Netflix Movie and TV Shows

Rajan Srivastava
Data Science Trainee,
AlmaBetter, Bangalore

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 Flexible, 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, Exploratory 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 Flexible 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, rotten tomatoes can also provide many interesting findings.

Data Description

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 Releaseyear 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 : Genere
12. description: The Summary description

Objective

Netflix is a popular service that people across the world use for entertainment. In this EDA, I will explore the netflix-shows dataset through visualizations and graphs using Unsupervised learning algorithms and matplotlib and seaborn.

Exploratory Data Analysis

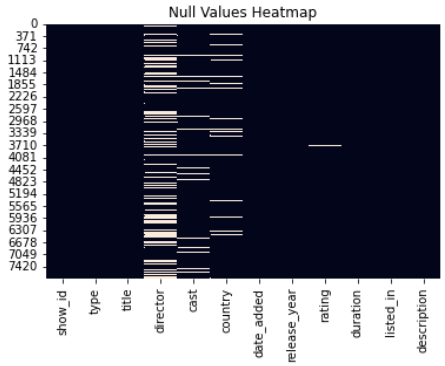
The first step involved in the analysis is to load the dataset into the pandas data frame. Before exploring the data using different libraries available in python we should if the dataset is ready to run the operations on it.

❖ 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 the rating column.

Handling Null Values

We can see that for each of the columns, there are a lot of different unique values for some of them. It makes sense that show_id is large since it is a unique key used to identify a movie/show. Title, director, cast, country, date_added, listed_in, and description contain many unique values as well.



```
[ ]  netflix.isnull().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
dtype: int64
```

Above in the heatmap and table, we can see that there are quite a few null values in the dataset. There are a total of 3,631 null values across the entire dataset with 2,389 missing points under 'director', 718 under 'cast', 507 under 'country', 10 under 'date_added', and 7 under 'rating'. We will have to handle all null data points before we can dive into EDA and modeling.

```
[ ]  netflix.isnull().any()
```

```
show_id         False
type            False
title           False
director        False
cast            False
country         False
date_added      False
release_year    False
rating          False
duration        False
listed_in       False
description     False
dtype: bool
```

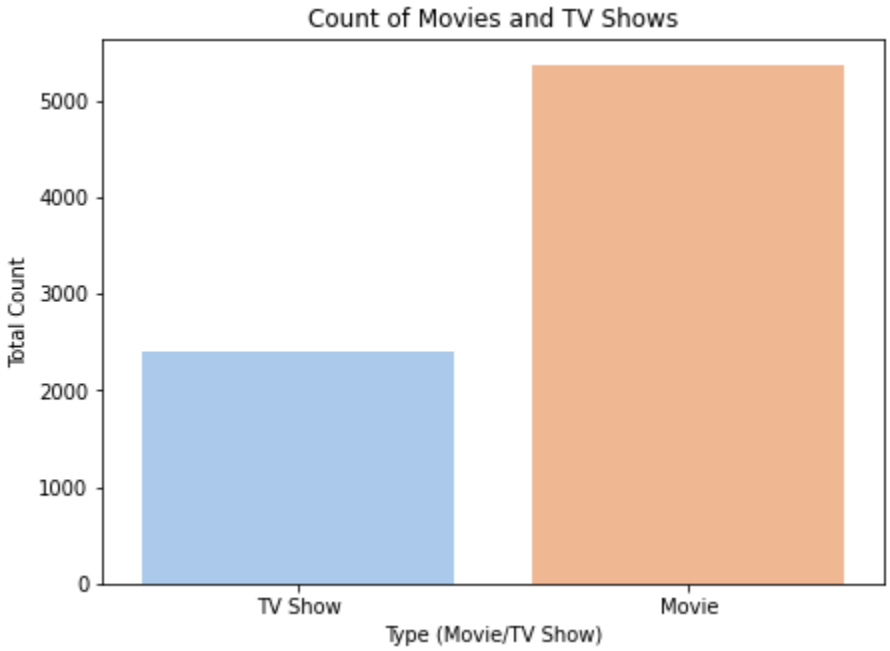
For null values, the easiest way to get rid of them would be to delete the rows with the missing data. However, this wouldn't be beneficial to our EDA since there is loss of information. Since 'director', 'cast', and 'country' contain the majority of null values, I will choose to treat each missing value as unavailable. The other two labels 'date_added' and 'rating' contain an insignificant portion of the data so I will drop them from the dataset. After, we can see that there are no more null values in the dataset.

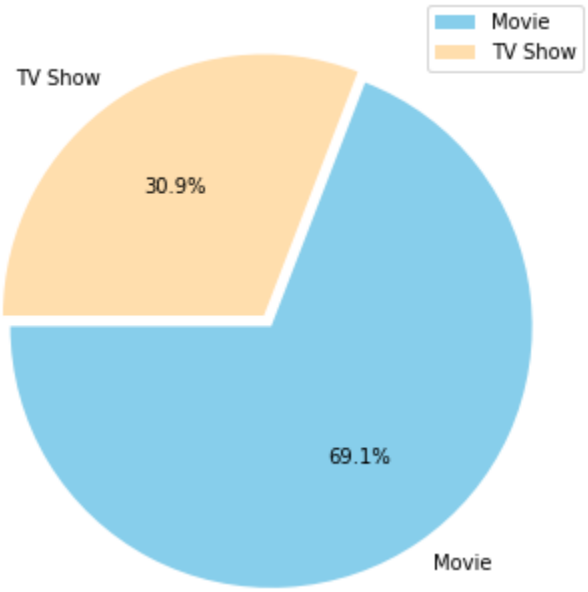
Data Preparation

In the duration column, there appears to be a discrepancy between movies and shows. Movies are based on the duration of the movie and shows are based on the number of seasons. To make EDA easier, I will convert the values in these columns into integers for both the movies and shows datasets.

Netflix Film Types: Movie or TV Show

It'd be interesting to see the comparison between the total number of movies and shows in this dataset just to get an idea of which one is the majority.

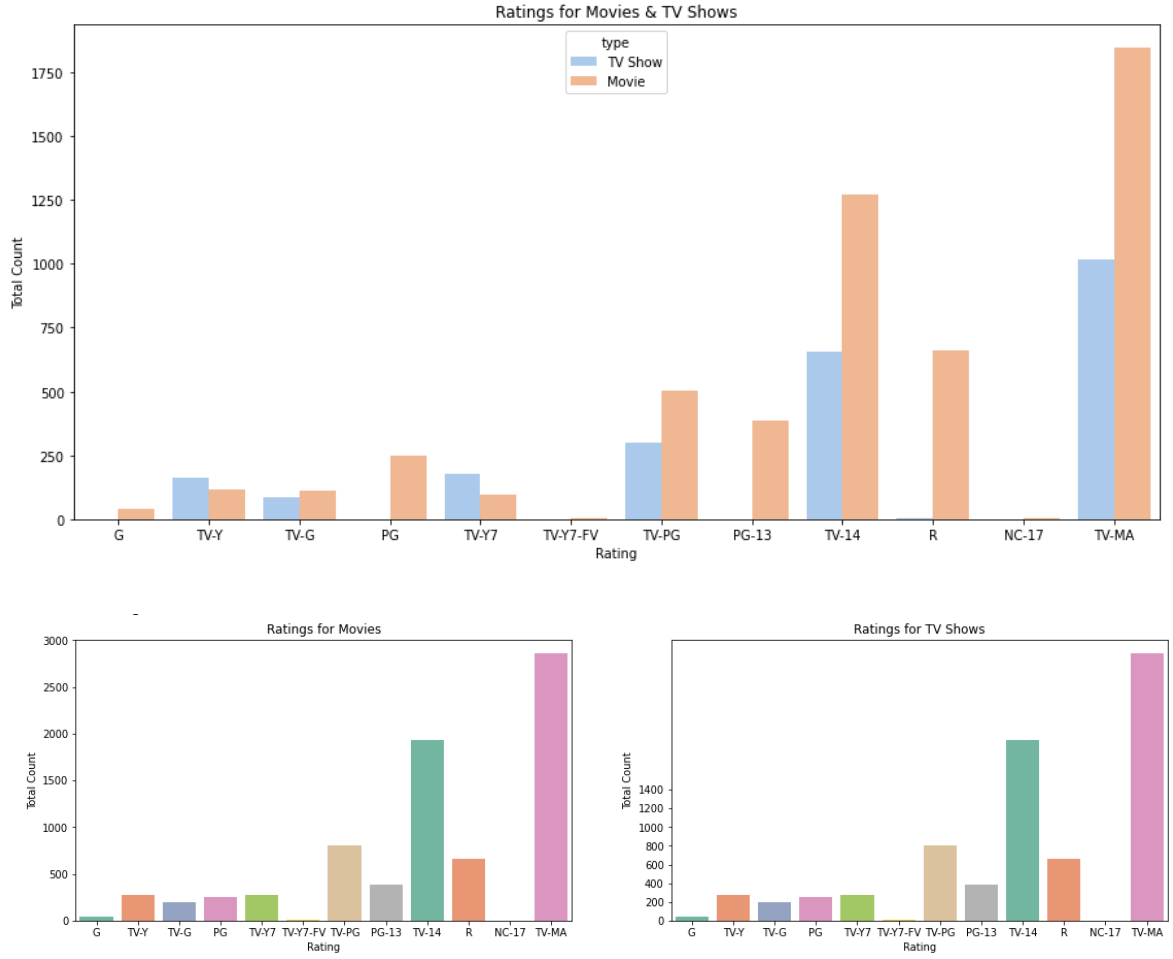




So there are roughly 5,000+ movies and almost 2,000+ shows with movies being the majority. This makes sense since shows are always an ongoing thing and have episodes. If we were to do a headcount of TV show episodes vs. movies, I am sure that TV shows would come out as the majority. However, in terms of title, there are far more movie titles (69.1%) than TV show titles (30.9%).

Netflix Film Ratings

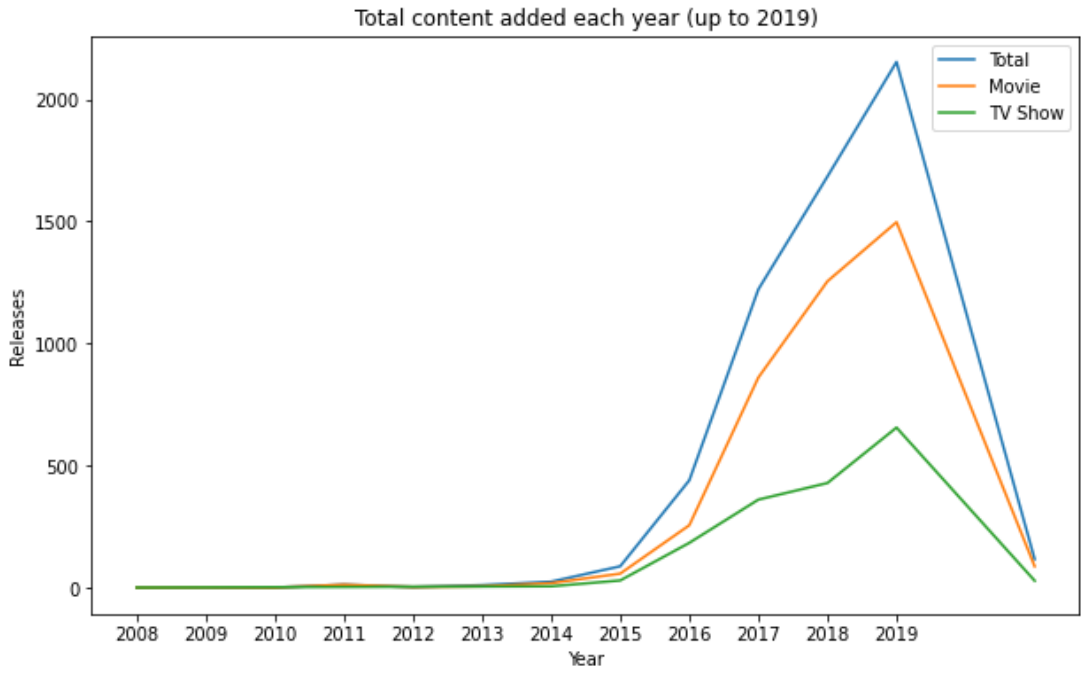
Now, we will explore the ratings which are based on the film rating system. The ordering of the ratings will be based on the age of the respective audience from youngest to oldest. We will not include the ratings 'NR' and 'UR' in the visuals since they stand for unrated and non-rated content.



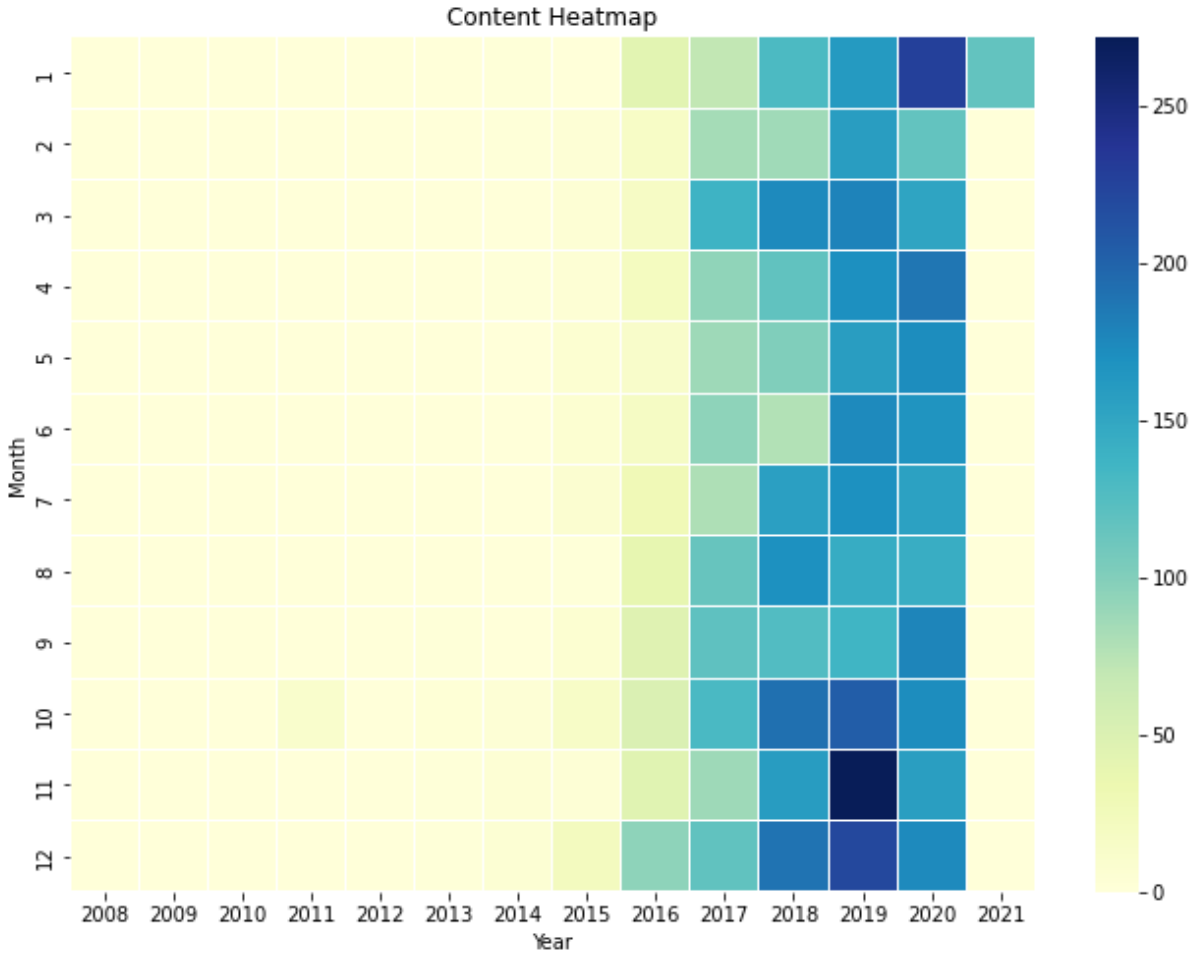
Overall, there is much more content for a more mature audience. For the mature audience, there is much more movie content than there are TV shows. However, for the younger audience (under the age of 17), it is the opposite, there are slightly more TV shows than there are movies.

Content added each year

Now we will take a look at the amount of content Netflix has added throughout the previous years. Since we are interested in when Netflix added the title onto their platform, we will add a 'year_added' column showing the year of the date from the 'date_added' column as shown above.

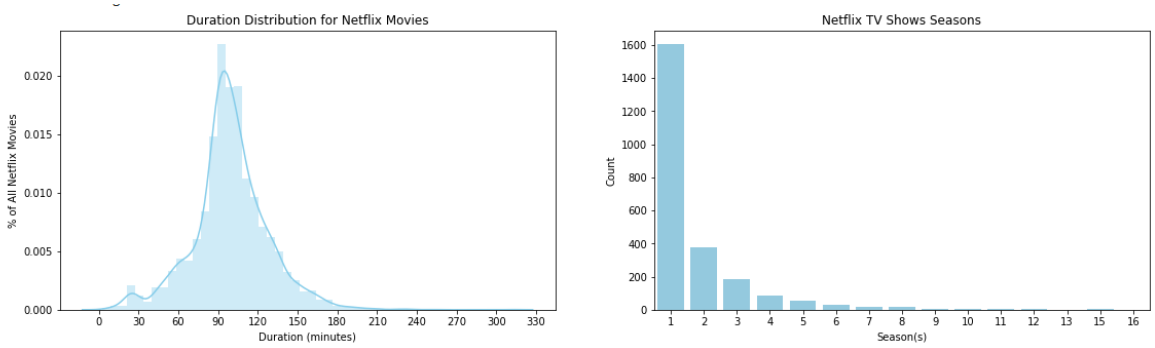


Based on the above timeline, we can see that the popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been tremendous. I decided to exclude content added during 2020 since the data does not include a full year's worth of data. We can see that there has been a consistent growth in the number of movies on Netflix compared to shows.



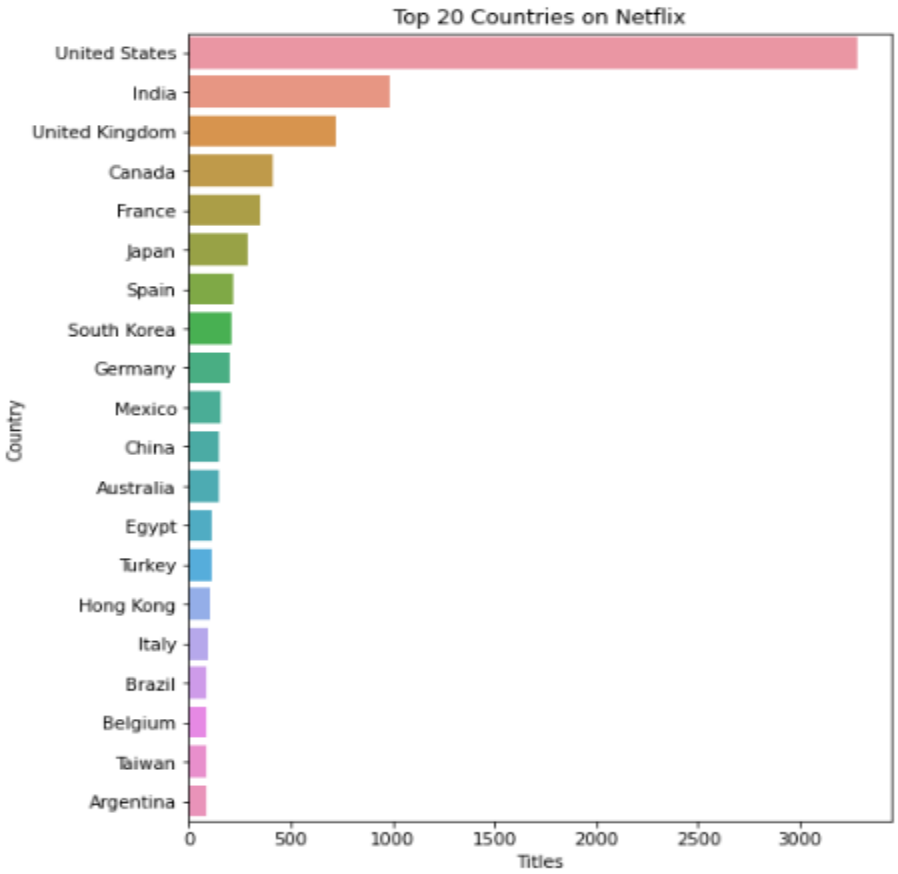
In the above heatmap, we can see that around 2014 is when Netflix began to increase their content count. We can see over the years and months, Netflix continues to slowly increase the amount of content that is being added into their platform. We can see in 2020, the data stops at January since that is the latest month available in the dataset.

Netflix Film Duration

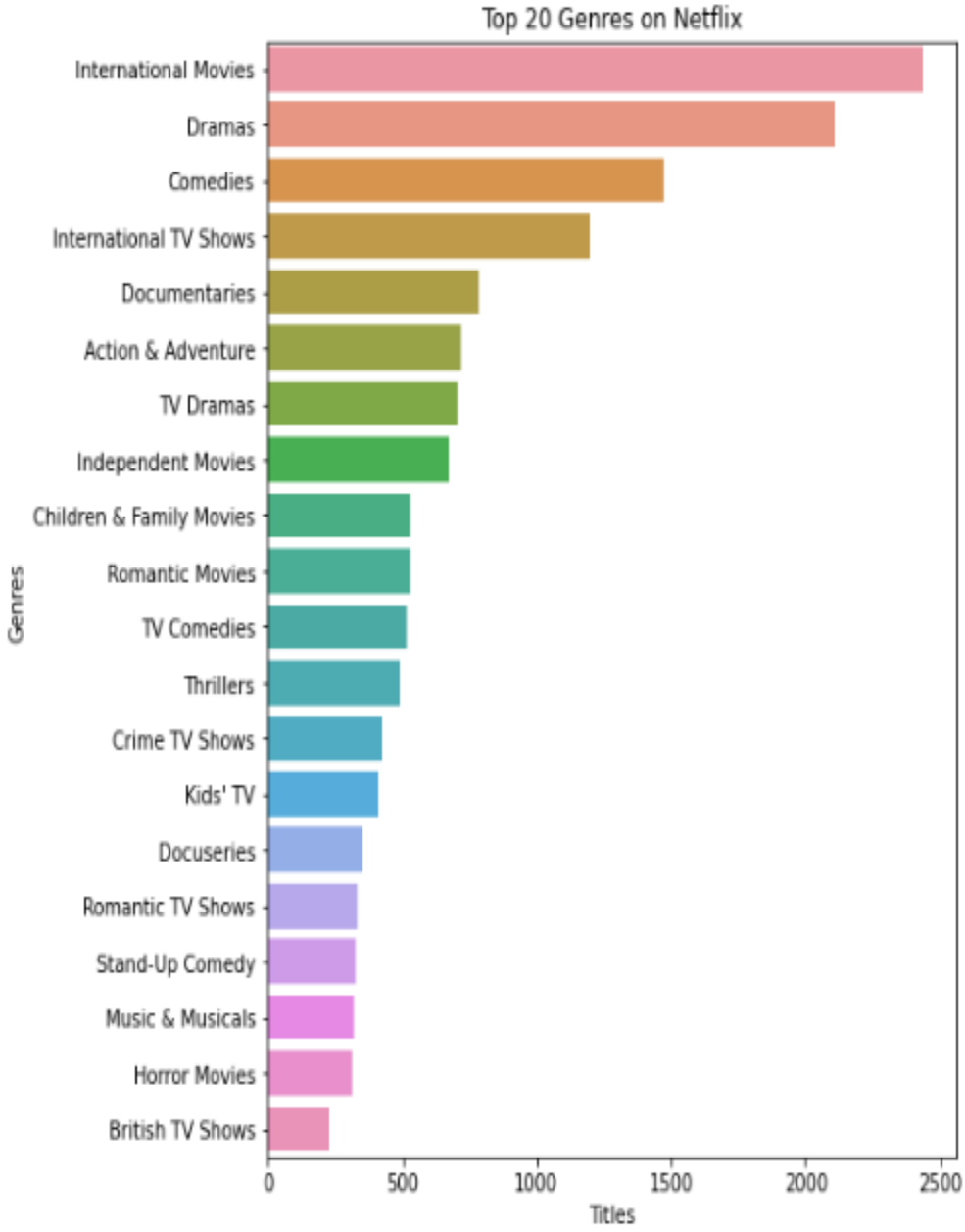


Now we will look into the duration of Netflix films. Since movies are measured in time and shows are measured by seasons, we need to split the dataset between movies and TV shows. Above on the left, we can see that the duration for Netflix movies closely resembles a normal distribution with the average viewing time spanning about 90 minutes which seems to make sense. Netflix TV shows on the other hand seems to be heavily skewed to the right where the majority of shows only have 1 season.

Countries with the most content available



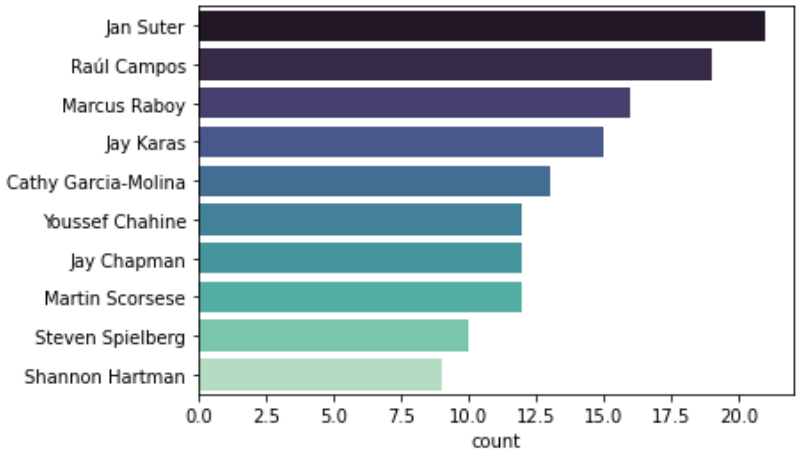
Now we will explore the countries with the most content on Netflix. Films typically are available in multiple countries as shown in the original dataset. Therefore, we need to seperate all countries within a film before we can analyze the data. After separating countries and removing titles with no countries available, we can plot a Top 20 list to see which countries have the highest availability of films on Netflix. Unsurprisingly, the United States stands out on top since Netflix is an American company. India surprisingly comes in second followed by the UK and Canada. China interestingly is not even close to the top even though it has about 18% of the world's population. Reasons for this could be for political reasons and the banning of certain applications which isn't uncommon between the United States and China.



In terms of genres, international movies take the cake surprisingly followed by dramas and comedies. Even though the United States has the most content available, it looks like Netflix has decided to release a ton of international movies. The reason for this could be that most Netflix subscribers aren't actually in the United States, but rather the majority of viewers are actually international subscribers.

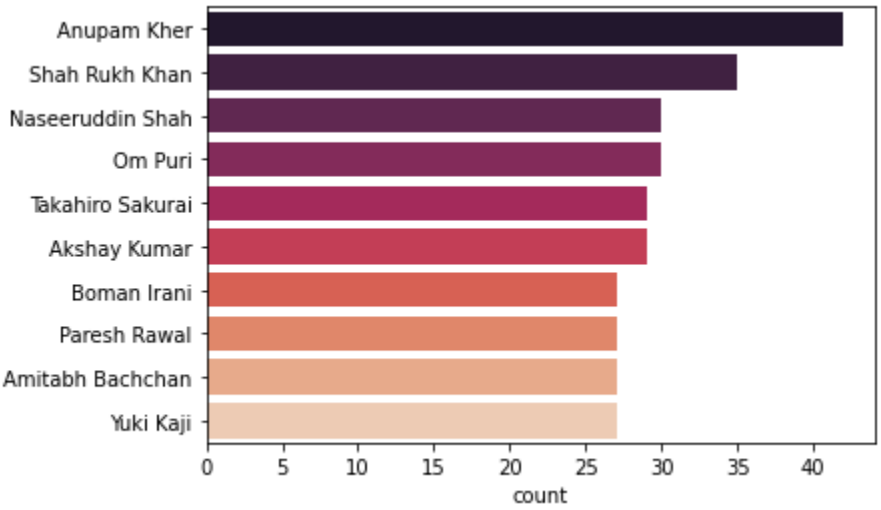
Asking and Answering Questions

Who are the top 10 directors on Netflix with the most releases?



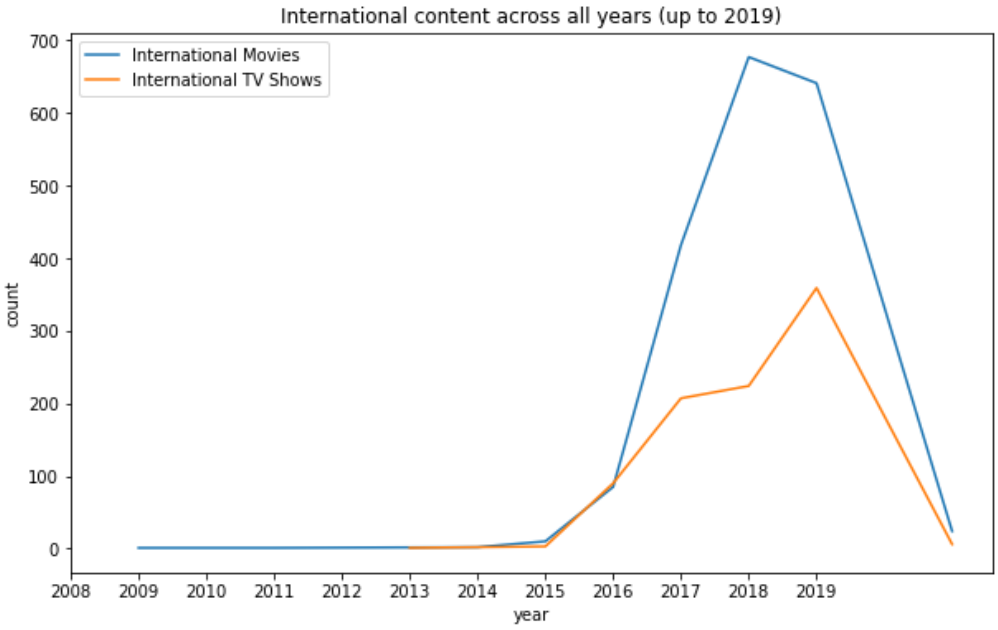
As stated previously regarding the top genres, it's no surprise that the most popular directors on Netflix with the most titles are mainly international as well.

Who are the top 10 actors on Netflix based on the number of titles?

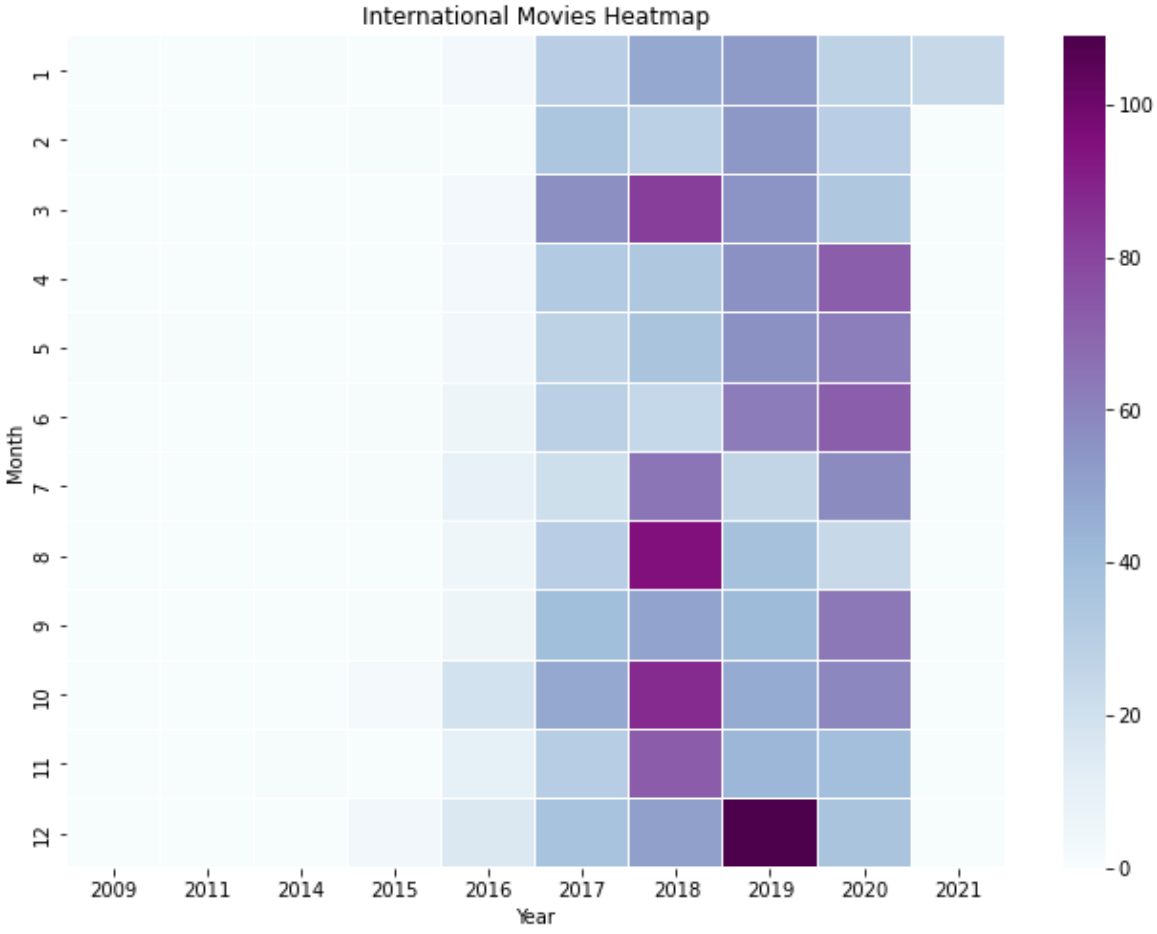


In this list, we can see that the most popular actors on Netflix based on the number of titles are all international as well. This reinforces the sentiment that the majority of Netflix subscribers are international.

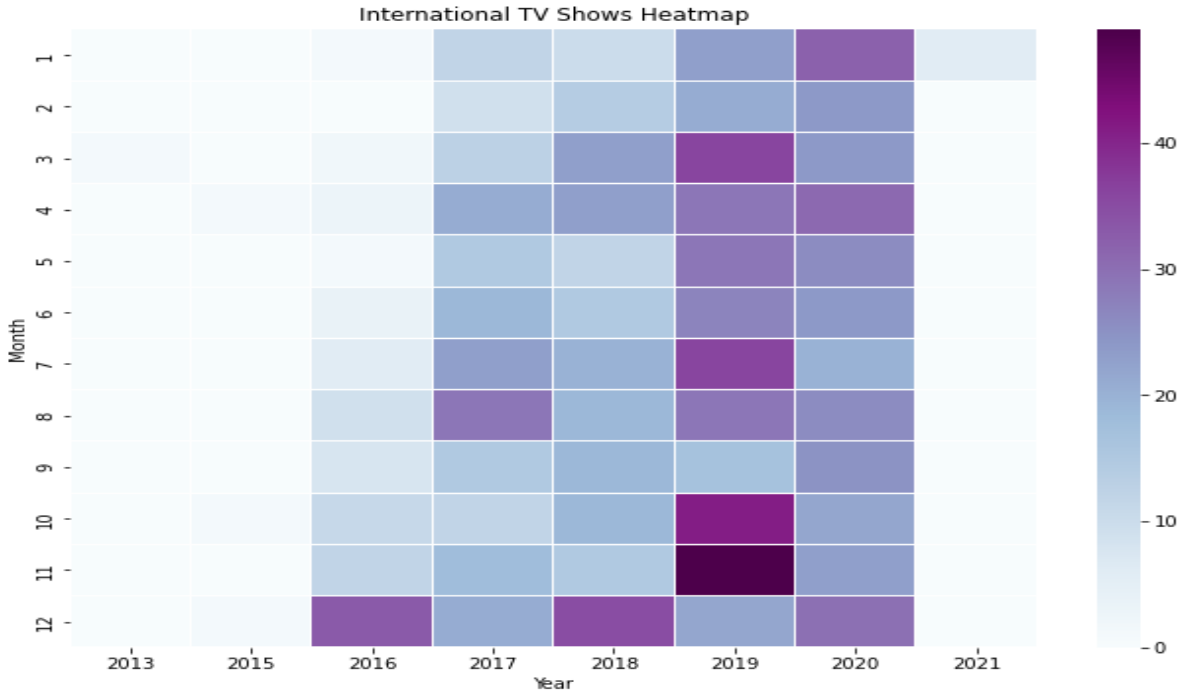
What does the timeline look like for the addition of International Movies compared to International TV Shows?



Based on the timeline, we can see that there are far more international movie releases than there are international tv show releases. However, near 2018, the growth of international movies started to decline while international tv shows constantly showed significant growth in the past few years.



In the heatmap above, we can see that a majority of international movies were added throughout the year in 2018. Then in December 2019, Netflix added the most international movie content.



In the above heatmap, we can see that the majority of international TV shows were added throughout the year 2019.

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 TFIDF, rather than cosine similarity)

Summary-

Got a dataset having 7787 records and 12 features, which is not labeled. Then I started by understanding the data followed by missing/null value treatment, then some feature engineering for data visualization and then data visualization performed. After getting insights from data I did some feature engineering like text cleaning, TF-IDF, PCA on data. Then after selecting no. of clusters, I built the KMeans clustering model. After that I assigned the clusters in our dataset and created word clouds for each cluster and after that calculated silhouette score for evaluation purposes.

Conclusions

It's clear that Netflix has grown over the years. We can see from the data that the company took certain approaches in their marketing strategy to break into new markets around the world. Based on an article from Business Insider, Netflix had about 158 million subscribers worldwide with 60 million from the US and almost 98 million internationally. Netflix's original subscriber base was based solely in the United States following its IPO. A large part of its success was due to the decision to expand to international markets. The popular market prioritizes what content the company will release. In this case, we can see that a good amount of international movies and TV shows were added over the years as part of Netflix's global expansion.

References

AlmaBetter
Kaggle
Github
Analytical Vidya
Other Sources