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

PURPOSE OF THE PROJECT

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.

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.

```
In [5]: df.describe()
```

Out[5]:

	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

-- This dataset consist of a single row with numerical values.

```
In [6]: 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
```

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

```
In [7]: df.isnull().sum()

Out[7]: 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.

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 'Unknown' or 'Not 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.

2. Cast: With 825 null values in the 'Cast' attribute, a similar approach can be applied. Filling the missing values with 'Unknown' or 'Not 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 'Unknown' can be a feasible approach. Another strategy is to crossreference with the title or other metadata to infer the country of production based on the content's origin or production company

4. 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 referencing the release year or utilizing the median or mode date added from the available data to maintain consistency.

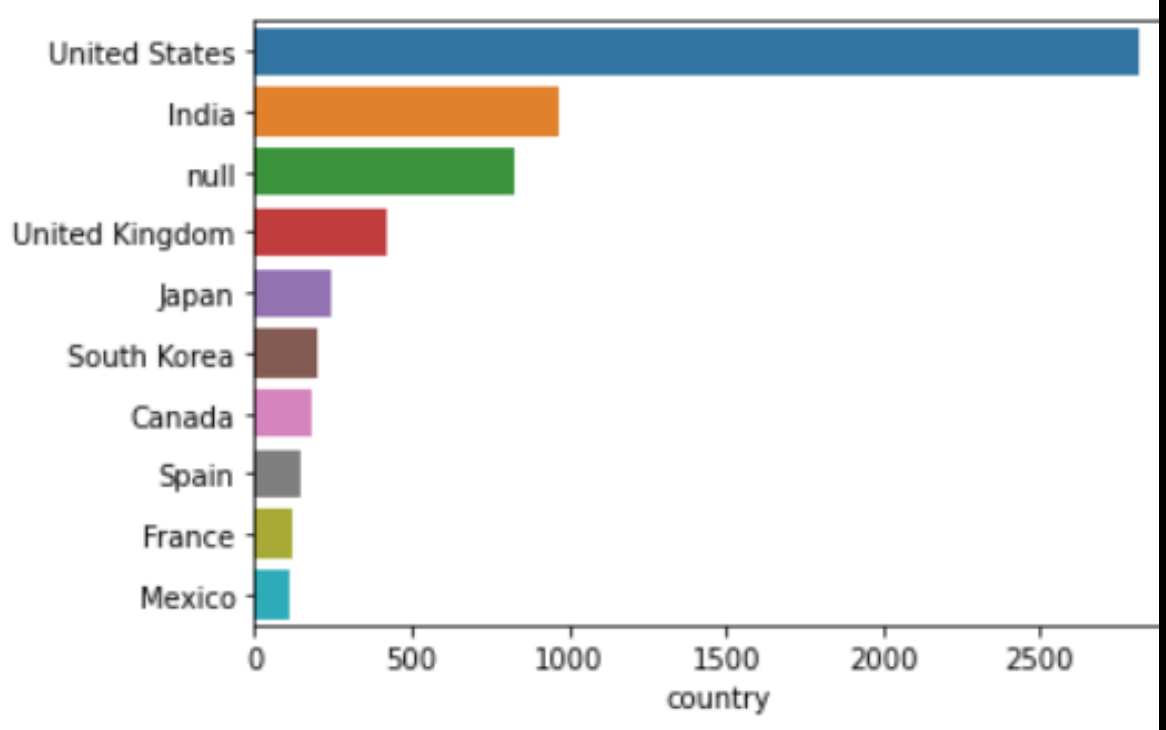
5. Rating: For the 'Rating' attribute with 4 null values, filling the missing values with the mode or most frequent rating from the dataset can be a suitable approach. Alternatively, you can infer the rating based on the content type (movie/TV show), genre, or other metadata attributes to assign a relevant rating.

6. Duration: With only 3 null values in the 'Duration' attribute, I have deleted the data which was having null values because these 3 will not harm our findings .

TOP 10 CONTENT CREATING COUNTRIES ARE AS FOLLOWS-

```
df.country.value_counts().head(10)
```

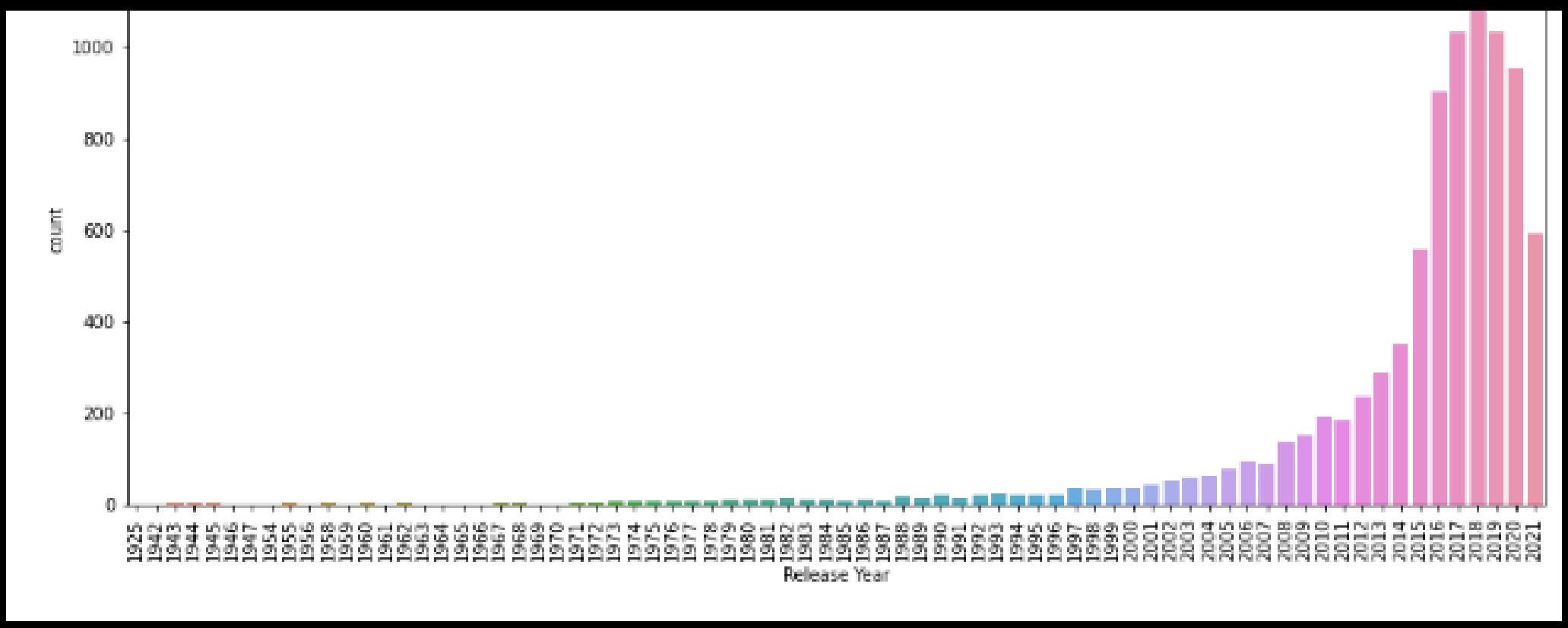
```
United States     2818
India              972
null               831
United Kingdom     419
Japan              245
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Name: country, dtype: int64
```



Observations and insights from the above derived graph:-

USA id dominating the market of netflix by capturing the 53.3% area. This shows the dominance of both technology and talent.
India has grown at a very steady rate after the formation of Bollywood industry and is at second in the list. The world is recognizing the Indian talent. India covers nearly 20% of the market.
United Kingdom is on the third number with 8 percent of the netflix market and have produced 418 shows and movies.
Japan is on the 4th number with 4.6 percent and 243 movies.A potential reason for this number could be explored in further research. Changes in viewer preferences, industry dynamics, or other factors might contribute to this observed decline.
We can also see from the pie chart that there are 829 shows (app 15%)of which the entries of country are not known. The reason of this could be anything like Error of non-response, error of misinterpretation or error of sampling bias. Though the above unknown data entry also suggest that netflix serves as a platform of a diverse range of creators. irrespective of there societal standing or production scale.

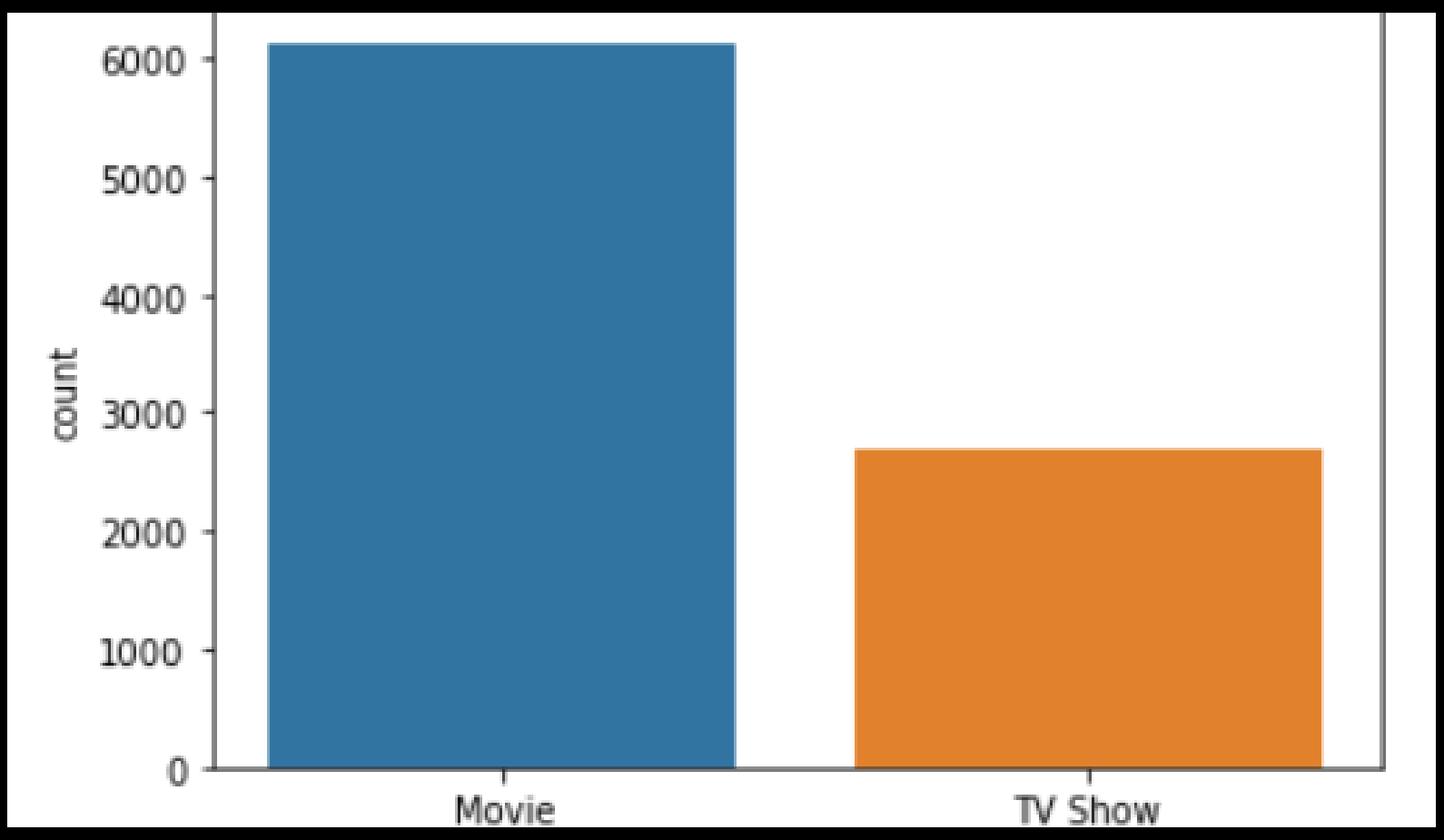
ANALYZING THE TREND OF RELEASE YEAR



From the above graph we can interpret following results:-

1. From 1997 to 2014 there is a gradual increase in the release of content.In these years there is a boom in internet consumption and therefore we can see the boom.
2. 2015 was a big year for Netflix. To start, the video streaming behemoth dropped more than a dozen new, original shows. Like these gem
   --The delightful Unbreakable Kimmy Schmidt
   --Marvels gripping Jessica Jones
   --Aziz Ansaris hilarious Master of None
3. In 2016 to 2020 there was a increase in Asia pacific cosumers therfore we can see a boom in netflix.
4. In year 2021 we can see the graph going down because of the deadly covid pandemic.

COMPARISON BETWEEN MOVIES AND TV SHOW.

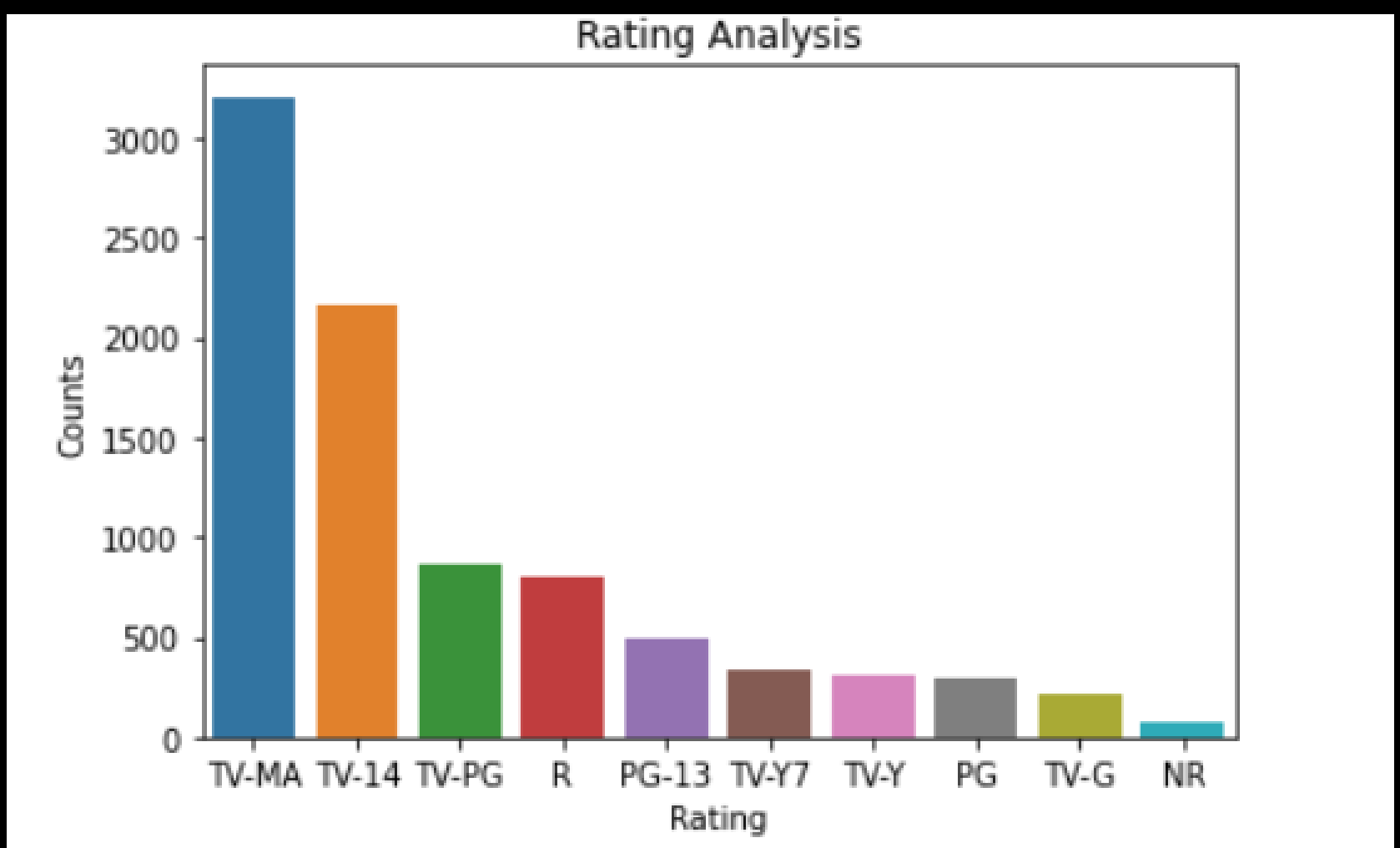


There are total 6131 movies and 2676 TV show present on netflix based on given data set. Above is the Bar graph representing the same thing in drawings.

RATING ANALYSIS OF TOP 10 COUNTRIES

```
df.rating.value_counts().head(10)
```

```
TV-MA    3207
TV-14    2160
TV-PG     863
R         799
PG-13     490
TV-Y7     334
TV-Y      307
PG        287
TV-G      220
NR         80
Name: rating, dtype: int64
```



1.T.V-MA (Matured Audience)
TV-MA leads the ratings with 3205 entries, demonstrating a significant market for entertainment geared towards adults. Coarse language, graphic or extreme violence, nudity, and explicit sexual content are frequently included in this category. The large number indicates that there is a sizable readership that is older than 18.

2. TV-14(Parents Strongly cautioned)
Closely behind, TV-14 (2157 entries) denotes material for viewers 14 years of age and up. These shows target a large teenage audience, since they may contain content that parents deem inappropriate for children under the age of fourteen. Purchasing material for this audience may be a wise strategic move.

3. TV-PG(Parental Guidance Suggested)
With 861 listings, TV-PG indicates material appropriate for kids with parental supervision. This rating points to a strong family and kid audience base. Netflix may take this into account when deciding how much more family-friendly programming to offer.

4. R(Restricted)
R-rated material (799 entries) is meant for viewers older than eighteen and is not appropriate for younger audiences. The noteworthy number suggests that adult viewers have a great appetite for mature entertainment with dramatic sequences. Knowledge of this consumer behaviour can help direct the creation of content

5. PG-13(Parents strongly cautioned)
With 490 entries, the PG-13 grade denotes content with a Parents Strongly Cautioned warning because some of it might be too mature for young readers. This implies that preteens and early adolescents have a noticeable need for content that might involve bolder language, prolonged violence, sexual encounters, or drug usage. Acknowledging this need enables Netflix to provide material that is appropriate for a wide range of age groups.

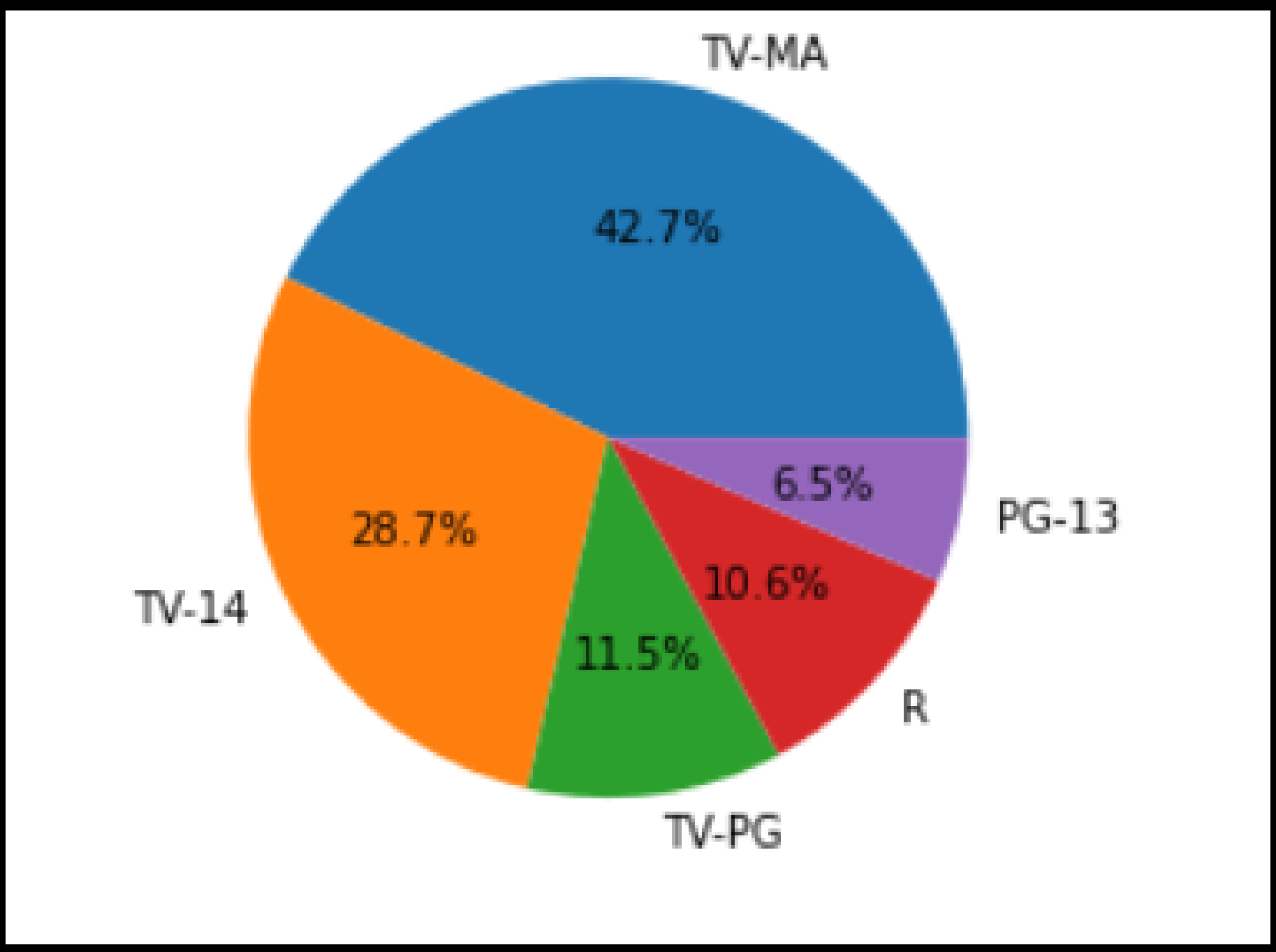
6. TV-Y(All Children*
TV-Y7, which has 333 entries, is aimed at older kids. The decreasing number may be explained by a preference for information catered to a particular age group or by a decreased emphasis on content creation for older kids.

7. PG(Parental Guidance)
With 287 entries, PG advises parental supervision. A shift in audience preferences towards more mature or specialised content may have contributed to the fall by lowering the demand for programming that calls for parental supervision.

8. TV-G(Genral Audience)
A potential need for more specialised or age-specific content could lead to a decline in TV-G, which has 220 entries for a broad audience, and a move away from programming for general audiences.

9. NR(Not Rated)
With 79 entries classified as not rated (NR), the drop may be attributed to a predilection for explicit-rated content, or it may simply reflect a lesser portion of the content library overall.

VISUALISING THE MARKET SHARE OF RATED AUDIENCE.

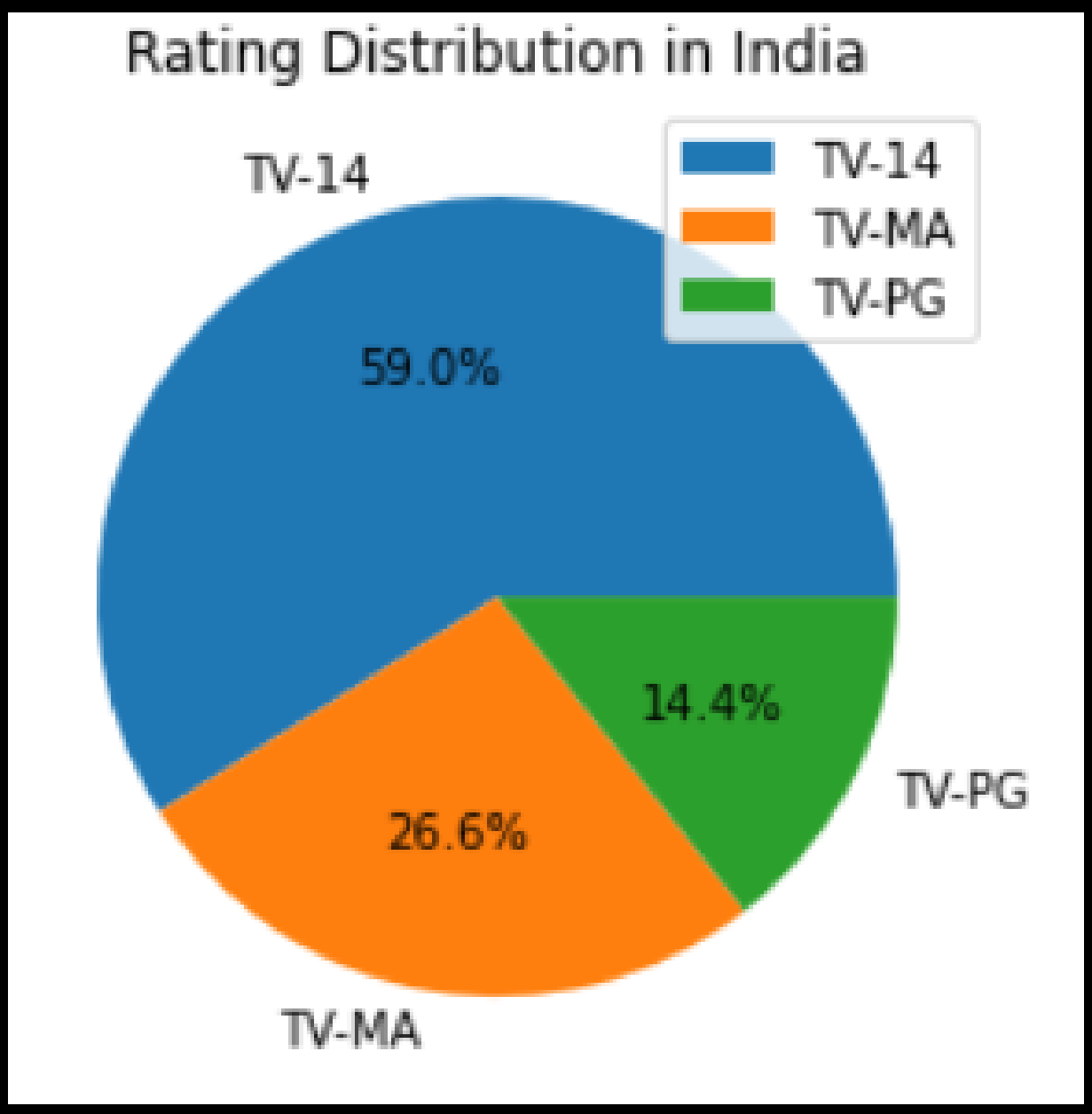


TV-MA is the network with the biggest market share, accounting for about 42.7% of the top 5 ratings. This dominance demonstrates a sizable audience base for explicit and intense content and highlights the enormous demand for adult content.

TV-14 is in second place with a notable share, suggesting that viewers looking for content appropriate for viewers 14 years and older find it appealing. TV-MA and TV-14 together have a big impact on Netflix's content lineup.

RATING DISTRIBUTION IN INDIA

```
TV-14     550
TV-MA     248
TV-PG     134
TV-Y7      14
TV-G        9
Name: rating, dtype: int64
```



TV-14

Almost 60 percent indians with 550 entries dominates this range of rating. This indicates a substantial demand for content suitable for viewers aged 14 and older who are teenager.

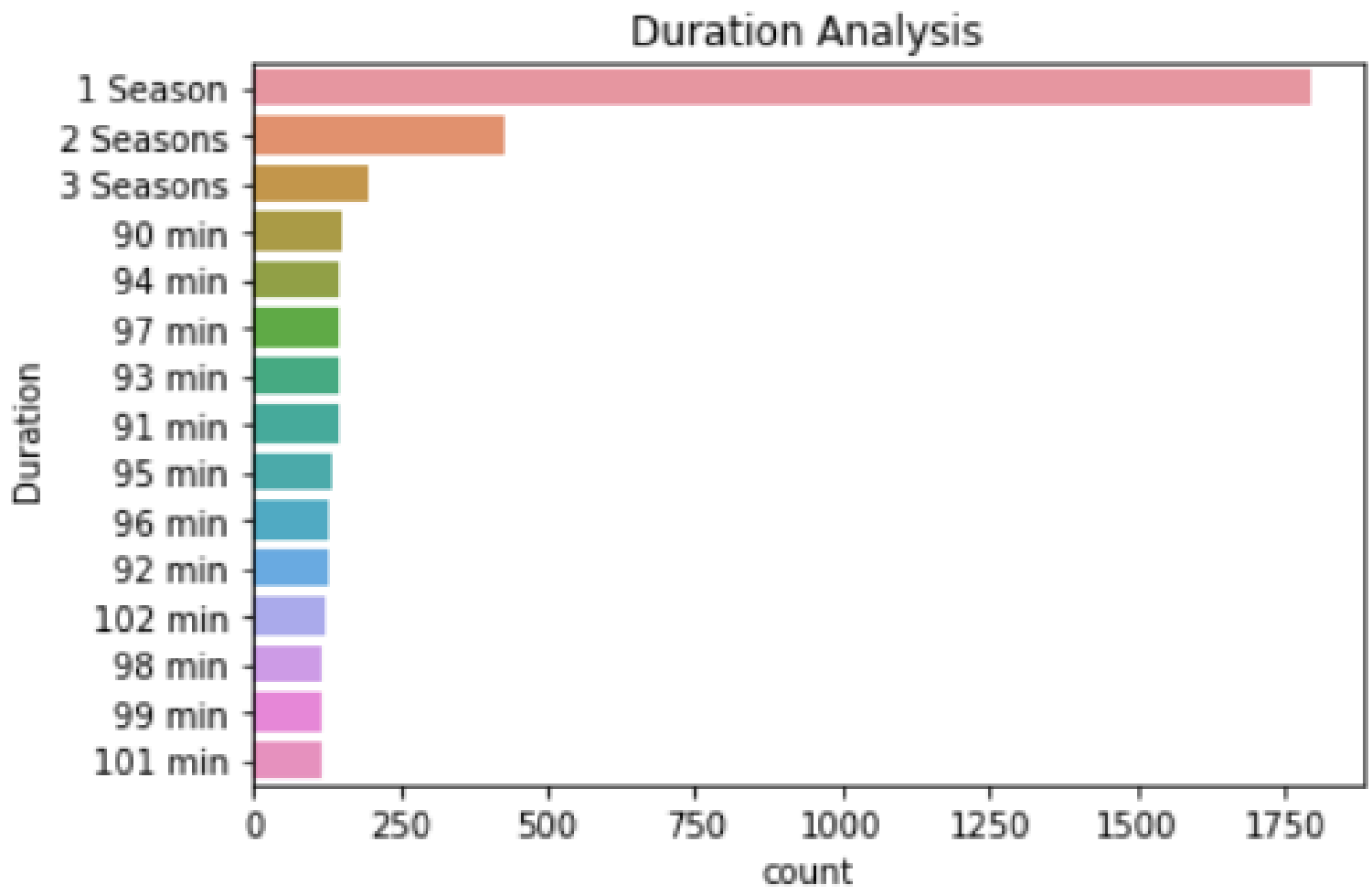
TV-MA

TV-MA has 248 entries and 27 percent of the market share. This shows the demand of the matured audience group content, emphasizing the Indian market in this criteria.

TV-PG

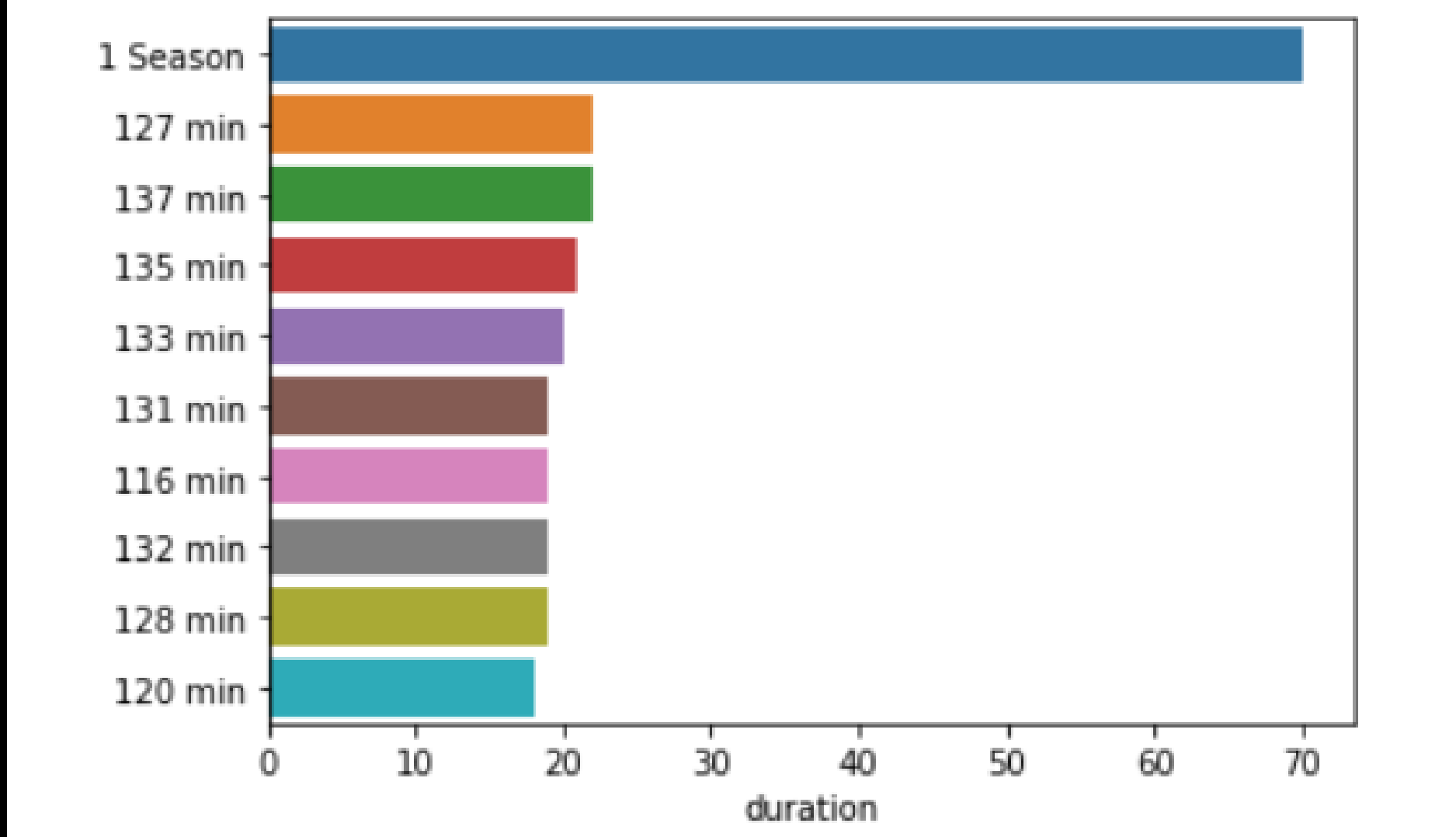
It indicates the audience segment in India that wants content suitable for children.

DURATION ANALYSIS



--From the above graph and we can see that series are preferred over the movies.

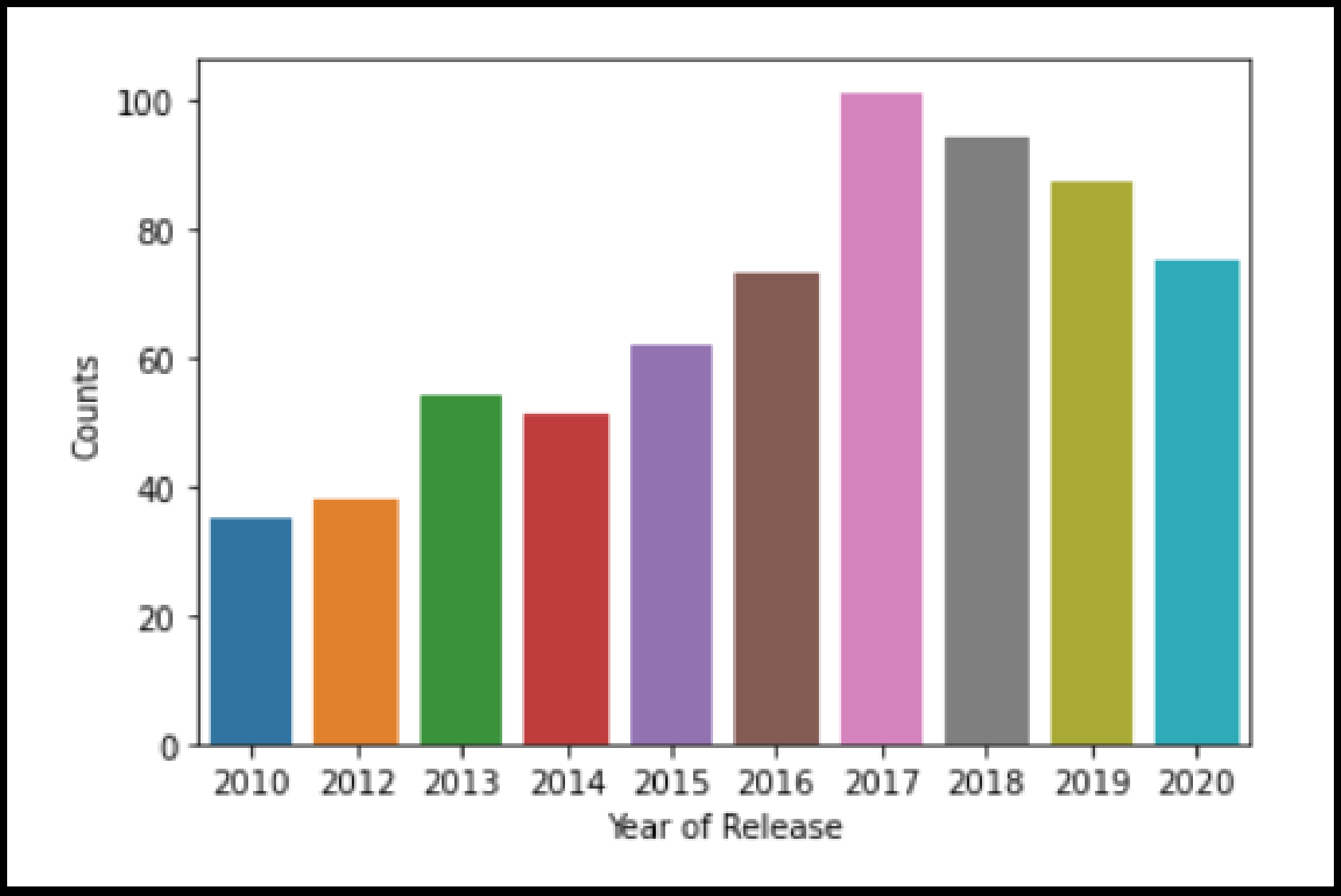
DURATION ANALYSIS OF INDIA



In India maximum people like the first season of the TV show which is quite concerning. The category of season 2 and season 3 are no where in the top 10 list which means that very few Indian are liking the sason 1 and then watching the further season. Netflix have to work on it.

MOST RELEASE IN INDIA YEAR WISE

```
2017    101
2018     94
2019     87
2020     75
2016     73
2015     62
2013     54
2014     51
2012     38
2010     35
Name: release_year, dtype: int64
```



From the above graph we can see that before 2016 there was a very slow growth of netflix in the Indian market. Netflix was launched in India in 2016 but started operating and streaming content in India in April 2017. Netflix initially focused on the English-speaking audience in India who were interested in popular international TV series and Hollywood movies. However, after understanding the lingual diversity of India, Netflix started streaming vernacular content for the Indian audience. In partnership with leading local production houses, Netflix developed original content in vernacular languages for Indian consumers. Prominent production houses with whom Netflix partnered include T-Series, Red Chillies Entertainment, Pooja Entertainment, Viacom 18 Studios, Luv Films, Reliance Entertainment, RSVP Movies, Benaras Media Works, Maddock Films, Junglee Pictures, Balaji Telefilms and Matchbox Shots

FINAL REPORT

INTRODUCTION

The results of an exploratory data analysis (EDA) project on a Netflix content dataset are compiled in this report. Understanding user preferences, content trends, and the kinds of films and TV series that are available on the platform were the main objectives.

DATA EXPLORATION

The data contained details about TV series and films, such as ratings, length, genre, release year, and (perhaps) geographic distribution in addition to user reviews and viewing information.
Understanding the data structure, determining the types of variables, and examining the data for errors, duplication, and inconsistencies were the first steps in the process. After that, the data was cleaned and made ready for analysis.

DESCRIPTIVE STATISTICS

To comprehend key tendencies and the distribution of the data, basic statistics like mean, median, mode, range, and standard deviation were computed for pertinent variables like ratings and duration.

FINAL REPORT

CONTENT ANALYSIS

User ratings were used to determine the highest rated films and television programmes.
Trends in genre popularity throughout time were examined.
Further investigation was done into the distribution of content among various nations and areas (if applicable).

Conclusions and Recommendations

Important information regarding Netflix content and user preferences was obtained via this EDA project. Important recommendations can be made in light of the findings for:
methods for acquiring material that maximise diversity of content and satisfy user preferences.
Using personalisation algorithms, content is recommended that is specific to each user's preferences.
techniques for creating content that prioritise popular formats and genres while encouraging the discovery of obscure ones.

FINAL REPORT

Future Work

Further analysis could involve:

Predicting user ratings or viewership using machine learning models. Analyzing user demographics to understand preferences based on factors like age, location, etc. Recommendation system development based on the findings of this EDA project.

Note: This report provides a general framework for the Netflix content EDA project. The specific findings and recommendations will depend on the actual data analyzed.

SUMMARY

OVERVIEW

We dug deep into a massive dataset containing details about Netflix films and television series. We set out to investigate the data, clean it up, and analyse different parts of it in order to find interesting patterns and trends.

Content Trends

Popularity by Genre: The most popular genres were found to be drama and comedy, which were followed by thriller and action.

Patterns of Release:

Over time, content has grown consistently, with a notable uptick in the last few years.

Engagement of the Audience

User Ratings: 'Good' to 'Very Good' rating categories are occupied by most content, suggesting a generally positive response from the audience.

User Engagement:

Popular shows and material received a lot of views and watch times.

Content Characteristics Duration:

The average duration of movies is around 120 minutes, while TV episodes tend to be around 30- 40 minutes. Language and Region: English remains the dominant language, but there's a growing diversity with content from various countries and languages.

SUMMARY

PROJECT SUMMARY: ANALYSING NETFLIX CONTENT TRENDS

Insights and Implications Content Strategy: Focus on producing more Drama and Comedy content, considering their popularity. User Engagement: Enhance user engagement by promoting top-rated and trending content. Localization: Explore opportunities to diversify content by introducing more regional and languagespecific shows and movies.

CONCLUSION

Our analysis provided valuable insights into the content landscape on Netflix, revealing trends in genres, audience preferences, and content characteristics. Understanding these patterns can guide content creators, marketers, and decisionmakers in making informed choices to enhance user experience and drive engagement on the platform.