In [2]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:

```
!wget "https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/origi
```

```
--2024-07-24 05:32:09--  https://d2beiqkhq929f0.cloudfront.net/public_assets/asset
s/000/000/940/original/netflix.csv
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)... 18.160.
146.106, 18.160.146.28, 18.160.146.45, ...
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)|18.16
0.146.106|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3399671 (3.2M) [text/plain]
Saving to: ‘netflix.csv’

netflix.csv         100%[===================>]   3.24M  --.-KB/s    in 0.09s

2024-07-24 05:32:09 (37.5 MB/s) - ‘netflix.csv’ saved [3399671/3399671]
```

In [4]:

```
df=pd.read_csv("netflix.csv")
df
```

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA
...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy	United States	January 11, 2020	2006	PG

	show_id	type	title	director	cast	country	date_added	release_year	rating
					Chase, Kate Ma...				
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14

8807 rows × 12 columns

In [5]:
```
df.describe()
```

Out[5]:

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

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

In [7]:
```
df.isna().sum()
```

Out[7]:

```
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
```

In [8]:

```
#As the Null values are very few dropping them for better analysis
df.dropna(subset= ["date_added", "rating", "duration"], inplace=True)
```

In [9]:

```
df1=df.loc[:,["title","director","cast","country","listed_in"]]
df1.head()
```

Out[9]:

	title	director	cast	country	listed_in
0	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	Documentaries
1	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	International TV Shows, TV Dramas, TV Mysteries
2	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	Crime TV Shows, International TV Shows, TV Act...
3	Jailbirds New Orleans	NaN	NaN	NaN	Docuseries, Reality TV
4	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	International TV Shows, Romantic TV Shows, TV ...

What does 'good' look like

1. Find the counts of each categorical variable both using graphical and non-graphical analysis.

Non-Graphical analysis

In [10]:

```
col=["director","cast","country","listed_in"]
for i in df1.columns:
  if i in col:
    df1[i]=df1[i].replace(", ",",")
```

In [11]:

```
for i in df1.columns:
  if i in col:
    df1[i]=df1[i].str.split(",")
```

In [12]:

```
df1=df1.explode("director")
df1=df1.explode("cast")
```

```
df1=df1.explode("country")
df1=df1.explode("listed_in")
```

In [13]:

```
df1
```

Out[13]:

	title	director	cast	country	listed_in
0	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	Documentaries
1	Blood & Water	NaN	Ama Qamata	South Africa	International TV Shows
1	Blood & Water	NaN	Ama Qamata	South Africa	TV Dramas
1	Blood & Water	NaN	Ama Qamata	South Africa	TV Mysteries
1	Blood & Water	NaN	Khosi Ngema	South Africa	International TV Shows
...	...	...	...	...	...
8806	Zubaan	Mozez Singh	Anita Shabdish	India	International Movies
8806	Zubaan	Mozez Singh	Anita Shabdish	India	Music & Musicals
8806	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Dramas
8806	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	International Movies
8806	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	Music & Musicals

201837 rows × 5 columns

In [14]:

```
df.drop(columns=["director","cast","country","listed_in"],inplace=True)
```

In [15]:

```
df_final=df.merge(df1,on="title")
df_final.head()
```

Out[15]:

	show_id	type	title	date_added	release_year	rating	duration	description	director	
0	s1	Movie	Dick Johnson Is Dead	September 25, 2021	2020	PG-13	90 min	As her father nears the end of his life, filmm...	Kirsten Johnson	
1	s2	TV Show	Blood & Water	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	NaN	Qar
2	s2	TV Show	Blood & Water	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	NaN	Qar
3	s2	TV Show	Blood & Water	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	NaN	Qar
4	s2	TV Show	Blood & Water	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	NaN	K Ng

In [16]:

```
# converting data type of date_added column
df_final['date_added']=pd.to_datetime(df_final['date_added'], errors='coerce')
df_final['month_added'] = df_final['date_added'].dt.month
df_final['year_added'] = df_final['date_added'].dt.year
df_final['week_added'] = df_final['date_added'].dt.isocalendar().week
df_final
```

Out[16]:

	show_id	type	title	date_added	release_year	rating	duration	description	directo
0	s1	Movie	Dick Johnson Is Dead	2021-09-25	2020	PG-13	90 min	As her father nears the end of his life, filmm...	Kirste Johnso
1	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Na
2	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Na
3	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Na
4	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Na
...	...	...	...	...	...	...	...	...	.
201832	s8807	Movie	Zubaan	2019-03-02	2015	TV-14	111 min	A scrappy but poor boy worms his way into a ty...	Moze Sing
201833	s8807	Movie	Zubaan	2019-03-02	2015	TV-14	111 min	A scrappy but poor boy worms his way into a ty...	Moze Sing
201834	s8807	Movie	Zubaan	2019-03-02	2015	TV-14	111 min	A scrappy but poor boy worms his way into a ty...	Moze Sing
201835	s8807	Movie	Zubaan	2019-03-02	2015	TV-14	111 min	A scrappy but poor boy worms his way into a ty...	Moze Sing
201836	s8807	Movie	Zubaan	2019-03-02	2015	TV-14	111 min	A scrappy but poor boy worms his way into a ty...	Moze Sing

201837 rows × 15 columns

```
In [17]: df_final.fillna(value={"director":"Unknown Director","country":"Unknown Country","c
```

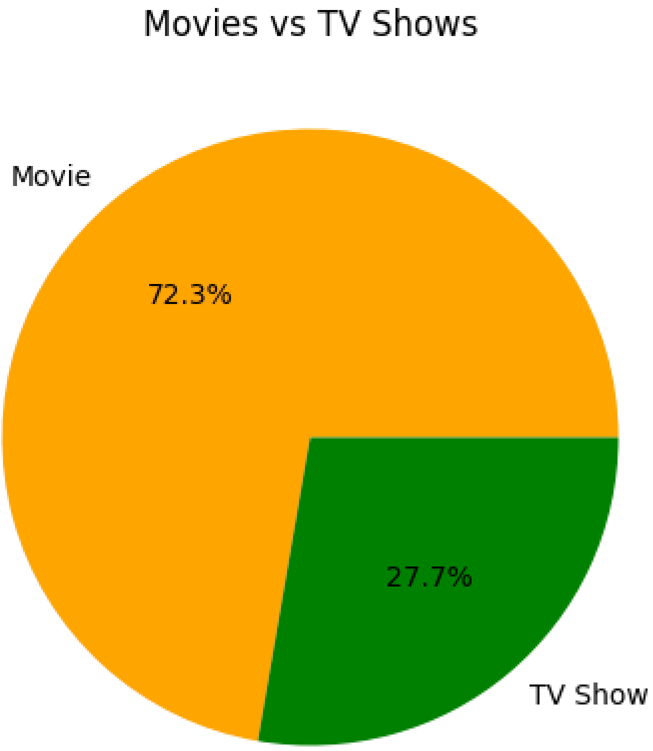
```
In [18]: df_final.head()
```

Out[18]:

	show_id	type	title	date_added	release_year	rating	duration	description	director	
0	s1	Movie	Dick Johnson Is Dead	2021-09-25	2020	PG-13	90 min	As her father nears the end of his life, filmm...	Kirsten Johnson	Ur
1	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	C
2	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	C
3	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	C
4	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	

Graphical analysis

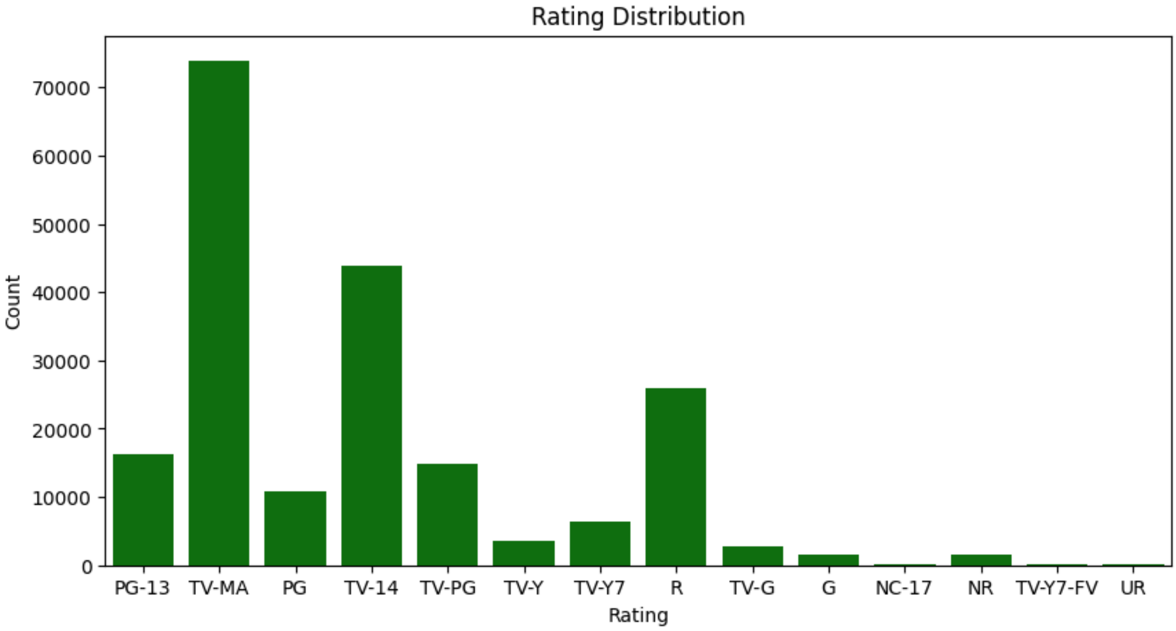
```
In [19]: mylabel = df_final['type'].value_counts()
         label = mylabel.index
         sizes = mylabel.values
         plt.figure(figsize=(10,5))
         plt.pie(sizes, labels=label, autopct='%1.1f%%', colors=['orange', 'green'])
         plt.title('Movies vs TV Shows')
         plt.show()
```



Insight -

- Movies make up 72.3% of the content, indicating a strong focus on feature films over TV shows.
- TV shows constitute only 27.7% of the content, highlighting an opportunity for growth in serialized content.

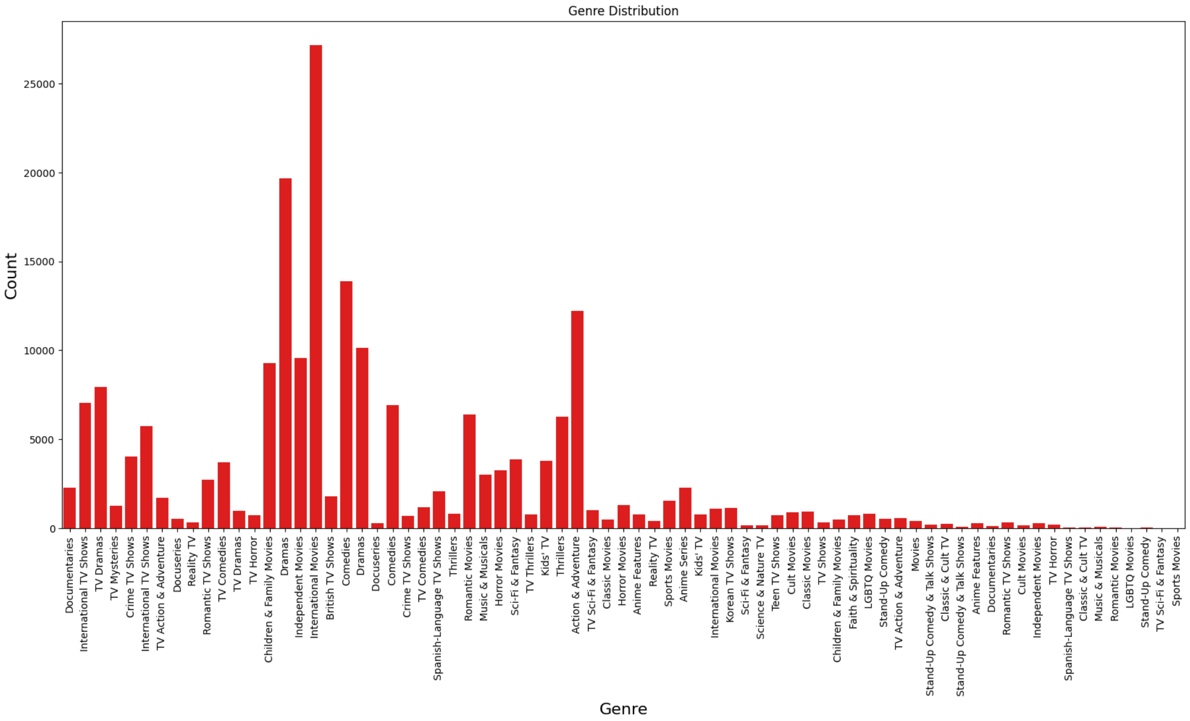
In [20]:
```
#Rating count distribution
plt.figure(figsize=(10,5))
sns.countplot(data=df_final,x="rating",color="green")
plt.xlabel("Rating")
plt.ylabel("Count")
plt.title("Rating Distribution")
plt.show()
```



Insight -

- TV-MA content has the highest count, indicating strong viewer preference for mature-rated content.
- Ratings like TV-Y, TV-Y7, and G have significantly lower counts, suggesting less emphasis on family-friendly content.

In [21]:
```
#Different genres using countplot
fig = plt.figure(figsize = (20,20))
plt.subplot(2,1,1)
plt.xticks(rotation = 90)
sns.countplot(data = df_final, x = 'listed_in',color="red")
plt.xlabel("Genre",fontsize=16)
plt.ylabel("Count",fontsize=16)
plt.title("Genre Distribution")
plt.show()
```



Insight -

- Dramas, international TV shows, and children & family movies are among the most popular genres.
- Genres like science & nature TV, teen TV shows, and anime movies have lower counts, indicating potential areas for content expansion

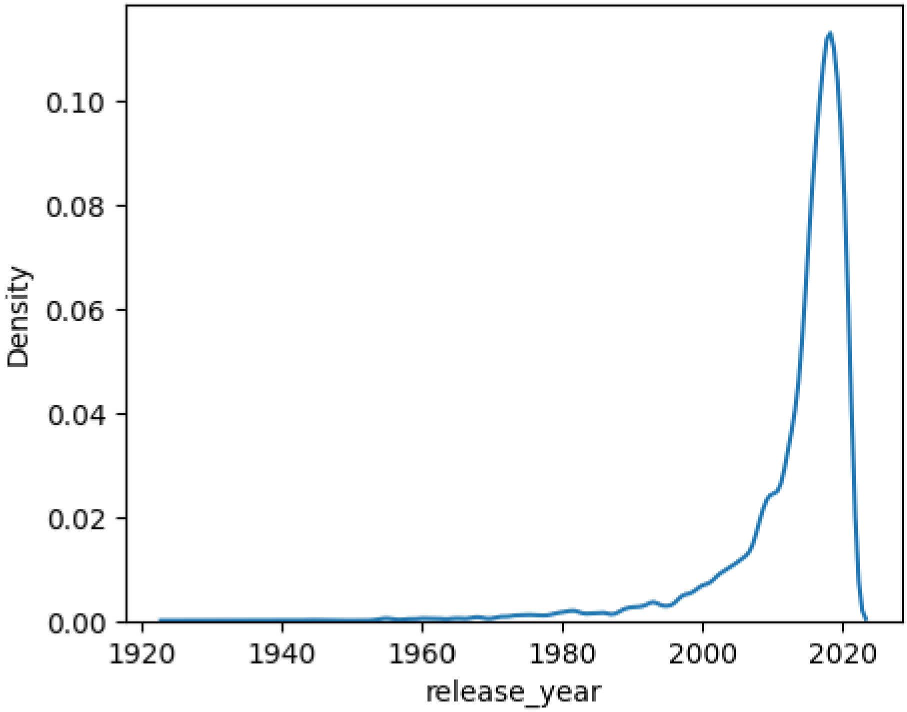
In [22]:
```
# getting top 10 directors
# dropping of the unknown rating to get better visualisation, and for that creating
df_new= df_final.loc[df_final['director']!='Unknown Director']
df_new.groupby('director')['title'].nunique().sort_values(ascending  = False)[0:11]
```

Out[22]:

	director	title
0	Rajiv Chilaka	22
1	Jan Suter	18
2	Raúl Campos	18
3	Suhas Kadav	16
4	Marcus Raboy	16
5	Jay Karas	15
6	Cathy Garcia-Molina	13
7	Jay Chapman	12
8	Martin Scorsese	12
9	Youssef Chahine	12
10	Steven Spielberg	11

In [29]:

```
# Movies and TV shows released trend analysis
plt.figure(figsize=(5,4))
sns.kdeplot(df_final['release_year'])
plt.show()
```

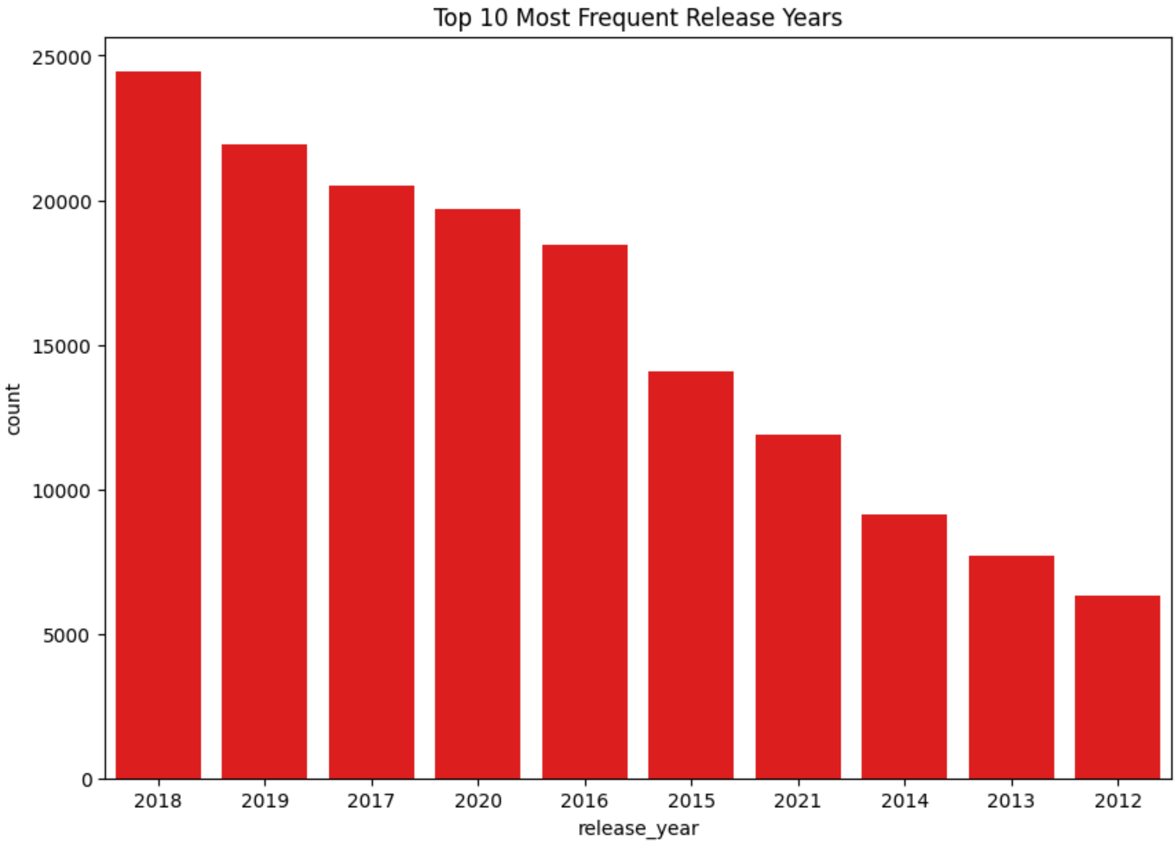


In [24]:

```
#Count of Movies,TV Shows releases,which year has most releases on netflix
plt.figure(figsize=(10,7))
sns.countplot(data=df_final, x='release_year',order=df_final['release_year'].value_
plt.title('Top 10 Most Frequent Release Years')
```

Out[24]:

```
Text(0.5, 1.0, 'Top 10 Most Frequent Release Years')
```



In [25]:

```
df_final["country"]=df_final["country"].str.strip()
```

2. Comparison of TV Shows vs. Movies.

Find the number of movies produced in each country and pick the top 10 countries.

In [26]:

```
#Movies count by Various Countries
df_co=df_final.loc[df_final["country"]!="Unknown Country"]
df_co[(df_co['type'] == 'Movie')]['country'].value_counts().head(11)
```

Out[26]:

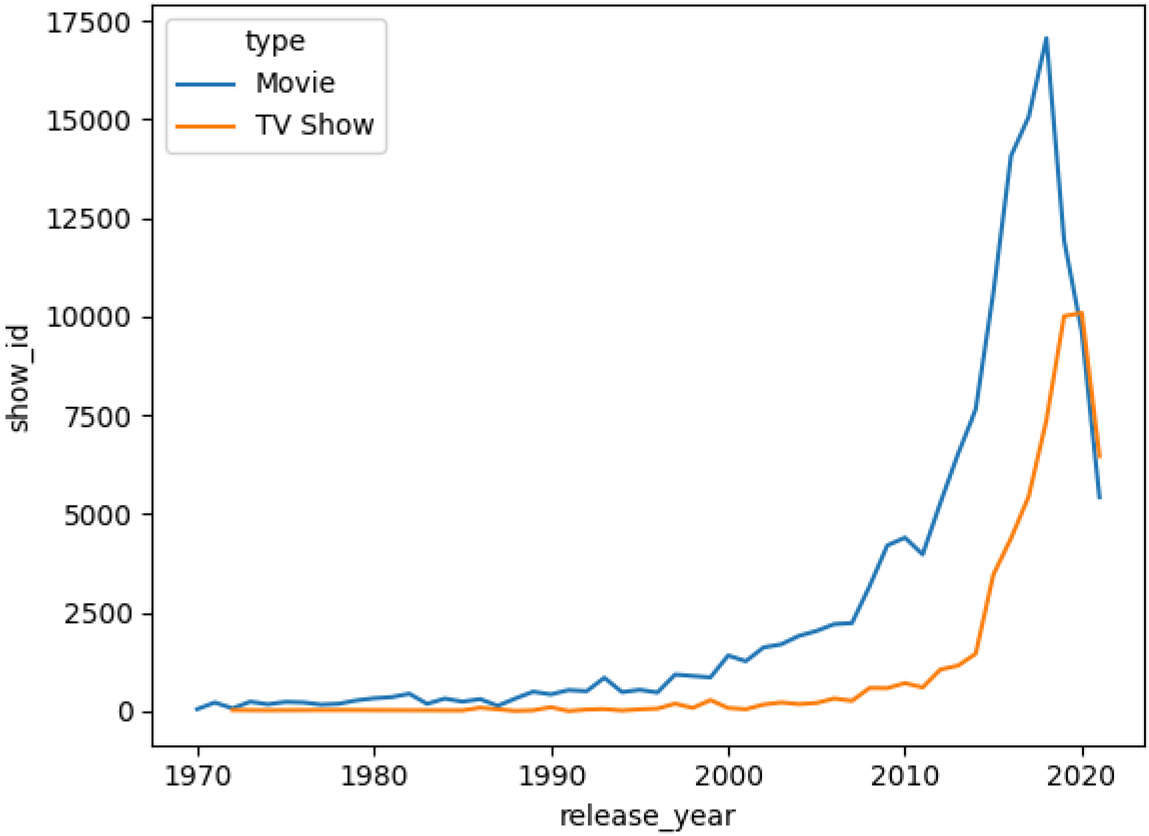
```
country
United States     45814
India             21411
United Kingdom     8580
France             6607
Canada             5738
Japan              3525
Spain              3469
Germany            3427
China              2377
Nigeria            2236
Hong Kong          2205
Name: count, dtype: int64
```

In [27]:

```
#TV Shows count by Various Countries
df_co=df_final.loc[df_final["country"]!="Unknown Country"]
df_co[(df_co['type'] == 'TV Show')]['country'].value_counts().head(11)
```

```
Out[27]: country
         United States     13449
         Japan              5074
         United Kingdom     4358
         South Korea        3754
         Canada             2177
         Mexico             2018
         Spain              1846
         Taiwan             1719
         France             1647
         India              1403
         Colombia           1284
         Name: count, dtype: int64
```

```
In [28]: #Movies Vs TV Shows release trend analysis
         s = df_final.groupby(['release_year','type'])['show_id'].count()
         s = s.reset_index()
         s = s[s['release_year'] >= 1970]
         sns.lineplot(data = s, x = 'release_year', y = 'show_id',hue='type')
         plt.show()
```



Insight -

- The graph shows a sharp increase in the release of both movies and TV shows from 2000 onwards, peaking around 2018.

3. What is the best time to launch a TV show?

a. Find which is the best week to release the Tv-show or the movie. Do the analysis separately for

Tv-shows and Movies

```
In [ ]: best_week = df_final.groupby('week_added')['type'].value_counts()
        best_week
```

```
Out[ ]: week_added  type    
        1           Movie       8456
                    TV Show      938
        2           Movie       1618
                    TV Show      585
        3           Movie       2031
                                ... 
        51          TV Show     1173
        52          Movie       1840
                    TV Show     1111
        53          Movie       1413
                    TV Show     1038
        Name: count, Length: 106, dtype: int64
```

b. Find which is the best month to release the Tv-show or the movie. Do the analysis separately for Tv-shows and Movies

```
In [ ]: best_month = df_final.groupby('type')['month_added'].value_counts()
        best_month
```

```
Out[ ]: type     month_added
        Movie    7.0            15075
                 1.0            13945
                 10.0           13541
                 9.0            13219
                 12.0           12768
                 4.0            12537
                 8.0            11923
                 6.0            11616
                 3.0            11500
                 11.0           11065
                 5.0             9579
                 2.0             9137
        TV Show  12.0            5297
                 7.0             5129
                 8.0             5029
                 6.0             4959
                 9.0             4818
                 4.0             4460
                 11.0            4428
                 3.0             4201
                 10.0            4199
                 5.0             4111
                 1.0             3941
                 2.0             3772
        Name: count, dtype: int64
```

4. Analysis of actors/directors of different types of shows/movies.

a. Identify the top 10 directors who have appeared in most movies or TV shows.

In [31]:

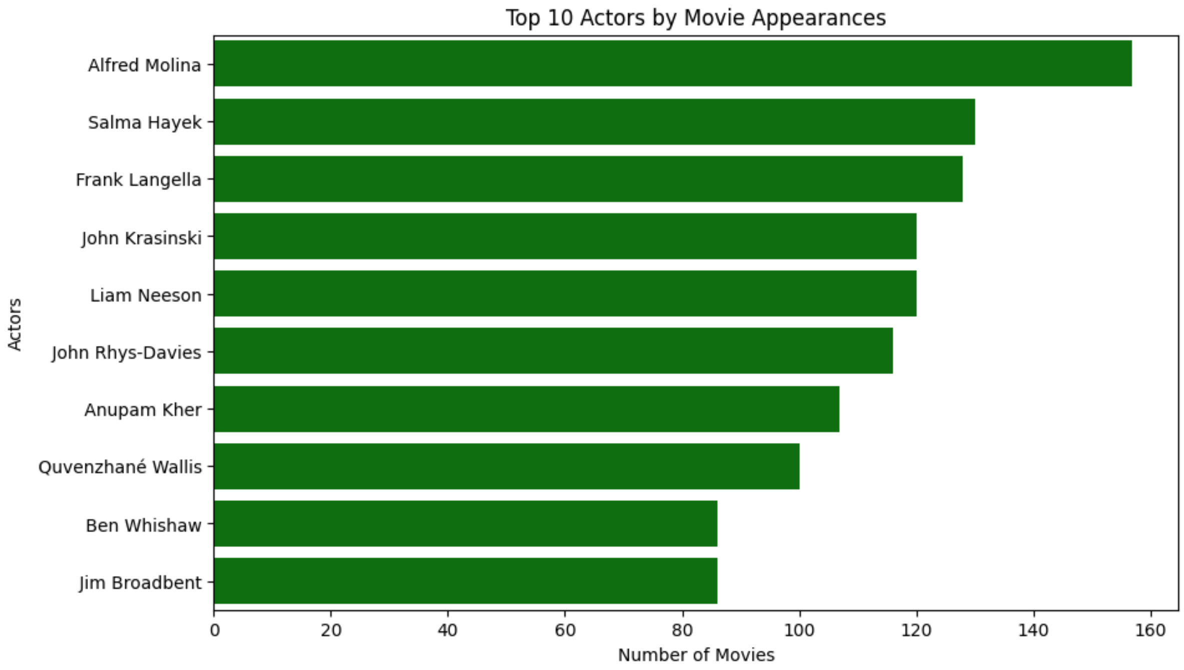
```
top_actors = df_final[df_final['type'] == 'Movie']['cast'].value_counts()[1:11].res
top_actors
```

Out[31]:

	cast	count
0	Alfred Molina	157
1	Salma Hayek	130
2	Frank Langella	128
3	John Krasinski	120
4	Liam Neeson	120
5	John Rhys-Davies	116
6	Anupam Kher	107
7	Quvenzhané Wallis	100
8	Ben Whishaw	86
9	Jim Broadbent	86

In [33]:

```
plt.figure(figsize=(10, 6))
sns.barplot(x='count', y='cast', data=top_actors, color='green')
plt.title('Top 10 Actors by Movie Appearances')
plt.xlabel('Number of Movies')
plt.ylabel('Actors')
plt.show()
```



In [34]:

```
df_new= df_final.loc[df_final['director']!='Unknown Director']
df_new.groupby('director')['title'].nunique().sort_values(ascending  = False)[0:11]
```

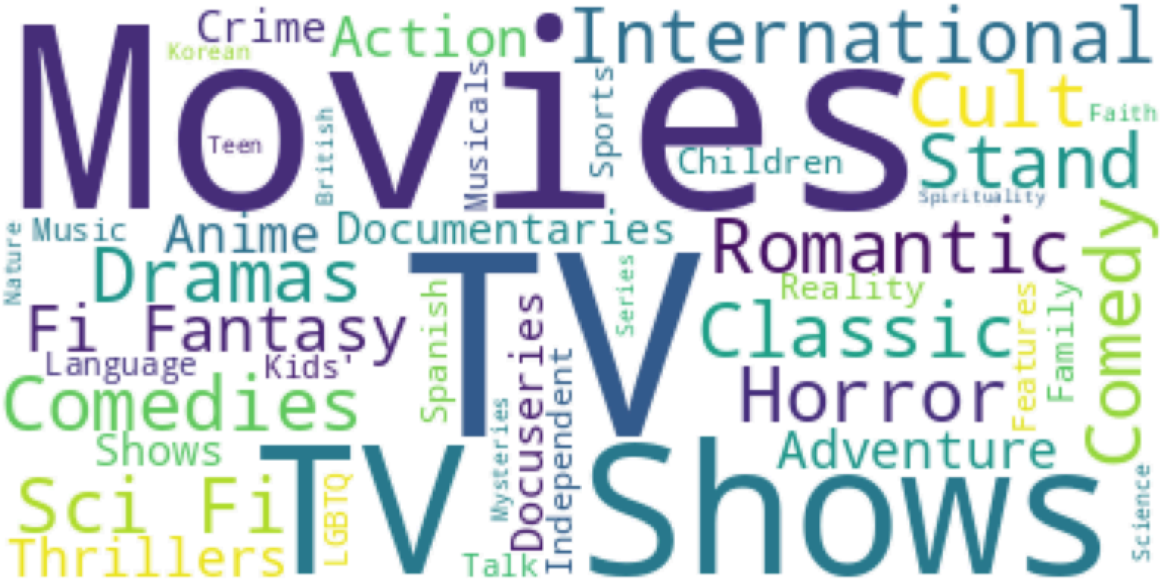
Out[34]:

	director	title
0	Rajiv Chilaka	22
1	Jan Suter	18
2	Raúl Campos	18
3	Suhas Kadav	16
4	Marcus Raboy	16
5	Jay Karas	15
6	Cathy Garcia-Molina	13
7	Jay Chapman	12
8	Martin Scorsese	12
9	Youssef Chahine	12
10	Steven Spielberg	11

5. Which genre movies are more popular or produced more

In [ ]:

```
a = df_final['listed_in'].unique()
from wordcloud import WordCloud
word = WordCloud(background_color  = 'white').generate(' '.join(a))
plt.figure(figsize = (10,8))
plt.imshow(word)
plt.axis("off")
plt.show()
```



Insight -

- The word cloud shows "Movies," "TV Shows," and "International" as dominant interests. For a balance of genres, try "Sci-Fi Thrillers" for excitement and "Comedies" for a lighter mood.

6. Find After how many days the movie will be added to Netflix after the release of the movie (you can consider the recent past data)

In [41]:

```
# calculating the difference of release year and date added
df_final['release_date'] = pd.to_datetime(df_final['release_year'].astype(str) + '-
df_final['date_added'] = pd.to_datetime(df_final['date_added'])
df_final['days_to_netflix'] = (df_final['date_added'] - df_final['release_date']).c
```

In [43]:

```
df_final.head()
```

Out[43]:

	show_id	type	title	date_added	release_year	rating	duration	description	director	
0	s1	Movie	Dick Johnson Is Dead	2021-09-25	2020	PG-13	90 min	As her father nears the end of his life, filmm...	Kirsten Johnson	Ur
1	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	C
2	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	C
3	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	C
4	s2	TV Show	Blood & Water	2021-09-24	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...	Unknown Director	

Insight -

- Negative days in this context occur when the date_added is earlier than the release_date. This means the movie was added to Netflix before it was officially released.

Recommendations --

- Continue producing and acquiring TV-MA content to cater to the strong demand among mature audiences.
- Invest in creating more TV-Y, TV-Y7, and G-rated content to attract and retain families and younger viewers.

- Continue producing and acquiring content in high-demand genres like dramas and international TV shows to maintain strong viewership.

- Invest in expanding content in underrepresented genres to attract diverse audience segments and fill gaps in the content library.

- Increase offerings in underrepresented genres like documentary, independent films, and foreign cinema to attract a broader audience.

- Given the popularity of true crime documentaries and mystery thrillers, prioritize production in these genres to maintain high viewer interest and retention.