Capstone Project Unsupervised

NETFLIX MOVIES AND TV SHOWS CLUSTERING

TEAM MEMBERS

Abhishek Mishra

Kurva Mallesh

Arunesh Mishra

WHAT IS NETFLIX MOVIES AND TV SHOWS CLUSTERING

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.

Objective of The Project

The main objective of this project is to find the business insider from the data. Also create the model which can create Clusters for given dataset.

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Fixable which is a third-party Netflix search engine.

Business Problems

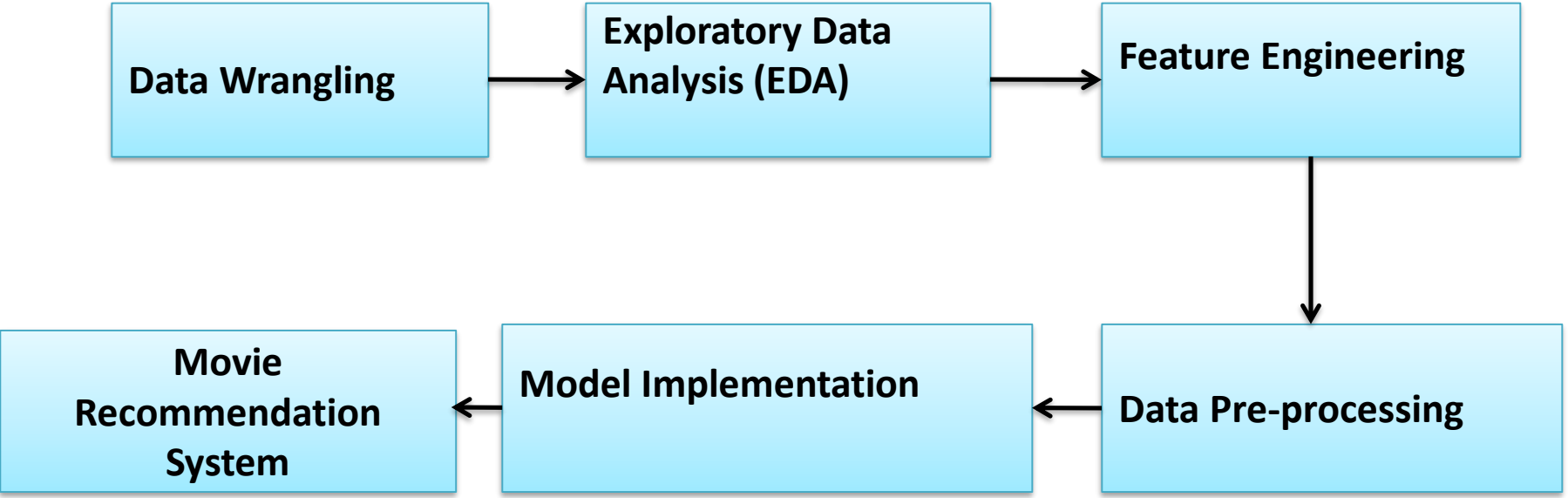
Dealing with null values in this dataset.

Check Duplicate values in the dataset?

Understanding what type content is available in different countries.

Clustering similar content by matching text-based features

Road map of the project



Data Wrangling

- ✓ We have 7787 rows and 12 columns provided in the data.
- ✓ In the dataset we have 11 object columns and 1 integer column as release year.
- ✓ Fist we have 2389 null values in director column.We have almost 30% null values in this colum n so we can not use this column in model training but we can use it in EDA.
- ✓ We have 718 null values in cast column. and it can be replaced with 'unknown'.
- ✓ we have 507 null values in country column. Replacing nulls with 'mode'.
- ✓ Also we have 10 null values in date_added column.we have few rows of date_added so we can 'drop' these rows.

Exploratory Data Analysis (EDA)

- ✓ According to the graph we have 5377(69.14%) movies.
- ✓ And 2400(30.86%) as TV Show in this dataset.
- ✓ According to plot we can say **Raul Campos** and **Jan Sulter** collectively have the most content o n Netflix.
- ✓ **Marcus Raboy** have the second most content on Netflix.
- ✓ Now we can say in this data **Anupam Kher** having 38 number of listing.
- ✓ According to the plot we can understanding United States have 2080 Movies and 975 TV Show
- ✓ We have so many content release in October(785), November(738), December(833) and Janu ary(757) may be it is because of Holiday season.

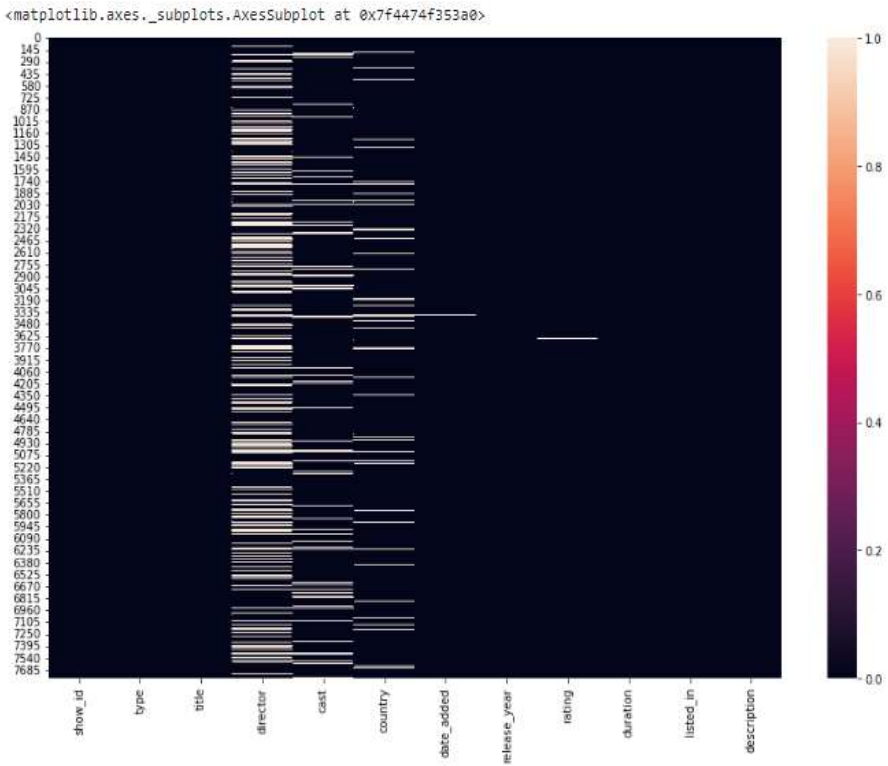
Feature Engineering

- ✓ For train the model we use description column, listed_in column, rating column, country column, title column, director column, cast column.
- ✓ convert all words in lowercase.
- ✓ We remove all stopwords. Also use stemming function.
- ✓ We performed TF-IDF vectorizer
- ✓ Applying PCA-Principal Component Analysis to reduce dimensions

Model Implementation

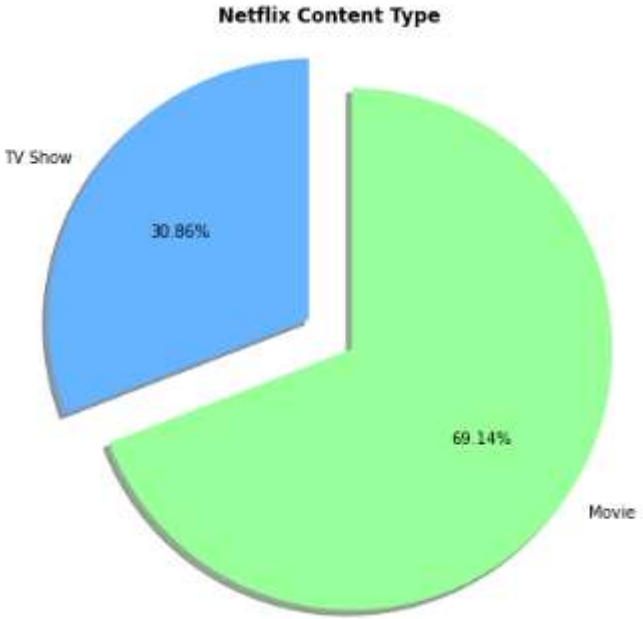
- ✓ Find the value of clusters.
- ✓ WE use Elbow method for finding k values.
- ✓ Also use Silhouette Score for best score.
- ✓ Also use Dendogram for finding the value of clusters.
- ✓ Use Agglomerative Clustering.
- ✓ Use KMeans Clustering.
- ✓ Analysis of cluster.

Null value treatment



- ✓ Fist we have 2389 null values in director column.We have almost 30% null values in this column so we can not use this column in model training but we can use it in EDA.
- ✓ We have 718 null values in cast column. and it can be replaced with 'unknown'.
- ✓ we have 507 null values in country column. Replacing nulls with 'mode'.

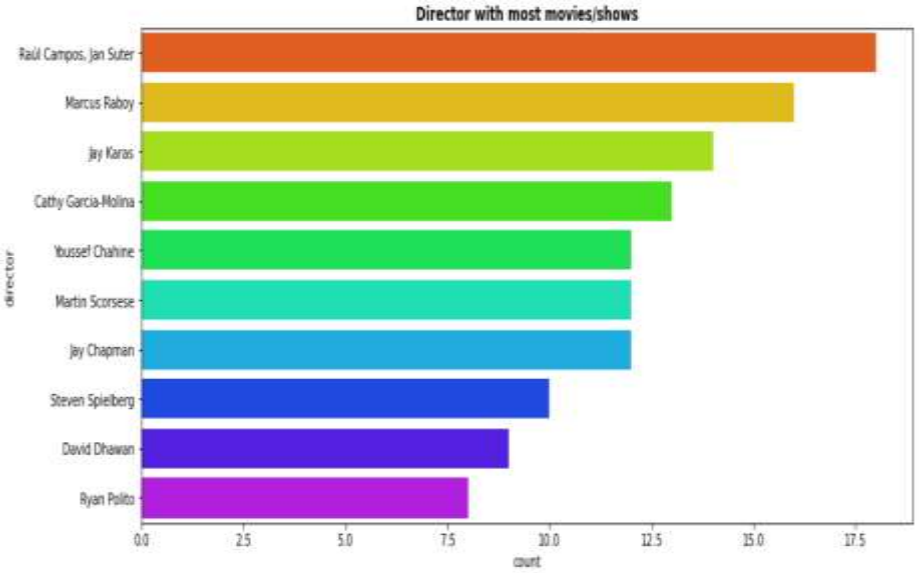
Learn about “Type" Column



✓According to the graph we have 5377 (69.14%) movies

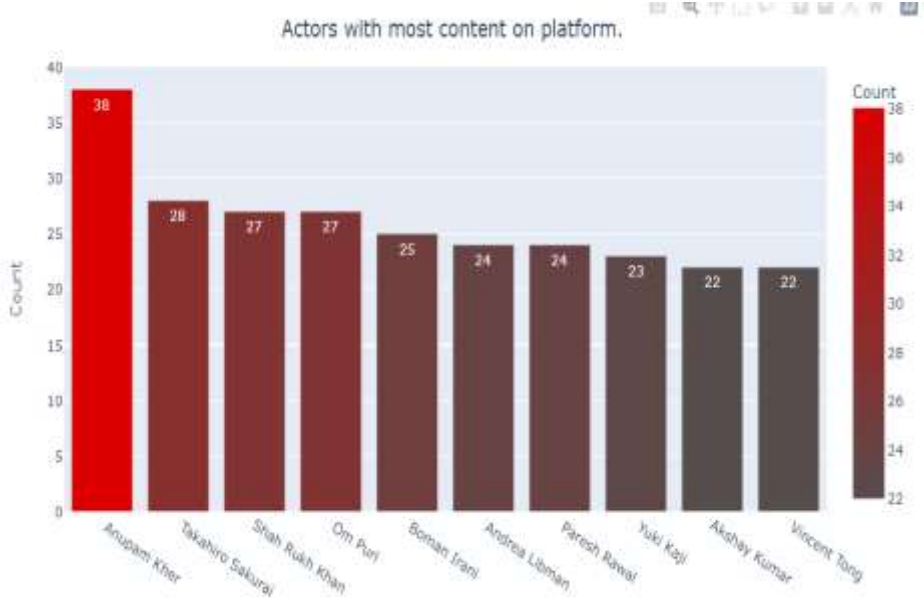
✓And 2400(30.86%) as TV Show in this dataset.

Learn about ‘director’ column



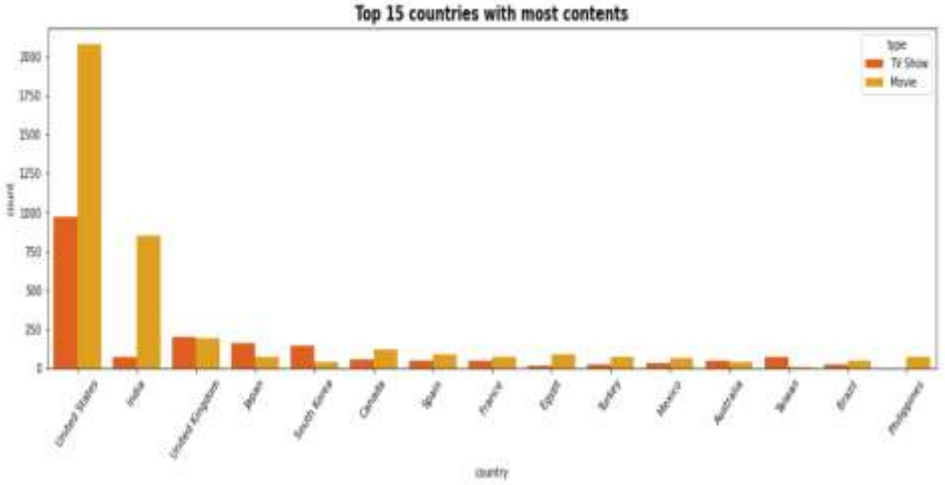
- ✓ According to plot we can say Raul Campos and Jan Sulter collectively have the most content on Netflix.
- ✓ Marcus Raboy have the second most content on Netflix.

Learn about 'cast' column



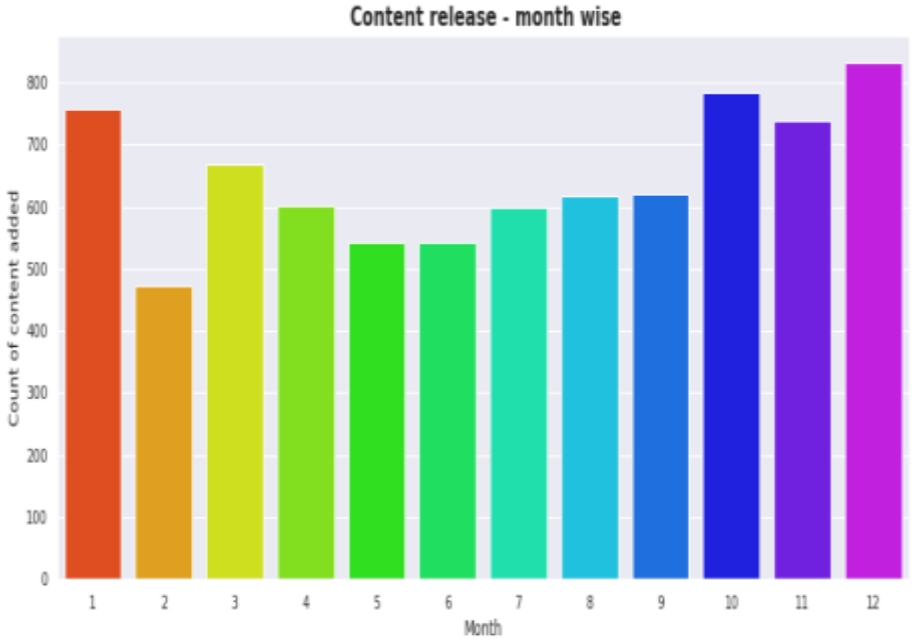
- ✓ Now we can say in this data Anupam Kher having 38 number of listing.
- ✓ Takahiro Sakurai is the second most listed actor on netflix.
- ✓ Shah Rukh Khan is the 3rd most listed actor on netflix.

Learn about 'country' column



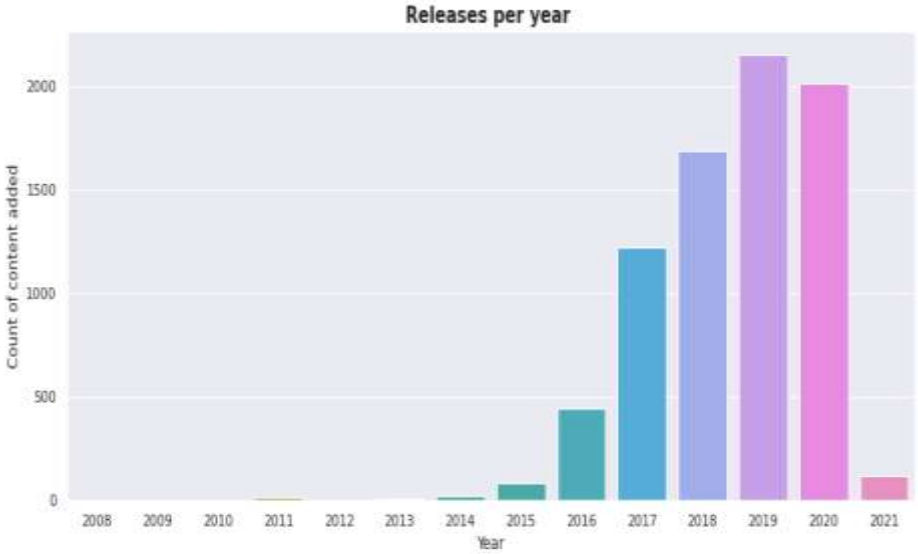
- ✓ According to the plot we can understanding United States have 2080 Movies and 975 TV Show.
- ✓ INDIA have second most listed country with 852 movies and 71 TV Show on Netflix.

Month wise Content release analysis



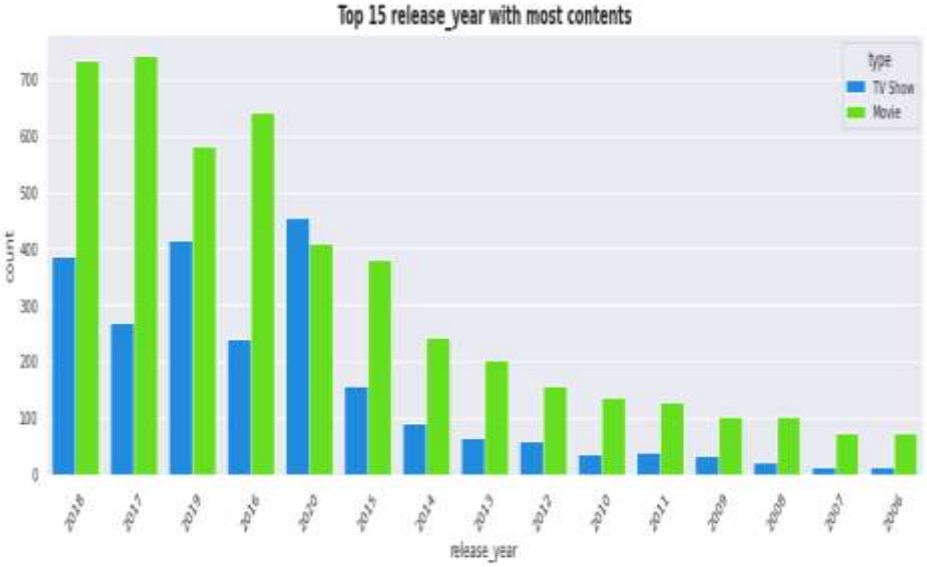
- ✓ We have so many content release in October(785), November(738), December(833) and January(757) may be it is because of Holiday season.

year wise Content release analysis



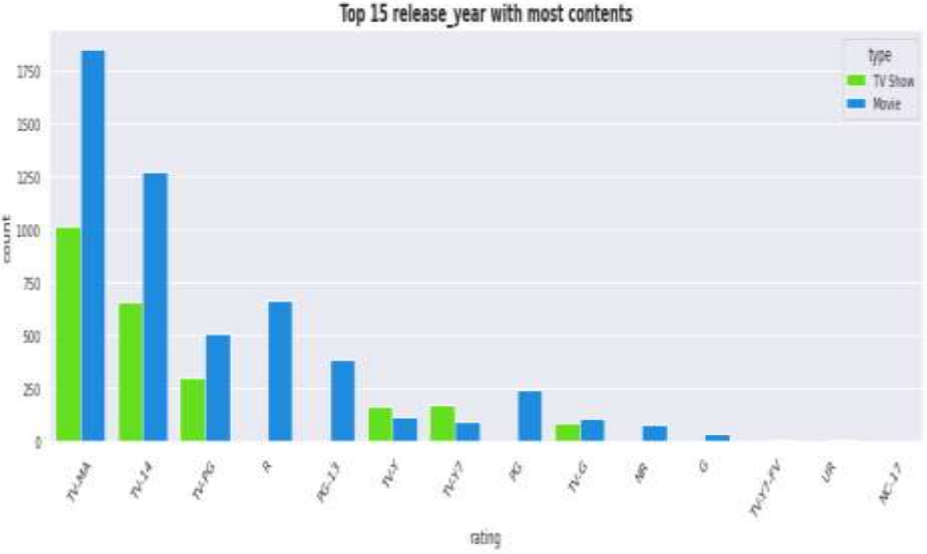
- ✓ The number of release have significantly increased after 2015 to 2020.
- ✓ But sudden drop in 2021 may be it is because of covid 19.

Learn about release year' column



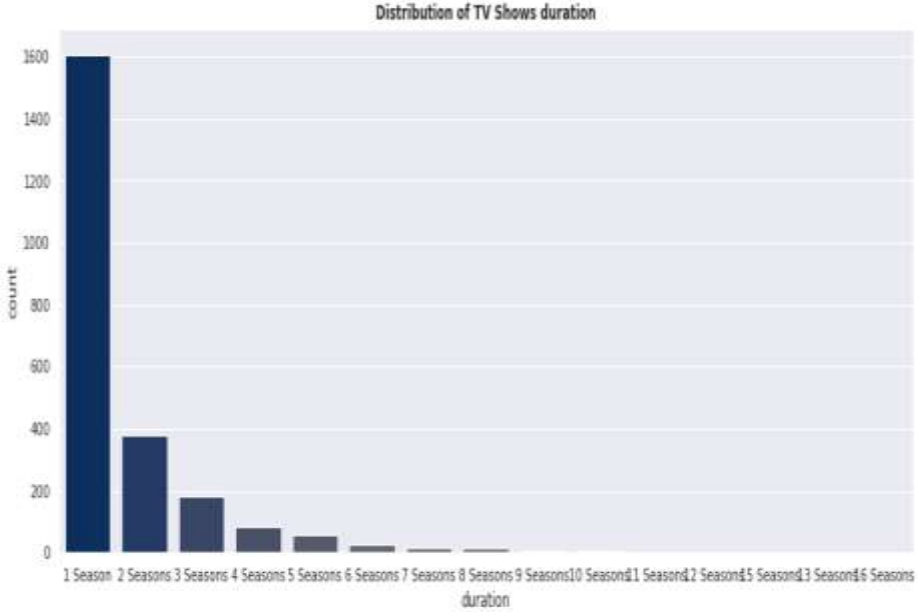
- ✓ We have 744 movies and 268 TV Show release in 2017.
- ✓ Also 734 movies and 386 TV Show release in 2018.
- ✓ 82%(6431) of the content was released between 2010 and 2021.
- ✓ 17.28%(1346) of the content was released before 2010.

Learn about 'Rating' column



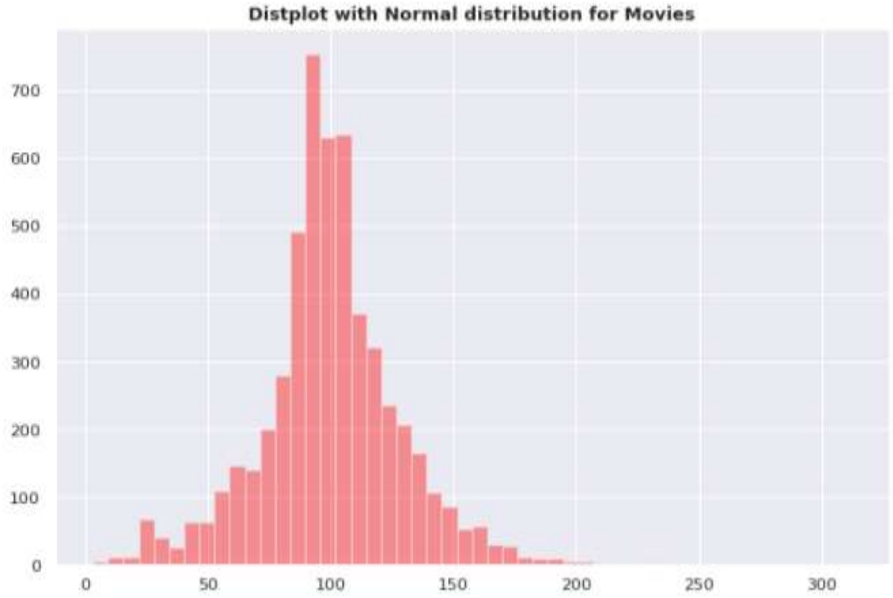
- ✓ Most number of movies rated TV-MA i.e. Adult Rating.
- ✓ Most number of TV Shows rated TV-MA i.e. Adult Rating.

Learn about 'duration' column for TV Show



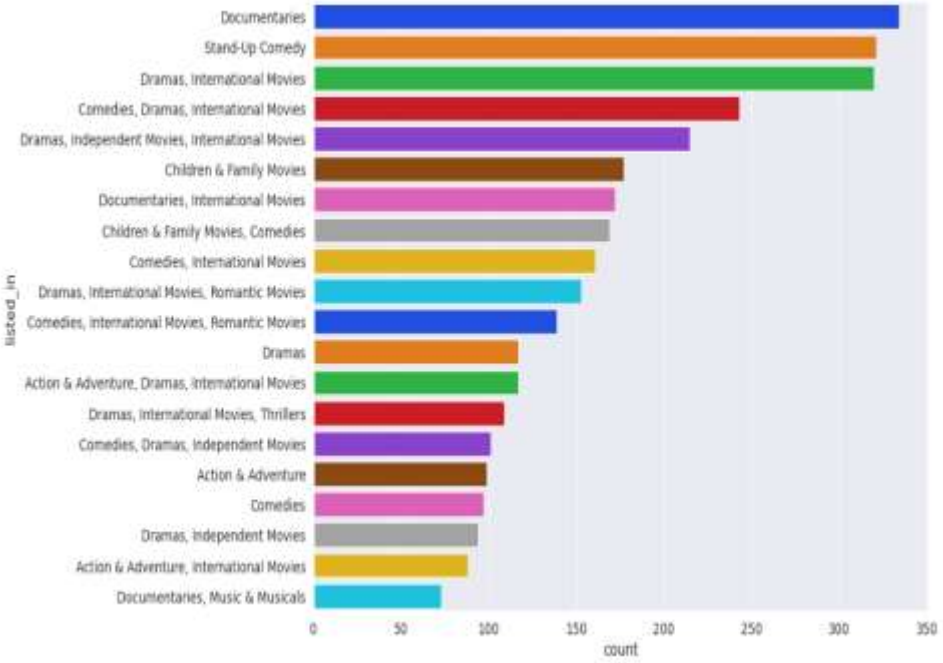
- ✓ We have most listed duration as season 1 with 1608 listing.
- ✓ We have second most listed duration as season 2 with 378 listing.

Learn about 'duration' column for movies



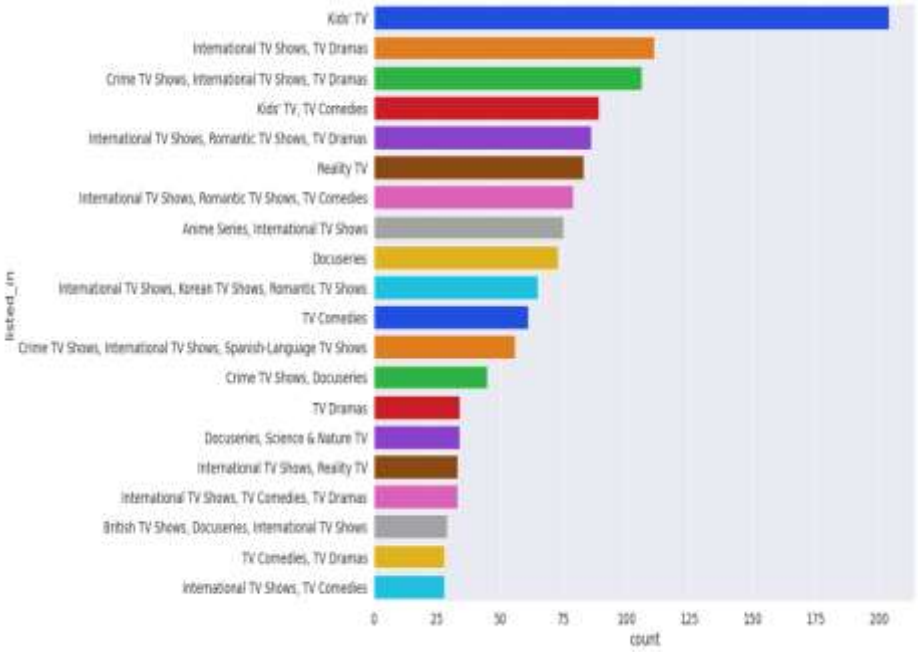
- ✓ Mainly the movie duration is in b/w 55 to 150 minutes.
- ✓ Most of the movies list for 90 to 120 minutes.

Learn about ' genera ' for movies column



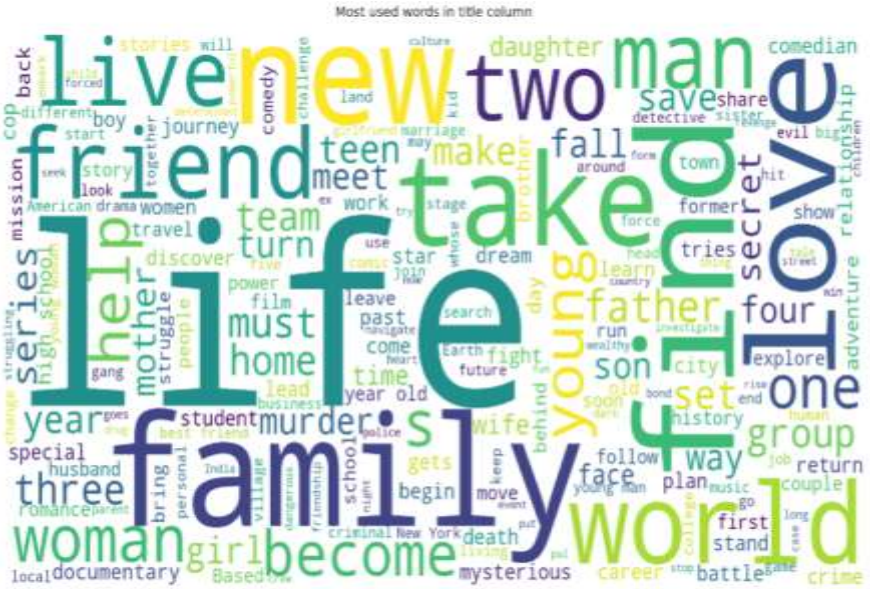
- ✓ In Movies Documentaries is the most popular genera on Netflix.
- ✓ Comedy is the second most popular genera on Netflix.

Learn about ' genera for TV-Show ' column



- ✓ In TV Shows Drama is the most popular genera.
- ✓ International TV shows is the second most popular genera.

Learn about 'title' column



- ✓ Most repeated words in title column are love, Christmas, World, Man, and life.

Learn about 'title' column

- ✓ Most repeated words in the description of the TV shows and movies are Family, new, Love, Life, mother, find.

EDA Conclusion

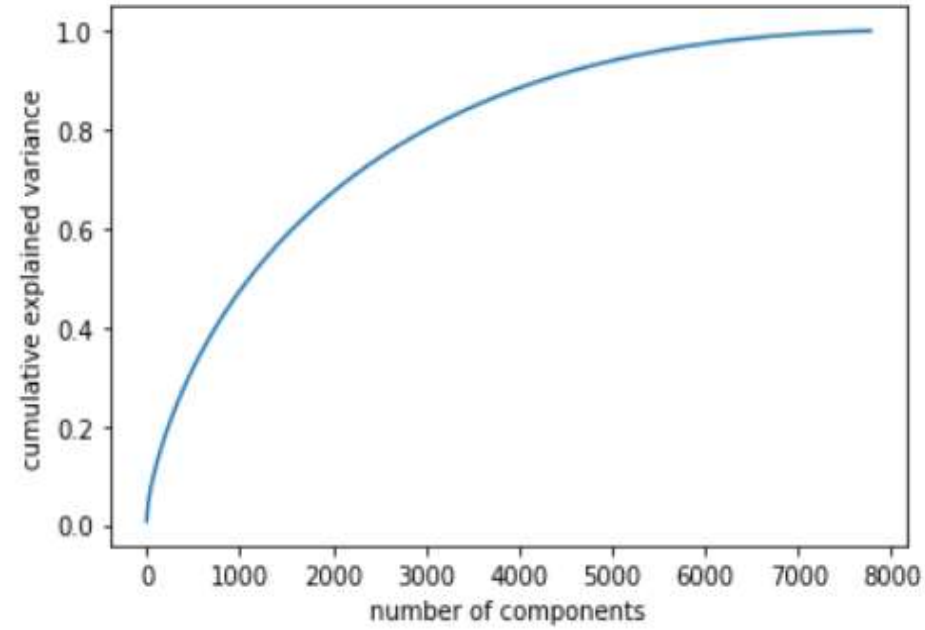
- ✓ The number of release have significantly increased after 2015 to 2020.
- ✓ But sudden drop in 2021 may be it is because of covid 19.
- ✓In TV Shows Drama is the most popular genera.
- ✓Most repeated words in title column are love, Christmas, World, Man, and life.
- ✓We have 744 movies and 268 TV Show release in 2017.
- ✓Also 734 movies and 386 TV Show release in 2018
- ✓ 82%(6431) of the content was released between 2010 and 2021
- ✓17.28%(1346) of the content was released before 2010.
- ✓Most number of movies rated TV-MA i.e. Adult Rating
- ✓Most number of TV Shows rated TV-MA i.e. Adult Rating
- ✓We have most listed duration as season 1 with 1608 listing.
- ✓We have second most listed duration as season 2 with 378 listing.
- ✓Mainly the movie duration is in b/w 55 to 150 minutes.
- ✓Most of the movies list for 90 to 120 minutes.
- ✓In Movies Documentaries is the most popular genera on Netflix.
- ✓Comedy is the second most popular genera on Netflix.

Feature Engineering

- For train the model we use description column, listed_in column, rating column, country column, title column, director column, cast column.
- convert all words in lowercase
- We remove all stop words.
- Also use stemming function.

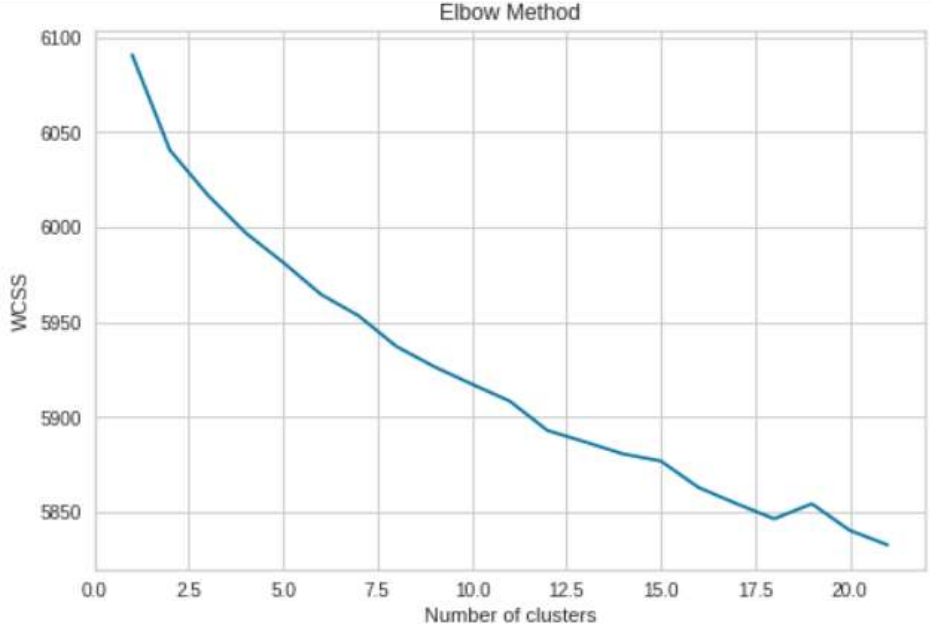
Cumulative Explained Variance

```
Text(0, 0.5, 'cumulative explained variance')
```



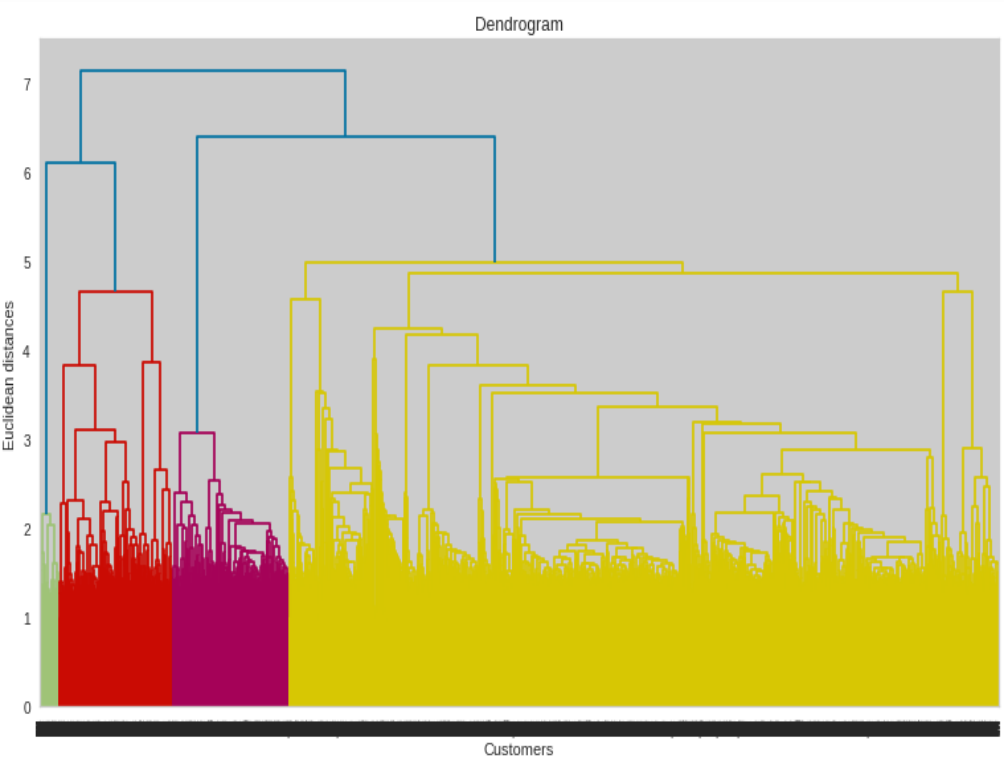
✓We will use 3000 components

Elbow Method for KMeans Clustering



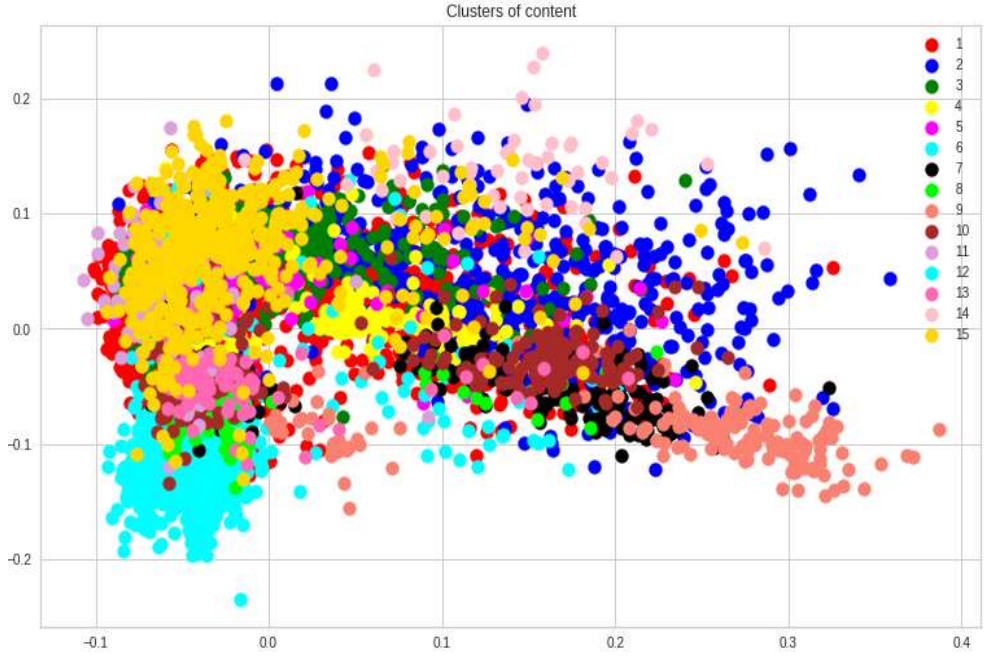
✓we will take no. of clusters as 15

Dendogram



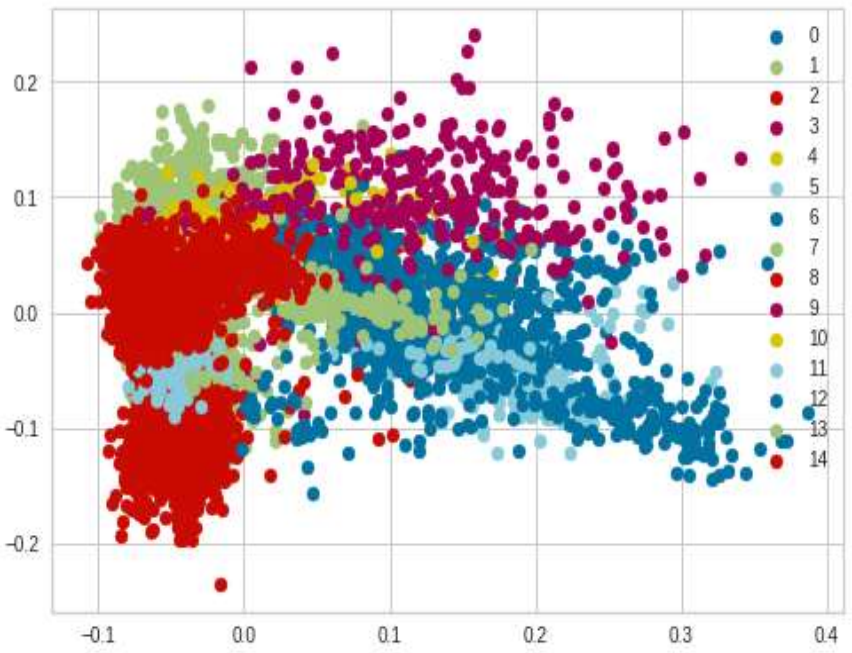
✓we will take no. of clusters as 15

Agglomerative Clustering



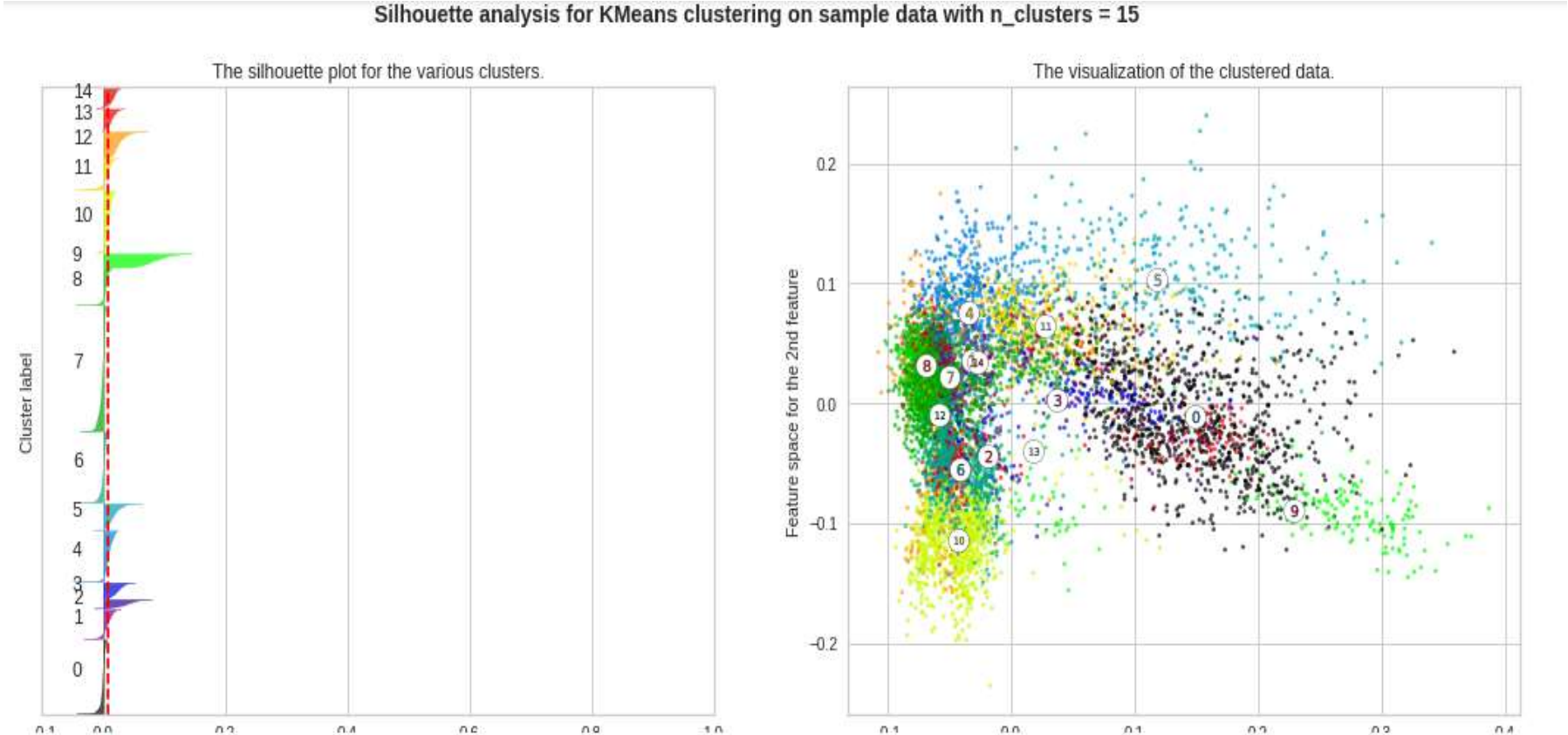
- ✓For n_clusters = 2,silhouette score 0.001154
- ✓ For n_clusters = 3, silhouette score is 0.001837
- ✓For n_clusters = 4, silhouette score is -0.004009
- ✓For n_clusters = 5, silhouette score is -0.003056
- ✓For n_clusters = 6, silhouette score is -0.002319
- ✓For n_clusters = 7, silhouette score is -0.001574
- ✓For n_clusters = 8, silhouette score is -0.001121
- ✓For n_clusters = 9, silhouette score is -0.0004279
- ✓For n_clusters = 10, silhouette score is 0.000220
- ✓For n_clusters = 11, silhouette score is -0.000294
- ✓For n_clusters = 12, silhouette score is 0.000230
- ✓For n_clusters = 13, silhouette score is 0.000521
- ✓For n_clusters = 14, silhouette score is 0.000385
- ✓ For n_clusters = 15, silhouette score is 0.000817

KMeans Clustering



✓Silhouette Coefficient: 0.004

Silhouette Score for KMeans Clustering



Movie Recommendation System

```
Enter your text here:ghost
the movies suggest for you:

1 - Eugenie Nights
2 - Al Hayba
3 - Because We're Heading Out
4 - 122
5 - El-Khawaga's Dilemma
6 - الف مبروك
7 - Son Of Adam
8 - The Land of Hypocrisy
9 - Juman
10 - The Dealer
11 - The Platform
12 - Secret of the Nile
13 - More to Say
14 - Game Over
15 - The Thief and the Imbecile
16 - An Hour and a Half
17 - The Land
18 - Cairo Station
19 - Convict
20 - Warda
21 - Border Security: America's Front Line
22 - My Pride
23 - Paranormal
24 - The Road to El Camino: Behind the Scenes of El Camino: A Breaking Bad Movie
25 - Return of the Prodigal Son
26 - El desconocido
27 - Disappearance
28 - Valentino
29 - Scarecrow
30 - Find Yourself
```

Try to create a recommendation system if I search ghost in this recommendation system so it suggest these 30 titles from the dataset of Netflix

Conclusion from Model

WE use Elbow method for finding k values.
Also use Silhouette Score for best score.
Also use Dendogram for finding the value of clusters.
Here are few clusters with there word cloud graph

Analysis of cluster 0

Type - Movie, TV Show
Title- Naruto, high, girl, low, movie, dragon, bleach, fate, battle
Countries- Japan, US, India
Ratings- TV-MA, PG, Y7
Genres- International TV series- Anime
Description- family, world, human, friend

Conclusion from Model

Analysis of cluster 1

Type - Movies, TV Show

Title- master, love, Dorgan, aur, Mumbai, Singh

Countries- India, China, Honcong

Ratings- TV-MA, pg

Genres- International movies, Dramas, Action

Description- family, man, love, India, woman, find

Analysis of cluster 2

Type - Movies, TV Show

Title- club, Spain, live, holy

Countries- Spain, Mexico, France

Ratings- TV-MA, NR, PG

Genres- Dramas International , show International

Description- family, young, secret, story

Conclusion from Model

Analysis of cluster 3

Type - Movies, TV Show

Title- Girl, man, love, monster, holiday etc

Countries- United states, United kingdom, Japan etc

Ratings- TV-MA,PG etc

Genres- family movies, movie comedies etc

Description- find, save, find, new, school, etc

Analysis of cluster 4

Type - TV Shows, Movies etc.

Title- Power rangers, adventure, stories, rescue, bheem, little, monster etc.

Countries- US, France, UK, Japan etc.

Ratings- TV-Y7 etc.

Genres- Kids shows-comedy,korean etc.

Description-adventure, friend, world, anime etc.

Thank You