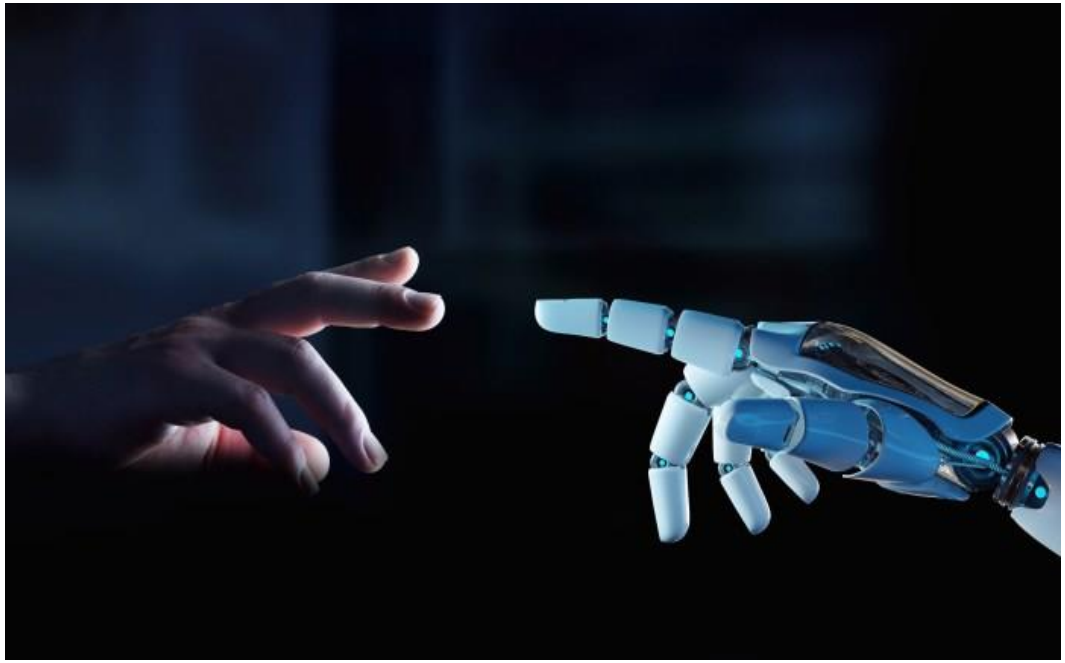
Capstone Project-4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Team members:
Prince Jain
Rishabh Patidar
Vikas Shrivas

Point in touch:

- Introduction
- Problem Statement
- Data Description
- Null Value
- Exploratory Data Analysis
- Data Cleaning
- Model Implementation
- K- Means
- Cluster Analysis



Introduction

Netflix: Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time.

Methodology:

- Unsupervised Machine Learning (Clustering)

Database:

- Netflix Movies and TV Shows
- 7787 rows and 12 columns
- Data from last decade

Problem Statement

- 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.
- 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.
- Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

In this project, you are required to do

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

Data Description

The data was collected from Flixable which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 rows of data.
The dataset consists of eleven textual columns and one numeric column.

Attribute Information :

1. **show_id :** Unique ID for every Movie / Tv Show
2. **type :** Identifier - A Movie or TV Show
3. **title :** Title of the Movie / Tv Show
4. **director :** Director of the Movie
5. **cast :** Actors involved in the movie / show
6. **country :** Country where the movie / show was produced
7. **date_added :** Date it was added on Netflix
8. **release_year :** Actual Release year of the movie / show
9. **rating :** TV Rating of the movie / show
10. **duration :** Total Duration - in minutes or number of seasons
11. **listed_in :** Genre
12. **description:** The Summary description

Null Value

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.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64 
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

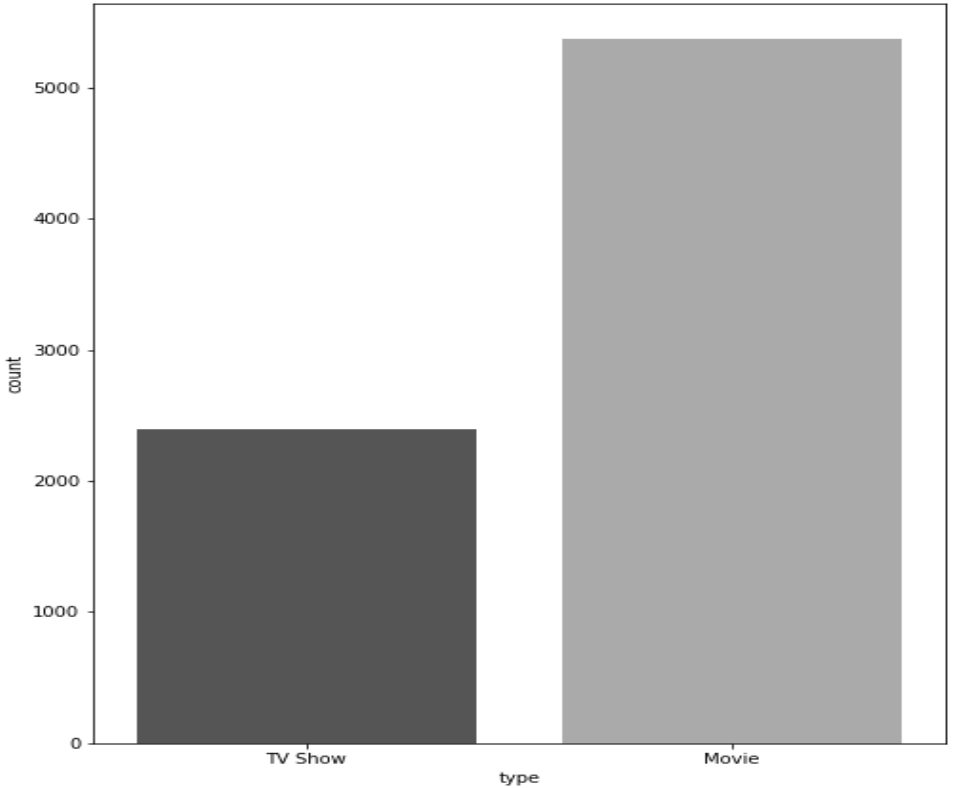

```
show_id          0
type             0
title            0
director      2389
cast           718
country        507
date_added      10
release_year     0
rating           7
duration         0
listed_in        0
description      0
dtype: int64
```

Exploratory Data Analysis

Type of content available on Netflix

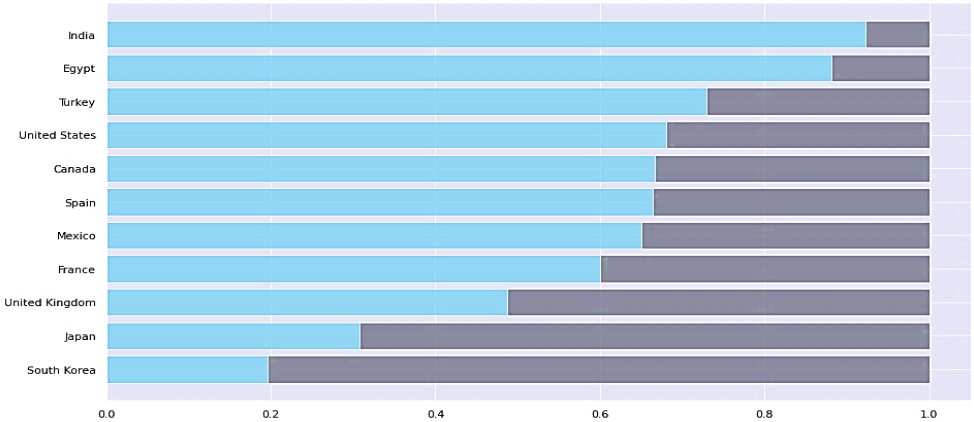
- It is evident that there are more movies on Netflix than TV shows.
- Netflix has 5372 movies, which is more than double the quantity of TV shows.

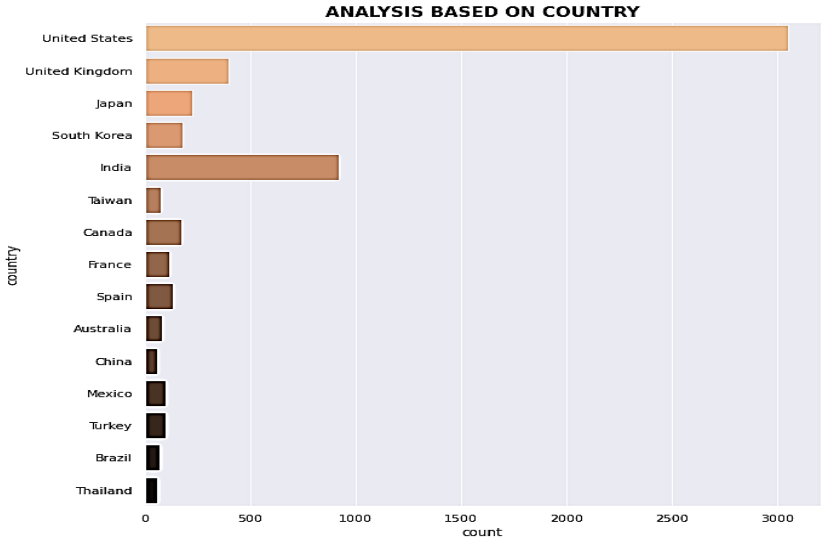
```
Movie      5372
TV Show    2398
Name: type, dtype: int64
```



Exploratory Data Analysis

Analysis based on country:



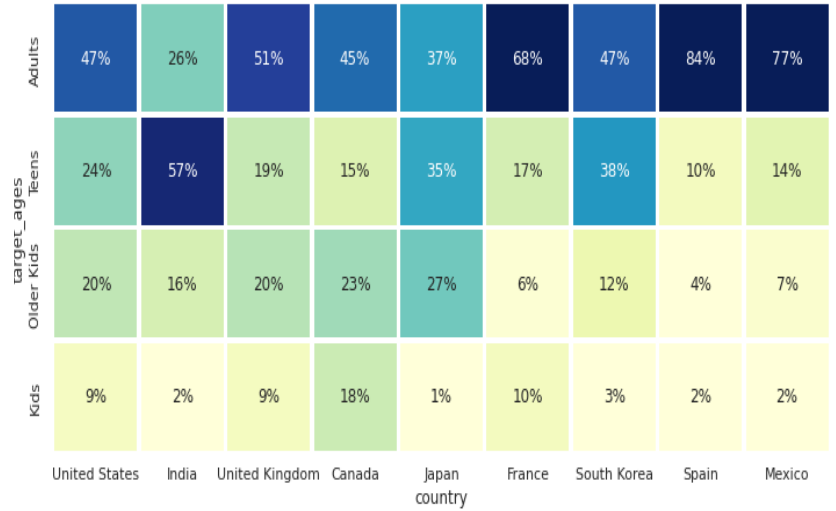


- India is the country having maximum numbers of movie on Netflix.
- **USA having maximum numbers of TV shows followed by India**

Exploratory Data Analysis

Analysis based on country: Heatmap

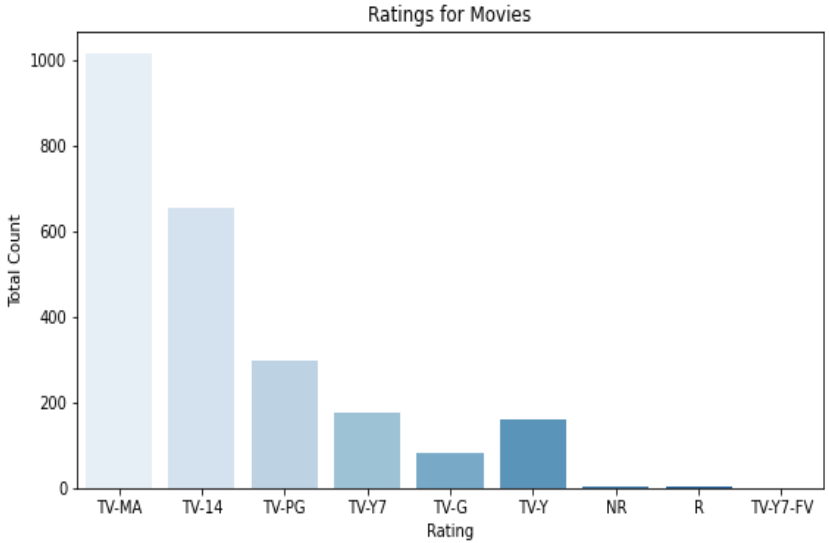
```
United States                                                  3051
India                                                           923
United Kingdom                                                  396
Japan                                                           224
South Korea                                                     183
                                                                ...
Russia, United States, China                                      1
Italy, Switzerland, France, Germany                               1
United States, United Kingdom, Canada                             1
United States, United Kingdom, Japan                              1
Sweden, Czech Republic, United Kingdom, Denmark, Netherlands      1
Name: country, Length: 681, dtype: int64
```

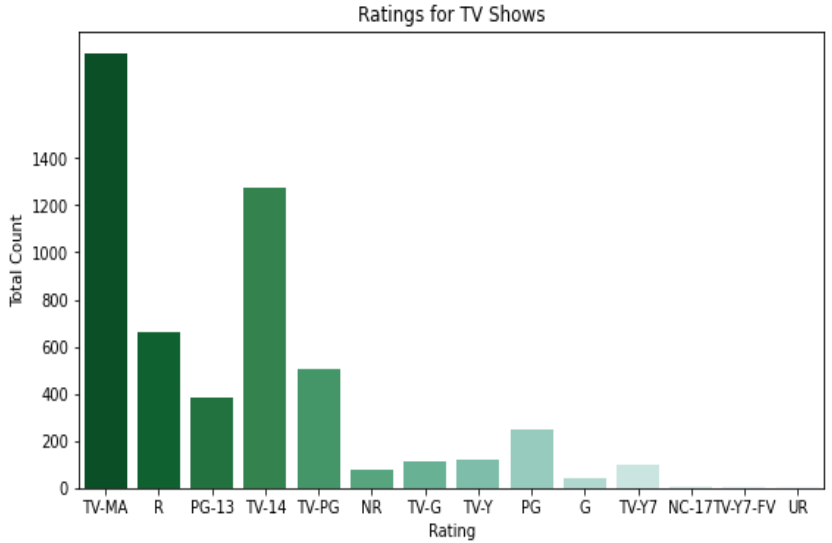


- **We group data into Four category and find that there are 57% uses are teens in India which maximum teens percentage.**
  In Adults user Spain have 84% followed my Mexico which is 77%.

Exploratory Data Analysis

According to rating :

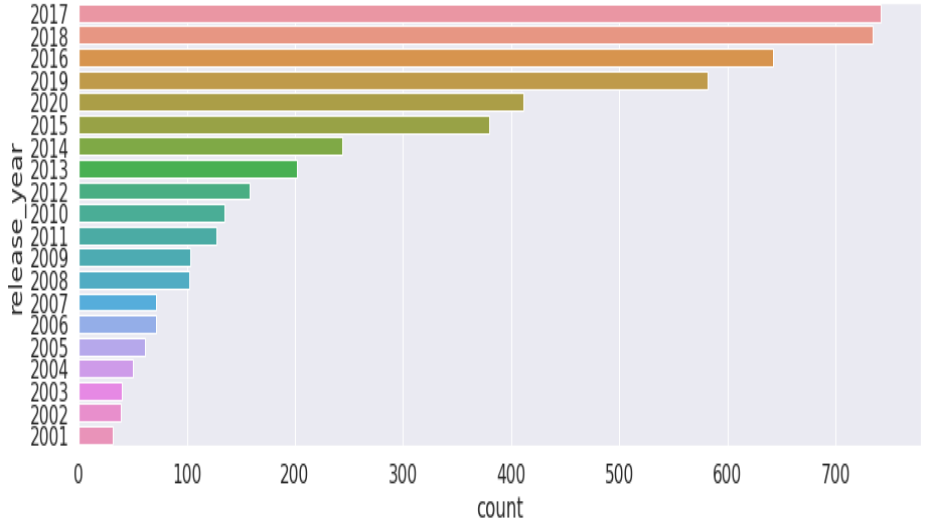




TV-MA has the highest number of ratings
for tv shows i.e. adult ratings

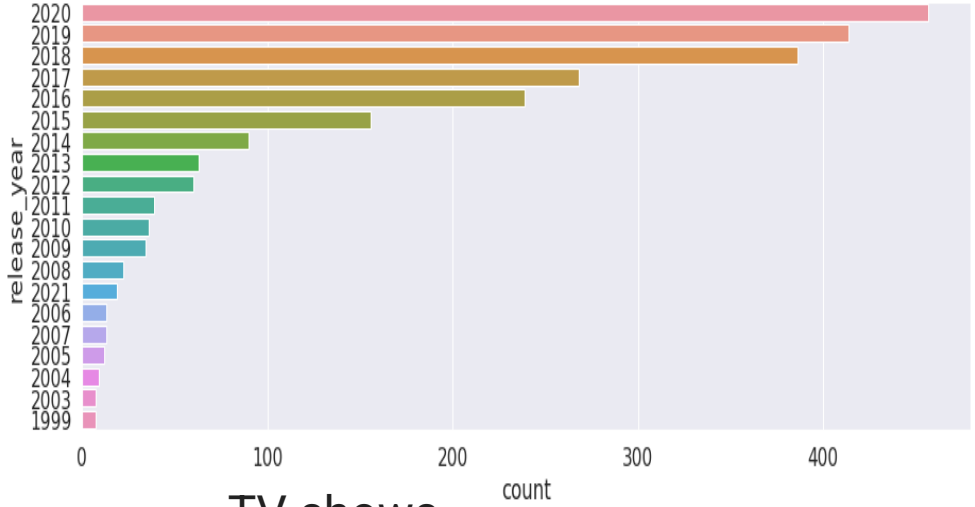
Exploratory Data Analysis

Releases over the year



Movies

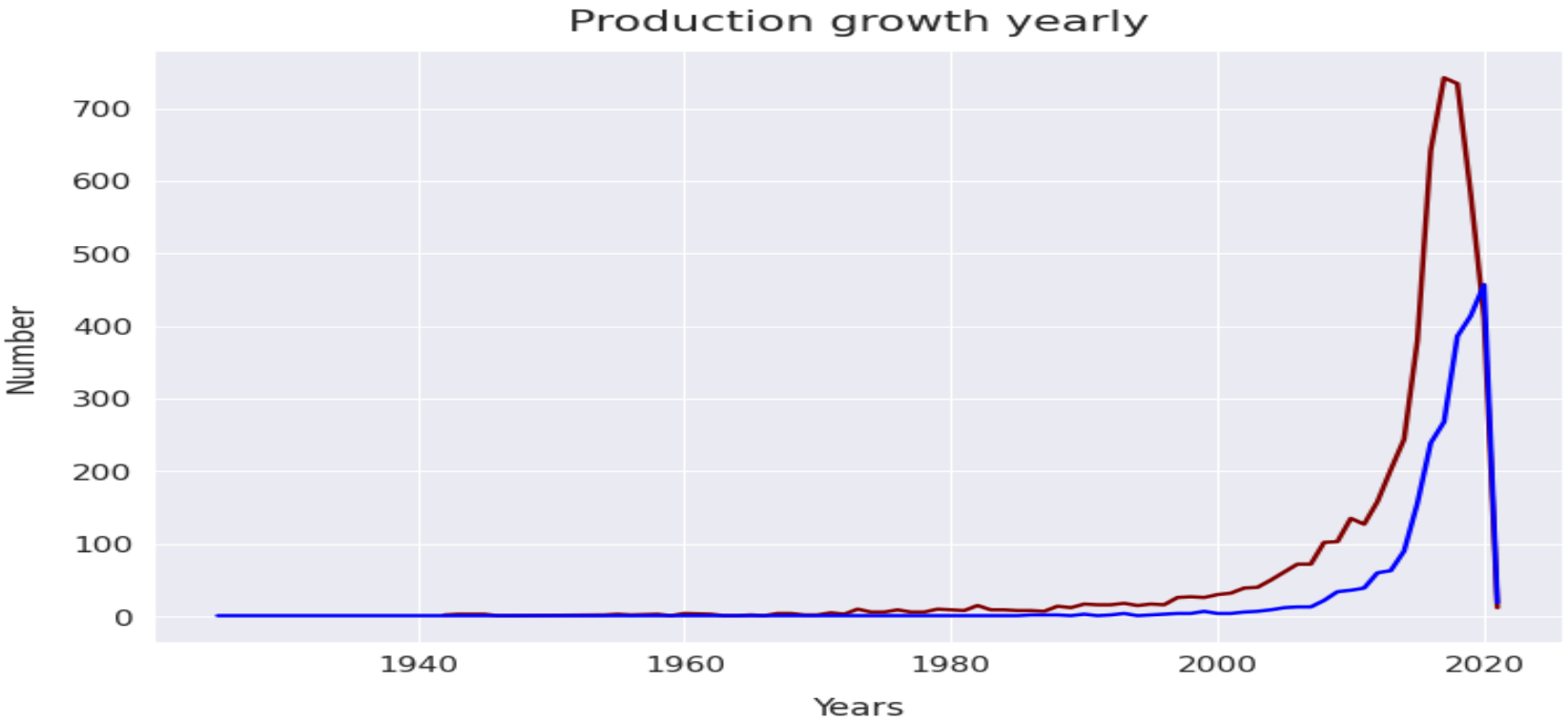
Highest number of movies released in 2017 and 2018



TV shows.

- Highest number of movies released in 2020.
- The number of movies on Netflix is growing significantly faster than the number of TV shows.
- We saw a huge increase in the number of movies and television episodes after 2015.
- There is a significant drop in the number of movies and television episodes produced after 2020.

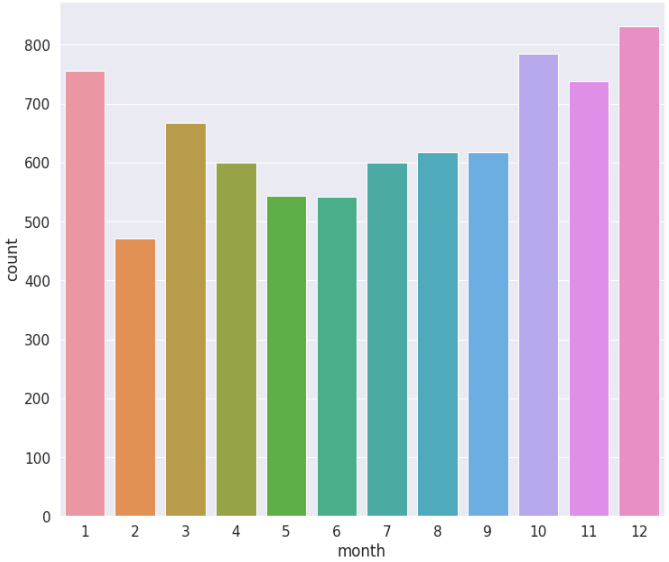
Exploratory Data Analysis



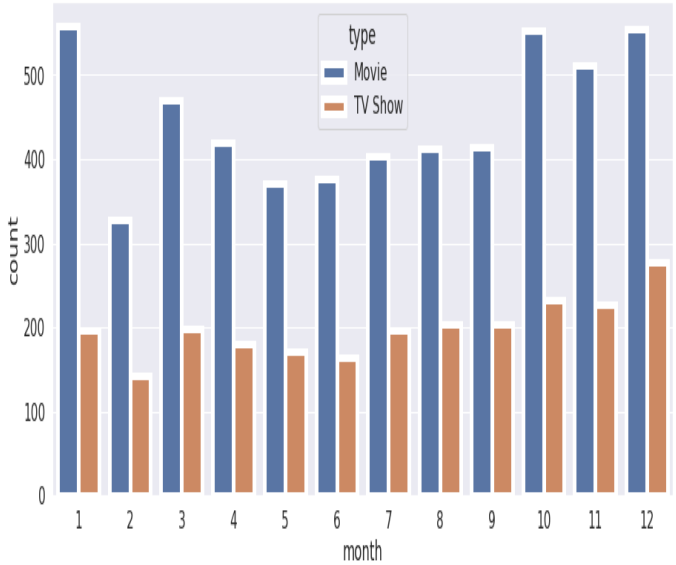
The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19

Exploratory Data Analysis

Month wise Analysis:

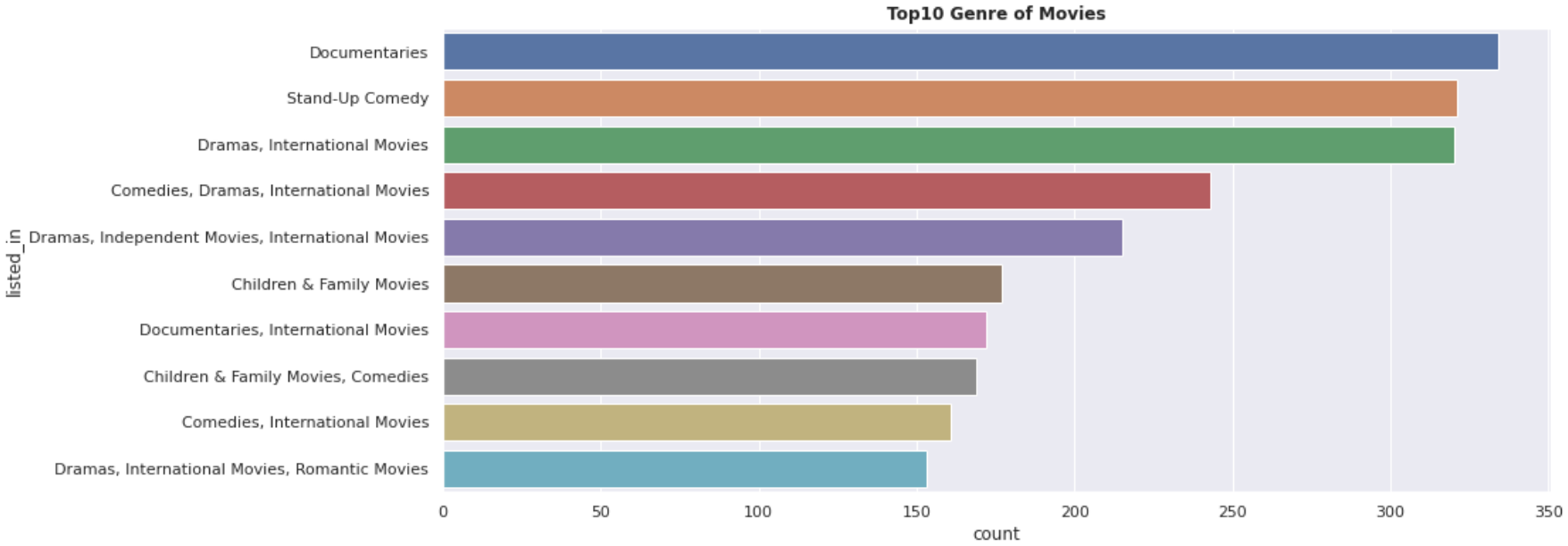


	month	count
5	January	745
4	February	465
8	March	660
1	April	595
9	May	537
7	June	538
6	July	592
2	August	611
12	September	613
11	October	780
10	November	730
3	December	816



From October to January, maximum number of movies and TV shows were added.

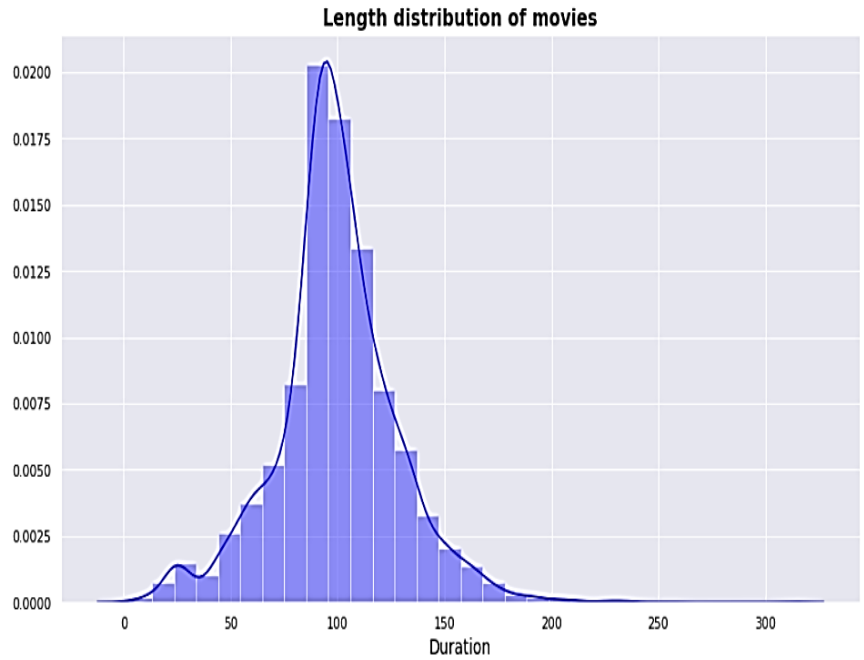
Exploratory Data Analysis



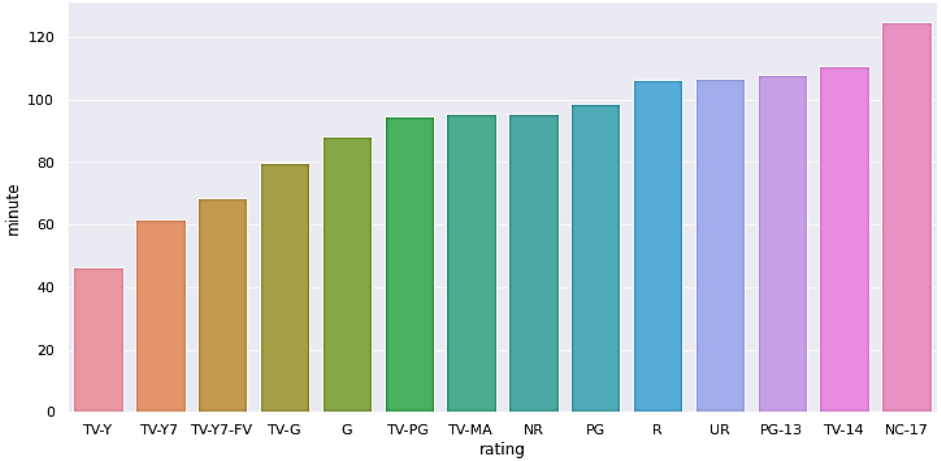
Documentaries are highest in numbers
followed by **standup comedy** then **Dramas**

Exploratory Data Analysis

Duration distribution of Movies



- Maximum movies duration have 60min to 80 min.

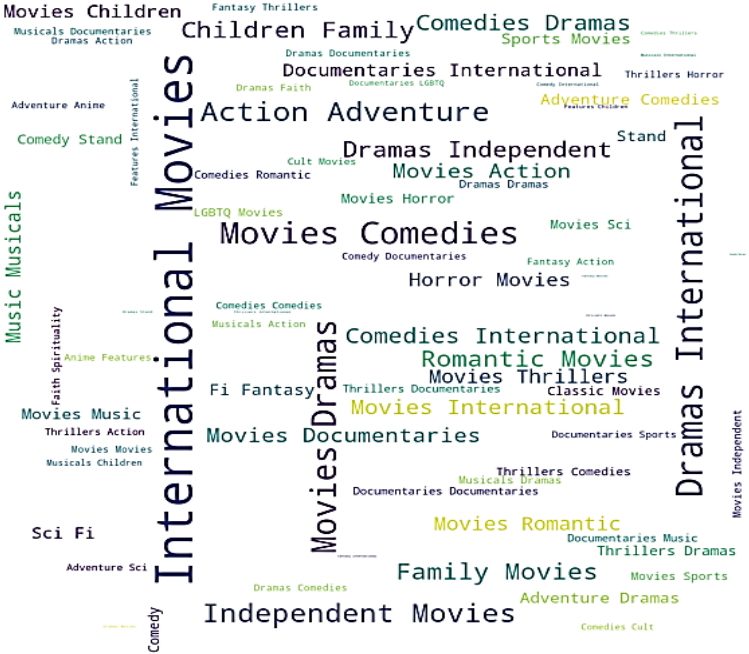


- Those movies that have a rating of NC-17 have the longest average duration.
- When it comes to movies having a TV-Y rating, they have the shortest runtime on average

Word Cloud for TV shows and Cast

TV Shows types

Which Cast to Choose?



Data Cleaning

- Label Encoding

- Lemmatisation- Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. ... For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.

- Removing Stop words - To remove stop words from a sentence, you can divide your text into words and then remove the word if it exits in the list of stop words provided by NLTK.

- Min-max Scaling - For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It preserves shape of original distribution.

K - Means

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved

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. Elbow Curve:

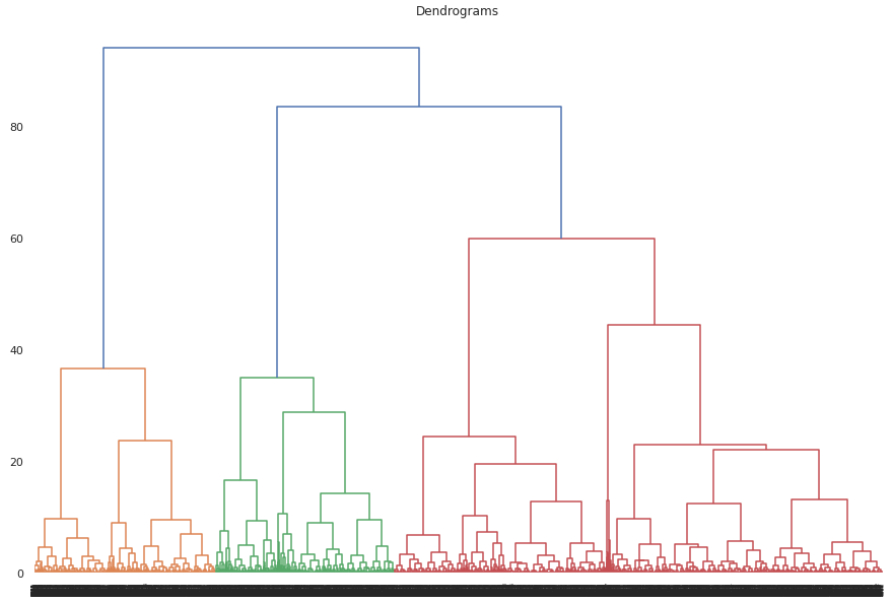
- The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- The elbow curve uses the sum of squared distance (SSE)to choose an ideal value of k based on the distance between the data points and their assigned clusters.

2. 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.

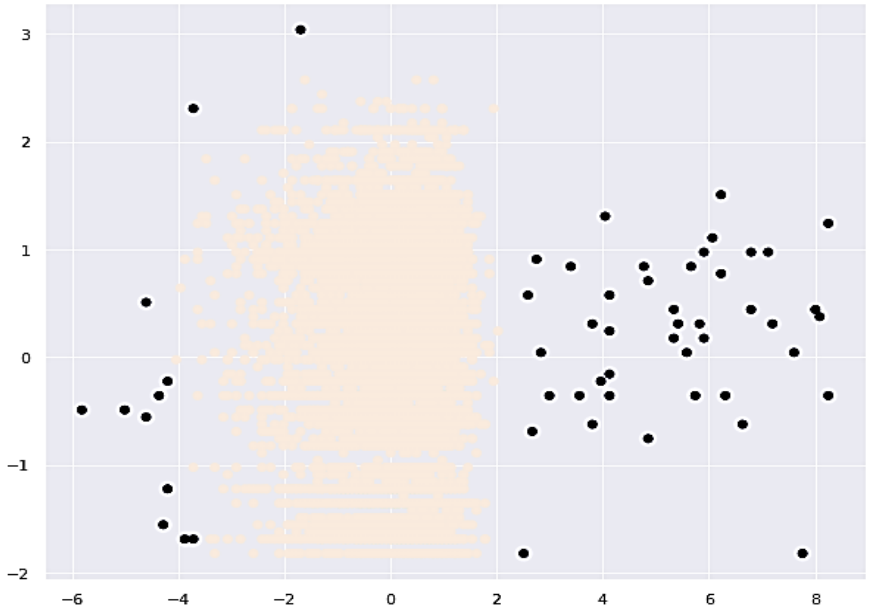
Model Implementation

1. Hierarchical Clustering



Hierarchical clustering is the most popular and widely used method **to analyze social network data**. In this method, nodes are compared with one another based on their similarity.

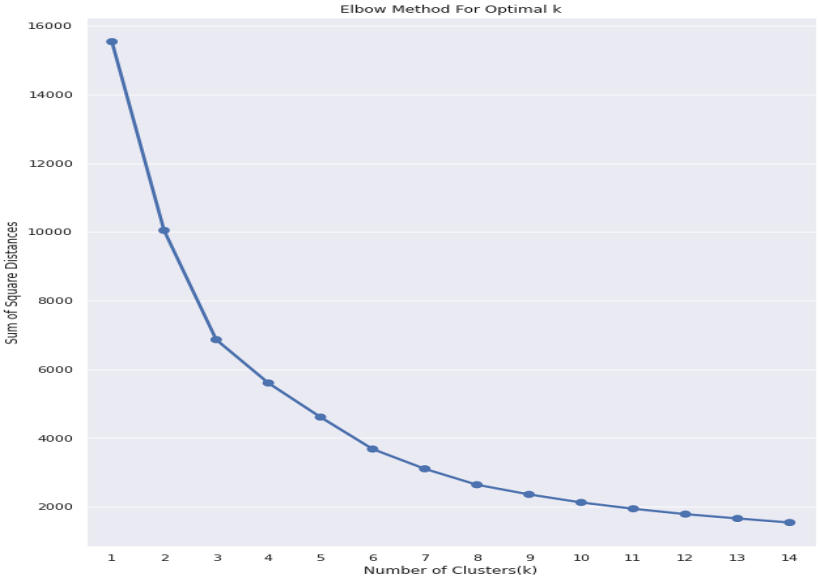
2. DBSCAN



DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

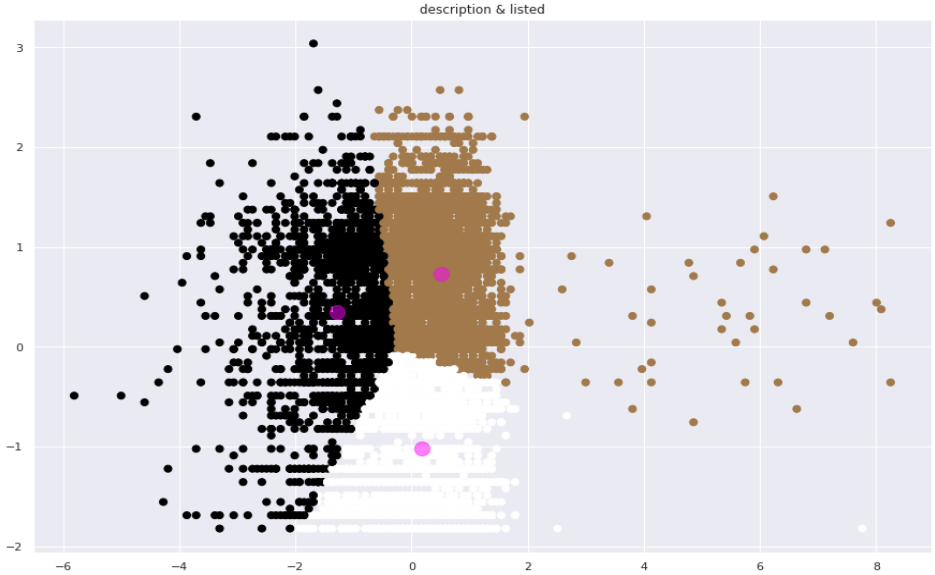
Model Implementation

3. Elbow Method



In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set.

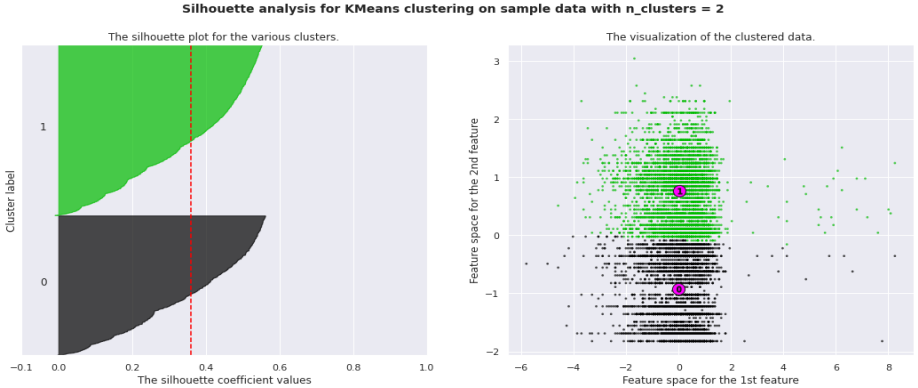
4. K-means Clustering unsupervised ML algorithm

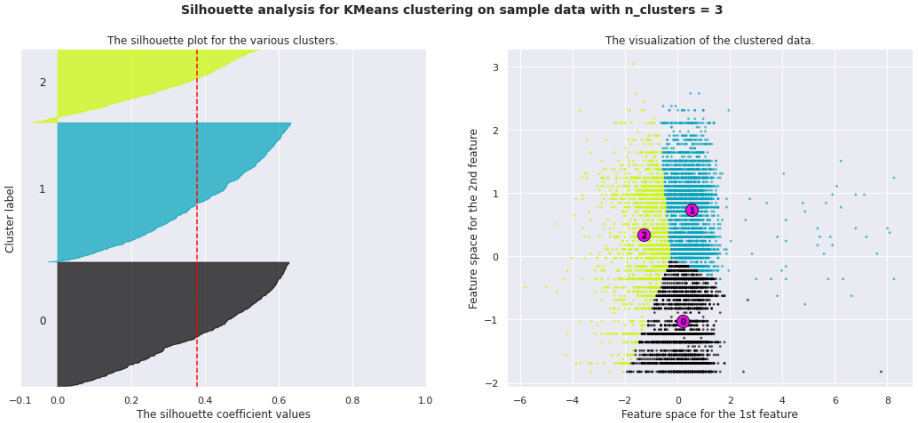


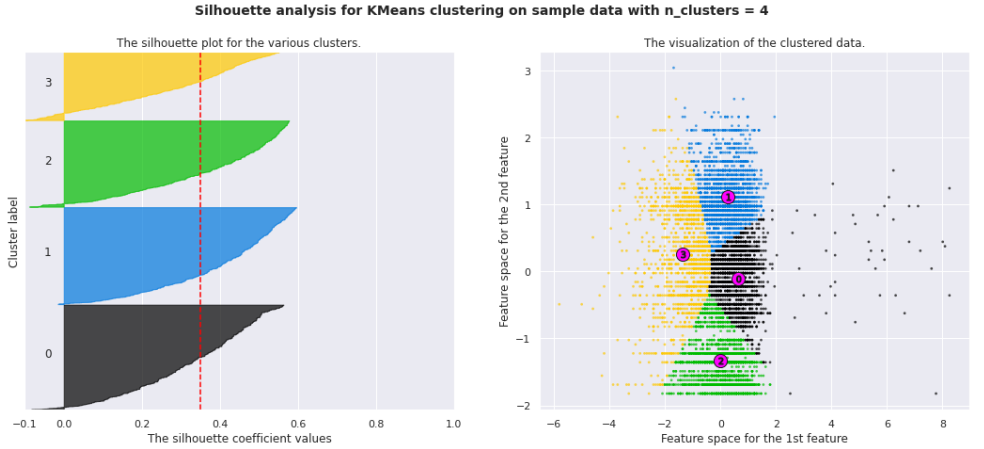
K-means clustering is an unsupervised technique that requires no labeled response for the given input data.

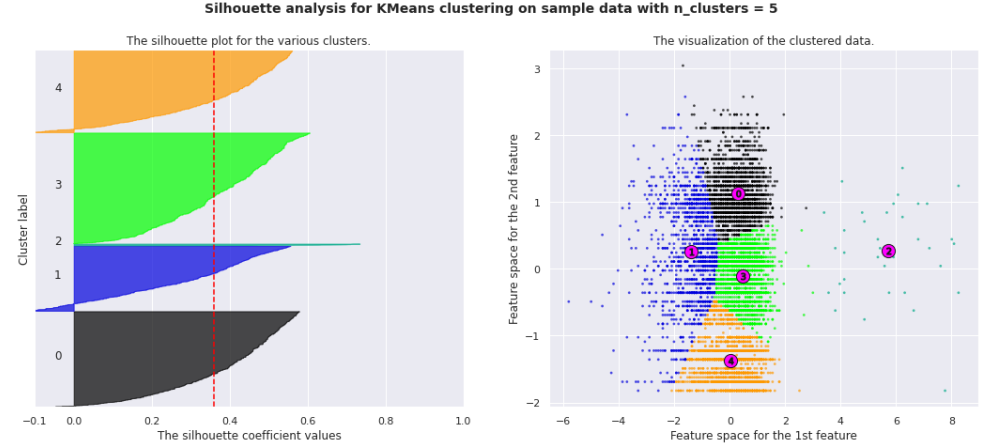
Model Implementation

3. k-means clustering



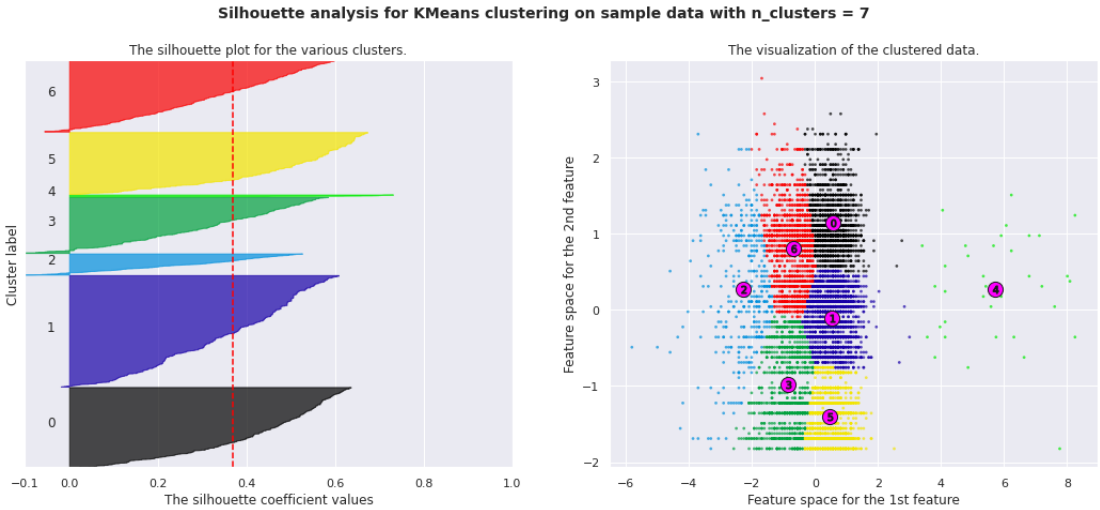






Model Implementation

3. k-means clustering



- Here is the Silhouette analysis done on the above plots to select an optimal value for n_clusters.
- The value of 6 and 7 for n_clusters looks to be the optimal one. The silhouette score for each cluster is above average silhouette scores.

For n_clusters = 2 The average silhouette_score is 0.3601761941410064

For n_clusters = 3 The average silhouette_score is : 0.3786424530063482

For n_clusters = 4 The average silhouette_score is : 0.34764733391208535

For n_clusters = 5 The average silhouette_score is 0.35819475414428864

For n_clusters = 6 The average silhouette_score is 0.36830181782871185

For n_clusters = 7 The average silhouette_score is 0.36688539987717955

Conclusion

- Netflix has 5372 movies and 2398 TV shows, there are more number movies on Netflix than TV shows.
- TV-MA has the highest number of ratings for tv shows i.e. **adult ratings**
- Highest number of **movies** released in **2017 and 2018.**
- We saw a huge increase in the number of movies and television episodes after 2015.
- There is a significant drop in the number of movies and television episodes produced after 2020.
- It appears that Netflix has focused more attention on increasing Movie content that TV Shows. Movies have increased much more dramatically than TV shows.
- **India is the country having maximum numbers of movie on Netflix.**
- **USA having maximum numbers of TV shows followed by India.**
- 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
- **October to January, maximum number of movies and TV shows were added**.
- The USA, India, the United Kingdom, Canada, and Egypt are the top five producer countries.

Conclusion

KMeans cluster score

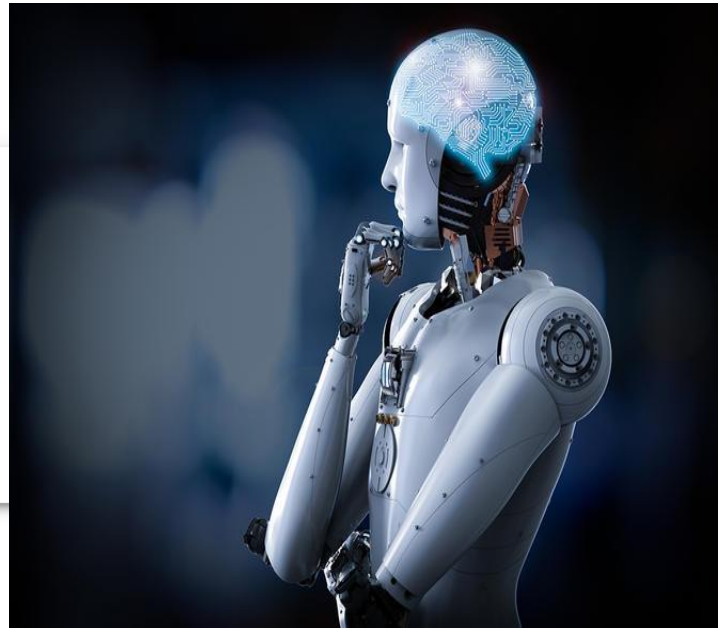
n_clusters = 2 average silhouette_score : 0.3601761941410064

n_clusters = 3 average silhouette_score : 0.3786424530063482

n_clusters = 4 average silhouette_score : 0.3490757756268031

n_clusters = 5 average silhouette_score : 0.3601906332511891

n_clusters = 6 average silhouette_score : 0.36897493498221406



K-means Clustering unsupervised ML algorithm

- 1-2992
- 2-2538
- 3-2240

Elbow Method For Optimal k as the numbers of clusters increases the Sum of squares distance decreases.

THANKS FOR WATCHING

S E E Y A .

AND NOW GO CHILL ON NETFLIX

Time for Q&A!!