# Movie and TV Show Exploration for Popular Online Streaming Sites

**Final Project For: IST 652 Syracuse University**

**By: Jonathan (Jon) Durbin**
**& Patricia (Patty) Meadows**

# Table of Contents:



## Summary


In this project, datasets for Amazon Prime, Disney+, HBO Max, Hulu, Netflix, and Paramount streaming services were combined into main files in order to explore the Movie and TV Show offerings of the sites as of July 2022 (for Netflix) and May 2022 (for all the other platforms). In the review, instead of monetary data, ratings from The Internet Movie Database (IMDb) and The Movie Database (TMDb) online databases were used to explore the popularity and offerings of the platforms. This project utilized the PathLib, Pandas, CSV, Regular Expressions, and Matplotlib packages to import the data, clean it, explore it, and display the results. Jon led the project in bringing together the raw data, we worked as a team to review the data and compile questions to explore, Patty completed questions 1, 2, 4, and 5 and Jon did the rest, Jon then worked on the Jupyter file cleanup and Patty did the final write up and put together the presentation for class.


## Data Details

Victor Soeiro uploaded two CSV files for each of the platforms on Kaggle. One CSV contained credit data and housed data on the credits of the actors and directors of the offerings per platform. The second CSV file contained title data and housed data on the movie and TV show offerings per platform. In total, 6 files for the credits and 6 files for the titles were combined into two master files. These were then explored in the questions below to better understand what each platform had to offer and for overall online streaming information.

The importing and cleaning occurred at the same time utilizing the PathLib package. First, all the files were downloaded in a “raw” folder on our computers and then a “cleaned”

folder was created. Then regular expressions were established to be used when reading in the data to establish patterns that are acceptable for use in each column and filter out those whose formatting would cause issues later in the project. Next, we iterated through each raw file, passing each file through the regular expressions as well as replacing formatting in each column and adding a column for the platform of the data before saving the cleaned file in the cleaned folder on our desktops. A dictionary was then established containing each of the files for each platform and these were then added to the master files for the credits and the titles. The credits master file contained 366,429 lines of data and the titles master file contained 25,430 lines of data. The findings for each question will be outlined below in the Data Exploration section. For now, we want to explore the processing details for each question.

For question 1, the platforms were grouped in the titles main file and the average of the runtime for the offerings were calculated using dataframe aggregations. This sorted dataframe summary was then graphed in a horizontal bar chart.

For question 2, the credits master file was separated into a dataframe of Actors and Directors using a Boolian sort. An aggregation was then performed to count the number of actors and directors per platform. Finally, a merge of the aggregated data was used to graph a side-by-side bar chart of the data.

For question 3, a loop was used to iterate through the genre's column of the titles main file and add in a category of "no genre" if there was nothing listed and to break apart each record with multiple genres and then count the number of records in each genre and display these counts sorted in descending order.

For question 4, a series of loops and dictionaries were used to iterate through the genre and IMDb and then the TMDb columns of the titles master file to then calculate the average of the ratings per genre and display the results in a descending list.

For question 5, a dataframe was created for each platform containing the type of offering and the release year of the offering. Then a function was created to iterate through each dataframe's type column to count the number of movies and TV shows per platform. These were then put into a dataframe and side-by-side bar charts were used to display the results. Next a function was created to count the release year for each offering per platform in 10 year increments from 1900-2029 (because the data runs from 1901-2022) and place the results in unique dataframe's per platform. These 6 dataframes were then joined on the platform to make one master dataframe of release year counts per platform. This master dataframe was then used to make a stacked bar chart of the counts of offerings per release ear bucketing per platform.

For question 6, histograms of the distribution of the IMDb and TMDb scores from the main titles file were plotted for analysis.

For questions 7, the platforms were again used as a grouping and the average IMDb and TMDb scores were aggregated for each platform. This dataframe was then used to make a side-by-side bar plot to compare the averages per platform and between platforms.

# Data Exploration

## Question 1: What is the Average Runtime per Platform?

In Image 1 we see that Amazon has the largest average run time (>80 minutes) for their offerings whereas Disney+ has the smallest average run time (~60 minutes). This makes sense in regards to the offerings and audiences of the platforms. Disney is tailored to children and families with less attention and time for watching, whereas Amazon is meant for an adult audience and offers more movies than TV shows. If you review the rankings of the platforms in Image 1 they align with the platforms offerings.

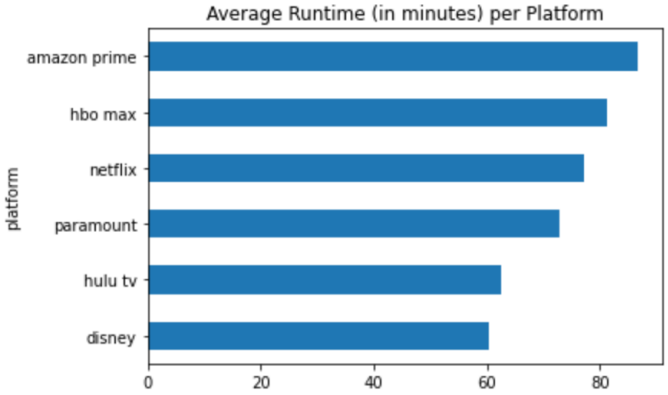


Image 1: Average runtime, in minutes, per platform.

## Question 2: What is the Count of Actors and Directors per Platform?

See Image 2 below for a visual of the findings for this question. We are seeing that Amazon not only has the longest average run time, it also has the greatest variety and count of actors and directors. As Amazon is more generic and tailors to a wider range of audiences this makes sense. It is interesting to see that Hulu, Paramount, and HBO Max have a smaller number of directors as they are just starting to make their own content and are newer platforms. You can see Netflix increasing their director count which aligns with more offerings coming from their platform of late.

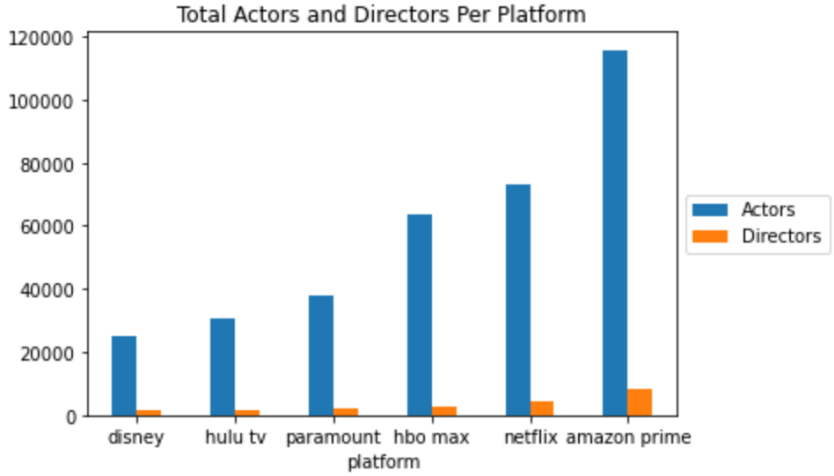


Image 2: Count of Actors and Directors Per Platform Based on Credits

## Question 3: What is the Distribution of Genres?

See Image 3 below for a sorted list of the count of offerings per genre for all the platforms combined. It is interesting to see how much larger the count of drama offerings there are. Also, it is surprising that reality is above the sport genre and catching up to westerns. In further reviews it would be interesting to see how many offerings had a single genre category and how many had several categories applied. Perhaps drama is so large because it is a common genre that is combined with others for movies and TV shows.

```
drama         : 11608
comedy        : 8867
thriller      : 5011
action        : 4833
romance       : 4241
documentation : 3925
crime         : 3433
family        : 2975
scifi         : 2580
animation     : 2506
fantasy       : 2495
horror        : 2193
european      : 1804
music         : 1180
history       : 1090
western       : 855
reality       : 799
war           : 773
sport         : 693
no genre      : 423
```

Image 3: Counts of offerings per genre across all platforms

## Question 4: What is the Average TMDb and IMDb Score per Genre?

See Image 4 below for a sorted list of the average scores per genre of offerings from all the platforms combined. Here it is important to note that IMDb is a privately accessible database offered from Amazon and TMDb is an open-source database for rating the same type of data. From this question we see that the users of the rating platforms show a difference in preference of genre. We are also seeing a bit wider of a spread of the average score for IMDb ratings.

| | | | |
|---|---|---|---|
| animation | : 7.17 | history | : 7.90 |
| reality | : 7.15 | music | : 7.10 |
| documentation | : 6.93 | european | : 7.10 |
| history | : 6.91 | sport | : 7.00 |
| family | : 6.82 | fantasy | : 6.90 |
| fantasy | : 6.79 | comedy | : 6.90 |
| scifi | : 6.65 | family | : 6.90 |
| war | : 6.65 | animation | : 6.90 |
| sport | : 6.63 | scifi | : 6.10 |
| music | : 6.55 | romance | : 6.00 |
| drama | : 6.52 | crime | : 5.10 |
| european | : 6.52 | drama | : 5.10 |
| crime | : 6.49 | thriller | : 5.10 |
| comedy | : 6.46 | horror | : 4.80 |
| action | : 6.43 | reality | : 4.60 |
| romance | : 6.36 | western | : 4.40 |
| thriller | : 6.21 | action | : 4.40 |
| horror | : 5.78 | war | : 4.40 |
| western | : 5.66 | documentation | : 4.10 |

Image 4: average TMDb Scores (left) and IMDb Scores (right) per genre

## Question 5: What is the Distribution of Movies and TV Shows Per Platform?

See Image 5 and 6 below for a review of the count of movies and TV shows per platform as well as the count of offerings based on the release year and bucketed in 10-year increments. We can first see that the findings in earlier questions are supported here with more movies and TV shows being offered in amazon than the other platforms which supports the larger counts of credited actors and directors. Of note is that Hulu has more TV shows than movies which is to be expected with the platforms reputation. In image 6 we see that all of the platforms focus their offerings on things released between 2010-2019. The interesting find here is that the 2020-2029 bucket is already quite full and this data only goes through 2022. Even with the pandemic offerings were being created. This could align with solely digital releases becoming popular.

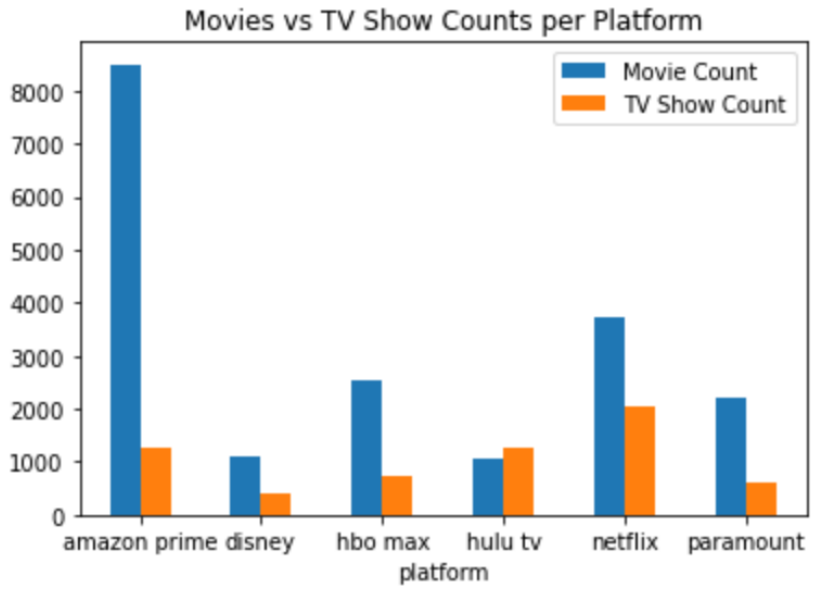


Image 5: Count of movies and TV shows per platform.

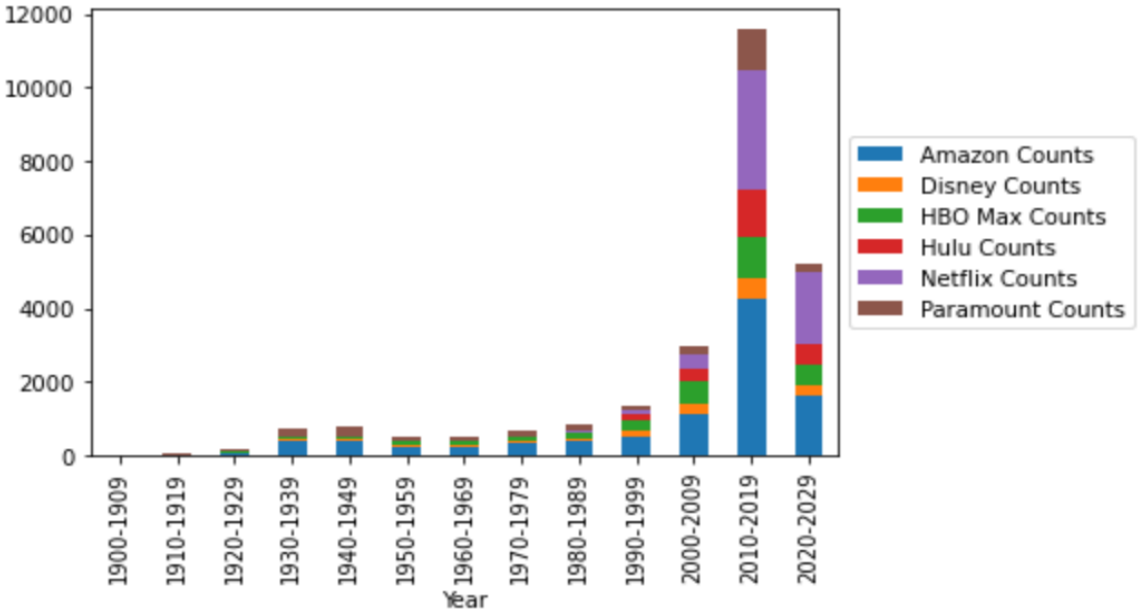


Image 6: Count of offerings per release year for each platform (note this data is from 1901-2022)

## Question 6: What are the TMDb and IMDb Score Distributions?

See Image 7 below for a visual on the ratings of each database. Here we see that each site is spread pretty evenly. However, the TMDb have more reviews and the IMDb have a higher average. With TMDb being more open-source it is not surprising to see a more bell shape as it's presumed to be less monitored/filtered than IMDb. Additionally we see a right skew in both histograms which aligns with the theory that individuals tend to rate higher rather than lower and tend to rate somewhere in the middle of the scale.

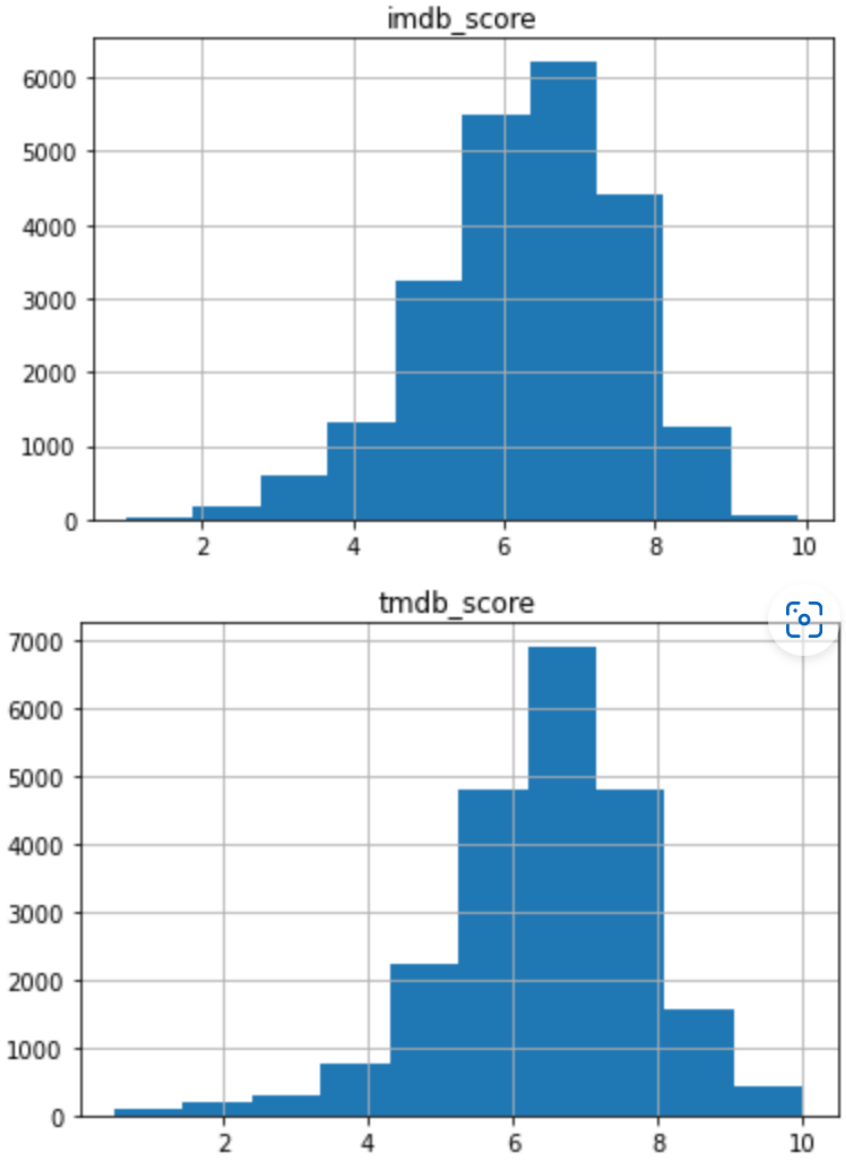


Image 7: Histogram of the IMDb and TMDb scores for all platforms

## Question 7: What are the Average TMDb and IMDb Scores Per Platform?

See Image 8 below for a comparison of the average of the ratings of offerings per platform. In this analysis it is interesting to note that Amazon and Paramount have the lower averages and are nearly equal between rating sites. Disney, Hulu, and Netflix are all rated higher on average by TMDb users which makes sense as they are more popular and familiar to most people. HBO Max is the only platform where the IMDb rating is above the TMDb rating and that also makes sense as HBO Max is not a mainstream platform.

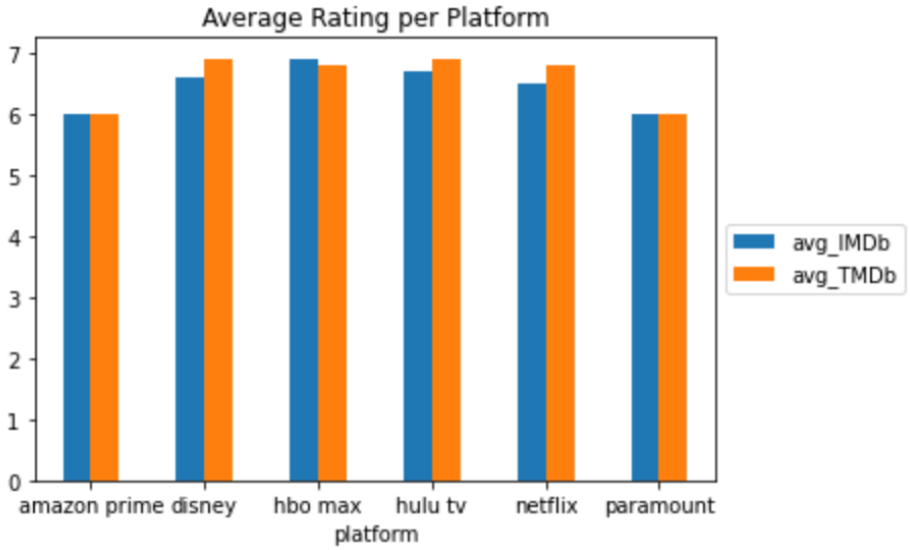


Image 8: Average of IMDb and TMDb ratings per platform

## Conclusions

This project was a success at applying many principals taught in IST 652. We were able to read in and combine 12 files for 6 online streaming platforms into 2 master files to explore similarities and differences in the platforms. We were able to compare and contrast the platforms in regard to their offering's runtimes, variety of performers, distribution of genres, distribution of movies and TV shows, and the IMDb and TMDb ratings of the offerings available.

To take this analysis further, sentiment analysis could have been utilized on the movie descriptions to see if there was commonality in offerings across platforms. More analysis could have been done utilizing the ratings of the offerings to see about the appropriateness of the platform for audiences. It would have been beneficial to see more about the budgets and earnings of the offerings to explore how movies and TV shows are chosen per platform. It also would have been interesting to see streaming data for each platform to look into actual usage of offerings over the popularity of the offerings.

Overall, this project was a great demonstration of both of our abilities and gave each of us an opportunity to apply concepts learned. Jon having more Python experience was able to combine the files seamlessly and Patty having no Python experience was able to understand and soak in these steps and then apply aggregations, create functions, and join datasets to explore the master data and use graphing to display results.