NETFLIX_DATA

December 23, 2024

0.0.1 Netflix Data Analysis

[ ]:

[1]:
```
# import required libraries

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

[2]:
```
DF = pd.read_csv('netflix1.csv')
```

[3]:
```
DF.shape
```

[3]: (8790, 10)

[4]:
```
DF.head()
```

[4]:

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act…
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	

```
4   Dramas, Independent Movies, International Movies
```

[5]:
```
DF.tail()
```

[5]:
```
     show_id     type              title   director        country  \
8785   s8797  TV Show          Yunus Emre  Not Given         Turkey
8786   s8798  TV Show           Zak Storm  Not Given  United States
8787   s8801  TV Show  Zindagi Gulzar Hai  Not Given       Pakistan
8788   s8784  TV Show                Yoko  Not Given       Pakistan
8789   s8786  TV Show                 YOM  Not Given       Pakistan

      date_added  release_year rating   duration  \
8785   1/17/2017          2016  TV-PG  2 Seasons
8786   9/13/2018          2016  TV-Y7  3 Seasons
8787  12/15/2016          2012  TV-PG   1 Season
8788   6/23/2018          2016   TV-Y   1 Season
8789  06-07-2018          2016  TV-Y7   1 Season

                                          listed_in
8785                 International TV Shows, TV Dramas
8786                                      Kids' TV
8787  International TV Shows, Romantic TV Shows, TV …
8788                                      Kids' TV
8789                                      Kids' TV
```

[6]:
```
DF.describe()
```

[6]:
```
       release_year
count   8790.000000
mean    2014.183163
std        8.825466
min     1925.000000
25%     2013.000000
50%     2017.000000
75%     2019.000000
max     2021.000000
```

[7]:
```
DF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
```

```
 4   country       8790 non-null   object
 5   date_added    8790 non-null   object
 6   release_year  8790 non-null   int64
 7   rating        8790 non-null   object
 8   duration      8790 non-null   object
 9   listed_in     8790 non-null   object
dtypes: int64(1), object(9)
memory usage: 686.8+ KB
```

[8]:
```
DF.isnull().sum()
```

[8]:
```
show_id         0
type            0
title           0
director        0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```

[9]:
```
DF.drop_duplicates(inplace=True)
```

[10]:
```
DF.dropna()
```

[10]:
```
     show_id     type                              title         director  \
0         s1    Movie               Dick Johnson Is Dead  Kirsten Johnson
1         s3  TV Show                          Ganglands  Julien Leclercq
2         s6  TV Show                      Midnight Mass    Mike Flanagan
3        s14    Movie  Confessions of an Invisible Girl    Bruno Garotti
4         s8    Movie                            Sankofa     Haile Gerima
…       …       …                                 …               …
8785   s8797  TV Show                         Yunus Emre        Not Given
8786   s8798  TV Show                          Zak Storm        Not Given
8787   s8801  TV Show                 Zindagi Gulzar Hai        Not Given
8788   s8784  TV Show                               Yoko        Not Given
8789   s8786  TV Show                                YOM        Not Given

            country  date_added  release_year rating   duration  \
0     United States   9/25/2021          2020  PG-13     90 min
1            France   9/24/2021          2021  TV-MA   1 Season
2     United States   9/24/2021          2021  TV-MA   1 Season
3            Brazil   9/22/2021          2021  TV-PG     91 min
4     United States   9/24/2021          1993  TV-MA    125 min
…              …         …             …     …        …
8785         Turkey   1/17/2017          2016  TV-PG  2 Seasons
```

```
8786  United States   9/13/2018          2016  TV-Y7  3 Seasons
8787       Pakistan  12/15/2016          2012  TV-PG   1 Season
8788       Pakistan   6/23/2018          2016   TV-Y   1 Season
8789       Pakistan  06-07-2018          2016  TV-Y7   1 Season

                                                listed_in
0                                           Documentaries
1     Crime TV Shows, International TV Shows, TV Act…
2                   TV Dramas, TV Horror, TV Mysteries
3                   Children & Family Movies, Comedies
4      Dramas, Independent Movies, International Movies
…                                                     …
8785                 International TV Shows, TV Dramas
8786                                          Kids' TV
8787  International TV Shows, Romantic TV Shows, TV …
8788                                          Kids' TV
8789                                          Kids' TV

[8790 rows x 10 columns]
```

[17]:
```
DF['date_added'] = pd.to_datetime(DF['date_added'], infer_datetime_format=True,␣
↪errors='coerce')
```

```
C:\Users\ARCHANA CHOUGALE\AppData\Local\Temp\ipykernel_15508\623934212.py:1:
UserWarning: The argument 'infer_datetime_format' is deprecated and will be
removed in a future version. A strict version of it is now the default, see
https://pandas.pydata.org/pdeps/0004-consistent-to-datetime-parsing.html. You
can safely remove this argument.
  DF['date_added'] = pd.to_datetime(DF['date_added'],
infer_datetime_format=True, errors='coerce')
```

[18]:
```
DF.dtypes
```

[18]:
```
show_id                 object
type                    object
title                   object
director                object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
dtype: object
```

[19]:
```
DF.head()
```

[19]:

	show_id	type	title	director
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson
1	s3	TV Show	Ganglands	Julien Leclercq
2	s6	TV Show	Midnight Mass	Mike Flanagan
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti
4	s8	Movie	Sankofa	Haile Gerima

	country	date_added	release_year	rating	duration	listed_in
0	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	France	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act…
2	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	Brazil	2021-09-22	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	United States	2021-09-24	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies

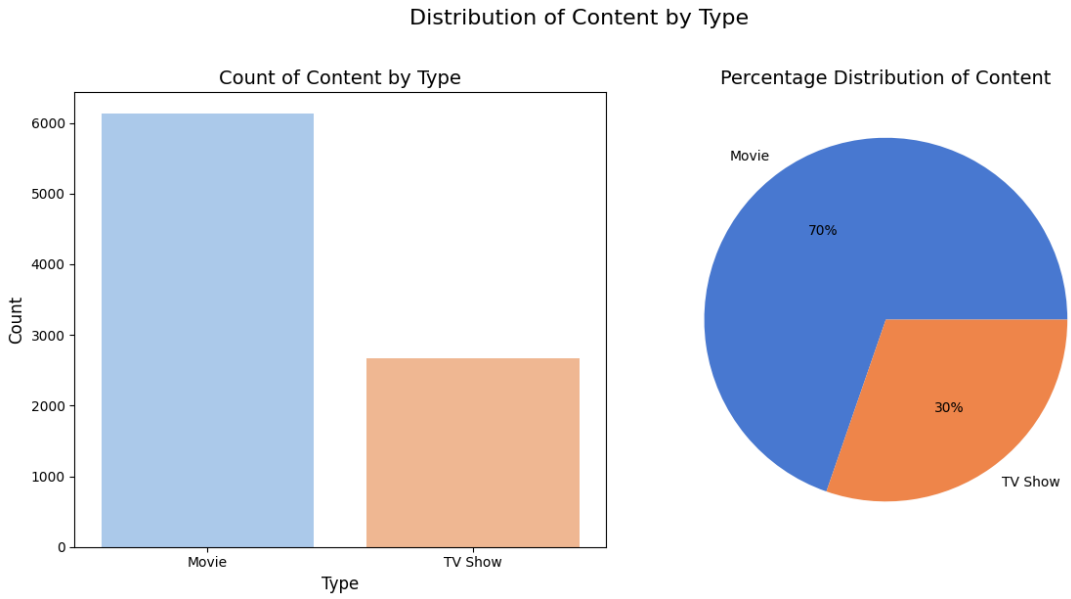
0.0.2 1.Distribution of content by type

[ ]:

[38]:
```
# Count the number of Movies and TV Shows
type_counts = DF['type'].value_counts()

# Create the figure and axes
fig, axes = plt.subplots(1, 2, figsize=(12, 6))
colors = sns.color_palette("muted", len(type_counts))
# Plot a countplot
sns.countplot(data=DF, x='type', ax=axes[0], palette="pastel")
axes[0].set_title('Count of Content by Type', fontsize=14)
axes[0].set_xlabel('Type', fontsize=12)
axes[0].set_ylabel('Count', fontsize=12)

# Plot a pie chart
axes[1].pie(type_counts, labels=type_counts.index, autopct='%.0f%%',␣
 ↪colors=colors)
axes[1].set_title('Percentage Distribution of Content', fontsize=14)

# Adjust layout
plt.suptitle('Distribution of Content by Type', fontsize=16, y=1.02)
plt.tight_layout()
plt.show()
```

```
C:\Users\ARCHANA CHOUGALE\AppData\Local\Temp\ipykernel_15508\1547861761.py:8:
```

```
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data=DF, x='type', ax=axes[0], palette="pastel")
```



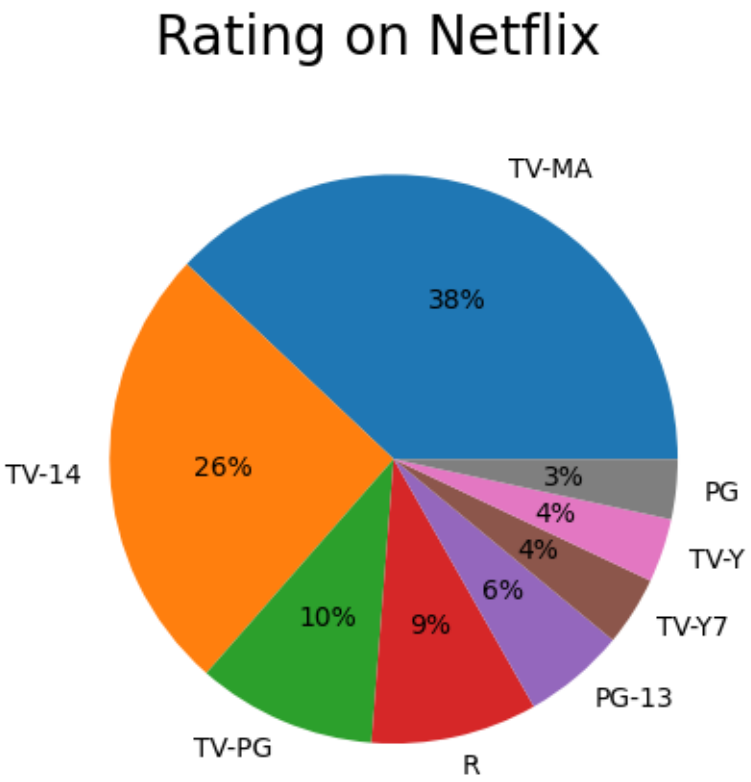
0.0.3

Using the dataset, we calculated the proportions of each content type. The analysis shows that: Movies dominate the platform, making up 70% of all titles.TV Shows account for the remaining 30%.

0.0.4 2. Rating on Netflix

[21]:
```
ratings=DF['rating'].value_counts().reset_index().sort_values(by='count',␣
  ↪ascending=False)
rating_counts = DF['rating'].value_counts()
plt.pie(ratings['count'][:8], labels=ratings['rating'][:8], autopct='%.0f%%')
plt.suptitle('Rating on Netflix', fontsize=20)
```

[21]:
```
Text(0.5, 0.98, 'Rating on Netflix')
```



Observations

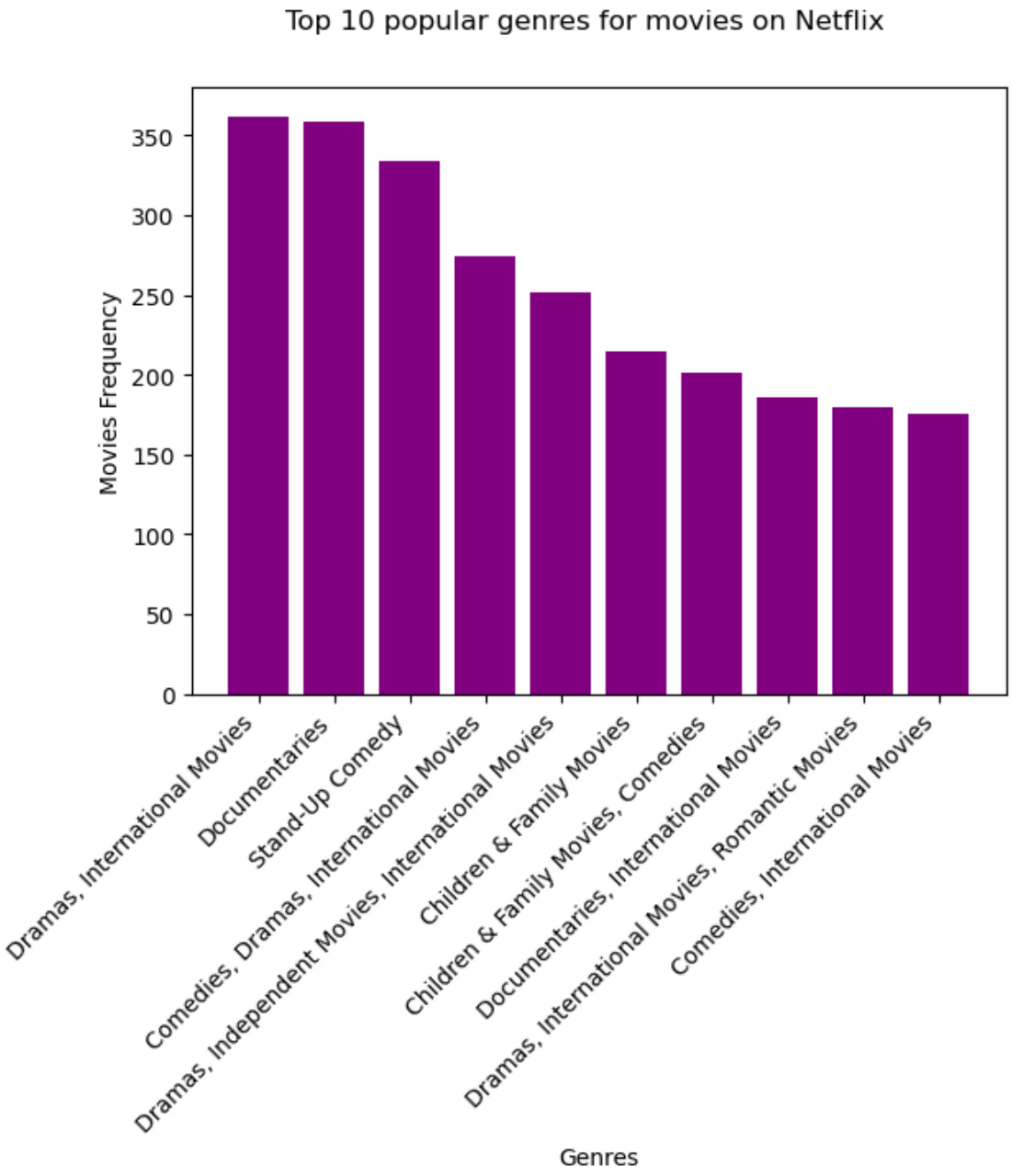
Overall Distribution: TV-MA leads the ratings, making up a significant portion of Netflix's catalog. TV-14 and TV-PG follow, showing that Netflix appeals to a wide audience but primarily targets mature viewers.

Ratings by Type: Movies tend to have a broader range of ratings, including R and PG-13, reflecting traditional film ratings. TV Shows are dominated by TV-MA and TV-14, highlighting a focus on episodic content for adults and teens.

0.0.5 TOP 10 MOVIE GENRES

[26]:
```
popular_movie_genre=DF[DF['type']=='Movie'].groupby("listed_in").size().
  ↪sort_values(ascending=False)[:10]
popular_series_genre=DF[DF['type']=='TV Show'].groupby("listed_in").size().
  ↪sort_values(ascending=False)[:10]


plt.bar(popular_movie_genre.index, popular_movie_genre.values,color='purple')
plt.xticks(rotation=45, ha='right')
plt.xlabel("Genres")
plt.ylabel("Movies Frequency")
```

```
plt.suptitle("Top 10 popular genres for movies on Netflix")
plt.show()
```

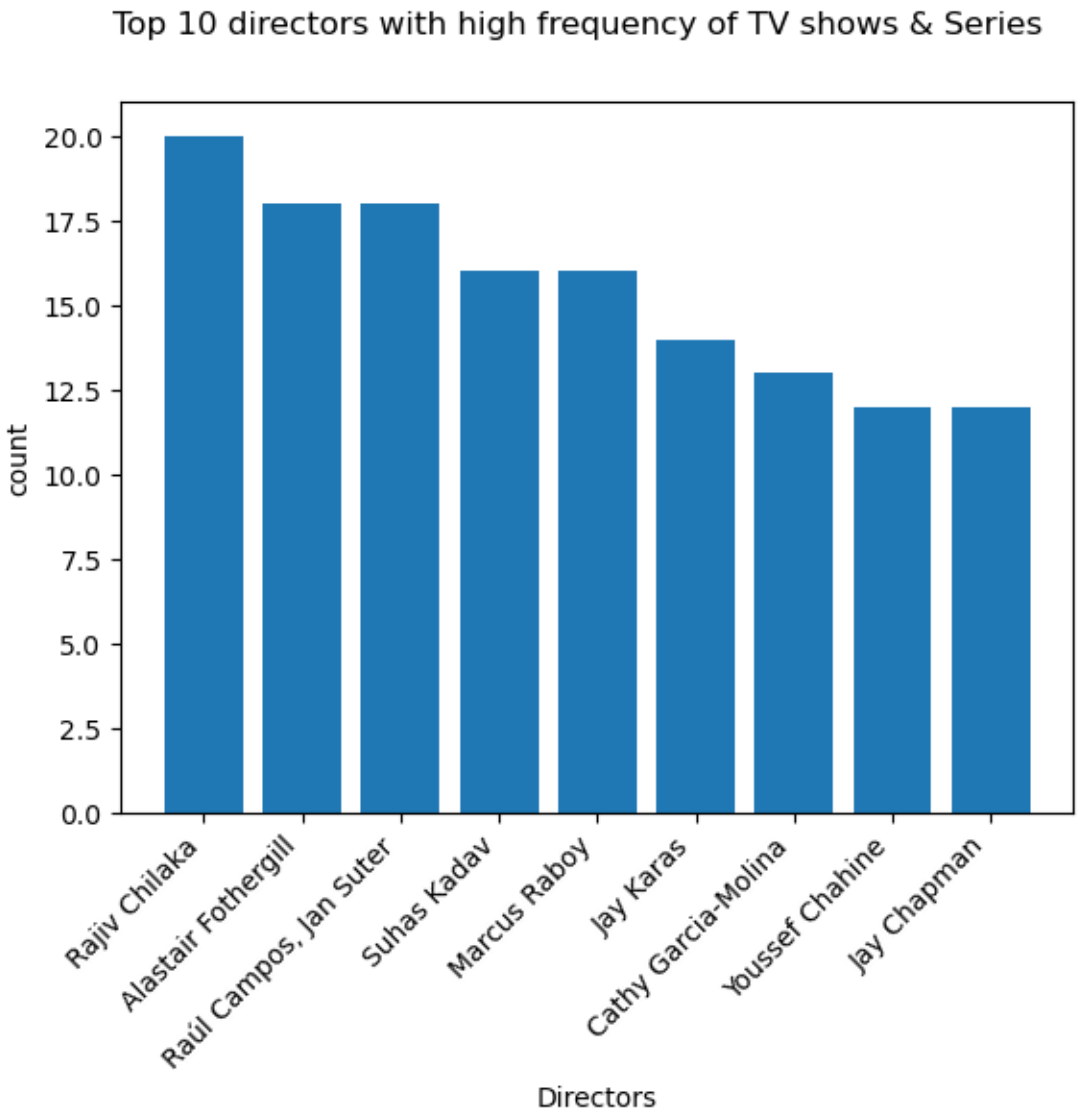


Observation- Top Genres:

The most frequently occurring genres on Netflix are: International Movies Dramas Comedies International TV Shows Documentaries

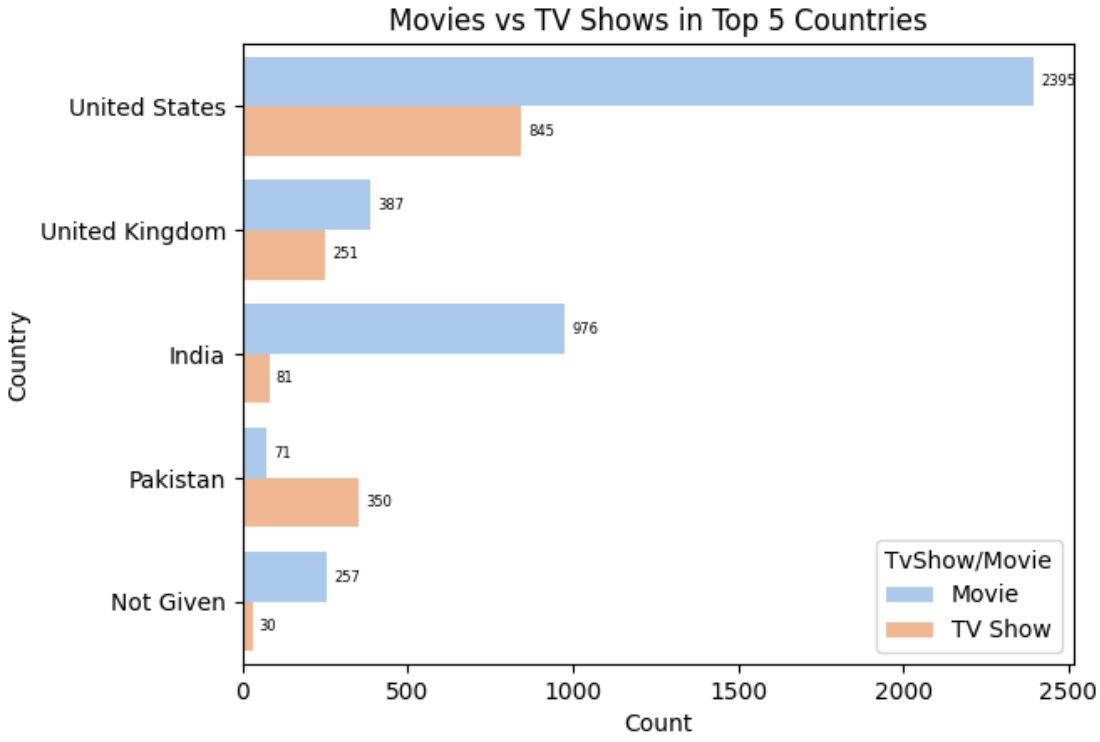
1 Top 10 directors with high frequency of TV shows & Series

[27]:
```
directors=DF['director'].value_counts().reset_index().sort_values(by='count',␣
↪ascending=False)[1:10]

plt.bar(directors['director'], directors['count'])
plt.xticks(rotation=45, ha='right')
plt.xlabel("Directors")
plt.ylabel("count")
plt.suptitle("Top 10 directors with high frequency of TV shows & Series ")
plt.show()
```



2 Movies vs TV Shows in Top 5 Countries

[39]:
```
top_countries = DF['country'].value_counts().head(5).index
filtered_df =DF[DF['country'].isin(top_countries)]

ax=sns.countplot(y='country', hue='type', data=filtered_df, palette='pastel')
# Add value annotations
for container in ax.containers:
    ax.bar_label(container, label_type='edge', fontsize=6, padding=3)

plt.title('Movies vs TV Shows in Top 5 Countries')
plt.xlabel('Count')
plt.ylabel('Country')
plt.legend(title='TvShow/Movie')
plt.show()
```



Observation-

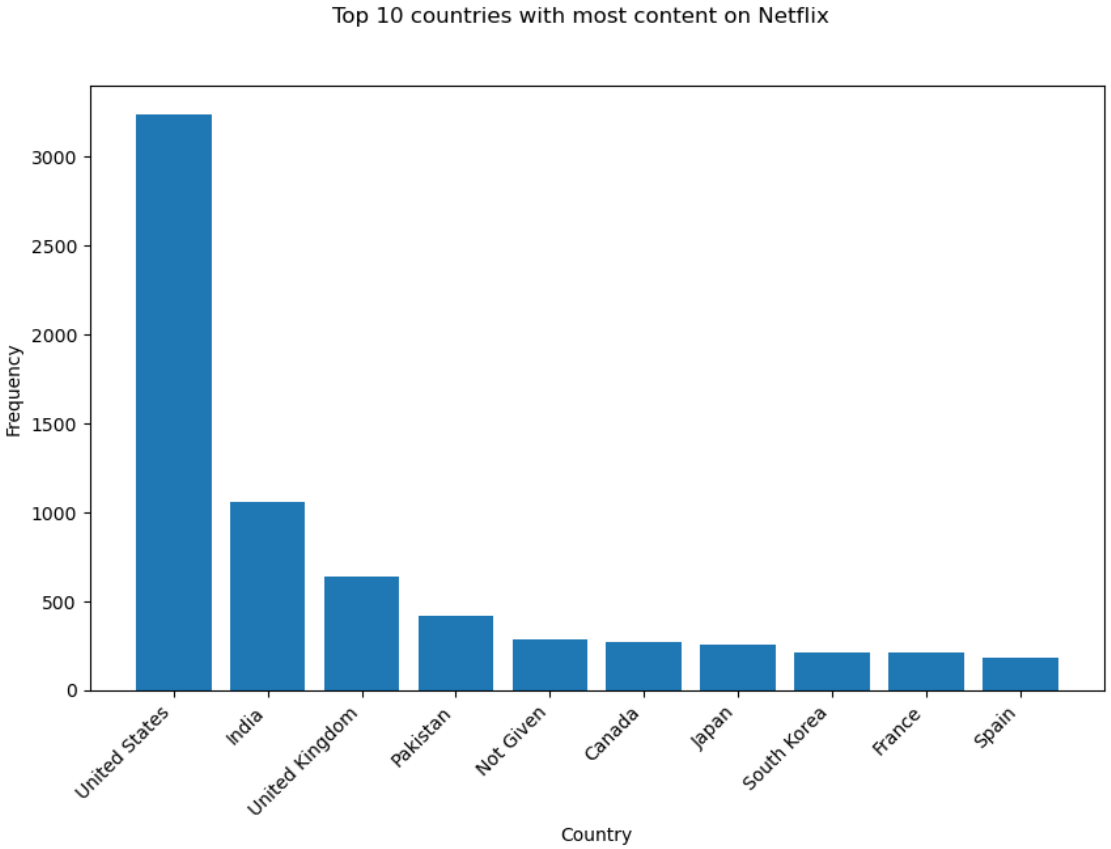
The dataset reveals that the following directors have the highest number of productions available on Netflix:

Rajiv Chilaka,Alastair Fothergill,Raul campos,jansuter etc.

3 Top 10 countries with most content on Netflix

[ ]:

[29]:
```
top_five_countries=DF['country'].value_counts().reset_index().
 ↪sort_values(by='count', ascending=False)[:10]
plt.figure(figsize=(10, 6))
plt.bar(top_five_countries['country'], top_five_countries['count'])
plt.xticks(rotation=45, ha='right')
plt.xlabel("Country")
plt.ylabel("Frequency")
plt.suptitle("Top 10 countries with most content on Netflix")
plt.show()
```

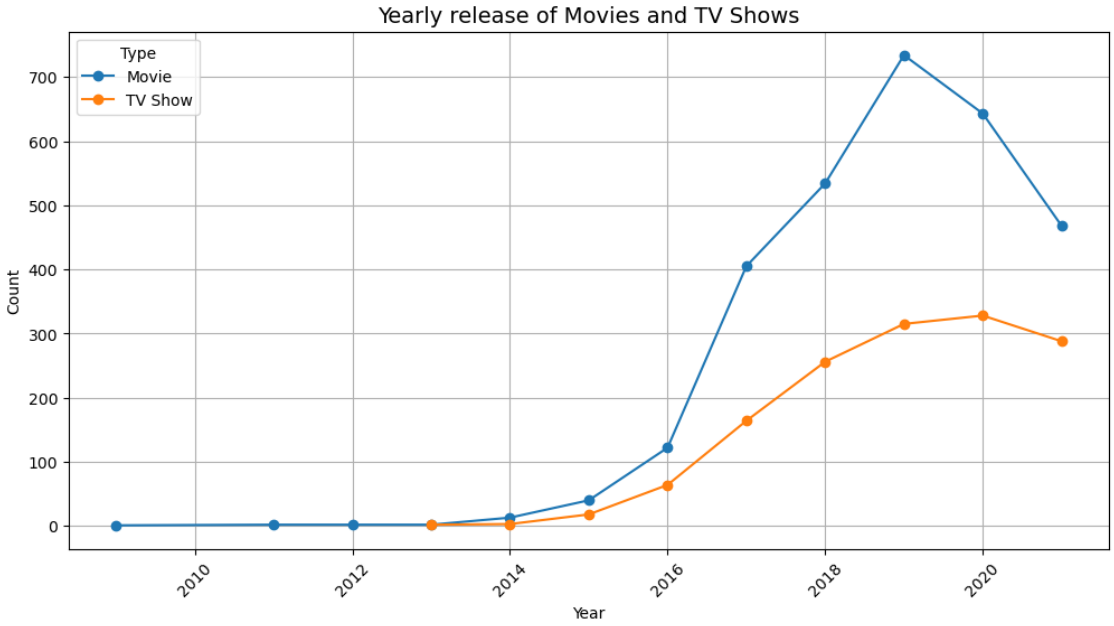


Observation-

From above graph we can analyse that United Staes,United Kingdom,India and pakistan are top countries where most of the movies and TV shows watched on netfilix.

4 Yearly release of Movies and TV Shows

[32]:
```
# Plot content added over the years using a line chart

DF['year']=DF['date_added'].dt.year
plt.figure(figsize=(12, 6))

DF.groupby(['year', 'type']).size().unstack().plot(kind='line', marker='o',␣
 ↪ax=plt.gca())

plt.title('Yearly release of Movies and TV Shows',fontsize=14)

plt.xlabel('Year')

plt.ylabel('Count')

plt.xticks(rotation=45)

plt.legend(title='Type')

plt.grid(True)

plt.show()
```



observations– After 2016 the movies and TV shows release goes on increasing.

Insights –

1.Movies dominate the platform, making up 70% of all titles.TV Shows account for the remaining 30%.

2.Netflix's catalog primarily targets mature audiences, as shown by the dominance of TV-MA. However, the platform maintains a diverse library, ensuring that viewers of all age groups can find suitable content. Regional content ratings, such as NR (Not Rated), could benefit from further exploration to understand their implications for global markets.

3.The most frequently occurring genres on Netflix are: International Movies Dramas Comedies International TV Shows Documentaries. The popularity of International Movies and TV Shows underscores Netflix's focus on expanding its global audience.

4.The highest number of productions available on Netflix:Rajiv Chilaka,Alastair Fothergill,Raul campos,jansuter etc.

5.United Staes,United Kingdom,India and pakistan are top countries where most of the movies and TV shows watched on netfilix.

6.Ratings by Type: Movies tend to have a broader range of ratings, including R and PG-13, reflecting traditional film ratings. TV Shows are dominated by TV-MA and TV-14, highlighting a focus on episodic content for adults and teens.

[ ]:

[ ]: