**Mary Gama**

**Professor Li**

**STA 3000**

**5/22/2024**

**Streaming Platform Study**

In today's digital age, the landscape of entertainment has been transformed by the proliferation of online streaming services. With many options available to consumers, the sheer abundance of choice can often lead to decision paralysis and uncertainty. While the diversity of offerings may initially seem good, the reality is that each streaming platform comes with its own set of costs and content overlaps, adding complexity to the decision-making process for potential subscribers. Using a dataset containing TV show rating data from four major streaming platforms: **Netflix**, **Hulu**, **Disney+,** and **Prime Video**, my project aims to alleviate the burden of choice by providing valuable insights into TV show ratings across different streaming platforms, coupled with demographic data analysis. The goal of this project is to equip the modern consumer, me included, with the knowledge needed to navigate the streaming landscape with confidence and clarity.

I first set out to get a decent feel for the data and bit of exploratory analysis; I gained a small amount of insight into the demographic data, seeing most recorded TV show data was for the 16+ age group. Very little had been recorded for the 13+ age group, and a large amount of data wasn’t recorded at all, seeing the high number of “NA” entries.

A graph with numbers and a bar

Description automatically generated

I decided that going forward into my analysis, I would make sure to omit the very large NA column so I can get a clearer idea of the actual intended viewer demographics.

The very next thing I decided to explore was the data I had on the release dates of the shows that have been rated. Looking at the plot, the vast majority of the TV shows recorded in this dataset had been released after 1975, with the bulk of it being skewed far left toward the 2000s and beyond.

A graph with numbers and a line

Description automatically generated

After getting a very basic feel for the data I’m working with, I decided to answer one of my first questions heading into the project: ***Which streaming platform has the highest rated TV shows overall?***

Before calculating the overall score for the platforms, I wanted to see if there were any potentially significant differences between the two recorded rating platforms for the dataset: “IMDb” and “Rotten Tomatoes”. When calculating the correlation coefficient, we can see it is about 0.46. The conclusion I drew from this is generally, as the IMDb score increases, so will the Rotten Tomatoes score. The correlation isn’t a perfect one “1” however, so there is a bit of wiggle room between the two ratings. When plotting the rating distribution between IMDb and Tomatoes, we can see they have very different rating distributions.

A graph of a function

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated

It appears IMDb ratings have a considerable left skew. The distribution of Tomatoes scores is concentrated heavily around the 50/100 score. If I had to guess, the reason for this would be the fact that the Tomatometer takes critic scores into account, and therefor Rotten Tomatoes might have more professional critics rating the TV shows a bit harsher. After realizing the distribution for both rating sites were a bit different, I decided to see how much the average score per streaming platform differed for both IMDb and Tomatoes.

While it took me a while to figure out how to properly average this data, after some consideration, I realized that since the platform columns are binary (with 1 indicating the presence of a show on the platform and 0 indicating its absence), I could easily calculate the mean scores for each platform. To do this, I filtered the rows where each platform's column equals 1 and then calculated the mean of the IMDb and Rotten Tomatoes scores for those rows. This approach allowed me to group the data by platform and compute the average ratings for each. I created two data frames: one for the average IMDb scores and another for the average Rotten Tomatoes scores, each with the platform names and their respective average ratings, and came up with these two plots:

A graph showing different colored rectangles

Description automatically generated

A graph showing a number of different colored squares

Description automatically generated

Looking at the average ratings per platform, it looks like if you trust IMDb scores more, Amazon Prime barely edges out the competing platform, and is the number one platform for best overall IMDb score. Conversely, if you think that Rotten Tomatoes has the better scoring system, the very opposite is true-- your best bet would probably be to subscribe to Hulu and avoid Prime Video at all costs. While comparing the two rating systems has been interesting, I decided to scale the IMDb scores, so they were using a 100 point scale like Rotten Tomatoes. After scaling the data, I combined the both of them to better gauge “overall score”.

A graph of multiple colored squares

Description automatically generated

After doing this, it seems like it is fair to consider Hulu the overall reigning champion of TV shows offered, and so the average consumer would probably be better off sticking with them as their monthly subscription of choice.

Now that the best overall streaming service by overall ratings has been found, I moved on to my next question: *Which platform has the best TV shows for a given age demographic (ie, 7+,16+,18+)?* This question is aimed at helping customers select a subscription service best suited to their age-specific content preferences-- particularly for those seeking family-friendly programming suitable for younger audiences.

A graph of different colored squares

Description automatically generated

With that in mind, I produced the above plot. I had a bit of an inkling Disney+ would lead in the 7+ age demographic, but it also seems like they lead very far in every demographic that they have available on their platform. Unfortunately, I believe they are beaten out in overall streaming score due to not having a 13+ category. Speaking of that specific demographic category, Prime Video has the absolute worst 13+ offerings out of all the streaming platforms by a *large* amount. Out of curiosity, I had tested the correlation between age ratings and overall score—the test returned a correlation of around .20. While this isn’t a perfect 1 correlation, it does show that overall score tends to go up as age rating does, weakly. I still think this is useful to know, as I can generally feel great knowing that the content for my age range (18+) tends to be rated higher.

As I wrapped up my analysis of streaming platforms and their content ratings, I pondered a larger question: Are TV shows getting better or worse over time? With many speculating on declining quality in entertainment, I decided to investigate. My final task was to examine the average ratings of TV shows over the years. This simple exploration aimed to uncover any noticeable trends in show quality over time. By doing so, I hoped to provide some clarity in the ongoing debate about the state of television. The results may or may not shock you:

A graph with red lines and numbers

Description automatically generated

At first glance, it seems like TV show quality might be declining over time, judging by the IMDb and Rotten Tomatoes scores provided. However, we need to keep in mind a key finding from earlier in the analysis. I discovered that most of the recorded scores are from after roughly 1975, with the majority after 2000. This is crucial because it means there's more recent data available, which includes both positive and negative scores. As a result, the trend line may appear to slope downwards due to these newer, potentially lower scores. Ultimately and unfortunately, the dataset lacks enough historical data to draw a definitive conclusion about TV quality over time. To get a deeper understanding of the trend line, I crafted a linear regression model, once again examining the effects of Overall Score ~ Year. As the plot visualization shows, the summary of the linear model gives a negative estimate with a p-value of 0.00289, which is low enough to show that there is a strong relationship between score and year. However, as noted before, this is very deceptive due to a lack of available data in the dataset. So, while I think the idea of this model is interesting, it is probably not representative of reality.

To conclude this report, I believe my exploration into streaming platforms and TV show ratings has revealed intriguing insights overall. Hulu emerges as a frontrunner for viewers seeking quality content across various genres and age groups. However, individual preferences may vary. I've also noted a mild trend: older age demographics tend to correlate with higher show ratings, offering valuable guidance for those seeking age-appropriate content. Regarding the overall trajectory of TV show quality over time, while initial observations hinted at a decline, closer analysis highlighted the impact of skewed data towards recent years. As a result, definitive conclusions remain elusive, calling for further research. In essence, I feel my study lays a foundation for understanding streaming dynamics, and I hope it can empower whoever reads it to make informed choices amidst the abundance of digital entertainment options—I, for one, will tend to see Hulu more favorably from now on, and I know my bank will thank me for the newfound decisiveness.

**APPENDIX**

The only extra tidbit I want to append is the [dataset source](https://www.kaggle.com/datasets/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney).