Data Report

Ethan Weiland

2024-03-07

Introduction

The contemporary United States is experiencing historically low levels of trust. Surveys show that Americans do not trust newspapers, universities, the Supreme Court, and other key institutions in society. This analysis studies this anti-establishment zeitgeist in 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 posited:

1. What is the association between critic score and public opinion for films in the 21st century?
2. Has the association between critic score and public opinion for films changed over the course of the 21st century?

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 score. Each movie contains an associated "TomatoMeter", which is an aggregate of the critic scores. 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. Each review was converted into a percentage. Numeric scores (e.g., "3/4") were simply calculated arithmetically (e.g., "3/4" became 75%). Letter-grade scores were converted to percentages using a standard grading scale (e.g., "B+" became 87.5%). The scraping algorithm resulted in some uninterpretable scores (e.g., "2.1/2"), which were coded as missing. To avoid one critic's review dominating a modified TomatoMeter score, only movies with at least two top critic scores were included. Then, for each film fitting this criterion, the critic scores were simply averaged.

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. Finally, due to the scope of the research question, any film released before 2000 was 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 12779.

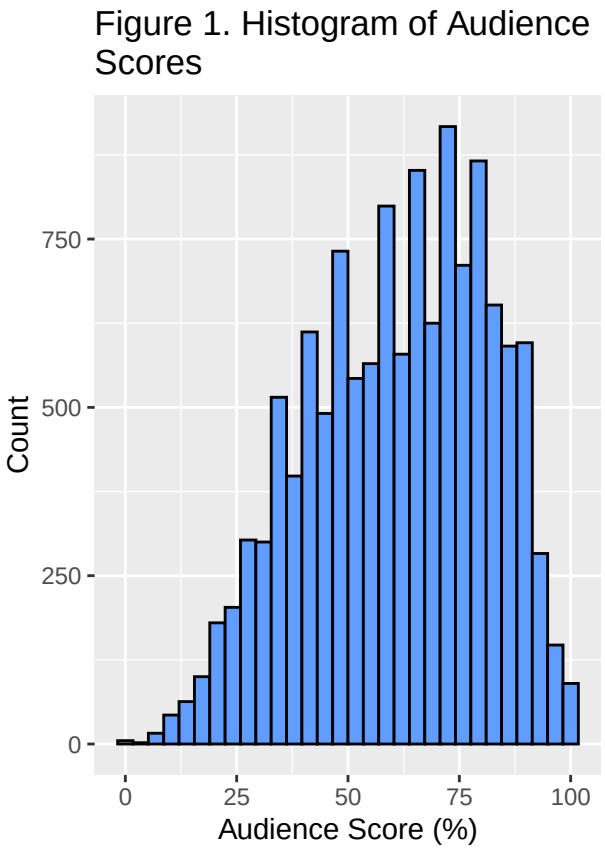
Descriptive Statistics

Table 1 reports the descriptive statistics (mean, median, standard deviation, and percent missing) of audience score, critic score, release date, box office, language, run time, and review count.

Table 1: Descriptive Statistics (n = 12779)

	Mean	Median	SD	Percent_Missing
Audience Score	61.28	63	20.20	0
Critic Score	62.08	63.53	14.32	0
Release Date	15,518.36	15,688		0
Box Office (Millions $)	17.90	0.17	53.56	11.20
Language (1=English)	0.78			0.11
Run Time (Minutes)	102.29	99	18.98	0.08
Review Count	16.42	11	14.07	0

Figure 1 and Figure 2 below plot the univariate distributions of audience score and critic score, respectively. Both variables have slight non-Normal distributions with a left-skew. However, the skew is greater for audience score, evidence by its greater standard deviation (20.20 percentage points) compared to critic score (14.32 percentage points). Both variables are centered around 60%. The mean and median of critic score (62.08 percentage points, 63.53 percentage points) are both slightly higher than the mean and median of audience score (61.28 percentage points, 63.00 percentage points).



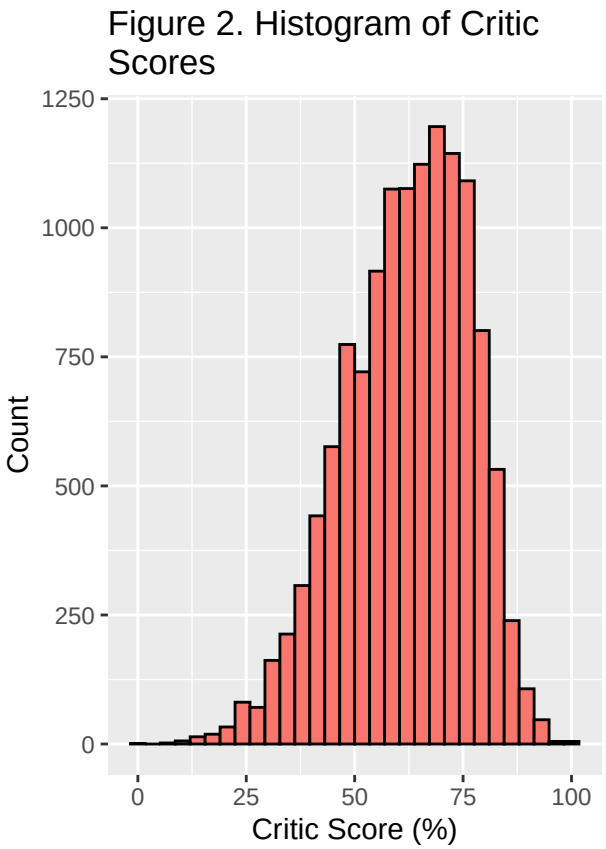
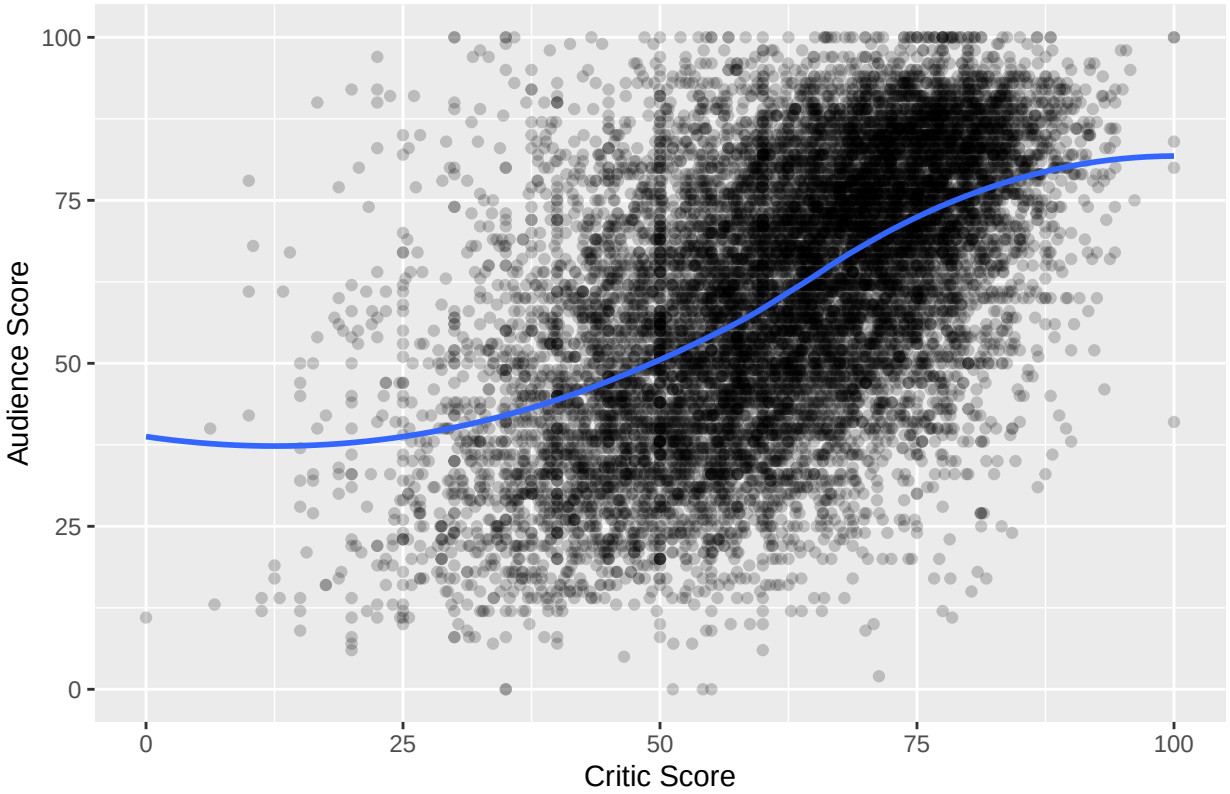


Figure 3 displays the bivariate association between audience score and critic score across all release dates. In doing so, Figure 3 provides a preliminary answer to the first research question. The blue line is a LOESS smoother (a local regression), which is a tool to help assess patterns in scatter plots with many points. Overall, there is a moderate linear association between audience score and critic score. This is evidenced by the smoother tracing a linear pattern through the main cluster of points. The smoother is flat at low levels of critic score and high levels of critic score, but this is due to a sparsity of data points rather than the relationship between the two variables. It is possible that the relationship is cubic, rather than strictly linear, which will be tested in future analyses. The correlation between the two variables is 0.54.

Figure 3. Audience Score vs. Critic Score



To begin answering the second research question, Figure 4 below plots the association between critic score and audience score in the 21st century, this time faceted out by five year time intervals. Again, LOESS smoothers are plotted (the blue lines) to assist in discerning patterns. The patterns from 2000 - 2019 are constant: a moderately linear relationship between critic score and audience score. The patterns in the lowest and highest audience scores differ, but again this is due to sparse data. However, the relationship for films from 2020 to present is different. There is still a linear relationship, but it is weaker. Correlations calculated for each time period confirm this (Table 2). From 2000-2019, the correlation between critic score and audience score is about 0.6. From 2020 to present, the correlation drops to 0.5.

Figure 4. Audience Score vs. Critic Score Over Time

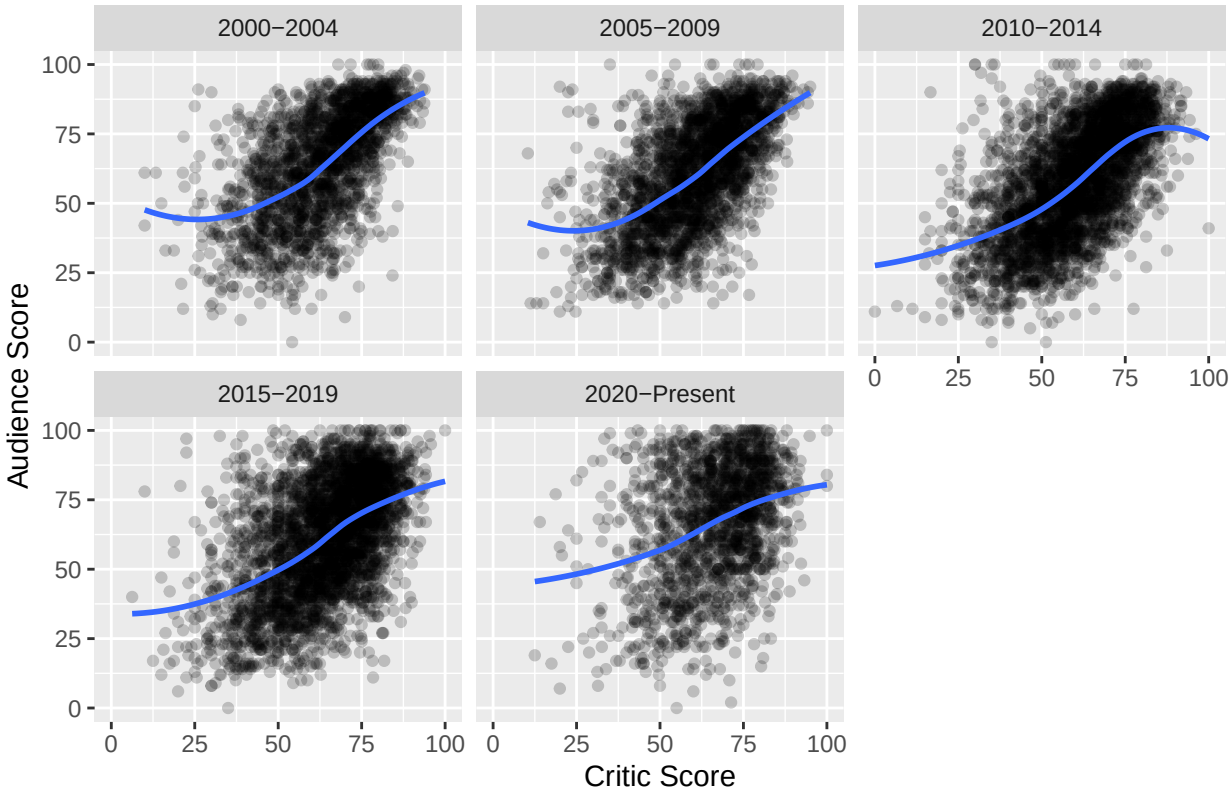


Table 2: Correlations Between Audience Score and Critic Score in Different Time Periods

	Period	Correlation
1	2000-2004	0.59
2	2005-2009	0.59
3	2010-2014	0.59
4	2015-2019	0.59
5	2020-Present	0.50

GitHub Link

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