Clustering Analysis on Netflix movies and Tv shows

Sahil pardeshi , kirtesh verma , pravin bejjo
Data science, Trainee Almabetter nashik

❖ Introduction :-

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device. You can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection.
The dataset is collected from Flexible which is a third-party Netflix search engine.
It has 7787 entries and 12 attributes.

Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions, known as Netflix Originals.

As of March 31, 2022, Netflix had over 221.6 million subscribers worldwide, including 74.6 million in the United States and Canada, 74.0 million in Europe, the Middle East and Africa, 39.9 million in Latin America and 32.7 million in Asia-Pacific. It is available worldwide aside from Mainland China, Syria, North Korea, and Russia. Netflix has played a prominent role in independent film distribution, and it is a member of the Motion Picture Association (MPA).

Netflix was founded on the aforementioned date by Reed Hastings and Marc Randolph in Scotts Valley, California. Netflix initially both sold and rented DVDs by mail, but the sales were eliminated within a year to focus on the DVD rental business.In 2007, Netflix introduced streaming media and video on demand. The company expanded to Canada in 2010, followed by Latin America and the Caribbean. Netflix entered the content-production industry in 2013, debuting its first series *House of Cards*. In January 2016, it expanded to an additional 130 countries and then operated in 190 countries.

We have used featured engineering columns for better intuition and followed data preprocessing steps to make data read for model building. We have converted the data type of certain columns. We have used it by handling null values. Maximum null values were present in the ‘director’ and ‘cast’ columns, as this are not important for model building,we dropped them.

We have used various methods like data visualization, data understanding, k - mean clustering , TF - IDF vectorizer , feature engineering , Handling null and missing values , Elbow method , Dendogram , performing EDA, Data preprocessing , Silhouette score etc.

➢ Problem Statement :- This datasets consists of Tv shows and movies shows available on Netflix as of 2019. The datasets 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 been decreased by more than 2000 title’s since 2010, while its number of Tv shows has nearly tripled.

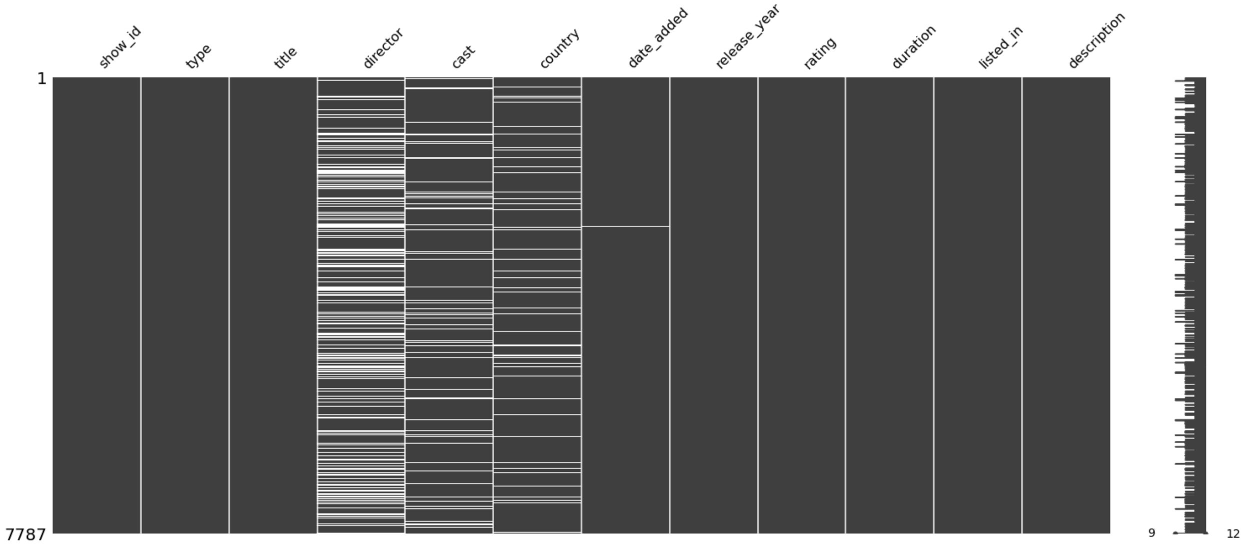
- ➢ **In this project, you are required to do**
    1) **EDA(exploratory data analysis)**
    2) **Understanding what type of content is available in different countries.**
    3) **Netflix has increasingly focused on Tv rather than movies in recent years.**
    4) **Clustering similar content by matching text - based features.**

- ➢ **Data set description**

The data set has 7787 rows and 12 attributes to work with.

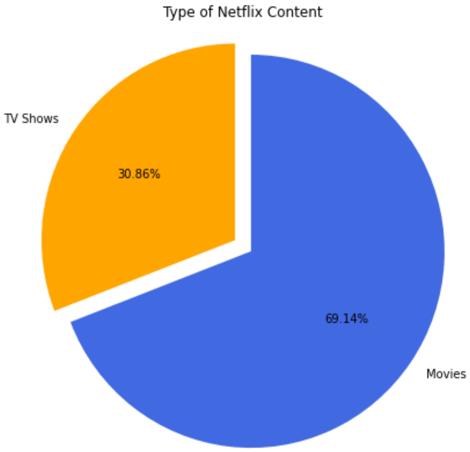
- ➢ **Steps involved performing EDA and Data preprocessing**

1) **Exploring head and tail of the data to get insight of the given data.**
2) **Checking Null values or missing values present in the dataset or not.**
3) **Checking duplicates values.**
4) **Creating dataframes which helps in drawing insights from the datasets.**

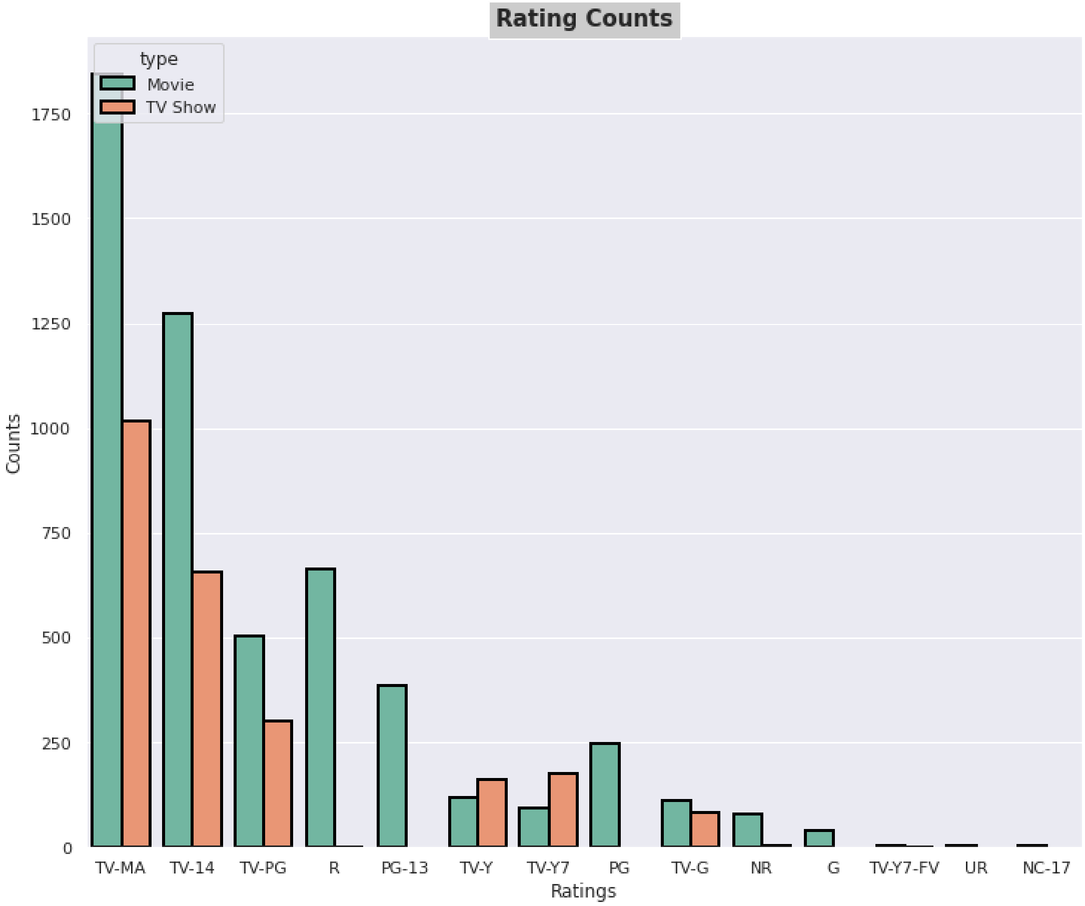


➢ **Drawing conclusion from the Data.**

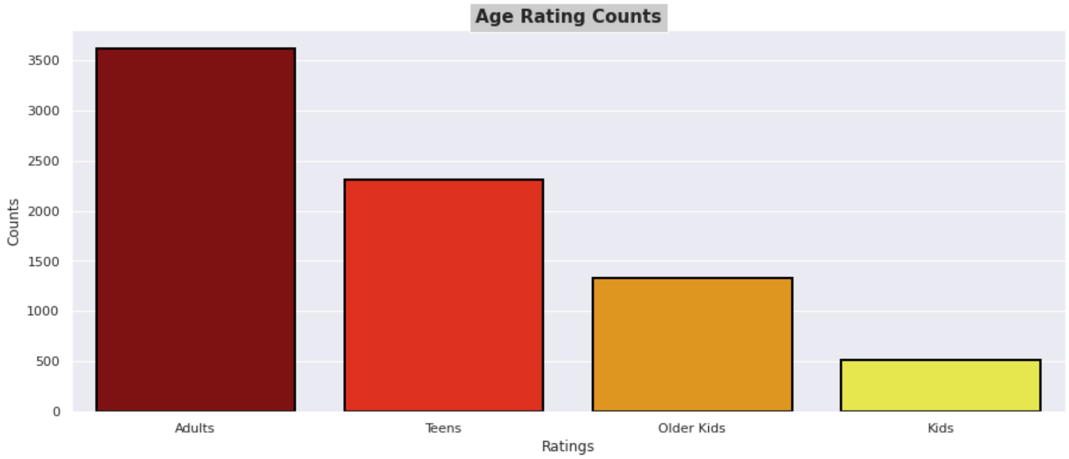
1) **Type of content on Netflix EDA on rating analysis on netflix movies and Tv shows**



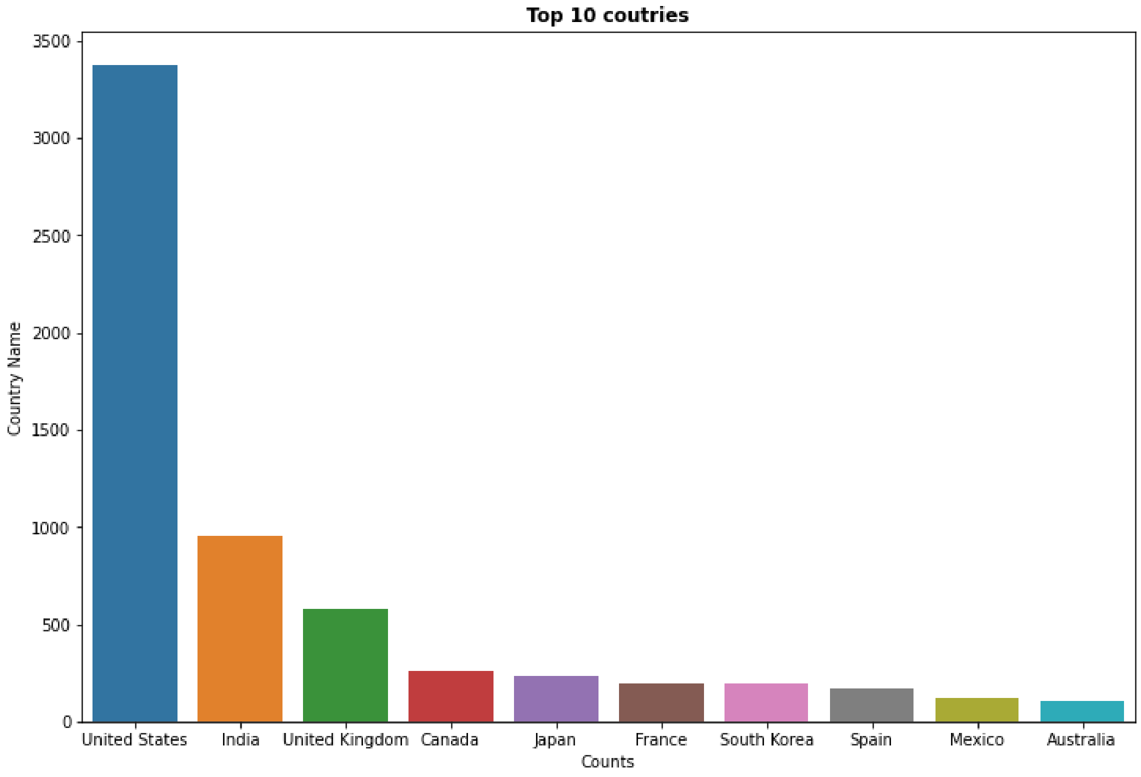
2)EDA on rating analysis on netflix movies and Tv shows



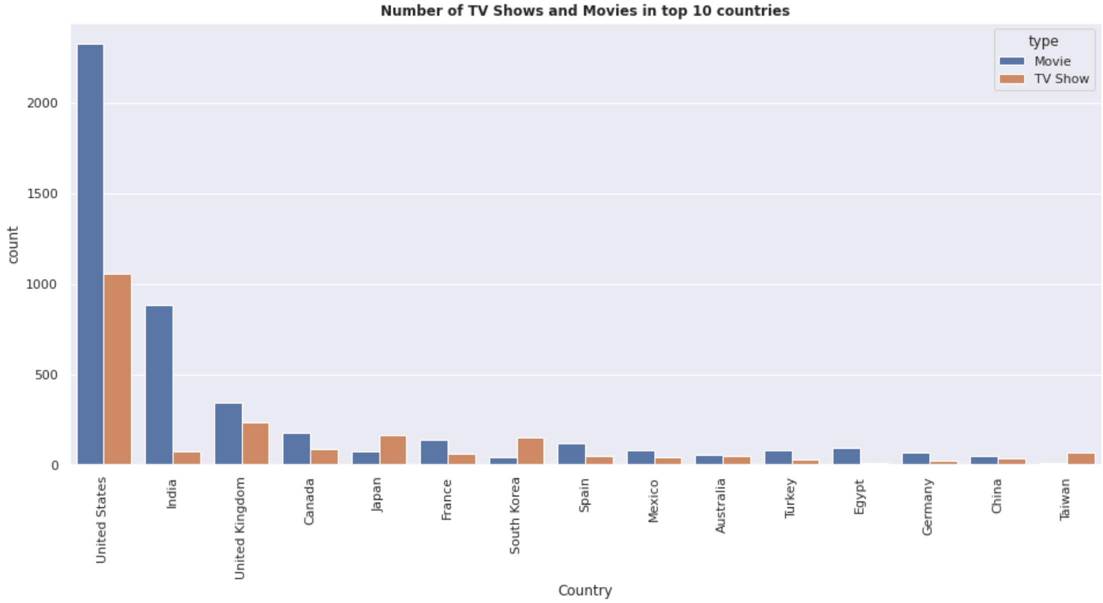
3) EDA on age rating count



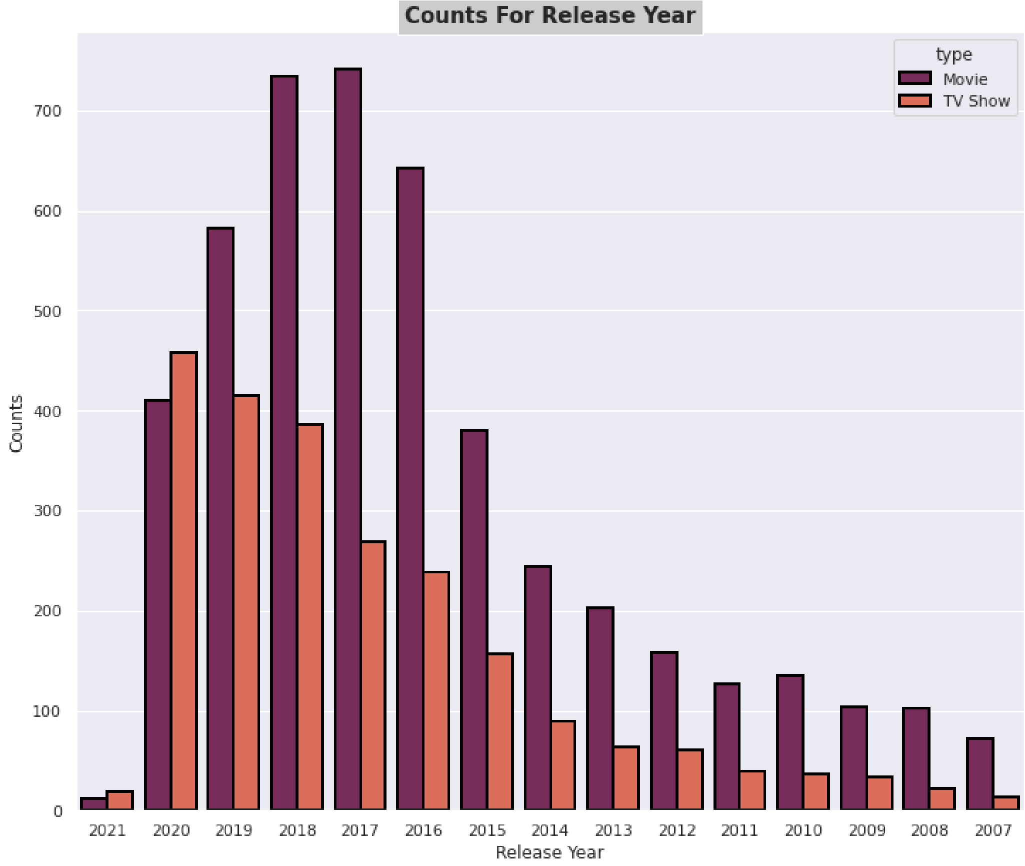
4)EDA on top 10 countries on Netflix



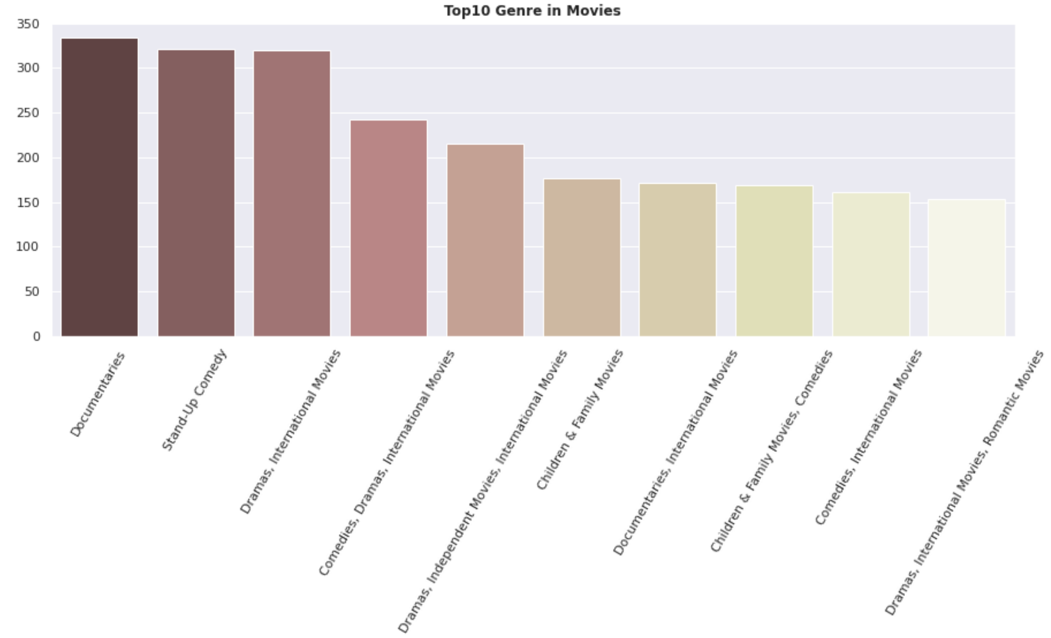
5) EDA on number of TV shows and movies content in top 10 countries with maximum content



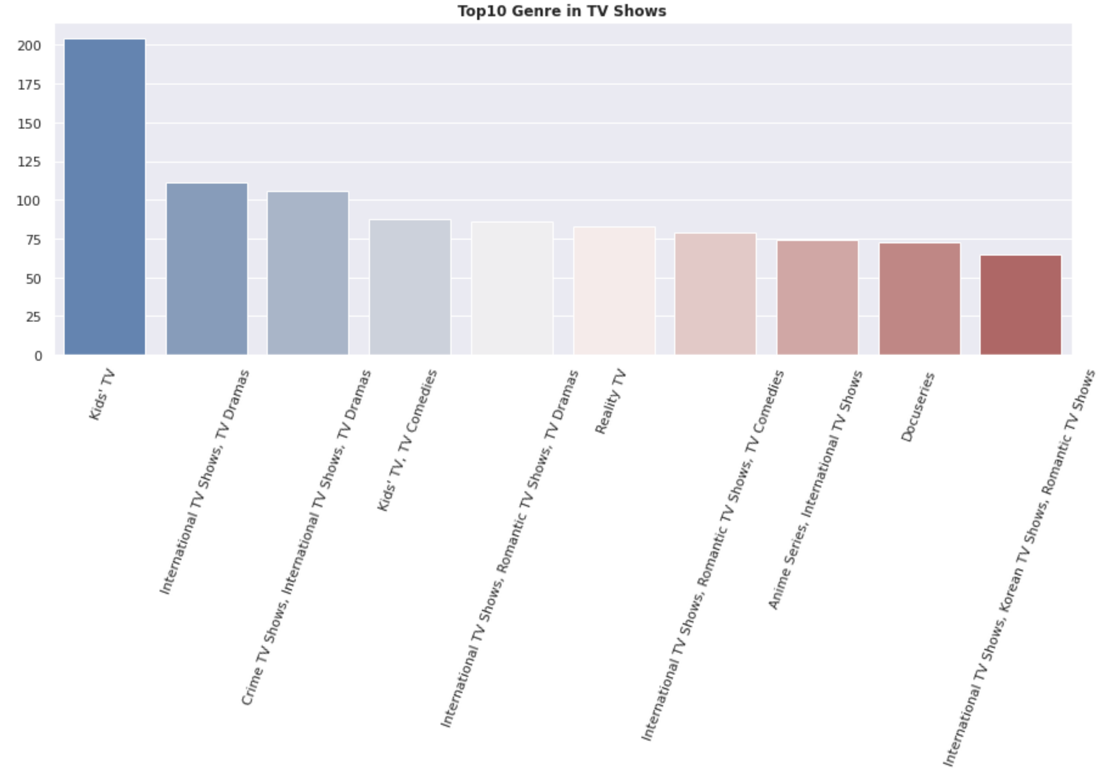
6) plotting the count plot for release year analysis



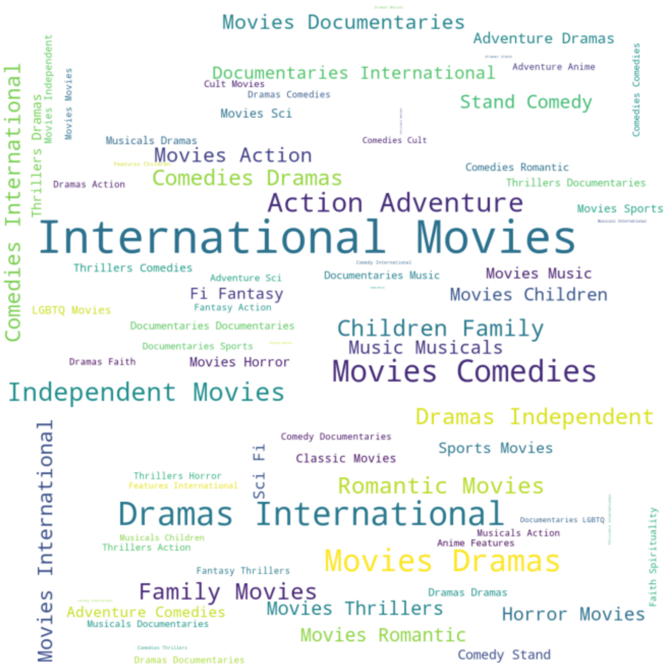
7) Top 10 genres in movies



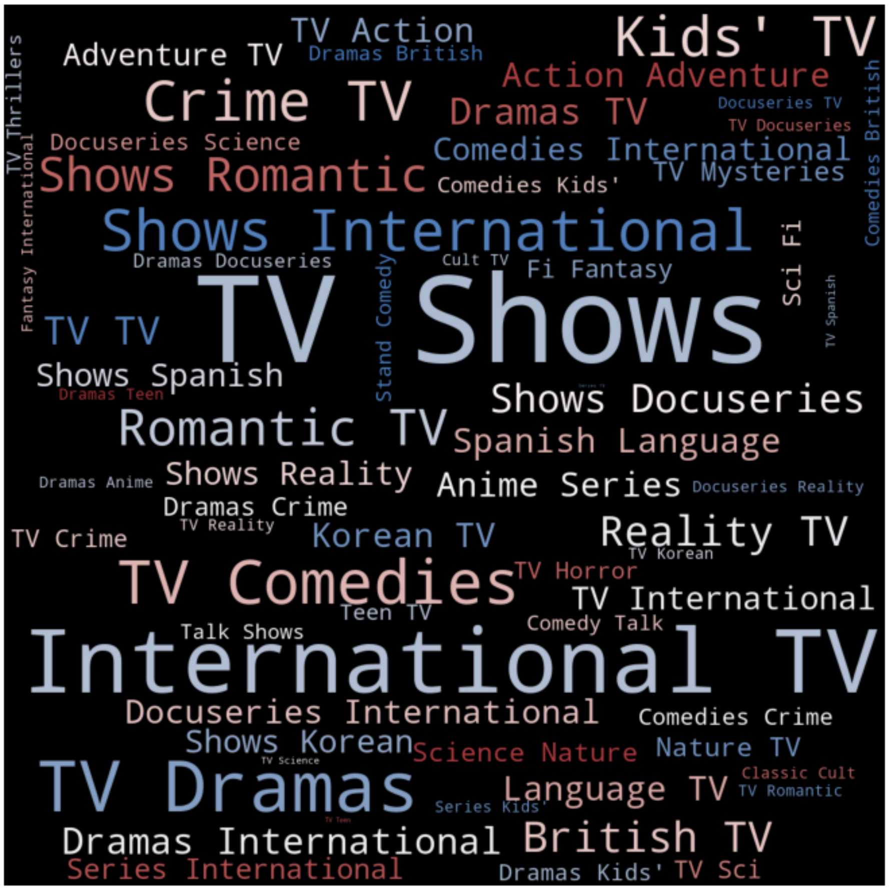
8) Top 10 genres in TV shows



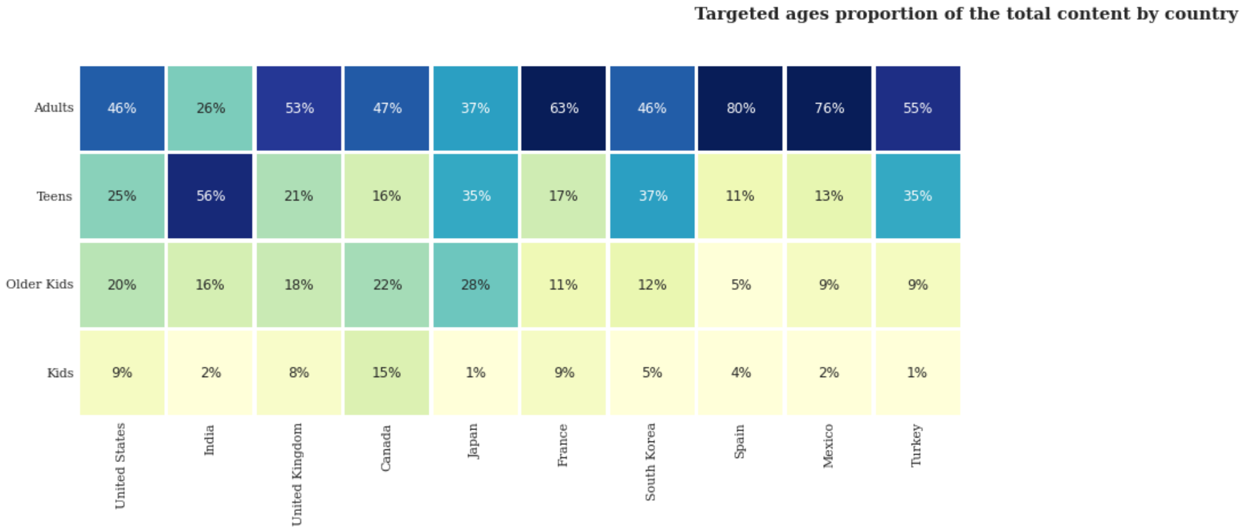
9) Word clouds for movies



10) word clouds for TV shows



11) plotting heat map



Targeted ages proportion of the total content by country

- **Method’s which we used:**
  Data understanding, data visualization, K - mean clustering, Elbow method, silhouette score, Dendogram , data preprocessing etc.

- **Scaling the data:**

We have used the standard scale method to scale the datasets.

- **Building the clustering model**

Clustering models allow you to categorize records into a certain number of clusters. This can help you identify natural groups in your data. Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong.

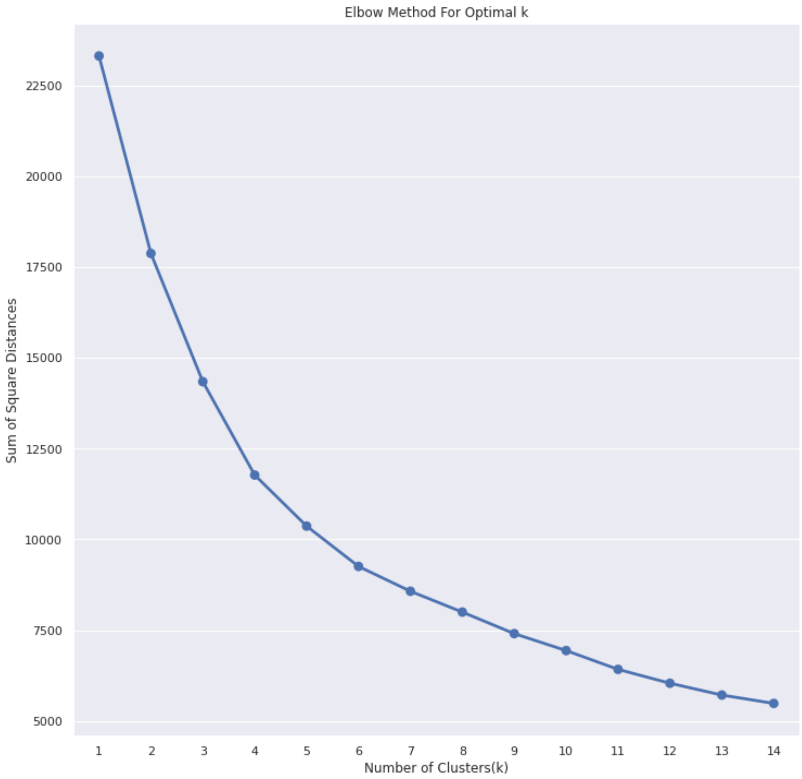
- **Metric used Silhouette coefficient and silhouette score:**

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. 1: Means clusters are well apart from each other and clearly distinguished.
The value of the silhouette coefficient is between [-1, 1]. A score of 1 denotes the best meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1. Values near 0 denote overlapping clusters.

- ➢ **Model Implementation:**

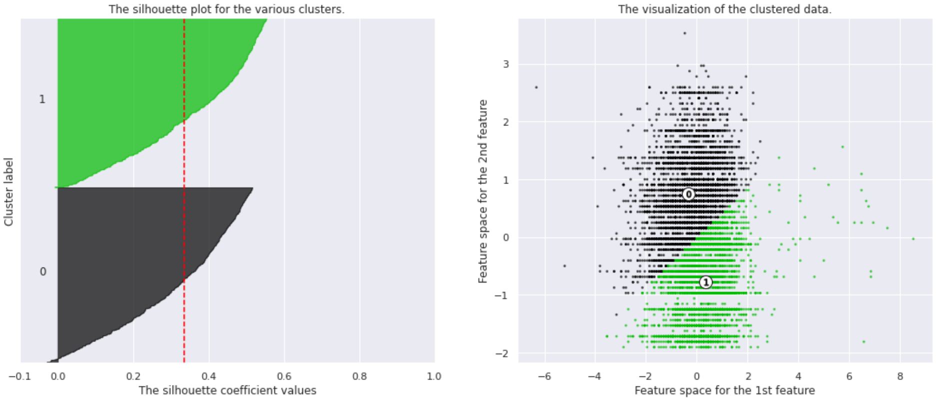
1) **K-Mean clustering : k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters.K-means clustering algorithm computes the centroids and iterates until it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.**

2) **Elbow Method: In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.WCSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1.**



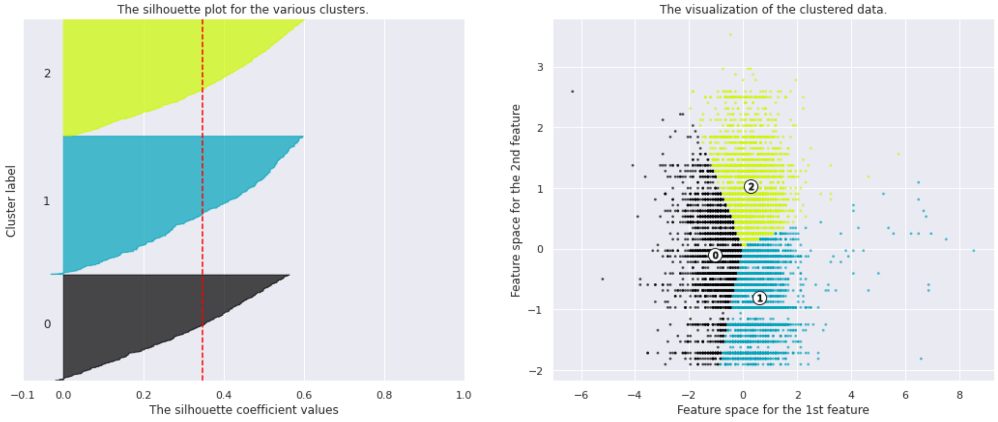
3) Silhouette score and visualization:

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2


The silhouette plot for the various clusters.

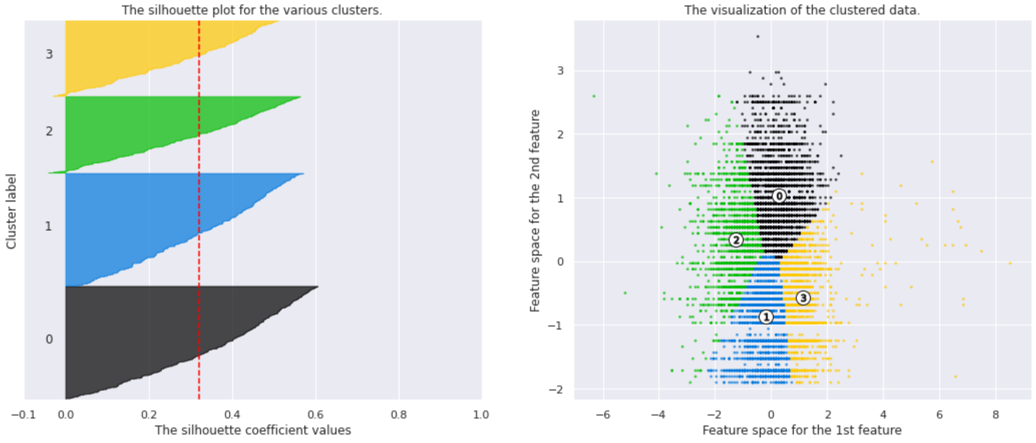
The visualization of the clustered data.

Silhouette analysis for KMeans clustering on sample data with n_clusters = 3


The silhouette plot for the various clusters.

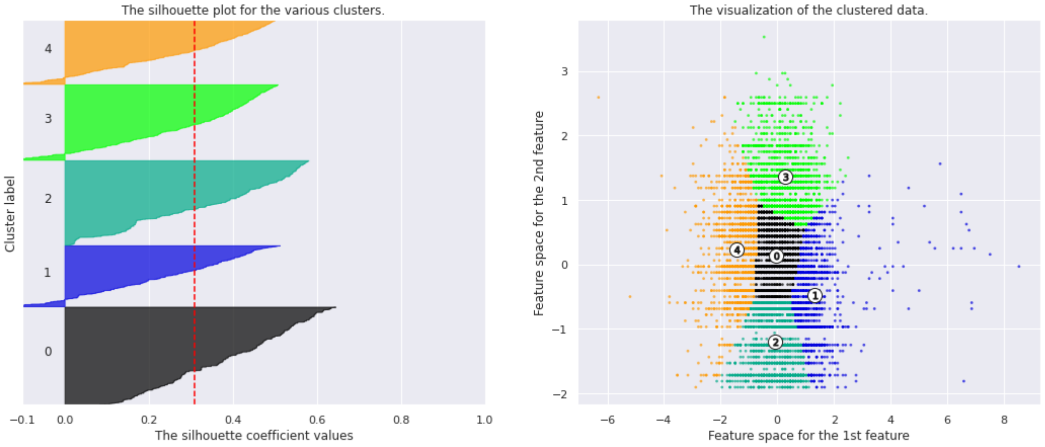
The visualization of the clustered data.

Silhouette analysis for KMeans clustering on sample data with n_clusters = 4


The silhouette plot for the various clusters.

The visualization of the clustered data.

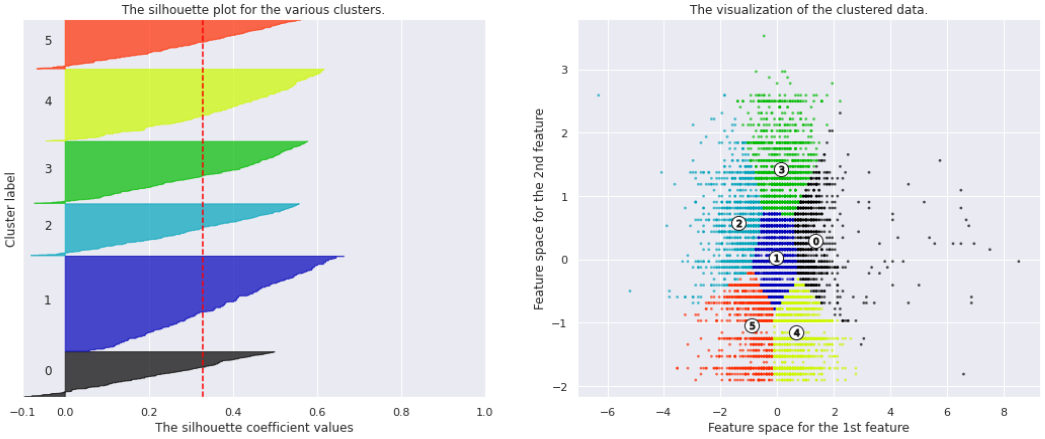
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



The silhouette plot for the various clusters.

The visualization of the clustered data.

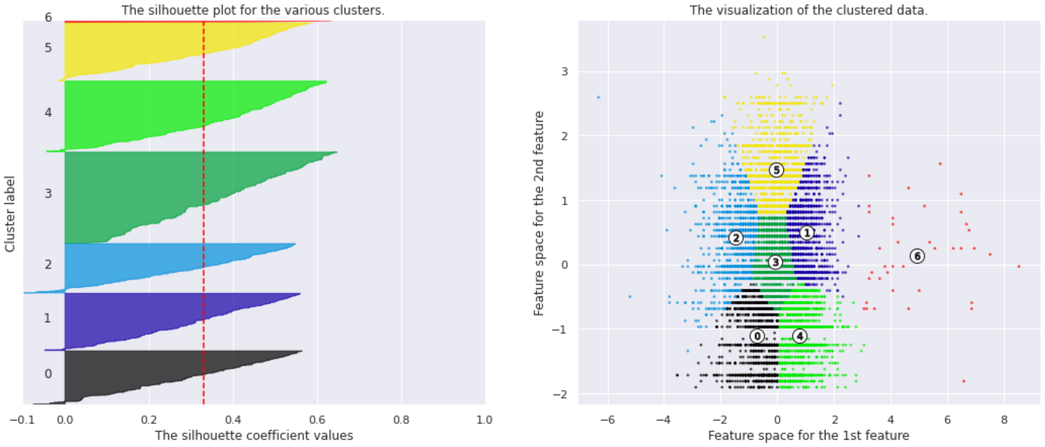
Silhouette analysis for KMeans clustering on sample data with n_clusters = 6



The silhouette plot for the various clusters.

The visualization of the clustered data.

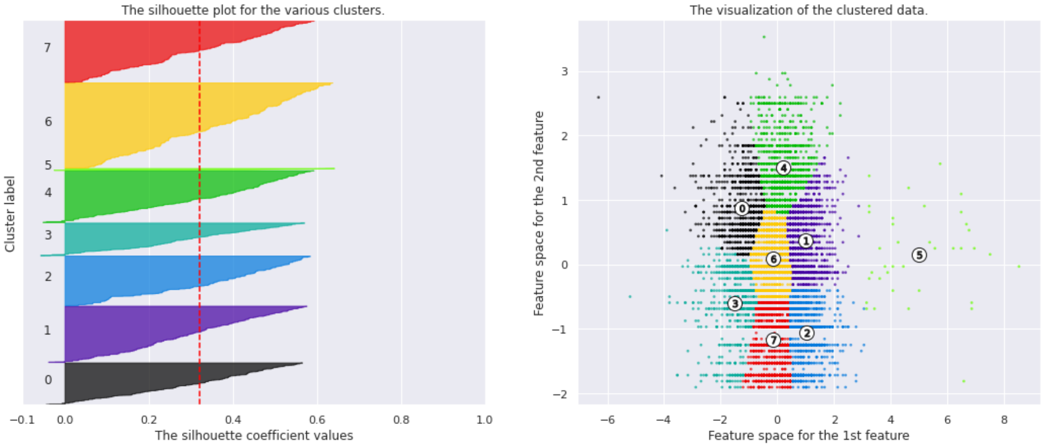
Silhouette analysis for KMeans clustering on sample data with n_clusters = 7



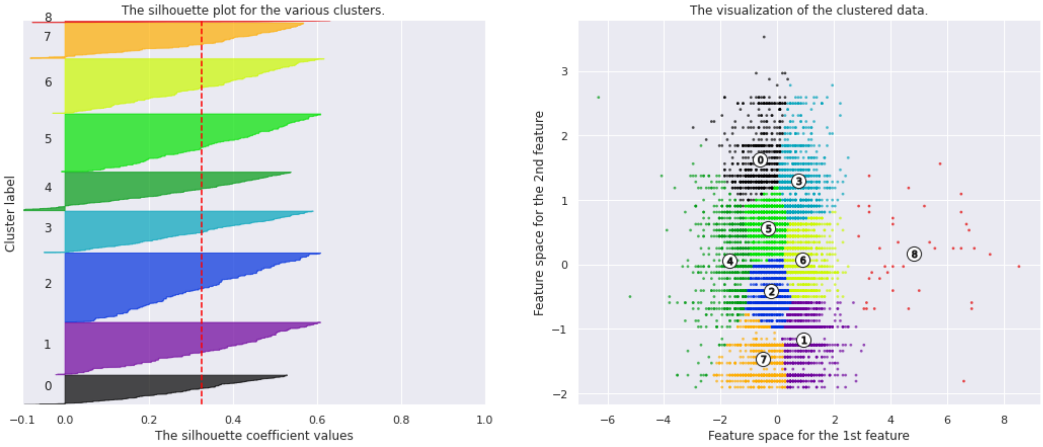
The silhouette plot for the various clusters.

The visualization of the clustered data.

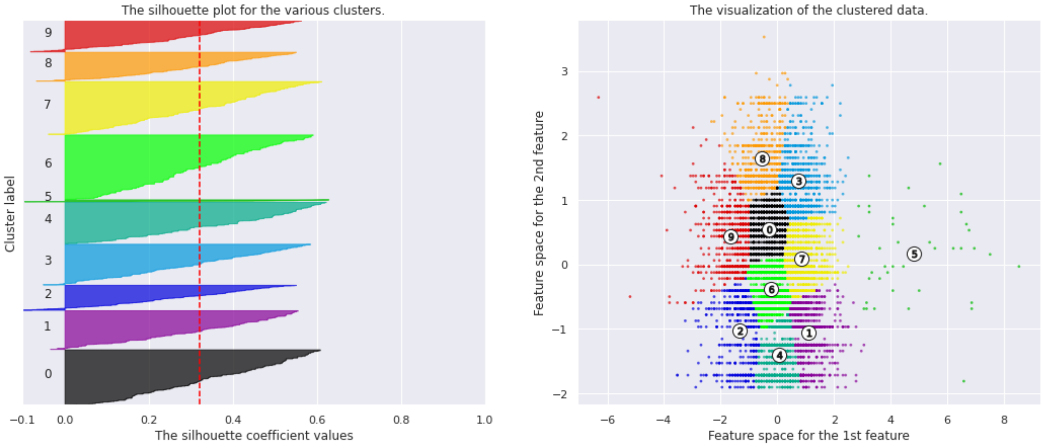
Silhouette analysis for KMeans clustering on sample data with n_clusters = 8



Silhouette analysis for KMeans clustering on sample data with n_clusters = 9



Silhouette analysis for KMeans clustering on sample data with n_clusters = 10



Silhouette clusters values:

```
For n_clusters = 2, silhouette score is 0.3367119899886218
For n_clusters = 3, silhouette score is 0.3486376954616092
For n_clusters = 4, silhouette score is 0.317329701195747
For n_clusters = 5, silhouette score is 0.30776212221665256
For n_clusters = 6, silhouette score is 0.32796114387833786
For n_clusters = 7, silhouette score is 0.3255282169856485
For n_clusters = 8, silhouette score is 0.32058273307230845
For n_clusters = 9, silhouette score is 0.3219115069956129
For n_clusters = 10, silhouette score is 0.3231660286685175
```

- The Silhouette ranges from -1 to +1, where the high values indicates that the object is well matched to its own clusters and poorly match to its neighbors clusters.

➢ **Challenges we faced :**

1) **Preprocessing data was one of the big challenge we faced which includes handling missing values and filling missing values.**
2) **Features engineering.**
3) **Removing punctuation and removing stopwords.**
4) **Model implementation.**

➢ **Conclusion:**

1) **Netflix is most popular in U.S .India lie in 2nd position in popularity list.**
2) **In most of the countries content available on the netflix is mostly on movies type except in south korea and japan.**
3) **Clustering was done during ‘length’ listed and ‘type’ columns.**
4) **Netflix has increasingly focus on movies than TV shows. It has been producing more movies than TV shows since 2014.**