Exploratory Data Analysis

NETFLIX

Introduction

Netflix, Inc. is one of the most popular OTT platform around the globe. It’s an American subscription streaming service and production company. Launched on August 29, 1997, it offers a library of films and television series through distribution deals as well as own production Netflix Originals.

Dataset

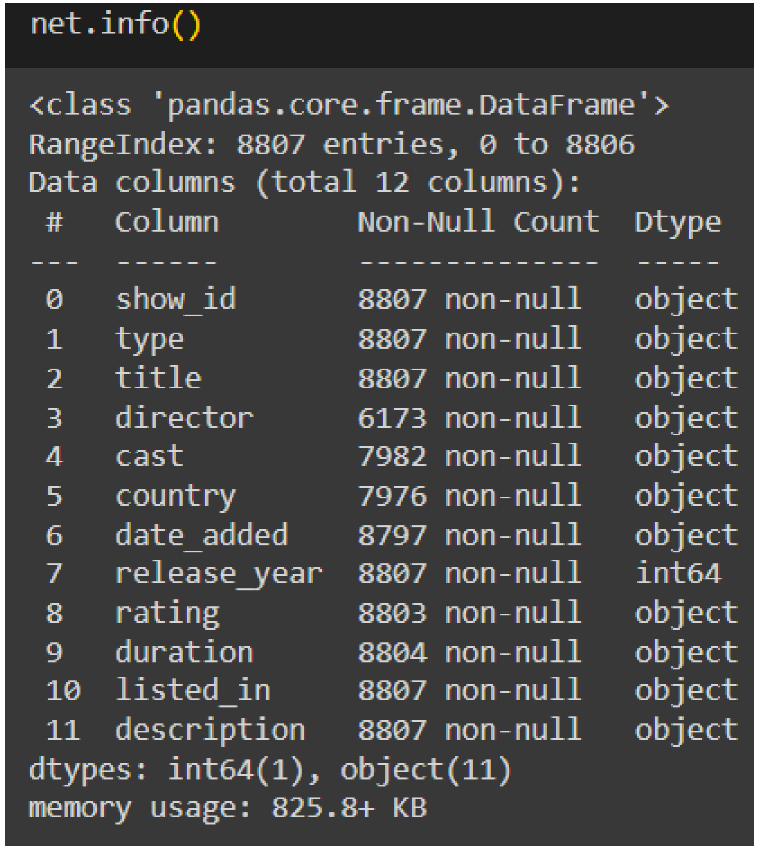
I used Netflix Movies and TV shows dataset. This dataset is widely used by beginner to learn EDA. It contains 8807 unique TV Shows and Movies.

Purpose of the project

The goal of the Netflix EDA project is to conduct a comprehensive exploration and analysis of Netflix's content dataset. This includes understanding the data structure, ensuring data integrity by handling missing values and duplicates, deriving descriptive statistics, and visualizing content distribution across genres and release years. Additionally, the project aims to identify temporal trends, analyze content attributes like ratings and duration, and assess audience engagement metrics. By synthesizing these insights, the project aims to draw meaningful conclusions and provide actionable recommendations to enhance Netflix's content offerings and user experience.

Description of Dataset

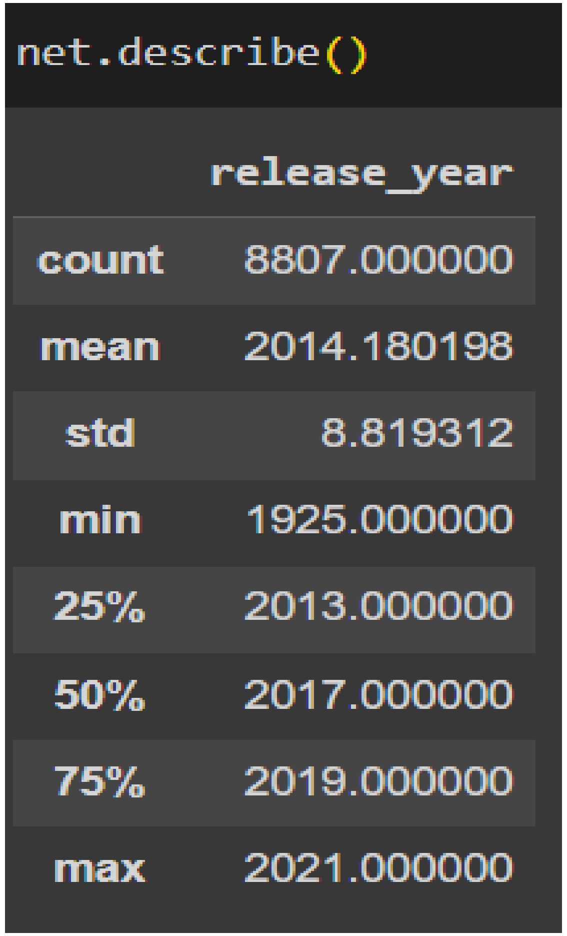
- Introduction and Objectives: The Netflix EDA project is designed to delve deep into the Netflix content dataset, aiming to gain comprehensive insights into the platform's content landscape. The primary objective is to understand, analyze, and derive meaningful insights from the dataset to inform strategic decisions, optimize content strategies, and enhance user experience on the platform.
- Data Exploration and Cleaning: The project commences with a meticulous data exploration phase to understand the dataset's structure, variables, and general patterns. This involves identifying and handling missing values, duplicates, and inconsistencies to ensure data integrity and reliability.
- Descriptive Statistics: Through descriptive statistics computation, the project aims to provide a quantitative perspective on the dataset. Key metrics such as mean, median, mode, range, and standard deviation will be calculated for relevant variables to gain insights into the content's distribution, characteristics, and variability.
- Data Visualization: The project includes creating compelling visualizations to represent the distribution of content across different genres, release years, and potentially geographical locations. These visualizations facilitate intuitive understanding, pattern recognition, and insight generation from the dataset.



```
net.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
```

Description of Dataset

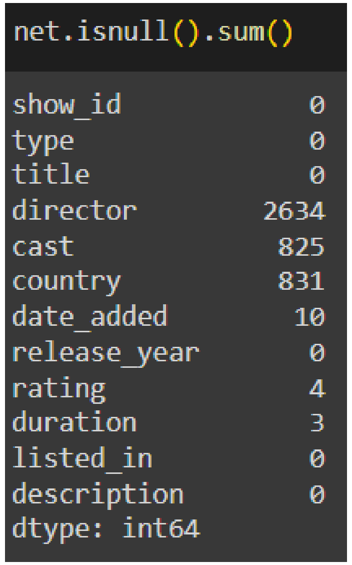
- Time Series Analysis: With a temporal component in the dataset, the project conducts time series analysis to identify evolving trends, patterns, and fluctuations over time, shedding light on the dynamic nature of Netflix content offerings.
- Content Analysis: The project delves into content- specific analyses, exploring attributes such as ratings, duration, and content variety. This analysis offers insights into the diversity, popularity, and characteristics of the content available on the platform.
- Audience Engagement Analysis: The project assesses audience engagement through user reviews, sentiment analysis, and engagement metrics like views or watch time. This analysis provides valuable data on audience behavior, interaction, and sentiment towards Netflix content.
- Conclusions and Recommendations: In conclusion, the project synthesizes the insights gained from the analysis to draw meaningful conclusions and provide actionable recommendations. These recommendations aim to enhance Netflix's content offerings, user experience, and engagement on the platform, based on the identified insights and trends.



```
net.describe()

       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
```

Data Cleaning and Preparation

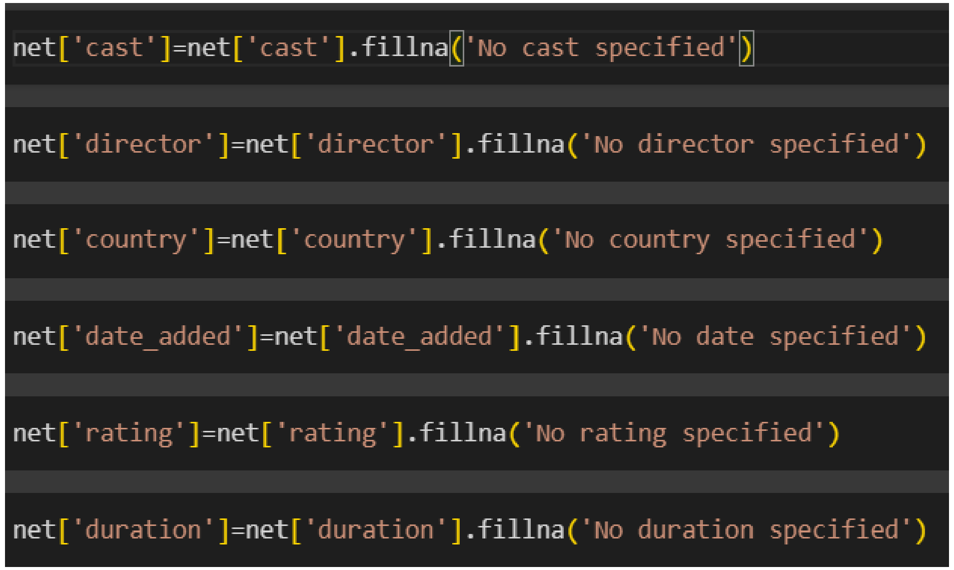
- **Director** : For the 'Director' attribute with 2634 null values, one approach is to fill these missing values with a placeholder such as 'Unknown' or ‘No director specified'. This allows retaining the data records while indicating the absence of director information. Alternatively, for more accurate data, you can research and populate missing director information by referencing external sources or databases related to the movies or TV shows.
- **Cast** : With 825 null values in the 'Cast' attribute, a similar approach can be applied. Filling the missing values with 'Unknown' or 'No cast specified' can help maintain data completeness. Alternatively, you can leverage external databases or IMDb (Internet Movie Database) to populate missing cast information for each movie or TV show.
- **Country** : For the 'Country' attribute with 831 null values, filling the missing values with the most common country of production or 'No country specified' can be a feasible approach. Another strategy is to cross- reference with the title or other metadata to infer the country of production based on the content's origin or production company.



```
net.isnull().sum()
```

```
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
```

There are total 4307 null values present in the data set.



```
net['cast']=net['cast'].fillna('No cast specified')
```

```
net['director']=net['director'].fillna('No director specified')
```

```
net['country']=net['country'].fillna('No country specified')
```

```
net['date_added']=net['date_added'].fillna('No date specified')
```

```
net['rating']=net['rating'].fillna('No rating specified')
```

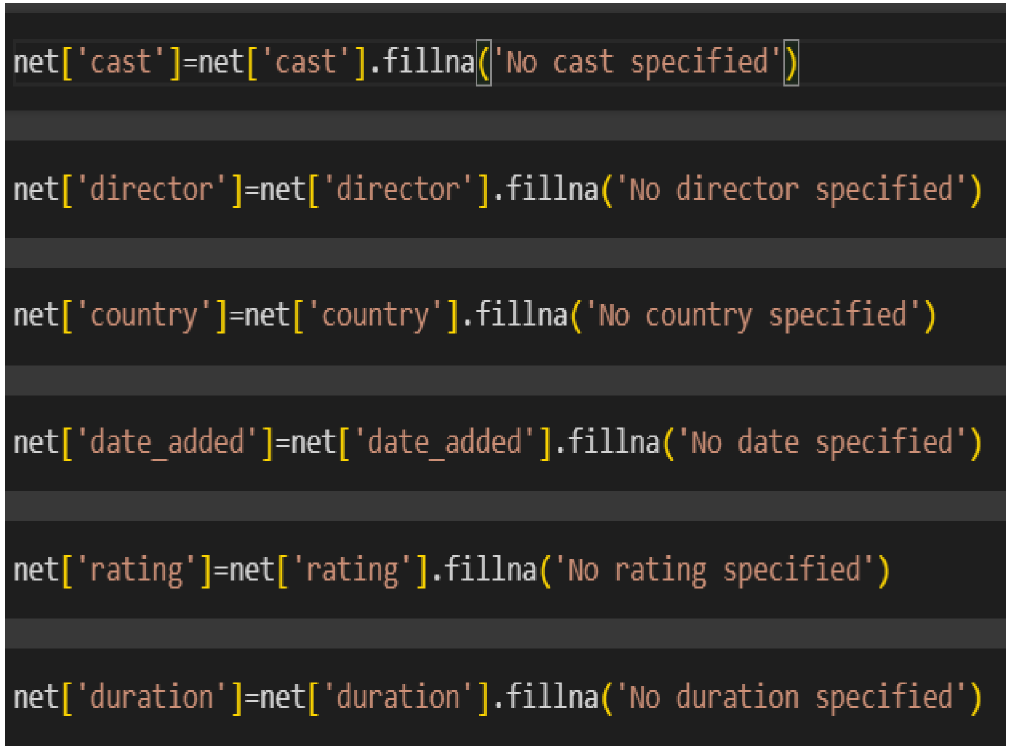
```
net['duration']=net['duration'].fillna('No duration specified')
```

Data Cleaning and Preparation

- **Date Added** : With only 10 null values in the 'Date Added' attribute, filling these missing values can be straightforward. You can impute the missing dates by filling 'No date specified'.

- **Rating** : For the 'Rating' attribute with 4 null values . I replaced the values with 'No rating specified' because I don't want to assume the ratings because it will be unfair with the dataset.

- **Duration** : For the 'Duration' attribute with 3 null values do I replaced the values with 'No duration specified' because I don't want to assume the duration because it will be unfair with the dataset.

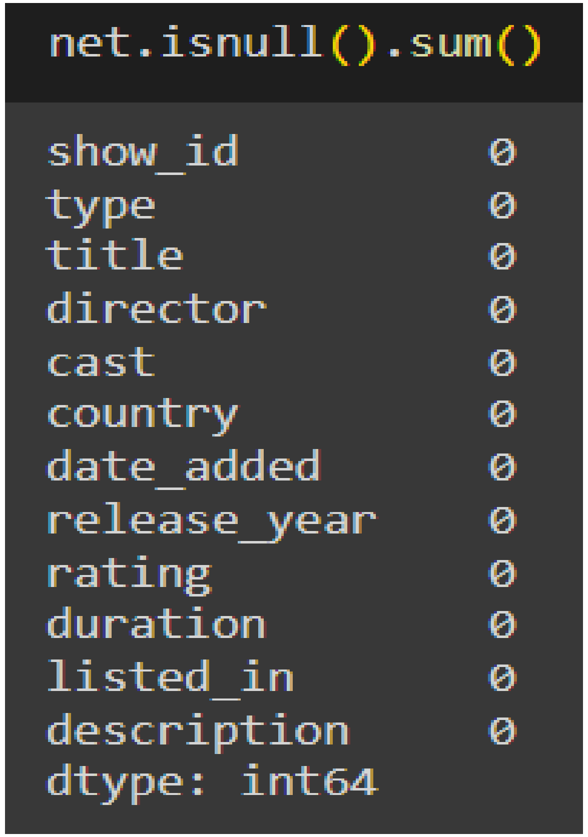


```
net['cast']=net['cast'].fillna('No cast specified')

net['director']=net['director'].fillna('No director specified')

net['country']=net['country'].fillna('No country specified')

net['date_added']=net['date_added'].fillna('No date specified')

net['rating']=net['rating'].fillna('No rating specified')

net['duration']=net['duration'].fillna('No duration specified')
```

Data Cleaning and Preparation

Conclusion :

In conclusion, dealing with null values requires a systematic approach based on the nature of the data and the specific attributes. By employing the described strategies, you can effectively handle and fill the missing values in the dataset, ensuring data completeness, integrity, and reliability for subsequent analysis and insights derivation.

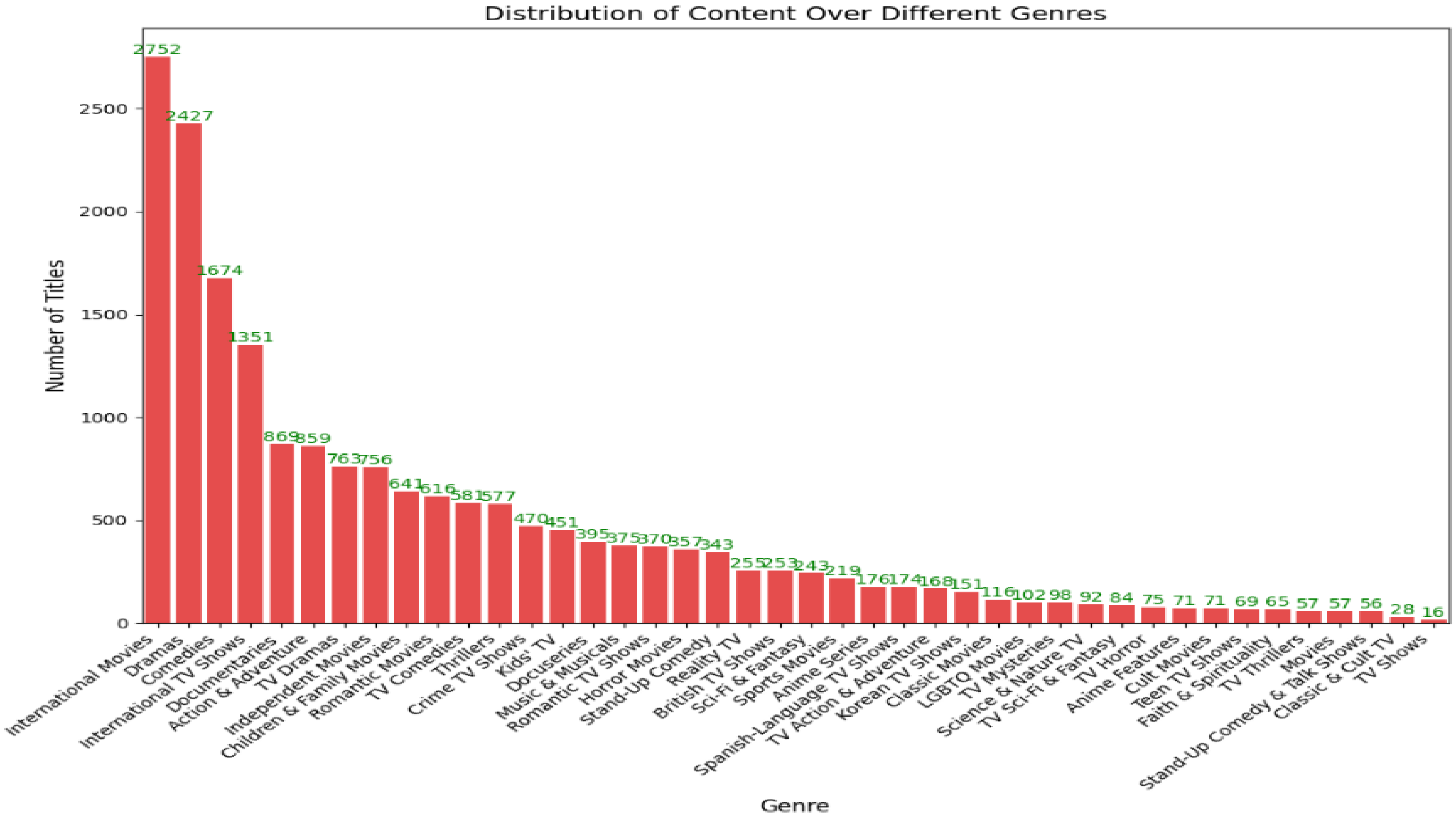


```
net.isnull().sum()
```

```
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
```

Data Visualization and Insights

Distribution of content over different genres :



Distribution of Content over Different Genres

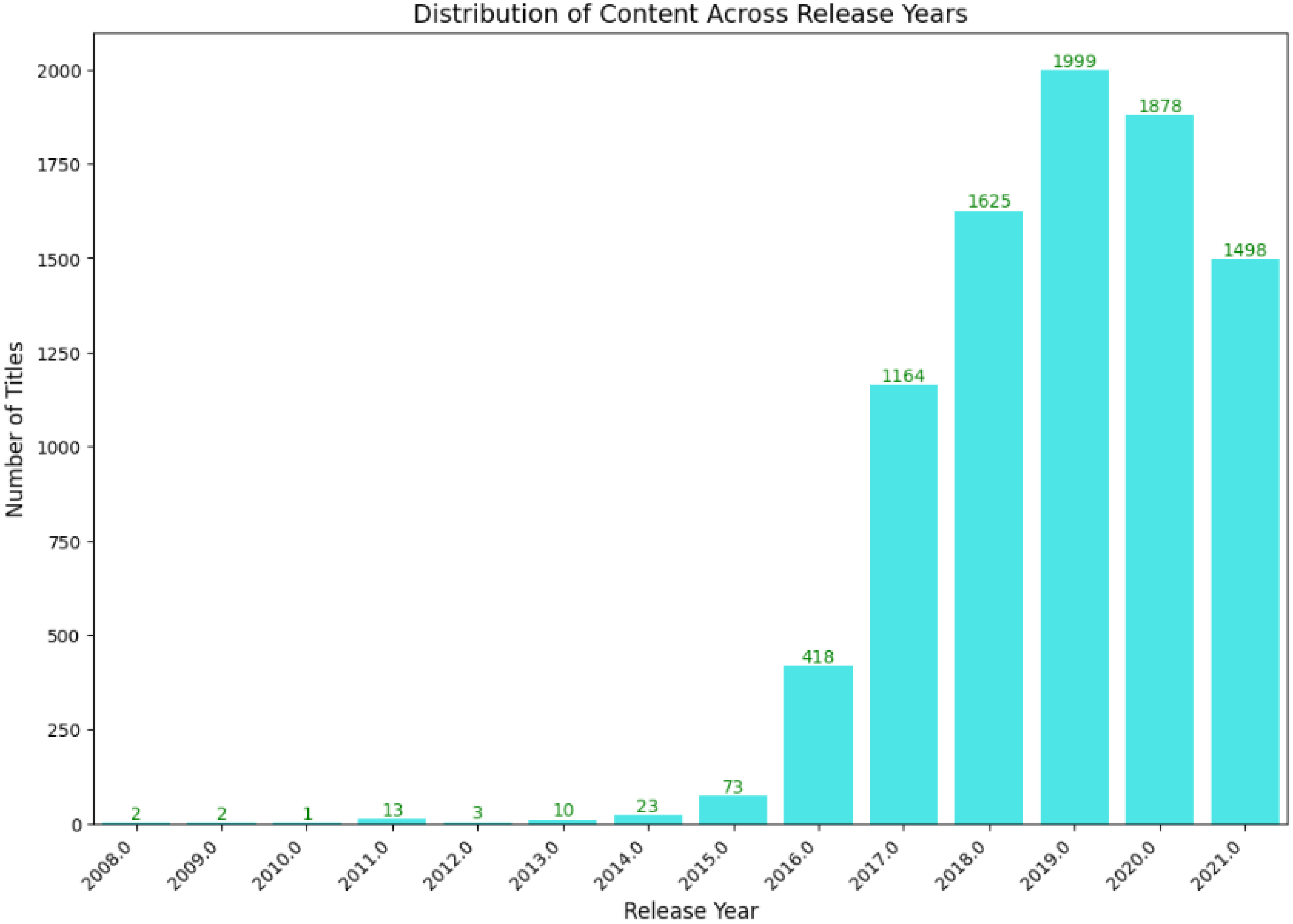
Key Findings :

- The most common genres on Netflix are "International Movies", "Dramas", and "Comedies".
- Genres like "Classic Movies", "TV Shows", and "Independent Movies" are less represented.
- There is a significant difference in the number of titles between the most and least popular genres.

Conclusion :

Netflix's content library is heavily skewed towards international movies, dramas, and comedies. This suggests that these genres are most popular among their audience. However, there is also a demand for niche genres, as indicated by the presence of categories like "Classic Movies" and "Independent Movies".

Distribution of Content Across Release Years



Distribution of Content Across Release Years

Key Findings :

- Recent Years Dominate: The majority of content available on Netflix was added in recent years, with a significant peak observed around 2019-2020.

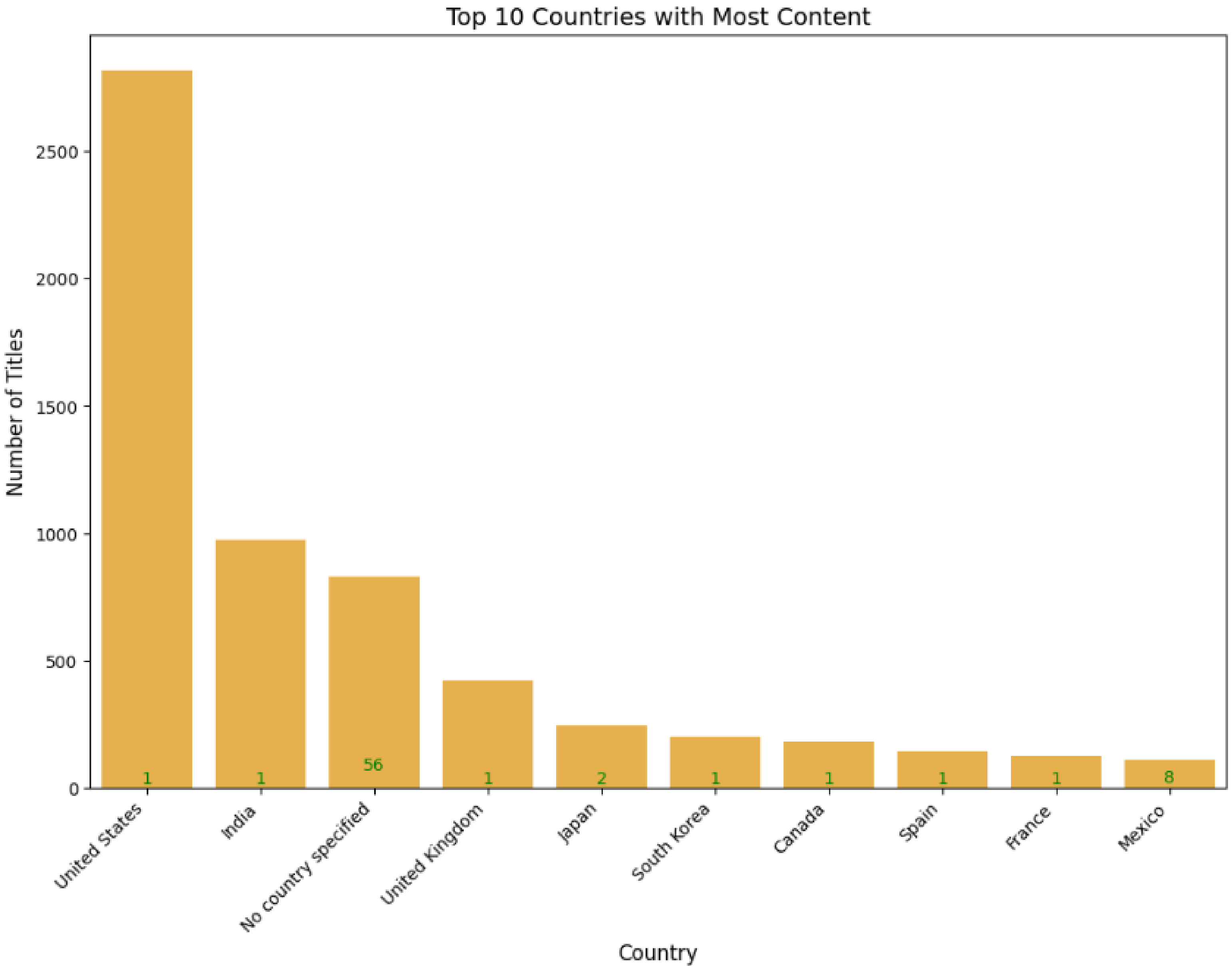
- Steady Growth: There's a general trend of increasing content additions over the years, reflecting Netflix's expansion and investment in original programming.

- Older Content Presence: While newer content is predominant, there's still a notable amount of content from earlier years, indicating a diverse catalog.

Conclusion :

Netflix's content library is heavily focused on recent releases, suggesting a strategy to attract viewers with fresh and current programming. However, the presence of older titles caters to a wider audience and provides a diverse selection.

Geographical Distribution of Content



Geographical Distribution of Content

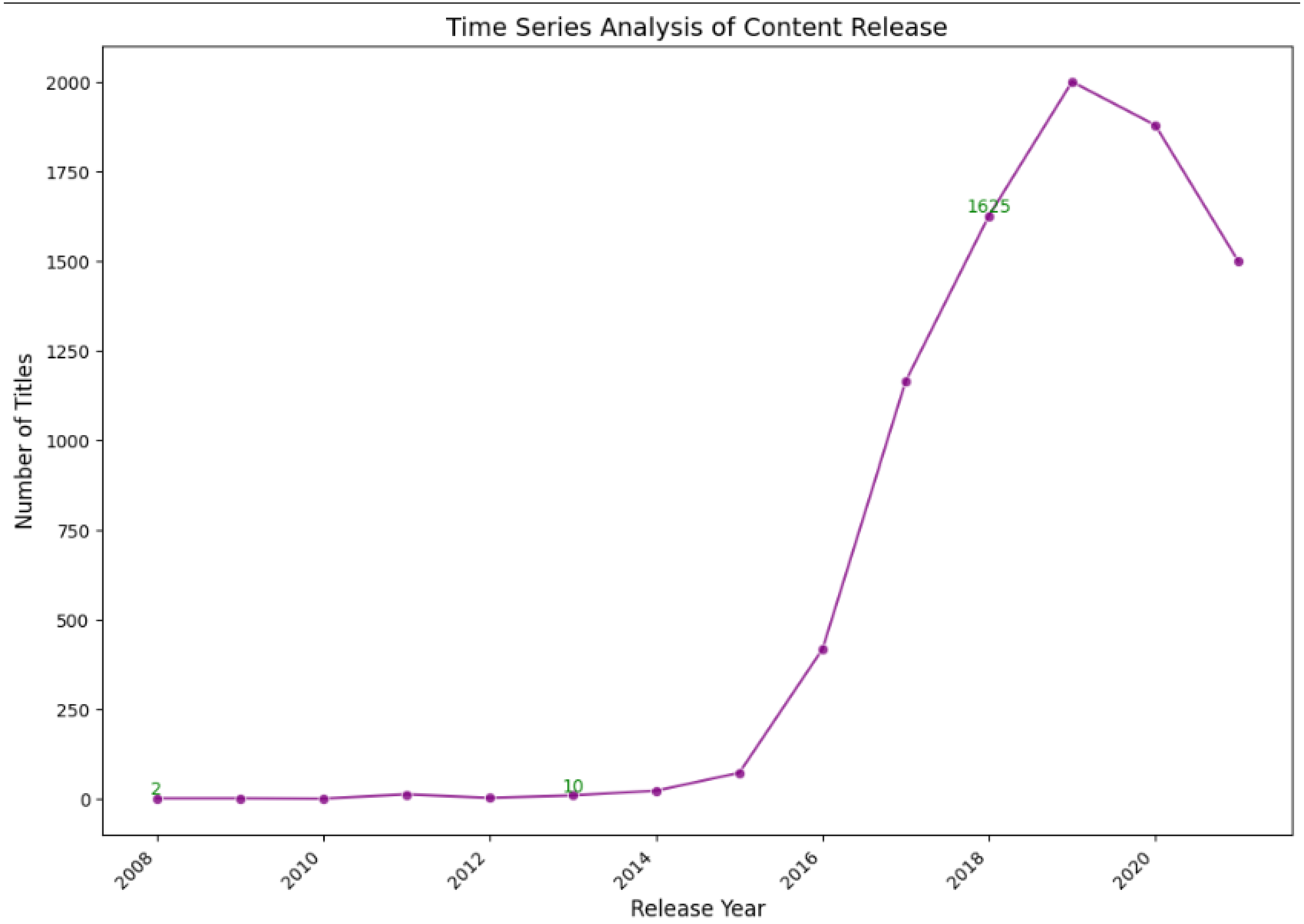
Key Findings :

- United States dominates content production: The US is the primary contributor of content on Netflix.
- Global representation is increasing: While the US leads, there's a notable presence of content from other countries, indicating Netflix's effort to cater to a global audience.
- Emerging markets are contributing: Countries like India, UK, Canada, etc. are becoming significant content providers.

Conclusion :

Netflix's content library reflects a globalized approach, with a strong US presence complemented by growing contributions from various countries. This strategy aligns with Netflix's goal of being a global entertainment platform.

Time Series Analysis to Identify Trends and Patterns over Time



Time Series Analysis to Identify Trends and Patterns over Time

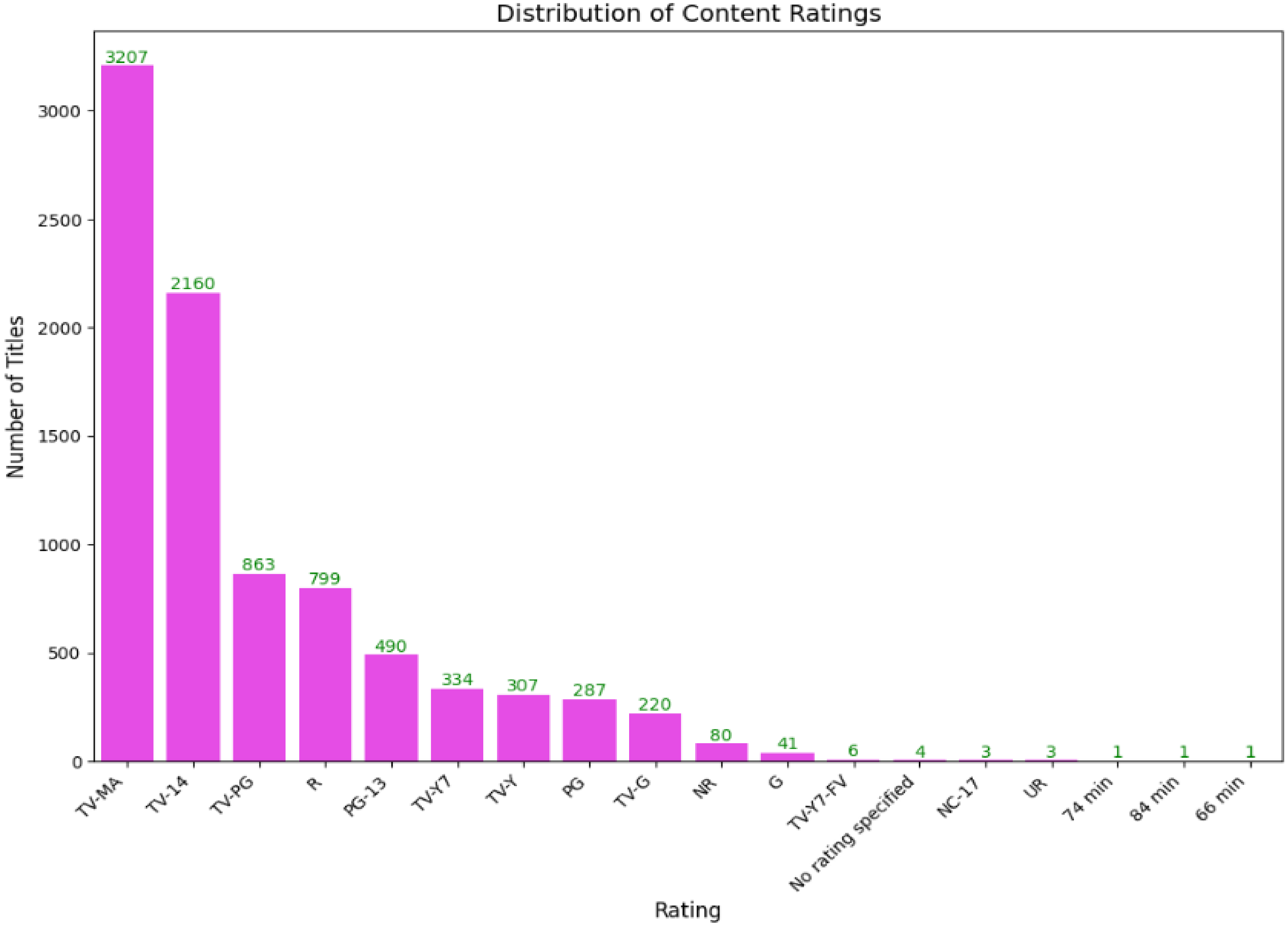
Key Findings :

- There's a general upward trend in the number of titles released on Netflix over the years.
- There might be some seasonality patterns, but the data provided doesn't allow for detailed seasonality analysis.
- There might be specific years with significant increases or drops in content release, which could be further investigated.

Conclusion :

- Netflix's content library has been expanding over time, indicating a commitment to providing more options to viewers.
- Understanding potential seasonality could help optimize content release strategies.
- Investigating the reasons behind significant yearly fluctuations could provide insights into market trends or internal strategic decisions.

Distribution of Content Ratings



Distribution of Content Ratings

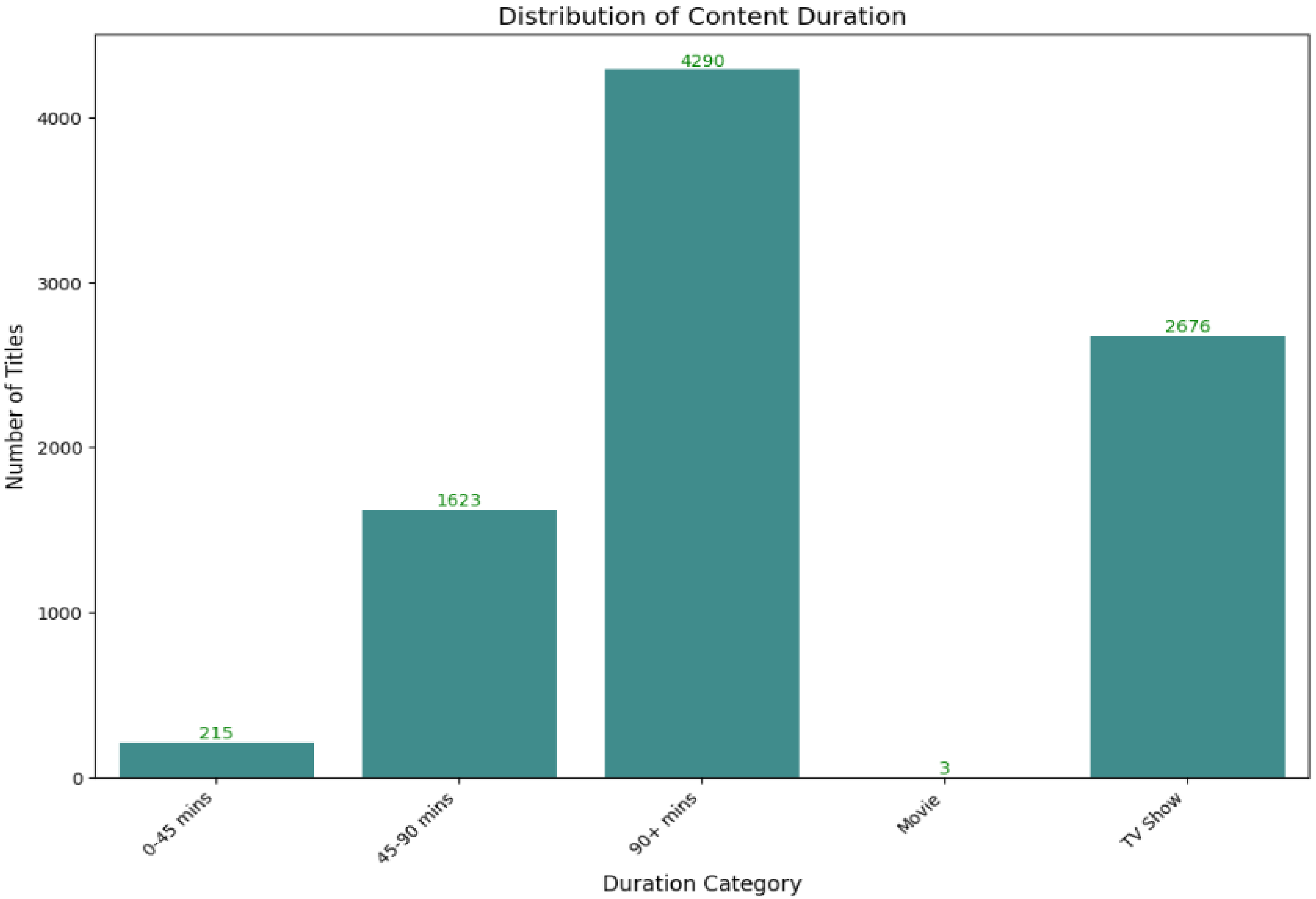
Key Findings :

- The distribution of content ratings on Netflix is not uniform.
- Certain ratings like "TV-MA" (mature audience) and "TV-14" (parents strongly cautioned) are more prevalent, indicating a larger portion of content geared towards older audiences.
- This suggests that Netflix caters significantly to adults and young adults. Content for younger viewers (e.g., "TV-Y", "TV-G") is less common.

Conclusion :

Understanding this distribution is crucial for content creators and Netflix to tailor their offerings to their target audience.

Distribution of Content Duration over Number of Titles



Distribution of Content Duration over Number of Titles

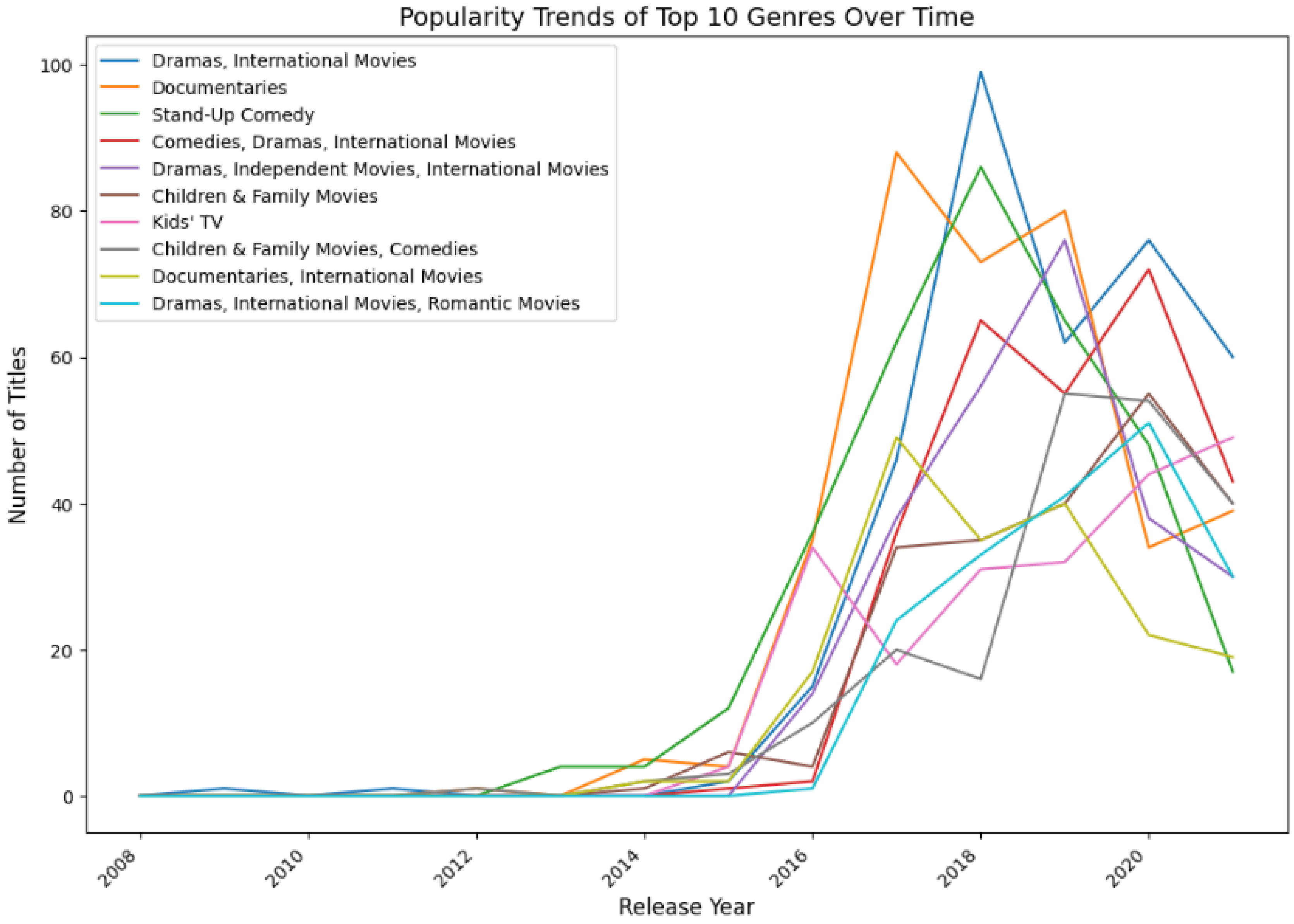
Key Findings :

- The majority of content on Netflix falls under the 'Movie' category.
- Among movies, the most common duration is '90+ mins', indicating a preference for longer films.
- TV Shows constitute a significant portion of the content, reflecting the growing popularity of serialized content.
- There's a smaller but notable presence of shorter movies under 90 minutes

Conclusion :

- Netflix's content library demonstrates a focus on providing a diverse range of durations to cater to varying viewer preferences.
- The platform caters to both movie enthusiasts who enjoy longer narratives and viewers who prefer shorter, more digestible content.
- The substantial presence of TV Shows highlights Netflix's commitment to offering serialized storytelling.

Popularity trends of Top 10 Genres over Time



Popularity trends of Top 10 Genres over Time

Key Findings :

- International Movies, Dramas, and Comedies: These genres have consistently remained popular over the years.

- International Movies have seen a significant surge in recent years, reflecting Netflix's global expansion. Dramas and Comedies remain staples in content production, catering to a wide audience.

- Rise of Docuseries and Stand-Up Comedy: These genres have experienced a notable increase in popularity, especially in recent years. This suggests a growing interest in non-fiction and comedic content.

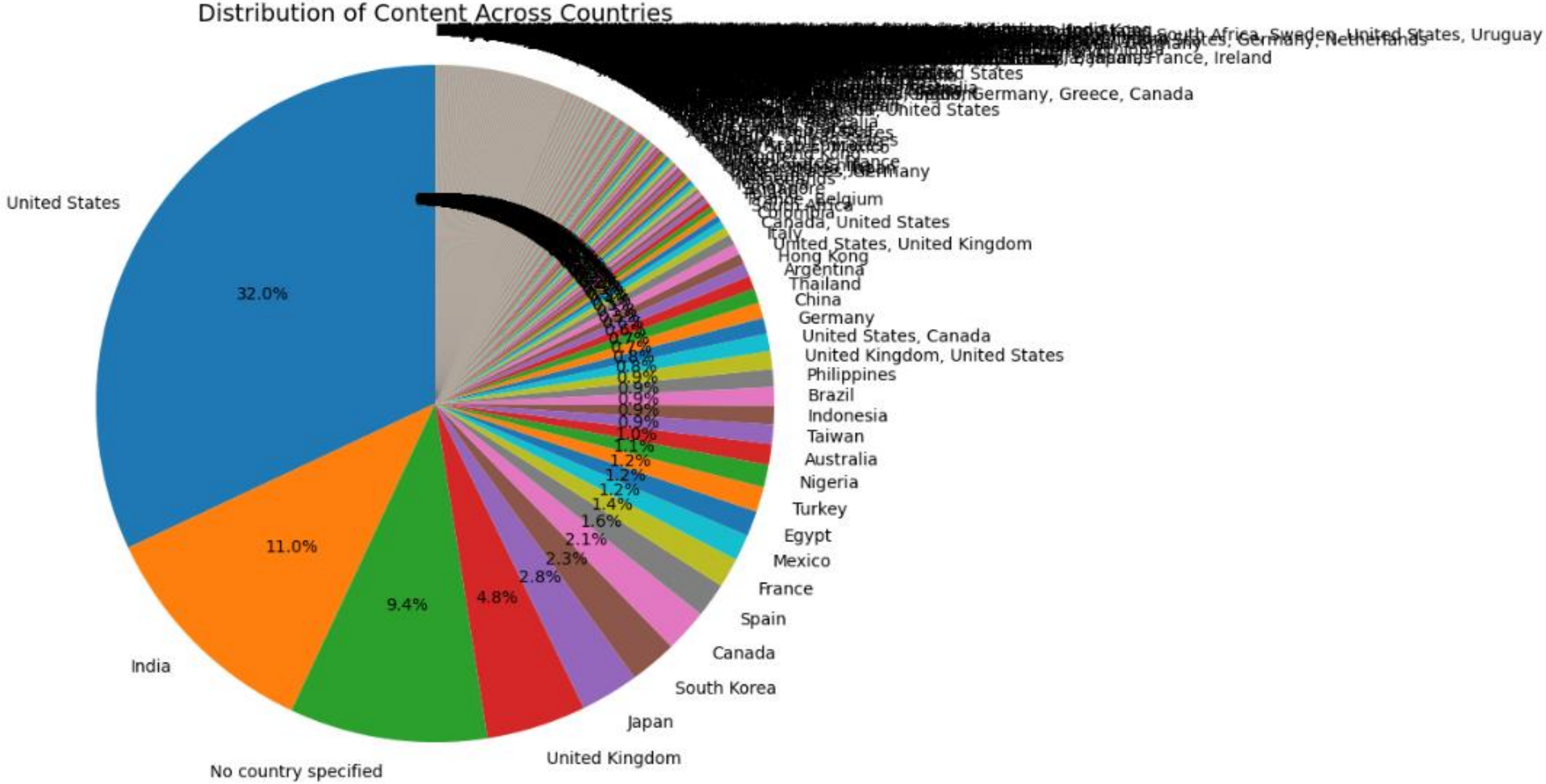
- Decline of Classic Movies and TV Shows: These genres have seen a decline in popularity, possibly due to changing viewer preferences and the abundance of newer content.

- Fluctuations in Other Genres: Genres like Action & Adventure, Independent Movies, and Children & Family Shows have experienced fluctuations in popularity, indicating changing trends and potential competition from other platforms.

Conclusion :

The popularity of different genres on Netflix is dynamic and influenced by various factors, including global trends, viewer preferences, and the platform's content strategy. Understanding these trends can help Netflix make informed decisions about content acquisition and production.

Distribution of Content Different Countries and Region



Distribution of Content Different Countries and Region

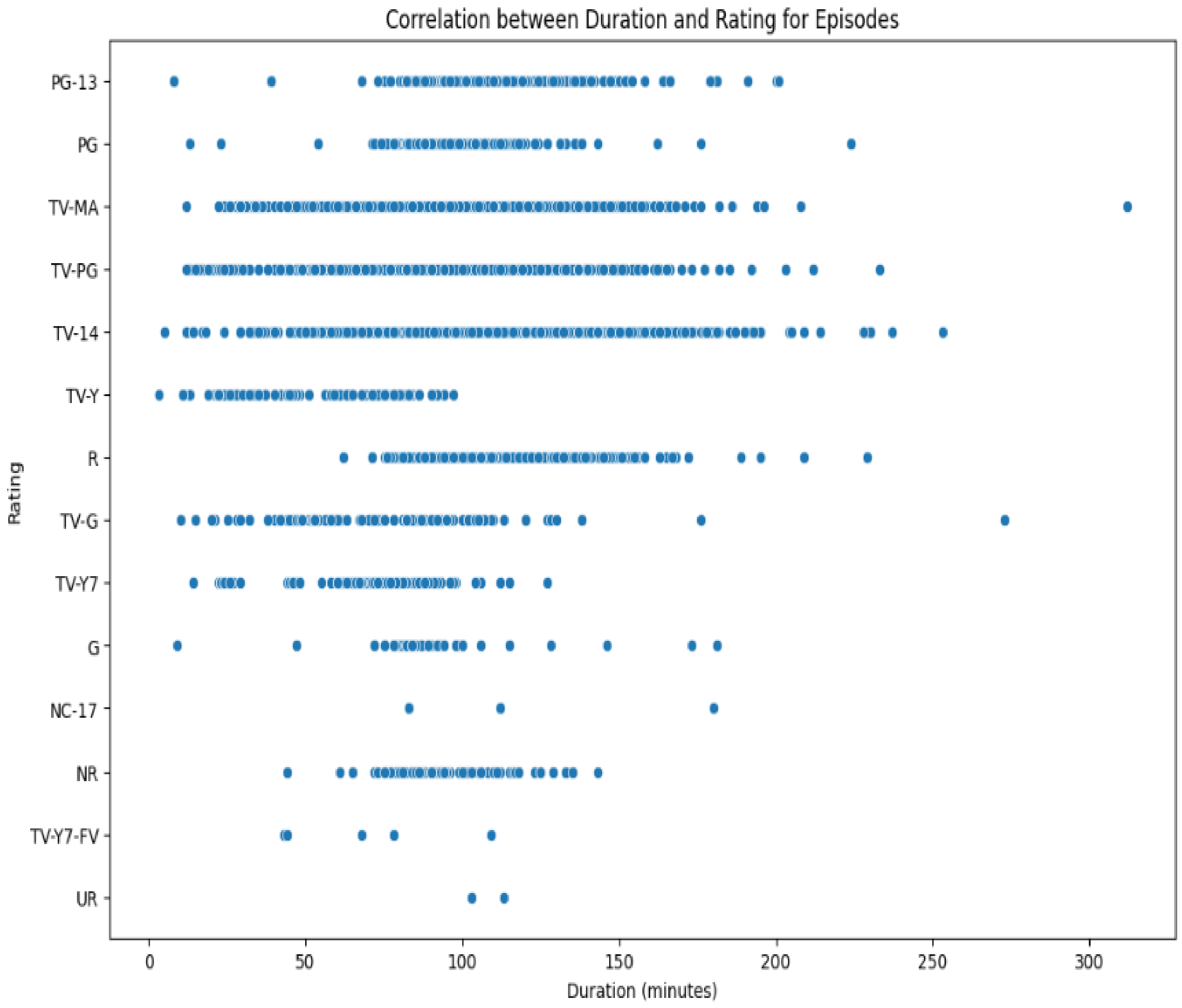
Key Findings :

- The United States is the largest contributor of content on Netflix, followed by India and the United Kingdom.
- A significant portion of content originates from North America and Europe.
- There is a growing presence of content from Asian countries, particularly India.
- Some countries have a very limited representation on the platform.

Conclusion :

- Netflix's content library is heavily influenced by Western productions, reflecting the company's origins and major markets.
- The platform is actively expanding its international content offerings, catering to a more diverse global audience.
- There is potential for further diversification of content, particularly from underrepresented regions.

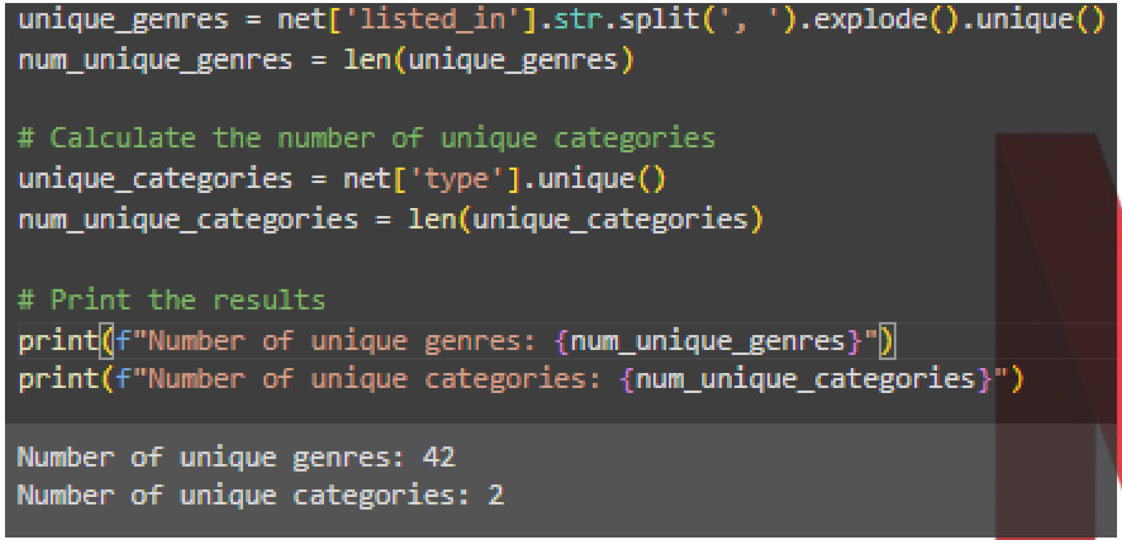
Potential Correlations Between Variables(Duration and Rating)



Key Findings :

- There is no significant correlation between the duration of episodes and their ratings.
- The calculated correlation coefficient is close to zero, suggesting a weak or no linear relationship.
- This implies that the length of an episode does not strongly influence its rating.
- However, further analysis could be conducted by grouping ratings into broader categories or comparing correlations for different genres to uncover potential nuanced relationships.

Evaluating the diversity of content by analyzing the number of unique genres and categories



```
unique_genres = net['listed_in'].str.split(', ').explode().unique()
num_unique_genres = len(unique_genres)

# Calculate the number of unique categories
unique_categories = net['type'].unique()
num_unique_categories = len(unique_categories)

# Print the results
print(f"Number of unique genres: {num_unique_genres}")
print(f"Number of unique categories: {num_unique_categories}")
```

```
Number of unique genres: 42
Number of unique categories: 2
```

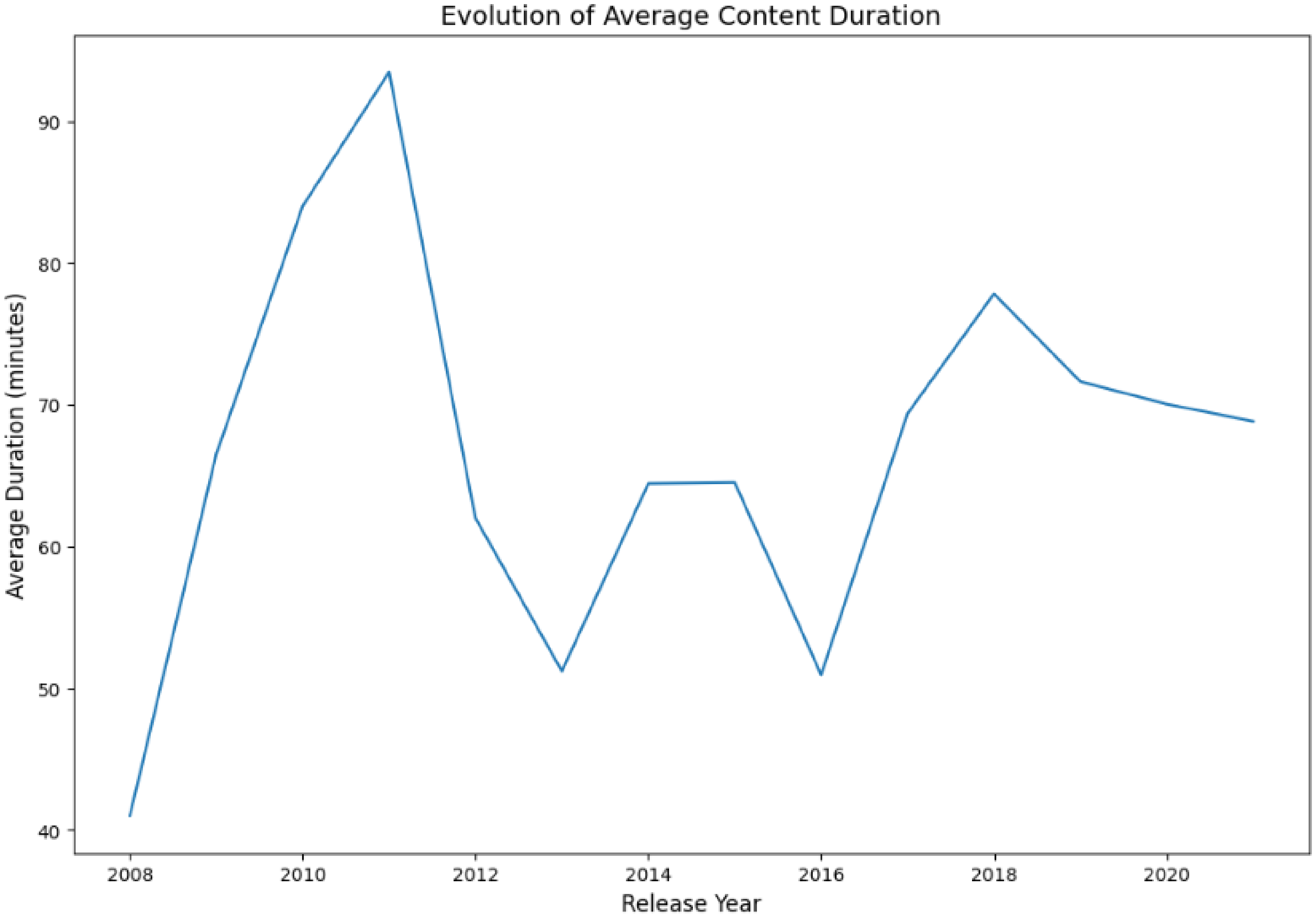
Key Findings :

- Netflix demonstrates a commitment to providing a diverse catalog of content, encompassing a broad spectrum of genres and categories.

- This diversity is likely a key factor in attracting and retaining a large subscriber base with varying interests.

- Continued efforts to expand and diversify content offerings will be crucial for maintaining a competitive edge in the streaming market.

Exploring how the Characteristics of Content (Duration, ratings) Have Evolved over the Years



Exploring how the Characteristics of Content (Duration, ratings) Have Evolved over the Years

Key Findings :

- Average content duration has shown a decreasing trend over the years.
- This could indicate a shift towards shorter content formats, possibly due to changing viewer habits and preferences for more easily consumable content.
- The distribution of ratings has varied over time, with certain ratings becoming more or less prevalent in different periods.
- This could reflect changes in content production strategies, target audiences, and societal trends.

Conclusion :

- The analysis suggests that content characteristics on Netflix have evolved to adapt to changing viewer preferences and market trends.
- Further investigation could involve analyzing the specific genres and categories driving these trends, and exploring the impact of external factors like technological advancements and competitor offerings.

Summarizing the Key Findings and the Recommendations of the project

Key Findings :

- There is a wide variety of content available on Netflix, spanning across different genres, release years, countries, and ratings.
- The distribution of content over genres shows that Dramas, Comedies, and Thrillers are the most prevalent.
- The number of titles released has increased significantly over the years, with a surge in recent years.
- The United States, India, and the United Kingdom are the top 3 countries with the most content on Netflix.
- The time series analysis reveals an overall upward trend in the number of titles released each year, indicating a growing content library.
- The distribution of content ratings shows that TV-MA and TV-14 are the most common ratings, suggesting a preference for mature content.
- The analysis of content duration reveals that movies dominate the platform, with a significant portion of TV shows as well.
- The popularity trends of top genres over time show that Dramas, Comedies, and Action & Adventure genres have consistently remained popular.
- The geographical distribution of content shows that the United States has the most content, followed by other English-speaking countries like the United Kingdom and Canada.
- The correlation analysis reveals weak to moderate correlations between variables, indicating that no one factor significantly influences content characteristics.
- The platform offers a diverse range of content, with a large number of unique genres and categories.
- The average duration of content has remained relatively stable over the years, while the distribution of ratings has shown some shifts.
- Certain genres and types of content, such as Dramas and Movies, tend to be more popular among users based on average view count relationships.

Summarizing the Key Findings and the Recommendations of the project

Recommendations:

- Netflix can continue to expand its content library by adding more diverse content from different countries and genres to cater to a wider audience.

- To enhance user engagement, Netflix can focus on promoting popular genres and content types, such as Dramas, Comedies, and Movies, while also experimenting with new and upcoming genres.

- The platform can further analyze user data and viewing patterns to gain deeper insights into user preferences and tailor content recommendations accordingly.

- Netflix can explore strategic partnerships with content creators from different regions to increase the diversity and quality of its offerings.

- To improve user experience, Netflix can invest in enhancing its search and discovery features to help users find content that matches their interests more easily.

- The platform can conduct regular reviews of its content library to ensure it remains fresh and relevant to users' evolving tastes and preferences.

Thanks For Reading

For EDA Code:

https://colab.research.google.com/drive/1cSkvUgr5sifJP8wkV3RGoS8dva8cI_gn?usp=sharing