Data Cleaning, Analysis and Visualization

Using python

- NAHEEDA PARVEEN

Netflix Data Analysis

Data Structure

The Netflix dataset consist of 8790 rows and 10 columns. The columns are show id, type, title, director, country, date added, listed in , year, month , day, rating , duration, release year

Data Overview

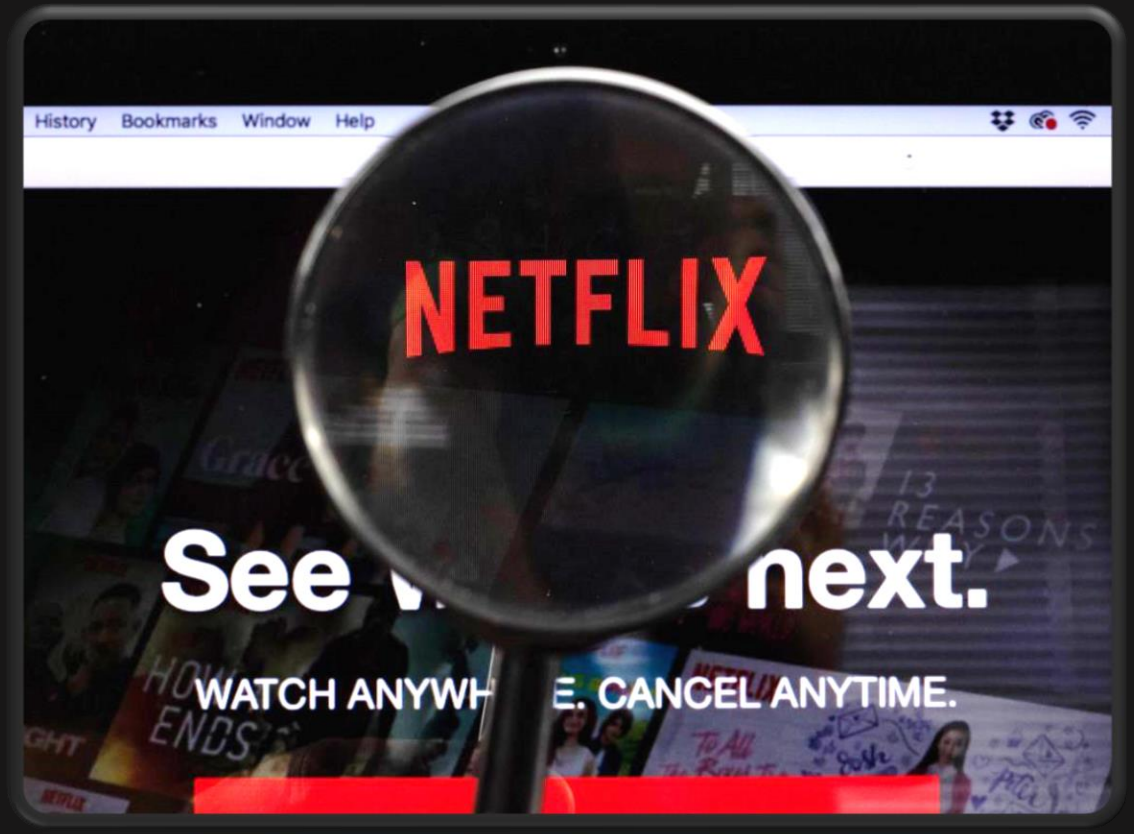
- The head() function displays the first 5 records of the dataset. And describe() function shows the description of the table of the dataset.
- The shape() function shows the total rows and columns in the dataset and the size() function shows the particular column shows the total size of the records in the dataset.

Introduction

This project involves loading, cleaning, analysing, and visualizing data from a Netflix dataset. Using some crucial libraries from python to focus on the data analysis process.

Objectives

- ❑ Import the libraries
- ❑ Load the dataset
- ❑ Handling the data
- ❑ Implementing Exploratory Data Analysis(EDA)
- ❑ Drive strategic Decision making



Scope

This project analyses the Netflix dataset using python libraries such as pandas, matplotlib to explore and gain insights from this content

Workflow


Data collection and Loading
Data cleaning and Transformation
Data storage and Management
Data Analysis and Insights

Data Collection and Loading

Data sources

- The Netflix dataset contains the information of the movies and tv shows on the Netflix , directors, genres, countries, duration and ratings of the shows in it.

Data loading process

- Import the panda library into python
- The dataset loaded as data frame in python using the pandas data frame as pd.read_csv() function.


Data Cleaning

Improvements in Netflix Dataset

- Identifying and remove the null values using the is null() function in python
- Identifying and removing the duplicate values from the dataset using drop. Duplicates() function
- Identifying the missing rows and drop the rows using dropna() function
- Converting the data column to datetime datatype using pd.to_datatime() function in pandas.
- Creating a new column if needed as per data analysis
- Ensuring the data quality before proceeding to the process of visualizations and explorations.

Data Loading and processing

- Importing the libraries from python
- Reading the Netflix dataset using the pandas function as given below

Import Required Libraries

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
```

Load Dataset

In [3]:
```
data=pd.read_csv('netflix1 - netflix1.csv')
data
```

Out[3]:

	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	Dramas, Independent Movies, International Movies

Data Cleaning

- Identifying the null values from the dataset
- Dropping duplicate values from the dataset
- And displaying the dataset using head() function

```
print(nf.isnull().sum())  # Identifying null values from the dataset
```

```
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
```

```
[10] nf.drop_duplicates(inplace=True)  #Droping duplicate values from the dataset
```

```
nf.head() # Displaying the value
```

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in	year	month	day
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 min	Documentaries	2021	9	25
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	2021	9	24
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	2021	9	24
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22	2021	TV-PG	91 min	Children & Family Movies, Comedies	2021	9	22
4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	2021	9	24

Data Cleaning

- ❑ Update the date column to datetime datatype
- ❑ Counting the types of shows in Netflix dataset using values_count() function

```
pd.to_datetime(nf['date_added']) # Converting the datatype to datetime
```

	date_added
0	2021-09-25
1	2021-09-24
2	2021-09-24
3	2021-09-22
4	2021-09-24
...	...
8785	2017-01-17
8786	2018-09-13
8787	2016-12-15
8788	2018-06-23
8789	2018-06-07

8790 rows × 1 columns

dtype: datetime64[ns]

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

```
type
Movie      6126
TV Show    2664
Name: count, dtype: int64
```

Data Cleaning

- Split function is used the split column example date time is splited into day, month and year

```
data['year']=data['date_added'].dt.year
data['month']=data['date_added'].dt.month
data['day']=data['date_added'].dt.day
```

```
data.head()
```

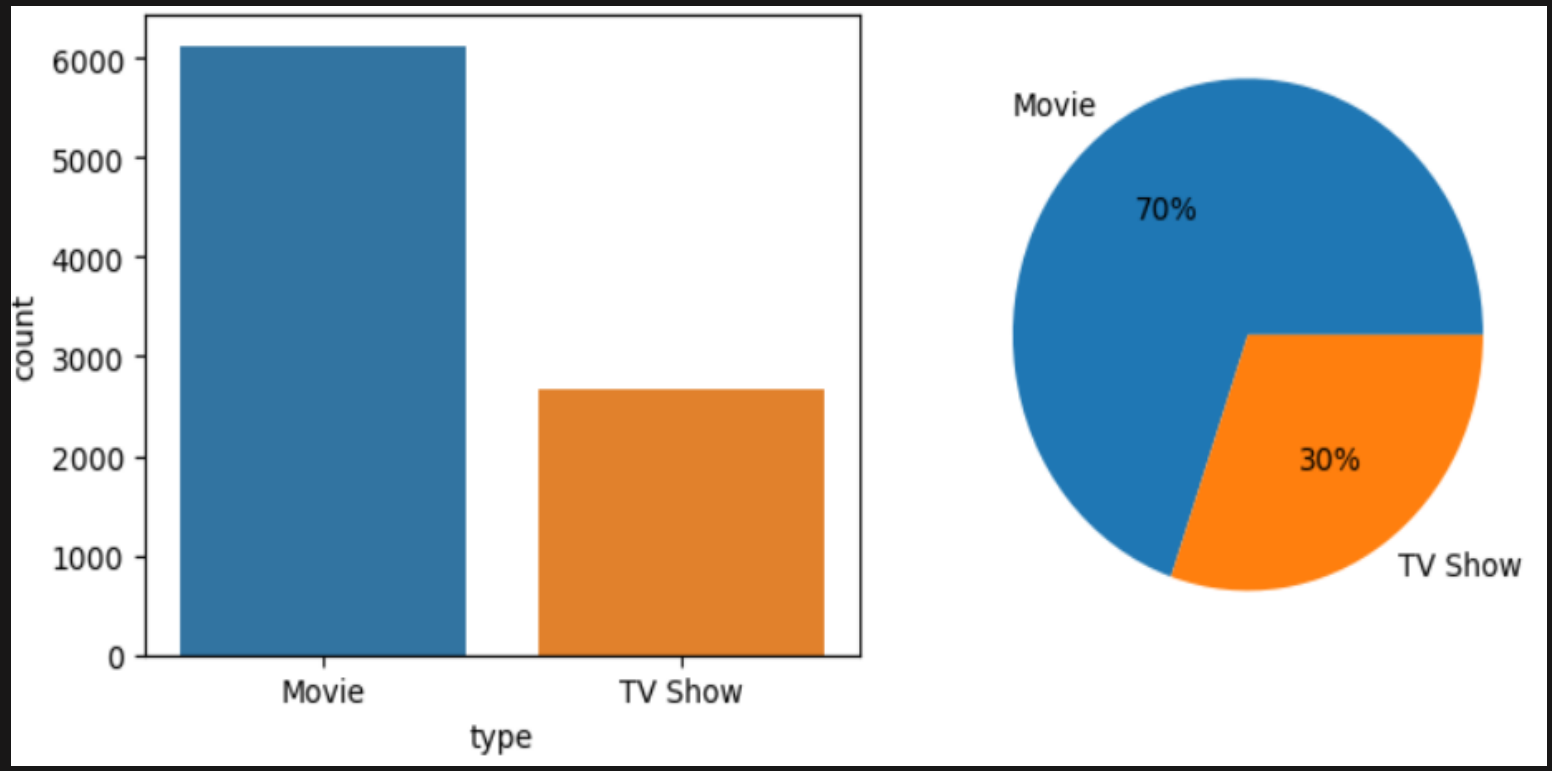
	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in	year_added	year	month	day
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 min	Documentaries	2021	2021	9	25
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	2021	2021	9	24
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	2021	2021	9	24
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22	2021	TV-PG	91 min	Children & Family Movies, Comedies	2021	2021	9	22
4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	2021	2021	9	24

Data Visualization

Total content on Netflix

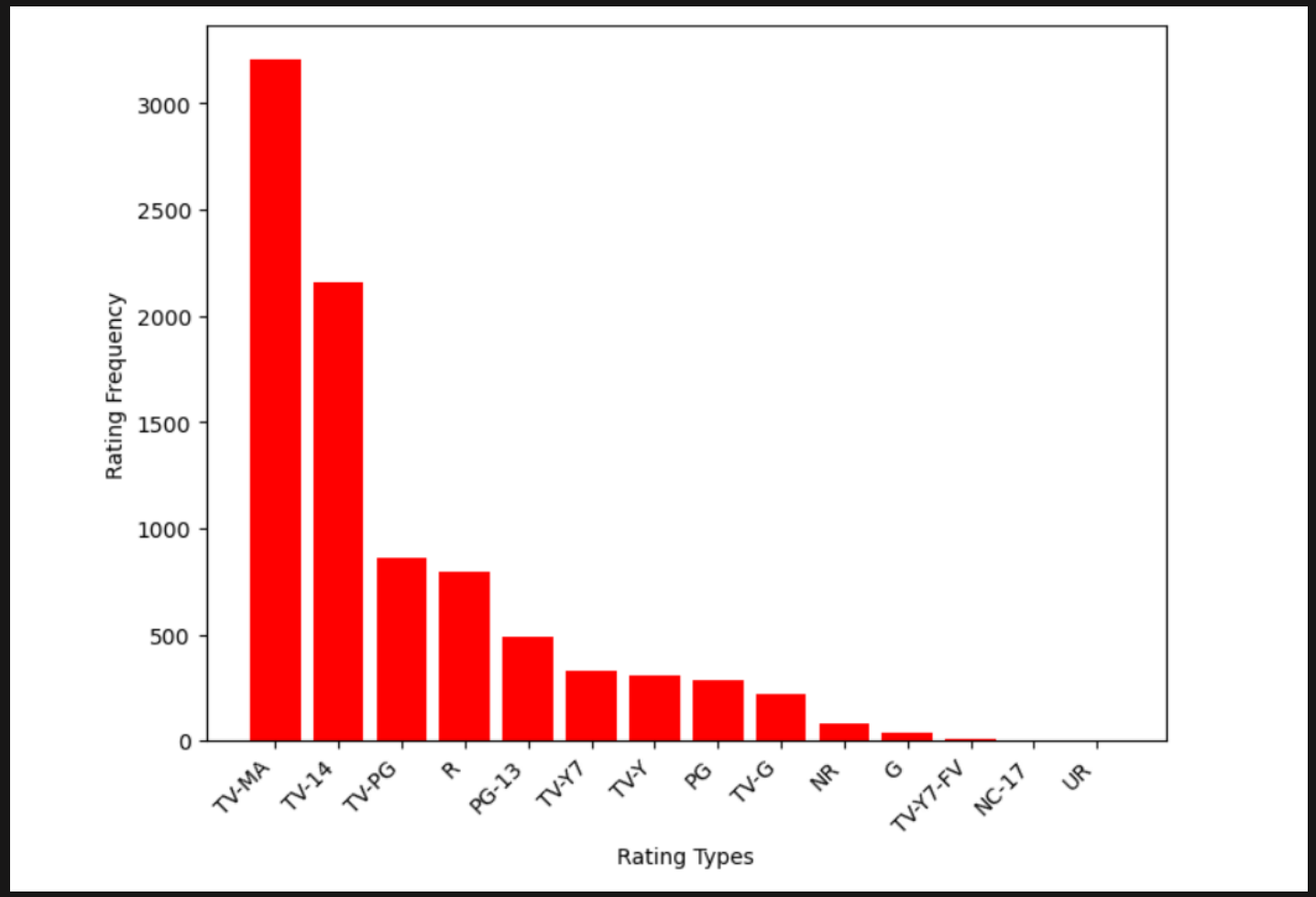
- This bar graph and pie chart shows the content type on the Netflix by counting the values on the movies and tv shows on Netflix

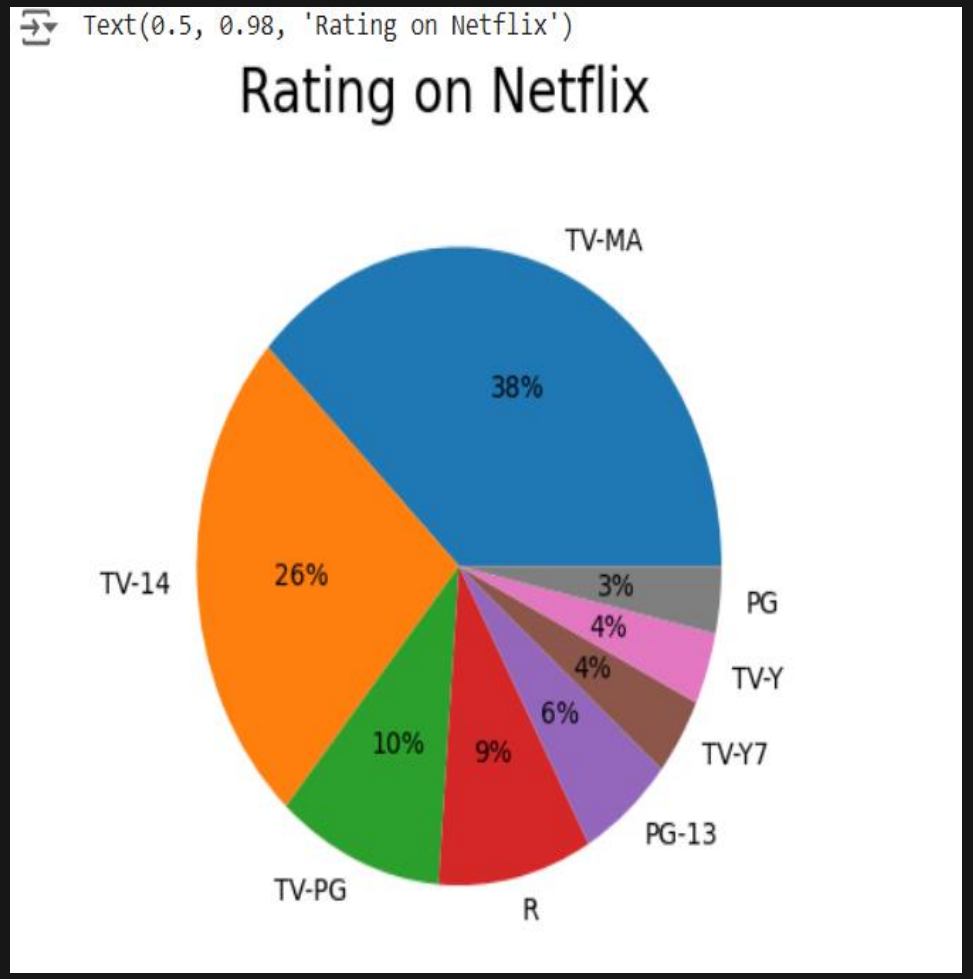
```
freq = data['type'].value_counts()
fig, axes = plt.subplots(1, 2, figsize=(8, 4))
sns.countplot(x='type', data=data, ax=axes[0])
axes[1].pie(freq, labels=freq.index, autopct='%.0f%%')
plt.suptitle('Total Content on Netflix', fontsize=20)
plt.tight_layout()
plt.show()
```



Total Rating on Netflix

- This bar graph and pie chart displays the total count of rating on Netflix dataset based on the rating types.





Top 10 countries with most content on Netflix

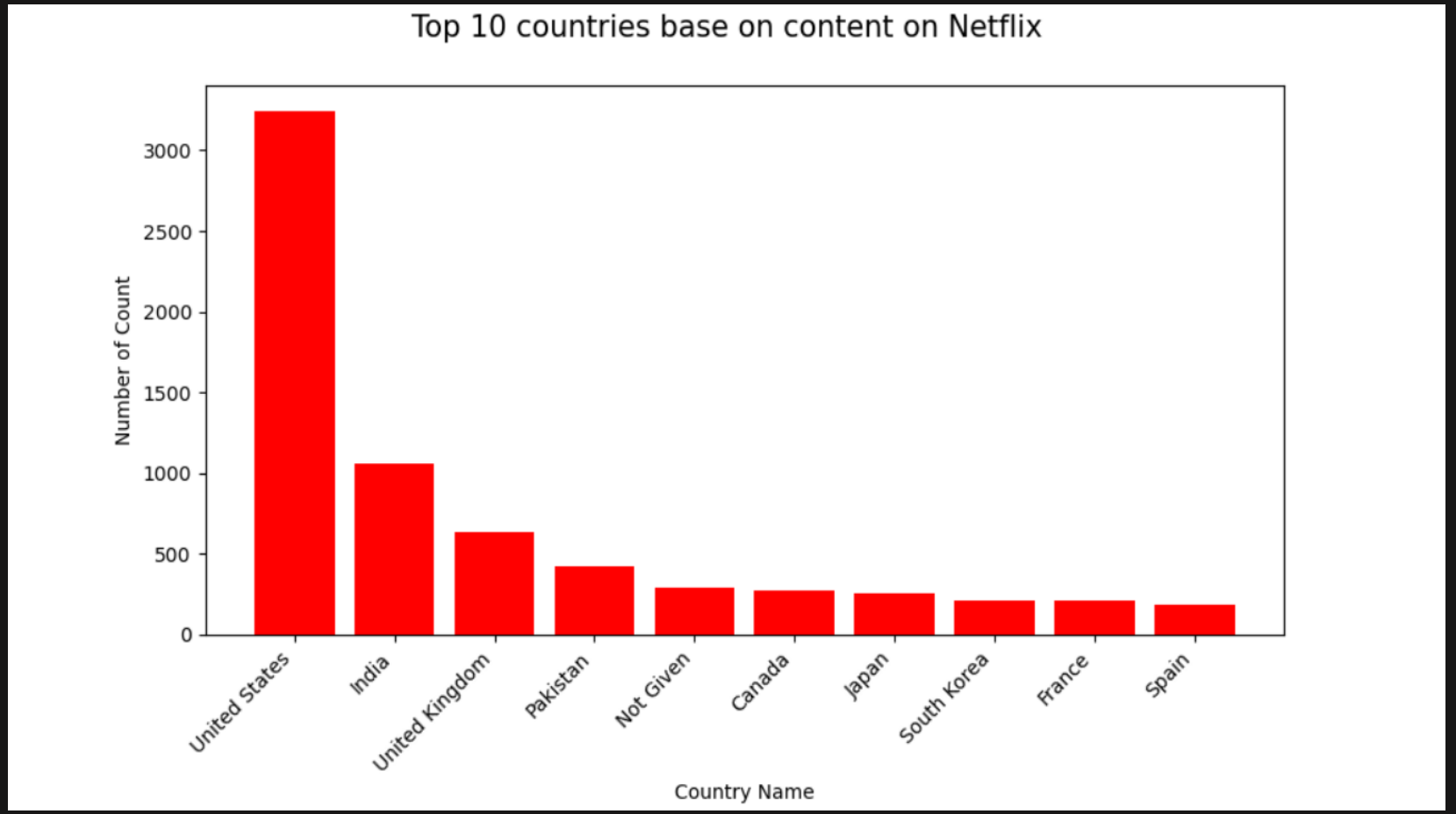
- This bar graph shows the top10 countries with most content on the Netflix dataset by the countries' frequency

```
nf['country'].value_counts() # Total countries values count
```

country	count
United States	3240
India	1057
United Kingdom	638
Pakistan	421
Not Given	287
...	...
Iran	1
West Germany	1
Greece	1
Zimbabwe	1
Soviet Union	1

86 rows × 1 columns

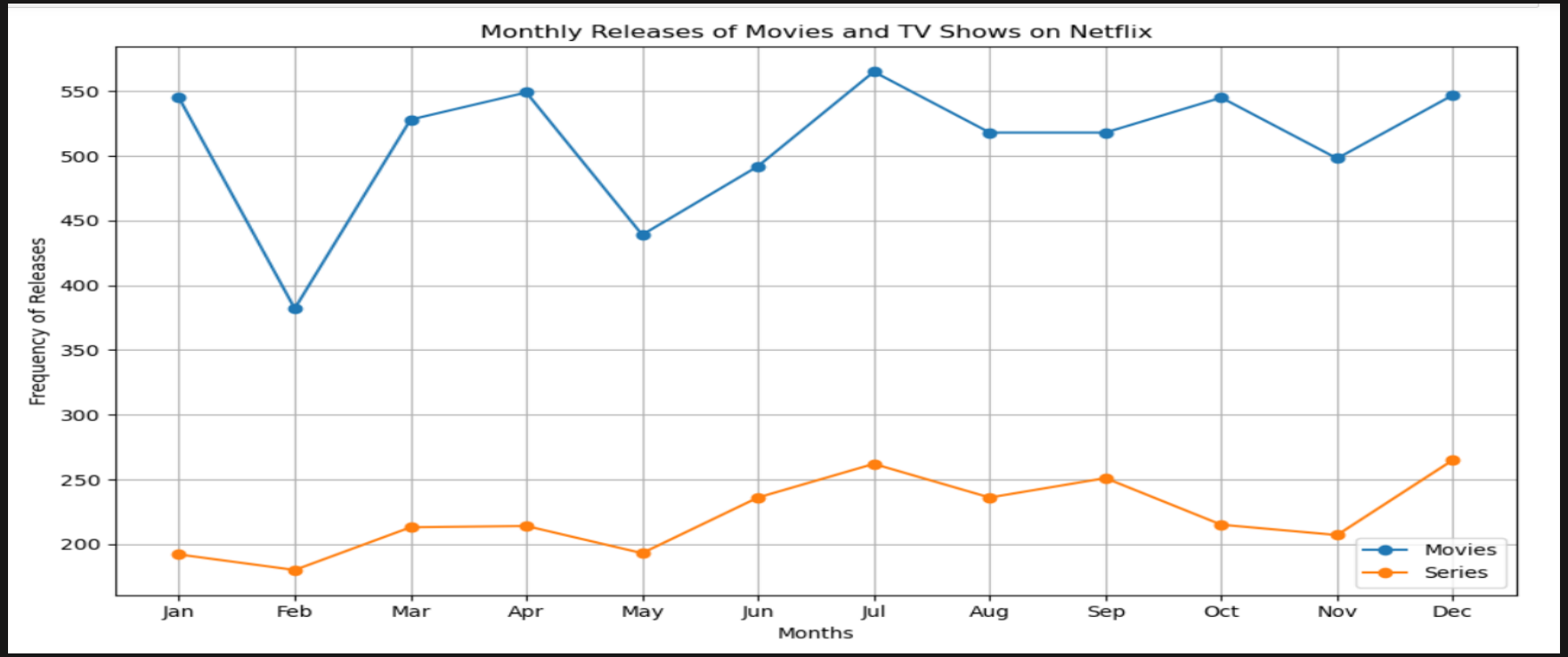
dtype: int64



Monthly releases of Movies and TV shows

- This line graph shows the frequency of the monthly and yearly releases of movies and tv shows on Netflix.

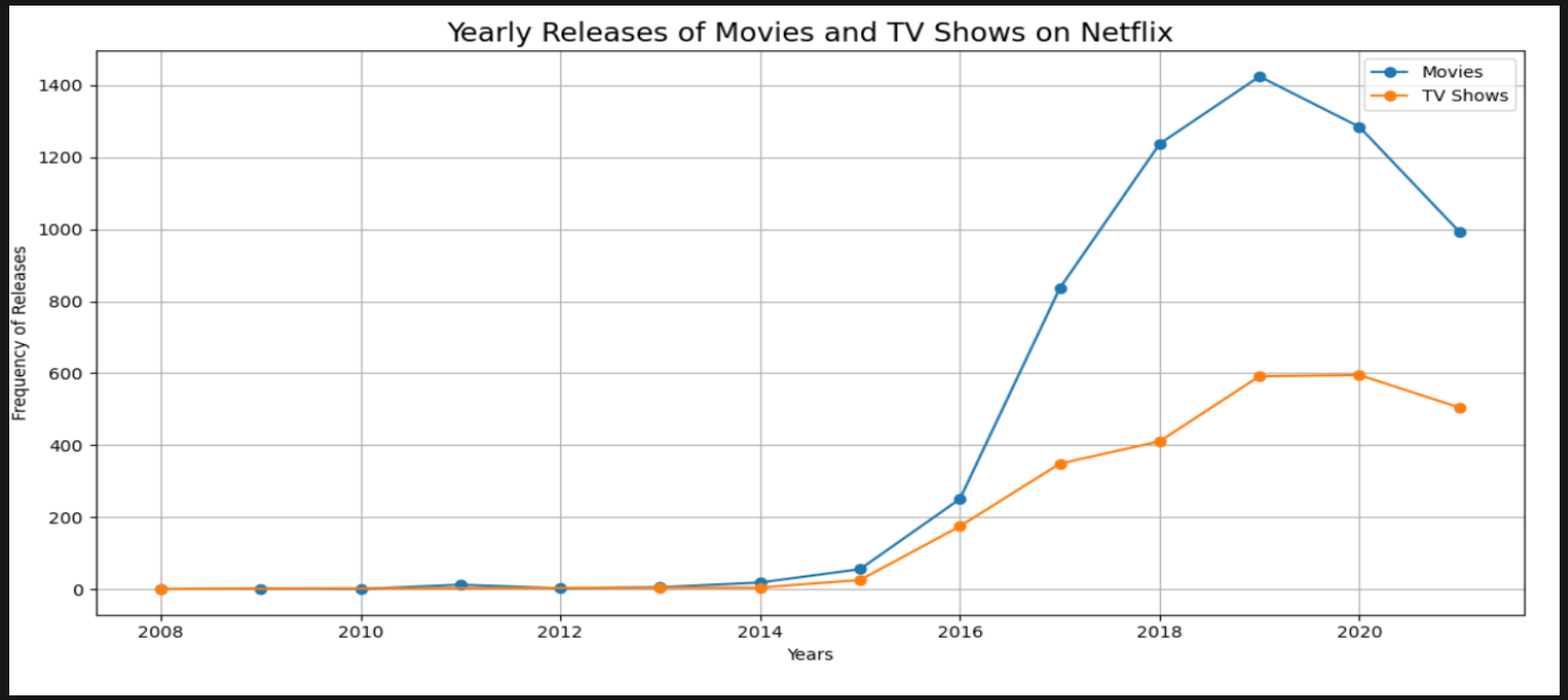
```
# Plot the data
plt.figure(figsize=(10, 6))
plt.plot(monthly_movie_release.index, monthly_movie_release.values, label='Movies', marker='o')
plt.plot(monthly_series_release.index, monthly_series_release.values, label='Series', marker='o')

# Customize the plot
plt.xlabel("Months")
plt.ylabel("Frequency of Releases")
plt.xticks(range(1, 13), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.title("Monthly Releases of Movies and TV Shows on Netflix")
plt.legend()
plt.grid(True)

# Display the plot
plt.tight_layout()
plt.show()
```



Yearly releases of Movies and TV shows

- This line graph shows the frequency of the monthly and yearly releases of movies and tv shows on Netflix.

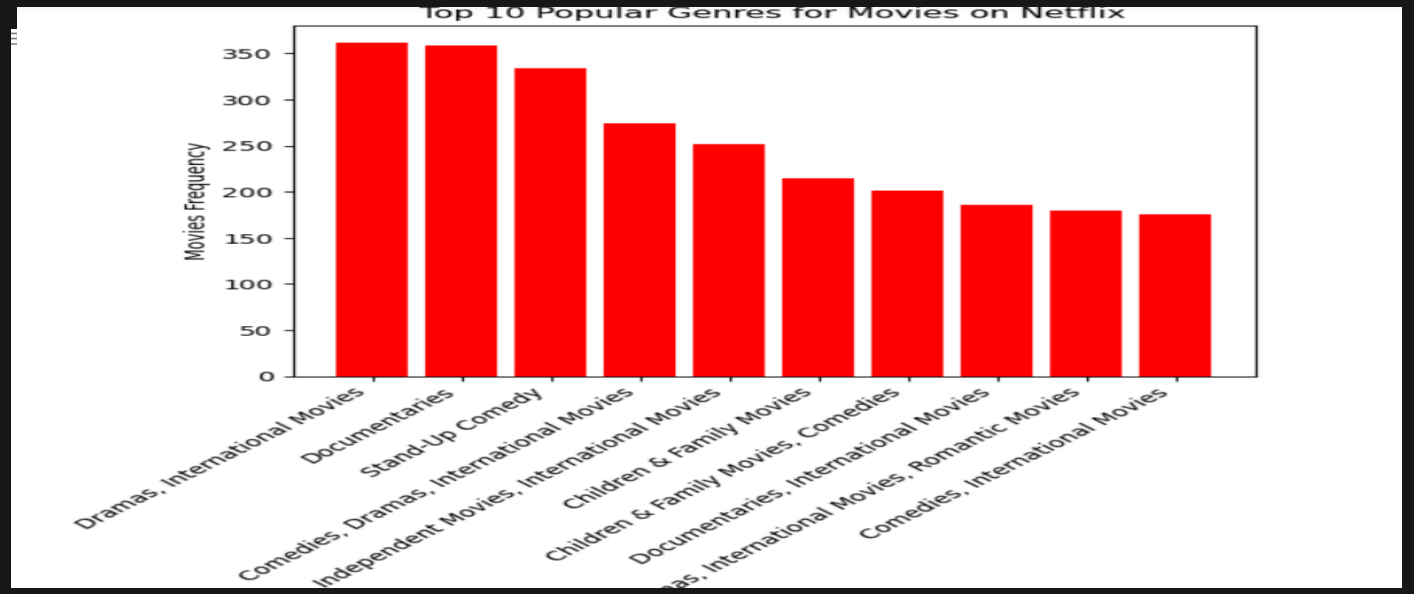
```
# Plot the data
plt.figure(figsize=(12, 6))
plt.plot(yearly_movie_releases.index, yearly_movie_releases.values, label='Movies', marker='o')
plt.plot(yearly_series_releases.index, yearly_series_releases.values, label='TV Shows', marker='o')

# Customize the plot
plt.xlabel("Years")
plt.ylabel("Frequency of Releases")
plt.title("Yearly Releases of Movies and TV Shows on Netflix", fontsize=16)
plt.grid(True)
plt.legend()

# Display the plot
plt.tight_layout()
plt.show()
```



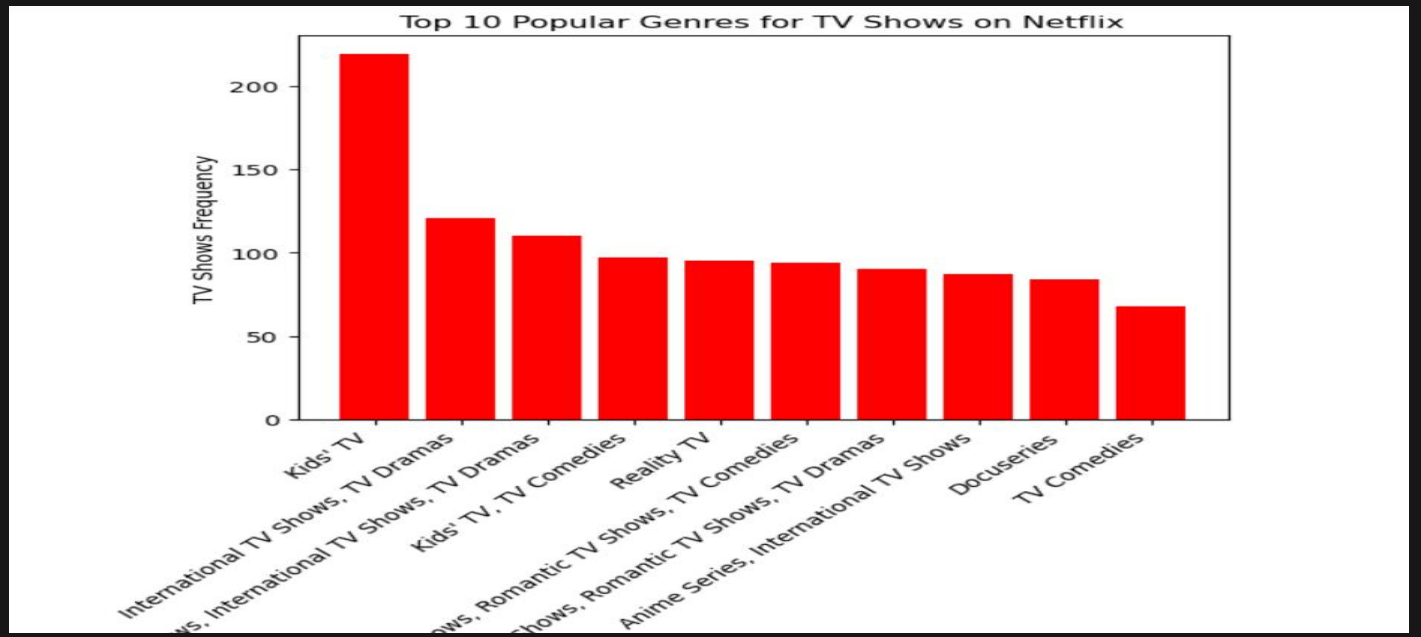
Top 10 movies and tv shows genres on Netflix

- This bar graph shows the top 10 popular genres of movies and tv shows on Netflix

```
# Plot the top 10 genres for Movies
plt.bar(popular_movie_genre.index, popular_movie_genre.values, color='red')
plt.xticks(rotation=45, ha='right')
plt.xlabel("Genres")
plt.ylabel("Movies Frequency")
plt.title("Top 10 Popular Genres for Movies on Netflix")
plt.show()
```



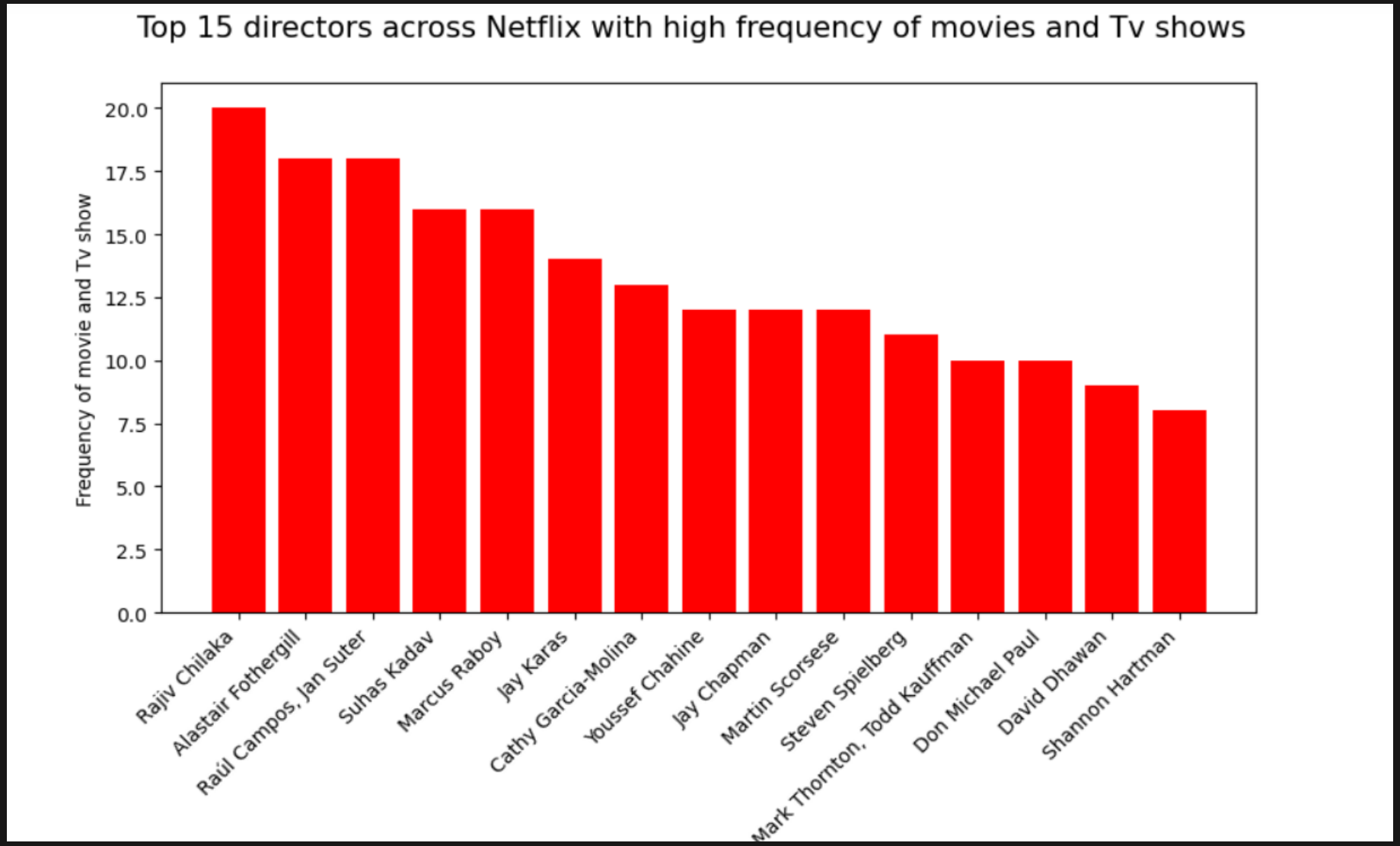

```
# Plot the top 10 genres for TV Shows
plt.bar(popular_series_genre.index, popular_series_genre.values, color='red')
plt.xticks(rotation=45, ha='right')
plt.xlabel("Genres")
plt.ylabel("TV Shows Frequency")
plt.title("Top 10 Popular Genres for TV Shows on Netflix")
plt.show()
```



Top 15 directors with high frequency of movies and TV shows

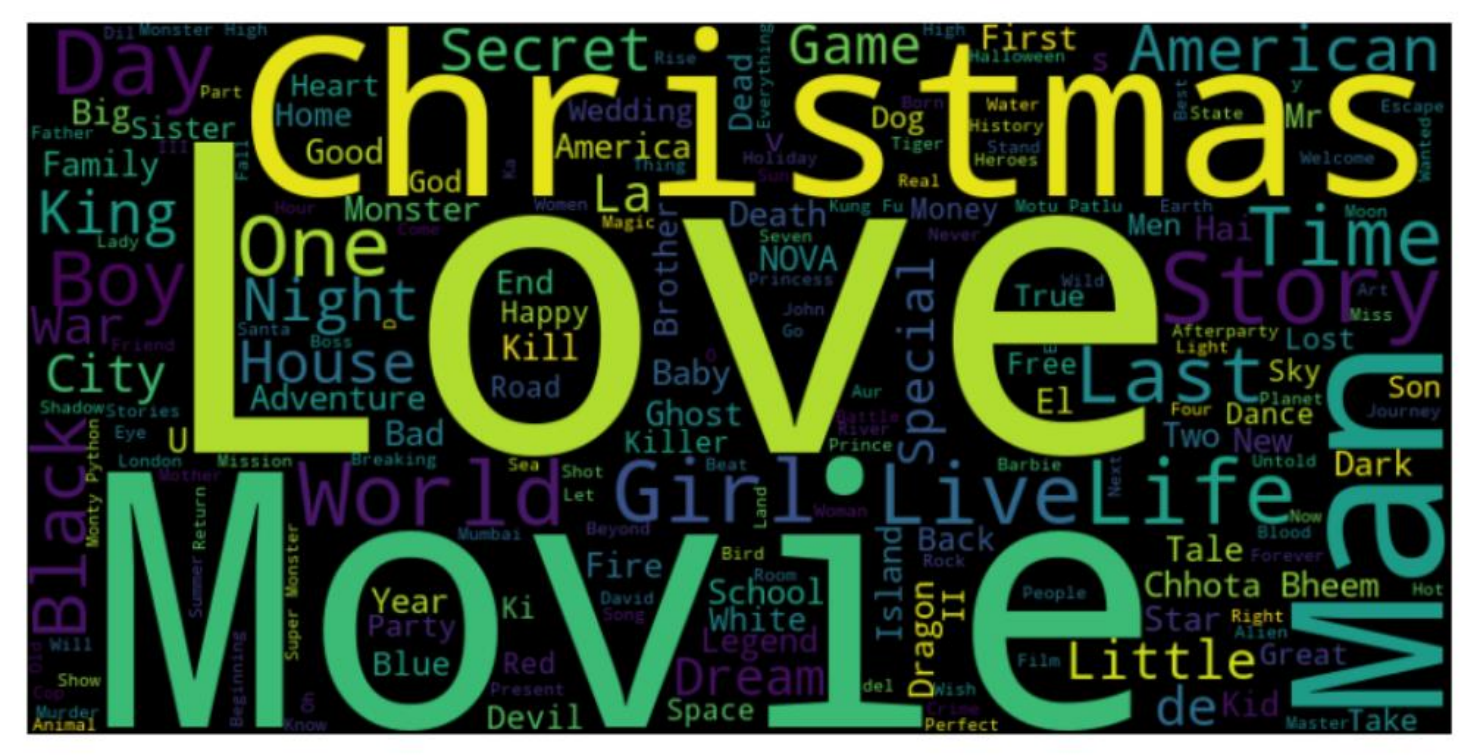
- This bar graph shows the top 15 directors with high frequency of movies and tv shows on Netflix.

	director	count
1	Rajiv Chilaka	20
2	Alastair Fothergill	18
3	Raúl Campos, Jan Suter	18
4	Suhas Kadav	16
5	Marcus Raboy	16
6	Jay Karas	14
7	Cathy Garcia-Molina	13
9	Youssef Chahine	12
10	Jay Chapman	12
8	Martin Scorsese	12
11	Steven Spielberg	11
12	Mark Thornton, Todd Kauffman	10
13	Don Michael Paul	10
14	David Dhawan	9
20	Shannon Hartman	8



World Cloud of Movie Titles

```
# Filter movie titles from the dataset
movie_titles = data.loc[data['type'] == 'Movie', 'title']

# Generate a word cloud
wordcloud = WordCloud(
    width=800,
    height=400,
    background_color='black'
).generate(' '.join(movie_titles))

# Display the word cloud
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



Key Insights

1. Movies constitute 70% (6126 titles), and TV shows make up 30% (2664 titles) of the Netflix catalog.

2. TV – MA has the highest content rating.

2. Top genres include International Dramas (362), Documentaries (359), and Stand-Up Comedy (334).

3. Significant growth in content production is observed post-2000.

4. The United States, India, and the United Kingdom are the top three countries producing content for Netflix.

5. Rajiv Chilaka, Alastair Fothergill, Raul Campos and Jan Suter are among the most featured directors on Netflix.

6. Movies are released more frequently than TV series each month. TV series release remain steady with slight increase in July and December, Feb has the lowest movie and series released, possibly due to seasonal trends.

7. Netflix significantly increased content release after 2015, with a sharp rise in both movie and TV shows.

8. Netflix focused more on movie than TV shows over the year.

Conclusion

This project showcases how data-driven insights can shape content strategies, improve audience engagement, and support platform expansion. This analysis reveals a significant growth in content production post-2000, driven by the platform's global expansion.

THANKS