AI

CAPSTONE PROJECT

Clustering Model

TITLE: Netlfix Tv shows and Movies Clustering

Prepared By → PRAVEEN.S

(COHORT EVEREST)

NETFLIX

Let’s go through the Defaulters:

1. Defining the Problem Statement
2. EDA and feature selection
3. Data Visualization
4. Detection of Outliers
5. Preparing dataset for modeling
6. Applying the Cluster models
7. Validation of Models

Problem Statement:

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.
This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable 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.

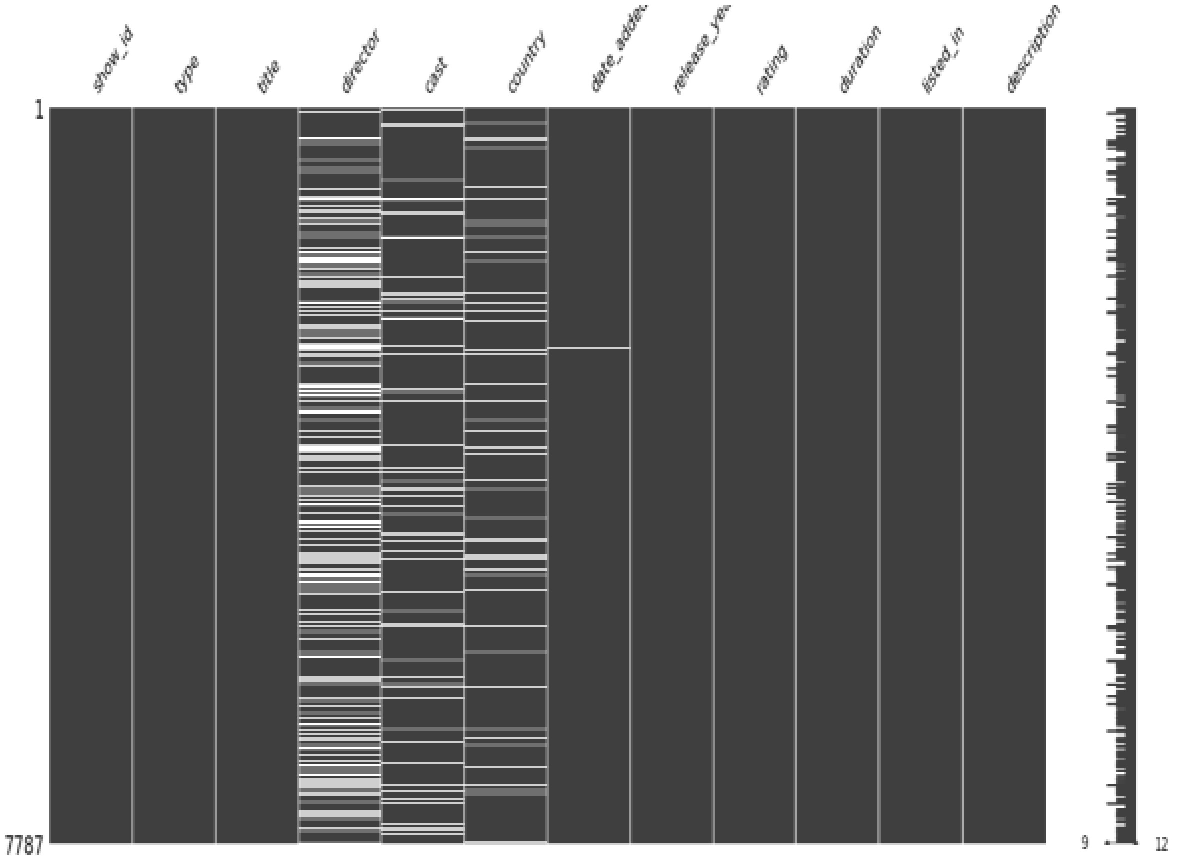
Dataset Understanding:

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 Movie
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 Releaseyear of the movie / show
rating : TV Rating of the movie / show
duration : Total Duration - in minutes or number of seasons
listed_in : Genere
description: The Summary description

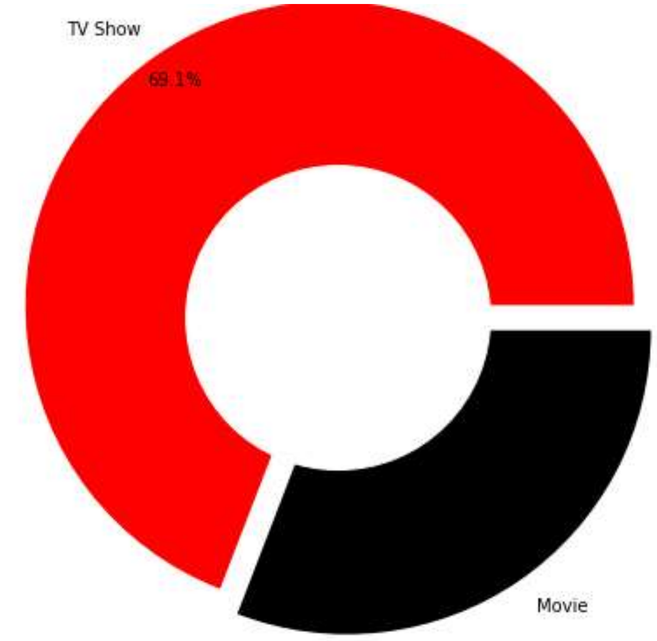
Missing Values Treatment:

- The Percentage of missing values to be around 46%, so it is not practical to drop the missing values.
- There are totally 3631 missing values found in the dataset which were distributed among the director,cast and country features.
- We have replaced the missing values with "No <<Feature>>"

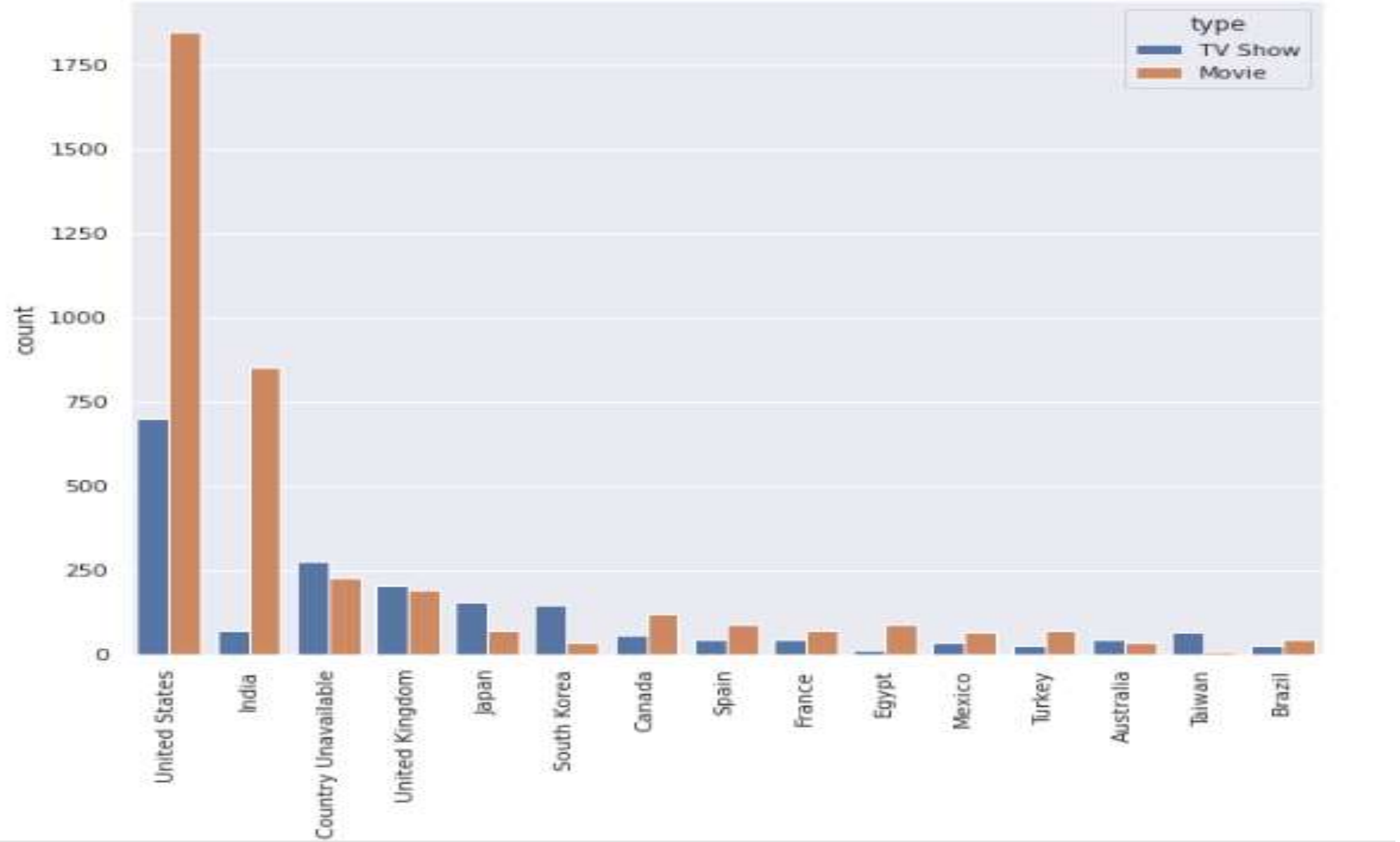
  for example: the missing value in the director column is replaced with "No Director"



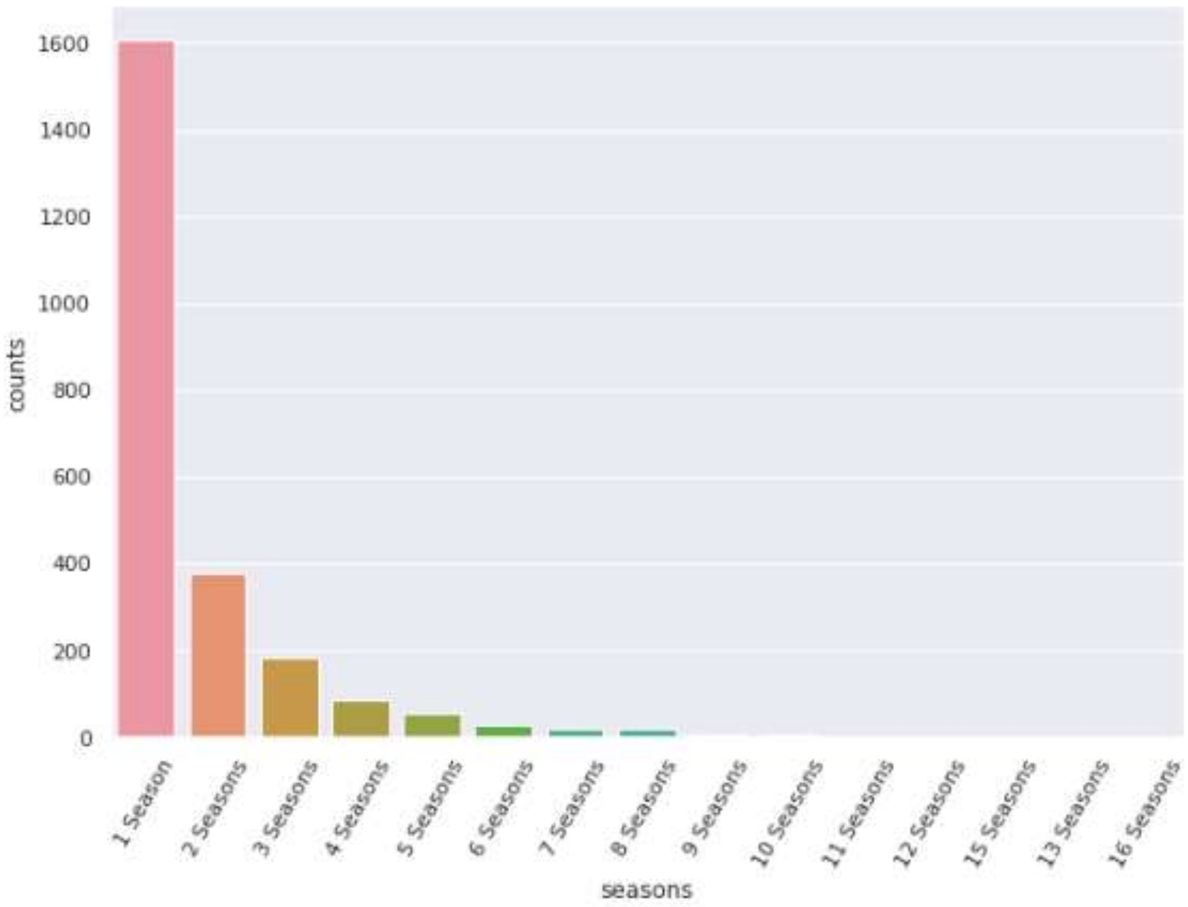
Distribution of TV shows and Movies:

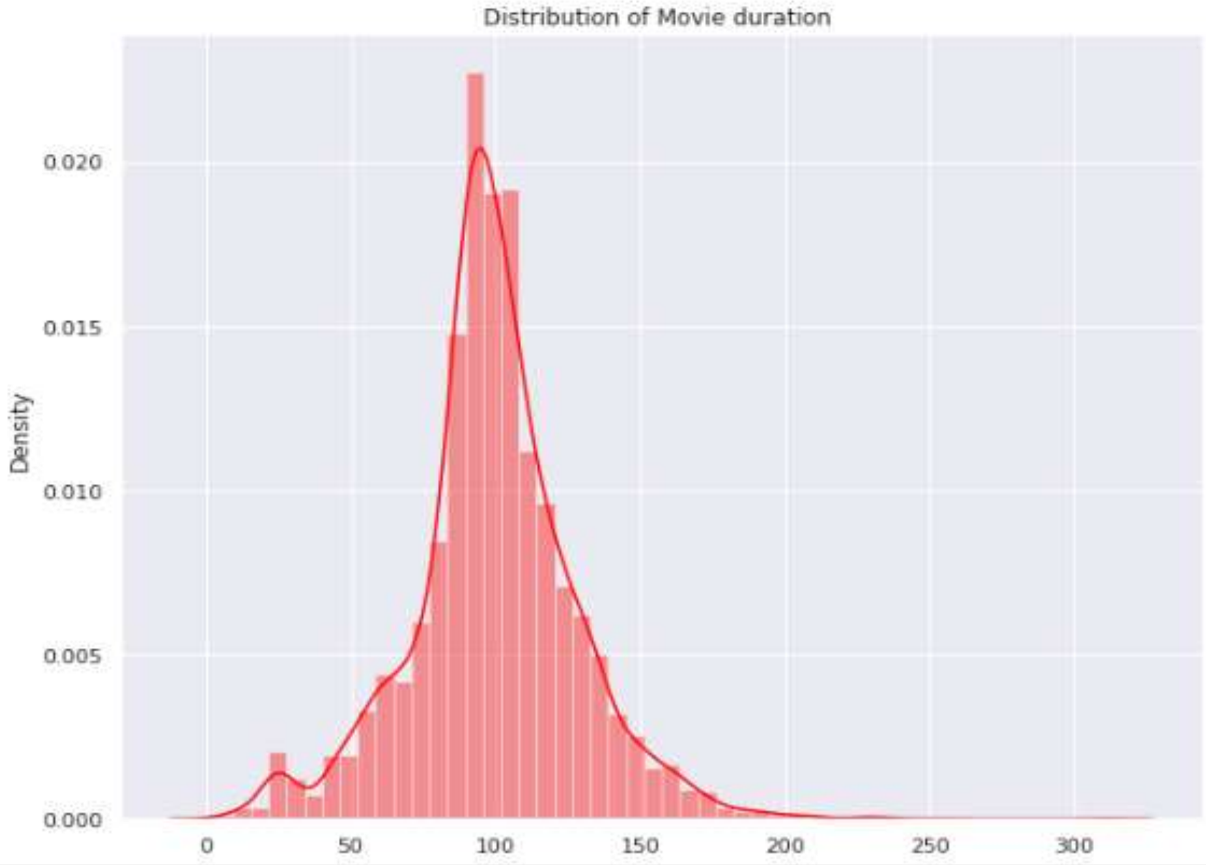


```
Movie       5372
TV Show     2398
```

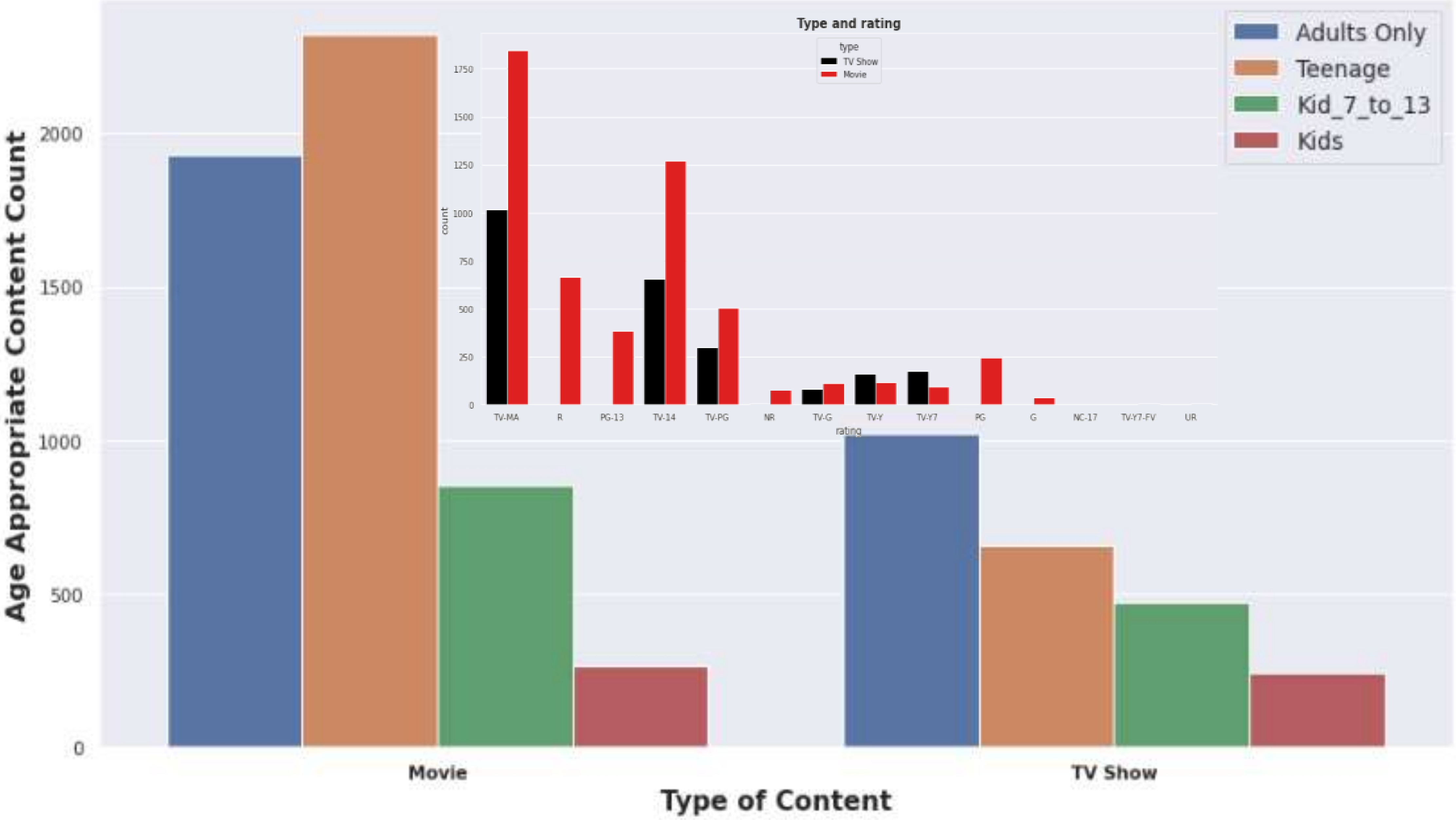


Distribution of TV shows and Movies Duration:

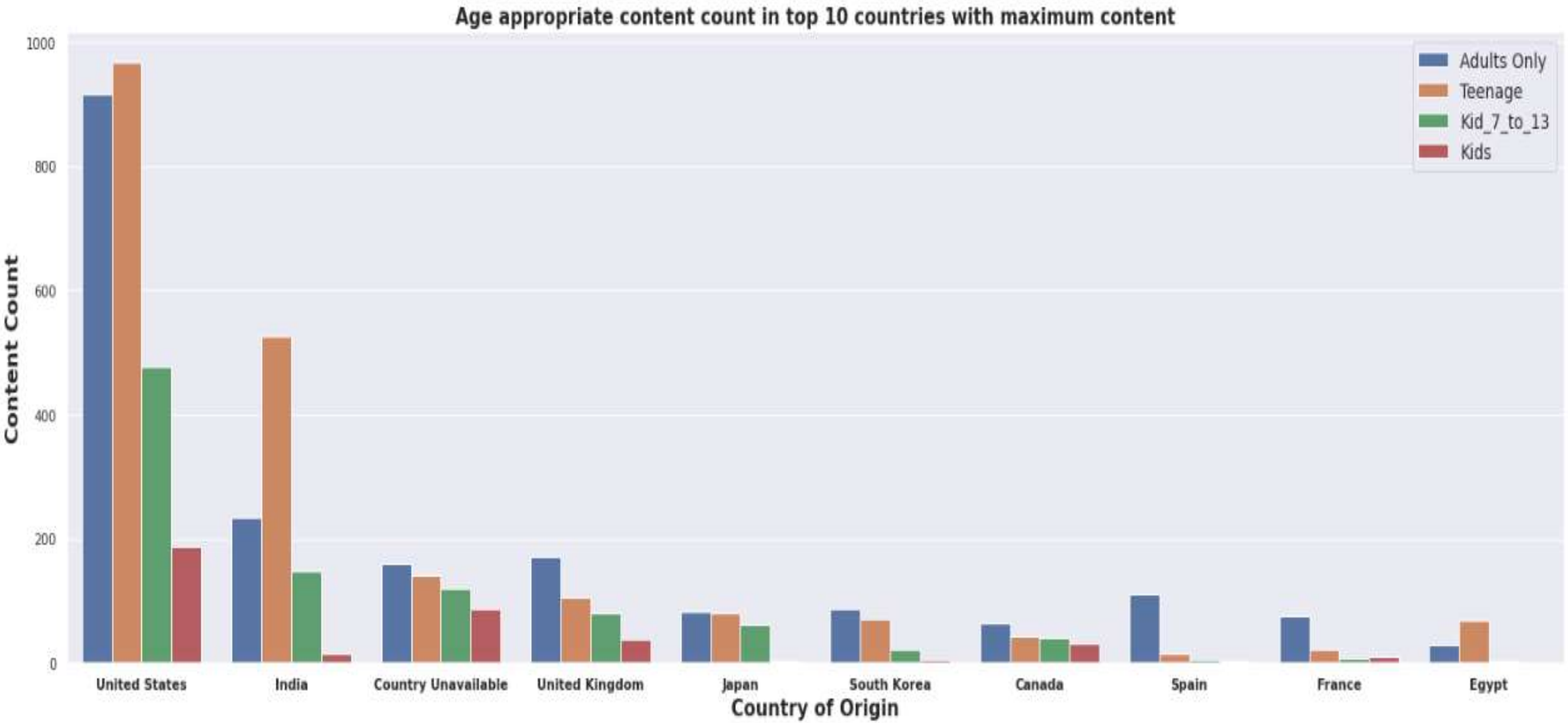




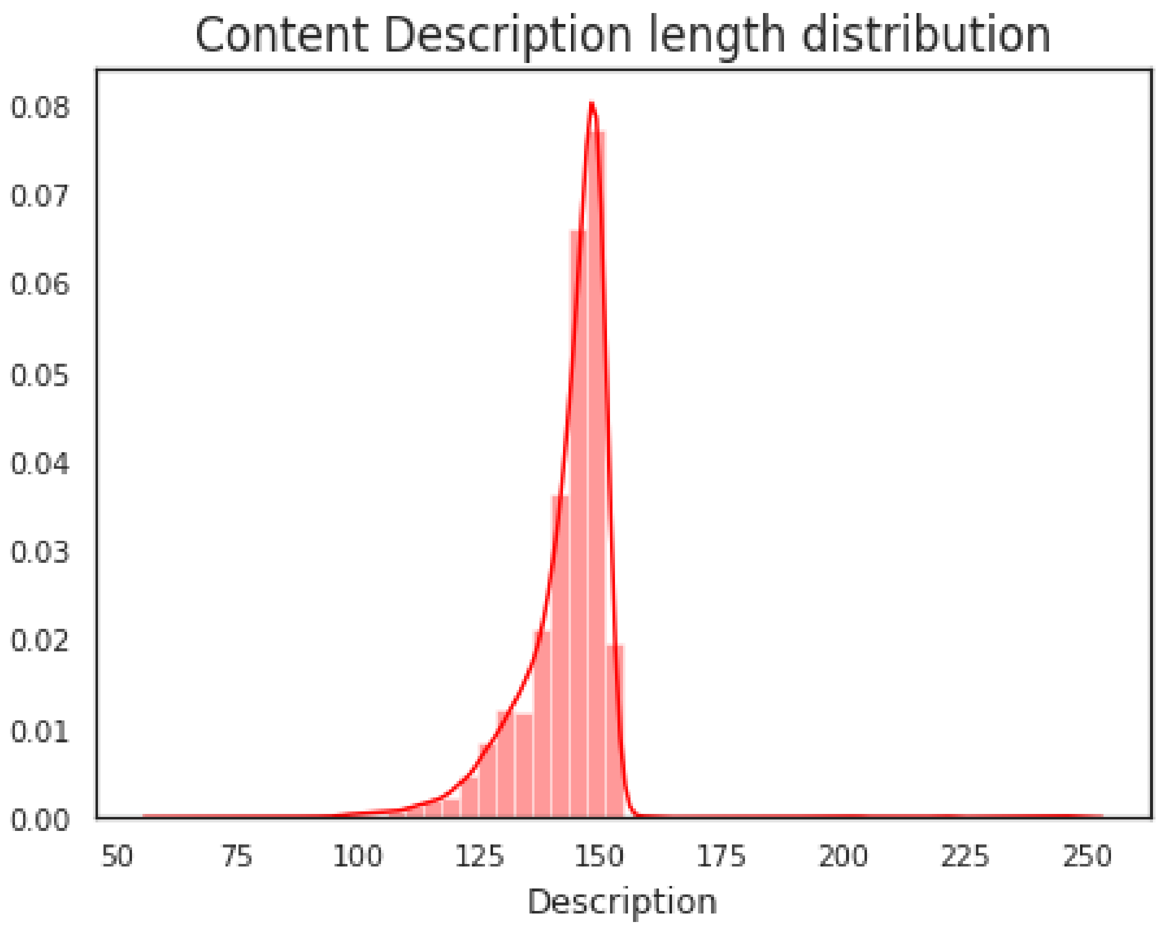
Proportion of TV Shows and Movies content appropriate for different ages

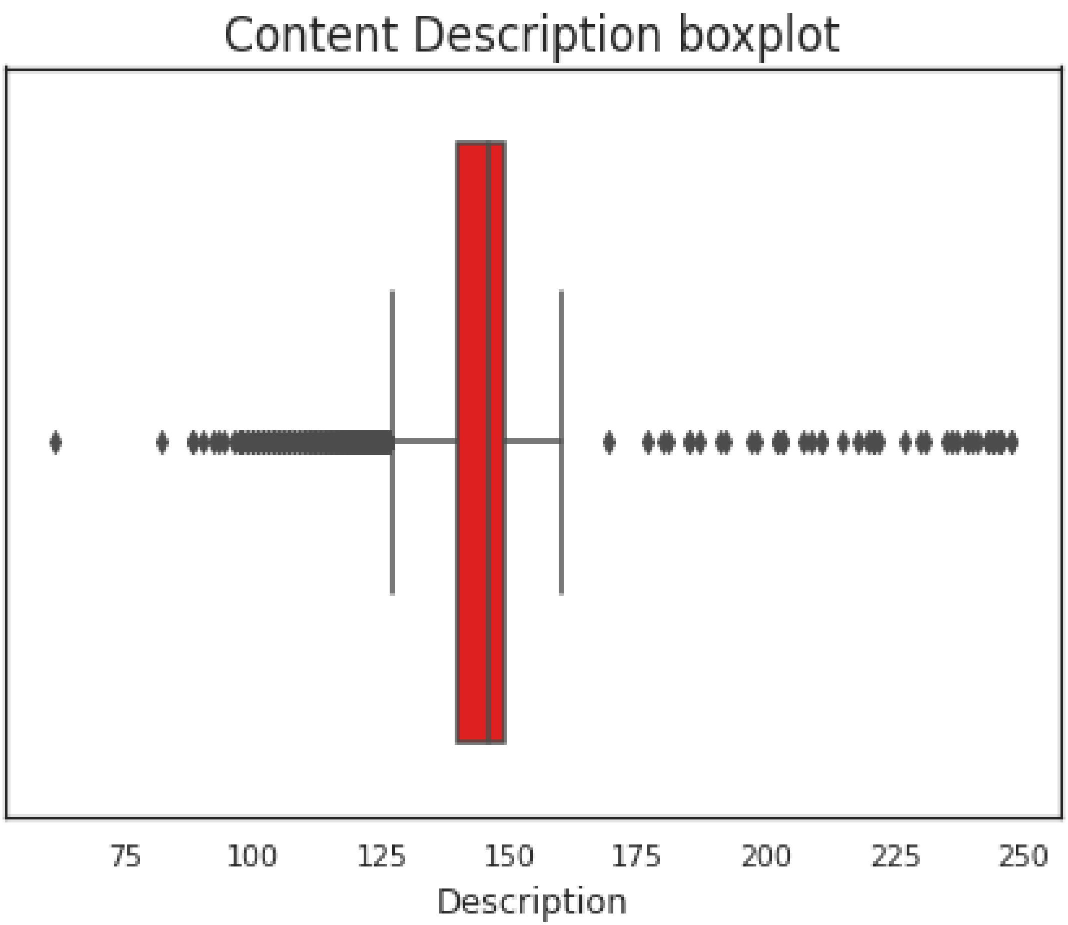


Content Rating from each country of Origin

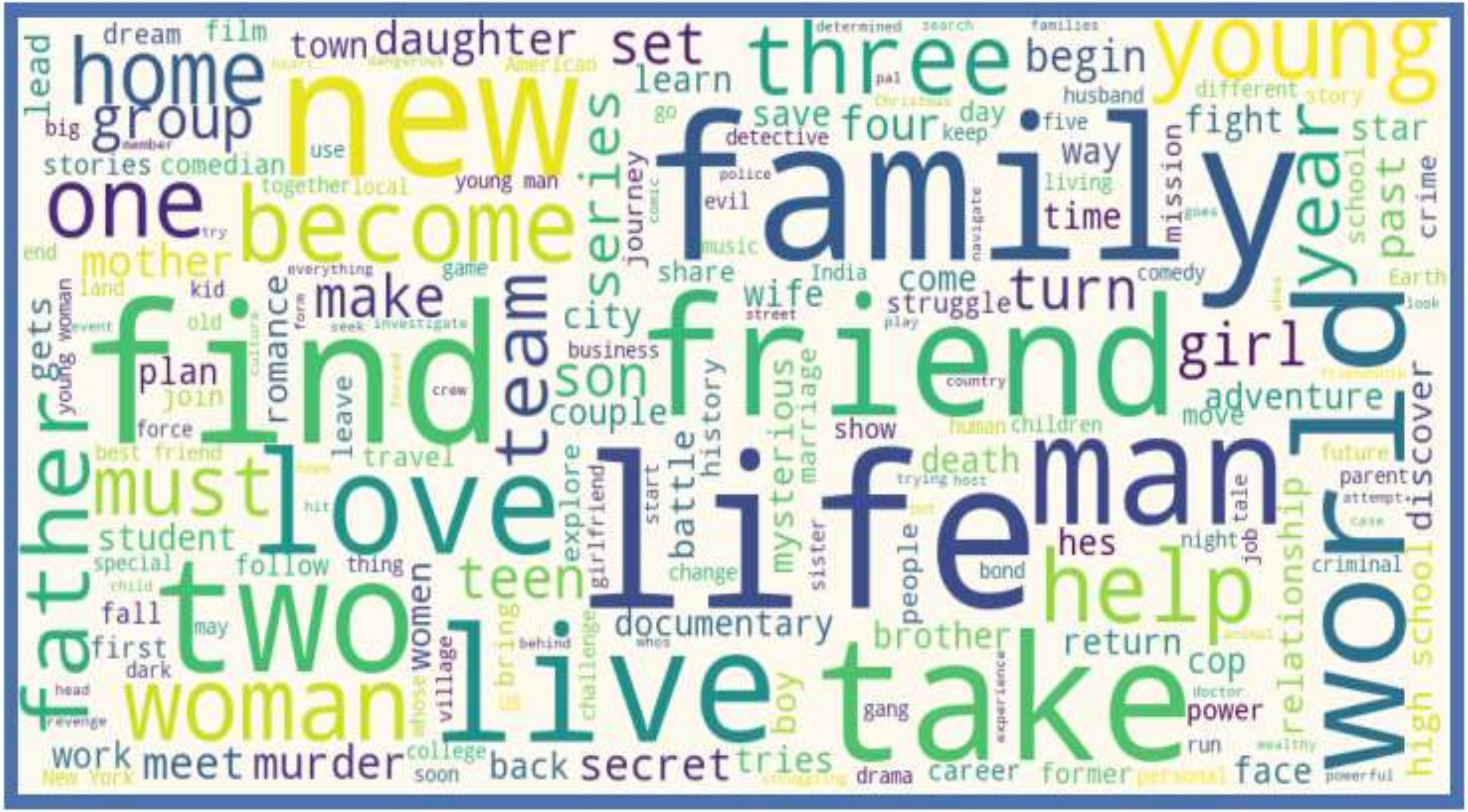


Length Distribution of words for Content Description:

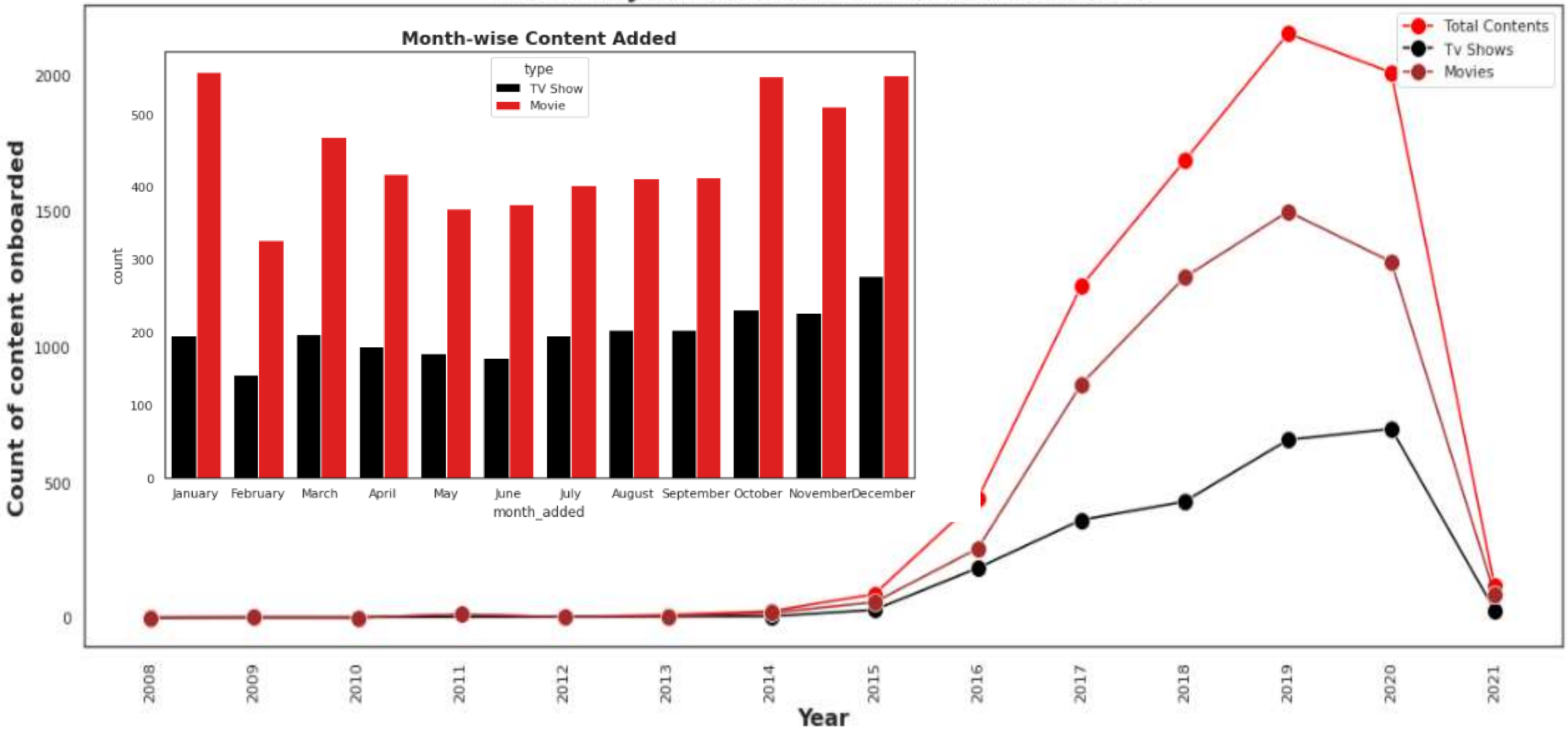




Frequency of words in the Content Description:

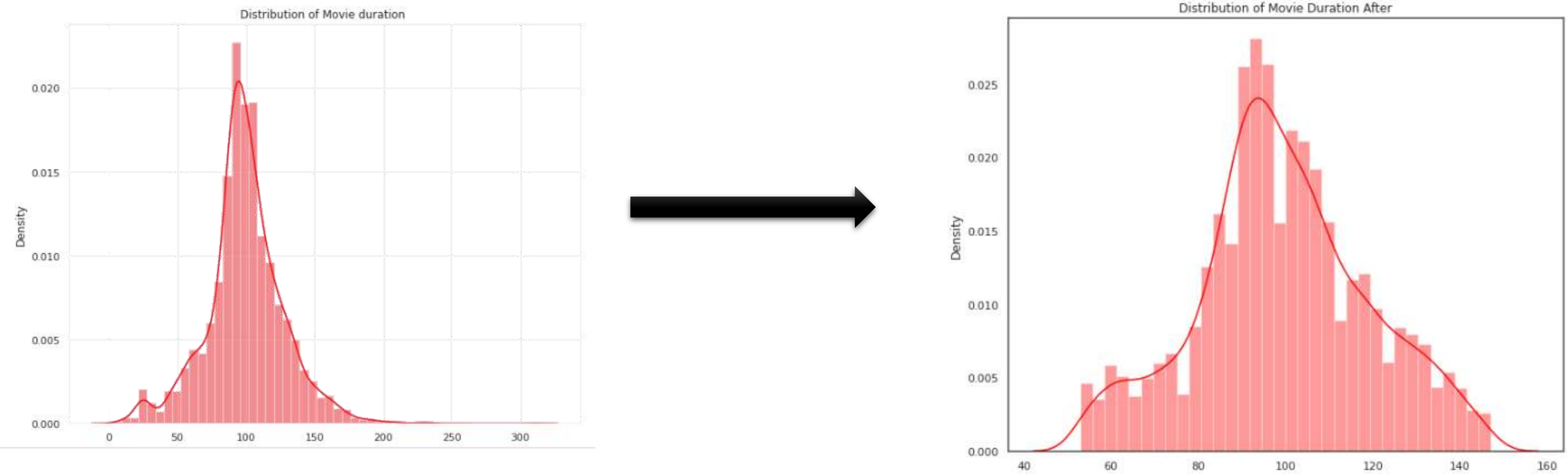


Trend of year-wise content added on Netflix



Outlier Detection:

- We Undergone various outlier method but the method eventually gave great results was the Inter-quantile method.
- As we have used this method, we had removed rows that are not betweeen 5 and 95% of the movie duration

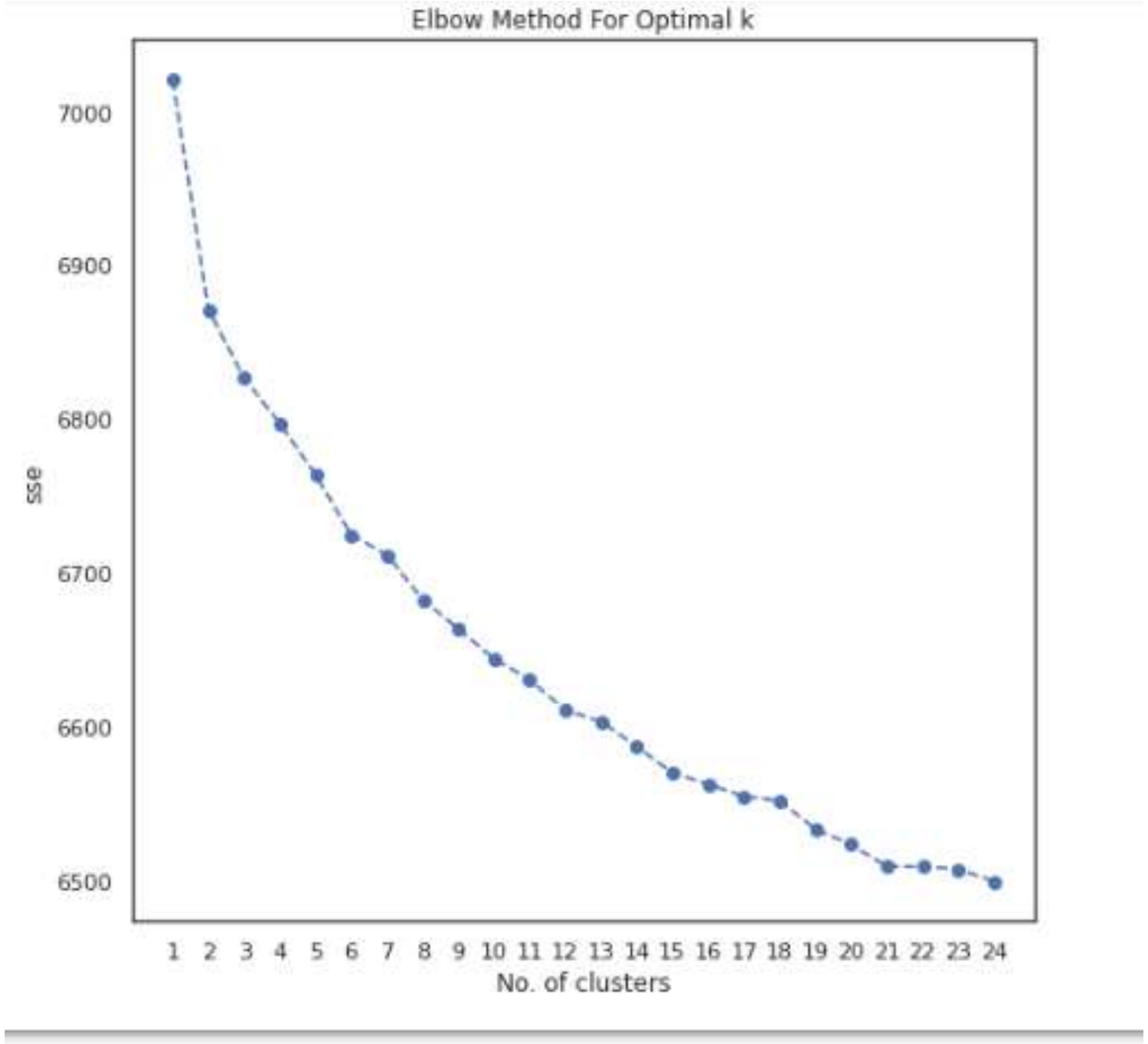


K-means Clustering (Elbow Method)

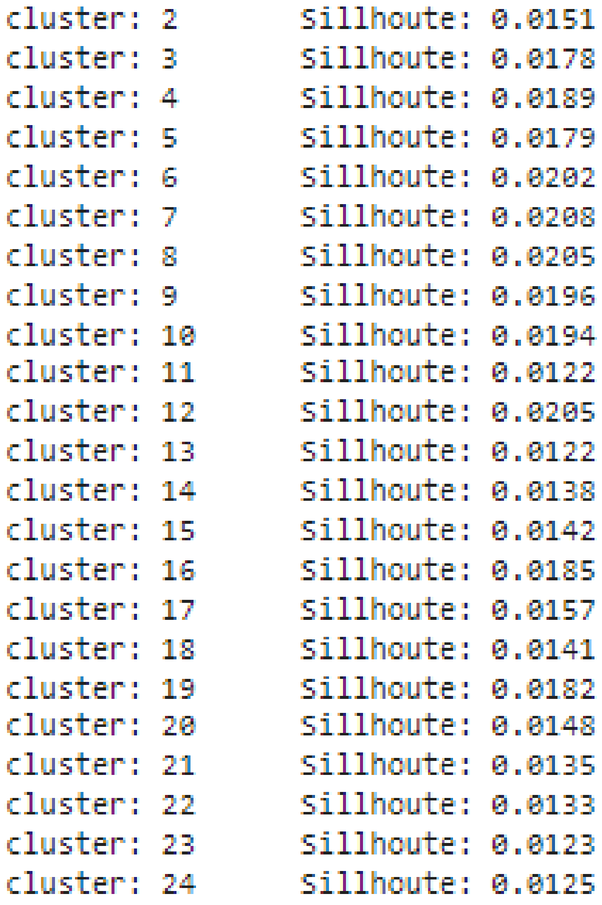
- k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster

- We initially imported K-means Clustreing through sklearn.cluster and set it to a range of about 25 clusters

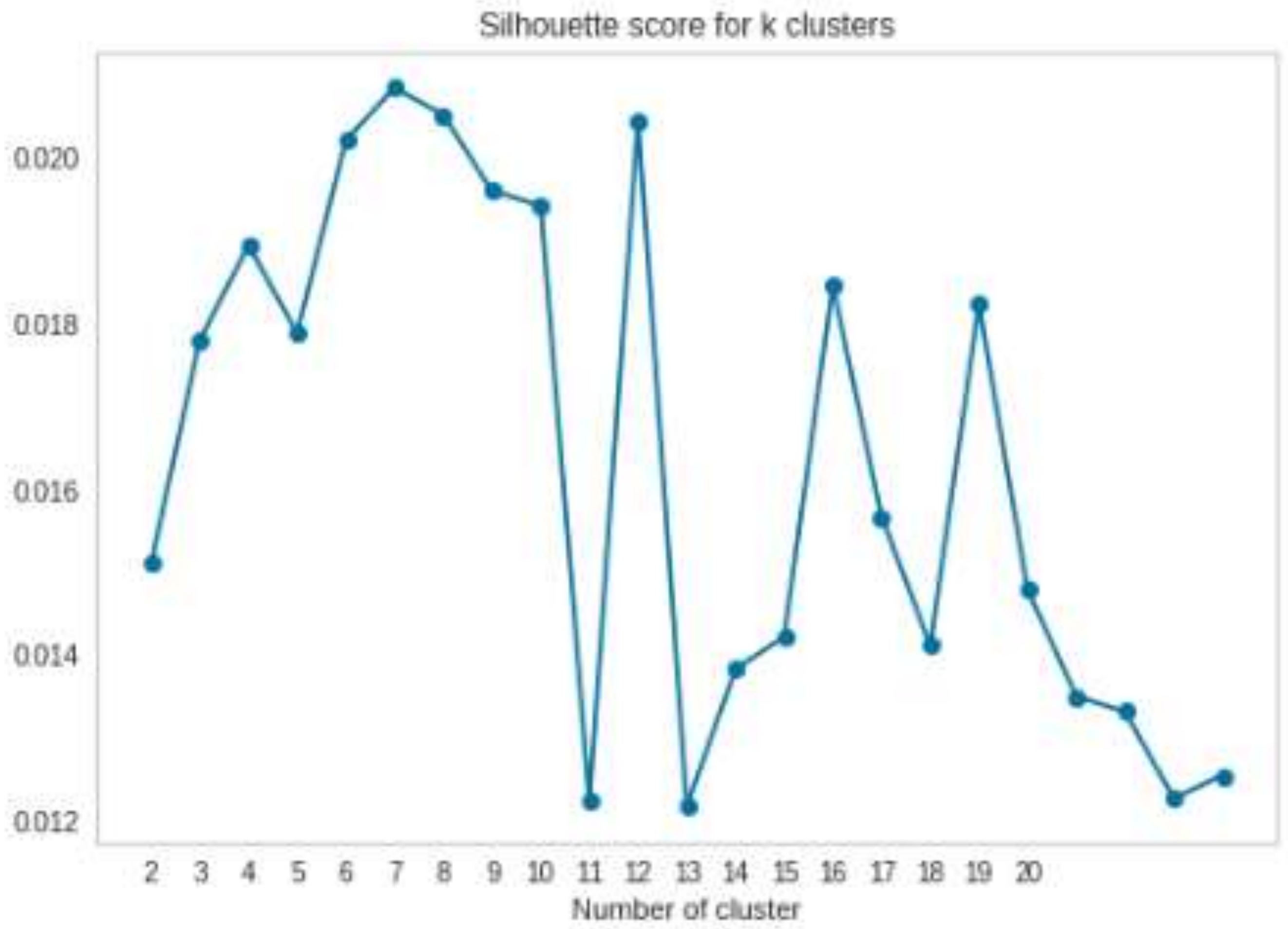
- The Line plot where the line takes a deviation in the form a “L” is found to be optimal which was also validated using the Sillhouette Score



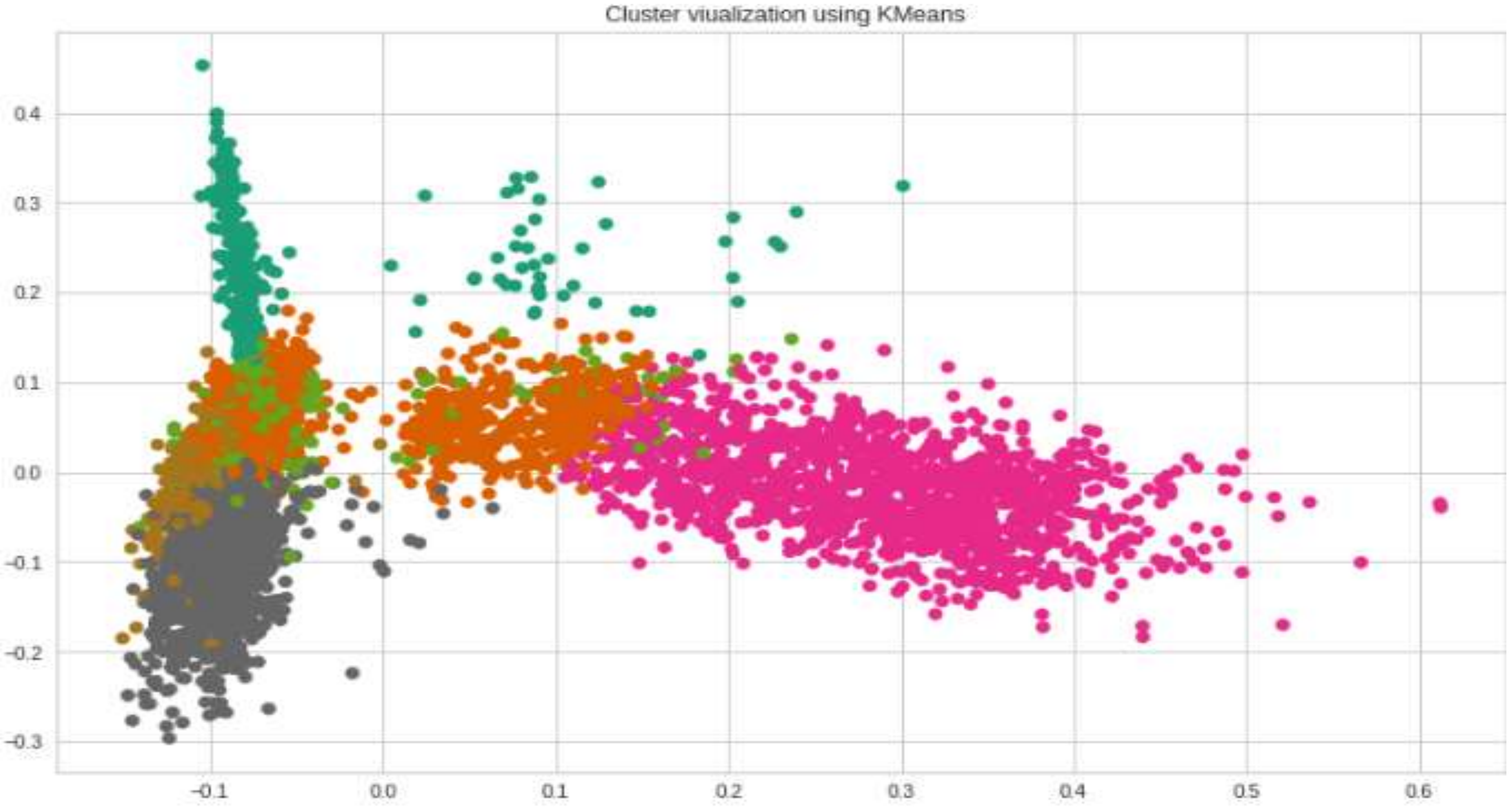
Sillhouette Score method:



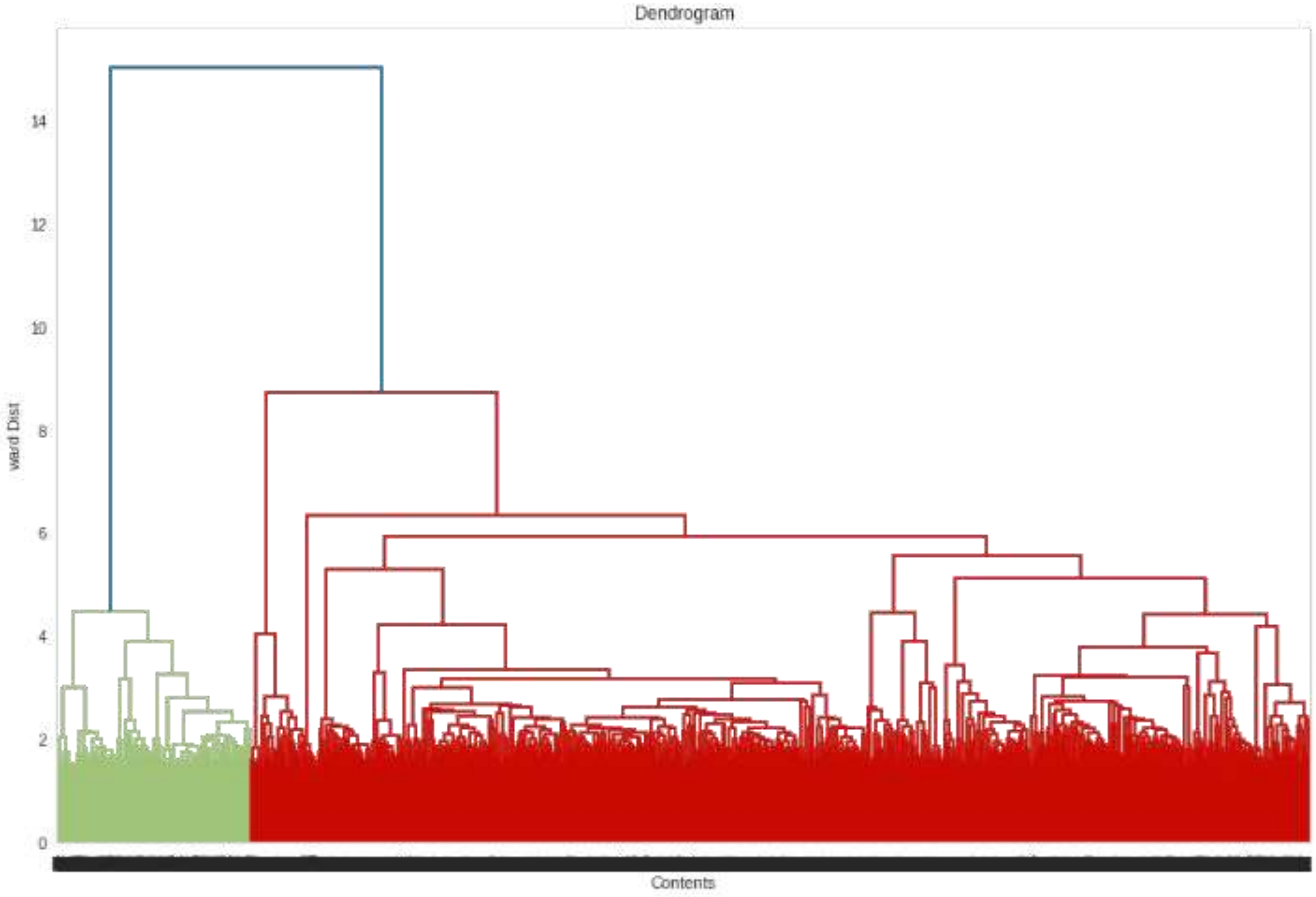
```
cluster: 2       Sillhoute: 0.0151
cluster: 3       Sillhoute: 0.0178
cluster: 4       Sillhoute: 0.0189
cluster: 5       Sillhoute: 0.0179
cluster: 6       Sillhoute: 0.0202
cluster: 7       Sillhoute: 0.0208
cluster: 8       Sillhoute: 0.0205
cluster: 9       Sillhoute: 0.0196
cluster: 10      Sillhoute: 0.0194
cluster: 11      Sillhoute: 0.0122
cluster: 12      Sillhoute: 0.0205
cluster: 13      Sillhoute: 0.0122
cluster: 14      Sillhoute: 0.0138
cluster: 15      Sillhoute: 0.0142
cluster: 16      Sillhoute: 0.0185
cluster: 17      Sillhoute: 0.0157
cluster: 18      Sillhoute: 0.0141
cluster: 19      Sillhoute: 0.0182
cluster: 20      Sillhoute: 0.0148
cluster: 21      Sillhoute: 0.0135
cluster: 22      Sillhoute: 0.0133
cluster: 23      Sillhoute: 0.0123
cluster: 24      Sillhoute: 0.0125
```



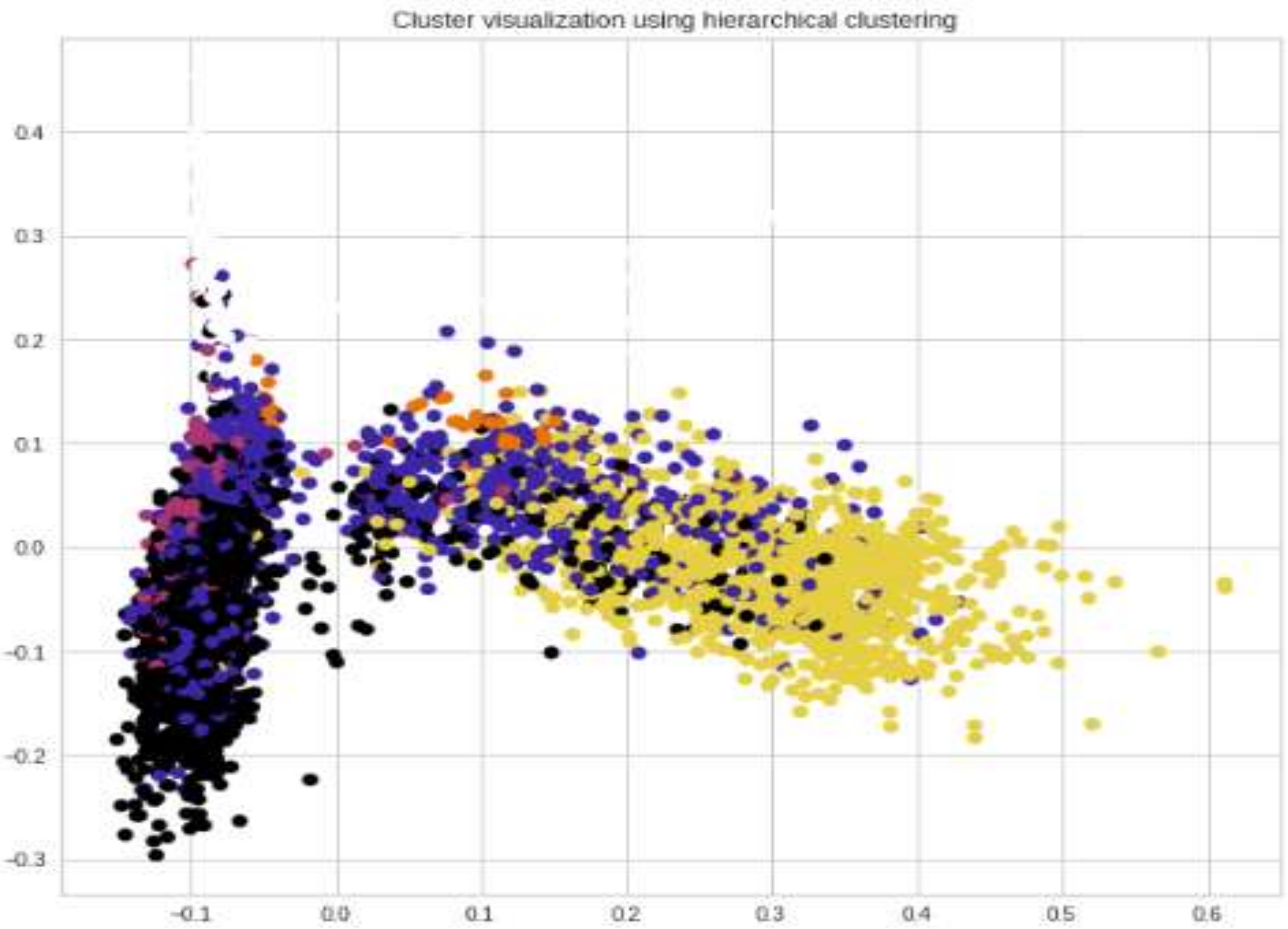
Applying the Model:



Dendrogram:



Hierarchial Clustering:



Conclusion:

Insights from Visualization

- Movies and TV Shows on Netflix are in ratio 7/3. And TV Shows have long way to catch upto the number of Movies.
- Most content on Netflix is from United States followed by India and United Kingdom. Netflix has a little over 90% of its total Content from top 12 countries, countries are namely, United States, India, United Kingdom, Canada, France, Japan, Spain, South Korea, Germany, Mexico, Chinaand Australia.
- Netflix has a lot of Japanese and South Korean TV Shows than the countries' movies and opposite in case of India.
- Majority of the TVshows have only one season and Also Most of the movies are of length ~90 mins
- Most content of Netflix is rated for mature audience only and majority of the movies are made for Teenage where as it is for Adults in the case of TVshows
- The average content description is around 140 words and the most used frequent words are life and Family

Insights from Clustering

- Using the Elbow method, we found that the optimal no. of clusters to be 6 which we also validated using the sillhoutte score method.
- We also have deployed Hierarchical Clustering with the same number of clusters obtained using the dendrogram

Challenges:

- ❖ Choosing the right plot for effective visualization consumed a lot of time.
- ❖ Choosing the correct method for outlier Detection was found tedious
- ❖ Since we had mostly textul/categorical features which became an obstacle while applying the clustering model.

Thank You