CLUSTERING ANALYSIS ON NETFLIX MOVIES AND TV SHOWS

Kirtesh Verma, Pravin Bejjo, and
Sahil Pardeshi
Data Science Trainees,
Almabetter, Nashik

➢ Introduction :

Netflix is an American subscription streaming service and production company. It is the one of the largest Platform which provides the collection of TV shows and movies, streaming via online means. The monthly subscription by user makes netflix a profitable business and the flexibility in subscription users can cancel it anytime. So to engage customers to this platform Netflix must keep their content interesting that can hook users on the platform. That's why the recommendation system which provides valuable suggestions to users is essential.

Netflix's recommendation system gives the idea to them about the popularity of their services provides as it help to increase the sold the subscriptions as more as possible,which offers a varieties of items for selections, this help to get them a user satisfaction,and their loyalty to platform and get them a better understanding of what the user wants.

Then it's easier to get the user to make better decisions from a wide variety of movie products.

With over 139 million paid subscribers(total viewer pool -300 million) across 190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the leading Internet television network and the most-valued largest streaming service in the world. The success behind the amazing story of Netflix is incomplete without the mention of its recommender systems that focus on personalization according to users. According to your preferences,there are several methods to create a list of recommendations.You can use (Collaborative-filtering) and (Content-based Filtering) for recommendation.

➢ 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 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) Netflix has increasingly focused on TV rather than movies in recent years.
4) Clustering similar content by matching text-based features

➢ Dataset Description:

The dataset has 7787 rows and 12 attributes to work with.

Attribute Information:

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 Release Year 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 : Genre

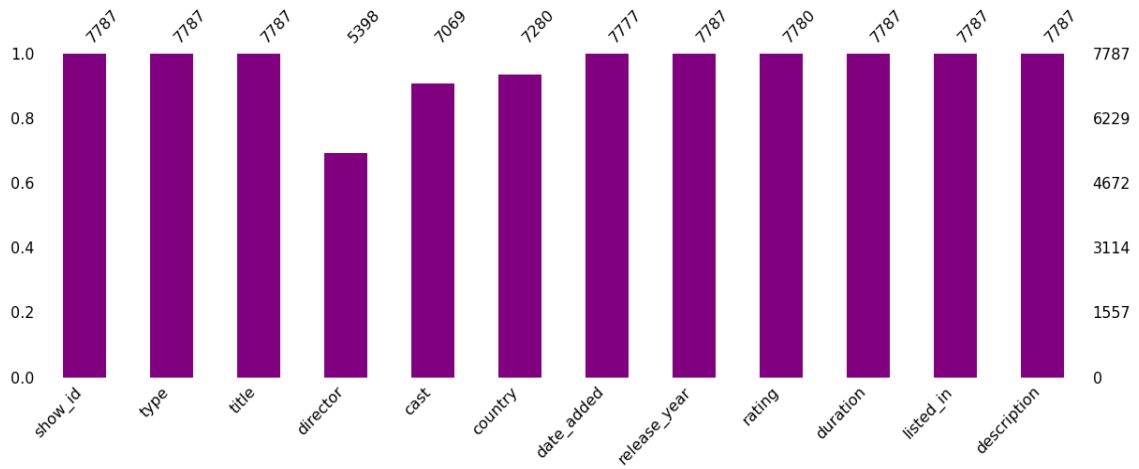
12. description: The Summary description

➢ **Steps Involved:**

Performing EDA (Exploratory data analysis) and data Preprocessing

1) Exploring head and tail of the data to get insights on the given data.

2) Checking if the Null values or missing values are present in the dataset or not.



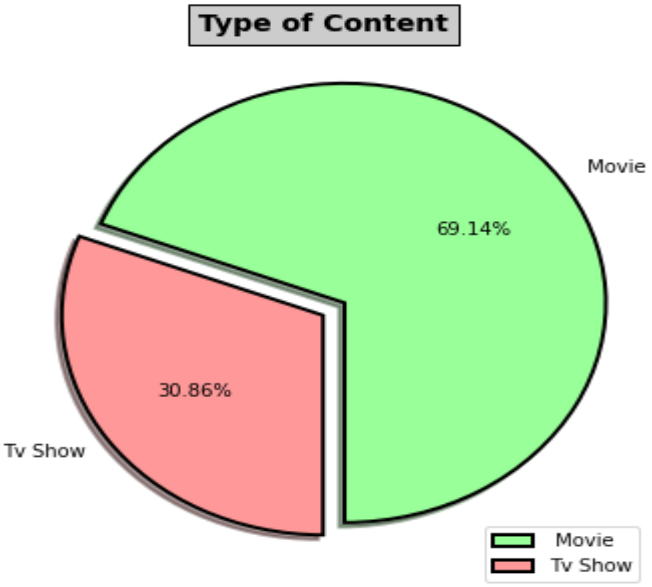
3) Check the duplicate values.

4) Creating data frames which help in drawing insights from the dataset.

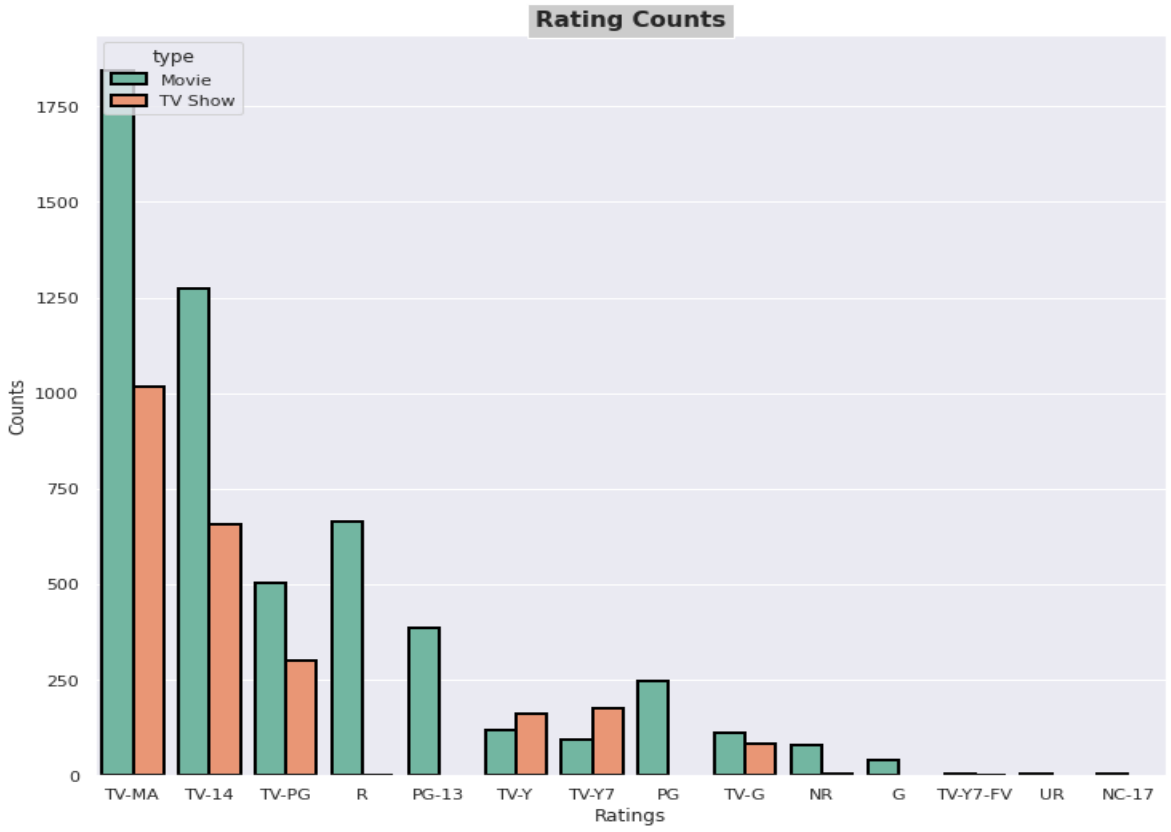
Drawing conclusion from the data

Plotting necessary graphs which provides relevant information on our data like:

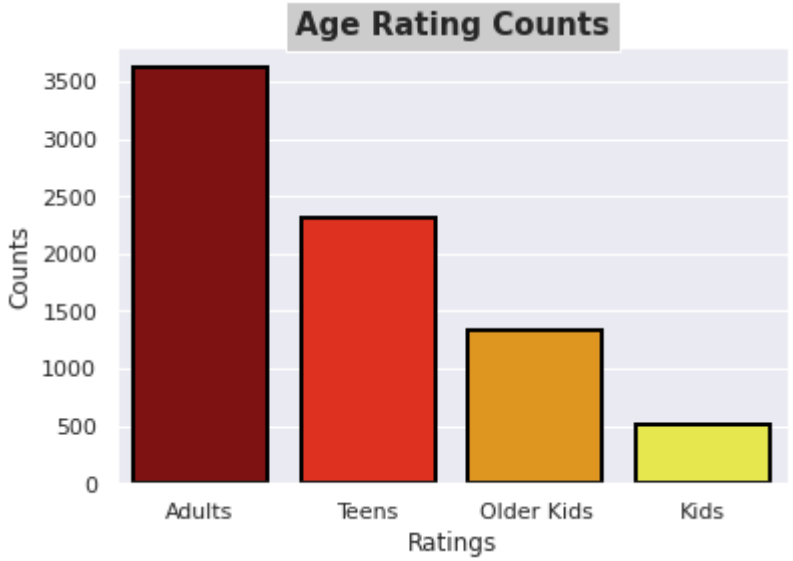
1) Let us plot the type of content available on netflix



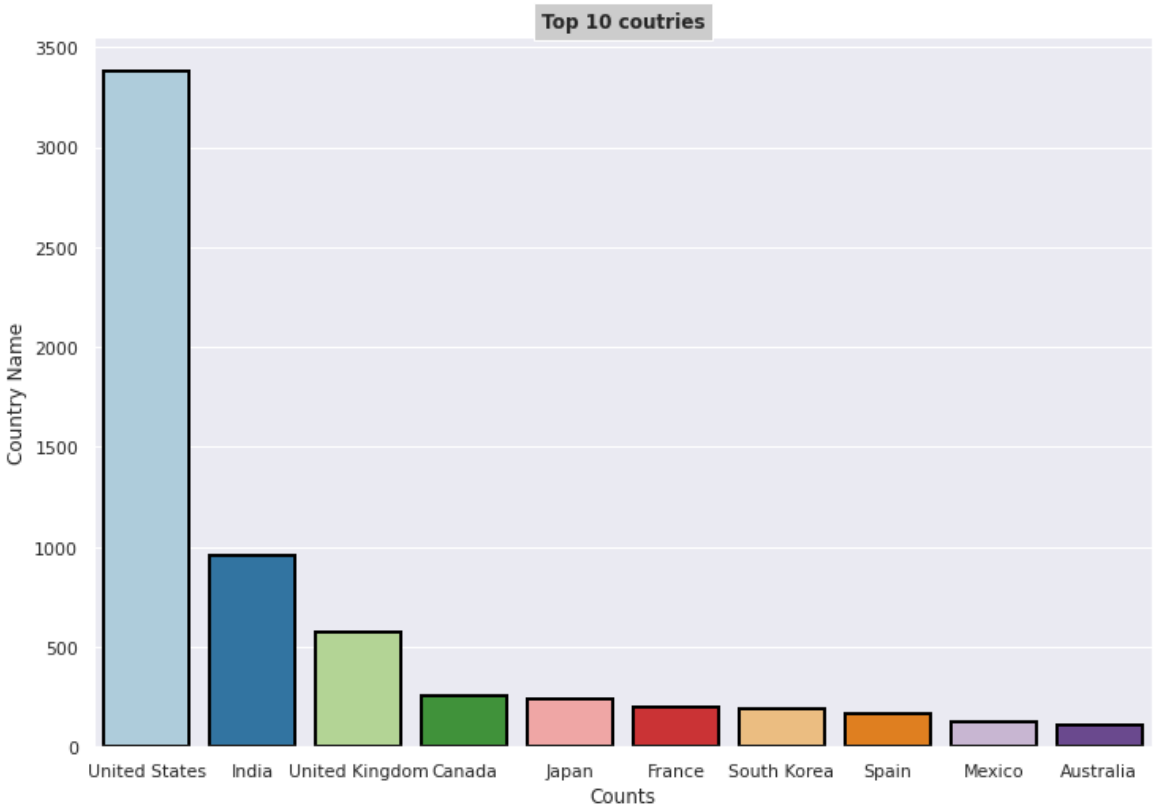
2) EDA on Rating analysis on netflix movies and TV shows



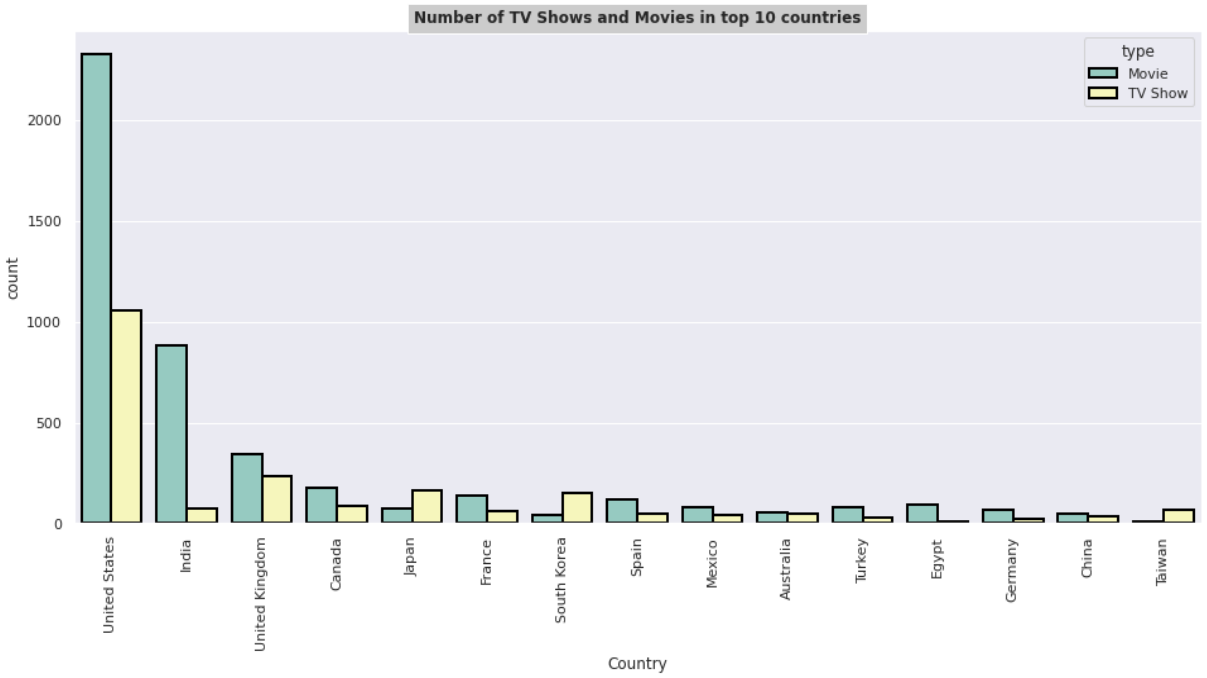
3) EDA on age rating counts



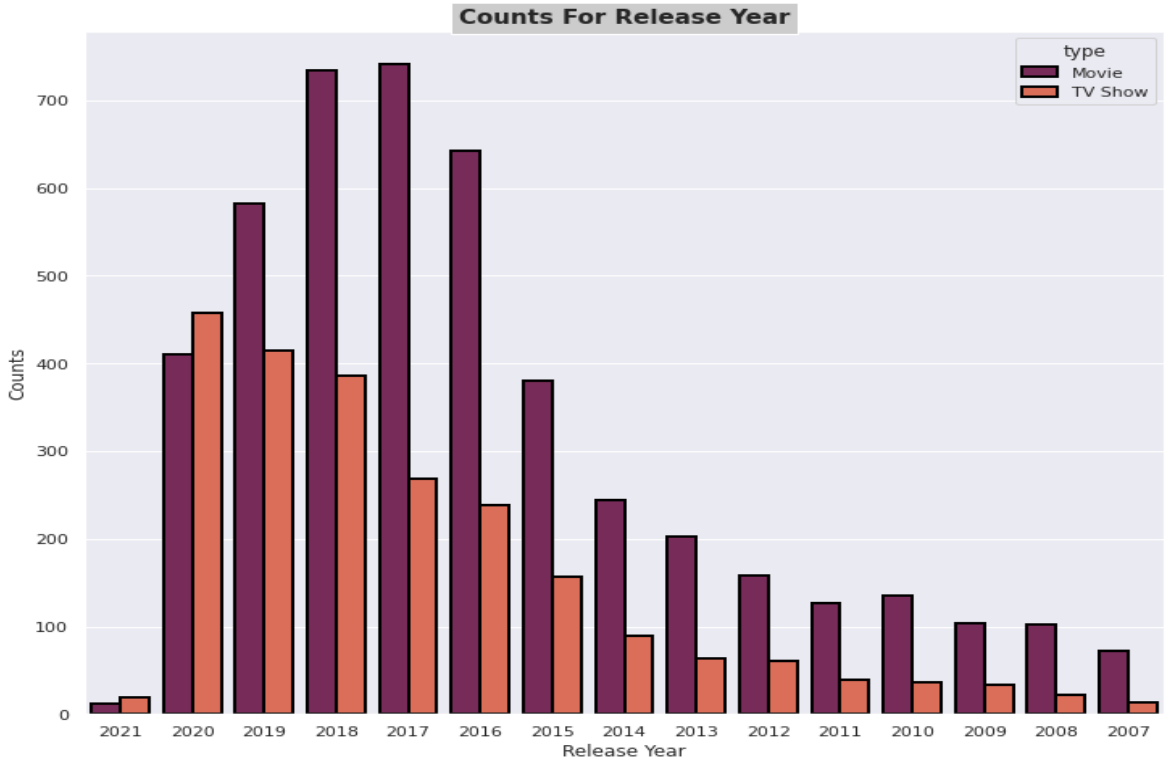
4) EDA on top 10 countries on netflix



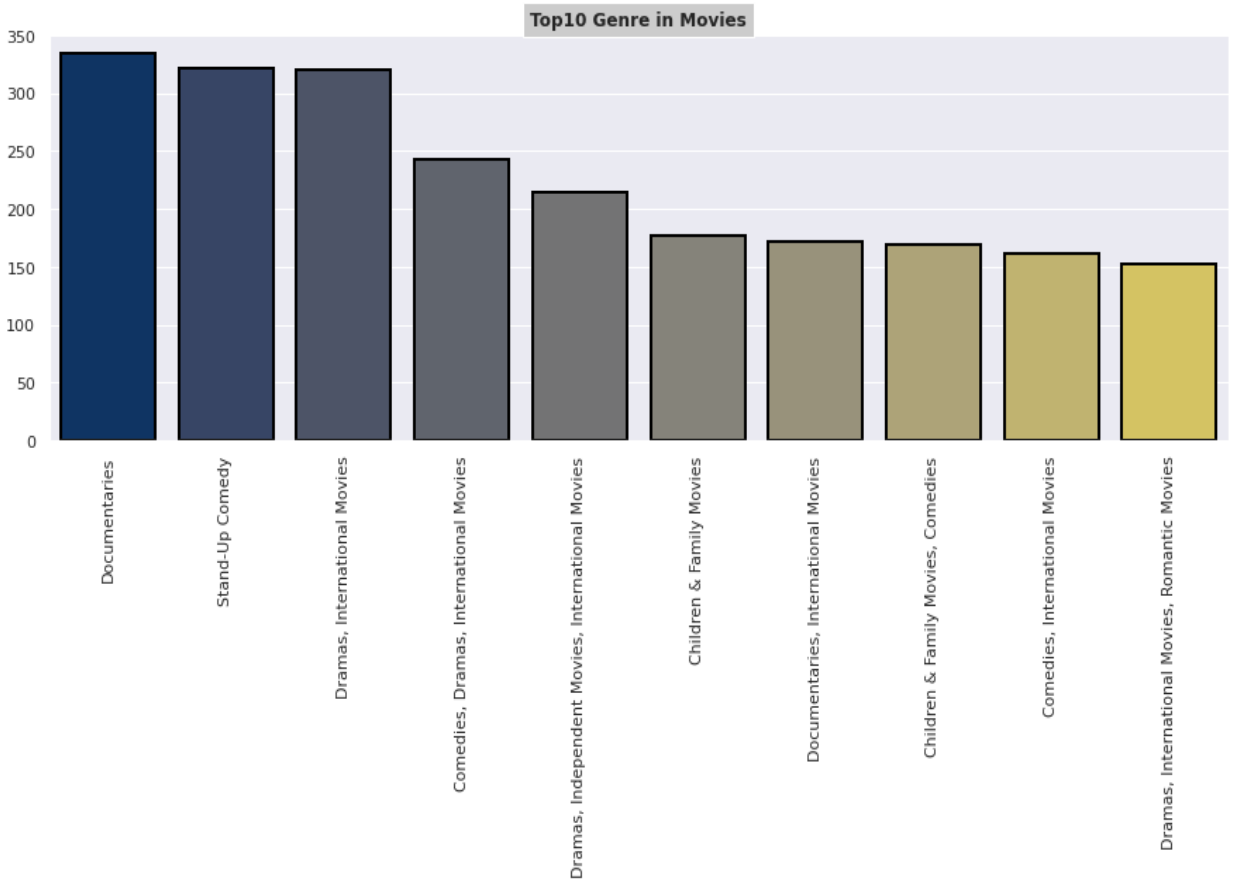
5) EDA on number of TV shows and movie contents in top 10 countries with maximum content



6) Plotting the countplot for release year analysis

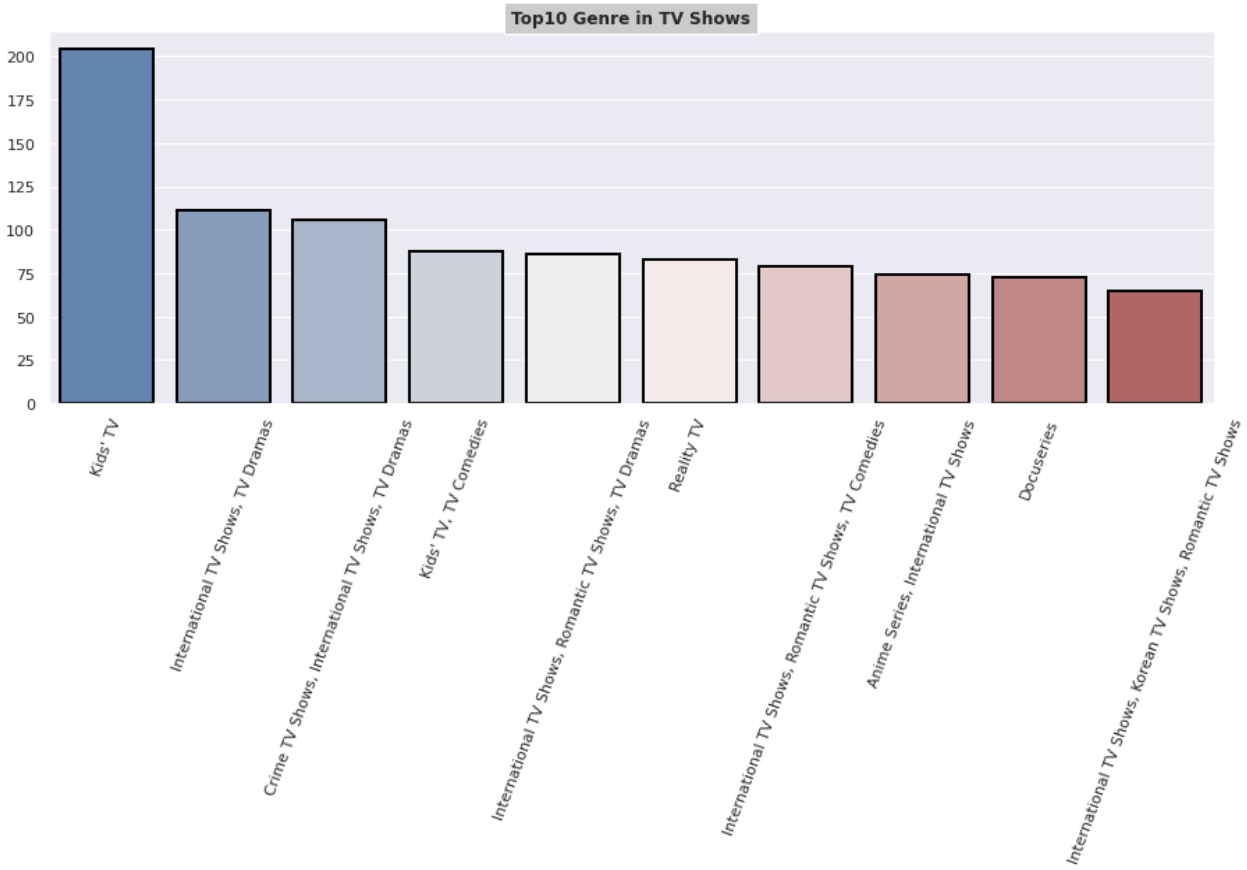


7) Top 10 genres in movies



8) Word Cloud for the movies

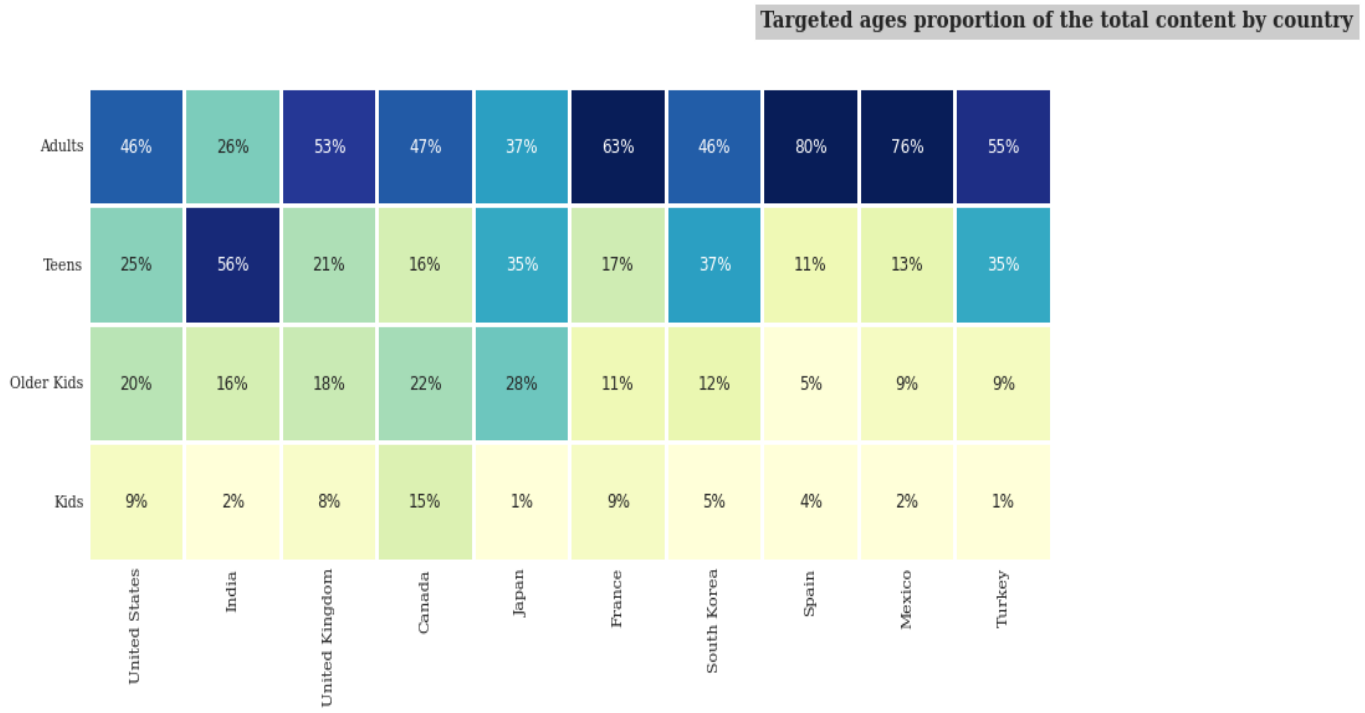
9) Top 10 genres in TV shows



10) Word cloud for Tv shows

11) Netflix Content for different age groups in top 10 countries

Plotting the heatmap:



➢ Approach

As per the problem statement, Understanding what type of content is available in different countries and Is Netflix increasingly focused on TV rather than movies in recent years we have to do clustering on similar content by matching text-based features. For that we used Countvectorizer , TF-IDF vectorizer, and K-means Clustering, Elbow method.

➢ Scaling the data

We have used the Standard Scale method to scale the dataset.

➢ Building a clustering model

Clustering models allow you to categorise 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 labelling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for. This is what distinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict. These models are often referred to as **unsupervised learning** models, since there is no external standard by which to judge the model's classification performance.

➢ Metrics used

Silhouette coefficient or silhouette score

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

➢ Model Implementation

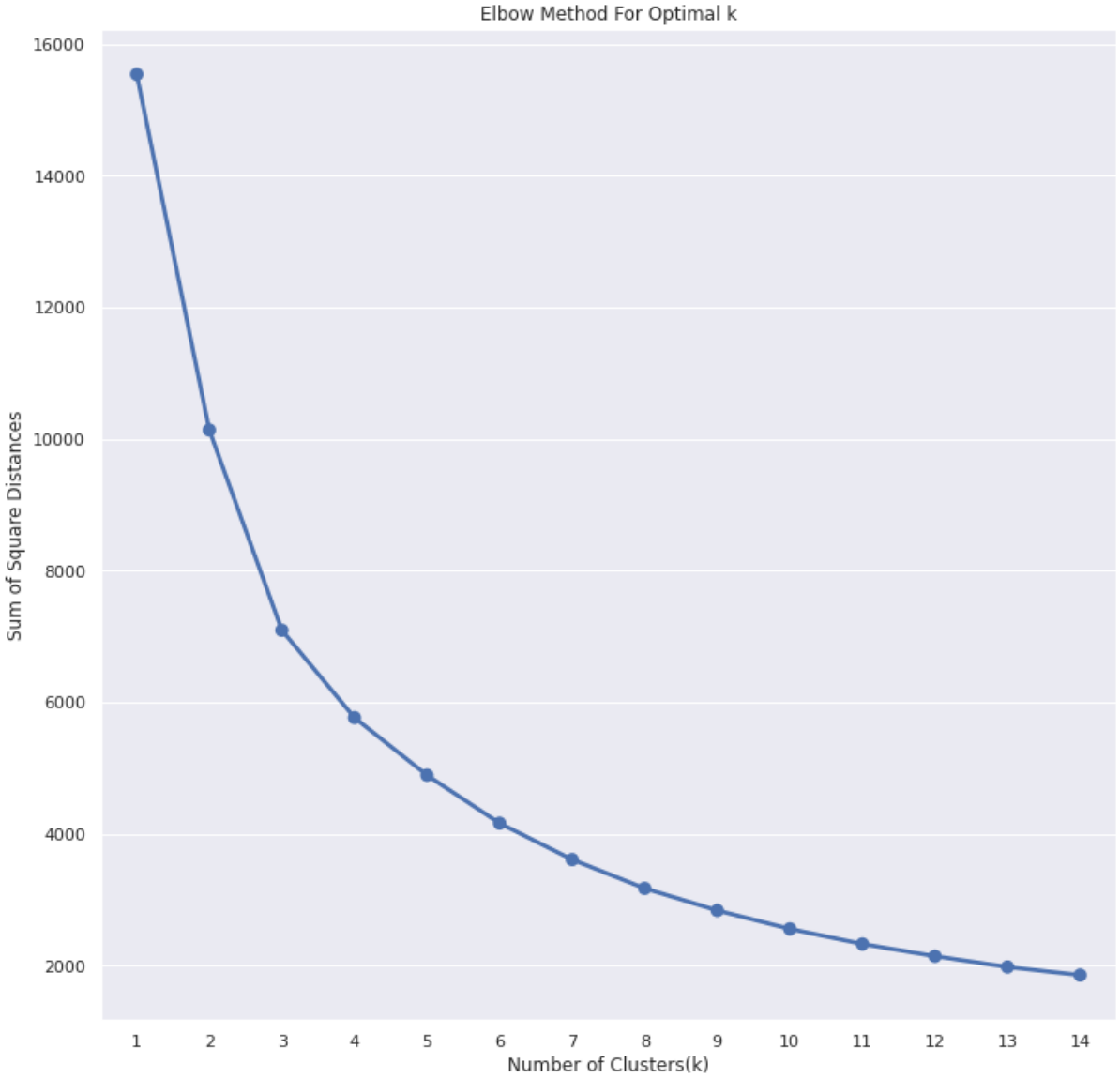
K-means Clustering

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster.

We created the sample data using build blobs and used range_n_clusters to specify the number of clusters we wanted to utilise in k means.

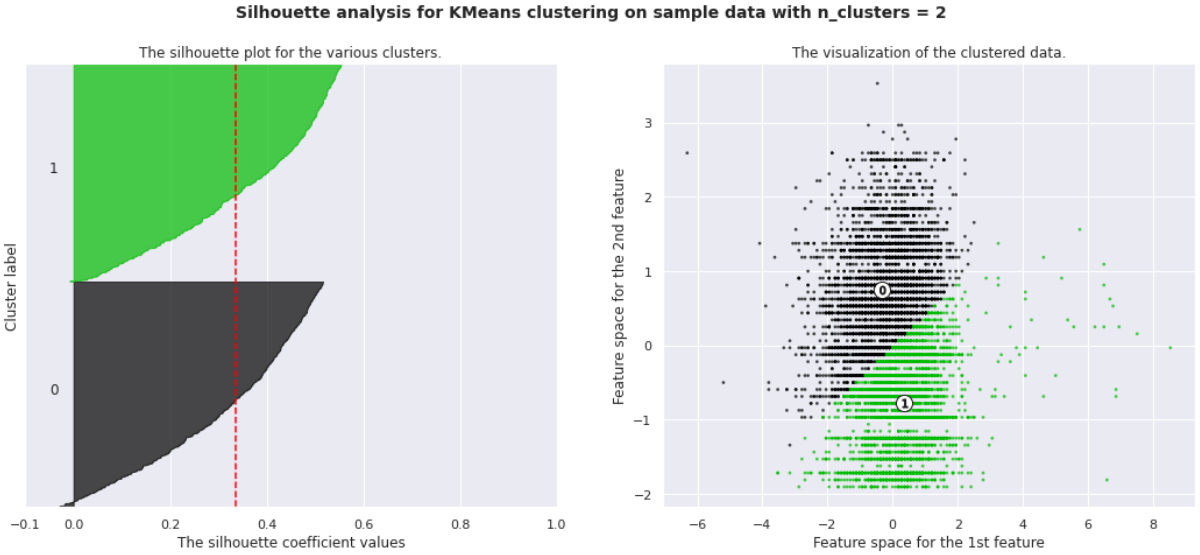
Elbow Method

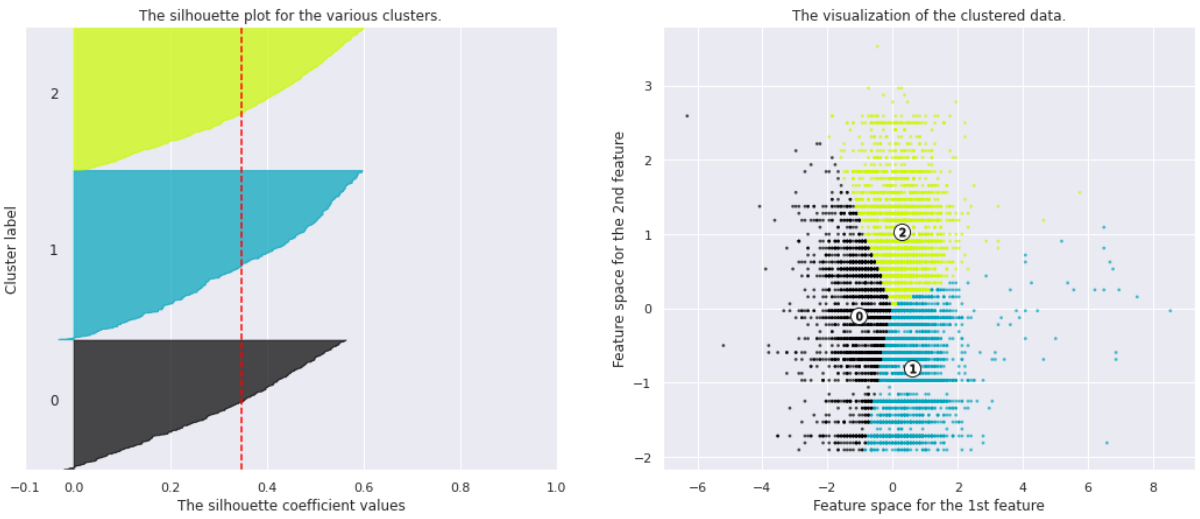
The Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.

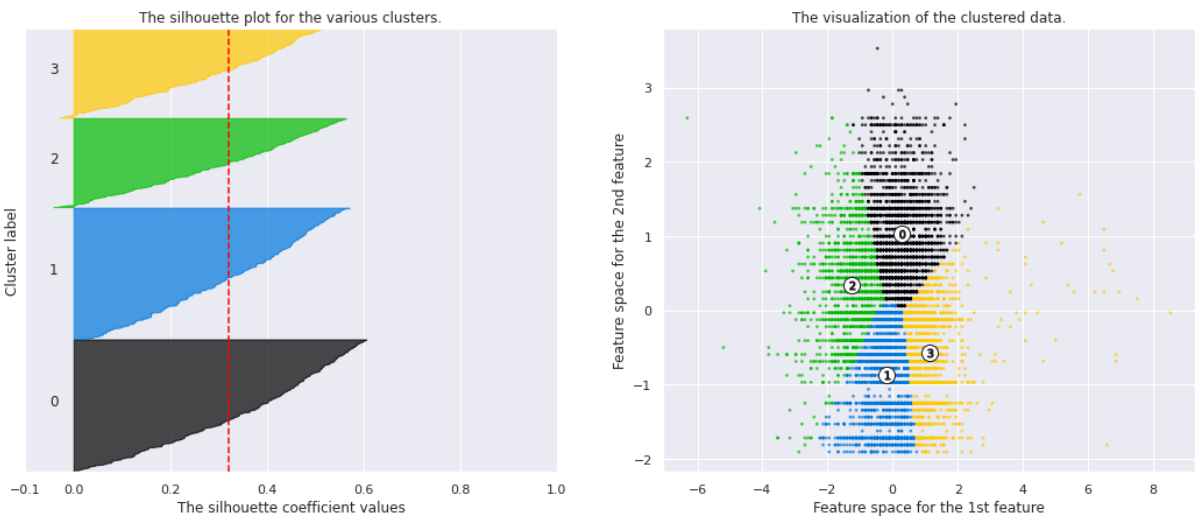


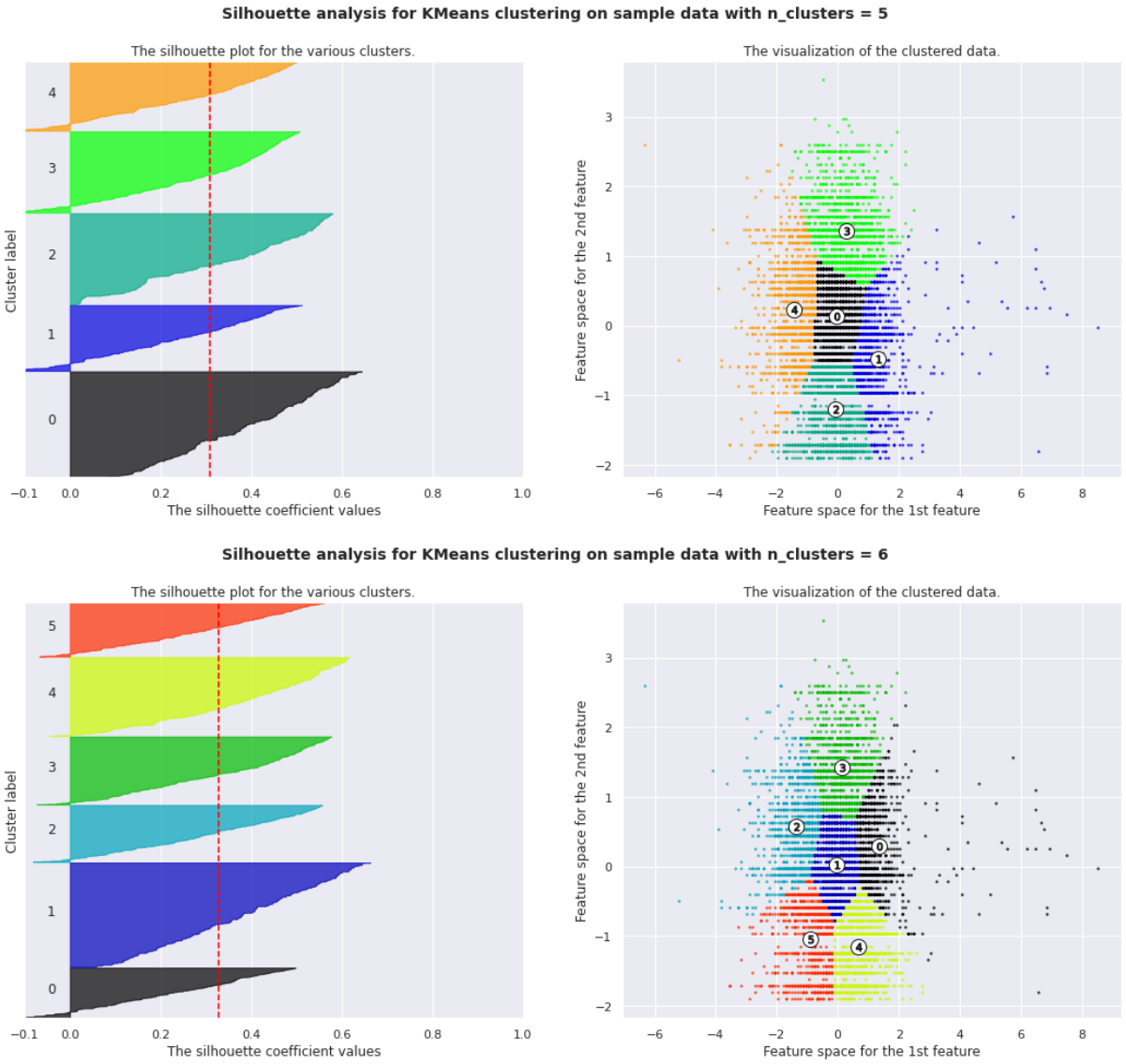
With an increase in the number of clusters (k), the average SSE decreases.To select the best value of k we use Silhouette score as below-

➢ Silhouette score and visualisation-


Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature


Silhouette analysis for KMeans clustering on sample data with n_clusters = 3
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature




```
For n_clusters = 2 The average silhouette_score is :
0.3367875569876181
For n_clusters = 3 The average silhouette_score is :
0.3481431878723329
For n_clusters = 4 The average silhouette_score is :
0.3207442149237176
For n_clusters = 5 The average silhouette_score is :
0.3079420368105537
For n_clusters = 6 The average silhouette_score is :
0.32881670294216747
```

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. If most objects have a high

value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

➢ Challenges Faced

The following are the challenges faced in the data analysis:

1) Pre-processing the data was one of the challenges we faced which includes handling missing values and filling the missing values
2) Feature engineering
3) Removing Punctuation and removing Stopwords
4) Model Implementation

❖ Conclusion

- **We've done null value treatment, feature engineering, and EDA since loading the dataset, and then we've completed some tasks that were assigned to us.**
- **We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies)**
- **Netflix has increasingly focused on movies than TV shows. It has been producing more movies than tv shows since 2014.**
- **Netflix is most popular in the United States. India lies at 2 positions in the popularity list.**
- **In most of the countries the content available on netflix is mostly of movie type except in South Korea and Japan.**
- **Clustering was done using 'length' and 'length_listed" columns.**
- **Using the elbow method and silhouette score the best number of clusters turned out to be 3 with silhouette score of 0.34 which is great indicating our clusters are homogeneous but heterogeneous to one another.**