In [ ]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import warnings
import plotly.graph_objects as go
```

In [ ]:

```
df=pd.read_csv(r'C:\Users\DELL\Downloads\netflix_titles.csv~\netflix_titles.csv')
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 [ ]:

```
df["date_added"]=pd.to_datetime(df["date_added"])
```

In [ ]:

```
df.info()
```

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

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

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year	rati
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	T N
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	T N
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	T N
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	T N

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

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

**** Show data columns

In [ ]:
```
df.columns
```

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

In [ ]:

```
df.describe()
```

Out[ ]:

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

Inspect missing values in the data set

In [ ]:

```
df.isnull().sum().sort_values(ascending=False)
```

Out[ ]:

```
director        2634
country          831
cast             825
date_added        10
rating             4
duration           3
show_id            0
type               0
title              0
release_year       0
listed_in          0
description        0
dtype: int64
```

*****check percentage of null data

In [ ]:

```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[ ]:

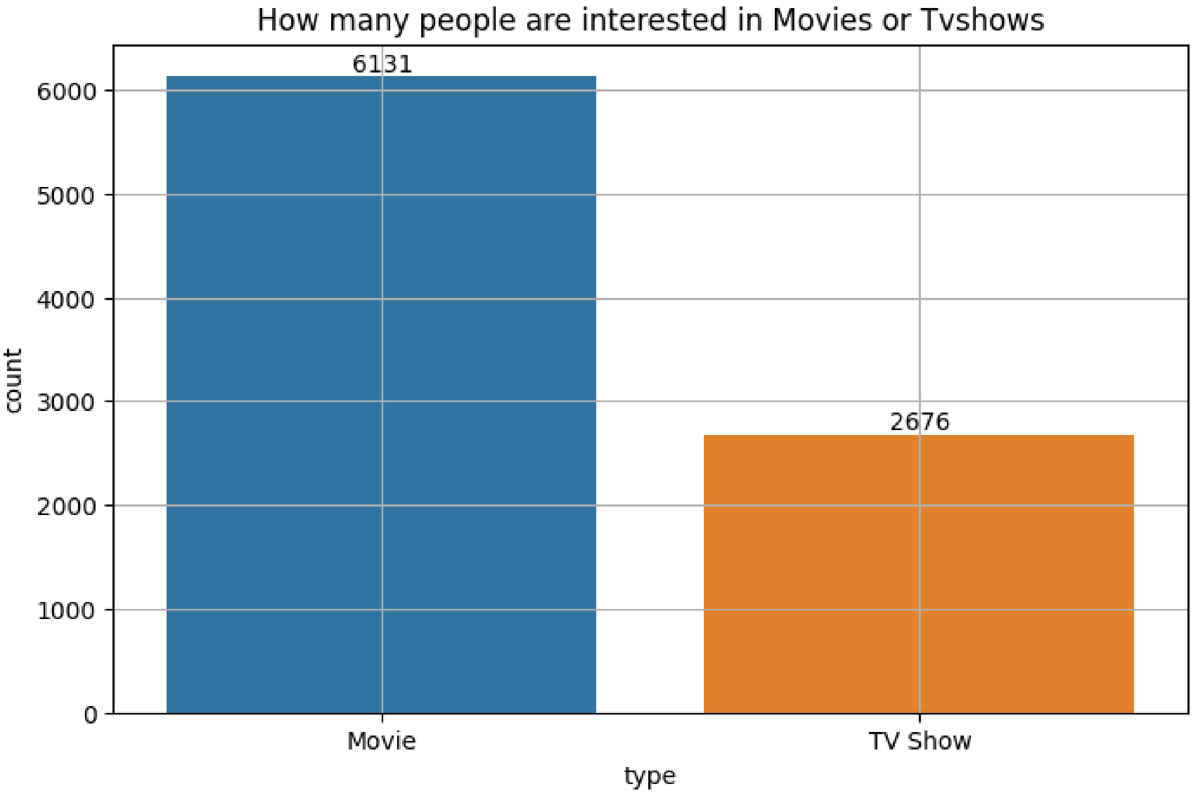
```
director        29.91
country          9.44
cast             9.37
date_added       0.11
rating           0.05
duration         0.03
show_id          0.00
type             0.00
title            0.00
release_year     0.00
listed_in        0.00
description      0.00
dtype: float64
```

In [ ]:
```
df["director"].value_counts().head(10)
```

Out[ ]:
```
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Marcus Raboy              16
Suhas Kadav               16
Jay Karas                 14
Cathy Garcia-Molina       13
Martin Scorsese           12
Youssef Chahine           12
Jay Chapman               12
Steven Spielberg          11
Name: director, dtype: int64
```

Movie v/s Tvshows......Data Visulizations

In [ ]:
```
plt.figure(figsize=(8,5))
c=sns.countplot(data=df,x=df["type"])
plt.title("How many people are interested in Movies or Tvshows")
plt.grid()
for count in c.containers:
 c.bar_label(count)
```



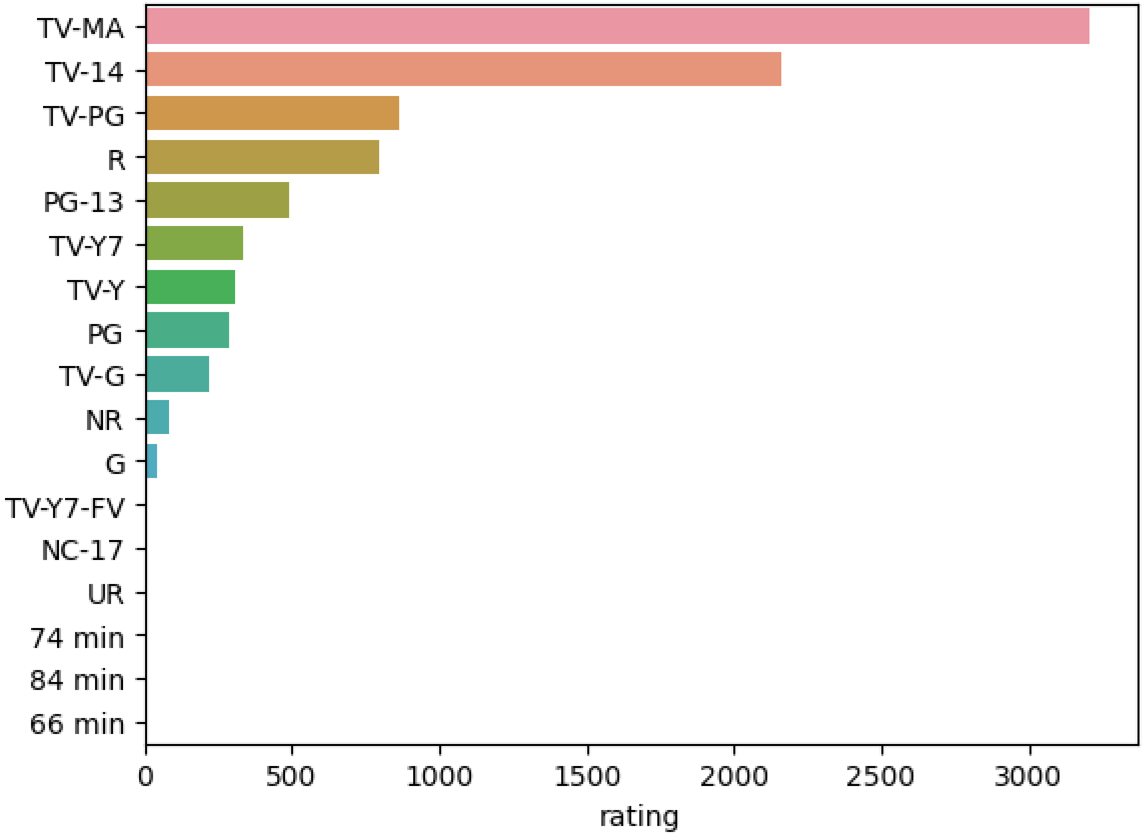
In [ ]:
```
pie_chart = go.Figure(data=[go.Pie(labels=df['type'].value_counts(normalize=True).i
                                   values=df['type'].value_counts(normalize=True).v
                                   hole=0.5)])
pie_chart.update_layout(title="Movies v/s TV Shows")
```

In [ ]:
```
df.rating.value_counts()
```

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

```
In [ ]: sns.barplot(data=df, x=df.rating.value_counts(),y=df.rating.value_counts().index,or
```

```
Out[ ]: <Axes: xlabel='rating'>
```



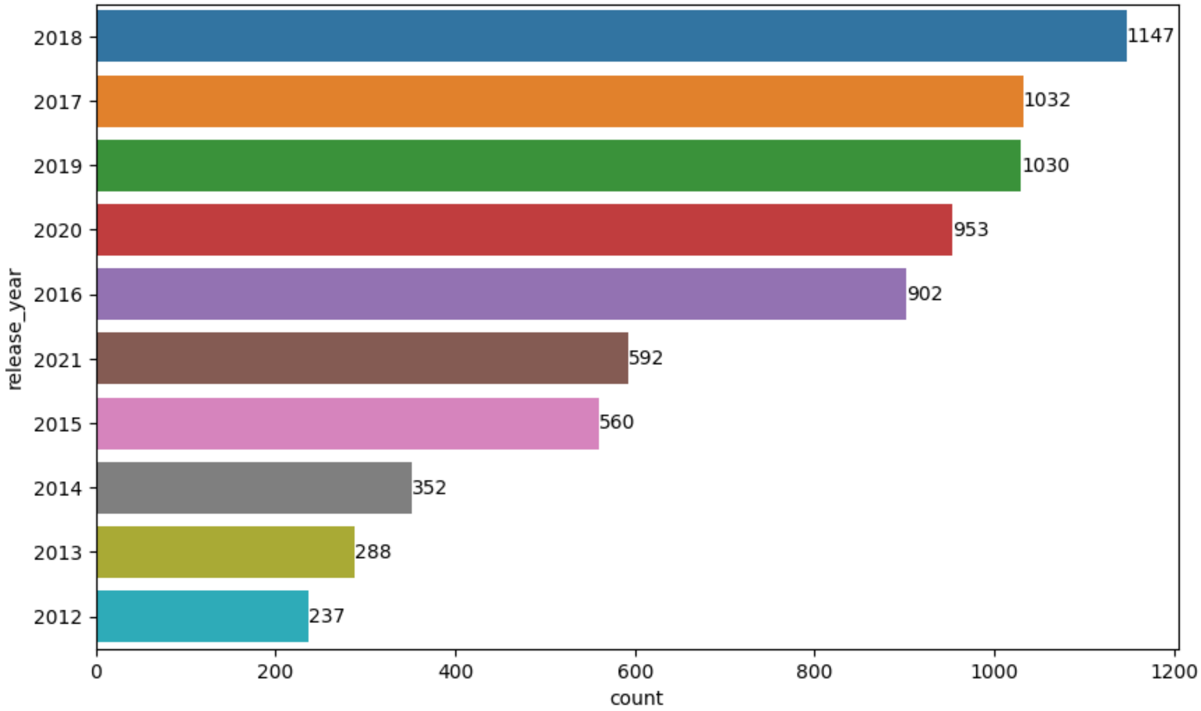
```
In [ ]: ndf=df.country.value_counts().reset_index().head(10)
        ndf=ndf.rename(columns={"index":"Country","country":"Count"})
        ndf
```

Out[ ]:

	Country	Count
0	United States	2818
1	India	972
2	United Kingdom	419
3	Japan	245
4	South Korea	199
5	Canada	181
6	Spain	145
7	France	124
8	Mexico	110
9	Egypt	106

In [ ]:

```
plt.figure(figsize=(10,6))
a=sns.countplot(data=df, y=df["release_year"],order=df.release_year.value_counts().
for count in a.containers:
    a.bar_label(count)
```



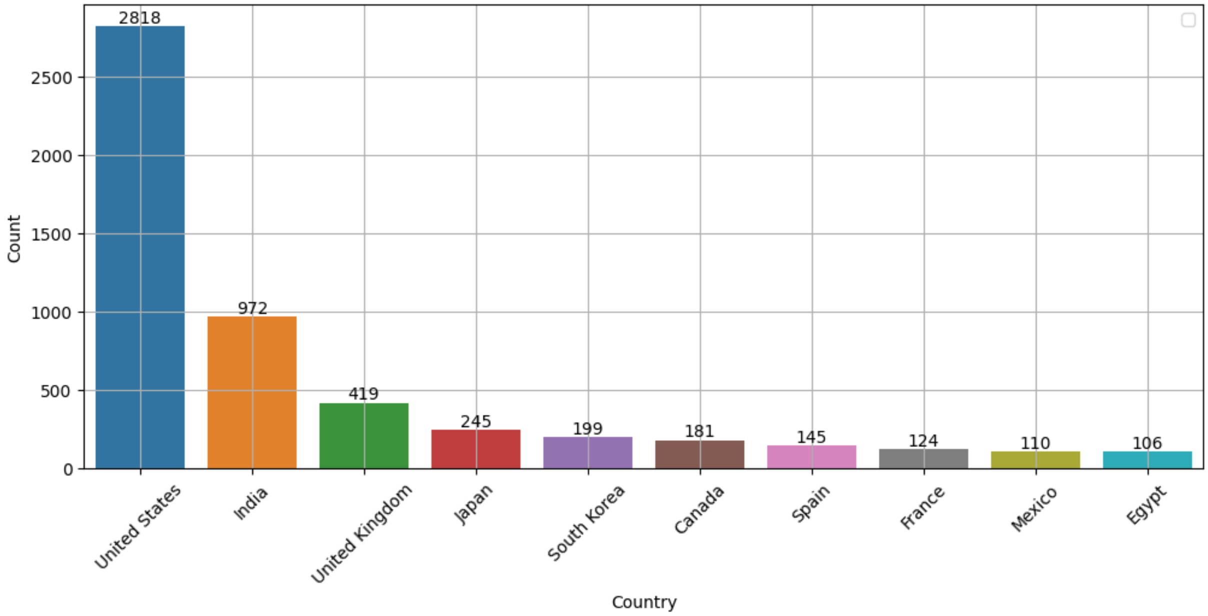
****** Higest release in 2018 followed by 2017 and 2019 ******

In [ ]:

```
plt.figure(figsize=(12,5))
r=sns.barplot(data=ndf,x=ndf.Country,y=ndf.Count)
plt.xticks(rotation=45)
plt.grid()
plt.legend()
```

```
for count in r.containers:
 r.bar_label(count)
```

```
No artists with labels found to put in legend.  Note that artists whose label start
with an underscore are ignored when legend() is called with no argument.
```



Top 10 Listed shows

In [ ]:
```
df.listed_in.value_counts().head(10)
```

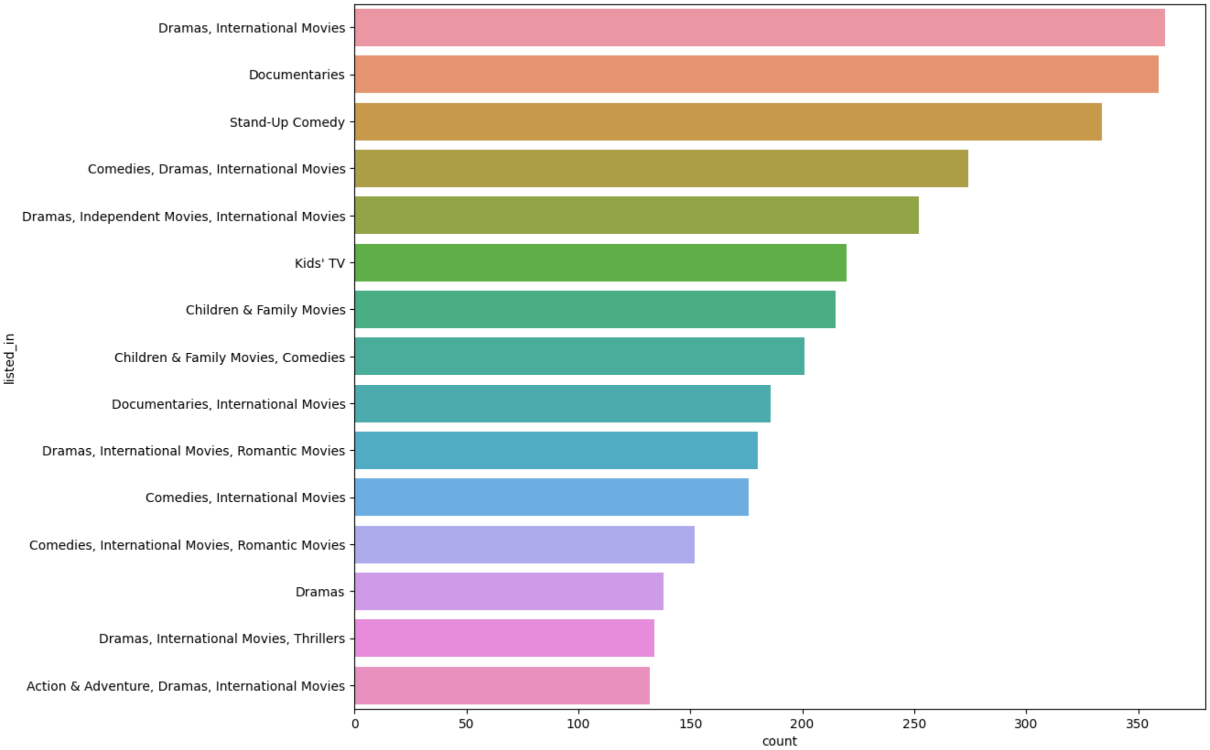
Out[ ]:
```
Dramas, International Movies                        362
Documentaries                                       359
Stand-Up Comedy                                     334
Comedies, Dramas, International Movies              274
Dramas, Independent Movies, International Movies    252
Kids' TV                                            220
Children & Family Movies                            215
Children & Family Movies, Comedies                  201
Documentaries, International Movies                 186
Dramas, International Movies, Romantic Movies       180
Name: listed_in, dtype: int64
```

In [ ]:
```
df.listed_in.value_counts().tail(10)
```

Out[ ]:
```
Docuseries, Reality TV, Teen TV Shows                   1
Crime TV Shows, International TV Shows, Reality TV      1
Anime Features, Romantic Movies                         1
Anime Features, Music & Musicals                        1
British TV Shows, Kids' TV, TV Thrillers                1
Kids' TV, TV Action & Adventure, TV Dramas              1
TV Comedies, TV Dramas, TV Horror                       1
Children & Family Movies, Comedies, LGBTQ Movies        1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows      1
Cult Movies, Dramas, Thrillers                          1
Name: listed_in, dtype: int64
```

In [ ]:

```
plt.figure(figsize=(12,10))
ax=sns.countplot(y="listed_in", data=df,order=df.listed_in.value_counts().index[0:1
```



Heading missing values

In [ ]:

```
round(df.isnull().sum()/df.shape[0]*100,4).sort_values(ascending=False).head(7)
```

Out[ ]:

```
director      29.9080
country        9.4357
cast           9.3675
date_added     0.1135
rating         0.0454
duration       0.0341
show_id        0.0000
dtype: float64
```

In [ ]:

```
df.dropna(subset=["rating","duration","date_added"],axis=0, inplace=True)
df.shape
```

Out[ ]:

```
(8790, 12)
```

In [ ]:

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

Out[ ]:

```
show_id            0
type               0
title              0
director        2621
cast             825
country          829
date_added         0
release_year       0
rating             0
duration           0
listed_in          0
description        0
dtype: int64
```

In [ ]:

```
# replace missing values for country,cast , director"

df["country"].replace(np.nan,"Unknown", inplace=True)
df["cast"].replace(np.nan,"No Cast", inplace=True)
df["director"].replace(np.nan,"No Director", inplace=True)
```

In [ ]:

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

Out[ ]:

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

In [ ]:

```
cast_show=df.loc[df["cast"]!="No Cast"]
cast_show.cast.value_counts().reset_index()
```

Out[ ]:

	index	cast
0	David Attenborough	19
1	Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jig...	14
2	Samuel West	10
3	Jeff Dunham	7
4	David Spade, London Hughes, Fortune Feimster	6
...	...	...
7673	Sanjay Dutt, Arjun Kapoor, Kriti Sanon, Zeenat...	1
7674	Lika Berning, Bobby van Jaarsveld, Marlee van ...	1
7675	Lisa Vicari, Dennis Mojen, Walid Al-Atiyat, Ch...	1
7676	Piotr Cyrwus, Mikołaj Kubacki, Anna Radwan, Ma...	1
7677	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	1

7678 rows × 2 columns

In [ ]:

```
year=df.release_year.value_counts().head(10)
ndf=year.reset_index().rename(columns=({"index":"Year","release_year":"No of Count"
ndf=ndf.sort_values("Year")
ndf
```

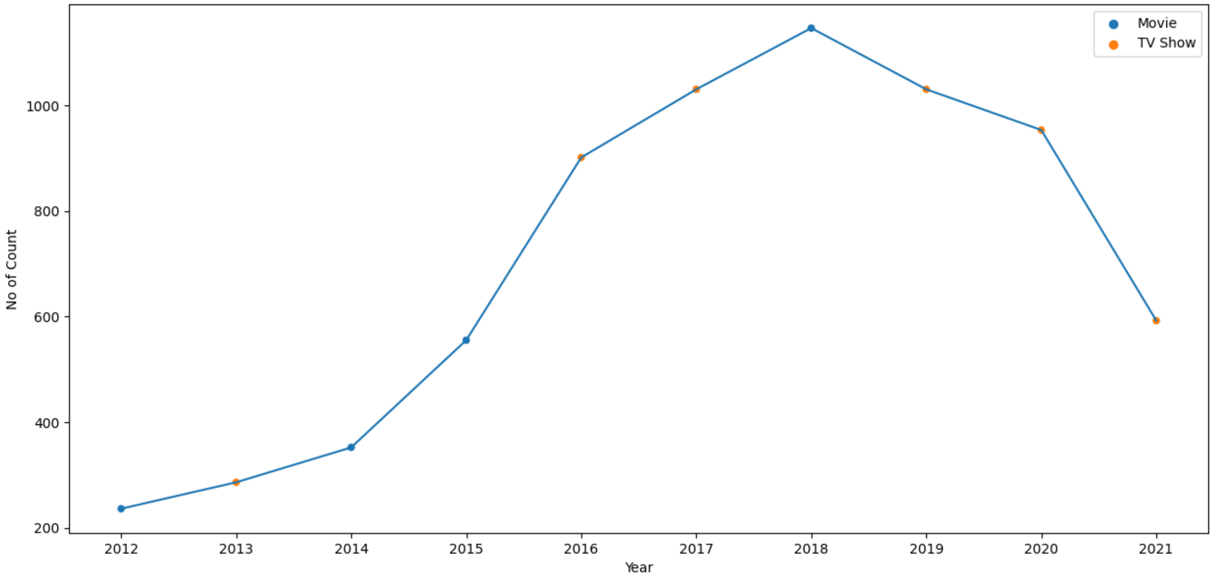
Out[ ]:

	Year	No of Count
9	2012	236
8	2013	286
7	2014	352
6	2015	555
4	2016	901
1	2017	1030
0	2018	1146
2	2019	1030
3	2020	953
5	2021	592

In [ ]:

```
plt.figure(figsize=(15,7))
sns.scatterplot(data=ndf,x=ndf["Year"],y=ndf["No of Count"], hue=df["type"])
plt.xticks(np.array(range(2012,2022)))
sns.lineplot(data=ndf,x=ndf["Year"],y=ndf["No of Count"])
```

```
Out[ ]: <Axes: xlabel='Year', ylabel='No of Count'>
```



```
In [ ]: cdf=df.loc[df["country"]!="Unknown"]
        cdf.head()
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year	r
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	2021-09-25	2020	F
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	
4	s5	TV Show	Kota Factory	No Director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	2021-09-24	1993	
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	2021-09-24	2021	-

In [ ]:

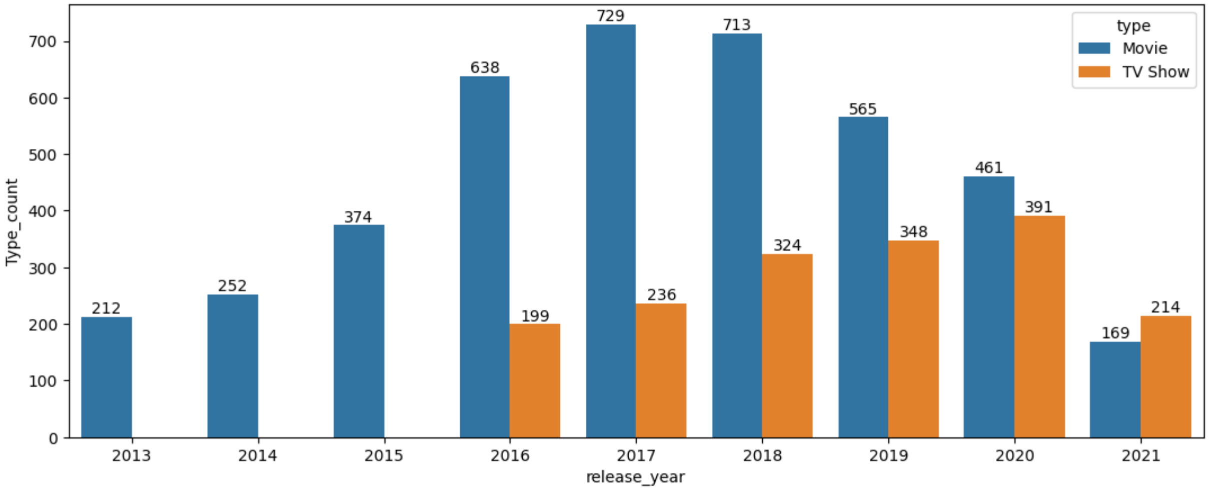
```
a=cdf.groupby(["release_year","type"]).agg(
    Type_count=("type", "count")
).reset_index().sort_values("Type_count",ascending=False).head(15)
a
```

Out[ ]:

	release_year	type	Type_count
107	2017	Movie	729
109	2018	Movie	713
105	2016	Movie	638
111	2019	Movie	565
113	2020	Movie	461
114	2020	TV Show	391
103	2015	Movie	374
112	2019	TV Show	348
110	2018	TV Show	324
101	2014	Movie	252
108	2017	TV Show	236
116	2021	TV Show	214
99	2013	Movie	212
106	2016	TV Show	199
115	2021	Movie	169

In [ ]:

```
plt.figure(figsize=(13,5))
b=sns.barplot(data=a,x=a["release_year"],y=a["Type_count"],hue="type")
for count in b.containers:
    b.bar_label(count)
```



In [ ]:

```
df.loc[df["duration"]<"1 min"]
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown	2021-09-24	2021
3	s4	TV Show	Jailbirds New Orleans	No Director	No Cast	Unknown	2021-09-24	2021
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	Unknown	2021-09-24	2021
10	s11	TV Show	Vendetta: Truth, Lies and The Mafia	No Director	No Cast	Unknown	2021-09-24	2021
11	s12	TV Show	Bangkok Breaking	Kongkiat Komesiri	Sukollawat Kanarot, Sushar Manaying, Pavarit M...	Unknown	2021-09-23	2021
...	...	...	...	...	...	...	...	...
8775	s8776	TV Show	Yeh Meri Family	No Director	Vishesh Bansal, Mona Singh, Akarsh Khurana, Ah...	India	2018-08-31	2018
8780	s8781	TV Show	Yo-Kai Watch	No Director	Johnny Yong Bosch, J.W. Terry, Alicyn Packard,...	United States	2016-04-01	2015
8783	s8784	TV Show	Yoko	No Director	Eileen Stevens, Alyson	Unknown	2018-06-23	2016

	show_id	type	title	director	cast	country	date_added	release_year
					Leigh Rosenfeld, Sarah ...			
8785	s8786	TV Show	YOM	No Director	Sairaj, Devyani Dagaonkar, Ketan Singh, Mayur ...	Unknown	2018-06-07	2016
8800	s8801	TV Show	Zindagi Gulzar Hai	No Director	Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen ...	Pakistan	2016-12-15	2012

In [ ]:

```
movie=df.loc[df["type"]=="Movie"]
movie=movie["duration"].value_counts().reset_index().sort_values("index",ascending=
movie
```

Out[ ]:

	index	duration
10	99 min	118
9	98 min	120
3	97 min	146
6	96 min	130
5	95 min	137
...	...	...
13	103 min	114
8	102 min	122
11	101 min	116
15	100 min	108
188	10 min	1

205 rows × 2 columns

In [ ]:

```
# shortest movie
movie=df.loc[df["type"]=="Movie"]
min=movie.loc[movie["duration"]=="10 min"]
min
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year
3535	s3536	Movie	American Factory: A Conversation with the Obamas	No Director	President Barack Obama, Michelle Obama, Julia ...	United States	2019-09-05	2019

In [ ]:

```
# Largest movie
movie=df.loc[df["type"]=="Movie"]
min=movie.loc[movie["duration"]=="99 min"].head(1)
min
```

Out[ ]:

	show_id	type	title	director	cast	country	date_added	release_year	ra
51	s52	Movie	InuYasha the Movie 2: The Castle Beyond the Lo...	Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Mieko Harada...	Japan	2021-09-15	2002	T

In [ ]:

```
df.to_csv("new_data.csv",index=False)
```

In [ ]: