Capstone Project

Netflix Movies and TV Shows Clustering

TEAM NAME : New World

Team Members

1. Agam Singh 2.Jyoti Patel

About Netflix Movies and TV Shows Clustering

- Netflix Movies and TV Shows is a popular streaming service that offers an extensive collection of diverse and captivating content to its subscribers. From blockbuster movies to binge-worthy TV series, the platform caters to a wide range of interests and preferences. With a user-friendly interface, Netflix allows viewers to explore a vast library of content from various genres, languages, and cultures, ensuring an immersive and personalized entertainment experience. As a pioneer in the streaming industry, Netflix continues to revolutionize the way people consume media, becoming a household name synonymous with quality entertainment.
- Netflix Movies and TV Shows Clustering is an innovative approach to enhance user experience and content recommendation on the popular streaming platform. With a vast library of movies and TV shows, it becomes crucial to organize the content efficiently, allowing users to discover content that aligns with their preferences effortlessly. Clustering algorithms play a significant role in this process by grouping similar movies and TV shows based on various attributes such as genre, language, release year, and user ratings. By employing cutting-edge machine learning techniques, Netflix can create clusters of content, making it easier for users to navigate through the vast selection and discover hidden gems that align with their tastes. This personalized content recommendation system not only enhances user engagement but also helps Netflix gain valuable insights into user preferences, enabling them to continuously improve their content library and provide a top-notch entertainment experience.

CONTENTS

1. Objective
2. Introduction
3. Problem Statement
4. Data Description
5. Exploratory Data Analysis
6. Data preprocessing
7. Clusters implementation
8. Content-based recommender system

Objective

- The main objective of the "Netflix Movies and TV Shows Clustering" project is to leverage unsupervised machine learning techniques to analyze and categorize the extensive catalog of movies and TV shows offered on the Netflix streaming platform. By identifying meaningful patterns and grouping similar titles based on their distinct characteristics, the project aims to derive valuable insights that will enhance content recommendation and management, ultimately improving the overall user experience and platform performance.

Introduction

The "Netflix Movies and TV Shows Clustering" project embarks on a transformative exploration, harnessing the power of unsupervised machine learning techniques to revolutionize the categorization of the vast collection of movies and TV shows available on the Netflix streaming platform. Through the skillful application of advanced preprocessing methods, cutting-edge clustering algorithms, and insightful feature extraction, our endeavor aims to reveal meaningful clusters of titles that share unique characteristics. The outcomes of this project hold the promise of invaluable insights, capable of enhancing content recommendation, optimizing content curation, and shaping content production strategies. Above all, our ultimate goal is to empower users with seamless access to relevant content while refining Netflix's content organization and delivery, thus creating an unparalleled and personalized entertainment experience. Step into the captivating world of "Netflix Movies and TV Shows Clustering" as we unlock the full potential of unsupervised machine learning, reshaping the way we immerse ourselves in the realm of entertainment.

Problem Statement

- The "Netflix Movies and TV Shows Clustering" project aims to utilize unsupervised machine learning techniques to analyze and categorize the vast collection of movies and TV shows available on the Netflix platform. By identifying meaningful patterns and grouping similar titles together based on their unique characteristics, the project seeks to derive valuable insights to enhance content recommendation and management. The ultimate goal is to improve the overall user experience and platform performance. The project will leverage a comprehensive dataset of Netflix movies and TV shows from 2019, along with the possibility of integrating external datasets such as IMDB ratings and Rotten Tomatoes to further enhance its findings and understanding of content on the Netflix platform.
- The project aims to leverage these clusters to derive valuable insights that will improve content recommendation and management, providing users with personalized suggestions and optimizing the overall user experience on the Netflix platform.
- By analyzing the vast collection of movies and TV shows, the project aims to uncover trends in content distribution, genre diversity, and popularity, contributing to strategic content curation decisions and ensuring a diverse and engaging content library for global audiences.

DATA SUMMARY

Name of the Dataset—----------Netflix Movies and TV Shows Clustering

Number of variables/Columns —----------12

Number of observations/Row—----------7787

Duplicate rows —----------0 (0.0%)

Total size in memory—---------- 2.86 MB

VARIABLE DATA TYPE

Data Type	Columns
Numeric – int64	7 release_year
Numeric – float64	Nil
String - object	0 show_id 1 type 2 title 3 director 4 cast 5 country 6 date_added 8 rating 9 duration 10 listed_in 11 description

Data Description

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

STEPS INVOLVED

1) Exploratory Data Analysis

2) Null values Treatment

3) Data Exploration

4) Data Visualization

5) Standardization of features

6) Clusters implementation

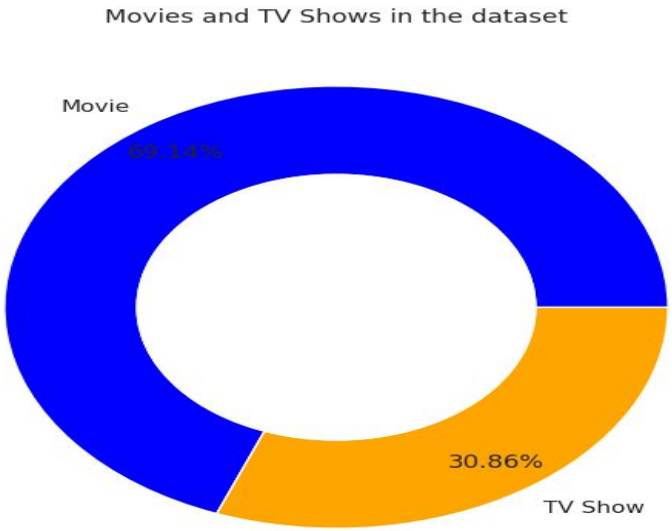
7) Content-based recommender system

8) Conclusion

Exploratory Data Analysis

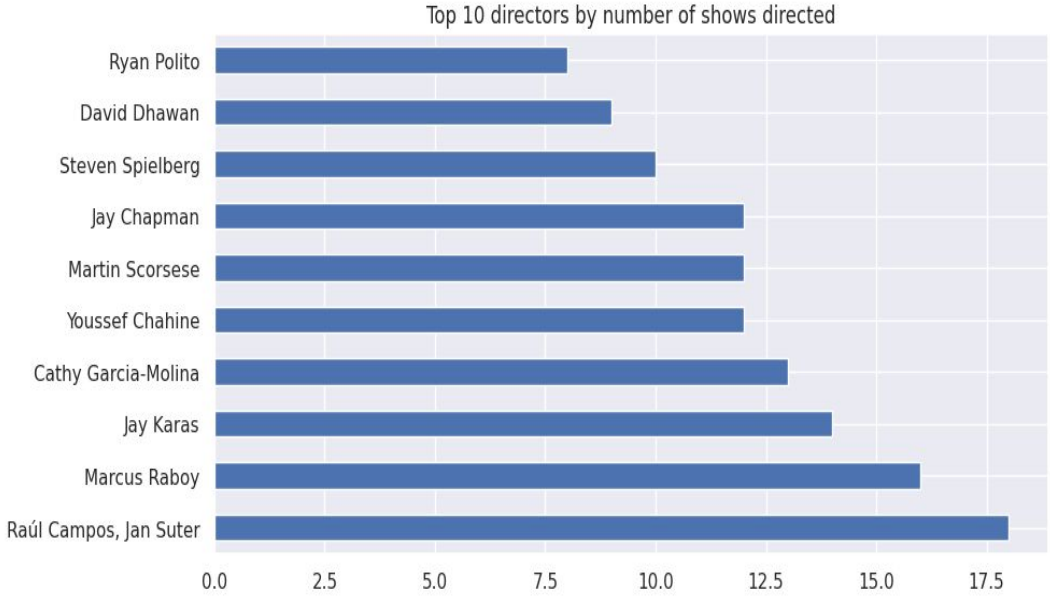
(a). Univariate Analysis

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



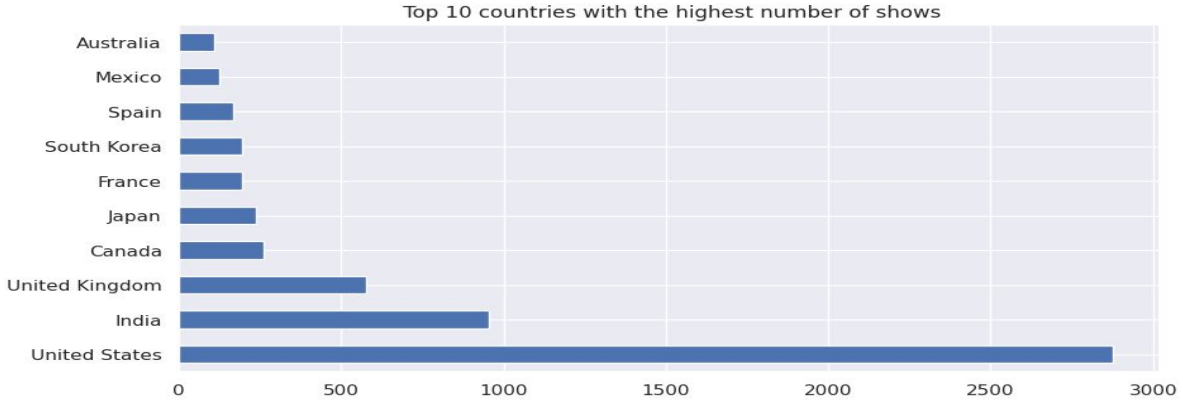
- From above graph we can see that there are more movies (69.14%) than TV shows (30.86%) in the dataset.

➢ **Top 10 directors in the dataset**



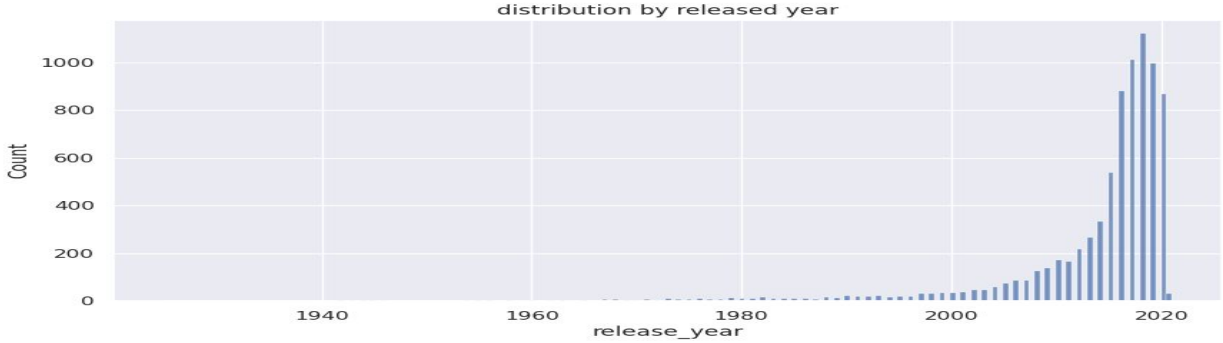
- Above graph show that Raul Campos and Jan Suter together have directed 18 movies / TV shows, higher than anyone in the dataset.

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



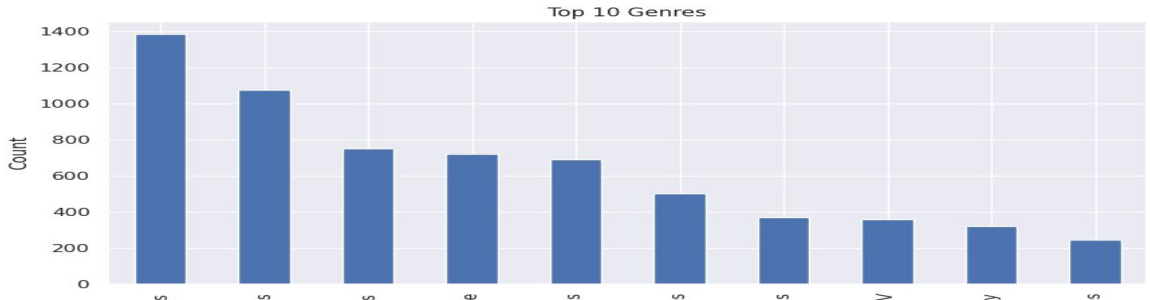
-The highest number of movies / TV shows were based out of the US, followed by India and UK.

➢ **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**

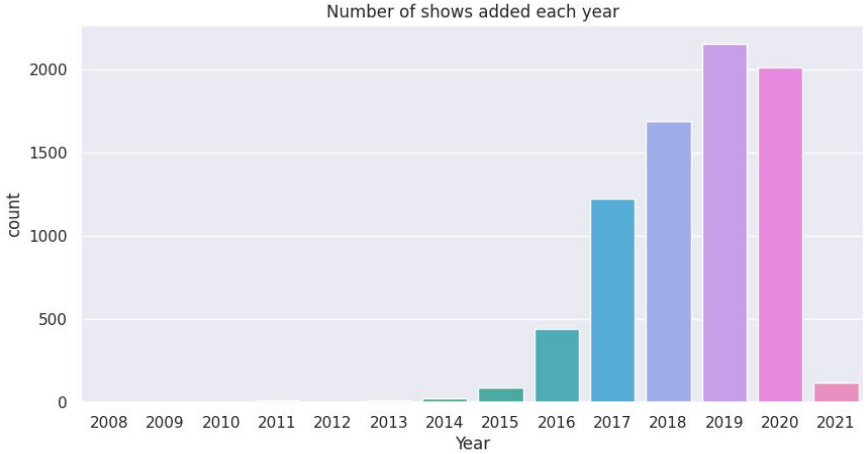


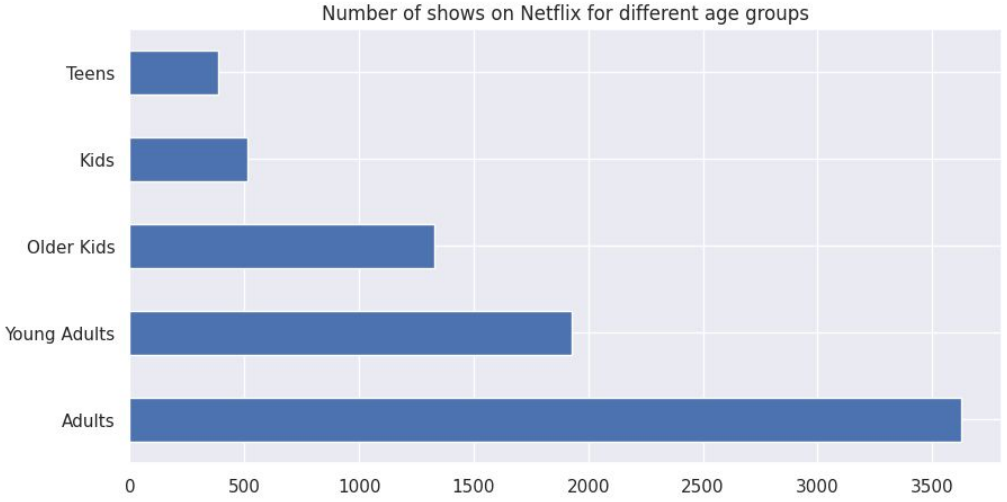
- **The dramas is the most popular genre followed by comedies and documentaries.**

-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**



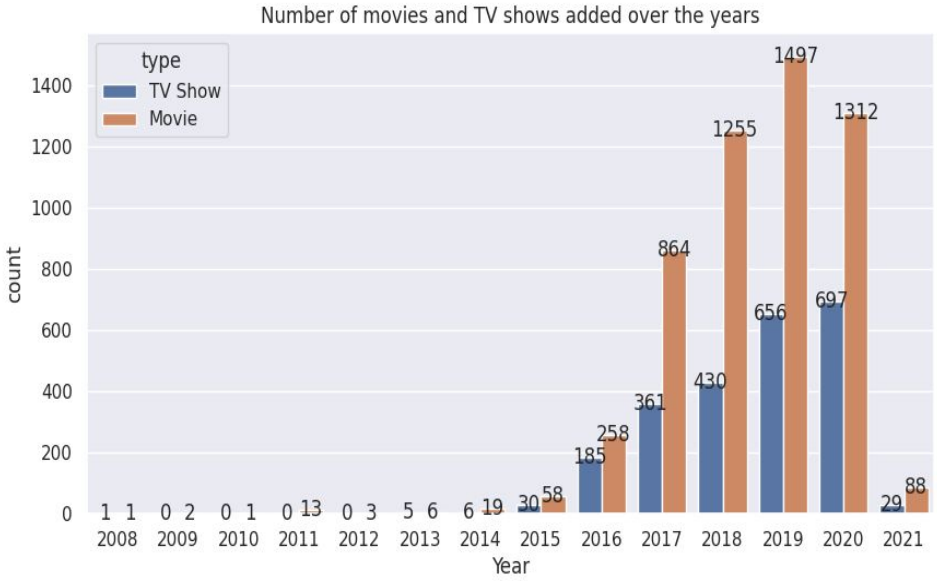


- **Netflix continuous to add more shows on its platform over the years.**
- **We have Netflix data only up to 16th January 2021,**

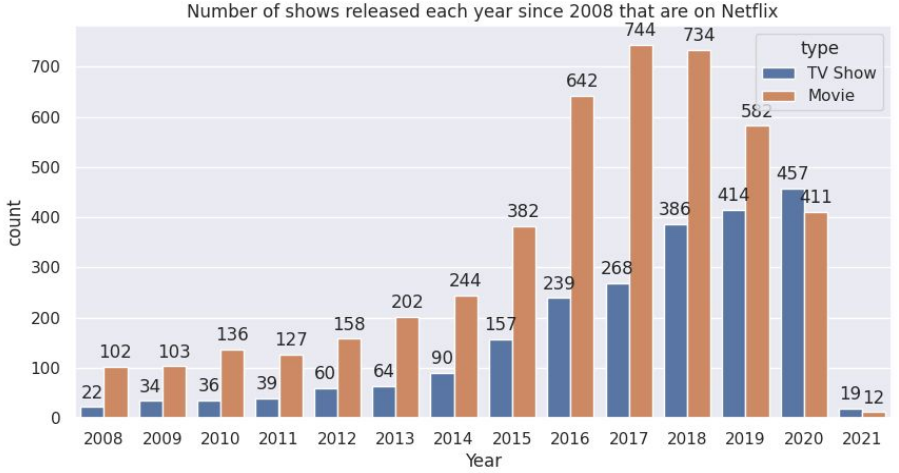
- **The majority of the shows on Netflix are catered to the needs of adult and young adult population.**

(b). Bivariate analysis:

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

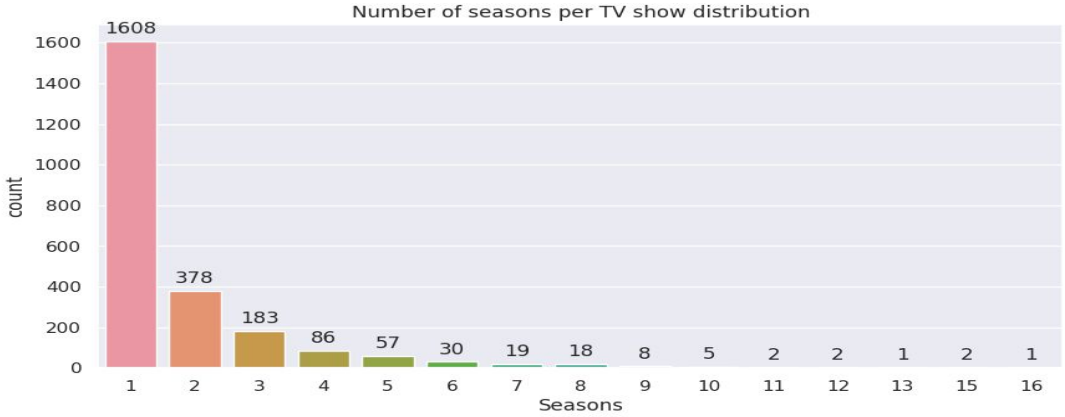


➢ **Number of shows released each year since 2008**



- **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.**

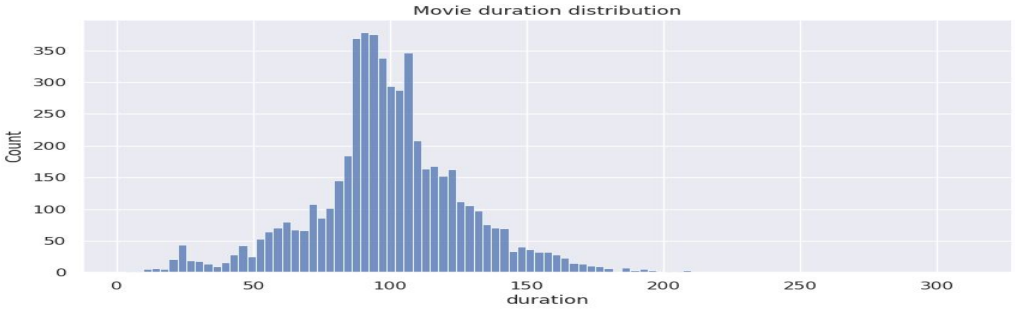
-Seasons in each TV show



-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.

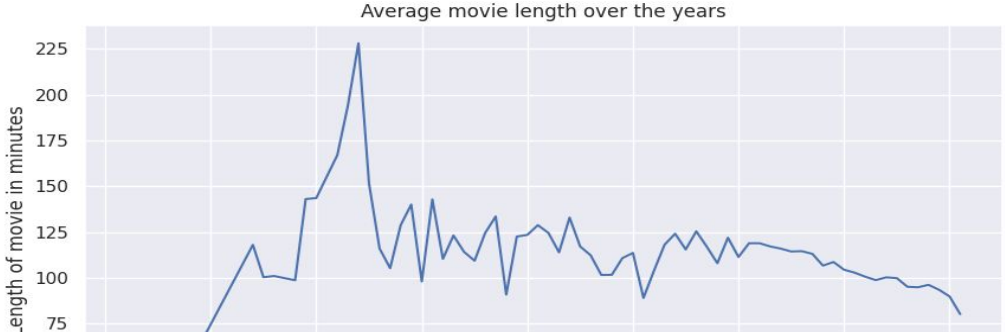
-There are very few TV shows that have more than 8 seasons.

-length of movie analysis



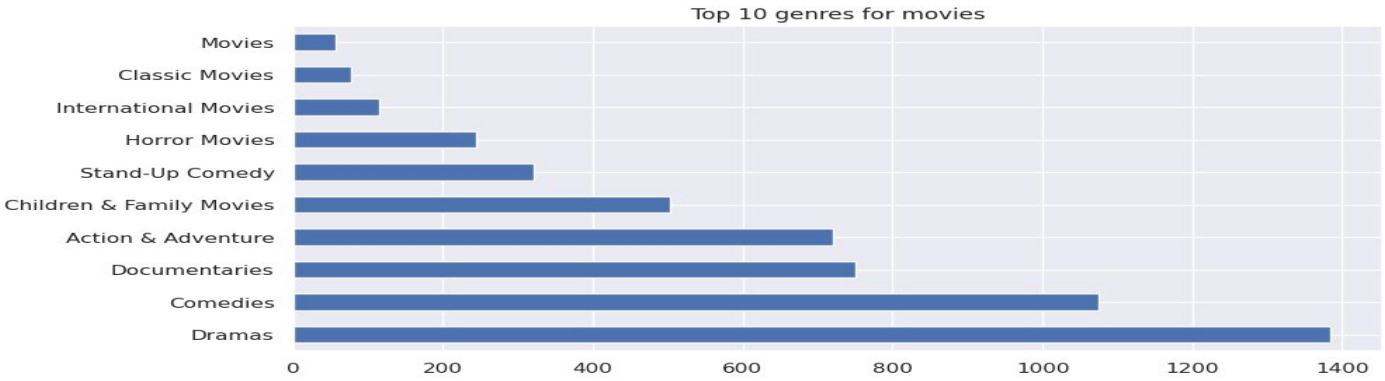
- The length of a movie may range from 3 min to 312 minutes, and the distribution is almost normally distributed.

-Average movie length over the years



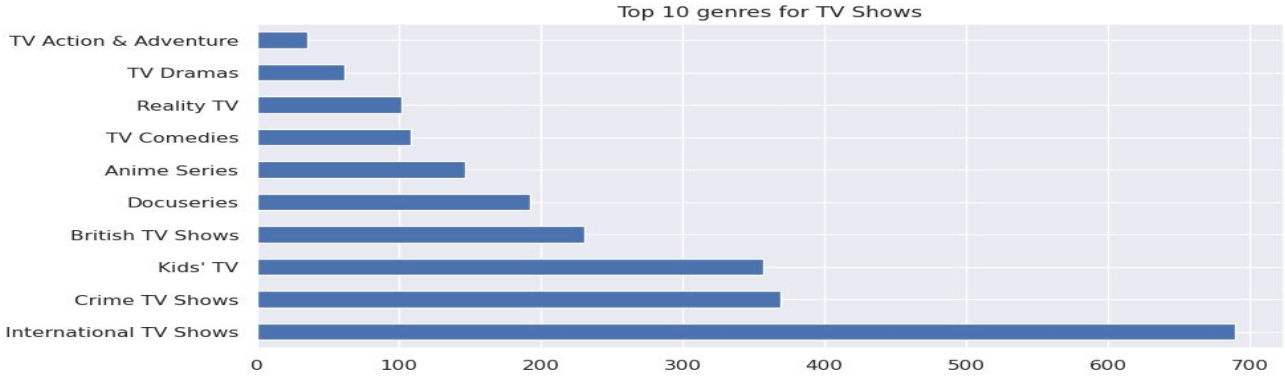
- Netflix has several movies on its site, including those that were released in way back 1942.
- As per the plot, movies made in the 1940s had a fairly short duration on average.
- On average, movies made in the

-Top 10 genre for movies



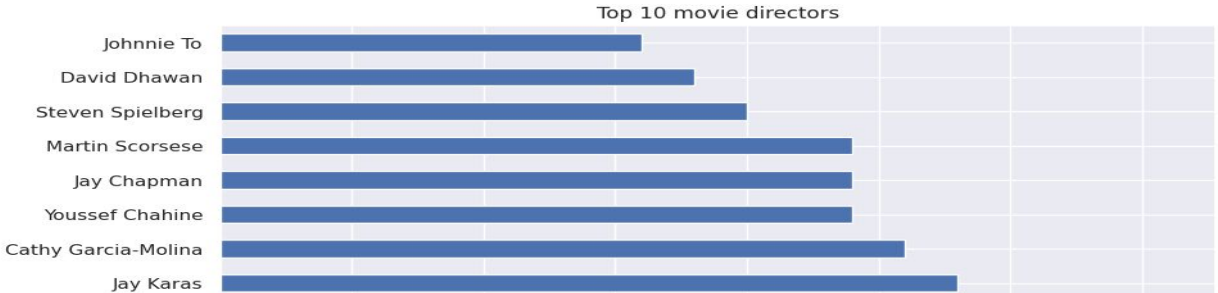
- Dramas, comedies, and documentaries are the most popular genre for the movies on Netflix.

-Top 10 genre for tv shows



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

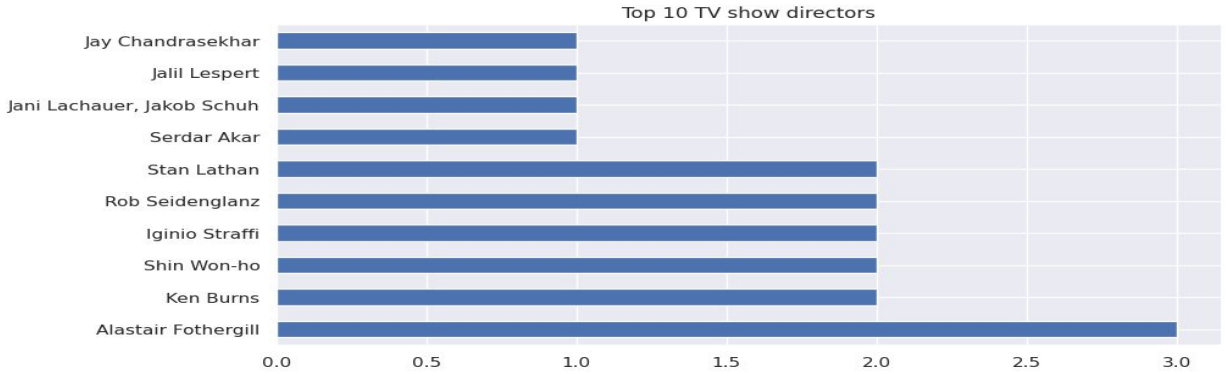
-Top 10 movie directors



-Raul Campos and Jan Suter have together directed in 18 movies, higher than anyone yet.

-This is followed by Marcus Raboy, Jay

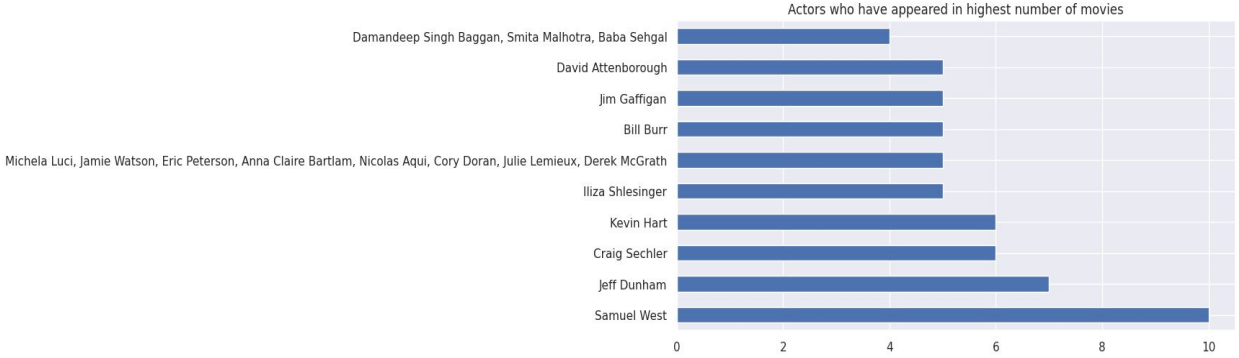
-Top 10 TV show directors



-Alastair Fothergill has directed three TV shows, the most of any director.

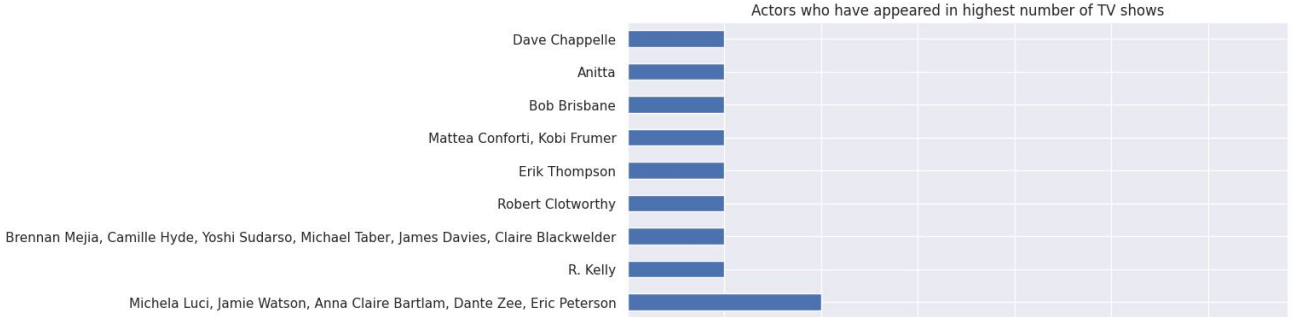
-Only six directors have directed more than one television show

-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

Wordcloud

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

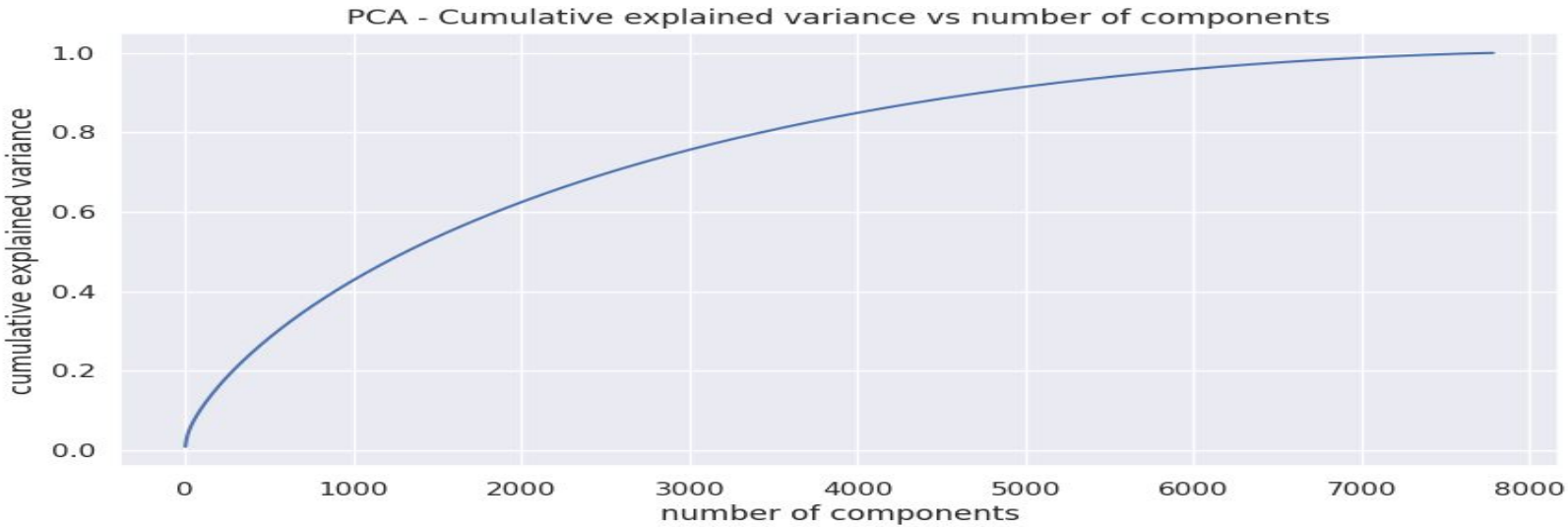
Data Preprocessing

Modelling Approach

- The "Netflix Movies and TV Shows Clustering" project follows a comprehensive modeling approach to effectively cluster shows based on selected attributes. The process begins with text preprocessing, including the removal of non-ASCII characters, stopwords, and punctuation marks, and converting all textual data to lowercase. Lemmatization is then applied to generate meaningful words from the corpus. Tokenization is employed to break down the corpus into individual words, which are then vectorized to represent their numerical values. Dimensionality reduction techniques are applied to simplify the dataset. The project explores various clustering algorithms to group the movies, determining the optimal number of clusters using different methodologies. Finally, the optimal clusters are built, and their contents are visualized using word clouds, providing valuable insights into the distinct characteristics of each cluster.

Dimensionality reduction using PCA

- **variance for different number of components**



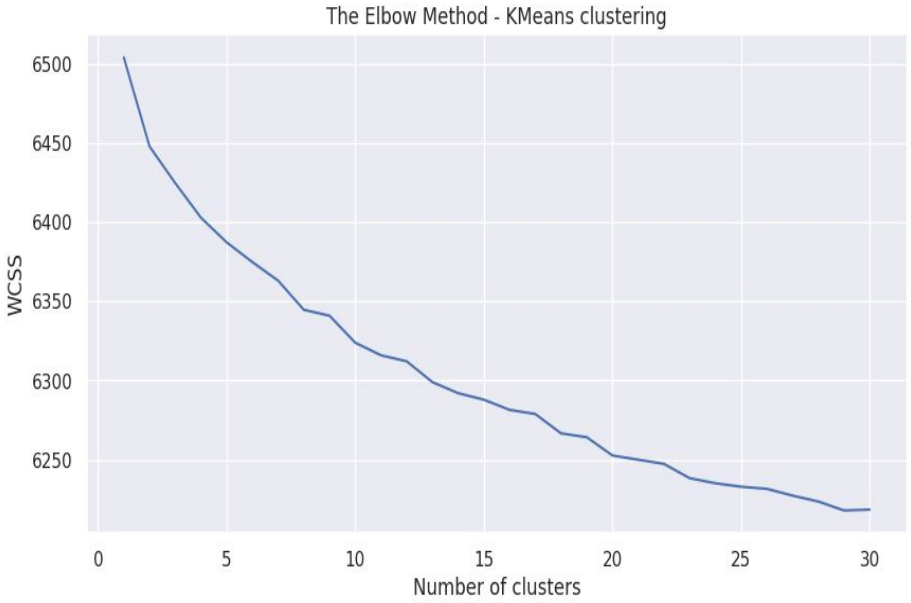
- **We find that 100% of the variance is explained by about ~7500 components.**
- **Also, more than 80% of the variance is explained just by 4000 components.**
- **Hence to simplify the model, and reduce dimensionality, we can take the top 4000 components, which will still be able to capture more than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

Clusters implementation

(a). K-Means Clustering

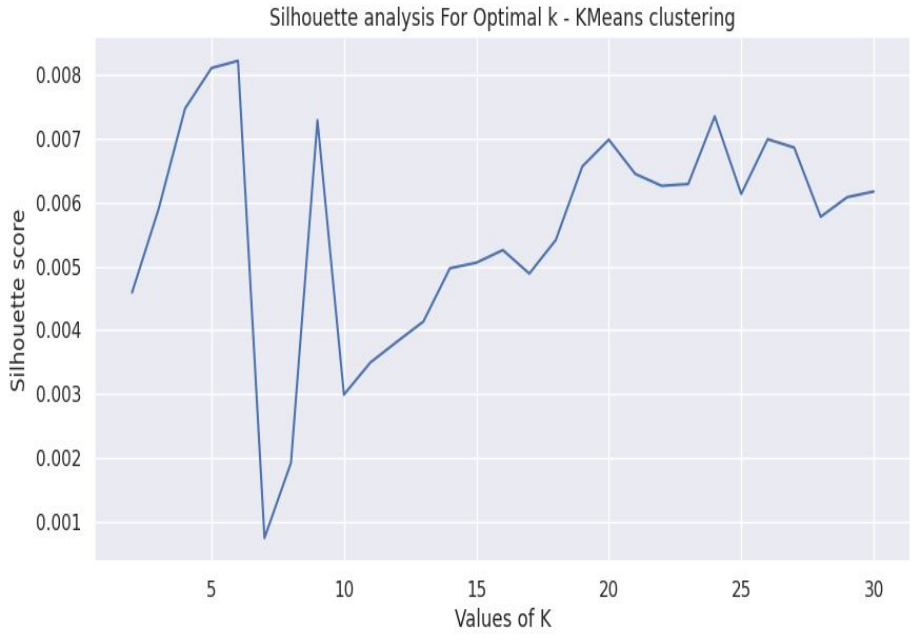
- In this project, we employ the K-means clustering algorithm to build clusters for the movies and TV shows in the Netflix dataset. To determine the optimal number of clusters, we visualize the elbow curve and Silhouette score, which help us make informed decisions about the best configuration for the K-means algorithm. These techniques enable us to effectively group the content and enhance our understanding of the underlying patterns and structures in the data, making content organization and recommendation more efficient and user-centric.

➢ **Elbow method to find the optimal value of k**



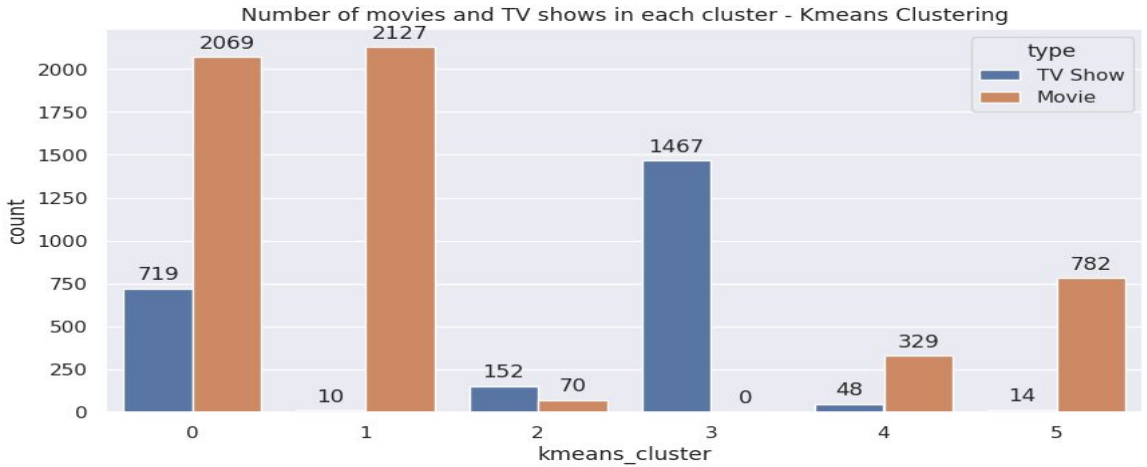
- **The sum of the squared distance between each point and the centroid in a cluster (WCSS) decreases with the increase in the number of clusters.**

➢ **Plotting Silhouette score for different number of clusters**



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

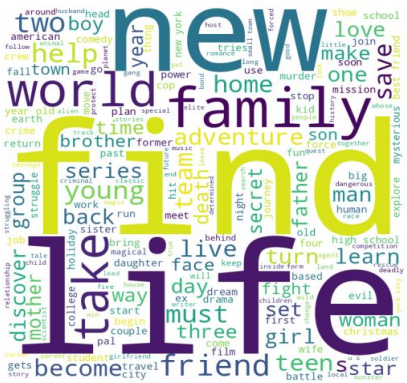
➢ Number of movies and tv shows in each cluster



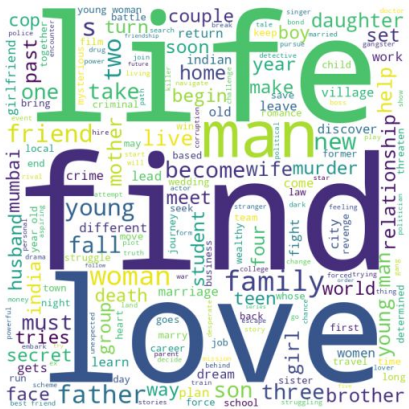
- **This graph represent Successfully built of 6 clusters using the k-means clustering algorithm.**

➢ Word clouds for different 6 clusters built

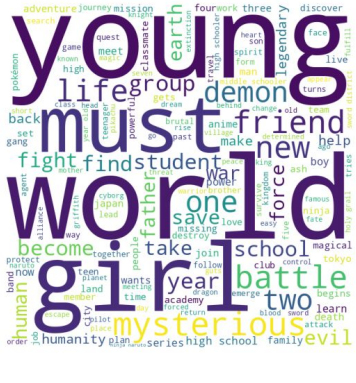
Word cloud for cluster 0



Word cloud for cluster 1



Word cloud for cluster 2



Word cloud for cluster 3

Word cloud for cluster 4

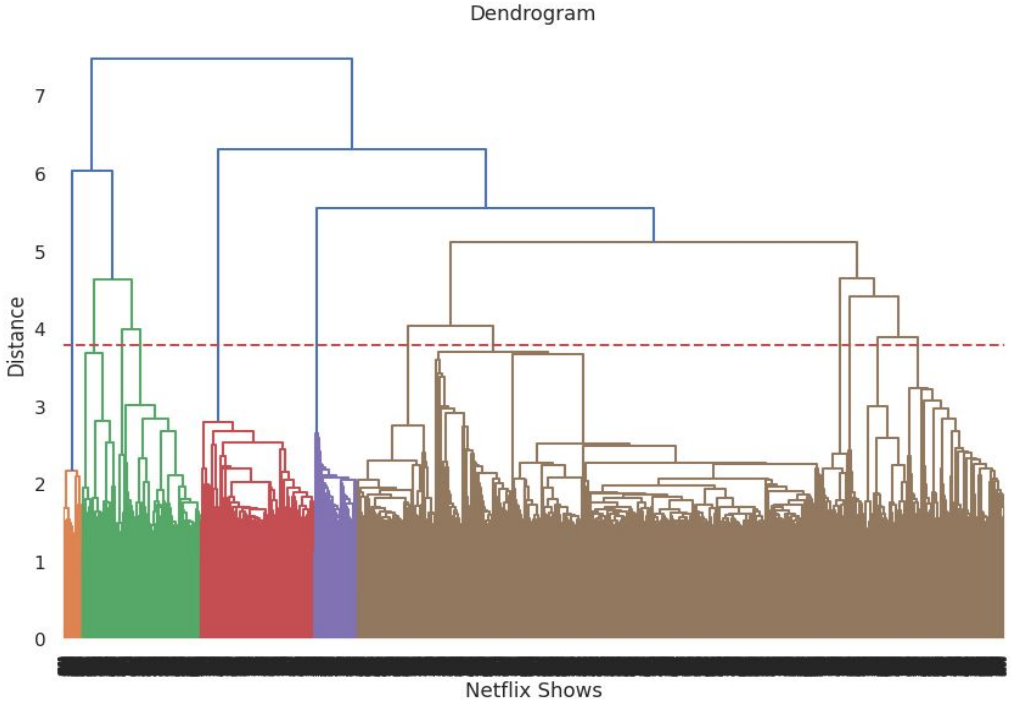
Word cloud for cluster 5

- As seen above, all six cluster clouds showcase distinct themes and content patterns in the Netflix dataset. Cluster 0 focuses on discovery and teenage experiences, Cluster 1 revolves around family relationships and coming-of-age stories, Cluster 2 features supernatural and student life elements, Cluster 3 centers on romantic relationships and adventures, Cluster 4 highlights stand-up comedy, and Cluster 5 is dedicated to documentaries exploring real-life experiences and stories. These clusters offer valuable insights for content organization and personalized recommendations on the platform.

(b). Hierarchical clustering

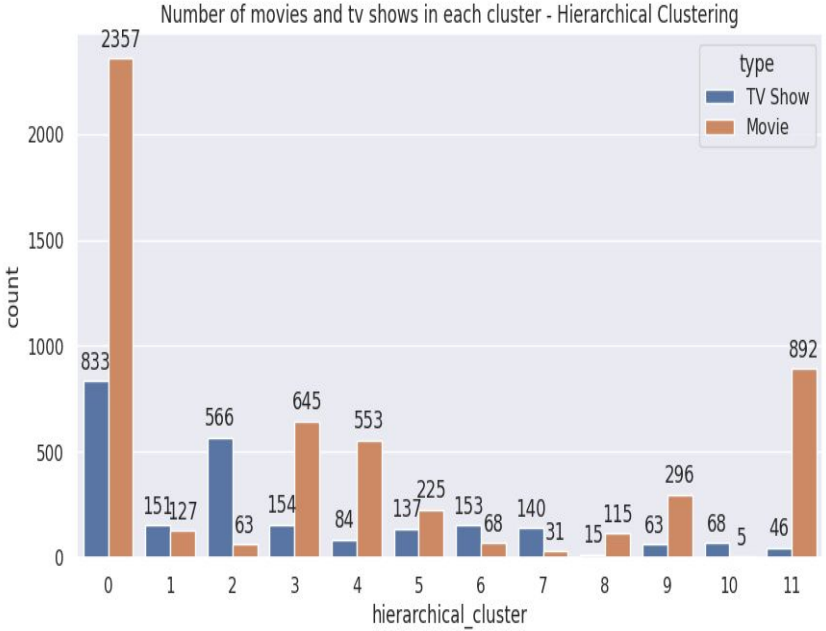
- In this project, we utilize the agglomerative (hierarchical) clustering algorithm to build clusters for the Netflix dataset. To determine the optimal number of clusters, we visualize the dendrogram, which helps us make informed decisions about the ideal configuration for the hierarchical clustering algorithm. These techniques enable us to effectively group the content and gain insights into the underlying patterns and structures in the data, facilitating content organization and personalized recommendations.

➢ Dendrogram to decide on the number of clusters



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

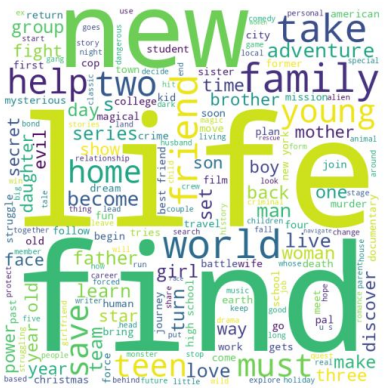
➢ Number of movies and tv shows in each cluster



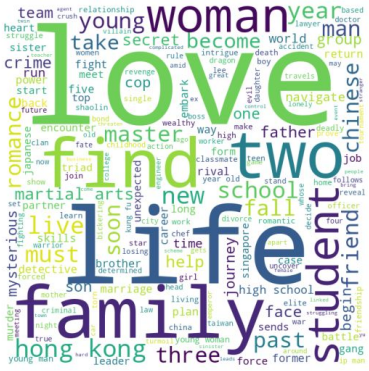
- **The above graph represents the Successfully built 12 clusters using the Agglomerative (hierarchical) clustering algorithm.**

Word clouds for different 12 clusters built

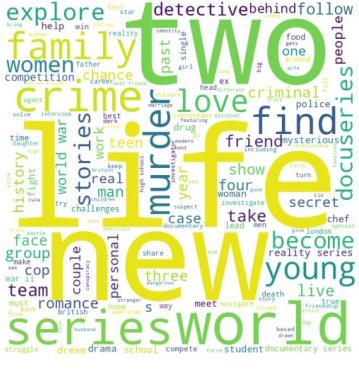
Word cloud for cluster 0



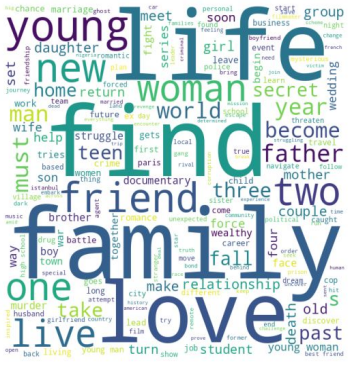
Word cloud for cluster 1



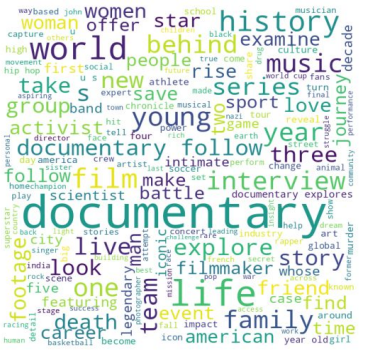
Word cloud for cluster 2



Word cloud for cluster 3

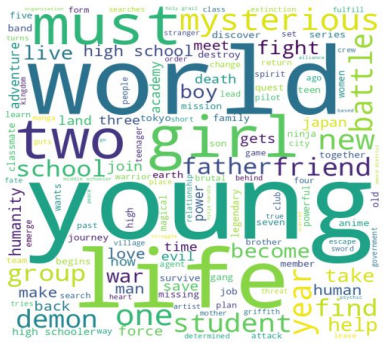


Word cloud for cluster 4



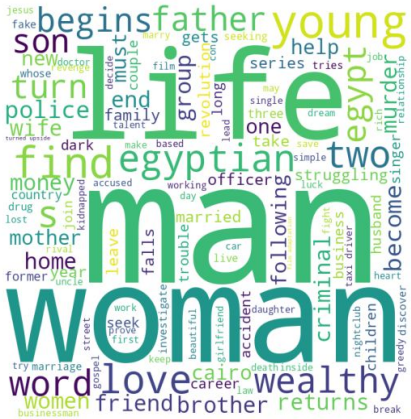
Word cloud for cluster 5

Word cloud for cluster 6



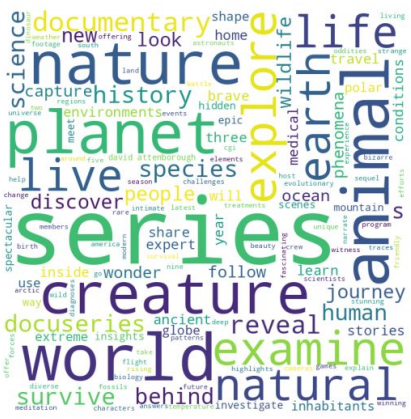
Word cloud for cluster 7

Word cloud for cluster 8

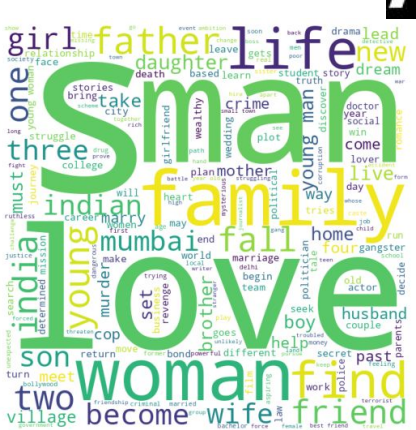


Word cloud for cluster 9

Word cloud for cluster 10



Word cloud for cluster 11



- As seen above, all 12 cluster clouds showcase distinct themes and content patterns in the Netflix dataset. These clusters offer valuable insights into the diverse range of content available on the platform. Cluster 0 presents themes of life, adventure, and discovery, while Cluster 1 revolves around love, family, and romance. Crime series and detective stories are prominent in Cluster 2, and Cluster 3 features content centered on family, friendship, and secrets. Cluster 4 highlights documentaries with a focus on music, history, and family, whereas Cluster 5 delves into family secrets, life in Mexico and Spain, and themes of death. Cluster 6 showcases content with young characters, mysterious elements, and demons, while Cluster 7 is themed around love, life, and secrets in Korea. In Cluster 8, viewers can explore stories of wealth, love, and revolution in Egypt. Stand-up comedy specials and hilarious stories take center stage in Cluster 9, while Cluster 10 is dedicated to nature, exploration, and the wonders of planet Earth. Finally, Cluster 11 features content centered around love, family, and life in India's vibrant city of Mumbai. These clusters provide valuable guidance for content organization and personalized recommendations, enriching the Netflix

Content-based recommender system

Content-based recommender systems are data-driven approaches that offer personalized content recommendations to users based on their preferences. In our project, we build a simple content-based recommender system focusing on the similarity of shows on Netflix. The system aims to recommend a list of similar shows to users who have already watched a particular show.

To determine the similarity score of shows, we employ the Cosine Similarity technique. The similarity between two vectors (A and B) is calculated by taking the dot product of the vectors and dividing it by the magnitude value, as shown in the equation $$Cos(\theta) = \frac{A \cdot B}{|A| \cdot |B|}$$. Essentially, the Cosine Similarity score increases as the angle between the vectors decreases, indicating higher similarity.

Our recommender system leverages Countvectorizer and Cosine Similarity techniques to find similar movies or TV shows based on clustering attributes. By calculating the similarity between shows, the system generates a list of 10 recommended titles for a given show. Users can utilize the 'recommend_10' function, providing the show title as input, to receive the personalized recommendations.

However, it's essential to note that the accuracy of recommendations relies on the quality and size of the dataset used to build the system. A robust and diverse dataset will enhance the system's ability to offer relevant and enjoyable content suggestions, enriching the user experience on Netflix.

Conclusions

- In this project, we addressed a text clustering problem by classifying Netflix shows into distinct clusters based on their similarities and dissimilarities. The dataset comprised approximately 7787 records with 11 attributes, revealing interesting trends such as Netflix hosting a larger number of movies compared to TV shows and exponential growth in the total count of shows over time. A significant portion of the content originated from the United States, with a focus on content targeting the adult and young adult age group.
- To perform clustering, we selected specific attributes, pre-processed, tokenized, and vectorized them using the TFIDF vectorizer, resulting in 20000 attributes. Principal Component Analysis (PCA) was utilized to handle high dimensionality, capturing over 80% of variance with 4000 components.
- The k-means clustering algorithm initially created 6 clusters, determined as optimal through the elbow method and Silhouette score analysis. Additionally, the Agglomerative clustering algorithm 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 system provides users with 10 personalized recommendations based on their previously watched shows, significantly enhancing their experience on the platform. Overall, this project successfully employed various clustering techniques, and the content-based recommender system offered valuable and relevant recommendations, contributing significantly to content management and user