Capstone Project

NETFLIX MOVIES & TV SHOWS CLUSTERING

Saransh Srivastava

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 whichour case is the description column which is a small plot summary of the contents.

HANDLING NULL VALUES

- show_id 0.000000
- type 0.000000
- title 0.000000
- director 30.679337
- cast 9.220496
- country 6.510851
- date_added 0.128419
- release_year 0.000000
- rating 0.089893
- duration 0.000000
- listed_in 0.000000
- description 0.000000

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%.

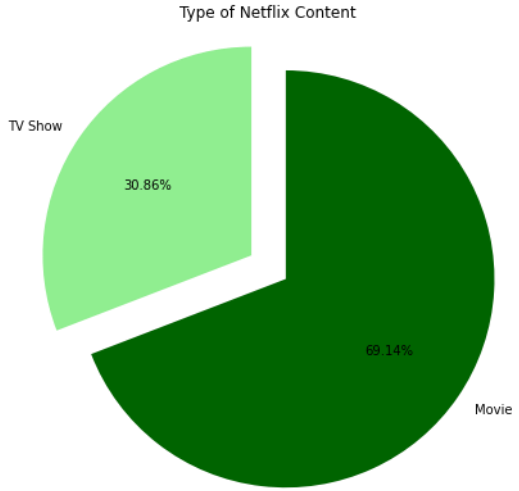
DATA MANIPULATION

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’}

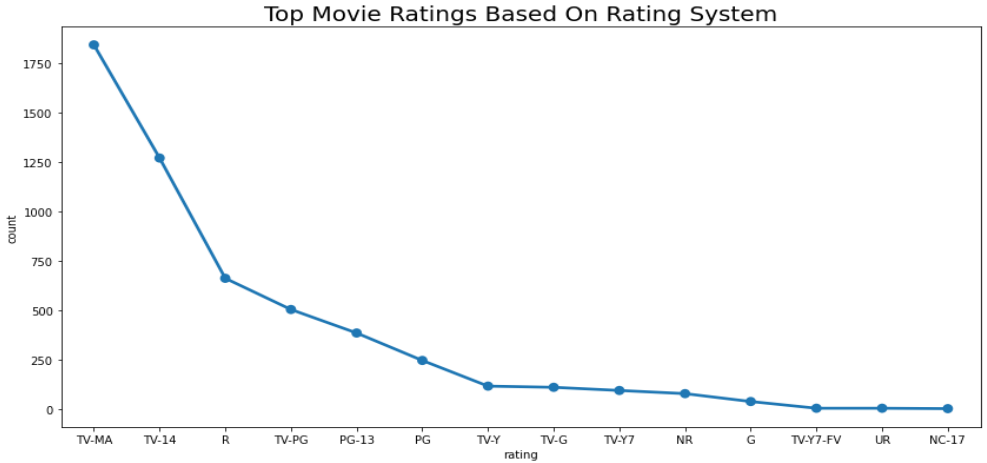
EDA

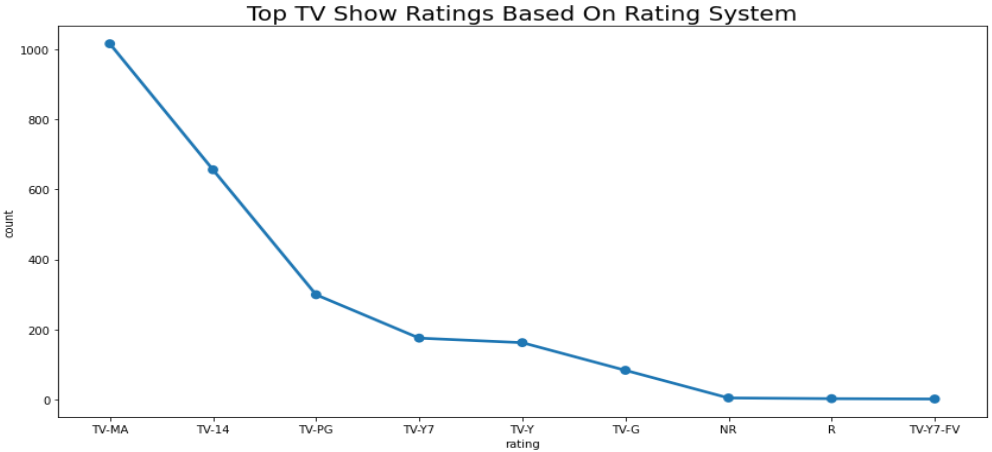
Distribution of type of contents :-



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

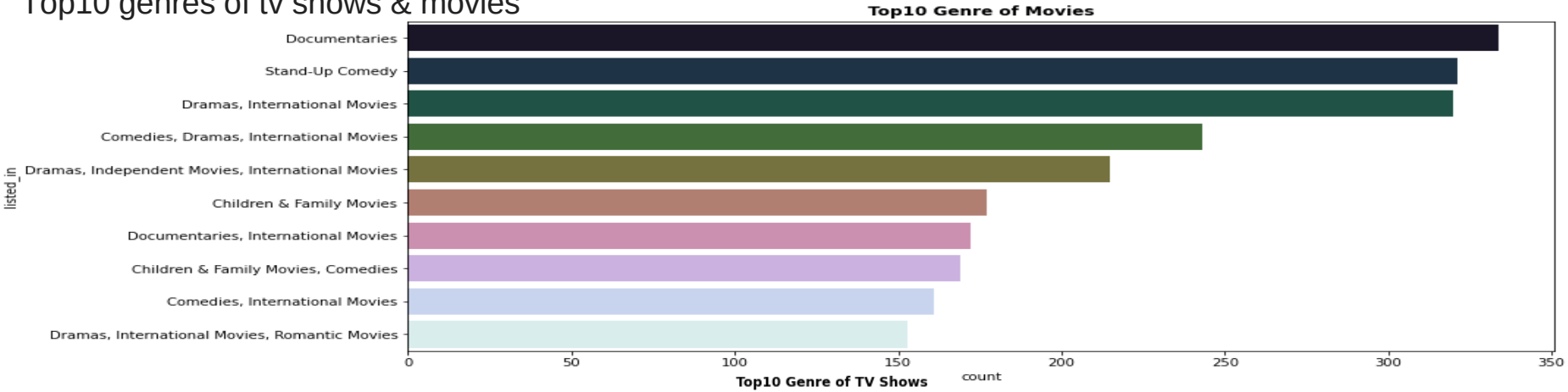
Top Movie & TV SHOWS Ratings Based On Rating System

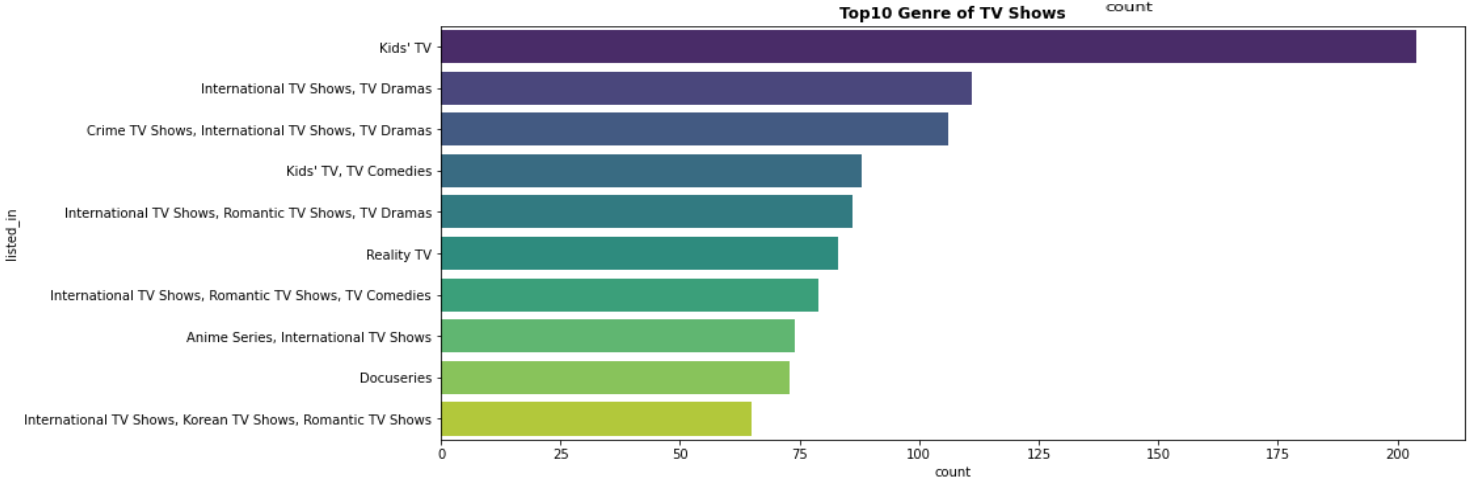




EDA

Top10 genres of tv shows & movies

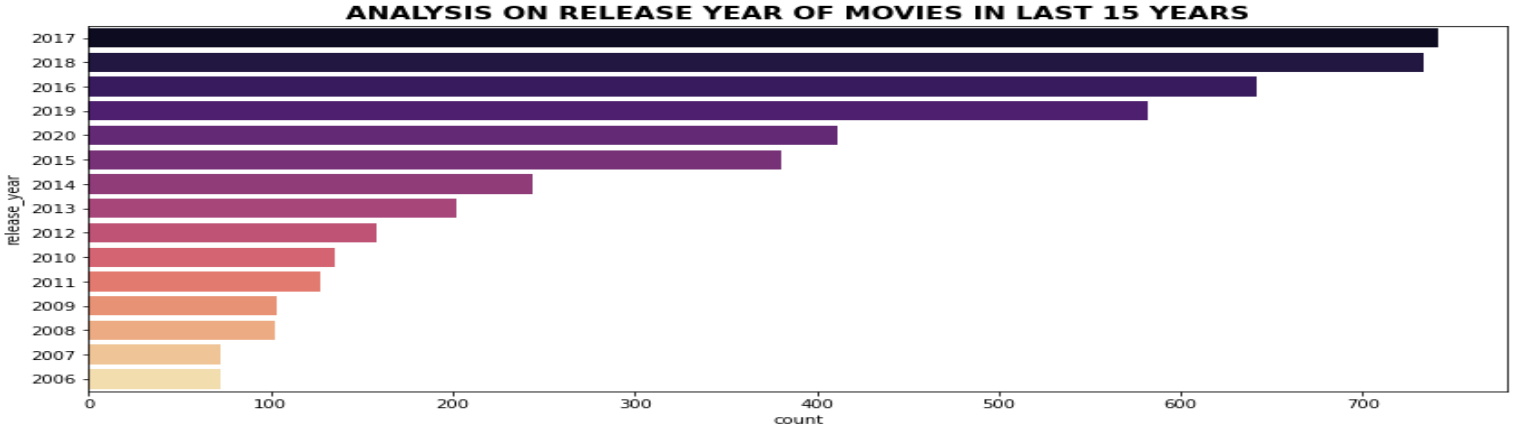


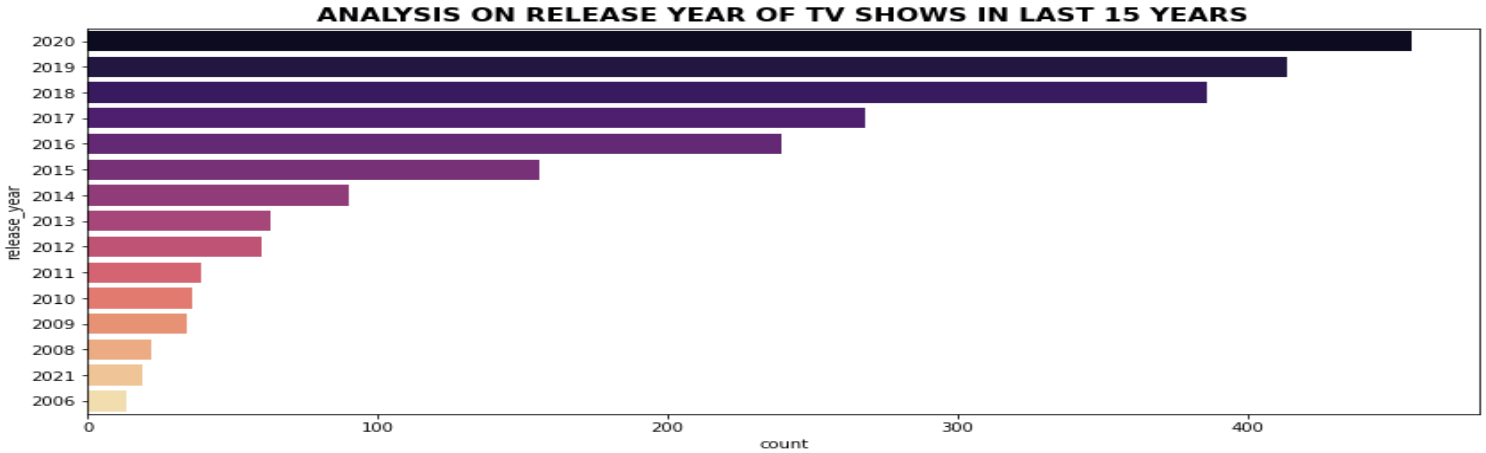


Number of documentaries and Kids' TV content is highest

EDA

Analysing at what year most number of movies and tv shows released in the past 15 years

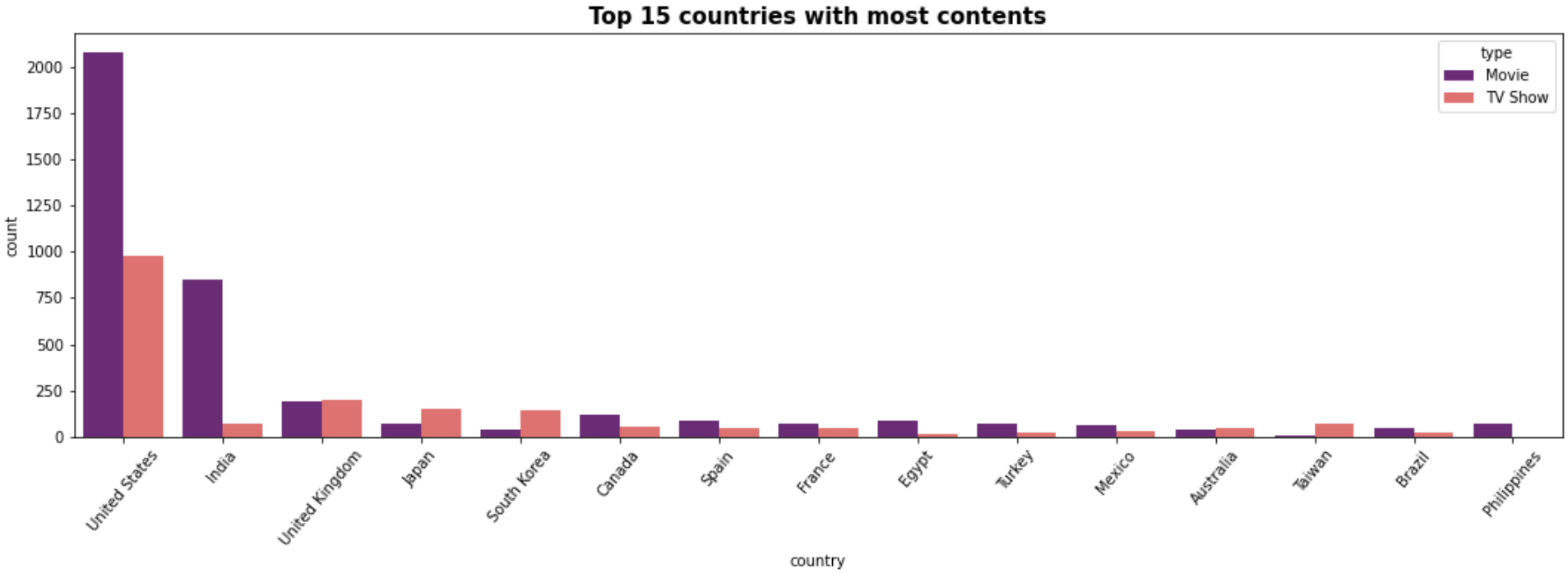




Most no. of movies released was in 2017 and Tv Shows was in 2020 in our dataset.

EDA

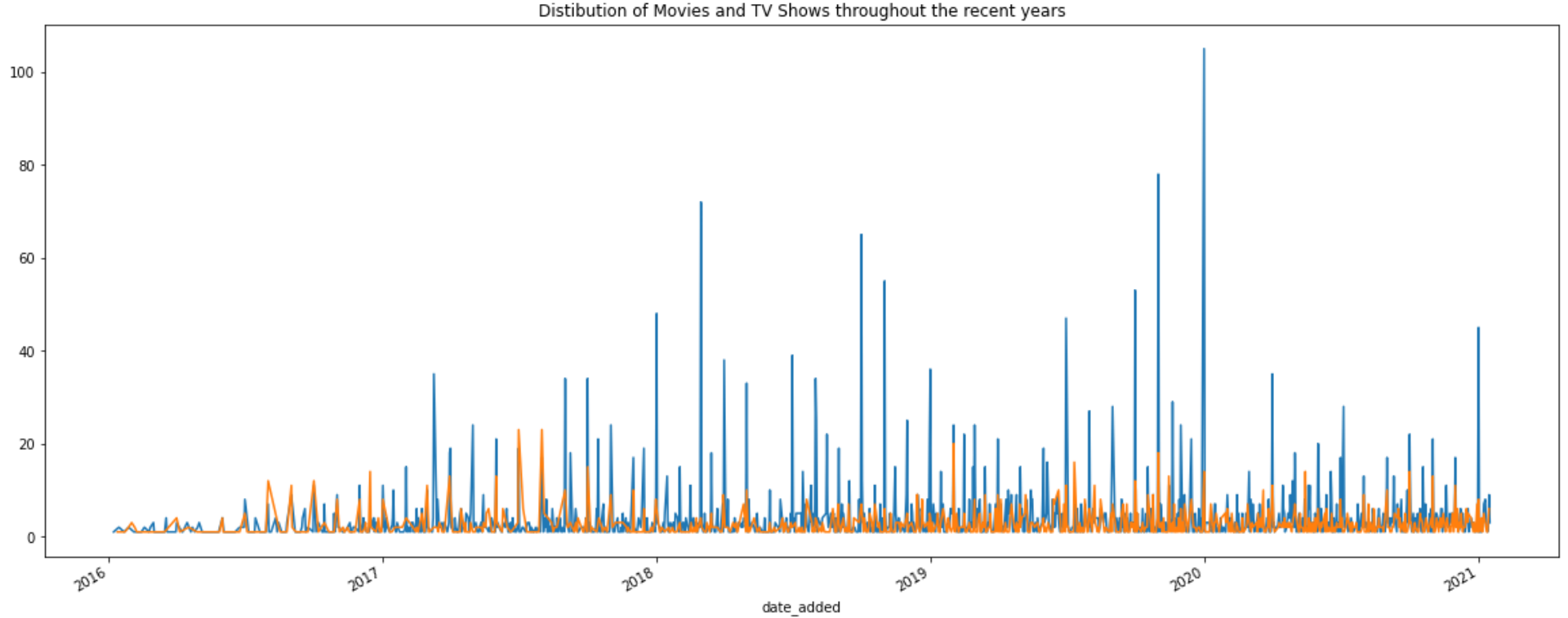
Understanding what type content is available in different countries



Movies are more available in different countries than TV SHOWS

EDA

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

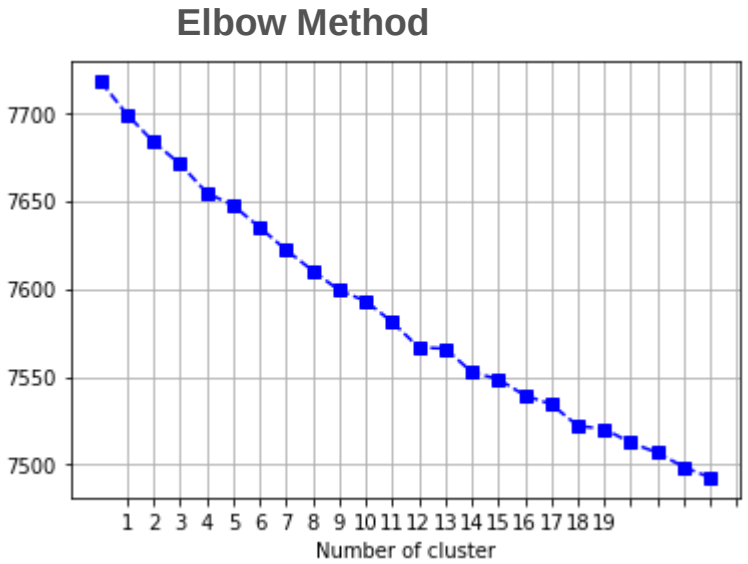
Feature Engineering

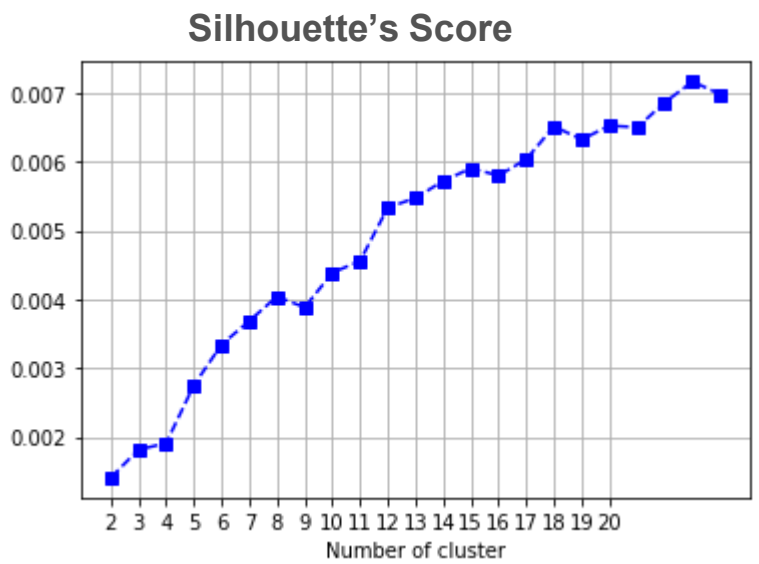
To perform Text Clustering and to cluster all the content :-

- 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 I used elbow method and Silhouette's score. After visualizing both, the best optimal number of clusters was 20.



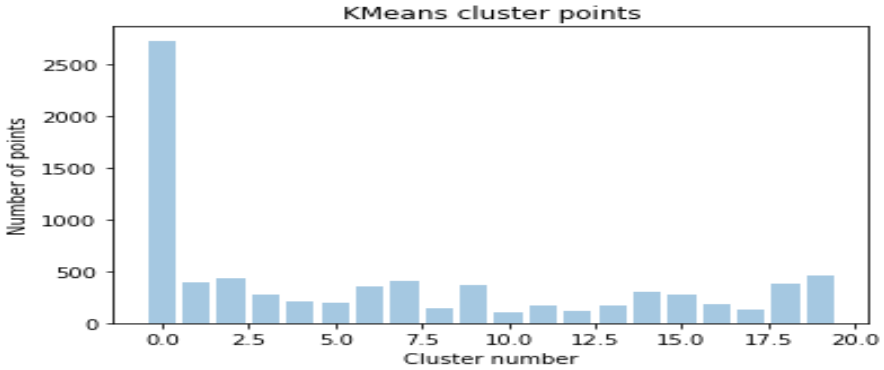


Implementing K Means Clustering Method

After fitting our model with 20 clusters in our tfidf array,
I've created another column where each row is assigned to their separate clusters.
After assigning clusters most of the number of points were assigned to cluster number 0 and rest of points were not evenly distributed among all the clusters.

Evaluation of our K Means model were –
Silhouette's score was – 0.00622
Calinski Harabasz score – 10.073
Davies Boulden Index- 10.065

Continued

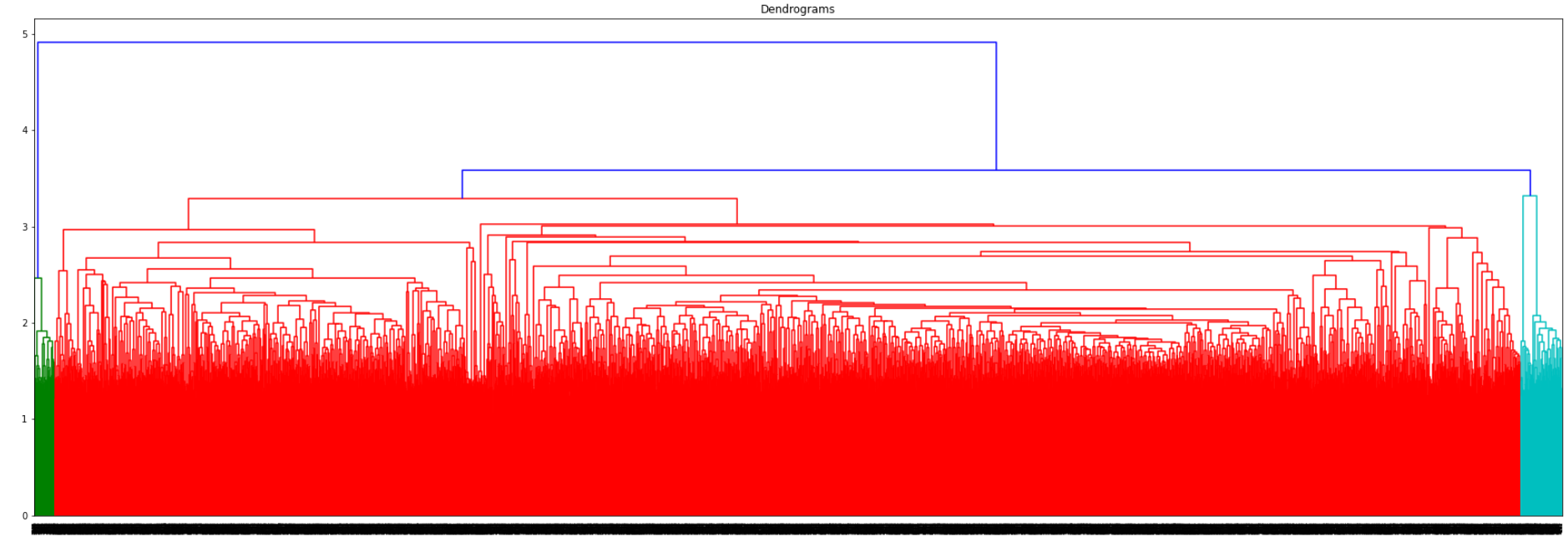


Titles of contents of cluster number 0, here we can see the most of the genres are similar here.

Show_id	title	listed_in
0	3%	International TV Shows, TV Dramas, TV Sci-Fi &...
2	23:59	Horror Movies, International Movies
5	46	International TV Shows, TV Dramas, TV Mysteries
17	22-Jul	Dramas, Thrillers
18	15-Aug	Comedies, Dramas, Independent Movies

Hierarchical Clustering

Finding number of clusters using Dendogram after visualizing it Optimal Number of clusters I got was 6.



After going with 6 clusters and fitting our variable in Agglomerative Clustering, after that predicting using our model here also I've created a separate column where each row is assigned to their respective clusters.

Continued

After using the evaluation metrics on our model
Silhouette Coefficient was -0.02
Calinski Harabasz score was 6.87
Davies Boulden Index was 15.43

Comparing it with our K Means model only Davies Boulden Index is better for Hierarchical Clustering. Here also our most number of points were assigned by our model to cluster number 0 and rest of the points were unevenly distributed among all the 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 20 clusters for K Means and from Dendrogram 6 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.

THANKYOU