Capstone Project

Netflix Movies And TV Shows Clustering

by

Ritik Vaidande

Computer Vision Practitioner

AlmaBetter, Bengaluru

POINTS FOR DISCUSSION

1. Problem Statement
2. Model Architecture
3. Data Summary
4. EDA
5. Text Preprocessing
6. Model Implementation
7. Model Explainability

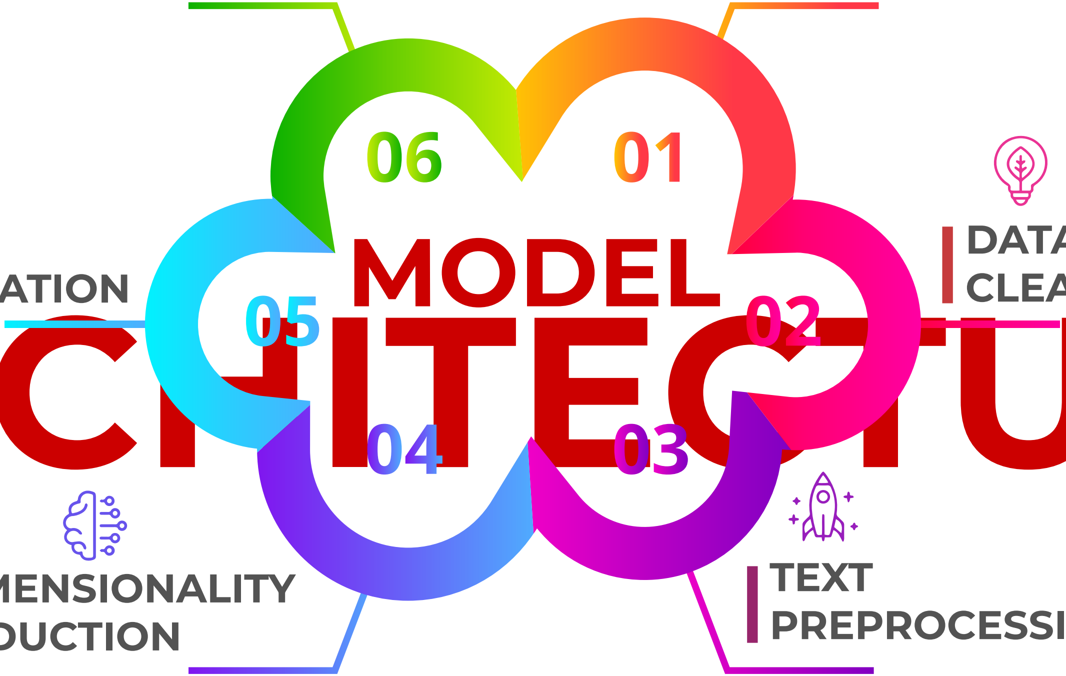
PROBLEM STATEMENT

As movie industry is evolving into streaming platforms, there’s no doubt that Netflix has become one of the important platforms for streaming. Our main objectives of this project is to do exploratory analysis and find useful insights such as what type content is available in different countries, also to find out is Netflix has increasingly focused on TV rather than movies in recent years and at last to do clustering of similar content by matching text-based features from dataset.

NETFLIX

Clustering Similar Content By Matching Text-Based Features

MODEL ARCHITECTURE



01 EDA

02 DATA CLEANING

03 TEXT PREPROCESSING

04 DIMENSIONALITY REDUCTION

05 MODEL IMPLEMENTATION

06 MODEL EXPLAINABILITY

DATASET

Shape - (7787, 12)

Columns Containing Null-Values -

1. director
2. cast
3. country
4. date_added
5. release_year

This dataset consists of tv shows and movies available on Netflix as of 2021.

country
country where the movie / show was produced

show_id
unique ID for every Movie / Tv Show

date_added
date it was added on Netflix

director
director of the Movie

duration
total Duration - in minutes or number of seasons

release_year
actual release year of the movie / show

type
identifier - A Movie or TV Show

title
title of the Movie / Tv Show

cast
actors involved in the movie / show

rating
TV Rating of the movie / show

listed_in
genre

description
the Summary description

CONTENT TYPE DISTRIBUTION

Movie :

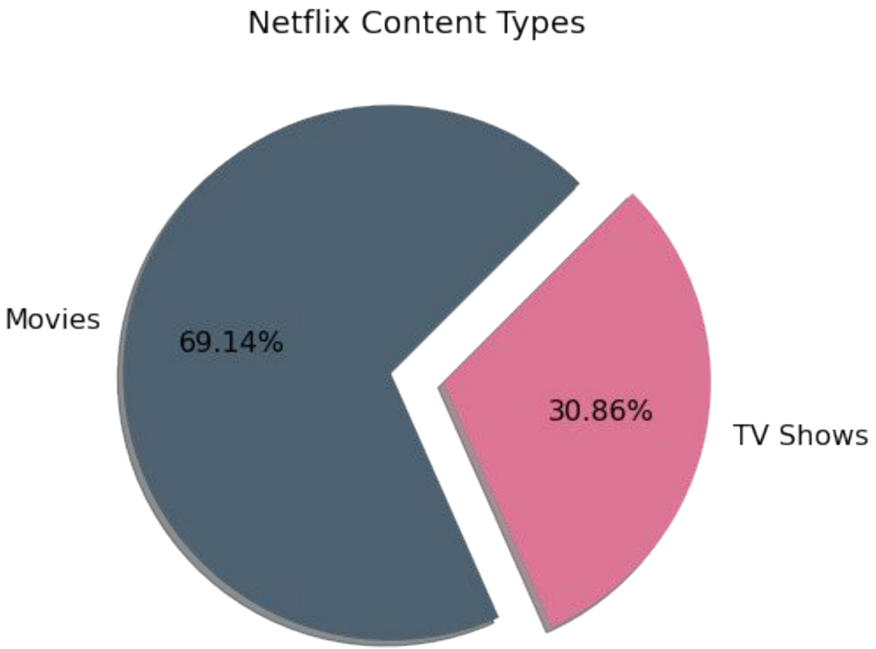
% - 69.14 %

Count - 5372

TV Show :

% - 30.86 %

Count - 2398



Majority of content available on Netflix are Movies.

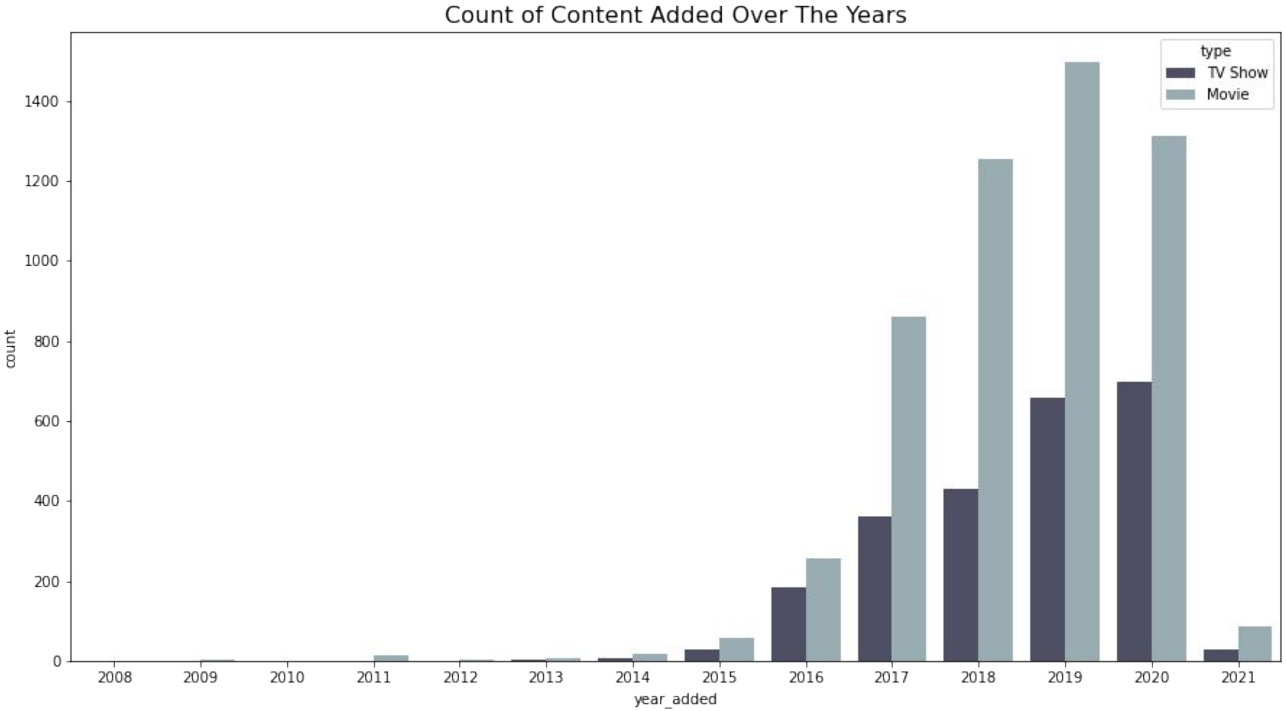
CONTENT ADDED OVER THE YEARS

Years of Data Available :

2008 - 2021

Rise of Content in Year :

2015



Large number of TV Shows and Movie got added in year 2019 and 2020.
Limited amount of data is available for the year 2021

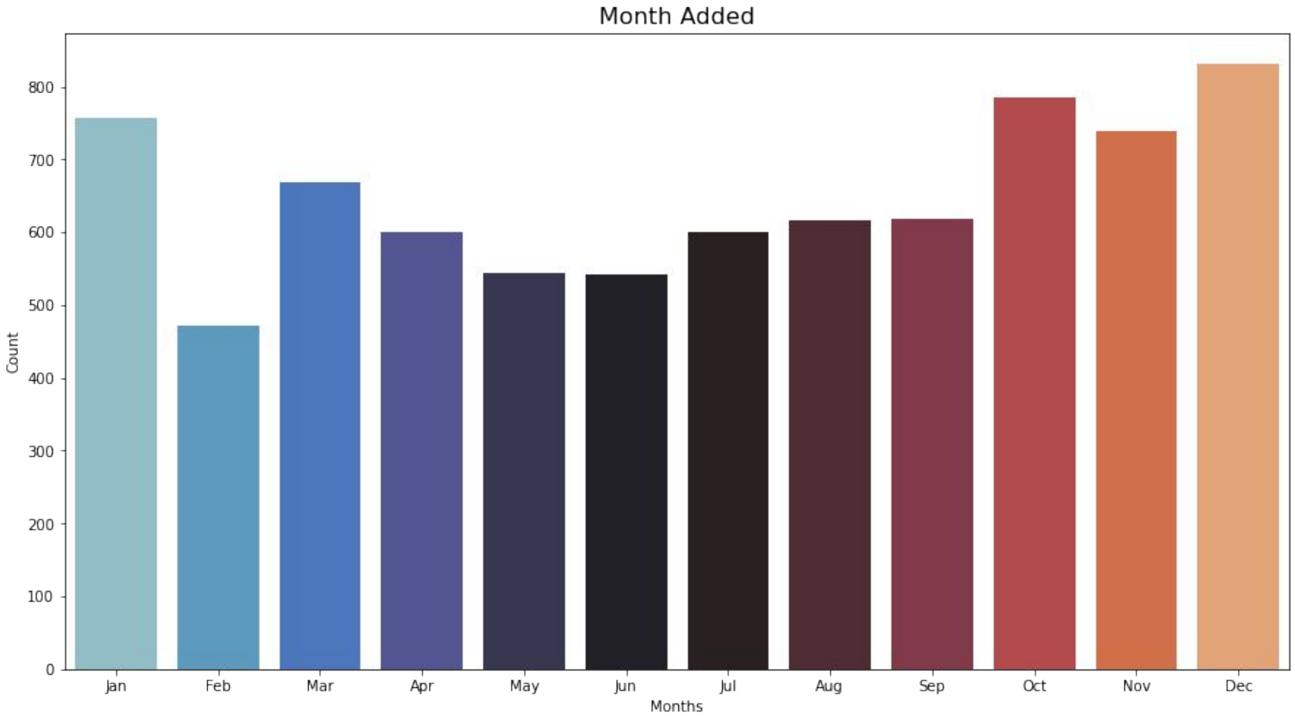
MONTH ADDED

Most Likely Month :

December

Least Likely Month :

February

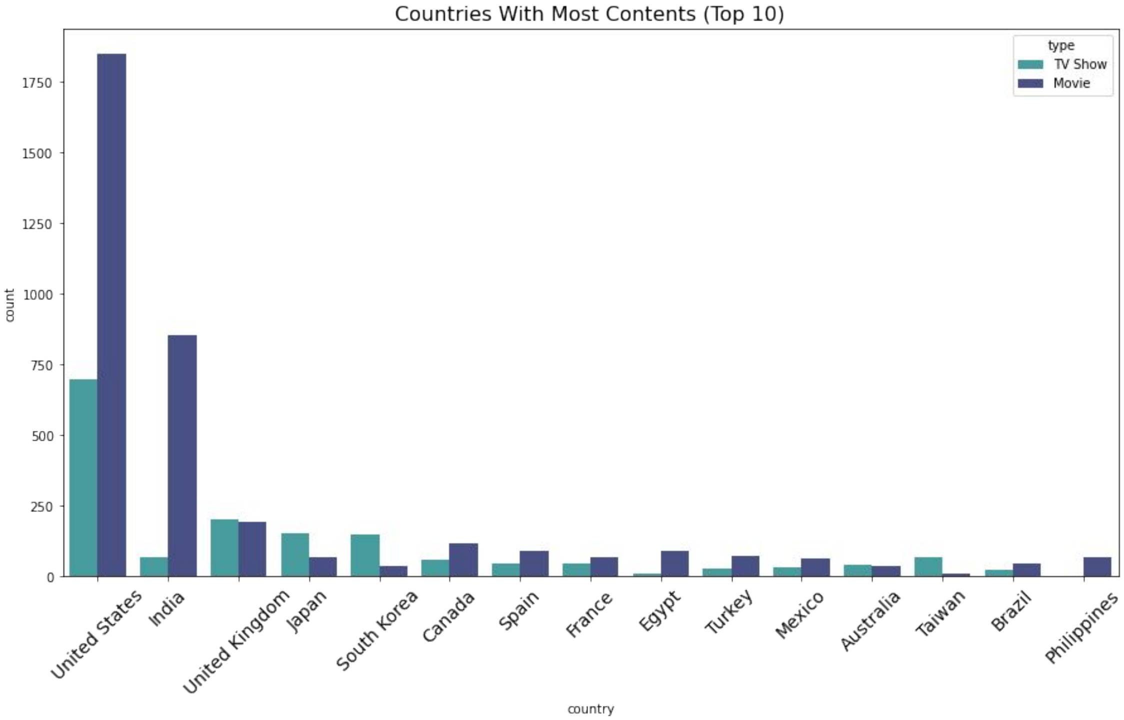


Most of the content is added on the platform during Winter Months.

COUNTRY WITH MOST CONTENTS

Top 5 Countries :

1. US
2. India
3. UK
4. Japan
5. South Korea



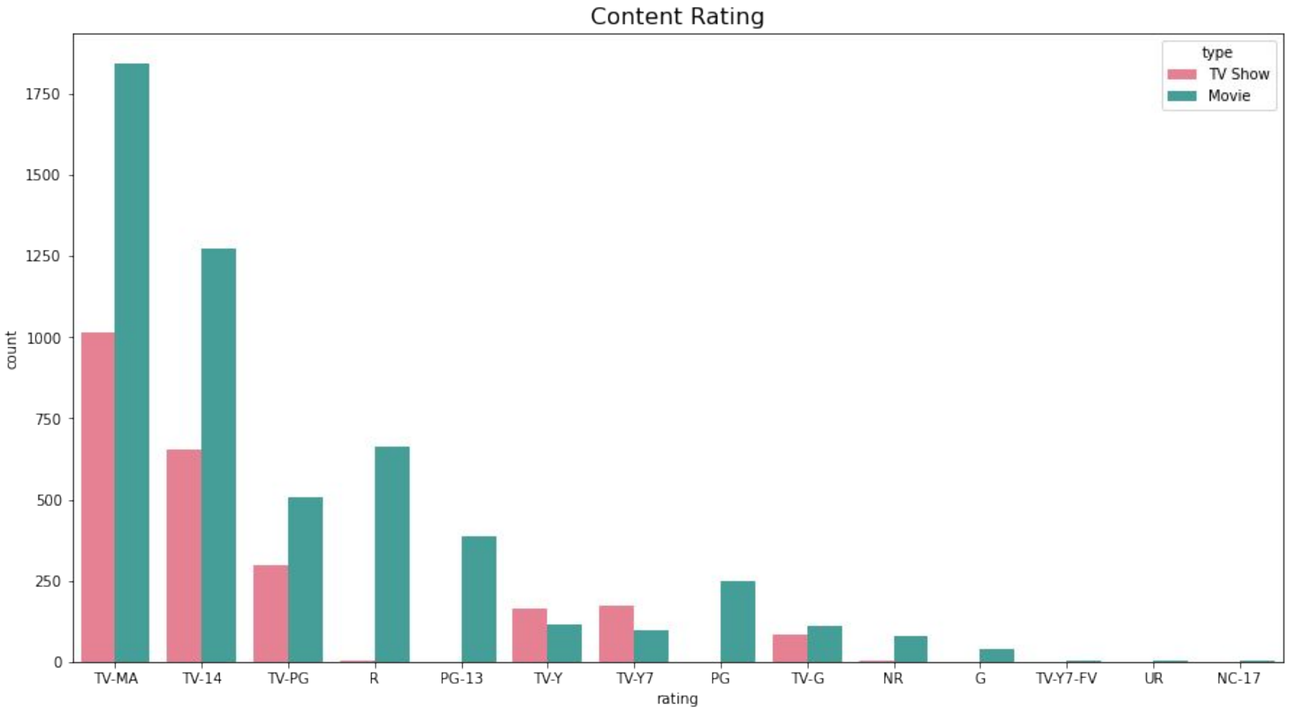
United States produces highest number of TV Shows and Movies.

India ranks 2nd in most content also the Movie counts are very high compared to their TV Shows.

CONTENT RATING

Top 5 Content Rating :

1. **TV-MA** (For Mature Audiences)
2. **TV-14** (Parental Guidance under 14)
3. **TV-PG** (Parental Guidance)
4. **R** (Restricted under 17)
5. **PG-13** (Parental Guidance under 13)

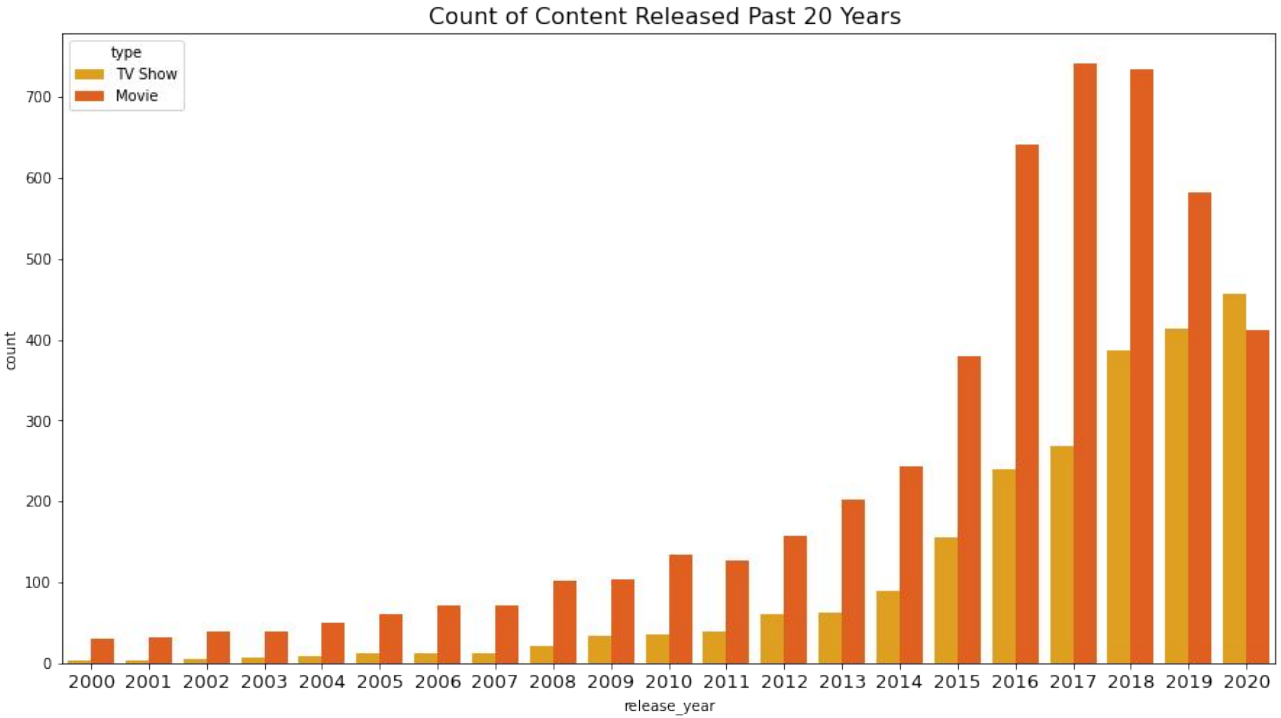


Large number of content are TV-MA rated, this indicates mature content is popular on Netflix.
TV-14 and TV-PG rated content are TV Shows and Movies popular among Teenagers.

CONTENT RELEASED OVER YEARS

Most Number of Content Released in Year :

2018

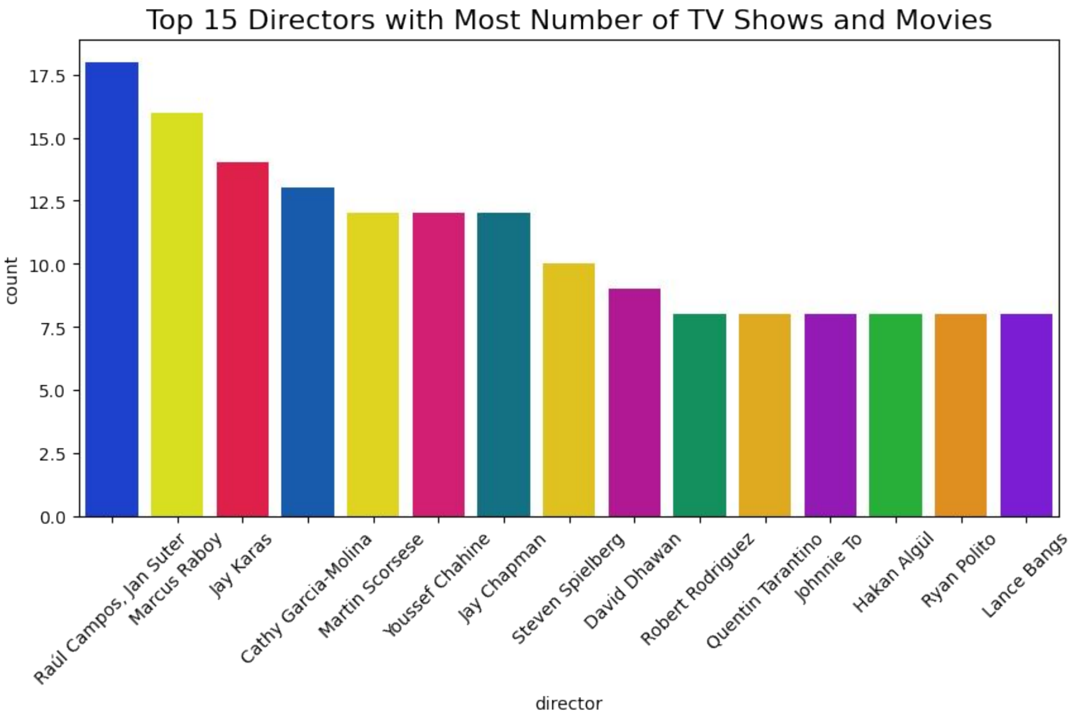


Immense amount of TV Shows and Movies were released over past 5 years.
TV Shows and Movies are following an consistent ratio among them over the years.

DIRECTORS

Director with Most Number of Content :

Raul Campos, Jan Suter (Mexico)



Raul Campos, Jan Suter, Marcus Raboy, Jay Karas, Cathy Garcia-Molina, Martin Scorsese are the top 5 directors with most content on Netflix.

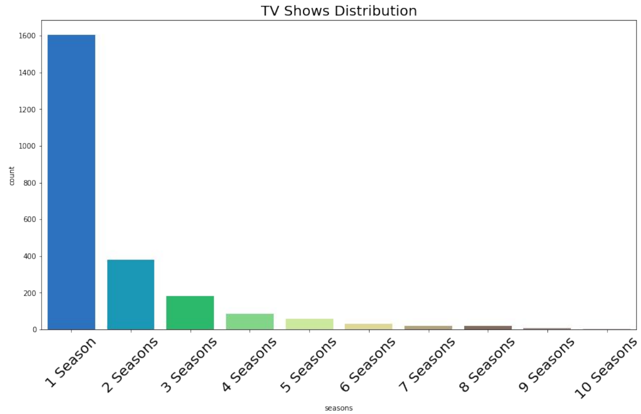
TV Show and Movie Distribution

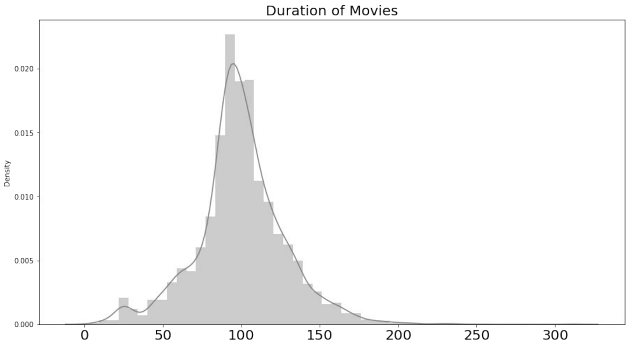
Number of Seasons a TV Show has Most :

1 Season

Average Movie Duration :

90 minutes



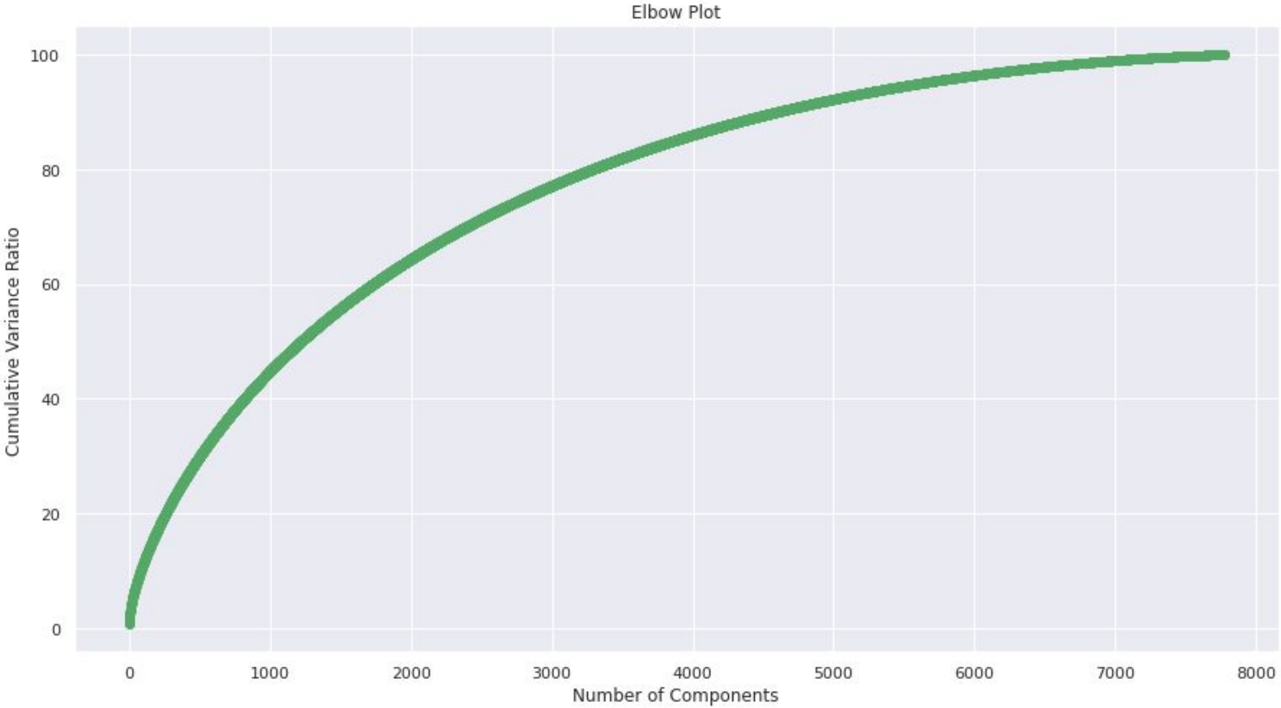


Very few TV Shows have more than 5 seasons.
Most of the Movies are around 90 to 120 minutes.

PCA

Why PCA?

Principal Component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

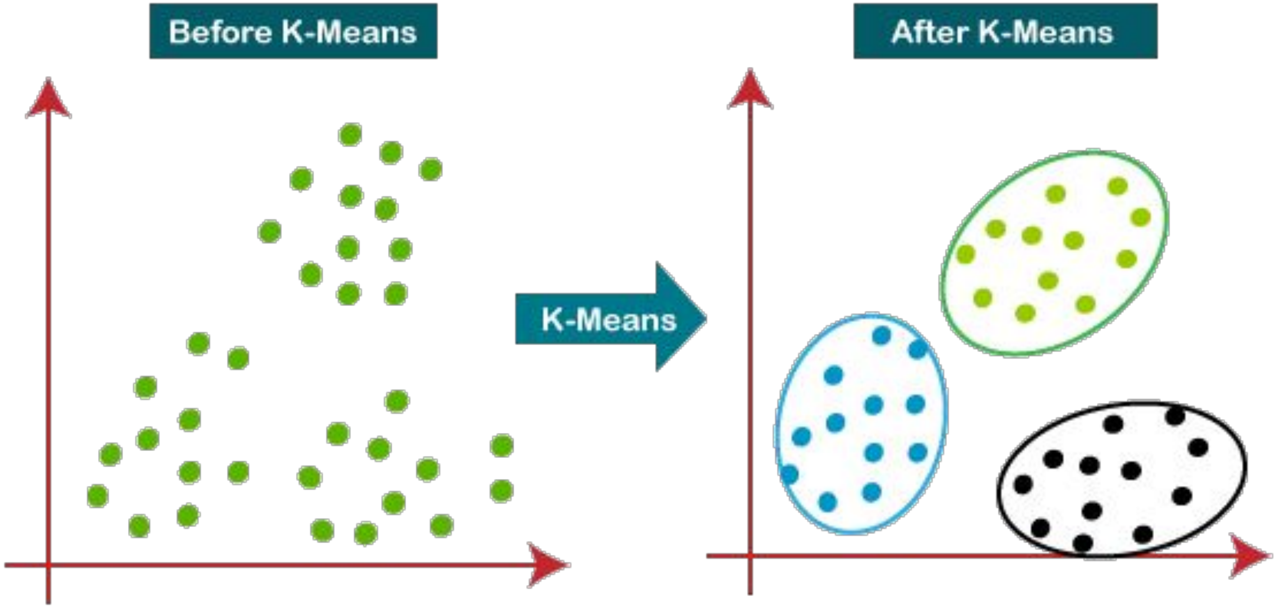


Almost 95% of variance is explained by 5500 components.

CLUSTERING

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

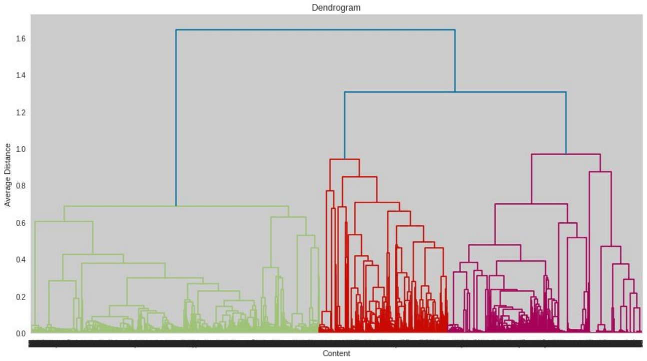
It is basically a collection of objects on the basis of similarity and dissimilarity between them.

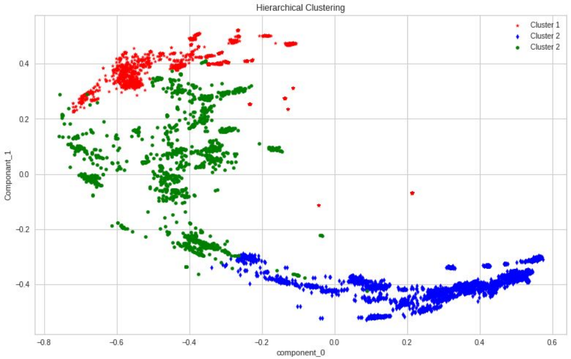


K-means Clustering Algorithm is the simplest unsupervised learning algorithm that solves clustering problem. It partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.

HIERARCHICAL CLUSTERING

Hierarchical clustering separates data into groups based on some measure of similarity, finding a way to measure how they’re alike and different, and further narrowing down the data.

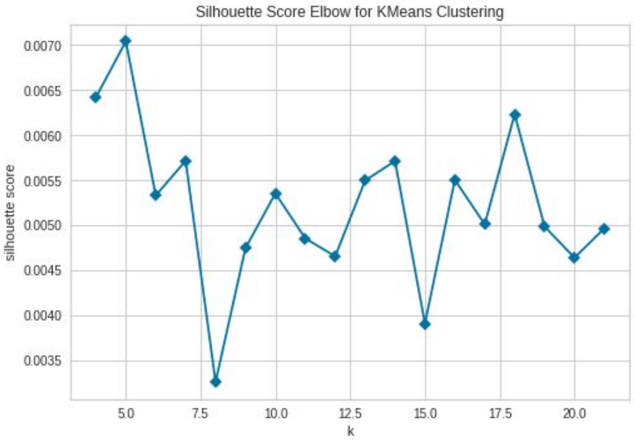


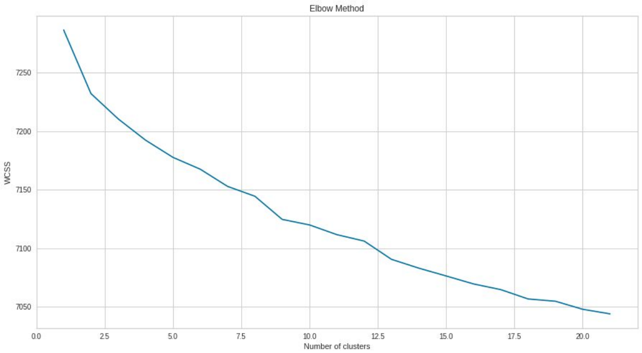


Dendrogram shows that 3 clusters would be suitable for the clustering the data by hierarchical clustering.

FINDING OPTIMAL NUMBER OF CLUSTERS FOR K-MEANS CLUSTERING

Optimal Clusters : 5



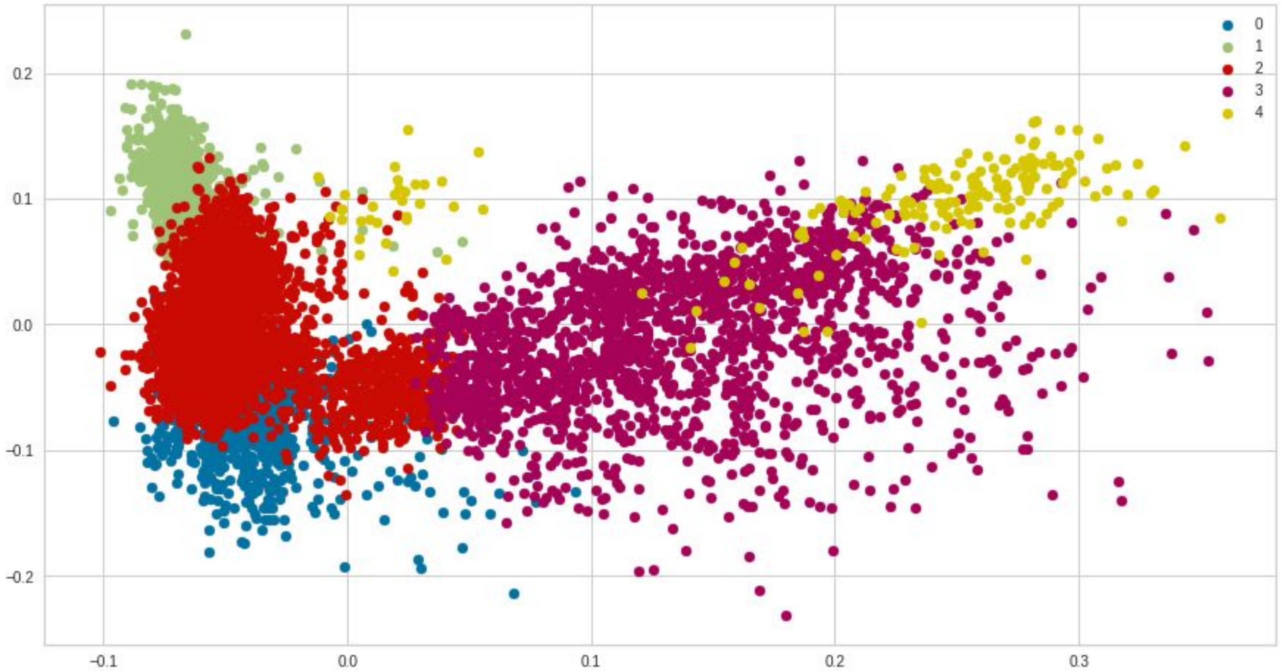


The optimal number of clusters selected here is 5 by using Silhouette Score and Elbow Method.

CLUSTERS VISUALIZATION

Cluster Label :

- ● 0
- ● 1
- ● 2
- ● 3
- ● 4



Total 5 distinct clusters are created by using K-means Clustering Algorithm.

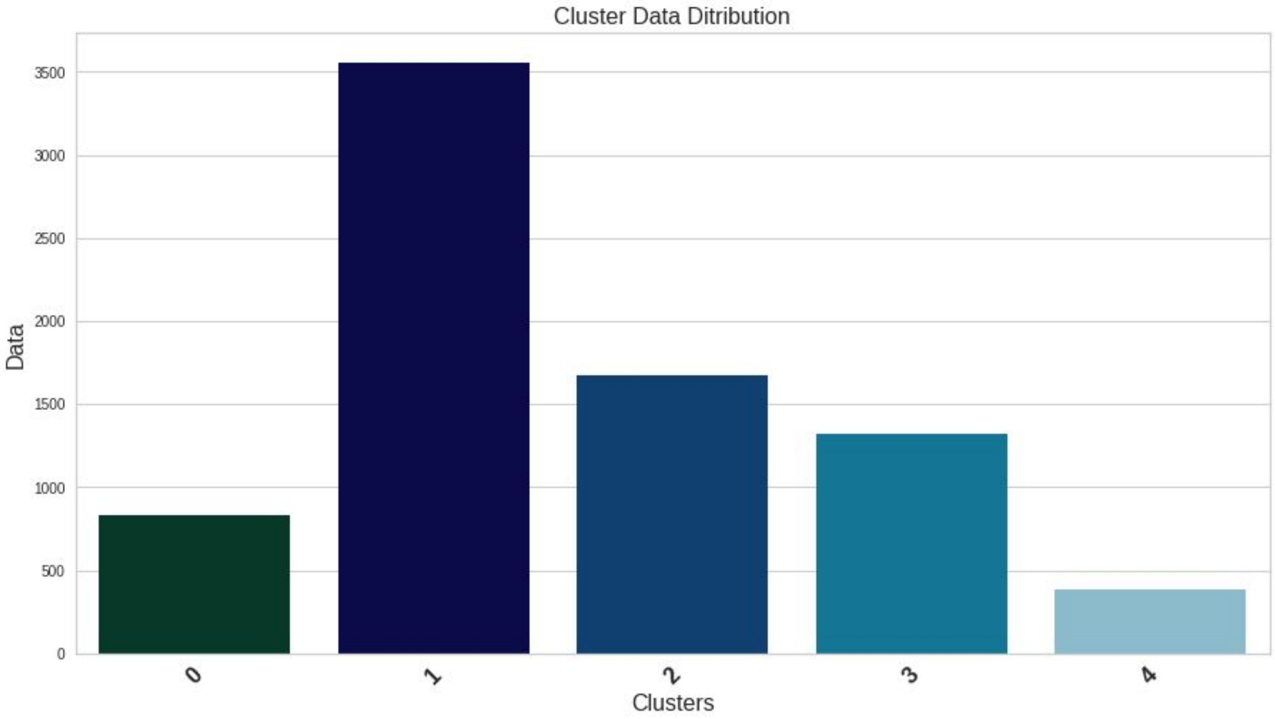
MODEL EXPLAINABILITY

Cluster with Highest Amount of Data :

#1

Cluster with Least Amount of Data :

#4



MODEL EXPLAINABILITY

Data Represented by Each Cluster :

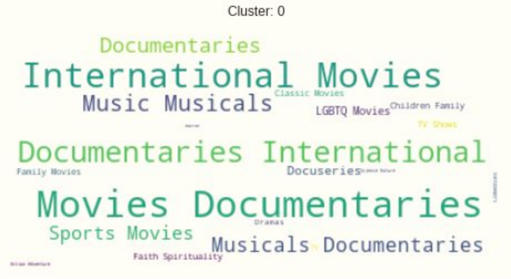
0 - Documentaries

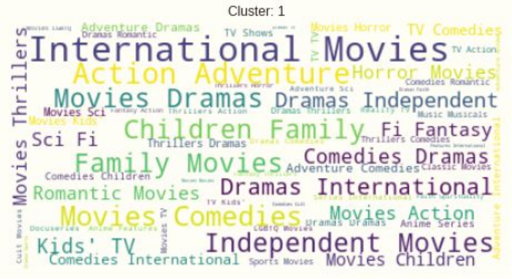
1 - Family and Children Movies

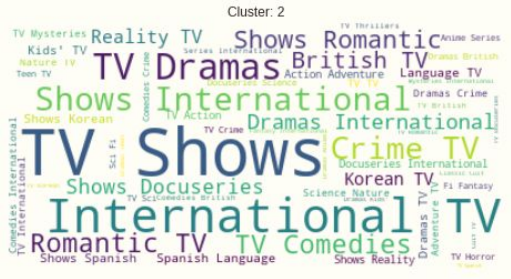
2 - International TV Shows

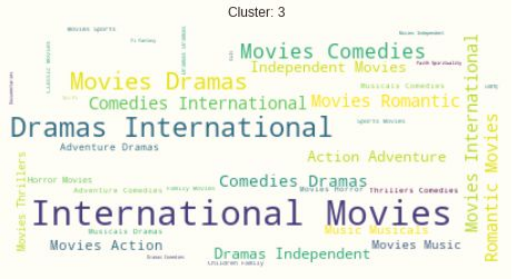
3 - International Movies and Dramas

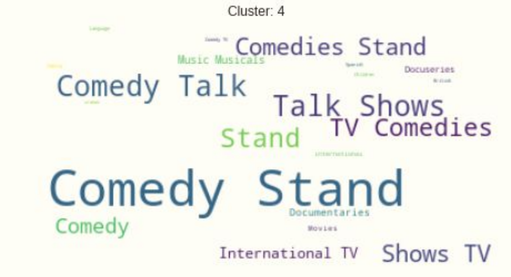
4 - Comedy Shows


Cluster: 0


Cluster: 1


Cluster: 2


Cluster: 3


Cluster: 4

CONCLUSION

Majority of content available on Netflix are movies.

Most of the TV Shows and Movies are added in the month October, November, December and January.

United States and India are the highest content producing countries.

Large number of content are for mature audiences.

Over past 5 years immense amount of TV Shows and Movies were released.

High percentage of TV-MA rating shows that Mature Content is more popular on Netflix.

CONCLUSION

TV Shows rarely go above 5 seasons and average time of a movie is around 90 to 120 minutes.

It was found that the optimal number of clusters was 5. Therefore total 5 distinct clusters were created using K-means Clustering Algorithm.

Documentaries, Family and Children Movies, International TV Shows, International Movies and Drama, Comedy Shows are the data represented in the clusters.

Thank You

As part of EDA Capstone Project by

AlmaBetter