Capstone Project

Netflix Movies and TV Shows Clustering

Unsupervised Machine Learning

Presented By:

Aman Verma

Table Of Contents

1. Defining problem statement
2. Data Cleaning & visualization
3. Data Pre-processing
4. Feature Selection
5. Applying different clustering methods
6. Applying Clustering Models
7. Conclusion

Problem Statement

Netflix is the world's largest online streaming service provider, with over 238.4 million subscribers as of Q2 2023. It is crucial that they effectively cluster the shows that are hosted on their platform in order to enhance the user experience, thereby preventing subscriber churn.

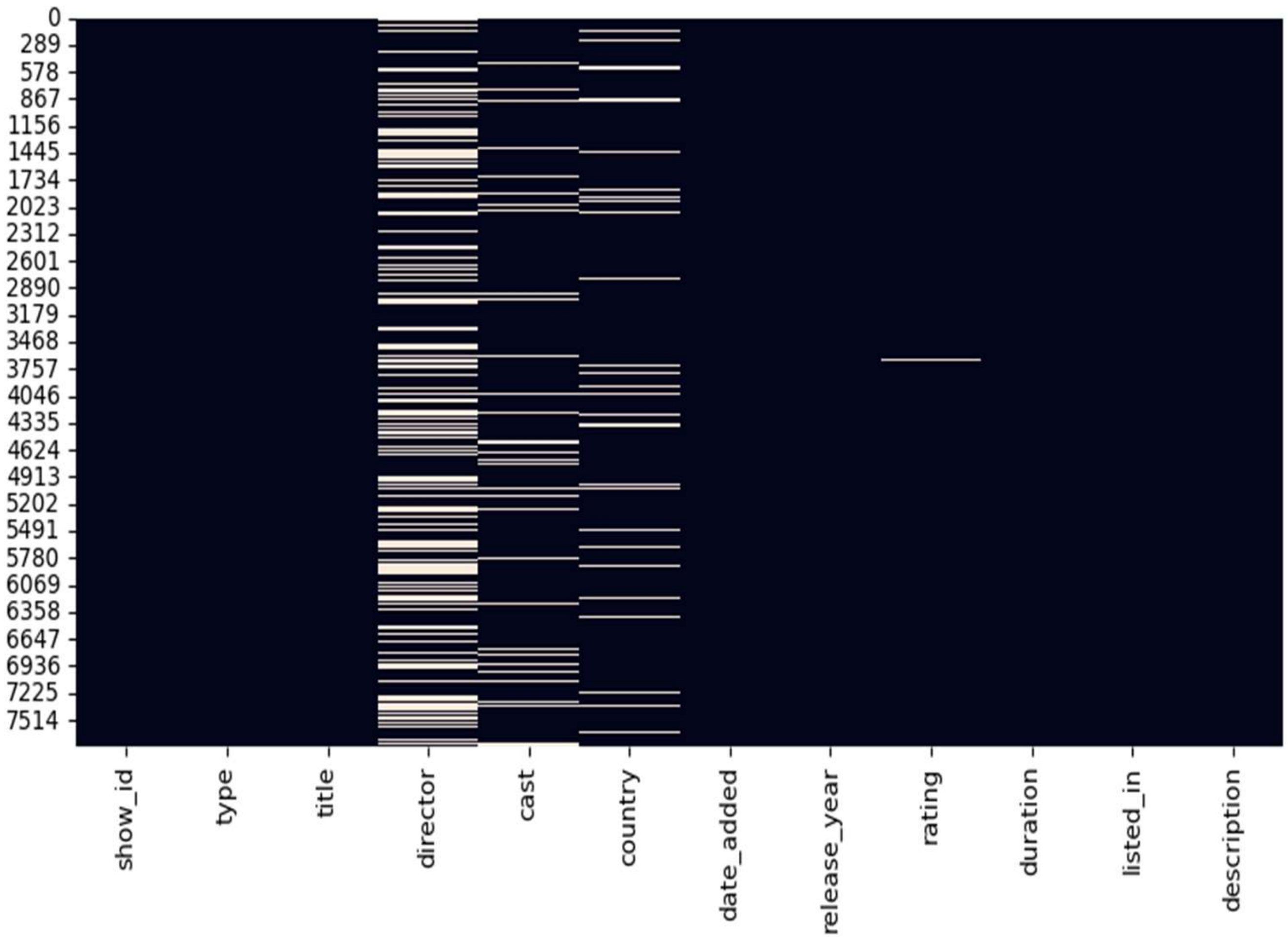
We will be able to understand the shows that are similar to and different from one another by creating clusters, which may be leveraged to offer the consumers personalized show suggestions depending on their preferences.

The goal of this project is to classify/group the Netflix shows into certain clusters such that the shows within a cluster are similar to each other and the shows in different clusters are dissimilar to each other.

Data Summary

- show_id: Unique ID for every Movie / Tv Show
- type : 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 : Genres
- description: The Summary description

EDA (Checking NaN values)



- **Null values present in these columns**

○ director

○ cast

○ country

○ Rating

- **No missing value present in these columns**

○ show_id

○ type

○ title

○ date_added

○ release_year

○ duration ○ listed_in

○ description

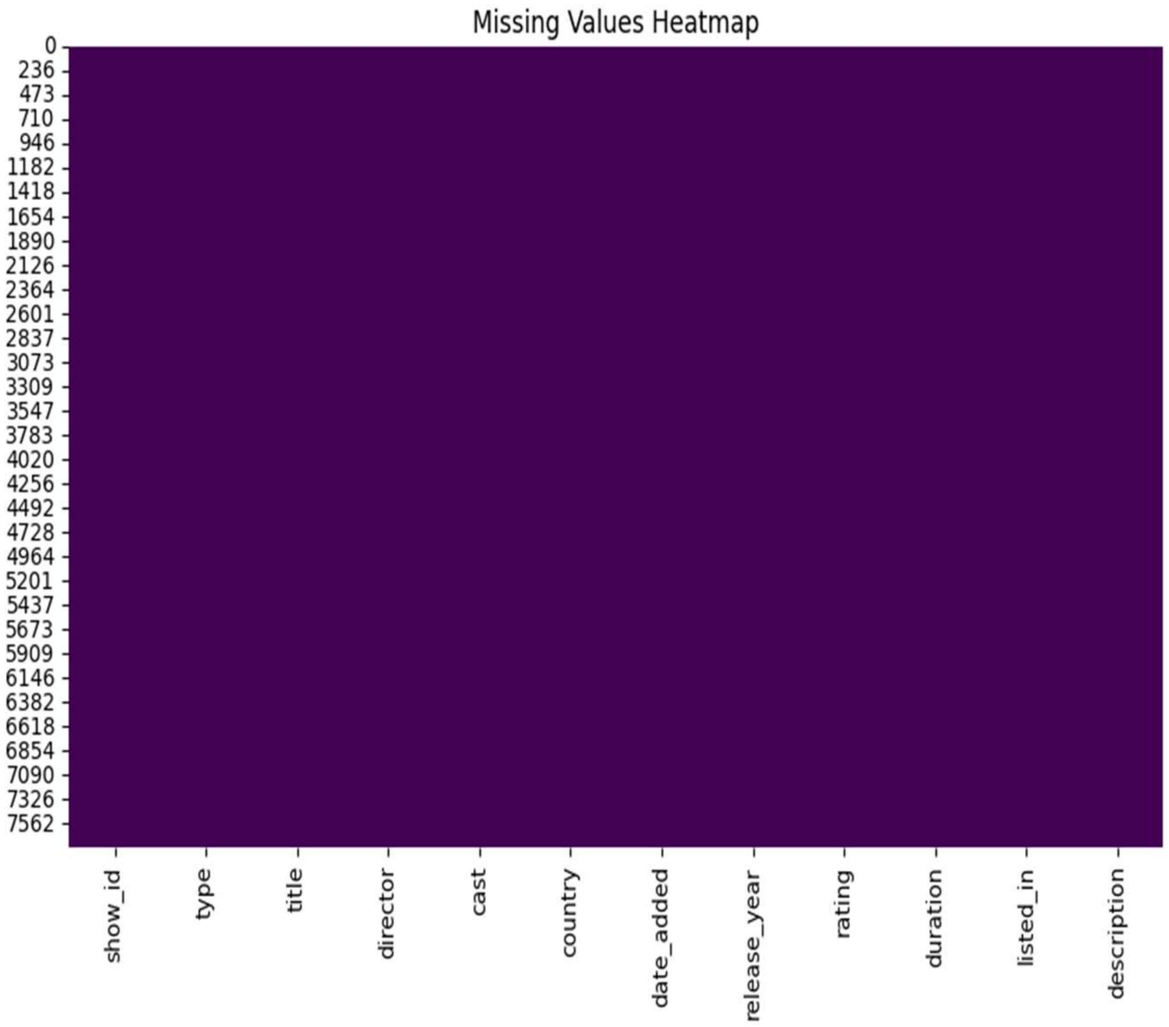
Dealing with missing values

1. Fill missing values in 'director', 'cast', and 'country' columns with 'Unknown’.

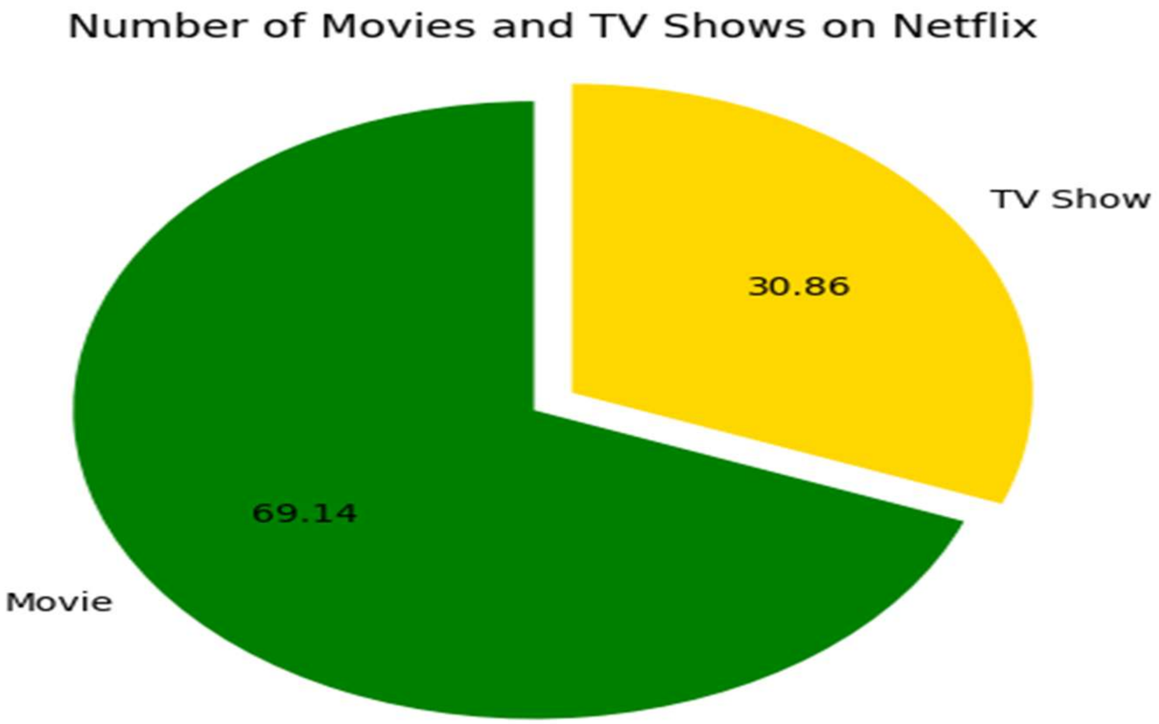
2. Fill missing values in the 'rating' column with the mode (most frequent value) from the same column.

3. Drop any remaining rows with missing values from the DataFrame.

4. Clearly we can see in this Heatmap, we’ve deal all the missing values

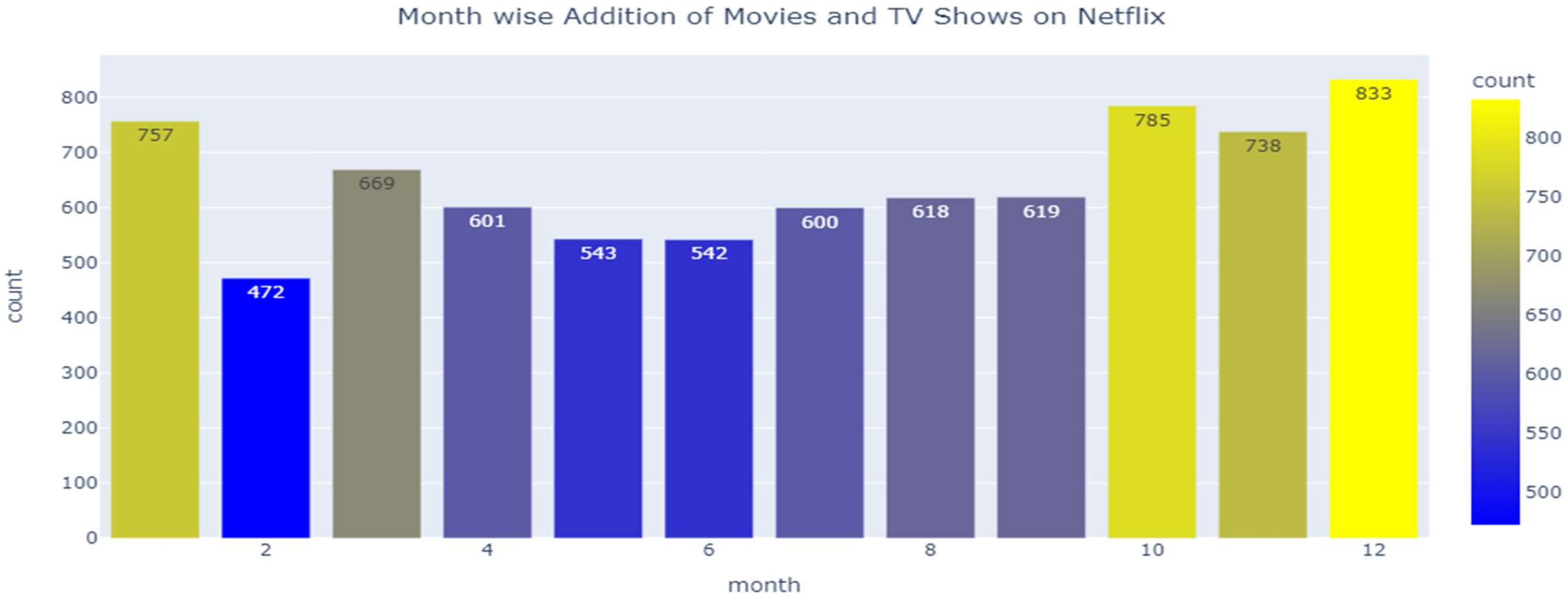


TV shows or Movies ??



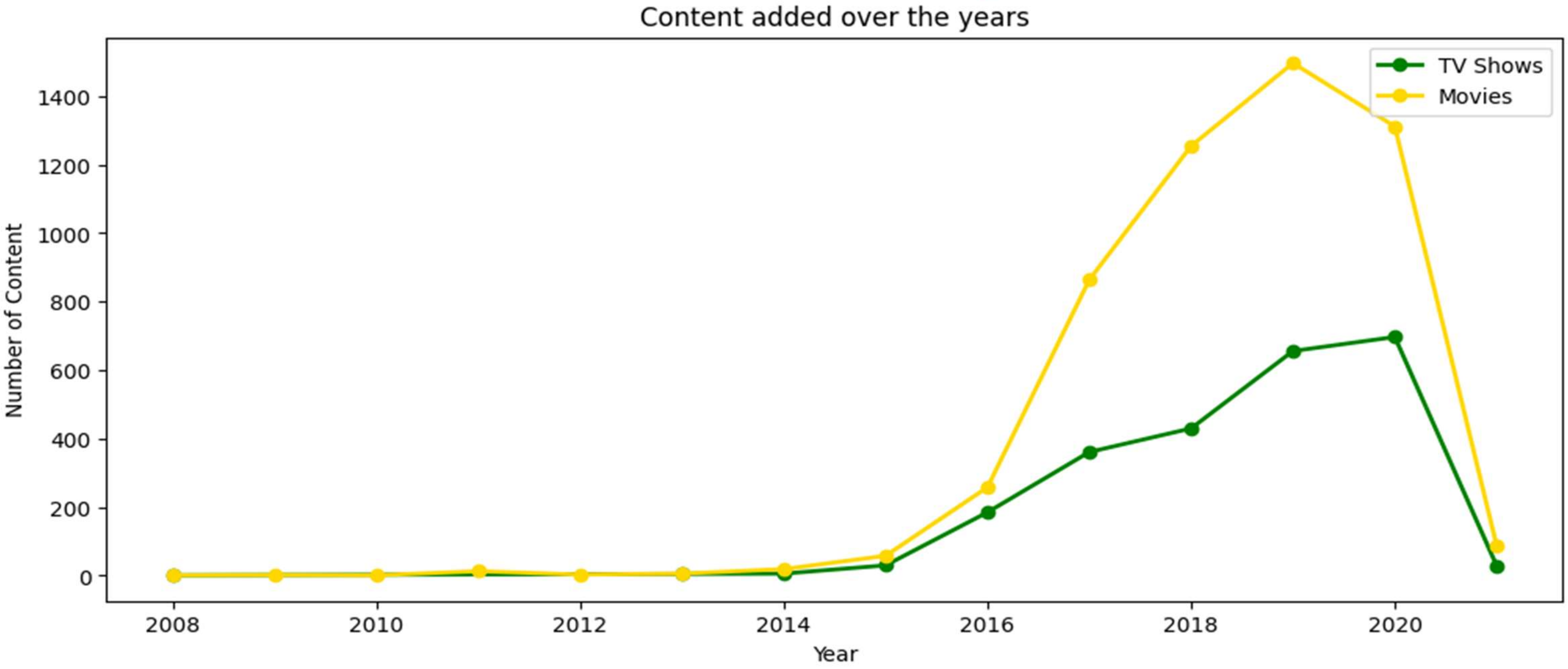
- Most of the contents are Movies
- Less than ⅓ content are TV Shows

The month-wise addition of movies and TV shows on Netflix



Netflix experiences a significant rise in TV shows and movie releases during the holiday months of October to December, coinciding with various festivities like Halloween, Diwali, Dusherra, and Christmas, when people tend to spend more time at home and seek entertainment.

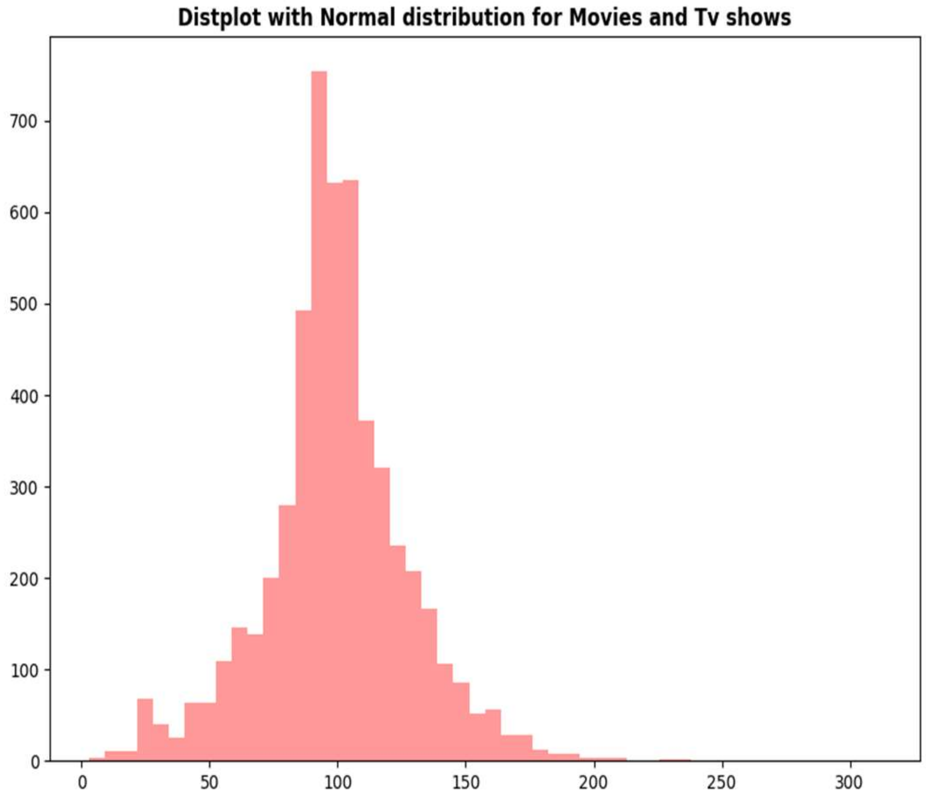
Content added over the years



From 2008 to 2015, Netflix had fewer new TV shows and movies. In 2016, additions started to increase. 2019 saw a big spike in movies, while TV shows also rose but less significantly than movies.

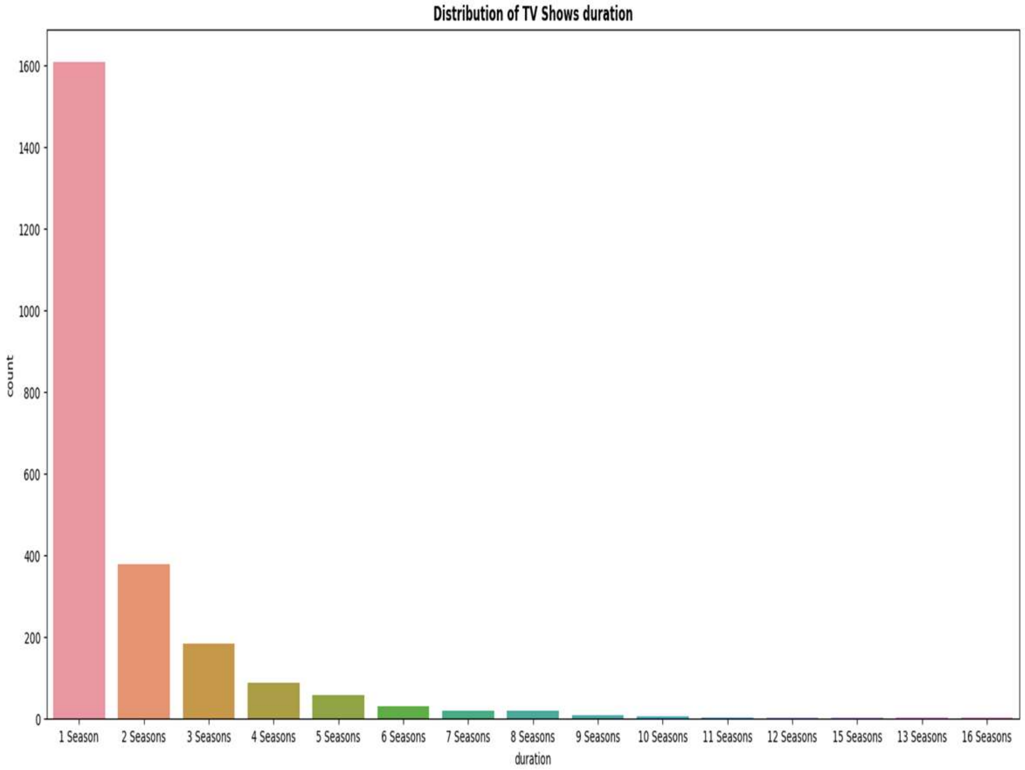
Netflix's content demand is on the rise, providing an opportunity to boost user satisfaction and engagement.

Duration



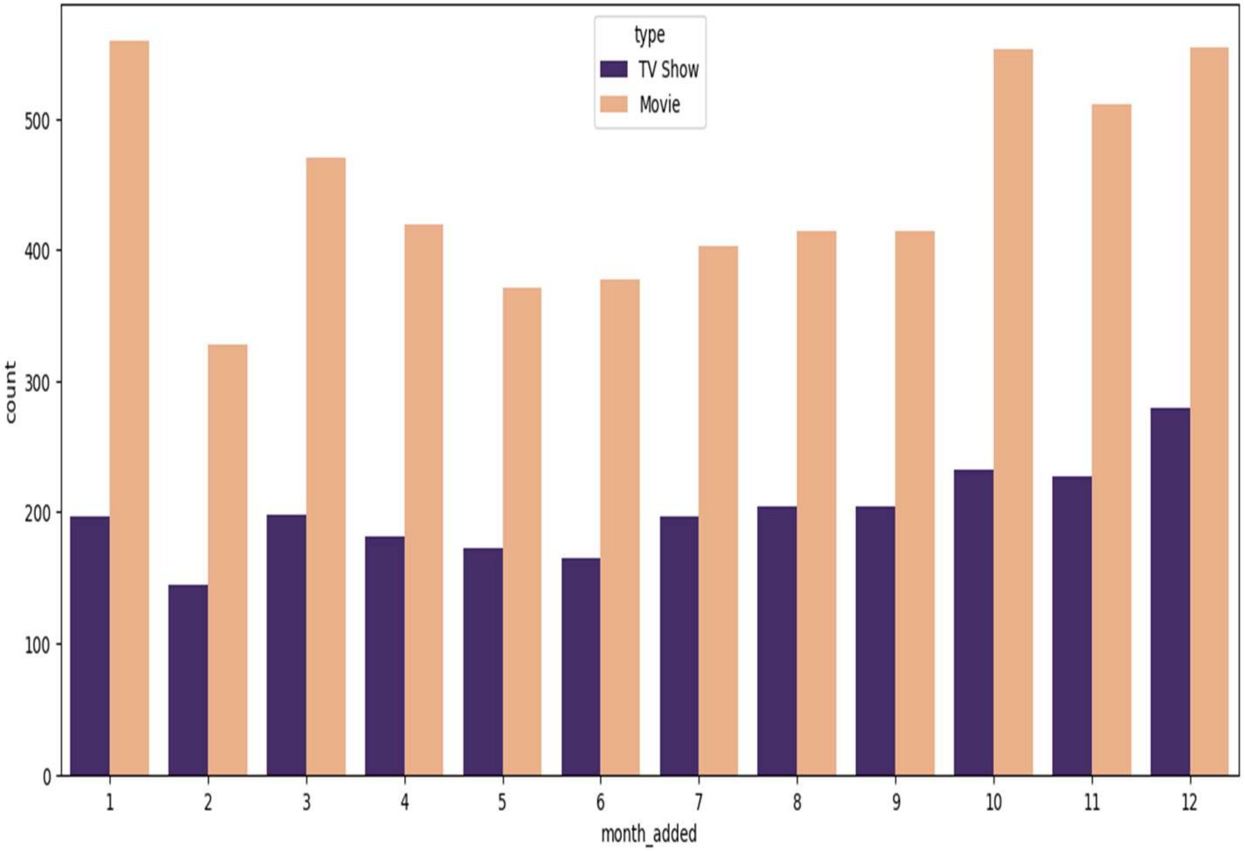
- most of the movies have duration of between 50 to 150

TV Show Duration Distribution



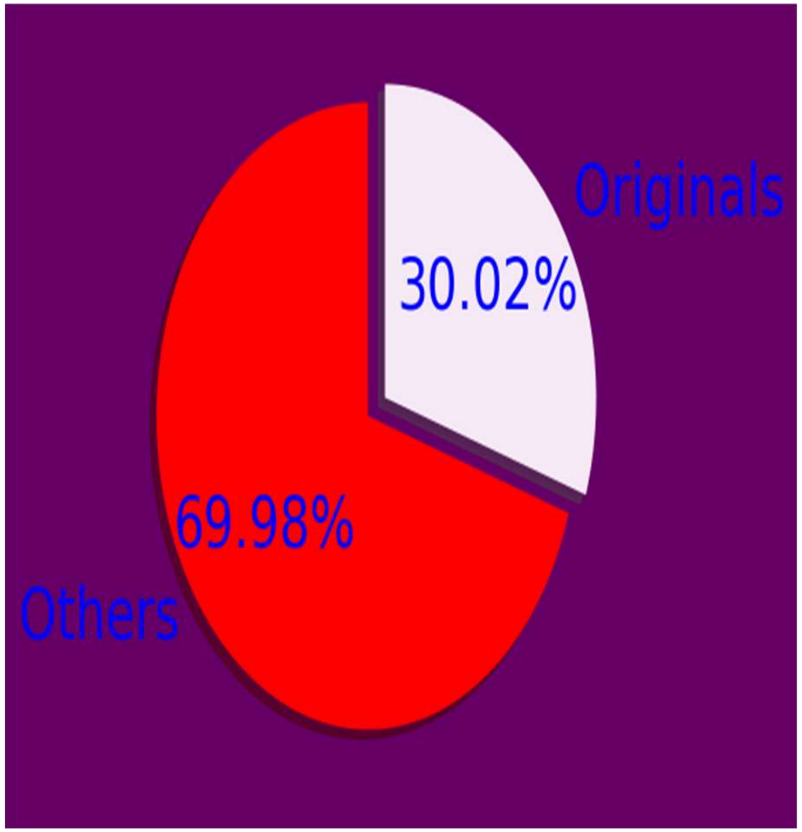
- highest number of tv_shows consisting of single season

Netflix content addition by month – movie show & TV show



January, October, and December are the prime months for Netflix movie additions, while for TV shows, it's October, November, and December that stand out as trending months.

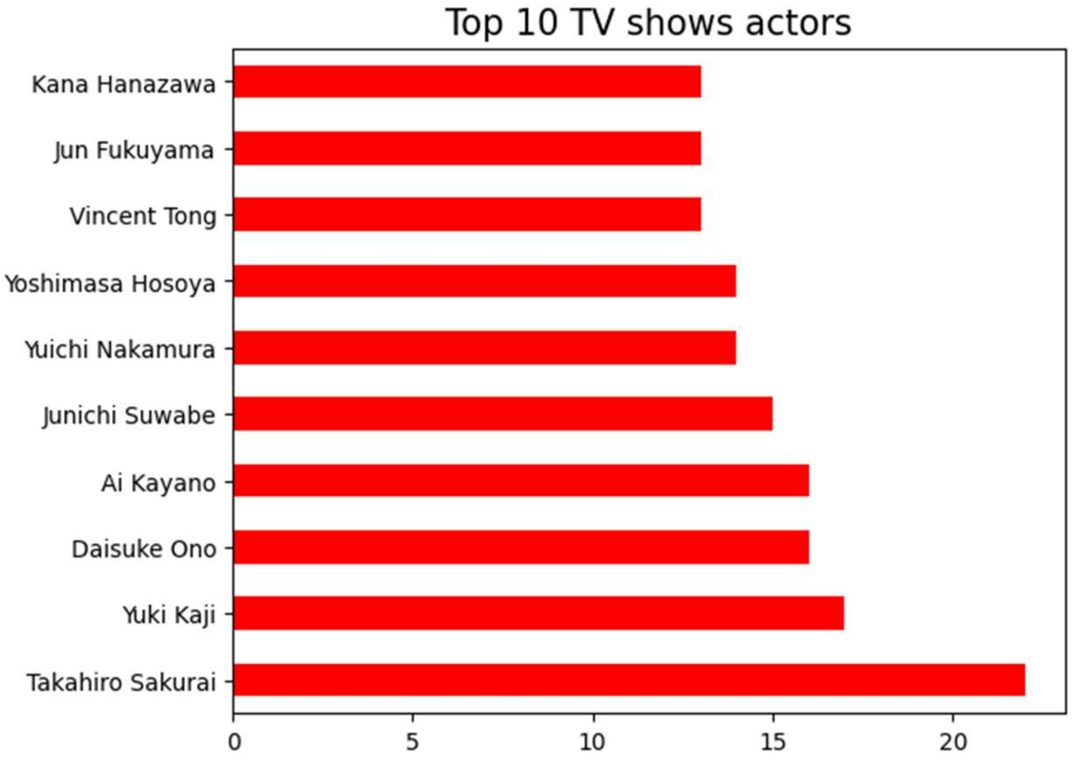
Originals

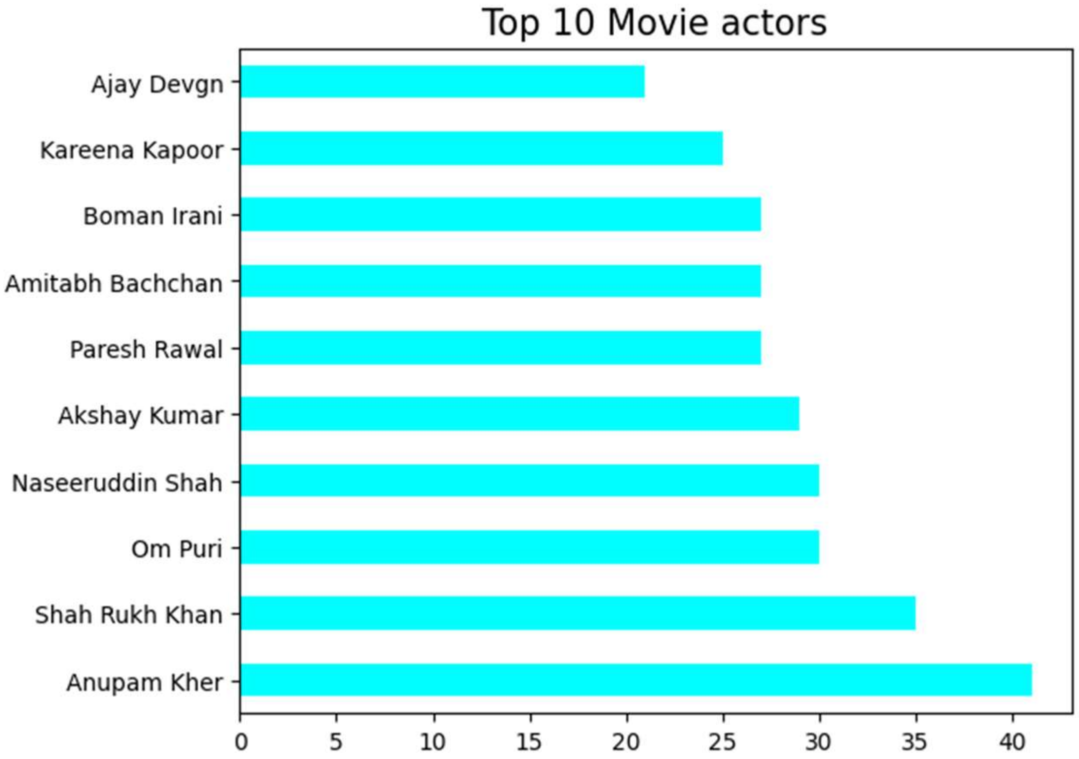


30% movies released on Netflix.

70% movies added on Netflix were released earlier by different mode.

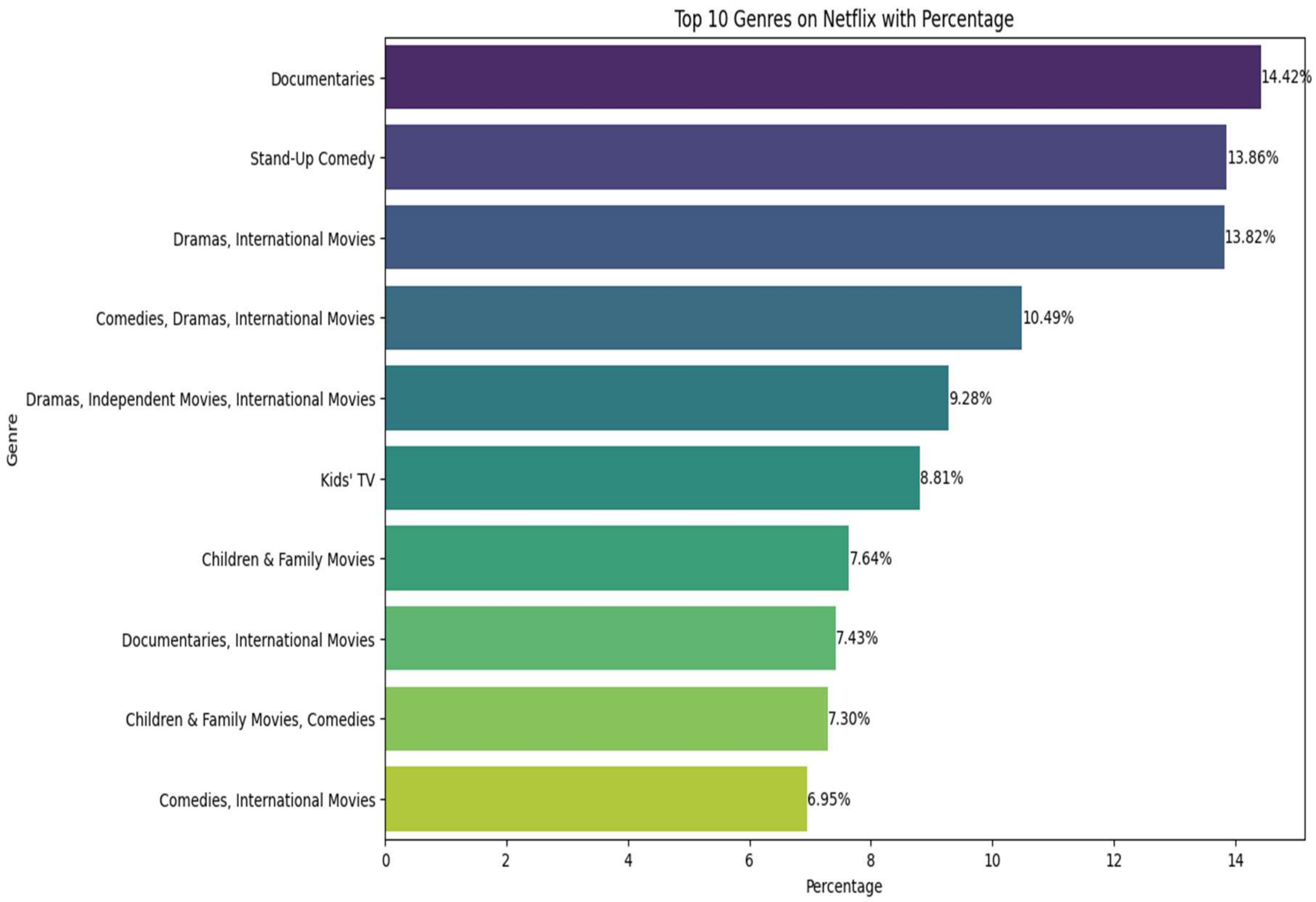
Top 10 Actors in Netflix TV Shows and Movies





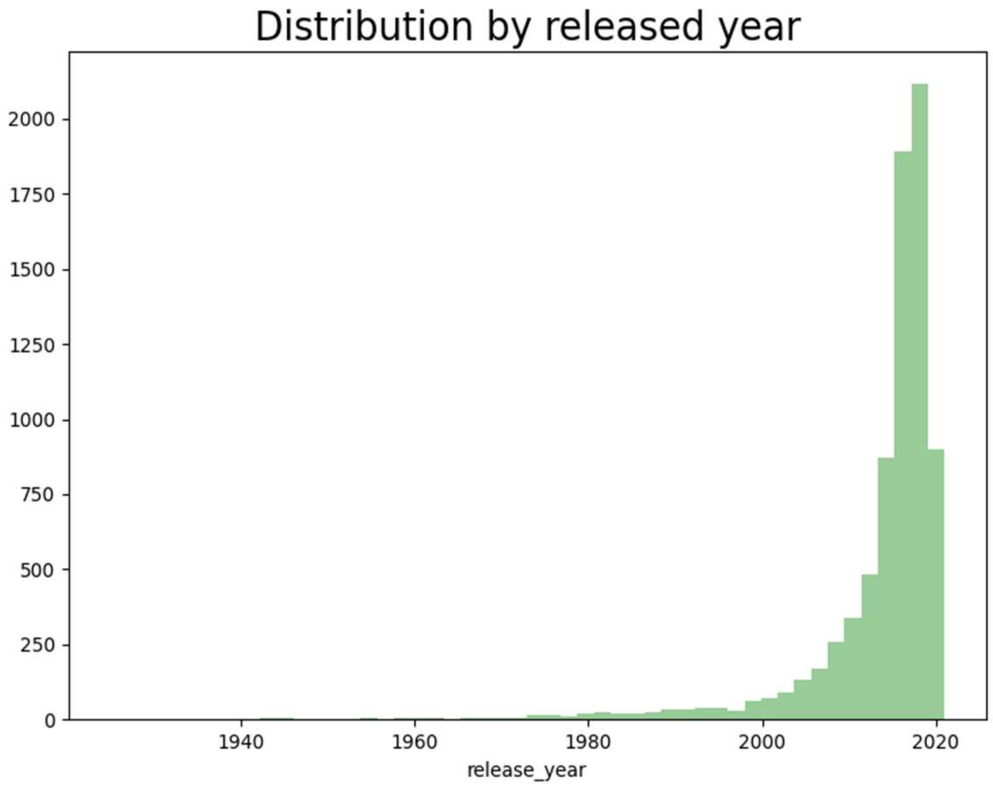
Takahiro Sakurai tops the list for TV shows, while Anupam Kher leads in the movies category with the most appearances on Netflix.

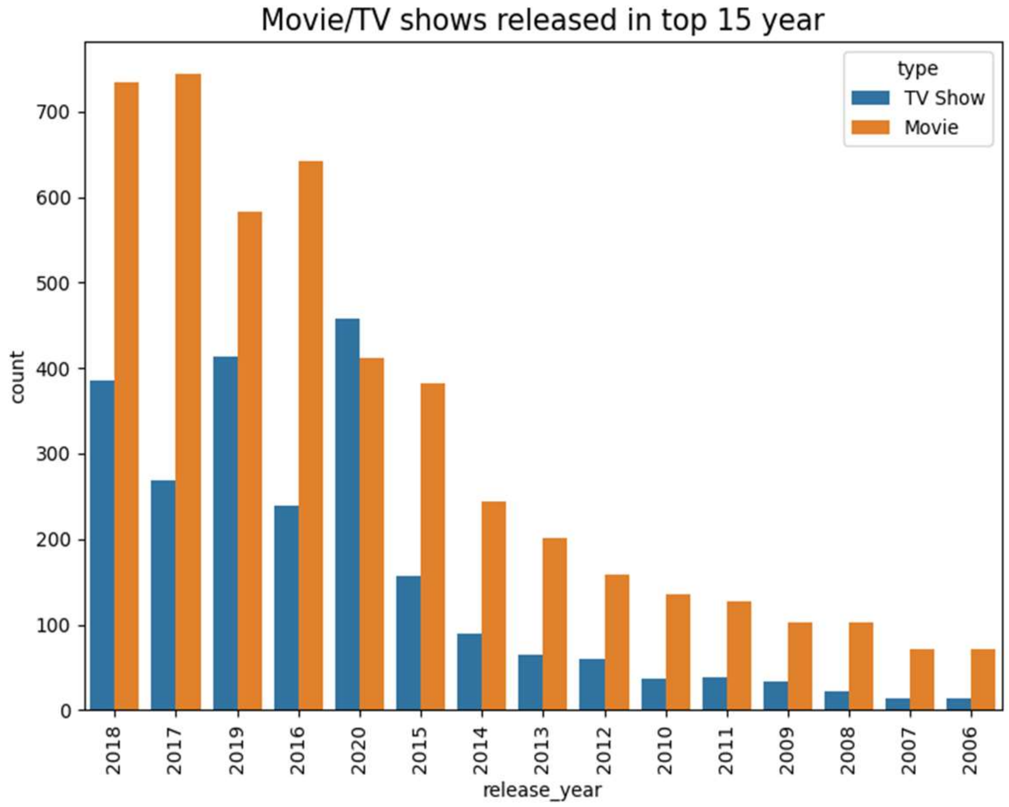
Genre Top 10



Documentaries hold the top position as the most prevalent genre on Netflix, followed closely by stand-up comedy, dramas, and international movies.

Netflix Content Distribution by Release Year and Top 15 Years for Movies and TV Shows



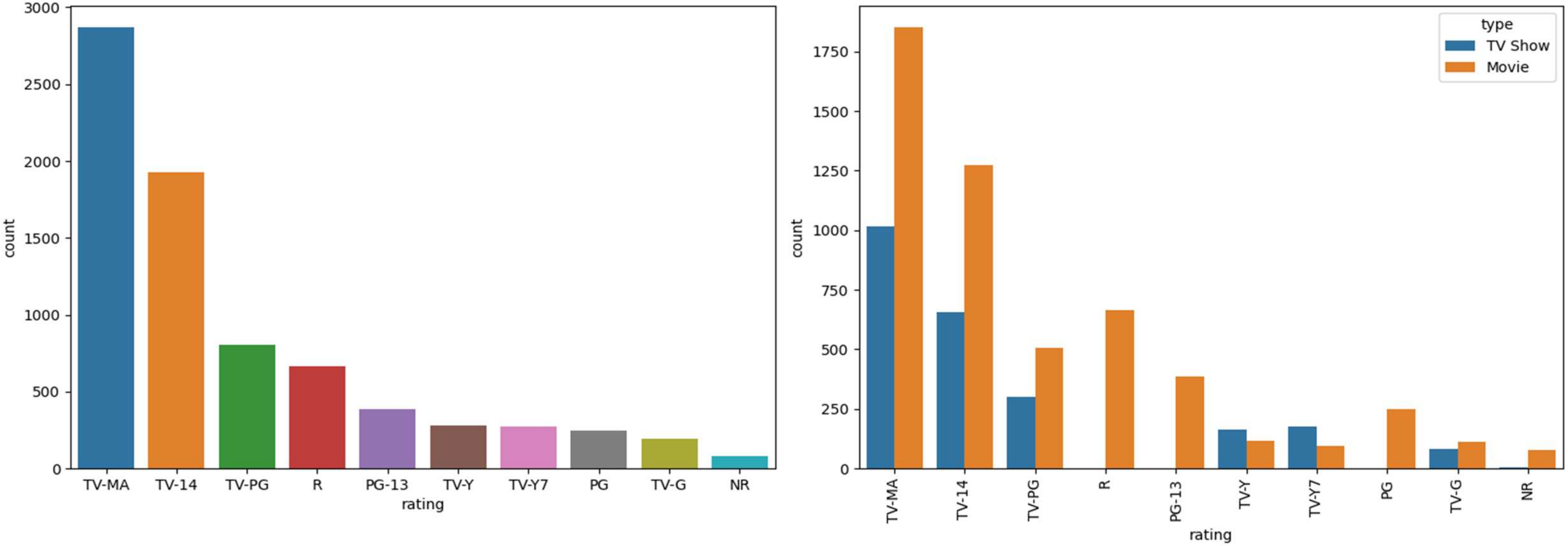


The histogram suggests a growing trend in Netflix movie releases from around 1980, with a notable surge from 2000 onwards.

The bar graph shows peaks for both movies and TV shows in 2017 and 2020, indicating a substantial amount of content added during those years.

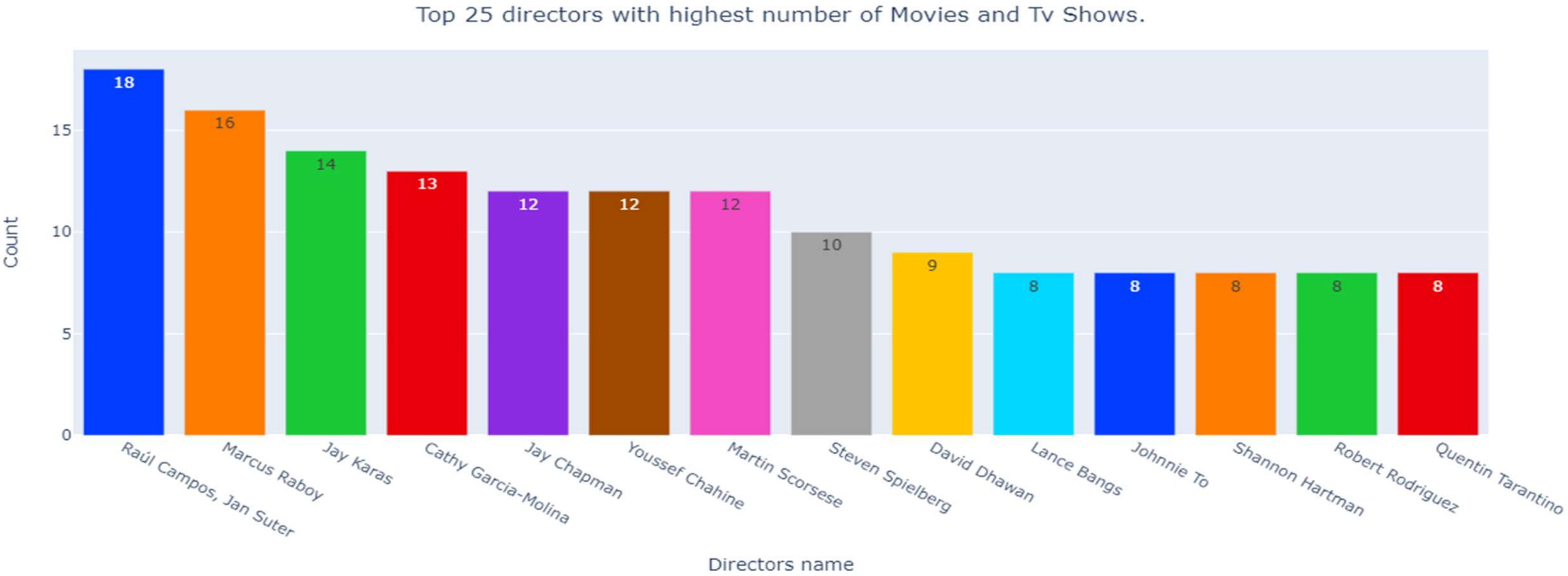
Netflix Content Ratings: Movies vs. TV Shows

Top 10 rating for different age groups and audiences & Rating based on Movie and Tv_Shows



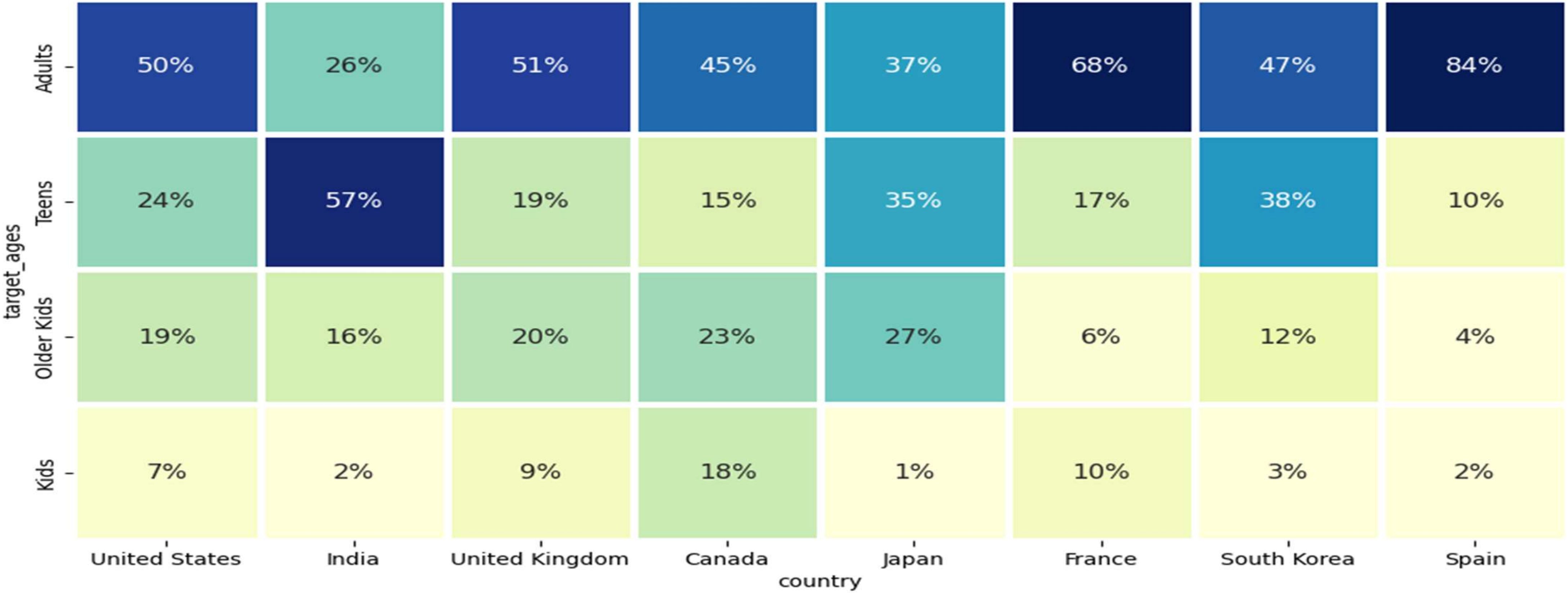
- The chart combination allows for a comprehensive analysis of ratings. TV-MA, which signifies content for mature audiences, emerges as the predominant rating for both movies and TV shows.
- In terms of ratings, the most common rating is **TV-MA**, which applies to both movies and TV shows.

Top Directors by Netflix Movie/TV Counts



The directors Raúl Campos and Jan Suter have the highest count in terms of overall Movies and TV shows on Netflix.

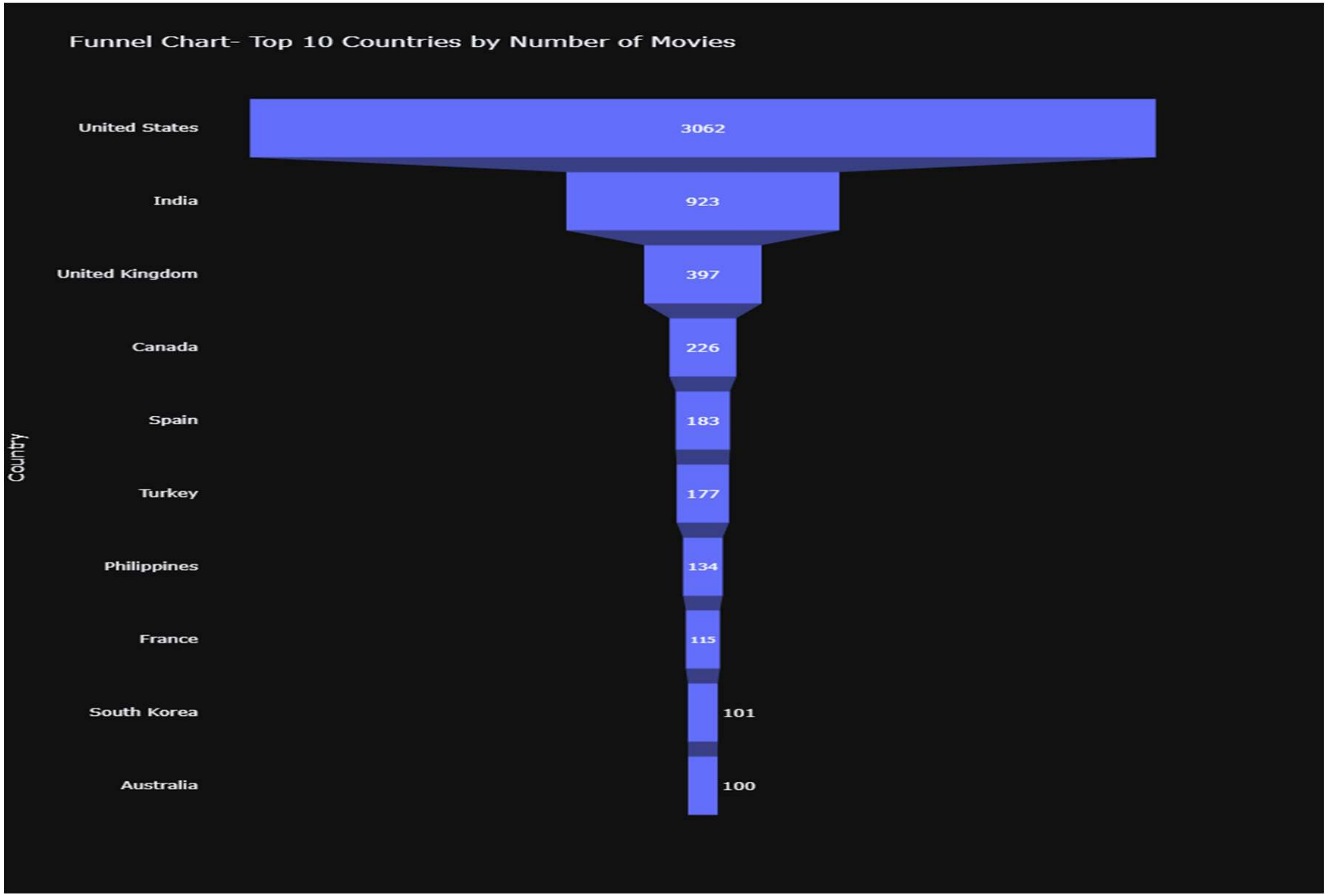
Heatmap



Interest in the subject varies across countries and age groups.

Spain and France lead with strong adult engagement, while India tops in teenage interest. The data underscores diverse preferences, with Mexico and Canada representing significant variations in subject appeal.

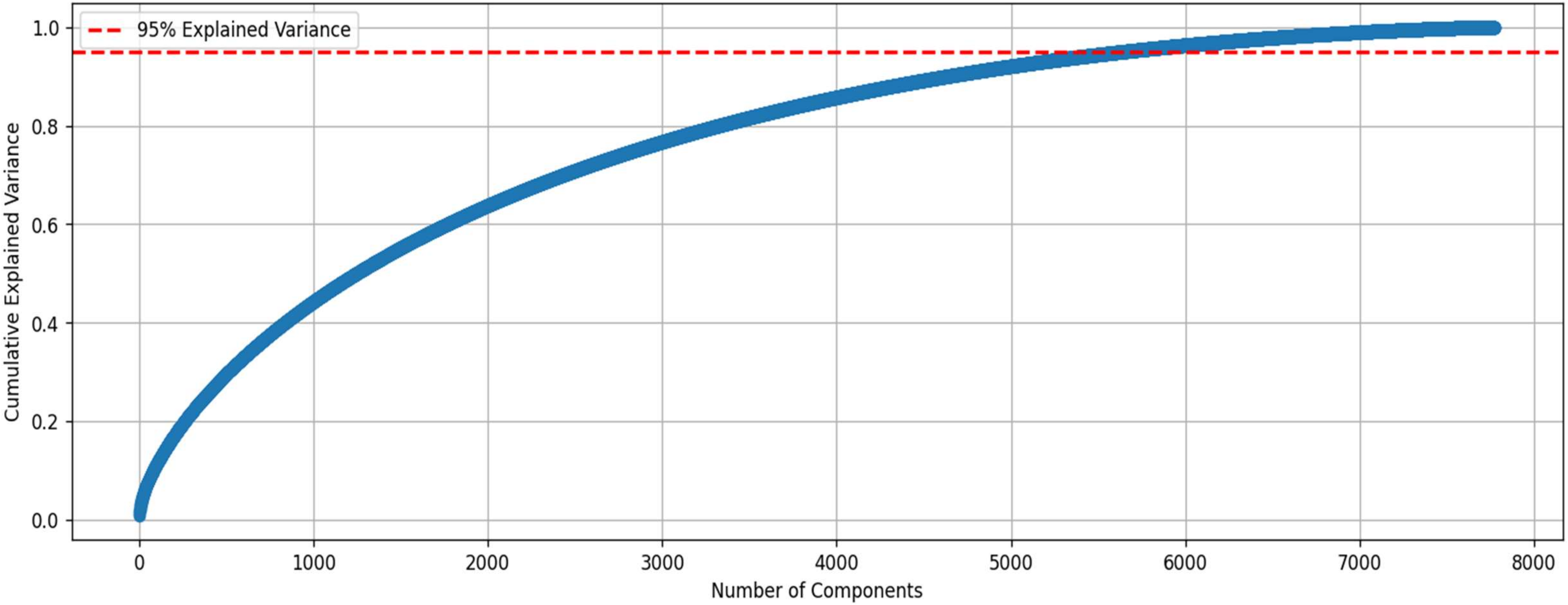
Top 10 Countries by Number of Movies - A Funnel Chart



The United States has the highest number of movies, with 3062 films, indicating a dominant presence in the film industry.

India is the second-highest contributor with 923 movies, demonstratin

Explained Variance Analysis using PCA

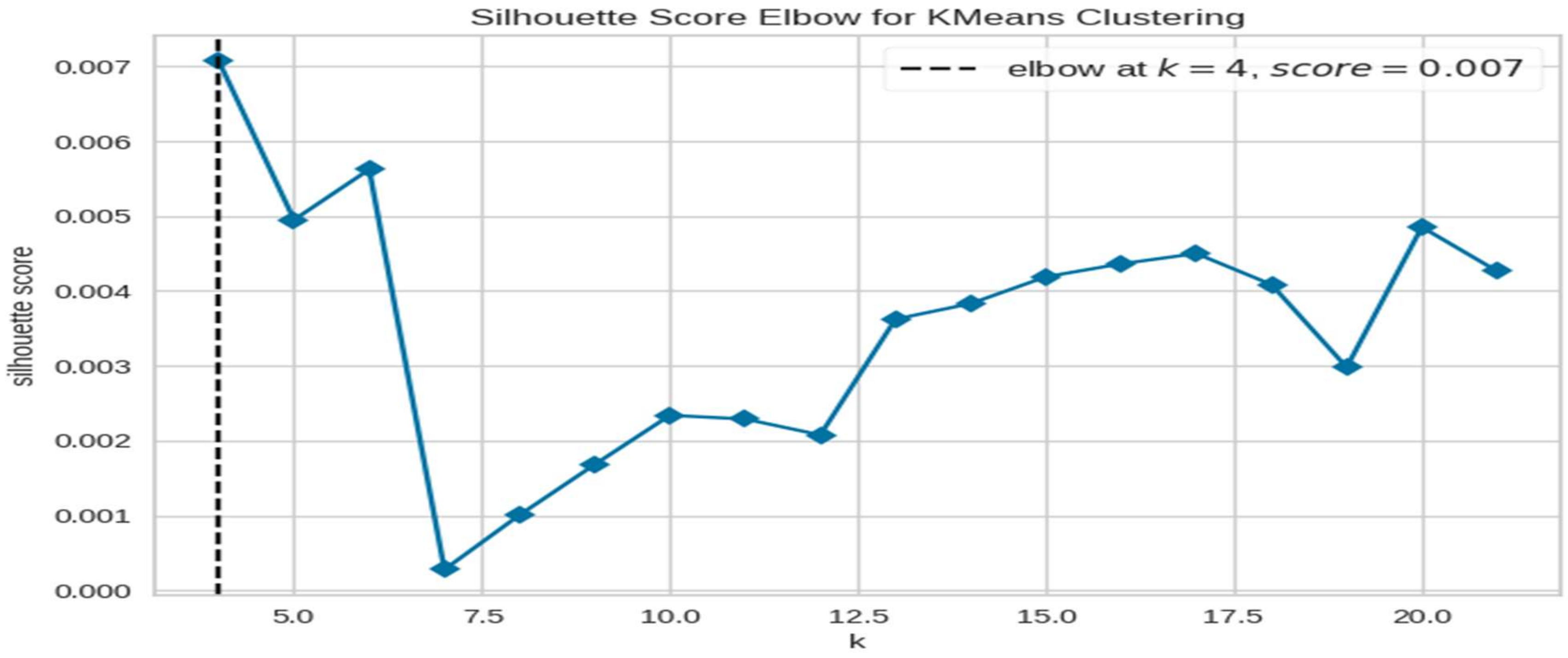


Use the plot to identify the ideal number of components for dimensionality reduction.

Aim for a satisfactory cumulative explained variance, typically set at 95%. Look for the point where the curve meets or is closest to this threshold line to guide your component selection.

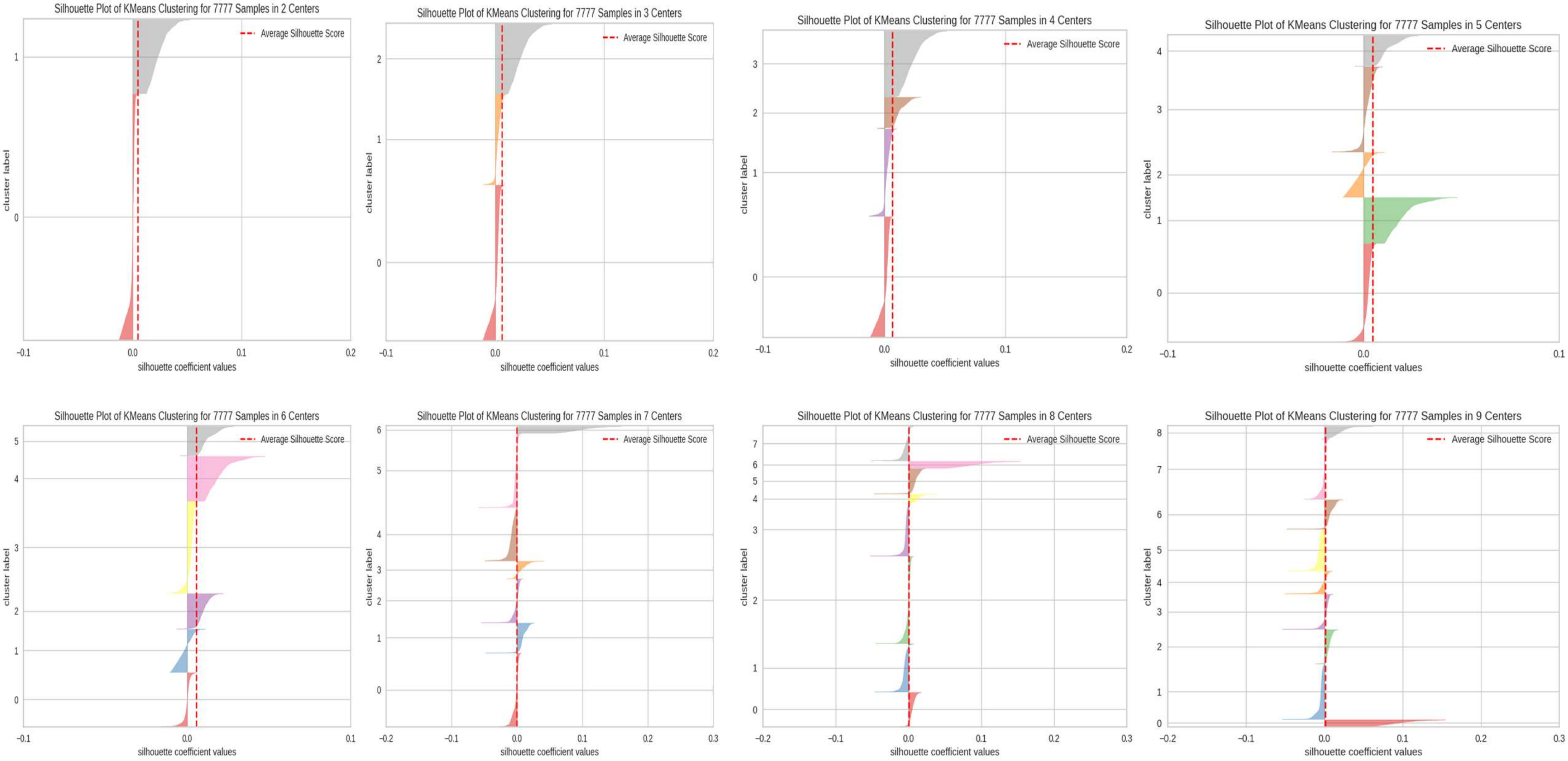
ML Model Implementation

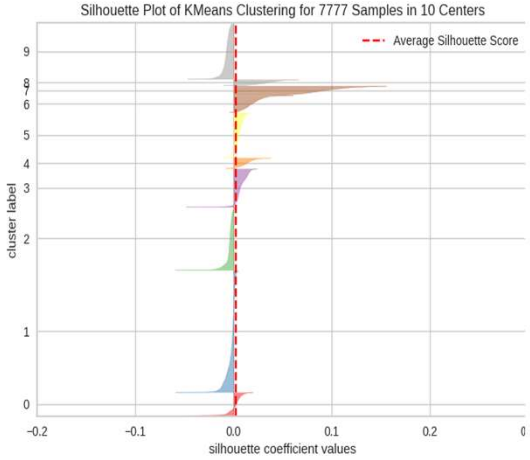
Optimal Cluster Determination

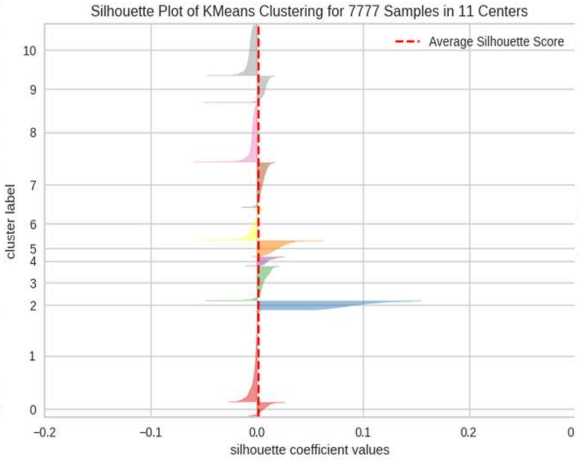


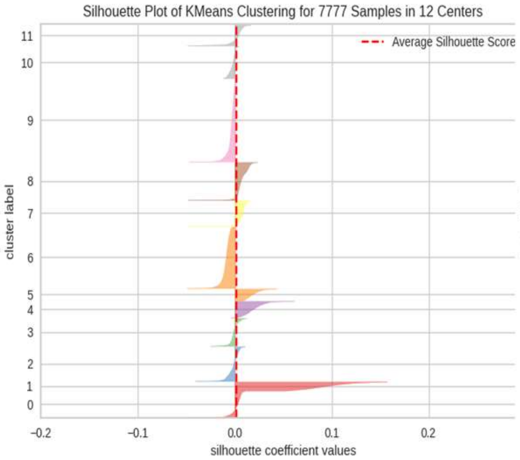
The plot will reveal the 'elbow' point, signifying the recommended number of clusters according to the chosen metric. In this instance, the elbow plot suggests the ideal number of clusters is 5

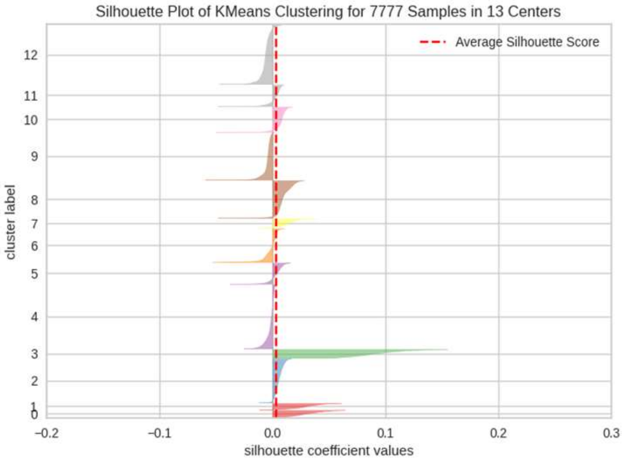
Clustering Analysis with Silhouette Scores: Finding Optimal Clusters and Visualizing Cluster Separation

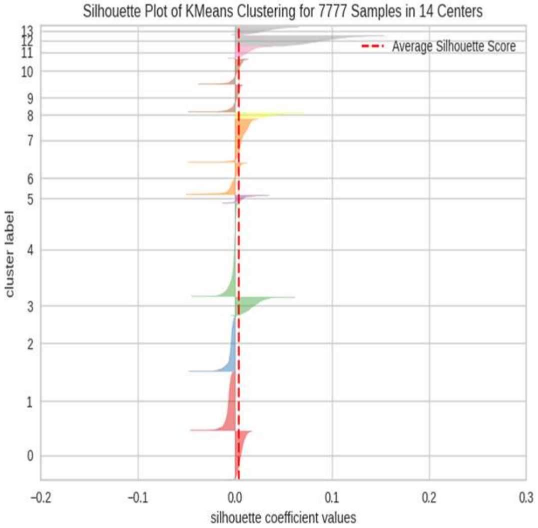










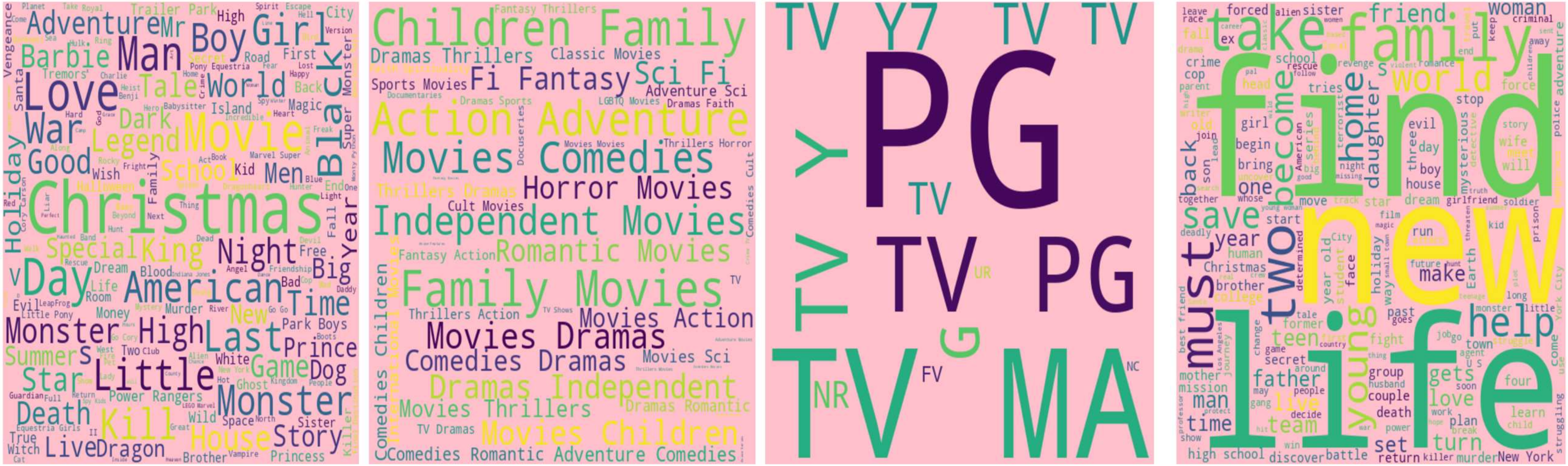


1.Silhouette Score Analysis: The code iterates through a range of cluster numbers (from 2 to 'n') and calculates the silhouette score for each number of clusters. The silhouette score provides a measure of the quality of clustering, with higher scores indicating better-defined clusters.

2.Optimal Number of Clusters: By printing the silhouette scores for different cluster numbers, you can identify the number of clusters that yields the highest silhouette score. The cluster number corresponding to the highest silhouette score is often considered the optimal number of clusters for your dataset.

3.Silhouette Visualizations: The code also generates silhouette visualizations for each clustering configuration. These visualizations provide a graphical representation of how well data points within each cluster are separated from data points in neighboring clusters. This can help you understand the compactness and separation of clusters.

Cluster-Based Word Cloud Analysis For Netflix movie & TV show



Cluster 9 in a dataset contains a total of 232 words. The most frequently occurring words in this cluster are as follows:

Type - Movie & Tv shows

Title - Broadway, Remastered, Christmas, Friends Orchestra

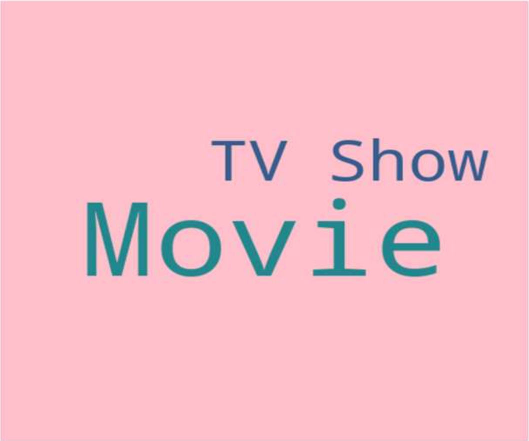
Country- United Kingdom,Argentina,United States,India

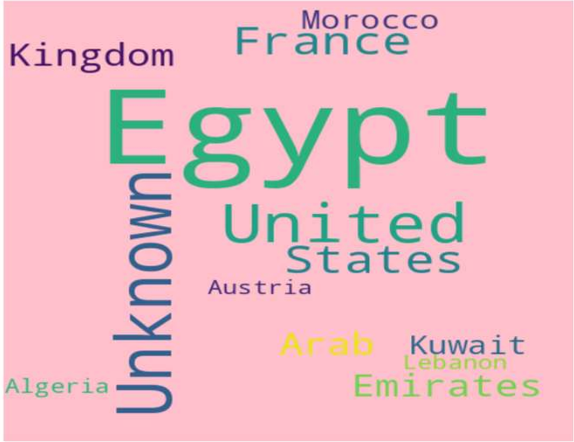
Rating -TV-MA, PG-TV

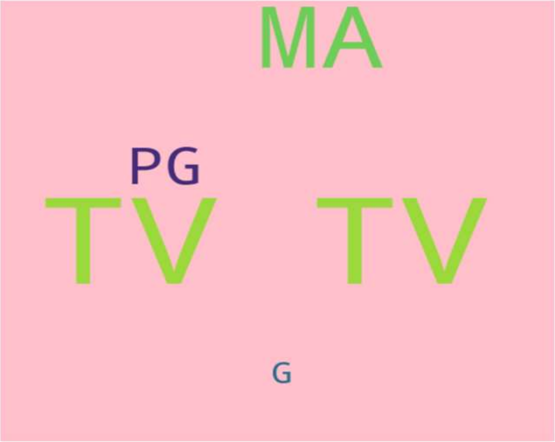
Listed_in - Family movies, Dramas International,Musical Dramas,Musicial Documentaries, Comedies International

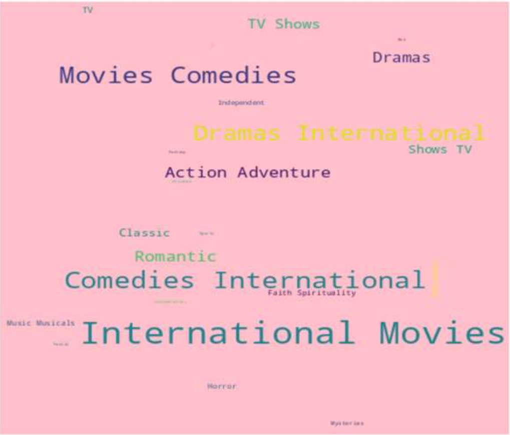
Description- Documentary, Music, One, Bad, Tour, Love.

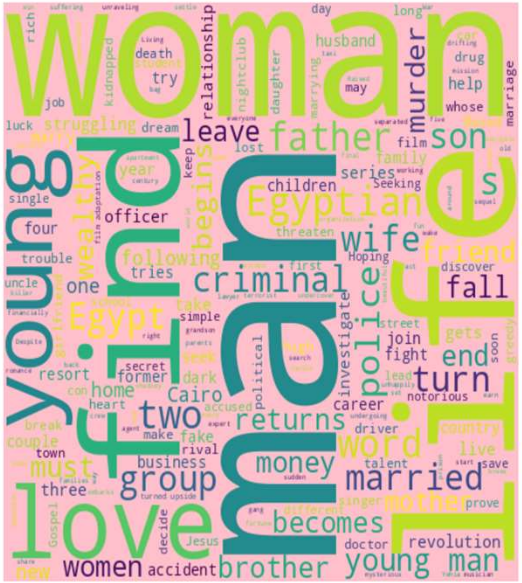
Cluster - Based Word Cloud Analysis

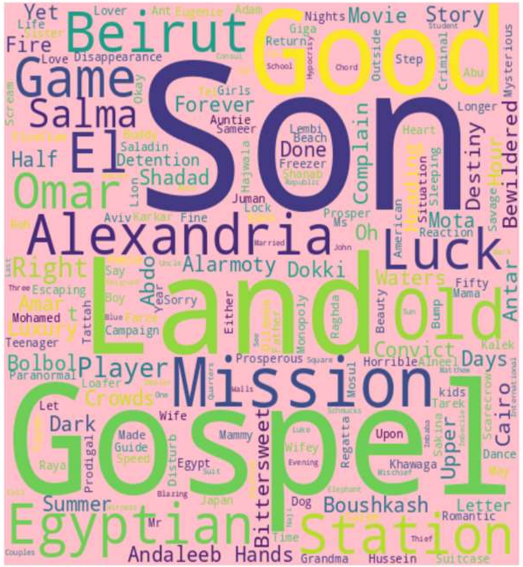












The most frequently occurring words in this cluster are as follows:

Type - Movie & Tv shows

Title - Special, America,Time,Live,Comedy, Netflix Alive, Martin

Country - United States,Brazil,Mexico,Italy

Rating -TV-MA,TV-PG

Listed_in - Tv-Comedies, Comedy Stand, Talk shows

Description- Stand Comedy, Comic, Take, Life, Live, Share,Stories.

Conclusion:

1- Exploring the dataset consist of 7777 records and 12 attributes, with a focus on missing value imputation and exploratory data analysis (EDA).

2- The analysis revealed that Netflix has a greater number of movies than TV shows, with a rapidly growing collection of shows from the United States.

3- It is interesting to note that the majority of the content available on Netflix consists of movies. However, in recent years, the platform has been focusing more on TV shows.

4- Most of these shows are released either at the end or the beginning of the year.

5- The United States and India are among the top five countries that produce all of the available content on the platform. Additionally, out of the top ten actors with the maximum content, six of them are from India.

6- When it comes to content ratings, TV-MA tops the charts, indicating that mature content is more popular on Netflix.

7- The value of k=15 was found to be optimal for clustering the data, and it was used to group the content into ten distinct clusters.

8- Using this data, a Content based recommender system was created using cosine similarity, which provided recommendations for Movies and TV shows.

Thank

You