01.30.2024 Presentation

Ethan Weiland

2024-01-30

1. Research Question

The research question for this project is: What is the association between how critics rate a movie and how general audiences rate a movie? Also, does this relationship vary, if at all, by factors like box office success, genre, and release date?

2. The Data

A fantastic source to answer my research question is the popular website Rotten Tomatoes. Rotten Tomatoes contains information on a plethora of movies, and for each movie it reports a "Tomatometer" (average critic score) and an "Audience Score" (average audience score from verified users). Besides these two main variables of interest, Rotten Tomatoes also contains information on other movie characteristics like rating, release date, box office gross, etc.

I do not have experience with web scraping, but fortunately a Kaggle user has already scraped Rotten Tomatoes and made the data freely available at the following link: https://www.kaggle.com/datasets/andrezaza/clapper-massive-rotten-tomatoes-movies-and-reviews/. This data contains information on over 140,000 movies. However, once I remove movies with missing data on either audience score or critic score, this is reduced to **30,323 movies**. The dataset contains the following variables:

- id = unique identifier for each movie in the dataset (i.e., primary key)
- title
- audienceScore
- tomatoMeter = average critic score
- rating
- ratingContents = why a movie is rated the way that it is
- releaseDateTheaters
- releaseDateStreaming
- runtimeMinutes
- genre
- originalLanguage
- director
- writer
- boxOffice = gross U.S. box office
- distributor
- soundMix (e.g., "Surround", "Dolby")

Not every film has information on each of these variables. Table 1 displays the percentage of **non-missing** data on each variable:

Table 1: Percentage of Non-Missing Data Among Each Variable

Variable	% Non-Missing
id	100
title	100
audienceScore	100
tomatoMeter	100
director	99.68
genre	98.22
runtimeMinutes	97.72
originalLanguage	97.35
releaseDateStreaming	89.65
writer	80.32
releaseDateTheaters	62.02
distributor	54.97
boxOffice	42.97
soundMix	31.76
rating	30.04
ratingContents	30.04

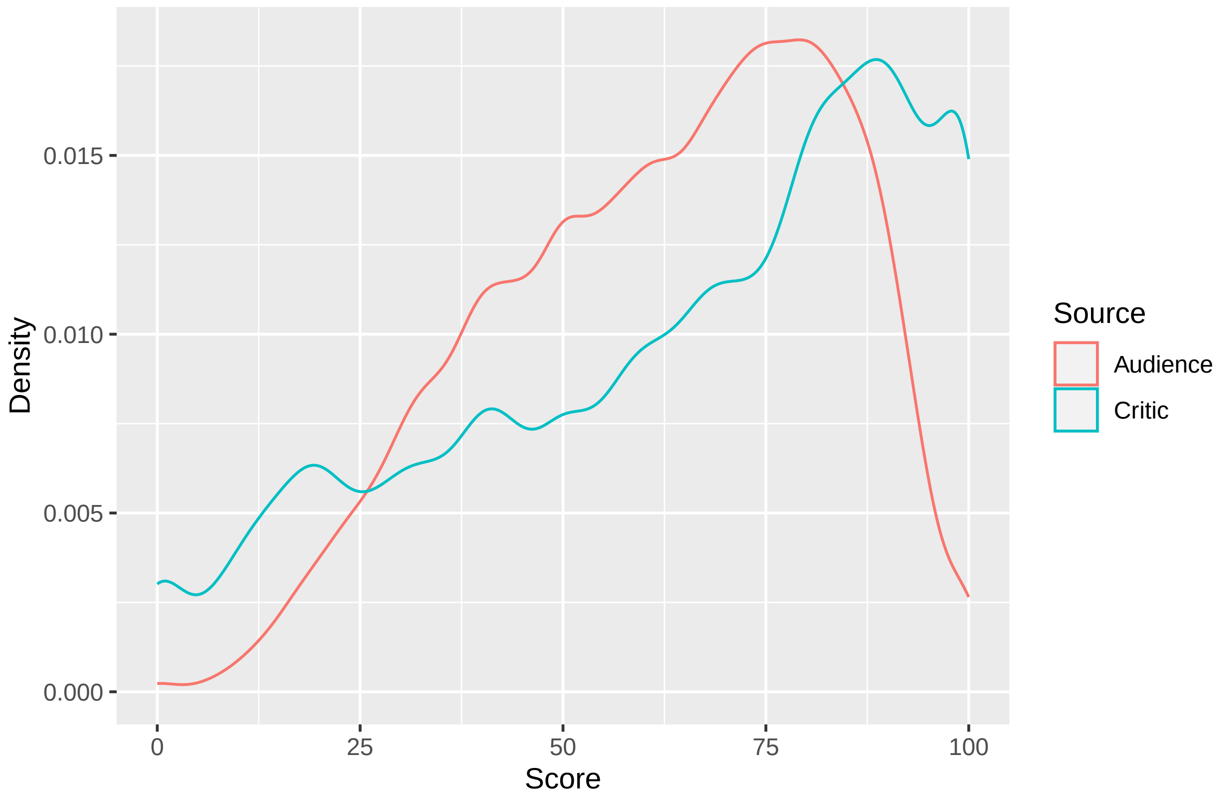
3 Bivariate Association Between Audience Score and Critic Score

The two main variables of interest are audienceScore (average audience score from verified users) and tomatoMeter (average critic score). Table 2 reports the summary statistics of each of these variables and Figure 1 plots their respective distributions.

Table 2: Descriptive Statistics of Critic Score and Audience Score

	Mean	Median	SD
Critic Score	64.53	71	28.05
Audience Score	62.32	65	20.86

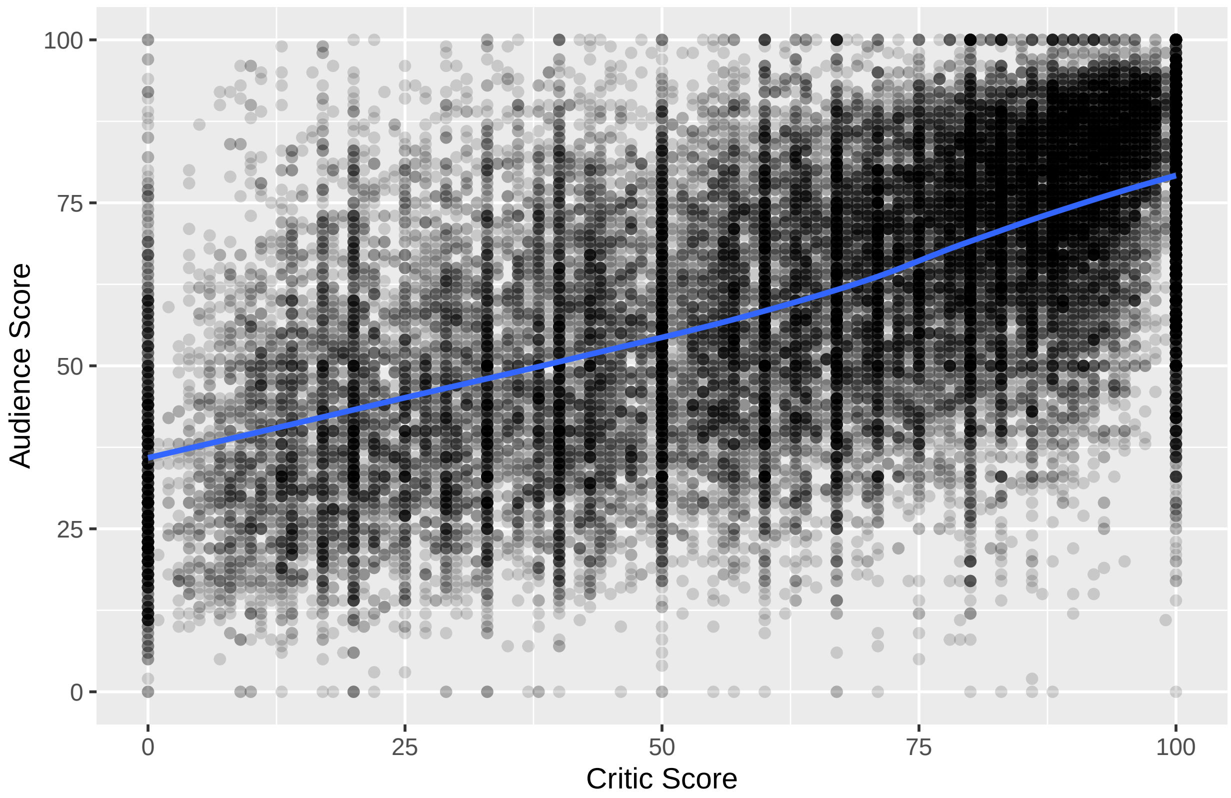
Figure 1. Distribution of Scores by Source



The respective distributions of audience score and critic score are somewhat similar. Both sources are slightly left-skewed, with a greater proportion of favorable than unfavorable ratings. However, critic scores are more likely to be very low and very high, while the mode of audience scores is at ~75%. Critic scores also have a greater spread, evidenced by the higher standard deviation value.

Now, turning to the bivariate relationship between these two variables. Figure 2 plots the association between audience score and critic score, as well as a LOESS smoother to help distinguish the overall pattern.

Figure 2. Audience Score vs. Critic Score



The sheer amount of movies makes discerning the overall relationship with the naked eye difficult. This is where the LOESS smoother comes in useful. The LOESS smoother (in blue) is remarkably linear across the range of scores. The correlation between these two variables is 0.6023454, which indicates a moderate-strong linear relationship. A linear model seems best moving forward, but it will be interesting to see how this relationship changes when controlling for other variables.

One interesting note is that there are more critic scores of 0% and 100% than audience scores of 0% and 100%, respectively.

4. Consulting Needs

I anticipate the following consulting needs:

1. Adjusting my research question. I am happy with the general research question exploring the association between audience score and critic score, but I am open to supplementing it with analyses the class thinks are worth pursuing.
2. Help with Git/Github. One of the skills I want to practice in this project is version control. I have successfully set up the repository on my GitHub (https://github.com/ethanphilipweiland/rotten-tomatoes). I will likely have questions on Git/Github as I work on my project.
3. Text analysis? The Kaggle repository also contains information on 1,400,000 reviews, including the reviews themselves as well as additional information (reviewer, publication, etc.). I have never done any text analysis, but am open to exploring this additional data.

5. Code

The code to produce this document, as well as the Rotten Tomatoes data, is available at https://github.com/ethanphilipweiland/rotten-tomatoes