In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
import re
import plotly.express as px
import plotly.graph_objs as go
import plotly.figure_factory as ff

warnings.filterwarnings('ignore')
%matplotlib inline
```

In [2]:

```
netflix=pd.read_csv(r'C:\Users\DELL\Downloads\netflix1.csv')
```

In [3]:

```
netflix.head()
```

Out[3]:

	show_id	type	title	director	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Docu
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Int TV S
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV TV H
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	C Famil C
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	Ind Int

In [4]:

```
netflix.tail()
```

Out[4]:

	show_id	type	title	director	country	date_added	release_year	rating	duration	liste
8785	s8797	TV Show	Yunus Emre	Not Given	Turkey	1/17/2017	2016	TV-PG	2 Seasons	Internati TV Sh TV Dra
8786	s8798	TV Show	Zak Storm	Not Given	United States	9/13/2018	2016	TV-Y7	3 Seasons	Kids
8787	s8801	TV Show	Zindagi Gulzar Hai	Not Given	Pakistan	12/15/2016	2012	TV-PG	1 Season	Internati TV Sh Roma TV Sh T
8788	s8784	TV Show	Yoko	Not Given	Pakistan	6/23/2018	2016	TV-Y	1 Season	Kids
8789	s8786	TV Show	YOM	Not Given	Pakistan	6/7/2018	2016	TV-Y7	1 Season	Kids

In [5]:

```
netflix.shape
```

Out[5]:

```
(8790, 10)
```

In [6]:

```
netflix.isnull().sum()
```

Out[6]:

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

In [7]:

```
netflix.describe()
```

Out[7]:

	release_year
count	8790.000000
mean	2014.183163
std	8.825466
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

In [8]:

```
netflix.columns
```

Out[8]:

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

In [9]:

```
netflix.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
 4   country       8790 non-null   object
 5   date_added    8790 non-null   object
 6   release_year  8790 non-null   int64 
 7   rating        8790 non-null   object
 8   duration      8790 non-null   object
 9   listed_in     8790 non-null   object
dtypes: int64(1), object(9)
memory usage: 686.8+ KB
```

Inspect Missing values in the dataset

In [10]:

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

Out[10]:

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

In [11]:

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

Out[11]:

```
listed_in       0.0
duration        0.0
rating          0.0
release_year    0.0
date_added      0.0
country         0.0
director        0.0
title           0.0
type            0.0
show_id         0.0
dtype: float64
```

In [12]:

```
netflix["director"].value_counts()
```

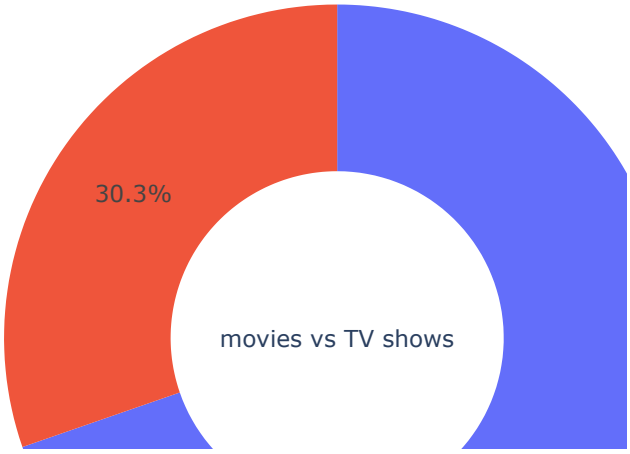
Out[12]:

```
Not Given                2588
Rajiv Chilaka              20
Alastair Fothergill        18
Raúl Campos, Jan Suter     18
Marcus Raboy               16
                         ...
Rai Yuvraj Bains            1
Todd Standing               1
Kees Van Oostrum            1
Adam Alleca                 1
Praveen Nischol             1
Name: director, Length: 4528, dtype: int64
```

movies vs TV Shows

In [13]:

```
#Checking Gender imbalance
go.Figure(data=[go.Pie(labels=netflix.type.value_counts(normalize=True).index,
                       values=netflix.type.value_counts(normalize=True).values,hole=.5,
                      title= 'movies vs TV shows')])
```



In [14]:

```
netflix.type.value_counts()
```

Out[14]:

```
Movie      6126
TV Show    2664
Name: type, dtype: int64
```

In [15]:

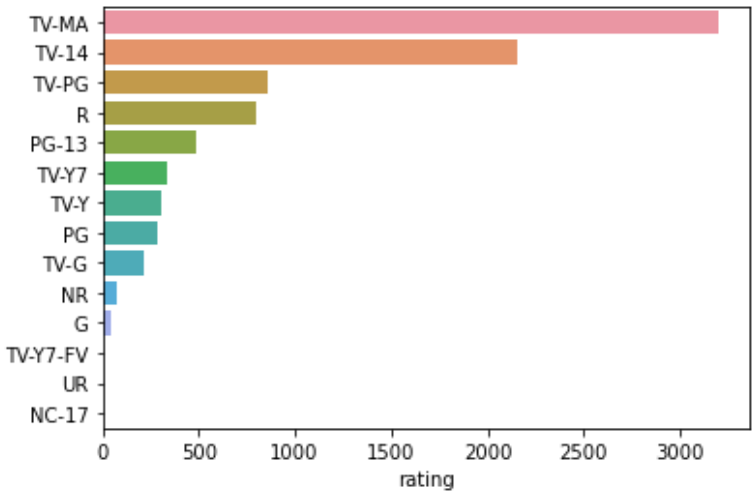
```
netflix.rating.value_counts()
```

Out[15]:

```
TV-MA       3205
TV-14       2157
TV-PG        861
R            799
PG-13        490
TV-Y7        333
TV-Y         306
PG           287
TV-G         220
NR            79
G             41
TV-Y7-FV       6
UR             3
NC-17          3
Name: rating, dtype: int64
```

In [16]:

```
sns.barplot(x=netflix.rating.value_counts(),y=netflix.rating.value_counts().index, data= netfl
plt.show()
```



In [17]:

```
netflix.country.value_counts()
```

Out[17]:

```
United States     3240
India             1057
United Kingdom     638
Pakistan           421
Not Given          287
                  ...
Paraguay             1
Puerto Rico          1
Guatemala            1
Luxembourg           1
Mozambique           1
Name: country, Length: 86, dtype: int64
```

In [18]:

```
netflix.country.value_counts().head(10)
```

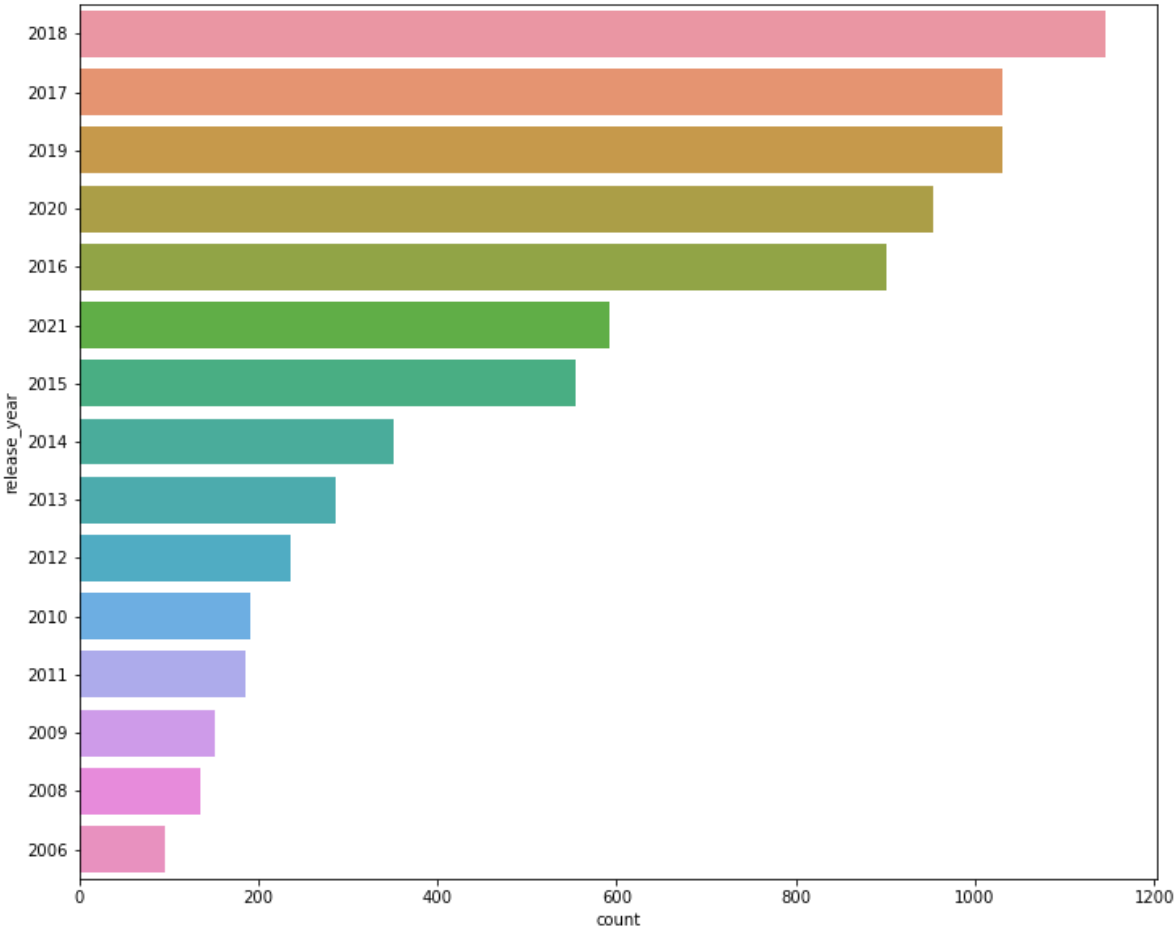
Out[18]:

```
United States     3240
India             1057
United Kingdom     638
Pakistan           421
Not Given          287
Canada             271
Japan              259
South Korea        214
France             213
Spain              182
Name: country, dtype: int64
```

year wise count

In [19]:

```
plt.figure(figsize=(12,10))
ax= sns.countplot(y="release_year",data=netflix,order=netflix.release_year.value_counts().inde
```



Top 10 Directors

In [20]:

```
netflix.director.value_counts().head(10)
```

Out[20]:

```
Not Given                 2588
Rajiv Chilaka               20
Alastair Fothergill         18
Raúl Campos, Jan Suter      18
Marcus Raboy                16
Suhas Kadav                 16
Jay Karas                   14
Cathy Garcia-Molina         13
Jay Chapman                 12
Martin Scorsese             12
Name: director, dtype: int64
```

In [21]:

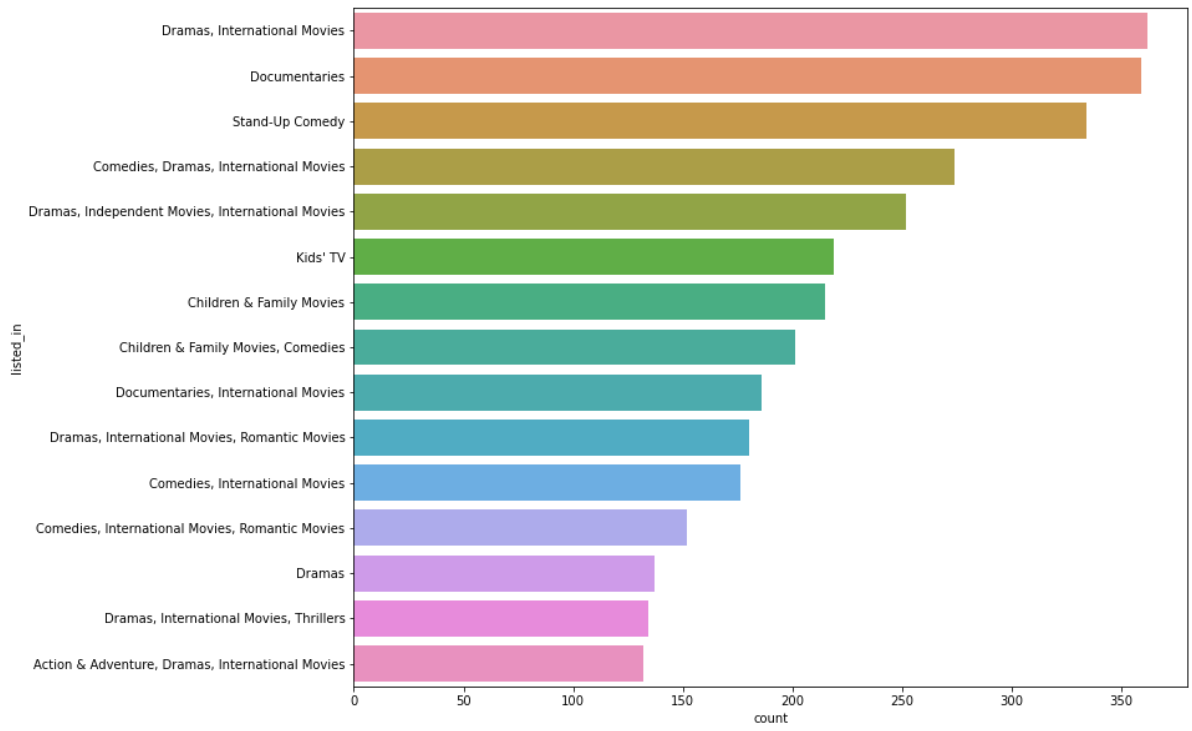
```
netflix.listed_in.value_counts().head(10)
```

Out[21]:

```
Dramas, International Movies                          362
Documentaries                                         359
Stand-Up Comedy                                       334
Comedies, Dramas, International Movies                274
Dramas, Independent Movies, International Movies      252
Kids' TV                                              219
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 [22]:

```
plt.figure(figsize=(12,10))
ax=sns.countplot(y="listed_in",data=netflix,order=netflix.listed_in.value_counts().index[0:15]
```



Handling Missing values

In [23]:

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

Out[23]:

```
listed_in       0.0
duration        0.0
rating          0.0
release_year    0.0
date_added      0.0
country         0.0
director        0.0
title           0.0
type            0.0
show_id         0.0
dtype: float64
```

In [24]:

```
round(netflix.isnull().sum())
```

Out[24]:

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

In [25]:

```
#dropping rows for small percentages for null
netflix.dropna(subset=["rating","duration"],axis=0,inplace=True)
```

In [26]:

```
netflix.shape
```

Out[26]:

```
(8790, 10)
```

In [27]:

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

Out[27]:

```
listed_in       0.0
duration        0.0
rating          0.0
release_year    0.0
date_added      0.0
country         0.0
director        0.0
title           0.0
type            0.0
show_id         0.0
dtype: float64
```

In [28]:

```
netflix.dropna(subset=["date_added"],axis=0,inplace=True)
```

In [29]:

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

Out[29]:

```
listed_in       0.0
duration        0.0
rating          0.0
release_year    0.0
date_added      0.0
country         0.0
director        0.0
title           0.0
type            0.0
show_id         0.0
dtype: float64
```

In [ ]:

```

```