Data Science Report Draft

Austin Araneo

# Introduction

**## Why this dataset?**

The purpose of a dataset like this is to examine in greater detail who the target audience is for Netflix. From the data the company could make decisions on future content to add to their library. The company can also look at where in the world the majority of their revenue is coming from as well as be able to tell on what devices their content is most popular to be streamed from. Being able to know this information can be critical when making decisions on future improvements to the services and where to add them.

## Research questions

1. Does users age contribute to different Netflix subscription tiers?
2. Do different countries prefer different ways of enjoying the entertainment.?
3. What demographic of people have the most Netflix subscriptions( male in their 40s? female in their 20s?

#Selection of Data

**##Source of Data Set:**

The data was first pulled from analytics on Netflix then posted to Kaggle on July of 2023 by the user ARNAV.

**##Munging, imputation, or feature engineering?**

As downloaded the dataset was complete. There was no outliers that did not make sense nor was there modifications needed to be able to make the data useable. The usability is well laid out with easy-to-read rows and columns that makes filtering and creating methods and visualizations easy.

**## What is in the Data Set?**

There is a total of 2500 users collected. The dataset consists of demograplics of each user including their age, gender and country of origin. More information is also available on the subscription tier of each account and the monthly price being paid. Another thing being tracked is the most popular device the account is being used on.

**## How is it filtered?**

More data that has been filtered out is the plan duration length. As Netflix does not have any annual or other options in duration length besides monthly this column was left out. The last part of filtering done to the data was the join and last payment dates. After going through the data, it is not relevant to the questions that are going to be answered.

## What tools will be used?

* Numpy, pandas,
* Coderunner as IDE
* Github to host the final deployment
* Graphs to show trends
* Excel to help with data cleaning.

## Results

1. Does users age contribute to different Netflix subscription tiers?

To be able to answer this question I started by making a pie chart of the number of users that has what subscription tier (Fig 1). This information was important so that later the more detailed graph can be determined. Next the data needed to be searched and modified. The age column was duplicated and a search function was executed to delete all duplicates in the coloumn to leave with only one copy of each age. After that there was a counter set and the program searched for all instances of (“plantype” + “age”). This went through and counted all the different ages and all the different plan types combinations. (fig 2) Shows the results displayed in a bar graph. The top three ages were 39,31, 29 and all of them were on the basic plan. In fact the top 8 ages were all on the basic plan and the least popular age was 38 and 36. Taking the top ages who were subscribed to the basic plan were 39,31, 29. The top Premium age subscriptions were. 41,28,30. And the top standard subscribers were. 42,41,35. This shows there really does not appear to be any correlation from Users age to what subscriptions they have.

fig 1

This shows the users age and what plan they have.

Fig 2

1. Do different countries prefer different ways of enjoying the entertainment.?

Each user is tracked with the way they stream Netflix. The most popular method is tracked and recorded in (fig3). Every users account also has a country they are streaming in this is the other part of the data pulled for this graph. There needed to be another modification to the data to concatenate some terms. First the column of countries was taken an put into another one. All duplicates were removed to show only one instance of each country. The same is then done with the devices list. A concatenation is then done to do every instance of both the devices and countries. After all of this a search unction is initiated to count every instance that matched the concatenated terns in the original data set. Fig 3 was the result after graphed into a bar graph. In order to simplify things even further (Fig 4) shows the other one simplified to only the top devices used in each country. Showing this It is clear while streaming on a tv is the least popular overall from all countries 4 countries prefer laptops, 3 prefer smartphones and 2 prefer tablets. While Laptops, tablets and phones are different devices it is clear that most Netflix users no matter the country prefer to use portable devicse rather than using a TV.

A graph of blue bars with white text

Description automatically generated

Fig 3

After looking at all of the results a second graph of only the most popular devices used in each country was created.

A graph of blue rectangular bars with white text

Description automatically generated

Fig 4

1. What demographic of people have the most Netflix subscriptions( male in their 40s? female in their 20s?

Demographics can be one pf the best ways for a company to tell the most about their users. In the data set the first thing done was I took a method to count total males and total number of females in the sheet. (Fig 5) shows that there was an even amount of both in the data. Here I modified a previous example from the subscription type and subsatuted it for gender. The method counted the number of users that matched each age and if they were male or female. The notable results were there was less than average 34 year old makes with Netflix and there were more than average 42 and 47 year old females with Netflix accounts. The most of any age was 59 users of any age. I believe this could be due to the scale of the data because that was an odd result as well that came up. With this being a mostly balanced dataset whether that is done on purpose or not it makes there is no obvious demographic male or female or age.

This shows the split of males and females is the same in the dataset for context to other graphs.

Fig 5

Fig 6

Summary

There were some shocking results in this data set of Netflix Users. It appeared to have more Users from the US and Spain than some of the other countries. This could be a contributing factor to some of the results. Netflix users did tend to prefer the Basic plan as opposed to the Premium or the Standard. It would be interesting if more data was available to investigate why people have Netflix over other streaming services. One reason could end up being because Netflix is watched mostly on tablets, phones, and laptops those do not require the higher quality bitrate the other plans offer. Another statistic that is not included bit worth noting is there is no data how many people are using each account. If most accounts are only one person then they would not need the feature to stream on multiple devices at once.

This Project gave me the ability to take a dataset and take some deeper dives into what data scientists do with it. By manipulating the data even further there is the possibility to get even more understanding of the users in this study. The ways that the questions ae laid out ae built o be able to expended even if there was a dataset much larger.

References

<https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset/data>