In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df=pd.read_csv(r'\Users\Home\Downloads\Dataset_link.csv')
df
```

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

1. Defining Problem Statement and Analysing basic metrics

Netflix is a media streaming platform with more than 1000s of movies, TV shows, documentaries, etc. The above 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.

This project aims at exploring the above data about Netflix in order to make observations and give useful insights and recommendations to help the business to grow. Further, this project aims to observe viewers' behavior by analysing the popularity of movies and TV shows across different countries.

2. Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary

In [3]:
```
#Shape of the data
df.shape
```

Out[3]:
```
(8807, 12)
```

In [4]:
```
#Data types of all attributes
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
```

We can observe that the data type is object and integer. There are also some null values in the dataset. The range index is 0 to 8806 with 12 columns.

Statistical Summary

This includes the unique count, frequency, mean, min, max and other important values of the dataset given below.

In [5]:
```
df.describe(include='all')
```

Out[5]:

	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 [6]:
```
#Missing Value Detection
df.isna().sum()
```

Out[6]:
```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

Pre-processing of Data

In [7]:
```
#unnesting of the data
Director=df['director'].apply(lambda x:str(x).split(', ')).tolist()
df_new1=pd.DataFrame(Director,index=df['title'])
df_new1=df_new1.stack()
df_new1=pd.DataFrame(df_new1.reset_index())
df_new1.rename(columns={0:'Directors'},inplace=True)
df_new1.drop(['level_1'],axis=1,inplace=True)
df_new1
```

Out[7]:

	title	Directors
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	nan
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	nan
4	Kota Factory	nan
...	...	...
9607	Zodiac	David Fincher
9608	Zombie Dumb	nan
9609	Zombieland	Ruben Fleischer
9610	Zoom	Peter Hewitt
9611	Zubaan	Mozez Singh

9612 rows × 2 columns

In [8]:
```
Cast=df['cast'].apply(lambda x:str(x).split(', ')).tolist()
df_new2=pd.DataFrame(Cast,index=df['title'])
df_new2=df_new2.stack()
df_new2=pd.DataFrame(df_new2.reset_index())
df_new2.rename(columns={0:'Actors'},inplace=True)
df_new2.drop(['level_1'],axis=1,inplace=True)
df_new2
```

Out[8]:

	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
...	...	...
64946	Zubaan	Manish Chaudhary
64947	Zubaan	Meghna Malik
64948	Zubaan	Malkeet Rauni
64949	Zubaan	Anita Shabdish
64950	Zubaan	Chittaranjan Tripathy

64951 rows × 2 columns

In [9]:
```
Country=df['country'].apply(lambda x:str(x).split(', ')).tolist()
df_new3=pd.DataFrame(Country,index=df['title'])
df_new3=df_new3.stack()
df_new3=pd.DataFrame(df_new3.reset_index())
df_new3.rename(columns={0:'Countries'},inplace=True)
df_new3.drop(['level_1'],axis=1,inplace=True)
df_new3
```

Out[9]:

	title	Countries
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	nan
3	Jailbirds New Orleans	nan
4	Kota Factory	India
...	...	...
10840	Zodiac	United States
10841	Zombie Dumb	nan
10842	Zombieland	United States
10843	Zoom	United States
10844	Zubaan	India

10845 rows × 2 columns

In [10]:
```
genre=df['listed_in'].apply(lambda x:str(x).split(', ')).tolist()
df_new4=pd.DataFrame(genre,index=df['title'])
df_new4=df_new4.stack()
df_new4=pd.DataFrame(df_new4.reset_index())
df_new4.rename(columns={0:'Genre'},inplace=True)
df_new4.drop(['level_1'],axis=1,inplace=True)
df_new4
```

Out[10]:

	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
...	...	...
19318	Zoom	Children & Family Movies
19319	Zoom	Comedies
19320	Zubaan	Dramas
19321	Zubaan	International Movies
19322	Zubaan	Music & Musicals

19323 rows × 2 columns

In [11]:
```
#merging the unnested directors data (df_new1) and actors data (df_new2)
df_new5=df_new1.merge(df_new2, on=['title'], how='inner')
#merging df_new5 with the unnested genre data (df_new3)
df_new6=df_new5.merge(df_new3, on=['title'], how='inner')
#merging df_new6 with the unnested countries data (df_new4)
df_new=df_new6.merge(df_new4, on=['title'], how='inner')
df_new
```

Out[11]:

	title	Directors	Actors	Countries	Genre
0	Dick Johnson Is Dead	Kirsten Johnson	nan	United States	Documentaries
1	Blood & Water	nan	Ama Qamata	South Africa	International TV Shows
2	Blood & Water	nan	Ama Qamata	South Africa	TV Dramas
3	Blood & Water	nan	Ama Qamata	South Africa	TV Mysteries
4	Blood & Water	nan	Khosi Ngema	South Africa	International TV Shows
...	...	...	...	...	...
201986	Zubaan	Mozez Singh	Anita Shabdish	India	International Movies
201987	Zubaan	Mozez Singh	Anita Shabdish	India	Music & Musicals
201988	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Dramas
201989	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	International Movies
201990	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Music & Musicals

201991 rows × 5 columns

In [12]:
```
#replacing nan values in the unnested data
df_new['Actors'].replace(['nan'],['Unknown Actor'], inplace=True)
df_new['Directors'].replace(['nan'],['Unknown Director'], inplace=True)
df_new['Countries'].replace(['nan'],['Unknown Country'], inplace=True)
df_new['Genre'].replace(['nan'],['Unknown Genre'], inplace=True)
df_new
```

Out[12]:

	title	Directors	Actors	Countries	Genre
0	Dick Johnson Is Dead	Kirsten Johnson	Unknown Actor	United States	Documentaries
1	Blood & Water	Unknown Director	Ama Qamata	South Africa	International TV Shows
2	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Dramas
3	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Mysteries
4	Blood & Water	Unknown Director	Khosi Ngema	South Africa	International TV Shows
...	...	...	...	...	...
201986	Zubaan	Mozez Singh	Anita Shabdish	India	International Movies
201987	Zubaan	Mozez Singh	Anita Shabdish	India	Music & Musicals
201988	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Dramas
201989	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	International Movies
201990	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Music & Musicals

201991 rows × 5 columns

In [13]:
```
#merging unnested data with original data
df_final=df_new.merge(df[['show_id', 'type', 'title', 'date_added', 'release_year',
                          'rating', 'duration']], on=['title'], how='left')
df_final
```

Out[13]:

	title	Directors	Actors	Countries	Genre	show_id	type	date_added	release_year	rating	duration
0	Dick Johnson Is Dead	Kirsten Johnson	Unknown Actor	United States	Documentaries	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Unknown Director	Ama Qamata	South Africa	International TV Shows	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Dramas	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Mysteries	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Unknown Director	Khosi Ngema	South Africa	International TV Shows	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
...	...	...	...	...	...	...	...	...	...	...	...
201986	Zubaan	Mozez Singh	Anita Shabdish	India	International Movies	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201987	Zubaan	Mozez Singh	Anita Shabdish	India	Music & Musicals	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201988	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Dramas	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201989	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	International Movies	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201990	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Music & Musicals	s8807	Movie	March 2, 2019	2015	TV-14	111 min

201991 rows × 11 columns

In [14]:
```
#renaming columns
df_final.rename(columns = {'title':'Title','show_id':'ID','type':'Type','date_added':'Date_Added',
                           'release_year':'Release_Year','rating':'Rating','duration':'Duration'},inplace=True)
df_final
```

Out[14]:

	Title	Directors	Actors	Countries	Genre	ID	Type	Date_Added	Release_Year	Rating	Duration
0	Dick Johnson Is Dead	Kirsten Johnson	Unknown Actor	United States	Documentaries	s1	Movie	September 25, 2021	2020	PG-13	90 min
1	Blood & Water	Unknown Director	Ama Qamata	South Africa	International TV Shows	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
2	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Dramas	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
3	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Mysteries	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
4	Blood & Water	Unknown Director	Khosi Ngema	South Africa	International TV Shows	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons
...	...	...	...	...	...	...	...	...	...	...	...
201986	Zubaan	Mozez Singh	Anita Shabdish	India	International Movies	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201987	Zubaan	Mozez Singh	Anita Shabdish	India	Music & Musicals	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201988	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Dramas	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201989	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	International Movies	s8807	Movie	March 2, 2019	2015	TV-14	111 min
201990	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Music & Musicals	s8807	Movie	March 2, 2019	2015	TV-14	111 min

201991 rows × 11 columns

3. Non-Graphical Analysis: Value counts and unique attributes

In [15]:
```
df_final.Title.value_counts()
```

Out[15]:
```
Kahlil Gibran's The Prophet    700
Holidays                       504
Movie 43                       468
The Eddy                       416
Narcos                         378
                              ...
Thackeray                        1
The 2000s                        1
Miniforce: Super Dino Power      1
Dancing with the Birds           1
Dick Johnson Is Dead             1
Name: Title, Length: 8807, dtype: int64
```

In [16]:
```
df_final.Directors.value_counts()
```

Out[16]:
```
Unknown Director       50643
Martin Scorsese          419
Youssef Chahine          409
Cathy Garcia-Molina      356
Steven Spielberg         355
                        ...
Richard Maurice            1
Richard E. Norman          1
Spencer Williams           1
Oscar Micheaux             1
Kirsten Johnson            1
Name: Directors, Length: 4994, dtype: int64
```

In [17]:
```
df_final.Actors.value_counts()
```

Out[17]:
```
Unknown Actor     2146
Liam Neeson        161
Alfred Molina      160
John Krasinski     139
Salma Hayek        130
                  ...
Dario Yazbek         1
Corinne Foxx         1
Jacob Craner         1
Laila Berzins        1
Richard Ryan         1
Name: Actors, Length: 36440, dtype: int64
```

In [18]:
```
df_final.Countries.value_counts()
```

Out[18]:
```
United States      59349
India              22814
United Kingdom     12945
Unknown Country    11897
Japan               8679
                   ...
Palestine              2
Kazakhstan             1
Nicaragua              1
United States,         1
Uganda                 1
Name: Countries, Length: 128, dtype: int64
```

In [19]:
```
df_final.Genre.value_counts()
```

Out[19]:
```
Dramas                          29775
International Movies            28211
Comedies                        20829
International TV Shows          12845
Action & Adventure              12216
Independent Movies               9834
Children & Family Movies         9771
TV Dramas                        8942
Thrillers                        7107
Romantic Movies                  6412
TV Comedies                      4963
Crime TV Shows                   4733
Horror Movies                    4571
Kids' TV                         4568
Sci-Fi & Fantasy                 4037
Music & Musicals                 3077
Romantic TV Shows                3049
Documentaries                    2407
Anime Series                     2313
TV Action & Adventure            2288
Spanish-Language TV Shows        2126
British TV Shows                 1808
Sports Movies                    1531
Classic Movies                   1434
TV Mysteries                     1281
Korean TV Shows                  1122
Cult Movies                      1077
TV Sci-Fi & Fantasy              1045
Anime Features                   1045
TV Horror                         941
Docuseries                        845
LGBTQ Movies                      838
TV Thrillers                      768
Teen TV Shows                     742
Reality TV                        735
Faith & Spirituality              719
Stand-Up Comedy                   540
Movies                            412
TV Shows                          337
Classic & Cult TV                 272
Stand-Up Comedy & Talk Shows      268
Science & Nature TV               157
Name: Genre, dtype: int64
```

In [20]:
```
df_final.ID.value_counts()
```

Out[20]:
```
s7165    700
s6985    504
s7516    468
s2554    416
s5306    378
        ...
s8174      1
s8176      1
s937       1
s3387      1
s1         1
Name: ID, Length: 8807, dtype: int64
```

In [21]:
```
df_final.Type.value_counts()
```

Out[21]:
```
Movie      145843
TV Show     56148
Name: Type, dtype: int64
```

In [22]:
```
df_final.Date_Added.value_counts()
```

Out[22]:
```
January 1, 2020       3730
November 1, 2019      2229
July 1, 2021          2219
October 1, 2017       1899
September 1, 2021     1756
                      ...
September 19, 2017       1
August 8, 2017           1
October 10, 2017         1
February 4, 2008         1
September 25, 2021       1
Name: Date_Added, Length: 1767, dtype: int64
```

In [23]:
```
df_final.Release_Year.value_counts()
```

Out[23]:
```
2018    24414
2019    21931
2017    20516
2020    19679
2016    18465
        ...
1947        8
1946        6
1942        6
1943        5
1925        1
Name: Release_Year, Length: 74, dtype: int64
```

In [24]:
```
df_final.Rating.value_counts()
```

Out[24]:
```
TV-MA       73867
TV-14       43931
R           25860
PG-13       16246
TV-PG       14926
PG          10919
TV-Y7        6304
TV-Y         3665
TV-G         2779
NR           1573
G            1530
NC-17         149
TV-Y7-FV       86
UR             86
74 min          1
84 min          1
66 min          1
Name: Rating, dtype: int64
```

In [25]:
```
#As Rating can't be in 'min', hence the same have been removed
df_final.loc[df_final['Rating'].str.contains('min', na=False), 'Rating']='NR'
df_final['Rating'].fillna('NR', inplace= True)
pd.set_option('display.max_rows',None)
```

In [26]:
```
df_final.Rating.value_counts()
```

Out[26]:
```
TV-MA       73867
TV-14       43931
R           25860
PG-13       16246
TV-PG       14926
PG          10919
TV-Y7        6304
TV-Y         3665
TV-G         2779
NR           1643
G            1530
NC-17         149
TV-Y7-FV       86
UR             86
Name: Rating, dtype: int64
```

In [27]:
```
df_final.Duration.value_counts()
```

Out[27]:
```
1 Season      35035
2 Seasons      9559
3 Seasons      5084
94 min         4343
106 min        4040
97 min         3624
95 min         3560
96 min         3484
93 min         3480
90 min         3305
105 min        3209
107 min        3103
101 min        3048
102 min        3017
103 min        2985
98 min         2984
99 min         2956
91 min         2915
92 min         2863
104 min        2822
88 min         2781
110 min        2711
100 min        2697
108 min        2614
112 min        2594
85 min         2486
89 min         2420
86 min         2213
4 Seasons      2134
```

116 min	2122
118 min	2119
119 min	2075
87 min	2063
109 min	2020
113 min	1990
120 min	1845
117 min	1770
121 min	1728
5 Seasons	1698
111 min	1667
124 min	1590
114 min	1529
127 min	1505
115 min	1444
123 min	1398
125 min	1299
122 min	1298
84 min	1267
128 min	1241
130 min	1216
126 min	1205
81 min	1203
83 min	1192
133 min	1169
137 min	1122
82 min	1100
136 min	1092
132 min	1047
131 min	913
135 min	851
7 Seasons	843
129 min	837
75 min	794
148 min	671
140 min	658
6 Seasons	633
79 min	629
139 min	617
143 min	608
80 min	586
134 min	572
145 min	549
149 min	540
138 min	540
74 min	516
78 min	506
141 min	495
72 min	470
142 min	464
46 min	451
77 min	447
150 min	442
172 min	432
158 min	424
76 min	408
73 min	408
151 min	395
147 min	379
163 min	371
154 min	356
146 min	342
162 min	333
54 min	323
144 min	303
153 min	300
71 min	297
70 min	289
8 Seasons	286
157 min	284
155 min	275
68 min	263
9 Seasons	257
24 min	252
161 min	230
166 min	228
10 Seasons	220
156 min	214
58 min	197
176 min	192
152 min	186
168 min	178
165 min	177
171 min	174
160 min	169
185 min	166
22 min	162
69 min	160

44 min	149
173 min	144
181 min	144
63 min	141
180 min	133
13 Seasons	132
159 min	132
26 min	128
170 min	120
177 min	117
23 min	116
60 min	114
64 min	113
28 min	113
12 Seasons	111
164 min	110
200 min	108
59 min	107
51 min	105
66 min	105
30 min	104
61 min	100
52 min	99
65 min	98
15 Seasons	96
62 min	92
33 min	91
25 min	86
47 min	81
187 min	78
182 min	76
42 min	74
67 min	74
56 min	73
48 min	73
186 min	72
57 min	71
40 min	68
179 min	66
32 min	64
27 min	62
224 min	60
208 min	60
29 min	60
53 min	57
55 min	56
205 min	54
174 min	53
192 min	51
201 min	48
45 min	48
209 min	40
229 min	40
195 min	36
169 min	34
50 min	34
190 min	34
36 min	33
11 Seasons	30
194 min	30
203 min	30
189 min	30
17 Seasons	30
204 min	29
214 min	27
21 min	26
35 min	25
38 min	25
193 min	24
228 min	24
178 min	24
13 min	23
14 min	23
212 min	21
253 min	21
15 min	20
167 min	20
233 min	18
237 min	18
49 min	16
37 min	16
43 min	16
312 min	15
12 min	14
31 min	13
191 min	13
230 min	12
41 min	11

```
19 min            8
273 min           7
34 min            6
17 min            5
39 min            5
10 min            4
16 min            4
196 min           4
20 min            4
18 min            4
3 min             4
5 min             3
11 min            2
8 min             2
9 min             2
Name: Duration, dtype: int64
```

Separating TV Shows and Movies into different categories

In [59]:
```
TV_Show = df_final[df_final['Type']=='TV Show']
TV_Show.Duration.value_counts()
```

Out[59]:
```
1 Season      35035
2 Seasons      9559
3 Seasons      5084
4 Seasons      2134
5 Seasons      1698
7 Seasons       843
6 Seasons       633
8 Seasons       286
9 Seasons       257
10 Seasons      220
13 Seasons      132
12 Seasons      111
15 Seasons       96
17 Seasons       30
11 Seasons       30
Name: Duration, dtype: int64
```

In [60]:
```
Movie = df_final[df_final['Type']=='Movie']
Movie.Duration.value_counts()
```

Out[60]:
```
94 min     4343
106 min    4040
97 min     3624
95 min     3560
96 min     3484
93 min     3480
90 min     3305
105 min    3209
107 min    3103
101 min    3048
102 min    3017
103 min    2985
98 min     2984
99 min     2956
91 min     2915
92 min     2863
104 min    2822
88 min     2781
110 min    2711
100 min    2697
108 min    2614
112 min    2594
85 min     2486
89 min     2420
86 min     2213
116 min    2122
118 min    2119
119 min    2075
87 min     2063
109 min    2020
113 min    1990
120 min    1845
117 min    1770
121 min    1728
111 min    1667
124 min    1590
114 min    1529
127 min    1505
115 min    1444
123 min    1398
125 min    1299
122 min    1298
84 min     1267
128 min    1241
130 min    1216
126 min    1205
```

81 min	1203
83 min	1192
133 min	1169
137 min	1122
82 min	1100
136 min	1092
132 min	1047
131 min	913
135 min	851
129 min	837
75 min	794
148 min	671
140 min	658
79 min	629
139 min	617
143 min	608
80 min	586
134 min	572
145 min	549
149 min	540
138 min	540
74 min	516
78 min	506
141 min	495
72 min	470
142 min	464
46 min	451
77 min	447
150 min	442
172 min	432
158 min	424
73 min	408
76 min	408
151 min	395
147 min	379
163 min	371
154 min	356
146 min	342
162 min	333
54 min	323
144 min	303
153 min	300
71 min	297
70 min	289
157 min	284
155 min	275
68 min	263
24 min	252
161 min	230
166 min	228
156 min	214
58 min	197
176 min	192
152 min	186
168 min	178
165 min	177
171 min	174
160 min	169
185 min	166
22 min	162
69 min	160
44 min	149
181 min	144
173 min	144
63 min	141
180 min	133
159 min	132
26 min	128
170 min	120
177 min	117
23 min	116
60 min	114
28 min	113
64 min	113
164 min	110
200 min	108
59 min	107
51 min	105
66 min	105
30 min	104
61 min	100
52 min	99
65 min	98
62 min	92
33 min	91
25 min	86
47 min	81
187 min	78

```
182 min      76
67 min       74
42 min       74
56 min       73
48 min       73
186 min      72
57 min       71
40 min       68
179 min      66
32 min       64
27 min       62
224 min      60
208 min      60
29 min       60
53 min       57
55 min       56
205 min      54
174 min      53
192 min      51
45 min       48
201 min      48
229 min      40
209 min      40
195 min      36
190 min      34
169 min      34
50 min       34
36 min       33
189 min      30
194 min      30
203 min      30
204 min      29
214 min      27
21 min       26
38 min       25
35 min       25
193 min      24
178 min      24
228 min      24
13 min       23
14 min       23
212 min      21
253 min      21
15 min       20
167 min      20
237 min      18
233 min      18
43 min       16
49 min       16
37 min       16
312 min      15
12 min       14
31 min       13
191 min      13
230 min      12
41 min       11
19 min        8
273 min       7
34 min        6
17 min        5
39 min        5
196 min       4
18 min        4
3 min         4
10 min        4
16 min        4
20 min        4
5 min         3
9 min         2
8 min         2
11 min        2
Name: Duration, dtype: int64
```

4. Visual Analysis - Univariate, Bivariate after pre-processing of the data

In [29]:
```
#Separating year and month from date added
df_final['Date_Added']=pd.to_datetime(df_final['Date_Added'])
```

In [30]:
```
df_final['Year_Added']=df_final['Date_Added'].dt.year
df_final['Month_Added']=df_final['Date_Added'].dt.month
```

In [31]:
```
df_final.head()
```

Out[31]:

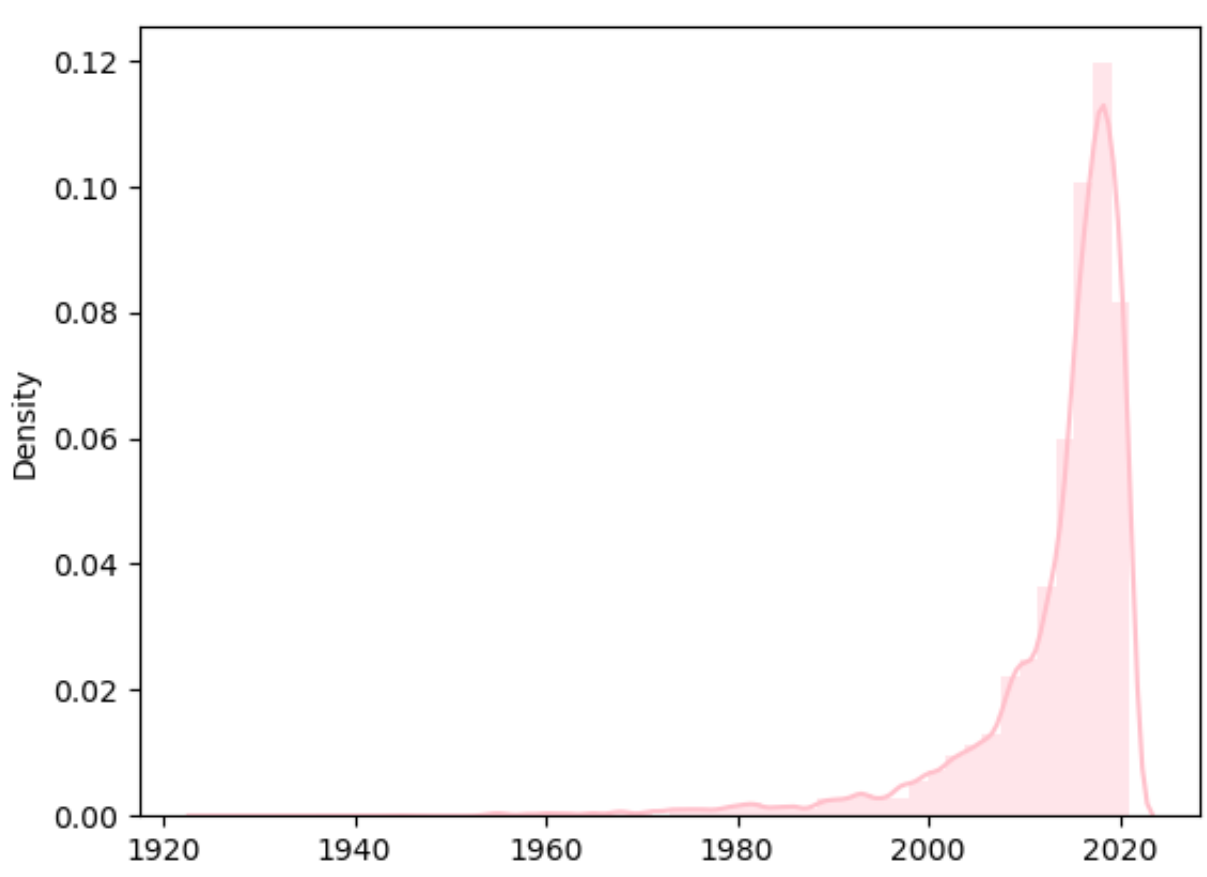
	Title	Directors	Actors	Countries	Genre	ID	Type	Date_Added	Release_Year	Rating	Duration	Year_Added	Month_
0	Dick Johnson Is Dead	Kirsten Johnson	Unknown Actor	United States	Documentaries	s1	Movie	2021-09-25	2020	PG-13	90 min	2021.0	
1	Blood & Water	Unknown Director	Ama Qamata	South Africa	International TV Shows	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	
2	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Dramas	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	
3	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Mysteries	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	
4	Blood & Water	Unknown Director	Khosi Ngema	South Africa	International TV Shows	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	

For Univariate-Continuous

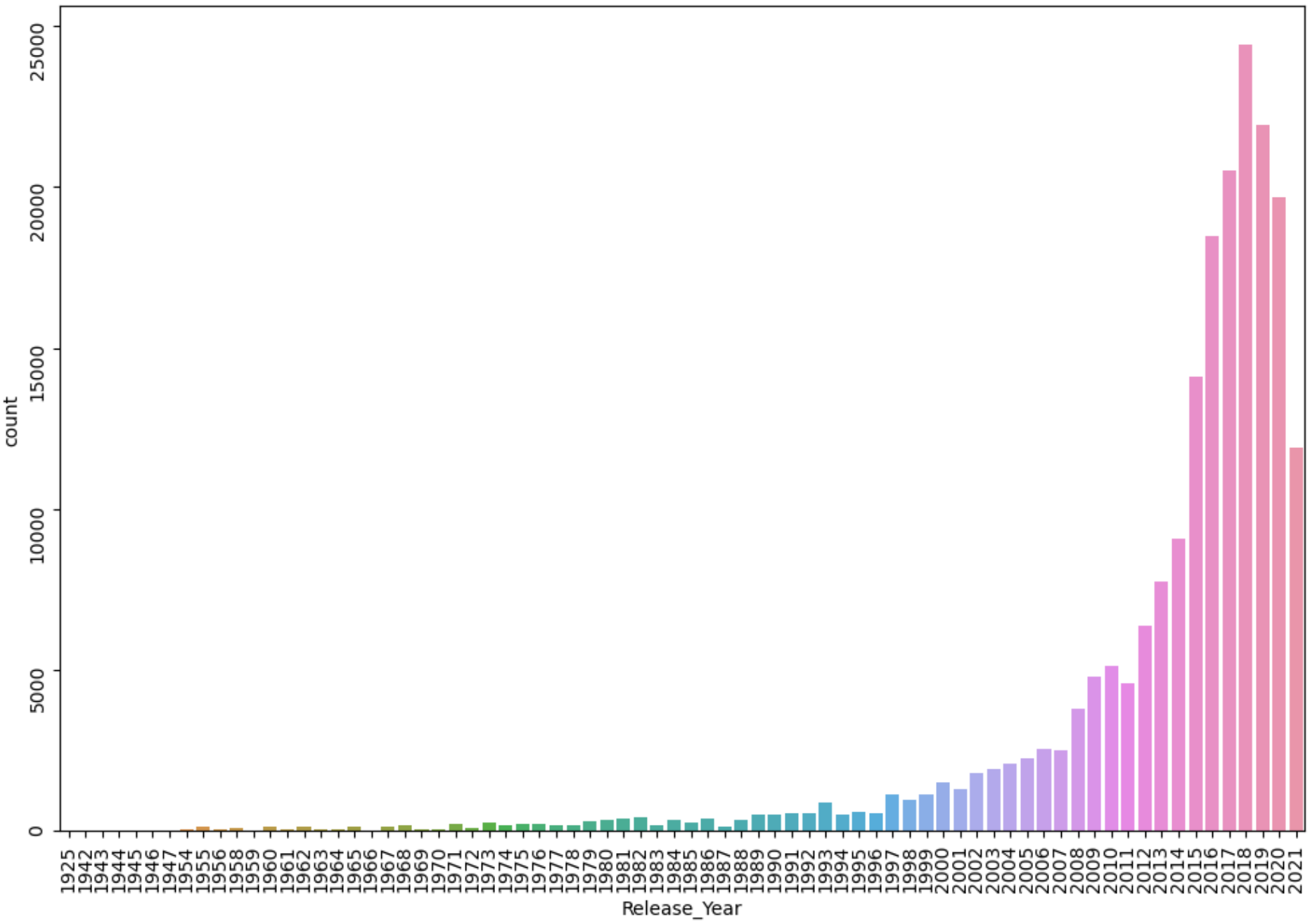
In [32]:
```
sns.distplot(x=df_final['Release_Year'],color='pink')
```

```
C:\Users\Home\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[32]:
```
<AxesSubplot:ylabel='Density'>
```

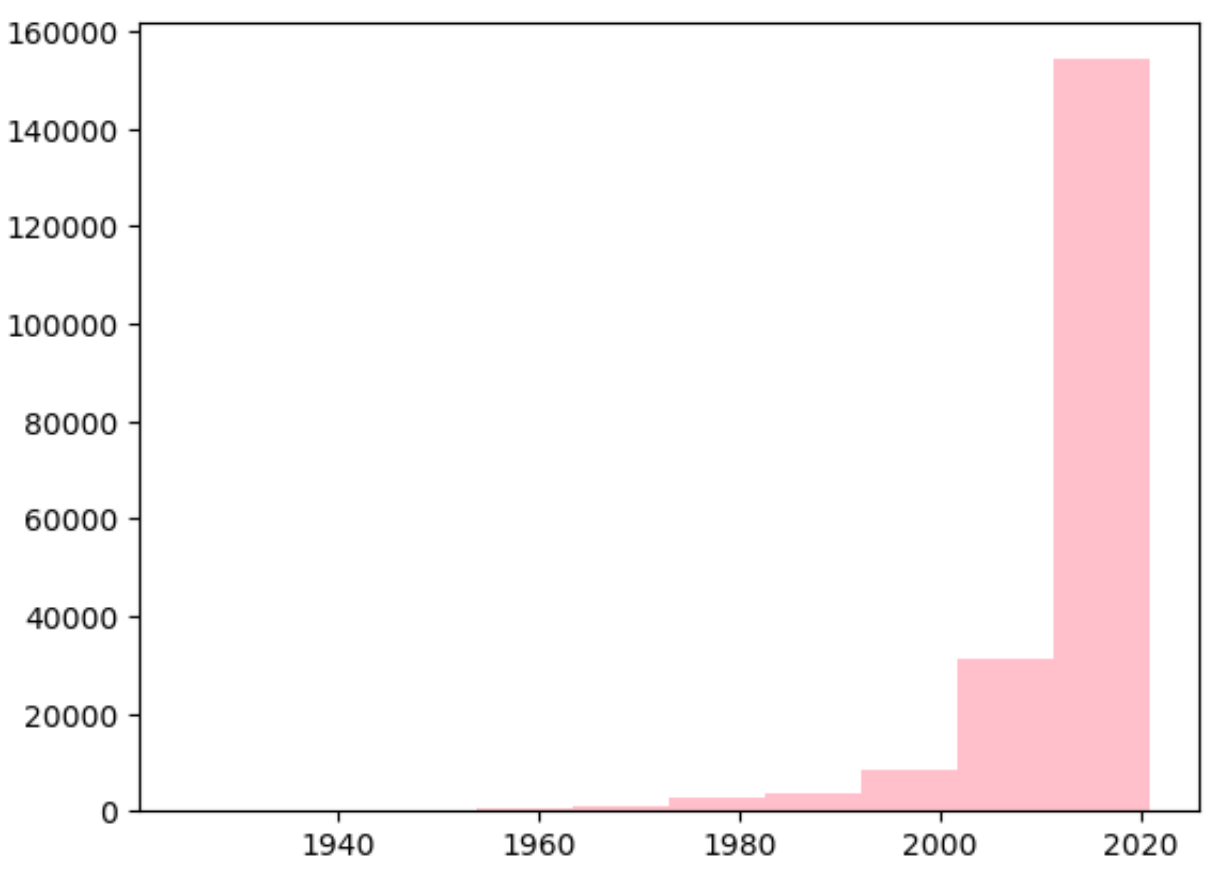


In [33]:
```
plt.figure(figsize=(12,8))
sns.countplot(x=df_final['Release_Year'])
plt.xticks(rotation=90)
plt.yticks(rotation=90)
plt.show()
```

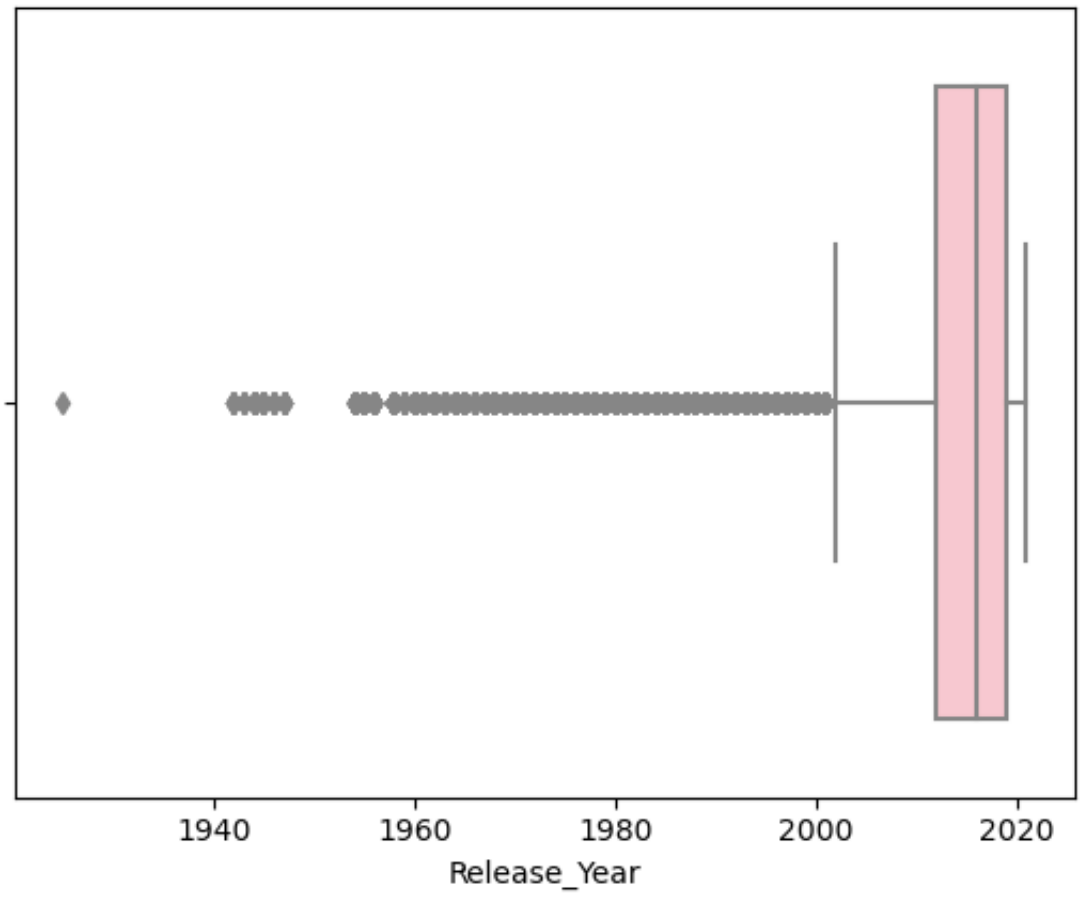


We can observe that the maximum number of TV Shows and movies were released in 2018. However, the number then started decreasing. One of the reasons for this could be spread of COVID-19, making it difficult for production of TV Shows and movies.

In [34]:
```
plt.hist(x=df_final['Release_Year'],color='pink')
plt.show()
```



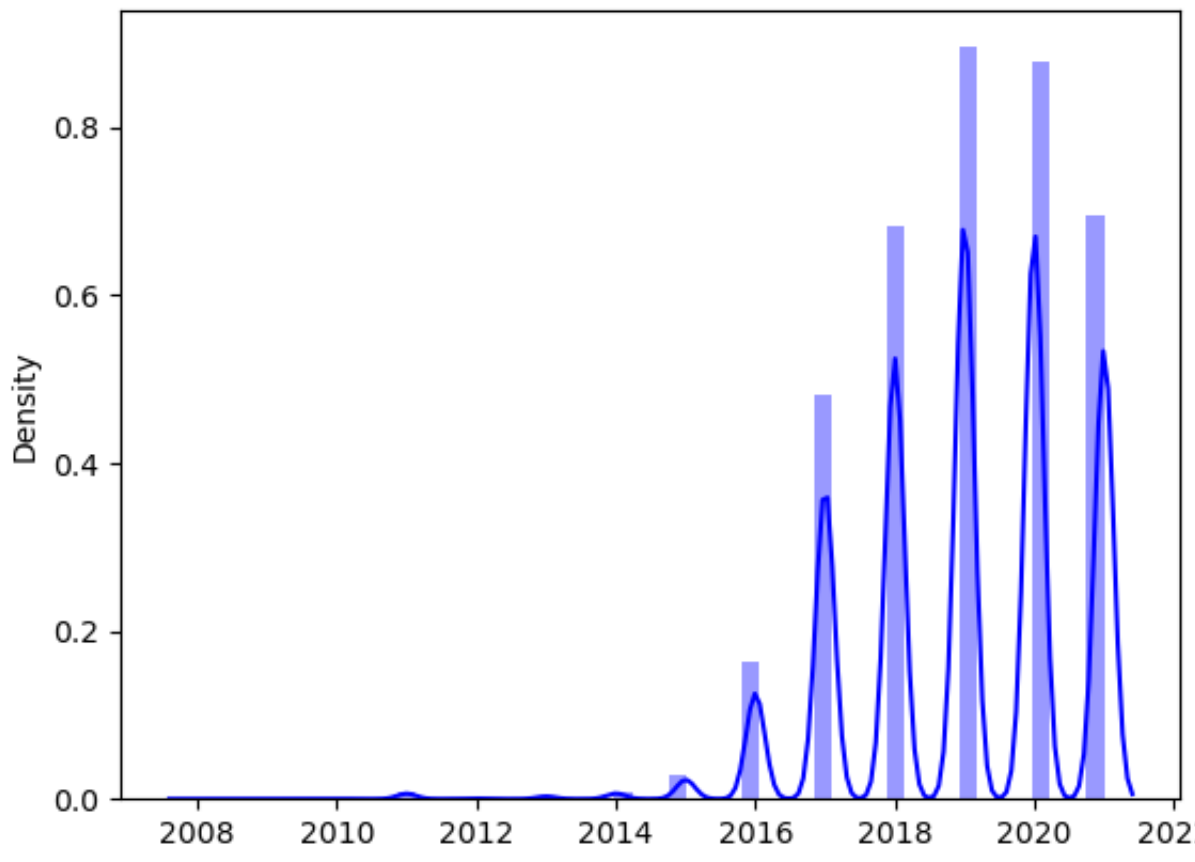
In [35]:
```
sns.boxplot(data=df_final,
            x="Release_Year", color='pink')
plt.show()
```



In [36]:
```
sns.distplot(x=df_final['Year_Added'],color='blue')
```

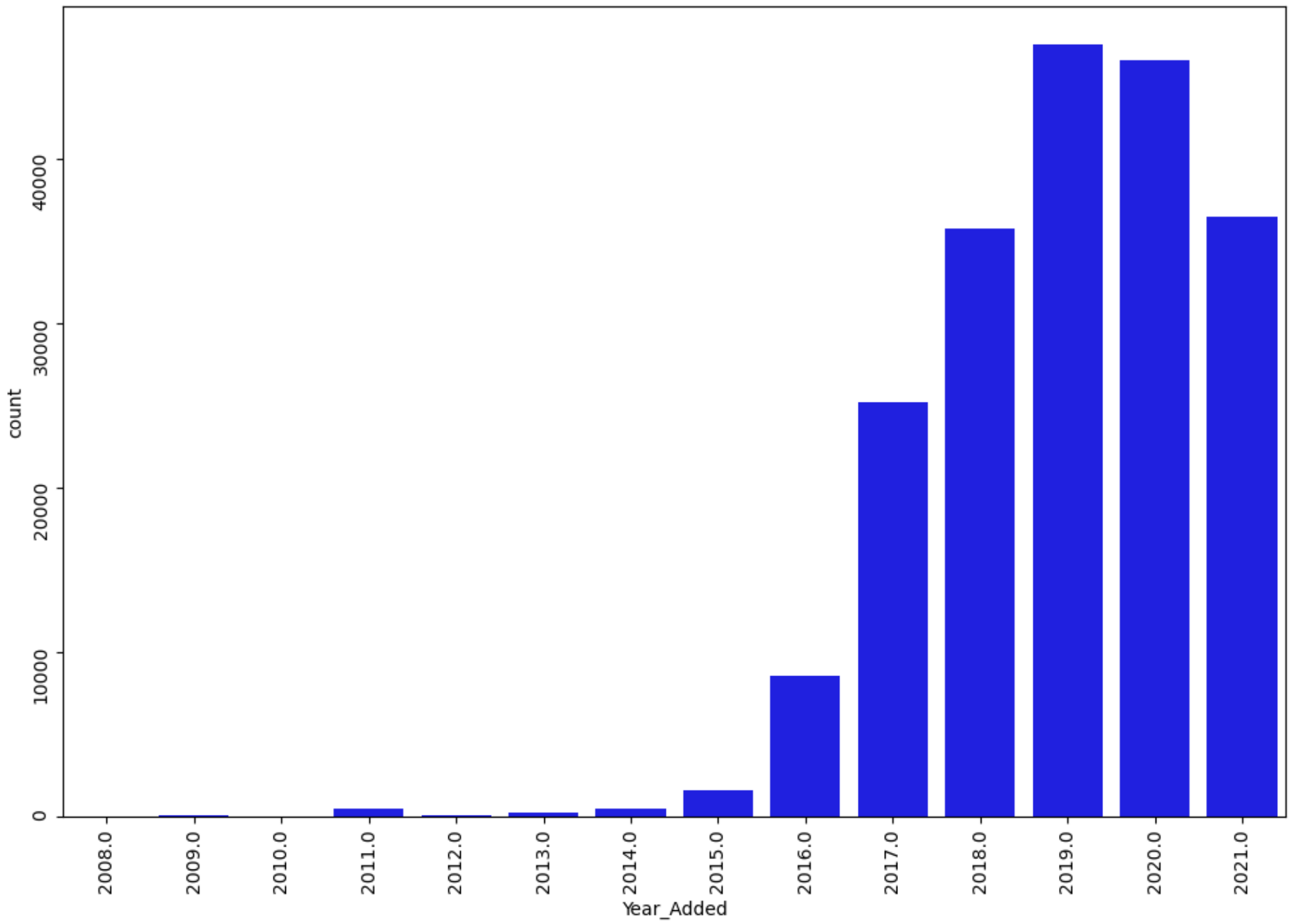
```
C:\Users\Home\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[36]:
```
<AxesSubplot:ylabel='Density'>
```

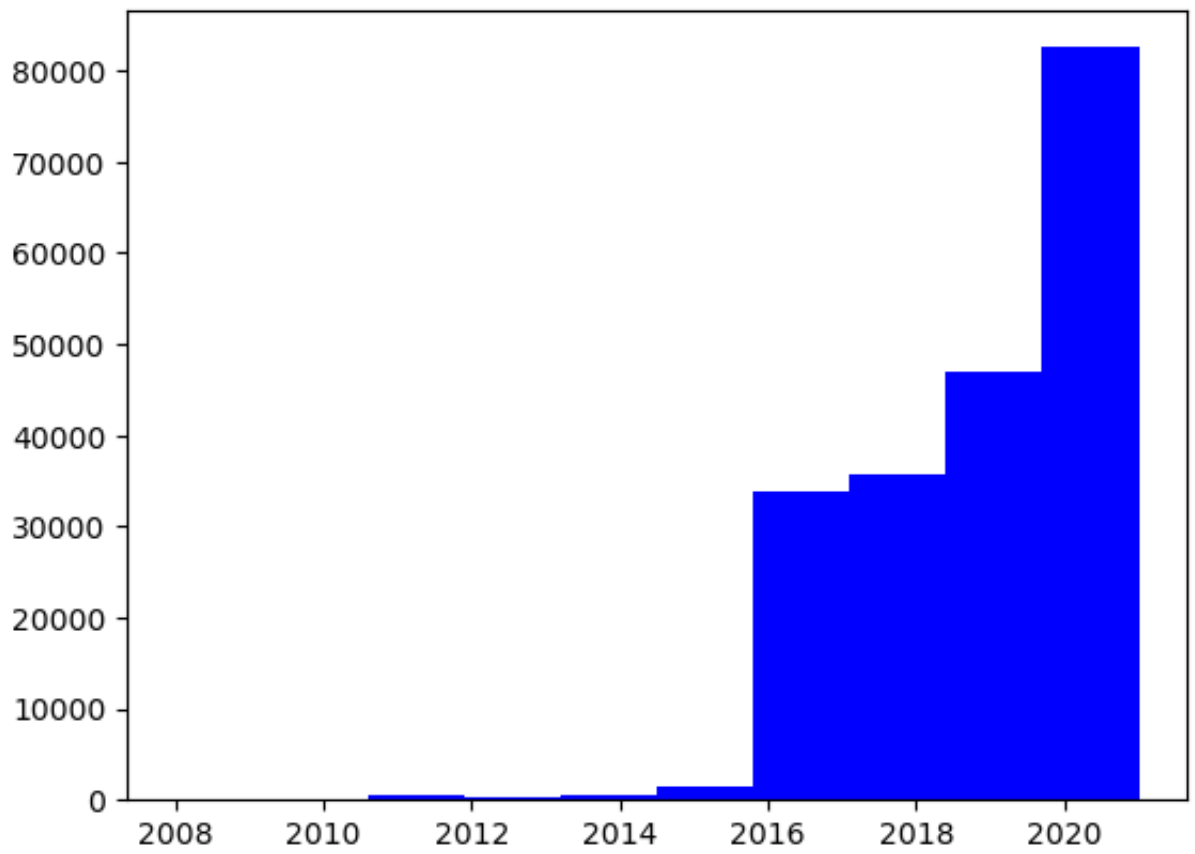


We can observe from above that the maximum number of TV Shows and movies were added in 2019 and then in 2020. This number saw a dip in 2021. The reason for this could be control of COVID-19 and people moving back to their jobs and schools, etc. During the lockdown in 2020, people had a lot of time at hand and Netflix saw this as an opportunity to increase its viewership. Thus, the number of addition of TV Shows and movies on the platform went up during this period.

In [37]:
```
plt.figure(figsize=(12,8))
sns.countplot(x=df_final['Year_Added'],color='blue')
plt.xticks(rotation=90)
plt.yticks(rotation=90)
plt.show()
```



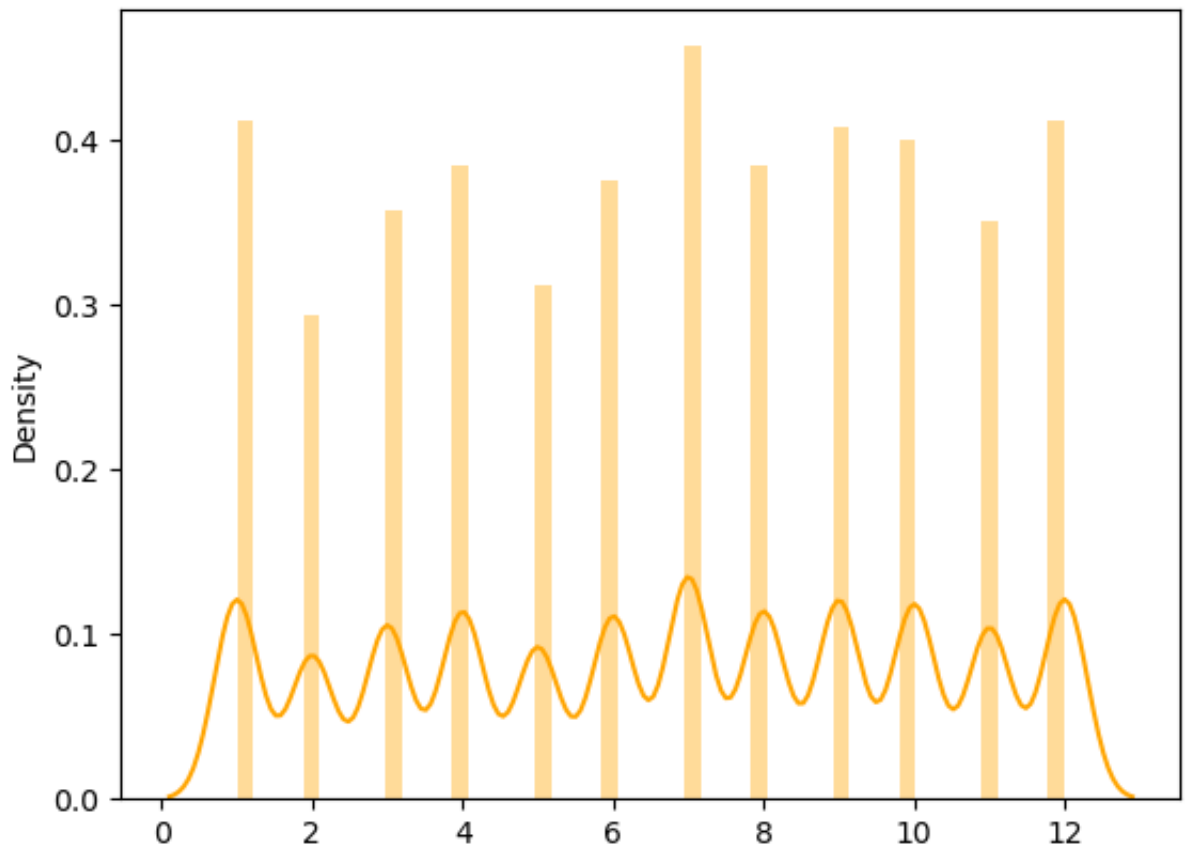
In [38]:
```
plt.hist(x=df_final['Year_Added'],color='blue')
plt.show()
```



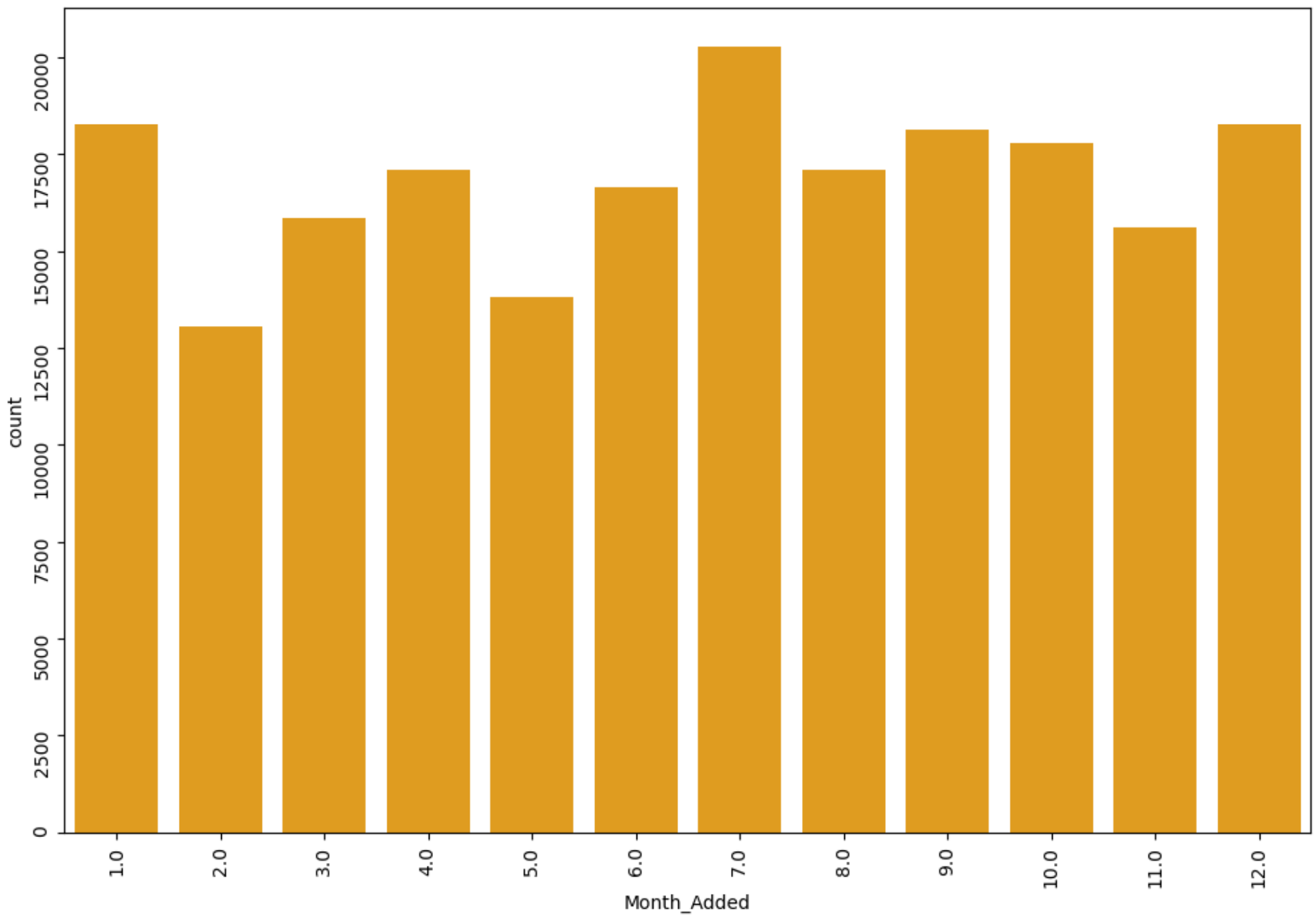
In [39]:
```
sns.distplot(x=df_final['Month_Added'],color='orange')
```

```
C:\Users\Home\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[39]:
```
<AxesSubplot:ylabel='Density'>
```



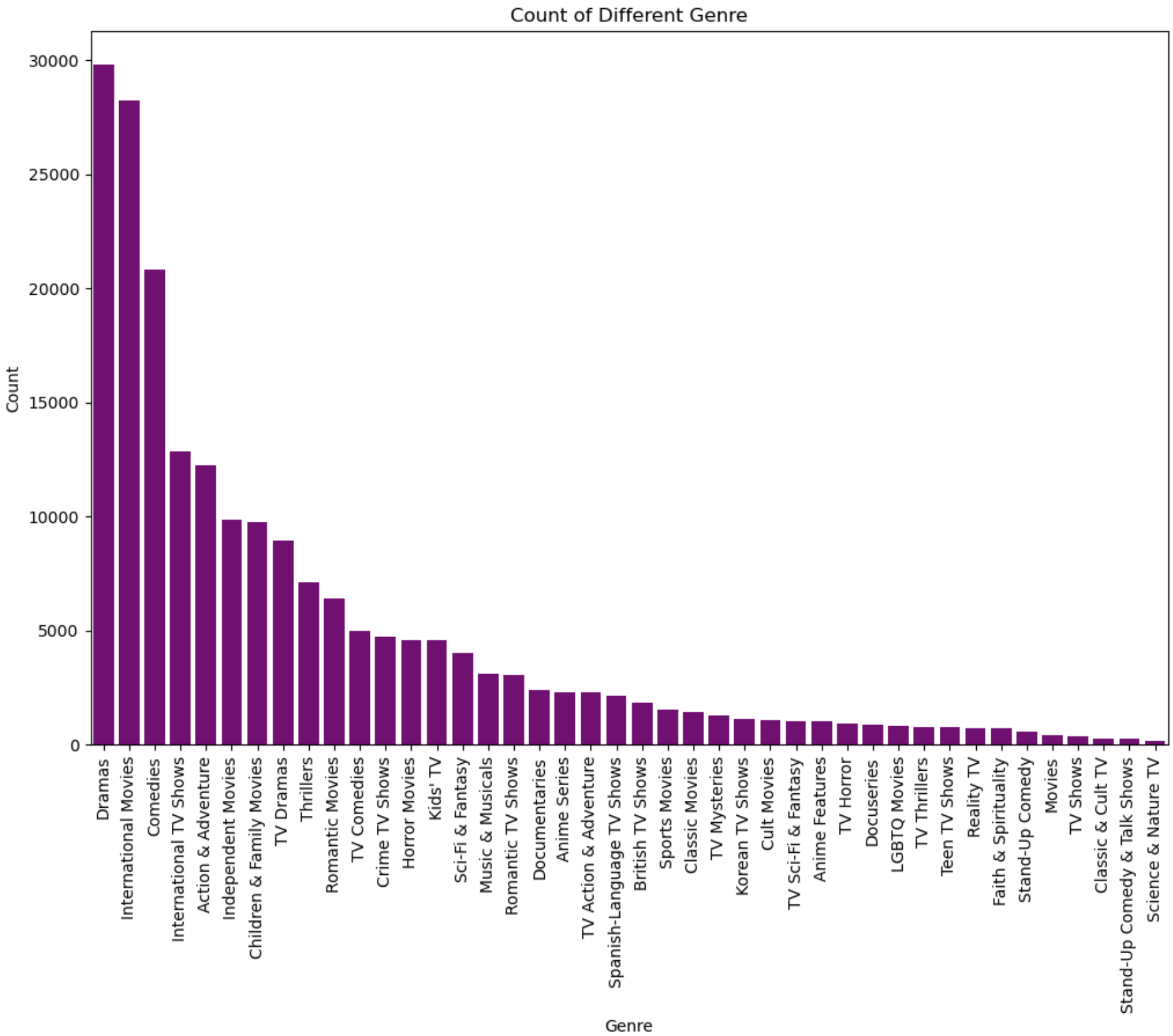
In [40]:
```
plt.figure(figsize=(12,8))
sns.countplot(x=df_final['Month_Added'],color='orange')
plt.xticks(rotation=90)
plt.yticks(rotation=90)
plt.show()
```



From the above, we can notice that the maximum number of TV shows and movies were added in the 7th month i.e. July, followed by the number of movies added in December and January. The reason for this could be holiday season as people have free time on their hand and would like to watch their favorite movies and shows.

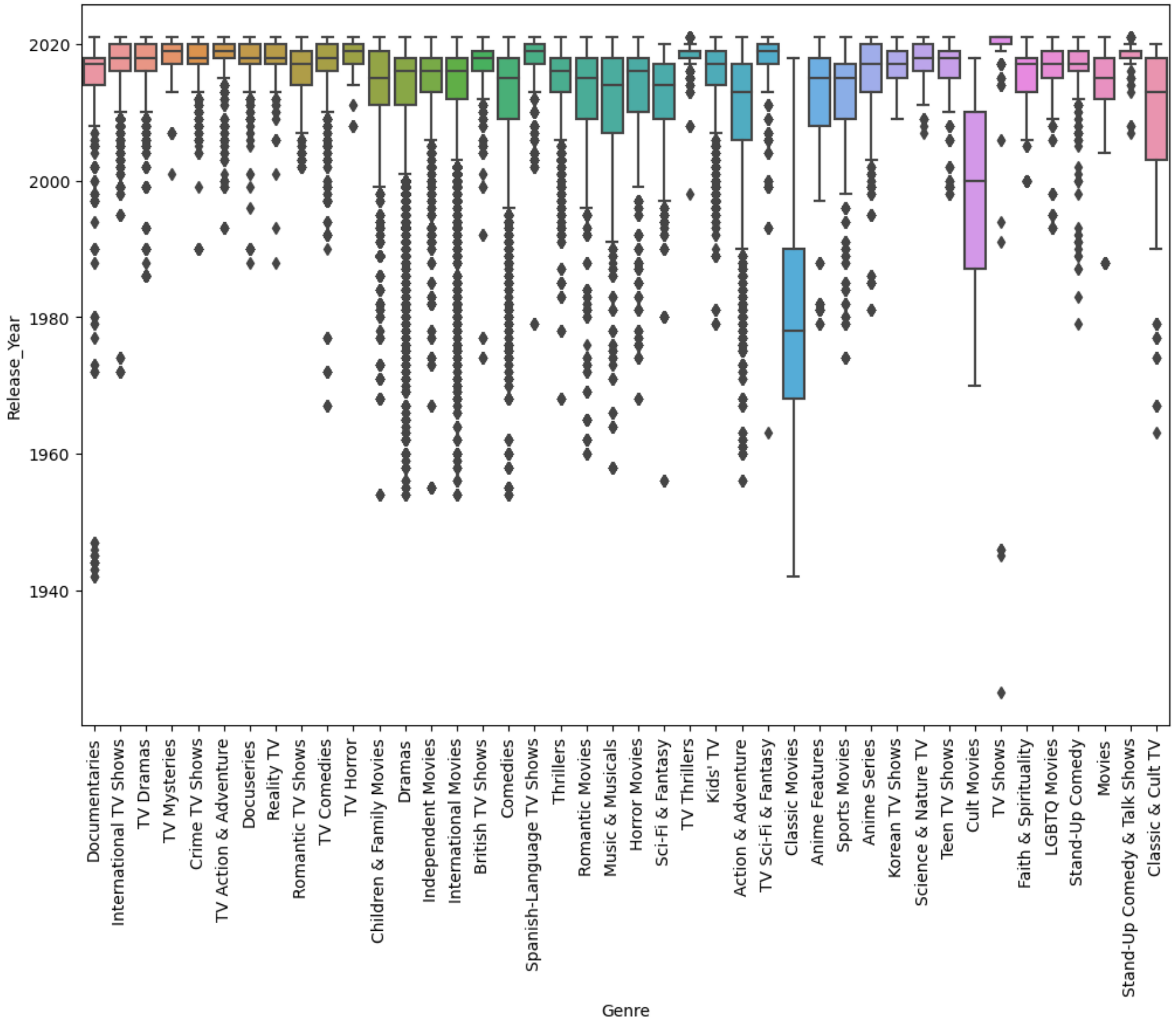
For Univariate-Categorical

In [43]:
```
plt.figure(figsize=(12,8))
sns.countplot(data=df_final,
 x="Genre",
 order=df_final["Genre"].value_counts().index,
 color="purple")
plt.xticks(rotation=90)
plt.xlabel("Genre")
plt.ylabel("Count")
plt.title("Count of Different Genre")
plt.show()
```

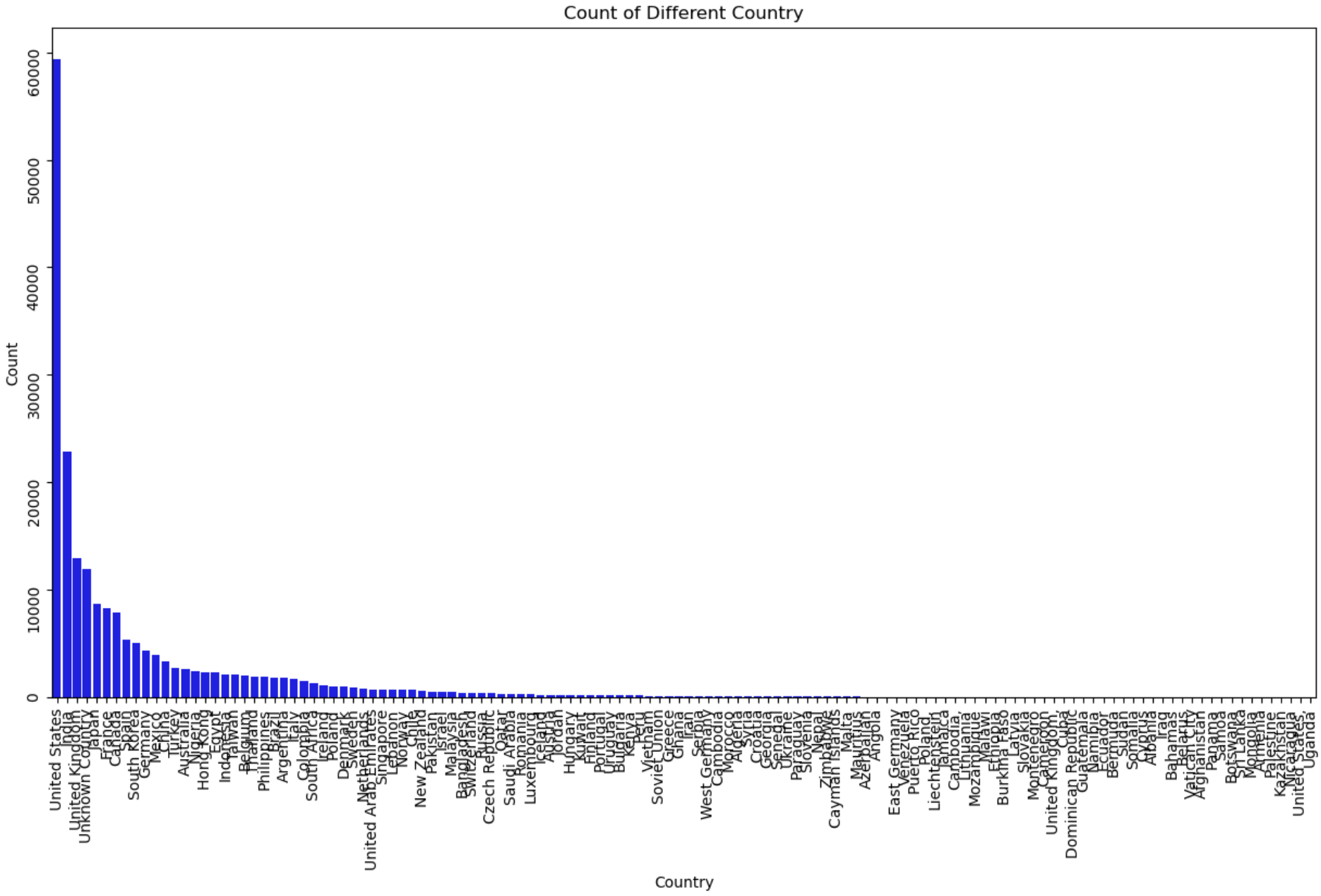


From the above, we can observe that 'Dramas' are the most popular genre among the viewers. The second most popular genre among the viewers is 'International Movies'.

In [44]:
```
plt.figure(figsize=(12,8))
sns.boxplot(data=df_final,
 x="Genre",
 y="Release_Year")
plt.xticks(rotation=90)
plt.show()
```

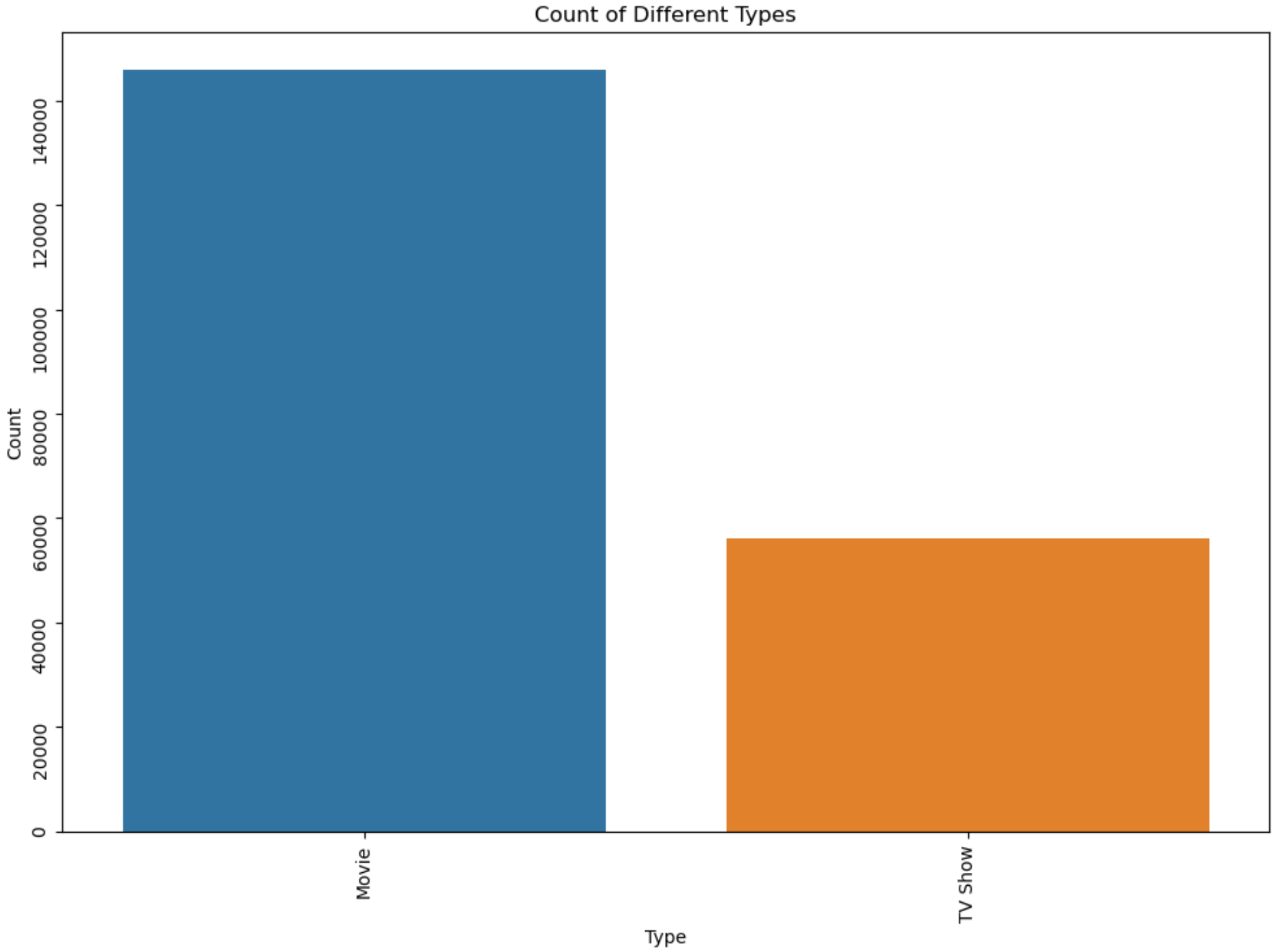


In [45]:
```
plt.figure(figsize=(15,8))
sns.countplot(data=df_final,
 x="Countries",
 order=df_final["Countries"].value_counts().index,
 color="blue")
plt.xticks(rotation=90)
plt.yticks(rotation=90)
plt.xlabel("Country")
plt.ylabel("Count")
plt.title("Count of Different Country")
plt.show()
```



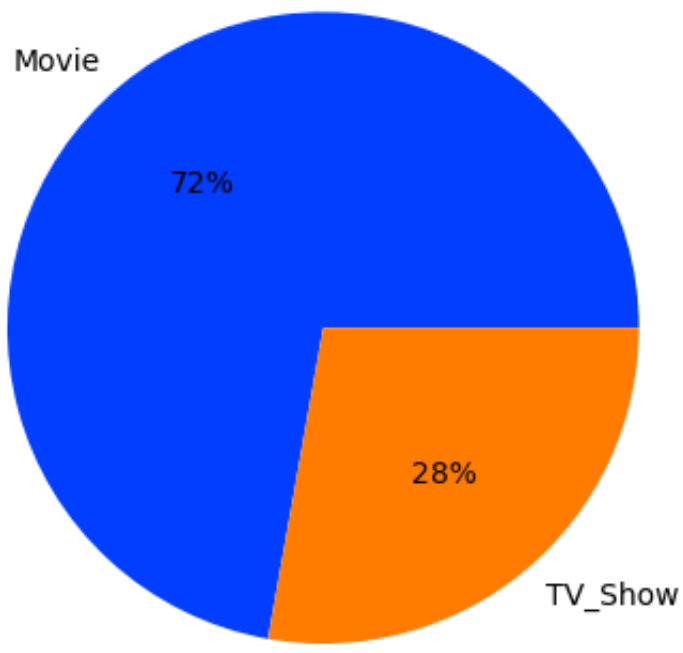
From the above graph, we can observe that the highest number of TV shows and movies are available in United States, followed by India.

In [46]:
```
plt.figure(figsize=(12,8))
sns.countplot(data=df_final,
 x="Type",
 order=df_final["Type"].value_counts().index)
plt.xticks(rotation=90)
plt.yticks(rotation=90)
plt.xlabel("Type")
plt.ylabel("Count")
plt.title("Count of Different Types")
plt.show()
```



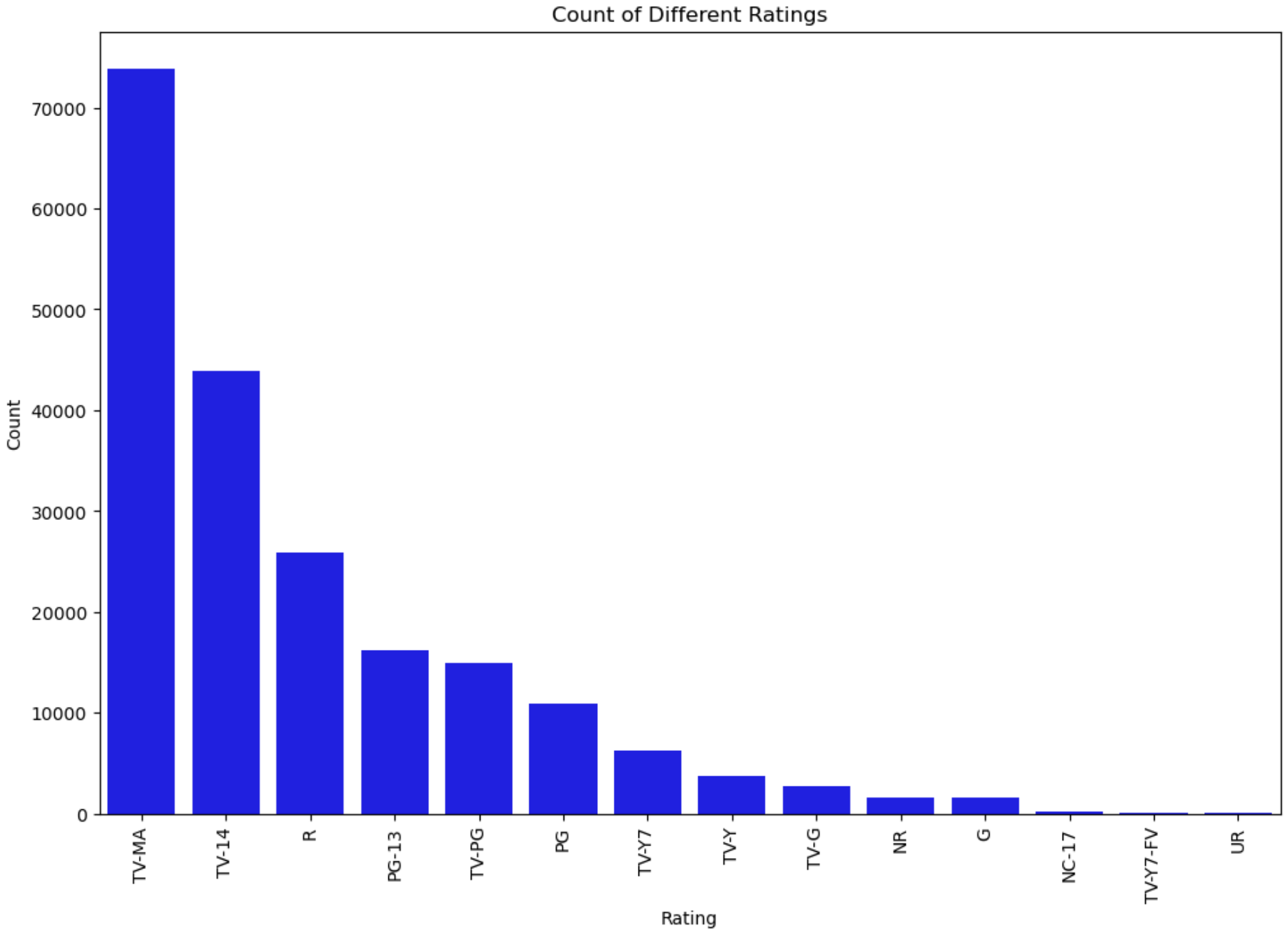
In [47]:
```
palette_color = sns.color_palette('bright')
keys=['Movie','TV_Show']
plt.pie(df_final.Type.value_counts(), labels=keys, colors=palette_color, autopct='%.0f%%')

# displaying chart
plt.show()
```



From the above, we can observe that out of the content available on Netflix, 72% are movies and 28% are TV shows.

In [48]:
```
plt.figure(figsize=(12,8))
sns.countplot(data=df_final,
 x="Rating",
 order=df_final["Rating"].value_counts().index,
 color="blue")
plt.xticks(rotation=90)
plt.xlabel("Rating")
plt.ylabel("Count")
plt.title("Count of Different Ratings")
plt.show()
```



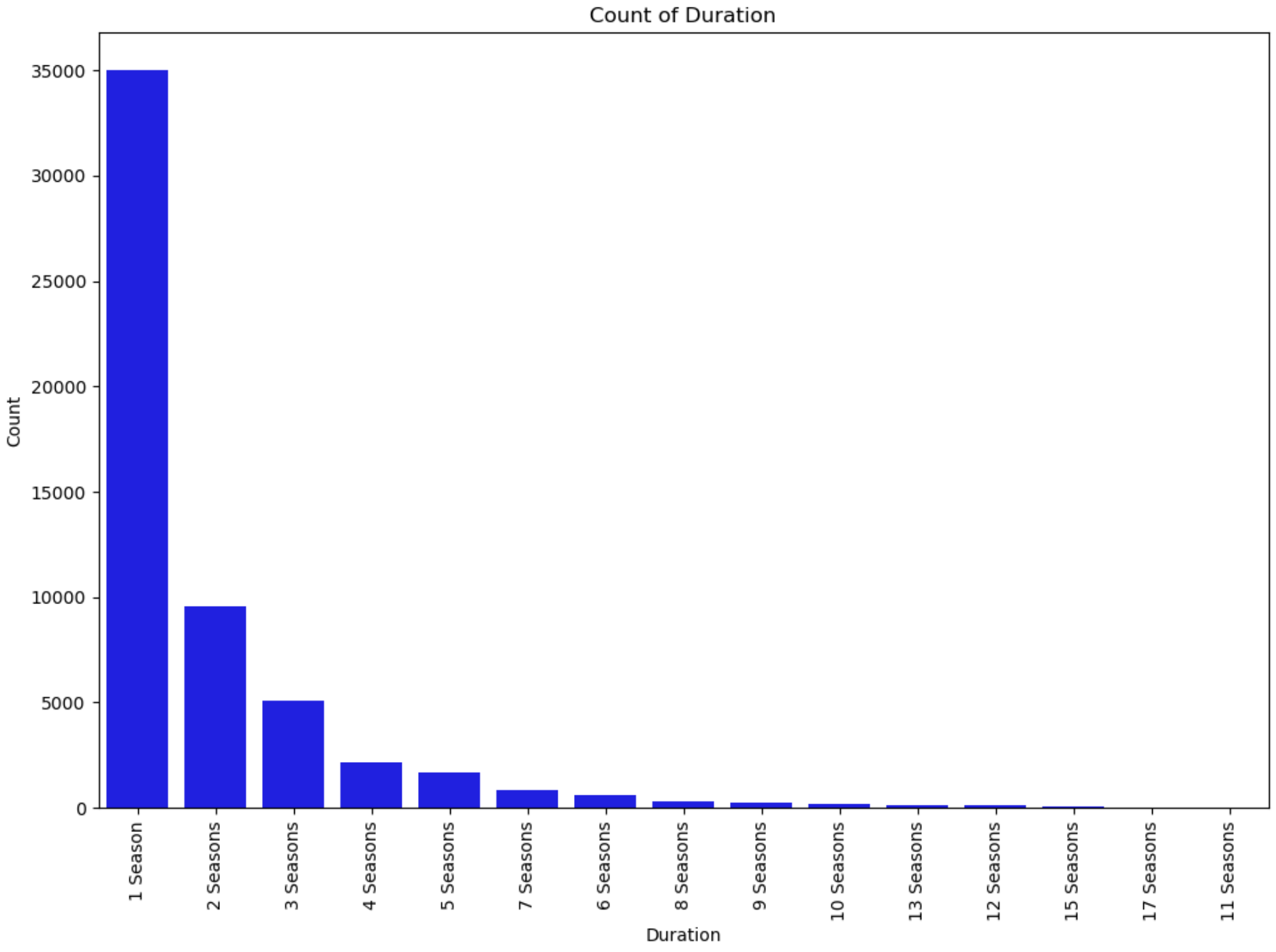
The most number of movies and TV shows on Netflix have TV-MA rating, followed by TV-14. This means that it aims to target mature audience as well as children, keeping its audience base wide.

In [49]:
```
TV_Show = df_final[df_final['Type']=='TV Show']
TV_Show.head()
```

Out[49]:

	Title	Directors	Actors	Countries	Genre	ID	Type	Date_Added	Release_Year	Rating	Duration	Year_Added	Month_Added
1	Blood & Water	Unknown Director	Ama Qamata	South Africa	International TV Shows	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	9.0
2	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Dramas	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	9.0
3	Blood & Water	Unknown Director	Ama Qamata	South Africa	TV Mysteries	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	9.0
4	Blood & Water	Unknown Director	Khosi Ngema	South Africa	International TV Shows	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	9.0
5	Blood & Water	Unknown Director	Khosi Ngema	South Africa	TV Dramas	s2	TV Show	2021-09-24	2021	TV-MA	2 Seasons	2021.0	9.0

In [50]:
```
plt.figure(figsize=(12,8))
sns.countplot(data=TV_Show,
 x='Duration',
 order=TV_Show["Duration"].value_counts().index,
 color="blue")
plt.xticks(rotation=90)
plt.xlabel("Duration")
plt.ylabel("Count")
plt.title("Count of Duration")
plt.show()
```

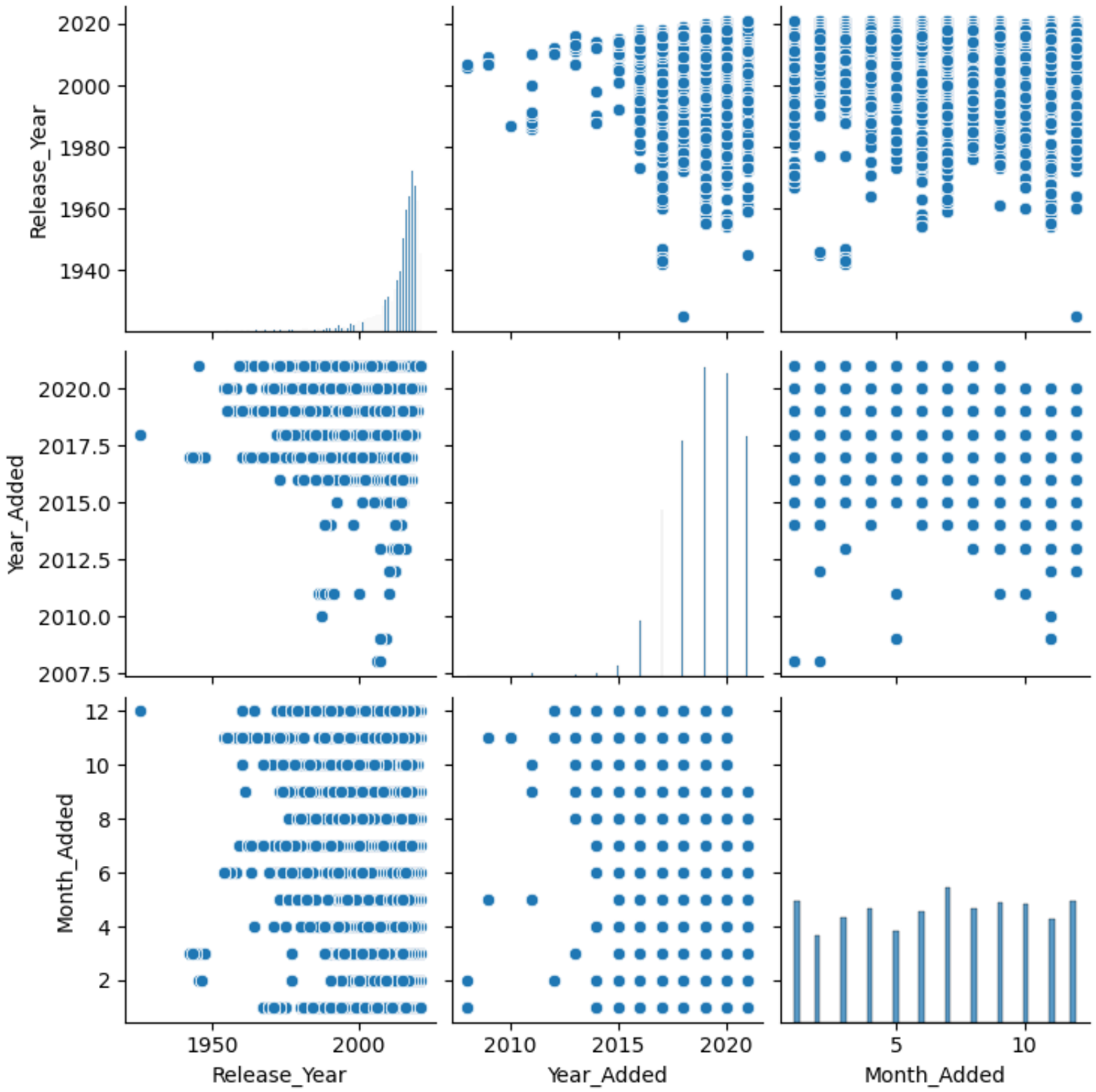


From the above, we can observe that the maximum TV shows on Netflix only have 1 season. The reason for this could be lack of interest of the audience in that particular TV show or the TV shows on Netflix could be recent and their season 2 might be under production.

For correlation: Heatmaps, Pairplots

In [52]:
```
sns.pairplot(df_final)
```

Out[52]:
```
<seaborn.axisgrid.PairGrid at 0x16784576160>
```



In [54]:
```
sns.heatmap(df_final.corr())
```

Out[54]: <AxesSubplot:>

5. Missing Value & Outlier check

In [61]:
```
df_final.isna().sum()
```

Out[61]:
```
Title             0
Directors         0
Actors            0
Countries         0
Genre             0
ID                0
Type              0
Date_Added      158
Release_Year      0
Rating            0
Duration          3
Year_Added      158
Month_Added     158
dtype: int64
```

6. Insights based on Non-Graphical and Visual Analysis

After Non-Graphical and Visual Analysis, we observe the following:

a. We observe that the maximum number of TV Shows and movies were released in 2018. However, the number then started decreasing. One of the reasons for this could be spread of COVID-19, making it difficult for production of TV Shows and movies.

b. We observe from above that the maximum number of TV Shows and movies were added in 2019 and then in 2020. This number saw a dip in 2021. The reason for this could be control of COVID-19 and people moving back to their jobs and schools, etc. During the lockdown in 2020, people had a lot of time at hand and Netflix saw this as an opportunity to increase its viewership. Thus, the number of addition of TV Shows and movies on the platform went up during this period.

c. We notice that the maximum number of TV shows and movies were added in the 7th month i.e. July, followed by the number of movies added in December and January. The reason for this could be holiday season as people have free time on their hand and would like to watch their favorite movies and shows.

d. We observe that 'Dramas' are the most popular genre among the viewers. The second most popular genre among the viewers is 'International Movies'.

e. We observe that the highest number of TV shows and movies are available in United States, followed by India.

f. We observe that out of the content available on Netflix, 72% are movies and 28% are TV shows.

g. We obseve that the most number of movies and TV shows on Netflix have TV-MA rating, followed by TV-14. This means that it aims to target mature audience as well as children, keeping its audience base wide.

h. We observe that the maximum TV shows on Netflix only have 1 season. The reason for this could be lack of interest of the audience in that particular TV show or the TV shows on Netflix could be recent and their season 2 might be under production.

7 & 8. Business Insights and Recommendations

We observe that the audience had more free time during lockdown. However, now that the COVID situation is under control and people have less time to spend on OTT platforms, it is suggested that Netflix add those movies and TV shows that are of shorter duration. Further, since dramas and international movies are more popular among the audience, netflix should more of such content on its platform. Adding movies and shows with TV-MA and TV-14 rating is a good way to keep a wide audience base and this will also help in more revenue generation. It is also observed that TV shows with lesser number of seasons are easily available on Netflix. It is advisable that Netflix adds all the seasons of a particular show to retain its customers.

In [ ]: