Analysis of Streaming Services and their Contents

Kelly Chu and Khoa Tran

Summary of Research Questions and Results:

1. Which streaming service provides the most diverse selection of movies and tv shows?

In the first part of our project, we will look at the different streaming services and analyze which service provides the most diverse selection of movies. In order to achieve this goal, we looked at the genres that are provided across the streaming services. In this research, we looked at Netflix, Hulu, Disney+, and Prime Video. We found that Netflix, Hulu, and Prime Video all have the same amount of unique genres at 27. However, they are different genres as each streaming service caters to their audience. Disney+ has the lowest amount of unique genres at 24. We then looked at the different genres and determined the percentage of movies for each genre in a particular streaming service. We also determined the average percentage of movies for a genre, and found that Disney+ has the highest average percentage of movies for each genre. However, this data is skewed with the high percentage of movies with the Family genre. Overall, between Hulu, Netflix, and Prime Video, the average percentage of movies for each genre is about 8%, but Hulu is a bit higher with an average of 9%. As a result, we determined that Hulu has the most diverse selection of movies due to the high genre count, and a high average percentage count of movies in a genre.

2. What streaming service has the highest rated selection of movies and tv shows?

To examine this question, we will look at the merged movies and tv shows dataset and its IMDb and Rotten Tomatoes ratings. Creating a uniform 1-10 rating scale, we will then find the average of all the ratings for each platform's content. After completing the necessary steps, we found that, of the four platforms, Hulu holds the highest average rating of available movies and tv shows at 6.95. Closely following, Netflix has an average of 6.81, Disney+ has an average of 6.55, and Prime has an average of 6.47.

3. How accurately can we predict the ratings for a movie based on its qualities?

Training our machine learning model on various combinations of features, we will search for the most suitable set to accurately predict the IMDb ratings of a given movie. On our test set, our decision tree regressor machine learning model can predict a movie's IMDb rating accurately within a mean squared error of 0.006 to 0.02 between our predicted ratings and given training and test ratings. The mean squared error range for our training set was approximately 0.03 to 0.05. The ratings are on a 1-10 scale, which also contributes to the small error margins.

Motivation and Background:

The issue is based on comparing the different streaming services with the content that each provides and the quality of the content. There are many different streaming services, and it is

ridiculously difficult to pick one due to the different content that each provides. Due to most families paying for multiple streaming services, the motivation is to narrow the choices of streaming service, while maximizing quality service. We want to analyze the ratings and genres provided by each streaming service, in order to provide the user with more information about a particular streaming service. This problem is worth computing because it allows users to analyze the different streaming services and purchase the one that fits them the most with the type of movies or tv shows that they prefer to watch. This project's goal is to observe the diversity of each streaming service's content and the quality of each in order to effectively cater to different audiences.

Dataset:

IMDb Movies Extensive Dataset

Link: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

This dataset contains over 81,000 movies and 22 attributes each such as genre, year, and country. It also contains reviews from users and critics and a number of votes.

Selection of Movies for Streaming Services

Link: https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

This dataset contains a list of movies with their responding rating on IMDB and Rotten Tomatoes along with the corresponding streaming service that provides the movie.

Selection of TV Shows for Streaming Services

Link: https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney

This dataset has information on a collection of tv shows from what streaming service they are available on, IMDb ratings, target age group, and other aspects.

Methodology:

Prepping our Datasets

Before we analyzed any data, we loaded our three csv files, IMDb Movies Extensive Dataset (IMDb), Selection of Movies for Streaming Services (Movies), and Selection of TV Shows for Streaming Services (Tv Shows). Then, we created two methods to merge two pairs of DataFrames. One resulting merged DataFrame contains all the original rows and columns from

both Movies and Tv Shows. The other merged DataFrame contains only rows with movie titles found in both the IMDb and Movies datasets.

In these datasets, we focused on the columns "Title", "Genres", "IMDb", "Rotten Tomatoes", "Netflix", "Hulu", "Prime Video", "Disney+", "Year", and "duration".

Counts

The number of movies and tv shows available is one element to consider when reviewing the selection of a streaming platform. For each platform, we totaled the number of available movies and tv shows separately.

Genre

Genre was the key component we based our analysis of platform content variety and selection on. We created a series of "Genres" for movies and tv shows available in the given platform. The data in the "Genres" column is stored as strings, and each movie or tv show can fit under multiple genres. Consequently, we dropped NaN values in "Genres", split the strings by the "," delimiter, and used explode() to transform list elements to individual rows. This resulted in each row of the dataset having one genre and the movies would be duplicated. This allows us to find the unique genres of all the different streaming services. As a result, we resulted in a list of unique genres for each streaming service and the total count of genres. With this information, we created a list that references the unique genre list by having the count of movies in a specific streaming service referenced with each genre on the unique genre list. This way, we could compare the amount of content on each streaming service for each genre. We also calculated the standard deviation of the count of movies for each genre to view if the amount of content for each genre is evenly distributed. Afterwards, we created a list of percentages of movies in a specific streaming service associated with all the different genres. Then, we could compare the diversity of each streaming service by the percent of movies for each genre, analyzing if a streaming service skewed their content towards a particular group of genres. This was done by creating a bar chart with the percentages of movies and the different types of genres. However, this wasn't enough. We realized that percentages and averages can be wrong if the amount of movies for a streaming service is less than the other. As a result, we created a box chart of the count of movies for each streaming service in order to show the averages of movie count for a genre and the size of each streaming service's content.

Average Ratings

Finding which platform has the highest rated movies and tv shows required us to develop a method to incorporate both ratings from "IMDb" and "Rotten Tomatoes". IMDb ratings are on a 1-10 scale, and Rotten Tomatoes ratings are based on percentages from 0-100. We decided to

convert Rotten Tomatoes ratings from percentages to a 1-10 scale like IMDb's. This method uses "Title", "IMDb", "Rotten Tomatoes", "Netflix", "Hulu", "Prime Video", and "Disney+". First, we dropped rows with NaNs in the studied columns. Rotten Tomatoes values were converted to ints and divided by 10. This transformed its scale to 0-10. Accommodating the lack of < 1 ratings in IMDb, we set every Rotten Tomatoes value that was less than 1 to 1.

We wanted to include as many data points as possible to determine fair movie ratings. So we created a new column and set it equal to the average between IMDb and RT. Finally, we took the average of that average column for each platform to compare the overall average rating of all their available movies and tv shows. We combined these averages because IMDb represents the general audience's perspective, and Rotten Tomatoes represents more seasoned movie critics. Thus, the average ratings of these platforms' content takes into account both audiences for an overall rating.

Machine Learning

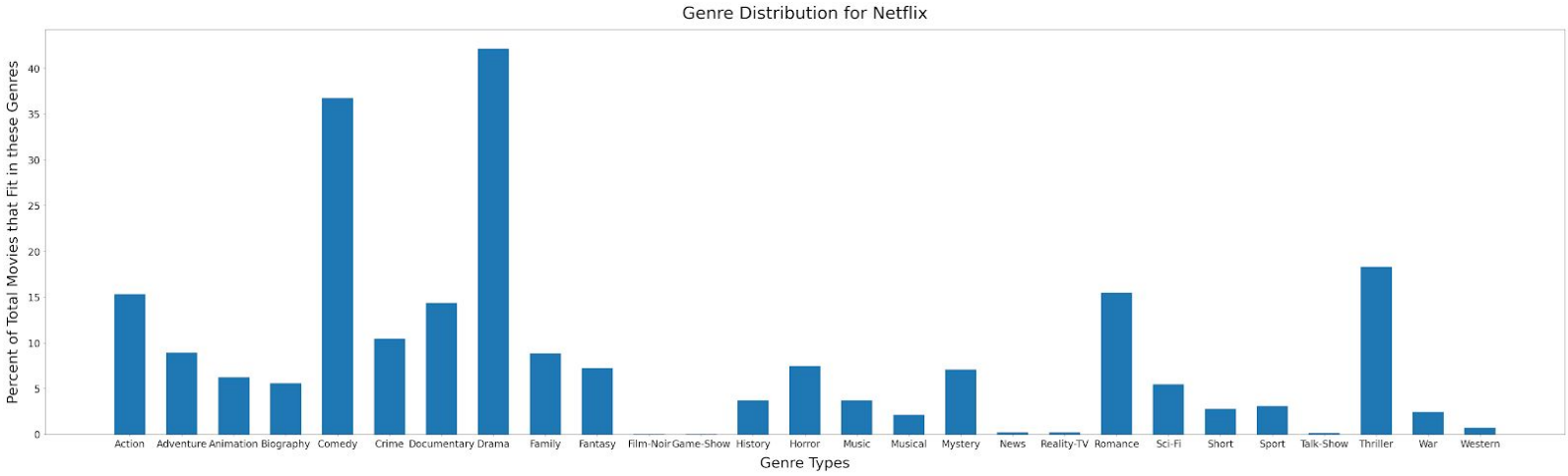
To predict the IMDb rating of a movie given its characteristics, we decided to test various combinations of features to train our model until we were satisfied with the resulting mean squared error. Since our rating predictions are floats, we trained a DecisionTreeRegressor. We chose to use 20% of our merged imdb_movs dataset as our test set to ensure that we have enough data to train our model. Taking in a features parameter, the model will train itself based on the given columns with numeric information. The "Genres" column is always one of the features. However, as addressed in the genres section, data in "Genres" is stored as strings, and can contain multiple genres that characterize a movie. Combating this, we split "Genres" by the "," delimiter in order to implement one hot encoding. We dropped the "Genres" column from the dataset to add back columns corresponding to all the unique genres. A 1 in these columns indicate the examined movie falls under that genre. Next, we fitted our model to our training features, predicted the ratings, and calculated the mean squared error. Then, the same was done to our test set. Returning, both the training MSE and test MSE, we can examine the accuracy of our model.

Tuning the model, we tested a few combinations of imdb_movs columns as features. The first features list consisted of the streaming platforms and genres. The second had the same columns plus "duration" of movies. The third and final trial included all the previous columns plus the "Year" the movie was released. Testing if the availability of a movie on streaming platforms affects the rating prediction, we decided to train a model without the platforms, with only "Year", "Genres", and "duration" columns. The mean squared error allowed us to review the accuracy of the predictions based on a movie's given characteristics.

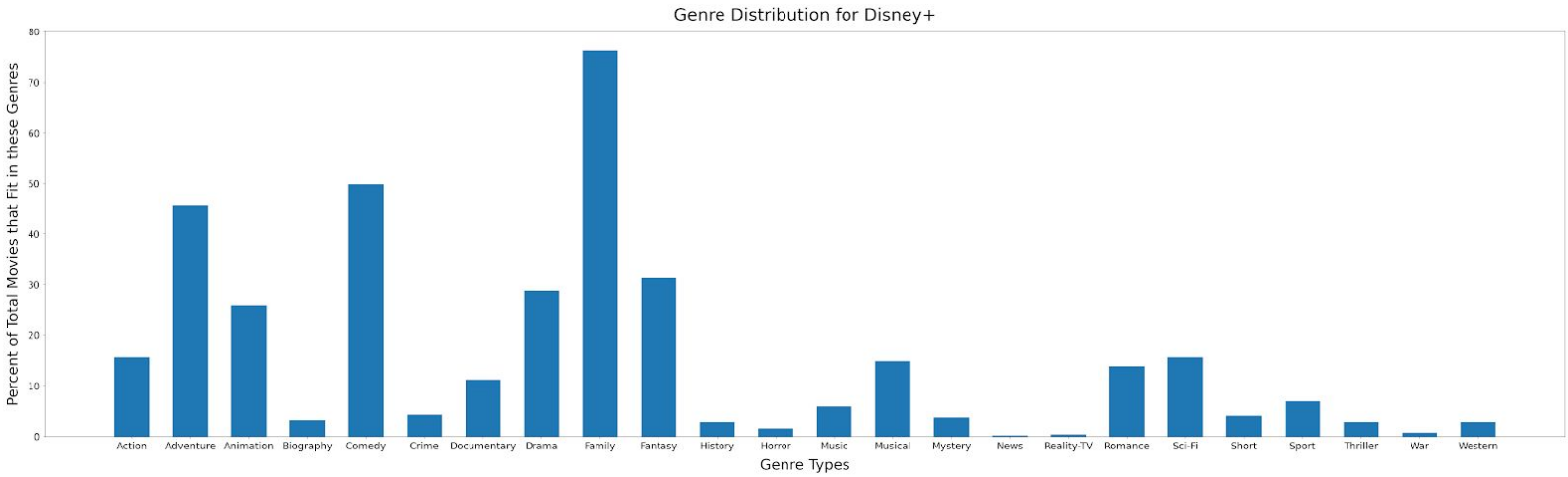
Results

Diverse Selection

For our first research question on the diversity of movies provided on the four different streaming services, we analyzed the genre that is present in each of the streaming services. Through obtaining the unique genres that each streaming service provides, we calculated the percentage of movies with each genre that a particular streaming service has. From this data, we were able to plot bar charts with unique genres on the x axis and percentage of total movies that fall under each genre category. From the bar charts below, we can see the percent of movies that

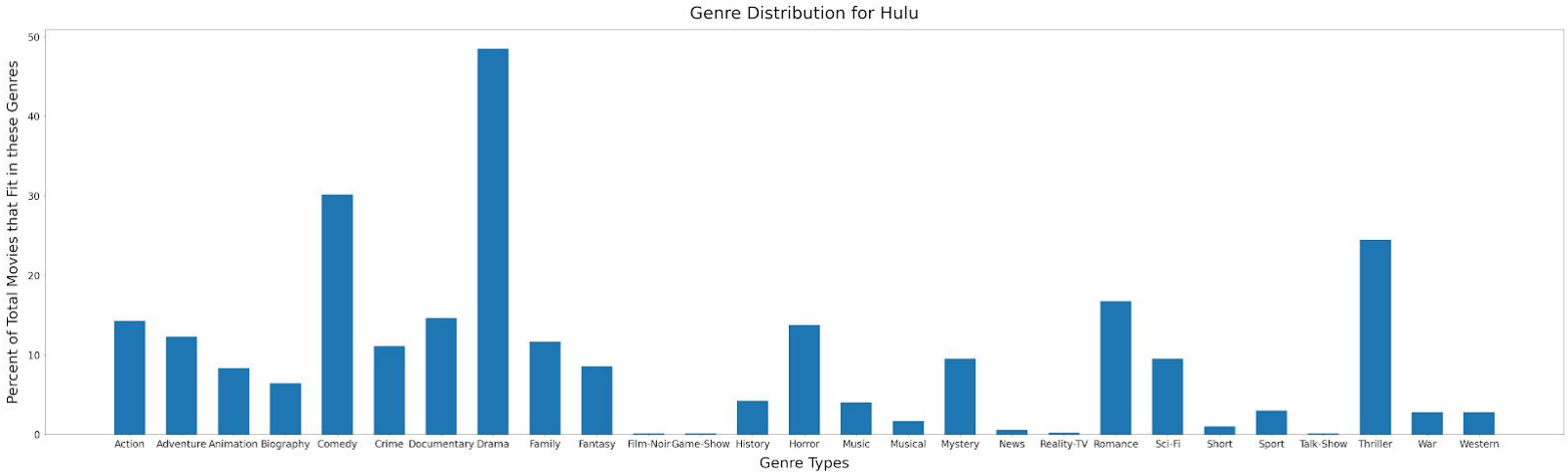


fit in each genre on Netflix, Disney+, Hulu, and Prime Video. Looking at the charts, above and below, it appears that these found platforms have very similar distributions of movie genres. All platforms except Disney+ have 27 different genres. Disney+ has one less genre, and Family is

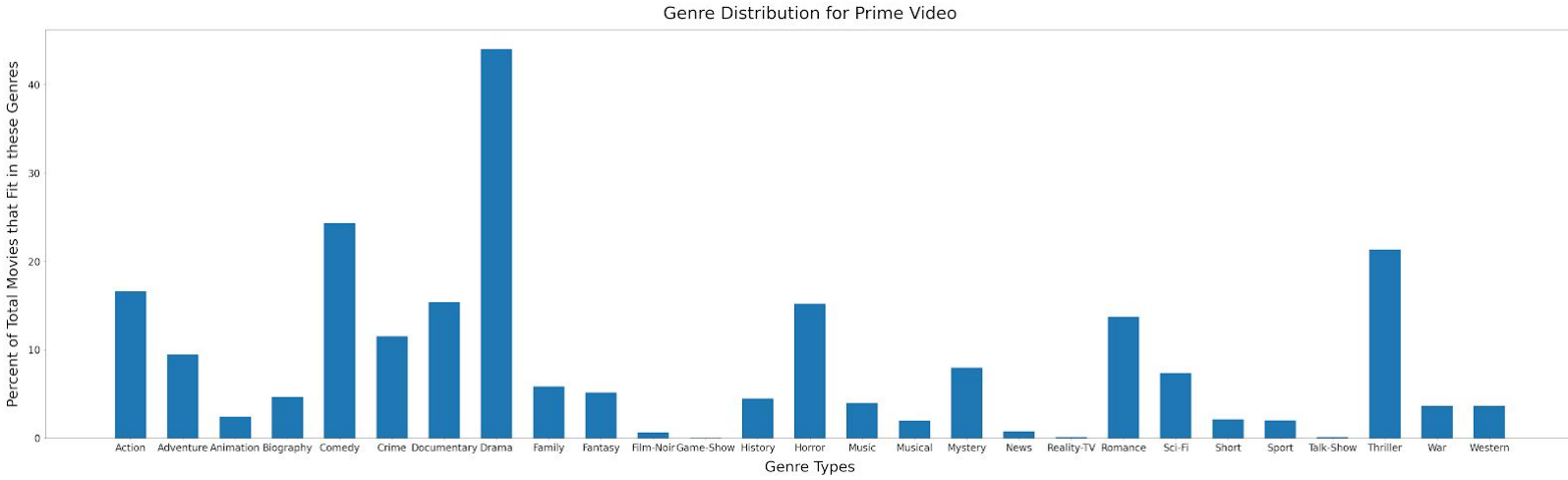


the leading genre unlike the other three with Drama on top. This is to be expected, as Disney+ is catered more towards family and kid-friendly movies.

Overall, these bar charts indicate a relatively uniform genre variety among Netflix, Hulu, and Prime Video. On Netflix, Hulu, and Prime Video, Comedy, Drama, and Thriller are the top three



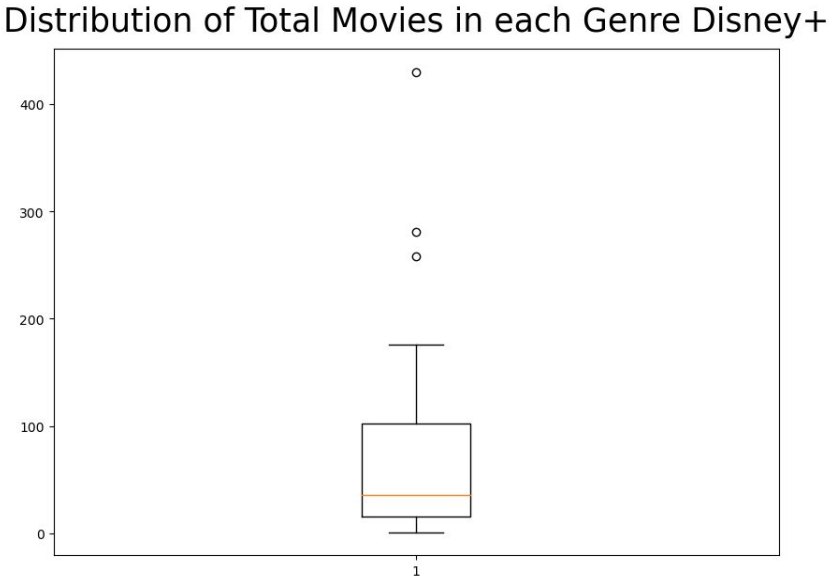
genres that movies fall under. Surprisingly, on all platforms, Reality Tv lands in the bottom five genres, in terms of total platform genre make-up. Admittedly, one misleading aspect of these bar charts is how total percentages do not add up to 100. Because each movie can have multiple genre labels, these percentages actually show the fraction of total movies that fall under each genre category. For example, over 40% of movies on Netflix and Hulu can be categorized as Drama. The graph does not indicate that Drama makes up 40% of the genre selection.

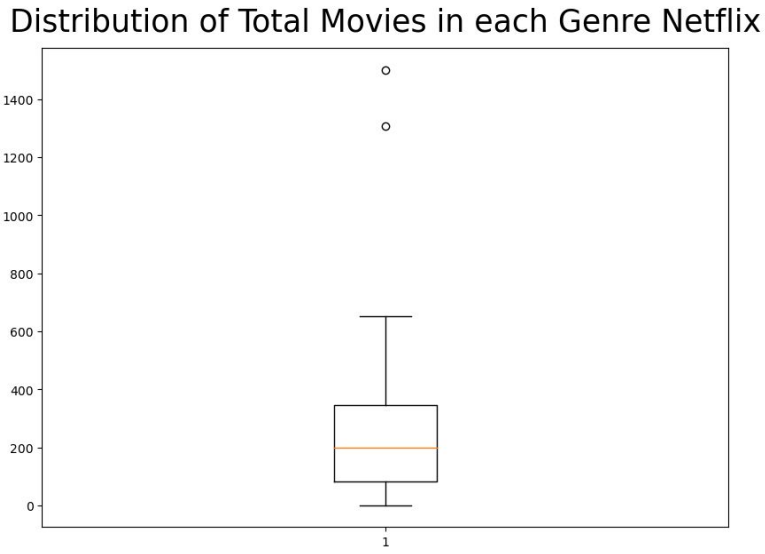


To further discuss the overall spread of these streaming services, we calculated the average of all genre percentages. Reviewing this data, it is unclear what specifically these averages represent

on their own. However, related information can be useful to examine the distribution of movies across genres. Surprisingly, Disney had the highest average genre percentage of 15.33%. The second highest percentage was Hulu at 10.89%. However, the standard deviation for distribution of genres was 19 compared to 10 across the other three platforms. We interpreted this to indicate that Disney had more movies with overlapping genres with a greater disparity among genre percentages as shown by the high standard deviation. However, the high average could very well also be caused by having one less genre than the other platforms. If genre percentages totaled to 100, this average would be strictly dependent on the number of genres available. But, it is not for this case.

Since our percentages do not add up to 100%, as each movie has more than one genre, we cannot only look at percentages because Netflix has more movies than Hulu, which brings their percentages down. To combat this, we also charted a box chart of the count of movies for each genre for Netflix, Hulu, and Prime Video. Looking at the box plot, the movie counts of each genre for Disney is positively skewed.





The median is much closer to the lower quartile vs Netflix which has a middle approximately in the center. This means the mean of movies in each genre is greater than the median for Disney+. We interpreted this to be due to having more genres with movie counts on the lower end, but those that had high movie counts had significantly more. We found out that Netflix has a higher average movie count for each genre which was expected because they have more total movies (3560 vs Disney+'s 564).

In conclusion, we determined that Disney+ has a strong selection of movies with the Family and Fantasy genres, while Netflix has a higher count of movies for each genre, and Hulu's content being the most evenly distributed among all the genres. If a user focuses more on Family and Fantasy movies, they should pick Disney+. While a user that likes a few genres and wants a bigger selection range should go with Netflix, and a user that prefers to watch all of the wide selection of genres should go with Hulu.

Highest Rated

Evaluating the ratings of streaming platforms in our datasets, we found that Hulu had the highest overall average rating of available movies and tv shows. Average ratings are from 1-10, one being the lowest and ten being the highest rated. Among Netflix, Hulu, Prime Video, and Disney+, Hulu came on top with an average of 6.95. Next comes Netflix at 6.81, Disney+ at 6.55, and finally Prime at 6.47. With only a 0.48 difference between the highest and lowest average rated, the range of the average ratings among these sites is relatively small. Though Hulu has the highest rated content, the difference between its rating and its runner ups', Netflix, is 0.14, or otherwise worded Hulu has 2.06% higher rated content than Netflix overall. Initially reviewing the results, it was surprising to see Hulu come first and Disney+ in second to last place. By only viewing the rankings, without any numbers, these comparisons can be misleading because of its small range. This small range may be attributed to the overlapping of many movies and tv shows. Movies and tv shows shared across these sites could have increased the uniformity of results. However, this does not detract from their value in the average calculation, as they are still valid ratings regardless of how widespread the content is. This ranking and these averages do not directly correlate to user's satisfaction or enjoyment of the respective platforms. These results may also be interpreted as indicative of the overall quality of content provided in each platform; however, many aspects could contribute to content quality, and our analysis of this research question strictly reviews the general response of viewers to the movies and tv shows themselves not in relation to these platforms.

Predicting IMDb Movie Ratings

Our final machine learning model predicted our test set's movie IMDb ratings within a 0.006 and 0.02 range of mean squared error between our predicted ratings and given training and test ratings. Without considering any other elements, these appear to be extremely small error values. However, labels, the IMDb ratings, are on a 1-10 scale, so the difference in the predicted value to the mean value, and subsequent error values, should be considered on a similarly-sized scale. On our first test trial of training our decision tree regressor, we used the year, genres, and duration of movies as our features. This resulted in an approximately 0.08 MSE value for the training set and 0.025 MSE value for the test set. While we were surprised and satisfied with these low values, we wanted to see how low we could get our error. So, we trained our model on streaming

platforms and genres. Our error increased to around 0.92 and 0.8 for training and test sets respectively. This showed us that which platform a movie is available on was not as significant in identifying patterns in IMDb ratings than the duration and year a movie was released. Testing on streaming platforms, genres, and durations, our error went back down to around 0.3 and 0.15 for training and test sets respectively. Finally, with platforms, genres, year, and duration as features, we trained our model to have around a 0.04 and 0.015 mean squared error for training and test sets respectively. With each addition or deletion, we discovered which qualities of movies were more indicative of its success with viewers rating-wise. We found that year and duration had a substantial impact on our model's accuracy. By adding duration to the set of features already encompassing streaming platforms and genres, we decreased MSE by approximately 67% in our training set and 81% in our test set. With all that in mind, we were still amazed at how small our error value was. We did not expect the year a movie was released and a movie's duration to have such a large impact on our model's ability to predict IMDb ratings.

Challenge Goals

Multiple Datasets

We met our challenge goal of utilizing multiple datasets. We hoped by incorporating more datasets, we would form a more well-rounded analysis of streaming platform's content selection. We achieved this by creating multiple merged datasets based on our research questions. One included Movies and Tv Shows carried on the same four streaming platforms. Using this dataset, we were able to examine the genre selection of each platform as well as the number of movies and tv shows available. Taking into account some streaming platforms may carry more movies than tv shows or vice versa, we focused on each platforms' overall genre variety, information from our merged dataset. We also looked at these datasets for the ratings of each movie and tv show from a variety of websites in order to get a better gauge of the population's thoughts on the movie. While the Movies and Tv Shows streaming datasets detail IMDb ratings, we found another dataset that had more extensive information on movies, including Rotten Tomatoes ratings. We successfully merged the IMDb and Movies datasets to answer the question of which platform has the highest average ratings. Furthermore, this dataset was the center of our machine learning portion of the analysis, which is our second challenge goal. By Implementing multiple datasets, we were able to perform analysis on each streaming service and as well as the tv shows and movies provided within each service.

Machine Learning

Initially, we did not have a clear direction for our machine learning challenge goal. We knew we wanted to incorporate it but could not figure out what we wanted our model to predict. One of

our brainstormed ideas included using machine learning to predict which platform a movie would be available to watch on. The second involved predicting the rating of a tv show or movie based on its qualities. We decided to elaborate on the latter goal. We hoped to train a Decision Tree Regressor on features like streaming platforms to predict the IMDb ratings of a movie. Halfway through the project, we made it our goal to train a model to have a mean squared error value of below 0.1. We successfully did that, training our model on our merged imdb_movs dataset to have a mean squared error value of only a few hundredths. More specifically, our mean squared error value for our test set generally landed between 0.015 and 0.08.

Work Plan Evaluation

Sharing access to source code: 20 minutes

We will share access through github allowing each other to clone the project and work on it simultaneously while we are on a call

Evaluation: This part took longer than expected because we both had to install Python and the libraries with it. However, we were able to clone the project and work on it together once we got everything to install.

1. Merging datasets: 1 hour
    a. Merge imdb extensive ratings to corresponding streaming selection datasets
    b. Merge together but include all rows ie outer join

Evaluation: This part didn't take too long and we overestimated it because our datasets matched with each other well, as the titles of the movies and tv shows were easy to match with the IMDB dataset. We also didn't merge the tv shows and movies as we just used concat to join them together in order to keep all the movies and tv shows for each streaming service without creating new columns.

2. Compare genre variety across streaming sites: 1 hour
    a. Group by sites
    b. Count unique genres and total movies/tv shows

Evaluation: This part took longer than expected because we found out that getting the unique genres for each streaming service was more difficult than expected. Since, each movie had more than one genre, separated by commas, we had to find a way to duplicate the movies and list the genres out so that there is only one genre per row. After doing this, we got the count of unique genres, but we realized that wasn't enough as we got the count of movies for each genre and the standard deviation for the counts for each streaming service. From then, we also had to get the percentage of movies for all the genres for each streaming service and chart a bar chart showing

the percentages and the associated genres. We also charted a box plot of the count of movies for each genre in order to show the average amount of movies of each genre for all four different streaming services.

3. Compare average rating for each site: 2 hours
    a. Convert rotten tomatoes to 0-10 scale
    b. Average rotten tomatoes and imdb rating
    c. Group by sites
    d. Sum ratings and divide by total movies and tv shows
    e. Average the ratings of the films of each streaming services

Evaluation: We were close to accurate on computing average rating as it had to do with pandas and finding the right filters. We were able to do this task within the time we estimated.

4. Graph: 1 hour
    a. Bar chart displaying each streaming service and their average film rating
    b. Pie chart for each genre percentage from the different streaming services

Evaluation: This part was difficult and took longer than expected because the pie chart was hard to look at due to the high amount of genres for each streaming service. We tried to manipulate the pie chart so the numbers and labels can be read but it was too close to each other. As a result, we made a box chart of the count of movies for each genre and a bar chart of the percentages of movies of each genre for the different streaming services.

5. Machine learning: 2 hours
    a. Model training with labels as the streaming services and the features as the tv shows or movies and their genre and rating
    b. Test for accuracy

Evaluation: We found that training with the labels as the streaming services to be more difficult than expected as it is hard to predict the streaming service that a specific show or movie is on. As a result, we changed up the training for the title, genre, and streaming service, and the labels as the ratings of the movie. This resulted in a longer than expected challenge but we were able to accurately predict the ratings of the movies that were given to the model, allowing the user to predict ratings of future shows or movies.

Testing

To make sure our results are correct, we ran tests on multiple parts of our code. For our genre analysis, we tested if our code worked with splitting the genres up and duplicating the movies. Since each movie has multiple genres, we had to duplicate the rows of movies and split the

genres in order to have one genre for each row. To test this, we obtained a small dataset of our current movie and tv show dataset. From this small dataset, we can easily count the number of unique genres and we tested if our code would produce the same result. We found that the code was accurate in splitting up the genre list for each movie and counting the number of unique ones. This allows us to continue analyzing the genre and diversity of each streaming service as we were able to easily get the count of movies for each genre as well as the percentage. With the test, we were able to prove that our code is correct as it was able to split the list of genres correctly and create the list of unique genres.

Additionally, we also tested the amount of movies and tv shows for each streaming service by employing a small dataset that is part of the big dataset. We were able to individually count the number of tv shows and movies for each streaming service on the small dataset. The test code ran without any issues and accurately produced the number of movies and tv shows for each streaming service. This proves that our code was correct in splitting the content of each streaming service in order for us to further analyze the average rating of movies and tv shows for each streaming service. The average rating was much more difficult to test but we were able to individually input the small dataset to the average rating and found that the number that was produced for each streaming service is correct. As a result, we tested our data and code fairly often and made sure that the results on this report were correct and accurate.

Collaboration

For help with this assignment, we didn't ask any other students or teachers for help. We were able to collaborate easily and help each other out as we worked together on a Zoom call. If anyone of us ran into an issue, we would ask each other and help each other figure it out. If we weren't able to figure out the issue, we would search the internet for any help possible. We utilized the python documentation website as well as w3schools.com for information on useful methods and libraries. During our genre analysis, we weren't sure how to duplicate the movies and have one genre per row. However, we were able to do a google search and found the explode function in order to manipulate the dataset as we desired it to be. Besides this one instance, we didn't ask for any other help as we were able to work on the project together with good teamwork and collaboration from each other.