CAPSTONE PROJECT 4

Netflix Movies and T.V Shows Clustering

Created by Sakshi Dhyani

Project Details

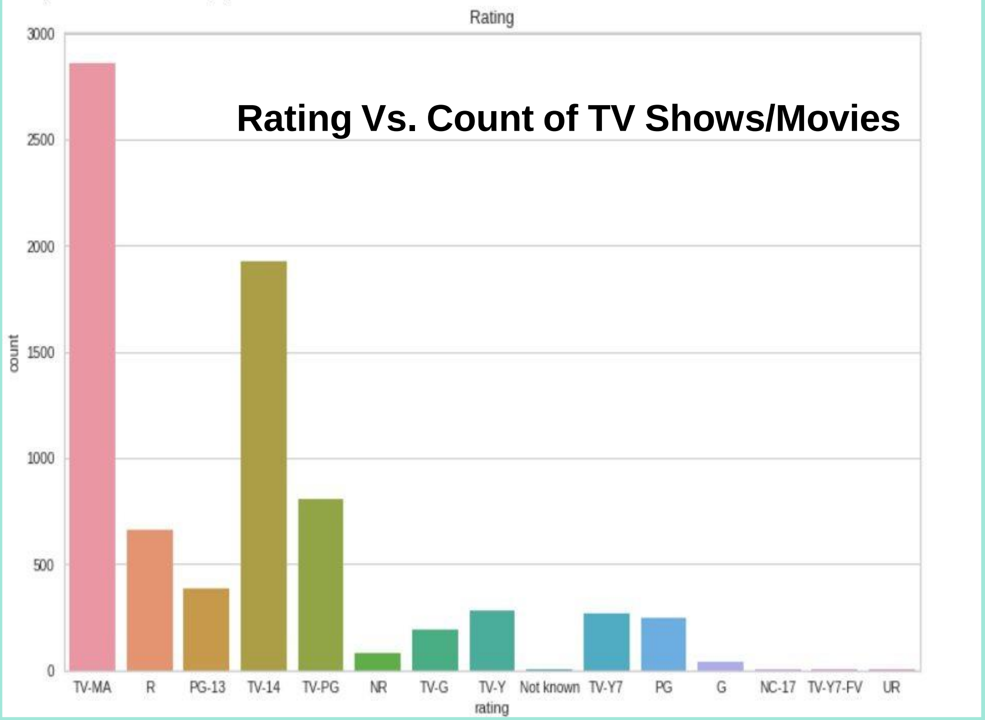
Netflix is a streaming service with wide variety of content. The idea of this project is to analyze and perform clustering to determine various patterns related to the content available in Netflix. The data is gathered from a third party engine.

Based on the attributes related to the Tv shows or movies, we will be implementing different clustering algorithms which comes under unsupervised Machine learning category.

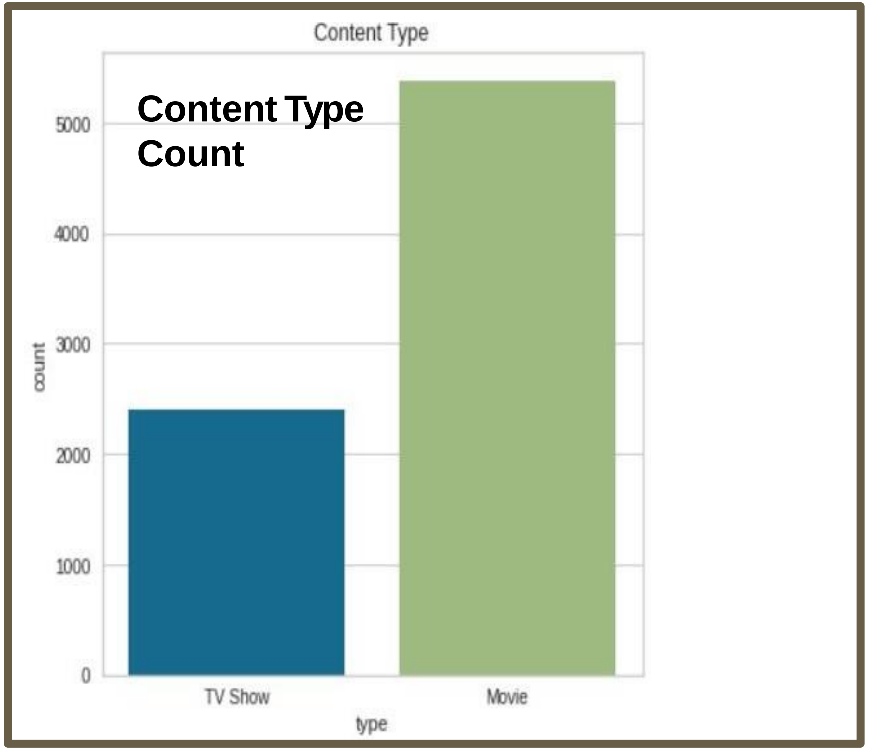
DataSet Details

- **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

Exploratory Data Analysis



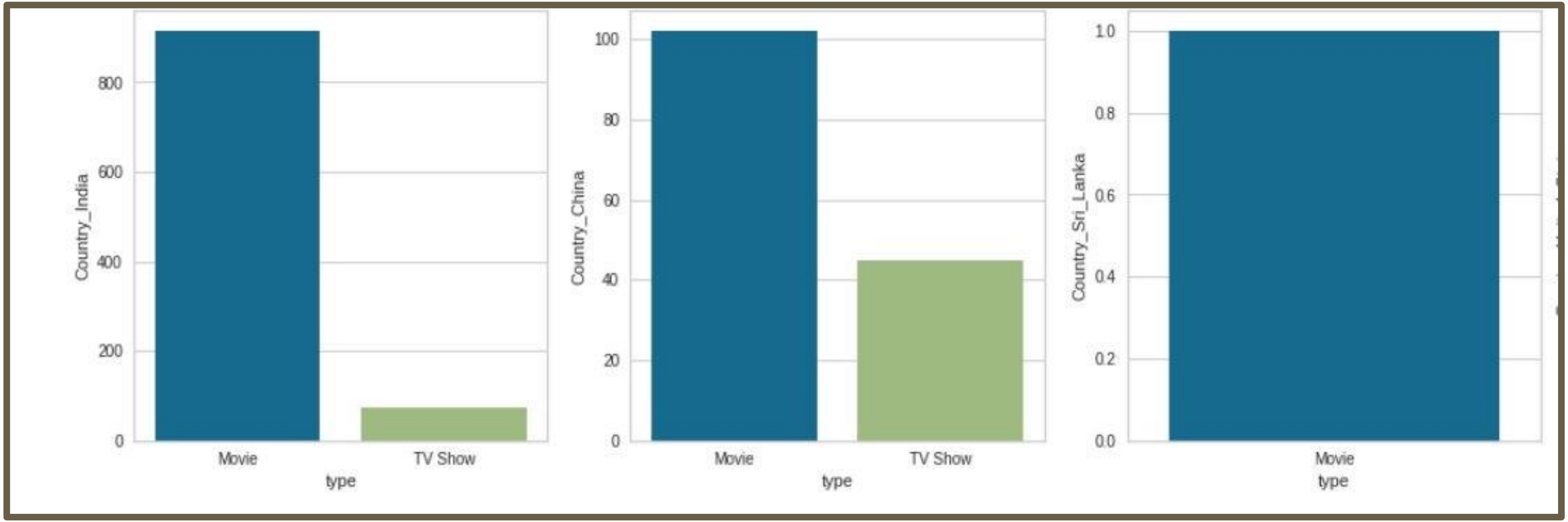
Rating Vs. Count of TV Shows/Movies



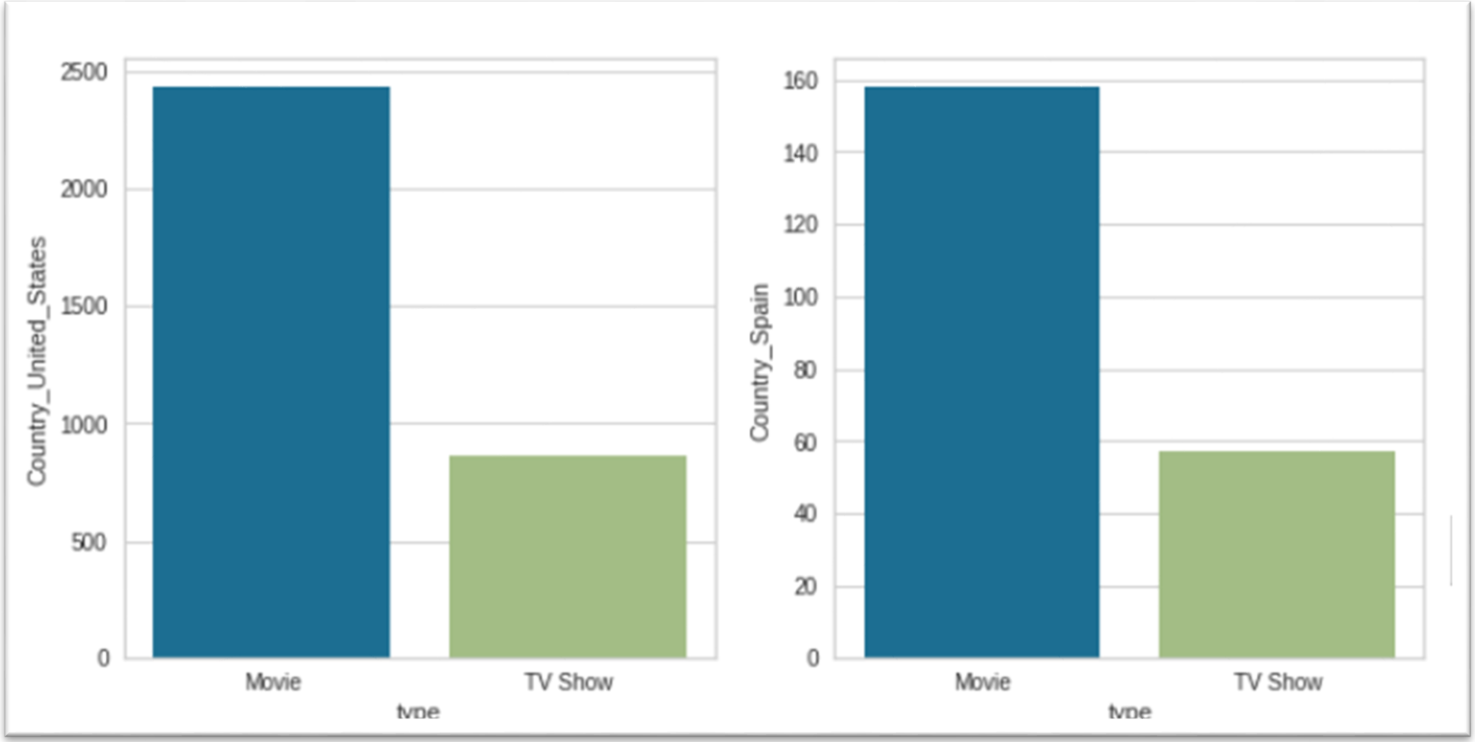
Content Type Count

Bar Plots for different countries based on content type

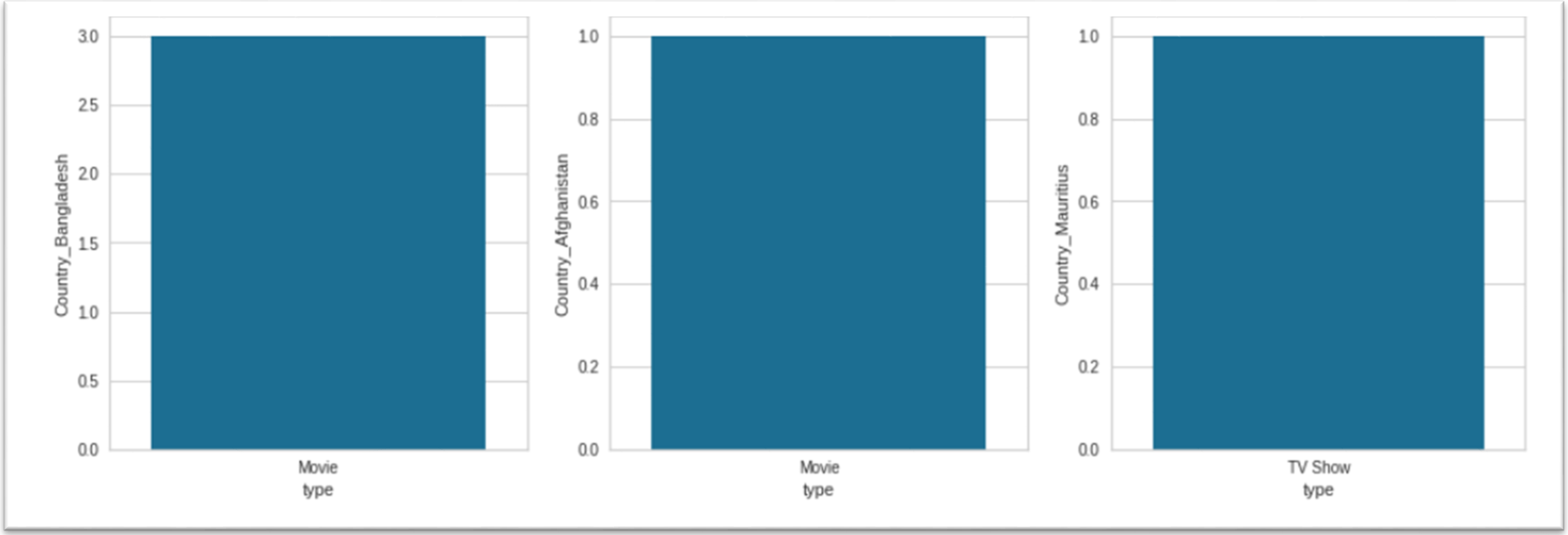
Selectively showing bar plots for countries India, China and Sri Lanka. Similar bar plots were plotted for other countries as well.



Bar Plots for different countries based on content type

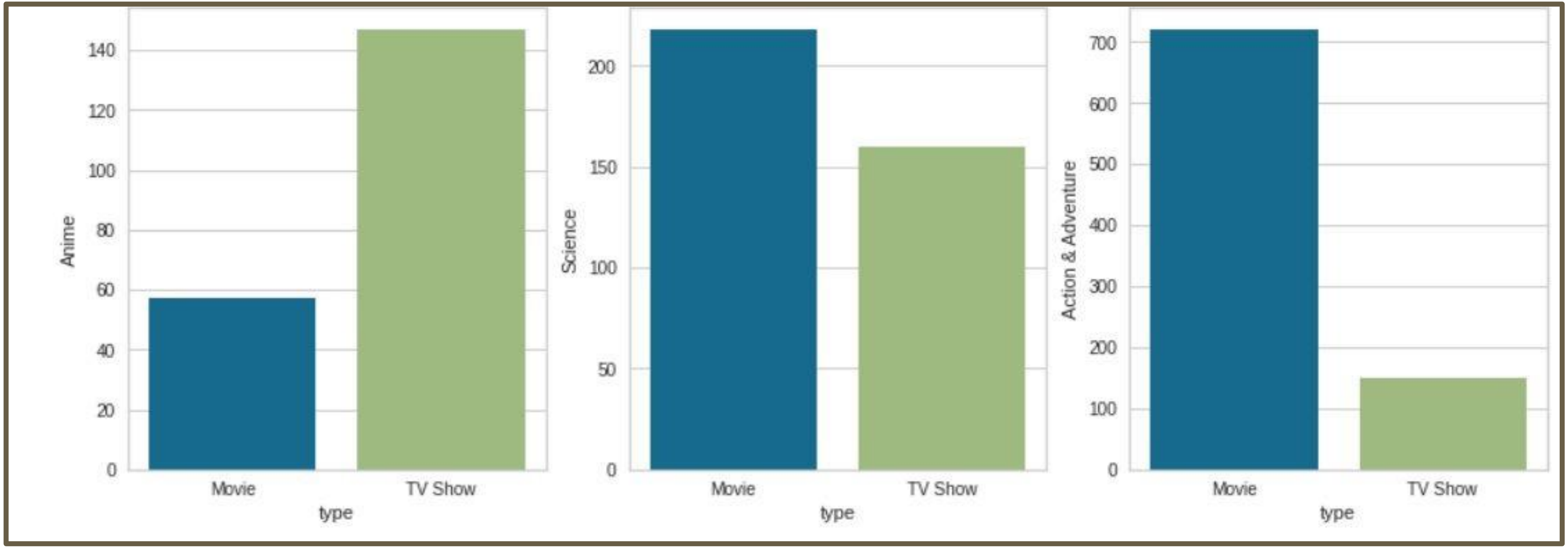


Bar Plots for different countries based on content type

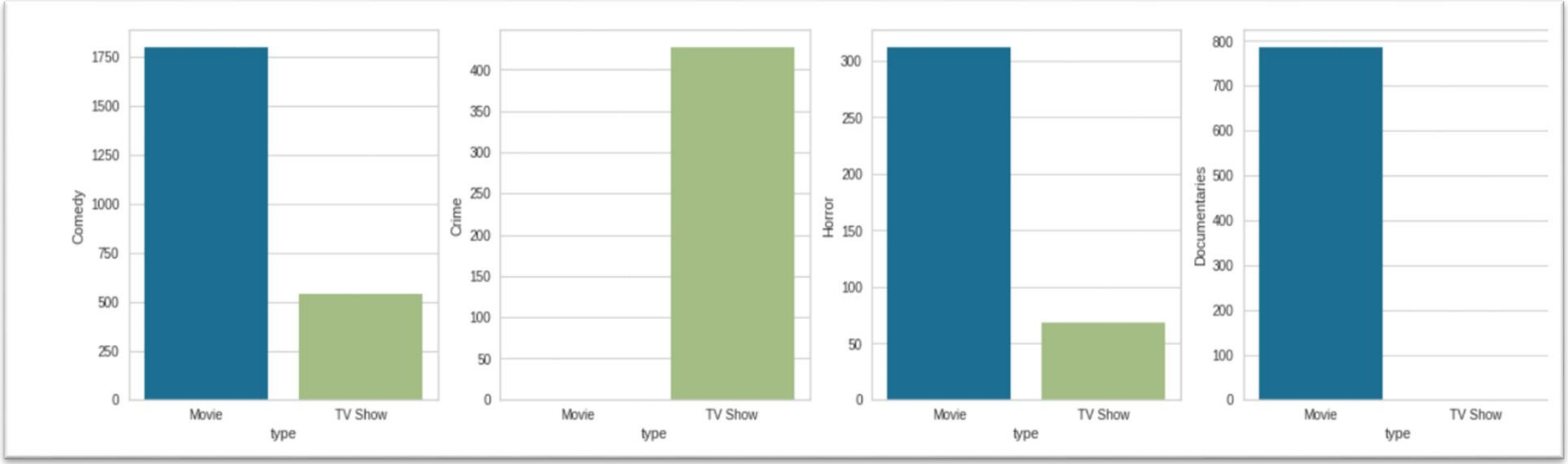


Bar Plots for different genres based on content type

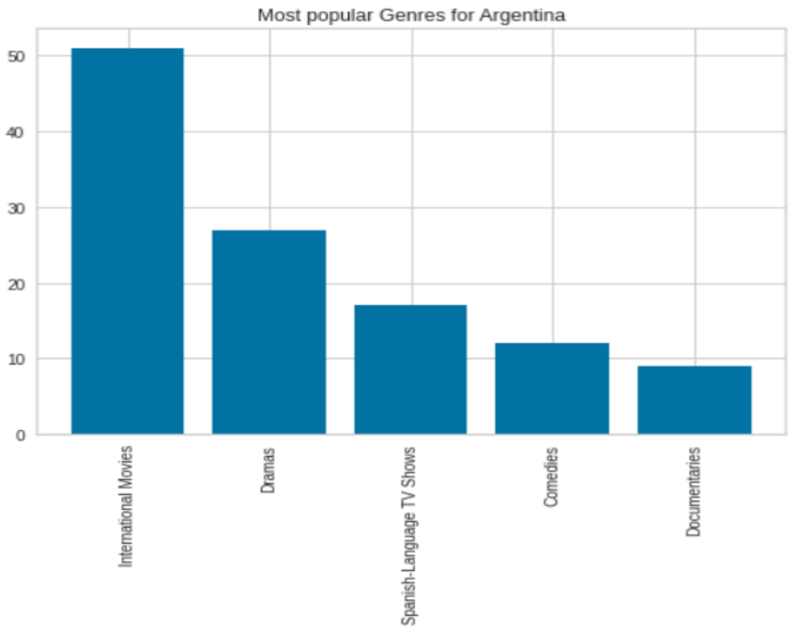
Selectively showing bar plots for 3 genres Anime, Science, Action & Adventure here. Similarly for all genres, bar plots were plotted

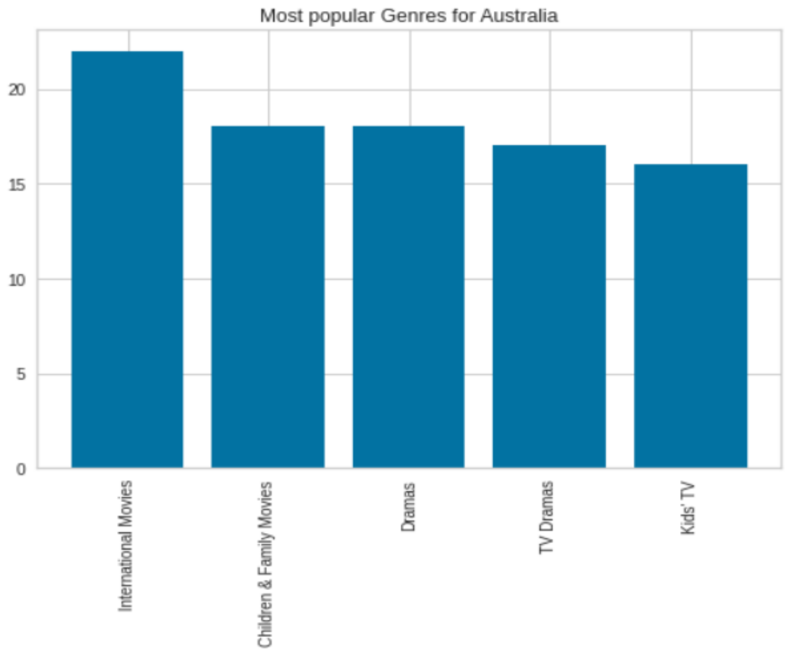


Bar Plots for different genres based on content type

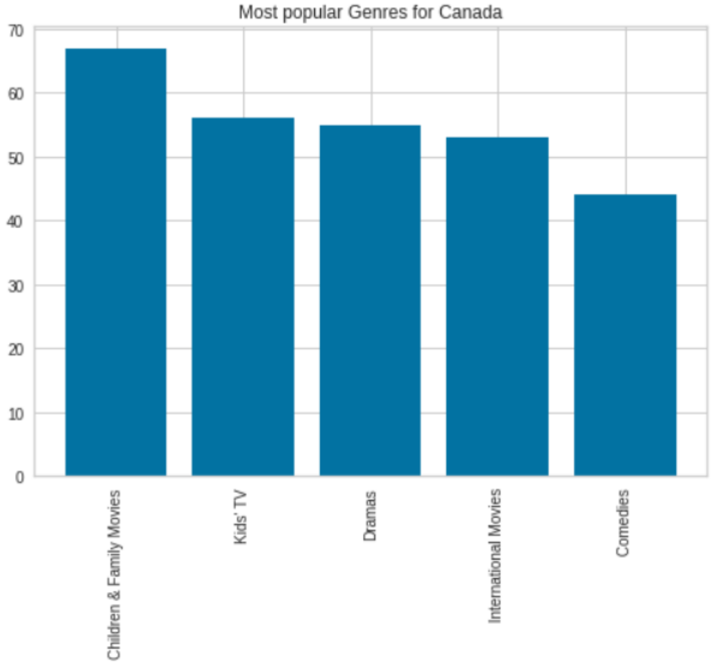


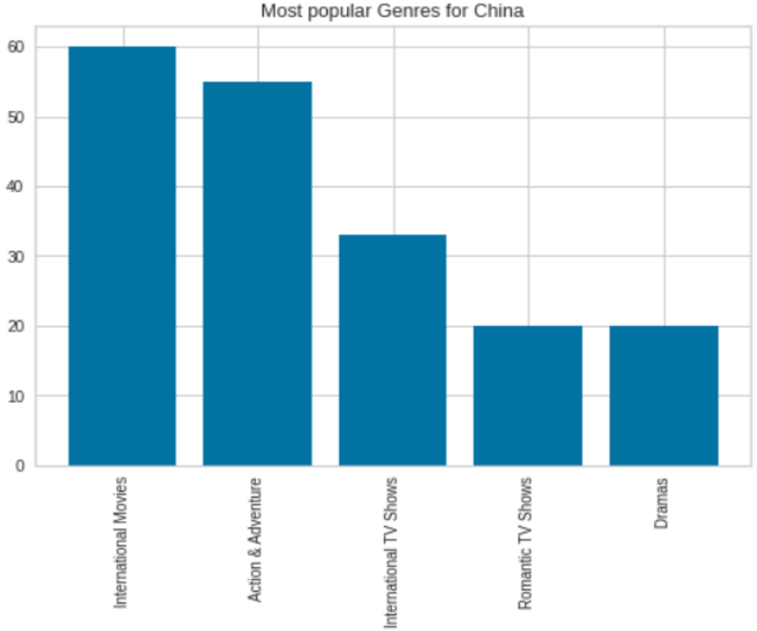
Bar Plots for most popular genre country wise.



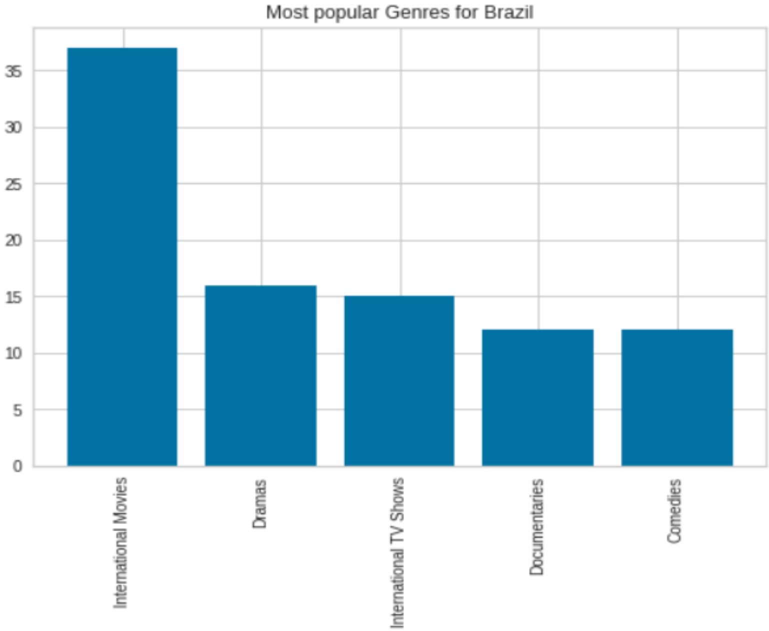


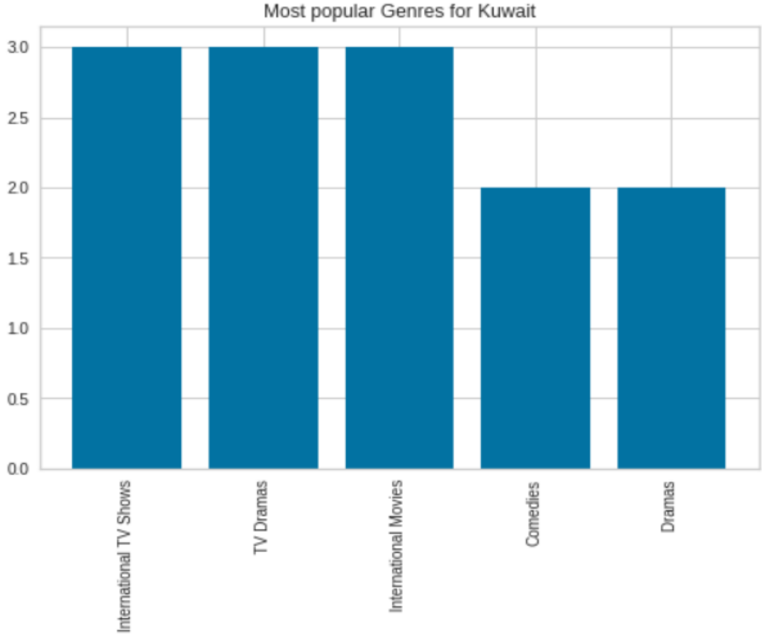
Bar Plots for most popular genre country wise.





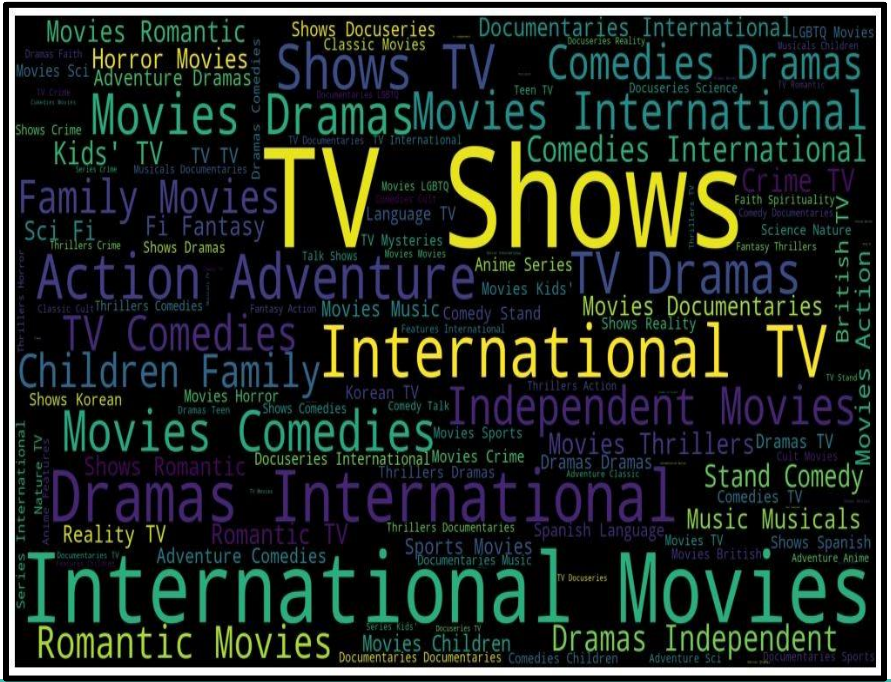
Bar Plots for most popular genre country wise.

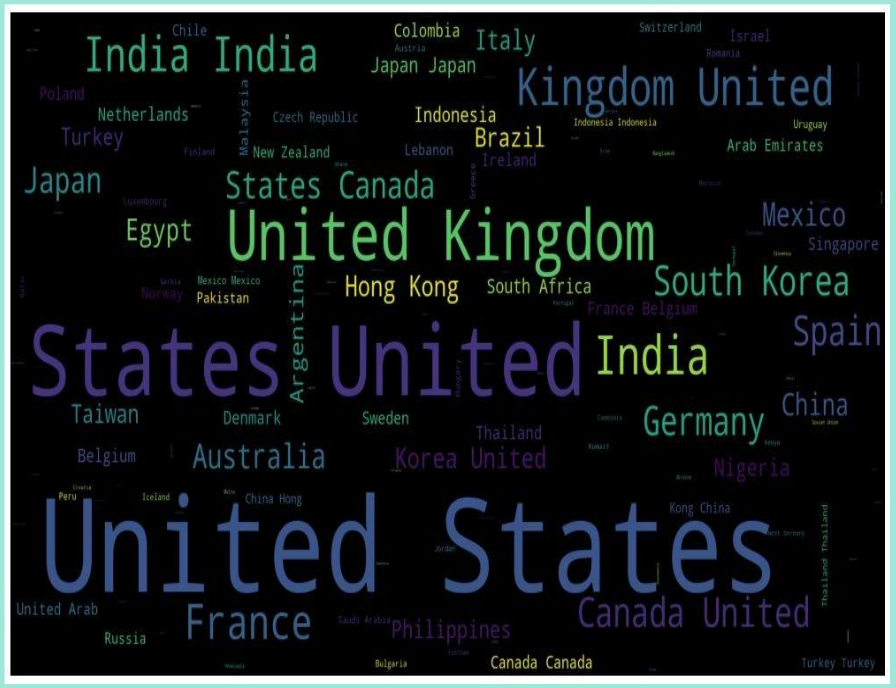




Exploratory Data Analysis

Word cloud for genres and different countries for the content in netflix

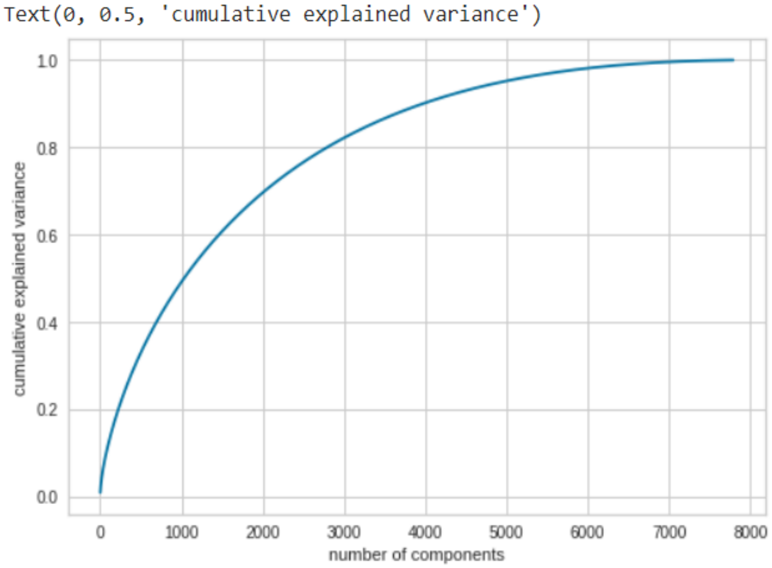




Data transformation

TF-IDF : Term Frequency, Inverse Document Frequency. It indicates the significance/relevance of word in a corpus. Term Frequency represents the number of instances of a given word and inverse document frequency tests how relevant the word is.

PCA: PCA is a dimensionality reduction technique. In this principal components are computed, and lot of information is also retained. Graph shows the explained variance value for different number of PCA components. n_components 4000 was chosen in this project.



```
Text(0, 0.5, 'cumulative explained variance')
```

Implementation of Clustering Algorithm

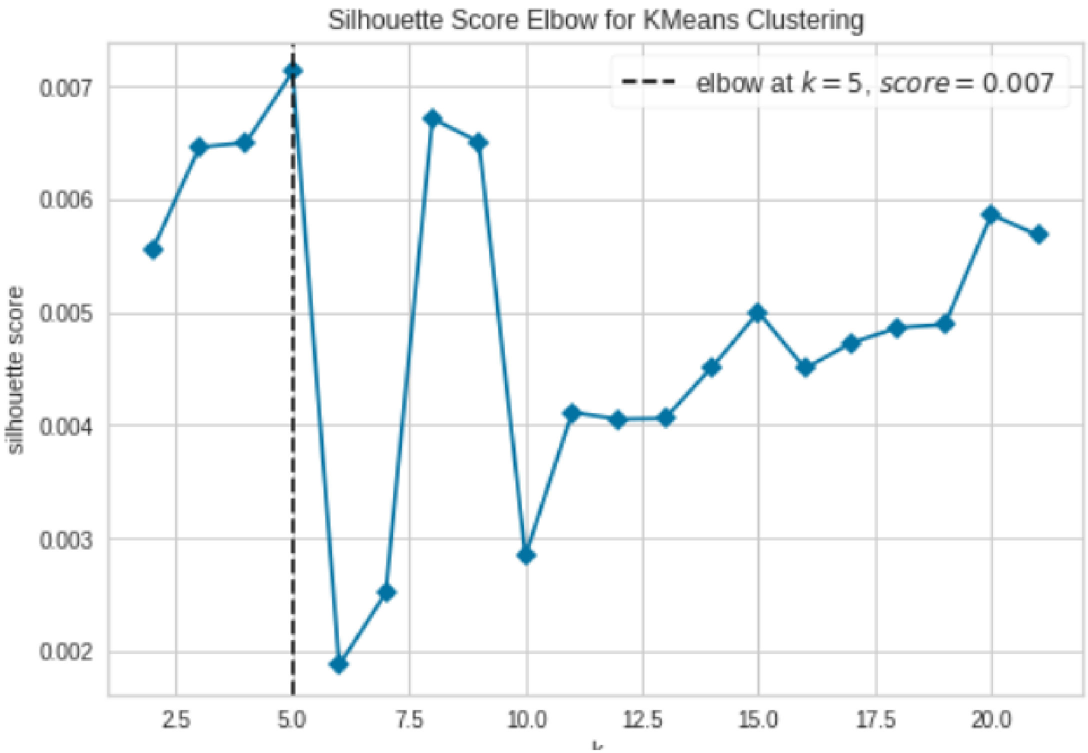
Textual data was combined and converted into numerical data using TF-IDF. Applied PCA to perform dimensionality reduction. Data was converted to 4000 dimensional data. K-Means is used in this project.

Some Clustering algorithms:

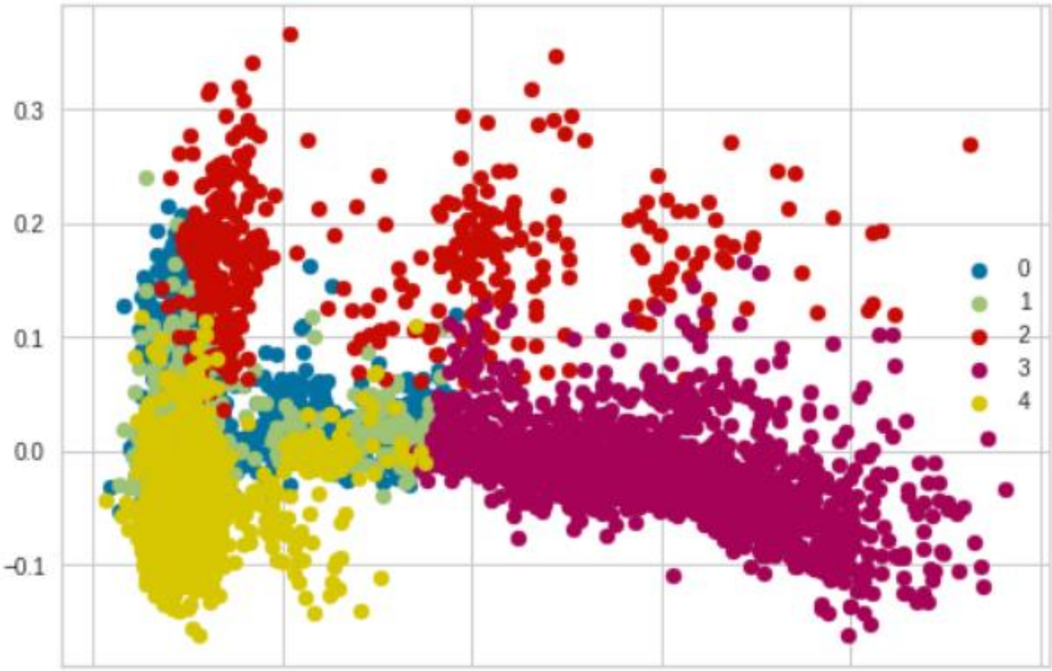
- ➢ K-Means Clustering
- ➢ Gaussian Clustering
- ➢ Agglomerative clustering

K-Means Clustering

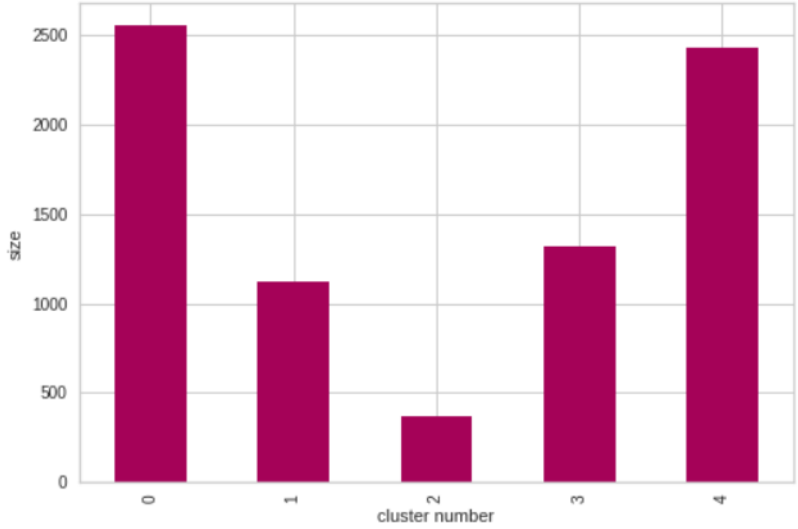
Silhouette score is maximum for clusters 5, and elbow is also at 5. 5 clusters chosen as optimal number of clusters.



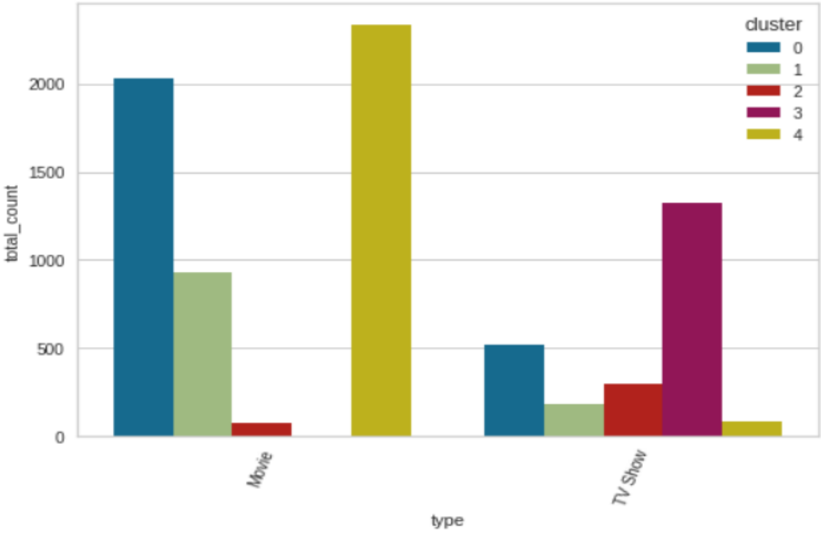
Clusters for K-Means Clustering



Cluster Analysis

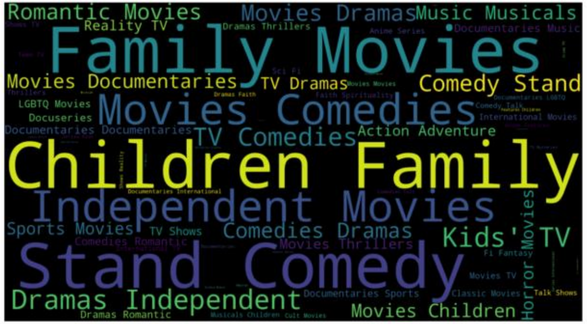


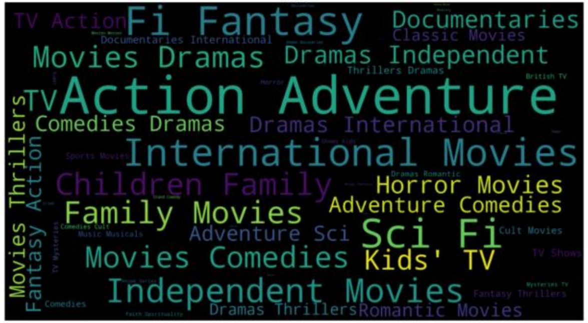
Size of clusters

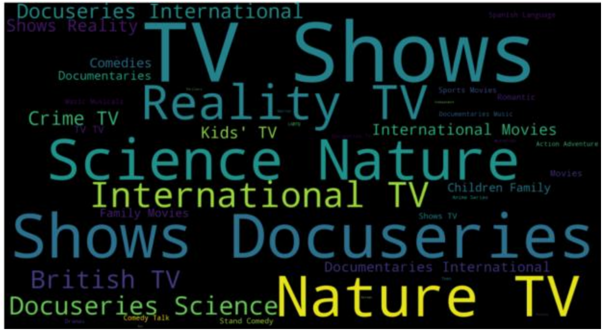


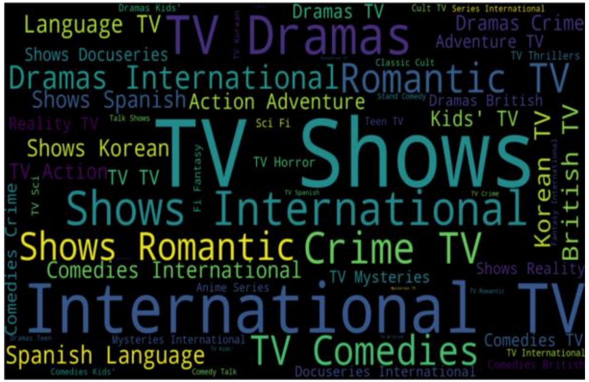
Cluster-wise content Type

Cluster Analysis

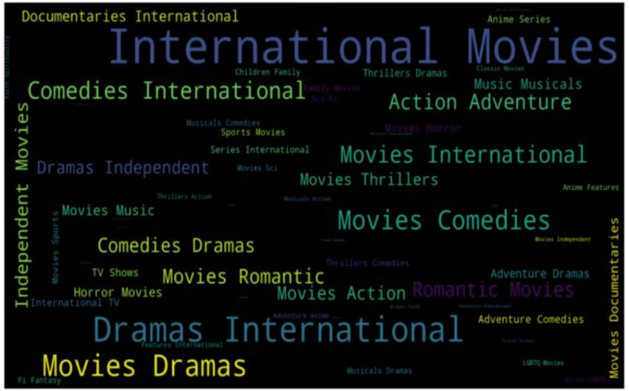






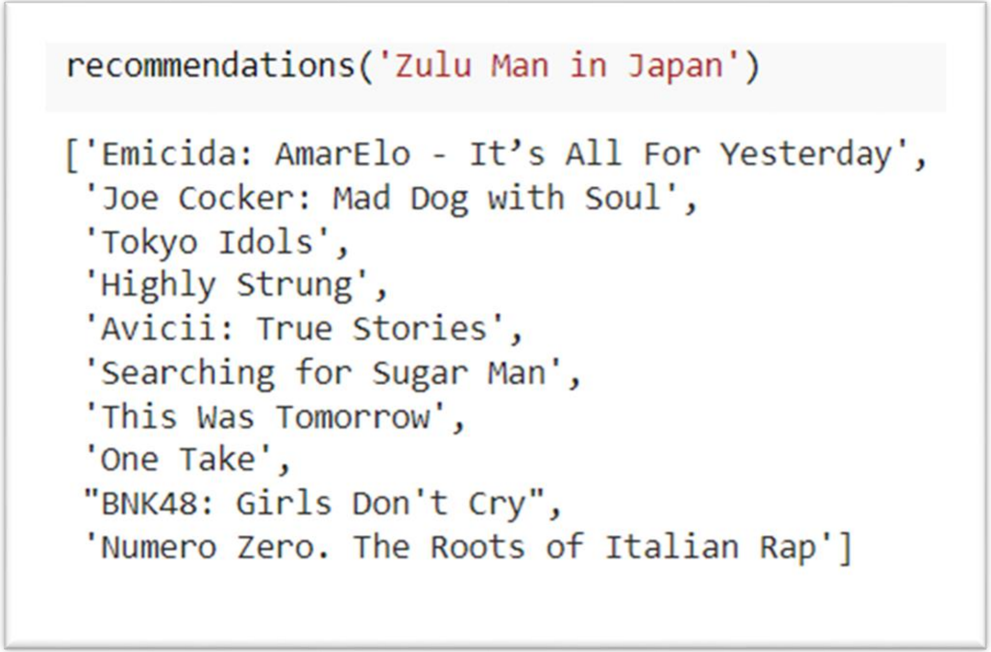


Cluster wise genres distribution via world cloud



Content Based Recommender System

Count Vectorizer : Used Count vectorizer to transform textual data.
Cosine Similarity : Used Cosine similarity to find similar movies/tv shows.



```
recommendations('Zulu Man in Japan')
```

```
['Emicida: AmarElo - It’s All For Yesterday',
 'Joe Cocker: Mad Dog with Soul',
 'Tokyo Idols',
 'Highly Strung',
 'Avicii: True Stories',
 'Searching for Sugar Man',
 'This Was Tomorrow',
 'One Take',
 "BNK48: Girls Don't Cry",
 'Numero Zero. The Roots of Italian Rap']
```

Conclusion

- ➔ EDA was performed to analyze the data
- ➔ Stemming was performed for converting textual data. Data cleaning was performed for the textual data.
- ➔ TF-IDF was used for transforming textual data.
- ➔ PCA was implemented to reduce dimensionality. Components were chosen using explained variance Vs components graph.
- ➔ K-means clustering was implemented.
- ➔ Optimal number of clusters was found using Silhouette score and elbow curve graph.
- ➔ Features for different clusters were compared.
- ➔ Content Based recommendation system was created based on cosine similarity using count vectorizer. It recommends similar 10 movies/tv shows for any movie name provided.

THANK YOU