Exploring the top 250 and bottom 100 movies of all time according to IMDB users

Jack Agren and Emily McBride
A02267469 and A02291674

Introduction

In this project, we looked at the top 250 movies of all time and the lowest rated 100 movies of all time as rated by IMDB users. The data (found here and here) has basic information like number of reviews, average rating, release year, and duration. We analyzed how these affected a movie's place on the two lists, and how this information might affect other information such as how the length and number of ratings affects a movie's star rating. For our analysis we used averages, means, t-tests, Pearson Correlation tests, scatterplots, and Kernel Distribution Estimation Plots to find any possible correlations and make observations on the data we found. We were able to find possible correlations for the top 250 movies length and their star rating, the movies length and whether it was in the top 250 movies or the lowest rated 100 movies, and the number of reviews vs the average rating. We also found that movies with different MPAA ratings do not have different average reviews.

Link to Presentation
Link to GitHub Repository

Dataset

For our dataset, we used BeautifulSoup and the requests library to get web content from IMDb's top 250 movies (here) and lowest rated 100 movies (here). To gather the data itself, we had to find the tags and class names specific to the content we were looking for, extract the contents of the tag, and pass the list of information gotten from each of the tags found to a DataFrame. Each DataFrame (for the top 250 and bottom 100 movies) contains the name of the movie, the year it came out, its length in minutes, the MPAA rating, the number of stars it's received through reviews, and the number of reviews. These pieces of information are relevant for analysis because they can tell us about how people view these movies and why they are placed on the charts in the location they are.

Analysis Techniques

For this project, we used many techniques to analyze this dataset. We used averages and means to figure out the most common movie length and star rating for both datasets and be able to compare them against each other. The t-test was used to compare the means of the top 250 movies and the bottom 100 movies. The Pearson Correlation test was used to find a possible correlation between the length of a movie and its star rating. Scatter plots and KDE plots were used to visualize the data and find possible correlation. We used a t-test to test differences between MPAA ratings, and a correlation test to test if movies with more reviews are rated higher.

Results

For the first analysis, we looked into whether or not the runtime had a significant difference between the top 250 movies and the lowest rated 100 movies. First we found the mean runtime for both charts which resulted in the top 250 movies having a mean runtime of 130.184 minutes and the lowest rated 100 movies having a mean runtime of 99.73 minutes. Next, we performed a t-test to compare the mean runtime of both charts which resulted in a statistical value of 8.905 and a p-value of 3.016e-17 which would indicate a difference in the mean between the two charts. A KDE plot was also used to visualize the distribution of different movie lengths (see **Figure 1** below). This information could be used to inform movie producers on how long their movie should be. An additional possible reason behind these results, by viewing these results and the movies these results are associated with, runtime may not be the only factor. It could be that more thought, detail, and higher quality content was put into the longer movies.

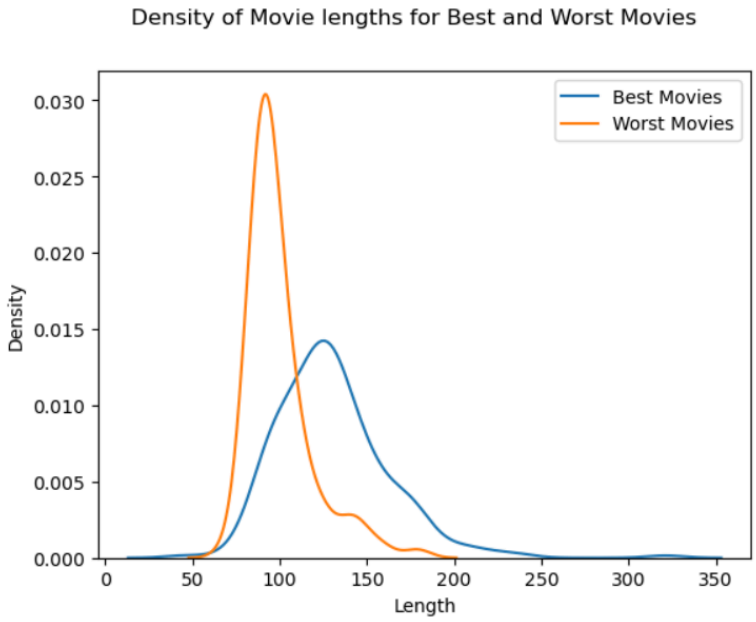


Figure 1: Density of movie lengths for best and worst movies.
y-axis = density, x-axis = length in minutes.

For our second analysis, we looked into how the length of a movie might affect how many stars it receives. First we found the average runtime and average star rating for the top 250 movies and the lowest rated 100 movies, this resulted in runtimes of 130.184 and 99.73 minutes respectively, and star ratings of 8.311 and 2.861 stars respectively. Then, we performed a Pearson Correlation test for the best movies and another test for the worst movies, resulting in statistical values of 0.2745 (best) and -0.0959 (worst), and p-values of 1.065e-05 (best) and 0.3424 (worst). This would indicate that the top 250 movies had a positive possible correlation and we were able to reject the null hypothesis for these movies runtimes and star rating. However, the lowest rated 100 movies did not have a significant correlation and we were not able to reject the null hypothesis. These results were also shown to be true through a scatterplot for both charts (see **Figures 2** and **3** below).

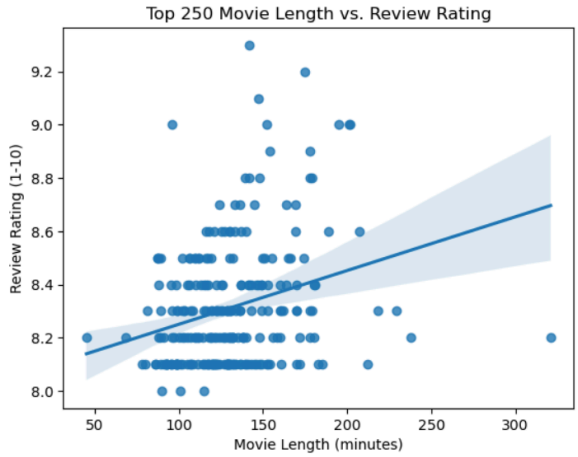


Figure 2: Top 250 movie length vs. Star Rating.
Y-axis = Star rating, x-axis = movie length in minutes

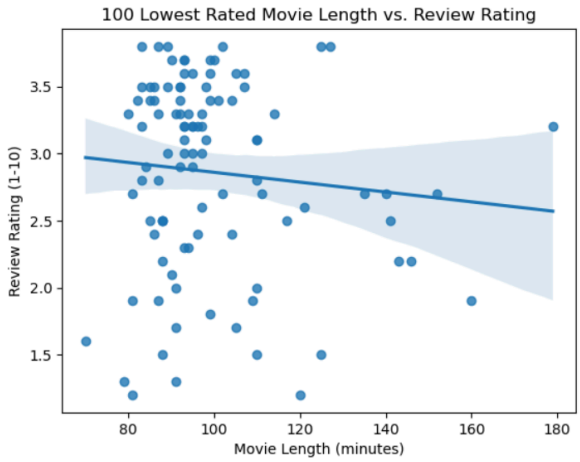
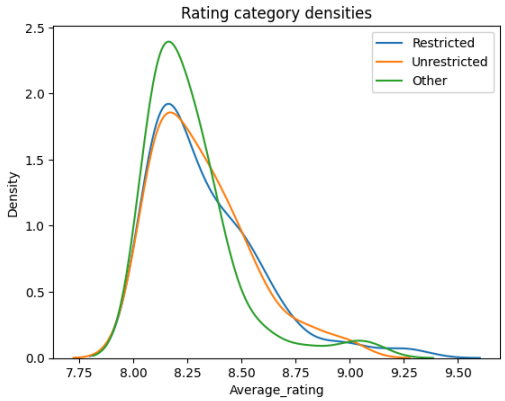
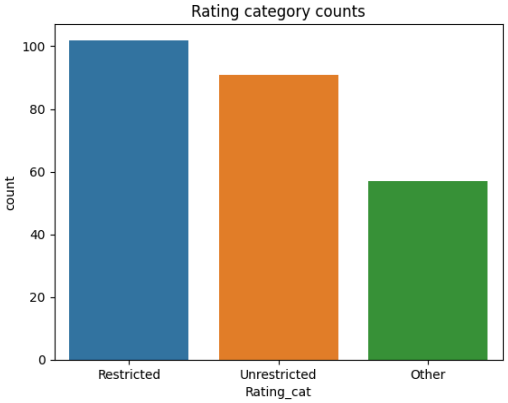


Figure 3: 100 lowest rated movie length vs. Star Rating.
Y-axis = Star rating, x-axis = movie length in minutes

For the 3rd analysis, we used a t-test to see if the average user rating was different between MPAA ratings. Since there were many more rated R movies than any others, we grouped the movies into ‘restricted’, ‘unrestricted’, and ‘other’, where restricted indicates restrictions on who can enter a movie, unrestricted means no restrictions on entry, and other contains foreign films and films that were rated before the modern MPAA rating system. We elected to not analyze the ratings we put in the ‘other’ category since ‘Not Rated’ or ‘Unrated’ could contain a wide spectrum of content. We also found that, interestingly, there are more restricted type movies in the top 250 list; however, after a t-test (p = .10), we found that the MPAA rating does not affect the average rating of a movie *within the top 250 list.* It is possible a difference would be found in a longer list.





The 4th analysis was a test of correlation between number of reviews and average rating. There was a significant correlation (statistic=0.57, p-value=1.63e-23) between the 2, but, as seen in the plot below, this is likely strongly influenced by the movies that have the most reviews at each user rating. It is also limited by the rough estimation of the number of reviews provided by top 250 list. The number of reviews was given by a string like “2.5M” or “684K”.

Technical

While preparing the data, we had to use BeautifulSoup and the requests library in order to get our data. We then had to find tags and class names of the content we were looking for and extract that content from the tags. After this we could then put this content into a DataFrame. Many of the number values had to be converted to floats for analysis against each other. For the movie lengths, we had to convert the times from being in the format of “2h 30m” to the total number of minutes, “150”, that way we could use this information in a more meaningful way by comparing it against other values.

For our analysis, we used means and averages to visualize the differences and relations in our two datasets. We also used t-test to measure the correlation between two groups, being our two datasets, and p-tests to measure correlations between different categories within the same group, such as movie length and star rating. Scatter plots and KDE plots were used to visualize our data and see the relations between our datasets or the relations between categories

During our analysis process, we started out trying to do natural language processing on the comment reviews for each movie in the top 250 movies. This proved to be very difficult and required more time than we could give to the project. We then moved onto our current project idea. This process involved getting and cleaning our data and figuring out ways to analyze it in a meaningful way. We decided on our analysis techniques because they seemed relevant for the data we were using and, on the visual side, would be able to tell the story that the analysis provided.