BAIS:3250 – Data Wrangling

Reed Ulses & Erin Schultz

Project Report
Movies IMDb Ratings and Availability on Streaming Services

1. Introduction

Movies scores and ratings can be a good reference for consumers when making movie-related decisions, but not *all* movies are available to consumers in one place. In this project, we plan to use Kaggle's "Movies Dataset: Netflix, Prime Video, Disney+" and the IMDb website to find the streaming platform with the best selection of highly rated movies to aid consumers in their subscription decision-making process. This will help consumers make informed choices based on content quality and quantity. Streaming platform analysts and executives may also be able to use this information to identify gaps in their offerings and provide more "popular" movies.

2. Data

This project uses two sources of data: IMDb[1] movie data on all movies released between 2000 and 2021 and Kaggle's[2] dataset about streaming availability of movies.

2.1 IMDb Scores Website

We collected data from IMDb's website that displayed data about all movies released between January 1, 2000, and December 31, 2021. The entire site contained data on 89,781 movies over 360 pages. Because this dataset is so large, we knew we needed to trim down the number of observations that we were going to collect given the amount of time the process would take (reference section 4). With the help of our professor, we wrote a web crawling script to scrape data from the IMDb site on all movies on this large website but ended with information on movies that were released in the years 2019, 2020, and 2021. This returned information on 13,244 movies where we collected the title of the movie, the year it was released, and the IMDb rating. After the scraping was completed, we created a data frame and saved it into a csv file (*raw_imdb_53.csv)* to use when merging with our second data set. The scraped data did not need much cleaning aside from some movies not having a rating listed. We later had to fill all these blank cells with 0's in our analysis.

2.2 Streaming Services Availability

Our second dataset came from Kaggle and is titled "Movie Dataset (Netflix, Prime Video, Disney+)". This data gave us a list of movies with information about each including release

[1] www.imdb.comelease_date=2000-01-01,2021-12-31

[2] www.kaggle.commovies-dataset-netflix-prime-video-disney

year, the recommended age group for the movie's audience, its Rotten Tomatoes score, and whether it was a movie or TV show. The dataset also included a column for Netflix, Hulu, Prime Video, and Disney+. Each movie had a 1, displaying availability on that platform, or a 0 displaying not being available on that platform in these columns.

The dataset includes data on 9,515 unique movies. There are 11 columns in the dataset but for our project we are only focused on the movie title, release year, Rotten Tomatoes score, and availability on each streaming platform. All other columns were ignored in our merging with the IMDb data. The only cleaning of the dataset we had to perform was converting Rotten Tomatoes score from a number out of 100 to simplify the number.

2.3 Combining IMDb Scores and Availability

Since both data sets included the same movie title and movie release year, we merged based on those two factors. After determining the same column names and data types existing in both dataframes, we used an inner join to ensure that only the rows with matching keys in both dataframes were kept. In total, 483 movies overlapped between the scraped IMDb dataframe and the imported streaming services dataframe. Table 1 contains the data dictionary for this merged dataframe.

Table 1 Data Dictionary

Field	**Type**	**Source**	**Description**
ID (id)	Numeric	Kaggle	Unique identifier for each movie
Title (title)	Text	Both	Full title of the movie as it appears on the streaming platforms
Year (year)	Date	Both	Release year of the movie, when the movie was first made available to the public
Rotten Tomatoes (rotten_tomatoes_percentage)	Numeric	Kaggle	The movie's score on Rotten Tomatoes on a scale of 0 to 100
Netflix (Netflix)	Categorical	Kaggle	A binary indicator (0 or 1) of whether the movie is available on Netflix, with 1 indicating availability
Hulu (hulu)	Categorical	Kaggle	A binary indicator (0 or 1) of whether the movie is available on Hulu, with 1 indicating availability

Prime Video (prime_video)	Categorical	Kaggle	A binary indicator (0 or 1) of whether the movie is available on Amazon Prime Video, with 1 indicating availability
Disney+ (disney_plus)	Categorical	Kaggle	A binary indicator (0 or 1) of whether the movie is available on Disney+, with 1 indicating availability
IMDb Rating (imdb_ratings)	Numeric	IMDb	The movie's average rating on IMDb on a scale of 0.0 to 10.0

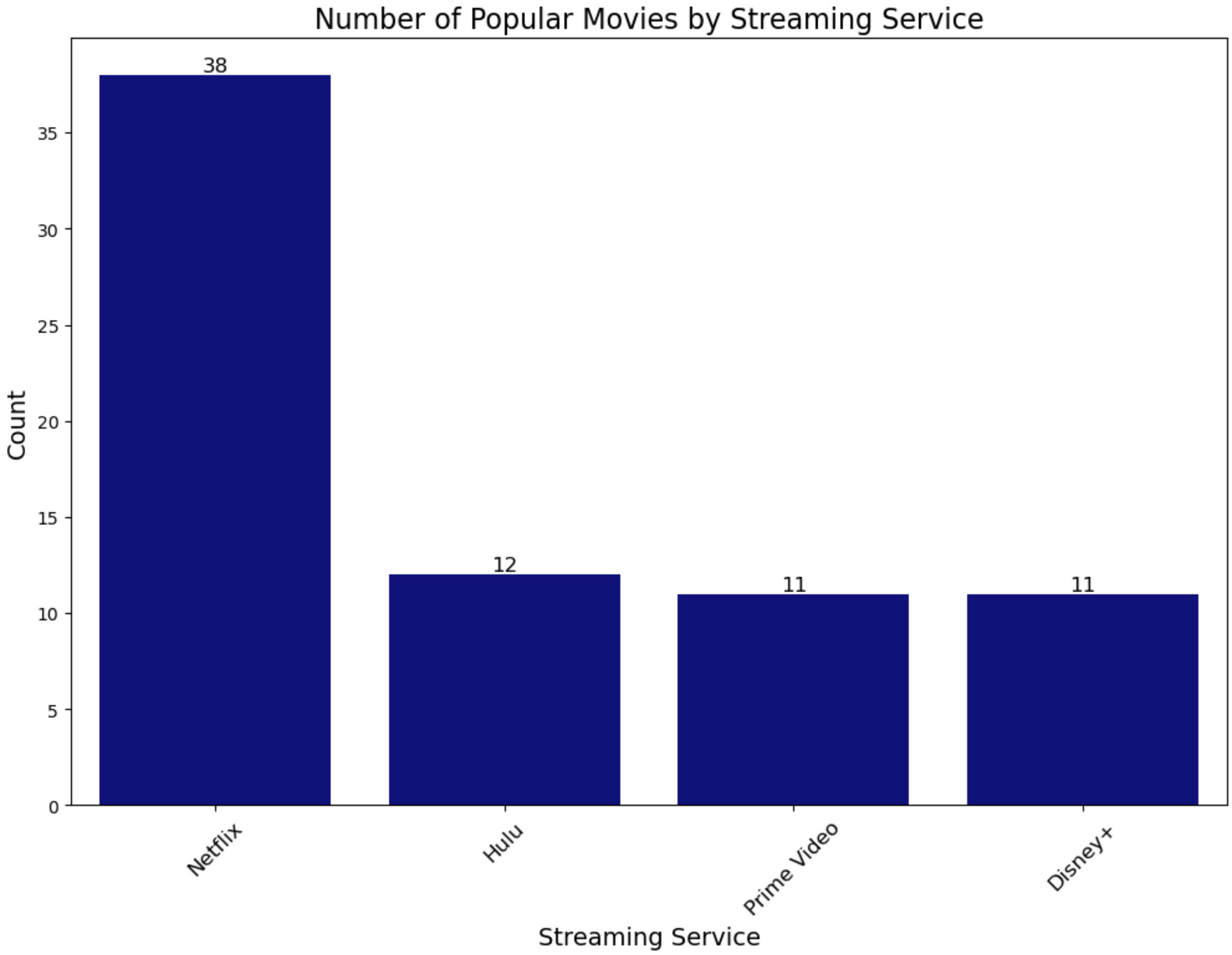
3. Analysis

3.1 Popular Movies

To begin our analysis, we decided to find out which streaming platform had the most popular movies. To do this, we had to set thresholds for what constituted a popular movie. According to Rotten Tomatoes' website, a score of 60% or higher is consider "fresh", which is their highest rating. We based our popularity threshold for Rotten Tomatoes score off this information. For IMDb rating, we decided that the top 25 percent of movie ratings would be considered popular.

In our merged data frame, Rotten Tomatoes score was listed as a number out of 100 (i.e. 65/100). Because of this, we had to create a new column with the score listed as a single number (i.e. 65), then we were able to drop the original Rotten Tomatoes column. Our next step was calculating what the score would place a movie in the top 25 percent of IMDb scores. This score turned out to be a 6.6 out of 10. After these tasks were completed, we were easily able to filter for movies that had a Rotten Tomatoes score of 60 or higher and an IMDb rating of 6.6 or higher on each streaming platform. We found Netflix to have the highest number of movies meeting or criteria with 38. Hulu had 12, and both Prime Video and Disney+ had 11. To display the variation between Netflix and the other streaming platforms, we created a bar chart that is shown below in Figure 1.

Figure 1 – Number of Popular Movies by Streaming Platform



3.2 Platform and Overall IMDb Ratings and Rotten Tomatoes Scores

To compare average Rotten Tomatoes scores and IMDb ratings by platform, we had to calculate the average score and rating overall as well as each individual streaming platform's averages. We began by looking at IMDb scores. Because our goal was to compare each streaming platform to the overall average, we began by calculating the average IMDb rating for the entire dataset. This came out to be 5.64. We then calculated the average IMDb rating for each individual streaming platform. The results came out to be Disney+ at 6.72, Netflix at 5.83, Hulu at 5.61, and finally Prime Video at 5.16.

Since Rotten Tomatoes score is on a different scale from IMDb, we had to repeat the process to find the average Rotten Tomatoes scores. Once again, we started by calculating the overall Rotten Tomatoes score which came out to be 56.17. Next, we calculated the average Rotten Tomatoes score for each streaming platform. Disney+ and Netflix had the highest averages at 60.36 and 60.12, respectively. Hulu had a slightly lower average rating at 57.82, and Prime Videos was much lower at 49.06.

Based on this information, we created a bar chart for IMDb (Figure 2) and Rotten Tomatoes (Figure 3). These charts show the overall average as well as each platforms average. As you can see that Disney+ has the highest scores on both rating systems and Prime Video had the lowest on both. On both rating systems, all platforms had very similar or higher averages compared to the overall except for Prime Video.

Figure 2 – Average IMDb Rating by Streaming Platform

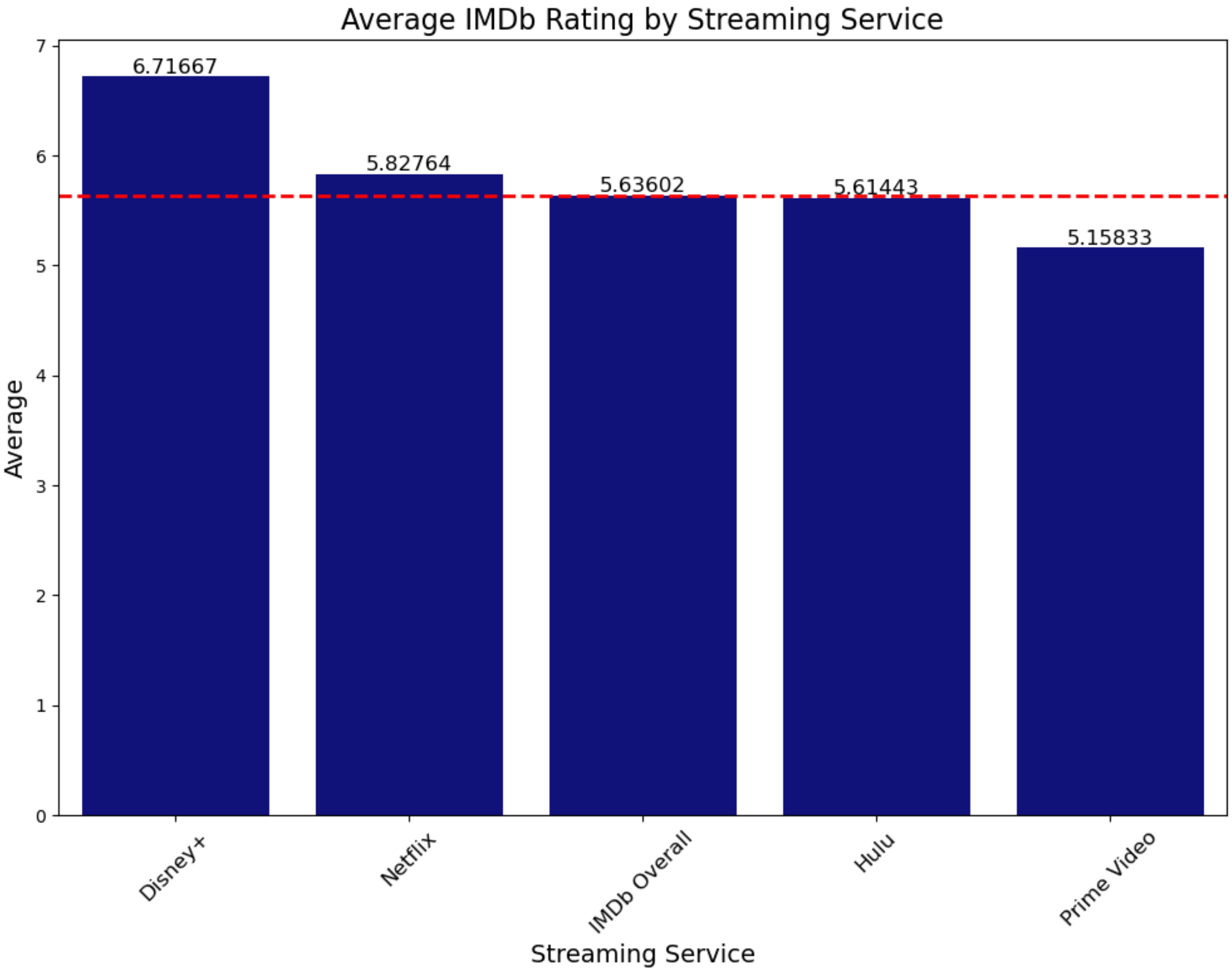
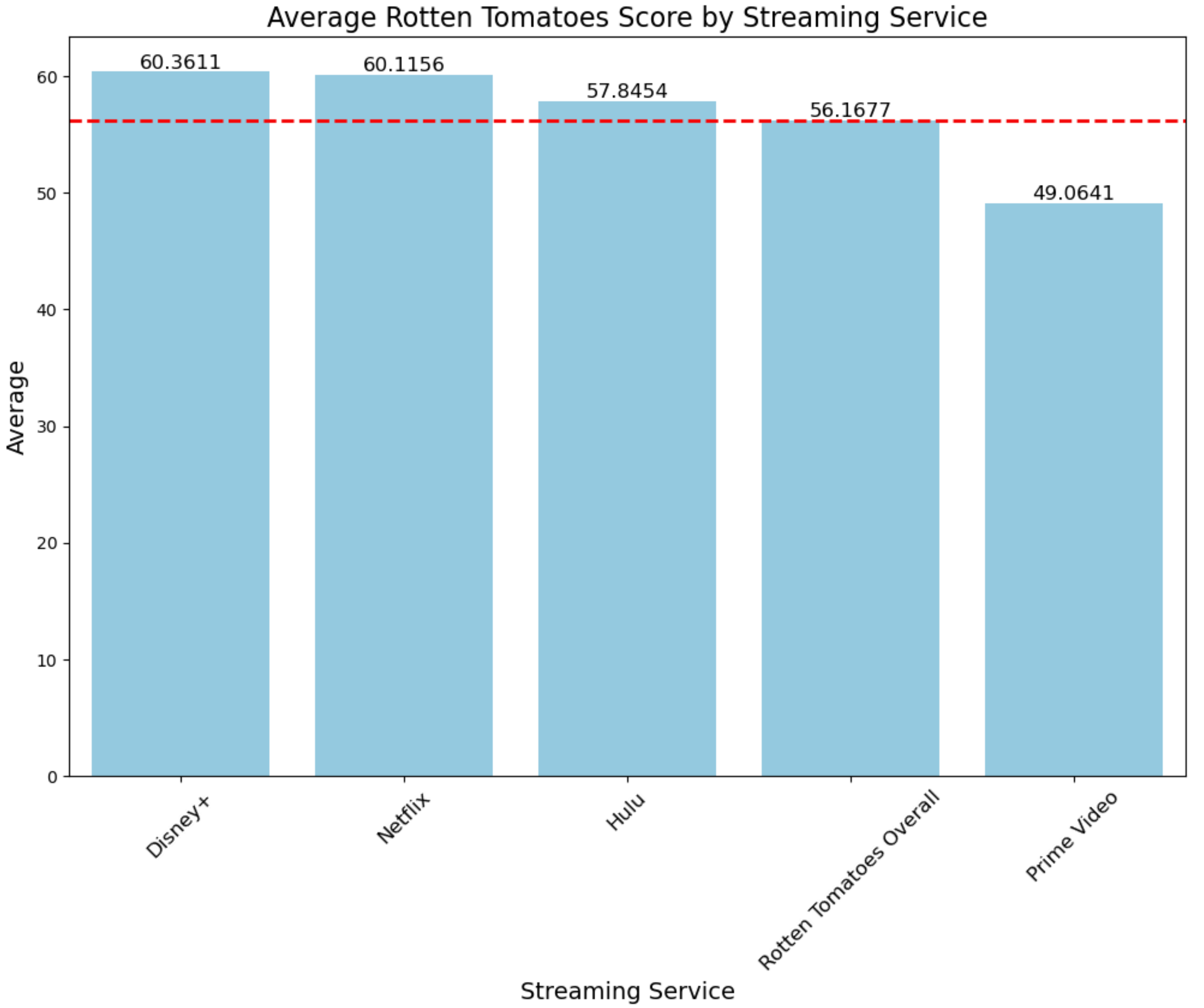


Figure 3 – Average Rotten Tomatoes Score by Streaming Platform

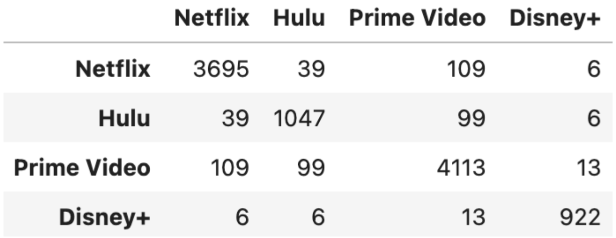


3.3 Multiple Platform Movie Availability

To find out the number of movies that are available across multiple streaming services, we had to start by finding the number of streaming services each movie is available on. To do this using the categorical 1s and 0s that indicate the availability of a movie for each of the four streaming platforms, we created a new column in the merged dataframe (movies_kaggle) that summed up the 1s and 0s in the Netflix, Hulu, Prime Video, and Disney+ columns. Where this new, calculated column was greater than 1, we knew that that specific movie was available on multiple platforms. This resulted in 253 movies that are available on multiple streaming platforms.

Once we knew that 253 movies could be accessed via multiple streaming platforms, we then wanted to understand which of these platforms had the most overlap, i.e. the highest number of movies that users could watch elsewhere and see which streaming platforms share the highest number of movies. To do this, we went through each combination of streaming platforms (Netflix and Hulu, Netflix and Prime Video, etc.) to see where each combination both had a "1" in their respective streaming platform columns. Each of these was then summed up and we needed to find the maximum given these sums. To execute this, we defined a maximum variable and went through each combination using IF/ELSE statements to find the platform with the most overlap. This gave us an answer of Netflix and Prime Video with 109 movies overlapping. This may indicate to users that having a subscription to both platforms may not be worth their money and indicate to the streaming platforms who their biggest competitor is for movie offerings.

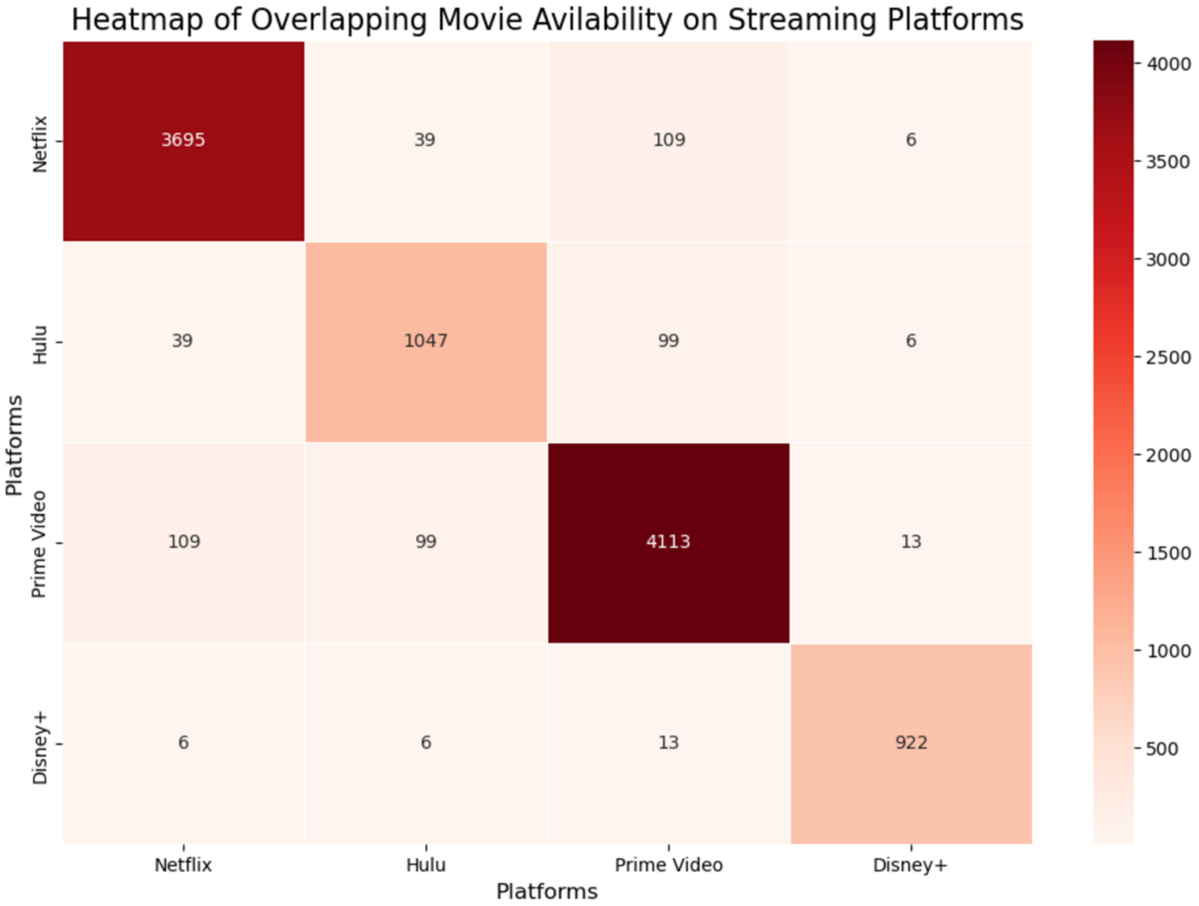
Figure 4 – Overlapping Movies DataFrame



	Netflix	Hulu	Prime Video	Disney+
Netflix	3695	39	109	6
Hulu	39	1047	99	6
Prime Video	109	99	4113	13
Disney+	6	6	13	922

With this newfound information, we created a heat map to visualize the relationship between each of these streaming platform combinations. This heat map can be found below in figure 5. Figure 4 (above) displays the DataFrame used to create this heatmap and was made using a dictionary.

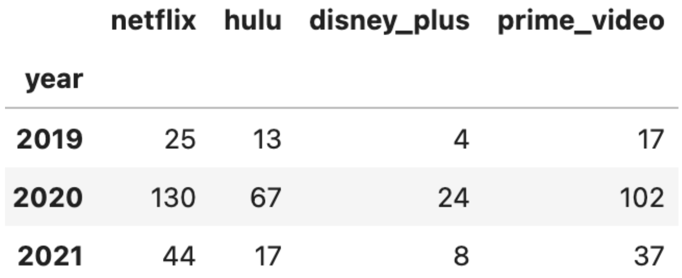
Figure 5 – Heatmap of Overlapping Movie Availability



3.4 Expanded Movie Offerings

To analyze the age of movies offered on each streaming platform, we first had to group by the year for each platform in the merged dataset. We then created a DataFrame that housed each platform as a column, and the count of movies from 2019, 2020, and 2020 as the rows. Figure 6 shows the displayed DataFrame.

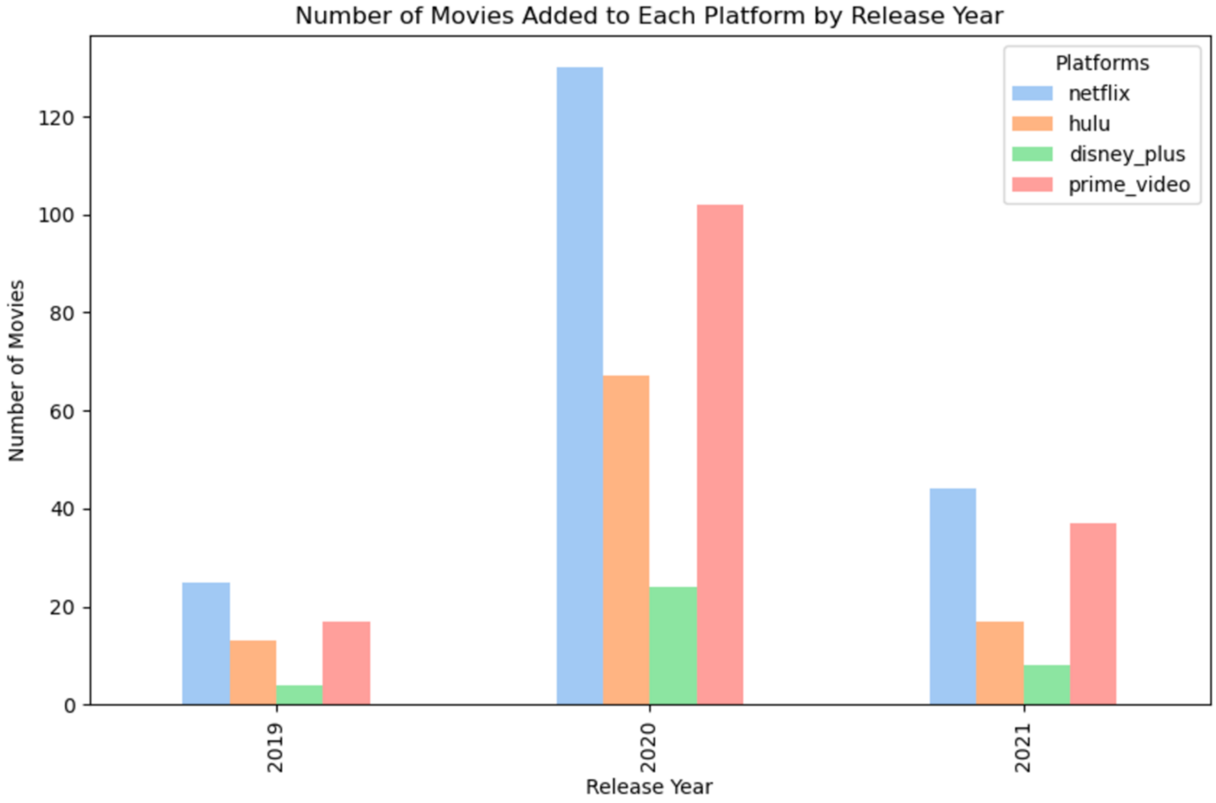
Figure 6 – Count of Movies from Each Release Year by Streaming Service



	netflix	hulu	disney_plus	prime_video
year				
2019	25	13	4	17
2020	130	67	24	102
2021	44	17	8	37

To better visualize this distribution and gain insight as to which platforms had more recent movie offerings, we plotted a bar chart. Figure 7 displays the number of movies that were released in 2019, 2020, and 2021 that are available on each streaming platform.

Figure 7 – Number of Movies Added to Each Platform by Release Year



4. Conclusion

In this project, we analyzed the effects of both IMDb and Rotten Tomatoes scores of movies, and display which streaming services (between Netflix, Hulu, Prime Video, and Disney+) have the highest rated movies. In summary, based on the analysis questions presented in our project proposal and added questions as our analysis progressed and changed, we found the following results.

1. *Which streaming platform has the most "popular" movies based on the given threshold (top 25% of IMDb ratings and Rotten Tomatoes score of >=60%)?*

   Netflix has the highest number of "popular" movies given the researched threshold. There are 38 popular movies on Netflix, compared to just 12, 11, and 11 movies on Hulu, Prime Video, and Disney+, respectively.

2. *How does each streaming platform's average IMDb rating and Rotten Tomatoes score compared to the overall average of all movies released between 2019-2021?*

   The overall average IMDb score is 5.64 for all movies release between 2019-2021. Netflix had an average of 5.83, Hulu had 5.61, Prime Video had 5.16, and Disney+ had 6.72 averages.

   The overall average Rotten Tomatoes score is 56.17 for all movies released between 2019-2021. Netflix had an average of 60.12, Hulu had 57.85, Prime Video had 49.06, and Disney+ had 60.36 averages.

   Disney+ and Netflix had higher average IMDb ratings than the overall average, while Disney+, Netflix, and Hulu had higher average Rotten Tomatoes scores than the average.

3. *How many movies are available on multiple platforms? Which platforms overlap the most in movie offerings?*

   There are 253 movies that are available on multiple streaming platforms. The pair with the most overlap (highest number of movies available on both platforms) is Netflix and Prime Video, with 109 movies overlapping.

4. *How many movies were added to each platform in each release year (2019-2021)?*

   Netflix had 25, 130, and 44 movies added to its services in 2019, 2020, and 2021. Hulu had 13, 67, and 17 movies added to its services in 2019, 2020, and 2021. Prime Video had 17, 102, and 37 movies added to its services in 2019, 2020, and 2021. And Disney+ had 4, 24, and 8 movies added to its services in 2019, 2020, and 2021.

This project has several limitations, including the fact that our original plan was to scrape around 89,418 movies released between January 1, 2000, to December 31, 2021. However, due to a time limit and computer capacity, we had to forcefully stop the scraping after ~45 hours of running, or 53/358 completed loops. This left us with 13,244 movies before dropping any null values. But once the scraped data from the IMDb website was merged with the data from the Kaggle dataset, we were left with 483 movies that were available on one of the four streaming platforms and had an IMDb rating available.

Future work on this project may include finishing the scraping with a computer that has a larger capacity and with the appropriate amount of time needed to get a more complete picture for the answers to our business questions. This would help users of streaming services made more informed decisions when deciding which services to subscribe to. This more complete picture would also help platform administrators evaluate their decisions based on some competitive analysis to identify gaps in their offerings and provide more "popular" movies.