Problem Statement:-

1) Analyzing the netflix data to get an idea about type of movies and shows released and their country origin

In [3]:
```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.graph_objects as go
import plotly.express as px
import pandas_profiling
#from pandas_profiling import ProfileReport
import seaborn as sns
%matplotlib inline

import matplotlib
sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9,5)
```

Data Acquisition and Description:-

In [4]:
```
netflix_df = pd.read_csv('netflix_titles.csv.zip' )
netflix_df
```

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Docum
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	Inte TV S Dra M
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	(Inte TV S
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Do R
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	Inte T Rom Sho

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	K S C
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	C Horro
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Ch Family C
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Inte Movi &

8807 rows × 12 columns

In [5]:
```
netflix_df.sample(5)
```

Out[5]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	lis
7982	s7983	TV Show	Sensitive Skin	NaN	Kim Cattrall, Don McKellar, Nicolas Wright, Jo...	Canada	December 1, 2019	2016	TV-MA	1 Season	Cor
2814	s2815	Movie	Steam Team to the Rescue	Joey So	Joseph May, Keith Wickham, Yvonne Grundy, Jule...	NaN	March 15, 2020	2019	TV-Y	23 min	Chil

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	lis
99	s100	TV Show	On the Verge	NaN	Julie Delpy, Elisabeth Shue, Sarah Jones, Alex...	France, United States	September 7, 2021	2021	TV-MA	1 Season	Con TV D
7192	s7193	Movie	Kickboxer: Vengeance	John Stockwell	Alain Moussi, Jean-Claude Van Damme, Dave Baut...	United States	December 8, 2016	2016	TV-MA	90 min	Ac Adv
3641	s3642	Movie	The Son	Sebastián Schindel	Joaquín Furriel, Martina Gusmán, Luciano Cácer...	Argentina	July 26, 2019	2019	TV-MA	93 min	D Indepe M Intern M

In [9]:
```
netflix_df.shape
```

Out[9]:
```
(8804, 12)
```

Data Description:-

In [10]:
```
netflix_df.describe()
```

Out[10]:

	release_year
count	8804.000000
mean	2014.180259
std	8.820647
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

Data Information:-

In [11]:
```
netflix_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8804 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8804 non-null   object
 1   type          8804 non-null   object
```

```
 2   title         8804 non-null   object
 3   director      6170 non-null   object
 4   cast          7979 non-null   object
 5   country       7973 non-null   object
 6   date_added    8794 non-null   object
 7   release_year  8804 non-null   int64
 8   rating        8800 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8804 non-null   object
 11  description   8804 non-null   object
dtypes: int64(1), object(11)
memory usage: 894.2+ KB
```

Data Pre-profiling:- 1) There are 4304 missing cells. 2) 11 Categorical values and 2 numeric values

In [14]:

```
netflix_Profile=pandas_profiling.ProfileReport(netflix_df)
netflix_Profile.to_file("netflixdata_Before_Processing.html")
netflix_Profile
```

Overview

Dataset statistics

Number of variables	13
Number of observations	8804
Missing cells	4304
Missing cells (%)	3.8%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	894.3 KiB
Average record size in memory	104.0 B

Variable types

Numeric	2
Categorical	11

Alerts

`show_id` has a high cardinality: 8804 distinct values	**High cardinality**
`title` has a high cardinality: 8804 distinct values	**High cardinality**
`director` has a high cardinality: 4527 distinct values	**High cardinality**
`cast` has a high cardinality: 7691 distinct values	**High cardinality**

Out[14]:

In [17]:
```
netflix_df.dropna(inplace=True)
```

In [15]:
```
#drop duplicate value
netflix_df.drop_duplicates(inplace=True)
netflix_df.shape
```

Out[15]:
```
(8804, 12)
```

In [16]:
```
#Tofind NAN values
```

```
netflix_df.isna().any()
```

Out[16]:
```
show_id         False
type            False
title           False
director         True
cast             True
country          True
date_added       True
release_year    False
rating           True
duration        False
listed_in       False
description     False
dtype: bool
```

In [6]:
```
# drop 3 incorrect values in rating
netflix_df.drop(netflix_df[netflix_df['rating']=='66 min'].index, inplace = True)
netflix_df.drop(netflix_df[netflix_df['rating']=='74 min'].index, inplace = True)
netflix_df.drop(netflix_df[netflix_df['rating']=='84 min'].index, inplace = True)
netflix_df['rating'].value_counts()
```

Out[6]:
```
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
Name: rating, dtype: int64
```

Data Post Profiling:-

In [18]:
```
netflix_Profile=pandas_profiling.ProfileReport(netflix_df)
netflix_Profile.to_file("netflixdata_Post_Processing.html")
netflix_Profile
```

Overview

Dataset statistics

Number of variables	13
Number of observations	5332
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	541.7 KiB
Average record size in memory	104.0 B

Variable types

Numeric	2
Categorical	11

Alerts

`show_id` has a high cardinality: 5332 distinct values	**High cardinality**
`title` has a high cardinality: 5332 distinct values	**High cardinality**
`director` has a high cardinality: 3945 distinct values	**High cardinality**
`cast` has a high cardinality: 5200 distinct values	**High cardinality**

Out[18]:

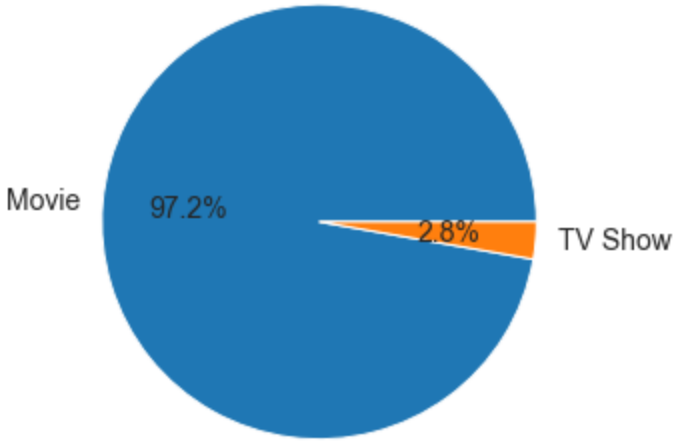
Are movies mostly streamed on netflix or TV shows?

In [19]:
```
#More number of movies are watched on netflix as compared to TV Shows

y=netflix_df['type'].value_counts()
y
mylabels = ["Movie", "TV Show"]
plt.pie(y, labels = mylabels, autopct='%1.1f%%')
```

Out[19]:
```
([<matplotlib.patches.Wedge at 0x171da8f6e20>,
  <matplotlib.patches.Wedge at 0x171dcac0df0>],
```

```
[Text(-1.0958766901652248, 0.09515398022422511, 'Movie'),
 Text(1.0958766840403136, -0.09515405076404879, 'TV Show')],
[Text(-0.5977509219083044, 0.051902171031395515, '97.2%'),
 Text(0.5977509185674437, -0.051902209507662965, '2.8%')])
```



Which country releases maximum number of movies and shows on netflix?

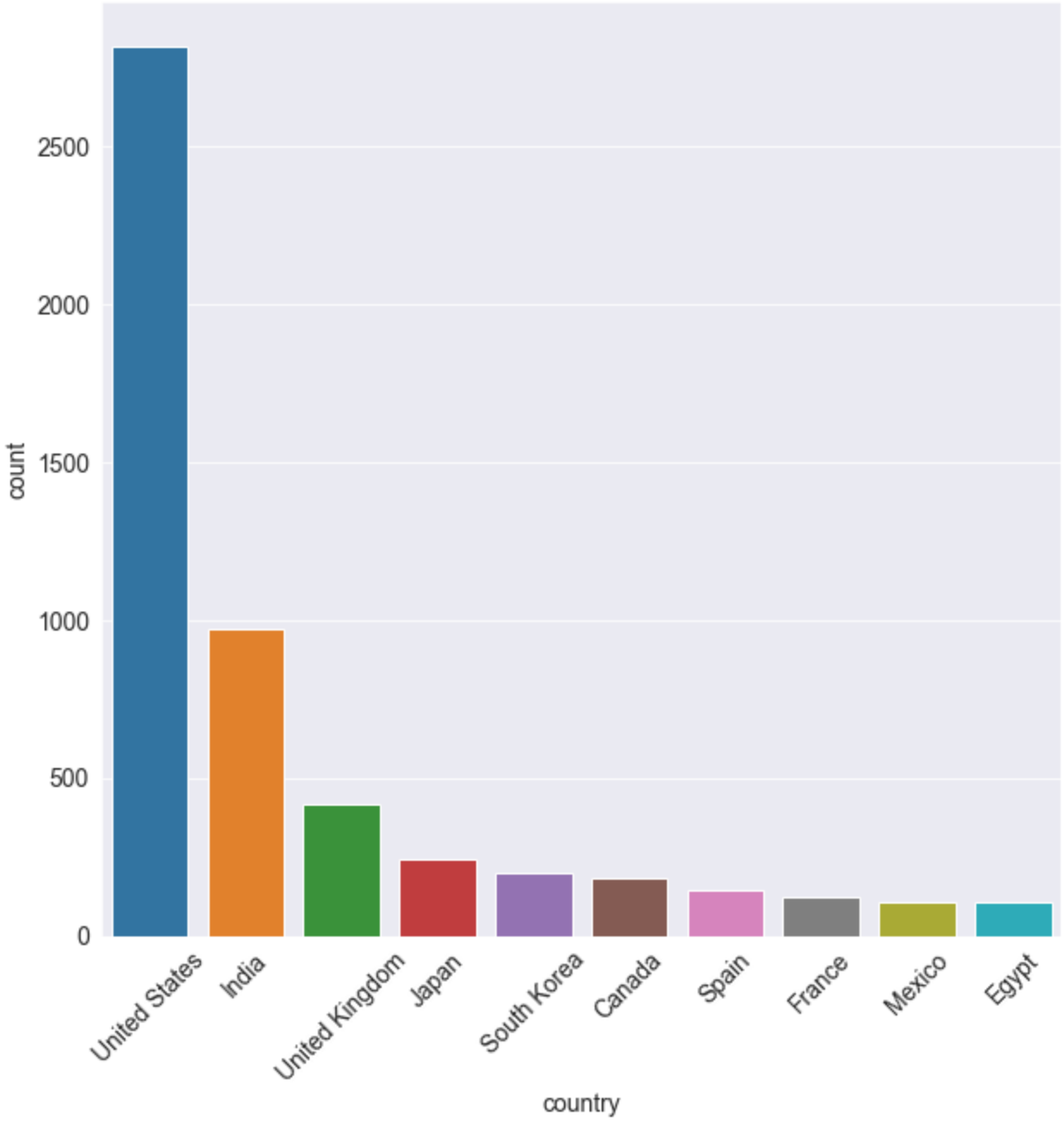
In [7]:
```
##Most movies or shows originated from United States

top_country= netflix_df['country'].value_counts()
top_country
```

Out[7]:
```
United States                                2818
India                                         972
United Kingdom                                419
Japan                                         245
South Korea                                   199
                                             ...
Romania, Bulgaria, Hungary                      1
Uruguay, Guatemala                              1
France, Senegal, Belgium                        1
Mexico, United States, Spain, Colombia          1
United Arab Emirates, Jordan                    1
Name: country, Length: 748, dtype: int64
```

In [8]:
```
plt.figure(figsize=(10,10))
country = sns.countplot(x='country',data=netflix_df, order=netflix_df['country'].value_cou
plt.xticks(rotation=45)
```

Out[8]:
```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]),
 [Text(0, 0, 'United States'),
  Text(1, 0, 'India'),
  Text(2, 0, 'United Kingdom'),
  Text(3, 0, 'Japan'),
  Text(4, 0, 'South Korea'),
  Text(5, 0, 'Canada'),
  Text(6, 0, 'Spain'),
  Text(7, 0, 'France'),
  Text(8, 0, 'Mexico'),
  Text(9, 0, 'Egypt')])
```



Which type of movies were released in recent years?

In [86]:
```
##TV-MA rated movies were mostly released in recent years

movie_release_year=netflix_df.groupby('release_year')['rating'].value_counts(ascending=Tru
movie_release_year.tail(50)
```

Out[86]:
```
release_year  rating
2017          TV-G         26
              TV-Y         31
              PG-13        32
              TV-Y7        37
              R            73
              TV-PG       111
              TV-14       251
              TV-MA       451
2018          NC-17         1
              NR            1
              TV-Y7-FV      1
              G             2
              TV-G         26
              PG-13        30
```

```
                PG          31
                TV-Y7       40
                TV-Y        41
                R           52
                TV-PG      105
                TV-14      268
                TV-MA      549
2019            G            1
                PG          12
                PG-13       19
                TV-G        23
                TV-Y7       36
                R           39
                TV-Y        50
                TV-PG       98
                TV-14      252
                TV-MA      500
2020            G            1
                PG          15
                PG-13       21
                TV-Y7       41
                TV-G        45
                R           48
                TV-Y        59
                TV-PG       80
                TV-14      174
                TV-MA      469
2021            PG          11
                PG-13       14
                R           21
                TV-G        21
                TV-Y        26
                TV-Y7       33
                TV-PG       45
                TV-14      151
                TV-MA      270
Name: rating, dtype: int64
```

In [42]:

```
rating =netflix_df['rating'].value_counts().reset_index().rename(columns={'index':'rating'
rating
```

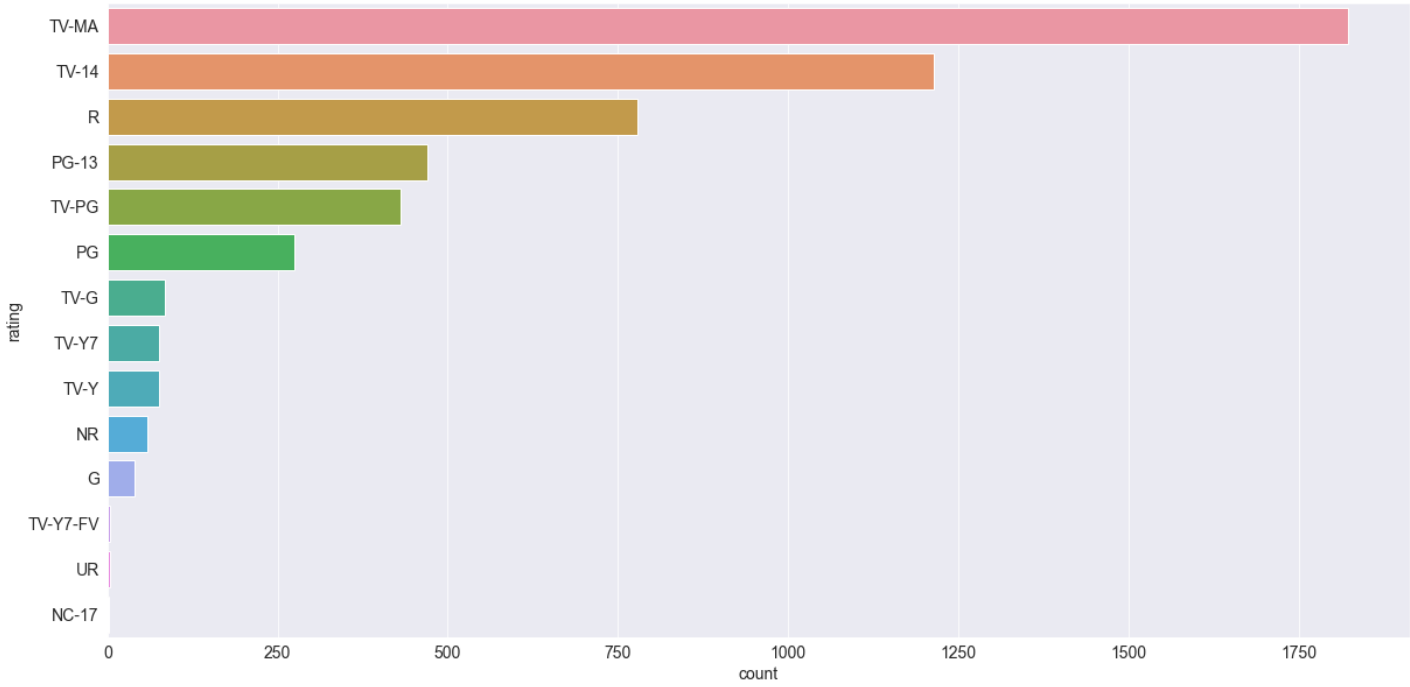
Out[42]:

	rating	count
0	TV-MA	1822
1	TV-14	1214
2	R	778
3	PG-13	470
4	TV-PG	431
5	PG	275
6	TV-G	84
7	TV-Y7	76
8	TV-Y	76
9	NR	58
10	G	40
11	TV-Y7-FV	3

	rating	count
12	UR	3
13	NC-17	2

In [84]:
```
plt.figure(figsize=(20,10))
sns.barplot(x='count', y='rating', data=rating)
```

Out[84]:
```
<AxesSubplot:xlabel='count', ylabel='rating'>
```



In which year there were maximum number of releases?

In [30]:
```
release= netflix_df['release_year'].value_counts()
release.head(15)
```

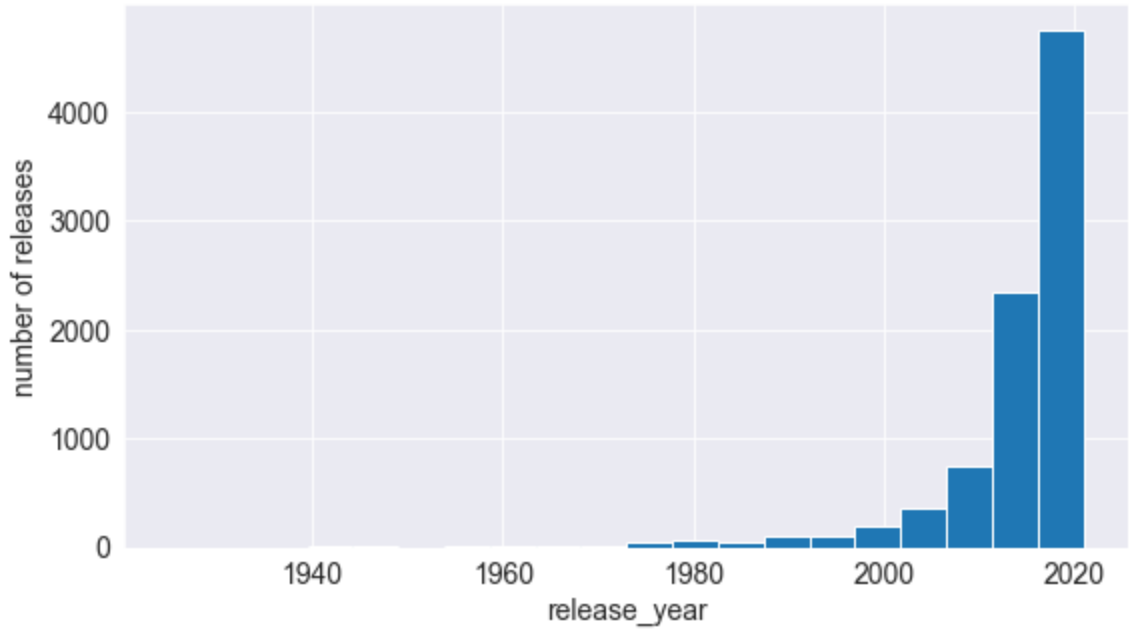
Out[30]:
```
2017    657
2018    648
2016    577
2019    519
2020    442
2015    349
2014    242
2013    197
2012    163
2021    161
2010    140
2011    135
2009    112
2008    110
2006     83
Name: release_year, dtype: int64
```

In [9]:
```
##between 2010 and 2020 there were maxium number of movie and shows released

data=netflix_df['release_year']
data
plt.hist(data, bins = 20)
plt.xlabel('release_year')
plt.ylabel('number of releases')
```

Out[9]:

```
Text(0, 0.5, 'number of releases')
```



Which genre of movie is the most popular?

In [6]:

```
netflix_df['listed_in'] = netflix_df['listed_in'].apply(lambda x: x.split(",")[0])
netflix_df.head(20)
```

Out[6]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Doc
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	I
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	I

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 Season	
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	NaN	September 24, 2021	2021	PG	91 min	Fa
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min	
10	s11	TV Show	Vendetta: Truth, Lies and The Mafia	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	
11	s12	TV Show	Bangkok Breaking	Kongkiat Komesiri	Sukollawat Kanarot, Sushar Manaying, Pavarit M...	NaN	September 23, 2021	2021	TV-MA	1 Season	
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	September 23, 2021	2021	TV-MA	127 min	
13	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Klara Castanho, Lucca Picon, Júlia Gomes, Marc...	NaN	September 22, 2021	2021	TV-PG	91 min	Fa

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
14	s15	TV Show	Crime Stories: India Detectives	NaN	NaN	NaN	September 22, 2021	2021	TV-MA	1 Season	
15	s16	TV Show	Dear White People	NaN	Logan Browning, Brandon P. Bell, DeRon Horton,...	United States	September 22, 2021	2021	TV-MA	4 Seasons	T
16	s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in ...	Pedro de Echave García, Pablo Azorín Williams	NaN	NaN	September 22, 2021	2020	TV-MA	67 min	Do
17	s18	TV Show	Falsa identidad	NaN	Luis Ernesto Franco, Camila Sodi, Sergio Goyri...	Mexico	September 22, 2021	2020	TV-MA	2 Seasons	
18	s19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Marshall-Green, Robert Joh...	NaN	September 22, 2021	2021	TV-14	94 min	
19	s20	TV Show	Jaguar	NaN	Blanca Suárez, Iván Marcos, Óscar Casas, Adriá...	NaN	September 22, 2021	2021	TV-MA	1 Season	I

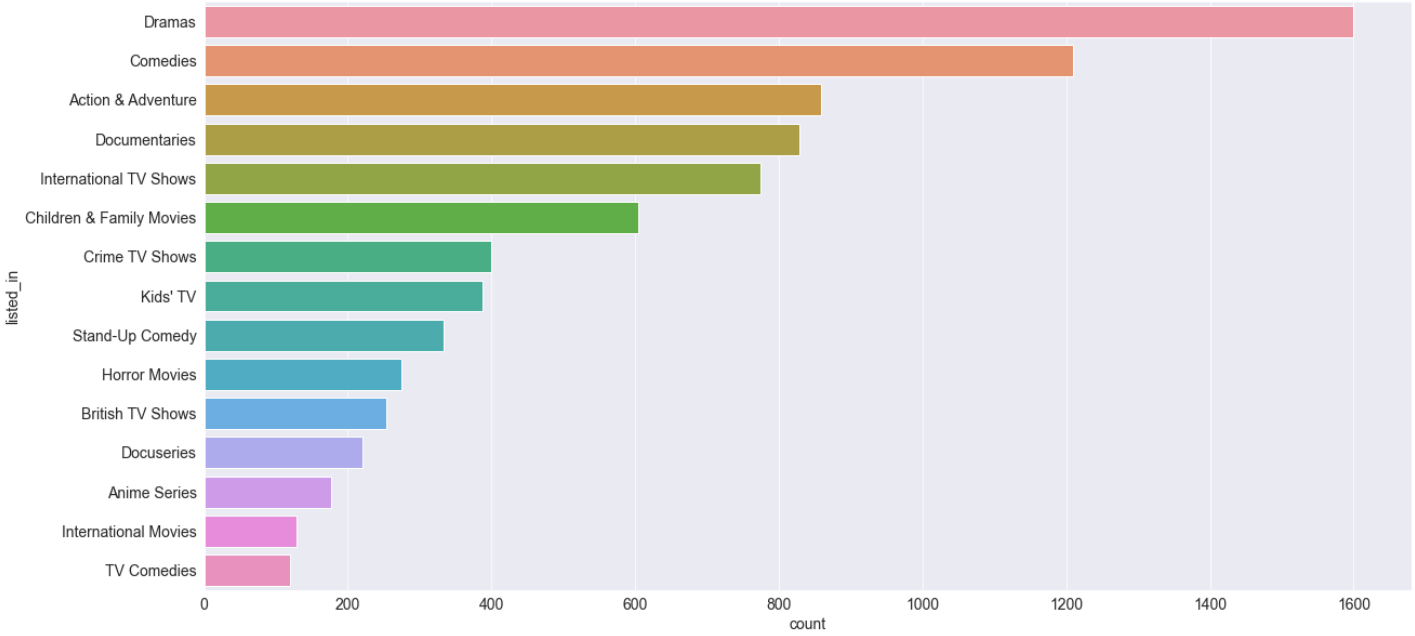
In [7]:

```
# Dramas were the most popular type of genre streamed on netflix
listed_in =netflix_df['listed_in'].value_counts().reset_index().rename(columns={'index':'
listed_in

plt.figure(figsize=(20,10))
sns.barplot(x='count', y='listed_in', data=listed_in.head(15))
```

Out[7]:

```
<AxesSubplot:xlabel='count', ylabel='listed_in'>
```



In which month most of the movies/ shows were released?

In [11]:
```
netflix_df['date_added'] = pd.to_datetime(netflix_df['date_added'])
netflix_df['Month_name'] = netflix_df['date_added'].dt.month_name()
netflix_df
```

Out[11]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Docum
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	Inte T
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	(
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Do
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	Inte T

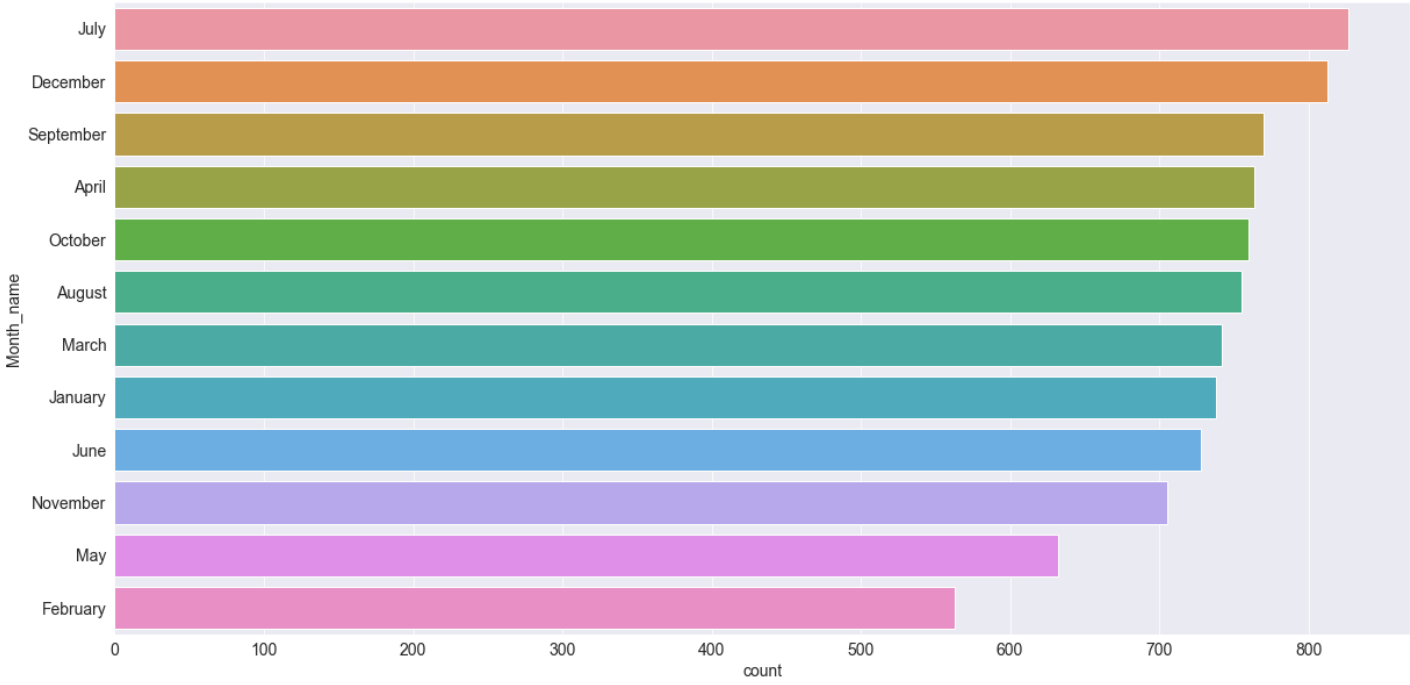
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
...	...	...	...	...	...	...	...	...	...	...	
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	R	158 min	Cul
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018	TV-Y7	2 Seasons	
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	R	88 min	C
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006	PG	88 min	Ch Famil
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-14	111 min	

8807 rows × 13 columns

In [12]:
```
# July and December are most suitable for releases
Month_name =netflix_df['Month_name'].value_counts().reset_index().rename(columns={'index':
Month_name

plt.figure(figsize=(20,10))
sns.barplot(x='count', y='Month_name', data=Month_name.head(15))
```

Out[12]:
```
<AxesSubplot:xlabel='count', ylabel='Month_name'>
```



In [14]:

```
mont= netflix_df['Month_name'].value_counts()
mont
```

Out[14]:

```
July          827
December      813
September     770
April         764
October       760
August        755
March         742
January       738
June          728
November      705
May           632
February      563
Name: Month_name, dtype: int64
```

In [ ]: