Bryan D'Amico, Sintia Stabel, Matthew Smith
IST652-Scripting for Data Analysis
Final Project Report

What Streaming Service is Best for Me?

Introduction

When Netflix first switched from mailing DVDs to having an on-demand streaming platform, that business model was seen as the solution to cable television. It was so popular, however, that it seems that every network is starting their own streaming service. These days, with so many streaming services, there is a temptation to subscribe to them all, making the solution to cable just as expensive. Each service tries to brand itself as being essential to have, but is this really true for everyone? If people were better informed about how the content on each service aligns with their personal interests, they would be able to save money by only choosing the ones that would bring them the most value.

This report is intended to provide a detailed analysis of several major streaming services to help a reader answer the question "What streaming service is best for me?" for themselves.

About the Data

The data was drawn from csv files of information for Amazon Prime, Disney+, HBOMax, Hulu, Netflix, and Paramount+, which were all found on Kaggle. These csv files were uploaded by Victor Soeiro in 2022 and contain data collected from the JustWatch database. The data reflects what was available on each service in the United States around the midpoint of 2022.

- Amazon Prime data set: https://www.kaggle.com/datasets/victorsoeiro/amazon-prime-tv-shows-and-movies
- Disney+ data set: https://www.kaggle.com/datasets/victorsoeiro/disney-tv-shows-and-movies
- HBOMax data set: https://www.kaggle.com/datasets/victorsoeiro/hbo-max-tv-shows-and-movies
- Hulu data set: https://www.kaggle.com/datasets/victorsoeiro/hulu-tv-shows-and-movies
- Netflix data set: https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies
- Paramount+ data set: https://www.kaggle.com/datasets/victorsoeiro/paramount-tv-shows-and-movies
- JustWatch database: https://www.justwatch.com/us

Each dataset is set up in a similar fashion so that they can be concatenated into a larger data frame. They contain 15 columns of information on each movie or TV show available in the U.S. on each streaming service. The following table summarizes the data found in the data sets as they were downloaded for the analysis. All variable definitions were taken from the Kaggle dataset page.

Variable Name	Description
id	The title ID on JustWatch.
title	The name of the title.
show_type	Indicates whether the title is a TV show or movie.
description	A brief text description of the program.
release_year	The release year.
age_certification	The age certification such as G, PG, TV14, R, etc.
runtime	The length of the episode for a show, or the length of the movie.
genres	A list of the program's genres.
production_countries	A list of countries that produced the title.
seasons	The number of seasons if the program is a show.
imdb_id	The title ID on IMDB.
imdb_score	The score on IMDB.
imdb_votes	The number of votes for the score on IMDB.
tmdb_popularity	Popularity on TMDB. (https://developers.themoviedb.org/3/getting-started/popularity)
tmdb_score	Score on TMDB.

Data Cleaning and Exploratory Data Analysis

Importing the Data and Creating the Data Set

The first task was to combine the six data sets into a single data frame so all the data could be worked with at once. To begin, each of the data sets were imported as separate data frames and named according to the streaming service each represents. Within each data frame there was no way to tell which titles belonged to which service, so before joining the data frames together, a "service" variable was added to each data set containing the name of the service for every title. The columns were also reordered so that this new "service" column was on the left side of the data frame, so each title's streaming platform is easily identifiable when viewing the data frame. Next, all six of the data frames were concatenated vertically to create a single large data frame containing all the available data. The index for the data frame was reset so that it is sequential throughout the entire list of titles and then the number of titles for each streaming service was compared to the length of that individual service's data frame to ensure that the concatenation worked as intended. With the data frame accurately created, it was saved to an external csv file to permanently preserve the changes.

Cleaning the Data Set

The newly created data set containing all the titles from all six streaming services contains 16 columns of information on a total of 25,773 titles. This data frame was first checked for missing values of which there were many. The following output shows how many missing values were present in each column.

The row with the missing value for title contained missing values for almost all columns so this row was dropped from the data set. There were a mix of movies and TV shows that were missing descriptions. From personal experience with streaming platforms, it was known that sometimes programs simply do not have a description published so these missing values were left as missing. There were almost 13,000 titles that did not have age certification rating. Since they do not have an age certification rating available, they can be treated as though they did not receive a rating. All these missing values were replaced with the string “Not Rated”. To investigate the titles with missing values for the number of seasons, the data frame was filtered for when the seasons value was missing and then a unique list of the entries in the “type” variable was printed. This showed that the only types of programs without a seasons value were movies. Since movies do not have seasons like TV shows do, these values were left as missing. Lastly, the missing values related to the IMDB and TMDB databases were also left as missing. Since there was no data for these titles on those sites, it did not make sense to try to replace them with particular values. These titles will just be left out of any analyses focused on the information contained within those databases.

```
id                        0
service                   0
title                     1
type                      0
description             181
release_year              0
age_certification       12847
runtime                   0
genres                    0
production_countries      0
seasons                 19183
imdb_id                 2092
imdb_score              2680
imdb_votes              2721
tmdb_popularity          789
tmdb_score              3418
dtype: int64
```

The next task involved checking the data types of the columns in the data frame and then making appropriate changes so that the data type of each column would support the types of analyses that would be run. All the columns containing numeric values already had accurate data types, containing either integer or float values. The columns such as service, title, type, and description that contained purely text data all had the object data type. These were manually converted to the string data type. The genres column and the production_countries column contained lists of values, so those columns were not changed.

```
id                      object
service                 object
title                   object
type                    object
description             object
release_year             int64
age_certification       object
runtime                  int64
genres                  object
production_countries    object
seasons                float64
imdb_id                 object
imdb_score             float64
imdb_votes             float64
tmdb_popularity        float64
tmdb_score             float64
dtype: object
```

The genres column was the next issue that needed to be addressed with the data frame. Since it contained a list of string values denoting which genres described each title, it would not be possible to aggregate the data by genre without altering the data frame. To solve this issue, a custom function was created. The function split the lists in the genres variable by stripping the square brackets and spaces out of the values and then splitting that resulting string into a list of genres for each title. The function then iterates over that list, adding a new column to the data frame for each possible genre.

genre_0	genre_1	genre_10	genre_2	genre_3	genre_4	genre_5	genre_6	genre_7	genre_8	genre_9
'music'	'comedy'	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
'war'	'comedy'	NaN	'drama'	NaN	NaN	NaN	NaN	NaN	NaN	NaN
'comedy'	'family'	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
'drama'	'comedy'	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

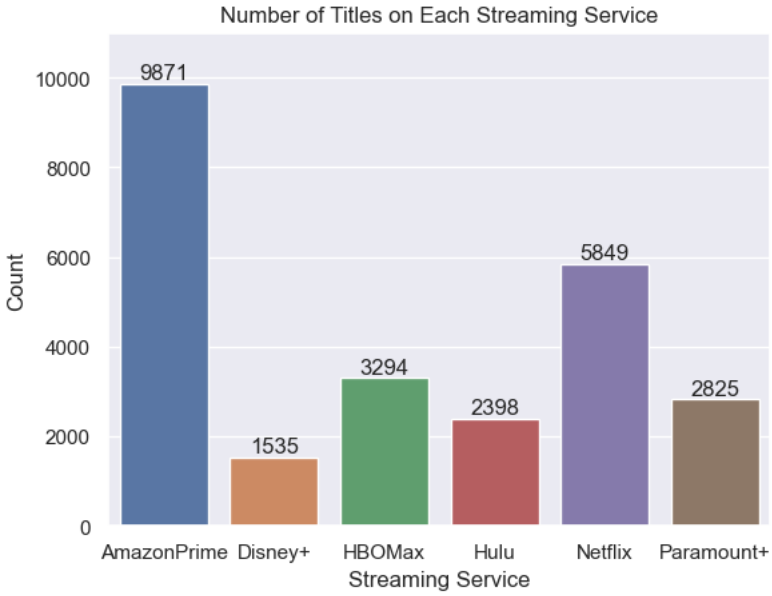
This allows each title to be included in each genre group it should belong to. So, if a show is a comedy intended for family audiences, it will be present in both those groups, just like how when browsing by genres on a streaming service one can find the same title in multiple genre pages.

Exploratory Data Analysis

To give the reader a good sense of the information contained within the data frame, the distributions of several of the important variables were visualized. While the following visualizations do not answer any of the specific questions of interest regarding this data, they could still help someone make a decision about what streaming service would be best for them.

Distribution of Titles Across Streaming Services

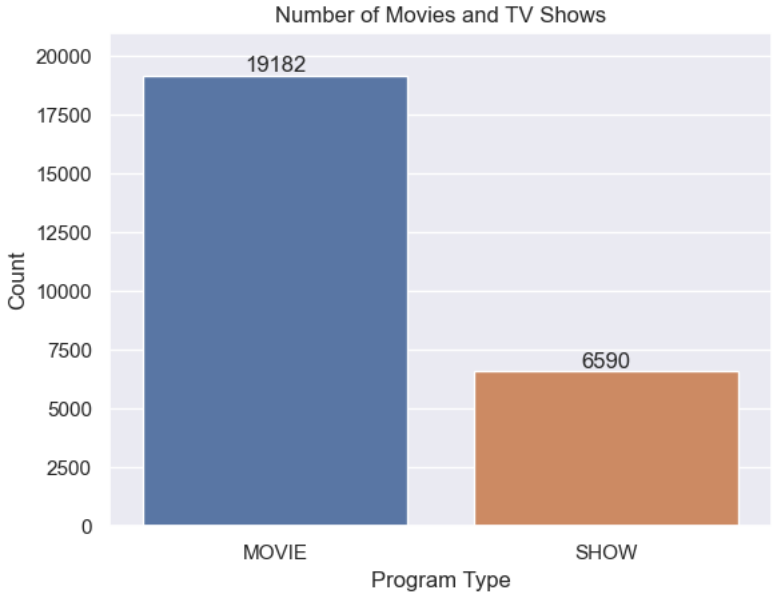
The following bar plot shows the number of titles belonging to each streaming service in the data frame.



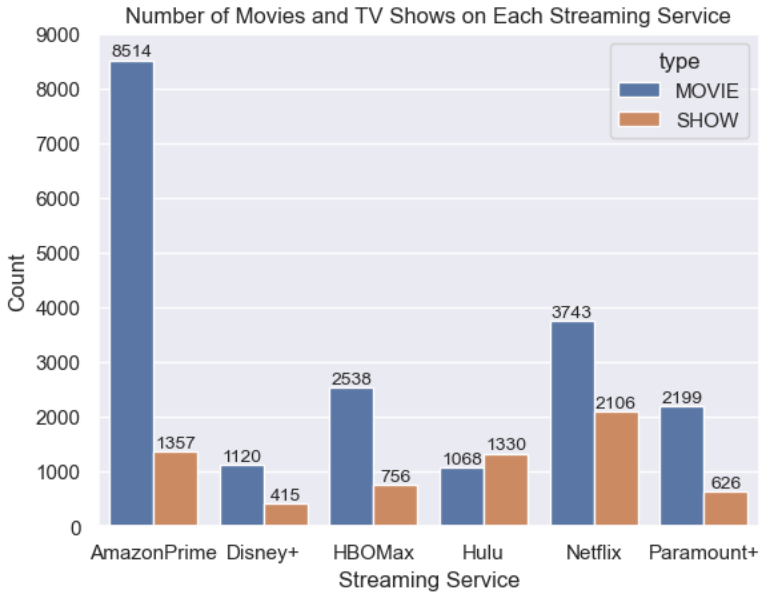
There is a large disparity in the number of titles available on each streaming service. This is due to many factors, including how long the service has been around for and how large of a budget each has for acquiring more titles. Amazon Prime has the most titles by far, having approximately 4,000 more than Netflix, the service with the next largest library. Disney+ has the smallest selection with just over 1,500 titles.

Distribution of TV Shows vs. Movies

The following bar plot shows the number of movies and TV shows available across all the streaming services in the data frame.



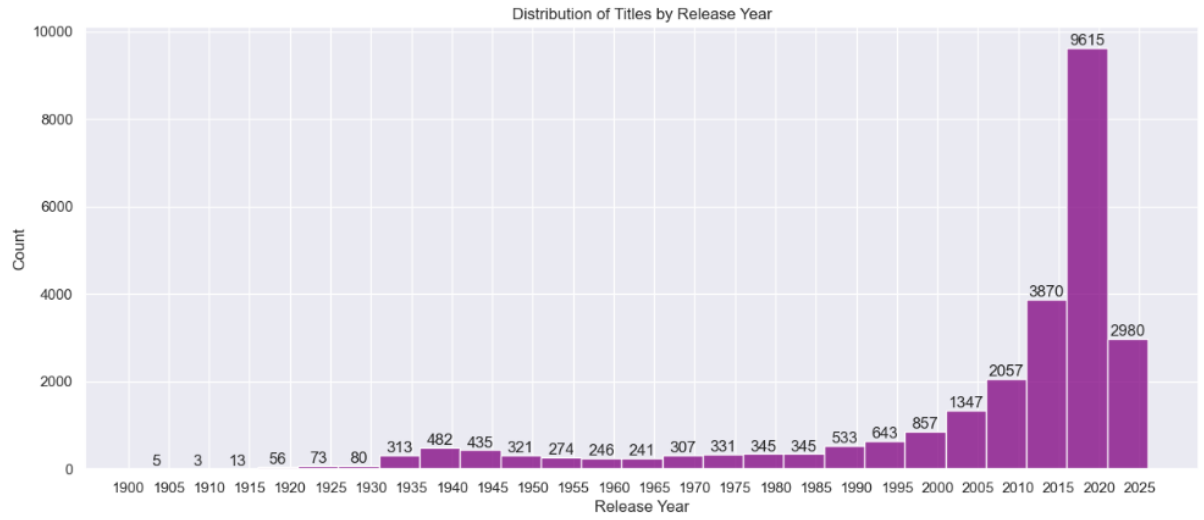
Among all the streaming services, there are about three times as many movies as shows. Next, the data frame will be further broken down by each streaming service to see how the number of movies and shows compare in each.



All but one of the streaming services have more movies than shows. Hulu has just under 300 more shows than it does movies. This means Amazon Prime, Disney+, HBOMax, Netflix, and Paramount+ have a similar makeup where they are more focused on providing movie content to viewers. Hulu may be intentionally providing more shows than movies to differentiate itself from the rest of the field.

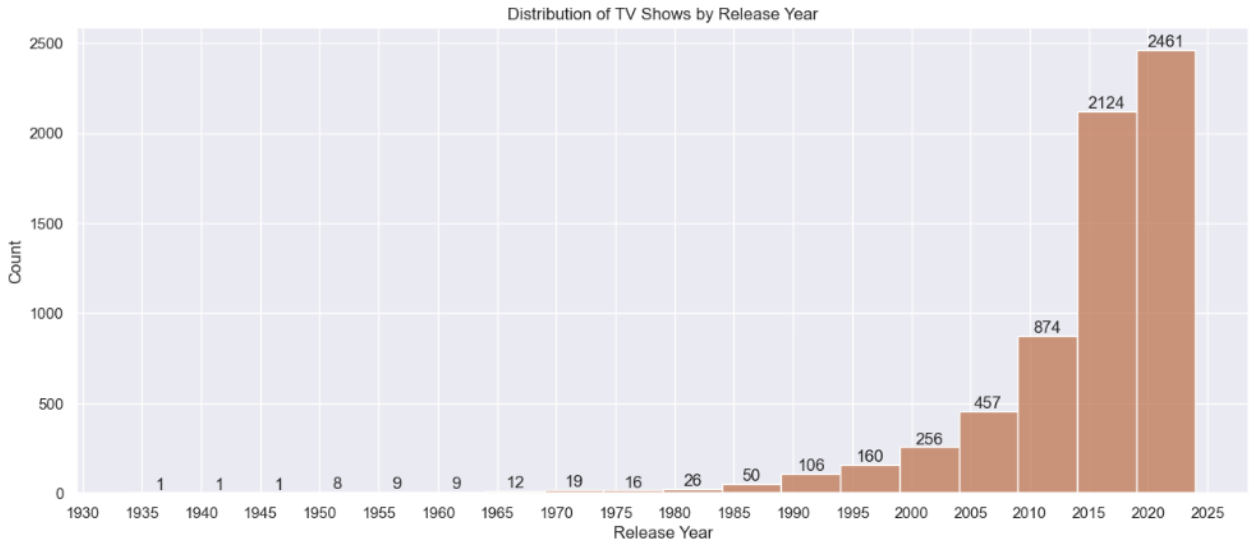
Distribution of Release Years

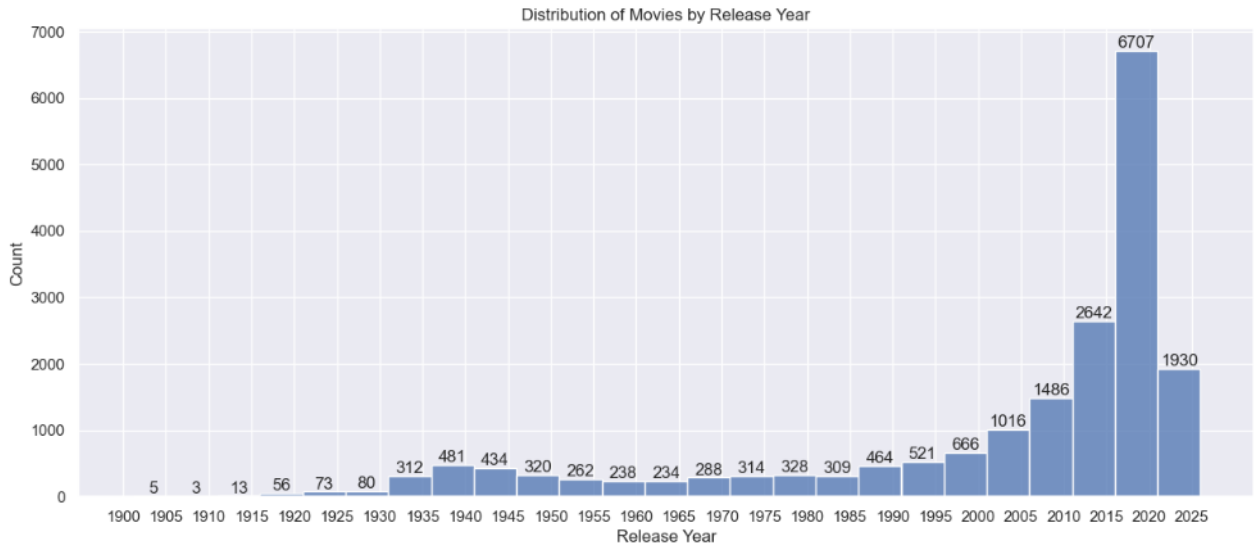
The distribution of the number of titles by their release year was visualized by placing the titles into 5-year bins while creating a histogram. The first histogram includes all the titles across all the streaming services.



While there is a good amount of old content available to be streamed, the great majority of the content is from the last decade or so. There are significantly more titles released after 2010 than from all the years before then.

Next, the distributions will be separated by their type to show the release years of the TV shows and movies separately.



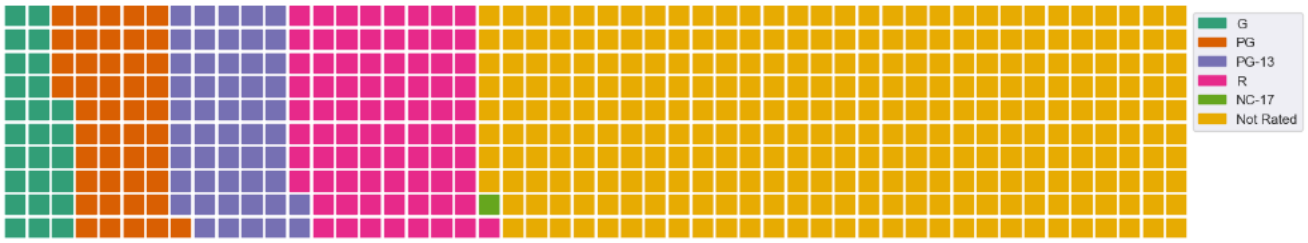


The distributions of both the TV shows and movies follow the same pattern as the overall distribution. While there are very few TV shows available that predate the 1990s, there are more old movies available. There are hundreds of movies available for streaming that date all the way back to the 1930s. The difference in these distributions is probably due to the fact that movies as an entertainment medium began well before TV shows.

Distribution of Age Certifications

The age certification values will certainly be of interest to parents who are choosing a streaming service that their children will also have access to. Since there are different age certifications for movies and TV shows, the two types of programs are visualized separately.

The following waffle chart shows the proportions of the total number of movies with each age certification.



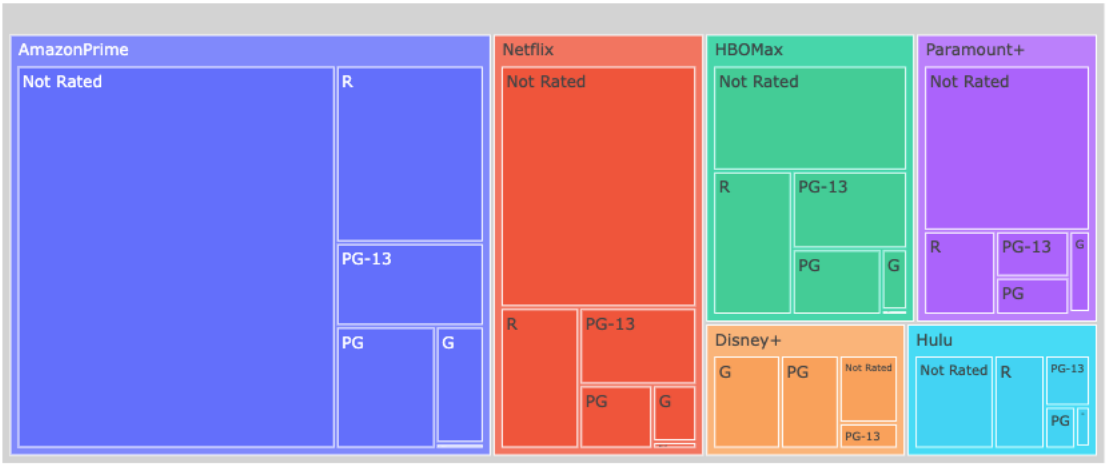
The majority great majority of movies available on these streaming services do not have a rating. There are very few NC-17 rated movies. Out of the movies that are rated, the majority are R and there are fairly equal numbers of PG and PG-13 movies.

The next waffle chart shows the proportions of the total number of TV shows with each age certification.

Most TV shows are for older audiences, rated either TV-14 or TV-MA. There are still a substantial number of TV shows that do not have ratings, but proportionally much less than for the movies. The number of TV shows rated as suitable for children (TV-G, TV-PG, TV-Y, and TV-Y7) is about the same as the number of shows rated TV-MA.

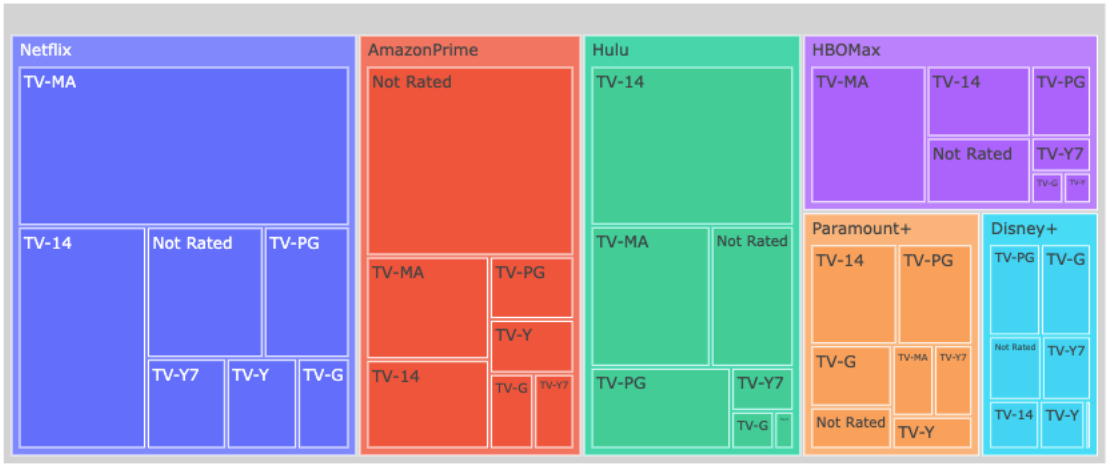
Since the age certification makeup could vary greatly by streaming service, the movie and TV show data will be grouped by service and age certification before finding the counts of each category. This data was visualized within the program using interactive treemaps and screenshots of those will be provided in this report.

Number of Movies by Age Certification for Each Streaming Service



Most movies do not have a rating for each of the streaming services, except for Disney+. Keeping with their family friendly image, the majority of movies are rated either G or PG, with no movies rated R.

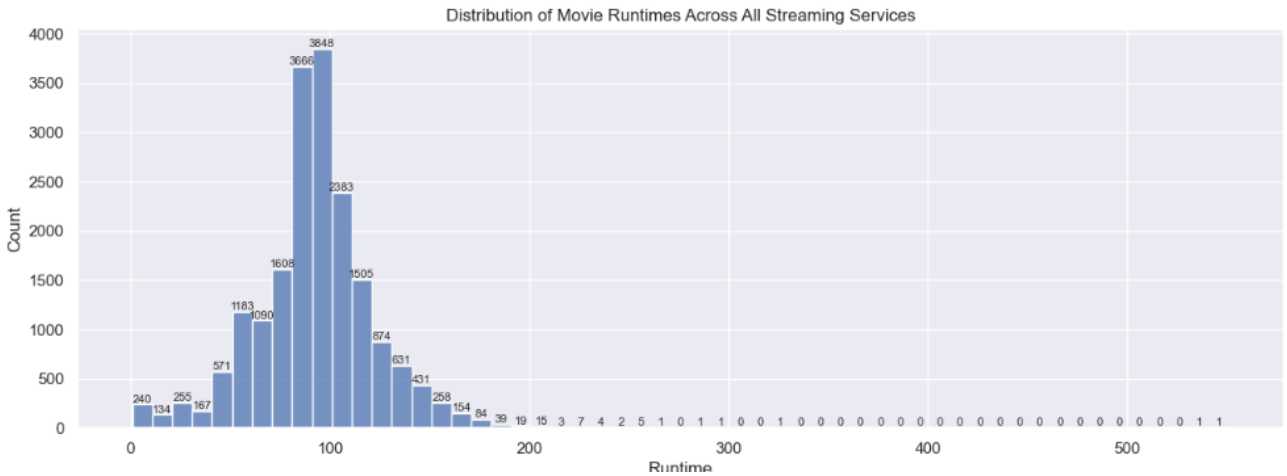
Number of TV Shows by Age Certification for Each Streaming Service



Netflix and HBOMax have the greatest proportion of their TV shows rated for mature audiences. Hulu and Paramount+ have most of their shows rated TV-14, while Amazon Prime contains mostly shows that are not rated. Similar to the age certifications for the movies, Disney+ has mostly family friendly shows with a small segment of them rated TV-14 and no shows rated TV-MA.

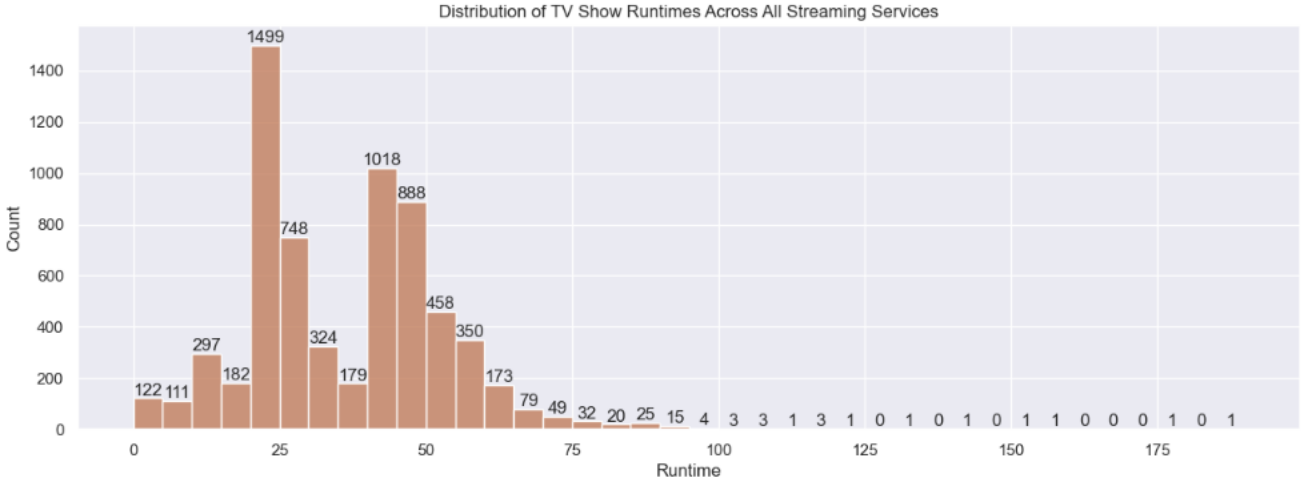
Distribution of Runtimes

Since not everyone has the same amount of time to watch TV or movies during the day, a reader might be interested in how long the TV shows and movies are. The movie data is visualized using a histogram featuring 10-minute bins.



The majority of movies are about 80 to 110 minutes in length. There are a couple hundred movies that are very short and a few movies that are very long, with two movies longer than 500 minutes.

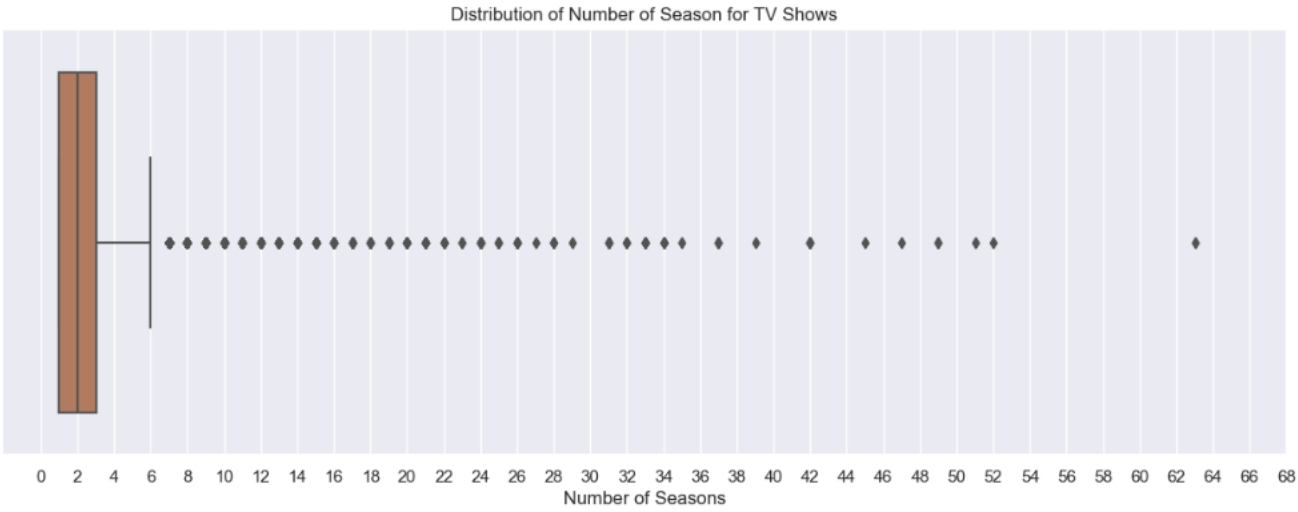
The TV show data is now visualized using a histogram featuring 5-minute bins. The decreased width of the bins reflects how TV shows have shorter runtimes than movies.



The distribution of show times has some bimodal characteristics, which makes sense since most U.S. shows are made to fit in either 30-minute or 60-minute blocks including ads. Most TV shows are either around 25 minutes or 45 to 55 minutes in length.

Distribution of Number of Seasons for TV Shows

In a similar fashion to runtimes, someone without a lot of time to watch streaming content might want to know how much of their time they may have to dedicate to completing any random series before committing to subscribing to a streaming service.



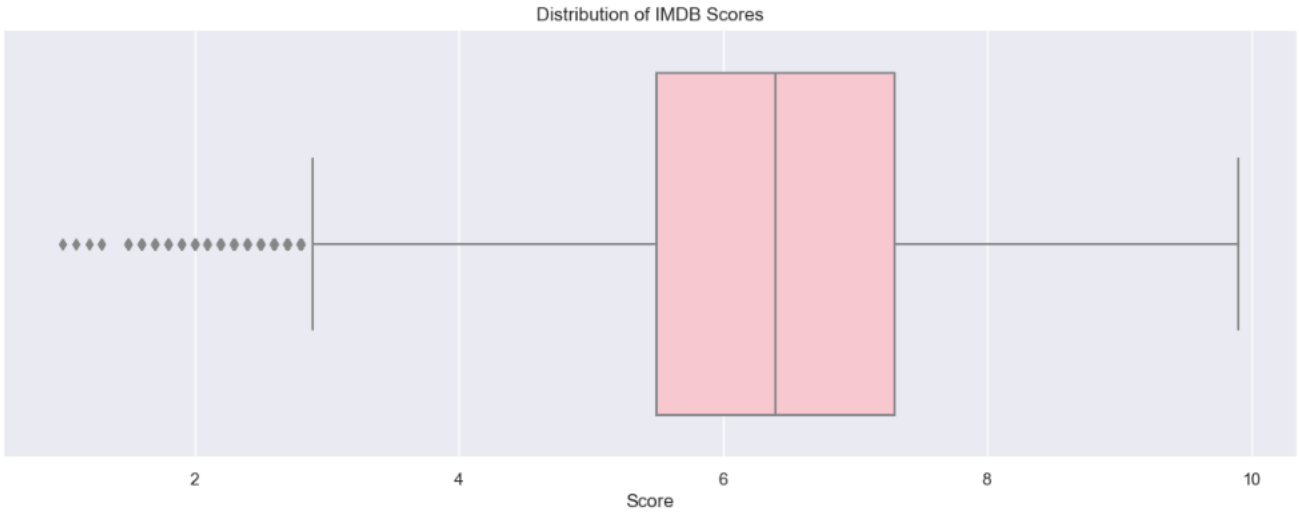
The following box plot shows the distribution of the number of seasons for all the TV shows across all the streaming services.

The boxplot reveals that more than 50% of the TV shows do not last for more than two seasons. It also shows the outliers, which would be a collection of very successful shows that have run for an extremely long time.

Distributions of IMDB and TMDB Scores

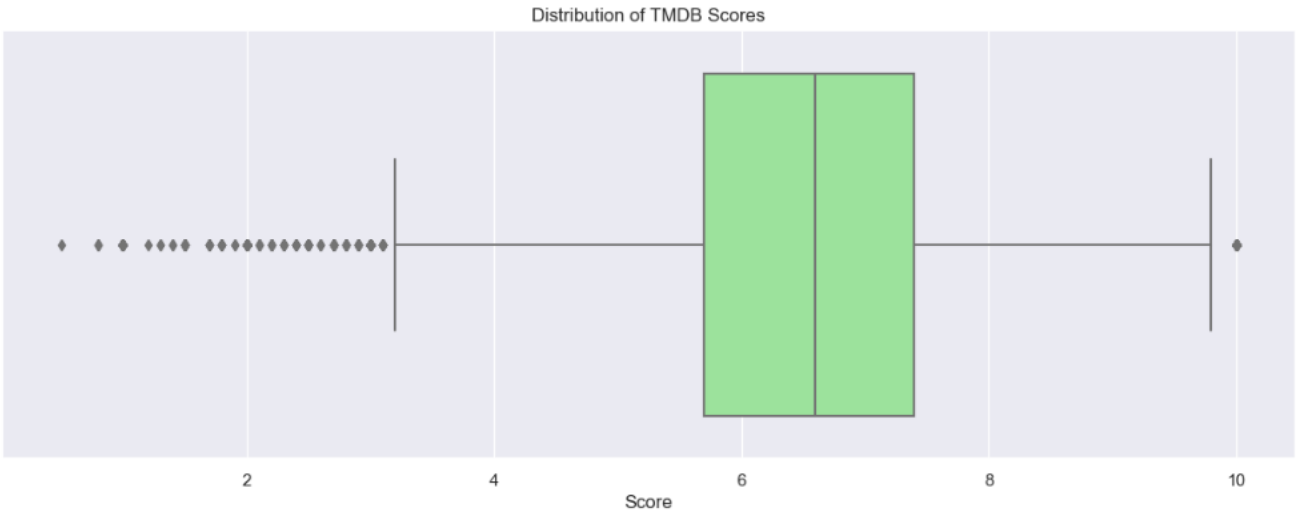
To understand if there are any innate differences between the ratings on the two movie databases, the distributions will be visualized.

The first box plot visualizes the distribution of IMDB scores.



The distribution of IMDB scores across all the streaming content is centered around the 6 to 7 score range, with the middle 50% of the data between scores of 5 and 8. This seems likely as most titles would probably settle around an "average" rating, which in the U.S. is likely a 6 or 7.

The second box plot visualizes the distribution of TMDB scores.



The distribution of the TMDB scores is extremely similar to the distribution of IMDB scores, with the middle 50% of the data again lying between scores of 5 and 8.

Questions

Five questions were developed and answered based on the available streaming data. The answers to these questions are intended to help a reader understand what streaming service is best for them.

Question 1: What service has the genres of movies and shows that a potential subscriber might be interested in?

Unit of Analysis: Genre and Streaming Service

Descriptive Statistics/Comparison/Visualization: For each streaming service, the number of titles available in each genre was calculated. This was used to create visualizations to compare the different streaming services based on the variety of content available on each. The proportion of each genre on each streaming service and the proportion of each streaming service's titles making up one genre were also calculated to create stacked proportion bar plots and a heatmap.

Question 2: What service has the highest rated movies and TV shows?

Part 1

Unit of Analysis: IMDB votes, Streaming Service, and Type

Summarization: The five movies and TV shows with the highest IMDB scores on each streaming service are displayed to give the reader a sense of what is the best each service has to offer. These titles all received more than 7,800 votes on IMDB placing them in the top 25% of the distribution of IMDB votes. These were deemed to be "high profile" shows for the services due to their popularity.

Part 2

Unit of Analysis: Type and Streaming Service

Aggregation/Comparison/Visualization: The mean IMDB and TMDB scores across all titles were calculated for each group of movies and TV shows on each of the streaming services. These were visualized using bar plots and compared to determine the best scoring streaming service on average.

Part 3

Unit of Analysis: IMDB Votes, Type, and Streaming Service

Aggregation/Comparison/Visualization: The mean IMDB and TMDB scores across all titles with IMDB votes greater than 7,800 were calculated and visualized for each group of movies and TV shows on each of the streaming services.

Part 4

Unit of Analysis: Type, Streaming Service, and Primary Genre

Aggregation/Comparison/Visualization: The mean IMDB scores were calculated for each combination of streaming service, movie or TV show, and primary genre. The primary genre of a title is the genre that was listed first in the original genre list variable. These were visualized using side-by-side bar plots and heatmaps to determine which streaming service has the best titles in any particular genre.

Question 3: Which streaming services are improving or declining over time?

Part 1

Unit of Analysis: Release Year and Streaming Service

Summarization/Descriptive Statistics/Comparison: The number of titles on each streaming service are displayed for each release year after the streaming service started. This is intended to show if a service has been working on acquiring newly released programs to continue expanding its library of content.

Part 2

Unit of Analysis: Release Year, Type, and Streaming Service

Summarization/Aggregation/Visualization/Comparison: The mean IMDB and TMDB scores were calculated for each combination of service, movie or TV show, and release year. The mean IMDB scores were visualized using a line plot, showing the changes in scores for both movies and TV shows by release year for each streaming service.

Part 3

Unit of Analysis: Release Year, Streaming Service

Descriptive Statistics/Aggregation/Visualization/Comparison: The mean IMDB scores for all the titles by release year on each streaming service are visualized using line plots. These plots also feature a 10-year rolling average overlaid on the raw data to give a smoother look at the trends by release year on each service.

Question 4: Is a program's score on IMDB related to the number of votes it receives?

Unit of Analysis: Streaming Service

Correlation/Comparison/Regression: A regression model was created for all titles and all titles within each streaming service to reveal the relationship between a title's IMDB score and the number of votes it received. The Pearson correlation coefficient was also calculated for the relationship between IMDB score and votes for each streaming service.

Question 5: Can age certifications be determined using the TV show or movie descriptions or genres?

Unit of Analysis: Tokenized text

Association Rule Mining: Apriori association rule mining was used to determine if the words in tokenized versions of the descriptions or genres would be strongly linked to particular age certifications.

Description of the Program

The program was created in the form of a report style Jupyter notebook. It guides the reader through the data importing, cleaning, and exploration steps previously described. The program then answers each of the five questions of interest by creating new data frames featuring slices, aggregations, groupings, descriptive statistics, and summarizations that are used to visualize the data and compare the different streaming services. The number of titles with each genre tag are counted and separated by streaming service to visualize the make-up of each streaming service. These counts are converted to percentages to compare the prevalence of each genre in each service's library as well as how much of all the available titles in each genre each streaming service contains. The top five high profile movies and TV shows are revealed for each service. All the information for each of five titles is printed in the form of a data frame. Then the mean IMDB and TMDB scores are calculated for each group of movies and TV shows on each of the streaming services. This analysis is run a second time, focusing only on the high profile movies and TV shows. The data set is then further broken down by groups of the movies' or TV shows' primary genre on each streaming service. The next section of the program displays how many titles released after the inception of each streaming service have been added to the corresponding service. The mean IMDB and TMDB scores are calculated by streaming service, program type, and release year to uncover whether the most recently available titles on each streaming service are seen as better or worse than the older content. This analysis also used a 10-year rolling average to smooth out the changes over time and get a better picture of how the scores are trending. Next, the relationship between the number of votes a title receives on IMDB and its score was investigated. Regression models were created for all the titles and for all the titles on each streaming service, and correlation coefficients were created for each of those groups as well. Lastly, the age certifications were analyzed according to their corresponding descriptions and genre tags. The descriptions were tokenized to see if there were any relevant key words. The tokenized descriptions and genre tags were related to the age certifications through apriori association rule mining to find any strong relationships.

The program concludes with a discussion of how this information could be used through a few brief case studies of possible subscriber profiles.

Output and Analysis

Question 1: What service has the genres of movies and shows that a potential subscriber might be interested in?

Different streaming services have varying selections of genres of movies and shows available. Some services may have a larger selection of certain genres, such as action, romance, or horror, while others may focus more on documentaries or romantic films.

Researching the offerings of different streaming services and comparing them with the potential subscriber's preferences can help determine which service has the genres of movies and shows that may be of interest.

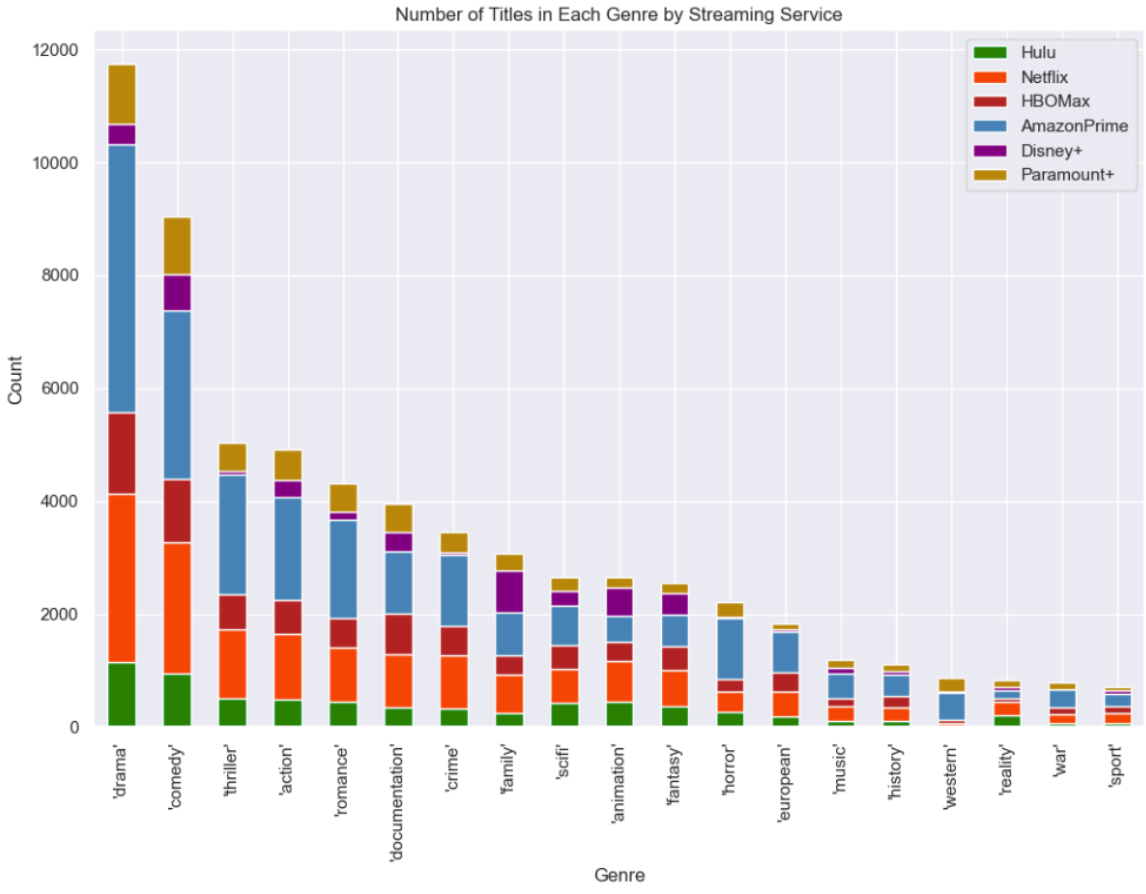
Since the individual genres were previously split across several columns in the data cleaning steps, they can be summarized by streaming service to help determine which has the most variety of movies and shows based on genres. With this knowledge the user can select a service based on genre preference.

A custom function was created to count the number of titles in each genre. It initializes an empty dictionary and an empty set. The function iterates over each row of the input data frame, extracts the service and genres columns, and adds each genre to the initialized set. For each genre and service combination, the function checks if the combination already exists in the dictionary. If it does, the function increments the count for the combination. If it does not, the function creates a new entry in the dictionary with a count of 1. If an error warning is raised when attempting to access the dictionary for a given genre, the function creates a new entry in the dictionary for the genre and service combination with a count of 1.

After looping over all rows, the function loops over each genre in the set and checks if the genre exists in the dictionary. The resulting dictionary contains counts for each genre and service provider in the input data frame. This dictionary was converted to a data frame so the information could be more easily accessed. The data frame to the right displays the number of titles in each genre on each streaming service.

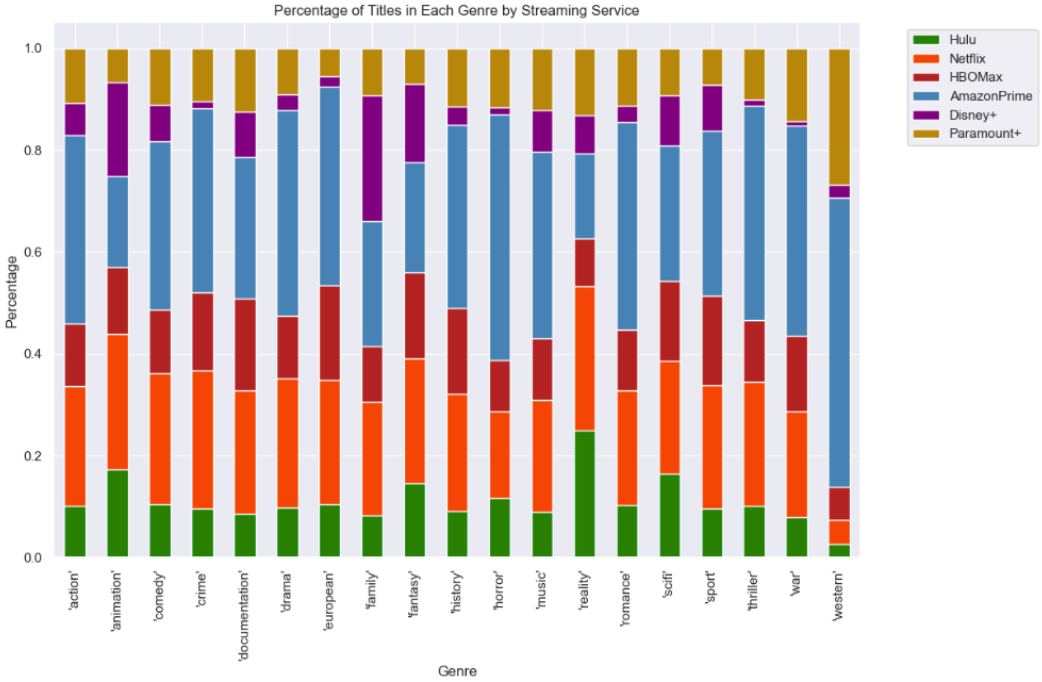
This data frame was visualized using a stacked bar plot. Each color band in each bar shows the number of titles in that genre that are available on a particular streaming service.

	Hulu	Netflix	HBOMax	AmazonPrime	Disney+	Paramount+
'drama'	1156	2968	1443	4764	347	1072
'comedy'	943	2325	1120	2987	649	1007
'thriller'	509	1228	612	2119	57	511
'action'	498	1157	598	1820	304	533
'romance'	441	971	514	1752	138	490
'documentation'	340	952	716	1096	351	495
'crime'	332	936	532	1251	47	359
'family'	255	682	336	751	754	286
'scifi'	438	589	416	705	265	244
'animation'	455	705	345	475	486	176
'fantasy'	373	630	431	554	392	181
'horror'	258	378	221	1065	29	259
'european'	191	443	339	712	35	101
'music'	107	262	145	438	98	146
'history'	101	254	187	396	41	126
'western'	23	41	55	490	23	231
'reality'	207	234	77	138	62	109
'war'	62	163	116	324	6	113
'sport'	68	170	123	228	63	51



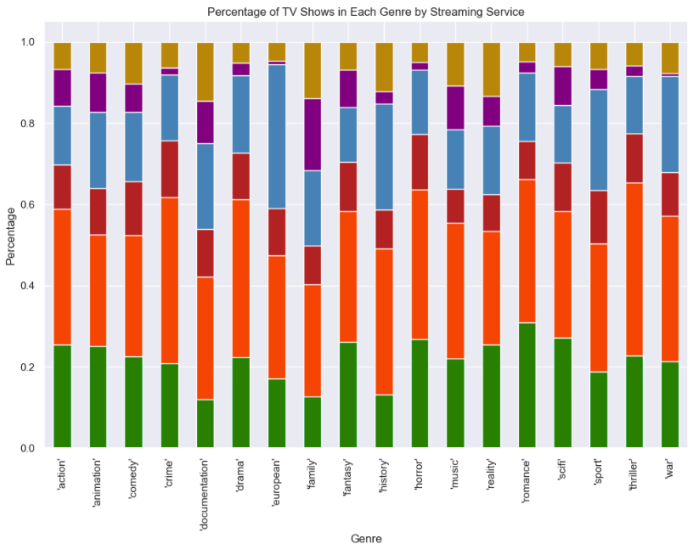
Drama and comedy are the two most frequent genres, with much higher counts than all the rest. Looking at the color bands they follow a fairly proportional distribution throughout each of the columns. This may be due to Amazon Prime having more titles than the other streaming services, so it is likely to have the highest count of titles in each genre. There are a few exceptions to this, however. Disney+ makes up one of the bigger chunks for the family genre even though it is a much smaller portion of each other bar.

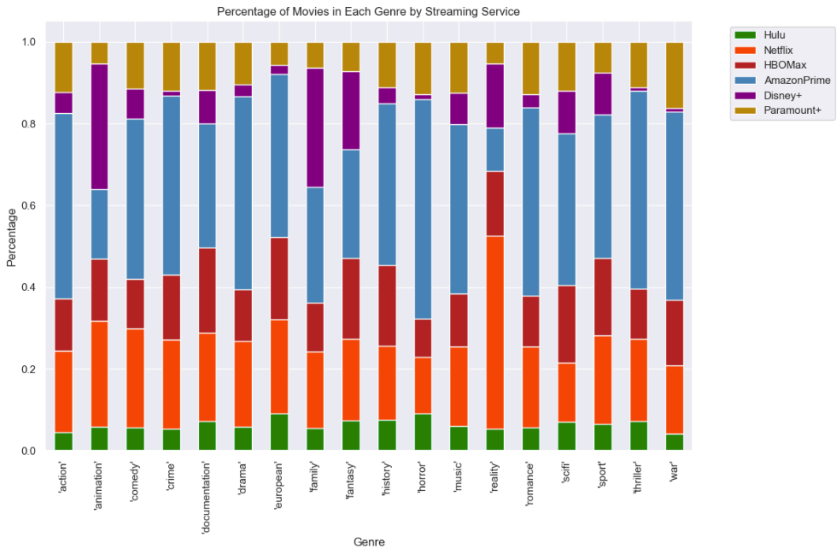
Since there are varying numbers of titles in each genre, the percentage of each genre contained within each streaming service were calculated. This will make the stacked bar chart have bars that are easier to compare to one another.



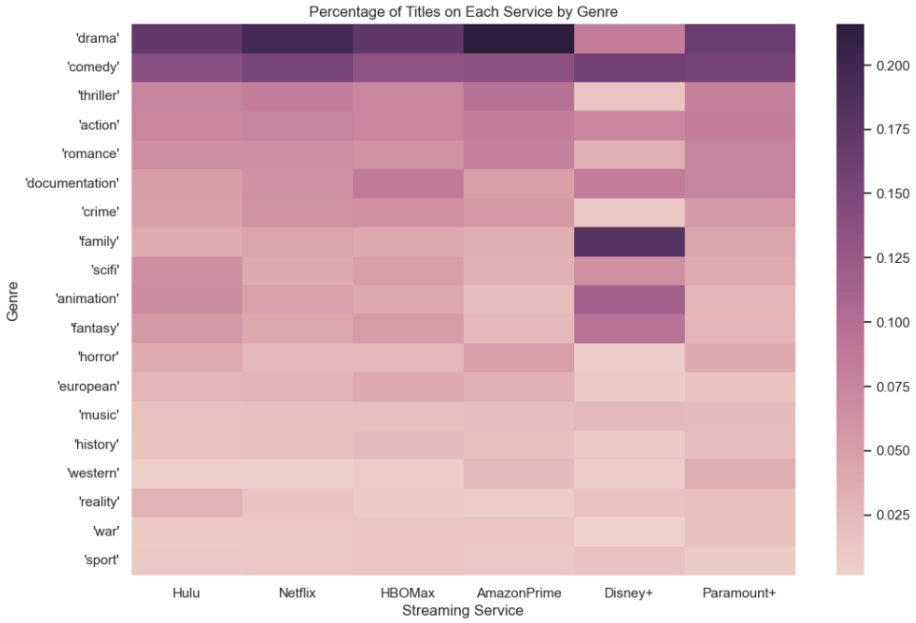
Even though there are a lot fewer titles in the sport and western genres for example, it is now easy to see which streaming services contain the most of those genres. Fans of reality TV will find much to enjoy on Hulu and Netflix, while Disney+ is the best service for those seeking family friendly content. Paramount+ has a large percentage of western titles, especially considering the smaller number of titles available on that service overall. HBOMax seems to have a fairly even split among all of the genres considered in this data set.

Two alternate versions of this visualization were created for readers who only care about either TV shows or movies.





The following heatmap provides a different take on this genre data, by showing the percentage of each genre as it relates to the total amount of content within each streaming service.



The most common genres across all streaming services are drama and comedy, although Disney+ has significantly more family content than any of the other services, while having significantly less crime content. More than 1 out of every 5 titles on Netflix and Amazon Prime are dramas. The high levels of drama, and comedy, may indicate that the streaming service providers see these genres as appealing to the largest markets.

Question 2: What service has the highest rated movies and TV shows?

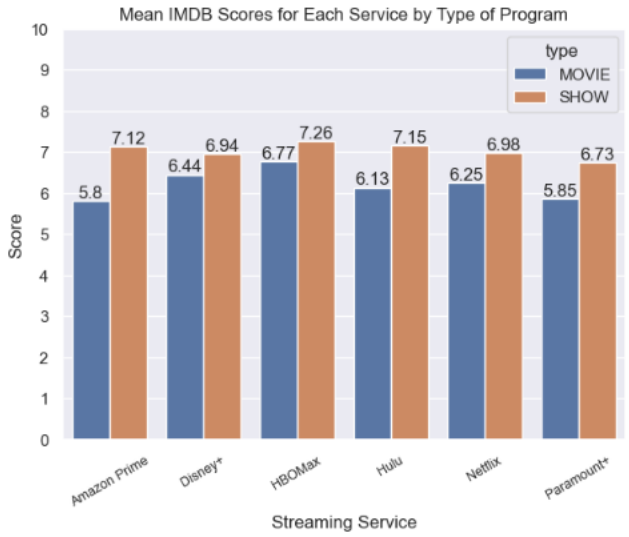
This question is intended to help the reader who is just interested in watching the most acclaimed movies and shows. Those that are rated the most highly are likely to be those that are the most watched among a random audience of viewers.

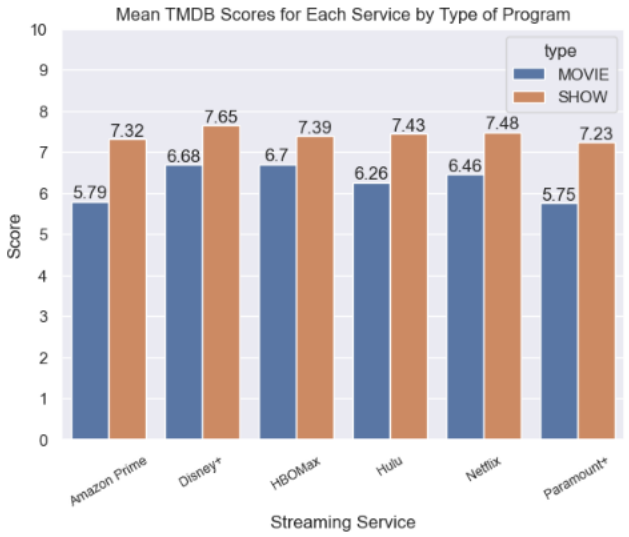
First, the highest rated movies and shows on each service were displayed. Looking at the top 5 rated movies and shows based on their IMDB scores on each service could help them narrow down their choice of service based on their interests. Since the average viewer would likely be more interested in the most popular titles on each service, only those with more than 7,800 votes were considered. This number would place those shows in the upper quartile of the data in terms of number of votes. This means these are the highest rated shows among only those with large popularity, which have been deemed as "high profile". While all the “Top 5s” can be found in the program output, part of the data frame containing the top five HBOMax high profile TV shows is shown here.

service	title	type	description	release_year	age_certification	runtime	genres	production_countries	seasons
HBOMax	Band of Brothers	SHOW	Drawn from interviews with survivors of Easy C...	2001	TV-MA	59	['war', 'drama', 'history', 'action']	['US']	1.0
HBOMax	Chernobyl	SHOW	The true story of one of the worst man-made ca...	2019	TV-MA	65	['drama', 'history', 'thriller', 'documentation']	['US']	1.0
HBOMax	The Wire	SHOW	Told from the points of view of both the Balti...	2002	TV-MA	59	['drama', 'thriller', 'crime']	['US']	5.0
HBOMax	Rick and Morty	SHOW	Rick is a mentally-unbalanced but scientifical...	2013	TV-MA	22	['scifi', 'action', 'animation', 'comedy']	['US']	5.0
HBOMax	Game of Thrones	SHOW	Seven noble families fight for control of the ...	2011	TV-MA	58	['scifi', 'action', 'drama', 'fantasy', 'roman...	['US']	8.0

There is a public perception of HBO being a platform for more mature programs, and the top 5 list of shows supports that opinion. All these shows are rated TV-MA, so clearly the best offerings on HBOMax are not for children. Most of these shows are dramatic action-oriented programs, although the animated comedy "Rick and Morty" is the fourth highest rated.

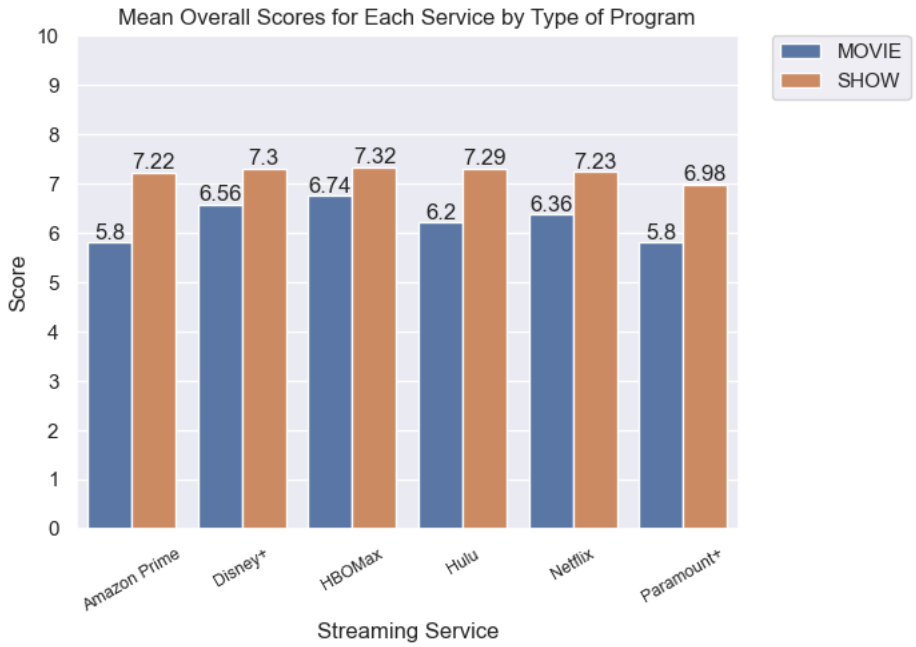
The mean scores of all the movies and TV shows on each service that have an IMDB or TMDB rating were then calculated. This data frame was then used to create side-by-side bar charts that allow for the comparison of the scores of movies and TV shows across the different streaming services.





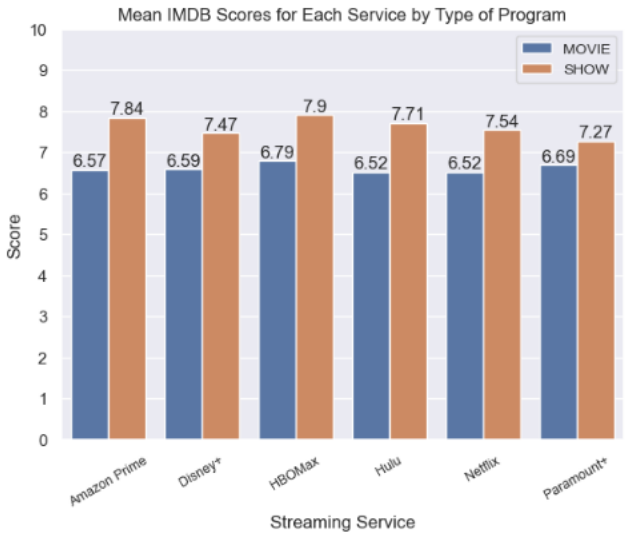
In general, the mean IMDB and TMDB scores are fairly similar. The mean scores for Amazon Prime movies only differ by 0.01. There are a few instances where there are larger differences. For example, Netflix shows have a mean rating on TMDB that is 0.5 points higher than on IMDB and Disney+ shows have a mean rating on TMDB that is 0.71 points higher than on IMDB. There are only three instances where the mean IMDB score is higher than the mean TMDB score. These are for Amazon Prime movies, HBOMax movies, and Paramount+ movies.

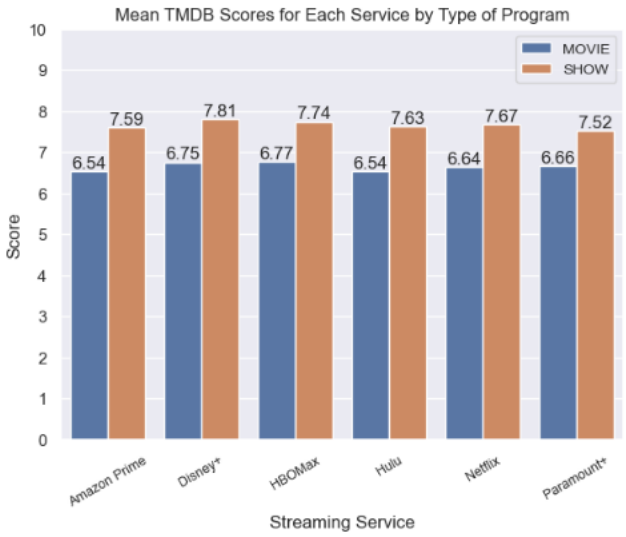
The mean IMDB and TMDB scores were then averaged to get an overall average score for each program type for each streaming service.



If someone is only interested in the service with highest rated TV shows or movies on average, they should choose HBOMax. It has the highest mean rating for its movies, beating Disney+ by 0.18 points, and the highest mean rating for its TV shows, again beating Disney+, this time by 0.02 points.

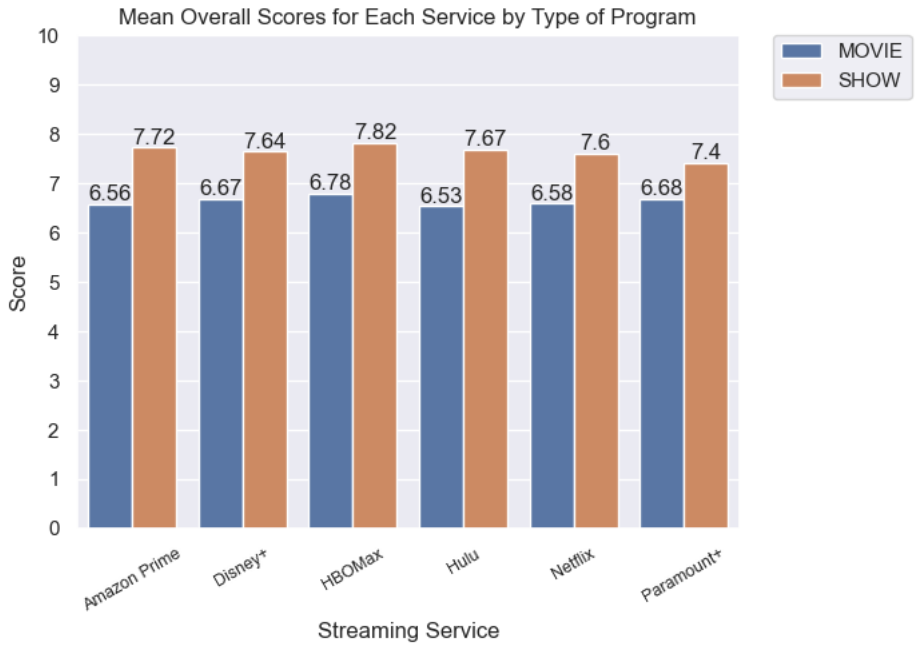
The analysis was then restricted to only the previously described high profile TV shows and movies to see if HBOMax would still be the streaming service with the highest rated content.





The high profile titles appear to have higher average scores than when all of the titles are included. Although the scores are now greater, the rankings of the services are the same. HBOMax still has the greatest mean scores for movies and TV shows on IMDB, although Disney+ now has the highest mean score for TV shows on TMDB.

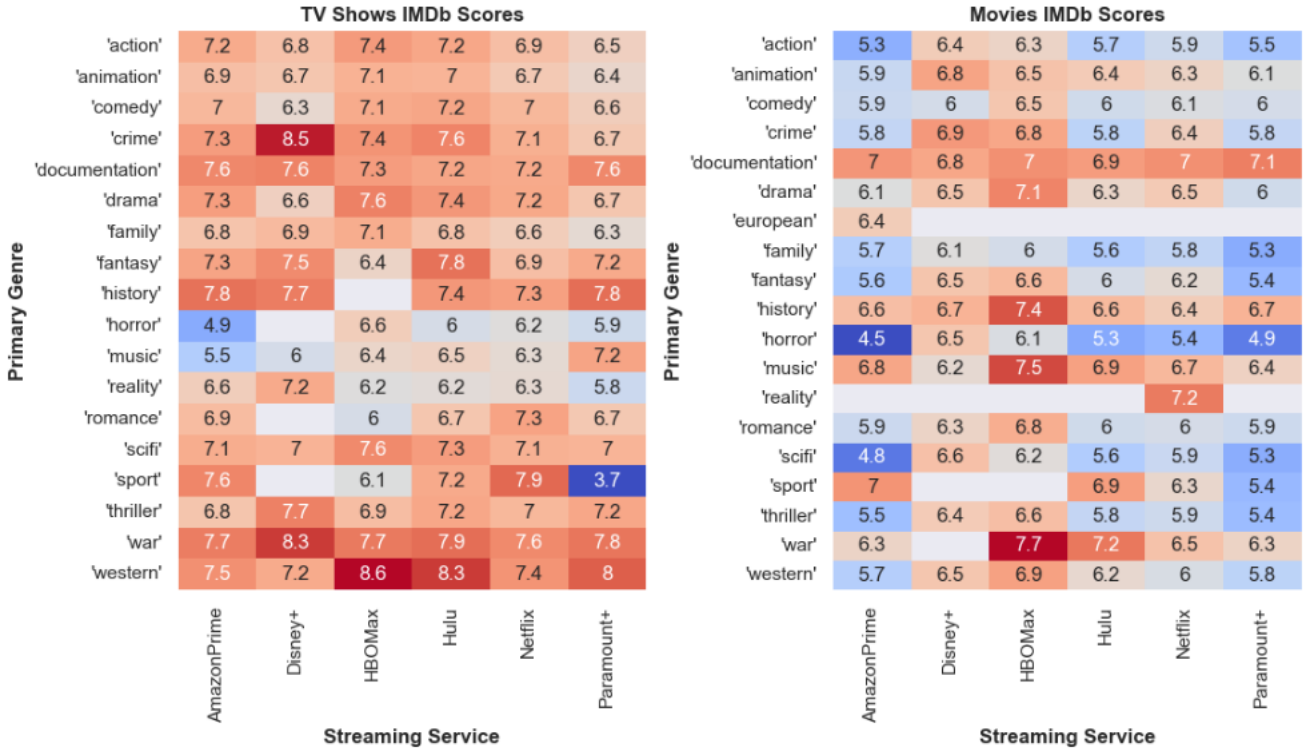
The mean IMDB and TMDB scores were averaged again to compare the overall mean scores for each streaming service.



Even when looking at just the high profile movies and shows, HBOMax still comes out on top. It has the highest scores again in both categories, meaning that much of its high ratings come from its popular shows. If one is interested in solely what is perceived to be the best, regardless of genre or content, HBOMax would be the best choice.

Since many viewers are only interested in certain genres, no matter how high scoring different titles might be, the data was separated based on each title's primary genre. The first genre tag listed in each title's genres value is interpreted as the primary genre for that title. The titles were once again separated by movies

and TV shows since a viewer might prefer different genres of each. The following heatmaps show the mean IMDB score for each primary genre within each streaming service.



In the heatmaps, darker red colors represent higher average scores, and darker blue colors represent lower average scores. The white spaces without numbers indicate that the streaming service does not have titles in that particular primary genre. Looking at the heatmaps across each row allows the reader to determine which streaming service has the highest average score in that genre. Looking at them along each column allows the reader to see what the highest and lowest rated genres on each service is.

Western TV shows on HBOMax have the highest average IMDB score overall, while sports TV shows on Paramount+ have the lowest. The highest rated movies on average are war movies on HBOMax, while the lowest are horror movies on Amazon Prime.

It should also be noted that this visualization reinforces that, overall, TV shows have higher average ratings than movies do for these streaming services.

Question 3: Which streaming services are improving or declining over time?

To answer this question, the changes in the services based on the release dates of programs were analyzed. The amount of improvement or decline was measured by changes in the ratings of the programs, the amount of content being released, and the variety of content available.

If someone is signing up for a subscription service, it is important to be confident that the service is growing and being updated. If new releases are not being added to the service, then one might quickly get bored of what content is available after they've watched what they were initially interested in.

Starting at the year each service was founded, the number of movies and shows were counted by release year. Since these programs were released after the streaming service began, the service shows a metric of growth by acquiring that much new content. It is also important to note that this data was pulled in the middle of 2022, so those numbers are representative of how much content each service added in the first half of that year.

The exact numbers of titles by release year for those time periods for each streaming service is detailed in the program. However, a summary is provided in this report.

Paramount+ is the youngest of the streaming services, starting in 2021. It has not had much time to show real trends of growth or decline, but it is still important to see if the amount of new content on the platform is similar or not to the other major streaming services. Since Paramount+ is brand new, there is not much to compare in this data. The good news is that the service is expanding in its second year since newer titles are being released for the service.

HBOMax is the next youngest, starting in 2020, however, this service evolved from HBO Now and HBO Go which were both previously available for several years. HBOMax has added a significant number of new titles being released each year, with 255 from 2020 and 258 from 2021. It will be interesting to see if the total number added from 2022 can keep pace with the previous two years of releases. Subscribers should be confident that HBOMax will have lots of new content though.

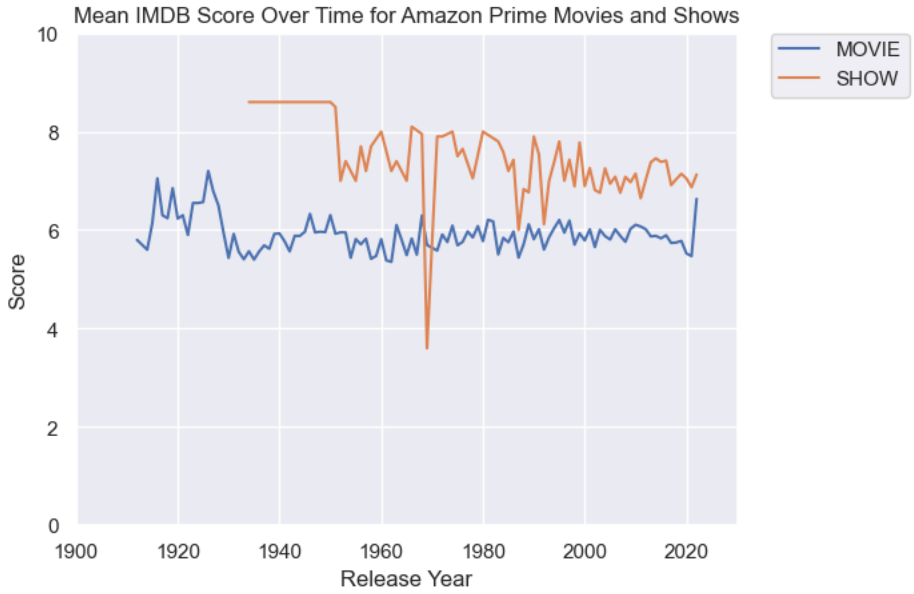
Disney+ launched in 2019. It has had a few years to show growth, but this streaming service is probably the most specific of these six as they are strictly focusing on content created by only a select number of studios. Even with a reduced pool of content to draw from, a good number of newly released titles have been added on to the service, with around 100 titles each from the past couple of years. Disney seems to be fully committing to expanding their streaming service.

Hulu launched in 2007, the youngest of the three much older streaming services. This has been a very popular streaming option for many years, so it will be interested to see how the greater levels of compartmentalization with a growing number of streaming platforms has affected their ability to get newly released content on the platform. Hulu appears to be doing a great job of adding newly released content. Hulu is matching the number of newly released titles being added on HBOMax. They have over 200 titles from each release year after 2019. There does not seem to be any risk of Hulu slowing down in the coming years.

Surprisingly, Netflix, the one that started it all, is not the oldest streaming service. Although they launched all the way back in 1997, it was originally just a DVD mailing service. They did not transition to on-demand streaming until 2007. Netflix has been adding a massive amount of newly released content to their available library, with over 800 titles from 2019 and 2020. With so many shows and movies, and many of them being developed directly by Netflix, it will be interesting to see if they are beginning to sacrifice quality in favor of quantity.

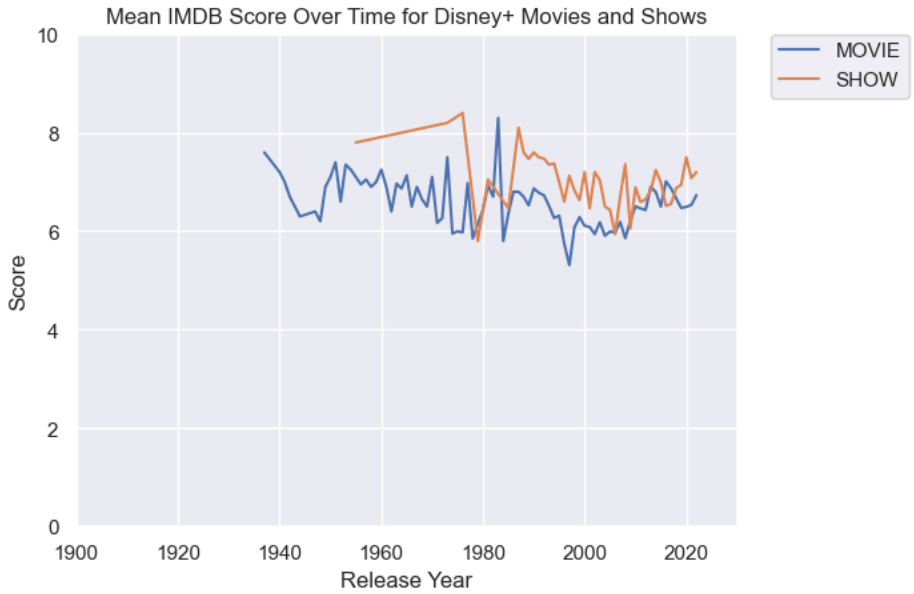
Amazon Prime Video has been around the longest, launching in 2006. The service was originally named Amazon Unbox and then renamed to Amazon Video on Demand in 2008, before settling on the Amazon Prime Video name in 2016. Amazon Prime has also been adding tons of newly released content to their platform as well, with roughly equal numbers to Netflix. It appears that Amazon has access to the greatest number of networks and studios and may be expanding the number of titles they are developing as well.

Each streaming service was also analyzed separately and the titles on each were grouped by whether they are shows or movies for each year they were released. The mean IMDB scores of the programs released each year are plotted over time. While these streaming services obviously did not exist 50 or more years ago, this allows the recent content they have licensed to be on their platform to be compared to older movies and shows. Focusing on the vertical direction of the rightmost part of the plot compared to any previous trends can reveal if the new titles on the service have started to decline, improve, or are holding steady. The following six line plots each visualize the mean IMDB scores by release year for each streaming service, starting with Amazon Prime. The rows of data for each streaming service are isolated and the mean IMDB scores are plotted by release year. To be consistent from service to service, the y-axis will span the entire range of possible scores from 0 to 10, and the x-axis will span from 1900 to present. This will also show how old the oldest content available on each service is.



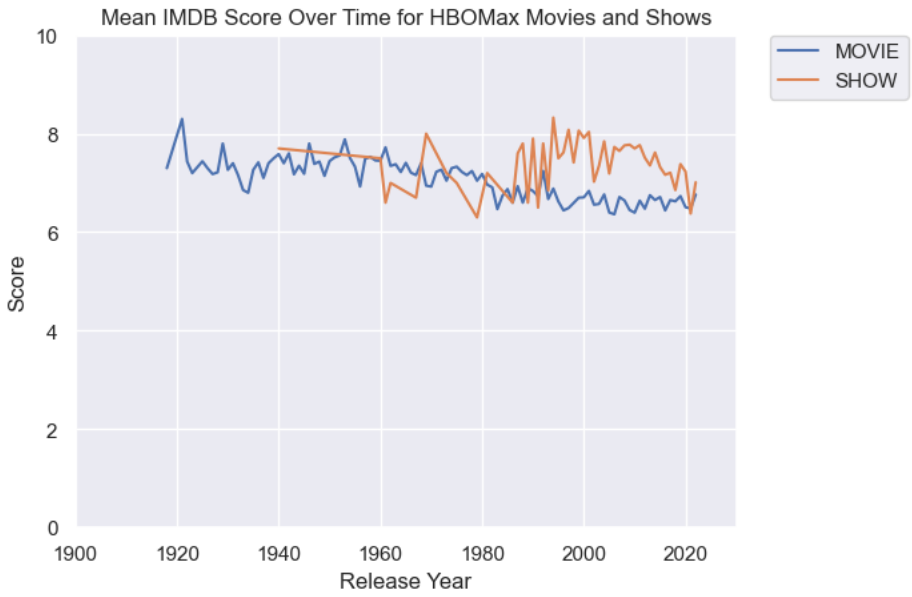
The mean IMDB scores for movies and shows on Amazon Prime are pretty steady from release year to release year. There is a bit more fluctuation in the mean scores for the shows compared to the movies. Most importantly, however, both the shows and the movies are trending upwards at the very end. This means the new content that Amazon is adding to Prime is starting to review better. Also, their available content goes far back, with the oldest movie released around 1910 and the oldest show released in the 1930s.

The line plot for Disney+ is shown next.



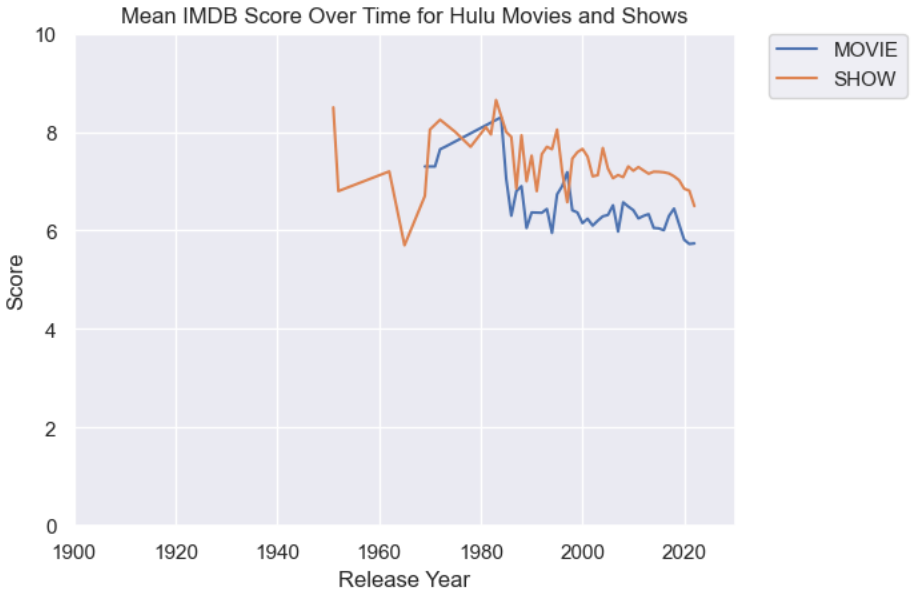
The mean movie and TV show scores, for the most part, stay within the 6 to 8 range across all the release years. Disney does not have as old of content as Amazon, with its oldest movie releasing in the 1930s and its oldest show releasing in the 1950s. Both movies and shows are showing an upward trend after 2020, but those are both occurring after some dips downward. It is unclear whether Disney has positive momentum or are actually stagnating as they add newly released content.

The following line plot is for HBOMax.



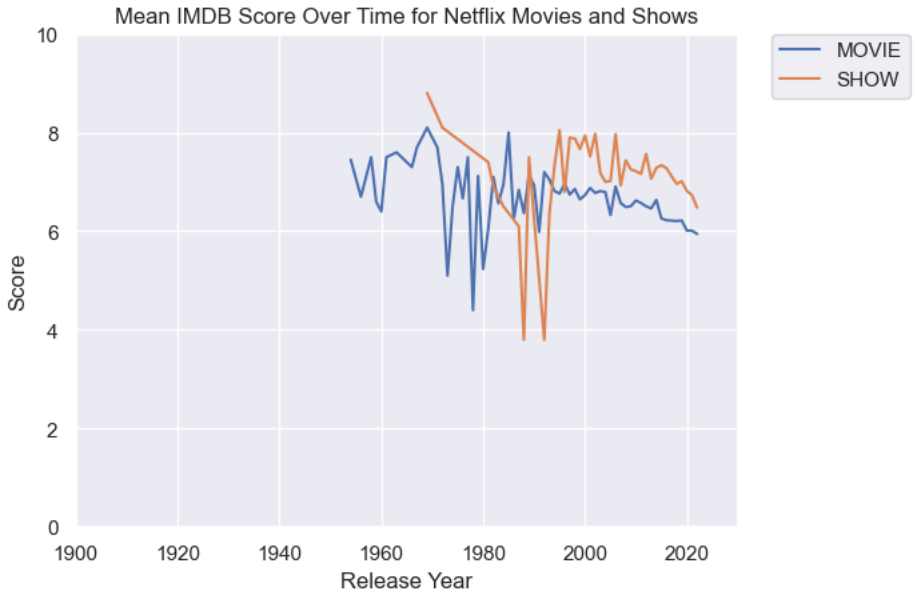
HBOMax features movies dating back before 1920 and TV shows starting at 1940. Their shows appear to be more highly rated than their movies over the last 30 to 40 years. The very end of the graph shows spikes upwards for both their movies and shows, although there is a bit of a drop in mean scores by release year for TV shows in the years leading up to that spike.

The following line plot shows the mean scores by release year for Hulu.



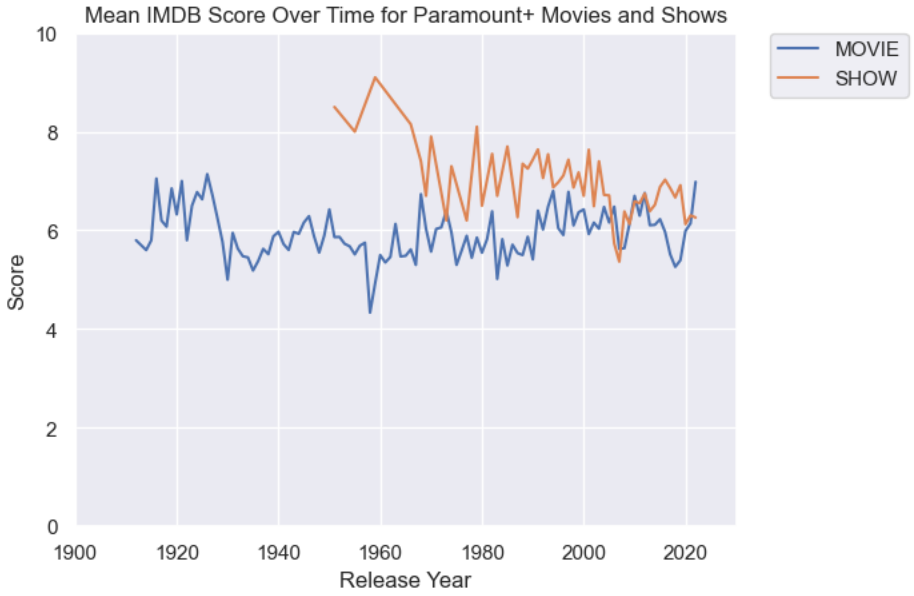
Hulu features a lot less older content when compared to the other streaming services. Their oldest shows are from around 1950 while their movies only date back to approximately 1970. This could be viewed as a significant drawback. Also, their mean IMDB scores by release year have been showing a decline over the last decade or so. This could be an indication that Hulu is struggling to provide new content at the same quality level as the older content on the service.

The next line plot shows the mean IMDB scores by release year for Netflix.



It is immediately obvious that Netflix has the least amount of older content compared to the other services. Their oldest movie was released in the 1950s, while their oldest show was released around 1970. This may indicate a trend of Netflix removing or not being willing to add older content to their service, while only adding newer content. Unfortunately for them, this newer content is not reviewing as favorably as older programs. The review scores of both movies and shows have been trending downward for more than the last decade. This may indicate that Netflix is on a decline and users are not as happy with what is being released on the service as they used to be.

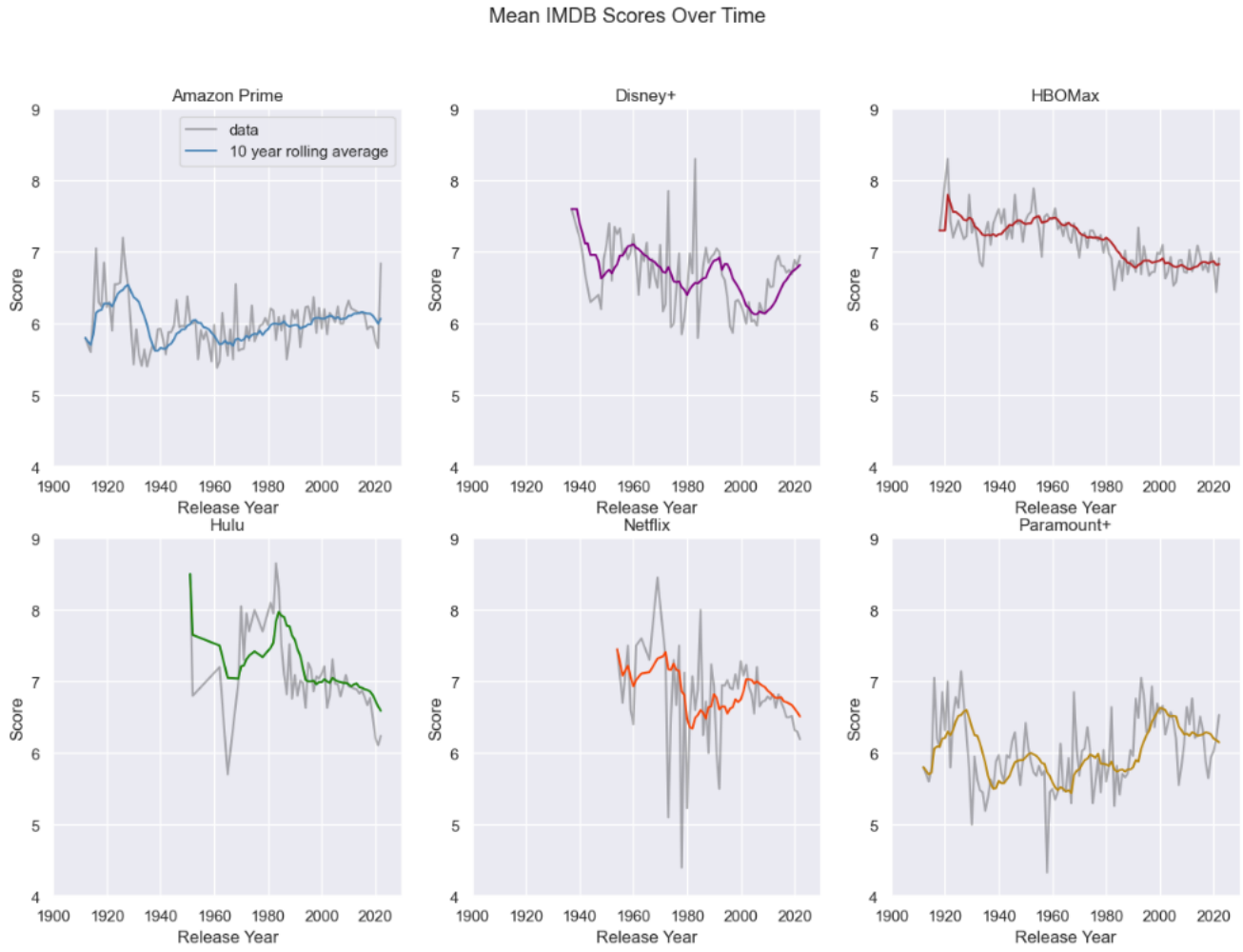
The final line plot visualizes the data for Paramount+.



Despite being the youngest service, Paramount+ has movies dating back to about 1910 and TV shows dating back to about 1950. The mean scores by release year for movies tend to fluctuate around a score of 6, with a bit of an upward spike appearing for those titles released in the last few years. TV shows on the other hand are not showing that kind of growth for the most recent titles. It is difficult to say whether Paramount+ is growing or not since it so new, but subscribers can be assured that the quality of their content is similar to the other services in terms of mean scores.

To better focus on any trends of growth or decline in the scores of titles by release year, a 10-year rolling average of the mean IMDB scores across all titles on each service will be overlaid on top of each line

plot of the mean IMDB scores by release year. This should give the viewer a clear picture on the quality of the content on each service that has been released over the most recent decade.



Disney+ is the only service whose 10-year rolling average is showing an increase in scores for those titles released in the last several years, and this is after a period of decline. Amazon Prime and HBOMax are holding steady with mean IMDB scores close to a 7 and 6 respectively. Hulu and Netflix are currently in periods of declining scores among their most recently released content. Both of their rolling averages have decreased to between a 6 and 7. Lastly, Paramount+ is showing a bit of decline in the ratings of its most recently released titles, but since the service is so new, it might be better to give the service a few more years to see how their upcoming titles perform with audiences.

Question 4: Is a program's score on IMDB related to the number of votes it receives?

The answer to this question examines the relationship between IMDB score and the number of votes that determined that score. Are the scores of titles that receive a lot of votes artificially inflated just because of their popularity or will a vast number of votes tend to drag the score down because there will be more unhappy viewers voting as well?

For this question, the first step was to determine if there was any correlation at all between the IMDB scores and the ratings given to the shows or movies on the different platforms. The scores and votes data frame was separated into six data frames based on the streaming service so each service can be analyzed individually as well. Using the data frame containing the titles on all the streaming services, a simple regression model was created to determine the rate of contribution that votes had on scores.

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2569	0.009	734.780	0.000	6.240	6.274
imdb_votes	2.543e-06	7.86e-08	32.347	0.000	2.39e-06	2.7e-06

The regression model reveals that increasing the IMDB votes count by 1 increases the IMDB score by 0.000002543. This means that the number of votes has close to no impact on the score the title receives. Regression models were created using each of the data frames for the titles on each streaming service.

Amazon Prime:

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9351	0.014	414.223	0.000	5.907	5.963
imdb_votes	4.852e-06	3.07e-07	15.815	0.000	4.25e-06	5.45e-06

Disney+:

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.4755	0.033	195.400	0.000	6.410	6.540
imdb_votes	1.662e-06	1.63e-07	10.199	0.000	1.34e-06	1.98e-06

HBOMax:

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.8012	0.020	332.237	0.000	6.761	6.841
imdb_votes	1.261e-06	1.09e-07	11.512	0.000	1.05e-06	1.48e-06

Hulu:

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.6043	0.027	248.659	0.000	6.552	6.656
imdb_votes	3.376e-06	3.2e-07	10.546	0.000	2.75e-06	4e-06

Netflix:

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.4563	0.016	401.401	0.000	6.425	6.488
imdb_votes	2.317e-06	1.63e-07	14.206	0.000	2e-06	2.64e-06

Paramount+:

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9880	0.025	241.813	0.000	5.939	6.037
imdb_votes	2.469e-06	2.1e-07	11.757	0.000	2.06e-06	2.88e-06

The regression models created for each service tell the same story. The change in the number of votes accounts for close to none of the change in IMDB scores.

Next, the Pearson correlation coefficient between the two variables is calculated using each streaming service's data set.

Amazon Prime:

	imdb_score	imdb_votes
imdb_score	1.000000	0.165892
imdb_votes	0.165892	1.000000

Disney+:

	imdb_score	imdb_votes
imdb_score	1.000000	0.293552
imdb_votes	0.293552	1.000000

HBOMax:

	imdb_score	imdb_votes
imdb_score	1.000000	0.208739
imdb_votes	0.208739	1.000000

Hulu:

	imdb_score	imdb_votes
imdb_score	1.000000	0.217997
imdb_votes	0.217997	1.000000

Netflix:

	imdb_score	imdb_votes
imdb_score	1.000000	0.190661
imdb_votes	0.190661	1.000000

Paramount+:

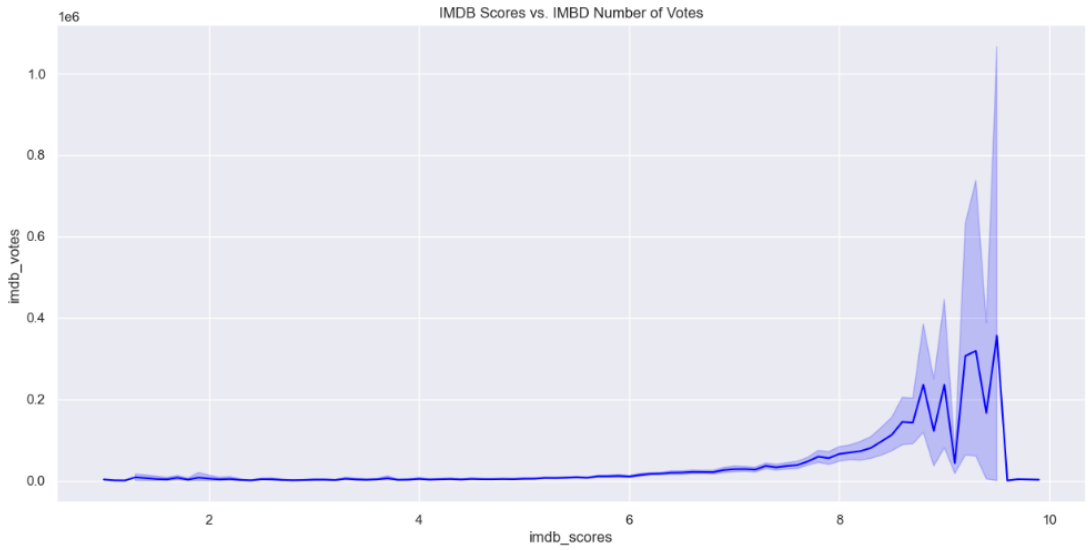
	imdb_score	imdb_votes
imdb_score	1.00000	0.22422
imdb_votes	0.22422	1.00000

There does happen to be some correlation between the two variables, which is in some contrast to the previous results, although this does not mean that the number of votes is directly affecting the score in some way, as correlation does not equal causation. All the streaming services exhibit some weak positive correlation between the number of votes and the IMDB score, with Disney+ having the strongest positive correlation with a correlation coefficient of approximately 0.29.

Next, the data was run through a standard scaler to showcase how the data for the correlation between scores and votes is very noisy and has no discernible pattern. Viewing the data in this light makes it no surprise that the previous correlations were low.

	index	0
index	1.000000	0.208378
0	0.208378	1.000000

The following plot reveals that there is not a high correlation between scores and number of votes until the titles have received very high overall scores. Around a score of 9, the correlation between the two variables very suddenly spikes upwards.



Titles with low scores, less than a 6, received almost no votes. But, once the scores increase past 6, the number of votes starts to increase. This may be because as a show begins to be more highly rated, it is more likely that someone else will also take action to vote for it. Maybe they would like to see this title pushed to the top of the ratings, or to the bottom, depending on how they feel about it. However, this does not explain why lower scores were associated with far less votes.

Question 5: Can age certifications be determined using the TV show or movie descriptions or genres?

The final question of interest will investigate if the age certifications are consistent with their descriptions and genres. This should uncover the trustworthiness of the age certifications as they are applied across the wide variety of titles on the streaming services. This will let subscribers with children know the level of confidence they should have in trusting the age certifications as the measure of what they allow their children to watch.

This question was answered using apriori association rule mining. First, the data needed to be prepared. The descriptions, and their accompanying age certifications, were isolated in their own data frame. Then those descriptions were tokenized, and the common English stop words were removed. This resulted in a new column with the tokenized versions of the descriptions. Also, the actual age certification was appended to the end of each tokenized description, so they were included in the overall basket of words that would be used for the association rule mining. Unfortunately, even when setting the support and confidence thresholds to very low levels, no significant association rules between the descriptions and age certifications were found.

The preparation process was repeated, this time with the goal of relating the genres of the TV shows and movies to the age certifications. The list of genres for each title also needed to be cleaned, with special characters being removed so that strings representing the same genre were not considered to be unique. While running the association rule mining, the right hand set of rules was fixed to be the unique values for the age certifications. This time, lists of association rules were generated. These were sorted and filtered so that only the strongest 3 rules were displayed for each age certification category, if that many rules existed.

Rules generated for genres associated with the 'TV-G' Rating:

Empty DataFrame
Columns: [antecedents, consequents, support, confidence, lift]
Index: []

Rules generated for genres associated with the 'Not Rated' Rating:

Empty DataFrame
Columns: [antecedents, consequents, support, confidence, lift]
Index: []

Rules generated for genres associated with the 'TV-14' Rating:

	antecedents	consequents	support	confidence	lift
96	(animation)	(TV-14)	0.012184	0.118849	1.942286
94	(action)	(TV-14)	0.014240	0.074745	1.221521
99	(comedy)	(TV-14)	0.023242	0.066327	1.083945

Rules generated for genres associated with the 'TV-Y' Rating:

	antecedents	consequents	support	confidence	lift
120	(animation)	(TV-Y)	0.010127	0.098789	8.239433

Rules generated for genres associated with the 'R' Rating:

	antecedents	consequents	support	confidence	lift
81	(crime)	(R)	0.032050	0.238935	2.030952
77	(action)	(R)	0.031041	0.162933	1.384929
79	(comedy)	(R)	0.036241	0.103422	0.879083

Rules generated for genres associated with the 'TV-PG' Rating:

	antecedents	consequents	support	confidence	lift
118	(comedy)	(TV-PG)	0.011641	0.033219	1.074175

Rules generated for genres associated with the 'NC-17' Rating:

Empty DataFrame
Columns: [antecedents, consequents, support, confidence, lift]
Index: []

Rules generated for genres associated with the 'TV-Y7' Rating:

	antecedents	consequents	support	confidence	lift
122	(animation)	(TV-Y7)	0.012417	0.121120	8.324037
126	(family)	(TV-Y7)	0.011641	0.097911	6.728982
125	(comedy)	(TV-Y7)	0.010709	0.030561	2.100342

Rules generated for genres associated with the 'PG-13' Rating:

	antecedents	consequents	support	confidence	lift
60	(action)	(PG-13)	0.021884	0.114868	1.527538
63	(comedy)	(PG-13)	0.027433	0.078286	1.041065
64	(crime)	(PG-13)	0.010205	0.076078	1.011698

Rules generated for genres associated with the 'G' Rating:

	antecedents	consequents	support	confidence	lift
2	(animation)	(G)	0.012766	0.124527	3.168121
5	(comedy)	(G)	0.018470	0.052707	1.340941
6	(drama)	(G)	0.012145	0.026638	0.677712

Rules generated for genres associated with the 'PG' Rating:

	antecedents	consequents	support	confidence	lift
49	(animation)	(PG)	0.010709	0.104466	1.575369
51	(comedy)	(PG)	0.029994	0.085594	1.290773
47	(action)	(PG)	0.013619	0.071487	1.078032

Rules generated for genres associated with the 'TV-MA' Rating:

	antecedents	consequents	support	confidence	lift
111	(crime)	(TV-MA)	0.019634	0.146370	2.138458
107	(action)	(TV-MA)	0.014046	0.073727	1.077151
109	(comedy)	(TV-MA)	0.024484	0.069870	1.020806

For the most part, the streaming services do stay true to their pairings of genre and age certification. Action and crime are most commonly associated with more mature certifications, while family and animation are for younger audiences.

Final Conclusions

This analysis revealed many aspects about the streaming services that the average subscriber would most likely not be aware of. Hopefully the reader can now make a more informed decision about what streaming service, or services, they feel are worth subscribing to.

When signing up for any streaming service other than Hulu, the reader should be aware that most of the content available will be movies and not TV shows. Almost all the content, regardless of type, was released after the turn of the century. The streaming content available is also mostly geared towards older audiences, with the one exception being Disney+. The runtimes of streaming content are what one would expect, with movies lasting around an hour and a half to two hours, and TV shows either being half-hour or one-hour episodes.

The different streaming services have diverse selections of genres for both their TV shows and movies, with drama and comedy being the most frequent by far. Each streaming service does seem to have their own focus as well. For example, Netflix has far more reality programming than the others, while Disney+ has the most animation and family content. HBOMax seems to specialize in trying to cater to everyone, with an almost even split of programming across all the genres that were analyzed.

If one is only interested in scores, then HBOMax and Hulu have the best TV shows, while HBOMax and Disney+ have the best movies. Limiting the analysis to the only the most popular content results in HBOMax having the highest rated TV shows and movies out of all six services. The score analysis was also broken down by genre with some standouts being HBOMax western TV shows with a mean score of 8.6, HBOMax war movies with a mean score of 7.7, Hulu fantasy shows with a mean score of 7.8, and Netflix sport shows with a mean score of 7.9. Some unfortunate standouts were Paramount+ sport shows with a mean score of 3.7, Amazon Prime horror movies with a mean score of 4.5, and Amazon Prime sci-fi movies with a mean score of 4.8.

While comparing the most recently released content on each service to the older content available, Disney+ and Amazon Prime seem to be the only services showing some level of improvement. Hulu and Netflix, in particular, have seen a significant decline in quality with their most recently released titles on a steep downward trend for the last 10 years.

There was not an overall correlation between votes and scores so it does not seem that scores should be viewed as untrustworthy depending on the number of votes. There was however a sharp increase in votes for the highest rated programs, those with about a score of 9 or more. This may be due to the immense popularity of the program inspiring people to go vote for the program as well. At this point it is not clear if the large number of votes led to the high score, or if the high score led to it getting more votes afterwards.

Lastly, while there was no strong association between the descriptions and the age certifications, there were some consistent patterns between the genres and age certifications. The genres do match with their intended audiences with action and crime programming being for adults, and animated programs being kid friendly.

Since this report was created with the intention of everyone drawing their own unique conclusions from it, this section will feature a few small case studies of example user profiles. The user will be described and then the relevant output will be referenced to show how that person could make a decision about what streaming service is best for them.

User Profile 1:

Mark is only interested in watching the best and most popular shows. He likes pretty much any genre of TV show or movie as long as he is keeping up with what the community at large is talking about. Mark is single and has plenty of time each night to watch TV if he wants to.

The best streaming services for Mark may be either Amazon Prime or HBOMax. Amazon Prime has the most content to watch and has the most titles available in every genre except for family and reality. Amazon Prime, however, does not have the highest rated movies and shows. HBOMax has roughly equal numbers of every genre and has the overall highest ratings for TV shows and movies. The top shows on HBOMax are also very popular including titles such as Chernobyl, Game of Thrones, and the Wire, as well as the Lord of the Rings movies.

User Profile 2:

Jennifer is a movie buff who enjoys a good scare, as well as watching old movies because she is interested in film history. She does not watch too many TV shows because she works different shifts throughout the week as a surgeon and finds it hard to commit to staying current with any series. She is also not worried about what others think of the movies, because even the “bad” movies are interesting, if only just to see what decisions led them to not be successful.

While Amazon Prime has the most movies, its horror movies are among the worst rated titles in all the streaming services. Even though Jennifer is not worried too much about ratings, she may want better offerings in her favorite genre. Hulu would also not be a good fit for her, as she does not watch many TV shows and

Hulu has more shows than movies. While HBOMax does not have the greatest number of horror movies, the ones it does have are the highest rated and their movies date back to 1920. HBOMax may be the best fit for Jennifer as well.

User Profile 3:

Ken is a college student and avid Dungeons and Dragons player who loves getting absorbed in great fantasy stories. He finds family or children-oriented titles to be too immature and rather just watch programs that are intended for older audiences. He likes action and sci-fi titles, thrillers, as well as anime. He has a lot of time to watch TV outside of his classes and homework, and wants to make sure he has plenty of content available to keep him entertained for a long time to come.

Hulu scores the highest for fantasy TV shows, is tied for the second highest for action, animated, and thriller shows, and scores the second highest for sci-fi shows. The greatest proportion of titles on Hulu are rated TV-14 , TV-MA, R, or Not Rated, so this service fits in with his preference towards content for older audiences as well. Also, three of the top five rated shows on Hulu are anime. Even though Hulu does not have the largest library of titles, this service seems like the perfect fit for Ken.

User Profile 4:

Makayla has two younger children. She enjoys drama and romance titles, but also wants to make sure there is plenty for her kids to watch. Her children are 6 and 8 and mostly enjoy watching cartoons. After her kids taking up the TV for most of the afternoon, she wants to make sure there is plenty for her to enjoy after the kids are asleep at night.

While Disney+ might seem like the obvious best choice for Makayla's kids, this service does not have many drama or romance titles for Makayla herself. Netflix has the greatest proportion of both drama and romance shows on their service, and many reality shows as well, which can encompass aspects of both those genres. Netflix also has the greatest percentage of available animated and family shows. Netflix's family content is also fairly well enjoyed, with family and animated titles both averaging a little above 6 which seems to be the norm for streaming titles in general. Netflix's vast library of almost 6,000 titles will certainly keep Makayla and her children entertained.

Group Tasks and Roles

Bryan D'Amico: Introduction, Data Importing and Creation, Data Cleaning, Exploratory Data Analysis, Question 1, Question 2, Question 3, Final Conclusions and user profiles, Project Report

Sintia Stabel: Genre splitting and variable creation, Question 1, Question 2, Final Presentation, Final Conclusions

Matthew Smith: Question 4, Question 5, Final Presentation