

NETFLIX

Netflix, a top streaming service, offers a wide range of TV shows, movies, and original content. Founded in 1997 by Reed Hastings and Marc Randolph in California, it started with DVD rentals online. In 2007, it shifted to streaming, changing how we watch entertainment. With its popular "Netflix Originals," it has become a global favourite, with millions of subscribers worldwide.

Business Case Problem

```
In [ ]: #Import Libraries
```

```
In [111... import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [ ]: #Loading Data
```

```
In [180... netflix_df = pd.read_csv('netflix1.csv')
```

```
In [185... netflix_df
```

Out[185]:

	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 l...
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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

Dataset is having 8807 rows of data with 12 attributes.

In [115... `#info`

In [186... `netflix_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 [187... `netflix_df.head()`

Out[187]:

	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 Mababane, 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 l...

In [188...

```
netflix_df.tail()
```

Out[188]:

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

In [189...

```
netflix_df.shape
```

Out[189]:

(8807, 12)

In [190...

```
netflix_df.ndim
```

Out[190]:

2

In [191...

```
netflix_df.dtypes
```

```
Out[191]: 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 [192... #Missing values
```

```
In [193... netflix_df.isnull().any()
```

```
Out[193]: show_id      False
          type        False
          title       False
          director    True
          cast        True
          country     True
          date_added  True
          release_year False
          rating      True
          duration    True
          listed_in   False
          description False
          dtype: bool
```

```
In [194... netflix_df.isnull().sum()
```

```
Out[194]: 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
```

The dataset contains missing values for the director, cast, country, rating, and duration. Among these, the director attribute has the most missing values.

```
In [195... #max and min year
```

```
In [196... netflix_df['release_year'].min(), netflix_df['release_year'].max()
```

```
Out[196]: (1925, 2021)
```

```
In [197... #Statistical summary
```

```
In [198... netflix_df.describe()
```

```
Out[198]:
```

	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

75% of the data corresponds to the years 2019-2021 and 25% of the data corresponds to the years 1925-2013.

```
In [199...
```

Non-Graphical Analysis

```
In [2]: #Counting the number of TV shows and movies
```

```
In [200... netflix_df['type'].value_counts()
```

```
Out[200]: type
Movie      6131
TV Show    2676
Name: count, dtype: int64
```

Counting the types, there are 6131 movies, compared to 2676 TV shows. This indicates a higher number of movies being made than TV shows.

```
In [3]: #Listed_in Count
```

```
In [201... netflix_df['listed_in'].value_counts()
```



```
Out[201]: listed_in
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: count, Length: 514, dtype: int64
```

```
In [4]: #Ratings Count
```

```
In [202... netflix_df['rating'].value_counts()
```

```
Out[202]: 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
NC-17        3
UR           3
74 min       1
84 min       1
66 min       1
Name: count, dtype: int64
```

```
In [203... #Unique values
```

```
In [204... netflix_df.nunique()
```

```
Out[204]: 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 [205... #Top Directors of Top 5 Countries
```

```
In [206... director_countrywise = netflix_df.groupby(['country', 'director']).size().reset_index(name='count')

# print the top directors for each country
country = director_countrywise['country'].value_counts()[:5].index.tolist() # Get top 5 countries
print('Top Directors of Top 5 Countries')
print('\n')
for val in country:
    if val != 'Unknown':
        print(f'***{val}***')
        top_directors = director_countrywise[(director_countrywise['country'] == val) & (director_countrywise['director'].notna())]
        top_directors = top_directors.sort_values(by='count', ascending=False)
        for idx, row in top_directors.iterrows():
            print(f"{row['country']: <20} {row['director']: <20} {row['count']: >5}")
        print('\n')
```

Top Directors of Top 5 Countries

United States

United States	Marcus Raboy	15
---------------	--------------	----

India

India	David Dhawan	9
-------	--------------	---

United Kingdom

United Kingdom	Edward Cotterill	4
----------------	------------------	---

Canada

Canada	Justin G. Dyck	6
--------	----------------	---

Spain

Spain	Alexis Morante	2
-------	----------------	---

Marcus Raboy stands out as the top director with 15 titles, reflecting a strong appeal to American audiences. David Dhawan closely follows as the second top director with 9 titles, contributing significantly to Bollywood's cinematic landscape.

In [207...] *#Top 5 Actors in Movies/TV shows*

```
In [208...] cast_df = netflix_df[['type', 'cast']].copy()
cast_df['cast'] = cast_df['cast'].apply(lambda x: x.split(', ') if isinstance(x, str) else [])
cast_df = cast_df.explode('cast')

# Group the data by type and count the occurrences of each actor
actor_counts_by_type = cast_df.groupby('type')['cast'].value_counts()

# top 5 actors for each type
for show_type in netflix_df['type'].unique():
    if show_type == 'Movie' or show_type == 'TV Show': # Filter out other types if present
        print(f"\nTop 5 Actors in {show_type}s:")
        top_actors = actor_counts_by_type[show_type].nlargest(5) # Get top 5 actors for the type
        print("Actor".ljust(25), "Appearance")
```

```
for idx, (actor, count) in enumerate(top_actors.items(), 1):
    print(f"{actor: <25} {count}")
```

Top 5 Actors in Movies:

Actor	Appearance
Anupam Kher	42
Shah Rukh Khan	35
Naseeruddin Shah	32
Om Puri	30
Akshay Kumar	30

Top 5 Actors in TV Shows:

Actor	Appearance
Takahiro Sakurai	25
Yuki Kaji	19
Junichi Suwabe	17
Ai Kayano	17
Daisuke Ono	17

Anupam Kher takes the top spot in Movies with 42 appearances and Takahiro Sakurai takes the top spot in TV shows with 25 appearances.

Cleanning the Data

In [210... *# Filling null values in the "country" column with mode country for each director*

```
In [211... director_mode_country = netflix_df.groupby('director')['country'].apply(lambda x: x.mode().iloc[0] if not x.dropna().empty else N
netflix_df['country'] = netflix_df['country'].fillna(netflix_df['director'].map(director_mode_country))
netflix_df
```

Out[211]:

	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...	France	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 l...
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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

Calculated the mode country for each director and stored this information in a dictionary. Then, filling the missing country values in the DataFrame by assigning the mode country corresponding to each director.

```
In [212... netflix_df['country'].isnull().sum()
```

Out[212]: 683

After filling the null values for country reduced to 683.

```
In [213... #Shifting the duration data to the rating columns
```

```
In [214... netflix_df[netflix_df['duration'].isnull()]
```

Out[214]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	April 4, 2017	2017	74 min	NaN	Movies	Louis C.K. muses on religion, eternal love, gi...
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	September 16, 2016	2010	84 min	NaN	Movies	Emmy-winning comedy writer Louis C.K. brings h...
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	August 15, 2016	2015	66 min	NaN	Movies	The comic puts his trademark hilarious/thought...

```
In [215]: null_duration_indices = netflix_df[netflix_df['duration'].isna()].index
netflix_df.loc[null_duration_indices] = netflix_df.loc[null_duration_indices].fillna(method='ffill', axis=1)
netflix_df.loc[null_duration_indices, 'rating'] = 'No Rating'
netflix_df.loc[null_duration_indices]
```

```
Out[215]:
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	April 4, 2017	2017	No Rating	74 min	Movies	Louis C.K. muses on religion, eternal love, gi...
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	September 16, 2016	2010	No Rating	84 min	Movies	Emmy-winning comedy writer Louis C.K. brings h...
5813	s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	August 15, 2016	2015	No Rating	66 min	Movies	The comic puts his trademark hilarious/thought...

3 missing data values were identified in the duration column, and it was found that these data had been accidentally placed into the rating column. So, relocated to the appropriate column.

```
In [216]: # Filling the remaing null values and converting date_added column to datetime type
```

```
In [217]: netflix_df.director.fillna("Director not specified", inplace=True)
netflix_df.cast.fillna("No Cast", inplace=True)
netflix_df.country.fillna("Country Unavailable", inplace=True)
netflix_df.date_added.fillna("Date Unavailable", inplace=True)
netflix_df.rating.fillna("No Rating", inplace=True)
netflix_df.dropna(subset=["date_added"], inplace = True)

# Converting date_added to datetime type
netflix_df['date_added'] = pd.to_datetime(netflix_df['date_added'], format="%B %d, %Y", errors='coerce')

# Extracting only the date part
netflix_df['date_added'] = netflix_df['date_added'].dt.date.astype(str)
netflix_df
```

Out[217]:

	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	No Cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	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...	France	2021-09-24	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	Director not specified	No Cast	Country Unavailable	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	Director not specified	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train l...
...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	Director not specified	No Cast	Country Unavailable	2019-07-01	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	United States	2019-11-01	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
					Harrelson, Emma Stone, ...							world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	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	2019-03-02	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

In [218... netflix_df.isnull().any()

Out[218]:

show_id	False
type	False
title	False
director	False
cast	False
country	False
date_added	False
release_year	False
rating	False
duration	False
listed_in	False
description	False
dtype:	bool

In [219... netflix_df.isnull().sum()

```
Out[219]: 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
```

Now no null values are found

Unnesting the data

```
In [221... netflix_df['cast'] = netflix_df['cast'].str.split(',')
netflix_df = netflix_df.explode('cast')
# Unnesting the 'Director' column
netflix_df['director'] = netflix_df['director'].str.split(',')
netflix_df = netflix_df.explode('director')
# Unnesting the 'Country' column
netflix_df['country'] = netflix_df['country'].str.split(',')
netflix_df = netflix_df.explode('country')
netflix_df['listed_in'] = netflix_df['listed_in'].str.split(',')
netflix_df = netflix_df.explode('listed_in')
# Reset index
netflix_df.reset_index(drop=True, inplace=True)
netflix_df
```

Out[221]:

	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	No Cast	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	Director not specified	Ama Qamata	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows	After crossing paths at a party, a Cape Town t...
2	s2	TV Show	Blood & Water	Director not specified	Ama Qamata	South Africa	2021-09-24	2021	TV-MA	2 Seasons	TV Dramas	After crossing paths at a party, a Cape Town t...
3	s2	TV Show	Blood & Water	Director not specified	Ama Qamata	South Africa	2021-09-24	2021	TV-MA	2 Seasons	TV Mysteries	After crossing paths at a party, a Cape Town t...
4	s2	TV Show	Blood & Water	Director not specified	Khosi Ngema	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows	After crossing paths at a party, a Cape Town t...
...
202336	s8807	Movie	Zubaan	Mozez Singh	Anita Shabdish	India	2019-03-02	2015	TV-14	111 min	International Movies	A scrappy but poor boy worms his way into a ty...
202337	s8807	Movie	Zubaan	Mozez Singh	Anita Shabdish	India	2019-03-02	2015	TV-14	111 min	Music & Musicals	A scrappy but poor boy worms his way into a ty...
202338	s8807	Movie	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	2019-03-02	2015	TV-14	111 min	Dramas	A scrappy but poor boy worms his way into a ty...
202339	s8807	Movie	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	2019-03-02	2015	TV-14	111 min	International Movies	A scrappy but poor boy worms his way into a ty...
202340	s8807	Movie	Zubaan	Mozez Singh	Chittaranjan Tripathy	India	2019-03-02	2015	TV-14	111 min	Music & Musicals	A scrappy but poor boy worms his way into a ty...

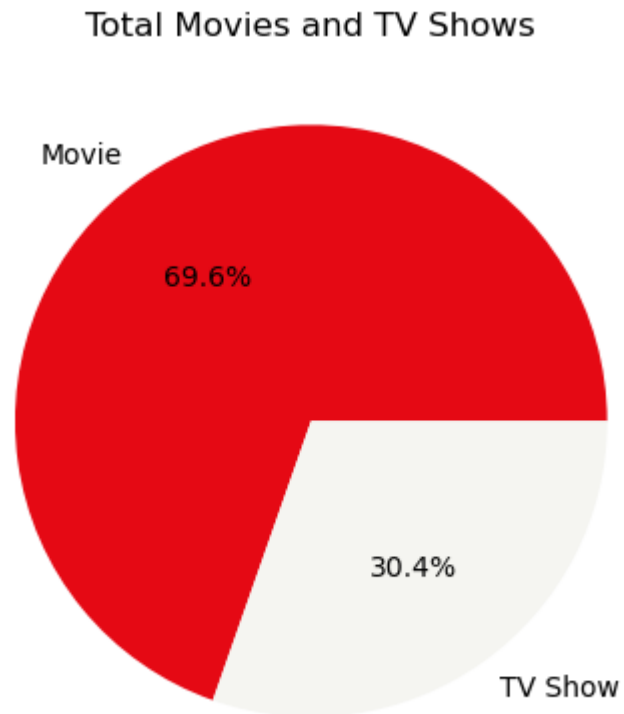
202341 rows × 12 columns

Splitting the values in the 'cast', 'director', 'country' and 'listed_in' columns by comma and space, then stacks them into separate rows, making the data more accessible and organized.

Visual Analysis - Univariate and Bivariate

```
In [153... #Piechart  
#Distribution of content across the different types
```

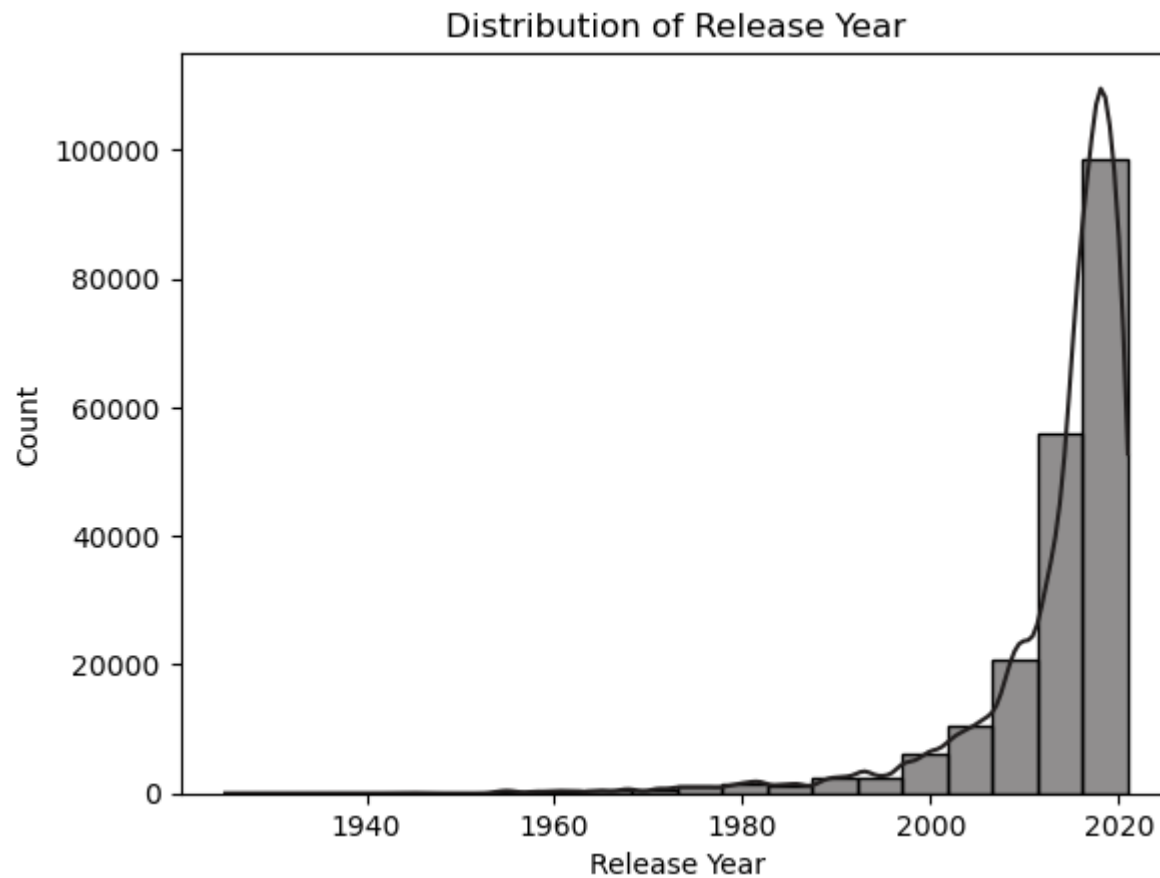
```
In [184... types = netflix_df.type.value_counts()  
netflix_colors = [ '#e50914', '#f5f5f1', '#221f1f', '#b20710', ]  
plt.pie(types, labels=types.index, autopct='%1.1f%%', colors=netflix_colors)  
plt.title('Total Movies and TV Shows')  
plt.show()
```



Around 70% of the content is Movies, while around only 30% of the content is TV shows.

```
In [155... #histplot  
#Distribution of Release Year
```

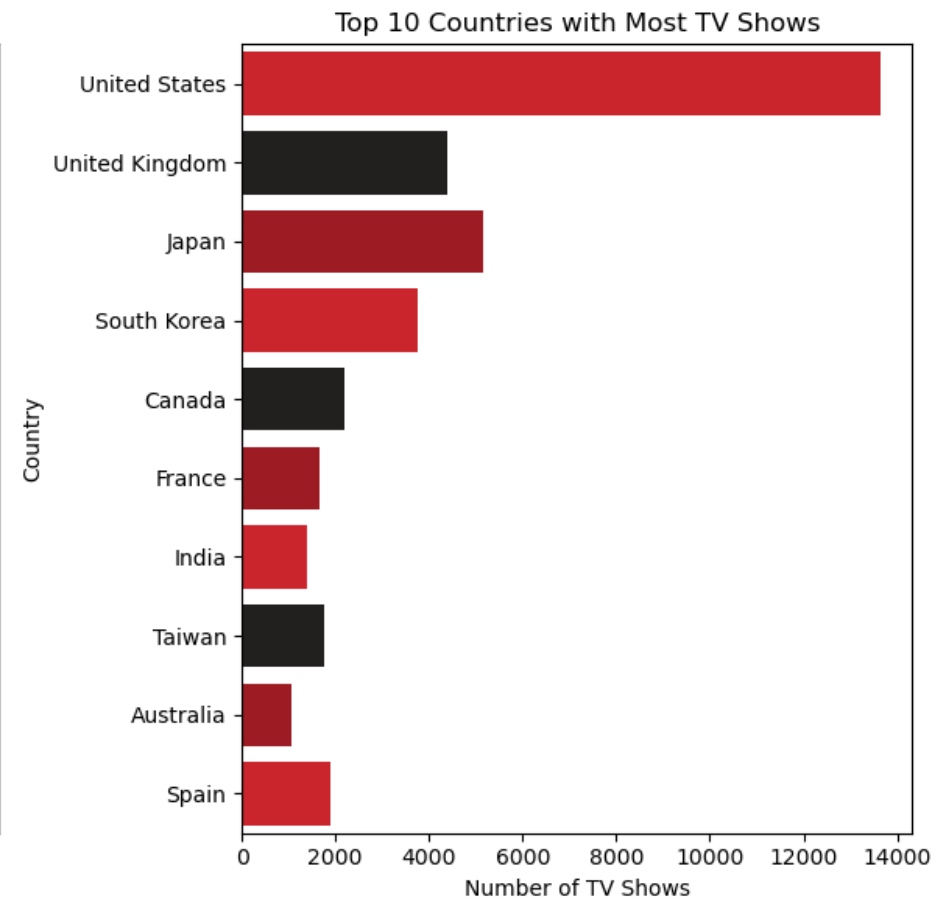
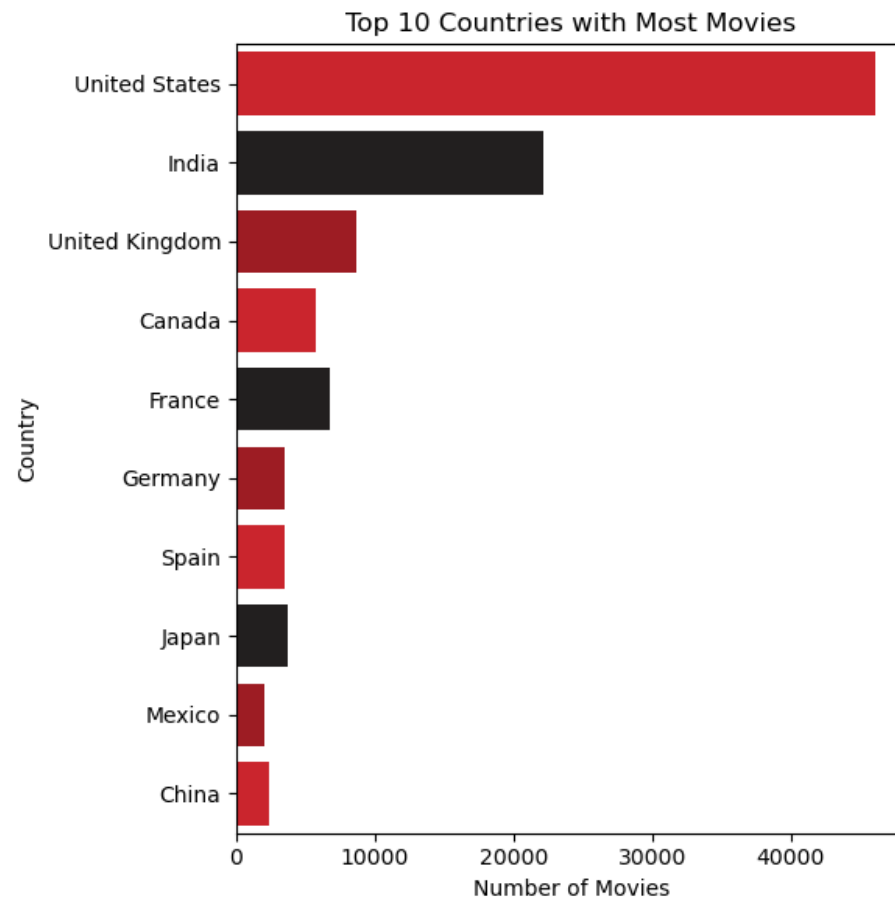
```
In [222... netflix_colors = ['#e50914', '#f5f5f1', '#221f1f', '#b20710']  
sns.histplot(netflix_df['release_year'], bins=20, kde=True, color=netflix_colors[2])  
plt.title('Distribution of Release Year')  
plt.xlabel('Release Year')  
plt.ylabel('Count')  
plt.show()
```



The data shows an increasing trend in the number of Movie and TV show releases since the year 2000.

```
In [157... #countplot  
#Top 10 Countries with most Movies and TV shows
```

```
In [223... netflix_df = netflix_df[netflix_df['country'] != 'Country Unavailable']  
  
movie_df = netflix_df[netflix_df['type'] == 'Movie']  
tv_show_df = netflix_df[netflix_df['type'] == 'TV Show']  
  
movie_count_country = movie_df.groupby('country')['title'].nunique().reset_index().sort_values(by='title', ascending=False).head(10)  
tv_show_count_country = tv_show_df.groupby('country')['title'].nunique().reset_index().sort_values(by='title', ascending=False).head(10)  
  
netflix_colors = ['#e50914', '#221f1f', '#b20710']  
  
plt.figure(figsize=(12, 6))  
  
# Countplot for movies  
plt.subplot(1, 2, 1)  
sns.countplot(y='country', data=movie_df, order=movie_count_country['country'], palette=netflix_colors)  
plt.title('Top 10 Countries with Most Movies')  
plt.xlabel('Number of Movies')  
plt.ylabel('Country')  
  
# Countplot for TV shows  
plt.subplot(1, 2, 2)  
sns.countplot(y='country', data=tv_show_df, order=tv_show_count_country['country'], palette=netflix_colors)  
plt.title('Top 10 Countries with Most TV Shows')  
plt.xlabel('Number of TV Shows')  
plt.ylabel('Country')  
  
plt.tight_layout()  
plt.show()
```



The data shows that the United States has the largest number of movies, followed by India. Similarly, for the TV shows, the United States holds the top position.

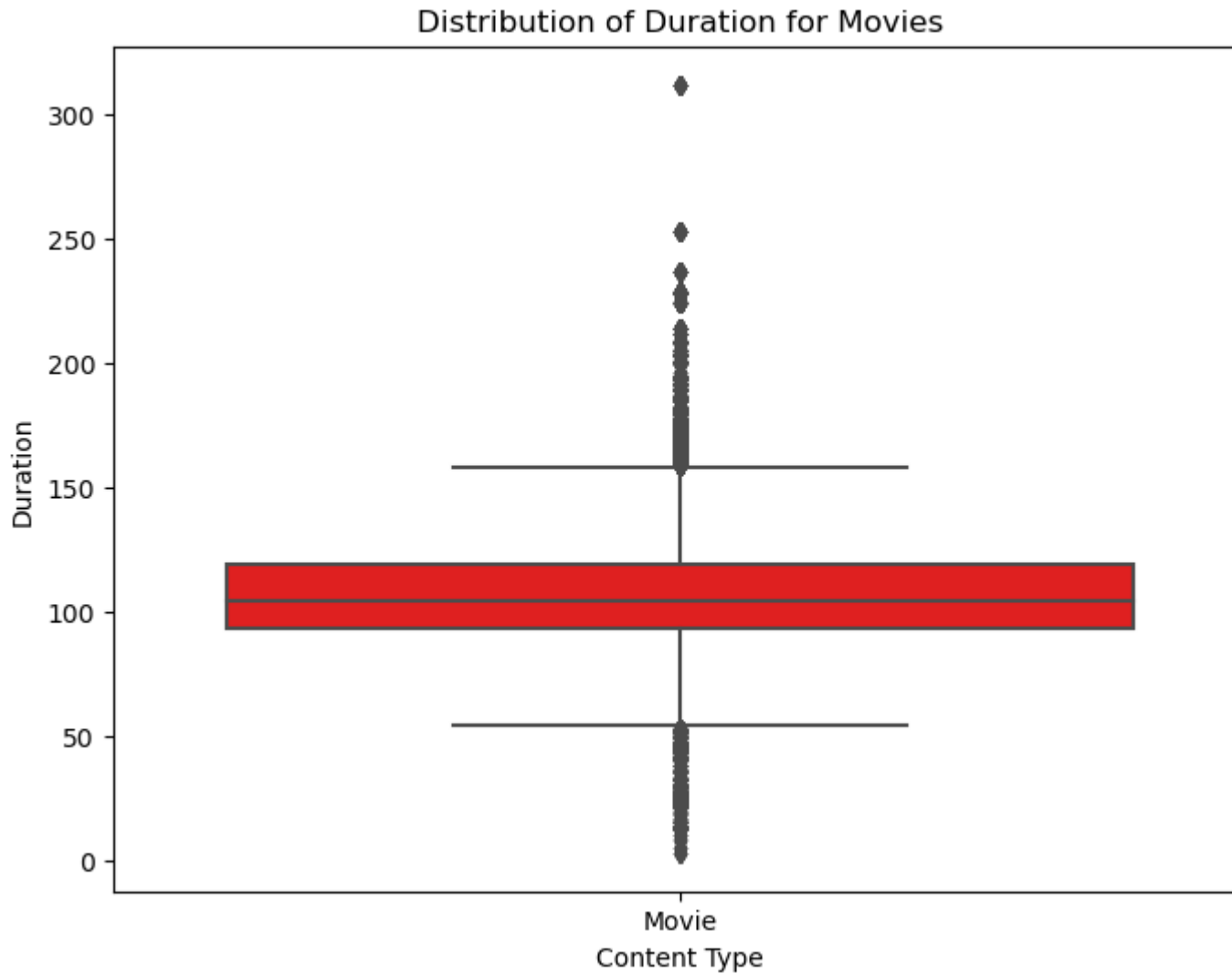
```
In [159... #boxplot
#Distribution of Duration for Movies and TV shows
```

```
In [224... movie_df = netflix_df[netflix_df.type.str.contains("Movie")].copy()

# Convert duration column to string before extracting numbers
movie_df['duration'] = movie_df['duration'].astype(str)
movie_df['duration'] = movie_df['duration'].str.extract('(\d+)', expand=False).astype(float)

# boxplot for movie duration
```

```
plt.figure(figsize=(8, 6))  
sns.boxplot(data=movie_df, x='type', y='duration', color='red')  
plt.xlabel('Content Type')  
plt.ylabel('Duration')  
plt.title('Distribution of Duration for Movies')  
plt.show()
```



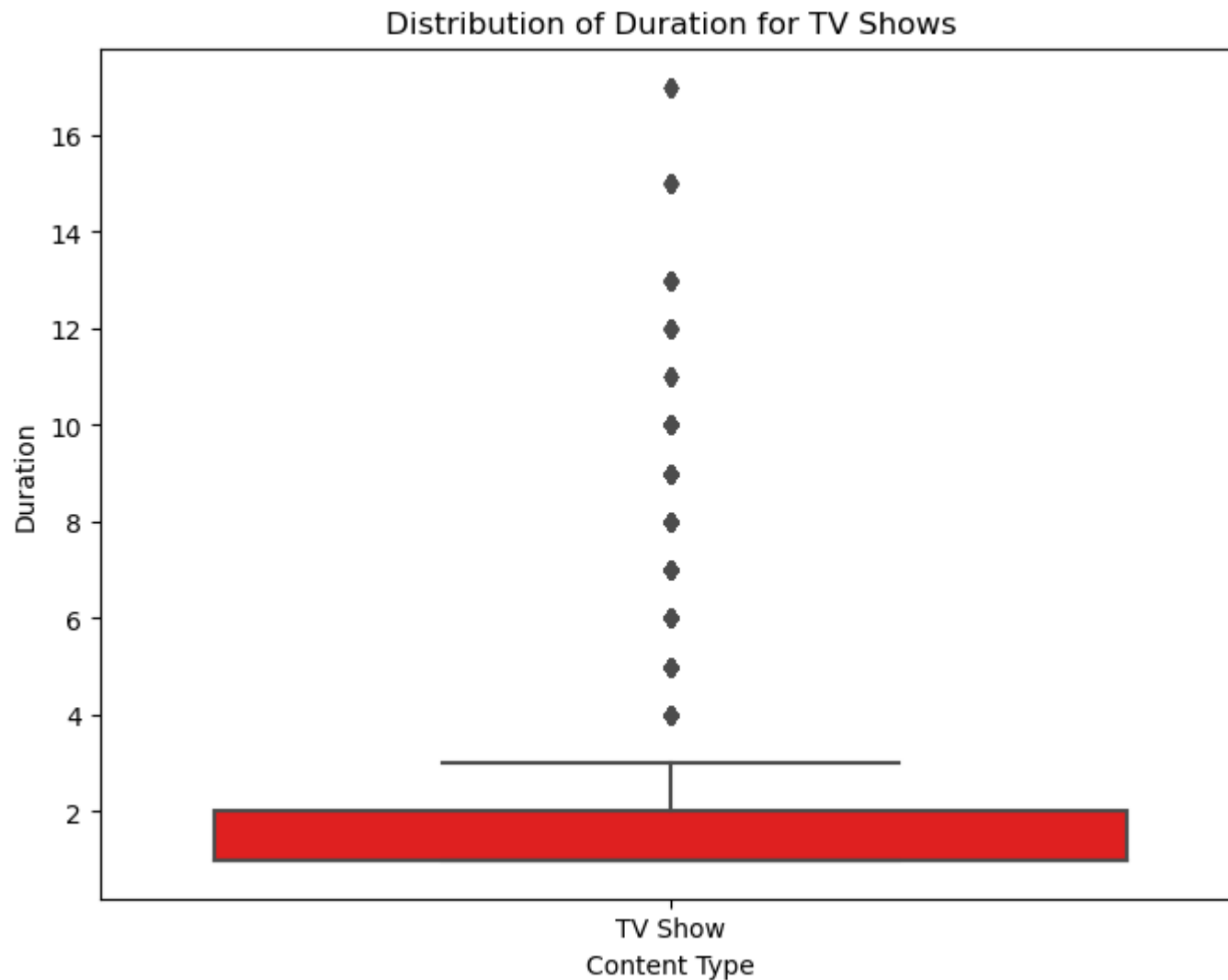
Most movies have durations around 100 min, this indicates that most Netflix movies adhere to a standard viewing time.

In [225...

```
tv_shows_df = netflix_df[netflix_df.type.str.contains("TV Show")].copy()

# Extracting numerical part from duration
tv_shows_df.loc[:, 'duration'] = tv_shows_df['duration'].str.extract('(\d+)', expand=False)
tv_shows_df['duration'] = tv_shows_df['duration'].astype(float) # Convert to float to handle potential NaN values

# boxplot for TV show duration
plt.figure(figsize=(8, 6))
sns.boxplot(data=tv_shows_df, x='type', y='duration', color='red')
plt.xlabel('Content Type')
plt.ylabel('Duration')
plt.title('Distribution of Duration for TV Shows')
plt.show()
```



For TV shows, the box plot indicates that the majority have between 1 to 2 seasons, with only a small number of outliers having longer durations. This confirms previous observations, suggesting that Netflix prioritizes shorter series formats.

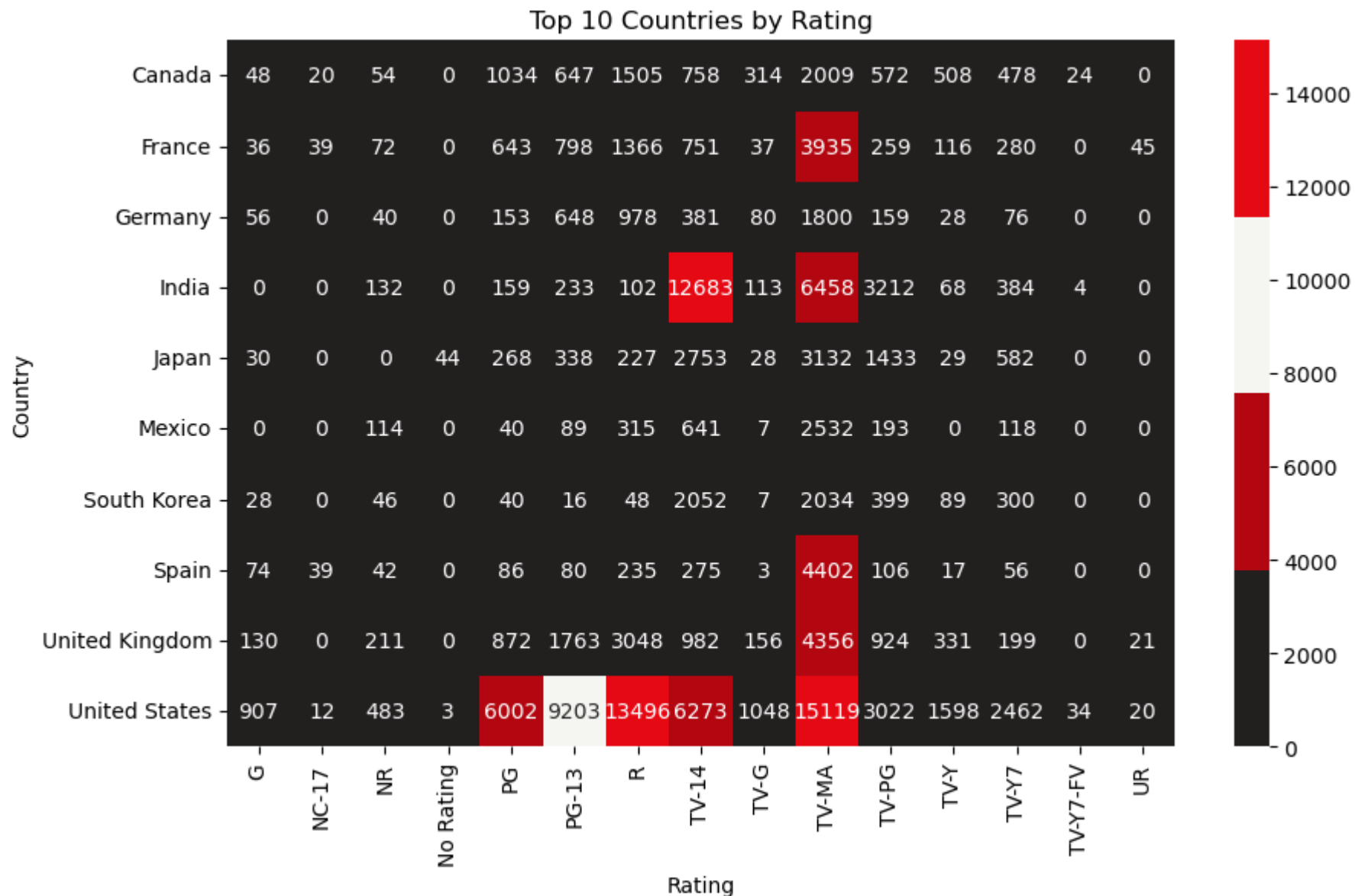
```
In [162... #heatmap  
##Distribution of Content Available in Top 10 Countries by Rating
```

```
In [235... top_countries = netflix_df['country'].value_counts().head(10).index.tolist()
```

```
# Filter the DataFrame to include only the top 10 countries
netflix_top_countries = netflix_df[netflix_df['country'].isin(top_countries)]
netflix_top_countries = netflix_top_countries[netflix_top_countries['country'] != 'Country Unavailable']
content_by_country_rating = netflix_top_countries.pivot_table(index='country', columns='rating', values='title', aggfunc='count',

netflix_colors = [ '#221f1f', '#b20710', '#f5f5f1', '#e50914' ]

# Plot heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(content_by_country_rating, cmap=netflix_colors, annot=True, fmt='d')
plt.title('Top 10 Countries by Rating')
plt.xlabel('Rating')
plt.ylabel('Country')
plt.show()
```

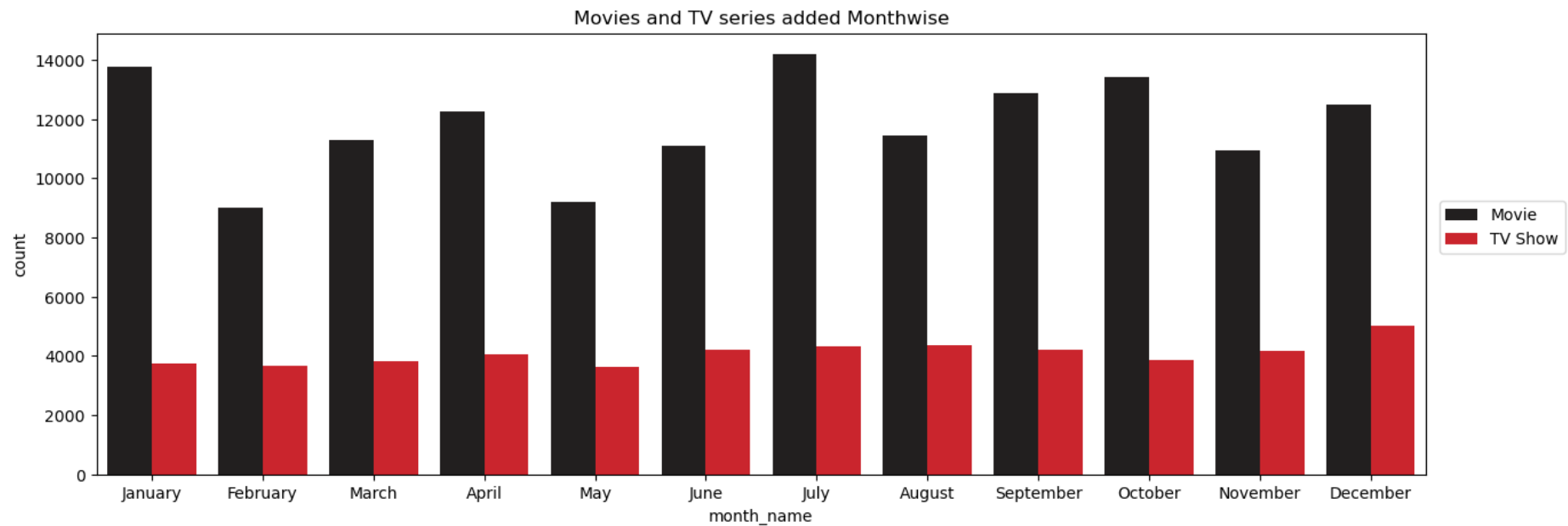


- TV-MA (TV Mature Audience) ratings dominate the content in the top 10 countries.
- India and the US have a higher number of TV-14 ratings.

```
In [164... #Monthly Distribution of TV Shows and Movies added in netflix
```

```
In [239... datetime_df['date_added'] = pd.to_datetime(datetime_df['date_added'])  
datetime_df['month'] = datetime_df['date_added'].dt.month  
month_df = datetime_df.sort_values(by="month")  
month_df['month_name'] = datetime_df['date_added'].dt.month_name()
```

```
In [240... netflix_colors = ['#221f1f', '#e50914']  
plt.figure(figsize=(15,5))  
sns.countplot(x = "month_name" , data = month_df , hue = "type", palette=netflix_colors)  
plt.title("Movies and TV series added Monthwise") #title name of the plot  
plt.legend(loc=(1.01,0.5))  
plt.show()
```



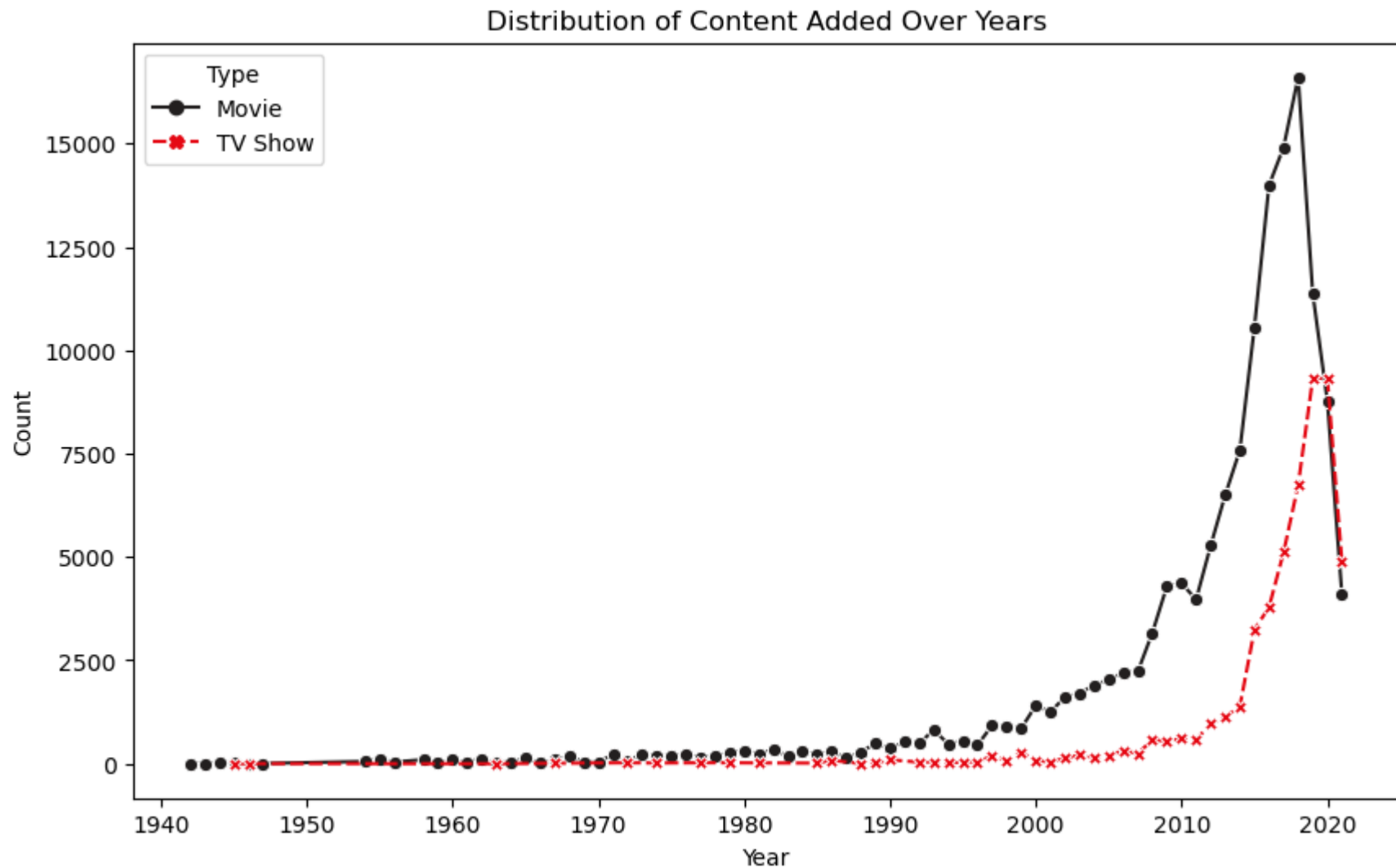
- July, April, and December have the highest number of movie releases, while February has the fewest.
- July and December boast the highest number of TV show releases, with February having the fewest.
- The number of movies added per month surpasses the number of TV shows added per month.

In [167... *#Distribution of Content Added Over Years*

```
In [241... # Grouping content by year for movies and TV shows
df_yearwise_trend = pd.DataFrame(netflix_df.groupby("release_year")["type"].value_counts())
df_yearwise_trend.reset_index(inplace=True)
df_yearwise_trend.columns = ['release_year', 'type', 'count']

df_content_count = df_yearwise_trend.pivot(index="release_year", columns="type", values="count")
netflix_colors = ['#221f1f', '#e50914']

# Plotting
plt.figure(figsize=(10, 6))
sns.lineplot(data=df_content_count, markers=True, palette=netflix_colors)
plt.xticks(np.arange(1940, 2025, 10))
plt.title("Distribution of Content Added Over Years")
plt.xlabel("Year")
plt.ylabel("Count")
plt.legend(title="Type")
plt.show()
```

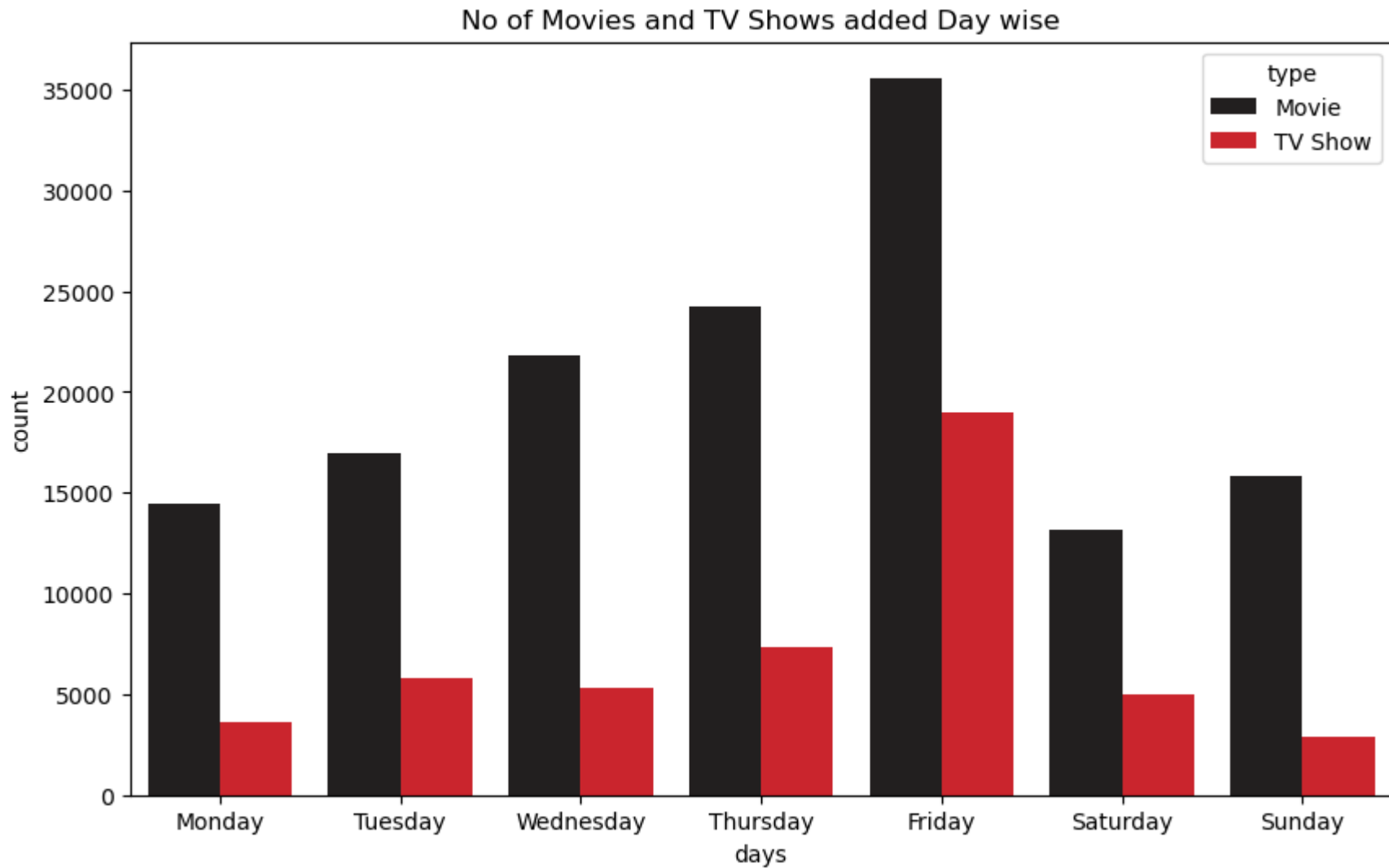


From 2010 onwards, there was a steady increase in both movies and TV shows, peaking in 2020, but then declining in 2021, possibly due to the COVID-19 pandemic.

```
In [169... #No of Movies and TV Shows added Day wise
```

```
In [244... netflix_df = netflix_df.copy()
netflix_df['date_added'] = pd.to_datetime(netflix_df['date_added'])
netflix_df.loc[:, 'days'] = netflix_df['date_added'].dt.day_name()
```

```
netflix_colors = ['#221f1f', '#e50914']  
plt.figure(figsize=(10, 6))  
sns.countplot(x="days", data=netflix_df, hue="type", order=["Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"],  
plt.title("No of Movies and TV Shows added Day wise")  
plt.show()
```



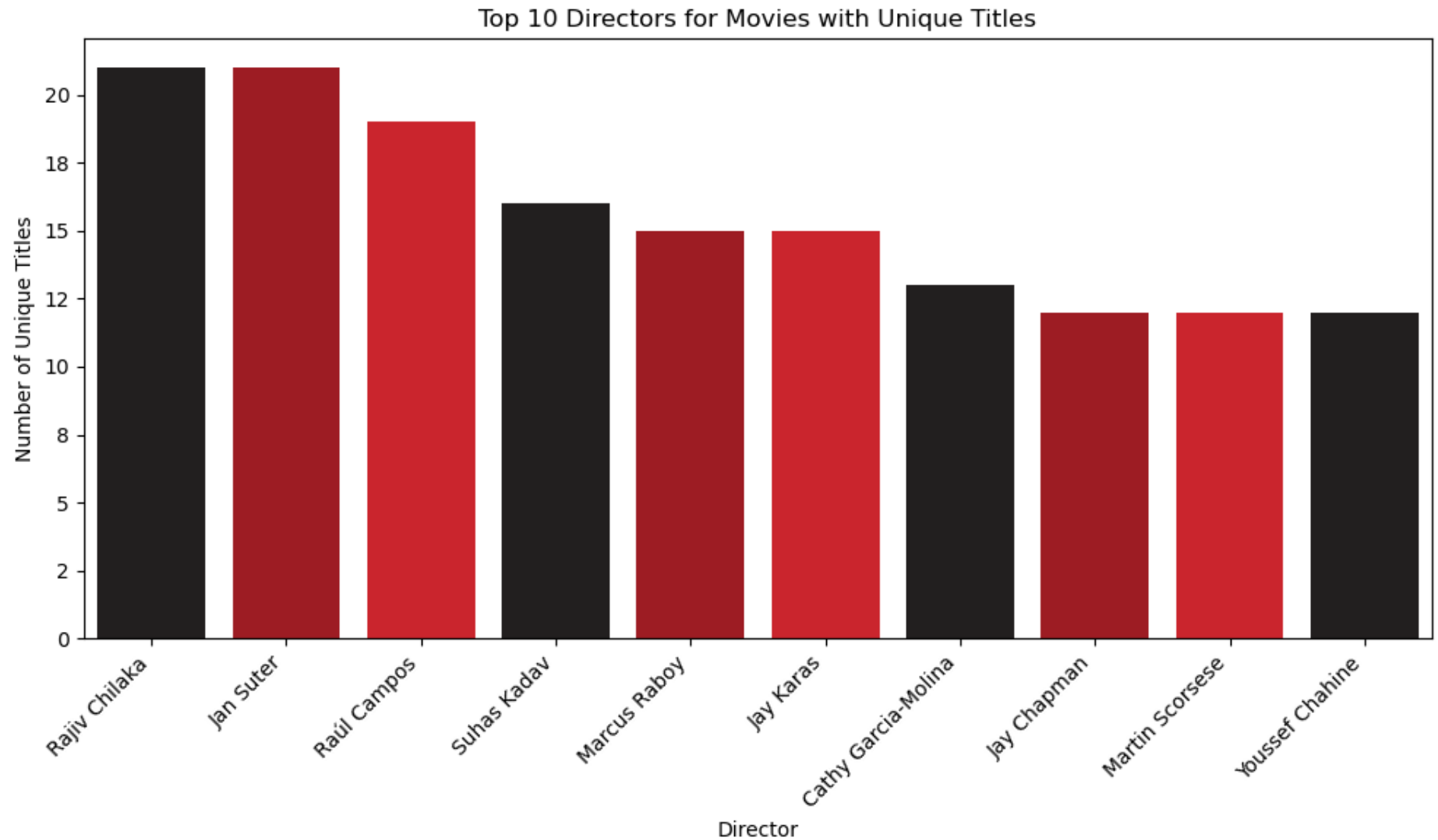
Many new releases on Netflix tend to drop on Fridays and Thursdays, likely because these days precede the weekend when people have more free time and are eager to watch.

In [171... *#Top 10 Directors for Movies with Unique Titles*

```
In [245... movies_df = netflix_df[(netflix_df['type'] == 'Movie') & (netflix_df['director'] != 'Director not specified')]
unique_titles_movies = movies_df.groupby('director')['title'].nunique().sort_values(ascending=False)

#top 10 directors by count of unique titles for movies
top_10_directors_movies = unique_titles_movies.head(10).reset_index()

netflix_colors = ['#221f1f', '#b20710', '#e50914']
# Plot top 10 directors for movies
plt.figure(figsize=(10, 6))
ax = sns.barplot(x='director', y='title', data=top_10_directors_movies, palette=netflix_colors)
plt.title('Top 10 Directors for Movies with Unique Titles')
plt.xlabel('Director')
plt.ylabel('Number of Unique Titles')
plt.xticks(rotation=45, ha='right')
ax.yaxis.set_major_formatter(plt.matplotlib.ticker.StrMethodFormatter('{x:.0f}'))
plt.tight_layout()
plt.show()
```



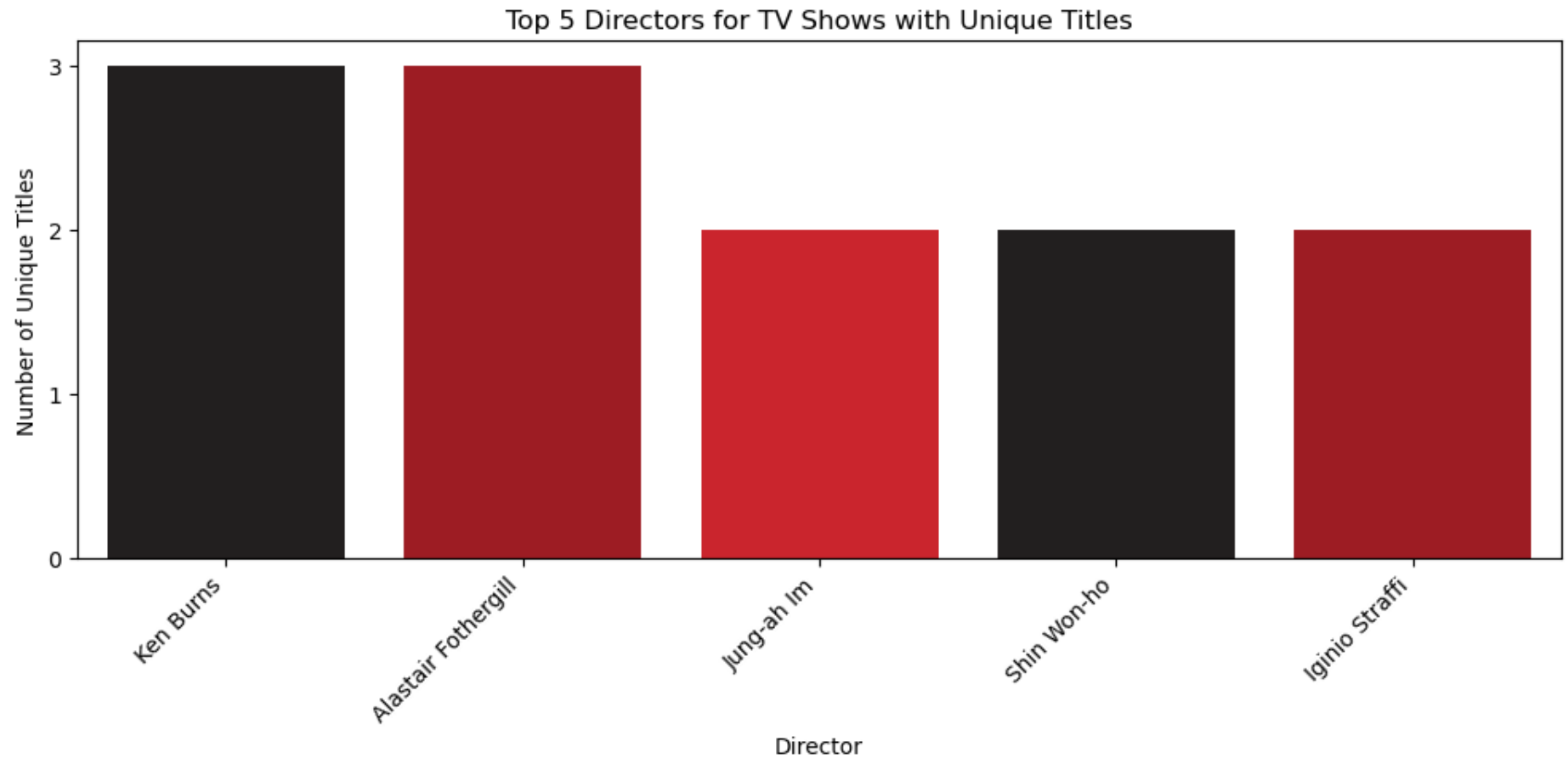
Rajiv Chilaka and Jan Suter leads with 21 unique movies, showcasing a diverse range of genres on Netflix. Raúl Campos follow closely behind with 19 titles each.

```
In [173... #Top 5 Directors for TV shows with Unique Titles
```

```
In [246... tv_shows_df = netflix_df[(netflix_df['type'] == 'TV Show') & (netflix_df['director'] != 'Director not specified')]
```

```
director_unique_titles_count_tv = tv_shows_df.groupby('director')['title'].nunique().sort_values(ascending=False)
top_5_directors_tv = director_unique_titles_count_tv.head(5).reset_index()

netflix_colors = ['#221f1f', '#b20710', '#e50914']
# Plot the top 5 directors for TV shows with unique titles
plt.figure(figsize=(10, 5))
ax = sns.barplot(x='director', y='title', data=top_5_directors_tv, palette=netflix_colors)
ax.set_ylim(bottom=0)
ax.yaxis.set_major_locator(plt.MaxNLocator(integer=True))
plt.title('Top 5 Directors for TV Shows with Unique Titles')
plt.xlabel('Director')
plt.ylabel('Number of Unique Titles')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



Ken Burns leads with 3 unique TV Shows on Netflix.

In [175...

#Top 10 Genres on Netflix

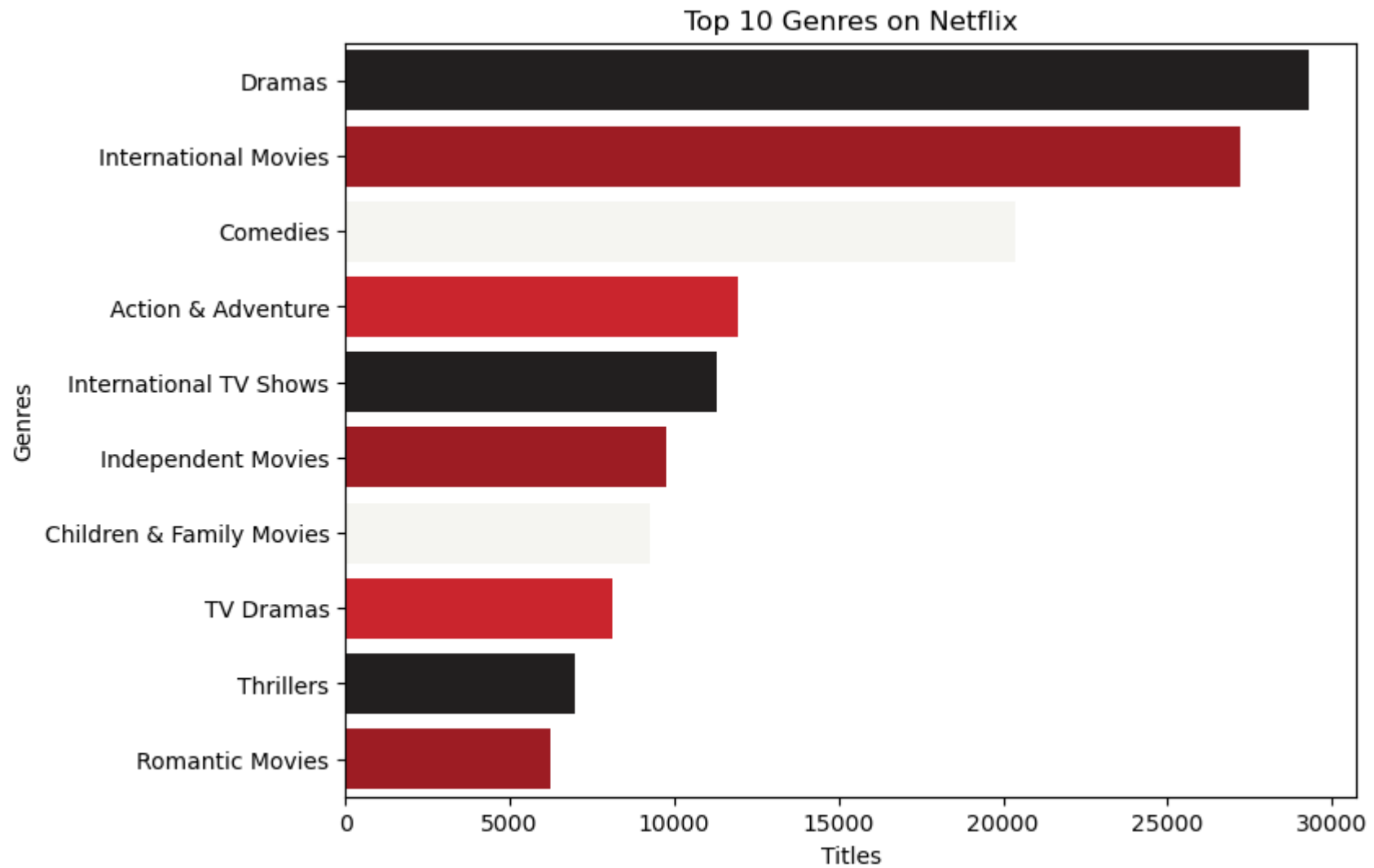
In [247...

```
filtered_genres = netflix_df.set_index('title')['listed_in'].str.split(', ', expand=True).stack().reset_index(level=1, drop=True)

netflix_colors = ['#221f1f', '#b20710', '#f5f5f1', '#e50914']

# Plotting
plt.figure(figsize=(8, 6))
g = sns.countplot(y=filtered_genres, order=filtered_genres.value_counts().index[:10], palette=netflix_colors)
plt.title('Top 10 Genres on Netflix')
plt.xlabel('Titles')
```

```
plt.ylabel('Genres')  
plt.show()
```



Dramas are the most produced on Netflix.

```
In [64]: #Popular Genre Movies  
         #word cloud
```

In [65]: `!pip install wordcloud`

```
Requirement already satisfied: wordcloud in c:\users\suriy\anaconda3\lib\site-packages (1.9.3)
Requirement already satisfied: numpy>=1.6.1 in c:\users\suriy\anaconda3\lib\site-packages (from wordcloud) (1.24.3)
Requirement already satisfied: pillow in c:\users\suriy\anaconda3\lib\site-packages (from wordcloud) (10.2.0)
Requirement already satisfied: matplotlib in c:\users\suriy\anaconda3\lib\site-packages (from wordcloud) (3.7.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\suriy\anaconda3\lib\site-packages (from matplotlib->wordcloud) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\suriy\anaconda3\lib\site-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\suriy\anaconda3\lib\site-packages (from matplotlib->wordcloud) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\suriy\anaconda3\lib\site-packages (from matplotlib->wordcloud) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\suriy\anaconda3\lib\site-packages (from matplotlib->wordcloud) (23.1)
Requirement already satisfied: pyparsing<3.1, >=2.3.1 in c:\users\suriy\anaconda3\lib\site-packages (from matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\suriy\anaconda3\lib\site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\suriy\anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
```

In [248... `from wordcloud import WordCloud`

```
# Combine all genre entries into a single string
all_genres = ' '.join(netflix_df['listed_in'])

netflix_colors = ['#221f1f', '#b20710', '#e50914']

# Create a function to generate color
def my_color_func(word, font_size, position, orientation, random_state=None, **kwargs):
    return np.random.choice(netflix_colors)

wordcloud = WordCloud(width=800, height=400, background_color='white', color_func=my_color_func).generate(all_genres)

# Plot the word cloud
plt.figure(figsize=(12, 8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.title('Movie Genres on Netflix')
plt.axis('off')
plt.show()
```

[illegible]

#content available in different countries

```
import geopandas as gpd
from matplotlib.colors import ListedColormap

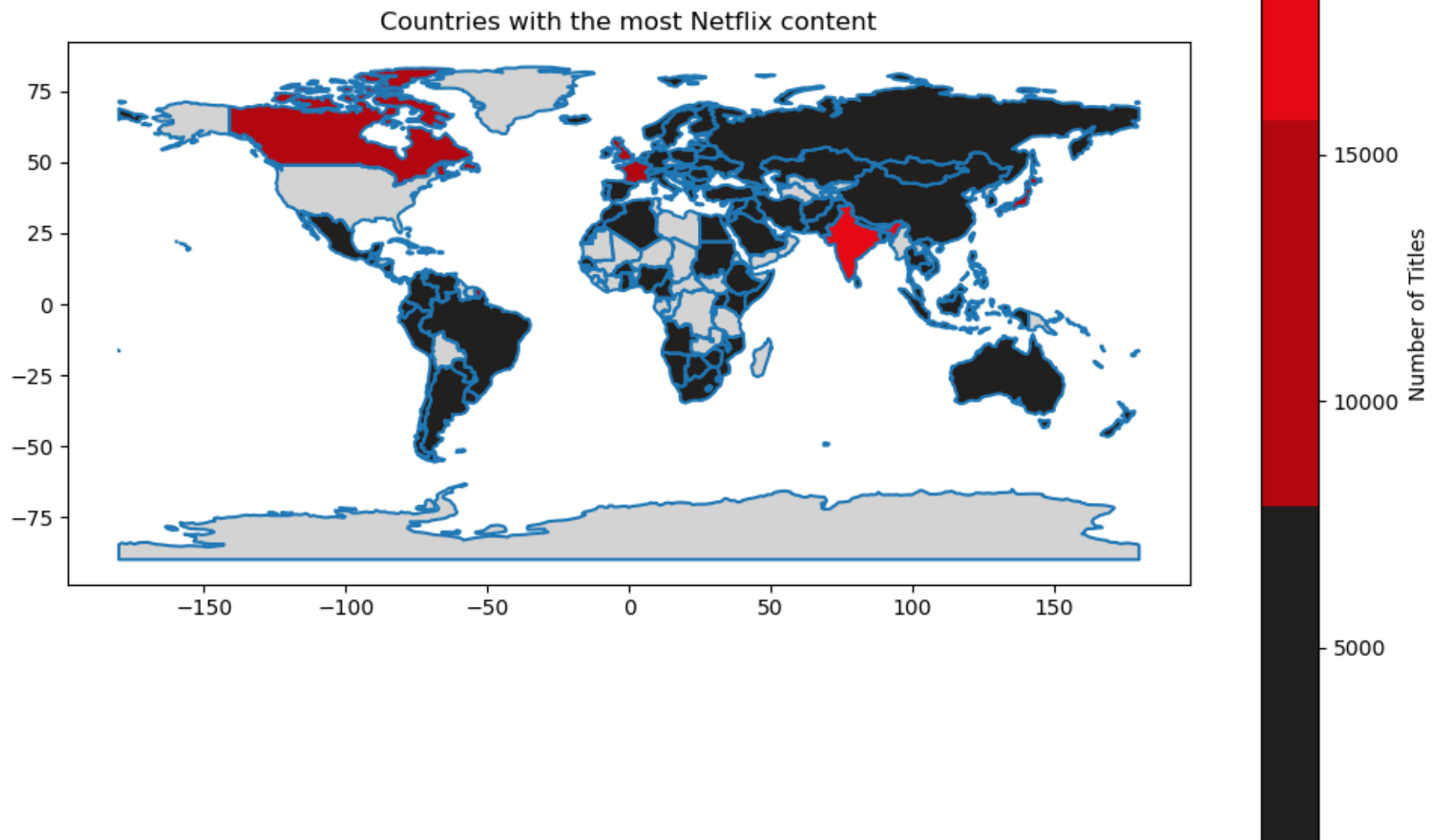
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
country_counts = netflix_df['country'].value_counts().reset_index()
country_counts.columns = ['country', 'num_titles']
world = world.merge(country_counts, how='left', left_on='name', right_on='country')
custom_cmap = ListedColormap(netflix_colors)

# Plot the map
```

```
fig, ax = plt.subplots(1, 1, figsize=(12, 10))
world.boundary.plot(ax=ax)
world.plot(column='num_titles', ax=ax, legend=True,
           legend_kwds={'label': "Number of Titles"},
           cmap=custom_cmap, missing_kwds={'color': 'lightgrey'})
ax.set_title('Countries with the most Netflix content')
plt.show()
```

C:\Users\suriy\AppData\Local\Temp\ipykernel_43808\2444040397.py:4: FutureWarning: The geopandas.dataset module is deprecated and will be removed in GeoPandas 1.0. You can get the original 'naturalearth_lowres' data from <https://www.naturalearthdata.com/downloads/110m-cultural-vectors/>.

```
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
```



India is leading with more content across the world.

Insights

1. Content Type Distribution: Netflix produces a higher proportion of movies compared to TV shows, with approximately 70% of the content being movies and 30% being TV shows.
2. Release Trends: The minimum release year in the dataset is 1925, indicating the historical breadth of content available on Netflix. Analysis reveals a slight increase in the number of movies and TV shows released starting from the year 2000, suggesting a growing emphasis on content production in recent decades.
3. Country Analysis: The United States and India emerge as the top producers of movies on Netflix, while the United States and the United Kingdom lead in TV show production.
4. Duration and Seasons: The average duration of movies on Netflix is around 100 minutes, indicating a preference for shorter formats among viewers. TV shows predominantly consist of 1 or 2 seasons.
5. Production and Casting Insights: Marcus Raboy from the United States has produced more content, while Anupam Kher has appeared in more movies, and Takahiro Sakurai has appeared in more TV shows.
6. Content Ratings: TV-MA (Mature Audience) ratings are prevalent in the content available on Netflix in the top 10 countries, indicating a preference for mature-themed content among subscribers. India and the United States exhibit a higher prevalence of TV-14 ratings compared to other rating categories, reflecting diverse viewer preferences across regions.
7. Release Timing: Analysis of release timing indicates that July, April, and December witness the highest number of movie releases, while July and December also see a peak in TV show releases. This suggests that Netflix strategically schedules content releases to coincide with peak viewing periods, such as holidays and seasonal breaks.
8. Trends Over Time: From 2010 onwards, there is a consistent upward trend in the number of movies and TV shows released on Netflix, peaking in 2020. However, there is a slight decline in 2021, possibly due to external factors such as the COVID-19 pandemic.

9. Director and Actor Insights: Rajiv Chilaka and Jan Suter leads with 21 unique movies on Netflix, showcasing a diverse range of genres. Raúl Campos closely follow with 19 titles each. Ken Burns leads with 3 unique TV shows on Netflix.
10. Genre Preferences: Dramas emerge as the most produced content category on Netflix because adults love watching dramas.
11. Global Analysis: India produces more content and secondly the United States.

Recommendations

1. Given the higher proportion of movies (70%) in Netflix's content, it might be beneficial to continue investing in movie production. However, considering the growing popularity of TV shows, especially with the advent of binge-watching culture, increasing the proportion of TV shows could attract a wider audience.
2. As there's been an increase in content production since 2000, Netflix should continue this trend. Producing more original content can help differentiate Netflix from competitors.
3. Given that the US and India are the top producers of Netflix content, it would be beneficial to continue investing in these markets. Additionally, exploring partnerships with production houses in other high-potential markets could help diversify the content library.
4. Netflix should schedule the release of their major shows and movies during holidays or vacation periods when audience engagement is typically higher. This can lead to increased viewership and greater popularity for their content.
5. Netflix should continue making movies that are around 100 minutes in length, as this duration tends to be appealing to viewers. Additionally, producing shorter content for on-the-go audiences can further cater to diverse viewing preferences. TV shows with 1 or 2 seasons suggest that Netflix may prioritize shorter series to maintain viewer engagement.
6. Given the prevalence of TV-MA ratings, producing more mature-themed content could cater to the existing viewer base. However, to attract a wider audience, including families and younger viewers, Netflix could also invest in more TV-PG and TV-14 rated content.
7. Netflix should collaborate with talented directors and actors who have a track record of creating successful content. By partnering with experienced professionals, they can produce more compelling and popular content.
8. Netflix should be prepared to adapt their plans and strategies in response to rapid changes, such as those experienced during the COVID-19 pandemic. Flexibility and agility in decision-making can help them navigate uncertainties and maintain relevance.

9. When promoting new shows or movies, Netflix should employ creative and engaging advertising techniques. This can generate excitement among viewers and contribute to the success and popularity of the content.

By doing these things, Netflix can keep growing and making shows and movies that people love to watch.