NETFLIX MOVIES AND TV SHOWS CLUSTERING

VINIT LADSE, PRATIKSHA KHARODE, GAURAV BHAKTE

CAPSTONE PROJECT-IV

ALMABETTER,BANGLORE

Abstract:

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 the OTT platforms 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 the dataset to get the insights from the information however the principal invalid qualities are taken care of. There are 12 features and around 7787 observations in the dataset and are mostly textual features. Clustering is a useful technique to achieve the best possible recommendations and increase the viewership of the platform.

Problem Statements:

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

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:

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. The dataset contains following 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 content

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

METHODOLOGY

A. Handling missing values :

We will need to replace blank countries with the mode (most common) country. It would be better to keep director because it can be fascinating to look at a specific filmmaker's movie. As a

result, we substitute the null values with the word 'unknown' for further analysis.

There are very few null entries in the date_added fieldsthus we delete them.

B. Duplicate Values Treatment:

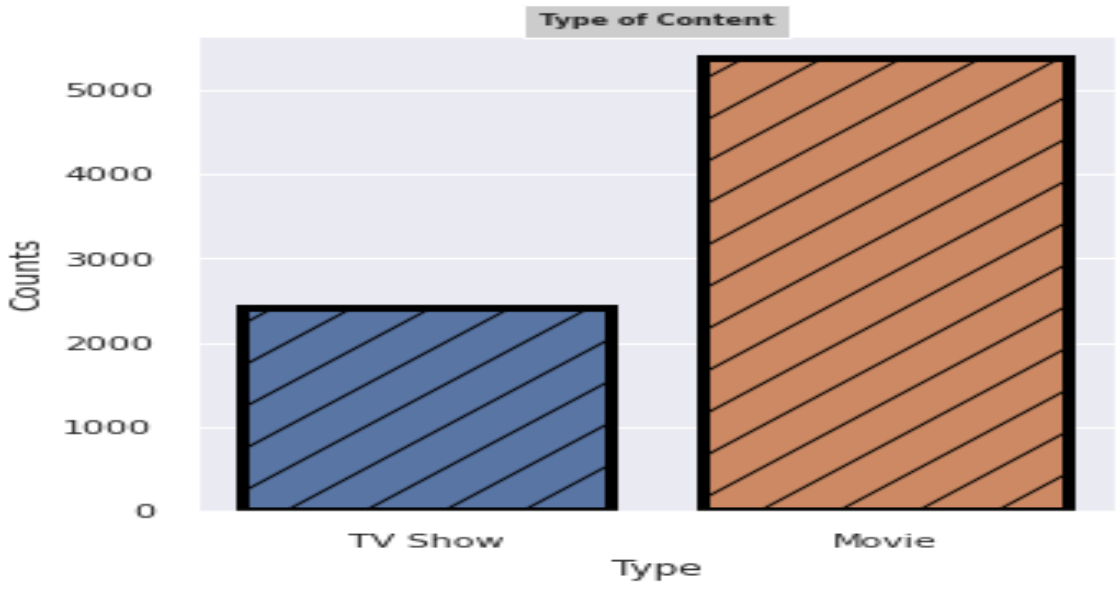
Duplicate values dose not contribute anything to accuracy of results. Our dataset dose not contains any duplicate values.

C. Exploratory Data Analysis:

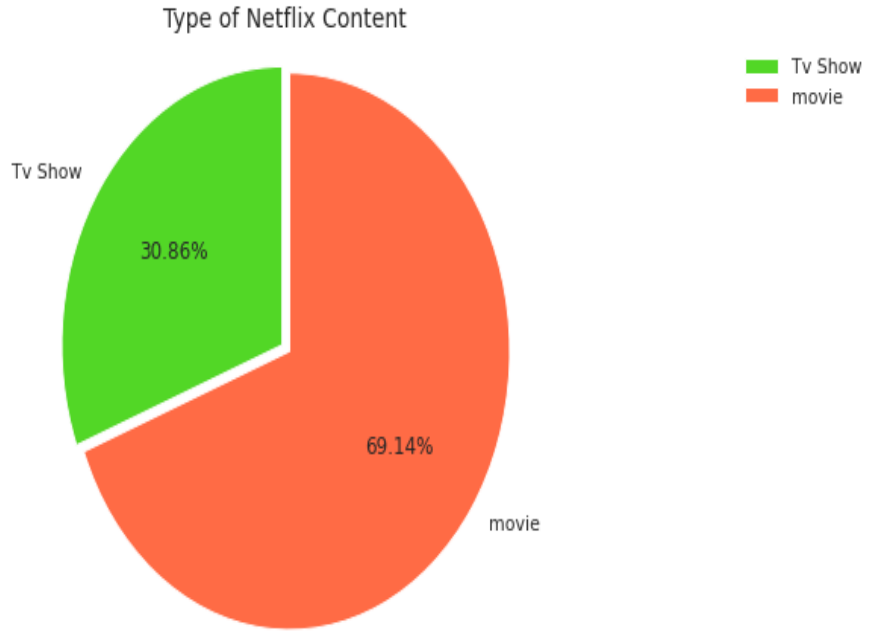
After mounting our drive and fetching and reading the dataset given, we performed the Exploratory Data Analysis for it.

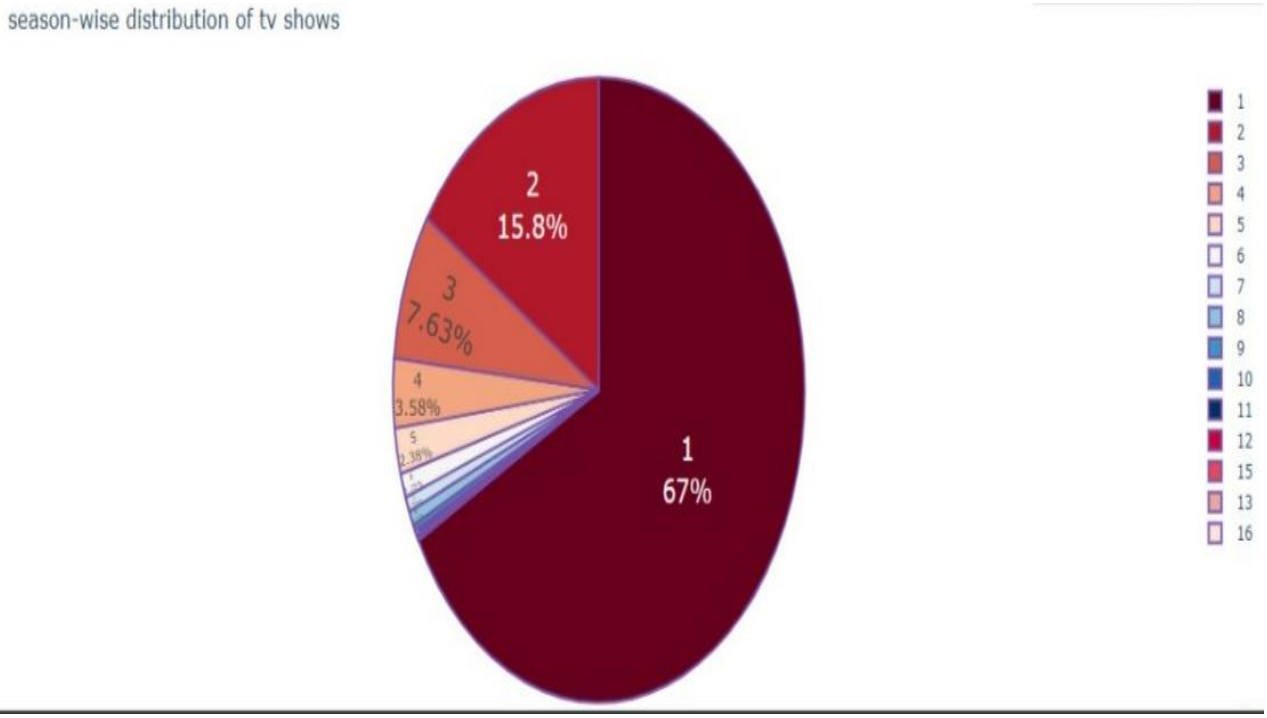
To get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step.

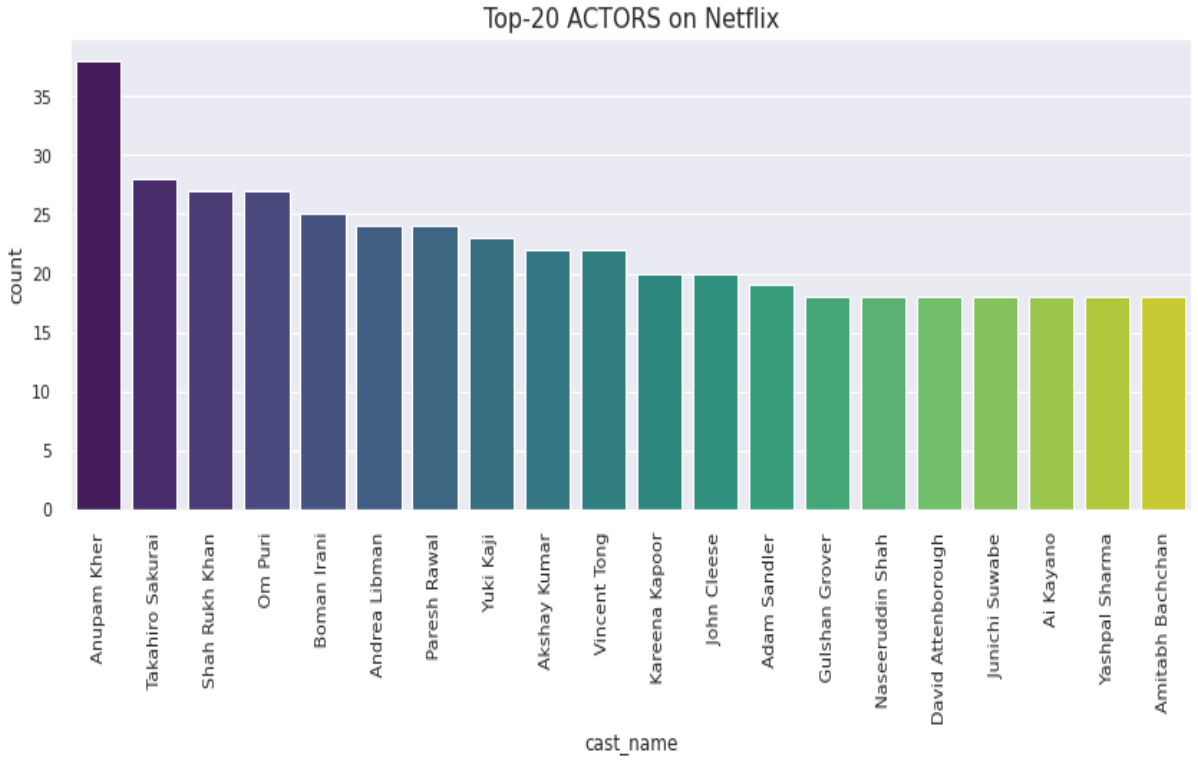
The United States is the most prolific generator of Netflix content, with India and the United Kingdom trailing far behind.

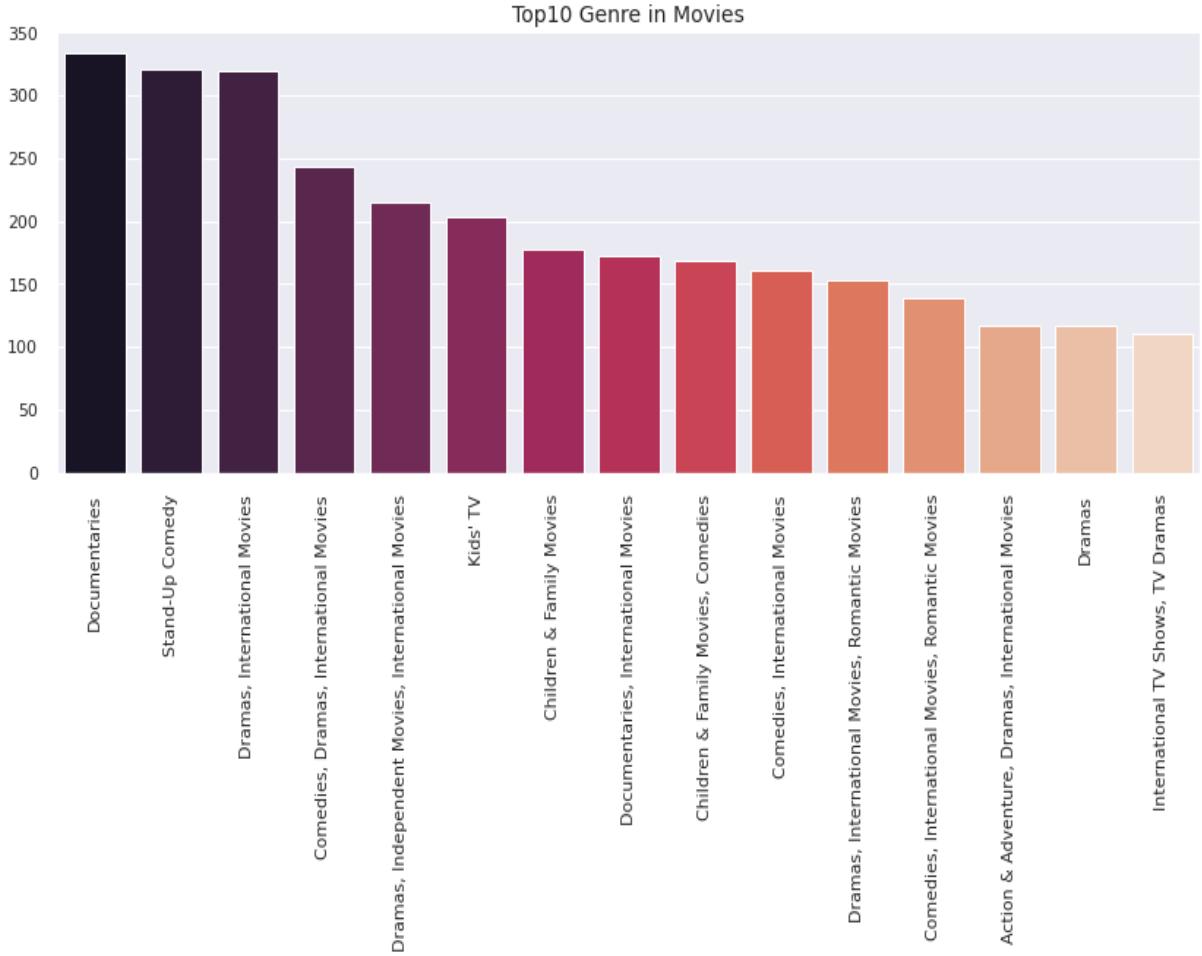


Types of Netflix content:

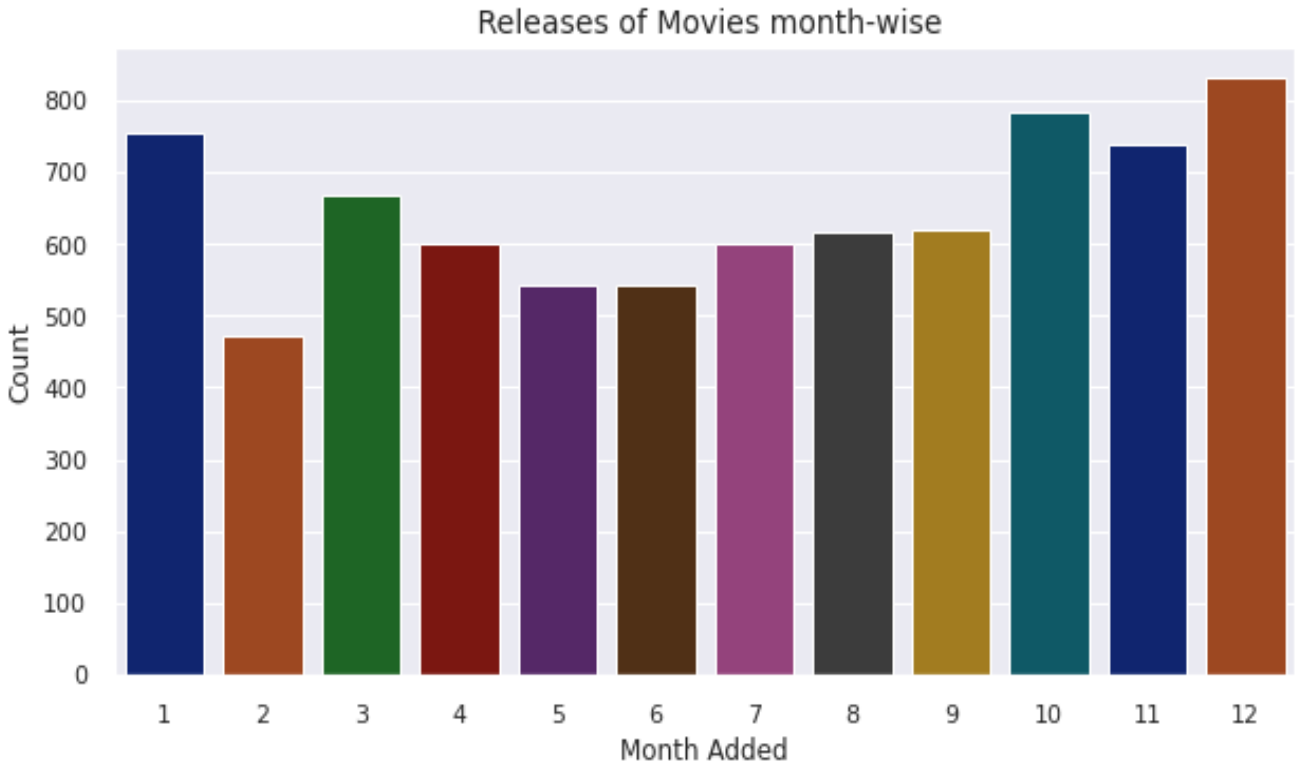




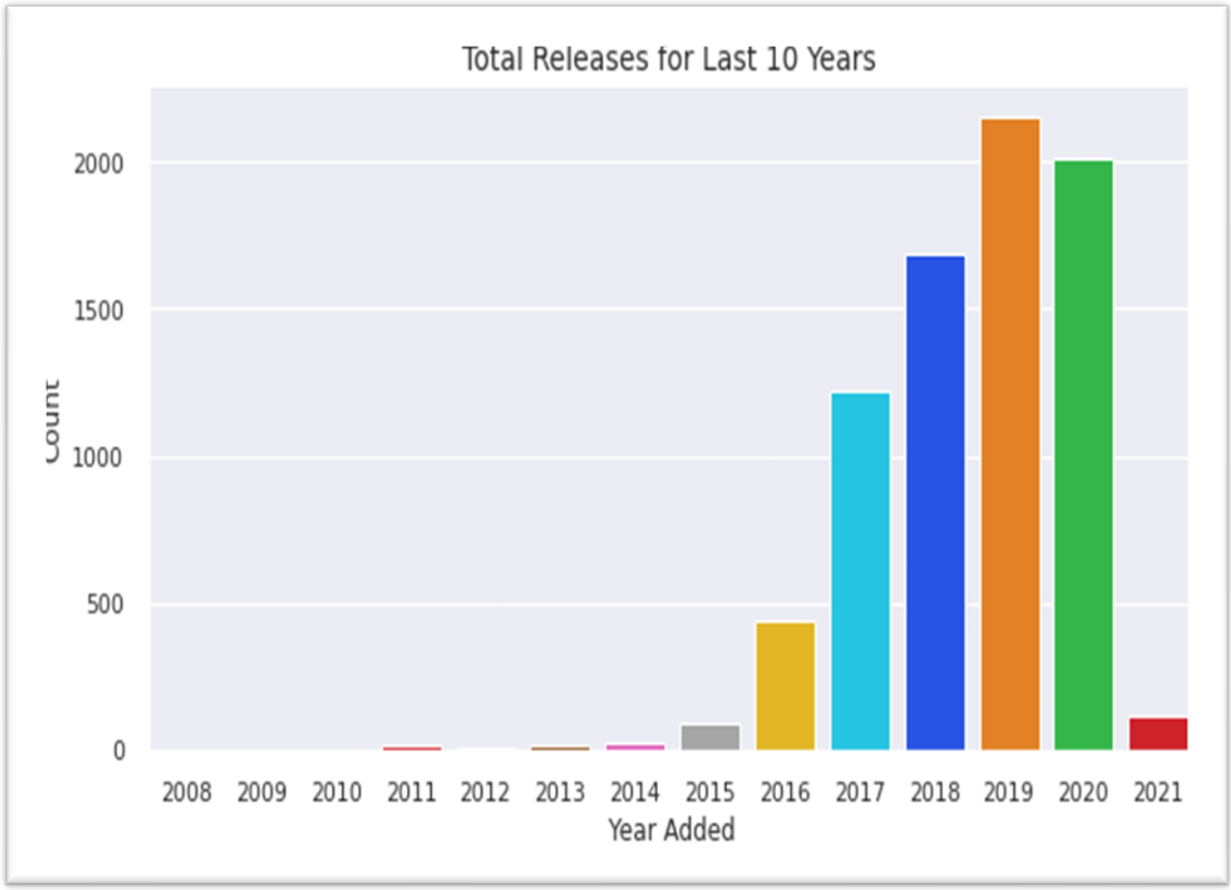

Top-20 ACTORS on Netflix
count
0
5
10
15
20
25
30
35
Anupam Kher
Takahiro Sakurai
Shah Rukh Khan
Om Puri
Boman Irani
Andrea Libman
Paresh Rawal
Yuki Kaji
Akshay Kumar
Vincent Tong
Kareena Kapoor
John Cleese
Adam Sandler
Gulshan Grover
Naseeruddin Shah
David Attenborough
Junichi Suwabe
Ai Kayano
Yashpal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

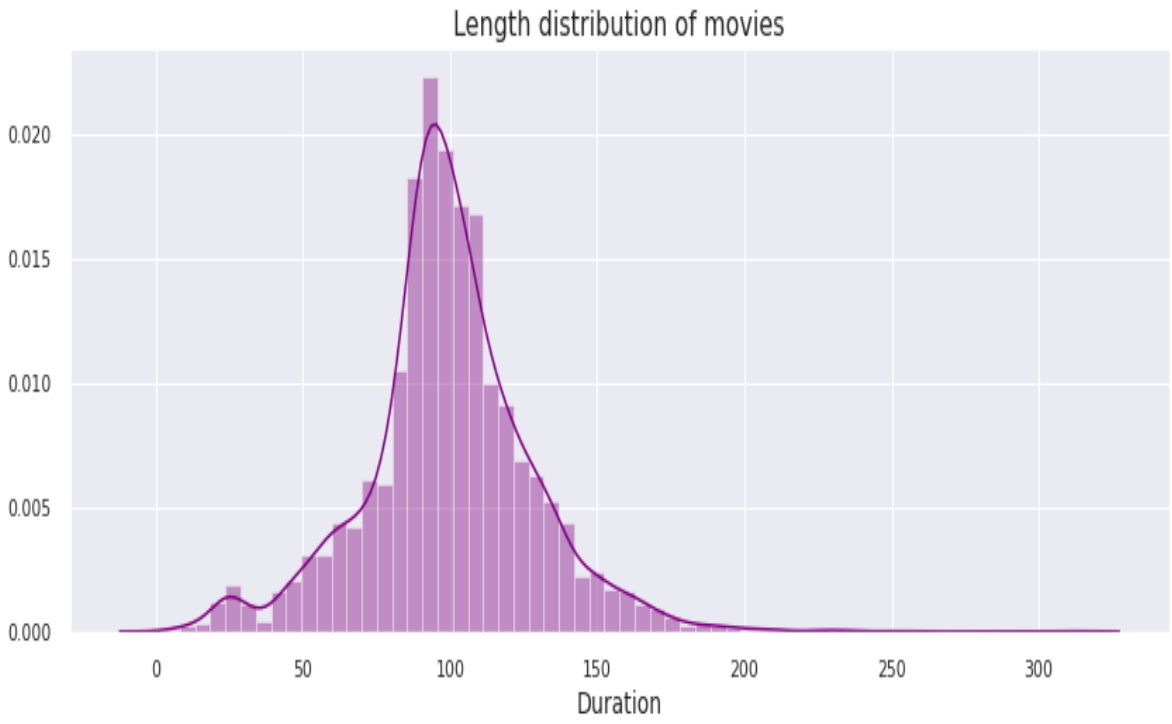

Top10 Genre in Movies
0
50
100
150
200
250
300
350
Documentaries
Stand-Up Comedy
Dramas, International Movies
Comedies, Dramas, International Movies
Dramas, Independent Movies, International Movies
Kids' TV
Children & Family Movies
Documentaries, International Movies
Children & Family Movies, Comedies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

Month Wise Analysis:



Year Wise Analysis:





D. Data Preprocessing :

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

Removing stop-words : 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.

Stemming : Stemming is the process of removing a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

Clustering:

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

for example:

• **Data analysis:** often a huge dataset contains several large clusters, analyzing which separately, you can come to interesting insights.

• **Anomaly detection:** as we saw before, data points located in the regions of low density can be considered as anomalies

• **Semi-supervised learning:** clustering approaches often helps you to automatically label partially labeled data for classification tasks.

• **Indirectly clustering tasks (tasks where clustering helps to gain good results):** recommender systems, search engines, etc.

• **Directly clustering tasks**: customer segmentation, image segmentation, etc .

Topic Modeling:

- **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.

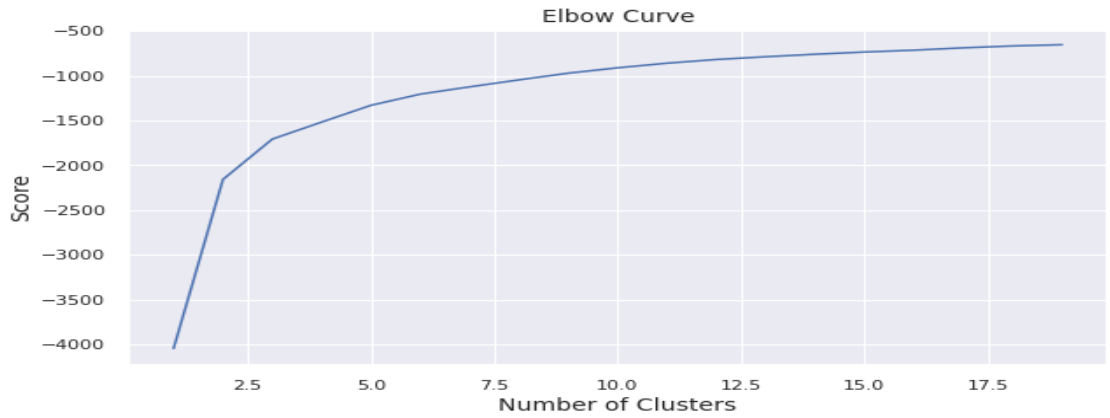
Therefore, the learning of LSA for latent topics includes matrix decomposition on the document-term matrix using Singular value decomposition. It is typically used as a dimension reduction or noise-reducing technique.

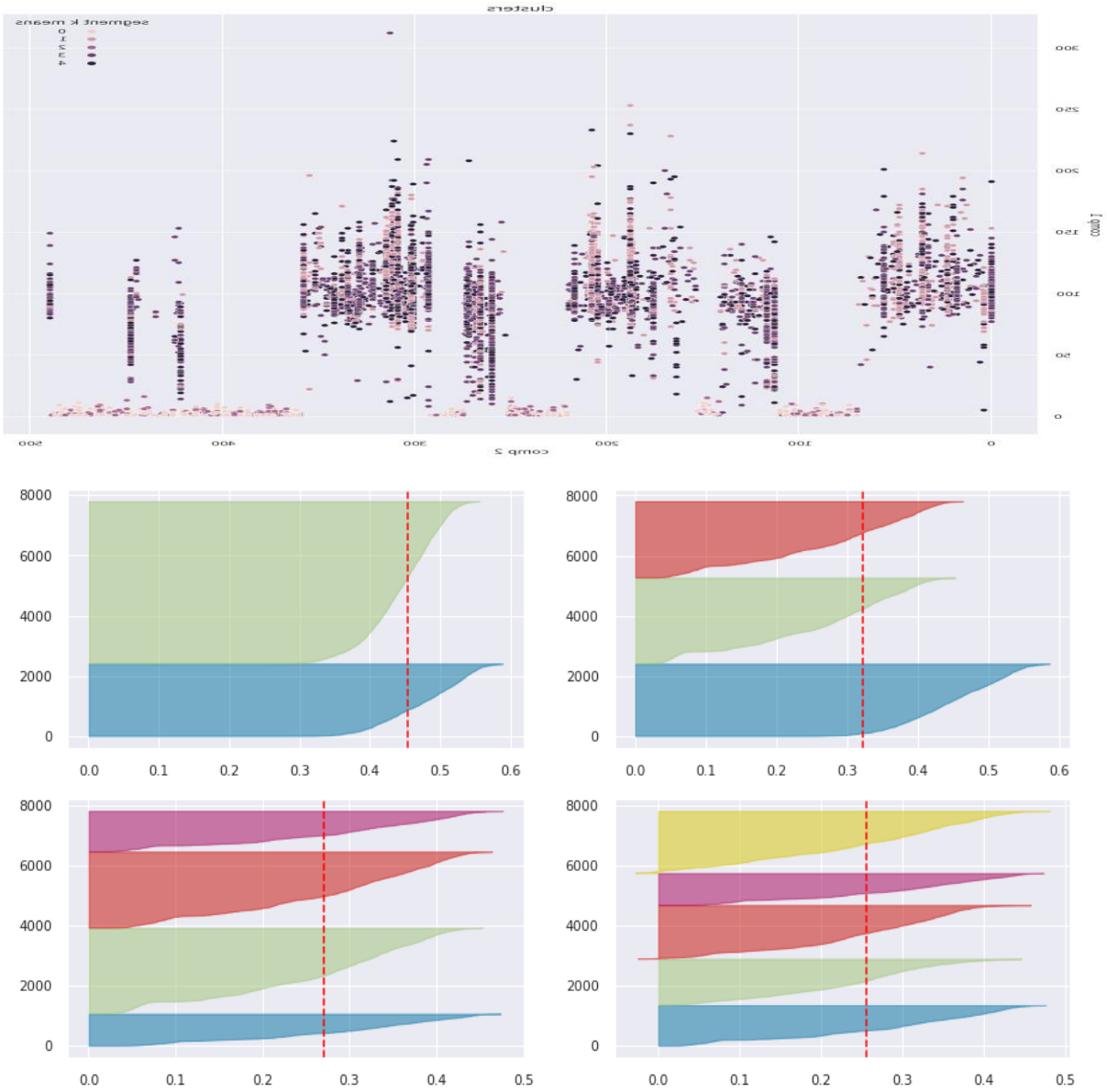
- **Latent Dirichlet Allocation (LDA)**

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.

E. K- means Clustering:

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, outcomes.





Conclusion:

- There are about 70% movies and 30% TV shows on Netflix.
- The United States has the highest number of content on Netflix by a huge margin followed by India.
- LDA and LSA has sorted much more similar titles in a group of genre.
- Recommendation system works perfectly well with description column.
- 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.