In [1]:

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

In [2]:

```
df = pd.read_csv('netflix.csv')
df.shape
```

Out[2]:

```
(8807, 12)
```

In [3]:

```
df.columns
```

Out[3]:

```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'dat
e_added',
       'release_year', 'rating', 'duration', 'listed_in', 'descriptio
n'],
      dtype='object')
```

Data Cleaning Preprocessing

In [23]:

```
df.head(3)
```

Out[23]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	d
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	S
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	

In [11]:

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

Out[11]:

```
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
```

In [14]:

```
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 [16]:

```
df['date_added'] = pd.to_datetime(df['date_added'])
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 [20]:

```
df.duplicated().any()
```

Out[20]:

```
False
```

In [21]:

```
# df.drop_duplicates(subset = ['show_id'])
```

In [22]:

```
# descriptive statistics
df.describe()
```

Out[22]:

	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

Univariate Analysis

In [5]:

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

Out[5]:

```
United States
2818
India
972
United Kingdom
419
Japan
245
South Korea
199

...
Canada, United States, Mexico
1
United States, Ireland, United Kingdom, India
1
United Kingdom, Czech Republic, United States, Germany, Bahamas
1
France, Brazil, Spain, Belgium
1
United States, Ghana, Burkina Faso, United Kingdom, Germany, Ethiopia
1
Name: country, Length: 748, dtype: int64
```

In [ ]:

```

```

In [66]:

```
df.head(3)
```

Out[66]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	d
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	

In [30]:

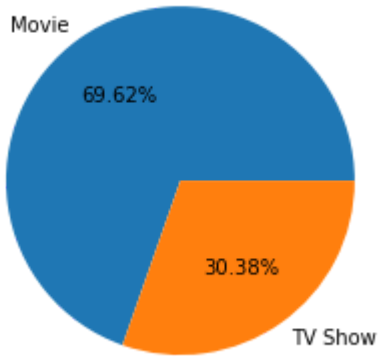
```
df['type'].value_counts()
```

Out[30]:

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

In [34]:

```
plt.pie(df['type'].value_counts(), labels= df['type'].value_counts().index, autopct=
```



Duration of movie/ No. of Seasons

In [ ]:

```
# for i in range(df.shape[0]):
#     df['duration'].str.split()[i][0]
```

In [49]:

```
df['duration'] = df['duration'].str.split().apply(pd.Series)[0]
df['duration']
```

Out[49]:

```
0         90
1          2
2          1
3          1
4          2
        ...
8802     158
8803       2
8804      88
8805      88
8806     111
Name: duration, Length: 8807, dtype: object
```

In [51]:

```
df['duration'].isna().sum()
```

Out[51]:

```
3
```

In [53]:

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

Out[53]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	durat
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	74 min	N
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	84 min	N
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	N

In [55]:

```
df['duration'].fillna(75, inplace = True)
```

In [57]:

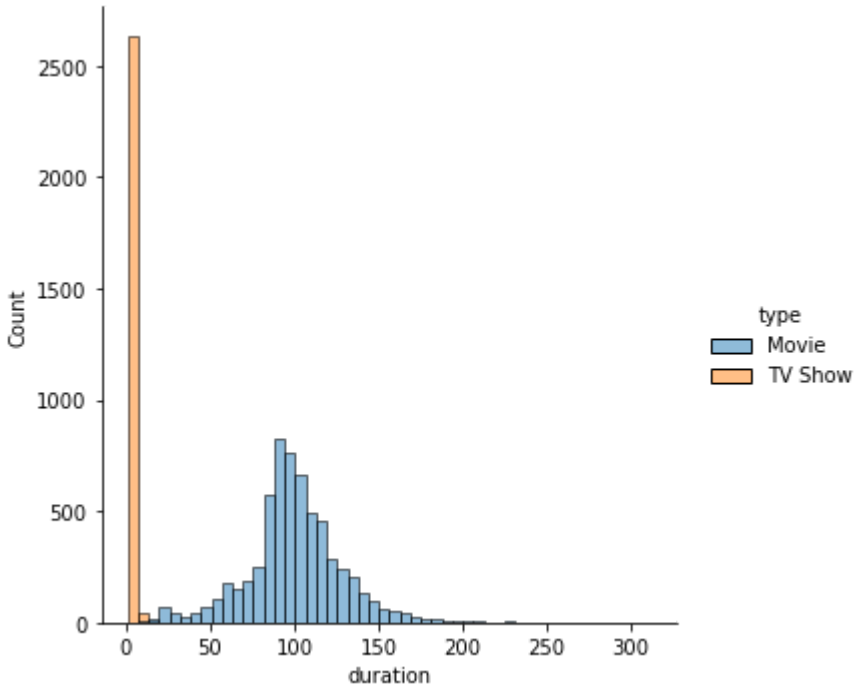
```
df['duration'] = df['duration'].astype('int')
df['duration']
```

Out[57]:

```
0         90
1          2
2          1
3          1
4          2
        ...
8802     158
8803       2
8804      88
8805      88
8806     111
Name: duration, Length: 8807, dtype: int64
```

In [64]:

```
sns.displot(x='duration', data=df, bins=50, hue='type');
```



In [ ]:

```

```

In [ ]:

```

```

In [68]:

```
df.head(3)
```

Out[68]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	d
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	

In [70]:

```
df['date_added'].isna().sum()
```

Out[70]:

```
10
```

In [76]:

```
df['date_added'].fillna(pd.to_datetime('01/01/1900'), inplace=True)
```

In [77]:

```
df['date_added'].isna().sum()
```

Out[77]:

```
0
```

In [82]:

```
df['month_added'] = df['date_added'].dt.month
df['year_added'] = df['date_added'].dt.year
```

In [84]:

```
df.head(3)
```

Out[84]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	d
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	

In [ ]:

```

```

In [91]:

```
data_movies = df[df['type'] == 'Movie'][['year_added', 'month_added']]
data_shows = df[df['type'] == 'TV Show'][['year_added', 'month_added']]
```

In [92]:

```
data_movies.shape
```

Out[92]:

```
(6131, 2)
```

In [93]:

```
data_shows.shape
```

Out[93]:

```
(2676, 2)
```

In [96]:

```
data_movies.head()
```

Out[96]:

	year_added	month_added
0	2021	9
6	2021	9
7	2021	9
9	2021	9
12	2021	9

In [95]:

```
count = pd.DataFrame(columns=['Movies', 'Shows'])
count
```

Out[95]:

	Movies	Shows

In [99]:

```
data_movies.groupby(by='year_added').count()['month_added']
```

Out[99]:

```
year_added
2008       1
2009       2
2010       1
2011      13
2012       3
2013       6
2014      19
2015      56
2016     253
2017     839
2018    1237
2019    1424
2020    1284
2021     993
Name: month_added, dtype: int64
```

In [103]:

```
data_shows.groupby(by='year_added').count()['month_added'].iloc[1:,]
```

Out[103]:

```
year_added
2008      1
2013      5
2014      5
2015     26
2016    176
2017    349
2018    412
2019    592
2020    595
2021    505
Name: month_added, dtype: int64
```

In [104]:

```
count['Movies'] = data_movies.groupby(by='year_added').count()['month_added']
count['Shows'] = data_shows.groupby(by='year_added').count()['month_added'].iloc[1:,
```

In [106]:

```
count.head()
```

Out[106]:

year_added	Movies	Shows
2008	1	1.0
2009	2	NaN
2010	1	NaN
2011	13	NaN
2012	3	NaN

In [107]:

```
count.fillna(0, inplace=True)
```

In [111]:

```
count = count.astype('int')
count
```

Out[111]:

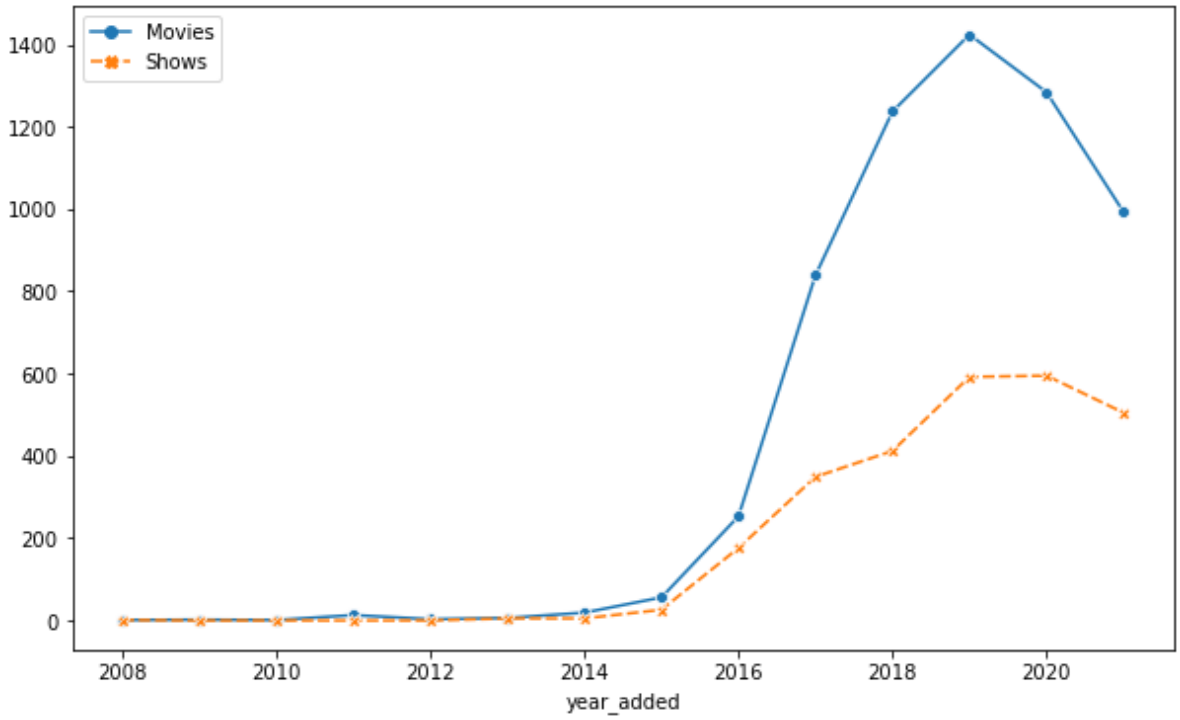
year_added	Movies	Shows
2008	1	1
2009	2	0
2010	1	0
2011	13	0
2012	3	0
2013	6	5
2014	19	5
2015	56	26
2016	253	176
2017	839	349
2018	1237	412
2019	1424	592
2020	1284	595
2021	993	505

In [116]:

```
plt.figure(figsize=(10,6))
sns.lineplot(data=count, markers=True)
```

Out[116]:

```
<AxesSubplot:xlabel='year_added'>
```



In [118]:

```
# pd.pivot_table()
```

In [132]:

```
month_wise_count = df.groupby(by=['year_added', 'month_added']).count().iloc[1:,-1].
month_wise_count
```

Out[132]:

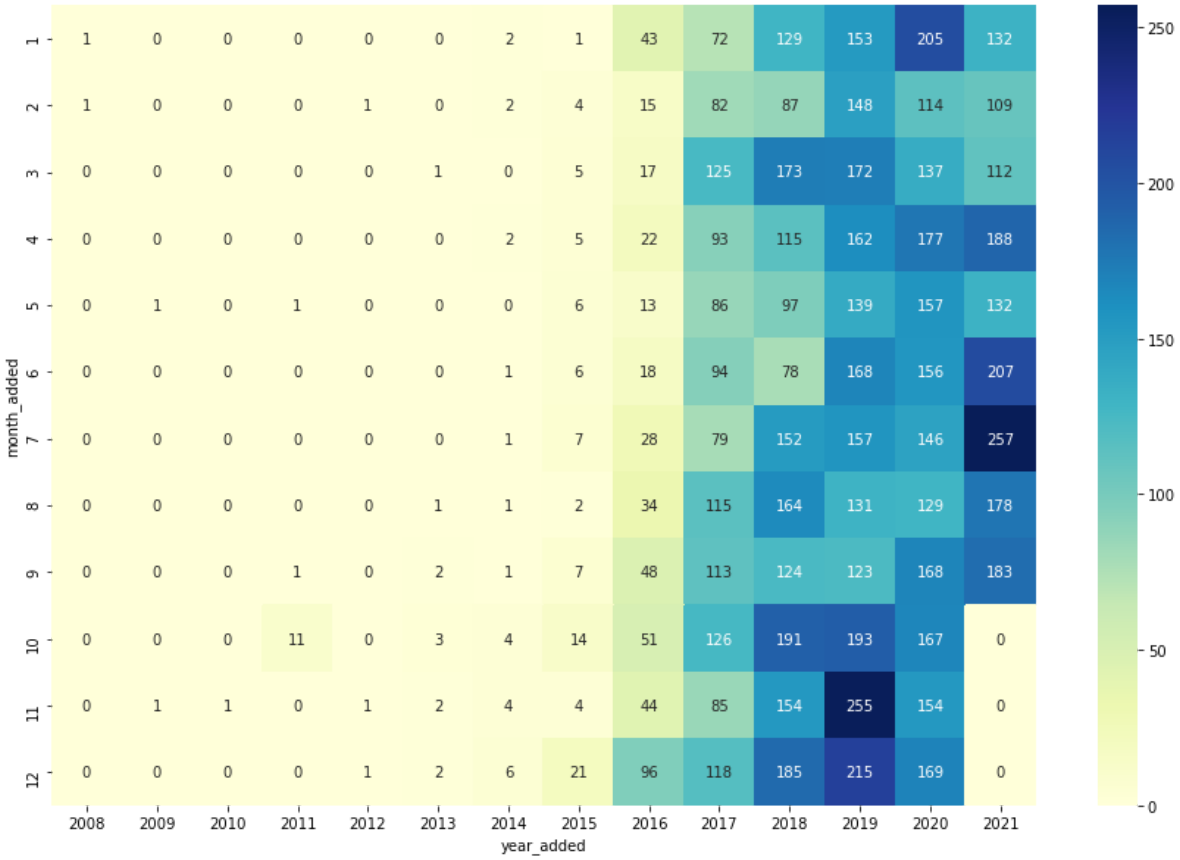
year_added	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
month_added													
1	1	0	0	0	0	0	2	1	43	72	129	153	205
2	1	0	0	0	1	0	2	4	15	82	87	148	114
3	0	0	0	0	0	1	0	5	17	125	173	172	137
4	0	0	0	0	0	0	2	5	22	93	115	162	177
5	0	1	0	1	0	0	0	6	13	86	97	139	157
6	0	0	0	0	0	0	1	6	18	94	78	168	156
7	0	0	0	0	0	0	1	7	28	79	152	157	146
8	0	0	0	0	0	1	1	2	34	115	164	131	129
9	0	0	0	1	0	2	1	7	48	113	124	123	168
10	0	0	0	11	0	3	4	14	51	126	191	193	167
11	0	1	1	0	1	2	4	4	44	85	154	255	154
12	0	0	0	0	1	2	6	21	96	118	185	215	169

In [137]:

```
plt.figure(figsize=(15,10))
sns.heatmap(month_wise_count, cmap='YlGnBu', annot=True, fmt="d")
```

Out[137]:

```
<AxesSubplot:xlabel='year_added', ylabel='month_added'>
```



In [ ]:

```

```

In [141]:

```
top_10_countries = df['country'].value_counts()[:10]
top_10_countries
```

Out[141]:

```
United States     2818
India              972
United Kingdom     419
Japan              245
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Egypt              106
Name: country, dtype: int64
```

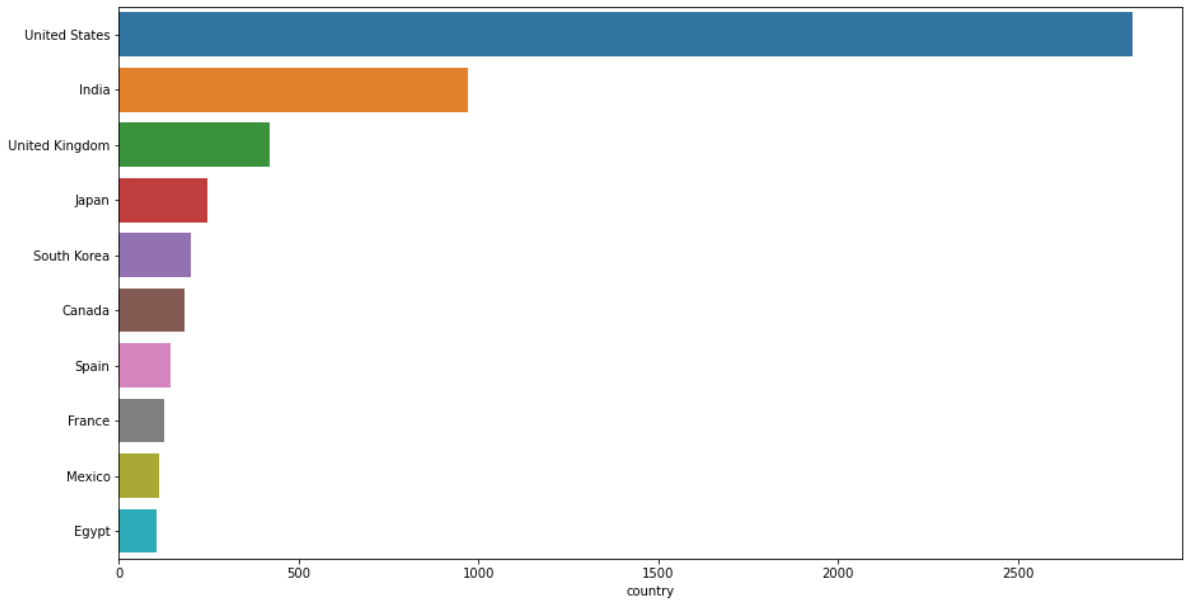
In [ ]:

```

```

In [144]:

```
plt.figure(figsize=(15,8))
sns.barplot(x=top_10_countries, y=top_10_countries.index)
plt.show()
```



In [ ]:

```

```

Actors

In [145]:

```
df.head(3)
```

Out[145]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	d
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	

In [ ]:

```

```

In [152]:

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

Out[152]:

```
1       Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...
2       Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...
4       Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...
5       Kate Siegel, Zach Gilford, Hamish Linklater, H...
6       Vanessa Hudgens, Kimiko Glenn, James Marsden, ...
                              ...
8801    Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...
8802    Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...
8804    Jesse Eisenberg, Woody Harrelson, Emma Stone, ...
8805    Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...
8806    Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...
Name: cast, Length: 7982, dtype: object
```

In [154]:

```
actors = " ".join(df['cast'].dropna())
```

In [161]:

```
actors= list(map(lambda x: x.strip(), actors.split(",")))
actors[:5]
```

Out[161]:

```
['Ama Qamata',
 'Khosi Ngema',
 'Gail Mabalane',
 'Thabang Molaba',
 'Dillon Windvogel']
```

In [162]:

```
from collections import Counter
```

In [164]:

```
Counter(actors).most_common(10)
```

Out[164]:

```
[('Anupam Kher', 37),
 ('Rupa Bhimani', 31),
 ('Takahiro Sakurai', 29),
 ('Julie Tejwani', 28),
 ('Om Puri', 27),
 ('Paresh Rawal', 25),
 ('Andrea Libman', 24),
 ('Jigna Bhardwaj', 23),
 ('Rajesh Kava', 22),
 ('Vincent Tong', 22)]
```

In [ ]:

```

```

In [166]:

```
df.head(3)
```

Out[166]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	d
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	

In [ ]:

```

```

In [171]:

```
df[df['type']=='TV Show']['duration'].value_counts()
```

Out[171]:

```
1     1793
2      425
3      199
4       95
5       65
6       33
7       23
8       17
9        9
10       7
13       3
12       2
11       2
15       2
17       1
Name: duration, dtype: int64
```

In [ ]:

Bivariate Analysis

In [ ]:

In [ ]:

In [ ]: