

Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By: Vikash Kumar

Problem Description

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.

Integrating this dataset with other external datasets such as IMDB ratings, and rotten tomatoes can also provide many interesting findings.

Problem Description

In this project, you are required to do

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

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

Data Description

The dataset includes over 7787 records and 12 attributes.

- 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: Genere
- description: The Summary description

Data Cleaning

Prior to EDA, cleaning the data is essential since it will get rid of any ambiguous information that can have an impact on the results.

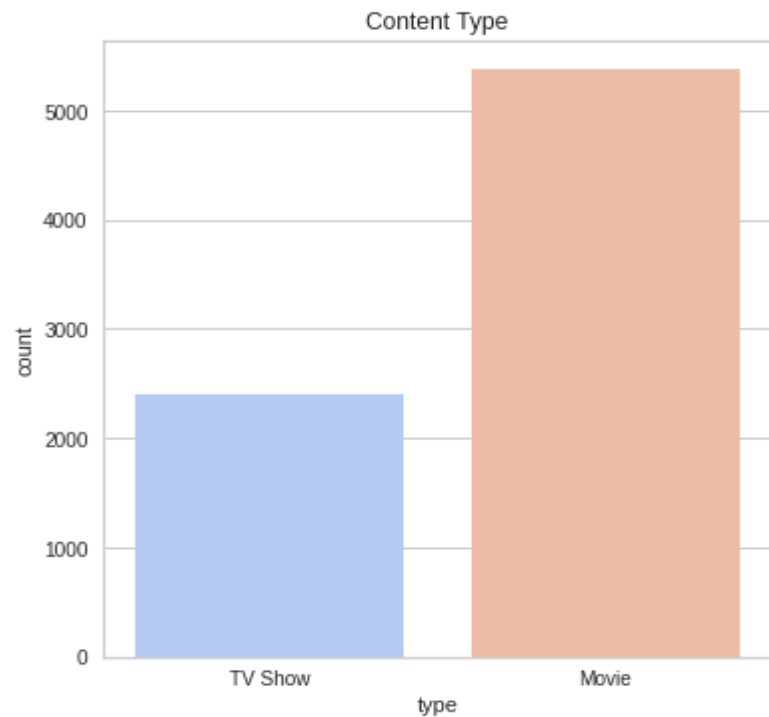
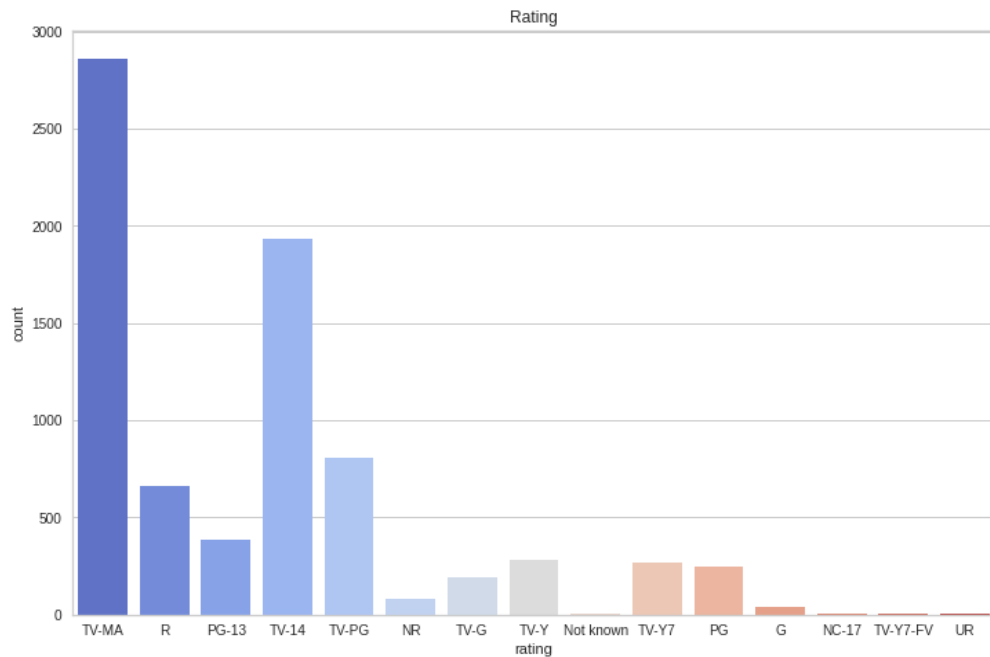
director, cast, country, date_added, and rating columns have missing or null values. I have replaced all the null values of the columns with 'Not Known'.

I removed the show id column because it doesn't offer any useful information.

data.isna().sum()

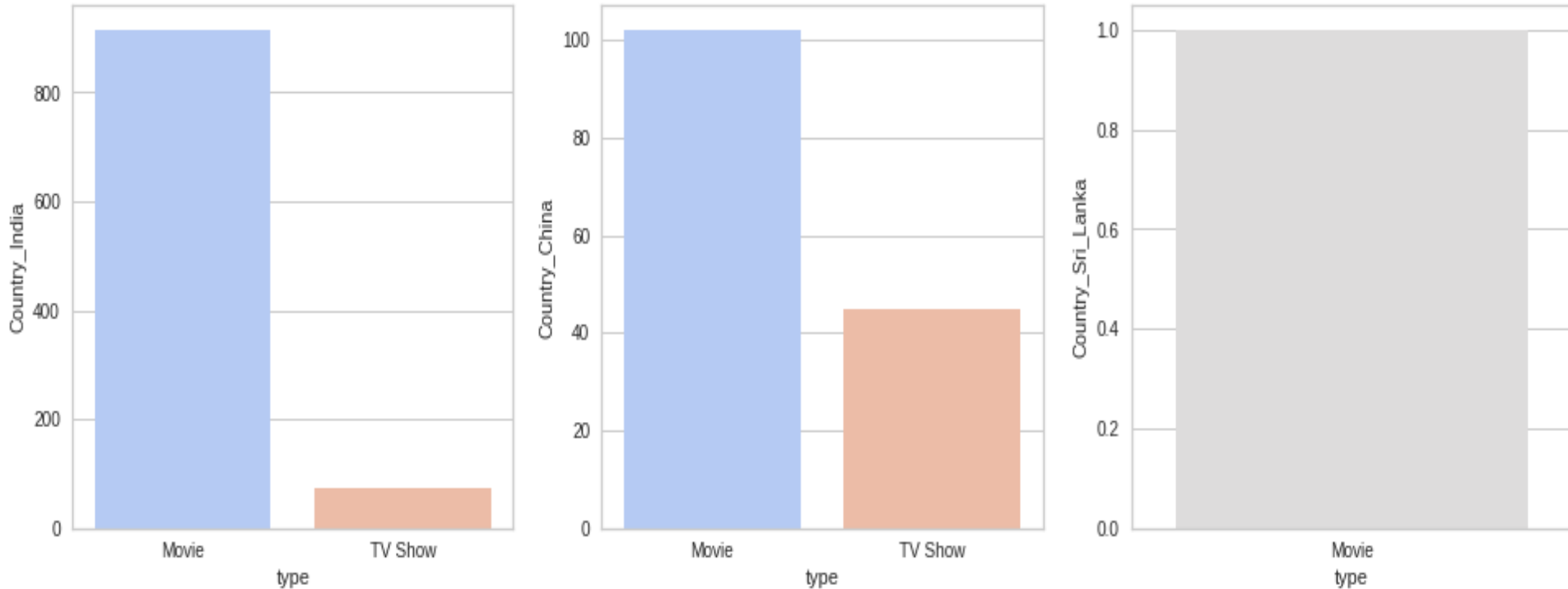
show_id	0
type	0
title	0
director	2389
cast	718
country	507
date_added	10
release_year	0
rating	7
duration	0
listed_in	0
Description	0

Exploratory Data Analysis

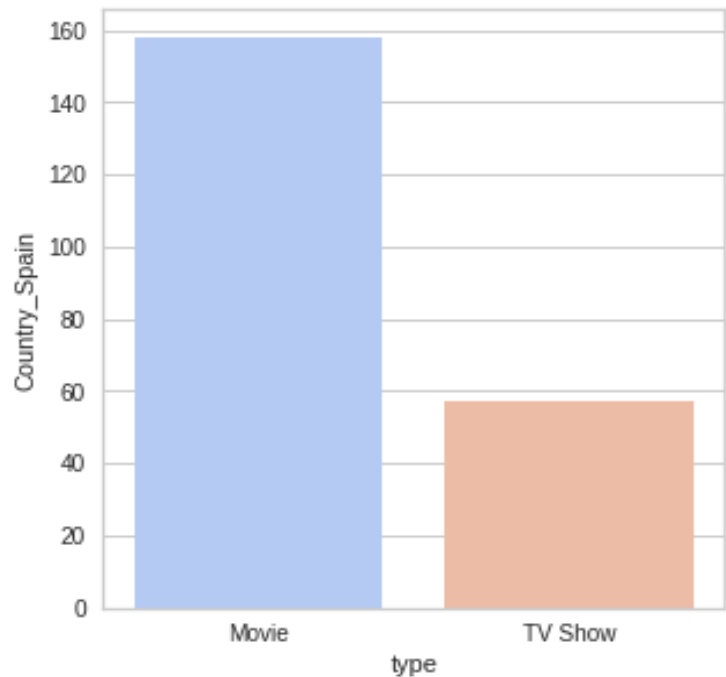
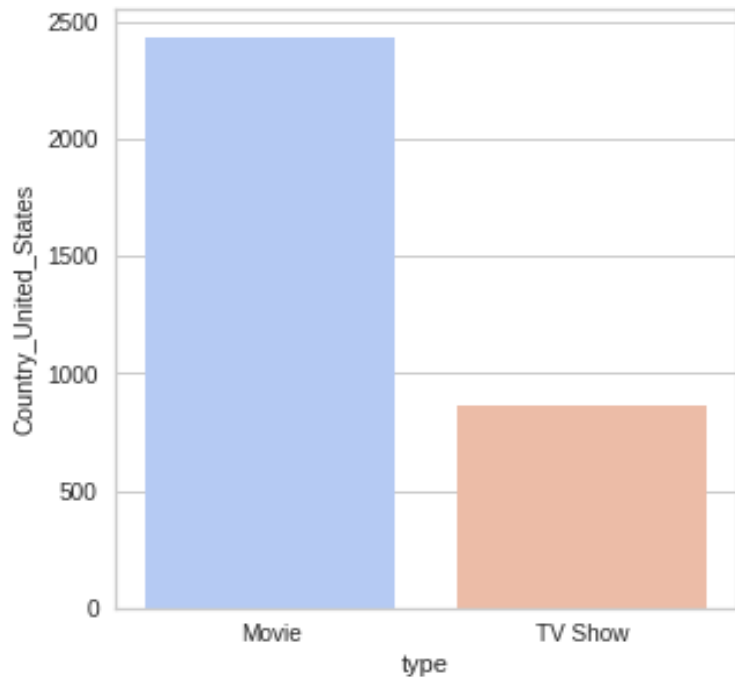


Bar plots for various countries based on the type of content.

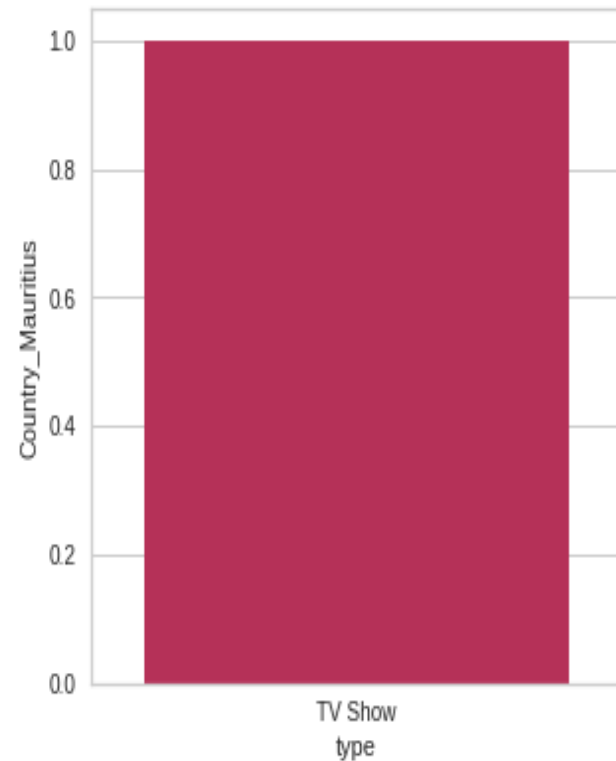
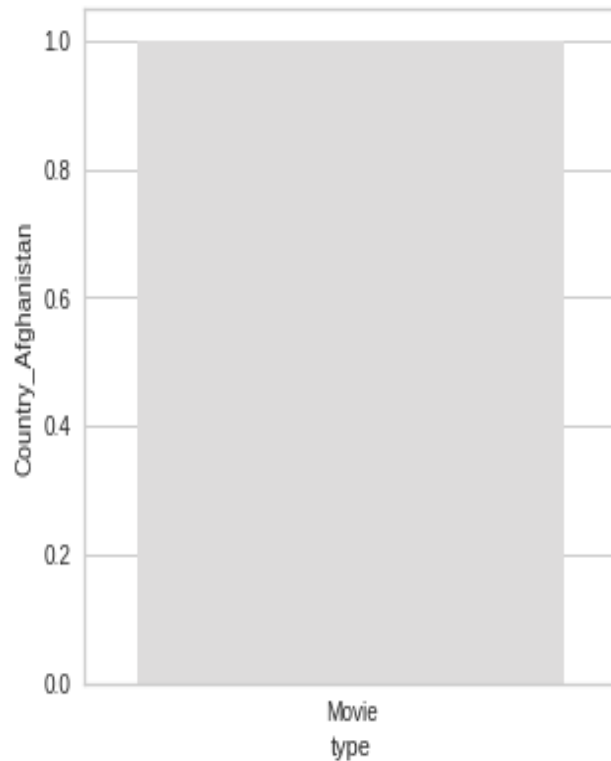
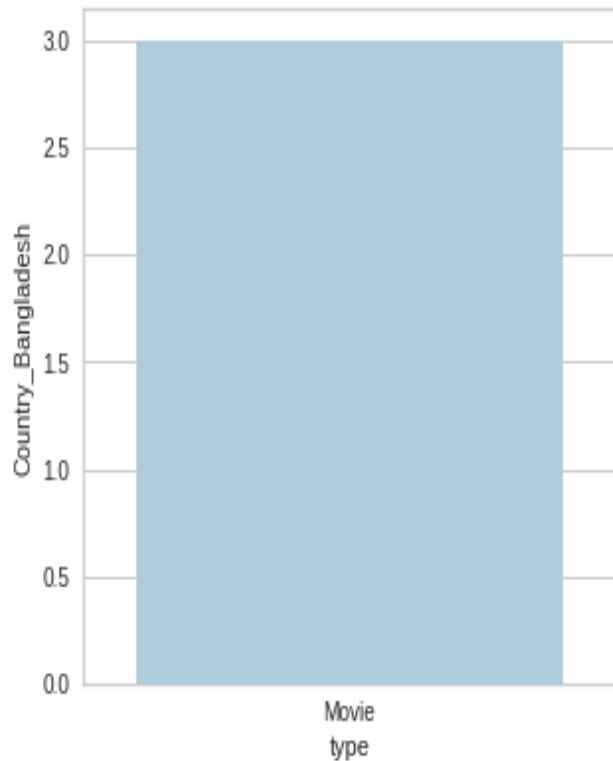
showing only the bar plots for Sri Lanka, China, and India. For other nations, similar bar plots were also created.



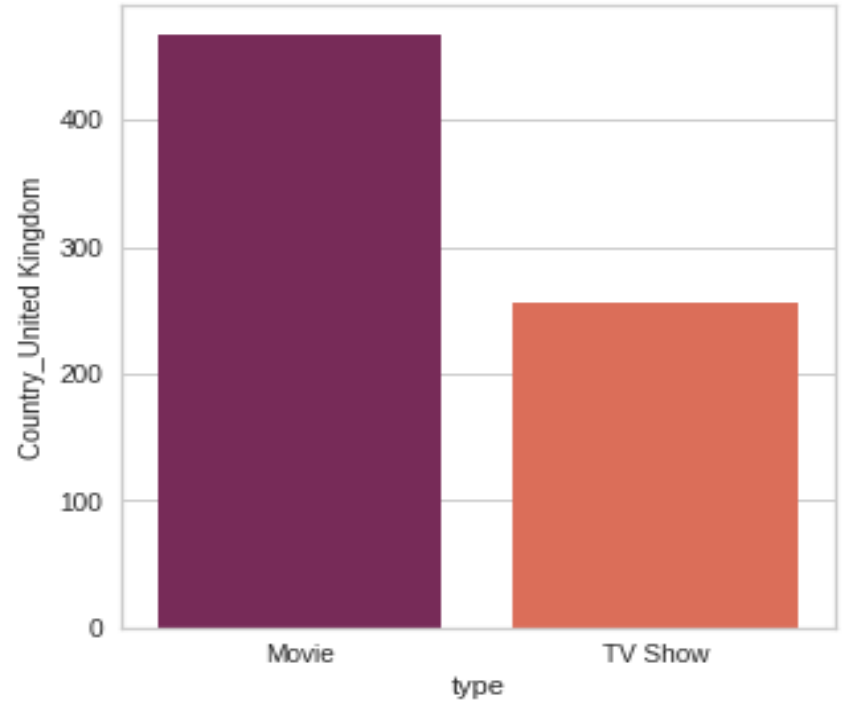
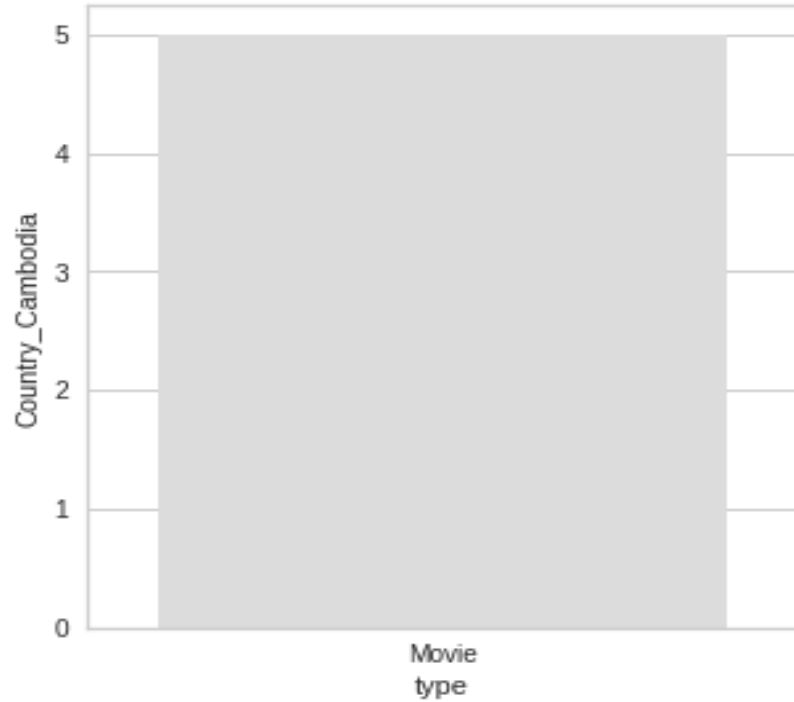
Bar plots for various countries based on the type of content.



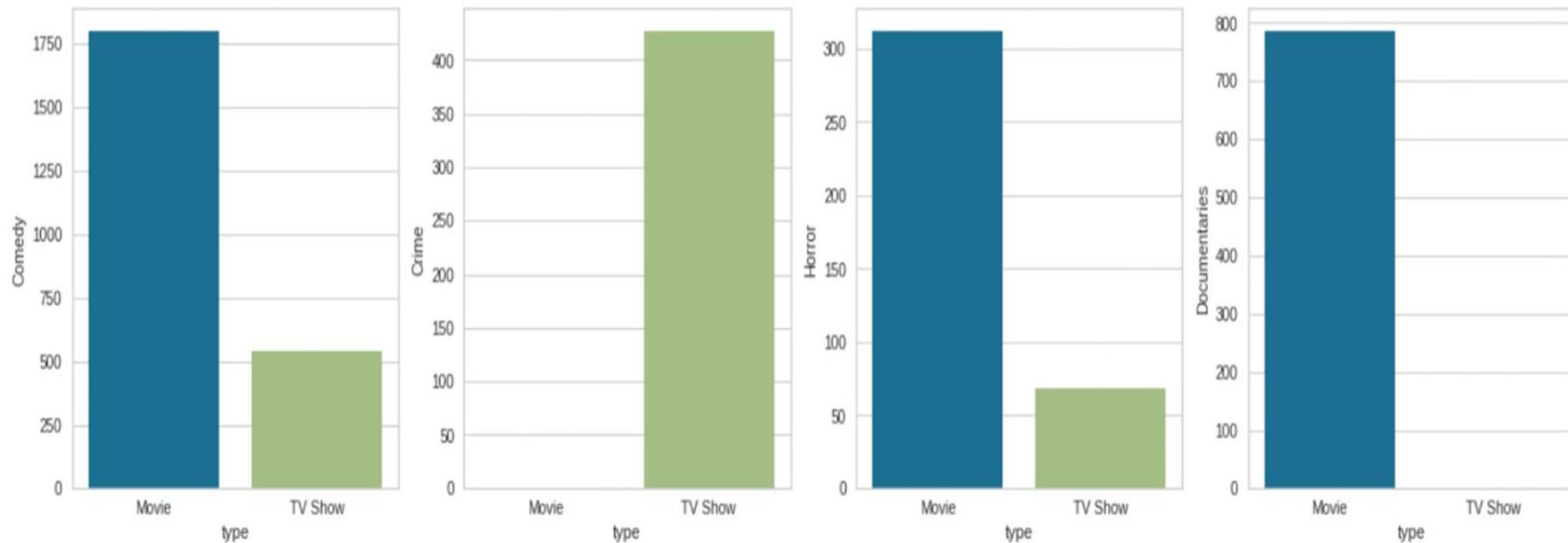
Bar plots for various countries based on the type of content.



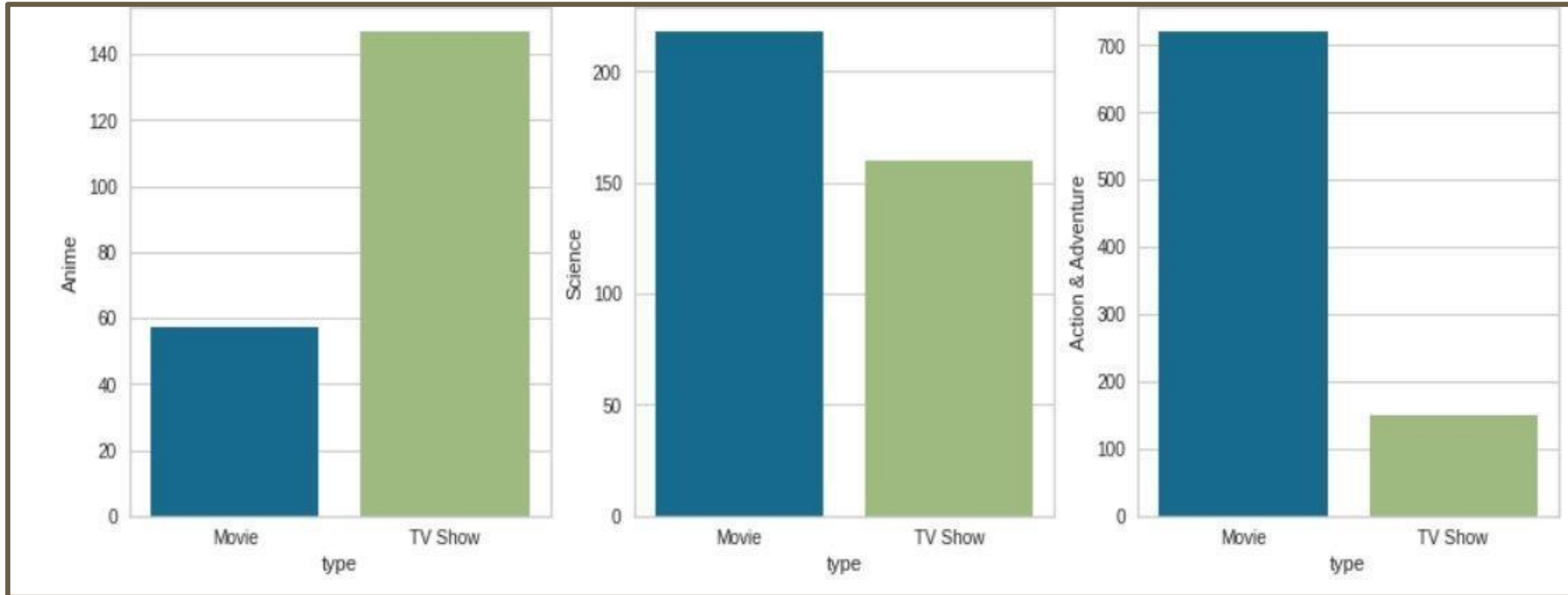
Bar plots for various countries based on the type of content.



Bar plots for various countries based on the type of content.

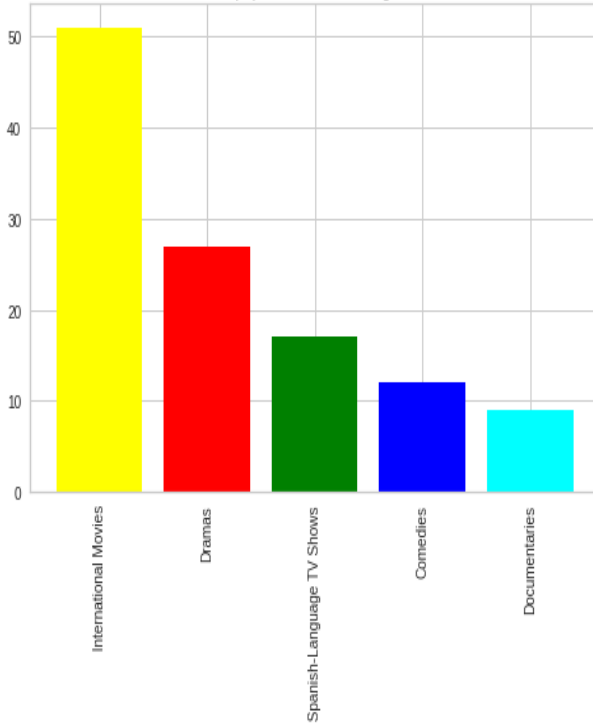


Choosing to display bar plots for three types of anime, science fiction, action, and adventure here. Similar bar plots were created for all genres.

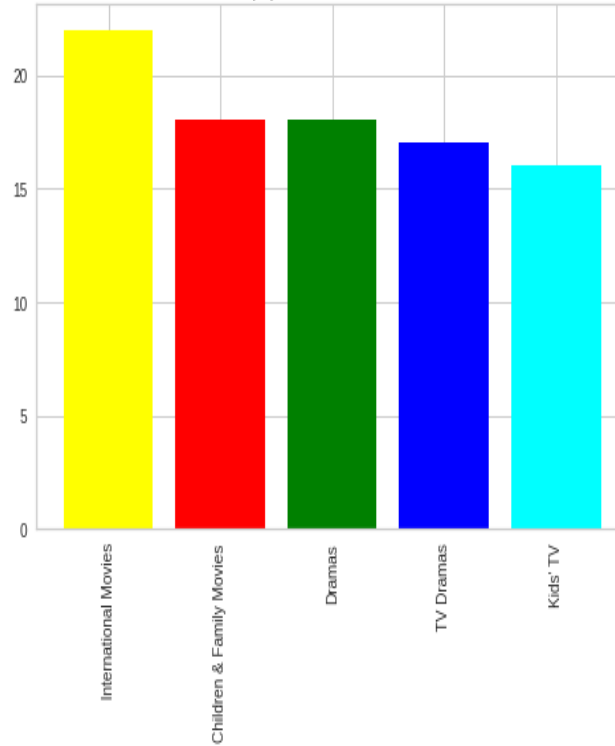


Bar Plots represent the most popular genres in each country.

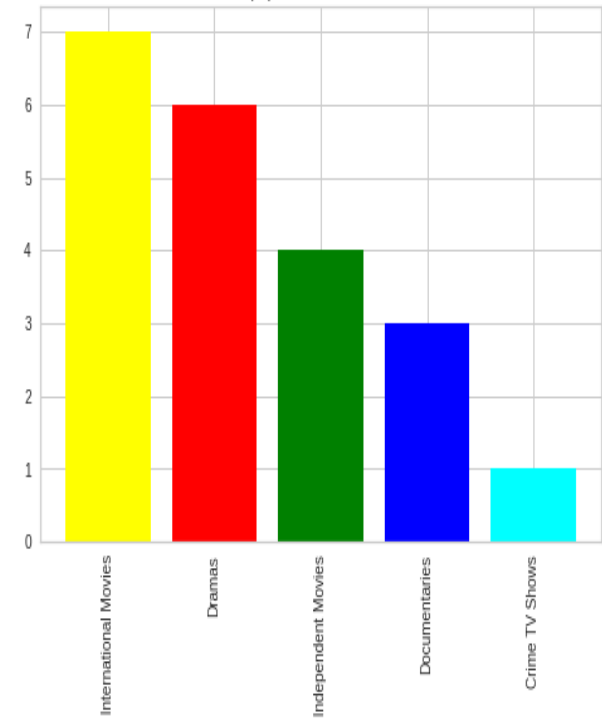
Most popular Genres for Argentina



Most popular Genres for Australia

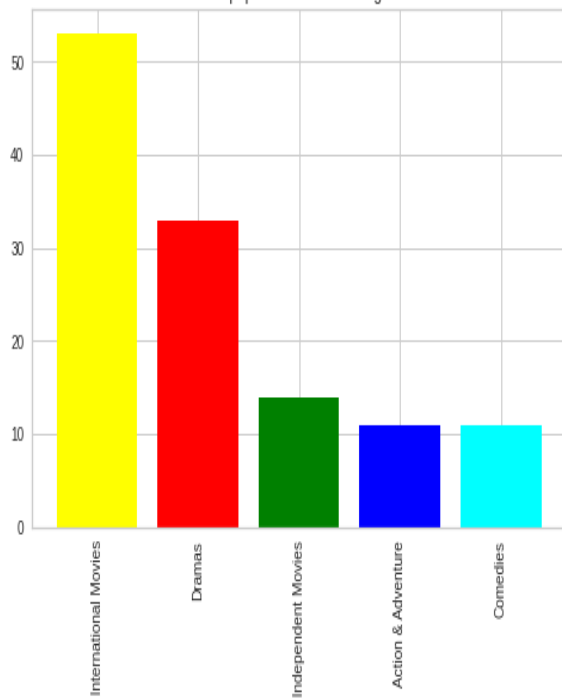


Most popular Genres for Austria

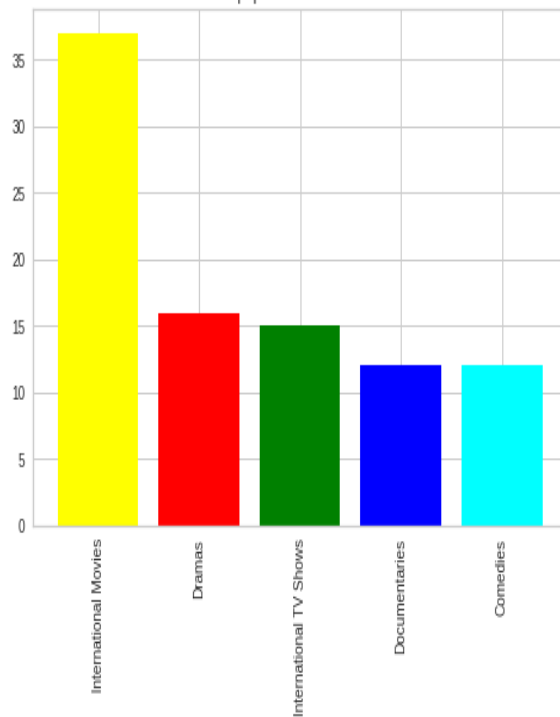


Bar Plots represent the most popular genres in each country.

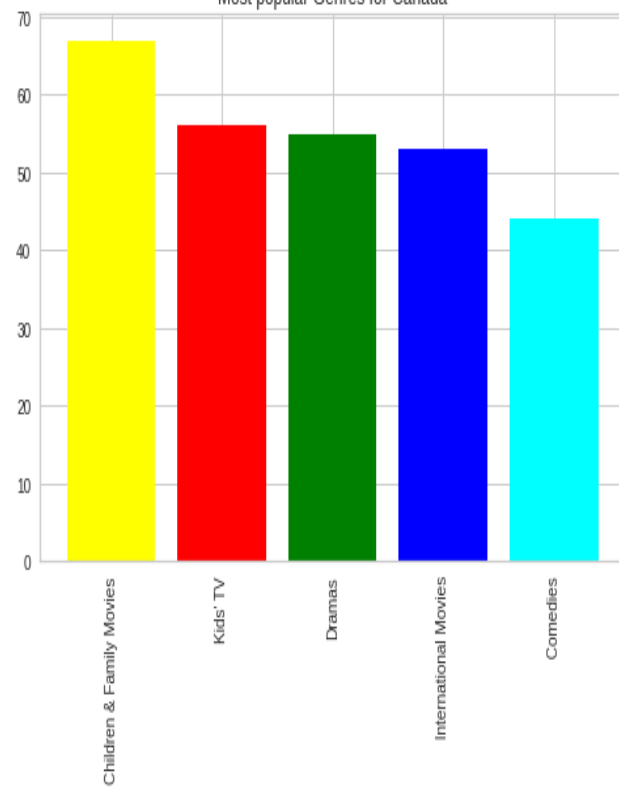
Most popular Genres for Belgium



Most popular Genres for Brazil

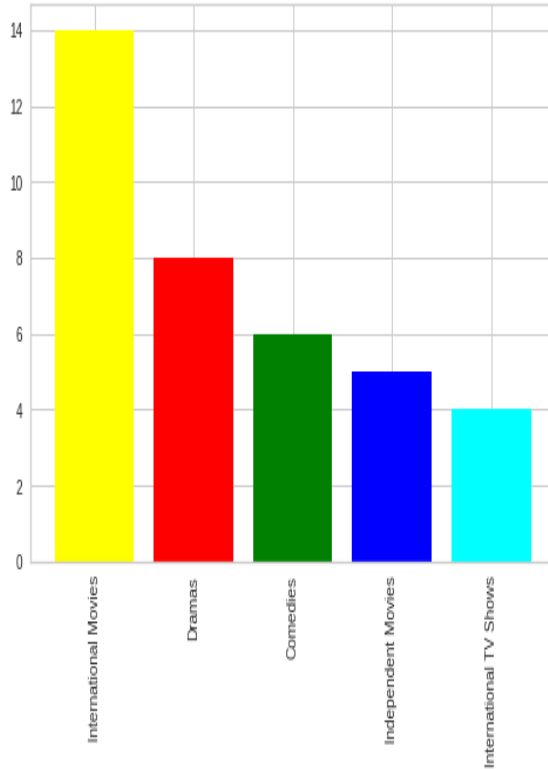


Most popular Genres for Canada

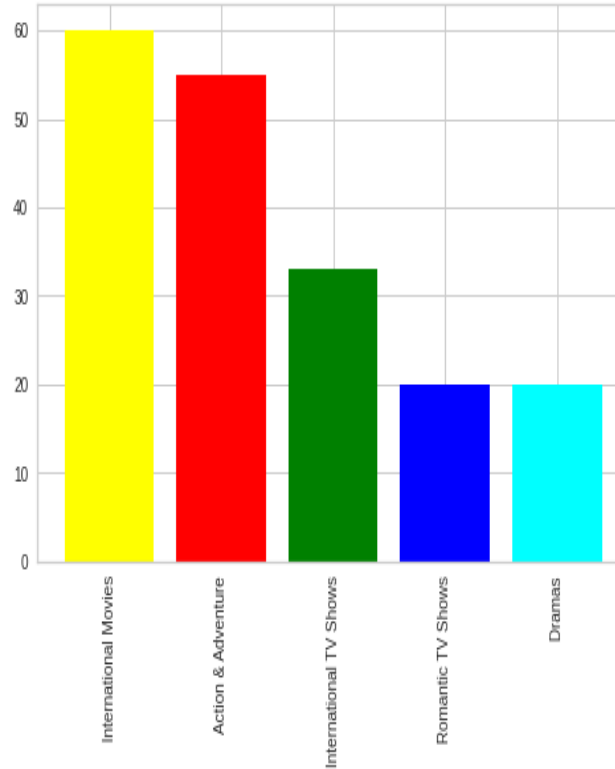


Bar Plots represent the most popular genres in each country.

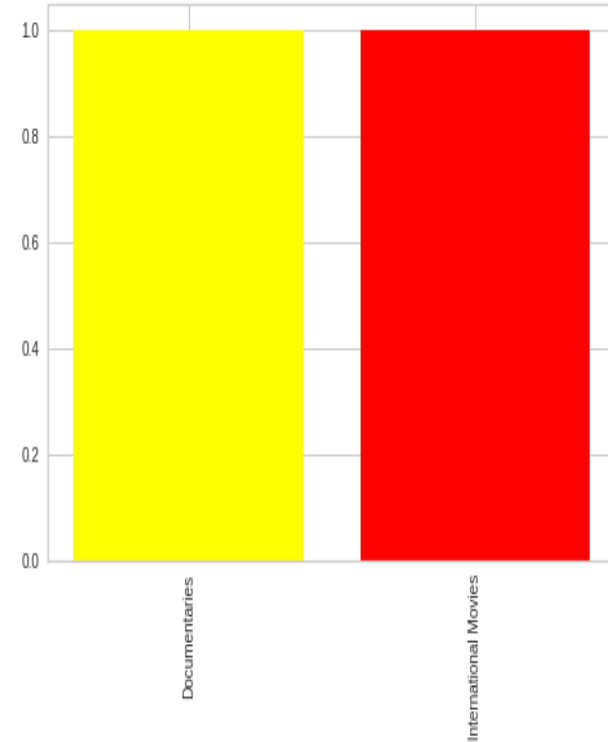
Most popular Genres for Chile



Most popular Genres for China

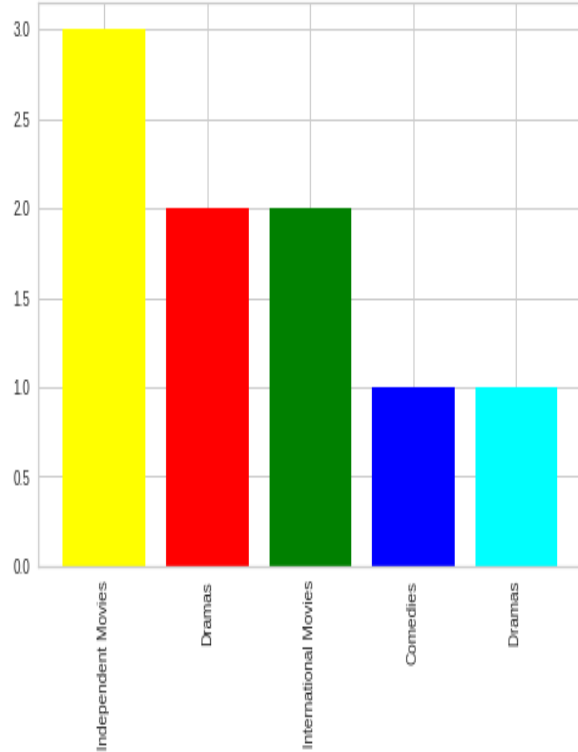


Most popular Genres for Afghanistan

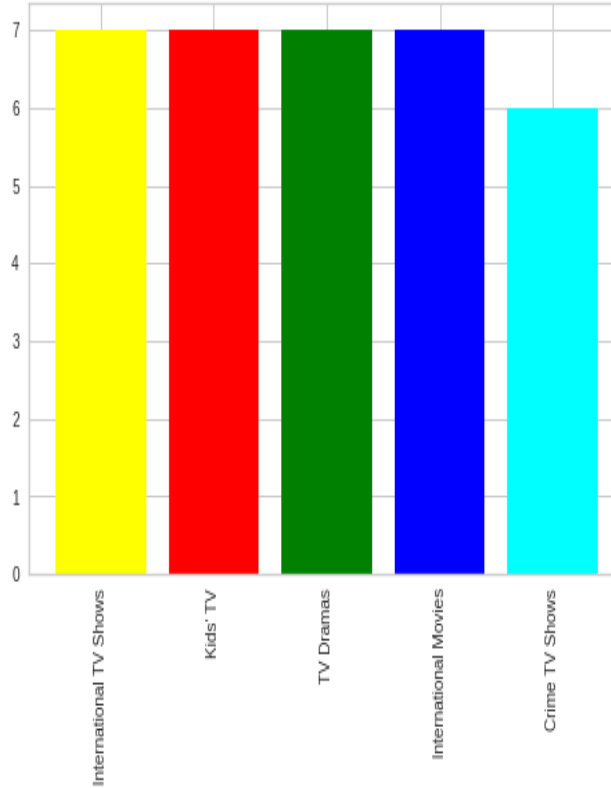


Bar Plots represent the most popular genres in each country.

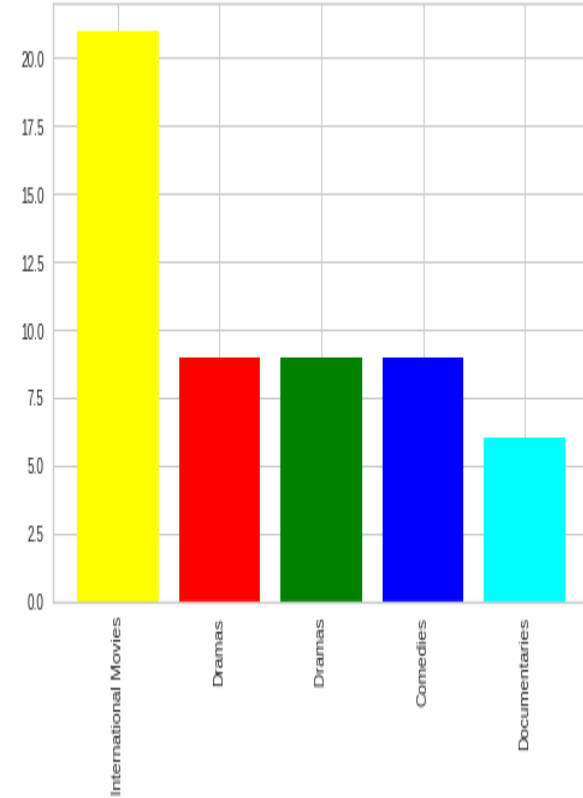
Most popular Genres for Bangladesh



Most popular Genres for Russia

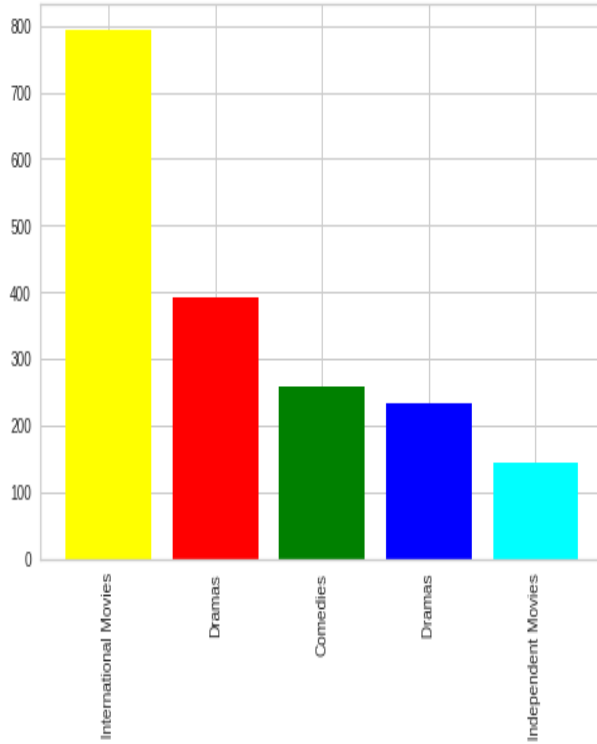


Most popular Genres for United Arab Emirates

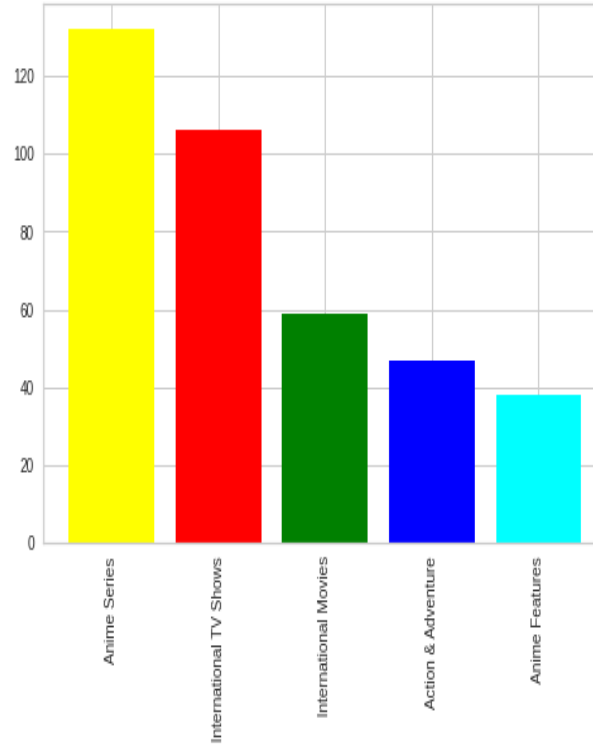


Bar Plots represent the most popular genres in each country.

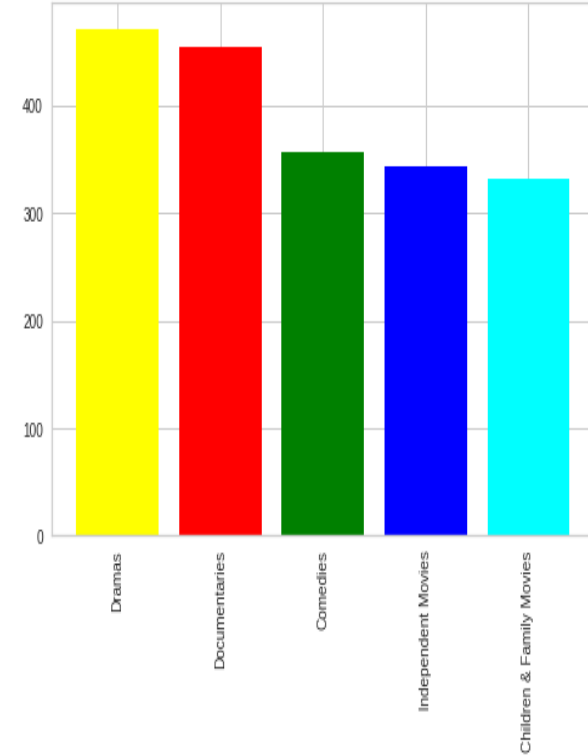
Most popular Genres for India



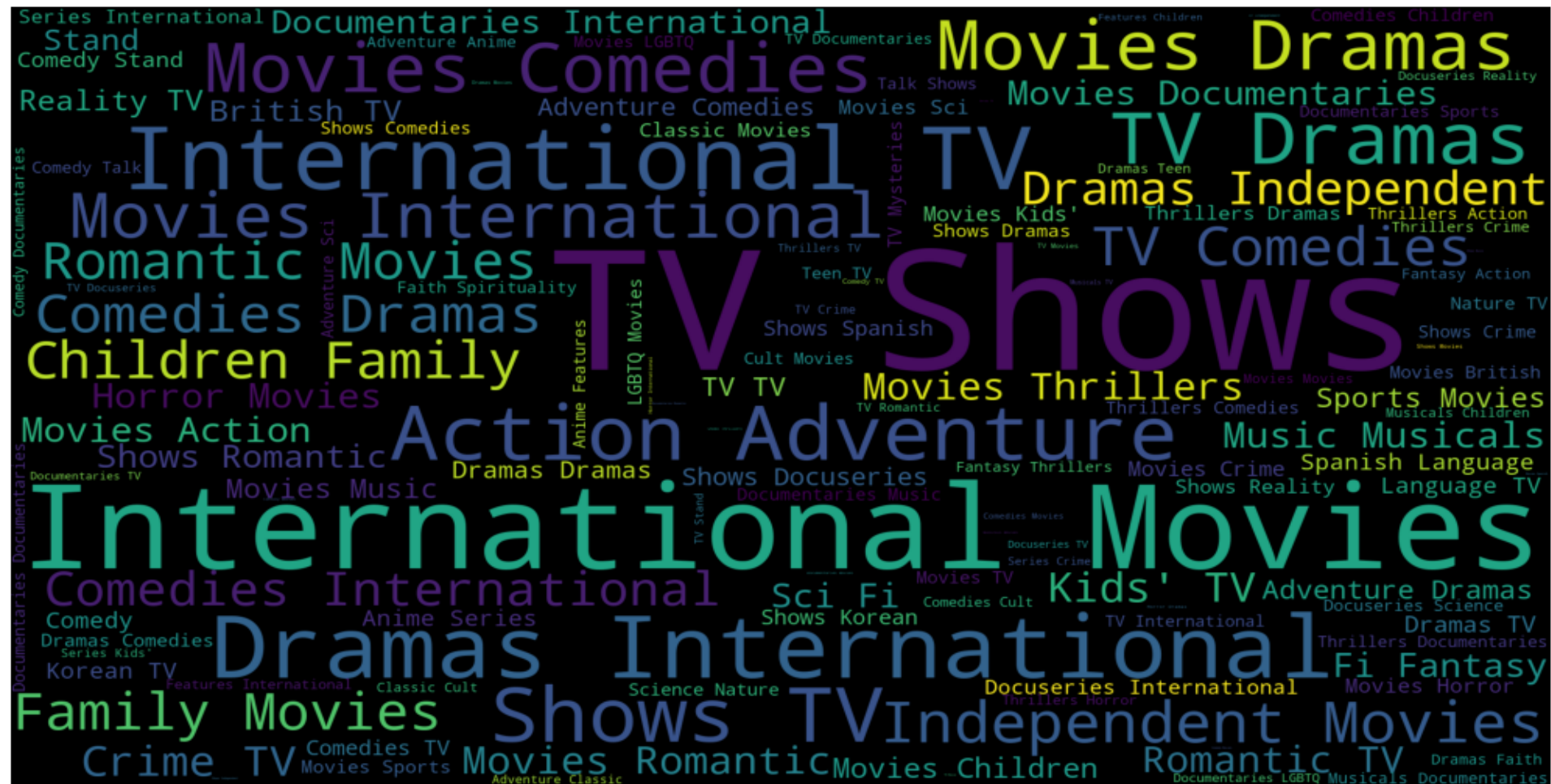
Most popular Genres for Japan



Most popular Genres for United States

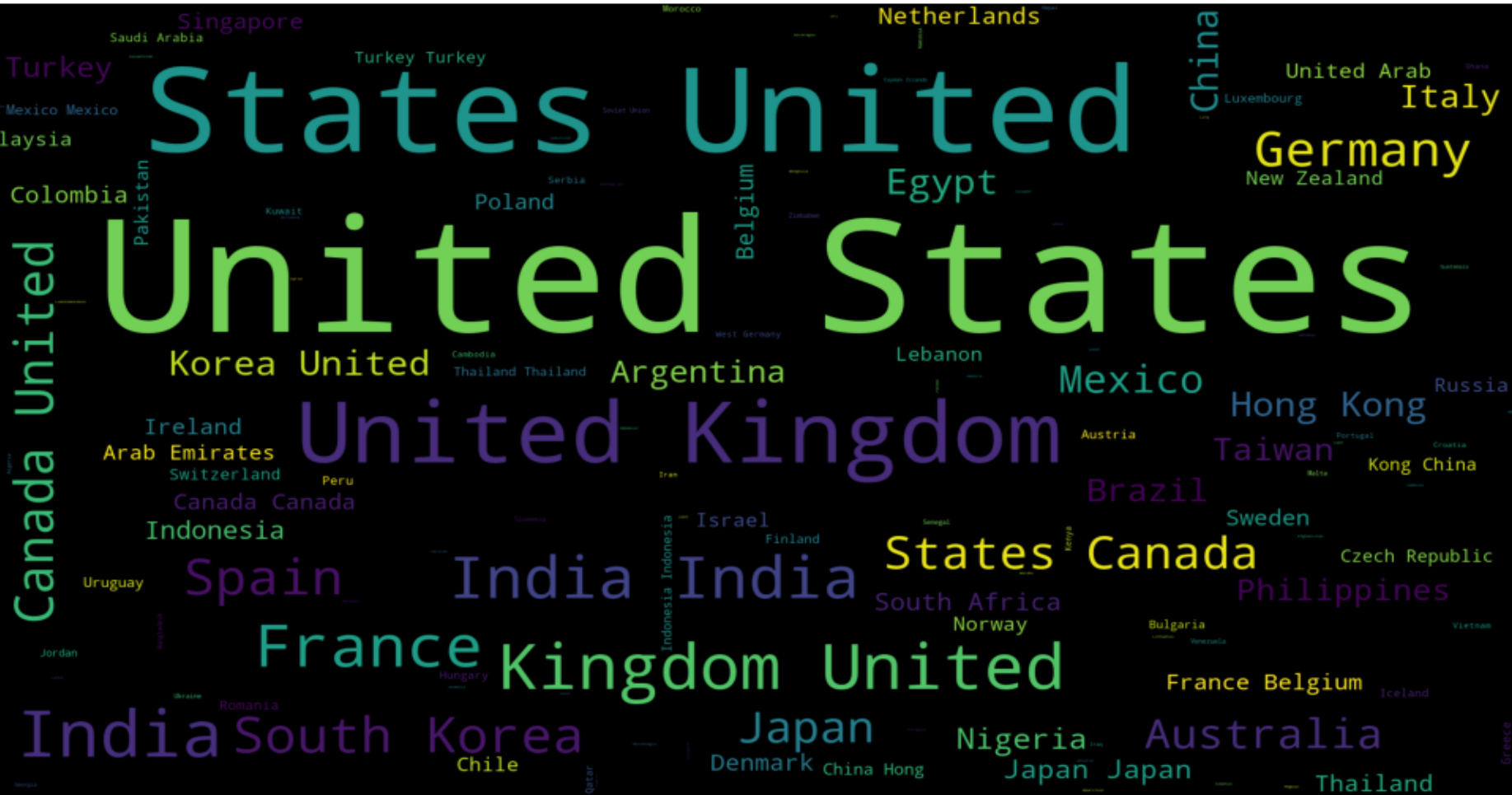


AI



Word cloud for the Netflix content for various countries.

AI

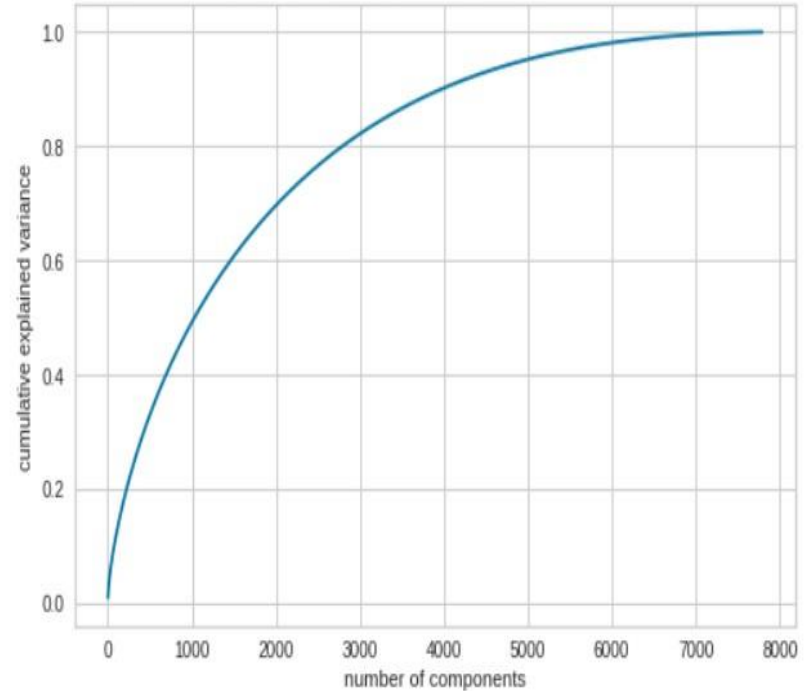


Data transformation

TF-IDF: Term Frequency, Inverse Document Frequency. It indicates the significance/relevance of a 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 a lot of information is also retained. The graph shows the explained variance value for the different number of PCA components. n_components 4000 was chosen in this project.

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



Implementation of Clustering Algorithm

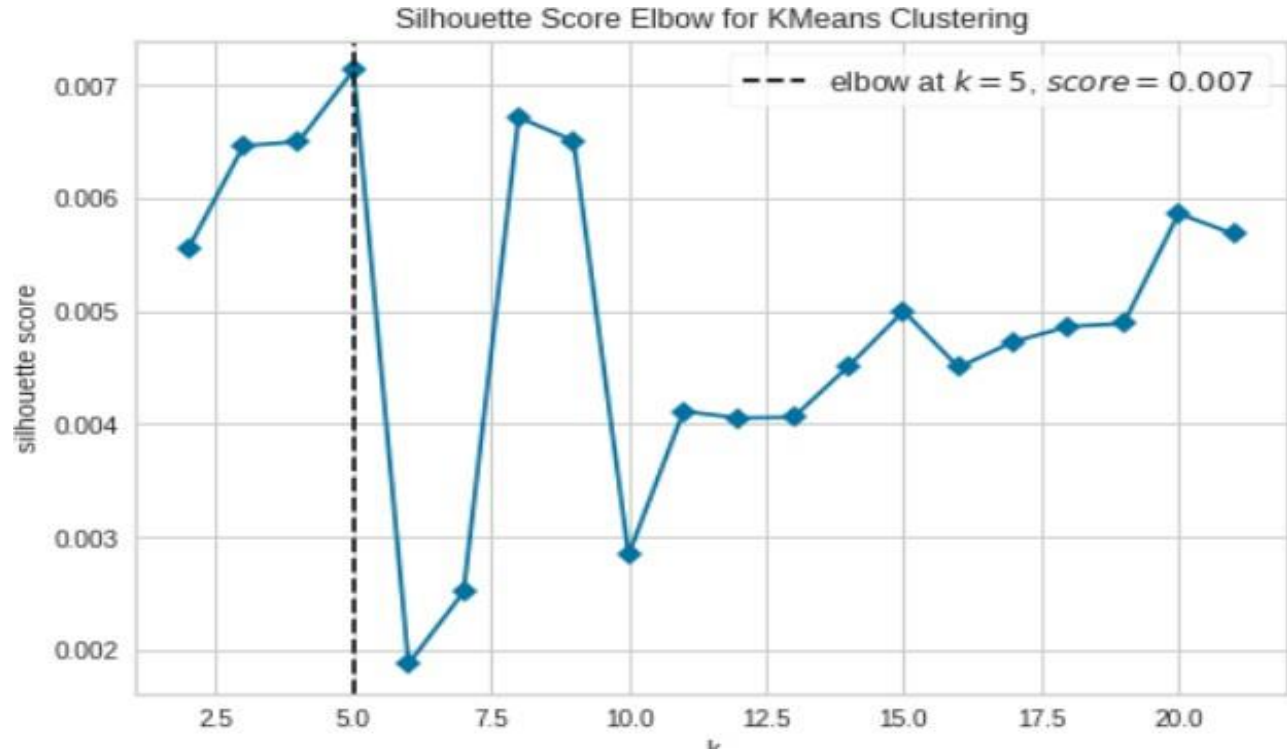
Textual data were combined and converted into numerical data using TF-IDF. Applied PCA to perform dimensionality reduction. Data were converted to 4000-dimensional data. K-Means is used in this project.

➤ Some Clustering algorithms:

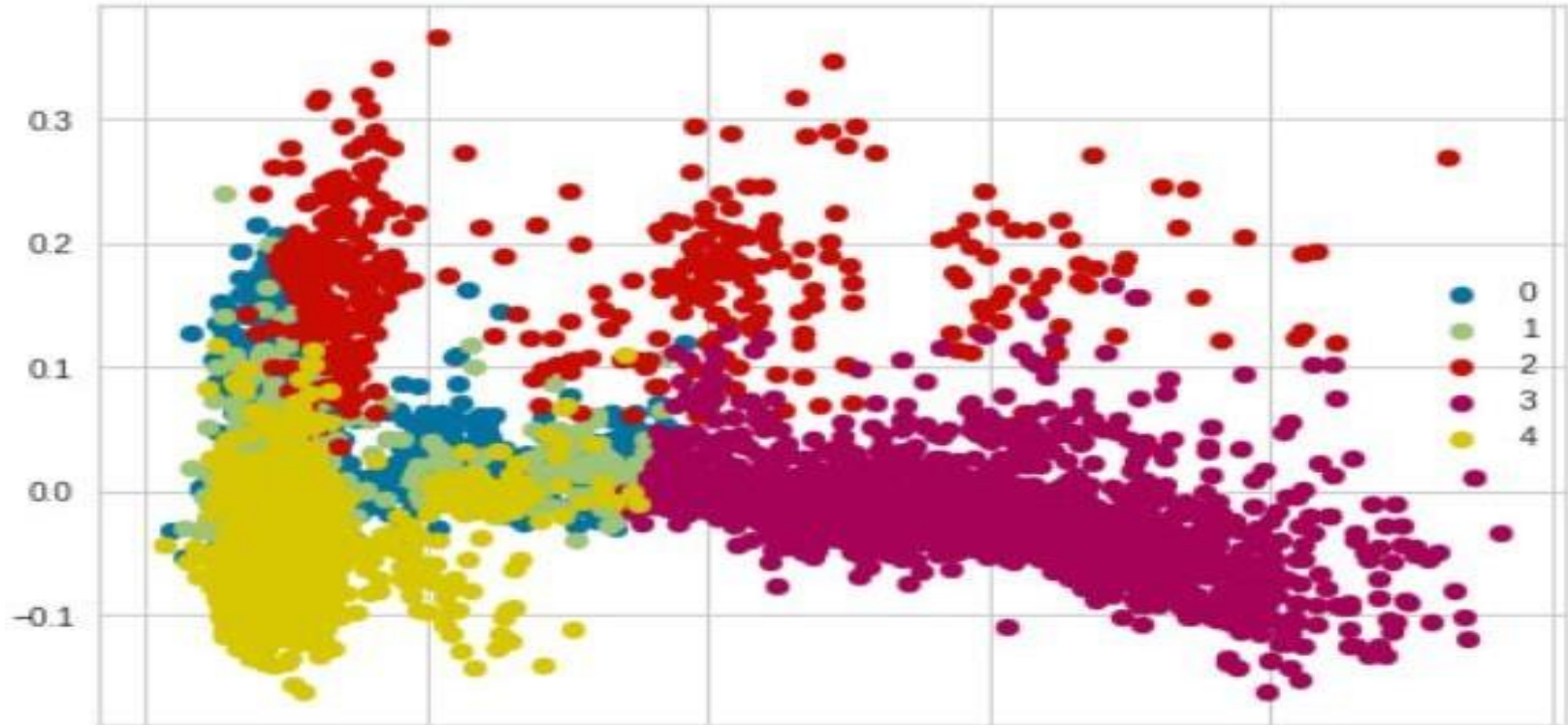
1. K-Means Clustering
2. Gaussian Clustering
3. Agglomerative clustering

K-Means Clustering

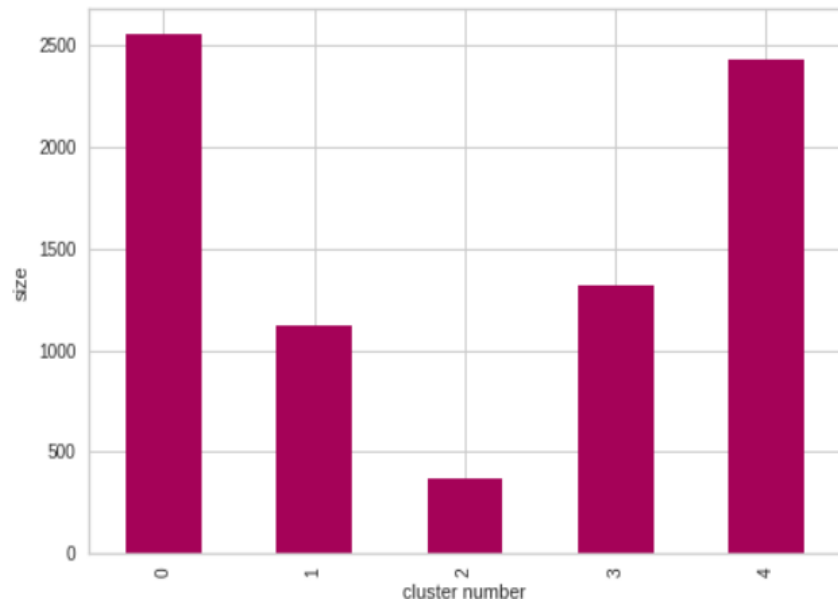
The maximum score for the silhouette and elbow are both at cluster number 5, which was regarded as the ideal number of clusters.



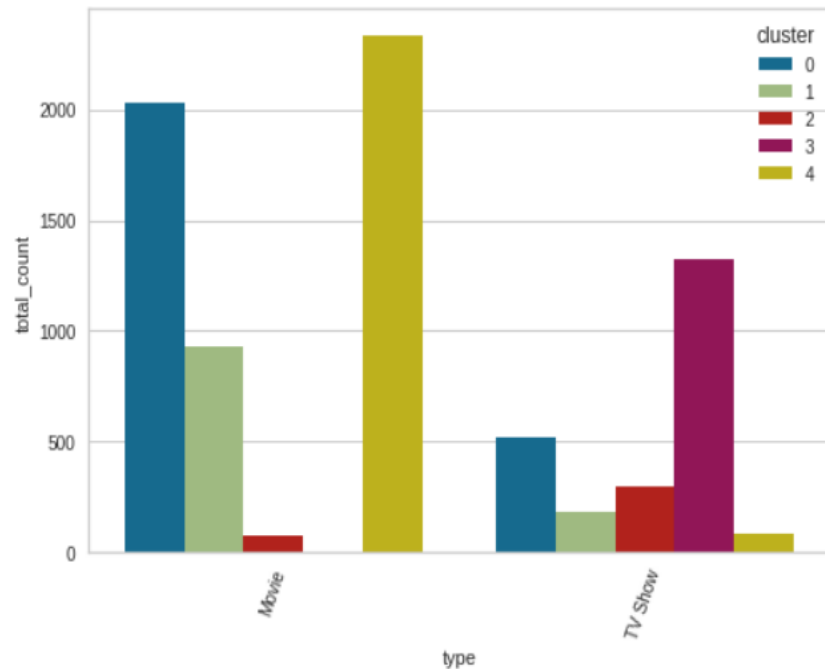
Clusters for K-Means Clustering



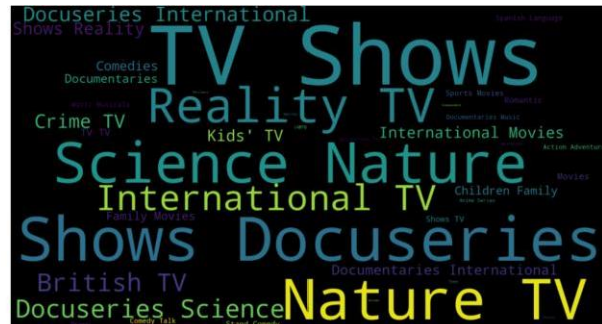
Cluster Analysis



Size of clusters



Cluster-wise content Type



Content Based Recommender System

Count Vectorizer: Textual data was transformed using Count Vectorizer.

Cosine Similarity: Cosine similarity was used to locate related movies/TV series.

```
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

1. EDA was used to examine the data.
2. Textual data were converted using stemming. Data cleansing for the textual data was done.
3. Textual data were transformed using TF-IDF.
4. Dimensionality was reduced by using PCA. The explained variance vs. components graph was used to choose the components.
5. Clustering using K-means was used.
6. Using the elbow curve graph and the Silhouette score, the ideal number of clusters was discovered.
7. The characteristics of several clusters were compared.
8. Utilizing a count vectorizer, a content-based recommendation system was developed. For any movie name entered, it suggests 10 other films or TV shows that are similar.

THANK YOU