Capstone Project

NETFLIX MOVIES & TV SHOWS CLUSTERING

Mihir Kulkarni

Ram Bakale

CONTENT

ABSTRACT

- The goal was to predict clusters similar content by matching text-based features.
- Exploratory Data Analysis is done on the dataset to get the insights from the data but first null values handled. Also, some hypothesis testing also performed from the insights from EDA. After that description column is our target variable has to be feature engineered where NLP operations such as removing symbols, stop words, punctuations, tokenizing performed on it and after that vectorized by using TFIDF. After that all left was to find the clusters and fitted our models by knowing number of clusters and further the model is evaluated using the metrics.

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.

  So, the goal is to predict clusters by similar content by matching text-based features whichever case is the description column which is a small plot summary of the contents.

HANDLING NULL VALUES

Treatment of Null Values

- Filling the rows which has higher than 5% null and lower than 30% null values.
- Dropping the rows which has lower than 5% null values.
- Dropping the column which has null values higher tha 30%.

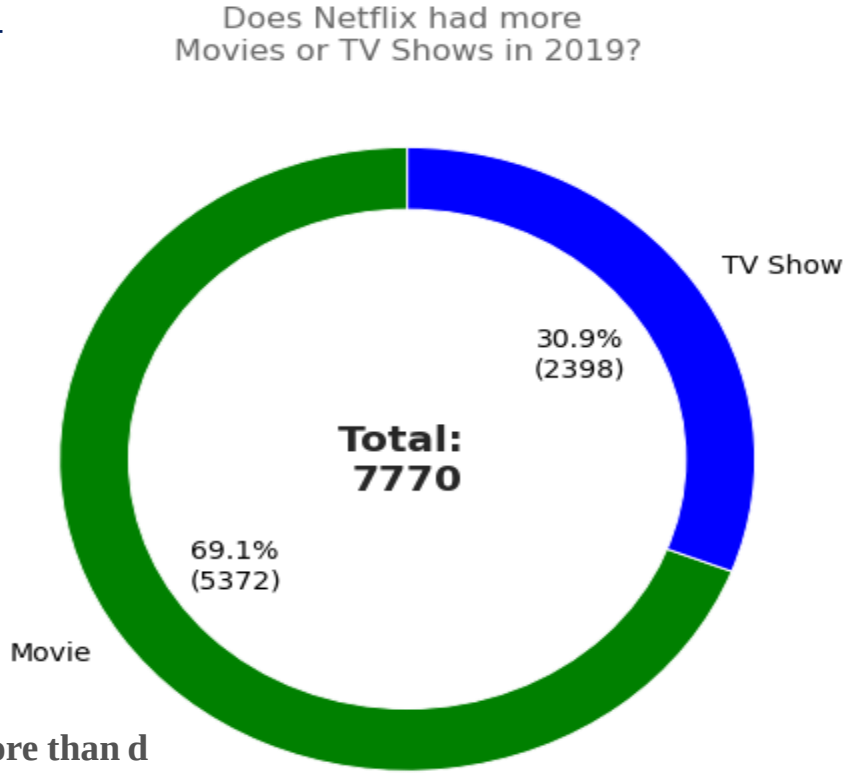
DATA MANIPULATION

Added year and month column obtained from year date added column.
Assigned the Ratings into grouped categories.
Such as –

- **ratings_ages = {**
- **'TV-PG': 'Older Kids’, 'TV-MA': 'Adults’, 'TV-Y7-FV': 'Older Kids’, 'TV-Y7': 'Older Kids’, 'TV-14': 'Teens', 'R': 'Adults’, 'TV-Y': 'Kids’, 'NR': 'Adults’, 'PG-13': 'Teens', 'TV-G': 'Kids’, 'PG': 'Older Kids’, ’G': 'Kids’, 'UR': 'Adults', 'NC-17': 'Adults’}**

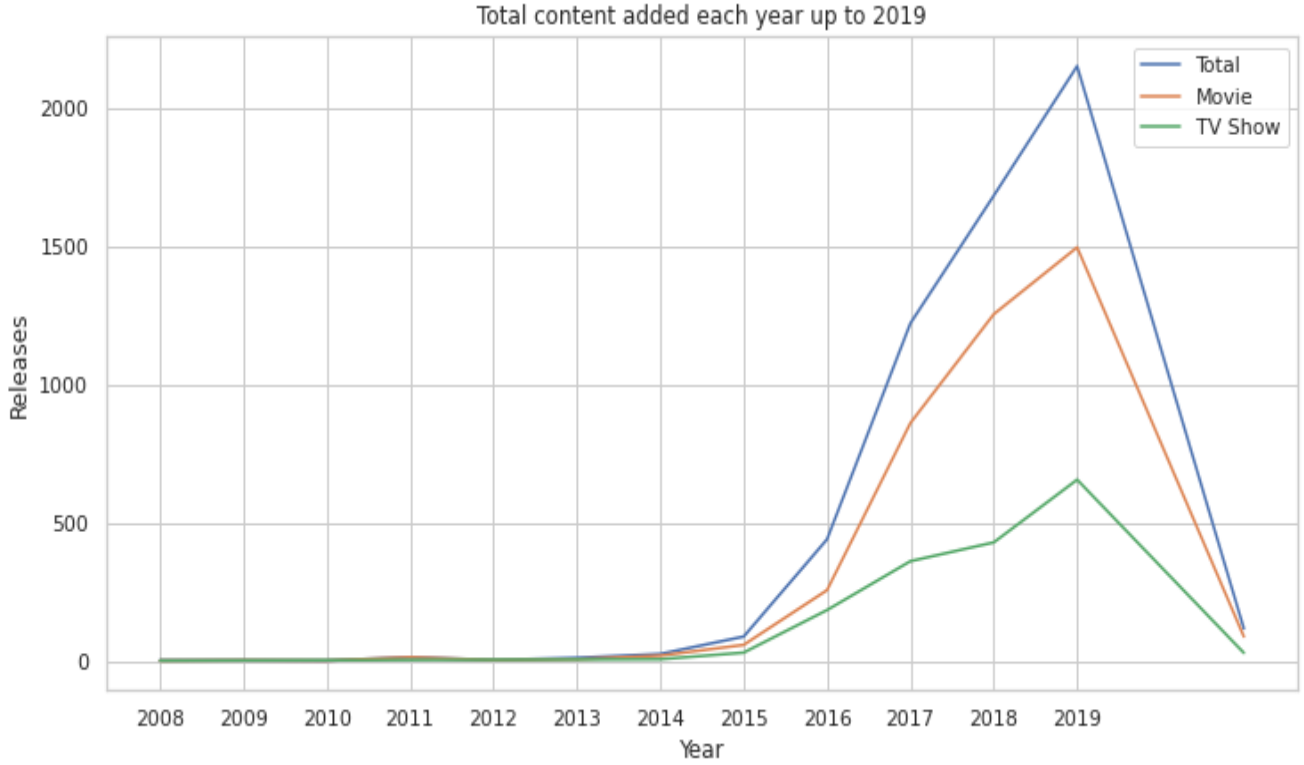
Exploratory Data Analysis

Distribution of type of contents :-



No. of movies content is more than d ouble than Tv show content

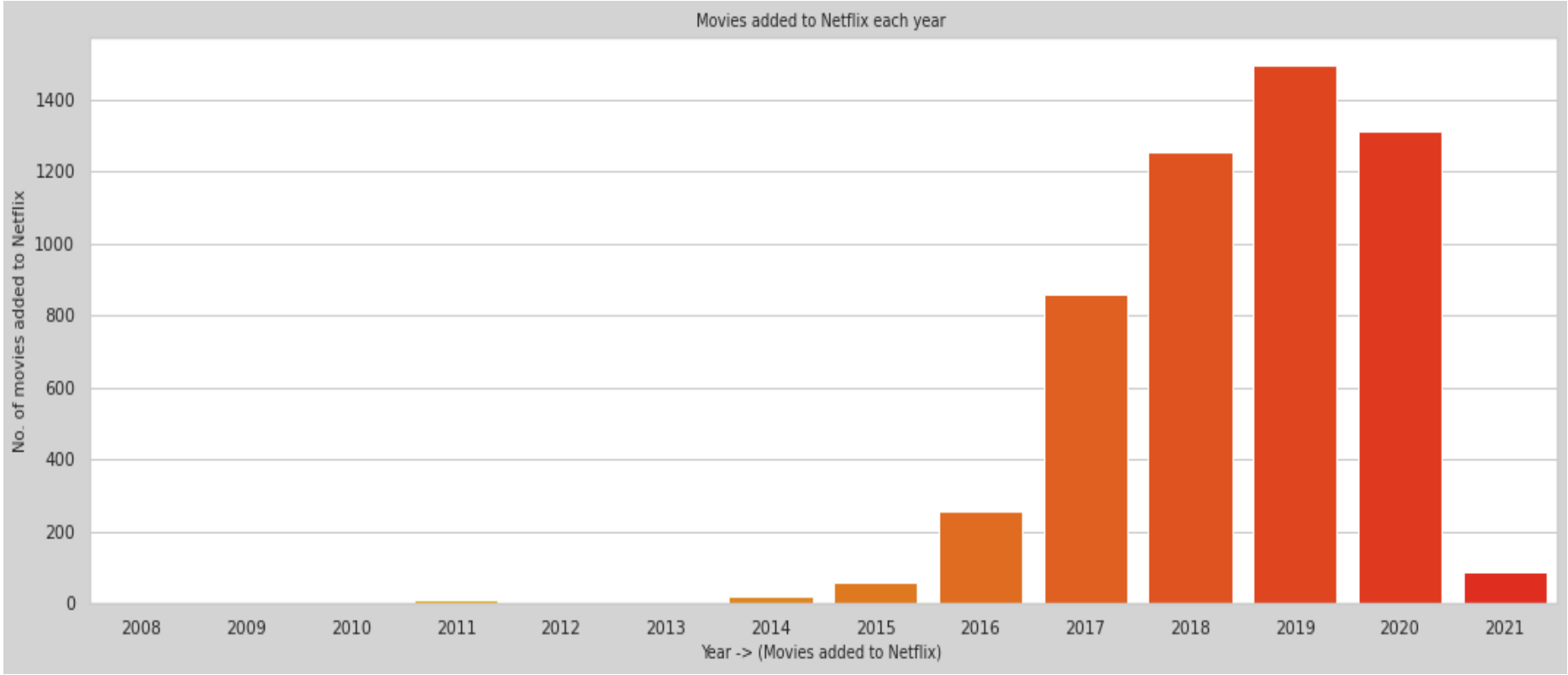
Exploratory Data Analysis



We can see from plot that since 2014 the amount of content added has been tremendous.

Exploratory Data Analysis

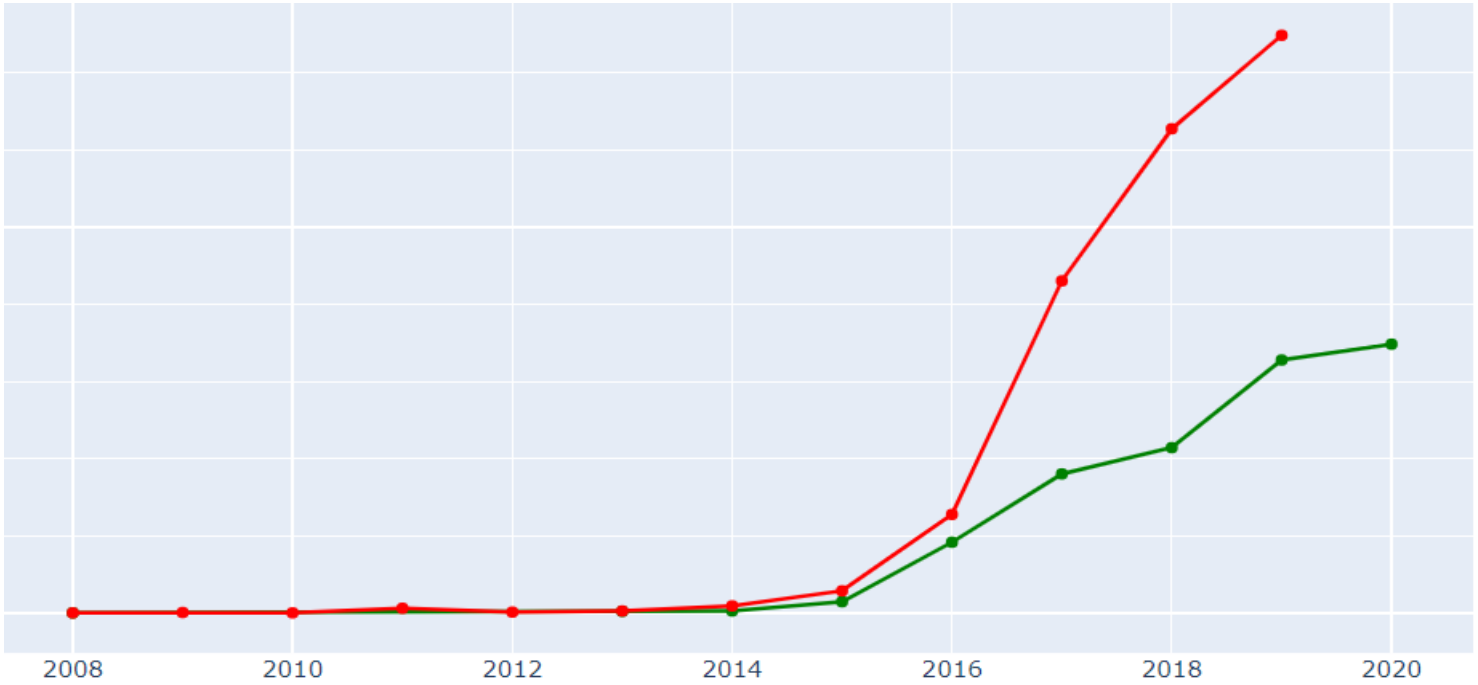
No of movies added yearly.



We can see from graph that popularity of OTT has boomed in last 5 years.
Highest no. of movies and TV shows were added in 2019.

Exploratory Data Analysis

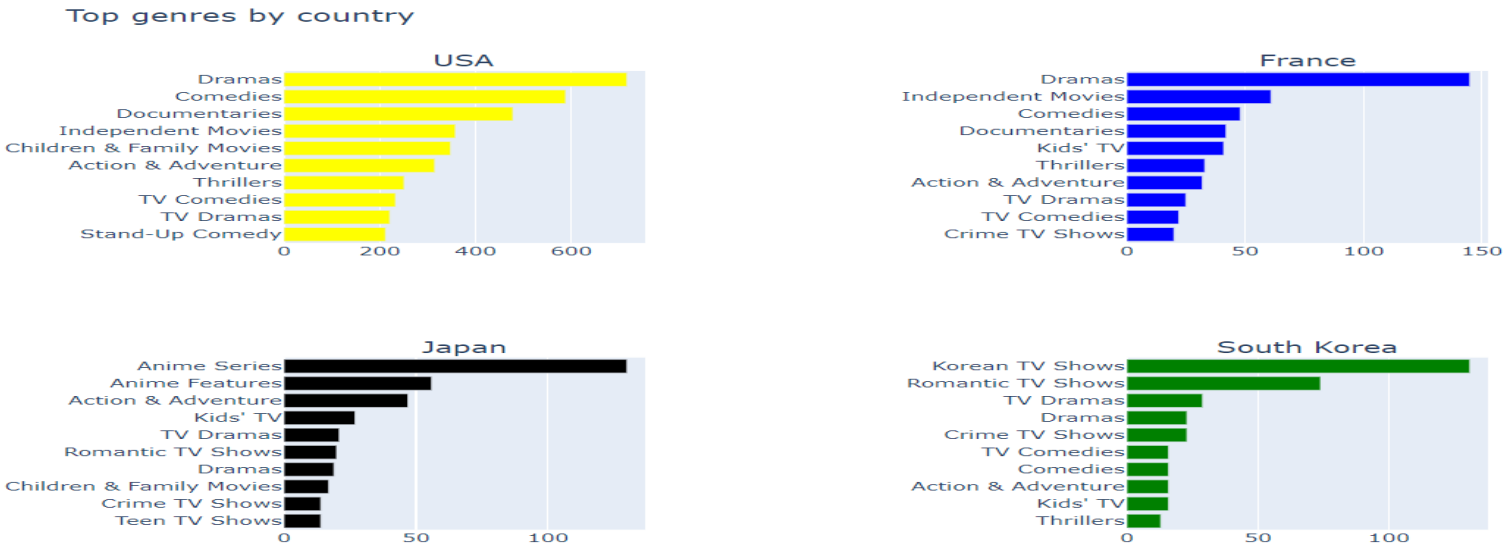
Is Netflix has increasingly focusing on TV rather than movies in recent years.



The growth in number of movies on Netflix is much higher than that of TV shows

Exploratory Data Analysis

What type of content is available in different countries?



The top 10 genre of Netflix movies varies from country to country which reflects different tastes of audience from different countries.

Feature Engineering

To perform Text Clustering and to cluster all the content :-

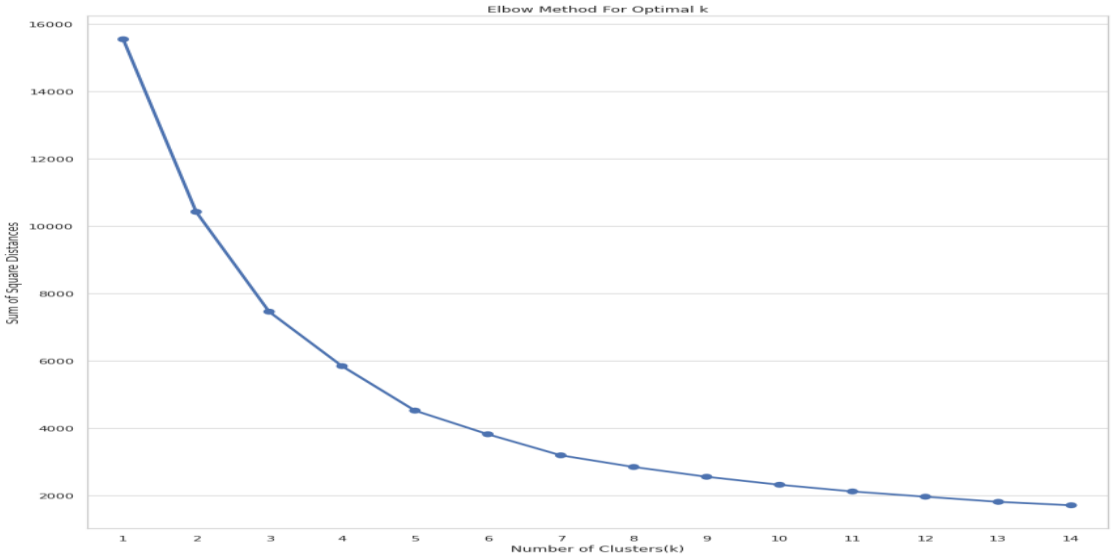
- Converted all the text to lower case
- Handled all the URLs
- Handled all the symbols
- Tokenized the text
- Removed Punctuation, and stop words
- After all that transformed the text by using TFIDF Vectorizer.

Finding Number of clusters

To find the number of clusters we used elbow method and Silhouette's score.

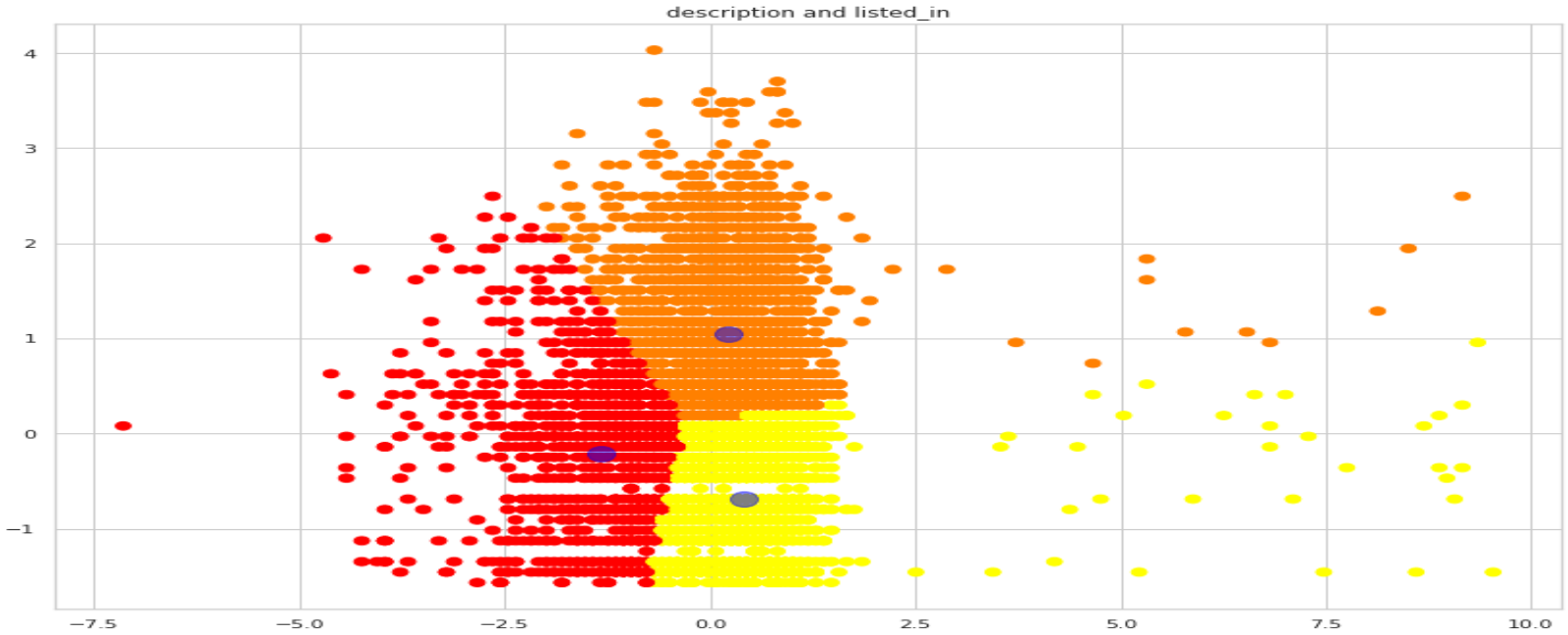
After visualizing both, the best optimal number of clusters was 3.The elbow method uses sum of squared distances to choose an ideal value of no. of clusters.

Elbow Method



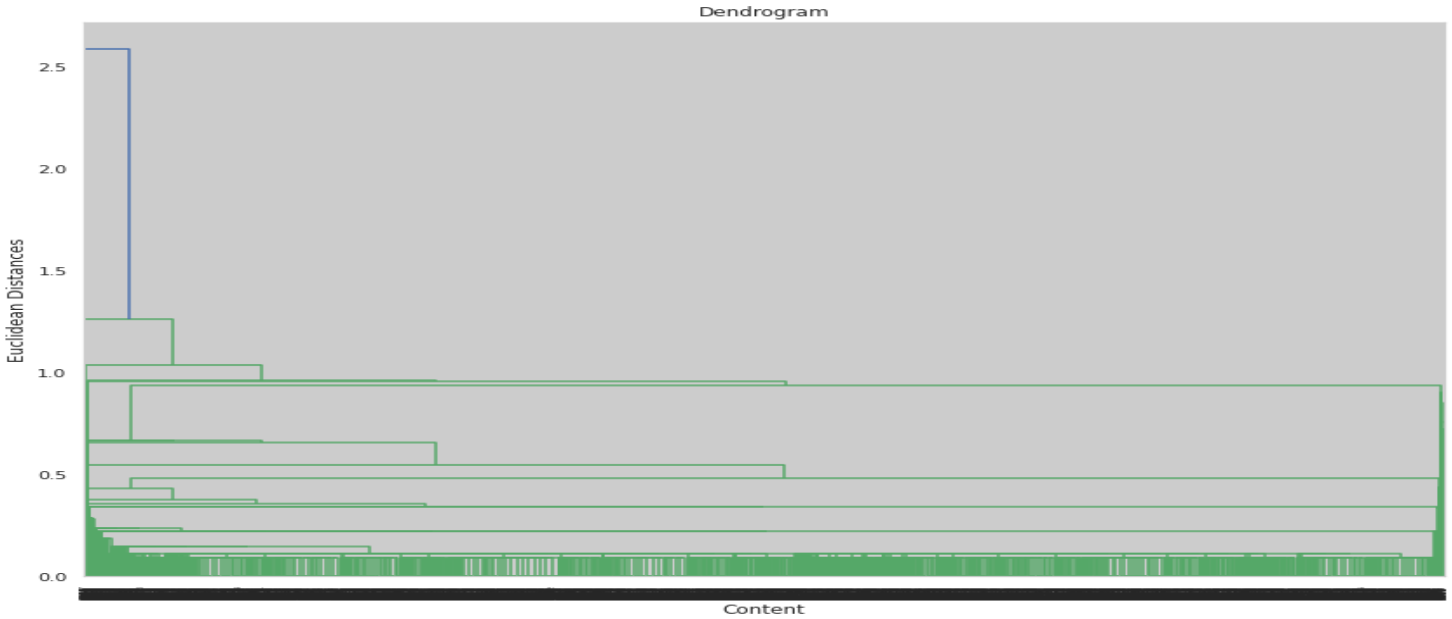
Implementing K Means Clustering Method

We have used elbow method for text based clustering, where we will use 3 clusters
As we can see from the fig 3 clusters are identified.



Hierarchical Clustering

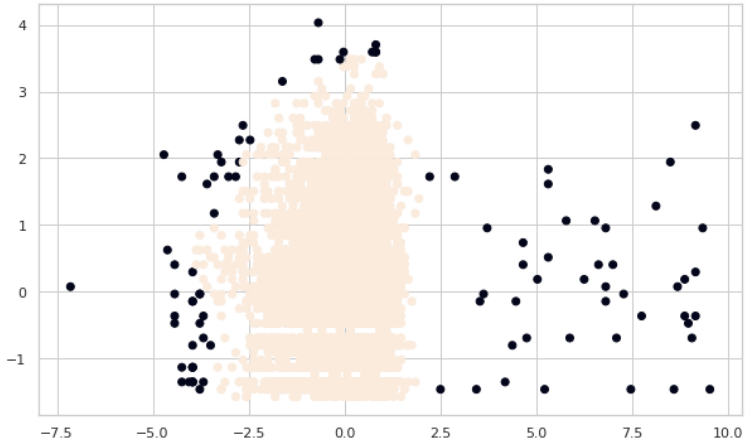
Finding number of clusters using Dendogram after visualizing it Optimal Number of clusters I got was 3.



Number of clusters will be number of vertical lines which are intersected by the line drawn using threshold.

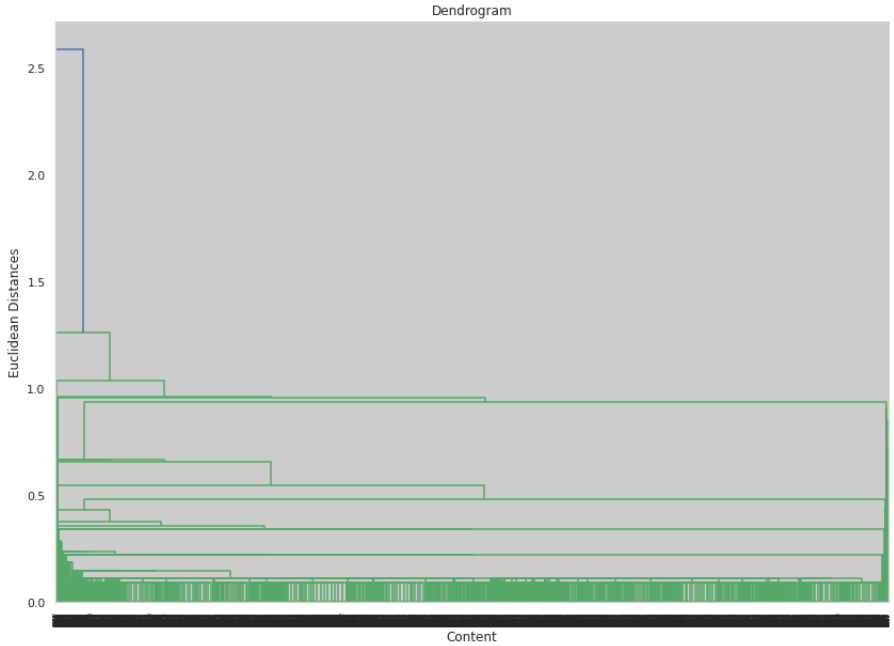
DBSCAN

DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density. Given that DBSCAN is a density based clustering algorithm, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. DBSCAN can sort data into clusters of varying shapes as well, another strong advantage.



Dendrogram

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.



Conclusion

We have reached at the end of our project. Few questions asked which we have to understand from EDA such as what type of content is available in different countries, so to answer that Movies are still more than TV Shows, another question was that whether Netflix is focusing more on TV Shows in recent years or not so again this question was answered from EDA and the answer is No, as from 2016 still there are a lot of Movies added than TV Shows. We can say K Means clustering algorithm fits well in our case with better evaluation metrics.

By using Silhouette's score and Elbow Method we generated optimal of 3 clusters for K Means and from Dendrogram 3 clusters were generated.

In both cases, one clusters has more than 3000 points whereas other points were unevenly distributed.

For Tfidf K Means is giving best results so K Means is our final model.

THANK YOU