Netflix Movies & TV Shows Clustering

Ashik Kumar

Abstract

With the advent of streaming platforms, there's no doubt that Netflix has become one of the important platforms for streaming. The dataset that we have used for EDA and clustering has been collected by Flixable, a third-party Netflix search engine. There are 12 features and around 7700 observations in the dataset and are mostly textual features.

Through univariate and multivariate analysis, we found trends that will help in understanding what content is being consumed country-wise, depending on some categorical features like rating, type, genres, cast, directors, etc. Clustering was performed along with NLP on textual columns and then a mini-recommendation system was built out of it.

Keywords—Machine Learning, Explanatory Data Analysis, Netflix, TV Shows, Movies, Genre, Clustering, K Means.

Introduction

Unsupervised Learning is a machine learning technique in which the models are not supervised by the training set instead we find hidden patterns and insights from the given data. It is a machine learning technique in which models are trained on the unlabeled data set without any supervision. A cluster is a collection of elements that are similar to each other but dissimilar to the elements belonging to other clusters. Clustering can be done using various kinds of distances such as euclidean distance, manhattan distance, gomer distance, etc. We can do different kinds of clustering based on the data pattern in space such as spherical clustering, K-means clustering, etc.

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.

Our goal here is to make an unsupervised clustering model, which will help in garnering insights on Netflix and how its content is being consumed.

A brief summary of the dataset is given below:

Feature Name	Feature Information
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 Releaseyear 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**

Design and Methodology

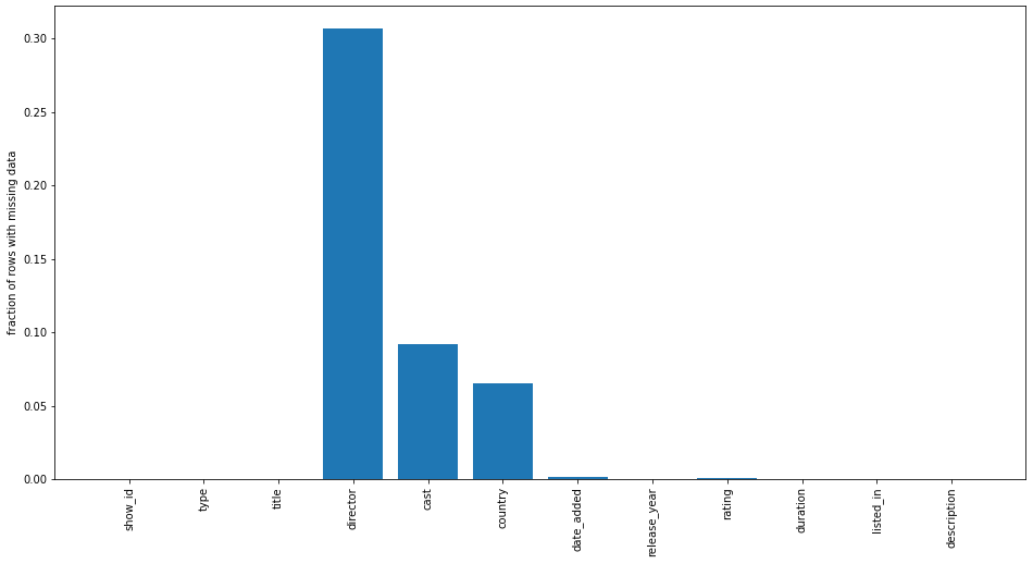
In this section, we will discuss the framework, extraction and preprocessing features, feature selection, and clustering algorithms.

Exploratory Data Analysis

The first step involved in the analysis is to load the dataset into the pandas data frame. Before exploring the data using different libraries available in python we should if the dataset is ready to run the operations on it.

- **Data Cleaning:** Data Cleaning is one of the important steps before we start building models, in fact, there will be a significant increase in Model Performance when we have a clean, rich dataset. So here, we decided to replace null values with an empty string.
  - There are 2389 null values in Director column
  - There are 718 null values in cast column
  - There are 507 null values in country column
  - There are 10 null values in date added column

- There are 7 null values in rating column



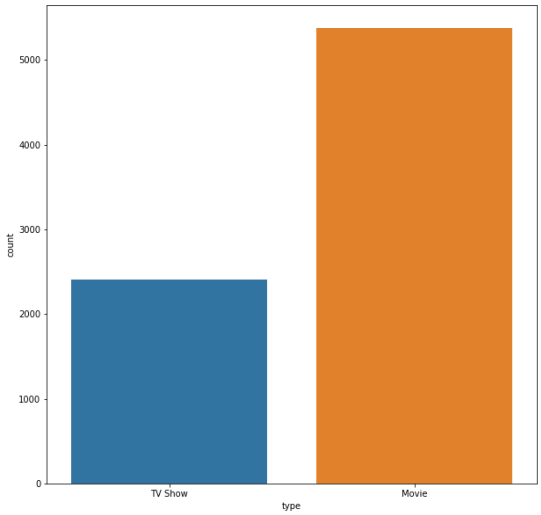
Null value visualisation

❖ **Handling Duplicates:** There were no duplicates.

❖ **Univariate Analysis**

As the name indicates the analysis was done on each one of the columns. We can check the unique values of the features and their frequency in the dataset.

1. **Type Column:**

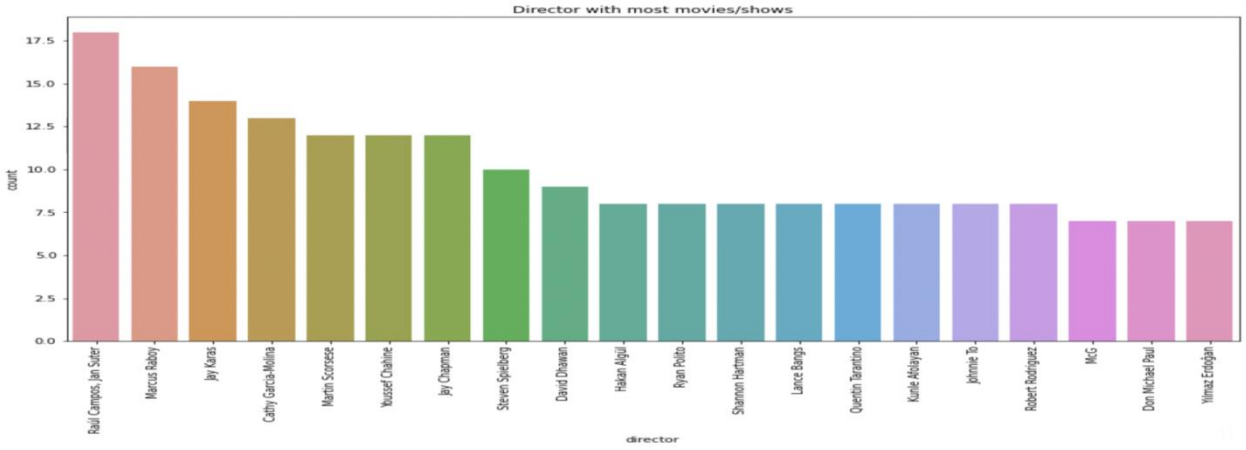


Inference:Almost 70% of datapoints belong to Movie, rest 30% to TV Show

2. **Title Column:**

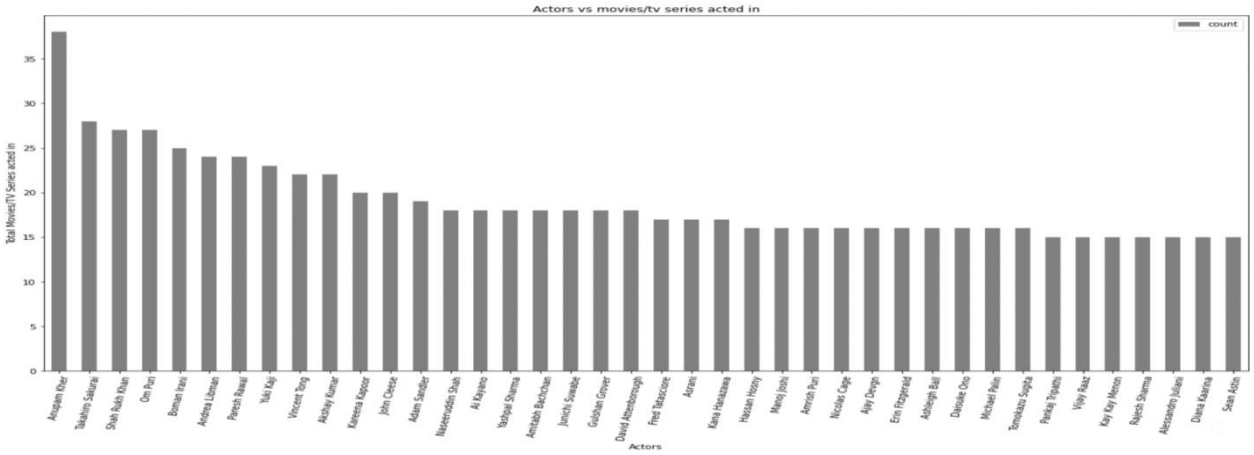
Inference: It seems like words like "Love", "Man", "World", "Story" are very common in titles.However, we are surprised by the overwhelming number of content having "Christmas" in their title.We are suspecting "Christmas" titles to be a very seasonal thing with most of the shows likely to be released during December. However, we do not have the release month in this data to confirm our hypothesis.

3. Director Column



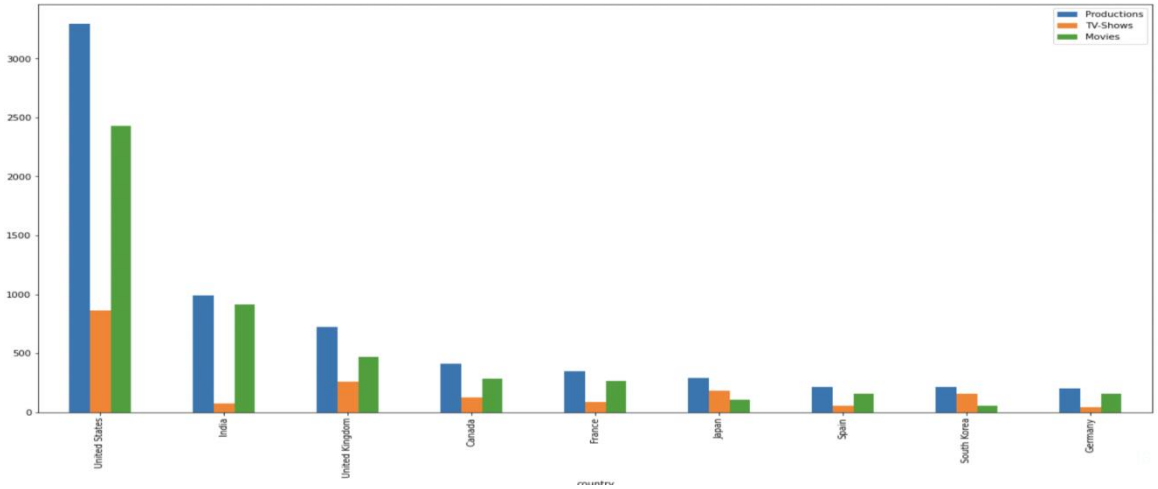
Top 20 Directors with most number of Movies/Shows on Netflix

4. Cast column



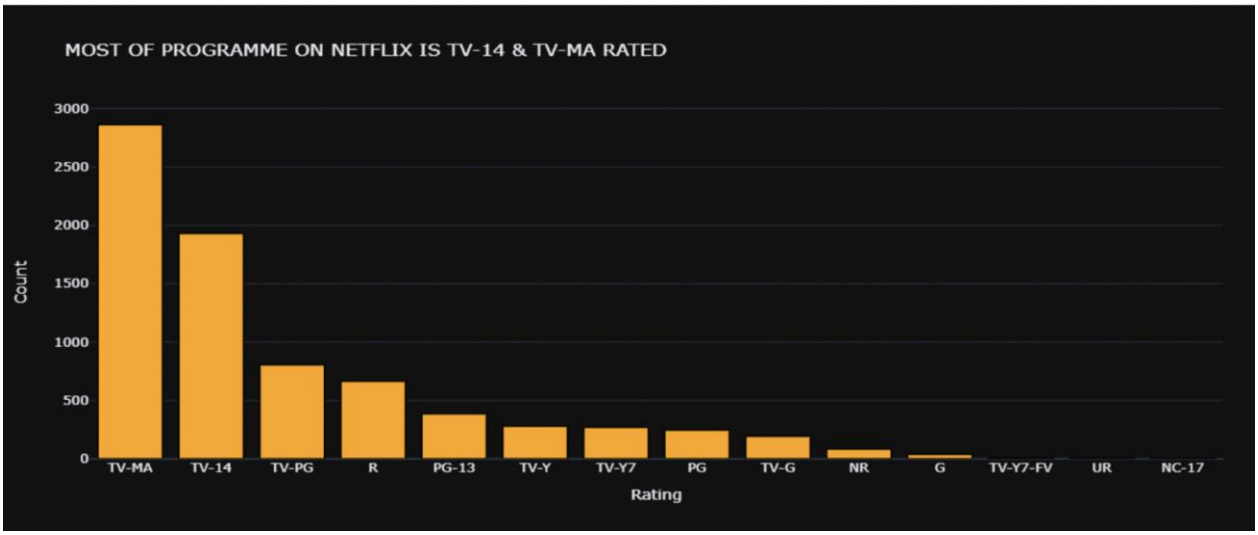
Top Actors with highest count of Movies

5. Country Column



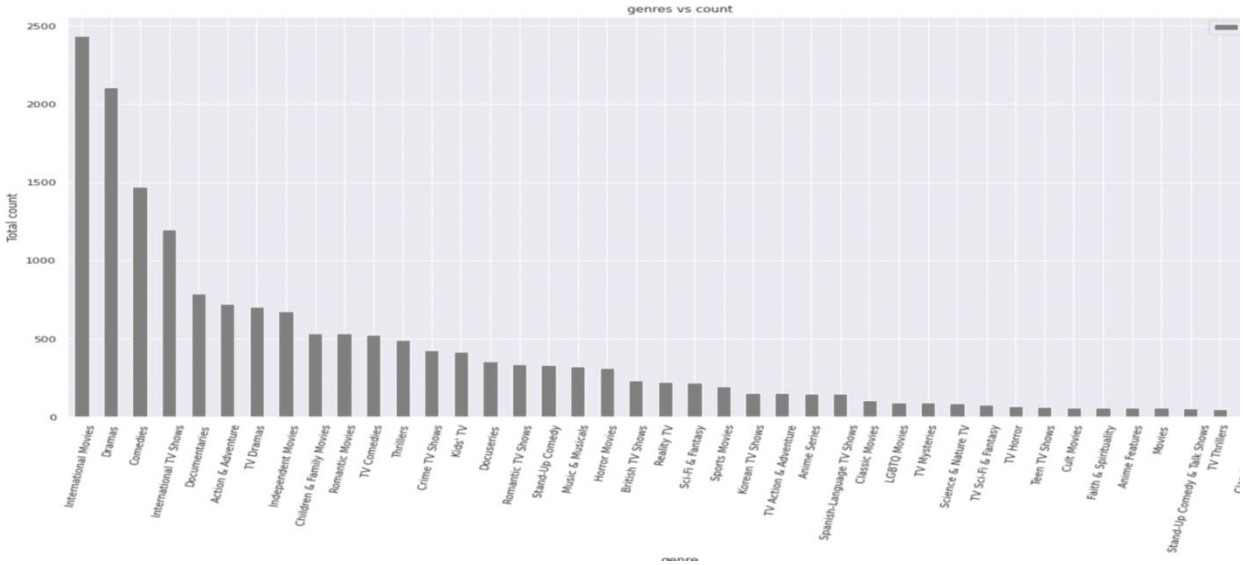
Top 20 Countries with more number of Productions

6. Content Rating Column



Most of the programmes on Netflix are TV-14 & TV-MA Rated.

7. Listed In Column

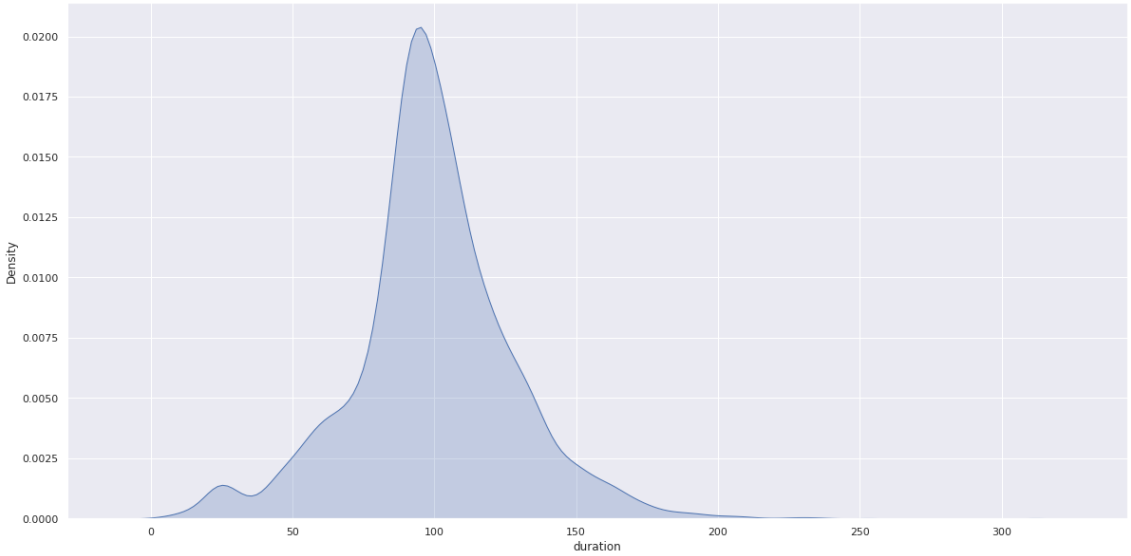


Inference: International Movies make up the top most genre!

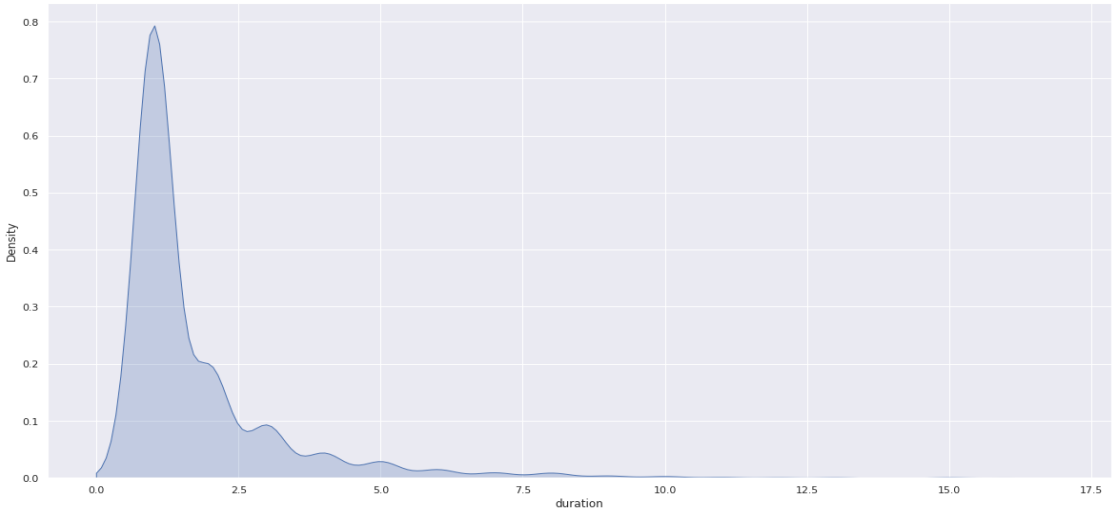
8. Description Column

Inference:Most words like Life, family popping up!

9. Duration Column



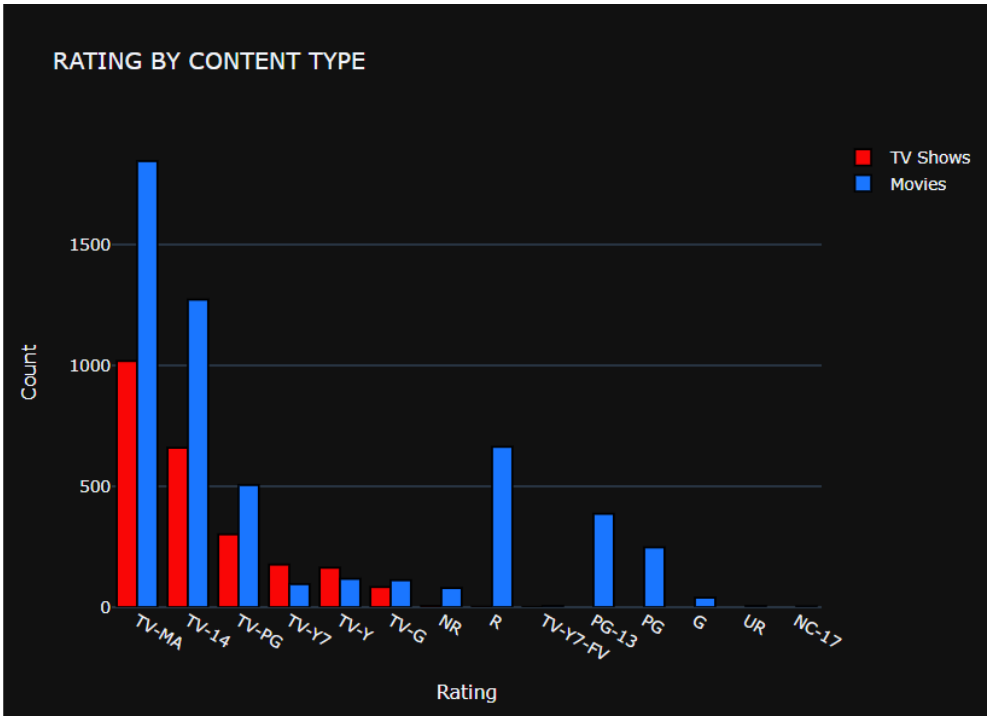
Inference: Most content are about 70 to 120 min duration for movies



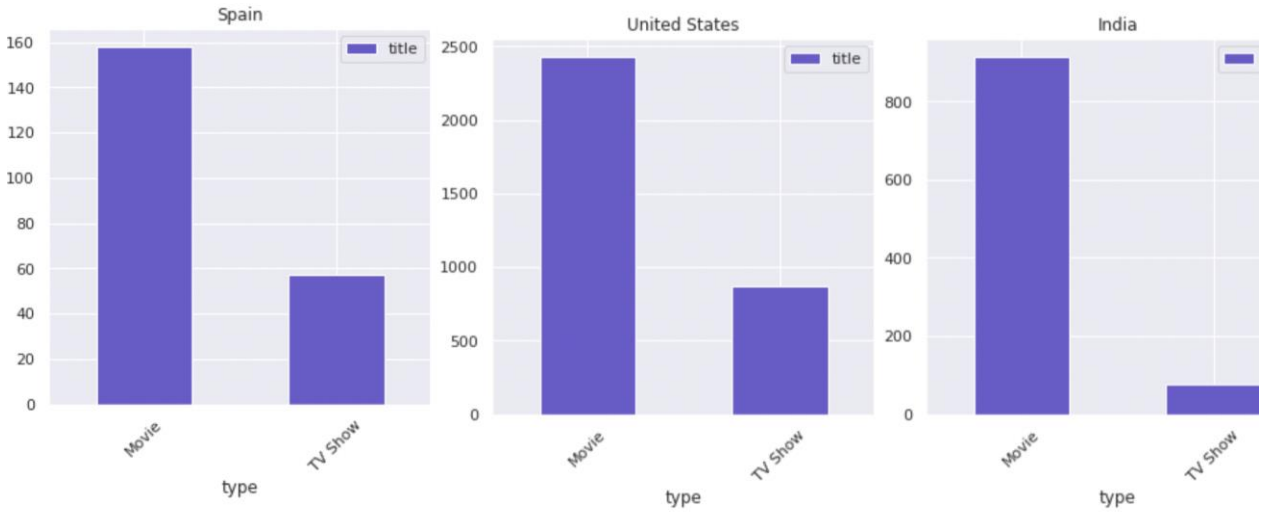
Inference: Most of the shows are 1 to 2 seasons long!

❖ Multivariate Analysis

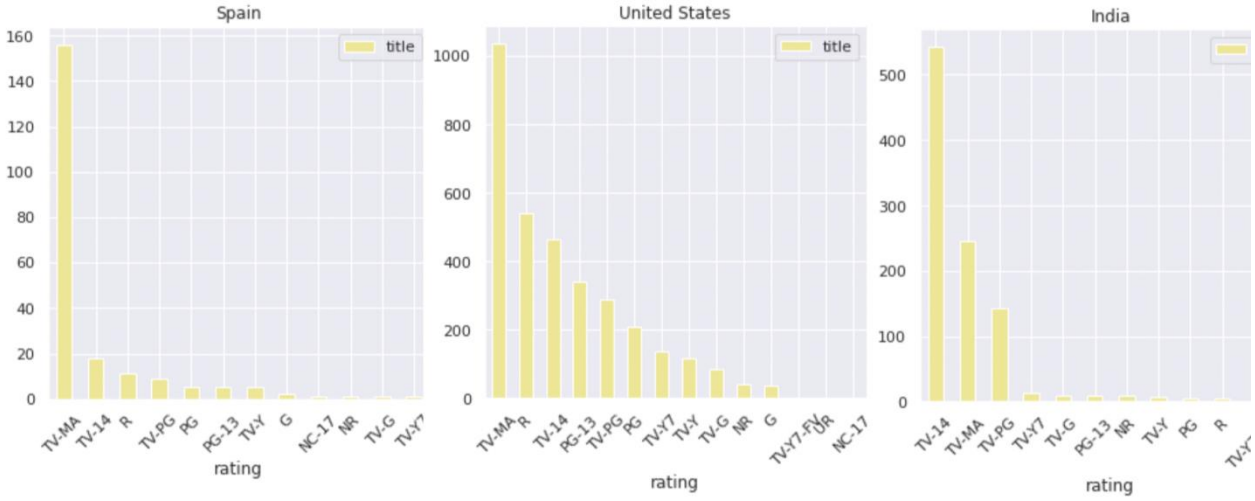
1) Understand content rating split between Movies & TV shows



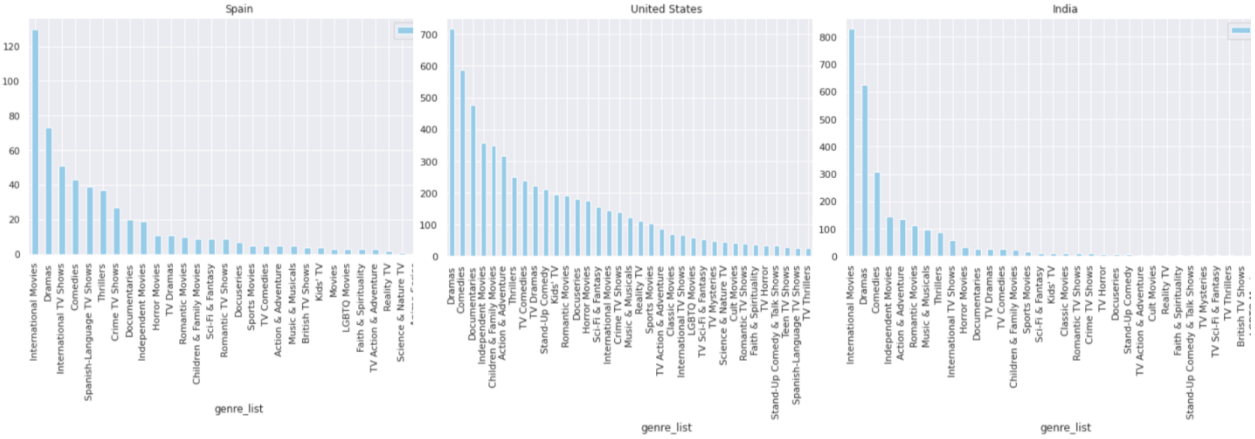
2) Understanding what type of content is available in different Countries



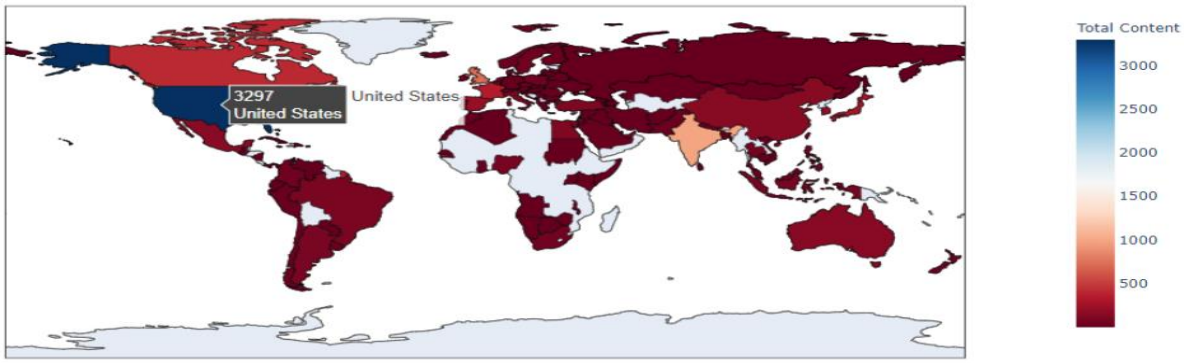
All the countries have preferred Movies over TV shows!!



Spain and US prefers TV-Mature content, while India prefers TV -14 (unsuitable for children under the age of 14)

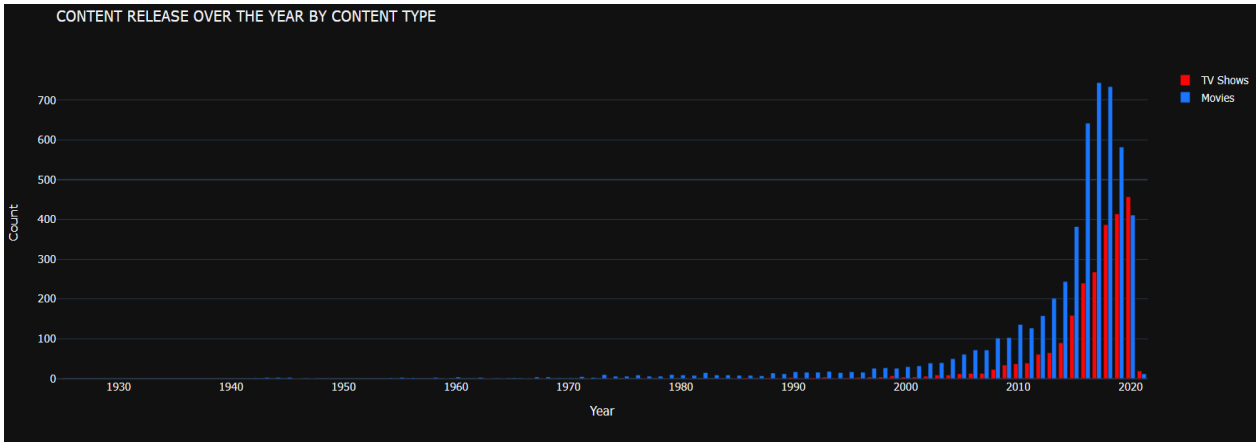


Spain & India has more International Movies, while US has more Drama Content.



Number of content produced by different countries

3) Is Netflix increasingly focused on TV rather than Movies in recent years?



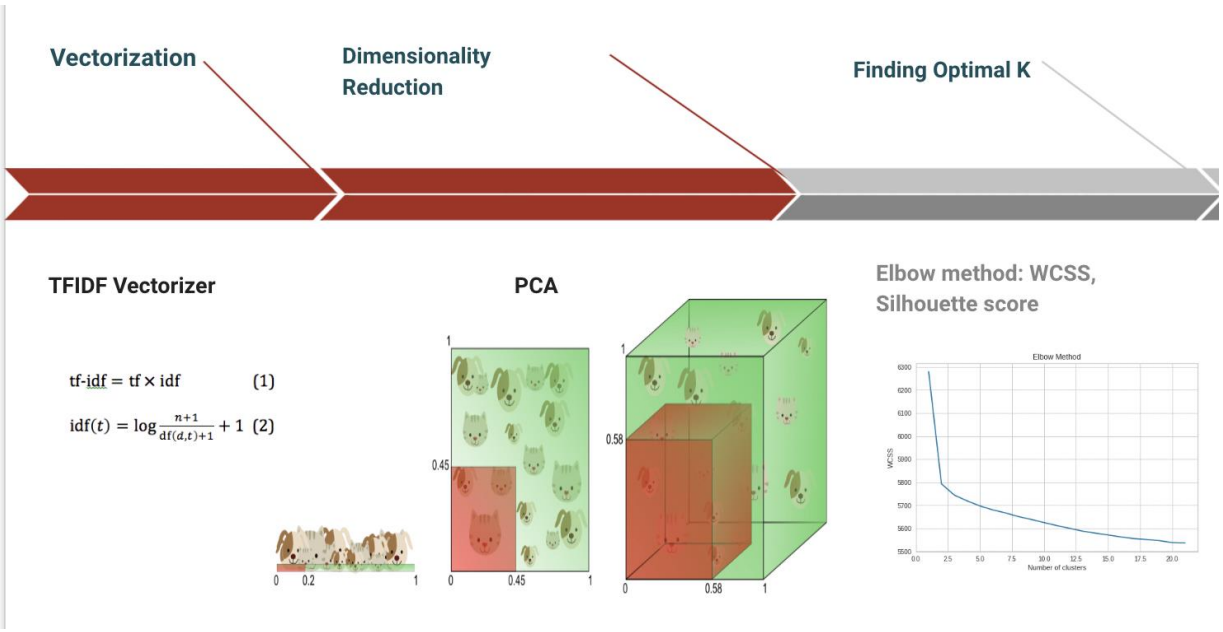
Yes, Netflix is increasingly focusing on TV Shows now, which is clear from the graph, in 2020, there were more Shows than Movies. Also, Movies preference shows a declining graph, while shows are increasing.

❖ Textual Columns

We clubbed 6 textual columns together and merged them into one final feature, which we used for clustering. We first started off by replacing null values in the columns with an empty string, followed by the removal of stopwords, tokenization, and stemming.

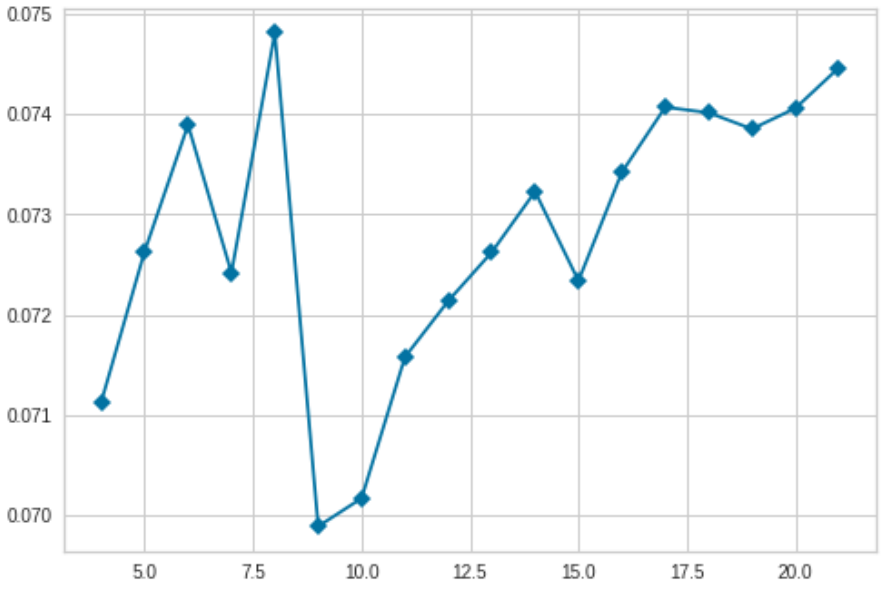
1. CLEANING	2. STOPWORDS	3. TOKENIZATION	4. STEMMING
• Cleaned Null values • All Columns: Only characters selected by regex • All words to lowercase • Merged text columns	• Removed Stop words • Normal english words & problem specific	• Splitted sentences to tokens • Used word_tokenise from nltk	• Transformed words to roots • Used Snowball Stemmer

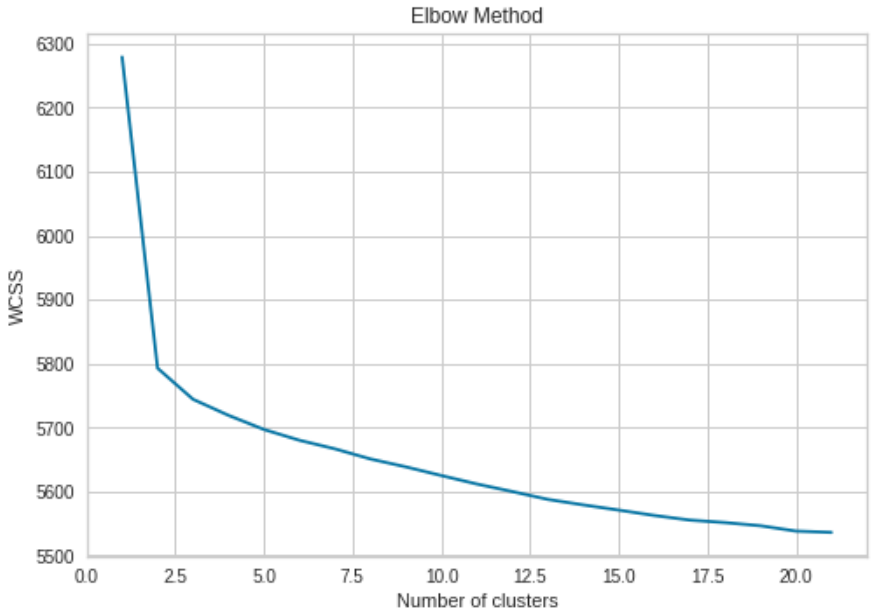
Finally, after we were done with textual preprocessing, we performed vectorization of the final text column using TFIDF followed by dimensionality reduction using PCA.



❖ Clustering

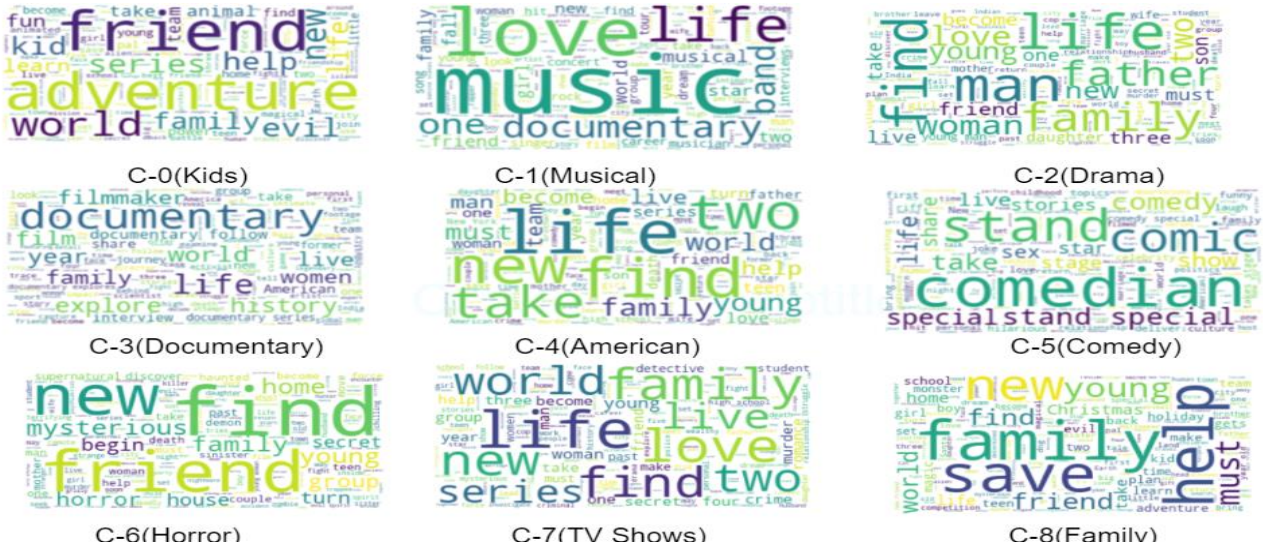
We have used a K Means clustering with an optimal value of k=9 for clustering the Text Based Columns. The value of k was chosen from the elbow method of WCSS score and silhouette score.





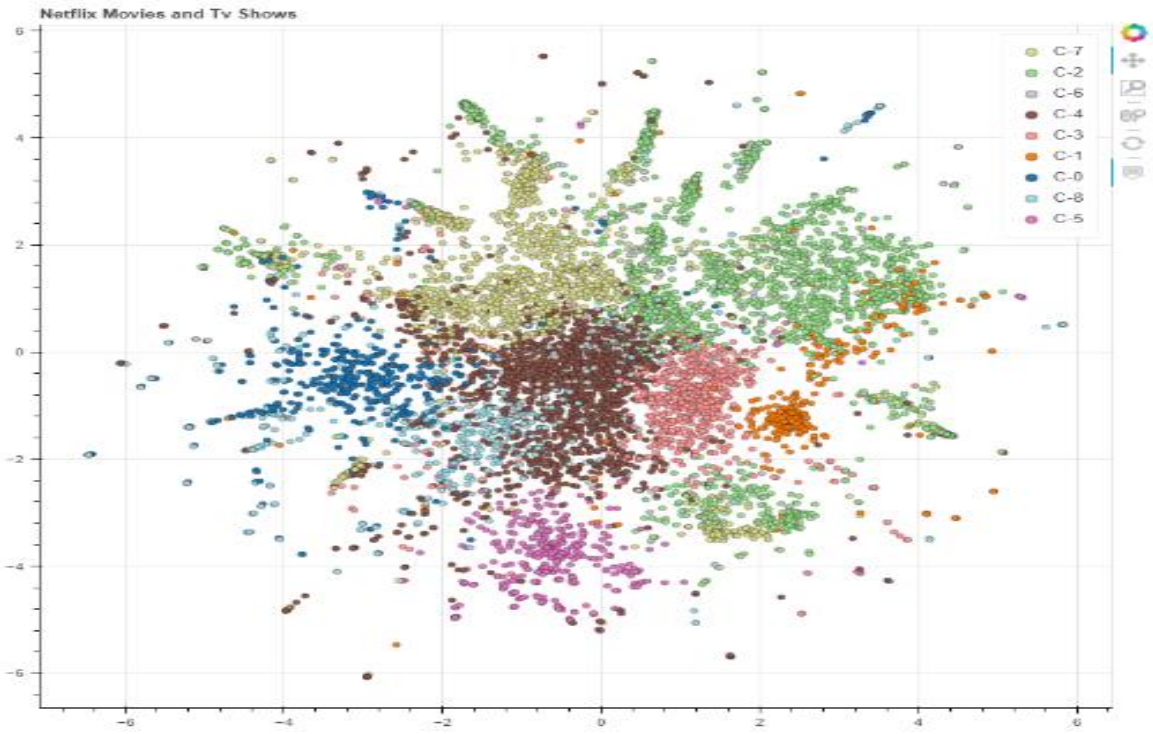
We were able to come up with the following clusters and name it accordingly.

1. Cluster 0: Kids, Animation & Anime
2. Cluster 1: Musical International & Indian
3. Cluster 2: Drama, International, Indian
4. Cluster 3: Documentaries, Sports
5. Cluster 4: Drama, American, Adventure
6. Cluster 5: Comedy
7. Cluster 6: Horror
8. Cluster 7: International TV Shows
9. Cluster 8: Family Movies



We have also done an interactive clustering visualisation in the notebook.

Find similar movies / tv shows in corresponding Cluster



❖ **Recommendation System**

A simple recommendation system was also made as an additional project on cosine similarity of description column and movie column, it also experimented in the notebook.

```
Chosen Movie/TV Show
Behind Enemy Lines: After dire setbacks in 1940, Winston Churchill commissions a new kind of fighting force: commandos trained to

Top Recommendations
Thunderbolt: A P-47 Thunderbolt squadron is shown in preparation, at play and in bombing raids aiming to halt Nazi supply lines ar

A Bridge Too Far: This wartime drama details a pivotal day in 1944 when an Allied task force tried to win World War II by seizing

The Outpost: A group of vastly outnumbered U.S. soldiers at a remote Afghanistan base must fend off a brutal offensive by Taliban

The Siege of Jadotville: Besieged by overwhelming enemy forces, Irish soldiers on a U.N. peacekeeping mission in Africa valiantly

Mission of Honor: As Hitler's Nazis threaten to take command of Britain's skies, a squadron of Polish pilots arrives to aid the Rc
```

Future Scopes

- More Post Cluster Analysis
- Integrate the Netflix dataset with other datasets and present more insights and clusters.

- We could have done some more research on the recommendation system. (Based on TF IDF, rather than cosine similarity)

Conclusion

After the data building and preprocessing we came up with 12 features and 7.7 k records. We clubbed textual features together and found 9 optimal clusters based on silhouette score and elbow graph and performed K-means clustering and named those clusters afters inferring the data we got in each one of them.