Netflix Data Analysis

Netflix is a global streaming service platform that offers a wide variety of movies, TV shows, documentaries, and more across a wide range of genres and languages. Netflix uses advanced streaming technology to deliver high-quality video content across different devices, including smart TVs, smartphones, tablets, and computers.

Importing Python Libraries:

Import libraries Pandas and NumPy for data manipulation, numerical operations and analysis. Data visualization libraries like Matplotlib, Seaborn and Plotly.

```
[2]: import numpy as np
     import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     import plotly.express as px

[3]: import warnings
     warnings.filterwarnings('ignore')
```

Dataset:

Using Netflix Dataset for EDA Project. This Dataset contains 8807 unique TV Shows and Movies. This column has different columns.

- `show_id`: Unique identifier for each show.
- `type`: Type of content (e.g., Movie, TV Show).
- `title`: Title of the show.
- `director`: Director of the show.
- `cast`: Main cast members.
- `country`: Country where the show was produced.
- `date_added`: Date when the show was added to Netflix.
- `release_year`: Year the show was released.
- `rating`: Rating of the show (e.g., PG, R).

```
[5]: 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
```

- `duration`: Duration of the show (e.g., number of seasons for TV shows, minutes for movies).
- `listed_in`: Genres the show is listed under.
- `description`: Brief description of the show.

```
[10]: df.describe()
```

[10]:

	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

We can see the mean, min, max of release_year.

Dealing with Null Values:

First, identify which columns have null values. We have 2634, 825, 831,10 null values in director, cast, country, date_added and 4,3 null values in rating and duration.

```
[9]: df.isnull().sum()
```

```
[9]: 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
```

- **Decide on a strategy**: drop or fill.
- **Drop rows/columns** if null values are insignificant or irrelevant.
- **Fill null values** with specific values or statistical values according as per need.

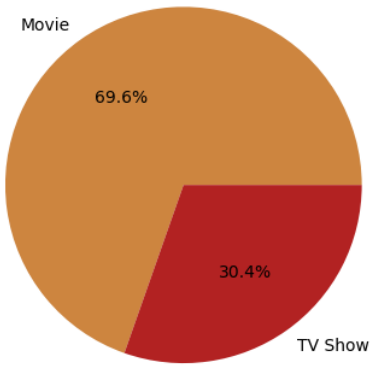
As shown in picture we change all null values of cast, country, date_added, rating with 'Unknown'.

```
[12]: df.director.fillna(value='Unknown',inplace=True)
      df.director
```

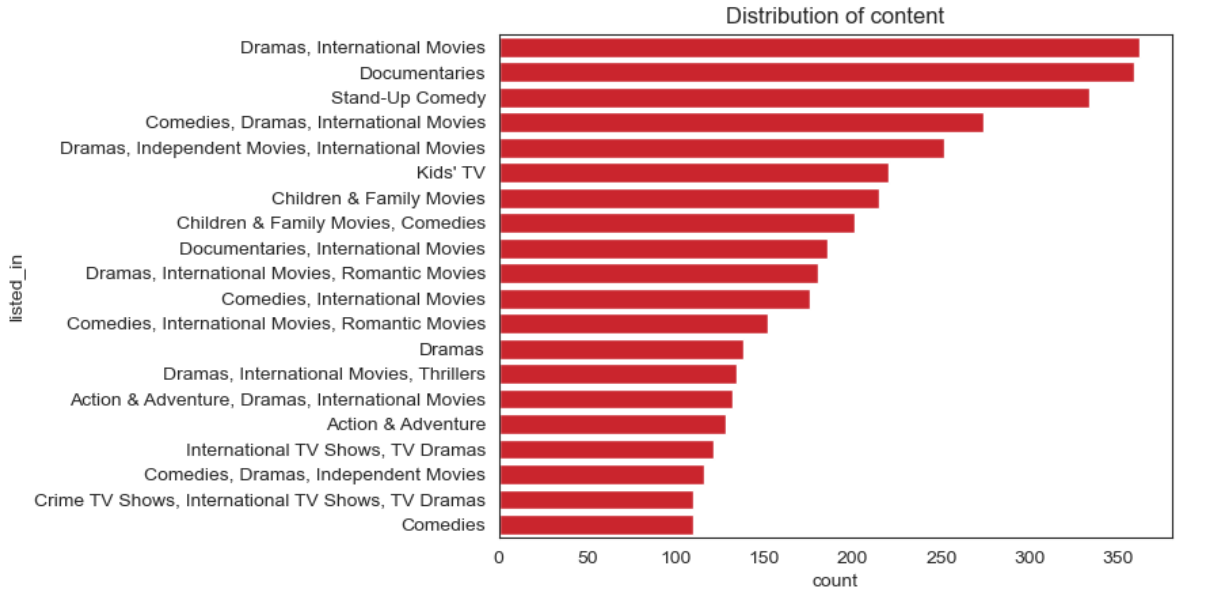
```
[12]: 0       Kirsten Johnson
      1               Unknown
      2       Julien Leclercq
      3               Unknown
      4               Unknown
                    ...
      8802      David Fincher
      8803            Unknown
      8804    Ruben Fleischer
      8805       Peter Hewitt
      8806        Mozez Singh
      Name: director, Length: 8807, dtype: object
```

Data Visualization

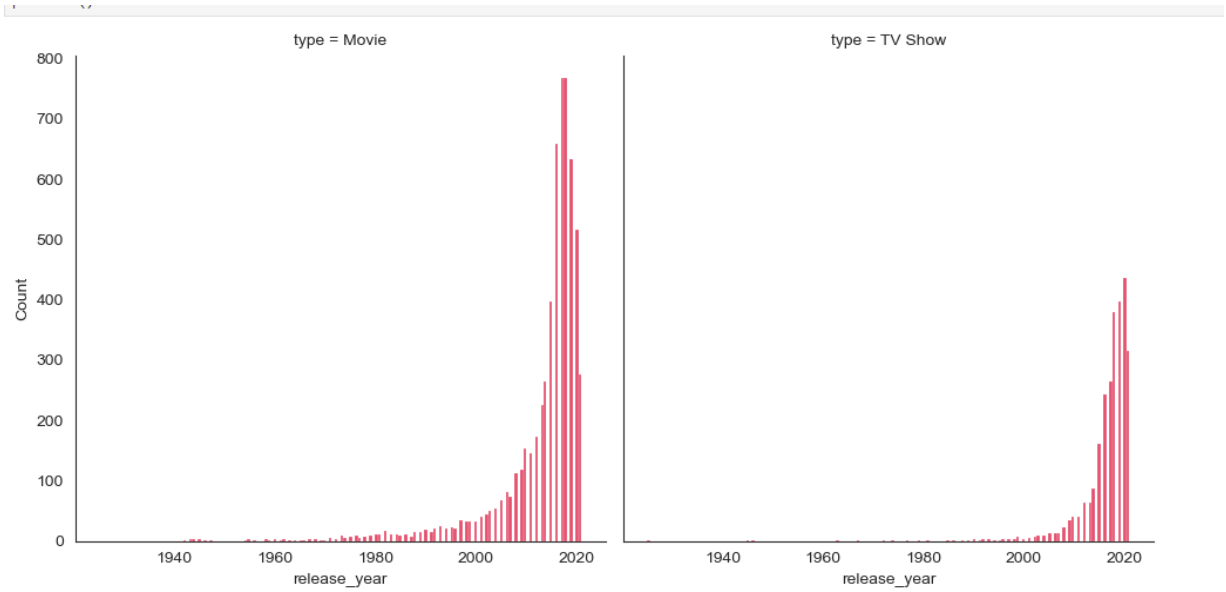
1. Content Type on Netflix: Netflix offers a significant number of both movies and TV shows, allowing users to choose from a diverse range of content types. We can observe that percentage of movies is 69.6% which is more than TV Show with percentage of 30.4%. As per audience choice movies are more famous.



2. Create visualizations to represent the distribution of content over different genres: Popular genres on Netflix include Drama, Internatinal Movies, Standup Comedy, Documentary, and others, providing insights into viewer preferences and content diversity. We observe that genres like Drama, International Movies and Comedy are highly represented niche genres with count of 350. This chart also displays the average ratings for each genre like children, family and romantic movies with count range from 170 to 200 and shows crime and international TV shows, TV dramas and comedies are less in production as compare to other content as per viewer interest with count of 100. Netflix should make most of content in top genres to scale the business.



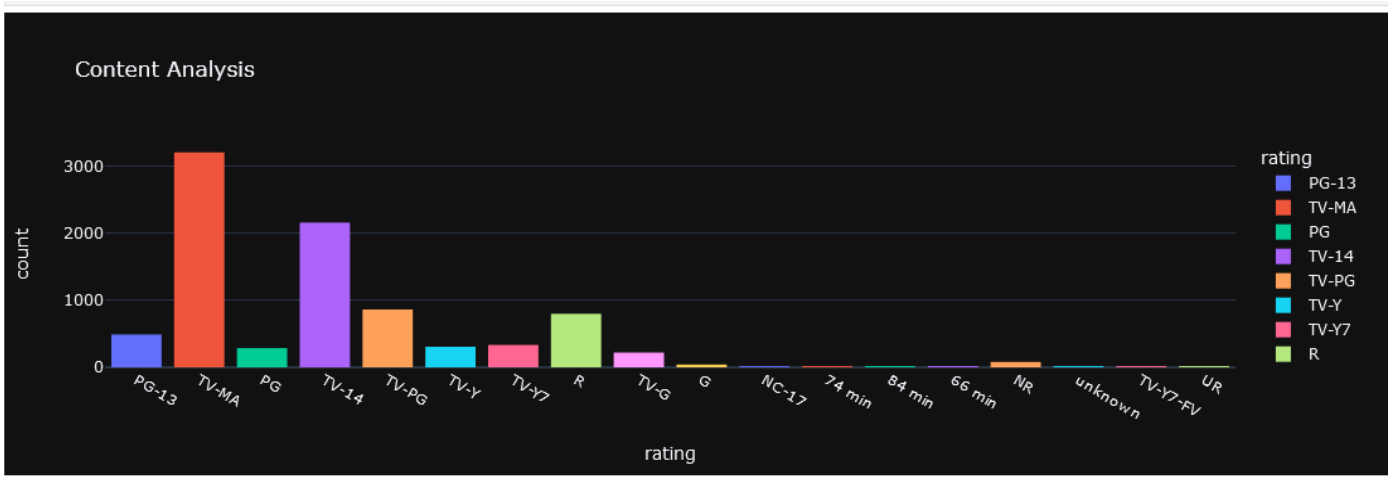
3. Visualize the distribution of content across release years: The line shows a steady upward trend, it indicates that the number of shows or movies released each year is increasing and increase in quantity. From 1940-2000 is increase in content is very less but after 2000 there is massive increase in making content and we can say 2017 year of Netflix with highest content making year. Before 2000 TV Shows are rarely made but after 2000 making TV Shows is also increase.



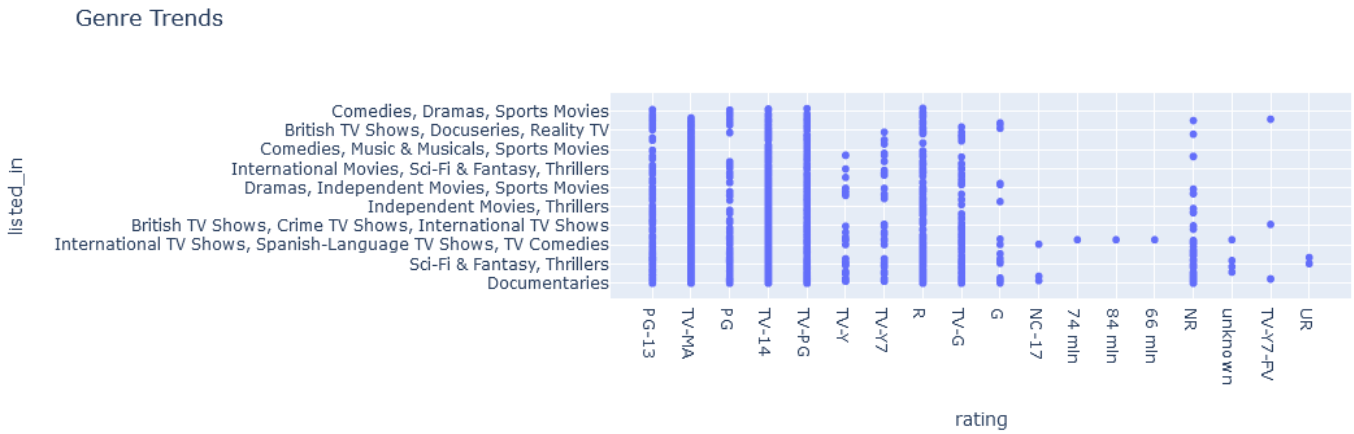
4. Analyze the distribution of content ratings:

- **TV-MA (Mature Audience Only)**: Suitable for mature audiences only. May contain very strong language, strong sexual content, or graphic violence.
- **TV-14**: Parents strongly cautioned. May contain some material parents might find unsuitable for children under 14 years of age.
- **TV-PG**: Parental guidance suggested. May contain some material parents might find unsuitable for younger children.
- **TV-Y7**: Directed to older children. Suitable for children 7 and older. May include mild violence or themes.
- **TV-Y**: Suitable for all children. This program is designed to be appropriate for all children.
- **TV-G (General Audience)**: Suitable for all ages. Contains little or no violence, no strong language, and little or no sexual dialogue or situations.
- **G (General Audience)**: Suitable for all ages. Contains nothing in theme, language, nudity, sex, violence, or other matters that would offend parents whose younger children view the motion picture.
- **PG (Parental Guidance)**: Parental guidance suggested. Some material may not be suitable for children.
- **PG-13**: Parents strongly cautioned. Some material may be inappropriate for children under 13.
- **R (Restricted)**: Restricted Contains some adult material. Parents are urged to learn more about the motion picture before taking their children to see it.
- **NC-17 (Adults Only)**: Adults Only. No one 17 and under admitted.
- **NR (Not Rated)**: Not rated by the Motion Picture Association.

It reveals TV-MA are most popular in audience preferences with rating of 3000+ followed by TV-14 with rating of 2100 then TV- PG with rating of 800 and R with 700. At G, NC-17, TV- Y7-FV are lowest in terms of rating. Understanding the spread of ratings helps Netflix target specific viewer demographics more effectively.



5. Analyze trends in the popularity of different genres over time: It provides information about different genres like comedies, Drame, Sports, Movies, British TV shows, documentaries according to ratings who has the highest rating according to different rating categories. We can see that PG-13 rating have more genres like romantic, horror, action movies and comedy. TV ma has more varieties Like science fiction and fantasy thriller adventures genres. We can see UR, unknown, NC 17 etc, have less listed in genres as compared to other ratings.



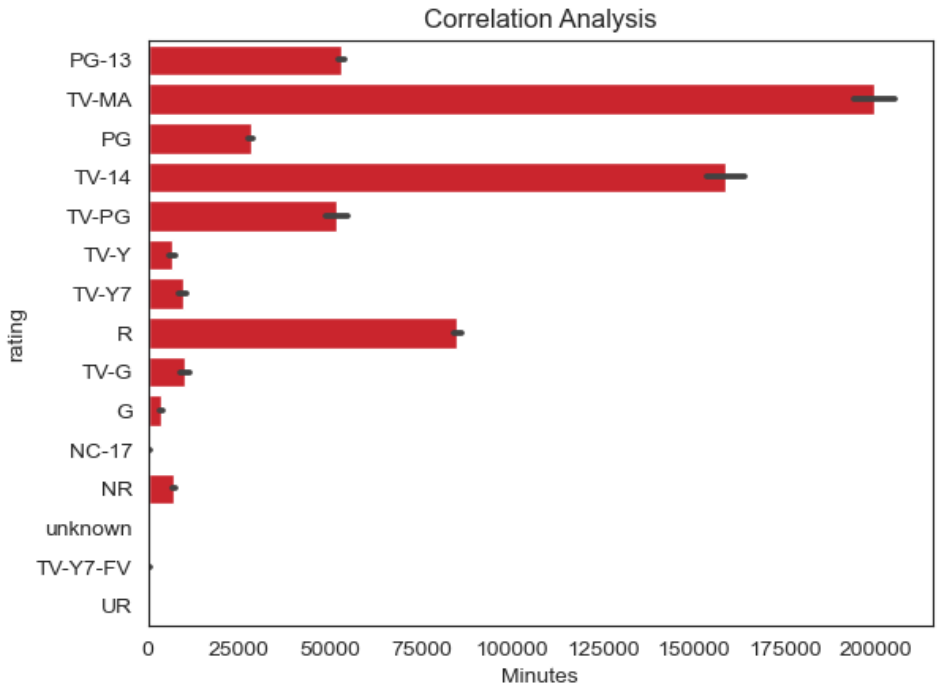
Investigate potential correlations between variables (e.g., ratings and duration):

Firstly, we split duration as per our need, numbers in one column and stings another column. After that we will drop Null values and convert Minutes column in int type. Then do correlation analysis between rating and Minutes.

```
df[['Minutes','Unit']] = df['duration'].str.split(' ', expand =True)
```

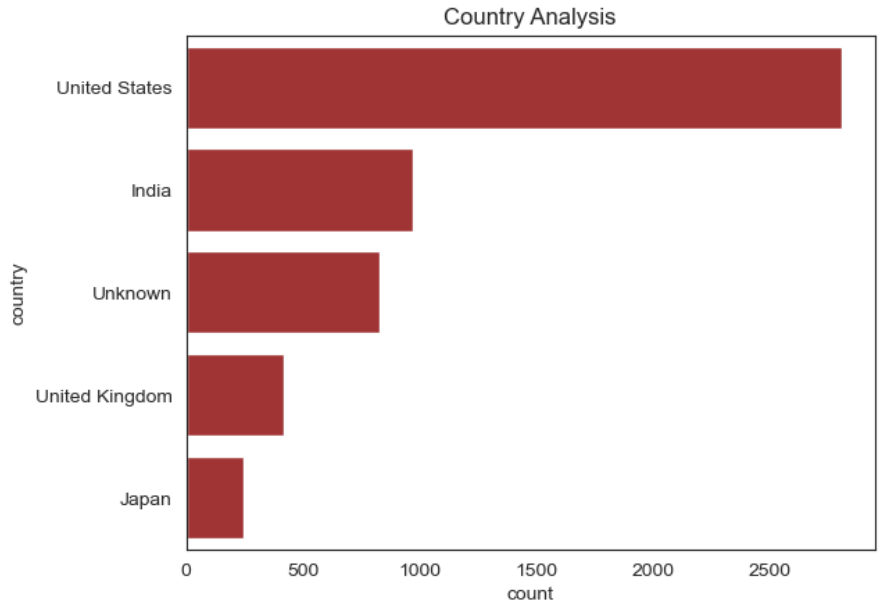
```
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	Minutes	Unit
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	90	min
1	s2	TV Show	Blood & Water	Unknown	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	Seasons
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown	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...	1	Season



It reveals different categories and their duration in minutes PG -13, PG, TV- PG have 30000 to 50000 and the highest duration is TV-MA, TV-14 and R rating categories. G, NC-17, UR With lowest duration.

6. Explore the geographical distribution of content: Netflix content originates from various countries, with the USA leading in production of 2500+, followed by other countries like India with 1000+ and the UK with 800, Japan indicating Netflix's global reach and diversity in content sourcing. These are the top 5 countries which make majority of content.



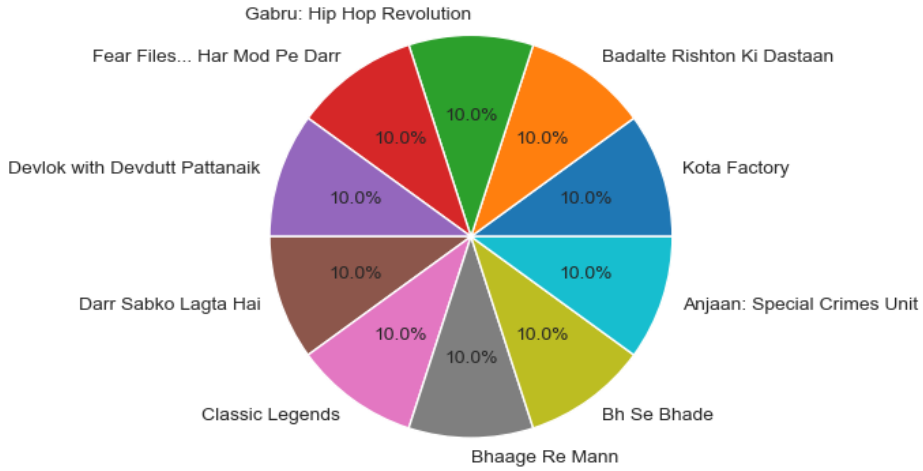
7. Title of TV Show that were released in India only: Firstly, using temp we separate TV Show which are made in India.

```
[47]: temp2 = df[(df['type']=='TV Show') & (df['country']=='India')]['title']

[49]: temp2.value_counts().head(10)

[49]: title
      Kota Factory                    1
      Badalte Rishton Ki Dastaan      1
      Gabru: Hip Hop Revolution       1
      Fear Files... Har Mod Pe Darr   1
      Devlok with Devdutt Pattanaik   1
      Darr Sabko Lagta Hai            1
      Classic Legends                 1
      Bhaage Re Mann                  1
      Bh Se Bhade                     1
      Anjaan: Special Crimes Unit     1
      Name: count, dtype: int64
```

This approach allows you to identify specific TV shows that were released exclusively in India like Kota Factory, Devlok with Devdutt Pattnayak, Classic Legend, Bhag re mann, Dar Sabko Lagta Hai etc.



Conclusion

Analyzing a Netflix CSV dataset provides valuable insights into the types of content available, trends in release dates, popular genres, geographic distribution, and content duration. These insights can inform decisions related to content acquisition, audience targeting, and platform strategy, enhancing the overall viewer experience and platform performance.