Capstone Project

Netflix Movies and TV ShowsClustering

Technical Documentation

Agam Singh

Jyoti Patel

Data science trainees

AlmaBetter

Table of Content:-

Abstract :

The "Netflix Movies and TV Shows Clustering" project is an unsupervised machine learning initiative aimed at analyzing and categorizing the extensive content library of Netflix. The objective is to apply clustering algorithms to discover hidden patterns and group similar titles together, providing valuable insights for content recommendation and management. By preprocessing the dataset, applying clustering algorithms, and extracting features, the project identifies meaningful clusters of movies and TV shows. The outcomes of this project have practical applications in content recommendation, content curation, and content production strategies, enhancing user experience and platform performance.

Introduction:

The "Netflix Movies and TV Shows Clustering" project utilizes unsupervised machine learning techniques to categorize the vast collection of Netflix movies and TV shows. By preprocessing the data, applying clustering algorithms, and extracting features, the project identifies meaningful clusters of titles sharing similar characteristics. The results offer valuable insights for content recommendation, content curation, and content production strategies, empowering users to discover relevant content and optimizing Netflix's content organization and delivery.

Problem Statement:

The primary objective of the "Netflix Movies and TV Shows Clustering" project is to employ unsupervised machine learning techniques to analyze and categorize the vast collection of movies and TV shows available on the Netflix streaming platform. Through this analysis, the project aims to identify meaningful patterns and group similar titles together based on their unique characteristics. The ultimate goal is to derive valuable insights that can enhance content recommendation and management, thereby improving the overall user experience and platform performance.

To carry out this endeavor, the project will utilize a comprehensive dataset of Netflix movies and TV shows from the year 2019, sourced from the reputable third-party Netflix search engine, Flixable. This dataset has already revealed interesting trends, including a notable surge in the number of TV shows on Netflix since 2010, while the count of movies has decreased by over 2,000

titles during the same period. The project team intends to delve deeper into this dataset to uncover additional valuable insights. Furthermore, they aim to explore the possibility of integrating external datasets, such as IMDB ratings and Rotten Tomatoes, to reveal further intriguing findings. By leveraging these data sources and employing advanced machine learning techniques, the project aspires to make substantial contributions to the understanding and management of content on the Netflix platform.

Approach:

1. Data Collection: The first step of the approach involves collecting a comprehensive dataset of Netflix movies and TV shows. The dataset should encompass a wide range of genres, release years, ratings, and countries of origin. It should also include essential features such as title, description, genre, duration, cast, and production details.

2. Data Preprocessing: Before applying clustering algorithms, the dataset undergoes careful preprocessing. This includes handling missing values appropriately, converting text-based features like title and description into a numerical format using natural language processing techniques (e.g., TF-IDF or Word Embeddings), and encoding categorical variables. Numeric features are also scaled to ensure consistency in data representation.

3. Feature Extraction: As the dataset contains both text-based and numerical features, a combination of techniques is employed for feature extraction. Text features are transformed into numerical vectors using methods like TF-IDF or Word Embeddings, while numeric features are used in their original form after scaling.

4. Clustering Algorithms: Several unsupervised clustering algorithms are considered for the project, such as K-means, hierarchical clustering, and DBSCAN. Each algorithm is evaluated to determine its effectiveness in partitioning the Netflix content dataset into coherent and meaningful clusters. The optimal number of clusters is determined through techniques like the Elbow method, Silhouette analysis, or Davies-Bouldin index.

5. Model Training: The chosen clustering algorithm is then trained on the preprocessed and feature-extracted dataset. The algorithm aims to group similar movies and TV shows into clusters based on their inherent features.

6. Cluster Interpretation: After the clustering process, the resulting clusters are analyzed and interpreted. This involves examining the titles within each cluster, identifying common genres, popular actors or directors, and regional preferences. Understanding the themes and patterns within each cluster is crucial to deriving valuable insights from the clustering outcomes.

7. Evaluation: Since clustering is an unsupervised learning task, there are no predefined labels to assess the model's performance directly. However, internal evaluation metrics like the Silhouette score and Davies-Bouldin index can provide an indication of the clustering quality.

8. Application: The insights obtained from the clustering exercise have practical applications for Netflix. They can be leveraged to enhance the content recommendation system, enabling users to discover content aligned with their preferences more effectively. Content managers can also utilize the clustering results to curate and organize the content library, creating specialized collections for different themes and interests. Additionally, the information can guide content production strategies to cater to diverse global audiences.

9. Visualization: To aid in understanding the clustering results, the clusters, and their associated features can be visualized using plots, charts, and interactive visualizations. The visual representation can provide a more intuitive understanding of the content distribution within each cluster.

10. Iteration and Refinement: The clustering process might require multiple iterations, fine-tuning, and parameter adjustments to achieve optimal results. Continuous refinement of the approach and cluster analysis can lead to better insights and enhanced applications for Netflix's content management and recommendation system.

Our Goal :

Our Goal in this Project:

The primary goal of the "Netflix Movies and TV Shows Clustering" project is to leverage the power of unsupervised machine learning to gain valuable insights into the vast catalog of TV shows and movies available on the Netflix streaming platform as of 2019. By applying clustering techniques and integrating external datasets,

our aim is to achieve the following objectives:

(a). Uncover Hidden Patterns and Trends: Our first goal is to delve into the Netflix dataset and discover hidden patterns and trends related to the content library. By exploring the growth of TV shows and movies over the years, we seek to understand how the content composition has evolved and identify any significant shifts or preferences among users.

(b). Cluster Similar Titles: We aim to use unsupervised clustering algorithms to group similar TV shows and movies together based on various features such as genres, ratings, and production details. This clustering will enable us to create meaningful content categories, facilitating better content organization and management.

(c). Enhance Content Recommendation: By utilizing the clustering results, our objective is to enhance Netflix's content recommendation system. By suggesting relevant TV shows and movies to users based on their interactions within different clusters, we aim to provide a more personalized and engaging viewing experience.

(d). Optimize Content Curation: We aim to utilize the clustering outcomes to curate and organize Netflix's extensive content library more effectively. Through the creation of specialized collections based on clustered content, we seek to improve user content discovery and facilitate easier navigation through the vast content offerings.

(**e). Integration with External Datasets:** Another goal is to explore the integration of the Netflix dataset with external datasets such as IMDB ratings and Rotten Tomatoes. This integration aims to provide additional insights into content popularity, critical acclaim, and user preferences, enhancing the overall understanding of content quality and relevance.

(f). Assess Impact and Performance: Throughout the project, we will continually assess the impact of our clustering approach and the integration of external datasets on Netflix's user engagement, satisfaction, and platform performance. Our goal is to demonstrate tangible improvements in content recommendation, user experience, and platform effectiveness.

(g). Promote Data-Driven Decision-Making: Ultimately, our overarching goal is to promote data-driven decision-making at Netflix. By extracting valuable insights from the dataset and applying machine learning techniques, we aim to empower content managers and stakeholders with the information needed to make informed content production and curation decisions, leading to a more engaging and satisfying user experience.

The main goal is to use cutting-edge machine-learning techniques to unlock valuable insights from the Netflix content library. By clustering similar titles and integrating external data sources, we strive to optimize content recommendation, enhance content curation, and make Netflix's platform more effective and user-centric.

Data Description:

Explaining the features present:-

The dataset consists of 7787 rows and 12 columns. There are some columns ('director', 'cast', 'country', 'date_added')with null values present in the dataset.

The columns present in the dataset are:

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: The 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 : Genere

12. description: The Summary description

Data Wrangling:

A) Data Cleaning: Data cleaning is a crucial step in the data preprocessing process that aims to ensure the dataset's quality and integrity. It involves identifying and resolving issues with the data to prepare it for analysis and modeling. Common data cleaning tasks include handling missing values, removing duplicates, correcting data types, and addressing outliers.

B) Null Values Treatment: Null values, also known as missing values, are placeholders in the dataset that lack actual data. Null values can disrupt analysis and modeling, so it is essential to handle them appropriately. In this project, null values are addressed using specific techniques. The columns 'director', 'cast', and 'country' with missing values are filled with the string 'Unknown' using the fillna() method. The 'rating' column's missing values are replaced with the mode (most frequent value) of the 'rating' column. Additionally, rows with missing values are dropped using the dropna() method with axis=0 to ensure the dataset contains complete and reliable data.

➢ **The top country in which a movie was produced**

```
United States      2877
India               956
United Kingdom      576
Unknown             506
Canada              259
                    ...
Zimbabwe              1
Namibia               1
Soviet Union          1
Iran                  1
Somalia               1
```

- The above table shows the top country in which movies were produced

➢ Top and bottom Genre of shows

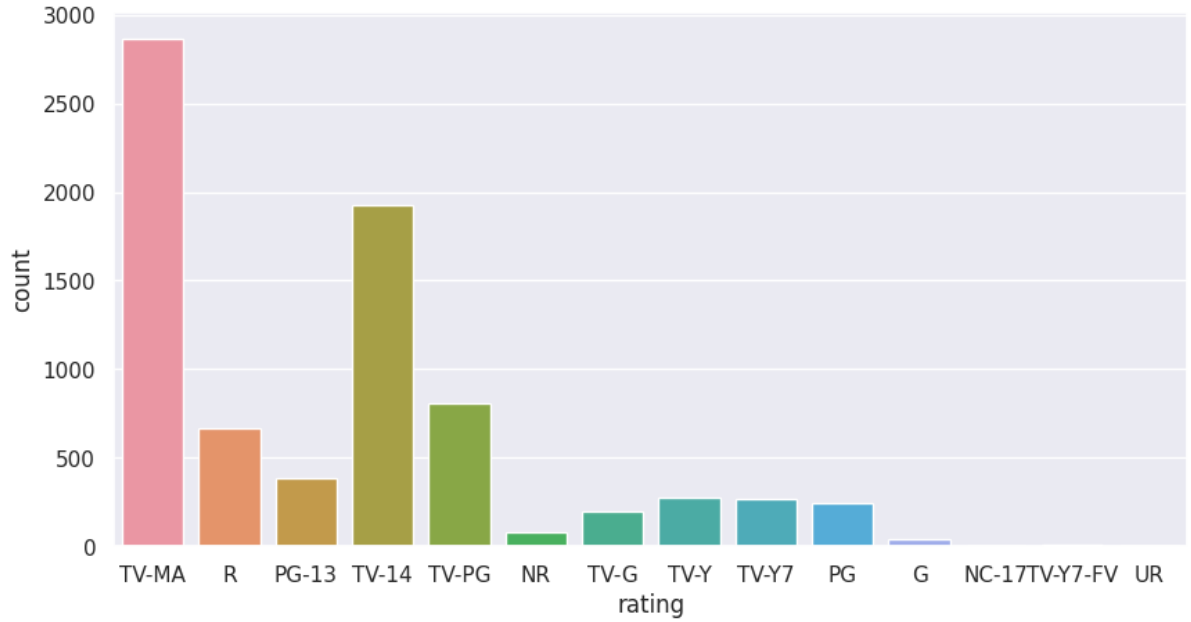
- **Top Genre of show**

Dramas	1384
Comedies	1074
Documentaries	751
Action & Adventure	721
International TV Shows	689
Children & Family Movies	502
Crime TV Shows	369
Kids' TV	357
Stand-Up Comedy	321
Horror Movies	244
British TV Shows	231
Docuseries	193
Anime Series	147
International Movies	114
TV Comedies	109
Reality TV	102
Classic Movies	77

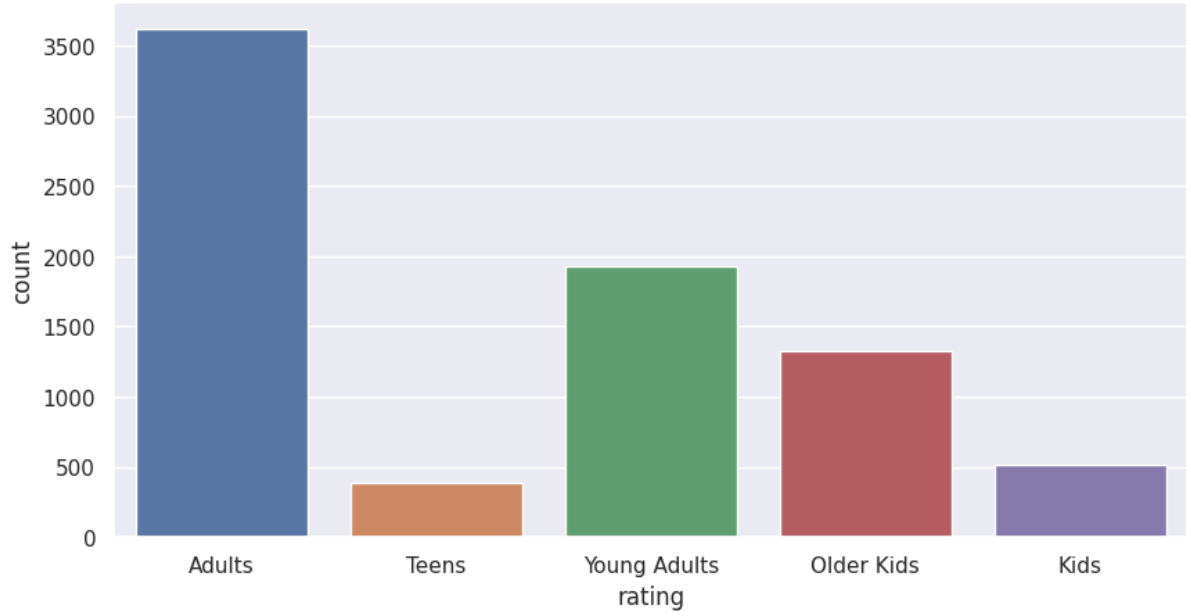
- **Bottom Genre of shows**

```
TV Dramas                          62
Movies                             56
Thrillers                          49
TV Action & Adventure              36
Stand-Up Comedy & Talk Shows       33
Romantic TV Shows                  28
Independent Movies                 20
Anime Features                     19
Classic & Cult TV                  19
Music & Musicals                   17
TV Shows                           12
Cult Movies                        12
Sci-Fi & Fantasy                   11
TV Horror                          10
Romantic Movies                     3
Spanish-Language TV Shows           2
Sports Movies                       1
TV Sci-Fi & Fantasy                 1
LGBTQ Movies                        1
```

➢ Age ratings for shows in the dataset



- From the above graph, we can see that the highest number of shows on Netflix are rated by TV-MA, followed by TV-14 and TV-PG.



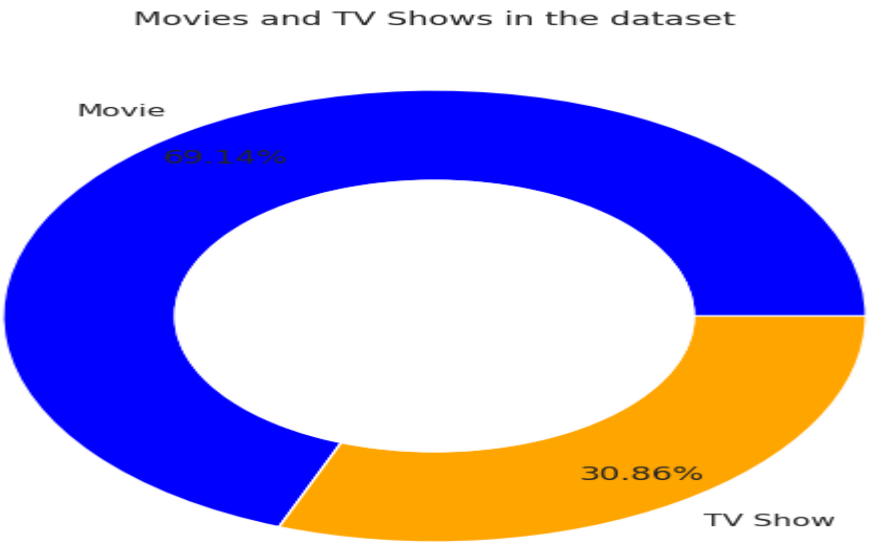
- Around 50% of shows on Netflix are produced for adult audiences. Followed by young adults, older kids, and kids. Netflix has the least number of shows that are specifically produced for teenagers than other age groups.

Exploratory Data Analysis:

A. Data Visualization:

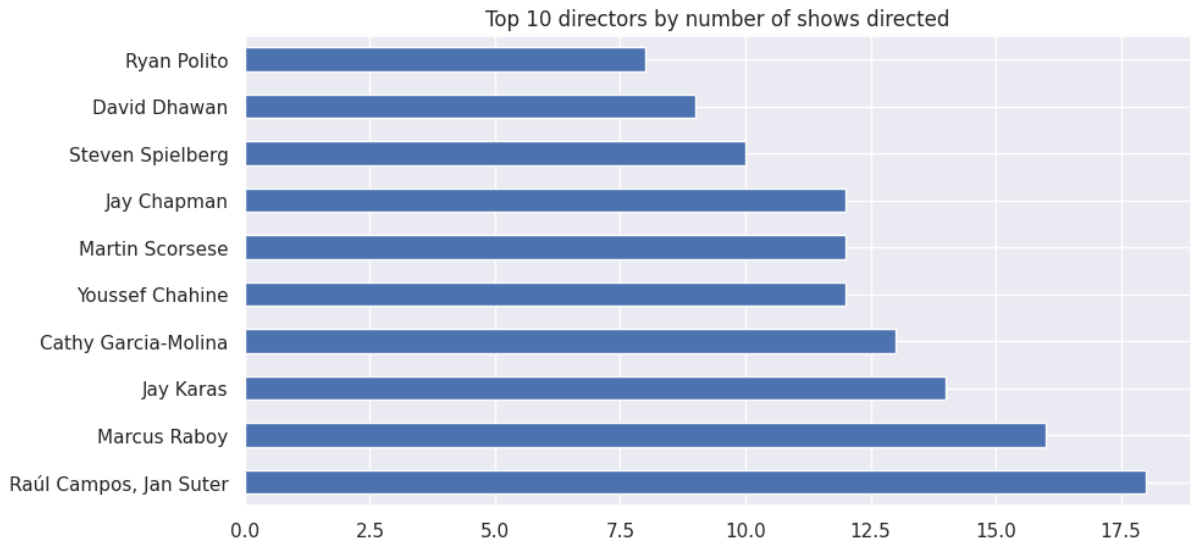
1. Univariate Analysis:

➢ **Number of Movies and TV Shows in the dataset**



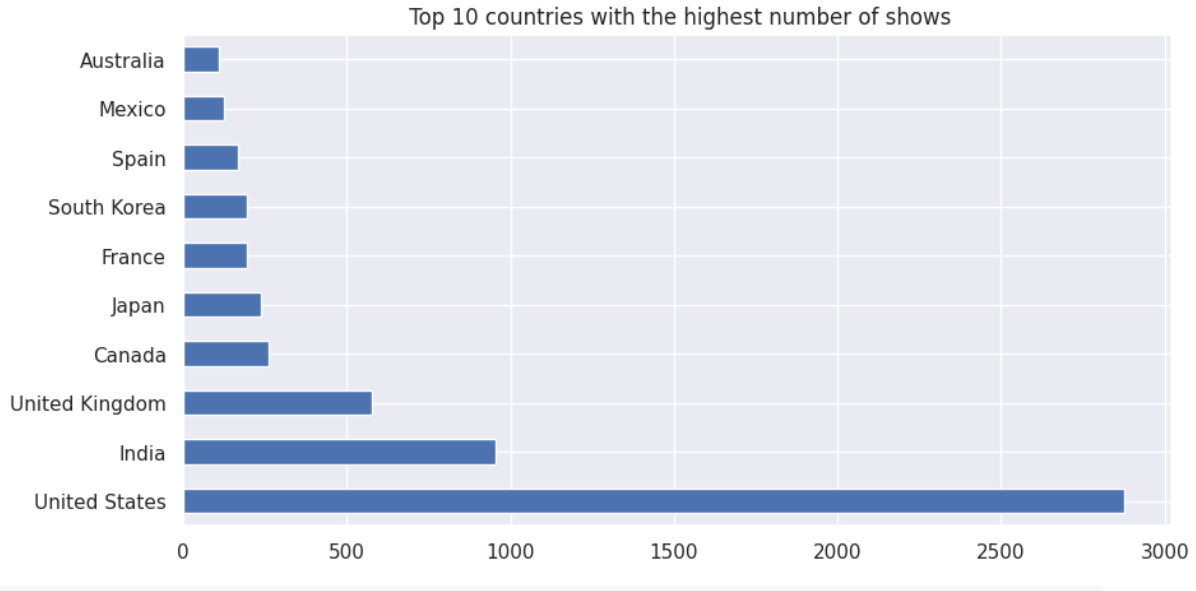
- **There are more movies (69.14%) than TV shows (30.86%) in the dataset.**

➢ **Top 10 directors in the dataset**



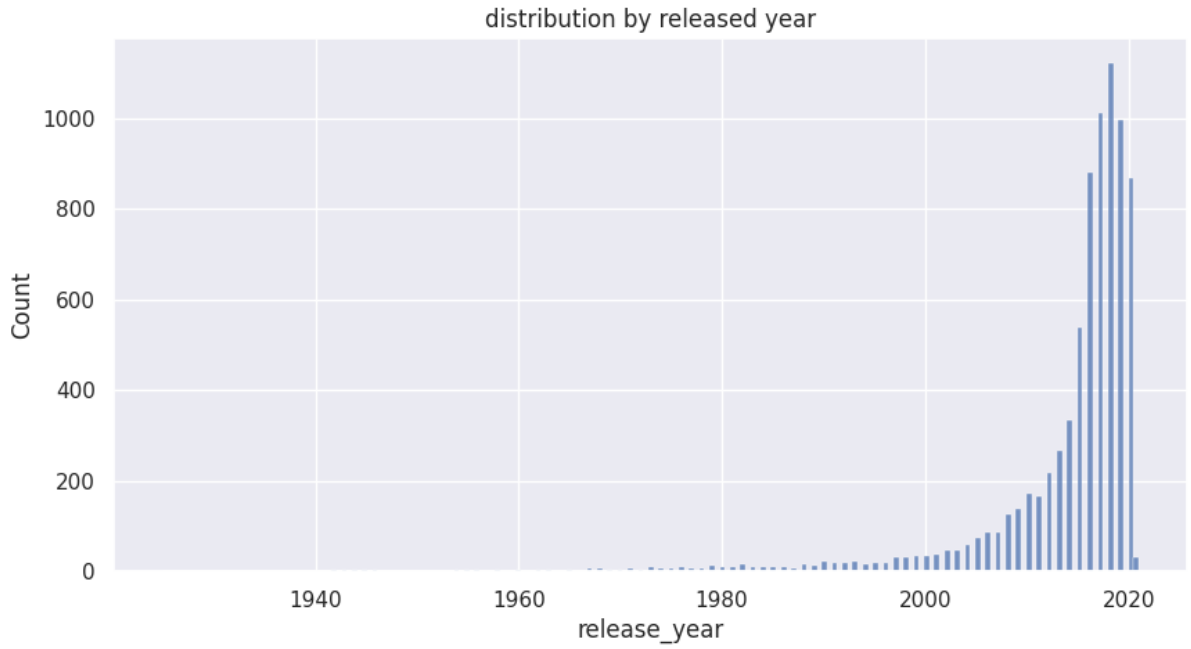
- **Raul Campos and Jan Suter together have directed 18 movies / TV shows, higher than anyone in the dataset.**

➢ **Top 10 countries with the highest number of movies / TV shows in the dataset**



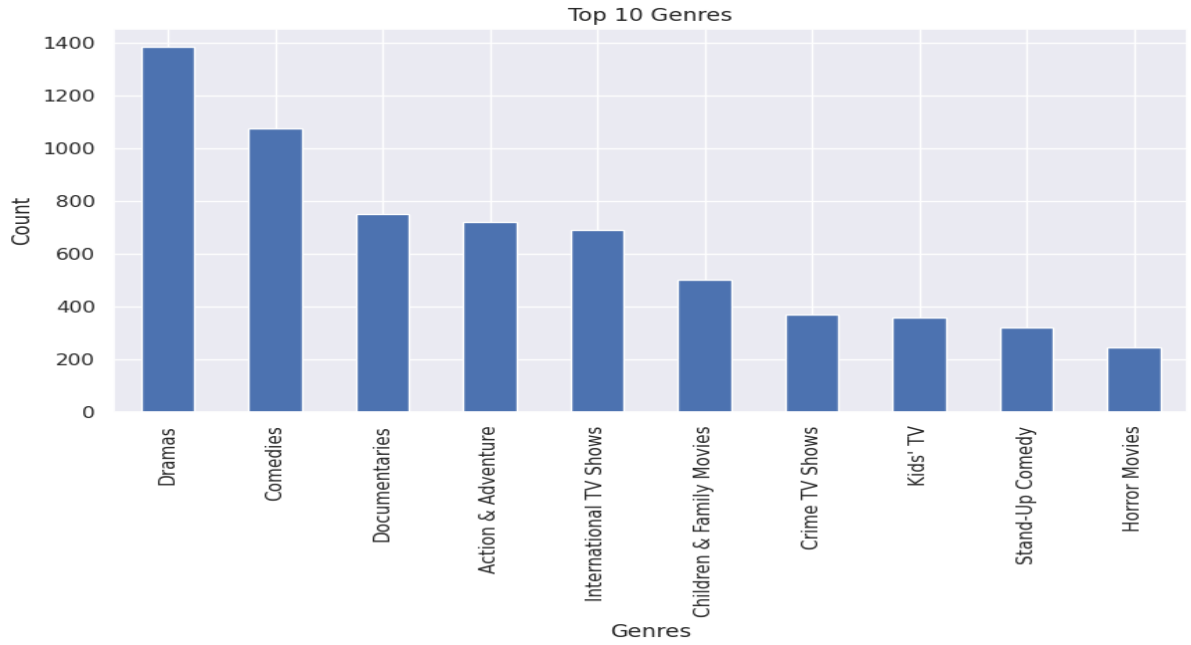
- **The highest number of movies / TV shows were based out of the US, followed by India and UK.**

➢ **Visualizing the year in which the movie / tv show was released**



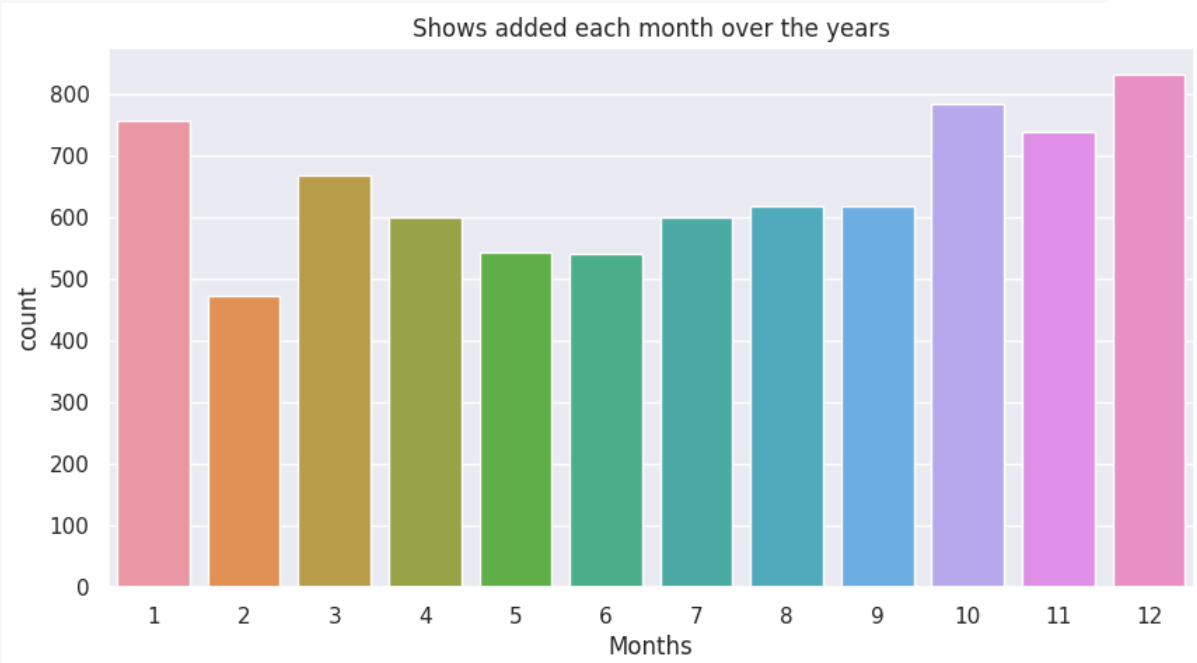
- **Netflix has greater number of new movies / TV shows than the old ones.**

➢ Top 10 genres



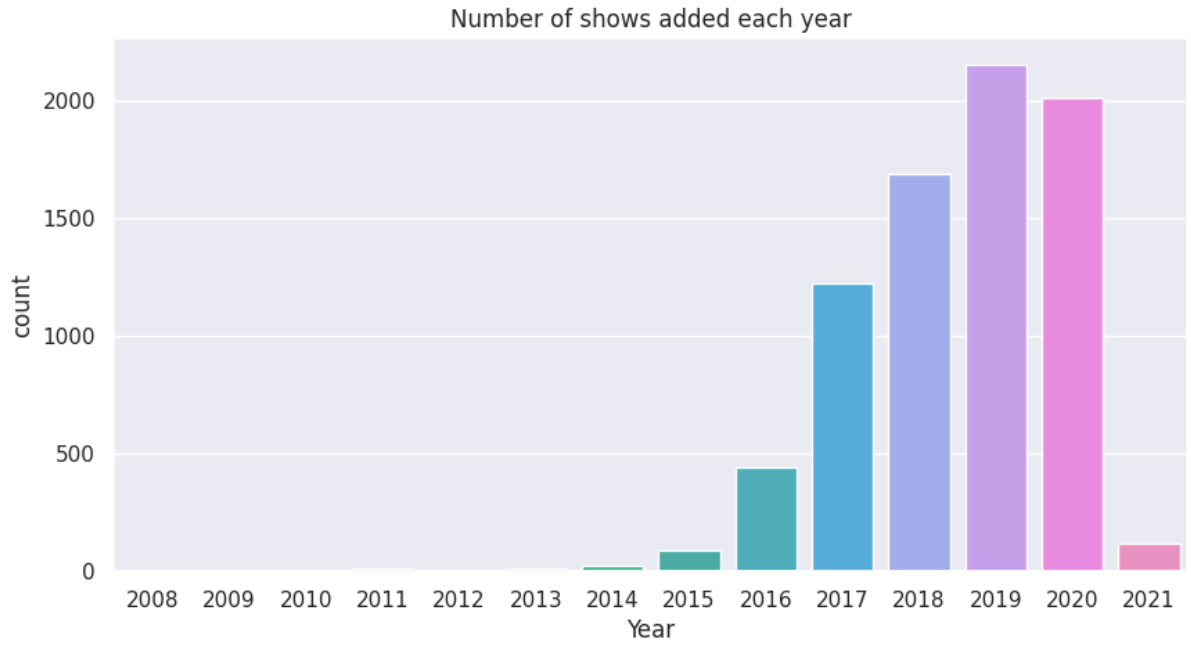
- **The dramas is the most popular genre followed by comedies and documentaries.**

➢ Number of shows added on different months



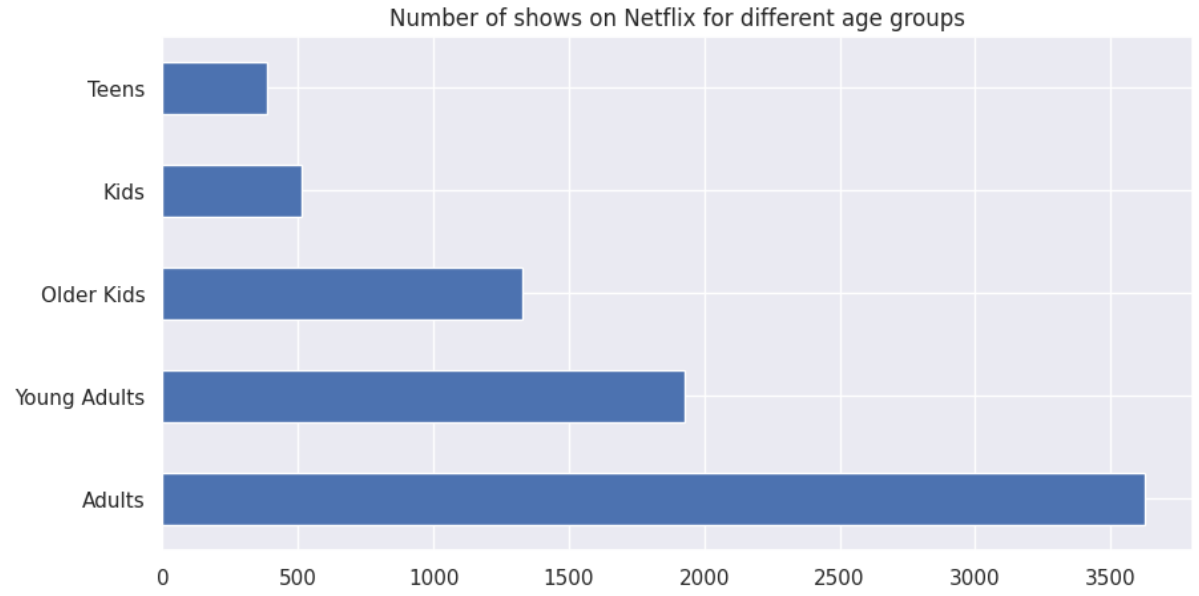
- **Over the years a greater number of shows were added in the months of October, November, December, and January.**

➢ Number of shows added over the years



- **Netflix continuous to add more shows on its platform over the years.**
- **There is a decrease in the number of shows added in the year 2020, which might be attributed to the covid-19-induced lockdowns, which halted the creation of shows.**

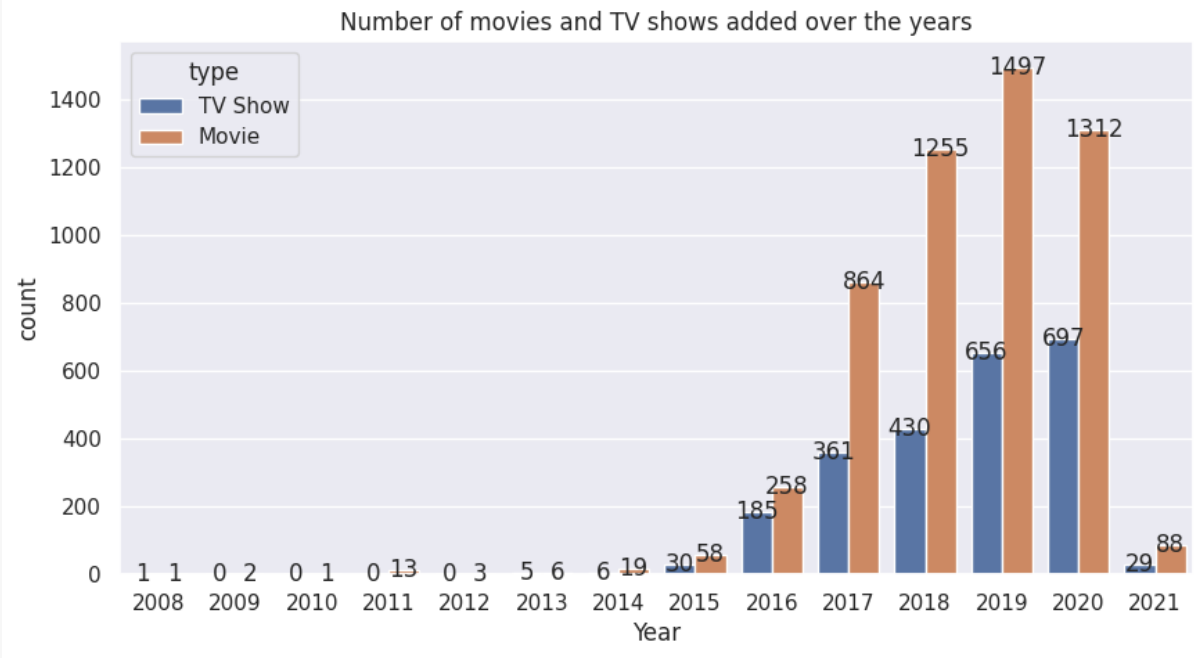
- **We have Netflix data only up to 16th January 2021, hence there are less movies added in this year.**



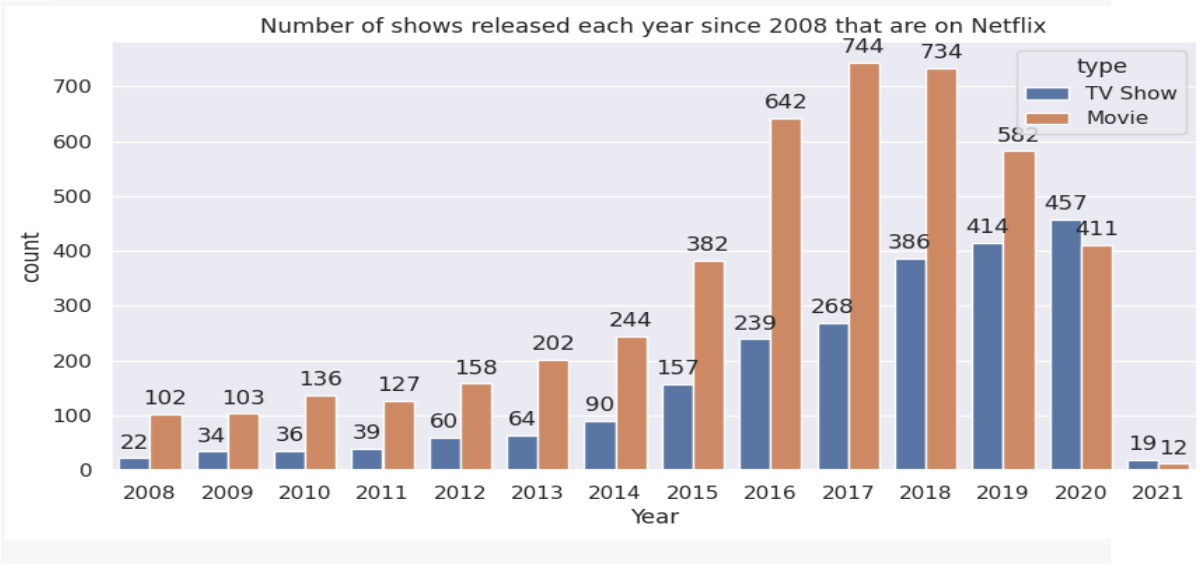
- **The majority of the shows on Netflix are catered to the needs of adult and young adult population.**

2. Bivariate analysis:

➢ **Number of movies and TV shows added over the years**

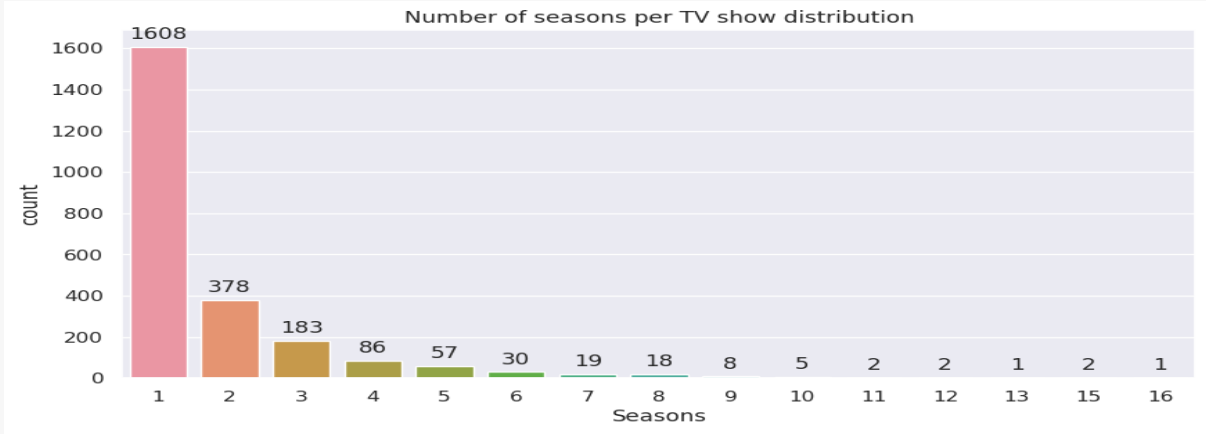


➢ Number of shows released each year since 2008



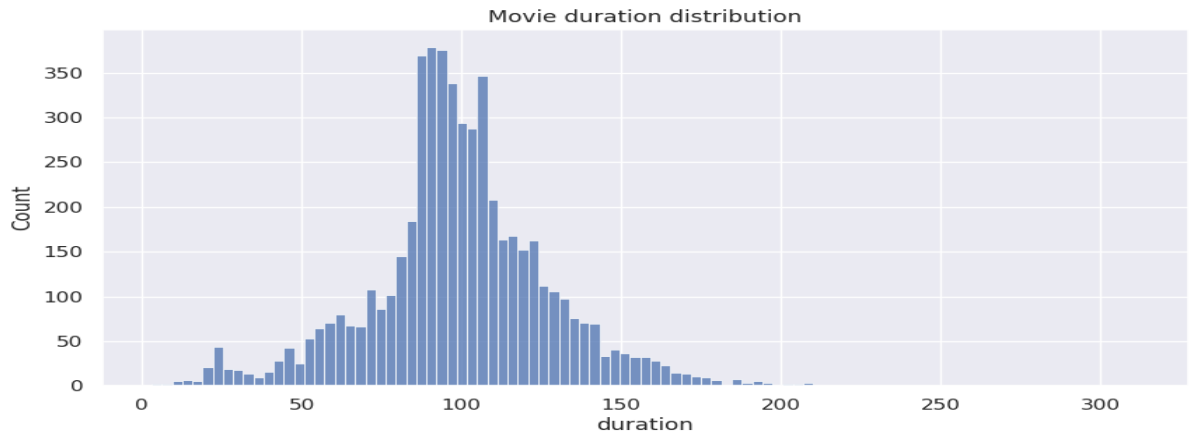
- **Over the years, Netflix has consistently focused on adding more shows in its platform.**
- **Though there was a decrease in the number of movies added in 2020, this pattern did not exist in the number of TV shows added in the same year.**
- **This might signal that Netflix is increasingly concentrating on introducing more TV series to its platform rather than movies.**

➢ Seasons in each TV show



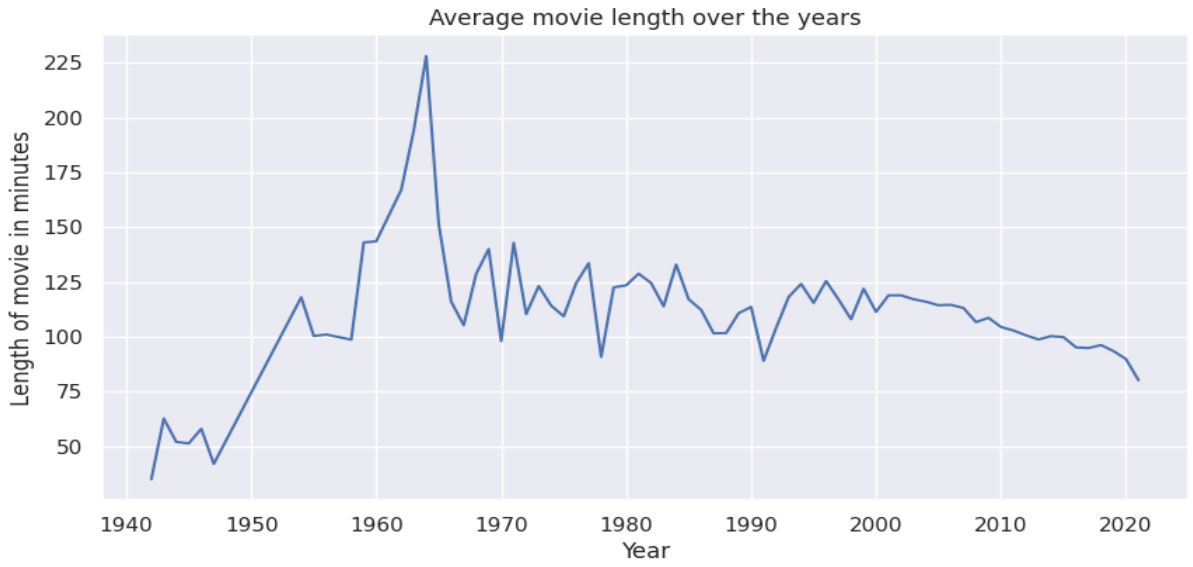
- **The TV series in the dataset have up to 16 seasons, however the bulk of them only have one. This might mean that the majority of TV shows has only recently begun, and that further seasons are on the way.**
- **There are very few TV shows that have more than 8 seasons.**

➢ length of movie analysis



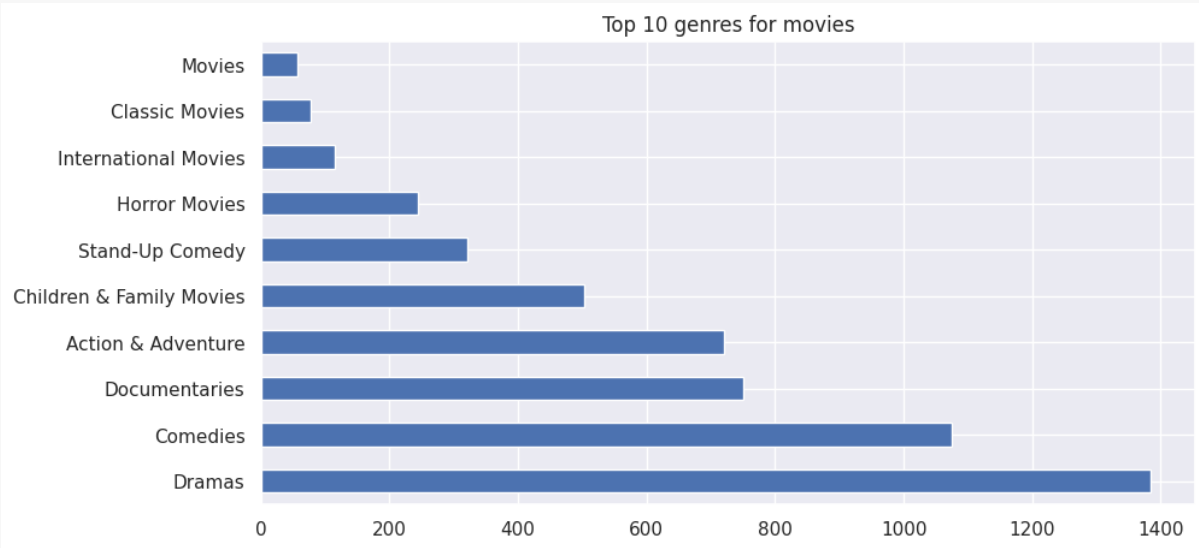
- **The length of a movie may range from 3 min to 312 minutes, and the distribution is almost normally distributed.**

➢ Average movie length over the years



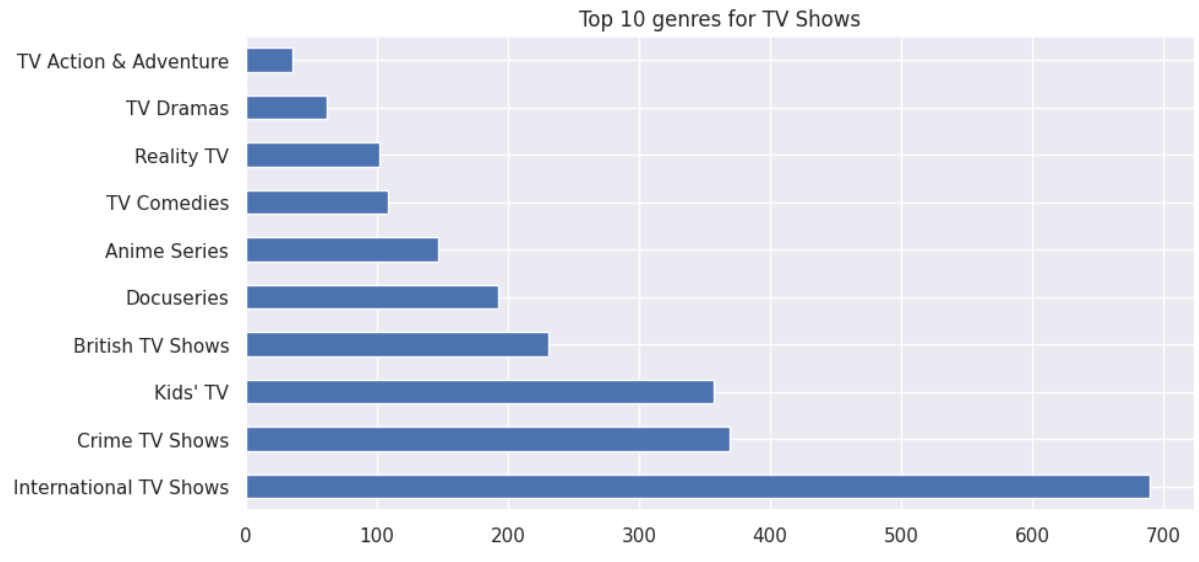
- **Netflix has several movies on its site, including those that were released in way back 1942.**
- **As per the plot, movies made in the 1940s had a fairly short duration on average.**
- **On average, movies made in the 1960s have the longest movie length.**
- **The average length of a movie has been continuously decreasing since the 2000s.**

➢ Top 10 genre for movies



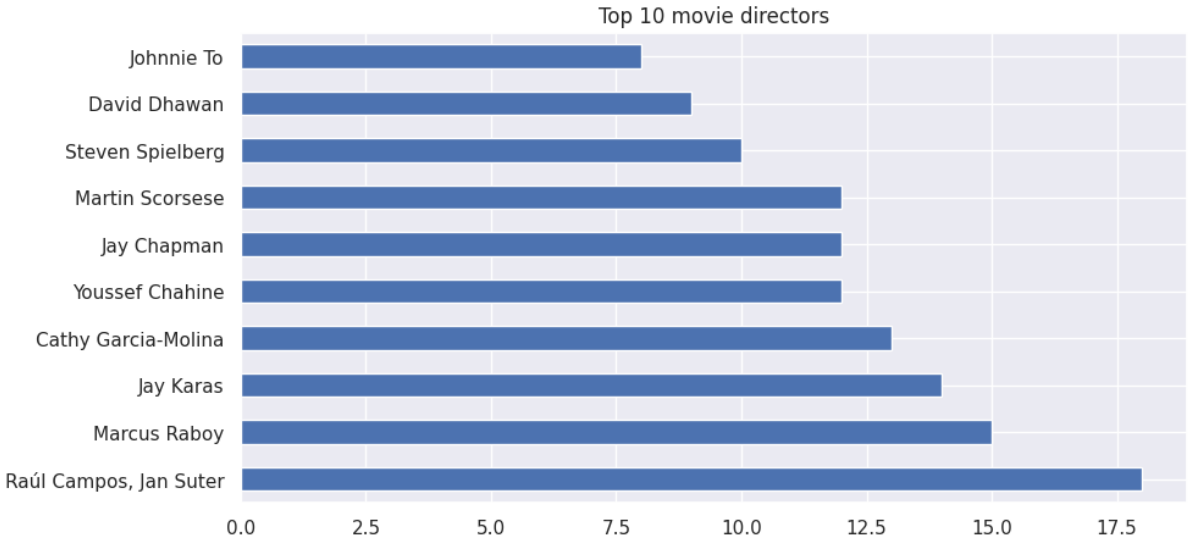
- **Dramas, comedies, and documentaries are the most popular genre for the movies on Netflix.**

➢ Top 10 genre for tv shows



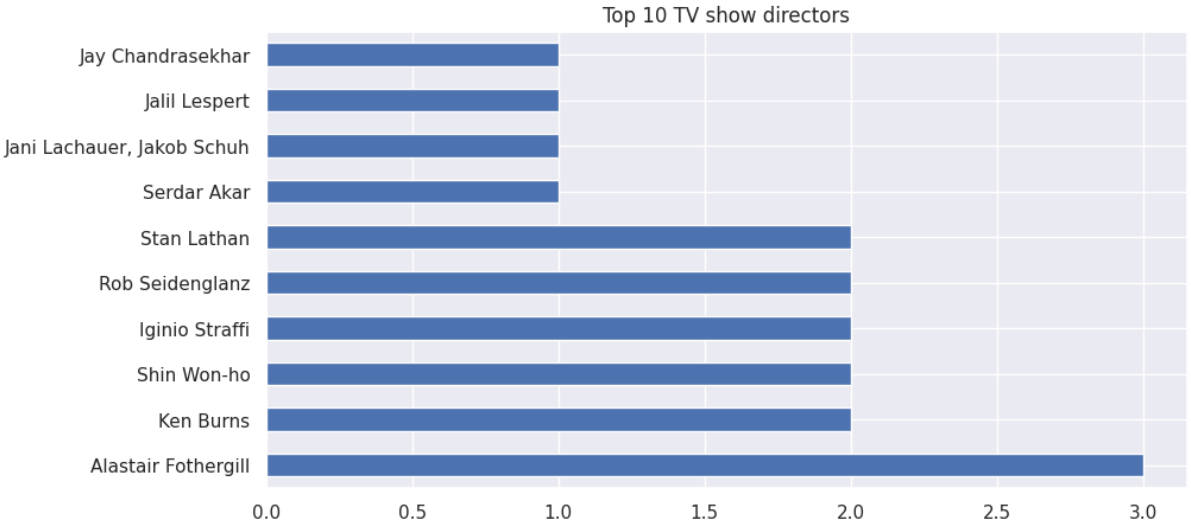
- **International, crime, and kids are the most popular genre for TV shows on Netflix.**

➢ Top 10 movie directors



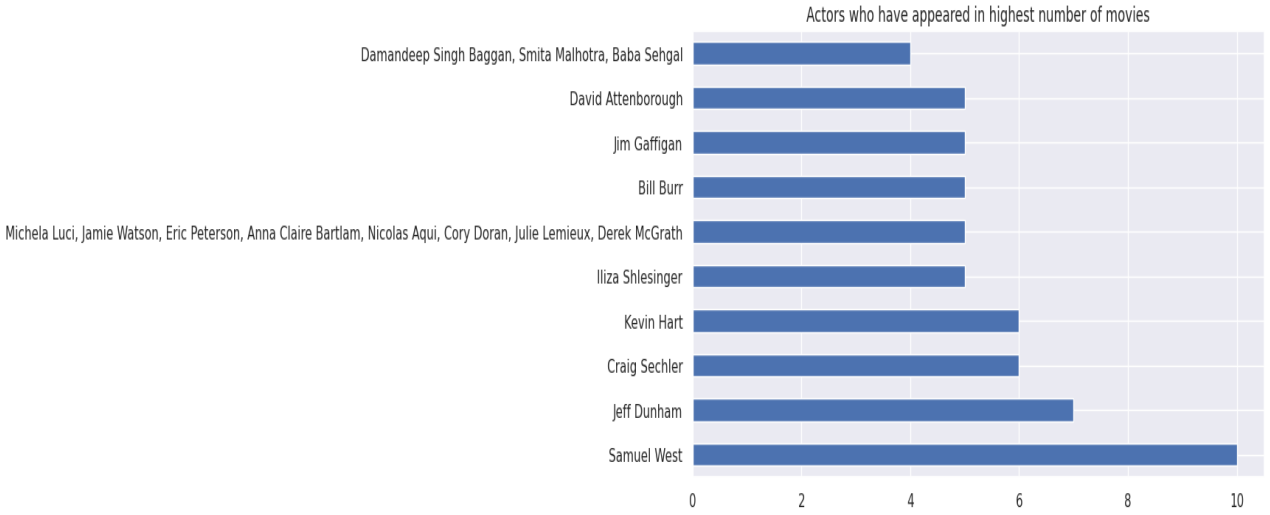
- **Raul Campos and Jan Suter have togather directed in 18 movies, higher than anyone yet.**
- **This is followed by Marcus Roboy, Jay Karas, and Cathy Gracia-Molina.**

➢ Top 10 TV show directors



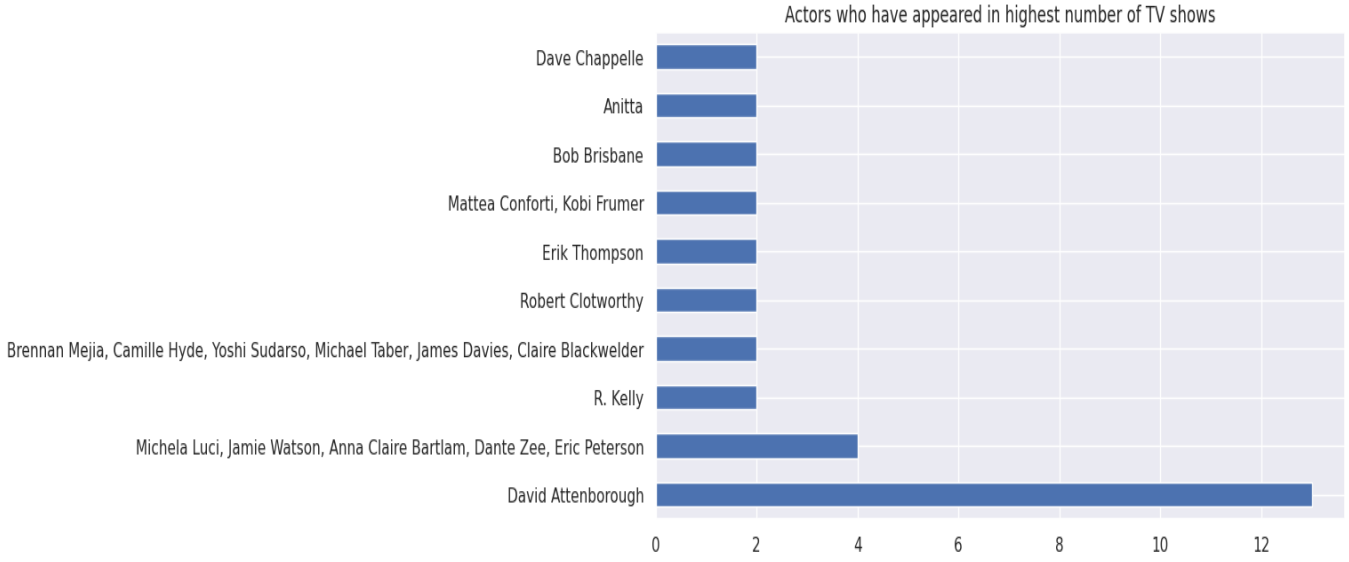
- **Alastair Fothergill has directed three TV shows, the most of any director.**
- **Only six directors have directed more than one television show.**

➢ Top actors for movies



- **Samuel West has appeared in 10 movies, followed by Jeff Dunham with 7 movies.**

➢ Top actors for TV shows



- **David Attenborough has appeared in 13 TV shows, followed by Michela Luci, Jamie Watson, Anna Claire Bartlam, Dante Zee, Eric Peterson with 4 TV shows.**

Building a wordcloud

- **Some keywords in Netflix show descriptions: life, family, new, love, young, world, group, death, man, woman, murder, son, girl, documentary, secret.**

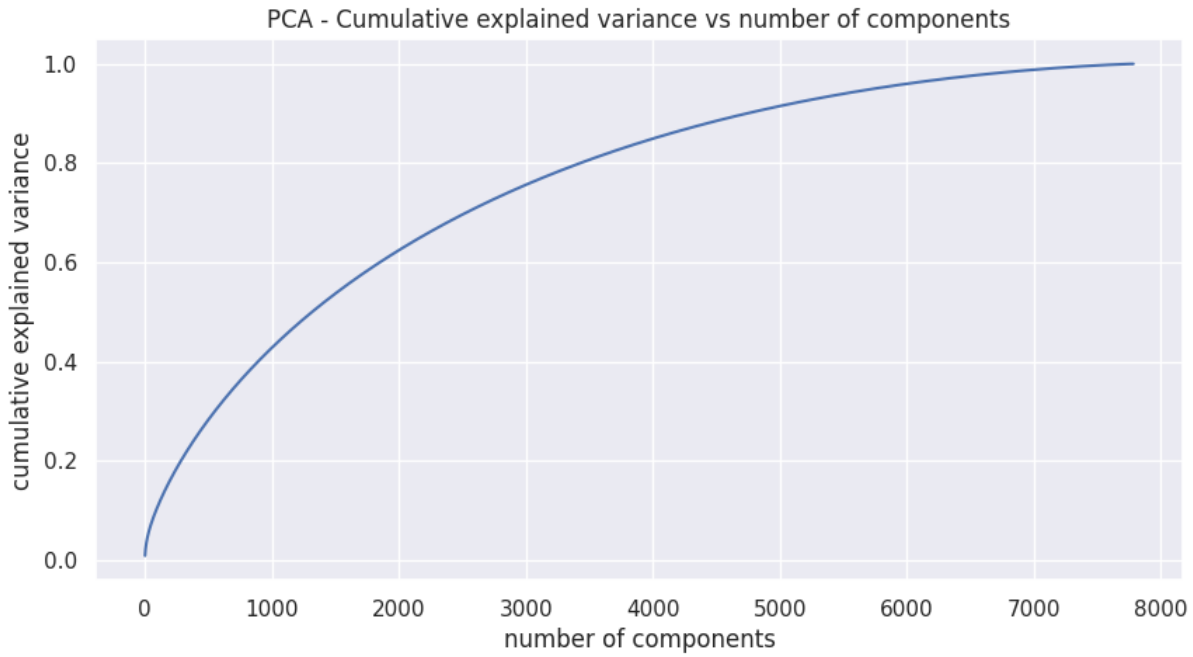
Data preprocessing:

A. Modelling Approach:

1. Select the attributes based on which you want to cluster the shows
2. Text preprocessing: Remove all non-ascii characters, stopwords and punctuation marks, convert all textual data to lowercase.
3. Lemmatization to generate a meaningful word out of corpus of words
4. Tokenization of corpus
5. Word vectorization
6. Dimensionality reduction
7. Use different algorithms to cluster the movies, obtain the optimal number of clusters using different techniques
8. Build optimal number of clusters and visualize the contents of each cluster using wordclouds.

Dimensionality reduction using PCA

- ➢ **`Explained variance for different number of components`**



- **We find that 100% of the variance is explained by about ~7500 components.**
- **Also, more than 80% of the variance is explained just by 4000 components.**
- **Hence to simplify the model, and reduce dimensionality, we can take the top 4000 components, which will still be able to capture more than 80% of variance.**

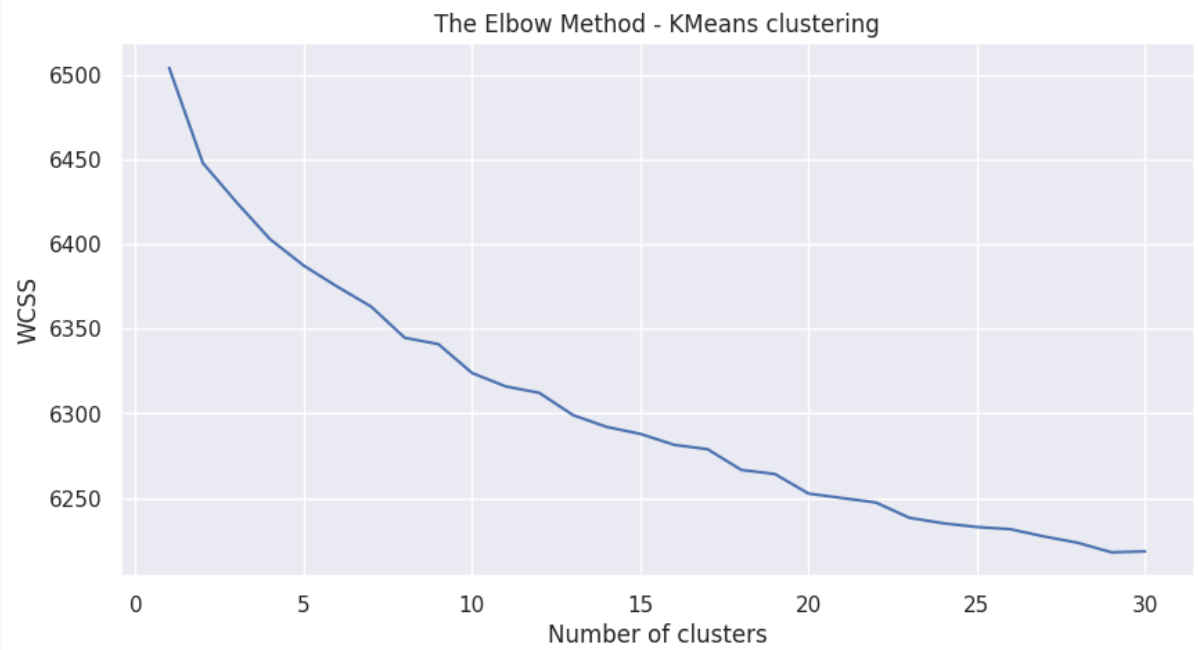
Clusters implementation:

A. K-Means Clustering

Building clusters using the K-means clustering algorithm.

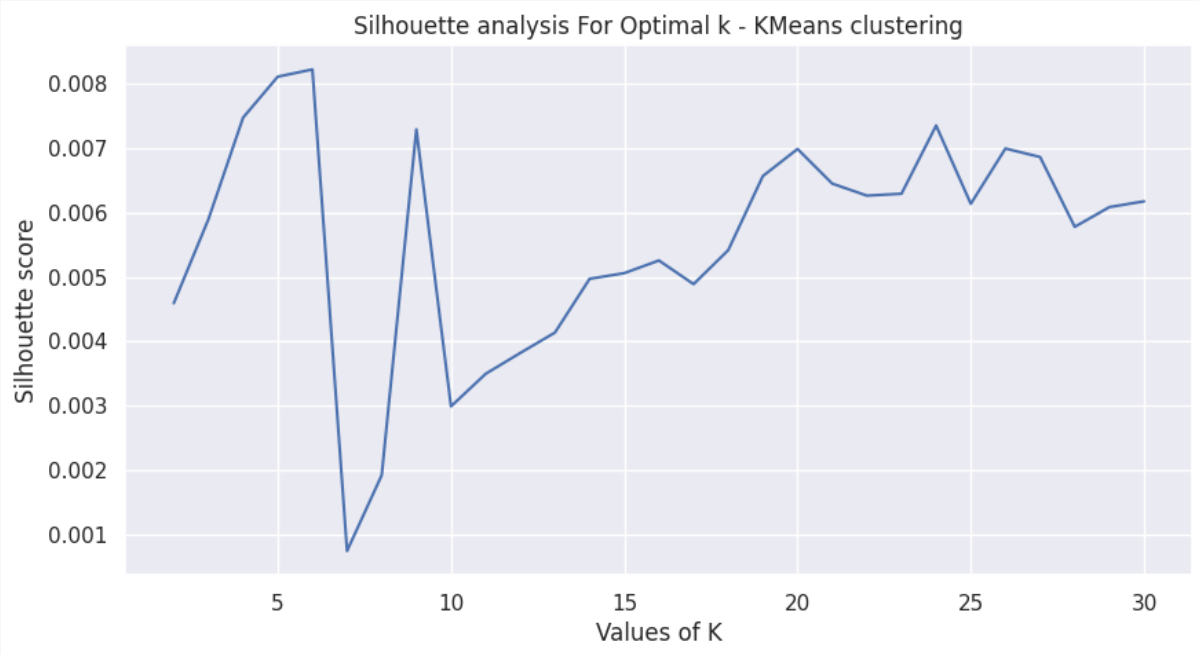
Visualizing the elbow curve and Silhouette score to decide on the optimal number of clusters for K-means clustering algorithm.

➢ Elbow method to find the optimal value of k



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

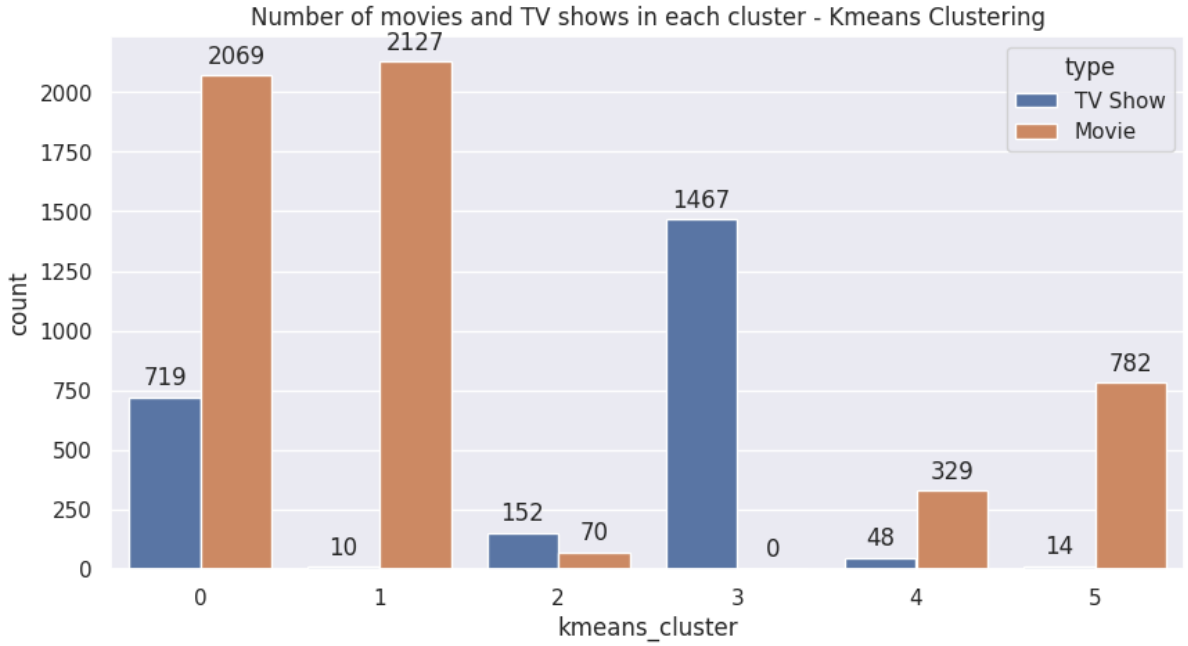
➢ Plotting Silhouette score for different umber of clusters



- **The highest Silhouette score is obtained for 6 clusters.**

Building 6 clusters using the k-means clustering algorithm:

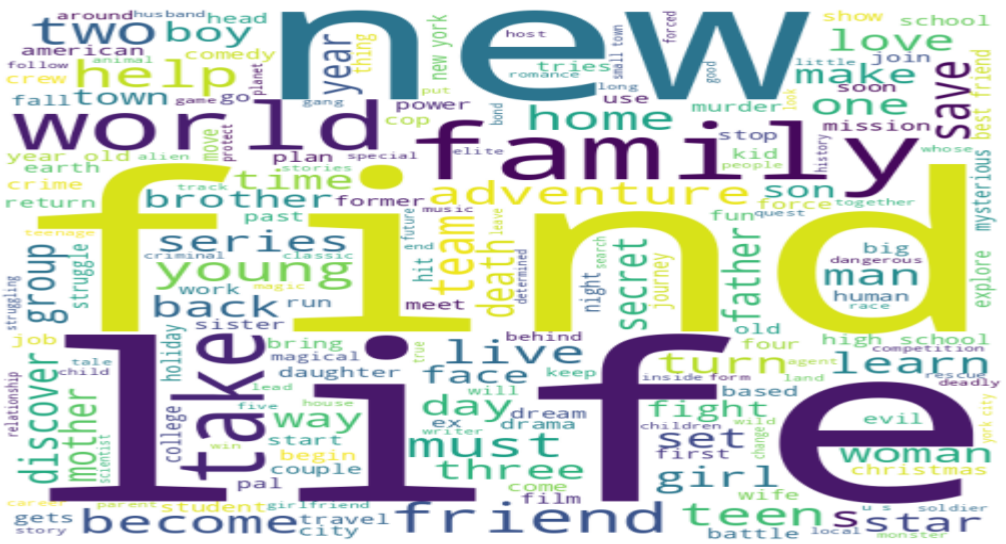
➢ **Number of movies and tv shows in each cluster**



- **Above graph represent Successfully built of 6 clusters using the k-means clustering algorithm.**

Building wordclouds for different 6 clusters built:

➢ **Wordcloud for cluster 0**

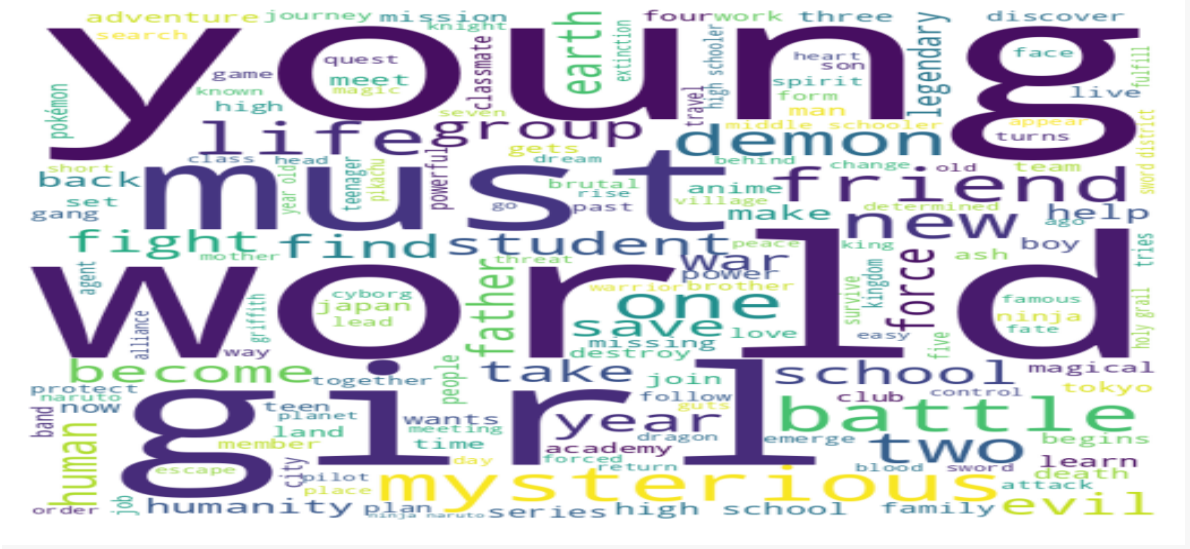


- **Keywords observed in cluster 0: life, new, family, friend, save, help, discover, home, teen**

➢ Wordcloud for cluster 1

- **Keywords observed in cluster 1: life, love, family, father, young, girl, man, woman, friend, daughter**

➢ Wordcloud for cluster 2



- **Keywords observed in cluster 2: young, world, girl, mysterious, humanity, life, student, school, battle, demon, force**

➢ Wordcloud for cluster 3

- **Keywords observed in cluster 3: love, life, family, romance, crime, murder, world, adventure**

➢ **Wordcloud for cluster 4**

- **Keywords observed in cluster 4: comedian, special, stand, comic, stage, sex, joke**

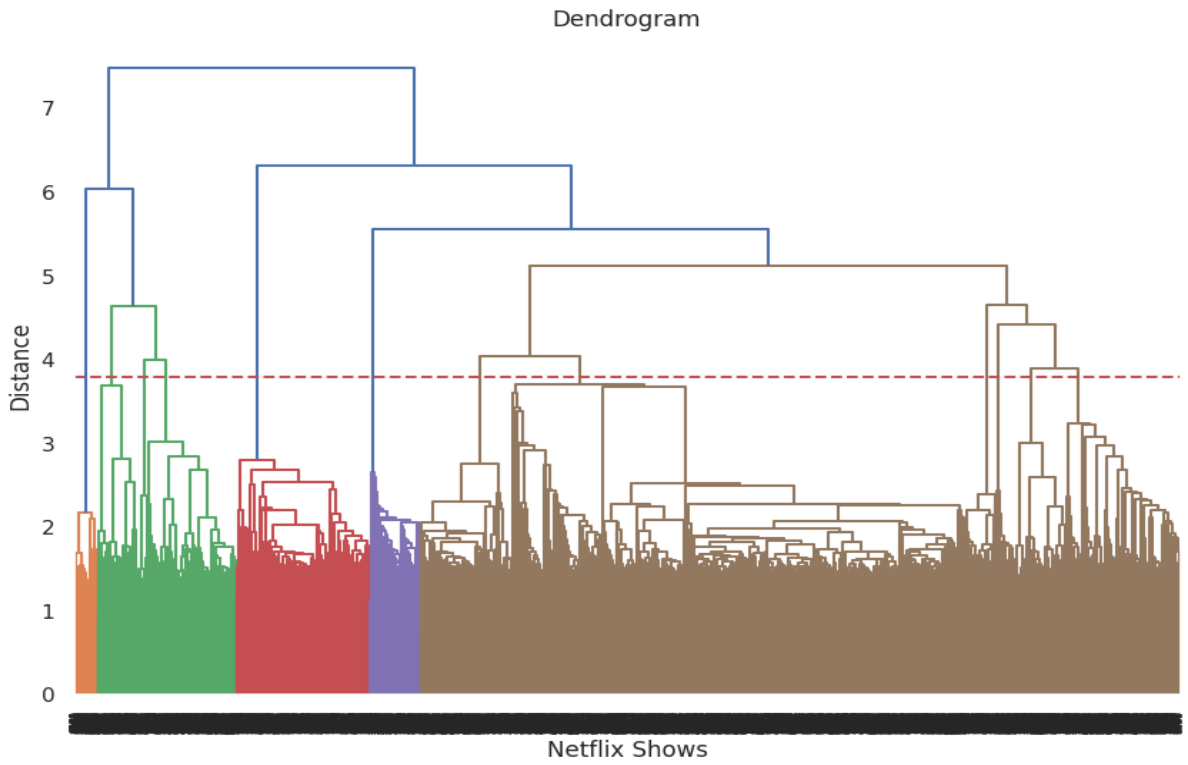
➢ **Wordcloud for cluster 5**

- **Keywords observed in cluster 5: documentary, world, life, filmmaker, american, life**

B. Hierarchical clustering:

Building clusters using the agglomerative (hierarchical) clustering algorithm. Visualizing the dendrogram to decide on the optimal number of clusters for the agglomerative (hierarchical) clustering algorithm:

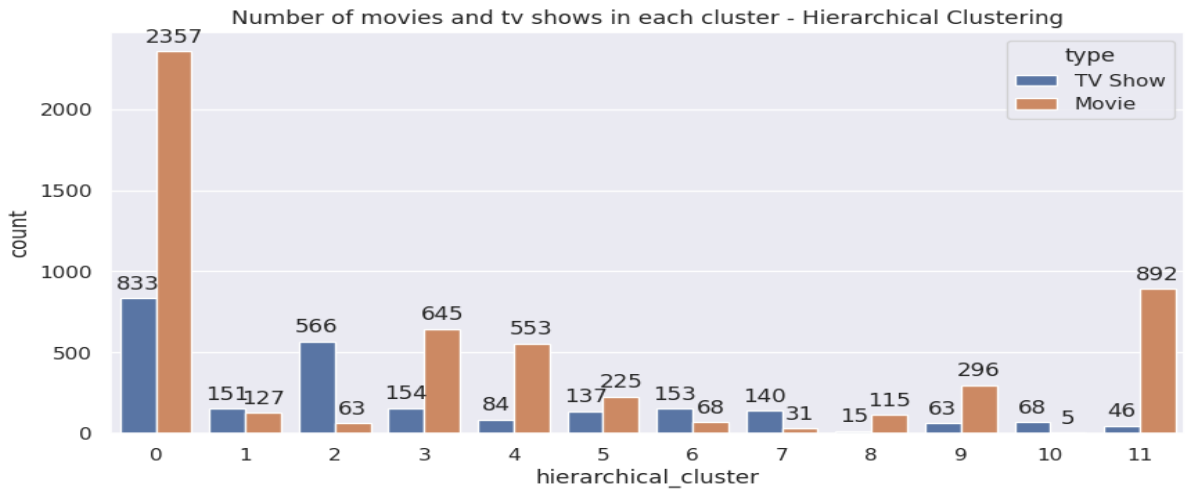
➢ **Building a dendogram to decide on the number of clusters**



- **At a distance of 3.8 units, 12 clusters can be built using the agglomerative clustering algorithm.**

Building 12 clusters using the Agglomerative clustering algorithm:

➢ **Number of movies and tv shows in each cluster**



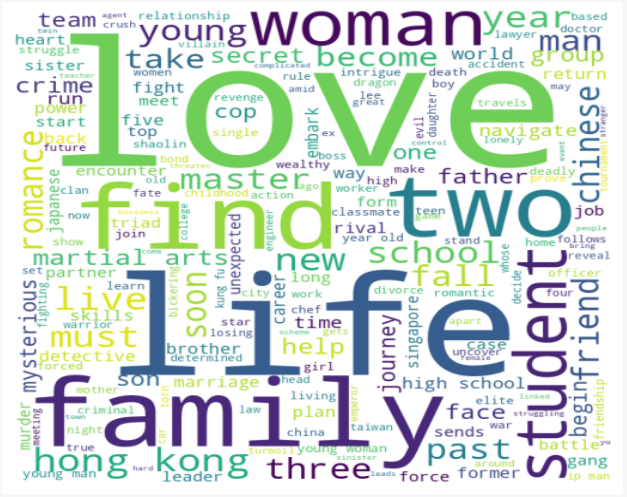
- **The above graph represents the Successfully built 12 clusters using the Agglomerative (hierarchical) clustering algorithm.**

Wordclouds for different clusters built:

Wordcloud for cluster0

Keywords observed in cluster 0: life, new, find, family, save, friend, young, teen, adventure

Wordcloud for cluster1

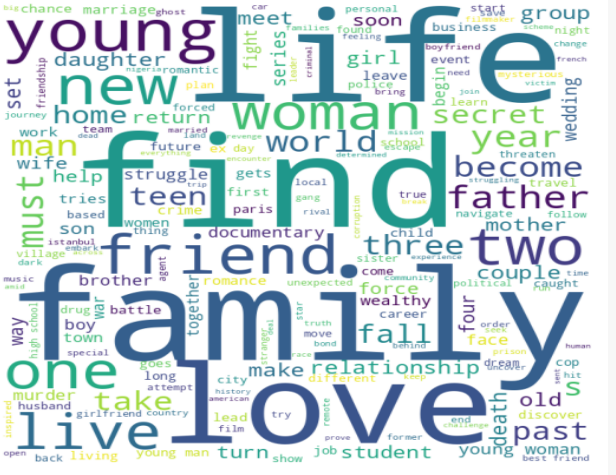


Keywords observed in cluster 1: love, family, life, student, romance, school, woman, master, father

Wordcloud for cluster2

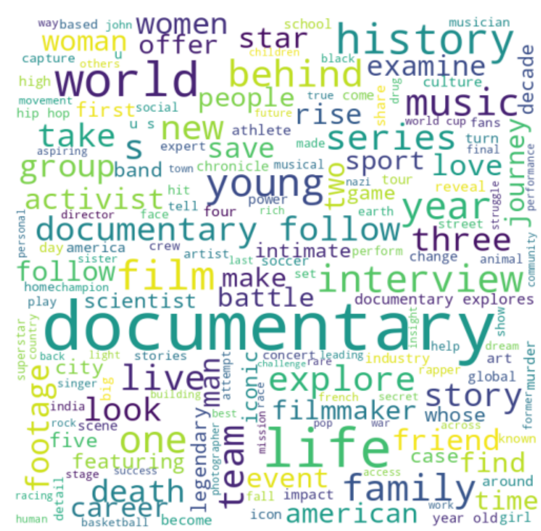
Keywords observed in cluster 2: life, new, series, crime, world, murder, history, detective

Wordcloud for cluster3



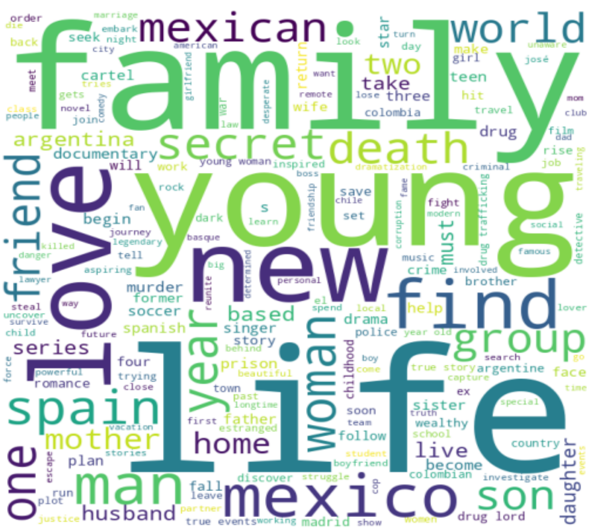
Keywords observed in cluster 3: family, life, love, friend, teen, woman, man, young, world, wedding, secret

Wordcloud for cluster4



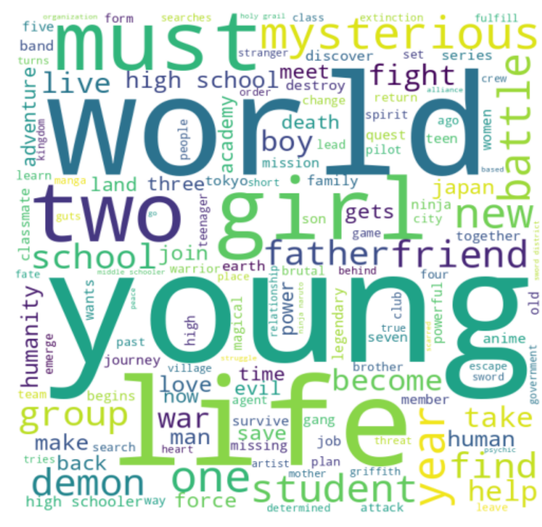
Keywords observed in cluster 4: documentary, music, world, team, interview,history, family, career, battle, death

Wordcloud for cluster 5



Keywords observed in cluster 5: family, life, mexico, young, new, woman, man, secret, spain, death, singer

Wordcloud for cluster6

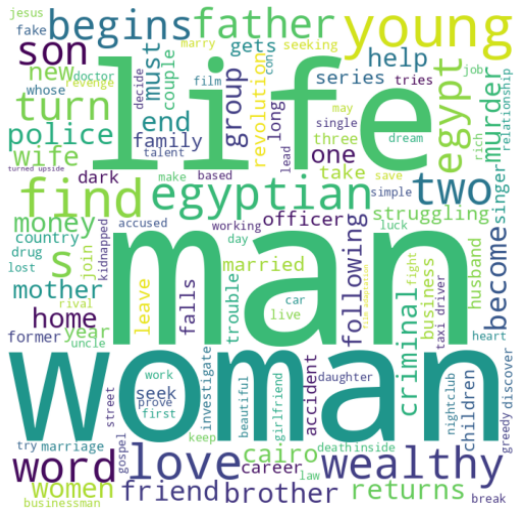


Keywords observed in cluster 6: young, life, girl, world, friend, mysterious, demon, student, school, father.

Wordcloud for cluster7

Keywords observed in cluster 7: love, life, woman, new, student, family, korea, secret, detective, young

Wordcloud for cluster8



Keywords observed in cluster 8: woman, man life, egypt, wealthy, money, young, love, revolution, Struggling

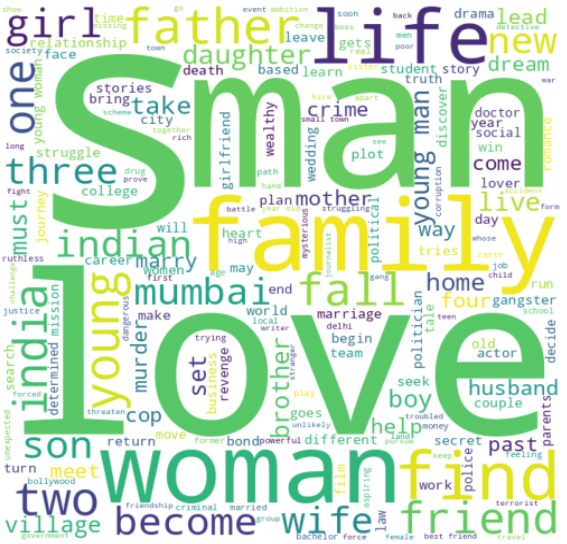
Wordcloud for cluster9

Keywords observed in cluster 9: comedian, stand, life, comic, special, show, live, star, stage, hilarious, stories

Wordcloud for cluster10

Keywords observed in cluster 10: animal, nature, explore, planet, species, survive, natural, life, examine, earth

Wordcloud for cluster11



Keywords observed in cluster 11: love, man, woman, india, father, friend, girl, mumbai, city, learn, young

Content-based recommender system:

- We build a simple content-based recommender system based on the similarity of the shows.
- If a person has watched a show on Netflix, the recommender system must be able to recommend a list of similar shows that s/he 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 as shown in the equation below. We can simply say that the CS score of two vectors increases as the angle between them decreases.

$$Cos(\theta) = \frac{A \,.\, B}{|A| \,.\, |B|}$$

The Recommender System is a data-driven approach that provides personalized content recommendations to users based on their preferences. It utilizes the Count Vectorizer and Cosine Similarity techniques to find similar movies or TV shows based on clustering attributes. By calculating the similarity between shows, the system offers 10 recommendations for a given show title. Users can call the **'recommend_10'** function with the show title as input to receive the recommended show titles. However, it's important to note that the accuracy of recommendations depends on the dataset's quality and size used to build the system.

Conclusions:

In this project, we tackled a text clustering problem involving the classification of Netflix shows into distinct clusters based on their similarities and dissimilarities. The dataset consisted of approximately 7787 records with 11 attributes. One of the key observations was that Netflix hosted a larger number of movies compared to TV shows on its platform, and the total count of shows added to Netflix showed exponential growth over time. Additionally, a significant portion of the shows originated from the United States, with a focus on content targeting the adult and young adult age group.

To perform the clustering, we selected specific attributes, including director, cast, country, genre, and description. These attributes were pre-processed, tokenized, and then vectorized using TFIDF vectorizer, resulting in a total of 20000 attributes. To tackle the high dimensionality of the data, Principal Component Analysis (PCA) was employed, and 4000 components were chosen to capture over 80% of the variance.

The k-means clustering algorithm was initially used to create clusters, with the optimal number of clusters determined to be 6 through the elbow method and Silhouette score analysis. Additionally, the Agglomerative clustering algorithm was applied to build a hierarchical clustering model, which resulted in 12 optimal clusters based on dendrogram visualization.

Furthermore, we developed a content-based recommender system using the similarity matrix obtained from cosine similarity. This recommender system provides users with 10 personalized recommendations based on the type of show they have previously watched.

Overall, this project successfully employed various clustering techniques to categorize Netflix shows, and the content-based recommender system offered valuable and relevant recommendations to enhance the user experience on the platform. These findings have the potential to contribute significantly to content management and user satisfaction in the context of Netflix's vast library of shows and movies.