

Capstone Project 4

Netflix Movies and TV Shows Clustering

by

mohammad Zabi ur rahman

Komal

mohd abdul samad

Content

- Introduction
- Problem Statement
- Data Description
- EDA
- Text Pre- Processing
- Feature Selection
- Performance Metrics
- Observations
- Conclusion

Introduction

- Netflix was founded in 1997 by Reed Hastings and Marc Randolph in a small California city called Scotts Valley in Santa Cruz county.
- Netflix is famed for its plethora of content and it's subscription based services in entertainment industry.
- After a few iterations in its first few years, Netflix eventually crafted a successful business model: a subscription-based service with no due dates or late fees and unlimited access to content at \$19.95.
- The Netflix Recommendation Engine's precise recommendations account for 80% of the Netflix viewer activity. Clustering plays a significant role in building recommendation engines helping group similar content.

Problem Statement

- In this project we'll be working with Netflix data to interpret latest trends and gain insights on the content listed, the dataset is collected from Flixable which is a third-party Netflix search engine.
- As TVShows popularity has been increasing on Netflix, it was about time that a recommended system was created.
- To deliver it, we are going to analyse the data and Cluster similar content by matching text-based features

Data Description

- `show_id` : Unique ID for every Movie / Tv Show
- `type` : Identifier - A Movie or TV Show
- `title` : Title of the Movie / Tv Show
- `director` : Director of the Movie
- `cast` : Actors involved in the movie / show
- `country` : Country where the movie / show was produced
- `date_added` : Date it was added on Netflix
- `release_year` : Actual Releaseyear of the movie / show
- `rating` : TV Rating of the movie / show
- `duration` : Total Duration - in minutes or number of seasons
- `listed_in` : Genre
- `description`: The Summary description

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 7787 entries, 0 to 7786
```

```
Data columns (total 12 columns):
```

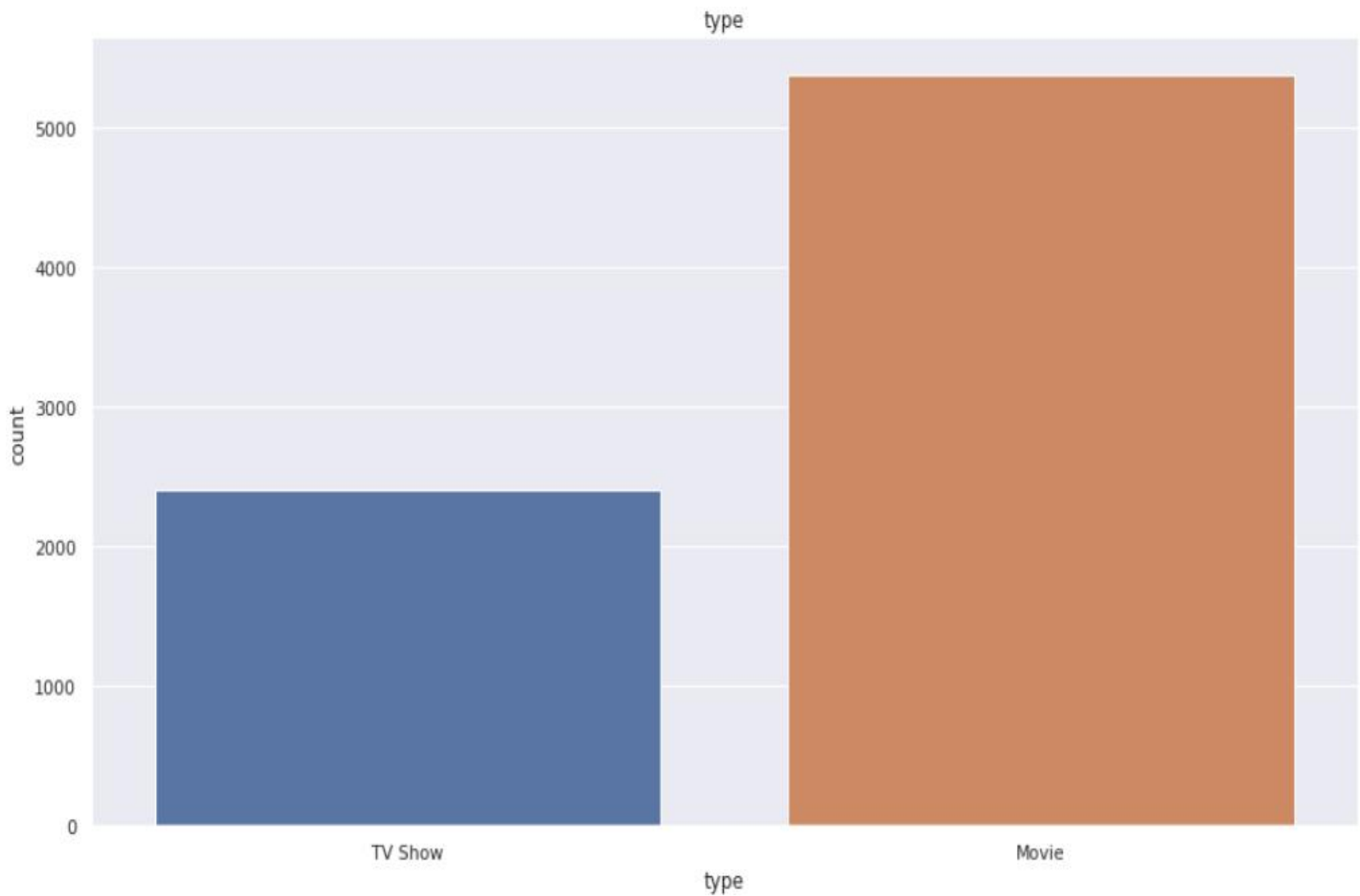
#	Column	Non-Null Count	Dtype
0	show_id	7787 non-null	object
1	type	7787 non-null	object
2	title	7787 non-null	object
3	director	5398 non-null	object
4	cast	7069 non-null	object
5	country	7280 non-null	object
6	date_added	7777 non-null	object
7	release_year	7787 non-null	int64
8	rating	7780 non-null	object
9	duration	7787 non-null	object
10	listed_in	7787 non-null	object
11	description	7787 non-null	object

```
dtypes: int64(1), object(11)
```

```
memory usage: 730.2+ KB
```

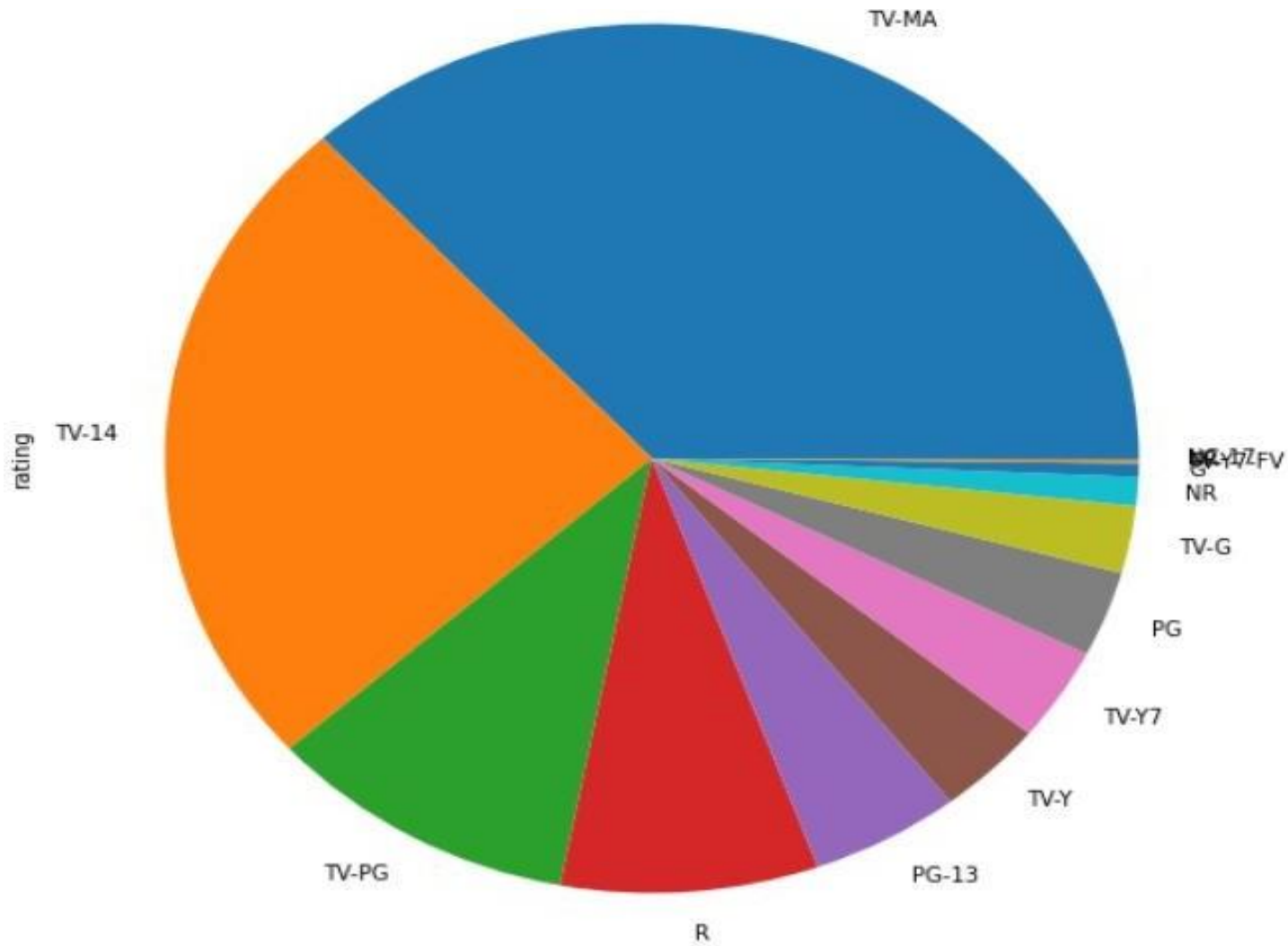
Exploratory Data Analysis

Distribution of Content



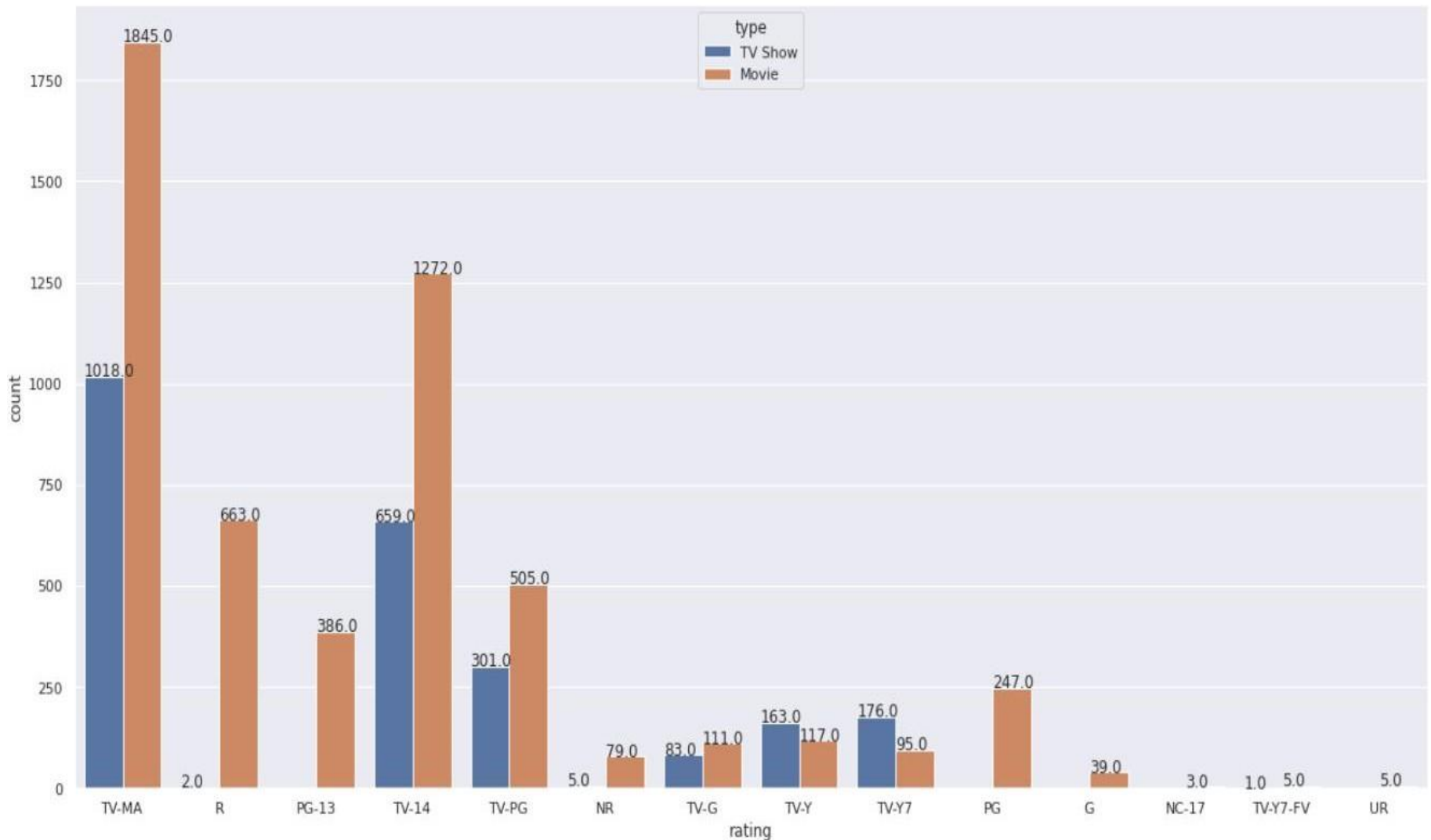
The content on Netflix is distributed between Movies and TVShows, Clearly it favours Movies as TVShows came into popularity later in the decade.

Distribution of Ratings on Netflix



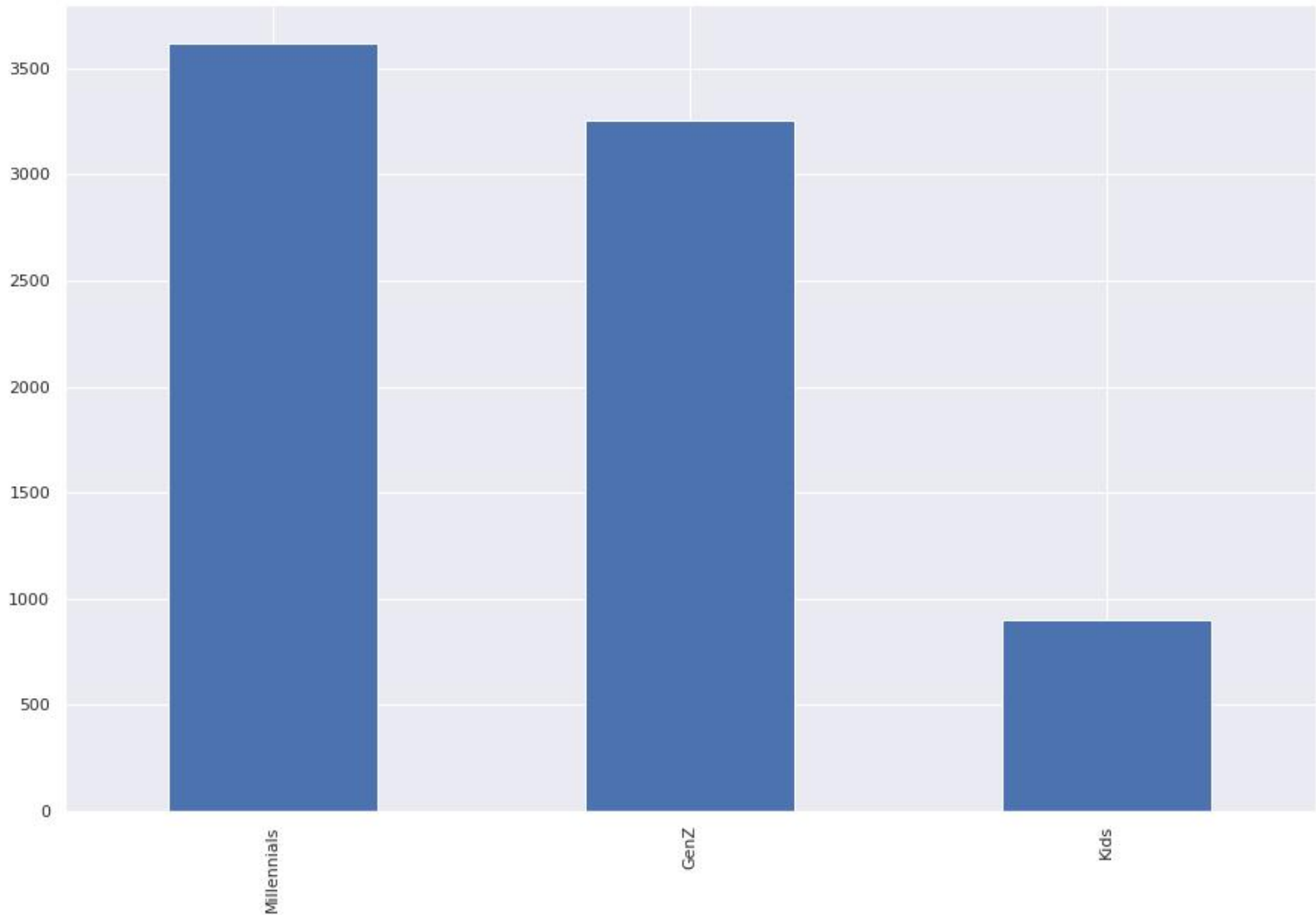
The majority of the Netflix audience are adults and teenagers so it is evident that the ratings like TV-MA and TV-14 are more in number.

Distribution of Ratings based on type of content



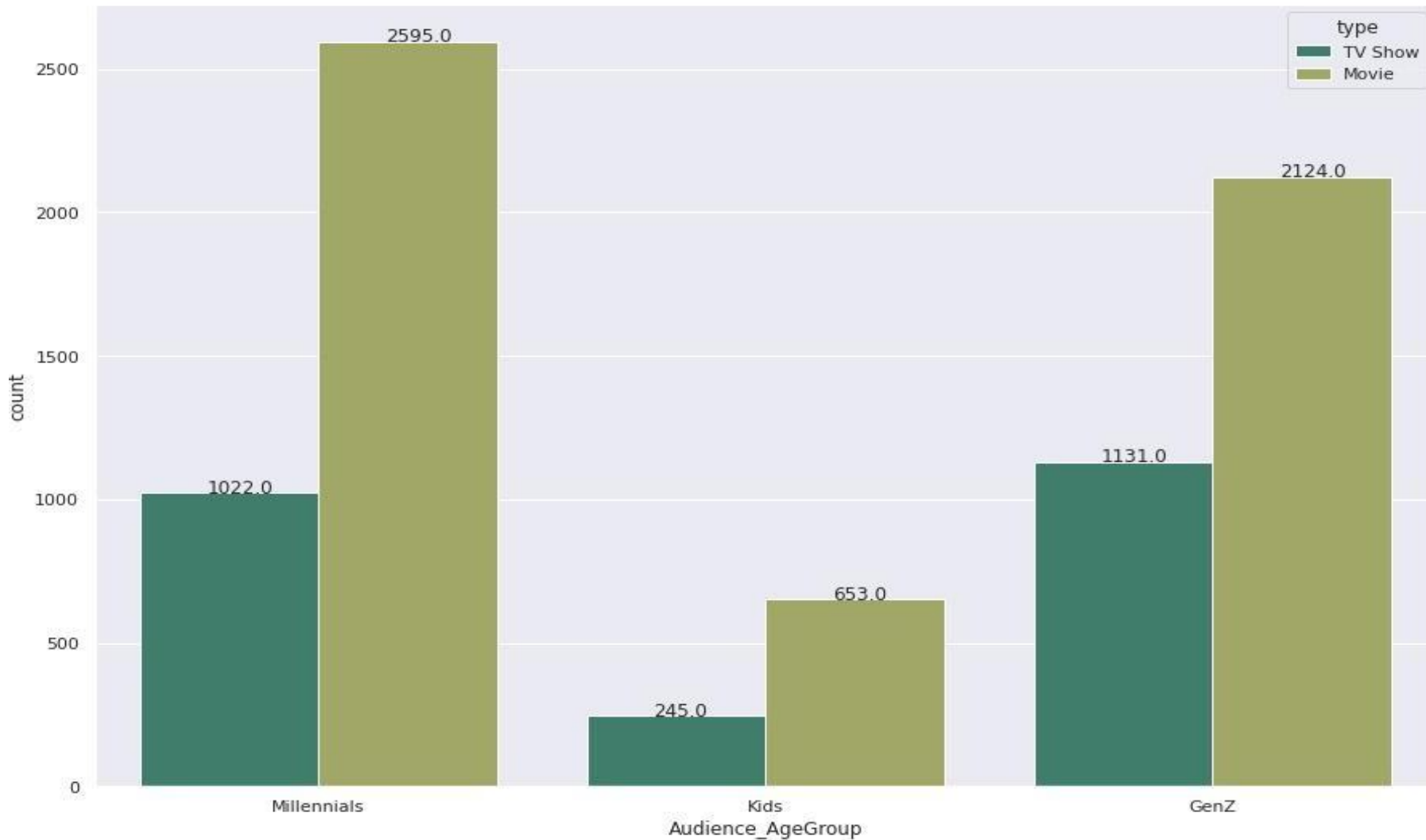
- As we can see that the distribution is quite uneven, One reason might be that there are less amount of TVShows than Movies on Netflix.
- There are greater number of TV-MA rated content on Netflix than any other and its lowest content count is for UR rating.

Distribution of Agegroup on Netflix



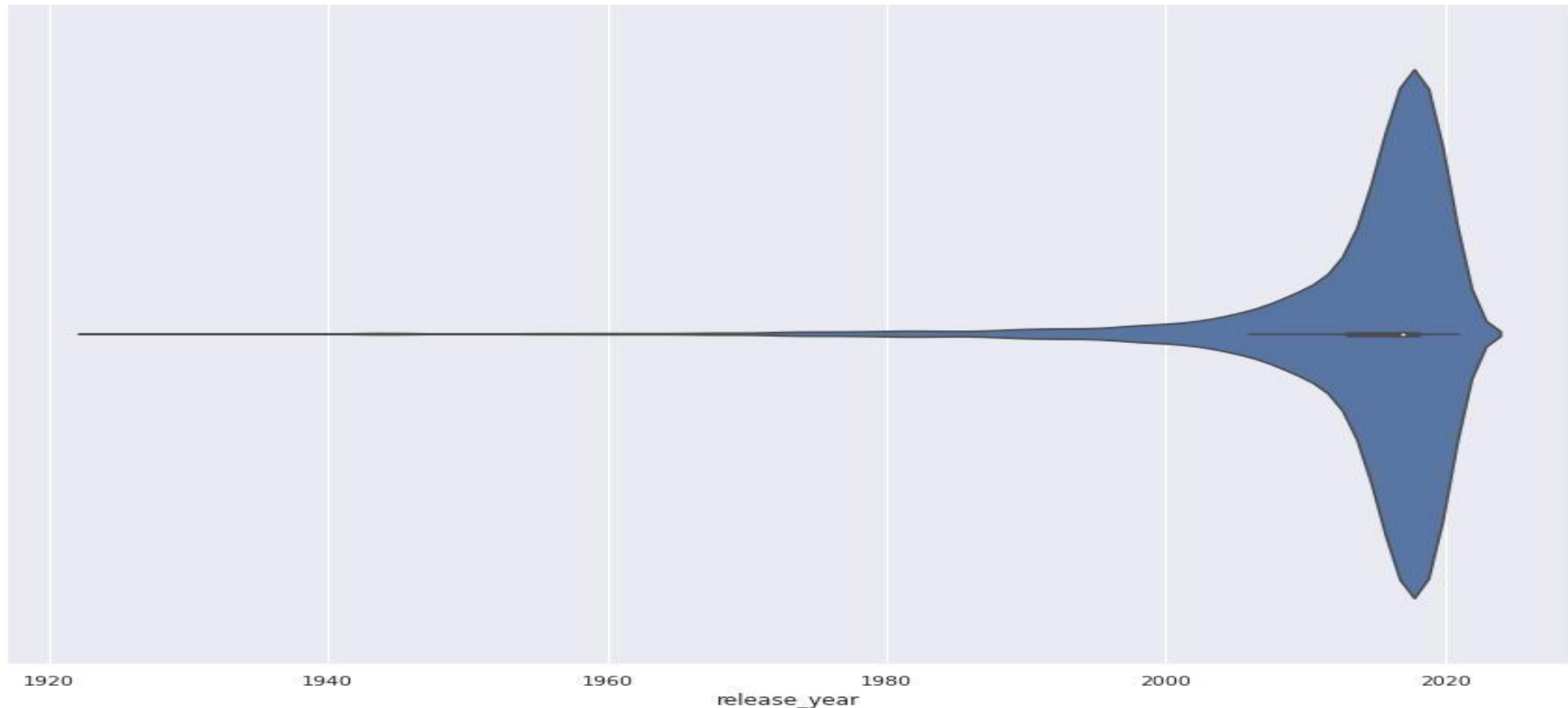
From the above plot, there is great disparity of especially regarding distribution of agegroups Millennials and Kids who consume content on Netflix which is understandable as the type of content on Netflix is more appealing for Millennials.

Type of content consumed by different Agegroups



Netflix has highest content count for individuals of age group Millennials and lowest content for Kids. The database contains wide-reaching number of movies for Millennials and fairly equivalent number of TV Shows for both Millennial and GenZ age groups.

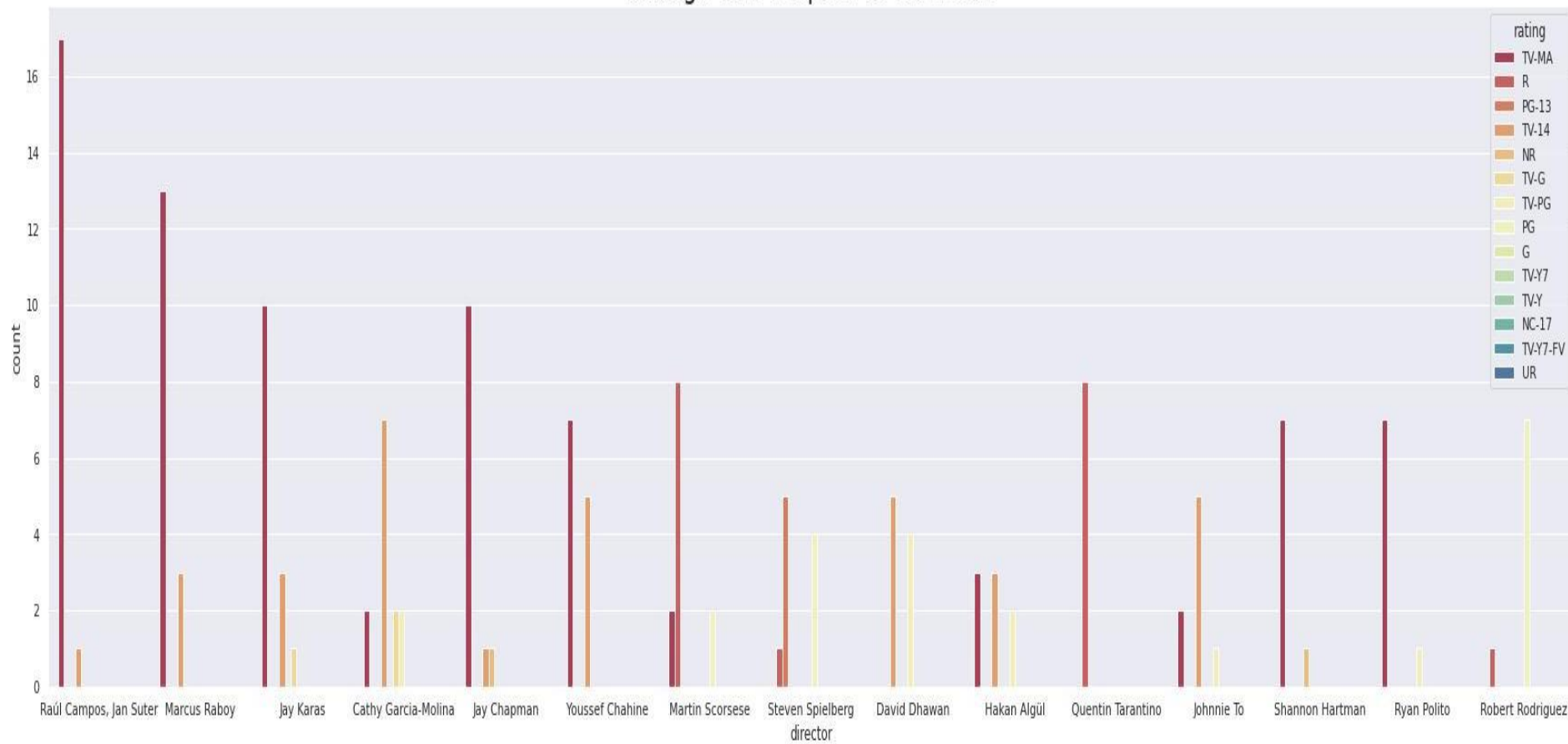
Top Netflix video content producing Countries



1. Highest Number of movies and TV shows were produced in the years 2015-2019.
2. Because of the digitization of the world, people are just one click away from streaming content so majority of the audience in these periods were more interested in Netflix so producers preferred to release their content on netflix.
3. People started taking interest in OTT platforms from 2017 and as Netflix is the largest OTT services provider so theres a peak from that period.

Top 15 Directors on Netflix

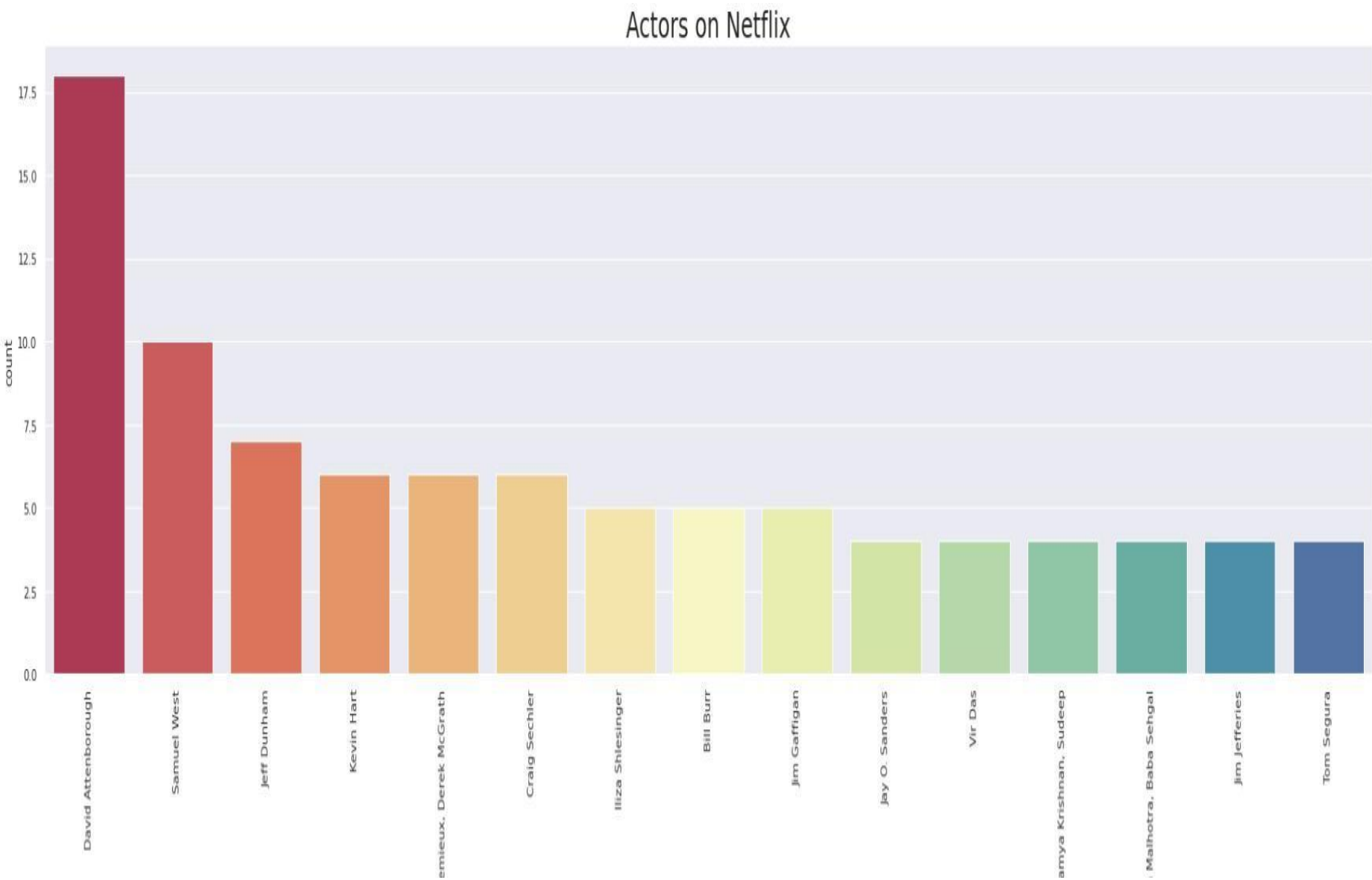
Ratings with respect to Directors



Top 5 directors are :

- Raul Campos
- Jan Suter
- Marcus Raboy
- Jay Karas
- Cathy-Garcia Molina

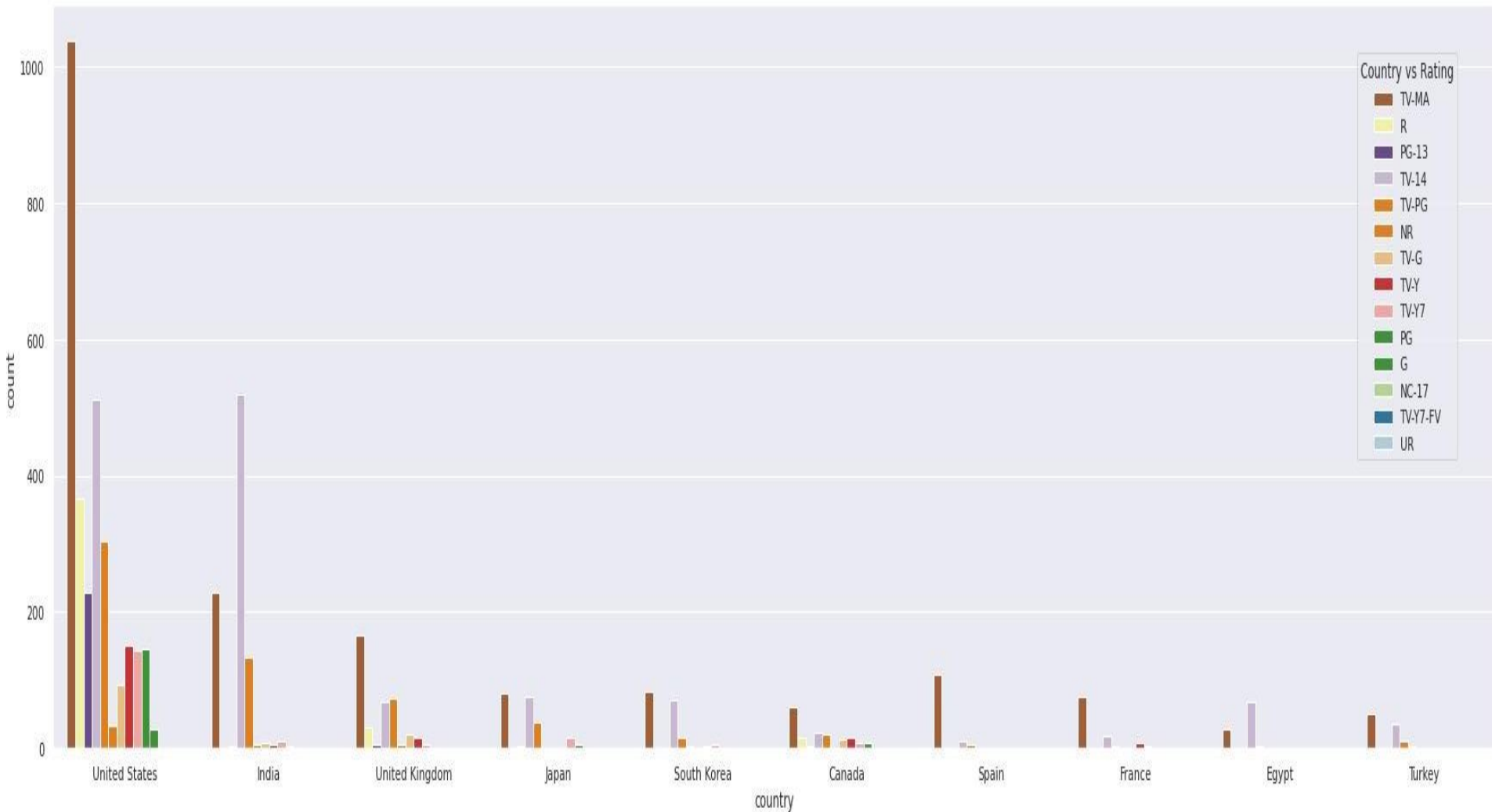
Top 15 Actors on Netflix



Top 5 Actors on Netflix :

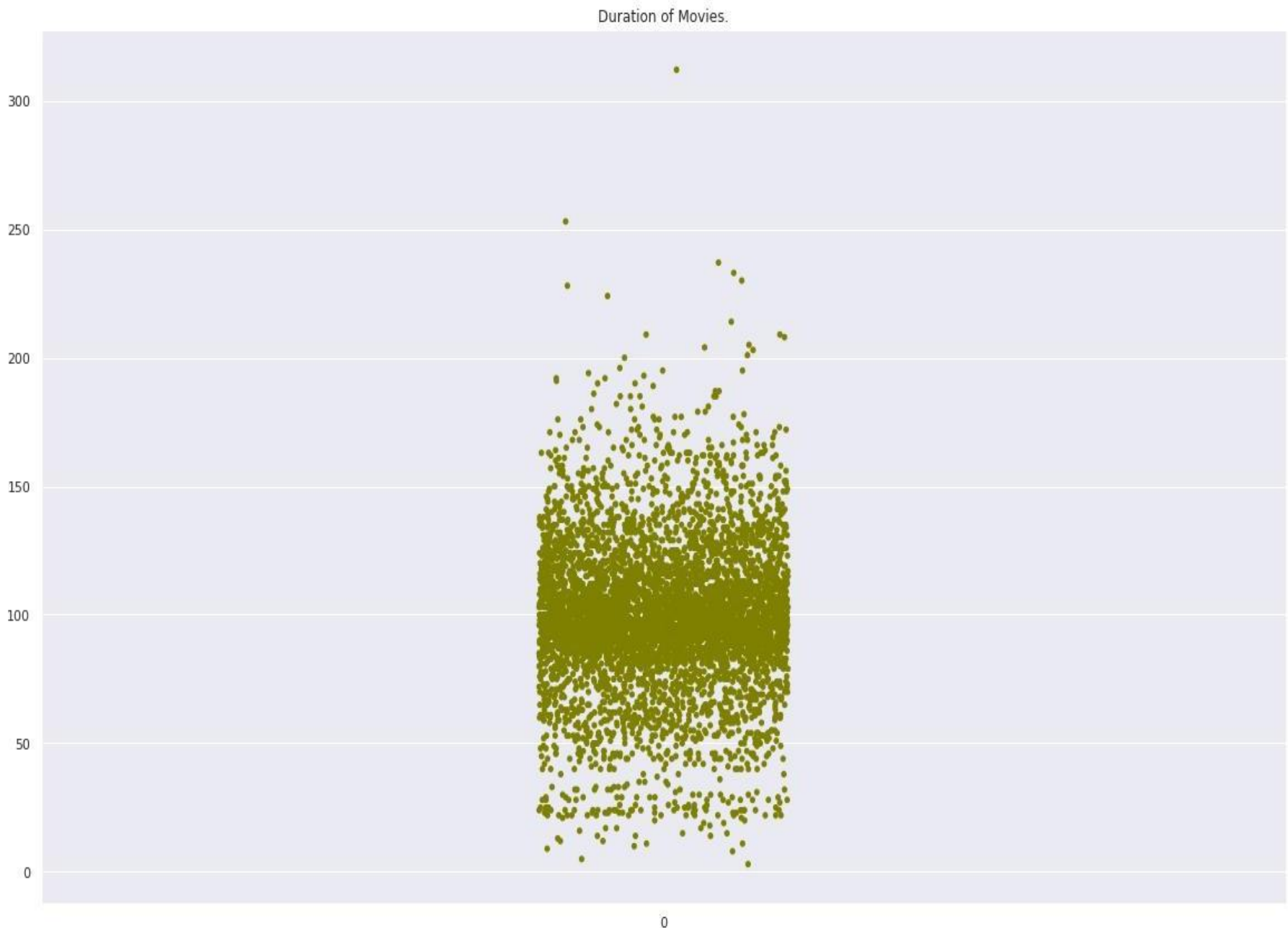
- David Attenborough
- Samuel West
- Jeff Dunham
- Kevin Hart
- Craig Sechler

Video Content in Top Countries



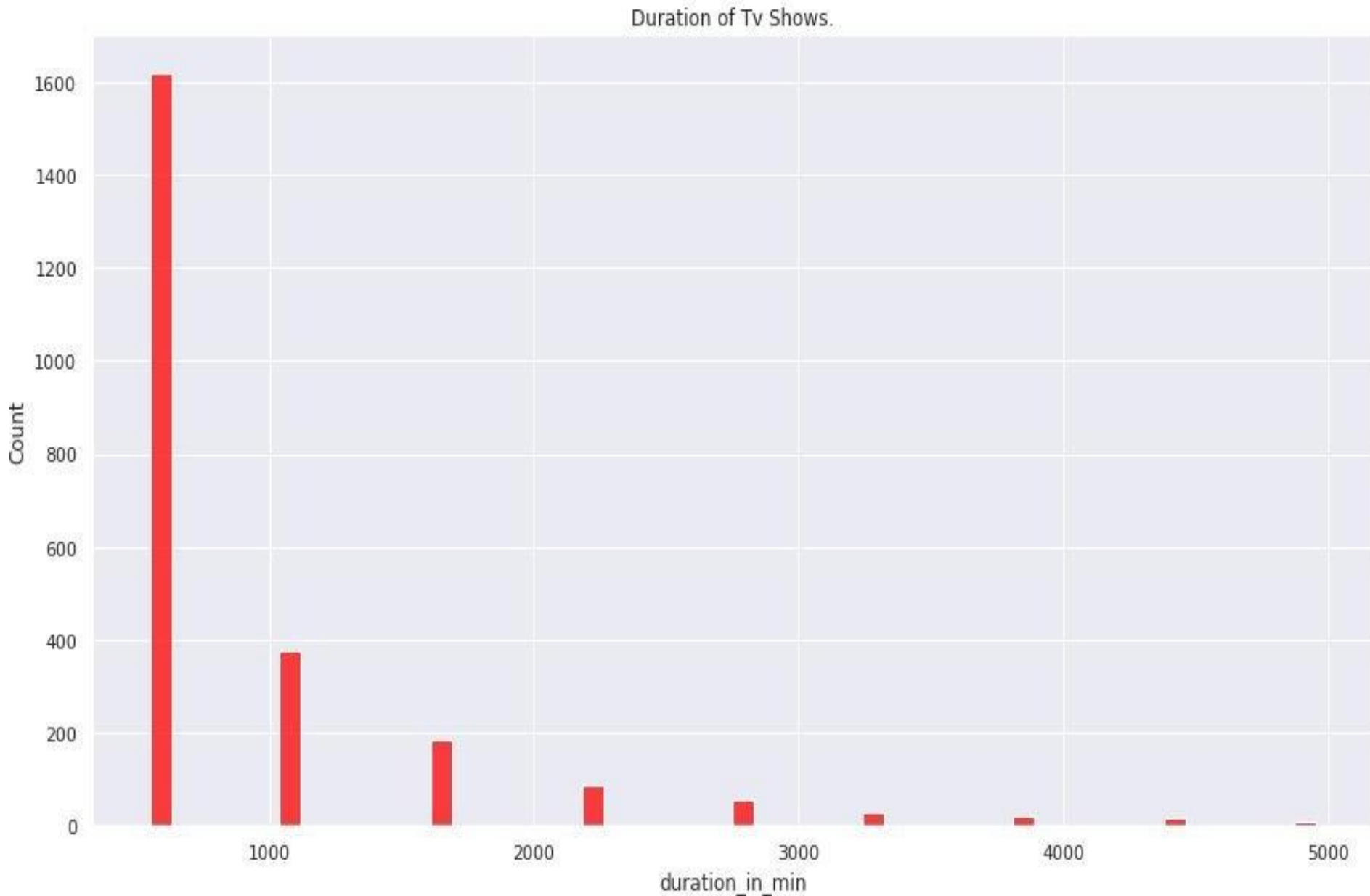
Here we can see the distribution of content is highest in US reason being there are a myriad of cultures that's why US produces variety of content on Netflix. As the total content count of a country decreases, the range of content ceases to exist.

Duration of Movies



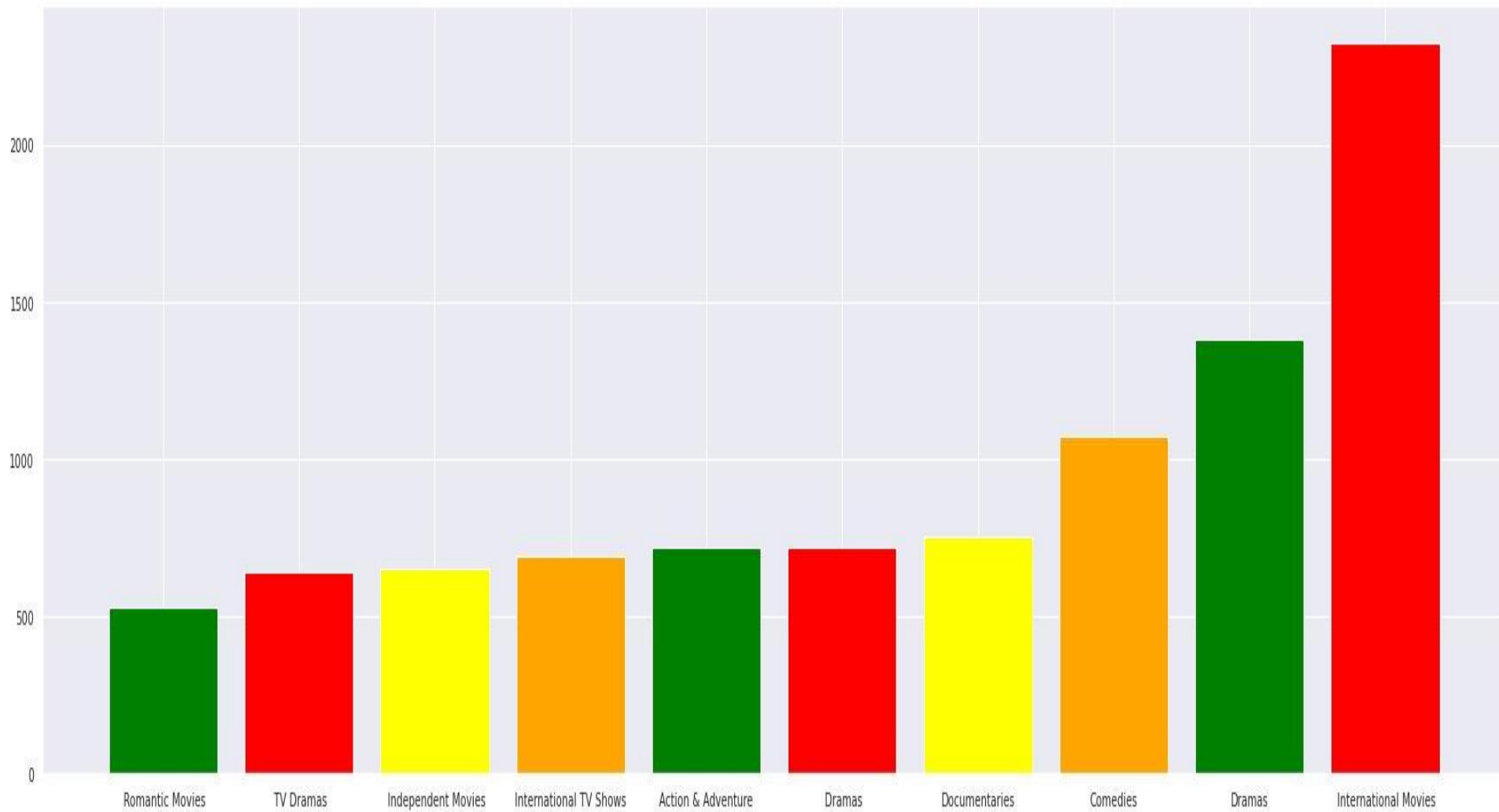
As we can see most of movies are around 100 to 120 minute mark which accounts to around 2hrs. Directors prefer to tell the story in 2hrs because it is the ideal period of time to get into crucial details and not dive too much into intricacies. Most of the people find it tiring if its a long movie.

Duration Of TVShows



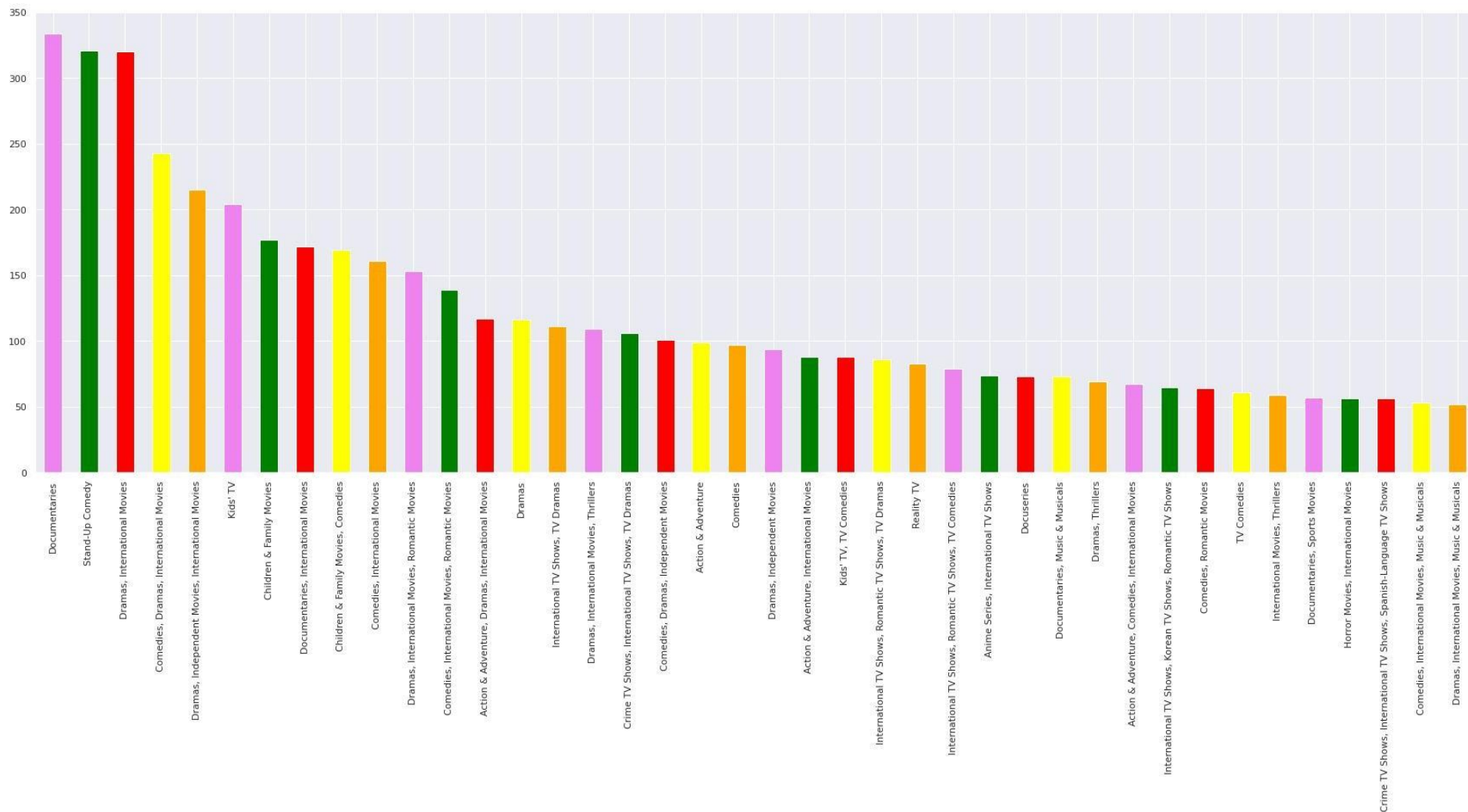
Most of the TV Show content duration is around 500 followed by 1000 minutes. 550 minutes comprise of single season and two seasons in 1000 minutes. Most of Shows may have been stopped producing halfway due to lack of financial resources and viewers.

Individual Genre Count for Netflix



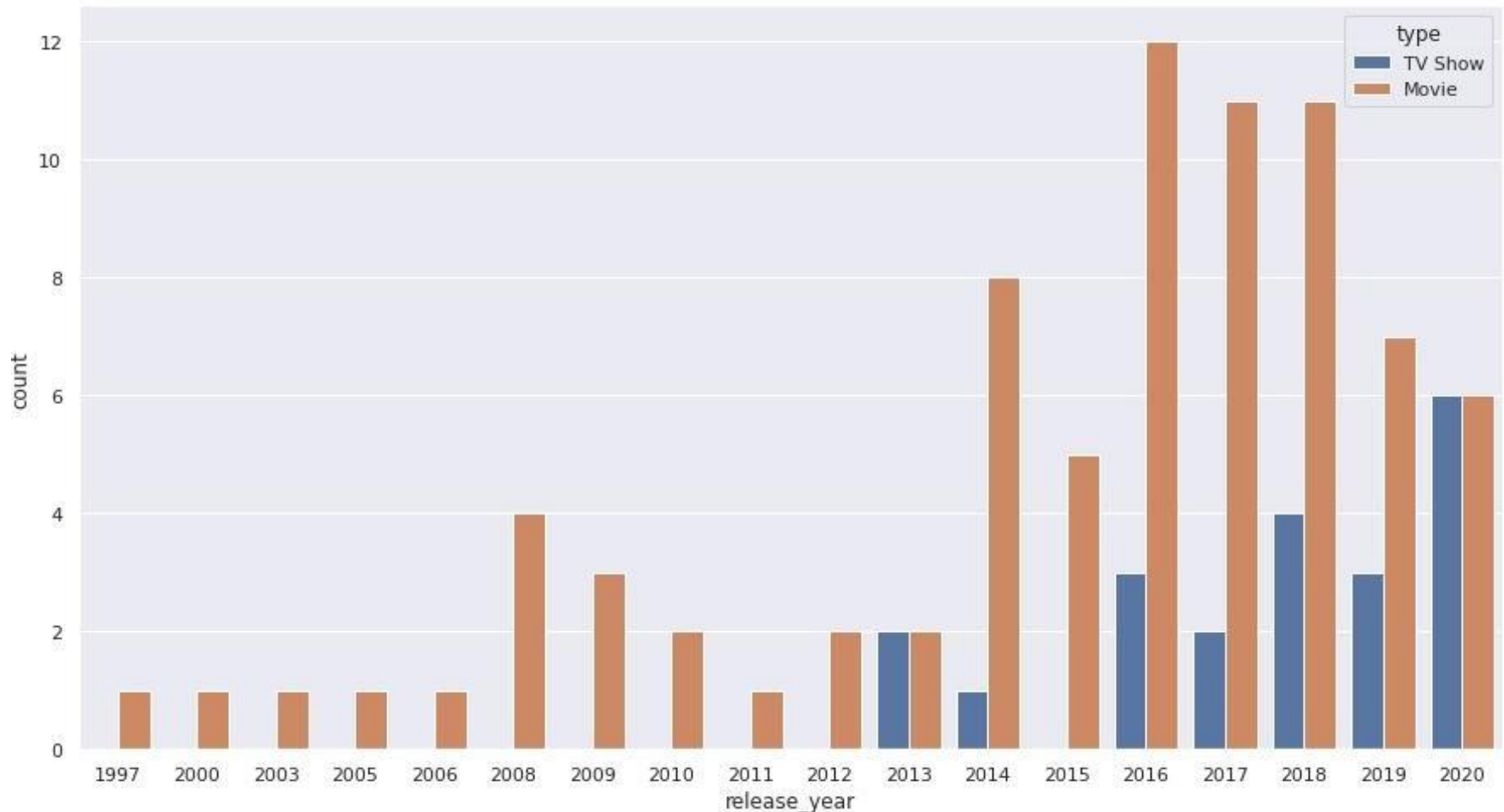
International Movies is the most popular genre followed by Dramas on Netflix. The reason might be that the content was filmed outside of U.S or was filmed by third-party production studios.

Content Count of Netflix which had multiple Genres



Top 40 genres of content across Netflix. Media count for Documentaries, Stand-Up Comedy, Dramas and International Movies was the highest. People are more inclined towards documentaries because it has knowledgeable content and its a fun way to learn and on the other hand Stand-Up Comedy is a form of stress buster for the general audience

Content of Movies and TVShows by Release Year



As we can see there weren't any TvShows until 2013. Then we can see that there is a steady increase in the TvShow content on Netflix. Now the TVShows are neck to neck with Movies which shows Netflix has been increasingly focusing on TV rather than movies in recent years

Preprocessing Algorithms

- Count Vectorizer
- TD-IDF Vectorizer
- Snowball Stemmer for Stemming Text
- Dimensionality Reduction using Principal Component Analysis

Models Used

- K-means Clustering
- AgglomerativeClustering

Different Algorithms used for the value K(no. of clusters) in K-means

- Elbow Method
- Silhouette Score
- Dendrogram

Conclusion

1. Majority of the content on Netflix is movies
2. There weren't any TVShows until 2013 on Netflix
3. The rate of TVShow content has been greater than that of movies and now in the year 2020, there are equal number of movies and TVShows on Netflix
4. Most of the content on Netflix is from United States and
5. There are a wide range of movies with respect to ratings on Netflix but Highest number of ratings are TV-MA and TV-14 from Movie and TV Show.
6. Netflix has the highest content count for individuals of the age group Millennials and lowest content for Kids.
7. Highest Number of movies and TV shows were produced in the years 2015-2019
December was the month where the most amount of content was added on Netflix followed by October
8. $k=15$ was found to be an optimal value for the number of clusters using which we grouped our data into 10 distinct clusters.
9. Using the given data a simple recommender system was created using cosine_similarity and recommendations for Movies and Tv Shows were obtained.

Thank You