CLUSTERING ANALYSIS ON NETFLIX MOVIES AND TV SHOWS

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

❖ INTRODUCTION

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. 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.

Netflix's recommendation system gives the idea to them about the popularity of their services provides as it helps to increase the sold the subscriptions as more as possible, which offers a variety 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.

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*. Netflix ended 2021 with 221.8 million global paid subscribers. Netflix's subscriber growth at this point is mostly from outside of the U.S. and Canada.

❖ 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

❖ DATA SUMMARY

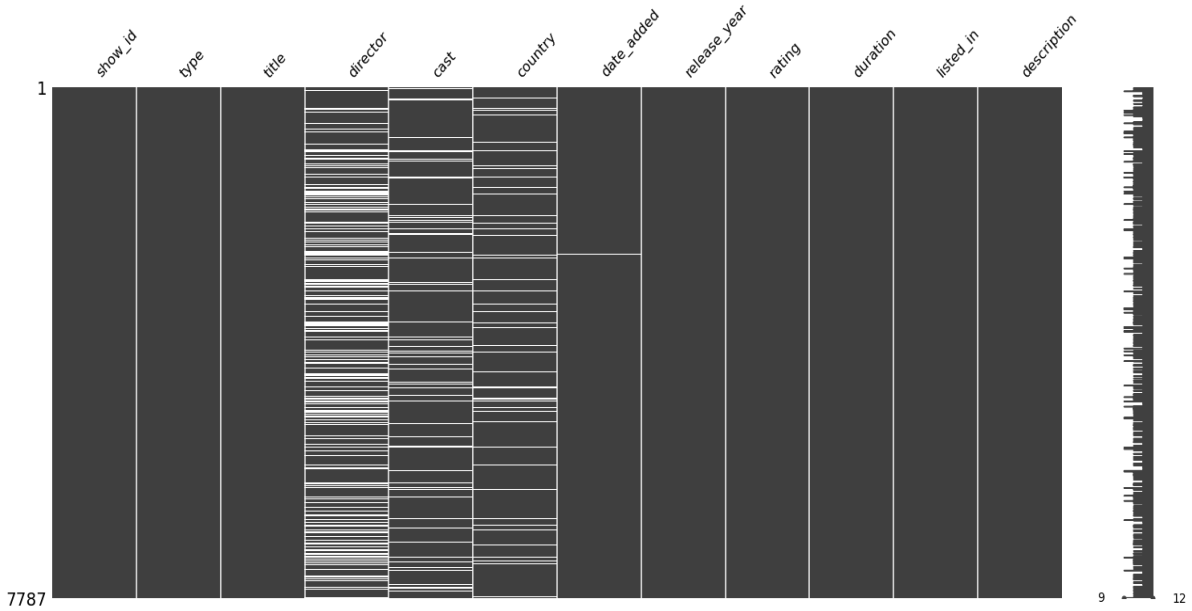
The dataset has 7787 rows and 12 attributes to work with.

- **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

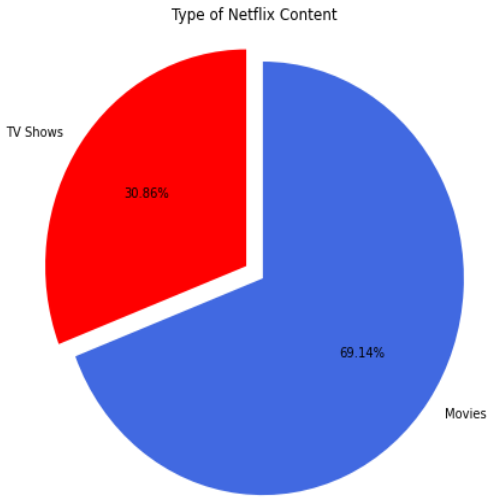
- **NULL AND DUPLICATE VALUES**
  The given dataset has no duplicate values but some of the attributes contain null values in it.



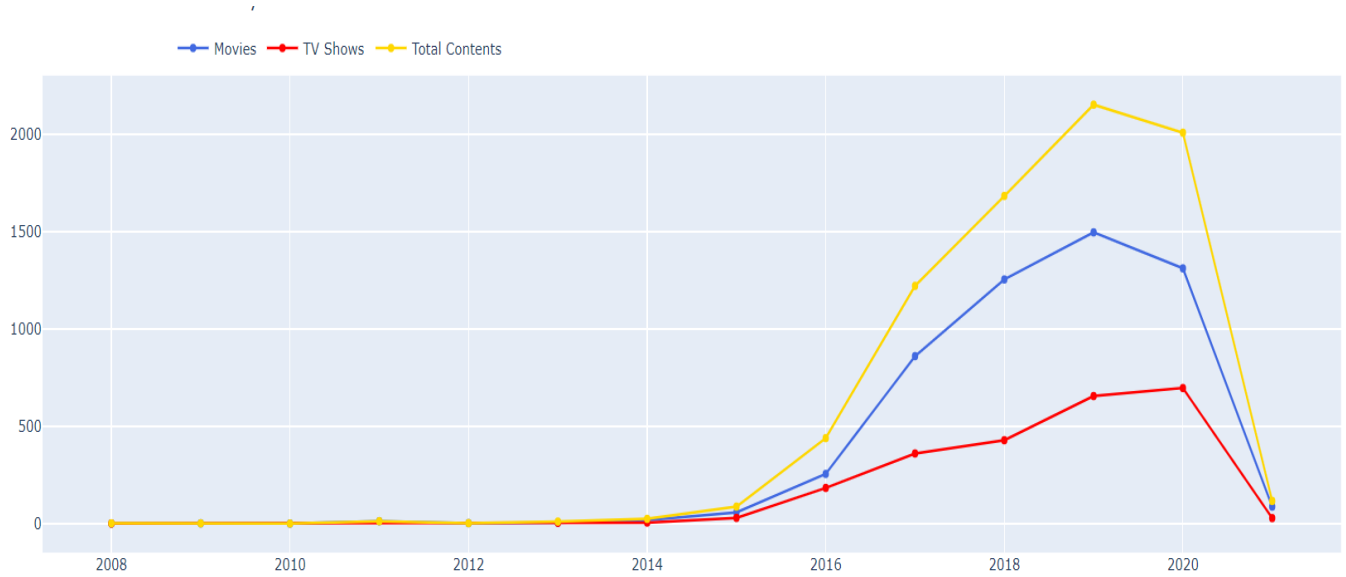
- **EXPLORATORY DATA ANALYSIS(EDA):**

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

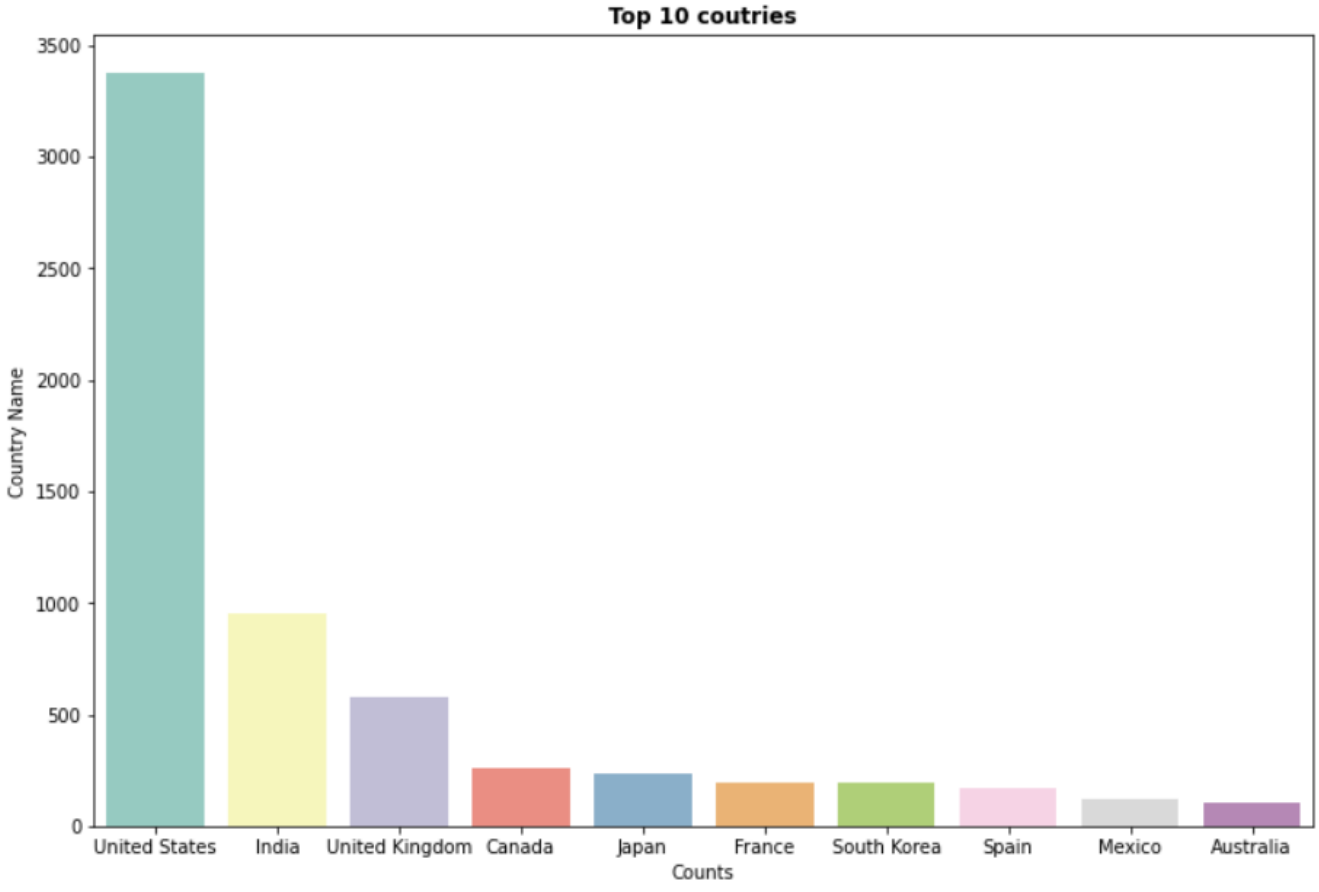
1. Type of content available on Netflix



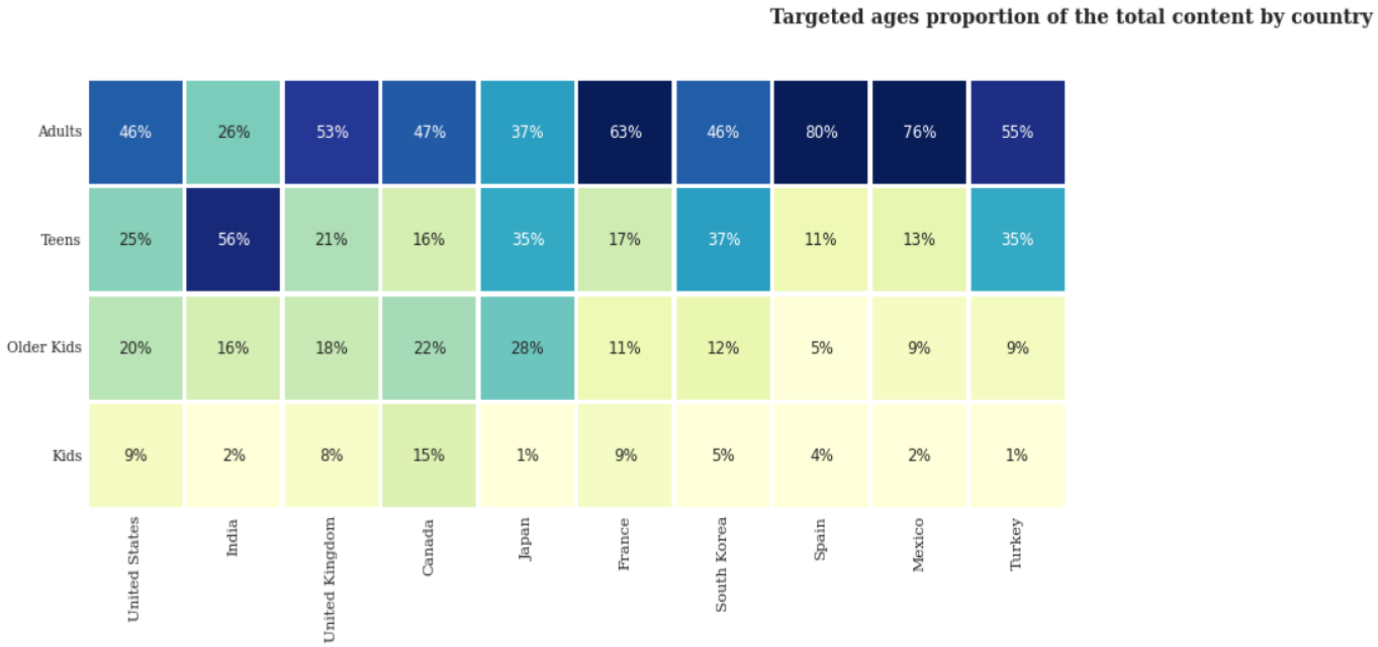
2. content added over the years



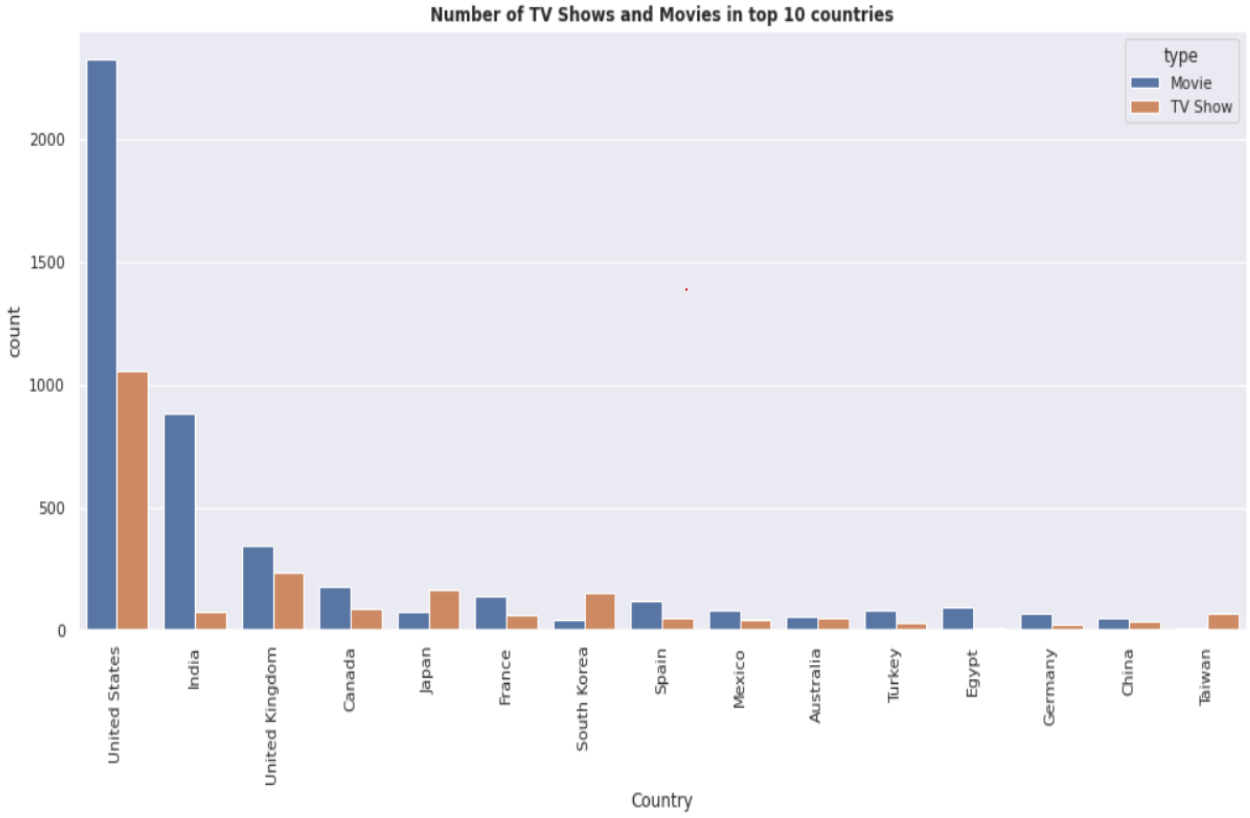
3. Countries with most Netflix popularity



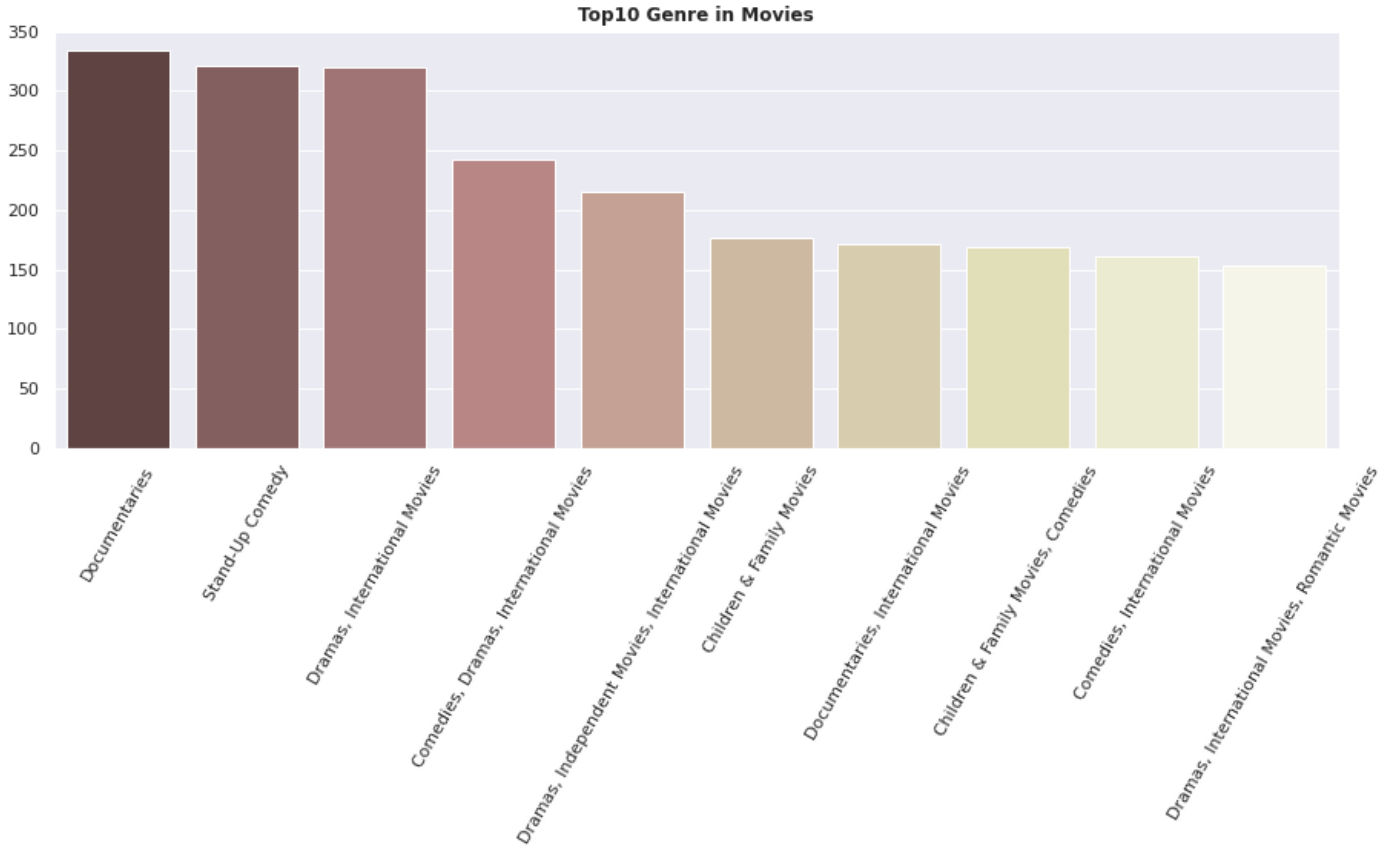
4. Targeted ages



5. Content in different countires

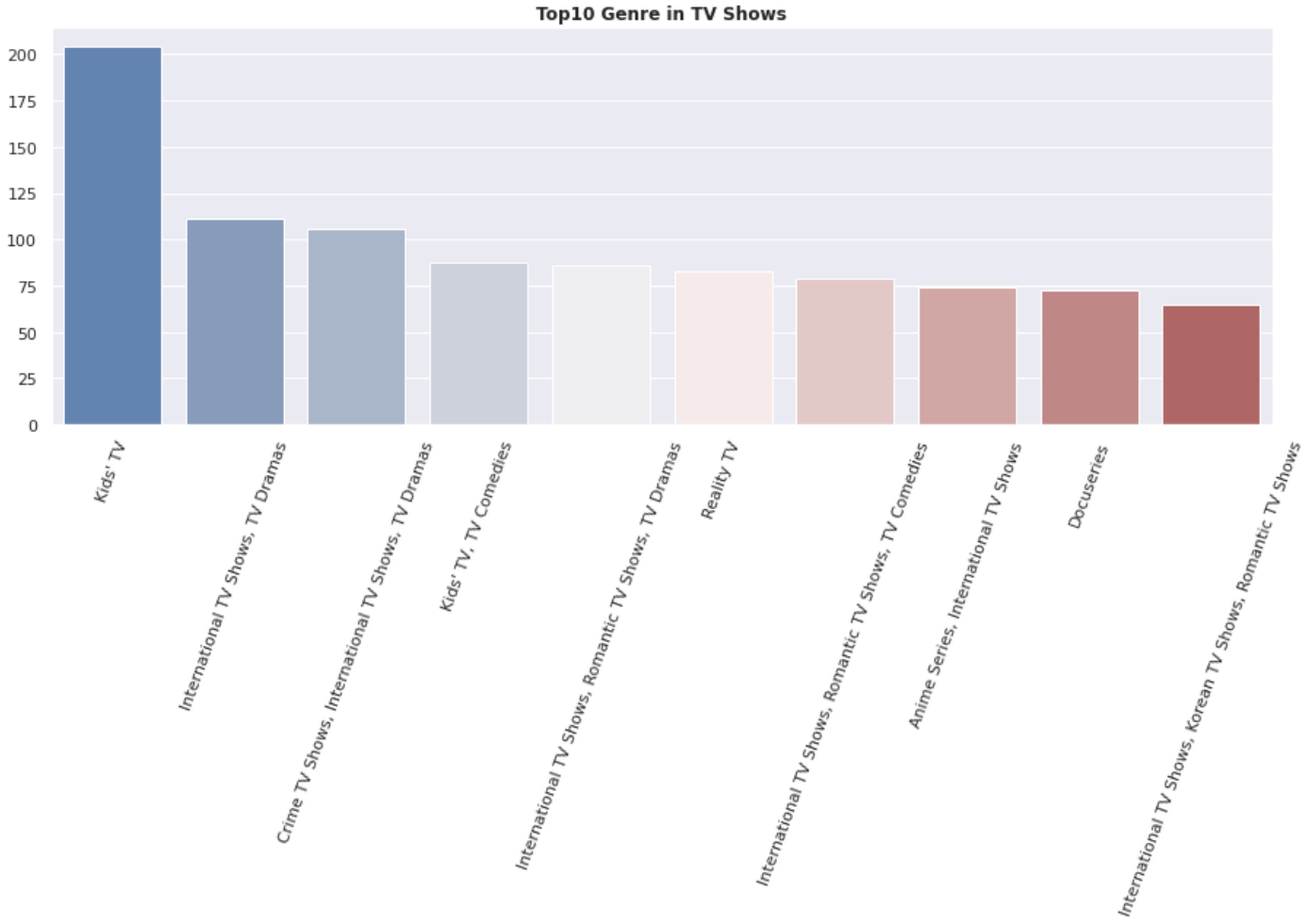


6. Top 10 genres in movies



Word Cloud for the movies

7. Top 10 genres in TV shows



Word cloud for Tv shows

❖ 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 Count vectorizer, 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.

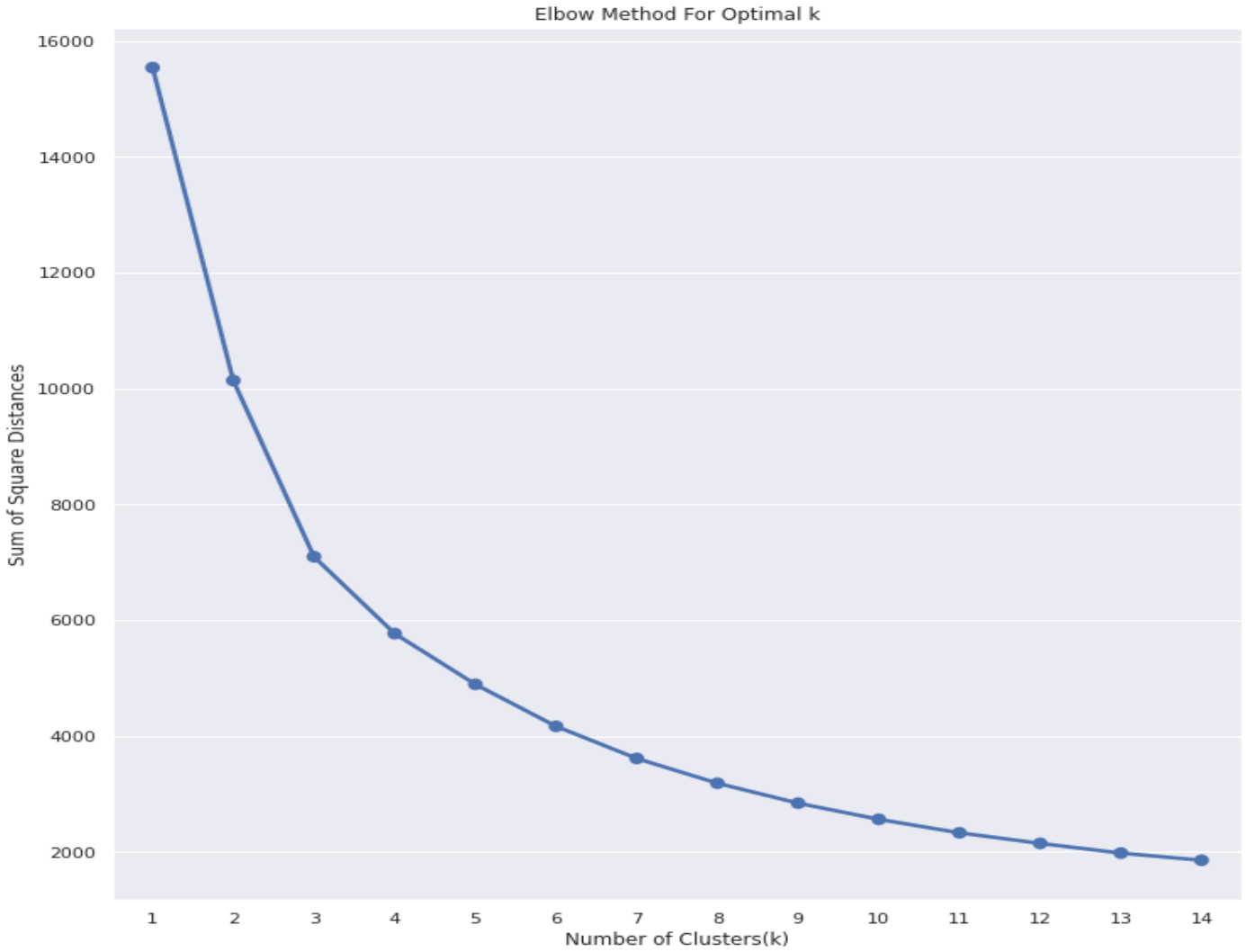
❖ MODEL IMPLEMENTATION

- **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.

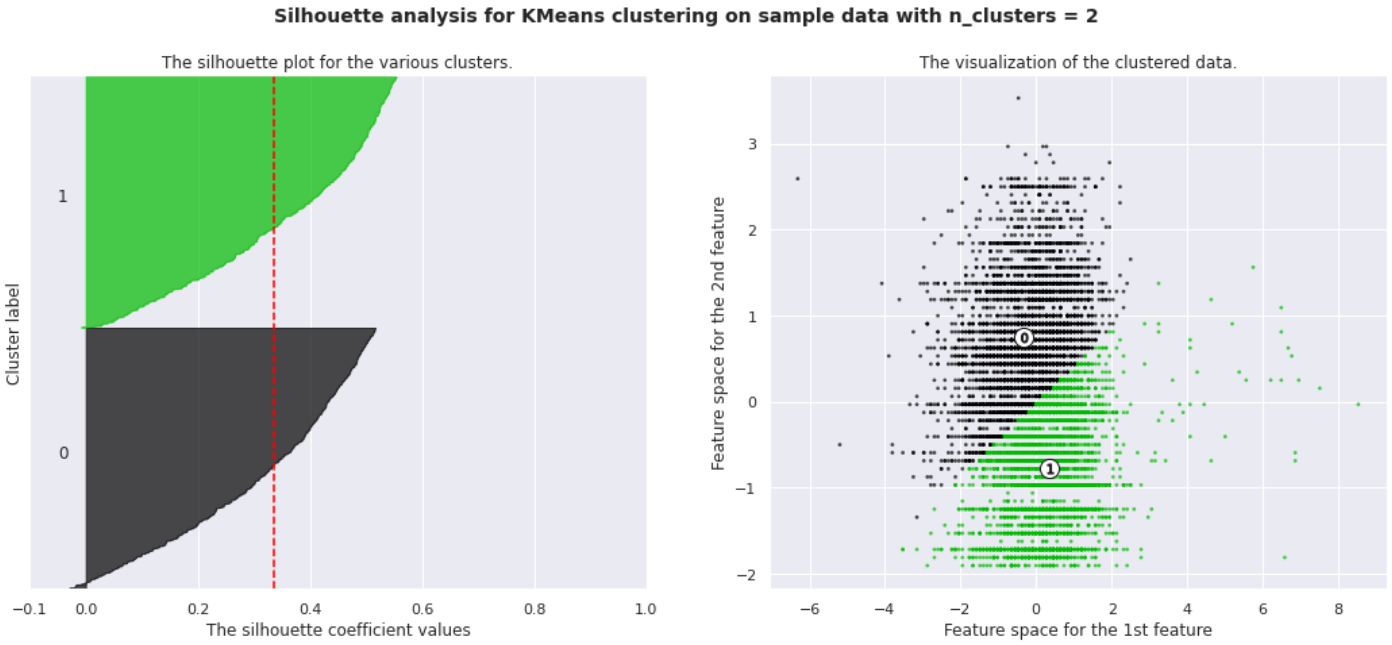
- **Elbow Method**

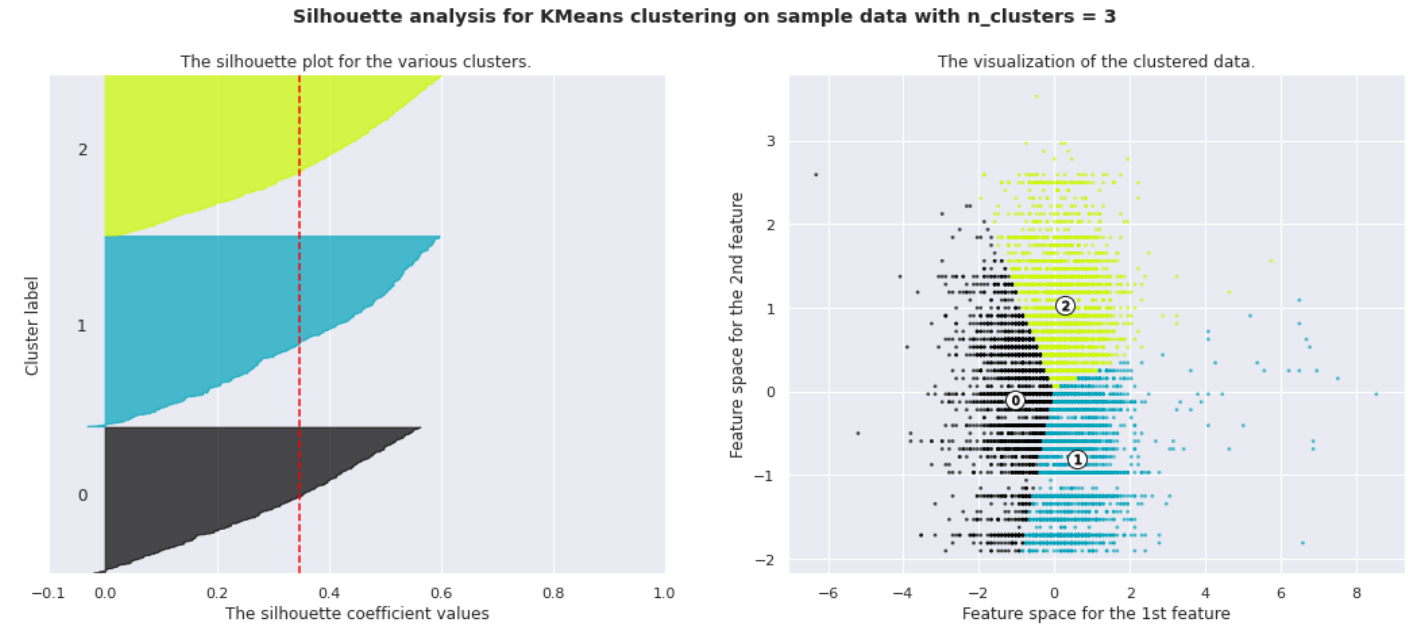
The Elbow Method is an empirical method to find the optimal number of clusters for a dataset. In this method, we pick a range of candidate values of k, then apply K-Means clustering using each of the values of k. Find the average distance of each point in a cluster to its centroid, and represent it in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



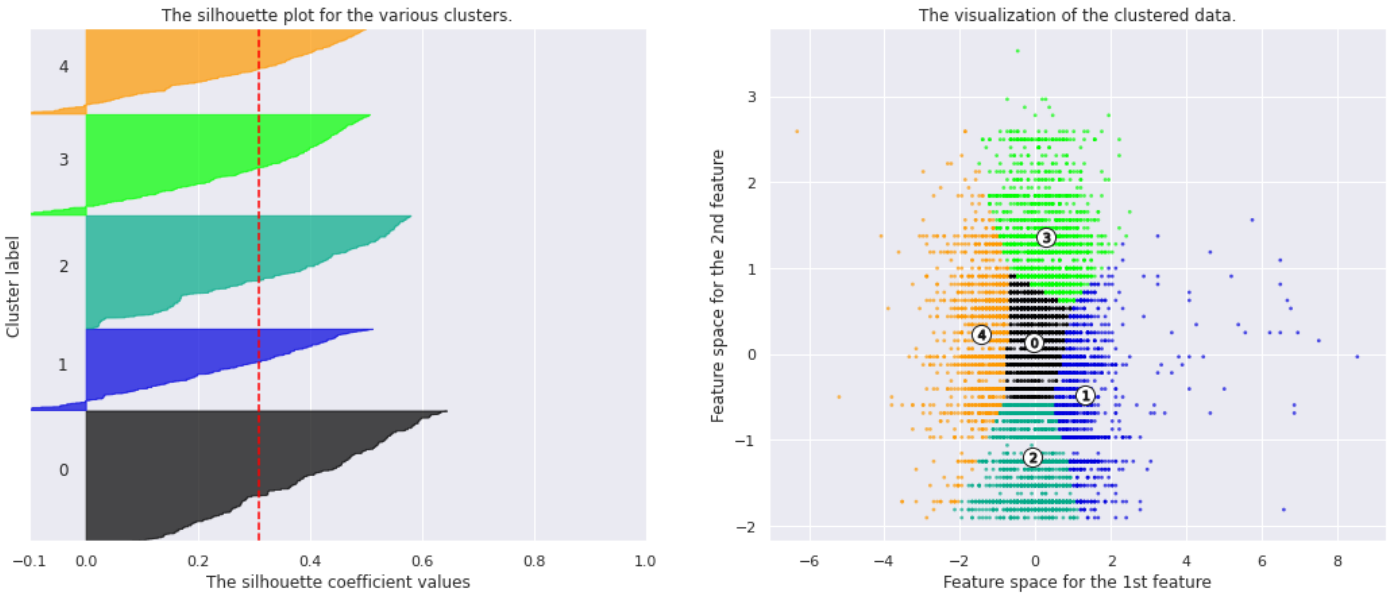
With an increase in the number of clusters (k), the average SSE decreases.To select the best value of k we use Silhouette score as below-

- **Silhouette score and visualisation-**

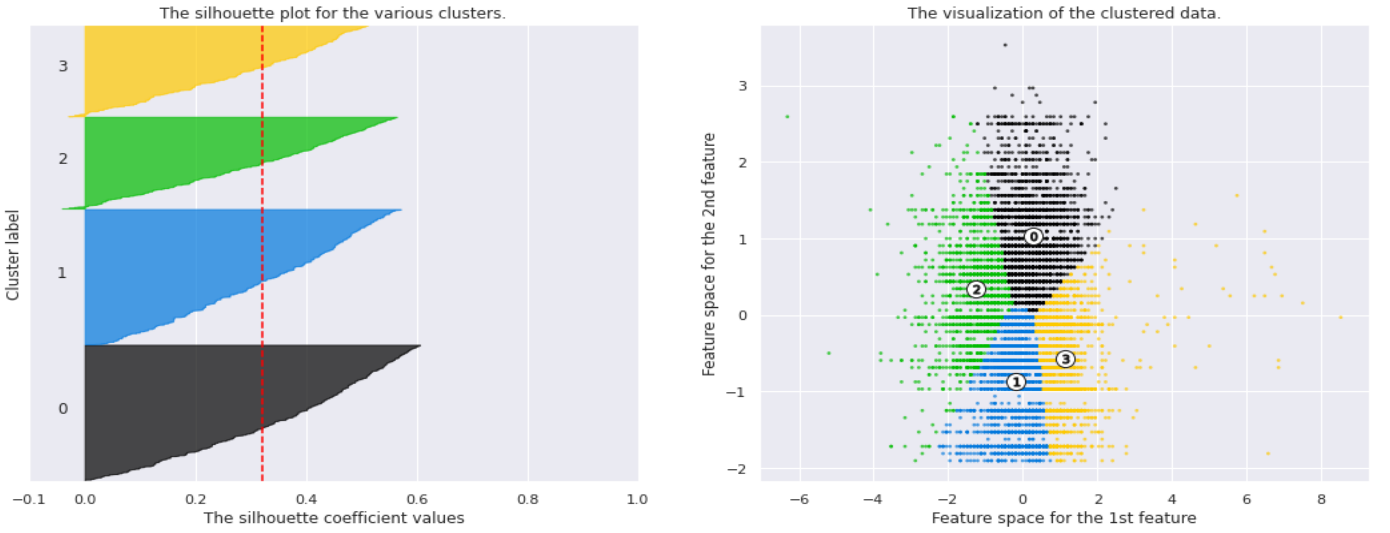




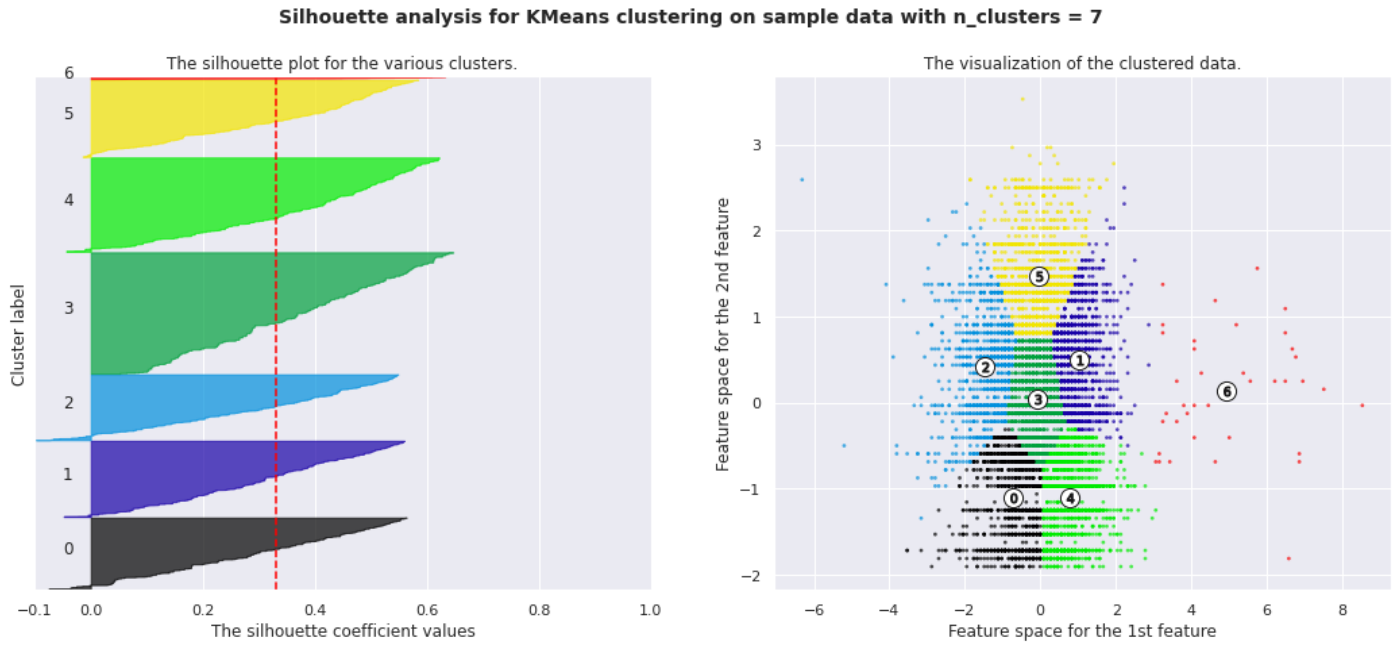
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



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



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




```
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
For n_clusters = 7 The average silhouette_score is : 0.3303332658812144
For n_clusters = 8 The average silhouette_score is : 0.32086474097662543
For n_clusters = 9 The average silhouette_score is : 0.3258331644784623
For n_clusters = 10 The average silhouette_score is : 0.32183497342458306
```

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.

Libraries used:

We used a number of libraries for data analysis purpose as well as for data pre-processing and model building and handling textual data

```
[ ]  # importing standard liabraries
     import numpy as np
     import seaborn as sns
     import pandas as pd
     import matplotlib.pyplot as plt
     import plotly.graph_objects as go
```

```
from sklearn import preprocessing
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split, KFold
from nltk.corpus import stopwords
from nltk.stem.snowball import SnowballStemmer

import nltk
```

```
from scipy.cluster.hierarchy import linkage, dendrogram
from sklearn.metrics import silhouette_score
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples, silhouette_score
from sklearn.cluster import KMeans
```

❖ 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. Handling textual data as there were many steps involved such as removing punctuations and stop word, calculating tf-idf.
4. Finding proper number of clusters

❖ CONCLUSION

- Netflix has a greater number of movies than TV-shows.
- Almost 70% of the content is movies while rest 30% is tv-shows.
- 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 and United Kingdom lies at 2 and 3 positions respectively 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' and 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.