

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.

NETFLIX

PROJECT OBJECTIVES

Thorough investigation and analysis of Netflix's content dataset is the aim of the Netflix EDA project. This entails comprehending the data structure, maintaining data integrity by managing duplicates and missing values, calculating descriptive statistics, and visualising the distribution of content among categories and release dates. The initiative also intends to evaluate audience engagement data, analyse content features like duration and ratings, and spot temporal trends. The project's goal is to synthesise these insights in order to make significant findings and propose doable suggestions for improving Netflix's user experience and content selection.

NETFLIX

DESCRIPTION OF THE DATA

All the statistics of the release year column of Netflix is given below , since other column are no integers so there are no statistics present.

```
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
```

```
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
```

There are total 8807 rows and 12 columns present in our dataset. There is only one numeric column and rest are non-numeric or categorical in nature. We will use describe command to find this.

DATA CLEANING AND PREPARATION

DEALING WITH NULL VALUES

```
[ ] 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
```

```
[10] net.isnull().sum().sum()

⇒ 4307
```

There are total 4307 null values present in the data set. Now we have to deal with these null values and we will remove them or replace them by some other value.

1. DIRECTOR: For the 'Director' attribute with 2634 null values, one approach is to fill these missing values with a placeholder such as '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

2. CAST : With 825 null values in the 'Cast' attribute, a similar approach can be applied. Filling the missing values with '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.

3. 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 meta-data to infer the country of production based on the content's origin or production company.

NETFLIX

4. DATA ADDED : With only 10 null values in the 'Date Added' attribute, filling these missing values with 'No date specified'. You can impute the missing dates by referencing the release year or utilizing the median or mode date added from the available data to maintain consistency.

5.RATING : For the 'Rating' attribute with 4 null values, filling the missing values with the 'No rating specified' from the dataset can be a suitable approach. Alternatively, you can infer the rating based on the content type (movie/TV show), genre, or other metadata attributes to assign a relevant rating.

6. DURATION : With only 3 null values in the 'Duration' attribute,. filling these missing values with 'No duration specified'.

```
[ ] 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

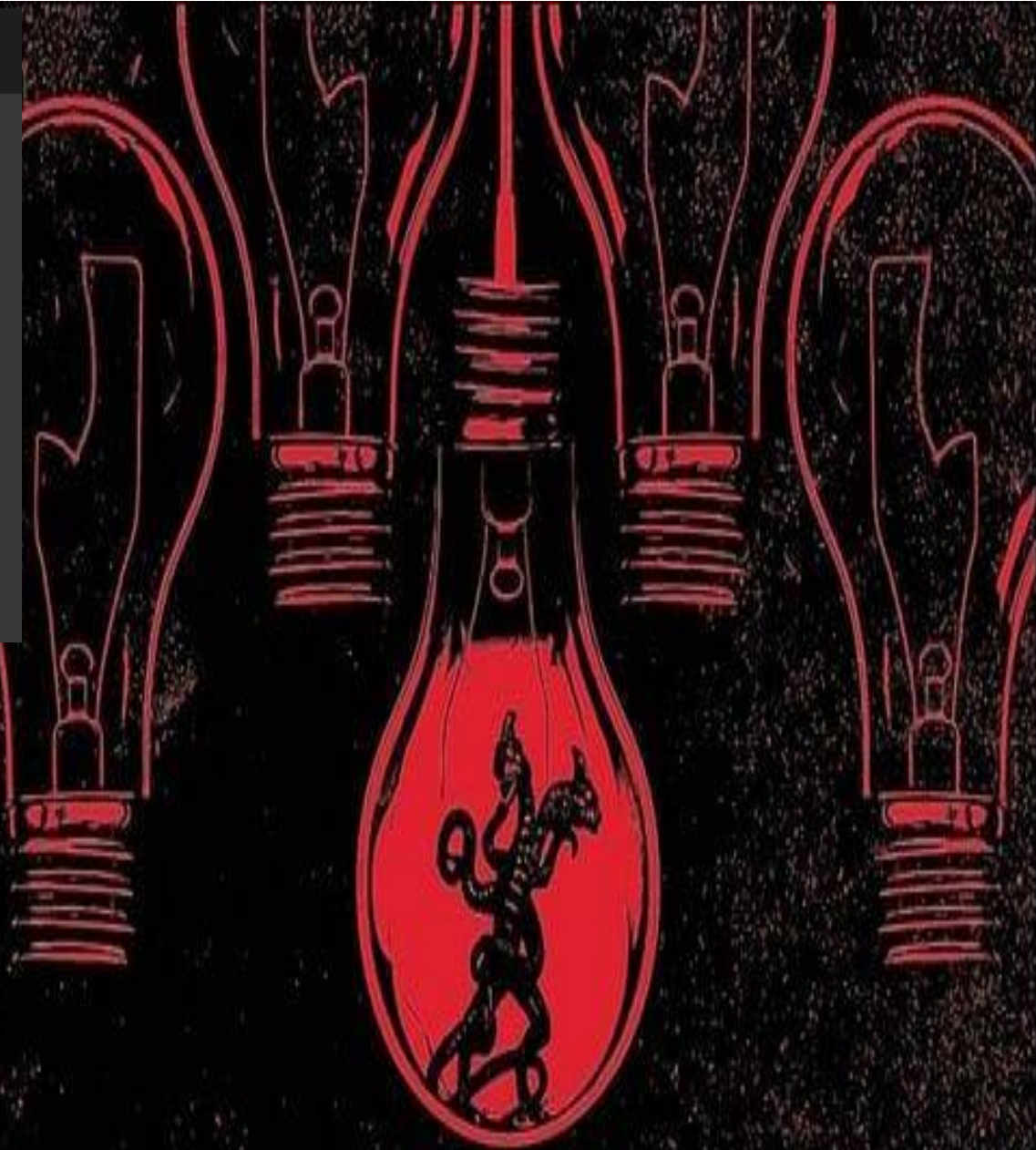
Conclusion :

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()
```

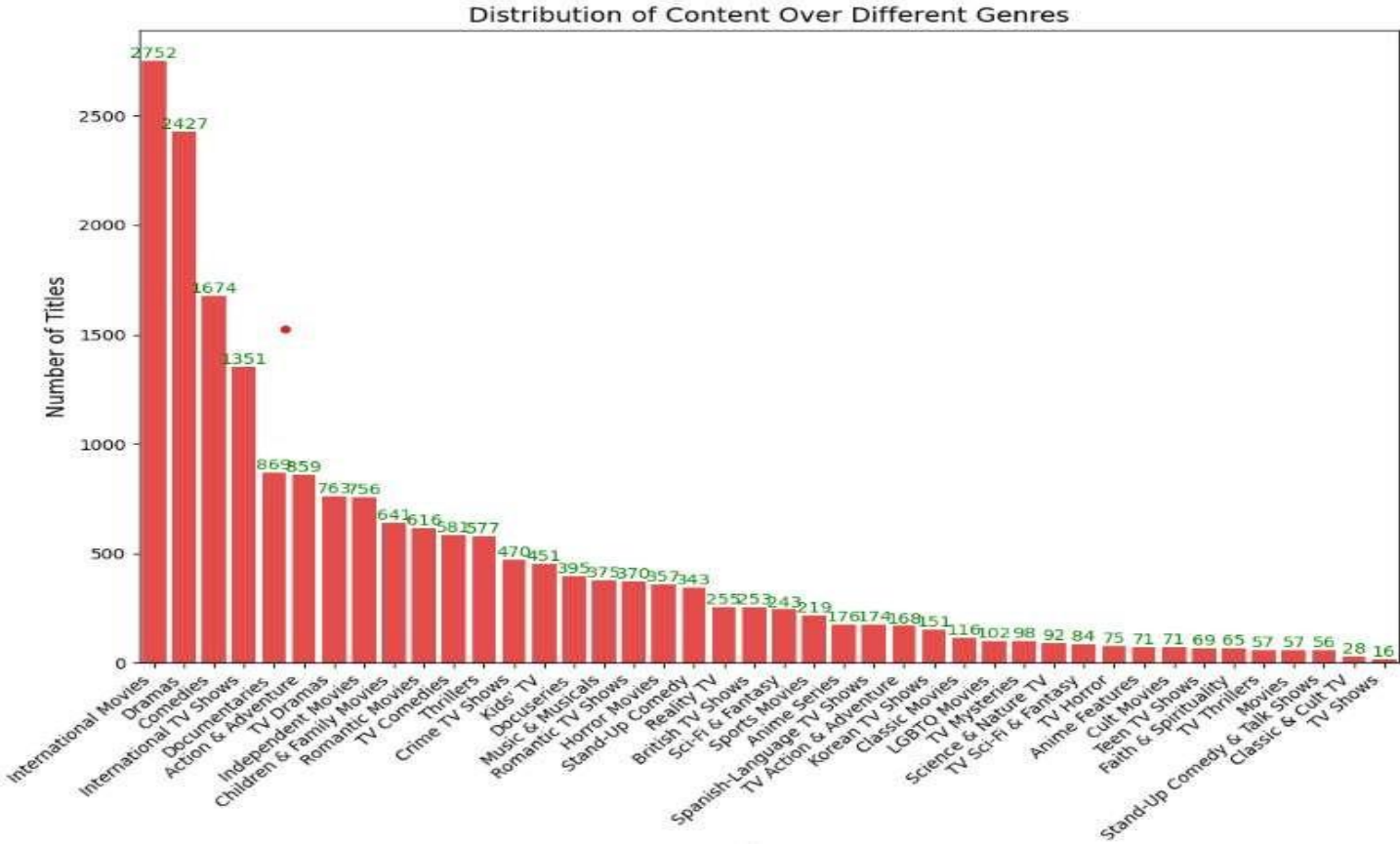
```
⇒ show_id      0  
   type        0  
   title       0  
   director    0  
   cast        0  
   country     0  
   date_added  0  
   release_year 0  
   rating      0  
   duration    0  
   listed_in   0  
   description  0  
   dtype: int64
```

STRANGER
THINGS



DATA VISUALISATION & INSIGHTS

Distribution of content over different genres:



NETFLIX

Distribution of content over different genres

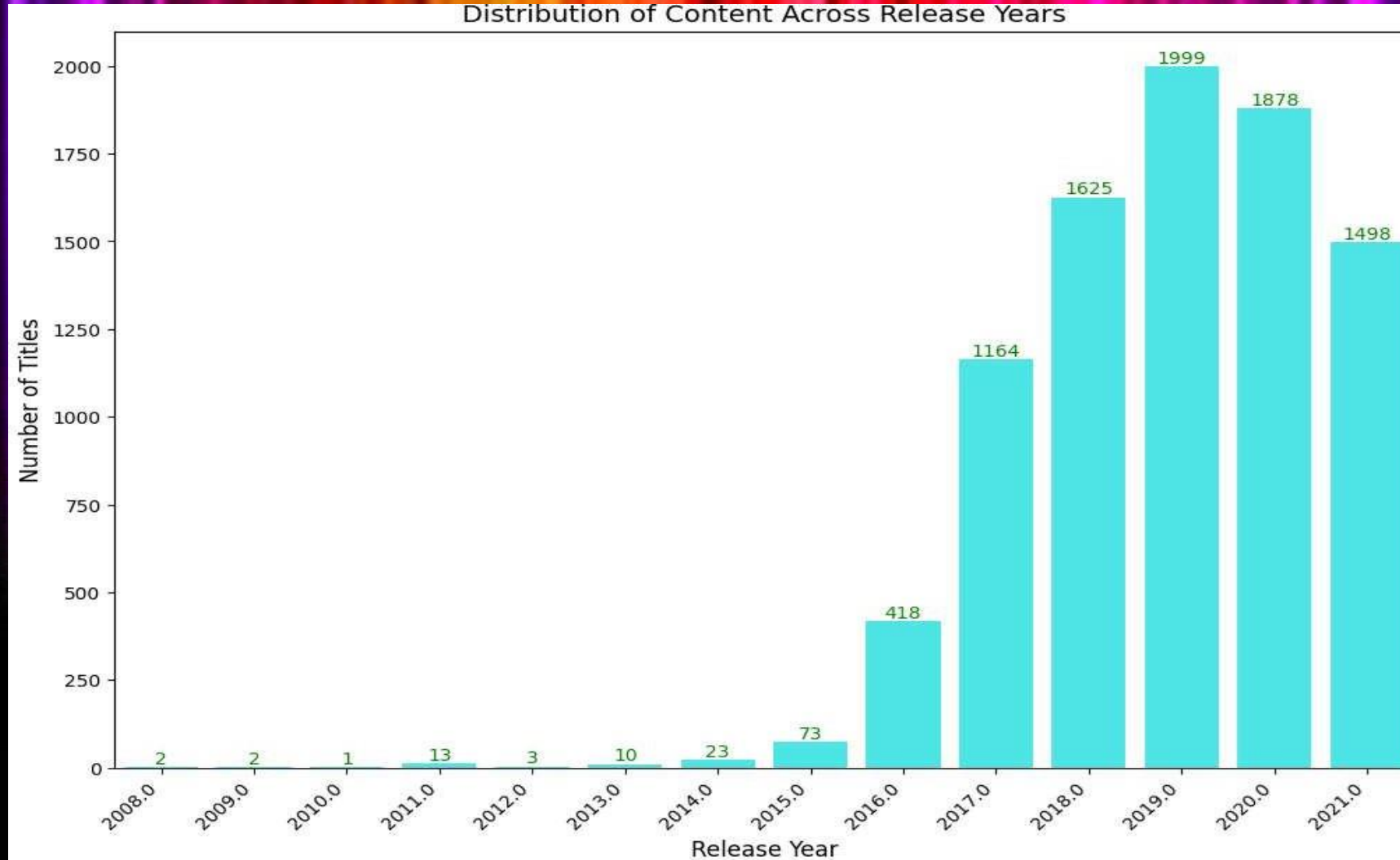
KEY FINDINGS :

- ❑ 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 the categories like “classic movies” and “independent movies”.

Distribution of Content Across Release Years



DISTRIBUTION OF CONTENT ACROSS RELEASE YEARS

Key findings :

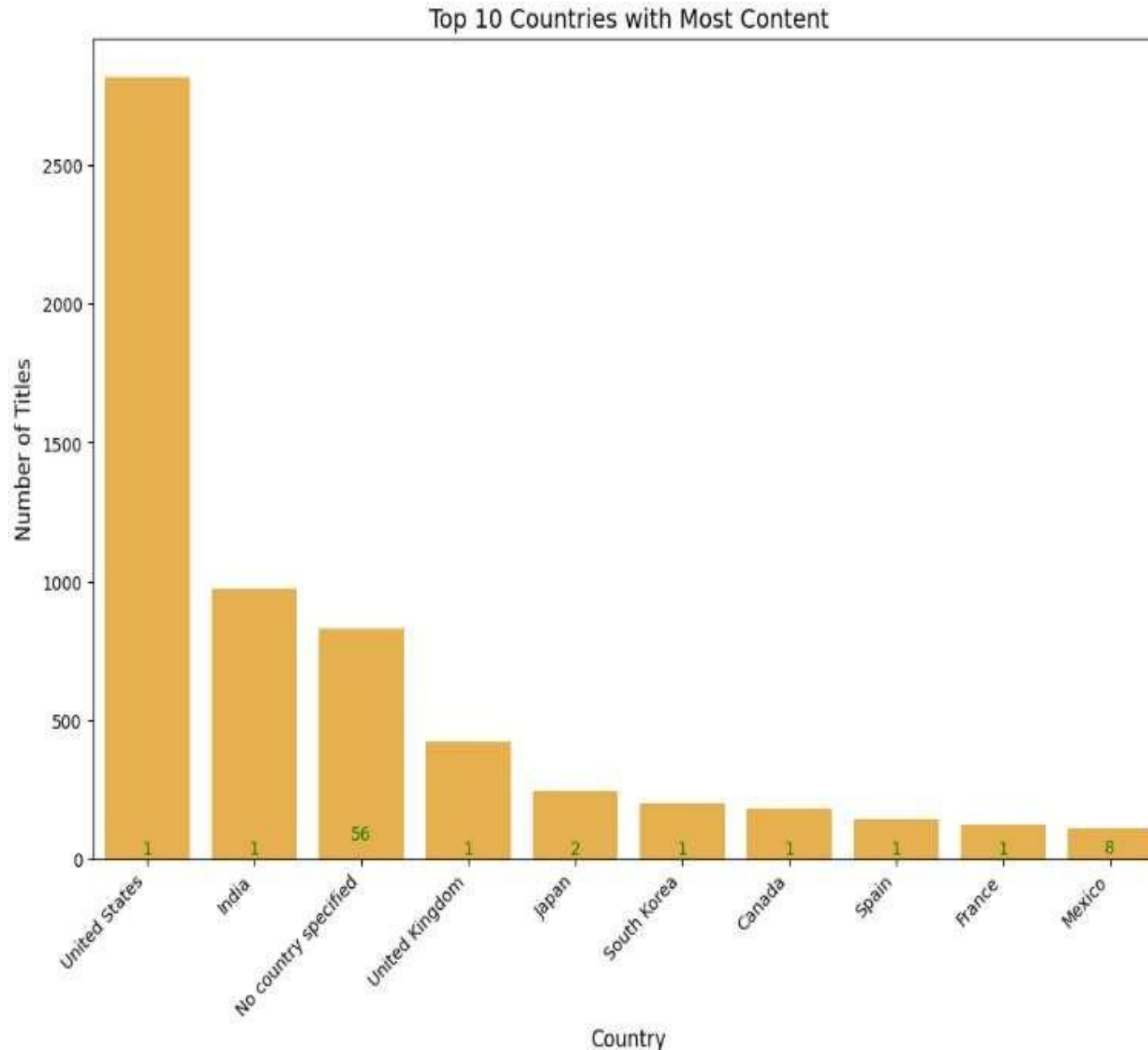
- ❑ Recent year dominate the majority of the content available on Netflix was added in recent years ,with a significant peak around 2019-2020.
- ❑ Steady growth: there is a general trend of increasing content additions over the years, reflecting Netflix's expansion and investment in the original programming.
- ❑ Older content presence while the newer content is predominating , there is still a notable amount of content from earlier years , indicating a diverse catalog.

The Netflix logo is displayed in a large, bold, red, sans-serif font. The letters are closely spaced and the overall color is a vibrant red.

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 cater to a wider audience and provides a diverse selection.

GRAPHICAL DISTRIBUTION OF CONTENT



GEOGRAPHICAL DISTRIBUTION OF CONTENT

Key findings:

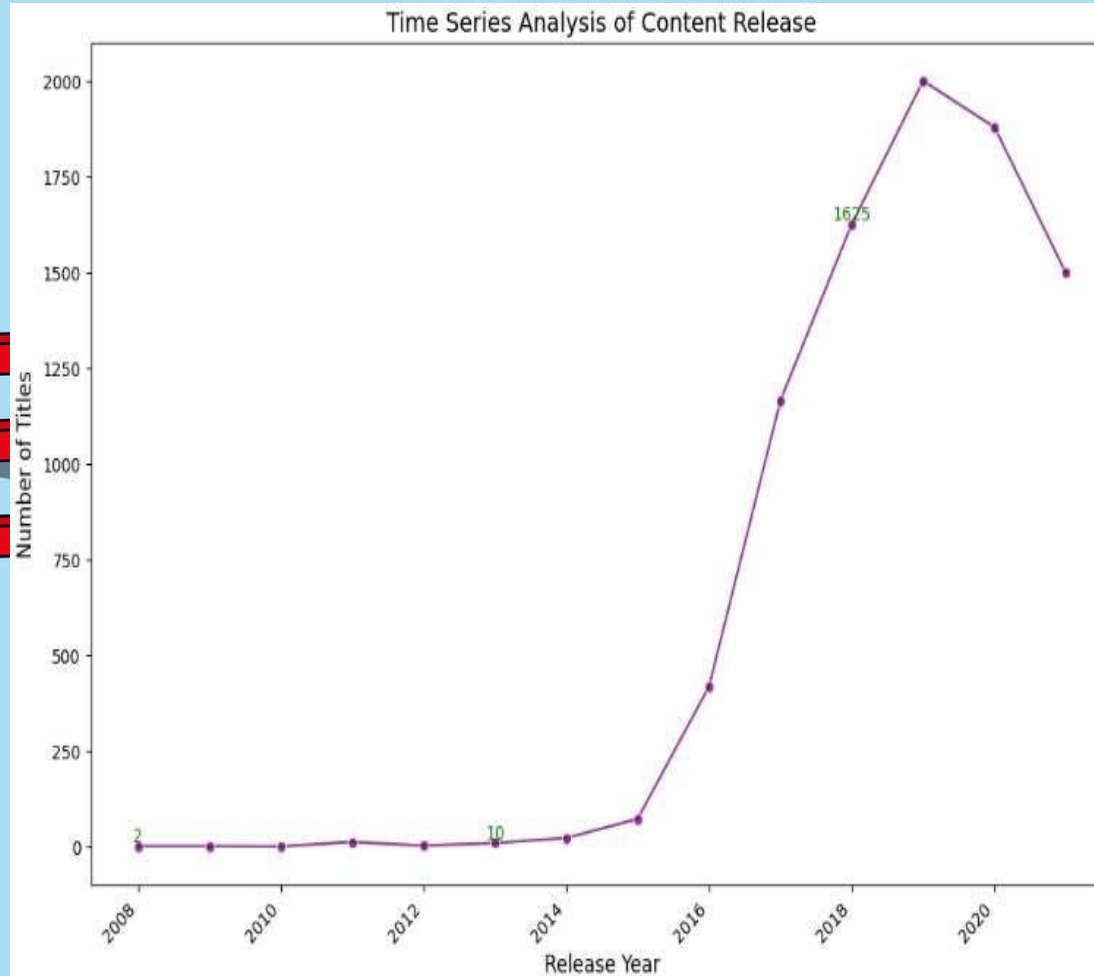
- ❑ United states dominates content production: The US is the primary contributor of the content on Netflix.
- ❑ Global representation is increasing: While the US leads, there is a notable presence of content from other countries , indicating Netflix' s effort to cater to a global audience.
- ❑ Emerging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

The Netflix logo is displayed in a large, bold, red, sans-serif font. The letters are slightly shadowed, giving it a 3D appearance. It is positioned horizontally across the middle of the slide, partially overlapping the 'Key findings' list.

Conclusion:

- ❑ Netflix's content library reflects a globalized approach ,with a strong US presence complemented by growing contributions from various countries. The 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:

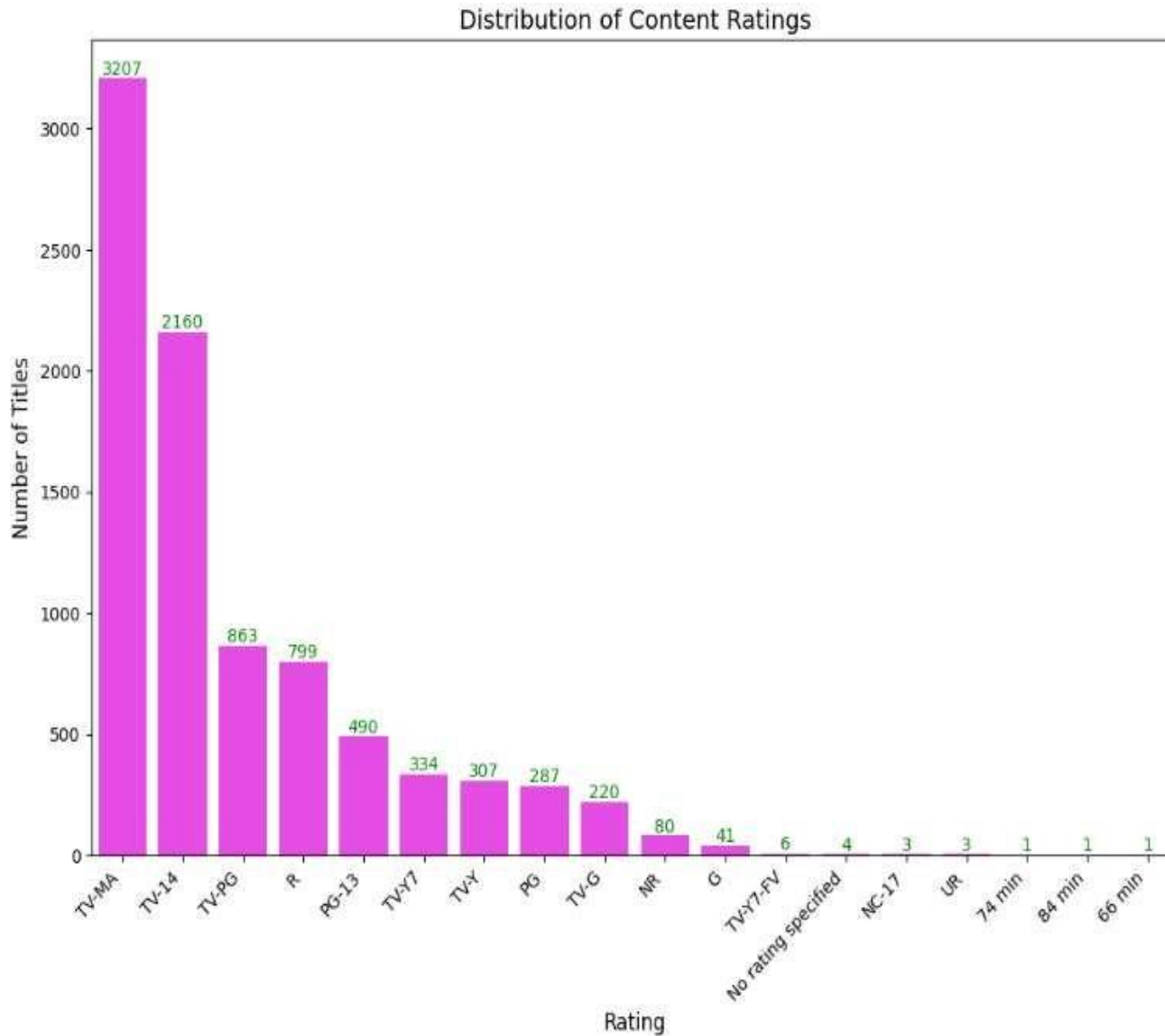
- ☐ There is general upward trend in the number of titles released on Netflix over the years.
- ☐ There might be some seasonality patterns , but the data provided does not allow for the detailed seasonality analysis.
- ☐ There might be specific years with significant or drops in content release, which would be further investigated.

The Netflix logo is displayed in a large, bold, red, sans-serif font. The letters are slightly shadowed, giving it a 3D appearance as if it's floating above a white surface.

CONCLUSION:

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

Distribution of Content Ratings



Harry



Potter

NETFLIX
&
CHILL
ELITE
STRANGER
THINGS

DISTRIBUTION OF CONTENT RATINGS

KEY FINDINGS:

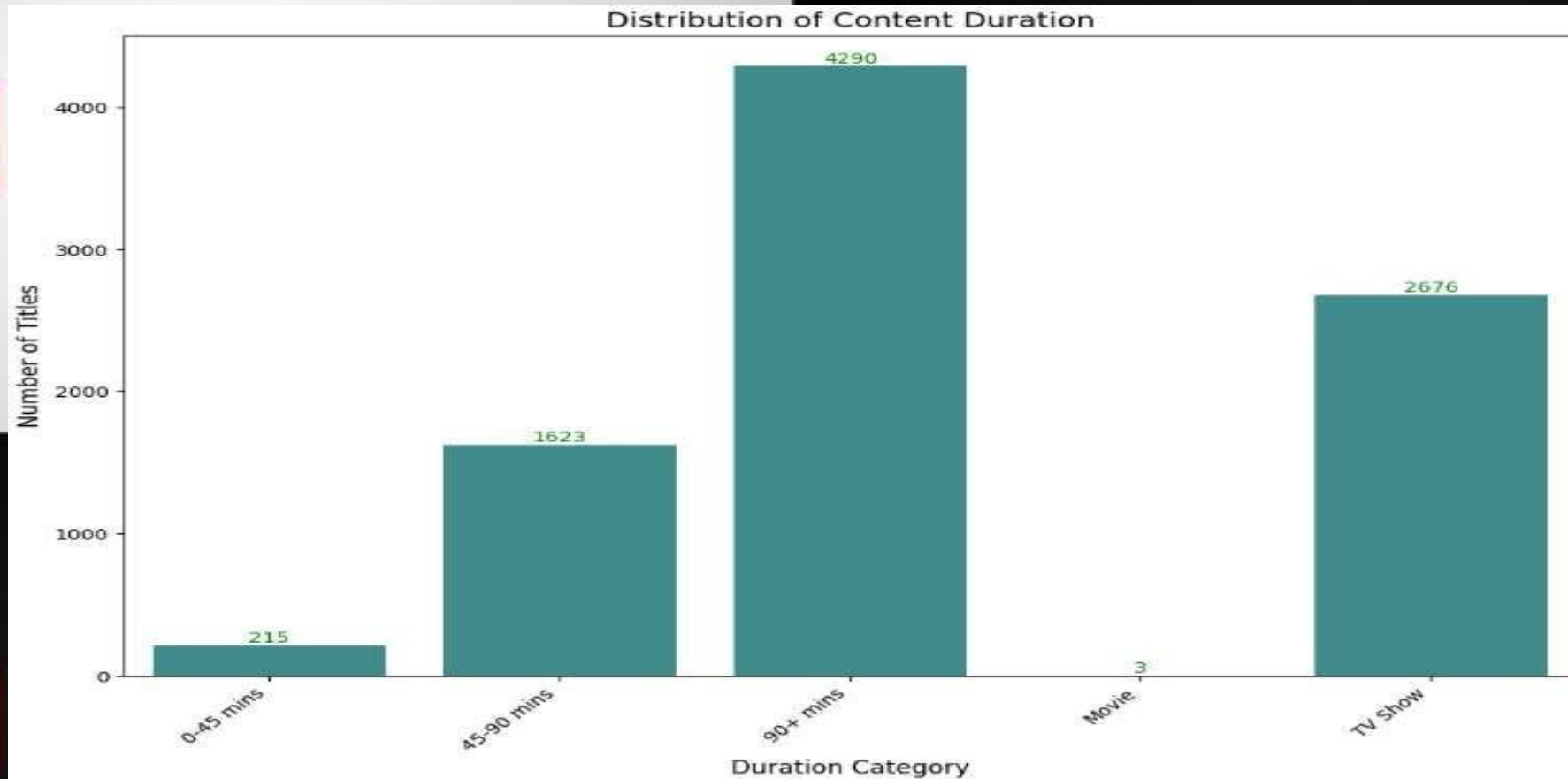
- ☐ The distribution of the content ratings on Netflix is not uniform.
- ☐ Certain ratings like “TV-MA”(mature audience) and “TV-14”(parents strongly cautioned) are more prevalent , indicating a larger proportion of content geared towards older audiences.
- ☐ This suggests that Netflix caters significantly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

The Netflix logo is displayed in a large, bold, red, sans-serif font. The letters are slightly shadowed, giving it a 3D appearance. The word "NETFLIX" is centered horizontally and takes up most of the width of the slide.

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 NO. OF TITLES

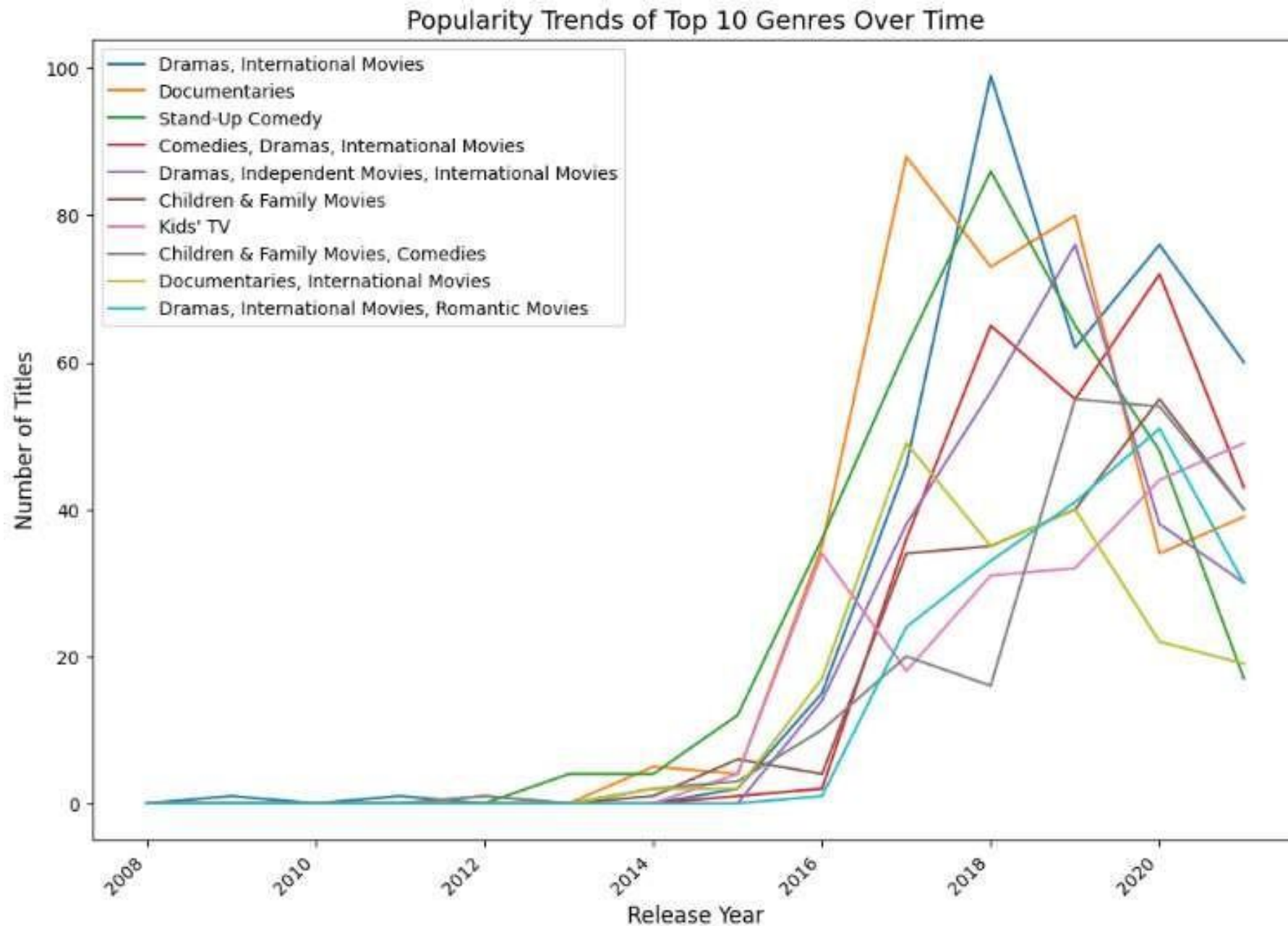
Key Findings :

- 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 not able 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 in different countries & regions.

Key Findings :

- The United States is the largest contributor of content on Netflix, followed by India and then UK.
- 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 :

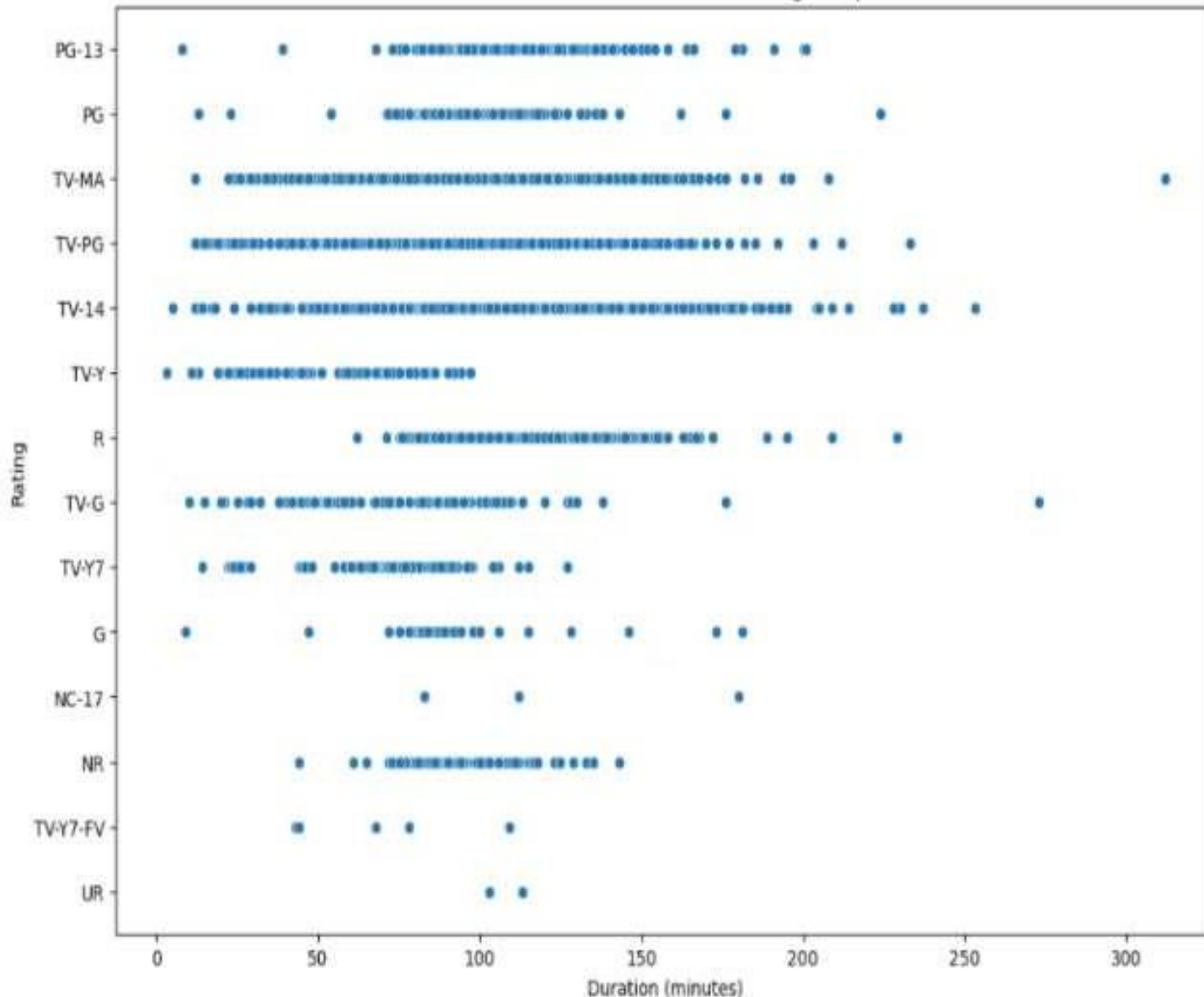
Netflix's content library is heavily influenced by Western productions, reflecting the company's origins and major markets.

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)

Correlation between Duration and Rating for Episodes



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 analysing the number of unique genres and categories

Key Findings :

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

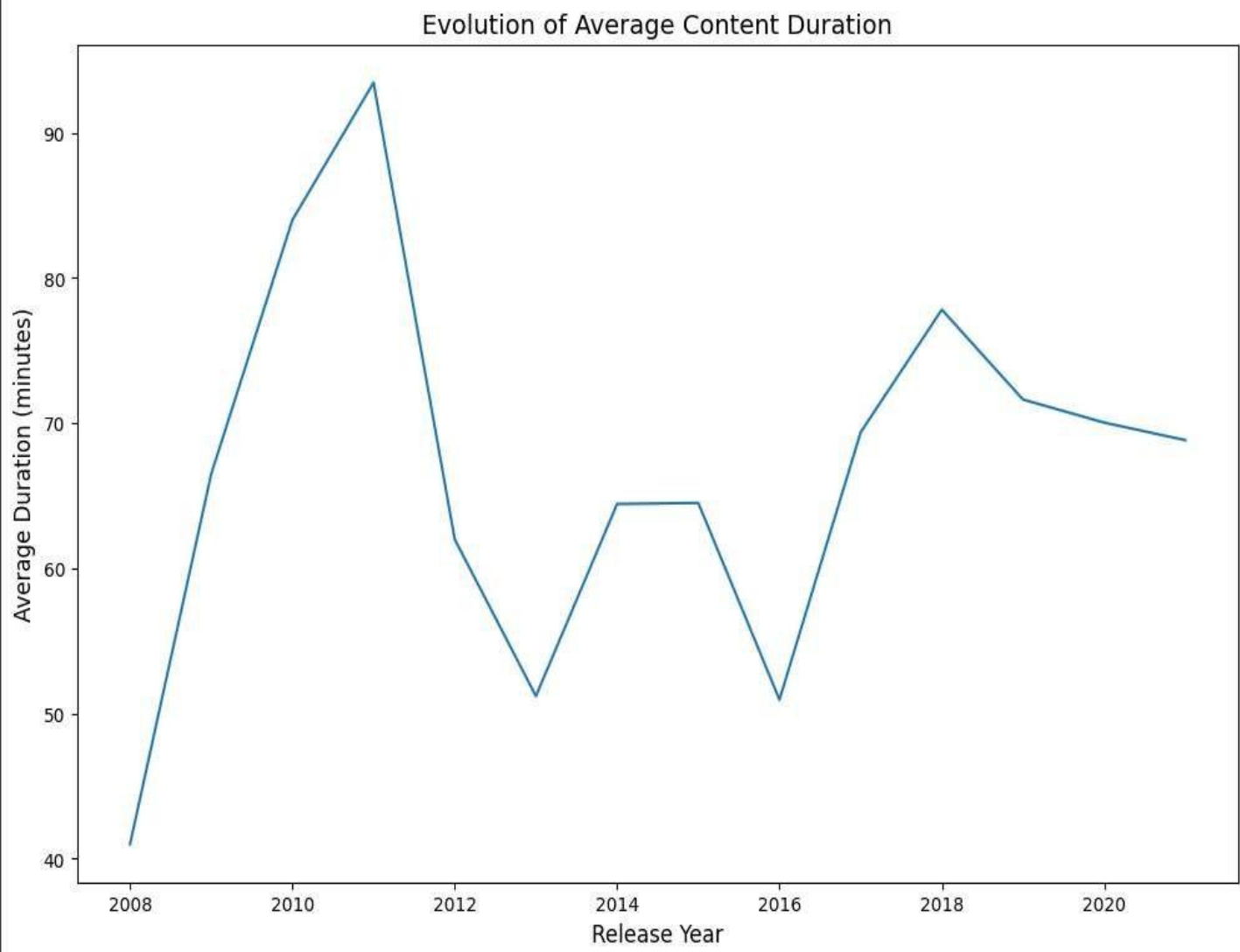
# Calculate the number of unique categories
unique_categories = net['type'].unique()
num_unique_categories = len(unique_categories)

# 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
```

- 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 analysing the specific genres and categories driving these trends, and exploring the impact of external factors like technological advancements and competitor offerings.

NETFLIX

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 analyse 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.

THANK YOU FOR READING

-BY PARAS CHAUHAN

FOR CODING PART KINDLY VISIT:-

https://colab.research.google.com/drive/1r_EcQy3BzAMXYGACm8t0EZC3p_ucVldZ?usp=sharing