Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

(Unsupervised – Clustering)

Team members:

Abhijeet Kulkarni , Kundan Lal

Pankaj Ganjare , Mohd sharik

Introduction to Netflix

Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television shows library through distribution deals as well as its own productions, known as Netflix Originals.

Problem Statement

In 2018, Netflix 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 :

- ***Exploratory Data Analysis***
- ***Understanding what type content is available in different countries***
- ***Is Netflix has increasingly focusing on TV rather than movies in recent years.***
- ***Clustering similar content by matching text-based features***

Overview


Data Introduction
Features Insights
What we get from features
We can Suggest you
Clustering
Challenges & Conclusions

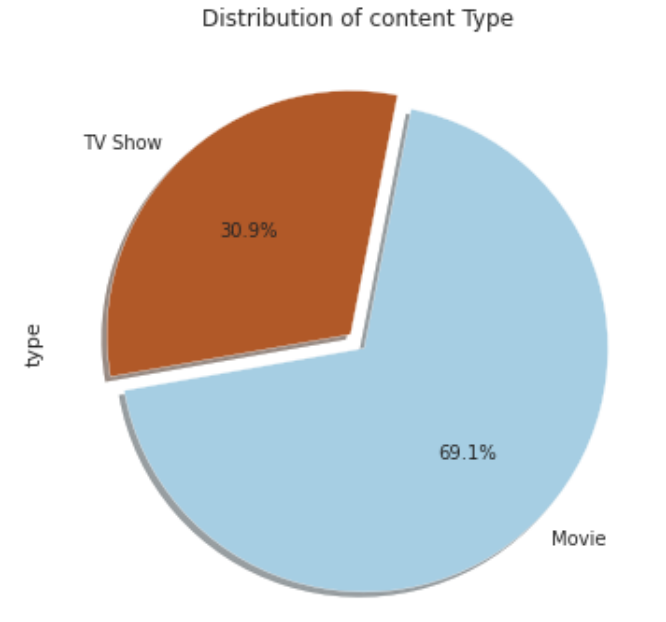
Data Introduction

Column	Description
show_id	Identifier - A Movie or 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 Release year 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

Feature Insights

Features Insights

Movies are more on Netflix and US is a winner



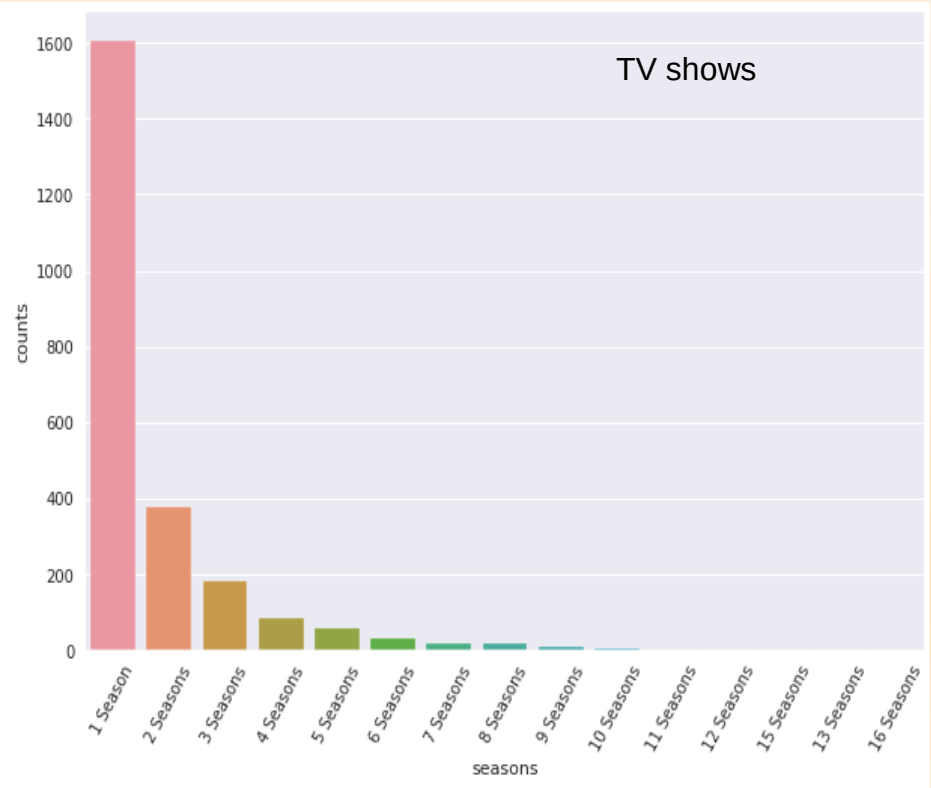
	country	shows counts	country	movies counts
0	United States	860	United States	2427
1	United Kingdom	255	India	915
2	Japan	182	United Kingdom	466
3	South Korea	157	Canada	286
4	Canada	126	France	265
5	France	84	Spain	158
6	India	75	Germany	157
7	Taiwan	70	Japan	103
8	Australia	58	China	102
9	Spain	57	Mexico	101
10	Mexico	53	Egypt	97
11	China	45	Hong Kong	97
12	Germany	42	Australia	84
13	Brazil	29	Turkey	80
14	Colombia	28	Philippines	77

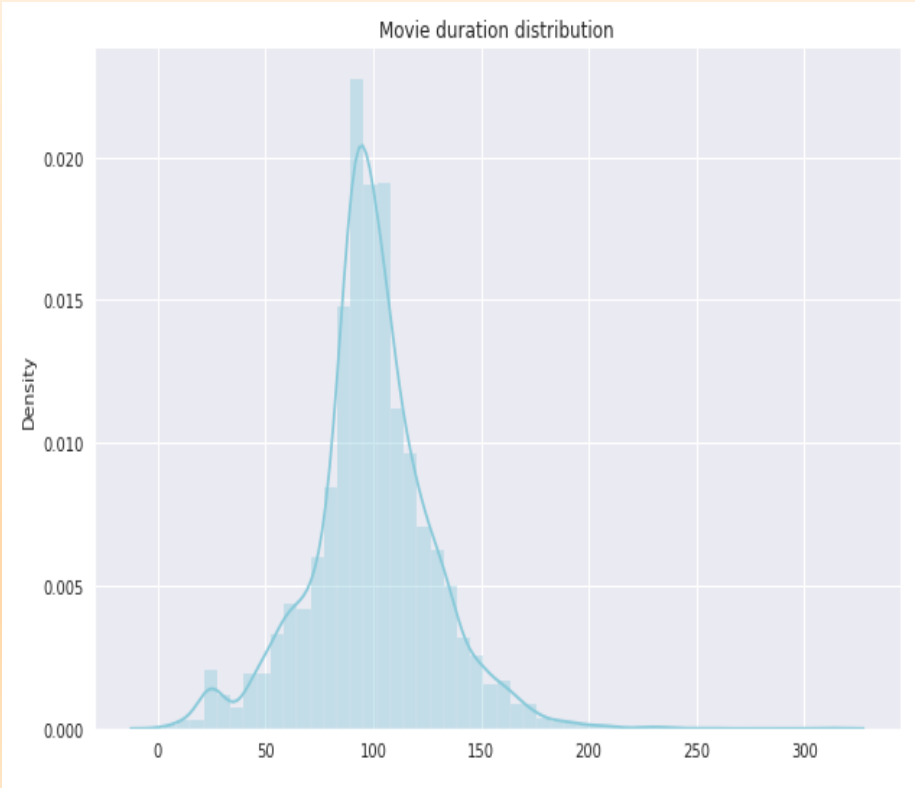
As per the data 69% of all content was occupied by movies.

We can see that US has given us the most content

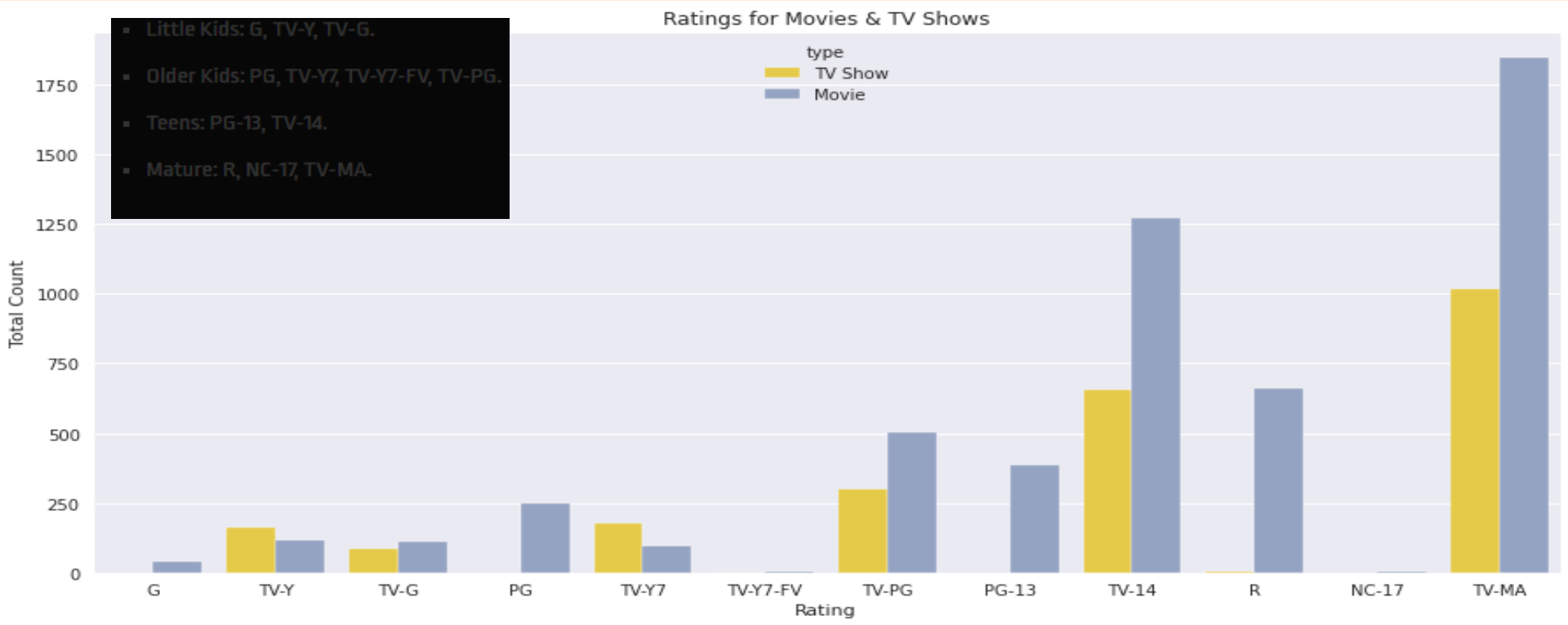
Higher the seasons lower the shows counts

Very few movies beyond the range 50-150

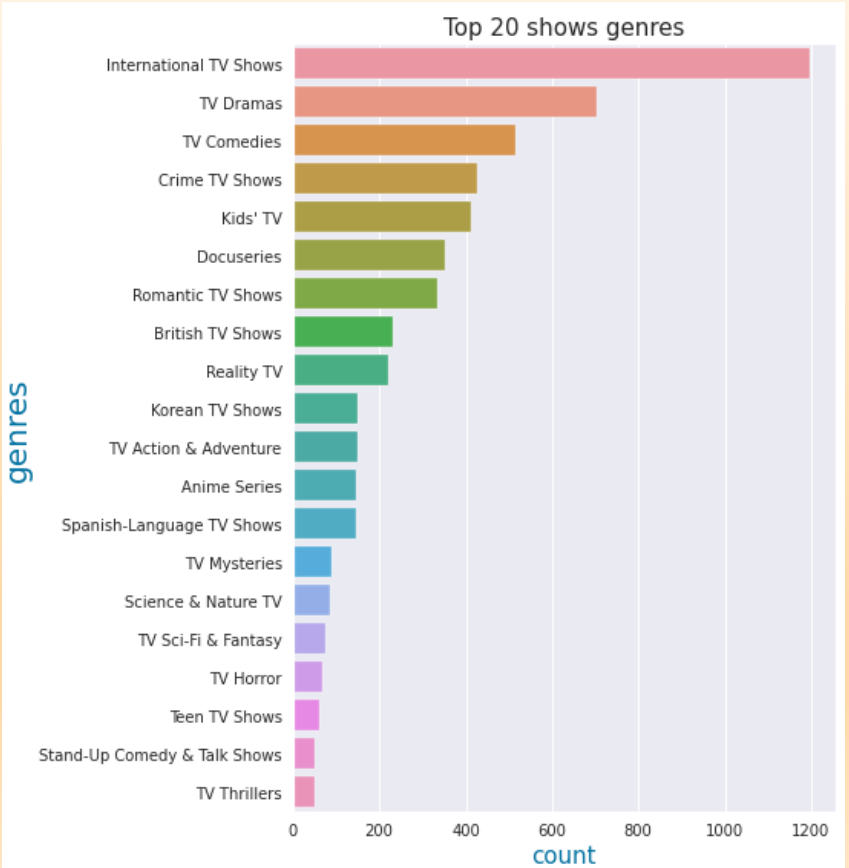


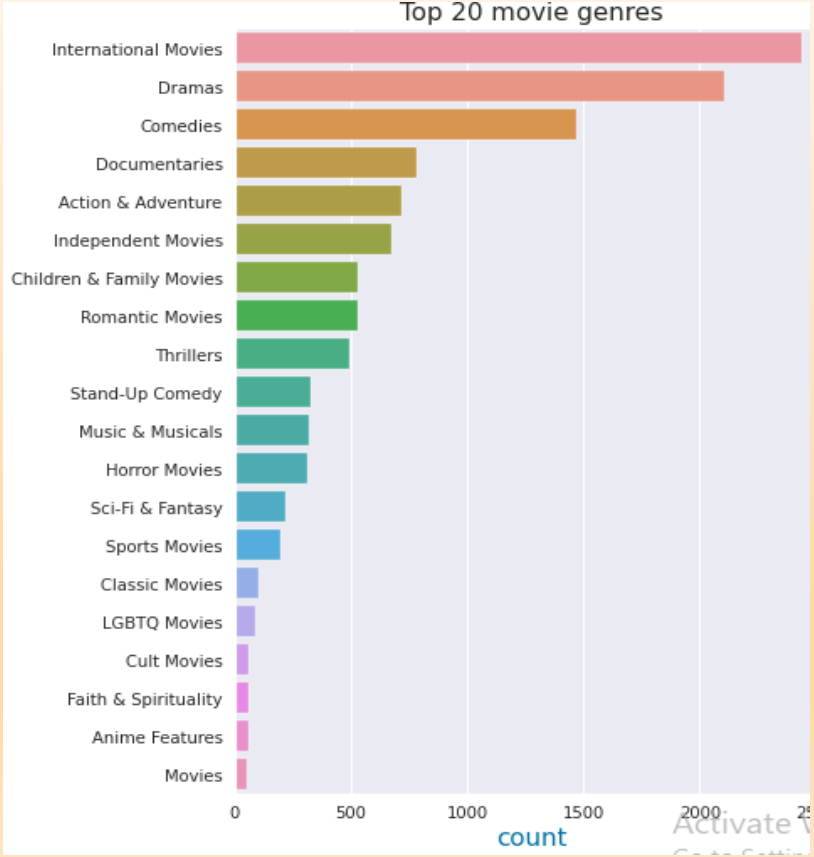


Older you are more content you will get on Netflix

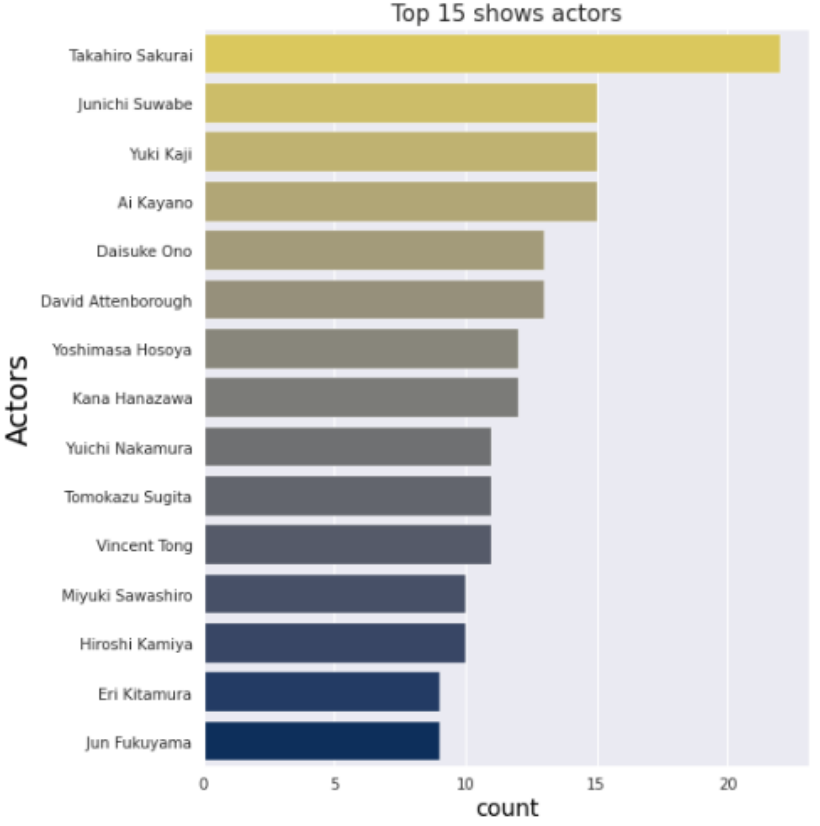


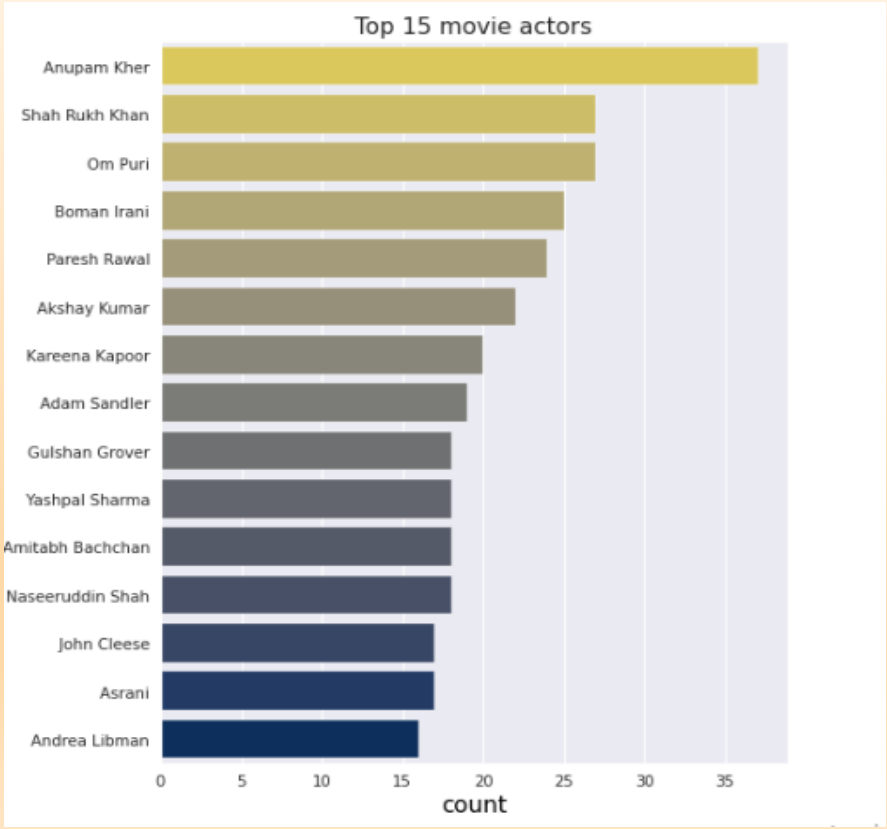
People like international, Drama & Comedies more





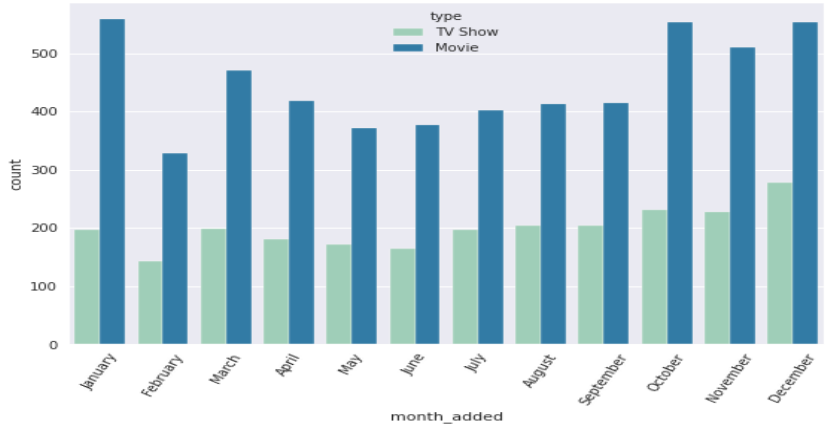
Do you know Takahiro Skurai & Anupam kher

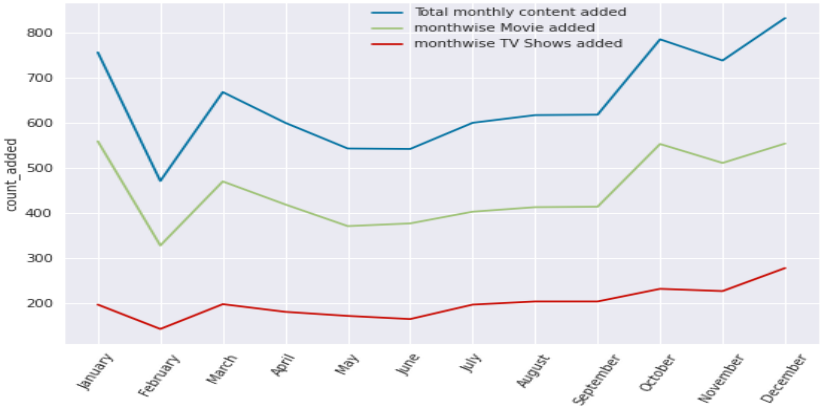




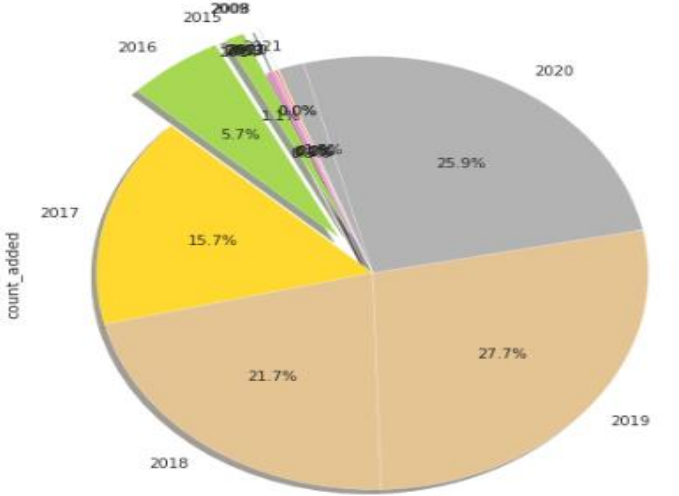
We have increased adding TV Shows on Netflix

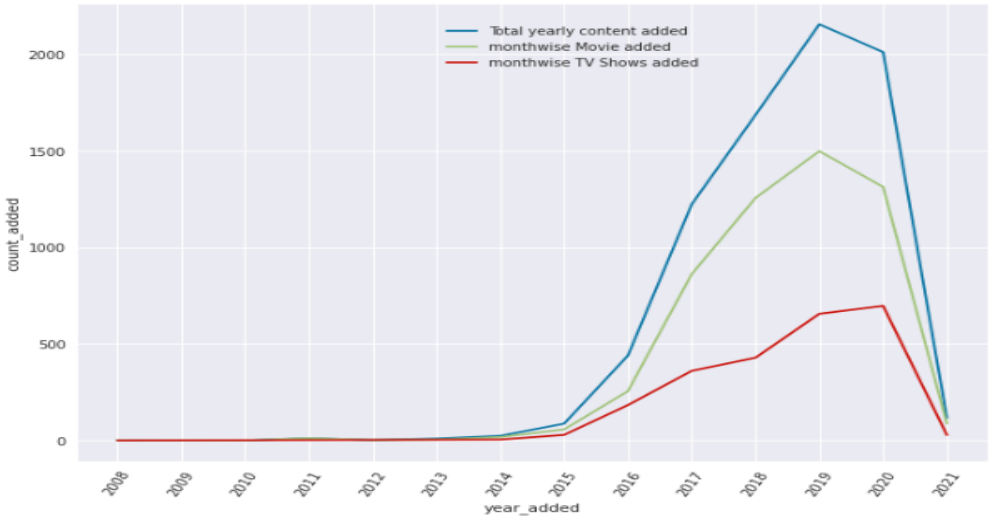
Monthly





Yearly





You watched it. no worries ,we have more similar content !

Similar to‘68 kill’ movies

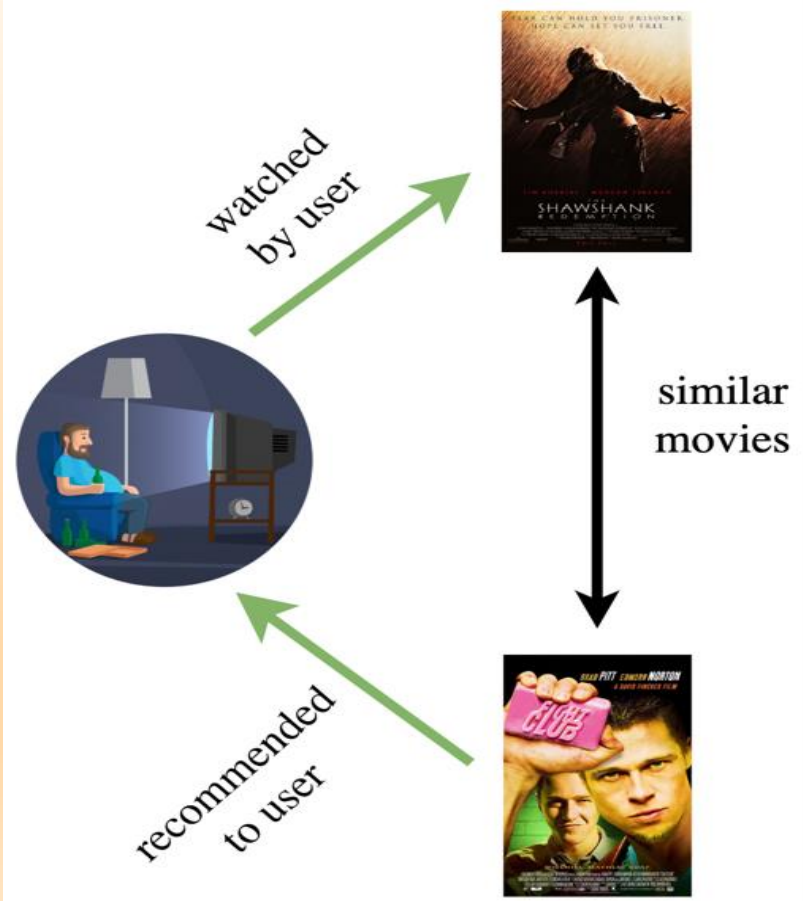
```
get_recommendations('68 Kill')
```

```
808                    Before the Flood
7667                World Trade Center
5675                      Smart People
5030                             Queen
2969             In Search of Fellini
5738                           Sparkle
2341                         Game Over
2847    How to Make an American Quilt
6844                       The Ruthless
5596               Sicilian Ghost Story
```

Similar to ‘1994’ TV shows

```
get_recom_shows('1994')
```

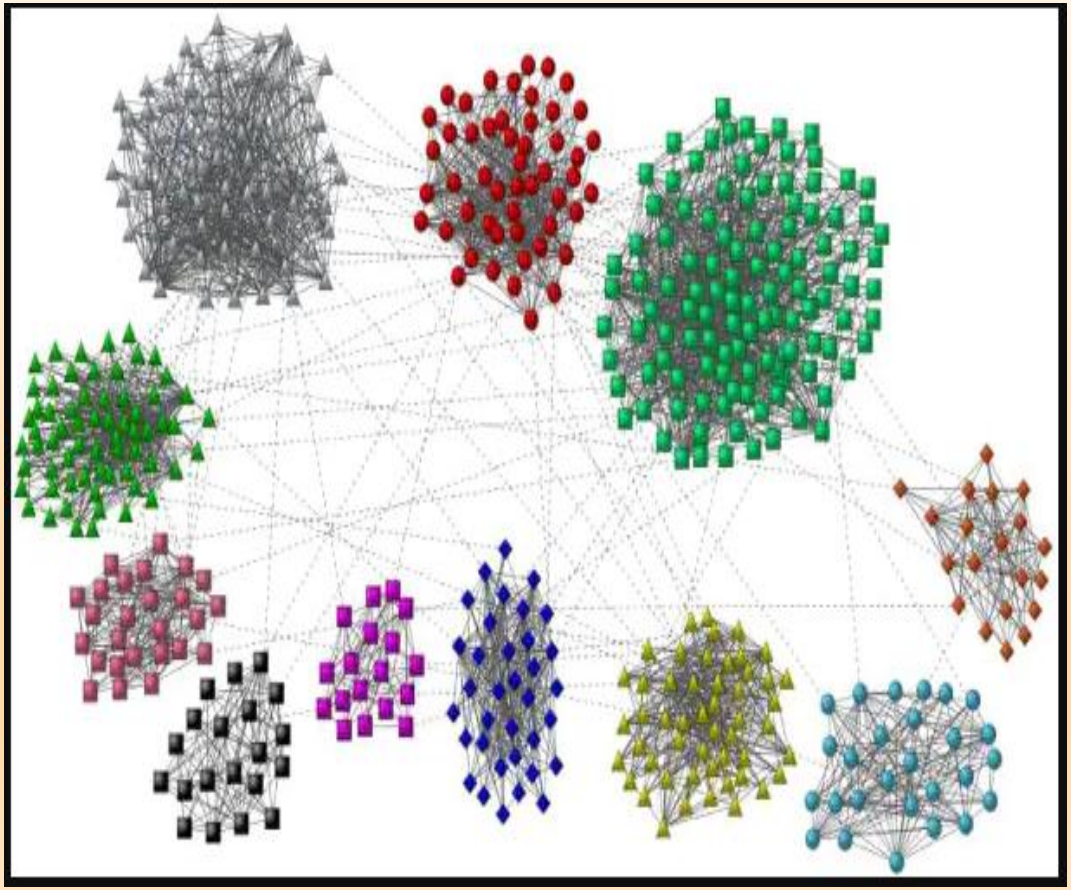
```
5130                            Record of Youth
5119                           Reality of Dream
4170                                Momo Salon
1924                                      Edha
4325                             My Hotter Half
7402                           Velvet Colección
2771    Holiday Home Makeover with Mr. Christmas
2958                                 Imposters
2441                                   Glow Up
2427                                Girlfriends
```



Clustering

AI

Finding similar content and cluster them



K-Modes Clustering

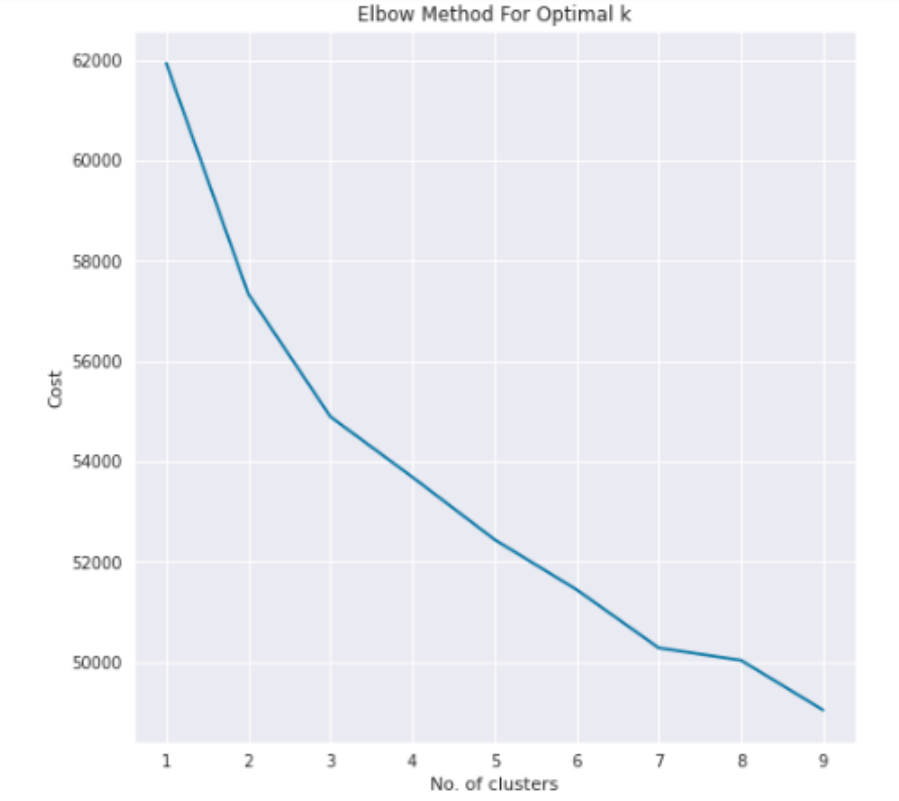
AI

Why K-modes ?

Since we had categorical data and K-modes clustering algorithm was established for categorical data. It works on the modes of categorical data. i.e most occurred observations and build clusters for most similar points.

3 clusters are best no. according to K-Modes:

An Elbow plot was plotted using K-modes algorithm to decide the best suitable no. of cluster for the given dataset. Using this plot we can say that the most suitable no. of clusters is 3 .

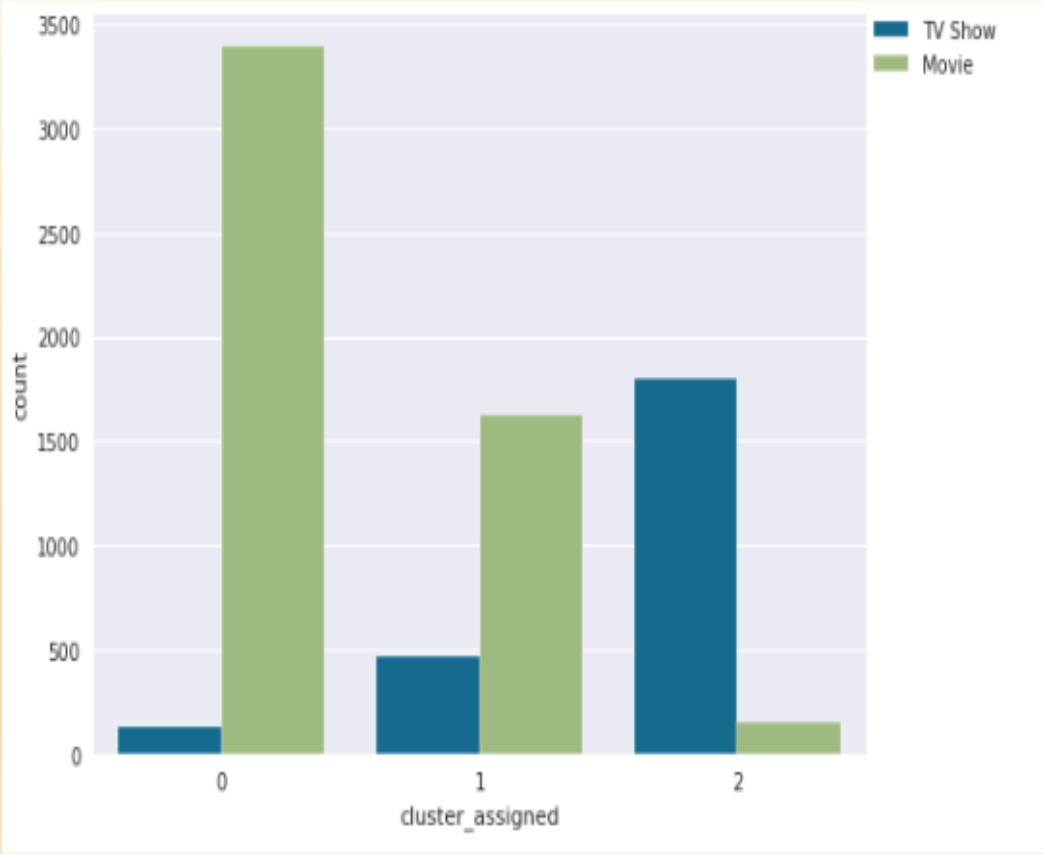


K-Modes Clustering

How the no. of points(content) falls in each cluster ?

After algorithm implementation a count plot was plotted to visualize no. of points falls in each cluster.

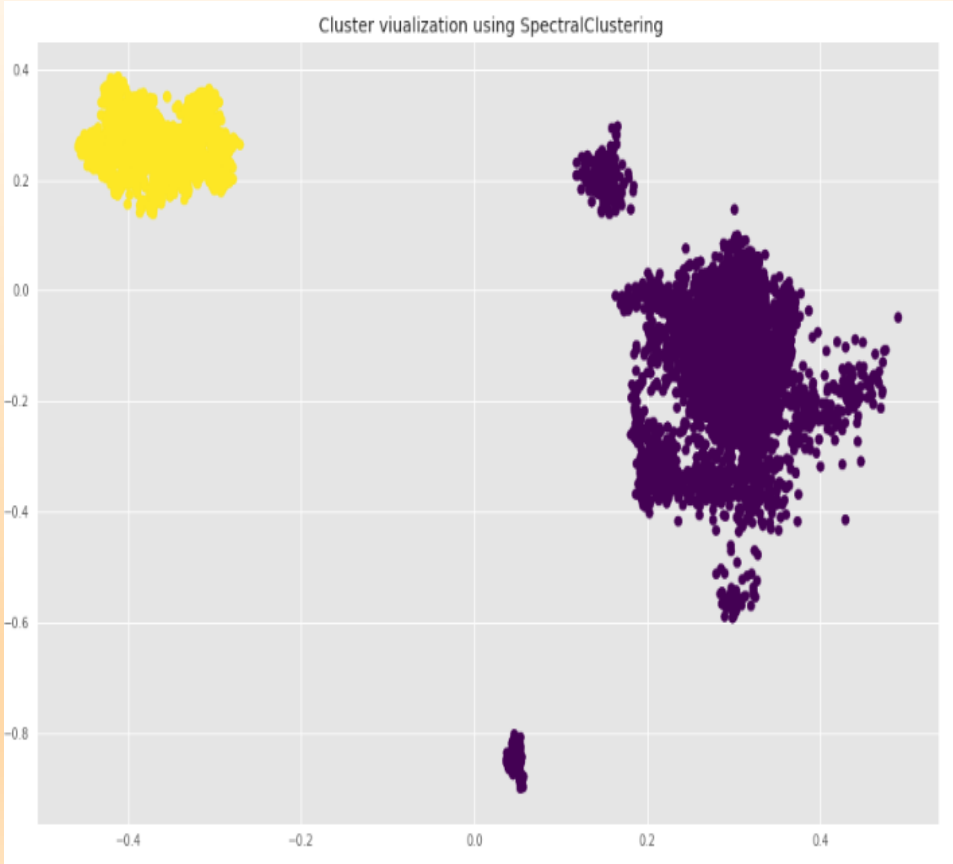
1. Mostly movies are in cluster 0
2. Mostly movies are in cluster1 but the ratio has changed from cluster 0
3. Mostly TV shows are in cluster 2 having only few movies



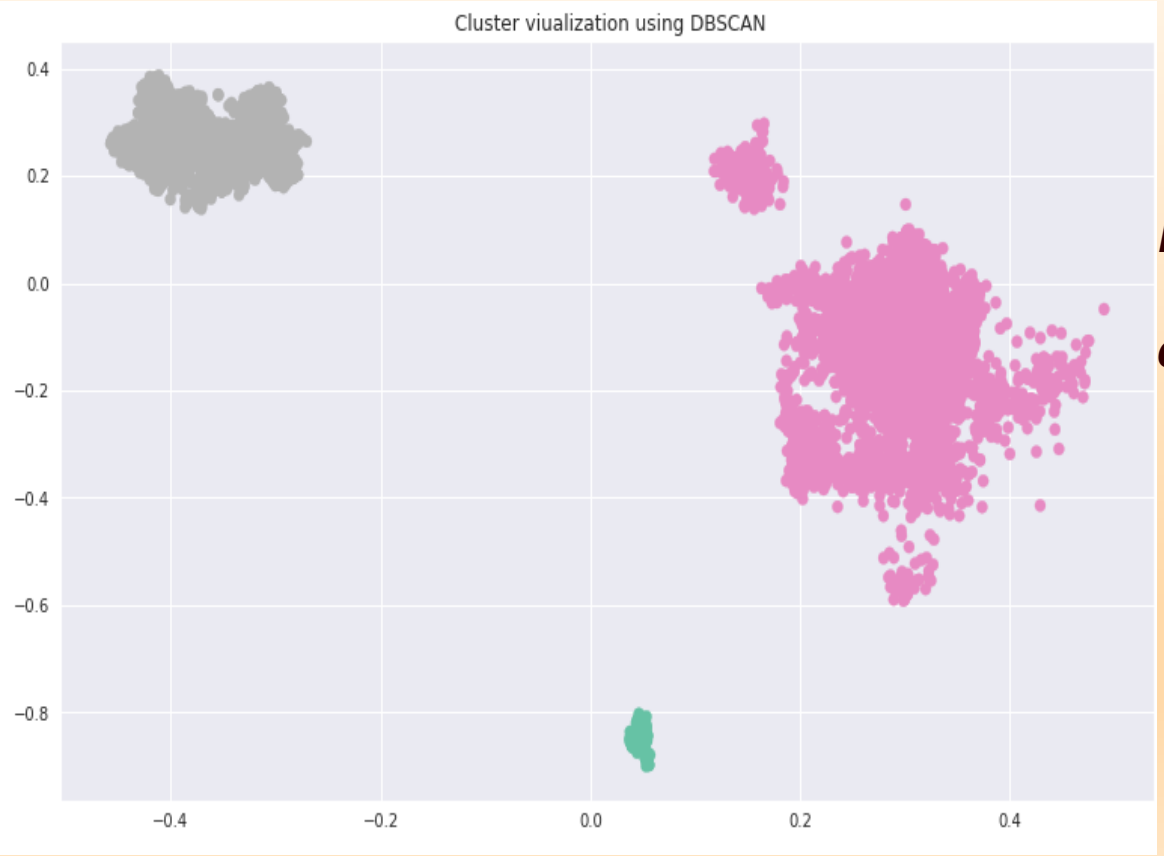
Spectral Clustering

Using Spectral clustering on transformed data we were able to built 2 clusters which is a different results from other algorithms.

We can see that how the clusters are separate from each other.



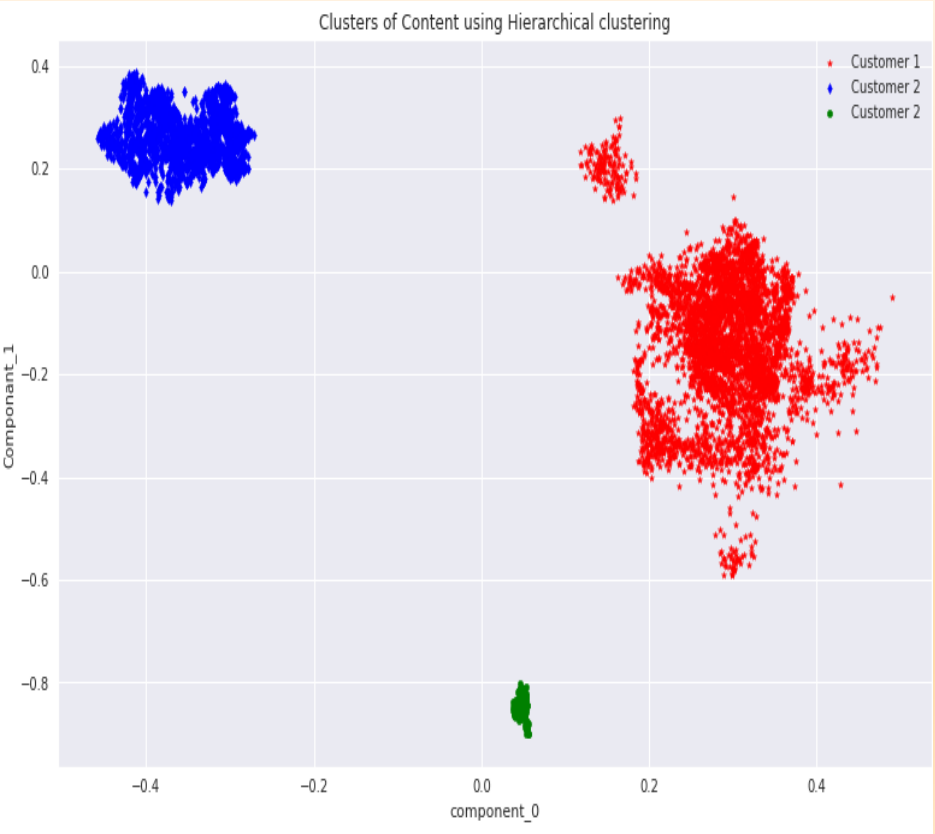
DBSCAN

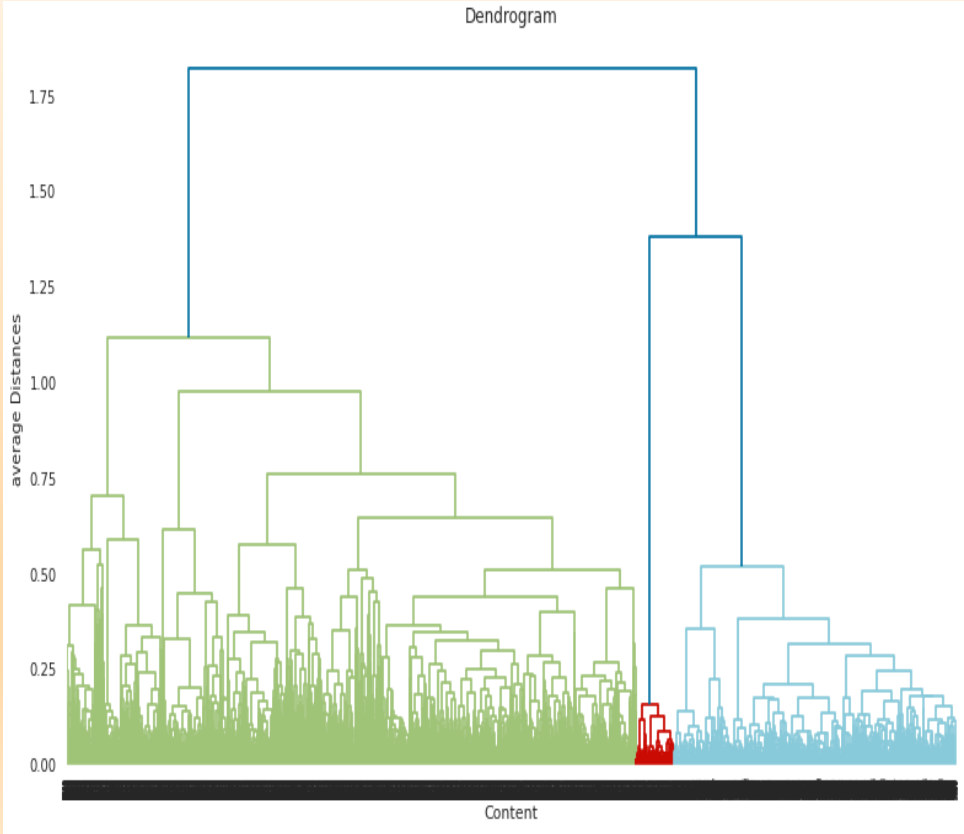


DBSCAN is also a method for clustering. When we use it to cluster the results it give us 3 clusters

Hierarchical Clustering

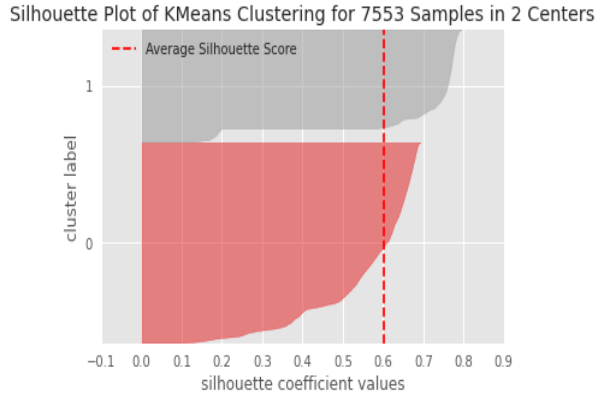
This method of clustering also suggest us to build 3 cluster



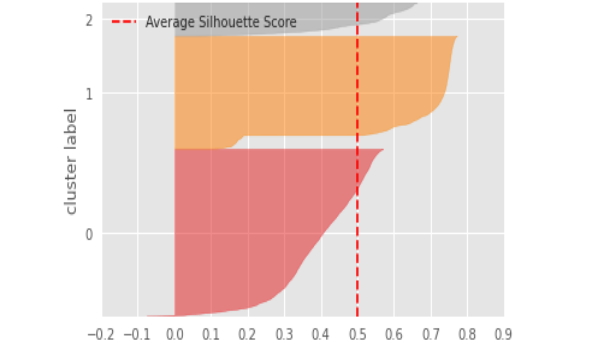


K-means is suggesting 2 clusters

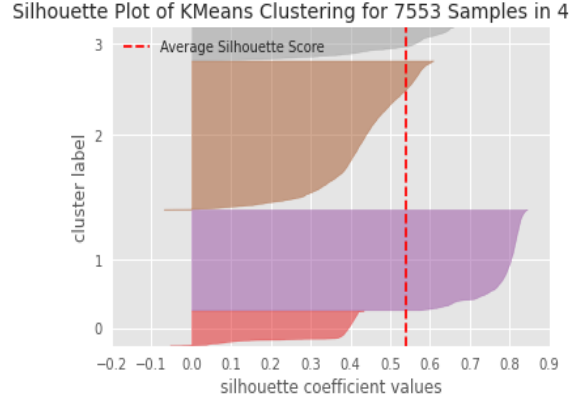
For n_clusters = 2, silhouette score is 0.6014919281005859



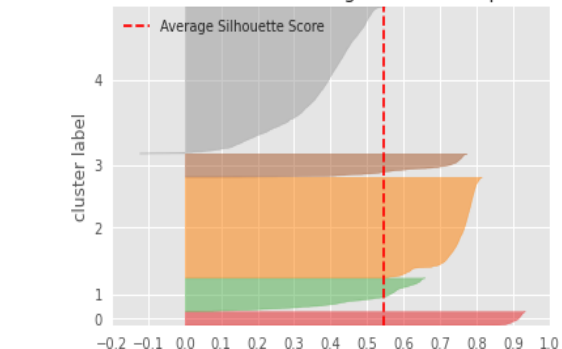
For n_clusters = 3, silhouette score is 0.5002426505088806



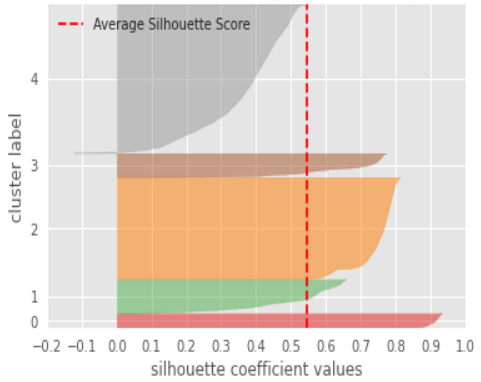
For n_clusters = 4, silhouette score is 0.5387254953384399



For n_clusters = 5, silhouette score is 0.5442843437194824

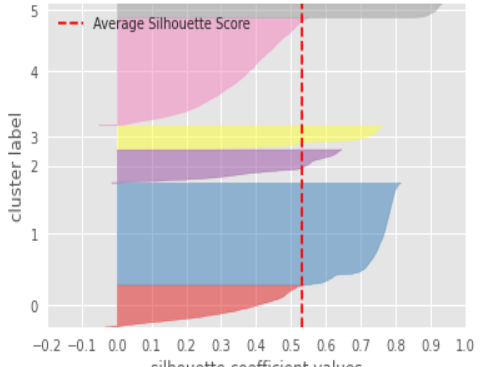


Silhouette Plot of KMeans Clustering for 7553 Samples in 5 Centers



For n_clusters = 6, silhouette score is 0.5323868989944458

Silhouette Plot of KMeans Clustering for 7553 Samples in 6 Centers



What should be the no. of optimal clusters

As most of the data was given in the form of categorical variable hence we have firstly used k-modes algorithm to perform clustering later on using UMAP we have converted our data into numeric and then applied some clustering methods. Results are given below in the table.

We conclude that the optimal no. of clusters is 3 as suggested by most of the algorithms

SL No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Modes with Elbow method	categorical	3
2	DBSCAN	Transformed(numeric)	3
3	Hierachical	Transformed(numeric)	3
4	Spectral Clustering	Transformed(numeric)	2
5	K-Means with Silhouette score	Transformed(numeric)	2

Challenges and conclusions

Challenges

- ***While we were performed analysis on data we come across some nested features(more than 1 value contained by a record in feature). It was a difficult task to per EDA on these features.***
- ***Choosing a right plot for effective visualization was a challenging task.***
- ***Selecting right feature came out to be a challenge for us.***
- ***Since we had mostly textual/categorical data which became an obstacle while selecting the clustering algorithm.***

Conclusions

- ***As per the data 69% of all content was occupied by movies.***
- ***US create highest content that is being uploaded on Netflix.***
- ***There are less shows having higher no. of seasons***
- ***Mostly movies fall in the range of 50-150 minsutes.***
- **People like international, Drama & Comedies hence it is available on Netflix in highest amount.**
- **Takahiro Skurai & Anupam kher have worked in most shows and movies respectively as per available data on Netflix.**
- **Netflix is increasingly focusing on TV shows in comparison of movies.**
- **The optimum no. of clusters that we found is 3.**

Thank You