Capstone Project-4

NETFLIX MOVIES & TV SHOWS CLUSTERING

Submitted by

Shubham Kadu

Data science trainee, Almabetter

Contents:

- Introduction
- Problem Statement
- Data Summary
- Data wrangling
- Exploratory data analysis
- Text Preprocessing
- ML-Model(Clustering)
- Conclusion

Introduction:

- Netflix is a media distribution company and a prominent OTT platform with a wide variety of content to view from a variety of nations and genres. It started with DVD distribution via mail but has evolved substantially throughout its existence. Today, Netflix is focused on streaming video.

- Netflix originally focused on movies, but television shows are probably the more common format today. Its works on a subscription model, where users get unlimited access to content with a paid subscription.

- 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.

- This project aims to analyze and perform clustering to determine patterns related to the content available on Netflix. Based on the attributes related to the Tv shows or movies, we will implement different clustering algorithms that come under the unsupervised Machine learning category.

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 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, and rotten tomatoes can also provide many interesting findings.

Dataset Preview:

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/show
- country: The country where the movie/show was produced
- date_added: Date it was added on Netflix
- release_year: Actual Releaseyear of the movie/show
- rating: TV Rating of the movie/show
- duration: Total Duration - in minutes or number of seasons
- listed_in : Genere
- description: The Summary description

Data Summary :

- In the Netflix movies & Tv shows project there is a dataset that contains Netflix Tv shows & movie information.

- Netflix has 7787 rows and 12 columns

- Data contains information on rating, duration, Genre, country, director, Title, and release year.

```
<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
```

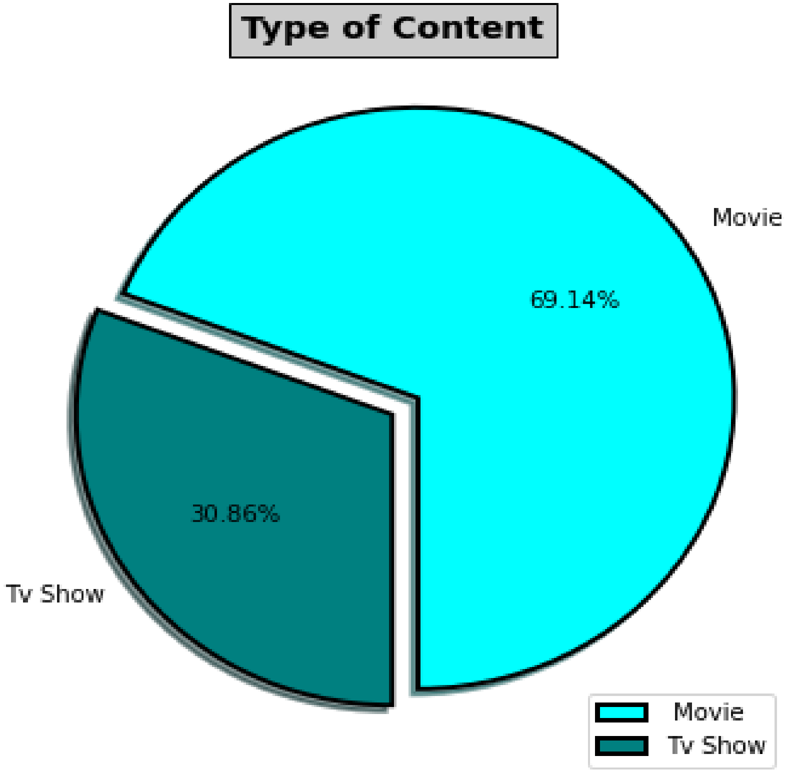
Exploratory Data Analysis (EDA):

Type :

- Netflix has 69.14% of the content available on Netflix is movies; the remaining 30.86% is TV Shows.

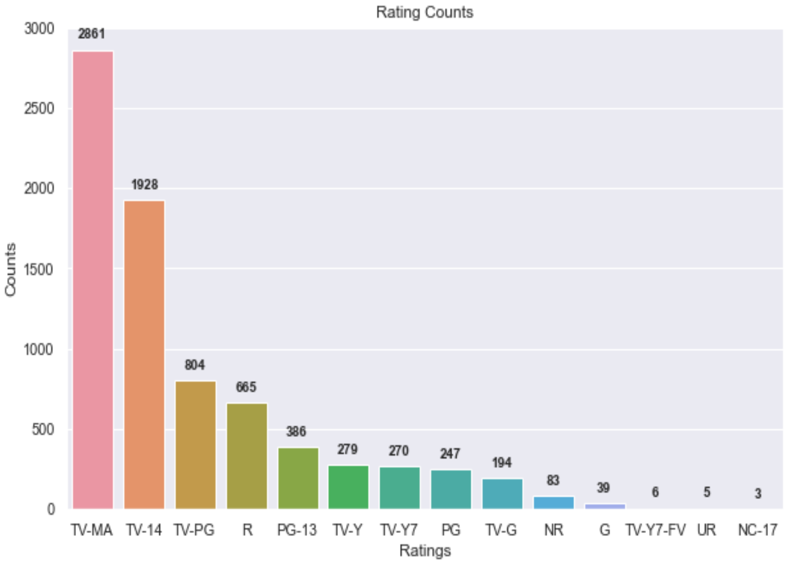
- It is evident that there are more movies on Netflix than on TV shows.

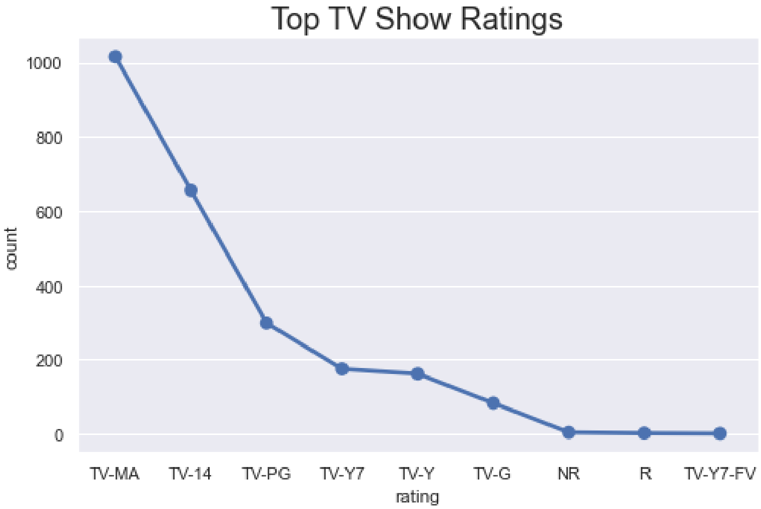
.

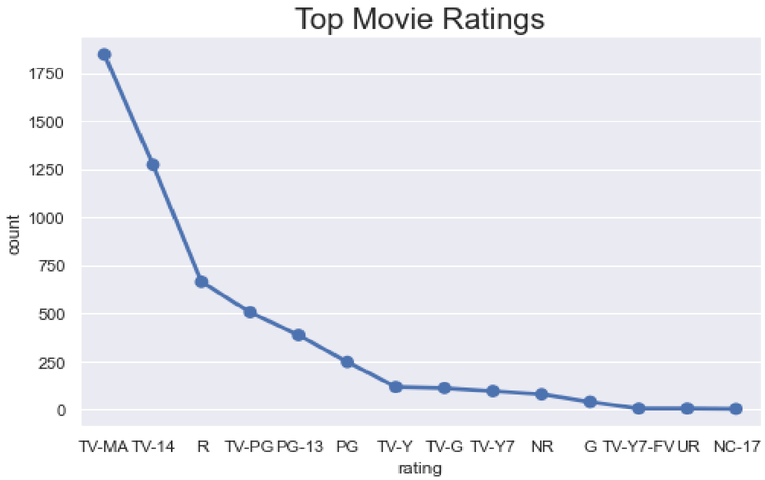


Ratings

- From the below graph, we can see that TV-MA has the highest rating after following TV-14.
- From the right top graph TV-MA has the highest number of ratings for TV shows, i.e. adult ratings.
- From the right top graph TV-MA has the highest number of ratings for movies, i.e. Adult ratings.



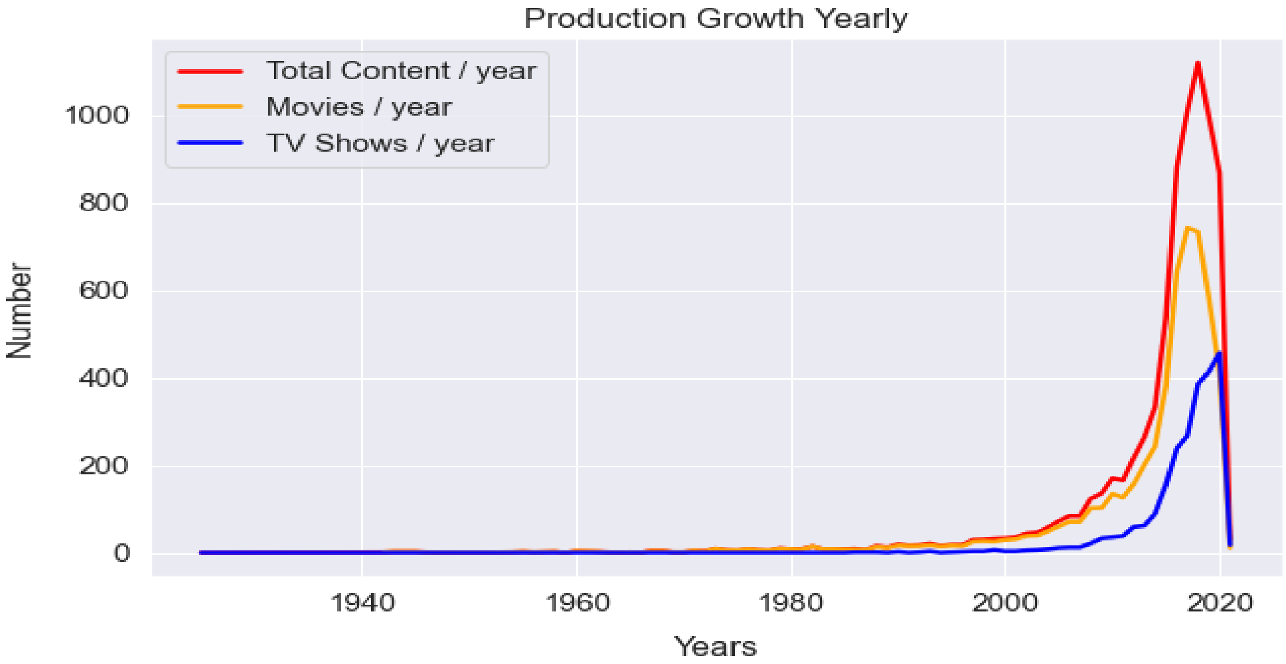




Production Growth

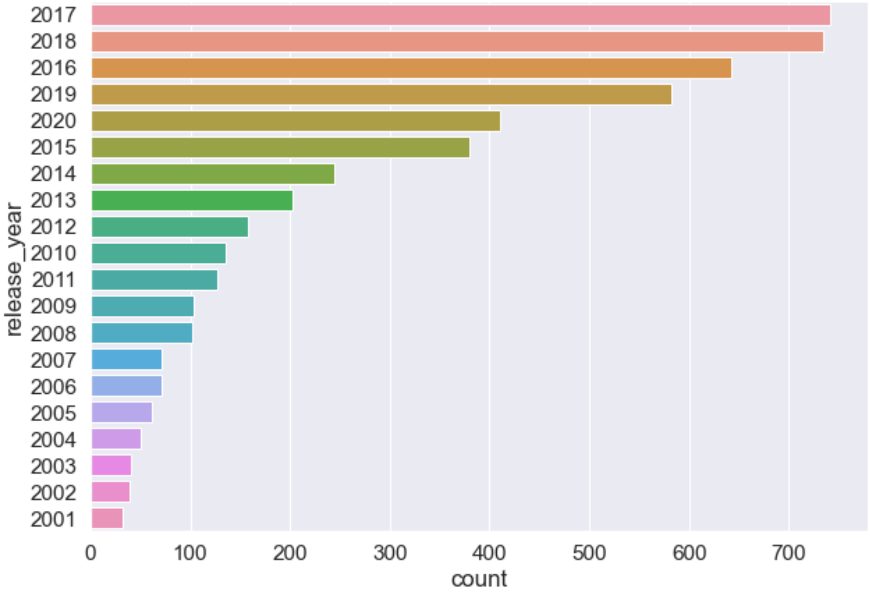
The below plot clearly see that in the year 2000 after that total content yearly increased and then movie content yearly increased but in the year 2020 suddenly fall down

It appears that Netflix has focused more after the year 2000 more attention on increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows.



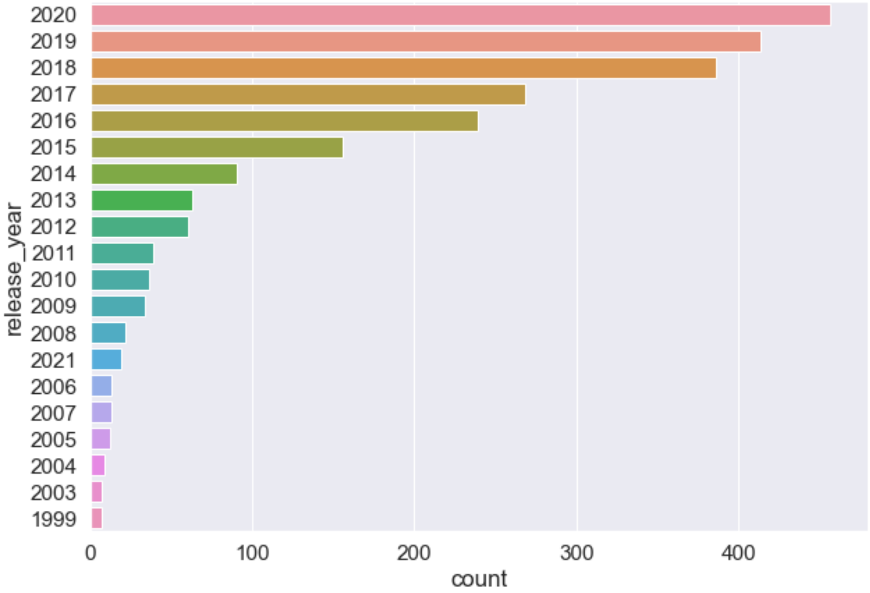
Production Growth based on Movies

- The highest number of movies were released in 2017 and 2018.
- We saw a huge increase in the number of movies after 2015.
- The number of movies on Netflix is growing significantly faster than the number of TV shows.



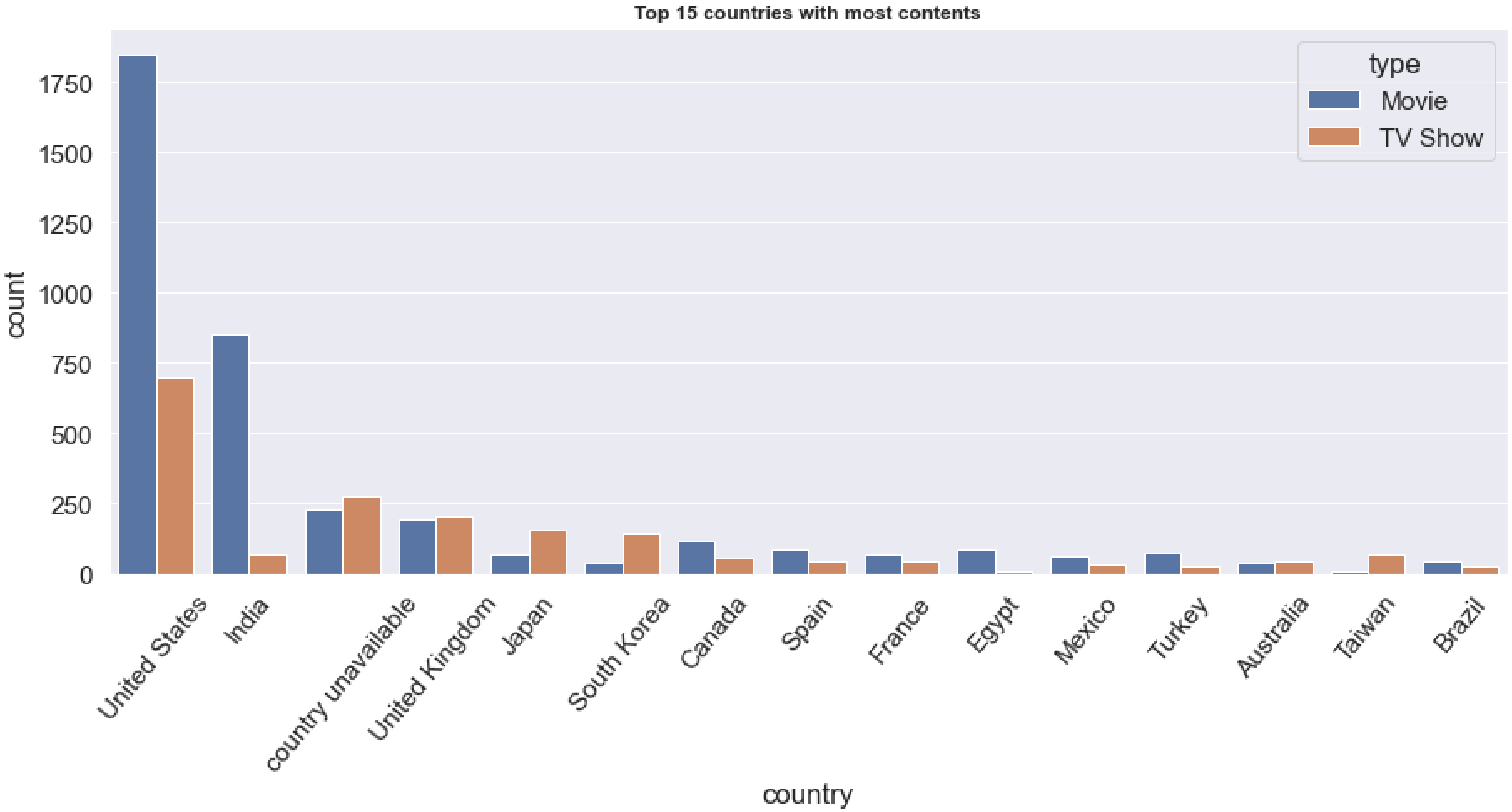
Production Growth based on TV Shows

- The highest number of TV Shows released in 2020
- We saw a huge increase in the number of tv shows after 2015.
- there is a significant drop in the number of tv show produced after 2020.



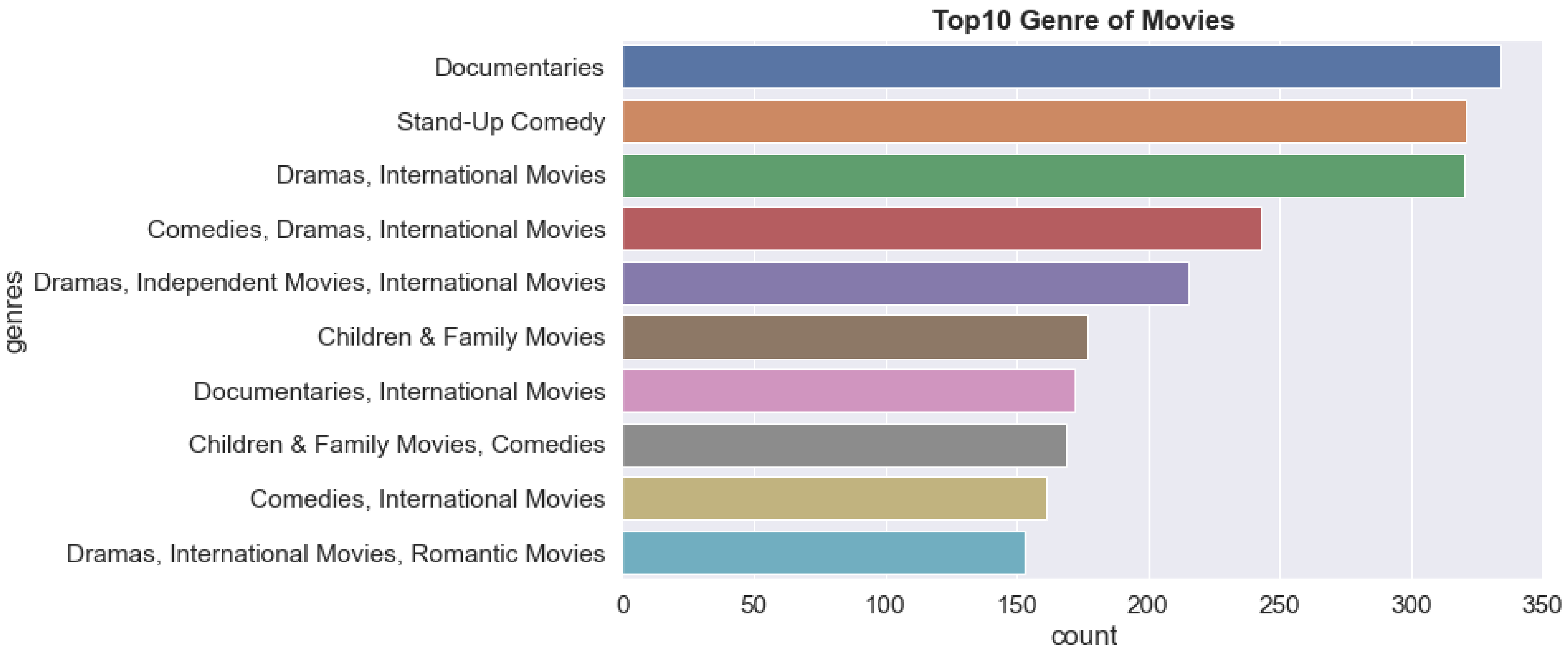
COUNTRY

The United States has the highest number of content on Netflix for watching more movies than TV shows, with India and the United Kingdom trailing far behind.



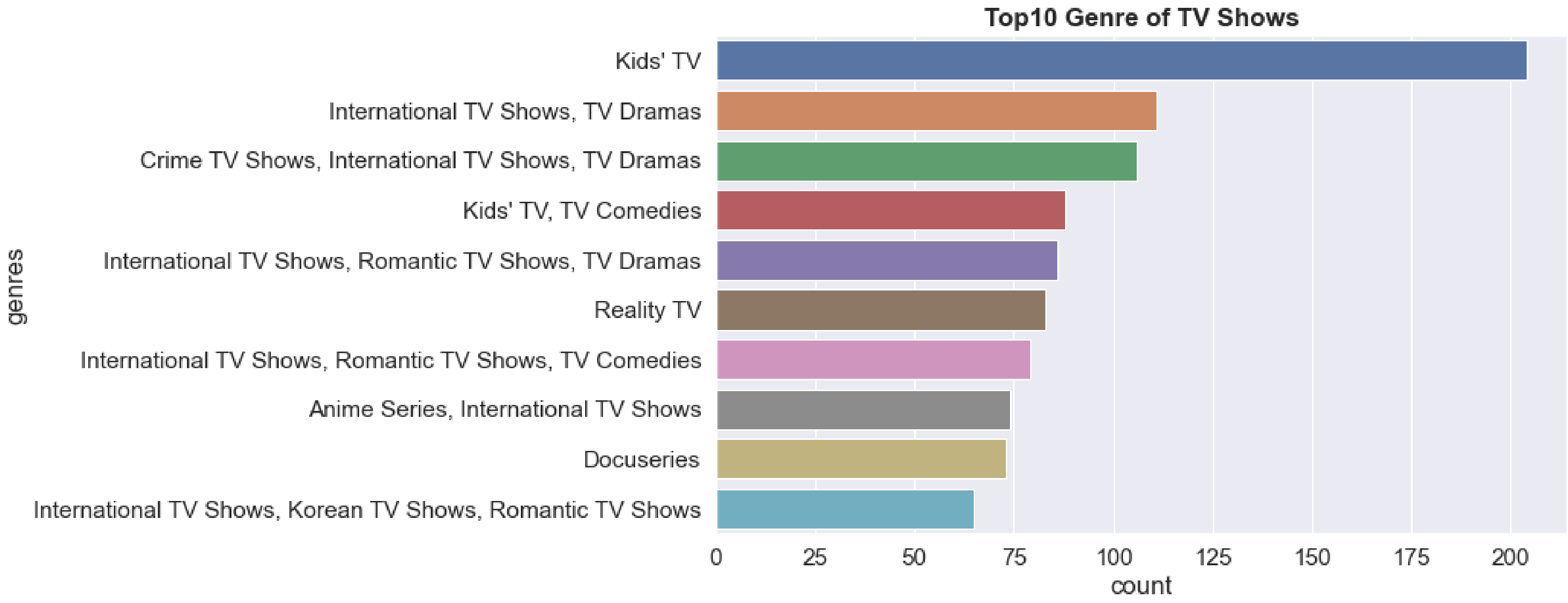
Top 10 Genre of Movies

Documentaries are the top most genre on Netflix which is followed by standup comedy dramas and international movies.



Top 10 Genre of TV shows

kids tv is the top show genre on Netflix.

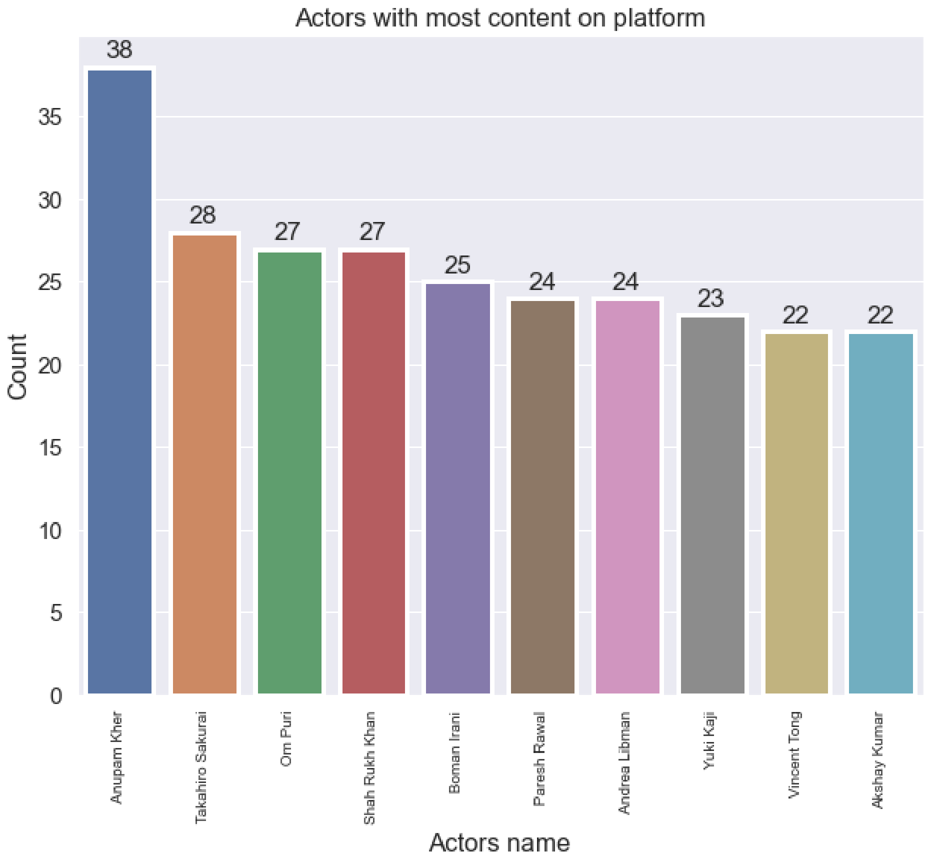


Top 10 Actors

There are 6 actors in the top ten list of most numbers of tv shows and movies from India.

According to the above bar plot, Anupam Kher has worked in over 38 films.

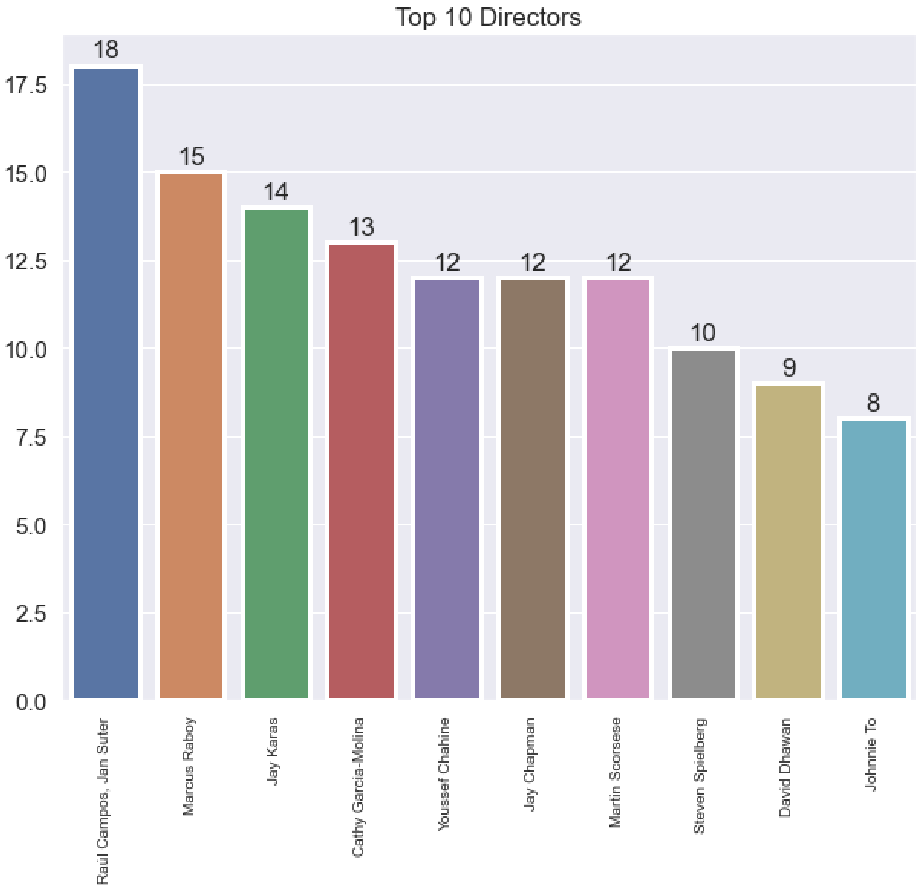
After Anupam Kher, Takahiro Sakurai is ranked second, with 28 films under his belt.



Top 10 Directors

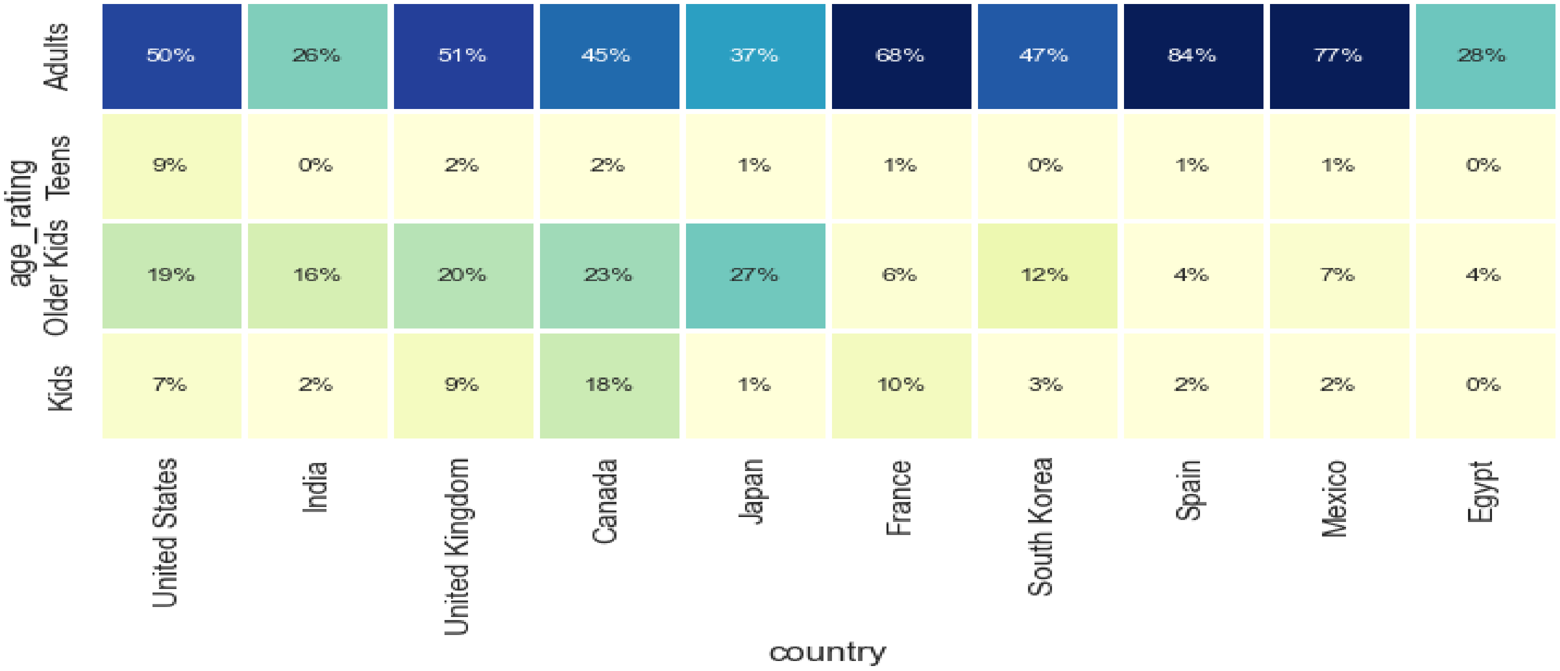
The Directors who produce the most material are Raul Campos and Jan Sutler. They work in 18 movies as a director.

Marcus Raboy is ranked second among top directors, having directed 15 films.



Netflix Content for different age groups in the countries

US and UK are closely aligned with their Netflix target ages, but radically different from, for example, India or Japan!
Also, Mexico and Spain have similar content on Netflix for different age groups.

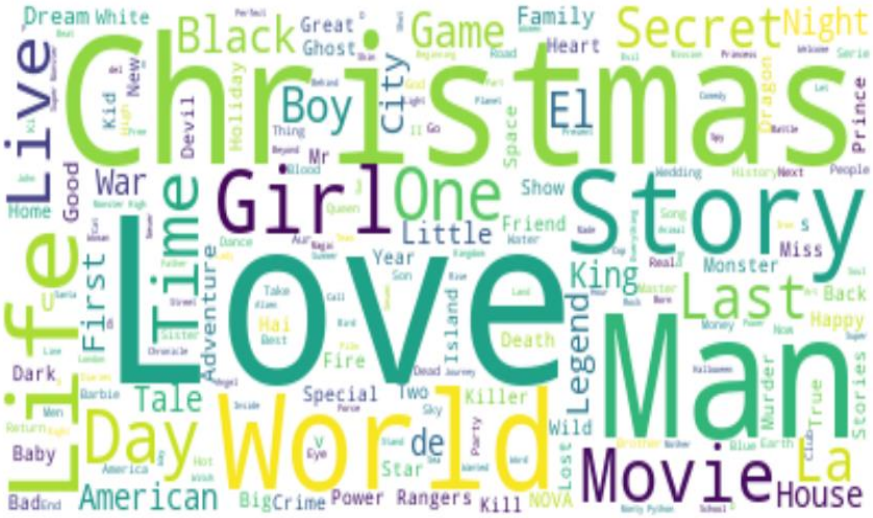


	United States	India	United Kingdom	Canada	Japan	France	South Korea	Spain	Mexico	Egypt
Adults	50%	26%	51%	45%	37%	68%	47%	84%	77%	28%
Teens	9%	0%	2%	2%	1%	1%	0%	1%	1%	0%
Older Kids	19%	16%	20%	23%	27%	6%	12%	4%	7%	4%
Kids	7%	2%	9%	18%	1%	10%	3%	2%	2%	0%

Title

It seems like words like “Love”, “Man”, “World”, “Story”, and "Christmas" are very common in titles.

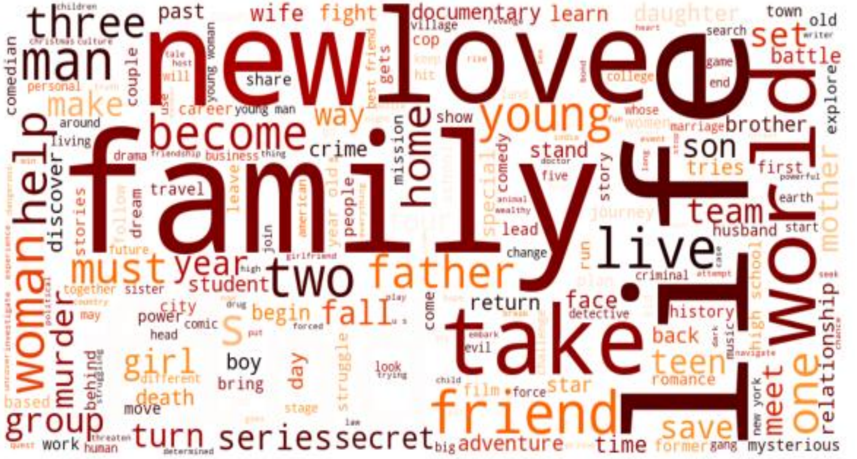
I have been surprised to see "Christmas’’ occur so many times. The reason may be that movie was s released in the month December.



Description

The most occurring words in the description of the tv shows and movies are Family, Friend, Love, Life, Woman, and Man.

Most used words in description



Text Pre-processing

1. Removing Punctuation:

- Punctuations do not carry any meaning in clustering.
- So, removing punctuations helps to get rid of unhelpful parts of the data, or noise.

2. Removing Stopwords:

- Stopwords are basically a set of commonly used words in any language, not just in English.
- If we remove the words that are very commonly used in a given language, we can focus on the important words instead

3. Stemming :

- Stemming is the process of removing a part of a word or reducing a word to its stem or root.
- Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

CLUSTERING

- Clustering is a type of unsupervised learning method of machine learning .in the unsupervised learning method. The inferences are drawn from the datasets which not contain labeled output variables. It is an exploratory data analysis technique that allows us to analyze multivariate data sets.
- Define: 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

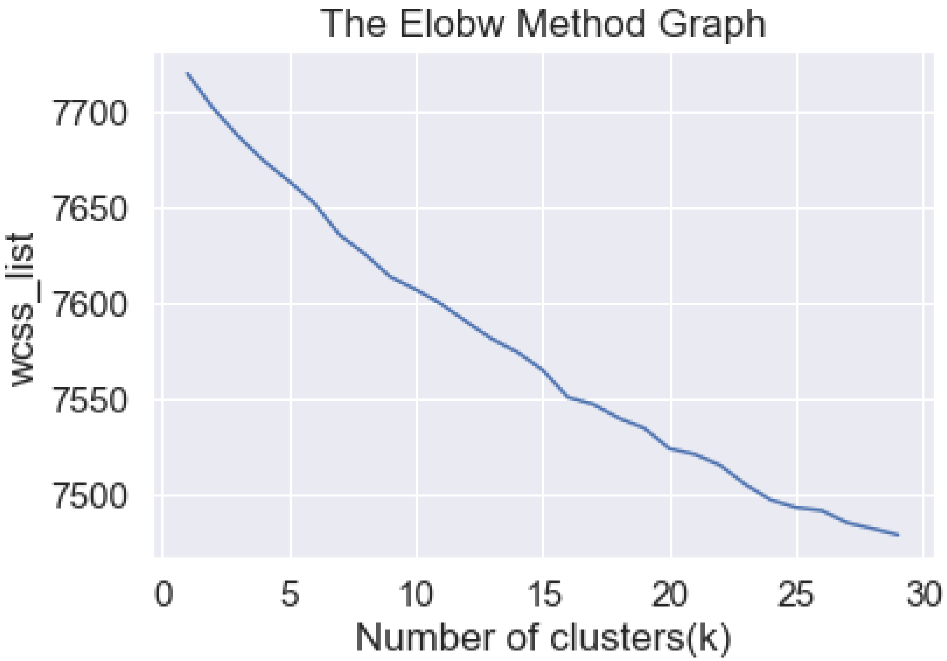
.

- 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 object belonging to another cluster.

- K-means clustering which is an iterative process in which the dataset is grouped into k number of predefined non-overlapping clusters or subgroups, making the inner points of the cluster as similar as possible while trying to keep the clusters at distinct space it allocates the data points to a cluster so that the sum of the squared distance between the clusters centroid and the data point is at a minimum.

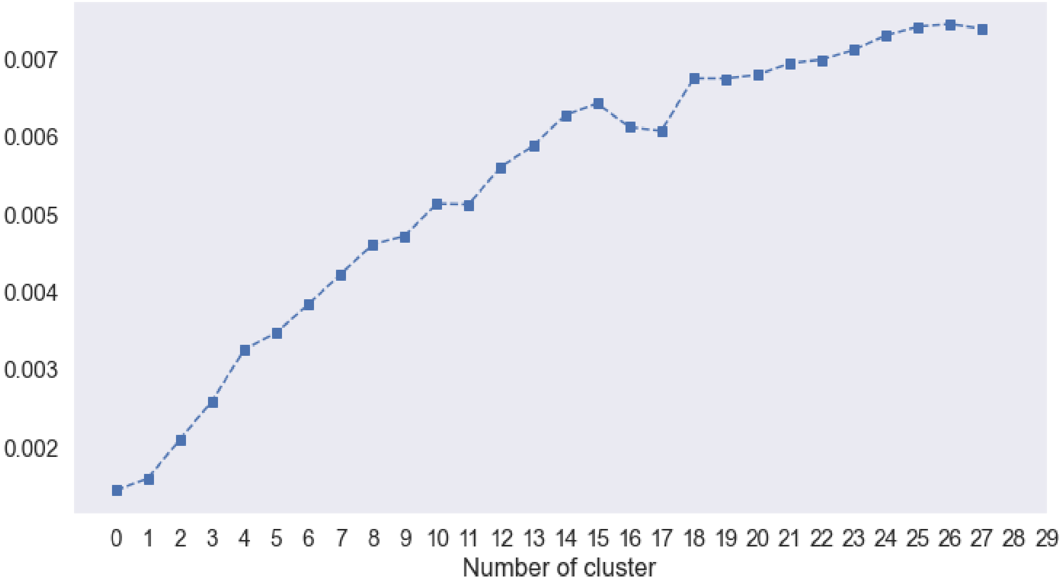
Elbow Method: Determining optimal value for k

- The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.
- Using the Elbow Method we select the optimal number of clusters to be 10.



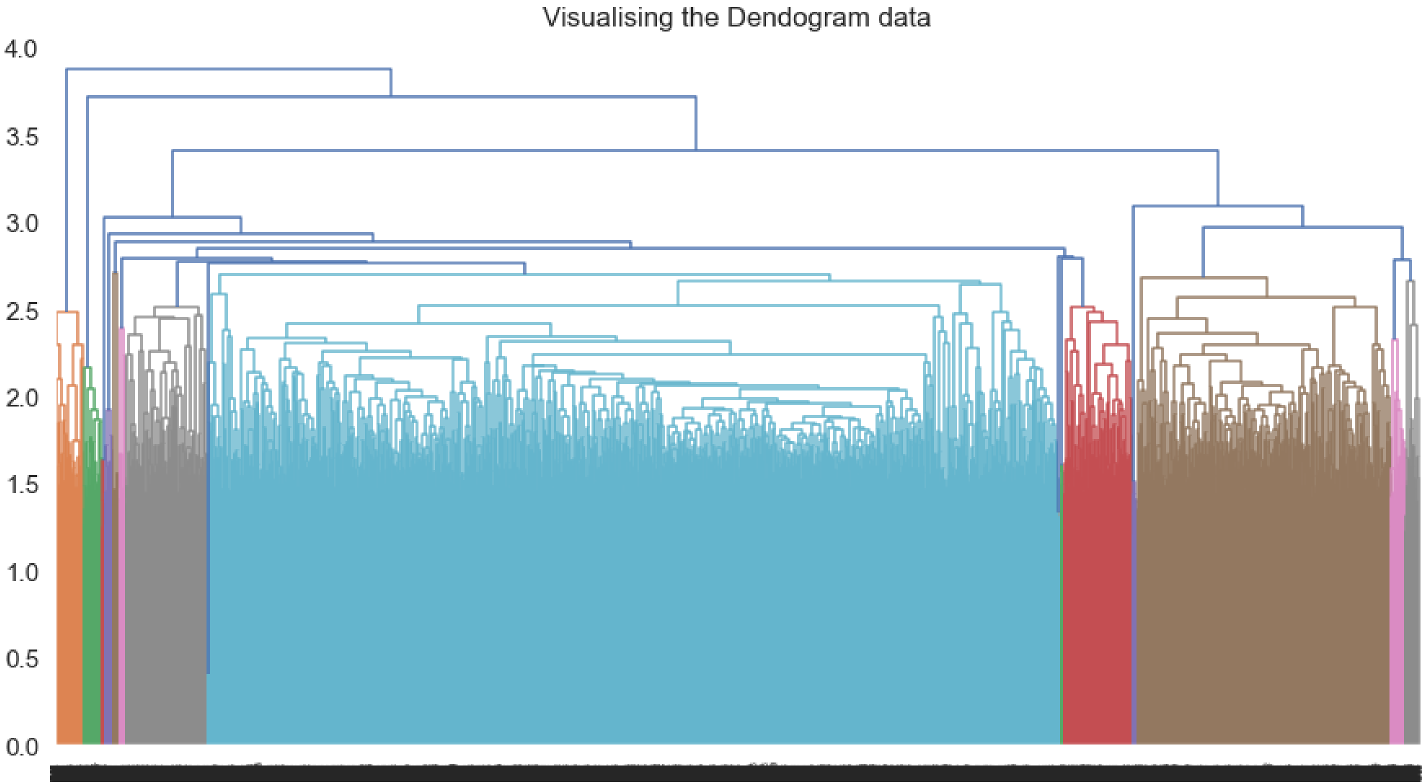
Silhouette Score

- The Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.
- The value of the silhouette coefficient is between [-1, 1]
- If the score is 1 denotes the best meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1
- If the score is 0 denotes overlapping clusters



Dendrogram

- Above Dendrogram The number of clusters will be the number of vertical lines which are being intersected by the line drawn using threshold No. of Cluster



Agglomerative Clustering

Agglomerative clustering is the most well-known kind of variable leveled clustering used to gather in bunches based on their comparability it’s otherwise called agglomerative clustering.

Advantages & disadvantages:

- No need for any information about how many clusters are required And easy to use on another side we can not take a step back in this algorithm and the Time complexity is higher by at least 0(n^2logn).
- It can produce an ordering of objects, which may be informative for the display.
- The time and space complexity of agglomerative clustering is more than K-means clustering, and in some cases, it is prohibitive.

The silhouette score of Agglomerative hierarchical Clustering is -0.002. so its worst score.

❑ Conclusion:

1) Data set contains 7787 rows and 12 columns in that cast and director features contain a large number of missing values so we can fill it and other features like ‘date_added’ and ‘rating’ contain an insignificant portion of the data so we will drop them from the dataset.
2) We have two types of content TV shows and Movies (30.86% contains TV shows and 69.14% contains Movies), there are more number movies on Netflix than TV shows.
3) TV-MA has the highest number of ratings for tv shows, i.e. adult ratings
4) The most number of movies and TV shows released in 2017 & 2018 or 2020 respectively.
5) United States has the highest number of content on Netflix, followed by India and India has the highest number of movies on Netflix.
6) The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a huge increase in the number of movies and television episodes after 2015 n our datasets.
7) Kids tv is the top TV show genre on Netflix.
8) Most of the movies have a duration of between 50 to 150 minutes long.
9) The most content is added to Netflix from October to January.
10) Documentaries are the top most genre on Netflix which is followed by standup comedy, Dramas, and international movies.
11) When it comes to movies having a TV-Y rating, they have the shortest runtime on average.
12) By applying the elbow and silhouette score, the optimal of 10 clusters formed, K Means is best for identification than Hierarchical as the evaluation metrics also indicate the same.

Thank you!