NETFLIX MOVIES AND TV SHOWS CLUSTERING

Minal Kharbade , Deveshya Gupta

CAPSTONE PROJECT – 4

ALMABETTER, BANGLORE

ABSRACT:

- Netflix is a subscription based streaming service that allows our members to watch TV shows and movies commercials on an on an internet connected device.
- 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 OTT platform 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.
- Exploratory data analysis is done on dataset to get the insights from the information however the principal invalid qualities are taken care of .
- Clustering is the useful technique to achieve the best possible recommendation and increase the viewership of the platform.

PROBLEM STATEMENT:

The dataset consist of TV shows and movies available Netflix as of 2019. In 2018 third party released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The services number of movies has decreased by more than 2000 titles since 2010, while its number of TV shows has 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:

- There are 12 features and around 7787 observation in the dataset and are mostly textual features.
- It will be interesting to explore what all other insights can be obtained from the same dataset. The dataset contains following column:

    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
    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 Releaseyear 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 : Genere
12. description: The Summary description

STEPS INVOLVED IN ANALYSIS:

- Handling Missing Values:

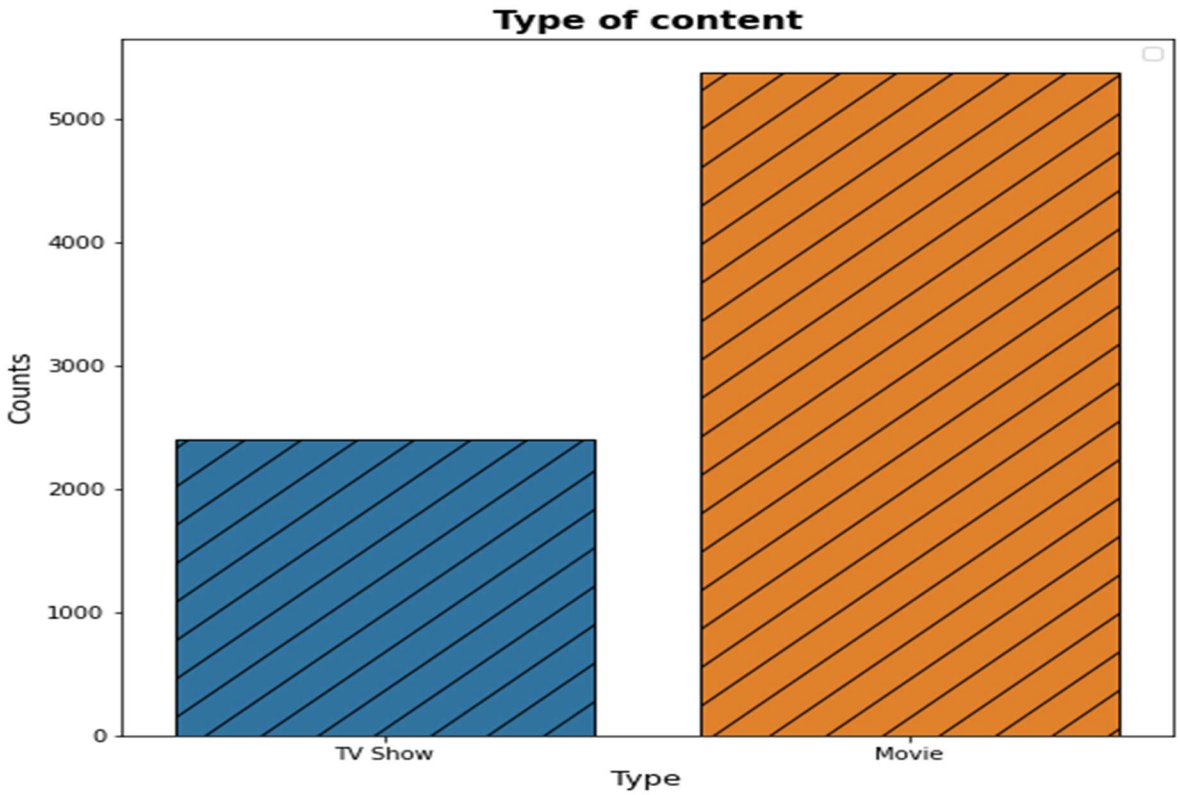
  We substitute the null values with word 'unkown' for further analysis. There are very few null entries in the date_added fields thus we delete them.

- Duplicate values Treatment:

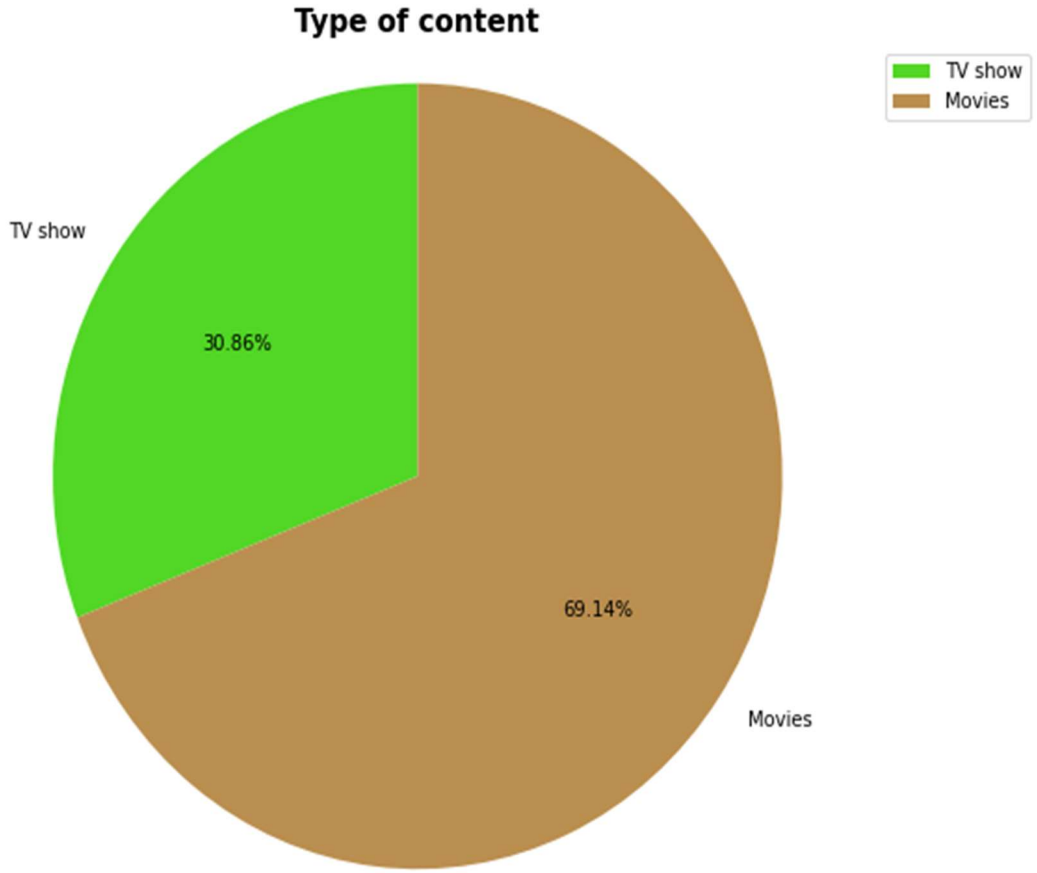
  Duplicate values does not contribute anything to accuracy of result. Our dataset does not contains any duplicate values.

- Exploratory Data Analysis:

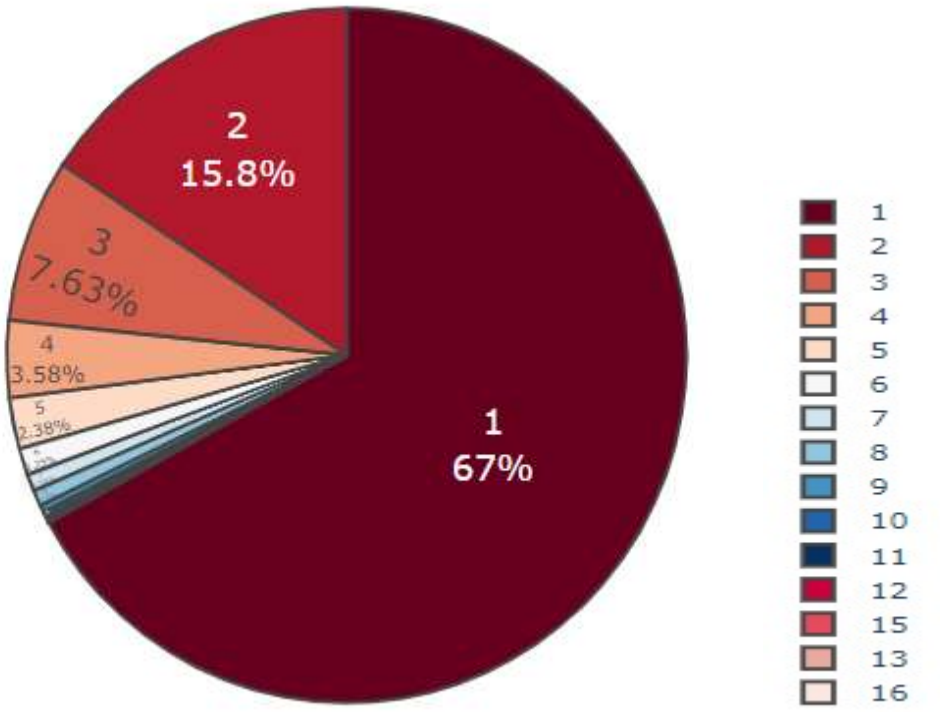
To get understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step.



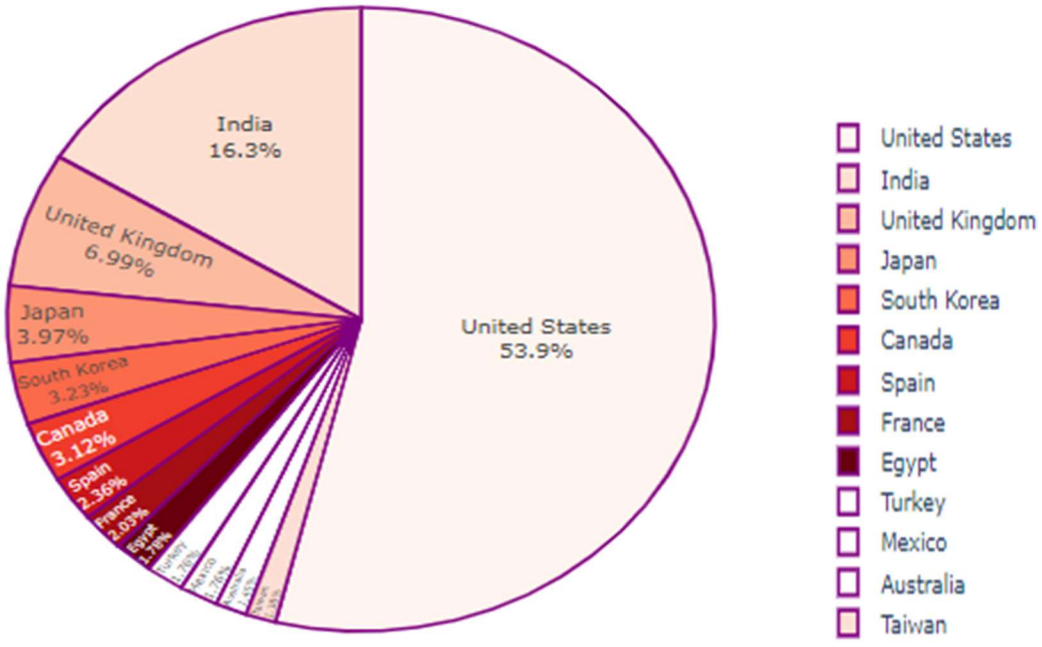
Type of Netflix Content:



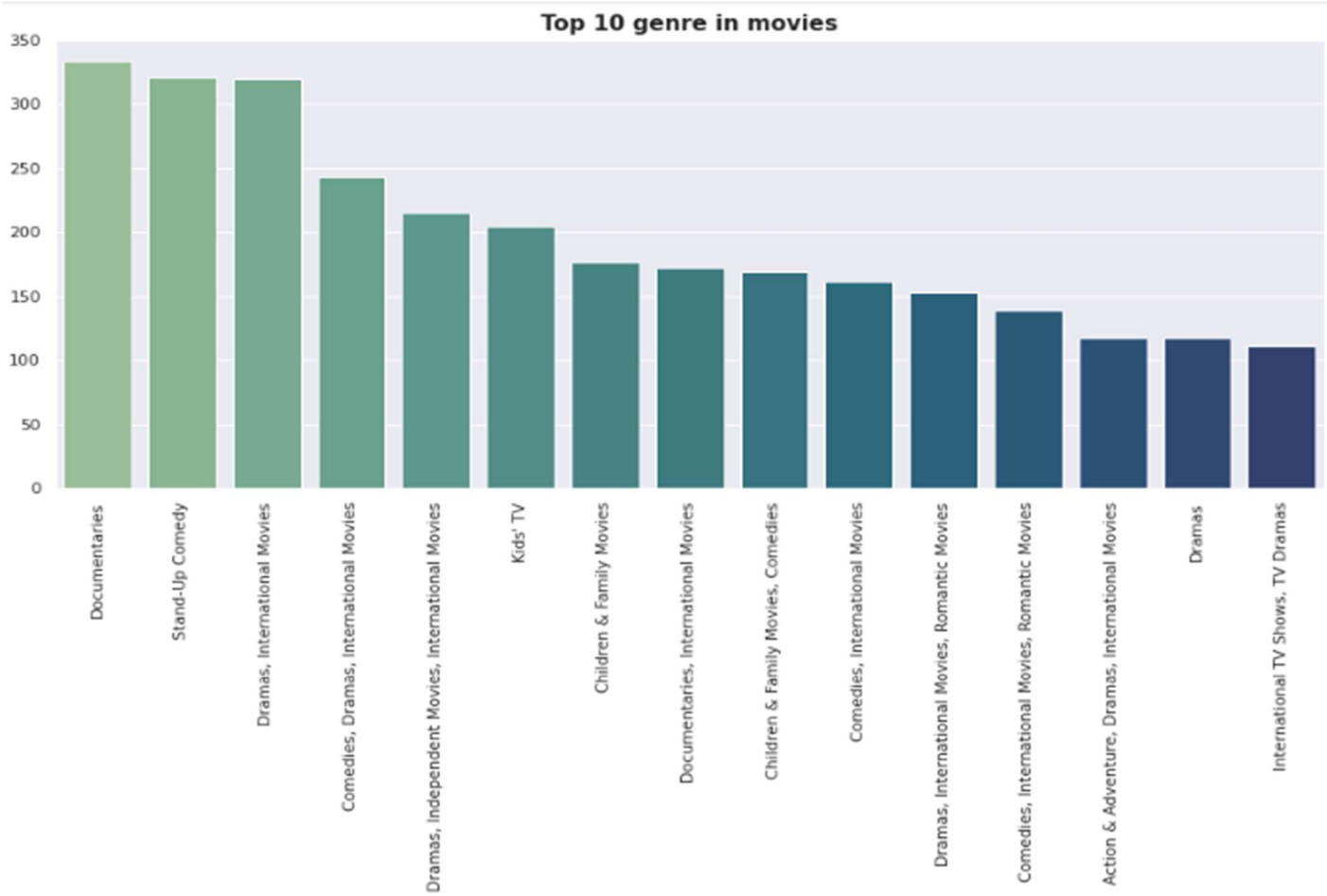
Season wise contribution of TV shows:

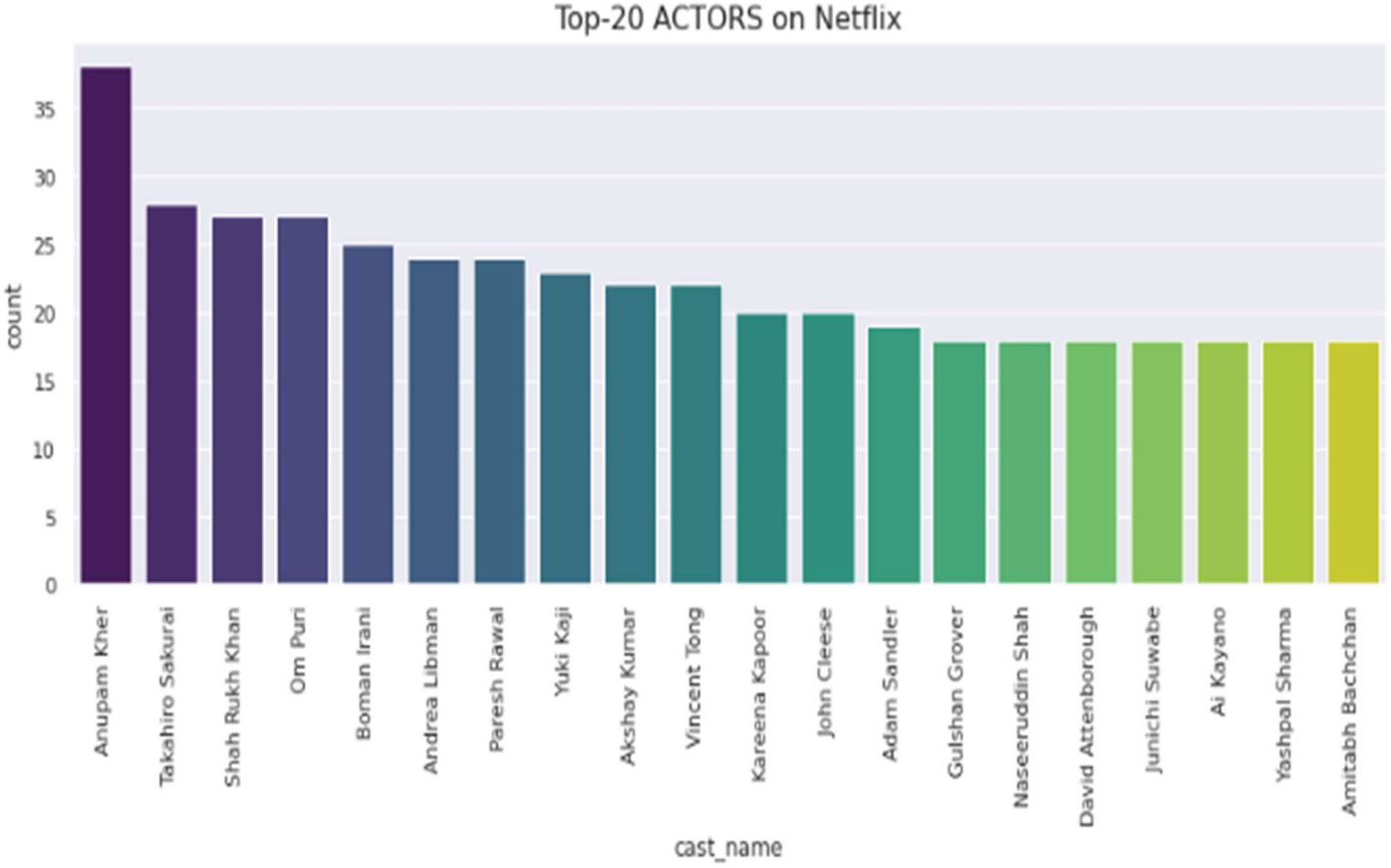


Contribution of different counties:

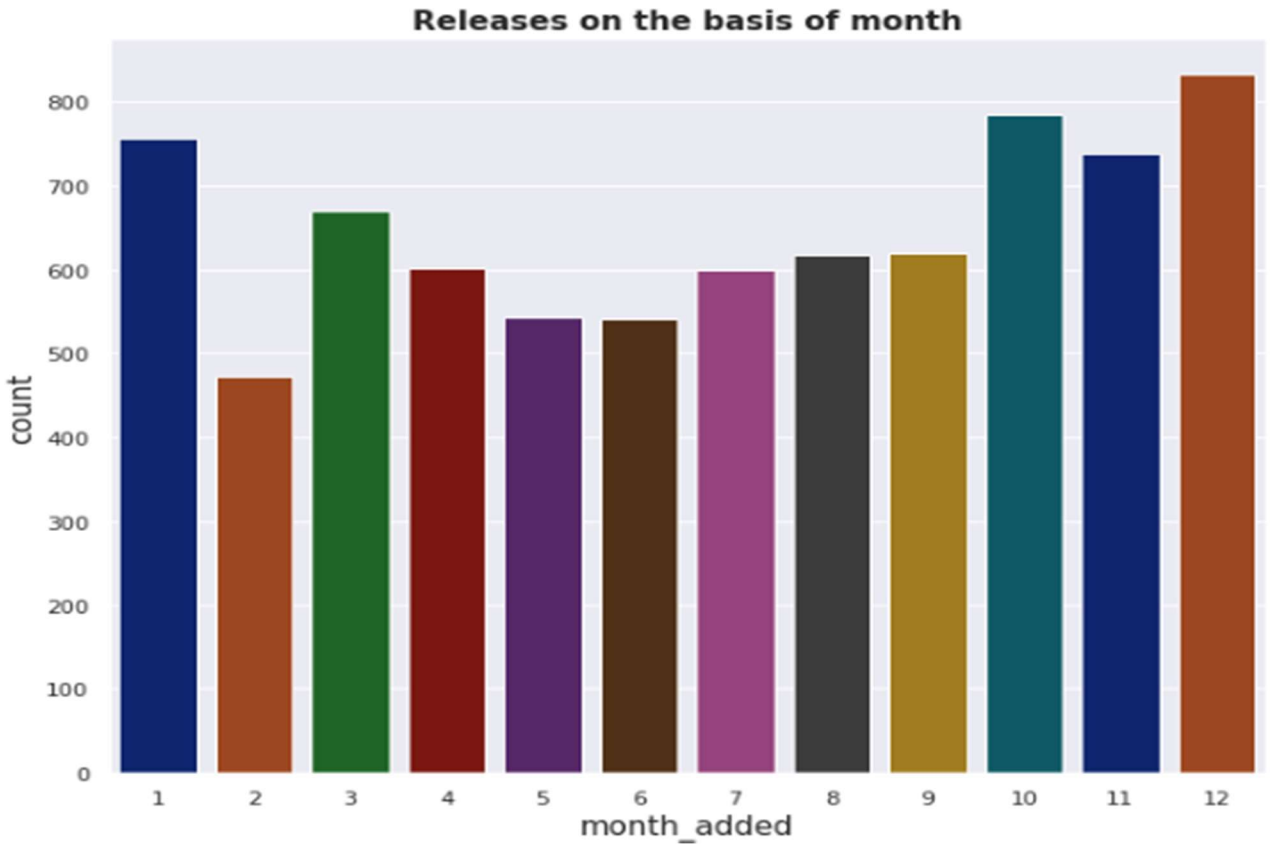


United States is the mostly generator of Netflix content, with India and UK trailing far behind.

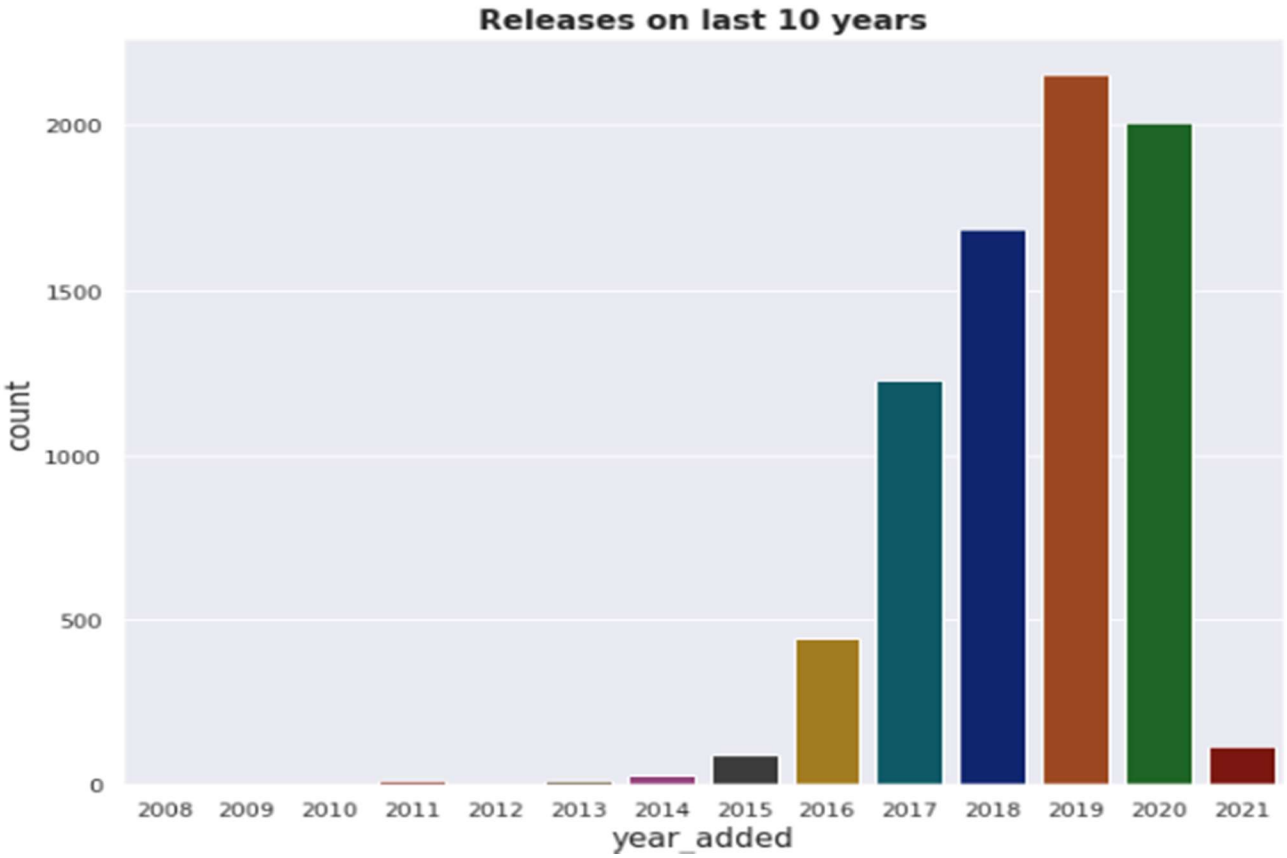

Top 10 genre in movies
350
300
250
200
150
100
50
0
Documentaries
Stand-Up Comedy
Dramas, International Movies
Comedies, Dramas, International Movies
Dramas, Independent Movies, International Movies
Kids' TV
Children & Family Movies
Documentaries, International Movies
Children & Family Movies, Comedies
Comedies, International Movies
Dramas, International Movies, Romantic Movies
Comedies, International Movies, Romantic Movies
Action & Adventure, Dramas, International Movies
Dramas
International TV Shows, TV Dramas

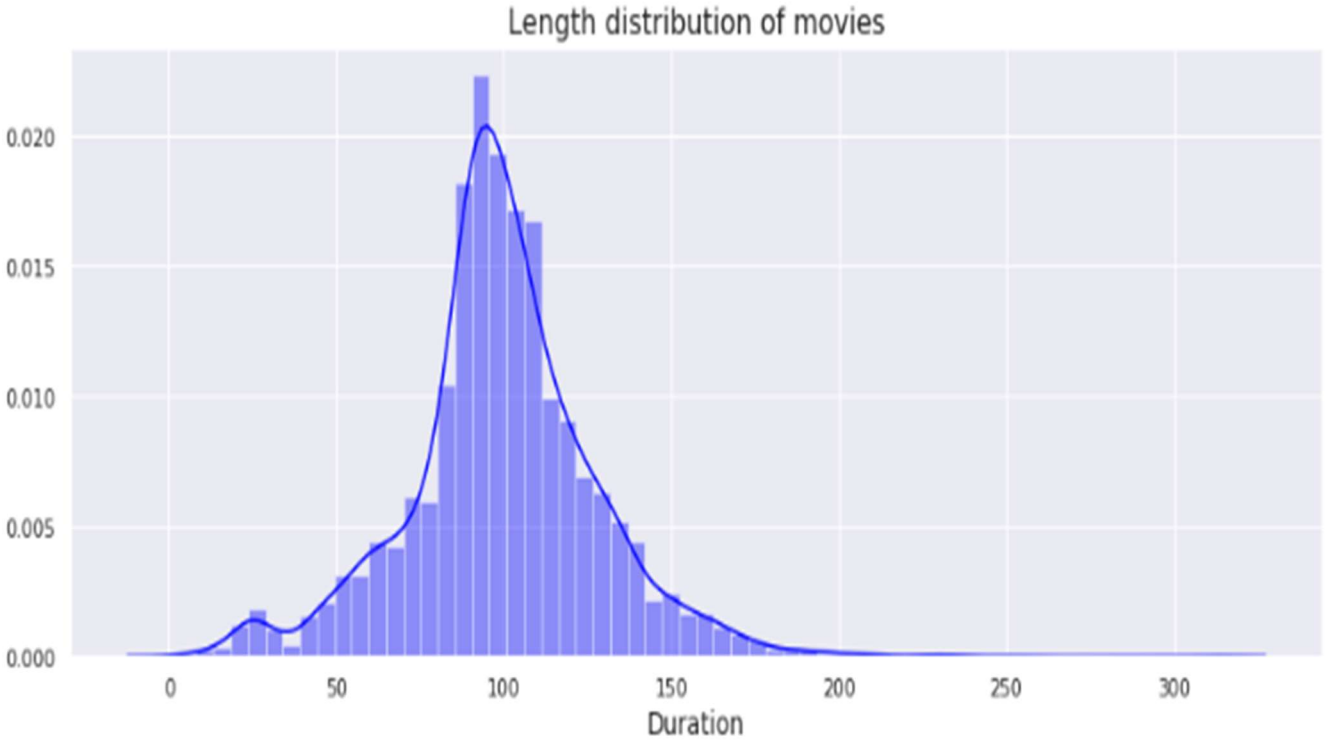

Top-20 ACTORS on Netflix
count
35
30
25
20
15
10
5
0
Anupam Kher
Takahiro Sakurai
Shah Rukh Khan
Om Puri
Boman Irani
Andrea Libman
Paresh Rawal
Yuki Kaji
Akshay Kumar
Vincent Tong
Kareena Kapoor
John Cleese
Adam Sandler
Gulshan Grover
Naseeruddin Shah
David Attenborough
Junichi Suwabe
Ai Kayano
Yashpal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

Month-wise analysis:



Year-wise analysis:





- Data processing:

  Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data , or noise. 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.

  Clustering

  Clustering (also called cluster analysis) is a task of grouping similar instances into clusters. More formally, clustering is the task of grouping population of unlabeled data points into clusters in a way that data points in the same cluster are more similar to each other clusters. The clustering task is probably the most important in unsupervised learning since it has many applications.

  Topic modeling:

  - **Latent Semantic Analysis:**

    LSA stands for latent semantic analysis is one of the foundational technique used in topic modeling. Latent semantic analysis is a natural language

processing method that analyzes relationship between set of documents and term contain within. The core idea is to take matrix of documents and term and try to decompose it into separate two matrices

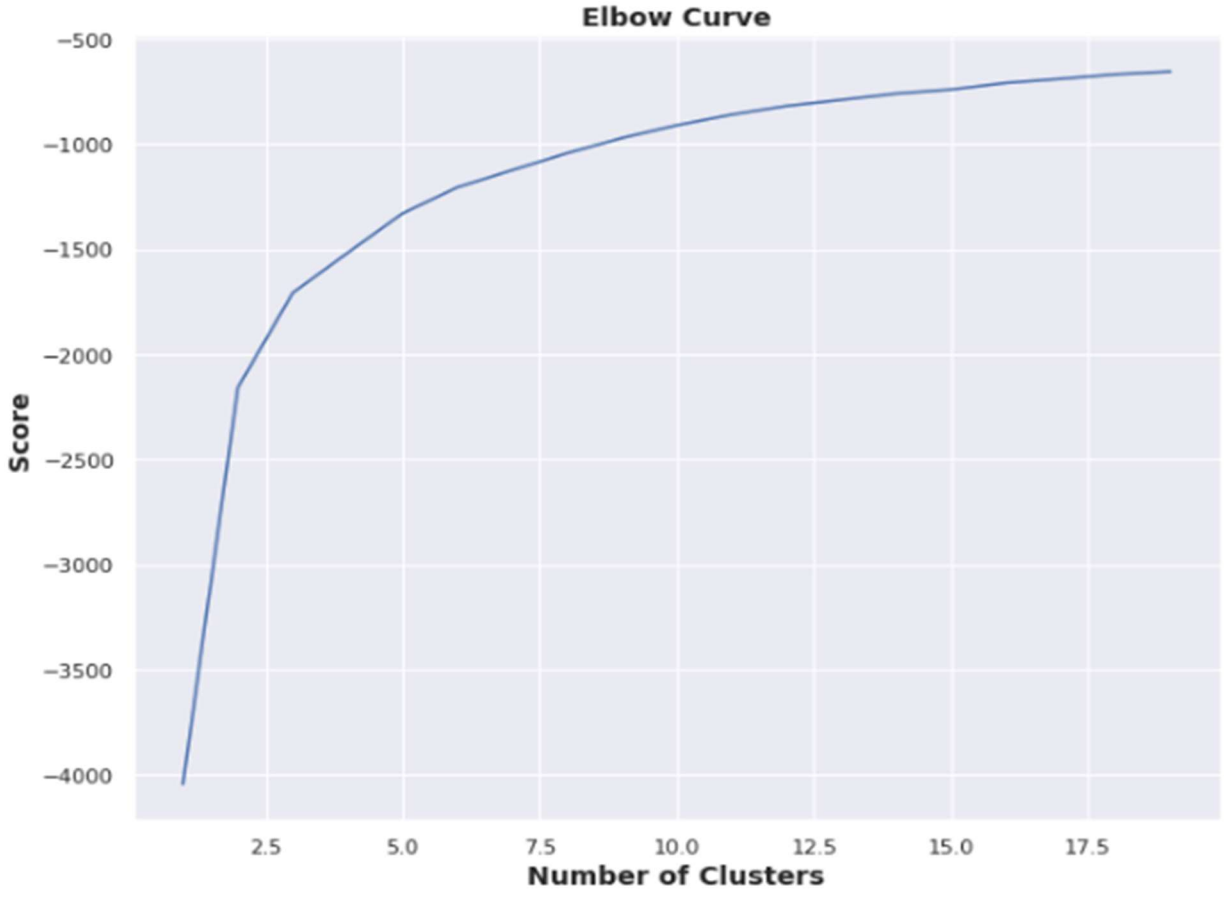
1. Document topic matrix
2. Topic term matrix

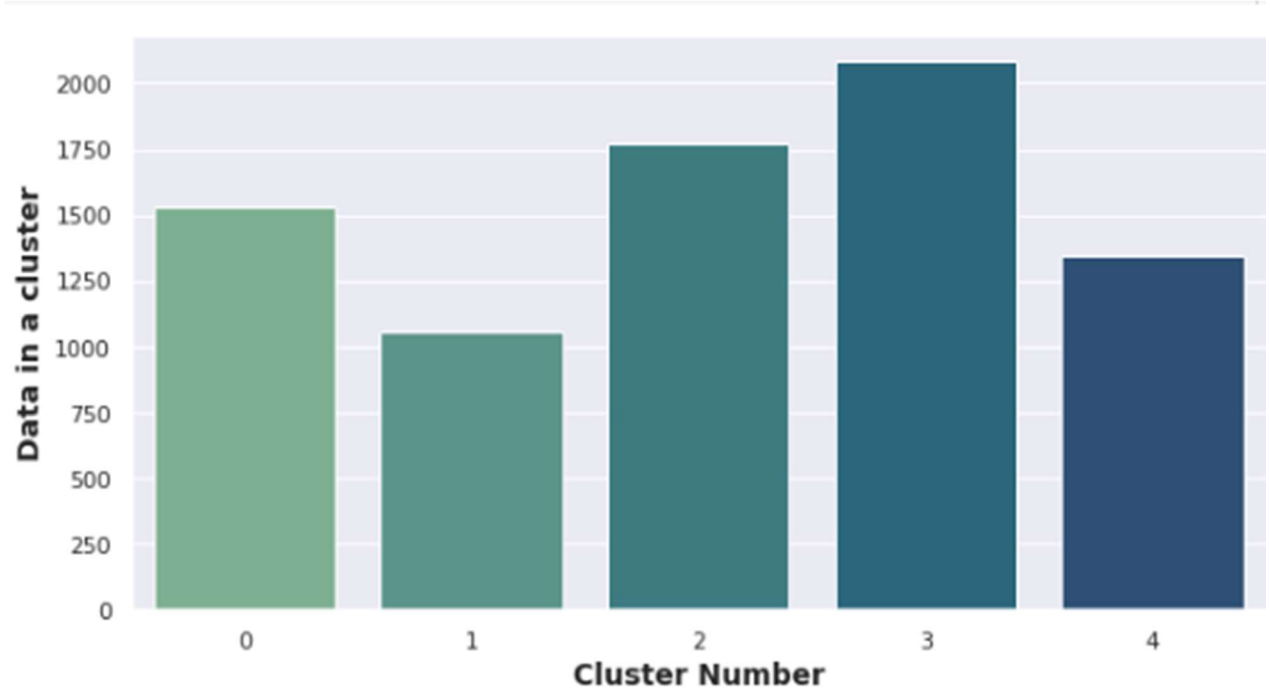
- **<u>Latent Dirichlet Allocation(LDA)</u>**

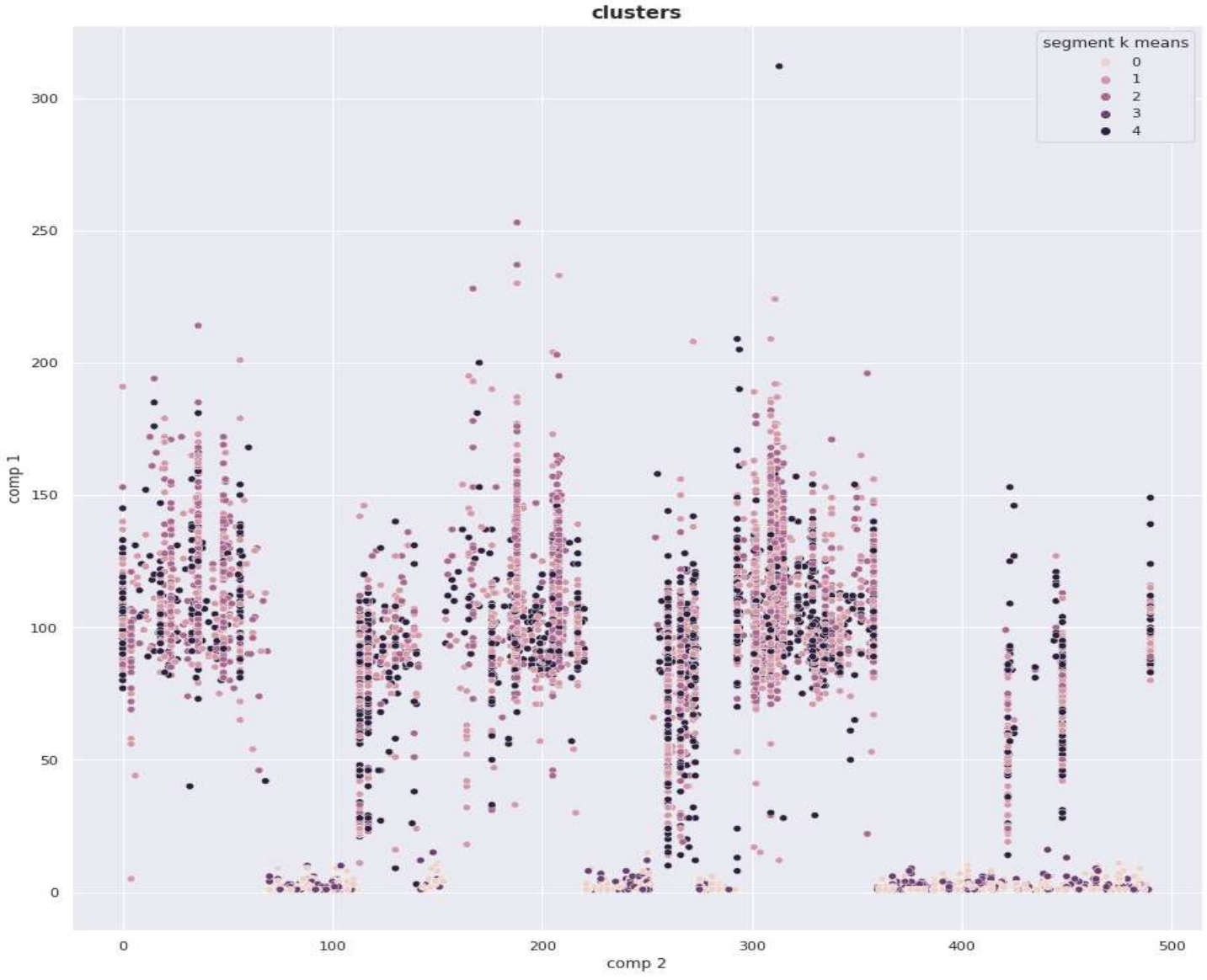
  LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

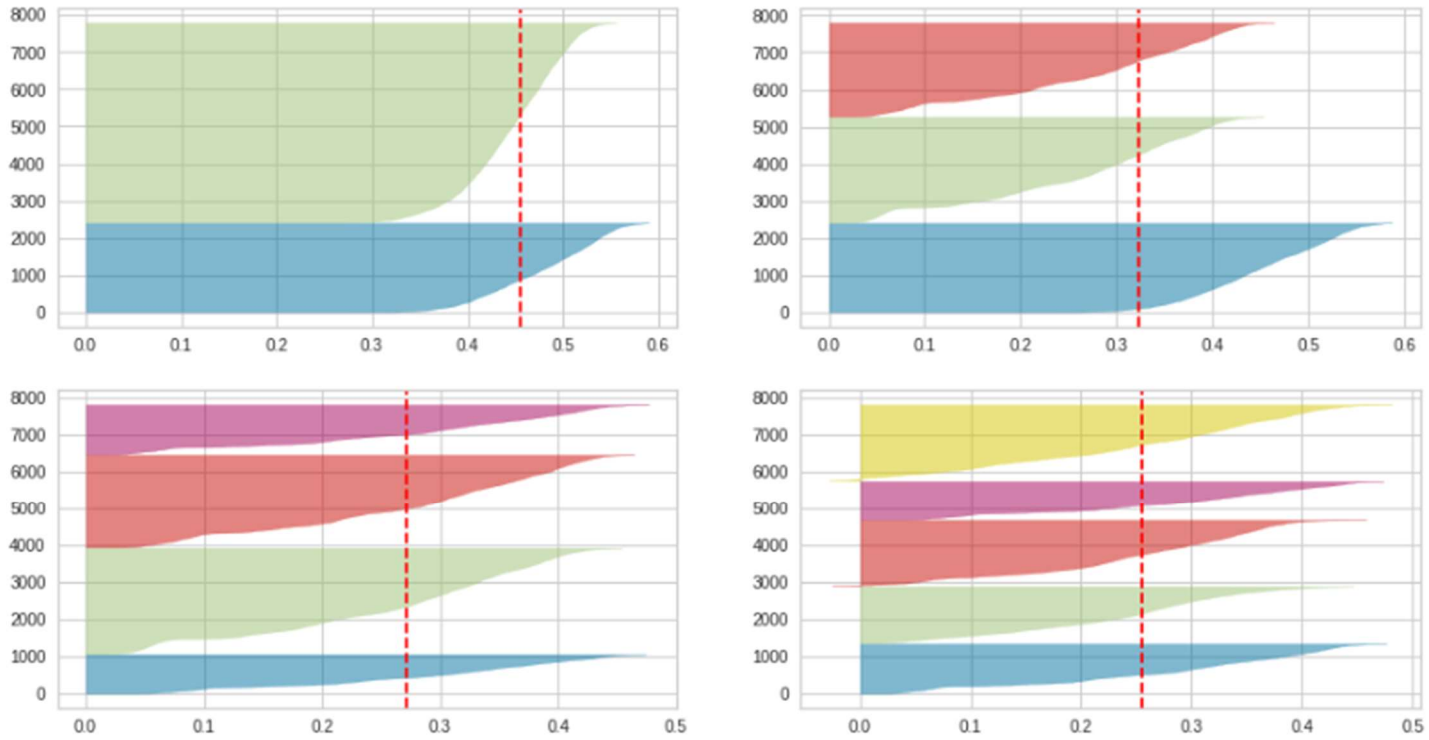
- <u>K-Means Clustering:</u>

  K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcome.




Data in a cluster
2000
1750
1500
1250
1000
750
500
250
0
0
1
2
3
4
Cluster Number


clusters
segment k means
0
1
2
3
4
comp 1
300
250
200
150
100
50
0
0
100
200
300
400
500
comp 2



CONCLUSION:

- From the above the analysis there are 70% movies and 30% TV shows.
- The United States has the 54% of content on Netflix then followed by India with 16%.
- LDA and LSA has sorted much more similar titles in a group of genre.
- Recommendation system works well with description column.
- After applying K-Means clustering the optimal value of number of cluster is 5.
- Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are