STAT 3280: The Earth & I

Ani Ponugoti

2023-04-19

For this project, I decided to compare my Netflix streaming history, which I was able to download from my Netflix profile, to the Netflix weekly top 10 charts for both the United States and India, which are available on the Netflix Top 10 website and at the time that I downloaded them, had rankings for every week from June 28th, 2021 to April 9th, 2023. I also decided to briefly compare my Spotify streaming history, with a focus on the artists I listen to the most, to the weekly top artists on Spotify in India and the United States. I requested my streaming data from Spotify, which ranges from April 8th, 2022 to April 9th, 2023, and the weekly top artists are available on the Spotify Charts website. Because the dataset of my watch history that I downloaded from Netflix only had the title of the content, season, episode name, and date I watched it, I decided to merge my data with a dataset of the shows and movies available on Netflix in the United States as of March 2023, which was collected from JustWatch, a streaming guide website. This dataset has a lot of attributes of the content that helped me to make these visualizations, such as the type of content, genres, production countries, and IMDb score, among others.

I did have to do a lot of data cleaning while merging my data with this dataset though because some titles would be present in both datasets but would be spelled slightly differently or have different special characters and I would have to manually change the names to match. Some of the content I have watched has been recently added to Netflix as well (after March 2023) and wasn't available in the dataset I was merging with, so I manually filled out the attributes from JustWatch necessary for those ones, and I also had to take care of duplicate content titles. After doing this, I removed any content that did not merge from my data because it meant it had been removed from Netflix and I did not want to consider those. Most of the genre values for the content had multiple genres, so I decided to use the primary genre and production country (which only had a few titles with multiple) rather than the multiple genres and/or production countries because after looking through much of the content, the first genre seemed to accurately depict the genre of the content and the first production country is the primary country that the content was produced in. I also merged the top 10 weekly Netflix content in India and the United States with the Netflix dataset I merged my data with. I had to do the same process with changing titles, manually inputting attributes from JustWatch, and taking care of duplicates titles, but with all of these titles, it was much more tedious and time consuming. I decided that although the dataset I was merging with was content available in the United States and some content from India may not be available in the United States and may not be in this dataset, I only wanted to look at content that I would also have access to in the United States, so I dropped them. I also made sure that for all of the data I collected, I only looked at content made in the United States or India, as that is what my project is focused on. I also looked at only unique titles for some of the plots because I did not want shows to repeat for every episode watched in some instances– I only wanted to consider them once.

I used Plotly for all of the Netflix plots because I thought it would be useful for the plots to be interactive and see other related information from hovering. I also decided to color elements representing India green, one of India's flag's colors, and America blue, one of America's flag's colors. For all of the Netflix plots, I also decided to remove the legend because the elements are hovered over, the country name is displayed, and it is nice when these supplemental elements are integrated into the graph rather than being off to the side. I also made all of these Netflix plots into subplots so that it would be easy to make comparisons across my different data sources. Additionally, I gave all the plots active titles to emphasize anything interesting that may be going on in the plot rather than just stating what the plot is about. In the first plot, I decided to show percentages of the unique Indian vs. American content consumed in each year because I wanted the

y-axis to be the same for comparability and there were much larger counts for the Netflix top 10 weekly data for both the United States and India when compared to my Netflix watch history. I did not want to mask this though, so I put both the counts and percentage for each of the bars. For the second plot, I wanted to compare the percentage of unique American content consumed to Indian content each year by the type of content– movies or shows. Thus, I faceted the plots by the content type and this helped to compare across types for the different data. I also added labels of which Netflix data was being compared next to the respective rows for clarity and I included count along with percentage and country in the hover text for transparency. I also faceted the third plot by genre for similar reasons and I wanted to show how the percentages of genre watched changed monthly because I thought yearly might be too broad and I wanted to see if I could see any more specific trends. I also made sure that instead of looking at just unique titles, I looked at the unique titles for every week in that month, so as to allow for titles that were streamed or made the charts more than once in a month to be included without counting them more than once in a week. I included the month, percentage, and count in the hover text across the lines. For the fourth plot, I wanted to look at the change in average IMDb score for Indian and American content both yearly and weekly in order to see if there might be any specific trends month to month, and I included the country, month, average IMDb score, and count in the hover text.

In order to create my last set of plots regarding my top artists on Spotify and their rankings weekly compared to the Spotify weekly top artists charts for India and the United States, I had to do a bit of data cleaning first. I subsetted my Spotify listening data to only include songs that I listened to from February 10th, 2023 to April 6th, 2023 because I wanted to look at my data across the most recent 8 weeks I had data for in order to make an effective and readable plot of my top artist ranks over time. Because the Spotify weekly top charts consider a new week as starting on Friday, I followed that too and made sure to start on a Friday and compare my data to the data for the weekly top artists from India and the United States during the same time period. I also only considered songs that I listened to for at least 30 seconds as plays because that is how Spotify counts streams. I was then able to get my top artist for each of these weeks along with their ranks, but I had to take care of ties. I decided to break ties by giving the higher rank to whoever I streamed more in this 8 week time period, and for my case, I had to take care of two-way and three-way ties. With my top artists for each week and their ranks figured out, I was then able to create my bump chart. I wanted to use a bump chart because it nicely plots the ranks as a point and connects the grouped points to show trends across time. For my top artists, I colored the Indian artists in shades of orange, one of India's flag's colors, and the American artist in blue to represent one of America's flag's colors. I also left the points that represented artists who were not consistently in my top 6 artists colored gray because I wanted to emphasize and focus on the artists who are consistently among my top artists. I similarly created the plots for the weekly top artists in the United States and India on Spotify, but I did not have to do any ranking or tie breaking because the ranks were already given to me on the website for each week. I colored the points on the United States plot in various shades of the colors on the American flag, and I did the same for the India plot. I decided to lay these plots on top of each other because laying them side by side was compressing the plots a lot and I did not want the purpose of the plots to be lost in that.

Sources