Capstone Project

NETFLIX MOVIES & TV SHOWS CLUSTERING

Ajinkya Shingote

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 fromthe 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

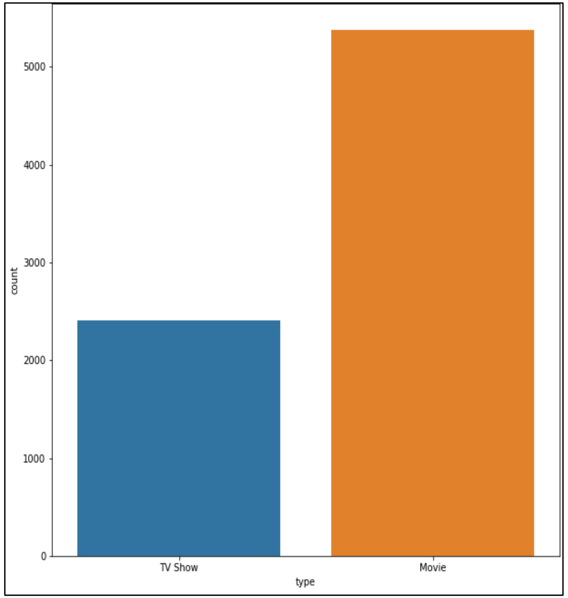
- We will need to replace blank countries with the mode (most common) country. It would be better to keep director because it can be fascinating to look at a specific filmmaker's movie. As a result, we substitute the null values with the word 'unknown' for further analysis.
- There are very few null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

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.

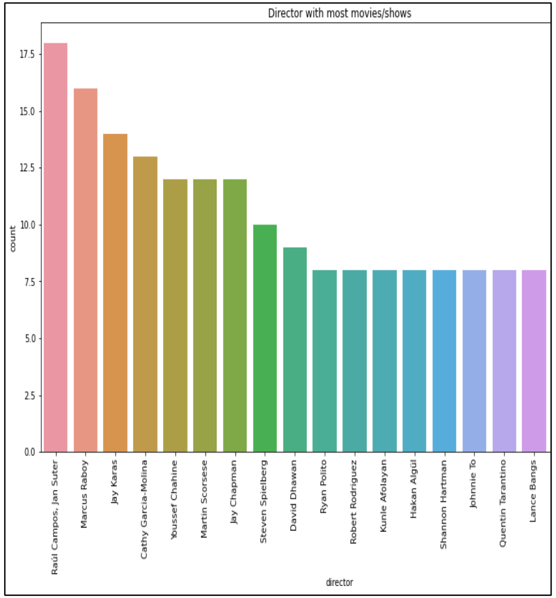
- There are very few null entries in the date_added fields thus we delete them.

- Duplicate values dose not contribute anything to accuracy of results.
- Our dataset dose not contains any duplicate values.

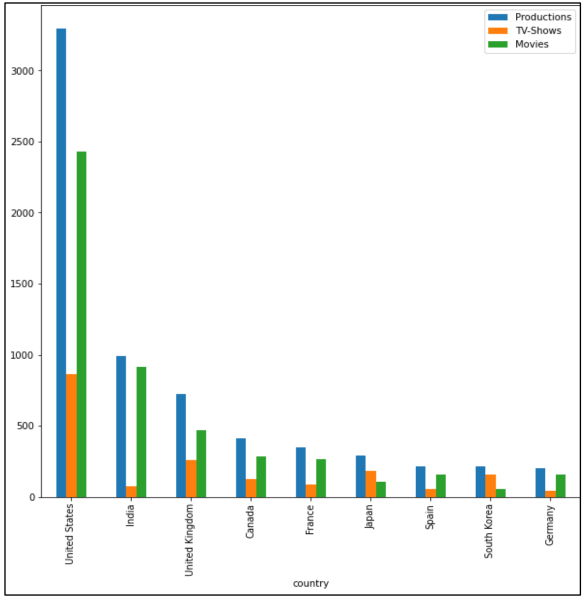
Exploratory Data Analysis



The above figure is count plot of movie and tv shows and the from the visualization we can draw the conclusion that almost 70% of datapoints belong to movie, rest 30% to TV show.

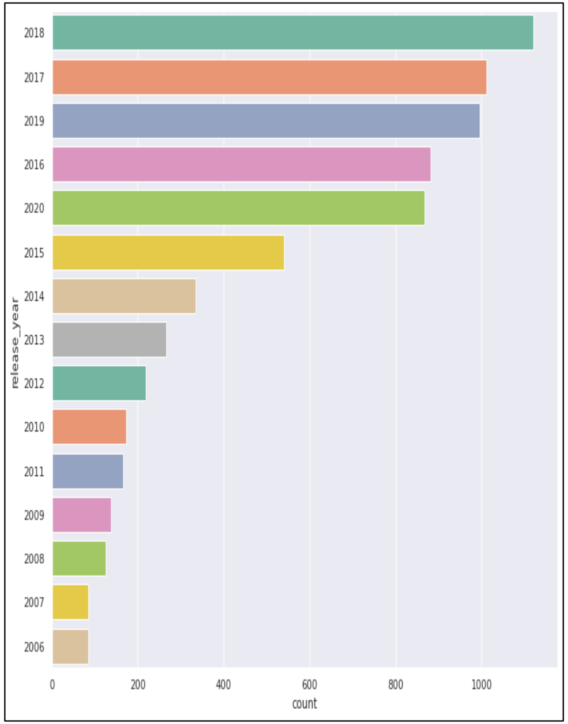


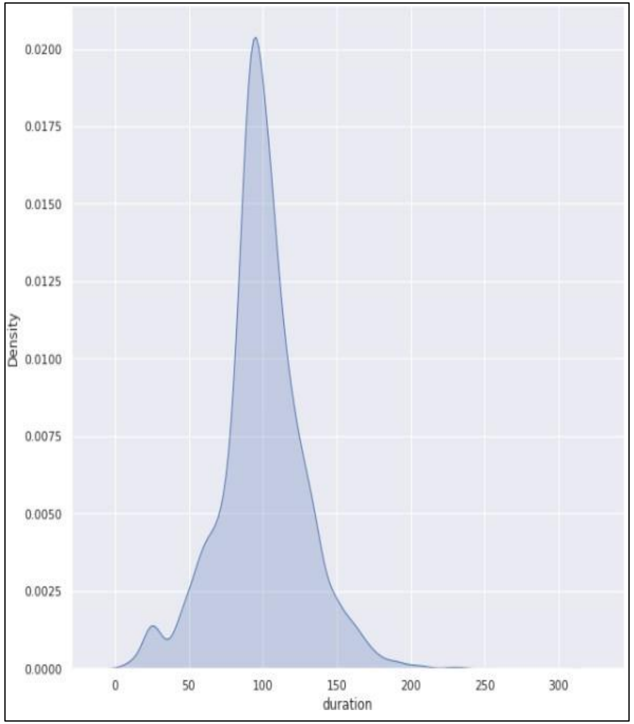
The above figure is a bar plot of count vs director which tells about director with the greatest number of movies or tv shows. The visualization depicts that Raul Campos, Jan Suter are the director with most tv shows or movies followed by Marcus Raboy.

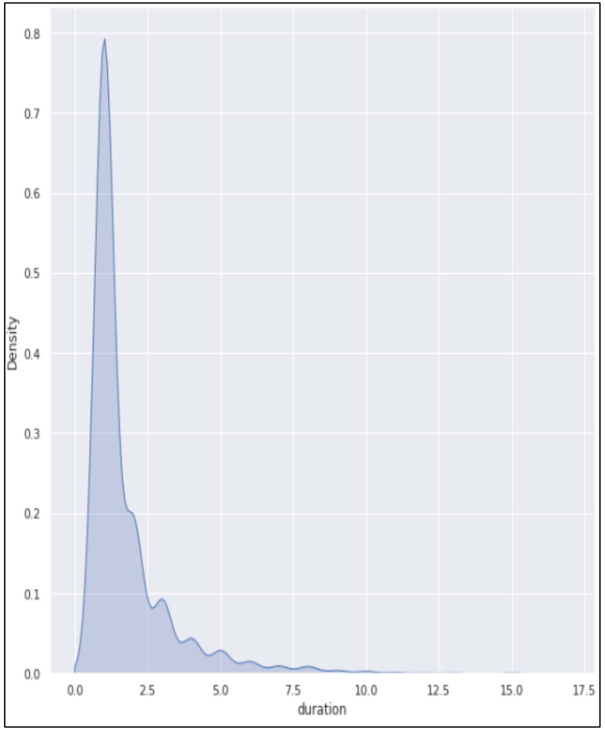


The top countries with Tv shows and movies along with production is plotted and the from visualization it is clear that United States tops the chart followed by India and Germany has the least TV shows and movies with a smaller number of productions.

Continued…..





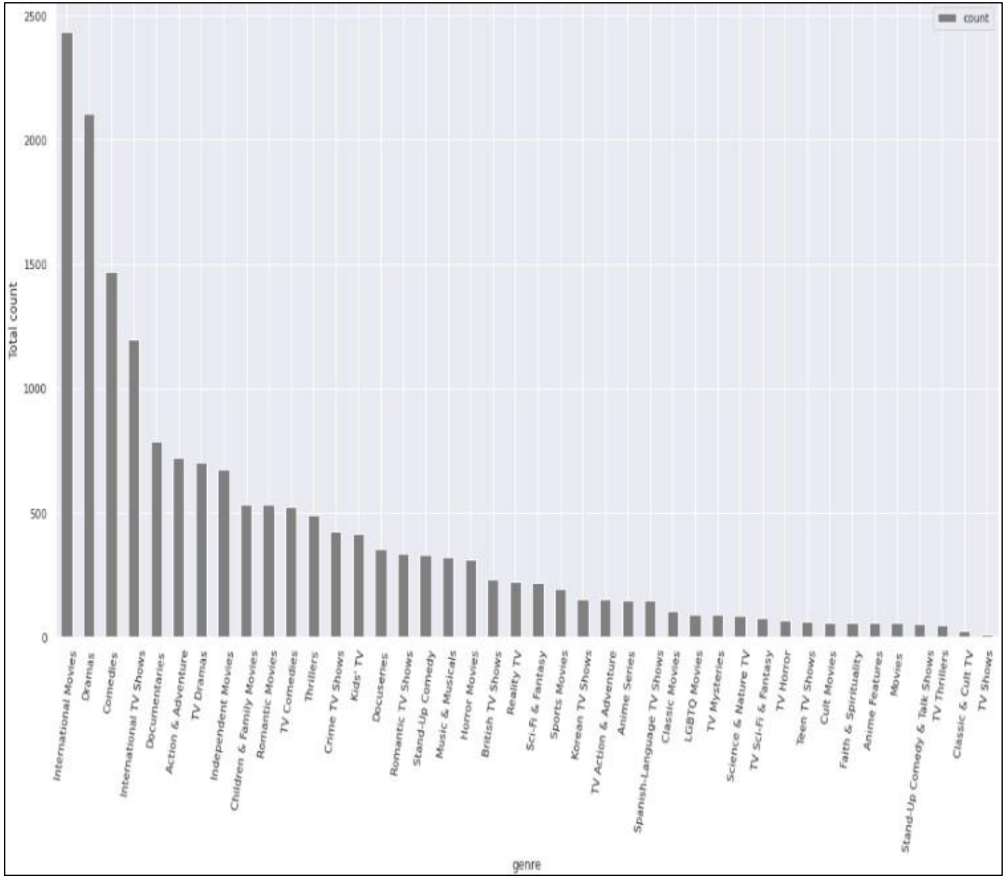


From the count plot it is clear that year 2018 is the year with the greatest number of movies and TV shows with an approximate count value of 1400 and the least movies and TV shows in the year 2006.

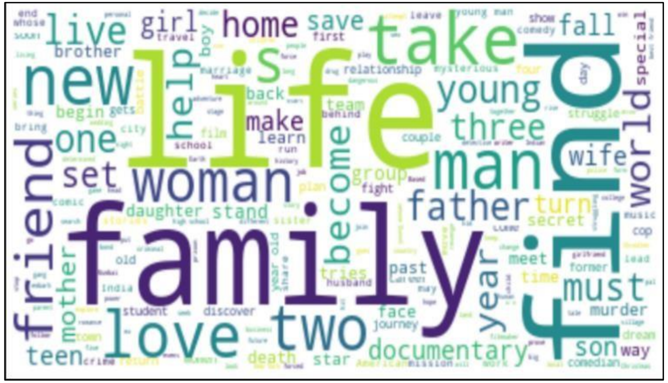
The below plot is a density plot for duration of movies and from the plot it is clear that most of the content is about 70 to 120 minutes duration for movies.

The below plot is a density plot for duration of show for no of seasons and from the plot it is clear that most of the shows are 1 to 2 seasons long.

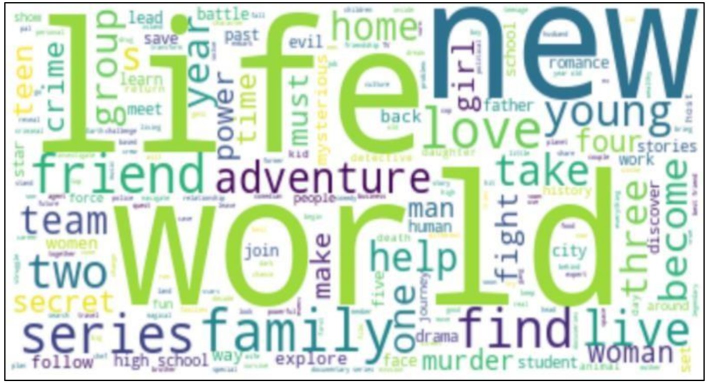
Continued…..



From the above count plot for genre distribution, it is pretty evident that 'international movies' is the genre with highest count followed by 'Dramas' and the least is tv shows.



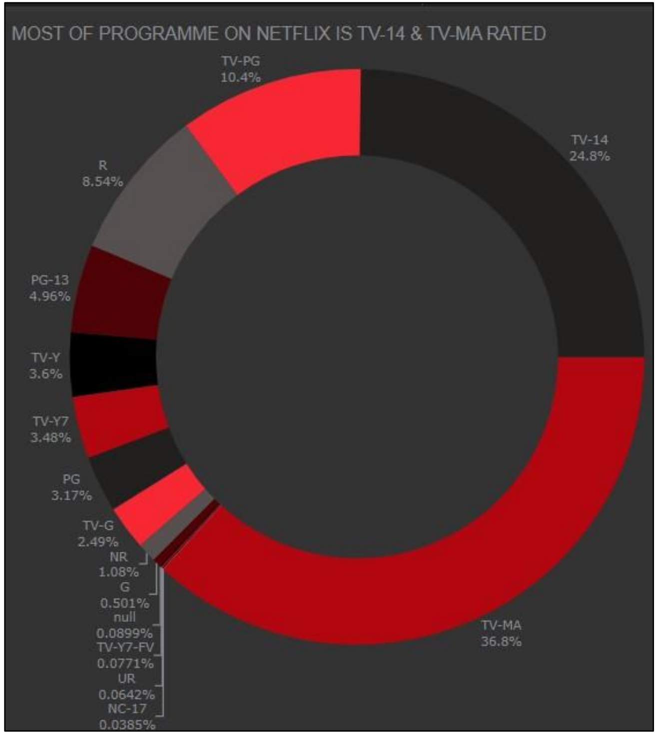
From the word cloud for movies most words like life, family popped up.

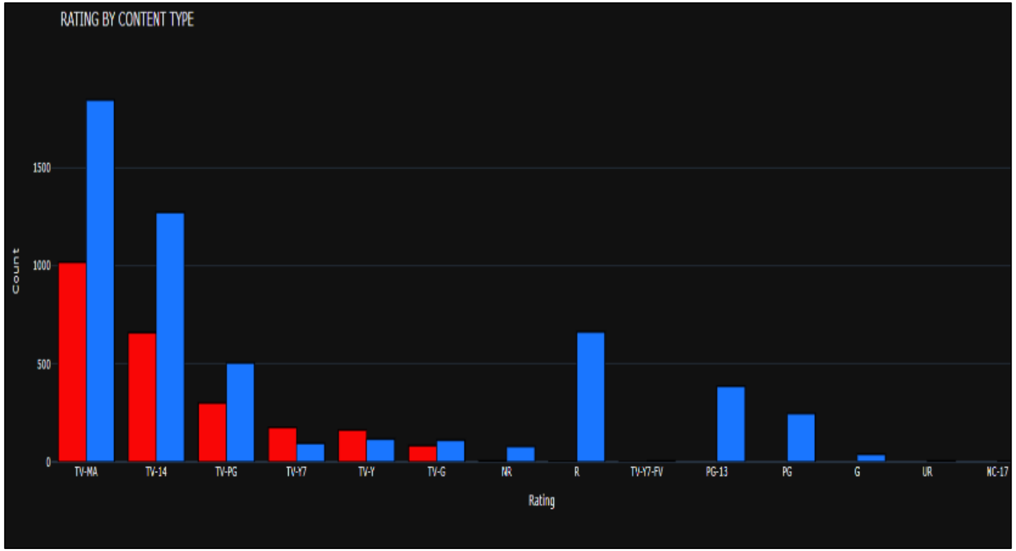


From the word cloud for shows most words like 'life','world','new','adventure','friend','family' popped up.

Continued…..

- Rating distribution by content type

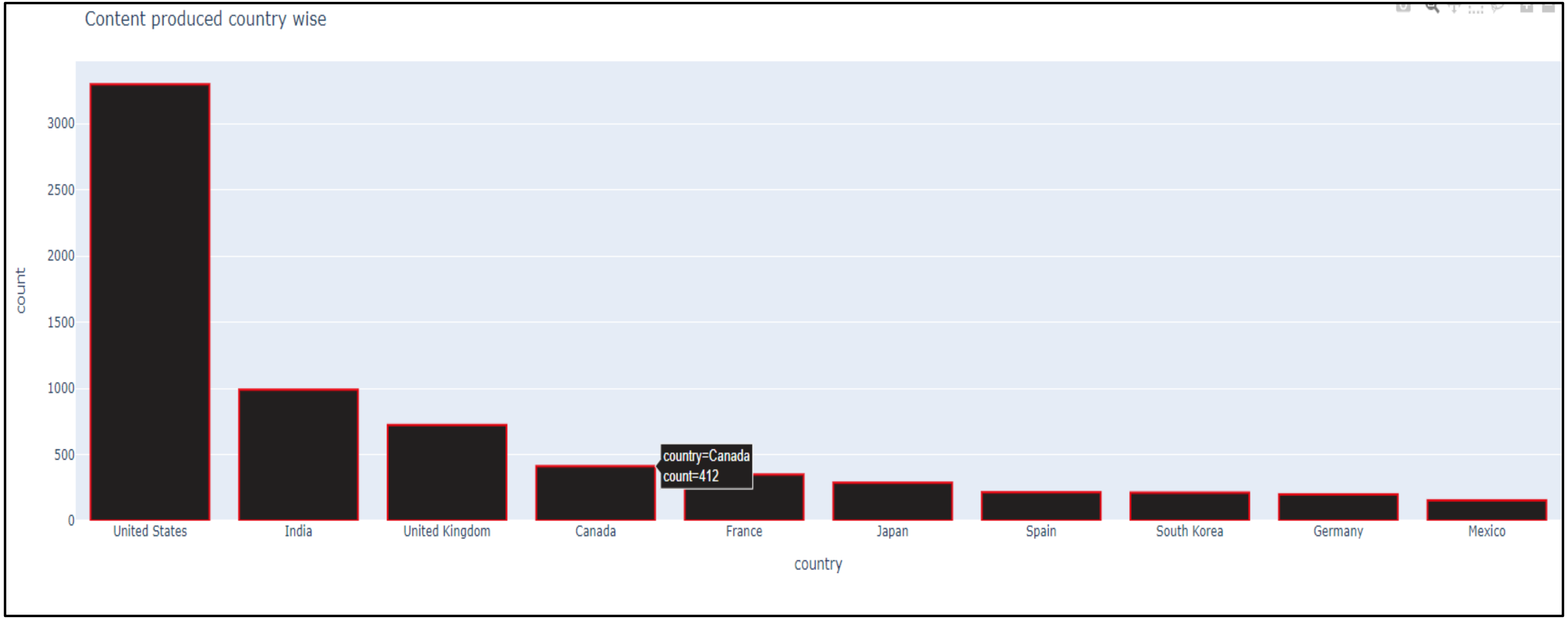




- From the pie chart it is clear that most of the programs on Netflix are TV-14 and TV-MA rated we can say that more content with mature content is available on Netflix.

Continued…..

- Content produced country wise



- From the above figure united states is the country that has produced the most content.

Continued…..

- Genre distribution by content type

	count
International Movies	2437
Dramas	2106
Comedies	1471
International TV Shows	1199
Documentaries	786
Action & Adventure	721
TV Dramas	704
Independent Movies	673
Children & Family Movies	532
Romantic Movies	531
TV Comedies	525
Thrillers	491
Crime TV Shows	427
Kids' TV	414
Docuseries	353
Romantic TV Shows	333
Stand-Up Comedy	329
Music & Musicals	321
Horror Movies	312

British TV Shows	232
Reality TV	222
Sci-Fi & Fantasy	218
Sports Movies	196
Korean TV Shows	150
TV Action & Adventure	150
Anime Series	148
Spanish-Language TV Shows	147
Classic Movies	103
LGBTQ Movies	90
TV Mysteries	90
Science & Nature TV	85
TV Sci-Fi & Fantasy	76
TV Horror	69
Teen TV Shows	60
Cult Movies	59
Faith & Spirituality	57
Anime Features	57
Movies	56
Stand-Up Comedy & Talk Shows	52
TV Thrillers	50
Classic & Cult TV	27
TV Shows	12

- From the above table we can see genre of international movie is the most available content.

Continued…..

- Top 20 Countries with more number of productions

	country	Productions	TV-Shows	Movies
0	United States	3297	866	2431
1	India	990	75	915
2	United Kingdom	723	256	467
3	Canada	412	126	286
4	France	349	84	265
5	Japan	287	184	103
6	Spain	215	57	158
7	South Korea	212	157	55
8	Germany	199	42	157
9	Mexico	154	53	101
10	China	147	45	102
11	Australia	144	60	84
12	Egypt	110	13	97
13	Turkey	108	28	80
14	Hong Kong	102	5	97
15	Italy	90	23	67
16	Brazil	88	29	59
17	Belgium	85	11	74
18	Taiwan	85	70	15
19	Argentina	82	18	64

Data preprocessing

Removing punctuation:
Punctuation has no meaning in clustering, so removing punctuation helps to get rid of useless bits of data or noise.

Removing stop words:
Stop-words are basically a collection of commonly used words in any language, not just English. If we remove words that are very commonly used in a given language, we can focus on important words instead.

Clusters Model Implementation

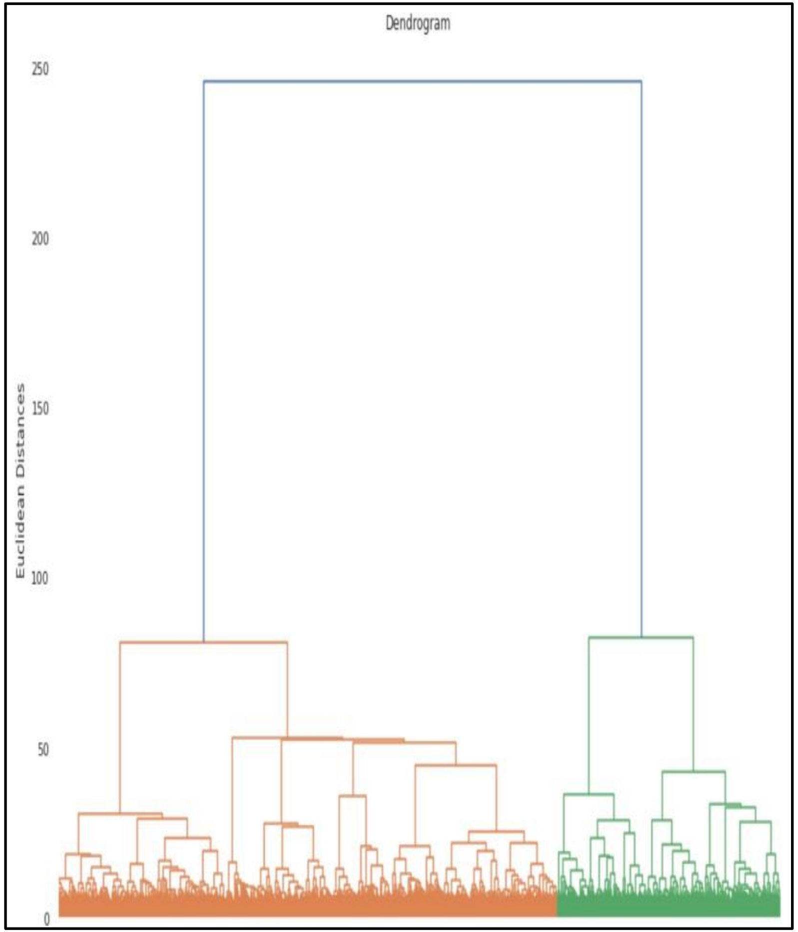
Agglomerative Clustering:

It is a type of hierarchical clustering algorithm. It is an unsupervised machine learning technique that divides the population into several clusters such that data points in the same cluster are more similar and data points in different clusters are dissimilar. Points in the same cluster are closer to each other.

Dendrogram

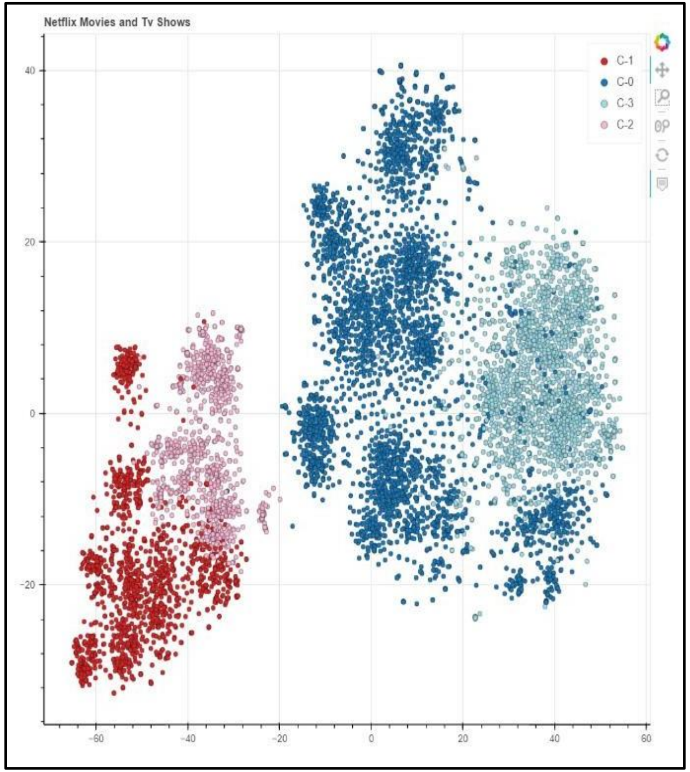
It is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data.

Continued…..



From the Figure we got after using the hierarchical clustering, we concluded that the best value for the number of clusters based on the Euclidian distances in the given figure was 4.
Using this value we tried to cluster the given dataset and got the results accordingly.

Continued…..



Cluster groups for movies or TV shows

- The above figure consists of different cluster groups for Movies and Tv shows over different parts of globe.

Conclusion

1. Exploratory Data Analysis was done for all the attributes to study the deep insights from the Given Dataset.

2. Univariate & multivariate analysis.

3. Visualized Data, inferred insights

4.Analysed various trends in Countries and the corresponding analysis was visualized to get a clear picture of the analysis.

5.TV Shows or Movies? Of course over the period of time the trend has been moving towards Netflix series instead of Movies. We tried to analyze this with graphical representation as well on yearly basis.

6.We used TFIDF Vectorizer and Sigmoid Kernel in order to recommend movies based on the similarities in the Textual Attributes.

7.Identified 4 distinct clusters and used Interactive Visualizations to dive deeper into the clustered data.

Thank you