# Capstone Project

## NETFLIX MOVIES AND TV SHOWS CLUSTERING

## (Unsupervised – Clustering)

*Created By-*
*Mohd Sharik*

# Introduction to Netflix

***Netflix, Inc. is an American subscription streaming service and production company. Launched in 1997, it offers a film and TV show library through distribution deals as well as its own productions, known as Netflix Originals.***

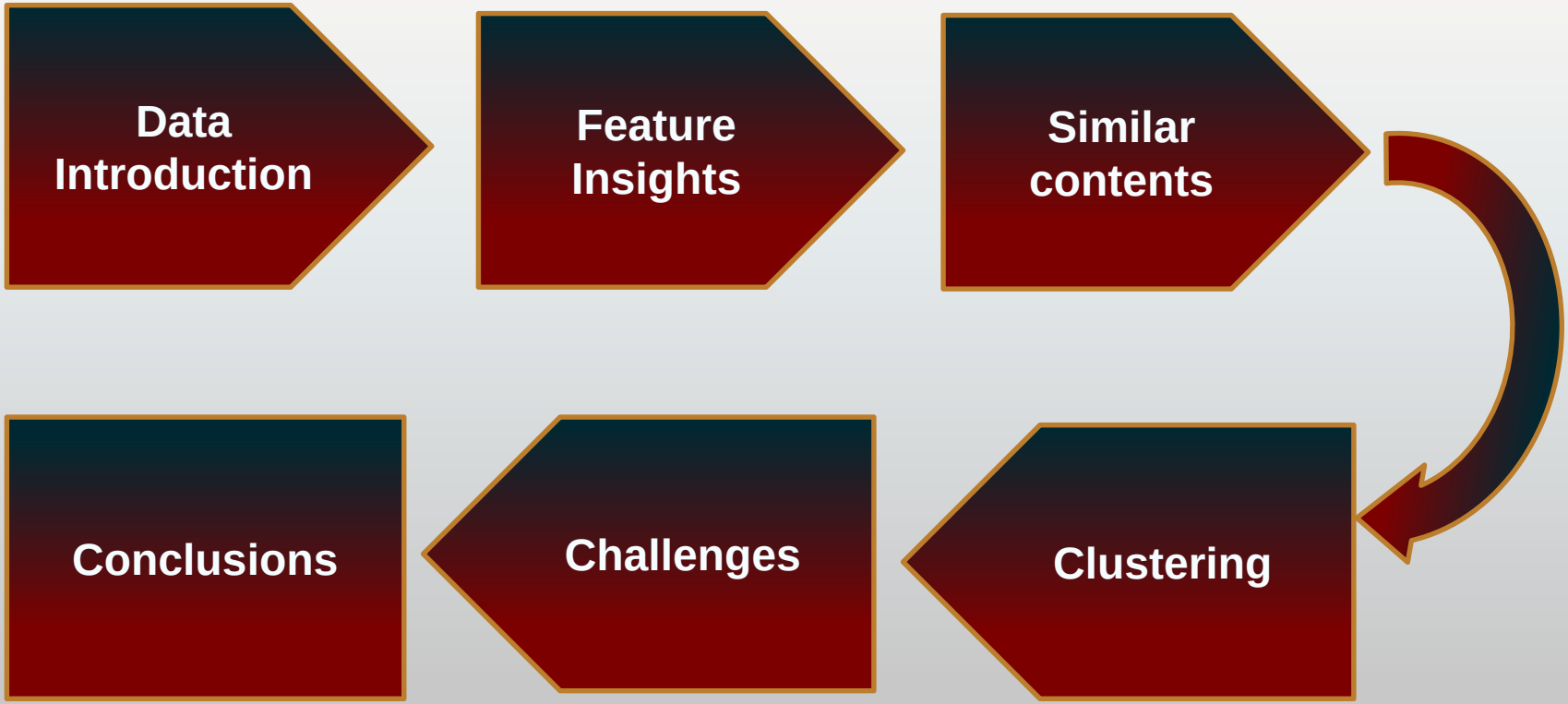
# Problem Statement

***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
Feature Insights
Similar contents
Clustering
Challenges
Conclusions

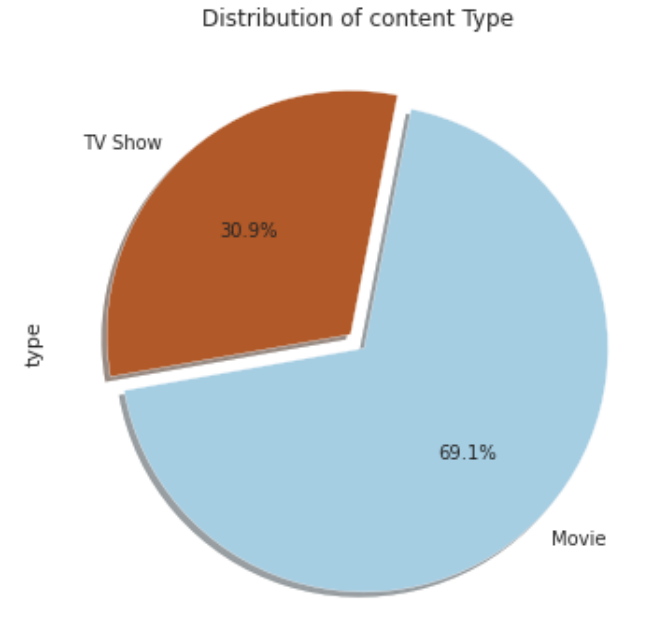
# Data Introduction

| 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



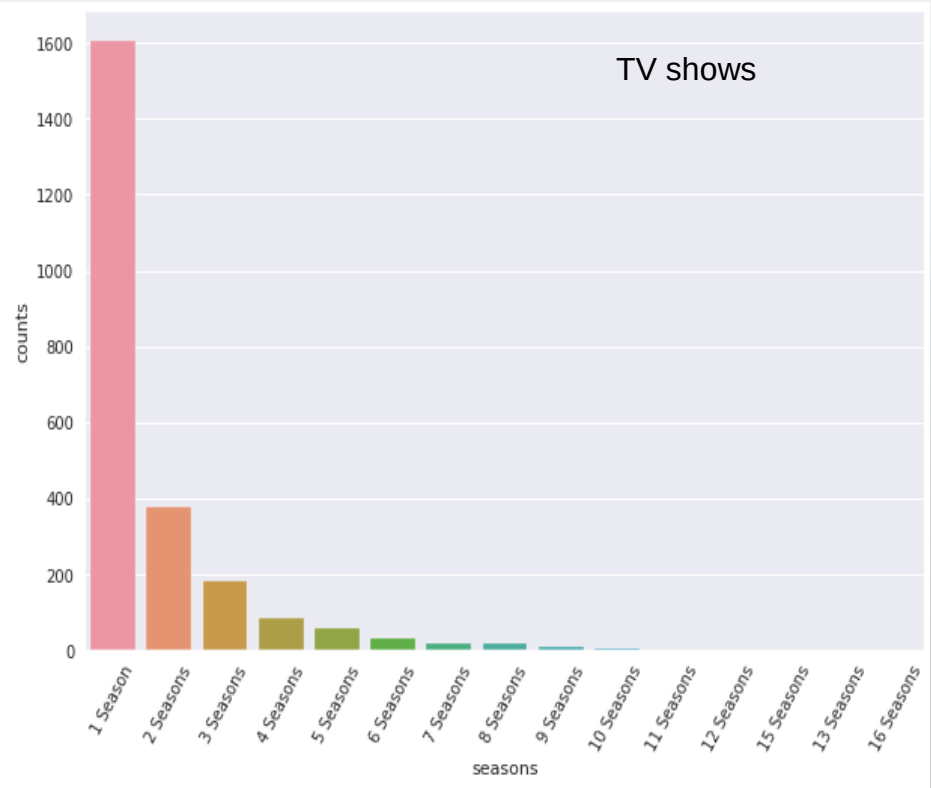

| | country | shows counts | country | movies counts |
|---|---|---|---|---|
| 0 | United States | 860 | United States | 2427 |
| 1 | United Kingdom | 255 | India | 915 |
| 2 | Japan | 182 | United Kingdom | 466 |
| 3 | South Korea | 157 | Canada | 286 |
| 4 | Canada | 126 | France | 265 |
| 5 | France | 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 |

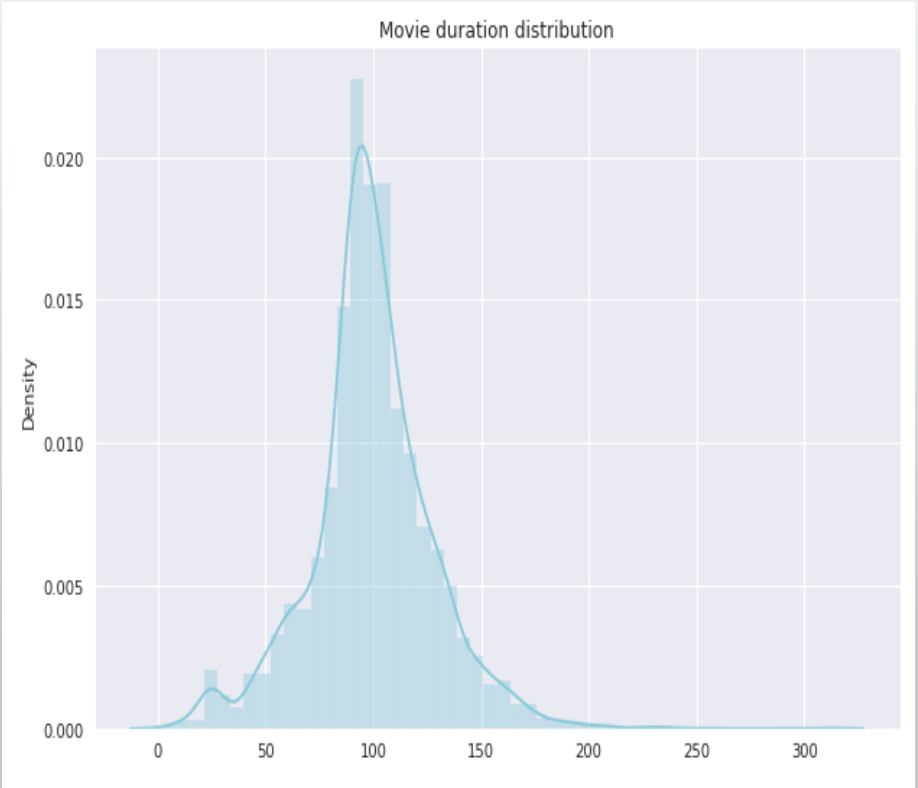
***As per the data 69% of all content was occupied by movies.***

***We can see that US has given us the most content***

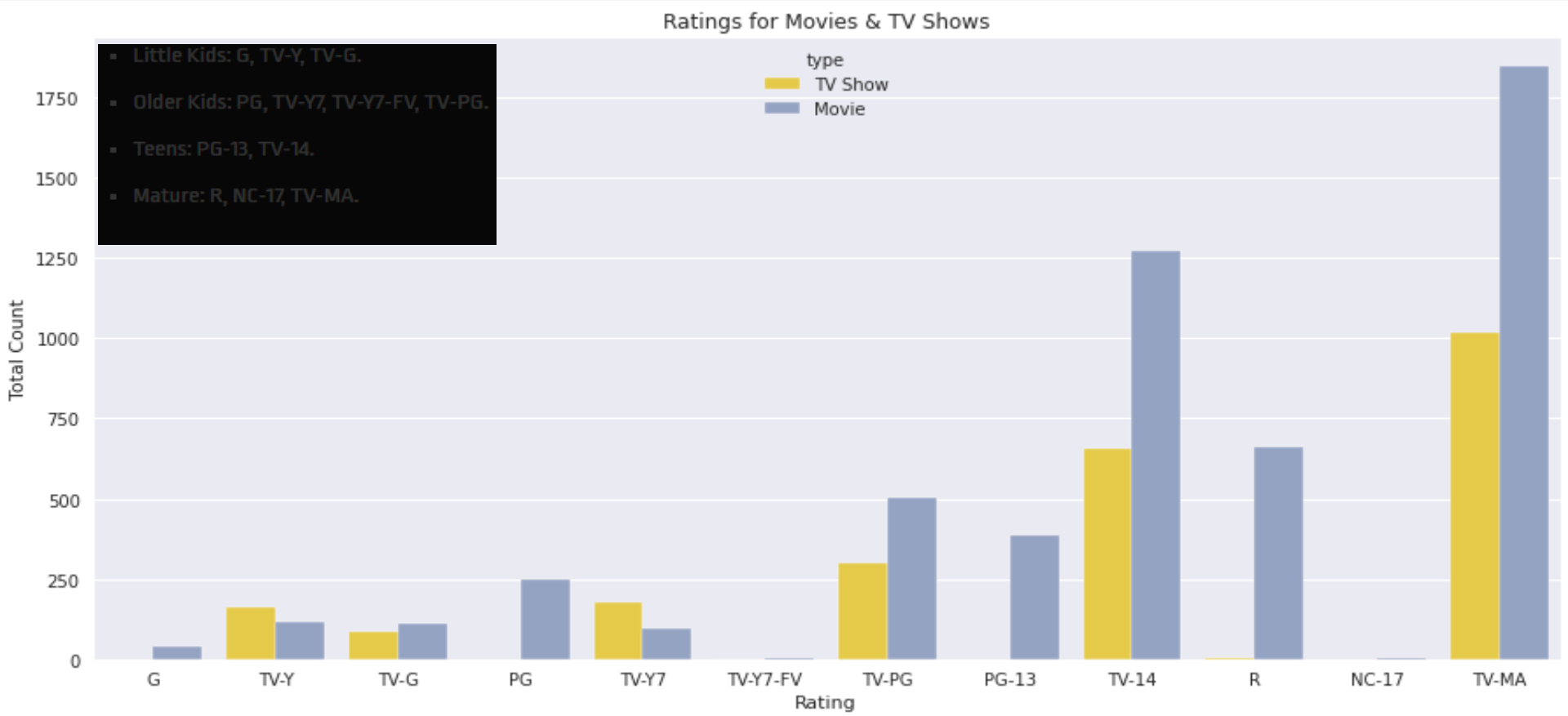
# *Higher the seasons lower the shows counts.*

# *Very few movies beyond the range 50-150.*

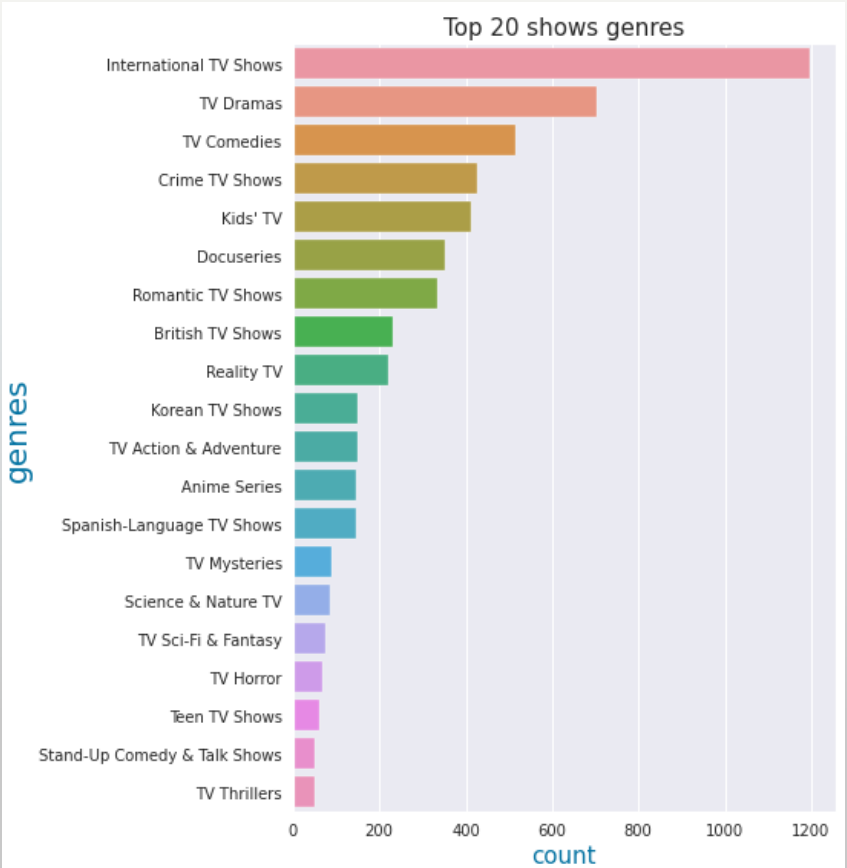


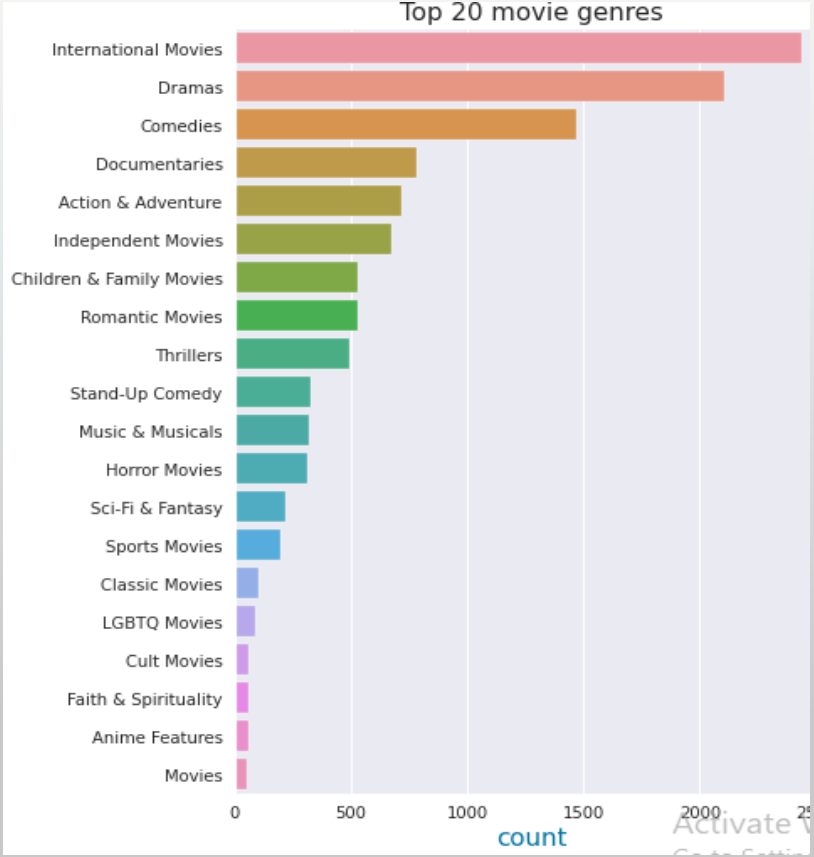



# Older you are more content you will get on Netflix

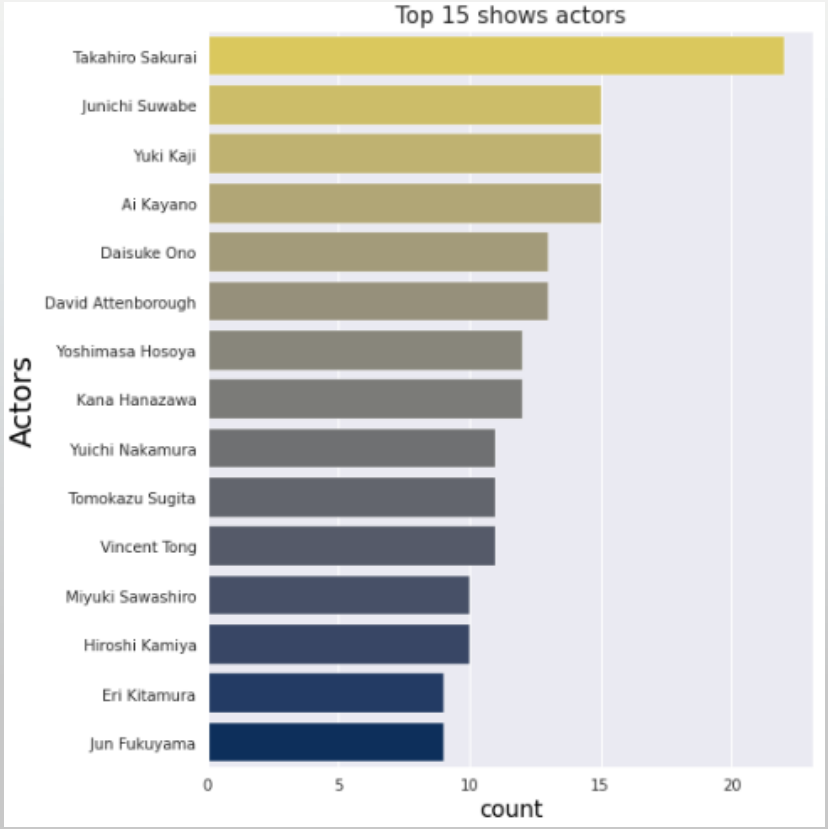


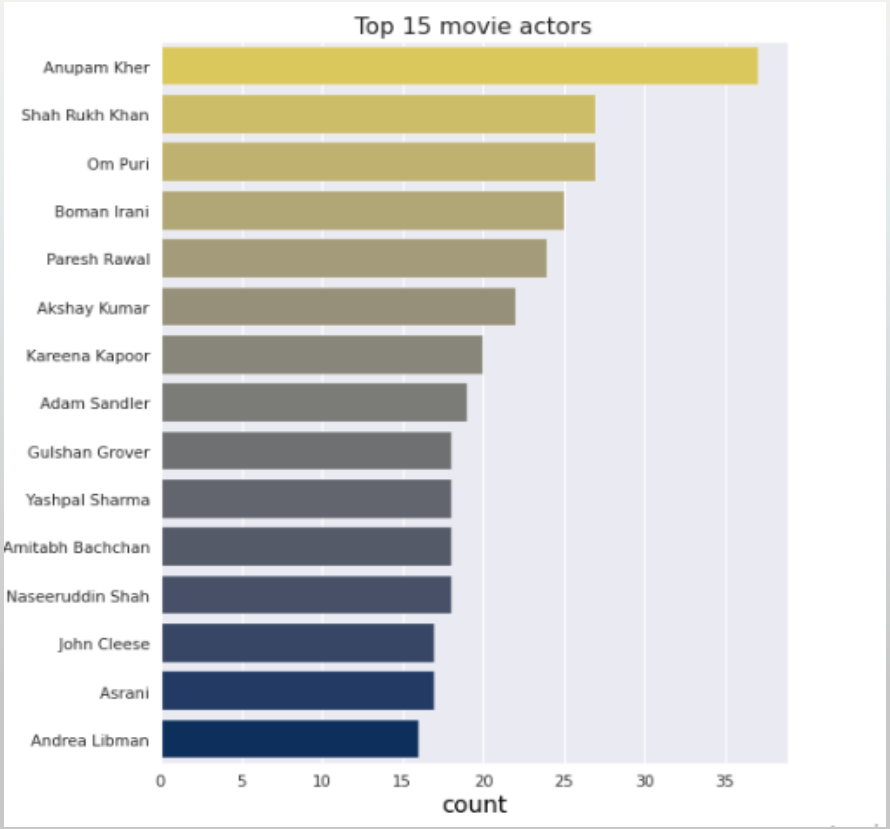
# People like international, Drama & Comedies more



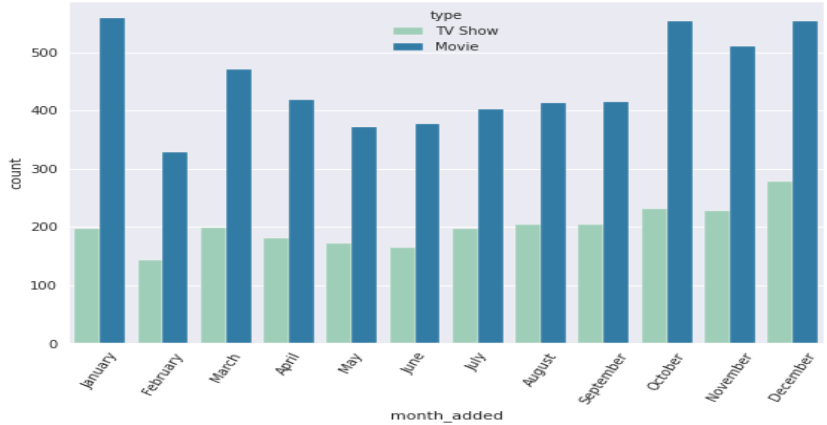
# Do you know Takahiro Skurai & Anupam kher

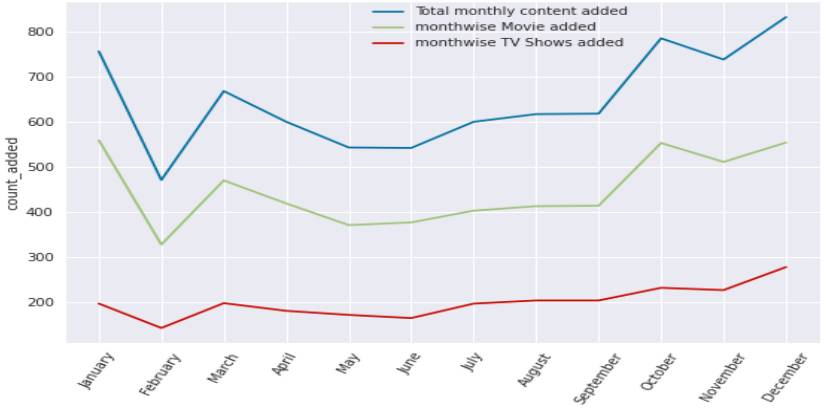


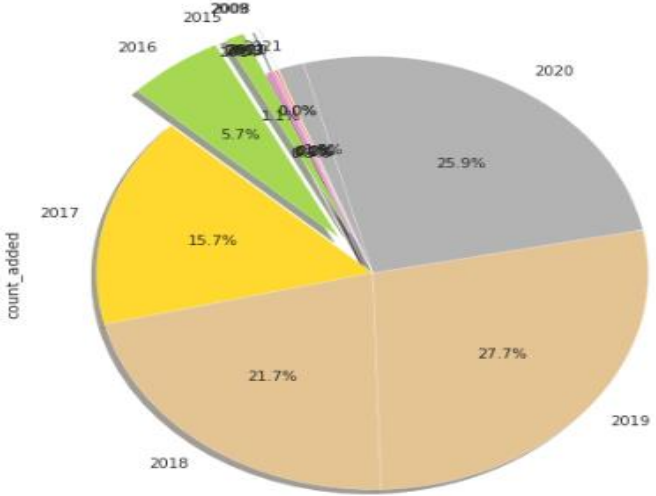
# We have increased adding TV Shows on Netflix

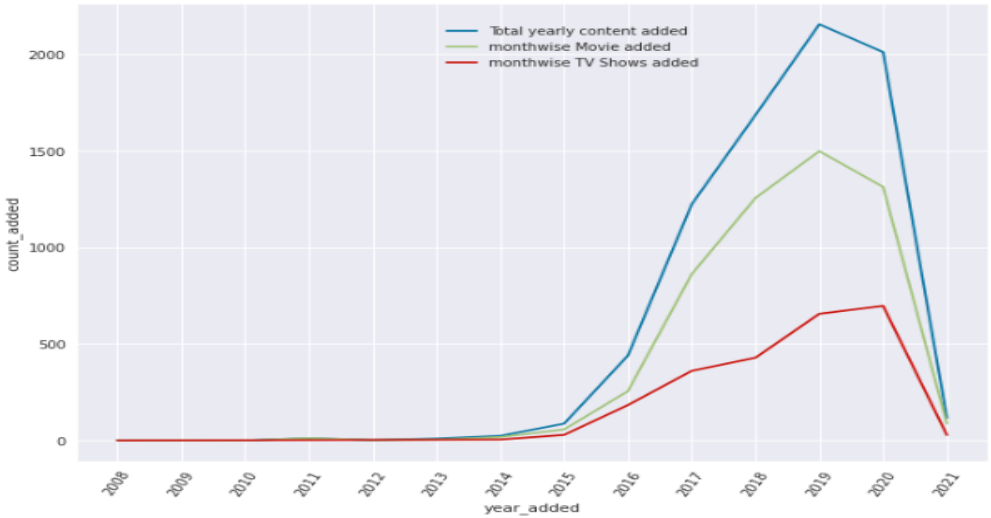
Monthly




Yearly



# Similar contents

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

***Similar to‘68 kill’ movies***

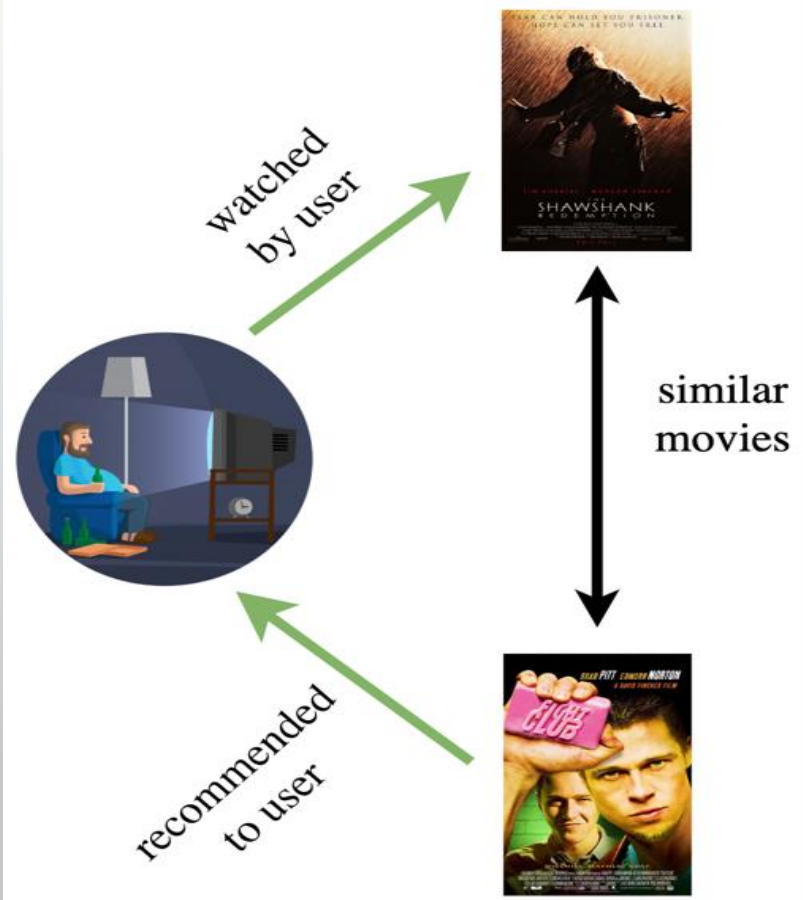
```
get_recommendations('68 Kill')
```

```
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
```

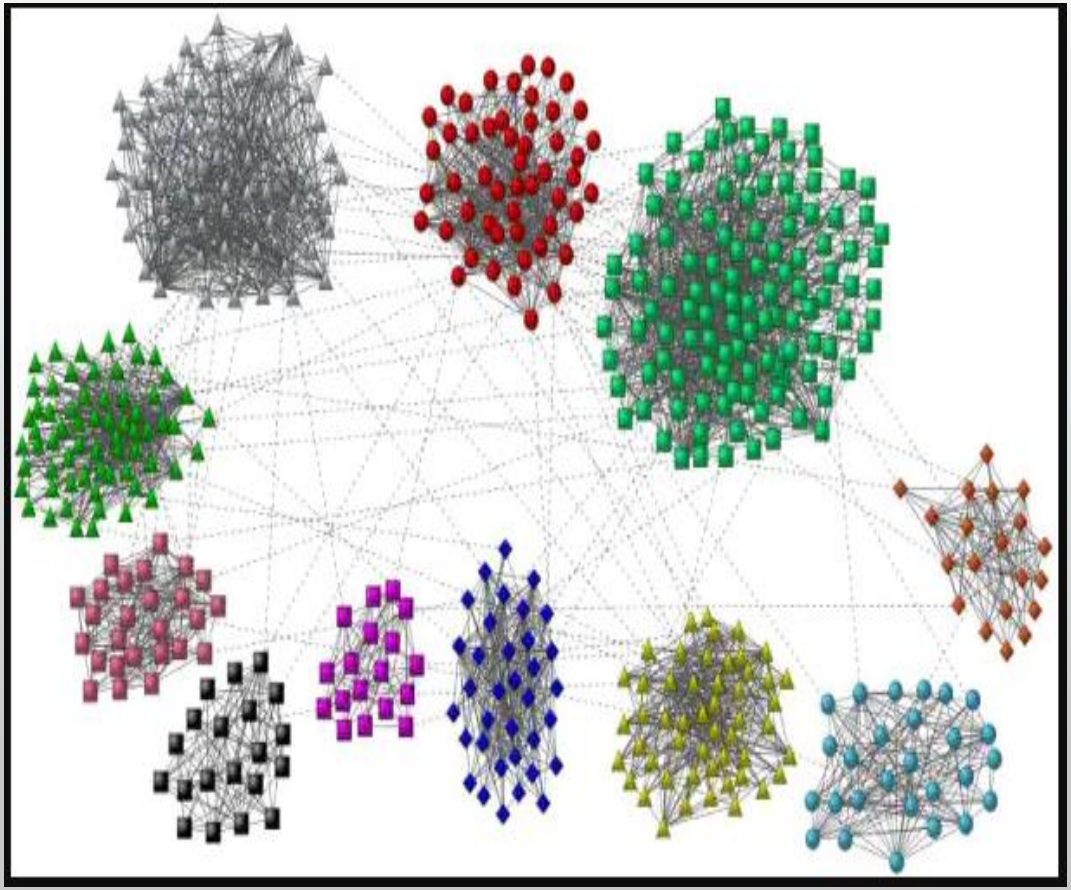
***Similar to ‘1994’ TV shows***

```
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
```

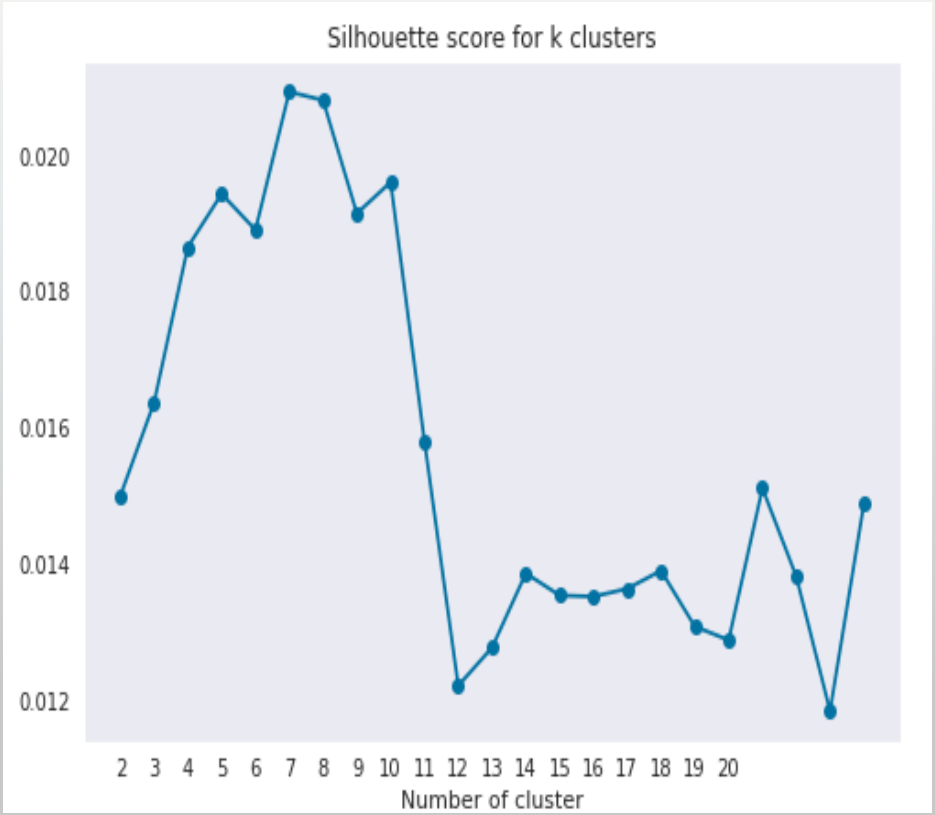


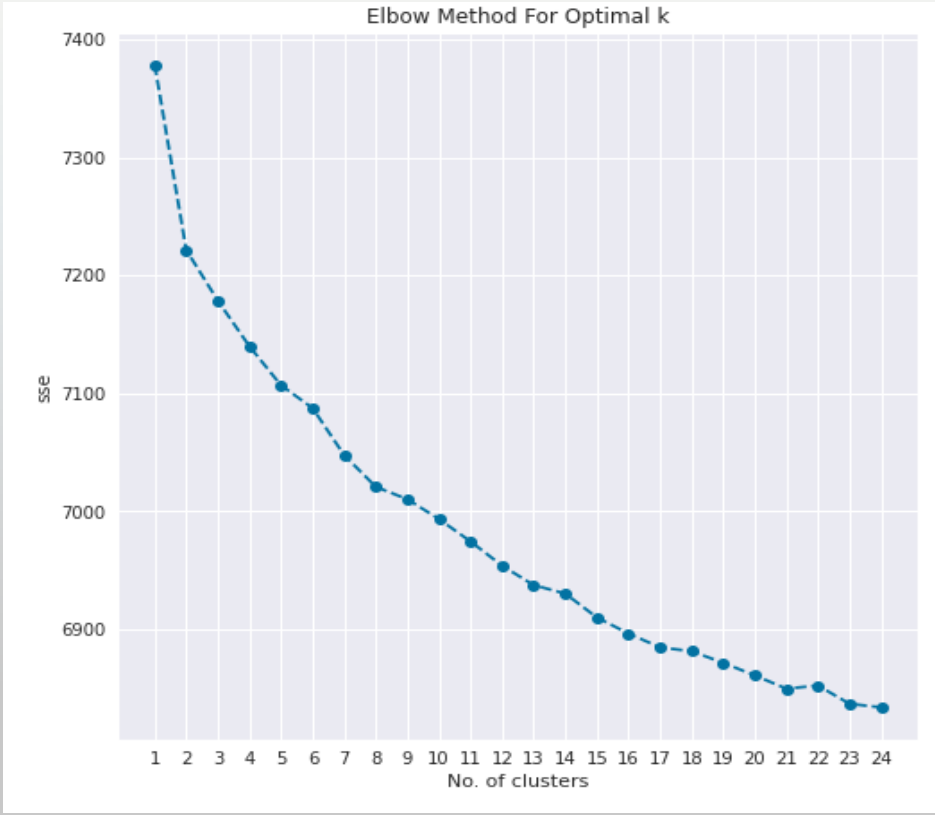
**Finding similar content and cluster them**



# Elbow plot for K-Means Clustering

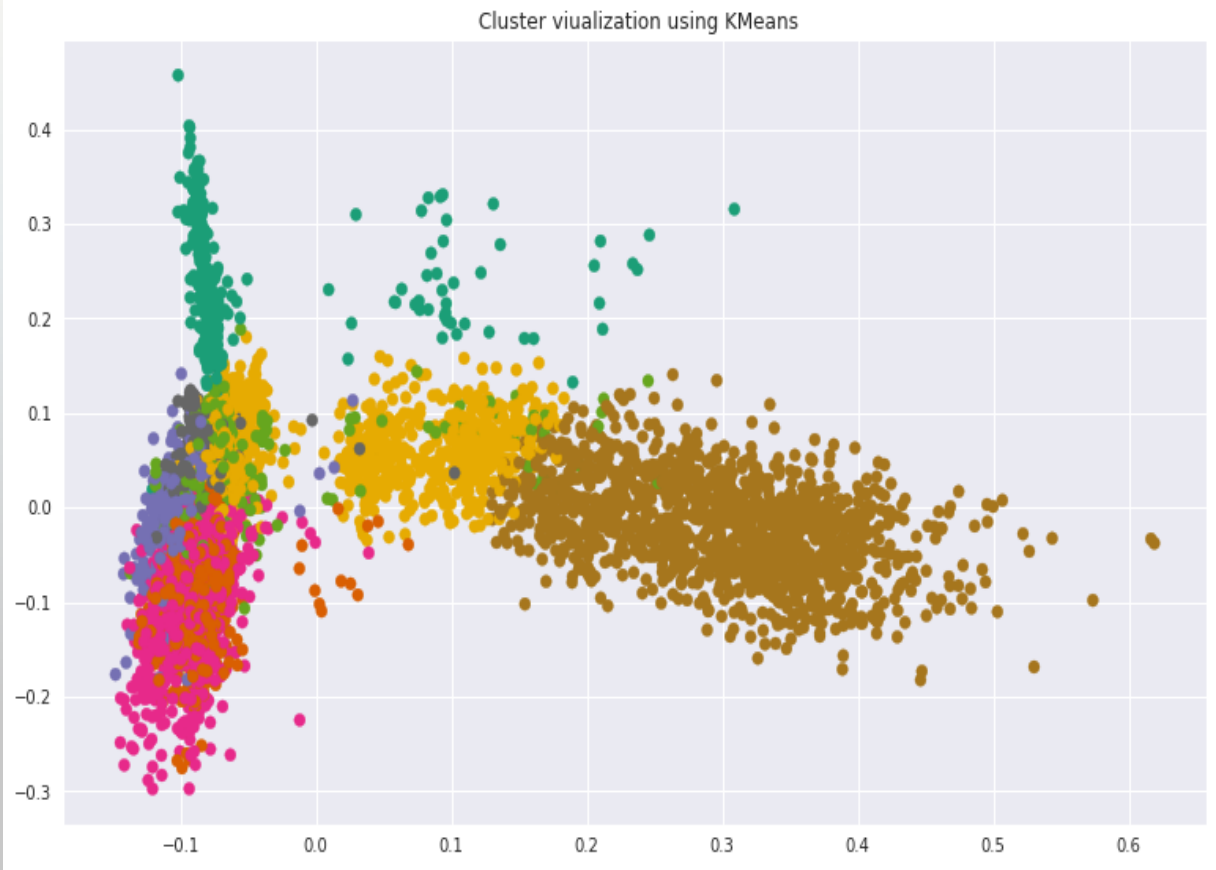
***We have plotted Silhouette score plot and elbow plot(SSE) and found that we have the best optimal no. of clusters can be 8 as the Silhouette score is high and SSE is low at this point in both plot.***



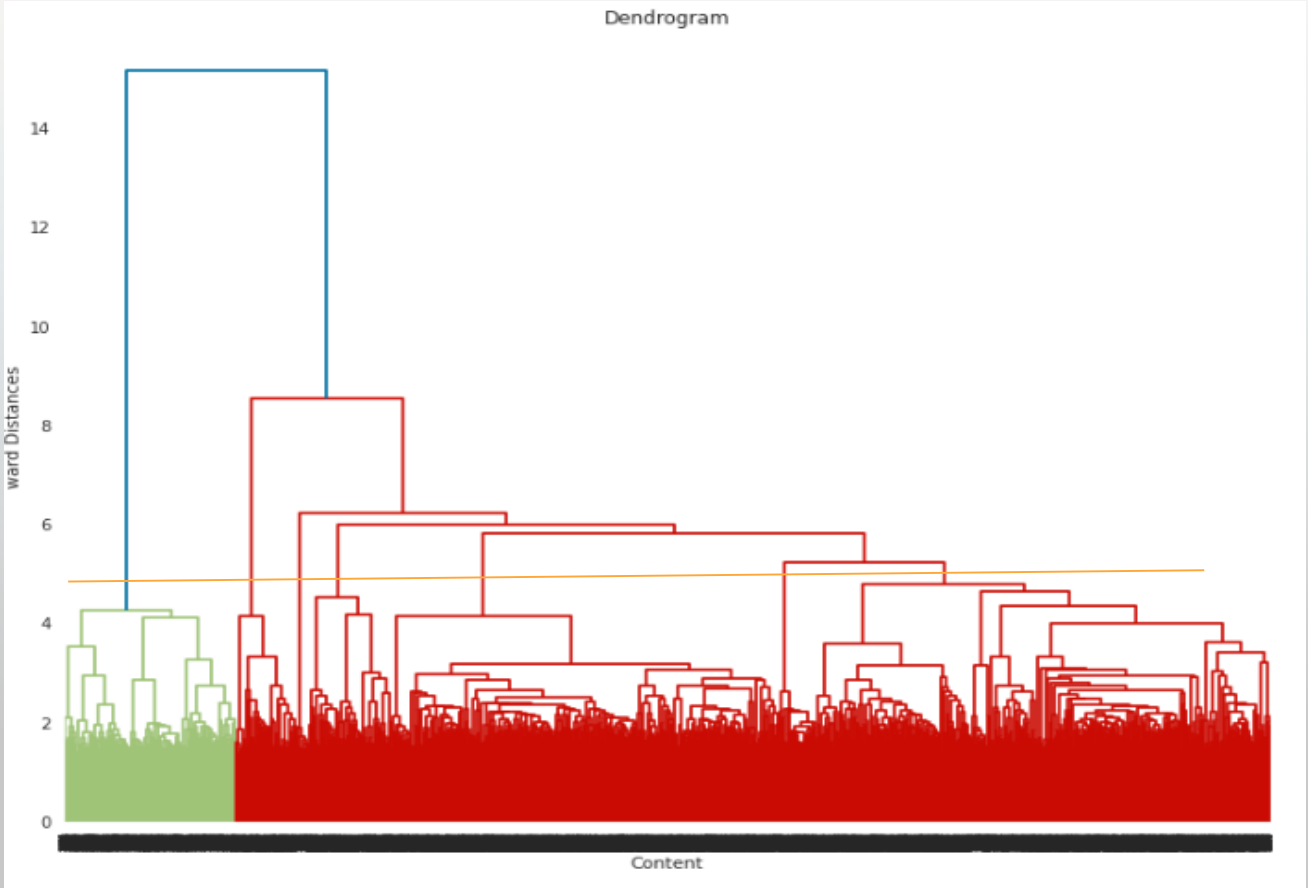
# Visualizing the resultant clusters of KMeans Algorithms

***A visual representation in two dimensions of all 8 clusters can be seen in the plot*.**



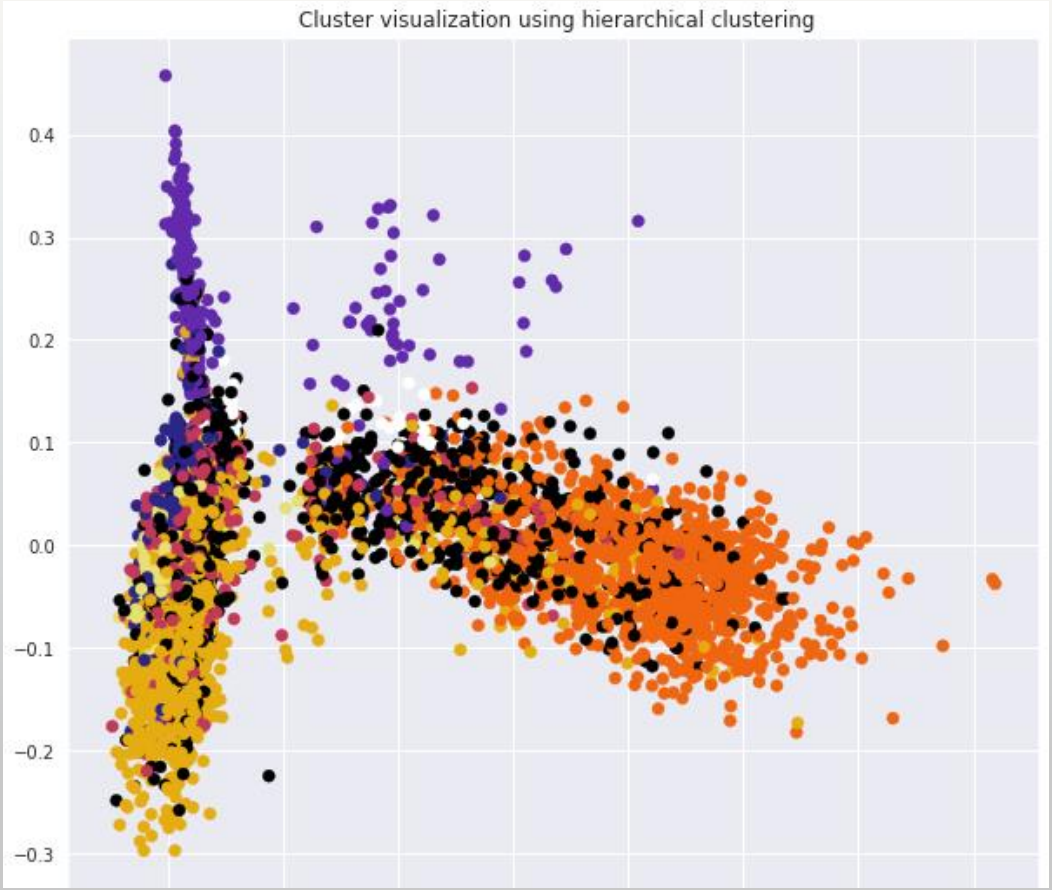
# Dendogram to decide the no. of clusters in hierarchical clustering

***Dendogram plot is the key plot to decide the no. of clusters in hierarchical clustering. In this plot we can see that cluster 6 and 8 can be taken into consideration. However keeping in mind Kmeans result we will take 8 as no of our clusters.***



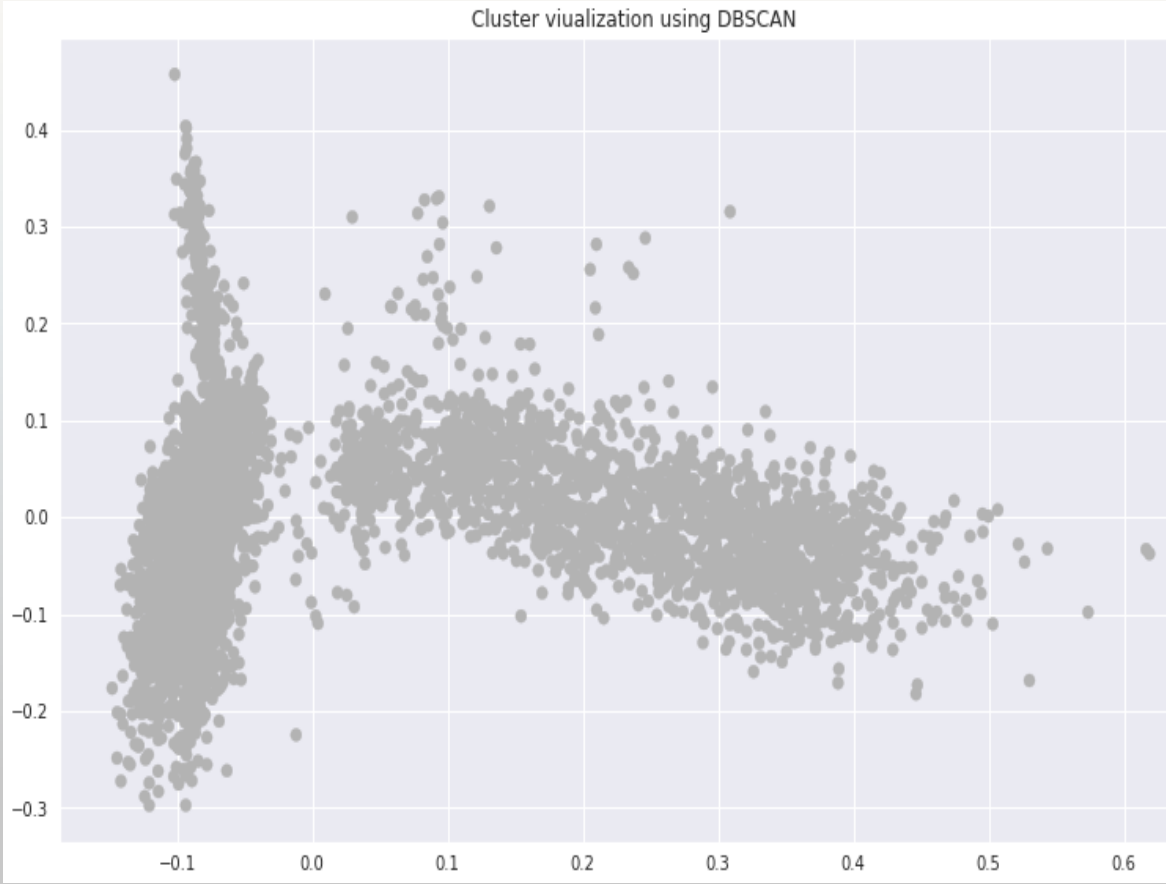
# Visualizing the results of hierarchical clustering

***We can visualize the results of hierarchical clustering. Here the results that we are visualizing are in two dimensions.***



# DBSCAN is not a good choice for us

***We also have tried to use DBSCAN however we couldn't get more than one cluster hence we can say that DBSCAN is not a good choice for us.***



# 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 minutes.***
- **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 8.**

# Thank You