Capstone Project

Netflix Movies and TV Shows Clustering.

Submitted By

Aadarsh Pandey

Data Science Trainee, AlmaBetter

Agenda

1. Problem Overview
2. Dataset Overview
3. Data Analysis
4. Data Preparation
5. Proposed Models
6. Conclusion

Problem Overview

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 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.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

Dataset Overview

As the information displayed above we can see that there is only one column with integer datatype, but the important thing is we have got some null values in the dataset. Also we have to check for the duplicate values.

We have got **Directors** with **2389 null values** in it then **cast** with **718 null values** and **country** with **507 null values** in it, also, we have **date_added** and **rating** with **10** and **7 null values** in it.

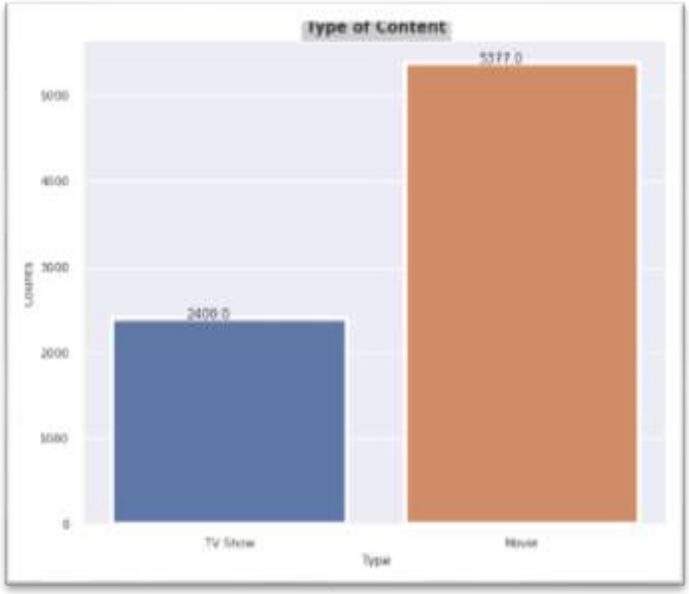
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

```
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
dtype: int64
```

Data Analysis.

According to above graph, content provided by Netflix is more Movies compare to TV Show.

Here we can see that there are total around 5400 Movies on Netflix and 2400 TV Shows on Netflix.



Data Analysis.

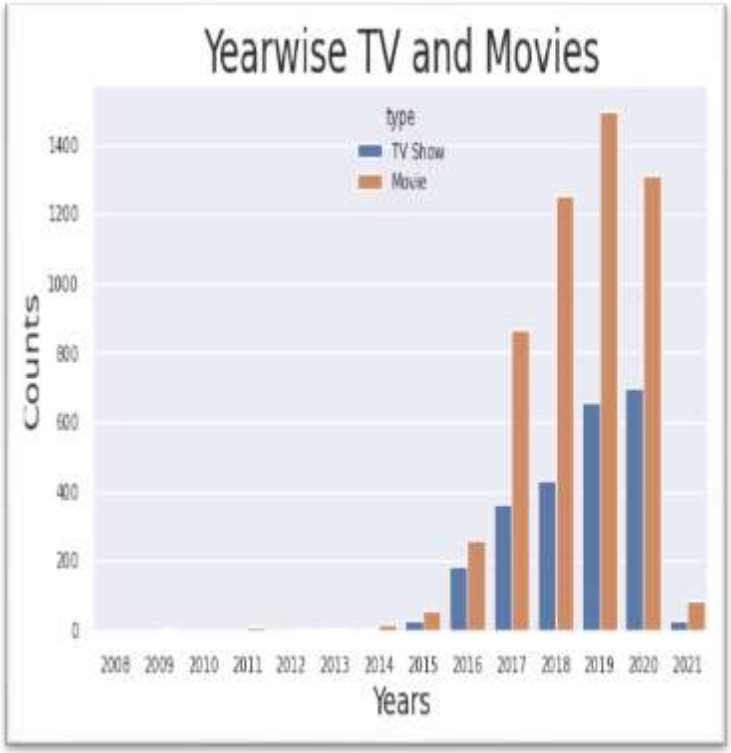
This graph is about year wise TV and Movies i.e.,

TV Shows are less compare to Movies and in every year.

By this we can also say that count of TV Shows and Movies has increased but they certainly decreased in 2021.

2019 was the year which has the highest number of Movies.

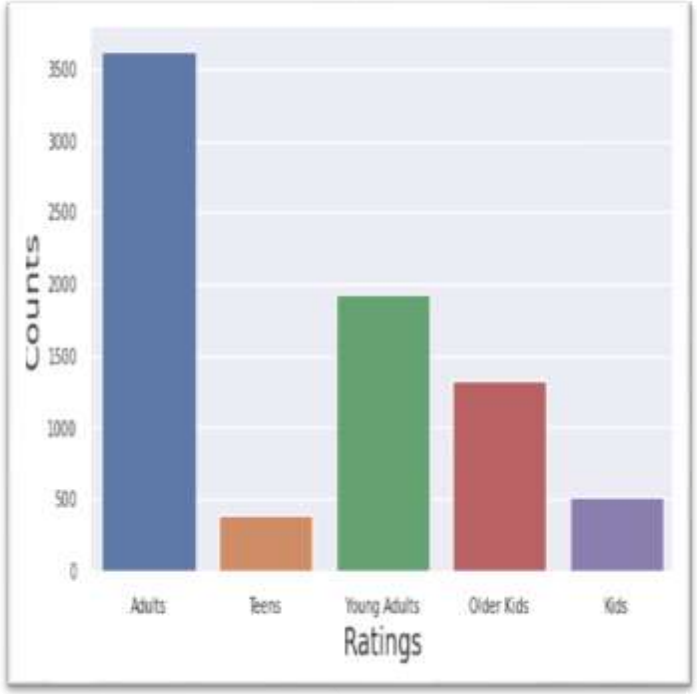
2020 was the year which has the highest number of TV Shows.



Data Analysis.

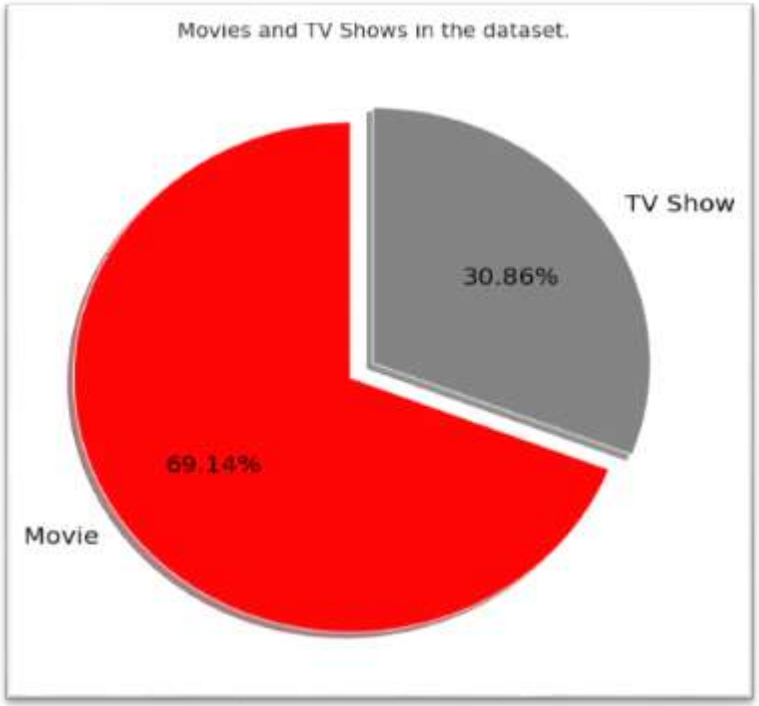
This chart is about the what is the categories of the shows in which they fall in.

Highest number of shows produced is for Adults followed by Young Adults and then Older Kids and Kids. The lowest number of shows are produced for Teens.



Data Analysis.

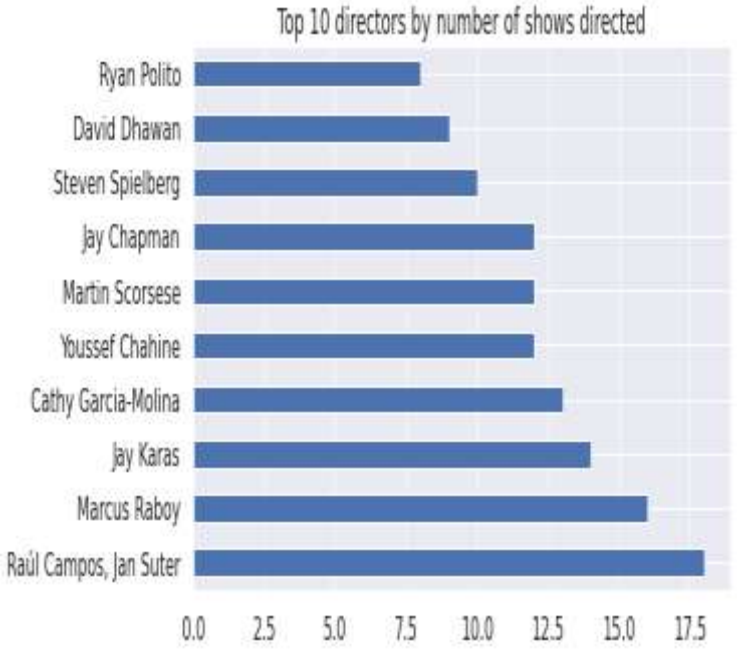
According to this Pie Chart we get that there are high number of Movies(69.14%) as compared to TV Shows(30.86%).



Data Analysis.

This chart is about top 10 directors having their highest number of shows on Netflix.

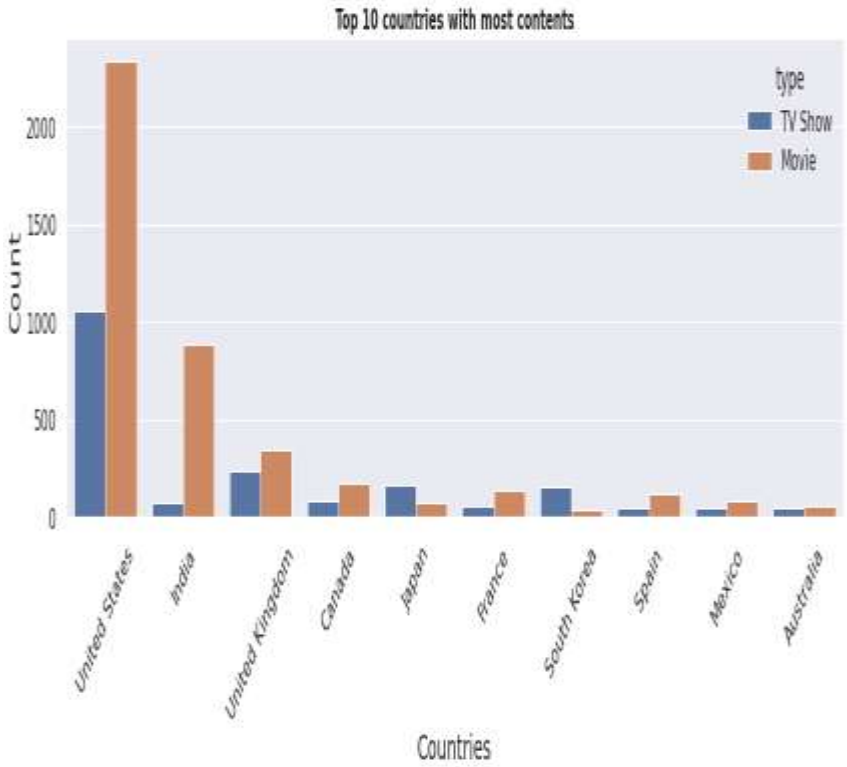
Raul Campos and Jan Suter are the 2 directors who have directed around 18 movies and they have directed highest movies compare to any other directors.



Data Analysis.

This Graph is about top 10 countries having the highest number of shows on Netflix platform.

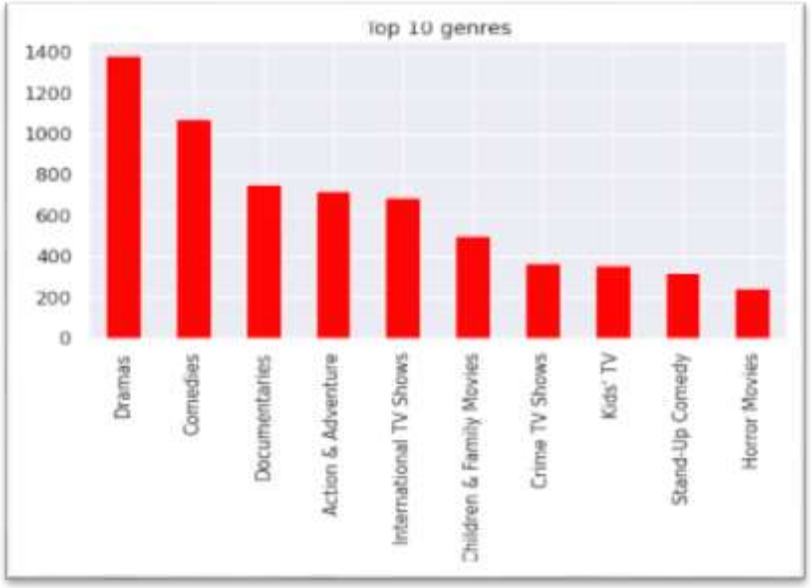
United States is on the top of all the countries followed by India on 2nd highest.



Data Analysis.

This Graph tells us about top 10 genres available on Netflix.

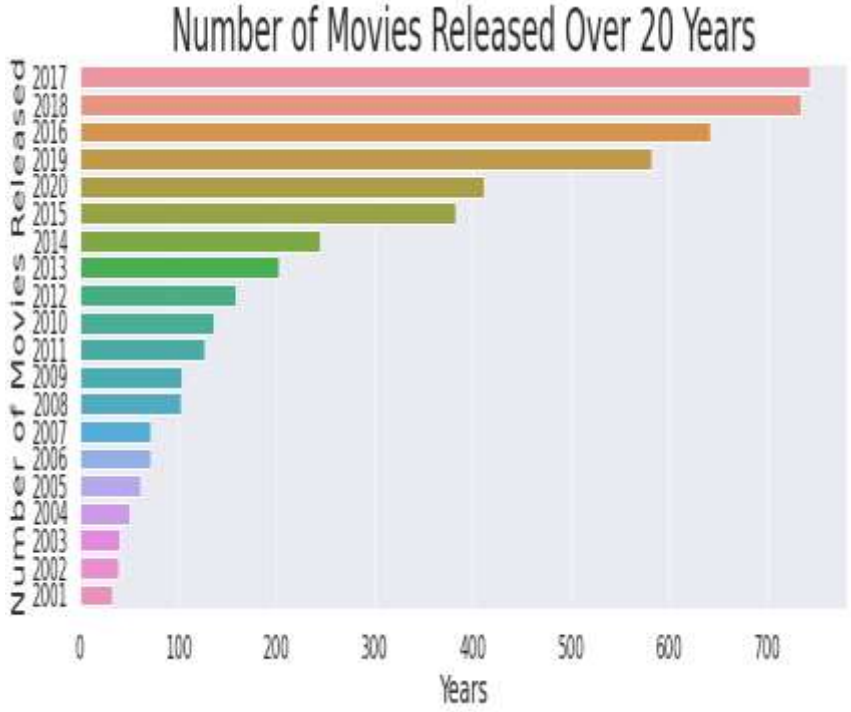
Dramas is the genre having the highest content on Netflix and then we have Comedies with 2nd highest genre.



Data Analysis.

This chart is about checking in which year the highest number of Movies was released.

On 2017 and 2018 there was around more than 700 Movies released.



Data Analysis.

This Chart represents Number of Movies and TV Shows added over the years.

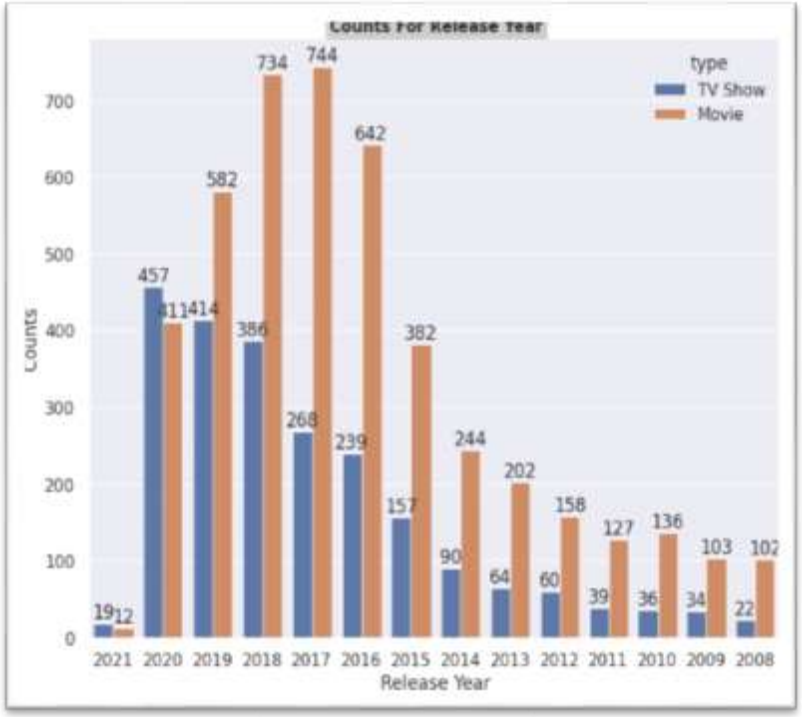
And in 2019 there was around 1500 Movies added and in 2020 around 700 TV Shows was added which is highest in overall the years.

Data Analysis.

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.

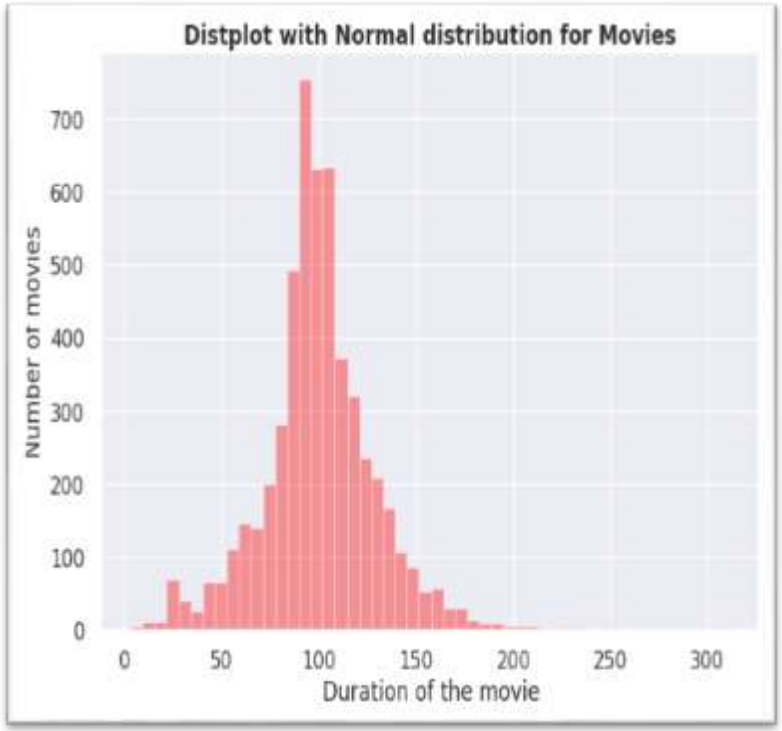
This might signal that Netflix is increasingly concentrating on introducing more TV series to its platform rather than movies.



Data Analysis.

This is a Distribution plot to find the average duration of Movies.

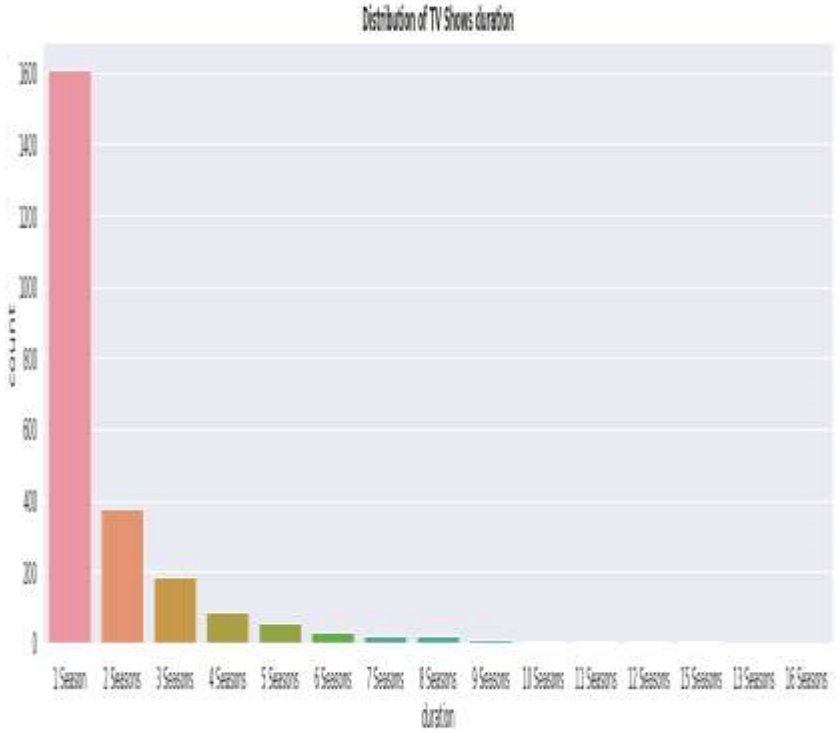
With this we can see that most of the Movies are of duration between 50 to 150 minutes.



Data Analysis.

This chart is about the duration of TV Shows available on Netflix.

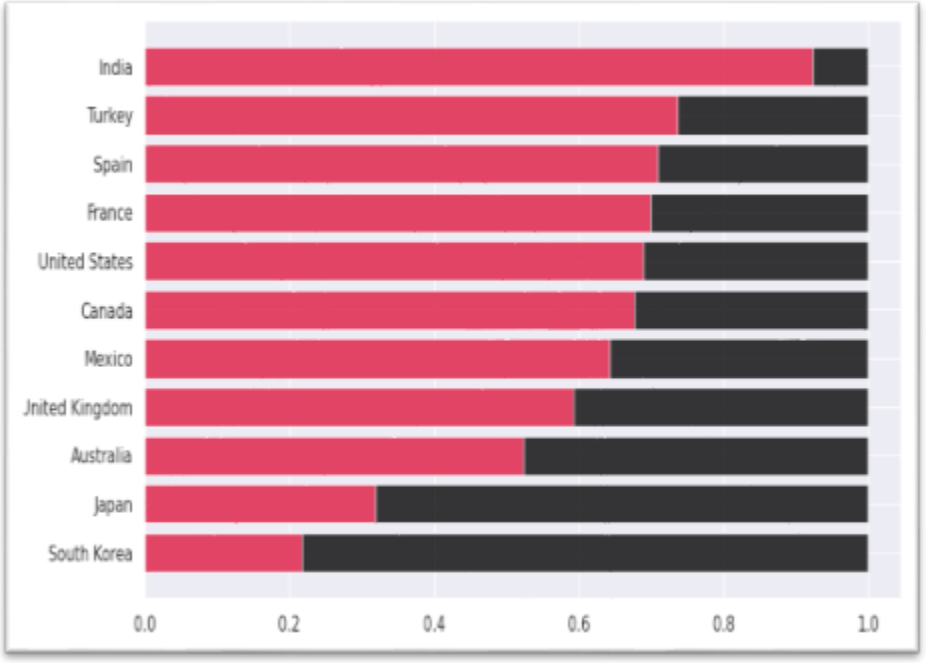
Most of the TV Shows have the duration of 1 Season.



Data Analysis.

This chart represents top countries having highest Number of Movies on Netflix.

According to this chart India has the highest number of Movies on Netflix.



Data Preprocessing.

Text Preprocessing:

Stopwords: Stop words are a set of commonly used words in any language. For example, in English, “the”, “is” and “and”, would easily qualify as stop words. In NLP and text mining applications, stop words are used to eliminate unimportant words, allowing applications to focus on the important words instead.

Lemmatization: Lemmatization is a text normalization technique used in Natural Language Processing (NLP), that switches any kind of a word to its base root mode. Lemmatization is responsible for grouping different inflected forms of words into the root form, having the same meaning.

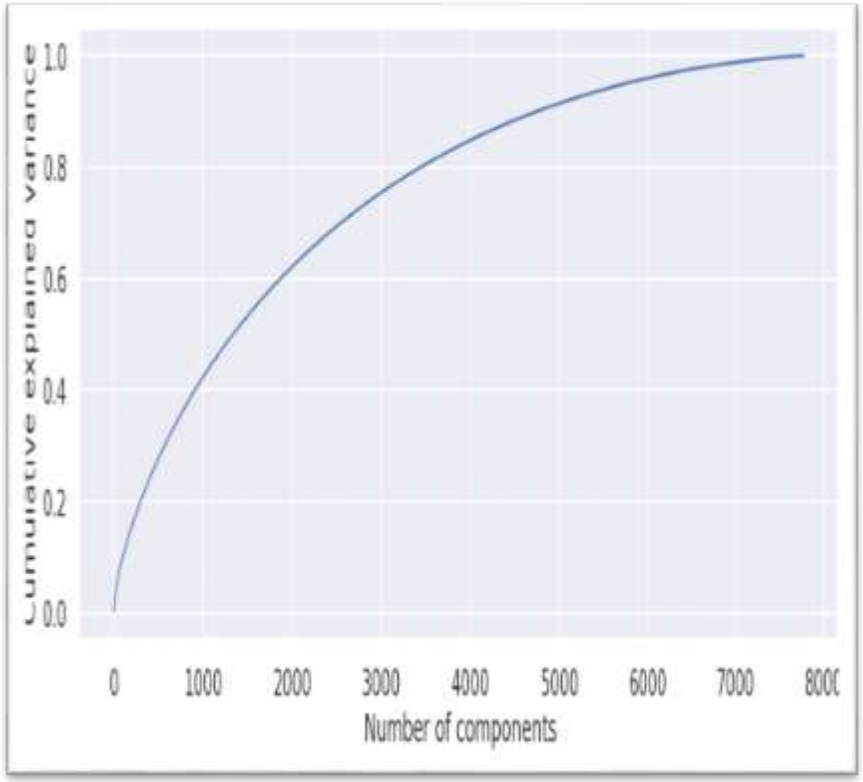
Data Preprocessing.

Word Vectorization:

Word Embeddings or Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics. The process of converting words into numbers are called Vectorization.

Dimensionality Reduction Using PCA.

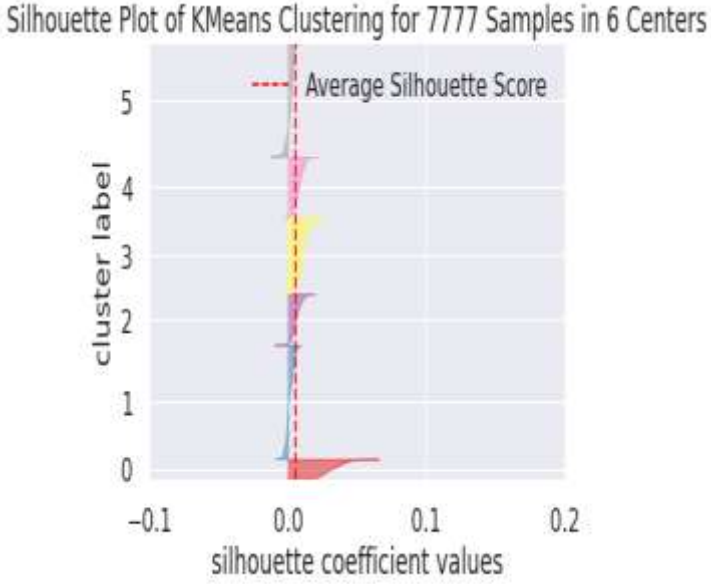
Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.



Proposed Models.

K-means Clustering:

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called flat clustering algorithm. The number of clusters identified from data by algorithm is represented by ‘K’ in K-means.



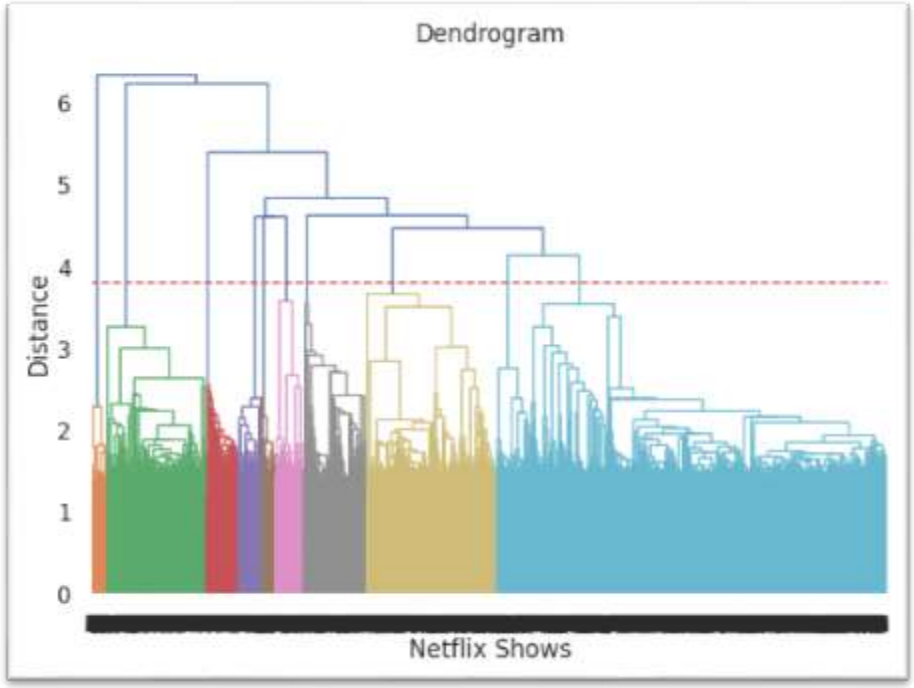
Silhouette score for this dataset is:
0.008198631568232029
Davies Bouldin Score for our cluster is:
9.508420074893262

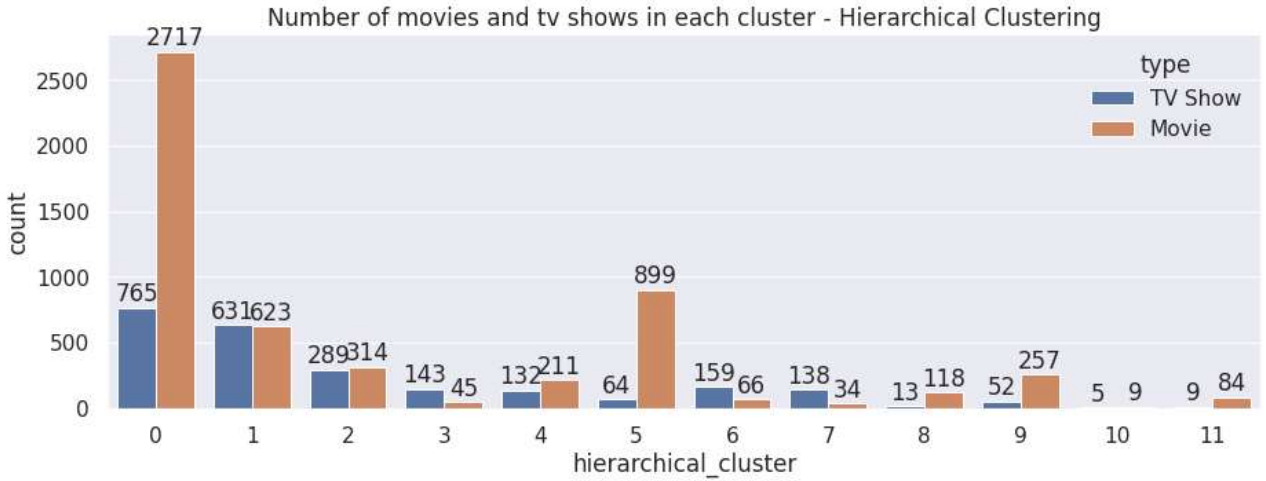
Proposed Models.

Hierarchical Clustering:

Hierarchical clustering is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics.

Agglomerative hierarchical algorithms − In agglomerative hierarchical algorithms, each data point is treated as a single cluster and then successively merge or agglomerate (bottom-up approach) the pairs of clusters. The hierarchy of the clusters is represented as a dendrogram or tree structure.



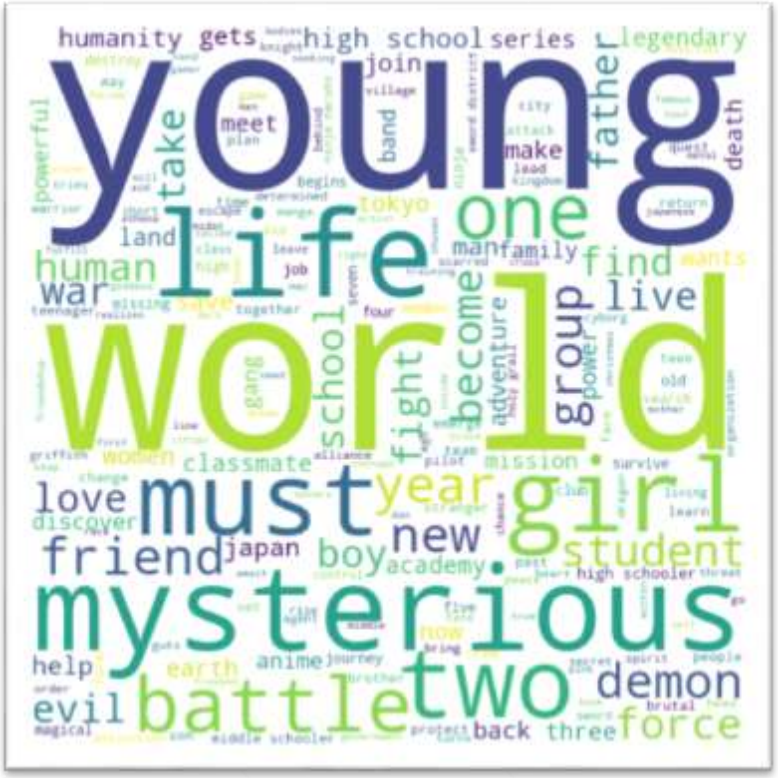


- Out of 12 clusters we have Cluster 0 with the Highest Number of Movies and TV Shows.

Proposed Models.

Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is.



Content Based Recommendation System.

A Content-Based Recommender works by the data that we take from the user, either explicitly (rating) or implicitly (clicking on a link). By the data we create a user profile, which is then used to suggest to the user, as the user provides more input or take more actions on the recommendation, the engine becomes more accurate.

	Recommendations
0	Bad Boys II
1	GoldenEye
2	Tortilla Soup
3	Madam Secretary
4	Operation Odessa
5	Slow West
6	Dollar
7	War on Everyone
8	Act of Valor
9	Tremors 5: Bloodline

Conclusion.

1. Dataset contains total 7787 rows and 12 columns.
2. The dataset contains total 3631 null values.
3. United States contains highest contents on Netflix i.e., 3055 followed by India with 923 contents on Netflix.
4. Netflix contains 5377 Movies and 2400 TV Shows.
5. 2019 was the year which has the highest number of Movies and 2020 was the year which has the highest number of TV Shows.
6. In 2021 the count of TV shows and Movies certainly decreased. Contents available on Netflix is more Adult followed by Young Adults.
7. Netflix contains around 1400 Drama content and around 1100 comedies
8. On Netflix actors which has highest content is Anupam Kher with 38 contents followed by Takahiro Sakurai with 28 contents.
9. Most of the Movies has the duration between 50 to 150 minutes.
10. Most of the TV Shows has the duration on 1 season.
11. India has the highest number of Movies on Netflix.
12. We used Principal Component Analysis (PCA) to handle the curse of dimensionality. 4000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 4000.
13. Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 12. This was obtained after visualizing the dendrogram.
14. A content-based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched.

THANK YOU.