NETFLIX

Type *Markdown* and LaTeX: α^2

In [27]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Loading dataset

In [28]:
```
df=pd.read_csv("C:/Users/NAMRITA/Downloads/scaler/netflix_data.csv")
```

Some basic operations over dataset

In [29]:
```
df.shape      # to show the no. of rows and columns
```

Out[29]:
```
(8807, 12)
```

In [30]:
```
df.head()          # to show top 5 rows of the dataset
```

Out[30]:

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

In [31]:
```
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 [32]:
```
df['type'] = df['type'].astype('category')            #conversion of categori
df['country'] = df['country'].astype('category')
df['rating'] = df['rating'].astype('category')
```

In [33]:
```
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   category
 2   title         8807 non-null   object  
 3   director      6173 non-null   object  
 4   cast          7982 non-null   object  
 5   country       7976 non-null   category
 6   date_added    8797 non-null   object  
 7   release_year  8807 non-null   int64   
 8   rating        8803 non-null   category
 9   duration      8804 non-null   object  
 10  listed_in     8807 non-null   object  
 11  description   8807 non-null   object  
dtypes: category(3), int64(1), object(8)
memory usage: 676.6+ KB
```

In [34]:
```
df.columns      # to show the list of columns in the dataset
```

Out[34]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_adde
d',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

In [35]:
```
df.dtypes    # to show the data type of each column
```

Out[35]:
```
show_id          object
type           category
title            object
director         object
cast             object
country        category
date_added       object
release_year      int64
rating         category
duration         object
listed_in        object
description      object
dtype: object
```

In [36]:
```
type_counts = df['type'].value_counts()
country_counts = df['country'].value_counts()
rating_counts = df['rating'].value_counts()
```

In [37]:
```
type_counts
```

Out[37]:
```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

In [38]:
```
country_counts
```

Out[38]:
```
United States
2818
India
972
United Kingdom
419
Japan
245
South Korea
199

...
Ireland, Canada, Luxembourg, United States, United Kingdom, Philippines, Indi
a            1
Ireland, Canada, United Kingdom, United States
1
Ireland, Canada, United States, United Kingdom
1
Ireland, France, Iceland, United States, Mexico, Belgium, United Kingdom, Hon
g Kong       1
Zimbabwe
1
Name: country, Length: 748, dtype: int64
```

In [39]:
```
rating_counts
```

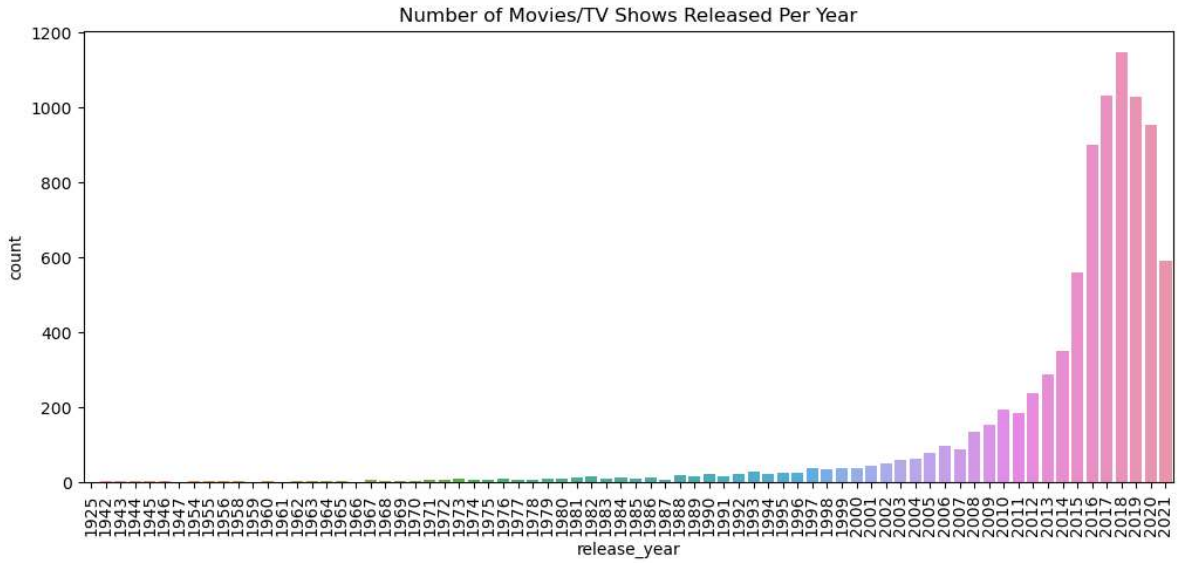
Out[39]:
```
TV-MA       3207
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
UR             3
NC-17          3
74 min         1
84 min         1
66 min         1
Name: rating, dtype: int64
```

In [40]:
```
df.describe()   #gives statistical data for the numerical column "Release year
```

Out[40]:

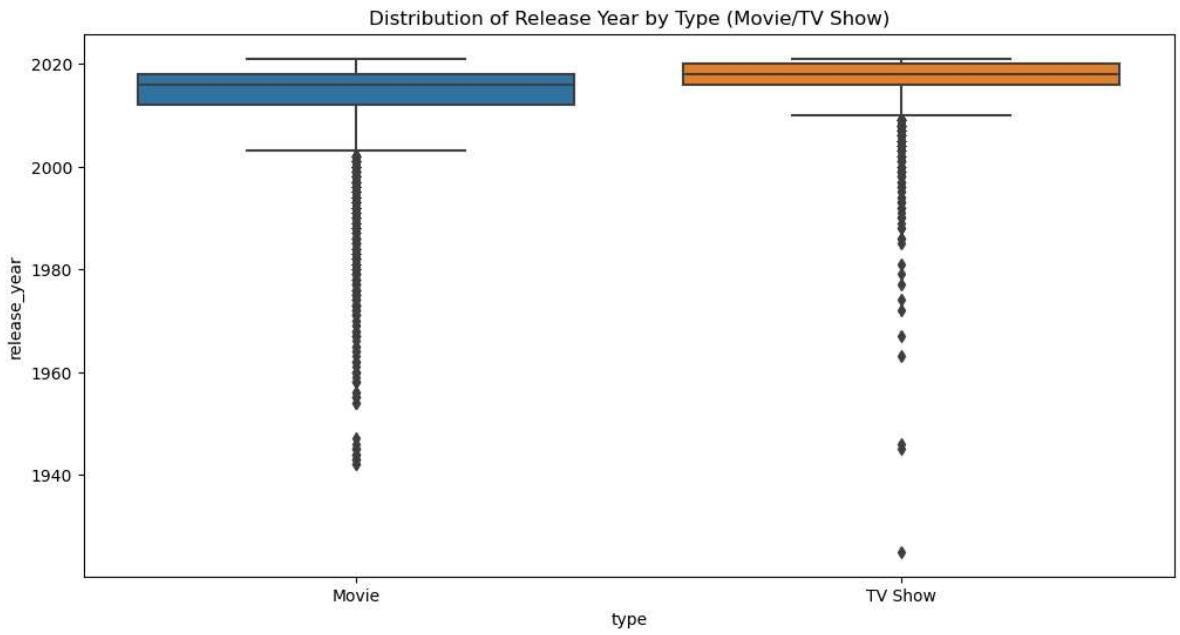
	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

In [41]:
```
plt.figure(figsize=(12, 5))                    # Univariate Analysis
sns.countplot(x='release_year', data=df)
plt.xticks(rotation=90)
plt.title('Number of Movies/TV Shows Released Per Year')
plt.show()
```



In [42]:
```
# The countplot shows a gradual increase in the number of movies and TV shows
```

In [43]:
```
plt.figure(figsize=(12, 6))                                    # Bivariate Analysis
sns.boxplot(x='type', y='release_year', data=df)
plt.title('Distribution of Release Year by Type (Movie/TV Show)')
plt.show()
```

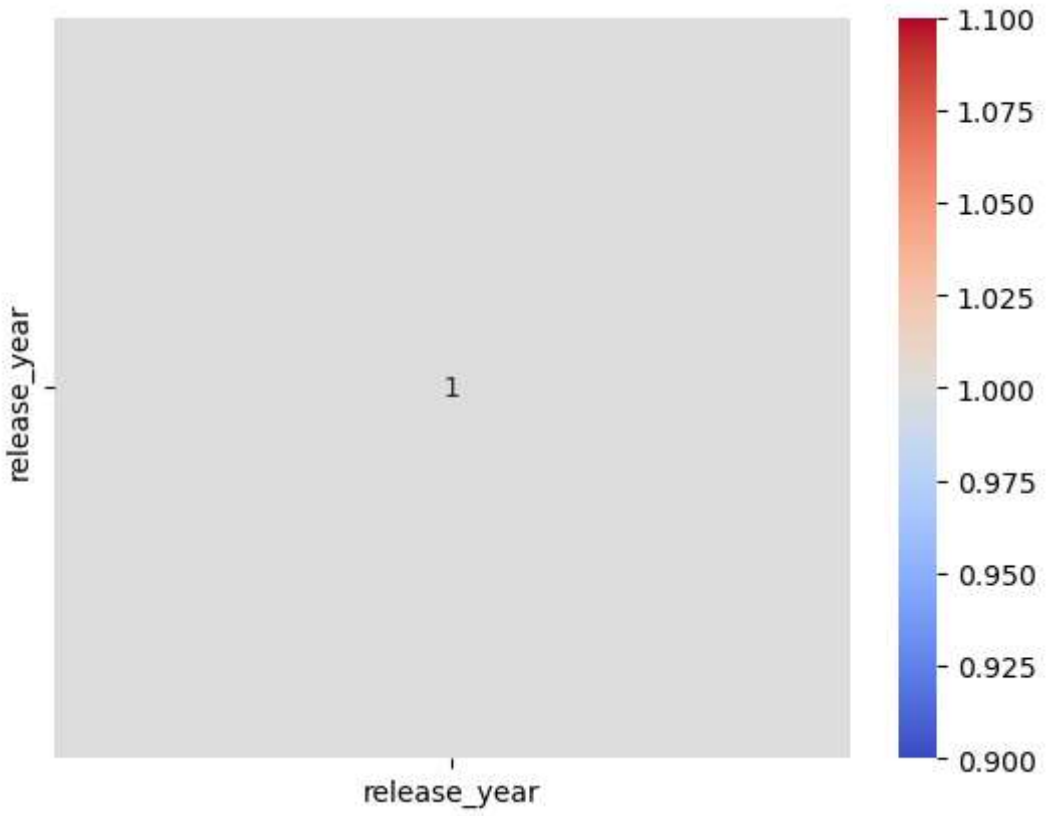


In [44]:
```
# The boxplot highlights that movies have been released in a wider range of ye
```

In [45]:
```
correlation_matrix = df.corr()                                    # Correla
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm")
```

```
C:\Users\NAMRITA\AppData\Local\Temp\ipykernel_6224\1272314064.py:1: FutureWar
ning: The default value of numeric_only in DataFrame.corr is deprecated. In a
future version, it will default to False. Select only valid columns or specif
y the value of numeric_only to silence this warning.
  correlation_matrix = df.corr()                                    # Corr
elation Analysis
```

Out[45]: <Axes: >



In [46]:
```
missing_values = df.isnull().sum()
outliers = df(np.abs(df['release_year'] - df['release_year'].mean()) / df['rel
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[46], line 2
      1 missing_values = df.isnull().sum()
----> 2 outliers = df(np.abs(df['release_year'] - df['release_year'].mean())
/ df['release_year'].std()) > 3

TypeError: 'DataFrame' object is not callable
```

In [ ]:
```
df.duplicated()    # to check any duplicate row is present in the dataset or n
```

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

In [ ]:
```
# Hence we got no duplicate rows
```

Now dealing with the Null values in the data

In [ ]:
```
df.isnull()     # check where null value is in the data
```

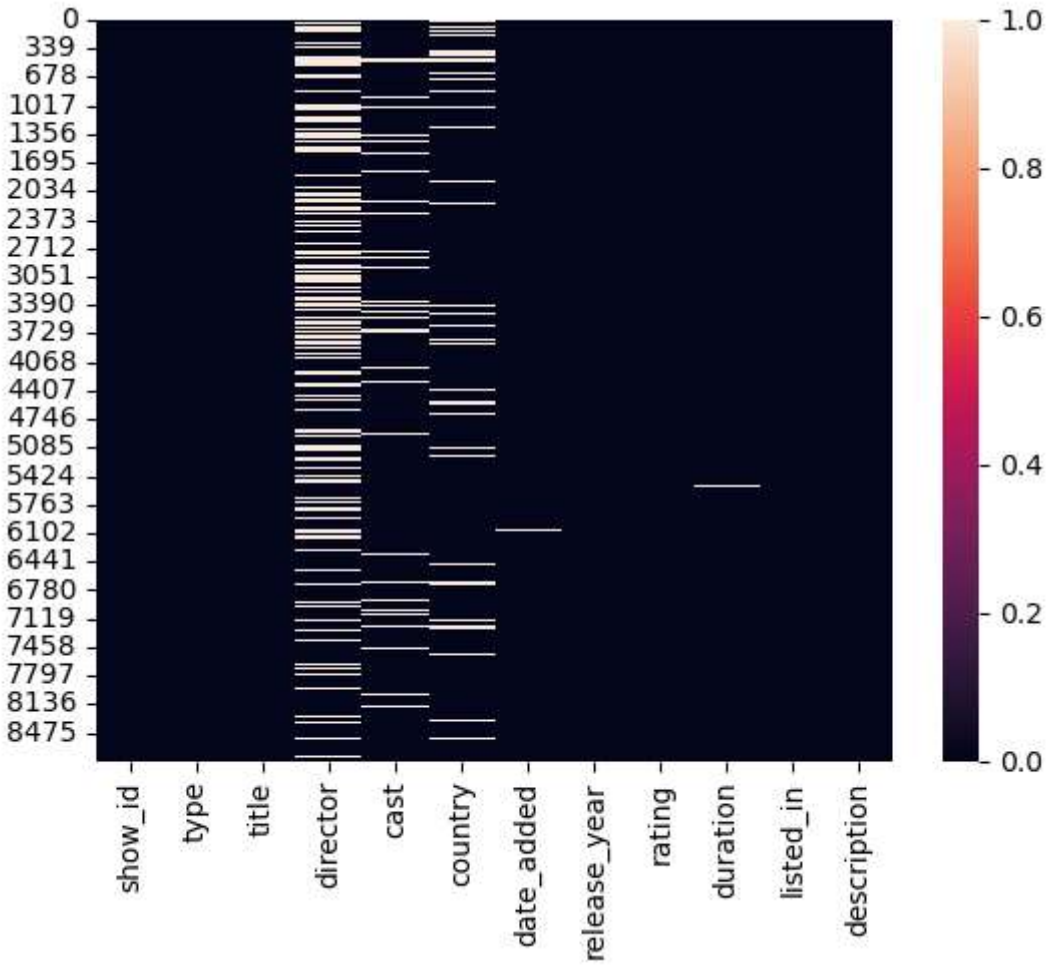
In [ ]:
```
df.isnull().sum().sort_values(ascending=False)              # to count total nu
```

In [47]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)    #to
```

Out[47]:
```
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 [48]:
```
sns.heatmap(df.isnull())    # to show nullvalue counts using heat map
```

Out[48]: <Axes: >



In [49]:
```
x = df.groupby(["rating"]).size().reset_index(name='counts')    # to check cou
```

In [50]:
```
x
```

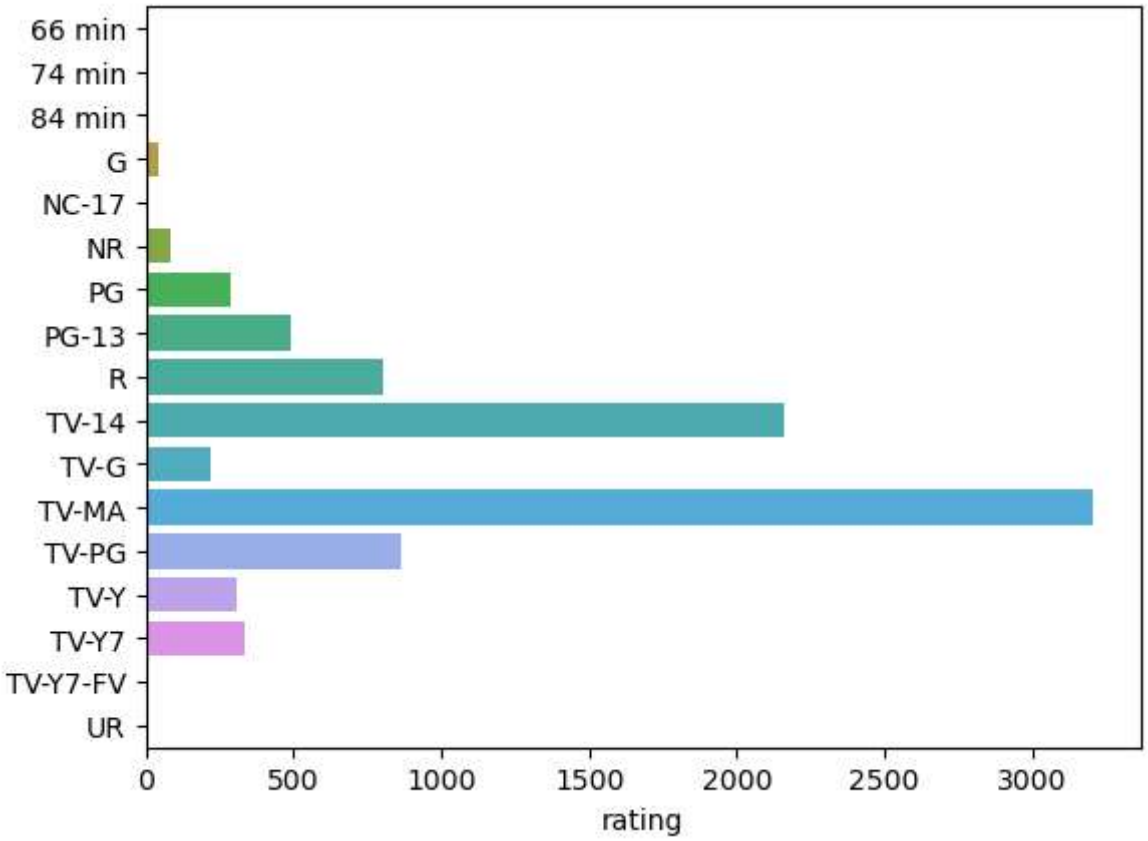
Out[50]:

	rating	counts
0	66 min	1
1	74 min	1
2	84 min	1
3	G	41
4	NC-17	3
5	NR	80
6	PG	287
7	PG-13	490
8	R	799
9	TV-14	2160
10	TV-G	220
11	TV-MA	3207
12	TV-PG	863
13	TV-Y	307
14	TV-Y7	334
15	TV-Y7-FV	6
16	UR	3

Type *Markdown* and LaTeX: α^2

In [51]:
```
sns.barplot(x=df.rating.value_counts(), y=df.rating.value_counts().index,data=
```

Out[51]:
```
<Axes: xlabel='rating'>
```



largest count = TV-MA Mature Audience Only This program is specifically designed to be viewed by adults and therefore may be unsuitable for children under 17. second largest = TV-14. This program contains material that most parents would find unsuitable for children under 14 years of age. Programs contain material that parents or adult guardians may find unsuitable for children under the age of 14. third = TV-PG. This program contains material that parents may find unsuitable for younger children. fourth = Rated R refers to movies (and also to TV shows and video games in certain systems) that have been given a "restricted" rating by the film rating systems.

** In which year the highest number of the TV Shows & Movies were released ? Show with Bar Graph.

In [52]:
```
df.dtypes
```

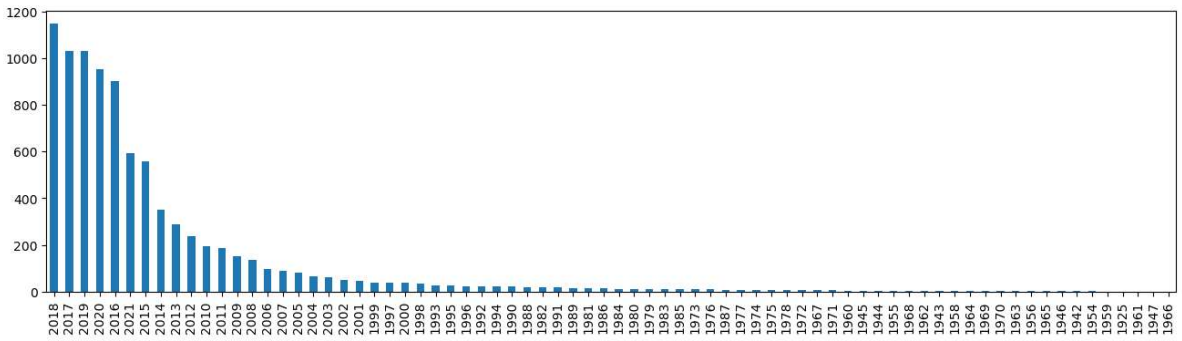
Out[52]:
```
show_id          object
type           category
title            object
director         object
cast             object
country        category
date_added       object
release_year      int64
rating         category
duration         object
listed_in        object
description      object
dtype: object
```

In [53]:
```
df["release_year"].value_counts()   # count all occurences of individual years
```

Out[53]:
```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
        ...
1959       1
1925       1
1961       1
1947       1
1966       1
Name: release_year, Length: 74, dtype: int64
```

In [54]:
```
plt.figure(figsize=(16,4))
df["release_year"].value_counts().plot(kind="bar")
```

Out[54]:
```
<Axes: >
```



In [55]:
```
# Split values in the "Cast" column and create a new row for each cast member

df['cast'] = df['cast'].str.split(', ')
df = df.explode('cast')
```

In [106]:
```
df
```

Out[106]:

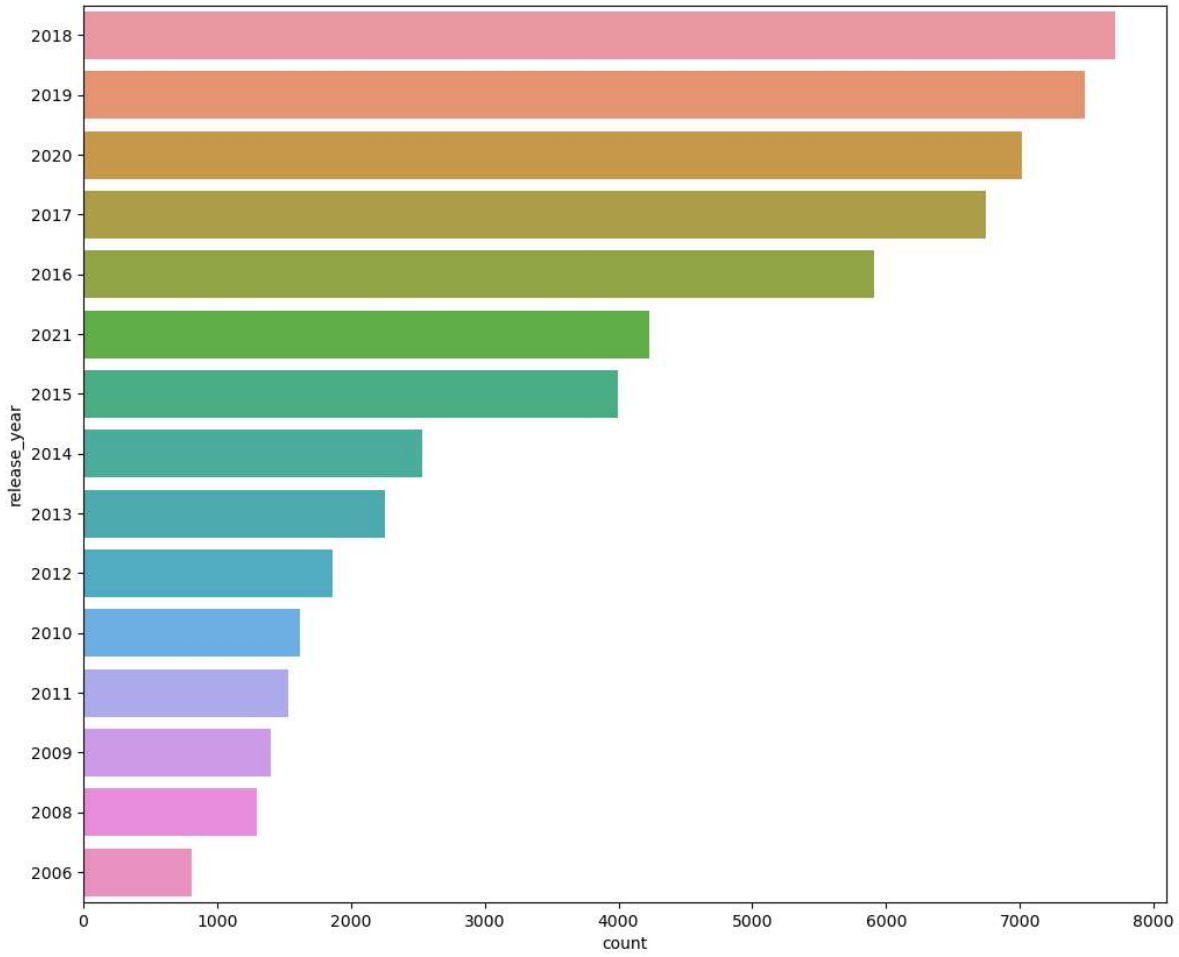
	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020	PG-13
1	s2	TV Show	Blood & Water	No Director	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	No Director	Khosi Ngema	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	No Director	Gail Mabalane	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	No Director	Thabang Molaba	South Africa	September 24, 2021	2021	TV-MA
...	...	...	...	...	...	...	...	...	...
8806	s8807	Movie	Zubaan	Mozez Singh	Manish Chaudhary	India	March 2, 2019	2015	TV-14
8806	s8807	Movie	Zubaan	Mozez Singh	Meghna Malik	India	March 2, 2019	2015	TV-14
8806	s8807	Movie	Zubaan	Mozez Singh	Malkeet Rauni	India	March 2, 2019	2015	TV-14
8806	s8807	Movie	Zubaan	Mozez Singh	Anita Shabdish	India	March 2, 2019	2015	TV-14
8806	s8807	Movie	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	March 2, 2019	2015	TV-14

64910 rows × 12 columns

In [ ]:

YEAR WISE COUNT

In [56]:
```
plt.figure(figsize=(12,10))
ax=sns.countplot(y="release_year",data=df,order=df.release_year.value_counts()
```

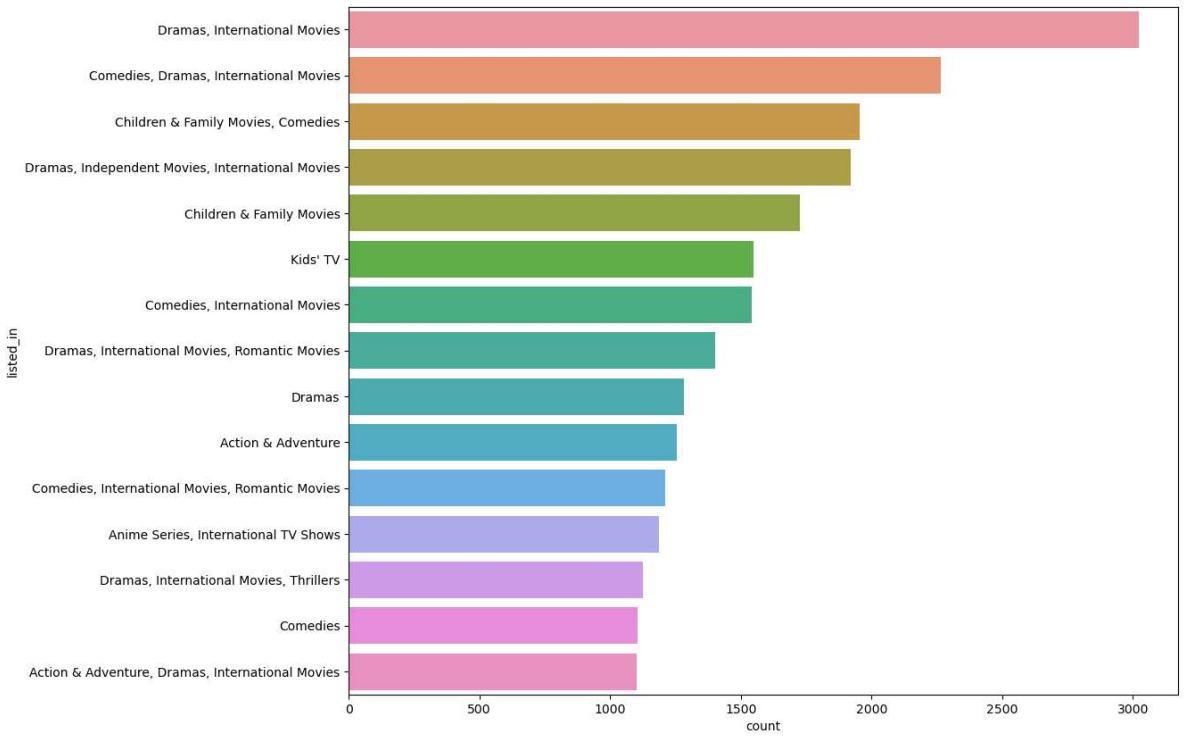


In [57]:
```
# 2018 has highest release followed by 2017 and 2019
```

In [58]:
```
df.listed_in.value_counts().head(10)
```

Out[58]:
```
Dramas, International Movies                          3022
Comedies, Dramas, International Movies                2266
Children & Family Movies, Comedies                    1956
Dramas, Independent Movies, International Movies      1919
Children & Family Movies                              1725
Kids' TV                                              1548
Comedies, International Movies                        1542
Dramas, International Movies, Romantic Movies         1402
Dramas                                                1284
Action & Adventure                                    1256
Name: listed_in, dtype: int64
```

In [59]:
```
plt.figure(figsize=(12,10))
ax=sns.countplot(y="listed_in",data=df, order=df.listed_in.value_counts().inde
```



How many Movies & TV Shows are in the dataset ? Show with Bar Graph.

In [60]:
```
df.head(2)
```

Out[60]:

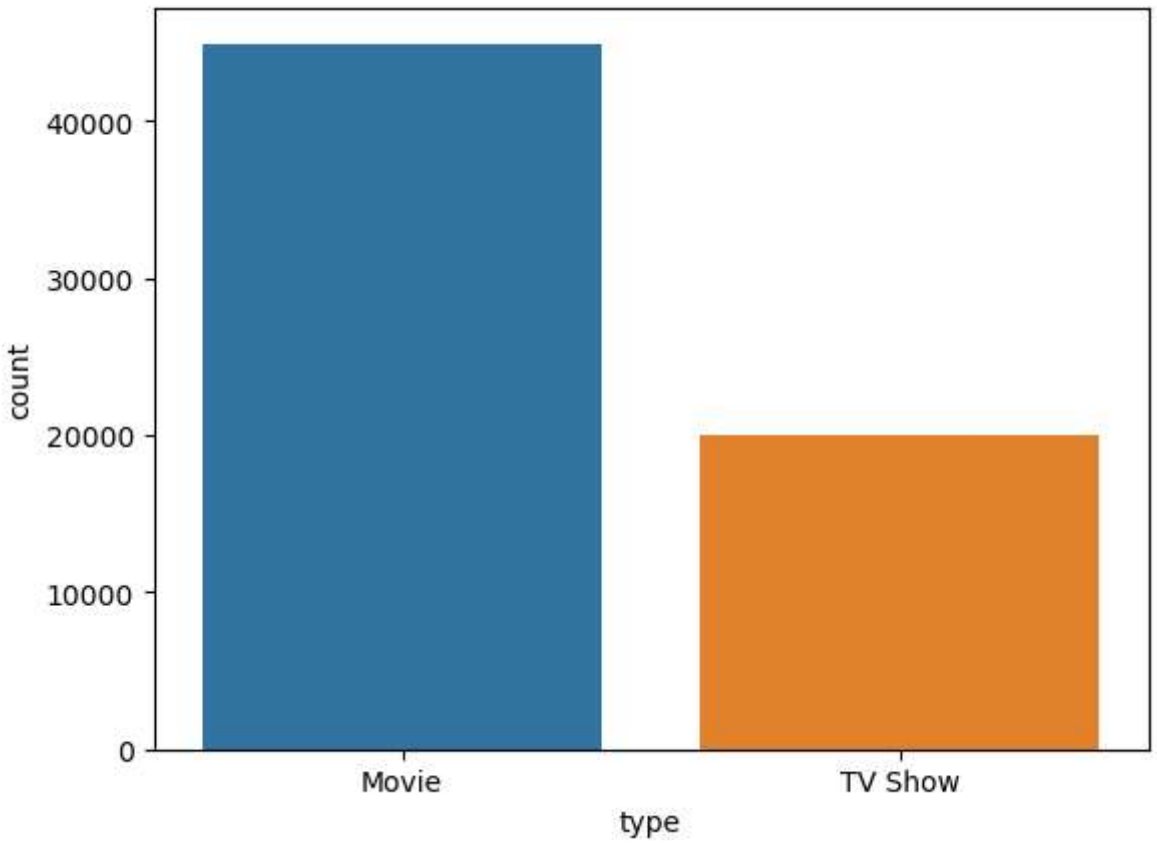
	show_id	type	title	director	cast	country	date_added	release_year	rating	duratic
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 m
1	s2	TV Show	Blood & Water	NaN	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	Seasor

In [61]:
```
df.groupby('type').type.count()
```

Out[61]:
```
type
Movie      44950
TV Show    20001
Name: type, dtype: int64
```

In [62]:
```
sns.countplot(x=df['type']) # to show count of movies and tv shows using bar g
```

Out[62]:
```
<Axes: xlabel='type', ylabel='count'>
```



Titles of all TV Shows that were released in India only.

In [63]:
```
df[(df['type']=='TV Show')&(df['country']=='India')]['title']
```

Out[63]:
```
4          Kota Factory
4          Kota Factory
4          Kota Factory
4          Kota Factory
4          Kota Factory
             ...
8775    Yeh Meri Family
8775    Yeh Meri Family
8775    Yeh Meri Family
8775    Yeh Meri Family
8775    Yeh Meri Family
Name: title, Length: 411, dtype: object
```

Show Top 10 Directors, who gave the highest number of TV Shows & Movies to Netflix ?

In [64]:
```
df.head(2)
```

Out[64]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duratic
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 m
1	s2	TV Show	Blood & Water	NaN	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	Seasor

In [65]:
```
df['director'].value_counts().head(10)
```

Out[65]:
```
Martin Scorsese        139
Cathy Garcia-Molina    125
Rajiv Chilaka          121
Steven Spielberg       121
Youssef Chahine        104
Quentin Tarantino       94
Robert Rodriguez        92
David Dhawan            90
Don Michael Paul        88
McG                     88
Name: director, dtype: int64
```

Which individual country has the Highest No. of TV Shows ?

In [66]:
```
data_TVshow=df[df['type']=='TV Show']
```

In [67]:
```
data_TVshow
```

Out[67]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
1	s2	TV Show	Blood & Water	NaN	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	S
1	s2	TV Show	Blood & Water	NaN	Khosi Ngema	South Africa	September 24, 2021	2021	TV-MA	S
1	s2	TV Show	Blood & Water	NaN	Gail Mabalane	South Africa	September 24, 2021	2021	TV-MA	S
1	s2	TV Show	Blood & Water	NaN	Thabang Molaba	South Africa	September 24, 2021	2021	TV-MA	S
1	s2	TV Show	Blood & Water	NaN	Dillon Windvogel	South Africa	September 24, 2021	2021	TV-MA	S
...	...	...	...	...	...	...	...	...	...	
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Samina Peerzada	Pakistan	December 15, 2016	2012	TV-PG	S
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Waseem Abbas	Pakistan	December 15, 2016	2012	TV-PG	S
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Javed Sheikh	Pakistan	December 15, 2016	2012	TV-PG	S
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Hina Khawaja Bayat	Pakistan	December 15, 2016	2012	TV-PG	S

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	S

20001 rows × 12 columns

In [68]:
```
data_TVshow.country.value_counts()
```

Out[68]:
```
United States
5085
Japan
1971
South Korea
1188
United Kingdom
1141
Spain
528

...
Iran, France
0
Ireland, Canada
0
Ireland, Canada, Luxembourg, United States, United Kingdom, Philippines, Indi
a       0
Ireland, Canada, United Kingdom, United States
0
Zimbabwe
0
Name: country, Length: 748, dtype: int64
```

In [69]:
```
data_TVshow.country.value_counts().head(1)
```

Out[69]:
```
United States    5085
Name: country, dtype: int64
```

Handling Missing Values

In [70]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[70]:
```
director        29.27
country          7.78
cast             1.27
date_added       0.11
rating           0.06
show_id          0.00
type             0.00
title            0.00
release_year     0.00
duration         0.00
listed_in        0.00
description      0.00
dtype: float64
```

In [71]:
```
#Dropping rows for small percentages of null
```

In [72]:
```
df.dropna(subset=["rating","duration"],axis=0,inplace=True)
```

In [73]:
```
df.shape
```

Out[73]:
```
(64910, 12)
```

In [74]:
```
round(df.isnull().sum()/df.shape[0]*100,2).sort_values(ascending=False)
```

Out[74]:
```
director        29.24
country          7.78
cast             1.27
date_added       0.11
show_id          0.00
type             0.00
title            0.00
release_year     0.00
rating           0.00
duration         0.00
listed_in        0.00
description      0.00
dtype: float64
```

In [75]:
```
df.country.value_counts().head()
```

Out[75]:
```
United States     19926
India              7246
Japan              2729
United Kingdom     2126
South Korea        1478
Name: country, dtype: int64
```

In [76]:
```
df.cast.value_counts().head(10)
```

Out[76]:
```
Anupam Kher          43
Shah Rukh Khan       35
Julie Tejwani        33
Takahiro Sakurai     32
Naseeruddin Shah     32
Rupa Bhimani         31
Om Puri              30
Akshay Kumar         30
Yuki Kaji            29
Amitabh Bachchan     28
Name: cast, dtype: int64
```

In [77]:
```
df["cast"].replace(np.NaN,"No Cast",inplace=True) #replace missing values in c
```

In [78]:
```
df["director"].replace(np.NaN,"No Director",inplace=True) #replace missing val
```

In [79]:
```
round(df.isnull().sum()/df.shape[0]*100,4).sort_values(ascending=False)
```

Out[79]:
```
country         7.7815
date_added      0.1063
show_id         0.0000
type            0.0000
title           0.0000
director        0.0000
cast            0.0000
release_year    0.0000
rating          0.0000
duration        0.0000
listed_in       0.0000
description     0.0000
dtype: float64
```

In [80]:
```
#check whether all null values are treated
```

In [81]:
```
df["title"]
```

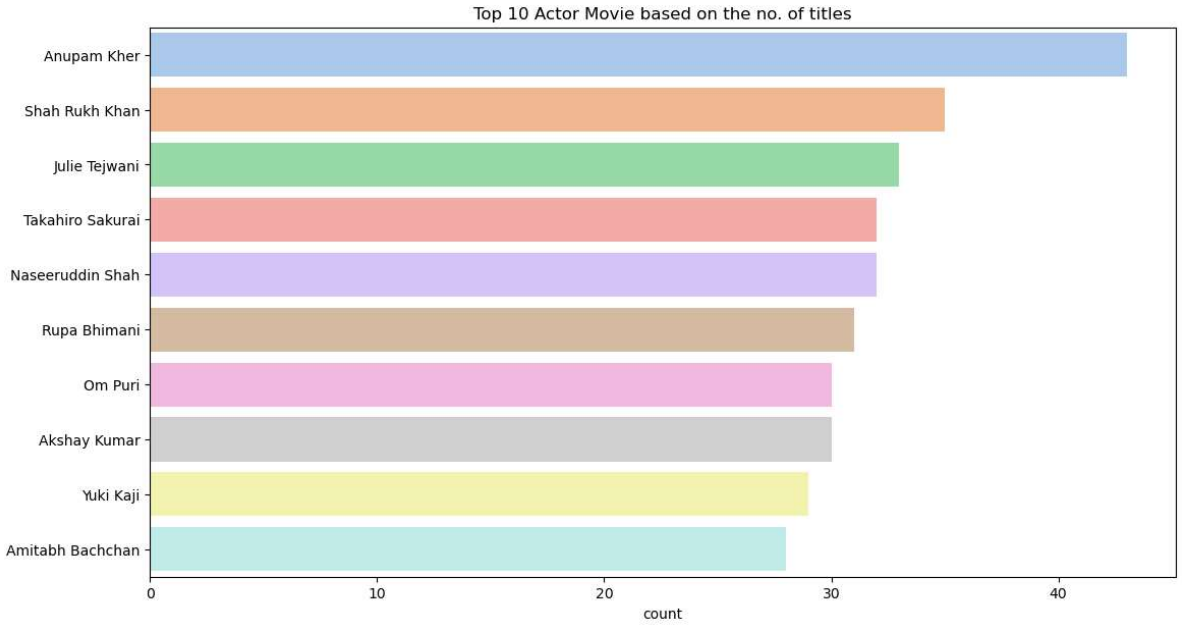
Out[81]:
```
0       Dick Johnson Is Dead
1              Blood & Water
1              Blood & Water
1              Blood & Water
1              Blood & Water
                ...
8806                  Zubaan
8806                  Zubaan
8806                  Zubaan
8806                  Zubaan
8806                  Zubaan
Name: title, Length: 64910, dtype: object
```

In [82]:
```
import warnings
import re
import plotly.express as px
import plotly.graph_objs as go
import plotly.figure_factory as ff

cast_shows=df[df.cast!='No Cast'].set_index('title').cast.str.split(', ',expan
cast_shows
cast_shows.value_counts()
```

Out[82]:
```
Anupam Kher              43
Shah Rukh Khan           35
Julie Tejwani            33
Takahiro Sakurai         32
Naseeruddin Shah         32
                         ..
Helena Zengel             1
Daniel Valenzuela         1
Mónica Ayos               1
Ignacio Quesada           1
Chittaranjan Tripathy     1
Length: 36428, dtype: int64
```

In [83]:
```
plt.figure(figsize=(13,7))
plt.title('Top 10 Actor Movie based on the no. of titles')
sns.countplot(y=cast_shows,order=cast_shows.value_counts().index[:10],palette=
plt.show()
```



In [84]:
```
movies_df=df.loc[(df['type']=="Movie")]  #dividing movies into movies_df datas
movies_df.head(2)
```

Out[84]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020	PG-13	9
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens	NaN	September 24, 2021	2021	PG	9

In [85]:
```
show_df=df.loc[(df['type']=="TV Show")]   #dividing movies into movies_df data
show_df.head(2)
```

Out[85]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
1	s2	TV Show	Blood & Water	No Director	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	2 Seasons
1	s2	TV Show	Blood & Water	No Director	Khosi Ngema	South Africa	September 24, 2021	2021	TV-MA	2 Seasons

In [86]:
```
movies_df.duration=movies_df.duration.apply(lambda x: x.replace("min","") if '
movies_df.head(2)  # in movies duration is mentioned as '90 min', so we will r
```

Out[86]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	du
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020	PG-13	
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens	NaN	September 24, 2021	2021	PG	

In [87]:
```
movies_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44938 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype   
---  ------        --------------  -----   
 0   show_id       44938 non-null  object  
 1   type          44938 non-null  category
 2   title         44938 non-null  object  
 3   director      44938 non-null  object  
 4   cast          44938 non-null  object  
 5   country       42227 non-null  category
 6   date_added    44938 non-null  object  
 7   release_year  44938 non-null  int64   
 8   rating        44938 non-null  category
 9   duration      44938 non-null  object  
 10  listed_in     44938 non-null  object  
 11  description   44938 non-null  object  
dtypes: category(3), int64(1), object(8)
memory usage: 3.6+ MB
```

We have changed it to numerical but still it shows datatype as object

In [88]:
```
movies_df.loc[:,["duration"]]=movies_df.loc[:,["duration"]].apply(lambda x:x.a
```

```
C:\Users\NAMRITA\AppData\Local\Temp\ipykernel_6224\1051877576.py:1: Deprecati
onWarning: In a future version, `df.iloc[:, i] = newvals` will attempt to set
the values inplace instead of always setting a new array. To retain the old b
ehavior, use either `df[df.columns[i]] = newvals` or, if columns are non-uniq
ue, `df.isetitem(i, newvals)`
  movies_df.loc[:,["duration"]]=movies_df.loc[:,["duration"]].apply(lambda x:
x.astype('int64',errors='ignore'))
```

In [89]:
```
movies_df.describe() # so we have changes the data type as describe() only sho
```

Out[89]:

	release_year	duration
count	44938.000000	44938.000000
mean	2012.299435	104.943144
std	9.834683	26.370296
min	1942.000000	3.000000
25%	2010.000000	91.000000
50%	2016.000000	102.000000
75%	2018.000000	118.000000
max	2021.000000	312.000000

In [90]:
```
missing_values = df.isnull().sum()
outliers = df(np.abs(df['release_year'] - df['release_year'].mean()) / df['rel
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[90], line 2
      1 missing_values = df.isnull().sum()
----> 2 outliers = df(np.abs(df['release_year'] - df['release_year'].mean())
/ df['release_year'].std()) > 3

TypeError: 'DataFrame' object is not callable
```

In [ ]:
```
missing_values
```

In [ ]:
```
missing_values.describe()
```

In [ ]:
```
range_of_attributes = df.describe()
range_of_attributes
```

In [ ]:
```
# Shortest Movie
```

In [ ]:
```
shortest_movie=movies_df.loc[(movies_df['duration']==np.min(movies_df.duration
shortest_movie
```

In [ ]:
```
longest_movie=movies_df.loc[(movies_df['duration']==np.max(movies_df.duration)
longest_movie
```

In [91]:
```
# to get the movie name with a specified duration, lets say 200 minutes
```

In [92]:
```
movie_random=movies_df.loc[(movies_df['duration']>=200)]
movie_random
```

Out[92]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
166	s167	Movie	Once Upon a Time in America	Sergio Leone	Robert De Niro	Italy, United States	September 1, 2021	1984	R
166	s167	Movie	Once Upon a Time in America	Sergio Leone	James Woods	Italy, United States	September 1, 2021	1984	R
166	s167	Movie	Once Upon a Time in America	Sergio Leone	Elizabeth McGovern	Italy, United States	September 1, 2021	1984	R
166	s167	Movie	Once Upon a Time in America	Sergio Leone	Treat Williams	Italy, United States	September 1, 2021	1984	R
166	s167	Movie	Once Upon a Time in America	Sergio Leone	Tuesday Weld	Italy, United States	September 1, 2021	1984	R
...	...	...	...	...	...	...	...	...	...
8404	s8405	Movie	The Lord of the Rings: The Return of the King	Peter Jackson	Bernard Hill	New Zealand, United States	January 1, 2020	2003	PG-13
8404	s8405	Movie	The Lord of the Rings: The Return of the King	Peter Jackson	Billy Boyd	New Zealand, United States	January 1, 2020	2003	PG-13
8404	s8405	Movie	The Lord of the Rings: The Return of the King	Peter Jackson	Dominic Monaghan	New Zealand, United States	January 1, 2020	2003	PG-13

	show_id	type	title	director	cast	country	date_added	release_year	rating	
8404	s8405	Movie	The Lord of the Rings: The Return of the King	Peter Jackson	Orlando Bloom	New Zealand, United States	January 1, 2020	2003	PG-13	
8404	s8405	Movie	The Lord of the Rings: The Return of the King	Peter Jackson	Hugo Weaving	New Zealand, United States	January 1, 2020	2003	PG-13	

170 rows × 12 columns

In [93]:
```
show_df.duration=show_df.duration.apply(lambda x: x.replace("Season","") if 'S
show_df.head(6)
```

Out[93]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duratio
1	s2	TV Show	Blood & Water	No Director	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	2
1	s2	TV Show	Blood & Water	No Director	Khosi Ngema	South Africa	September 24, 2021	2021	TV-MA	2
1	s2	TV Show	Blood & Water	No Director	Gail Mabalane	South Africa	September 24, 2021	2021	TV-MA	2
1	s2	TV Show	Blood & Water	No Director	Thabang Molaba	South Africa	September 24, 2021	2021	TV-MA	2
1	s2	TV Show	Blood & Water	No Director	Dillon Windvogel	South Africa	September 24, 2021	2021	TV-MA	2
1	s2	TV Show	Blood & Water	No Director	Natasha Thahane	South Africa	September 24, 2021	2021	TV-MA	2

In [94]:
```
show_df.duration=show_df.duration.apply(lambda x: x.replace(" s","") if 's' in
```

In [95]:
```
show_df.duration
```

Out[95]:
```
1        2
1        2
1        2
1        2
1        2
        ..
8800    1
8800    1
8800    1
8800    1
8803     2
Name: duration, Length: 19972, dtype: object
```

In [96]:
```
show_df.head(6)
```

Out[96]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duratio
1	s2	TV Show	Blood & Water	No Director	Ama Qamata	South Africa	September 24, 2021	2021	TV-MA	
1	s2	TV Show	Blood & Water	No Director	Khosi Ngema	South Africa	September 24, 2021	2021	TV-MA	
1	s2	TV Show	Blood & Water	No Director	Gail Mabalane	South Africa	September 24, 2021	2021	TV-MA	
1	s2	TV Show	Blood & Water	No Director	Thabang Molaba	South Africa	September 24, 2021	2021	TV-MA	
1	s2	TV Show	Blood & Water	No Director	Dillon Windvogel	South Africa	September 24, 2021	2021	TV-MA	
1	s2	TV Show	Blood & Water	No Director	Natasha Thahane	South Africa	September 24, 2021	2021	TV-MA	

In [97]:
```
show_df.describe()
```

Out[97]:

	release_year
count	19972.000000
mean	2016.647306
std	5.319832
min	1925.000000
25%	2016.000000
50%	2018.000000
75%	2020.000000
max	2021.000000

In [98]:
```
show_df.loc[:,["duration"]]=show_df.loc[:,["duration"]].apply(lambda x: x.asty
show_df.describe()
```

```
C:\Users\NAMRITA\AppData\Local\Temp\ipykernel_6224\2731820753.py:1: Deprecati
onWarning: In a future version, `df.iloc[:, i] = newvals` will attempt to set
the values inplace instead of always setting a new array. To retain the old b
ehavior, use either `df[df.columns[i]] = newvals` or, if columns are non-uniq
ue, `df.isetitem(i, newvals)`
  show_df.loc[:,["duration"]]=show_df.loc[:,["duration"]].apply(lambda x: x.a
stype('int64',errors='ignore'))
```

Out[98]:

	release_year	duration
count	19972.000000	19972.000000
mean	2016.647306	1.968256
std	5.319832	1.851314
min	1925.000000	1.000000
25%	2016.000000	1.000000
50%	2018.000000	1.000000
75%	2020.000000	2.000000
max	2021.000000	17.000000

In [99]:
```
show_df.duration.value_counts().tail(10)
```

Out[99]:
```
7     305
6     278
8     147
9     102
10     66
13     65
12     26
15     23
17     15
11     15
Name: duration, dtype: int64
```

In [100]:
```
#show with highest no. of seasons
```

In [101]:
```
longest_shows=show_df.loc[show_df["duration"]>=13]
longest_shows
```

Out[101]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	c
548	s549	TV Show	Grey's Anatomy	No Director	Ellen Pompeo	United States	July 3, 2021	2020	TV-14	
548	s549	TV Show	Grey's Anatomy	No Director	Sandra Oh	United States	July 3, 2021	2020	TV-14	
548	s549	TV Show	Grey's Anatomy	No Director	Katherine Heigl	United States	July 3, 2021	2020	TV-14	
548	s549	TV Show	Grey's Anatomy	No Director	Justin Chambers	United States	July 3, 2021	2020	TV-14	
548	s549	TV Show	Grey's Anatomy	No Director	Patrick Dempsey	United States	July 3, 2021	2020	TV-14	
...	...	...	...	...	...	...	...	...	...	
7847	s7848	TV Show	Red vs. Blue	No Director	Matt Hullum	United States	NaN	2015	NR	
7847	s7848	TV Show	Red vs. Blue	No Director	Dan Godwin	United States	NaN	2015	NR	
7847	s7848	TV Show	Red vs. Blue	No Director	Kathleen Zuelch	United States	NaN	2015	NR	
7847	s7848	TV Show	Red vs. Blue	No Director	Yomary Cruz	United States	NaN	2015	NR	

	show_id	type	title	director	cast	country	date_added	release_year	rating	c
7847	s7848	TV Show	Red vs. Blue	No Director	Nathan Zellner	United States	NaN	2015	NR	

103 rows × 12 columns

In [102]:
```
longest_shows.rating.value_counts()
```

Out[102]:
```
TV-MA       47
TV-14       46
NR          10
66 min       0
TV-Y7-FV     0
TV-Y7        0
TV-Y         0
TV-PG        0
TV-G         0
R            0
74 min       0
PG-13        0
PG           0
NC-17        0
G            0
84 min       0
UR           0
Name: rating, dtype: int64
```

In [103]:
```
netflix_date=df[['date_added']].dropna()
netflix_date
```

Out[103]:

	date_added
0	September 25, 2021
1	September 24, 2021
1	September 24, 2021
1	September 24, 2021
1	September 24, 2021
...	...
8806	March 2, 2019
8806	March 2, 2019
8806	March 2, 2019
8806	March 2, 2019
8806	March 2, 2019

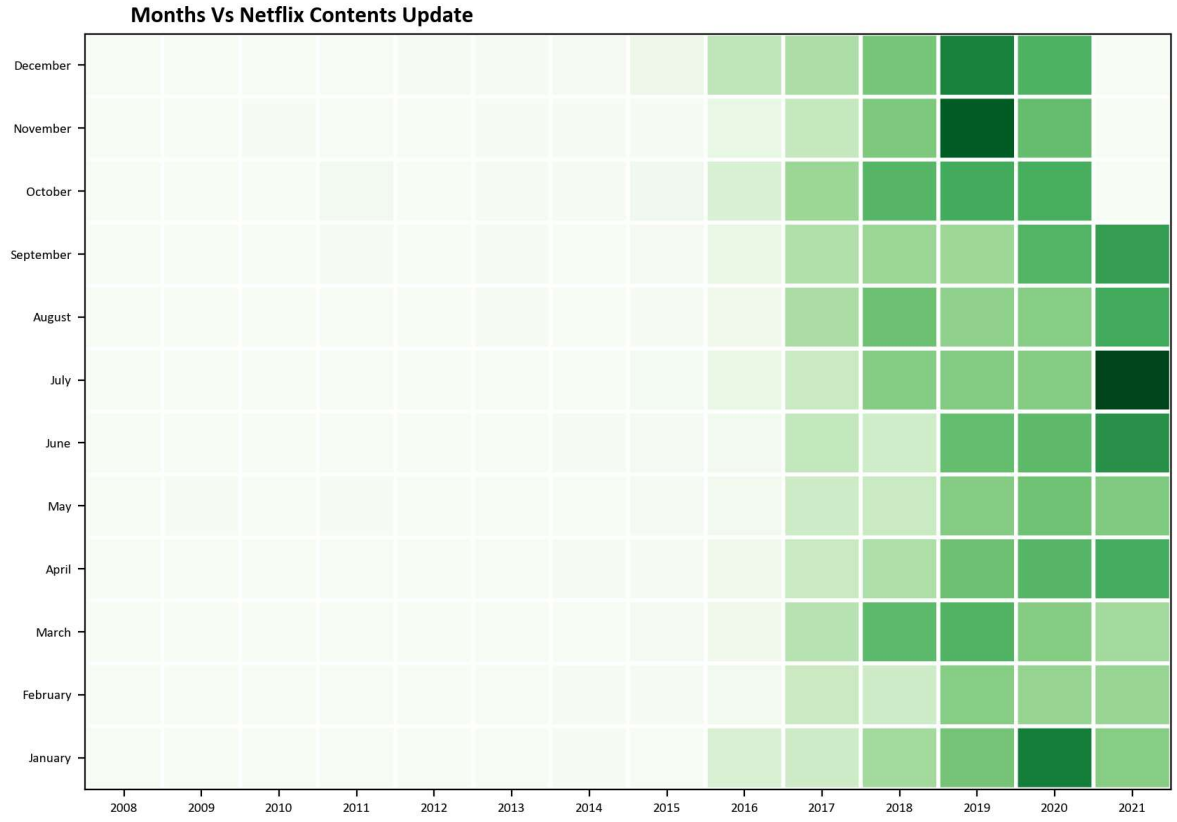
64841 rows × 1 columns

In [104]:
```
netflix_date['year']=netflix_date['date_added'].apply(lambda x: x.split(', ')[
netflix_date['month']=netflix_date['date_added'].apply(lambda x: x.lstrip().sp

month_order=['January','February','March','April','May','June','July','August'
new_df=netflix_date.groupby('year')['month'].value_counts().unstack().fillna(0
plt.figure(figsize=(10, 7), dpi=200)

plt.pcolor(new_df, cmap='Greens',edgecolors='white',linewidths=2)      #heatma
plt.xticks(np.arange(0.5,len(new_df.columns),1), new_df.columns, fontsize=7, f
plt.yticks(np.arange(0.5,len(new_df.index),1), new_df.index, fontsize=7, fontf

plt.title('Months Vs Netflix Contents Update', fontsize=12, fontfamily='calibr
plt.show()
```



In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: