Capstone Project

AI

Netflix Movies and TV Shows Clustering

By

Ravi Kumar

Introduction

AI

With more than 83 million subscribers and presence in more than 190 countries, Netflix is the most popular Internet television network in the world. Its users watch more than 125 million hours of TV and movie content daily, including original series, documentaries, and feature films. On almost any screen that is linked to the Internet, members can watch as much as they want, whenever and wherever. Without interruptions or obligations, members can play, pause, and resume watching at any time.

Index

AI

- Problem statement and their attributes
- Checking for missing values and cleaning the data
- Exploratory data analysis
- Handling Outliers
- Data preprocessing
- Model implementation
- Building content based recommender system
- Summary and Conclusions

Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming services number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Attribute Information:

Show_id : Unique ID for every Movie / Tv Show

Type : Identifier - A Movie or TV Show

Title : Title of the Movie / Tv Show

Director : Director of the Movie

Cast : Actors involved in the movie / TV show

Country : Country where the movie / TV show was produced

Attribute Information:

Date_added : Date it was added on Netflix

Release_year : Actual Release year of the movie / tv show

Rating : TV Rating of the movie / tv show

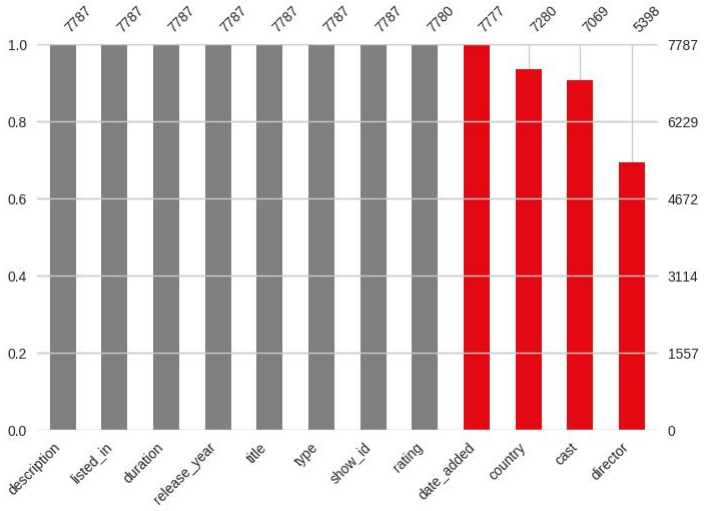
Duration : Total Duration - in minutes or number of seasons

Listed_in : Genres

Description: The Summary description

Checking for missing values and cleaning the data

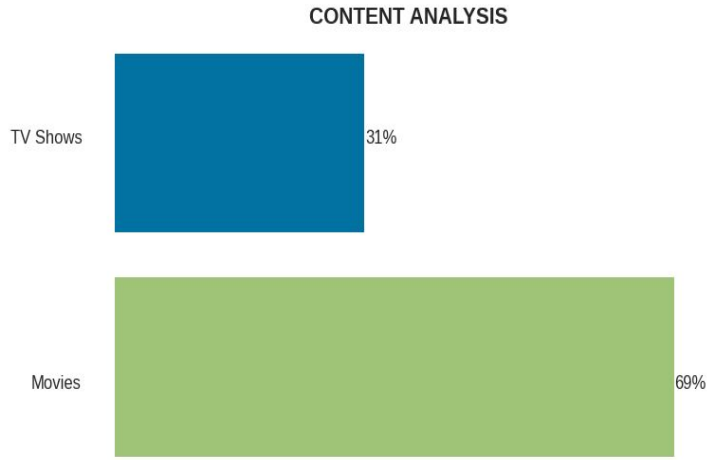
- Director's has the most missing values in our dataset which is followed by cast, country and date_added.
- The null values in the director and country columns are filled with the string "unknown," the cast column is filled with 'no cast', and the mode value is used to fill the null values in the rating column. Finally, the records with null values in the "date_added" column have been removed.



Exploratory Data Analysis

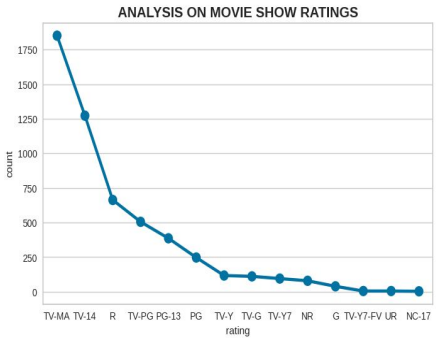
Netflix Content

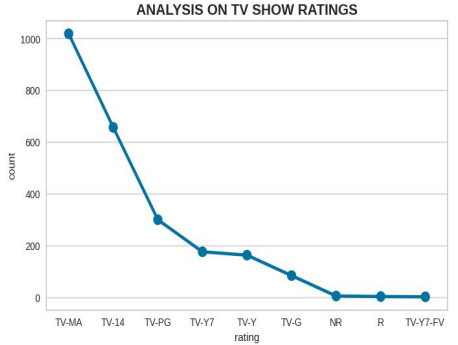
- **Netflix contains 69% movies and 31% television shows, indicating that movie shows have more content.**

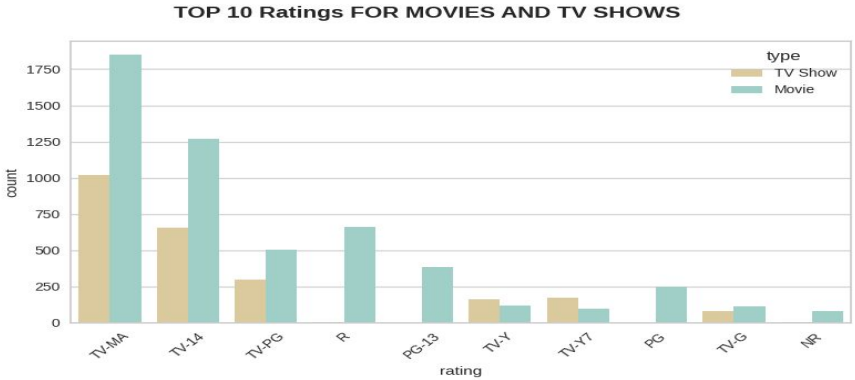


Netflix Ratings

- The most common rating for movies and tv show is TV-MA, which stands for "Mature Audience," followed by TV-14, which stands for "Younger Audience." Since the number of movies is higher than the number of TV shows, as we saw earlier in the type column, movies receive the highest rating when compared to TV shows, which is pretty good.

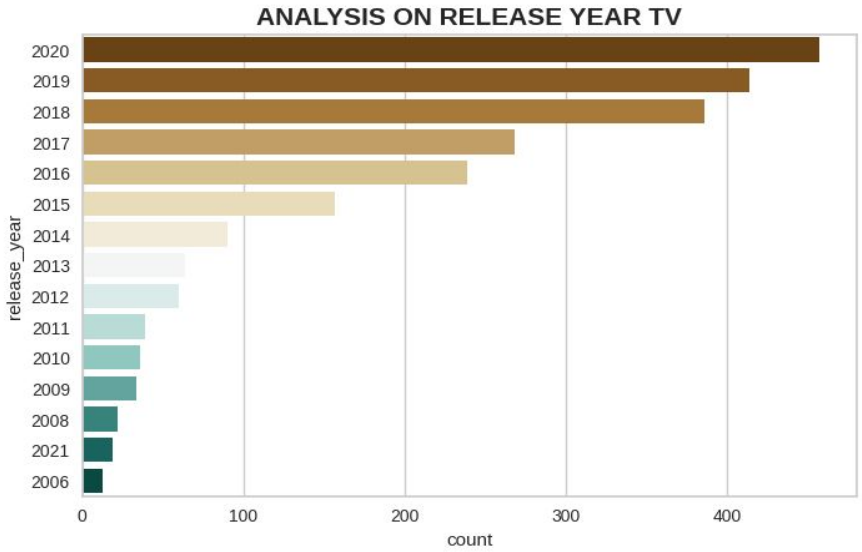






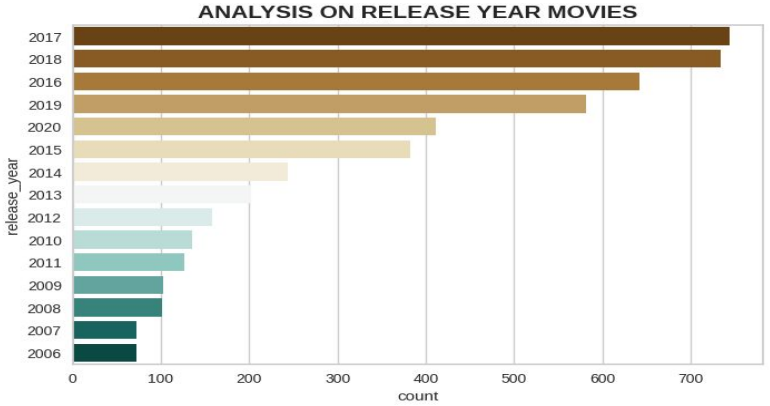
Netflix Release year

- Netflix continues to add more shows on its platform over the years. Highest number of movies released in 2017 and 2018. Highest number of tv shows released in 2019 and 2020.
- There is a decrease in the number of movies 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 2021, hence there are less movies added



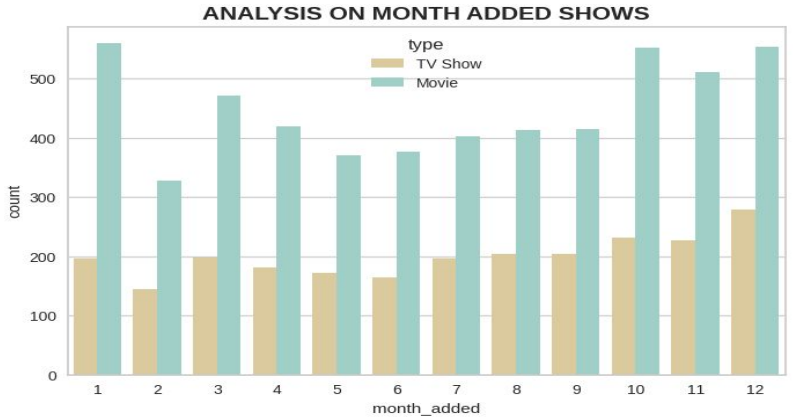
Netflix Release year

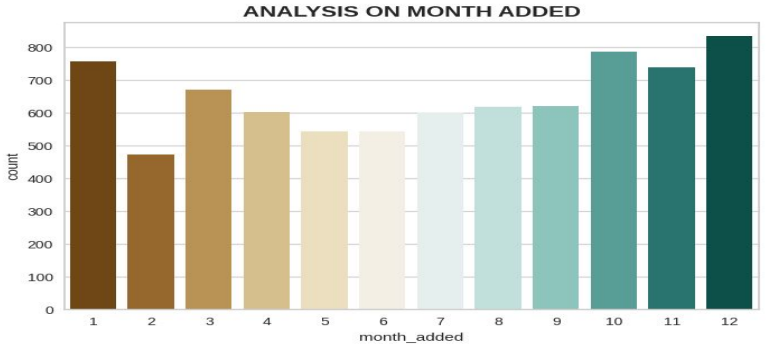
- The number of movies on Netflix is growing significantly faster than the number of TV shows. It appears that Netflix has focused more attention on increasing Movie content that TV Shows. Movies have increased much more dramatically than TV shows.



Netflix Release Month

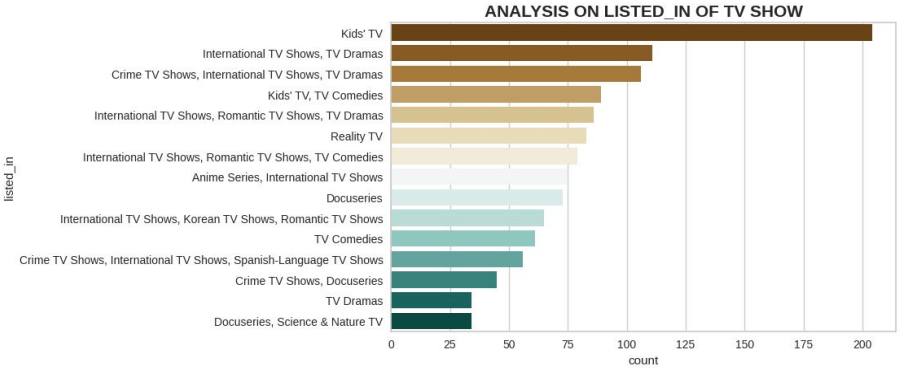
- **The Christmas season (October, November, December, and January) sees a greater amount of content released. Compared to TV shows, more movies are released each month.**

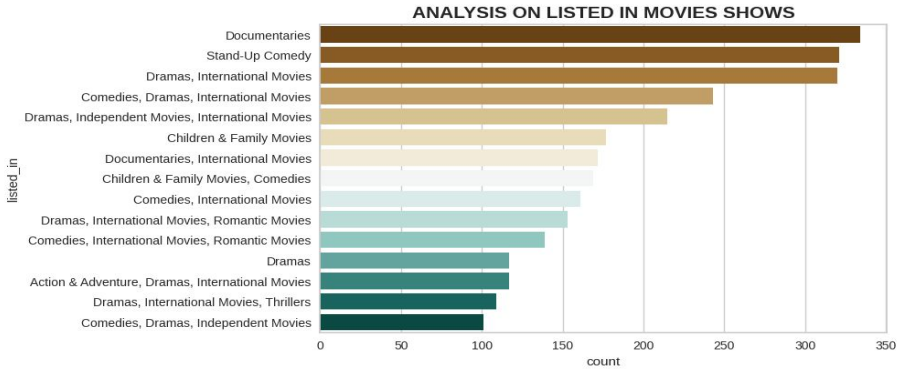




Netflix Genres

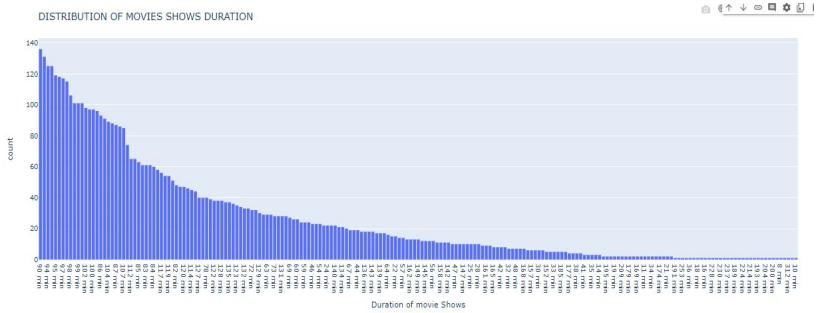
- The most popular Netflix category is documentaries, which are followed by stand-up comedy, Drams, and foreign films.
- The most popular Netflix TV show category is kids TV.

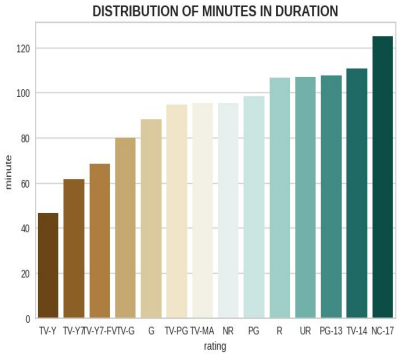


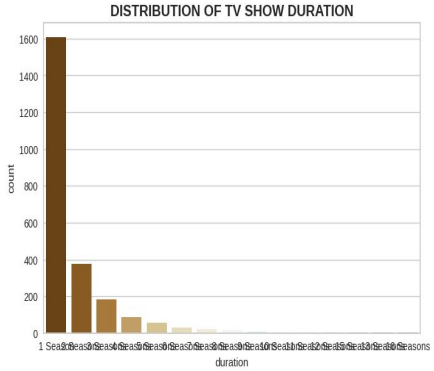


Netflix Duration

- The majority of movies have a duration between 90 and 120 minutes.
- The Majority of tv shows consisting of single season.
- The lengthiest average runtimes are found in NC-17 rated movies(adults only like dreamers, lust and caution). The average duration of movies with a TV-Y rating(childrens of all ages like cartoons or animated films) is the shortest.

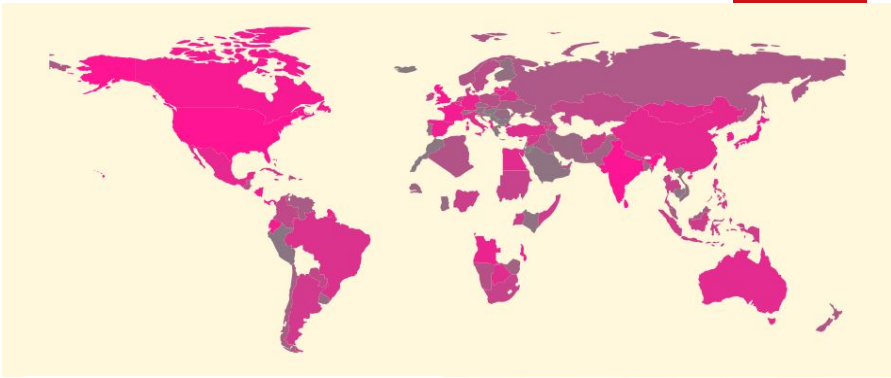


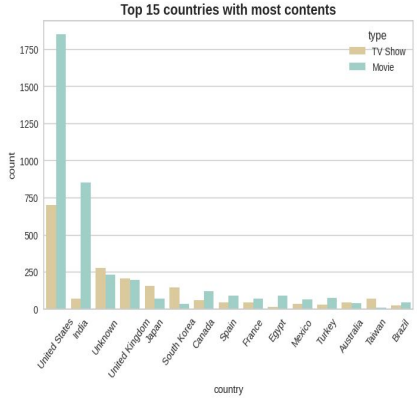


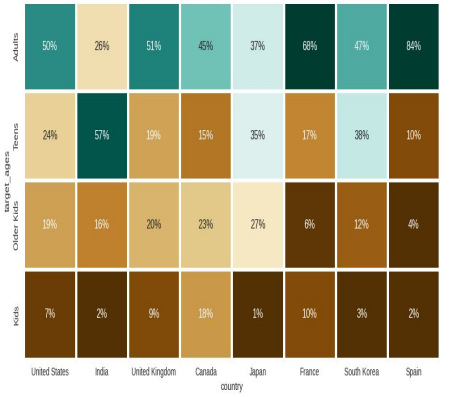


Netflix country wise

- The graph visualisations show the content's country of origin, which include both Movies and TVs shows. Top of the list of nations were the US and India. A few countries, including Australia, Taiwan, and Brazil, produce little Netflix content.
- From the heatmap,the US and UK are very similar to the Netflix target age group, although they differ greatly from such as India or Japan.

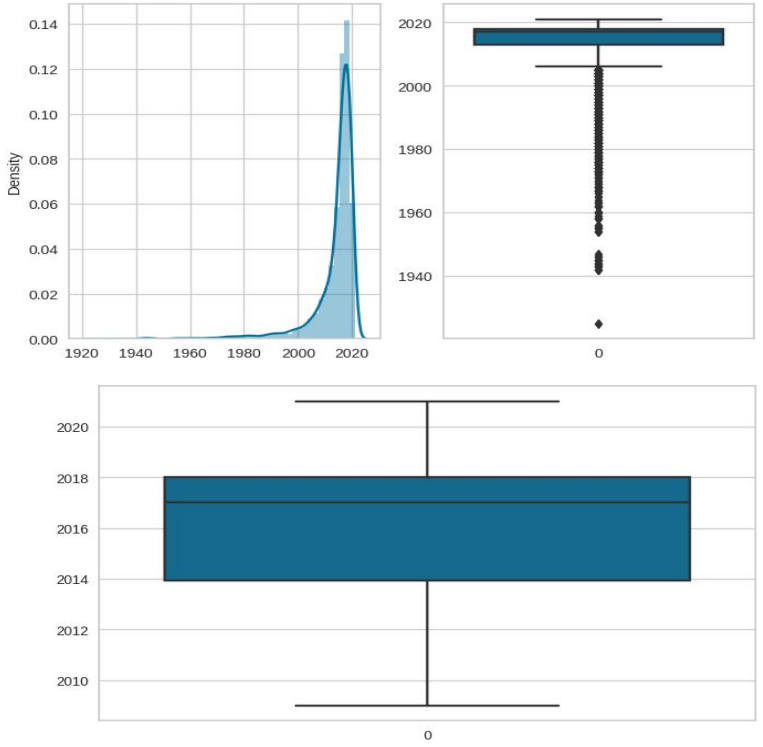






Handling Outliers

- The figures (release_year less than 2009) are being displayed as outliers
- Replacing outliers with mean value



Data Preprocessing

AI

Modelling Approach:

- **Select the attributes based on which you want to cluster the shows**
- **Text preprocessing: Remove all stopwords and punctuation marks, convert all textual data to lowercase**
- **Lemmatization to generate a meaningful word out of corpus of words**
- **Tokenization of corpus**
- **Word vectorization**
- **Dimensionality reduction**
- **Use different algorithms to cluster the movies, obtain the optimal number of clusters using different techniques**
- **Build an optimal number of clusters and visualise the contents of each cluster using word clouds**

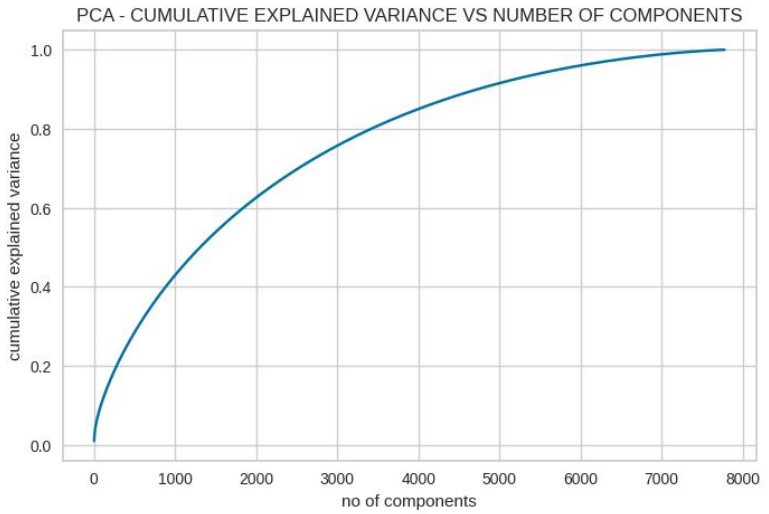
Data Preprocessing

AI

- The director, cast, country, listed_in, and description are chosen as the attributes to cluster
- Removing non-ascii characters
- Removing stop words and converting to lowercase
- Removing punctuation marks
- Lemmatization, tokenization and text vectorization

Dimensionality reduction using PCA

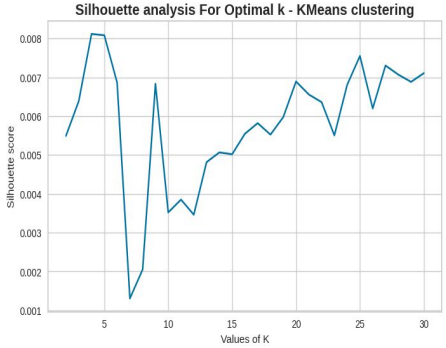
- We find that around 7500 components account for 100% of the variance
- Also, just 4000 components comprise more than 80% of the variation
- As a result, we can pull the top 4000 components out of the model to make it simpler and less dimensional while still being able to account for more than 80% of variance.

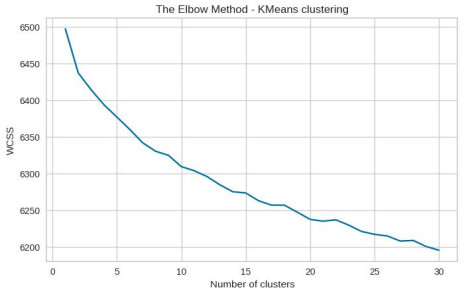


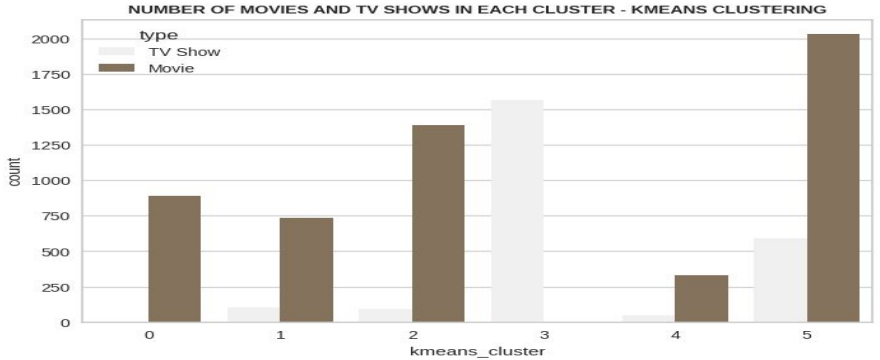
Model Implementation

K-Means Clustering

- A well-liked unsupervised machine learning method for combining comparable data points is called K-means clustering. K-means clustering aims to divide a dataset into k clusters, each of which is represented by its centroid and contains similar data points.
- Finding the ideal number of clusters for the K-means clustering algorithm requires visualising the elbow curve and Silhouette score.

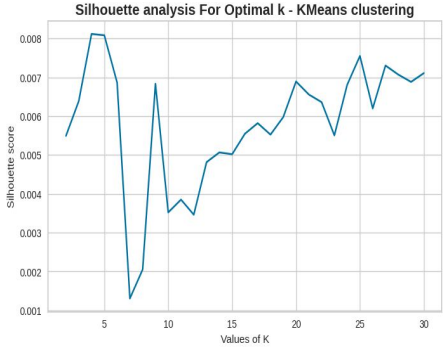


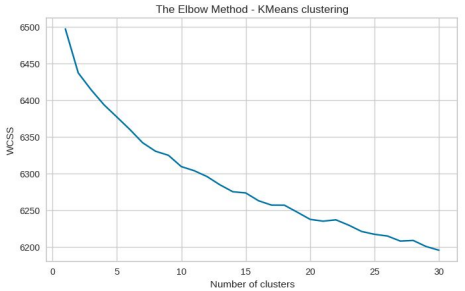


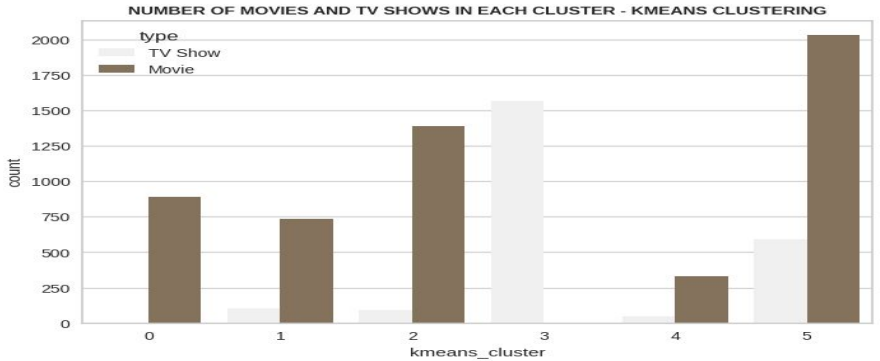


K-Means Clustering

- With more clusters, there is a decrease in the sum of squared distances between each point and the centroid.
- Six clusters receive the highest Silhouette score overall.

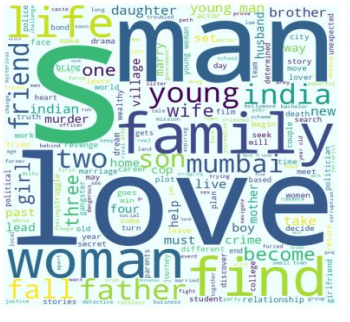






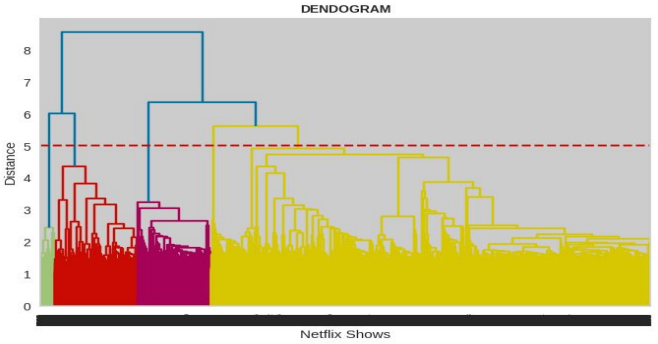
Word Clouds: K Means clusters

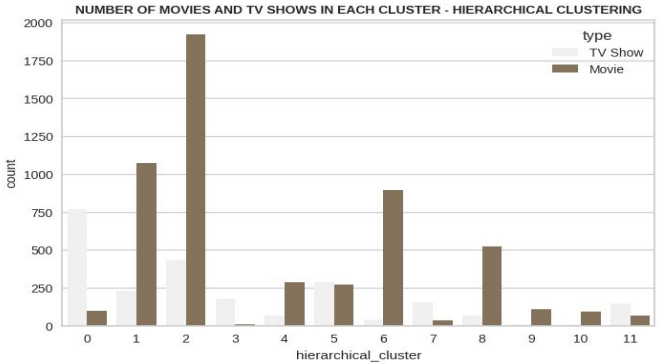
- **Important keywords observed in clusters**



Hierarchical clustering

- the agglomerative (hierarchical) clustering process for building clusters. Using the dendrogram to visualize the agglomerative (hierarchical) clustering process to determine the ideal number of clusters.
- Utilising the Agglomerative (hierarchical) clustering algorithm, 12 clusters were successfully built

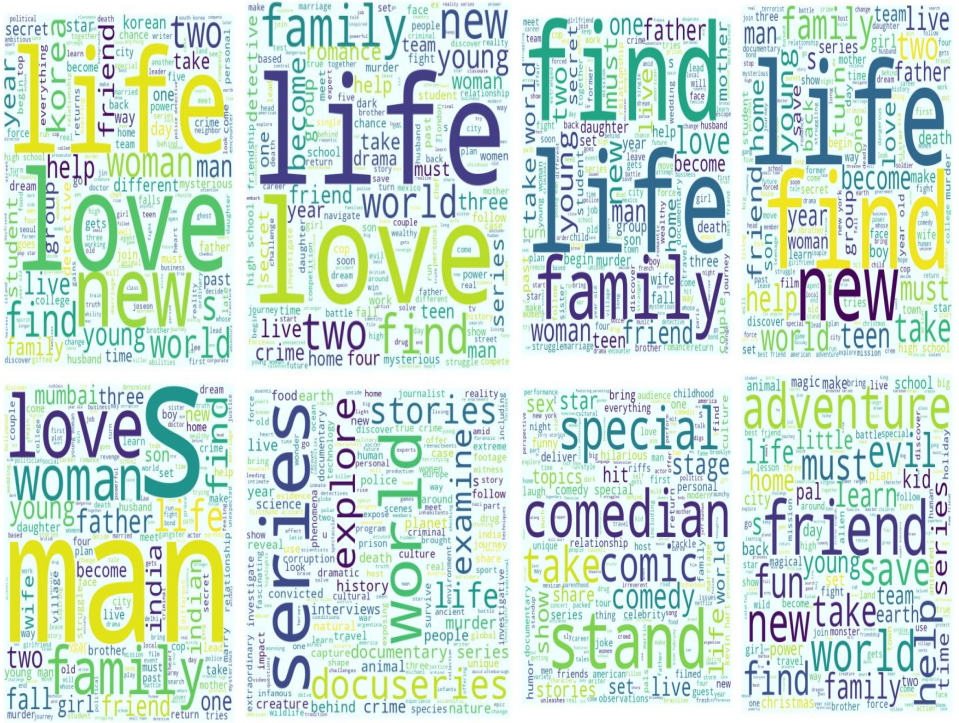




Word Clouds: Hierarchical clustering

AI

- **Important keywords observed in clusters**



Building Content based recommender system

- The objective of a Recommender System is to recommend relevant items for users, based on their preference. If a person has watched a show on netflix, then the recommender system must be able to recommend a list of similar shows.
- We have used cosine similarity to recommend top 10 movies.

```
# Recommendations for 'Lucifer'
recommend_top10('Lucifer')
```

```
If you like this 'Lucifer', you may also enjoy:

['Get Shorty',
 'The Good Cop',
 'Rica, Famosa, Latina',
 'The Expanding Universe of Ashley Garcia',
 'Better Call Saul',
 'Jack Taylor',
 'Dramaworld',
 'Love Rhythms - Accidental Daddy',
 'L.A.’s Finest',
 "Marvel's Iron Fist"]
```

```
# Recommendations for 'abc'
recommend_top10('fun')
```

```
'Invalid Entry'
```

Summary and Conclusion:

AI

In this project, we tackled a text clustering issue where we had to categorize Netflix shows into specific clusters such that the shows within a cluster are similar to one another and the shows in different clusters are dissimilar to one another.

- **Once our dataset is loaded, and then we search for duplicates and missing values. No duplicate values were discovered, and any missing values were used to fill them in. In our dataset, the director column contains the most missing entries, followed by cast, country, and date_added. The string "unknown" is used to fill missing values in the director and country columns, "no cast" is used fill in the cast column, and the mode value is used to fill missing values in the rating column. the records that had null entries in the "date_added" column were deleted.**

Summary and Conclusion:

- 31% of Netflix's content is television shows, while 69% of it is movie show, demonstrating that movie shows have greater content. TV-MA, which stands for "Mature Audience," is the most frequently used classification for movie and tv shows, followed by TV-14, which stands for "Younger Audience." Since the number of movie shows is higher than the number of TV shows, movie shows receive the highest rating when compared to TV shows, from this we can say people like to watch movie show than compare to tv shows.
- Over the years, Netflix has added more shows to its platform. Most movies were released in 2017 and 2018. Most television shows were broadcast in 2019 and 2020. The covid-19-induced lockdowns that stopped the production of shows may be to blame for the decline in the number of movies added in the year 2020. There are fewer movies uploaded this year because the Netflix data we have only extends through 2021.

Summary and Conclusion:

- Netflix's movie show library is expanding much more quickly than its TV show library. It looks that Netflix has prioritised adding more movie material over TV shows. The growth of movies has been significantly more pronounced than that of TV shows.More content is released over the Christmas season (October, November, December, and January). There are more movies released each month compared to TV shows.Documentaries are the most popular Netflix category, followed by stand-up comedy, dramas, and foreign films. Kids TV is the most well-liked Netflix TV shows.
- The majority of movies durations last between 90 and 120 minutes. Most tv shows have just one season. The lengthiest average runtimes are found in NC-17 rated movies. The average duration of movies with a TV-Y rating is the shortest. The geograph visualisations show that the United States and India are the two countries that produce the most content.

Summary and Conclusion:

AI

- The director, cast, country, genre, and description are chosen as the attributes to cluster the data based on. These attributes' values underwent tokenization, preprocessing, and vectorization using TFIDF vectorizer. A total of 20000 characteristics were produced through TFIDF vectorization. For the purpose of overcoming the dimensionality curse, we applied Principal Component Analysis (PCA). 4000 components were able to capture more than 80% of variance.
- The ideal number of clusters was found to be six when we first created clusters using the k-means clustering technique. The elbow method and Silhouette score analysis were used to get this result.The Agglomerative clustering technique was then used to create clusters, with 12 being the optimum number. The dendrogram was visualised to achieve this.The similarity matrix acquired after utilising cosine similarity was used to construct a content-based recommender system. Based on the sort of show the user viewed, this recommender system will provide them with 10 recommendations.

Thank you