NETFLIX EDA

In [1]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import seaborn as sns
sns.set()
```

TASK

1. Understand the data set, data types and missing values
2. Clean dataset and handle missing values
3. Perform data visualisation
4. Final Summary

In [2]:
```python
data = pd.read_csv(r"D:\project netflix\netflix_titles.csv")
```

In [3]:
```python
data.head()
```

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 I...

In [4]:
```python
data.shape #rows and collomns
```

Out[4]:
```
(8807, 12)
```

In [5]:
```python
data.describe() #basic details about the data
```

Out[5]:

	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

In [6]:
```python
data.info() #counts and data types of collomns
```

```
<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
```

Missing values

In [7]:
```python
data.isna().sum() #to check missing values
```

Out[7]:
```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

Adjust data type and fill missing values data type of date added doesnt make sense its data type should be date-time the following require filling of missing values:

1. director
2. cast
3. country
4. date_added
5. rating
6. duration

In [8]:
```python
#convert date_added data type from object to datetime64
data['date_added'] = pd.to_datetime(data['date_added'])
```

In [9]:
```python
data.head()
```

Out[9]:

	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	2021-09-25	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	2021-09-24	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	2021-09-24	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	2021-09-24	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	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Handling Missing Values

In [10]:
```python
data.fillna({'rating' : 'unavailable', 'cast' : 'unavailable', 'country' : 'unavailable', 'director' : 'unavailable'} , inplace = True)
data.isna().sum()
```

Out[10]:
```
show_id          0
type             0
title            0
director         0
cast             0
country          0
date_added      10
release_year     0
rating           0
duration         3
listed_in        0
description      0
dtype: int64
```

Replacing the missing values of date_added with most recent date from date_added.

In [11]:
```python
data[data.date_added.isnull()]
```

Out[11]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
6066	s6067	TV Show	A Young Doctor's Notebook and Other Stories	unavailable	Daniel Radcliffe, Jon Hamm, Adam Godley, Chris...	United Kingdom	NaT	2013	TV-MA	2 Seasons	British TV Shows, TV Comedies, TV Dramas	Set during the Russian Revolution, this comic ...
6174	s6175	TV Show	Anthony Bourdain: Parts Unknown	unavailable	Anthony Bourdain	United States	NaT	2018	TV-PG	5 Seasons	Docuseries	This CNN original series has chef Anthony Bour...
6795	s6796	TV Show	Frasier	unavailable	Kelsey Grammer, Jane Leeves, David Hyde Pierce...	United States	NaT	2003	TV-PG	11 Seasons	Classic & Cult TV, TV Comedies	Frasier Crane is a snooty but lovable Seattle ...
6806	s6807	TV Show	Friends	unavailable	Jennifer Aniston, Courteney Cox, Lisa Kudrow, ...	United States	NaT	2003	TV-14	10 Seasons	Classic & Cult TV, TV Comedies	This hit sitcom follows the merry misadventure...
6901	s6902	TV Show	Gunslinger Girl	unavailable	Yuuka Nanri, Kanako Mitsuhashi, Eri Sendai, Am...	Japan	NaT	2008	TV-14	2 Seasons	Anime Series, Crime TV Shows	On the surface, the Social Welfare Agency appe...
7196	s7197	TV Show	Kikoriki	unavailable	Igor Dmitriev	unavailable	NaT	2010	TV-Y	2 Seasons	Kids' TV	A wacky rabbit and his gang of animal pals hav...
7254	s7255	TV Show	La Familia P. Luche	unavailable	Eugenio Derbez, Consuelo Duval, Luis Manuel Áv...	United States	NaT	2012	TV-14	3 Seasons	International TV Shows, Spanish-Language TV Sh...	This irreverent sitcom featues Ludovico, Feder...
7406	s7407	TV Show	Maron	unavailable	Marc Maron, Judd Hirsch, Josh Brener, Nora Zeh...	United States	NaT	2016	TV-MA	4 Seasons	TV Comedies	Marc Maron stars as Marc Maron, who interviews...
7847	s7848	TV Show	Red vs. Blue	unavailable	Burnie Burns, Jason Saldaña, Gustavo Sorola, G...	United States	NaT	2015	NR	13 Seasons	TV Action & Adventure, TV Comedies, TV Sci-Fi ...	This parody of first-person shooter games, mil...
8182	s8183	TV Show	The Adventures of Figaro Pho	unavailable	Luke Jurevicius, Craig Behenna, Charlotte Haml...	Australia	NaT	2015	TV-Y7	2 Seasons	Kids' TV, TV Comedies	Imagine your worst fears, then multiply them: ...

In [12]:
```python
most_recent_entry_date=data['date_added'].max()
data.fillna({'date_added' : most_recent_entry_date},inplace = True)
```

In [13]:
```python
data[data.show_id == 's6067']
```

Out[13]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
6066	s6067	TV Show	A Young Doctor's Notebook and Other Stories	unavailable	Daniel Radcliffe, Jon Hamm, Adam Godley, Chris...	United Kingdom	2021-09-25	2013	TV-MA	2 Seasons	British TV Shows, TV Comedies, TV Dramas	Set during the Russian Revolution, this comic ...

In [14]:
```python
data[data.duration.isnull()]
```

Out[14]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	74 min	NaN	Movies	Louis C.K. muses on religion, eternal love, gi...
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	84 min	NaN	Movies	Emmy-winning comedy writer Louis C.K. brings h...
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	The comic puts his trademark hilarious/thought...

All movies with missing durtion are movies by Louisc C.K.

In [15]:
```python
data[data.director == 'Louis C.K.']
```

Out[15]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	74 min	NaN	Movies	Louis C.K. muses on religion, eternal love, gi...
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	84 min	NaN	Movies	Emmy-winning comedy writer Louis C.K. brings h...
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	The comic puts his trademark hilarious/thought...

There is a mismatch in rating and duration so we replace data

In [16]:
```python
#loc helps us easily access the colomns by name
data.loc[data['director'] == 'Louis C.K.', 'duration']= data['rating']
data[data.director == 'Louis C.K.'].head()
```

Out[16]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	74 min	74 min	Movies	Louis C.K. muses on religion, eternal love, gi...
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	84 min	84 min	Movies	Emmy-winning comedy writer Louis C.K. brings h...
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	66 min	Movies	The comic puts his trademark hilarious/thought...

In [17]:
```python
data.loc[data['director']=='Louis C.K.','rating'] = 'unavailable'
data[data.director == 'Louis C.K.'].head()
```

Out[17]:

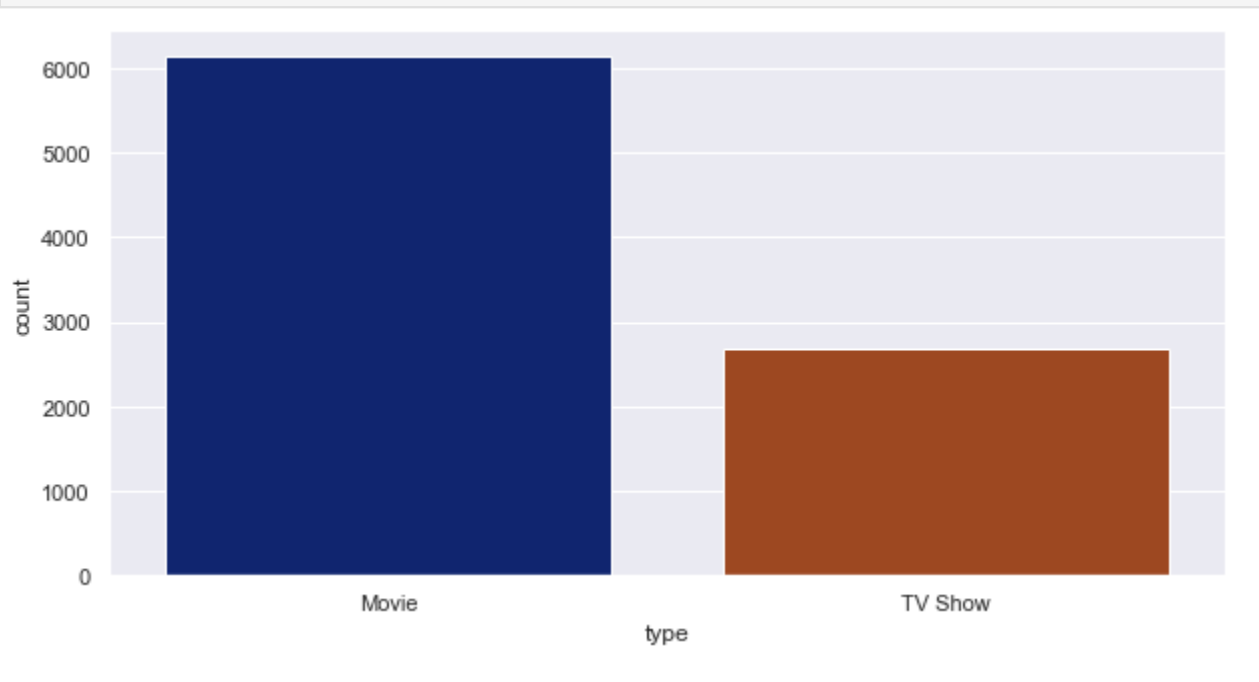
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	unavailable	74 min	Movies	Louis C.K. muses on religion, eternal love, gi...
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	unavailable	84 min	Movies	Emmy-winning comedy writer Louis C.K. brings h...
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	2016-08-15	2015	unavailable	66 min	Movies	The comic puts his trademark hilarious/thought...

Visualisation

In [18]:
```python
data.type.value_counts() #value_count shows the counts of different categories in a given collomn
```

Out[18]:
```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

In [19]:
```python
plt.rcParams['figure.figsize'] = 10,5
sns.countplot(data=data, x="type",palette="dark")
plt.show()
```



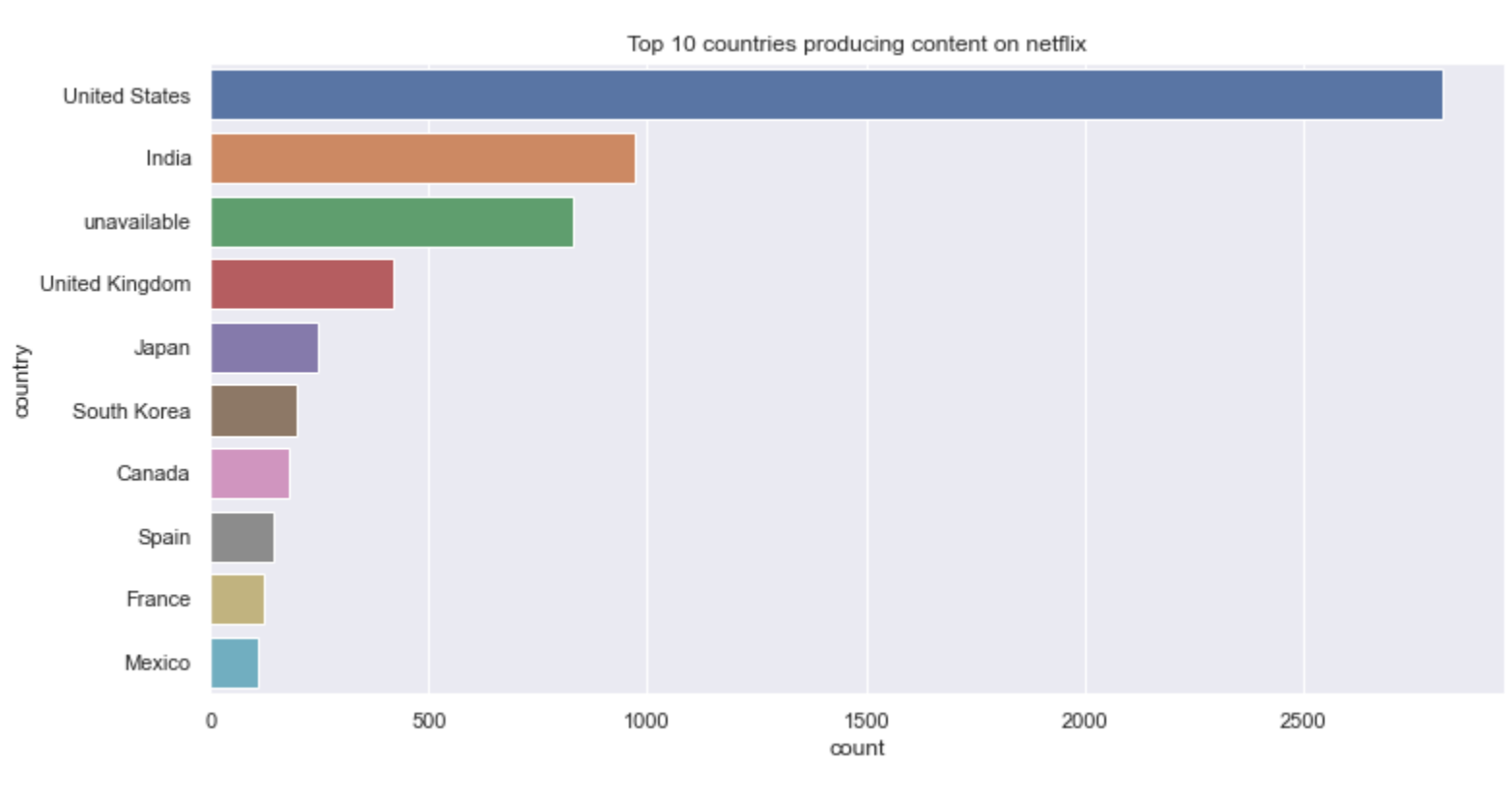
more no. of movies than tv shows

In [20]:
```python
data['country'].value_counts().head(10)
```

Out[20]:
```
United States     2818
India              972
unavailable        831
United Kingdom     419
Japan              245
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Name: country, dtype: int64
```

In [21]:
```python
plt.figure(figsize = (12,6))
sns.countplot(y='country',order = data['country'].value_counts().index[0:10],data = data)
plt.title('Top 10 countries producing content on netflix')
```

Out[21]:
```
Text(0.5, 1.0, 'Top 10 countries producing content on netflix')
```

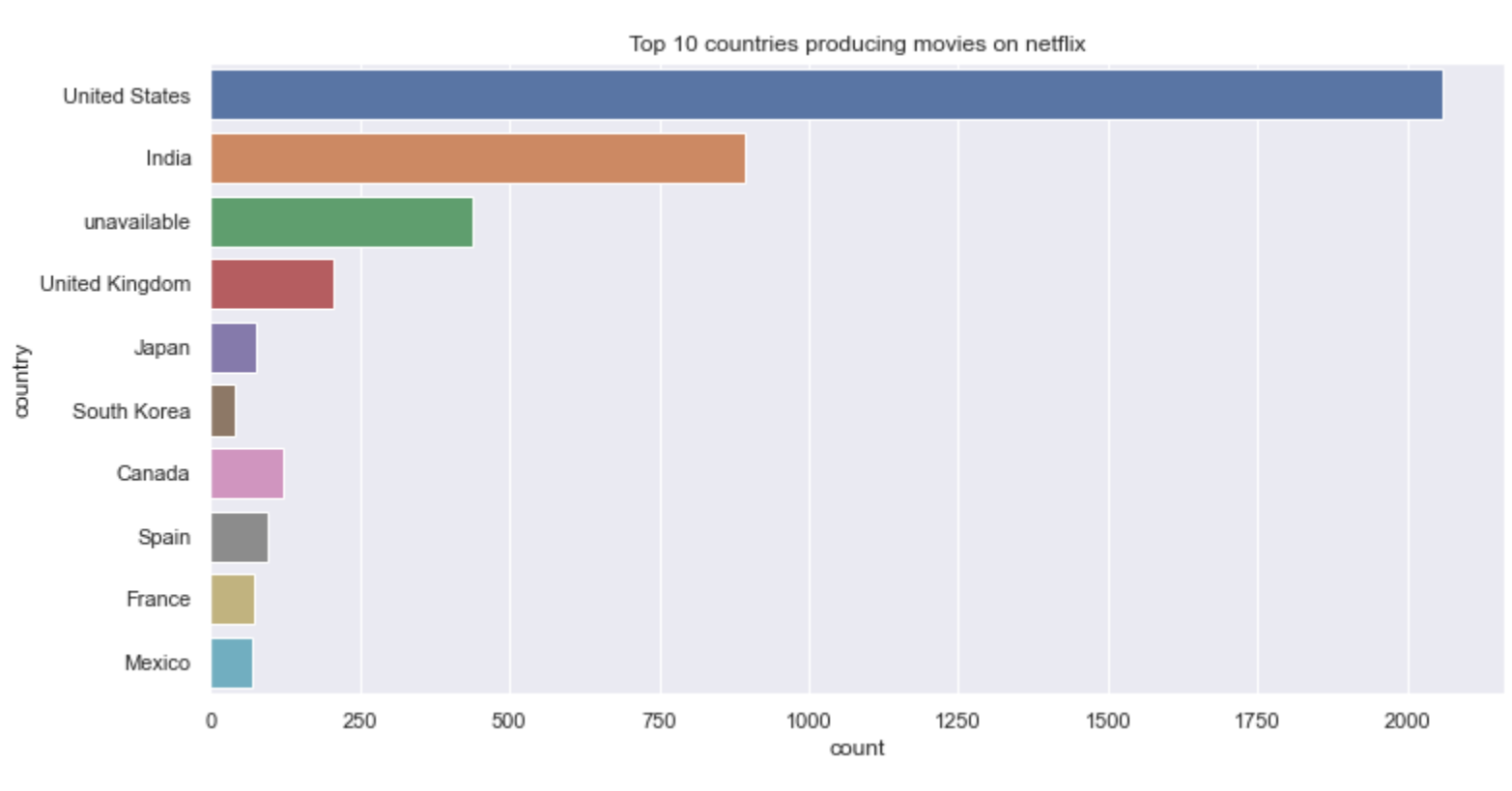


United States has the maximum content

In [22]:
```python
#now checking country based on type of content
movie_countries = data[data['type']=='Movie']
tv_shows_countries = data[data['type']=='TV Show']
```

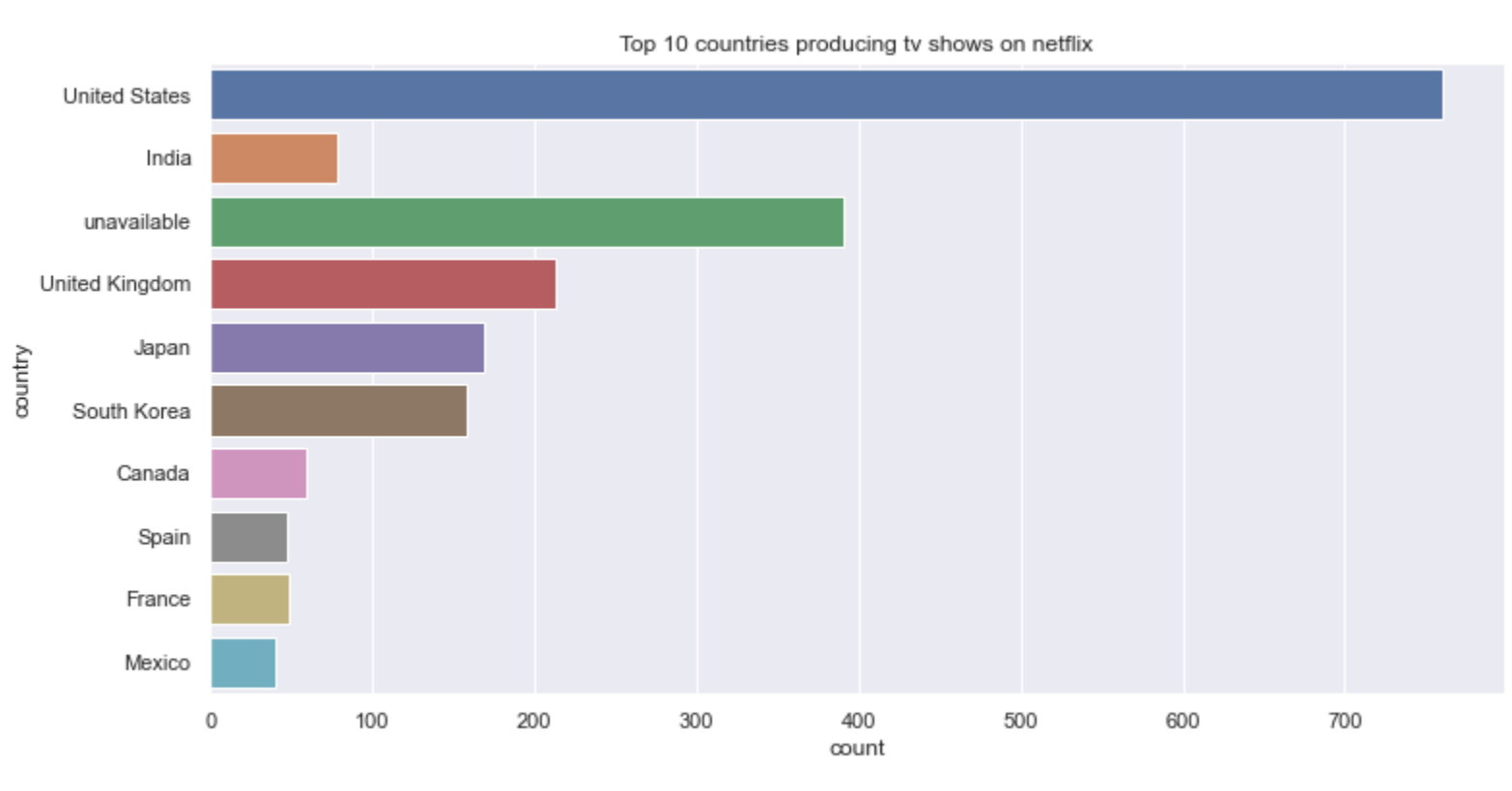
In [23]:
```python
plt.figure(figsize = (12,6))
sns.countplot(y='country',order = data['country'].value_counts().index[0:10],data = movie_countries)
plt.title('Top 10 countries producing movies on netflix')
```

Out[23]:
```
Text(0.5, 1.0, 'Top 10 countries producing movies on netflix')
```



In [24]:
```python
plt.figure(figsize = (12,6))
sns.countplot(y='country',order = data['country'].value_counts().index[0:10],data = tv_shows_countries)
plt.title('Top 10 countries producing tv shows on netflix')
```

Out[24]:
```
Text(0.5, 1.0, 'Top 10 countries producing tv shows on netflix')
```



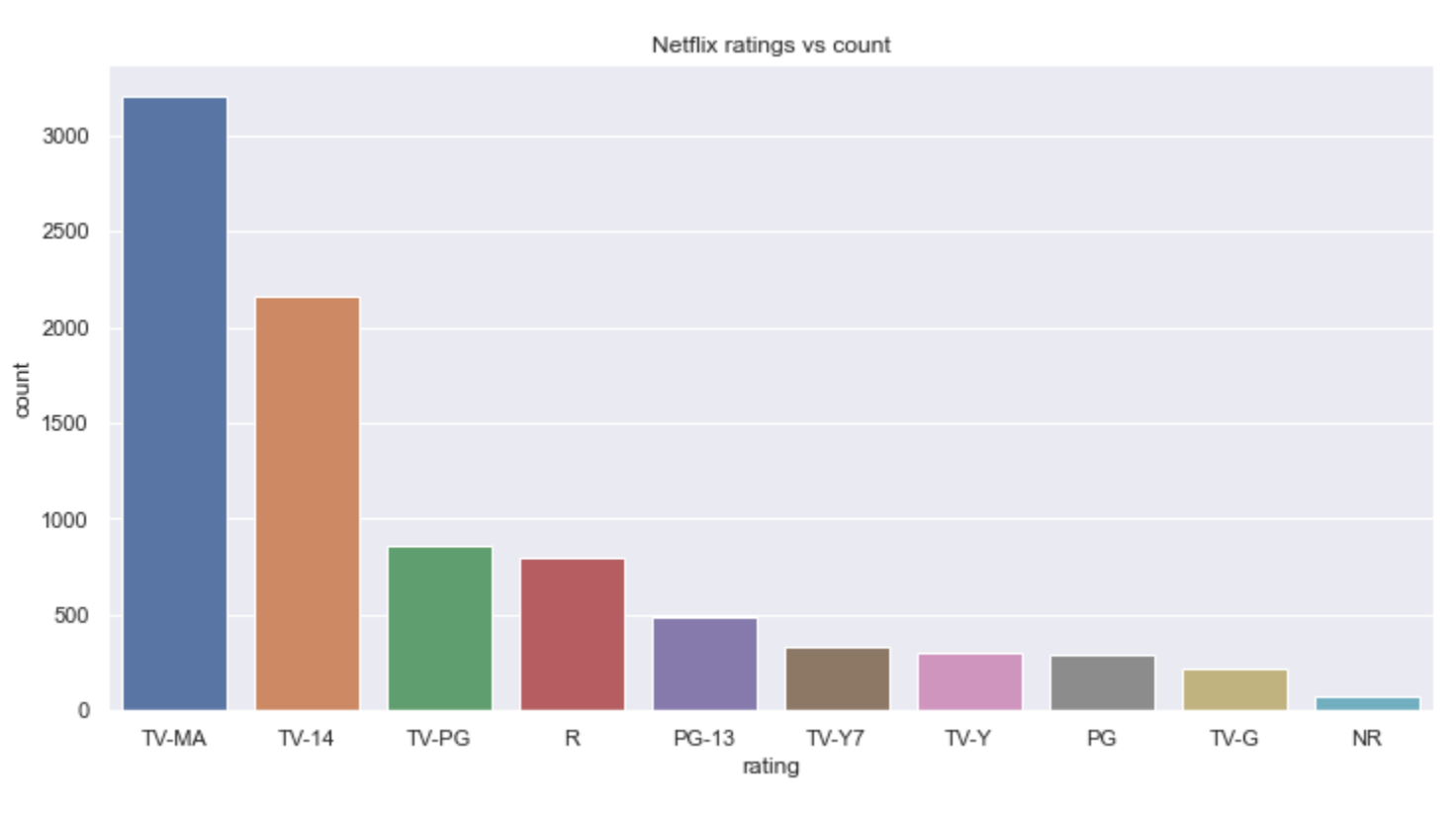
check ratings

In [25]:
```python
data.rating.value_counts()
```

Out[25]:
```
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
unavailable       7
TV-Y7-FV          6
NC-17             3
UR                3
Name: rating, dtype: int64
```

In [26]:
```python
plt.figure(figsize = (12,6))
sns.countplot(x='rating',order = data['rating'].value_counts().index[0:10],data = data)
plt.title('Netflix ratings vs count')
```

Out[26]:
```
Text(0.5, 1.0, 'Netflix ratings vs count')
```



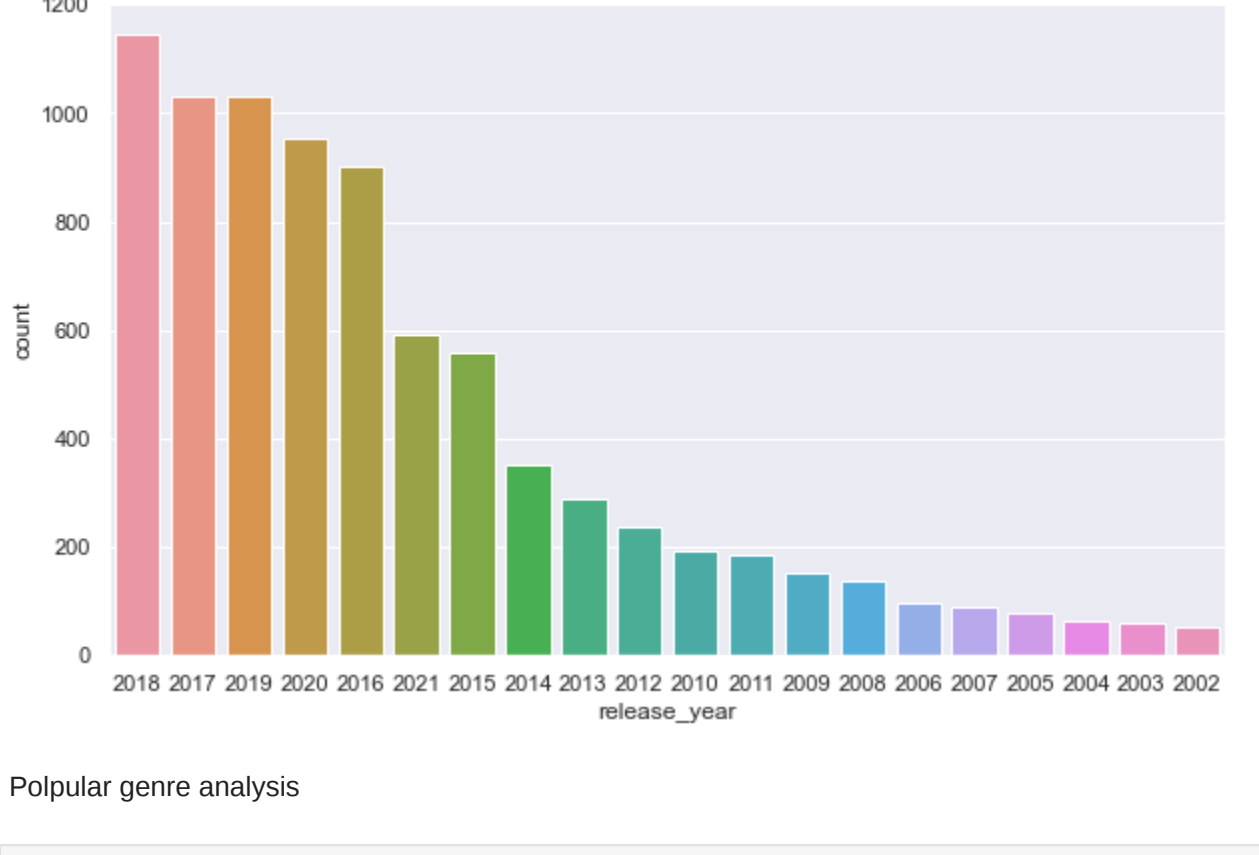
Most of the shows has TV-MA and TV-14 ratings.

In [27]:
```python
data.release_year.value_counts()[:20]
```

Out[27]:
```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
2021     592
2015     560
2014     352
2013     288
2012     237
2010     194
2011     185
2009     152
2008     136
2006      96
2007      88
2005      80
2004      64
2003      61
2002      51
Name: release_year, dtype: int64
```

In [28]:
```python
plt.figure(figsize = (10,6))
sns.countplot(x='release_year',order = data['release_year'].value_counts().index[0:20],data = data)
plt.title('Content release in Years on Netflix vs count')
```

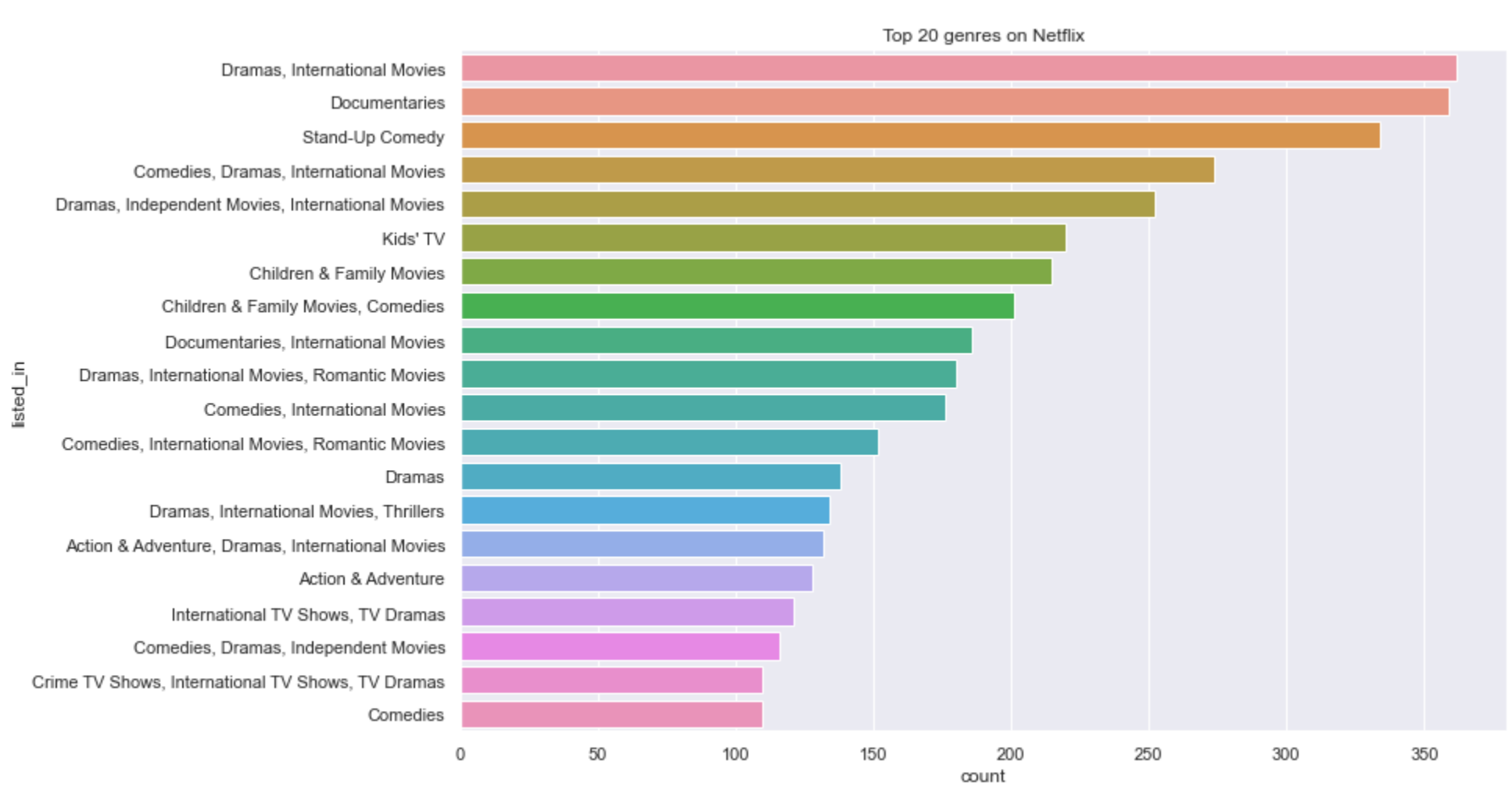
Out[28]:
```
Text(0.5, 1.0, 'Content release in Years on Netflix vs count')
```



Polpular genre analysis

In [29]:
```python
plt.figure(figsize = (12,8))
sns.countplot(y='listed_in',order = data['listed_in'].value_counts().index[0:20],data = data)
plt.title('Top 20 genres on Netflix')
```

Out[29]:
```
Text(0.5, 1.0, 'Top 20 genres on Netflix')
```



Summary

Some of the important information that we can dig out from the data are:

1. Netflix has more movies than tv shows.
2. United States is the largest producer of content (movies and tv shows) followed by India.
3. Most of the content on netflix is rated TV-MA followed by TV_14.
4. Maximum content was created in 2017-2020 in which 2018 had maximum content released.
5. top 3 genres will be international movies, documentries and stand up comedy

In [ ]: