**Netflix Movies & TV Shows Clustering**  
**Harshjyot Singh Chawla**

**Ajinkya Dakhale**

**Suvir Kapse**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

In the modern world, OTT platforms are now a common and accepted type of entertainment. Several well-known OTT services, such Amazon Prime, VOOT Netflix, Discovery+, etc., are in competition with one another to attract more users.

Without a question, Netflix has emerged as one of the key streaming platforms since the emergence of these services... Netflix oversees a vast library of TV series and films that are available for watching at any time online.

Due to the monthly fees that customers pay to access the platform, this business is profitable. Customers, however, are free to stop their memberships at any moment. As a result, the business needs to maintain users' interest in the platform. The diversity of content and proper user recommendations is two factors that determine the success of OTT platforms.

A third-party Netflix search tool called Flixable gathered the information that we utilized for EDA and clustering. The dataset contains approximately 7700 observations and 12 features, the majority of which are textual features.

We discovered patterns via exploratory data analysis, based on categorical parameters like rating, kind, genres, actors, directors, etc., that would aid in understanding what content is consumed across countries. The next step is to highlight our objective variable, conduct NLP operations on it, and then vectorize the variable by using TFIDF.From that point on, all that remained was to locate the clusters, fit our models using various clusters, and then construct a small-scale recommendation system from it.

Additionally, the model is evaluated using the metrics.

***Keywords—Machine Learning, Explanatory Data Analysis, Netflix, TV Shows, Movies, Genre, Clustering, K Means.***

**Introduction:**

Both a streaming service and a production enterprise, Netflix charges subscribers.

One of the most sought-after Media companies for a variety of entertainment content to watch that includes diverse themes from different countries is Netflix, which is very well-liked across the globe. As they help improve sales, provide a broad selection, boost customer happiness and loyalty to the brand, and are highly helpful in learning more about what customers want, Netflix's suggestion system helps them become more well-liked among service providers. Then, it will be simpler to influence the customer to choose wisely from the many movie items available.

Now that everything cannot be displayed on the landing page, viewers cannot be recommended all TV shows or movies. We must consider the preferences of the client and suggest programmes and films that s/he could enjoy. Understanding the measures that will aid in distinguishing and clustering them is necessary in order to accomplish that.

Building a clustering algorithm that can group related TV shows and films is the aim of this project. The platform's recommender system can be created using these clusters.

**Problem Statement:**

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Flixable 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.

**In this project, you are required to do**

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

**Attribute Information:**

The dataset contains following columns:

* 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 content
* 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

**Steps involved:**

**1.Initial stage:**

Immediately after importing the dataset, we used some standard head(), tail(), and describe() procedures to check the meaningful information about the data.

**2.Missing or Null value and treatment:**

Missing values can occur in datasets for a variety of reasons, including errors or data handling mistakes.

We looked for any null values in our data, and we found that there are several in the dataset.

● There are 2389 null values in Director Column

● There are 718 null values in cast column

● There are 507 null values in country column

● There are 10 null values in date added column

● There are 7 null values in rating column

To handle the null values, we did some imputations in director, cast and country columns. We replace the null values with the word 'unknown' for further analysis and drop date\_added','rating' column because of minimum null value.

**3.Duplicate Values Treatment:**

Duplicate values do nothing to improve the accuracy of the outcomes. Our dataset doesn't contain any values that are duplicates.

**4.Exploratory Data Analysis:**

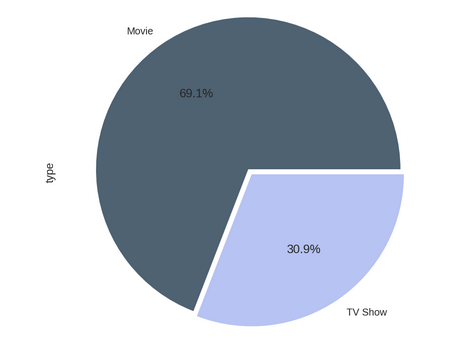
Exploratory Data Analysis (EDA), as the name implies, uses data visualisation techniques to evaluate, research, and summarise dataset’s key features. It makes it simpler for data analysts to find patterns, identify anomalies, test hypotheses, or verify assumptions by determining how to modify data sources to achieve the answers they need.

Its nature and specifics, such as which nations view more material and what kind of content is in demand, etc., have been studied in this step to gain an understanding of the data and how the content is dispersed in the dataset.

Understanding the connections between both the variables (if any) is also beneficial for feature engineering. Before making any judgments, it is helpful to thoroughly comprehend the data in order to spot obvious mistakes, better grasp patterns within the data, identify outliers and unusual occurrences, and discover intriguing relationships between the variables.

**Explorations and visualizations are as follows:**

* **Proportion of type of content**



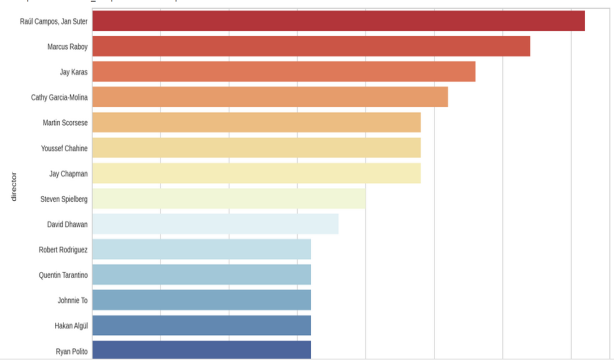
### ****There is more number of Movies than TV shows on Netflix. The percentages are 69.1% for Movies and 30.9% for TV shows****

* **Most occurred word in title**

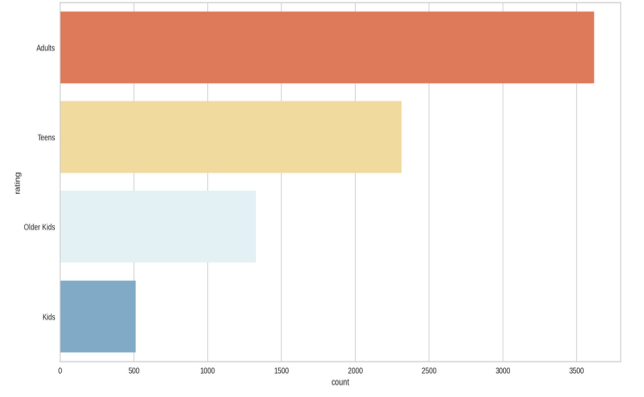


### Christmas", "love", "world", "man" and "life" are few of the words which are used frequently in title of movies and TV shows.

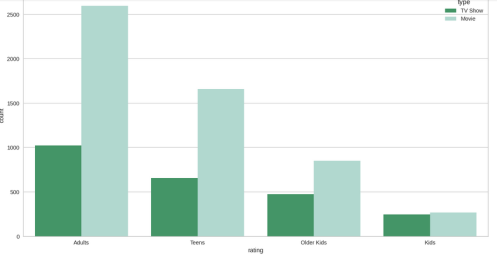
* **Count of movie and TV show by each director**.



* **Count of content appropriate for different ages**



* **Proportion of movies and TV shows content appropriate for different ages**



These are few types of EDA performed. We had a deep analysis on all the features such as Country-wise count of content, Top 10 countries with highest content production, Proportion of TV shows and movies in top 10 countries with maximum content, Rating-wise content count, Count of content appropriate for different ages, , year wise analysis, Proportion of movies and TV shows content appropriate for different ages, Trend of year-wise content release, Trend of year-wise content on-boarded on Netflix, Relation between month and content on boarding on Netflix,.

# ****5.Objective Analysis - Understanding what type content is available in different countries:****

# C:\Users\admin\Desktop\Capture.PNG

**●**Most of the countries have

  more movies than TV

  Shows.

**●**The United States is a

  leading producer of both

  types of content; this makes

  sense since Netflix is a US

company.

**●**The influence of

Bollywood in India explains

the type of content available,

and perhaps the main focus

of this industry is Movies

and not TV Shows.

**●**For South Korea and

 Japan it's the opposite. It

Maybe because K-Dramas

And Anime are more

Popular in these two

countries respectively.

# ****6.Objective Analysis-Is Netflix has increasingly focusing on TV rather than movies in recent years?****

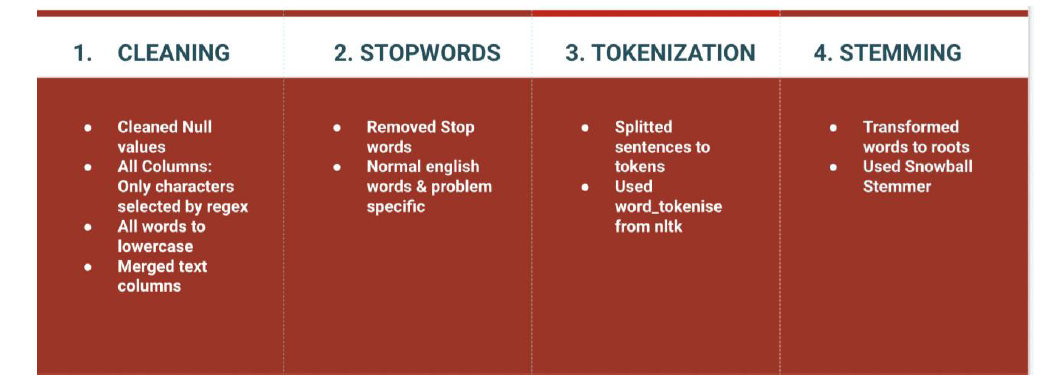
# C:\Users\admin\Desktop\Capture.PNG

The graph shows that, Till 2019 we can see there is gradual rise in both  movies and TV show but in 2020 we can see that is a bit drop in movies whereas the growth of TV shows remains the same.**So yes we can say  Netflix has increasingly focusing on TV rather than movies in recent years**

**7.Data Pre-processing / Feature Engineering:**

Three textual columns were combined, integrated, and then used as one final characteristic for clustering.

Text-based aspects like the description and type of the TV show or movie were what I was expected to work on. In this case, we cleaned the data by removing punctuation, stop words, and stems from the text.



The remaining data was used to validate the models and examine the cluster formations, while the text data was used to train the clustering model.

We then vectorized the textual information after this method.

**8.Tfidf vectorization:**

Term Frequency Inverse Document Frequency is referred to as TF-IDF. It is quite common to utilise this approach to convert text into meaningful numerical representations that can be used to fit machine learning algorithms for prediction.

We also used PCA since it allows us to enhance efficiency at a very little cost to model correctness. The capacity to extract distinct, statistically independent features from the information and the decrease of random noise are other advantages of PCA.

Therefore, in order to fit our text into our model, we must first transform it into a tfidf vectorizer and then into an array.

**9.Clustering**

We went on to determine the ideal number of clusters needed for the algorithm once we had some guidance with the data.

Five clusters should be the ideal number of clusters, according to our analysis using the Elbow approach and the Silhouette score.

It was finally time to build the clusters once I got the ideal amount of clusters, so I fitted the KMeans model to the data.

**10.Fitting into model**

We have used a K means clustering technique in this work. Unsupervised machine learning can make use of the data clustering method K-means. It can group unlabeled data based on similarities into a preset number of clusters (k).

**Algorithms:**

**K Means Clustering:**

One of the most straightforward and well-liked unsupervised machine learning techniques is K-means clustering.

The goal of k-means clustering, a vector quantization technique that originated in signal processing, is to divide n observations into k clusters, where each study gathered to the cluster that has the closest mean (also known as the cluster centroid or cluster centre), which serves as a prototype for the cluster.

**K-means algorithm works:**

The K-means technique in data mining uses a first set of centroids that are randomly chosen as the starting points for each cluster to process the learning data. The programme then conducts repeated (repetitive) calculations to optimise the positions of the centroids.

When one of the following occurs:

• The cluster centres have stabilised — there is no movement in their values as a result of successful clustering;

• The specified amount of iterations has indeed been reached.

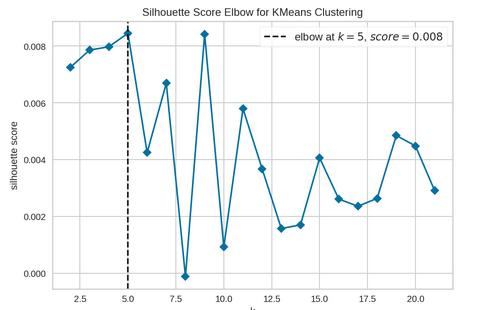
**Model performance:**

**Silhouette score:**

The silhouette score measures how effectively samples were grouped with other examples that are comparable to them in order to evaluate the quality of groups produced by clustering algorithms like K-Means. Each sample of various clusters receives a Silhouette score. The following distances must be determined for each observation belonging to each cluster in order to determine the Silhouette score for that observation/data point.

**2. Elbow Curve:**

One of the most widely used techniques for figuring out this ideal value of k is the elbow curve.

The elbow curve determines the optimal value of k depending on the distance between the data points and the clusters to which they are assigned using the sum of squared distance (SSE).

# Conclusion:

* Till 2019 we can see there is gradual rise in both  movies and TV show but in 2020 we can see that is a bit drop in movies whereas the growth of TV shows remains the same.So yes we can say  Netflix has increasingly focusing on TV rather than movies in recent years
* The United States and India are the two nations that generate the most content.
* The three most popular genres on Netflix are international movies, drama, and comedy. Further research revealed that Netflix had more movies than TV shows.
* The majority of TV shows terminate after the third season, while most Netflix movies last between 75 and 120 minutes.
* In order to build clusters, we employed K-Means Clustering. To determine the ideal number of clusters, we used the elbow curve and the silhouette score.
* The KMEANS clustering technique has been used, and we have defined 5 clusters. Then we discovered that cluster number three has the most clusters.

**Reference:**

[1] Applied Science Article MDPI

[2] GeeksforGeeks

[3] Wikipedia

[4] Data Camp