Capstone Project - IV

Unsupervised ML

Netflix Movies and TV shows Clustering

BY

Yashwant B. Raul
(yashwantraul24@gmail.com)
Mayur S. Marathe
(marathemayu1990@gmail.com
Sanket Gawali
(sanketgawali23@gmail.com)

Content

□ Project Overview

□ Problem Description

□ Feature Summary

□ Exploratory Data Analysis

□ Data Preprocessing

□ K-Means Clustering

□ Clustering Technique For Finding Value Of K

□ Conclusion

Project Overview

Netflix is one of the leading OTT platforms, not only in India but also internationally Netflix manages a large collection of TV shows and movies, streaming it anytime via online. The success of the OTT platforms depends on two things- the variety of content and appropriate recommendations to the users. This business is profitable because users make a monthly payment to access the platform.

In this project, we focused on

- 1. Exploratory Data Analysis
- 2. Understanding what type content is available in different countries
- 3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
- 4. Clustering similar content by matching text-based features

Problem Description

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Fixable 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 services 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.

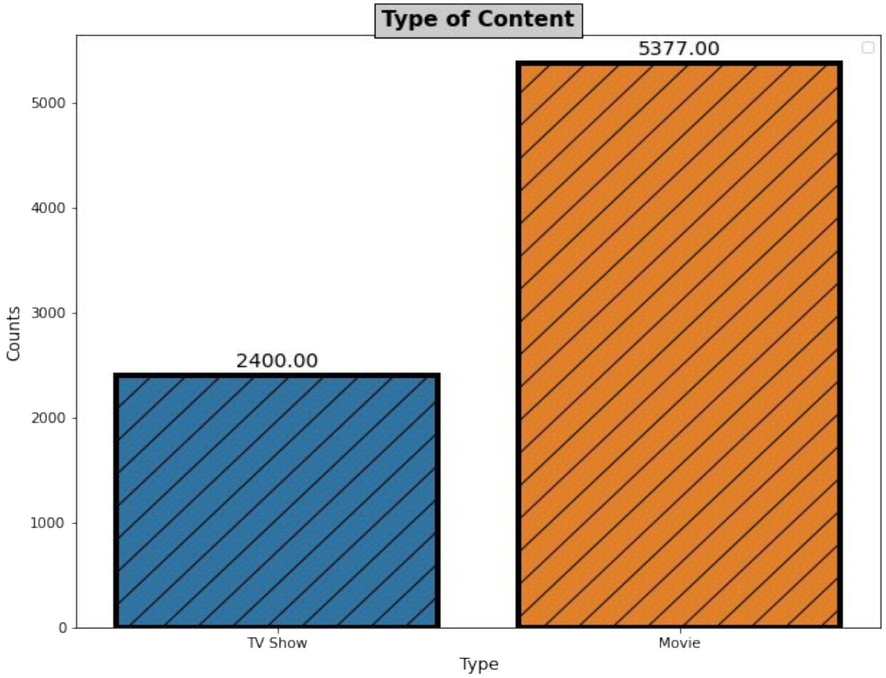
Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

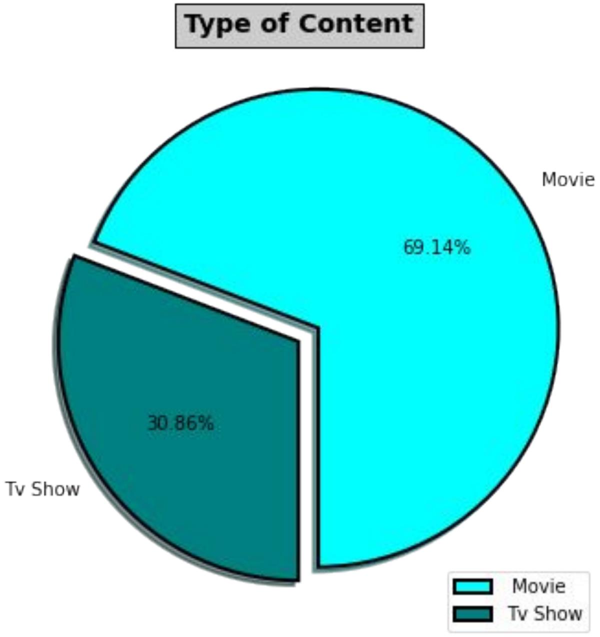
Feature Summary

- 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 Release year 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

Exploratory Data Analysis

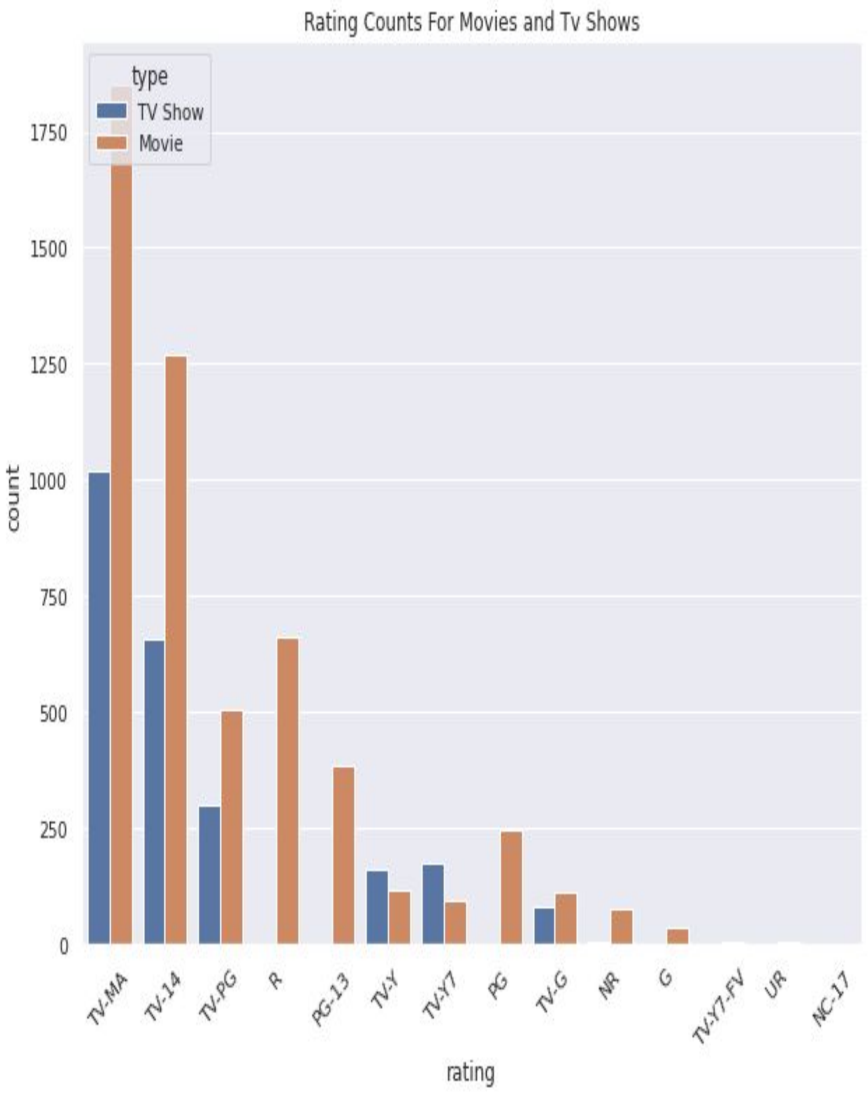
Type of content available on Netflix





It is evident that there are more movies on Netflix than TV shows.
Netflix has 5377 movies, which is more than double the quantity of TV shows

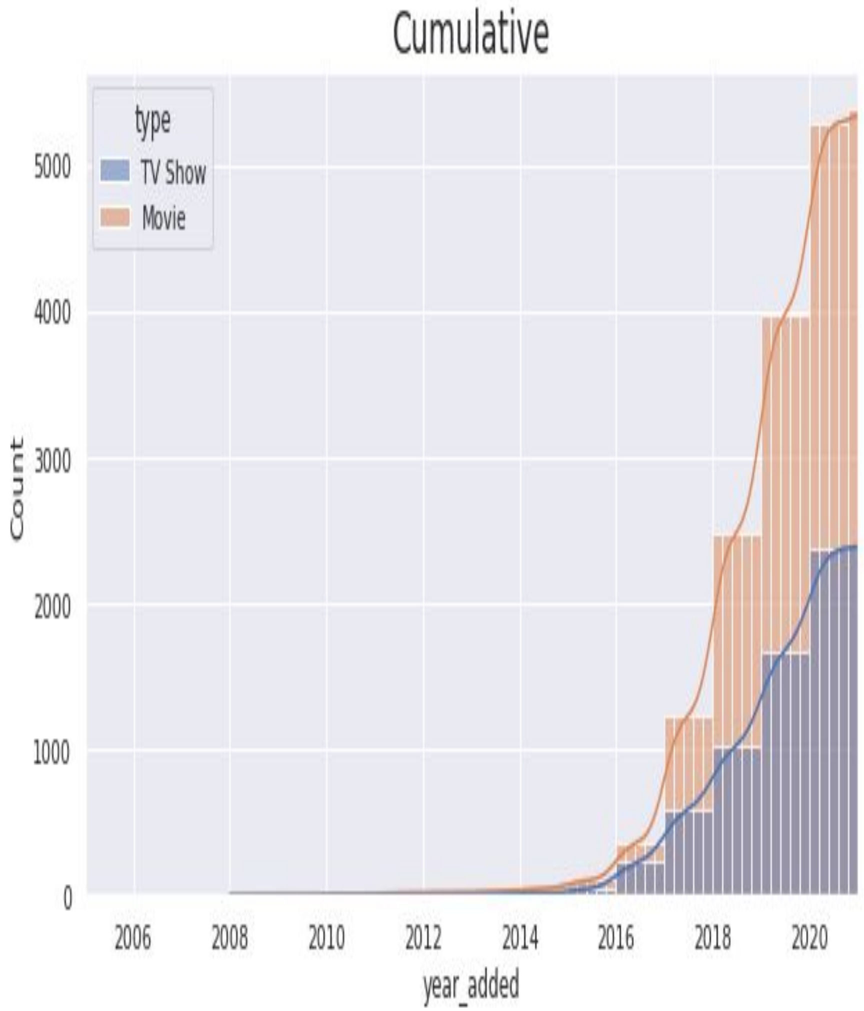
Movie ratings analysis



The 'TV-MA' rating is used in the majority of the film. The TV Parental Guidelines provide a "TV-MA" classification to a television programmed that is intended solely for mature audiences.

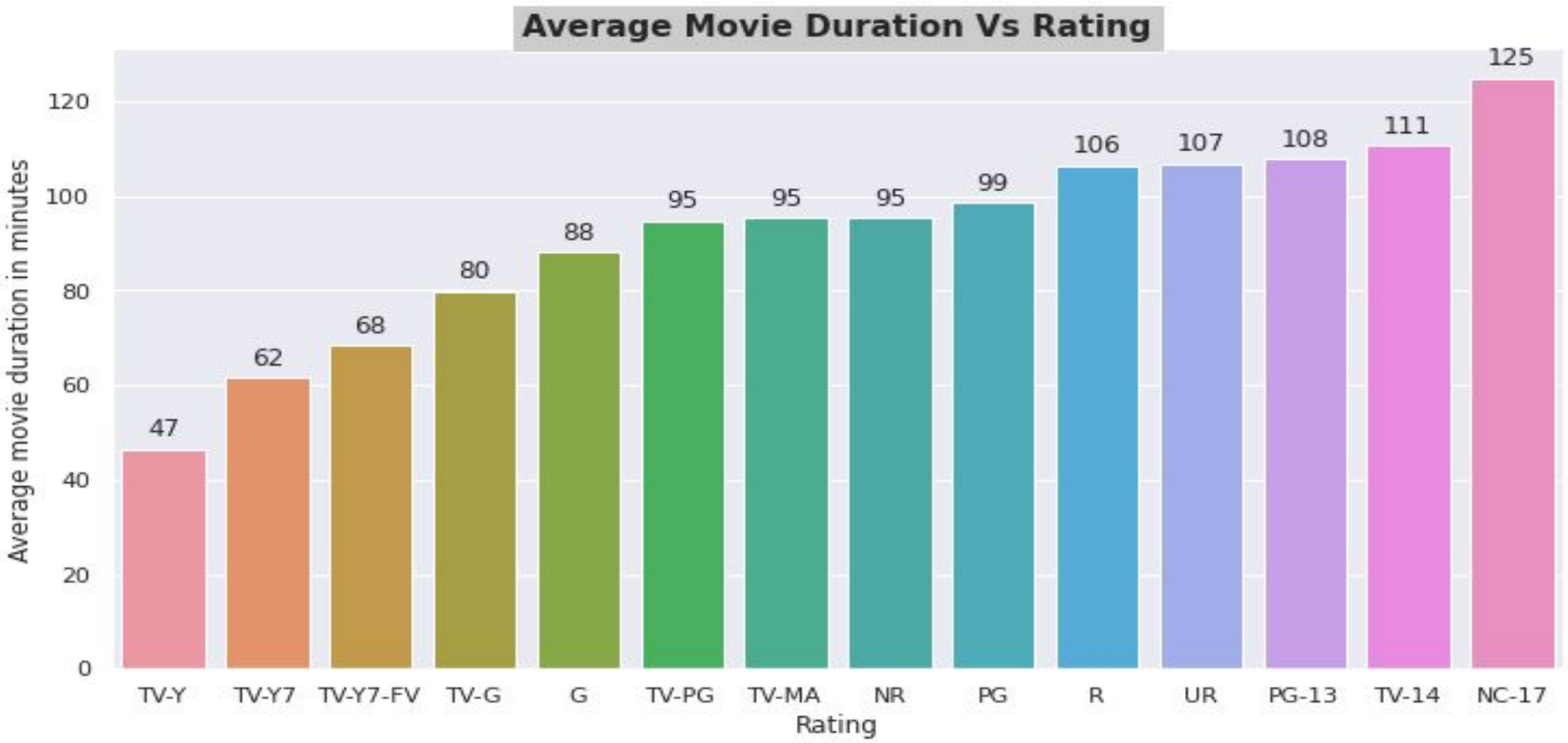
- The second largest is 'TV-14,' which stands for content that may be inappropriate for minors under the age of 14.
- The third most common is the extremely popular 'R' rating. The Motion Picture Association of America defines an R-rated film as one that contains material that may be inappropriate for children under the age of 17; the MPAA states that "Under 17 requires accompanying parent or adult guardian

Growth in content over the years



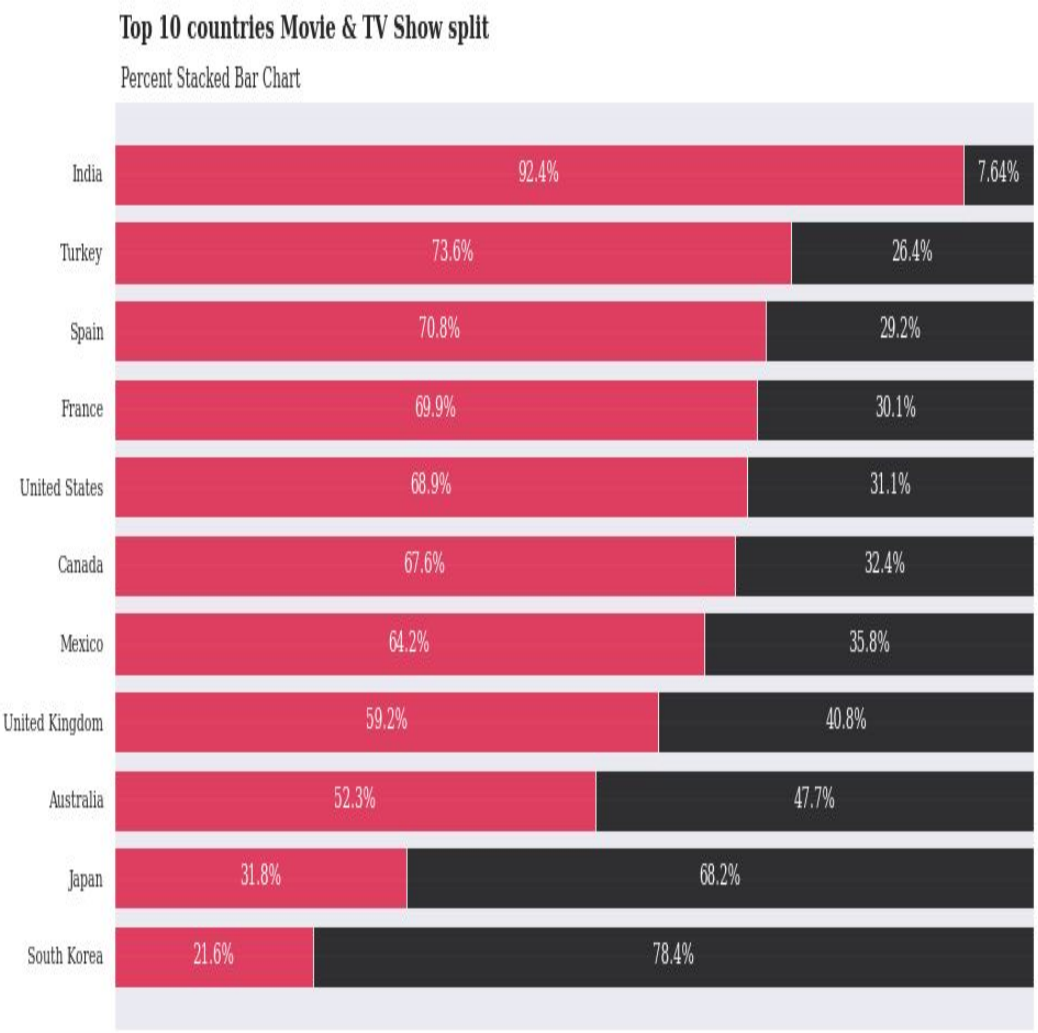
- The number of movies on Netflix is growing significantly faster than the number of TV shows.
- In both 2018 and 2019, approximately 1200 new movies were added.
- We saw a huge increase in the number of movies and television episodes after 2014.
- It appears that Netflix has focused more attention on increasing Movie content that TV Shows. Movies have increased much more dramatically than TV shows

Average Movie Duration Vs Rating



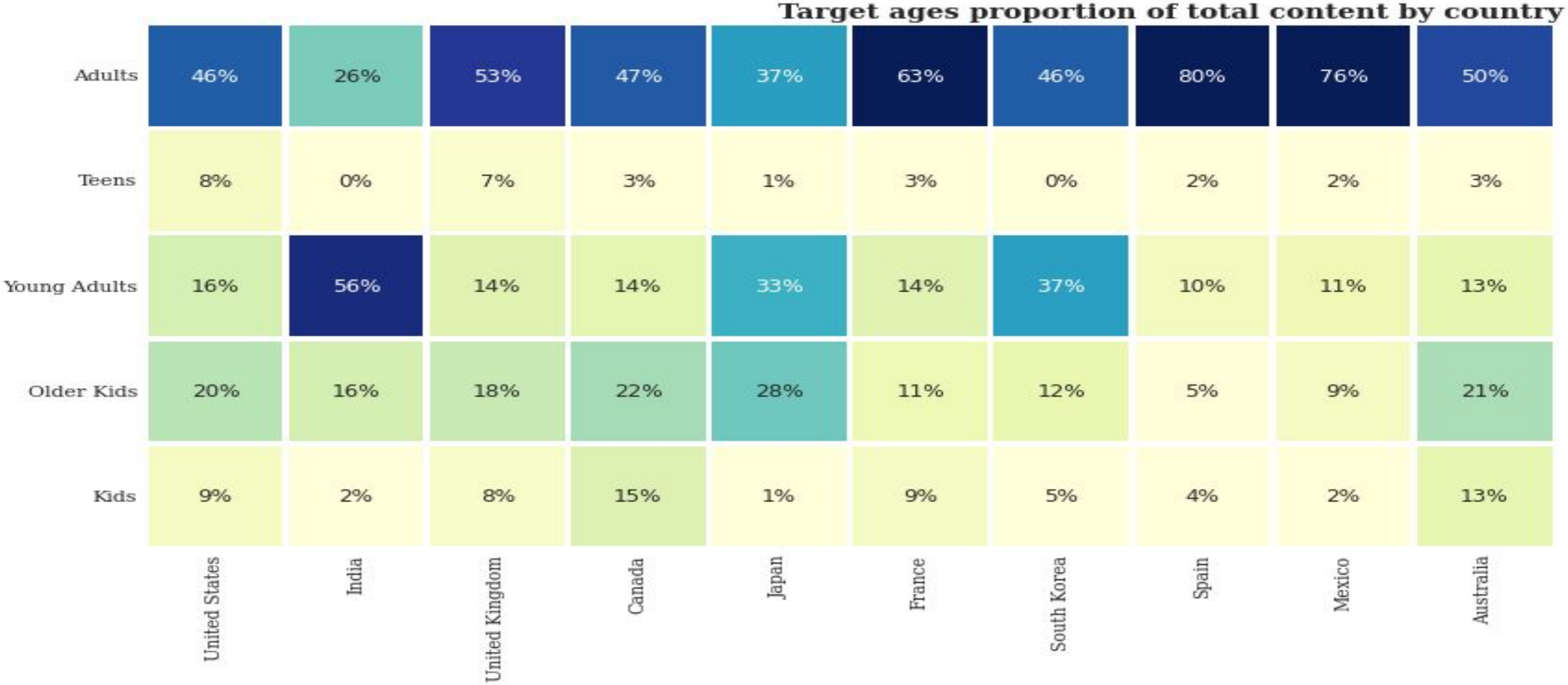
- Those movies that have a rating of NC-17 have the longest average duration.
- When it comes to movies having a TV-Y rating, they have the shortest runtime on average.

List of country differ by content



- The majority of the content on Netflix in India is comprised of movies.
- Bollywood is a significant business, and movies, rather than TV shows, may be the industry's major focus.
- South Korean Netflix on the other hand is almost entirely TV Shows.
- The fundamental reason for the variation in content must be due to market research undertaken by Netflix

Netflix Content for different age groups in top 10 countries

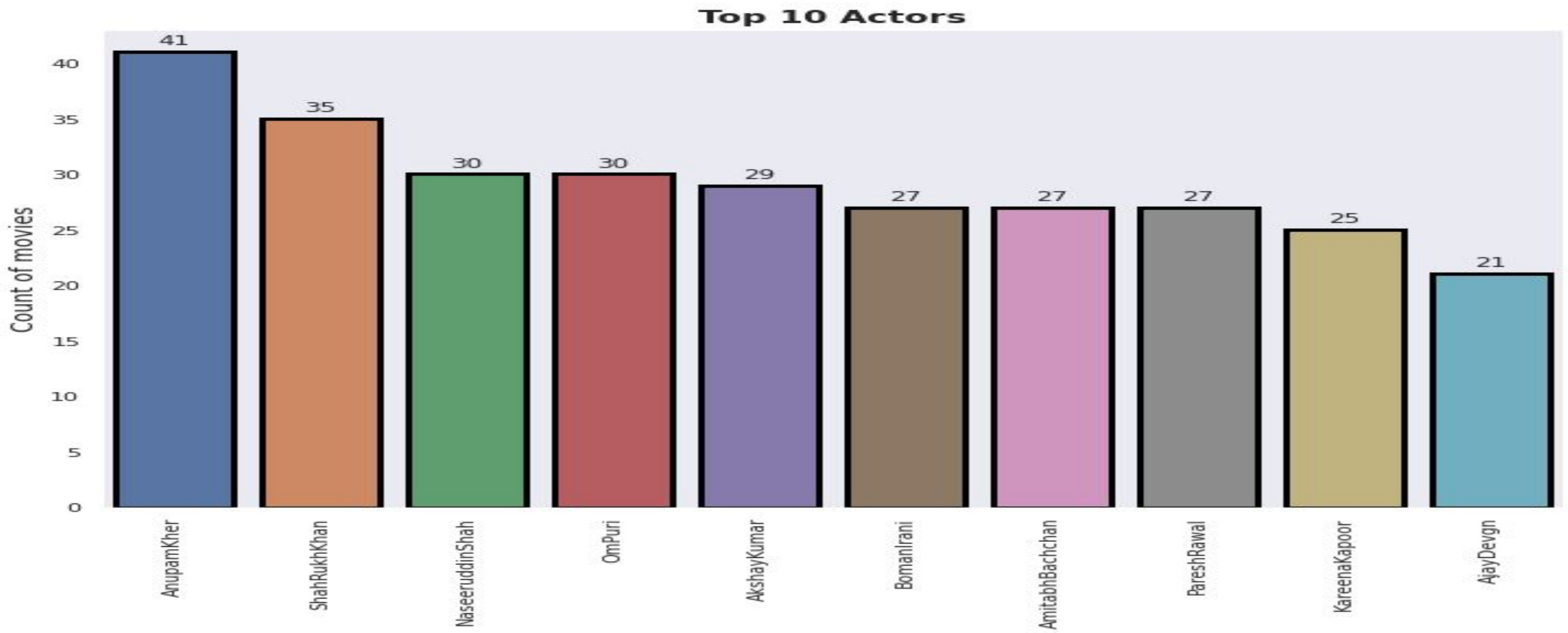


Target ages proportion of total content by country

	United States	India	United Kingdom	Canada	Japan	France	South Korea	Spain	Mexico	Australia
Adults	46%	26%	53%	47%	37%	63%	46%	80%	76%	50%
Teens	8%	0%	7%	3%	1%	3%	0%	2%	2%	3%
Young Adults	16%	56%	14%	14%	33%	14%	37%	10%	11%	13%
Older Kids	20%	16%	18%	22%	28%	11%	12%	5%	9%	21%
Kids	9%	2%	8%	15%	1%	9%	5%	4%	2%	13%

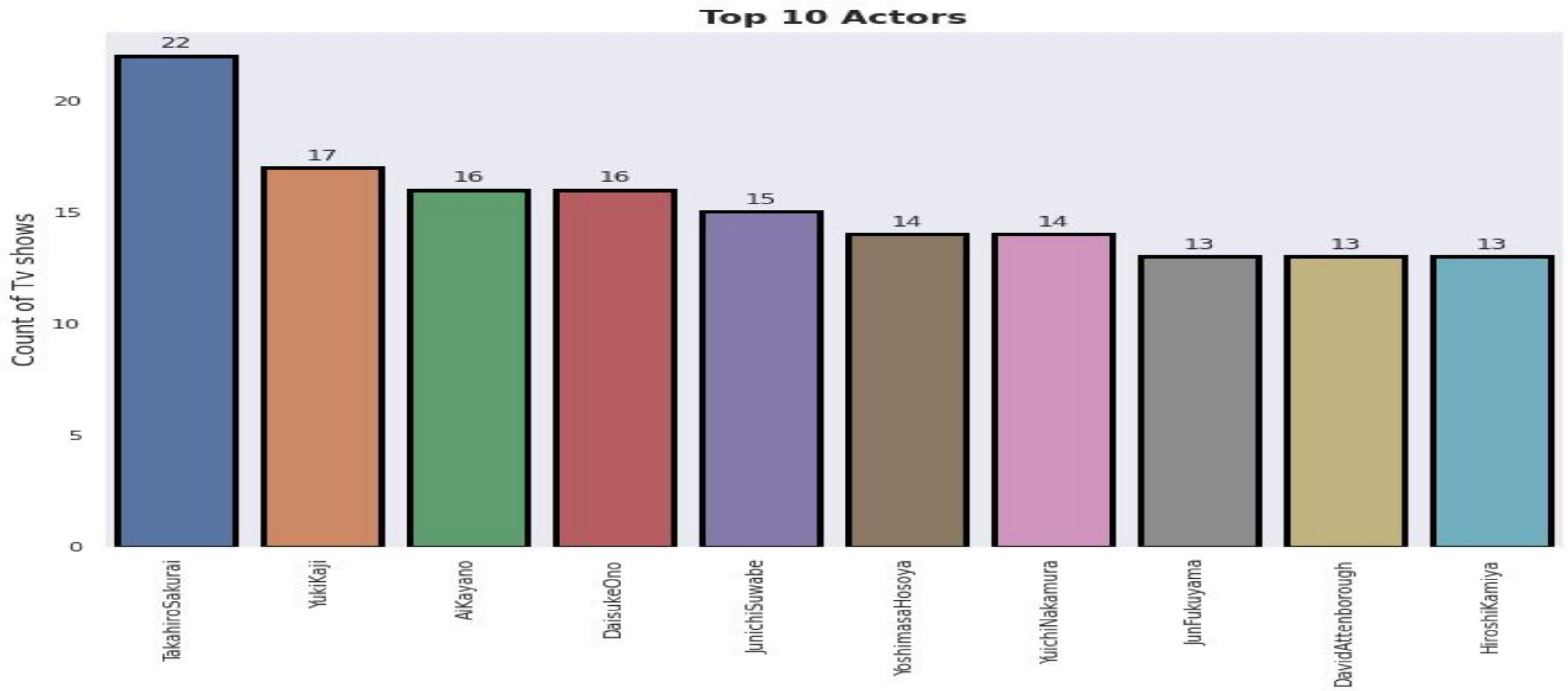
- It is also interesting to see parallels between culturally comparable nations - the US and UK are closely aligned with their Netflix target ages, but radically different from, example, India or Japan!
- Also, Mexico and Spain have similar content on Netflix for different age groups

Top 10 Actors who appear in the majority of films



- According to the above bar plot, Anupam Kher has worked in over 41 films.
- After Anupam Kher, Shahrukh Khan is ranked second, with 35 films under his belt.
- Naseeruddin Shah and Om Puri have worked in 30 films

Top 10 Actors who appear in the majority of TV Shows



- According to the above bar plot, Takahiro Sakurai has worked in over 22 TV shows.
- After Takahiro Sakurai, Yuki Kaji is ranked second, with 17 TV shows under his belt.
- Ai Kayano and Daisuke Ono have worked in 16 TV shows

Data Pre-processing

Removing Punctuation: Punctuations does not carry any meaning in clustering.

So, removing punctuations helps to get rid of unhelpful parts of the data, or noise.

Removing Stop words: Stop words are basically a set of commonly used words in any language, not just in English.

If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Stemming: It is the process of removing a part of a word, or reducing a word to its stem or root.

Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

K-Means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre defined distinct non overlapping subgroups where each data point belongs to only one group

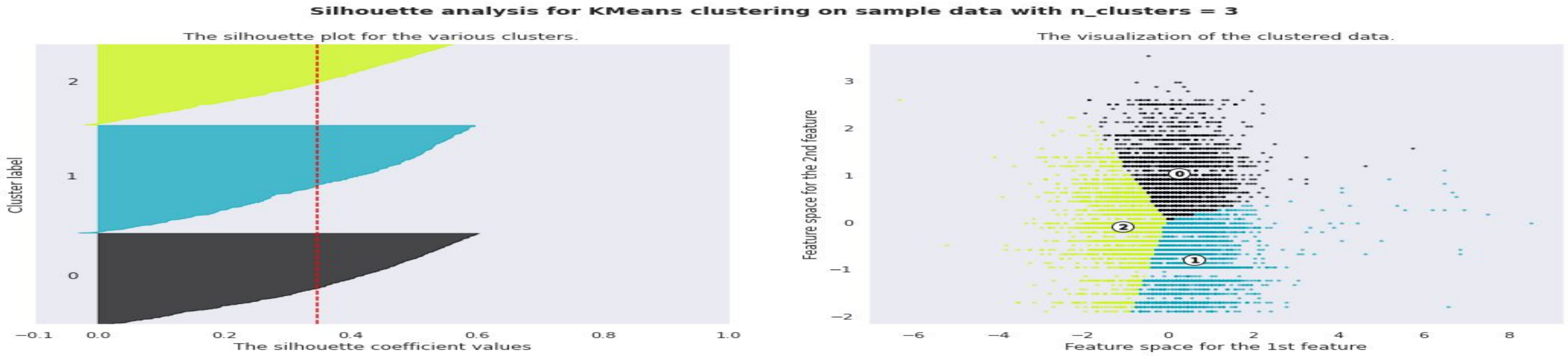
Vectorization: Clustering algorithms cannot understand textual data. So, we use vectorization technique to convert textual data to numerical vectors..

Elbow Curve: The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

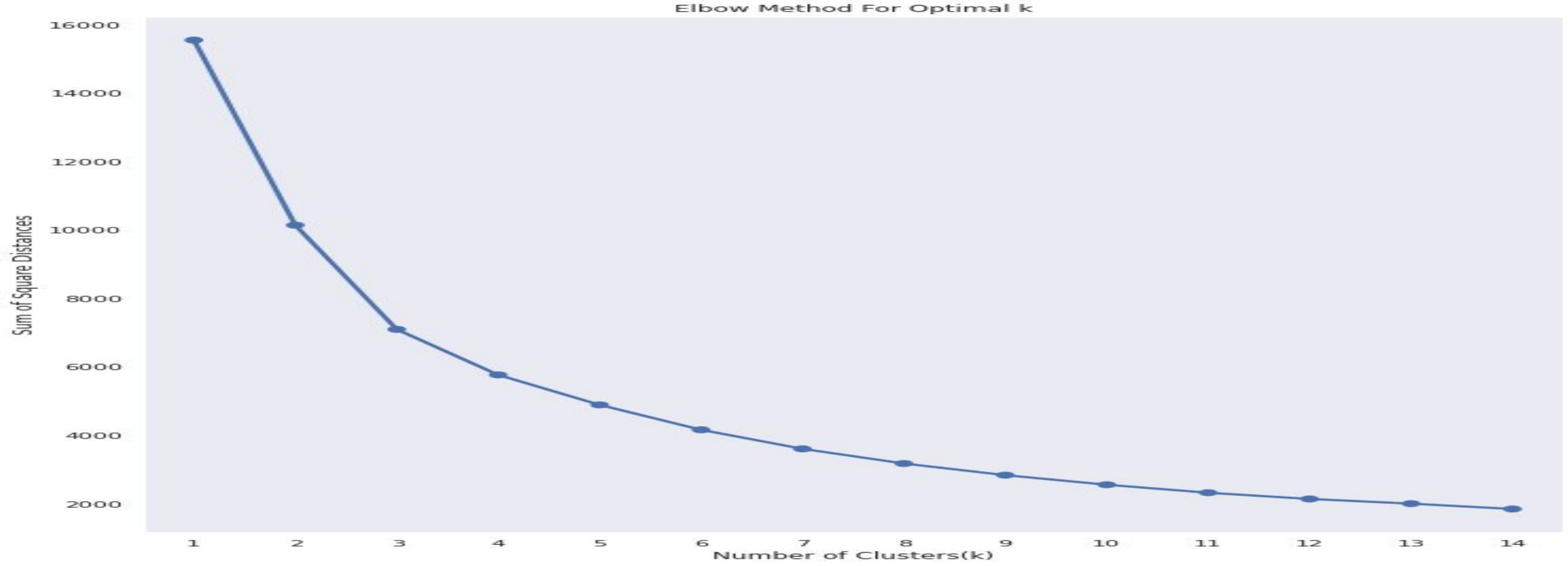
Silhouette score : Silhouette score is used to evaluate the quality of clusters created. Using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each

Clustering technique used to find out best value of K

Silhouette score

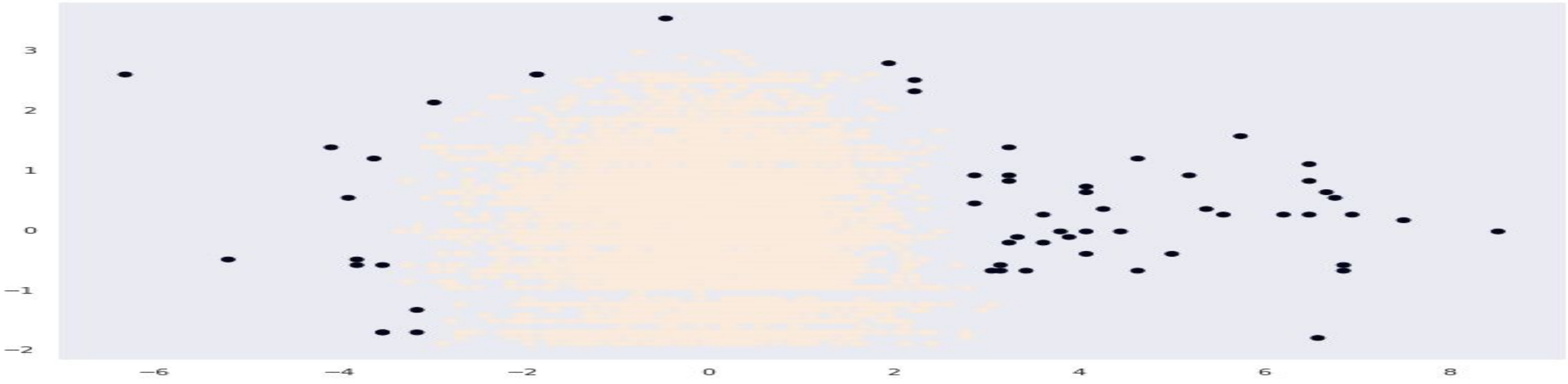


Elbow Method

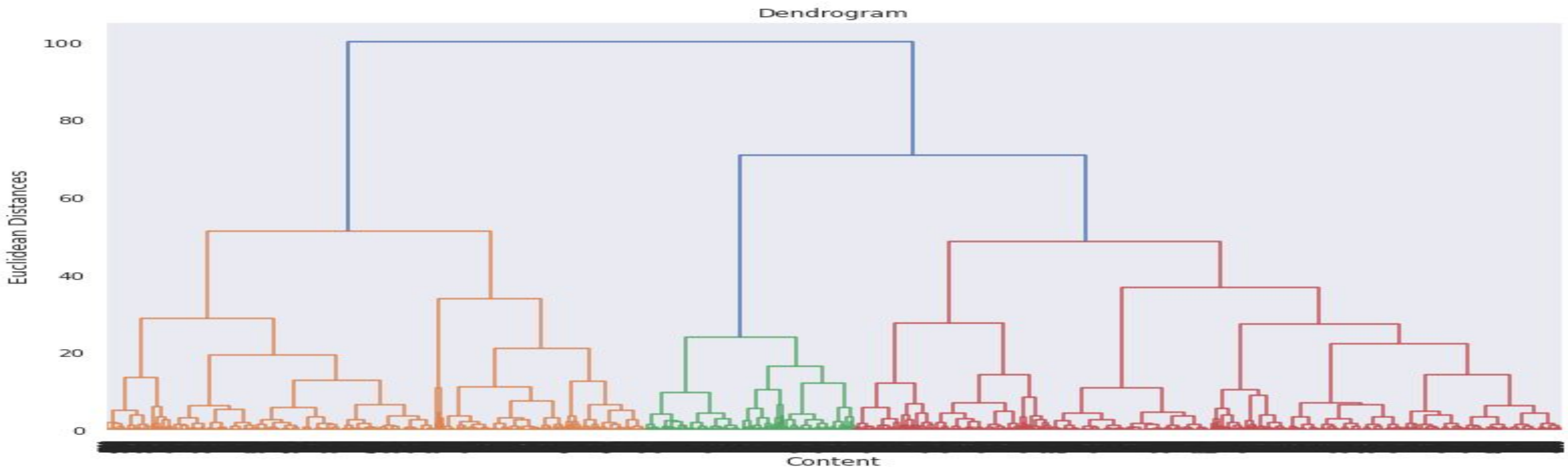


Clustering technique used to find out best value of K

DBSCAN



Dendrogram



Conclusion

- ✔ Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation.
- ✔ We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies)
- ✔ By analyzing the content added over years we get to know that in recent years Netflix is focusing movies than TV shows.
- ✔ The most number of the movies and TV shows release in 2017 and 2020 respectively and united nation have the maximum content on Netflix
- ✔ On Netflix, Dramas genre contains the maximum content among all of the genres and the most of the content added in December month and less content in February
- ✔ By applying the silhouette score method for n range clusters on dataset we got best score which is 0.348 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 curve gets linear it means k = 3 will be the best cluster
- ✔ By applying different clustering algorithms to our dataset, we get the optimal number of cluster is equal to 3