Netflix Movies and TV shows Clustering

V Rupesh Kumar Patro
Shashank Maindola

Abstract:

Netflix is a popular American subscription-based streaming service that provides a vast library of movies, television series, documentaries, and other forms of visual entertainment. It was founded in 1997 by Reed Hastings and Marc Randolph and has since grown to become one of the world's leading streaming platforms.

Netflix offers a wide range of content, including original productions and licensed content from various studios and distributors. Subscribers can access the service through the Netflix website or mobile applications on various devices, such as smartphones, tablets, smart TVs, gaming consoles, and streaming media players.

The platform has revolutionized the way people consume media by offering a convenient and on-demand way to watch TV shows and movies. Users can stream content instantly without having to wait for traditional television schedules or physical media like DVDs. Netflix's extensive library, personalized recommendations, and the ability to binge-watch entire series have contributed to its widespread popularity across the globe.

Introduction:

Unsupervised Learning is a machine learning technique in which the model is not supervised by the training set instead we find hidden patterns and insight from the given data. It is a machine learning technique in which model are trained on the unlabeled data set without any supervision. A cluster is a collection of elements that are similar to each other but dissimilar to the elements belonging to other clusters. In 2018, Netflix 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. Here we will use KNN clustering, hierarchical clustering and DBSCAN clustering for the same

Below is the info that is available in given dataset-

- **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 :** 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 for appliance energy prediction using machine learning

1. **EDA**
2. **Pre Processing**
3. **Model Implementation & Explaination**

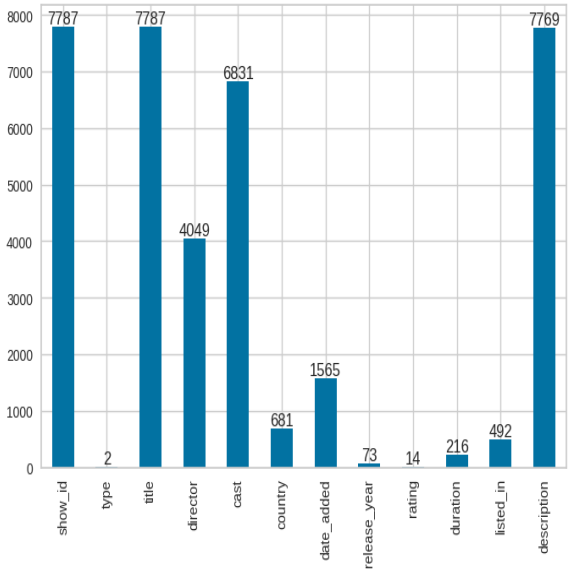
4. **Conclusion**

1. Exploratory Data Analysis

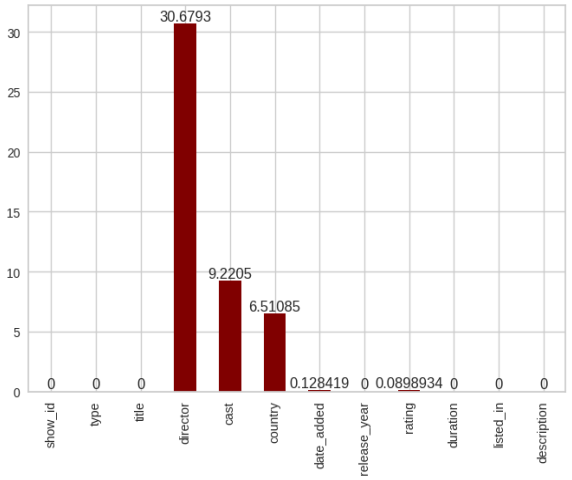
We first did a preliminary data analysis and these were the salient features of our data set that we found:-

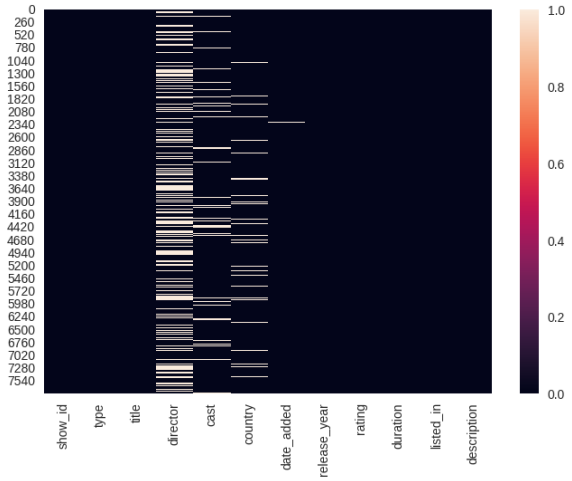
- We have 12 features and 7727 rows in our dataset
- There are movies and Tv show titles present and each show has its own ID . It also contains rating, cast, director, release year and date added to it along with the type of genre it belongs to.

Fields	Description
show_id	Unique ID for every Movie / Tv Show
type	Identifier - A Movie or TV Show
title	Title of the movie/show
director	Director of the show
cast	Actors involved
Country	Country of production
date_added	Date it was added on Netflix
release_year	Actual release year of the show
rating	TV rating of the show
duration	Total duration in minutes or number of seasons
listed_in	Genre
Description	The summary description



Then we looked for missing values





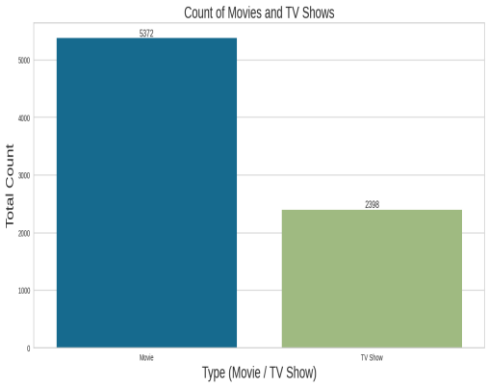
We found out:-

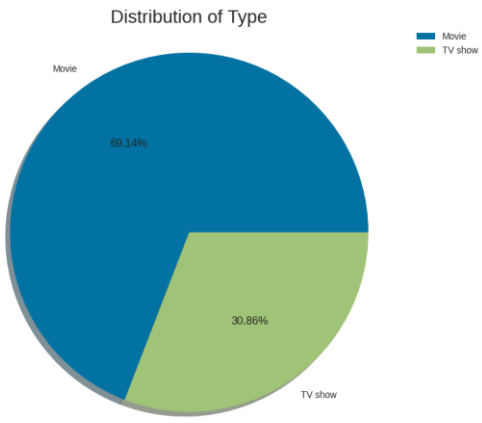
- Nan Values are more in Director , Cast and Country Columns
- **Director** column has highest NaN values 30.7% data is missing
- **Cast** column has 9% NaN values
- **Country , date_added , rating** columns also containing missing values

Since director have many null values if we drop them, we loss a lot data so, we replacing them with unknown.

Now that we have imputed the missing value we can start making observations. We broadly summarized our observations as follows:-

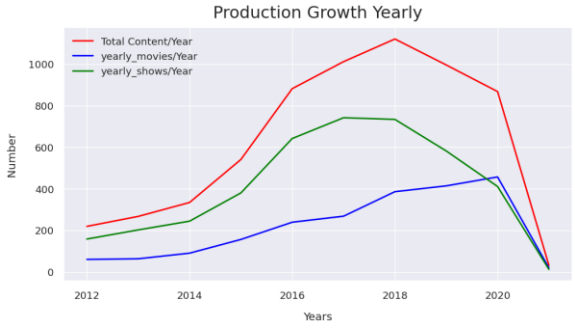
Content Type





- There are 5372 movies and 2398 TV shows present on Netflix platform.
- Almost ¾ of total content are movies.

Growth in the content release over years



Year 2018 saw maximum increase in content on Netflix and then there was a sudden drop in new content release around 2020 maybe due to covid

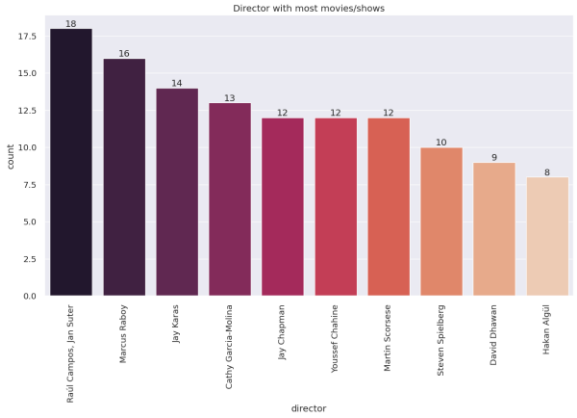
Title

It seems like words like "Love", "Man", "World", "Story" are very common in titles. However, we are surprised by the overwhelming number of content having "Christmas" in their title.

We are suspecting "Christmas" titles to be a very seasonal thing with most of the shows likely to be released during the month of December. However, we do not have the

release month in this data to conform our hypothesis.

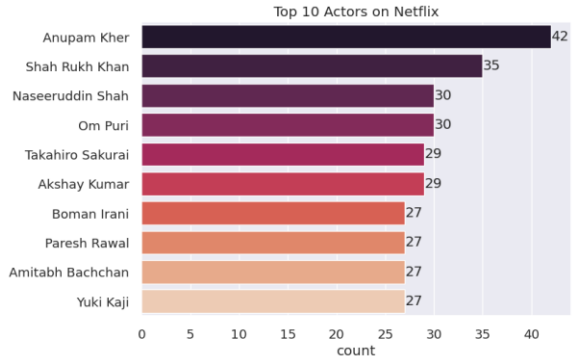
Top 10 directors on Netflix



Raul Campos and Jan Sulter collectively have the most content on Netflix. They have around 18 titles featured on Netflix closely followed by Marcus Raboy who has 16.

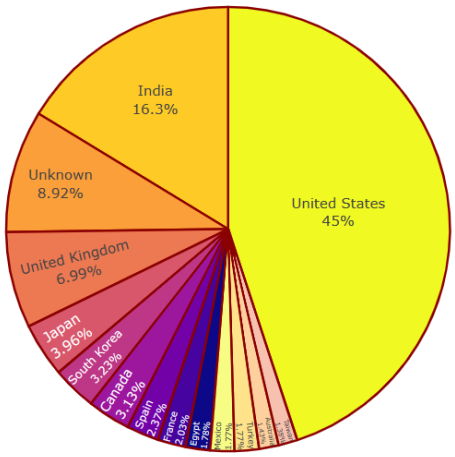
western countries tweeted more than other parts of world.

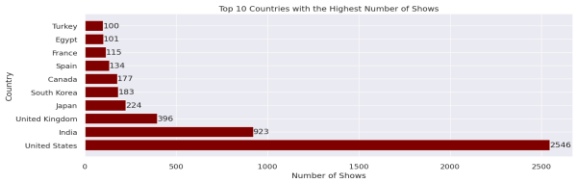
Top 10 actors on Netflix



- The top actor by the number of shows they appeared in is Anupam Kher, who appeared in 42 titles in the dataset.
- The second most popular actor is Shah Rukh Khan, who appeared in 35 shows.
- Most of the top 10 actors have appeared in between 25-30 shows on Netflix.
- The top 10 actors are mostly from India, with the exception of Takahiro Sakurai and Yuki Kaji from the Japan.

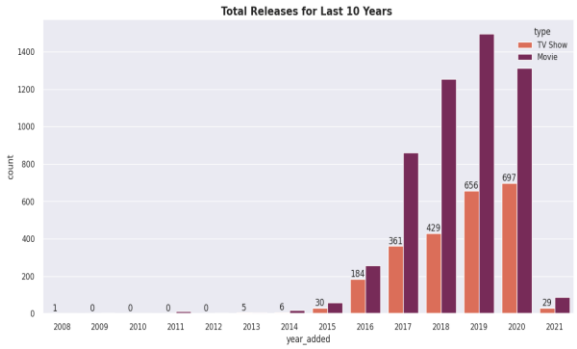
Content by Country

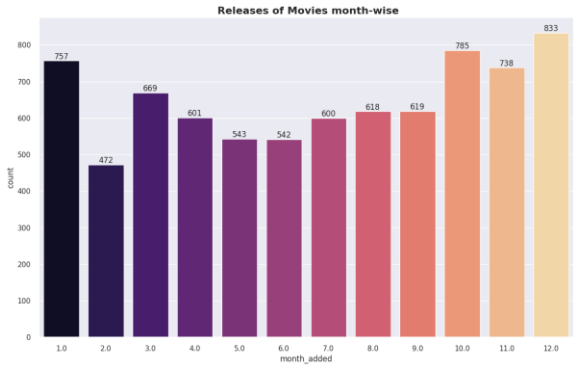




The United States is by far the largest content producer of movies and TV shows in the dataset, with over 2500 titles contributing to 45% of all titles. The next highest content producing countries are India at 17.8%, the United Kingdom 7.68 % share in total.

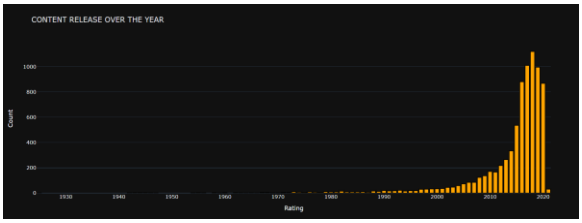
Date added





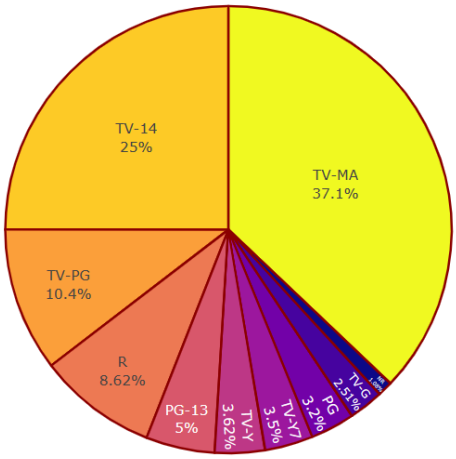
- The number of release have significantly increased till 2020 but after 2020 and have dropped in 2021 as the data which we have is till 2020 and also in covid era there were few production units which could follow the covid suite.
- Movie addition increased over the years but suddenly reduced in 2020 but TV show addition has increased over the year and hasn't seen decline until 2021
- More of the content is released in holiday season - October, November, December and January.

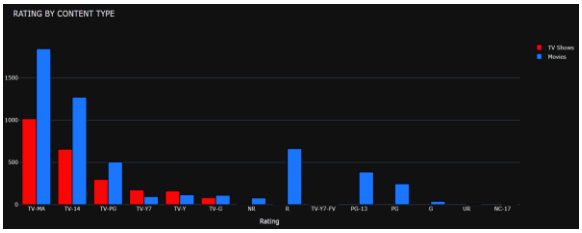
Release Year



We see a slow start for Netflix over several years , maybe Netflix was a small company back then. Things begin to pick up in 2015 and then there is a rapid increase from 2016. This rapid growth in contents on their platform which signifies there may be more investment and licensing deals

Rating





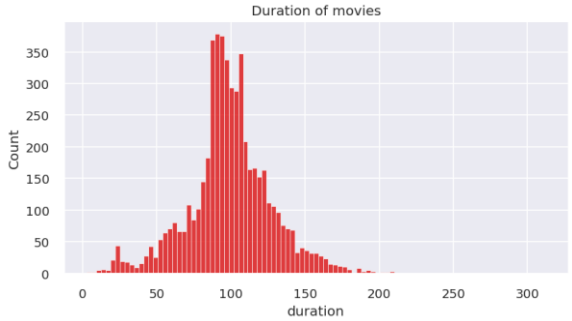
Most of the contents got ratings like:

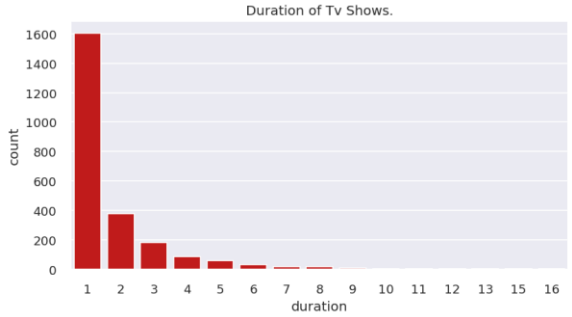
- TV-MA (For Mature Audiences)
- TV-14 (May be unsuitable for children under 14)
- TV-PG (Parental Guidance Suggested)

- NR (Not Rated)

We also observe that some ratings are only applicable to Movies. The most common rating for both Movies & TV Shows are Mature audience and under 14 years

Duration of Content

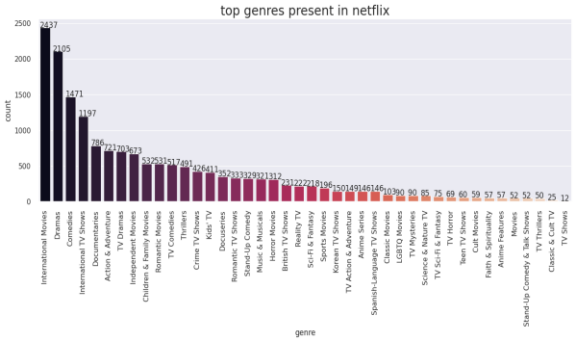




Duration of TV shows have been mentioned in the form of seasons. Most of the Tv Shows last for 1 or 2 seasons, it is rare for a show to have more than 5 seasons.

Most of the movies last for 90 to 120 minutes.

Top Genres



- Highest number of genre belong from International movies, Dramas, Comedies respectively.
- Least number of genre belong from Classic & cult TV, TV Thriller, Stand-Up comedy and TV shows.

2. Preprocessing

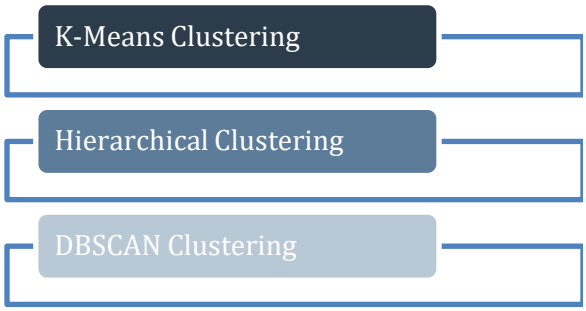
As part of preprocessing, we first expand contraction, lower casing, removed punctuation, removed URL's & digits and removed stopwords and white spaces.

Then we used other NLP algorithms, the methods which we used are as follows:-

- **Tokenization**
  We use this to split paragraphs and sentences into smaller units that can be more easily assigned meaning.
- **Text Normalization**
  Using this pre-processing step for improving the quality of the text and making it suitable for machines to process.
- **Text Vectorization**
  We use TF-IDF (Term Frequency-Inverse Document Frequency) text vectorization for text classification and information retrieval tasks. It assigns weights to each word in the document based on its frequency and rarity across the corpus.
- **Dimensionality Reduction**
  As the number of features (words in this case) is high, it is useful to apply dimensionality reduction to simplify the dataset and improve computational efficiency. We use PCA with 95% for Dimensionality reduction

3. Model Implementation and Explanation

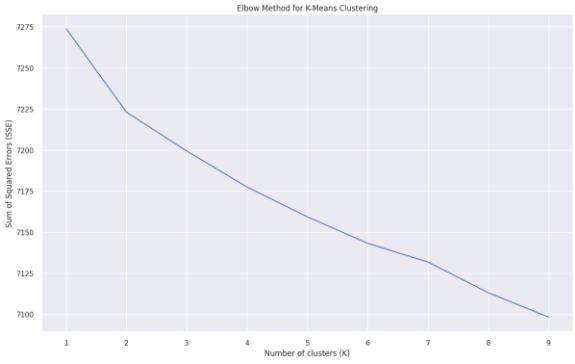
We are using three clustering algorithms for model implementation. We will use Silhouette Score, Calinski-Harabasz Score & Davies-Bouldin Score for performance evaluation of the models



K-MEANS CLUSTERING

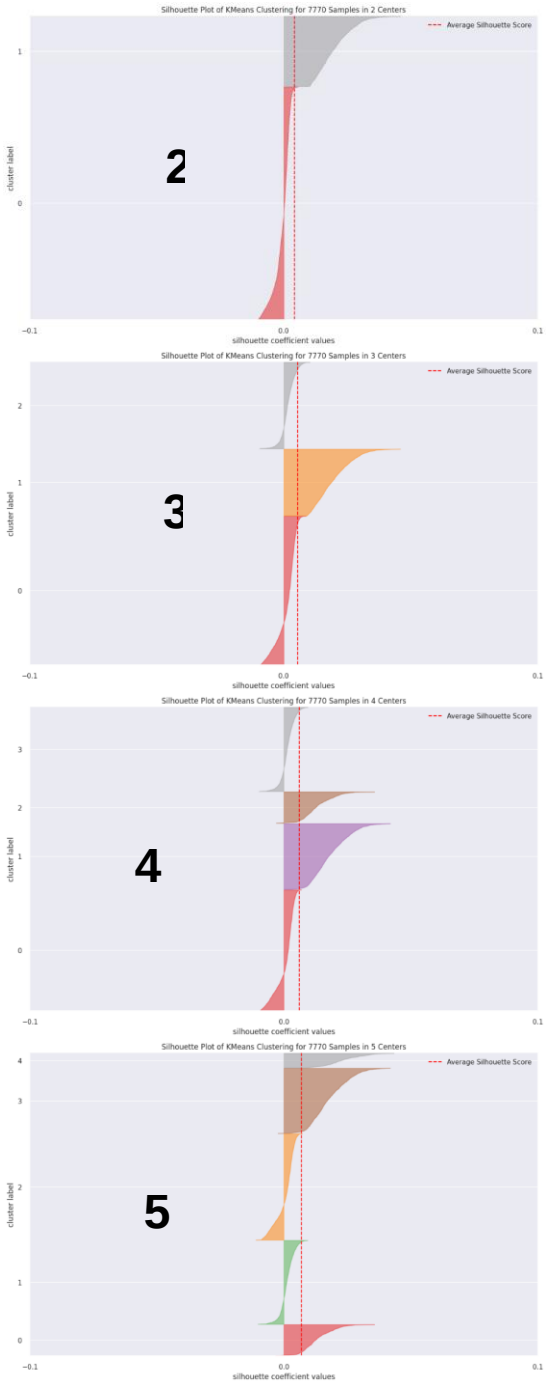
To Find Optimum Numbers of Clusters we use following methods

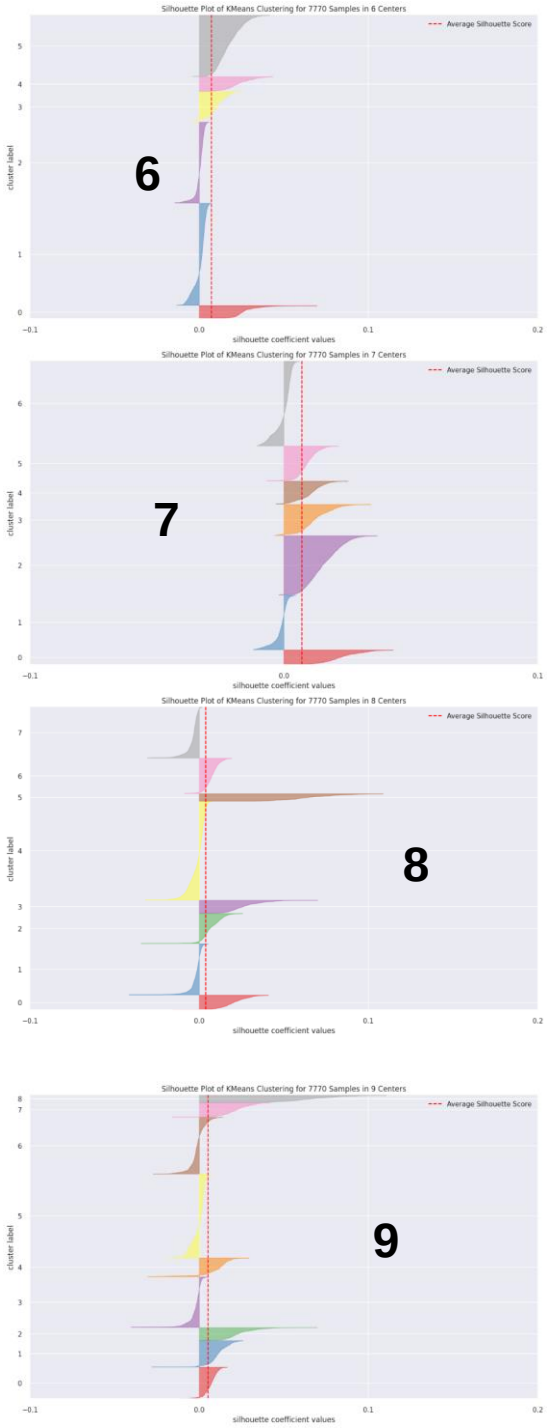
- Elbow Method
- silhouette Score

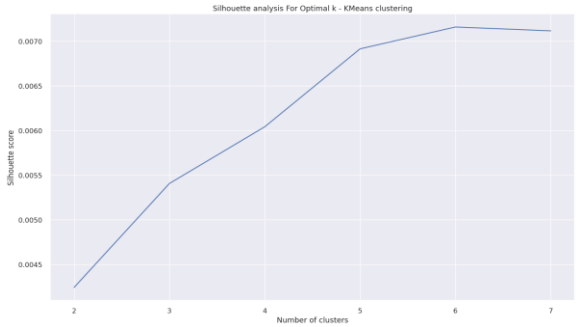


We have narrowed down the range of possible number of clusters to be between 4 to 7, as the slope of the elbow plot is steep at this range. To determine the optimal number of clusters, we will check the silhouette scores for each value in this range and choose the one with the highest score.

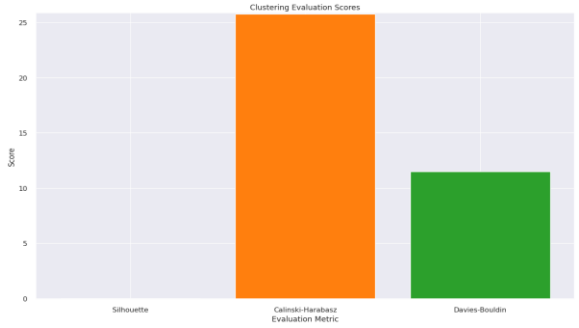
We can conclude that Logistic regression is the best model for our dataset, followed closely by Navies Bayes, KNN Classifier and Random Forest classifier did not give a good result compared to others.







We observe that 6 and 7 have best performance. There is slight difference between performance hence we will consider 7 clusters to be optimum.

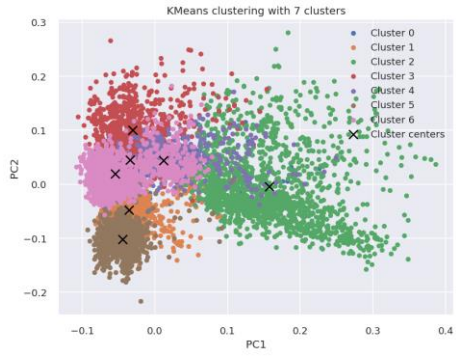


We ran performance evaluation algorithms and got following results.

```
Number of clusters: 7
Silhouette score: 0.0071
Calinski-Harabasz score: 25.7653
Davies-Bouldin score: 11.5074
```

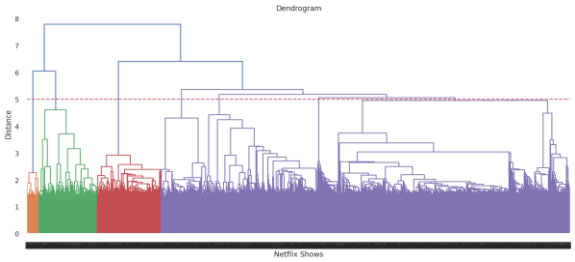
	Evaluation Metric	Score
0	Silhouette Score	0.00711593
1	Calinski-Harabasz Score	25.7653
2	Davies-Bouldin Score	11.5074

Model Explanation

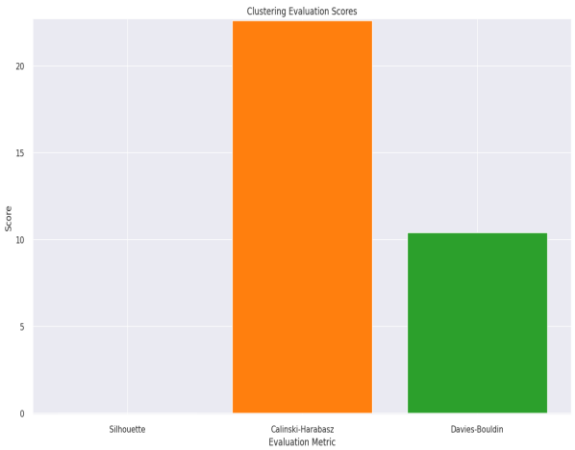


This is the clustering representation of our model. We can see that third cluster is the largest, but we will plot a word cloud map to have the better visual of the output from KNN clustering algorithm.

HIERARCHICAL CLUSTERING



We will select 5 as number of clusters as horizontal line intersects 5

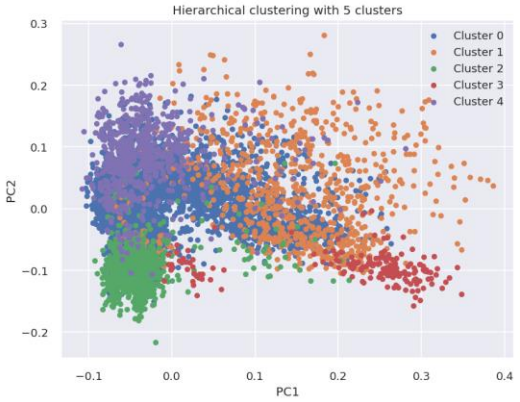


We ran performance evaluation algorithms and got following results.

```
Number of clusters: 5
Silhouette score: -0.0009
Calinski-Harabasz score: 22.5853
Davies-Bouldin score: 10.4136
```

	Evaluation Metric	Score
0	Silhouette Score	-0.000855075
1	Calinski-Harabasz Score	22.5853
2	Davies-Bouldin Score	10.4136

Model Explanation

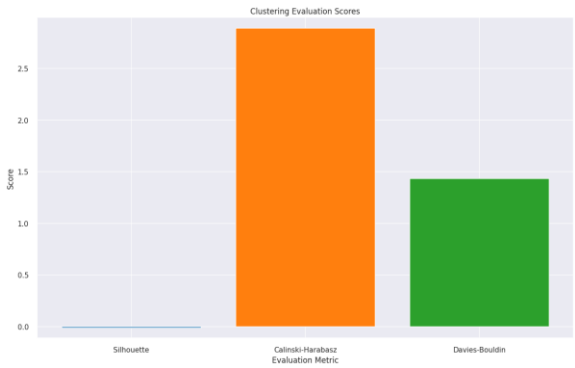


This is the clustering representation of our model. We can see that third cluster is the largest, but we will plot a word cloud map to have the better visual of the output from Hierarchical clustering algorithm.

DBSCAN CLUSTERING

This algorithm doesn't require specifying number of clusters it will automatically consider the number of clusters. Here the algorithm has decided that there can be 18 clusters. We will run the performance

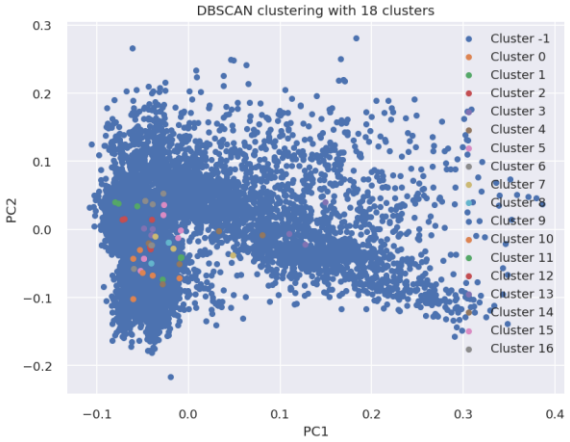
evaluation algorithms to evaluate the model performance.




```
Number of clusters: 18
Silhouette score: -0.0149
Calinski-Harabasz score: 2.8919
Davies-Bouldin score: 1.4322
```

	Evaluation Metric	Score
0	Silhouette Score	-0.0149461
1	Calinski-Harabasz Score	2.8919
2	Davies-Bouldin Score	1.43219

Model Explanation



This is the clustering representation of our model. We can see that third cluster is the largest, but we will plot a word cloud map to have the better visual of the output from DBSCAN clustering algorithm. Maximum density is till 5 clusters hence we will visualize 5 clusters with word cloud

Model Evaluation

	Model	Number of clusters	silhouette_score	calinski_harabasz_score	davies_bouldin_score
0	K-Means Clustering	7	0.007116	25.7653	11.50740
1	Hierarchical Clustering	5	-0.000855	22.5853	10.41360
2	DBSCAN Clustering	18	-0.014946	2.8919	1.43219

- After evaluating multiple machine learning models including K-Means Clustering, Hierarchical Clustering – Agglomerative and DBSCAN Clustering, we selected K-Means Clustering as our final prediction model.
- We chose K-Means Clustering because it performed well on our evaluation dataset in terms of accuracy and computational efficiency. The model was able to cluster similar movies and TV shows together based on their shared attributes, which allowed us to make better recommendations to our users. Additionally, K-Means Clustering was relatively easy to implement and maintain, which made it a practical choice for our project.
- While Hierarchical Clustering - Agglomerative and DBSCAN Clustering also showed promising results, it was computationally expensive and required more processing power and time to execute.
- K- Means Clustering Model gives better results for calinski_harabasz_score (higher than others) and Davies-Bouldin

score(lower than others) than Hierarchical Clustering and DBSCAN Clustering also gives good silhouette score.
- So, we chose K-Means Clustering as our final prediction model for its accuracy, efficiency, and practicality in making recommendations to our users.

Conclusion

CONCLUSION FROM EDA:

- Netflix has more movies than TV shows available on the platform.
- The majority of content on Netflix is suitable for mature audiences, with a TV-MA rating being the most common.
- The United States is the country with the highest number of productions available on Netflix, followed by India and the United Kingdom.

- Netflix has seen a steady increase in its content library since its inception in 2008.
- The most common genre of content on Netflix is Dramas, followed by Comedies and Documentaries.
- The Wordcloud visualization of movie descriptions shows that some of the most common words used in Netflix movie descriptions include love, family, young, life, and world.

CONCLUSION FROM MODEL IMPLEMENTATION:

- The data was clustered based on the attributes: director, cast, country, genre, rating, and description.
- TFIDF vectorizer was used to tokenize, preprocess, and vectorize the values in these attributes, creating a total of 10000 attributes.
- K-Means Clustering algorithm was used to build clusters with the optimal number of clusters being 7 based on the elbow method and Silhouette score analysis.
- Agglomerative clustering algorithm was used to build clusters with the optimal number of clusters being 5 based on the dendrogram visualization.
- DBSCAN clustering was built and it gives optimal number of clusters as 18 with very less metric score.

References-

1. Analyticsvidya
2. GeeksforGeeks
3. Stackoverflow
4. Kaggle