Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

BY-HRITIK SHARMA

DISCUSSION POINTS

- About Netflix
- Problem Statement
- Data Summary
- Exploratory data analysis
- K-means
- Dendogram
- Silhouette analysis for agglomerative clustering
- Interpretation based on model
- Conclusion

Netflix

Netflix is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through duration 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.

As of December 31, 2021, Netflix had over 221.8 million subscribers worldwide.

Netflix can be accessed via internet browsers on computers, or via application software installed on smart TVs, set-top boxes connected to television, tablet, computers, smartphone etc.

Problem Description

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 services 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.

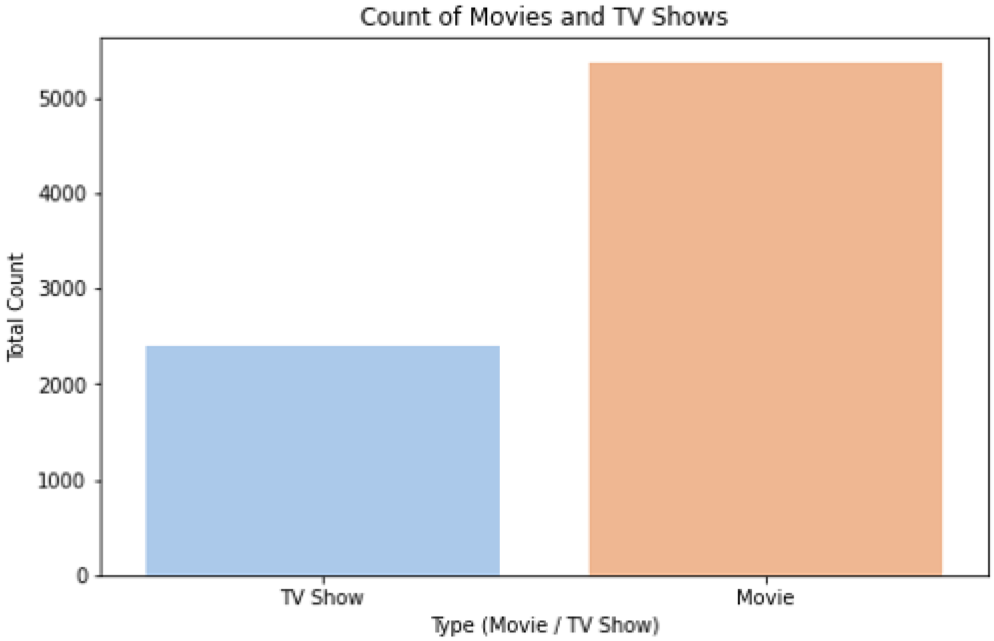
DATA SUMMARY

- 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

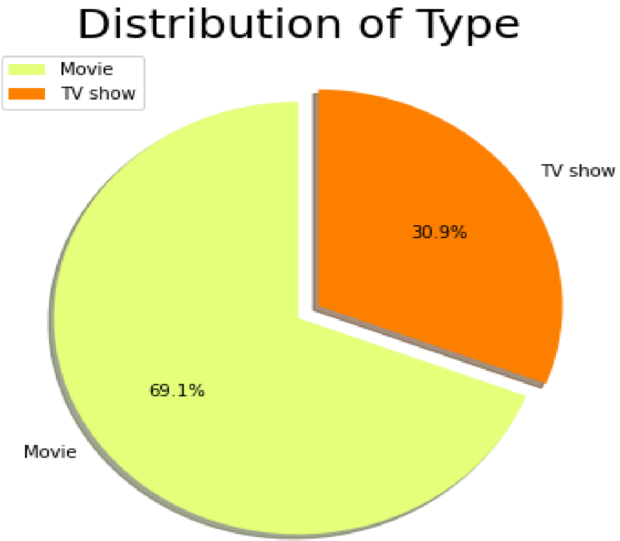
DATA SUMMARY

- date_added : Date it was added on Netflix
- release_year : Actual Release Year of the movie / show
- rating : TV Rating of the movie / show
- duration : Total Duration - in minutes or number of seasons
- listed_in : Genre
- Description: The Summary description

EDA

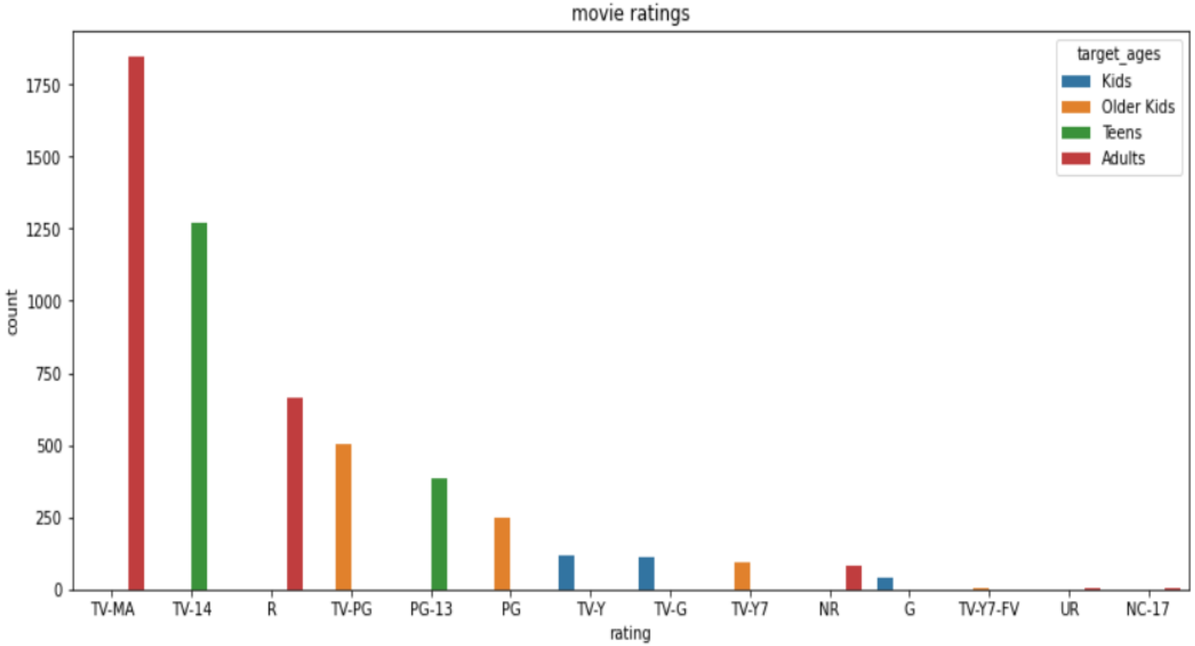


Count of Movies and TV Shows
Total Count
0
1000
2000
3000
4000
5000
TV Show
Movie
Type (Movie / TV Show)



Distribution of Type
Movie
TV show
TV show
30.9%
69.1%
Movie

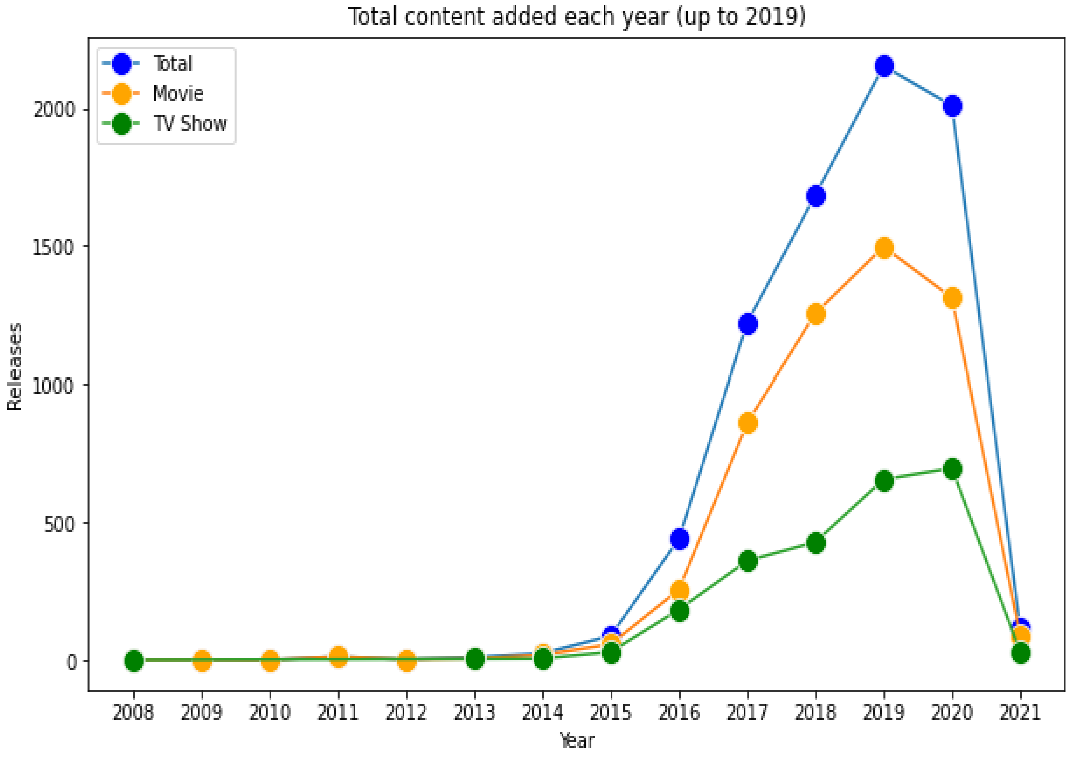
Continue……



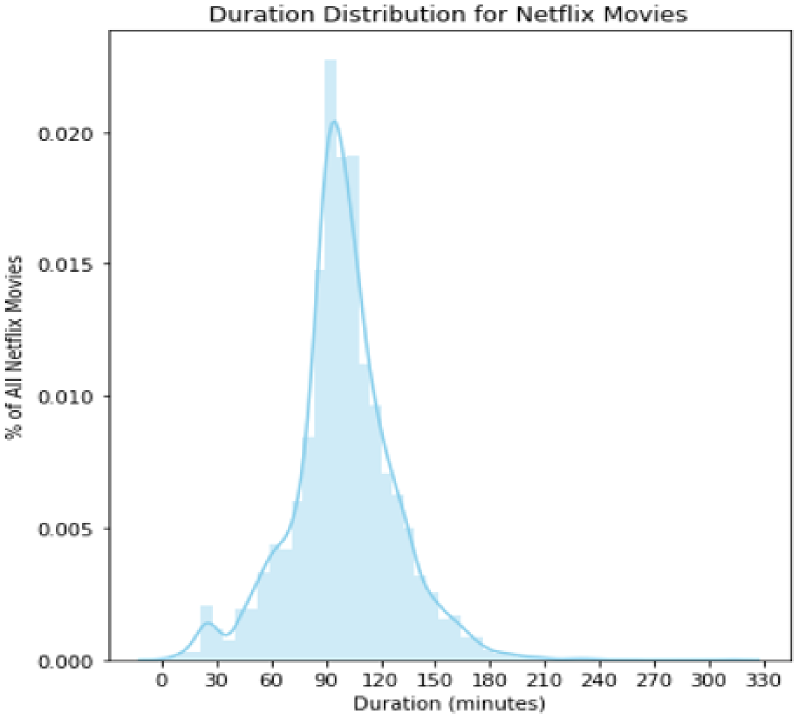
TV-MA has the highest number of ratings for tv shows i.e adult ratings in both the cases TV-MA has the highest number of ratings

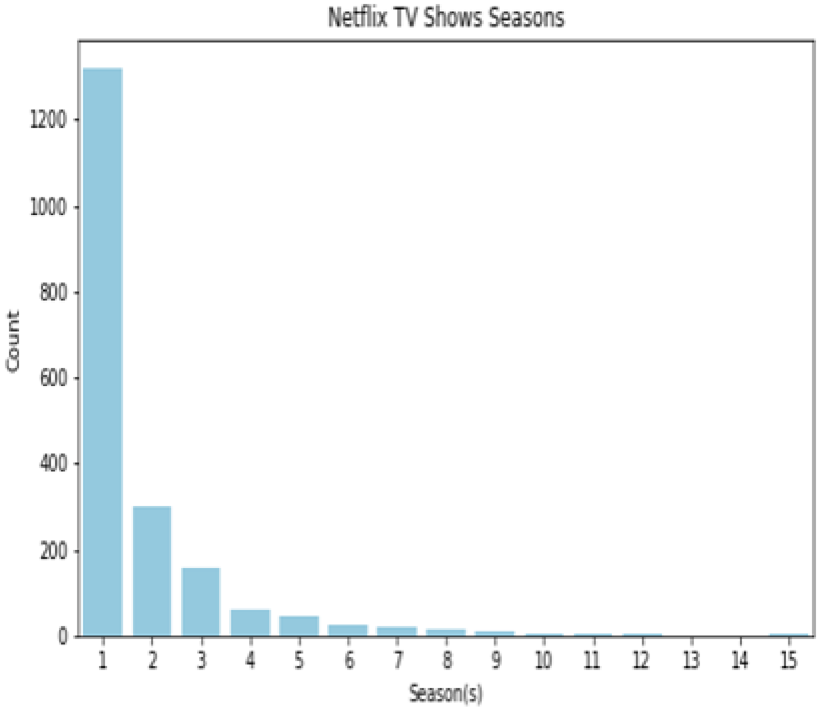
Continue……

- Based on this graph, we can see that the popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been tremendous

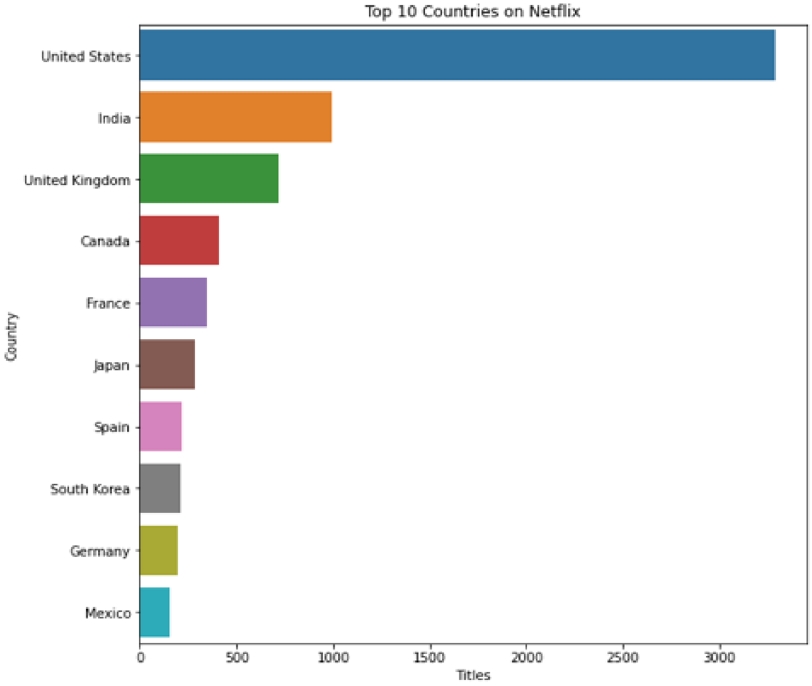


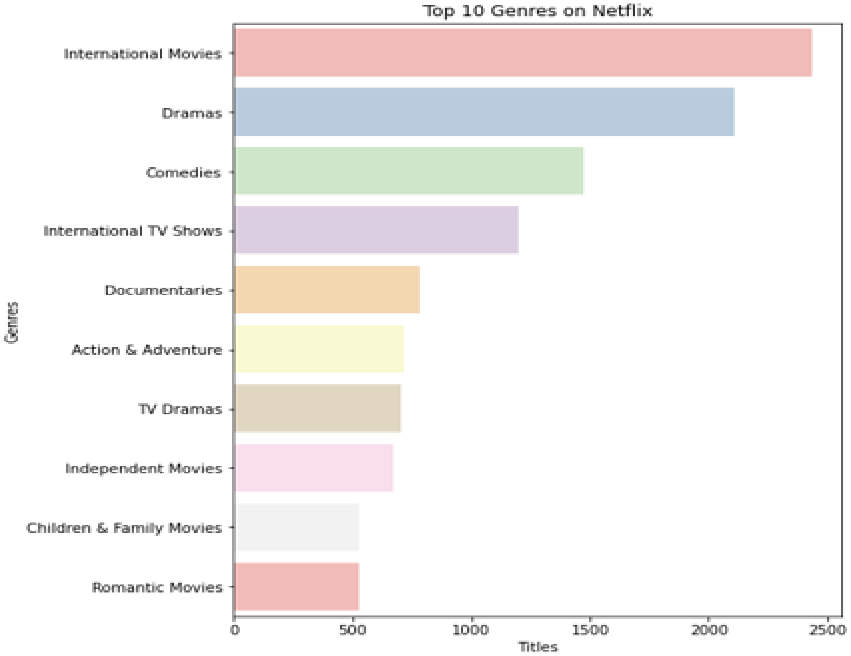
Continue……


Duration Distribution for Netflix Movies
% of All Netflix Movies
0.020
0.015
0.010
0.005
0.000
0
30
60
90
120
150
180
210
240
270
300
330
Duration (minutes)

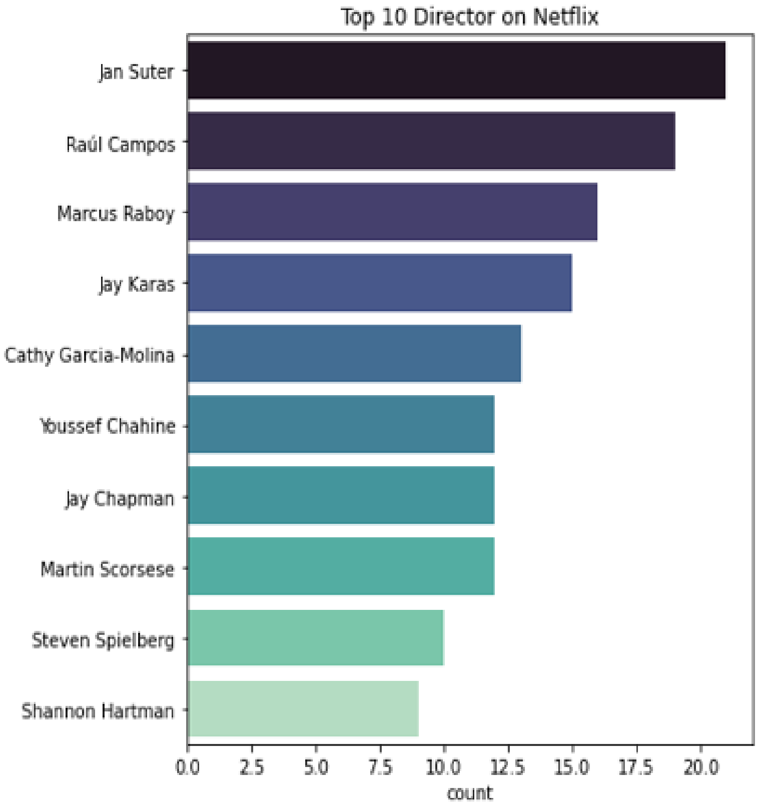

Netflix TV Shows Seasons
Count
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Season(s)

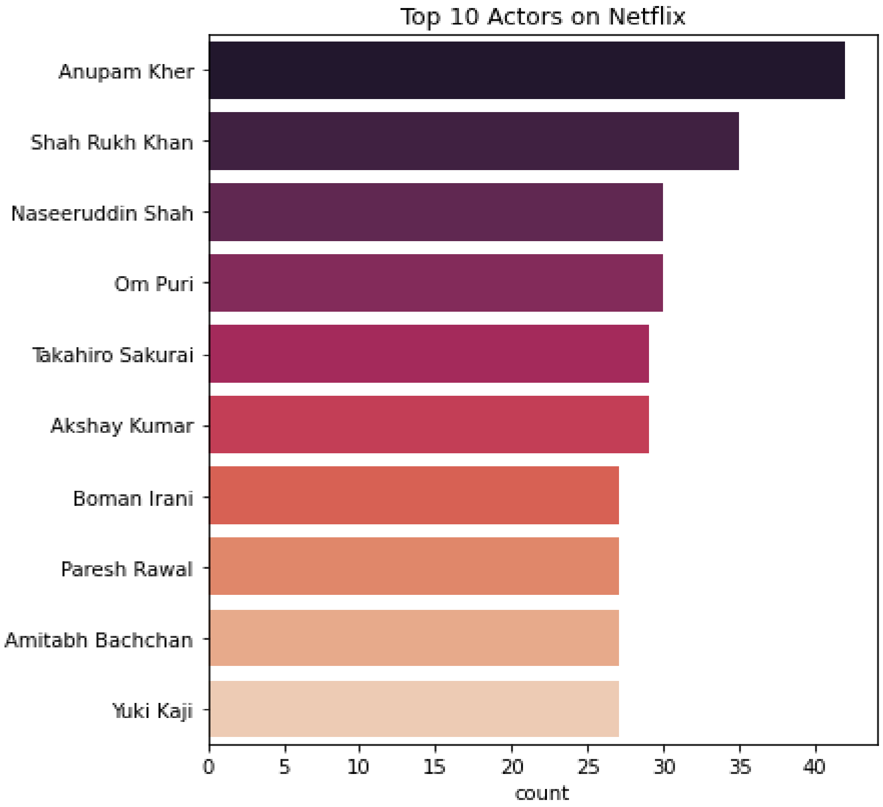
Continue……


Top 10 Countries on Netflix
Country
United States
India
United Kingdom
Canada
France
Japan
Spain
South Korea
Germany
Mexico
0
500
1000
1500
2000
2500
3000
Titles

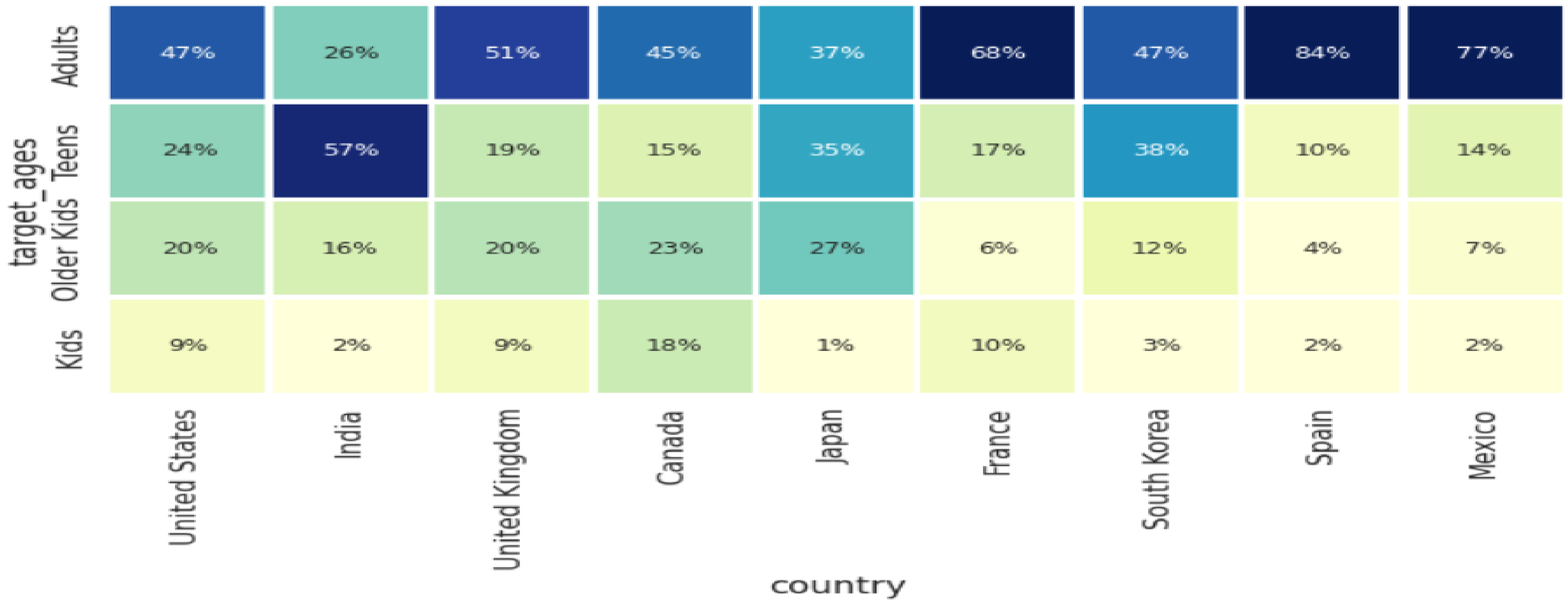

Top 10 Genres on Netflix
Genres
International Movies
Dramas
Comedies
International TV Shows
Documentaries
Action & Adventure
TV Dramas
Independent Movies
Children & Family Movies
Romantic Movies
0
500
1000
1500
2000
2500
Titles

Continue……

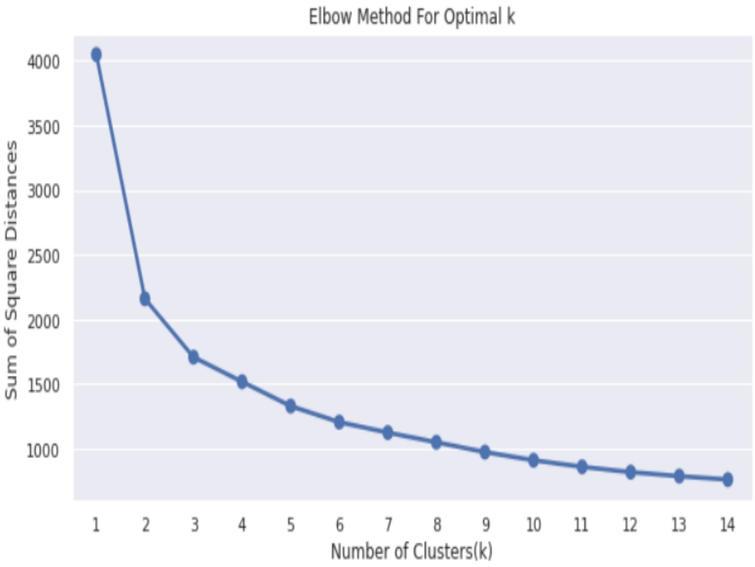

Top 10 Director on Netflix
Jan Suter
Raúl Campos
Marcus Raboy
Jay Karas
Cathy Garcia-Molina
Youssef Chahine
Jay Chapman
Martin Scorsese
Steven Spielberg
Shannon Hartman
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
count

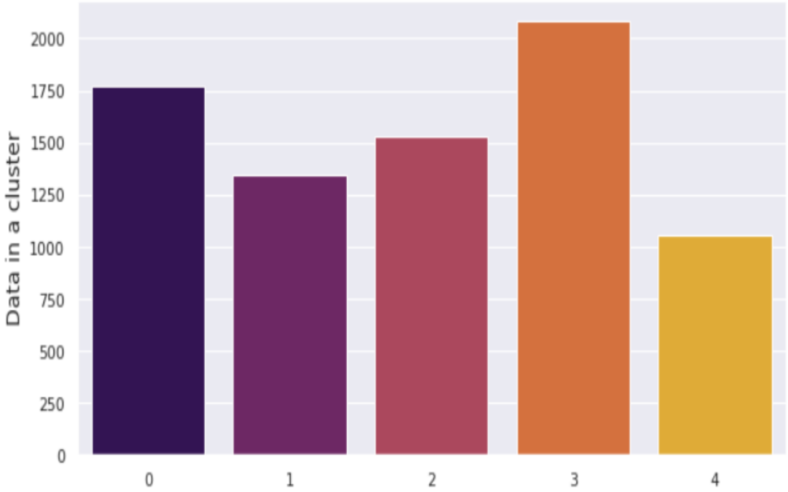

Top 10 Actors on Netflix
Anupam Kher
Shah Rukh Khan
Naseeruddin Shah
Om Puri
Takahiro Sakurai
Akshay Kumar
Boman Irani
Paresh Rawal
Amitabh Bachchan
Yuki Kaji
0
5
10
15
20
25
30
35
40
count

Continue……

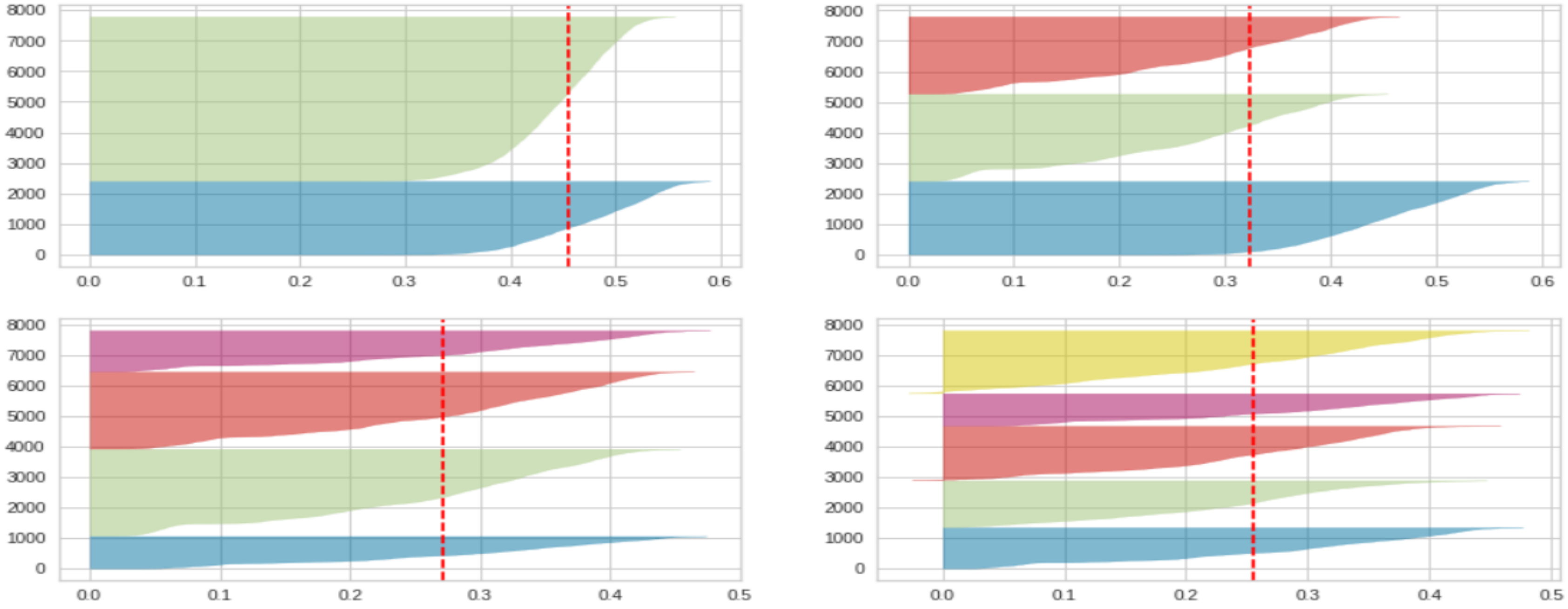

target_ages
Adults
Teens
Older Kids
Kids
47%
26%
51%
45%
37%
68%
47%
84%
77%
24%
57%
19%
15%
35%
17%
38%
10%
14%
20%
16%
20%
23%
27%
6%
12%
4%
7%
9%
2%
9%
18%
1%
10%
3%
2%
2%
United States
India
United Kingdom
Canada
Japan
France
South Korea
Spain
Mexico
country

GRAPHICAL REPRESENTATION OF K MEANS CLUSTERING

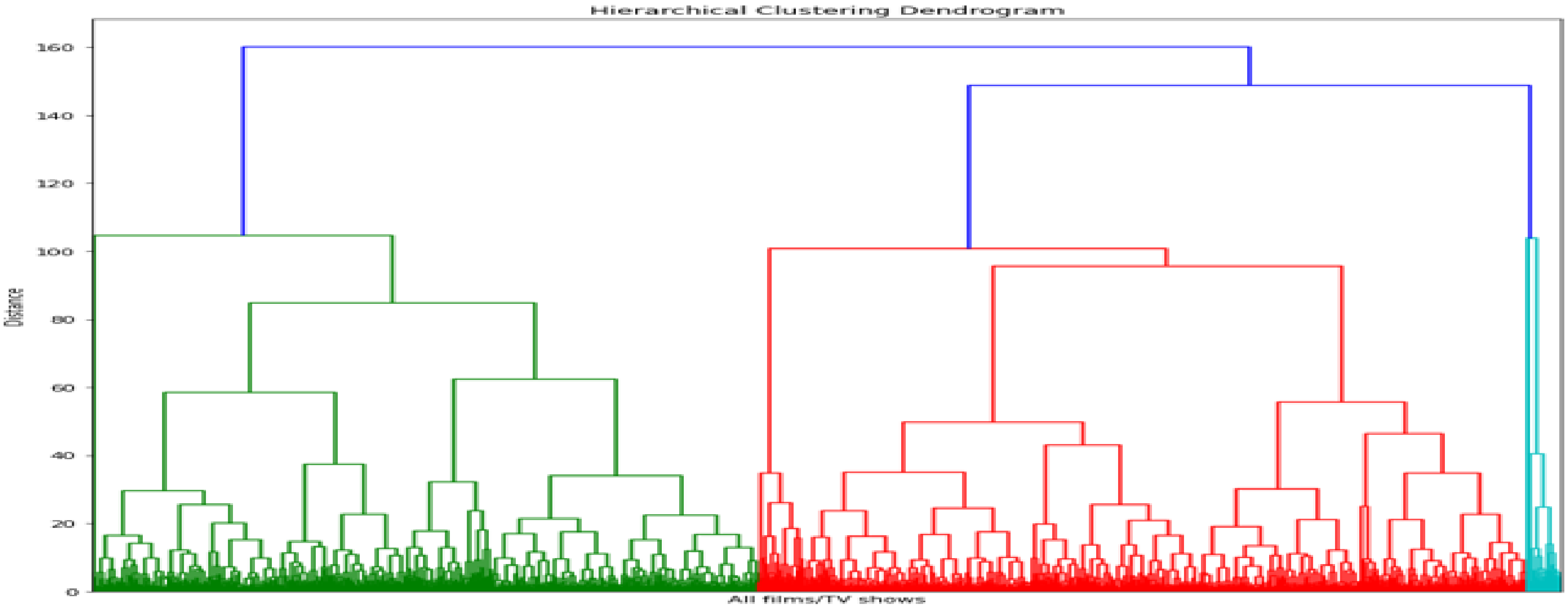




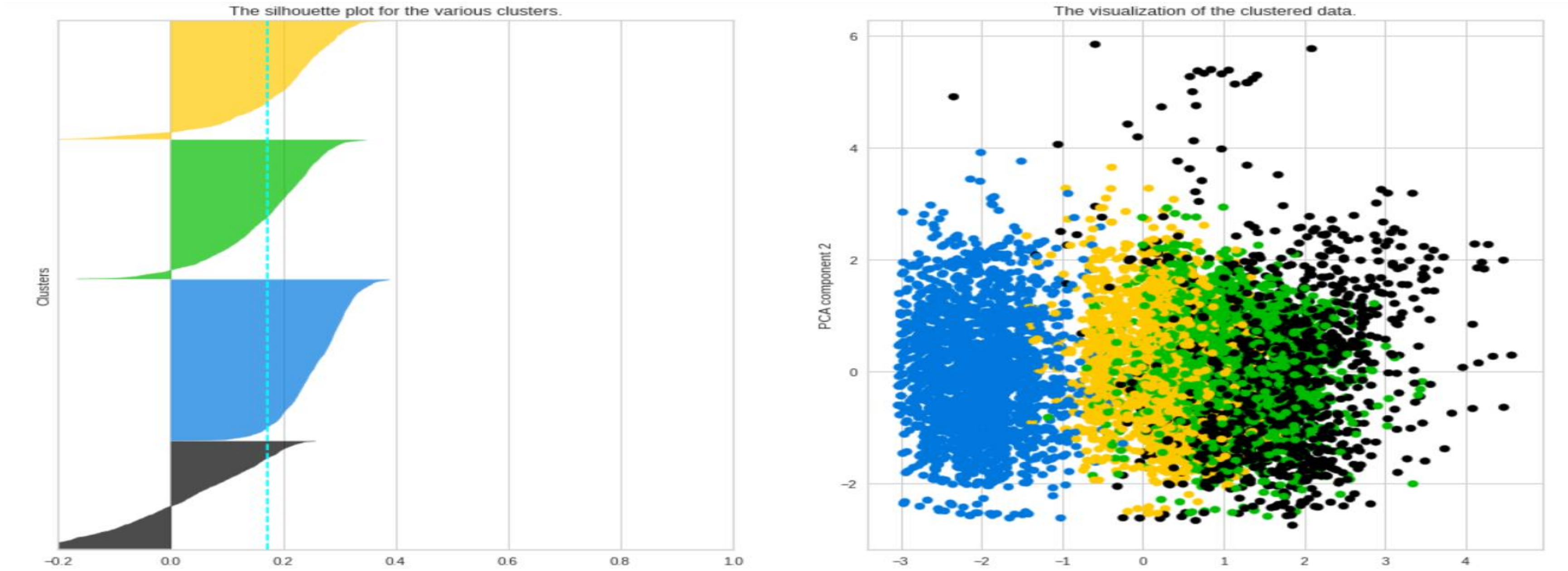
CLUSTER..


8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.1
0.2
0.3
0.4
0.5
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.1
0.2
0.3
0.4
0.5

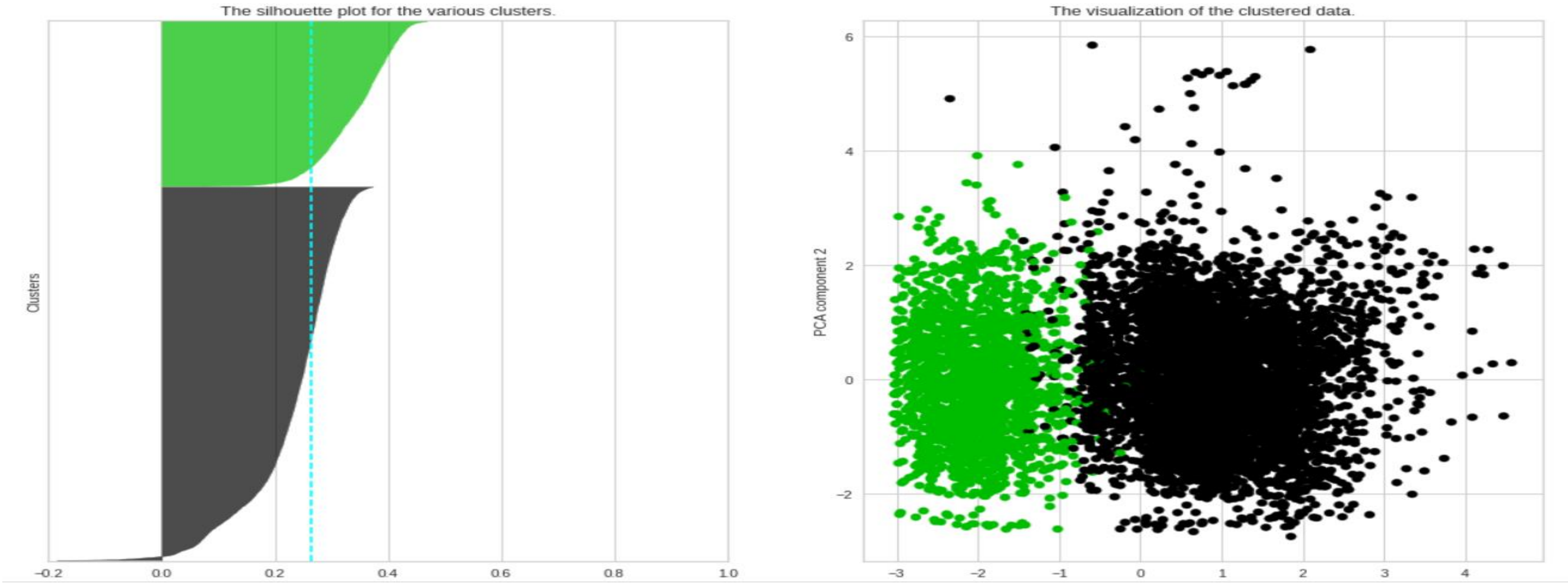
DENDOGRAM


Hierarchical Clustering Dendrogram
160
140
120
100
80
60
40
20
0
Distance
All films/TV shows

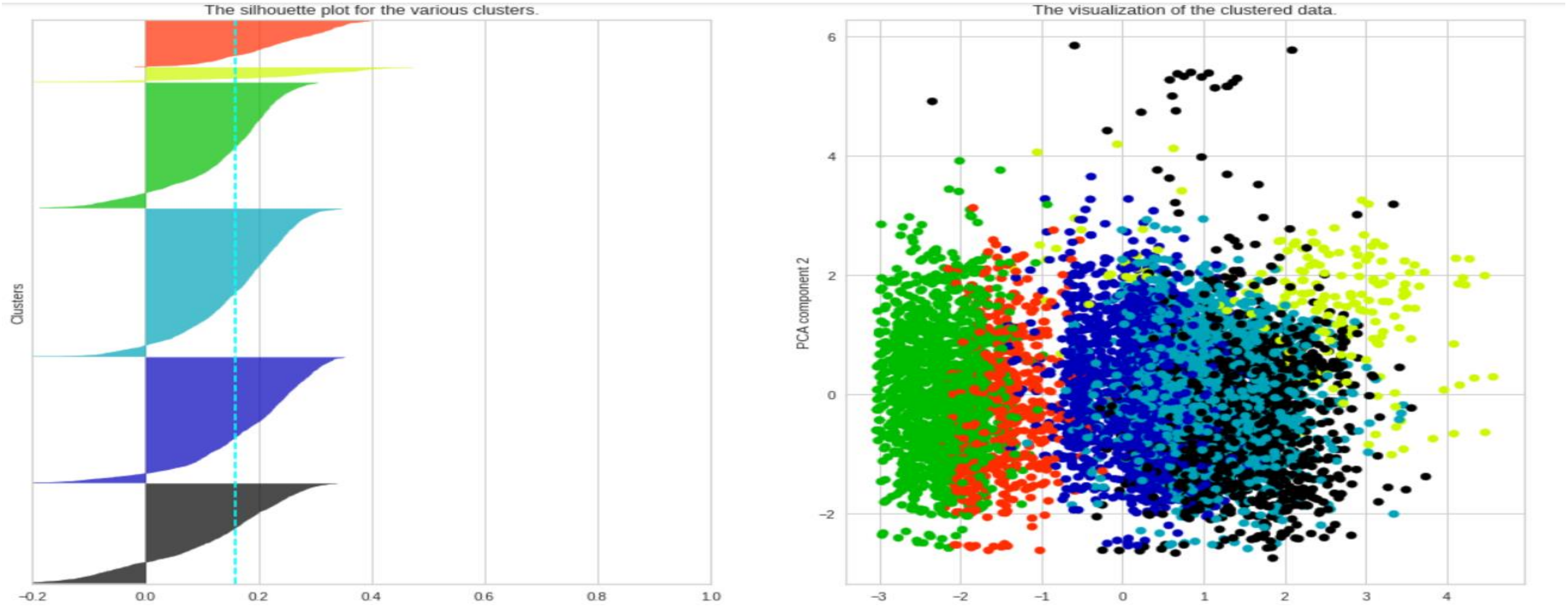
SILHOUETTE ANALYSIS FOR AGGLOMERATIVE CLUSTERING WITH N_CLUSTER = 4



SILHOUETTE ANALYSIS FOR AGGLOMERATIVE CLUSTERING WITH N_CLUSTER = 2



SILHOUETTE ANALYSIS FOR AGGLOMERATIVE CLUSTERING WITH N_CLUSTER = 6



INTERPRETATION BASED ON MODELS

1. **Here is the Silhouette analysis done on the above plots to select an optimal value for n_clusters.**
2. **The value of 4 and 5 for n_clusters looks to be the optimal one. The silhouette score for each cluster is above average silhouette scores.**

CONCLUSION

1) Director contains large number of null values so we will drop these columns

2) In these dataset there are two types of contents where 30.86% includes tv shows remaining 69.14% carries movies.

3) We have reached a conclusion from our analysis from the content added over years that Netflix is focusing movies and TV shows (Fom 2016 data we get to know that Movies is increased by 80% and TV shows is increased by 73% compare)

4) From the dataset insights we can conclude that the most number of TV Shows released in 2017 and for Movies it is 2020

5) On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows.

6) Most of the movies are belonging to 3 categories

7) TOP 3 content categories are International movies , dramas , comedies.

CONCLUSION

8) Applied different clustering models like Kmeans,Hierarchial,Agglomerative clustering on data we got the best arrangement
9) By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 2.