Topic : Netflix Movies and TV shows Clustering

Tripti Singh

Data Science Trainees

Alma Better

Abstract:

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

Introduction:

Netflix's recommendation system helps them increase their popularity among service providers as they help increase the number of items sold, offer a diverse selection of items, increase user satisfaction, as well as user loyalty to the company, and they are very helpful in getting a better understanding of what the user wants. Then it's easier to get the user to make better decisions from a wide variety of movie products. With over 139 million paid subscribers (total viewer pool -300 million) across 190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the world's leading Internet television network and the most-valued largest streaming service in the world. The amazing digital success story of Netflix is incomplete without the mention of its recommender systems that focus on personalization. There are several methods to create a list of recommendations according to your preferences. You can use (Collaborative-filtering) and (Content-based Filtering) for recommendation.

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

About the Data

The data I chose to go with for this demonstration was Kaggle's "Netflix Movies and TV Shows". This was based on a personal love for film and a burning desire to get to know more about them. This dataset consists of TV shows and shows available on Netflix as of 2019 and was collected from Flixable - a third-party search engine by Netflix.

The dataset is available Here

Inspiration

Maybe just one more episode ...

I intend to perform an Exploratory Data Analysis (EDA) on this data in a bid to get answers to some interesting questions/tasks. These include, but are not limited to -

1. Understanding what content is available in different countries
2. Identifying similar content by matching text-based features (This involves the creation of a Machine Learning Model to handle predictions, and as such will be covered in later iterations of this post.)
3. Whether Netflix has increasingly focused on TV rather than movies in recent years.

PS: Being a programmer and a data scientist to boot, I am intentionally going to use a non-programming approach to this task. This is to show that data driven decision making can be done without a programming background and to encourage individuals and businesses to embrace effective decision making. It is in this regard that throughout this task I will be using Google Sheets and Tableau in favor of Python.

Design

Data Description

The data-set consisted of 7787 Rows and 12 Columns. The columns, and their descriptions were as listed below:

1. **SHOW-ID** - Unique id of each show (not much of a use for us in this notebook)
2. **TYPE** - Show category. Could be either a Movie or a TV Show
3. **TITLE** - Name of the show
4. **DIRECTOR** - Name of the director(s) of the show
5. **CAST** - Names of Actors/ Actresses in the show
6. **COUNTRY** - Countries where the show is available to watch on Netflix
7. **DATE ADDED** - Date when the show was added on Netflix
8. **RATING** - Show rating on netflix
9. **RELEASE YEAR** - Release year of the show

10. **DURATION -** Time duration of the show
11. **LISTED IN -** Genre of the show
12. **DESCRIPTION - Brief insight into what the show is about**

Having a look at the columns, all save for the "Release Year" which was an integer, were of type "Object"

Data Cleaning

Data Cleaning process, involves identifying incorrect, incomplete, inaccurate, irrelevant, or missing pieces of data and then modifying, replacing, or deleting them as needed.

Missing Values

Using the formula *=COUNTIFS(original,"")* where "original" is the custom name for my entire data-set range, I found a total of 3631 blank entries across the data-set.

The following columns contained null values, "director," "cast," "country," "date_added," "rating." This had to be handled before I could proceed with further analysis. In order to deal with this, I had the option of either omitting/dropping the blank entries, filling them with values of the most commonly occurring column value or adding custom values for each empty cell. The easiest way to get rid of them would be to delete the rows with the missing data for missing values. However, this wouldn't be beneficial to me since it would result in loss of information. Hence I chose to treat each missing value as unavailable and input the value "**N/A**" for each blank cell.

I used *Find and Replace* with "*Search using regular expressions*" enabled to make this work on Google Sheets. I had to format the entire range as plain text (I had this changed back to Automatic in order to get the original format once I was done filling the blank cells). I used the regular expression **"^([\t]*)$"** to denote blank cells and replaced all matching values with "**N/A**"

Rerunning the *=COUNTIFS(original,"")* function outputs a zero meaning I had successfully tackled all the blank entries in the data-set.

The countries column contains some cells with more than 1 country value yet we need to prepare data on content distribution by country without having to group them. For us to effectively use this column, I needed to split this column using the *Split Text to Columns* formula. From this, I then had to use the *=UNIQUE()* formula to figure out unique values in the country column, then count all the instances of the unique country values using *=COUNTIF***.** The importance of this step will be evident when I will be querying the data for content distribution by country later on in the exploratory data analysis stage

Duplicate Values

The data-set contained no duplicate values.

New Columns

In order to get a clearer picture of the data, I further created the following columns

- *month_added* and *year_added* derived from the *date_added* column and
- For *month_added* , I first had to regularize the date format to conform to the "11-Feb-1997" format, then apply a *MID* function *(=MID(G2,4,3))* to derive the month (Feb) from the string. Or as I later came to learn, an easier faster way would have just been to use the *=MONTH("20/07/1969")* function to return the month (🤦)
- For *year_added*, I simply applied the *RIGHT* function. Or again as I later came to learn, just using the *=YEAR("20/07/1969")* function would return the exact year (🤦)
- *rating_age* derived from the *ratings* column to help decipher the codes. To do this, I grouped the ages as follows with their corresponding codes

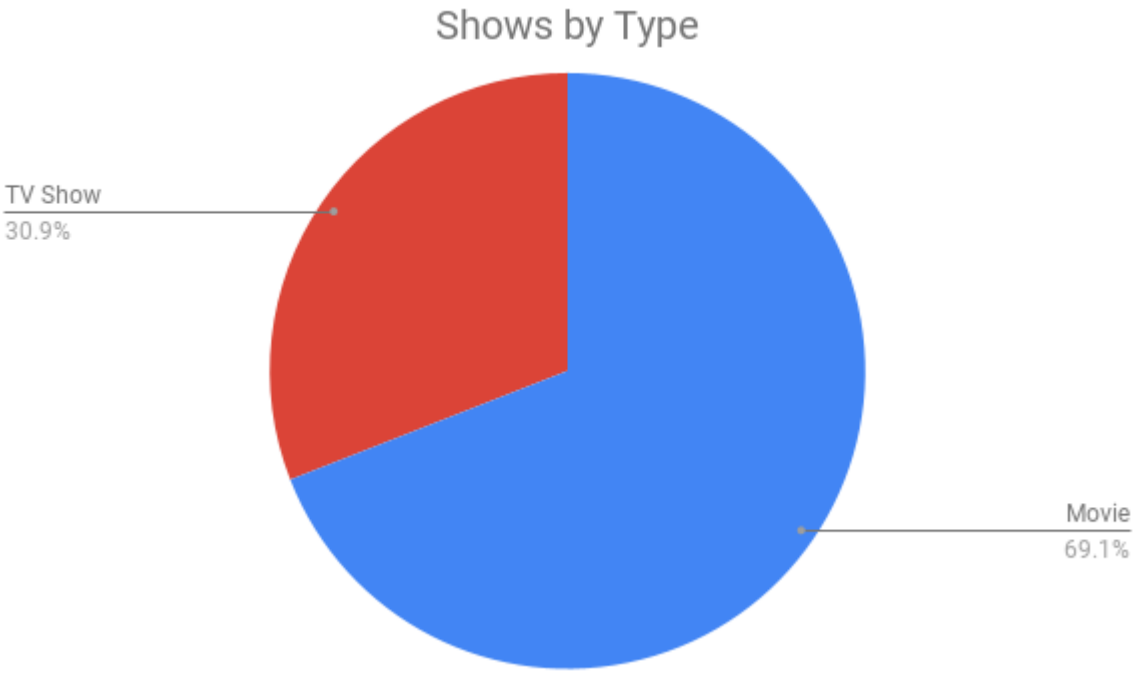
1. 'TV-PG': 'Older Kids',
2. 'TV-MA': 'Adults',
3. 'TV-Y7-FV': 'Older Kids',
4. 'TV-Y7': 'Older Kids',
5. 'TV-14': 'Teens',
6. 'R': 'Adults',
7. 'TV-Y': 'Kids',
8. 'NR': 'Adults',
9. 'PG-13': 'Teens',
10. 'TV-G': 'Kids',
11. 'PG': 'Older Kids',
12. 'G': 'Kids',
13. 'UR': 'Adults',
14. 'NC-17': 'Adults'

The below *SWITCH* function was then used to implement the assignment above

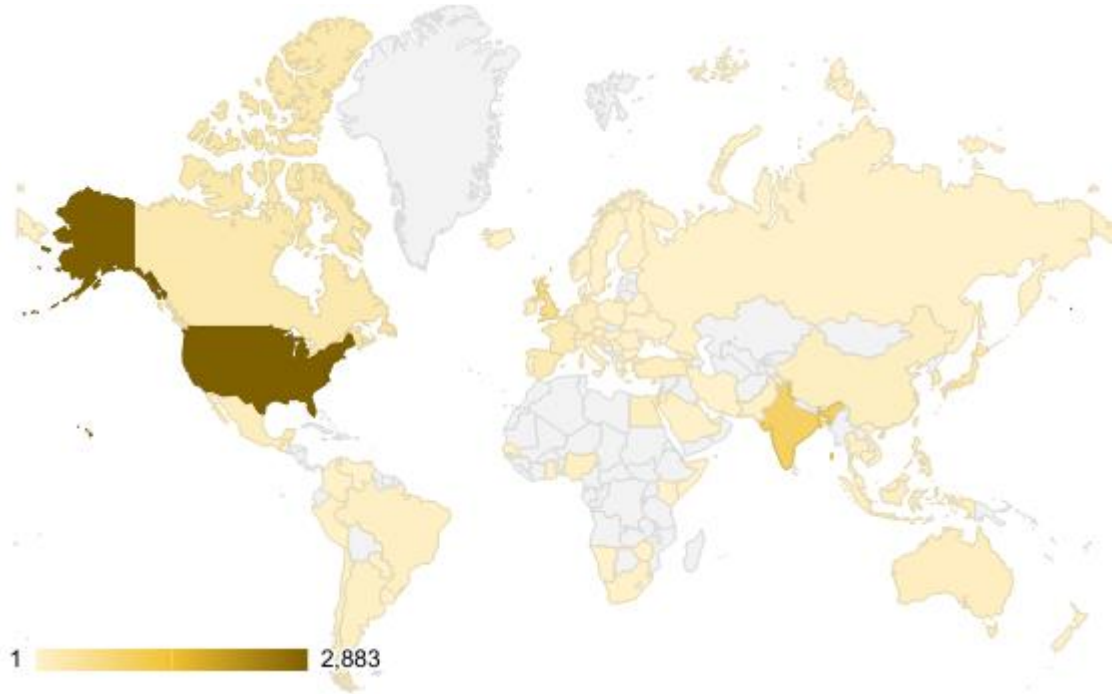
=SWITCH(K2,"TV-MA","Adults","R","Adults","NR","Adults","UR","Adults","NC-17","Adults","TV-14","Teens","PG-13","Teens","TV-Y7-FV","Older Kids","TV-Y7","Older Kids","PG","Older Kids","TV-PG","Older Kids","TV-Y","Kids","TV-G","Kids","G","Kids",)

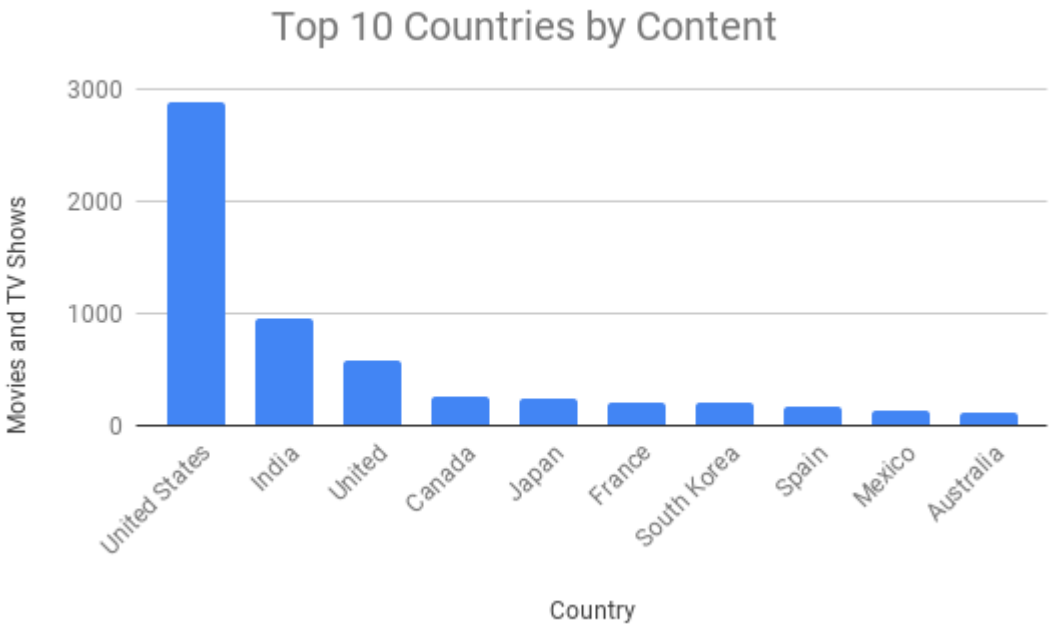
Exploratory Data Analysis and Visualization

Content distribution between Movies and TV Shows

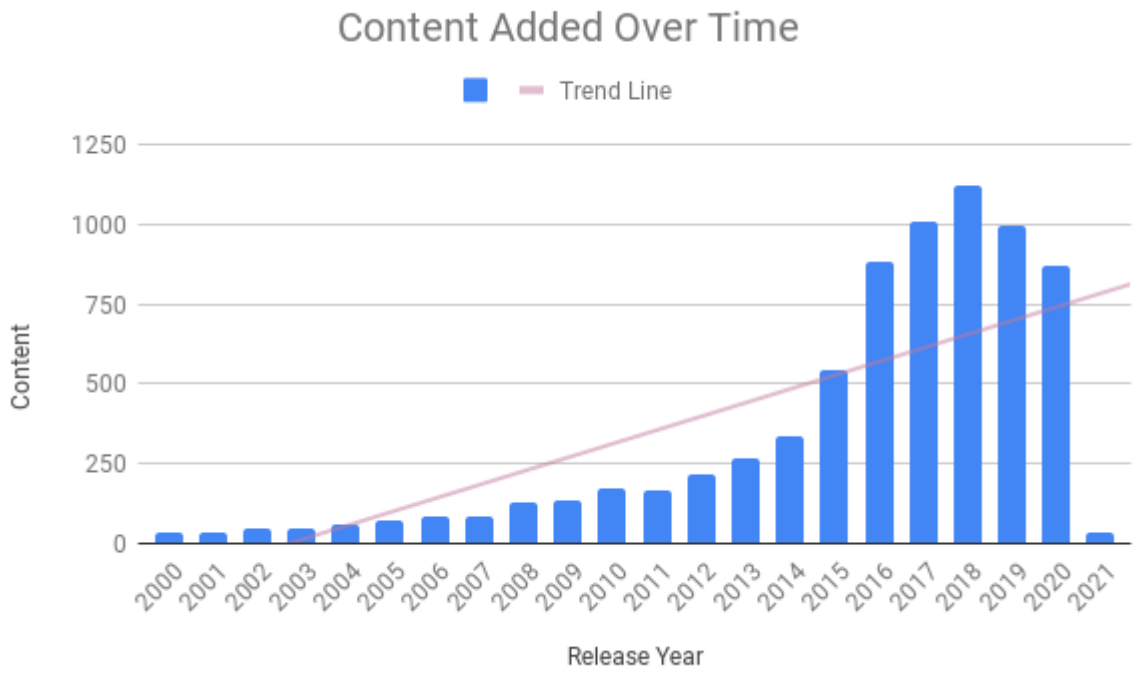


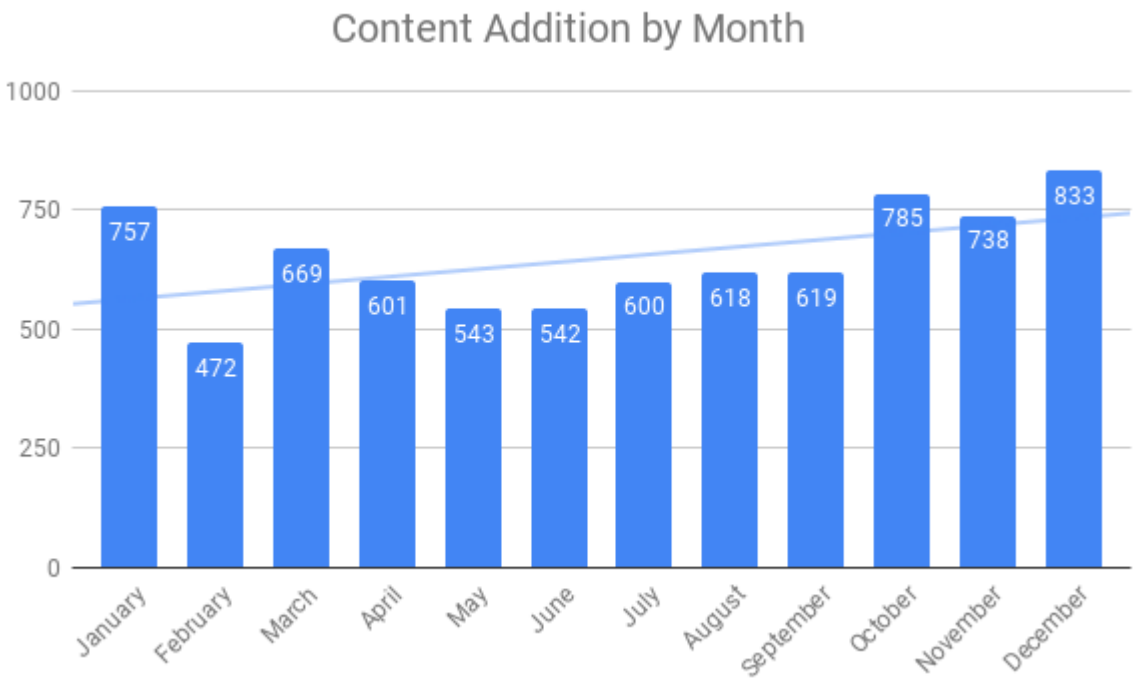
Content distribution per country where the films were allowed to air



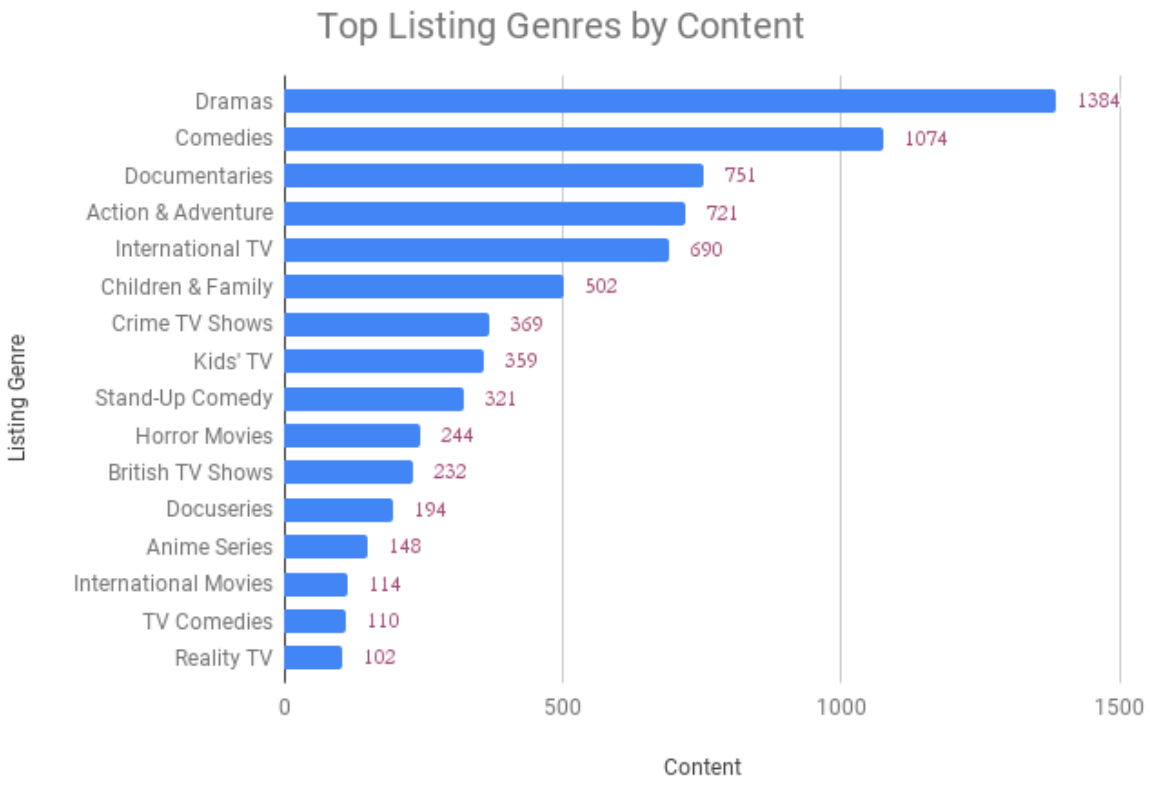


Content as a Function of Time

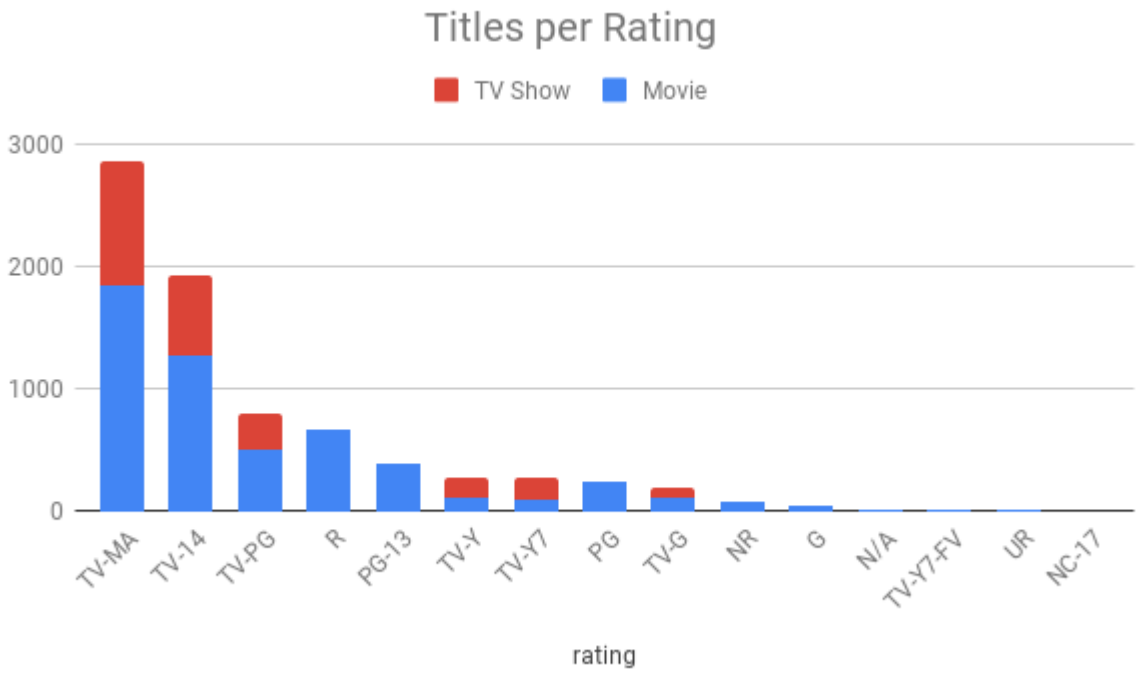


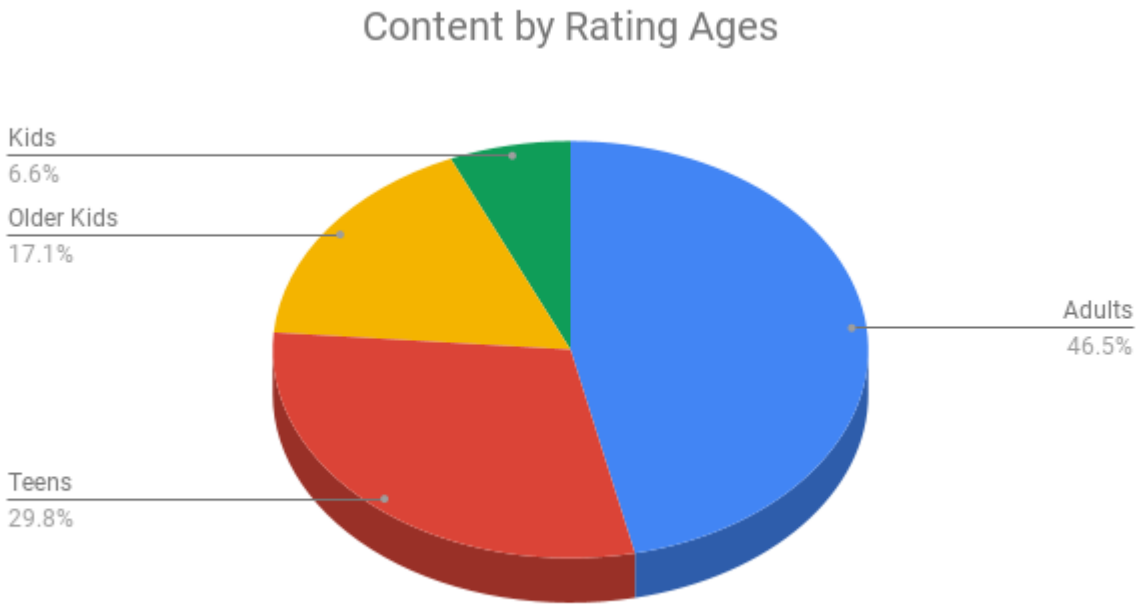


Top listing Genres



Count of titles for each rating





Word Cloud on Title

Conclusion

From this dataset, some interesting patterns have come out. Below is a summary of some of them:

1. The most popular content type on Netflix is movies,
2. Netflix started gaining ground in 2014 and has been significantly increasing its content portfolio ever since.
3. The country by the amount of the produced content is the United States,
4. The largest count of Netflix content is made with a “TV-MA” rating,
5. The Word Cloud on titles shows the most popular words used when describing shows, with top among them being Love, Story and Christmas further lending credence to the analysis that showed December to be the most popular release month as Christmas falls in December and is generally a month of love. And after all, ain't all movies and series just stories?
6. There was a steady rise in content production over the years till about 2019 when the content dipped. This could be attributed to the COVID-19 pandemic that hit the entertainment industry particularly hard during that period.
7. Over the years, there seems to be a steady increase on content production for Netflix from June through December spilling over to January before it starts to decline again.

For future iterations, however, I would like to make the following improvements;

1. I would like to go ahead and recreate this analysis using python to make it more of a predictor system. This would be done using ML concepts of regression.
2. In addition, integrating this dataset with other external datasets e.g IMDB ratings or rotten tomatoes could also provide interesting findings e.g from inputting a movie title, the system should be able to predict whether it will be a hit or a miss based on the title, the actors, producers, target genre, core themes etc and give a weighted value on why it thinks that might be so

That has been it from me. Be Safe, Be Kind. PEACE.