NETFLIX

NETFLIX MOVIES AND TV SHOWS CLUSTERING

NETFLIX, is world's largest on-demand internet streaming media and online DVD movie rental service provider. it was Founded on August 29, 1997, in Los Gatos, California by Marc and Reed. It has 69 million members in over 60 countries enjoying more than 100 million hours of TV shows and movies per day. Netflix is the world's leading internet entertainment service enjoying TV series, documentaries, and feature films across a wide variety of genres and languages.

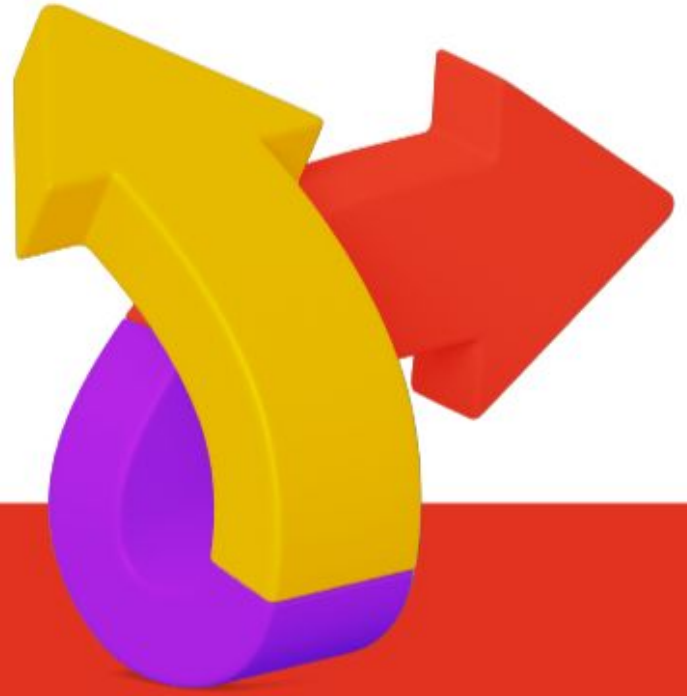
This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting

OBJECTIVES

Objectives:



01 Exploratory Data Analysis

02 Understanding what type content is available in different countries

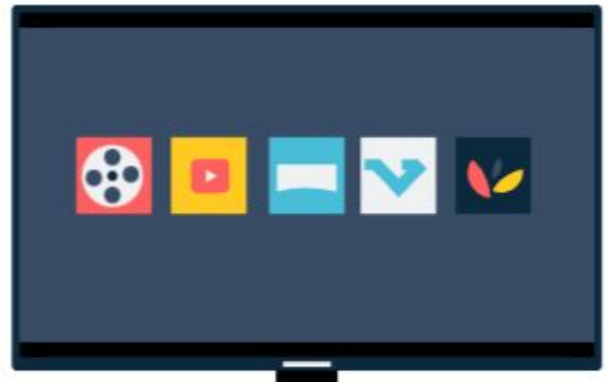
03 Is Netflix has increasingly focusing on TV rather than movies in recent years.

04 Clustering 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: Genere
description: The Summary description

Our dataset is composed of **7787 rows** and **12 Feature Columns**

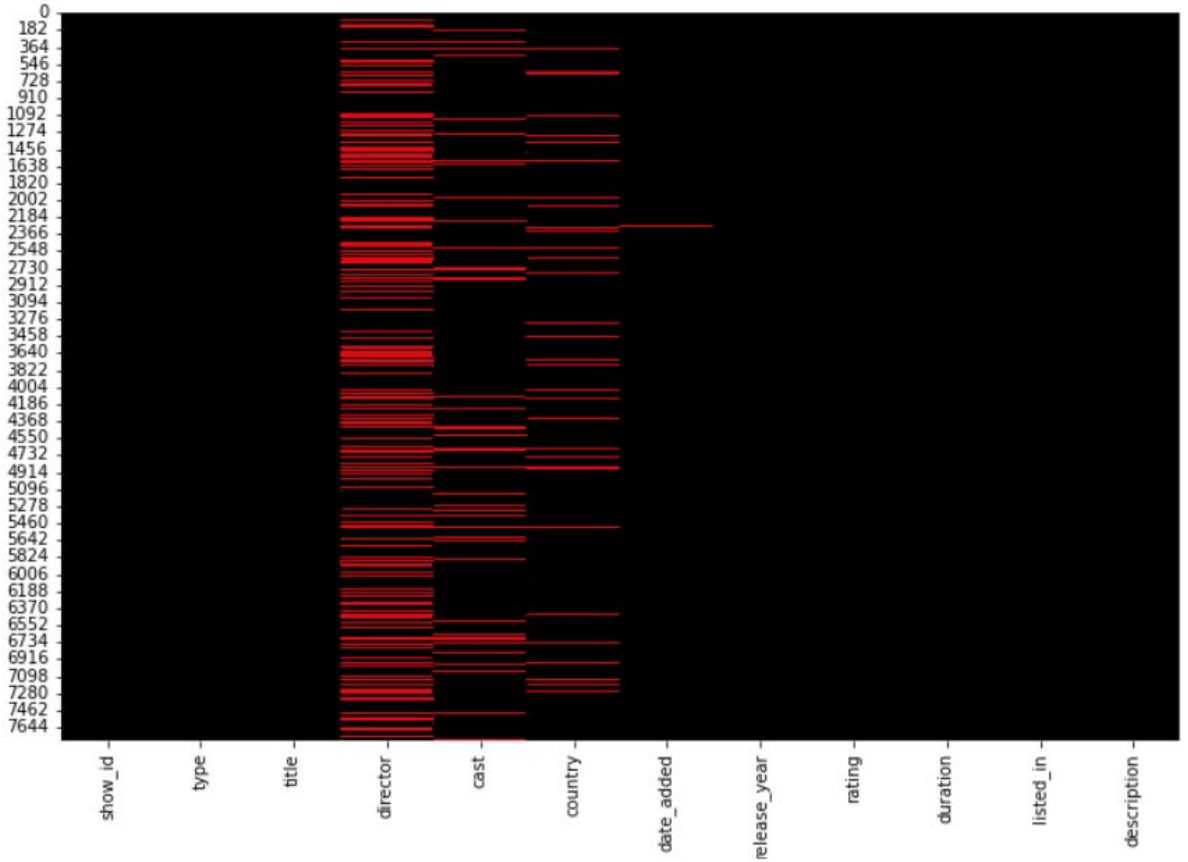


DATA CLEANING

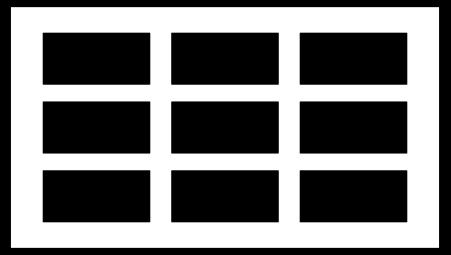
THERE ARE NULL VALUES IN OUR DATASET

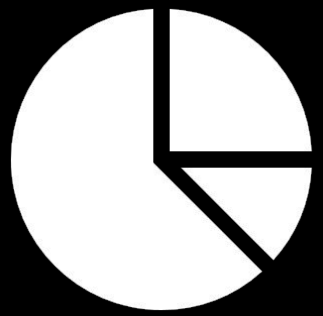
As we can see here,
Column: ***Director, Cast, Country and date_added contrains null values***
but as in date_added only 10 rows contain null value so we will remove only those rows

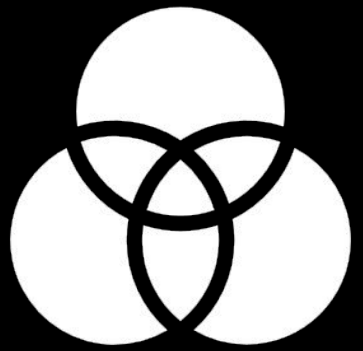
- Removing rows with null values in the date_added column
- Removing rows with null values in the rating column
- Introducing a new column - "year_added" column to get the year in which movie was added to netflix

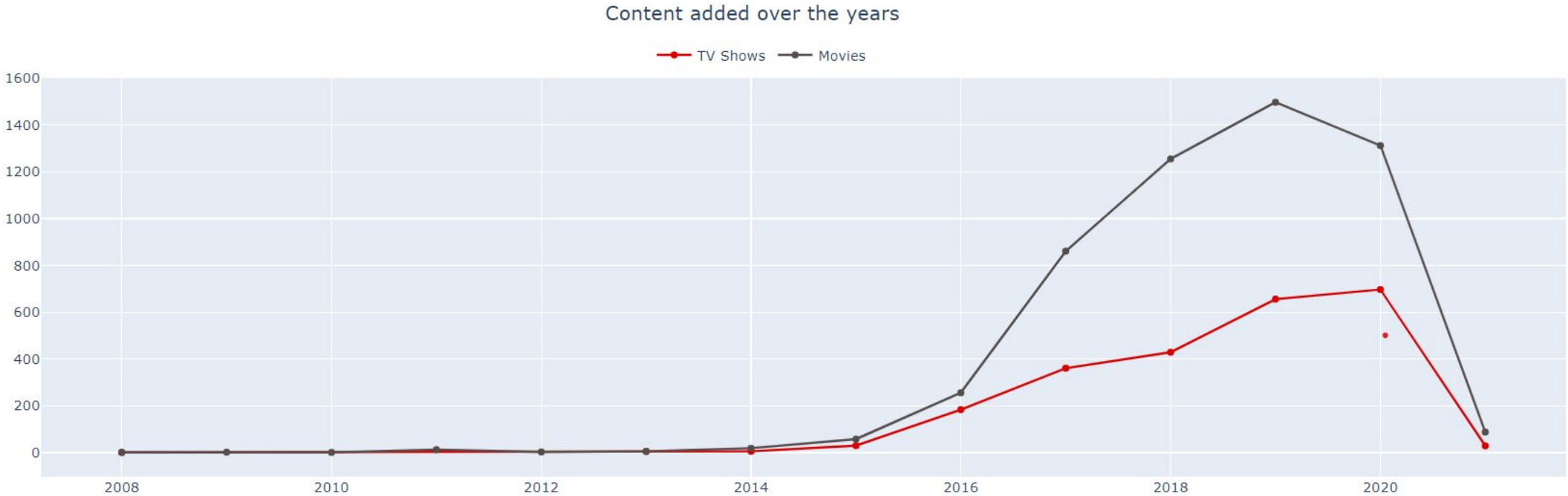


EXPLORATORY DATA ANALYSIS



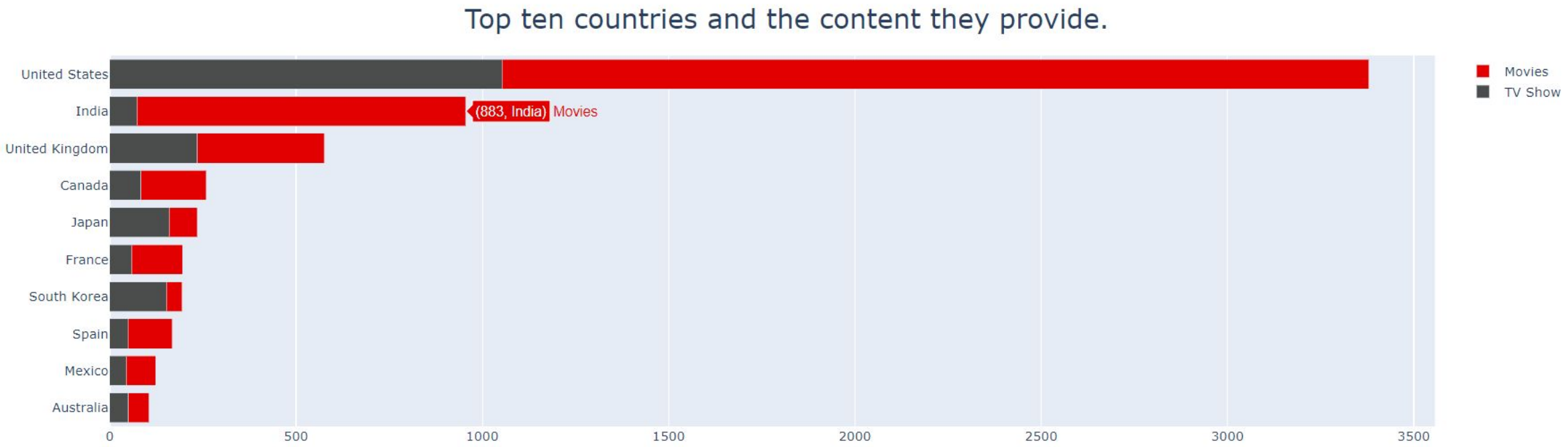




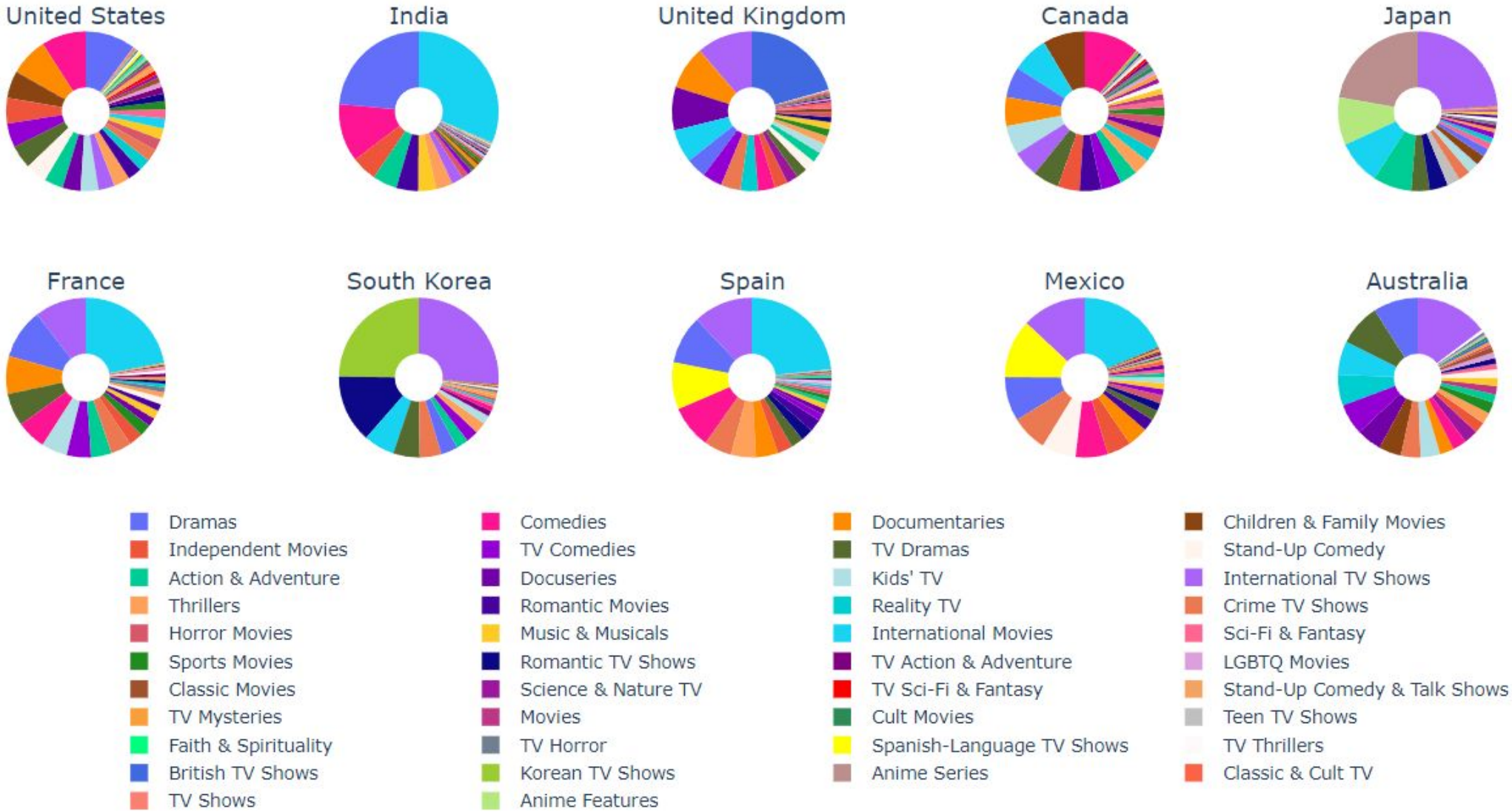


- Relative growth is seen in the number of movies on Netflix than TV shows
- Increase in number of movies and Tv shows is noticeable from year 2015
- In 2019 and 2020, highest number of movies and TV shows were added by Netflix on its platform.
- Very few movies and Tv shows were recorded in 2021, due to very less data recorded in the same year.

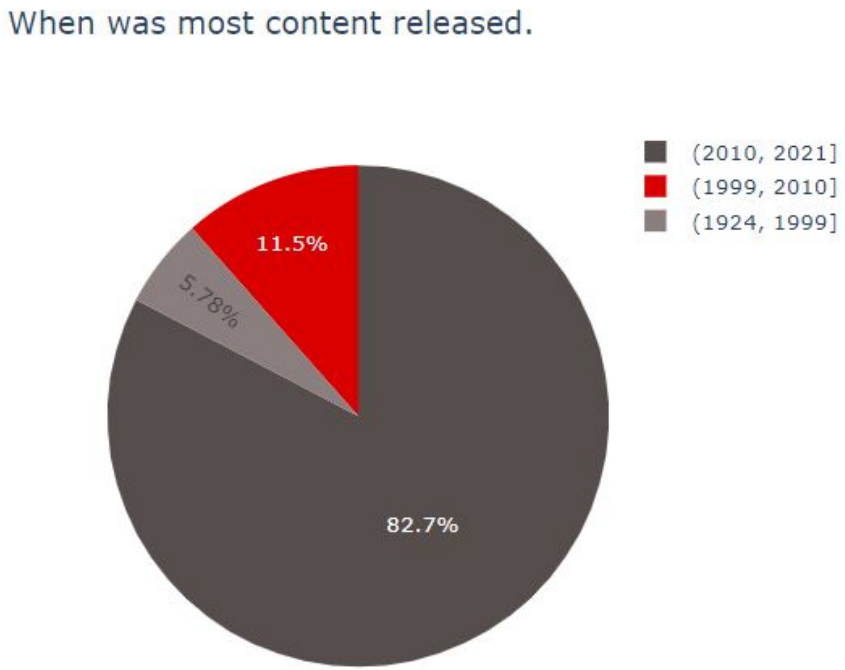
- United States is the leading producer of both movies and tv shows, since Netflix is a US company this makes sense.
- India stands seconds via the influence of Bollywood which explains the type of content available which mainly focuses on movies rather than tv shows.
- Japan and Korea are more frequent in Tv shows which explains the KDrama and Anime culture nowadays.



Top ten countries and the content they provide.



- Drama, International Movies, and Comedies seem popular choices in most countries.
- British and International Tv Shows dominate in the United Kingdom. Regional specialties such as Anime in Japan and Korean Tv shows in South Korea are more prominent in these countries; This makes sense as anime has always been popular in Japan, and the rising k-pop culture explains the increase in Korean Tv Shows.
- It's also observed that in the countries where the regional language is not English, International Tv Shows and Movies are more in demand.

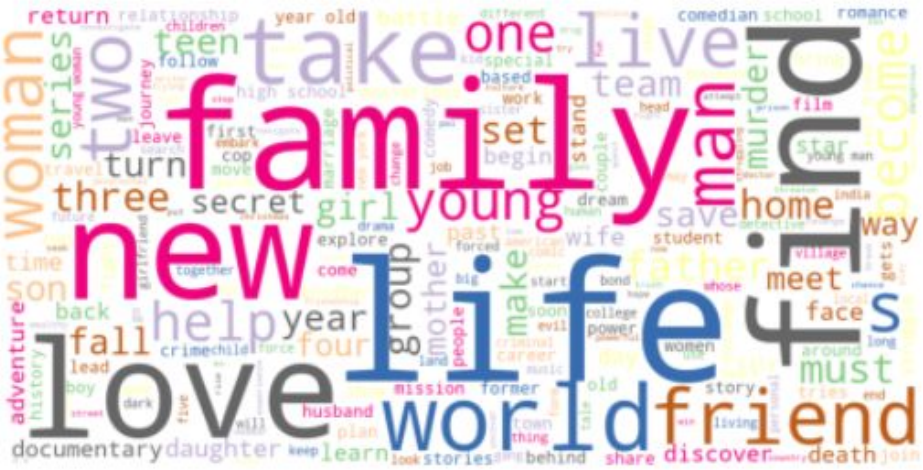


- 82.7% of the content was released between the year of 2010 - 2021
- Before 2010 only 17.28% content was released.

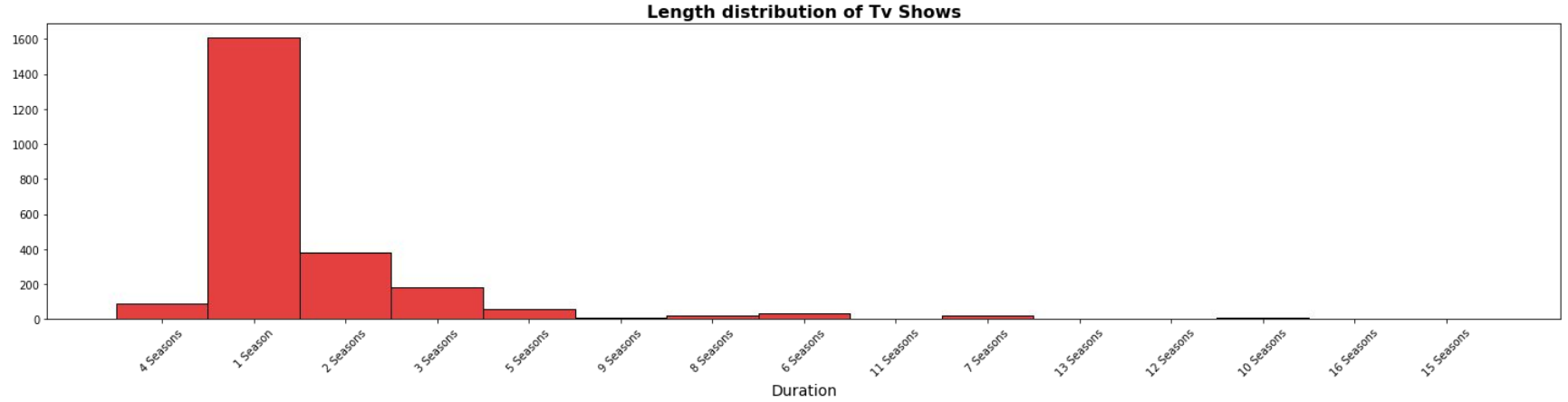
- Christmas, Love, World, Man, and Story are the most repeated words including in the title.
- Also we can say that christmas being appeared many times as a sign we saw earlier of most of the movies and tv shows got released during the winters.

Max number of words used in title:

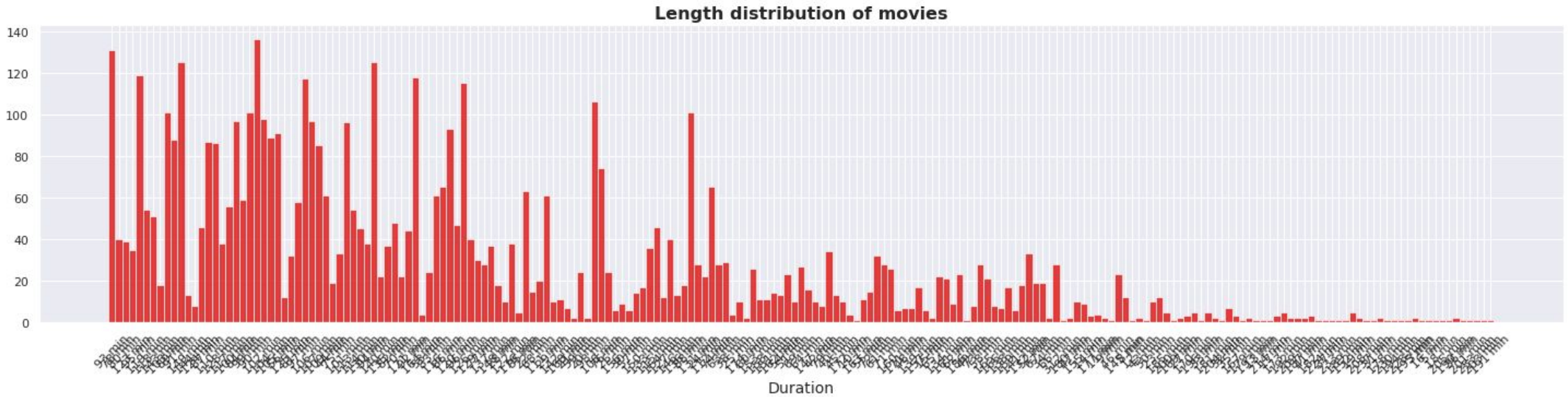
Max number of words used in description:



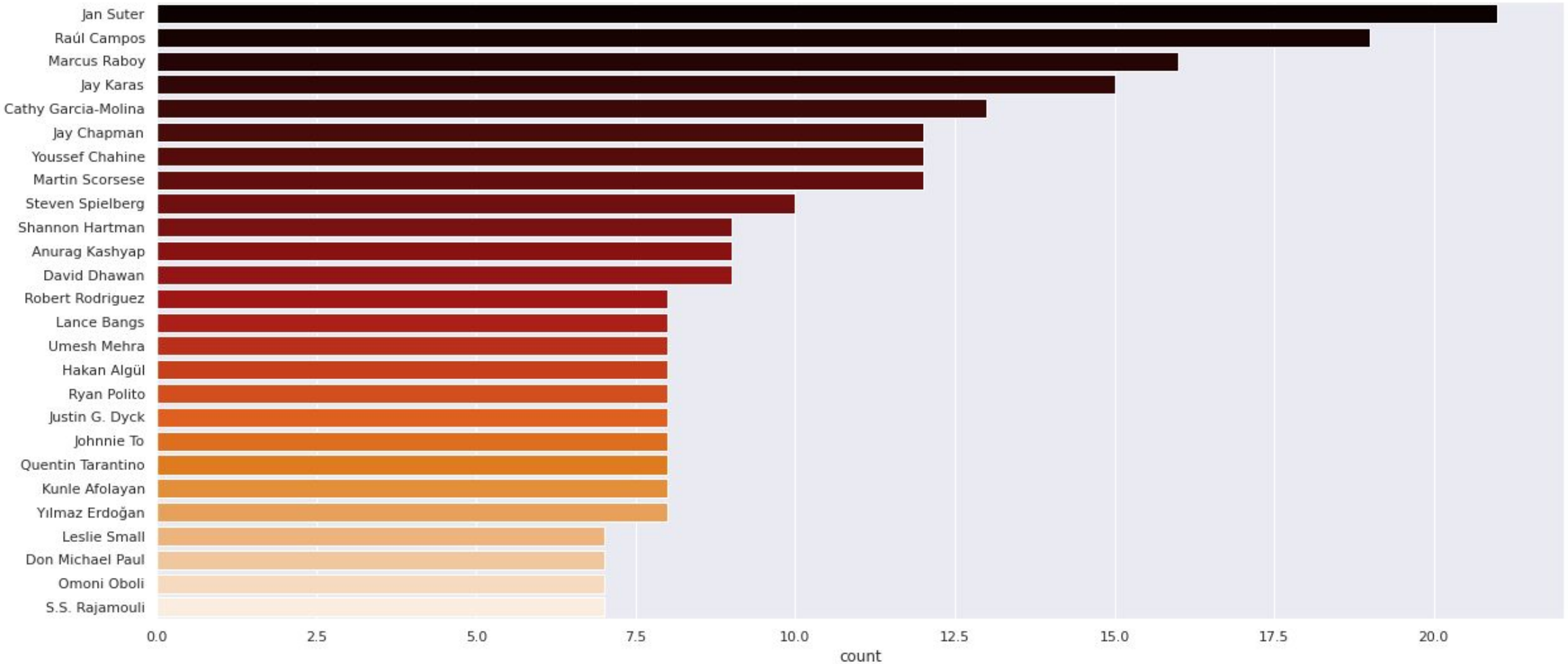
Family, Friend, Love, Life, Woman, Man are the most occurring words in the description of the tv shows and movies.



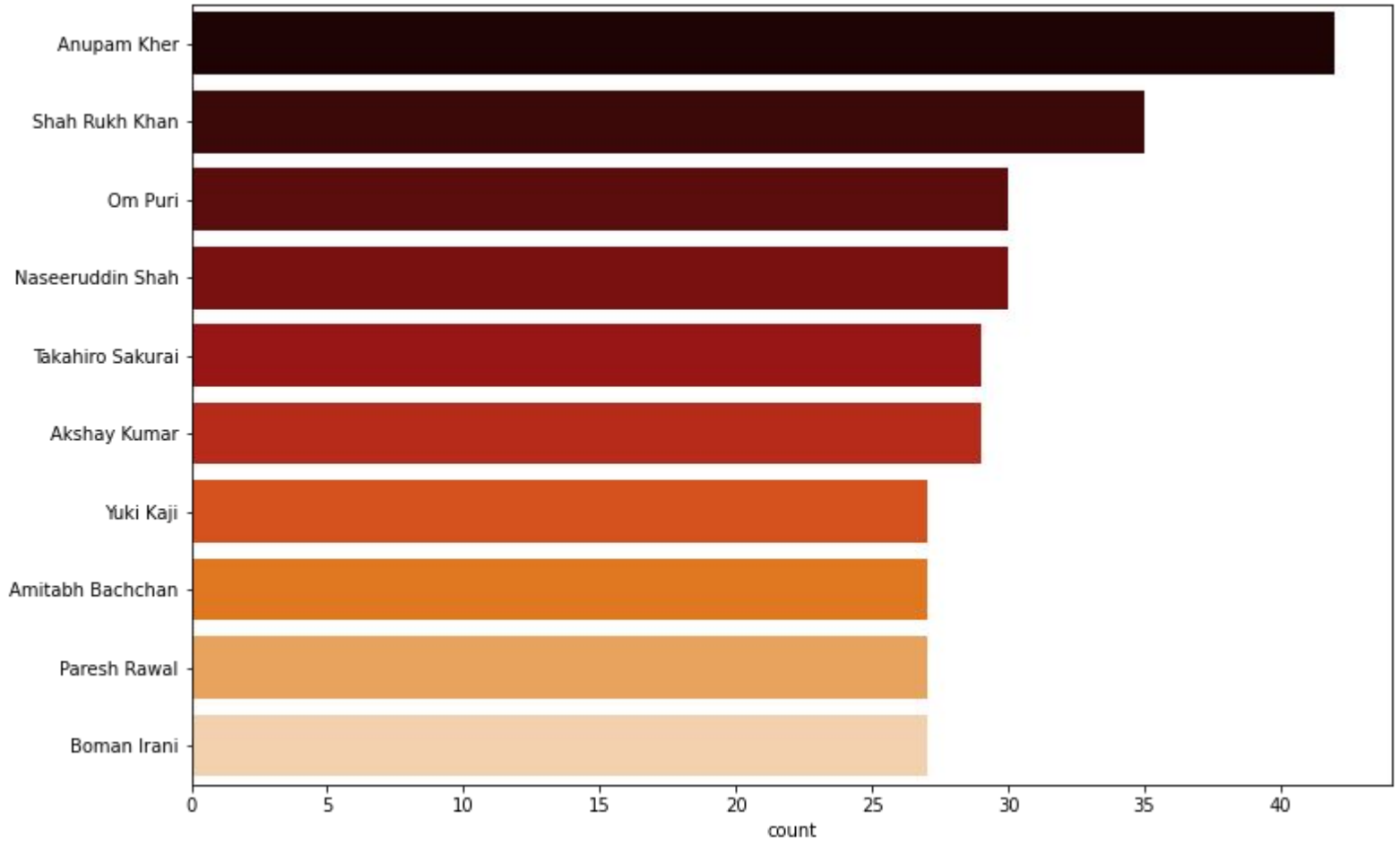
Most of the Tv Shows last for more than 1 or 2 seasons, very few tv shows are listed for more than 5 seasons.
Most of the movies last for 90 to 120 minutes



Top 25 Directors whose content is available on Netflix



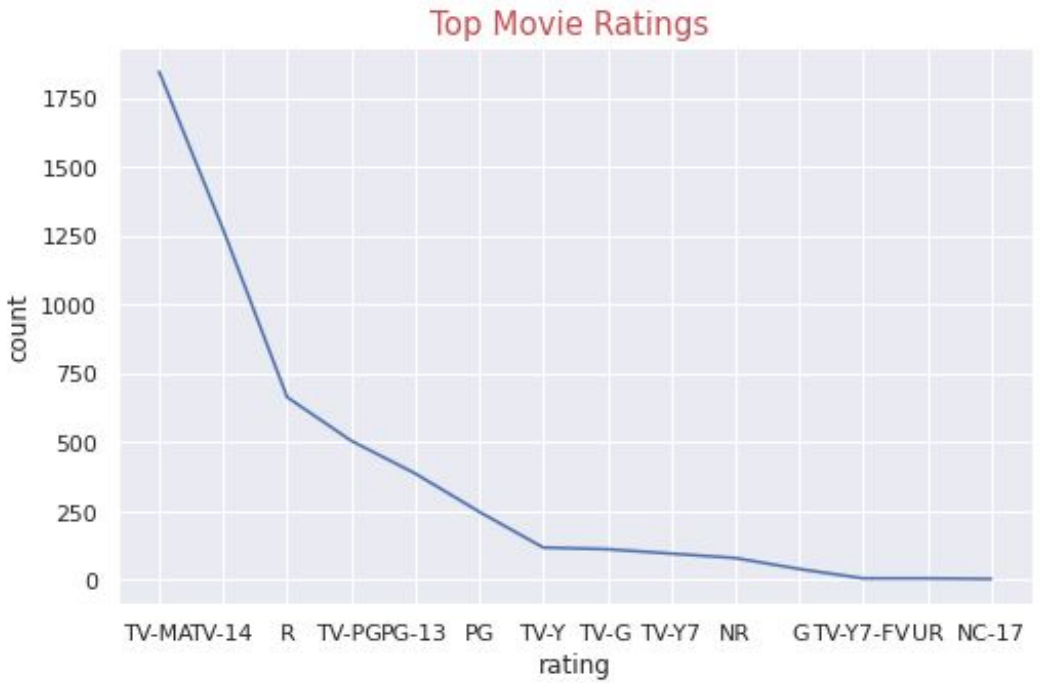
Top 10 actors whose content is available in Netflix



Yahhh!!! Delightful to see indian actors in the top 6.

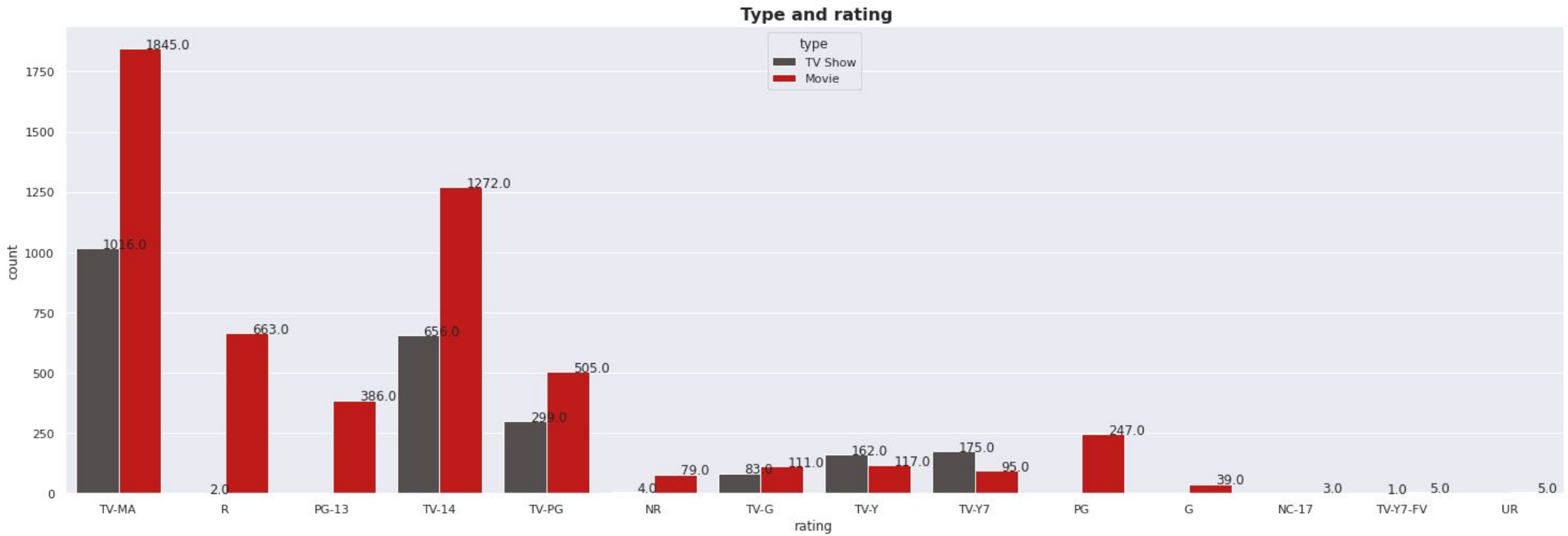
Top Tv Show ratings

Each TV show and movie on Netflix is assigned a maturity rating to help members make informed choices for themselves and their children.Netflix determines maturity ratings by the frequency and impact of mature content in a TV show or movie. TV show ratings reflect the overall maturity level of the whole series.



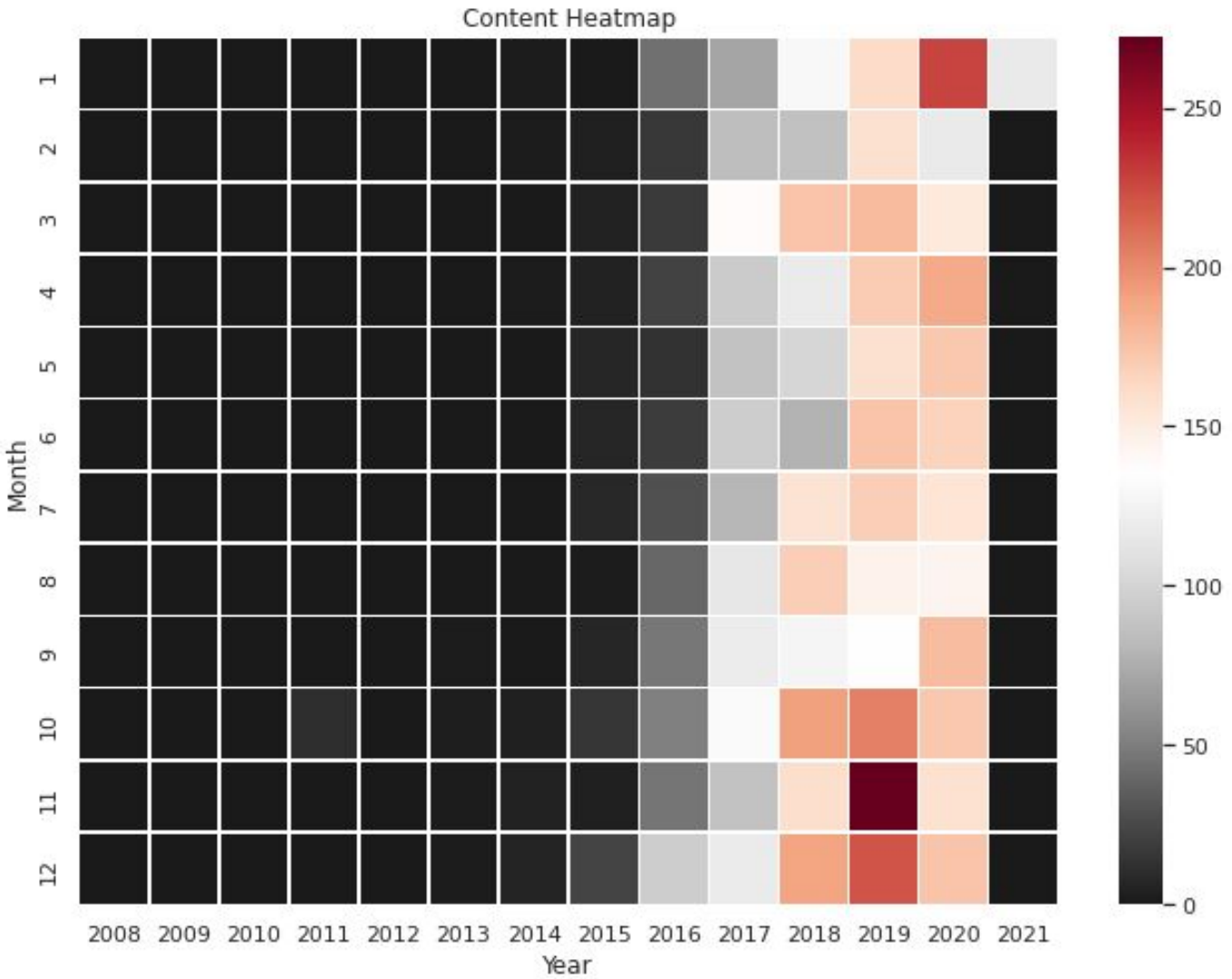
TV-MA tops the charts, indicating that mature content is more popular on Netflix.

- This popularity is followed by TV-14 and TV-PG, which are Shows focused on Teens and Older kids.
- Very few titles with a rating NC-17 exist. It can be understood since this type of content is purely for the audience above 17.



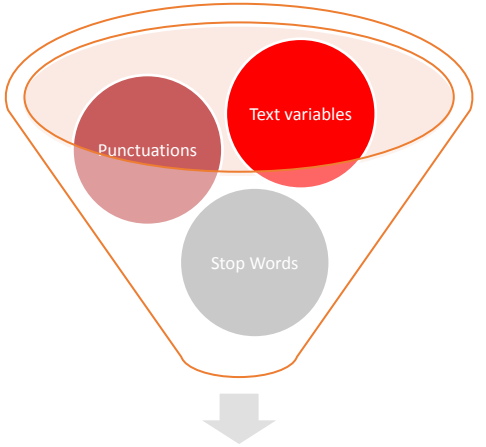
From around 2014 is when Netflix began to increase their content count. We can see over the years and months

- Netflix continues to slowly increase the amount of content that is being added into their platform.
- We can see in 2020, the data stops at January since that is the latest month available in the dataset.

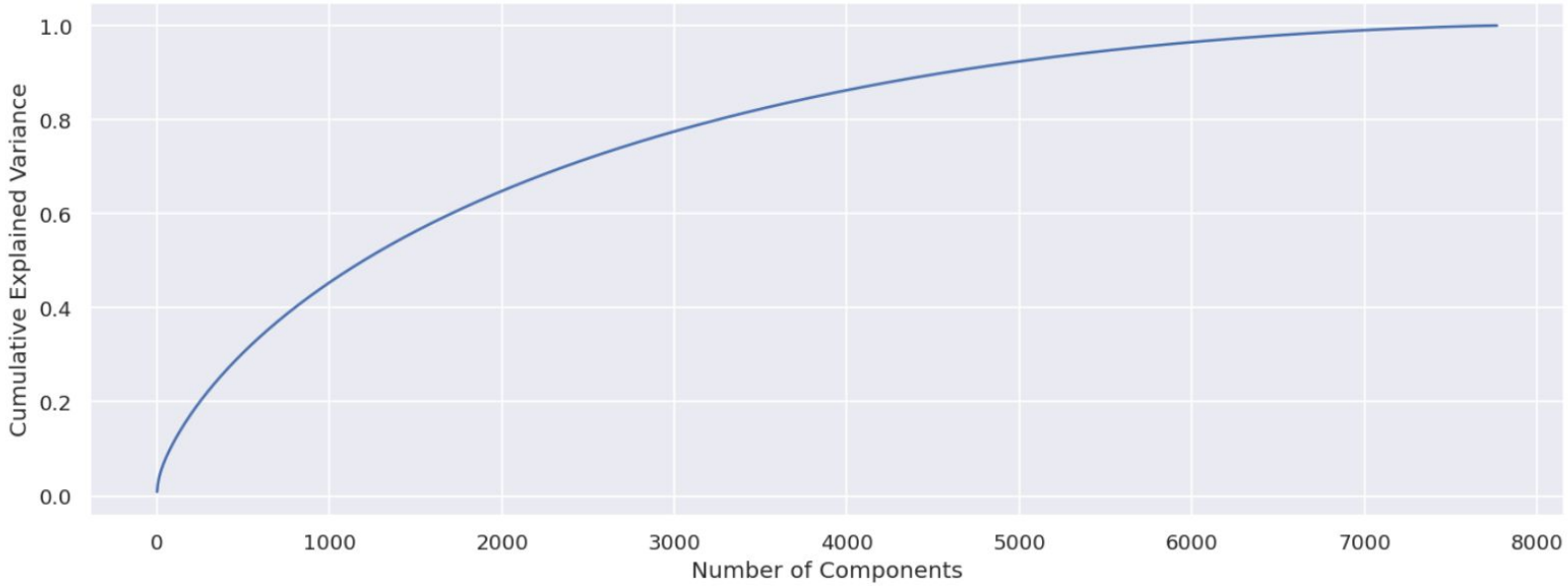


DATA PRE-PROCESSING

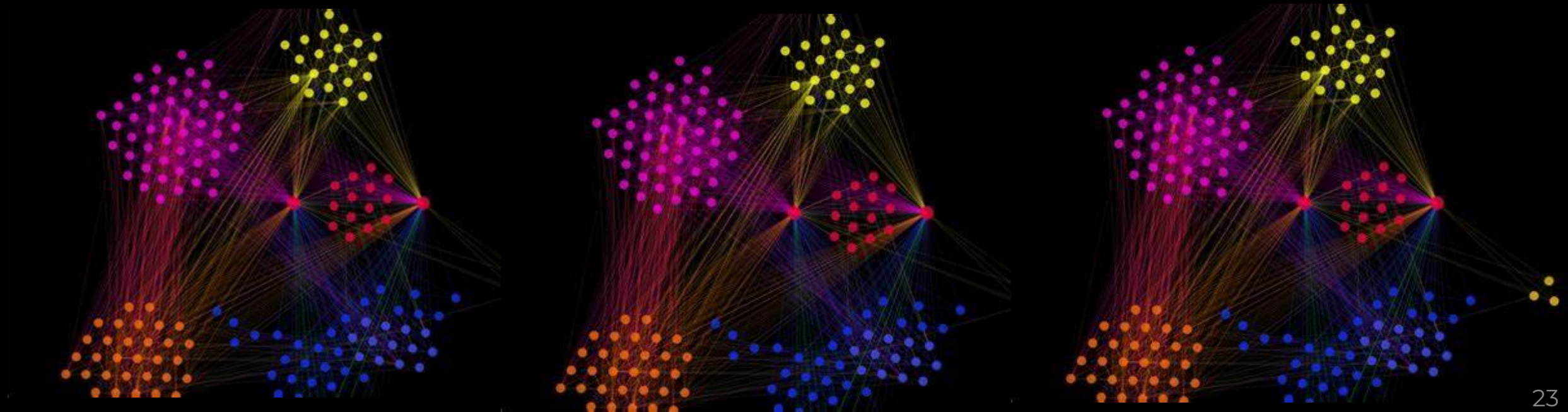
- Create clusters for our data now using text columns (Description, listed_in, Director, Cast, Country, Rating)
- Remove punctuations and stop words.
- Perform stemming to reduce words with similar meaning.
- Perform PCA to reduce dimensions, select number of components that explain 95% of the variance.



Stemming



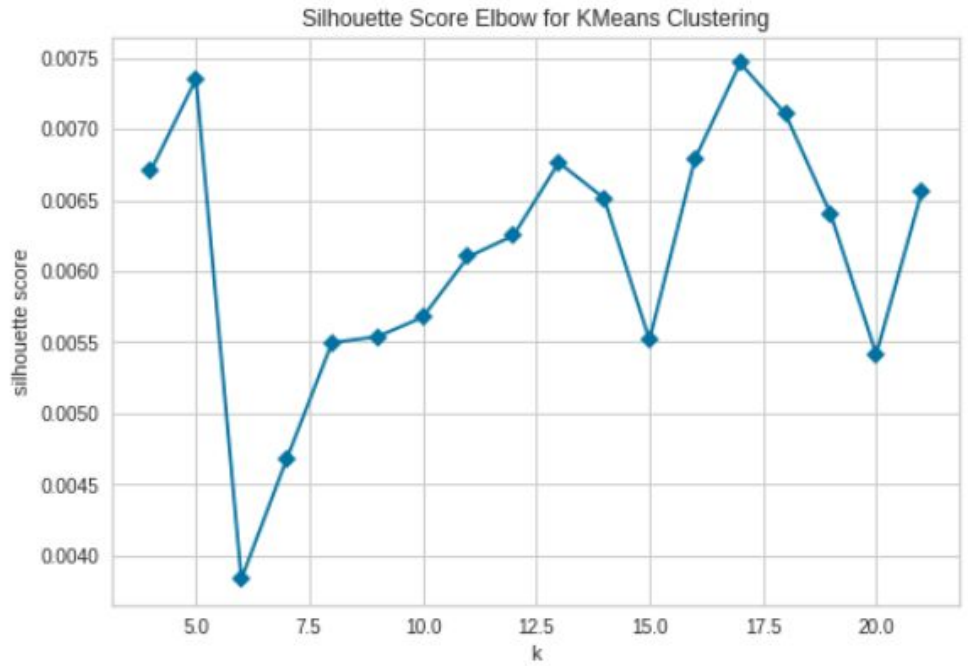
CLUSTERING

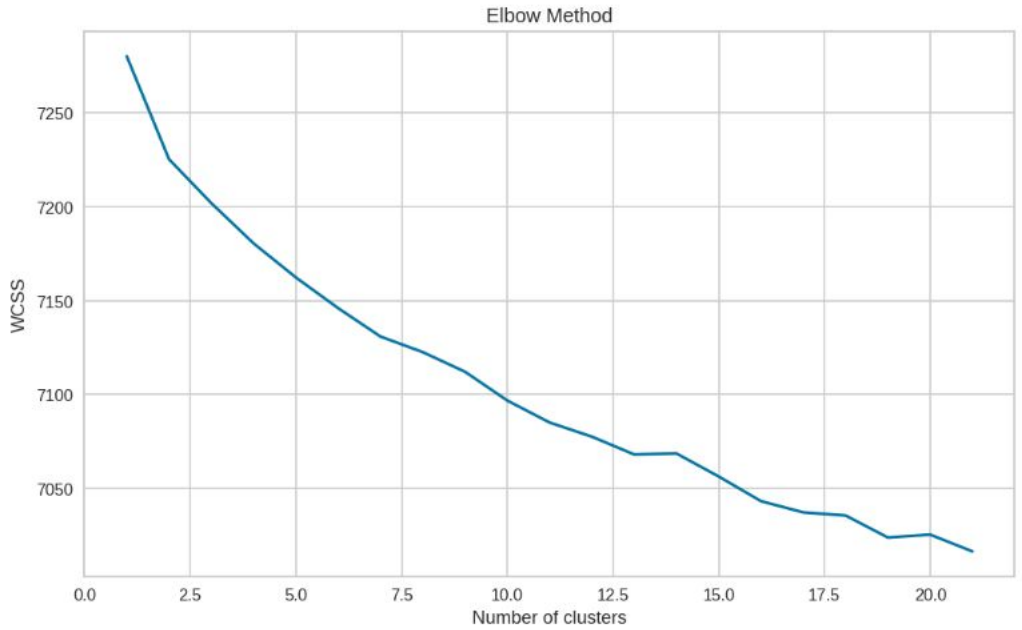


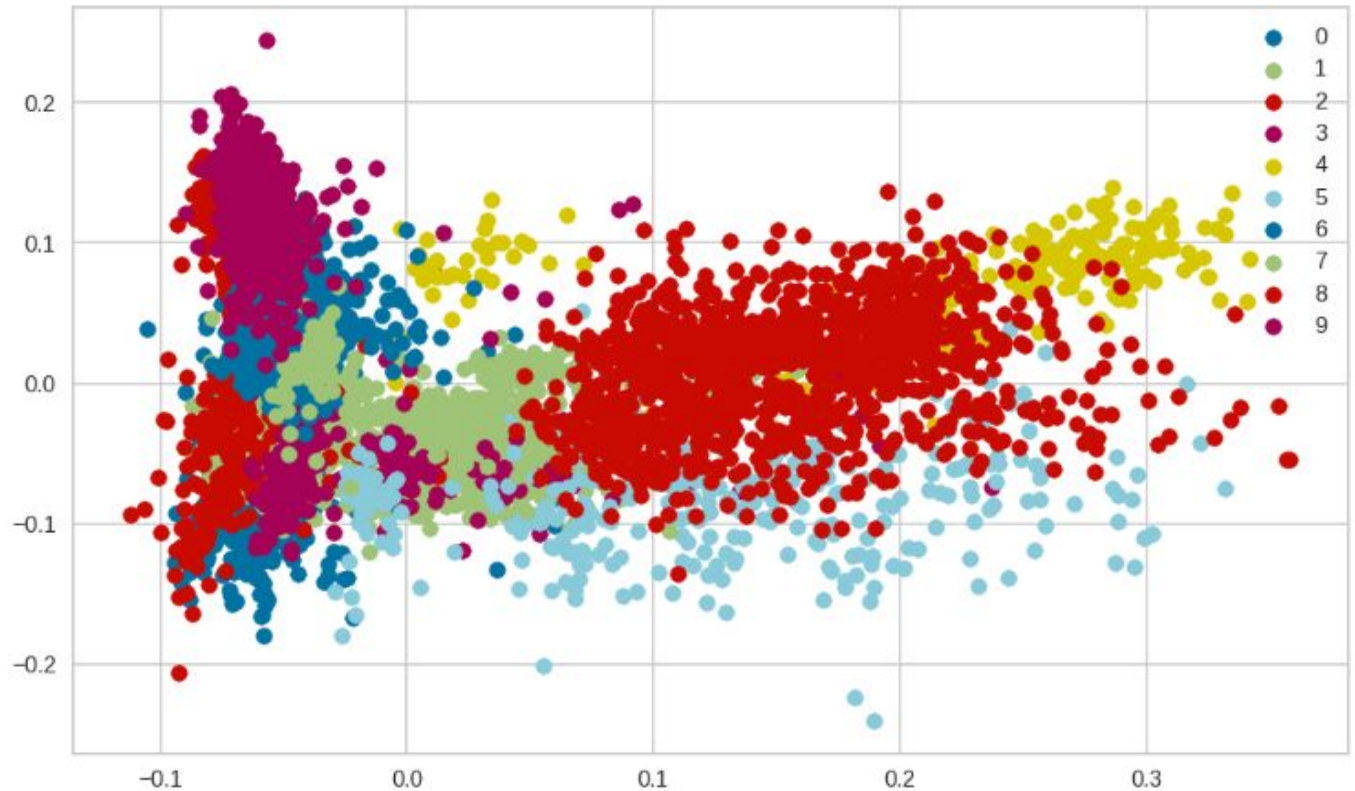
Data represented via Clusters

Cluster 0: Documentaries.
Cluster 1: Family and Children Movies.
Cluster 2: Musical Movies and Documentaries.
Cluster 3: Stand Up Comedy and Comedy Shows.
Cluster 4: Korean and Romantic Tv Shows.
Cluster 5: Science, Nature, Reality, Crime Tv Shows and Docuseries.
Cluster 6: International Movies.
Cluster 7: Anime Series and Tv Shows.
Cluster 8: International Tv Shows.
Cluster 9: Action, Adventure and Independent Movies.

Optimum K value using Metric : Silhouette Method and Elbow Method



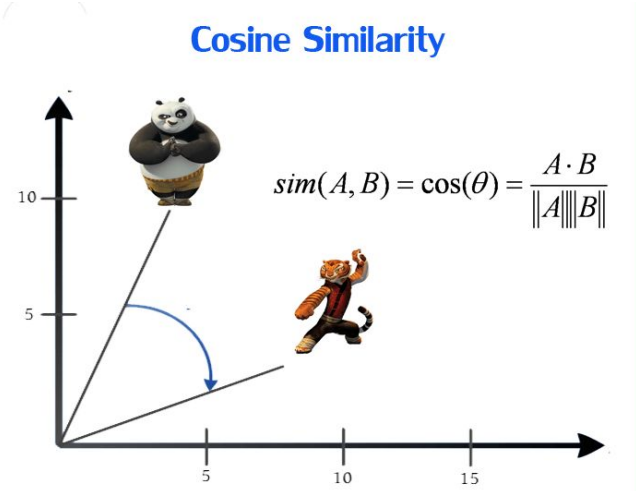




- Here we performed clustering considering K = 10.
- The numbers 0 to 9 represent 10-distinct clusters formed by K-means clustering.
- Each cluster contains data points similar to those in the same groups but varies from other groups.

GETTING RECOMMENDATIONS.....

Obtained recommendations using cosine similarity.




```
# Lets try getting recommendations for Movies.
movie_recommendations = pd.DataFrame(recommend('Naruto'), columns=['Recommendations'])
movie_recommendations.head(11)
```

	Recommendations
0	Naruto Shippûden the Movie: Bonds
1	Naruto Shippuden : Blood Prison
2	Naruto the Movie 2: Legend of the Stone of Gelel
3	Naruto Shippuden: The Movie
4	Naruto Shippûden the Movie: The Will of Fire
5	Naruto the Movie 3: Guardians of the Crescent ...
6	Naruto Shippuden: The Movie: The Lost Tower
7	Marvel Anime: Wolverine
8	Dragon's Dogma
9	Saint Seiya: The Lost Canvas

```
#checking for a tv show
movie_recommendations = pd.DataFrame(recommend('Sacred Games'), columns=['Recommendations'])
movie_recommendations.head(11)
```

	Recommendations
0	AK vs AK
1	Tukaram
2	Baazaar
3	Seven (Tamil)
4	Raman Raghav 2.0
5	GHOUL
6	The Indian Detective
7	Seven (Telugu)
8	Lust Stories
9	Bombay Talkies

CONCLUSION.....

NETFLIX

A Case of **NETFLIX MOVIES** AND **TV SHOWS CLUSTERING**

- First comes first, Information about our data set which comprises of 7787 rows and 12 columns, where columns like director, cast, country, date_added had some null values which were treated accordingly.
- Insights from our Exploratory Data analysis :

  ✔ 68.1% of the content available on Netflix are movies ands 30.9% of the content are TV shows.
  ✔ Relative Growth is observed here in the number of movies on Netflix than TV Shows
  ✔ 2015 was the year where the spike of growth began , 2019 and 2020 were the peak years where highest number of movies and tv shows were added on Netflix.
  ✔ It is noticed that US, India and UK majorly create movies on this platform.
  ✔ Jan Suter is the most popular director on Netflix
  ✔ Anupam Kher and Shahrukh Khan are the most popular actors on Netflix.

- By applying Silhoutte Score Method we found the optimum value of **K = 10**.
- Using the given data set a simple recommender system was also created using the cosine similarity and recommendations for TV shows and movies were obtained.

Future scope : If the dataset is integrated with IMDB dataset and Rotten Tomatoes Data set better and interesting insights can be noted.

Thank You !