Data Report

Ethan Weiland

2024-03-21

1. Introduction

The contemporary United States is experiencing low levels of societal trust. Surveys show that Americans do not trust newspapers, universities, the Supreme Court, and other key institutions. This analysis examines whether this anti-establishment sentiment extends into the cultural sphere. Specifically, I look at the opinions of the general public and elite critics on a key cultural good: movies. Using data from Rotten Tomatoes, two research questions are asked:

1. What is the association between critic opinion and public opinion for films?
2. Has the association between critic opinion and public opinion for films changed over time, especially during the 21st century?

2. Data

The data for this analysis comes from 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). Additionally, Rotten Tomatoes contains information on other movie characteristics like release date, box office gross, run time, etc. A Kaggle user 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 repository contains two tables used for this analysis: a first table of information on 140,000 movies and a second table of information on 1,400,000 individual critic reviews of these movies.

The dependent variable of interest is public opinion. Public opinion is operationalized as "Audience Score" for each movie, which is an aggregated score of individual user ratings of a movie on Rotten Tomatoes. Movies without an audience score were dropped.

The first independent variable of interest is critic opinion. Each movie contains an associated "TomatoMeter", which is an aggregate of the critic reviews. However, the definition of "critics" used in the TomatoMeter is too broad, so this analysis instead calculates a modified TomatoMeter which only considers reviews from critics classified as "Top Critics". Top Critics are critics that work for established and well-known publications (e.g., The New York Times), who best represents elite opinion. To calculate this modified TomatoMeter using only Top Critics, the table of individual reviews was used. Individual critics score films on different scales so directly comparing scores can result in measurement problems. However, for each review Rotten Tomatoes reports whether the review is "fresh" (a positive review) or "rotten" (a negative review). For each film, critic opinion is the percentage of "fresh" reviews. The overall critic opinion is "fresh" if greater than or equal to 50% of critic reviews are "fresh" and "rotten" otherwise. To prevent one critic's review from dominating, only movies with at least two top critic scores were included.

The second independent variable of interest is release date. Two potential dates are available from the Rotten Tomatoes data: the release date in theaters and the release date on streaming. The release date in theaters is chosen because films in theaters are more likely to be seen and reviewed by critics than

films released straight to streaming (especially before the COVID pandemic). To fill in missing values on release dates, the Rotten Tomatoes data was supplemented with data from The Movie DataBase (TMDB). A Kaggle user posts the complete dataset from TMDB daily, which is freely available at the following link (https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies). If a film had a valid release date from the Rotten Tomatoes data, this value was used. If a film had a missing release date from the Rotten Tomatoes data, and a non-missing release date from the TMDB data, the latter value was used. The films that still had a missing value on release date were dropped.

Various control variables are included. The box office variable refers to the monetary success of the movie in millions of dollars. Due to missing data, this variable was also supplemented with data from TMDB in the same manner as the release date variable. Language refers to the original language of the film. Language is a factor variable with two levels: "English" and "Non-English". Run time refers to the length of the film (in minutes). Review count refers to the number of Top Critic reviews for each film. As described above, the minimum review count value is 2. After cleaning the data and removing films without a valid critic score, audience score, or release date, the final sample size is 20196.

3. Descriptive Statistics

Table 1: Descriptive Statistics (n = 20196)

	Mean	Median	SD	Percent_Missing
Audience Score	63.03	66	20.62	0
Critic Score (% of 'Fresh' Reviews)	60.12	66.67	32.00	0
Release Date	12,389.40	14,608.50		0
Box Office (Millions $)	16.44	0.16	48.63	16.02
Language (1=English)	0.79			0.50
Run Time (Minutes)	102.95	99	20.54	0.10
Review Count	17.95	10	18.30	0

Figure 1. Distribution of Audience Scores

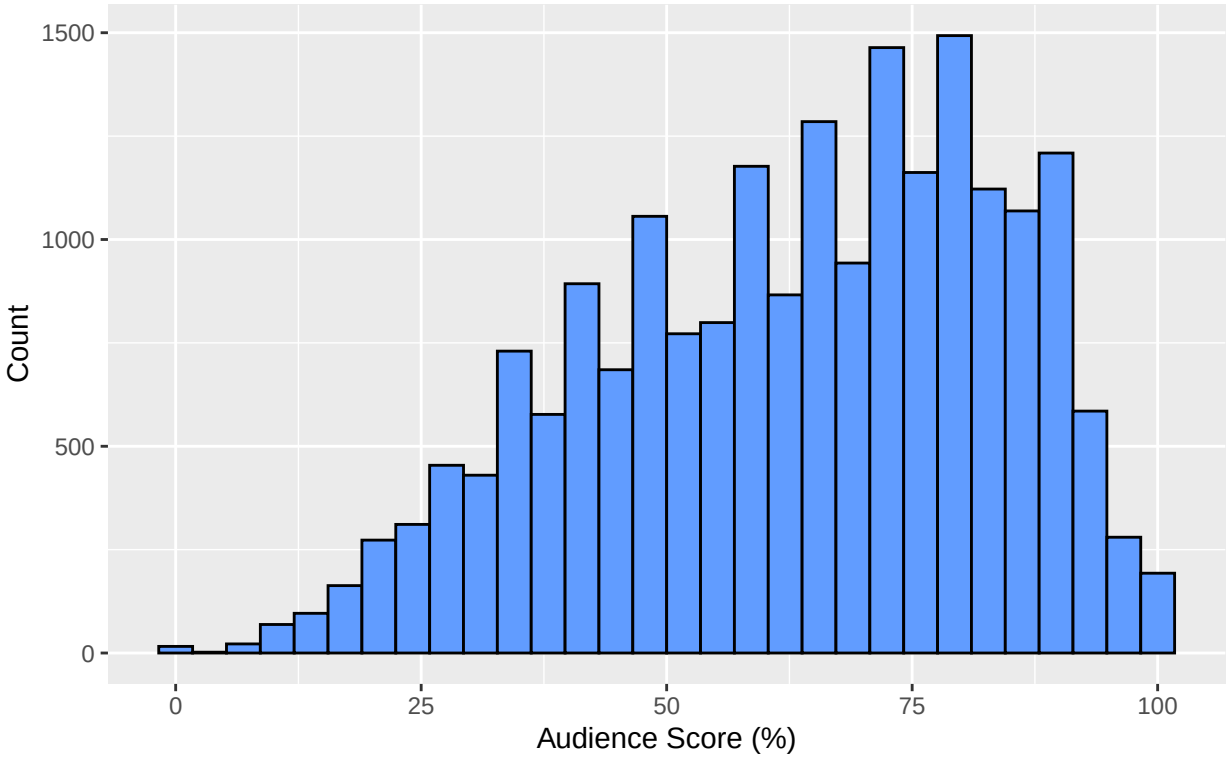


Figure 2. Distribution of Critic Scores (Continuous)

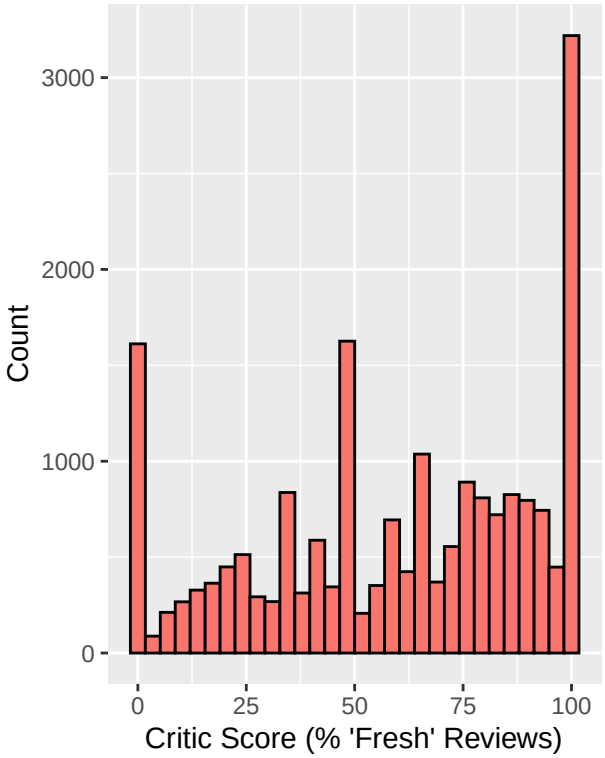
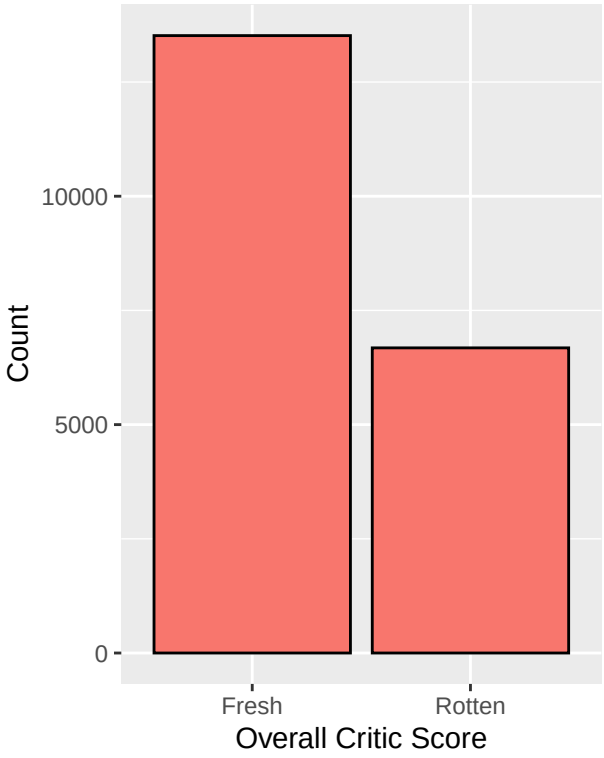


Figure 3. Distribution of Critic Scores (Discrete)



4. Association Between Audience Score and Critic Score

Figure 4. Audience Score vs. Critic Score (Continuous)

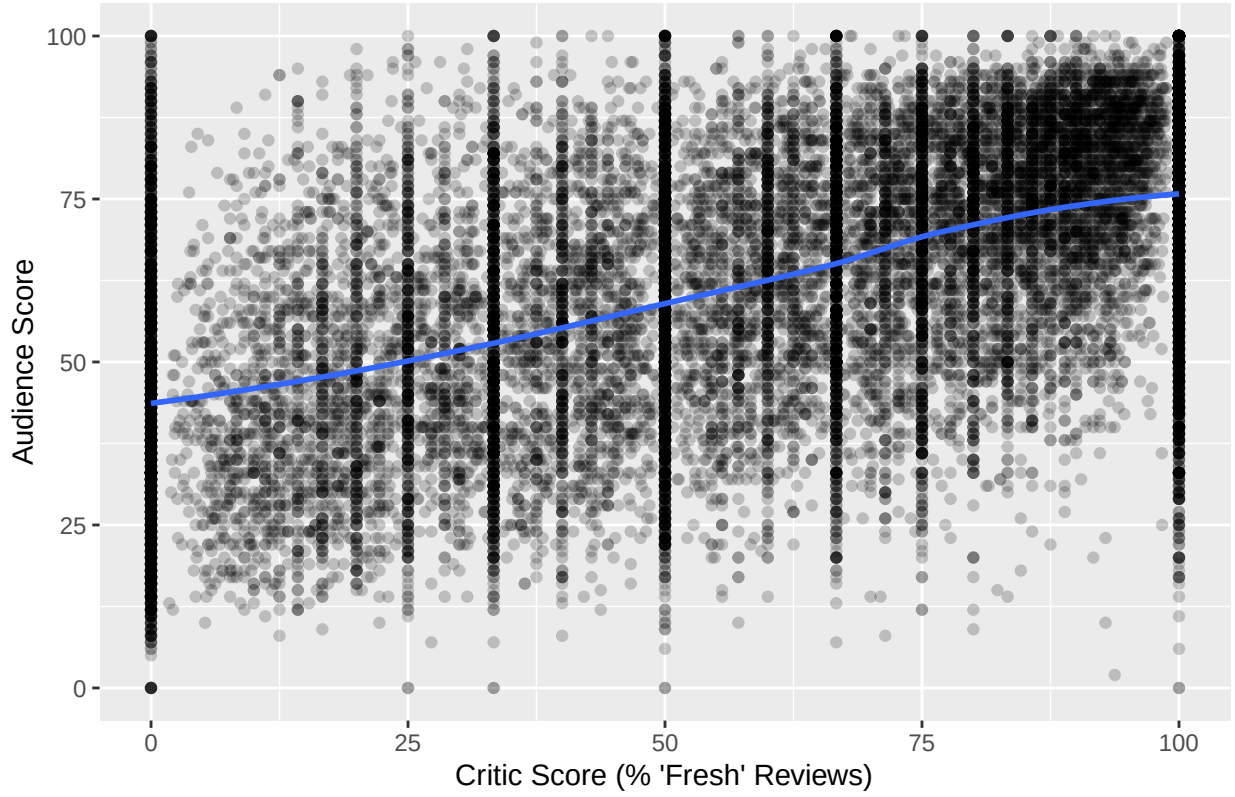


Figure 5. Audience Score vs. Critic Score (Discrete)

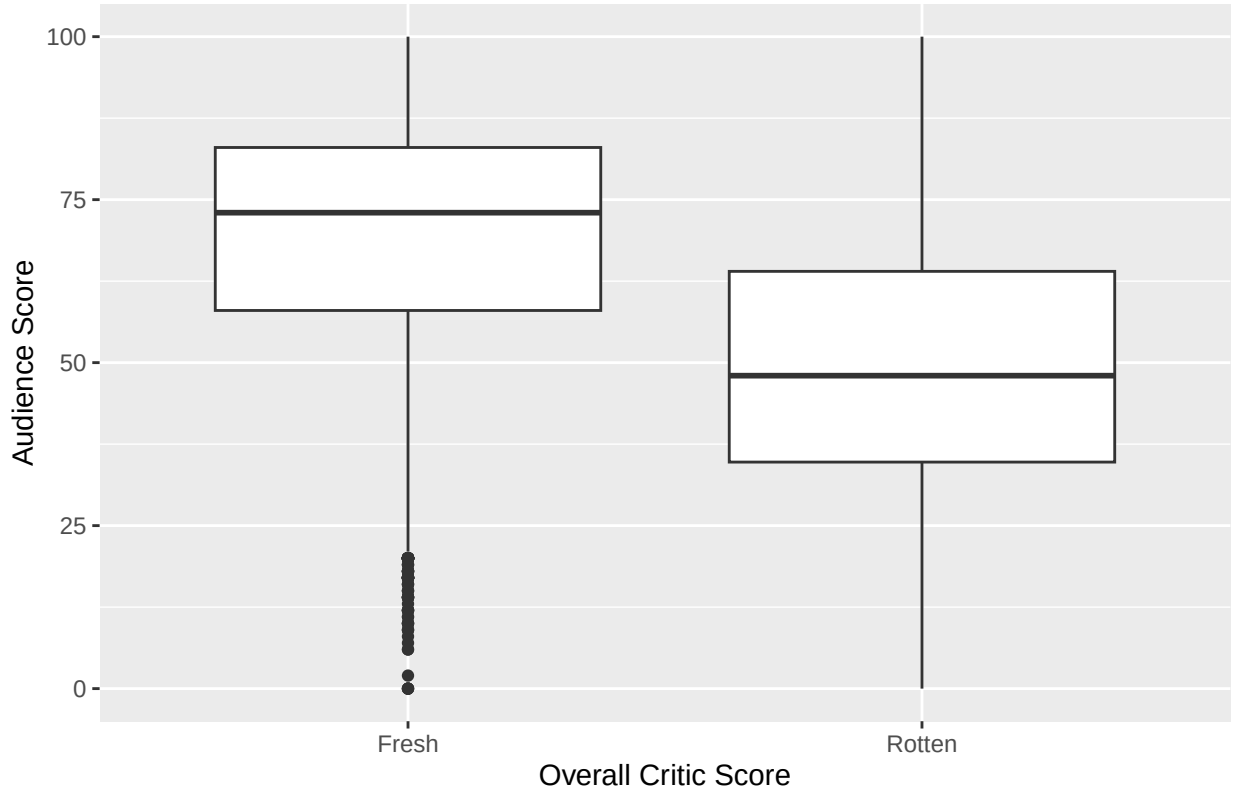


Table 2: Audience Score by Overall Critic Score

Overall Critic Score	Mean	Median	SD
Fresh	69.74	73	17.54
Rotten	49.46	48	19.68

5. Association Between Audience Score and Critic Score Over Time

Figure 6. Audience Score vs. Critic Score Over Time (Continuous)

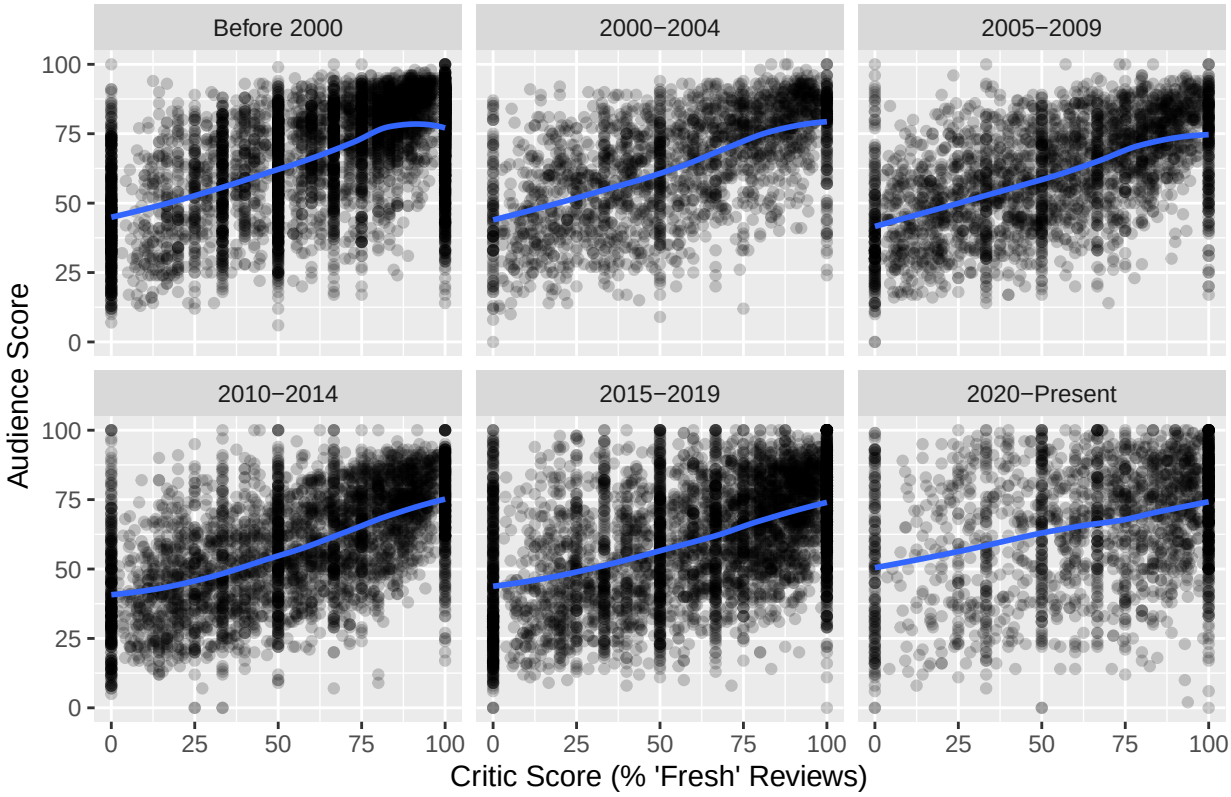


Figure 7. Audience Score vs. Critic Score Over Time (Discrete)

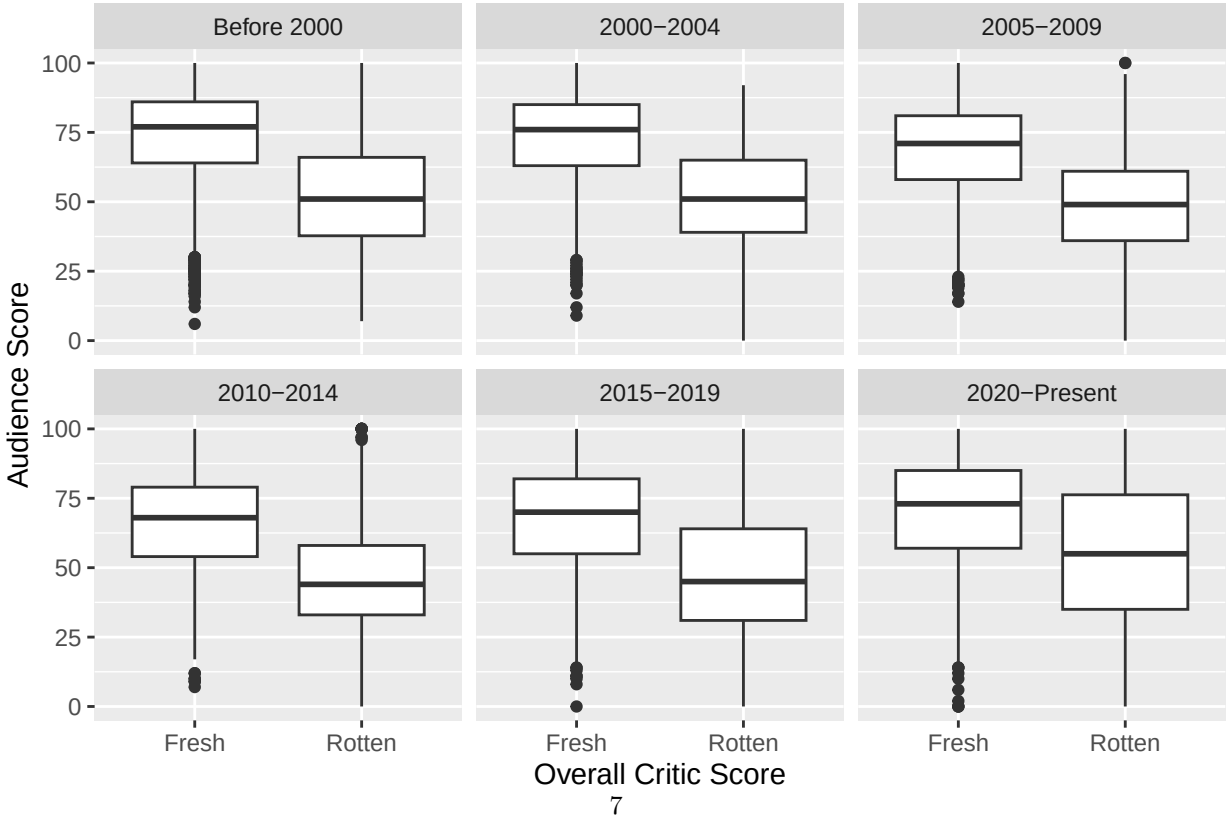


Table 3: Audience Score by Overall Critic Score Over Time

Period	Overall Critic Score	Mean	Median	SD
Before 2000	Fresh	73.3	77	16.44
Before 2000	Rotten	51.57	51	18.63
2000-2004	Fresh	72.08	76	16.19
2000-2004	Rotten	51.96	51	17.82
2005-2009	Fresh	68.39	71	16.26
2005-2009	Rotten	49.38	49	17.64
2010-2014	Fresh	66.01	68	17.01
2010-2014	Rotten	45.38	44	18.21
2015-2019	Fresh	67.43	70	18.31
2015-2019	Rotten	47.95	45	21.52
2020-Present	Fresh	69.97	73	19.76
2020-Present	Rotten	55.34	55	25.06

6. GitHub Link

https://github.com/ethanphilipweiland/rotten-tomatoes