Exploratory Analysis of Netflix movies and TV shows

Karthikeyan Chellamuthu

DSC530-T302 – Week 12.2 Term Project

Professor: Dr. Neerja Sahu **, Phd**

03/05/2022

Introduction

Especially in COVID pandemic situation Netflix needs no introduction, it is a streaming service that allows members to watch a wide variety of contents like movies, TV shows, documentaries and more. Netflix allows us to enjoy unlimited ad-free viewing of the contents. It was founded in 1997 and has become $176 billion company in 25 years.

Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, as of mid-2021, they have over 200M Subscribers globally. The tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration.

Due to pandemic many people sitting at home and started watching more tv and gaming stuffs. Additional important variables will be analyzed to perform exploratory data analysis and also predict the number of Movies and TV shows that will be added in 2022.

Also the following question has been addressed.

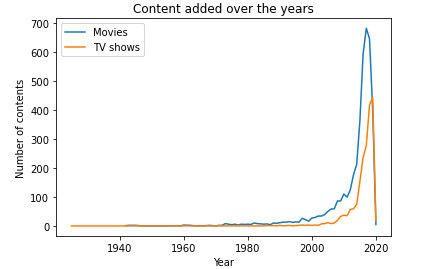
* Outcome of your EDA?
* What do you feel was missed during the analysis?
* Were there any variables you felt could have helped in the analysis?
* Were there any assumptions made you felt were incorrect?
* What challenges did you face, what did you not fully understand?

Statistical or Hypothetical questions

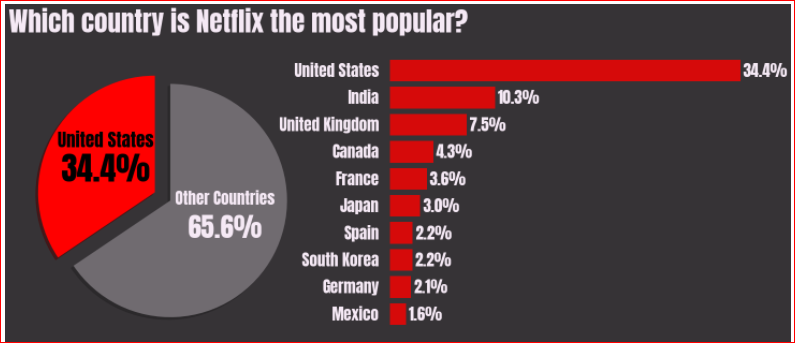
* Analyze the contents available in different countries and see which countries produce more movies and TV shows.
* In which month are the contents added more?
* What is the duration of the contents?
* What are the top categories or genre?
* What are the types of contents and their rating?
* How many contents will be added in 2022?

Outcome of EDA

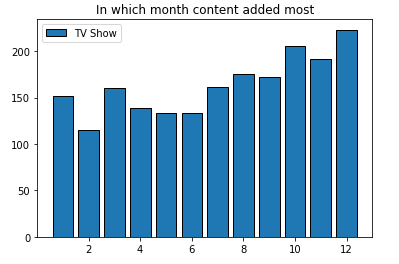
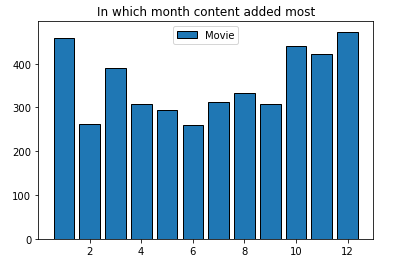
* Below shows content added over the years. Note that year 2022 shows a drop due to few data available in dataset for this year.



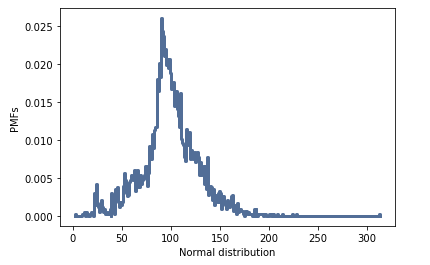
* Contents present in United States are more than any other country.



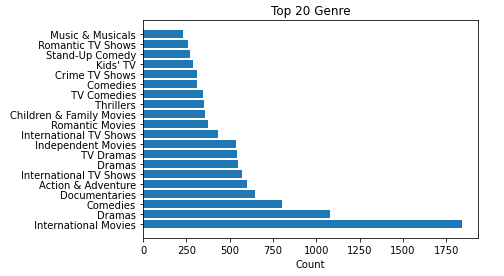
* More number of contents are added in second half of the year.



* Average duration for Movies is 100 minutes.



* The top genre is International Movies. It could one of the reasons why Netflix is very popular in many countries. Users get chance to watch contents from various countries at single place.

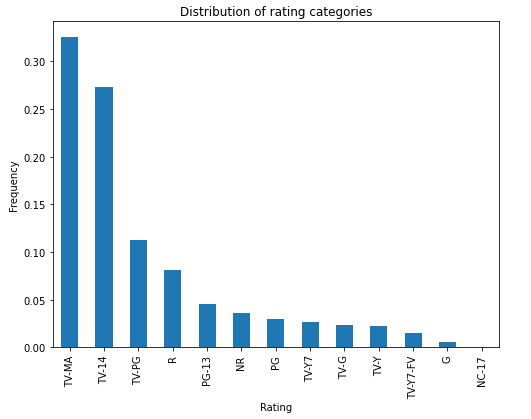


* Below are what top categories mean:

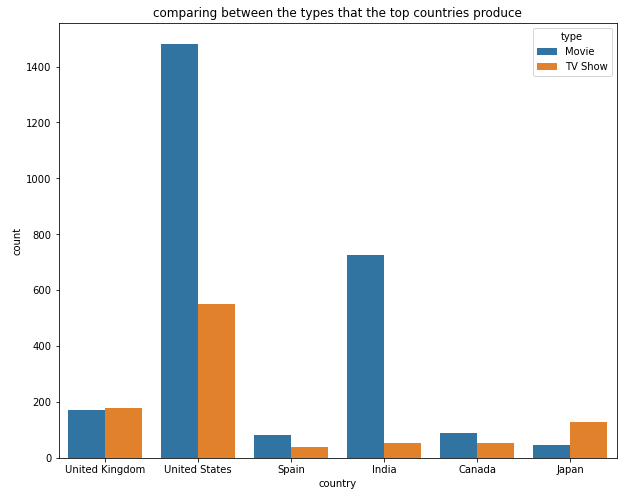
TV-MA: This program is specifically designed to be viewed by adults and therefore may be unsuitable for children under 17.

TV-14: This program contains some material that many parents would find unsuitable for children under 14 years of age.

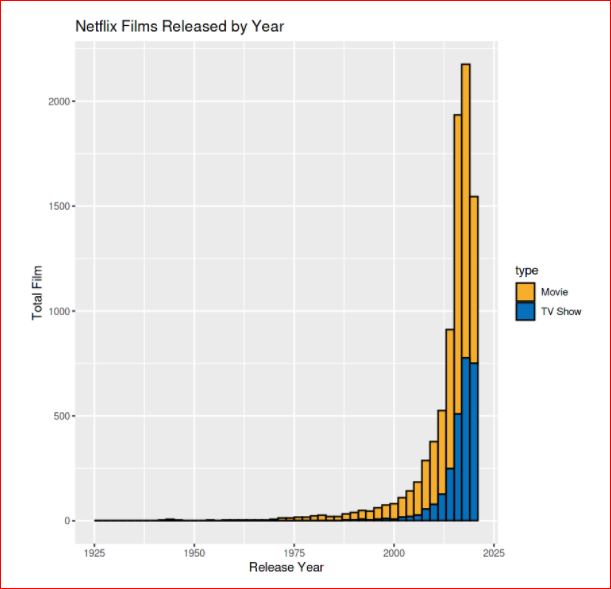
TV-PG: This program contains material that parents may find unsuitable for younger children.



* It is interesting to note that UK and Japan have more TV shows than Movies.



* Based on the prediction, 2,014 Movies and 933 TV Shows will be added in 2022.



Based on the prediction – 1023 Movies and 958 TV Shows will be added in 2022

* **Were there any variables you felt could have helped in the analysis?**

Type – It is the type of content Movies/TV Shows

Country – It is the place where contents are present

Date\_added – Date when the contents were added to Netflix

Release\_year – Year when the Movies/TV shows were released

Rating – It represents the content type. For example TV-MA is for adults only and may not be suitable for children under 17; TV-G is suitable for all ages.

Listed\_in – It is the genre. Example: Comedies, Kid’s TV

* **What do you feel was missed during the analysis?**

Another variable that I suppose to be missing in IMDB ratings value. It would be good to analyze the IMDB ratings of Netflix contents. I found it little challenging to plot hypothesis test with this data. Did some additional research about the variables and came up with CDF plot

* **What challenges did you face, what did you not fully understand?**

Especially in this problem I did face lot of challenges to come up with an idea of what I am going to do with given data set, there are some online samples given there but not accurate due to covid pandemic season the data sets published some of them were not accurate. I felt this may lead to lot of outliers and I was not able to fully understand those situations and it was assumed.

References:

<https://www.kaggle.com/kharismabima/netflix-data-exploration>

Downey, A. B. (2015). *Think Stats:* Beijing: OReilly.