Mini-Project 4 – Analyzing Data Sets

Zachary Moser

November 22, 2024

Abstract

In this project we will look at two data sets of iMDb movie ratings and rotten Tomato movie ratings. We will do this by creating different types of graphs in order to compare trends and differences in the data. The graphs will be created in python and then studied in order to make conclusions about the two data sets.

1 Introduction

Throughout this paper we will look at different graphs and tables. We will compare data from the iMDb data set and the rotten tomatoes data set separately and then merge the data sets and look at how the movies that are in both sets compare in ratings of critics and audiences. In order to compare these sets we will use two-way tables, pie charts, box plots, bi-variate histograms (heat maps), bar graphs, and histograms. We will strategically choose each type based on what kind of data we are comparing and how many things we are trying to compare at once.

In the following sections we will look at the each step in further detail and interpret the graphs and what they tell us about the data.

2 Problem Statement

The goals of this project are to compare movie ratings from rotten tomatoes and iMDb. We will also find what type of movies the critics prefer compared to the audience. This will allow us to better sort data into ways that we can easily interpret patterns and and interpret the data.

Our first problem will be to compare different categories inside the iMDb file to see trends over a large number of movies. The data from this file will be the most accurate because it has the most movies in it. When the two files are merged it we will only keep the movies that are in both files which will make the averages less accurate. For this comparison we will look at some box plots of audience scores and critic scores for each certificate. This will be followed by a heat map of runtime and scores. This will allow us to see how movie length effects the reviews. After that, we will look at a pie chart that will break down the genres and a bar graph that will look at certificates based on how often they occur. Finally, we will look at a histogram of critics reviews vs audience reviews for iMDb only. The iMDb file doesn't have all the categories that the rotten tomatoes data file has so we will also do some comparisons for the rotten tomatoes as well. These comparisons will consist of finding the most popular studios by way of a bar graph. We will complete our examination of the rotten tomatoes data set by looking at a histogram of critics reviews vs audience reviews.

Once we have compared both files separately, we will merge them to see how the reviews from both sites rate movies differently. We will complete this task by looking at a two-way chart of the different scores and statistics about them. More specifically, we will analyze the count, mean, standard deviation, maximum, and minimum for each scoring system. We will also look at a chart that will compare genre to rating across all four rating categories. Lastly, we will view a histogram of the iMDb critic reviews compared to the rotten tomatoes critics reviews.

3 Methodology

In order to merge the data set and build our graphs and charts we will use built in python commands such as: pandas for merging files and ".boxplot" for creating a box plot. The contents of each graph and type will be chosen so that the data is clear and easily deciphered.

4 Results

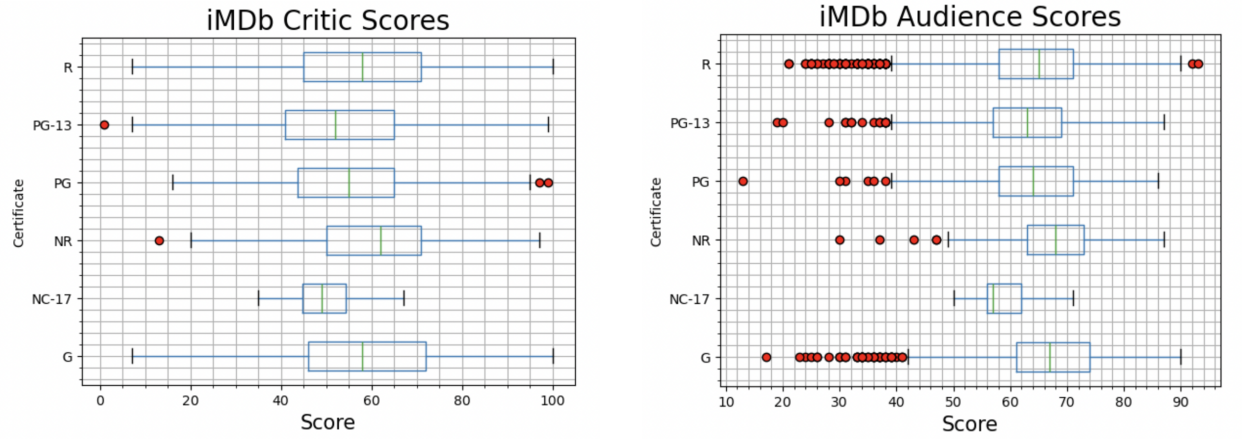


Figure 1: Boxplots with iMDb scores for each certificate.

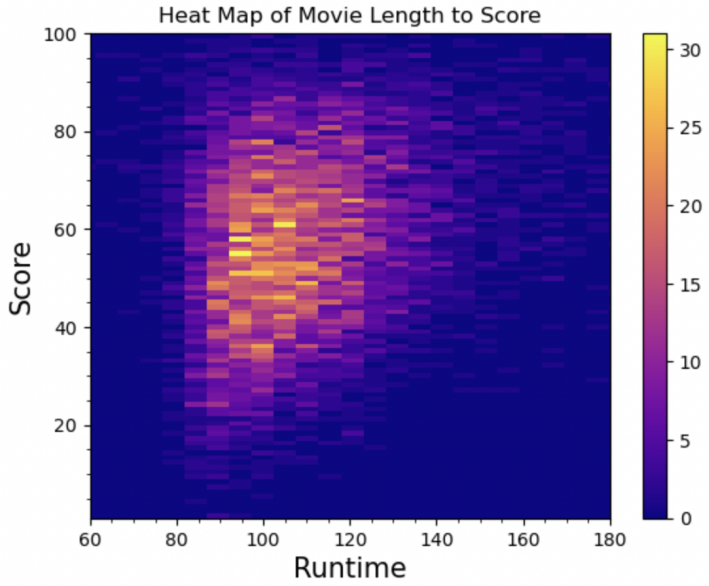
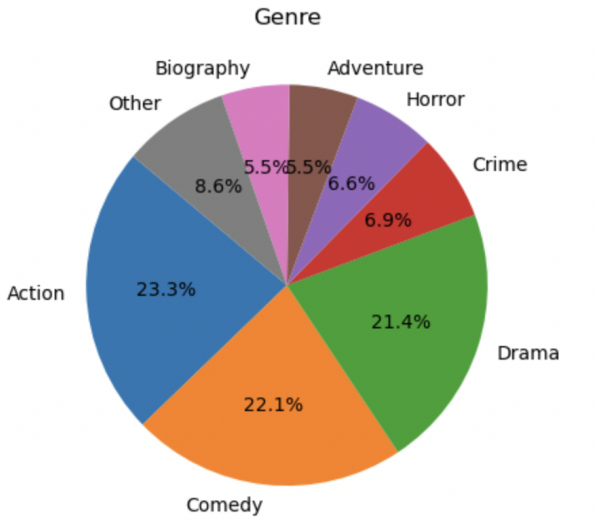
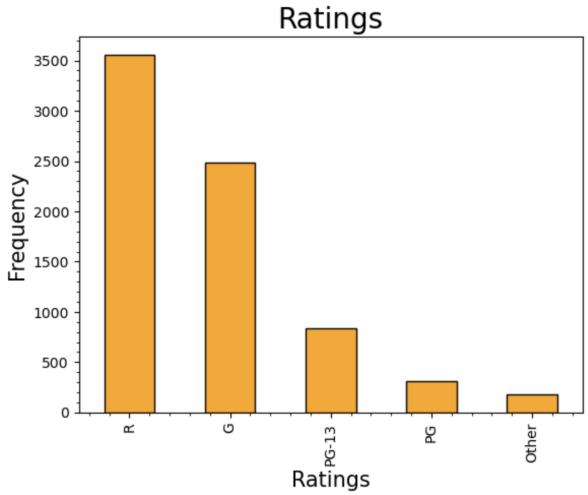


Figure 2: Heatmap of runtime vs. iMDb critic scores.



(a) Pie Chart of frequency of each genre.



(b) Bar graph of frequency of each rating.

5 Discussion

We will begin by talking about the iMDb data. Our first figure, 1, is a set of boxplots that look at iMDb critic scores and audience scores respectively. By looking at the iMDb Critic Scores box plot we can see large standard deviation and few outliers for every rating. We can exclude NC-17 from this observation because it has so few entries. This can be seen in 3b where NC-17 resides in the other category due to it's quantity of entries. Shifting focus to the Audience box plot, we can see many more outliers and less standard deviation. This is an observation between critic scores and audience scores that we will keep in mind later when comparing reviews.

In figure 2 we can see a heat map of runtime vs. iMDb critic scores. We can see that most movies have a runtime between 80-140 minutes. We can also see that the scores are pretty well spread out with the highest peaks being around a score of 60 for a movie that lasts around 100 minutes.

In figure 3a and 3b we can see a pie chart and a bar graph. The pie chart shows the frequency of each genre. We can see that the most common genres in the iMDb movie set are action, comedy, and drama. Similarly, we can see the most popular ratings are R and G from the bar graph.

Our last graph for iMDb can be found under figure 4. This is a histogram of the frequency critics and audience members put each rating down for a movie. The audience members frequently rated movies between 60 and 75, while the critics were slightly more balanced from 1-100 with their most popular rating being around 60. We will continue to compare critics and iMDb reviews in future graphs.

Moving on to rotten tomatoes data takes us to figure 5a and 5b. Here we can see a nice bar graph on the left of studio distribution. This shows us that most of the rotten tomato movies came from netflix which is an interesting side note. More importantly on the right in b, we can see a histogram of rotten tomato reviews. Similarly to the histogram of iMDb reviews we looked at we will compare critics' reviews to those of audience members. We can see the critics for rotten tomatoes are very nice in their reviews and have many above 80. We can recognize a correlation between their scores and the certified fresh chart. They might rate these movies highly in order to keep more movies certified fresh and therefore get more people to watch those movies. We can also see that the audience is evenly spread out with peaks on either side of 60. This is similar to what we saw in the histogram of iMDb scores in figure 4.

Finally, we will look at some data that compares rotten tomatoes to iMDb reviews. Looking at figure 7 we can see a bar graph with the average score of each rating for each category of review we are inspecting. We can see that G has the best reviews across the board with it's highest averages coming from the rotten tomatoes data while PG-13 is more well liked by the audience.

In our final figure, 8, we have a histogram where iMDb critics' reviews are compared to rotten tomatoes critics' reviews. We can see that rotten tomato critics give higher reviews on average. As

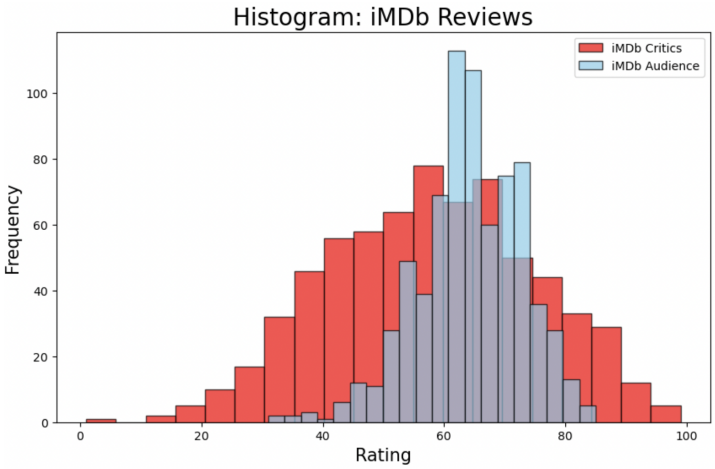
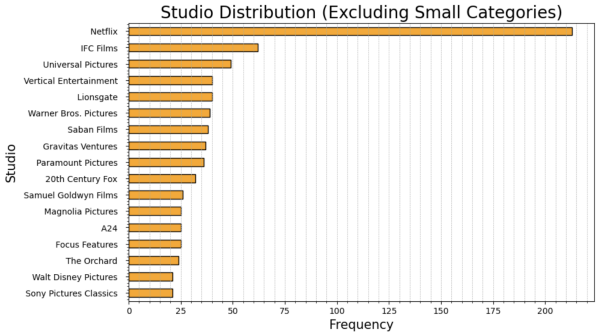
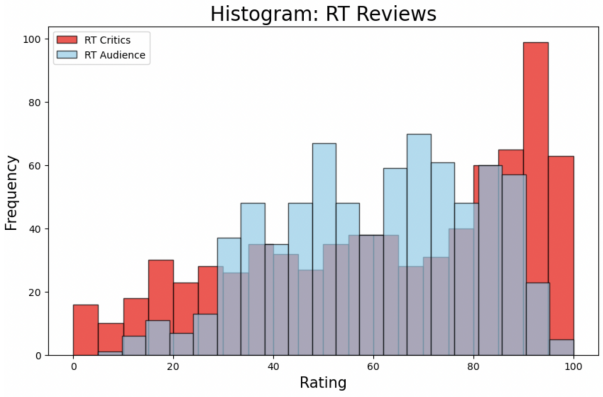


Figure 4: Histogram of iMDb critic scores and audience scores.



(a) Bar graph of studios and their frequency.



(b) Histogram of rotten tomatoes scores from critics and audience.

stated previously, this is likely due to their certified fresh system. We can also see that rotten tomato critics give more terrible reviews below 20.

6 Conclusion

Throughout this paper we have looked at different graphs that showed the correlation between the iMDB data set and the rotten tomatoes data set. We also looked at the differences between how critics and audiences rate movies. We were able to use a variety of different graphs and tables from python in order to make these comparisons.

To improve this project in the future, more data could be collected about the movies, more movies could be studied, or more comparisons could be made about the particular categories in the data. We could also utilize python to make more graphs or tables in order to better show our data. This will allow for the data to be clear and the takeaways to be more accurate.

	iMDb Critics Score	iMDb Metascore	RT Critics Score	RT Audience Score
count	738.0	683.0	742.0	742.0
mean	64.23	57.86	62.15	60.6
std	8.67	17.06	28.03	20.0
min	31.0	1.0	0.0	5.0
max	85.0	99.0	100.0	100.0

Figure 6: Two-way chart comparing data sets.

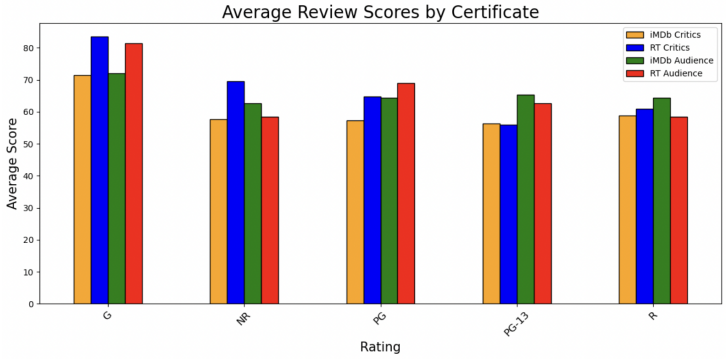


Figure 7: Bar graph of all rating categories and their average score.

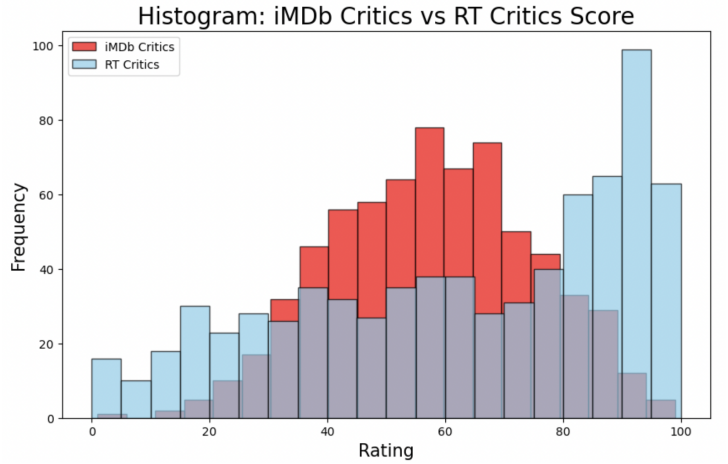


Figure 8: Histogram of iMDb critic scores plotted against rotten tomatoe's critics scores.