Capstone Project

NETFLIX-MOVIE-AND-TV-SHOWS-CLUSTERING

Team Members

Mohammad Jibran
Siddhi Thakur

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 increasingly focused on TV rather than movies in recent years?
4. Clustering similar content by matching text-based features.

Problem Statement.

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

Table of Content

1. Introduction
2. Define Problem Statement.
3. Data Preview
4. Dataset summary.
5. Data Cleaning & Data Visualization
6. Exploratory Data Analysis.
7. Feature Selection
8. Applying different clustering methods
9. Applying Clustering Models
10. Conclusion

Introduction

Netflix is an American subscription streaming service and production company, Founded on August 29, 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California.

A media distribution company is Netflix. It began with mail-order DVD distribution but has since undergone a significant evolution. Netflix focuses on streaming video these days. While part of its content is created internally, some of it is licenced.

It is the second largest entertainment/media company by market capitalization as of February 2022.

Although Netflix's primary concentration was on movies, the more popular format today is clearly television series. With a paid subscription, users of Netflix have unrestricted access to all of its material.

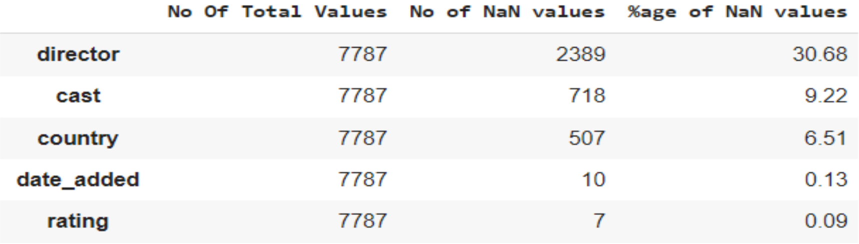
Data Preview

The dataset was taken through the third-party Netflix search engine flixable.This dataset includes of Netflix-eligible television series and motion pictures as of 2019.

- 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

Dataset summary

- Netflix dataset contain 12 columns and 7787 rows.
- In some columns of our datasets have exits Null values.these are the columns shown below.

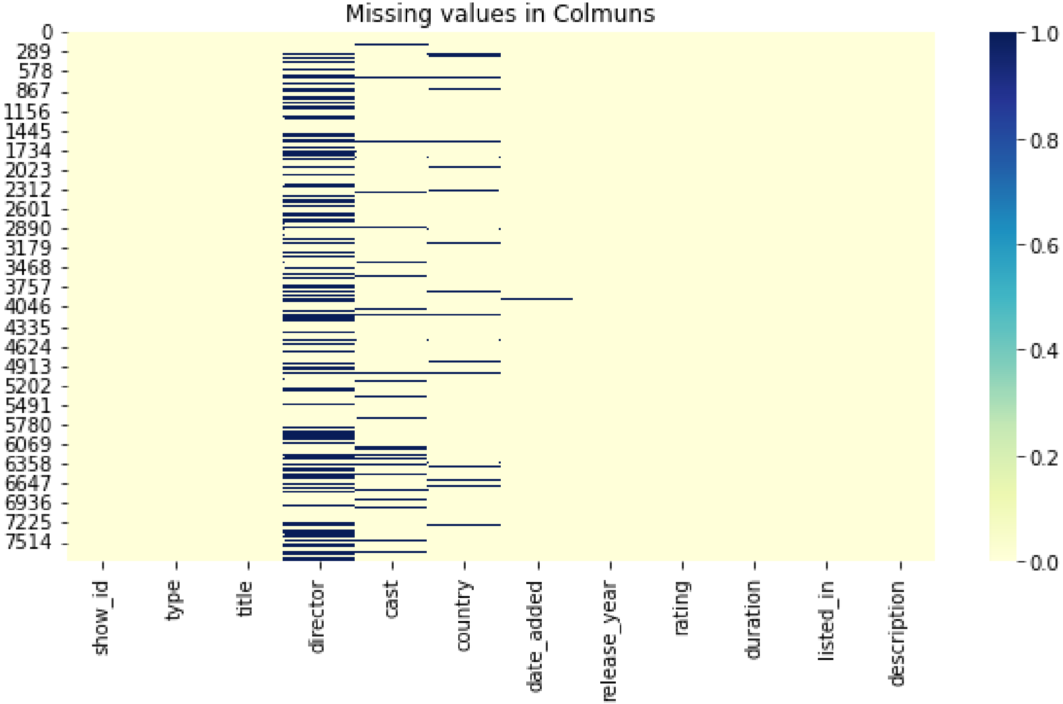


	No Of Total Values	No of NaN values	%age of NaN values
director	7787	2389	30.68
cast	7787	718	9.22
country	7787	507	6.51
date_added	7787	10	0.13
rating	7787	7	0.09

- Since there are very few null values in rating column,so we drop that rows from datasets.
- Finally we will do some feature engineering to create few new variables: Compute year, month and day from date_added after converting it into datetime variable.and then drop date_added column.
- Finally Netflix dataset contain 15 columns and 7780 rows.

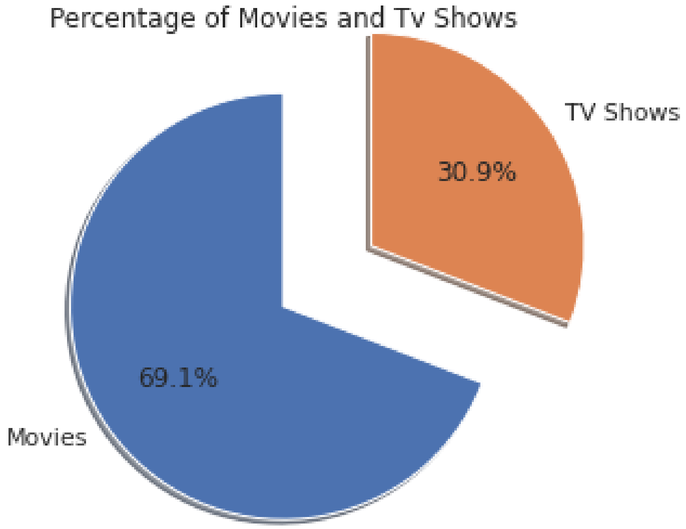
Data Cleaning & Data Visualization

This heatmap revel that columns director,cast,country are contain highly Null values,it can seen that very few null values in rating,date_added columns.and rest all are field with data.



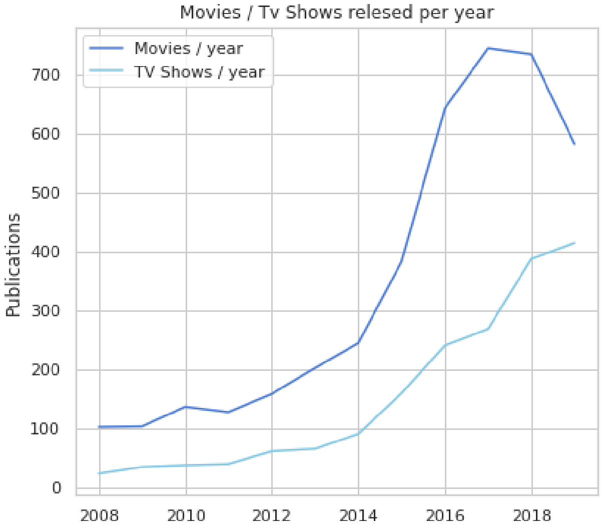
What type of content is available on Netflix.

This Pie plot shows, 69.1% of the content available on Netflix are movies; the remaining 30.9% are TV Shows.

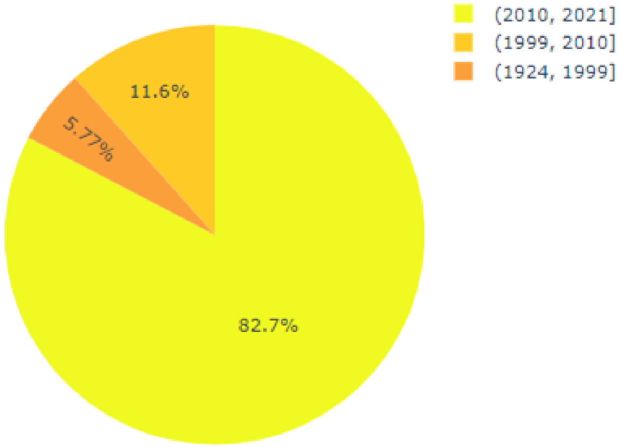


How many Movie/TV shows are released per year.

The lineplot reveals that released of movie and tv-shows per year.since 2014 to 2019 the releasing of movie is greater than TV-shows.



In Which year most of the content was added.



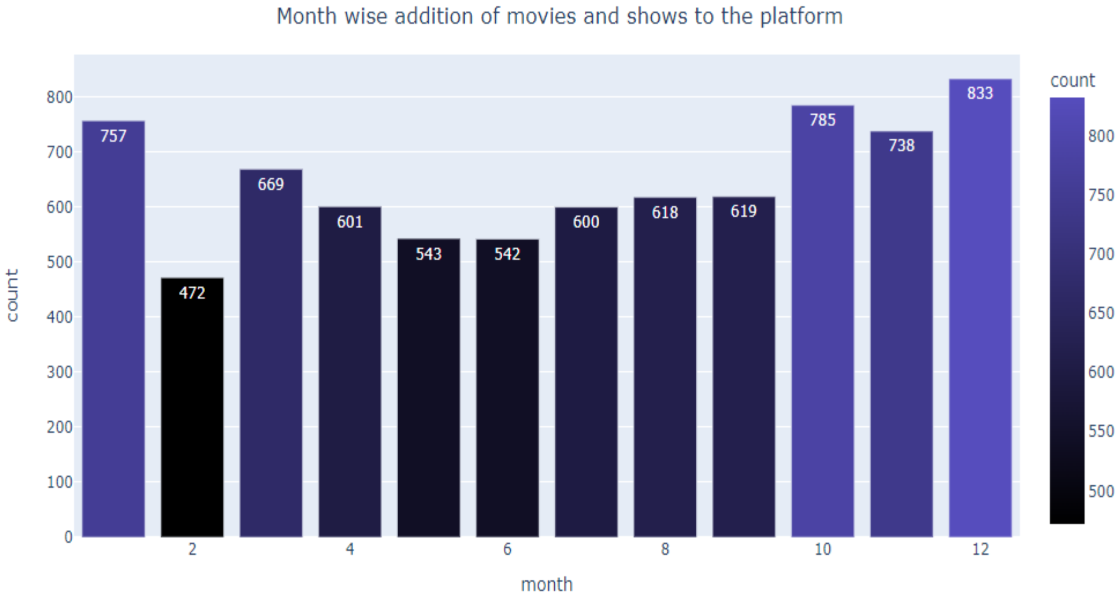
We can see that the majority of the content available was between 2010 and 2021 that is 82.7%.

Since 1924 to 2010,17.28% of the content available was released.

In which months do most movies and TV shows are added.

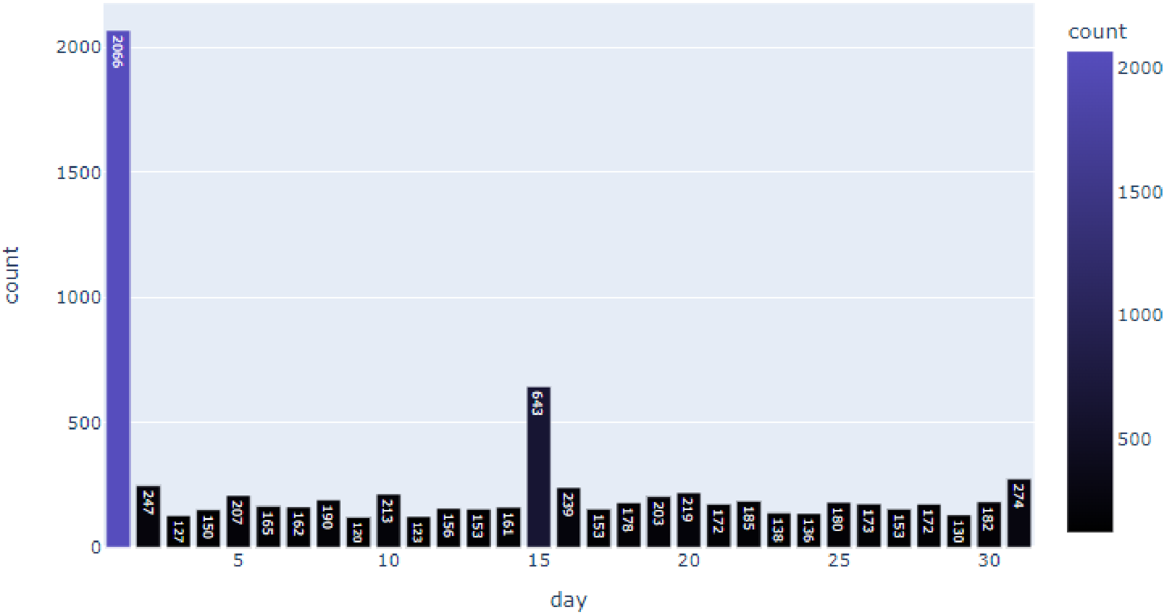
This bar graph shows most of the content is added either by year ending or beginning.

As we see the highlighted bars are October, November, December, and January are months in which many shows and movies get uploaded to the platform.



Which Days are Outstanding.

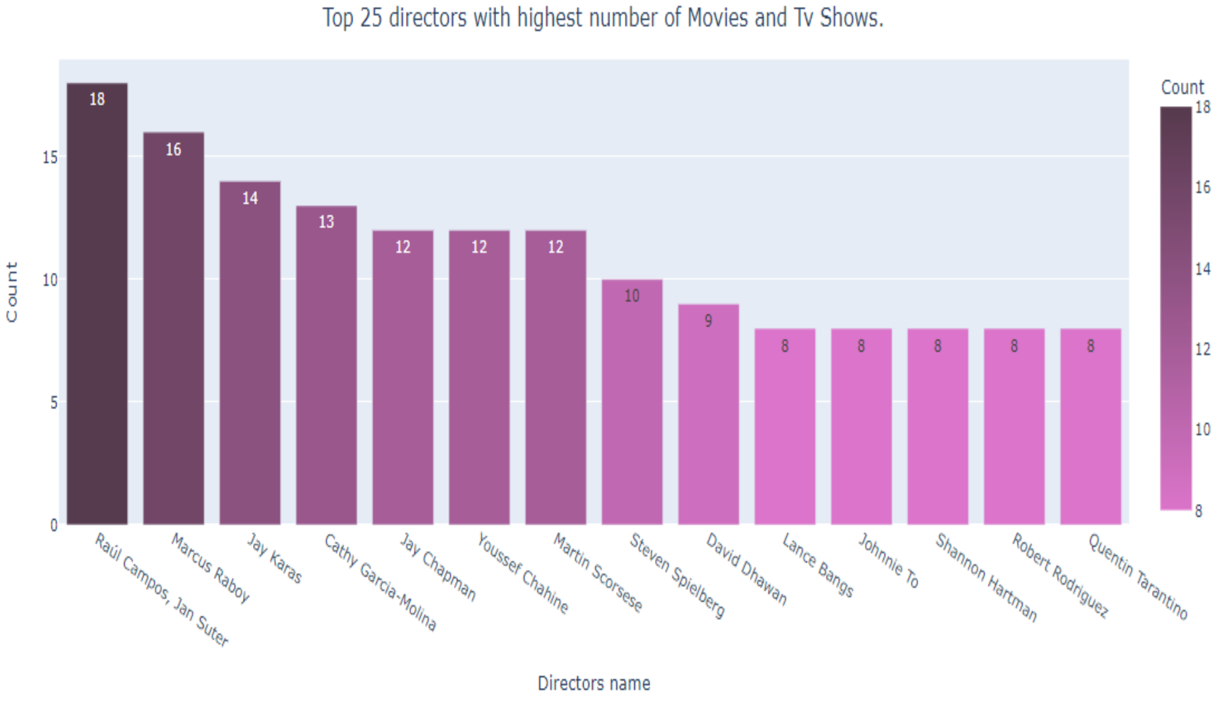
Day wise addition of movies and shows to the platform



This bar graph shows Most of the content is uploaded at the beginning, middle, or the end of a month.

Which makes 1st, 15th or 31st of a month more outstanding in getting new tv shows and movies.

Which director has directed the most movies and TV shows?

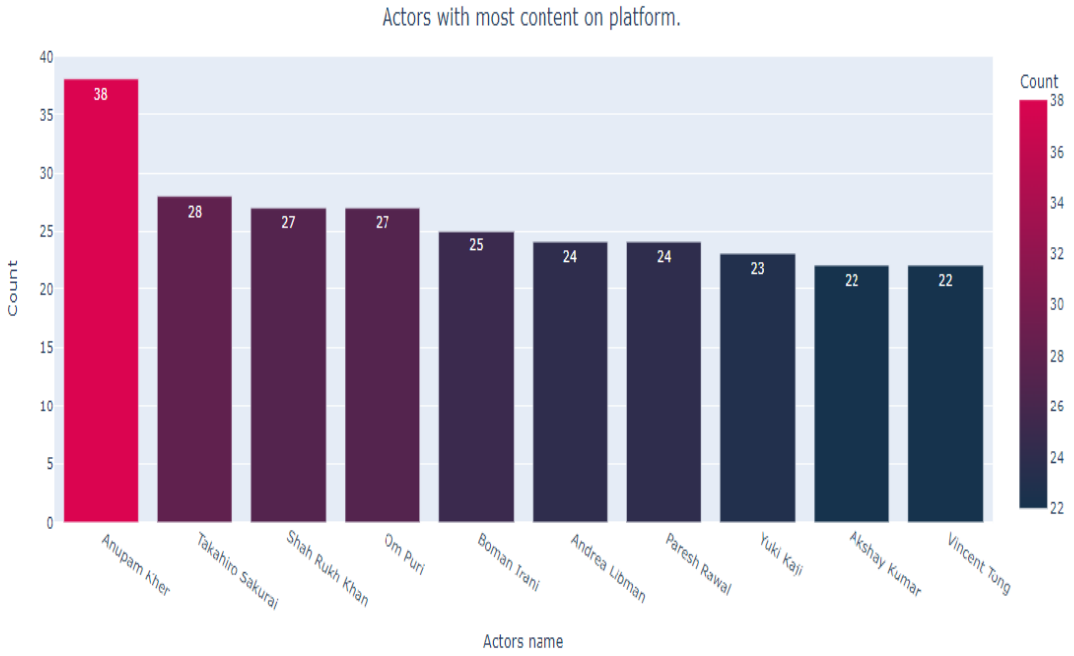


This graph shares a report of Top 5 directors who direct the majority of movies and TV-shows.

Raúl Campos, Jan Suter, Marcus Raboy, Jay Karas, Cathy Garcia-Molina, Jay Chapman are the top 5 directors.

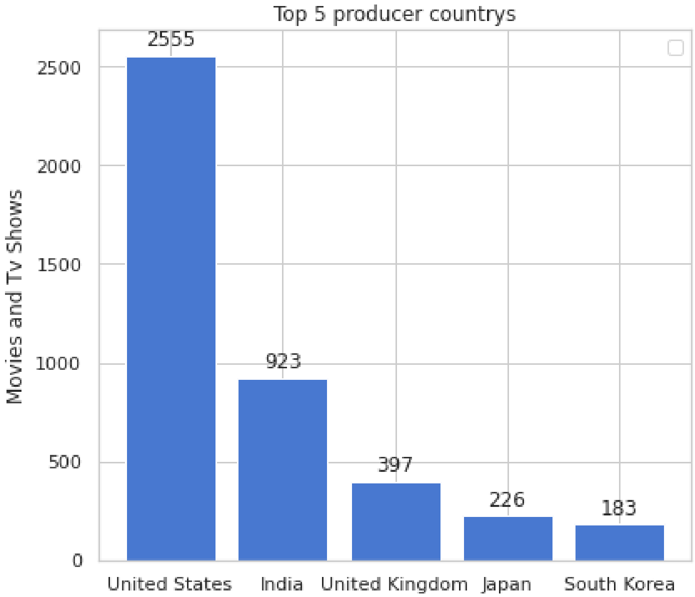
Which Actor/Actress have been cast in most of the movies and TV shows?.

- Anupam Kher is ranked first with 38 overall appearances in TV shows and films.
- Proud to see Six other Indian actors round up the top ten list.

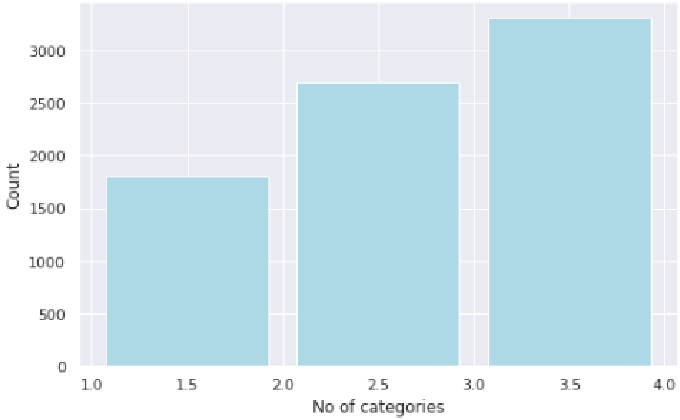


Countries producing most no of contents.

As we can see in the bar graph, the United States has the most content and then India and so on at last South Korea.



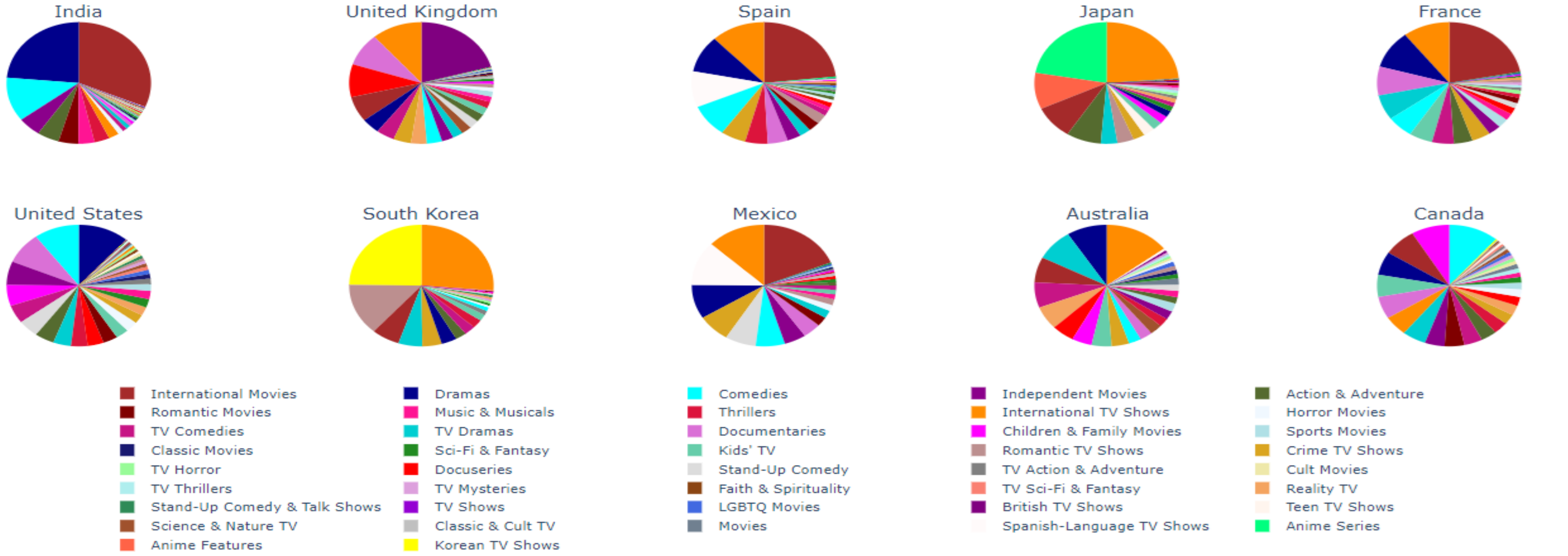
How many no of categories are present there in each content.



We can see that the majority of the contents are belonging to 3 categories.

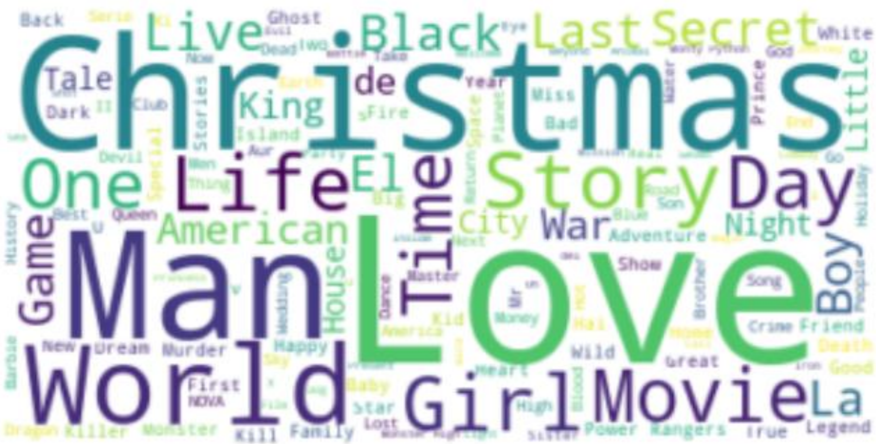
Which Genre is more popular in these countries.

- Some countries have a higher concentration of regional specialties, such as anime in Japan and Korean TV series in South Korea. This makes sense because Japanese people have long been fans of anime, and the growth of k-pop culture contributes for the popularity of Korean TV shows.
- In most nations, drama, foreign films, and comedies seem to be top on picks.
- However, it has been noted that international television shows and films are more popular in nations where English is not the native tongue.



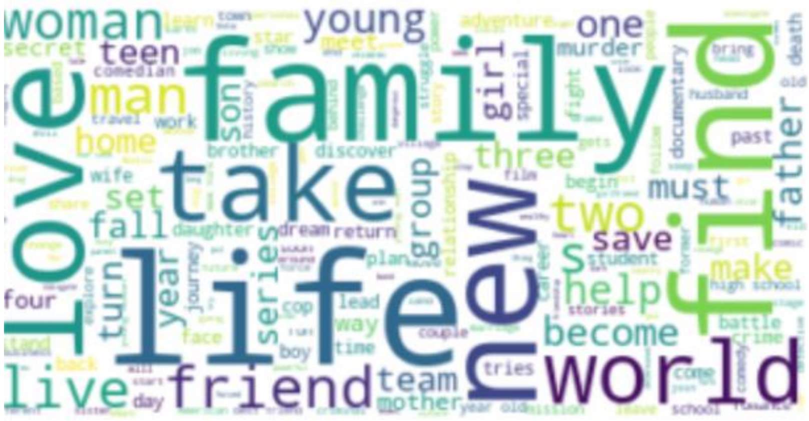
Applying WordCloud on Title and Description

WordCloud creators are used to highlight popular words and phrases based on frequency and relevance.

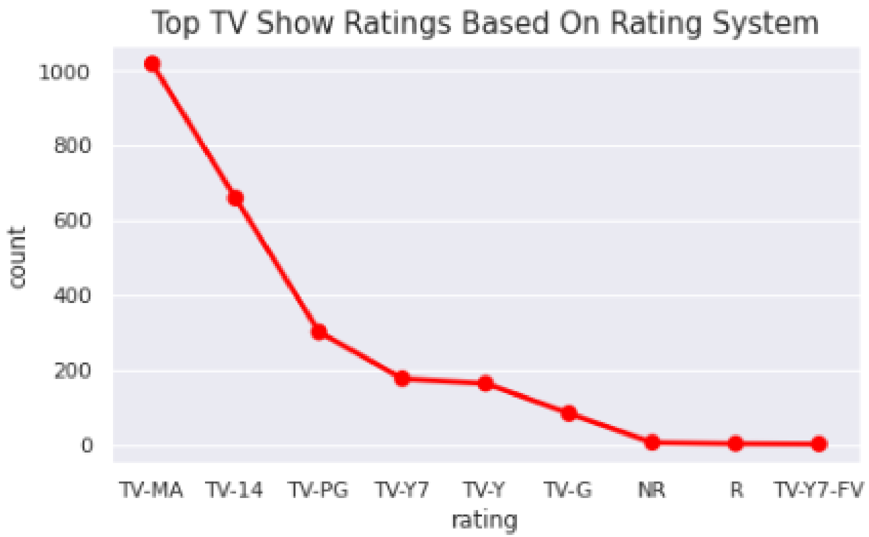


Unique and most occurring words in the title columns.

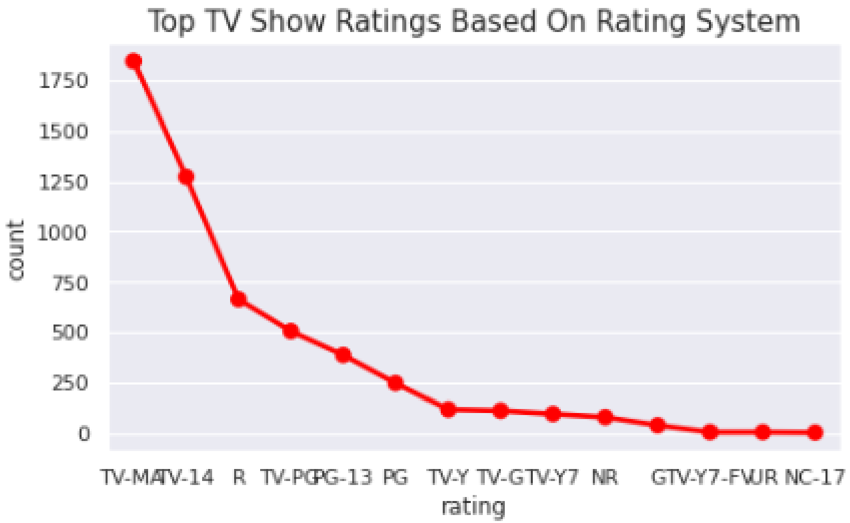
Unique and most occurring words in the description columns.



Analysis Most popular TV-Shows Rating.



Top TV Show Ratings Based On Rating System



Top TV Show Ratings Based On Rating System

Most of the contents got ratings like

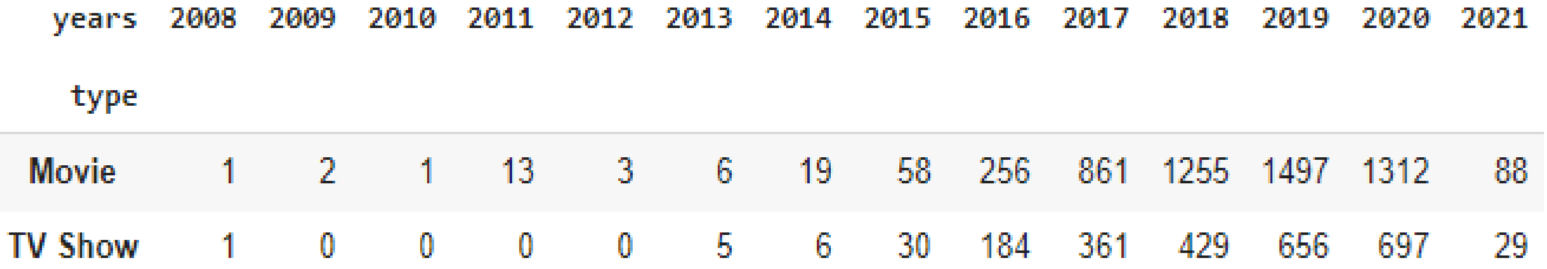
- TV-MA (For Mature Audiences)
- TV-14 (May be unsuitable for children under 14)
- TV-PG (Parental Guidance Suggested)
- NR (Not Rated)

Year v\s Types

Netflix has increasingly focused on TV rather than movies in recent years.

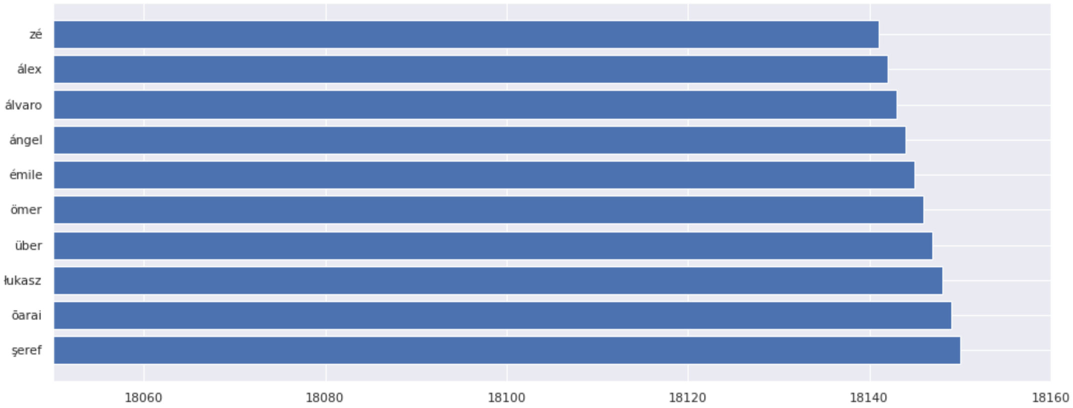
We use Hypothesis Testing to check if there is any relation between year and type.

- **Null Hypothesis :** year has no impact on the type of content that gets added to the platform.
- **Alternative Hypothesis :** year_added has an impact on the type of content that gets added to the platform.

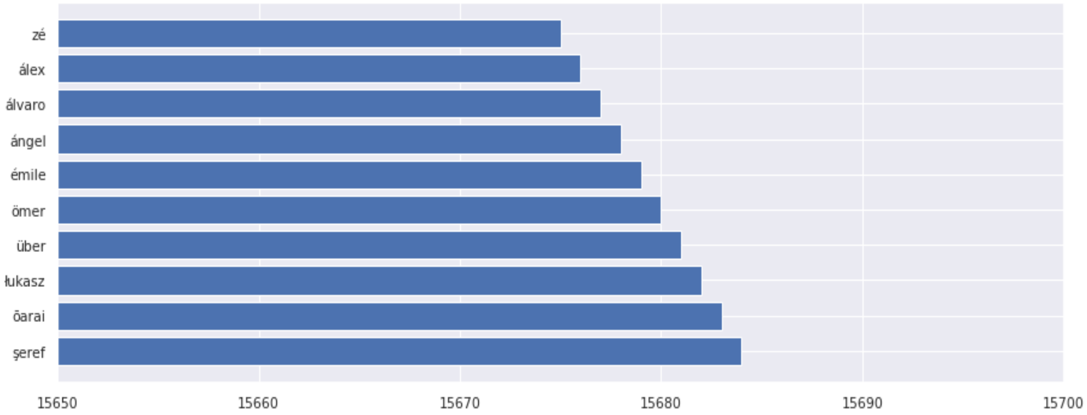


years	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
type														
Movie	1	2	1	13	3	6	19	58	256	861	1255	1497	1312	88
TV Show	1	0	0	0	0	5	6	30	184	361	429	656	697	29

Most occurred words In description.

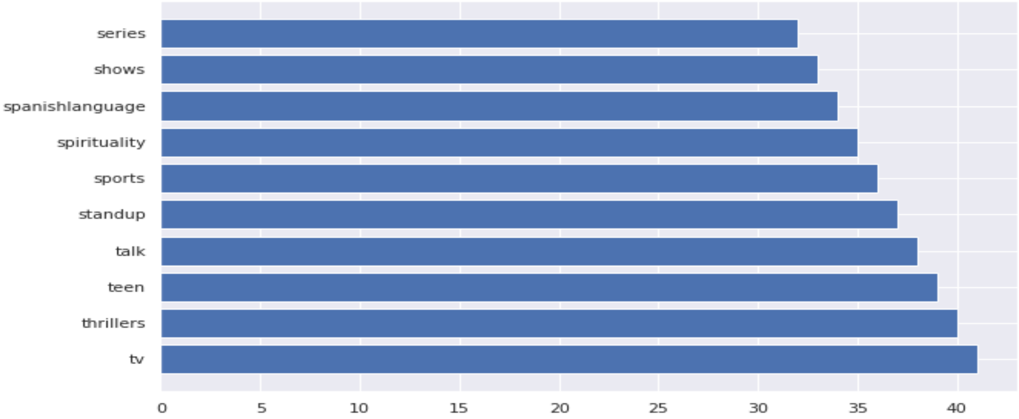


Before Stemming

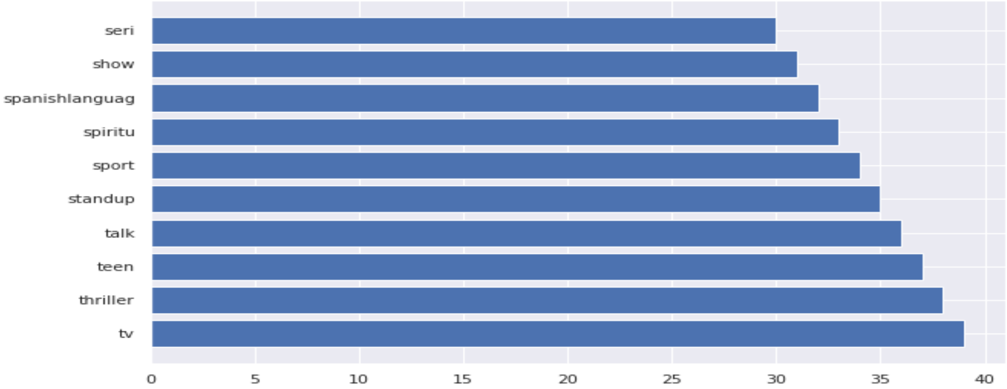


After Stemming

Most occurred words In listed_in.



Before Stemming



After Stemming

Feature Selection

Here we selected only 3 features for perform clustering.

- no_of_category
- description_length
- listed_in_length

After these we apply StandardScaler to standardize above features.

Apply Different clustering Algorithms.

1. Silhouette score
2. Elbow Method
3. DBSCAN

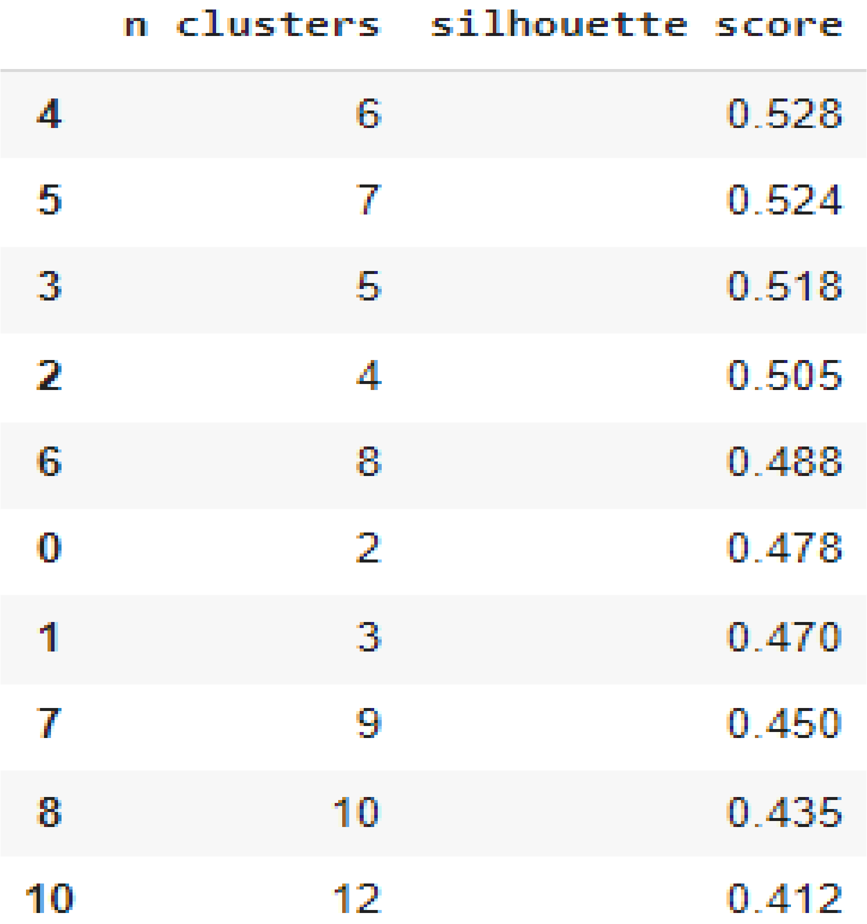
Silhouette score

Silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

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

mean intra-cluster distance (a) - mean Intra-cluster distance is the distance among members of a cluster, rather than the distance between two different clusters.

mean nearest-cluster distance (b) - The observation's average distance from every other data point in the subsequent closest cluster. This distance is also referred to as a.

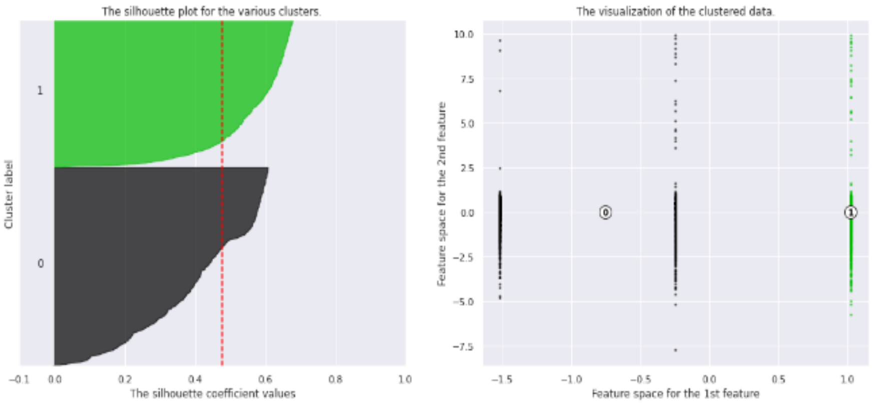


	n clusters	silhouette score
4	6	0.528
5	7	0.524
3	5	0.518
2	4	0.505
6	8	0.488
0	2	0.478
1	3	0.470
7	9	0.450
8	10	0.435
10	12	0.412

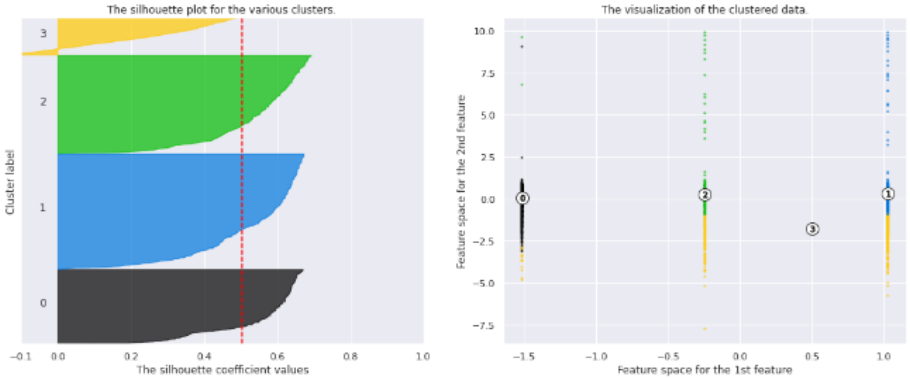
- 1: Means clusters are well apart from each other and clearly distinguished.
- 0: We can say that the distance between clusters in not significant.
- -1: Means clusters are assigned in the wrong way.
- Plot different cluster values using graph in next slide.

Continued….

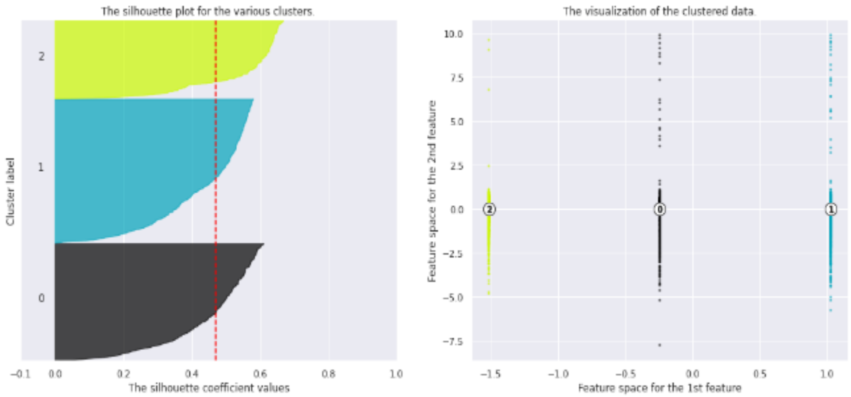
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



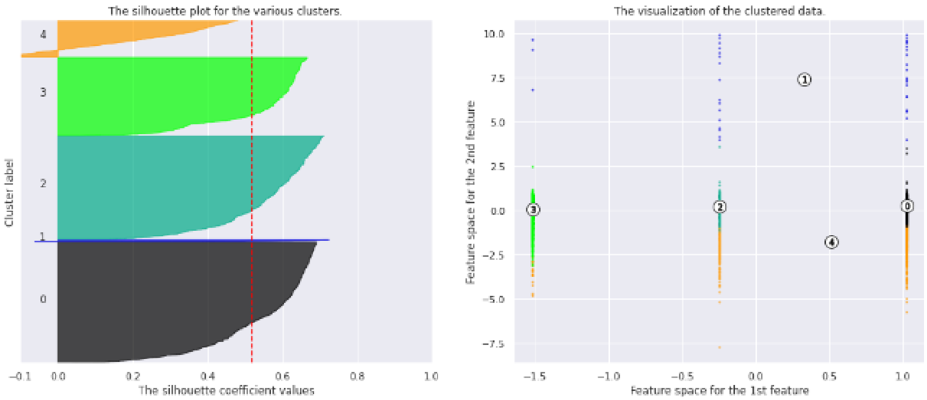
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



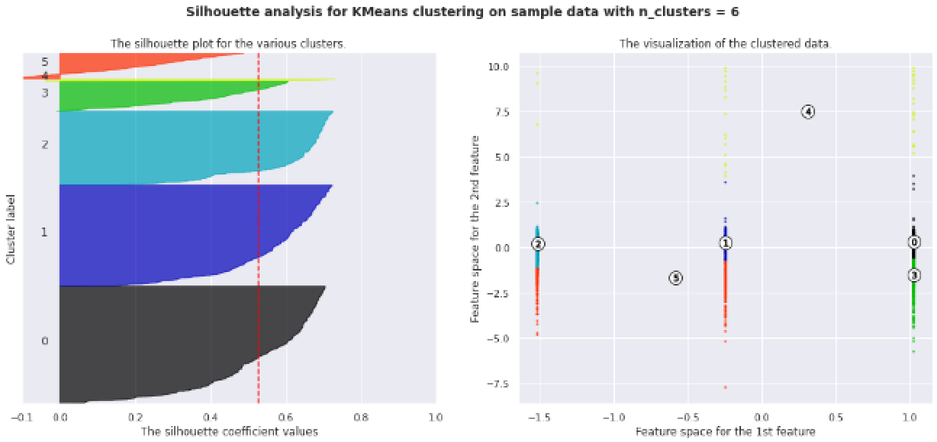
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

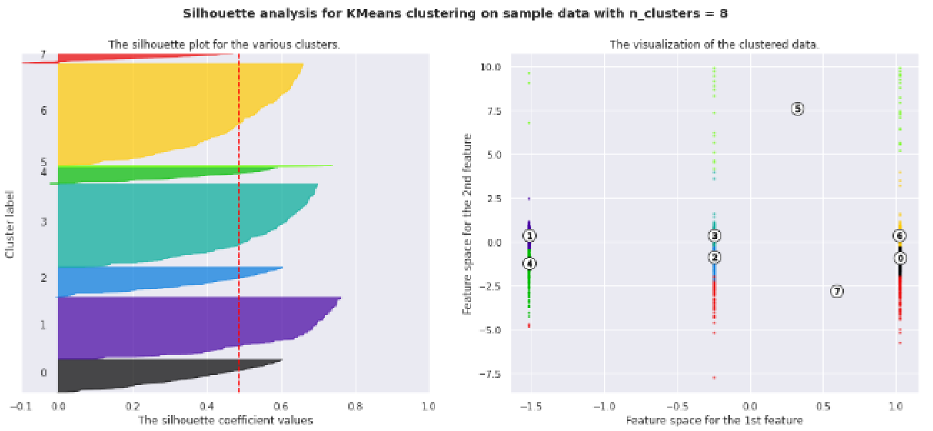


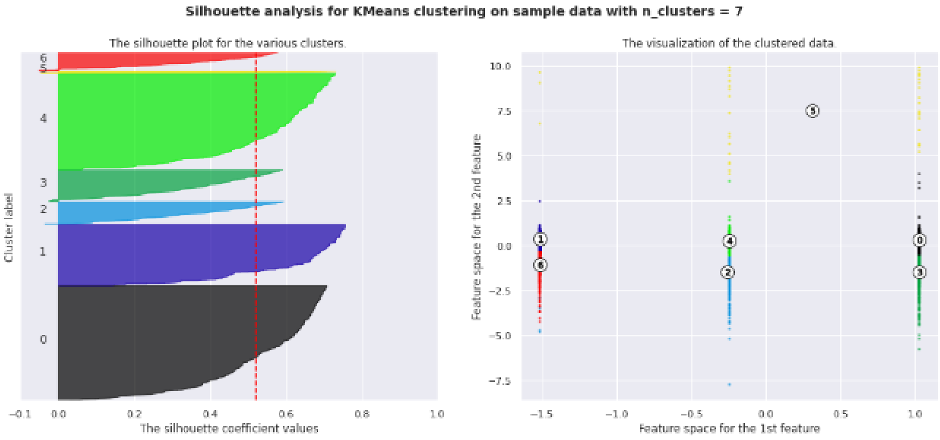
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5

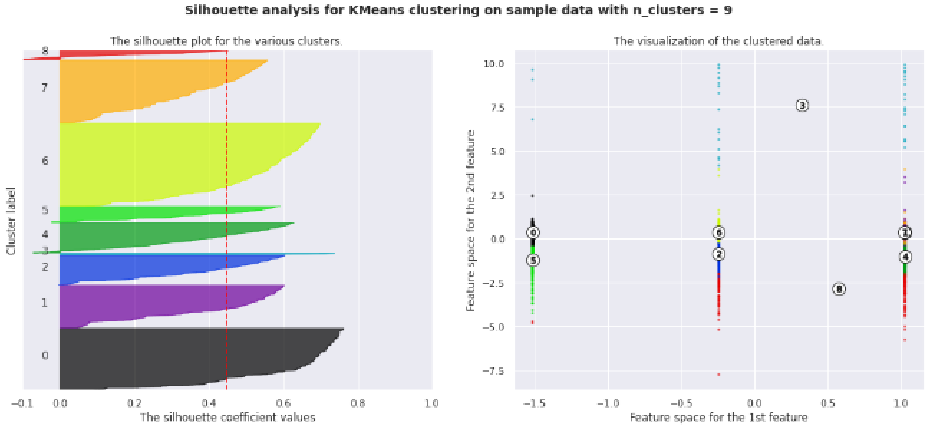


Continued….


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6

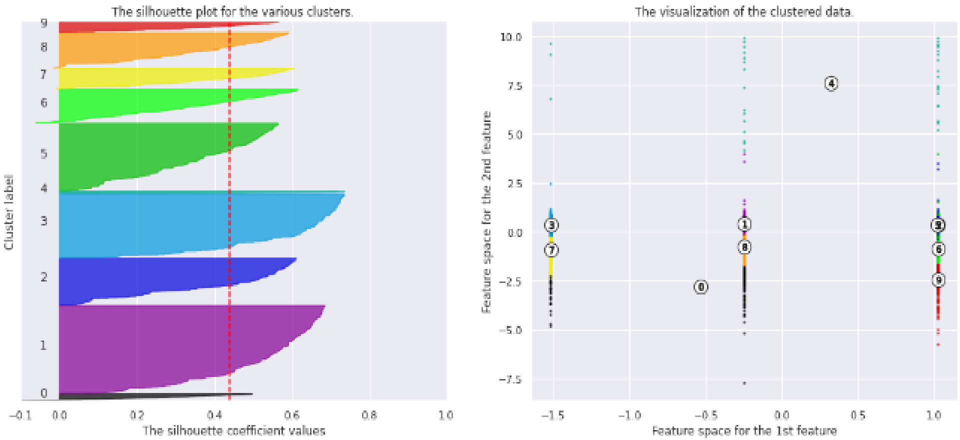

Silhouette analysis for KMeans clustering on sample data with n_clusters = 8


Silhouette analysis for KMeans clustering on sample data with n_clusters = 7

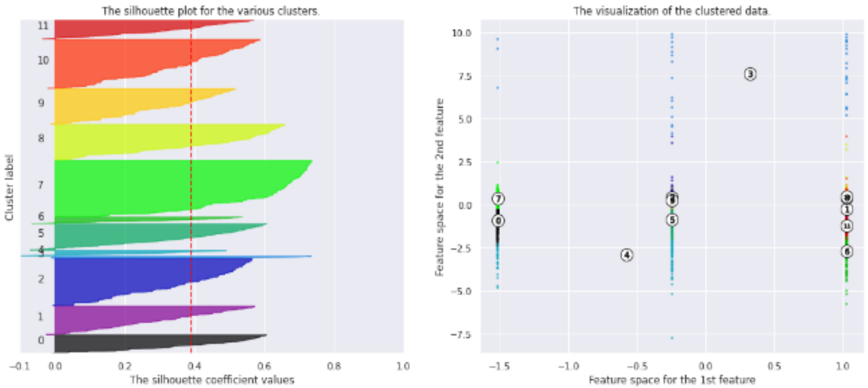

Silhouette analysis for KMeans clustering on sample data with n_clusters = 9

Continued….

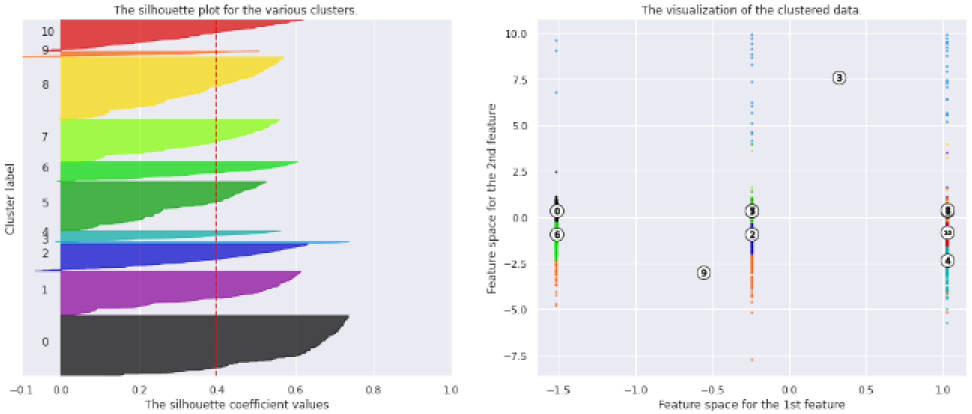
Silhouette analysis for KMeans clustering on sample data with n_clusters = 10



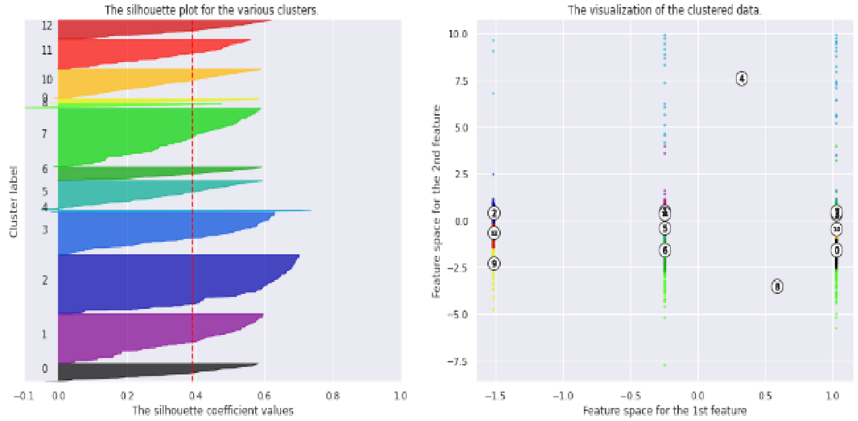
Silhouette analysis for KMeans clustering on sample data with n_clusters = 12



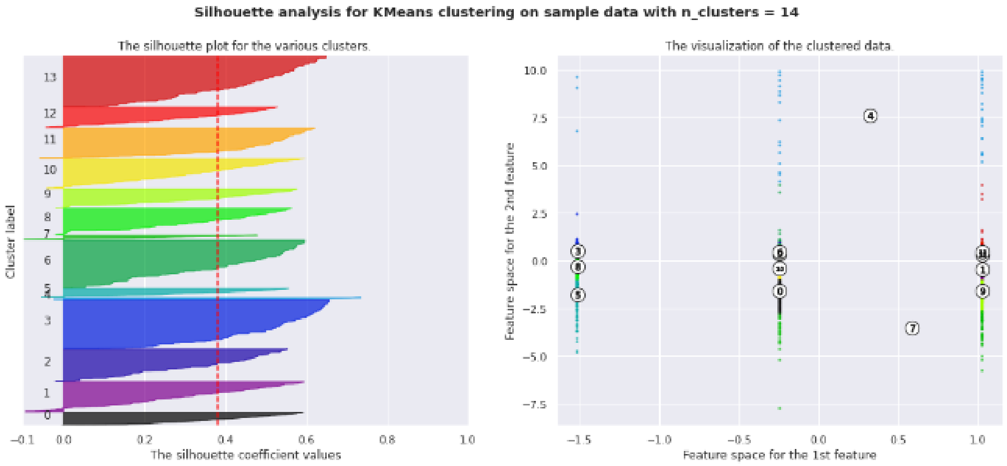
Silhouette analysis for KMeans clustering on sample data with n_clusters = 11

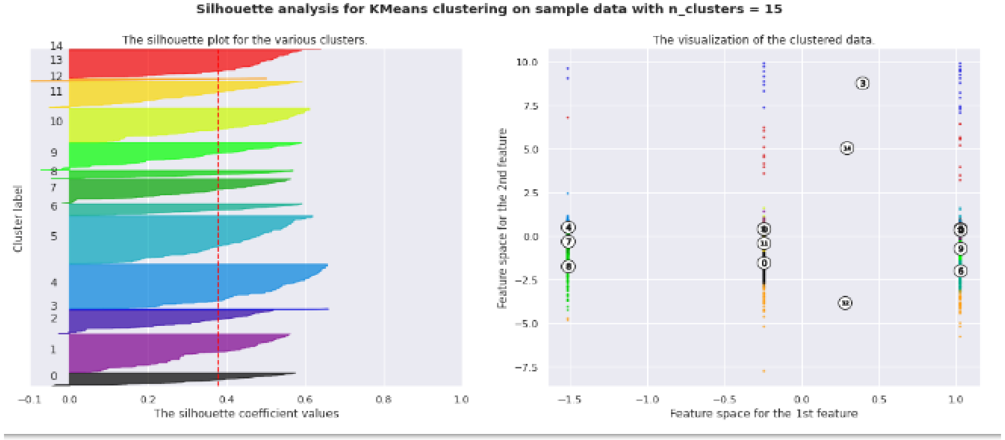


Silhouette analysis for KMeans clustering on sample data with n_clusters = 13

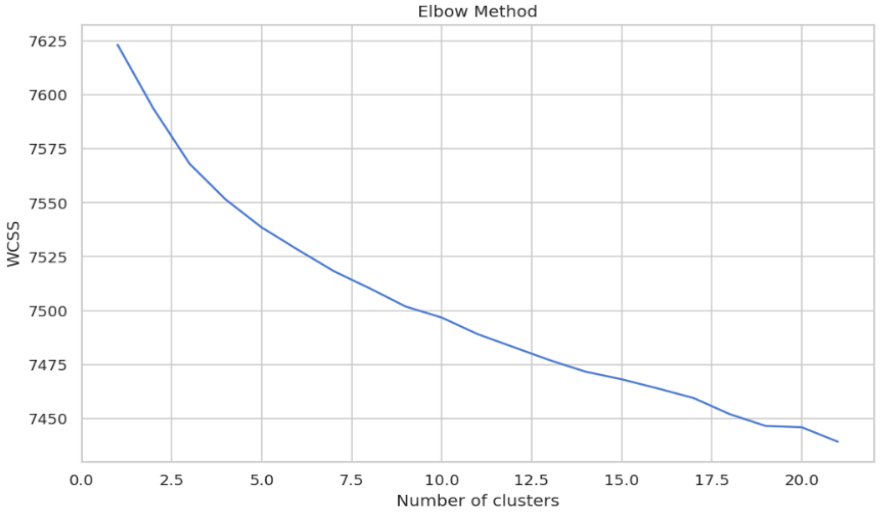


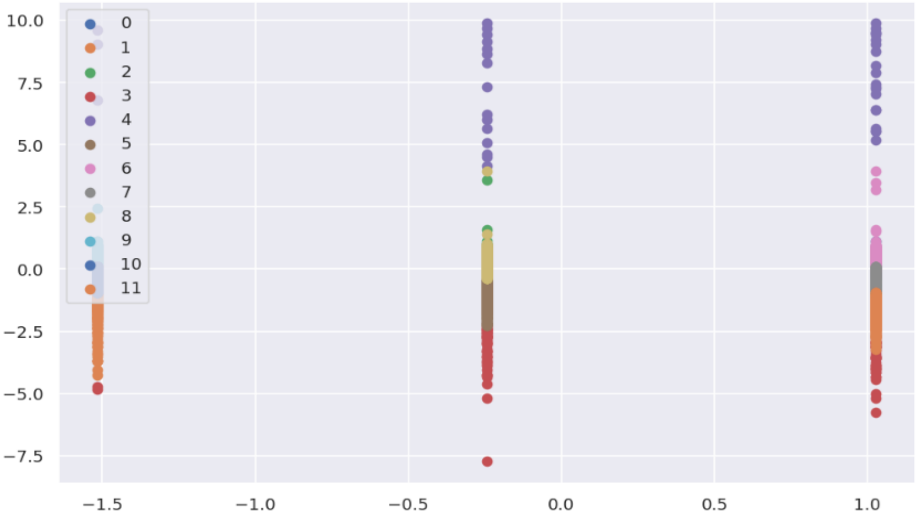
Continued….


Silhouette analysis for KMeans clustering on sample data with n_clusters = 14
The silhouette plot for the various clusters.
Cluster label
13
12
11
10
9
8
7
6
5
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
Feature space for the 1st feature


Silhouette analysis for KMeans clustering on sample data with n_clusters = 15
The silhouette plot for the various clusters.
Cluster label
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
Feature space for the 1st feature

Elbow Method

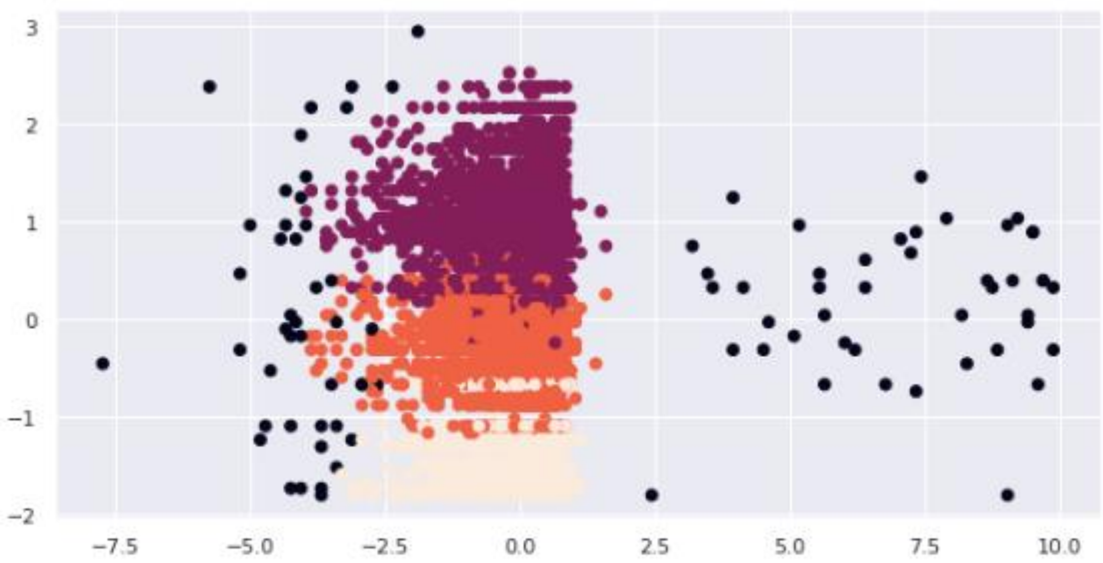




The elbow method applies k-means clustering on the dataset for a range of k values, such as 1 to 22, and then calculates the WCSS score for each value of k.

By default, to calculate the distortion score, the total of the square distances between each point and its designated centre is used.

DBSCAN

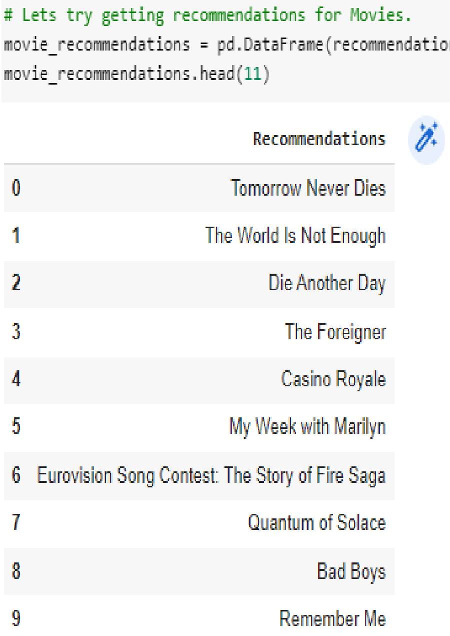


Density-based spatial clustering of applications with noise (DBSCAN) clustering method.

The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

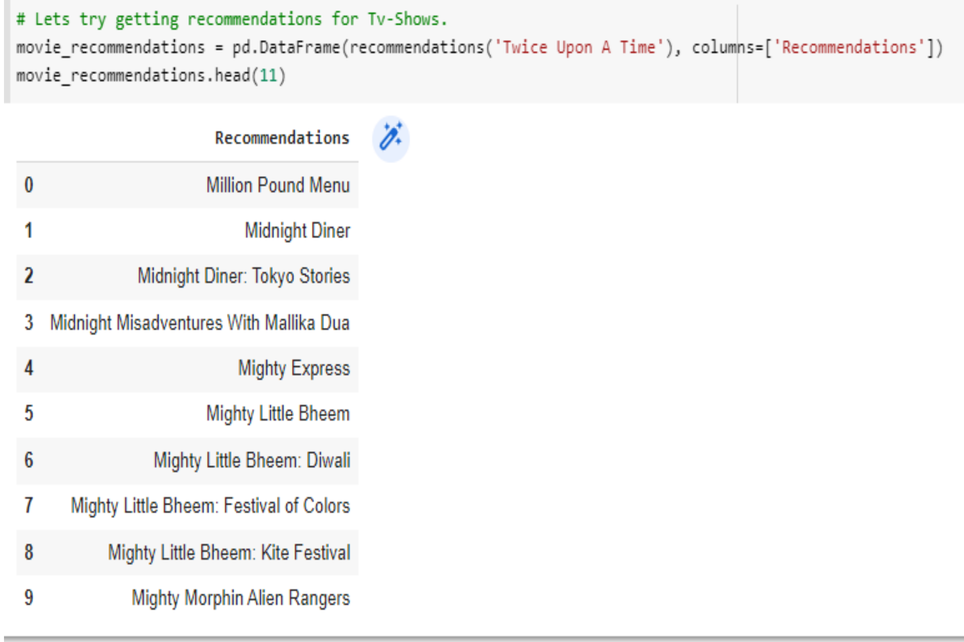
Recommendations

Here We obtained recommendations for Movies and Tv- Shows using Cosine similarity.



```
# Lets try getting recommendations for Movies.
movie_recommendations = pd.DataFrame(recommendations('GoldenEye'), columns=['Recommendations'])
movie_recommendations.head(11)
```

	Recommendations
0	Tomorrow Never Dies
1	The World Is Not Enough
2	Die Another Day
3	The Foreigner
4	Casino Royale
5	My Week with Marilyn
6	Eurovision Song Contest: The Story of Fire Saga
7	Quantum of Solace
8	Bad Boys
9	Remember Me



```
# Lets try getting recommendations for Tv-Shows.
movie_recommendations = pd.DataFrame(recommendations('Twice Upon A Time'), columns=['Recommendations'])
movie_recommendations.head(11)
```

	Recommendations
0	Million Pound Menu
1	Midnight Diner
2	Midnight Diner: Tokyo Stories
3	Midnight Misadventures With Mallika Dua
4	Mighty Express
5	Mighty Little Bheem
6	Mighty Little Bheem: Diwali
7	Mighty Little Bheem: Festival of Colors
8	Mighty Little Bheem: Kite Festival
9	Mighty Morphin Alien Rangers

Conclusions:

1. The majority of the Netflix content is movies, making it an interesting discovery.
2. There are two different forms of material in this dataset: movies (69.14%) and TV shows (30.86%).
3. We may infer from the dataset insights that the most TV shows were released in 2017, and the most movies were published in 2020.
4. But it has increasingly been concentrating more on television shows.
5. Among the top 5 countries that create all of the content that is made available on the site are the United States and India.
6. In fact, six of the top ten actors with the most content are Indian.
7. In text analysis (NLP) I used stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP.
8. k=10 was found to be an optimal value for clusters using which we grouped our data into 10 distinct clusters.
9. In text analysis (NLP) I used stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP.

Future Scope

1. Many exciting discoveries can be obtained by combining this dataset with other external datasets, such as IMDB ratings and rotten fruit.
2. A better recommender system might be developed with more time and then put online for users to use.

THANK YOU