# Netflix Data Analysis by Navjoth Singh

November 4, 2023

## 1 Data Collection: Importing Libraries & File

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from matplotlib import style
%matplotlib inline
data = pd.read_csv("C:/Users/hp/OneDrive/Desktop/Data Science/Data for Practice/
 ↪netflix_titles.csv")
```

[2]:
```
head_tail_concatenated = pd.concat([data.head(2), data.tail(2)])
# Display the concatenated DataFrame
head_tail_concatenated
```

[2]:
```
     show_id     type                 title         director  \
0         s1    Movie  Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show         Blood & Water              NaN
8805   s8806    Movie                  Zoom     Peter Hewitt
8806   s8807    Movie                Zubaan      Mozez Singh

                                                 cast        country  \
0                                                 NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban…   South Africa
8805  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma…  United States
8806  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan…          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
8805    January 11, 2020          2006     PG     88 min
8806       March 2, 2019          2015  TV-14    111 min

                                      listed_in  \
0                                   Documentaries
1     International TV Shows, TV Dramas, TV Mysteries
8805              Children & Family Movies, Comedies
```

```
8806   Dramas, International Movies, Music & Musicals

                                                 description
0     As her father nears the end of his life, filmm…
1     After crossing paths at a party, a Cape Town t…
8805  Dragged from civilian life, a former superhero…
8806  A scrappy but poor boy worms his way into a ty…
```

[3]:
```
data.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
```

[4]:
```
data.isnull().sum()    # some blanks cells are in data
```

[4]:
```
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
```

[5]:
```
data.nunique()
```

[5]:
```
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
```

## 2 Data Cleaning

[6]:
```
data['rating'].head(8)
```

[6]:
```
0    PG-13
1    TV-MA
2    TV-MA
3    TV-MA
4    TV-MA
5    TV-MA
6       PG
7    TV-MA
Name: rating, dtype: object
```

[7]:
```
data['rating'].fillna('TV-MA',inplace=True)    # Handling missing values
```

[8]:
```
data['duration'].head(8)
```

[8]:
```
0       90 min
1    2 Seasons
2     1 Season
3     1 Season
4    2 Seasons
5     1 Season
6       91 min
7      125 min
Name: duration, dtype: object
```

[9]:
```
data['duration'].fillna('90 min',inplace=True)  # Handling missing values
```

[10]:
```
data[['director','cast','country','date_added']].head(8)
```

[10]:
```
                      director  \
0               Kirsten Johnson
1                           NaN
2               Julien Leclercq
3                           NaN
4                           NaN
5                 Mike Flanagan
6  Robert Cullen, José Luis Ucha
7                  Haile Gerima

                                                cast  \
0                                                NaN
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban…
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi…
3                                                NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K…
5  Kate Siegel, Zach Gilford, Hamish Linklater, H…
6  Vanessa Hudgens, Kimiko Glenn, James Marsden, …
7  Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D…

                                        country          date_added
0                                 United States  September 25, 2021
1                                  South Africa  September 24, 2021
2                                           NaN  September 24, 2021
3                                           NaN  September 24, 2021
4                                         India  September 24, 2021
5                                           NaN  September 24, 2021
6                                           NaN  September 24, 2021
7  United States, Ghana, Burkina Faso, United Kin…  September 24, 2021
```

[11]:
```
data.fillna({'director':'unknown', 'cast':'unknown', 'country':'unknown'},␣
↪inplace=True)   #Handling missing values
```

[12]:
```
data.dropna(subset=['date_added'],inplace=True)    # used to remove rows with␣
↪missing values
```

[13]:
```
data.isnull().sum()   # All Missing values has been cleared
```

[13]:
```
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
```

[14]:
```
data.dtypes   # Changing Data Type
```

[14]:
```
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
```

## 3 Data Manipulation

[15]:
```
# Adding new columns
```

[16]:
```
data['date_added'] = pd.to_datetime(data['date_added'], format='%B %d, %Y',␣
 ↪errors='coerce')
```

[17]:
```
data['month'] = pd.to_numeric(data['date_added'].dt.month, errors='coerce').
 ↪astype('Int64')

# The errors='coerce' argument ensures that if there are any values that can't␣
 ↪be converted to numeric
#they will be set to NaN (Not a Number) instead of raising an error.
```

[18]:
```
data['year'] = pd.to_numeric(data['date_added'].dt.year, errors='coerce').
 ↪astype('Int64')
```

[19]:
```
data.head(3)
```

[19]:

| | show_id | type | title | director | \ |
|---|---|---|---|---|---|
| 0 | s1 | Movie | Dick Johnson Is Dead | Kirsten Johnson | |
| 1 | s2 | TV Show | Blood & Water | unknown | |
| 2 | s3 | TV Show | Ganglands | Julien Leclercq | |

| | cast | country | \ |
|---|---|---|---|
| 0 | unknown | United States | |

```
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban…   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi…        unknown

  date_added  release_year rating   duration  \
0 2021-09-25          2020  PG-13     90 min
1 2021-09-24          2021  TV-MA  2 Seasons
2 2021-09-24          2021  TV-MA   1 Season

                                          listed_in  \
0                                    Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act…

                                        description  month  year
0  As her father nears the end of his life, filmm…      9  2021
1  After crossing paths at a party, a Cape Town t…      9  2021
2  To protect his family from a powerful drug lor…      9  2021
```

[20]:
```
# Renaming the column
```

[21]:
```
data = data.rename(columns = {'listed_in' : 'genre'})
```

[22]:
```
data.head(3)
```

[22]:
```
  show_id     type                 title         director  \
0      s1    Movie  Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show         Blood & Water          unknown
2      s3  TV Show             Ganglands  Julien Leclercq

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

  date_added  release_year rating   duration  \
0 2021-09-25          2020  PG-13     90 min
1 2021-09-24          2021  TV-MA  2 Seasons
2 2021-09-24          2021  TV-MA   1 Season

                                          genre  \
0                                    Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act…

                                        description  month  year
0  As her father nears the end of his life, filmm…      9  2021
1  After crossing paths at a party, a Cape Town t…      9  2021
```

```
2  To protect his family from a powerful drug lor…      9  2021
```

## 4 Exploratory Data Analysis (EDA)

[23]:
```
data.describe()   # Work on Where Dtype is int
# Descriptive statistics
```

[23]:
```
                           date_added  release_year     month         year
count                            8709   8797.000000    8709.0       8709.0
mean   2019-05-23 01:45:29.452290560   2014.183472  6.653347  2018.887932
min              2008-01-01 00:00:00   1925.000000       1.0       2008.0
25%              2018-04-20 00:00:00   2013.000000       4.0       2018.0
50%              2019-07-12 00:00:00   2017.000000       7.0       2019.0
75%              2020-08-26 00:00:00   2019.000000      10.0       2020.0
max              2021-09-25 00:00:00   2021.000000      12.0       2021.0
std                              NaN      8.822191  3.431434     1.567961
```

[24]:
```
data.dtypes
```

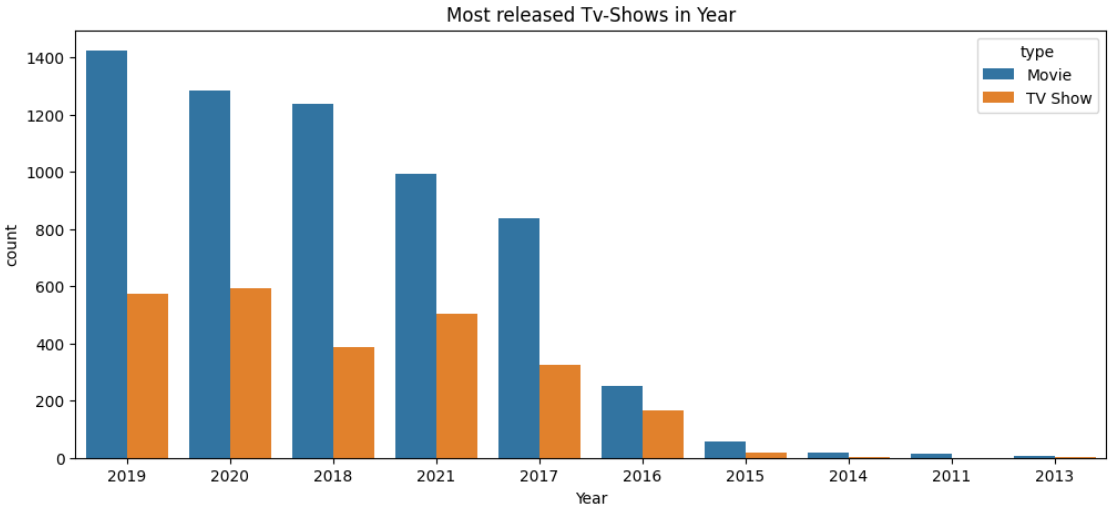
[24]:
```
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
genre                   object
description             object
month                    Int64
year                     Int64
dtype: object
```

1. Which year has the maximum number of released TV shows and movies?

[25]:
```
Movies = data[data['type'] == 'Movie']
tv_shows = data[data['type'] == 'TV Show']
```

[26]:
```
# Convert 'year' column to numeric
data['year'] = pd.to_numeric(data['year'], errors='coerce')

# Drop rows with missing values in 'year' column
data = data.dropna(subset=['year'])
```

[27]:
```
plt.figure(figsize=(12,5))
sns.countplot(x ='year', hue='type', data = data, order = data['year'].
 ↪value_counts().iloc[:10].index)
plt.title("Most released Tv-Shows in Year")
plt.ylabel('count')
plt.xlabel('Year')
plt.show()
```

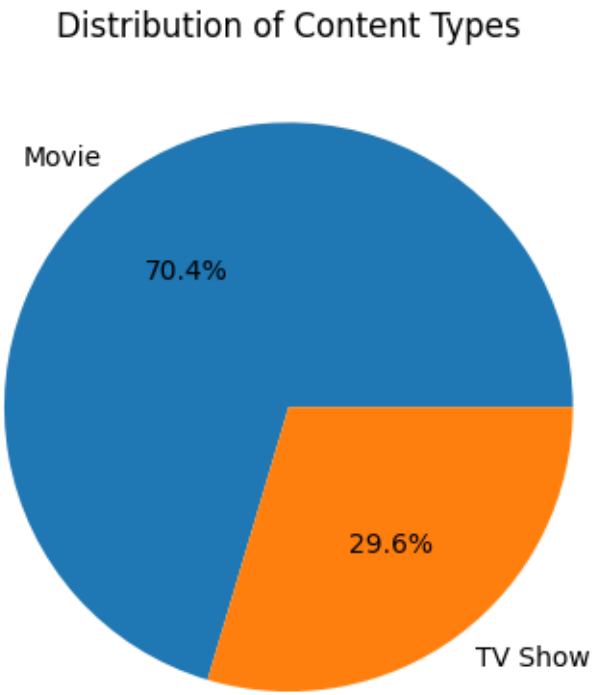



2. What is the distribution of content types (TV shows and movies) in the dataset?

[28]:
```
data['type'].value_counts()
```

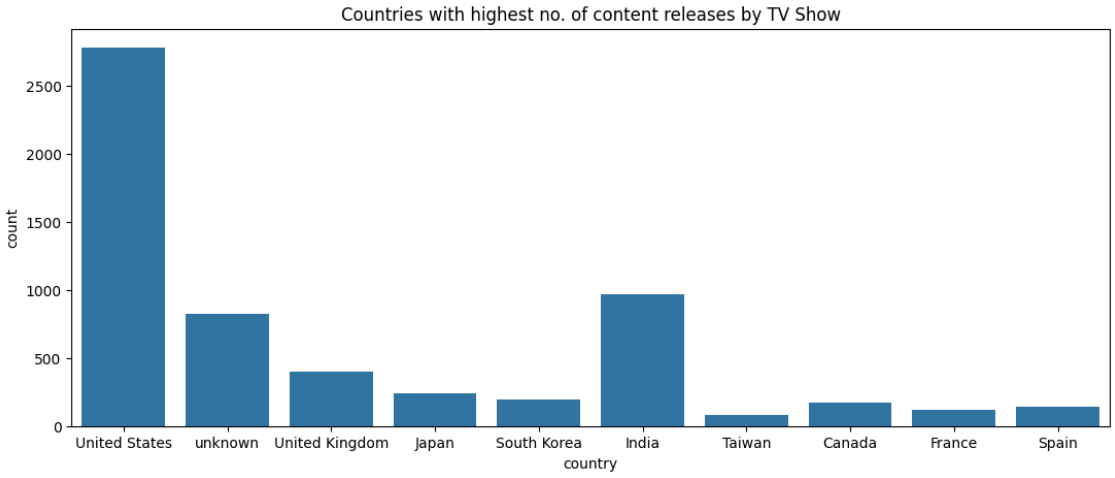
[28]:
```
type
Movie      6131
TV Show    2578
Name: count, dtype: int64
```

[29]:
```
Count = data['type'].value_counts()
plt.pie(Count,labels=Count.index, autopct='%1.1f%%')
plt.title('Distribution of Content Types')
plt.show()
```

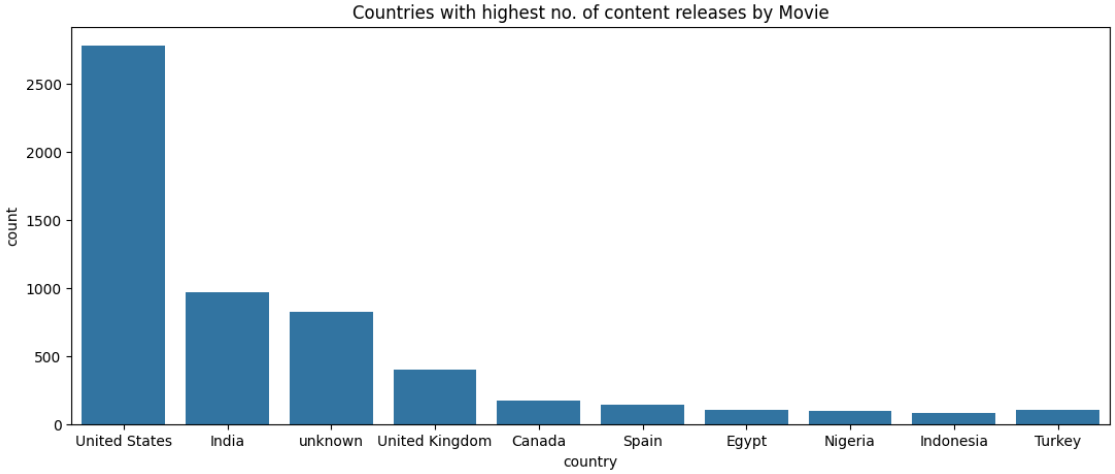



3. Which countries have the highest number of content releases Tv-shows and Movies?

[30]:
```
plt.figure(figsize=(13,5))
sns.countplot(x='country', data=data, order = tv_shows['country'].
 ↪value_counts().iloc[:10].index)
plt.title('Countries with highest no. of content releases by TV Show')
plt.show()
```

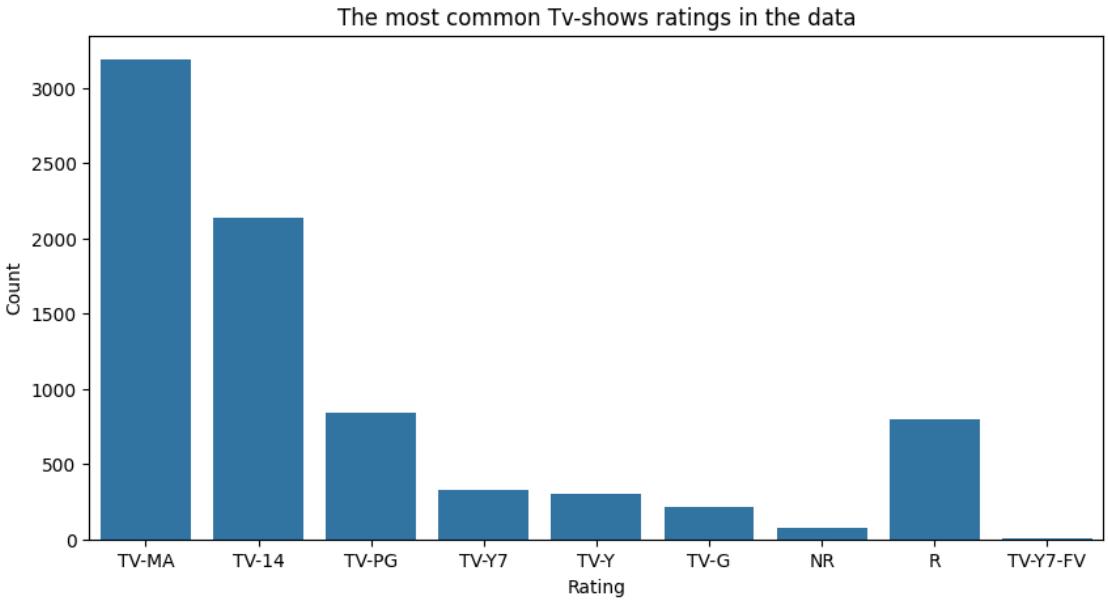


[31]:
```
plt.figure(figsize=(13,5))
sns.countplot(x='country',data=data,order= Movies['country'].value_counts().
  ↪iloc[:10].index)
plt.title('Countries with highest no. of content releases by Movie')
plt.show()
```

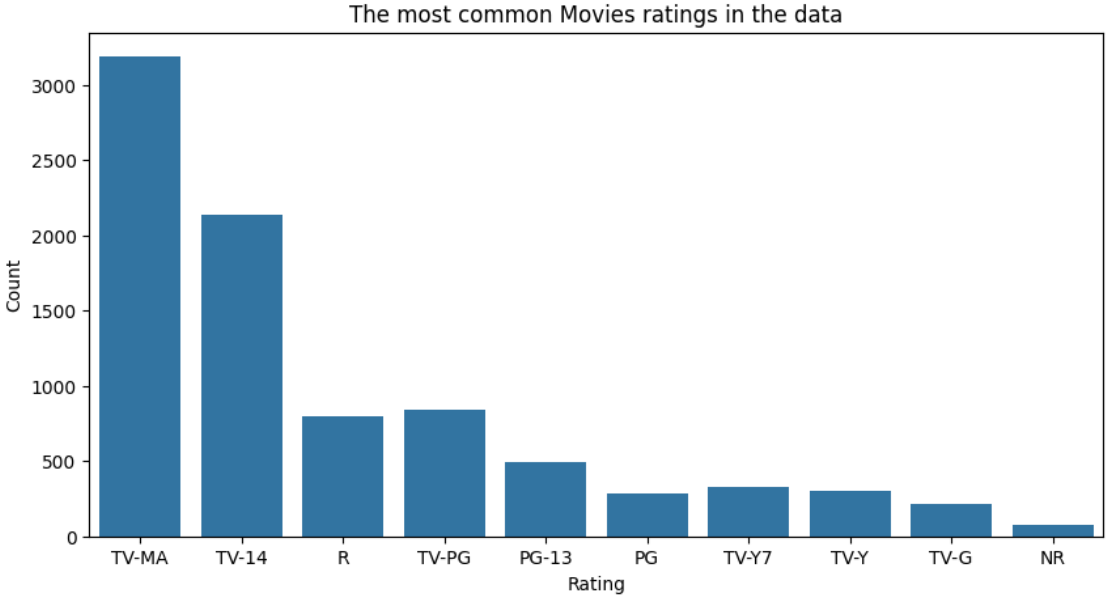



4. What are the most common content ratings in the dataset Tv-shows and Movies?

[32]:
```
plt.figure(figsize=(10,5))
sns.countplot(x='rating', data=data, order=tv_shows['rating'].value_counts().
  ↪index)
plt.xlabel('Rating')
plt.ylabel('Count')
plt.title('The most common Tv-shows ratings in the data')
plt.show()
```




[33]:
```
plt.figure(figsize=(10,5))
sns.countplot(x='rating', data=data, order=Movies['rating'].value_counts().
  ↪iloc[:10].index)
plt.xlabel('Rating')
plt.ylabel('Count')
plt.title('The most common Movies ratings in the data')
plt.show()
```



5. Which genres are the most popular in terms of content type [Tv-show and Movies] count?

[34]:
```
data['genre'].head(5)
```

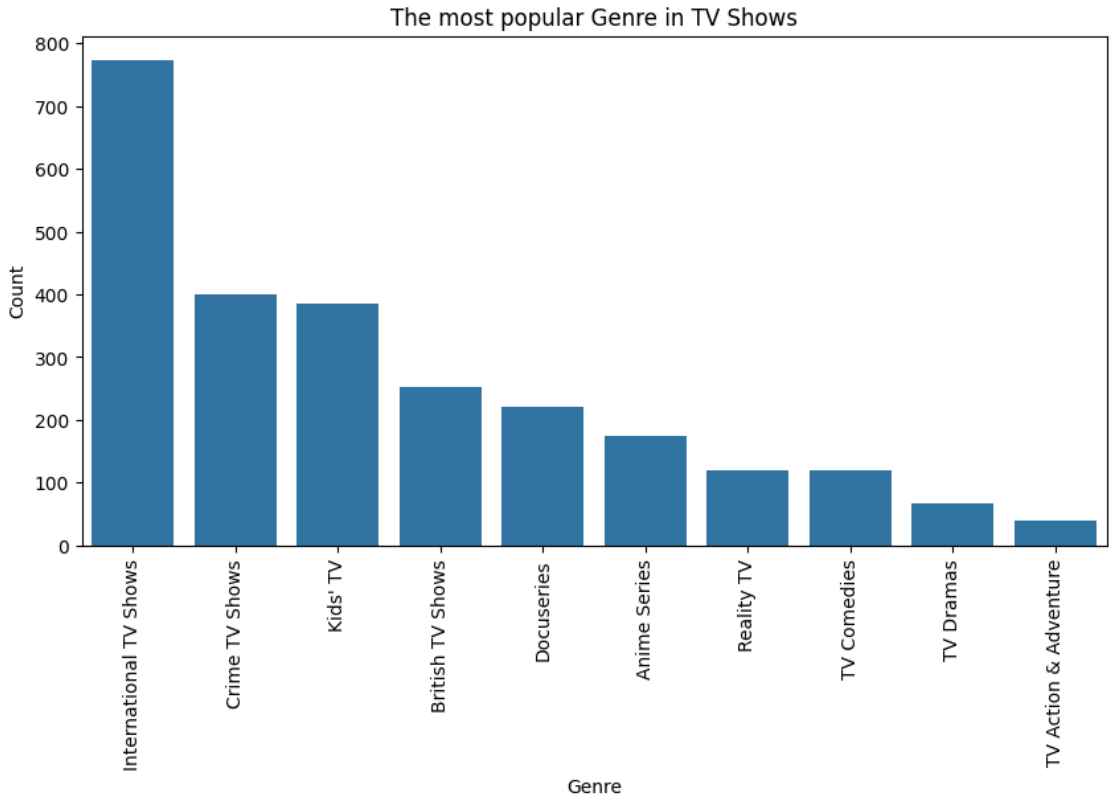
[34]:
```
0                                          Documentaries
1        International TV Shows, TV Dramas, TV Mysteries
2      Crime TV Shows, International TV Shows, TV Act…
3                                Docuseries, Reality TV
4      International TV Shows, Romantic TV Shows, TV …
Name: genre, dtype: object
```

[35]:
```
# Apply lambda function to get the first genre
data['genre'] = data['genre'].apply(lambda x: x.split(",")[0])
```

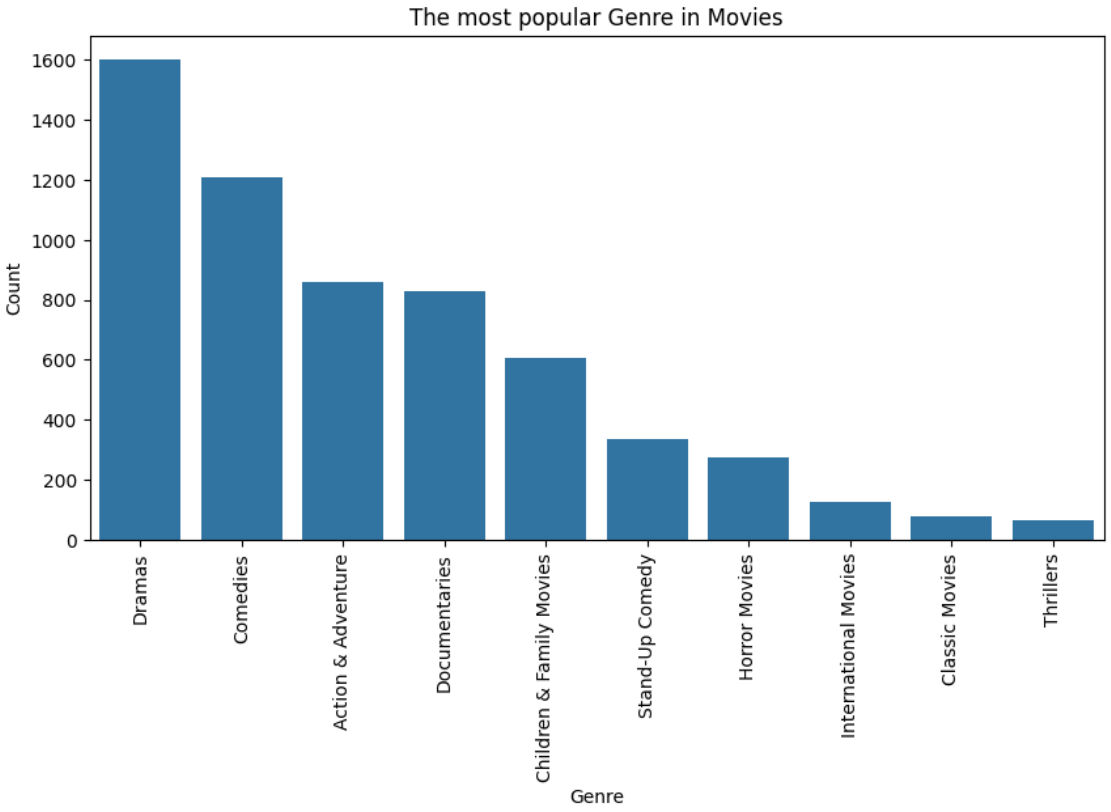
[36]:
```
data['genre'].head(5)
```

[36]:
```
0             Documentaries
1    International TV Shows
2             Crime TV Shows
3                 Docuseries
4    International TV Shows
Name: genre, dtype: object
```

[37]:
```
plt.figure(figsize=(10,5))
tv_shows['genre'] = tv_shows['genre'].str.split(',').str[0]
sns.countplot(x='genre', data=tv_shows, order=tv_shows['genre'].value_counts().
 ↪iloc[:10].index)
plt.xticks(rotation=90)
plt.xlabel('Genre')
plt.ylabel('Count')
plt.title('The most popular Genre in TV Shows')
plt.show()
```

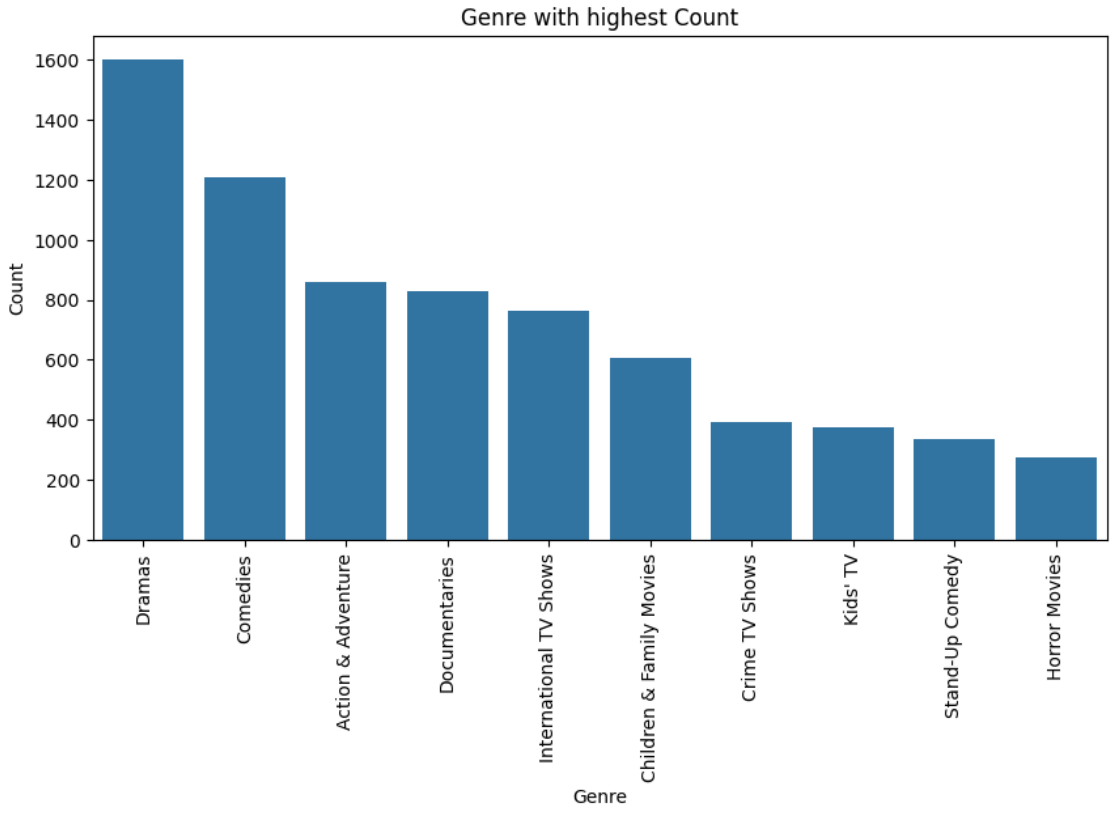



[38]:
```
plt.figure(figsize=(10,5))
Movies['genre'] = Movies['genre'].str.split(',').str[0]
sns.countplot(x='genre', data=Movies, order=Movies['genre'].value_counts().
  ↪iloc[:10].index)
plt.xticks(rotation=90)
plt.xlabel('Genre')
plt.ylabel('Count')
plt.title('The most popular Genre in Movies')
plt.show()
```




6. Which genre has highest number of counts in the dataset?

[39]:
```
plt.figure(figsize=(10,5))
sns.countplot(x='genre',data = data, order= data['genre'].value_counts().iloc[:
↪10].index)
plt.xticks(rotation=90)
plt.xlabel('Genre')
plt.ylabel('Count')
plt.title('Genre with highest Count')
plt.show()
```



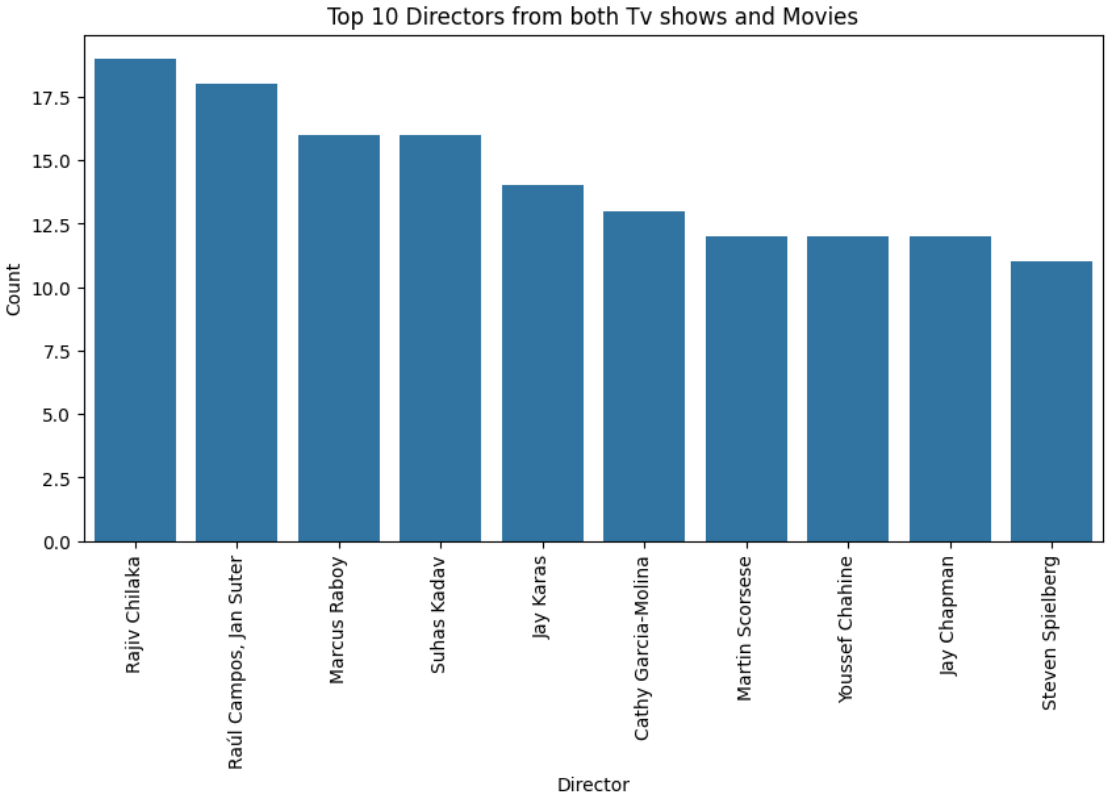

7. List the top 10 directors in the sheet, considering both TV shows and movies.

[40]:
```
data.head(2)
```

[40]:
```
  show_id     type                  title         director  \
0      s1    Movie  Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water          unknown

                                                cast        country  \
0                                            unknown  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban…   South Africa

  date_added  release_year rating   duration                     genre  \
0 2021-09-25          2020  PG-13     90 min             Documentaries
1 2021-09-24          2021  TV-MA  2 Seasons  International TV Shows

                                        description  month  year
0  As her father nears the end of his life, filmm…      9  2021
1  After crossing paths at a party, a Cape Town t…      9  2021
```

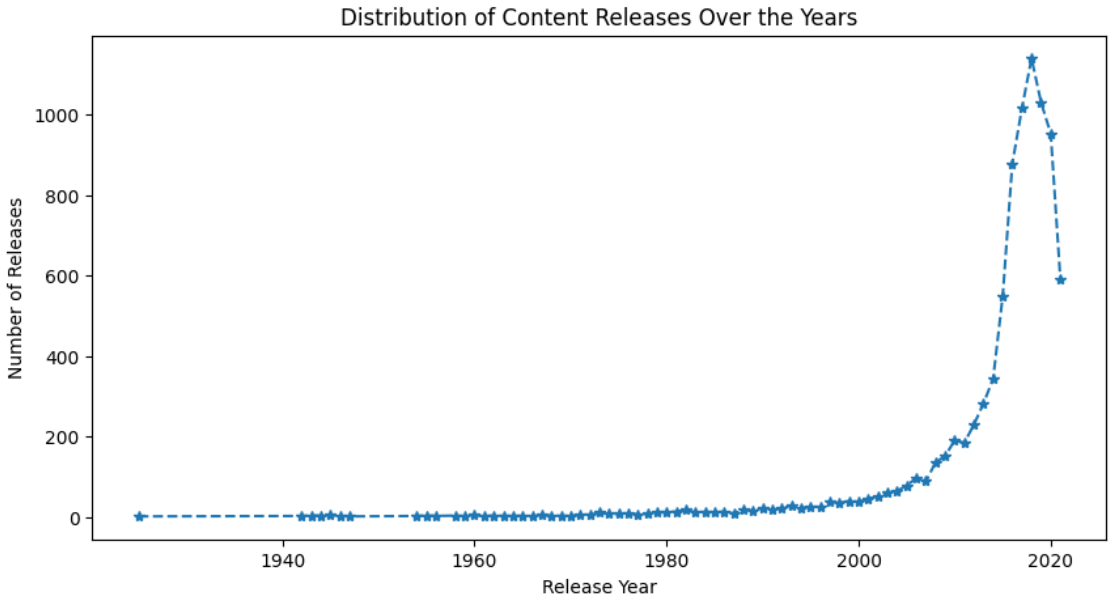
[41]:
```
plt.figure(figsize=(10,5))
sns.countplot(x='director',data = data, order= data['director'].value_counts().
  ↪iloc[1:11].index)
```

```
plt.xticks(rotation=90)
plt.xlabel('Director')
plt.ylabel('Count')
plt.title('Top 10 Directors from both Tv shows and Movies')
plt.show()
```




8. How has the distribution of content releases changed over the years?

[42]:
```
counts = data['release_year'].value_counts().sort_index()
plt.figure(figsize=(10,5))
plt.plot(counts.index, counts.values, '--*')
plt.xlabel('Release Year')
plt.ylabel('Number of Releases')
plt.title('Distribution of Content Releases Over the Years')
plt.show()
```




## 5 Conclusion

1. Content Releases by Year: In 2019, there was a surge in movie releases, while in 2020, there was a notable increase in TV show releases. This indicates a shift in content strategy over these years.

2. Distribution of Content Types: Approximately 29.6

3. Top Countries for Content Releases:The United States stands out as the top country for content releases in both TV shows and movies, indicating a strong presence in the American market.

4. Common Content Ratings: The most common content rating across both TV shows and movies is TV-MA (Mature Audience Only). This suggests that Netflix caters to a mature audience with content containing elements like graphic violence, explicit sexual activity, or indecent language.

5. Popular Genres: For TV shows, the most popular genre is "International TV Shows," while for movies, it is "Dramas." This information can help in content curation and recommendation algorithms.

6. Top Director(s): The top directors based on the dataset are Rajiv Chilaka, Jan Suter, and Marcus Raboy. Collaborations with these directors may yield successful content.

7. Content Releases Over the Years: The graph shows a peak in content releases during 2020, indicating a potential impact of global events (like the COVID-19 pandemic) on content production and consumption.

## 6 Suggestions for Growth

1. Diversify Content Library: While movies dominate, there's room for growth in the TV show category. Expanding and diversifying the TV show library can attract a wider audience.

2. Global Expansion: Since the United States is the primary market, focusing on international content and expanding into new regions can lead to significant growth.

3. Targeted Content for Mature Audiences: Given the popularity of TV-MA content, creating more tailored content for mature audiences can be a growth opportunity.

4. Collaborations with Top Directors: Further collaborations with top directors like Rajiv Chilaka, Jan Suter, and Marcus Raboy can lead to successful projects.

# 7 Predictions

1. Continued Emphasis on Original Content: - Netflix is likely to continue investing in original content to differentiate itself in a competitive market.

2. Increased Global Content: - Expect more international content releases as Netflix expands its global presence.

3. Diverse Genre Offerings: - Netflix may continue to diversify its genre offerings to cater to a wider audience.

4. Innovations in Viewer Experience: - Predictions may include innovations in viewer experience, potentially through interactive content or augmented reality.

# 8 Thank You by Navjoth Singh