Project Report: Netflix Data Analysis and Visualization

1. Introduction

Netflix has become a prominent platform for streaming movies, TV shows, and exclusive original content, that offers a wide range of catalogs that attracts a global audience. In this project, we will explore Netflix's content library by analyzing a dataset containing detailed information on titles available on the platform from 2008 to 2021. The dataset includes content from various countries and multiple languages, spanning genres and content types from dramas to documentaries. The oldest entry in the dataset dates back to 1925, while the most recent entries are from 2021, providing a broad historical perspective on the content offered on Netflix.

The primary aim of this project is to use data cleaning and exploratory data analysis (EDA) techniques to uncover meaningful patterns and trends in Netflix's catalog. This project also enhances data manipulation, cleaning, and visualization skills using Python in Jupyter Notebook. We hope to better understand Netflix's content strategies and user engagement patterns through data processing and visual exploration. Visualizations were created in Python to present a comprehensive picture of the findings.

2. Dataset Overview

The dataset contains 8790 rows and 10 columns representing attributes of Netflix titles. The structure of the dataset is as follows:

- **show_id:** A unique identifier for each title.
- **type:** Type of content, either a movie or a TV show.
- **title:** Title of the content.
- **director:** Name(s) of the director(s).
- **country:** The country where the content was produced.
- **date_added:** Date when the content was added to Netflix.
- **release_year:** Year the content was initially released.
- **rating:** Audience rating (e.g., PG-13, R).
- **duration:** Duration (in minutes for movies and seasons for TV shows).
- **listed_in:** Genres or categories under which the content is listed.

```
#7. To Check the Index, Columns, Data Type, and Memory at Once

data.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
```

This dataset was ideal for applying data cleaning and transformation techniques. In the project, Python was used for cleaning, transforming, and analyzing the dataset, while Jupyter Notebook served as the primary IDE for documenting and visualizing the analysis.

3. Data Cleaning Process

Cleaning data is essential for accurate and efficient analysis. Key data cleaning steps undertaken in this project include identifying and handling missing values, converting data types, and refining the dataset to ensure accuracy in analysis. Below is a breakdown of each step:

3.1 Checking for Duplicates

To maintain data integrity, we checked for duplicate entries in the data set. This ensured that each row in the dataset represented a unique Netflix title.

```
data[data.duplicated()] #To check for any duplicate values, existing in the dataset
```

show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in

```
#We see no duplicate data exist in this dataset, so we can proceed further with this.
```

Result: No duplicate records were found, allowing us to proceed with confidence in the dataset's uniqueness.

3.2 Handling Missing Values

Handling missing values was critical for ensuring that analyses were accurate and representative. However, the dataset that we are working on doesn't contain any missing values, as we can see in the following information.

```
#7. To Check the Index, Columns, Data Type, and Memory at Once

data.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
```

3.3 Data Type Conversion

To enable time-based analysis, we converted the date_added column into the DateTime data type. This conversion allowed us to perform more precise monthly and yearly analyses, critical for understanding Netflix's content release trends.

```
data['date_added'] = pd.to_datetime(data['date_added']) #To convert 'date_added' column into DateTime data type
```

```
data.dtypes #To Confirm the Changes
```

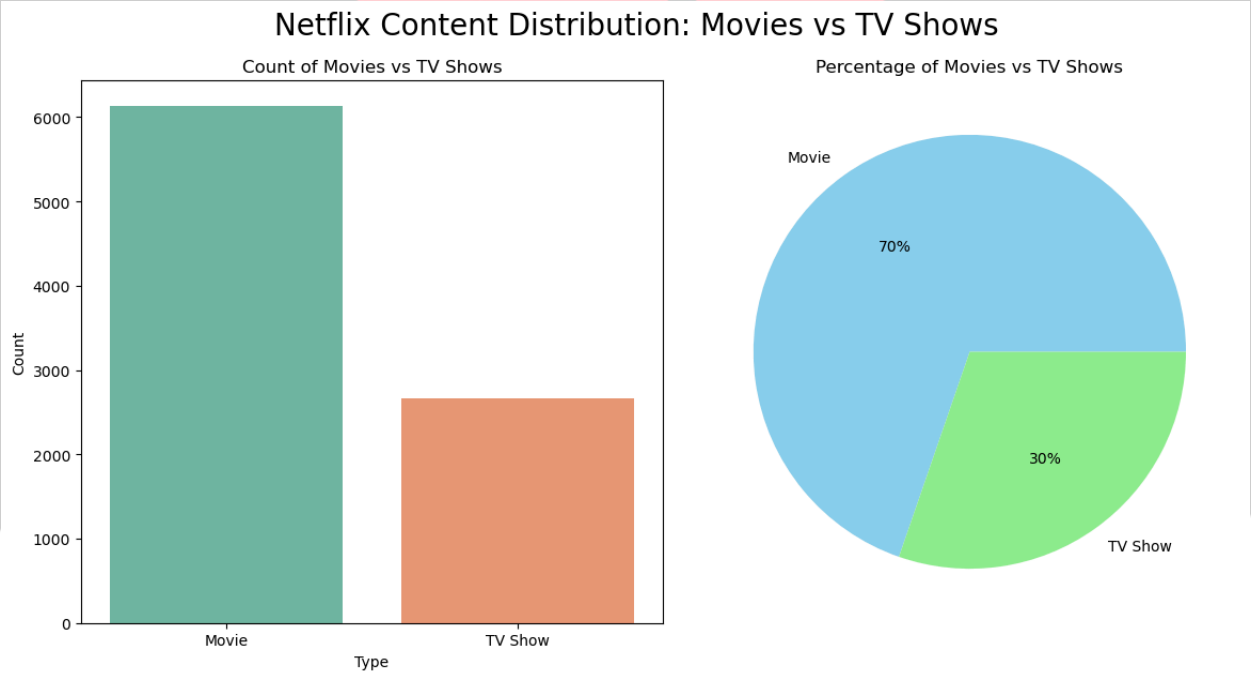
```
show_id                 object
type                    object
title                   object
director                object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in        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

4. Exploratory Data Analysis (EDA)

The EDA phase aimed to reveal patterns and insights into Netflix's content library, which focused on content type distribution, country-specific contributions, release patterns, and popular genres. Each analysis provides insights into viewer engagement trends, regional content focus, and genre preferences on Netflix.

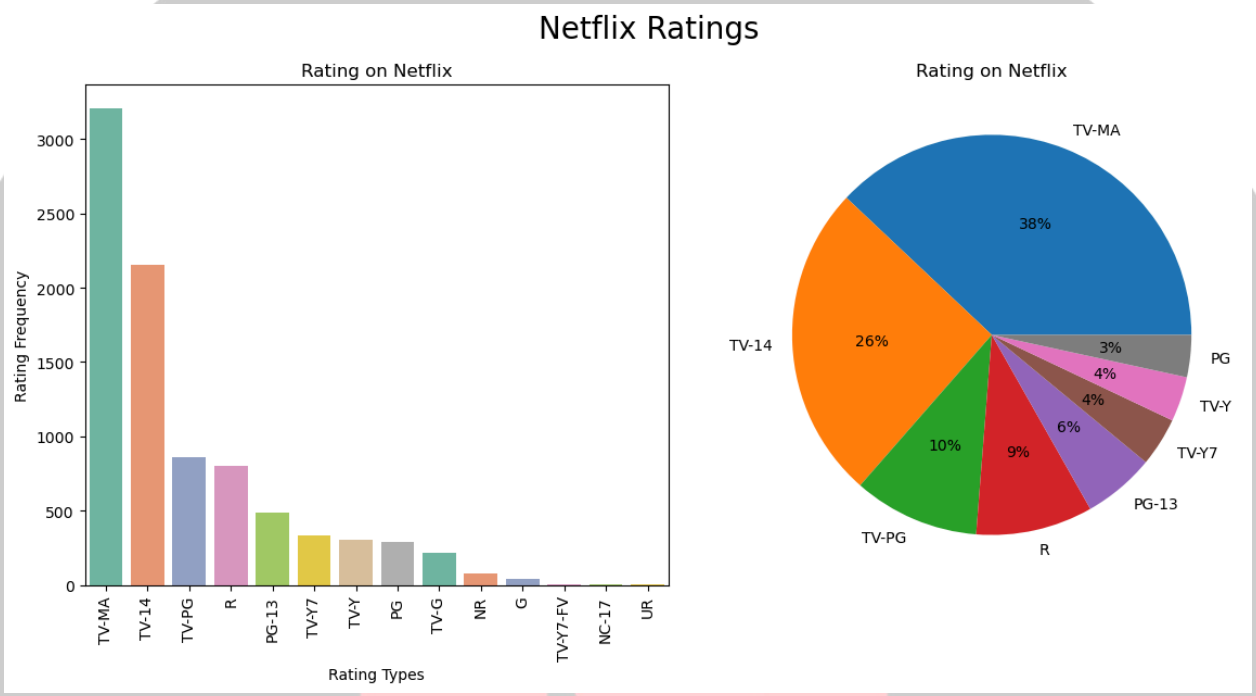
4.1 Content Distribution Analysis

This analysis compared the distribution of content types (movies vs. TV shows) to understand Netflix's content strategy. By plotting the frequency of movies and TV shows, we could see whether there was a focus on long-form content like TV shows or single-episode content like movies.



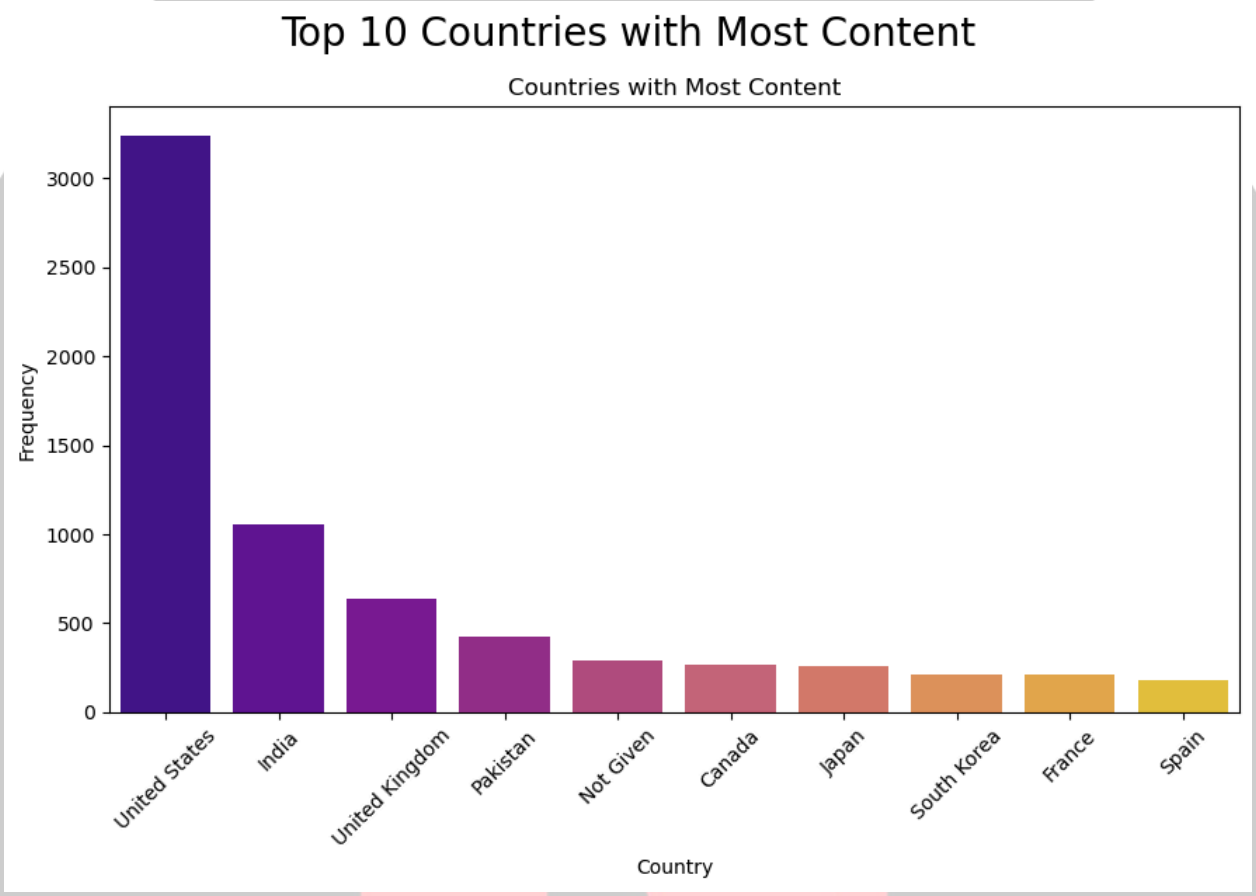
4.2 Rating Frequency of Movies and TV Shows

To analyze content ratings, we assessed how frequently each rating (such as PG, R, or TV-MA) appeared. This step provided insights into the audience demographics that Netflix targets with its content.



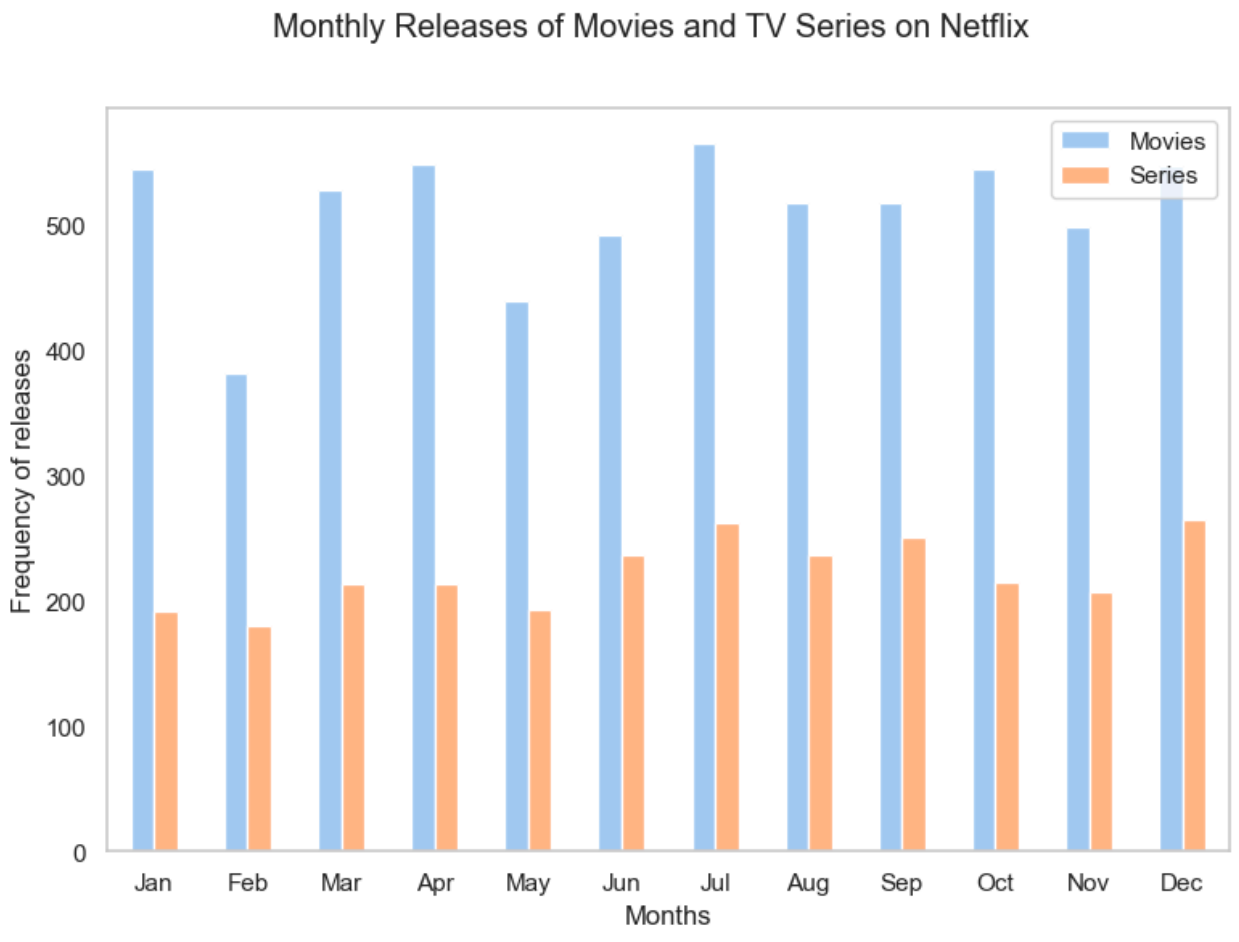
4.3 Top 10 Countries with Most Content

Using the country column, we identified the top 10 countries that contribute the most content to Netflix. This analysis offered insights into the geographic diversity of Netflix's library and highlighted the platform's efforts to feature content from various regions.



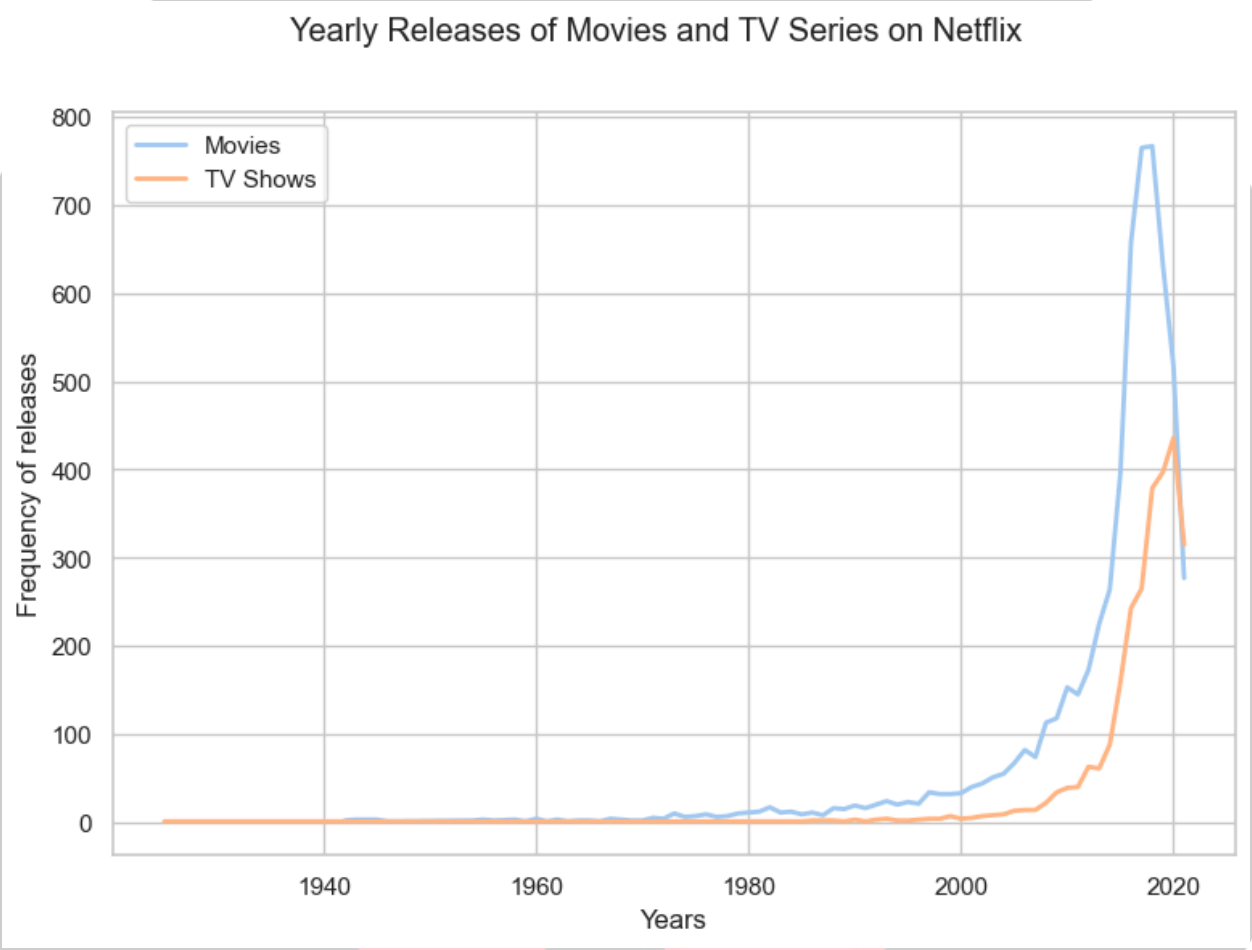
4.4 Monthly Release of Movies and TV Series

Analyzing monthly releases helped us identify trends in Netflix's content addition schedule, which pinpoints the peak months and slower months for content release. This analysis was possible by converting date_added to DateTime format and plotting the frequency of additions by month.



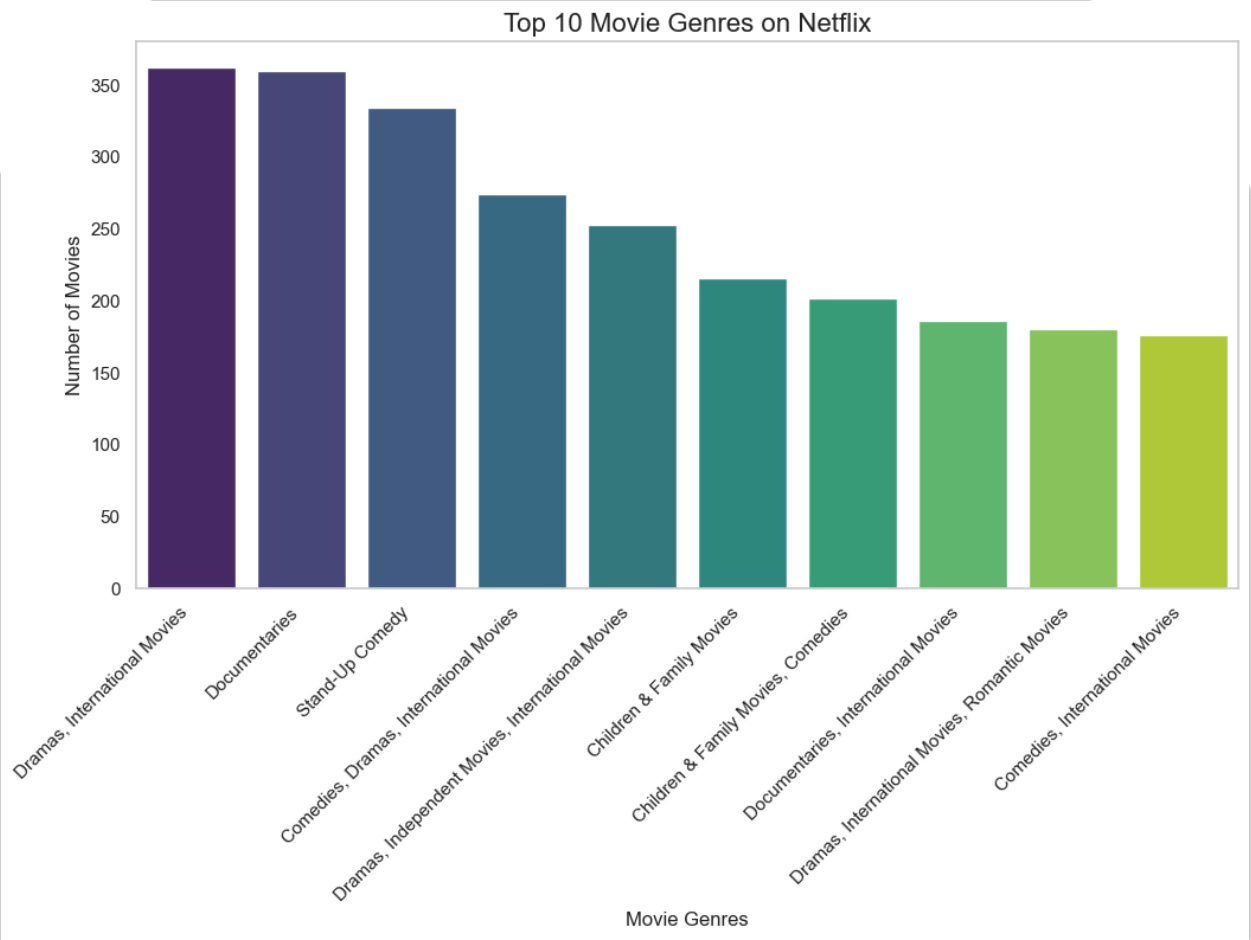
4.5 Yearly Release of Movies and TV Series

This analysis explored the year-over-year growth or decline in Netflix's content library. Yearly trends gave insight into Netflix's expansion, especially its content growth trajectory from 2008 to 2021, a period of rapid growth for the platform.



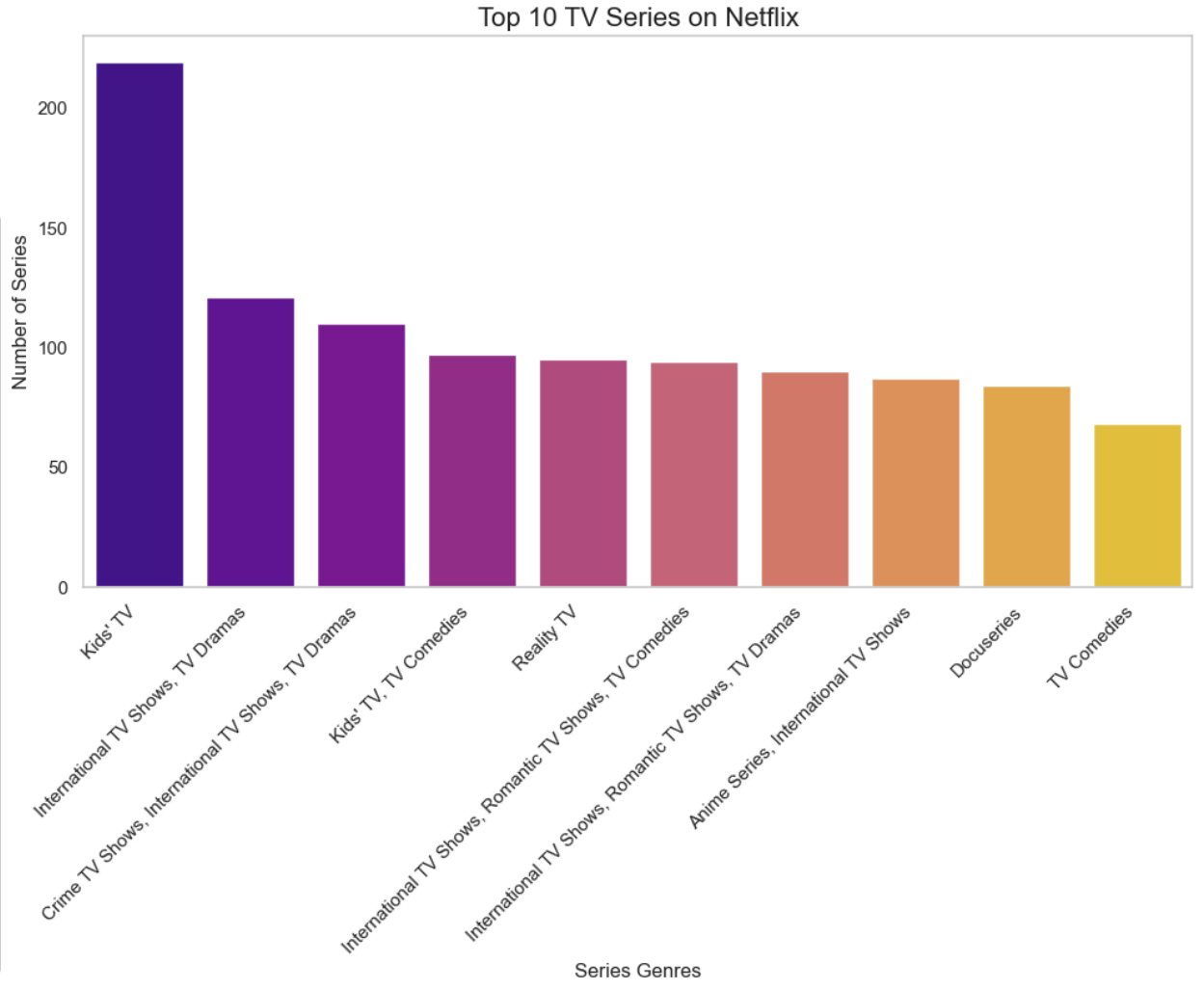
4.6 Top 10 Popular Movie Genres

We analyzed the listed_in column to determine the most popular genres among movies on Netflix. By counting occurrences of each genre, we identified genre preferences and focused content areas in the movie segment.



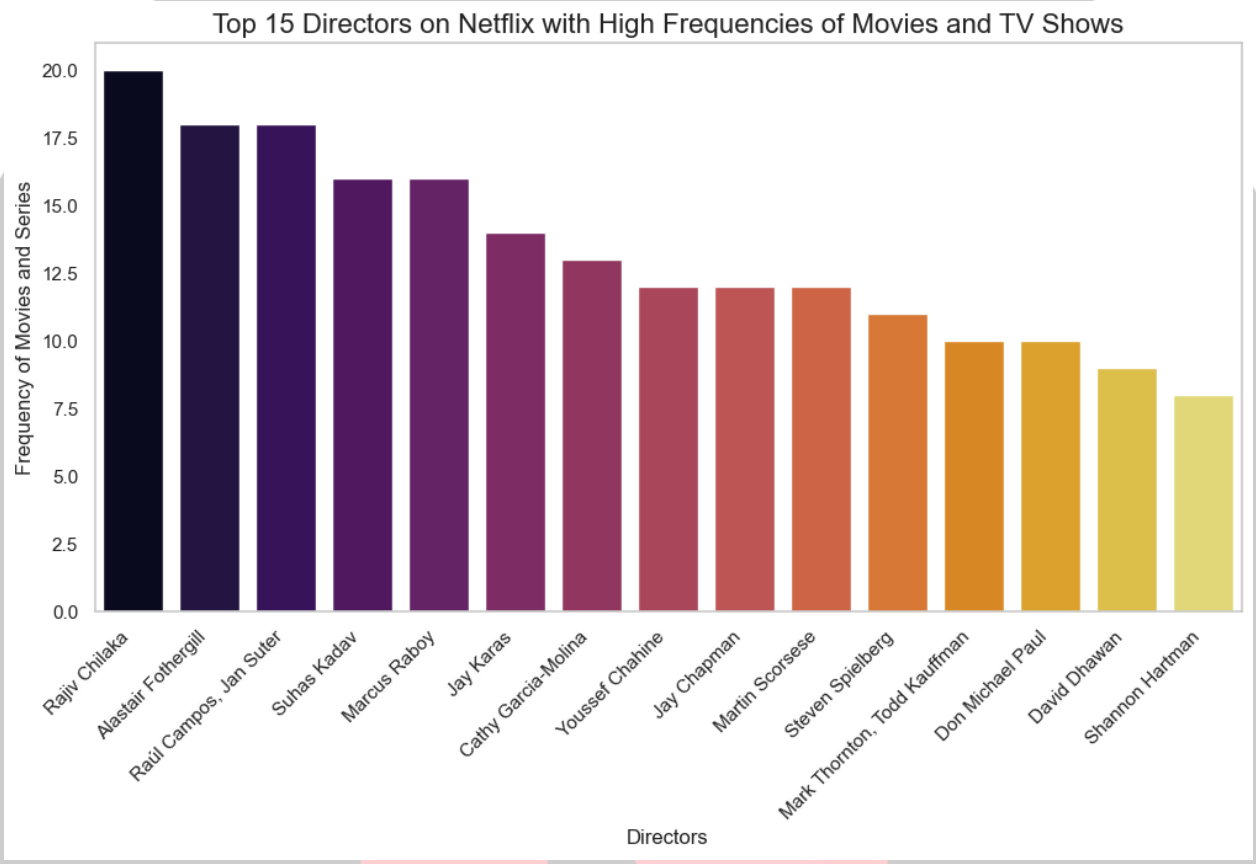
4.7 Top 10 TV Series Genres

Similarly, the top genres for TV series were analyzed, that provides a detailed insights into which genres draw the highest engagement for Netflix's series content.



4.8 Top 15 Directors with High Content Frequency

Using the director column, we identified the top 15 directors who have the most content on Netflix. It highlights the prolific contributors and the focus areas of Netflix's curated content library.



5. Visualization and Insights

Data visualizations were created using Python libraries such as Matplotlib and Seaborn to bring the insights to life and allow for easy interpretation of complex data. Each analysis from the EDA was supported by visual representations, that enables a clear presentation of trends. Key insights from the visualizations include:

- **Content Type Distribution:** Clear patterns showing Netflix's focus on movies or TV shows over the years.
- **Rating Frequency:** The popularity of content ratings indicates the audience demographics targeted by Netflix.
- **Country-Based Analysis:** Countries contributing the most content reflect Netflix's emphasis on regional diversity.
- **Monthly and Yearly Trends:** Monthly and yearly trends reveal strategic content additions, with significant growth periods highlighted.
- **Genre Preferences:** Analysis of genres across movies and TV series reveals popular viewer categories and genre diversity in Netflix's catalog.

6. Conclusion

This project demonstrated successful data cleaning, exploratory analysis, and visualization, providing a deep understanding of Netflix's content library. Using Python in Jupyter Notebook provided an effective and organized way to document each step, which ensures clarity and reproducibility.

By combining Python's data processing capabilities with visualization, we gained actionable insights into the platform's content trends and target demographics. The skills developed during this project are invaluable for handling real-world data and generating impactful visual representations, which showcases the importance of data cleaning and effective visualization in the analysis pipeline.