Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Sonali Kaushal
Somya Hingorani
Nitesh Verma
Harsh Vardhan
Prateek Gupta

Problem statement

- 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 findings.

In this project, we are required to do

A. Exploratory Data Analysis
B. Understanding what type content is available in different countries
C. Is Netflix increasingly focusing on TV rather than movies in recent years.
D. Clustering similar content by matching text-based features

Agenda

To discuss the analysis of given Netflix Movies and TV Shows dataset.

Topics covered for the project :

- Data Pipeline
- Data Summary
- Data Description
- Data Cleaning
- Feature engineering
- Exploratory Data Analysis
- Data Preprocessing
- Dimensionality reduction using PCA
- Clustering Analysis
- Topic Modelling using LDA
- Conclusion

Data Pipeline

- Data pre-processing: We have pre-processed the data by dealing with the null values. Null values were present in director, cast, country, date_added and rating columns. Since there were very few missing values in date_added and rating, we have removed them. We have checked for duplicate values also but didn't find any.

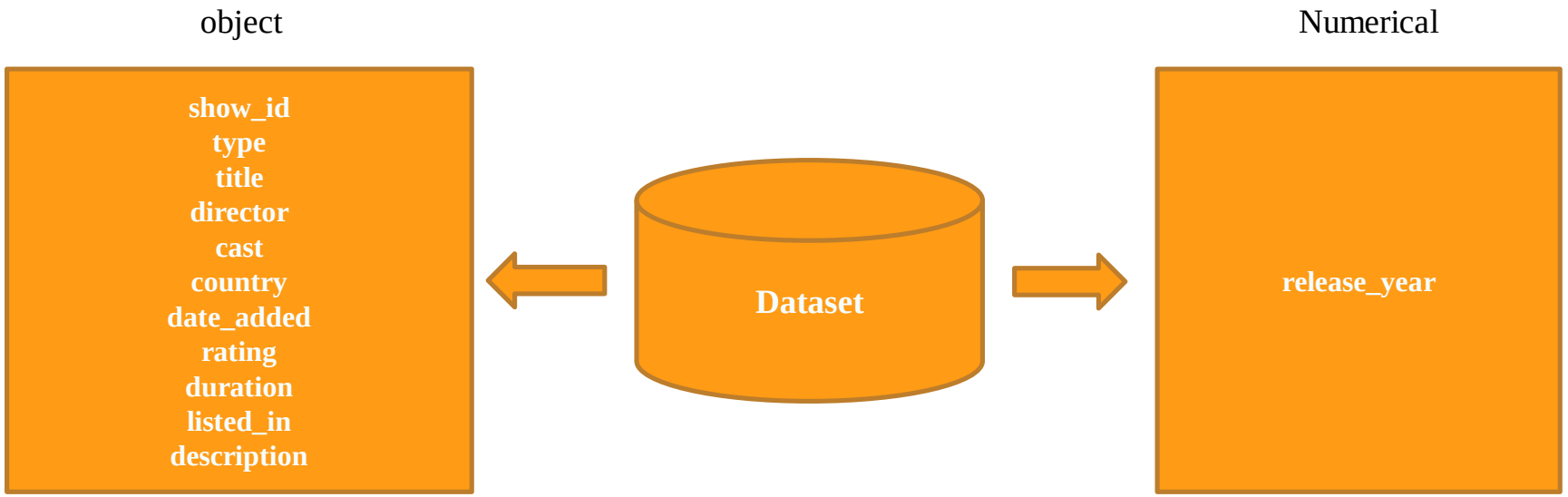
- Feature engineering: In this part some new features in the form of day_added, month_added , year_added from date_added column. Genres are extracted from the listed_in column and redefined accordingly.

- Exploratory Data Analysis (EDA): In this part we have done analysis such as type of content available, content growth over the years, top actors, top directors etc. to get some meaningful insights from the dataset.

- Data pre-processing for cluster formation - Methods used for this are Tokenization, Punctuation Removal, Stopwards Removal, Stemming, Text Vectorization.

- Clustering Analysis : Under this different clustering techniques are used and topic modelling has also performed using LDA technique.

Data Summary



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 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

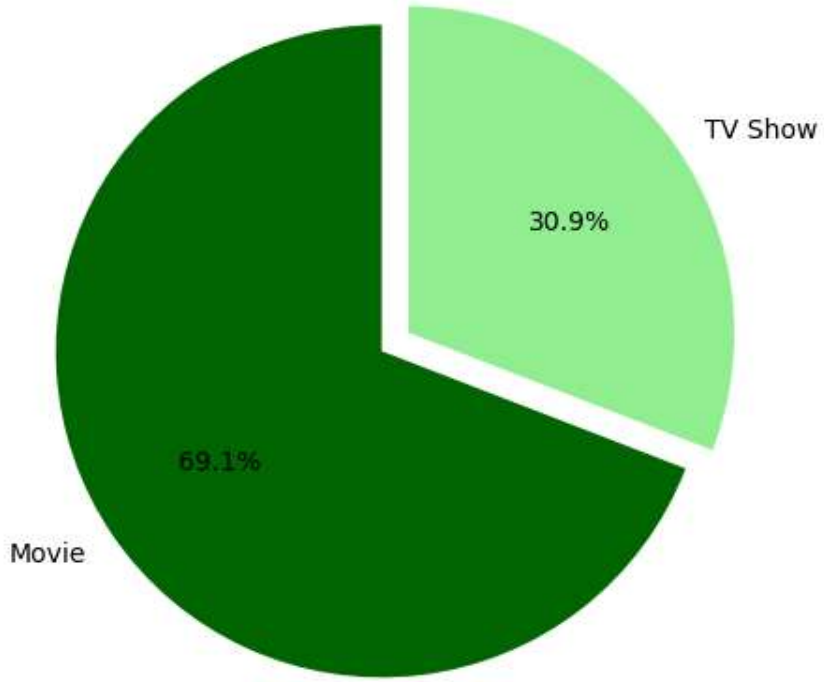
Feature engineering

Steps followed

- 3 new features were extracted from data_added column – day, month & year.
- Genres are extracted and re-defined accordingly. TV and Movie genres are separately defined.
- Topics like International TV Shows are removed as it could bring in a bias by displaying content in reference to American movies.

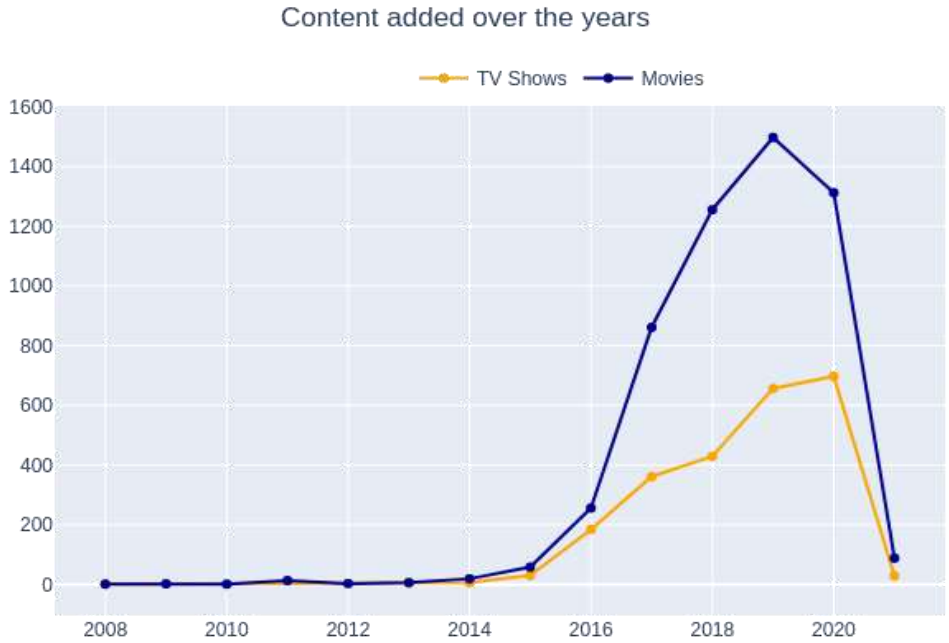
Exploratory Data Analysis

Content type on Netflix



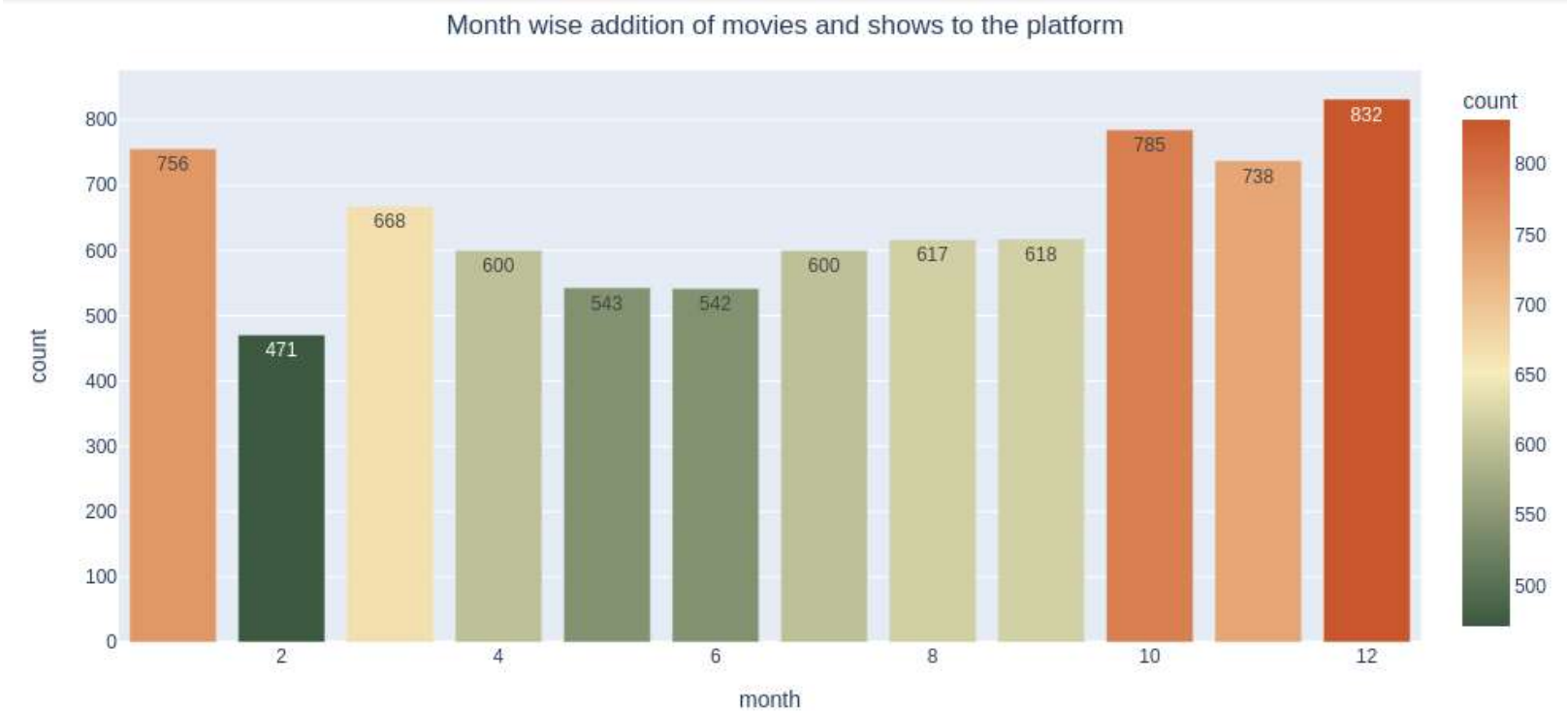
- 69.1% of the content available on Netflix are movies and remaining 30.9% are TV Shows.

Content growth over the years



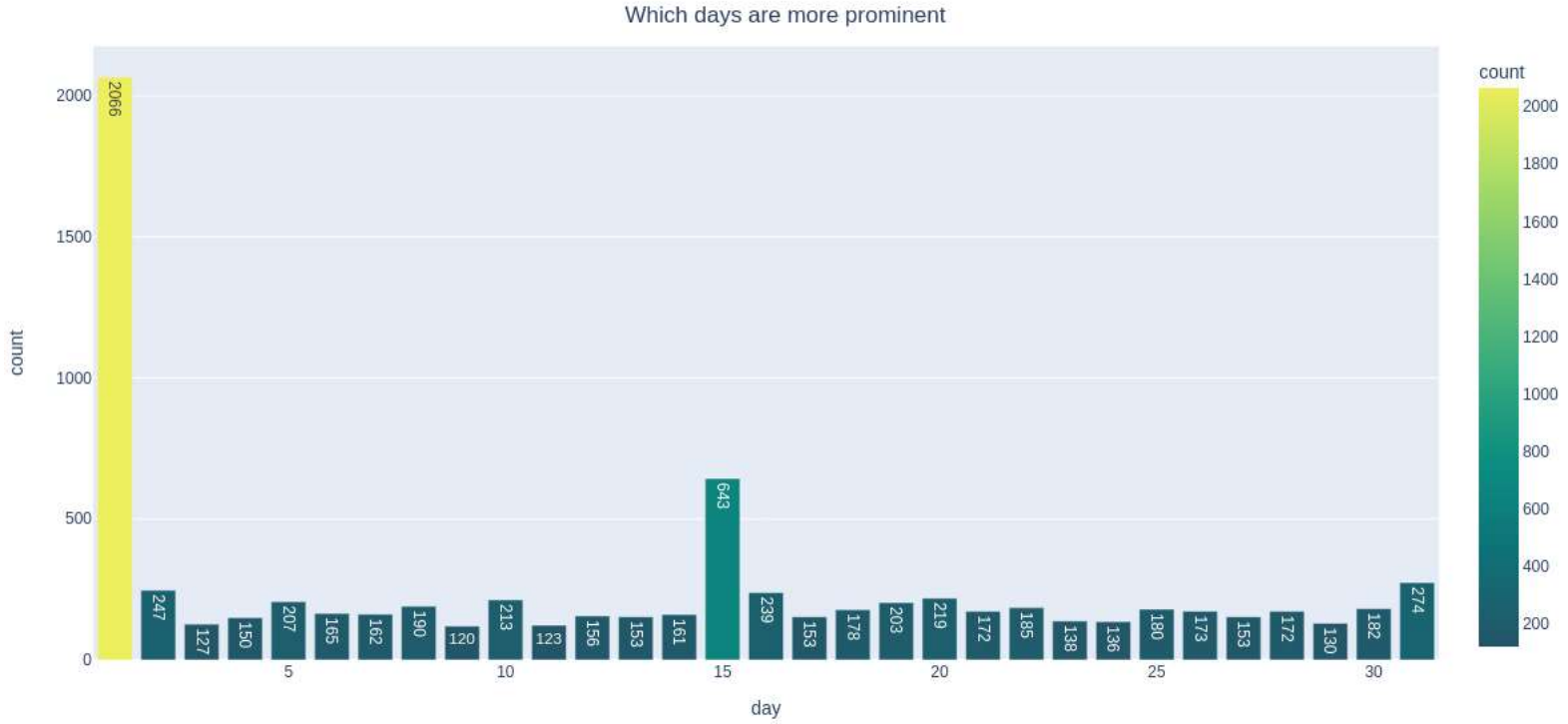
- Given Line Plot shows, growth in the number of movies on Netflix is much higher than TV shows.
- From 2015, we can see a noticeable growth in the number of movies and TV shows uploaded by Netflix on its platform.
- The highest number of movies and tv shows got added in 2019 and 2020.
- Also, very few movies and tv shows got added in 2021.

In which month do most movies and tv shows get added?



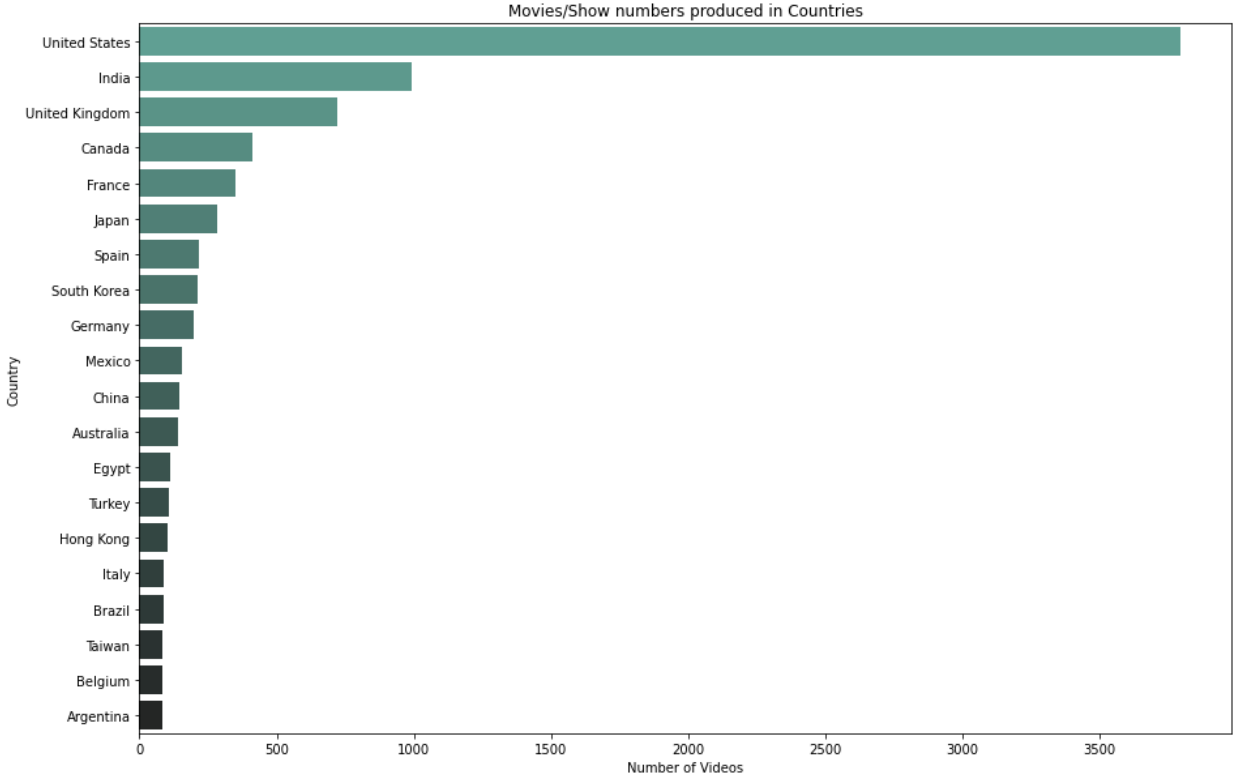
- Most of the content is uploaded either by year ending or beginning.
- October, November, December, and January are months in which many shows and movies get uploaded to the platform.
- It might be due to the winter, as in these months people stay at home and watch shows and movies in their free time.

Which days are more prominent?



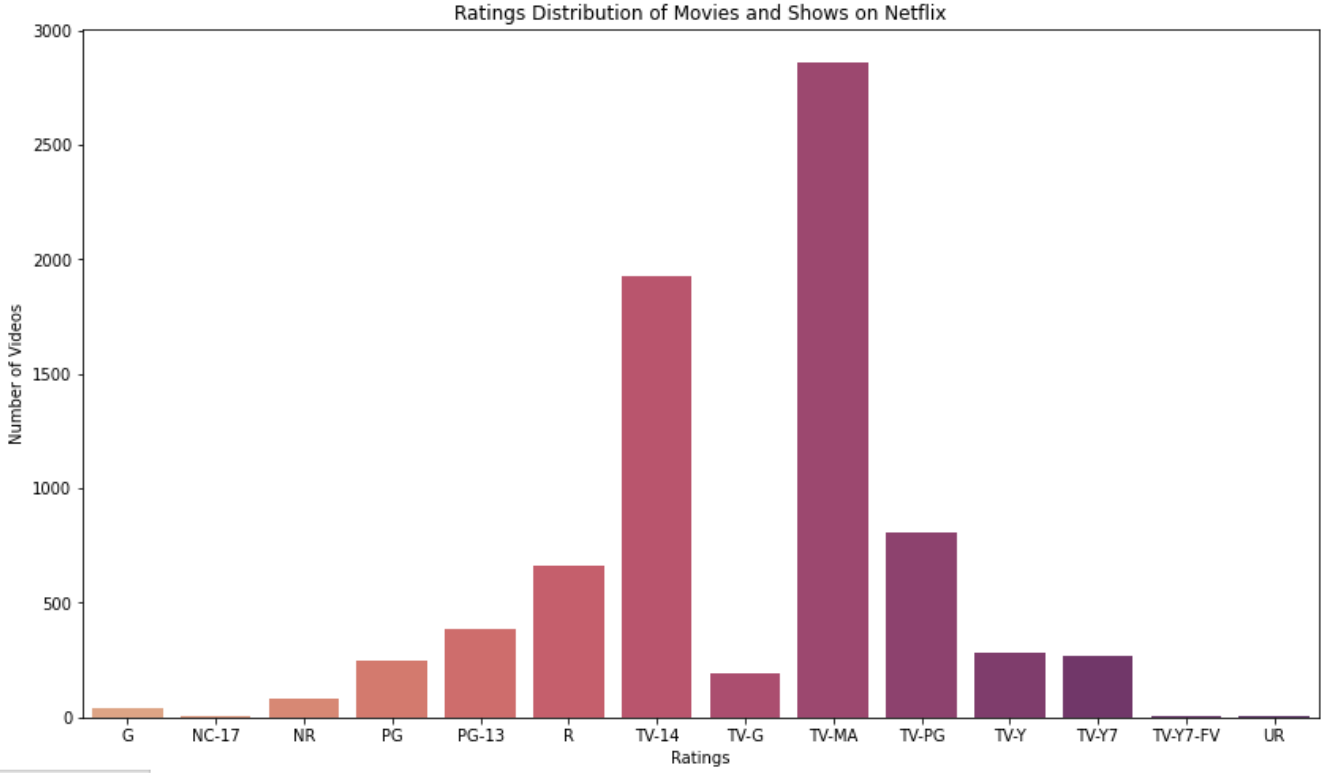
- Most of the content gets uploaded in the beginning and the middle of the month.

Top 10 Countries that produced content on Netflix



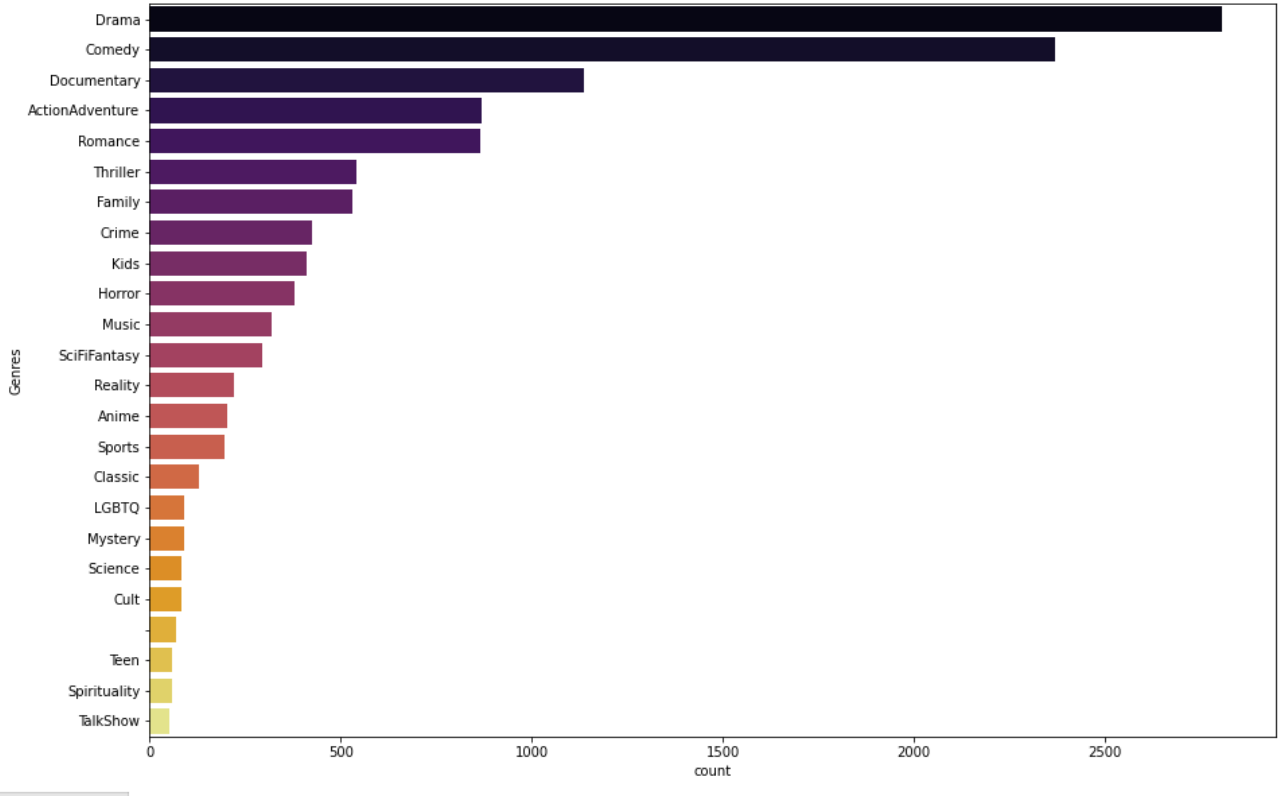
- The majority of the content providers are in the above top-ten countries.
- Among which USA, India, and UK create more than half of the tv shows and movies on the platform.

Different Ratings on the Platform



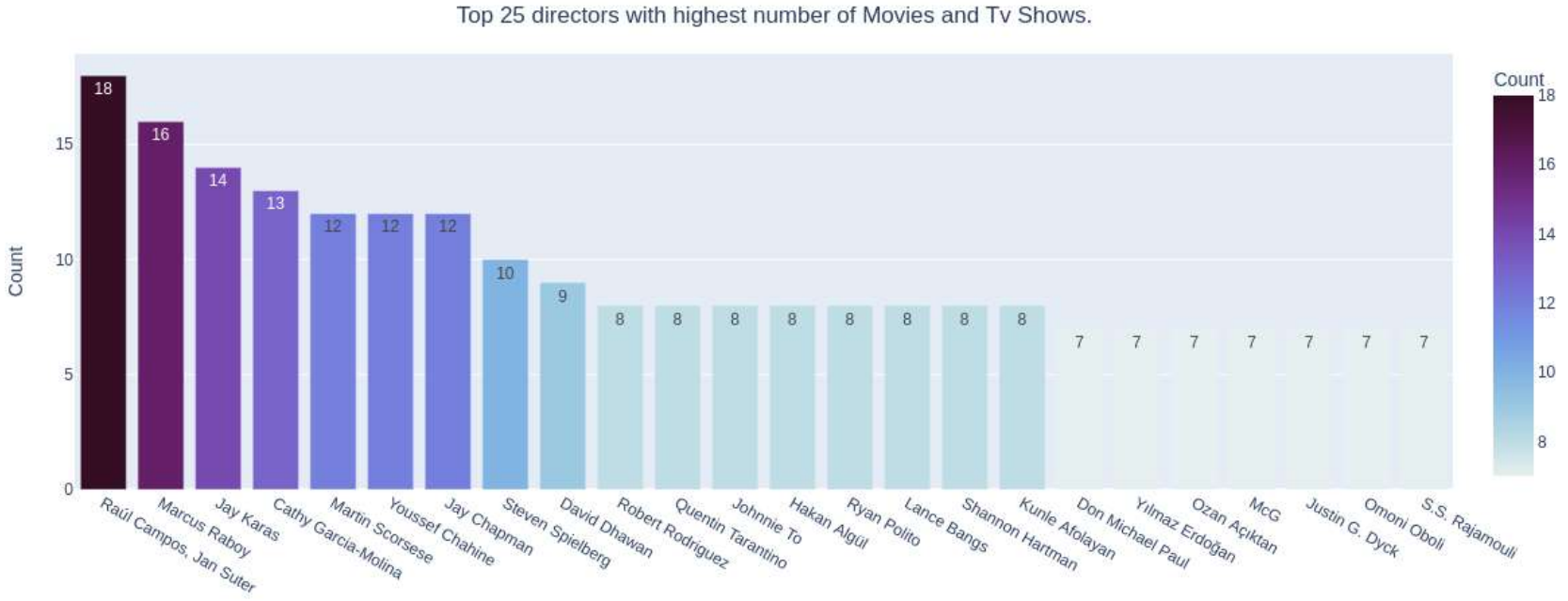
- Most content on Netflix is rated for Mature Audiences and over 14 years old

Top Genres on the Platform



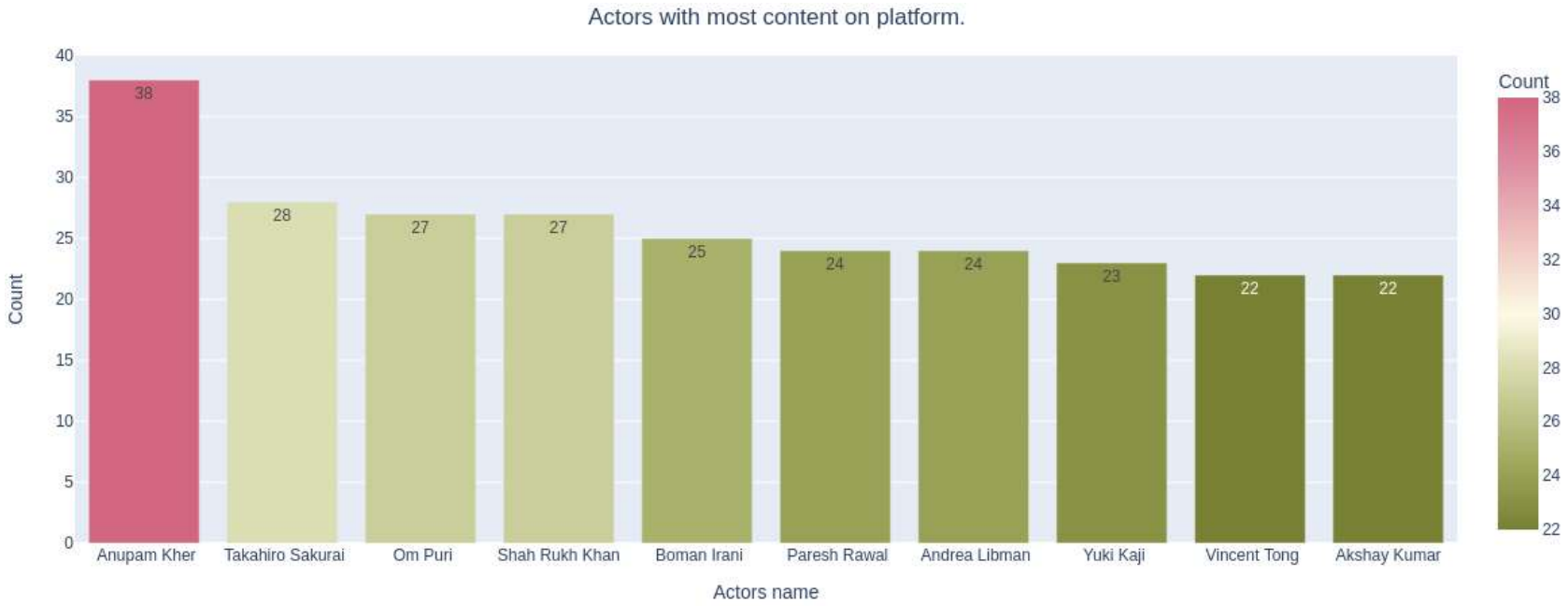
- Top Genres on Netflix are found to be: Drama, Comedy, Documentary, Action and Adventure, Romance etc.

Which directors have most number of the movies and tv shows?



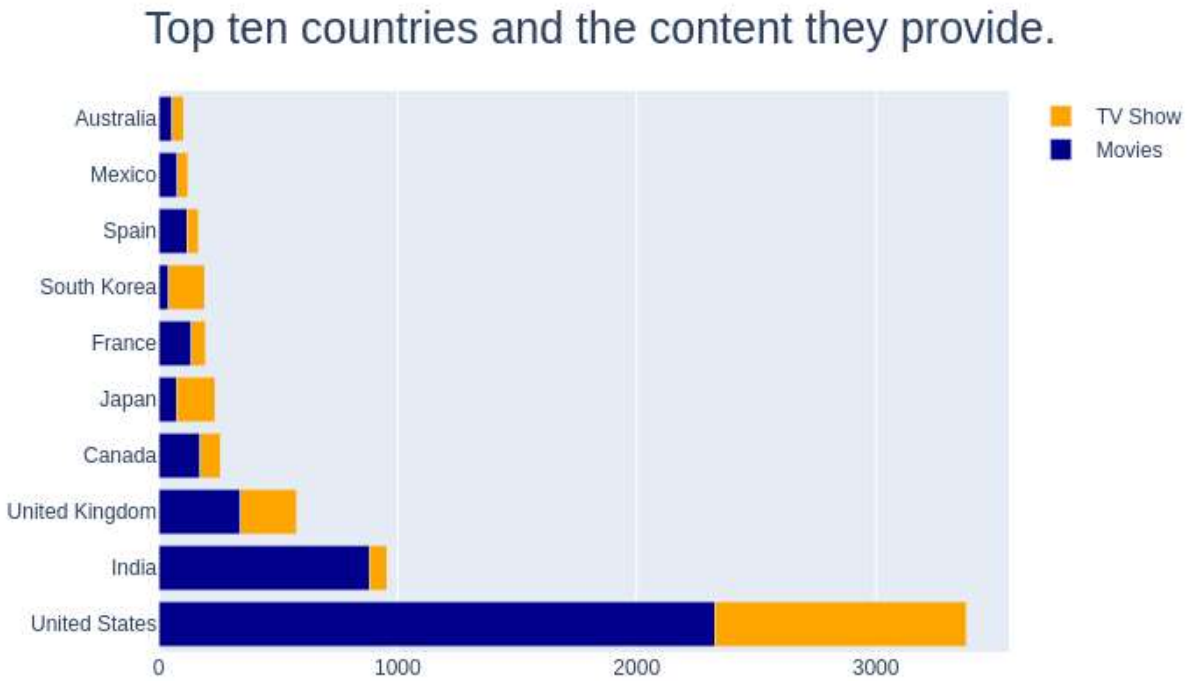
- Raúl Campos, Jan Suter, Marcus Raboy, Jay Karas, Cathy Garcia-Molina, Jay Chapman are the top 5 directors having most numbers of movies and TV shows.

Top 10 Actors



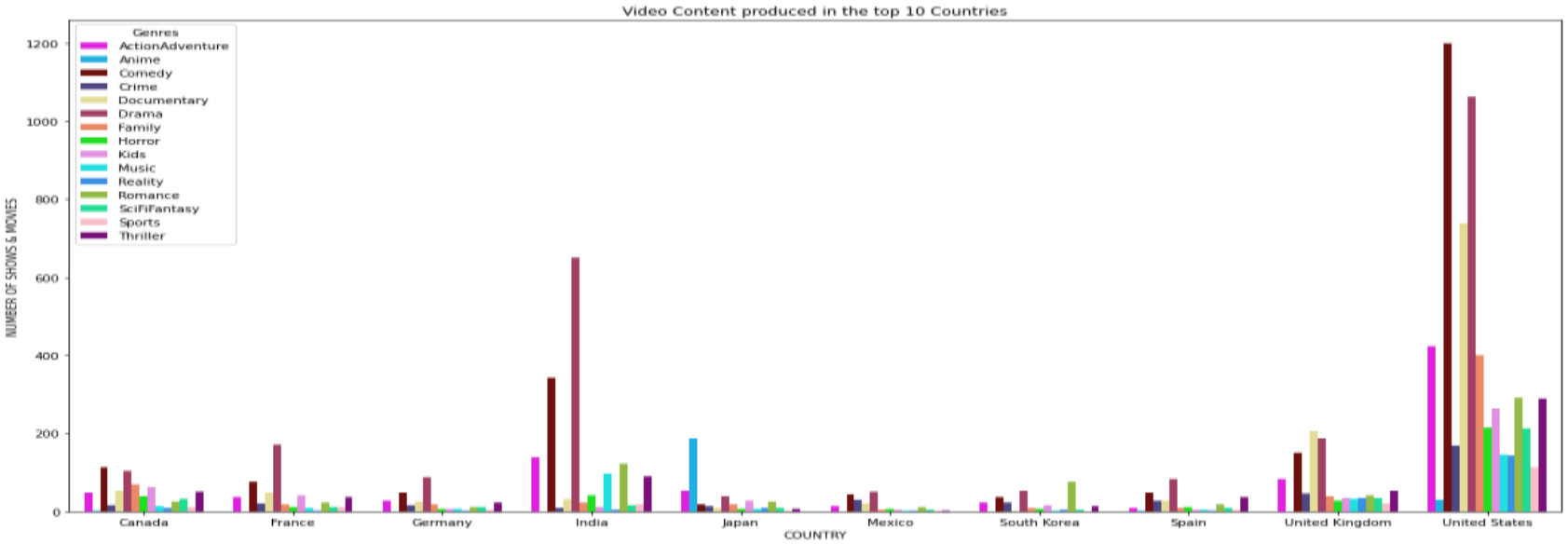
- Anupam Kher is the one who appeared in most of the movies and TV Shows.
- Also, its good to see that 6 of the actors in the top ten list of most numbers tv shows and movies are from India.

Content available in different countries



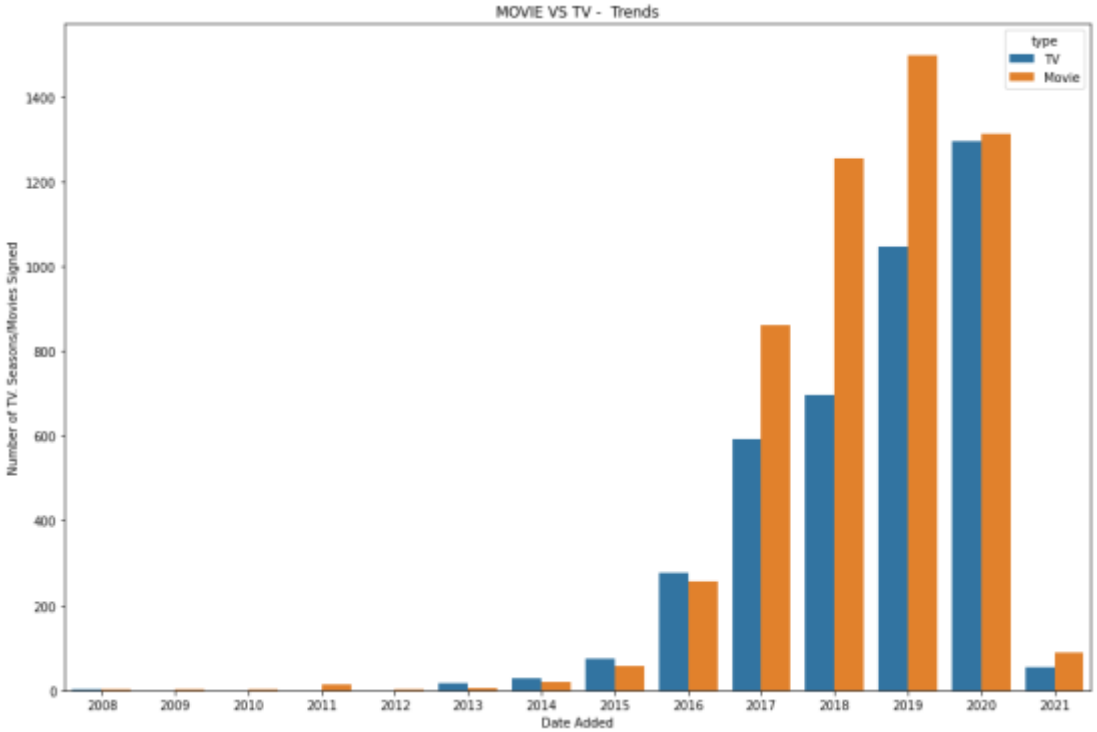
- The United States is a leading producer of both types of content which is obvious as Netflix is US Based company.
- It is followed by India where most of the content is in the form of movies.

Content produced in top 10 countries



- Drama is the most produced genre in a lot of Non-English speaking countries
- Comedy is the most produced genre in English speaking countries like United States of America and United Kingdom and Canada
- Japan is the biggest producer of Anime. Anime is also the most produced in genre in Japan
- Most South Korean content are from the Romance genre
- Documentaries are mainly produced in United Kingdom and United States of America
- Drama and Comedy are the most produced genres in the top countries with exceptions of Japan and South Korea

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



- We can observe that TV shows signed have been higher than movies in 2016.
- While the no of movies signed were higher, it can be seen that the TV shows signed per year is catching up with the movies signed year by year.

Data pre-processing for cluster formation

Methods used for Data pre-processing

- **Word tokenization**: It is the process of splitting a large sample of text into words. This is a requirement in natural language processing tasks where each word needs to be captured and subjected to further analysis like classifying and counting them for a particular sentiment etc.

- **Punctuation Removal**: The punctuation removal process will help to treat each text equally. For example, the word data and data! are treated equally after the process of removal of punctuations.

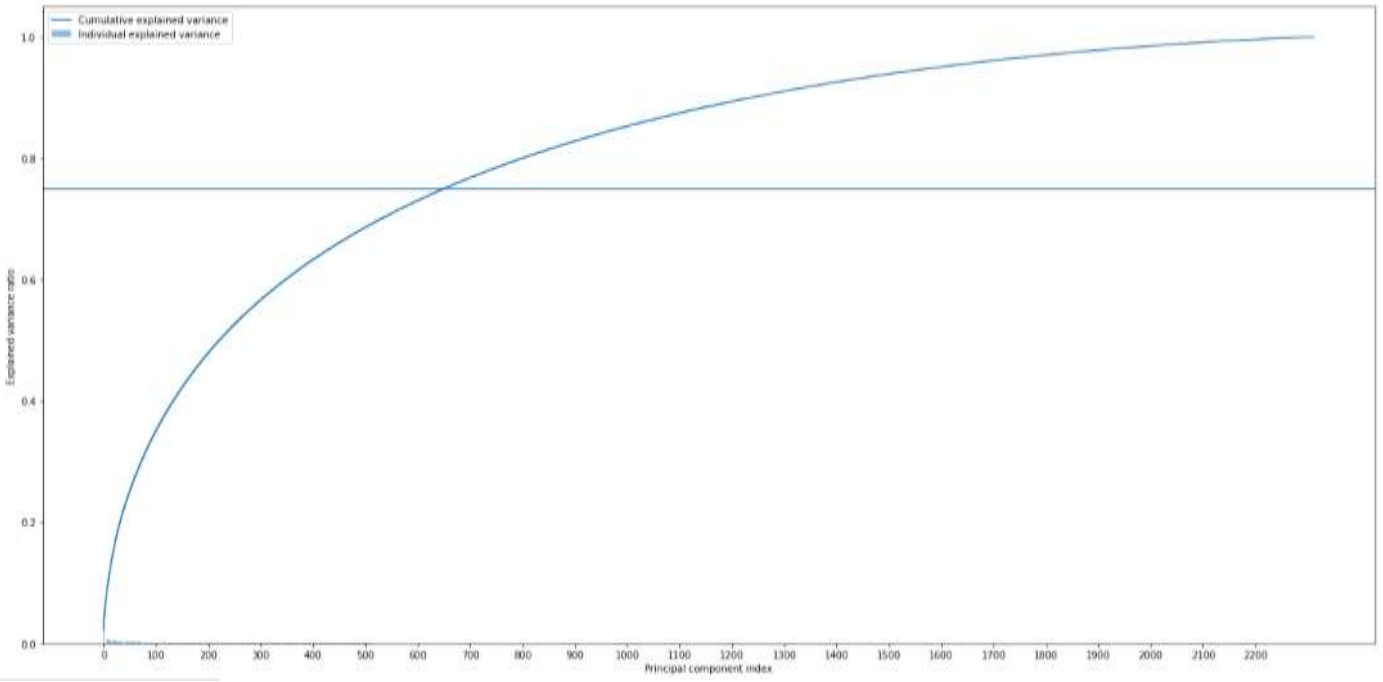
- **Stopwards Removal**: Stopwords are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For example, the words like the, he, have etc.

- **Stemming:** It is the process of reducing a word to its stem that affixes to suffixes and prefixes or to the roots of words known as "lemmas".

- **Text Vectorization:** It is the process of converting text into numerical format.

Dimensionality reduction using PCA

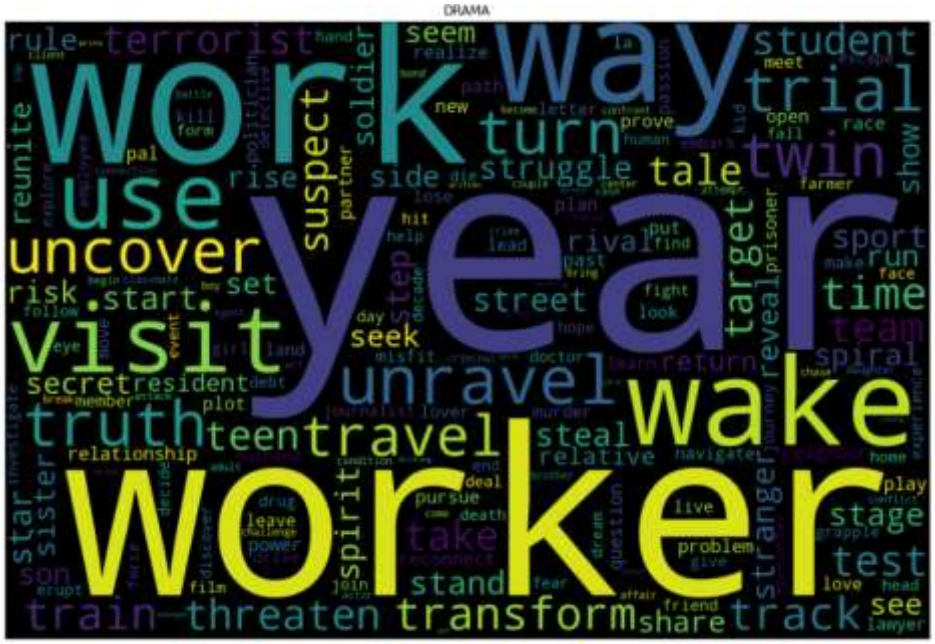
Dimensionality reduction using PCA



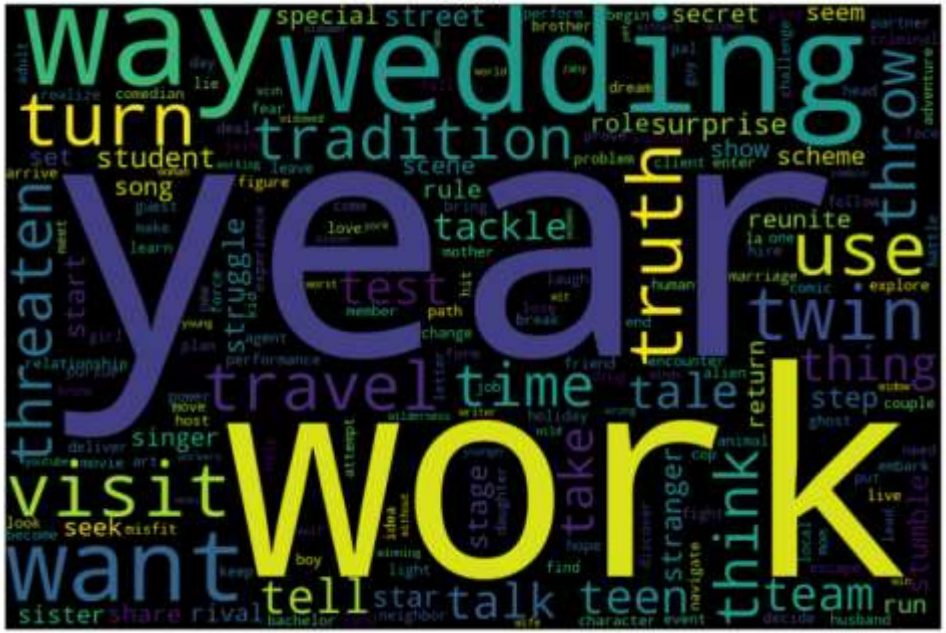
- Vectorising the preprocessed attributes Movie Deets and Description has sum total of 2121 + 192 = 2313 dimensions. These dimensions will have to be reduced using PCA, which would result in loss of information. Alternatively, the two attributes can be used to model the content into topics using Latent Dirichlet Allocation. This would makes sure that all the topical information about video content are captured without putting any available information to waste.

Bag Of Words for most commonly used Genres

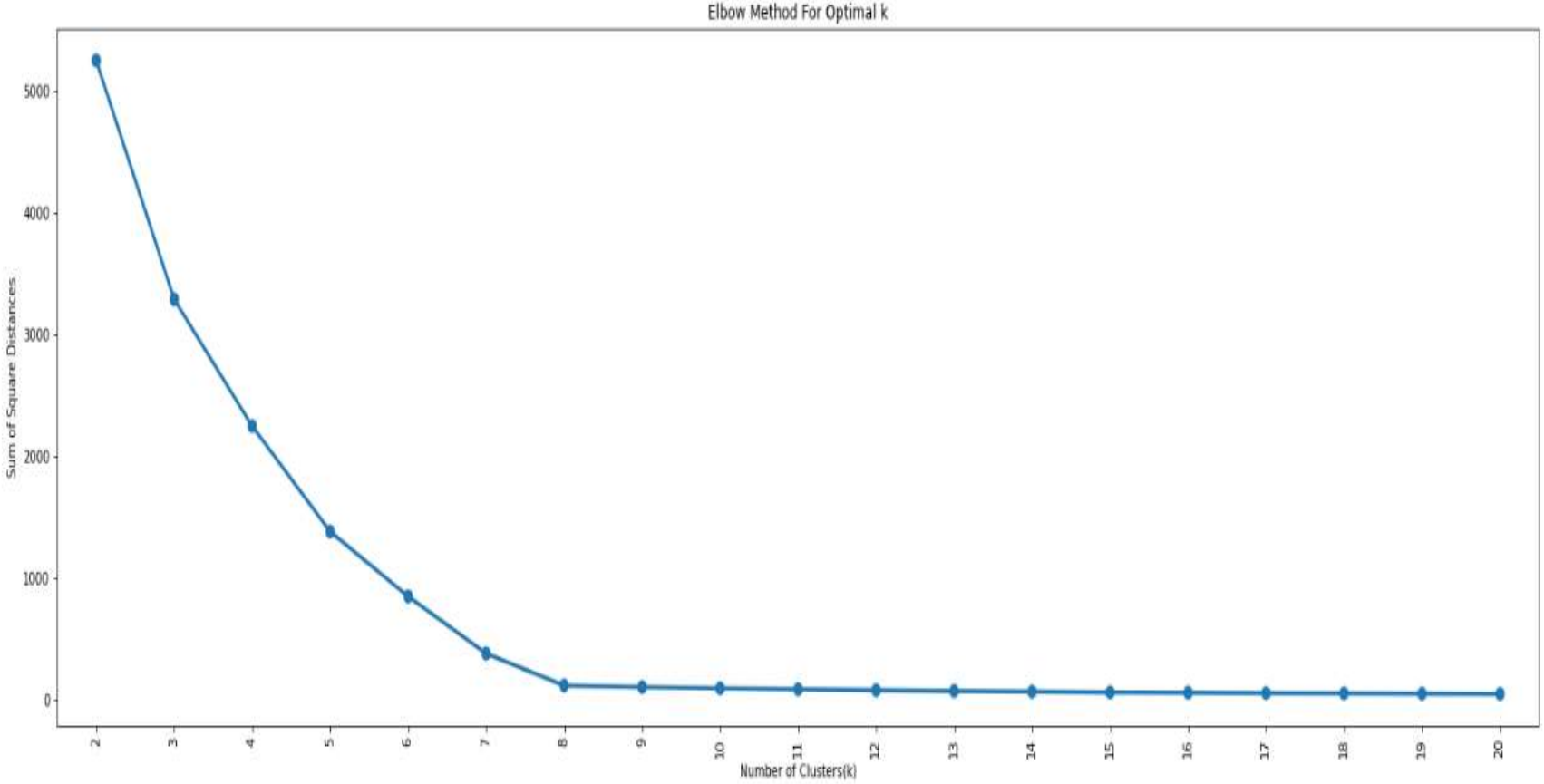
Drama



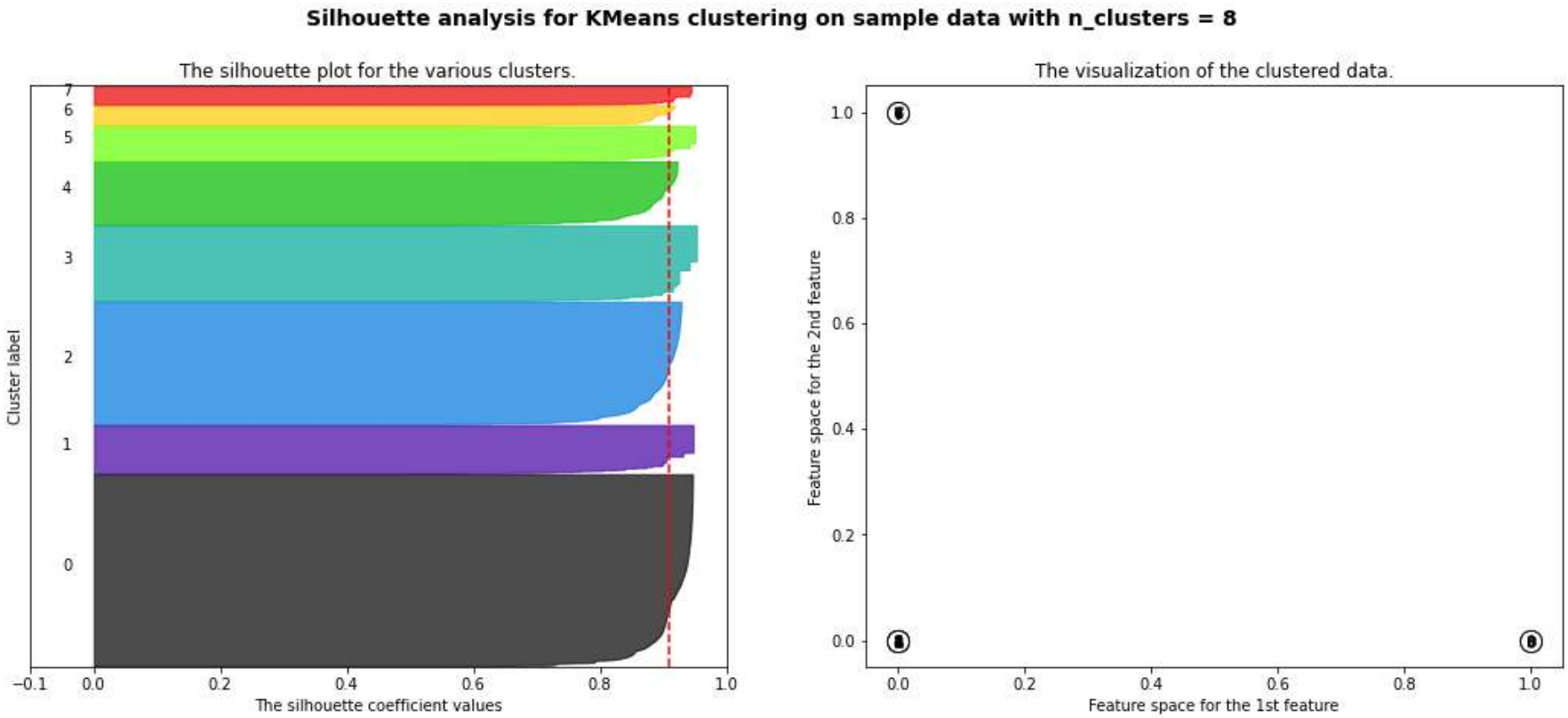
Comedy



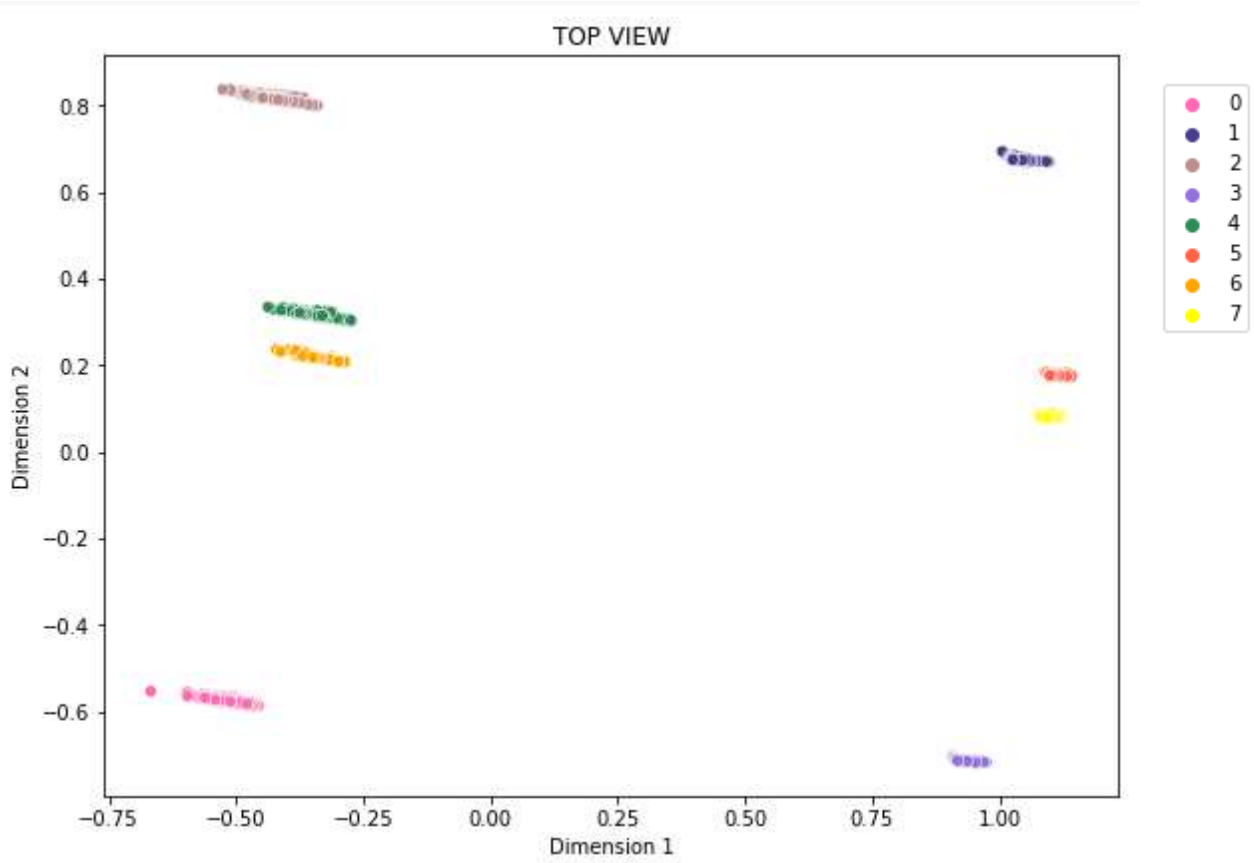
KMeans Clustering



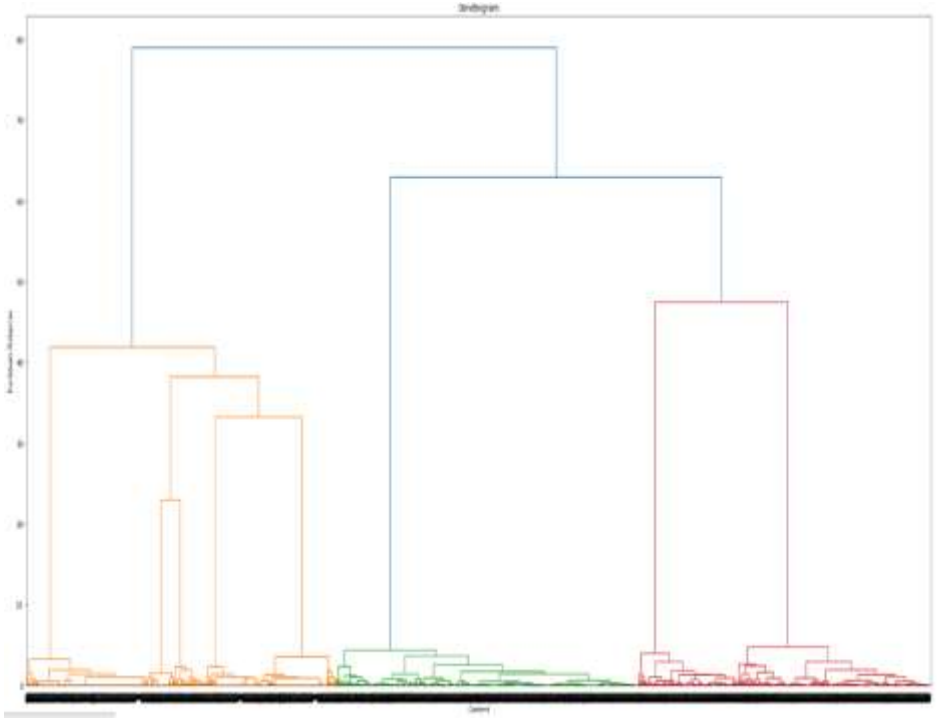
KMeans Clustering

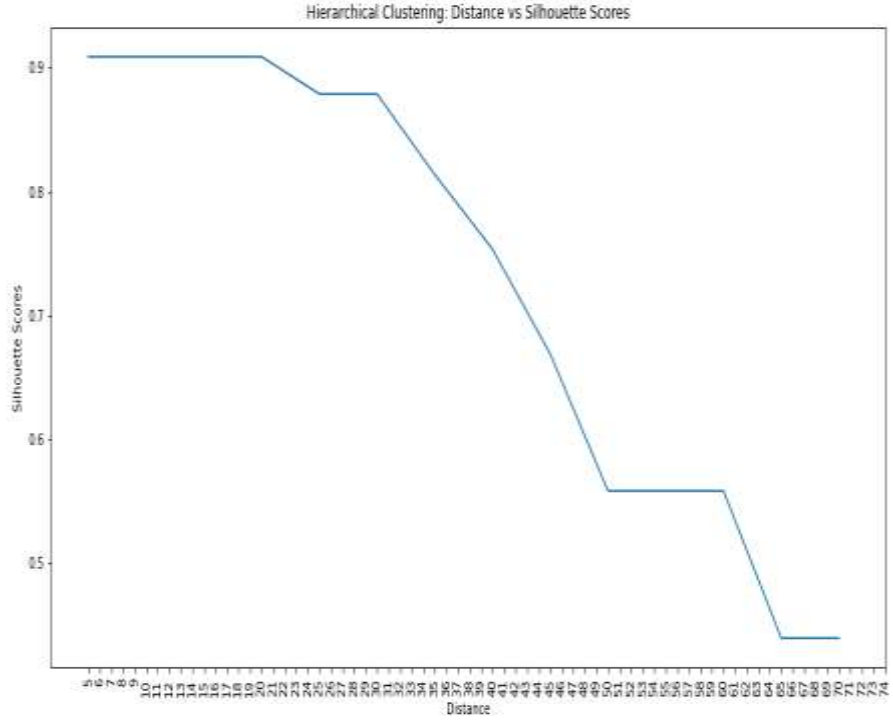


KMeans Clustering



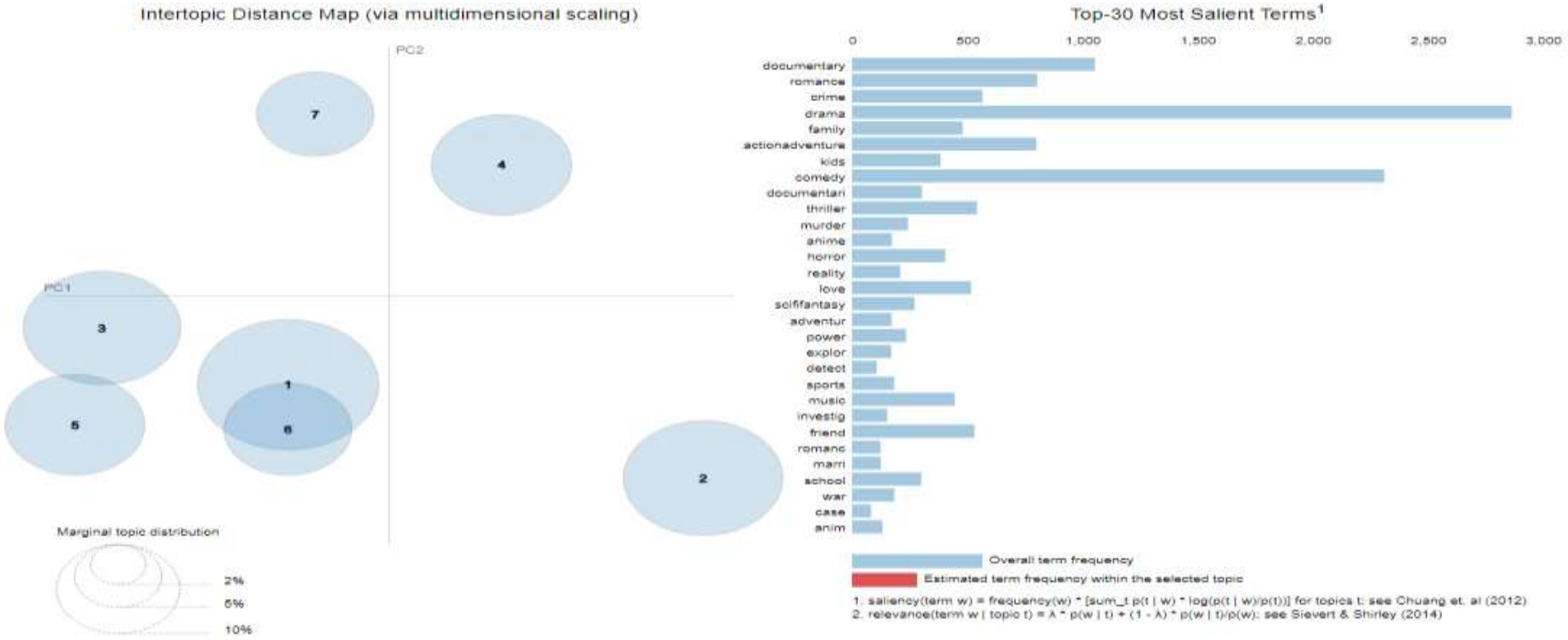
Hierarchical Clustering





- Here, we are getting the highest silhouette score of 0.90 at the distance 20 which corresponds to 8 clusters.

Topic Modelling Using LDA



- Seven topics are found to be the most suitable upon comparing coherence scores for different topic numbers.

Top words of different topics

Conclusion

1. Exploratory Data Analysis

- Majority of content available on Netflix is Movies.
- Growth in the number of movies on Netflix is much higher than tv shows. The highest number of movies and tv shows got added in 2019 and 2020
- October, November, December, and January are months in which many shows and movies get uploaded to the platform.
- Most of the content gets uploaded in the beginning and the middle of the month.
- United States and India are the top countries that produce all of the available content on the platform.
- TV-MA tops the charts, indicating that mature content is more popular on Netflix
- Top Genres on Netflix are found to be : Drama, Comedy, Documentary, Action and Adventure, Romance etc.

Conclusion

2. Analysis of Content produced in different countries

- The United States is a leading producer of both types of content which is obvious as Netflix is US Based company. It is followed by India where most of the content is in the form of movies
- Drama is the most produced genre in a lot of Non-English speaking countries
- Comedy is the most produced genre in English speaking countries like United States of America, United Kingdom and Canada
- Drama and Comedy are the most produced genres in the top countries with exceptions of Japan and South Korea
- Japan is the biggest producer of Anime. Anime is also the most produced in genre in Japan

Conclusion

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

- We have observed that TV shows signed have been higher than movies in 2016.
- While the no of movies signed were higher, it can be seen that the TV shows signed per year is catching up with the movies signed year by year.

4. Clustering

- k=8 is found to be an optimal value for clusters with highest silhouette score of 0.909 using which we grouped our data into 8 distinct clusters.
- Using dendrograms and comparing various distance thresholds, a distance of 20 produced the highest silhouette score of 0.908 with 8 clusters

Conclusion

5. Topic Modelling

- Latent Dirichlet Allocation is used to model textual data (description, genre, directors and cast) into topics.
- Seven topics are found to be the most suitable by comparing coherence scores for different topic numbers.
- The Topics and corresponding top words are given down below:

Topics	Corresponding Top Words
Topic 1	Drama > Comedy > Romance > Love > Life
Topic 2	Drama > Action & Adventure > Thriller > Horror > Young
Topic 3	Documentary > Music > World > Reality > Life
Topic 4	Crime > Drama > Murder > Thriller > Comedy

Thank You