Capstone Project-4

NETFLIX MOVIES & TV SHOWS CLUSTERING

Team Members

Debabrata Sahoo

Vinay Vijay Lanjewar

CONTENT

- Introduction
- Problem Statement
- Data Summary
- Data wrangling
- EDA
- Text Preprocessing
- K-means Clustering
- Feature Selection & ML algo used
- Conclusion

Introduction

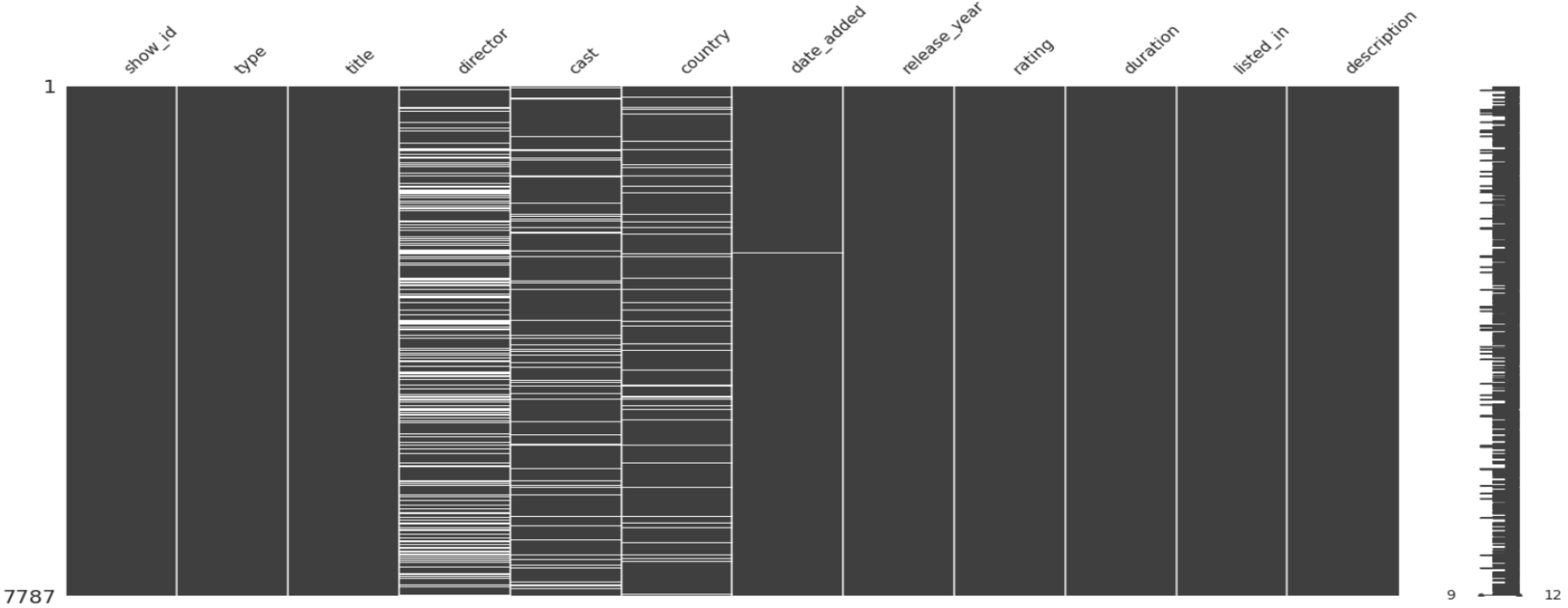
- Netflix is a prominent OTT platform with a wide variety of content to view from a variety of nations and genres, so keep an eye on it. This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine
- The idea of this project is to analyze and perform clustering to determine various patterns related to the content available in Netflix. Based on the attributes related to the Tv shows or movies, we will be implementing different clustering algorithms which comes under unsupervised Machine learning category.

DATA SUMMARY

The dataset has 7787 rows and 12 columns.

- **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
- **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

DATA WRANGLING



Missing values -

- “Director" has the most missing value followed by "cast" and "country".
- There are few missing value in "date_added" and "rating".

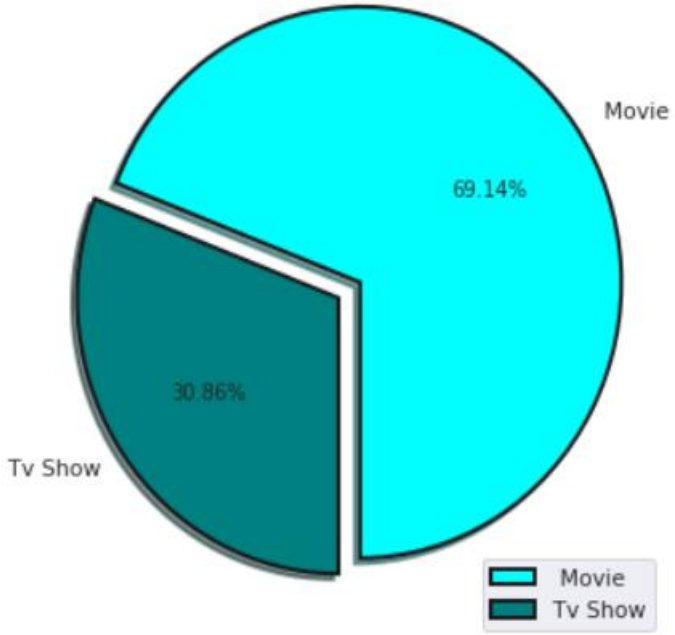
Data Cleaning

Null Value Treatment:

- • ***Director*** feature have more than ***30.68%*** of null values. Filling null values by 'unknown'.
- • ***Country*** feature have ***6.51%*** of null values. Filling null values by mode of feature.
- • ***Cast*** feature have ***9.22%*** of null values. Filling null values by 'unknown'.
- • ***Rating*** feature have ***0.09%*** of null values. Filling null values by mode of feature.
- •***Date_added*** feature have ***0.13%*** of null values. Dropping rows corresponding to null values.

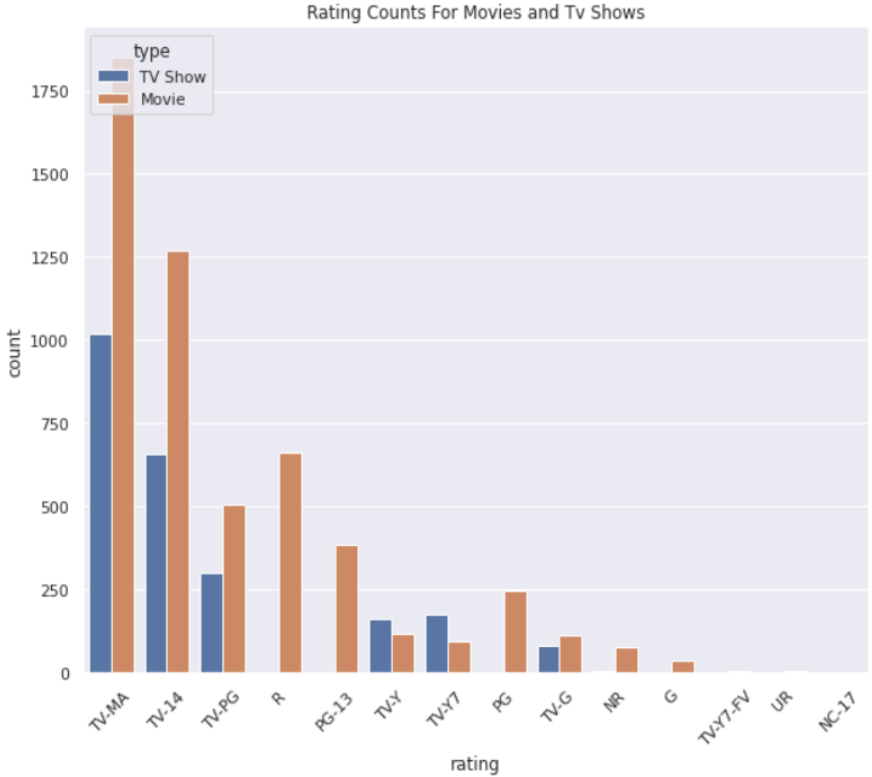
Exploratory Data Analysis(EDA)

Type of content available on Netflix



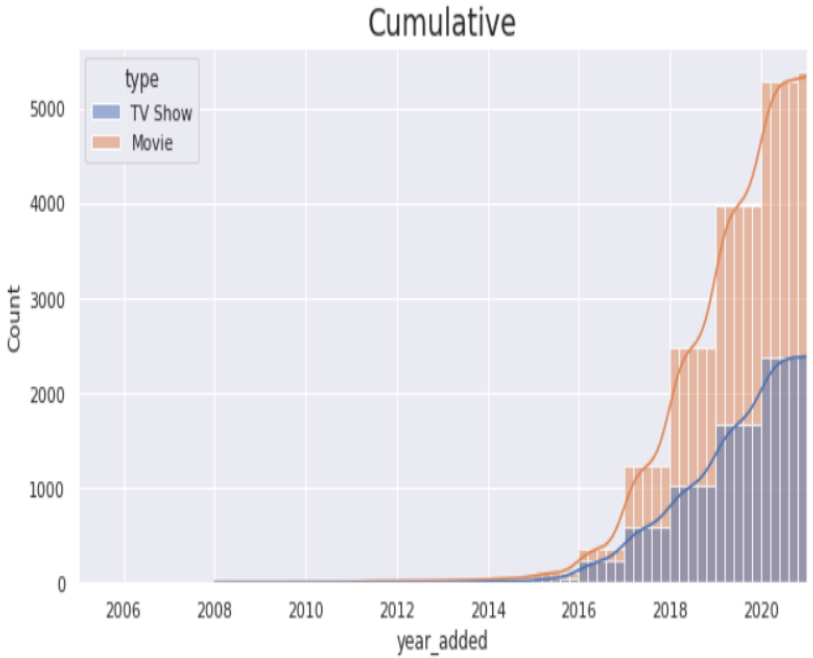
- It is evident that there are more movies on Netflix than TV shows.
- Netflix has 5377 movies, which is more than double the quantity of TV shows.

Movie ratings analysis



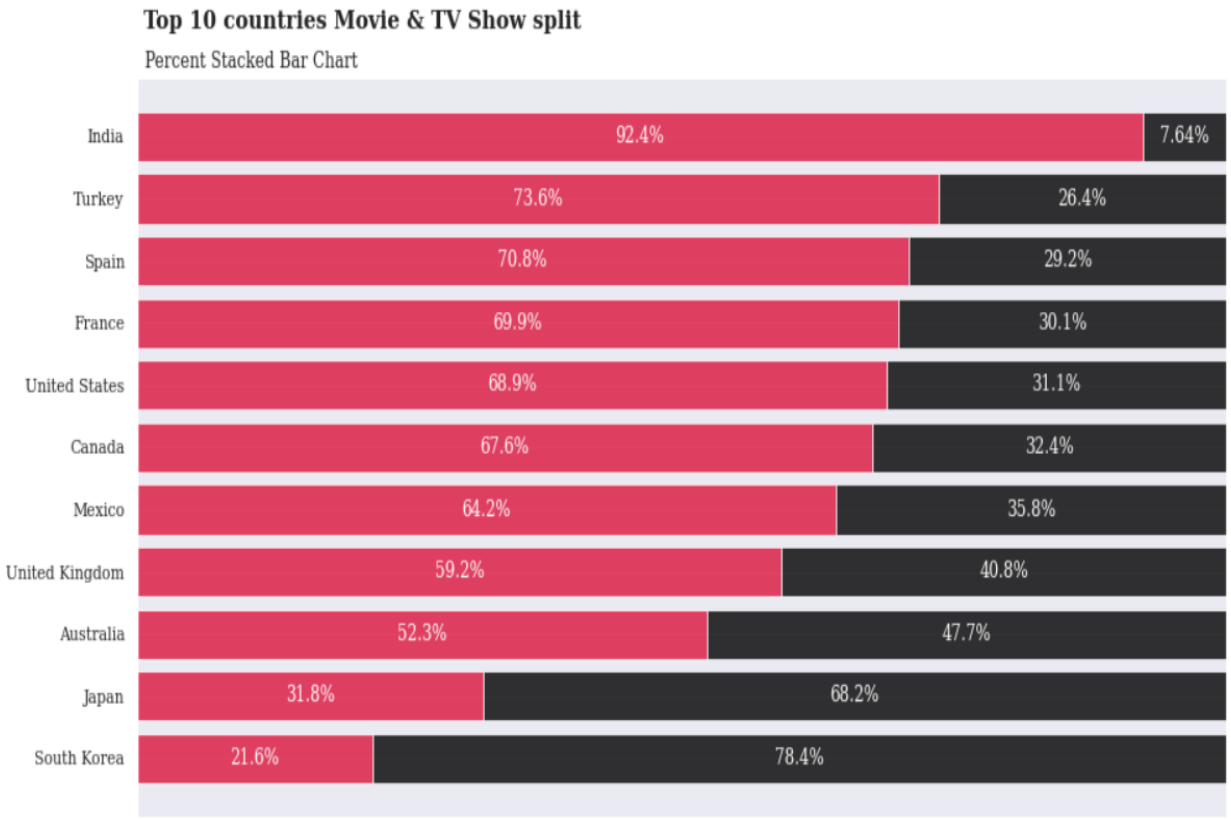
- The 'TV-MA' rating is used in the majority of the film. The TV Parental Guidelines provide a "TV-MA" classification to a television programme that is intended solely for mature audiences.
- The second largest is 'TV-14,' which stands for content that may be inappropriate for minors under the age of 14.
- The third most common is the extremely popular 'R' rating. The Motion Picture Association of America defines an R-rated film as one that contains material that may be inappropriate for children under the age of 17; the MPAA states that "Under 17 requires accompanying parent or adult guardian."

Growth in content over the years



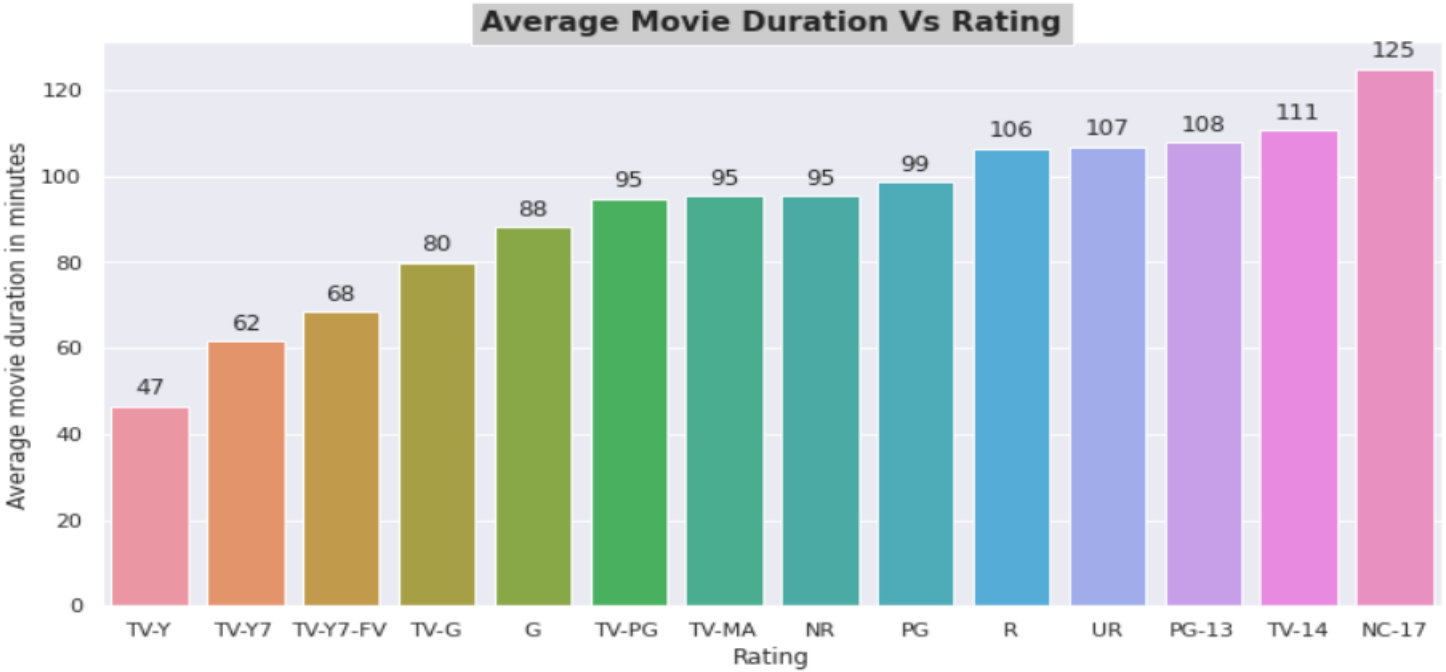
- The number of movies on Netflix is growing significantly faster than the number of TV shows.
- In both 2018 and 2019, approximately 1200 new movies were added.
- We saw a huge increase in the number of movies and television episodes after 2014.
- It appears that Netflix has focused more attention on increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows.

How does content differ by country in the top ten lists



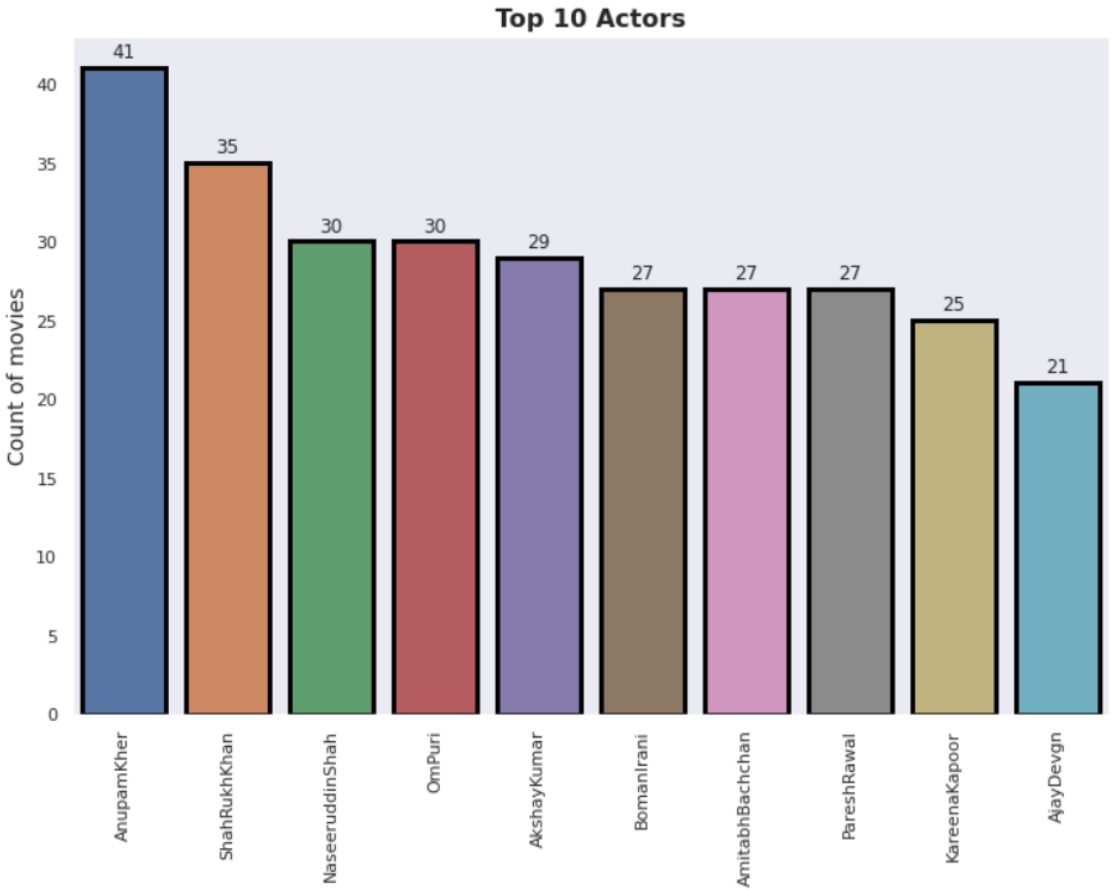
- The majority of the content on Netflix in India is comprised of movies.
- Bollywood is a significant business, and movies, rather than TV shows, may be the industry's major focus.
- South Korean Netflix on the other hand is almost entirely TV Shows.
- The fundamental reason for the variation in content must be due to market research undertaken by Netflix.

Average Movie Duration Vs Rating



- Those movies that have a rating of NC-17 have the longest average duration.
- When it comes to movies having a TV-Y rating, they have the shortest runtime on average.

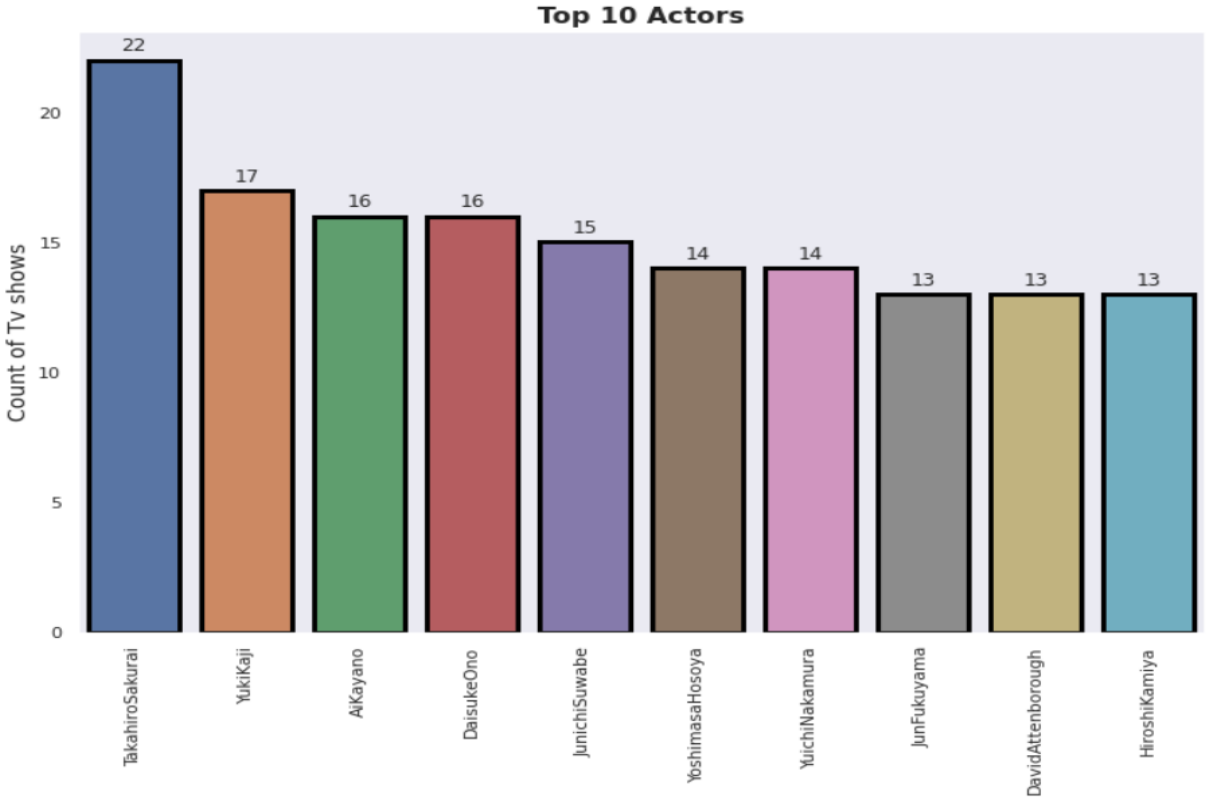
Top 10 Actors who appear in the majority of films



- According to the above barplot, Anupam Kher has worked in over 40 films.
- After Anupam Kher, Shahrukh Khan is ranked second, with 35 films under his belt.
- Naseeruddin Shah and Ompuri have worked in 30 films.

Top 10 Actors who appear in the majority of TV Shows

- According to the above barplot, Takahiro Sakurai has worked in over 20 tv shows.
- After Takahiro Sakurai, Yuki Kaji is ranked second, with 17 tv shows under his belt.
- Aikayano and Daisuke Ono have worked in 16 tv shows.

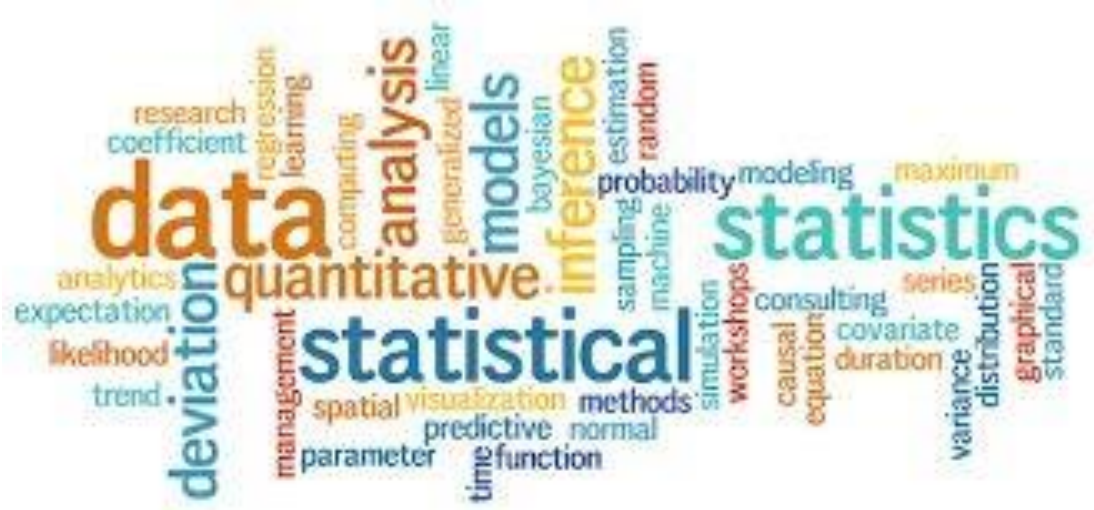


Word Cloud

What Is a Word Cloud?

A word cloud (also known as a tag cloud) is a visual representation of words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide you with quick and simple visual insights that can lead to more in-depth analyses.

Example ⟶



Text Pre-processing for Clustering

1. **Removing Punctuation:**
   - • Punctuations does not carry any meaning in clustering.
   - • So, removing punctuations helps to get rid of unhelpful parts of the data, or noise

2. **Removing Stopwords:**
   - • Stopwords are basically a set of commonly used words in any language, not just in English
   - • If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

3. Stemming:

1. • Stemming is the process of removing a part of a word, or reducing a word to its stem or root.

- • Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

K-Means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre defined distinct non overlapping subgroups where each data point belongs to only one group

1 . Vectorization:

- • Here we have textual data
- • Clustering algorithms cannot understand textual data
- • So, we use vectorization technique to convert textual data to numerical vectors.

So,weusevectorizationtechniquetoconverttextualdatatonumericalvectors.

2. Elbow Curve:

- • The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- • The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

3. Silhouette score :

- • Silhouette score is used to evaluate the quality of clusters created
- using clustering algorithms such as K Means in terms of how well
- samples are clustered with other samples that are similar to each
- other.

Feature Selection & ML algo used

- Only selected 3 features , to do clustering
  - no_of_category
  - Length(description)
  - Length(listed-in)

- Using StandardScaler

- Used 5 algo to find out best k value
  1. Silhouette score
  2. Elbow Method
  3. DBSCAN
  4. Dendrogram
  5. Agglomerative Clustering

1. Silhouette Score

Silhouette Coefficient Formula

$$S = \frac{(b-a)}{max(a,b)}.$$

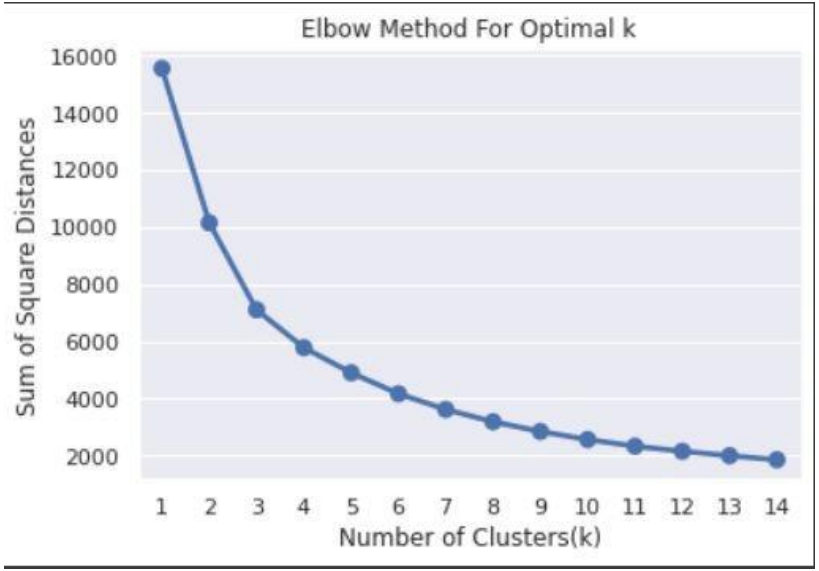
- **mean intra-cluster distance(a)**:- Mean distance between the observation and all other data points in the same cluster.
- **mean nearest-cluster distance (b)** :- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a.

The value of the silhouette coefficient is between [-1, 1]

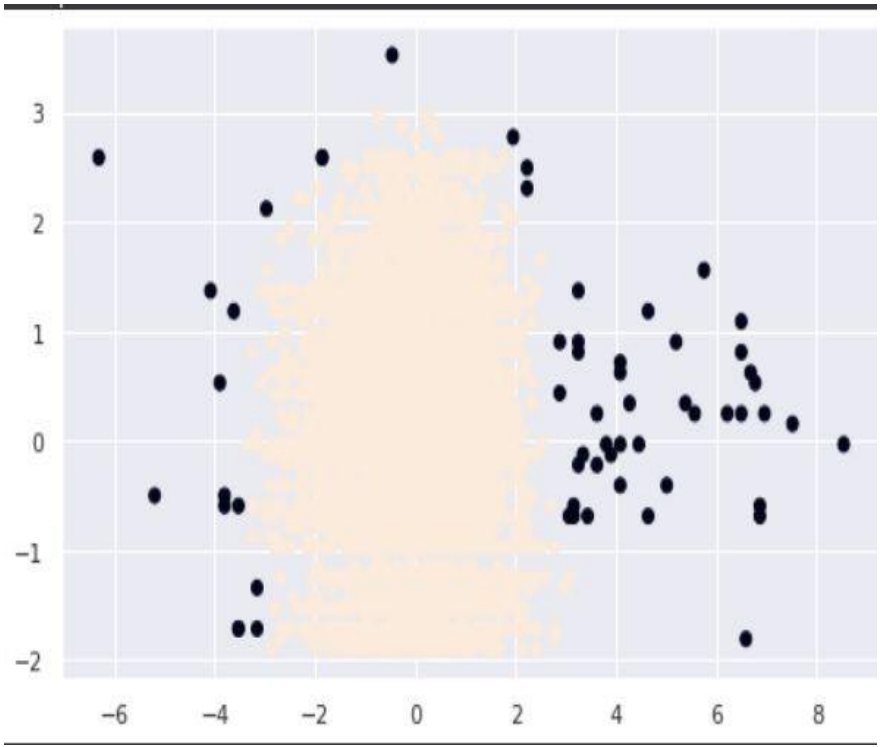
- If score is **1 denotes the best** meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- *The worst value is -1*
- If score is 0 denotes overlapping clusters

	n clusters	silhouette score
1	3	0.348
0	2	0.337
12	14	0.332
5	7	0.330
11	13	0.329
10	12	0.328
13	15	0.326
9	11	0.324
8	10	0.323
7	9	0.322
2	4	0.320
4	6	0.320
6	8	0.316
3	5	0.308

2. Elbow Method

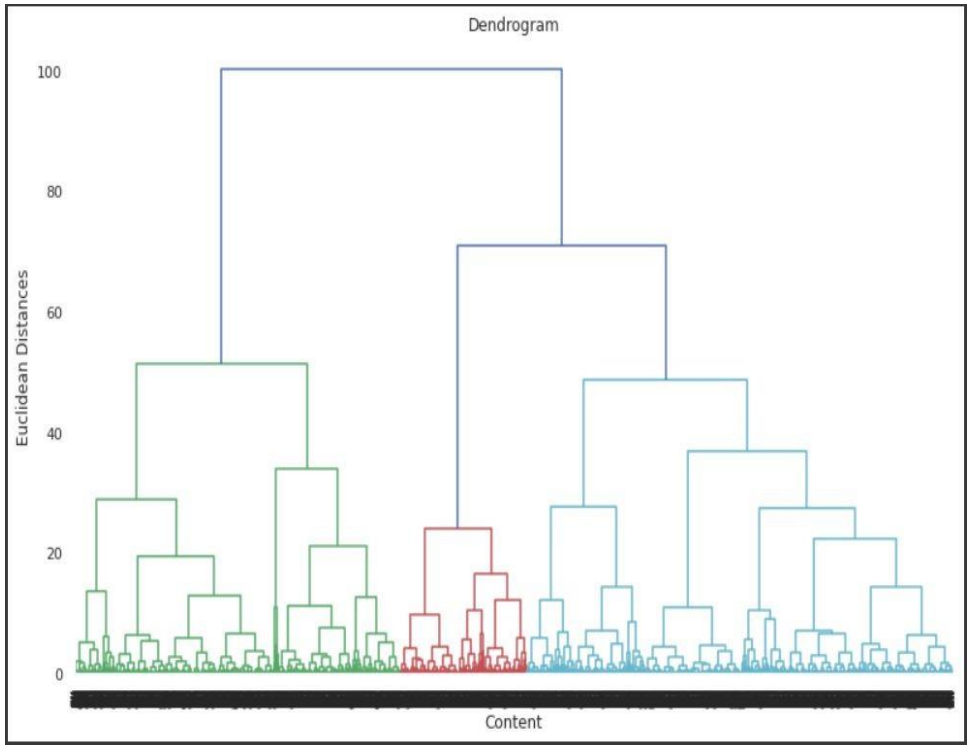


3 . DBSCAN



DBSCAN

4. Dendrogram



Dendrogram

Conclusion

- Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation
- We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies)
- By analysing the content added over years we get to know that in recent years netflix is focusing movies than TV shows (movies is increased by 80% and TV shows is increased by 73% compare to 2016 data)
- The most number of the movies and TV shows release in 2017 and 2020 respectively and united nation have the maximum content on netflix

- On Netflix, Dramas genre contains the maximum content among all of the genres and the most of the content added in december month and less content in february
- By applying the silhouette score method for n range clusters on dataset we got best score which is 0.348 for 3 clusters it means content explained well on their own clusters, by using elbow method after k = 3 curve gets linear it means k = 3 will be the best cluster
- Applied different clustering models Kmeans, hierarchical, Agglomerative clustering on data we got the best cluster arrangments
- By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 3

Thank you