Capstone Project -4

Netflix Movies and TV Shows Clustering

Submitted by

SANDIP DEY

Data science trainee, Alma better

Content :

- Introduction
- Problem Statement
- Data Summary
- Data Cleaning
- Data processing
- Exploratory Data Analysis (EDA)
- Hypothesis testing
- Feature Engineering
- Dimensionality Reduction
- Clustering
- Word cloud on Clusters
- Build Recommendation System
- Conclusions

Introduction :

NETFLIX

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time.

METHODOLOGY

Unsupervised Machine Learning (Clustering)

DATABASE

Netflix Movies and TV Shows
7787 rows and 12 columns
Data from last decade

Problem Statement :

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from able 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 service's 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.

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

Points to discuss

- Data description
- Exploratory data analysis
- Hypothesis testing
- Feature selection

Machine learning algorithms(unsupervised)

- 1. K-mean
- 2.agglomerative clustering

Model performance

Data Summary :

- **The dataset consists of listings of all the movies and TV shows available on Netflix, along with details such as – cast, directors, ratings, release year, duration.**

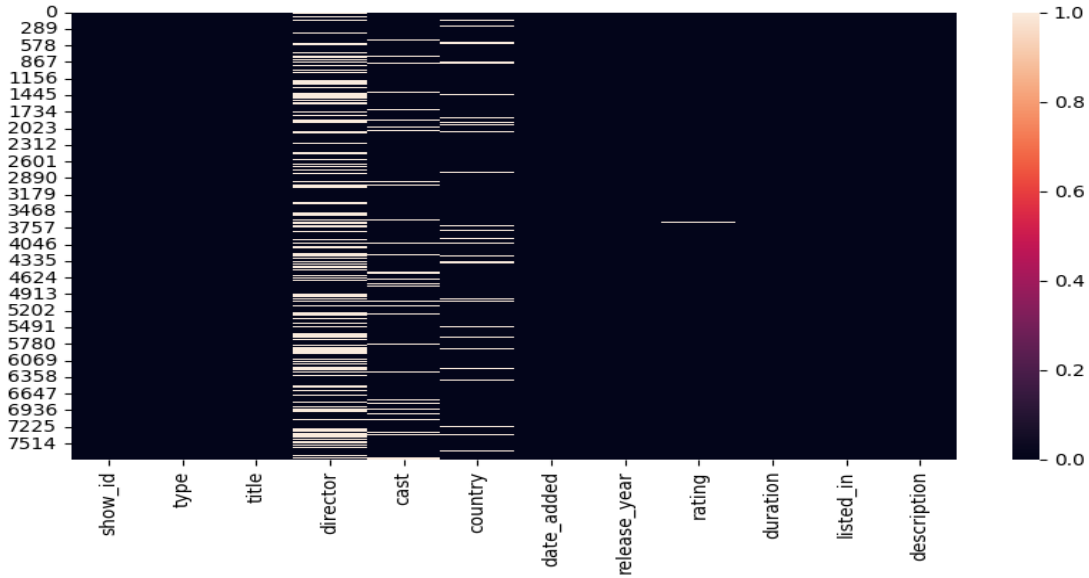
- **Show id** : Unique ID for every Movie / TV Show
- **type** : Identifier – A Movie or TV Show
- **title** : Title of the Movie / TV Show
- **director** : Director of the Movie
- **cast** : Actors involved in the movie / show
- **country** : Country where the movie / show was produced
- **Date added** : Date it was added on Netflix
- **Release year** : Actual Release Year of the movie / show
- **rating** : TV Rating of the movie / show
- **duration** : Total Duration – in minutes or number of seasons
- **Listed in** : Genre
- **description**: The Summary description

Data Cleaning :

Handling Missing values:

- Director(2389),cast(718),and country(507)–replace with **'Unknown'**
- Date added(10)–**dropped**.
- Rating(7)–**mode** imputation**.**

We have successfully handled all the missing values in the dataset

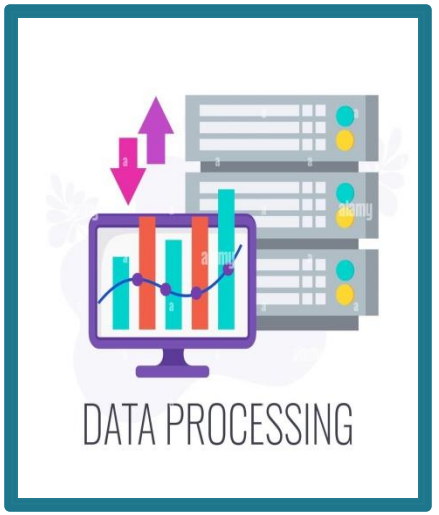


Null Values

Show-id	0
type	0
title	0
director	2389
cast	718
country	507
Date-added	10
Release-year	0
rating	7
duration	0
listed-in	0
description	0

Data processing :

- **Country, listed in:**
  - There are some movies / TV shows that were filmed in multiple countries, have multiple genres associated with it. we split it and stack in a new dataset for further analysis.



- **Typecasting 'duration' from string to integer**
  - Converted the data-type of duration column to INT using apply function.

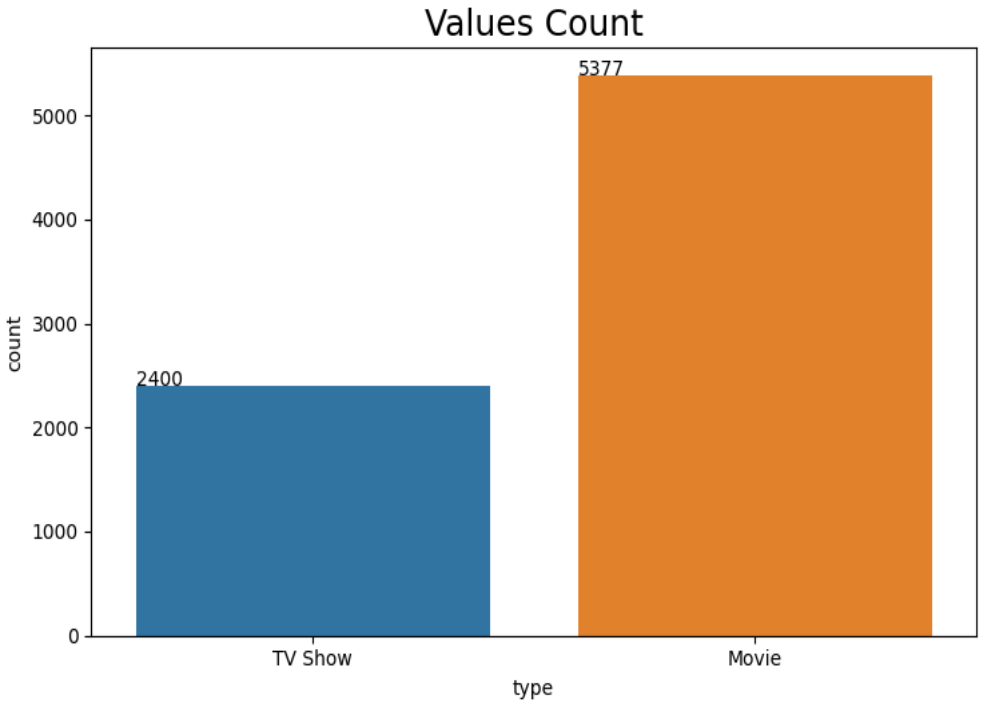
.

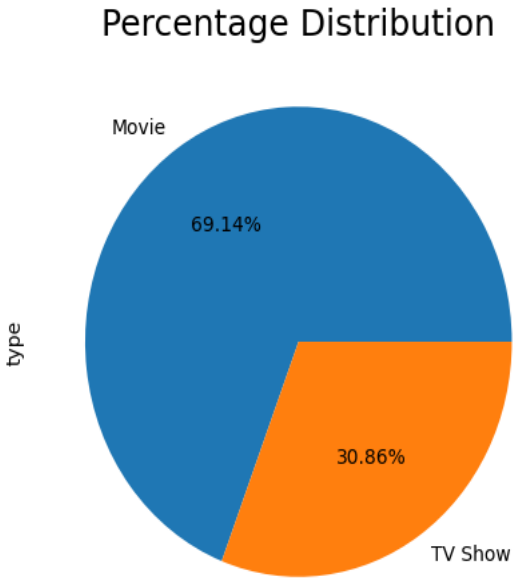
- **Typecasting 'date-added' from string to date time**
  - The shows were added on Netflix between 1st January 2008 to 16th January 2021, and we convert date_added column string to datetime fomate.

- **The ratings can be changed to age restrictions that apply on certain movies and TV shows.**
  - The data set contained separate age ratings for movies and TV shows and were replaced with values of:'Adults','Teens','YoungAdults','OlderKids','Kids'

EDA(Content Type column) :

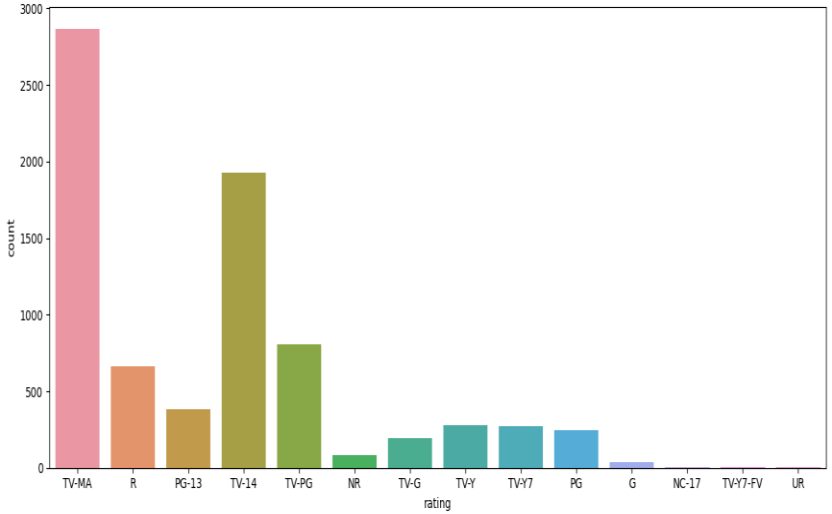
- Netflix has 5372 movies and 2398 TV shows, there are more number movies on Netflix than TV shows. 69% of data belong from Movie class and 31% of data belong from TV shows, Greater number of count belong from movie class than TV show class.

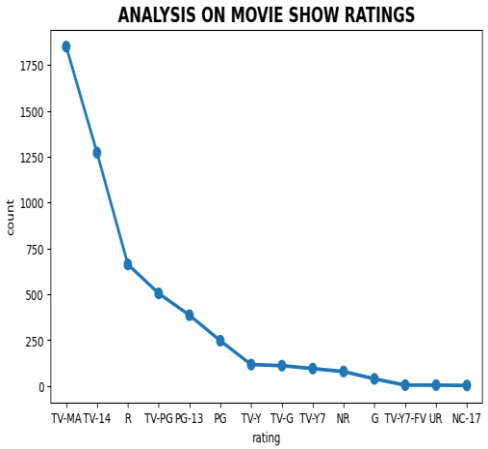


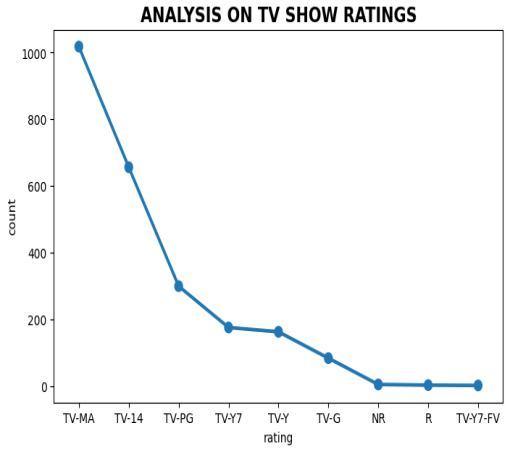


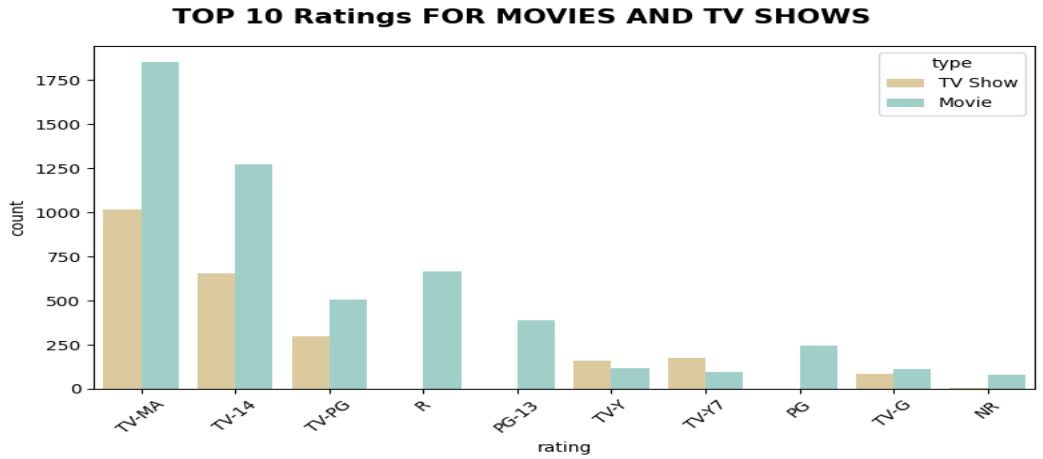
EDA(Netflix Ratings):

- The most common rating for movies and television show is TV-MA, which stands for "Mature Audience or adults," 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 obvious.





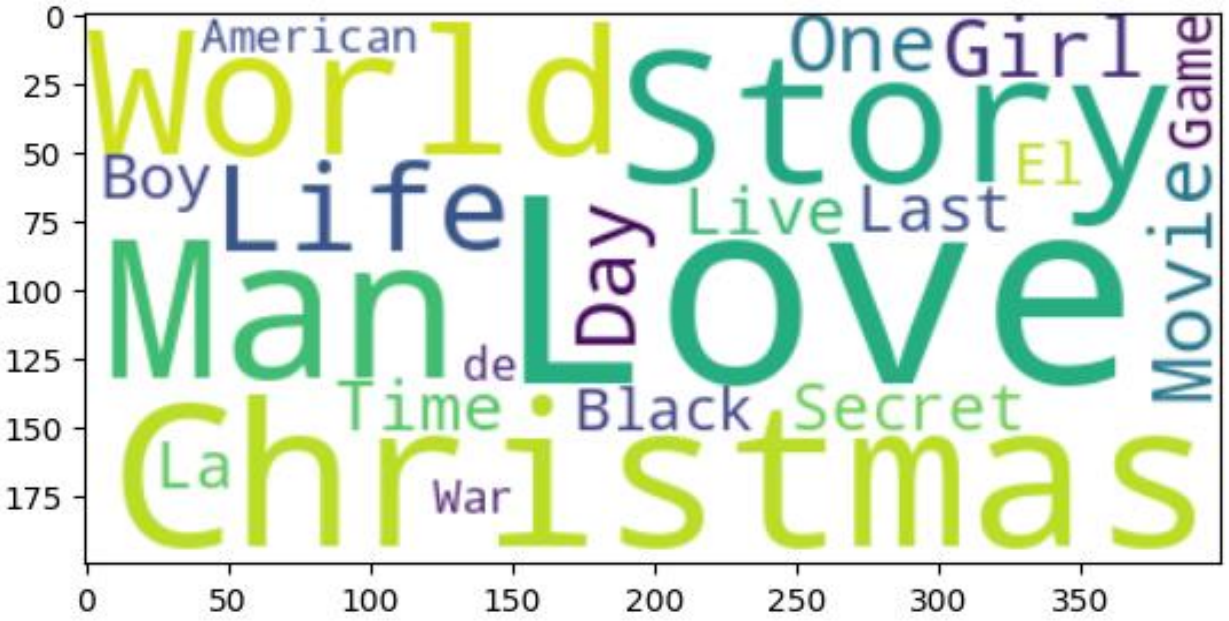




EDA(Most occurred word in Title column) :

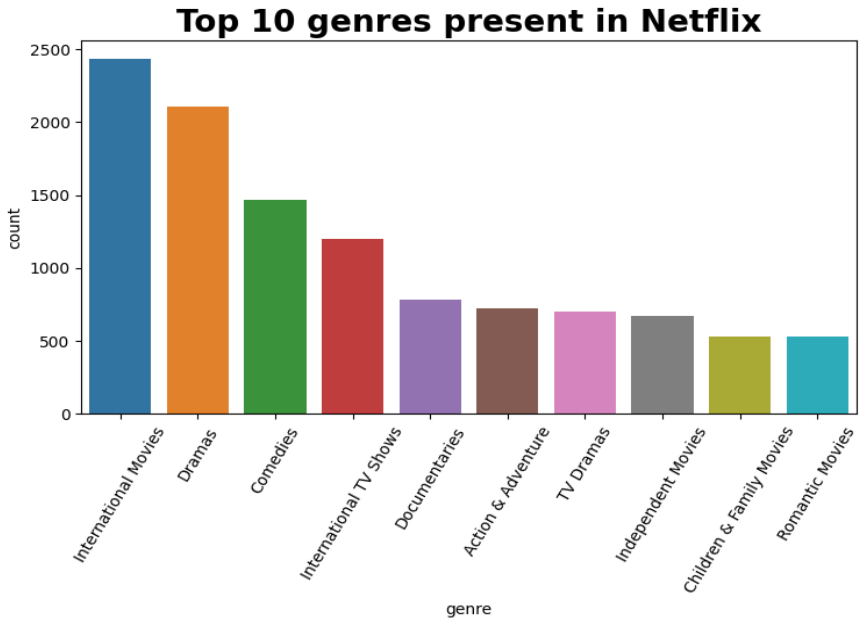
Words like 'Love', 'Christmas', 'Man', 'World', 'Life', 'Boy', 'Time', and 'Story' are frequently used in the movie title column.

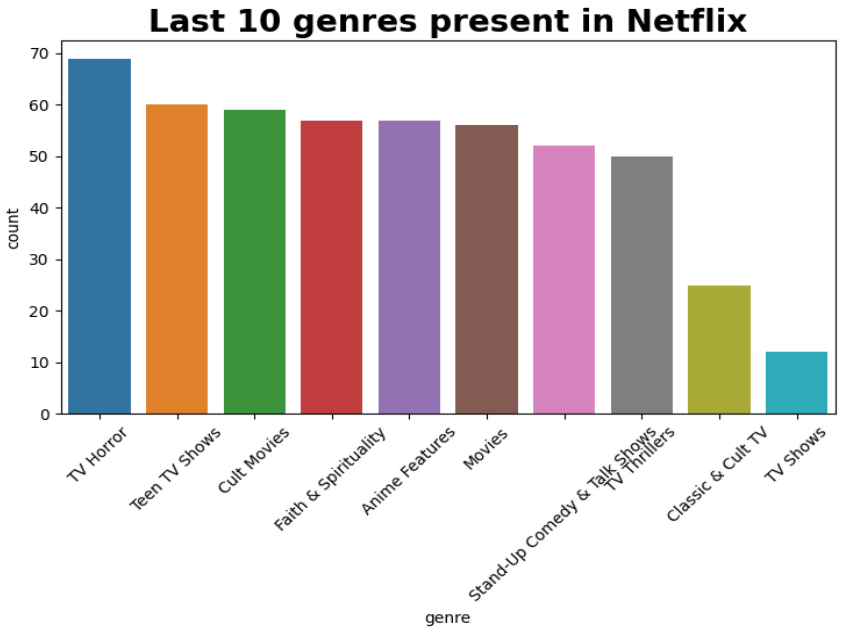
we can see that the most of the movie title contain 'Love', 'Christmas', 'Man' , ‘word-Its’ attract more people to watch the movie



EDA(Netflix Genres):

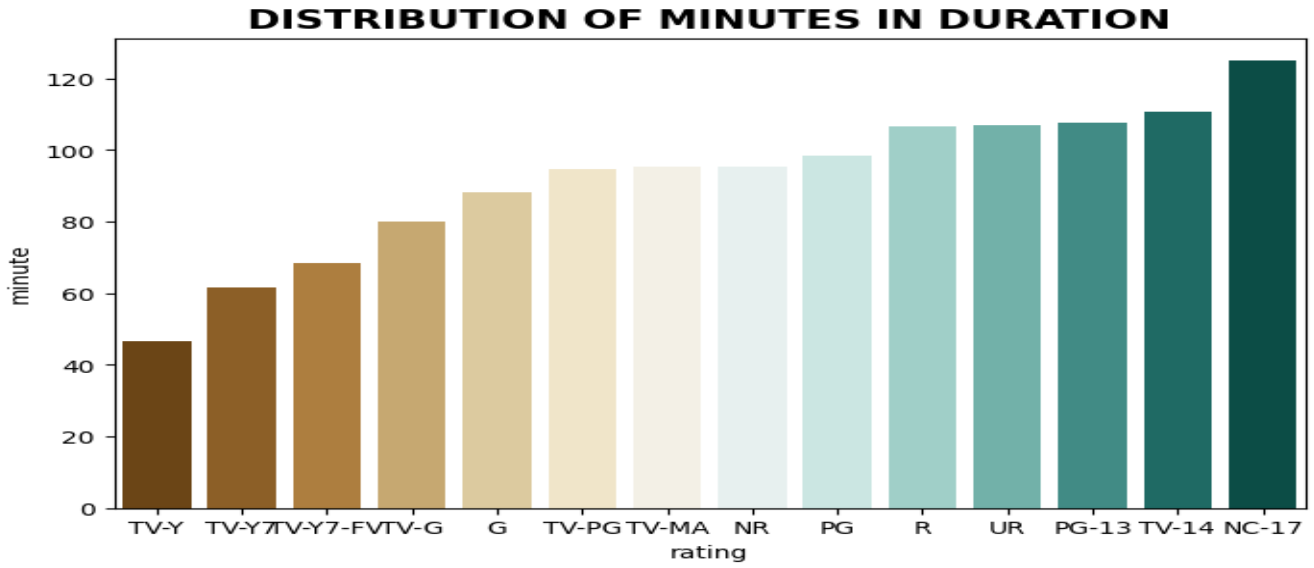
- International Movies, Dramas, and Comedies make up the majority of the genres.
- TV Shows, Classic and cult TV, TV thrillers, Stand-Up comedy, and Talk shows account for the least genres.





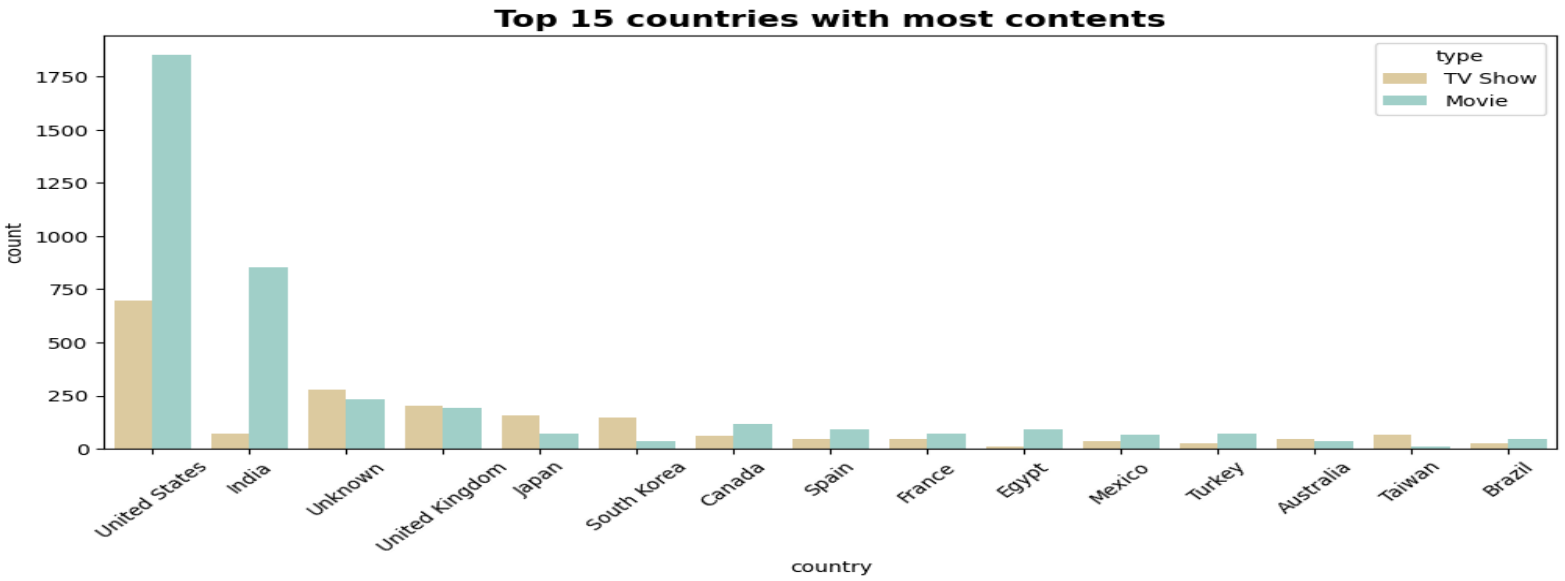
EDA(Netflix Duration):

- The majority of movies have a duration between 90 to 120 minutes.
- The Majority of TV shows consisting of single 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.



EDA(Netflix country wise):

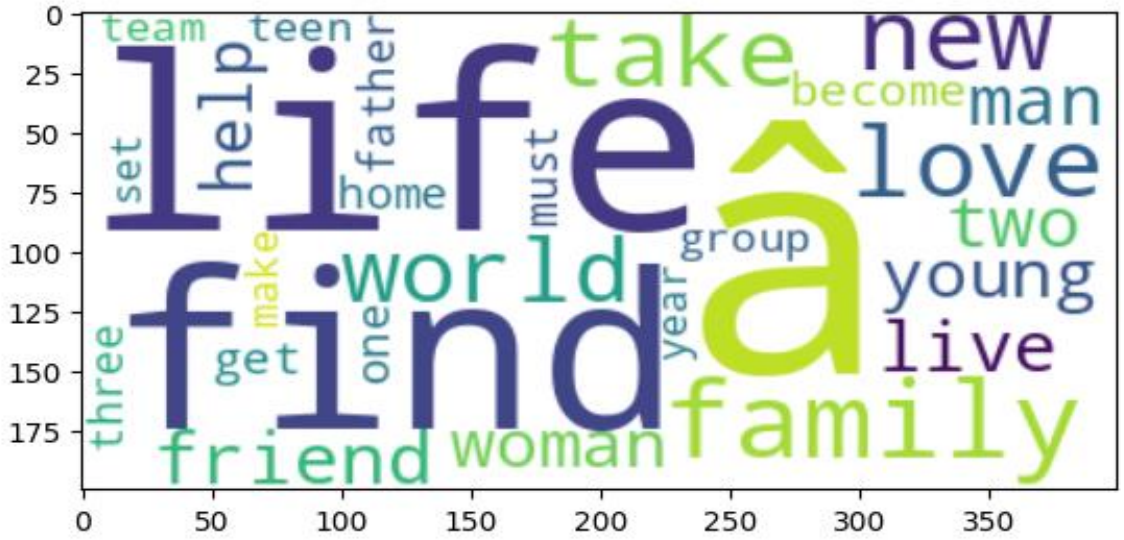
- The United States-based movies and TV shows were produced most, followed by India and the United Kingdom.
- In India and United State, a greater number of movies are present compared to TV shows.
- In the UK, Japan, and South Korea there are a greater number of TV shows than movies
- India and united State have produced very less TV show compare to moves so there have a big gap in TV show production. it gives a huge opportunity to TV show maker to fulfill the gap and make a big profit in the business purpose.



EDA(Most occurred word in Description):

Here we can see that the most of the movie and TV show description contain "family," "find", "life", “Love", "new world" and "friend" word.

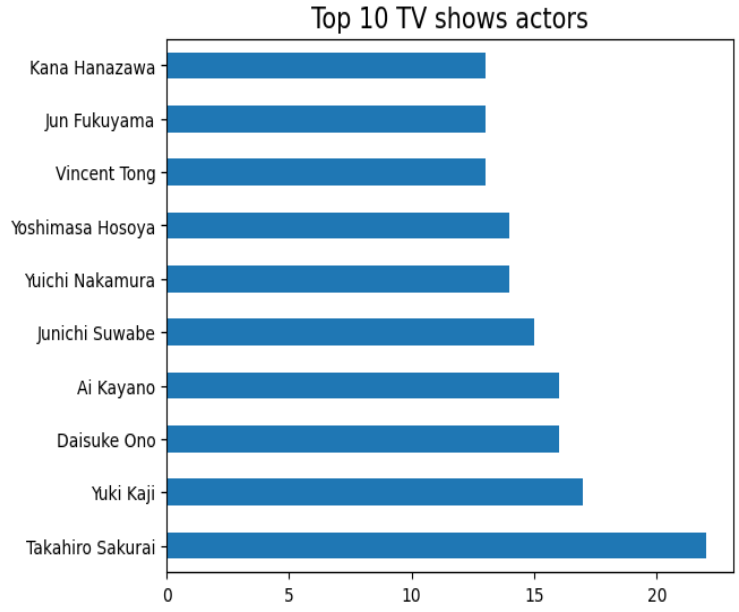
Its attract more people to watch the movie or TV show.

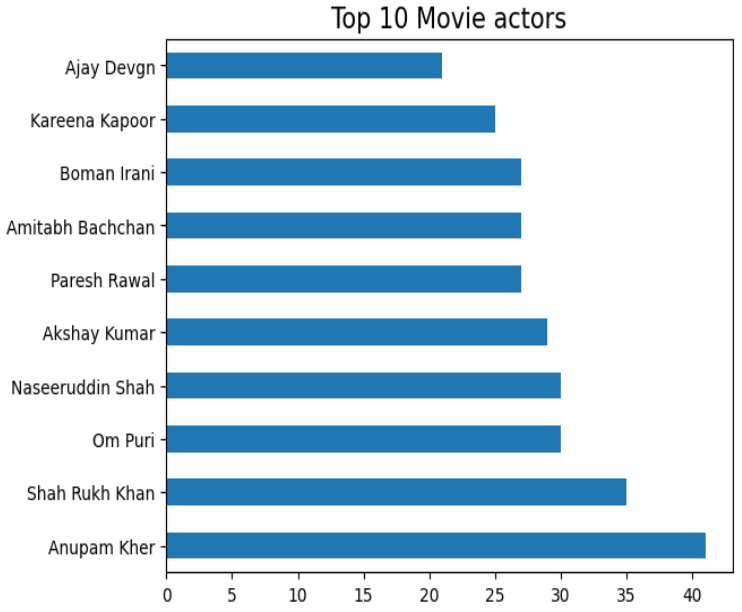


EDA(Netflix cast):

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

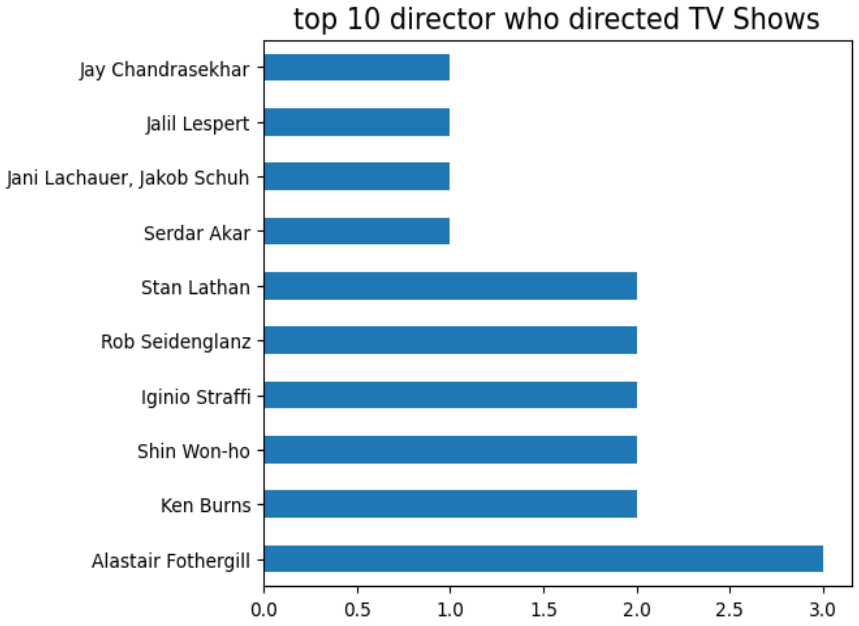
- In the shows, Takahiro Sakurai, Yuki Kaji, and Daisuke Ono played the most number of roles.

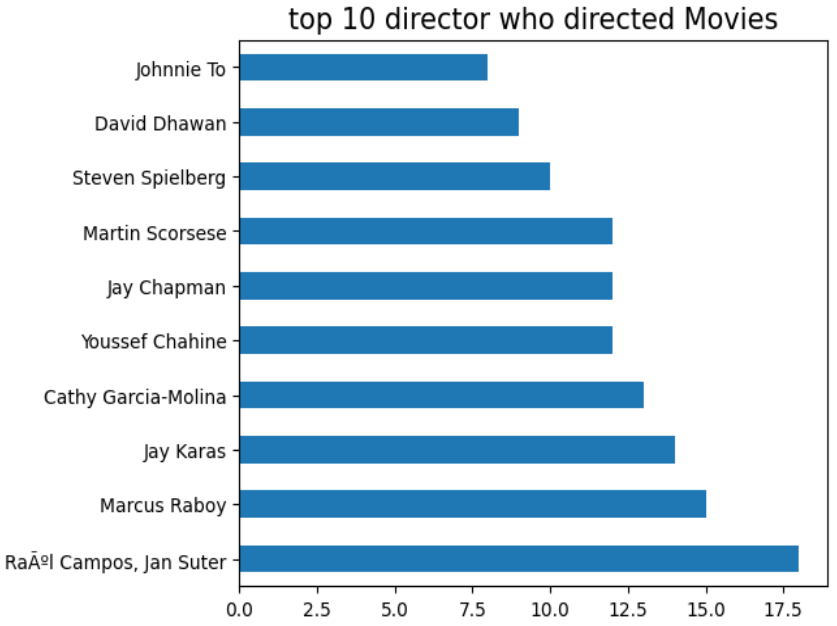




EDA(Netflix Director):

- The three shows directed by Alastair Fothergill are the highest on the data list.
- Both, Jansuter and Raul Campos have directed 18 films, more than anyone else in the dataset.





Hypothesis Testing :

- Hypothesis testing in statistics refers to analyzing an assumption about a population parameter.

HO : movies rated for kids and older kids are at least two hours long
H1:movies rated for kids and older kids are not at least two hours long.

	Target-ages	duration
0	Kids	66.486891
1	Older-kids	92.024648
2	Teens	107.772021
3	Adults	98.230769

Mean for movies rated for Kids duration 66.486891
Mean for movies rated for older kids duration 92.024648
Standard deviation for movies rated for Older Kids duration 31.182577
Standard deviation for movies rated for kids duration 31.739465

- To perform this method, we first formulate the Null and Alternate Hypotheses.
- The P-value method is used in Hypothesis Testing to check the significance of the given Null Hypothesis. Then, deciding to reject or support it is based upon the specified significance level or threshold.
- **the t-value is not in the range, the null hypothesis is rejected.**
- **As a result, movies rated for kids and older kids are not at least two hours long.**

Hypothesis Testing :

2. H1:The duration which is more than 90mins are movies
HO : The duration which is more than 90mins are NOT movies

	type	duration
0	Movie	99.307978
1	TV Show	1.760833

	Target-ages	duration
0	Kids	66.486891
1	Older Kids	92.024648
2	Teens	107.772021
3	Adults	98.230769

Mean for movies rated for Kids duration 99.307978
Mean for movies rated for older kids duration 1.760833
Standard deviation for movies rated for Older Kids duration 1.560603
Standard deviation for movies rated for kids duration 28.530881

- Because the t-value is not in the range, the null hypothesis is rejected.
- As a result, The duration which is more than 90mins are movies

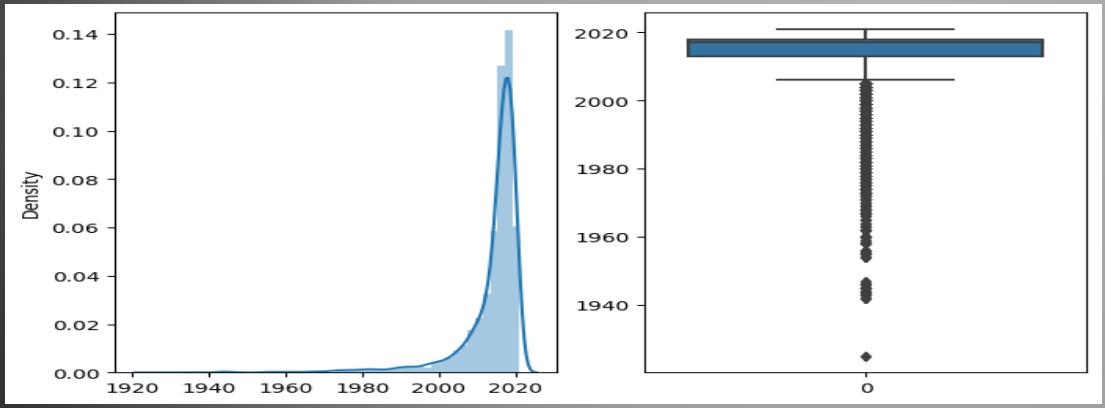
Feature Engineering:

Handling Missing Values –

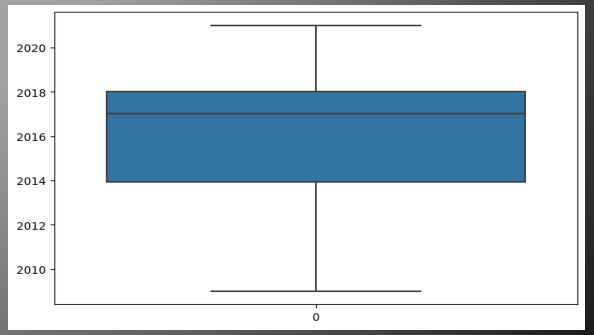
- we have all ready clear missing values or handling NAN values that why we can see here no missing values , if we drop these nan values it will not affect that much while building the model.

Handling Outliers–

- Since, the some of the data present in textual format except release year. and duration is all data clear shows and some unwanted values show.
- The data that we need to create cluster/building model are present in textual format. So, there is no need to perform handling outlier.



Release year outliers box plot



Modeling Approach:

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

Cluster :

We create one cluster column based on the following features:

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

Feature Engineering:

Before clusters implementation we need to pre-process the data. So that we filtered data with following steps :

1.Removing Stop words :

- Stop words are common words like “the”, “and” and “but” do not carry much meaning on their own and are often seen as noise in the data.

2. Removing Punctuation :

- Punctuation marks like periods, commas, and exclamation points can add noise to the data and can sometimes be treated as separate tokens, which can affect the performance of NLP models.

3. Stemming :

- used Snowball Stemmer to generate a meaningful word out of corpus of words.
- For example, the words "run," "runs," "ran," and "running" are all different inflected forms of the same word "run," and a stemmer can reduce them all to the base form "run."

Feature Engineering:

4. Lowercasing words :

• Lowercasing the words can also reduce the size of the vocabulary, which can make it easier to work with larger texts or texts in languages with a high number of inflected forms.

5. Tokenization of corpus and Word vector – TFIDF :

1. This is important in NLP tasks because most machine learning models expect numerical input and cannot work with raw text data directly. Word vector allows you to input the words into a machine learning model in a way that preserves the meaning and context of the words.

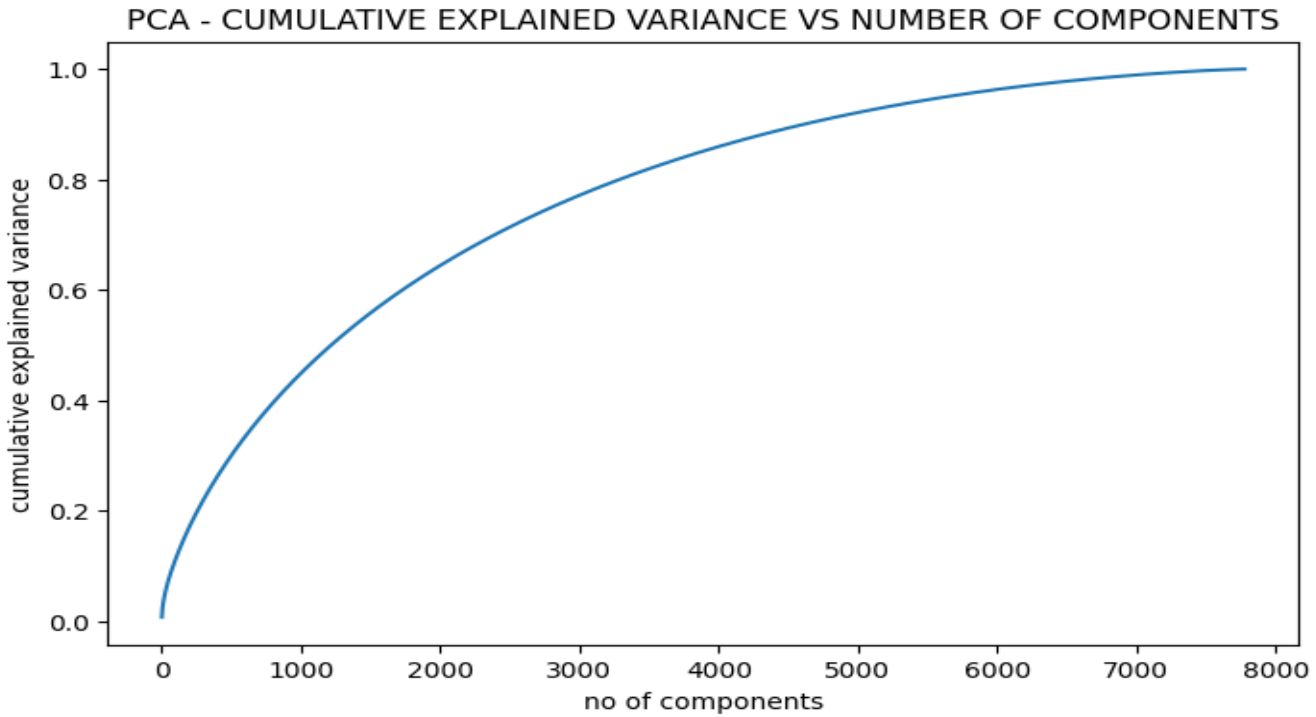
6. Lemmatization :

Lemmatize verbs in list of tokenized words.

7. Dimensionality reduction – PCA :

• Dimensionality reduction is the process of reducing the number of features or dimensions in a dataset while preserving as much information as possible. As high-dimensional datasets can be difficult to work with and can sometimes suffer from the curse of dimensionality

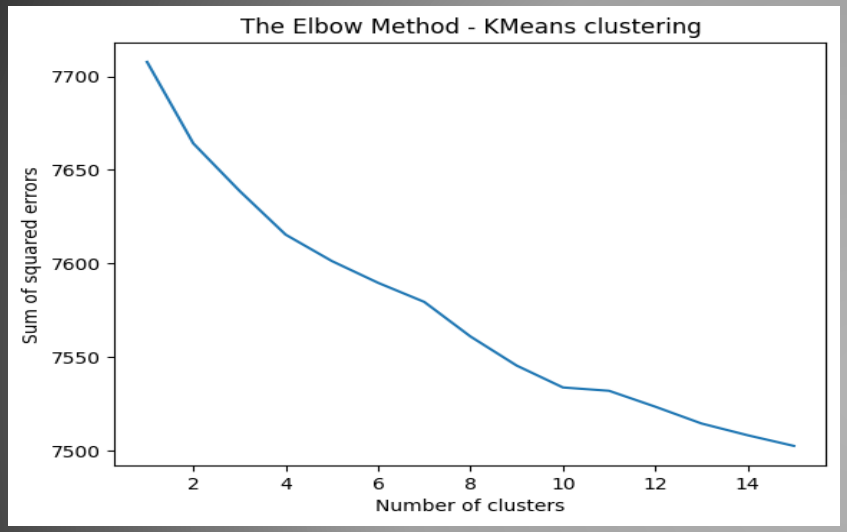
Principle Component Analysis:

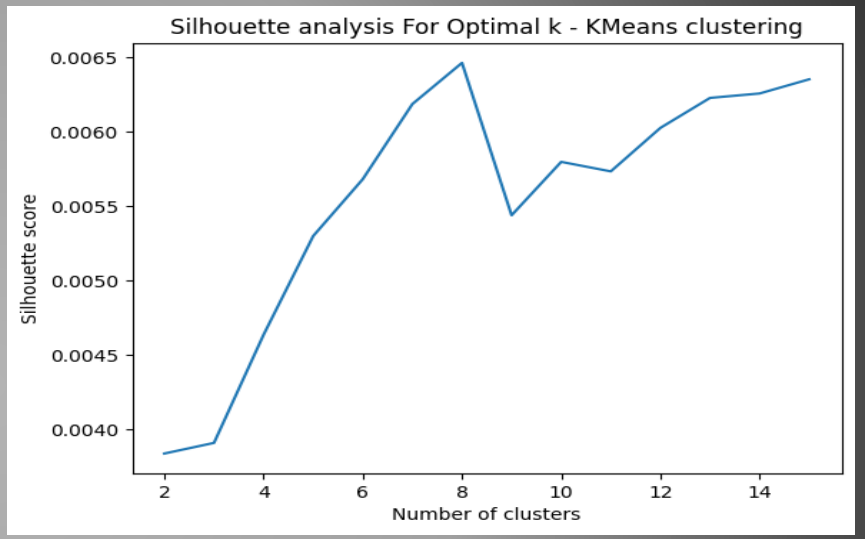


- We find that 100% of the variance is explained by about ~7500 components.
- Also, more than 80% of the variance is explained just by 4000 components.
- Hence to simplify the model, and reduce dimensionality, we can take the top 4000 components, which will still be able to capture more than 80% of variance.

Clusters Model Implementation

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

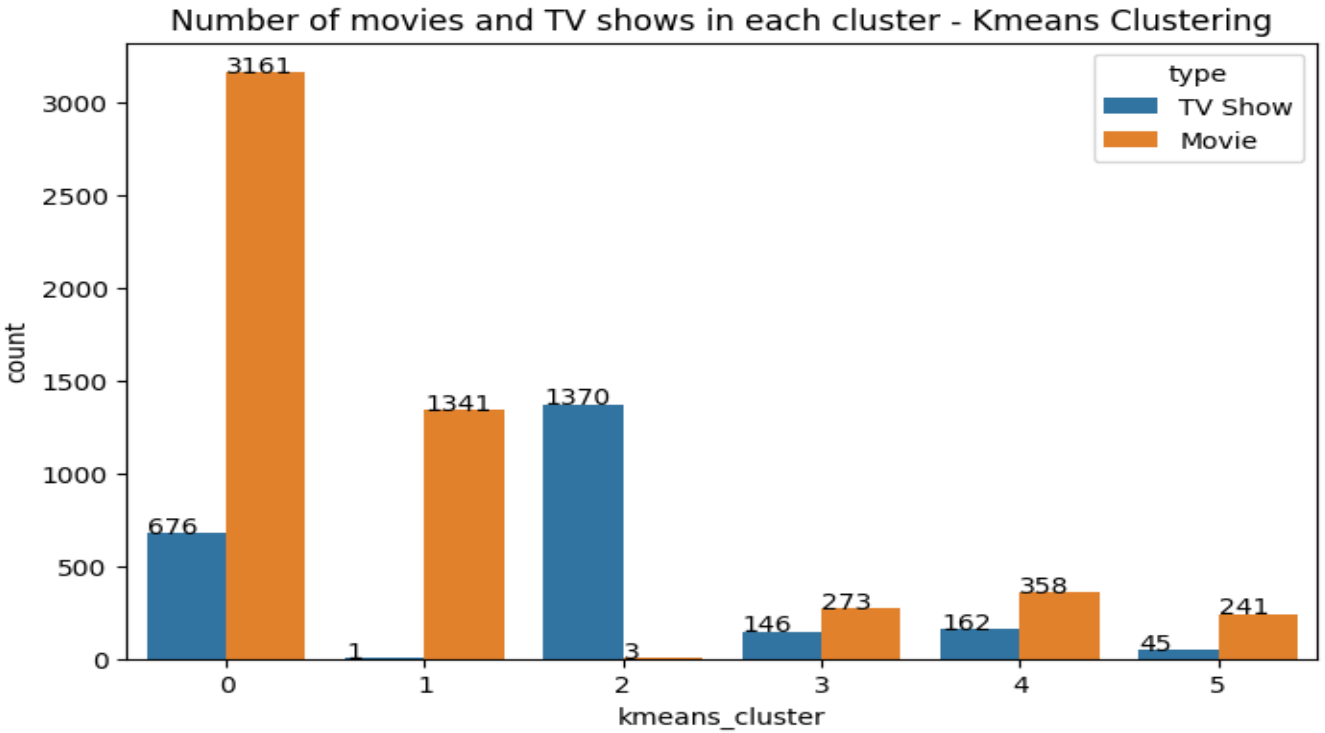




- The sum of squared distance between each point and the centroid in a cluster decreases with the increase in the number of clusters.

- The highest Silhouette score is obtained for 5 clusters.
- Building 5 clusters using the k-means clustering algorithm

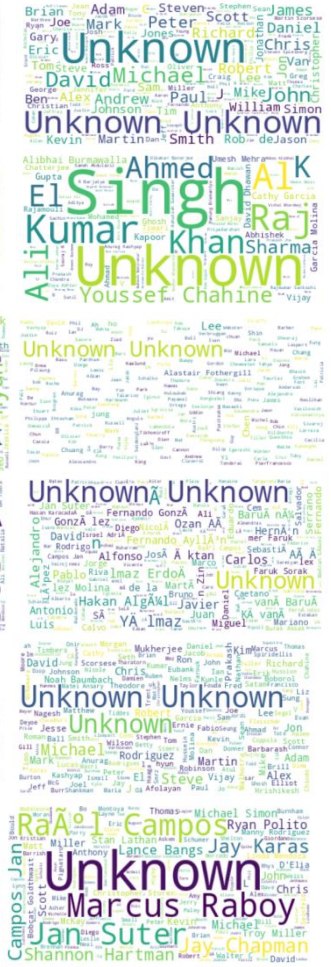
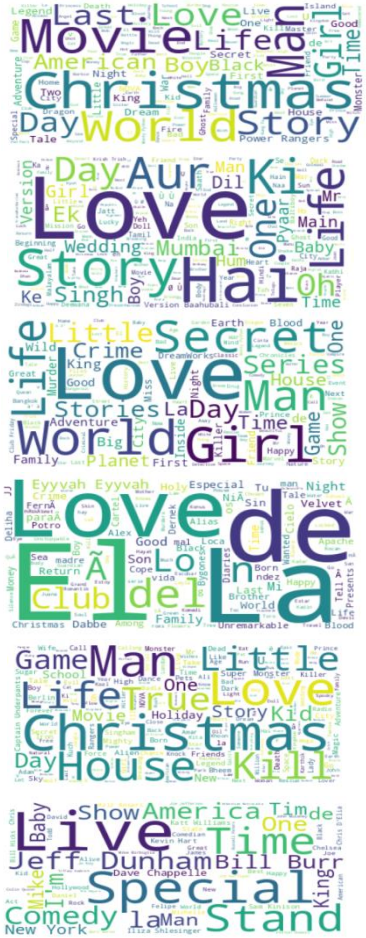
K-Means Clusters :



- Successfully built 5 clusters using the k-means clustering algorithm.
- In cluster 0, 1 & 4 highest number of count belong from Movie class.
- Cluster 2, 0 build on TV shows.

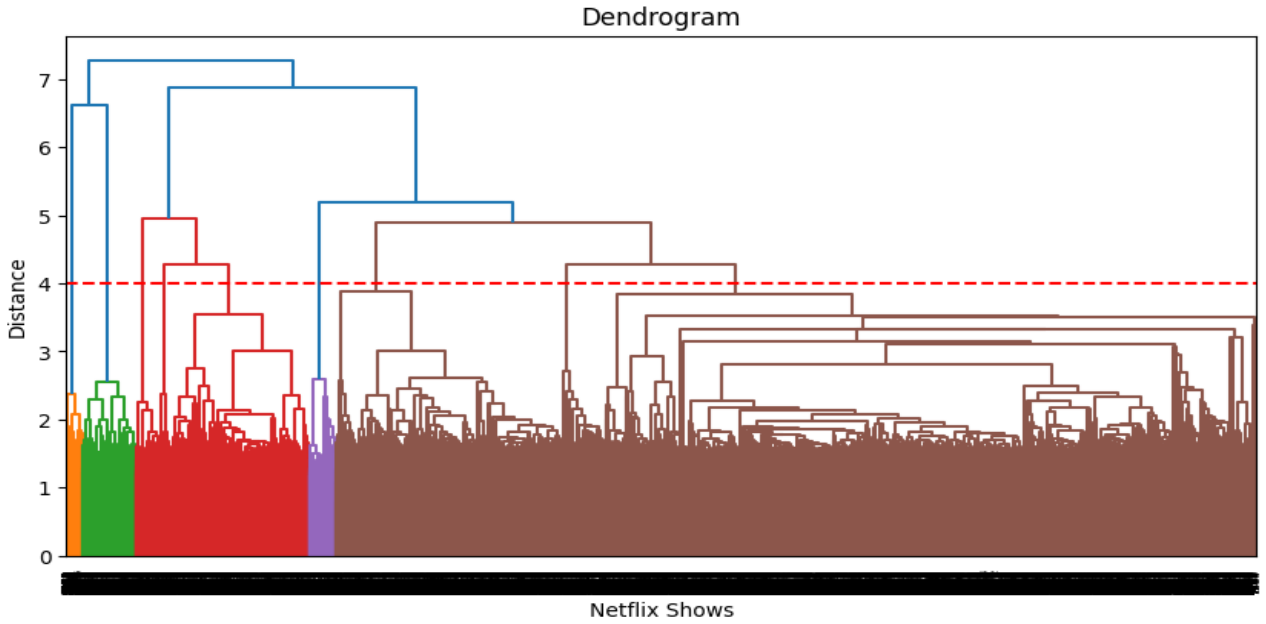
Word Cloud :

- In this here Implement and showing only 4 different word cloud cluster

Word Cloud on "description"	Word Cloud on "cast"	Word Cloud on “Director"	Word Cloud on listed-in	Word Cloud on "title"
				

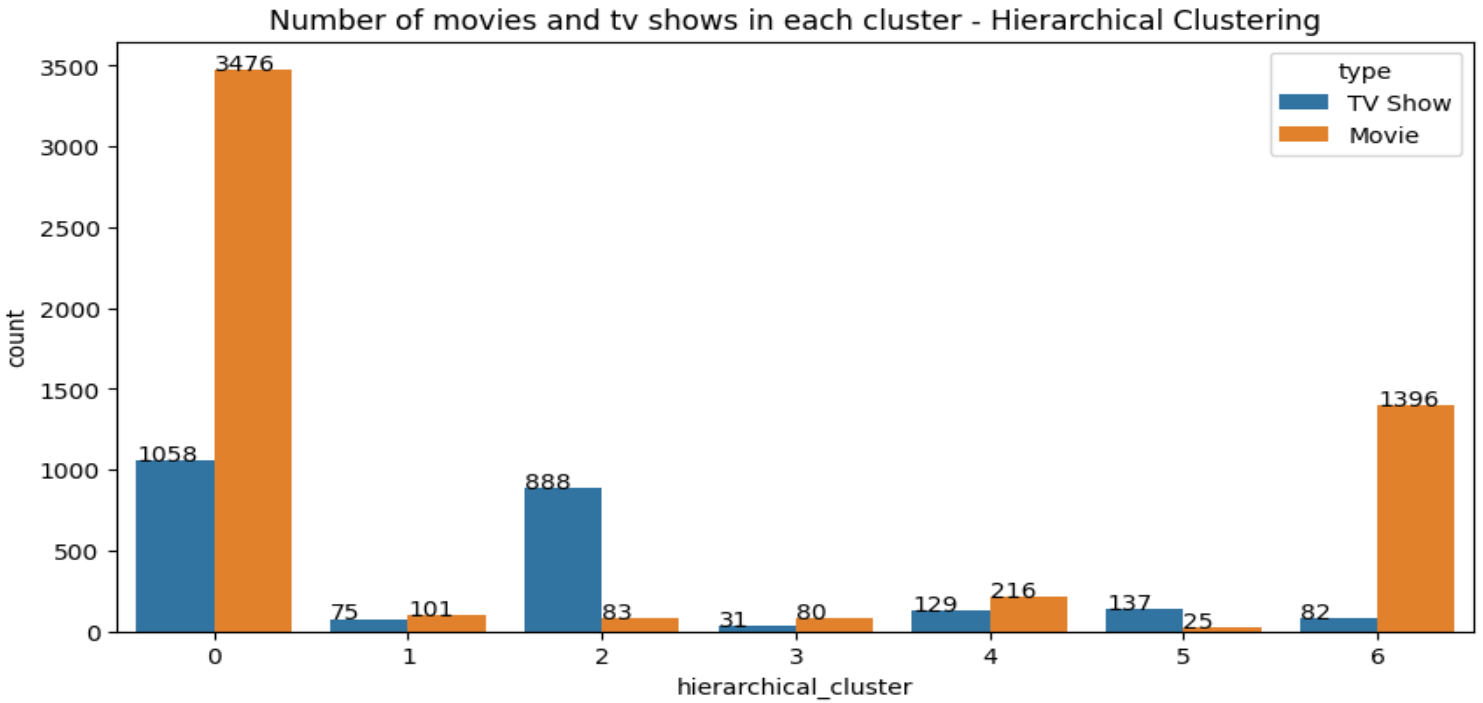
Hierarchical clustering:

- Building clusters using the **Agglomerative (hierarchical) clustering algorithm.**
  - Agglomerative hierarchical clustering is a method of clustering that is used to build a hierarchy of clusters. It is a bottom-up approach, where each sample is initially treated as a single-sample cluster and clusters are merged together as they are deemed similar.



- Visualizing the Dendorogram to decide on the optimal number of clusters for the agglomerative (hierarchical) clustering algorithm.
- At a distance of 4 units, 7 clusters can be built using the agglomerative clustering algorithm

Hierarchical clusters:



- Agglomerative Hierarchical clusters:

• Successfully built 7 clusters using the Agglomerative (hierarchical) clustering algorithm.

• Highest number of data point build on cluster 0

Word Cloud

- 5 Word cloud use

Word cloud ‘tittle’	Word cloud ‘description’	Word cloud “cast”	Word Cloud on "country“	Word Cloud on "listed-in (genre)"
				

Content Based Recommendation System :

- A content-based recommendation system is a type of recommendation system that suggests items to users based on their similarity to other items that the user has shown interest in. It uses the attributes or features of the items to determine the similarity between them.
- Based on how similar the movies and shows are, we can create a straightforward content-based recommender system. The recommender system needs to be able to suggest a list of similar shows that a person who has watched a show on Netflix likes. We can use cosine similarity to determine the shows' similarity scores. By dividing the dot product of the two vectors by their magnitude values, the similarity between A and B can be calculated. Simply put, the angle between two vectors decreases as the cosine similarity score increases.
- .

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

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

Golmaal Returns
Hattrick
Phir Hera Pheri
Maine Pyaar Kyun Kiya
Ishqiya
Singham
Jaoon Kahan Bata Ae Dil
Haseena Maan Jaayegi
Thank You
Tum Milo Toh Sahi
```

```
recommend('Lucifer')
```

```
If you liked 'Lucifer', you may also enjoy:

Beauty & the Beast
Rica, Famosa, Latina
On My Block
The Good Cop
Carmel: Who Killed Maria Marta?
Monty Python's Fliegender Zirkus
Cold Case Files
L.A.â□□s Finest
Hinterland
Undercover
If you liked 'Lucifer', you may also enjoy:

Beauty & the Beast
Rica, Famosa, Latina
On My Block
The Good Cop
Carmel: Who Killed Maria Marta?
Monty Python's Fliegender Zirkus
Cold Case Files
L.A.â□□s Finest
Hinterland
Undercover
```

CONCLUSION:

- 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.
- 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.
- 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.
- 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