Capstone Project

Netflix Movies and TV Shows Clustering

(Unsupervised Clustering and Recommendation System)

By- Navneet Keshri

Steps Involved

- Problem Statement
- Understanding Data
- Exploratory Data Analysis
- Data Cleaning
- Textual Data Preprocessing
- Model Implementation
- Build Recommendation System
- Conclusion

Project Goal:

The objective of the project is to classify or group Netflix Movies and TV shows into specific clusters in such a way that Movies and TV shows in different groups should have very different properties or features, while Movies and TV shows in the same cluster or group should have similar properties or features.

Problem Statement

- As of 2022-Q2, more than 220 million people had signed up for Netflix's online streaming service, making it the largest OTT provider worldwide. To improve the user experience and prevent subscriber churn, they must efficiently cluster the shows hosted on their platform.

- By creating clusters, we will be able to comprehend the shows that are alike and different from one another. These clusters can be used to provide customers with personalized recommendations based on their preferences.

Data Pipeline

1. **Analyze Data:** In this initial step, we will attempt to comprehend the data and search for various available features. We will look for things like the shape of the data, the data types of each feature, a statistical summary, etc. at this stage.

2. **EDA:** EDA stands for Exploratory Data Analysis. It is a process of analyzing and understanding the data. The goal of EDA is to gain insights into the data, identify patterns, and discover relationships and trends. It also helps to identify outliers, missing values, and any other issues that may affect our analysis and model building.

3. **Data Cleaning:** Data cleaning is the process of identifying and correcting or removing inconsistencies, and missing values in a dataset. We will inspect the dataset for duplicate values. The null value and outlier detection and treatment will be followed.

Data Pipeline

4. **Textual Data Preprocessing:** We will cluster the data based on the attributes at this stage. All punctuation marks and stop words are removed during data preprocessing, and all textual data is converted to lowercase. Stemming to construct a meaningful word from a word corpus. We will use vectorization and tokenization of the corpus and we will use Principal Component Analysis (PCA) to handle the curse of dimensionality.

5. **Clusters Implementation:** We will use the K-Means clustering and Agglomerative Hierarchical clustering to group the movies and TV shows. To determine the ideal number of clusters we will use different techniques.

6. **Build Content-Based Recommendation System:** The similarity matrix generated by applying cosine similarity will be used to construct a content-based recommender system. The user will receive 10 recommendations from this recommender system based on the type of movie or television show they have viewed.

Dataset Summary:

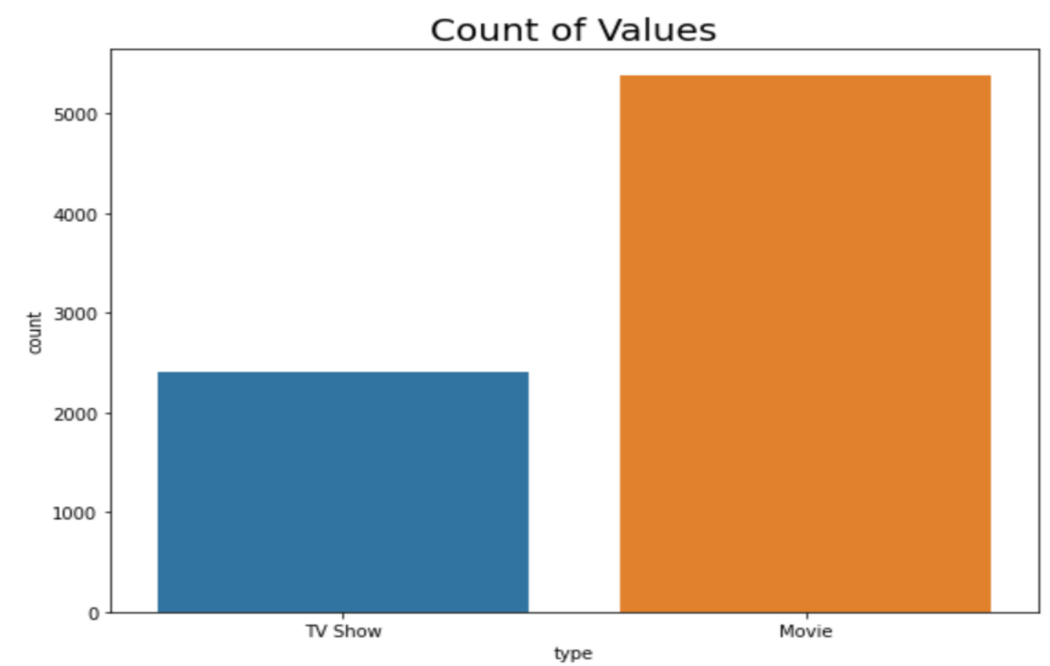
- This dataset includes Netflix movies and TV shows that are available until 2020.
- There are approximately 7787 rows and 12 columns available in the dataset.
- The majority of features are in textual form.
- The movie and the TV show are the two types of content in the dataset.

ATTRIBUTES INFORMATION:

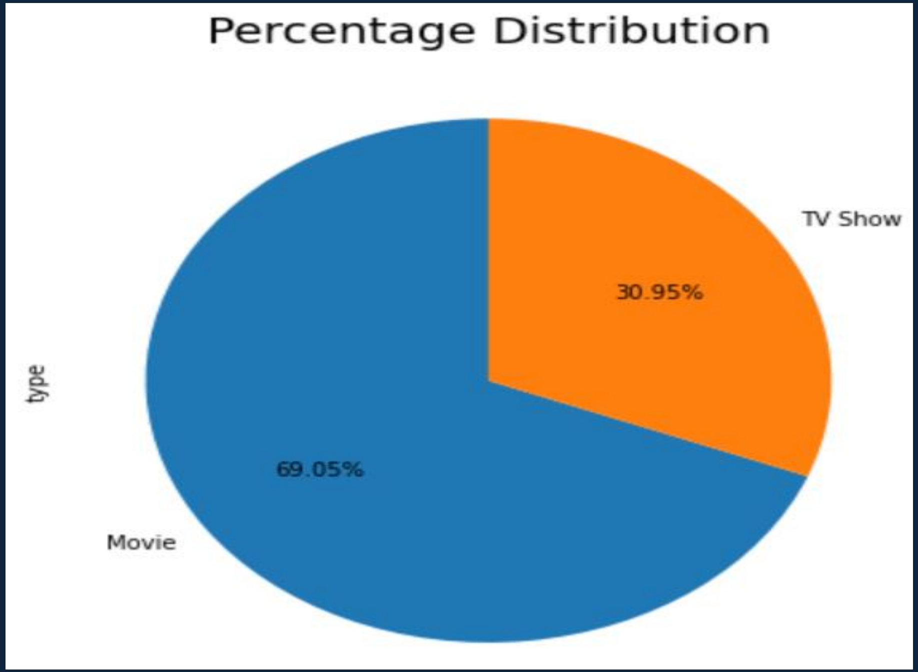
- show_id: Unique ID for every Movie/Show
- type: Identifier - Movie/Show
- title: Title of the Movie/Show
- director: Director of the Movie/Show
- cast: Actors involved in the Movie/Show
- country: Country where the Movie/Show was produced
- date_added: Date it was added on Netflix
- release_year: Actual 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

Exploratory Data Analysis:

Column: ‘type’



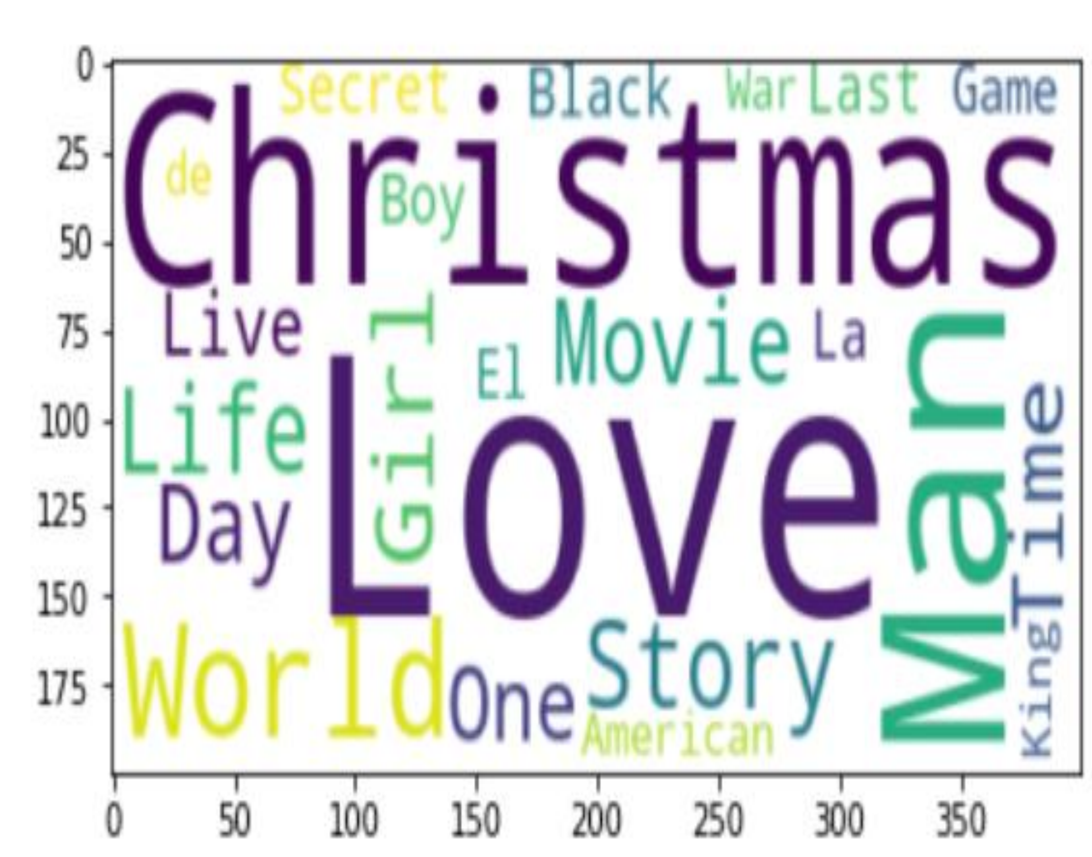
Movies have more counts than TV Shows.



31% of the data are from TV shows, while 69% of the data are from movies.

Exploratory Data Analysis:

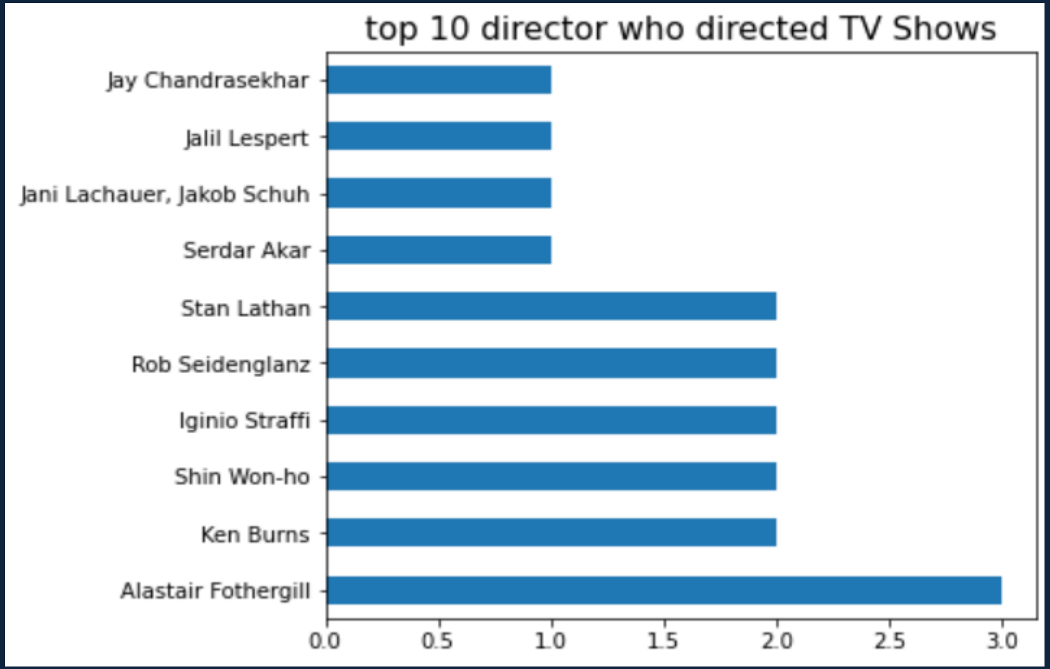
Column: ‘title’



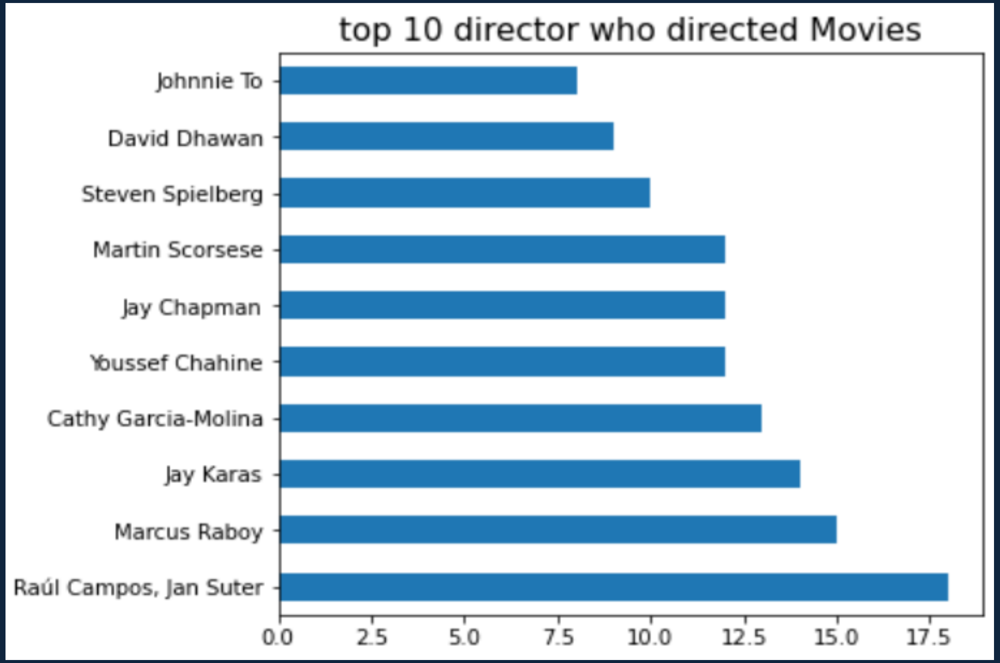
- Word cloud is a visual representation of the most frequently used words in a given text or set of texts.
- most frequently used words appearing in larger font sizes.
- Words like 'Love', 'Christmas', 'Man', 'World', 'Life', 'Girl', and 'Story' are frequently used in the movie title column.

Exploratory Data Analysis:

Column: ‘director’



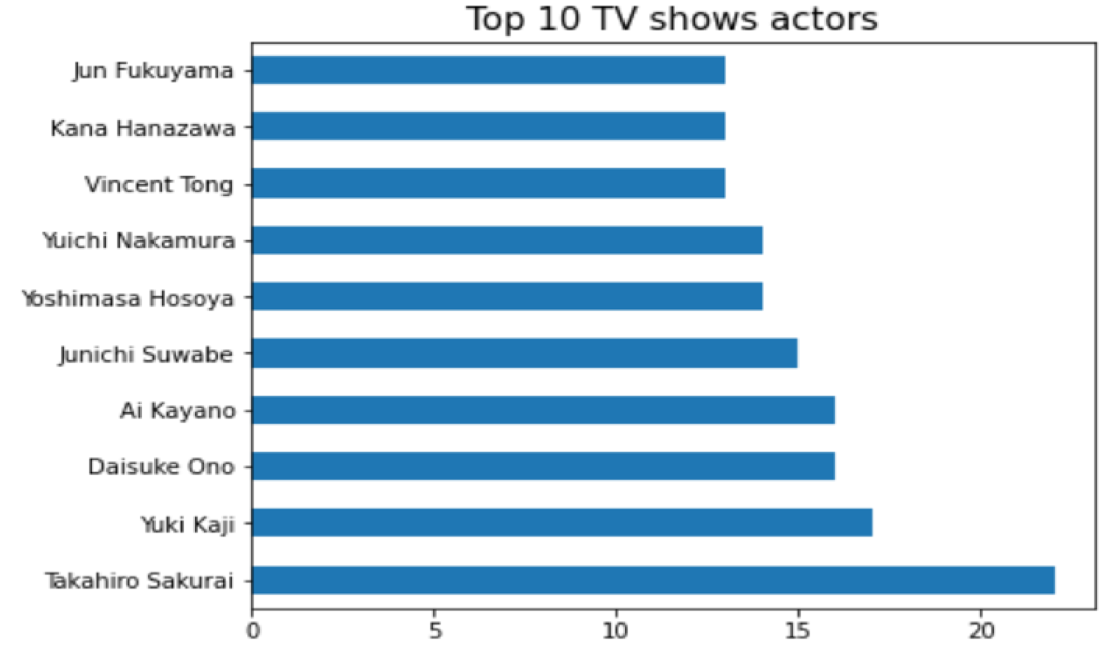
3 shows directed by Alastair Fothergill are the highest on the data list.



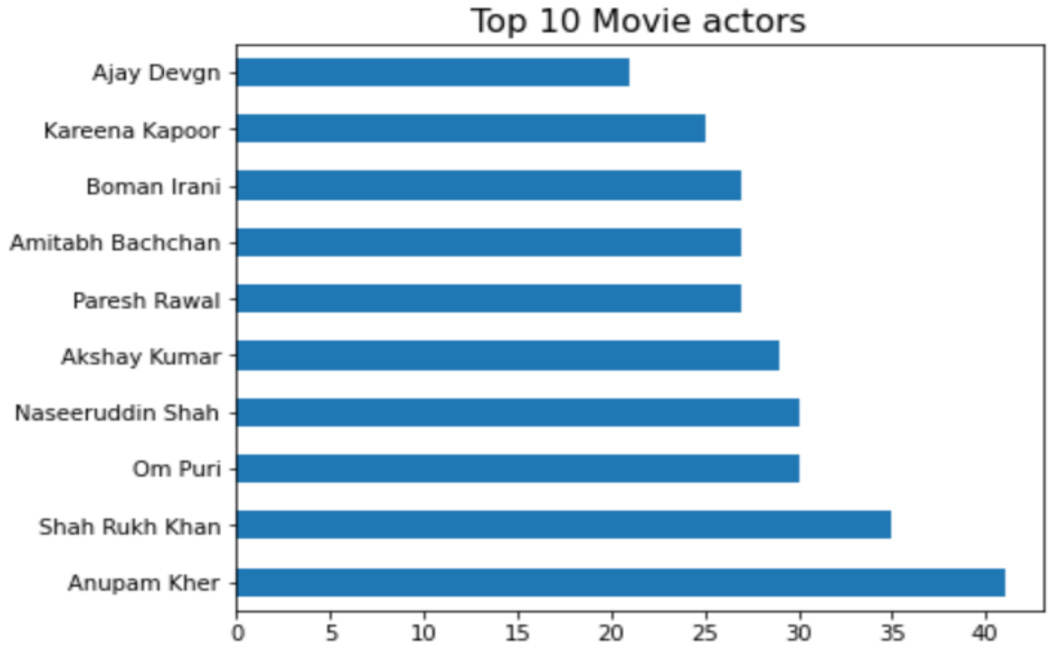
Both, Jan Suter and Raul Campos have directed 18 films, more than anyone else in the dataset.

Exploratory Data Analysis:

Column: ‘cast’



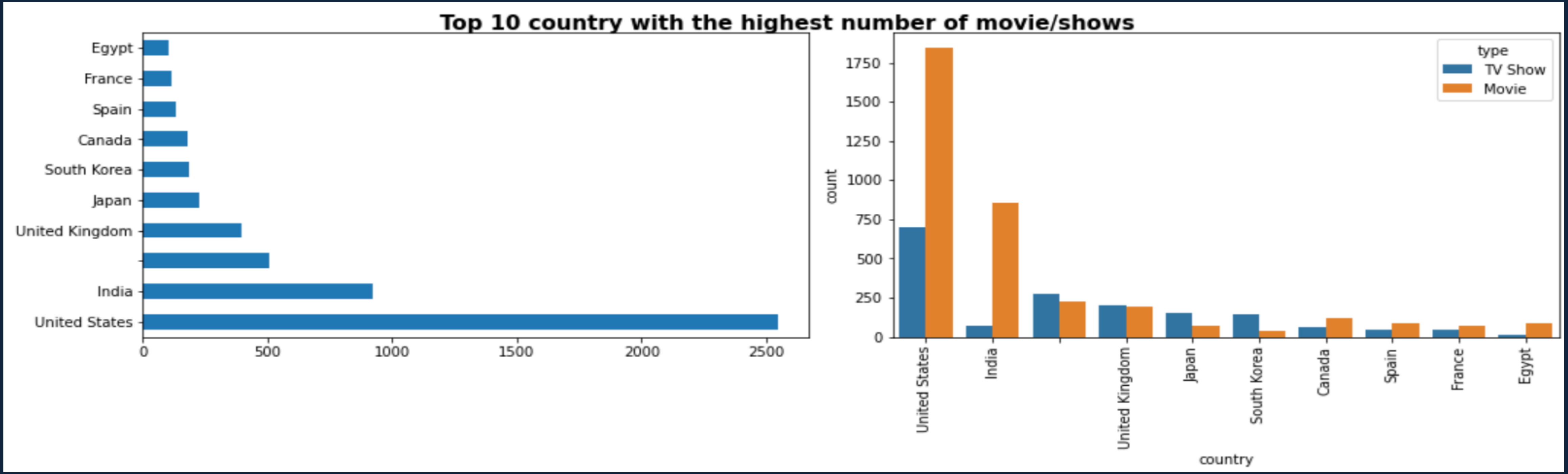
In the TV shows, Takahiro Sakurai, Yuki Kaji, and Daisuke Ono played the most number of roles.



The majority of the roles in the movies are played by Anupam Kher, Shahrukh Khan, and Om Puri.

Exploratory Data Analysis:

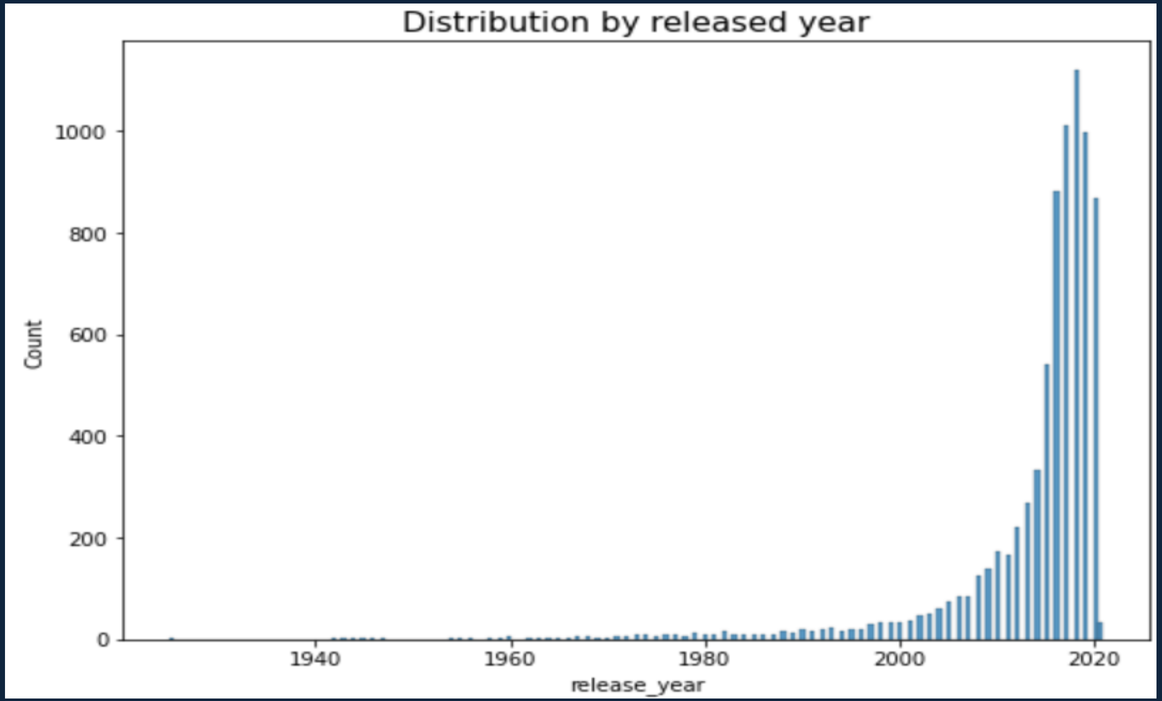
Column: ‘country’

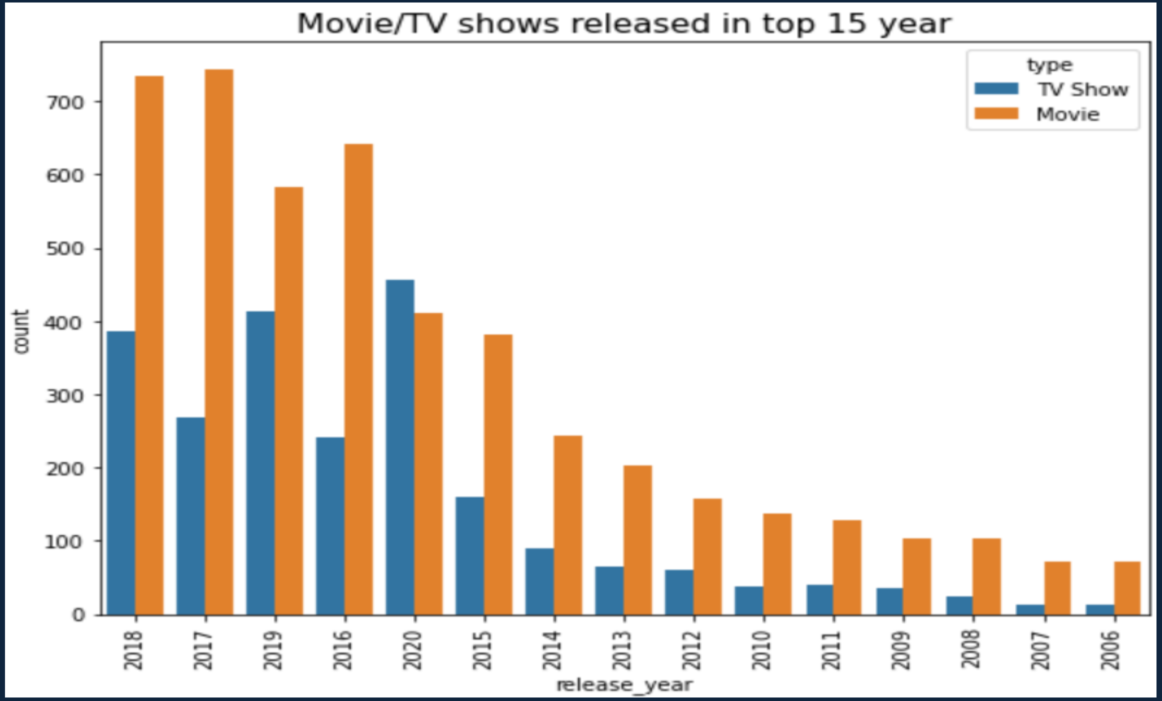


- The United States-based movies and TV shows were produced most, followed by India and the United Kingdom.
- In India and United State, a greater number of movies are present compared to TV shows.
- In the UK, Japan, and South Korea there are a greater number of TV shows than movies.

Exploratory Data Analysis:

Column: ‘release_year’



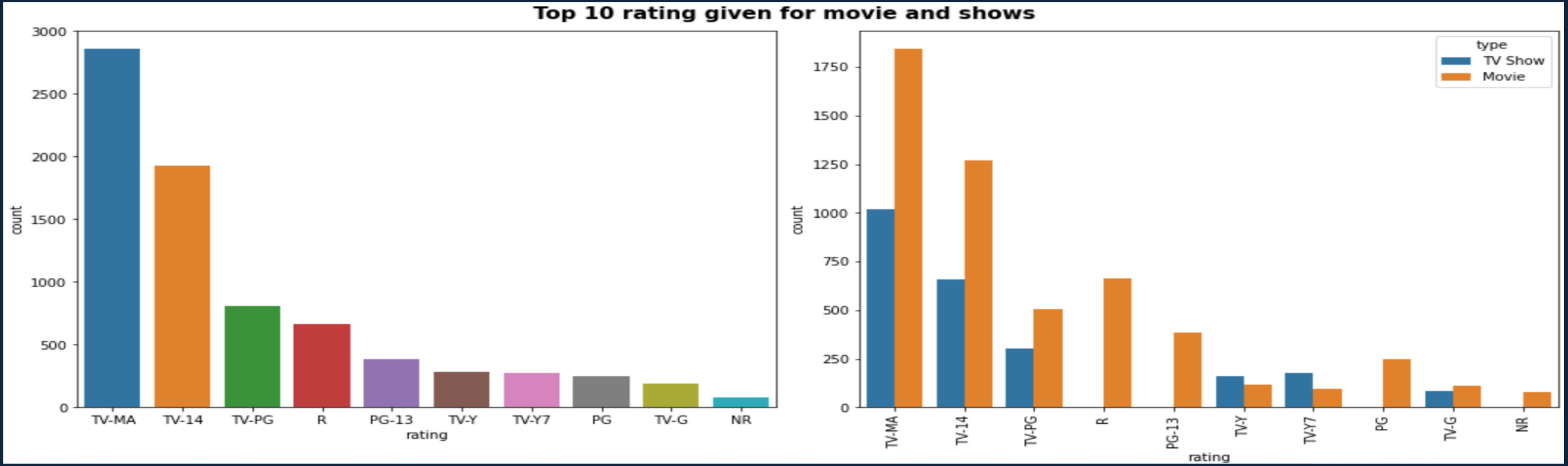


Netflix starts releasing more Movies/TV shows in recent years compared to old ones.

Most Movies and TV shows are available on Netflix between 2015 and 2020, and the highest are in 2018.

Exploratory Data Analysis:

Column: ‘rating’



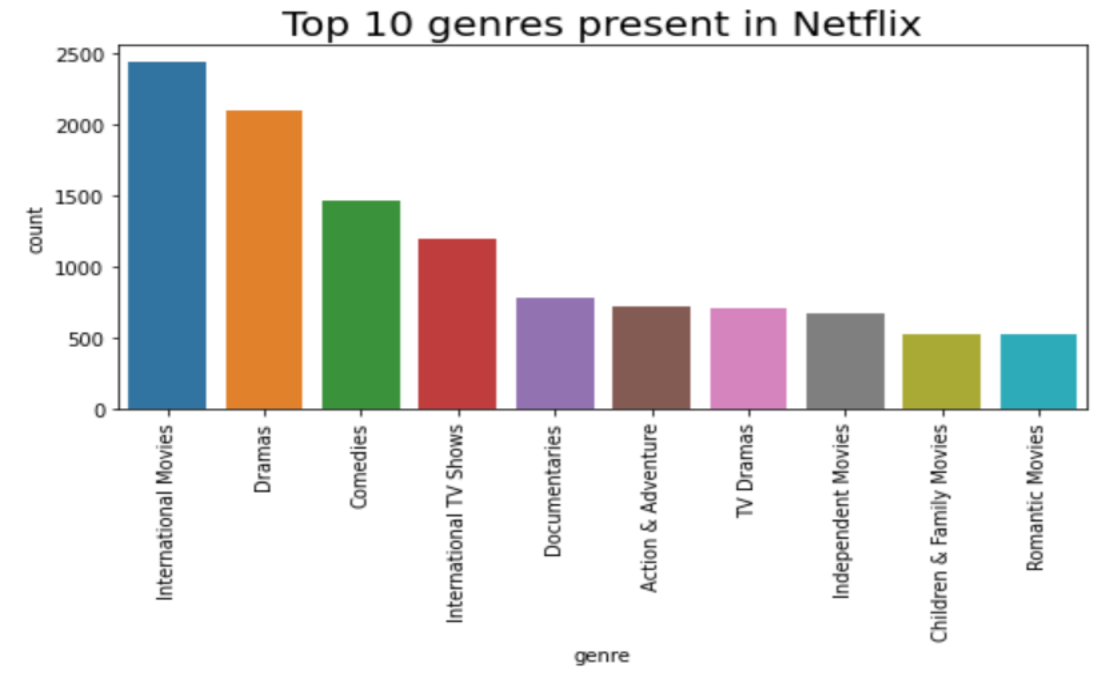
Top 10 rating given for movie and shows

The majority of Movies and TV shows have a rating of TV-MA, which stands for "Mature Audience," followed by TV-14, which stands for "Younger Audience."

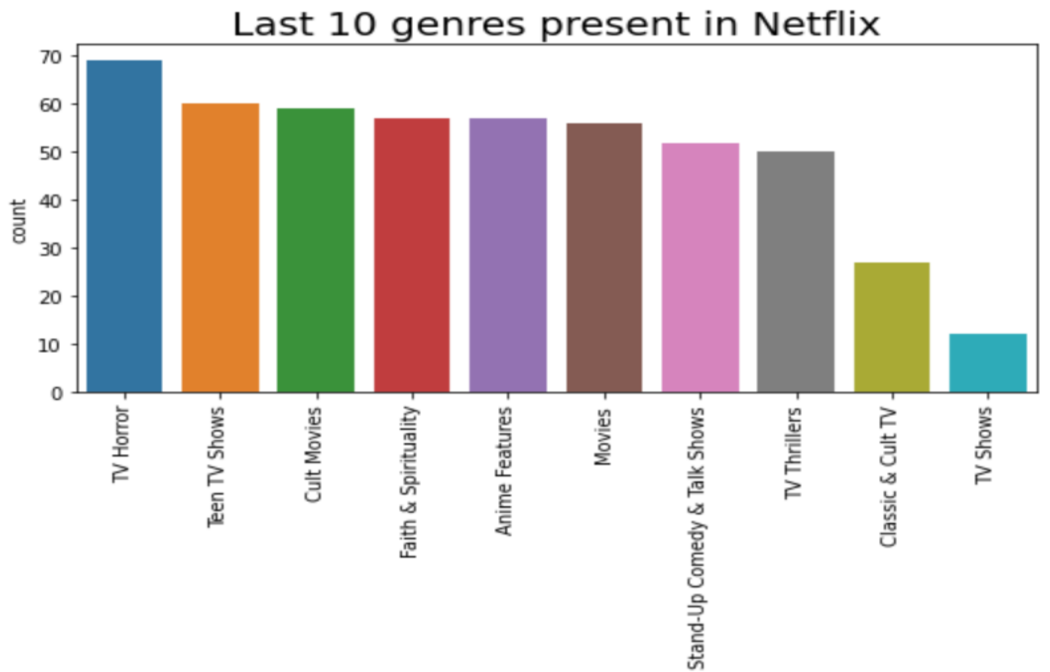
When compared to TV shows, Movies receive the highest rating, which is pretty obvious given that the number of Movies is higher compared to TV shows, as we saw earlier in the type column.

Exploratory Data Analysis:

Column: ‘listed_in (genre)’

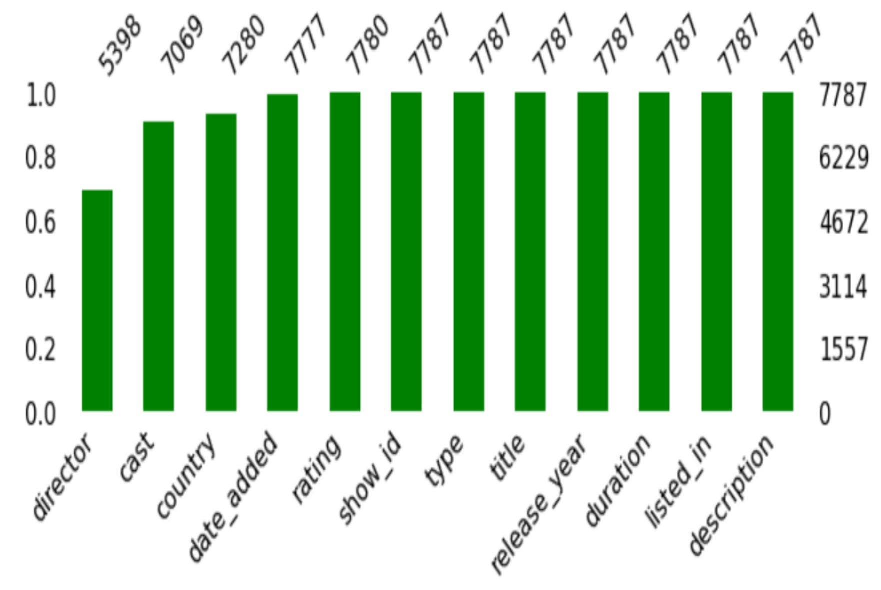


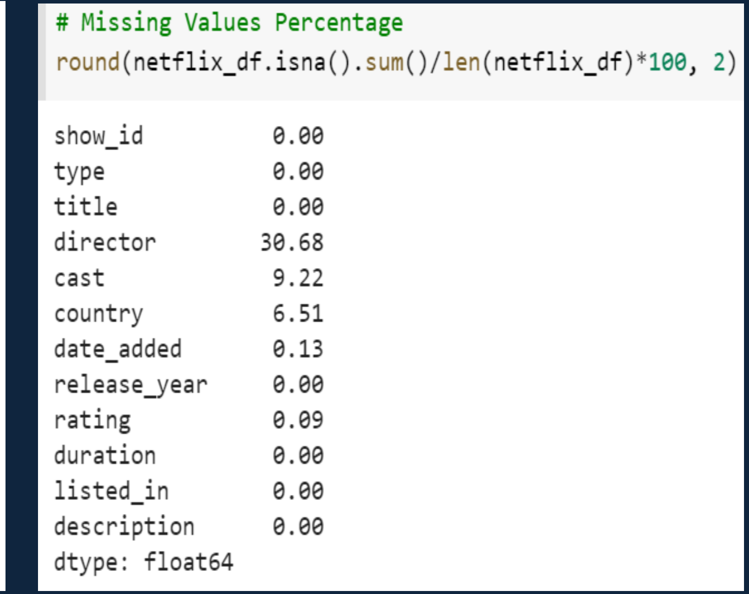
International Movies, Dramas, and Comedies are the most listed genres.



TV Shows, Classic and cult TV, TV thrillers, Stand-Up comedy, and Talk shows account for the least number of listing in genres.

Data Cleaning:





```
# Missing Values Percentage
round(netflix_df.isna().sum()/len(netflix_df)*100, 2)
```

```
show_id          0.00
type             0.00
title            0.00
director        30.68
cast             9.22
country          6.51
date_added       0.13
release_year     0.00
rating           0.09
duration         0.00
listed_in        0.00
description      0.00
dtype: float64
```

- The goal of data cleaning is to improve the quality of the data and make it suitable for further analysis and modeling.
- Null values present in the director, cast, country, date_added, and rating columns.
- All the data that we have is related to each specific movie. So, we can't impute null values with any method. Also, we don't want to lose any data since the data size is small.

1. The null values in the director, cast, and country columns can be replaced with an 'empty string'.

2. Small amount of null value percentage present in rating and date_added column, if we drop these nan values it will not affect that much while building the model. So, we simply drop the null values present in the rating and date_added columns.

Textual Data Pre-processing:

1. **Selecting Attributes:** We have clustered the Netflix movies and TV shows into groups based on the following textual characteristics:

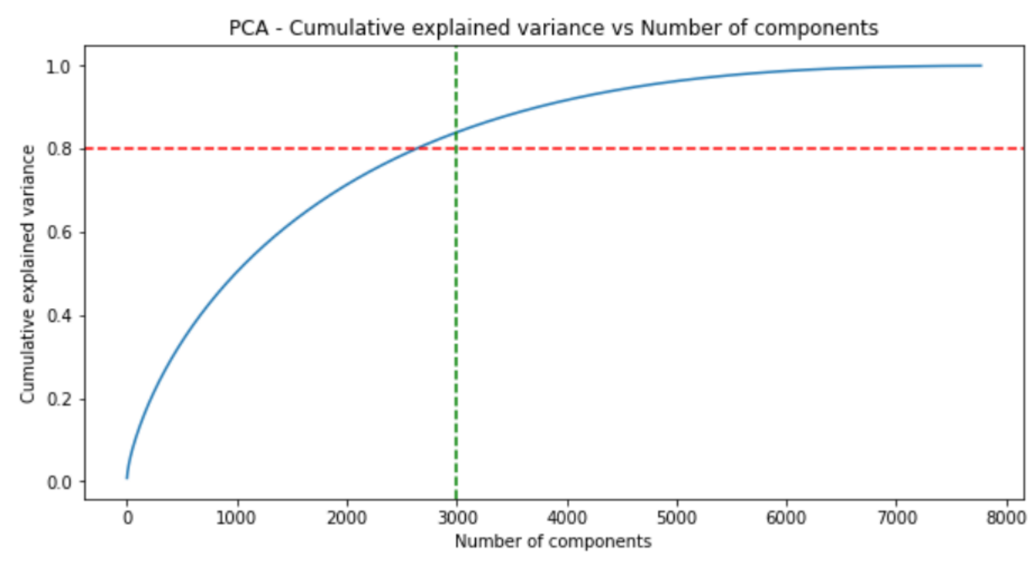
- Director
- Cast
- Country
- Rating
- Listed in (genres)
- Description

2. **Removing Stop words:** Words such as "a," "an," "the," and "is," are words that are commonly used in a language but do not convey meaningful information. These words can add noise to the data. So we removed stop words.
3. **Lowercasing words:** It is the process of converting all the words in a text to lowercase. Useful where case differences are not important and also can reduce the size of the vocabulary.
4. **Removing Punctuation:** Process of removing any punctuation marks (e.g., periods, commas, exclamation points, etc.) as punctuation marks often do not carry much meaning and can add noise to the data.

AI

Textual Data Pre-processing:

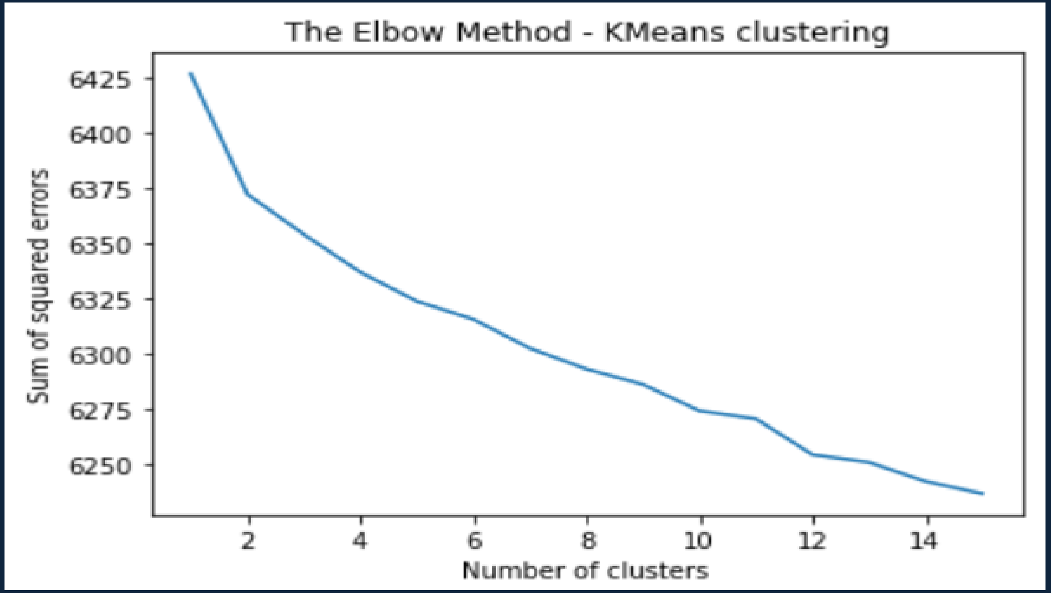
5. **Stemming:** Stemming is the process of reducing a word to its base or root form. For example, stemming would reduce "running," "runner," and "ran" to the base form “run”. We utilized Snowball Stemmer for this.
6. **Text Vectorization:** Process of converting text data into numerical vectors that can be used for machine learning. We used TF-IDF vectorizer for this process.
7. **Dimensionality Reduction:** Process of reducing the number of features or dimensions in a dataset while retaining as much information as possible. We used Principal Component Analysis (PCA) to reduce the dimensionality of the data.

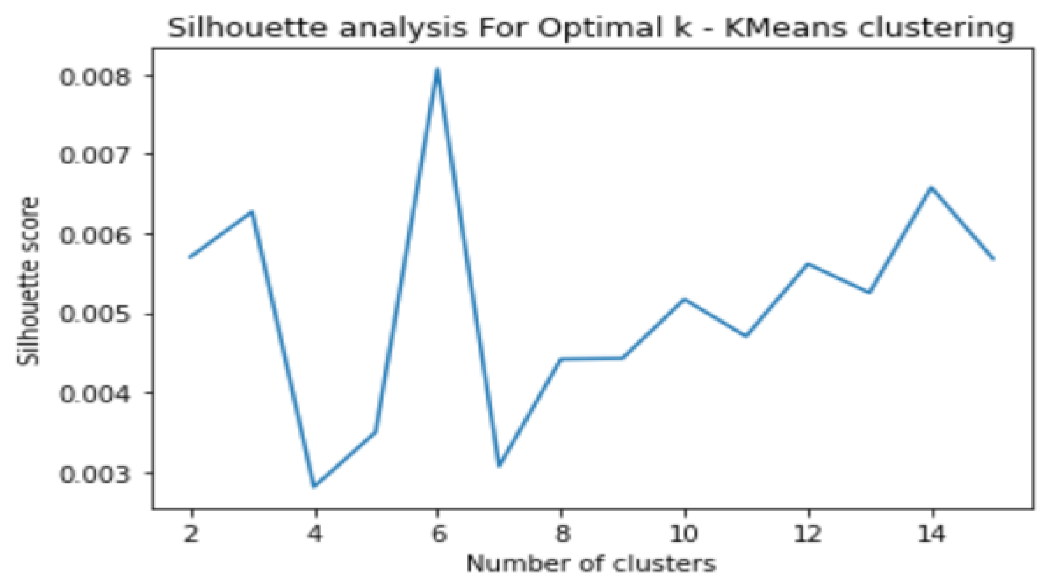


- We discovered that approximately 7500 components account for 100 percent of the variance.
- 3000 components alone account for more than 80% of the variance.
- Therefore, we can take the top 3000 components to reduce dimensionality and simplify the model while still being able to capture more than 80% of the variance.

Model Implementation:

K-Means Clustering: We visualized the elbow curve and Silhouette score to decide on the optimal number of clusters for the K-Means Clustering algorithm.



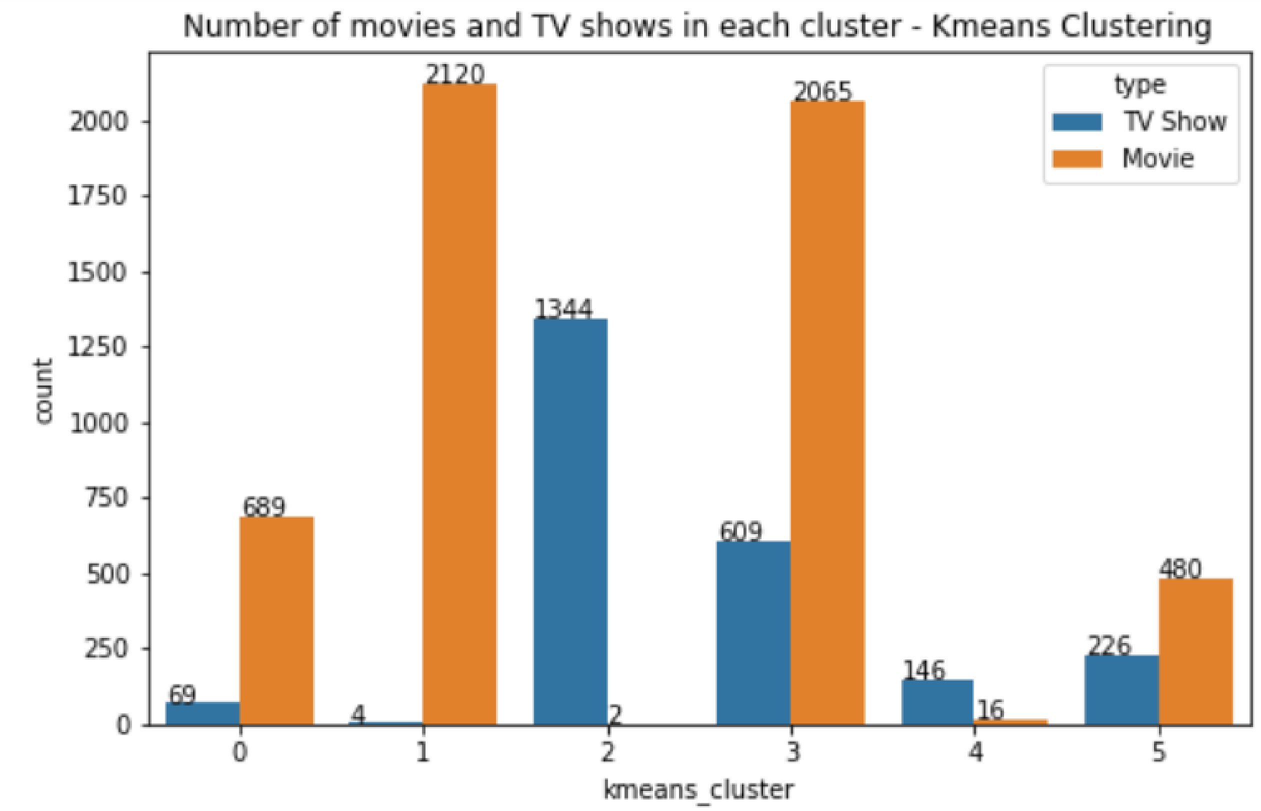


The sum of the squared distance between each point and the centroid in a cluster decreases with the increase in the number of clusters.

The highest Silhouette score is obtained for 6 clusters. Building 6 clusters using the k-means clustering algorithm.

Model Implementation:

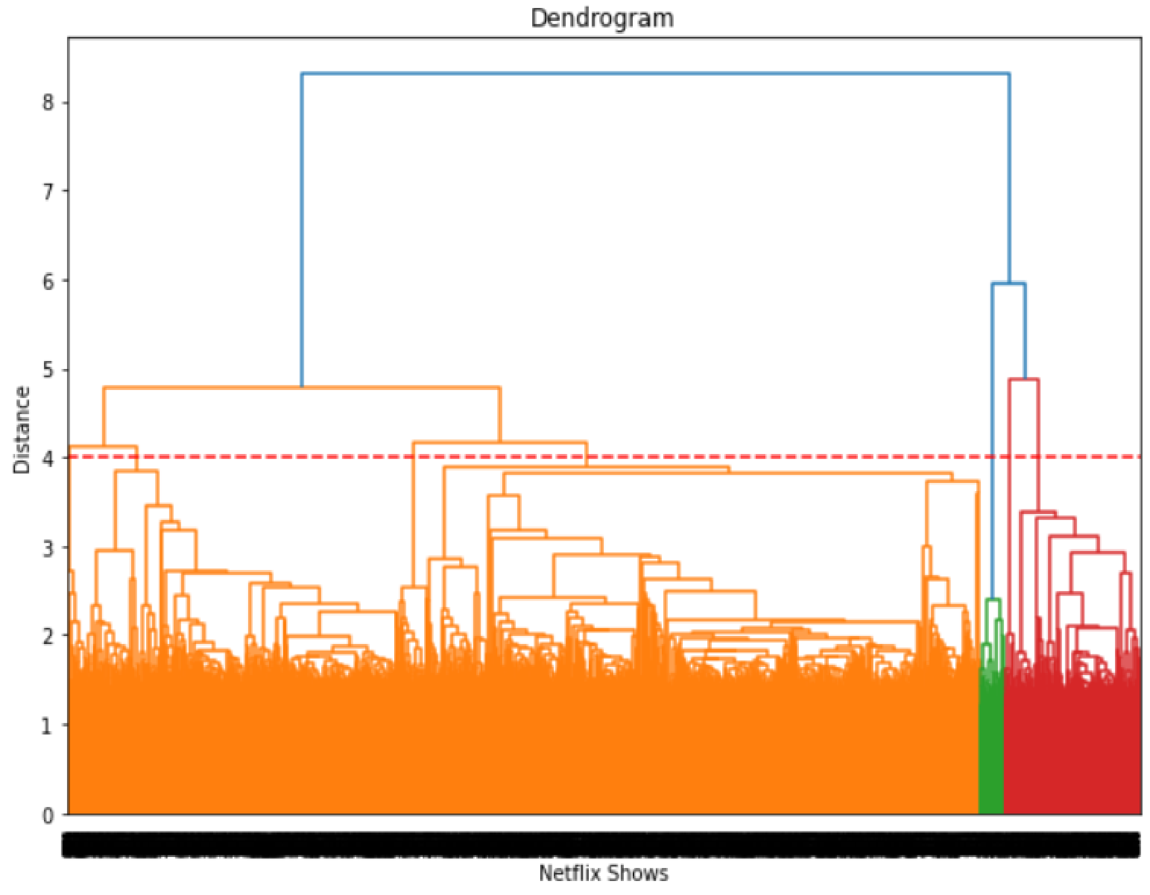
K-Means Clustering:



- We successfully built 6 clusters using the k-means clustering algorithm.
- In clusters 0, 1, 3, and 5 highest number of counts belong to the Movies class.
- Cluster 2 builds only on the TV show.
- We can use the word cloud to visualize the most important words in each cluster for each column.

Model Implementation:

Agglomerative Hierarchical Clustering:

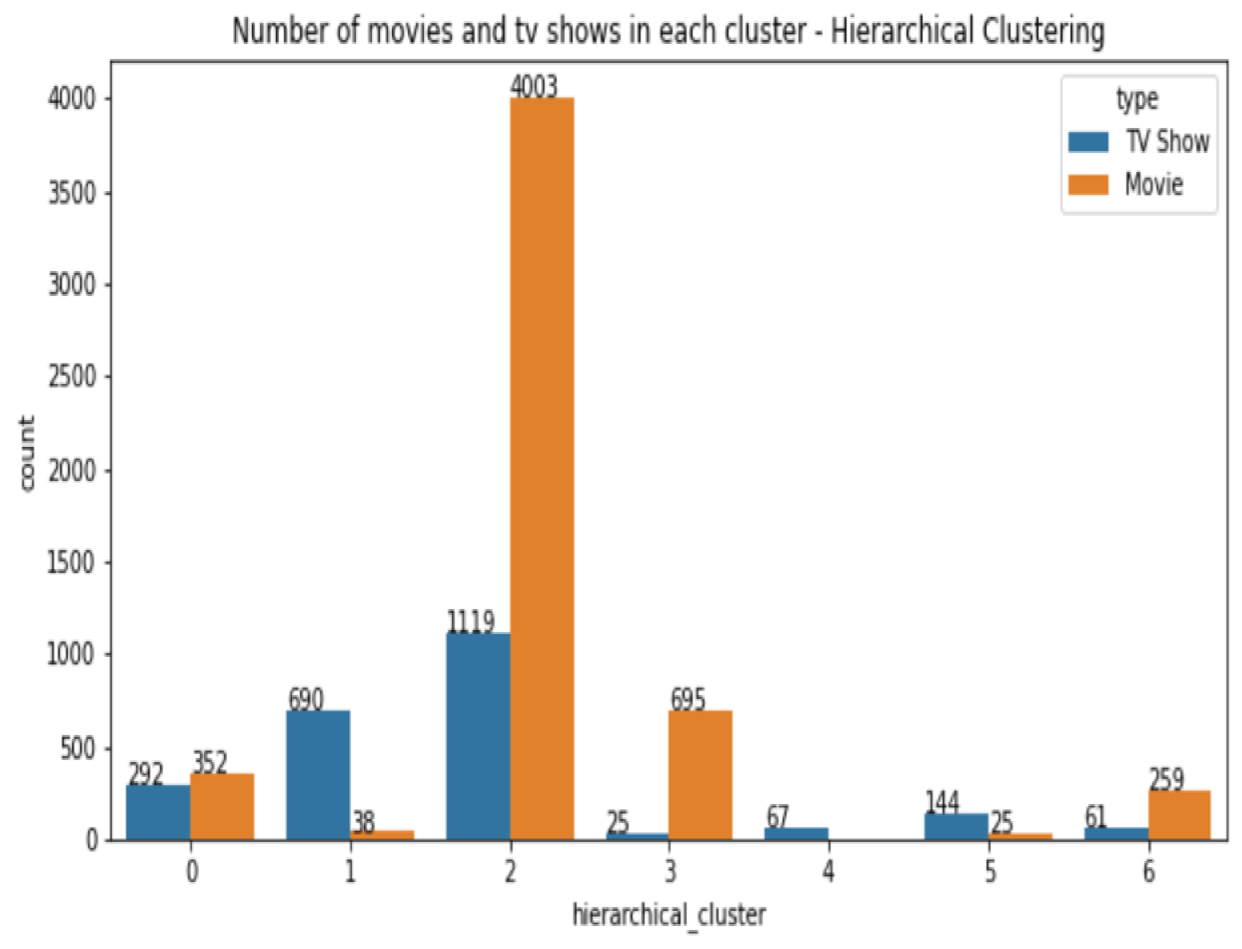


- We visualized the dendrogram to decide on the optimal number of clusters for the Agglomerative (hierarchical) clustering algorithm.

- At a distance of 4 units, 7 clusters can be built using the agglomerative clustering algorithm.

Model Implementation:

Agglomerative Hierarchical Clustering:

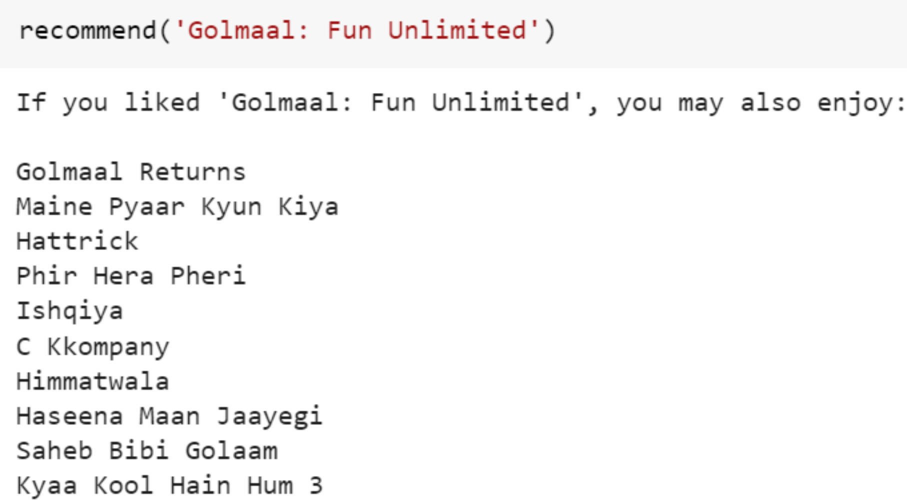


- We successfully built 7 clusters using the Agglomerative (hierarchical) clustering algorithm.

- Cluster 2 has the highest number of data points.

- Clusters 1, 4, and 5 have more TV shows than movies.

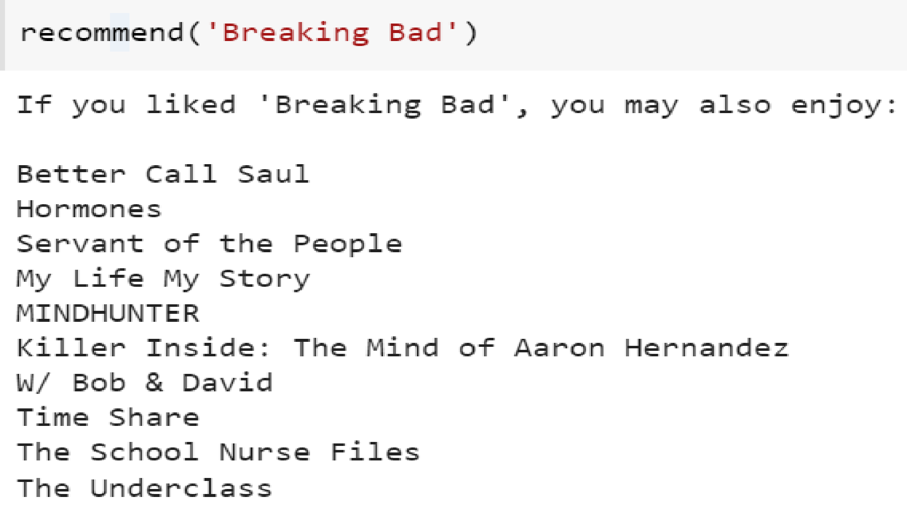
Recommendation System:



```
recommend('Golmaal: Fun Unlimited')
```

```
If you liked 'Golmaal: Fun Unlimited', you may also enjoy:

Golmaal Returns
Maine Pyaar Kyun Kiya
Hattrick
Phir Hera Pheri
Ishqiya
C Kkompany
Himmatwala
Haseena Maan Jaayegi
Saheb Bibi Golaam
Kyaa Kool Hain Hum 3
```



```
recommend('Breaking Bad')
```

```
If you liked 'Breaking Bad', you may also enjoy:

Better Call Saul
Hormones
Servant of the People
My Life My Story
MINDHUNTER
Killer Inside: The Mind of Aaron Hernandez
W/ Bob & David
Time Share
The School Nurse Files
The Underclass
```

- **Recommendation system suggests items to users based on their similarity to other items that the user has shown interest in.**

- **If a person has watched a show on Netflix, the recommender system must be able to recommend a list of similar shows that the user likes.**

- **To get the similarity score of the shows, we can use cosine similarity.**

- **The similarity between two vectors (A and B) is calculated by taking the dot product of the two vectors and dividing it by the magnitude value.**

AI

Conclusion:

In this project, we tackled a text clustering problem in which we had to categorize and group Netflix shows into specific clusters in such a way that shows in the same cluster are similar to one another and shows in different clusters are not.

- There were approximately 7787 records and 11 attributes in the dataset.
- We started by working on the missing values and conducted exploratory data analysis (EDA).
- It was discovered that Netflix hosts more movies than television shows on its platform, and the total number of shows added to Netflix is expanding at an exponential rate. Additionally, most of the shows were made in the United States.
- The attributes were chosen as the basis for the clustering of the data: cast, country, genre, director, rating, and description. The TF-IDF vectorizer was used to tokenize, preprocess, and vectorize the values in these attributes.
- 10000 attributes in total were created by TF-IDF vectorization.
- Principal Component Analysis (PCA) was used to solve the dimensionality issue. The total number of components was limited to 3000 because 3000 components were able to account for more than 80% of the variance.

Conclusion:

- We Utilized the K-Means Clustering algorithm, we first constructed clusters, and the optimal number of clusters was determined to be 6. The elbow method and Silhouette score analysis were used to get the optimal number of clusters.

- The Agglomerative clustering algorithm was then used to create clusters, and the optimal number of clusters was determined to be 7. This was obtained after visualizing the dendrogram.

- The similarity matrix generated by applying cosine similarity was used to construct a content-based recommender system. The user will receive 10 recommendations from this recommender system based on the type of show they watched.

Thank you