In [65]:
```
# Business Problem Statement: Analyze the data and generate insights that co
```

In [66]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from datetime import datetime
```

In [67]:
```
df = pd.read_csv(r"C:\Users\DELL\Desktop\Data Science Scaler\netflix.csv")
df.head(50)
```

Out[67]:

	show_id	type	title	director	cast	country	date_added	release_
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	

Q1.Defining Problem Statement and Analysing basic metrics

1.PROBLEM STATEMENT: a. Analyze the netflix dataset to provide data-driven recommendation on the type of content (movies or TV shows)to produce.

```
b. Explore how Netflix can expand and grow its business in differen
t countries.
```

2. BASIC METRICS ANALYSIS:
   a. Import the dataset and load it into a suitable data structure for analysis.

   b. Check the data for any missing values,duplicates and handle them .

   c. Analyse the overall distribution of content types(movies,tv.shows) to see it netflix has a preference.

   d. Calculate the total number of movies and tv shows available on Netflix.

   e. Analyze the tv ratings of the content to see if there is a particular rating that performs better.

   f. Determine the average duration (in minutes) of movies and the average number of seasons for TV shows.
3. Content by country: a. Analyze which countries produce the most content for Netflix.

b. Analyze which types of contents are most popular in specific countries.

4. Launch time or date for tv shows and movies: a. Examine the release dates and time of tv shows to determine if there is a season or time of the year that tends to perform better.
5. Actor and Director Analysis: a. Identify most polularly appearing actors and directors in Netflix content.

   b. Determine the specific actors and directors are associated with higher ratings.
6. Focus on TV Shows or Movies: a. Analyze which types of contents (TV Shows or Movies) is more producing in recent years.
7. Growth Strategies: a. Provide recommendation for Netflix on expanding its business in different countries based on content prefrences ,regional trends and potential market opportunities.

In [68]:
```
# fill null values with a specific values:

df['director'].fillna("Unknown director",inplace=True)
df['cast'].fillna("Unknown cast",inplace=True)
df['country'].fillna("Unknown country",inplace=True)
df['date_added'].fillna("January 1,1900",inplace=True)
df['duration'].fillna("Unknown duration",inplace=True)
df['rating'].fillna("Unknown rating",inplace=True)
df
```

Out[68]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	September 24, 2021	2021
3	s4	TV Show	Jailbirds New Orleans	Unknown director	Unknown cast	Unknown country	September 24, 2021	2021
4	s5	TV Show	Kota Factory	Unknown director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021
...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007
8803	s8804	TV Show	Zombie Dumb	Unknown director	Unknown cast	Unknown country	July 1, 2019	2018
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006

	show_id	type	title	director	cast	country	date_added	release_year
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015

8807 rows × 12 columns

Q2.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 [69]:

```
# shape of data
shape=df.shape
# display the shape
("shape of the dataframe(rows,columns):",shape)
```

Out[69]:

```
('shape of the dataframe(rows,columns):', (8807, 12))
```

In [70]:

```
# Data type of all attributes
print('Data Type of Attributes')
print(df.dtypes)
```

```
Data Type of Attributes
show_id         object
type            object
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
dtype: object
```

In [71]:

```
#Missing Value Detection
missing_values=df.isnull().sum()
print('Missing_values:')
print(missing_values)
```

```
Missing_values:
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
dtype: int64
```

In [72]:

```
df["type"]=df["type"].astype("category")
df["country"]=df["country"].astype("category")
df["rating"]=df["rating"].astype("category")
("Data type :",df.dtypes)
```

Out[72]:

```
('Data type :',
 show_id           object
 type            category
 title             object
 director          object
 cast              object
 country         category
 date_added        object
 release_year       int64
 rating          category
 duration          object
 listed_in         object
 description       object
 dtype: object)
```

In [73]:

```
#Statsistical Summary
summary=df.describe(include='all')
print('Statistical Summary:')
print(summary)
df.describe()
```

```
Statistical Summary:
       show_id   type                  title          director          cas
t  \
count     8807   8807                   8807              8807          880
7
unique    8807      2                   8807              4529          769
3
top         s1  Movie  Dick Johnson Is Dead  Unknown director  Unknown cas
t
freq         1   6131                      1              2634           82
5
mean       NaN    NaN                    NaN               NaN           Na
N
std        NaN    NaN                    NaN               NaN           Na
N
min        NaN    NaN                    NaN               NaN           Na
N
25%        NaN    NaN                    NaN               NaN           Na
N
50%        NaN    NaN                    NaN               NaN           Na
N
75%        NaN    NaN                    NaN               NaN           Na
N
max        NaN    NaN                    NaN               NaN           Na
N

              country       date_added  release_year rating  duration  \
count            8807             8807   8807.000000   8807      8807
unique            749             1768           NaN     18       221
top     United States  January 1, 2020           NaN  TV-MA  1 Season
freq             2818              109           NaN   3207      1793
mean              NaN              NaN   2014.180198    NaN       NaN
std               NaN              NaN      8.819312    NaN       NaN
min               NaN              NaN   1925.000000    NaN       NaN
25%               NaN              NaN   2013.000000    NaN       NaN
50%               NaN              NaN   2017.000000    NaN       NaN
75%               NaN              NaN   2019.000000    NaN       NaN
max               NaN              NaN   2021.000000    NaN       NaN

                           listed_in  \
count                           8807
unique                           514
top     Dramas, International Movies
freq                             362
mean                             NaN
std                              NaN
min                              NaN
25%                              NaN
50%                              NaN
75%                              NaN
max                              NaN

                                              description
count                                                8807
unique                                               8775
top     Paranormal activity at a lush, abandoned prope...
freq                                                    4
mean                                                  NaN
std                                                   NaN
min                                                   NaN
25%                                                   NaN
```

```
50%                                                      NaN
75%                                                      NaN
max                                                      NaN
```

Out[73]:

	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

Q3.Non-Graphical Analysis: Value counts and unique attributes

In [74]:

```
# Get value counts for a specific column,e.g.,'Rating','type'
rating_counts=df['rating'].value_counts()
type_counts=df['type'].value_counts()
```

In [75]:

```
# Get unique in a specific column,e.g.,"country"
country=df['country'].unique()
```

In [76]:
```
# # Display the value counts and unique attributes
print('Value counts for Rating:')
print(rating_counts)


print('Value counts for Type:')
print(type_counts)


print("\nunique Country:")
print(country)
```

```
Value counts for Rating:
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
Unknown rating       4
NC-17                3
UR                   3
74 min               1
84 min               1
66 min               1
Name: rating, dtype: int64
Value counts for Type:
Movie      6131
TV Show    2676
Name: type, dtype: int64

unique Country:
['United States', 'South Africa', 'Unknown country', 'India', 'United Stat
es, Ghana, Burkina Faso, United Ki..., ..., 'Russia, Spain', 'Croatia, Slo
venia, Serbia, Montenegro', 'Japan, Canada', 'United States, France, South
Korea, Indonesia', 'United Arab Emirates, Jordan']
Length: 749
Categories (749, object): [', France, Algeria', ', South Korea', 'Argentin
a', 'Argentina, Brazil, France, Poland, Germany, D..., ..., 'Venezuela, Co
lombia', 'Vietnam', 'West Germany', 'Zimbabwe']
```

In [77]:
```
df.country.value_counts().head()
```

Out[77]:
```
United States      2818
India               972
Unknown country     831
United Kingdom      419
Japan               245
Name: country, dtype: int64
```

```
In [78]: count_genre=df['listed_in'].value_counts()
         count_genre
```

```
Out[78]: 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 [79]: year_counts= df['release_year'].value_counts().sort_index()
         year_counts
```

```
Out[79]: 1925       1
         1942       2
         1943       3
         1944       3
         1945       4
                 ...
         2017    1032
         2018    1147
         2019    1030
         2020     953
         2021     592
         Name: release_year, Length: 74, dtype: int64
```

Q4.Visual Analysis - Univariate, Bivariate after pre-processing of the data

In [80]:

```
# unnesting the columns
df['cast_split'] = df['cast'].str.split(', ')
df = df.explode('cast_split')
df['director_split'] = df['director'].str.split(', ')
df = df.explode('director_split')
df['country_split'] = df['country'].str.split(', ')
df = df.explode('country_split')
df['listed_in_split'] = df['listed_in'].str.split(', ')
df = df.explode('listed_in_split')
df.head()
```

Out[80]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	September 25, 2021	2020	PG-13
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA

In [81]:

```
small_df=df[['show_id','title','release_year','type']]
small_df.head(5)
```

Out[81]:

	show_id	title	release_year	type
0	s1	Dick Johnson Is Dead	2020	Movie
1	s2	Blood & Water	2021	TV Show
1	s2	Blood & Water	2021	TV Show
1	s2	Blood & Water	2021	TV Show
1	s2	Blood & Water	2021	TV Show

In [82]:

```
small_df.drop_duplicates(inplace=True)
small_df.head(20)
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_14304\2201532.py:1: SettingWith
CopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  small_df.drop_duplicates(inplace=True)
```

Out[82]:

	show_id	title	release_year	type
0	s1	Dick Johnson Is Dead	2020	Movie
1	s2	Blood & Water	2021	TV Show
2	s3	Ganglands	2021	TV Show
3	s4	Jailbirds New Orleans	2021	TV Show
4	s5	Kota Factory	2021	TV Show
5	s6	Midnight Mass	2021	TV Show
6	s7	My Little Pony: A New Generation	2021	Movie
7	s8	Sankofa	1993	Movie
8	s9	The Great British Baking Show	2021	TV Show
9	s10	The Starling	2021	Movie
10	s11	Vendetta: Truth, Lies and The Mafia	2021	TV Show
11	s12	Bangkok Breaking	2021	TV Show
12	s13	Je Suis Karl	2021	Movie
13	s14	Confessions of an Invisible Girl	2021	Movie
14	s15	Crime Stories: India Detectives	2021	TV Show
15	s16	Dear White People	2021	TV Show
16	s17	Europe's Most Dangerous Man: Otto Skorzeny in ...	2020	Movie
17	s18	Falsa identidad	2020	TV Show
18	s19	Intrusion	2021	Movie
19	s20	Jaguar	2021	TV Show

In [83]:
```
movie_count_by_country=df[df['type']=='Movie'].groupby('country')['title'].r
movie_count_by_country.head(10)
```

Out[83]:
```
country
United States      2058
India               893
Unknown country     440
United Kingdom      206
Canada              122
Spain                97
Egypt                92
Nigeria              86
Indonesia            77
Turkey               76
Name: title, dtype: int64
```

In [84]:
```
Tv_Show_count_by_country=df[df['type']=='TV Show'].groupby('country')['title
Tv_Show_count_by_country.head(10)
```

Out[84]:
```
country
United States      760
Unknown country    391
United Kingdom     213
Japan              169
South Korea        158
India               79
Taiwan              68
Canada              59
France              49
Spain               48
Name: title, dtype: int64
```

In [85]:
```
df.head()
```

Out[85]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	September 25, 2021	2020	PG-13
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA

In [86]:
```
df.dtypes
```

Out[86]:
```
show_id                 object
type                  category
title                   object
director                object
cast                    object
country               category
date_added              object
release_year             int64
rating                category
duration                object
listed_in               object
description             object
cast_split              object
director_split          object
country_split           object
listed_in_split         object
dtype: object
```

In [87]:

```
df['date_added']=pd.to_datetime(df['date_added'])
df.dtypes
```

Out[87]:

```
show_id                   object
type                    category
title                     object
director                  object
cast                      object
country                 category
date_added        datetime64[ns]
release_year               int64
rating                  category
duration                  object
listed_in                 object
description               object
cast_split                object
director_split            object
country_split             object
listed_in_split           object
dtype: object
```

In [88]:

```
df['date_added']=pd.to_datetime(df['date_added'],format='%M %D,%Y')
df['week_added']=df['date_added'].dt.strftime('%Y-%U')
df
```

Out[88]:

	show_id	type	title	director	cast	country	date_added	release_year	ratin
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	2021-09-25	2020	PG 1
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV M
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV M
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV M
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV M
...	...	...	...	...	...	...	...	...	.
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-1
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-1
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-1
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-1

	show_id	type	title	director	cast	country	date_added	release_year	ratin
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV-1

201991 rows × 17 columns

In [89]:

```
tv_shows = df[df['type'] == 'TV Show']
movies = df[df['type'] == 'Movie']

# Split the 'cast' column to create a list of actors
tv_shows['cast_split'] = tv_shows['cast'].apply(lambda x: x.split(', ') if i
movies['cast_split'] = movies['cast'].apply(lambda x: x.split(', ') if isins

print(tv_shows['cast_split'])
print(movies['cast_split'])
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_14304\2992122032.py:5: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  tv_shows['cast_split'] = tv_shows['cast'].apply(lambda x: x.split(', ')
if isinstance(x, str) else [])

1       [Ama Qamata, Khosi Ngema, Gail Mabalane, Thaba...
1       [Ama Qamata, Khosi Ngema, Gail Mabalane, Thaba...
1       [Ama Qamata, Khosi Ngema, Gail Mabalane, Thaba...
1       [Ama Qamata, Khosi Ngema, Gail Mabalane, Thaba...
1       [Ama Qamata, Khosi Ngema, Gail Mabalane, Thaba...
                              ...
8800    [Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen...
8800    [Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen...
8803                                       [Unknown cast]
8803                                       [Unknown cast]
8803                                       [Unknown cast]
Name: cast_split, Length: 56148, dtype: object
0                                          [Unknown cast]
6       [Vanessa Hudgens, Kimiko Glenn, James Marsden,...
6       [Vanessa Hudgens, Kimiko Glenn, James Marsden,...
6       [Vanessa Hudgens, Kimiko Glenn, James Marsden,...
6       [Vanessa Hudgens, Kimiko Glenn, James Marsden,...
                              ...
8806    [Vicky Kaushal, Sarah-Jane Dias, Raaghav Chana...
8806    [Vicky Kaushal, Sarah-Jane Dias, Raaghav Chana...
8806    [Vicky Kaushal, Sarah-Jane Dias, Raaghav Chana...
8806    [Vicky Kaushal, Sarah-Jane Dias, Raaghav Chana...
8806    [Vicky Kaushal, Sarah-Jane Dias, Raaghav Chana...
Name: cast_split, Length: 145843, dtype: object

C:\Users\DELL\AppData\Local\Temp\ipykernel_14304\2992122032.py:6: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  movies['cast_split'] = movies['cast'].apply(lambda x: x.split(', ') if i
sinstance(x, str) else [])
```

In [90]:
```
all_actors = [actor for sublist in movies['cast_split'] for actor in sublist

actor_counts = pd.Series(all_actors).value_counts().reset_index()
actor_counts.columns = ['Actor', 'Appearances']

top_10_actors = actor_counts.head(11)

print("Top 10 Actors with frequently Appearances in movies:")
print(top_10_actors)
```

```
Top 10 Actors with frequently Appearances in movies:
             Actor  Appearances
0     Unknown cast         1328
1    Alfred Molina         1255
2      Liam Neeson         1244
3      Anupam Kher         1122
4      Salma Hayek         1092
5   John Krasinski         1072
6     James Franco         1058
7      Halle Berry         1057
8    Paul Giamatti         1026
9   Shah Rukh Khan         1007
10   Jim Broadbent          998
```

In [91]:
```
all_actors = [actor for sublist in tv_shows['cast_split'] for actor in subli

actor_counts = pd.Series(all_actors).value_counts().reset_index()
actor_counts.columns = ['Actor', 'Appearances']

top_10_actors = actor_counts.head(11)

print("Top 10 Actors with frequently Appearances in Tv Shows:")
print(top_10_actors)
```

```
Top 10 Actors with frequently Appearances in Tv Shows:
               Actor  Appearances
0   Takahiro Sakurai          843
1       Unknown cast          818
2    Yuichi Nakamura          732
3       Jun Fukuyama          679
4          Yuki Kaji          674
5     Junichi Suwabe          624
6     Hiroshi Kamiya          608
7        Raúl Méndez          597
8        Daisuke Ono          583
9      André Holland          536
10         Ai Kayano          535
```

In [92]:

```

tv_shows['director_split'] = tv_shows['director'].apply(lambda x: x.split(',
movies['director_split'] = movies['director'].apply(lambda x: x.split(', ')
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_14304\2335380343.py:1: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  tv_shows['director_split'] = tv_shows['director'].apply(lambda x: x.spli
t(', ') if isinstance(x, str) else [])
C:\Users\DELL\AppData\Local\Temp\ipykernel_14304\2335380343.py:2: SettingW
ithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  movies['director_split'] = movies['director'].apply(lambda x: x.split(',
') if isinstance(x, str) else [])
```

In [93]:

```
all_directors = [director for sublist in movies['director_split'] for direct

# Create a DataFrame with director counts
director_counts = pd.Series(all_directors).value_counts().reset_index()
director_counts.columns = ['Director', 'Appearances']

# Get the top 10 directors with the most appearances
top_10_directors = director_counts.head(11)

# Print the top 10 directors
print("\nTop 10 Directors with frequently Appearances in movies:")
print(top_10_directors)
```

```

Top 10 Directors with frequently Appearances in movies:
                Director  Appearances
0       Unknown director         1285
1           Roger Allers          935
2             Joann Sfar          700
3          Bill Plympton          700
4             Nina Paley          700
5             Tomm Moore          700
6    Mohammed Saeed Harib          700
7           Joan C. Gratz          700
8            Paul Brizzi          700
9          Gaëtan Brizzi          700
10        Michael Socha          700
```

In [94]:

```
all_directors = [director for sublist in tv_shows['director_split'] for dire

# Create a DataFrame with director counts
director_counts = pd.Series(all_directors).value_counts().reset_index()
director_counts.columns = ['Director', 'Appearances']

# Get the top 10 directors with the most appearances
top_10_directors = director_counts.head(11)

# Print the top 10 directors
print("\nTop 10 Directors with frequently Appearances in Tv Shows:")
print(top_10_directors)
```

```
Top 10 Directors with frequently Appearances in Tv Shows:
                 Director  Appearances
0        Unknown director        49358
1         Damien Chazelle          416
2         Laïla Marrakchi          416
3         Houda Benyamina          416
4               Alan Poul          416
5   Gautham Vasudev Menon          286
6            Priyadarshan          198
7                  Sarjun          198
8     Rathindran R Prasad          198
9            Arvind Swamy          198
10       Karthik Subbaraj          198
```

In [95]:

```
genres = df['listed_in'].str.split(', ').explode().str.strip()
df.head(10)
genres
```

Out[95]:

```
0                Documentaries
1       International TV Shows
1                    TV Dramas
1                 TV Mysteries
1       International TV Shows
                 ...
8806      International Movies
8806          Music & Musicals
8806                    Dramas
8806      International Movies
8806          Music & Musicals
Name: listed_in, Length: 506879, dtype: object
```

In [96]:

```
df1=pd.read_csv(r"C:\Users\DELL\Desktop\Data Science Scaler\netflix.csv")
df1.head(50)
```

Out[96]:

	show_id	type	title	director	cast	country	date_added	release_
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	

In [97]:

```
df1['director'].fillna("Unknown director" , inplace = True)
df1['cast'].fillna("Unknown cast" , inplace = True)
df1['country'].fillna("Unknown country" , inplace = True)
df1['date_added'].fillna("January 1, 1900" , inplace = True)
df1['rating'].fillna("Unknown rating" , inplace = True)
df1['duration'].fillna("Unknown duration" , inplace = True)
missing_values = df1.isnull().sum()
print("\nMissing Values:")
print(missing_values)
df1["type"] = df1["type"].astype("category")
df1["country"] = df1["country"].astype("category")
df1["rating"] = df1["rating"].astype("category")
```

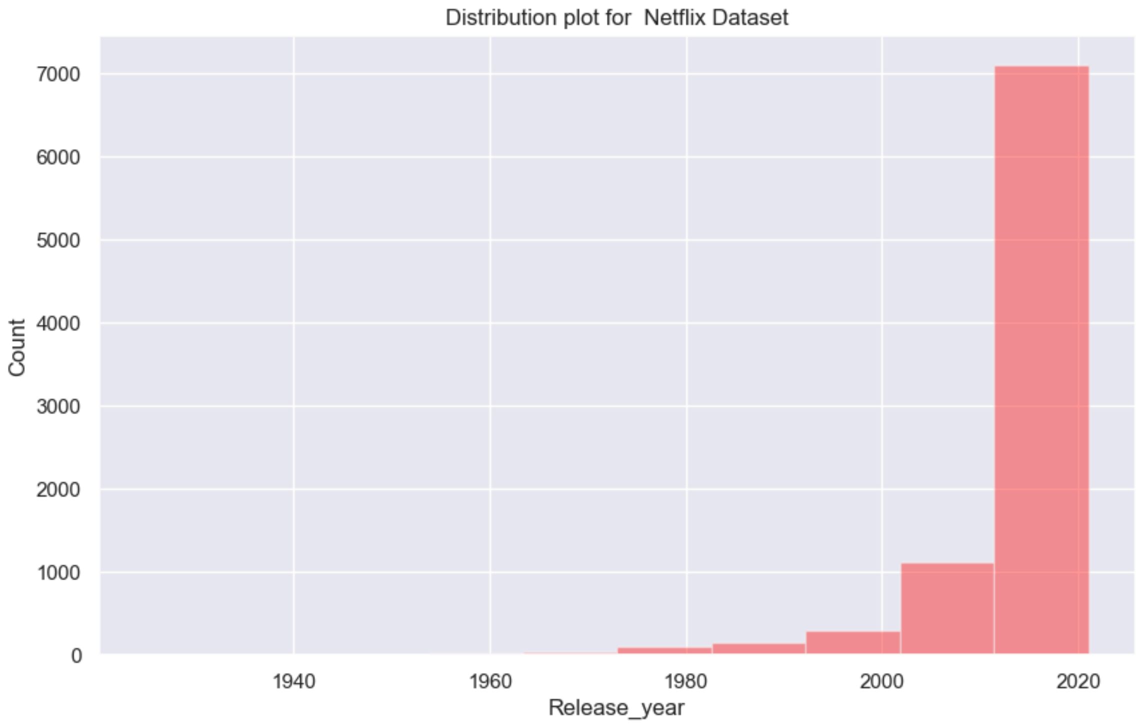
```

Missing Values:
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
dtype: int64
```

In [98]:
```
df1['date_added'] = pd.to_datetime(df1['date_added'])

# Calculate the difference in days between 'date_added' and 'release_year'
df1['days_to_add'] = (df1['date_added'] - pd.to_datetime(df1['release_year']

# Calculate the mode (most common) value for 'days_to_add'
mode_days_to_add = df1['days_to_add'].mode().iloc[0]

print(f"The most common time duration between release and addition to Netfli
```

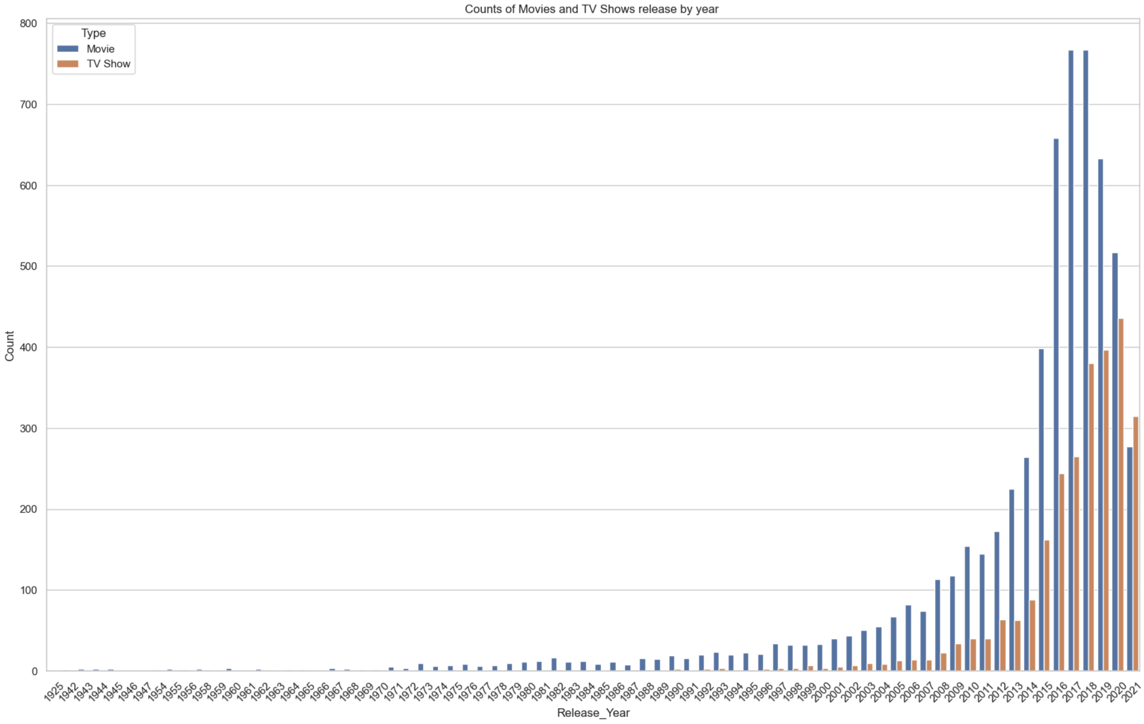
```
The most common time duration between release and addition to Netflix is a
pproximately 334 days.
```

In [99]:
```
# 1.For continuous variable(s): Distplot, countplot, histogram for univariat

# Distplot for 'release_year'

# Create a distribution plot for rating
plt.figure(figsize=(10,6))
sns.set(style="darkgrid") #set the plot style
sns.distplot(df1['release_year'],bins=10,kde=False,color='red')
# add lebels and a title
plt.xlabel('Release_year')
plt.ylabel('Count')
plt.title('Distribution plot for  Netflix Dataset')
# show the plot
plt.show()
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_14304\1136020693.py:8: UserWarn
ing:

`distplot` is a deprecated function and will be removed in seaborn v0.14.
0.

Please adapt your code to use either `displot` (a figure-level function wi
th
similar flexibility) or `histplot` (an axes-level function for histogram
s).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://
gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(df1['release_year'],bins=10,kde=False,color='red')
```



In [100]:
```
# for countplot

# Create a count plot for the "genre" column
sns.set(style="whitegrid") #set the style for the plot
plt.figure(figsize=(20,12)) #set the figure size
# Assuming 'Genre'is the name of the categorical variable
sns.countplot(data=small_df,x='release_year',hue='type')
# add lebel and a title
plt.xlabel('Release_Year')
plt.ylabel('Count')
plt.title('Counts of Movies and TV Shows release by year ')
plt.legend(title='Type',loc='upper left',labels=['Movie','TV Show'])
# Rotate x-axis labels for better readability
plt.xticks(rotation=45)
# show the plot
plt.show()
```

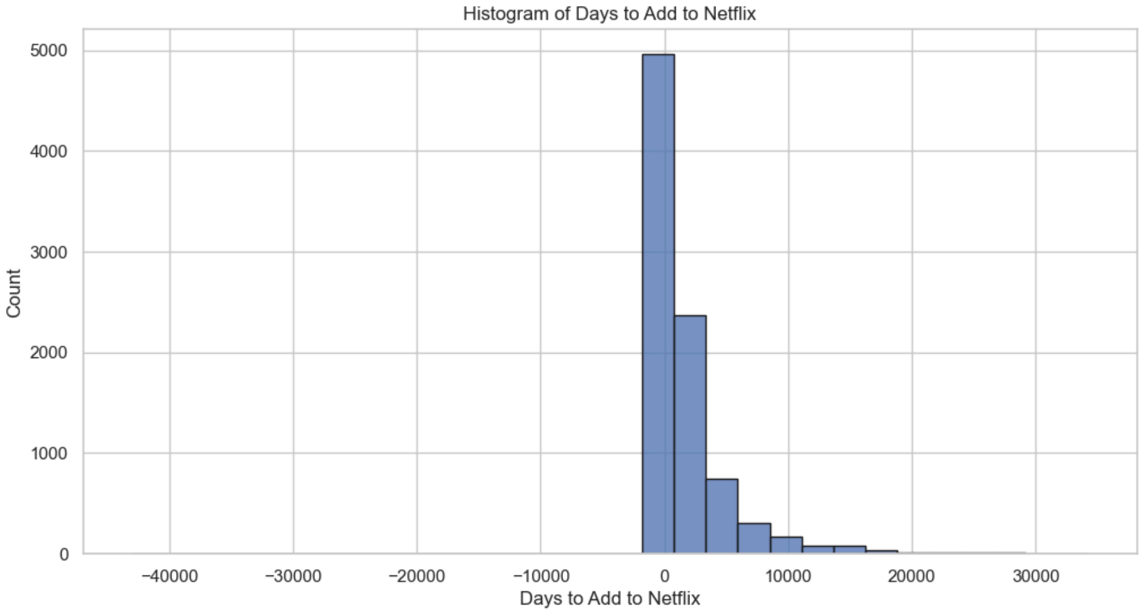


In [101]:
```
# df['date_added'] = pd.to_datetime(df1['date_added'], format='%B %d, %Y')

# Calculate the difference in days between 'date_added' and 'release_year'
df['days_to_add'] = (df['date_added'] - pd.to_datetime(df['release_year'], f

# Calculate the mode (most common) value for 'days_to_add'
mode_days_to_add = df['days_to_add'].mode().iloc[0]

print(f"The most common time duration between release and addition to Netfli
```

```
The most common time duration between release and addition to Netflix is a
pproximately 547 days.
```

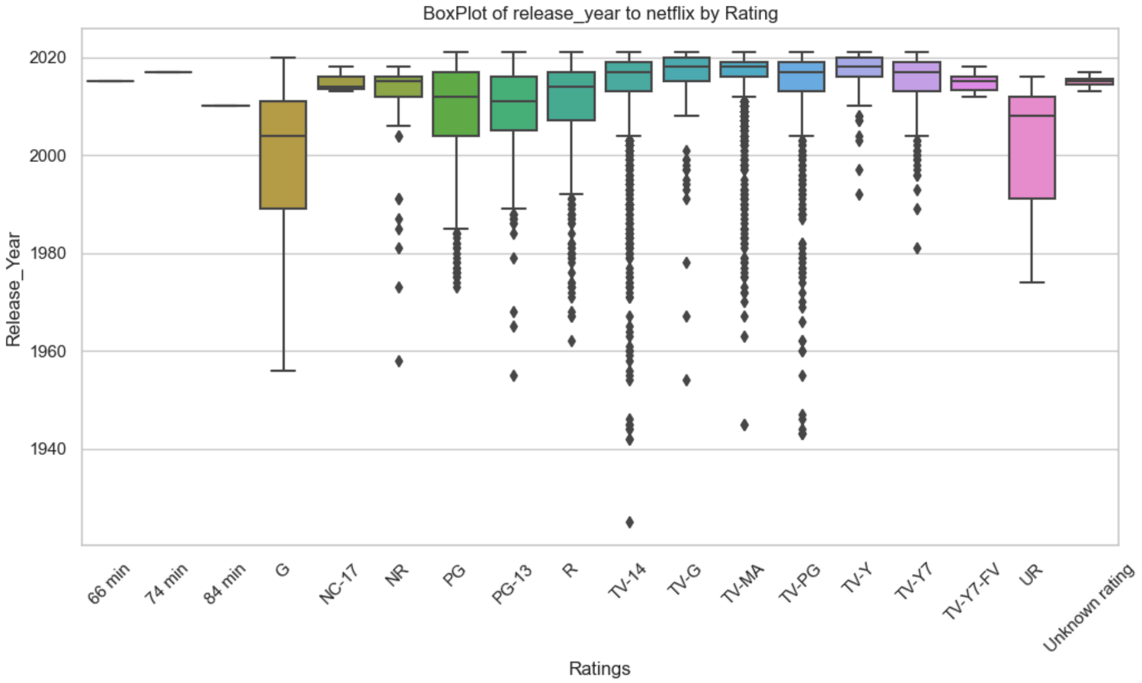
In [102]:

```
# Histogram for days_to_add
plt.figure(figsize=(12, 6))
sns.histplot(df1['days_to_add'], bins=30,edgecolor='k')
plt.title('Histogram of Days to Add to Netflix')
plt.xlabel('Days to Add to Netflix')
plt.ylabel('Count')
plt.grid(True)
plt.show()
```



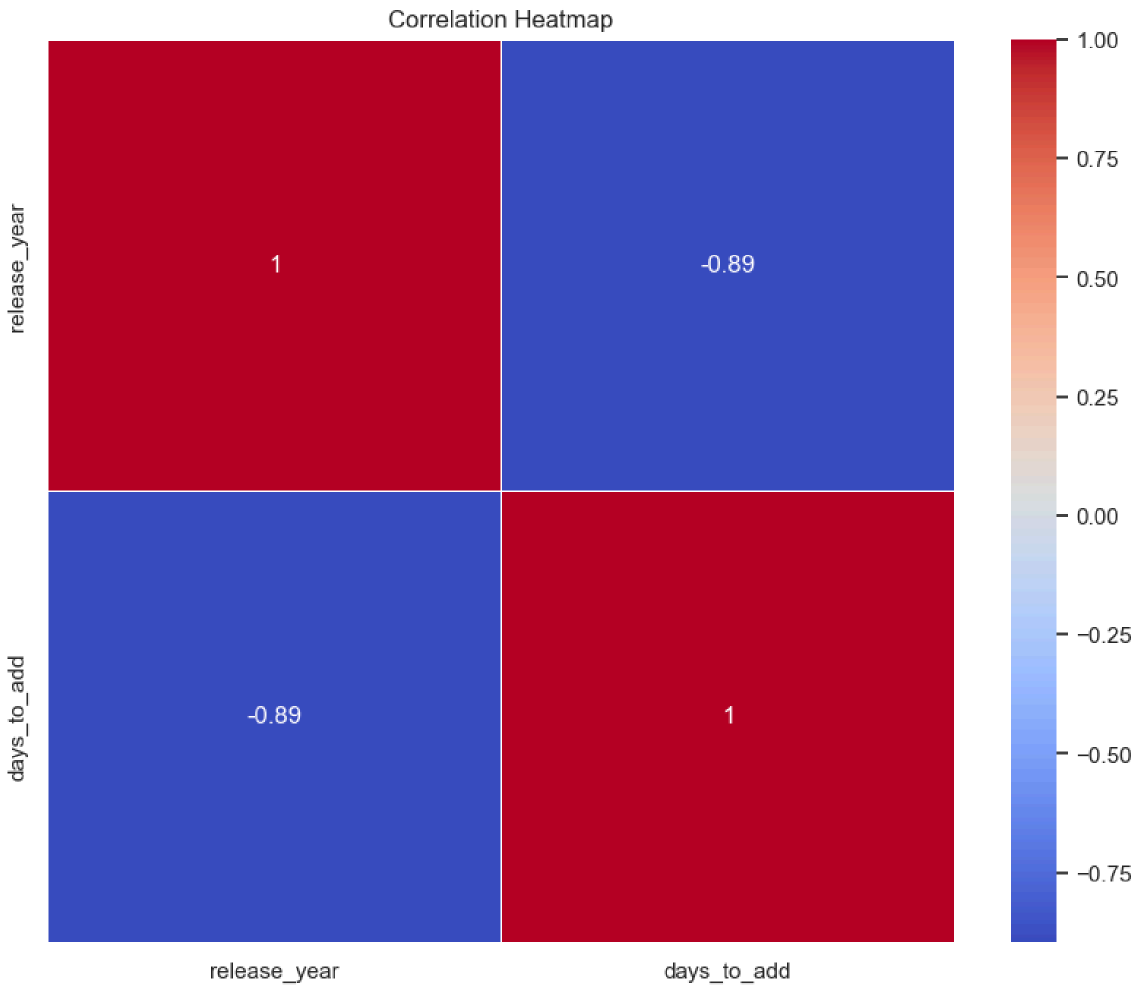
In [103]:

```
# Boxplot for Categorical Variables
plt.figure(figsize=(12,6))
sns.boxplot(data=df1,x='rating',y='release_year')
plt.title('BoxPlot of release_year to netflix by Rating')
plt.xlabel('Ratings')
plt.ylabel('Release_Year')
plt.xticks(rotation=45)
plt.show()
```



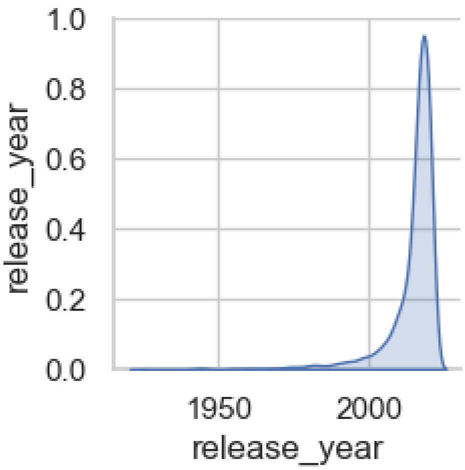
In [104]:
```
# Correlation Analysis
# Calculate the correlation matrix
correlation_matrix=df1.corr()
# Heatmap to visualize the correlations
plt.figure(figsize=(10,8))
sns.heatmap(correlation_matrix,annot=True,cmap='coolwarm',linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```

```
C:\Users\DELL\AppData\Local\Temp\ipykernel_14304\1233909943.py:3: FutureWa
rning: The default value of numeric_only in DataFrame.corr is deprecated.
In a future version, it will default to False. Select only valid columns o
r specify the value of numeric_only to silence this warning.
  correlation_matrix=df1.corr()
```



In [105]:

```
# Pairplot for selected numeric columns
import seaborn as sns
numeric_cols=['release_year','date_added']
sns.pairplot(df1[numeric_cols],diag_kind='kde')
plt.show()
```



Q5.Missing Value & Outlier check (Treatment optional)

In [106]:

```
missing_values=df1.isna().sum()
missing_values
```

Out[106]:

```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
days_to_add     0
dtype: int64
```

```
In [107]: df1.head(10)
```

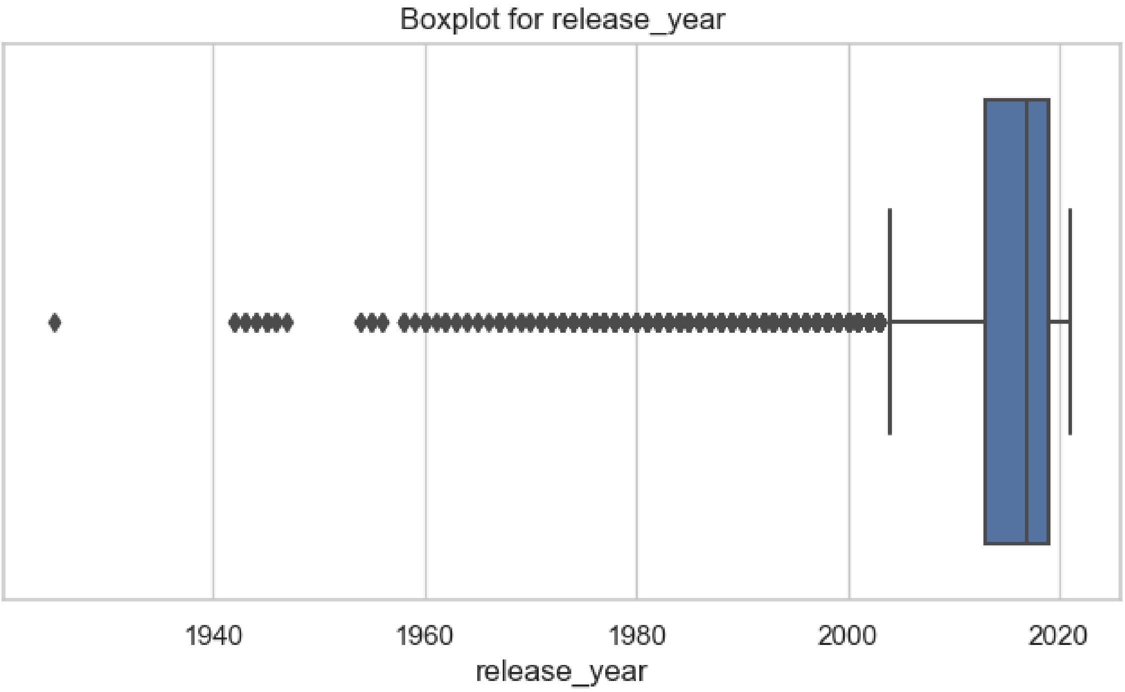
Out[107]:

	show_id	type	title	director	cast	country	date_added	release_year	ra
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown cast	United States	2021-09-25	2020	
1	s2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown country	2021-09-24	2021	
3	s4	TV Show	Jailbirds New Orleans	Unknown director	Unknown cast	Unknown country	2021-09-24	2021	
4	s5	TV Show	Kota Factory	Unknown director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	Unknown country	2021-09-24	2021	
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Unknown country	2021-09-24	2021	
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	2021-09-24	1993	
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	2021-09-24	2021	T

	show_id	type	title	director	cast	country	date_added	release_year	ra
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	2021-09-24	2021	

In [108]:

```
# Check for Outlier
plt.figure(figsize=(8,4))
sns.boxplot(x='release_year',data=df1)
plt.title('Boxplot for release_year')
plt.show()
```



Q6.Insights based on Non-Graphical and Visual Analysis (10 Points)

1 Comments on the range of attributes

2 Comments on the distribution of the variables and relationship between them

3 Comments for each univariate and bivariate plot

In [109]:

```
# Comments on the range of attributes date_added $ release_year

# For the date_added attributes:
# extract the minimum and maximum dates
min_date=df1['date_added'].min()
max_date=df1['date_added'].max()
```

In [110]:

```
# Print the result
print('Minimum Date:',min_date)
print('Maximum Date:',max_date)
```

```
Minimum Date: 1900-01-01 00:00:00
Maximum Date: 2021-09-25 00:00:00
```

In [111]:

```
# Find the minimum and maximum years
min_year=df1['release_year'].min()
max_year=df1['release_year'].max()
```

In [112]:

```
# Print the results
print("Minimum year:", min_year)
print("Maximum year:", max_year)
```

```
Minimum year: 1925
Maximum year: 2021
```

2 Comments on the distribution of the variables and relationship between them

In [113]:

```
# for rating_distribution
ratings_distribution=df1['rating'].describe()
ratings_distribution
```

Out[113]:

```
count      8807
unique       18
top       TV-MA
freq       3207
Name: rating, dtype: object
```

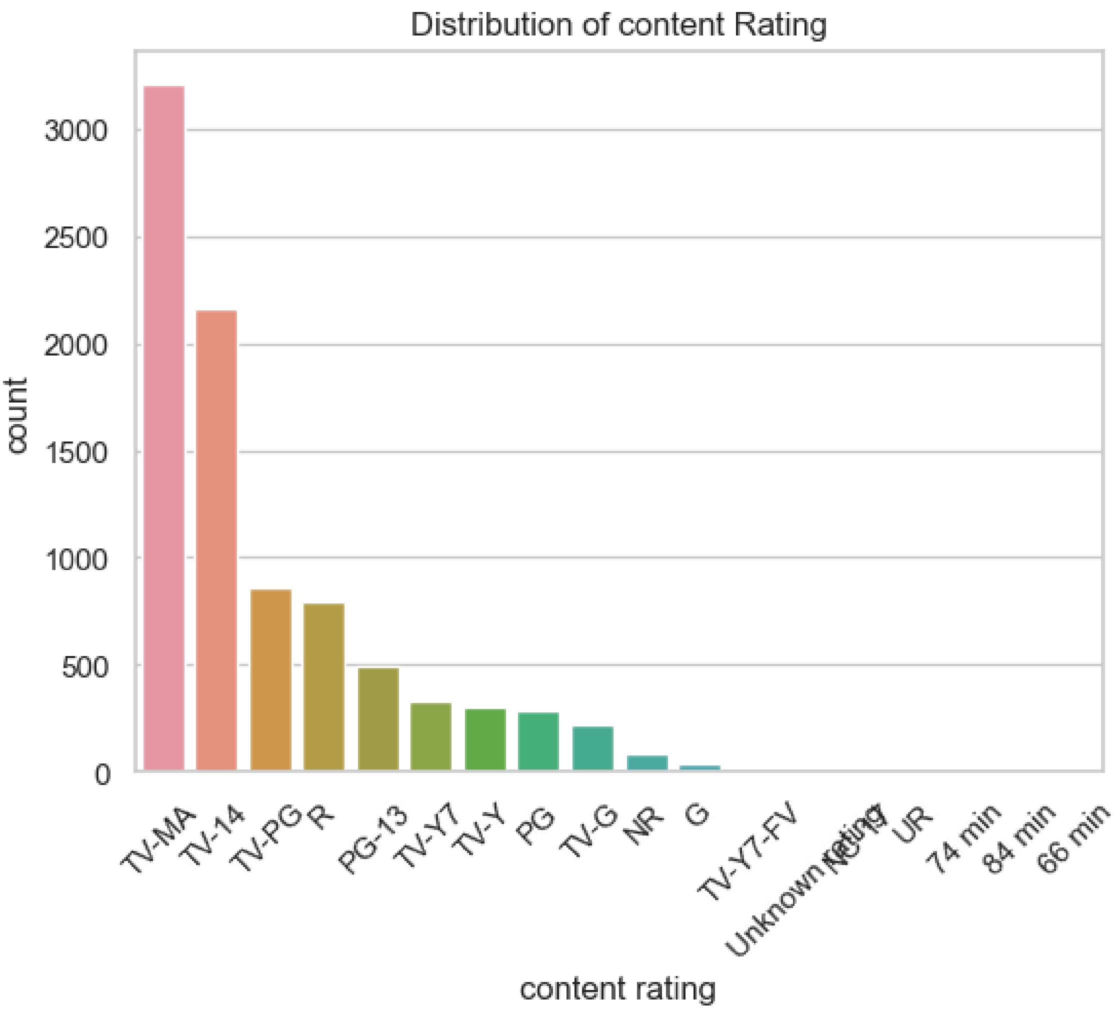
In [114]:

```
# For release_year distribution

release_year_distribution=df1['release_year'].value_counts().sort_index()
release_year_distribution
```

Out[114]:

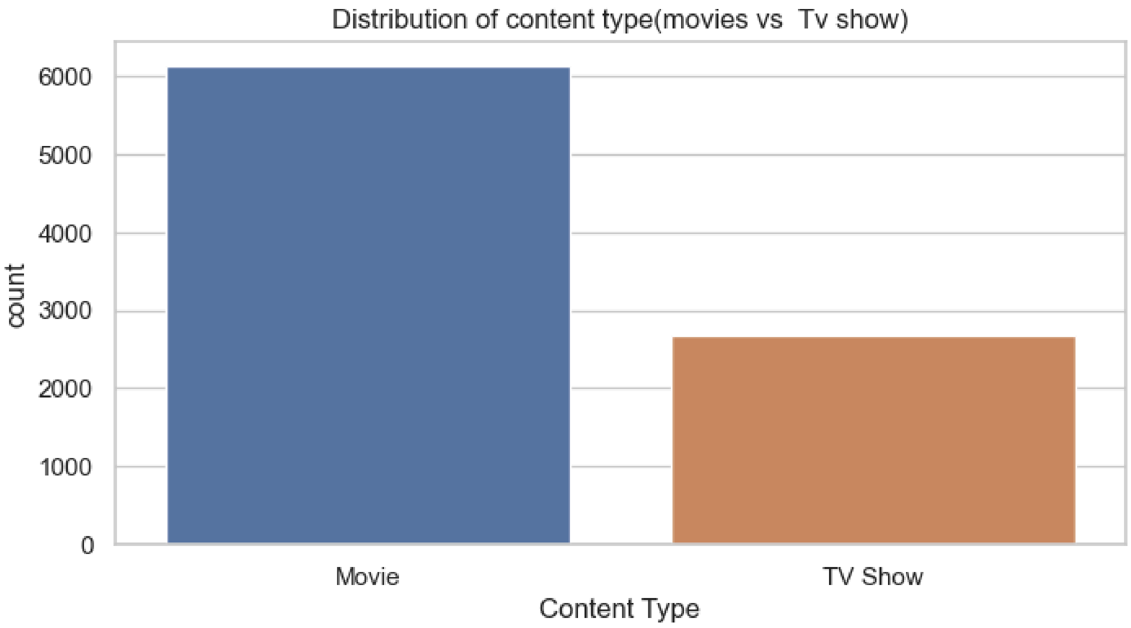
```
1925       1
1942       2
1943       3
1944       3
1945       4
        ... 
2017    1032
2018    1147
2019    1030
2020     953
2021     592
Name: release_year, Length: 74, dtype: int64
```

In [115]:
```
sns.countplot(x='rating',data=df1,order=df1['rating'].value_counts().index)
plt.title('Distribution of content Rating')
plt.xlabel('content rating')
plt.ylabel('count')
plt.xticks(rotation=45)
plt.show()
```



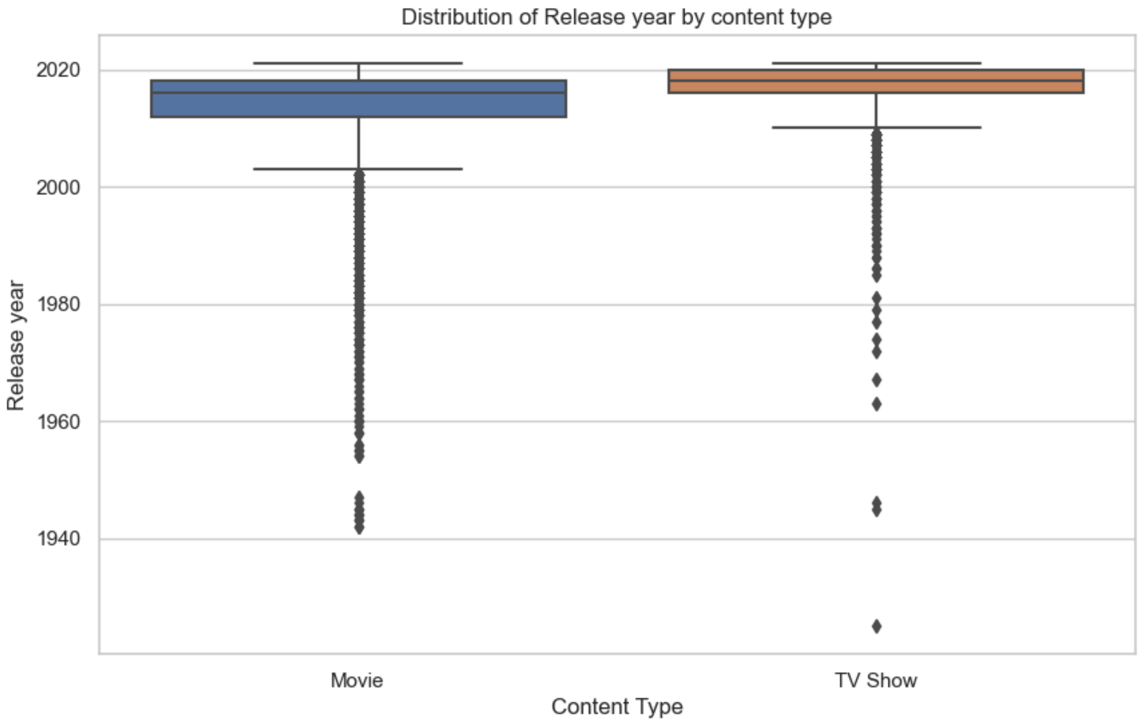
In [116]:

```
# Distribution of Content Type(Movies vs show)
plt.figure(figsize=(8,4))
sns.countplot(x='type',data=df1)
plt.title('Distribution of content type(movies vs  Tv show)')
plt.xlabel('Content Type')
plt.ylabel('count')
plt.show()
```



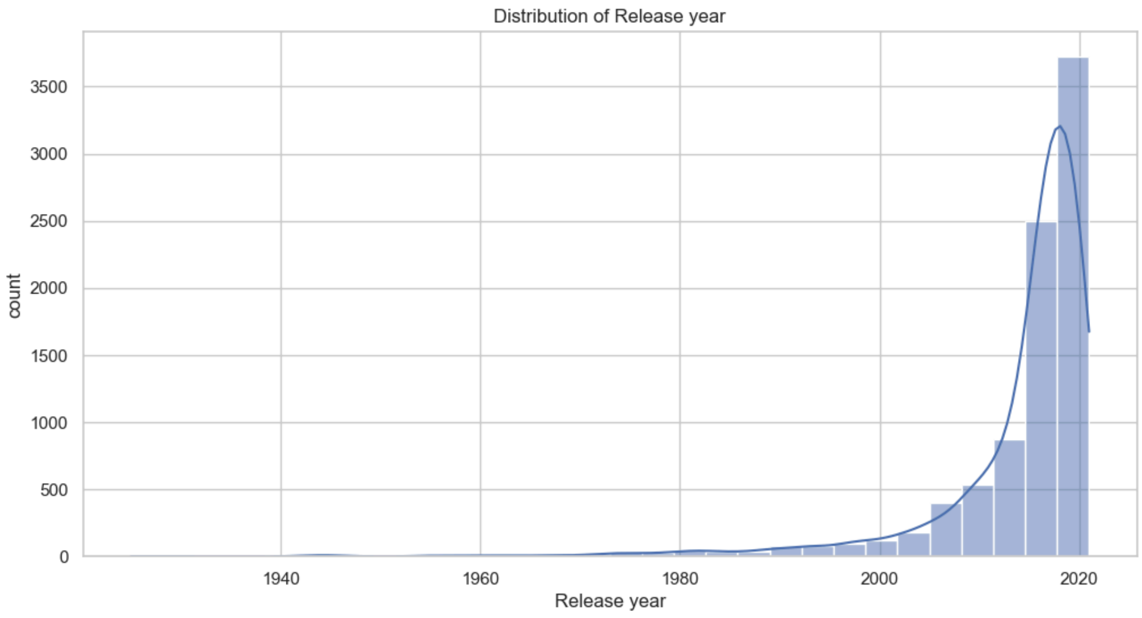
In [117]:

```
# Content Type and release year
plt.figure(figsize=(10,6))
sns.boxplot(x='type',y='release_year',data=df1)
plt.title('Distribution of Release year by content type')
plt.xlabel('Content Type')
plt.ylabel('Release year')
plt.show()
```



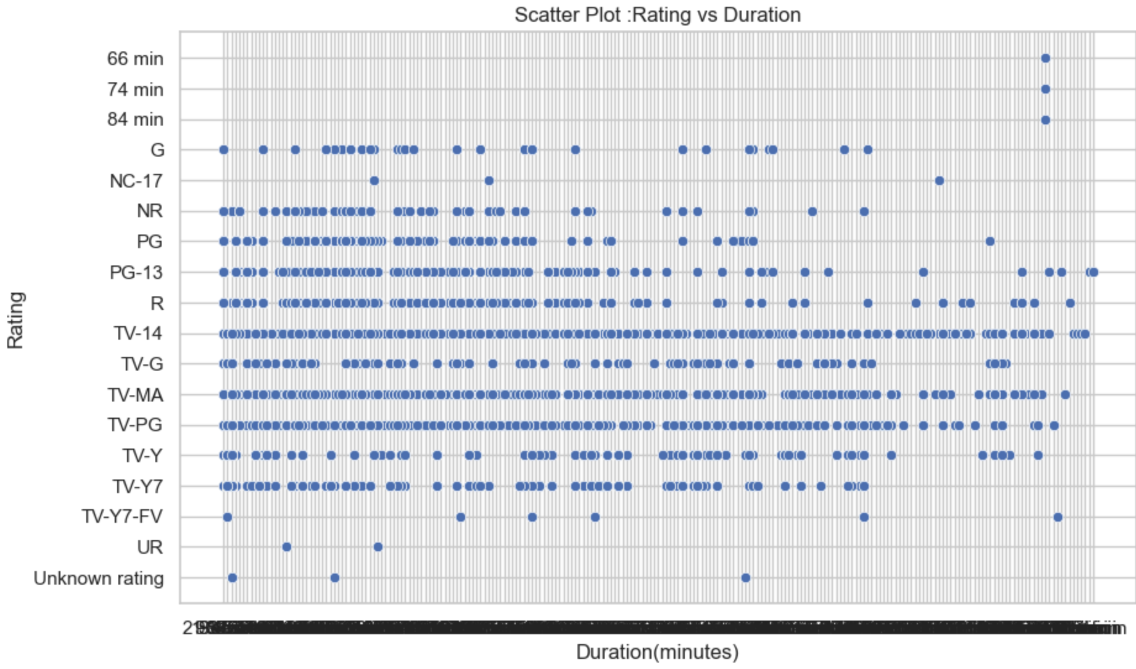
In [118]:

```
plt.figure(figsize=(12,6))
sns.histplot(x='release_year',bins=30,kde=True,data=df1)
plt.title('Distribution of Release year')
plt.xlabel('Release year')
plt.ylabel('count')
plt.show()
```



In [119]:

```
# relation between rating and duration
plt.figure(figsize=(10,6))
sns.scatterplot(x='duration',y='rating',data=df1)
plt.title('Scatter Plot :Rating vs Duration')
plt.xlabel('Duration(minutes)')
plt.ylabel('Rating')
plt.show()
```



7 Business Insights-Should include patterns observed in the data along with what you can

infer from it

1. Movies content are more than tv shows.
2. TV-MA has highest content rating.
3. Highest content releasing year is 2020-2021
4. TV shows are added much faster as compare to movies.
5. In initially the movies ratio is much higher as compare to Tv shows.
6. As the years are increasing the movies and shows are inreasing

8 Recommendations - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand.

1. We can add more international movies and Tv shows.
2. We can reduce the time duration between release and addition to Netflix.
3. We can add some more family friendly content on netflix for every age group .

```
In [ ]:
```

```
In [ ]:
```