Capstone Project

Netflix Movies And TV Shows Clustering

By

By- Baishnavee Mahato
Data Science Trainee
AlmaBetter, Bengaluru

Content

Introduction:

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies on an internet-connected device. Netflix content varies by region and may change over time. We can watch a variety of award-winning Netflix originals, TV shows, movies, documentaries, and more.

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

Problem Statement:

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

Project Scope:

The project is an **Unsupervised Machine Learning (Clustering)** of the Data set of Netflix Movies And TV Shows Clustering.csv

In this project, we 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 Summary

Data set name: Netflix Movies And TV Shows Clustering.csv

Shape :

- Rows: **7787**
- Columns: **12**

Columns: show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added','release_year', 'rating', 'duration', 'listed_in', 'description'.

Description of Dataset Columns:

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

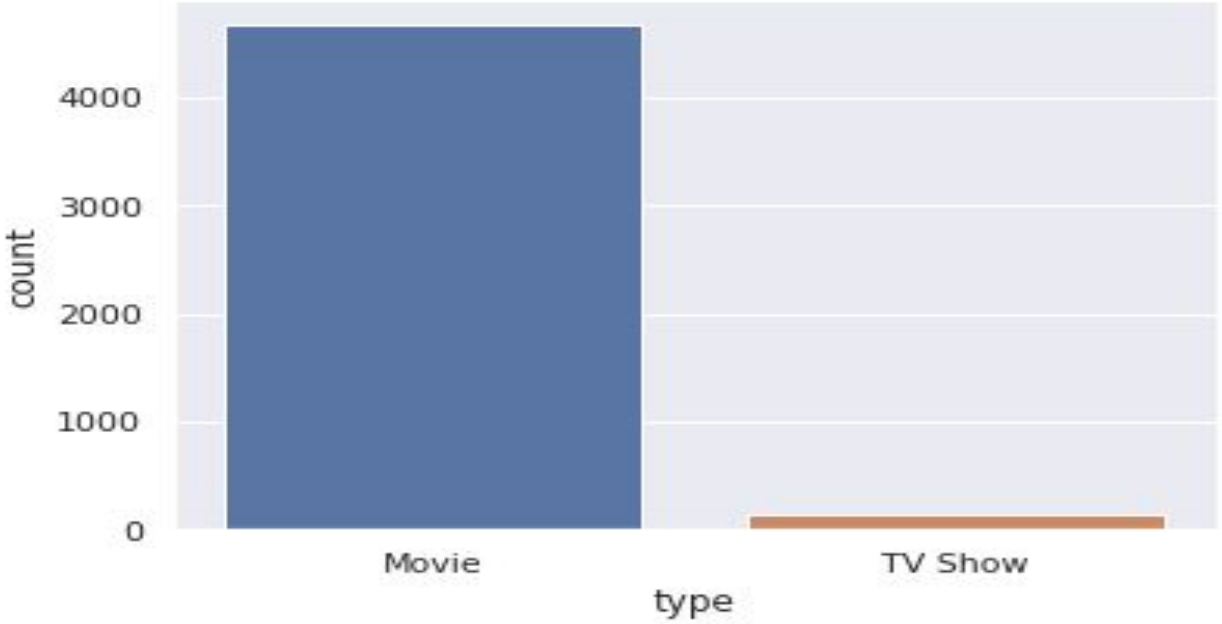
description: The Summary description

Exploratory

Data

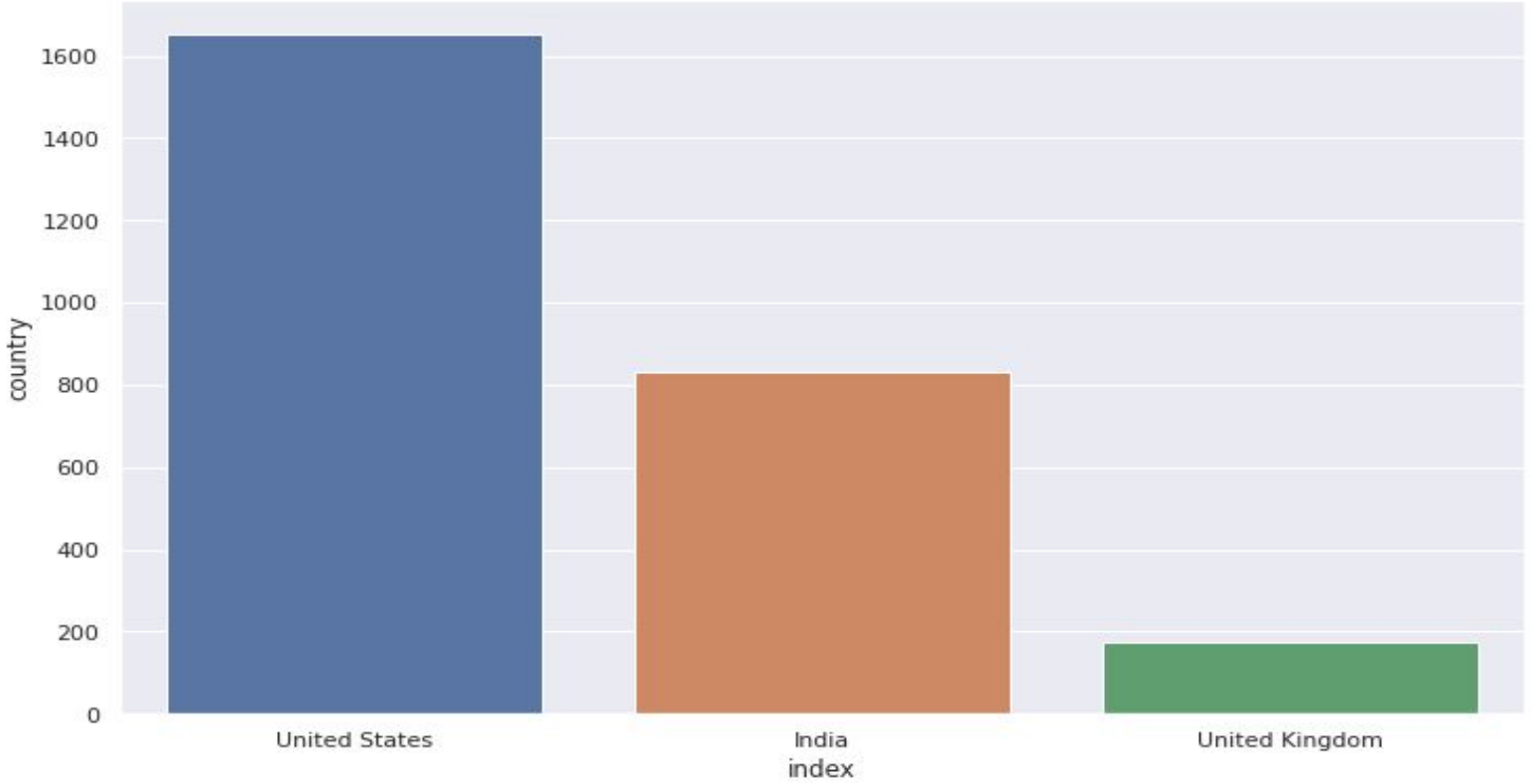
Analysis

No. Of TV Shows and Movies in Netflix

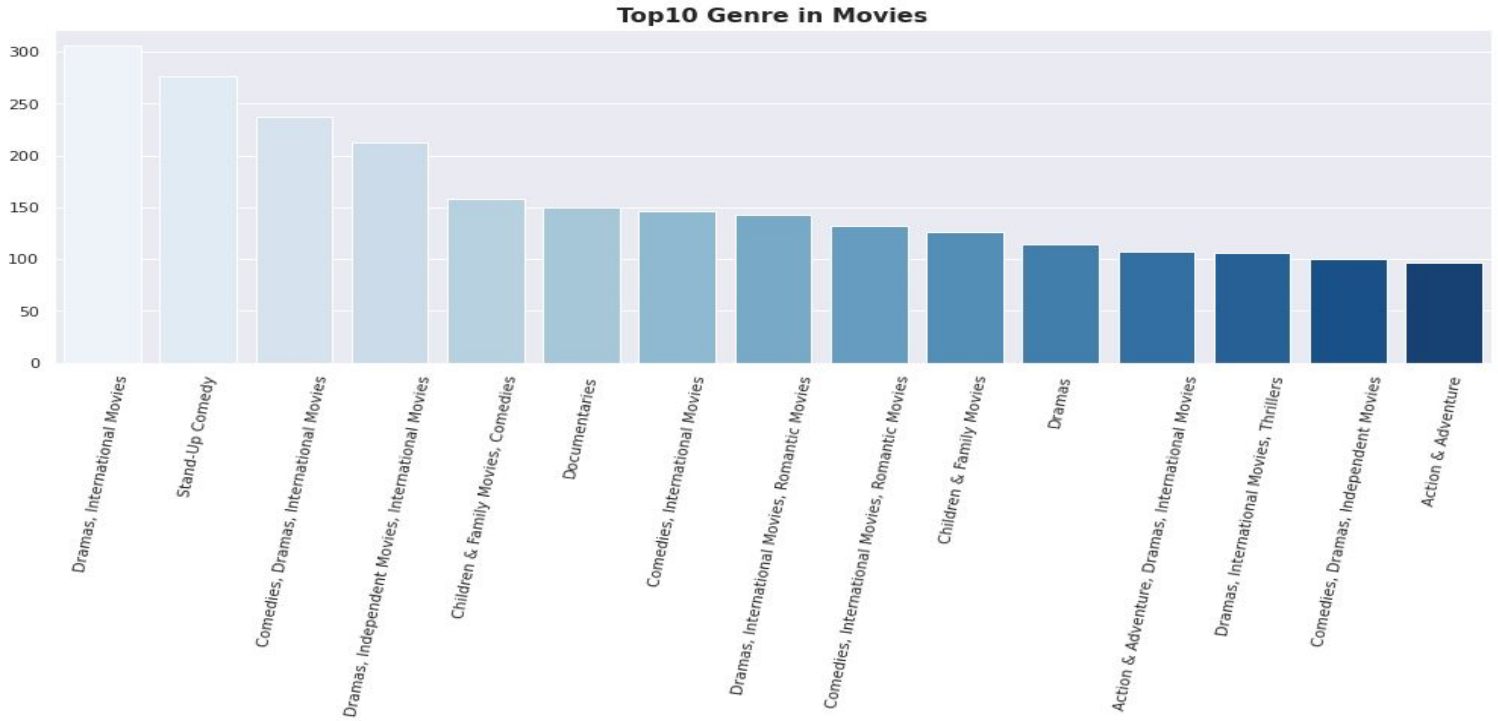


There are more number movies on Netflix than TV shows. Netflix has 4673 movies, which is more than double the quantity of TV shows.

Top three countries where Netflix is most popular?

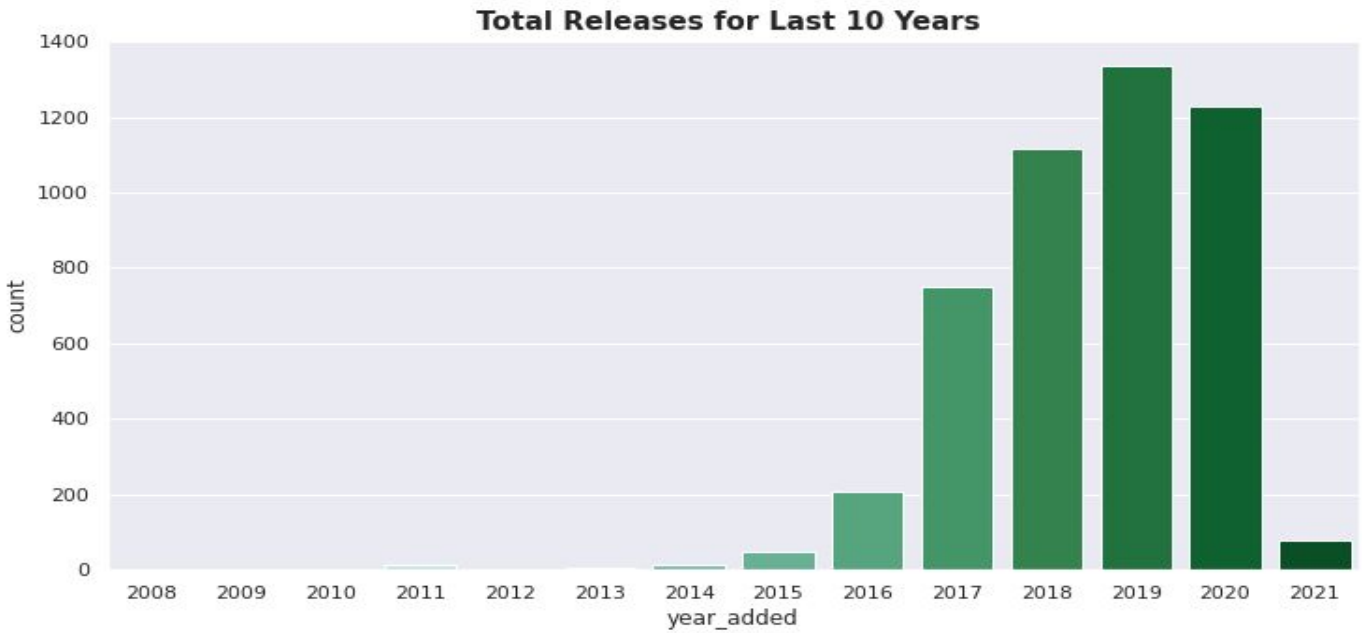


Top 10 Genre in movies



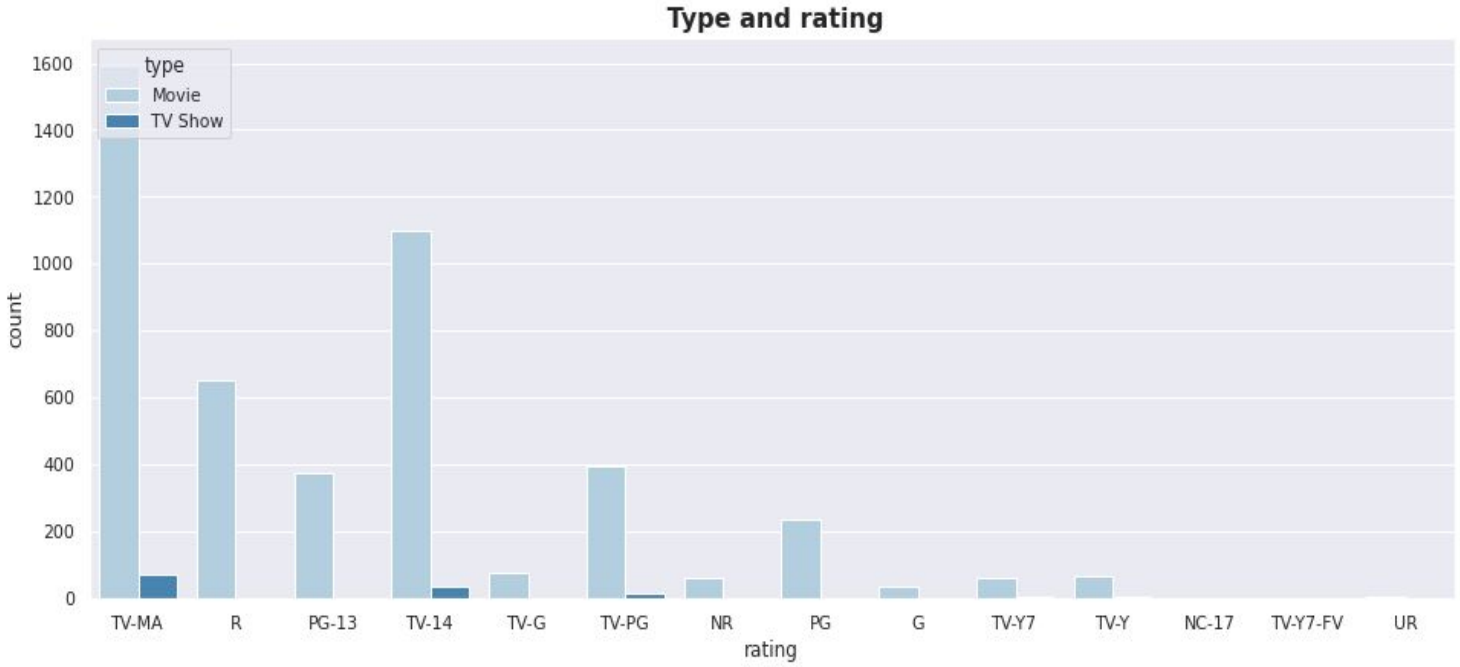
From the above graph we came to know that Documentaries take the first place, followed by Stand-up Comedy and Dramas,International Movies... and so on.

Total release for last 10 years



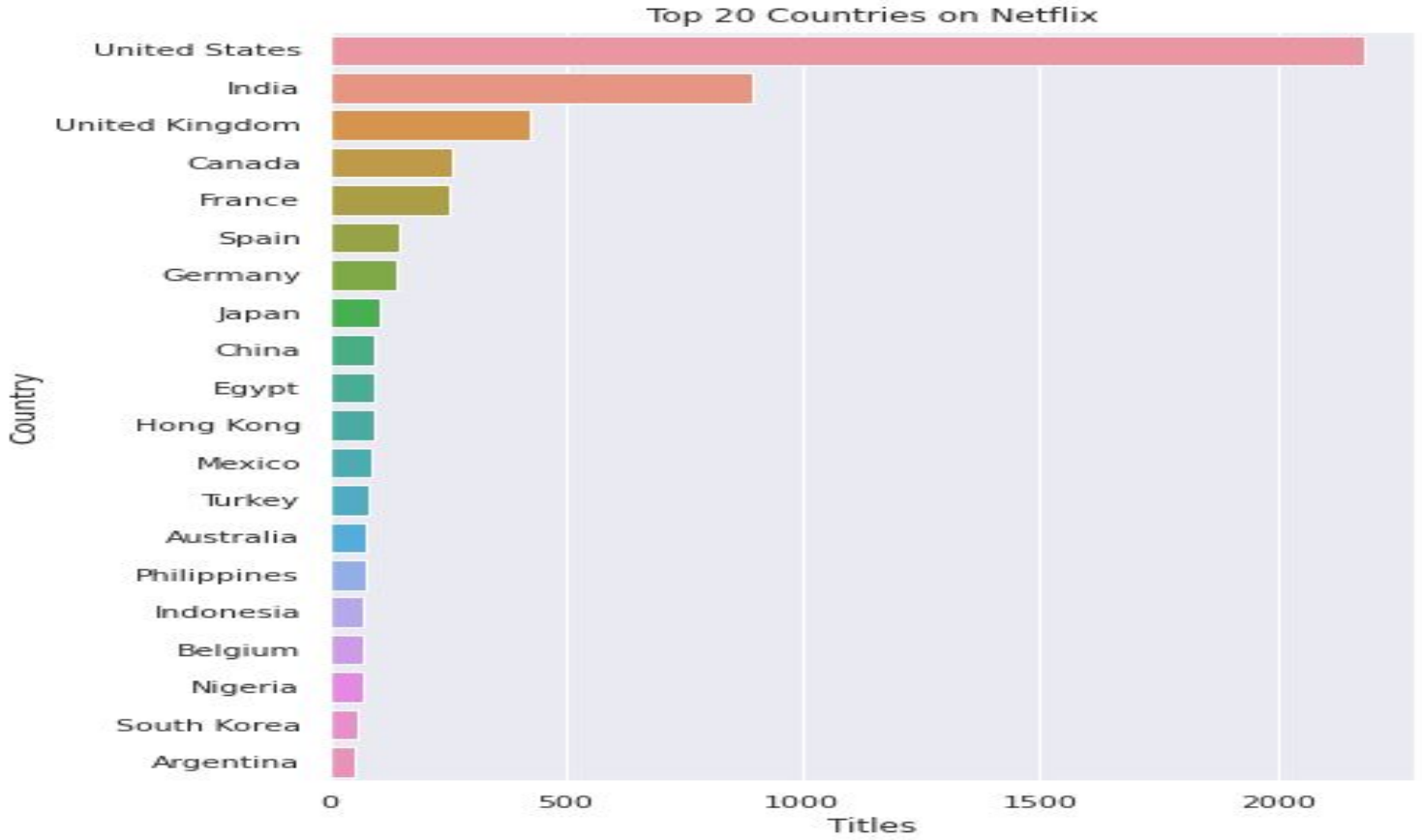
The number of release have significantly increased after 2015 and have dropped in 2021 due to Covid 19

Rating

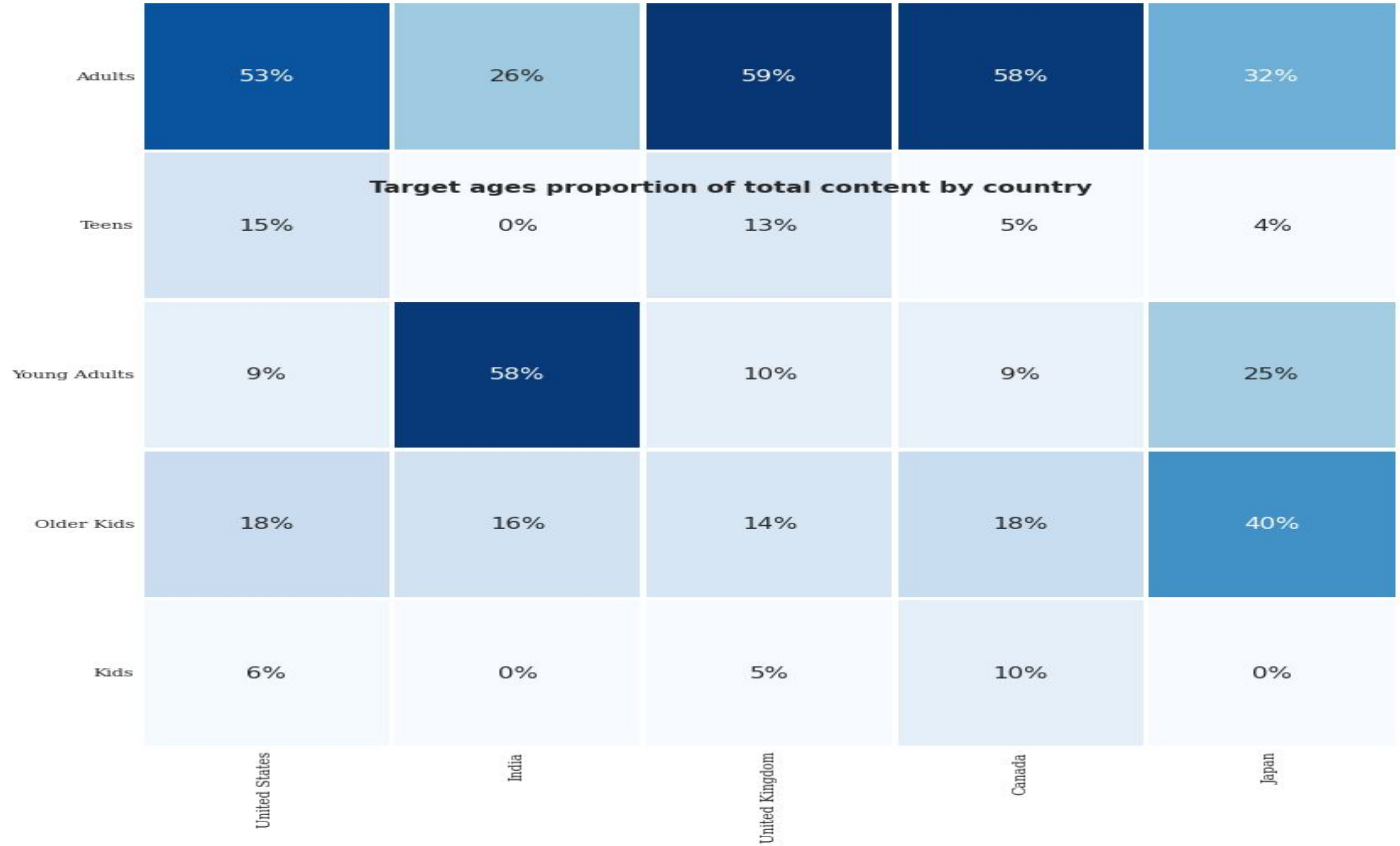


As shown in the plot above, TV-MA ratings are the highest in movies and TV shows, because TV-MA standing for Mature Audience Only. Because this programme is intended for adults, it may not be appropriate for children under the age of 17.

Countries with the most content available

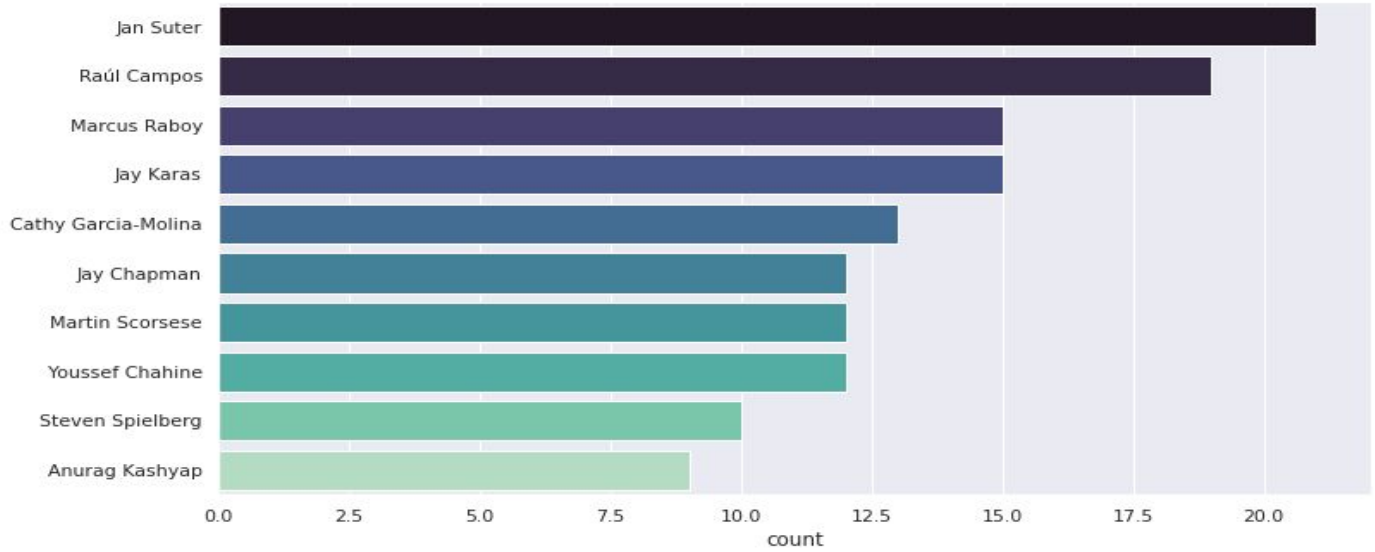


Netflix Content for different age groups.



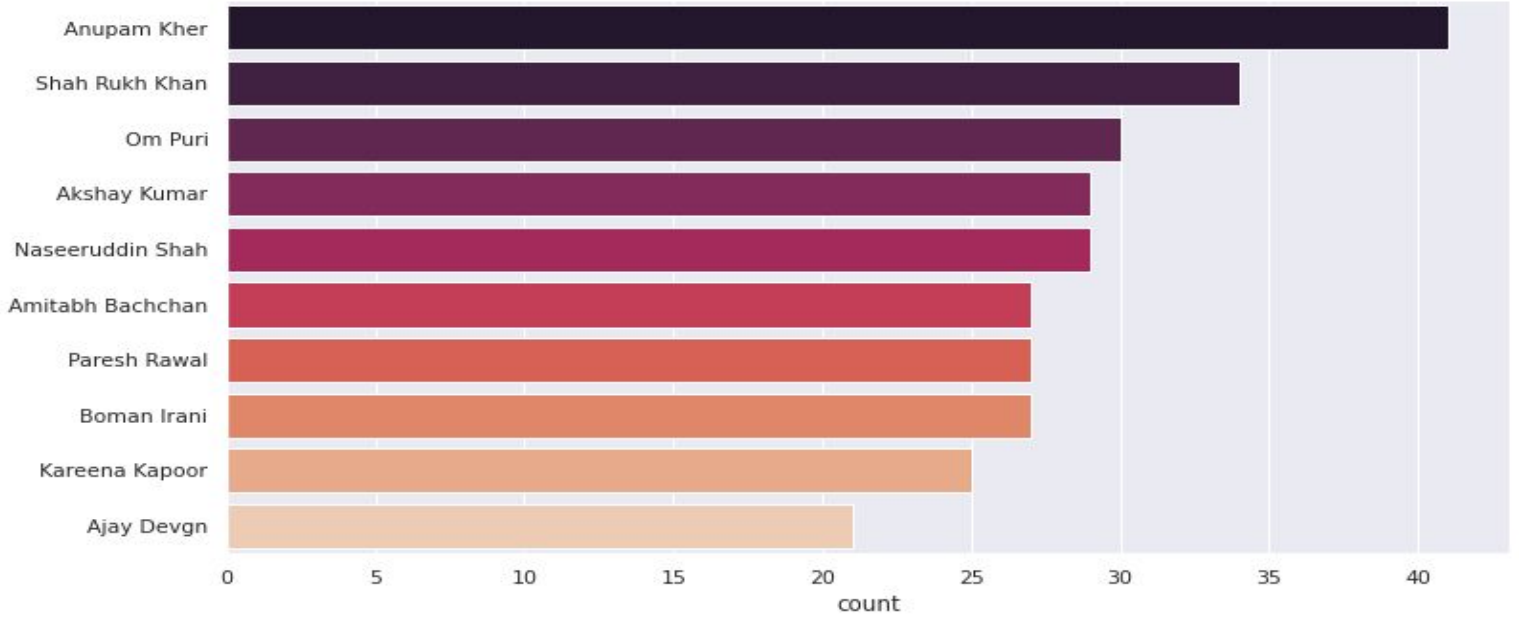
Asking and Answering Questions

Who are the top 10 directors on Netflix with the most releases?



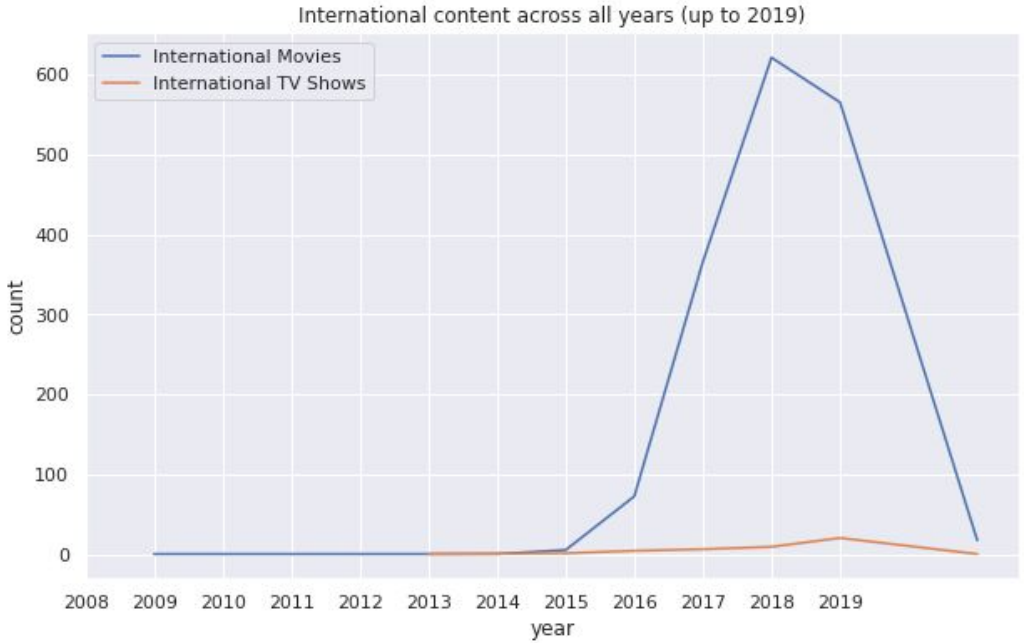
As stated previously regarding the top genres, it's no surprise that the most popular directors on Netflix with the most titles are mainly international as well.

Who are the top 10 actors on Netflix based on number of titles?



In this list, we can see that the most popular actors on Netflix based on the number of titles are all international as well. This reinforces the sentiment that the majority of Netflix subscribers are international.

How does the timeline look like for the addition of International Movies compared to International TV Shows?



Based on the timeline, we can see that there are far more international movie releases than there are international tv show releases. However, from 2015 onwards the growth of international movies has been increased till 2018 then it started to decline while international tv shows constantly showed significant growth in the past few years.

Data Preprocessing

Modelling Approach:

1.Select the attributes based on which you want to cluster the shows

2.Text preprocessing: Remove all, stopwords and punctuation marks, convert all textual data to lowercase.

3.Lemmatization to generate a meaningful word out of corpus of words

4.Tokenization of corpus

5.Word vectorization

6.Dimensionality reduction

7.Used different algorithms to cluster the movies, obtain the optimal number of clusters using different techniques.

We will cluster the shows on Netflix based on the following attributes:

Director

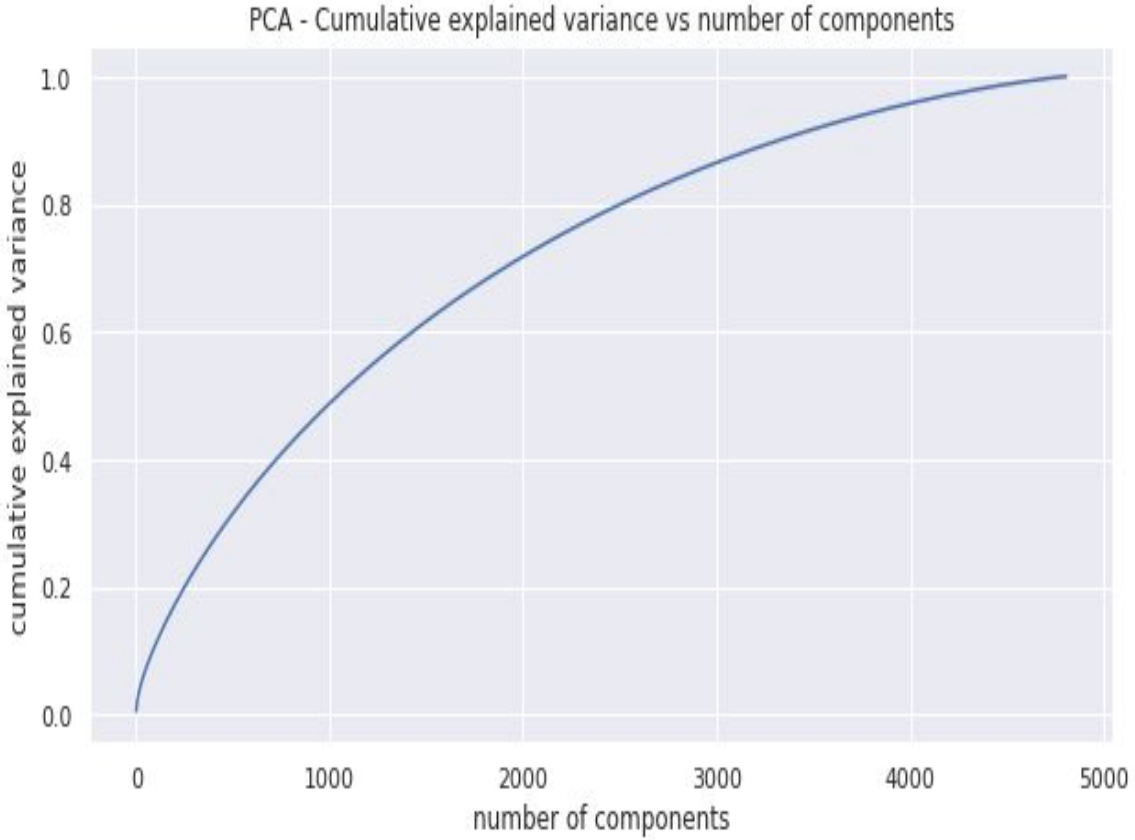
Cast

Country

Listed in (genres)

Description

Dimensionality reduction using PCA:



1.I found that 100% of the variance is explained by about ~4500 components.

2.Also, more than 80% of the variance is explained just by 3000 components.

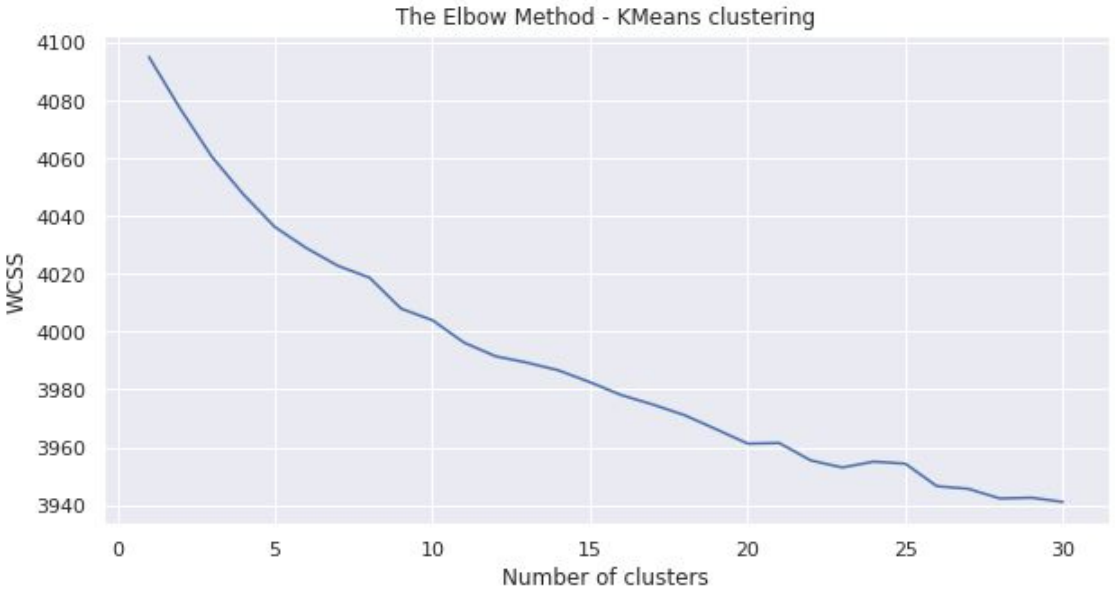
3.Hence to simplify the model, and reduce dimensionality, I can take the top 3000 components, which will still be able to capture more than 80% of variance.

Clusters implementation:

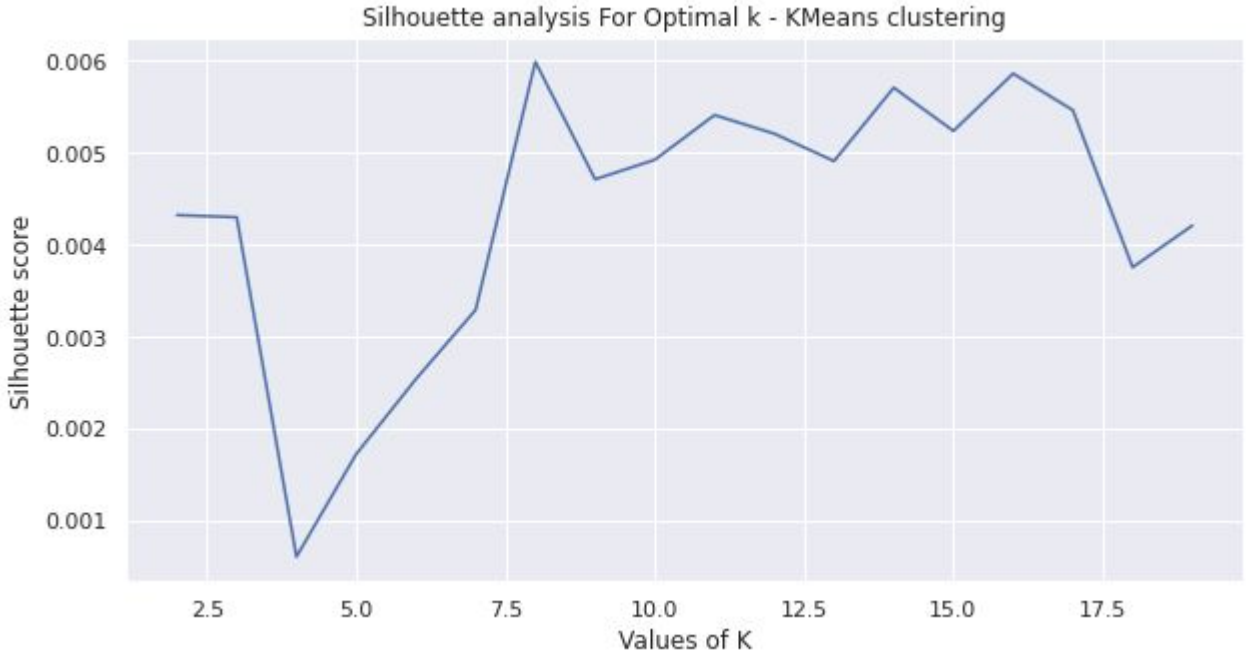
K-Means Clustering

Building clusters using the K-means clustering algorithm.

Visualizing the elbow curve and Silhouette score to decide on the optimal number of clusters for K-means clustering algorithm.



The sum of squared distance between each point and the centroid in a cluster (WCSS) decreases with the increase in the number of clusters.

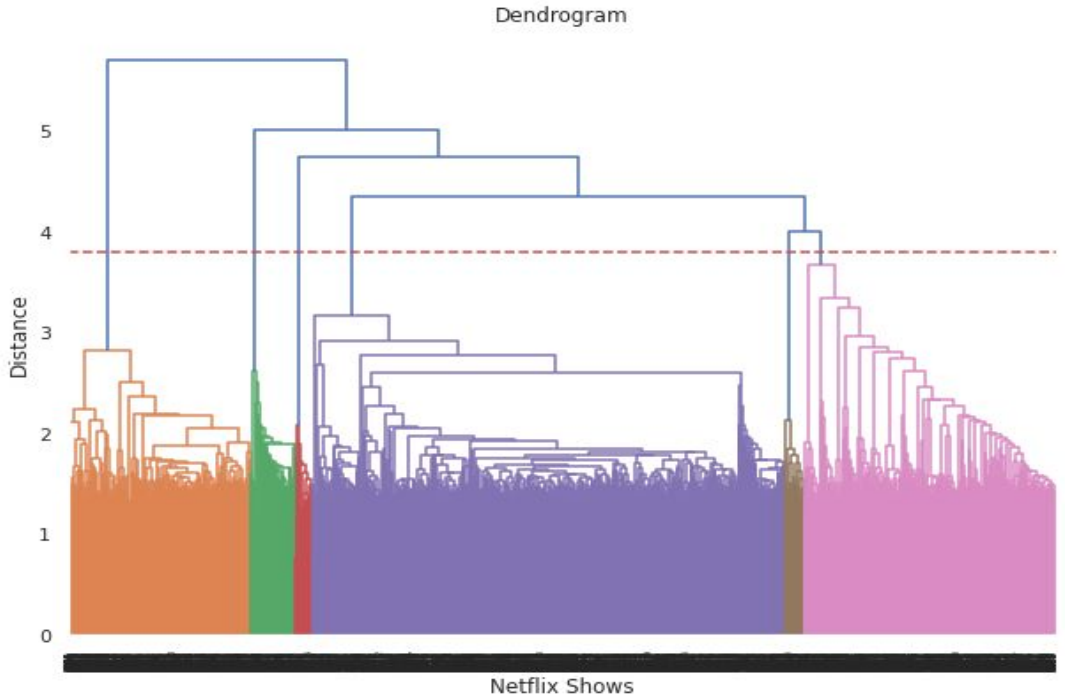


The highest Silhouette score is obtained for 8 clusters.

Building 8 clusters using the k-means clustering algorithm:

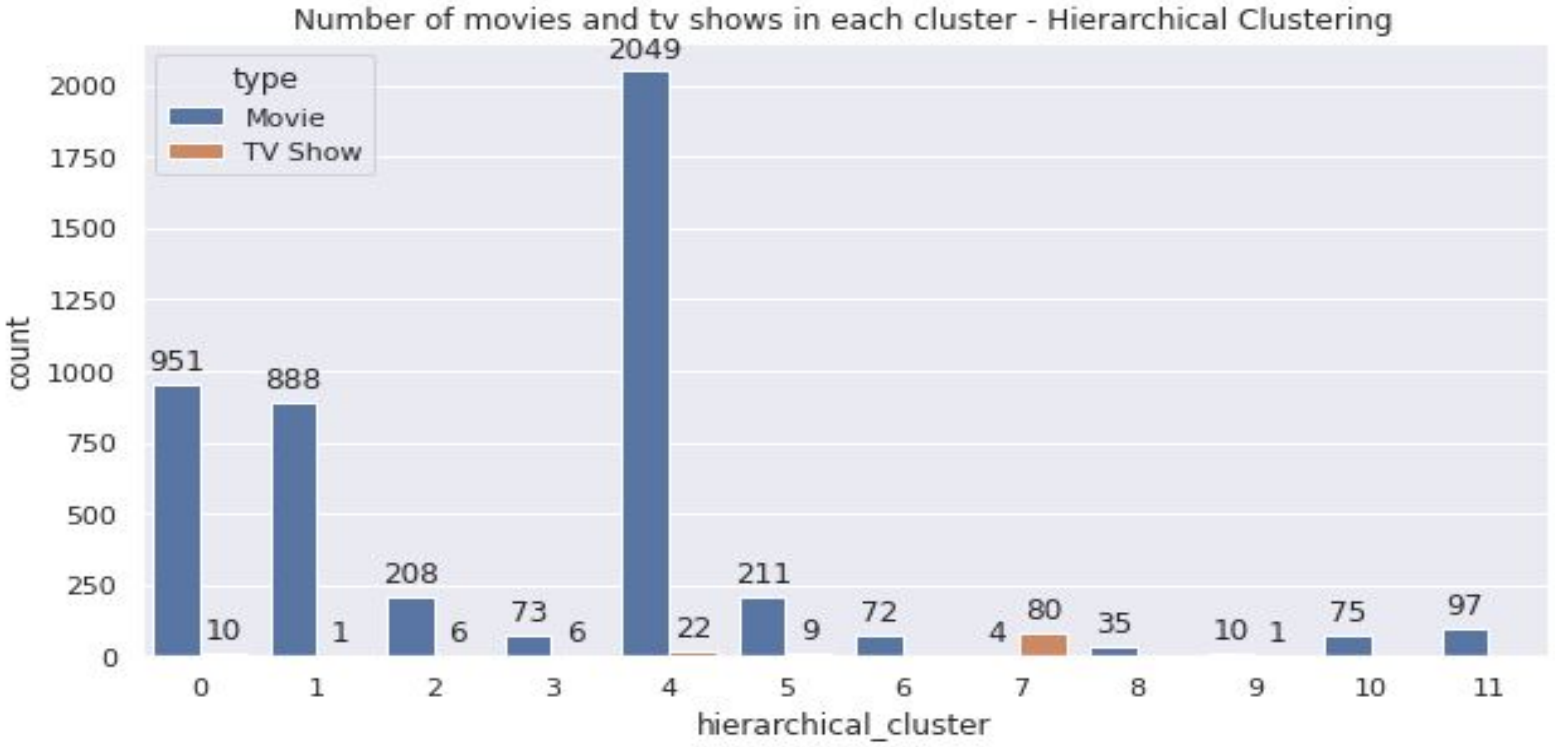
Hierarchical clustering

Dendrogram:



At a distance of 3.8 units, 12 clusters can be built using the agglomerative clustering algorithm.

Building 12 clusters using the Agglomerative clustering algorithm:



Conclusions

1.The dataset contained about 7787 rows, and 12 columns.

2.I began by dealing with the datasets missing values and doing exploratory data analysis (EDA).

3.There are more number movies on Netflix than TV shows. It is evident that there are more movies on Netflix than TV shows. Netflix has 4673 movies, which is more than double the quantity of TV shows.

4.The number of release have significantly increased after 2015 and have dropped in 2021 due to Covid 19.

5.TV-MA ratings are the highest in movies and TV shows, because TV-MA standing for Mature Audience Only. Because this programme is intended for adults, it may not be appropriate for children under the age of 17.

6.I used Principal Component Analysis (PCA) to handle the curse of dimensionality. 3000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 3000.

7.I first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 8. This was obtained through the elbow method and Silhouette score analysis.

8.Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 12. This was obtained after visualizing the dendrogram.

THANK YOU