Motivation

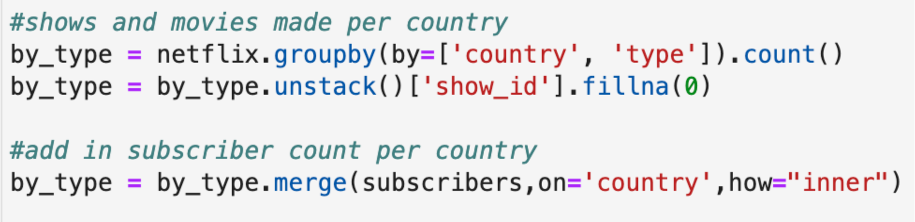
For my project, I wanted to take a closer look at Netflix data. I was particularly interested in the national origins of the content featured on Netflix as well as the subscriber count from each country. I was specifically trying to see if there were any differences in countries that produce media vs the countries that consume media. One of the countries I was most interested in looking at subscriber count vs the amount of media produced was India due to them having both Bollywood studios and a high poverty rate, meaning fewer people with subscriptions despite high media production.

Data Sources

1. "Netflix Movies and TV Shows," a CSV file downloaded from Kaggle. It contained records of all the media on Netflix through mid-2021, including the type of media (TV show vs movie) and the country of origin, two variables in which I was most interested. Other variables included things like title, duration, rating, etc.
2. I also used the "Netflix subscriber figures and revenue in 50 countries" from Comparitech which I read in as a CSV file. The table contained information on the revenue and subscriber count for each fiscal quarter of 2021 grouped by country. I decided to focus on the first quarter of the year for this project, focusing on the countries and subscriber counts.

Data Manipulation Methods

I manipulated the data by first cutting the data frame down only to the columns that were relevant to my calculations. I immediately renamed some of the long column names and made them all lowercase. I then joined the subscriber count data for each country onto the main data frame I was using with the data about tv shows and movies. Luckily, there was no missing or incomplete data to deal with. I decided to use the subscriber count of the first quarter of the year and discarded the other quarters' data due to it lining up best with the last update to the main Netflix media dataset.



```
#shows and movies made per country
by_type = netflix.groupby(by=['country', 'type']).count()
by_type = by_type.unstack()['show_id'].fillna(0)

#add in subscriber count per country
by_type = by_type.merge(subscribers,on='country',how="inner")
```

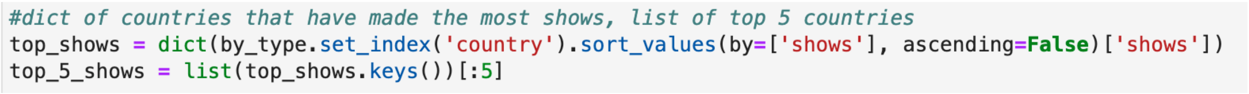
After cutting down the data frame, I created a copy of it and added columns with the count of total movies, total shows, total media, and ratios and grouped it by each country. I did this by taking the sum of each category's column after grouping them by country and type of media (tv show vs movie). For the proportion, I just divided it by the total count I had calculated earlier. I then rounded all the values in the data frame to 3 digits after the decimal point to keep everything clean and avoid extra clutter from long percentages.

Most of the challenges I encountered were due to extra clutter and disorganization. I found that once I cut down to only the variables I needed and annotated each step I wanted to do in the comments before getting started, everything went smoothly and I was able to avoid losing track of my work, especially in cases where I'd stop midway through and return hours later to finish.

Analysis and Visualization

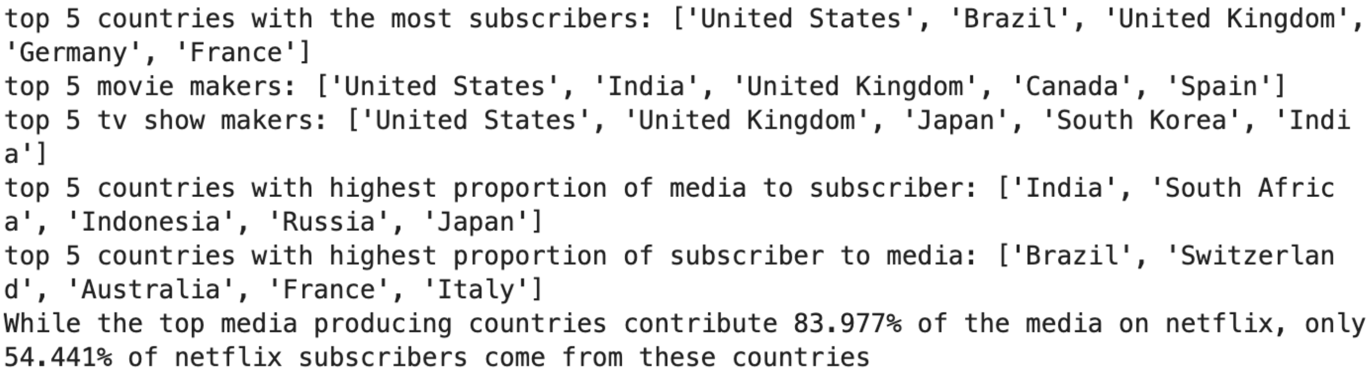
For my analysis and calculations, I started off by calculating the total amounts for each variable (subscribers, movies, shows, and all media regardless of type) and storing them into variables. I then used those values to calculate the percentage of how much each country contributed to the category. One thing I was particularly interested in was looking at the proportion of media made vs subscriber count per country, which I also calculated and added as a column to my data frame.

I also wanted to focus on the top 5 countries in each category. At first, I tried to do this by working within the constraints of what I had, but I realized it would be easier and cleaner to store the sorted values and their country into a dictionary and make a list of the first 5 keys (aka the top 5 countries).



```
#dict of countries that have made the most shows, list of top 5 countries
top_shows = dict(by_type.set_index('country').sort_values(by=['shows'], ascending=False)['shows'])
top_5_shows = list(top_shows.keys())[:5]
```

I was particularly interested in exploring the relationship between the countries that produce the most media and how many subscribers come from those countries. I printed a summary of my findings in my code:



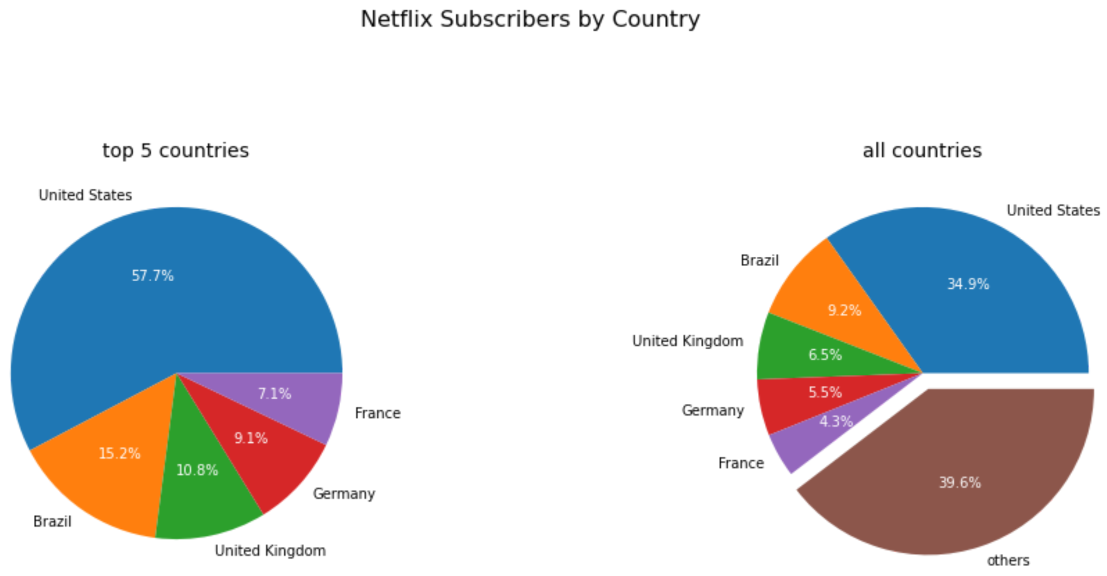
```
top 5 countries with the most subscribers: ['United States', 'Brazil', 'United Kingdom',
'Germany', 'France']
top 5 movie makers: ['United States', 'India', 'United Kingdom', 'Canada', 'Spain']
top 5 tv show makers: ['United States', 'United Kingdom', 'Japan', 'South Korea', 'Indi
a']
top 5 countries with highest proportion of media to subscriber: ['India', 'South Afric
a', 'Indonesia', 'Russia', 'Japan']
top 5 countries with highest proportion of subscriber to media: ['Brazil', 'Switzerlan
d', 'Australia', 'France', 'Italy']
While the top media producing countries contribute 83.977% of the media on netflix, only
54.441% of netflix subscribers come from these countries
```

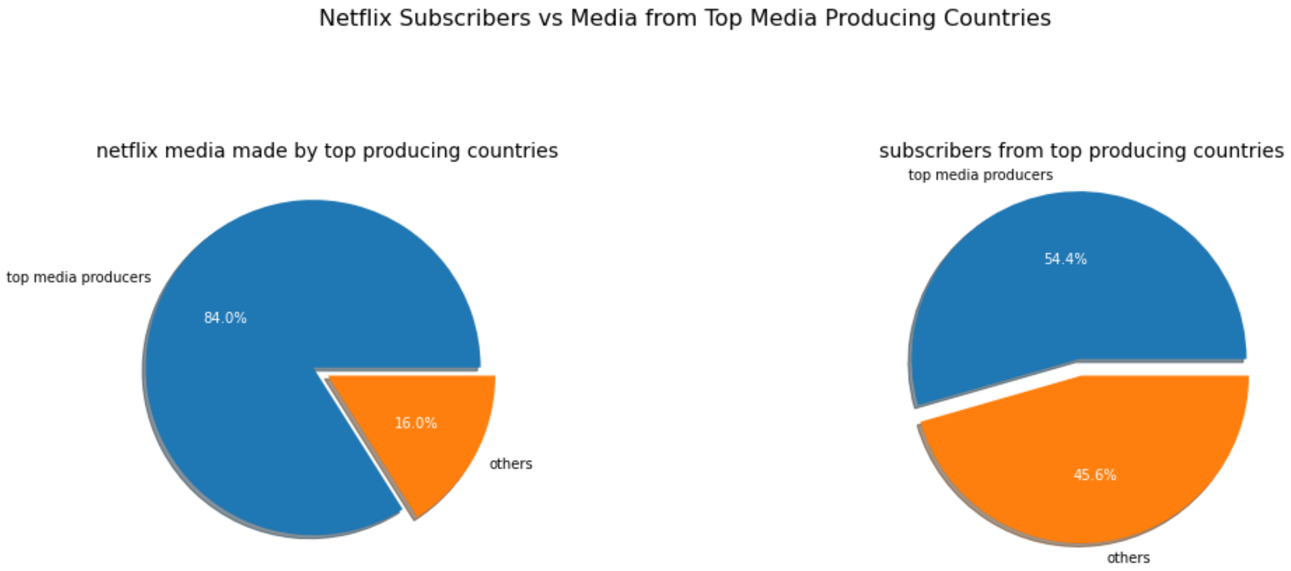
As I suspected, most of the subscribers came from western countries while a significant amount of media contributors came from Asian countries. While India was in both the top 5 movie and show makers and Japan and South Korea are in the top tv show makers, the countries with the most subscribers do not include any Asian countries. The countries with the highest proportion of media made compared to subscribers aligned with countries with big media production but relatively underprivileged populations. The USA was in the top 10 "givers" (aka low media:subscriber ratio) because Netflix is an American company and produces most content here. Other "giver" countries were due to other reasons (ex: India has Bollywood producing media but high poverty rates, meaning huge swaths of the population wouldn't be able to afford Netflix in the first place).

I think perhaps the most important conclusion I found through my calculations was that while the top media producing countries contribute 83.977% of the media on Netflix, only 54.441% of Netflix subscribers come from these countries, meaning almost half of the countries pretty much only consume media and contribute very little.

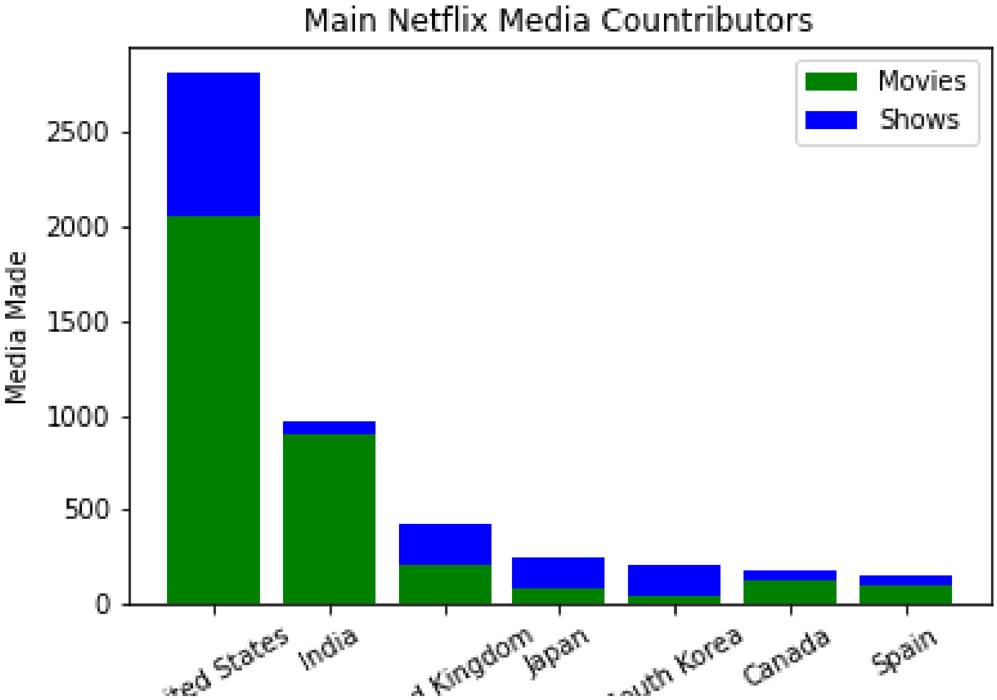
Other findings are summarized below in the visualizations, including information like how many tv shows vs movies the top contributing countries make, the Netflix subscribers by country, etc.

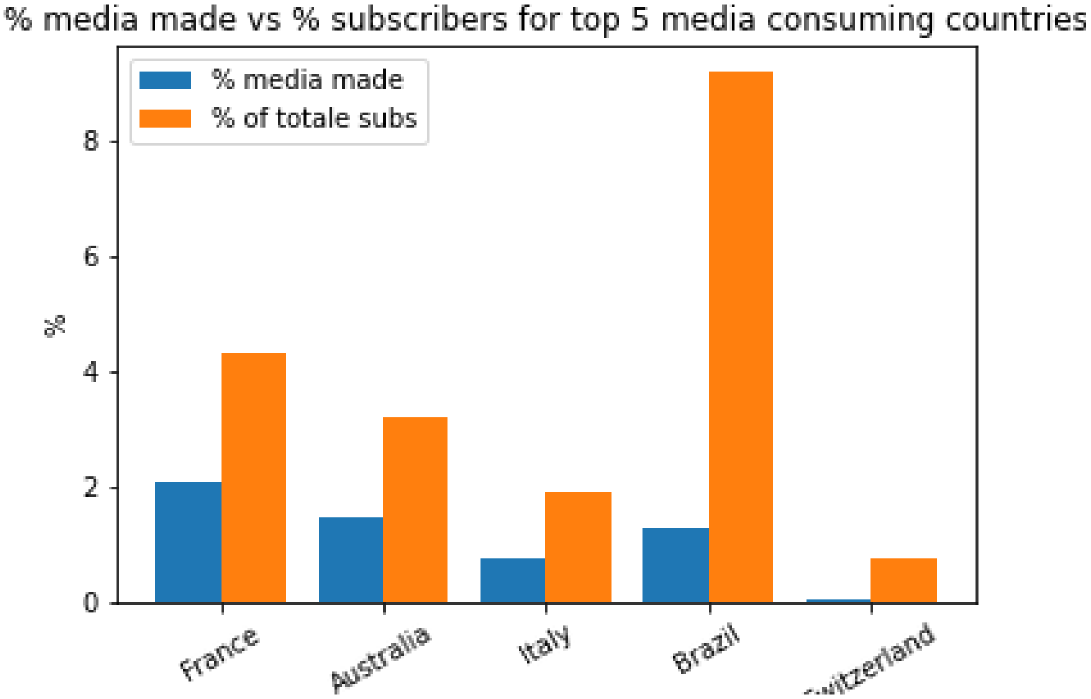
Visualizations



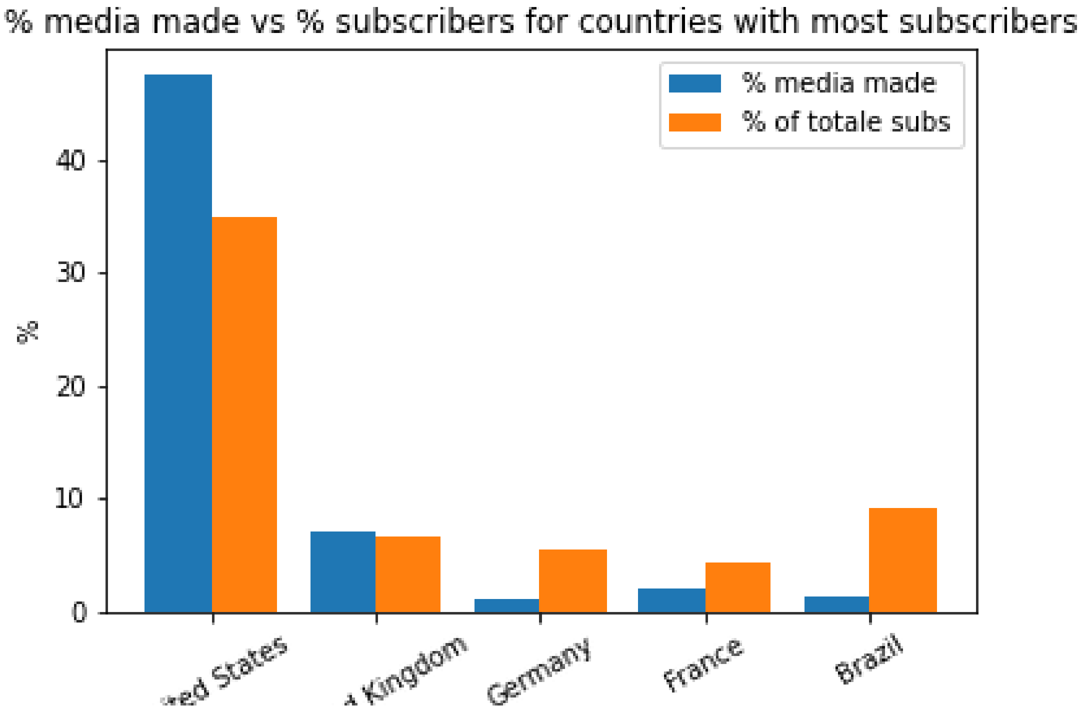


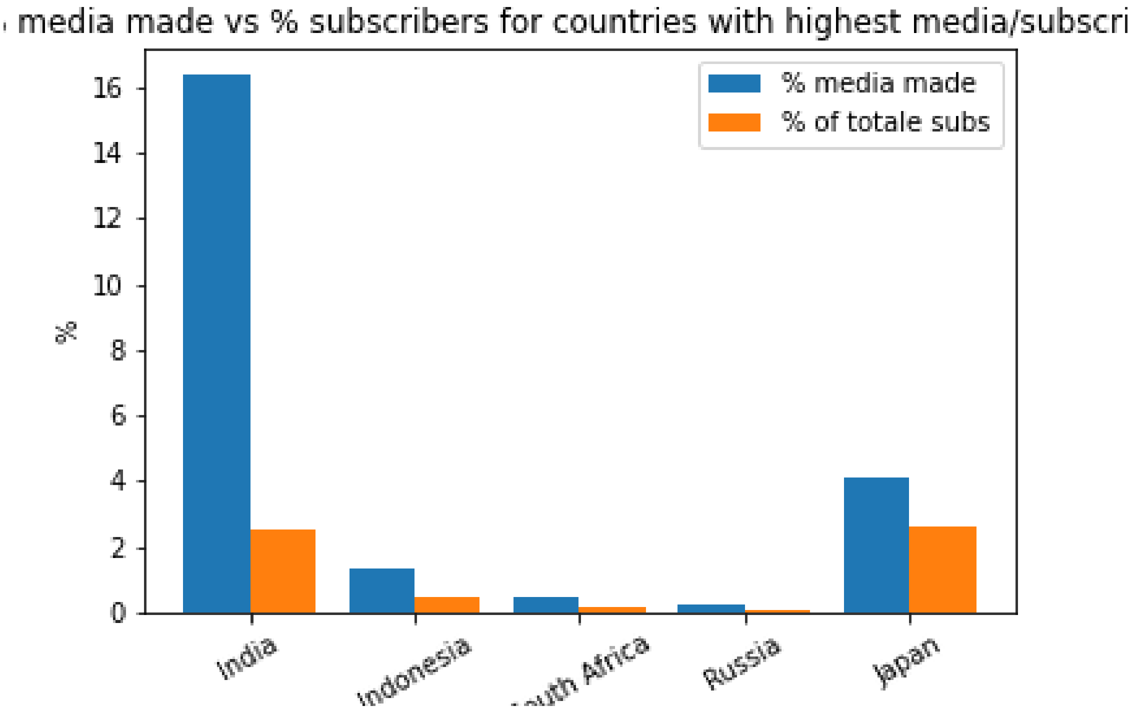
Visualizations





Visualizations


% media made vs % subscribers for countries with most subscribers
% media made
% of totale subs
ited States
d Kingdom
Germany
France
Brazil


media made vs % subscribers for countries with highest media/subscri
% media made
% of totale subs
India
Indonesia
outh Africa
Russia
Japan