Capstone Project

Netflix Movies and TV Shows Clustering

By-
Sourav Chowdhury
sourav.20497@gmail.com

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 services' 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.

In this project, we are required to do

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

Tools Used

Programming Language - Python

Data Analysis - Pandas

Data Visualization - Matplotlib, Seaborn

ML Algorithm - K-means Clustering and Hierarchical Clustering

Reading Input Dataset

To load the dataset in the colab notebook, first we mounted the notebook with google drive and then read the dataset using pandas built in function read_csv.

```
from google.colab import drive
drive.mount('/content/drive')


dataset = pd.read_csv ('/content/drive/MyDrive/Colab Notebooks/Capstone Projects/Unsupervised ML (Netflix Movies and TV shows Clustering) /netflix_titles (1).csv')
```

Data Pipeline

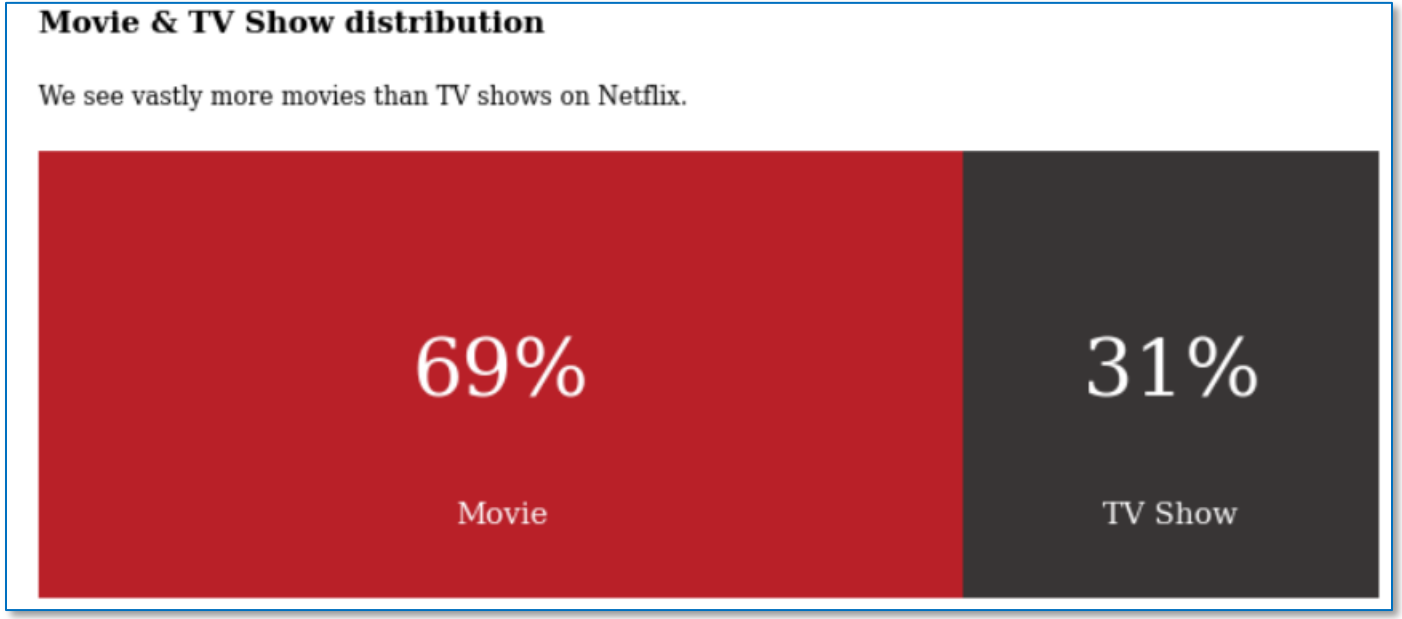
- **Data Processing - 1(Inspection):** In this part we have checked the dataset with its shape size and content.
- **Data Processing - 2:** In this part we have done the exploratory data analysis Univariate and Multivariate analysis and hypothesis over visualized data.
- **Data Preparation:** Done Feature engineering on the data. Text Pre-processing using stopwords removal, stemming, rescaling of data using standard scalar.
- **Model Implementation:** Finally, in this we have implemented K-means clustering and Hierarchical clustering.

Data Summary

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 : Genres
12. description: The Summary description

EDA

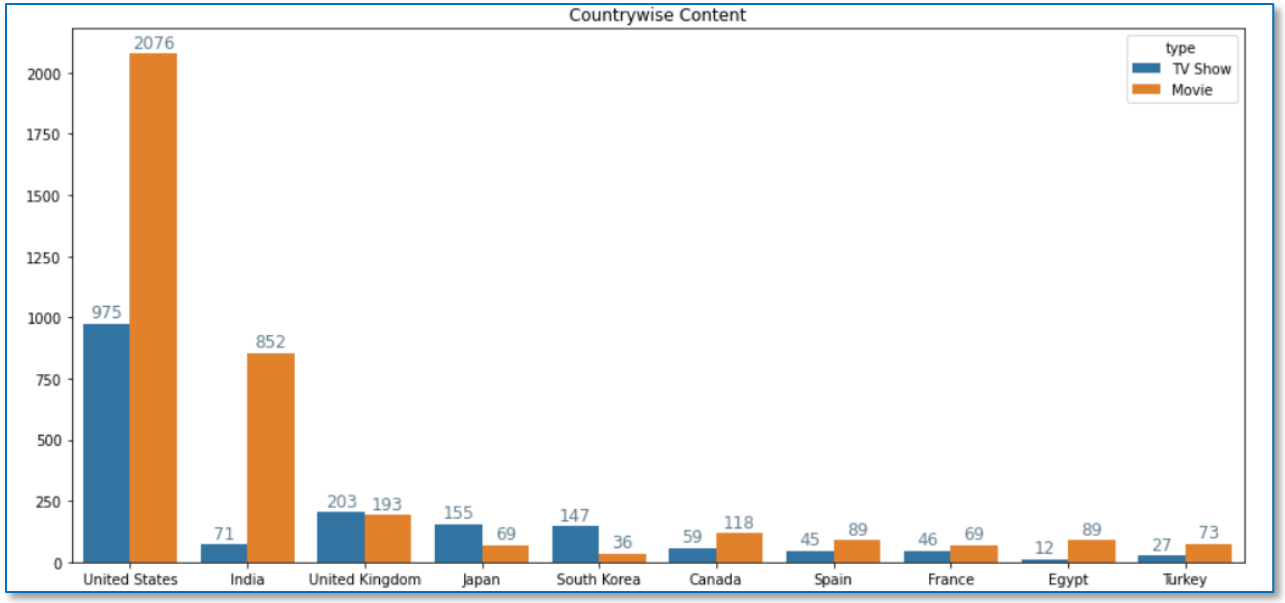
There are 69% movies and remaining 31% TV shows on Netflix.



EDA

Distribution of Content Countrywise

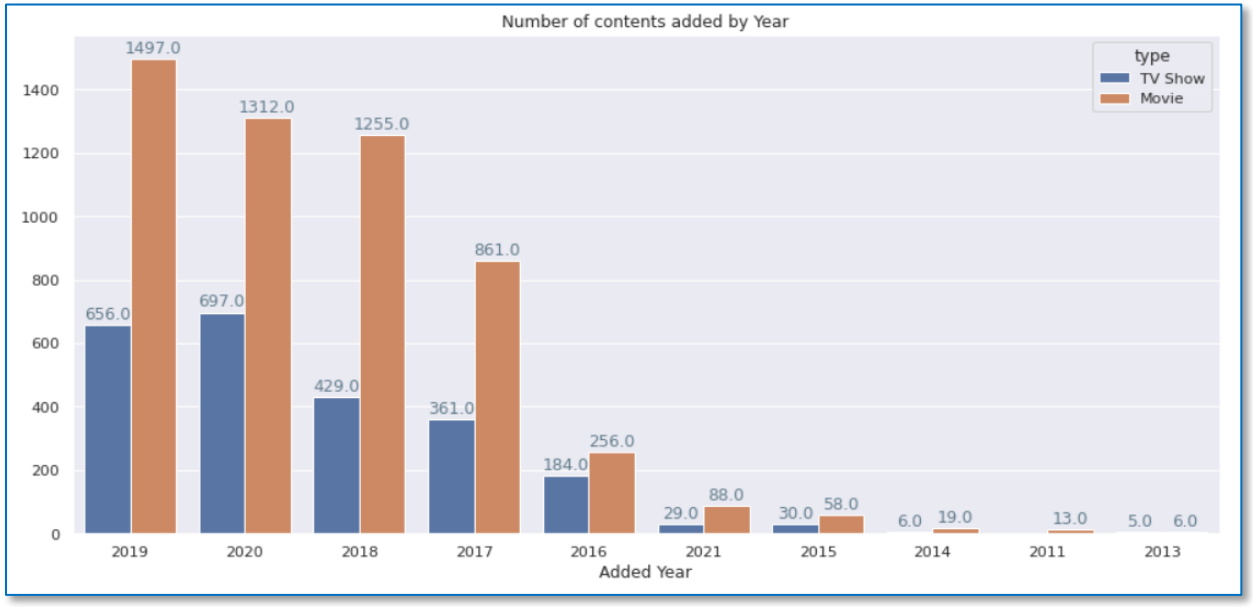
- United States is at top both in Movies and TV shows followed by India and then United Kingdom.



EDA

Year wise count of content added

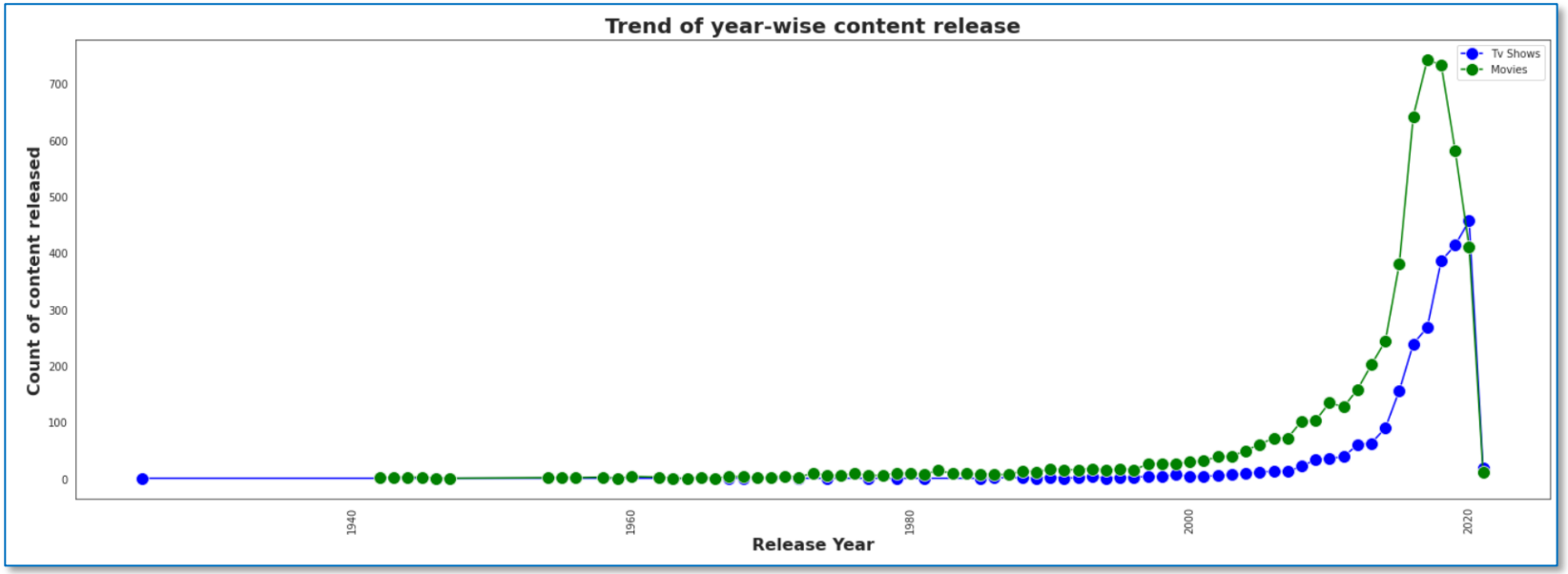
- Year on year basis content added got increased, drastic increase can be seen from 2016 to 2017.



EDA

- ## Year wise content released

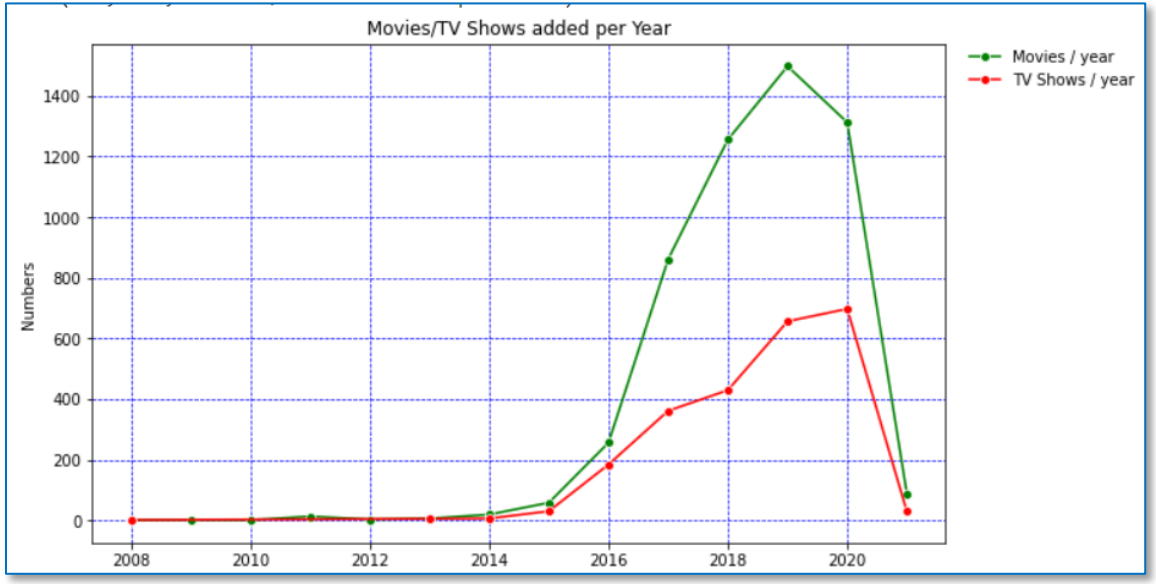
- The graph shows content released by Netflix since it’s starting, Movies was always in high number but in 2020 movies got slight dip by TV shows.



EDA

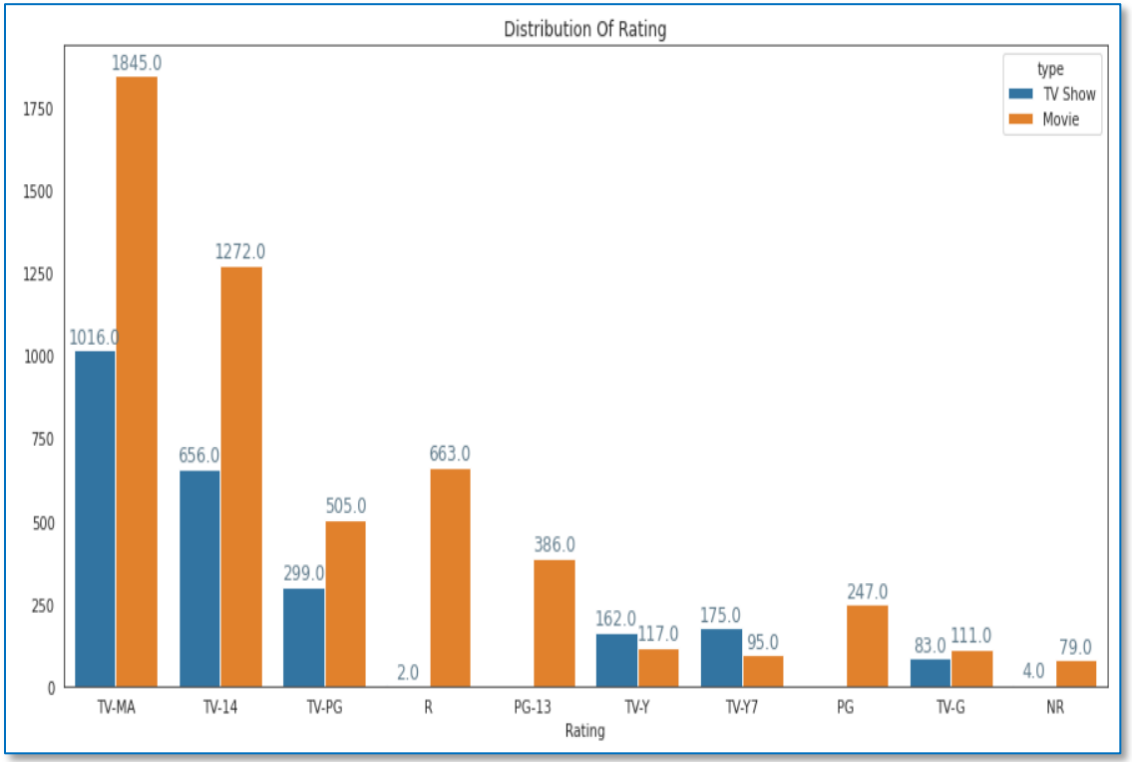
- ## Year wise content added

  It was moderate till 2015 and then it takes jump and goes on increasing till 2019.
  There is slight decrease in movies content in year 2020 but remain constant for TV Shows, pandemic also hits this year, so it could be the cause.



EDA

Distribution of Ratings of content available on Netflix



Description of Content Ratings

TV Content Ratings

TV-Y: Appropriate for all children

TV-Y7: For children aged 7 and above

TV-Y7-FV: For children aged 7 and above, program has moreintense fantasy violence

TV-G: General Audience

TV-PG: Parental Guidance Suggested

TV-14: Content might be unsuitable for children under 14 years age

TV-MA: Mature audience only (above 17 years of age)

Movie Ratings

G: General Audience

PG: Parental Guidance, some content may not be suitable for children

PG-13: Some material may be inappropriate for children under 13R:

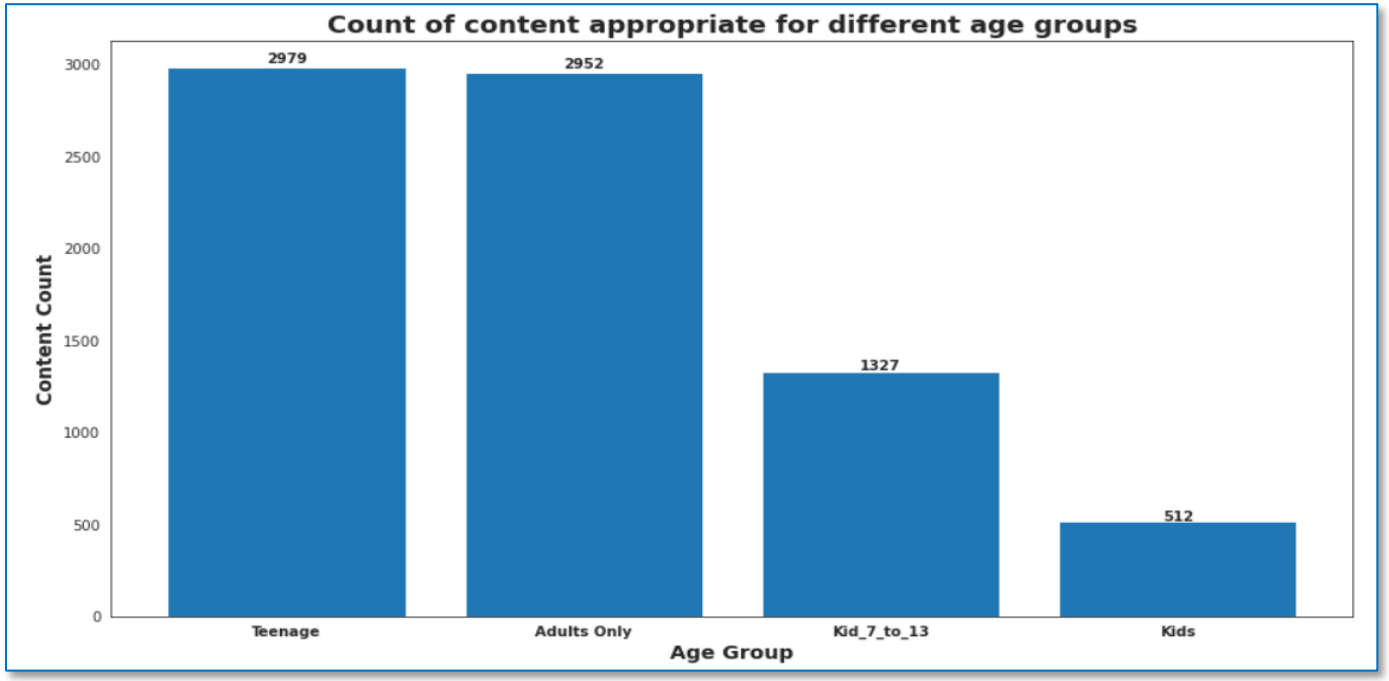
Under 17 accompanying with parent/guardian will be allowed NC-17:

No one 17 and under admitted

NR/UR: Not rated or unrated

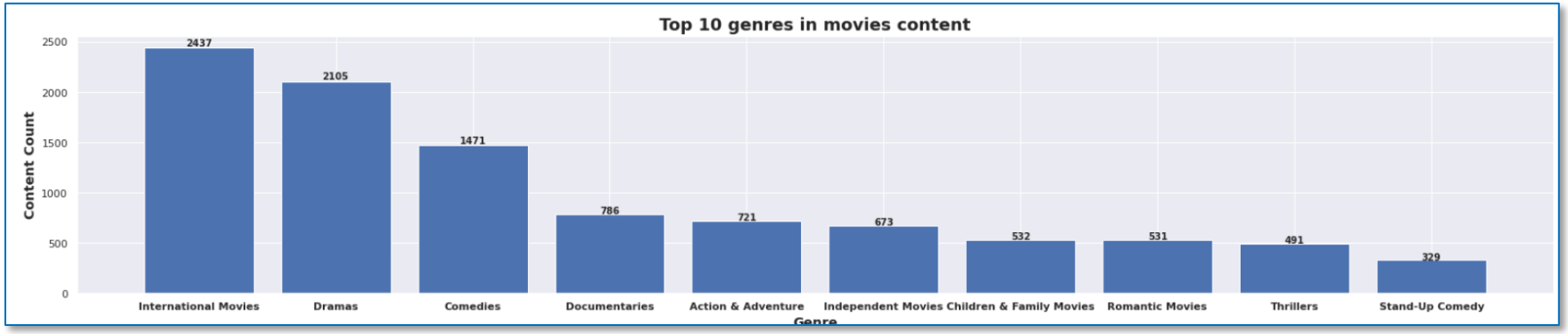
EDA

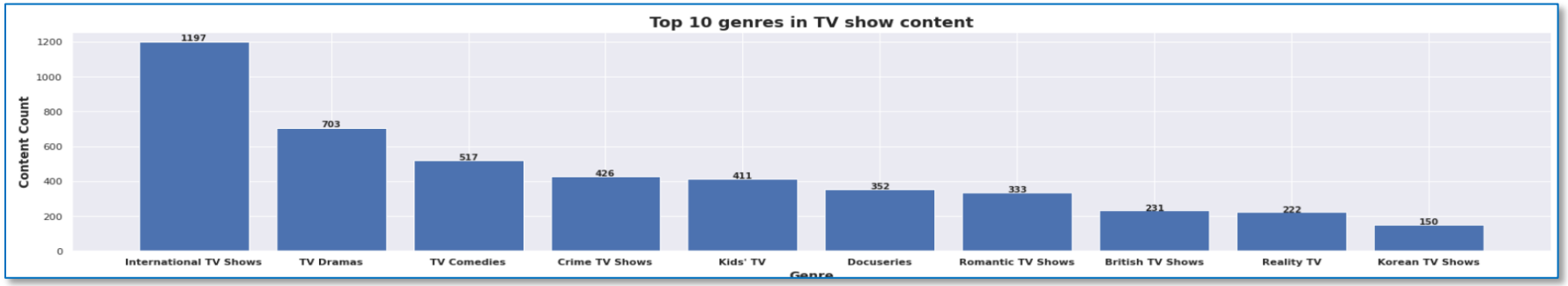
Count of Content Appropriate for different ages



EDA

Top 10 Genres in Movies and TV Shows content





Data Preparation

Data Pre-processing:

We first clean text, which means splitting it into words and handling punctuation. For clustering we choose “description” and “Listed_in” variables. Before clustering we need to pre-process the data. So that we filtered data with following steps:

Removing Punctuations

In this step, punctuations (.,”?/ etc) from sentences has been removed. So,as to make it free from unnecessarydata.

Removing StopWords

In this step, stopwords (is,are,this,that etc) from sentences has been removed. We make use of NLTK Library to remove it.

Stemming

Stemming is the process of producing morphological variants of a root/base word.

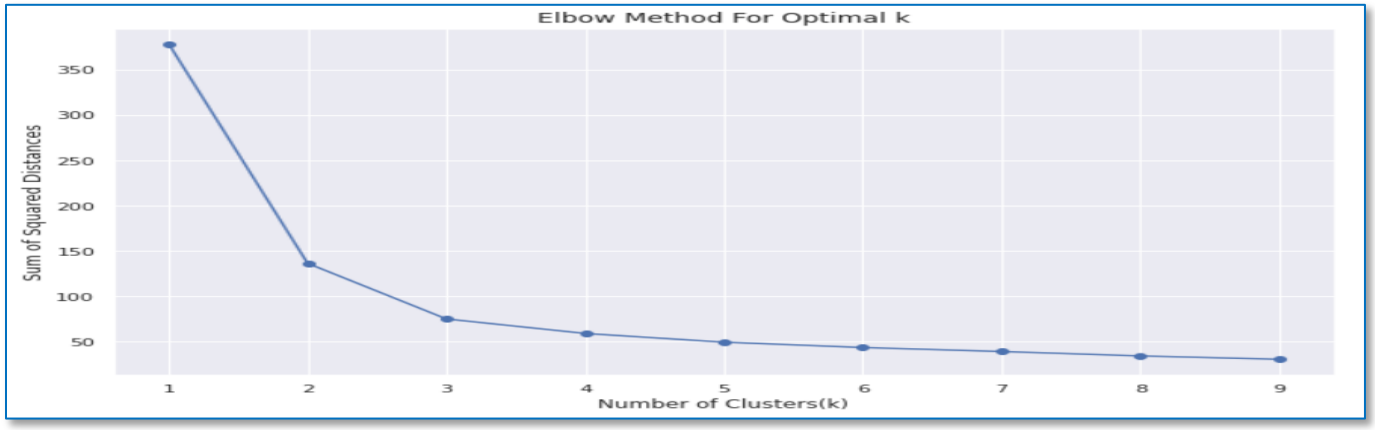
Eg: “chocolates”, “chocolatey”, “choco” reduces to the root word, “chocolate”

Creating New Variables for TextLength

Calculating the length of text we got from first three steps to do clustering

Applying Model (K-Means Clustering)

- The **K-Means** algorithm searches for a predetermined number of clusters within an unlabelled multidimensional dataset.
- The Elbow Method is one of the most popular method to determine this optimal value of k number of clusters.
- To determine the optimal number of clusters, we have to select the value of k at the “elbow” i.e. the point after which the distortion/inertiastart decreasing in a linear fashion.
- Thus from this chart we need to check, which would be the best number of clusters from 2,3,4,5,6 and 7.
- We found elbow formation at k = 3 and 6 but k = 6 we will take for optimal value.



Applying Model (Clusters for optimum number)

Silhouette Score for K-Means

Let's see Silhouette Score for different number of clusters:

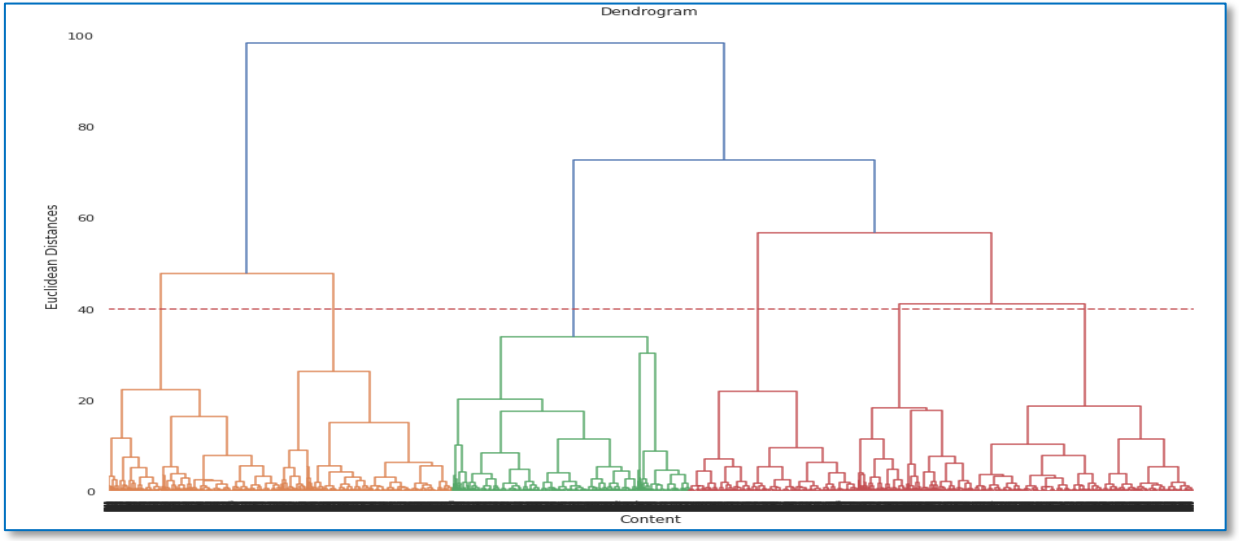
- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other.
- The Silhouette score is calculated for each sample of different clusters.
- The Silhouette Coefficient for a sample is

$$S = (b-a)/\max(a,b)$$

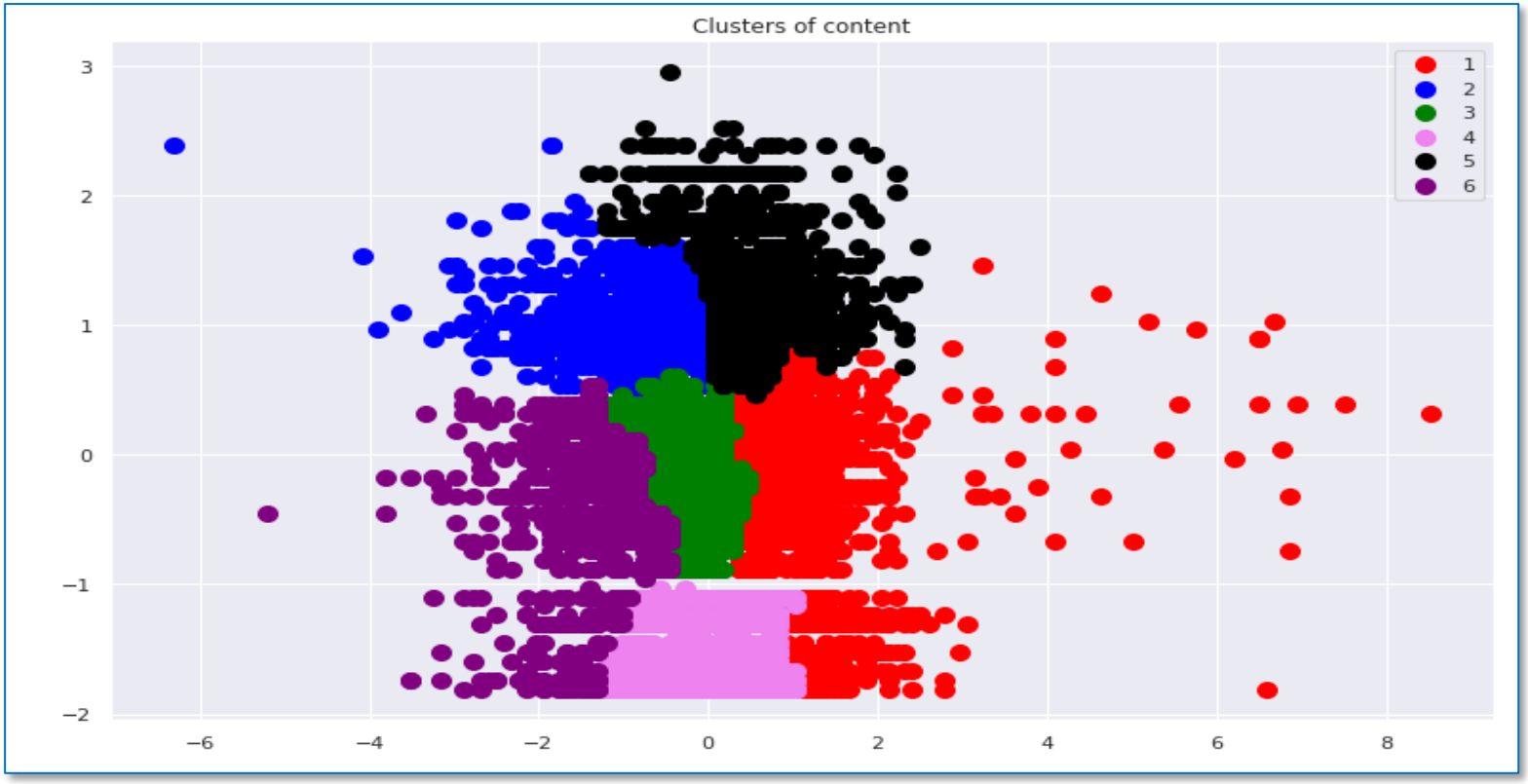
```
For n_clusters = 2, silhouette score is 0.35512756429120607
For n_clusters = 3, silhouette score is 0.3558545073559524
For n_clusters = 4, silhouette score is 0.3281130746442887
For n_clusters = 5, silhouette score is 0.33585923304123133
For n_clusters = 6, silhouette score is 0.3572596839713048
For n_clusters = 7, silhouette score is 0.35485725266624235
For n_clusters = 8, silhouette score is 0.35381821107922423
```

Applying Model (Hierarchical Agglomerative Clustering)

- Hierarchical Agglomerative clustering starts with treating each observation as an individual cluster, and then iteratively merges clusters until all the data points are merged into a single cluster.
- Dendrograms are used to represent hierarchical clustering results.
- The number of appropriate clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold. In this case it is at 6.



Hierarchical Agglomerative Clustering



Silhouette Score for HAC

After calculating the silhouette Score for HAC, we found that perfect number of cluster is 7.

```
For n_clusters = 2, silhouette score is 0.31733940335920435

For n_clusters = 3, silhouette score is 0.28576591463080314

For n_clusters = 4, silhouette score is 0.2970786181831897

For n_clusters = 5, silhouette score is 0.2818170622516962

For n_clusters = 6, silhouette score is 0.3004594443021899

For n_clusters = 7, silhouette score is 0.31843009608053685

For n_clusters = 8, silhouette score is 0.3102634162499398
```

Conclusion

After pre-processing the data, I started with EDA to understand the trends and features of the provided dataset. Major findings from EDA are as follows:

- There is 69% movie content and 31% TV show content in the dataset.
- The United States account for the majority of the content created on Netflix, numbering 3051 titles. India is the second largest with 923 titles.
- All the top 10 countries producing highest amount of content have higher proportion of movies than TV shows except Japan, South Korea and United Kingdom.
- Netflix is increasingly focusing on TV Shows now, which is clear from the graph, from 2019 to 2020, there was a decreasing trend of Movies. The TV shows from 2019 to 2020 remains constant.
- TV-MA, TV-14 and TV-PG are the top content rating to which highest number of content belongs.
- Among the top 10 countries with highest content volume, United States and India has highest volume of content appropriate for teens and most other countries have highest volume of content appropriate for adults.
- Content onboarding on Netflix started increasing from 2015 and reached to its peak in 2019, there is a drastic downfall in 2021. The downfall can be attributed to Covid pandemic.
- International movies is the highest content genre in movies and International TV Shows is the highest one in TV shows.
- TV shows with one season are highest in number.

Findings after applying k- means and Hierarchical clustering:

- K-means clustering is used to form the clusters of clusters of the content. To find out the optimal value of k, Elbow and Silhouette method was used. K=6 was appropriate value and can be visualized after clustering the content using k-means clustering.
- Hierarchical clustering implementation gives perfect score of silhouette at 7 clusters, same as visualized using dendrogram.

Thank You