Capstone Project-4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

PRINCE KUMAR JHA

Content:

- Introduction
- Problem Statement
- Data Description
- Null Value
- Exploratory Data Analysis
- Data Cleaning
- Topic modelling
- Model Implementation
- Data Pre-processing
- Model Implementation
- K- Means
- Cluster Analysis

Introduction

Netflix:

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time.

Methodology:

- Unsupervised Machine Learning (Clustering)

Database:

- Netflix Movies and TV Shows
- 7787 rows and 12 columns
- Data from last decade

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 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.

In this project, you are required to do

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

Data Description

- **The data was collected from Flixable which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 rows of data.**

- **The dataset consists of eleven textual columns and one numeric column.**

Attribute Information :

1. **show_id :** Unique ID for every Movie / Tv Show
2. **type :** Identifier - A Movie or TV Show
3. **title :** Title of the Movie / Tv Show
4. **director :** Director of the Movie

Data Description

5. **cast :** Actors involved in the movie / show
6. **country :** Country where the movie / show was produced
7. **date_added :** Date it was added on Netflix
8. **release_year :** Actual Release year of the movie / show
9. **rating :** TV Rating of the movie / show
10. **duration :** Total Duration - in minutes or number of seasons
11. **listed_in :** Genre
12. **description:** The Summary description

Null Value

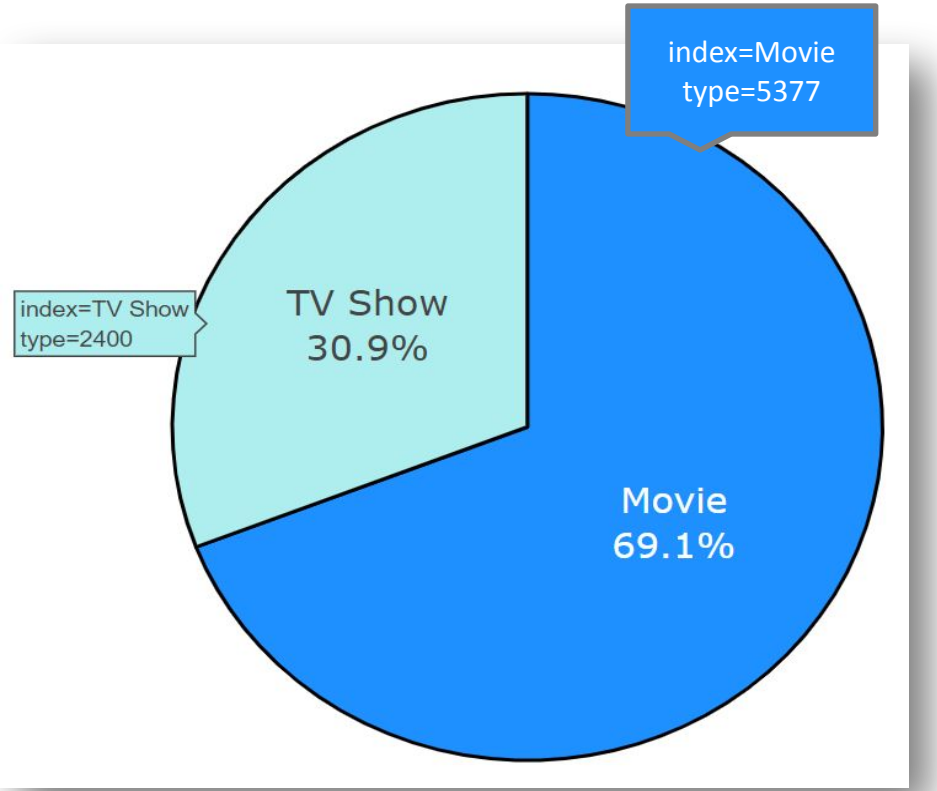
Null Value Treatment:

- ***Director*** feature have more than ***30.68%*** of null values. Filling null values by 'unknown'.
- ***Country*** feature have ***6.51%*** of null values. Filling null values by mode of feature.
- ***Cast feature*** have ***9.22%*** of null values. Filling null values by 'unknown'.
- ***Rating*** feature have ***0.09%*** of null values. Filling null values by mode of feature.
- ***Date_added*** feature have ***0.13%*** of null values. Dropping rows corresponding to null values.

Exploratory Data Analysis

Type of content available on Netflix

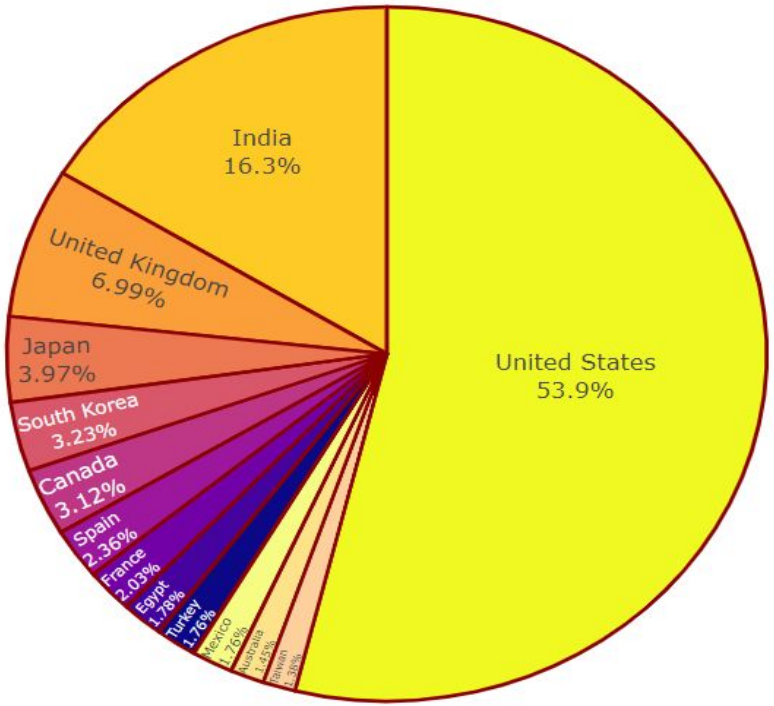
- It is evident that there are more movies on Netflix than TV shows.
- Netflix has 5377 movies, which is more than double the quantity of TV shows.



Exploratory Data Analysis

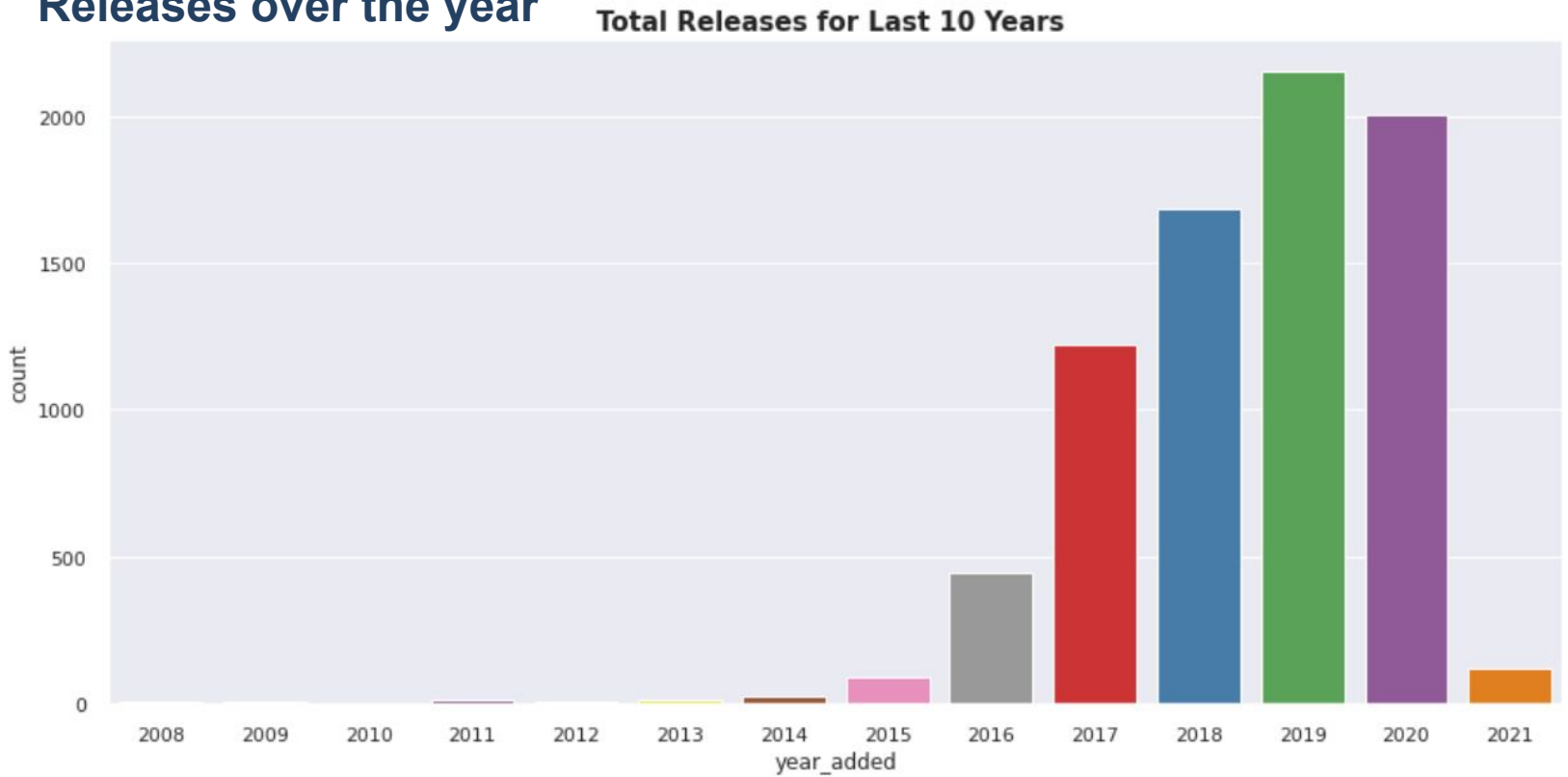
Top countries with highest content production

- United States has the most number of content on Netflix
- India has second highest content on Netflix
- Australia and Taiwan has least number of content on Netflix



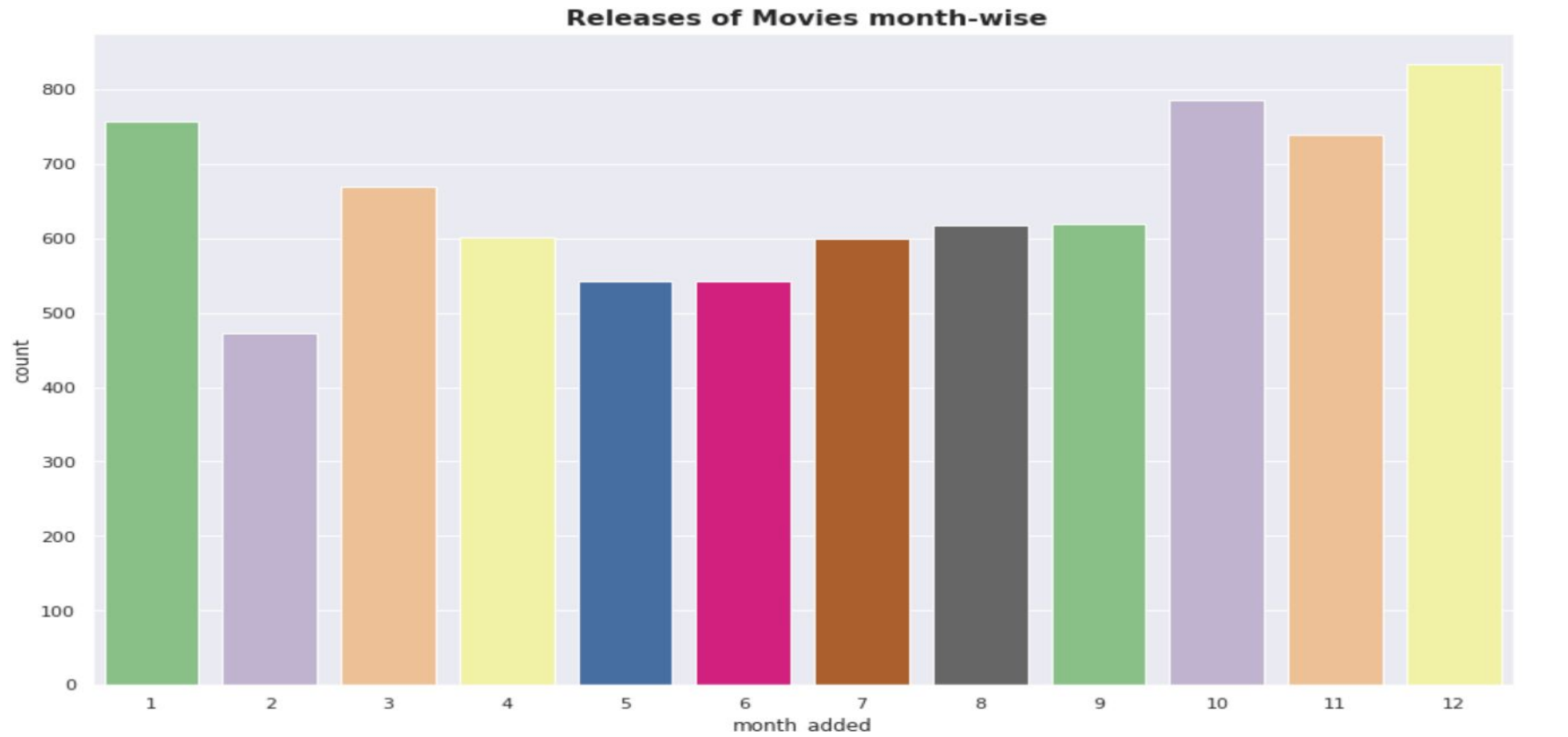
Exploratory Data Analysis

Releases over the year



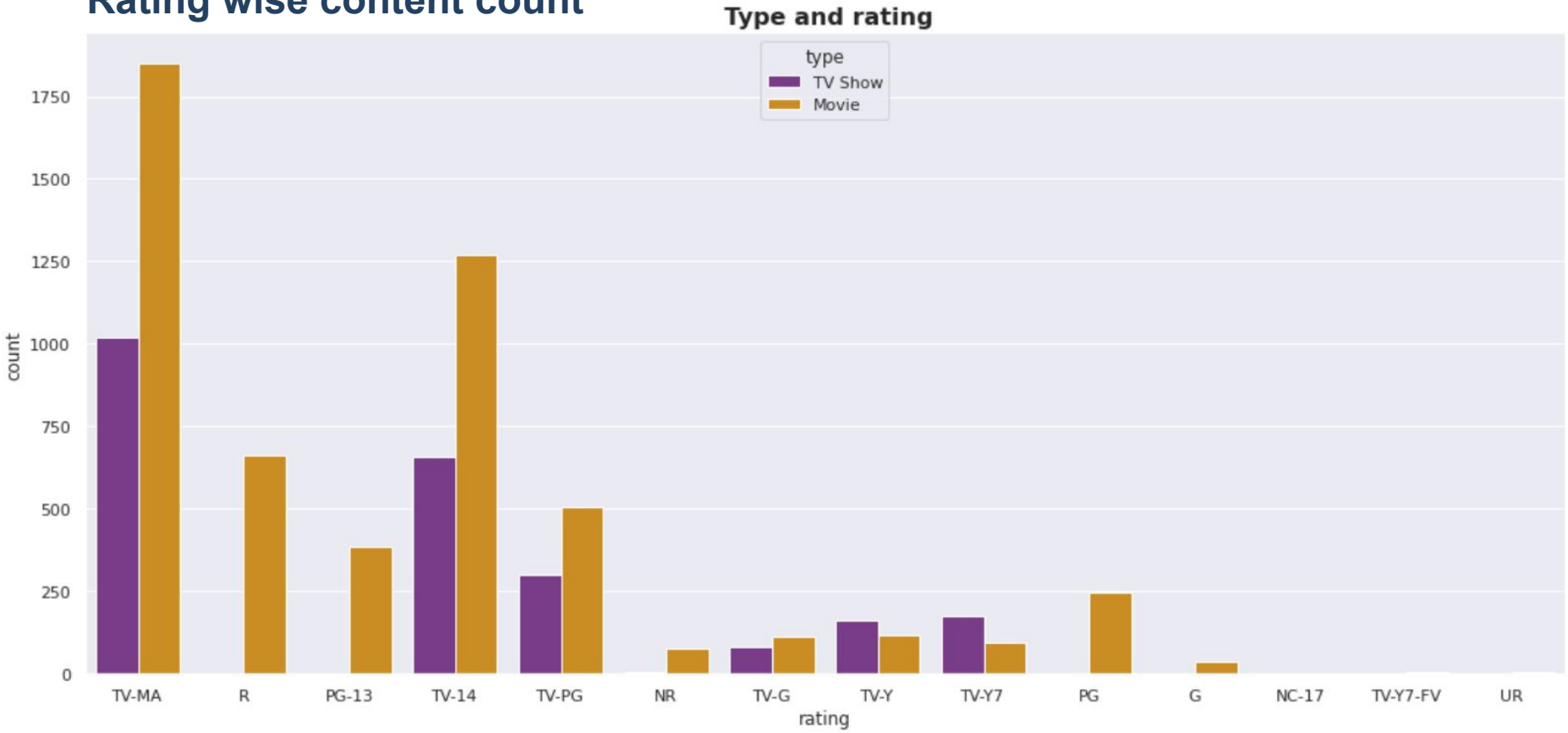
The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19

Exploratory Data Analysis

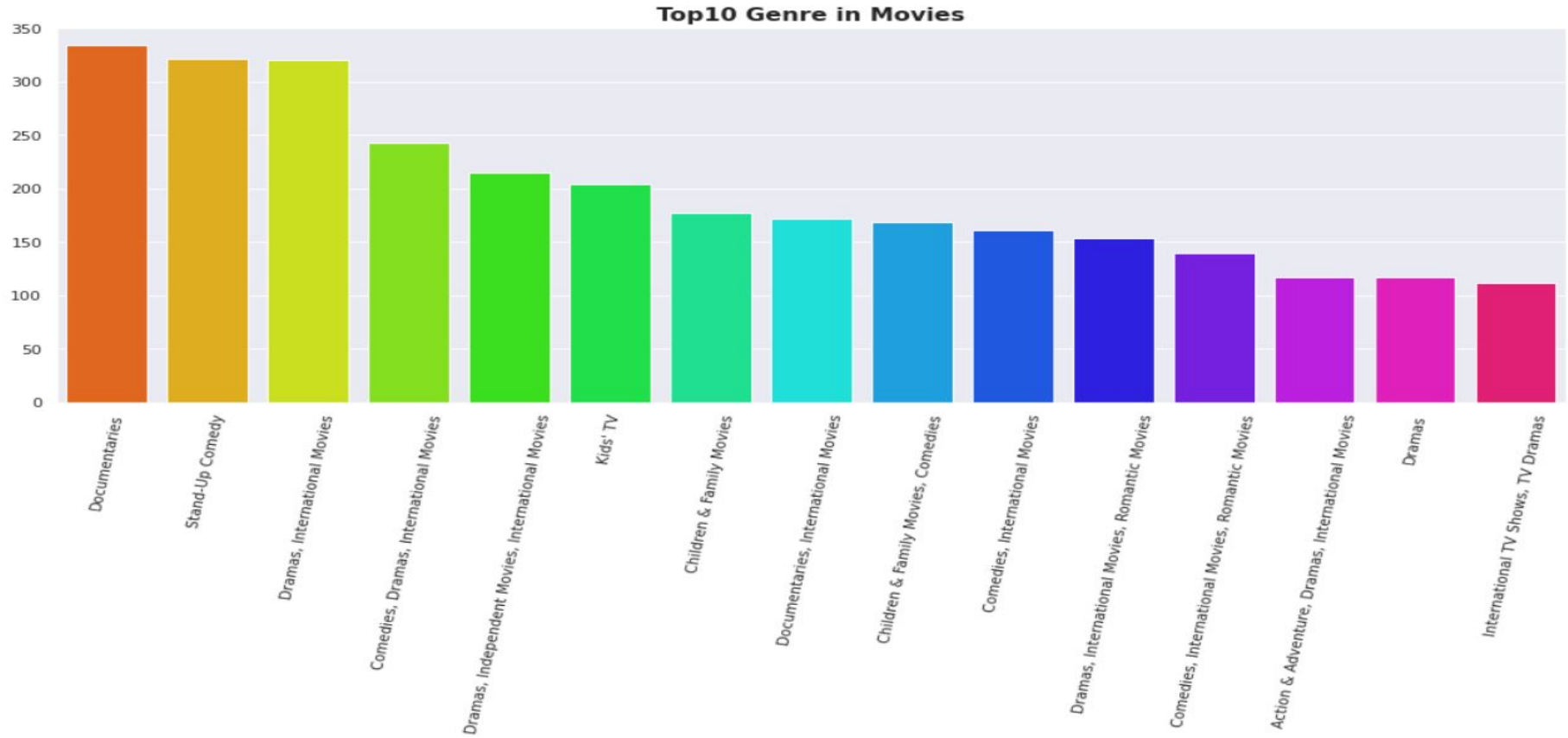


Exploratory Data Analysis

Rating wise content count

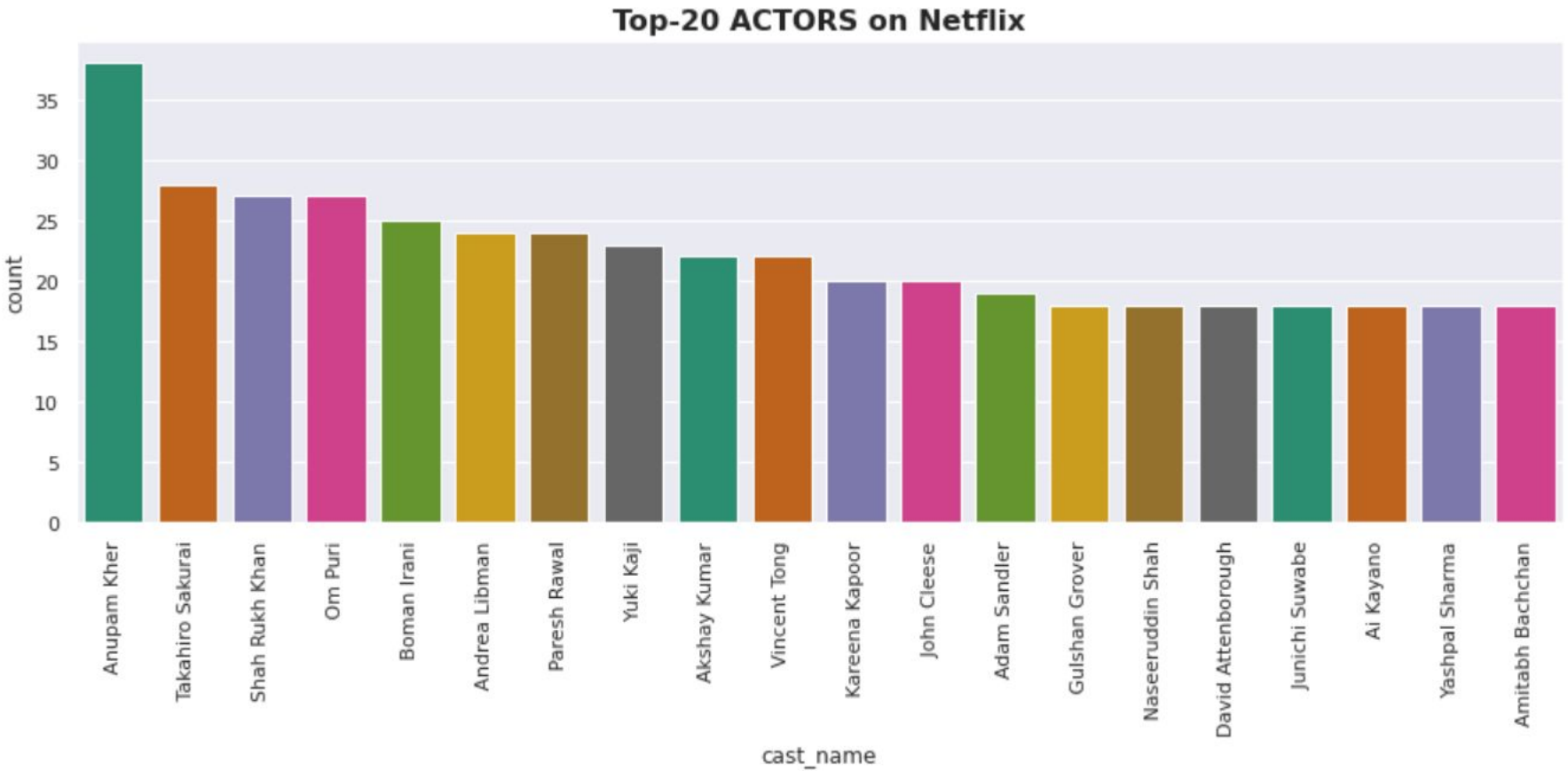


Exploratory Data Analysis



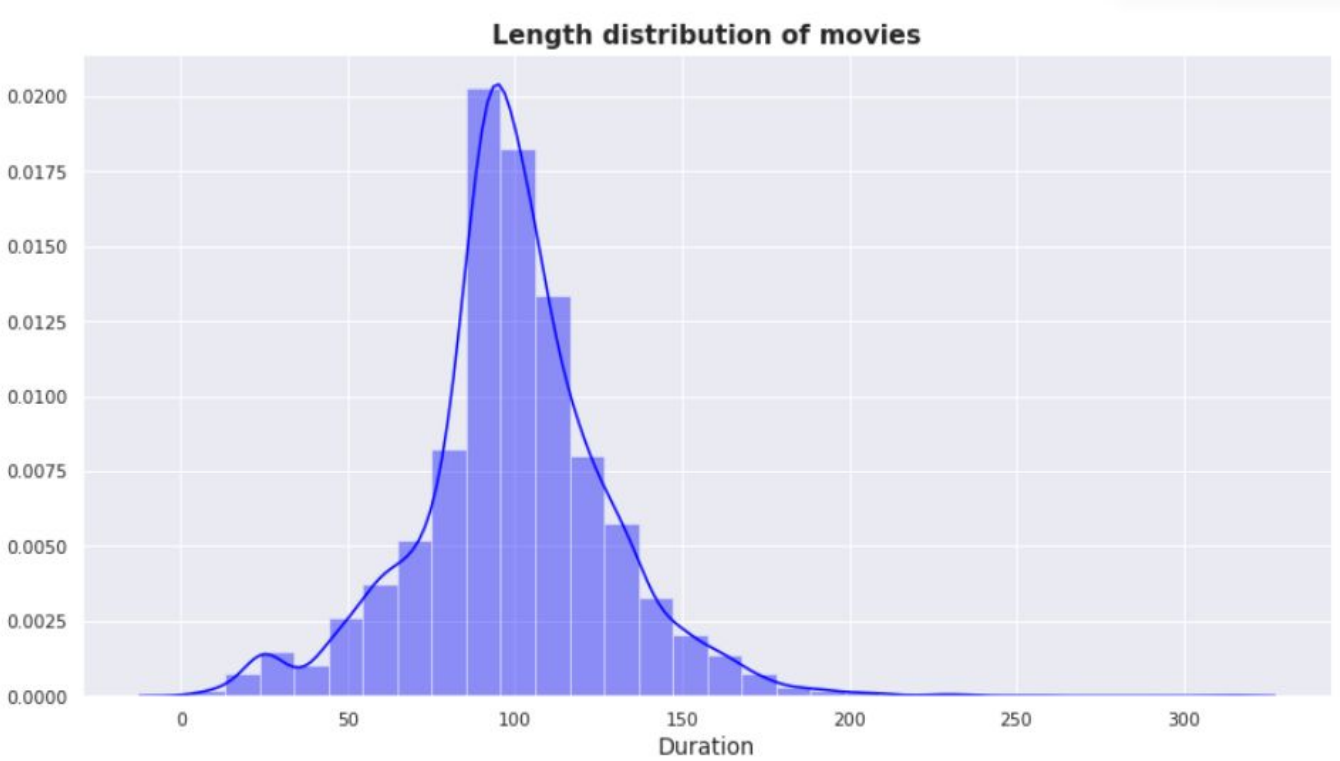
Documentaries is the most popular genre followed by comedy

Exploratory Data Analysis



Exploratory Data Analysis

Duration distribution of Movies



Data Cleaning

- Label Encoding
- Lemmatization- Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. ... For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.
- Removing Stop words - To remove stop words from a sentence, you can divide your text into words and then remove the word if it exits in the list of stop words provided by NLTK.
- Tf - idf Vectorization - TF-IDF stands for “Term Frequency — Inverse Document Frequency”. This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.
- Min-max Scaling - For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It preserves shape of original distribution.

Topic Modelling (LDA and LSA)

- **Latent Semantic Analysis**(LSA) is used to find the hidden topics represented by the document or text. This hidden topics then are used for clustering the similar documents together. LSA is an unsupervised algorithm and hence we don't know the actual topic of the document.
- In natural language processing, the **Latent Dirichlet Allocatio**n (LDA) is a generative statistical model that allows sets of observations to be explained

by unobserved groups that explain why some parts of the data are similar.

Topic Modelling (LDA and LSA)

NETFLIX Genre 0:
international shows movies dramas comedies romantic family life independent young

NETFLIX Genre 1:
shows crime british spanish language korean docuseries series reality romantic

NETFLIX Genre 2:
adventure action fi sci kids stand family children world series

NETFLIX Genre 3:
stand special comedy comedian comic talk family take show life

NETFLIX Genre 4:
family children movies shows save friend christmas comedieswhen music kids

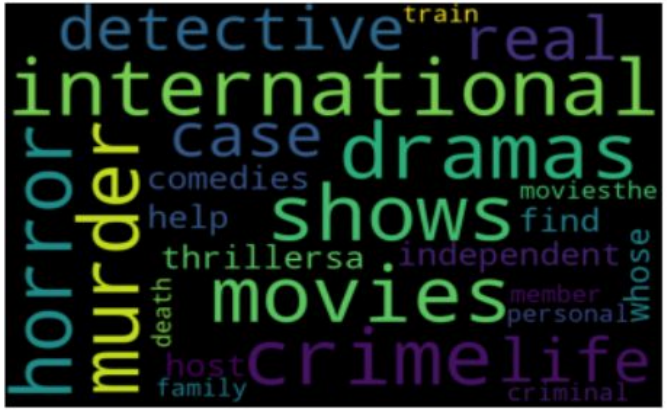
NETFLIX Genre 5:
documentaries documentary music world docuseries series moviesthis sports life international

NETFLIX Genre 6:
comedies romantic friend kids love life school best high series

NETFLIX Genre 7:
series kids independent docuseries dramas friend science nature anime world

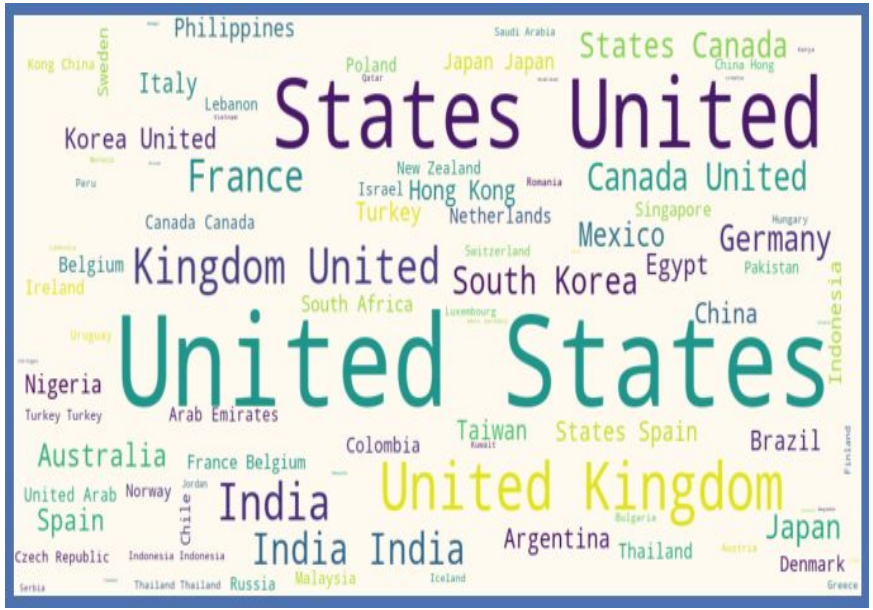
NETFLIX Genre 8:
horror movies fi sci romantic series reality kids thrillersa docuseries

NETFLIX Genre 9:
life docuseries young woman reality nature love science family romantic



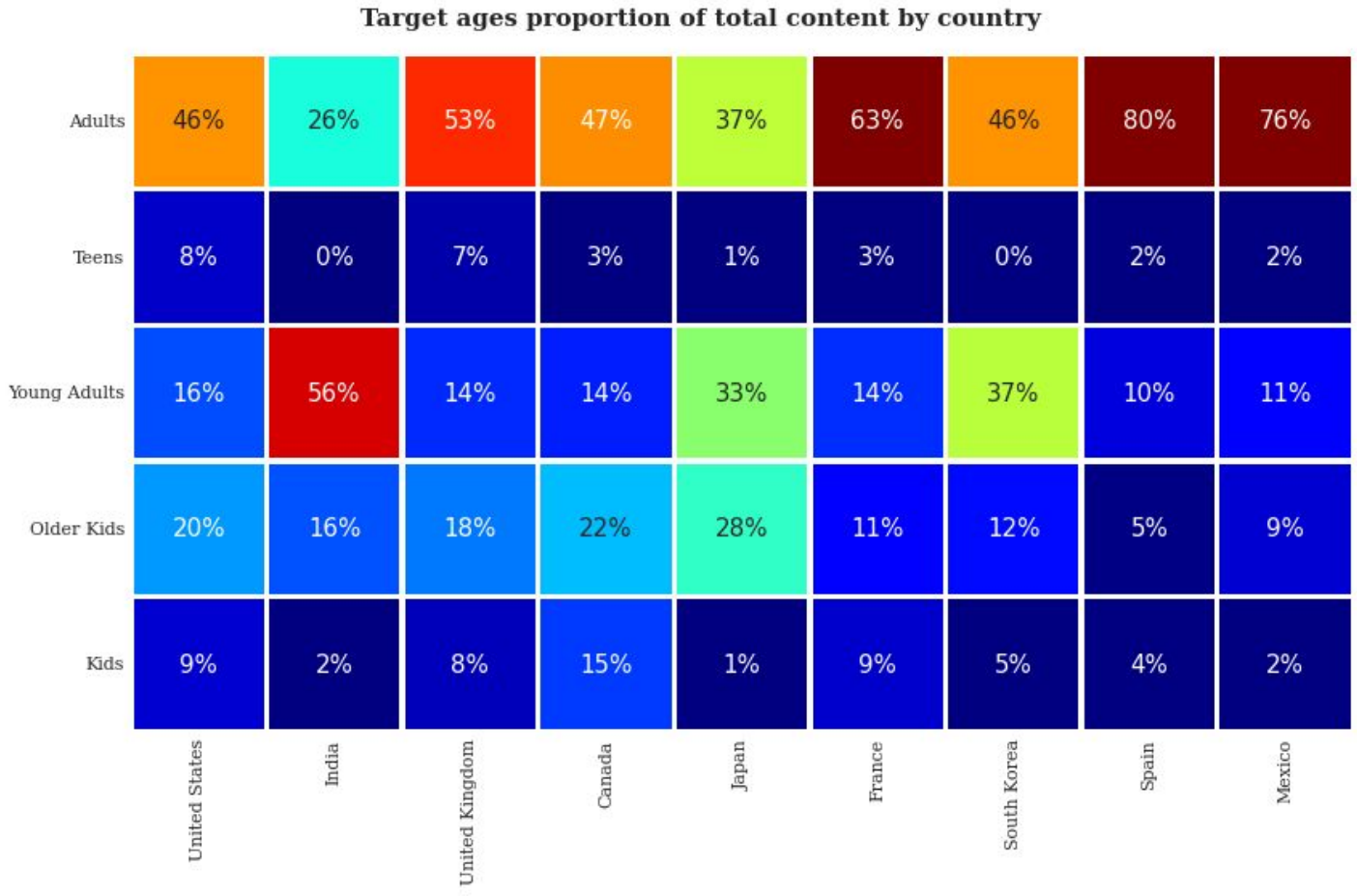
Word Cloud for Countries and Cast

Countries with the most content available



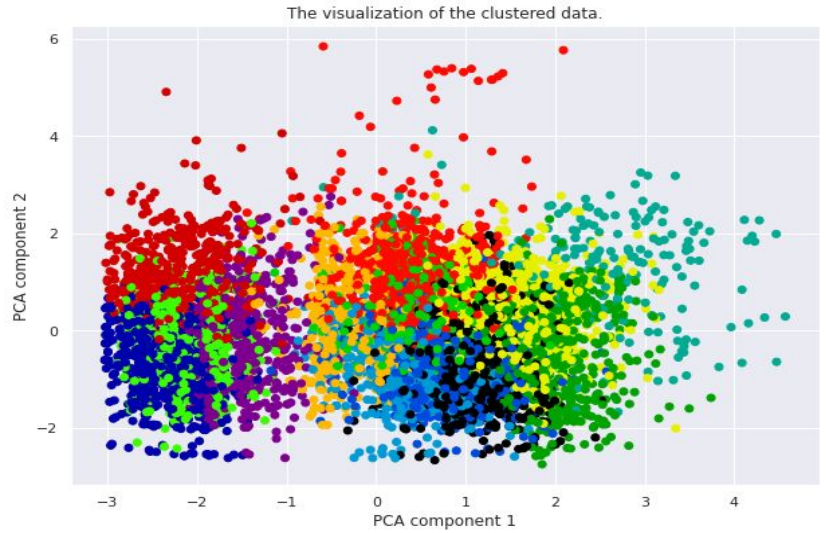
Which Cast to Choose?

Correlation Heatmap



Model Implementation

1. Affinity Propagation



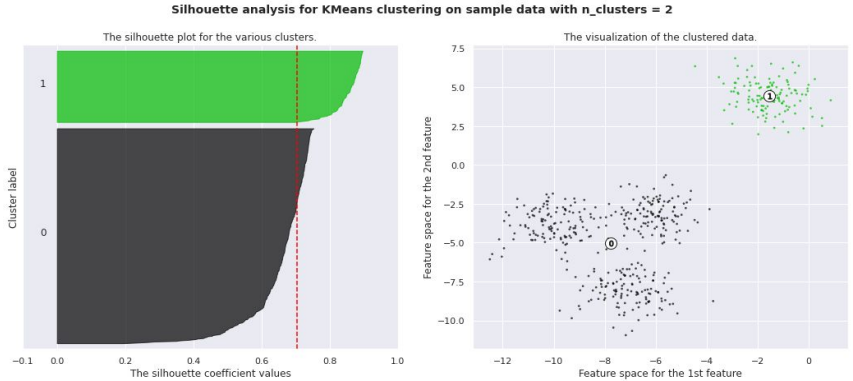
Converged after 81 iterations.
Estimated number of clusters: 13
Silhouette Coefficient: 0.244

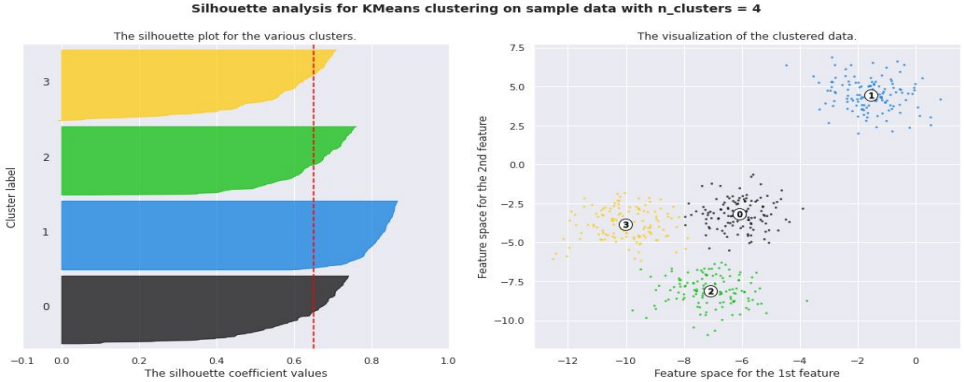
2. Agglomerative Propagation

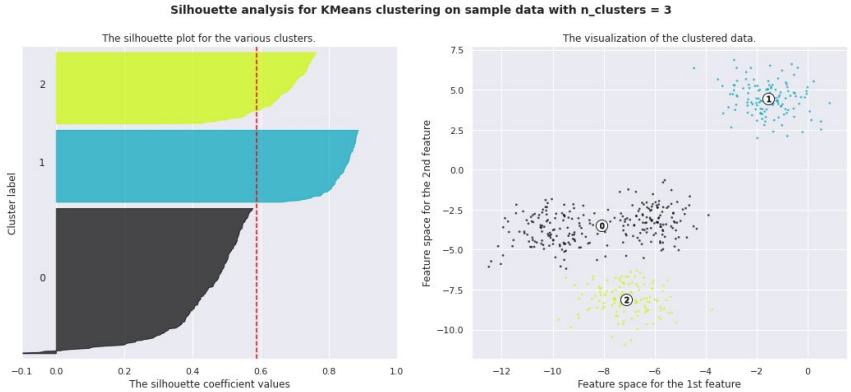
Assume we cut vertical lines with a horizontal line to obtain the number of clusters. Number of clusters = 4
The average silhouette_score is : 0.17296314851287742

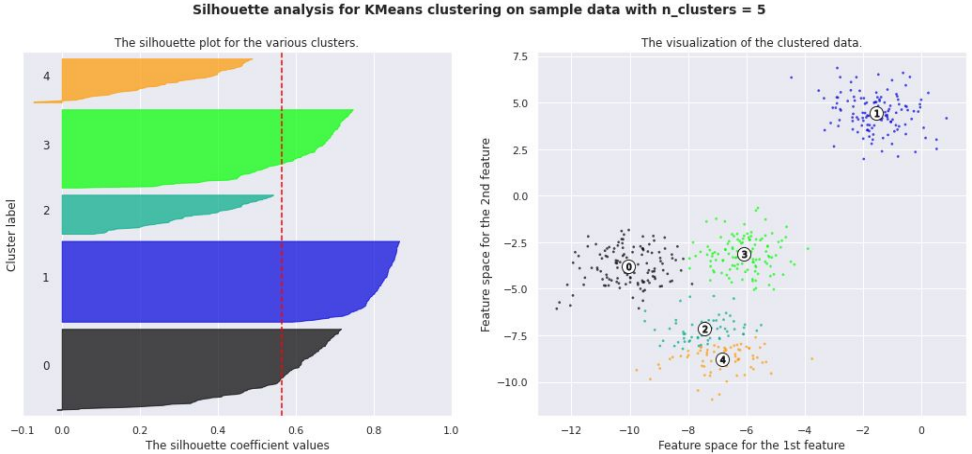
Model Implementation

3. k-means clustering



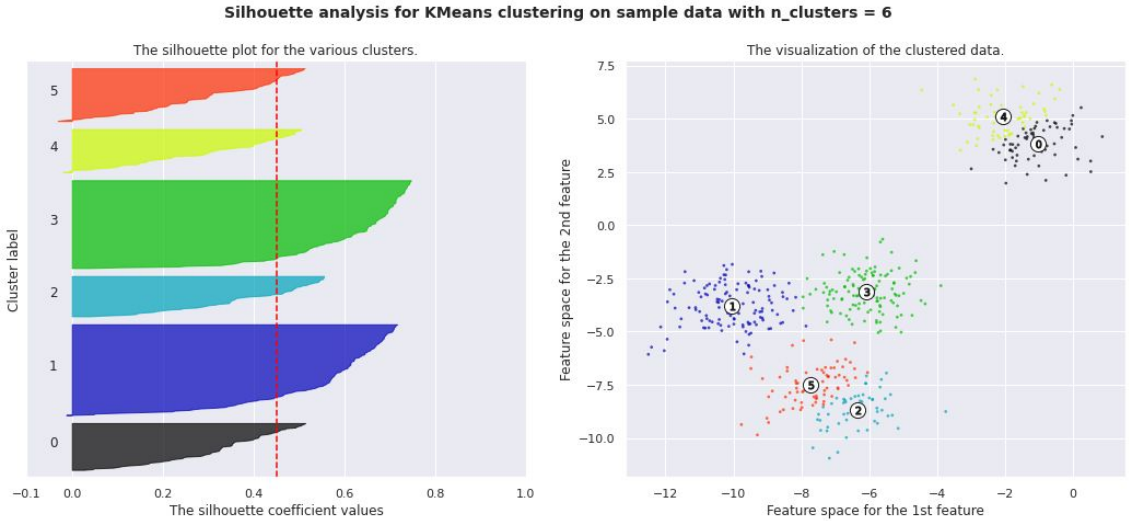






Model Implementation

3. k-means clustering



For n_clusters = 2 The average silhouette_score is : 0.7049787496083262

For n_clusters = 3 The average silhouette_score is : 0.5882004012129721

For n_clusters = 4 The average silhouette_score is : 0.6505186632729437

For n_clusters = 5 The average silhouette_score is : 0.56376469026194

For n_clusters = 6 The average silhouette_score is : 0.4504666294372765

- Here is the Silhouette analysis done on the above plots to select an optimal value for n_clusters.
- The value of 4 and 5 for n_clusters looks to be the optimal one. The silhouette score for each cluster is above average silhouette scores.

K - Means

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved

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.

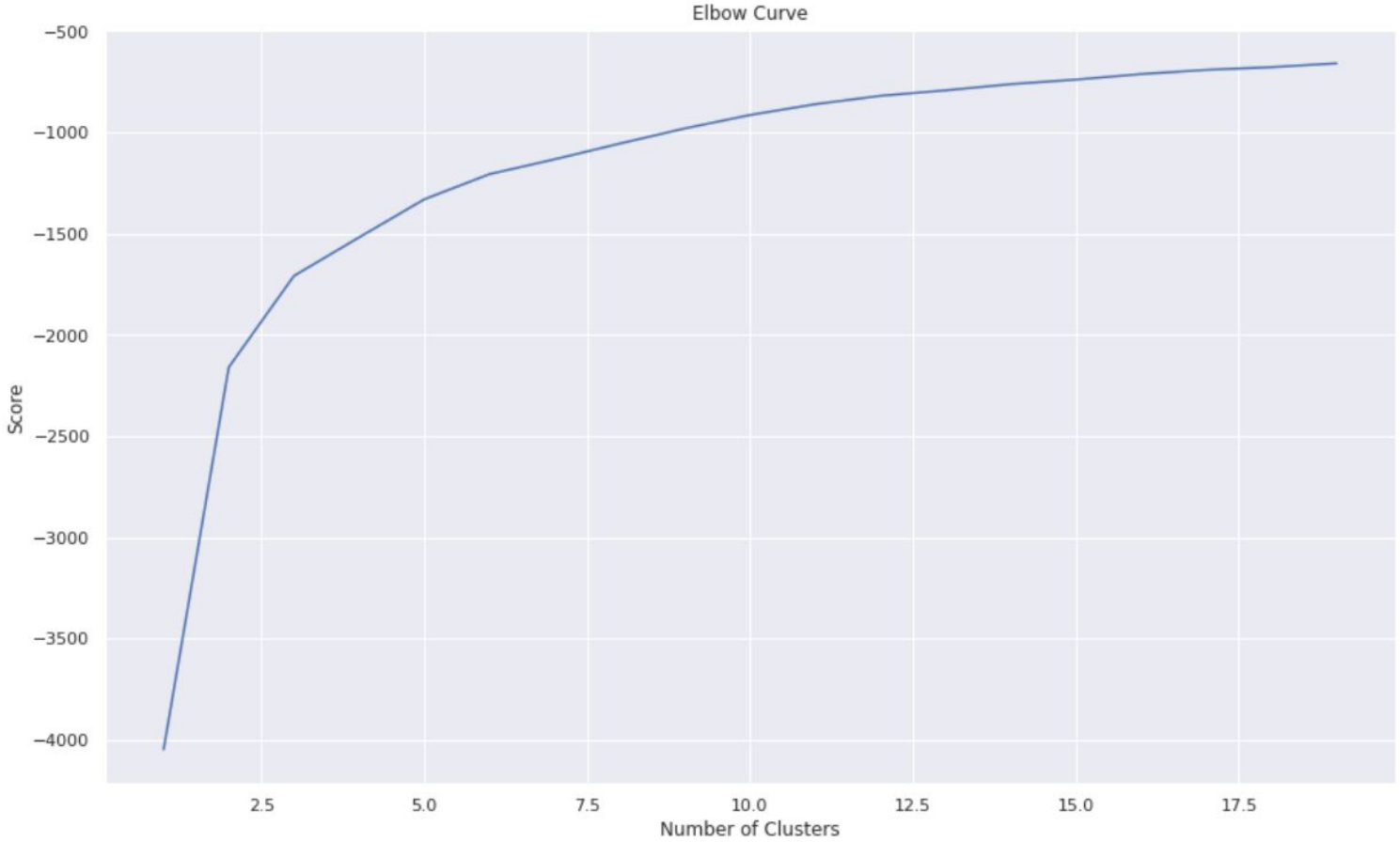
1. Elbow Curve:

- The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- 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.

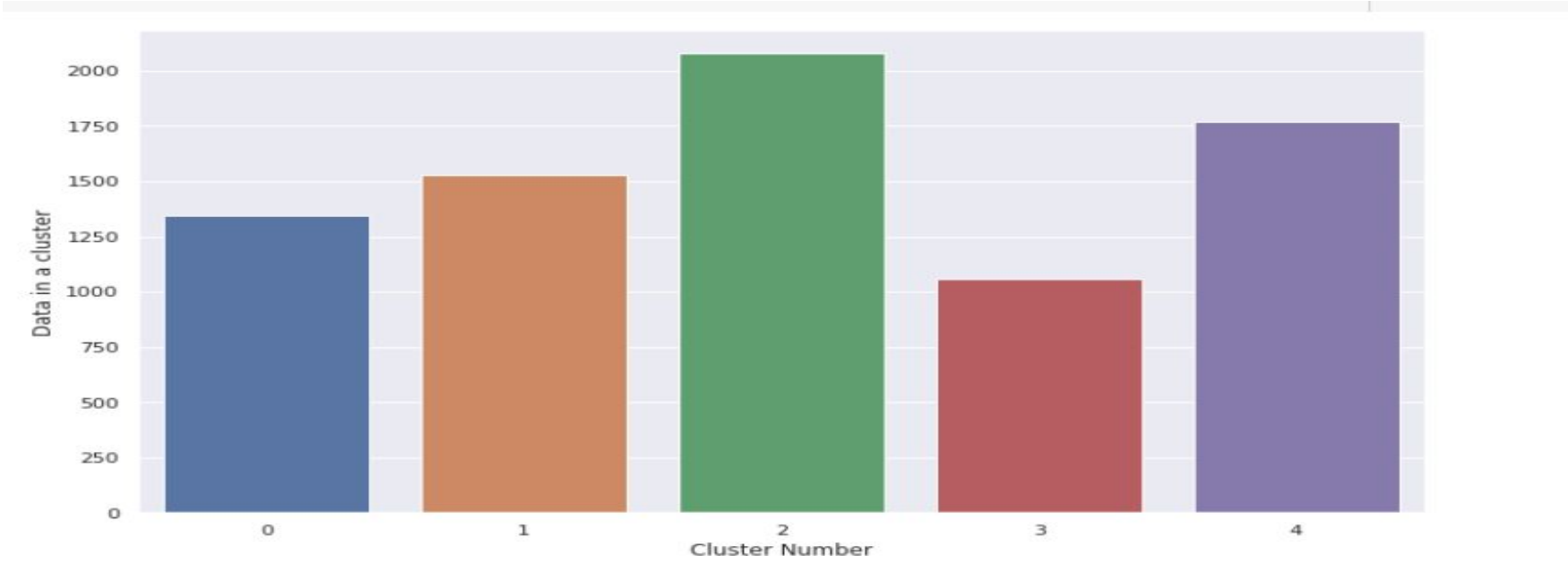
2. Silhouette score :

- 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 other.

K-Means


Elbow Curve
Score
−500
−1000
−1500
−2000
−2500
−3000
−3500
−4000
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Clusters

Clusters



We clearly see that one cluster is the largest and one cluster has the fewest number of movies.

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.
- We have two types of content TV shows and Movies (30.9% contains TV shows and 69.1% contains Movies).
- Most films were released in the years 2018, 2019, and 2020 and united states have the maximum content on Netflix.
- The months of October, November, December and January had the largest number of films and Tv-shows released.
- The USA, India, the United Kingdom, Canada, and Egypt are the top five producer countries.
- For the clustering algorithm, we utilized type, director, nation, released year, genre, and year.
- LDA and LSA has sorted much more similar titles in a group of genre.
- Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangements.
- In Affinity Propagation, we had 13 clusters and a Silhouette Coefficient score of 0.244.
- We cut vertical lines with a horizontal line to obtain the number of clusters in Agglomerative Clustering. There were four clusters, with an average silhouette score of 0.17296314851287742.
- The final model we used was k-means clustering, which consisted of 2,3,4,5,6 clusters. 4 numbers of clusters gives us good fitting.
- After applying K - means optimal value of number of clusters is 5
- Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are.

Thank you

Time for Q&A!!