Project Part 2

Gideon French

Data

There was not as much data as I originally imagined for the status of whether a movie was available on streaming platforms within a week of its cinematic release, which my original research questions were mainly investigating.

My original research questions were:

- Is there a relationship between box office revenues and streaming platform subscription counts over time?
- Is there a relationship between genre of movie and whether it is available to stream on a streaming platform within a week of release?
- Is there a relationship between a movie being available to stream on streaming platforms within a week of release and theater box office revenues for that movie?

I did, in exploring the data available to me, find good data to analyze and compare movies on select streaming platforms and movies in general in aspects such as gross revenue and movie scores.

My modified research questions which better fit the data I found in my research are as follows:

- Is there a relationship between movie scores and year of release?
- Is there a relationship between movie scores and which streaming platform it is available on?
- Is there a relationship between gross revenues of movies and which streaming platform it is available on?

General Information

IMDB Movies Dataset

The IMDB Movies Dataset on Kaggle consists of information relating to the Top 1000 Movies on the IMDB website by IMDB score. The data in the dataset was obtained by web scraping the IMDB website. IMDB (Internet Movie Data Base) is a website for aggregating, among other data, a movie's critical reception using scores. IMDB scores on movies are a weighted average of user votes on the quality of the movie from 1 to 10.

I subset this data by the movies available on select streaming platforms for analysis.

The data is then a subset of all movies on the online database, being only concerned with the movies that scored in the top 1000 by IMDB scores. It is a sample.

The rows consist of each entry in the dataset, which are individual movies. The columns are:

- Poster_Link - Link of the poster that imdb using
- Series_Title - Name of the movie
- Released_Year - Year at which that movie released
- Certificate - Certificate earned by that movie
- Runtime - Total runtime of the movie
- Genre - Genre of the movie
- IMDB_Rating - Rating of the movie at IMDB site
- Overview - mini story/summary
- Meta_score - Score earned by the movie
- Director - Name of the Director
- Star1, Star2, Star3, Star4 - Name of the Stars
- No_of_votes - Total number of votes
- Gross - Money earned by that movie

We are using the columns Series_Title, Released_Year, and IMDB_Rating.

A potential issue with this dataset is that though it covers a large number of movies, it does only cover the top thousand, and thus the average scores will be higher than the general population of all movies. Still, we can use it to compare how scores compare across streaming platforms or to gross revenue by subsetting it.

The data was last updated a year ago and contains movies dating from 1920 to 2020. It is appropriate to answer our research questions concerning how streaming platforms and movies are related to movie scores.

Source:

- "IMDB Movies Dataset" on Kaggle by user *Harshit Shankhdhar*

Netflix, Hulu, Disney Plus, Amazon Prime Movies and TV Shows

These datasets were all created by the same user on Kaggle. They contain data on the movies and shows available on four of the biggest streaming platforms for movies: Netflix, Hulu, Disney Plus, and Amazon Prime. We will only be using the data relating to the movies on these platforms. The data is sourced from Wikipedia, Unofficial DB Search, and FlixDatabase, and was obtained through web scraping and API calls. This data will be used to subset the other datasets by streaming platform.

Each dataset contains over a thousand entries. The data is of the entire population of movies available on each streaming platform up to the end of 2021.

The rows consist of each entry in the dataset, which are individual movies. The columns are:

- show_id - Unique ID for every Movie / Tv Show
- type - Identifier - A Movie or TV Show
- title - Title of the Movie / Tv Show
- director - Director of the Movie
- cast - Actors involved in the movie / show
- country - Country where the movie / show was produced
- date_added - Date it was added on Netflix
- release_year - Actual Release year of the move / show
- rating - TV Rating of the movie / show
- duration - Total Duration - in minutes or number of seasons

We are using the columns title, type, and release_year.

The data was last updated up to 6 months ago, varying on the dataset, and data about movies on the streaming platform up to 2021. It is appropriate to answer our questions about specific streaming platforms and to subset the other datasets of movies by platform.

Sources:

- "Disney+ Movies and TV Shows" on Kaggle by user *Shivam Bansal*
- "Netflix Movies and TV Shows" on Kaggle by user *Shivam Bansal*
- "Hulu Movies and TV Shows" on Kaggle by user *Shivam Bansal*
- "Amazon Prime Movies and TV Shows" on Kaggle by user *Shivam Bansal*

Top Box Office Revenue Data - English Movies

The Top Box Office Revenue Data - English Movies on Kaggle is a dataset of the top 1000 English movies by gross revenue. We are using the revenues of English language movies in the United States of America. The data was obtained by the screenscraping of boxofficemojo.com website. Box Office Mojo is a website that tracks the revenues of movies.

I subset this data by the movies available on select streaming platforms for analysis.

The data is then a subset of all movies on Box Office Mojo, being only concerned with the movies that placed in the top 1000 by gorss revenue. It is a sample.

The rows consist of each entry in the dataset, which are individual movies. The columns are:

- Rank - Rank of the movie by gross revenue
- Lifetime Gross - Lifetime gross revenue
- Year - Year of release

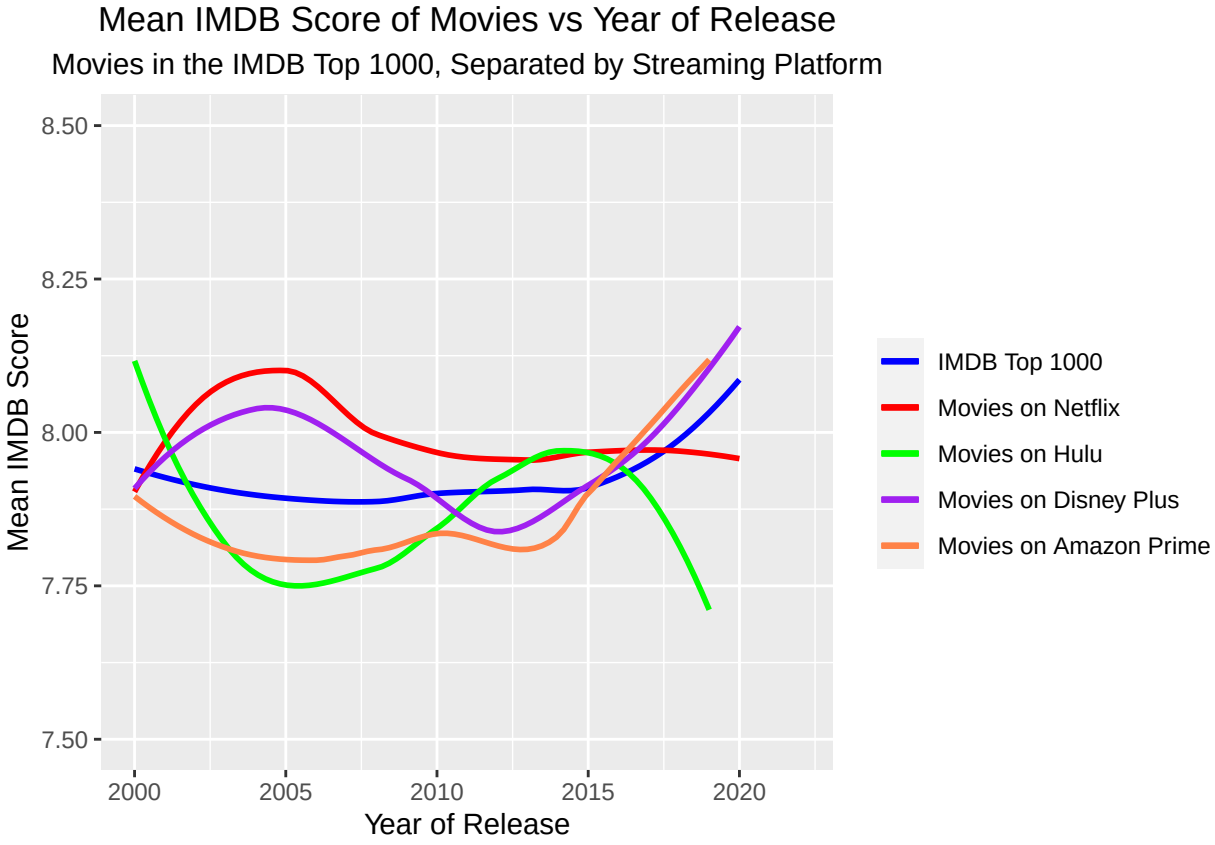
We are using the columns title, type, and release_year.

A potential issue with the data is that it is only of English movies, and we are using the top 1000 movies by gross revenue from the United States. Though the streaming companies that we are considering are created primarily for American audiences, this could exclude international trends outside of American revenue for English language movies. It also excludes American revenue for non-English language movies.

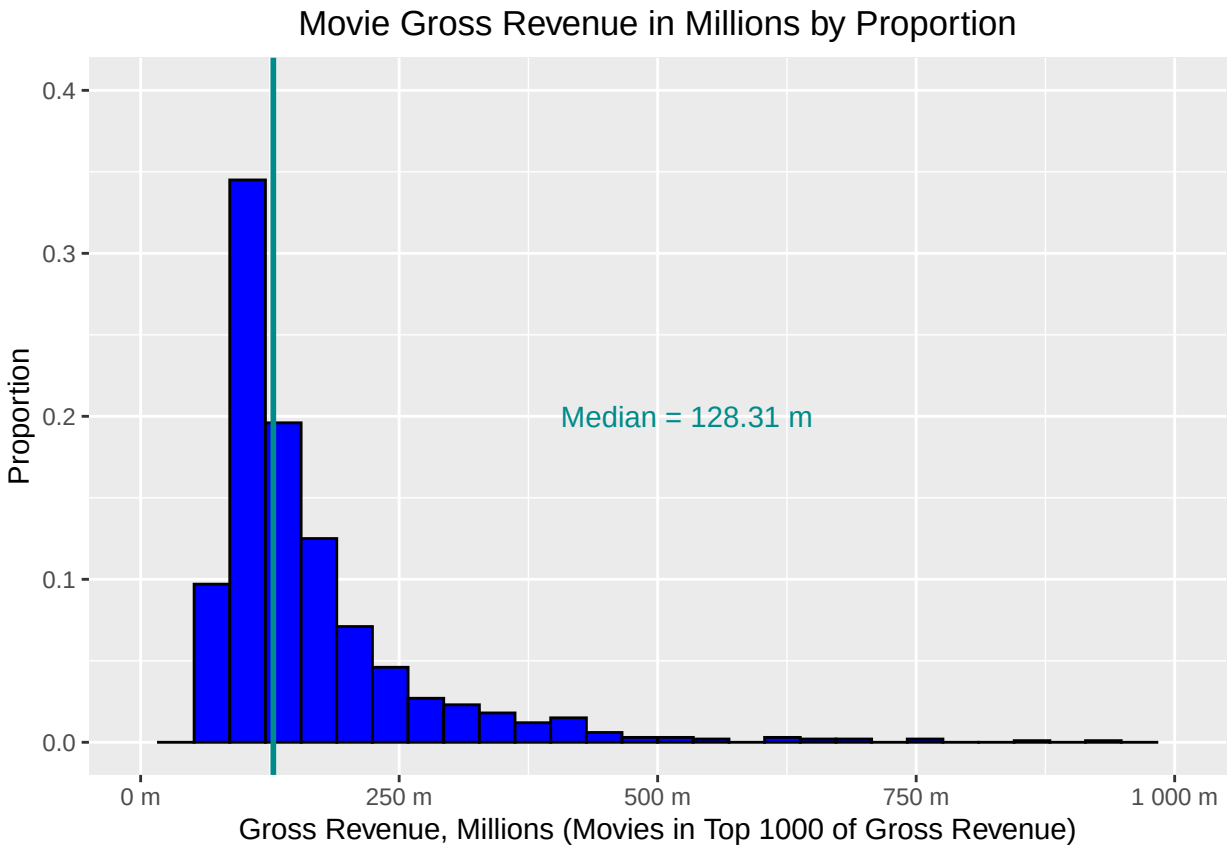
Source:

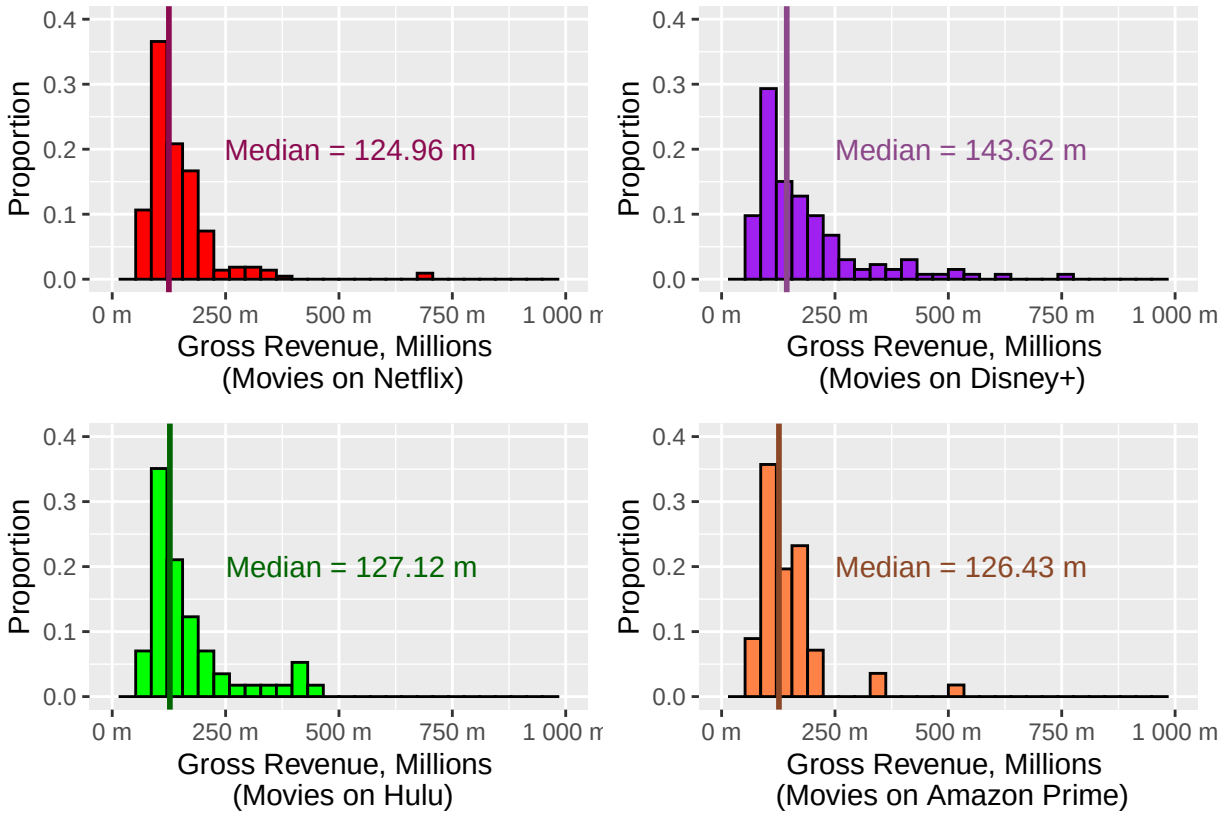
- "Top Box Office Revenue Data - English Movies" on Kaggle by user *Kalilur Rahman*

Summaries



The graph shows that the scores of movies in the IMDB Top 1000 have a general trend upwards the newer that the movies are. This trend is also reflected in the smoothed trend line of the subset of movies available on Amazon Prime and Disney Plus. The subset of movies available on Hulu have a general downward trend in IMDB scores the newer that they are, and the subset of movies available on Netflix have a generally stable IMDB score line.


Movie Gross Revenue in Millions by Proportion
Proportion
0.4
0.3
0.2
0.1
0.0
Median = 128.31 m
0 m
250 m
500 m
750 m
1 000 m
Gross Revenue, Millions (Movies in Top 1000 of Gross Revenue)



From the above histograms of the distribution of gross revenue for movies available on each streaming platform, we can see that the distributions and medians of gross revenues between movies on Hulu, Netflix, and Amazon Prime are roughly the same. It is interesting to note that these medians all fall under the median for all movies in the top 1000 of gross revenue, 128.31 million. Movies available on Disney Plus, however, have a median gross revenue that is around 15 million higher than the median for all movies in the top 1000 by revenue. The distribution of gross revenues for Disney Plus also has a longer tail to the right and more skew to the higher end of gross revenues.

References

1. IMDB Scoring Information https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#.
2. Netflix Movies and TV Shows https://www.kaggle.com/datasets/shivamb/netflix-shows.
3. Amazon Prime Movies and TV Shows https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows.
4. Hulu Movies and TV Shows https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows.
5. Disney+ Movies and TV Shows https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows.
6. Top Box Office Revenue Data - English Movies https://www.kaggle.com/datasets/kalilurrahman/top-box-office-revenue-data-english-movies.
7. Adding Mean to Histograms ggplot2 https://stackoverflow.com/questions/47000494/how-to-add-mean-and-mode-to-ggplot-histogram.