NETFLIX MOVIES & TV SHOWS CLUSTERING

Prince Jain, Vikas Shrivas,
Rishabh Patidar
Data science trainees,
Almabetter, Bangalore

Abstract:

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

Introduction:

Netflix's recommendation system helps them increase their popularity among service providers as they help increase the number of items sold, offer a diverse selection of items, increase user satisfaction, as well as user loyalty to the company, and they are very helpful in getting a better understanding of what the user wants. Then it's easier to get the user to make better decisions from a wide variety of movie products. With over 139 million paid subscribers (total viewer pool -300 million) across 190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the world's leading Internet television network and the most-valued largest streaming service in the world. The amazing digital success story of Netflix is incomplete without the mention of its recommender systems that focus on personalization. There are several methods to create a list of recommendations according to your preferences. You can use (Collaborative-filtering) and (Content-based Filtering) for recommendation.

1. 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:

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

2. Attribute Information:

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
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** : Genere
12. description: The Summary description.

3. Research Methods:

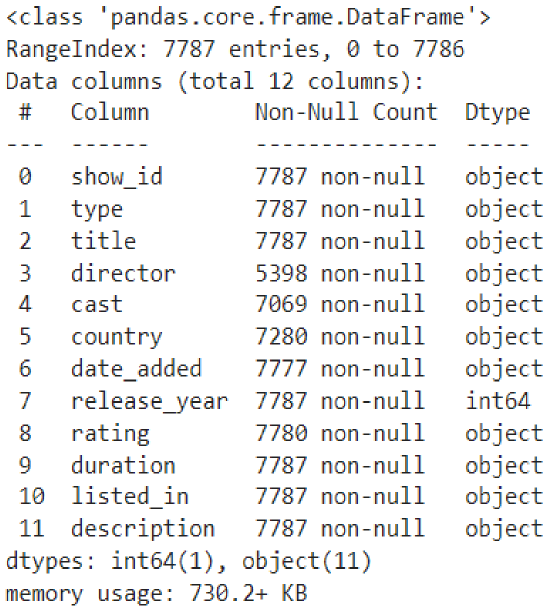
In this paper, we propose a classification framework based on ensemble learning to classify and predict terrorist organizations. The framework involved four steps, including data collection, and understanding, data preprocessing, data cleaning and Exploratory Data Analysis (EDA).

3.1. Data collection and understanding:

The data primarily contained the following attributes of information: Unique ID for every Movie / Tv Show, Identifier - A Movie or TV Show, Title of the Movie / Tv Show, Director of the Movie, Actors involved in the movie / show, Country where the movie / show was produced, Date it was added on Netflix, Actual Release year of the movie / show, TV Rating of the movie / show, Total Duration - in minutes or number of seasons, Genre, description: The Summary description.

3.2. Handling missing values:

We will need to replace blank countries with the mode (most common) country. It would be better to keep director because it can be fascinating to look at a specific filmmaker's movie. As a result, we substitute the null values with the word 'unknown' for further analysis.
There are very few null entries in the date added fields thus we delete them.



```
<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
```

3.3. Data cleaning: There are 3631 null values in the dataset, 2389 null values in director column, 718 null values in cast column ,507 null values in country column ,10 in date added and 7 in rating. so, we need to handle the null values cleaning.

Label Encoding

Lemmatization- Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. ... For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.

Removing Stop words - To remove stop words from a sentence, you can divide your text into words and then remove the word if it exits in the list of stop words provided by NLTK.

Min-max Scaling - For each value in a feature, MinMaxScaler subtracts the minimum value in
the feature and then divides by the range. It preserves shape of original distribution.

3.4. Exploratory Data Analysis (EDA):

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

4. Analysis:

This section consists of details regarding the visual results:

4.1 Type of content available on Netflix:

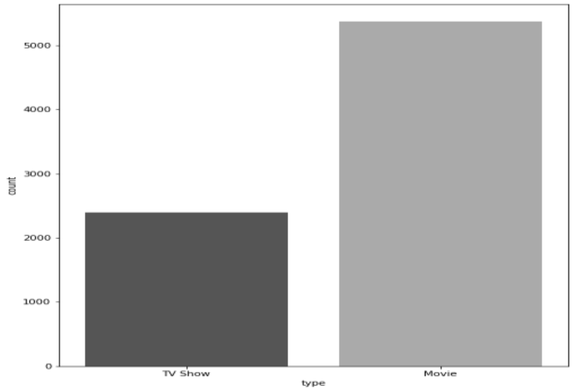
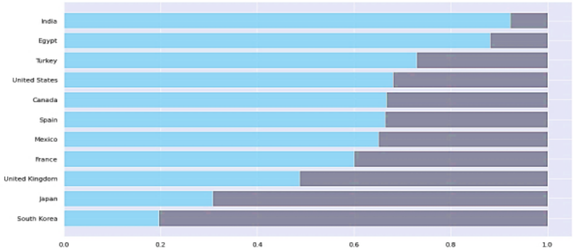
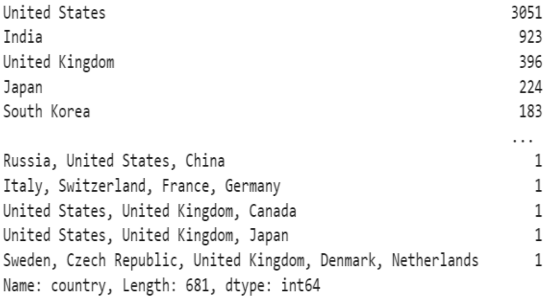


Figure 1: Shows the TV shows or movies
•It is evident that there are more movies on Netflix than TV shows.
•Netflix has 5372 movies, which is more than double the quantity of TV shows.

4.2 Country data Analysis:

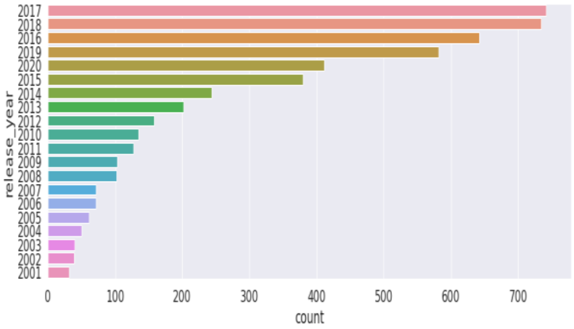


India is the country having maximum numbers of movie on Netflix.
USA having maximum numbers of TV shows followed by India.



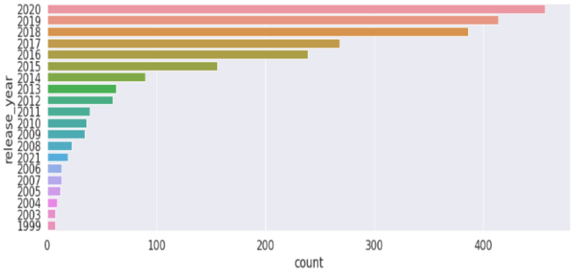
```
United States                                               3051
India                                                        923
United Kingdom                                               396
Japan                                                        224
South Korea                                                  183
                                                            ...
Russia, United States, China                                   1
Italy, Switzerland, France, Germany                            1
United States, United Kingdom, Canada                          1
United States, United Kingdom, Japan                           1
Sweden, Czech Republic, United Kingdom, Denmark, Netherlands   1
Name: country, Length: 681, dtype: int64
```

4.3 Releases over the year:



movies

Highest number of movies released in 2017 and 2018.



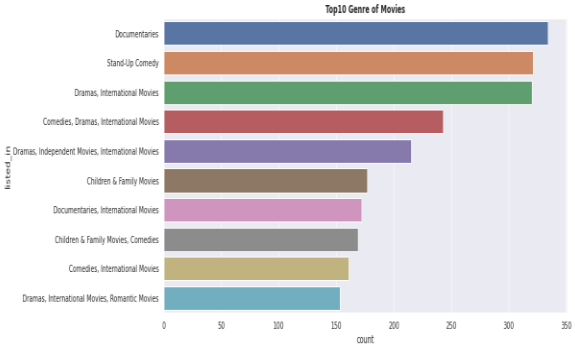
TV shows

Highest number of movies released in 2020. 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.

There is a significant drop in the number of movies and television episodes produced after 2020.

4.4 Top 10 Genre:



Documentaries are highest in numbers followed by standup comedy then Dramas.

5. Word cloud representation:

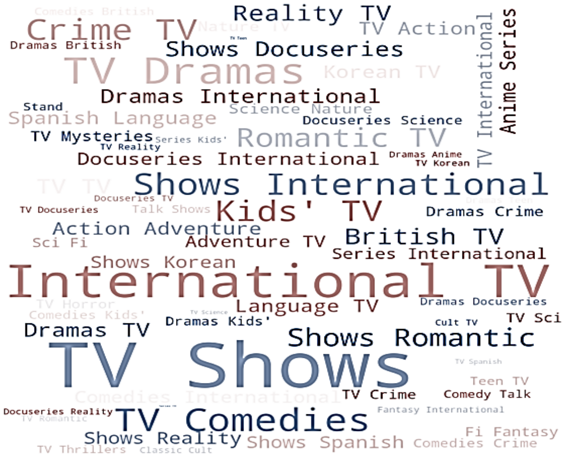


Figure: TV Shows types

Here we plot the Word cloud in Tv shows and find that as the texts are large in display which means the frequency of that shows is high and it is clearly seen that International TV are leading channel followed by Tv shows and then TV comedy.

5.1 Which Cast to Choose:

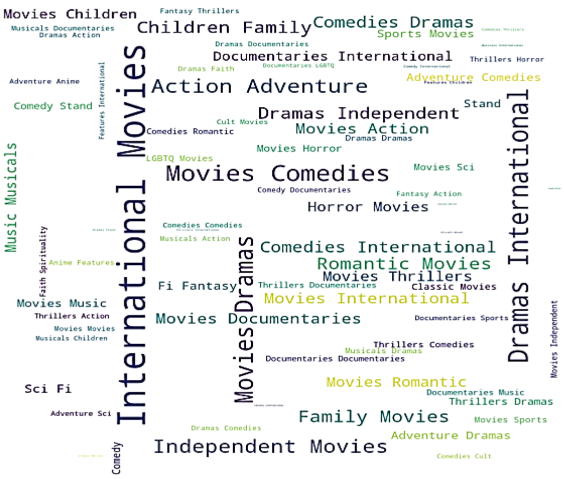
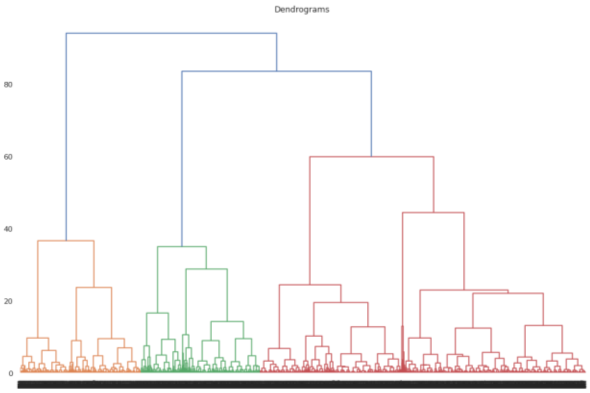


Figure: Which Cast to Choose?

We find that in Movies the International movies are leading after dramas International and comedies movies also making an impact in terms of movie in Netflix.

6. Models Implementation:

6.1. Hierarchical Clustering:



Hierarchical clustering is the most popular and widely used method **to analyze social network data**. In this method, nodes are compared with one another based on their similarity.

6.2. DBSCAN:

DBSCAN stands for density-based spatial clustering of applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

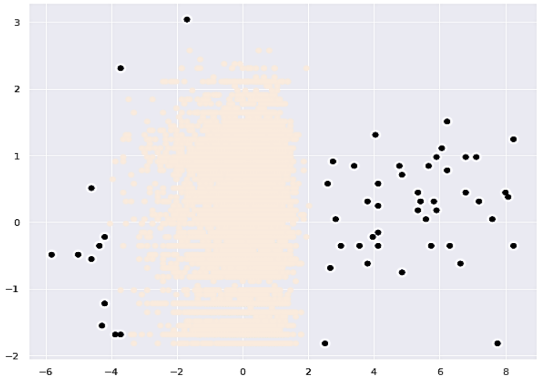
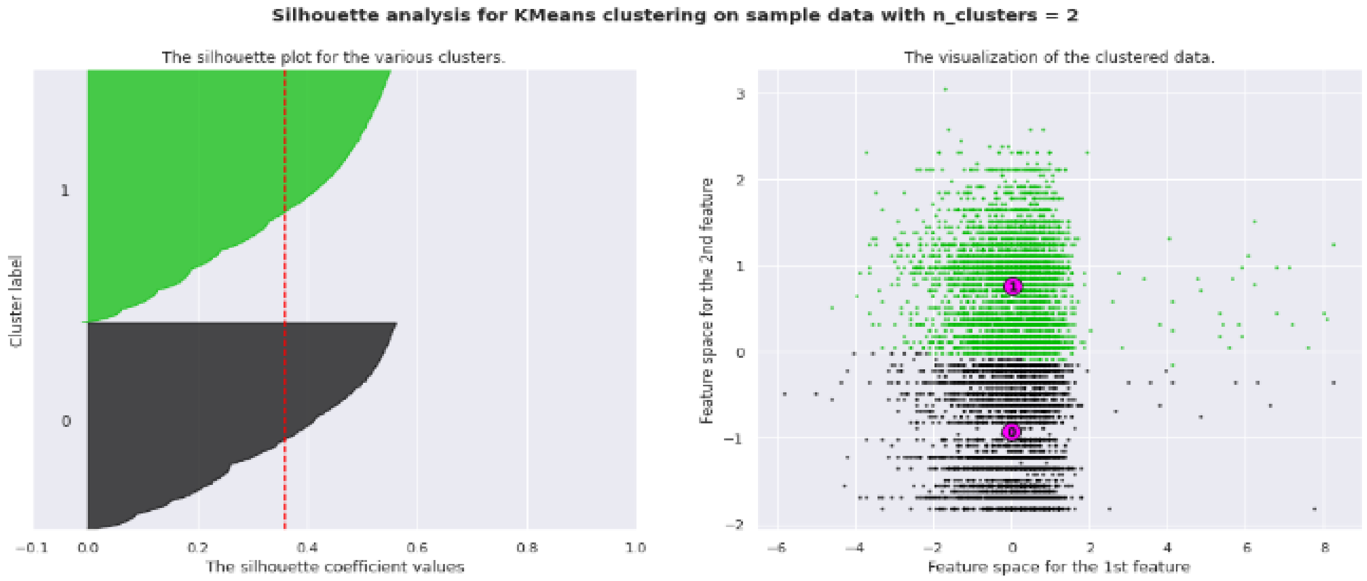


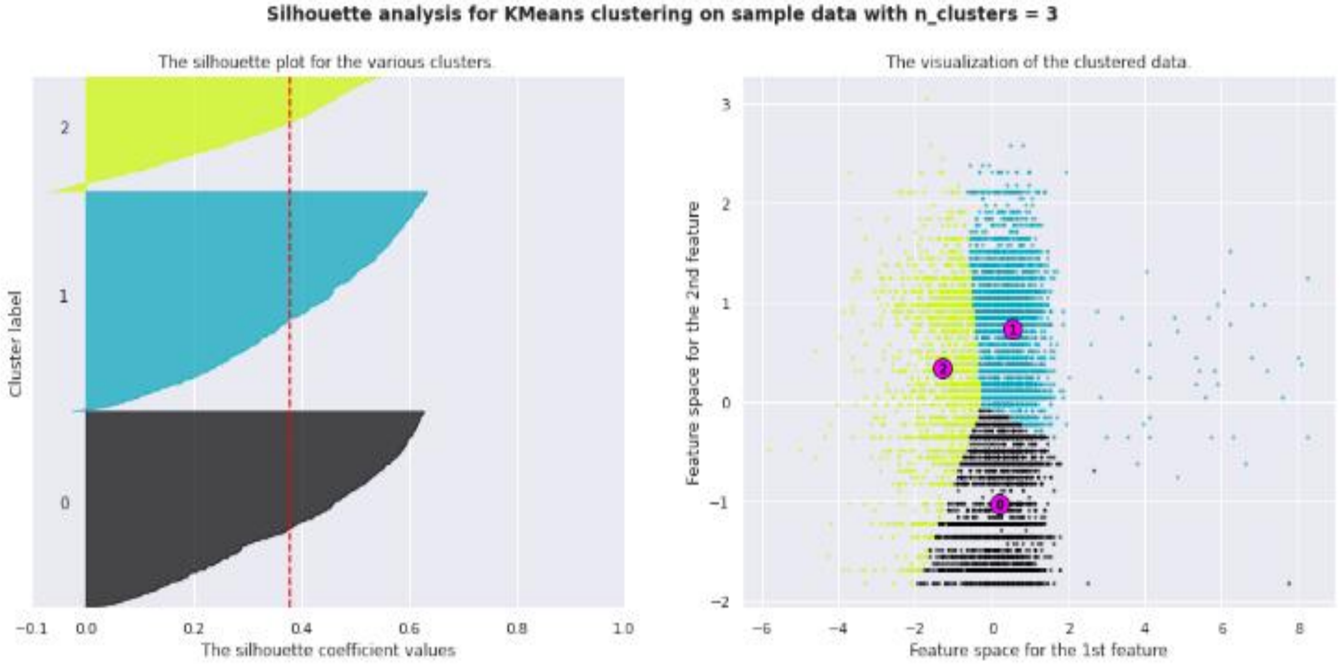
Figure: density-based spatial clustering of applications with noise

6.3. k-means clustering:


Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

The silhouette plot for the various clusters.

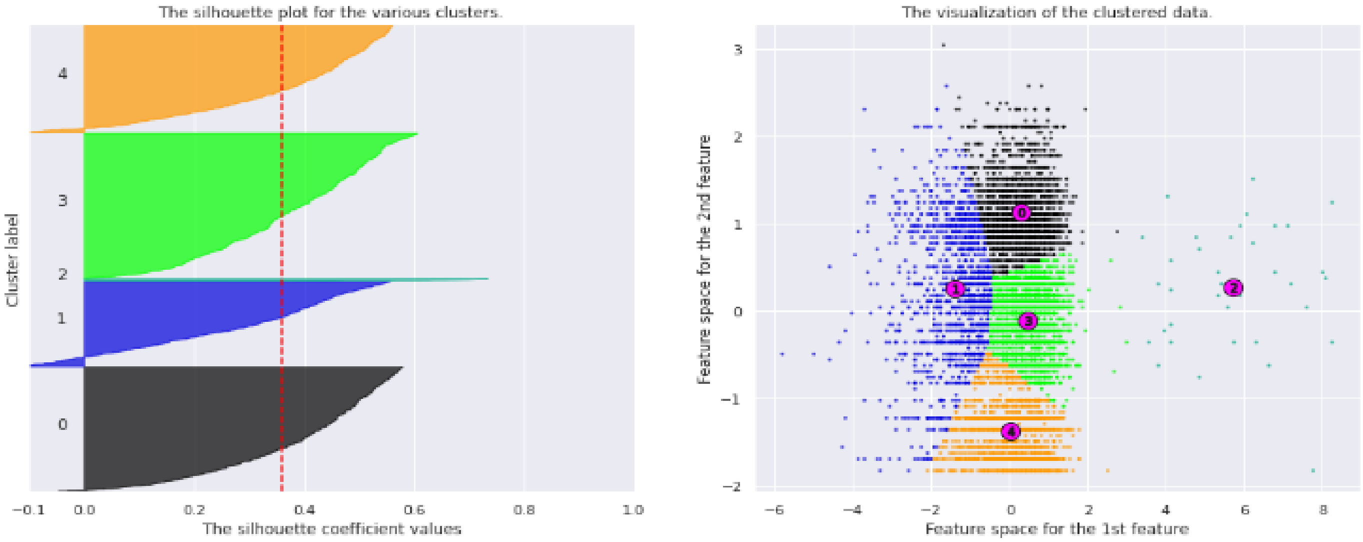
The visualization of the clustered data.


Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

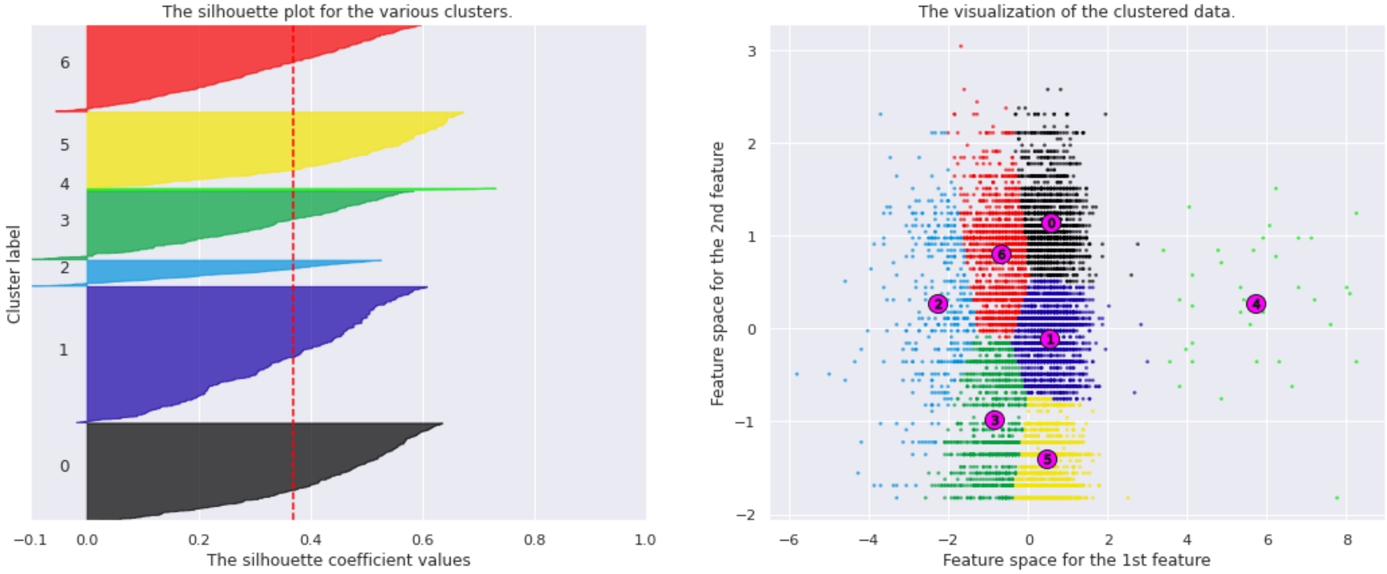
The silhouette plot for the various clusters.

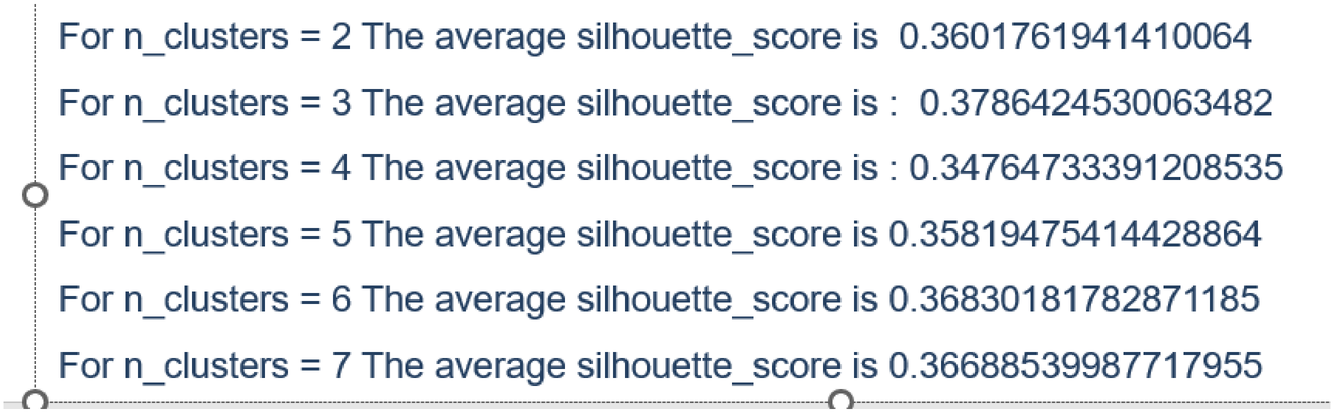
The visualization of the clustered data.

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



Silhouette analysis for KMeans clustering on sample data with n_clusters = 7





For n_clusters = 2 The average silhouette_score is 0.3601761941410064

For n_clusters = 3 The average silhouette_score is : 0.3786424530063482

For n_clusters = 4 The average silhouette_score is : 0.34764733391208535

For n_clusters = 5 The average silhouette_score is 0.35819475414428864

For n_clusters = 6 The average silhouette_score is 0.36830181782871185

For n_clusters = 7 The average silhouette_score is 0.36688539987717955

4. Technologies used:

Python: Python is a high-level interpreted language that supports different platforms like Windows, Linux, Mac, Raspberry Pi, etc. Python can be used for creating web applications, database systems, handle big data, perform complex mathematical calculations. Python can be treated in an object-oriented, functional or procedural way.

Google Colab: Collaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

Python packages: Following are some of the python packages used in this project.

Matplotlib: Matplotlib is an visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Pandas: Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

NumPy: It provides structures for multiple dimensional array objects and tools for related operations. NumPy is usually used for high performance scientific computational tasks.

Seaborn: Seaborn is an visualization library for statistical graphics plotting in Python. It provides default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset**.**

Word cloud: A word cloud (also known as a tag cloud or text cloud) is a visual representation of a text, in which the words appear bigger the more often they are mentioned. Word clouds are great for visualizing unstructured text data and getting insights on trends and patterns.

Clustering: In machine learning too, we often group examples as a first step to understand a subject (data set) in a machine learning system. Grouping unlabeled examples is called clustering. As the examples are unlabeled, clustering relies on unsupervised machine learning.

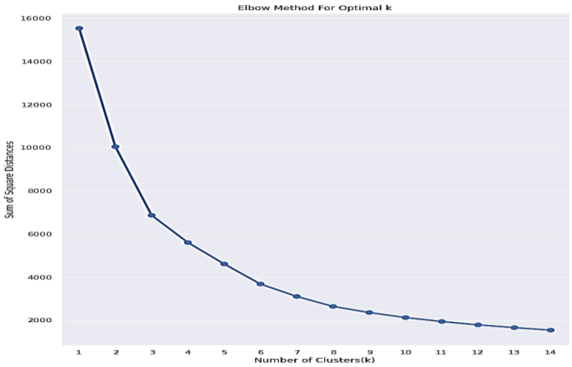
K-means clustering: It is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.
K-means algorithm works:
To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

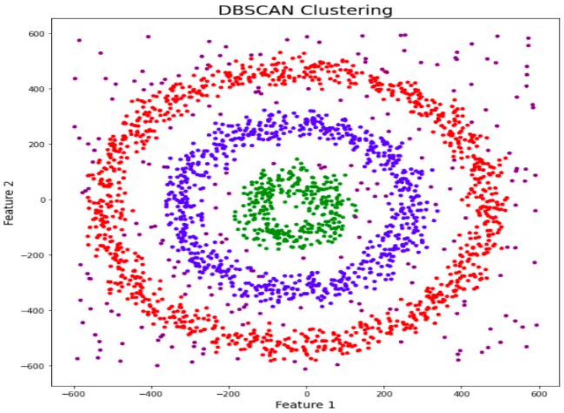
K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non overlapping subgroups where each data point belongs to only one group.

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**.**



DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.



Hierarchical clustering: A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps: Identify the 2 clusters which can be closest together, and. Merge the 2 maximum comparable clusters.

5. Conclusion:

- Netflix has 5372 movies and 2398 TV shows, there are more number movies on Netflix than TV shows.
- TV-MA has the highest number of ratings for tv shows i.e., adult ratings
- Highest number of movies released in 2017 and 2018.
- 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.
- There is a significant drop in the number of movies and television episodes produced after 2020.
- It appears that Netflix has focused more attention on increasing Movie content that TV Shows. Movies have increased much more dramatically than TV shows.
- India is the country having maximum numbers of movie on Netflix.
- USA having maximum numbers of TV shows followed by India.
- Those movies that have a rating of NC-17 have the longest average duration.
- When it comes to movies having a TV-Y rating, they have the shortest runtime on average
- October to January, maximum number of movies and TV shows were added.

References:

1. Google.com
2. GeeksforGeeks.
3. https://colab.research.google.com/drive/1I1P4ZV4mJirpLrd2AYEmb5bdx9SsbDRG#scrollTo=kYDuuNSBhE9W
4. https://github.com/apoorvaKR12695/Netflix-Movies-and-TV-Shows-Clustering/blob/main/final_notebook_NETFLIX_MOVIES_AND_TV_SHOWS_CLUSTERING.ipynb

5. https://www.kaggle.com/code/adityarawat10/netflix-visualization-eda

6. https://github.com/San13deep/NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING/blob/main/Capstone_4_NETFLIX_MOVIES_AND_TV_SHOWS_CLUSTERING.ipynb

7. https://www.google.com/search?q=K-means+Clustering+unsupervised+ML+algorithm&oq=K-means+Clustering+unsupervised+ML+algorithm&aqs=chrome..69i57.1482j0j7&sourceid=chrome&ie=UTF-8