Iteration 2

Team Stream

Group members: Huy Tran, Kelsey Kua, Kenneth Ven

Overview

Online video streaming is commonly used by people of all ages. One of the first big streaming platforms was Netflix, and after their success, other new streaming services popped up. Now there are several services to choose from, and it is not uncommon for an individual consumer to subscribe to multiple streaming services.

In this project, we will be exploring the availability of movies on the popular streaming platforms Netflix, Prime Video, and Hulu.

Our Approach

The question we want to answer is the following: **What kind of movie content does each major streaming platform tend to feature (e.g. genre, language, age group, year)?** To go about answering this question, we will be using public datasets found online. The current datasets that we will use can be found here from Kaggle (last updated: 5 months ago). This report explores the accuracy and intent of the current data.

From this dataset, we will examine the following variables:

1. Streaming platform
    a. Netflix
    b. Prime Video
    c. Hulu
2. Movie year
3. Age rating
4. Movie rating
    a. IMDb
    b. Rotten Tomatoes

With these variables, we hope to identify content patterns within each platform to provide a high-level summary of movie content. The features we can control include the streaming platform, movie year, and age rating. For example, we can see how the variety of movie year and age ratings differ by platform. The features we cannot control are ratings (IMDb and Rotten Tomatoes), which have been determined by users of these websites.

Data Pre-Processing

To process that data, we took the database on Kaggle and separated the database by its streaming platform to create three excel files (Netflix, Hulu, and Prime). Below are the steps we took to filter the data:

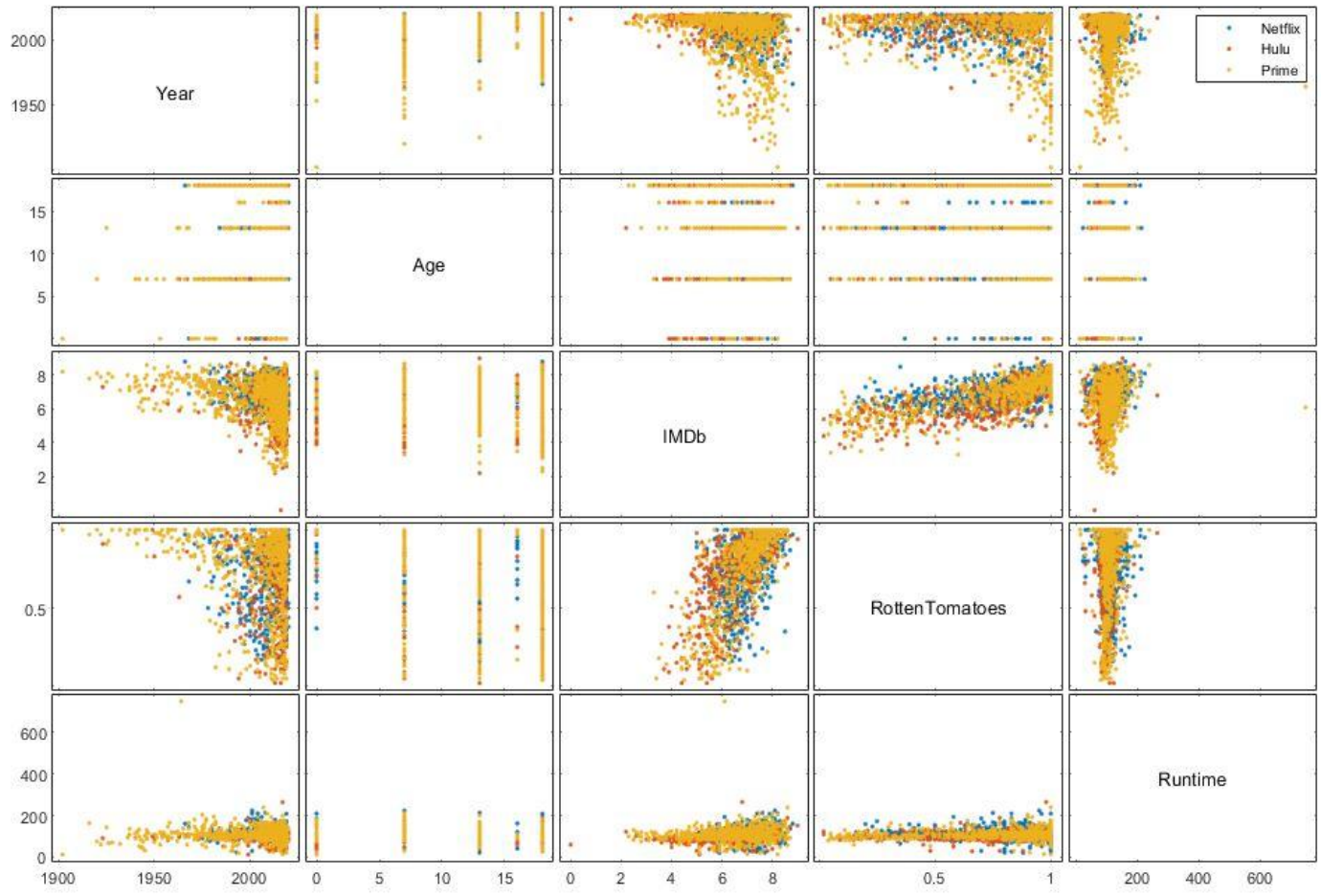
1. Delete the first column, ID, and Type
2. Filter the streaming platform that you want by making it be 0. This would create a sheet that contains movies that are not on that platform.
3. Delete the current sheet
4. Delete all the streaming platform type
5. Filter genres and delete all blanks
6. Filter age, IMDb, and runtime together to delete all blanks.
7. Freeze top row
8. Change age and IMDb to be recognized as a number

If you follow these steps, then it should recreate our current three excel files for each platform. Our goal for data preprocessing was to filter the original data by streaming platform and delete any information that was not needed.

Data Plots

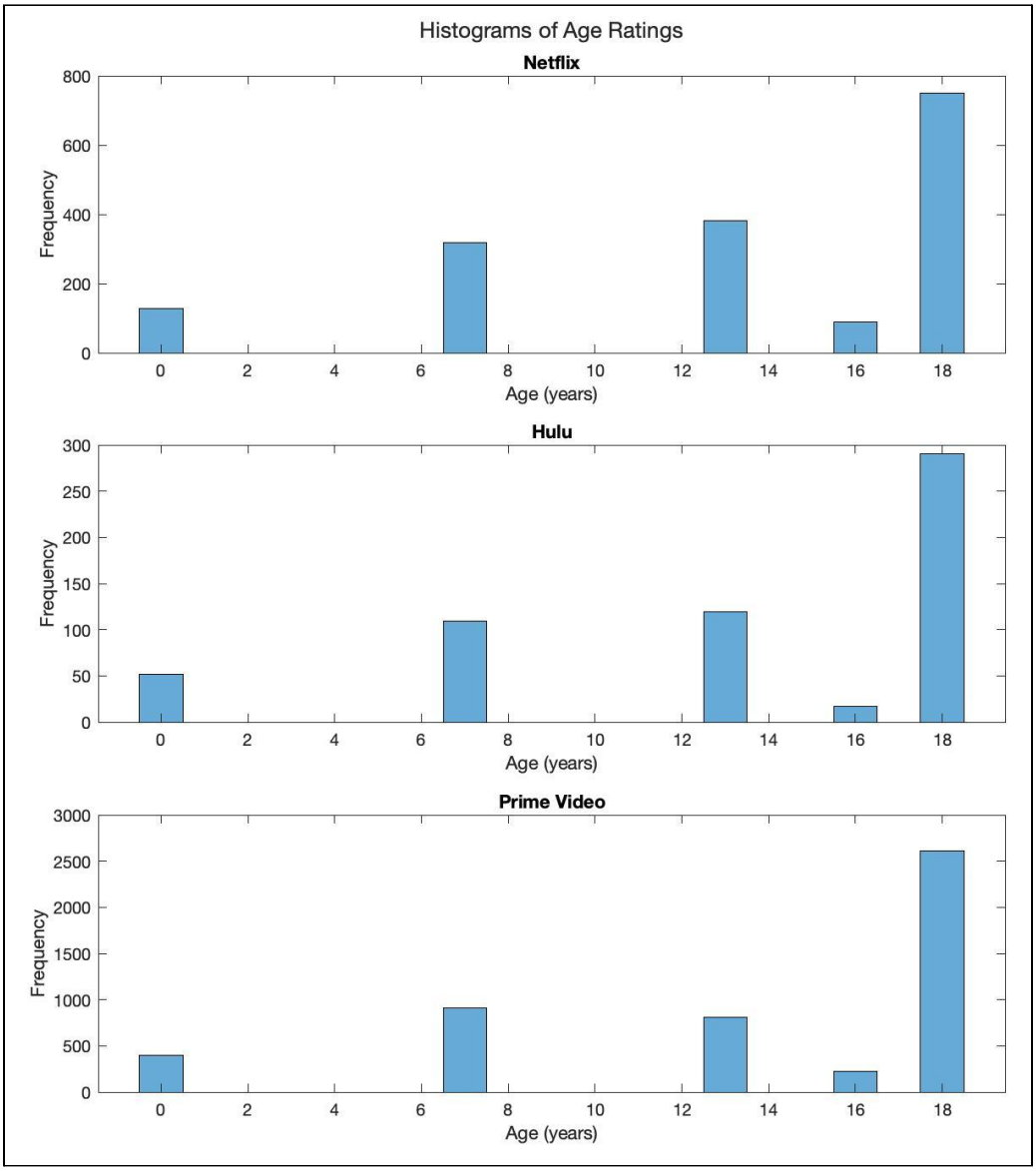
Of the plots we have learned about, we think a scatterplot matrix and histograms are the most useful. We also plotted normal distributions of each feature for each class to determine if we could assume normality. These plots are described in the following sections.

Scatterplot Matrix

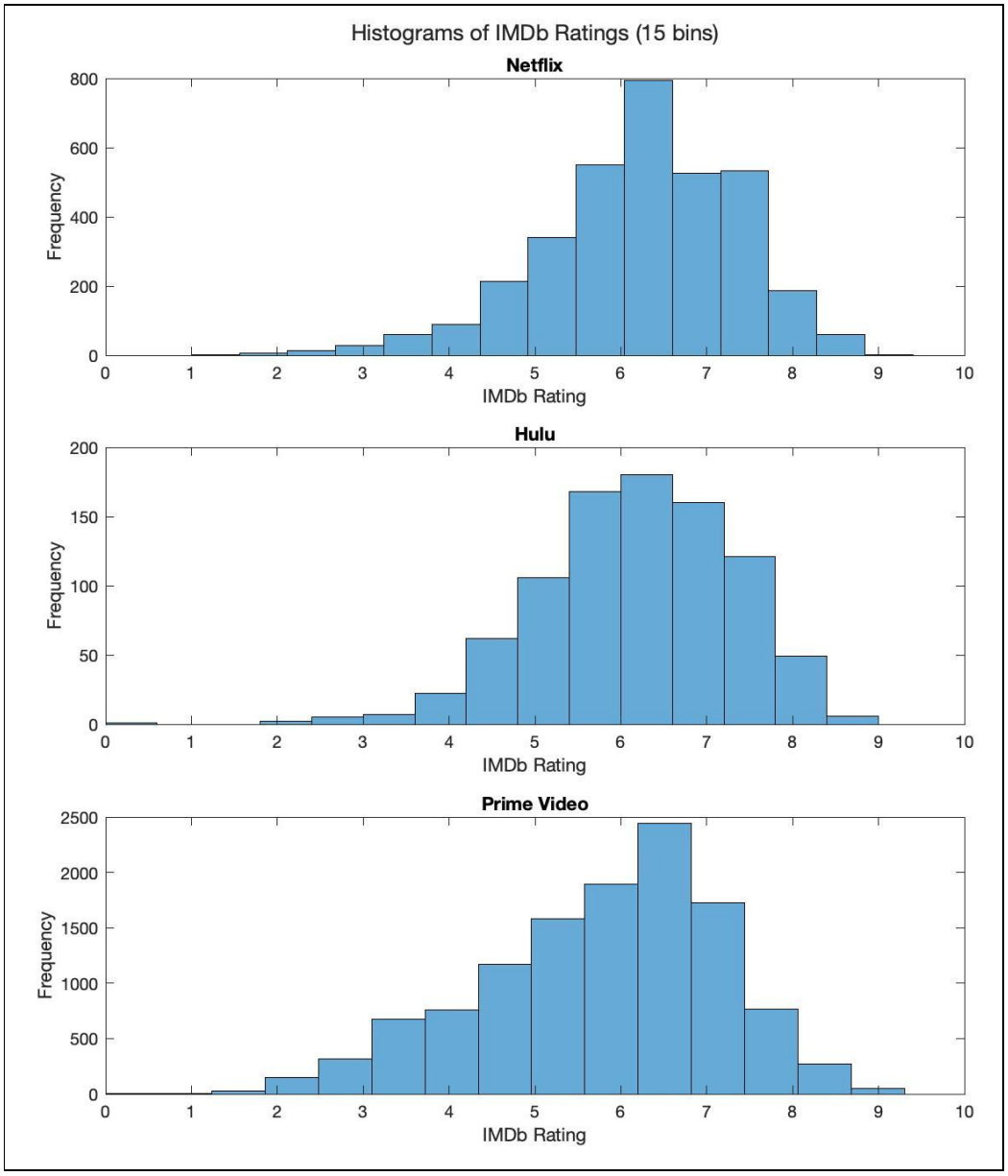


The Scatterplot Matrix shows that Netflix, Hulu, and Prime have a heavy correlation with each other. However, the reason that there is more data for Prime is that there is less missing data compared to the two other streaming platforms This resulted in more Prime data appearing on the matrix. It seems that primes stretch further compared to the two other platforms as it covers more of the space, while Hulu and Netflix seem to be grouped at. This lets us assume that Prime video seems to cover a variety of movies, while Hulu and Netflix like to cover a specific type as shown when observing the relationship between Year. Also, it seems that all three platforms do not care about the Rotten Tomatoes and IMDb ratings as the plot stretches far.

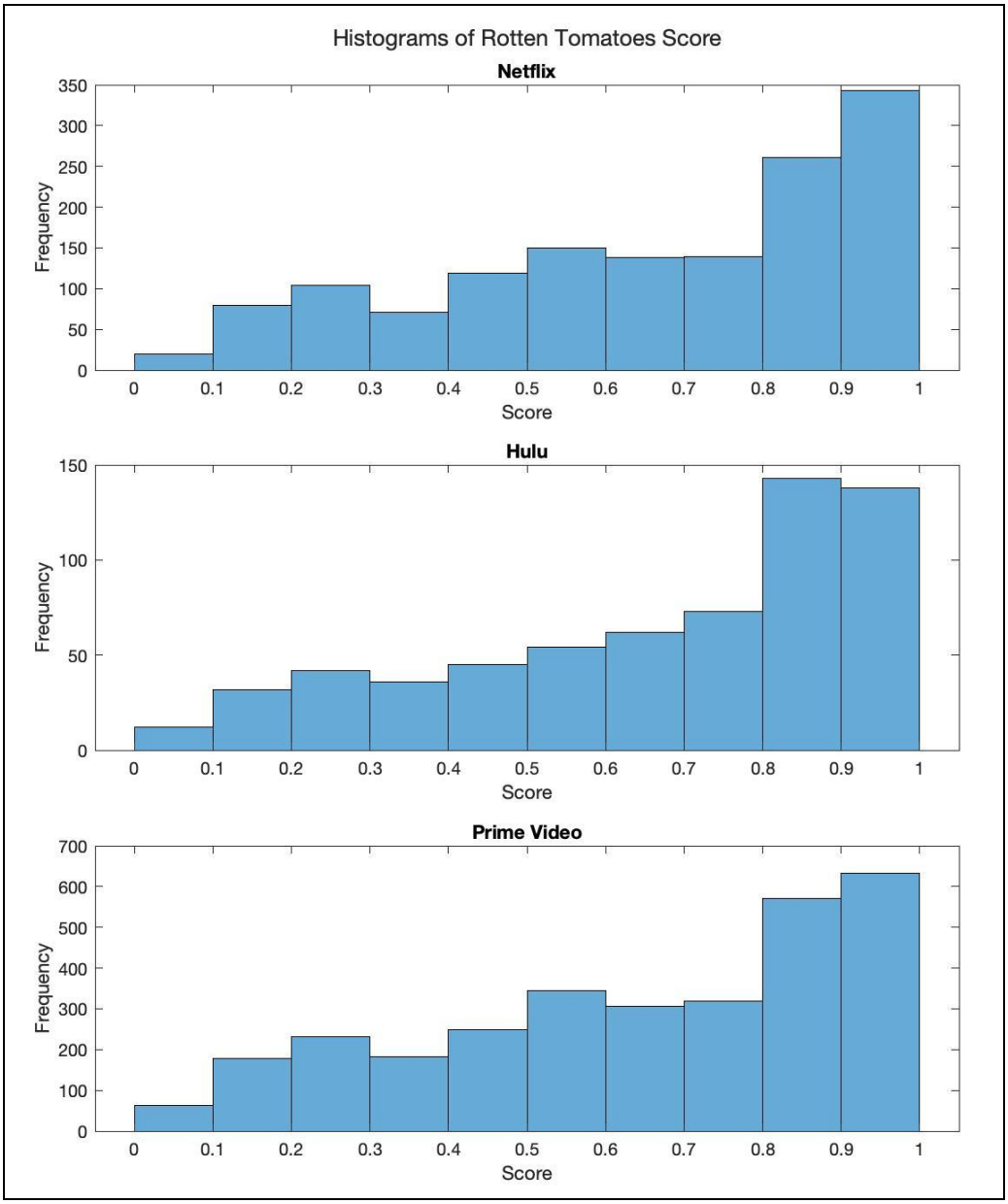
Histograms



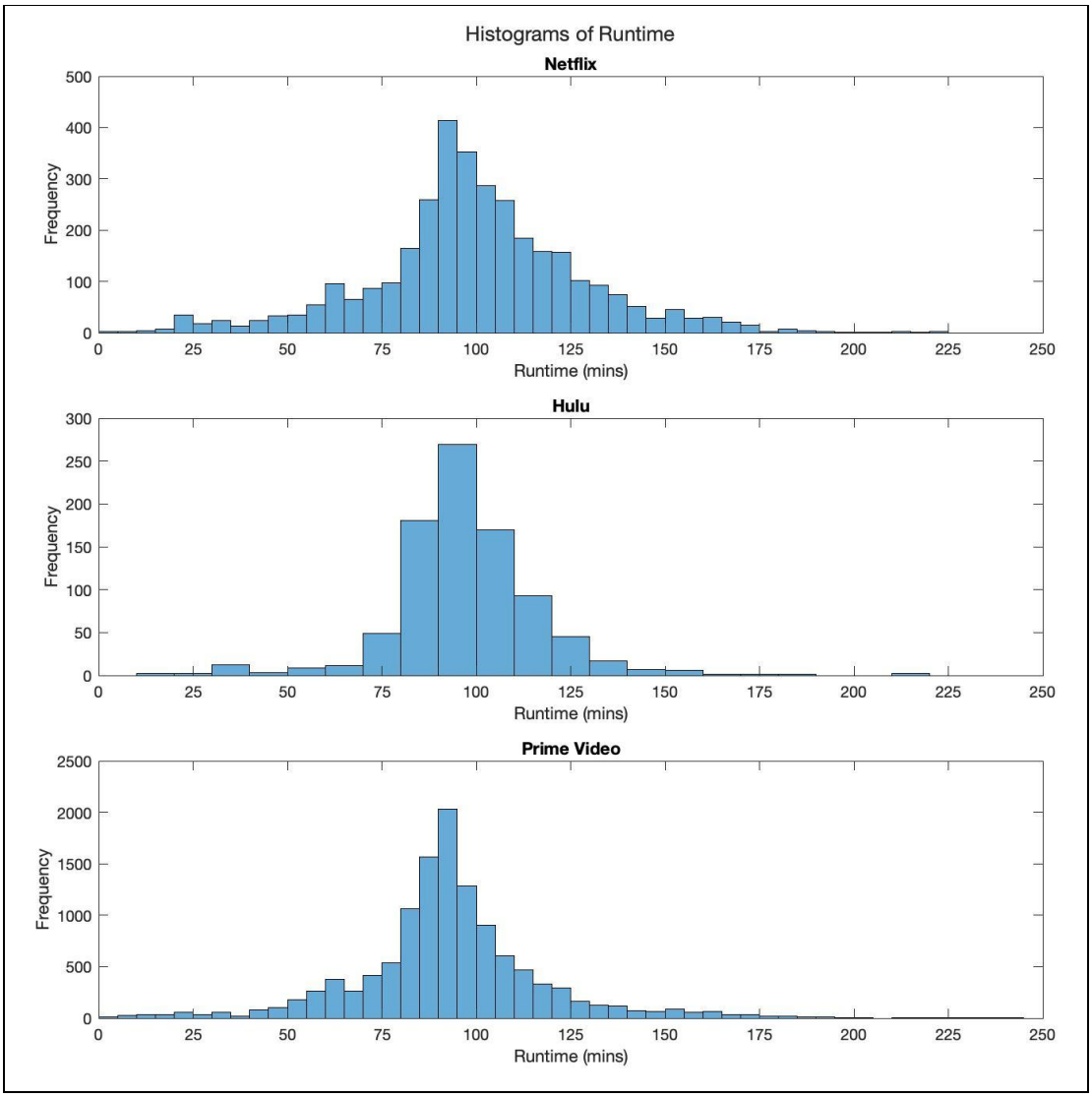
The five age rating categories of films are as follows: all ages (which we have corrected to be 0+ years), 7+ years, 13+ years, 16+ years, and 18+ years. The histograms for each streaming platform look similar: The older the rating, the higher the frequency. Each platform has the greatest frequency of films in the 18+ category, meaning they tend to feature films for more mature audiences.



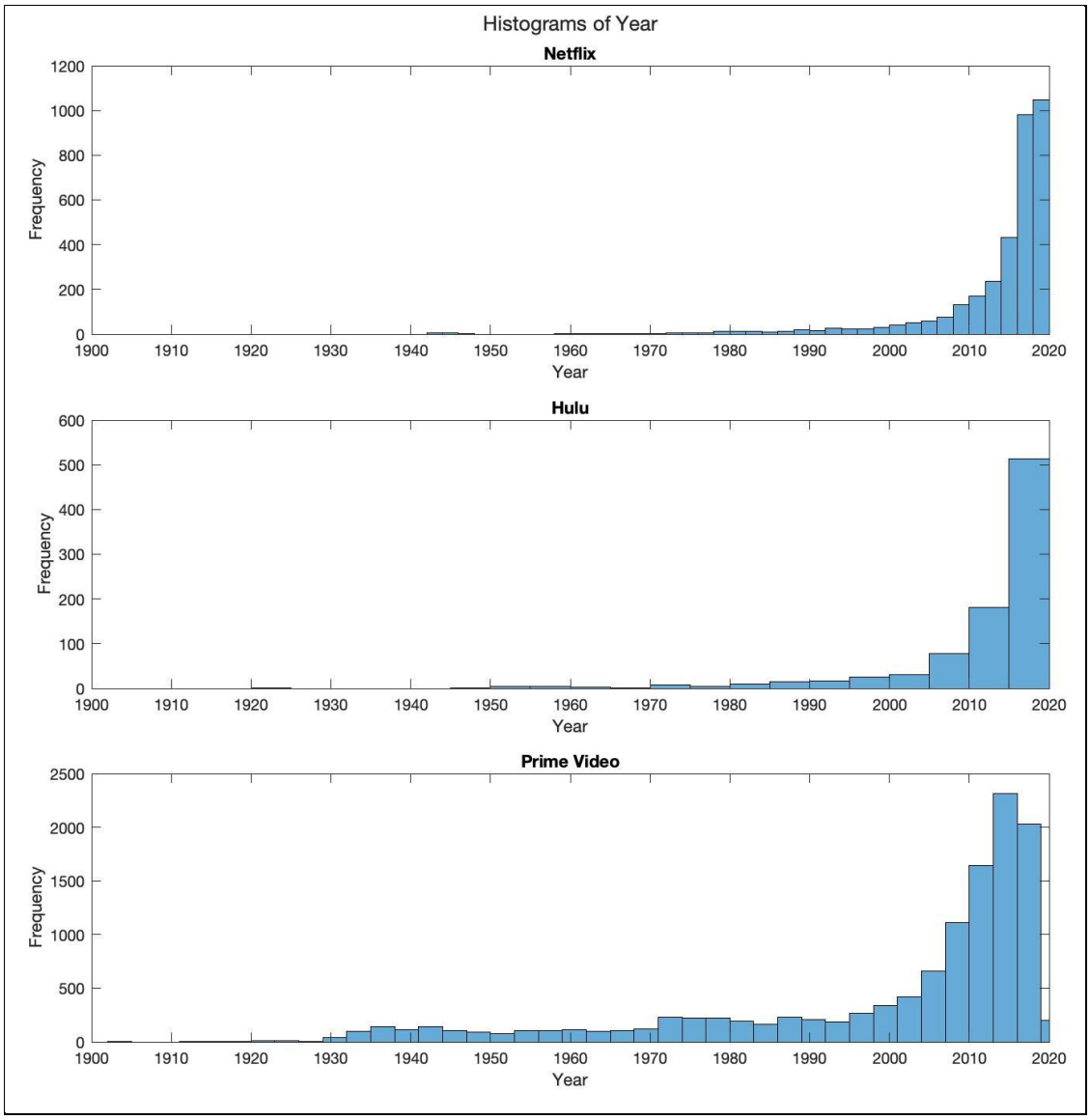
These histograms show the spread of films' IMDb ratings. For each platform, the majority of films have ratings between 5.5 and 7.5, which seems like a decently high rating. If we knew how films tend to be rated on IMDb (e.g. well-received movies tend to be rated between X and Y, mediocre movies tend to be rated between X and Y), we would have a better feel for the quality of films on each platform. To provide more context to the rating systems themselves, IMDb ratings can be contrasted with Rotten Tomatoes scores.



These histograms show the distribution of films by Rotten Tomatoes score. An interesting part of this plot is that, for each streaming platform, the frequency of films stays somewhat consistent below 80%. A high frequency of films has scored between 80% and 100%, which indicates that they were very well-received. In contrast with the IMDb ratings, Rotten Tomatoes scores tend to be higher. Perhaps this means that highly regarded films have 80%+ approval ratings on Rotten Tomatoes but have lower IMDb ratings (between 6.0 and 8.0).



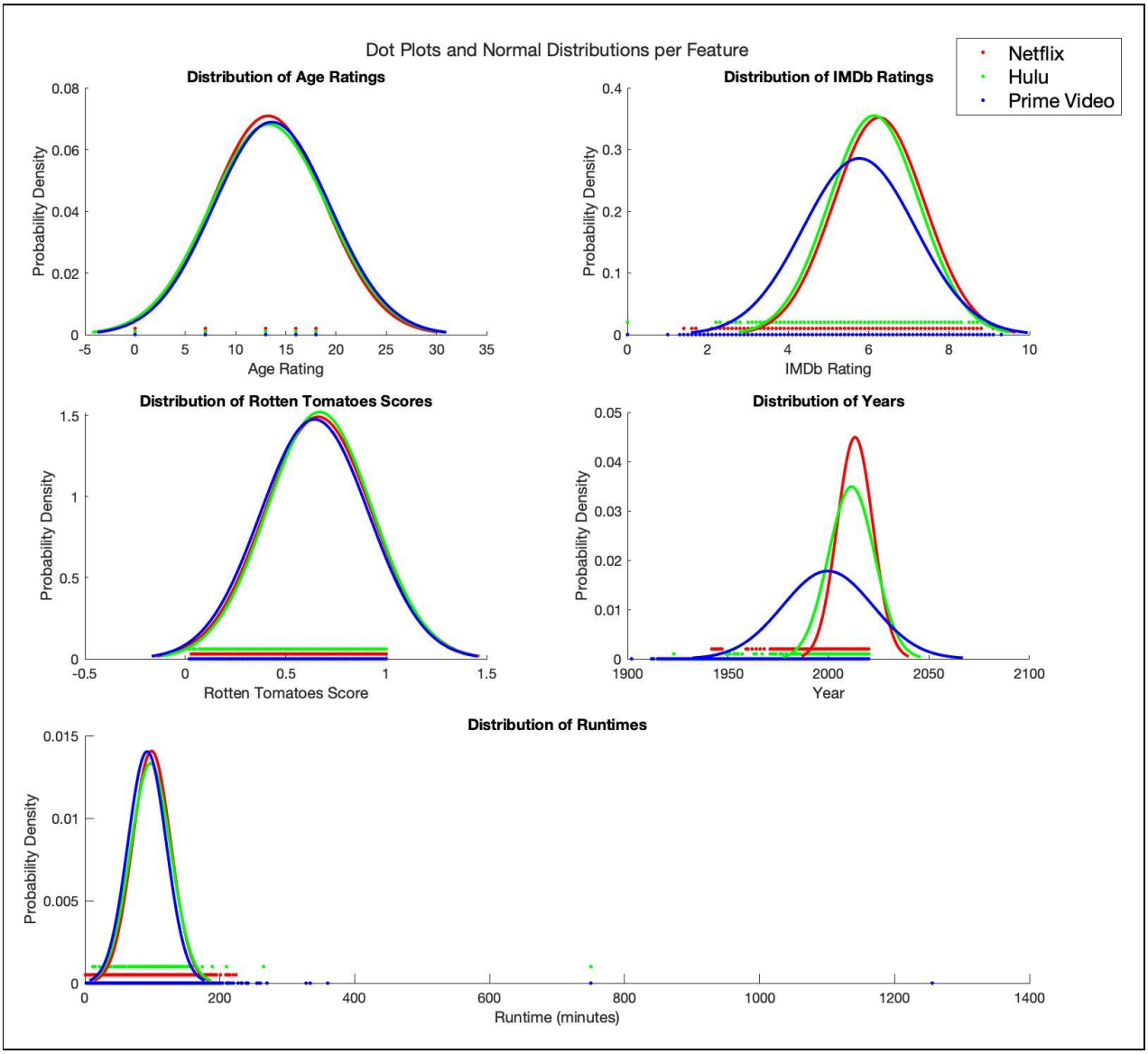
These histograms show the distribution of films' runtimes in minutes. Prime Video had films beyond the 250-minute mark, but we set the x-axis limit to be 250 minutes to provide fair comparisons among the streaming platforms. The center of the distribution seems to be just under 100 minutes, which is between 1.5 and 2 hours. From personal experience watching films, this is not surprising. For Netflix and Prime Video, however, there are quite a few films past the 2-hour (120 minutes) mark, which is longer than most casually viewed films today.



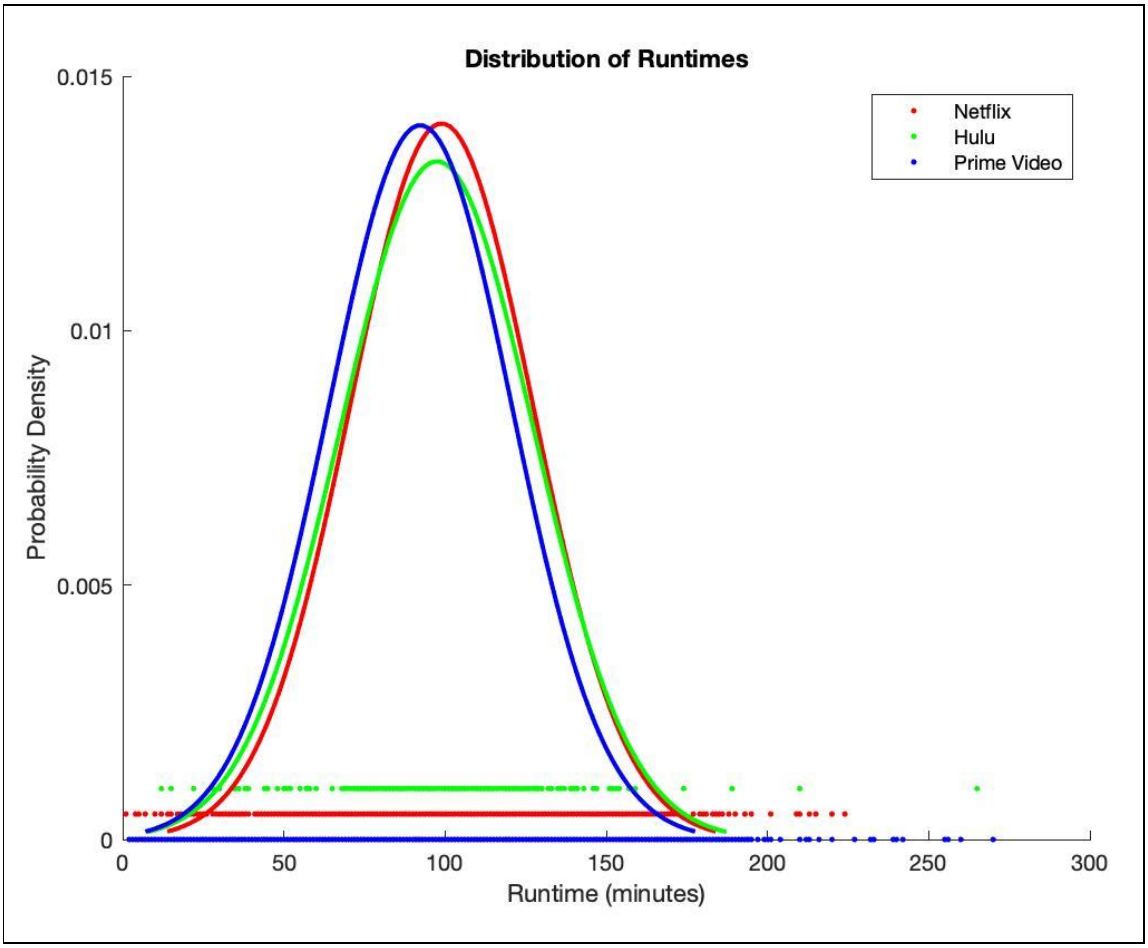
These histograms show the distribution of films by year released. On each platform, a heavy concentration of films was released between 2015 and the present. Compared to the other two streaming platforms, Prime Video features significantly more films between 1900 and 1980. This occurrence may be attributed to the fact that Prime Video has more observations in the dataset than the other two streaming platforms.

Normal Distributions

To determine if we can assume normality, dot plots and normal distributions were made for each class and each feature. Because the values of the dot plots often overlapped, the values of only one class were visible. To make all values visible, the dot heights of the classes were adjusted. Dot plots, however, are one-dimensional and have no value on the y-axis.



For almost every feature, the normal distributions of each streaming platform line up almost perfectly. This makes sense for the distribution of age ratings because there are only five possible values (0+, 7+, 13+, 16+, 18+), which can be seen in the dot plot values. This also makes sense for IMDb ratings and Rotten Tomatoes scores because the histograms for these features were very similar across all three platforms. The normal distributions for release year, however, have the most variation: Netflix and Hulu have similar distributions, but Prime Video's distribution is wider and centered about 15 years earlier. According to the histogram of the release year, Prime Video has a sizable number of older movies (between 1900 and 1980) compared to the other two platforms. Another interesting feature is runtime. According to the dot plot, Prime Video supposedly has films that run at 750 minutes and 1250 minutes. Including these points forces a zoomed-out view of the distribution, so we limited the x-axis to values between 0 and 300 minutes to get a clearer view.



With this clearer view of the runtime distributions, we see that the distributions for each class line up almost perfectly.

From these dot plots and normal distributions, it is reasonable to assume normality for all features except release year. Prime Video's distribution for the release year may differ from that of the other two platforms because Prime Video had ~12,000 observations compared to Netflix's ~3,500 observations and Hulu's ~900 observations.

Independence of data observations

We are assuming that each class(Netflix, Hulu, Prime) has multiple related movies that they have in common. For example, Netflix and Hulu will have IP Man 3.

For this reason, there isn't independence between the observations. This is because the comparisons of the movies depend on whether or not the other platform streams the movie. If the streaming platforms were to have all unique movies compared to each other then we could assume that our observations are independent of one another.

However there will be occasions where some platforms have unique movies, these observations could be assumed as independent because there are no other observations to compare it with and are not dependent on other factors.

Other factors to be considered when figuring out if we can assume dependency of observations:

- If the movie is on the same platform as the other, then the specific movie is dependant on the other one(s) from the different platforms
- The rating of the movie could be dependent on which genre the movie is categorized as. This is assuming that some people like horror movies more than comedy movies and will rate them higher because of it.
- The runtime could be dependent on the year the movie was made. We are assuming that most movies made in the 1900s have shorter runtimes than movies made in the 2000s.
- The runtime of the movies can be dependent on the age range of the movie. We are assuming that the duration of kid movies tend to be shorter than adult movies.

However, some factors can be independent of others. For example, the ratings of the movies are independent from the runtime of the movie. We are assuming that people don't base their ratings on how short/long the movie was. Another factor that we can observe is that ratings of a movie is independent of the year the movie was made. We are assuming that people don't base their ratings on how old the movie is.

In conclusion, even though there may be a few factors that could make our observations independent of one another, we assume that most of the observations are dependent on each other and their factors.

Team Assessment

Kelsey - I revised the overview and approach from the previous iteration. I created the histograms and their analysis, as well as the dot plots/normal distributions and their analysis.

Kenneth - I mostly worked on the analysis for the independence of our observations. I also helped my teammates look over the graphs to make sure it's an accurate representation of what we want to show.

Huy - I cleaned the original dataset to become three excel files that represent each streaming platform. I created the scatterplot matrix and wrote the analysis for it. I look over the document for any grammar mistakes.