Netflix Data Exploration Business Case

Importing the dependencies and downloading the dataset

In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
!gdown https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv
```

```
Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv
To: /content/netflix.csv
100% 3.40M/3.40M [00:00<00:00, 52.4MB/s]
```

Reading the dataset

In [3]:
```
df=pd.read_csv('netflix.csv')
df
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
...	...	...	...	...	...	...	...	...	...	...	...	...
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,	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
					Emma Stone, ...							
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

Basic Analysis

Basic Information

In [4]:
```
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 [ ]:
```
df.shape  #defines the shape of data frame
```

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

In [ ]:
```
df.head(10)    #displays top 10 rows
```

Out[ ]:

	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...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	NaN	September 24, 2021	2021	PG	91 min	Children & Family Movies	Equestria's divided. But a bright-eyed hero be...
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso,	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	On a photo shoot in Ghana, an American model s...

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
						United Kin...						
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A talented batch of amateur bakers face off in...
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min	Comedies, Dramas	A woman adjusting to life after a loss contend...

In [ ]:
```
df.nunique()    #gives count of unique values
```

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

In [122…
```
df.describe(include='all')   # gives the detailed interactive analysis like count,mean,min,max,etc.
```

```
<ipython-input-122-104ea69d96f3>:1: FutureWarning: Treating datetime data as categorical rather than numeric in `.describe` is d
eprecated and will be removed in a future version of pandas. Specify `datetime_is_numeric=True` to silence this warning and adop
t the future behavior now.
  df.describe(include='all')   # gives the detailed interactive analysis like count,mean,min,max,etc.
```

Out[122]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	duration_numeric
count	8797	8797	8797	8797	8797	8797	8797	8797.000000	8797	8797	8797	8797	8794.000000
unique	8797	2	8797	4529	7683	749	1714	NaN	18	221	513	8765	NaN
top	s1	Movie	Dick Johnson Is Dead	unknown director	unknown cast	United States	2020-01-01 00:00:00	NaN	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...	NaN
freq	1	6131	1	2624	825	2812	110	NaN	3205	1793	362	4	NaN
first	NaN	NaN	NaN	NaN	NaN	NaN	2008-01-01 00:00:00	NaN	NaN	NaN	NaN	NaN	NaN
last	NaN	NaN	NaN	NaN	NaN	NaN	2021-09-25 00:00:00	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014.183472	NaN	NaN	NaN	NaN	69.920173
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.822191	NaN	NaN	NaN	NaN	50.797005
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	NaN	NaN	NaN	NaN	1.000000
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	NaN	NaN	NaN	NaN	2.000000
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	NaN	NaN	NaN	NaN	88.000000
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	NaN	NaN	NaN	NaN	106.000000
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	NaN	NaN	NaN	NaN	312.000000

In [15]:
```
df.isnull().sum()   #  number of null values
```

Out[15]:
```
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
```

Percentage of null values

In [14]:
```
round((df.isnull().sum()/len(df))*100,2)
```

Out[14]:
```
show_id          0.00
type             0.00
title            0.00
director        29.91
cast             9.37
country          9.44
date_added       0.11
release_year     0.00
rating           0.05
duration         0.03
listed_in        0.00
description      0.00
dtype: float64
```

Percentage of missing data in director column is 29.91%, cast is 9.37%, country is 9.44%

Count of movies and TV shows directed by each director

In [24]:
```
df['director'].value_counts()
```

Out[24]:
```
Rajiv Chilaka                     19
Raúl Campos, Jan Suter            18
Marcus Raboy                      16
Suhas Kadav                       16
Jay Karas                         14
                                  ..
Raymie Muzquiz, Stu Livingston     1
Joe Menendez                       1
Eric Bross                         1
Will Eisenberg                     1
Mozez Singh                        1
Name: director, Length: 4528, dtype: int64
```

Removing the blank fields

In [21]:
```
df['director']=df['director'].fillna('unknown director')
df['rating']=df['rating'].fillna('unknown rating')
df['duration']=df['duration'].fillna('unknown duration')
df['country']=df['country'].fillna('unknown country')
df['cast']=df['cast'].fillna('unknown cast')
df.dropna(subset=['date_added'],inplace=True)
df.isnull().sum()
```

Out[21]:
```
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
```

Count of movies and TV shows directed by each director after removing empty fields

In [5]:
```
df[df['director']!='unknown director']['director'].value_counts()
```

Out[5]:
```
Rajiv Chilaka                     19
Raúl Campos, Jan Suter            18
Marcus Raboy                      16
Suhas Kadav                       16
Jay Karas                         14
                                  ..
Raymie Muzquiz, Stu Livingston     1
Joe Menendez                       1
Eric Bross                         1
Will Eisenberg                     1
Mozez Singh                        1
Name: director, Length: 4528, dtype: int64
```

Count of movies and shows released by year

In [21]:
```
by_year = df.groupby('release_year')['title'].count()
by_year
```

Out[21]:
```
release_year
1925       1
1942       2
1943       3
1944       3
1945       4
         ...
2017    1032
2018    1147
2019    1030
2020     953
2021     592
Name: title, Length: 74, dtype: int64
```

Count of Movies and TV Shows released

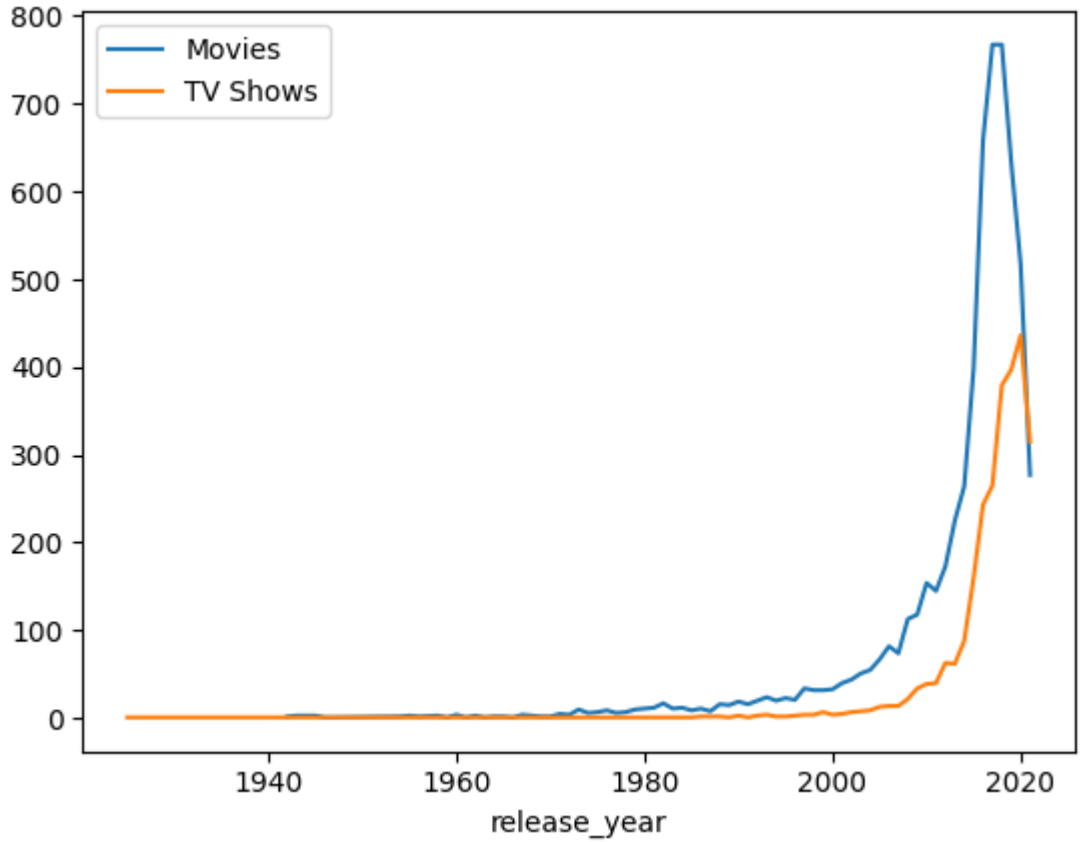
In [68]:
```
df['type'].value_counts()
```

Out[68]:
```
Movie      6131
TV Show    2666
Name: type, dtype: int64
```

Number of Movies/TV Shows by year

In [8]:
```
Movies_by_year = df[df['type'] == 'Movie'].groupby('release_year')['type'].count()
TVshows_by_year = df[df['type'] == 'TV Show'].groupby('release_year')['type'].count()
Movies_by_year.plot(label = 'Movies')
TVshows_by_year.plot(label = 'TV Shows')

plt.legend()
```

Out[8]:
```
<matplotlib.legend.Legend at 0x7893b07f8640>
```

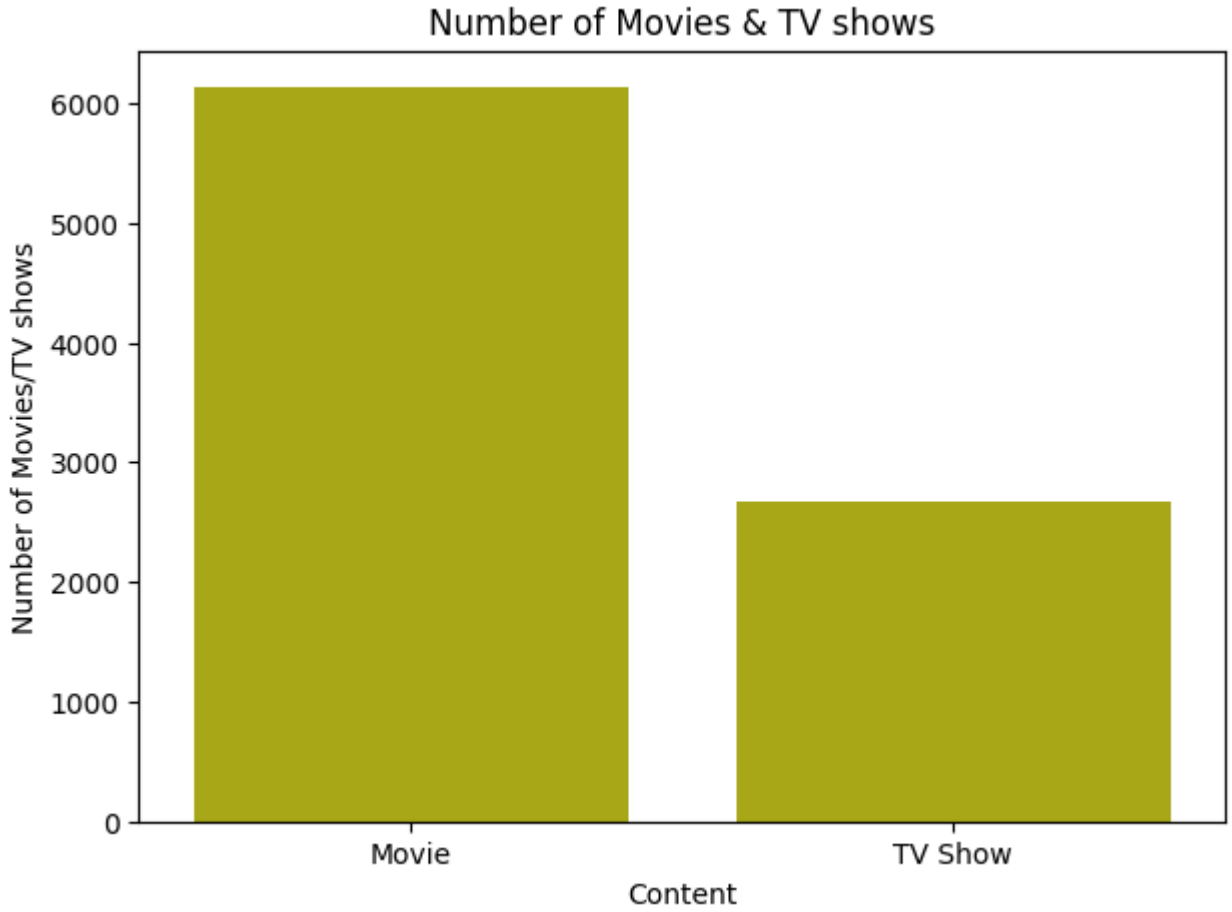


Number of Movies released is more as compared to TV Shows. Higher production is between 2000 and 2020

Count Plot

In [19]:
```
plt.figure(figsize=(7,5))
sns.countplot(data=df,x='type', color = 'y')
```

```
plt.title('Number of Movies & TV shows')
plt.xlabel('Content')
plt.ylabel('Number of Movies/TV shows')
plt.show()
```



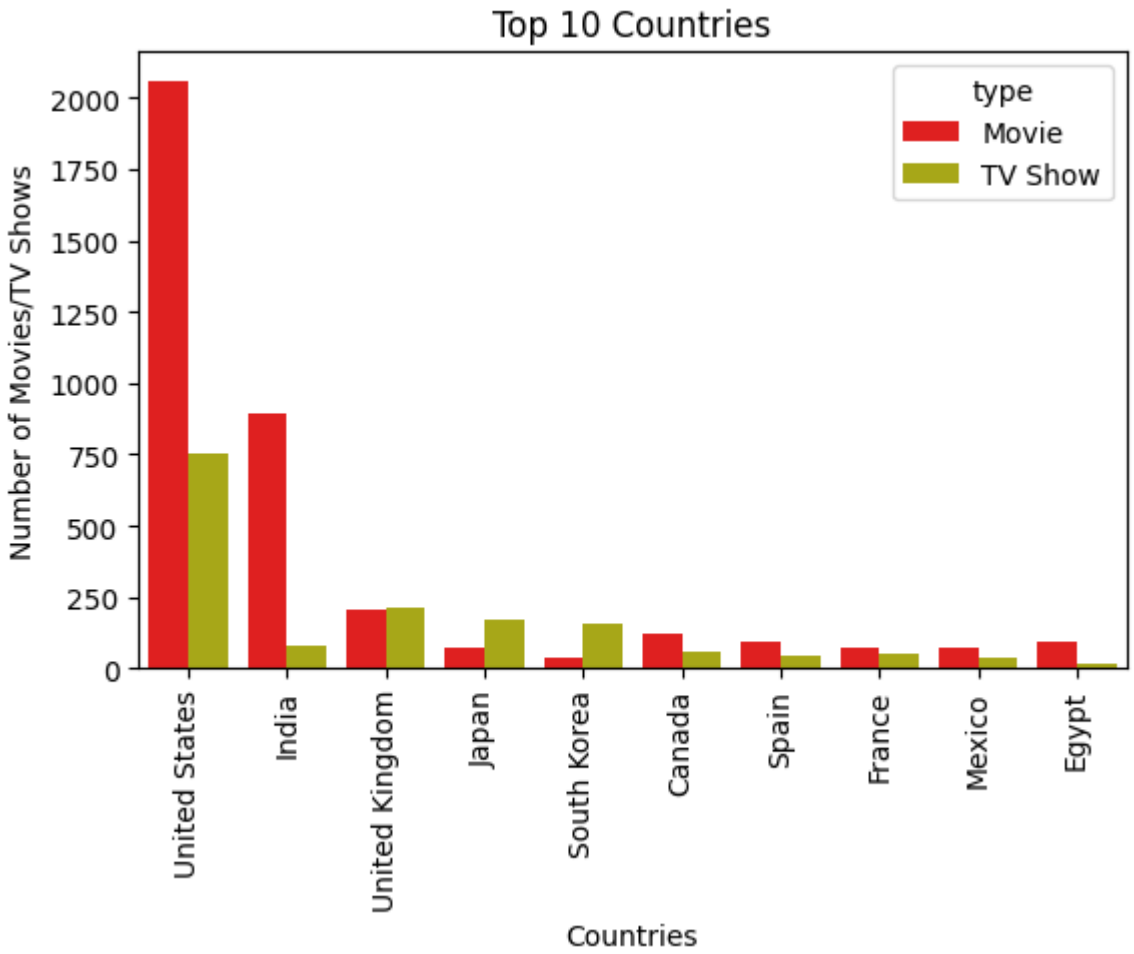
Top 10 countries for releasing movies

In [24]:
```
t_countries=df[df['country']!='unknown country']
top10countries = t_countries['country'].value_counts().iloc[:10]
top10countries
```

Out[24]:
```
United States     2812
India              972
United Kingdom     418
Japan              244
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Egypt              106
Name: country, dtype: int64
```

Count Plot for Top 10 Countries

In [38]:
```
plt.figure(figsize=(14,4))
plt.subplot(1,2,2)
sns.countplot(data=df[df['country'].isin(top10countries.index)],x='country',order=top10countries.index, hue='type',palette={'Movi
plt.xticks(rotation=90)
plt.xlabel('Countries')
plt.ylabel('Number of Movies/TV Shows')
plt.title('Top 10 Countries')
plt.show()
```

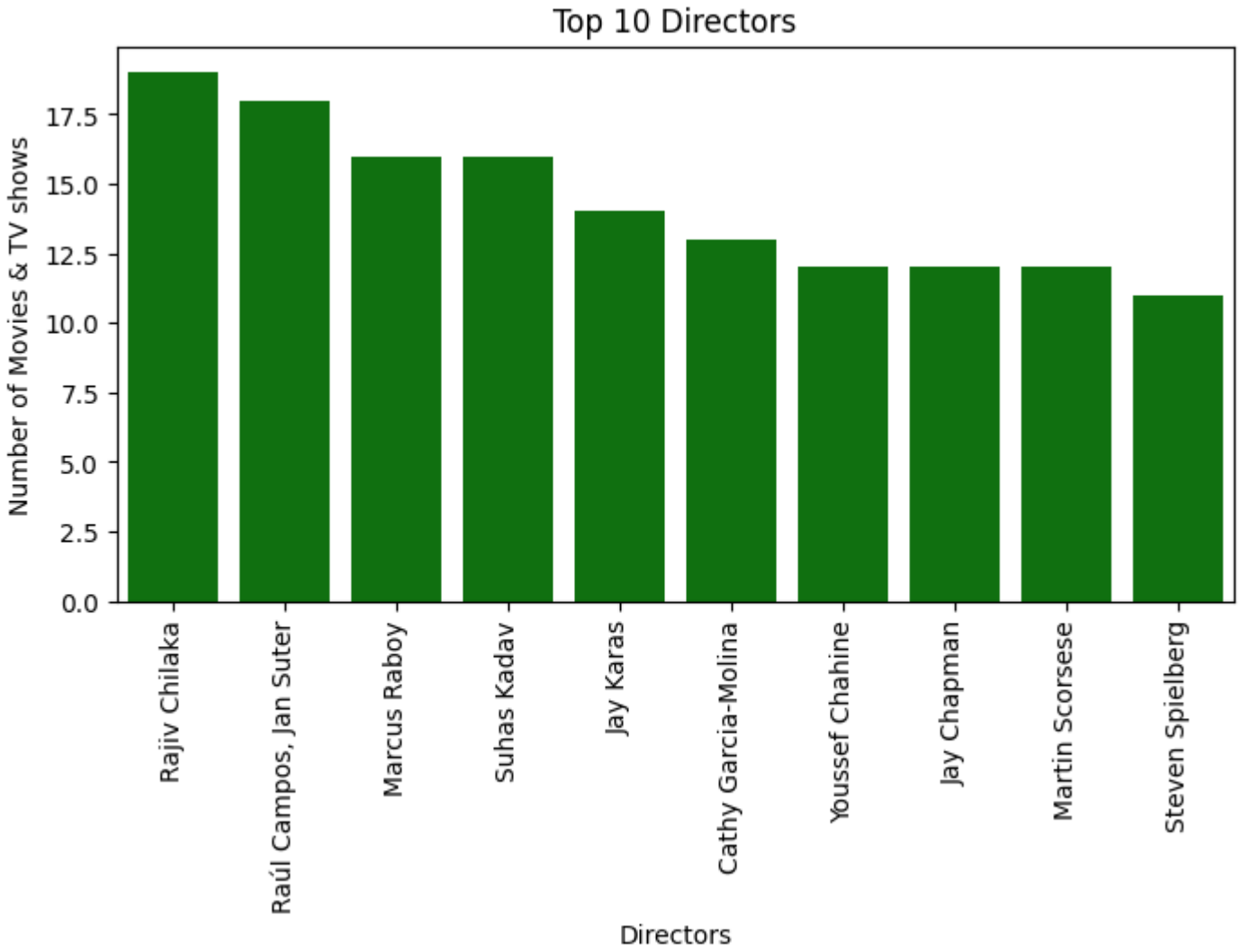


United States released highest number of Movies and TV Shows

Top 10 Directors

In [40]:
```
top10directors = df[df['director']!='unknown director']['director'].value_counts().sort_values(ascending = False).iloc[:10]
plt.figure(figsize=(8,4))
sns.countplot(data = df[df['director'].isin (top10directors.index)],x ='director',
order = top10directors.index, color = 'g')
plt.title('Top 10 Directors')
plt.xlabel('Directors')
plt.ylabel('Number of Movies & TV shows')
```

```
plt.xticks(rotation = 90)
plt.show()
```

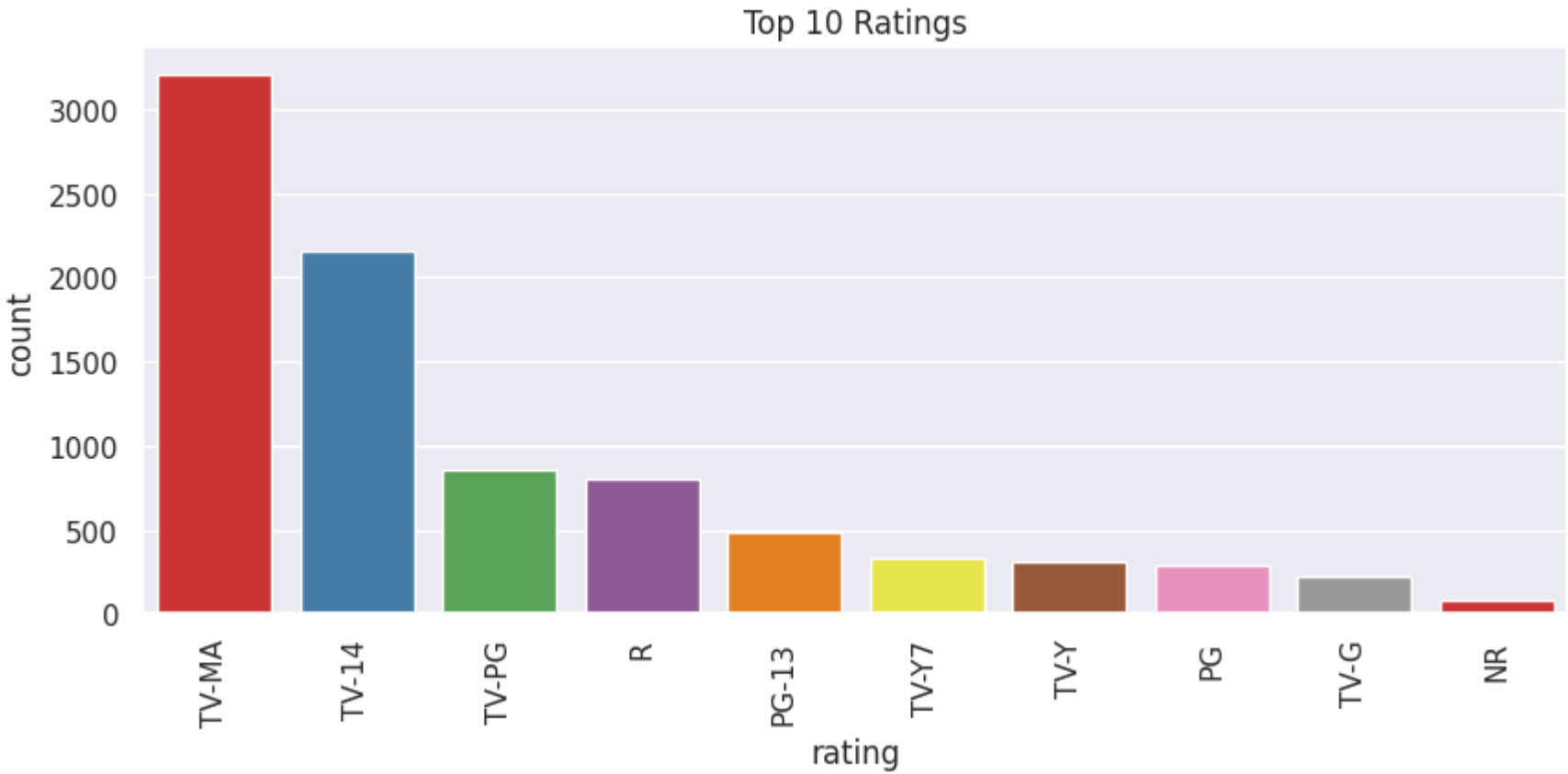


Rajiv Chilaka released highest number of movies and TV Shows

Most Popular Ratings

In [67]:
```
plt.figure(figsize=(10,4))
sns.set(style="darkgrid")
```

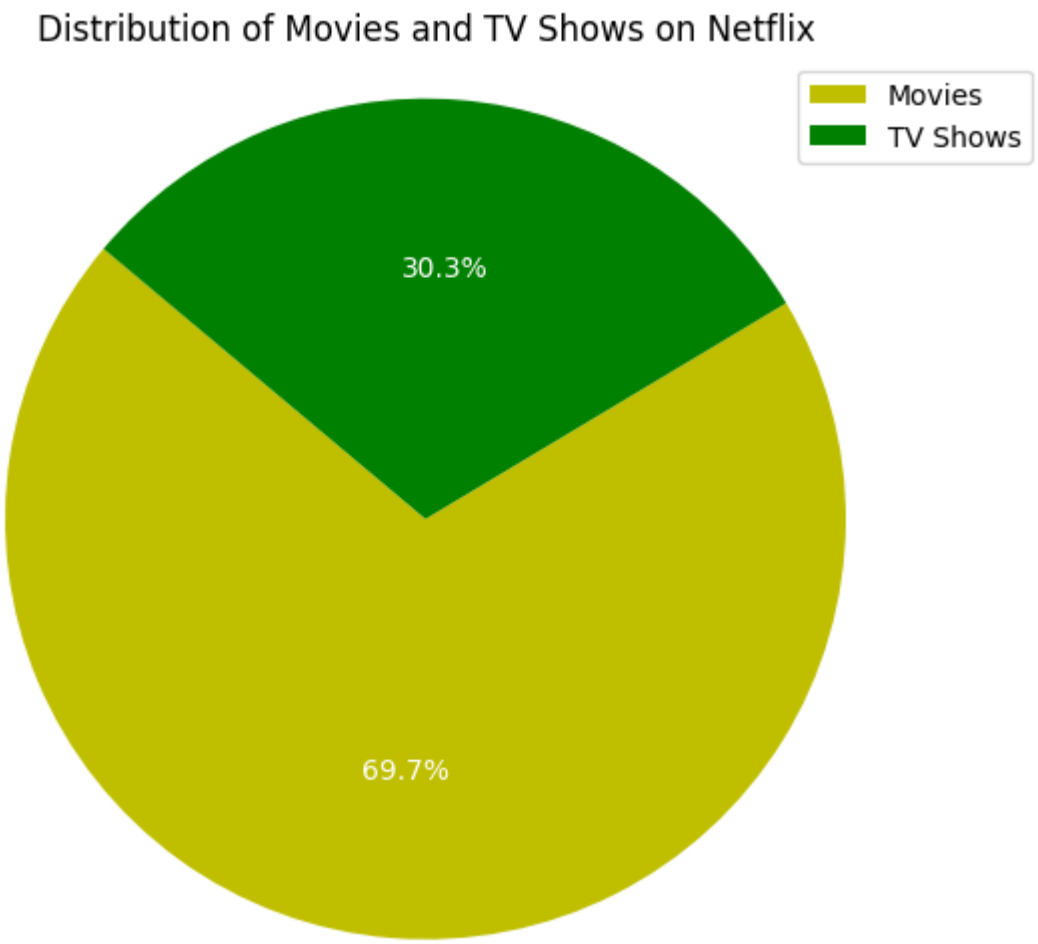
```
plt.title('Top 10 Ratings')
plt.xlabel('Rating')
plt.ylabel('Number of Movies & TV shows')
plt.xticks(rotation = 90)
ax = sns.countplot(x="rating", data=df, palette="Set1", order=df['rating'].value_counts().index[0:10])
```



Distribution of Movies and TV Shows on Netflix

In [55]:
```
unique_movies = df[df['type'] == 'Movie']['title'].nunique()
unique_tv_shows = df[df['type'] == 'TV Show']['title'].nunique()

labels = 'Movies', 'TV Shows'
sizes = [unique_movies, unique_tv_shows]
colors = ['y', 'g']

plt.figure(figsize=(8, 6))
```

```
plt.pie(sizes, labels=labels, colors=colors, autopct='%1.1f%%', startangle=140, textprops={'color':"white"})
plt.axis('equal')

plt.title('Distribution of Movies and TV Shows on Netflix')
plt.legend()
plt.show()
```



Number of Movies released is more as compared to TV Shows

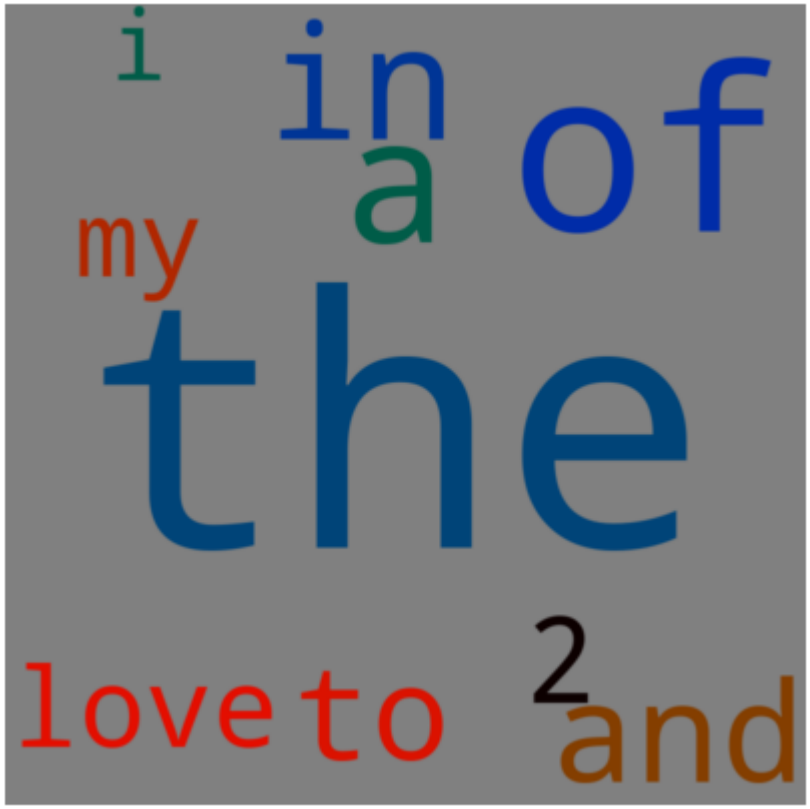
Top 10 words used in movie titles

In [73]:
```
from collections import Counter
import string

movie_titles = df[df['type'] == 'Movie']['title']
all_titles = ' '.join(movie_titles)
all_titles = all_titles.translate(str.maketrans('', '', string.punctuation)).lower()
words = all_titles.split()
word_counts = Counter(words)
top_10_words = word_counts.most_common(10)

top_10_words
```

Out[73]:
```
[('the', 1637),
 ('of', 505),
 ('a', 276),
 ('in', 222),
 ('and', 169),
 ('to', 148),
 ('2', 125),
 ('love', 105),
 ('my', 96),
 ('i', 80)]
```

Word Cloud

In [82]:
```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from matplotlib.colors import LinearSegmentedColormap

top_10_dict = dict(top_10_words)
colors = ['blue', 'green','red' , 'black']
custom_colormap = LinearSegmentedColormap.from_list("custom", colors)
wordcloud = WordCloud(width=800, height=800,
                      background_color='grey',
                      colormap=custom_colormap,
                      min_font_size=10).generate_from_frequencies(top_10_dict)
plt.figure(figsize=(10, 4), facecolor=None)
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.tight_layout(pad=0)
plt.show()
```



'the' is the highest used word in movie titles ie. 1637 times

Best week for releasing movies

In [118…

```
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
df['week_number'] = df['date_added'].dt.isocalendar().week

movies_data = df[df['type'] == 'Movie']
tv_shows_data = df[df['type'] == 'TV Show']

movies_weekly_count = movies_data.groupby('week_number')['title'].nunique().reset_index()
tv_shows_weekly_count = tv_shows_data.groupby('week_number')['title'].nunique().reset_index()

best_week_movies = movies_weekly_count.sort_values('title', ascending=False).iloc[0]
best_week_tv_shows = tv_shows_weekly_count.sort_values('title', ascending=False).iloc[0]
```

```
best_week_movies, best_week_tv_shows
```

Out[118]:
```
(week_number      1
 title          316
 Name: 0, dtype: Int64,
 week_number    27
 title          86
 Name: 26, dtype: Int64)
```

First week of month is best for releasing movies

Best Month for releasing movies

In [120…
```
tv_shows_data = df[df['type'] == 'TV Show'].copy()
movies_data = df[df['type'] == 'Movie'].copy()

movies_data['month'] = movies_data['date_added'].dt.month
tv_shows_data['month'] = tv_shows_data['date_added'].dt.month

movies_monthly_count = movies_data.groupby('month')['title'].nunique().reset_index()
tv_shows_monthly_count = tv_shows_data.groupby('month')['title'].nunique().reset_index()

best_month_movies = movies_monthly_count.sort_values('title', ascending=False).iloc[0]
best_month_tv_shows = tv_shows_monthly_count.sort_values('title', ascending=False).iloc[0]

best_month_movies, best_month_tv_shows
```

Out[120]:
```
(month      7
 title    565
 Name: 6, dtype: int64,
 month     12
 title    266
 Name: 11, dtype: int64)
```

7th month ie. July is best for releasing movies

Best day of week for releasing movies

In [121…
```
df['day_of_week_added'] = df['date_added'].dt.day_name()

shows_per_day_of_week = df.groupby('day_of_week_added')['show_id'].nunique()
most_popular_day = shows_per_day_of_week.idxmax()
```

```
most_popular_day_count = shows_per_day_of_week.max()

most_popular_day, most_popular_day_count, shows_per_day_of_week
```

Out[121]:
```
('Friday',
 2498,
 day_of_week_added
 Friday       2498
 Monday        851
 Saturday      816
 Sunday        751
 Thursday     1396
 Tuesday      1197
 Wednesday    1288
 Name: show_id, dtype: int64)
```

Friday has highest release of movies

Un-nesting

In [83]:
```
def unnest_dataframe(df, column):
    return (
        df.drop(column, axis=1)
        .join(
            df[column].str.split(',', expand=True)
            .stack()
            .reset_index(level=1, drop=True)
            .rename(column)
        )
    )

# Un-nesting the 'cast' column
unnested_cast = unnest_dataframe(df, 'cast')

# Un-nesting the 'country' column
unnested_country = unnest_dataframe(df, 'country')

# Un-nesting the 'listed_in' (genre) column
unnested_listed_in = unnest_dataframe(df, 'listed_in')

# Un-nesting the 'director' column
unnested_director = unnest_dataframe(df, 'director')
```

```
unnested_cast.head(), unnested_country.head(), unnested_listed_in.head(), unnested_director.head()
```

Out[83]:
```
(  show_id     type                 title          director        country  \
 0      s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  United States
 1      s2  TV Show         Blood & Water  unknown director   South Africa
 1      s2  TV Show         Blood & Water  unknown director   South Africa
 1      s2  TV Show         Blood & Water  unknown director   South Africa
 1      s2  TV Show         Blood & Water  unknown director   South Africa

           date_added  release_year rating   duration  \
 0  September 25, 2021          2020  PG-13     90 min
 1  September 24, 2021          2021  TV-MA  2 Seasons
 1  September 24, 2021          2021  TV-MA  2 Seasons
 1  September 24, 2021          2021  TV-MA  2 Seasons
 1  September 24, 2021          2021  TV-MA  2 Seasons

                                     listed_in  \
 0                               Documentaries
 1  International TV Shows, TV Dramas, TV Mysteries
 1  International TV Shows, TV Dramas, TV Mysteries
 1  International TV Shows, TV Dramas, TV Mysteries
 1  International TV Shows, TV Dramas, TV Mysteries

                                       description             cast
 0  As her father nears the end of his life, filmm...     unknown cast
 1  After crossing paths at a party, a Cape Town t...       Ama Qamata
 1  After crossing paths at a party, a Cape Town t...      Khosi Ngema
 1  After crossing paths at a party, a Cape Town t...    Gail Mabalane
 1  After crossing paths at a party, a Cape Town t...   Thabang Molaba  ,
   show_id     type                 title          director  \
 0      s1    Movie   Dick Johnson Is Dead   Kirsten Johnson
 1      s2  TV Show          Blood & Water  unknown director
 2      s3  TV Show              Ganglands   Julien Leclercq
 3      s4  TV Show  Jailbirds New Orleans  unknown director
 4      s5  TV Show           Kota Factory  unknown director

                                           cast          date_added  \
 0                                 unknown cast  September 25, 2021
 1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...  September 24, 2021
 2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...  September 24, 2021
 3                                 unknown cast  September 24, 2021
 4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...  September 24, 2021

    release_year rating   duration  \
 0          2020  PG-13     90 min
 1          2021  TV-MA  2 Seasons
```

```
2          2021  TV-MA   1 Season
3          2021  TV-MA   1 Season
4          2021  TV-MA  2 Seasons

                                      listed_in  \
0                                   Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                             Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...

                                        description          country
0  As her father nears the end of his life, filmm...    United States
1  After crossing paths at a party, a Cape Town t...     South Africa
2  To protect his family from a powerful drug lor...  unknown country
3  Feuds, flirtations and toilet talk go down amo...  unknown country
4  In a city of coaching centers known to train I...            India  ,
  show_id     type                 title          director  \
0      s1    Movie  Dick Johnson Is Dead   Kirsten Johnson
1      s2  TV Show         Blood & Water  unknown director
1      s2  TV Show         Blood & Water  unknown director
1      s2  TV Show         Blood & Water  unknown director
2      s3  TV Show             Ganglands   Julien Leclercq

                                           cast          country  \
0                                   unknown cast    United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...     South Africa
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...     South Africa
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...     South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...  unknown country

          date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
1  September 24, 2021          2021  TV-MA  2 Seasons
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season

                                        description              listed_in
0  As her father nears the end of his life, filmm...           Documentaries
1  After crossing paths at a party, a Cape Town t...  International TV Shows
1  After crossing paths at a party, a Cape Town t...               TV Dramas
1  After crossing paths at a party, a Cape Town t...            TV Mysteries
2  To protect his family from a powerful drug lor...          Crime TV Shows  ,
```

```
  show_id     type                  title  \
0      s1    Movie   Dick Johnson Is Dead
1      s2  TV Show          Blood & Water
2      s3  TV Show              Ganglands
3      s4  TV Show  Jailbirds New Orleans
4      s5  TV Show           Kota Factory

                                          cast          country  \
0                                  unknown cast    United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...     South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...  unknown country
3                                  unknown cast  unknown country
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...            India

         date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                    listed_in  \
0                                Documentaries
1  International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                          Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV ...

                                      description          director
0  As her father nears the end of his life, filmm...   Kirsten Johnson
1  After crossing paths at a party, a Cape Town t...  unknown director
2  To protect his family from a powerful drug lor...   Julien Leclercq
3  Feuds, flirtations and toilet talk go down amo...  unknown director
4  In a city of coaching centers known to train I...  unknown director  )
```

Average Movie Duration

In [84]:

```
movies_data = df[df['type'] == 'Movie'].copy()

movies_data['duration_min'] = movies_data['duration'].str.extract('(\d+)').astype(float)

mean_duration = movies_data['duration_min'].mean()
median_duration = movies_data['duration_min'].median()
```

```
mean_duration, median_duration
```

Out[84]:
```
(99.57718668407311, 98.0)
```

Average movie duration is 99 min

Top Directors in India

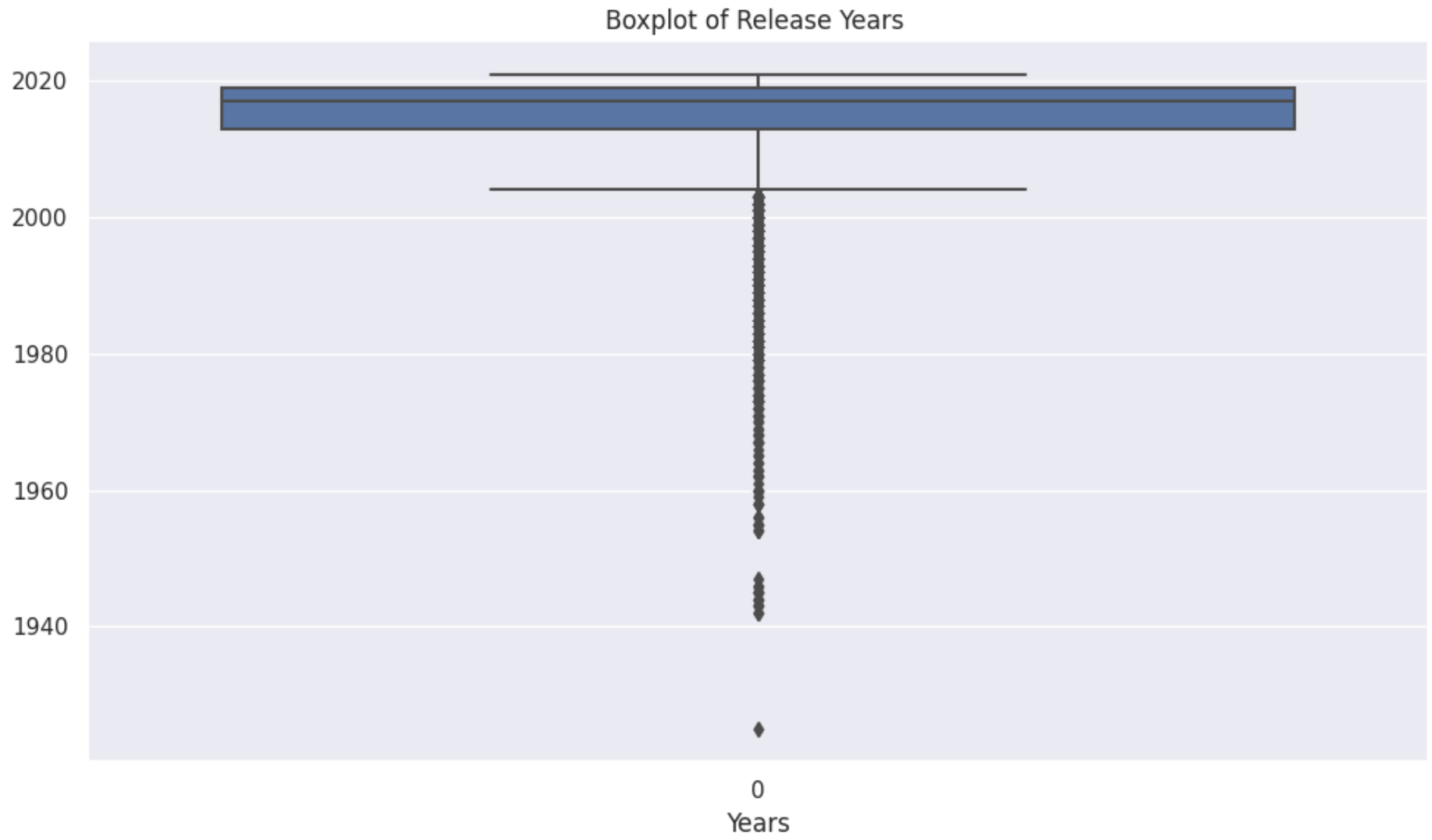
In [91]:
```
pop_dir = df.loc[(df['country'] == 'India') & (df['type'] == 'Movie')]
pop_dir = pop_dir[pop_dir['director'] != 'unknown director']
pop_dir.loc[: , 'director'].value_counts()
```

Out[91]:
```
David Dhawan          9
Ram Gopal Varma       7
Anees Bazmee          6
Sooraj R. Barjatya    6
Rajkumar Santoshi     6
                     ..
Manu Ashokan          1
Saurabh Sinha         1
Sunil Thakur          1
Rai Yuvraj Bains      1
Mozez Singh           1
Name: director, Length: 637, dtype: int64
```

David Dhawan is top most director in India

Boxplot of Release Years

In [123…
```
plt.figure(figsize=(10, 6))
sns.boxplot(df['release_year'])
plt.title('Boxplot of Release Years ')
plt.xlabel('Years')
plt.tight_layout()
plt.show()
```

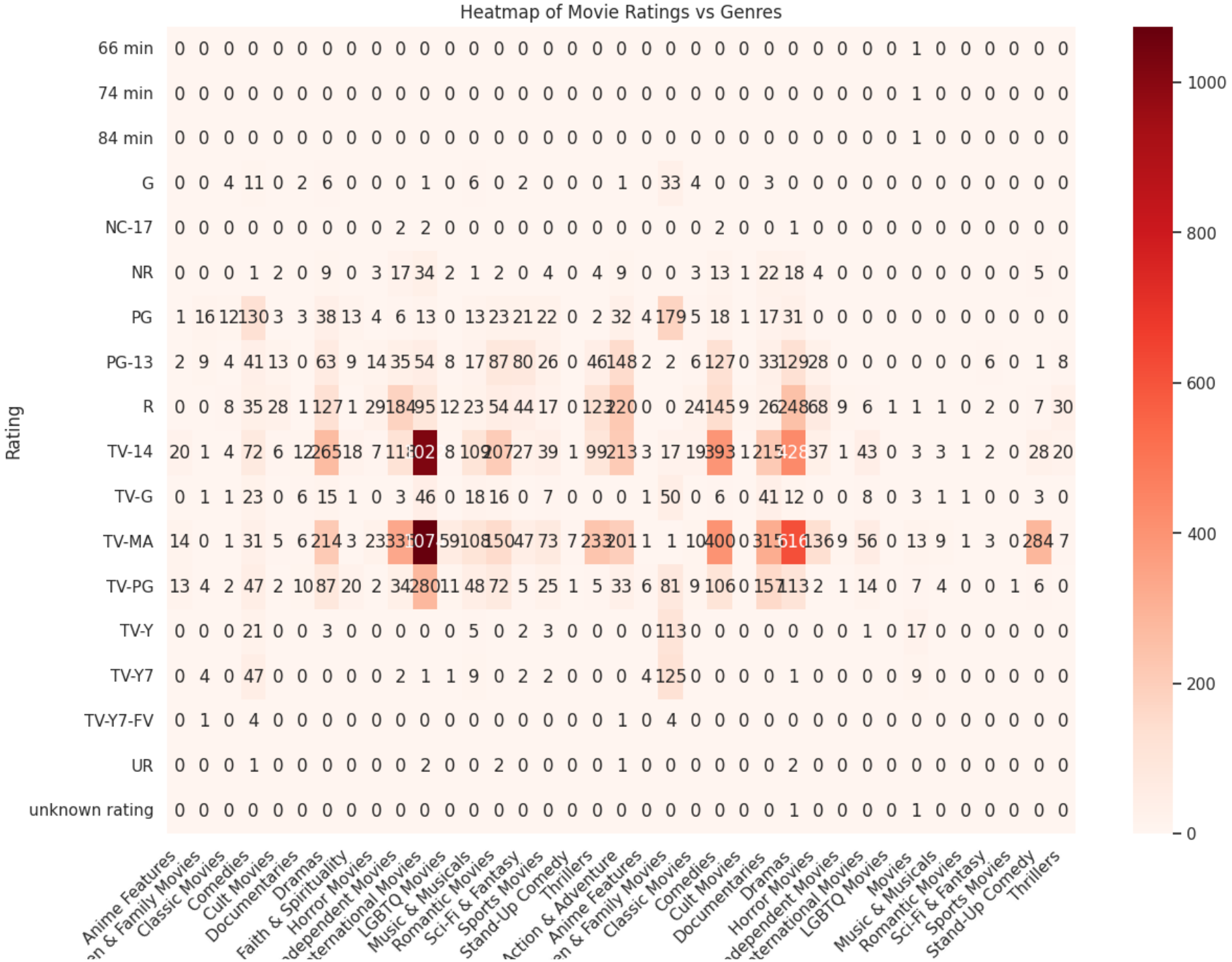


Heat Map

In [115…

```
movies_data['duration_numeric'] = movies_data['duration'].str.extract('(\d+)').astype(float)

genre_rating = unnest_dataframe(movies_data, 'listed_in').groupby(['rating', 'listed_in']).size().unstack().fillna(0)
duration_rating = movies_data.groupby('rating')['duration_numeric'].mean()

genre_rating, duration_rating
```

```
plt.figure(figsize=(14, 10))
sns.heatmap(genre_rating, cmap='Reds', annot=True, fmt=".0f")
plt.title('Heatmap of Movie Ratings vs Genres')
plt.xlabel('Genre')
plt.ylabel('Rating')
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.show()
```


Heatmap of Movie Ratings vs Genres
Rating
66 min
74 min
84 min
G
NC-17
NR
PG
PG-13
R
TV-14
TV-G
TV-MA
TV-PG
TV-Y
TV-Y7
TV-Y7-FV
UR
unknown rating
Anime Features
Classic Movies
Comedies
Cult Movies
Documentaries
Dramas
Faith & Spirituality
Horror Movies
Independent Movies
International Movies
LGBTQ Movies
Music & Musicals
Romantic Movies
Sci-Fi & Fantasy
Sports Movies
Stand-Up Comedy
Thrillers
Action & Adventure
Movies
1000
800
600
400
200
0

Recommendations:

Optimize Release Timing Strategically: Utilize time series analysis to refine the timing of new content releases on Netflix. Identifying seasonal patterns or specific periods when subscribers are most active can aid in planning strategic release schedules. Analysis indicates that Fridays are the prime day for releases, with the first week being ideal for Movies and the 27th week for TV Shows. Consider July as the optimal month for Movie releases and December for TV Shows.

Amplify Successful Genres in Key Ratings: If certain genres thrive in specific rating categories, contemplate expanding the production or acquisition of similar content to cater to the established audience. Notably, TV-MA & TV-14 in International Movies and TV-MA in Dramas form popular rating-genre combinations.

Tailor Rating Strategies by Genre: Examining movie duration by rating reveals distinct patterns, suggesting tailored strategies for content production. Utilize this insight to align the length of movies or shows with audience expectations for each rating. For example, NC-17 audiences prefer longer movies, averaging around 125 minutes, while TV-Y viewers prefer shorter durations suitable for children.

Diversify Genre Offerings: While international movies, dramas, and comedies enjoy popularity, consider diversifying into other genres such as documentaries, action, independent films, TV dramas, and romantic movies to cater to a broader audience.

Maintain a Balanced Mix of Movies and TV Shows: Acknowledge the preference for movies among most audience members and strive for equilibrium between movie and TV show content to cater to a diverse viewership.

Explore Popular Themes in Titles: Analyze the most frequently used words in movie and TV show titles to discern audience preferences. Awareness of trending words can inform marketing strategies and title creation, enhancing the appeal and discoverability of new content.

Adapt to External Factors: Stay agile and adaptive in content strategy, particularly in response to external influences like the COVID-19 pandemic, which may impact content production and consumption.

Refine Targeted Marketing and Recommendations: Tailor marketing strategies and personalized recommendations based on the popularity of specific genres, directors, or cast members. Directors like Rajiv Chilaka, Raúl Campos, and Jan Suter, along with cast members Anupam Kher, Shah Rukh Khan, Julie Tejwani, Takahiro Sakurai, Naseeruddin Shah, and Rupa Bhimani, are noteworthy influencers.

Strategic Content Planning and Development: Understand the average duration of content to plan and produce material aligned with audience expectations and viewing habits. Utilize this data to tailor recommendations, ensuring they resonate with users' preferences and engagement patterns.

Enhance Personalization Algorithms: With a growing library, continually improve content recommendation algorithms to consider not only viewers' past preferences but also potential interest in undiscovered older titles. This personalized approach enhances viewer engagement and satisfaction.

Welcome to Colab!

If you're already familiar with Colab, check out this video to learn about interactive tables, the executed code history view, and the command palette.

What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!