Capstone Project

Netflix Movies & TV Shows Clustering

by Sayan Bandopadhyay

Synopsis

Netflix is the world’s leading premium media streaming platform, operating in nearly every country in the world. It is one of the first players in the streaming industry and the bet has paid off with hundreds of millions of subscribers worldwide. It has even led people using phrases like "binge watching" and "Netflix and chill".

Here, I have a Netflix dataset with details of rating, cast, director, genre, description etc. The main objective here is to explore and analyze the data to extract some fruitful insights from it and to cluster similar content, based on text based clusters which further helps in generating recommendations.

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/ TV show
“cast” →	Actors involved in the movie / TV show
“country” →	Country where the movie / TV show was produced
“date_added” →	Date the content was added on Netflix
“release_year” →	Actual release year of the movie / TV show
“rating” →	Rating of the movie / TV show
“duration” →	Total duration in minutes or number of seasons
“listed_in” →	Genre
“description” →	Description of the content

Data Briefing

This dataset is a mid-size one which initially had 7787 rows and 12 columns. Focusing on data information, 11 columns is of ‘object’ dtype and only one column is of ‘int64’ dtype which implies that almost all the columns are categorical except one.

Total count of null values : 3631

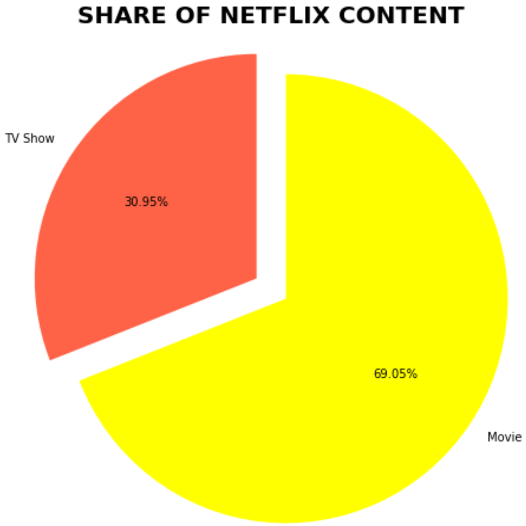
Column wise null value count:

- “director” column : 2389 (30.68% of total data)
- “cast” column : 718 (9.22% of total data)
- “country” column : 507 (6.51% of total data)
- “date_added” column : 10 (0.13% of total data)
- “rating” column : 7 (0.09% of total data)

Analysis Report

Content Type

There are basically two types of content in the dataset: Movies and TV shows. Count of movies and TV shows are 5372 and 2408 respectively.



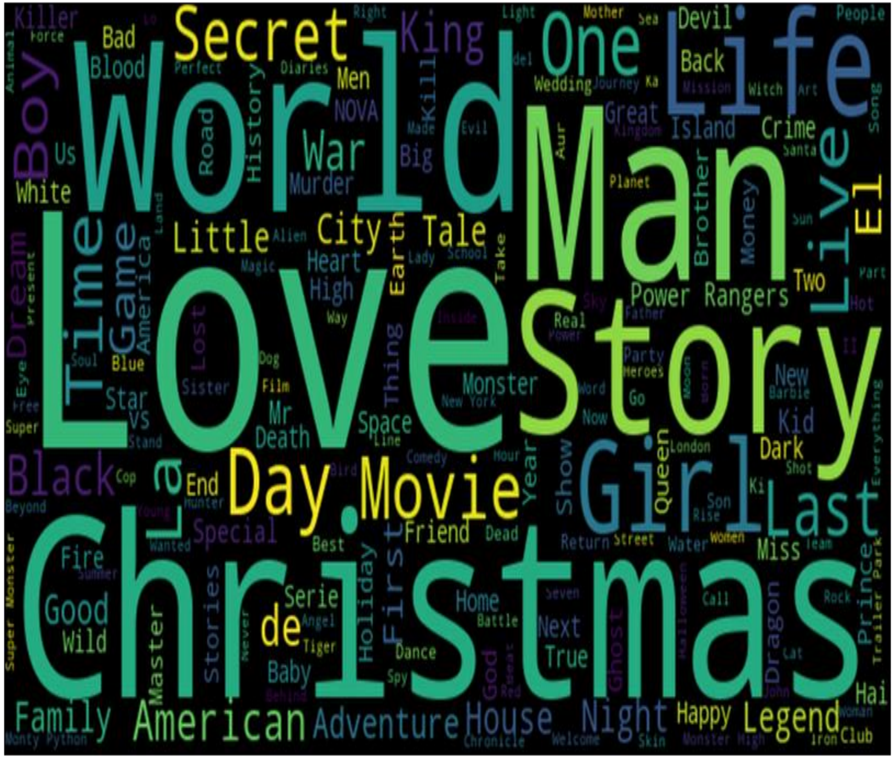
Its clear from the count that movies are more than TV shows.
Pie chart on the left helps denoting the percentage share of each content in the dataset.

- 69.05% of total content are movies.
- 30.95% of total content are TV shows.

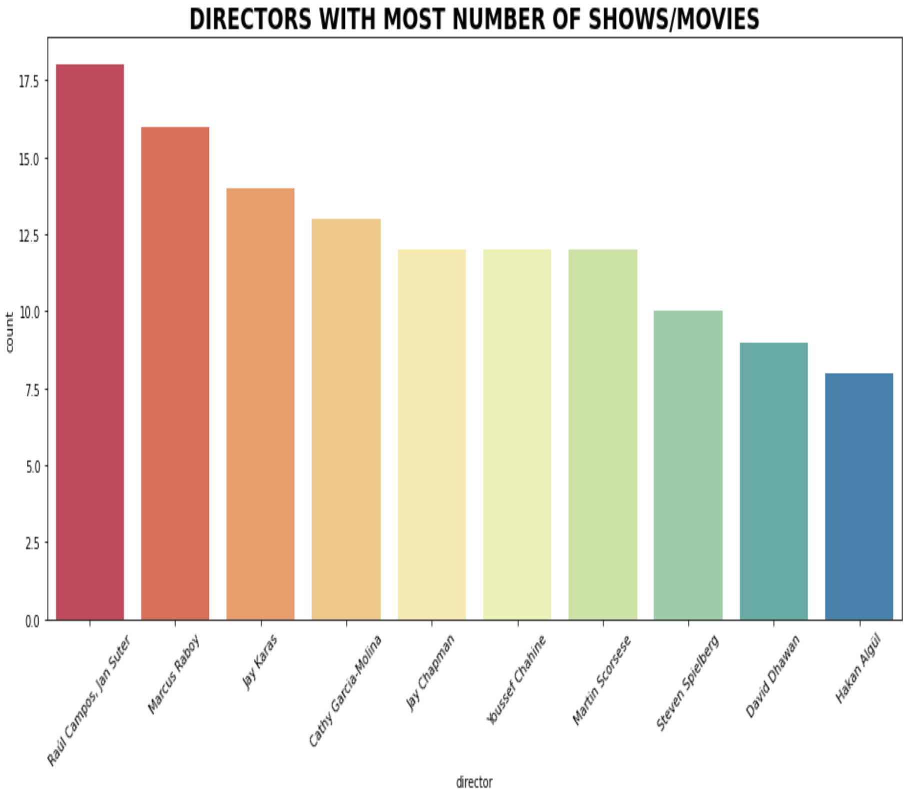
Content Titles

Performed analysis so as to find the most utilised words in content titles. On right hand side the word cloud depicts it all.

Clearly,
'Christmas', 'Love', 'Man', 'World' are some of the most utilized words for content titles.



Top Directors



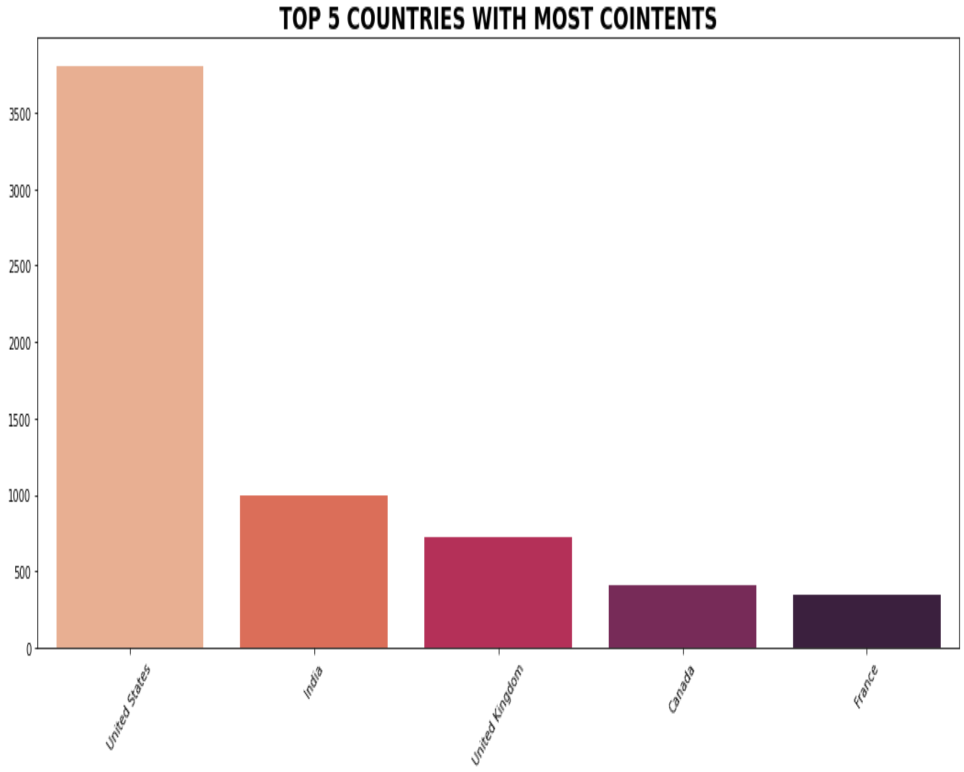
A lot of directors were listed in the dataset but for convenience the analysis has been done to find only the top 10 directors with most number of contents.

'Raul Campos' & 'Jan Suter' seized the top position with maximum number of contents out of all followed by 'Marcus Raboy', 'Jay Karas' and so on.

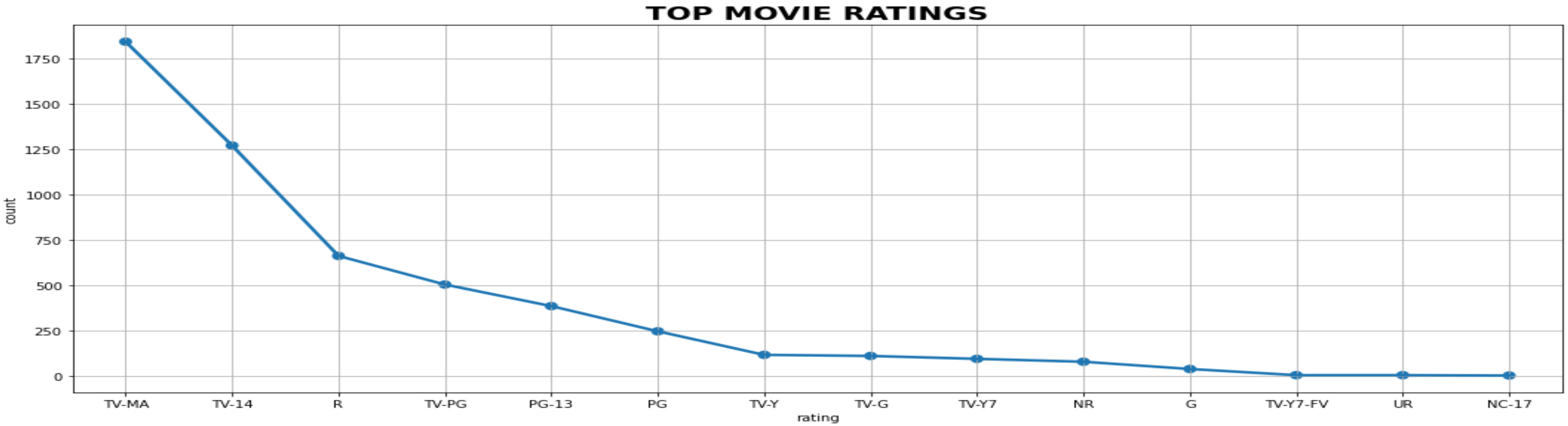
Top Countries

Its all about the analysis on the countries that has produced the maximum number of contents.

Considering only the top 5 countries, we get to know that 'United States' has maximum production followed by 'India', 'United Kingdom' and so on.



Content Ratings

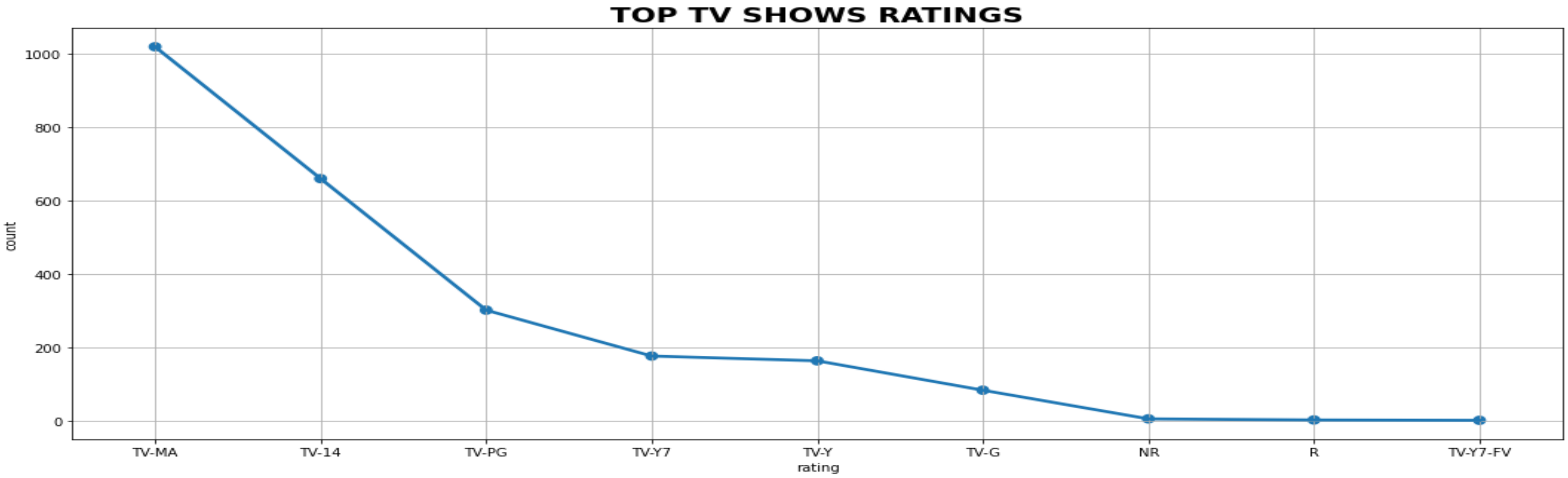


Above line plot depicts different movie ratings and also the rating given for most number of movies in the dataset.

No doubt most of the movies are ‘TV-MA’ rated.

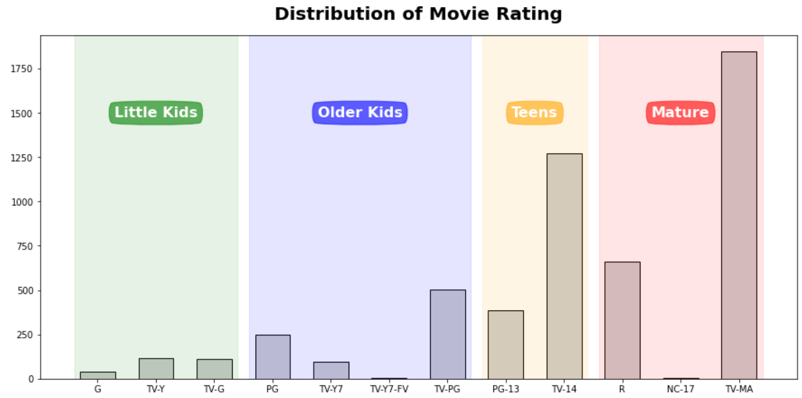
(TV-MA rating means that the movie is intended to be viewed by mature and adult audiences)

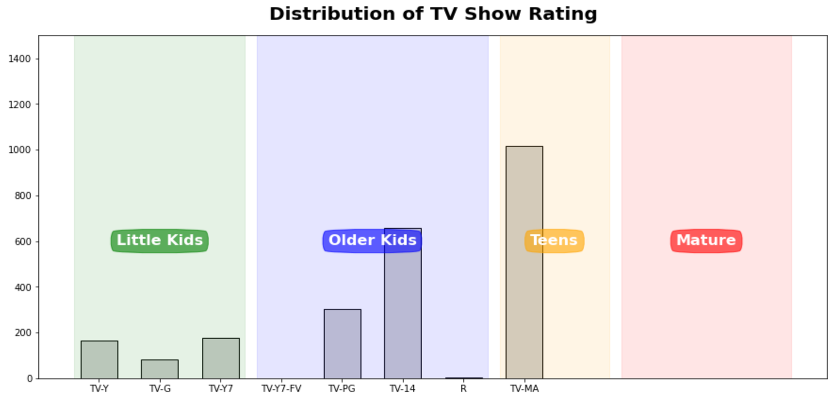
Continued.....



This one also looks similar to the previous one but this depicts the ratings given to most of the TV shows.
Same as movie ratings, most of the TV shows are also ‘TV-MA’ rated.

Continued.....


Distribution of Movie Rating


Distribution of TV Show Rating

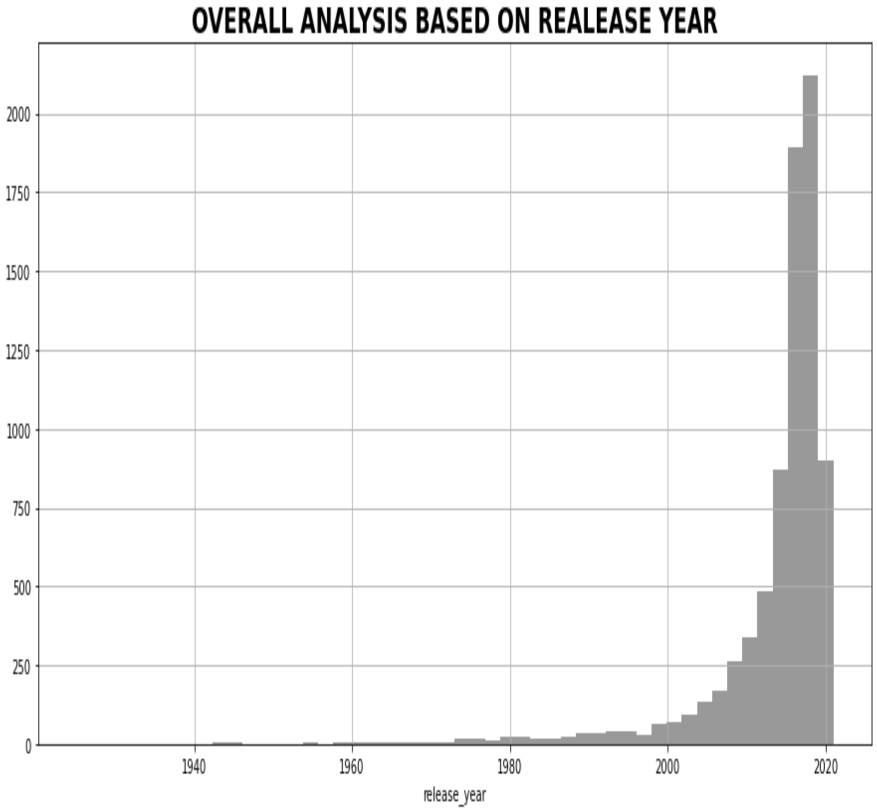
This analysis is purely to understand different ratings on which movies and TV shows has been rated. So, similar ratings is grouped under a specific label. Labels taken into consideration are ‘Little Kids’, ‘ Older Kids’, ‘Teens’ and ‘Mature’.

Label ‘Older Kids’ have the highest distribution with maximum rating category under it.

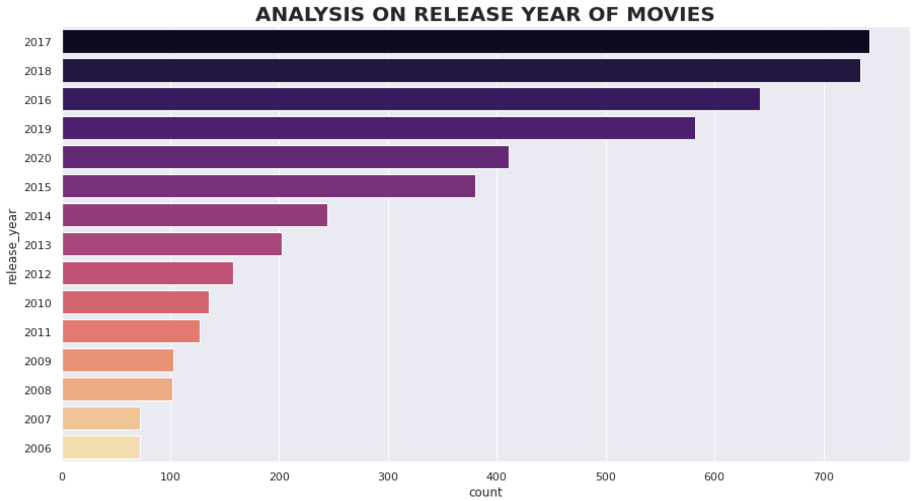
Year Of Release

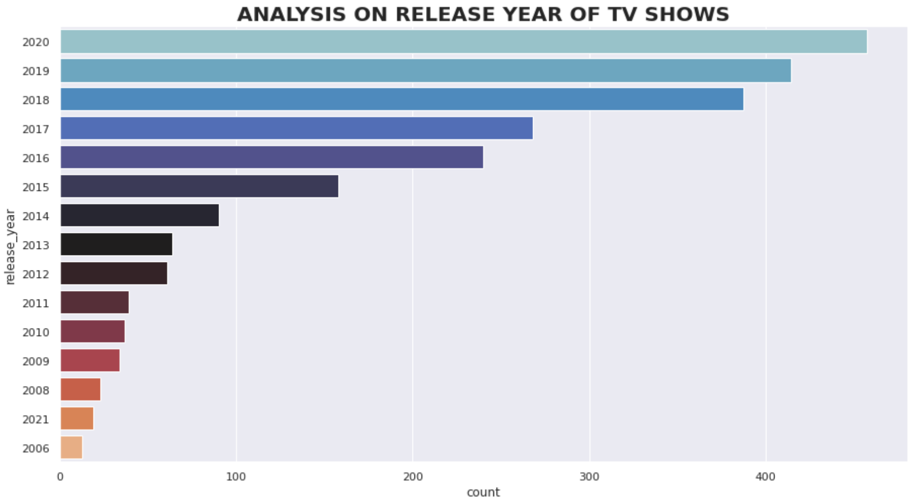
There has been a column which stored the respective release year of each movies / TV shows. Here, what I found post analysis of this column:

- Oldest content present was released in 1925.
- Most of the contents has been released during the last decade(2010-2020) .



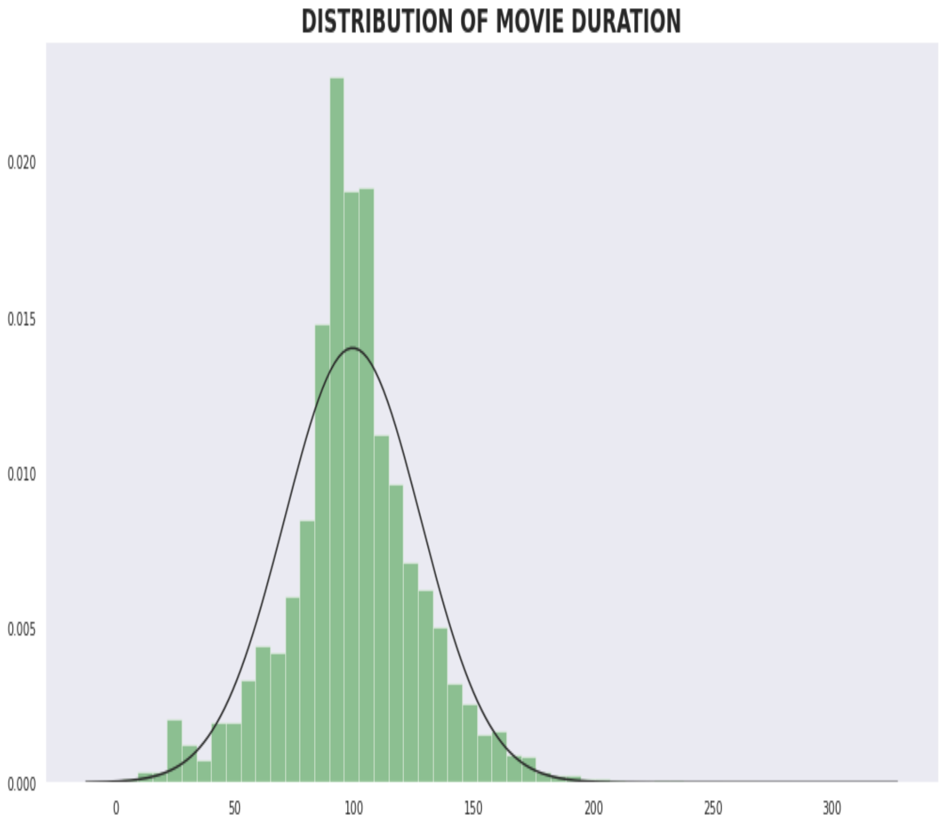
Continued.....





- Left hand plot clearly depicts that most of the movies were released in the year 2017 followed by 2018, 2016 and so on.
- Right hand plot depicts that most of the TV shows were released in the year 2020 followed by 2019, 2018 and so on.
- 2006 has been the year of least number of releases for both movies and TV shows.

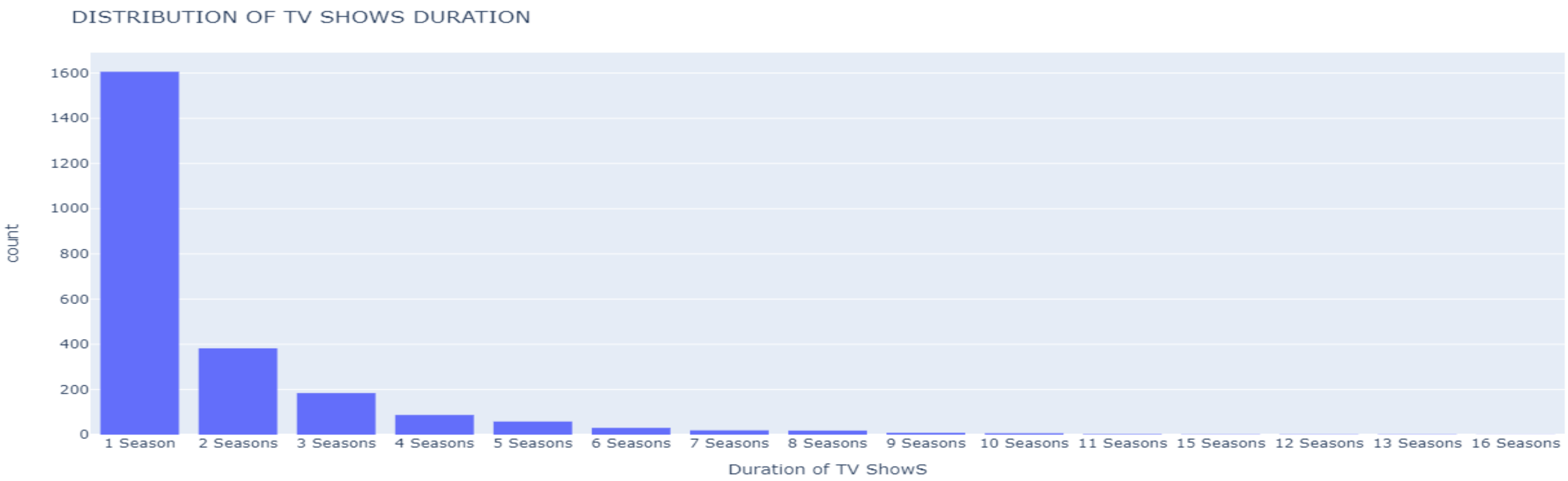
Content Duration



The plot on the left shows the distribution of movie duration. As the movie duration is in minutes which is continuous in nature so, I preferred to use this distribution plot instead.

Bell shaped curve states that duration of the movies is normally distributed and majority of the movies have duration ranging from 85 minutes to 120 minutes.

Continued.....



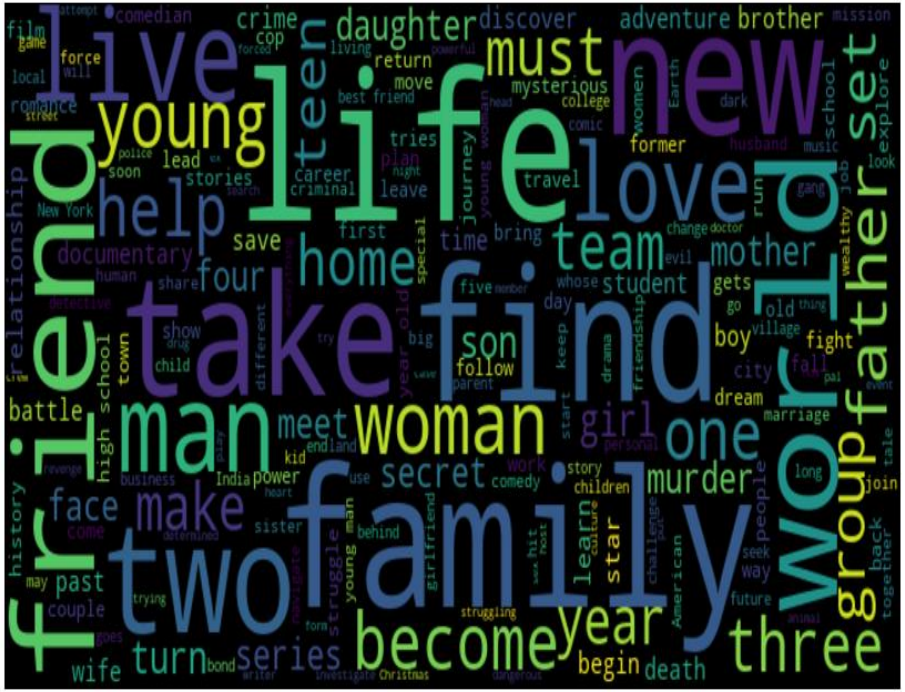
Above plot presents the graphical analysis of TV show’s duration which is listed in seasons. So, its clear that,

- Most of the TV shows streamed have only 1 season.
- Majority of the TV show ends by season 3.

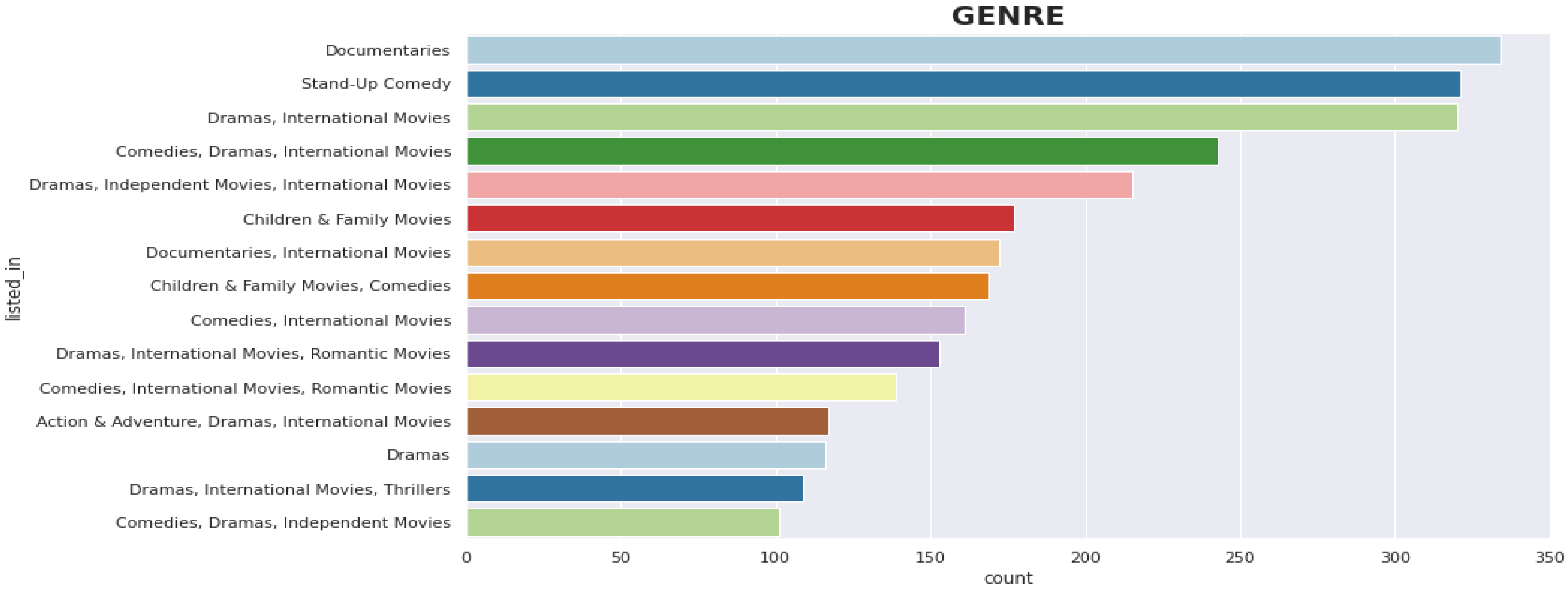
Content Description

Main purpose of this word cloud is to find the most utilized word in content description.

Clearly,
'Family', 'life', 'find', 'friend' are some of the most utilized words in content description.

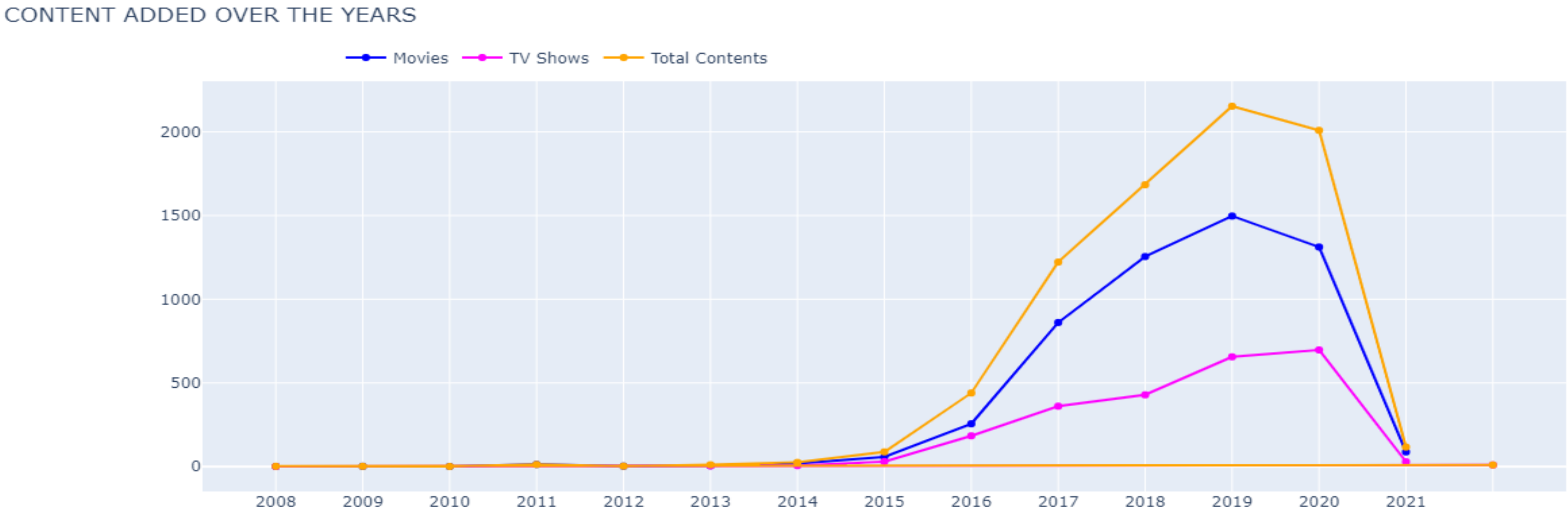


Genre Of Content



Most of the content in the dataset are basically ‘Documentaries’.

Netflix Industry Over Years



The growth in number of movies on Netflix is much higher than that of TV shows and the growth of contents took a plunge from 2014 onwards.

Data Preparation

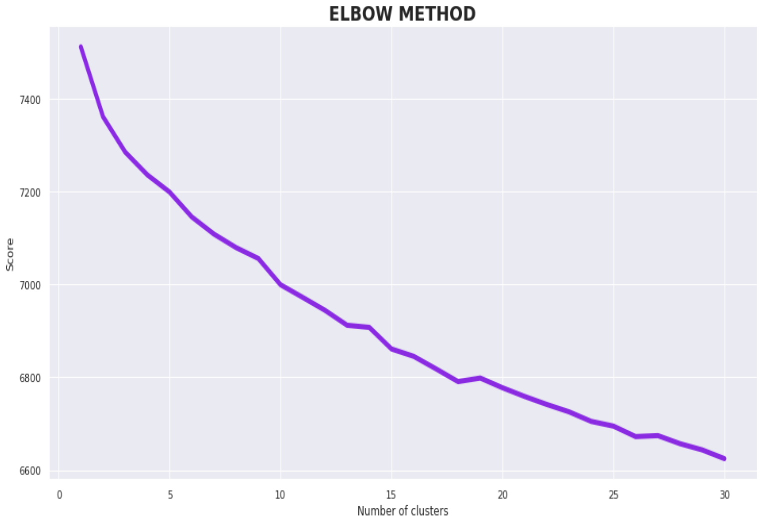
Post EDA, data preparation is one of the most crucial step to be taken care of before implementation of any machine learning model. This particular step basically includes text cleaning, feature engineering, organising the dataset etc.

Steps performed for data preparation in this particular case are:

- Split the information in ‘cast’, ‘director’ and ‘country column and only considering the primary information.
- Creating a new column combining selected features so as to perform vectorization.
- Lowercased all the information.
- Replaced some unnecessary symbols in the data with spaces.
- Removed stop words.
- Stemmed the words.

K-Means Clustering

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

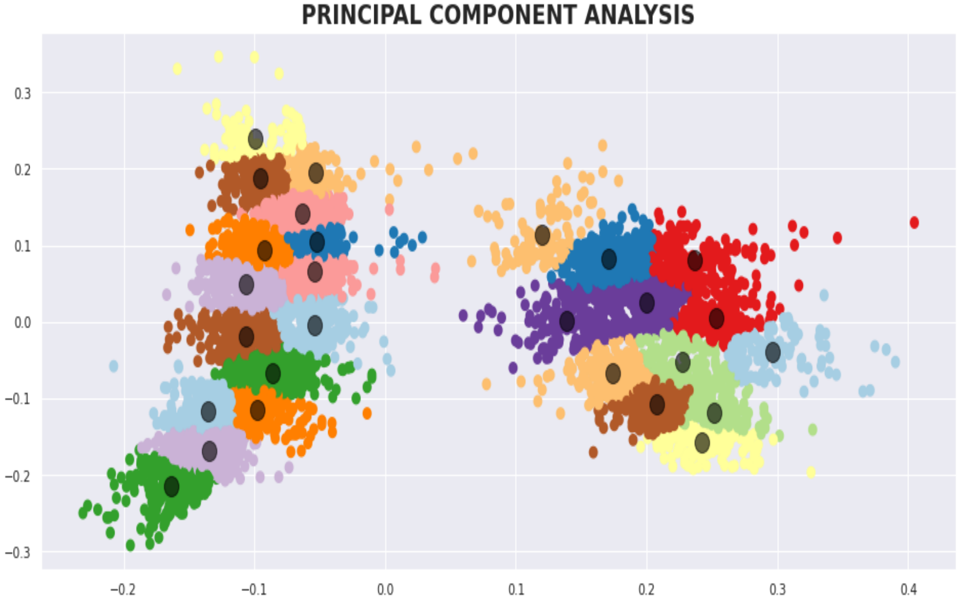


The 'Elbow method' is a heuristic method of interpretation and validation of consistency within cluster analysis designed to find the appropriate number of clusters. Left hand plot shows the implementation of the elbow method.

Continued.....

Now, its time to make use of ‘Silhouette method’ which is nothing but a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Silhouette value ranges from ‘-1’ to ‘+1’. So, optimal cluster number is found out to be 27.

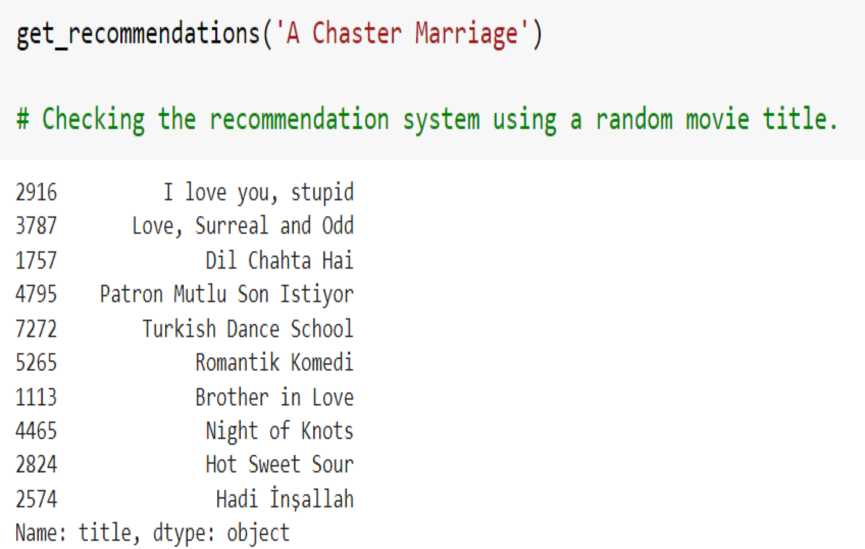
PCA is a method of analysis which involves finding the linear combination of a set of variables that has maximum variance and removing its effect, repeating this successively.



Silhouette score before PCA: 0.276
Silhouette score after PCA: 0.356

Recommendation System

After all the processes, now its ready to create a recommendation system so as to suggest recommendations based on the item metadata. In this particular case, I have used cosine similarity method on the same vectorized data.



```
get_recommendations('A Chaster Marriage')

# Checking the recommendation system using a random movie title.
```

```
2916          I love you, stupid
3787        Love, Surreal and Odd
1757              Dil Chahta Hai
4795     Patron Mutlu Son Istiyor
7272          Turkish Dance School
5265              Romantik Komedi
1113              Brother in Love
4465              Night of Knots
2824              Hot Sweet Sour
2574               Hadi İnşallah
Name: title, dtype: object
```

On the left hand side, I shared a glimpse of how the recommendation system worked in this case.

For example, finding recommendation similar to ‘A Chaster Marriage’ resulted the below list of contents as recommendations.

Conclusions

1.There are basically two types of content in the dataset: 'Movies' & 'TV shows'.
2.Netflix has more movies which constitutes 69.05% of total content than TV shows which is only 30.95% of total content.
3.'Christmas', 'Love', 'Man', 'World' are some of the most utilized words for movie titles.
4.'Raul Campos and Jan Suter' have most number of contents in this particular dataset.
5.'United States' tops the list with maximum number of contents.
6.Most of the movies are 'TV-MA' rated.
7.Most of the TV shows are 'TV-MA' rated as well.
8.Most movies were released during the last decade (2010 - 2020) compared to all other time periods.
9.Most of the movies were released in the year 2017 followed by 2018 and 2016.
10.Most of the TV Shows were released in 2020 followed by 2019 and 2018.
11.Majority of the movies have duration ranging from 85 minutes to 120 minutes.
12.'Family', 'life', 'find', 'friend' are some of the most utilized words in description.
13.Most of the streamed TV Shows have only one season.
14.Most of the contents are basically documentaries.
15.Optimal number of clusters were found out to be 27 with silhouette coefficient value of 0.02765.
16.Principal component analysis was performed in order to reduce the dimensionality which improved the silhouette coefficient to 0.3561.

THANK YOU