Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Done By:-

Saquib Neyaz
Rajni Shukla

Contents

- Problem Statement
- Data Summary
- EDA and Feature Engineering
- Data cleaning and Pre-processing
- Top Modelling(LSA and LDA)
- Recommendation
- Applying different Clustering models
- Conclusion

Problem Statement

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. In this project, you are required to do -

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

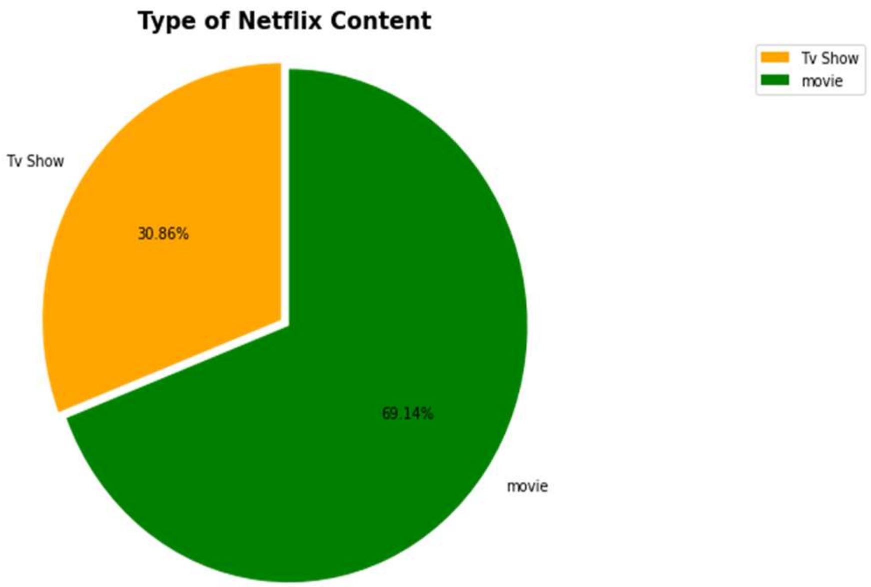
Data Summary

- 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 : Genres
- description: The Summary description

Exploratory Data Analysis

Different types of content present In the Netflix.

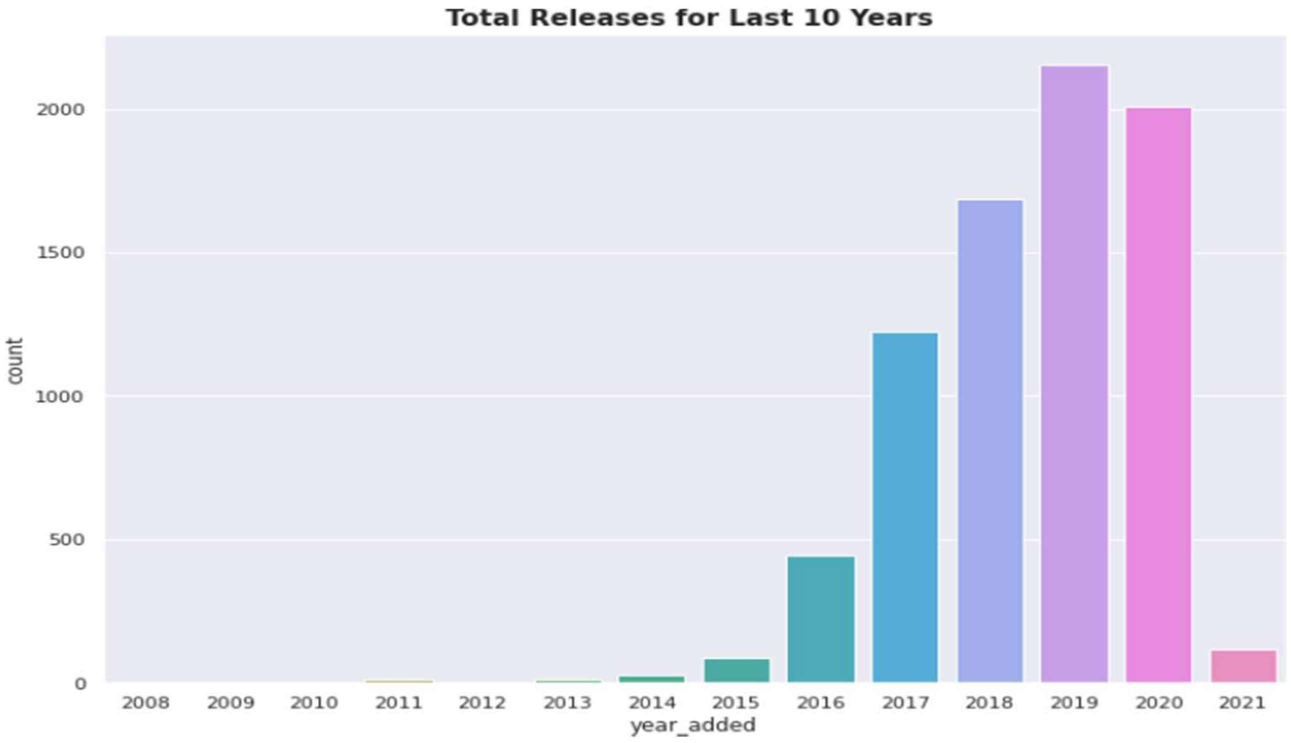
There are about 70% movies and 30% TV shows on Netflix. It means nearly 2/3rd of the Content on Netflix are "movies” while the rest are "TV Shows".



Continued…

Year wise Analysis :-

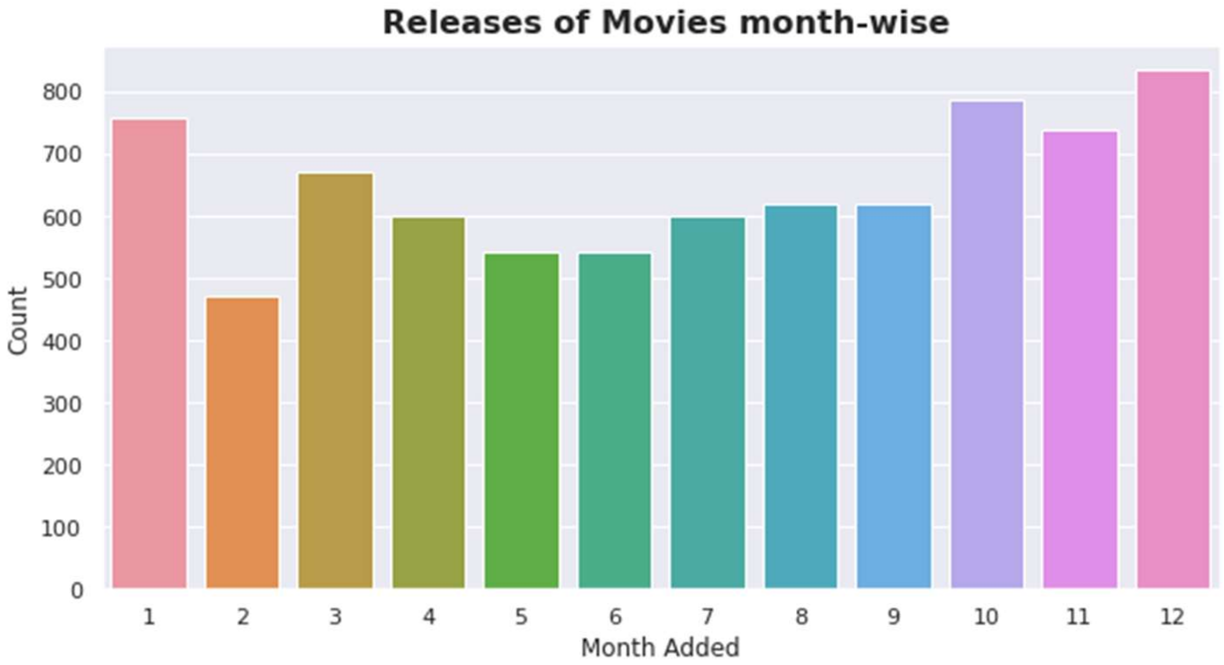
The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19.



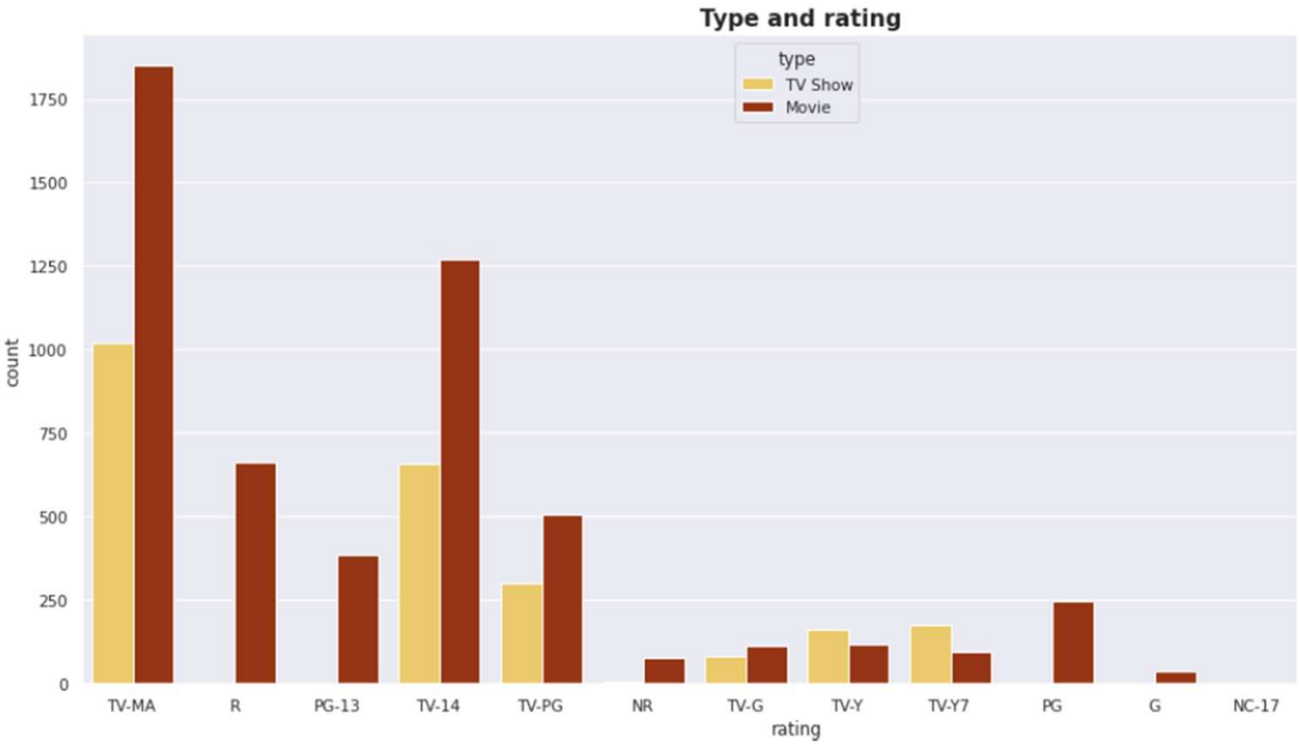
Continued…

<u>Month wise Analysis</u> :-

More of the content is released in holiday season - October, November, December and January

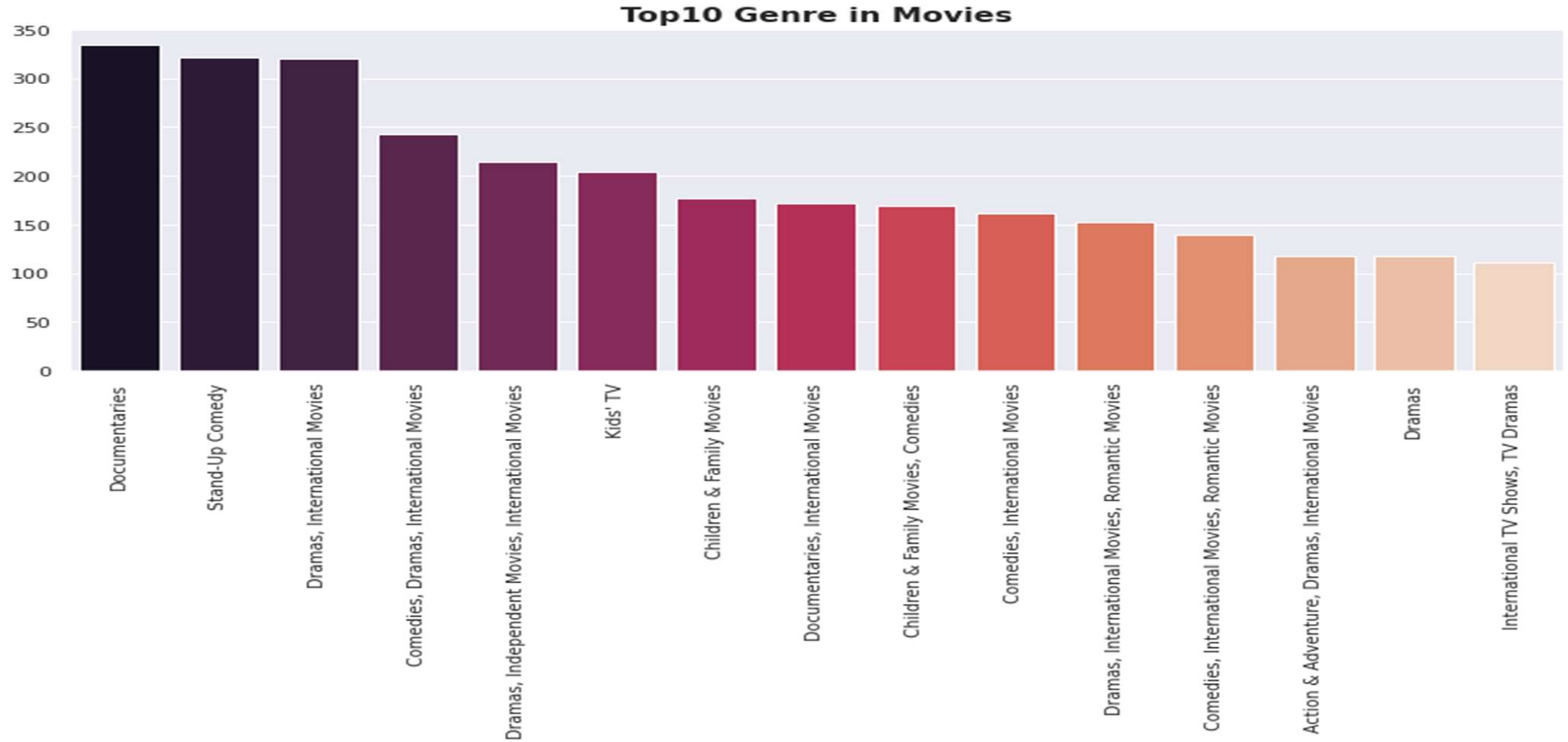


Continued…

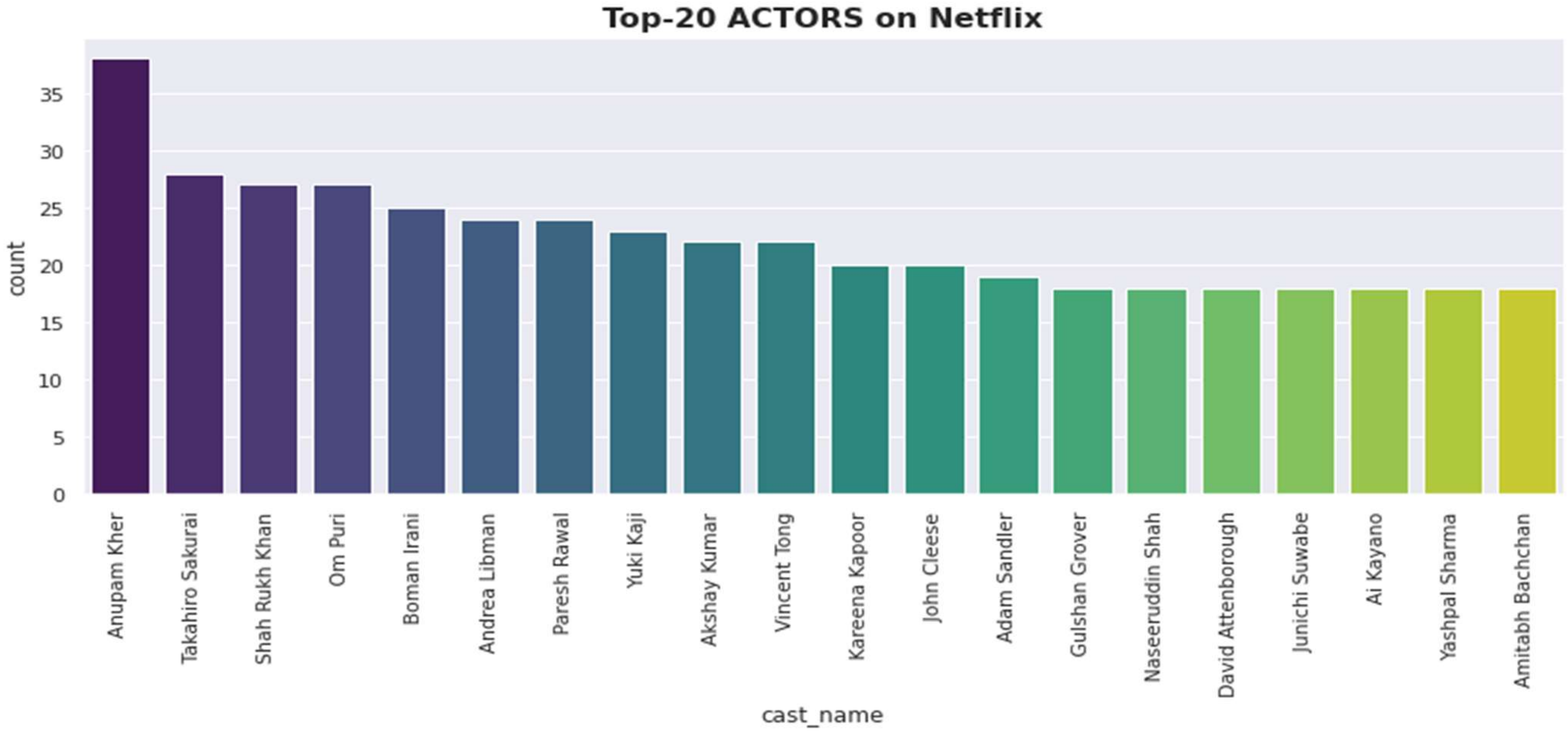


The 'TV-MA' rating is used in the majority of the film.

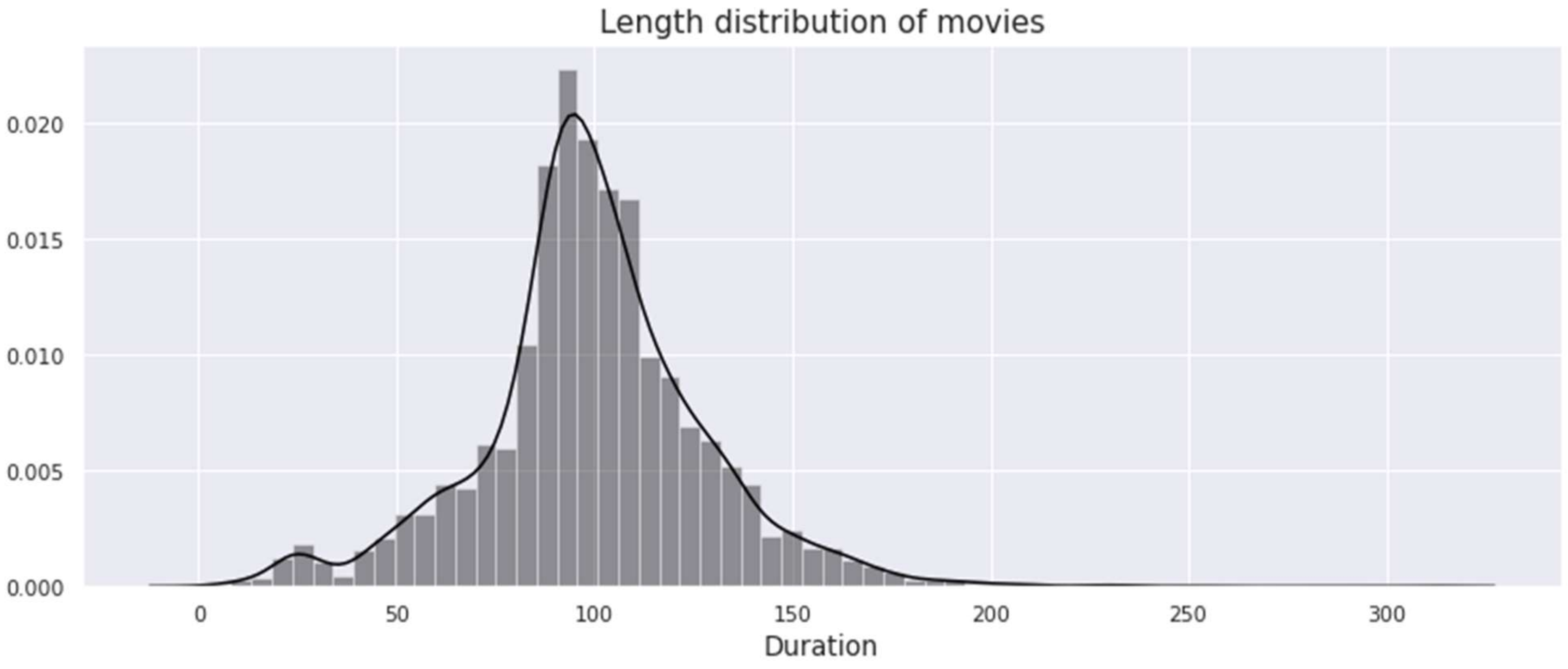
Continued…


Top10 Genre in Movies
350
300
250
200
150
100
50
0
Documentaries
Stand-Up Comedy
Dramas, International Movies
Comedies, Dramas, International Movies
Dramas, Independent Movies, International Movies
Kids' TV
Children & Family Movies
Documentaries, International Movies
Children & Family Movies, Comedies
Comedies, International Movies
Dramas, International Movies, Romantic Movies
Comedies, International Movies, Romantic Movies
Action & Adventure, Dramas, International Movies
Dramas
International TV Shows, TV Dramas

Continued…


Top-20 ACTORS on Netflix
count
0
5
10
15
20
25
30
35
cast_name
Anupam Kher
Takahiro Sakurai
Shah Rukh Khan
Om Puri
Boman Irani
Andrea Libman
Paresh Rawal
Yuki Kaji
Akshay Kumar
Vincent Tong
Kareena Kapoor
John Cleese
Adam Sandler
Gulshan Grover
Naseeruddin Shah
David Attenborough
Junichi Suwabe
Ai Kayano
Yashpal Sharma
Amitabh Bachchan

Continued…


Length distribution of movies
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
Duration

Data Cleaning and Pre-processing

- Label Encoding
- Lemmatisation- 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

Latent Semantic Analysis(LSA):-

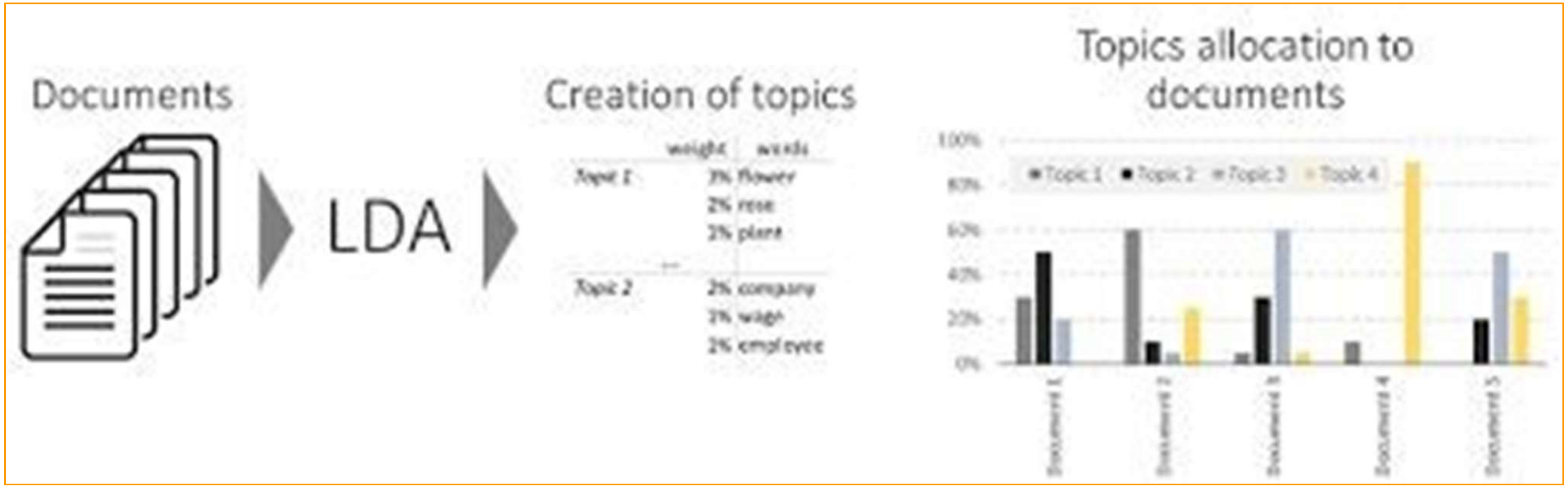
LSA, which stands for Latent Semantic Analysis, is one of the foundational techniques used in topic modeling. The core idea is to take a matrix of documents and terms and try to decompose it into separate two matrices –

- A document-topic matrix
- A topic-term matrix.

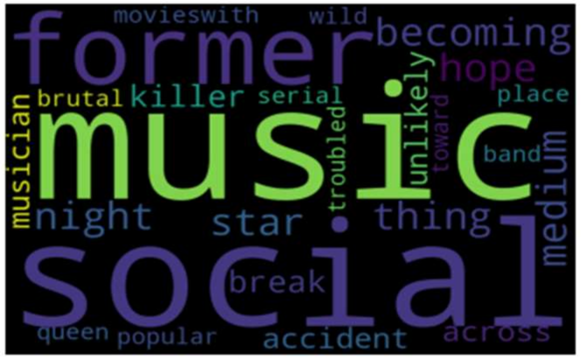
Therefore, the learning of LSA for latent topics includes matrix decomposition on the document-term matrix using Singular value decomposition. It is typically used as a dimension reduction or noise-reducing technique.

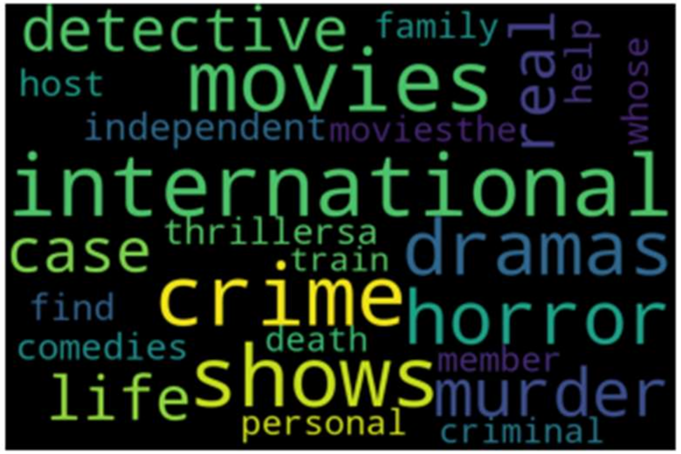
Latent Dirichlet Allocation(LDA):-

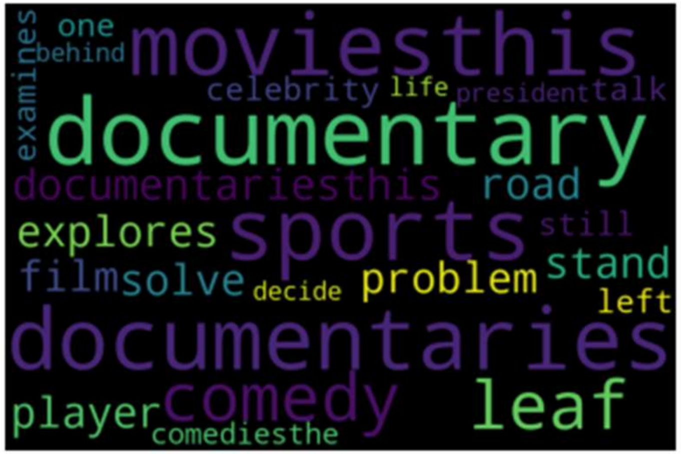
LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.



Genre Word cloud

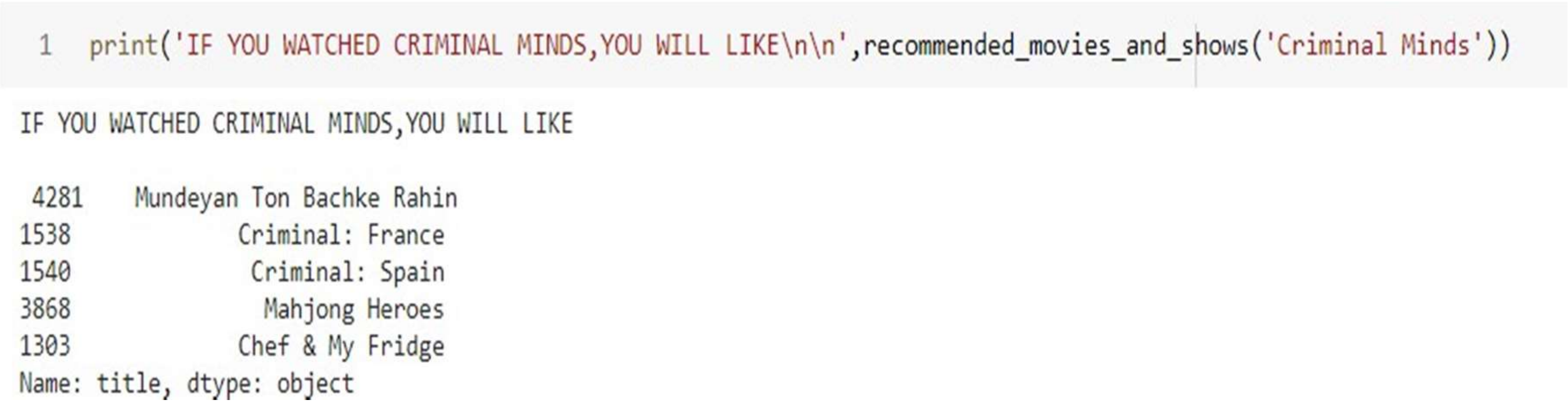






Recommendation

A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item.



```
print('IF YOU WATCHED CRIMINAL MINDS,YOU WILL LIKE\n\n',recommended_movies_and_shows('Criminal Minds'))
```

```
IF YOU WATCHED CRIMINAL MINDS,YOU WILL LIKE

 4281    Mundeyan Ton Bachke Rahin
1538             Criminal: France
1540              Criminal: Spain
3868               Mahjong Heroes
1303             Chef & My Fridge
Name: title, dtype: object
```

Different Clustering Models

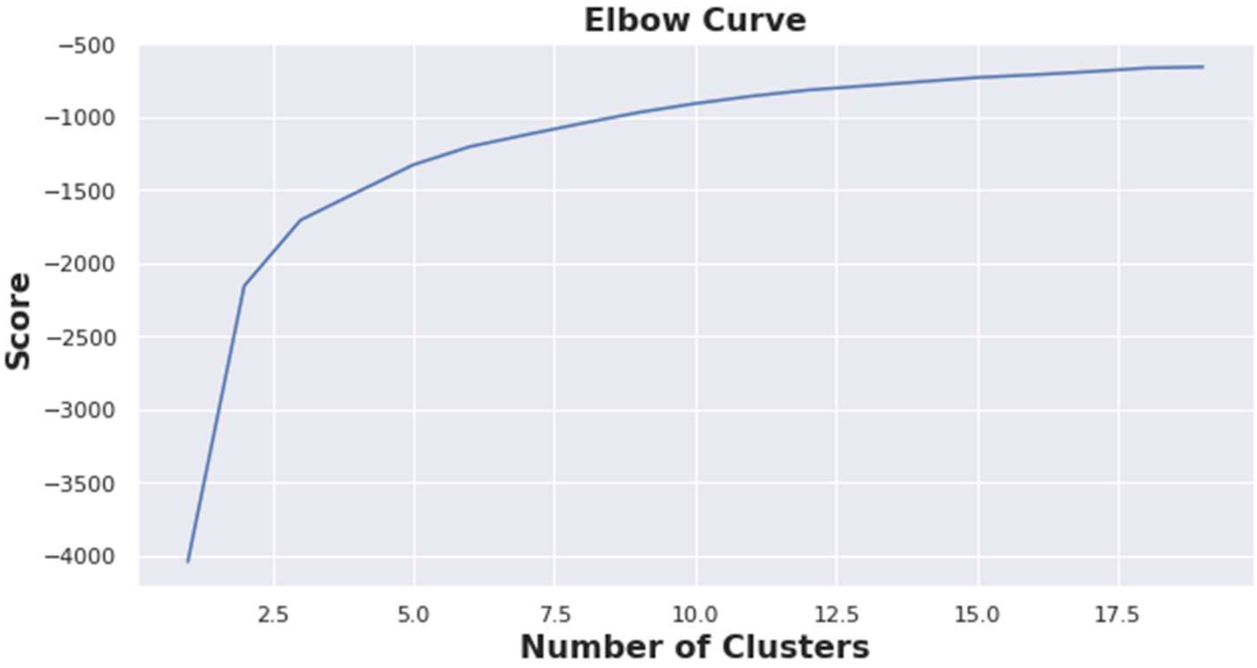
1) K – Means Clustering

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:

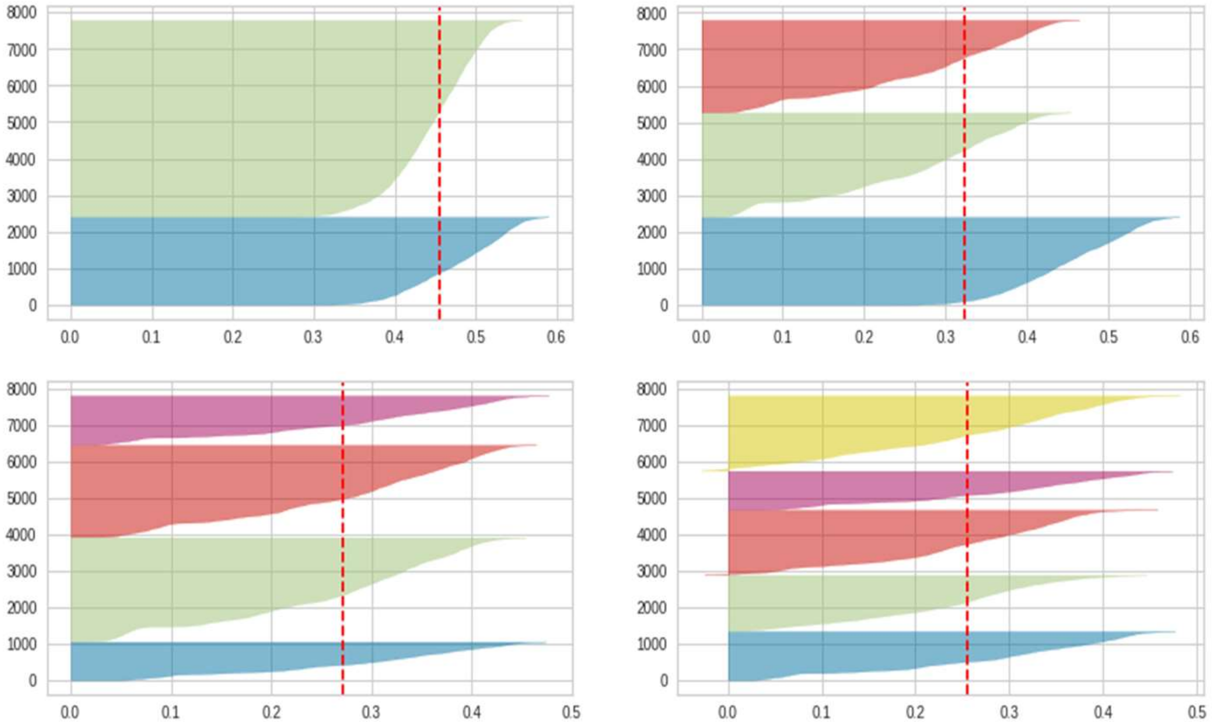
- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

Elbow Method:



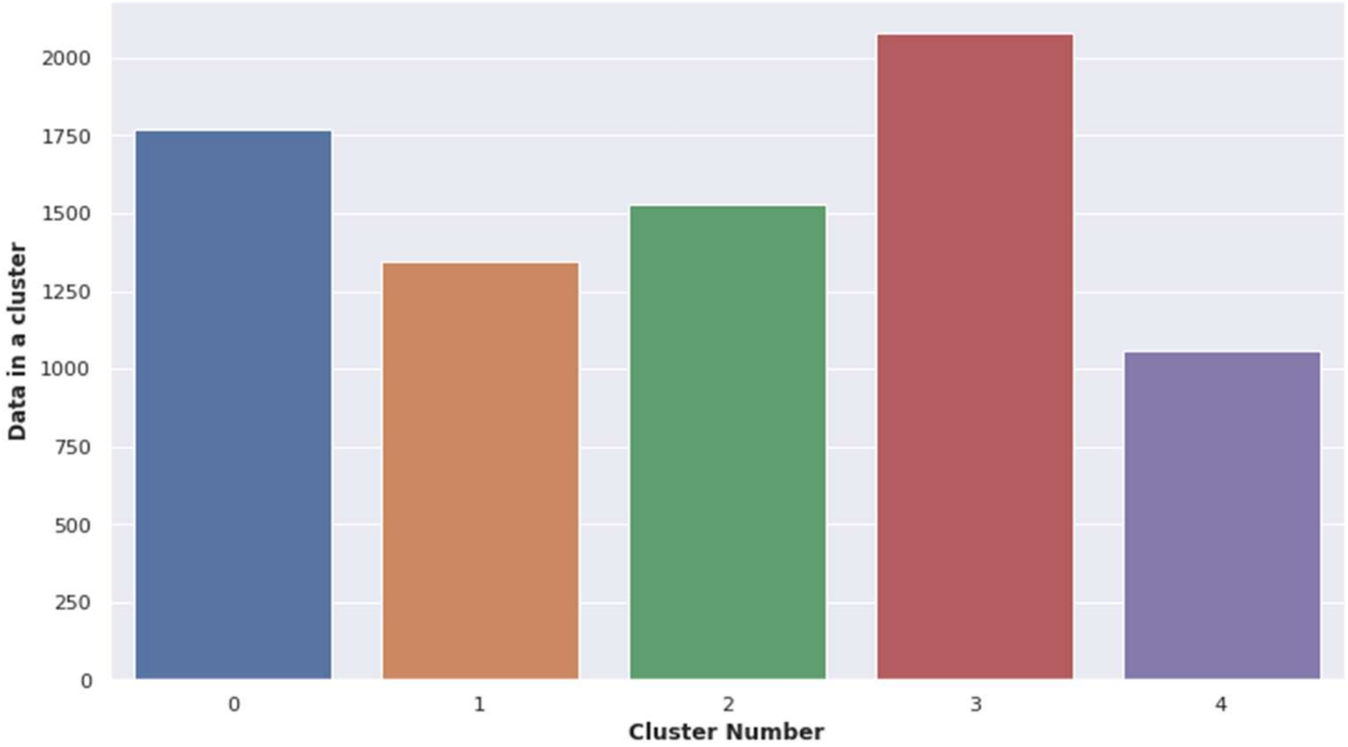
Silhouette Analysis :

Here is the Silhouette analysis done 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.

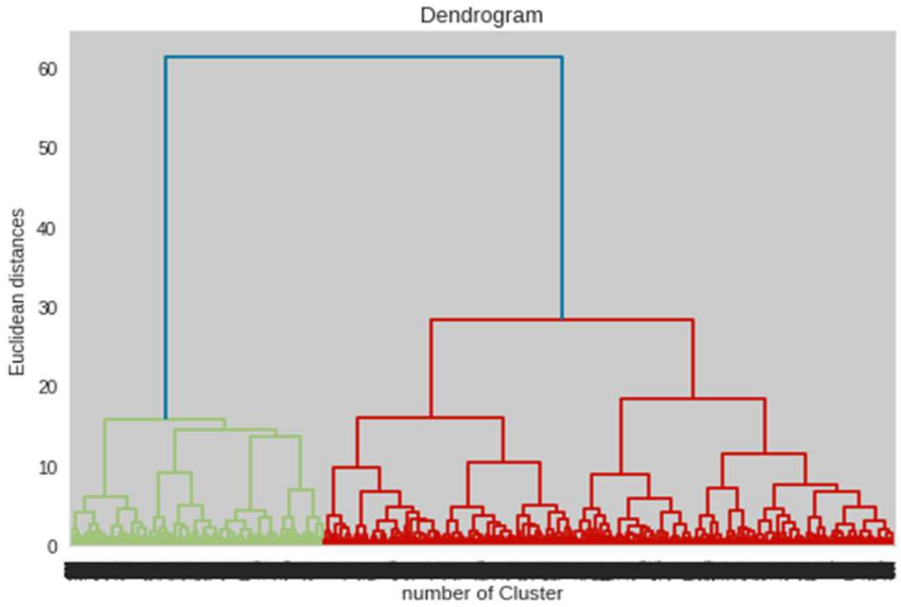


Data in a Cluster :

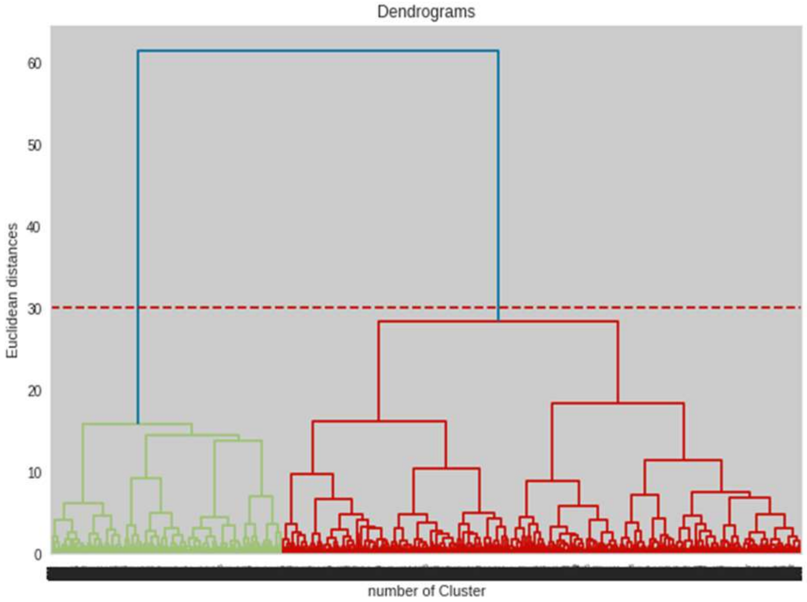
We clearly see that one cluster is the largest and one cluster has the fewest number of movies.



2) Hierarchical Cluster



The x-axis contains the samples and y-axis represents the distance between these samples. The vertical line with maximum distance is the blue line and hence we can decide a threshold of 30 and cut the dendrogram.



We only found 2 cluster, that why we decided to not going forward with hierarchal clustering.

CONCLUSION

1. There are about 70% movies and 30% TV shows on Netflix.
2. In this context, we've noticed that Netflix is increasingly focusing on movies rather than TV shows, especially after 2014.
3. We have also found that different types of content are available in different countries, but TV MA is the content that is available in the majority of countries. This could be because it shows that it is just for adult audiences, and the Netflix audience enjoys content like this.
4. The United States has the highest number of content on Netflix by a huge margin followed by India.

Continued :-

5. We have also defined different clusters based on their content; we've defined 5 clusters and implemented the K-MEANS clustering algorithm. we've also import Silhouette Visualizer which displays the silhouette coefficient for each sample on a per-cluster basis, visualizing which clusters are dense and which are not.
6. We have also apply Hierarchical cluster on our dataset, but that did not provides the best solution, it involved lots of arbitrary decisions, it did not work with missing data, it did not work well on very large data sets, and its main output, the dendrogram, was commonly misinterpreted.
7. After applying "**K – means**" optimal value of number of clusters is "**5**".
8. Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are.

THANK YOU