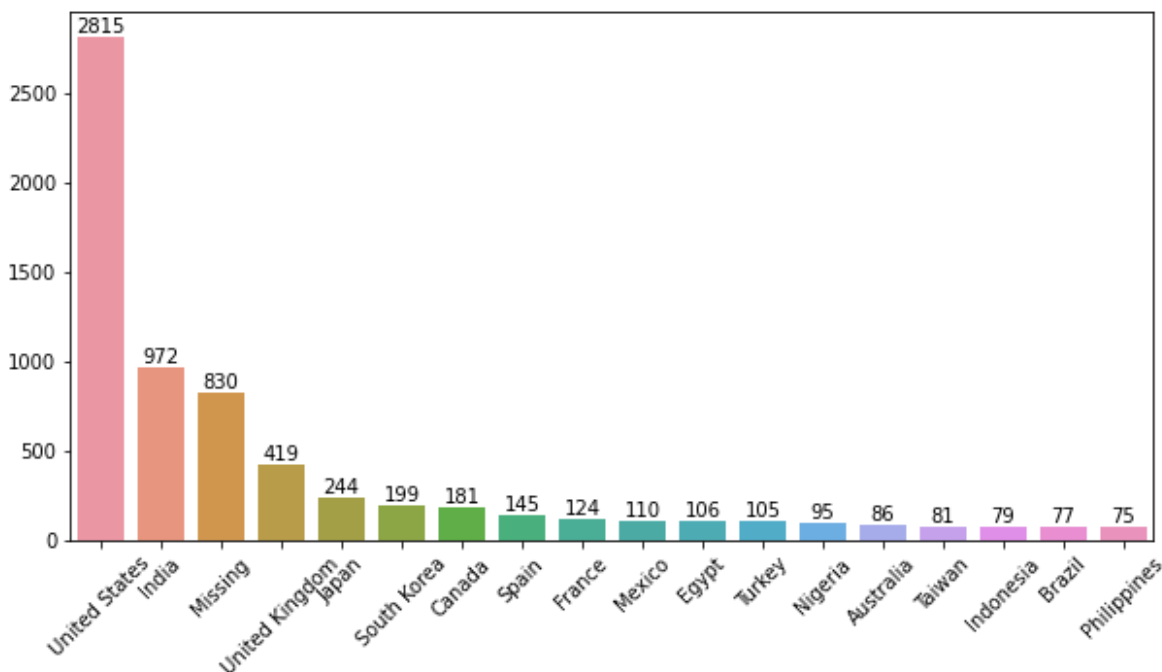
In [119]:

```python
# Best time to launch a TV show / Movie
```

In [144]:

```python
# finding the day in the week where most shows/movies have been released
```

In [231]:

```python
newdf["date_added"]=pd.to_datetime(newdf["date_added"])
newdf["day_of_week"]=newdf["date_added"].dt.dayofweek
```

In [147]:

```python
# Convert day_of_week into day "text"
```

In [232]:

```python
newdf.isnull().sum()
```

Out[232]:

```
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
day_of_week      10
dtype: int64
```

In [233]:

```python
# Function to convert day to text using dict
def convertday(x):
    if x==-1:
        return None
    ans={0:"Mon",1:"Tue",2:"Wed",3:"Thu",4:"Fri",5:"Sat",6:"Sun"}
    return ans[int(x)]
```

In [234]:

```python
newdf['day_of_week'] = newdf['day_of_week'].fillna(-1)
```

In [235]:

```python
newdf["day_of_week"]=newdf["day_of_week"].apply(convertday)
```
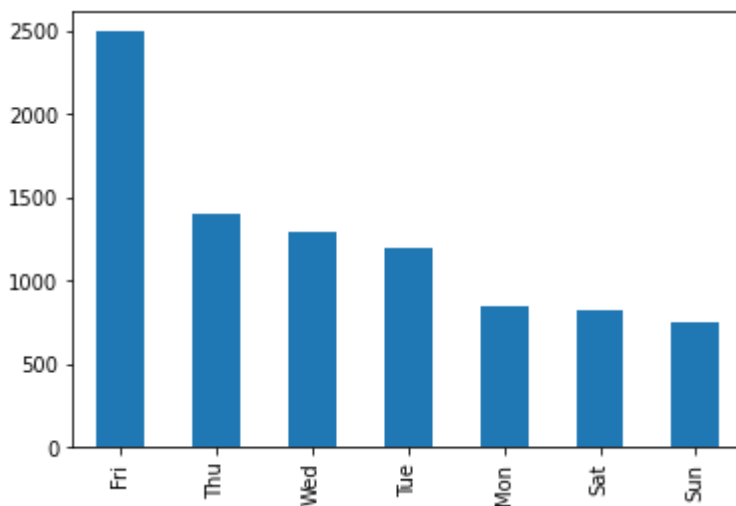
In [236]:

```python
newdf["day_of_week"].value_counts()
```

Out[236]:

```
Fri    2498
Thu    1396
Wed    1288
Tue    1197
Mon     851
Sat     816
Sun     751
Name: day_of_week, dtype: int64
```

In [237]:

```python
# Graph4 : Day in which most content is released in a week
newdf["day_of_week"].value_counts().plot(kind="bar")
plt.show()
```


In [168]:

```
s per the bar graph4 we can easliy say that most of the content that is released is on Frida
an binge watch the content :-D
```

In [238]:

```python
newdf.isnull().sum()
```

Out[238]:

```
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
day_of_week      10
dtype: int64
```

In [239]:

```python
newdf["date_added"]=pd.to_datetime(newdf["date_added"])
newdf["month_added"]=newdf["date_added"].dt.month
```

In [240]:

```python
# Function to convert month to text using dict
def convertmonth(x):
    if x==-1:
        return None
    ans={1:"Jan",2:"Feb",3:"Mar",4:"April",5:"May",6:"June",7:"July",8:"Aug",9:"Sept",10:"O
    return ans[int(x)]
```

In [241]:

```python
newdf['month_added'] = newdf['month_added'].fillna(-1)
newdf["month_added"]=newdf["month_added"].apply(convertmonth)
```

In [242]:

```python
newdf["month_added"].value_counts()
```

Out[242]:
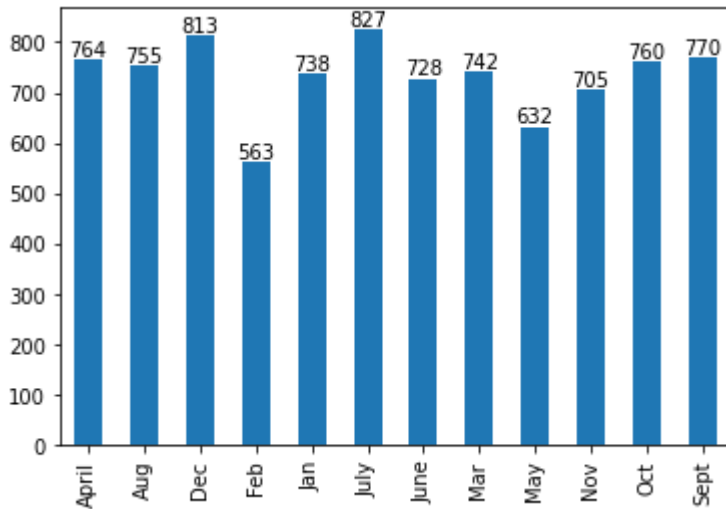
```
July      827
Dec       813
Sept      770
April     764
Oct       760
Aug       755
Mar       742
Jan       738
June      728
Nov       705
May       632
Feb       563
Name: month_added, dtype: int64
```

In [243]:

```python
# Graph5 : Month in which most content is released

graph5=newdf["month_added"].value_counts().sort_index().plot(kind="bar")
for i in graph5.containers:
    graph5.bar_label(i)

plt.show()
```


In [193]:

```python
# As per Graph5 we can see that "July" is the month where most content is released
```

In [204]:

```python
# Actor and Director Analysis
# Top 30 Actors and directors have the most movies released on netflix
```

In [246]:

```python
newdf1=newdf.copy(deep=True)
```

In [248]:

```python
newdf1['director'] = newdf1['director'].str.split(',')
```

In [267]:

```python
# graph6=Most directed movies and tvshows by a director
plt.figure(figsize=(10,6))
graph6=df_f1.explode('director')["director"].value_counts()[:30].plot(kind='bar')
graph6.set_xticklabels(df_f1.explode('director')["director"].value_counts()[:30].index,rota
for i in graph6.containers:
    graph6.bar_label(i)
plt.show()
```


In [ ]:
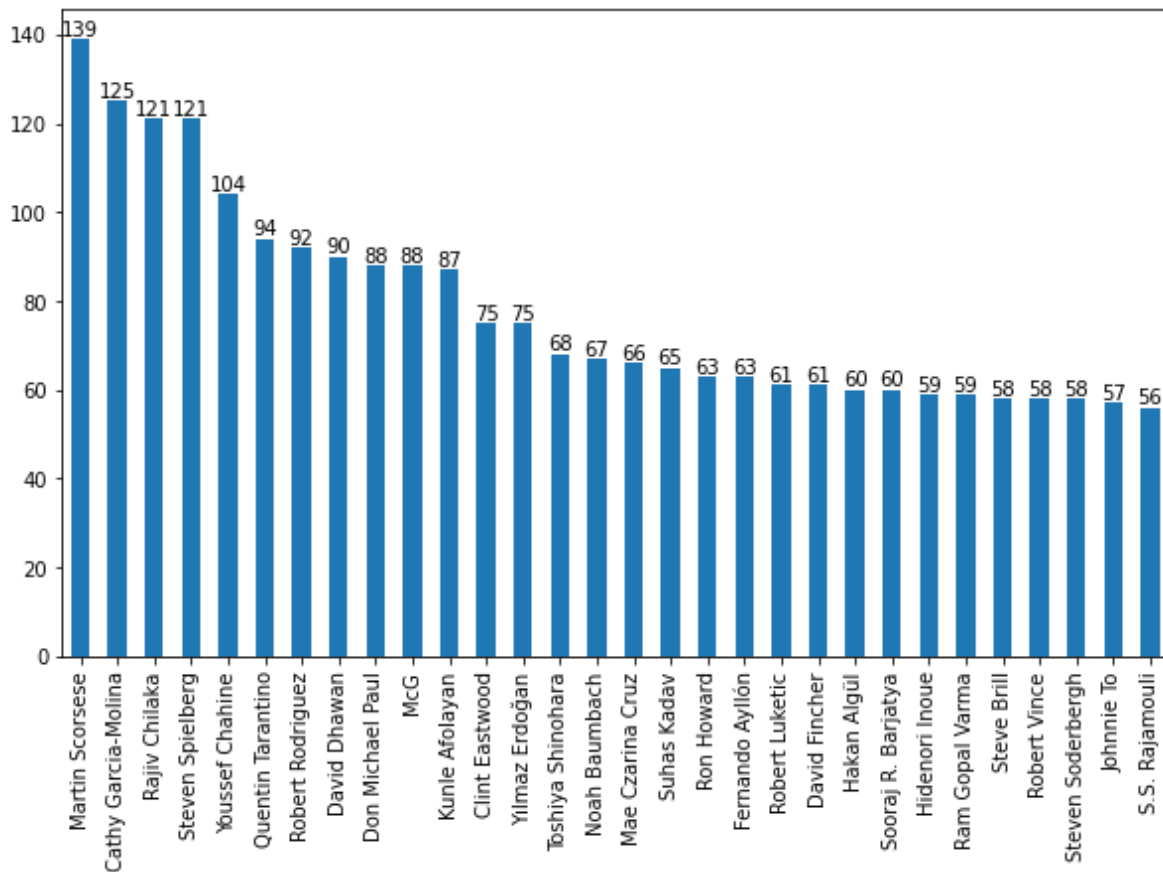
```python
# From above graph we can say that Martin Scorsese has the most directions
```

In [ ]:

```python
# Top 30 Actors
```

In [301]:

```python
# graph7=Most viewed Actor/Actress movies and tvshows
plt.figure(figsize=(10,6))
graph7=df_f1.explode('Actors')["Actors"].value_counts()[1:31].plot(kind='bar')
graph7.set_xticklabels(df_f1.explode('Actors')["Actors"].value_counts()[1:31].index,rotatio
for i in graph7.containers:
    graph7.bar_label(i)
plt.show()
```
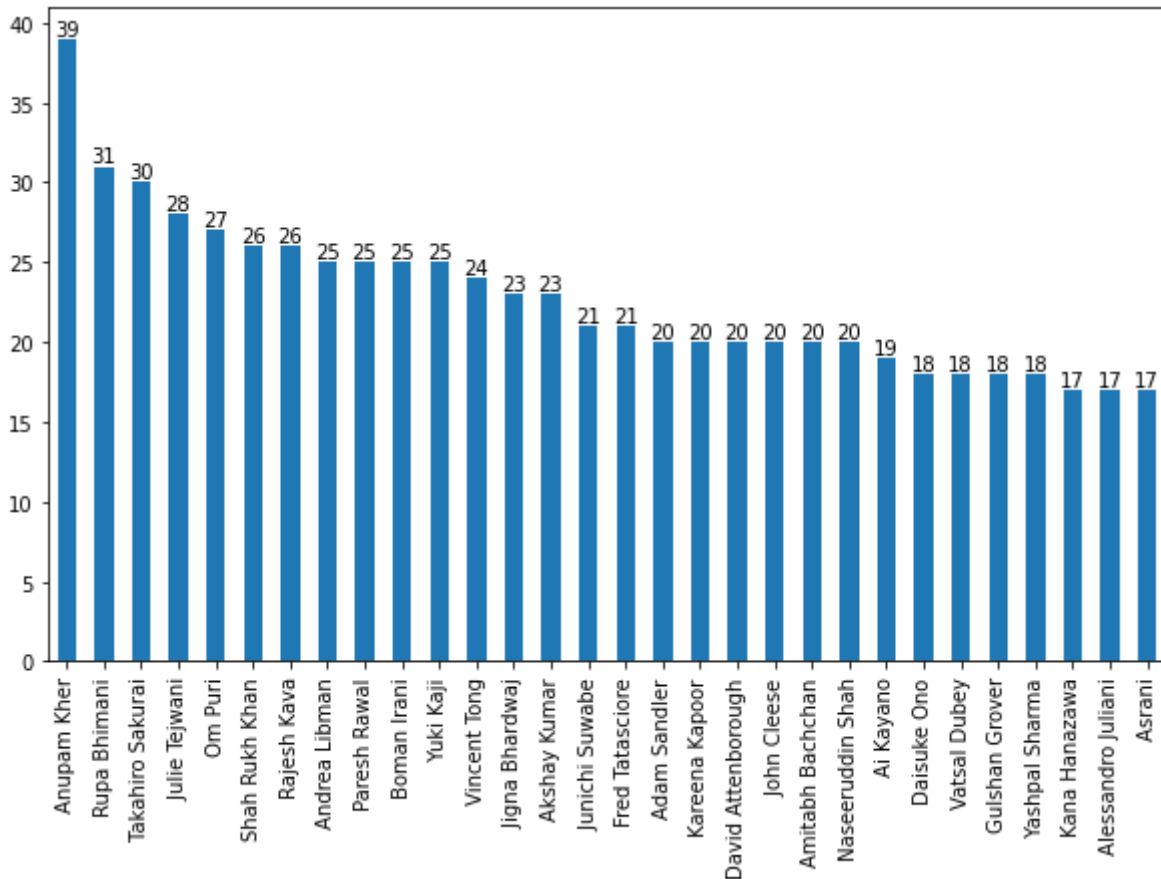

In [259]:

```python
df_f1.explode('Actors')["Actors"].value_counts()
```

Out[259]:

```
nan                   825
 Anupam Kher           39
 Rupa Bhimani          31
 Takahiro Sakurai      30
 Julie Tejwani         28
                      ...
 Vedika                 1
 Tedros Teclebrhan      1
 Maryam Zaree           1
 Melanie Straub         1
 Chittaranjan Tripathy  1
Name: Actors, Length: 39297, dtype: int64
```

In [284]:

```python
df_f1[["country","Category1","Category2","Category3"]].value_counts()
```

Out[284]:

```
country                                                        Category1
Category2            Category3
India                                                          Comedies
Dramas               International Movies       947
                                                               Dramas
Independent Movies   International Movies       787
United States                                                  Comedies
Dramas               Independent Movies        600
India                                                          Action & Adve
nture        Dramas              International Movies      475
South Korea                                                    International
TV Shows    Korean TV Shows       Romantic TV Shows       462

...
Australia                                                      Documentaries
International Movies   Music & Musicals              1

LGBTQ Movies                   1
South Africa                                                   International
TV Shows    TV Comedies          TV Dramas                1
United Kingdom, United States                                  British TV Sh
ows         Crime TV Shows       Docuseries               1
United Kingdom, United States, Spain, Germany, Greece, Canada  British TV Sh
ows         Docuseries           International TV Shows    1
Length: 1422, dtype: int64
```

In [286]:

```python
# Unnesting country
countrynew=newdf["country"].apply(lambda x:str(x).split(',')).tolist()
```

In [287]:

```python
countrynew
```

```
['Spain'],
['Turkey'],
['India'],
['India'],
['United States'],
['Turkey'],
['Turkey'],
['Turkey'],
['Turkey'],
['Turkey'],
['Turkey'],
['Spain'],
['Turkey'],
['Turkey'],
['Turkey'],
['nan'],
['Thailand'],
['Turkey'],
['Turkey'],
```

In [288]:

```python
newdf4=pd.DataFrame(countrynew,index=newdf["title"])
```

In [291]:

```
newdf4
```

Out[291]:

title	0	1	2	3	4	5	6	7	8	9	10	11
Dick Johnson Is Dead	United States	None	None	None	None	None	None	None	None	None	None	None
Blood & Water	South Africa	None	None	None	None	None	None	None	None	None	None	None
Ganglands	nan	None	None	None	None	None	None	None	None	None	None	None
Jailbirds New Orleans	nan	None	None	None	None	None	None	None	None	None	None	None
Kota Factory	India	None	None	None	None	None	None	None	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...	...	...
Zodiac	United States	None	None	None	None	None	None	None	None	None	None	None
Zombie Dumb	nan	None	None	None	None	None	None	None	None	None	None	None
Zombieland	United States	None	None	None	None	None	None	None	None	None	None	None
Zoom	United States	None	None	None	None	None	None	None	None	None	None	None
Zubaan	India	None	None	None	None	None	None	None	None	None	None	None

8807 rows × 12 columns

In [292]:

```
df_f2=df_f1.merge(newdf4,on="title",how="left")
```

In [293]:

```
df_f2
```

Out[293]:

Actors	show_id	type	director	country	listed_in	release_year	date_added	rating	...
nan	s1	Movie	Kirsten Johnson	United States	Documentaries	2020	September 25, 2021	PG-13	...
Ama Qamata	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	September 24, 2021	TV-MA	...
Khosi Ngema	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	September 24, 2021	TV-MA	...
Gail abalane	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	September 24, 2021	TV-MA	...
habang Molaba	s2	TV Show	NaN	South Africa	International TV Shows, TV Dramas, TV Mysteries	2021	September 24, 2021	TV-MA	...
...	...	...	...	...	...	...	...	...	...
Manish audhary	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	March 2, 2019	TV-14	...
Meghna Malik	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	March 2, 2019	TV-14	...
Malkeet Rauni	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	March 2, 2019	TV-14	...
Anita habdish	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	March 2, 2019	TV-14	...
taranjan Tripathy	s8807	Movie	Mozez Singh	India	Dramas, International Movies, Music & Musicals	2015	March 2, 2019	TV-14	...

mns

In [294]:

```
newdf
```

Out[294]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA
..	...	...	...	...	...	...	...	...	...
2	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	R
3	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018	TV-Y7
4	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	R

	show_id	type	title	director	cast	country	date_added	release_year	rating
5	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006	PG
6	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-14

7 rows × 14 columns

In [295]:

```python
newdf['listed_in']=newdf['listed_in'].apply(lambda x : x.split(','))
```

In [303]:

```python
newdf['listed_in'].explode().value_counts()[:30]
```
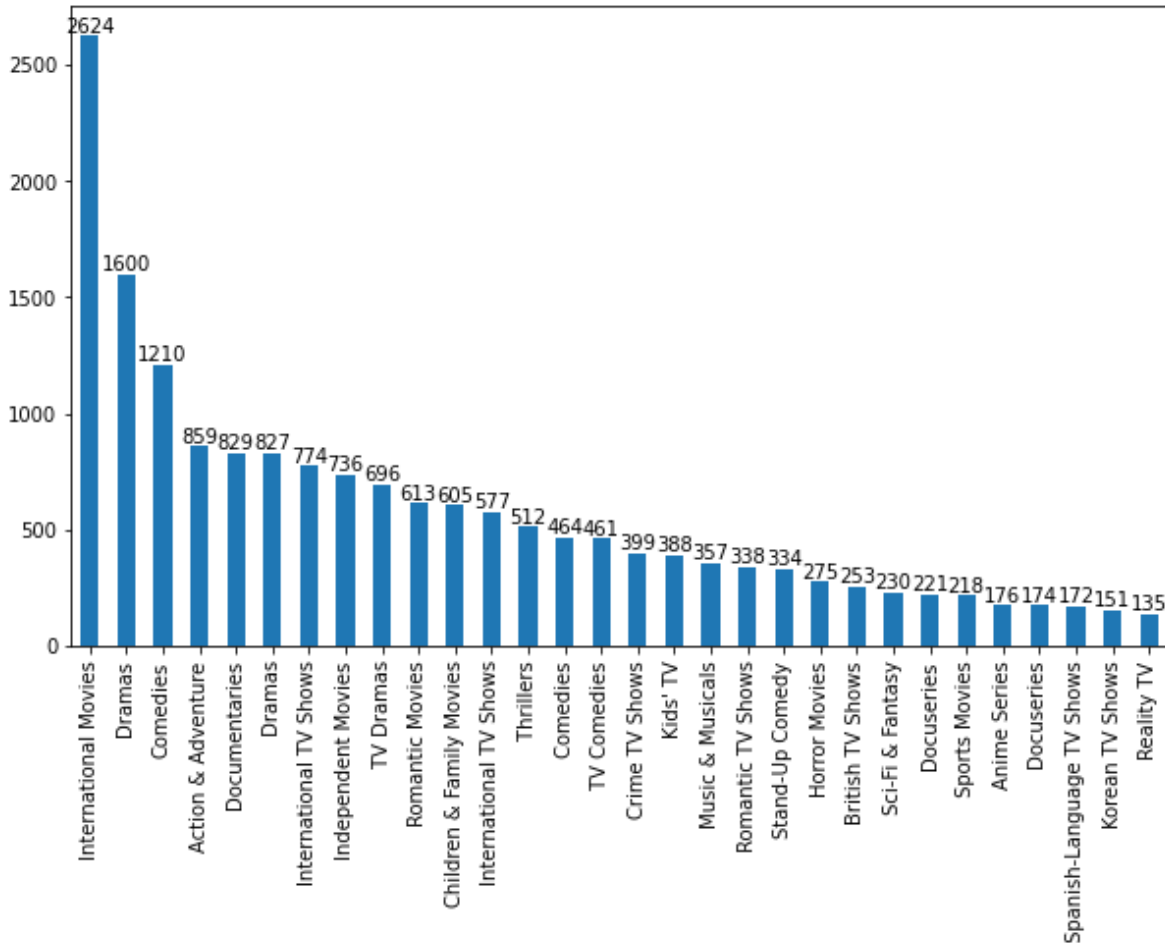
Out[303]:

30

In [ ]:

```python
# Top 30 frequent genres in Netflix
```

In [300]:

```python
# graph8=Most viewed genres movies and tvshows
plt.figure(figsize=(10,6))
graph8=newdf['listed_in'].explode().value_counts()[:30].plot(kind='bar')
graph8.set_xticklabels(newdf['listed_in'].explode().value_counts()[:30].index,rotation=90)
for i in graph8.containers:
    graph8.bar_label(i)
plt.show()
```


In [ ]: