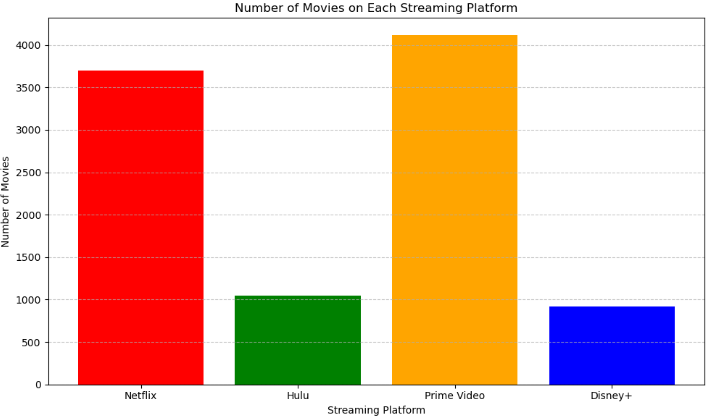
Streaming Services Movie Reviews Affecting Business

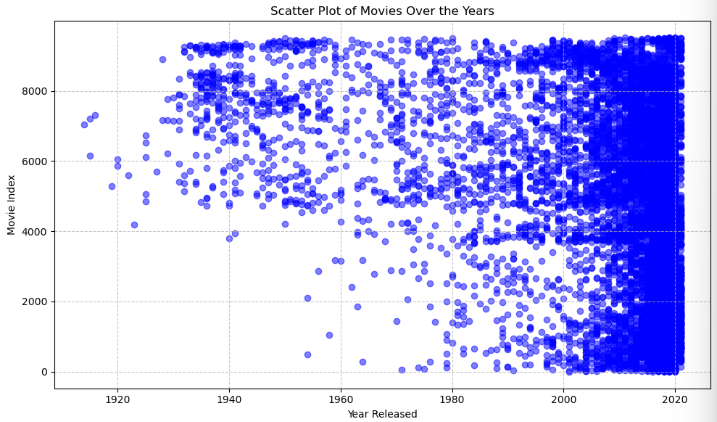
By: Taylor Woodington

In the early 2000’s streaming services such as Netflix and Hulu were created for the public to access their favorite movies or shows whenever they would like to. Since then, many more streaming services such as Prime Video, Disney+, and more have surfaced as a service to the common household. As of this year, 99% of U.S. households are subscribed to at least one streaming service. The price of all the streaming services can add up quickly and the average person will choose between one and three platforms to subscribe to with over 2,000 choices being available. (*Durrani, 2024*) This poses a problem for the streaming services, with there being so many choices, the streaming service companies need to make sure they are staying relevant and good enough to be one of the chosen platforms amongst the population. When pitching this problem to a group of stakeholders I would explain the statistics of the matter where Netflix, Disney+ and Prime Video are within the top five streaming services as far as subscribers go. For those in the top five, they may feel like they do not have to worry, however there are other services such as HBO Max and Paramount+ which are also included in this top five and above some of the others (*Durrani, 2024*). The problem for Hulu is that it is not in the top five, however, we are not sure of the changes now that they have merged with Disney+. All together, these four streaming services have been highly regarded by the public and need to stay relevant. If they do not have relevant or enjoyable movies, subscribers will start to look for what they want elsewhere which means competition for these streaming services. This problem is important to solve for the stakeholders because losing subscribers means losing money and experiencing a decline in the business. Finding out which streaming service could be more enjoyable for specific ages or overall, from movie ratings, could help those running the business understand what direction they may need to move in. For example, adding movies with higher ratings, or movies that cater to children could be beneficial for the companies depending on the analysis. This project aims to show which streaming service has the most highly rated movies and if there is a relationship between the ratings and platform, as well as the age range of most movies in amongst the services. With the dataset being acquired from Kaggle, the specific origin is unknown, but it lists three contributors with 30.4 thousand downloads of the dataset and 240 thousand views so the public does have access to this data and can see which platform has the best rated movies from the four (*Bhatia, 2021*).

After acquiring the data, I read the dataset to analyze its structure and decide what visuals would be beneficial for exploratory analysis. Upon seeing there are four types of streaming platforms I decided it would be beneficial to analyze how many of the movies in the dataset are from each platform. The results are shown in the graph below revealing that most of the movies apart of this data analysis are from Netflix and Prime Video.



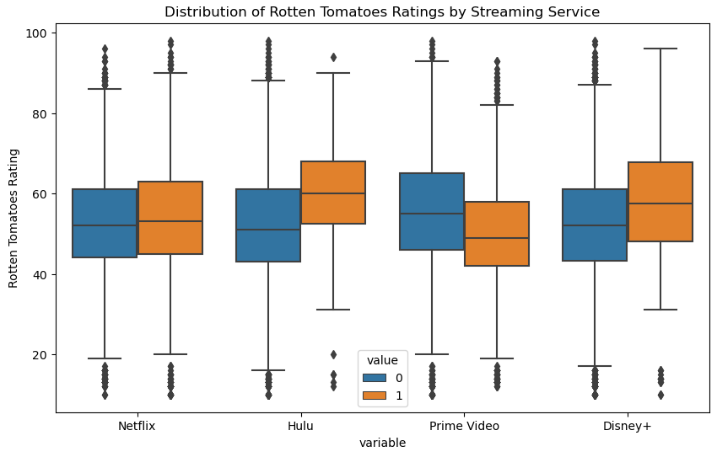
From there I wanted to see if there was a pattern of older or newer movies being a part of the dataset. I made a scatter plot of the number of movies released over the years. The results showed that most of the movies on all streaming platforms were released in the 2000’s, specifically in the late 2000’s so they appeal to those looking for newer releases movies.



While newer movies are appealing, the reviews will be something that can increase the demand for a service. To visualize the ratings data, it needed to be reformatted to be one solid number that would read better for a histogram. From there I made a histogram of the ratings across all movies in the dataset, dropping any rows with a NAN value if any. Those that include a NAN value will not be useful for the analysis seeing that it has such a heavy focus on the ratings. Overall, the histogram showed ratings to fall between 40% and 80% with few ratings being above 85%. Originally upon acquiring the data I planned to analyze data that had movies with an “excellent” rating with a score of 85% and above. However, if I were to do that, there would not be enough data to analyze, so I decided to focus on all the ratings ranging from 0% to 100%. Taking this information and the information from the first graphic, knowing that Netflix and Prime video has the most movies, I made two more histograms to show the count of movies across all ratings for both streaming platforms. Both Netflix and Prime Video were very comparable to the overall ratings histogram. However, Netflix seemed to have more movies with over 85% as a rating and a more normal distribution. Whereas, Prime Video had more of a slight skew towards negative scores showing that there is a difference between the platforms being analyzed. All of this allows for a better understanding when going into the preparation of the data for the future model and milestone two.

In milestone two I updated my analysis mission to include a focus on age and made it clear that my target for my analysis is the companies who own the streaming services, more specifically those in charge of increasing subscribers to the streaming service. My models should be able to show which streaming service has a higher relationship with better reviewed movies. It would also show which movie age groups have better reviews to indicate the kind of audience that the streaming service should be advertising to or to change their ways for the age groups they are not advertising to. In addition, I began to clean and adjust the data. First, I started with looking for duplicates through the title column to make sure there were not multiples of the same movie that would create bias through the platforms. After finding no duplicates I removed the “Title” and “Type” column to rid of having multiple columns which would identify the movies. There is not a reason to have multiple columns which serve the same purpose. After cleaning out those columns I decided to check for any missing values within important columns for analysis such as the rating, minimum age, and all the individual streaming platforms to make sure there was not anything missing that would create bias. While most of the columns were not missing any values, the age column was missing values. To adjust the age values, I split the values to rid of the ‘+’ character so the numbers could be used for calculation such as finding the average of all the ages and renamed that column to be “Minimum Age Required” for watching the specific movie. Once the average was calculated for those ages, I filled the “NAN” or missing values with the average minimum age required to watch a movie. While I’ve considered that this might create some bias, I think it would create even more bias if I replaced those missing values with the minimum required age or got rid of them enter from the dataset. Lastly, when cleaning up and preparing the data for a model, I originally planned to have a column which was labeled as “Streaming Service”. This was to combine the four streaming platform columns that are one hot encoded in one place as categorical data. I thought this would be easier to use when running a model which would compare the datasets and their ratings. However, moving to milestone three and exploring models I quickly changed my mind and removed this column, deciding to use the one hot encoded columns for further analysis and modeling.

In addition to the change back in column usage, I removed more unnecessary columns such as the “Unnamed: 0” column, “ID” and “Year” because I did not plan to use them for these specific models. Once this was done, I visualized the data through box and whisker plots seeing the range of all ratings across the four streaming platforms to give more of a visual reference to compare results of the models to.



For the models I decided to make a linear regression model and decision tree model because they can give me different perspectives when it comes to the data. The linear regression is simpler and can show me simple linear relationships in the data whereas the decision tree model is more complex and could show more factors affecting the ratings on the platform. To start, I split the data into a test and training set and then moved into training the linear regression model, predicting the test set, and evaluating it. Afterwards I did the same thing all over again with the decision tree model. From there I got the coefficients from the linear regression model and feature importances from the decision tree to get a better understanding of it. The linear regression model showed Hulu to have the highest impact on movie ratings with a score of 11.7 followed by Disney+ with 10.25 whereas the decision tree showed a significance with Prime Video at an importance of 0.146 and Disney+ at 0.066. While they both gave different results, there was a focus from both on Disney+, however, Hulu and Prime Video were shown to do extremely well in the different models which creates a little bit of confusion, so I looked to the analysis of the model performance. When looking at the MSE for either of the models it is shown to be an extremely large number with the linear regression being at 165.47 and the decision tree being 138.22, which means there is room for a lot of error for both models. Looking at the R squared value, the linear regression model shows 0.083 whereas the decision tree shows 0.234, so the decision tree model explains over double the amount of what the linear regression model does for the dataset. However, neither model explains much of the data so the results cannot be taken heavily to change actions within a company but can be action to get more information and try the models again.

After looking at the ratings relationship with the platforms I ran a linear regression model to look at the relationship between the minimum required age and the different streaming platforms. As expected, Disney+ showed a negative coefficient meaning that their target audience is those of the lower age so they may not appeal to a large range of ages. The model also shows Hulu to have the highest positive impact meaning that they appeal to a broader or higher age range. Now, this dataset has not been updated within the last year where Disney+ and Hulu have now merged, this could be beneficial for both platforms to keep their companies up and appealing to a wider audience. While this model does have a lower MSE at 9.103, the R squared value is 0.063, so the model does not do a great job at explaining the whole dataset.

Overall, these models did not do a great job of explaining the dataset the way that I had hoped. They show a lot of information about Disney+ and Hulu which are the platforms that had the least number of movies in the dataset. So much of the data between Netflix and Prime Video may not be explained and the models suggest that more data may be necessary to continue with this. I do not think that the model is ready to be deployed. There was no data included about the genre of the movies on either of the streaming platforms, or any data about shows that were on the platforms as well. These factors could affect the streaming platform and possibly the ratings of the movies. I think more data needs to be collected on the movies themselves such as the genre, movie length and more, in addition to the different shows that are offered on each service. Once more factors such as that are found, I think that a multiple linear regression or decision tree model could help explain more of the data. However, the movies and shows are always changing on the platforms and some of them have now merged that the current data would need to be updated which can be a challenge. It would be easy to pull this data from the service platforms, possibly a little tedious but it could be done. Another challenge would be knowing what information may be important to pull over into a new database. Overall, I think more data needs to be collected to readjust the models and see what factors are affecting the ratings and their relationship with the streaming services.

References

Bhatia, R. (2021, December 16). *Movies on Netflix, Prime Video, hulu and disney+*. Kaggle. <https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney>

Durrani, A. (2024, June 13). *Top streaming statistics in 2024*. Forbes. <https://www.forbes.com/home-improvement/internet/streaming-stats/#:~:text=An%20overwhelming%2099%25%20of%20U.S.,norm%20rather%20than%20the%20exception>.