AlmaBetter

Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By – Lavanya Shinde

WHAT IS NETFLIX?

Netflix is an American subscription-based streaming service that allows us to watch TV shows and movies without commercials on an internet-connected device. We can also download TV shows and movies to our iOS, Android, or Windows 10 device and watch without an internet connection. Netflix Founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. As of December 2022, Netflix had more than 200 million worldwide subscribers .

CONTENTS

01 INTRODUCTION

1. From 2006 Netflix start the analyzing user data to predict how much a viewer would like a movie, based on previous user Preferences.
2. Whenever we access the Netflix service, Netflix recommendations system strives to help us to find a show or movie to enjoy with minimal effort.
3. All of these done by using user data as an inputs that we process in our algorithms. (An algorithm is a process or set of rules followed in a problem-solving operation).
4. As well as we use clustering in this project. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.

02

PROBLEM STATEMENT

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.

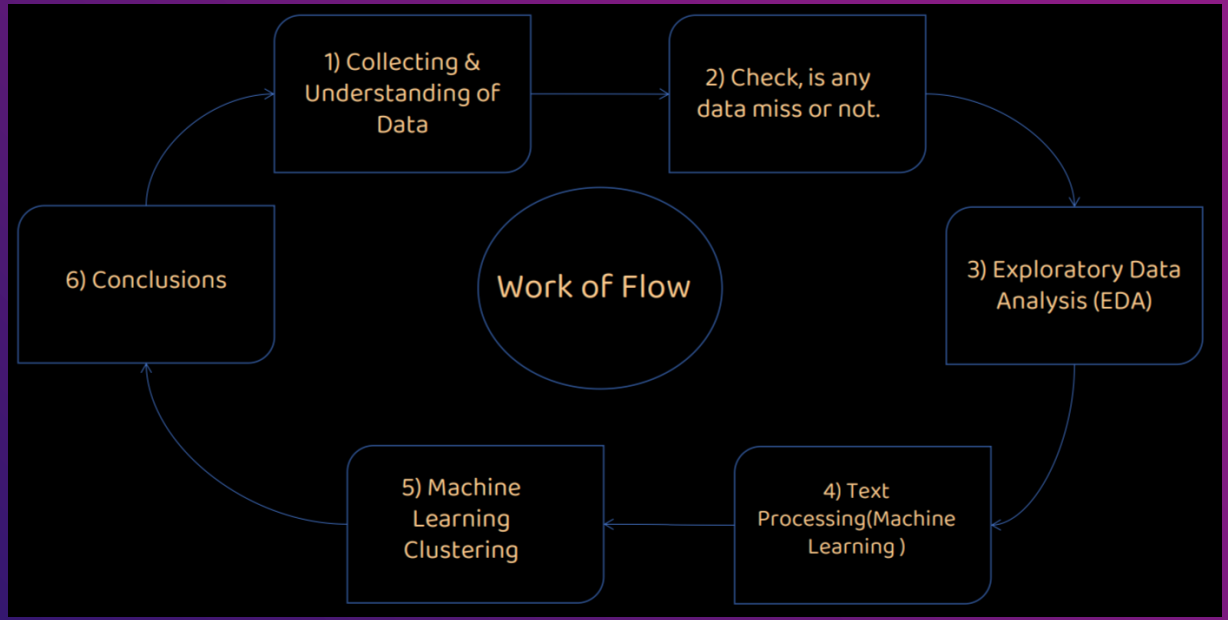
In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

While using this Netflix Data we Try to Understand :-

i) Understanding what type content is available in different countries.
ii) Is Netflix has increasingly focused on TV rather than movies in recent years.
iii) Clustering similar content by matching text-based features.

03 WORKFLOW

- Here is the workflow used for Project :-



04 DATA REVIEW

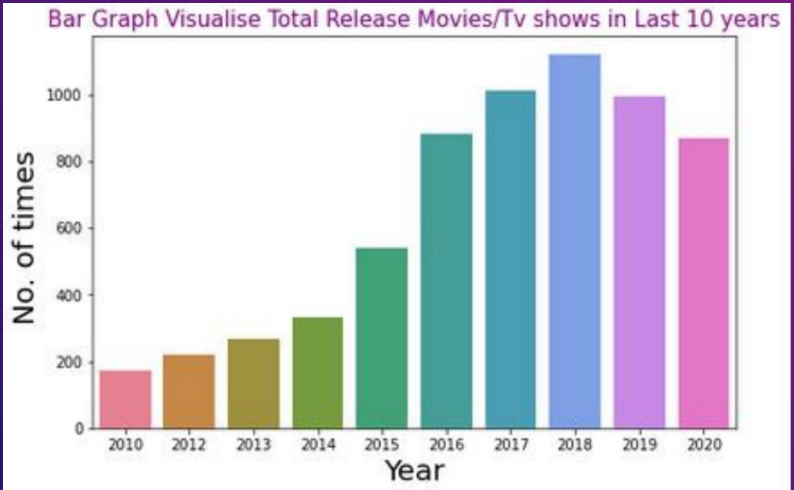
1) The shape of dataset is (7787 x 12)
2) Total Number of Rows are : 7787
3) Total Number of Columns are : 12
4) Missing Value in Columnswise :-
    i) director - 2389
    ii) cast - 718
    iii) country - 507
    iv) date_added - 10
    v) rating - 7
5) The Total number of duplicate values in the data set is = 0

Feature	Type	Samples
show_id	Continuous	s1,s2,s3...
title	Text	[3%, Ozark,...]
type	Categorical	Movie/ TV Show
rating	Categorical	TV-MA, TV-R, R, PG-13....
director	Text	Raúl Campos, Jan Suter
cast	Text	David Attenborough
country	Categorical	United States
date added	Categorical	August 14, 2020
release year	Numerical	1999,2000,2001..
duration	Categorical	1 season, 2 seasons… / 90 mins, 120 mins...
listed_in	Text	[International Movies, Drama..]
description	Text	In a future where the elite inhabit a..

05

EXPLORATORY DATA ANALYSIS

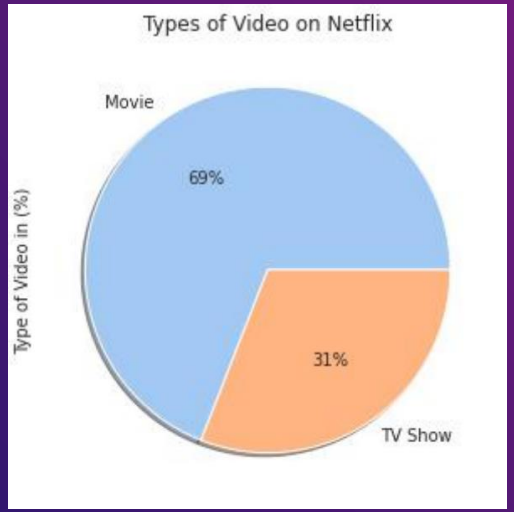
1) Visualize Total Release Movies/Tv shows in Last 10 years



Key Insights:

1) 2018 is a year were Maximum number of movies have been released which is Total 1121.
2) 2017 is a Second Highest year For Maximum number of movies have been released which is Total 1012.
3) 2019 is a Third Highest year where 996 number of movies have been released.

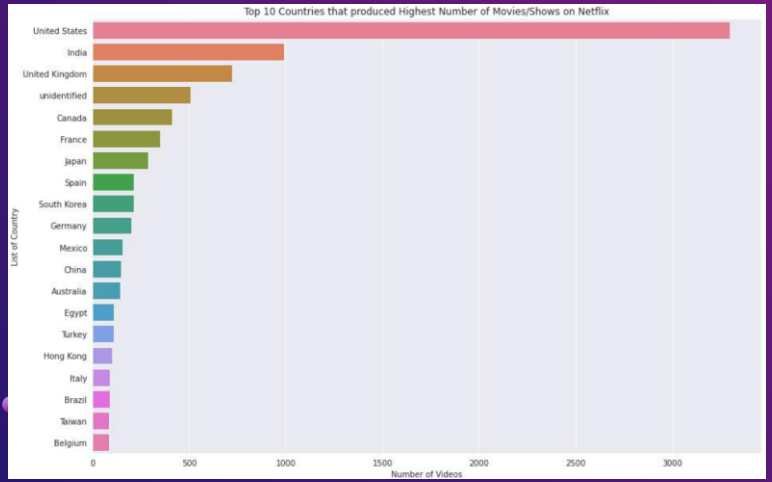
2) Visualize Types of Video on Netflix



Key Insights:

1. 69% are Movie type of Contents on Netflix which is Total 5377 Number of Movies.
2. 31% are TV Show type of Contents on Netflix which is Total 2410 Number of TV Shows.

3) Visualize the top 10 Countries that produced Highest Number of Movies/Shows on Netflix



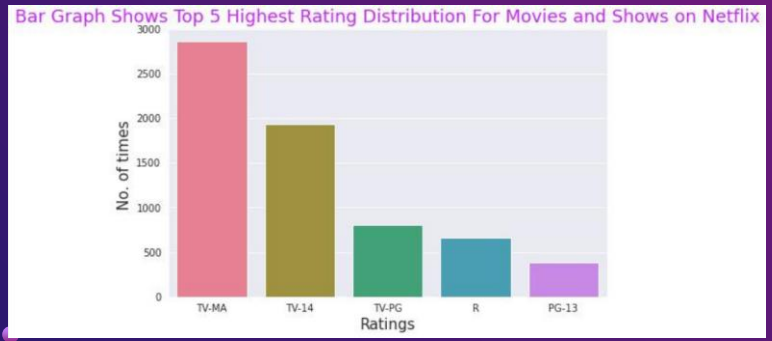
Key Insights:

1) United State is a Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by US is 3296.

2) India is a Second Highest Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by India is 990.

3) United Kingdom is a Third Highest Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by UK is 722.

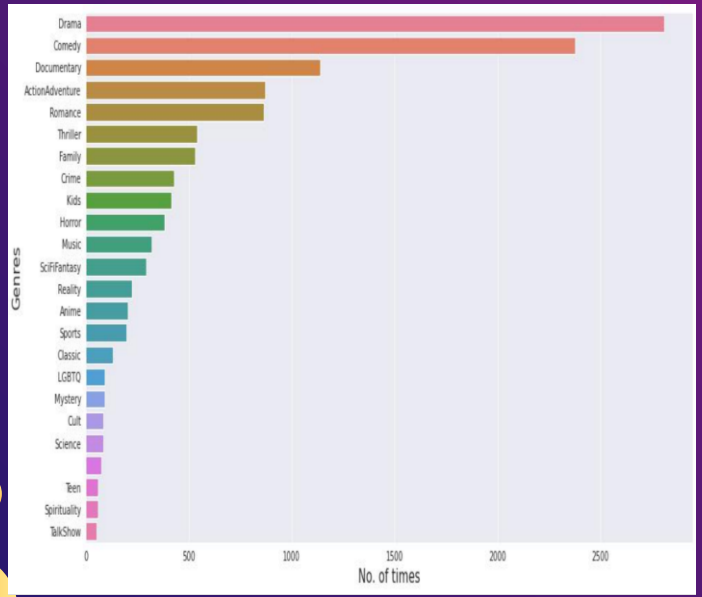
4) Visualize Top 5 Rating Distribution for Movies and Shows on Netflix



Key Insights:

1) Tv-MA is a Highest Rating Distribution For Movies and Shows on Netflix, which is Total 2863.
2) Tv-14 is a Second Highest Rating Distribution For Movies and Shows on Netflix, which is Total 1931.
3) Tv-PG is a Third Highest Rating Distribution For Movies and Shows on Netflix, which is Total 806.

5) Visualize the top Genres For Movies/TV-Shows on Netflix



Key Insights:

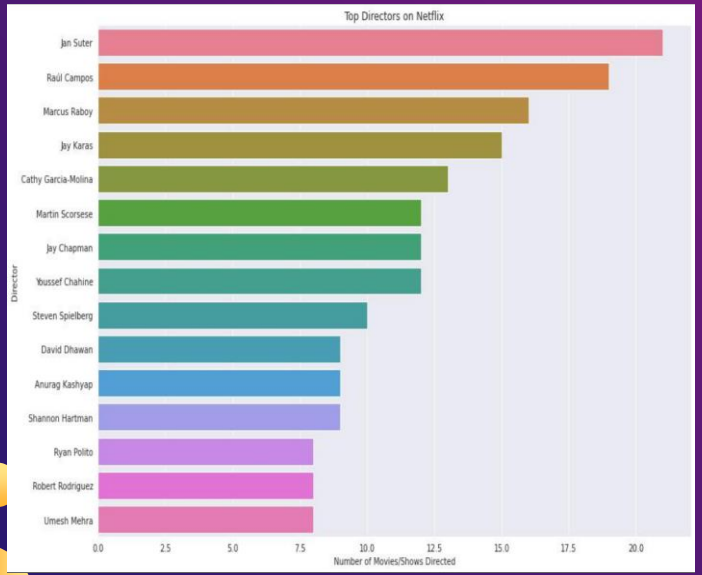
1) Drama is a Top Genres For Movies/TV Shows on Netflix. That is, the content of the movie in the drama genre has been produced the most which is 2810 Times.

2) Comedy is a Second Highest Genres For Movies/TV Shows on Netflix which is 2377 Times.

3) Documentary is a Third Highest Genres For Movies/TV Shows on Netflix which is 1139 Times.

4) Action Adventure are in 4th Position & Romance are in 5th Position.

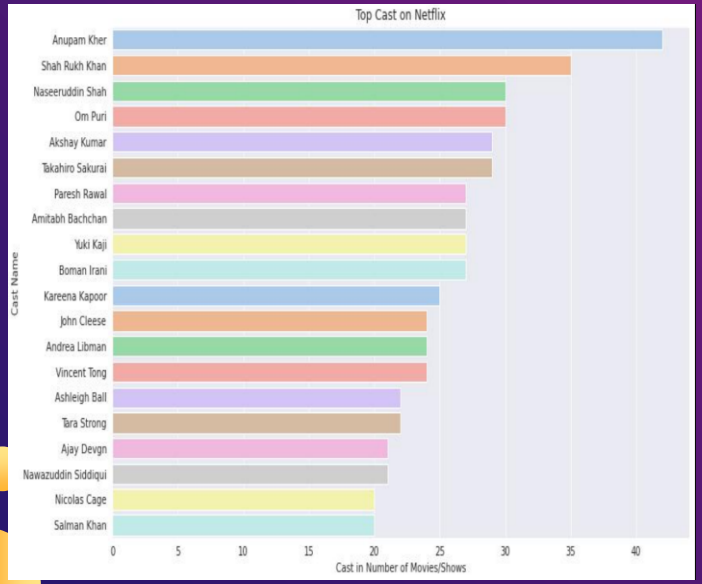
6) Visualize the Top Directors on Netflix



Key Insights:

1) Jan Suter is the top director in the Netflix industry & he has Directed 21 Movies/Shows.
2) Raul Compose is the Second top director in the Netflix industry & he has Directed 19 Movies/Shows.
3) Marcus Raboy is the Third top director in the Netflix industry & he has Directed 16 Movies/Shows.
4) Jay Karas is the 4th position & he has Directed 15 Movies/Shows and Cathy Garcia-Molina is the 5th Position & he has Directed 13 Movies/Shows

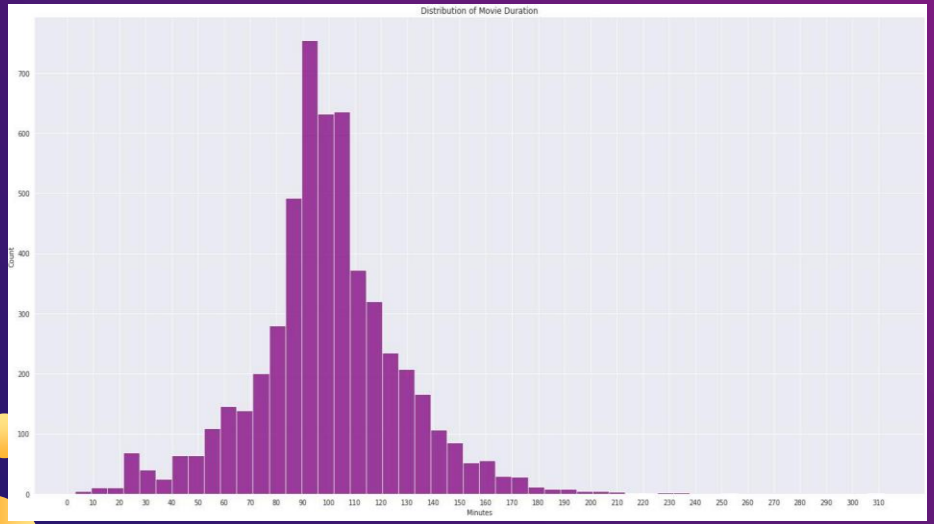
7) Visualize the Top cast on Netflix Till Year of 2020



Key Insights:

1) Anupam Kher is the Top Cast on Netflix As Our Visualization.
2) Shah Rukh Khan is the second highest cast on Netflix.
3) Naseeruddin Shah is the third highest cast on Netflix.
4) Om Puri is the 4th Highest cast on Netflix
5) Akshay Kumar is the 5th Highest cast on Netflix

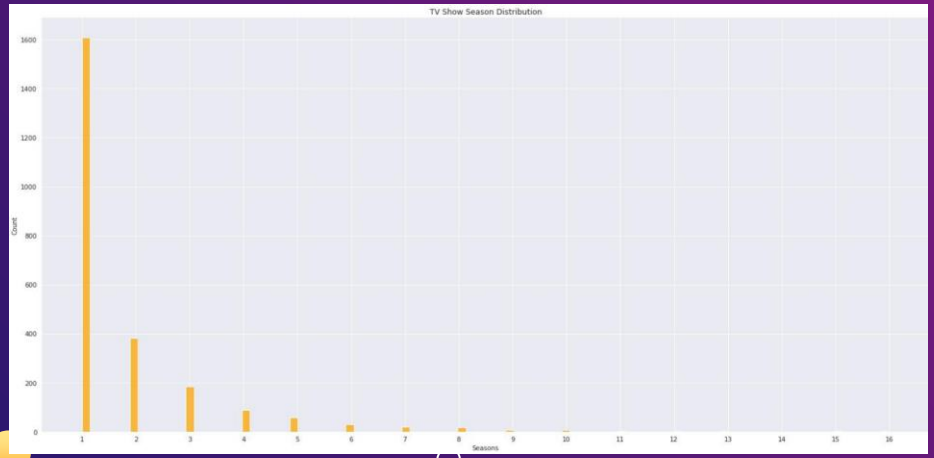
8) Top Duration of Movies on Netflix



Key Insights:

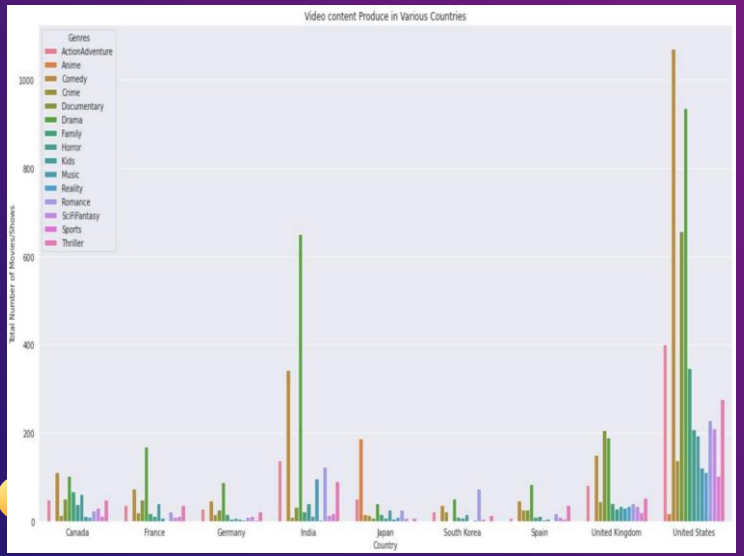
Most of the movies on Netflix have a duration range from 85 to 115 minutes.

9) Highest Duration of TV Shows on Netflix



Key Insights:
Most TV shows on Netflix have a length of 1 season only

10) What type content is available in different countries



Key Insights:

1) Drama is the most produced genre in a lot of Non-English-speaking countries.

2) Comedy is the most produced genre in English speaking countries like United States of America and United Kingdom and Canada.

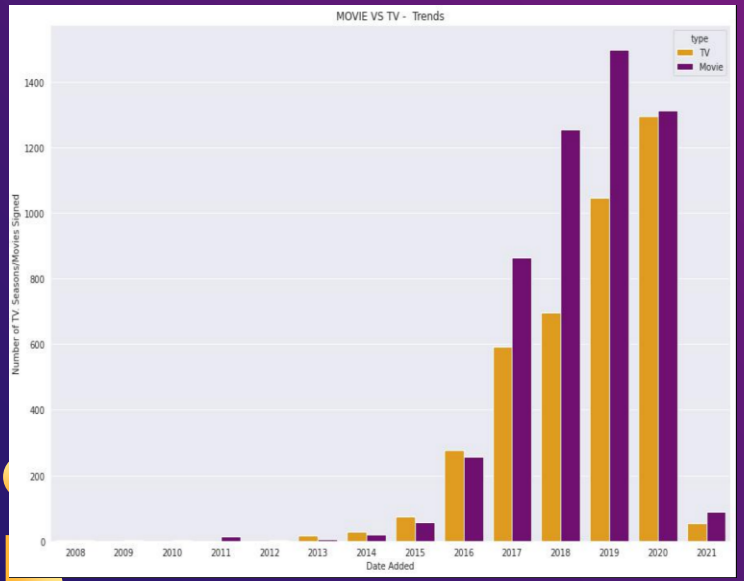
3) Drama and Comedy are the most produced genres in the top countries with exceptions of Japan and South Korea.

4) Japan is the biggest producer of Anime. Anime is also the most produced in genre in Japan.

5) Most South Korean content are from the Romance genre.

6) Documentaries are mainly produced in United Kingdom and United States of America

11) Is Netflix has increasingly focusing on TV rather than movies in recent years?



Key Insights:

1) The above graph depicts seasons of TV shows signed vs the movies signed.

2) This distinction gives contacts as TV shows require recurring investment for each seasons. So, the TV numbers have been increased in accordance with the seasons. As they were considered as one entity earlier.

3) We can observe that TV shows signed have been higher than movies in 2016. While the movies signed have been higher, it is blatantly visible that the TV shows signed per year is catching up to the movies signed by the year

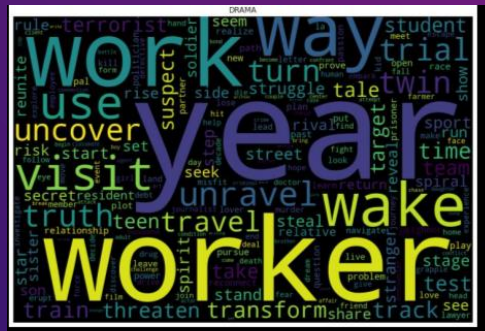
06

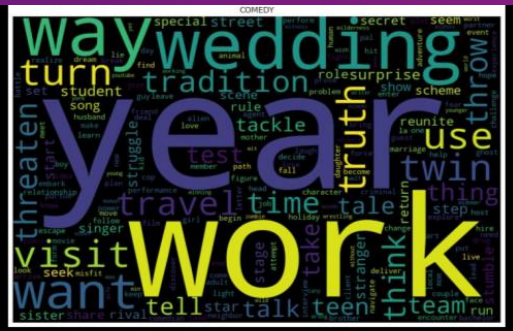
TEXT PROCESSING (MACHINE LEARNING)

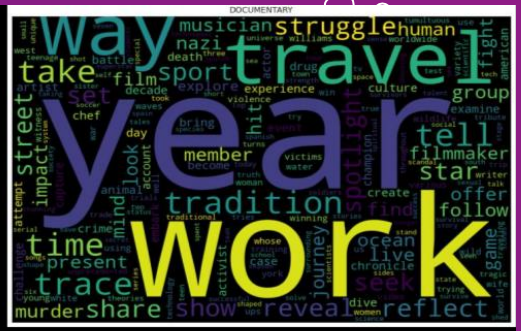
What is text Processing ?
:- It is the process of obtaining valuable insights from texts. This method also use for detect different patterns in data and break the text into clusters. We convert all text to lower case, remove punctuations as well as remove irrelevant words. Here similar
words are unified to save memory and processing time as well. Individual words and group of words are also collected to extract context related details.

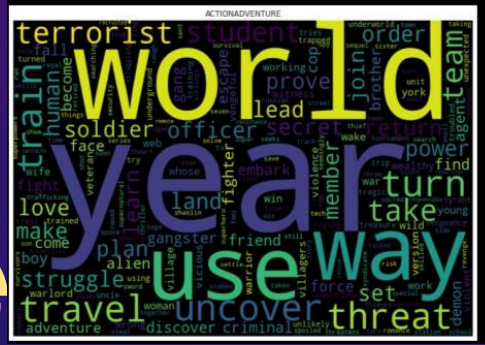
The steps involved in text preprocessing are :
● Tokenization :- Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or
developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.
● Punctuation Removal :- All the punctuations from the text are removed.
● Stopword Removal :- Stop word removal is one of the most used preprocessing steps across different NLP applications. The
idea is simply removing the words that occur commonly across all the documents in the corpus.
● Stemming Words :- Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of
words known as a lemma.

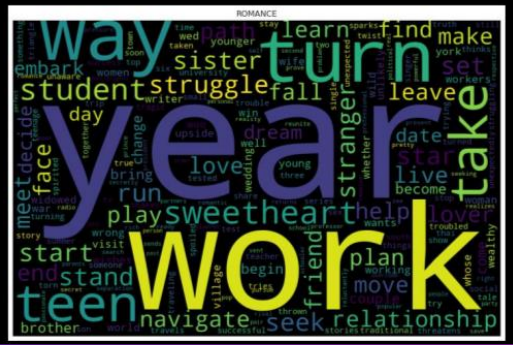
Visualization of Word Clouds For Top 6 Genres (Drama, Comedy, Documentary, Action-Adventure, Romance, Thriller)

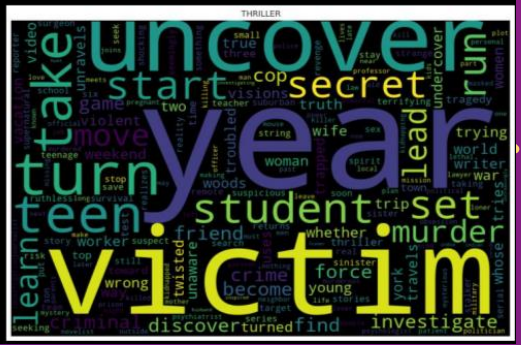




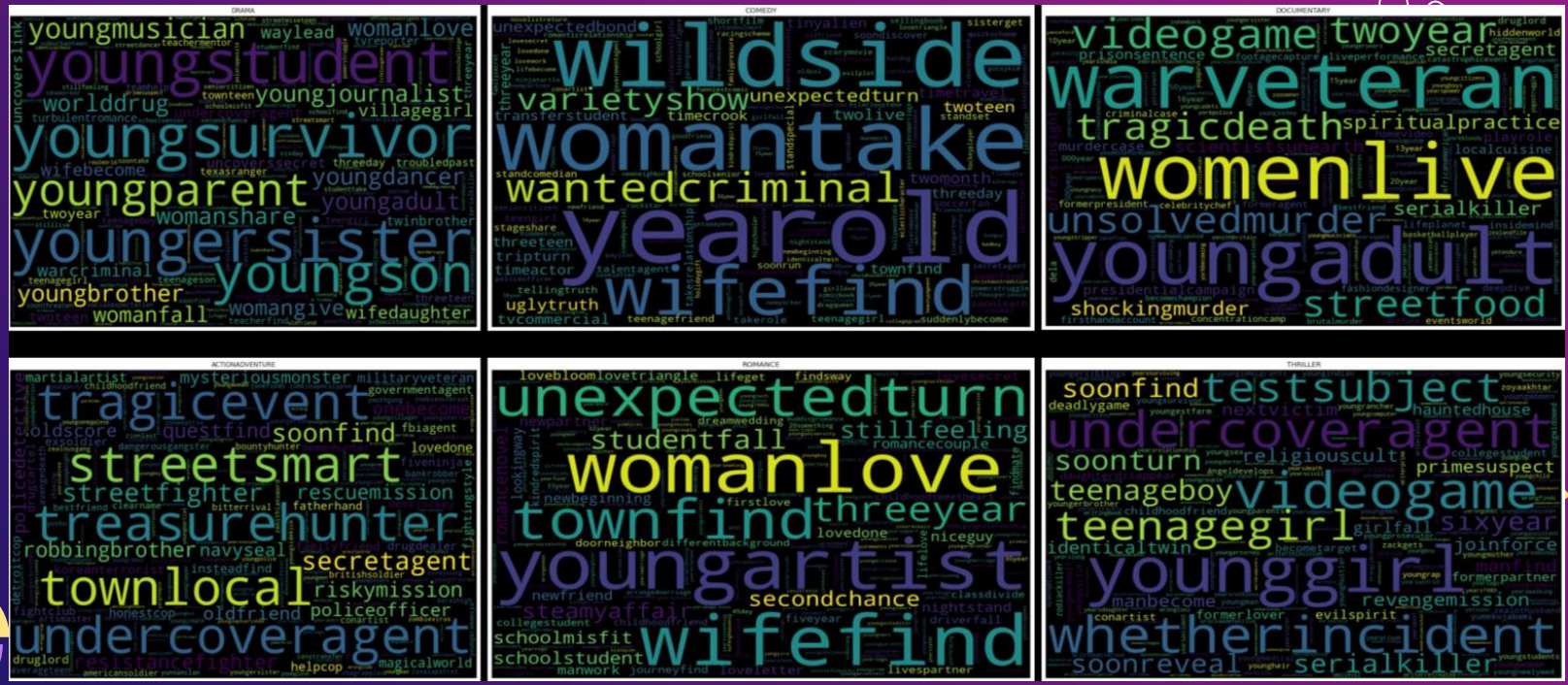




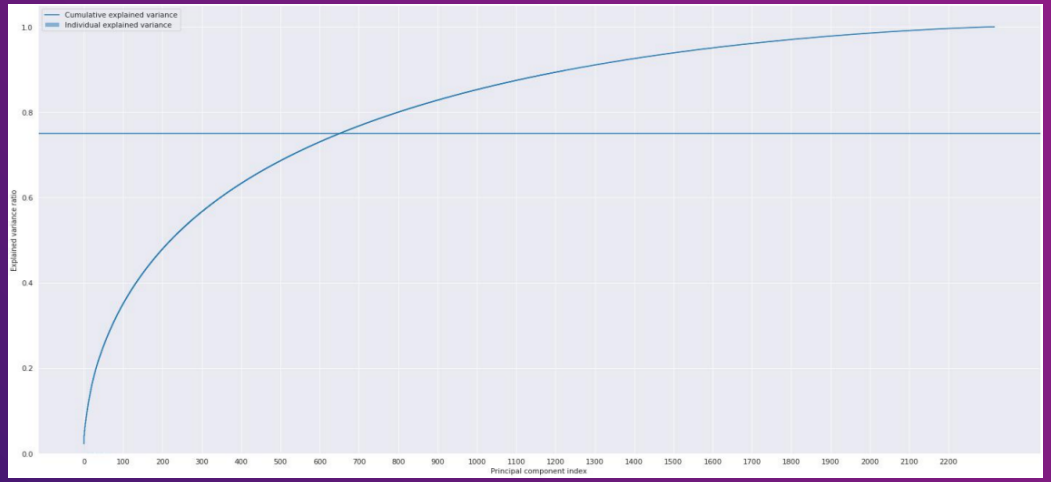




Visualization of Bigram Word Clouds For Top 6 Genres (Drama, Comedy, Documentary, Action-Adventure, Romance, Thriller)



Vectorizing Texts (Topic Modeling)



Vectorizing the preprocessed attributes Movie Deet's and Description has sum total of 2126 + 192 = 2318 dimensions. These dimensions will have to be reduced using PCA, which would result in loss of information. Alternatively, the two attributes can be used to model the content into topics using Latent Dirichlet Allocation. This would make sure that all the topical information about video content are captured without putting any available information to waste

Visualization of WORD Clouds For Topics (Drama, Comedy, Documentary, Action-Adventure, Romance, Thriller)

07

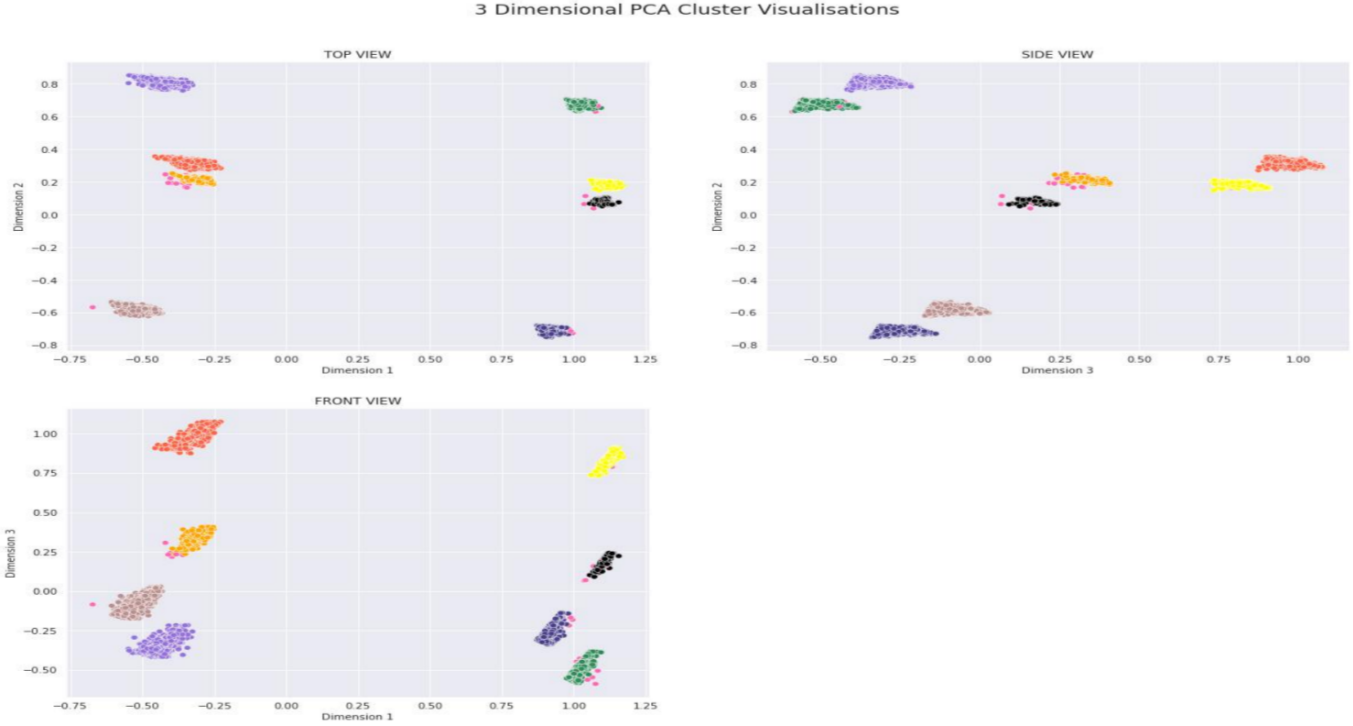
MACHINE LEARNING CLUSTERING

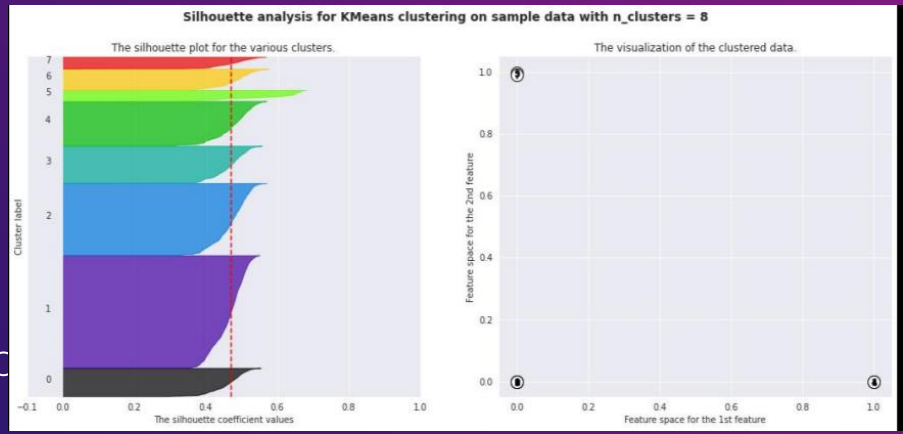
DBSCAN Clustering Algorithm

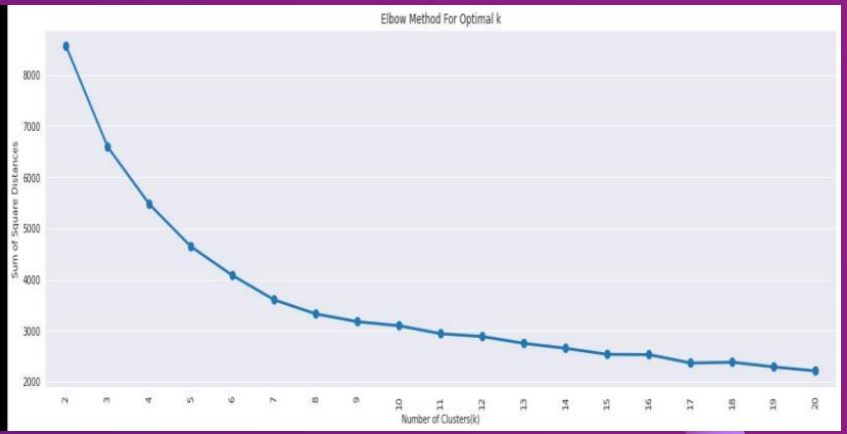
- DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. The DBSCAN algorithm uses two parameters:

- minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.
- * eps (ε): A distance measure that will be used to locate the points in the neighborhood of any point. After Performing DBSCAN clustered the data into 10 clusters with a silhouette score is 0.43875

- DBSCAN Clustering Algorithm Visualization PCA







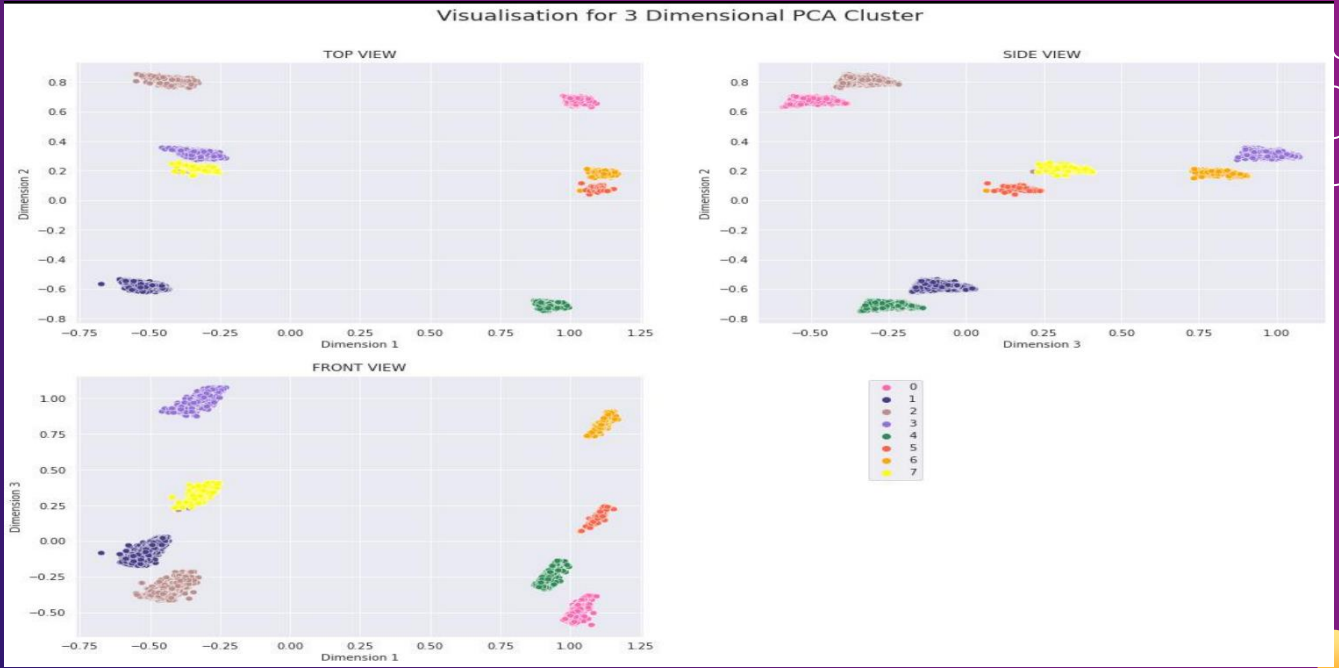
Silhouette Plot

From Silhouette Analysis and Elbow method, the optimal cluster is 8. This gives a clustering score of 0.354

Elbow Curve

From the above graph has plots of the sum squared inertia for K clusters trained on K means algorithm. In this graph, we decide the optimal number of clusters by locating the elbow of the graph which is at 8 clusters.

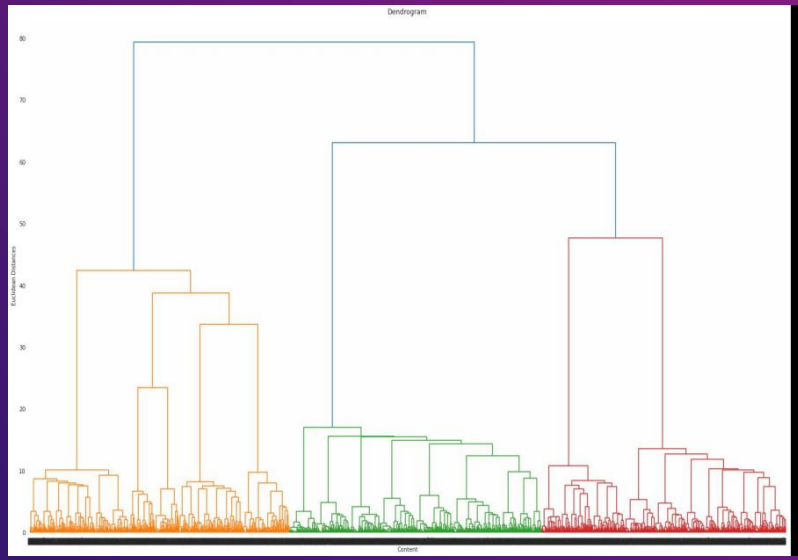
- **K-means Clustering Algorithm Using PCA**

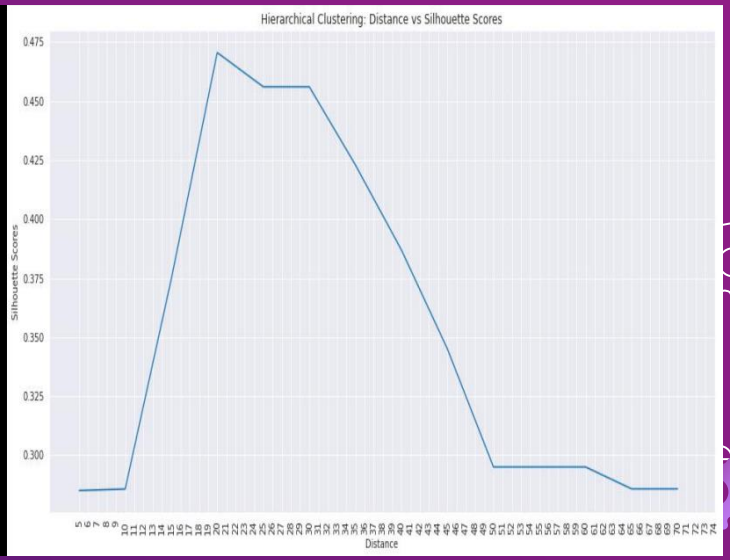


- **Hierarchical Clustering Algorithm**

Hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters. It is most popular and widely used method to analyze social network data. In this method, nodes are compared with one another based on their similarity. Larger groups are built by joining groups of nodes based on their similarity. The highly similar or close clusters are merged and the proximity matrix for each cluster is recalculated.

After Performing Hierarchical Clustering, the dendrogram distance was optimal at a distance of 20 with eight clusters producing a silhouette score of 0.4705, Davies-Bouldin Index of 0.8839 and Calinski-Harbaz Score of 2930.84



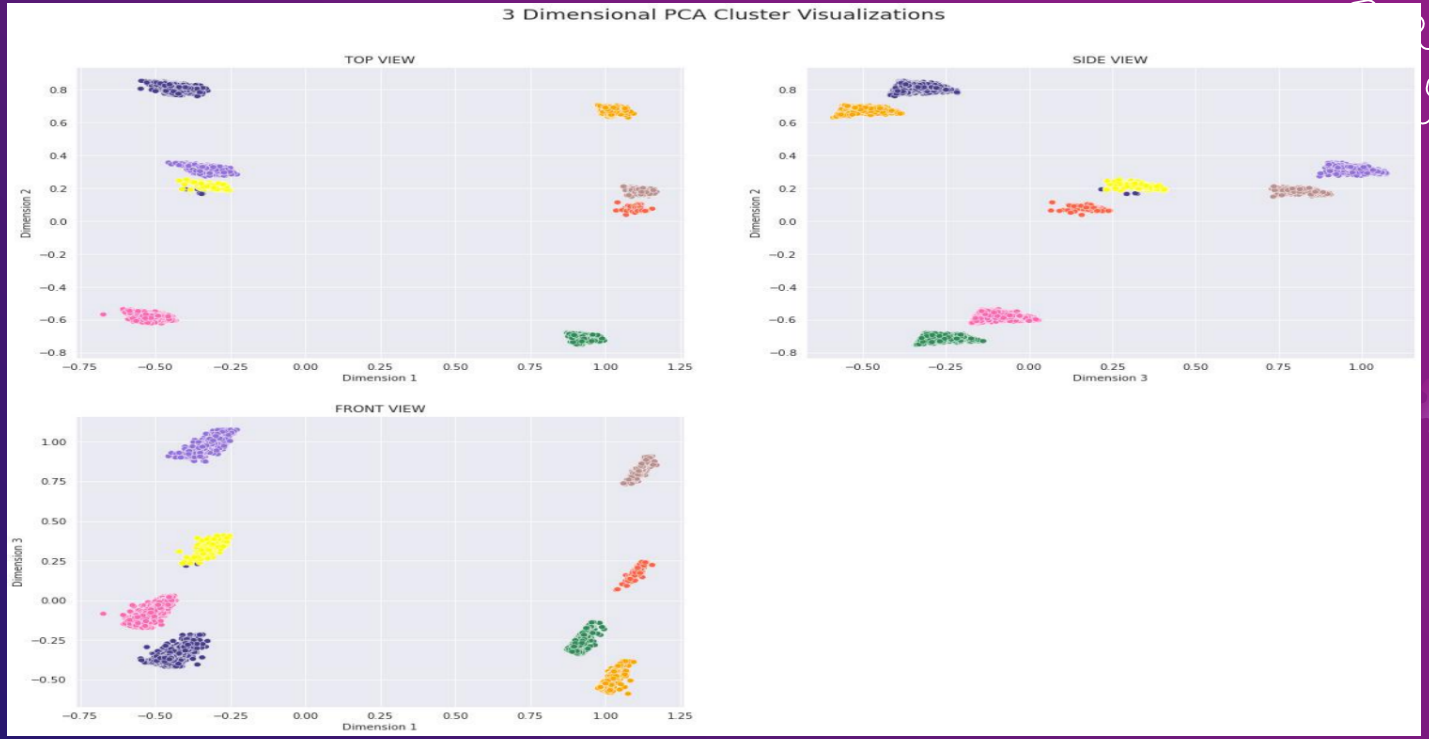


Dendrogram Plot

From Our Visualization, a Distance Of 20-30 appears to have well defined trees.

Silhouette Score vs Distance Plot

- 3 Dimensional PCA Hierarchical Cluster Visualizations



- After Optimizing and Selecting the Best Number of Clusters for the Three Models, the Final Report is Displayed Down Below.

	Algorithm	Parameters	Clusters	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harbaz Score
0	DBSCAN	Default	9	0.468302	1.617649	2538.137264
1	KMeans	Default	8	0.470426	0.884662	2932.283775
2	Hierarchical Agglomerative	Distance = 20	8	0.470549	0.883982	2930.849722

08 CONCLUSION

1) After Visualize Total Release Movies/Tv shows in Last 10 years, we find out :-

i) 2018 is a year where Maximum number of movies have been released which is Total 1121.

ii) 2017 is a Second Highest year For Maximum number of movies have been released which is Total 1012.

iii) 2019 is a Third Highest year where 996 number of movies have been released.

2) 69% are Movie type of Contents on Netflix which is Total 5377 Number of Movies. 31% are TV Show type of Contents on Netflix which is Total 2410 Number of TV Shows.

3) i) United State is a Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by US is 3296.

ii) India is a Second Highest Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by India is 990.

iii) United Kingdom is a Third Highest Country which is produced Highest Number of Movies/Shows on Netflix. Total Number of Movies/Show produced by UK is 722.

4) i) Tv-MA is a Highest Rating Distribution For Movies and Shows on Netflix, which is Total 2863.

ii) Tv-14 is a Second Highest Rating Distribution For Movies and Shows on Netflix, which is Total 1931.

iii) Tv-PG is a Third Highest Rating Distribution For Movies and Shows on Netflix, which is Total 806

5) i) Drama is a Top Genres For Movies/TV Shows on Netflix. That is, the content of the movie in the drama genre has been produced the most which is 2810 Times.
ii) Comedy is a Second Highest Genres For Movies/TV Shows on Netflix which is 2377 Times.
iii) Documentary is a Third Highest Genres For Movies/TV Shows on Netflix which is 1139 Times
iv) Action Adventure are in 4th Position & Romance are in 5th Position.

6) i) Jan Suter is the top director in the Netflix industry & he has Directed 21 Movies/Shows.
ii) Raul Compose is the Second top director in the Netflix industry & he has Directed 19 Movies/Shows.
iii) Marcus Raboy is the Third top director in the Netflix industry & he has Directed 16 Movies/Shows.
iv) Jay Karas is the 4th position & he has Directed 15 Movies/Shows and Cathy Garcia-Molina is the 5th Position & he has Directed 13 Movies/Shows.

7) i) Anupam Kher is the Top Cast on Netflix As Our Visualization.
ii) Shah Rukh Khan is the second highest cast on Netflix.
iii) Naseeruddin Shah is the third highest cast on Netflix.
iv) Om Puri is the 4th Highest cast on Netflix.
v) Akshay Kumar is the 5th Highest cast on Netflix.

8) Most of the movies on Netflix have a duration range from 85 to 115 minutes.

9) Most TV shows on Netflix have a length of 1 season only.

10) i) Drama is the most produced genre in a lot of Non-English speaking countries.
ii) Comedy is the most produced genre in English speaking countries like United States of America and United Kingdom and Canada.
iii) Drama and Comedy are the most produced genres in the top countries with exceptions of Japan and South Korea.
iv) Japan is the biggest producer of Anime. Anime is also the most produced in genre in Japan.
v) Most South Korean content are from the Romance genre.
vi) Documentaries are mainly produced in United Kingdom and United States of America.

11) i) The above graph depicts seasons of TV shows signed vs the movies signed.
ii) This distinction gives contacts as TV shows require recurring investment for each seasons. So, the TV numbers have been increased in accordance to the seasons. As they were considered as one entity earlier.
iii) We can observe that TV shows signed have been higher than movies in 2016. While the movies signed have been higher, it is blatantly visible that the TV shows signed per year is catching up to the movies signed by the year

12) After Performing DBSCAN clustered the data into 10 clusters with a silhouette score is 0.43875.

13) After Performing K-means Clustering the elbow and optimal silhouette score were found at 8 clusters with a silhouette score of 0.474, Davies-Bouldin Index of 0.884 and Calinski-Harbaz Score of 2932.28.

14) After Performing Hierarchical Clustering, the dendrogram distance was optimal at a distance of 20 with eight clusters producing a silhouette score of 0.4705, Davies-Bouldin Index of 0.8839 and Calinski-Harbaz Score of 2930.84