Exploratory Data Analysis: Netflix Service

Jennie Yang

Jonathan Tshimpaka

Dinh Bao Tran Nguyen

SP22: CS-122 Sec 02 - Adv Python Programming

Instructor: Wendy Lee Ph.D.

May 08, 2022

Link to Project:

https://colab.research.google.com/drive/1IzhWW96cqapJFrIQaEAxuFEUEeV_qFpC?usp=sharing

a) Introduction - states the overall goal of the project and the approach you took to accomplish the goal

Netflix is one of the top leading streaming services that has risen fastest in the past few years. The American giant has acquired more than 200 million subscribers worldwide by the end of 2020 with a market cap of 243 billion dollars. We now want to look at and answer the question of why Netflix became the number one choice for people before the COVID pandemic. Our goal is to analyze a dataset of Netflix, which will show the increase in popularity and answer why Netflix has become the number one streaming service in the 21st century. To accomplish the goal, we decided to use a dataset from Kaggle. We take a different approach by looking at the connection between the increase in subscribers and the types of movie content that Netflix provides. The approach is unprecedented, but it will shed light on the connection if there is any. We are going to consider some aspects related to movie genres, movie rating, movie casting, and anything else that is relevant. Not only that, but we also consider the aspect of demographics if it contributes to the increase in the number of subscribers in relation to the success of Netflix.

What motivates us to analyze the success of Netflix is the popularity that has been gaining over the past few years. We understand that there is a challenge using the dataset we choose to work with to explain the increase in subscribers because it is about the number of contents that Netflix rolled out. Looking at it carefully, we have noticed a correlation between the number of content rolling out and the increase in subscribers. Perhaps, there is a correlation that many people overlook. Thus, we take this opportunity to dig deeper into this dataset as it is

shown in Kaggle. As in the 2020s, Netflix has now faced competition from other streaming platforms such as HBO, DisneyPlus, and Amazon Prime. But, it still has its advantages because of the diversity of movie genres and services in more than 190 countries around the world. Netflix provides TV series, documentaries, and feature films across a wide variety of genres and languages. With over 200 million paid memberships, Users can easily watch as much as they want, anytime, anywhere, and on any platform. With the memberships, members can play, pause, resume watching, and download any content on the go. With that being said, we would love to analyze all the titles currently appearing on Netflix screens, especially with more competition from streaming entertainment services such as Hulu and HBOMax. With the dataset we use, we will analyze the various contents in this streaming service, so we can conclude which content Netflix should add in a way so that more subscribers are added.

b) Methods - explain all the components in your project and how they work together. Include flow diagrams and pseudocode in your explanation. Do not include any actual python code in the report, only use pseudo-code.

Step to Perform Exploratory Data Analysis:

1. Import libraries and load dataset

   Dataset use: https://www.kaggle.com/datasets/shivamb/netflix-shows

   The original file URL was not opened in the Pandas DataFrame. So, we downloaded the dataset to the Google Cloud and mounted google drive to the Colab.

   Python library and Framework use: Numpy, Pandas, Plotly, Seaborn, Matplotlib

   - Numpy and Pandas help process the dataset and read the file

- Use Matplotlib to visualize the dataset for clearer indicators
- Make statistical graphics With the Seaborn and Plotly library

2. Clean Dataset
    a. Check for missing values :
        i. The data set does not contain Null Values. However, we found mistyped values when we created the plot. Some time duration values were typed on the rating columns.
    b. Replace missing values
        i. Because we did not know the rating of the data, we set the values as NaN.
    c. Handle mistyped values
        i. However, we knew the time duration, so we fixed the values.
3. Asking Analytical Questions and Visualizations
    a. Analyze the positive and negative correlation between variables
    b. Determine the effect of one variable on other variables
    c. Use libraries and Python to configure the data frame if needed
    d. Draw the visualization
4. Conclusion
    - From the visualization, we can perform a prediction and draw a good conclusion from the dataset.

c) Results - you can include screenshots along with descriptions

1. The distribution of content rising from 2008 to 2021 started to appear around the world

This graph is the comparison by country of programs added every year.

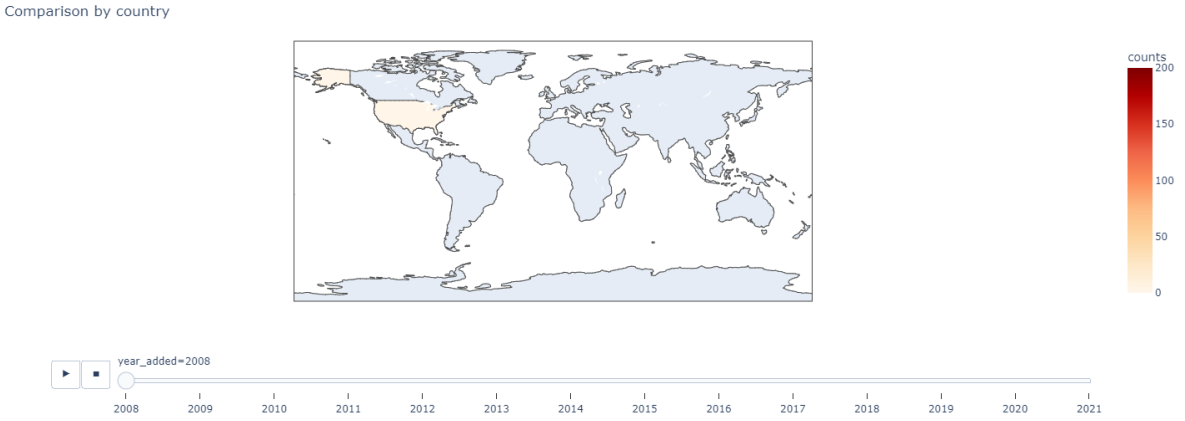

figure 1.

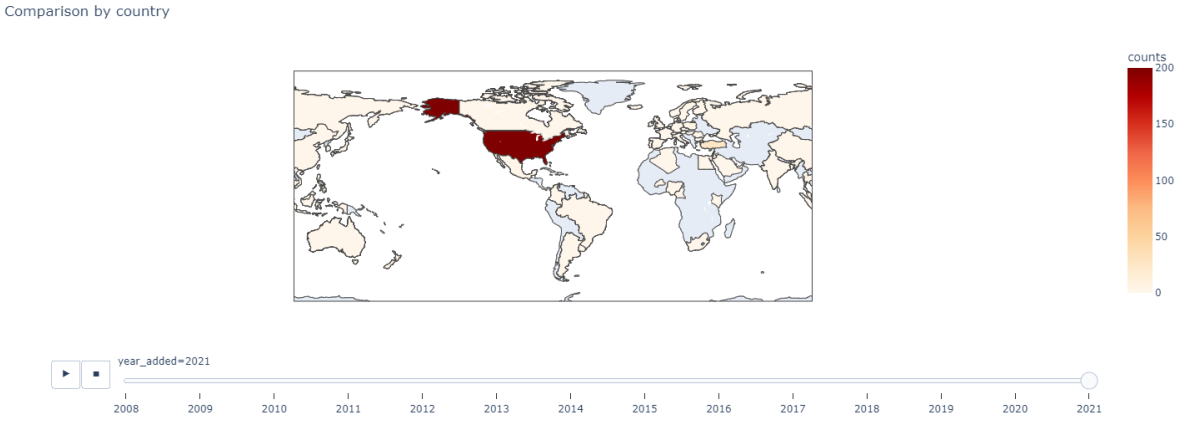

figure 2.

Starting in 2008, Netflix started a partnership with consumer electronics brands to allow streaming on Xbox 360, Blu-ray players, and TV set-top boxes. The rising content from 2008 to 2021 indicated the prosperity of this giant streaming service. From figures 1 and 2, we can see the increase of users that spans over ten years. It shows that people around the world are getting to know about Netflix and starting to use the service. In 2017-2018, we started to notice an increase in the number of movies added as well as the number of users. We suspect such a correlation because of the diversity in the movies that attracted more users to sign up with Netflix. Before the COVID pandemic, Netflix experienced an increase in the number of users. During the pandemic, the number increased greatly compared to the previous years. We believe

that the lockdown helped make this increase because people had to stay home for months. So they chose to use streaming services to entertain themselves. Thus, a huge leap in the number of subscribers was expected in the year 2020. For example, we can see that the United–States was the one that had many new signups on Netflix, compared to other countries because of the lockdowns.

2. Analysis by program type

Netflix program consists of 69.6% of movies and 30.4% of TV shows worldwide. In the United States, the Netflix program consists of 73% of Movies and 27% of TV Shows.

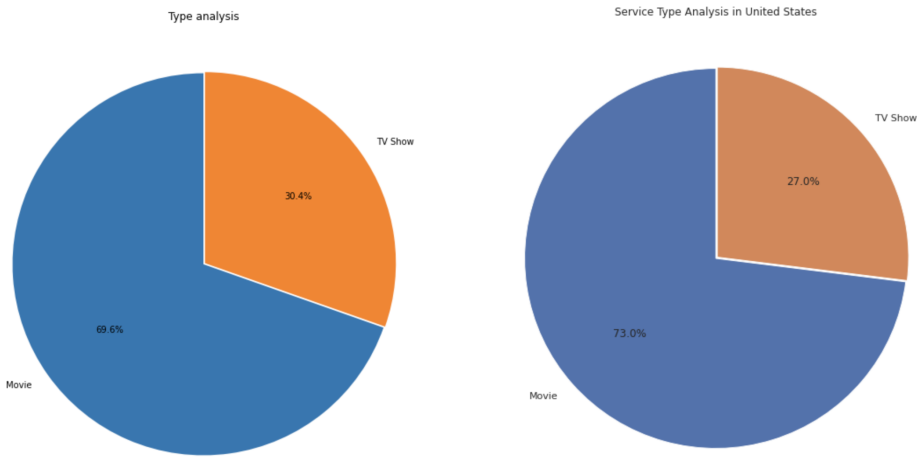

figure 2. program type

3. Frequency of Listed-in Type of Netflix Content

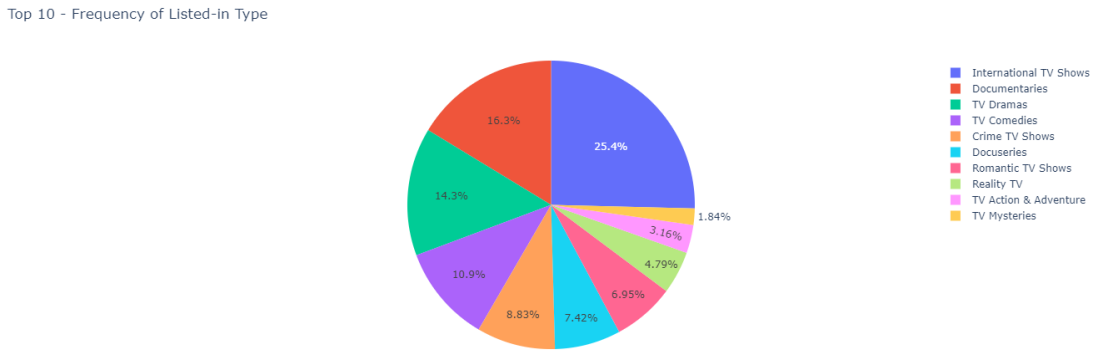

figure 3. Content-Type

From the graph, we observe the diversity of content types. The highest percentage in Netflix content type is the International TV shows which are 25.4%. This number means that Netflix has numerous global audiences that use Netflix for its streaming platform. It shows that the platform attracts worldwide users, not only American people. From figure 3, we can see that Netflix focuses on diversifying movie genres to attract more users as there are many other movie genres such as documentaries, tv dramas, etc… Netflix does offer such diversification in the movie genre to compete with other streaming platforms. In the 2020s, HBO and Disney take back a lot of movies that were created by them so that they can run those movies on their platforms. Netflix has sought out other alternatives to make sure that they do not lose subscribers so diversification is one of the methods that they have been using.

4. Service top 15 countries

The United States has the most programs among the Netflix service countries. Netflix provides 2818 programs in the United States.

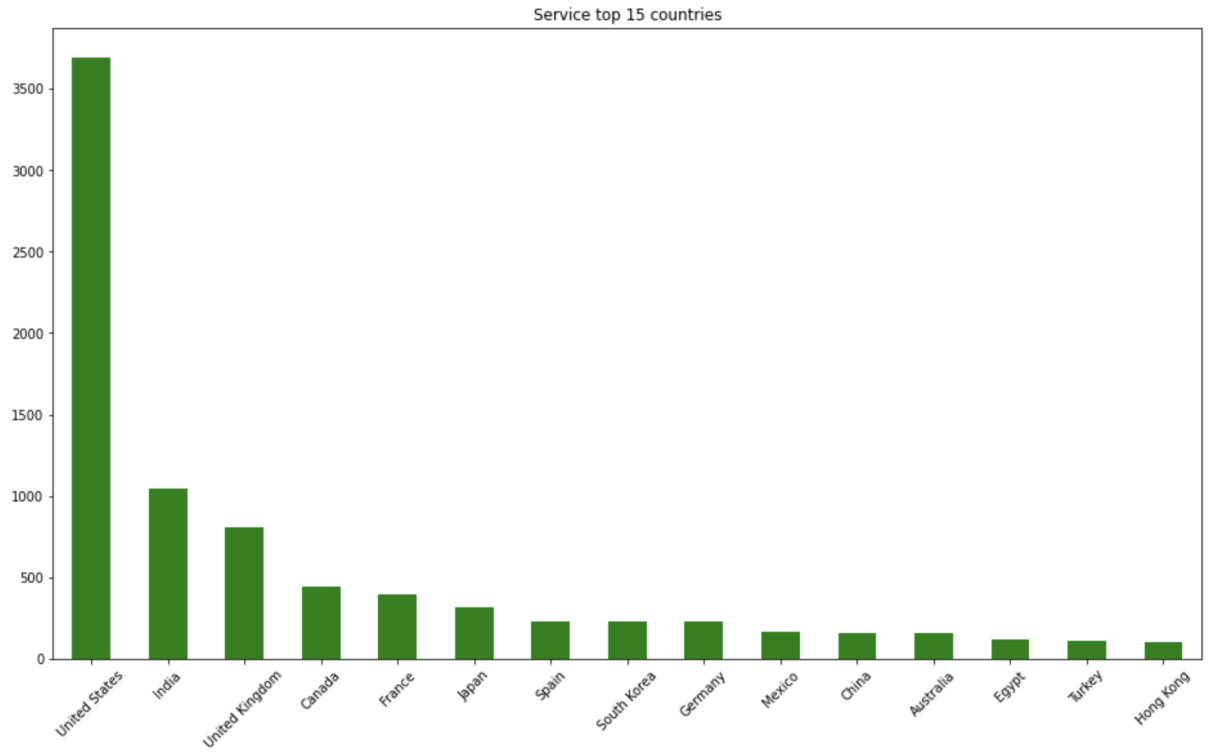

figure 4. Service top 15 countries

5. Year trends for adding Movies and TV shows on the Netflix

In 2017 & 2018, Netflix tends to add the most amount of movies. The number of movies released by year decreased since COVID-19. In contrast to Movies Trends, Netflix tends to add the most amount of TV shows by 2020. But, the number of TV shows released in the next year decreased as well because of COVID-19.

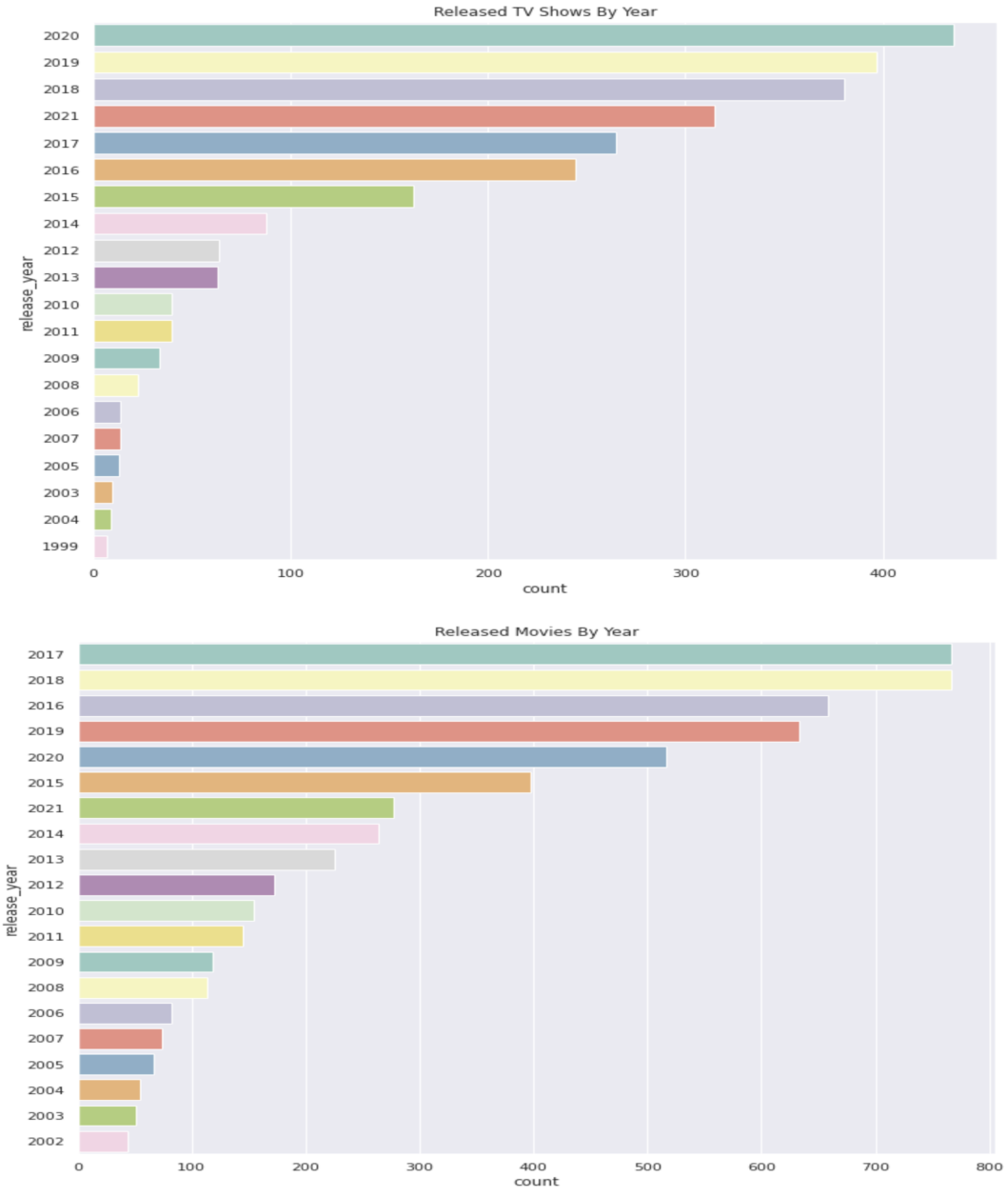

figure 5. Year trends for adding Movies and TV shows on the Netflix

6. Month trends for adding programs on the Netflix

In July, Netflix tends to add the most programs while adding the least in February.

In July, Netflix added 827 programs. In February, Netflix added 563 programs.

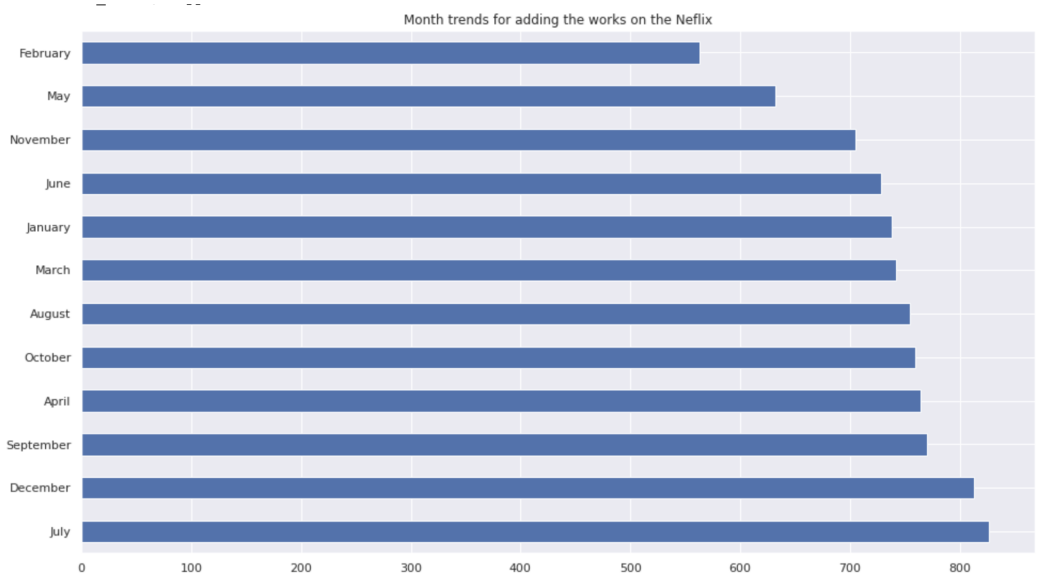

figure 6. Month trends for adding programs on the Netflix

7. Analysis by Rating

TV-MA(TV Mature Audience Only) is the highest, and UR is the lowest. The trends are also similar in the United States.

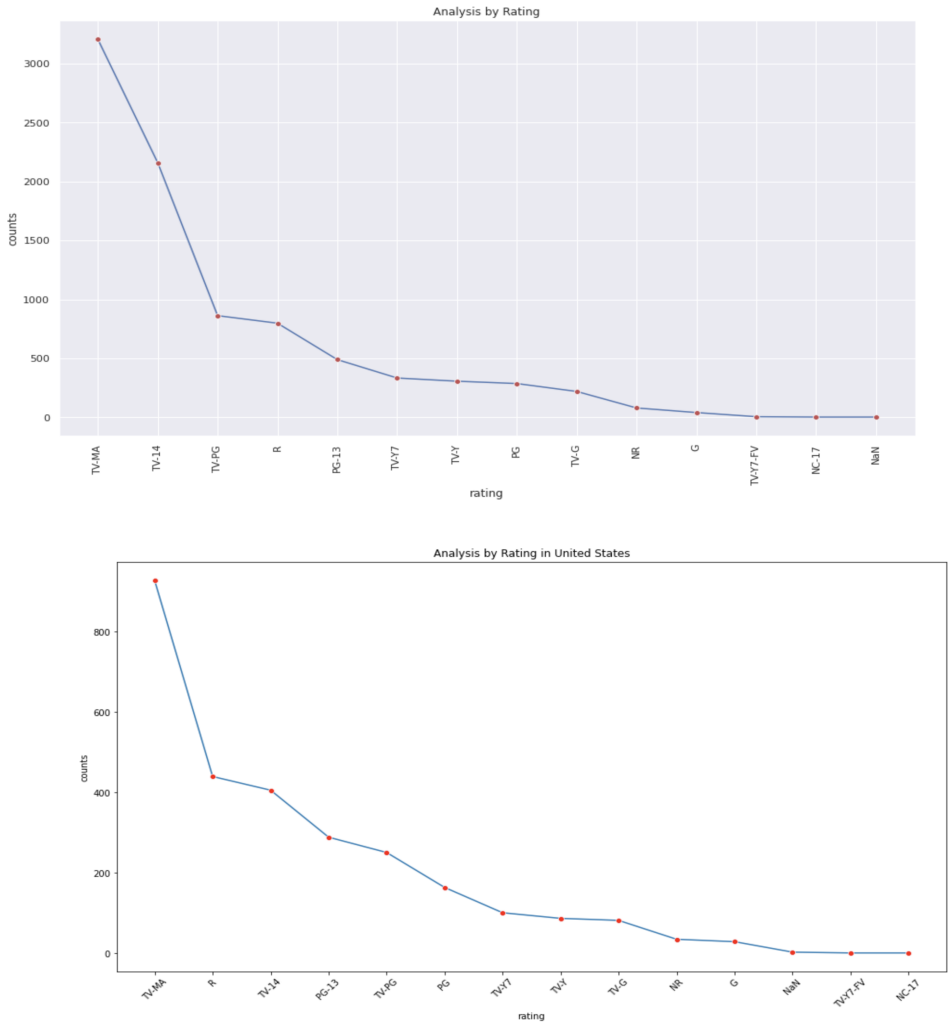

figure 7. Analysis by Rating

8. Netflix content classification is based on the audience

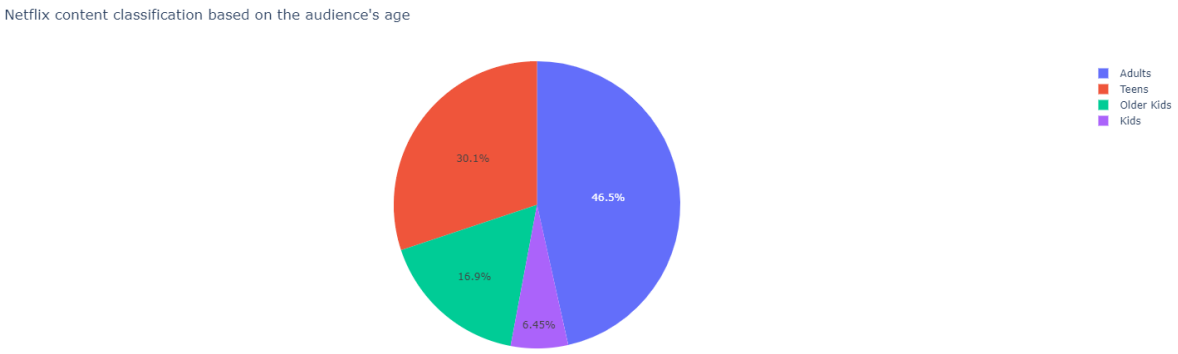

figure 8. Content-Type based on the age of users

We noticed that the majority of content on Netflix targets mainly adult members. With 46.9% of content that is suitable material for the adult group, we can know the group target that Netflix focuses on. We believe that adults watch more movies than children as expected because adults are more likely to seek out entertainment than children. So, Netflix adds more content that is suitable for adults so that they can keep adults watching and using their streaming service to counter the loss of users every month as people tend to cut down the cost of living or switch to other streaming services. So, having more content that is suitable for adults is one of their methods to keep up with the number of subscribers.

d) Discussion - you can talk about the technical challenges you faced in completing this project and what are the areas that can be improved on.

Starting this project, our team is not completely comfortable with Python and Python libraries.

The first aspect we have some difficulties with is how to implement the data visualization part with the first main goal. To visualize the dataset, we have to handle the dataset so that we can implement it into the libraries such as Matplotlib, Plotly, and Seaborn. We all agree that we could add some complex data visualization that can help us have a clearer understanding of this dataset. However, some of the planned graphs are needed to handle by using Python data structures, and we faced some errors while trying to visualize the data.

Not only that, as the submission of our preliminary report, we were going to use sklearn libraries to build the system content recommendation. However, we have some trouble implementing this dataset to sklearn.

e) detailed instructions on how to run the code for the project. Include any additional tools/modules installation steps needed before running your program.

This project mainly focuses on Exploration Data Analysis, we use Google Colab to help us achieve this. Since Google Colab provides all the installed libraries, we don’t need to install any distribution or packages outside. We can simply open the ipynb file on Google Colab and run it line by line.