Business Case: Netflix - Data Exploration and Visualisation

About NETFLIX:

1. NETFLIX is one of the most popular media and video streaming platforms.
2. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally.
3. This tabular dataset consists of listings of all the movies and tv shows available on NETFLIX, along with details such as - cast, directors, ratings, release year, duration, etc.

Business Problem:

Analyze the data and generate insights that could help NETFLIX decide which type of shows/movies to produce and how they can grow the business in different countries.

The dataset provided to you consists of a list of all the TV shows/movies available on NETFLIX:

```
1. Show_id: Unique ID for every Movie / TV show
2. Type: Identifier - A Movie or TV Show
3. Title: Title of the Movie / TV Show
4. Director: Director of the Movie
5. Cast: Actors involved in the movie/show
6. Country: Country where the movie/show was produced
7. Date_added: Date it was added on Netflix
8. Release_year: Actual Release year of the movie/show
9. Rating: TV Rating of the movie/show
10. Duration: Total Duration - in minutes or number of seasons
11. Listed_in: Genre
12. Description: The Summary Description
```

Problem Statements:

1. How has the number of movies released per year changed over the last 20-30 years?
2. Comparison of tv shows vs. movies.
3. What is the best time to lunch a TV show?
4. Analysis of actors/directors of different types of shows/movies.
5. Does Netflix has more focus on TV Shows than movies in recent years?
6. Understanding what content is available in different countries.

1. Analysing Basic Metics

Netflix is a popular service that people across the world use for entertainment.

In this Exploratory Analysis and Visualization,

I will explore the netflix dataset through visualizations and graphs using numpy, pandas,matplotlib and seaborn.

The Aim of this Business case study is to explore and analyze the Netflix shows data after filtering some of the columns. This Netflix movies and TV shows data . We will alter and filter some columns and perform some feature engineering after this we prepare data for analysis.We will perform univariate analysis so we can get a better understanding of every column and then bivariate and multivariate analysis so we will understand the relations between columns. In the end, we will conclude the result of the analysis.

IMPORTING THE LIBRARIES

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import matplotlib
import plotly.express as px
import plotly.graph_objs as go
import plotly.figure_factory as ff
import warnings
warnings.filterwarnings('ignore')
```

LOADING THE DATASET

In [2]:
```
df_A = pd.read_csv("netflix.csv")
df_A
```

Out[2]:

	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...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

8807 rows × 12 columns

After the loading dataset we can see that there is 8807 rows and 12 columns. Here also some "NaN" values.

In [3]:
```
df_A.head(5) # Top 5
```

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]:
```
df_A.tail(5) # Bottom 5
```

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

2. Observations:

a. Shape of Data
b. Data type of all attributes
c. missing value detection
d. statistcal summary

In [5]:
```
df_A.shape # shows like(rows, columns)
```

Out[5]:
```
(8807, 12)
```

there are 8807 rows and 12 columns.

In [6]:
```
df_A.describe(include = 'all') #statistical summary
```

Out[6]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8807.000000	8803	8804	8807	8807
unique	8807	2	8807	4528	7692	748	1767	NaN	17	220	514	8775
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	NaN	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	1	19	19	2818	109	NaN	3207	1793	362	4
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014.180198	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.819312	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	NaN	NaN	NaN	NaN

In [7]:
```
df_A.describe() #statistical summary
```

Out[7]:

	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

It give some statistical summary like Count, Mean, Minimum, Maximum, Standard Deviation, 25%tile, 50%tile, 75%tile.

In [8]:
```
df_A.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
```

There are 8807 entries and 12 columns to work with for Exploratory Data Analysis and Visualization. Right off the bat, there are a few columns that contain null values ('director', 'cast', 'country', 'date_added', 'rating')

In [9]:
```
df_A.columns # shows all the column names
```

Out[9]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

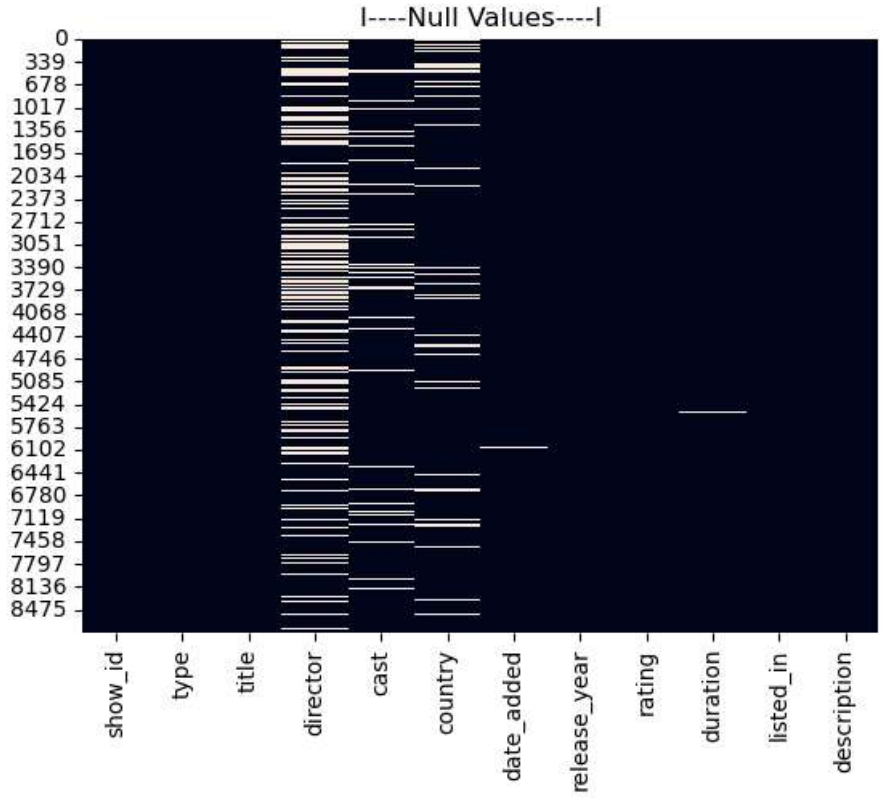
In [10]:
```
df_A.isnull().values.any()
```

Out[10]:
```
True
```

In [11]:
```
df_A.isnull().sum().sum()
```

Out[11]:
```
4307
```

In [12]:
```
sns.heatmap(df_A.isnull(), cbar=False)
plt.title("I----Null Values----I")
plt.show()
```



In [13]:
```
df_A.isnull().sum().sort_values(ascending = False)
```

Out[13]:
```
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
```

Above in the heatmap and table, we can see that there are many null values in the dataset. There are a total of 4,307 null values across the entire dataset with 2634 missing values in 'director', 825 in 'cast', 831 in 'country', 10 in 'date_added', 4 in 'rating', 3 in duration. We will have to handle all null data values before we can dive into Exploratory data and modeling.

In [14]:
```
round(df_A.isnull().sum()/df_A.shape[0]*100,2).sort_values(ascending = False)
```

Out[14]:
```
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
```

It shows the percentage null values.There is 29.91% null value in director, 9.44% in country, 9.37% in cast, 0.11% in date_added 0.05% in rating, 0.03% in duration.

3.Non Graphical Analysis: Value counts and unique attributes

In [15]:
```
df_A["director"].value_counts() # director value count
```

Out[15]:
```
Rajiv Chilaka                     19
Raúl Campos, Jan Suter            18
Marcus Raboy                      16
Suhas Kadav                       16
Jay Karas                         14
                                  ..
Raymie Muzquiz, Stu Livingston     1
Joe Menendez                       1
Eric Bross                         1
Will Eisenberg                     1
Mozez Singh                        1
Name: director, Length: 4528, dtype: int64
```

In [16]:
```
df_A.nunique() #unique values
```

Out[16]:
```
show_id         8807
type               2
title           8807
director        4528
cast            7692
country          748
date_added      1767
release_year      74
rating            17
duration         220
listed_in        514
description     8775
dtype: int64
```

We can see that for each of the columns, there are alot different unique values for some of them. It makes sense that show_id is large since it is a unique key used to identify a movies And TVshows. title, director, cast, country, date_added, release_year,rating, duration, listed_in, and description contain many unique values as well.

In [17]:
```
df_A.country.value_counts() # country value count
```

Out[17]:
```
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 [18]:
```
df_A["listed_in"].value_counts() # Genre value count
```

Out[18]:
```
Dramas, International Movies                          362
Documentaries                                         359
Stand-Up Comedy                                       334
Comedies, Dramas, International Movies                274
Dramas, Independent Movies, International Movies      252
                                                     ...
Kids' TV, TV Action & Adventure, TV Dramas              1
TV Comedies, TV Dramas, TV Horror                       1
Children & Family Movies, Comedies, LGBTQ Movies        1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows      1
Cult Movies, Dramas, Thrillers                          1
Name: listed_in, Length: 514, dtype: int64
```

In [19]:
```
df_A["date_added"].value_counts() # date_added value count
```

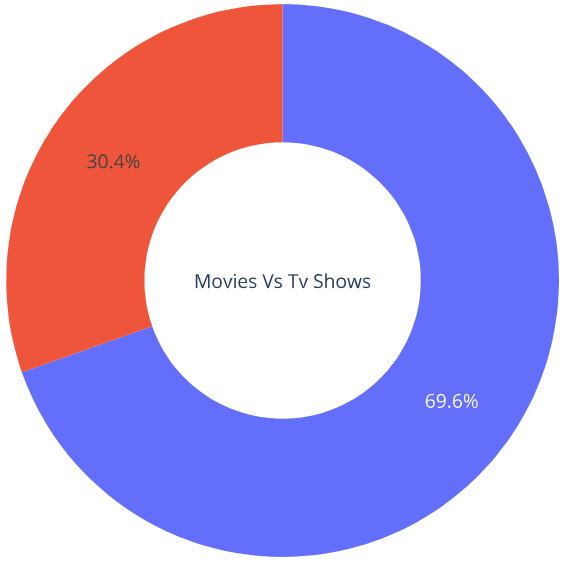
Out[19]:
```
January 1, 2020      109
November 1, 2019      89
March 1, 2018         75
December 31, 2019     74
October 1, 2018       71
                    ...
December 4, 2016       1
November 21, 2016      1
November 19, 2016      1
November 17, 2016      1
January 11, 2020       1
Name: date_added, Length: 1767, dtype: int64
```

In [20]:
```
df_A["release_year"].value_counts() # release year value count
```

Out[20]:
```
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
```

Some Basic Visual Analysis:

In [21]:
```
go.Figure(data = [go.Pie(labels = df_A.type.value_counts(normalize = True).index,
                         values = df_A.type.value_counts(normalize = True).values, hole =.5,
                         title = "Movies Vs Tv Shows")])
```



This data set contain around 70% of movies And 30% of TV shows

In [22]:
```
df_A["type"].value_counts()
```

Out[22]:
```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

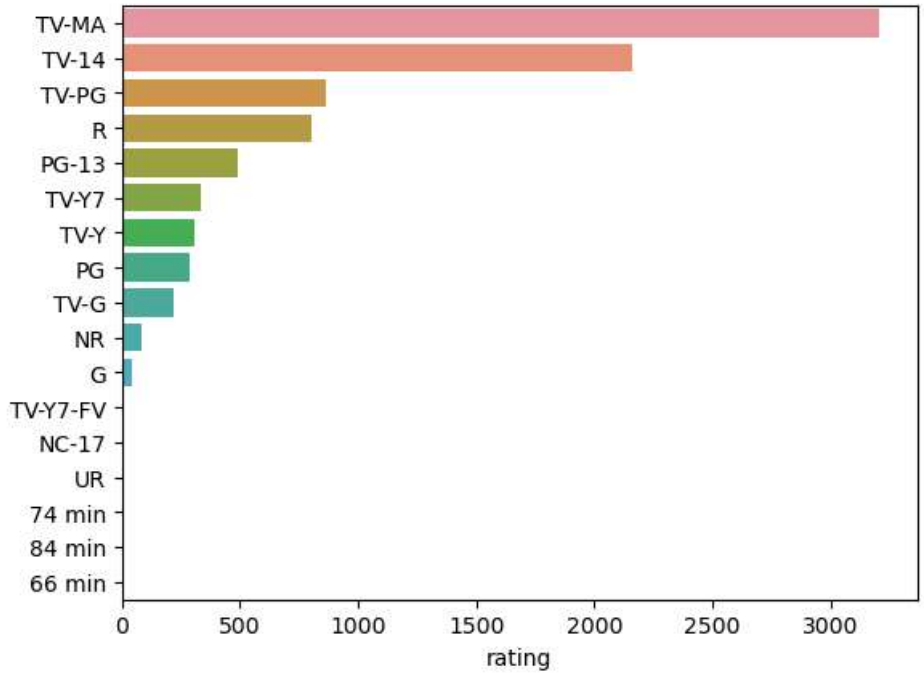
The total Movies is 6131 & Tv shows 2676.

In [23]:
```
df_A["rating"].value_counts()
```

Out[23]:
```
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
74 min         1
84 min         1
66 min         1
Name: rating, dtype: int64
```

In [24]:
```
sns.barplot(x=df_A.rating.value_counts(), y = df_A.rating.value_counts().index, data = df_A, orient = "h")
```

Out[24]:
```
<Axes: xlabel='rating'>
```



Highest Count: TV-MA is the rating that shows that a program is intended for adults. "MA" stands for "Mature Aduiances". Child aged 17 and younger should not view these programs.

Followed by Highest(2nd) is the "TV-14". "TV-14" program is meant for children over 14 years of age. It is generally not recommended to let children watchthe program without parental attendance or atleast whithout them vetting it first. It can contain crude humor, the use of harmful substances, strong language, violence, and complex or upsetting themes.

Third largest TV-PG Parental Guidance Suggested This program contains material that parents may find unsuitable for younger children

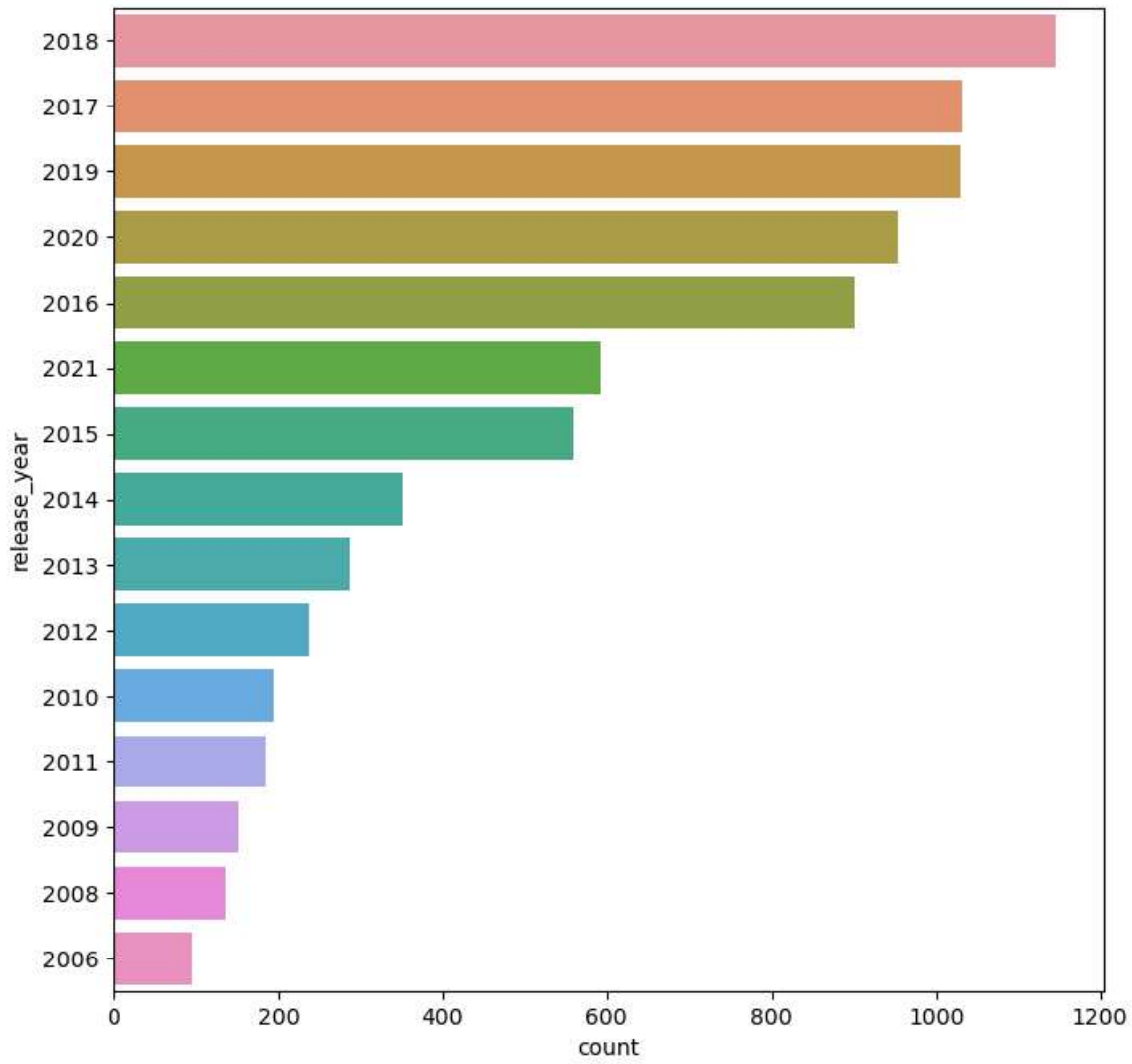
Fourth largest is the very popular "R" ratings. R is the short for restricted, so any young person under 17 should not watch.

In [25]:
```
# Top 10 countries of creation Movies & TV shows
df_A.country.value_counts().head(10)
```

Out[25]:
```
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 [26]:
```
#year wise count
plt.figure(figsize = (8, 8))
```

```
ax = sns.countplot(y = "release_year", data=df_A, order= df_A.release_year.value_counts().index[0:15])
```

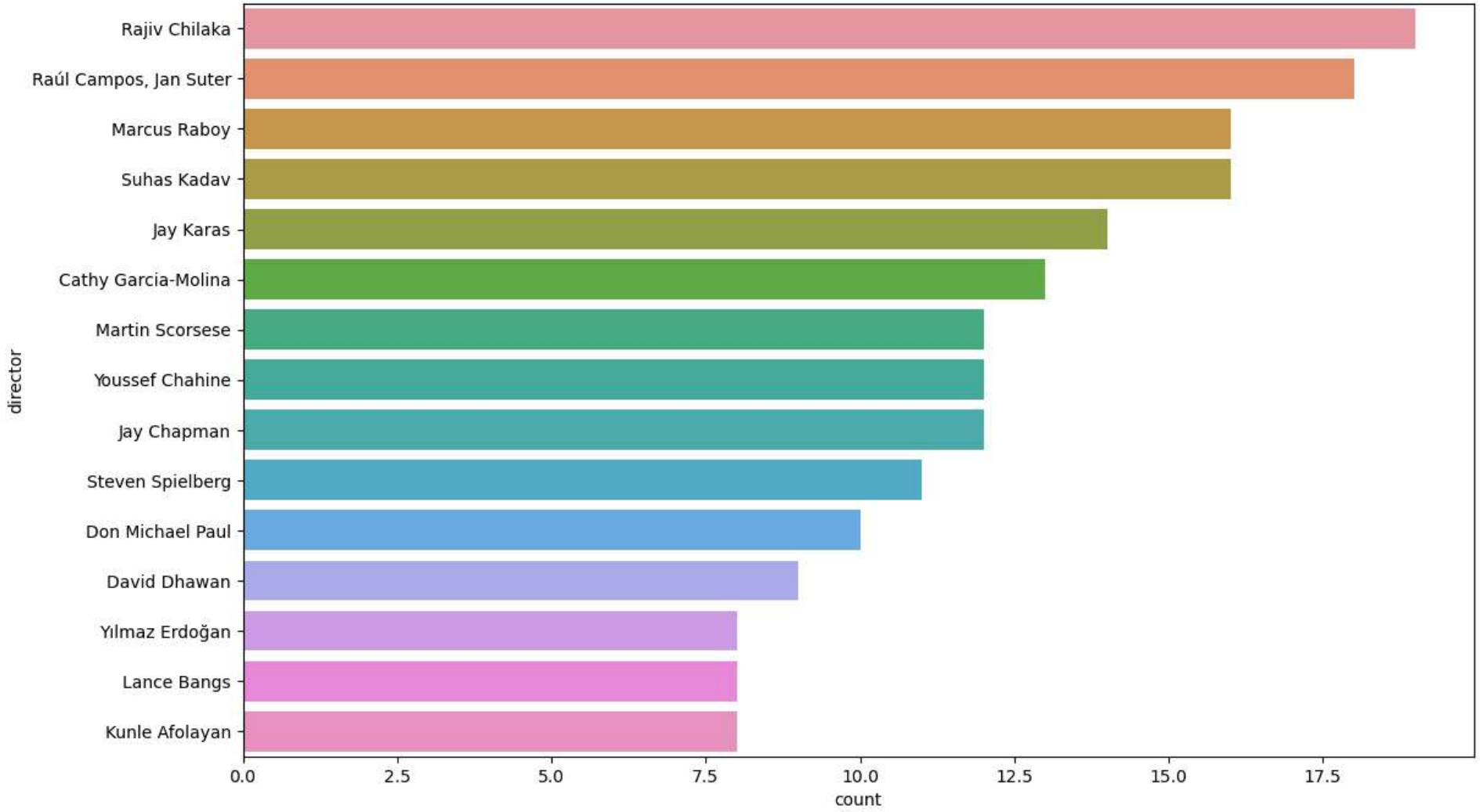


Highest release in 2018 followed by 2017 and 2019

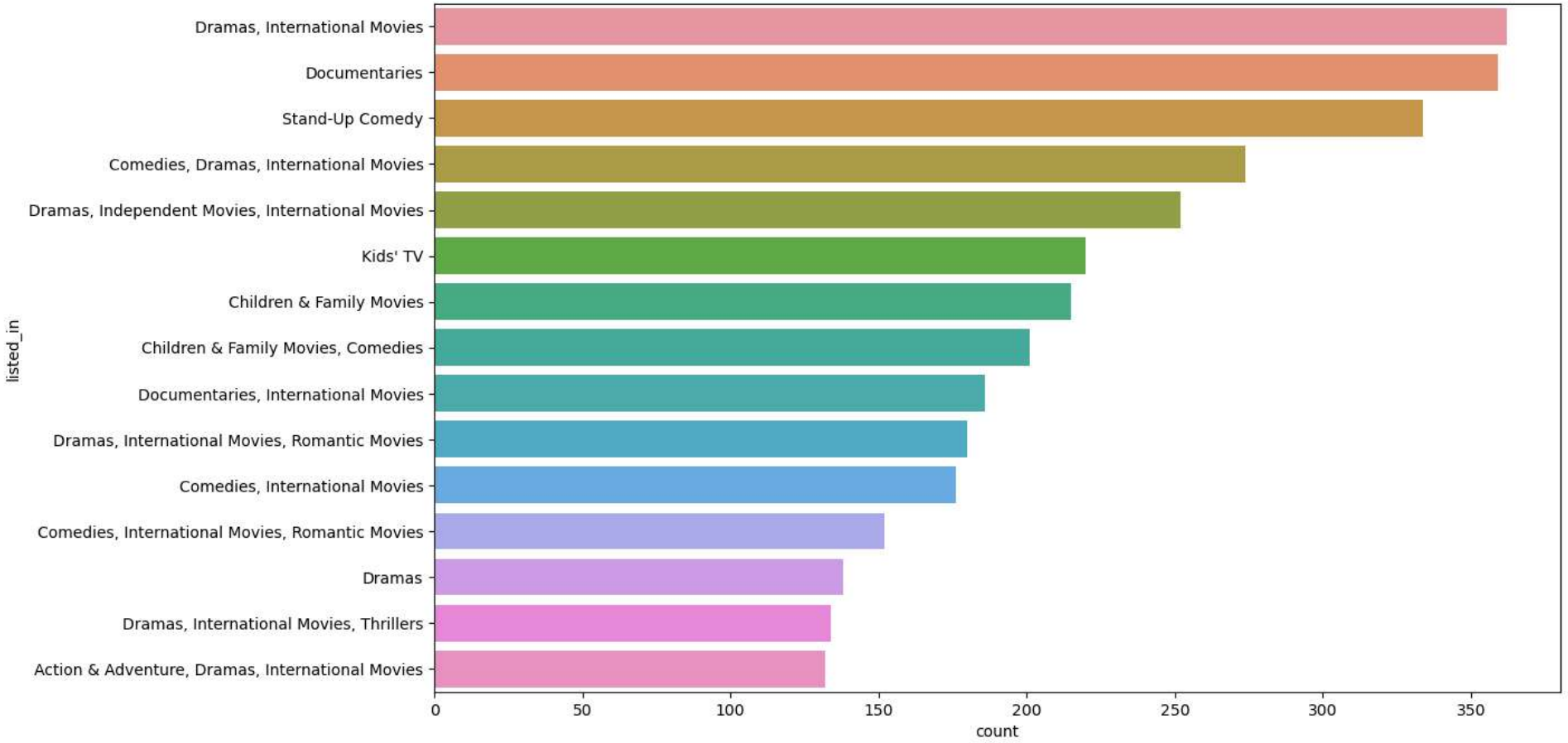
In [27]:
```
# Top 10 directors
df_A.director.value_counts().head(10)
```

Out[27]:
```
Rajiv Chilaka             19
Raúl Campos, Jan Suter    18
Marcus Raboy              16
Suhas Kadav               16
Jay Karas                 14
Cathy Garcia-Molina       13
Martin Scorsese           12
Youssef Chahine           12
Jay Chapman               12
Steven Spielberg          11
Name: director, dtype: int64
```

In [28]:
```
# Top 15 director
plt.figure(figsize = (13, 8))
ax = sns.countplot(y = "director", data=df_A, order= df_A.director.value_counts().index[0:15])
```



In [29]:
```
# Top 15 Genre
plt.figure(figsize = (13, 8))
bx = sns.countplot(y = "listed_in", data=df_A, order= df_A.listed_in.value_counts().index[0:15])
```



4. Visual Analysis - Univariate, Bivariate after pre-processingof the Data

Note: Pre-processing involves unnesting of the data in columns like Actor, Director & Country.

Unnesting the Cast Column

In [30]:
```
cast_constraint=df_A['cast'].apply(lambda x: str(x).split(', ')).tolist()
df_B = pd.DataFrame(cast_constraint, index = df_A['title'])
df_B = df_B.stack()
df_B = pd.DataFrame(df_B.reset_index())
df_B.rename(columns={0:'Actors'},inplace=True)
df_B = df_B.drop(['level_1'],axis=1)
df_B.head(15)
```

Out[30]:

	title	Actors
0	Dick Johnson Is Dead	nan
1	Blood & Water	Ama Qamata
2	Blood & Water	Khosi Ngema
3	Blood & Water	Gail Mabalane
4	Blood & Water	Thabang Molaba
5	Blood & Water	Dillon Windvogel
6	Blood & Water	Natasha Thahane
7	Blood & Water	Arno Greeff
8	Blood & Water	Xolile Tshabalala
9	Blood & Water	Getmore Sithole
10	Blood & Water	Cindy Mahlangu
11	Blood & Water	Ryle De Morny
12	Blood & Water	Greteli Fincham
13	Blood & Water	Sello Maake Ka-Ncube
14	Blood & Water	Odwa Gwanya

Unnesting the Director Column

In [31]:
```
dir_constraint=df_A['director'].apply(lambda x: str(x).split(', ')).tolist()
df_C = pd.DataFrame(dir_constraint, index = df_A['title'])
df_C = df_C.stack()
df_C = pd.DataFrame(df_C.reset_index())
df_C.rename(columns={0:'Director'},inplace=True)
df_C = df_C.drop(['level_1'],axis=1)
df_C.head(15)
```

Out[31]:

	title	Director
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	nan
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	nan
4	Kota Factory	nan
5	Midnight Mass	Mike Flanagan
6	My Little Pony: A New Generation	Robert Cullen
7	My Little Pony: A New Generation	José Luis Ucha
8	Sankofa	Haile Gerima
9	The Great British Baking Show	Andy Devonshire
10	The Starling	Theodore Melfi
11	Vendetta: Truth, Lies and The Mafia	nan
12	Bangkok Breaking	Kongkiat Komesiri
13	Je Suis Karl	Christian Schwochow
14	Confessions of an Invisible Girl	Bruno Garotti

Unnesting the country column

In [32]:
```
country_constraint=df_A['country'].apply(lambda x: str(x).split(', ')).tolist()
df_D = pd.DataFrame(country_constraint, index = df_A['title'])
df_D = df_D.stack()
df_D = pd.DataFrame(df_D.reset_index())
df_D.rename(columns={0:'Country'},inplace=True)
df_D = df_D.drop(['level_1'],axis=1)
df_D.head(15)
```

Out[32]:

	title	Country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	nan
3	Jailbirds New Orleans	nan
4	Kota Factory	India
5	Midnight Mass	nan
6	My Little Pony: A New Generation	nan
7	Sankofa	United States
8	Sankofa	Ghana
9	Sankofa	Burkina Faso
10	Sankofa	United Kingdom
11	Sankofa	Germany
12	Sankofa	Ethiopia
13	The Great British Baking Show	United Kingdom
14	The Starling	United States

Unnesting the listed_in column (Genre)

In [33]:
```
listed_constraint=df_A['listed_in'].apply(lambda x: str(x).split(', ')).tolist()
df_E = pd.DataFrame(listed_constraint, index = df_A['title'])
df_E = df_E.stack()
df_E = pd.DataFrame(df_E.reset_index())
df_E.rename(columns={0:'Genre'},inplace=True)
df_E = df_E.drop(['level_1'],axis=1)
df_E.head(10)
```

Out[33]:

	title	Genre
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Blood & Water	TV Dramas
3	Blood & Water	TV Mysteries
4	Ganglands	Crime TV Shows
5	Ganglands	International TV Shows
6	Ganglands	TV Action & Adventure
7	Jailbirds New Orleans	Docuseries
8	Jailbirds New Orleans	Reality TV
9	Kota Factory	International TV Shows

Collecting the all the unnested dataframes

In [34]:
```
df_F = df_B.merge(df_C,on=['title'],how='inner')

df_G = df_F.merge(df_E,on=['title'],how='inner')

df_H = df_G.merge(df_D,on=['title'],how='inner')

df_H.head(10)
```

Out[34]:

	title	Actors	Director	Genre	Country
0	Dick Johnson Is Dead	nan	Kirsten Johnson	Documentaries	United States
1	Blood & Water	Ama Qamata	nan	International TV Shows	South Africa
2	Blood & Water	Ama Qamata	nan	TV Dramas	South Africa
3	Blood & Water	Ama Qamata	nan	TV Mysteries	South Africa
4	Blood & Water	Khosi Ngema	nan	International TV Shows	South Africa
5	Blood & Water	Khosi Ngema	nan	TV Dramas	South Africa
6	Blood & Water	Khosi Ngema	nan	TV Mysteries	South Africa
7	Blood & Water	Gail Mabalane	nan	International TV Shows	South Africa
8	Blood & Water	Gail Mabalane	nan	TV Dramas	South Africa
9	Blood & Water	Gail Mabalane	nan	TV Mysteries	South Africa

In [35]:
```
df_H.shape # shows shape like(rows, columns)
```

Out[35]: (201991, 5)

Merging unnested data with the primary dataframe

In [36]:
```
df_A = df_H.merge(df_A[['show_id', 'type', 'title', 'date_added',
       'release_year', 'rating', 'duration']],on=['title'],how='left')
df_A.head(20)
```

Out[36]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	nan	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Ama Qamata	nan	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	nan	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	nan	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	nan	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
5	Blood & Water	Khosi Ngema	nan	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
6	Blood & Water	Khosi Ngema	nan	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
7	Blood & Water	Gail Mabalane	nan	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
8	Blood & Water	Gail Mabalane	nan	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
9	Blood & Water	Gail Mabalane	nan	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
10	Blood & Water	Thabang Molaba	nan	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
11	Blood & Water	Thabang Molaba	nan	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
12	Blood & Water	Thabang Molaba	nan	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
13	Blood & Water	Dillon Windvogel	nan	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
14	Blood & Water	Dillon Windvogel	nan	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
15	Blood & Water	Dillon Windvogel	nan	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
16	Blood & Water	Natasha Thahane	nan	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
17	Blood & Water	Natasha Thahane	nan	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
18	Blood & Water	Natasha Thahane	nan	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
19	Blood & Water	Arno Greeff	nan	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

In [37]:
```
df_A.shape # shows shape like(rows, columns)
```

Out[37]: (201991, 11)

Befor merging there is 201991 rows and 5 columns and after merging there are same rows but 11 columns are there. Now the dataset is fine to shows further visualization.

Handeling the Null values/ Missing values

In [38]:
```
df_A.isnull().sum().sort_values(ascending = False)
```

Out[38]:
```
date_added      158
rating           67
duration          3
title             0
Actors            0
Director          0
Genre             0
Country           0
show_id           0
type              0
release_year      0
dtype: int64
```

In [39]:
```
total_null_data = df_A.isnull().sum().sort_values(ascending = False)
percent = ((df_A.isnull().sum()/df_A.isnull().count())*100).sort_values(ascending = False)
print("Total records = ", df_A.shape[0])
Null_data = pd.concat([total_null_data,percent.round(2)],axis=1,keys=['Total Null Data','In % null value'])
Null_data.head(12)
```

```
Total records =  201991
```

Out[39]:

	Total Null Data	In % null value
date_added	158	0.08
rating	67	0.03
duration	3	0.00
title	0	0.00
Actors	0	0.00
Director	0	0.00
Genre	0	0.00
Country	0	0.00
show_id	0	0.00
type	0	0.00
release_year	0	0.00

Above table gives missing values summary in absolute value and in Percentage, date added has the maximum missing values 158 then 67 is rating

Missing Value Handeling

In [40]:
```
df_A['Actors'].replace(['nan'],['Unknown Actor'],inplace=True)
df_A['Director'].replace(['nan'],['Unknown Director'],inplace=True)
df_A['Country'].replace(['nan'],[np.nan],inplace=True)
df_A.head(20)
```

Out[40]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
5	Blood & Water	Khosi Ngema	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
6	Blood & Water	Khosi Ngema	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
7	Blood & Water	Gail Mabalane	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
8	Blood & Water	Gail Mabalane	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
9	Blood & Water	Gail Mabalane	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
10	Blood & Water	Thabang Molaba	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
11	Blood & Water	Thabang Molaba	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
12	Blood & Water	Thabang Molaba	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
13	Blood & Water	Dillon Windvogel	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
14	Blood & Water	Dillon Windvogel	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
15	Blood & Water	Dillon Windvogel	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
16	Blood & Water	Natasha Thahane	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
17	Blood & Water	Natasha Thahane	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
18	Blood & Water	Natasha Thahane	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
19	Blood & Water	Arno Greeff	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

In [41]:
```
total_null = df_A.isnull().sum().sort_values(ascending = False)
percent = ((df_A.isnull().sum()/df_A.isnull().count())*100).sort_values(ascending = False)
print("Total records = ", df_A.shape[0])

missing_data = pd.concat([total_null,percent.round(2)],axis=1,keys=['Total Missing','In Percent'])
missing_data.head(10)
```

```
Total records =  201991
```

Out[41]:

	Total Missing	In Percent
Country	11897	5.89
date_added	158	0.08
rating	67	0.03
duration	3	0.00
title	0	0.00
Actors	0	0.00
Director	0	0.00
Genre	0	0.00
show_id	0	0.00
type	0	0.00

After replacing string nan with np.nan, actual null values of country went upto 11897 or 5.89 %.

In [42]:
```
df_A[df_A.duration.isnull()]
```

Out[42]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration
126537	Louis C.K. 2017	Louis C.K.	Louis C.K.	Movies	United States	s5542	Movie	April 4, 2017	2017	74 min	NaN
131603	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	Movies	United States	s5795	Movie	September 16, 2016	2010	84 min	NaN
131737	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	Movies	United States	s5814	Movie	August 15, 2016	2015	66 min	NaN

In [43]:
```
df_A.loc[df_A['duration'].isnull(),'duration'] = df_A.loc[df_A['duration'].isnull(),'duration'].fillna(df_A['rating'])
df_A.loc[df_A['rating'].str.contains('min', na=False),'rating'] = 'NR'
df_A['rating'].fillna('NR',inplace=True)
df_A.isnull().sum()
```

Out[43]:
```
title              0
Actors             0
Director           0
Genre              0
Country        11897
show_id            0
type               0
date_added       158
release_year       0
rating             0
duration           0
dtype: int64
```

Filling missing values of date added column with mode value with respective release years

In [44]:
```
for i in df_A[df_A['date_added'].isnull()]['release_year'].unique():
    date = df_A[df_A['release_year'] == i]['date_added'].mode().values[0]
    df_A.loc[df_A['release_year'] == i,'date_added'] = df_A.loc[df_A['release_year']==i,'date_added'].fillna(date)
df_A[df_A.Country.isna()]
```

Out[44]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration
58	Ganglands	Sami Bouajila	Julien Leclercq	Crime TV Shows	NaN	s3	TV Show	September 24, 2021	2021	TV-MA	1 Season
59	Ganglands	Sami Bouajila	Julien Leclercq	International TV Shows	NaN	s3	TV Show	September 24, 2021	2021	TV-MA	1 Season
60	Ganglands	Sami Bouajila	Julien Leclercq	TV Action & Adventure	NaN	s3	TV Show	September 24, 2021	2021	TV-MA	1 Season
61	Ganglands	Tracy Gotoas	Julien Leclercq	Crime TV Shows	NaN	s3	TV Show	September 24, 2021	2021	TV-MA	1 Season
62	Ganglands	Tracy Gotoas	Julien Leclercq	International TV Shows	NaN	s3	TV Show	September 24, 2021	2021	TV-MA	1 Season
...	...	...	...	...	...	...	...	...	...	...	...
201424	YOM	Mayur Vyas	Unknown Director	Kids' TV	NaN	s8786	TV Show	June 7, 2018	2016	TV-Y7	1 Season
201425	YOM	Ketan Kava	Unknown Director	Kids' TV	NaN	s8786	TV Show	June 7, 2018	2016	TV-Y7	1 Season
201932	Zombie Dumb	Unknown Actor	Unknown Director	Kids' TV	NaN	s8804	TV Show	July 1, 2019	2018	TV-Y7	2 Seasons
201933	Zombie Dumb	Unknown Actor	Unknown Director	Korean TV Shows	NaN	s8804	TV Show	July 1, 2019	2018	TV-Y7	2 Seasons
201934	Zombie Dumb	Unknown Actor	Unknown Director	TV Comedies	NaN	s8804	TV Show	July 1, 2019	2018	TV-Y7	2 Seasons

11897 rows × 11 columns

Filling missing values of country column with mode value with respective directors

In [45]:
```
for i in df_A[df_A['Country'].isnull()]['Director'].unique():
    if i in df_A[~df_A['Country'].isnull()]['Director'].unique():
        country = df_A[df_A['Director'] == i]['Country'].mode().values[0]
        df_A.loc[df_A['Director'] == i,'Country'] = df_A.loc[df_A['Director'] == i,'Country'].fillna(country)
df_A.isnull().sum()
```

Out[45]:
```
title            0
Actors           0
Director         0
Genre            0
Country       4276
show_id          0
type             0
date_added       0
release_year     0
rating           0
duration         0
dtype: int64
```

In [46]:
```
for i in df_A[df_A['Country'].isnull()]['Actors'].unique():
    if i in df_A[~df_A['Country'].isnull()]['Actors'].unique():
```

```
        imp = df_A[df_A['Actors'] == i]['Country'].mode().values[0]
        df_A.loc[df_A['Actors'] == i,'Country'] = df_A.loc[df_A['Actors']==i,'Country'].fillna(imp)
```

In [47]:
```
df_A['Country'].fillna('Unknown Country',inplace=True)
df_A.isnull().sum().sort_values(ascending = False)
```

Out[47]:
```
title           0
Actors          0
Director        0
Genre           0
Country         0
show_id         0
type            0
date_added      0
release_year    0
rating          0
duration        0
dtype: int64
```

In [48]:
```
df_A.head(10)
```

Out[48]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
5	Blood & Water	Khosi Ngema	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
6	Blood & Water	Khosi Ngema	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
7	Blood & Water	Gail Mabalane	Unknown Director	International TV Shows	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
8	Blood & Water	Gail Mabalane	Unknown Director	TV Dramas	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
9	Blood & Water	Gail Mabalane	Unknown Director	TV Mysteries	South Africa	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons

In [49]:
```
df_A["date_added"] = pd.to_datetime(df_A['date_added'])
```

In [50]:
```
df_A['duration'] = df_A['duration'].str.replace("min","")
df_A.head(10)
```

Out[50]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
5	Blood & Water	Khosi Ngema	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
6	Blood & Water	Khosi Ngema	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
7	Blood & Water	Gail Mabalane	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
8	Blood & Water	Gail Mabalane	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
9	Blood & Water	Gail Mabalane	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons

In [51]:
```
df_A['duration2'] = df_A.duration.copy()
df_ = df_A.copy()
```

In [52]:
```
df_.loc[df_['duration2'].str.contains('Season'),'duration2'] = 0
df_['duration2'] = df_.duration2.astype('int')
df_.head()
```

Out[52]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration	duration2
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90	90
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	0

In [53]:
```
df_.duration2.describe()
```

Out[53]:
```
count    201991.000000
mean         77.152789
std          52.269154
min           0.000000
25%           0.000000
50%          95.000000
75%         112.000000
max         312.000000
Name: duration2, dtype: float64
```

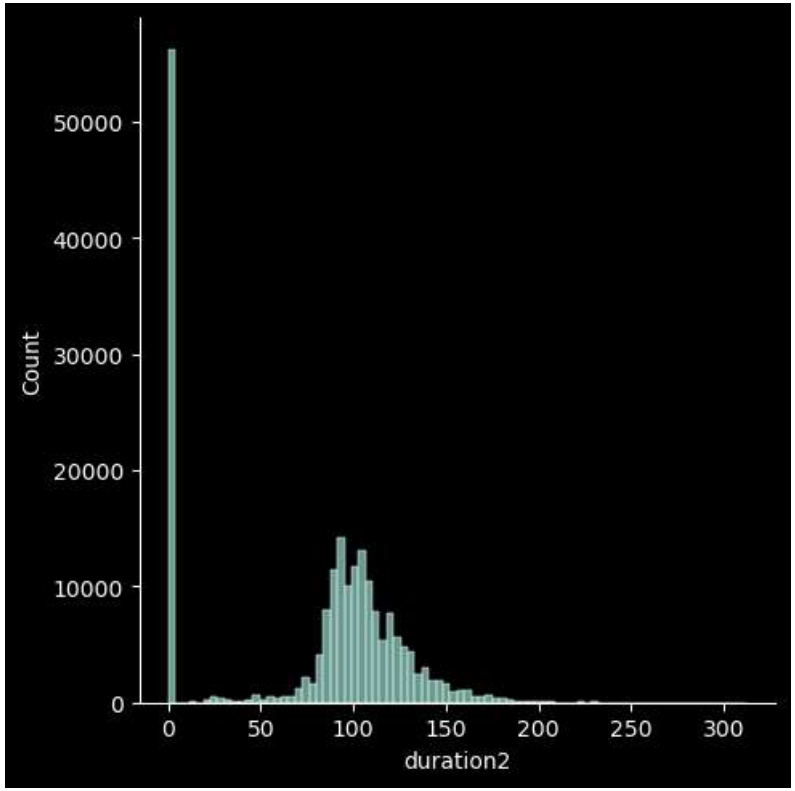
In [54]:
```
df_.T.apply(lambda x: x.nunique(), axis=1)
```

Out[54]:
```
title           8807
Actors         36440
Director        4994
Genre             42
Country          128
show_id         8807
type               2
date_added      1714
release_year      74
rating            14
duration         220
duration2        206
dtype: int64
```

Univariate analysis of duration column

In [55]:
```
plt.style.use('dark_background')
plt.figure(figsize=(10,2))
sns.displot(df_['duration2'])

plt.show()
```

```
<Figure size 1000x200 with 0 Axes>
```



In [56]:
```
bins = [-1,1,50,80,100,120,150,200,315]
labels = ['<1','1-50','50-80','80-100','100-120','120-150','150-200','200-315']
df_['duration2'] = pd.cut(df_['duration2'],bins = bins, labels = labels )
df_.head()
```

Out[56]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration	duration2
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	90	80-100
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	<1

In [57]:
```
df_.loc[~df_['duration'].str.contains('Season'),'duration'] = df_.loc[~df_['duration'].str.contains('Season'),'duration2']
df_.drop(['duration2'],axis=1,inplace=True)
df_.head()
```

Out[57]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons

In [58]:
```
from datetime import datetime
from dateutil.parser import parse
df_["year_added"] = df_['date_added'].dt.year
df_["year_added"] = df_["year_added"].astype("Int64")
df_["month_added"] = df_['date_added'].dt.month
df_['month_name'] = df_['date_added'].dt.month_name()
df_["month_added"] = df_["month_added"].astype("Int64")
df_["day_added"] = df_['date_added'].dt.day
df_["day_added"] = df_["day_added"].astype("Int64")
df_['Weekday_added'] = df_['date_added'].apply(lambda x: parse(str(x)).strftime("%A"))
df_.head()
```

Out[58]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration	year_added	month_added	month_name	day_added	Weekday_added
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100	2021	9	September	25	Saturday
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday

In [59]:
```
df_['title'] = df_['title'].str.replace(r"\(.*\)","")
df_.head()
```

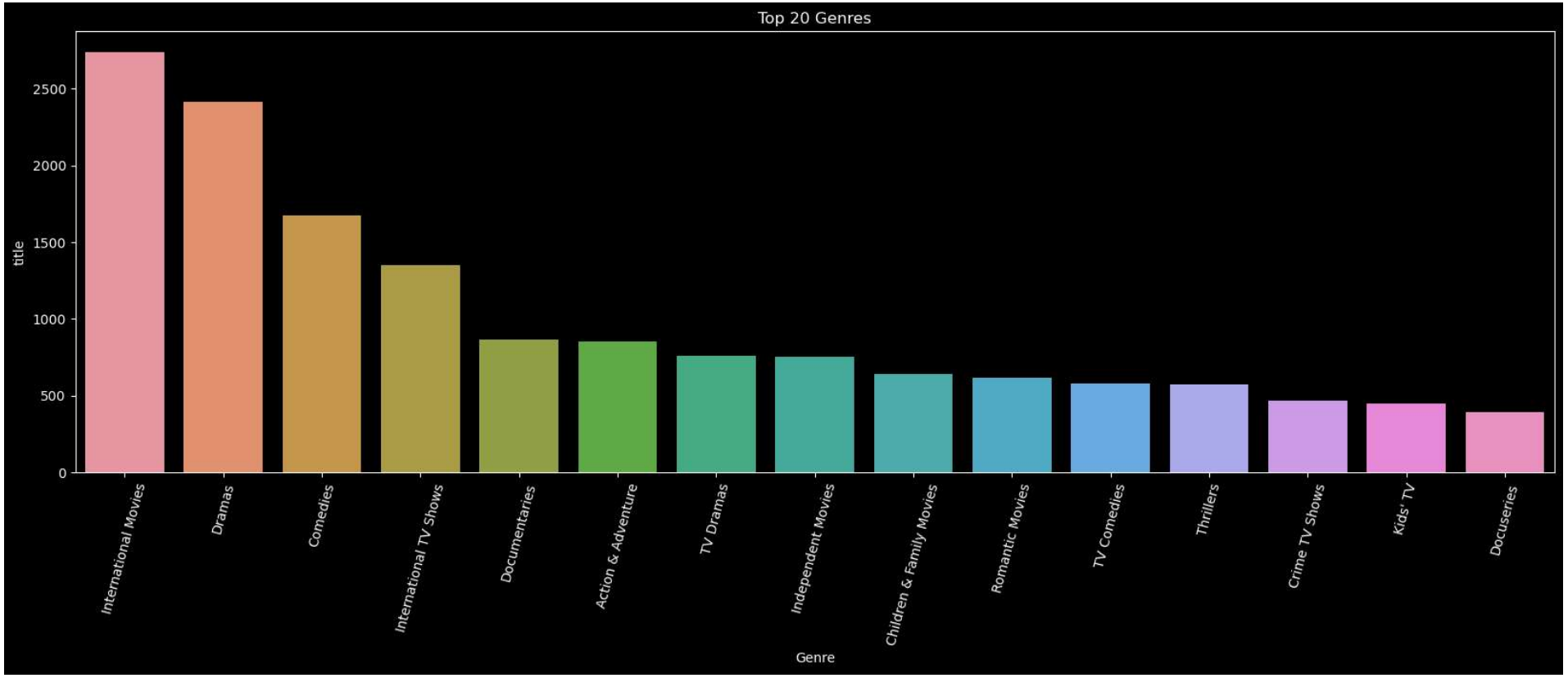
Out[59]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration	year_added	month_added	month_name	day_added	Weekday_added
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100	2021	9	September	25	Saturday
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday

Univariate Analysis

In [60]:

```
df_genre=df_.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
plt.figure(figsize=(20,6))
sns.barplot(x = "Genre",y = 'title', data = df_genre)
plt.xticks(rotation = 75)
plt.title('Top 20 Genres')
plt.show()
```

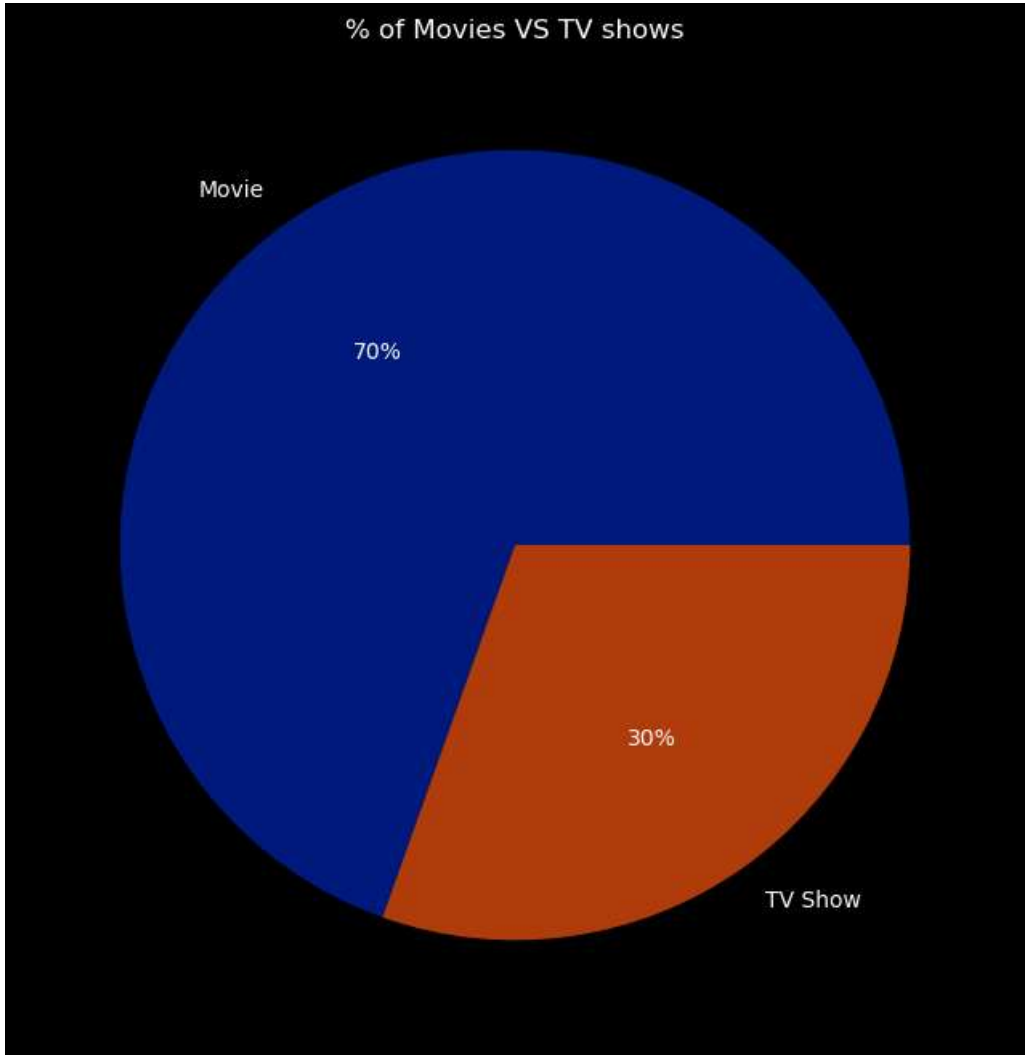


In [61]:
```
df_pie = df_.groupby(['type']).agg({'title':'nunique'}).reset_index()
df_pie
```

Out[61]:

	type	title
0	Movie	6115
1	TV Show	2676

In [62]:
```
colors = sns.color_palette('dark')[0:9]
plt.figure(figsize=(20,8))

plt.pie(df_pie['title'], labels = df_pie['type'], colors = colors, autopct='%.0f%%')
plt.title('% of Movies VS TV shows')
plt.show()
```



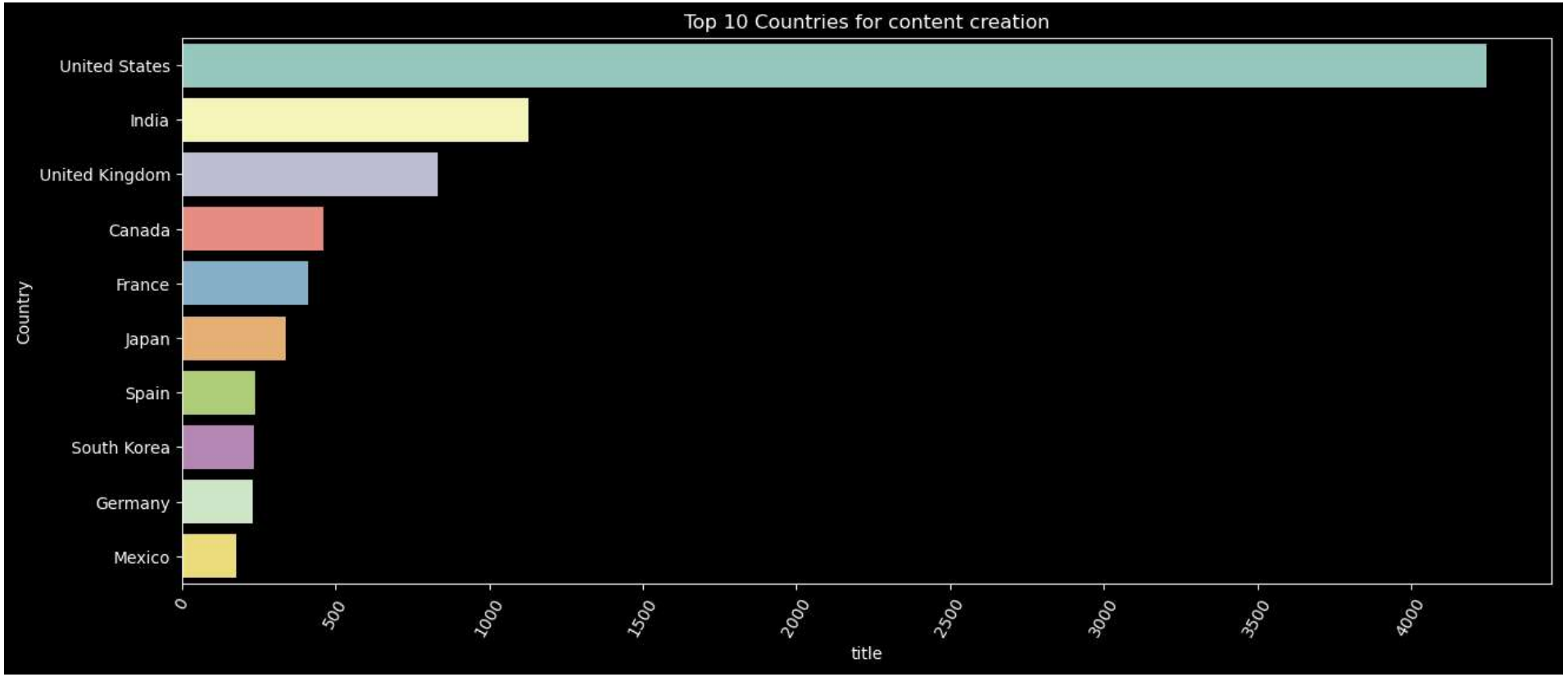
In [63]:
```
df_['Country'] = df_['Country'].str.replace(',', '')
df_.head()
```

Out[63]:

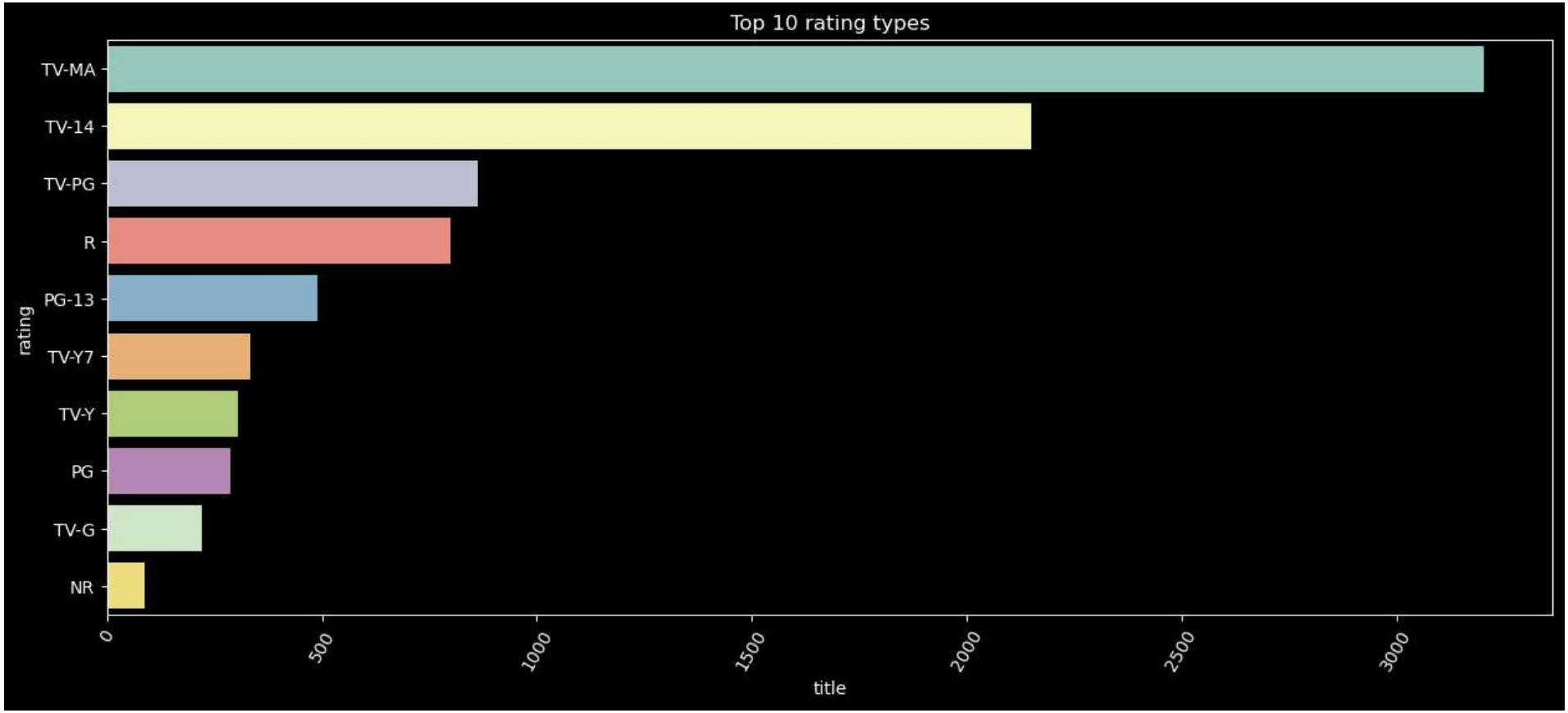
	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration	year_added	month_added	month_name	day_added	Weekday_added
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100	2021	9	September	25	Saturday
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday

In [64]:

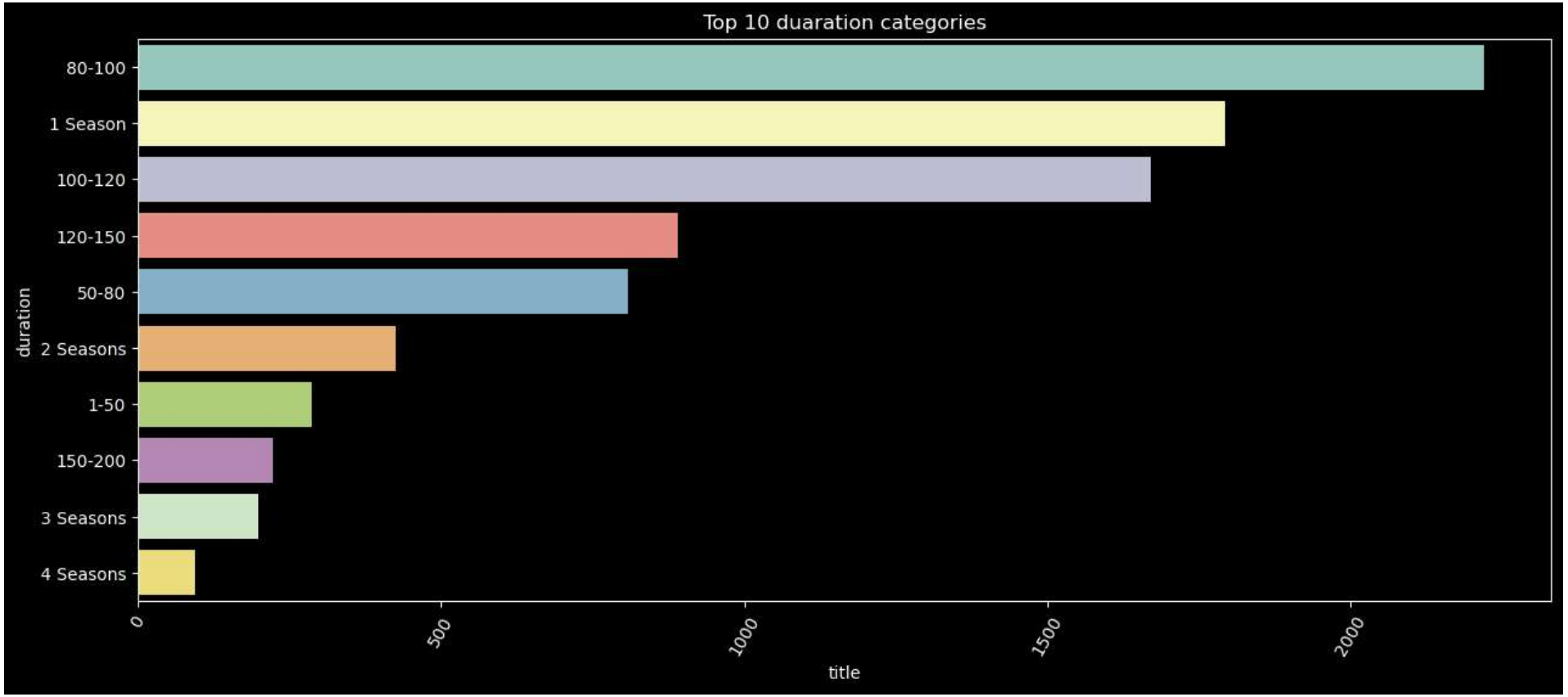
```
df_country = df_.groupby(['Country']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)[:10]
plt.figure(figsize=(15,6))
sns.barplot(y = "Country",x = 'title', data = df_country)
plt.xticks(rotation = 60)
plt.title('Top 10 Countries for content creation')
plt.show()
```



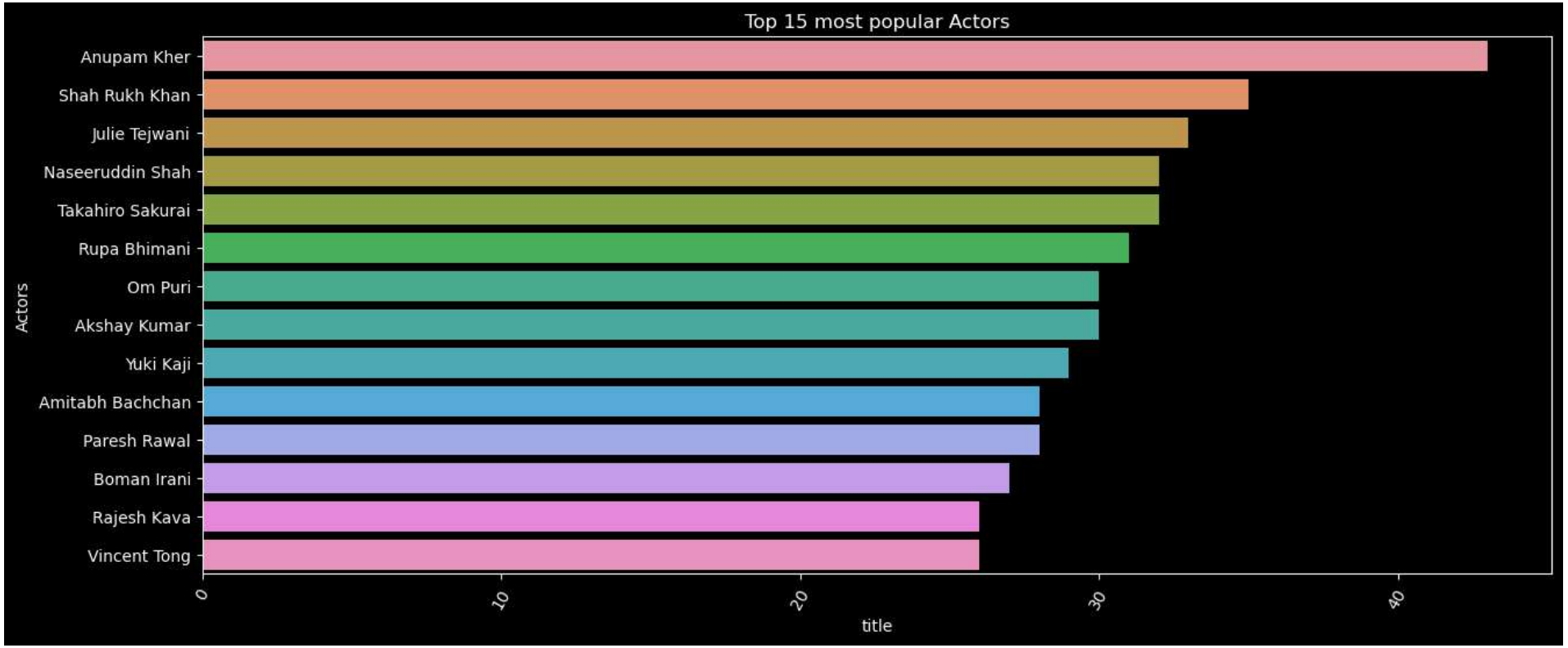
In [65]:
```
df_rating = df_.groupby(['rating']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)[:10]

plt.figure(figsize=(15,6))
sns.barplot(y = "rating",x = 'title', data = df_rating)
plt.xticks(rotation = 60)
plt.title('Top 10 rating types')
plt.show()
```



In [66]:
```
df_duration = df_.groupby(['duration']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)[:10]

plt.figure(figsize=(15,6))
sns.barplot(y = "duration",x = 'title', data = df_duration)
plt.xticks(rotation = 60)
plt.title('Top 10 duaration categories')
plt.show()
```

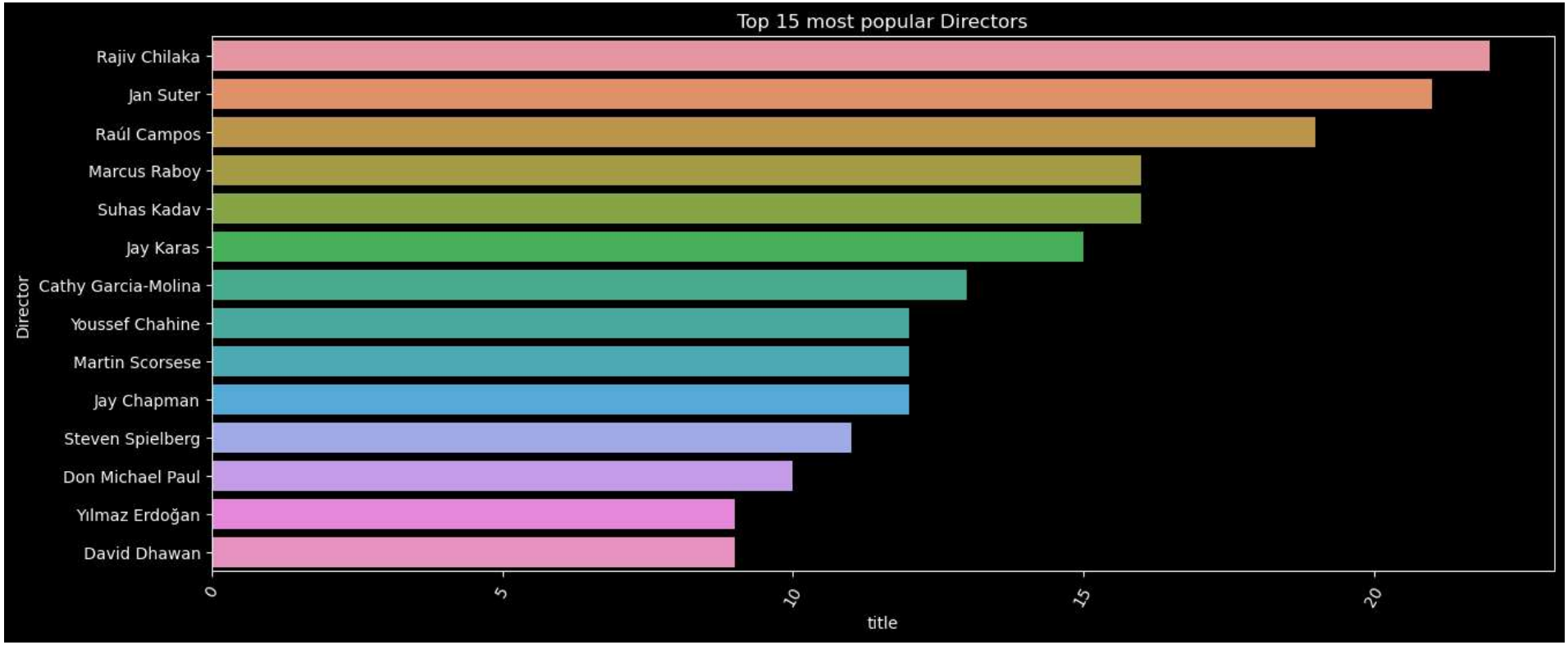


In [67]:
```
df_actors = df_.groupby(['Actors']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)[:15]
df_actors = df_actors[df_actors['Actors']!='Unknown Actor']
plt.figure(figsize=(15,6))
sns.barplot(y = "Actors",x = 'title', data = df_actors )
plt.xticks(rotation = 60)
plt.title('Top 15 most popular Actors')
plt.show()
```



In [68]:
```
df_directors = df_.groupby(['Director']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)[:15]
df_directors = df_directors[df_directors['Director']!='Unknown Director']
plt.figure(figsize=(15,6))
sns.barplot(y = "Director",x = 'title', data = df_directors )
plt.xticks(rotation = 60)
plt.title('Top 15 most popular Directors')
plt.show
```

Out[68]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



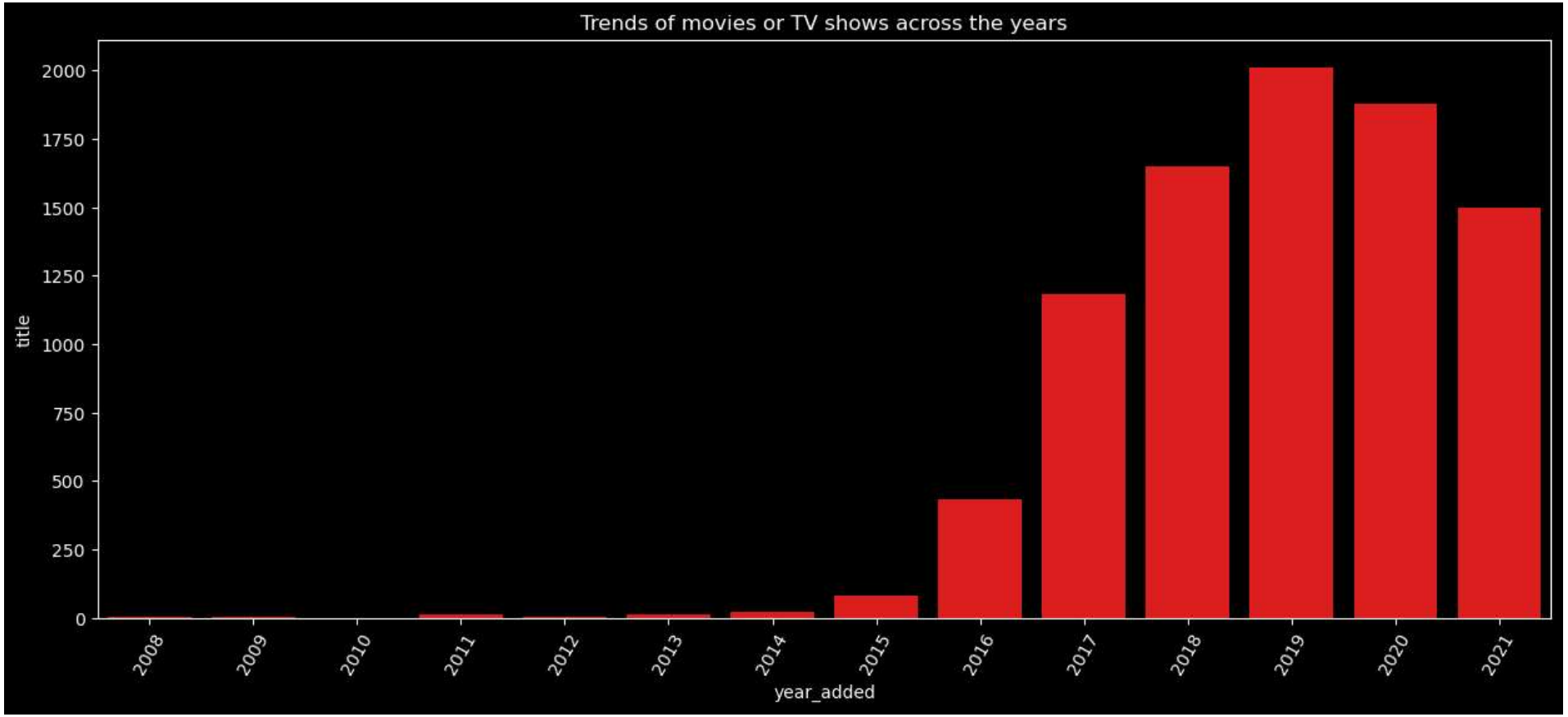
In [69]:
```
df_.head(15)
```

Out[69]:

	title	Actors	Director	Genre	Country	show_id	type	date_added	release_year	rating	duration	year_added	month_added	month_name	day_added	Weekday_added
0	Dick Johnson Is Dead	Unknown Actor	Kirsten Johnson	Documentaries	United States	s1	Movie	2021-09-25	2020	PG-13	80-100	2021	9	September	25	Saturday
1	Blood & Water	Ama Qamata	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
2	Blood & Water	Ama Qamata	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
3	Blood & Water	Ama Qamata	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
4	Blood & Water	Khosi Ngema	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
5	Blood & Water	Khosi Ngema	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
6	Blood & Water	Khosi Ngema	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
7	Blood & Water	Gail Mabalane	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
8	Blood & Water	Gail Mabalane	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
9	Blood & Water	Gail Mabalane	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
10	Blood & Water	Thabang Molaba	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
11	Blood & Water	Thabang Molaba	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
12	Blood & Water	Thabang Molaba	Unknown Director	TV Mysteries	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
13	Blood & Water	Dillon Windvogel	Unknown Director	International TV Shows	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday
14	Blood & Water	Dillon Windvogel	Unknown Director	TV Dramas	South Africa	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021	9	September	24	Friday

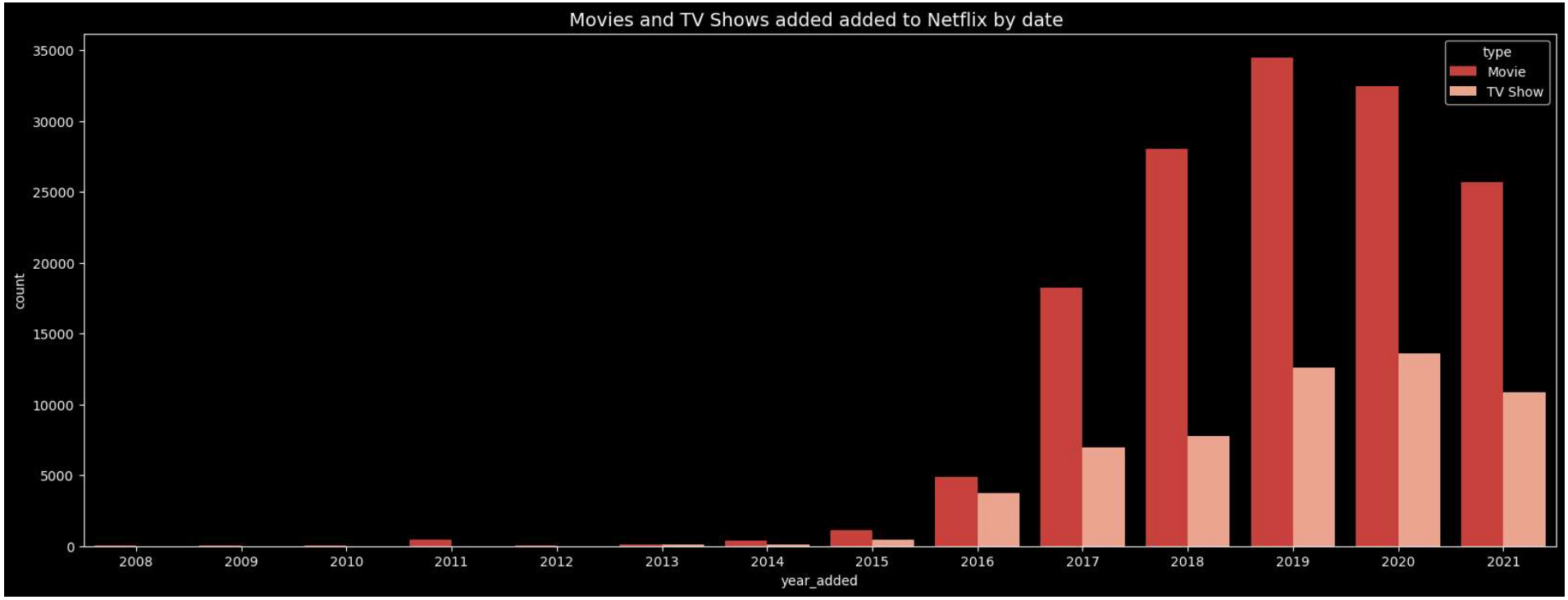
In [70]:

```
df_year = df_.groupby(['year_added']).agg({'title':'nunique'}).reset_index()

plt.figure(figsize=(15,6))
sns.barplot(x = "year_added", y = 'title', data = df_year, color = 'red')
plt.xticks(rotation = 60)
plt.title('Trends of movies or TV shows across the years')
plt.show()
```



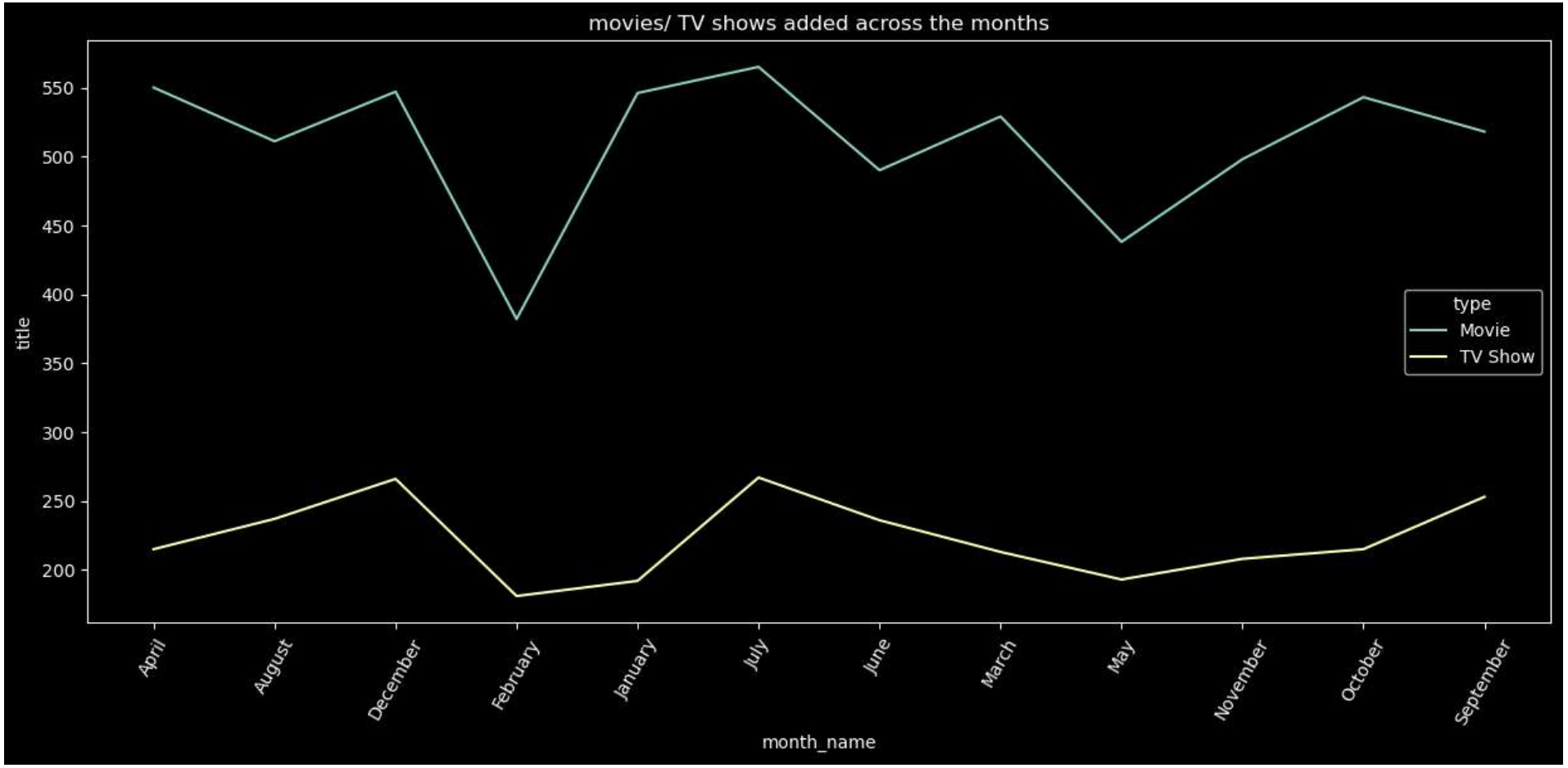
In [71]:
```
fig = plt.figure(figsize = (20,7))

plt.style.use('dark_background')
sns.countplot(data = df_,x = 'year_added',hue = 'type',palette ="Reds_r")
plt.title('Movies and TV Shows added added to Netflix by date ', fontsize=14)
```

Out[71]:
```
Text(0.5, 1.0, 'Movies and TV Shows added added to Netflix by date ')
```



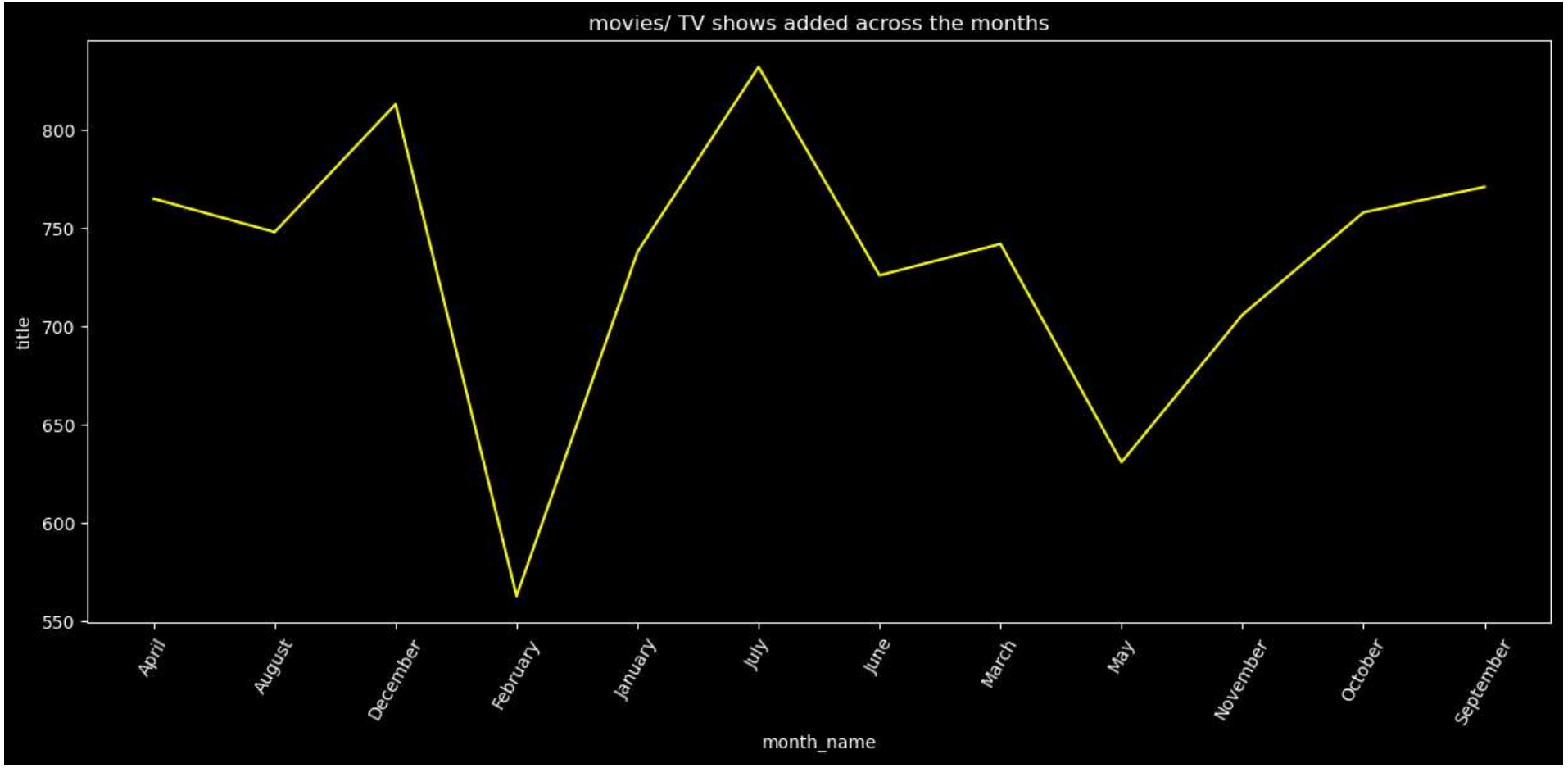
In [72]:
```
df_month = df_.groupby(['month_name', 'type']).agg({'title':'nunique'}).reset_index()

plt.figure(figsize=(15,6))
sns.lineplot(x = "month_name",y = 'title', data = df_month, color = 'red', hue = df_month.type )
plt.xticks(rotation = 60)
plt.title('movies/ TV shows added across the months')
plt.show
```

Out[72]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



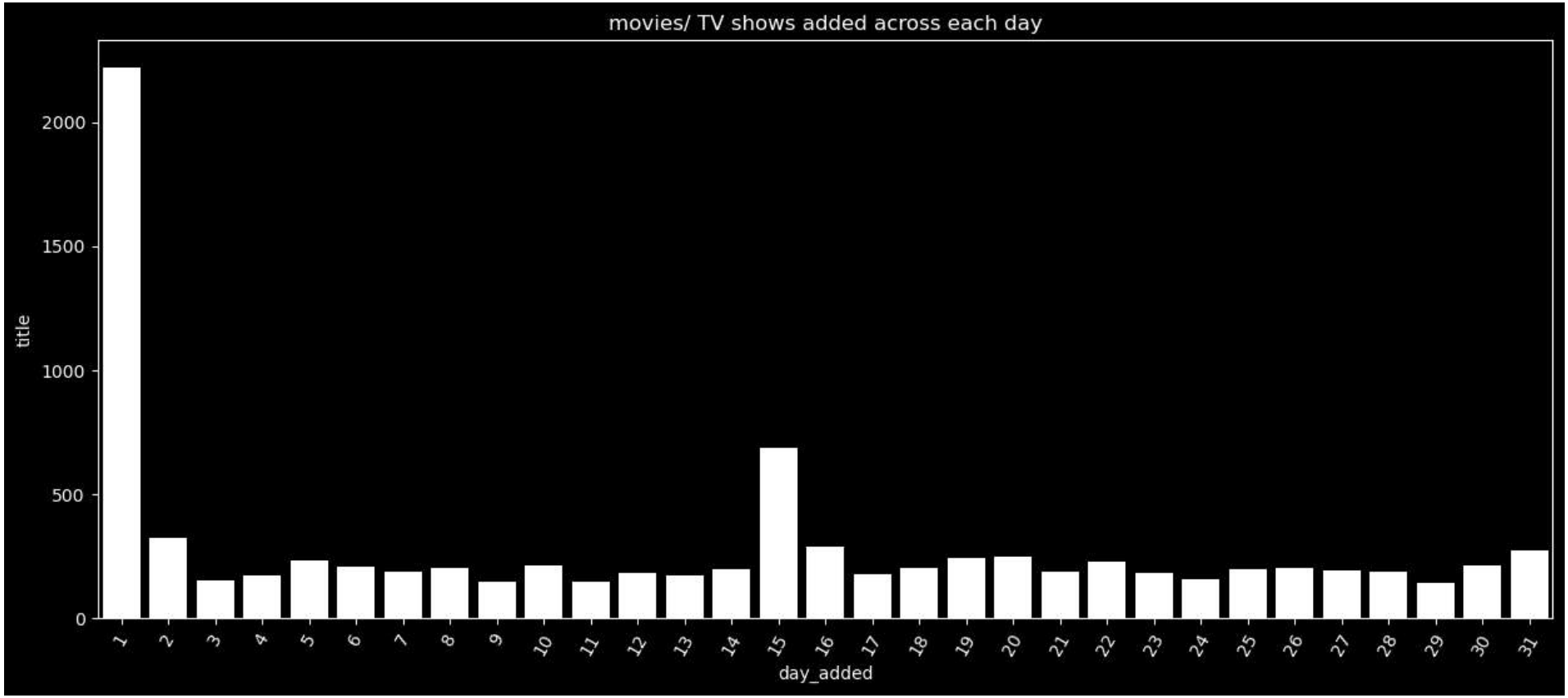
In [73]:
```
df_month = df_.groupby(['month_name']).agg({'title':'nunique'}).reset_index()

plt.figure(figsize=(15,6))
sns.lineplot(x = "month_name",y = 'title', data = df_month, color = 'yellow' )
plt.xticks(rotation = 60)
plt.title('movies/ TV shows added across the months')
plt.show
```

Out[73]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



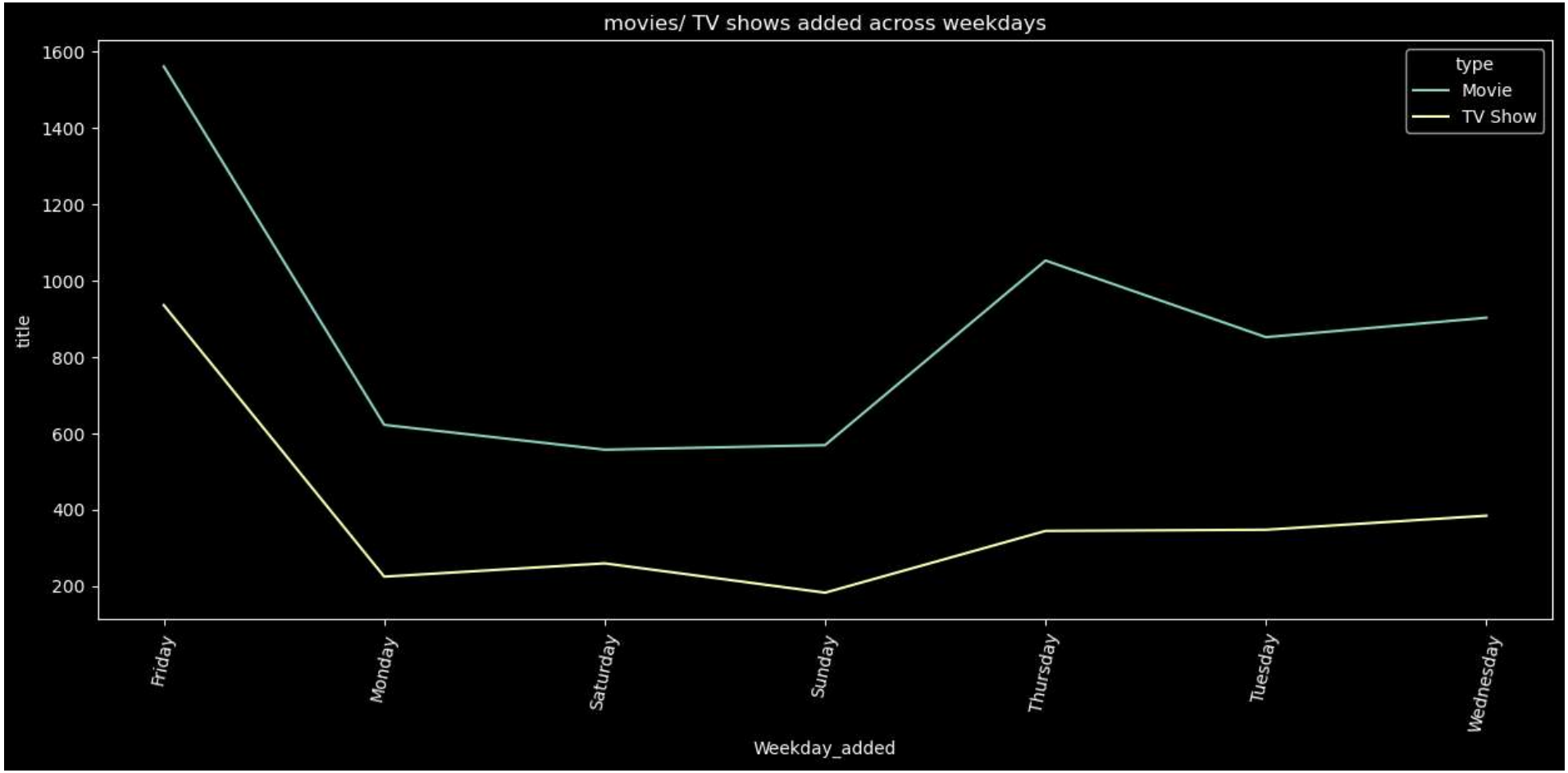
In [74]:
```
df_day = df_.groupby(['day_added']).agg({'title':'nunique'}).reset_index()

plt.figure(figsize=(15,6))
sns.barplot(x = "day_added",y = 'title', data = df_day, color = 'white' )
plt.xticks(rotation = 60)
plt.title('movies/ TV shows added across each day')
plt.show
```

Out[74]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



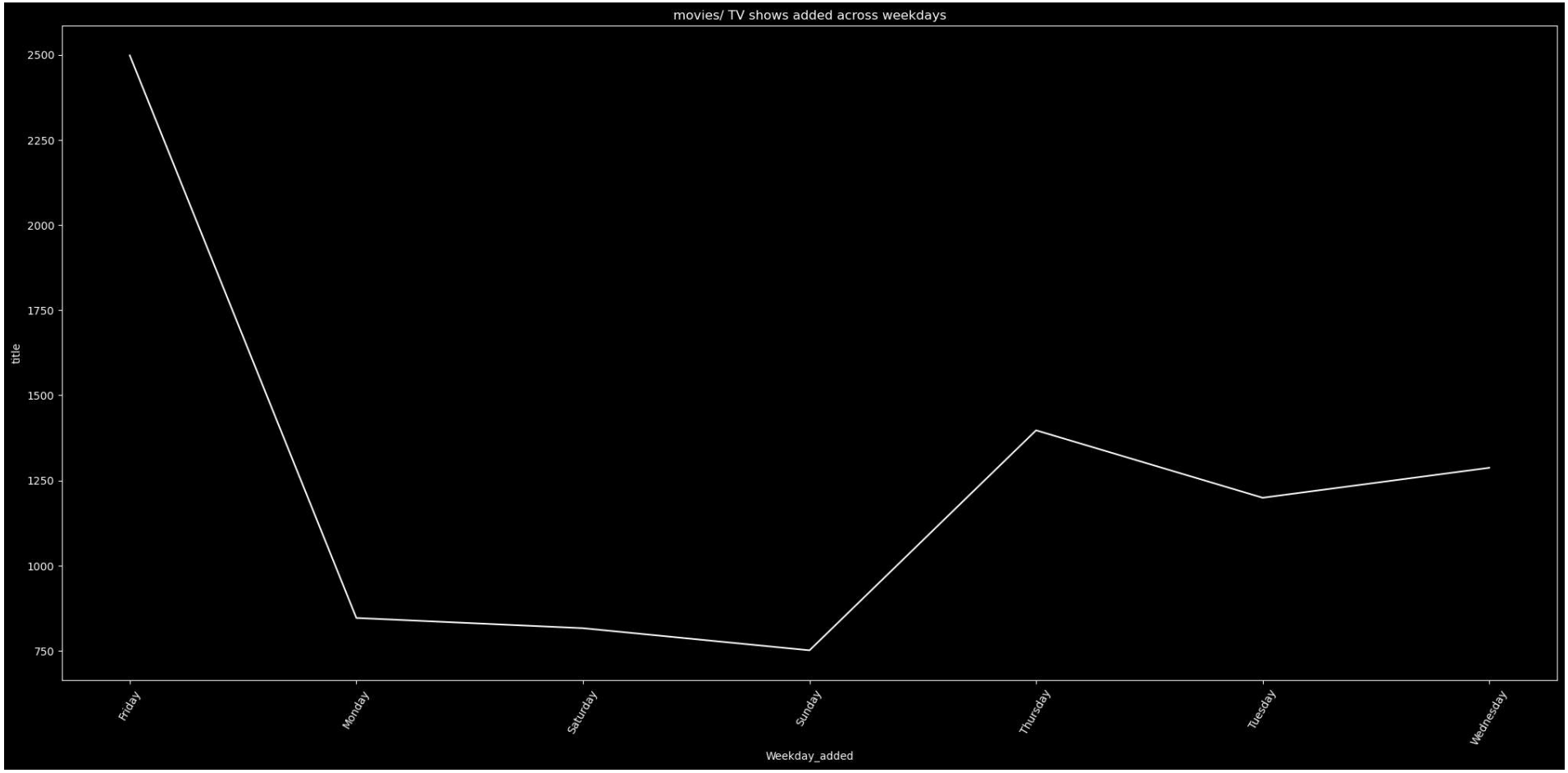
In [75]:
```
df_weekday = df_.groupby(['Weekday_added', 'type']).agg({'title':'nunique'}).reset_index()

plt.figure(figsize=(15,6))
sns.lineplot(x = "Weekday_added",y = 'title', data = df_weekday, color = 'blue' , hue = df_weekday.type)
plt.xticks(rotation = 78)
plt.title('movies/ TV shows added across weekdays')
plt.show
```

Out[75]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



In [76]:
```
df_weekday = df_.groupby(['Weekday_added']).agg({'title':'nunique'}).reset_index()

plt.figure(figsize=(25,11))
sns.lineplot(x = "Weekday_added",y = 'title', data = df_weekday, color = 'white' )
plt.xticks(rotation = 60)
plt.title('movies/ TV shows added across weekdays')
plt.show
```

Out[76]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```

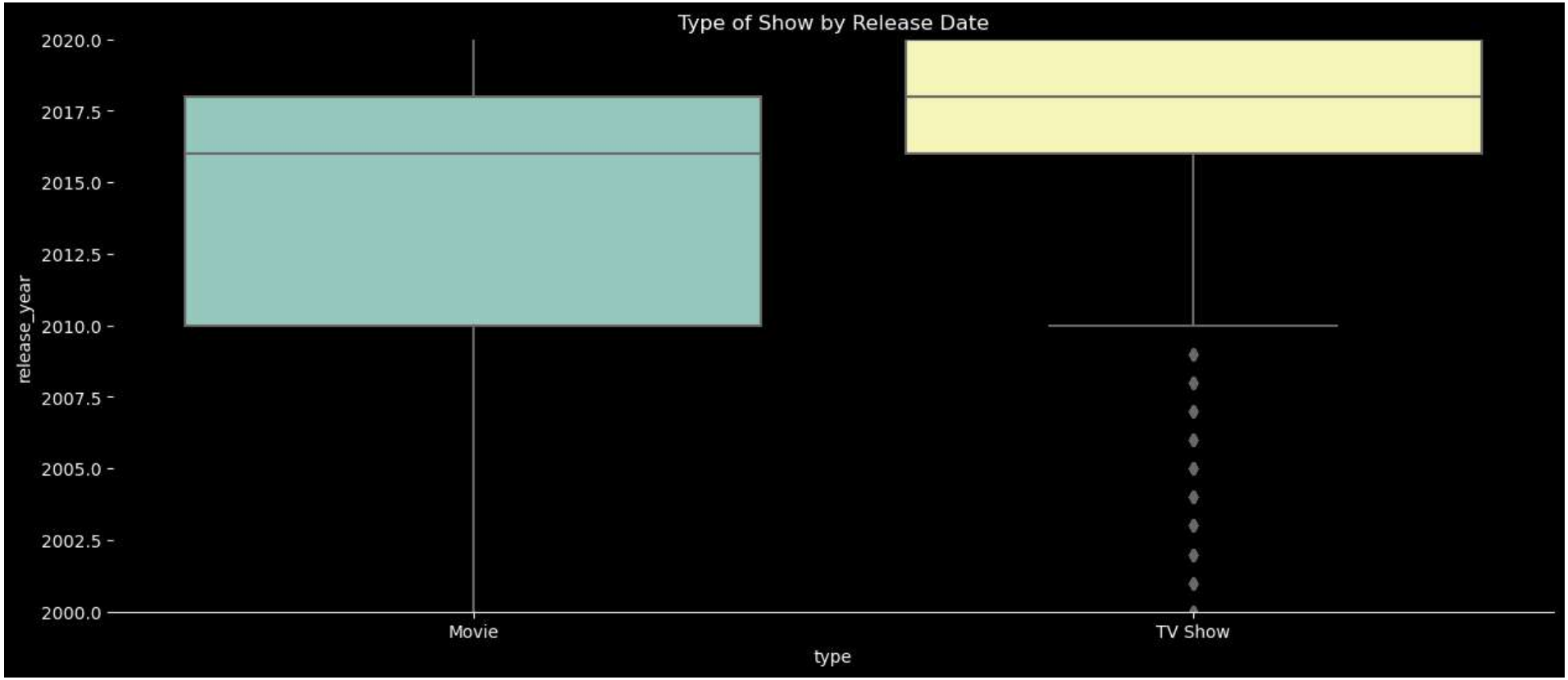


In [77]:
```
df_.columns
```

Out[77]:
```
Index(['title', 'Actors', 'Director', 'Genre', 'Country', 'show_id', 'type',
       'date_added', 'release_year', 'rating', 'duration', 'year_added',
       'month_added', 'month_name', 'day_added', 'Weekday_added'],
      dtype='object')
```

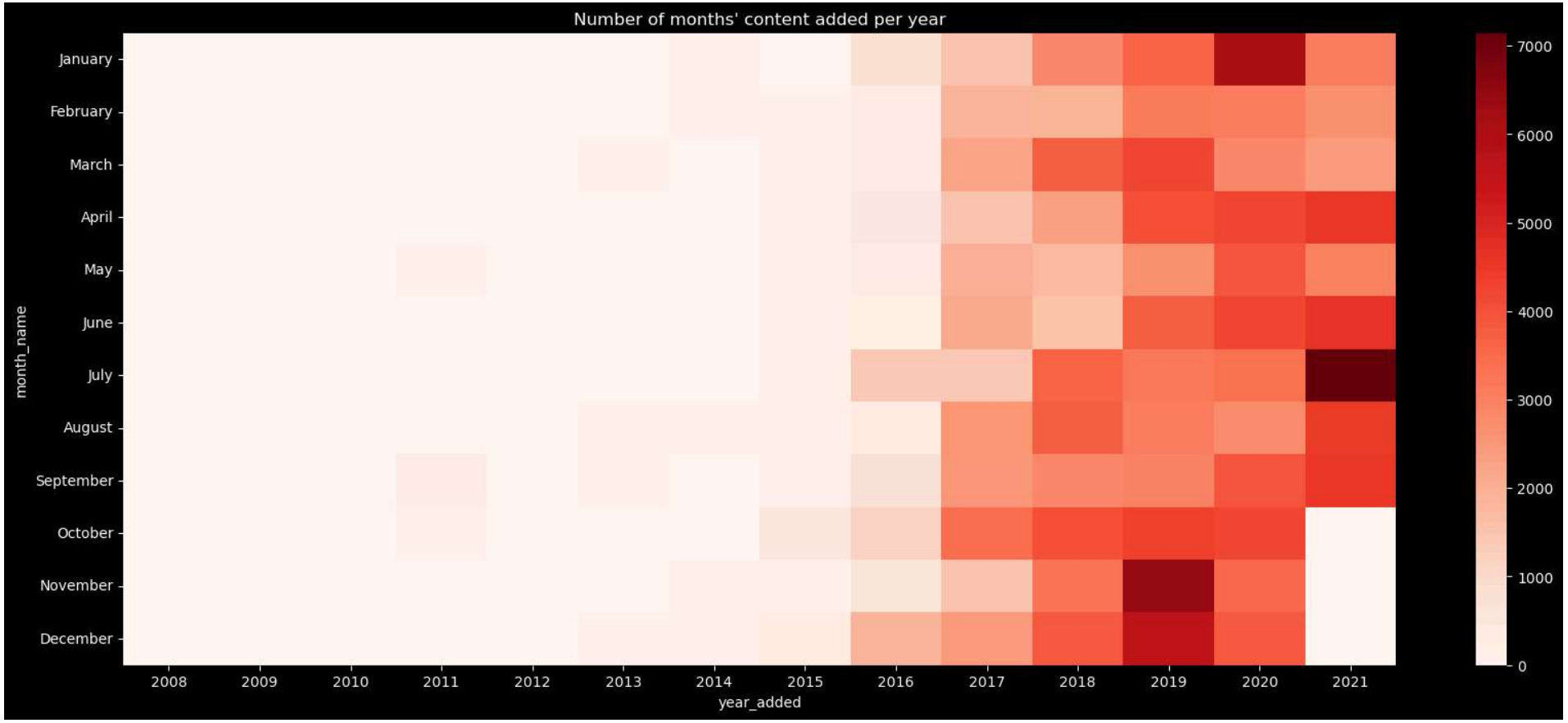
In [78]:
```
plt.figure(figsize=(15,6))
sns.boxplot(x='type', y='release_year', data=df_, )
sns.despine(left=True)
plt.title('Type of Show by Release Date')
plt.ylim(2000,2020)
```

Out[78]:
```
(2000.0, 2020.0)
```



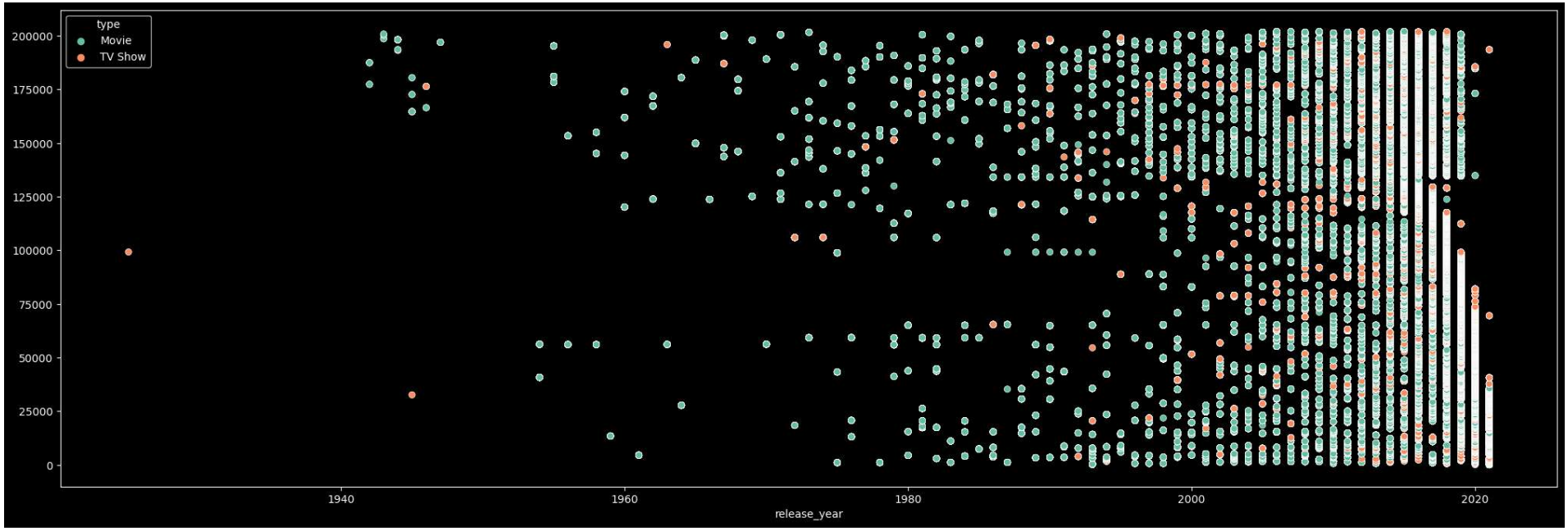
Bivariate Analysis

In [79]:
```
month_order = ['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September',
               'October', 'November', 'December']
content = df_.groupby('year_added')['month_name'].value_counts().unstack().fillna(0)[month_order].T

plt.figure(figsize=(20,8))
plt.title("Number of months' content added per year")
sns.heatmap(content , cmap = 'Reds')
plt.show()
```



In [80]:
```
plt.figure(figsize = (25,8))
sns.scatterplot(y = df_.index , x = df_.release_year , hue = df_.type , palette='Set2')
```

Out[80]:
```
<Axes: xlabel='release_year'>
```



In [81]:
```
df_.groupby(['day_added']).agg({"title":"nunique"})
```

Out[81]:

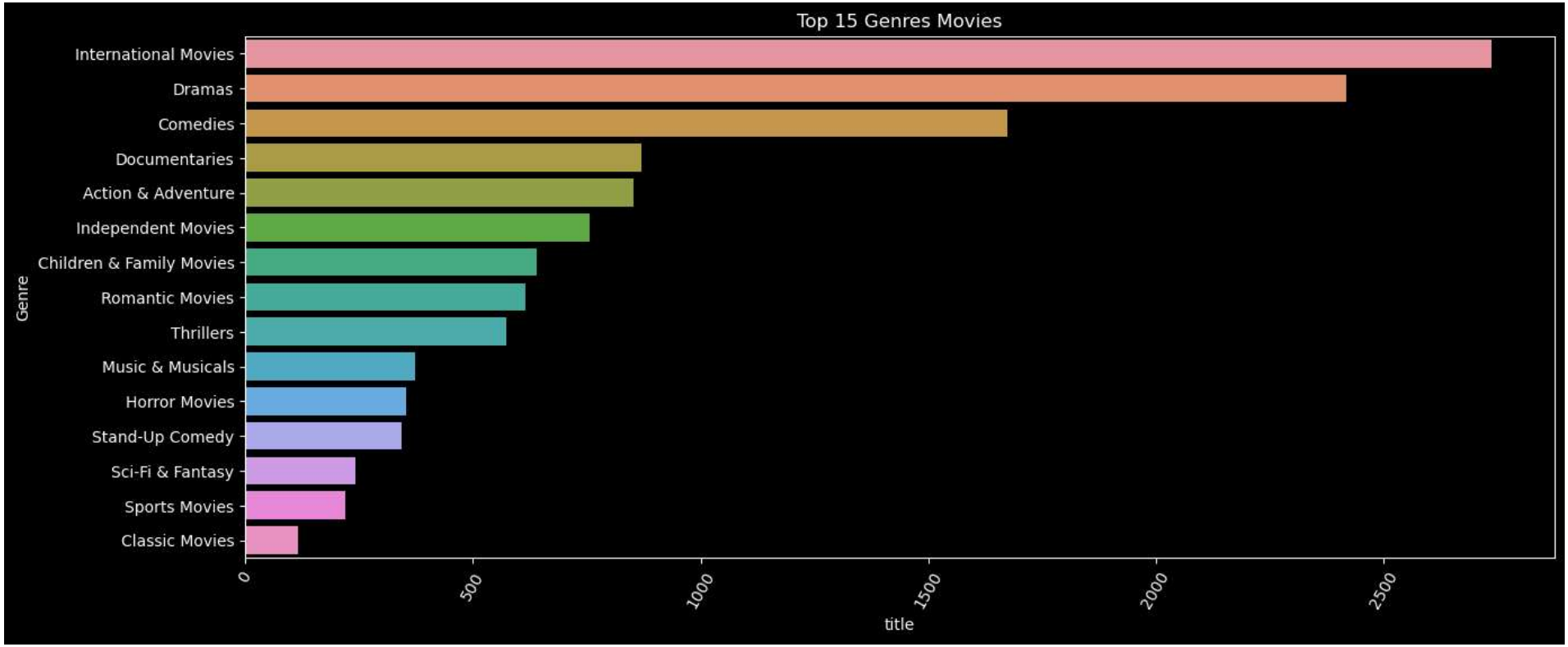
day_added	title
1	2219
2	325
3	151
4	175
5	231
6	210
7	190
8	201
9	148
10	214
11	149
12	181
13	175
14	198
15	688
16	289
17	180
18	205
19	243
20	249
21	190
22	230
23	182
24	159
25	196
26	205
27	195
28	190
29	141
30	211

day_added	title
31	274

Univariate Analysis separately for shows and movies

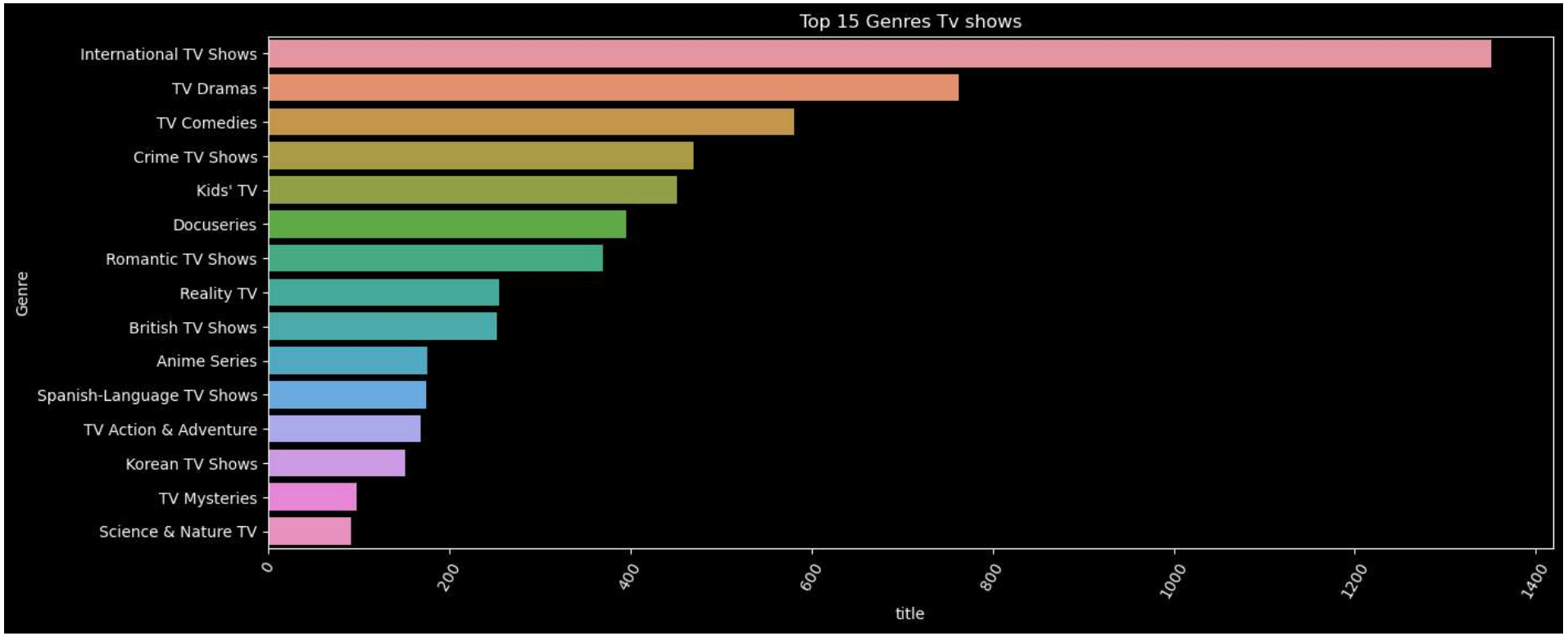
In [82]:
```
df_shows = df_[df_['type']=='TV Show']
df_movies = df_[df_['type']=='Movie']
```

In [83]:
```
df_genre = df_movies.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
plt.figure(figsize = (15,6))
sns.barplot(y = "Genre",x = 'title', data = df_genre)
plt.xticks(rotation = 60)
plt.title('Top 15 Genres Movies')
plt.show() # movies
```

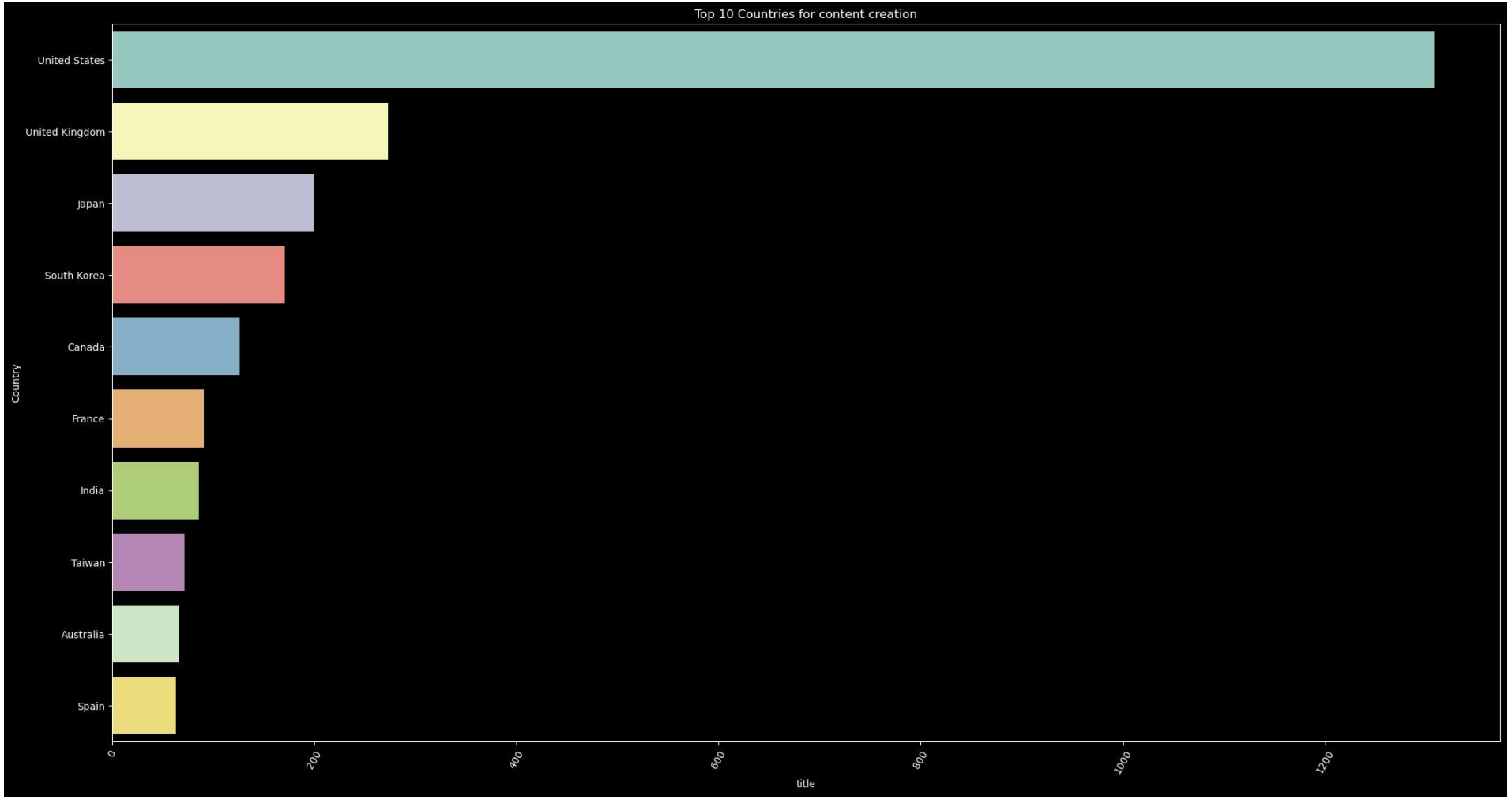


In [84]:
```
df_genre = df_shows.groupby(['Genre']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
plt.figure(figsize = (15,6))
sns.barplot(y = "Genre",x = 'title', data = df_genre)
```

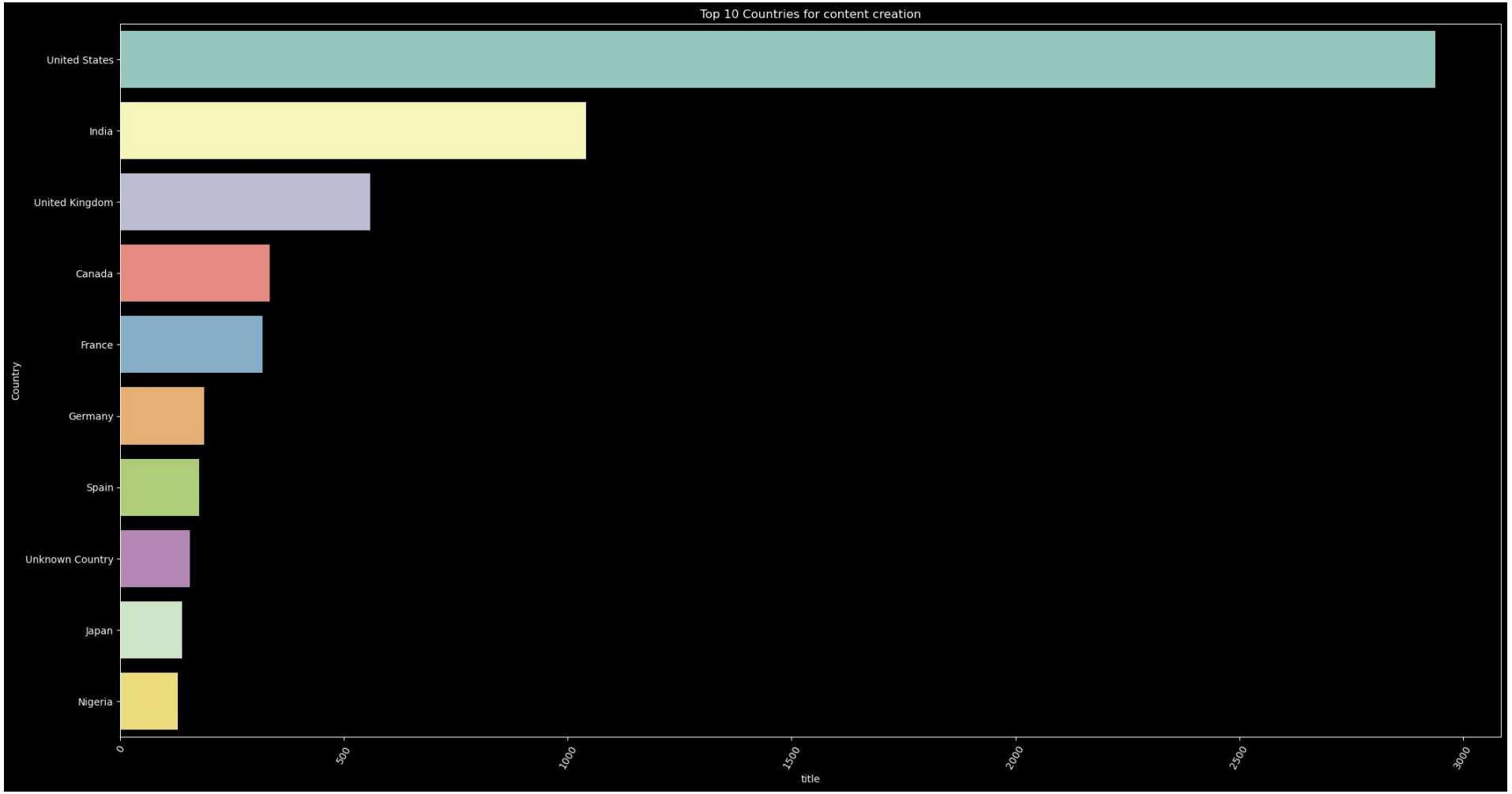
```
plt.xticks(rotation = 60)
plt.title('Top 15 Genres Tv shows')
plt.show() # Tv shows
```



In [85]:
```
df_country = df_shows.groupby(['Country']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)[:10]
plt.figure(figsize=(25,13))
sns.barplot(y = "Country",x = 'title', data = df_country)
plt.xticks(rotation = 60)
plt.title('Top 10 Countries for content creation')
plt.show()
```



In [86]:
```
df_country = df_movies.groupby(['Country']).agg({'title':'nunique'}).reset_index().sort_values(by=['title'],ascending=False)[:10]
plt.figure(figsize=(25,13))
sns.barplot(y = "Country",x = 'title', data = df_country)
plt.xticks(rotation = 60)
plt.title('Top 10 Countries for content creation')
plt.show()
```



United States is leading across both TV Shows and Movies, UK also provides great content across TV Shows and Movies. Surprisingly India is much more prevalent in Movies as compared TV Shows.

Moreover the number of Movies created in India outweigh the sum of TV Shows and Movies across UK since India was rated as second in net sum of whole content across Netflix.

Business Insights

1. Over the years both TV shows and movie contents addtion has increased till 2020, but after 2020 its started declining may be due to Covid relief, number of Movies added is more compare to TV shows over the years.
2. Most of the content get added in december and july month, for day wise, Friday is the best day followed by Thursday.

3. It was evident that 1st of every month was when the most content was added.
4. Anupam Kher,SRK,Julie Tejwani, Naseeruddin Shah and Takahiro Sakurai occupy the top stop in Most Watched content.
5. Rajiv Chilaka, Jan Suter and Raul Campos are the most popular directors across Netflix. Rajiv Chilaka director producing more movies.
6. Netflix is more focusing on movies compare to TV shows.
7. There is a 70% & 30% of Movies and TV Shows content in Netflix platform.
8. International Movies, Dramas and Comedies are the most popular are most popular Genre.
9. US,India,UK,Canada and France are leading countries in Content Creation on Netflix.
10. Most of the highly rated content on Netflix is intended for TV - Mature Audiences(TV-MA)
11. The duration of Most Watched content in our whole data is 80-120 mins. These must be movies and Shows having only 1 Season.
12. United States is leading across both TV Shows and Movies, UK also provides great content across TV Shows and Movies. Surprisingly India is much more prevalent in Movies as compared TV Shows.
13. India has second position in creating movies & UK has 2nd in Tv shows as well

Recommendations

1. The most popular Genres across the countries and in both TV Shows and Movies are Drama, Comedy and International TV Shows/Movies, so recommended to generate more content on these genres.
2. Add TV Shows/ movies in the month of July 1st / August 1st.
3. Add movies for Indian Audience, it has been declining since 2018.
4. While creating content, take into consideration the popular actors/directors for that country. Also take into account the director-actor combination which is highly recommended.
5. 80-120 mins is avagrage watch by audiance so eye on this also towads creating the content.
6. As per Most Rating TV-MA so we can create more content according to this followed by TV-14, TV-PG & R.