

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import wget
```

```
wget.download('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv')
```

```
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 l...
...
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...

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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]: `df.describe(include='all')`

Out[5]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8807.000000	8803	8804	8807	8807
unique	8807	2	8807	4528	7692	748	1767	NaN	17	220	514	8775
top	s7291	Movie	Mauli	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	NaN	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	1	19	19	2818	109	NaN	3207	1793	362	4
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014.180198	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.819312	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	NaN	NaN	NaN	NaN

In [6]: *# Here in this Netflix dataset the description column is of no use for to derive the insights,hence I am dropping the co*
`df.drop(['description'], axis=1, inplace=True)`

In [7]: *# we saw in the info of the data that the dtype of the date_added is object hence changing it to date_time*
`df['date_added'] = pd.to_datetime(df['date_added'])`

In [8]: df

Out[8]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries
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
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...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
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 ...
...
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
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	R	88 min	Comedies, Horror Movies
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
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

8807 rows × 11 columns

In [9]: df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 11 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
dtypes: datetime64[ns](1), int64(1), object(9)
memory usage: 757.0+ KB
```

```
In [10]: df.head(50)
```

```
Out[10]:
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mababane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries
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...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
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 [11]: df['rating'].value_counts()
```

```
Out[11]: 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
84 min     1
66 min     1
74 min     1
Name: rating, dtype: int64
```

```
In [12]: df[df['rating'].isin(['84 min', '74 min', '66 min'])]
```

```
Out[12]:
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	74 min	NaN	Movies
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	84 min	NaN	Movies
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	2016-08-15	2015	66 min	NaN	Movies

```
In [13]: df.loc[[5541,5794,5813],['duration']] = ['74 min', '84 min', '66 min']
```

```
In [14]: df.loc[[5541,5794,5813],['rating']] = np.nan
```



```
In [15]: df.loc[[5541,5794,5813]] ## Replaced the values of rating and duration with each other to balance the columns
```

```
Out[15]:
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	NaN	74 min	Movies
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	NaN	84 min	Movies
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	2016-08-15	2015	NaN	66 min	Movies

In [16]: df

Out[16]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries
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
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...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV
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 ...
...
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
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	R	88 min	Comedies, Horror Movies
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
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

8807 rows × 11 columns

```
In [17]: df['title'].nunique() ##
```

```
Out[17]: 8807
```

Dealing with nested data in director, cast, country and listed_in columns

```
In [18]: '''Here we have some nested data in the director column,  
so we will perform split and stackk functionalities on this column to deal with the nested data'''  
  
director_constraint = df['director'].apply(lambda x: str(x).split(', ')).to_list()
```

```
In [19]: df_new_director = pd.DataFrame(director_constraint, index=df['title'])
```

```
In [20]: df_new_director.head(50)
```

```
Out[20]:
```

	0	1	2	3	4	5	6	7	8	9	10	11	12
title													
Dick Johnson Is Dead	Kirsten Johnson	None	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	None
Ganglands	Julien Leclercq	None	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	None
Kota Factory	nan	None	None	None	None	None	None	None	None	None	None	None	None
Midnight Mass	Mike Flanagan	None	None	None	None	None	None	None	None	None	None	None	None
My Little Pony: A New Generation	Robert Cullen	José Luis Ucha	None	None	None	None	None	None	None	None	None	None	None
Sankofa	Haile Gerima	None	None	None	None	None	None	None	None	None	None	None	None
The Great British Baking Show	Andy Devonshire	None	None	None	None	None	None	None	None	None	None	None	None
The Starling	Theodore Melfi	None	None	None	None	None	None	None	None	None	None	None	None
Vendetta: Truth, Lies and The Mafia	nan	None	None	None	None	None	None	None	None	None	None	None	None
Bangkok Breaking	Kongkiat Komesiri	None	None	None	None	None	None	None	None	None	None	None	None
Je Suis Karl	Christian Schwochow	None	None	None	None	None	None	None	None	None	None	None	None
Confessions of an Invisible Girl	Bruno Garotti	None	None	None	None	None	None	None	None	None	None	None	None
Crime Stories: India Detectives	nan	None	None	None	None	None	None	None	None	None	None	None	None
Dear White People	nan	None	None	None	None	None	None	None	None	None	None	None	None
Europe's Most Dangerous Man: Otto Skorzeny in Spain	Pedro de Echave García	Pablo Azorín Williams	None	None	None	None	None	None	None	None	None	None	None
Falsa identidad	nan	None	None	None	None	None	None	None	None	None	None	None	None
Intrusion	Adam Salky	None	None	None	None	None	None	None	None	None	None	None	None
Jaguar	nan	None	None	None	None	None	None	None	None	None	None	None	None

		0	1	2	3	4	5	6	7	8	9	10	11	12
title														
Monsters Inside: The 24 Faces of Billy Milligan	Olivier Megaton		None	None	None	None	None	None	None	None	None	None	None	None
Resurrection: Ertugrul	nan		None	None	None	None	None	None	None	None	None	None	None	None
Avvai Shanmughi	K.S. Ravikumar		None	None	None	None	None	None	None	None	None	None	None	None
Go! Go! Cory Carson: Chrissy Takes the Wheel	Alex Woo	Stanley Moore		None	None	None	None	None	None	None	None	None	None	None
Jeans	S. Shankar		None	None	None	None	None	None	None	None	None	None	None	None
Love on the Spectrum	nan		None	None	None	None	None	None	None	None	None	None	None	None
Minsara Kanavu	Rajiv Menon		None	None	None	None	None	None	None	None	None	None	None	None
Grown Ups	Dennis Dugan		None	None	None	None	None	None	None	None	None	None	None	None
Dark Skies	Scott Stewart		None	None	None	None	None	None	None	None	None	None	None	None
Paranoia	Robert Luketic		None	None	None	None	None	None	None	None	None	None	None	None
Ankahi Kahaniya	Ashwiny Iyer Tiwari	Abhishek Chaubey	Saket Chaudhary	None	None	None	None	None	None	None	None	None	None	None
Chicago Party Aunt	nan		None	None	None	None	None	None	None	None	None	None	None	None
Sex Education	nan		None	None	None	None	None	None	None	None	None	None	None	None
Squid Game	nan		None	None	None	None	None	None	None	None	None	None	None	None
Tayo and Little Wizards	nan		None	None	None	None	None	None	None	None	None	None	None	None
The Father Who Moves Mountains	Daniel Sandu		None	None	None	None	None	None	None	None	None	None	None	None
The Stronghold	Cédric Jimenez		None	None	None	None	None	None	None	None	None	None	None	None
Angry Birds	nan		None	None	None	None	None	None	None	None	None	None	None	None
Birth of the Dragon	George Nolfi		None	None	None	None	None	None	None	None	None	None	None	None
Chhota Bheem	nan		None	None	None	None	None	None	None	None	None	None	None	None
He-Man and the Masters of the Universe	nan		None	None	None	None	None	None	None	None	None	None	None	None
Jaws	Steven Spielberg		None	None	None	None	None	None	None	None	None	None	None	None
Jaws 2	Jeannot Szwarc		None	None	None	None	None	None	None	None	None	None	None	None

	0	1	2	3	4	5	6	7	8	9	10	11	12
title													
Jaws 3	Joe Alves	None	None	None	None	None	None	None	None	None	None	None	None
Jaws: The Revenge	Joseph Sargent	None	None	None	None	None	None	None	None	None	None	None	None
My Heroes Were Cowboys	Tyler Greco	None	None	None	None	None	None	None	None	None	None	None	None
Safe House	Daniel Espinosa	None	None	None	None	None	None	None	None	None	None	None	None
The Smart Money Woman	Bunmi Ajakaiye	None	None	None	None	None	None	None	None	None	None	None	None
Training Day	Antoine Fuqua	None	None	None	None	None	None	None	None	None	None	None	None
Castle and Castle	nan	None	None	None	None	None	None	None	None	None	None	None	None

```
In [21]: df_new_director = df_new_director.stack()
```

```
In [22]: df_new_director
```

```
Out[22]: title
Dick Johnson Is Dead    0    Kirsten Johnson
Blood & Water          0             nan
Ganglands              0    Julien Leclercq
Jailbirds New Orleans  0             nan
Kota Factory           0             nan
...
Zodiac                0    David Fincher
Zombie Dumb           0             nan
Zombieland            0    Ruben Fleischer
Zoom                 0    Peter Hewitt
Zubaan               0    Mozez Singh
Length: 9612, dtype: object
```

```
In [23]: df_new_director = pd.DataFrame(df_new_director)
```

```
In [24]: df_new_director.reset_index(inplace=True)
```

```
In [25]: df_new_director = df_new_director[['title', 0]]
```

```
In [26]: df_new_director.columns = ['title', 'director']
```

```
In [27]: df_new_director
```

Out[27]:

	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
...
9607	Zodiac	David Fincher
9608	Zombie Dumb	nan
9609	Zombieland	Ruben Fleischer
9610	Zoom	Peter Hewitt
9611	Zubaan	Mozez Singh

9612 rows × 2 columns

```
In [28]: df_new_director[df_new_director['director'] == 'Steven Spielberg'] ## you can see here we have dealt with the nest data
```

Out[28]:

	title	director
46	Jaws	Steven Spielberg
361	Catch Me If You Can	Steven Spielberg
1303	The BFG	Steven Spielberg
7663	Indiana Jones and the Kingdom of the Crystal S...	Steven Spielberg
7664	Indiana Jones and the Last Crusade	Steven Spielberg
7665	Indiana Jones and the Raiders of the Lost Ark	Steven Spielberg
7666	Indiana Jones and the Temple of Doom	Steven Spielberg
7930	Lincoln	Steven Spielberg
8655	Schindler's List	Steven Spielberg
8912	The Adventures of Tintin	Steven Spielberg
9491	War Horse	Steven Spielberg

As the above steps were performed on the director column to deal with the nested data, same steps will be performed on the cast, country and list_in columns.

Cast nested Data

```
In [29]: cast_constraint = df['cast'].apply(lambda x: str(x).split(', ')).to_list()
```

```
In [30]: df_new_cast = pd.DataFrame(cast_constraint, index=df['title'])
```


In [31]: `df_new_cast.head(20)`

Out[31]:

	0	1	2	3	4	5	6	7	8	9	...
title											
Dick Johnson Is Dead	nan	None	None	None	None	None	None	None	None	None	...
Blood & Water	Ama Qamata	Khosi Ngema	Gail Mabalane	Thabang Molaba	Dillon Windvogel	Natasha Thahane	Arno Greeff	Xolile Tshabalala	Getmore Sithole	Cindy Mahlangu	...
Ganglands	Sami Bouajila	Tracy Gotoas	Samuel Jouy	Nabiha Akkari	Sofia Lesaffre	Salim Kechiouche	Noureddine Farihi	Geert Van Rampelberg	Bakary Diombero	None	...
Jailbirds New Orleans	nan	None	None	None	None	None	None	None	None	None	...
Kota Factory	Mayur More	Jitendra Kumar	Ranjan Raj	Alam Khan	Ahsaas Channa	Revathi Pillai	Urvi Singh	Arun Kumar	None	None	...
Midnight Mass	Kate Siegel	Zach Gilford	Hamish Linklater	Henry Thomas	Kristin Lehman	Samantha Sloyan	Igby Rigney	Rahul Kohli	Annarah Cymone	Annabeth Gish	...
My Little Pony: A New Generation	Vanessa Hudgens	Kimiko Glenn	James Marsden	Sofia Carson	Liza Koshy	Ken Jeong	Elizabeth Perkins	Jane Krakowski	Michael McKean	Phil LaMarr	...
Sankofa	Kofi Ghanaba	Oyafunmike Ogunlano	Alexandra Duah	Nick Medley	Mutabaruka	Afemo Omilami	Reggie Carter	Mzuri	None	None	...
The Great British Baking Show	Mel Giedroyc	Sue Perkins	Mary Berry	Paul Hollywood	None	None	None	None	None	None	...
The Starling	Melissa McCarthy	Chris O'Dowd	Kevin Kline	Timothy Olyphant	Daveed Diggs	Skyler Gisondo	Laura Harrier	Rosalind Chao	Kimberly Quinn	Loretta Devine	...
Vendetta: Truth, Lies and The Mafia	nan	None	None	None	None	None	None	None	None	None	...
Bangkok Breaking	Sukollawat Kanarot	Sushar Manaying	Pavarit Mongkolpisit	Sahajak Boonthanakit	Suthipongse Thatphithakkul	Bhasaworn Bawronkirati	Daweerit Chullasapya	Warattaya Wongchayaporn	Kittiphoom Wongpentak	Abhicha Thanachanun	...

	0	1	2	3	4	5	6	7	8	9	...
title											
Je Suis Karl	Luna Wedler	Jannis Niewöhner	Milan Peschel	Edin Hasanović	Anna Fialová	Marlon Boess	Victor Boccard	Fleur Geffrier	Aziz Dyab	Mélanie Fouché	...
Confessions of an Invisible Girl	Klara Castanho	Lucca Picon	Júlia Gomes	Marcus Bessa	Kiria Malheiros	Fernanda Concon	Gabriel Lima	Caio Cabral	Leonardo Cidade	Jade Cardozo	...
Crime Stories: India Detectives	nan	None	None	None	None	None	None	None	None	None	...
Dear White People	Logan Browning	Brandon P. Bell	DeRon Horton	Antoinette Robertson	John Patrick Amedori	Ashley Blaine Featherson	Marque Richardson	Giancarlo Esposito	None	None	...
Europe's Most Dangerous Man: Otto Skorzeny in Spain	nan	None	None	None	None	None	None	None	None	None	...
Falsa identidad	Luis Ernesto Franco	Camila Sodi	Sergio Goyri	Samadhi Zendejas	Eduardo Yáñez	Sonya Smith	Alejandro Camacho	Azela Robinson	Uriel del Toro	Géraldine Bazán	...
Intrusion	Freida Pinto	Logan Marshall-Green	Robert John Burke	Megan Elisabeth Kelly	Sarah Minnich	Hayes Hargrove	Mark Sivertsen	Brandon Fierro	Antonio Valles	Clint Obenchain	...
Jaguar	Blanca Suárez	Iván Marcos	Óscar Casas	Adrián Lastra	Francesc Garrido	Stefan Weinert	Julia Möller	Alicia Chojnowski	None	None	...

20 rows × 50 columns



```
In [32]: df_new_cast = df_new_cast.stack()
```

```
In [33]: df_new_cast = pd.DataFrame(df_new_cast)
```

```
In [34]: df_new_cast.reset_index(inplace=True)
```

```
In [35]: df_new_cast = df_new_cast[['title', 0]]
```

```
In [36]: df_new_cast.columns = ['title', 'cast']
```

```
In [37]: df_new_cast
```

Out[37]:

	title	cast
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

Country nested Data

```
In [38]: country_constraint = df['country'].apply(lambda x: str(x).split(', ')).to_list()
```

```
In [39]: df_new_country = pd.DataFrame(country_constraint, index=df['title'])
```

```
In [40]: df_new_country = df_new_country.stack()
```

```
In [41]: df_new_country = pd.DataFrame(df_new_country)
```

```
In [42]: df_new_country.reset_index(inplace=True)
```

```
In [43]: df_new_country = df_new_country[['title', 0]]
```

```
In [44]: df_new_country.columns = ['title', 'country']
```

```
In [45]: df_new_country
```

```
Out[45]:
```

	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
...
10840	Zodiac	United States
10841	Zombie Dumb	nan
10842	Zombieland	United States
10843	Zoom	United States
10844	Zubaan	India

10845 rows × 2 columns

```
In [46]: df_new_country[df_new_country['country'] == 'United States']
```

Out[46]:

	title	country
0	Dick Johnson Is Dead	United States
7	Sankofa	United States
14	The Starling	United States
21	Dear White People	United States
33	Grown Ups	United States
...
10825	Yours, Mine and Ours	United States
10831	Zak Storm	United States
10840	Zodiac	United States
10842	Zombieland	United States
10843	Zoom	United States

3689 rows × 2 columns

listed_in nested Data

```
In [47]: listed_in_constraint = df['listed_in'].apply(lambda x: str(x).split(', ')).to_list()
```

```
In [48]: df_new_listed_in = pd.DataFrame(listed_in_constraint, index=df['title'])
```

```
In [49]: df_new_listed_in = df_new_listed_in.stack()
```

```
In [50]: df_new_listed_in = pd.DataFrame(df_new_listed_in)
```

```
In [51]: df_new_listed_in.reset_index(inplace=True)
```

```
In [52]: df_new_listed_in = df_new_listed_in[['title', 0]]
```

```
In [53]: df_new_listed_in.columns = ['title', 'genre']
```

```
In [54]: df_new_listed_in
```

Out[54]:

	title	genre
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Blood & Water	TV Dramas
3	Blood & Water	TV Mysteries
4	Ganglands	Crime TV Shows
...
19318	Zoom	Children & Family Movies
19319	Zoom	Comedies
19320	Zubaan	Dramas
19321	Zubaan	International Movies
19322	Zubaan	Music & Musicals

19323 rows × 2 columns

```
In [55]: df_new_listed_in[df_new_listed_in['genre'] == 'International TV Shows']
```

Out[55]:

	title	genre
1	Blood & Water	International TV Shows
5	Ganglands	International TV Shows
9	Kota Factory	International TV Shows
25	Vendetta: Truth, Lies and The Mafia	International TV Shows
27	Bangkok Breaking	International TV Shows
...
19164	Who's the One	International TV Shows
19192	Winter Sun	International TV Shows
19251	Yeh Meri Family	International TV Shows
19296	Yunus Emre	International TV Shows
19304	Zindagi Gulzar Hai	International TV Shows

1351 rows × 2 columns

Here after dealing with nested data we have got 4 new mini DataFrames (df_new_director, df_new_cast, df_new_country, df_new_listed_in) which will be merged now

Now will be merging this DataFrames one by one

```
In [56]: df_new = pd.merge(df_new_director, df_new_cast, on='title')
```

In [57]: df_new

Out[57]:

	title	director	cast
0	Dick Johnson Is Dead	Kirsten Johnson	nan
1	Blood & Water	nan	Ama Qamata
2	Blood & Water	nan	Khosi Ngema
3	Blood & Water	nan	Gail Mabalane
4	Blood & Water	nan	Thabang Molaba
...
70807	Zubaan	Mozez Singh	Manish Chaudhary
70808	Zubaan	Mozez Singh	Meghna Malik
70809	Zubaan	Mozez Singh	Malkeet Rauni
70810	Zubaan	Mozez Singh	Anita Shabdish
70811	Zubaan	Mozez Singh	Chittaranjan Tripathy

70812 rows × 3 columns

In [58]: df_new2 = pd.merge(df_new, df_new_country, on='title')

In [59]: df_clean = pd.merge(df_new2, df_new_listed_in, on='title')


```
In [60]: df_clean
```

```
Out[60]:
```

	title	director	cast	country	genre
0	Dick Johnson Is Dead	Kirsten Johnson	nan	United States	Documentaries
1	Blood & Water	nan	Ama Qamata	South Africa	International TV Shows
2	Blood & Water	nan	Ama Qamata	South Africa	TV Dramas
3	Blood & Water	nan	Ama Qamata	South Africa	TV Mysteries
4	Blood & Water	nan	Khosi Ngema	South Africa	International TV Shows
...
201986	Zubaan	Mozez Singh	Anita Shabdish	India	International Movies
201987	Zubaan	Mozez Singh	Anita Shabdish	India	Music & Musicals
201988	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Dramas
201989	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	International Movies
201990	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Music & Musicals

201991 rows × 5 columns

merging this clean dataframe with the original dataframe

```
In [61]: df = pd.merge(df, df_clean, on='title')
```

In [62]: `df.head()`

Out[62]:

	show_id	type	title	director_x	cast_x	country_x	date_added	release_year	rating	duration	listed_in	director_y	cast_y	country_y
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries	Kirsten Johnson	nan	United States
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	nan	Ama Qamata	South Africa
2	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	nan	Ama Qamata	South Africa
3	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	nan	Ama Qamata	South Africa
4	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thabane...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	nan	Khosi Ngema	South Africa

In [63]: `df.drop(['director_x', 'cast_x', 'country_x', 'listed_in'], axis=1, inplace=True)`

```
In [64]: df.rename(columns= {'director_y': 'director', 'cast_y': 'cast', 'country_y': 'country'}, inplace=True)
```

```
In [65]: df.head()
```

```
Out[65]:
```

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

```
In [66]: df = df.reindex(columns=['show_id', 'type', 'title', 'genre', 'director', 'cast', 'country', 'rating', 'date_added', 're
```

```
In [67]: df.head()
```

```
Out[67]:
```

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

Got the orinal data set without nested data, now operating on NaN values

```
In [68]: df.replace('nan', np.nan, inplace=True) ## as nan value was string so replacing it with np.nan
```

```
In [69]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 201991 entries, 0 to 201990
Data columns (total 11 columns):
#   Column          Non-Null Count  Dtype
---  -
0   show_id         201991 non-null object
1   type            201991 non-null object
2   title           201991 non-null object
3   genre           201991 non-null object
4   director        151348 non-null object
5   cast            199845 non-null object
6   country         190094 non-null object
7   rating          201921 non-null object
8   date_added      201833 non-null datetime64[ns]
9   release_year    201991 non-null int64
10  duration         201991 non-null object
dtypes: datetime64[ns](1), int64(1), object(9)
memory usage: 18.5+ MB
```

```
In [70]: df[df.columns].isnull().sum()
```

```
Out[70]: show_id         0
type             0
title            0
genre            0
director         50643
cast             2146
country          11897
rating           70
date_added       158
release_year     0
duration         0
dtype: int64
```

```
In [71]: print('Percentage of NaN Values if present in the columns')
print('- '*50)
for i in df.columns:
    if df[i].isnull().sum() > 0:
        print(f"{i}: \t {round((df[i].isnull().sum() / len(df[i])) * 100, 2)}%")
```

Percentage of NaN Values if present in the columns

```
-----
director:      25.07%
cast:          1.06%
country:       5.89%
rating:        0.03%
date_added:    0.08%
```

```
In [72]: # as cast, rating and date_added columns have Nan values less than 3%, hence will be deleting the rows
df.dropna(axis=0, subset=['cast', 'rating', 'date_added'], inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 199617 entries, 1 to 201990
Data columns (total 11 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   show_id     199617 non-null  object
1   type        199617 non-null  object
2   title       199617 non-null  object
3   genre       199617 non-null  object
4   director    149930 non-null  object
5   cast        199617 non-null  object
6   country     188039 non-null  object
7   rating      199617 non-null  object
8   date_added  199617 non-null  datetime64[ns]
9   release_year 199617 non-null  int64
10  duration    199617 non-null  object
dtypes: datetime64[ns](1), int64(1), object(9)
memory usage: 18.3+ MB
```

```
In [73]: df['country'].value_counts()
```

```
Out[73]: United States    58520  
        India            22717  
        United Kingdom   12693  
        Japan            8575  
        France           8172  
        ...  
        Vatican City      3  
        Sri Lanka         2  
        Panama            2  
        Afghanistan       2  
        Nicaragua         1  
        Name: country, Length: 119, dtype: int64
```

In [74]: df

Out[74]:

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration
1	s2	TV Show	Blood & Water	International TV Shows	NaN	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
2	s2	TV Show	Blood & Water	TV Dramas	NaN	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
3	s2	TV Show	Blood & Water	TV Mysteries	NaN	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
4	s2	TV Show	Blood & Water	International TV Shows	NaN	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons
5	s2	TV Show	Blood & Water	TV Dramas	NaN	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons
...
201986	s8807	Movie	Zubaan	International Movies	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min
201987	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min
201988	s8807	Movie	Zubaan	Dramas	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min
201989	s8807	Movie	Zubaan	International Movies	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min
201990	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min

199617 rows × 11 columns

```
In [75]: # now as director and country column has Nan value between 3% - 30%, will replace with mode of the column
df['director'] = df['director'].fillna(df['director'].mode()[0])
df['country'] = df['country'].fillna(df['country'].mode()[0])
```

In [76]: df

Out[76]:

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration
1	s2	TV Show	Blood & Water	International TV Shows	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
2	s2	TV Show	Blood & Water	TV Dramas	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
3	s2	TV Show	Blood & Water	TV Mysteries	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
4	s2	TV Show	Blood & Water	International TV Shows	Martin Scorsese	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons
5	s2	TV Show	Blood & Water	TV Dramas	Martin Scorsese	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons
...
201986	s8807	Movie	Zubaan	International Movies	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min
201987	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min
201988	s8807	Movie	Zubaan	Dramas	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min
201989	s8807	Movie	Zubaan	International Movies	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min
201990	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min

199617 rows × 11 columns


```
In [77]: df.isna().sum()
```

```
Out[77]: show_id      0
         type        0
         title       0
         genre       0
         director    0
         cast        0
         country     0
         rating      0
         date_added  0
         release_year 0
         duration    0
         dtype: int64
```

Now as all the null values are handled we will deal with duplicate rows

In [78]: *## Checking for the duplicate rows*

```
df[df.duplicated()]
```

Out[78]:

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration
39336	s1632	Movie	Rust Creek	Thrillers	Jen McGowan	Micah Hauptman	United States	R	2020-11-30	2018	108 min
88516	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Oscar Martínez	Argentina	TV-MA	2019-06-21	2019	113 min
88517	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Oscar Martínez	Argentina	TV-MA	2019-06-21	2019	113 min
88518	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Oscar Martínez	Argentina	TV-MA	2019-06-21	2019	113 min
88519	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Oscar Martínez	United States	TV-MA	2019-06-21	2019	113 min
88520	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Oscar Martínez	United States	TV-MA	2019-06-21	2019	113 min
88521	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Oscar Martínez	United States	TV-MA	2019-06-21	2019	113 min
88522	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Dolores Fonzi	Argentina	TV-MA	2019-06-21	2019	113 min
88523	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Dolores Fonzi	Argentina	TV-MA	2019-06-21	2019	113 min
88524	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Dolores Fonzi	Argentina	TV-MA	2019-06-21	2019	113 min
88525	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Dolores Fonzi	United States	TV-MA	2019-06-21	2019	113 min
88526	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Dolores Fonzi	United States	TV-MA	2019-06-21	2019	113 min
88527	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Dolores Fonzi	United States	TV-MA	2019-06-21	2019	113 min
88528	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Diego Velázquez	Argentina	TV-MA	2019-06-21	2019	113 min
88529	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Diego Velázquez	Argentina	TV-MA	2019-06-21	2019	113 min

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration
88530	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Diego Velázquez	Argentina	TV-MA	2019-06-21	2019	113 min
88531	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Diego Velázquez	United States	TV-MA	2019-06-21	2019	113 min
88532	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Diego Velázquez	United States	TV-MA	2019-06-21	2019	113 min
88533	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Diego Velázquez	United States	TV-MA	2019-06-21	2019	113 min
88534	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Paulina Garcia	Argentina	TV-MA	2019-06-21	2019	113 min
88535	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Paulina Garcia	Argentina	TV-MA	2019-06-21	2019	113 min
88536	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Paulina Garcia	Argentina	TV-MA	2019-06-21	2019	113 min
88537	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Paulina Garcia	United States	TV-MA	2019-06-21	2019	113 min
88538	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Paulina Garcia	United States	TV-MA	2019-06-21	2019	113 min
88539	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Paulina Garcia	United States	TV-MA	2019-06-21	2019	113 min
88540	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Luis Gnecco	Argentina	TV-MA	2019-06-21	2019	113 min
88541	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Luis Gnecco	Argentina	TV-MA	2019-06-21	2019	113 min
88542	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Luis Gnecco	Argentina	TV-MA	2019-06-21	2019	113 min
88543	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Luis Gnecco	United States	TV-MA	2019-06-21	2019	113 min
88544	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Luis Gnecco	United States	TV-MA	2019-06-21	2019	113 min
88545	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Luis Gnecco	United States	TV-MA	2019-06-21	2019	113 min
88546	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Malena Sánchez	Argentina	TV-MA	2019-06-21	2019	113 min

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration
88547	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Malena Sánchez	Argentina	TV-MA	2019-06-21	2019	113 min
88548	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Malena Sánchez	Argentina	TV-MA	2019-06-21	2019	113 min
88549	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Malena Sánchez	United States	TV-MA	2019-06-21	2019	113 min
88550	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Malena Sánchez	United States	TV-MA	2019-06-21	2019	113 min
88551	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Malena Sánchez	United States	TV-MA	2019-06-21	2019	113 min
88552	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Emilio Vodanovich	Argentina	TV-MA	2019-06-21	2019	113 min
88553	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Emilio Vodanovich	Argentina	TV-MA	2019-06-21	2019	113 min
88554	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Emilio Vodanovich	Argentina	TV-MA	2019-06-21	2019	113 min
88555	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Emilio Vodanovich	United States	TV-MA	2019-06-21	2019	113 min
88556	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Emilio Vodanovich	United States	TV-MA	2019-06-21	2019	113 min
88557	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Emilio Vodanovich	United States	TV-MA	2019-06-21	2019	113 min
88558	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Norman Briski	Argentina	TV-MA	2019-06-21	2019	113 min
88559	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Norman Briski	Argentina	TV-MA	2019-06-21	2019	113 min
88560	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Norman Briski	Argentina	TV-MA	2019-06-21	2019	113 min
88561	s3719	Movie	Blood Will Tell	Dramas	Miguel Cohan	Norman Briski	United States	TV-MA	2019-06-21	2019	113 min
88562	s3719	Movie	Blood Will Tell	Independent Movies	Miguel Cohan	Norman Briski	United States	TV-MA	2019-06-21	2019	113 min
88563	s3719	Movie	Blood Will Tell	International Movies	Miguel Cohan	Norman Briski	United States	TV-MA	2019-06-21	2019	113 min

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration
135609	s6014	Movie	300 Miles to Heaven	Dramas	Maciej Dejczer	Adrianna Biedrzyńska	Denmark	TV-14	2019-10-01	1989	93 min
135610	s6014	Movie	300 Miles to Heaven	International Movies	Maciej Dejczer	Adrianna Biedrzyńska	Denmark	TV-14	2019-10-01	1989	93 min
135611	s6014	Movie	300 Miles to Heaven	Dramas	Maciej Dejczer	Adrianna Biedrzyńska	France	TV-14	2019-10-01	1989	93 min
135612	s6014	Movie	300 Miles to Heaven	International Movies	Maciej Dejczer	Adrianna Biedrzyńska	France	TV-14	2019-10-01	1989	93 min
135613	s6014	Movie	300 Miles to Heaven	Dramas	Maciej Dejczer	Adrianna Biedrzyńska	Poland	TV-14	2019-10-01	1989	93 min
135614	s6014	Movie	300 Miles to Heaven	International Movies	Maciej Dejczer	Adrianna Biedrzyńska	Poland	TV-14	2019-10-01	1989	93 min

```
In [79]: df.drop_duplicates(inplace=True)
```

In [80]: df

Out[80]:

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration
1	s2	TV Show	Blood & Water	International TV Shows	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
2	s2	TV Show	Blood & Water	TV Dramas	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
3	s2	TV Show	Blood & Water	TV Mysteries	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons
4	s2	TV Show	Blood & Water	International TV Shows	Martin Scorsese	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons
5	s2	TV Show	Blood & Water	TV Dramas	Martin Scorsese	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons
...
201986	s8807	Movie	Zubaan	International Movies	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min
201987	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min
201988	s8807	Movie	Zubaan	Dramas	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min
201989	s8807	Movie	Zubaan	International Movies	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min
201990	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min

199562 rows × 11 columns

In [81]: df['year_added'] = df['date_added'].dt.year.astype('int64')

In [82]: df['month_added'] = pd.to_datetime(df['date_added']).dt.month_name()

In [83]: df

Out[83]:

	show_id	type	title	genre	director	cast	country	rating	date_added	release_year	duration	year_added	month_added
1	s2	TV Show	Blood & Water	International TV Shows	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons	2021	September
2	s2	TV Show	Blood & Water	TV Dramas	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons	2021	September
3	s2	TV Show	Blood & Water	TV Mysteries	Martin Scorsese	Ama Qamata	South Africa	TV-MA	2021-09-24	2021	2 Seasons	2021	September
4	s2	TV Show	Blood & Water	International TV Shows	Martin Scorsese	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons	2021	September
5	s2	TV Show	Blood & Water	TV Dramas	Martin Scorsese	Khosi Ngema	South Africa	TV-MA	2021-09-24	2021	2 Seasons	2021	September
...
201986	s8807	Movie	Zubaan	International Movies	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min	2019	March
201987	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Anita Shabdish	India	TV-14	2019-03-02	2015	111 min	2019	March
201988	s8807	Movie	Zubaan	Dramas	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min	2019	March
201989	s8807	Movie	Zubaan	International Movies	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min	2019	March
201990	s8807	Movie	Zubaan	Music & Musicals	Mozez Singh	Chittaranjan Tripathy	India	TV-14	2019-03-02	2015	111 min	2019	March

199562 rows × 13 columns



Visualisations

```
In [84]: sns.set_style('darkgrid')
plt.rcParams['font.size'] = 14
plt.rcParams['figure.figsize'] = (15, 7)
plt.rcParams['figure.facecolor'] = '#00000000'
```

```
In [85]: df['type'].value_counts()
```

```
Out[85]: Movie      144448
TV Show      55114
Name: type, dtype: int64
```



```
In [86]: x = df['type'].value_counts().to_list()

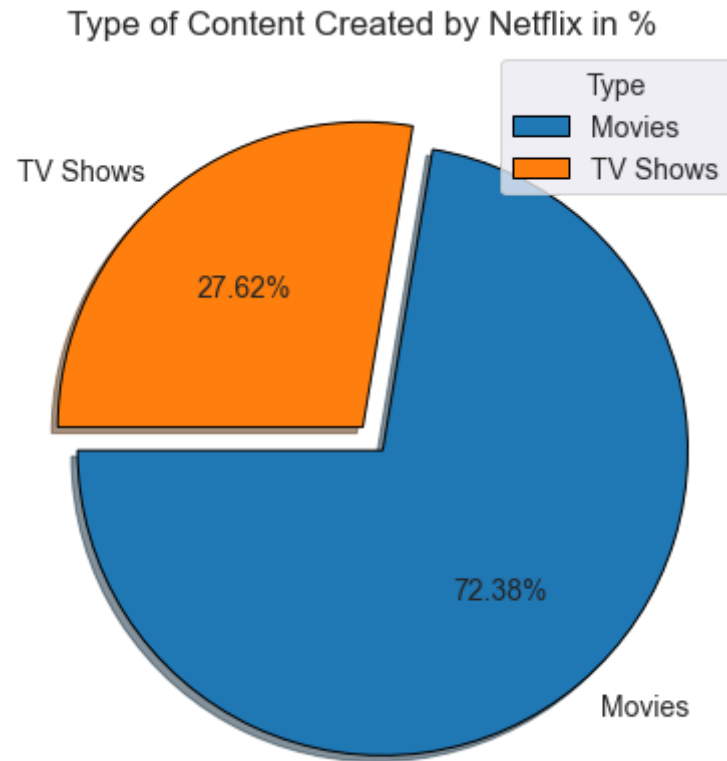
fig, ax = plt.subplots(figsize =(10, 7))

ax.pie(x,
      startangle=180,
      labels=['Movies', 'TV Shows'],
      radius=1,
      shadow=True,
      explode= [0.1, 0],
      autopct= '%.2f%%',
      wedgeprops= {'linewidth': 1, 'edgecolor': 'black'})

plt.legend(['Movies', 'TV Shows'],
          title='Type')

plt.title('Type of Content Created by Netflix in %')

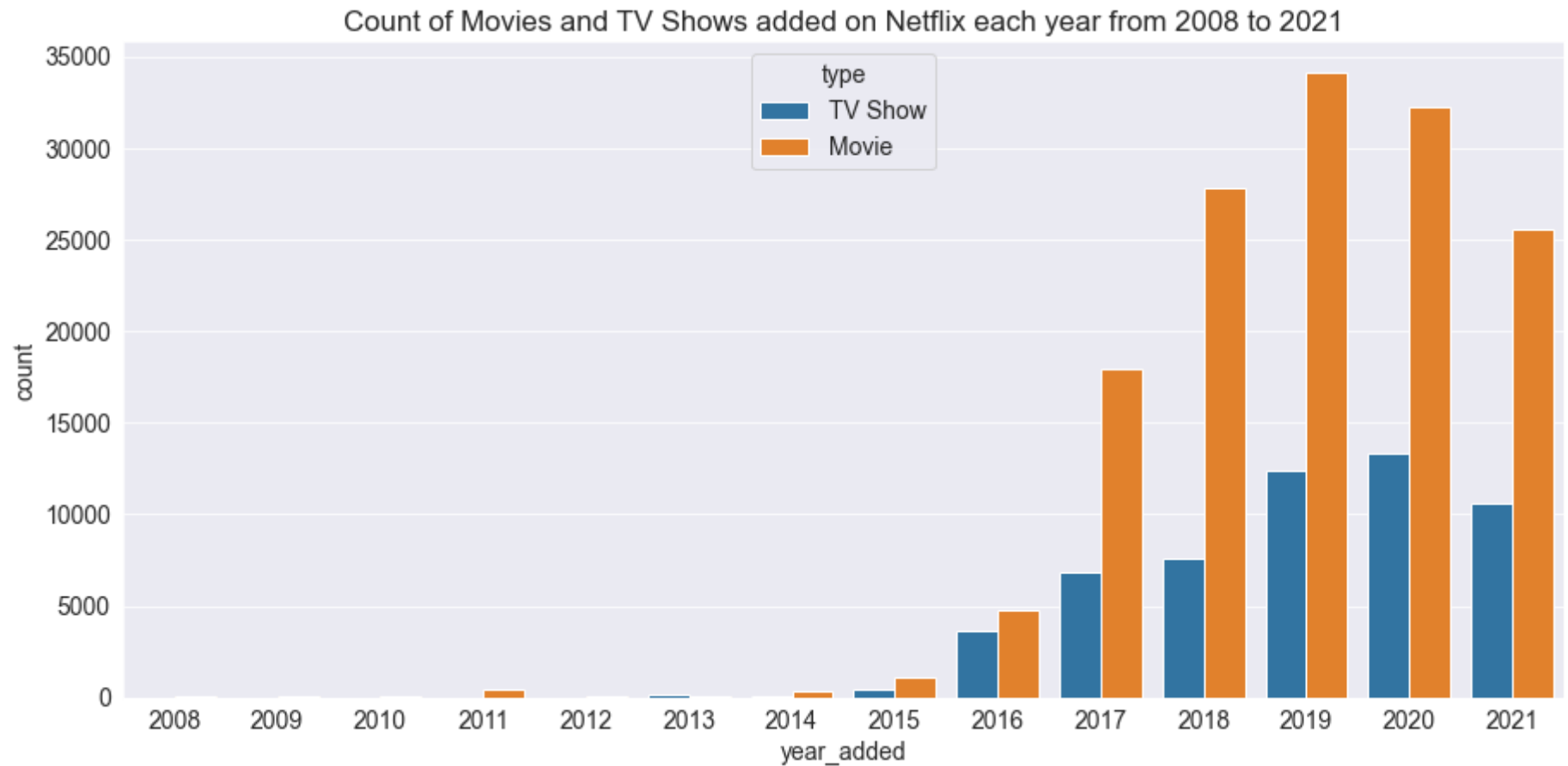
plt.show()
```



From the above Pie chart we can see that Netflix has focused more on Movies than TV Shows. We will have lot of options in Movies rather than TV Shows

Content added per year

```
In [87]: sns.countplot(data=df,  
                        x='year_added',  
                        hue='type')  
plt.title('Count of Movies and TV Shows added on Netflix each year from 2008 to 2021')  
plt.show()
```



Insight-1

In the above countplot we can see the count of Movies and TV Shows added in the Netflix from 2008 to 2021. Here we can see that Netflix started focusing on content addition on the platform from 2015, till 2016 they were focusing equally on the Movies and TV Shows, but from 2017 they suddenly started to focus on adding movies and it gradually increased by hitting its peak at 2019. Again in 2020 no. of movies started to fall gradually. On the other side TV Shows are getting added continuously, and we can see it in the chart that the number is increasing each year.

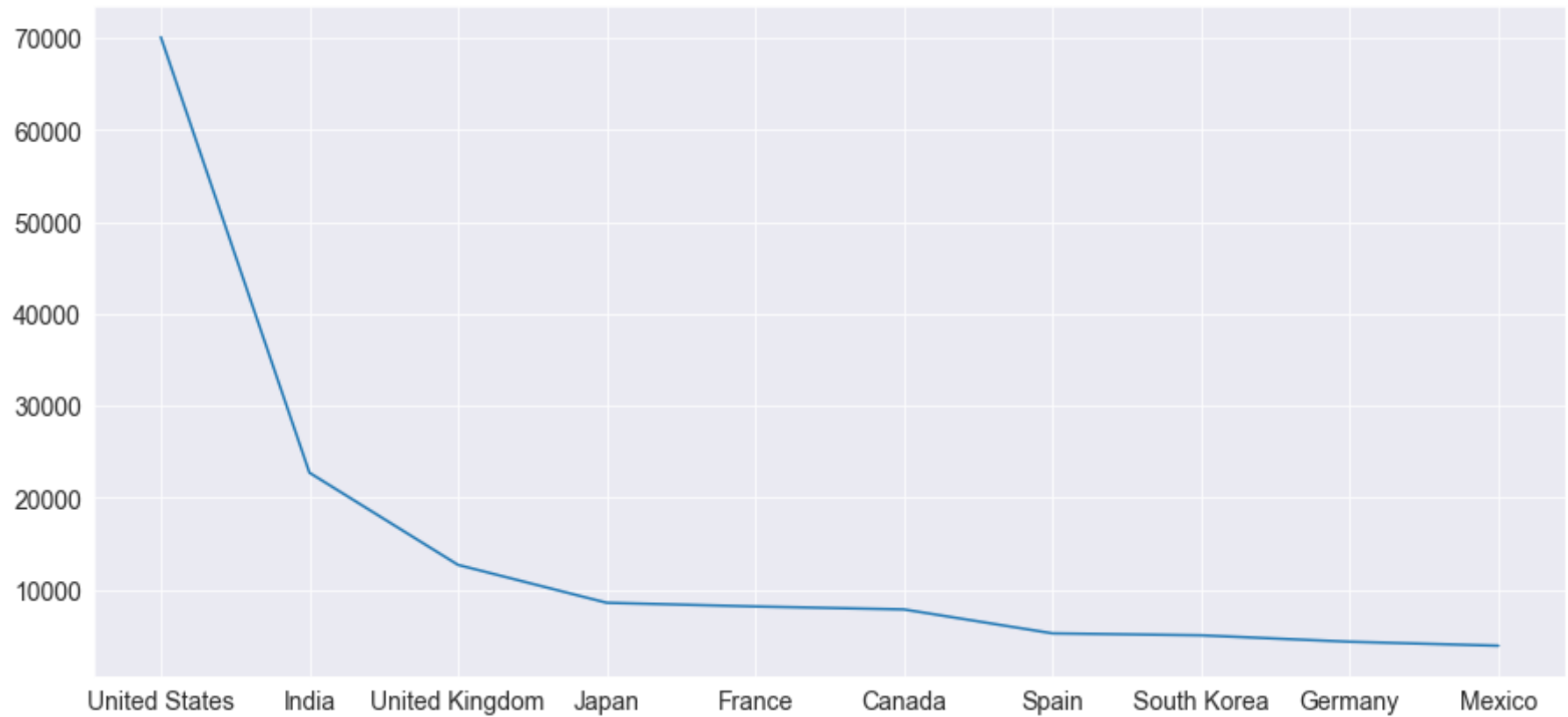
From 2017 to 2021 the number of Movies is almost twice the number of TV Shows, hence it can be concluded that in recent years there has been increase in no. of TV Shows and Movies on Netflix

Top 10 Countries Creating most Content

```
In [88]: df.groupby(['country']).count().sort_values(by='type', ascending=False)['show_id'].head(10)
```

```
Out[88]: country
United States    70073
India            22717
United Kingdom   12693
Japan            8575
France           8170
Canada           7849
Spain            5249
South Korea      5035
Germany          4335
Mexico           3905
Name: show_id, dtype: int64
```

```
In [89]: countries = ['United States', 'India', 'United Kingdom', 'Japan', 'France', 'Canada', 'Spain', 'South Korea', 'Germany',  
content_count = df.groupby(['country']).count().sort_values(by='type', ascending=False)['show_id'].head(10).to_list()  
  
sns.lineplot(x=countries, y=content_count)  
plt.show()
```



Insight - 2

Above are the Top 10 countries which has highest amount of content generated. It can be seen in the above line chart that United States has most amount of content followed by India and United Kingdom.

Recommendation

Netflix should also focus more on countries like Japan, France, Canada, Spain, South Korea, Germany, Mexico as the entertainment industry of these countries is developing and it will also help Netflix to increase its business globally.

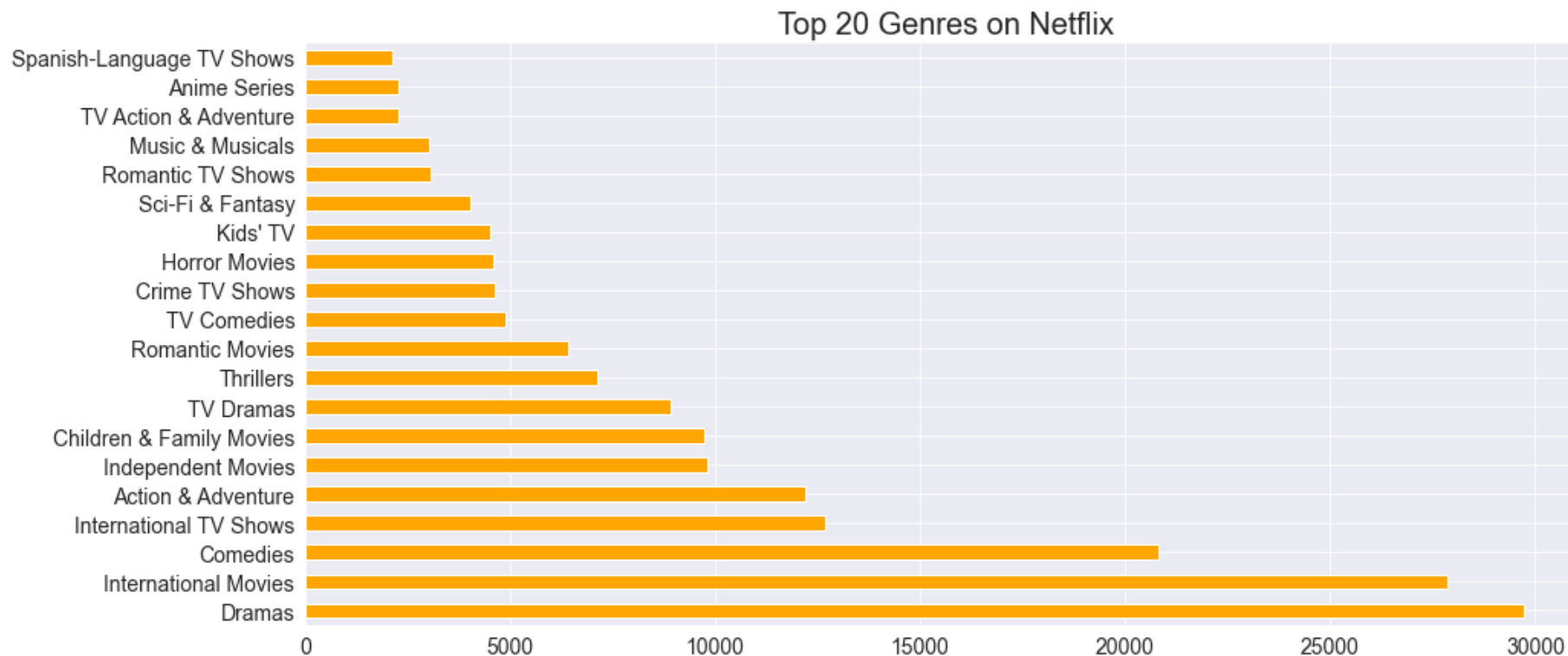
Top 20 Genres available on Netflix

```
In [90]: df['genre'].value_counts().head(20)
```

```
Out[90]: Dramas                29736
          International Movies  27858
          Comedies             20814
          International TV Shows 12691
          Action & Adventure    12206
          Independent Movies     9815
          Children & Family Movies 9728
          TV Dramas             8926
          Thrillers             7106
          Romantic Movies       6405
          TV Comedies           4883
          Crime TV Shows        4626
          Horror Movies         4567
          Kids' TV              4509
          Sci-Fi & Fantasy       4024
          Romantic TV Shows     3035
          Music & Musicals       3024
          TV Action & Adventure  2275
          Anime Series          2270
          Spanish-Language TV Shows 2103
          Name: genre, dtype: int64
```

```
In [105]: df['genre'].value_counts()[:20].plot(kind='barh', color='orange')  
plt.title('Top 20 Genres on Netflix', size=20)
```

```
Out[105]: Text(0.5, 1.0, 'Top 20 Genres on Netflix')
```



Insight - 3

From the above graph we can conclude that audience likes to watch Comedies, International Movies and Dramas.

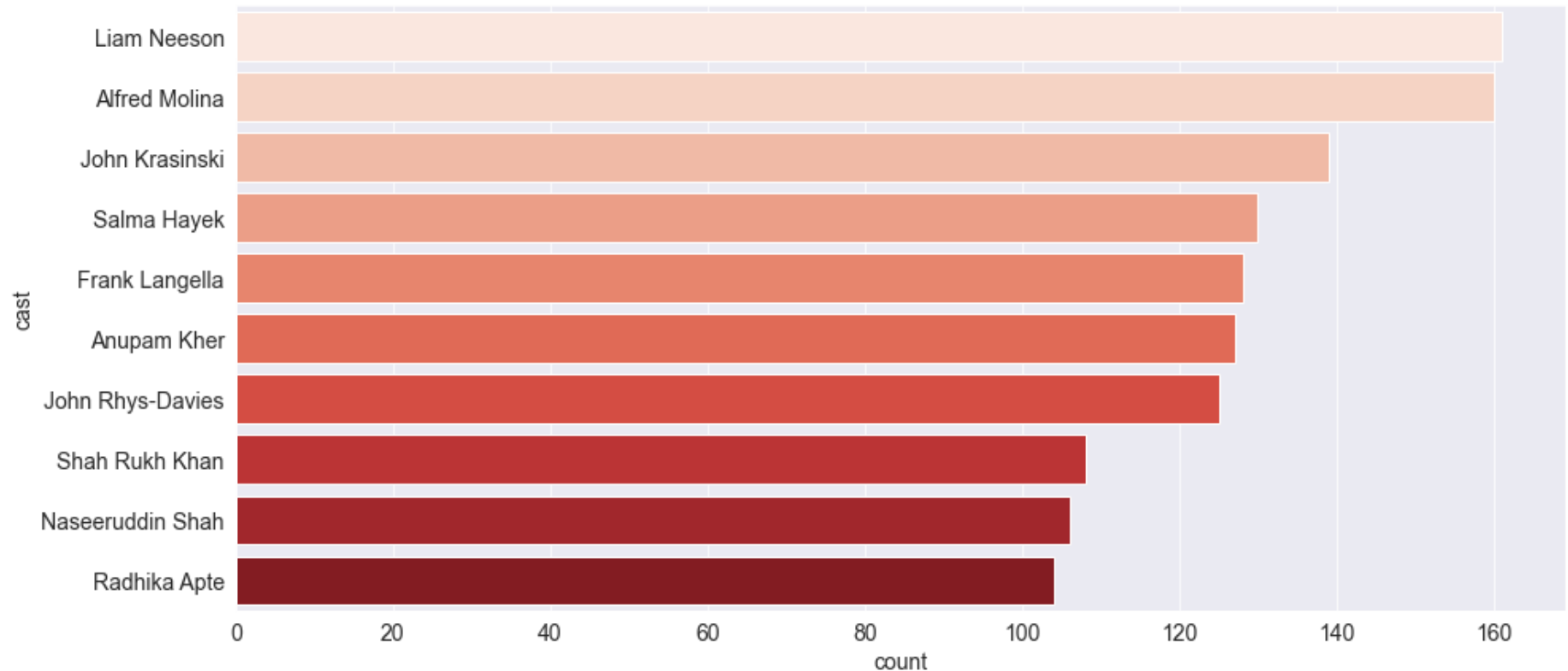
Recommendation

To increase its business Netflix should focus more on generating and adding content related to comedies, International Movies and Dramas genres.

Top 10 Actors based on number of Titles

```
In [116]: sns.countplot(data=df, y='cast', order=df['cast'].value_counts().index[:10], palette='Reds')
```

```
Out[116]: <AxesSubplot:xlabel='count', ylabel='cast'>
```

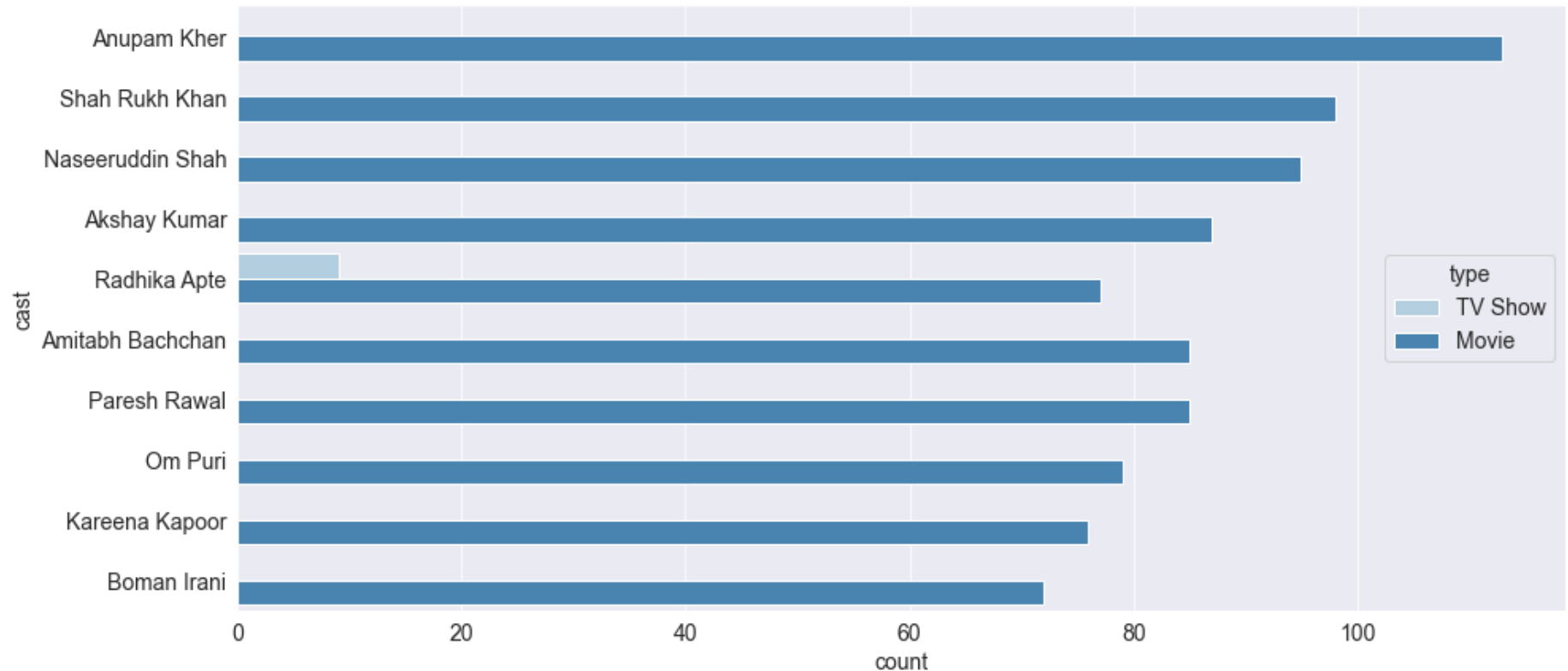


Insight - 4

Top Actor on the basis of total number of titles is Liam Neeson followed by Alfred Molina withh very less margin.


```
In [114]: df_indian_actor = df[df['country'] == 'India']  
  
sns.countplot(data=df_indian_actor, y='cast', order=df_indian_actor['cast'].value_counts().index[:10], palette='Blues',
```

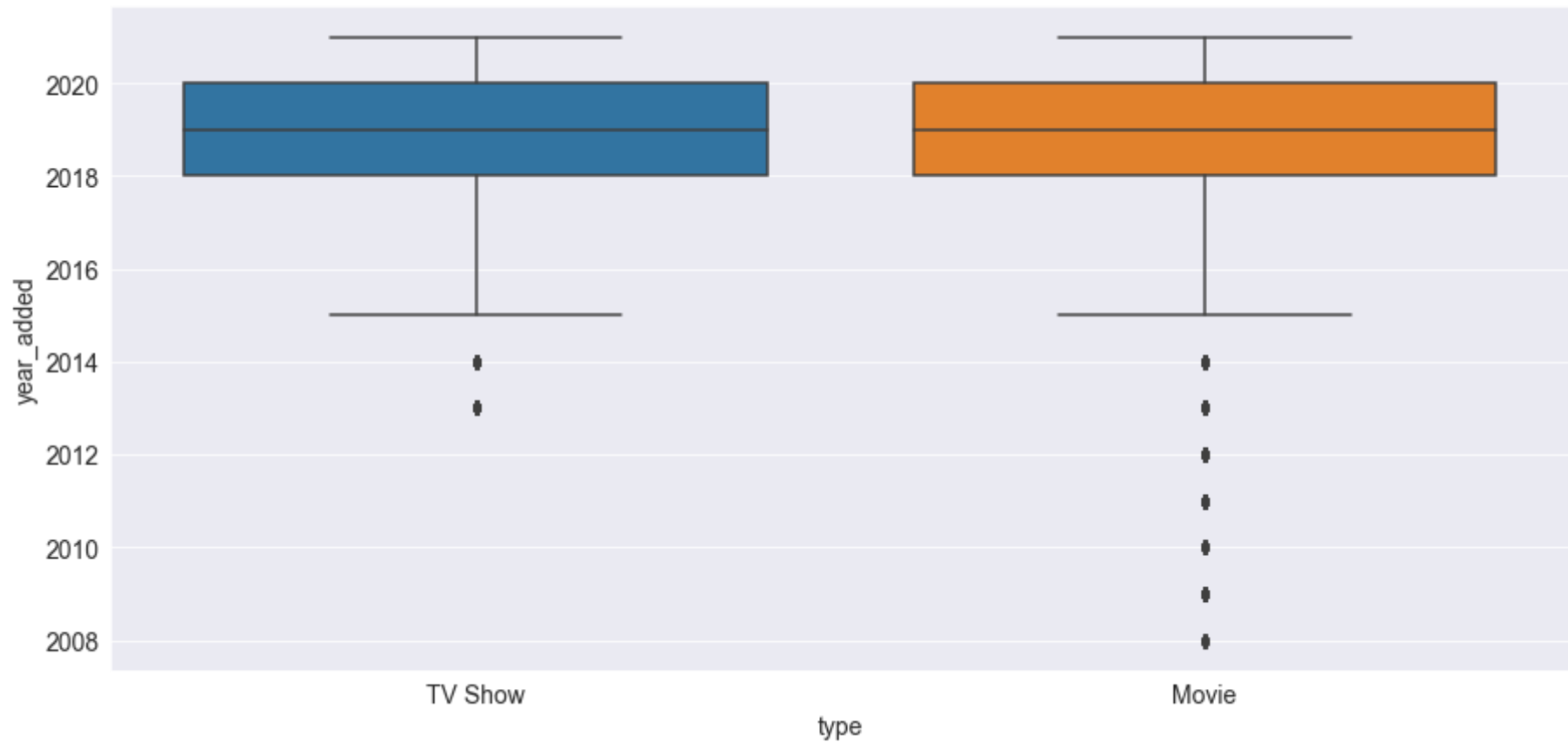
```
Out[114]: <AxesSubplot:xlabel='count', ylabel='cast'>
```



In Top 10 Actors from India we have Anupam Kher at Top with most number of titles. Also in Top 10 Actors from India there is only one actor Radhika Apte who has worked in both TV Show as well as Movie

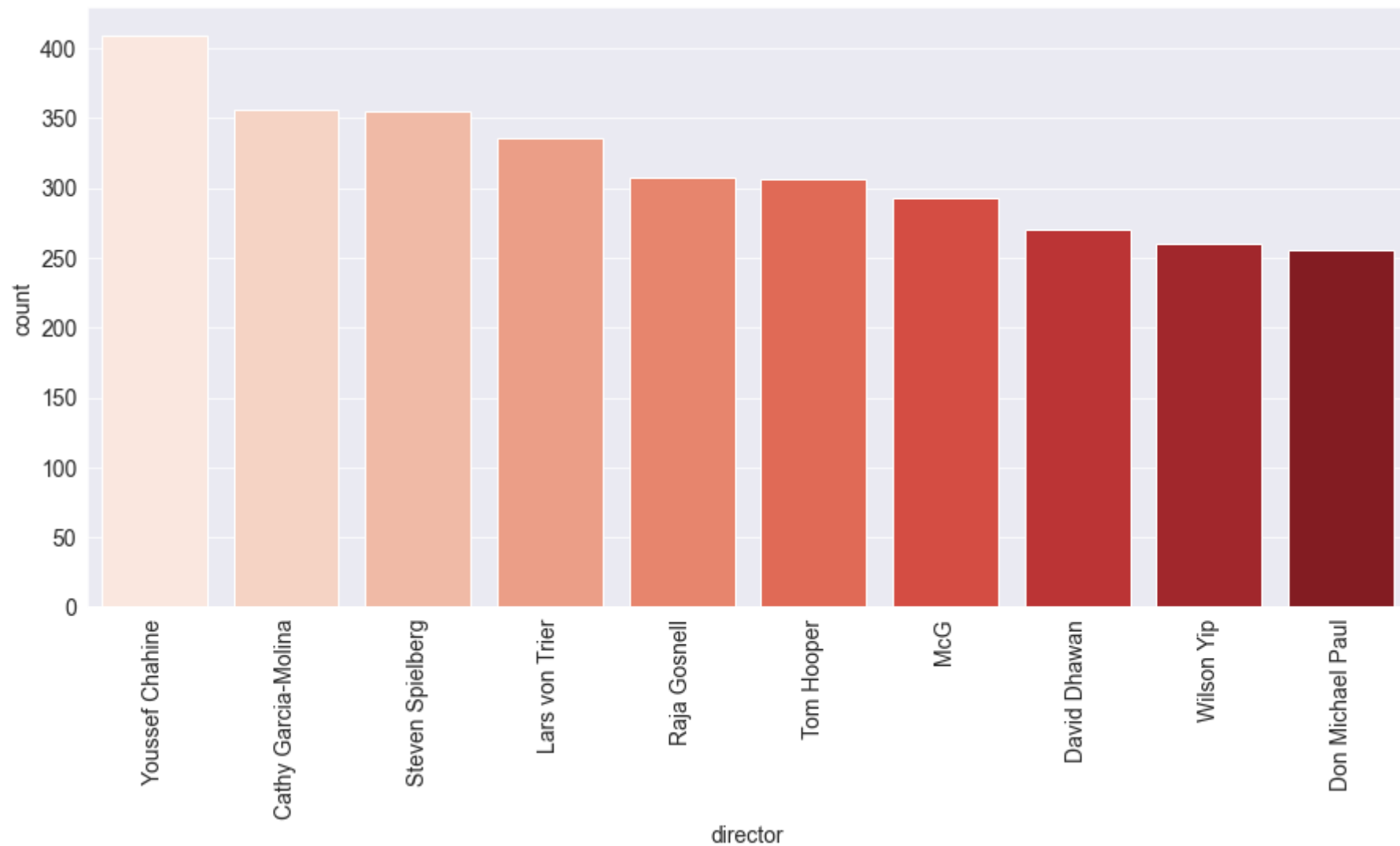
```
In [117]: sns.boxplot(data=df, x='type', y='year_added')
```

```
Out[117]: <AxesSubplot:xlabel='type', ylabel='year_added'>
```



Top 10 Directors based on number of Titles

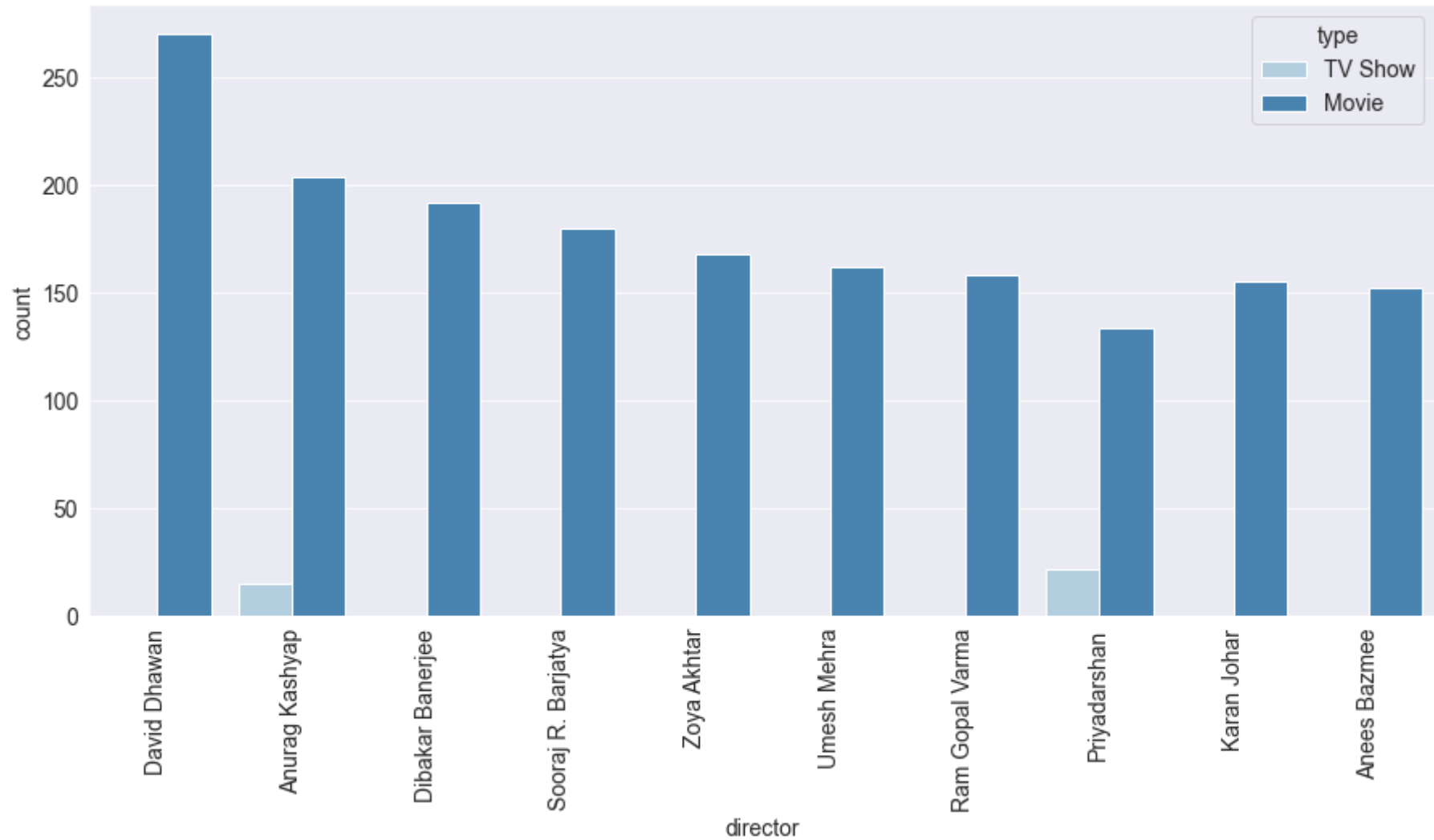
```
In [121]: sns.countplot(data=df, x='director', order=df['director'].value_counts().index[1:11], palette='Reds')  
plt.xticks(rotation=90)  
plt.show()
```



Insight - 5

Youssef Chahine is the top director according to number of titles globally.

```
In [124]: df_indian_actor = df[df['country'] == 'India']  
  
sns.countplot(data=df_indian_actor, x='director', order=df_indian_actor['director'].value_counts().index[1:11], palette=''  
plt.xticks(rotation=90)  
plt.show()
```



David Dhawan is the top director in India according to the number of titles followed by Anurag Kashyap. Also Anurag Kashyap and

Priyadarshan are the only directors in India who have directed a TV Show.

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

