Netflix Movies And TV Shows Clustering

Simran Dapke

Data science trainee, AlmaBetter

Abstract:

Netflix The firm was founded in 1999 and is now recognized as one of the largest international firms in this field. It has over 6.3 million subscribers and over 100,000 DVD titles in order for the customer to choose from. Netflix has maintained a competitive strategy by having a business model based upon fast delivery, no late-fees policy and a very useful return in the mail system.

OTT platform success depend on content available and recommendation to users profit of this business because of subscription fee.

EDA has to be done to know insights from data with business perspective. Clustering use for recommendation and also increase user viewership of OTT platforms by applying different algo.

Keywords: OTT, Clustering, EDA, Algo

1. Problem Statement

Media revenue from OTT is expected to surpass $210 billion by 2026. This is nearly double the amount of revenue in 2020, which was $106 billion. Netflix users worldwide watched 452,000 hours of content in 2021.

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset 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..

2. Introduction

Netflix,Inc. American subscription streamig service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions, known as Originals. It is the second largest entertainment/media company by market capitalization.

Unsupervised learning is a machine learning technique, where you do not need to supervise the model. It is machine learning technique in which model trained on unlabelled data. Clustering can done using eucledean distance, gomer distance.We do use different kind of cluster based on data pattern K-means clustering etc.

2.1 Netflix Dataset

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.

2.3 Python

Most of the info scientists use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is a simple programing language for select compared other languages. The most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is easy to use. That's one among the most reasons to use python. Specifically, for data scientists, the foremost popular data inbuilt open-source library is named panda. As we've seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, and 3-dimensional data python built-in library comes very helpful choose to wait a few minutes to see if the rates go back down.

3. Steps involved:

Exploratory Data Analysis

- **Data Exploration**
  After loading the dataset we started exploring about the data what we have at first I explored that our dataset is of shape 7787 rows and 12 columns. After knowing the shape we saw the datatypes of our the field and some info about our data where I got to know about null values which I need to treat.

- **Null values Treatment / Duplicate values Treatment**
  there are null values present in director, cast, country, date_added and rating. Around 30% of values as null in director so instead of dropping we can fill those values with 'No Director' same for cast and country. We can drop null values from date_added and rating as there are not much null values.

- **Data Visualisations**
  Data visualization is the representation of data through use of common graphics, such as charts, plots, info graphics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

- **Fitting into different models**
  For modelling, we tried various algorithms like:
  - K-Means Clustering
  - Principal component analysis
  - Hierarchical clustering
  - DBSCAN

3.1 Feature Description

The dataset contains following columns:

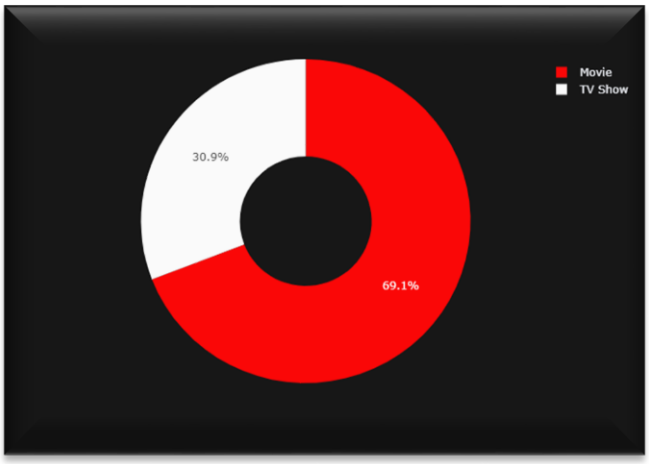
1. Show id: Unique ID for every Movie / TV Show
2. type – Identifier - A Movie or TV Show
3. title – Title of the Movie / TV Show
4. director-director of the content
5. cast –Actors involved in the movie / show
6. country – Country where the movie / show was produced
7. date_added – Date it was added on Netflix
8. release_year – Actual Release year of the movie / show
9. rating – TV Rating of the movie / show
10. duration – Total Duration - in minutes or number of seasons
11. listed_in – genre
12. description – The Summary description

3.2 EDA

Exploratory data analysis (EDA) plays vital role in analysis of data and gives idea of feature engineering. EDA help us to determine dependent and independent variables.

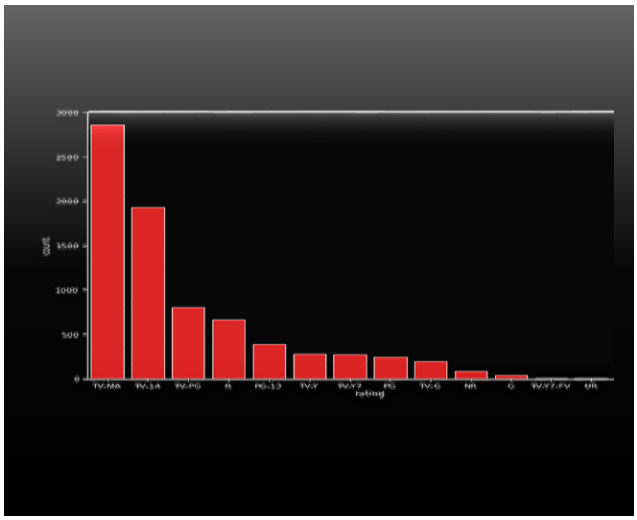
3.2.1 Distribution of Movies & TV shows

Pie chart clears pictures about distribution. Netflix has 69% of its content as movies. Movies are clearly more popular on Netflix than TV shows.



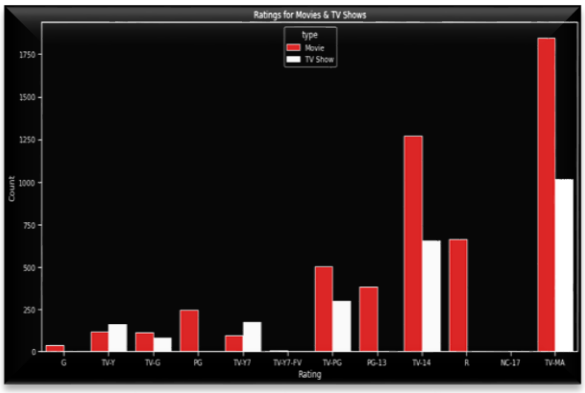
3.2.2 Countplot for Ratings

TV-MA is the most given rating then TV-14.That means most of the shows are for adults



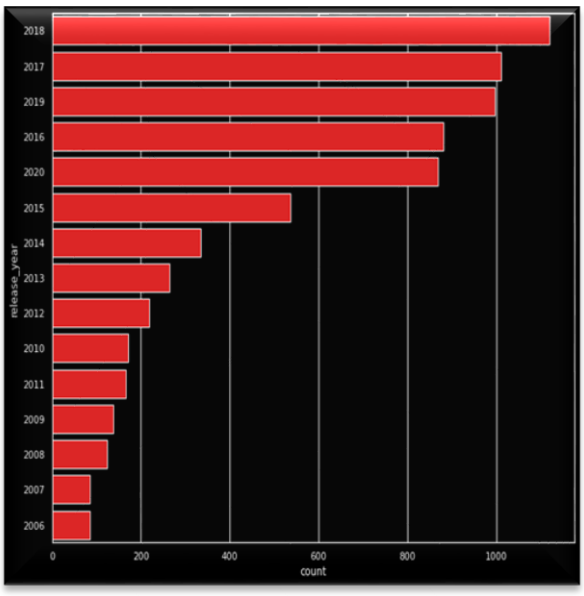
3.2.3 Countplot for Rating wrt Type

Again TV-MA and TV-14 are the most rated shows in that as well movies are having greater number of these rating than TV Shows

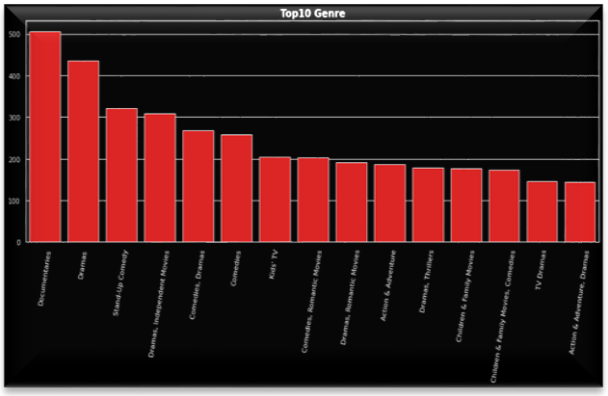


3.2.4 Year wise distribution

We can see after 2014 there is growth in the amount of content added.

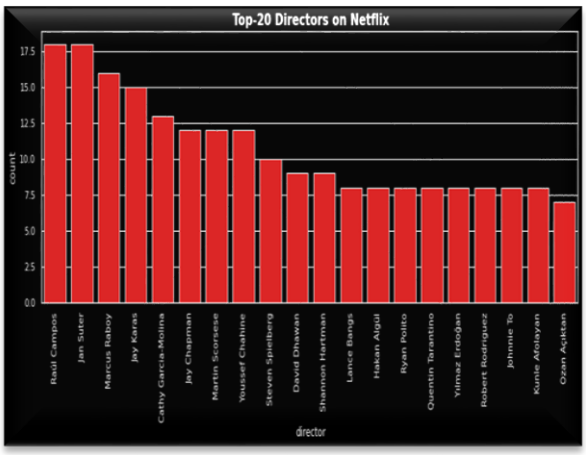


3.2.5 Top 10 Genres



Documentaries and Dramas are the most watched Genres.

3.2.6 Top 20 actors

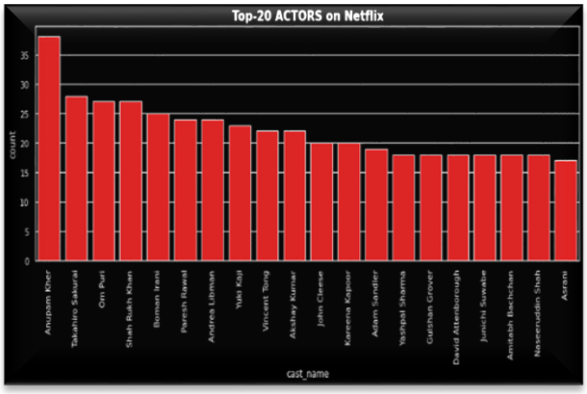


We can see Anupam Kher, Takahiro Sakurai, Om Puri, Shah Rukh Khan, Boman Irani are among the top 5 actors worked on shows/movies on Netflix.

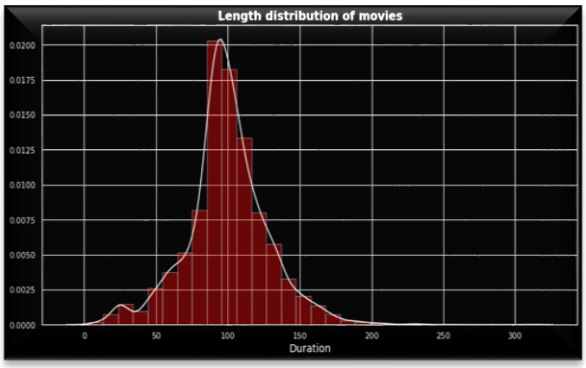
Majority of Netflix movies are having Indian actors.In this list, we can see that the most popular actors on Netflix based on the number of titles are international as well

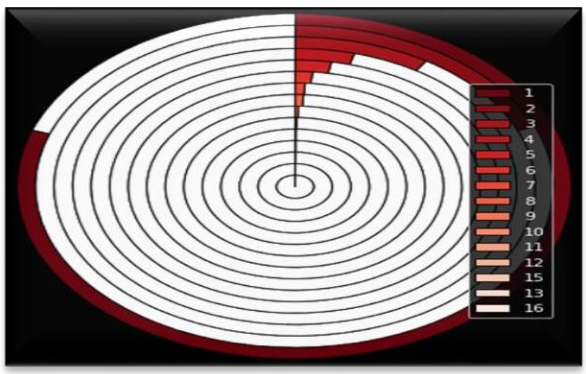
3.2.7 Top 20 directors

Here we can see the Raul Campos, Jan Suter, Marcus Raboy, Jay Karas, Cathy Garcia-Molina are the top 5 directors.



3.2.8 Netflix movies duration

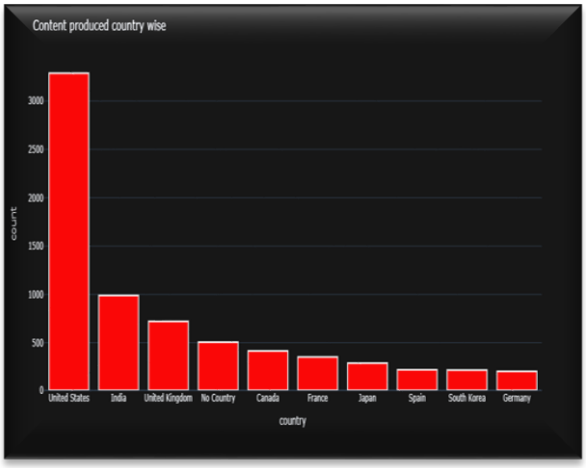




Most content are about 70 to 120 min duration for movies

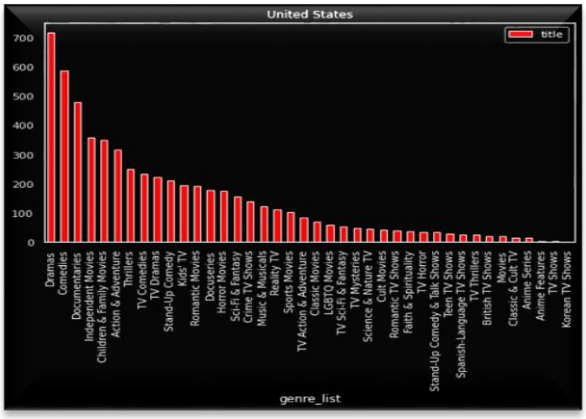
Most of the shows are 1 to 2 seasons long.

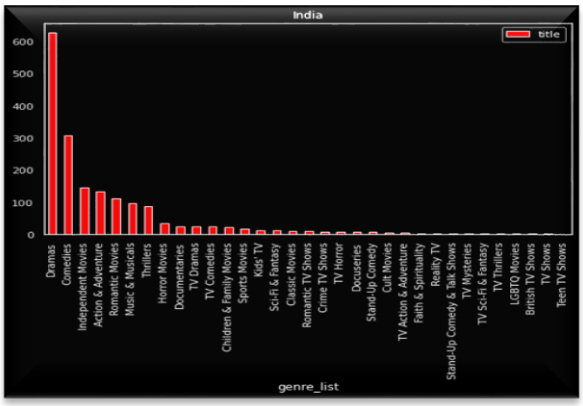
3.2.9 Content produce country wise

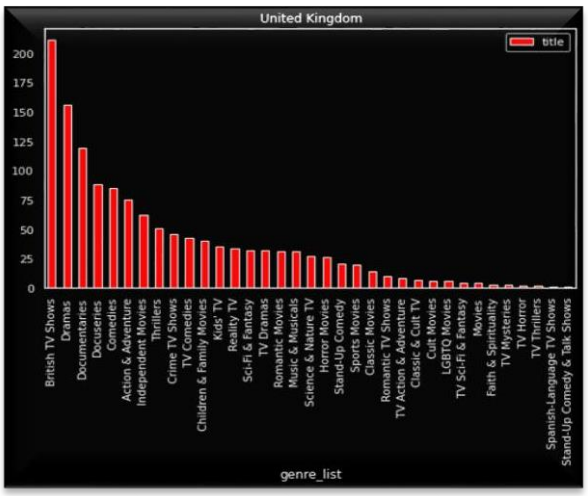


United States and India has the most number of content.

3.2.10 Country VS Listed_in

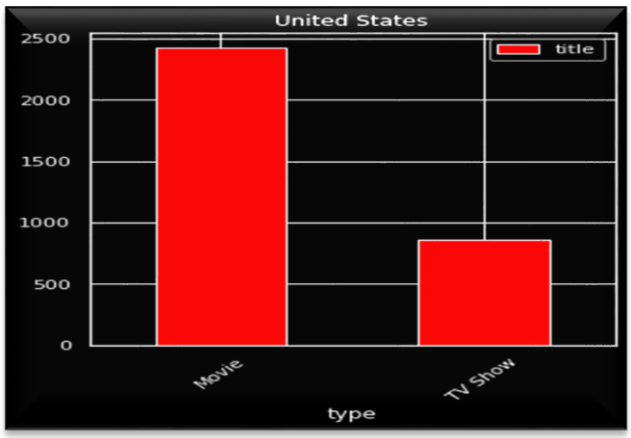


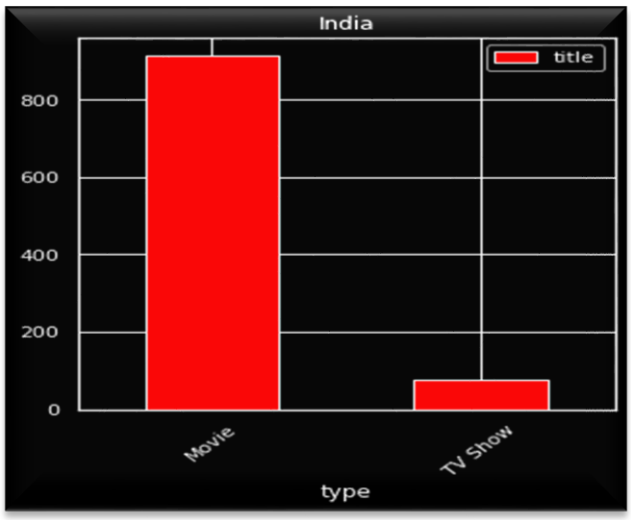


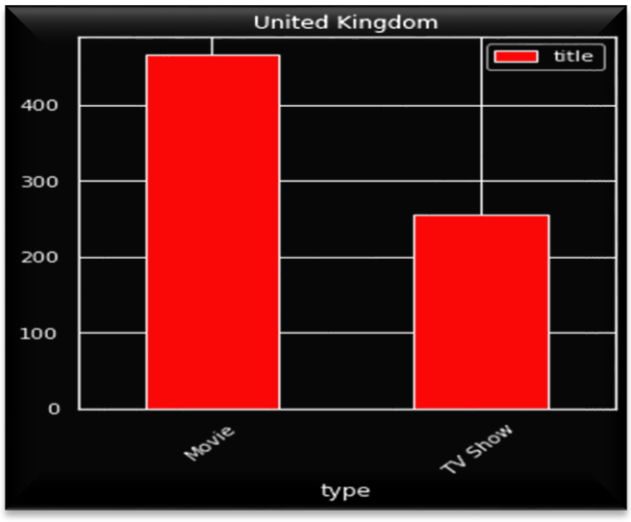


In every country we can see that Dramas are most popular genres then documentaries and Action and Thriller.

3.2.11 Country VS Type

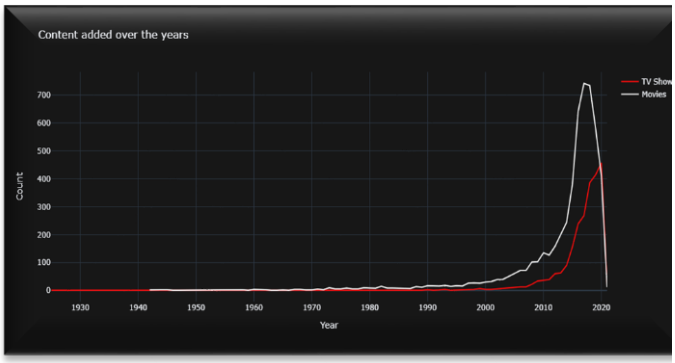






We can see that every country movies are shot more than TV Shows.

3.2.12 Netlix focusing more on tv or movies in recent years



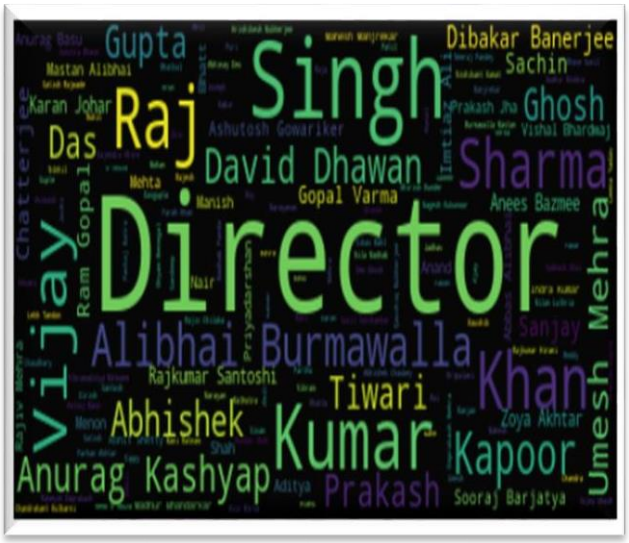
From the above plot we can see that the number of movies added to netflix is higher than that of TV shows. In 2018, netflix added 734 movies and 386 TV shows. So there we cannot conclude that netflix has switched focus from movies to TV shows.

4. WordClouds

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.Significant textual data points can be highlighted using a word cloud.

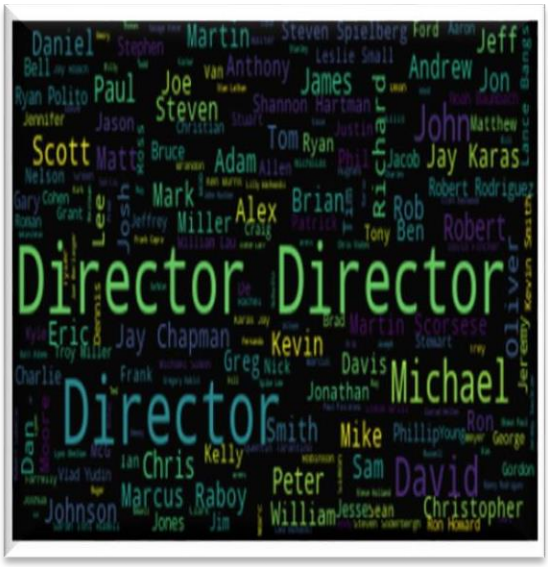
4.1 Genres

4.2 country vs director



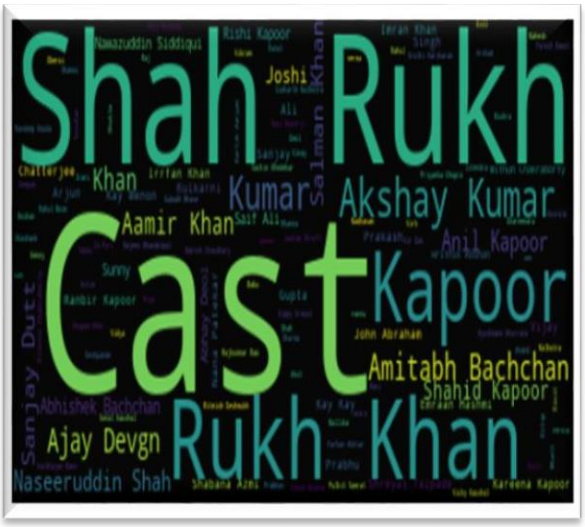
For India we can see Anurag Kashyab, Dhavid Dhawan, Gopal Varma, Dibakar Banerjee and many more.

For UK we can see Patricia, Alexis, Daniel, Calvo and many more.

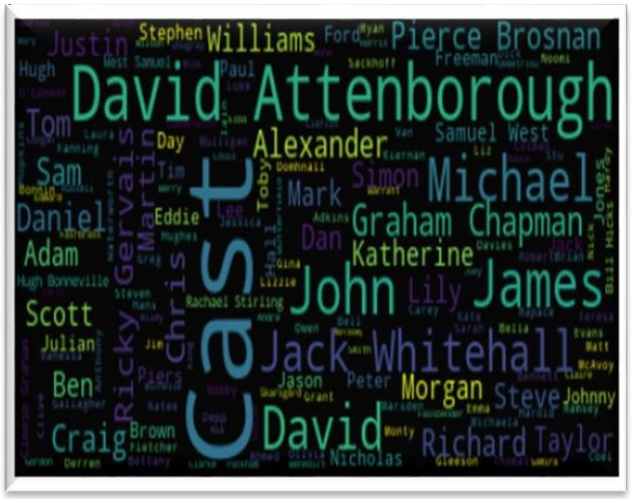


For US we can see David, Marcus, Scott, Adam, Johnson and many more.

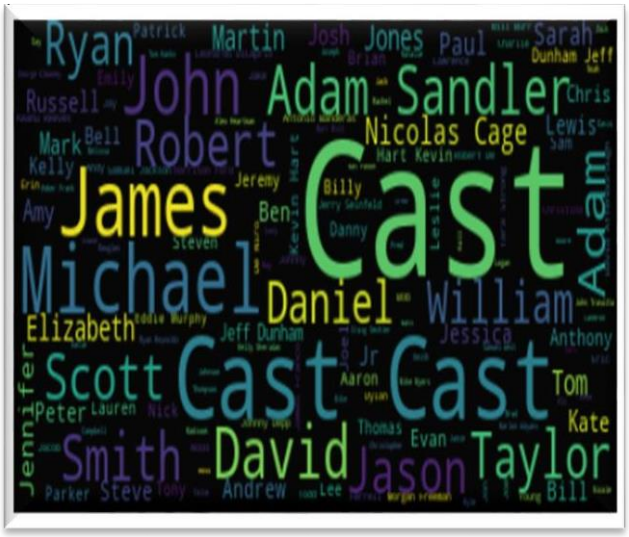
4.3 country vs cast



In India we can see Amitabh Bachchan, Shah Rukh Khan, Naseeruddin Shah, Kareena Kapoor, Anupam Kher and many more.



In Spain we can see Maria, Fernandez, Carlo, Ana and many more.



In US we can see Michael, Tom Adam, Scott, William, James and many more.

5.1 Data Pre-processing

- **Removing Punctuation**:
  Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data, or noise.
- **Removing stop-words** :
  Stop-words 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.
- **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.

- **Countvectorizer**
  Machines cannot understand characters and words. So when dealing with text data we need to represent it in numbers to be understood by the machine. Countvectorizer is a method to convert text to numerical data. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.

- **TF-IDF vectorizer**

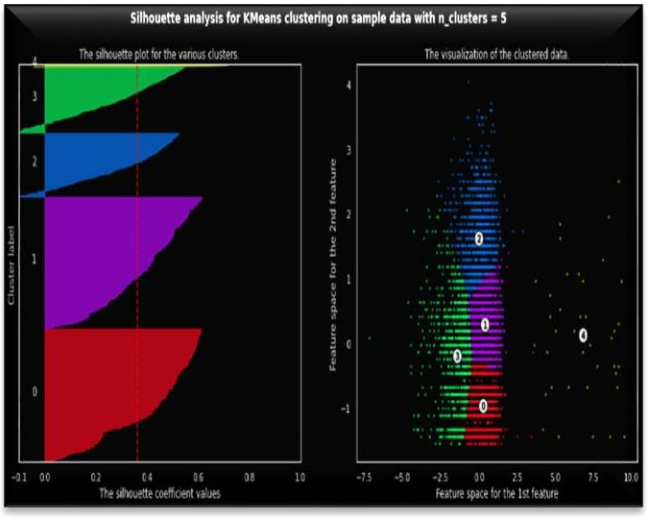
  TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.

5.3 Methods to find K value

1. Silhouette score :

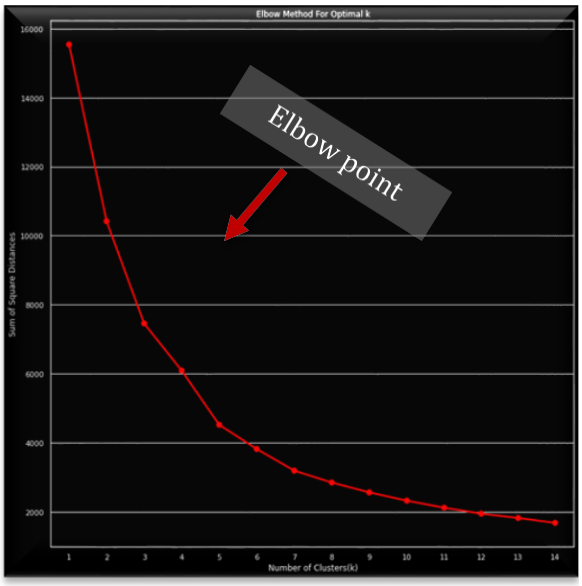
Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.

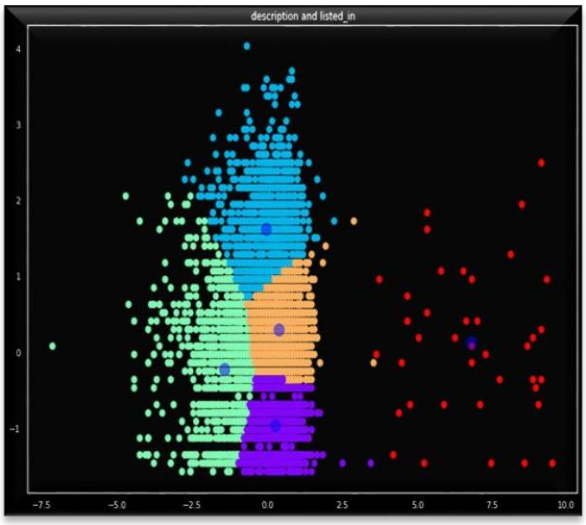
```
For n_clusters = 2, silhouette score is 0.34951465464796905
For n_clusters = 3, silhouette score is 0.37281514079850314
For n_clusters = 4, silhouette score is 0.3862368715405086
For n_clusters = 5, silhouette score is 0.3640633581237277
For n_clusters = 6, silhouette score is 0.34912075022261346
For n_clusters = 7, silhouette score is 0.356956590822019
For n_clusters = 8, silhouette score is 0.33799101630477646
For n_clusters = 9, silhouette score is 0.3371975698072569
For n_clusters = 10, silhouette score is 0.32935664062434744
For n_clusters = 11, silhouette score is 0.3307768469937211
For n_clusters = 12, silhouette score is 0.3364989158651138
For n_clusters = 13, silhouette score is 0.32976423299150764
For n_clusters = 14, silhouette score is 0.3352545921599981
For n_clusters = 15, silhouette score is 0.33762305481596533
```



2. Elbow curve :

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.





4) DBSCAN

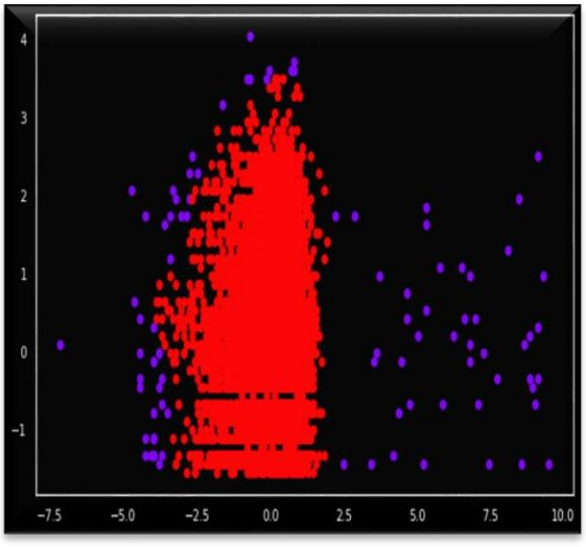
DBSCAN is the acronym for Density-Based Spatial Clustering of Applications with Noise. DBSCAN is extensively used for the identification of clusters in massive spatial populations or datasets by taking the local densities and properties of the data points into account.DBSCAN is especially useful for working with outliers or anomalies and correctly detecting these outliers and points that stand out.

5.2 Algorithms

Its time to apply different models on given dataset as follows.

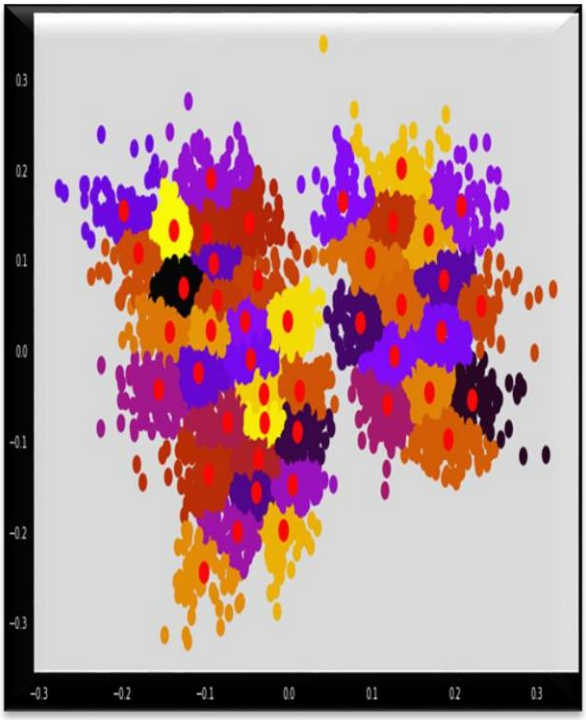
1) K- means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.



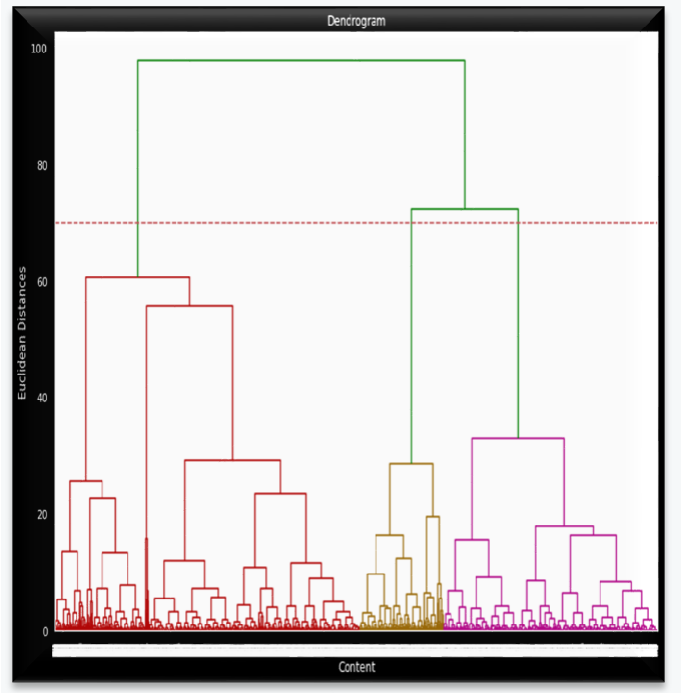
2) Principal component analysis

Principal Component Analysis is an unsupervised learning algorithm that is used for dimensionality reduction in machine learning. It is a statistical process that transforms the observations of correlated features into a collection of linearly uncorrelated features with the support of orthogonal data. These new transformed features are known as the Principal Components.



3) Hierarchical clustering

Hierarchical clustering is also known as Hierarchical Cluster Analysis (HCA) is unsupervised Machine Learning. It groups unlabeled data sets into groups also Known as clusters. In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the Dendogram .

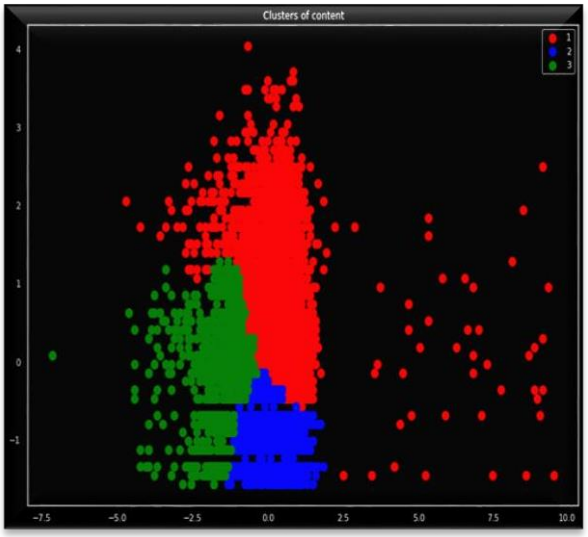


Dendogram

The root of the tree (usually the upper or left element) is one large cluster cluster that contains all data points. The leaves (bottom or right elements) are tiny clusters, each of which contains only one data point. According to the generated dendrogram, you can choose the desired separation into any number of clusters

Agglomerative clustering

Agglomerative Clustering Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to pre specify the number of clusters.



6. Recommendation System

```
get_recommendations('Dear Zindagi')

5197                Ricardo Quevedo: Hay gente así
497                      An Unremarkable Christmas
3795                               Loving is Losing
5595    Si saben cómo me pongo ¿pá qué me invitan?
4853                                   Pickpockets
4818                                  Penalty Kick
5386                                  Santo Cachón
2147                               Feo pero sabroso
375          Alejandro Riaño: Especial de stand up
1740                             Dhoondte Reh Jaoge
Name: title, dtype: object
```

A recommender system, or a recommendation system, is a subclass of information filtering system that seeks to predict the “rating” or “preference” a user would give to an item. They are primarily used in commercial applications.The famous The Netflix Prize is also a competition in the context of recommendation systems.

7. Conclusion

1. Netflix has 69% of its content as movies, so movies are more popular on Netflix than TV shows.
2. United States has the most number of movies and shows followed by India and United Kingdom.
3. TV-MA rated content is maximum in number in the dataset. This rating indicates that the content is for mature and adult audience above the age of 17.
4. There is an exponential raise in the number of TV shows and movies distributed by Netflix in the recent years.
5. Text cleaning and vectorization was done on the combined features of the dataset which includes origin country, leading cast member, rating type, content type and description for clustering analysis.
6. Optimal number of clusters were found out to be 25 with silhouette coefficient value of 0.0279

7. Principal component analysis was performed in order to reduce the higher dimensionality which improved the silhouette coefficient to 0.35. Even though there's improvement in the silhouette score, these cannot be compared as these are two different method of pre processing is involved.
8. Recommendation based on cosine similarity is also done on the same transformed data.

References-

1. Stackoverflow
2. GeeksforGeeks
3. Jovian