

Capstone Project-IV

NETFLIX MOVIES AND TV SHOWS CLUSTERING



JUEGOS

By- Mahima Phalkey
Sameer Satpute

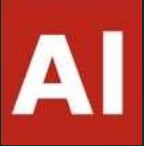
CONTENTS:



1. Agenda
2. Introduction
3. Problem Statement
4. Data Summary
5. Basic Observations
6. EDA
7. Text Classification
8. Algorithms used
9. Recommendation System
10. Conclusion
11. References

NETFLIX

AGENDA:



- Understanding what type of content is available in different countries
- Is Netflix increasingly focused on TV rather than movies recently.
- Clustering similar content by matching text-based features


INTRODUCTION:



- 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.

PROBLEM STATEMENT:



- 
- This dataset consists of tv shows and movies available on Netflix as of 2019.
 - In 2018, Flixable 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.
 - 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..

DATA SUMMARY:



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 Movie

5. Cast: Actors involved in the movie / show

6. Country: Country where the movie / show was produced

DATA SUMMARY(Contd.):



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: Genres

12. Description: The summary description

BASIC OBSERVATIONS:



- We have a dataset of 7787 rows and 12 columns.
- We need to change the data type of the `date_added` column and we can also see there are some comma-separated values we need to clean those columns.
- We can see there are some null values present in our dataset we will first treat those null values.

EDA (DATA CLEANING):



NULL VALUES TREATMENT:

- There were null values present in director, cast, country, date_added, and rating.
- We have 30% of values as null in a 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 many null values.

DUPLICATE VALUES TREATMENT:

- There were no duplicates present in the dataset.

EDA (DATA PREPROCESSING):



1. **Rating:** It appears like there is nothing wrong with 'rating'. But, to the viewer - ratings like 'TV-MA' or 'PG-13' means nothing. We just know that the rating means "for a specific audience". So, we want to conduct research to understand all the ratings, and then change the text into a more readable, appropriate text

Classifying the 'rating' feature into three categories. (Kids, Teenagers, Adults)

In Kids:

- TV-Y
- TV-Y7
- TV-Y7-FV
- G
- TV-G
- PG
- TV-PG

In Teenagers:

- PG-13
- TV-14

In Adults:

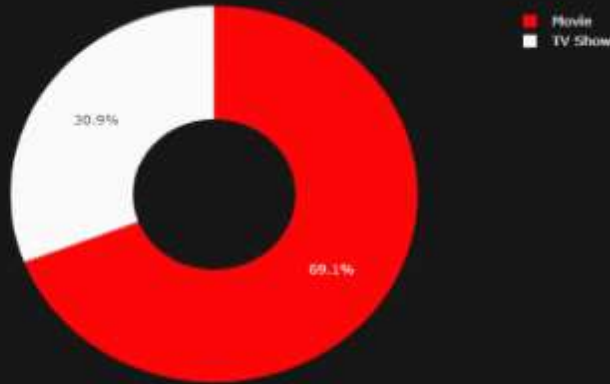
- TV-MA
- R
- NC-17
- NR
- UR

EDA (DATA PREPROCESSING):



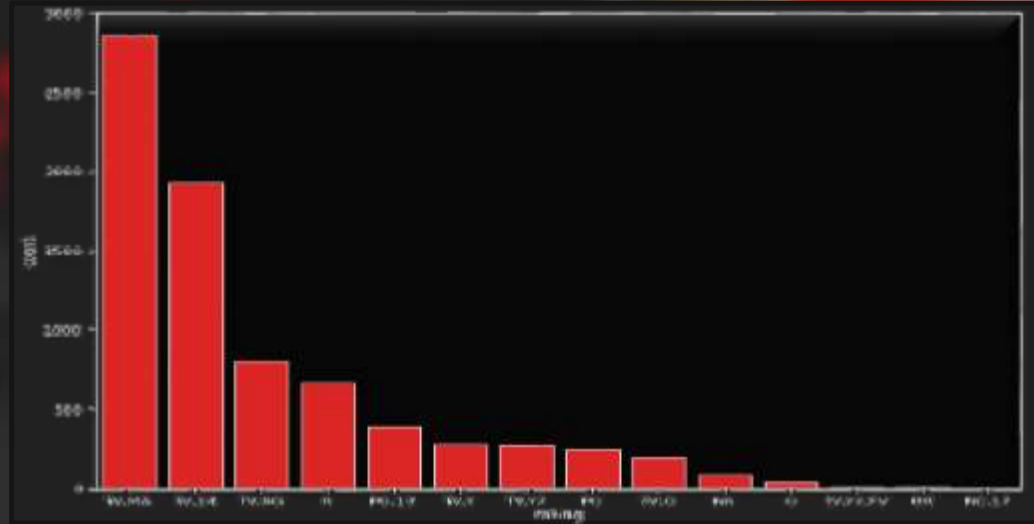
2. **"Listed In"**: We did notice something odd in the listed_in column. Values like "International TV Show" or "International Movie" are not genres. These are types of content. So, let's split all the "International" to a different column, and remove them from "listed_in"
3. **Data type**: Changing the data type of type, ratings_cat to categorical data type and also changed the data type of date_added to DateTime
4. **Handling Comma-Separated Values**: We require all comma-separated values are placed in the correct order/form

EDA (DATA VISUALIZATIONS):



Distribution of Movies & TV shows

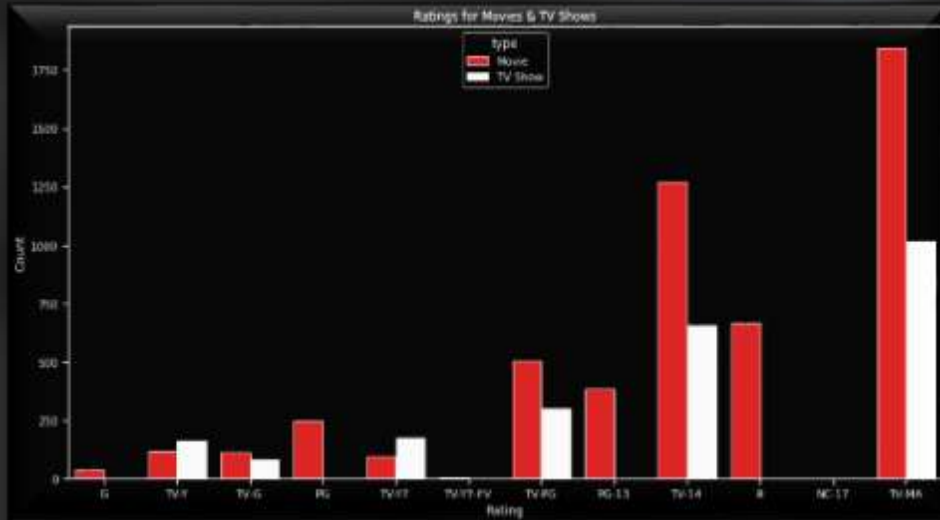
- Netflix has 69% of its content as movies
- Movies are clearly more popular on Netflix than TV shows.



Countplot for Ratings

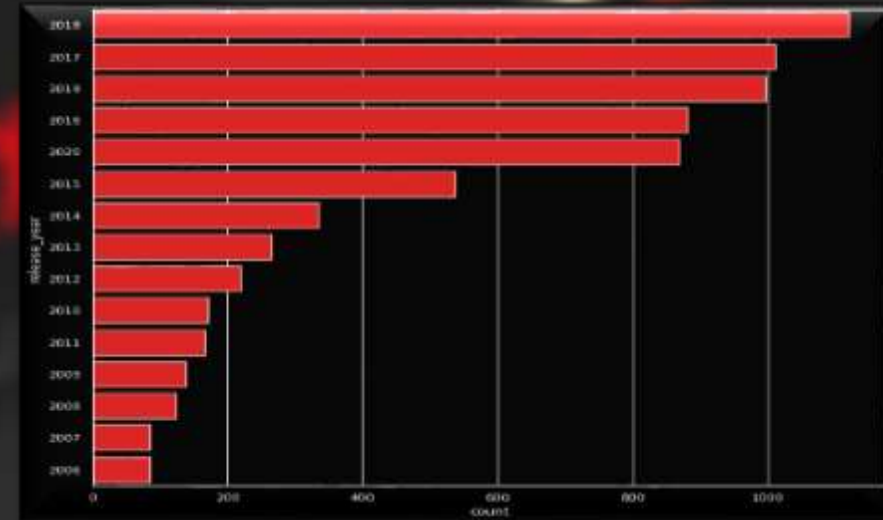
- TV-MA is the most given rating then TV-14.
- That means most of the shows are for adults.

EDA (DATA VISUALIZATIONS):



Countplot for Rating wrt Type

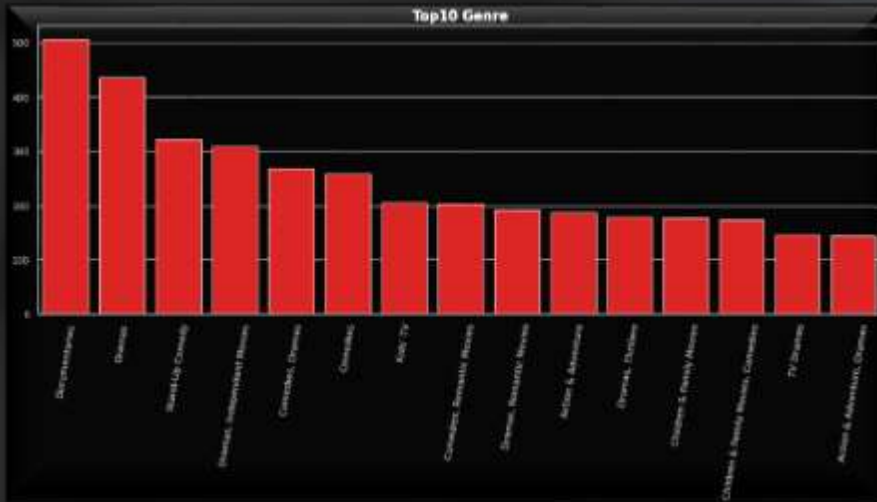
- Again TV-MA and TV-14 are the most rated shows in that as well movies are having greater number of these rating than TV Shows



Count plot for releases over the years

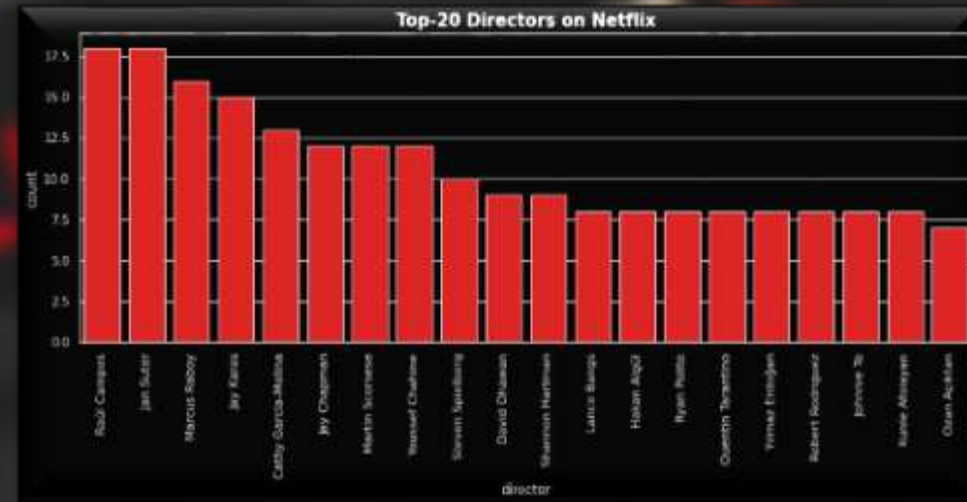
- We can see after 2014 there is growth in the amount of content added.

EDA (DATA VISUALIZATIONS):



Top 10 Genres for Movies/Shows on Netflix

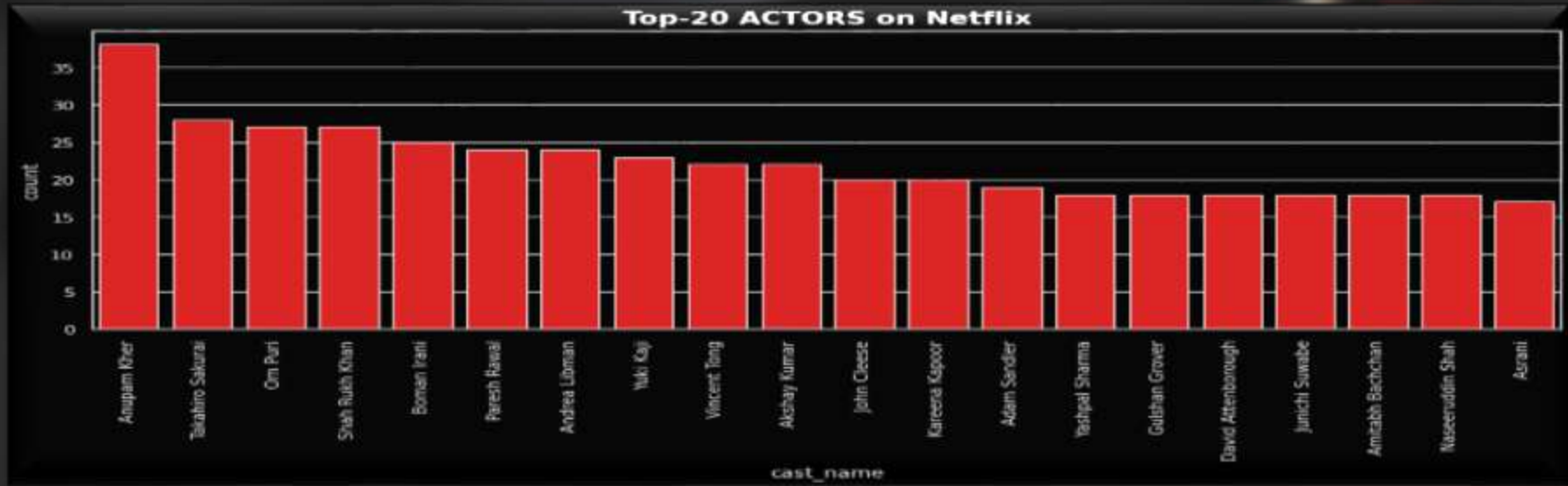
- Documentaries and Dramas are the most watched Genres.



Top 20 Directors on Netflix

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

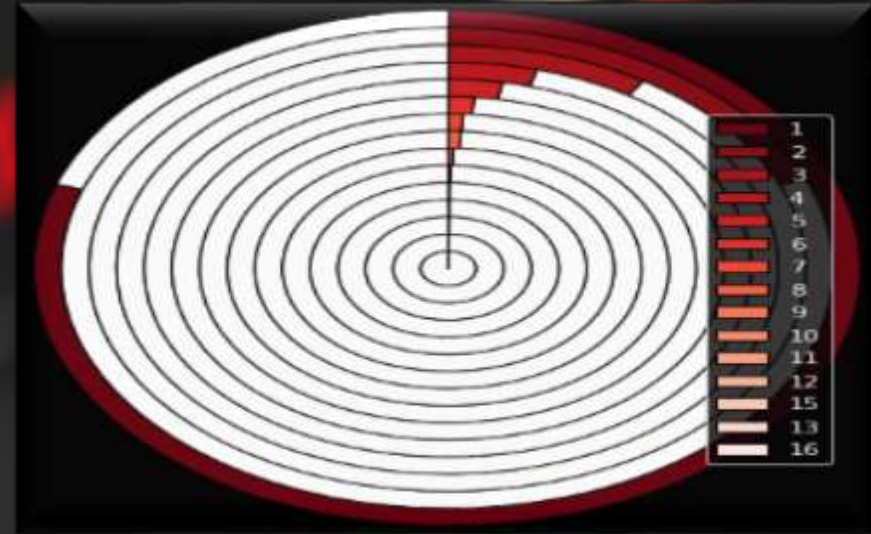
EDA (DATA VISUALIZATIONS):



Top 20 Actors with most number of movies/shows 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

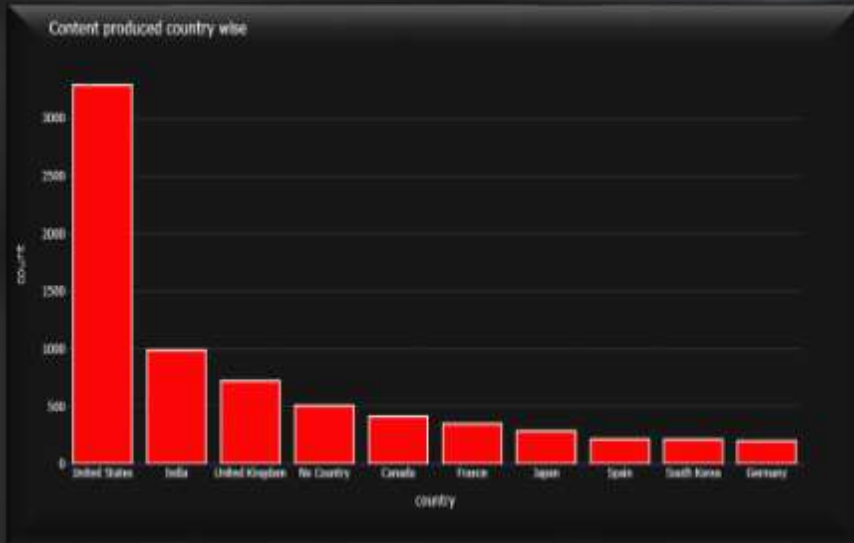
EDA (DATA VISUALIZATIONS):



Netflix Movie Duration Distribution:

- Most contents are about 70 to 120 min duration for movies
- Most of the shows are 1 to 2 seasons long.

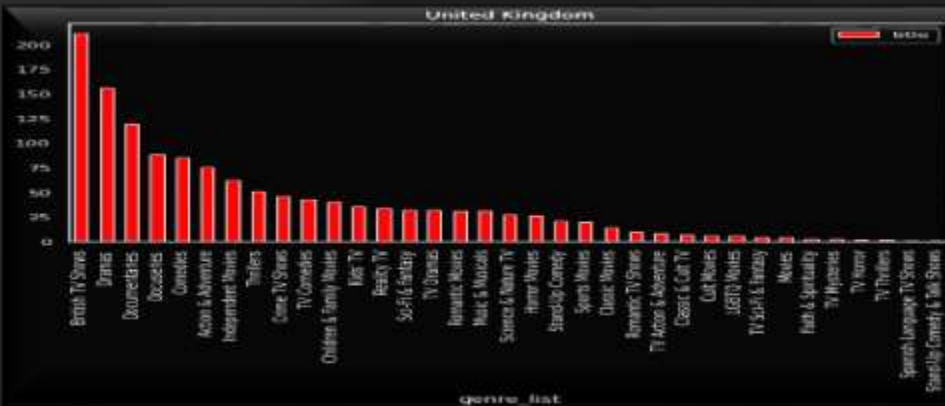
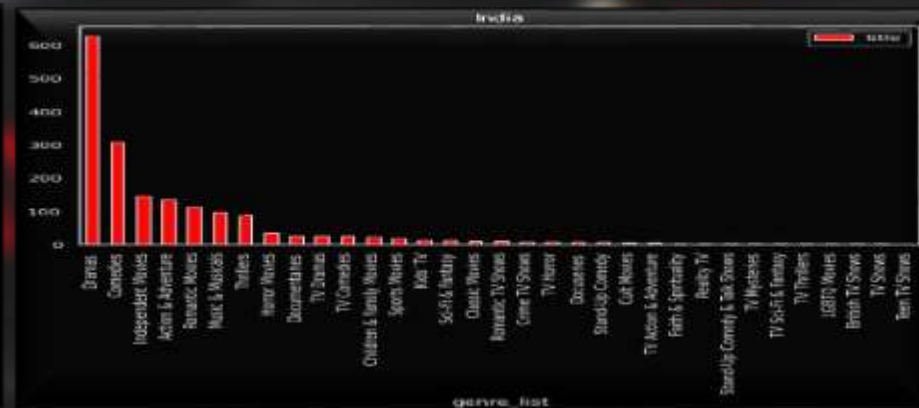
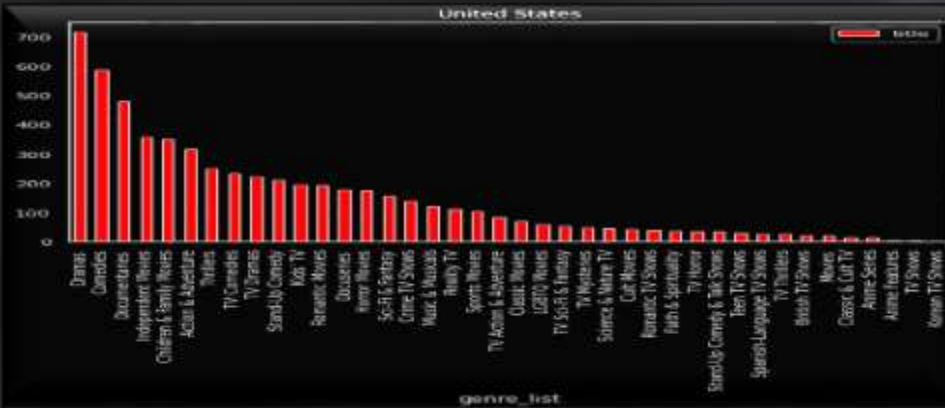
EDA (DATA VISUALIZATIONS):



Country wise Distribution:

- The United States, United Kingdom, and India have the most number of content.

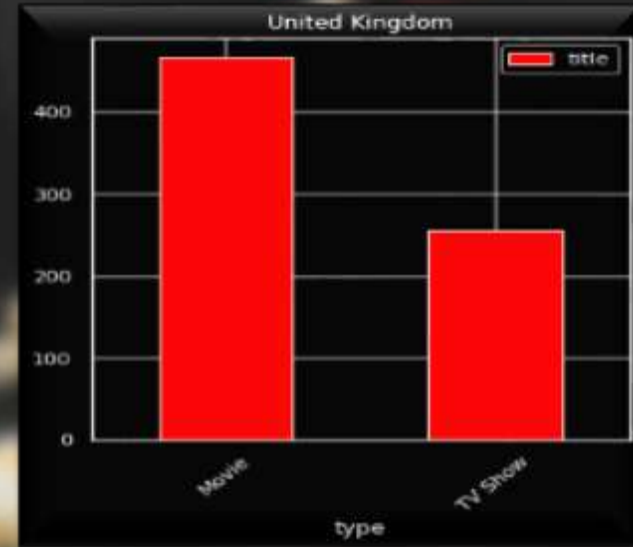
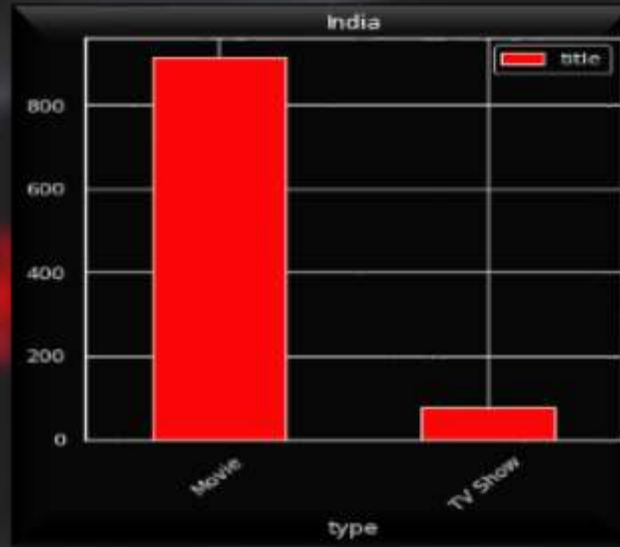
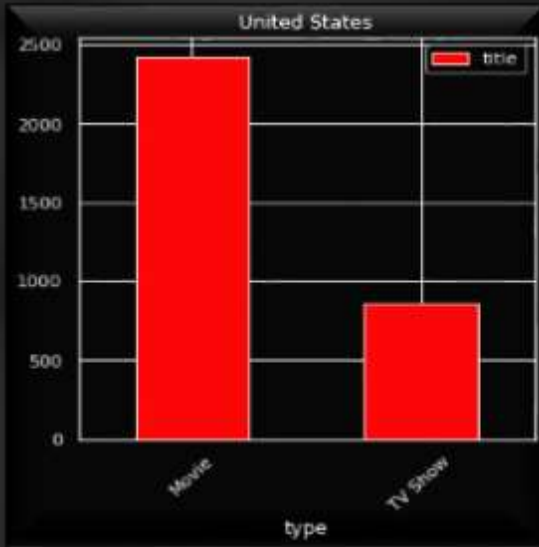
EDA (DATA VISUALIZATIONS):



TOP 3 Countries VS LISTED-IN

- In every country we can see that Dramas are the most popular genre then comedies, documentaries and Action, and thrillers.

EDA (DATA VISUALIZATIONS):



TOP 3 Countries VS Type

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

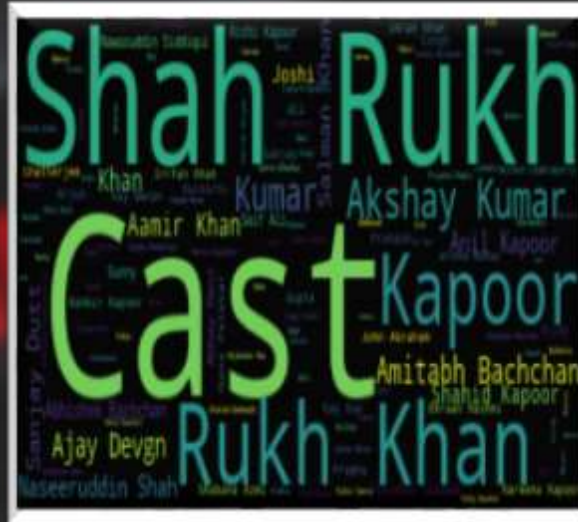
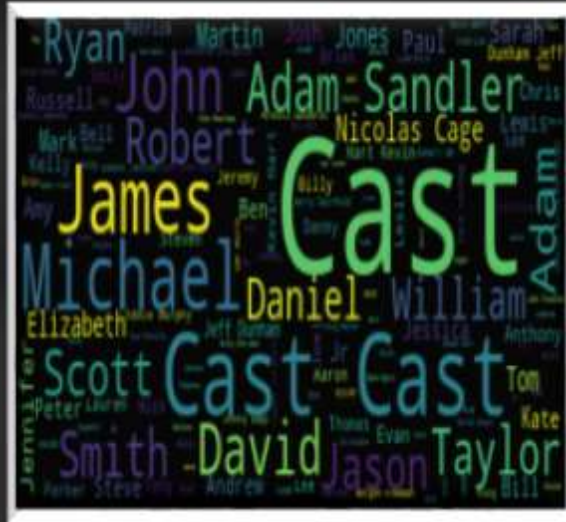
EDA (DATA VISUALIZATIONS):



TOP 3 Countries VS Director

- For India we can see Anurag Kashyap, David Dhawan, Gopal Varma, Dibakar Banerjee and many more.
- For the United Kingdom we can see Mathew, Adam, Michael, John, and many more.
- For the US we can see David, Marcus, Scott, Adam, Johnson, and many more.

EDA (DATA VISUALIZATIONS):



TOP 3 Countries VS Cast

- In India we can see Amitabh Bachchan, Shah Rukh Khan, Naseeruddin Shah, Kareena Kapoor, Anupam Kher, and many more.
- In the UK we can see John, James, David, Katherine and many more.
- In the US we can see Michael, Tom Adam, Scott, William, James and many more.

AI



WordCloud for Genres:

- We can see Dramas, Romantic Movies, Action, Adventure, Family, Horror and many more.

TEXT CLASSIFICATION



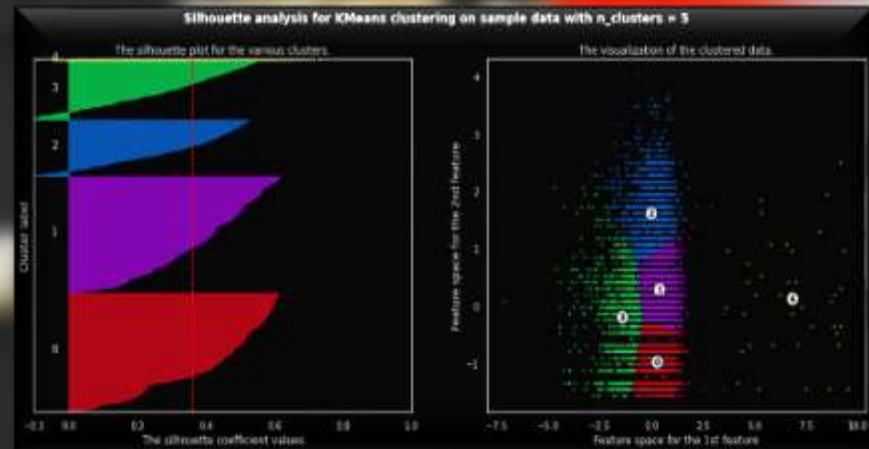
1.CLEANING	2.STOPWORDS	3.TOKENIZATION	4.STEMMING
<ul style="list-style-type: none">▪ Cleaned all null values.▪ All columns-only characters selected by regex.▪ All words to lower case.▪ Merged Text columns.	<ul style="list-style-type: none">▪ Removed stop words.▪ Normal English words & problem specific.	<ul style="list-style-type: none">▪ Splitted sentences to tokens.▪ Used word_tokenize from nltk.▪ Used Count Vectorizer and TF-IDF vectorizer.	<ul style="list-style-type: none">▪ Transformed words to roots.▪ Used Snowball Stemmer.

SILHOUTTE SCORE



- Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.
- 1: Means clusters are well apart from each other and clearly distinguished.
- 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
- -1: Means clusters are assigned in the wrong way.

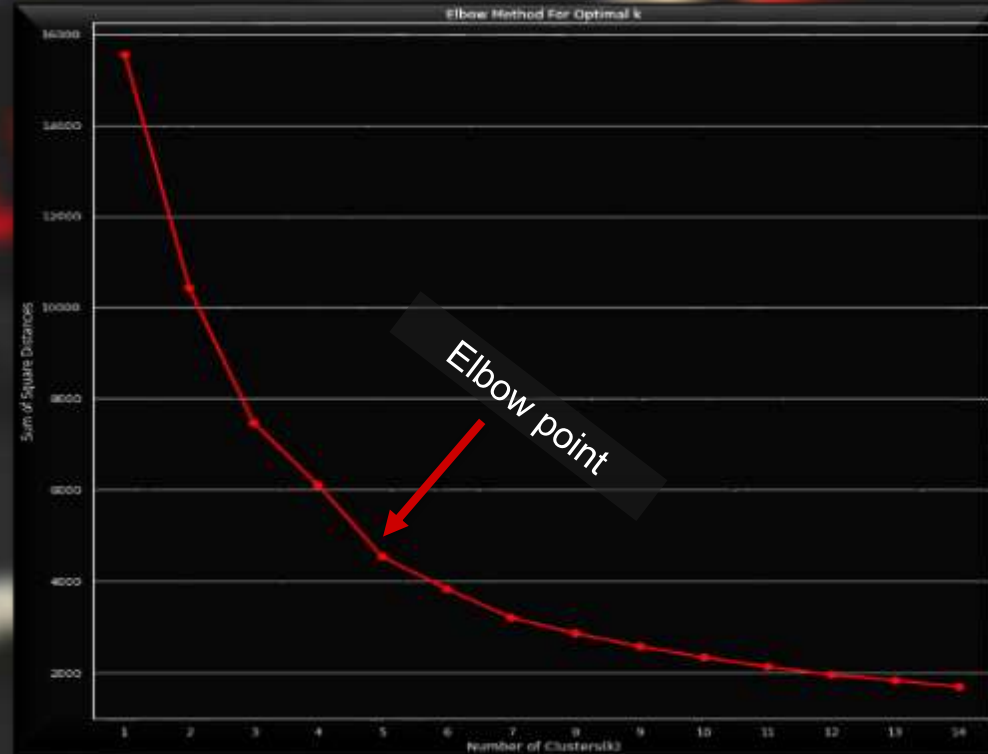
```
For n_clusters = 2, silhouette score is 0.34951465464796985
For n_clusters = 3, silhouette score is 0.37281514879850314
For n_clusters = 4, silhouette score is 0.3862368715405086
For n_clusters = 5, silhouette score is 0.3640633581237277
For n_clusters = 6, silhouette score is 0.34912075022261346
For n_clusters = 7, silhouette score is 0.356956598822019
For n_clusters = 8, silhouette score is 0.33799101638477646
For n_clusters = 9, silhouette score is 0.33719756988072569
For n_clusters = 10, silhouette score is 0.32935664062434744
For n_clusters = 11, silhouette score is 0.3387768469937211
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
```



ELBOW METHOD



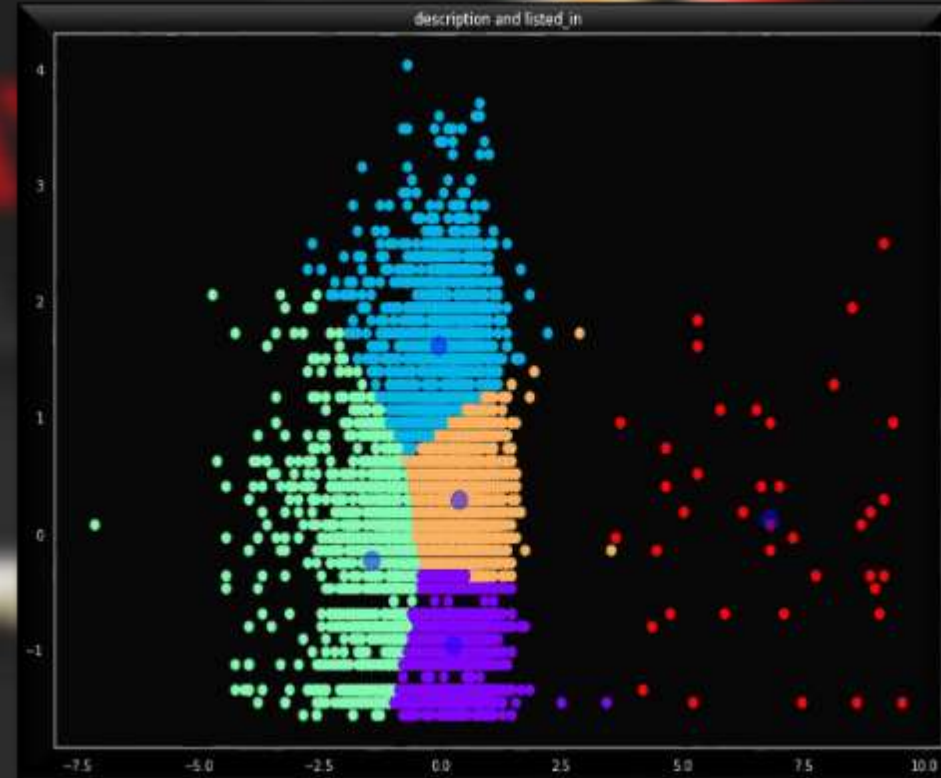
- The Elbow Curve is one of the most popular methods to determine this optimal value of k .
- The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.



K-MEANS



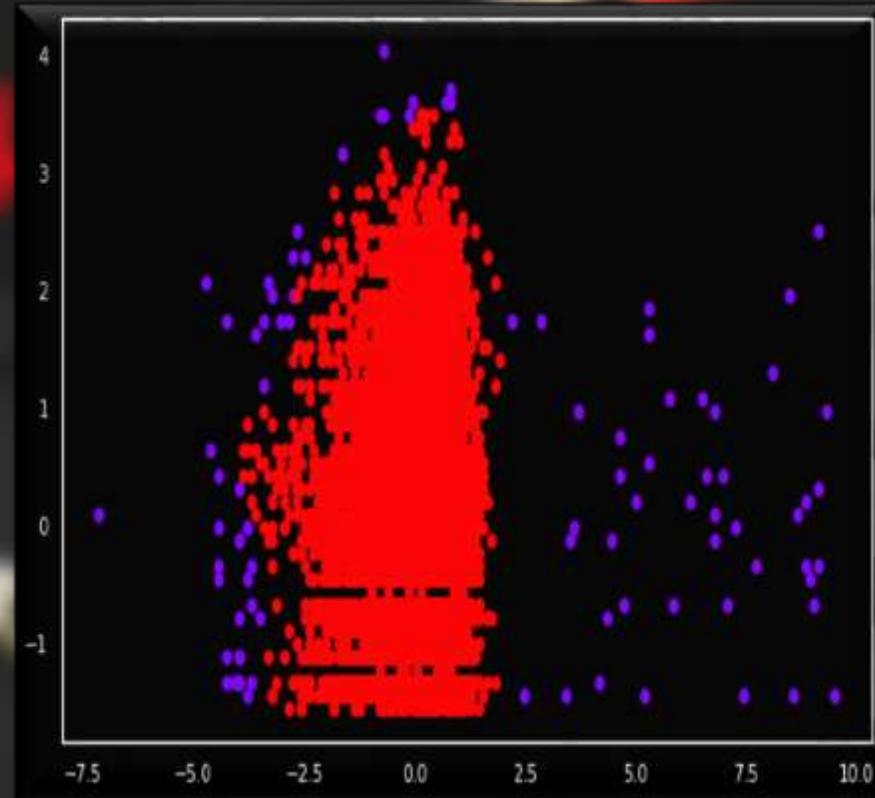
- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.



DBSCAN



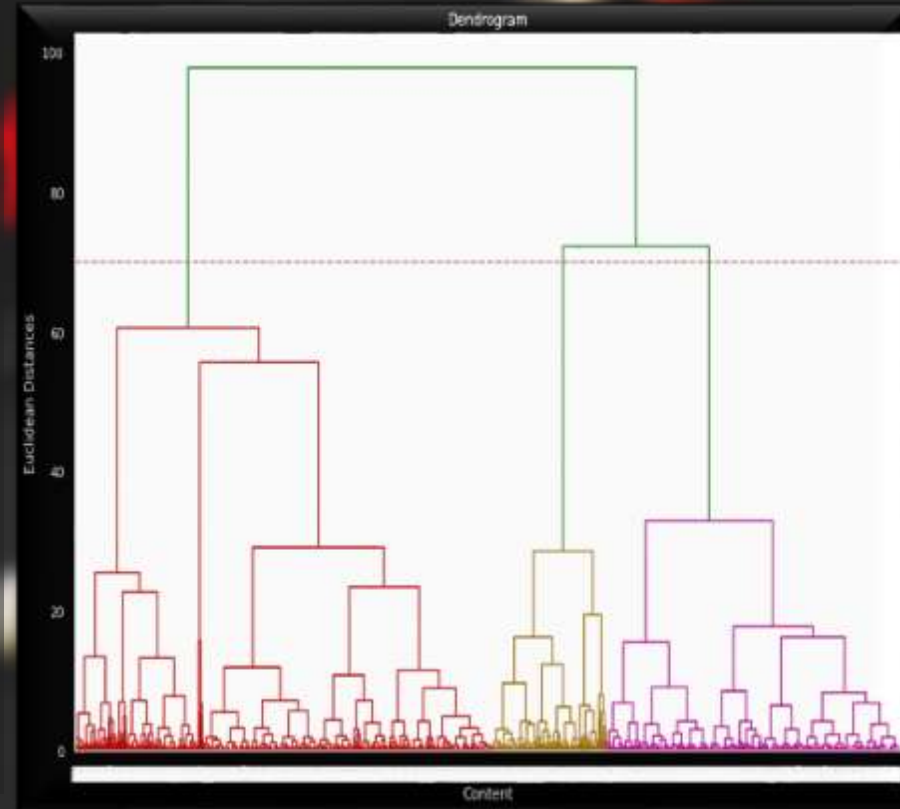
- DBSCAN is the acronym for Density-Based Spatial Clustering of Applications with Noise. DBSCAN is extensively used for the identification of clusters in massive spatial populations or datasets by taking the local densities and properties of the data points into account.
- DBSCAN is especially useful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



HIERARCHICAL CLUSTERING



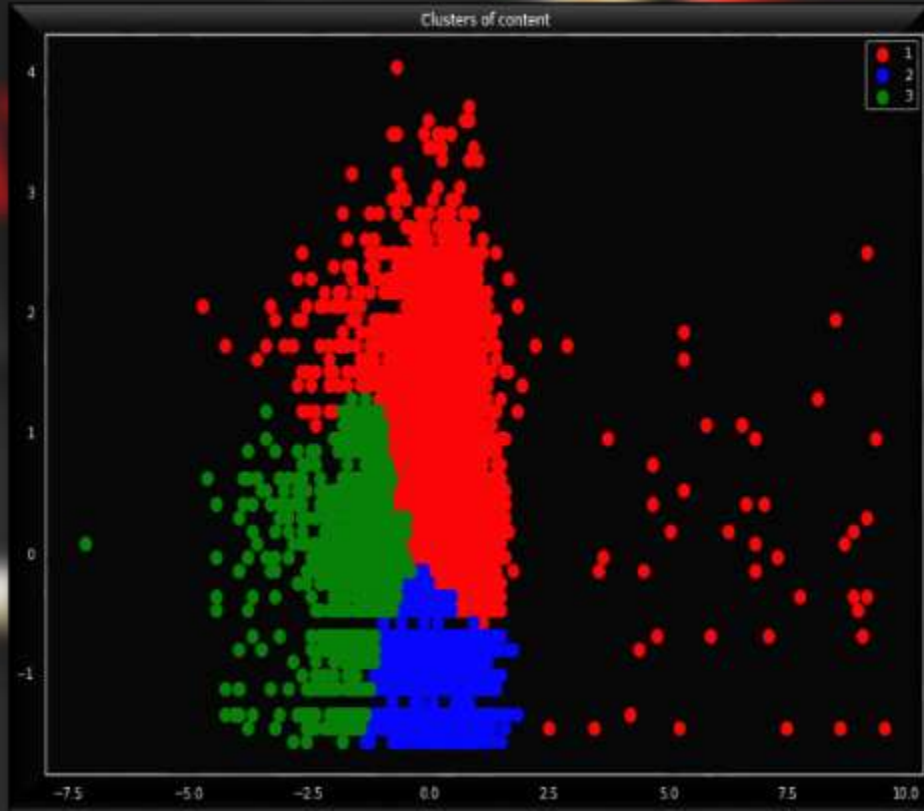
- Hierarchical clustering is also known as Hierarchical Cluster Analysis (HCA) is unsupervised Machine Learning.
- It groups unlabelled data sets into groups also Known as clusters.
- In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known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



AGGLOMERATIVE CLUSTERING



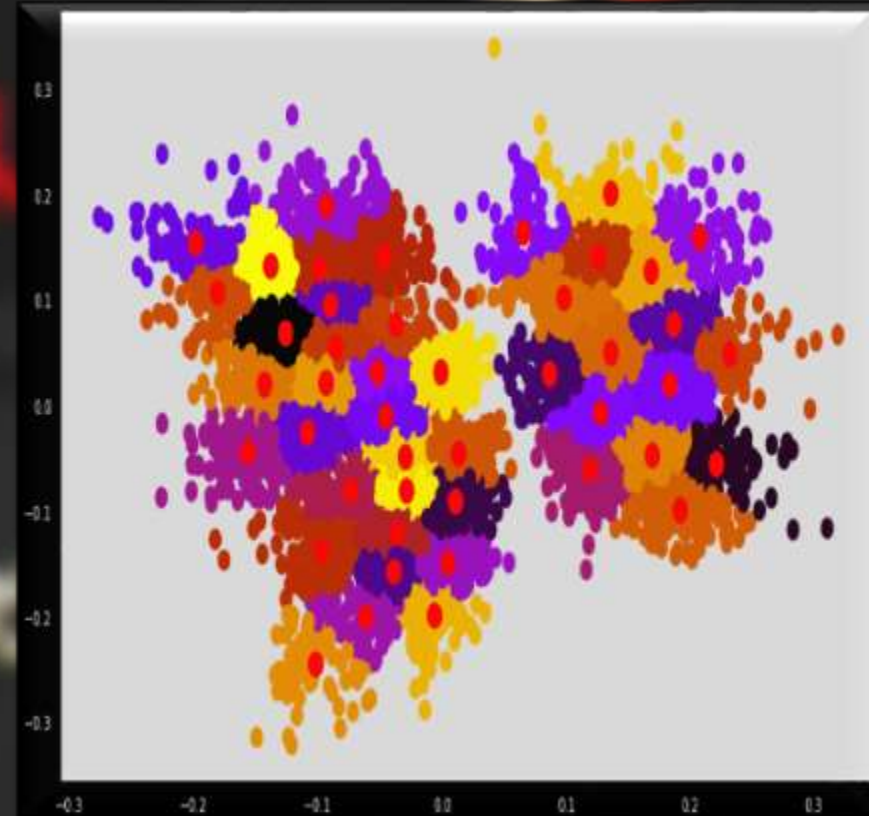
- **Agglomerative Clustering:** Also known as bottom-up approach or hierarchical agglomerative clustering (HAC).
- A structure that is more informative than the unstructured set of clusters returned by flat clustering.
- This clustering algorithm does not require us to prespecify the number of clusters.



PCA



- 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.



RECOMMENDATION SYSTEM



- 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 Netflix Prize is also a competition in the context of recommendation systems.

```
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
```


CONCLUSION



- 1.The dataset contains 7787 rows and 12 columns, cast and director columns have a lot of missing values so we dropped them and we have 10 features for the further analysis.
- 2.Netflix has 69% of its content as movies, so we can say that movies are clearly more popular on Netflix than TV shows. Netflix has more movies than TV Shows
- 3.United States provides the most number of movies and shows followed by India and United Kingdom.
- 4.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.
- 5.There is an exponential raise in the number of TV shows and movies distributed by Netflix in the recent years.

CONCLUSION



6. 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.

7. Optimal number of clusters were found out to be 5 with silhouette coefficient value of 0.36

8. 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 methods of preprocessing is involved.

9. Clusters are identified for each of the record in the dataset.

10. Recommendation based on cosine similarity is also done on the same transformed data.

REFERENCES



1. Stackoverflow
2. GeeksforGeeks
3. Jovian
4. Towards data science



THANK-YOU