THE UNIVERSITY OF TEXAS AT DALLAS

Is there a correlation between the drop rate of google trend searches between Day 1 and Day 21 of the movie being released and the movies rating on IMDb and Rotten Tomatoes?

School of Economic, Political and Policy Sciences

EPPS 6302: Methods of Data Collection and Production

Group leader: Oliver Myers

Group members: Calvin Hanebeck, Allen Hernandez & Lynn Kuhlwein

Instructor: Dr. Karl Ho

Dec 3, 2024

Table of Contents

1. Introduction

Google Trends is a handy and freely accessible tool to retrieve information about daily global search interests and trends. Ever since its establishment in 2006, the tool's predictive power was recognized by various research communities to enhance forecasting models regarding consumer behavior, financial markets' volatility or disease spread. But only very little research focused on examining the relationship between Google searches and IMDb movie ratings.

Building up on Demir et al.'s intuition “that for popular movies one should see the higher search volume of queries associated with the movie” (Demir et al., 2012, p. 1), we examine in what relation the most rated movies on the website IMDb stand to their search interest on Google. We gather data about movies with the most ratings for the year 2022 from the movie rating platform IMDb and combine it with data provided by Google Trends regarding the movies' search interest in that year. Hence, we formulate our research question as follows: “Is there a correlation between the drop rate of google trend searches between Day 1 and Day 21 of the movie being released the IMDB Rating?”. We calculate a drop rate in search interest for the 65 most rated movies on IMDb (more than 10,000 ratings) of the year 2022 by comparing the search interest' value on the day of release with the search interest' value 21 days after the movie's release. We hypothesize that a movie with a lower drop rate in search interest after 21 days is associated with having a higher number of movie ratings on the website of IMDb, one of the most renowned digital platforms for movie and tv reviews.

The importance of receiving a high amount of (good) ratings for a product cannot be understated in an era of digitized economy with a high reliance on platforms for product or information exchange. The global internet connectivity allows for network effects, i.e. the more users rely on a platform, the more valuable that platform or network becomes, which gives way to a feedback loop that can lead to market concentration. Following this logic, having established a good rating for a movie on platforms like IMDb or Rotten Tomatoes can be crucial to the movie's economic success but also to its long-term impact of shaping cultural perception. Shukla et al. (2022) point out that "Electronic Word-of-Mouth (e-WOM) has played a significant role in influencing the sale of products" (p.1) since e-WOM practices like consumer reviews or ratings are a form of free advertising and have a greater influence because of its credibility among other consumers as a reliable source and its relative independence. Analyzing the emergence of consumer ratings like the user movie ratings on IMDb is therefore of crucial importance.

The paper's content will be the following: the second section will provide the methodology section including a literature review on scholarly research using Google Trends data, our method of webscraping and an explanation of our regression analysis. The third section will provide a detailed description of our data production process using R Studio. The fourth section summarizes our main results from our regression analysis with descriptive statistics and regression tables using Stata testing and contextualizing our hypothesis with our collected data. The fifth section presents a discussion about the results obtained in the prior section and its limitations as well as a description of our difficulties and learnings during our data gathering process. Finally, the last section summarizes the project’s main points and the analysis’ main findings, thereby presenting options for future research.

2. Methodology

2.1 Webscraping with Rvest

To construct our dataset, we employed the R package **rvest** *Wickham H (2024)* to scrape movie information from relevant websites. This package allows users to extract specific HTML elements from a webpage, enabling the construction of structured data directly from online resources.

Initially, we attempted to scrape data from IMDb's advanced search page, filtering for movies released in 2022 with more than 50,000 IMDb votes and in English. However, this method encountered technical challenges (detailed in Chapter 3), prompting a pivot to Box Office Mojo, an IMDb subsidiary. Box Office Mojo provides detailed box office data, making it an ideal alternative.

Using rvest, we scraped key elements from Box Office Mojo’s domestic box office page for 2022 releases. Elements extracted included:

- Movie Title: The name of the movie.
- Release ID: A unique identifier assigned to each movie for URL referencing.
- Maximum Theaters: The maximum number of theaters in which the movie was shown.

From this dataset, we filtered out movies shown in fewer than 600 theaters, ensuring that our focus remained on widely distributed releases. The resulting list of movies served as the foundation for further data enrichment and analysis.

We applied the rvest package again to construct URLs dynamically using each movie's Release ID. These URLs allowed us to scrape daily box office earnings for each movie, including:

- Date: The calendar date of the earnings report.
- Days Since Release: The number of days since the movie's release.
- Daily Earnings: The revenue generated by the movie on a given day.

Each movie’s earnings data was saved as a separate CSV file for subsequent integration with Google Trends data. This systematic approach ensured comprehensive temporal coverage of box office performance.

2.2 IMDb Rating, omdb

The OMDb (Open Movie Database) API was utilized to collect detailed metadata for the movies identified in our initial dataset. This step enriched our data with additional variables, including:

- Genres: Categorization of the movie (e.g., drama, action).
- Runtime: The duration of the movie in minutes.
- IMDb Ratings: User-generated scores on IMDb.
- User Votes: The number of ratings a movie received.
- Awards: Accolades won by the movie.

- Rotten Tomatoes Ratings: Aggregate critical reception.

Data Retrieval and Cleaning

Using the httr package in R, we queried the OMDb API for each movie title. Each query returned data in JSON format, which was parsed into structured data frames using the jsonlite package.

However, challenges arose during this process:

Null Values: Some movies lacked complete metadata (e.g., missing ratings or runtime).

Duplicate Entries: Occasionally, multiple entries for the same movie were returned.

Formatting Issues: Certain fields required cleaning, such as converting IMDb vote counts from text to numeric values.

To address these issues, we implemented error-handling mechanisms in our R scripts, ensuring that incomplete or duplicate rows were flagged and removed. The cleaned dataset was further refined by applying filters to:

- Exclude movies with fewer than 10,000 IMDb votes.
- Retain only movies released in the United States and in English.

Preparing for Google Trends Data

To align with the requirements of Google Trends queries, the movie release dates were reformatted into a consistent structure. Additionally, a new column was created, adding 21 days to the release date, to capture the drop in search interest during the critical post-release period.

This finalized dataset was stored as a CSV file, ready for integration with Google Trends and box office data in the subsequent stages.

2.3 Google Trends

The scholarly research using Google Trends emerged in the years after the tool's establishment in May 2006 (Jun et al. 2018). It is primarily used to examine its ability to forecast events or predict behavior. Fritzsch et al. (2020) come to the conclusion that Google Trends data can enhance the prediction performance of conventional models (1409) supporting the notion of Google Trends' predictive power. Carneiro and Mylonakis (2009) use Google Flu Trends to track infectious diseases. The researchers found that Trends "can detect regional outbreaks of influenza 7–10 days before conventional Centers for Disease Control and Prevention surveillance systems" (1557), thereby showcasing great potential as a timely, robust, and sensitive surveillance system. Despite their findings, Carneiro and Mylonakis hint at the fact that prediction could only work in developed countries with high internet penetration because the forecasting requires large populations of web search users to be most effective (1563).

In another study, Woo and Owen (2018) analyzed the potential of Google Trends as a consumption indicator by demonstrating that applying Google Trends-augmented models improve forecasts of private consumption growth over forecasts that do not utilize Google Trends data (81). They hint at the potential that Google Trends data "will allow policy makers to respond to economic events in a timelier and more appropriate manner" (90) and at the tool's general benefits: easily accessible, contains a large sample size, no usage cost and daily data updates. Additionally, it is a valuable channel for obtaining data on consumers' pre-purchase research activities.

Regarding financial research, Petropoulos et al. (2022) used Google Trends indices to measure people's interest in financial news, combining it with a deep learning tool.

They argue that Google Trends “can provide useful input in the creation of crisis Early Warning Systems”, stressing the importance of the information conveyed by social data since it is “more responsive compared to official financial indicators” (353). Google searches convey information able to predict future market turbulence in a short time period of one month. Similar to the above mentioned phenomena of forecasting the emergence of financial crises, epidemics through infectious diseases or consumption behavior, Borup and Schütte (2022) that forecasting models equipped with Google trends data can be a “valuable tool for obtaining accurate real-time information on future employment growth and labor market conditions” (32) at horizons up to one year ahead. It is important to mention that the researchers did not use simple search queries but a combination of many Google Trends series, preferably of a non-linear manner, to reach a stronger forecasting power.

Regarding our IMDb Rating Google Trends nexus, we found three studies using Google Trends to predict either IMDb movie ratings (Demir et al. 2012), UK cinema admissions (Judge and Hand 2010) or movie ticket sales (Shukla et al. 2022). The latter study analyzed the impact of e-WOM (electronic word-of-mouth) on movie ticket sales. The researchers collected ticket sales from *Box Office Mojo*, an online box office reporting and analysis service by IMDbPro, and four e-WOM factors from *Rotten Tomatoes*, one of the biggest online sources of movie and TV reviews. They retrieved four explanatory variables, namely Tomatometer ratings, total count of ratings, audience score and user ratings (Schukla et al. 2022, 28), to do a regression analysis examining the correlation on gross sales of tickets of a given “Hollywood movies” (29). The researchers found that all four independent variables were positively associated with movie ticket sales. The researchers conclude that “consumers trust the opinion of peers as more valuable compared to expert opinion” (29) since the variable user ratings showcases the strongest correlation with movie ticket sales.

Similar to Shukla et al., Judge and Hand (2010), examined cinema admissions for movies in the UK using forecast models enhanced with Google Trends data on searches relevant to cinema visits. They found that the enhanced models have the potential to increase the accuracy of forecasting regarding cinema admissions (12). The researchers suggest further research to make use of Google Trends data regarding ”searches on individual movie titles, especially those that have attracted media attention because of Oscar nominations” (13).

Our data production project commences at this point with the aim to fill this research gap of using Google Trends data to examine its correlation with IMDb ratings. To the best of our knowledge, there is only one study from Demir et al. (2012) which uses Google Trends data to predict IMDb movie ratings. For their study the researchers collected two datasets for two experiments. For the first analysis, they collected search interest data on 120 movies in the time interval starting one month before the release date and ending four months after the release date (2). Data on four different features of every movie - title of the movie, directory of the movie, actor #1, actor #2 - was obtained through search queries for both experiments (4). The researchers actually find that “short term post-release search activities (4 weeks of search activity after the release) have less predictive power than those of pre-release and longer term post-release” (5). Since the difference in forecasting ability was small and their findings have not further been validated, our approach aims to build up on that finding.

In order to obtain data for our independent variable, the development, i.e. the drop rate, of search interest on the platform Google for a given movie, we used Google Trends. It is a free tool provided by Google itself that allows users to access information about the development of search interest for “a sample of search requests, by topic, over time, and down to city-level geography” (Google Trends 2024). This can be utilized to get a quick overview of which phenomena or news are currently trending. Further, the tool allows for five terms to be searched and compared simultaneously, also across different areas of the Google universe which are completely separated from each other. It allows search requests for *Web Search*, *Image Search*, *Google Shopping* and *Youtube Search*. We only took the *Web Search* for our data retrieval and analysis into account. Further, we did not make use of the option to compare search since our pool of movies was too big to allow any yielding comparison among them.

The most important thing regarding the data’s interpretation is that the output of a search on Google Trends is normalized and indexed. To determine when or whether a topic or a specific term is popular, one is looking for spikes in the output graph, i.e. “a sudden acceleration of search interest in a topic, compared to usual search volume.” Because of the indexing and normalization it is advised to “describe spikes as an increase in ‘search interest’, rather than ‘searches’” (Google Trends 2024). A topic’s search interest is normalized means that the output (ranging from 0 to 100) does not show the total number of searches at a certain point of time, but rather a “percentage of searches for that topic, as a proportion of all searches at that time and location” (Google Trends, 2024). Regarding the indexing, the data is taken from a random, unbiased sample of Google searches which does not reproduce exact number of searches but rather shows when a term reached its “maximum search interest for the time and location selected”.

If a spike in search interest for a single topic is observed, it is recommended to “sense-check” the output by adding a second, consistently highly-searched topic (e.g. the news, weather, leading politicians or recipes) to put the output of the first topic into context. Given the comparison, Google Trends will produce data (primarily shown in a graph) to show how popular the own search term was in comparison to the highly-searched term. The bigger the proportion of the own, often lesser searched term to the highly-searched term, the more certain one can be that the own search term actually has a general high search interest at the given time and location. Since we computed a drop rate for movies’ search interest around their publication date, the “sense-check”-process was not part of our project. Despite the fact that Google Trends only uses relative search interest as a measure, another shortcoming is the sampling method used (Fritsch et al., 2023, 812). Reports on search queries are drawn from a sample of searches which could lead to sampling error if the data pool is small. But the tool does offer a variety of benefits: it is a massive dataset (data is retrievable back to 2004) which is easily accessible at no usage cost providing daily data updates. Additionally, it provides information about what humans are really interested in since search queries are made in private and reflect what humans really think. Lastly, the data itself is “anonymized, categorized and aggregated, thereby overcoming privacy-related concerns” (Silva and Madsen, 2022, 446).

To retrieve the data from Google Trends, we used the gtrendsR package for running Google Trends queries in R. It is possible to retrieve and display data from Google

Trends similar to what is presented via a web browser. The package includes a complete list of categories that can be used to narrow requests. Search queries can be specified to a time, different geographic regions, using different Google products (e.g. web search or images) or multiple keywords for comparison.

2.4 Regression analysis

To be able to analyze our data and answer our research question - whether one can predict the rating of a movie by looking at the Google Trends drop rate - we will use a multiple linear regression. A general multiple linear regression model looks like the following:

$$y = \beta_0 + \beta_1 * x + \mu$$

It consists of an explanatory and independent variable x , the explained and dependent variable y, the constant β_0, the coefficient β_1 for the explanatory variable x and the error term μ. β_1 measures the effect a change in variable x has on variable y holding all other factors constant. For an OLS model there are six general assumptions that must hold:

- The model is linear – in coefficients and error terms.
- The sample is random.
- There is no perfect collinearity.
- The mean of the error terms is zero.
- The error terms are normally distributed.
- There is no serial correlation.
- There is homoscedasticity.

For further explanation of these assumptions see Wooldridge 2013.

In our case, the independent variable is the drop_rate that reflects the change in interest in the movie after 21 days, proxied by the difference in amount of Google searches between the release date and 21 days later, as measured by Google Trends. We calculate the drop rate with the following equation:

$$drop\_rate = \frac{hits\_day1 - hits\_day21}{hits\_day1} \times 100.$$

This drop rate reflects the rate of decrease in percentage. The explained variable is the IMDb rating of the movie or the Rotten Tomatoes rating. We will also use two control variables that will be added to both models: the runtime of the movie in minutes and the box office revenue. In model 1 the number of votes received on the IMDb website will also be included as a control variable. They are included because they might influence the rating. Regarding their influence one could expect the length of a movie to influence its rating as longer movies might be perceived as higher-quality productions compared to shorter movies. A higher box office revenue could lead to a higher ranking as it attracted more visitors. As the IMDb ratings are averages based on user votes, the number of votes also reflect the popularity and visibility of a movie. Movies with a greater number of votes may have more polarized or representative ratings compared to lesser-known films. By including these controls, we account for their potential influence on the ratings and/or the drop_rate. This also accounts for potential correlations with both the independent variable and the dependent variable. If we would not control for these variables, this could result in biased estimates. Also, it possibly increases the explanatory power of our model, which improves the model fit.

We are defining our two models as follows:

Model 1:

$$imdbrating_i = \beta_0 + \beta_1 * drop\_rate_i + \beta_2 * runtime_i + \beta_3 * boxoffice_i + \beta_4 * imdbvotes_i + \mu_i$$

Model 2:

$$rottentomatoes\_rating_i = \beta_0 + \beta_1 * drop\_rate_i + \beta_2 * runtime_i + \beta_3 * boxoffice_i + \mu_i$$

The subscript $_i$ stands for every individual movie. μ is the error term, which captures unobserved influences on the movie's rating. We use robust standard errors to adjust for potential heteroscedasticity in the model.

3. Data production

The data collection process for this project involved multiple stages, combining web scraping, API integration, and data cleaning methods in R-Studio to create a comprehensive dataset of movies released in 2022. This dataset incorporated metadata, daily earnings, and search interest trends, all of which were essential for our analysis.

To begin, we used the rvest package in R to scrape movie information from Box Office Mojo, a subsidiary of IMDb. The target URL was specifically designed to retrieve data for domestic movies released in the United States in 2022: https://www.boxofficemojo.com/year/2022/?grossesOption=totalGrosses&releaseScale=wide. From this webpage, we identified and extracted critical data elements, including movie titles, unique releaseID values (used for URL referencing in subsequent steps), and the maximum number of theaters in which each movie was shown. These elements were organized into a CSV file to form the initial list of movies. To refine the dataset, we filtered out movies that were shown in fewer than 600 theaters, focusing on widely released films.

Following the initial web scraping, we enriched the dataset by integrating metadata from the Open Movie Database (OMDb) API. Access to the API was obtained through an application process, and we used the httr package in R to send requests. Each query included the movie title and the release year, enabling us to collect additional metadata such as genres, runtime, awards, IMDb ratings, IMDb vote counts, and Rotten Tomatoes ratings. The responses from the API were returned in JSON format, which we parsed into structured data frames for further use.

Several challenges arose during this stage, including missing values in key columns and duplicate entries. To address these issues, we developed cleaning functions that handled null values and consolidated duplicate rows. Additionally, we split grouped ratings, such as those from Rotten Tomatoes, into individual columns to facilitate analysis. The release dates were reformatted into a consistent numeric format, ensuring compatibility with Google Trends' required input structure. An additional column was created to extend the release date by 21 days, establishing a defined timeframe for subsequent Google Trends analysis. To finalize the metadata dataset, we removed movies with fewer than 10,000 IMDb votes, non-English movies, and those not released in the United States. Rows with missing values in critical columns, such as ratings or votes, were also excluded.

To capture daily earnings data for each movie, we returned to the rvest package. Using the releaseID values from the refined metadata CSV, we dynamically constructed URLs to scrape additional data from Box Office Mojo. For each movie, we extracted the daily earnings, the number of days since release, and the corresponding dates. These data were saved as individual CSV files, with each file named according to the respective movie's releaseID. This approach ensured that the earnings data could be easily linked back to the main dataset.

Google Trends data was then collected using the gtrendsR package in R. Each movie title served as the search keyword, with the start date set to the release date and the end date set to 21 days later. The data captured included the search term (movie title), normalized search interest values (hits), and the dates within the specified timeframe.

Collecting Google Trends data presented several challenges due to API rate limitations and occasional errors in retrieving data. To address these, we implemented error-handling mechanisms and included a 20-second delay between queries to minimize disruptions. Each query generated a CSV file containing Google Trends data for both the United States (US) and Canada (CA). After completing the collection, we filtered out all Canadian data, retaining only trends specific to the United States. Each Google Trends file was updated to include releaseID and imdbID columns for consistency with the earnings and metadata datasets.

Once all data were collected, the integration process began. We merged the daily earnings and Google Trends datasets by aligning their date columns. Rows with missing data or daily earnings extending beyond the 21-day Google Trends scope were removed. The combined data for each movie was stored in individual CSV files, which were later appended into a single consolidated dataset.

The final step involved merging this consolidated dataset with the original metadata file, creating a unified dataset that included all relevant information. This dataset incorporated Google Trends search interest data, daily earnings for each movie (limited to 21 days post-release), and comprehensive metadata, including IMDb ratings, Rotten Tomatoes ratings, genres, runtime, and total box office earnings. This final dataset served as the foundation for our analysis, enabling a robust examination of the relationships between search trends, earnings, and movie metadata.

4. Stata Analysis

After the data scraping process, we proceeded to the data analysis in Stata. In the following the data preparation process will be explained as well as short descriptive statistics and a regression presented.

4.1 Data preparation

To be able to analyze the data further with a statistical regression, it was necessary to create one data file containing all the necessary information. For this purpose, we wanted to merge the ‘movies_filtered.csv’ file with the 79 Google Trends CSV files. Our first step was turning the ‘movies_filtered.csv’ file into a .dta file. The next steps turned out to be more complicated than expected. We started to name all the 79 Google Trends CSV files ‘trends_i.csv’, where i stands for the numbers from 1 to 79. Creating a loop would be the easiest way to convert and merge the files. We did not manage to create the loop without producing error messages, therefore we decided to pursue an alternative path. To turn these files into .dta files, we used the import and save option from Stata. Chat GPT was used here to create the commands for the files 2 to 79 in reference to the code we created with file number 1. This saved a lot of time.

```
import delimited "C:\Users\CXH240016\Desktop\Movie_GT_Money_Combined
copy\trends_1.csv"
save "C:\Users\CXH240016\Desktop\Movie_GT_Money_Combined
copy\trends_1.dta"
clear all
……
import delimited "C:\Users\CXH240016\Desktop\Movie_GT_Money_Combined
copy\trends_79.csv"
save "C:\Users\CXH240016\Desktop\Movie_GT_Money_Combined
copy\trends_79.dta"
```

To merge all these data-files we started merging trends_1.dta and trends_2.dta and then proceeded with the rest of the files. Again we used Chat GPT to produce the repetitive code with the command merge m:m. The result was a data file containing the Google Trends data for all the 79 movies. This file was then merged with the file ‘movies_filtered.dta’.

```
merge m:m title using Trends_2.dta
drop _merge
……
merge m:m title using Trends_79.dta
drop _merge
merge m:m title using MovieDataFirstFinal.dta
```

The final data file ‘MovieDataFirstFinal.dta’ was created, containing the movie information scraped from omDb as well as the Google trends data. After browsing the data, we noticed that one movie did not have Google Trends data. We dropped this movie. Also, we converted several variables, such as imdb_rating and day, to numeric and set the data to be panel data,with the different days of the Google Trends data as the time dimension. To be able to create the drop rate for each movie we first needed a variable containing the google trends hits of day 1 as well as a variable containing the hits of day 21. After creating the drop rate, we only kept one observation for each movie. The decision to take the observation of day 21 ensures that only movies with data for this day are kept. Also movies where a drop rate could not be created were dropped.

```
gen gthits_day1 = gt_hits if day == 1
gen gthits_day21 = gt_hits if day == 21

bysort title_num (day): replace gthits_day1 = gthits_day1[_n-1] if
missing(gthits_day1)
bysort title_num (day): replace gthits_day21 = gthits_day21[_n-1] if
missing(gthits_day21)

gen drop_rate = (gthits_day1 - gthits_day21)/ gthits_day1 * 100
keep if day == 21
drop if missing(drop_rate)
```

The dataset is now ready to be used for a regression.

4.2 Descriptive statistics

The final dataset contains 65 movies. The average rating on IMDb is 6.337, while ‘Top Gun: Maverick’ has the highest ranking with an 8.2 and ‘Firestarter’ has the lowest ranking with 4.6. The rating on Rotten Tomatoes is comparable. The average here is 64%, which almost corresponds to the 6.337 average IMDb rating when scaled to percentages. Also, the same movies are ranked highest and lowest, with ‘Firestarter’ scoring 10% and ‘Top Gun:Maverick’ scoring 96%. The summaries are displayed in Tables 1 and 2.

Descriptive Statistics imdbrating

Variable	Obs	Mean	Std. Dev.	Min	Max
imdbrating	65	6.337	.752	4.6	8.2

Table 1: Summary of imdbrating

Descriptive Statistics rottentomatoes_rating

Variable	Obs	Mean	Std. Dev.	Min	Max
rottentomatoes rat~g	65	64.262	24.759	10	96

Table 2: Summary of rottentomatoes_rating

On average, the drop rate is 64.036%. This means that, on average, Google Trends search hits are 64% lower after 21 days compared to the release date. There are also movies with a negative drop rate – so their hits were higher three weeks after the first theatrical release of the movie compared to its release date. This is the case for the movies ‘Terrifier 2’, ‘Fall’, and ‘Violent Night’. The highest drop rate belongs to the movie ‘Brahmastra Part One: Shiva’, which suggests that the interest in this movie dropped the most after only 21 days of being released.

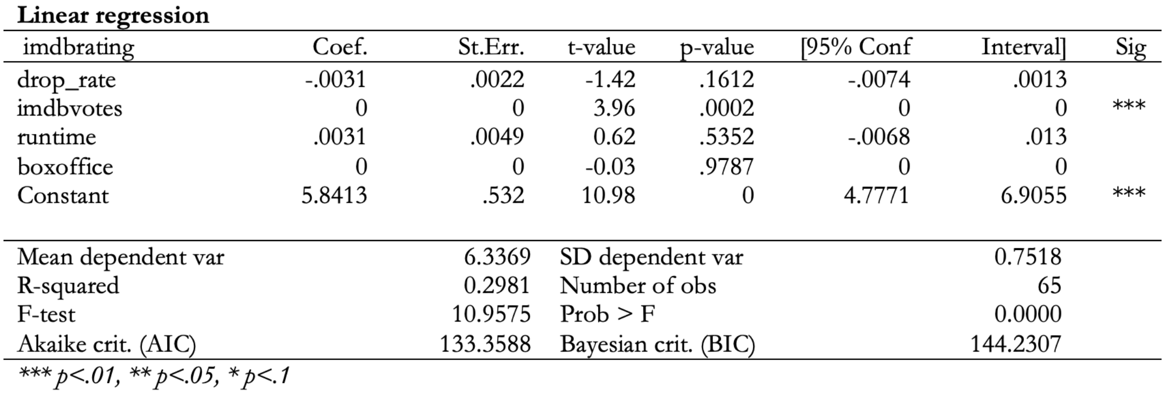
Descriptive Statistics Drop_rate

Variable	Obs	Mean	Std. Dev.	Min	Max
drop rate	65	64.036	37.247	-128.571	100

Table 3: Summary of drop_rate

4.3 Regression Analysis

We decided to perform two different regressions, as we had data for both the IMDb rating and the Rotten Tomatoes rating. Table 4 shows the regression with the IMDb rating as the dependent variable, the drop rate as the main explanatory variable, and the runtime, the number of votes and the box office revenue as control variables. The coefficient of -0.0031 can be interpreted as follows: Every 1% increase in a movie's drop rate is associated with a 0.0031-point decrease in the IMDb rating, holding all other factors constant.However, this effect is statistically not significant at any level, as the p-value is too high at 0.1612. The only significant variable in this model is the number of votes, as its p-value is 0.0002. But the coefficient is very small, due to rounding it is displayed as 0 in the table. Its original value is 0.00000233, indicating that for every additional IMDb vote, the IMDb rating increases by 0.00000233 points. This shows that IMDb ratings are influenced by the number of votes, although the real-world impact of this coefficient is very small. The R^2 0.2981 indicates that this model explains 29.81% of the variation in the IMDb rating.



Linear regression

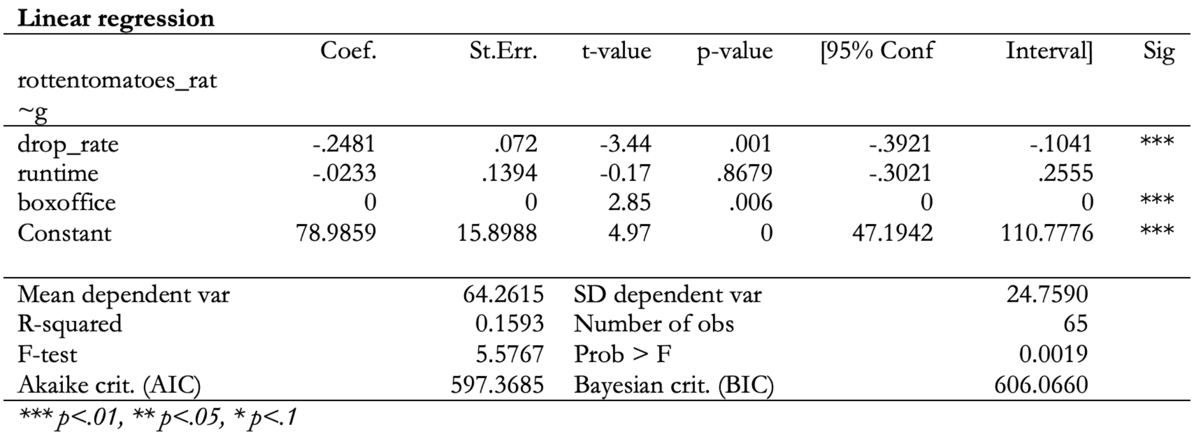
imdbrating	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
drop_rate	-.0031	.0022	-1.42	.1612	-.0074	.0013	
imdbvotes	0	0	3.96	.0002	0	0	***
runtime	.0031	.0049	0.62	.5352	-.0068	.013	
boxoffice	0	0	-0.03	.9787	0	0	
Constant	5.8413	.532	10.98	0	4.7771	6.9055	***

Mean dependent var	6.3369	SD dependent var	0.7518
R-squared	0.2981	Number of obs	65
F-test	10.9575	Prob > F	0.0000
Akaike crit. (AIC)	133.3588	Bayesian crit. (BIC)	144.2307

*** $p<.01$, ** $p<.05$, * $p<.1$

Table 4: Regression Model 1

Table 5 shows the regression with the Rotten Tomatoes rating as the dependent variable, the drop rate as the main explanatory variable, and the runtime and the box office revenue as control variables. We excluded the number of votes in this model, as it only accounts for the number of votes on the IMDb website. The coefficient of -0.2481 can be interpreted as follows: Every 1% increase in a movie's drop rate is associated with a 0.2481 percentage point decrease in the Rotten Tomatoes rating, holding all other factors constant. This effect is statistically significant at the 1% level, with a p-value 0.001. Regarding the other variables, similar to Model 1, only the box office revenue shows a very small significant effect on the rating. The R^2 0.1593 indicates that this model with its variables explains 15.93% of the variation in the Rotten Tomatoes ranking, making it weaker in predictive power compared to the first model.



Linear regression

	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
rottentomatoes_rat ~g							
drop_rate	-.2481	.072	-3.44	.001	-.3921	-.1041	***
runtime	-.0233	.1394	-0.17	.8679	-.3021	.2555	
boxoffice	0	0	2.85	.006	0	0	***
Constant	78.9859	15.8988	4.97	0	47.1942	110.7776	***
Mean dependent var		64.2615	SD dependent var			24.7590	
R-squared		0.1593	Number of obs			65	
F-test		5.5767	Prob > F			0.0019	
Akaike crit. (AIC)		597.3685	Bayesian crit. (BIC)			606.0660	

**** p<.01, ** p<.05, * p<.1*

Table 5: Regression Model 2

We also created the corresponding scatterplots for both models. The x-axis represents the drop rate and, while the y-axis represents either the IMDb rating or the Rotten

Tomatoes rating. Both plots show two extreme outliers, which could be excluded to potentially improve the results, as these outliers may distort the results. The plots visually confirm our regression results. There is no visible pattern in the variation of a movie's IMDb rating based on its drop rate. In contrast, the pattern in plot 2 is stronger, aligning with our regression result that the drop rate is significant for the Rotten Tomatoes rating.

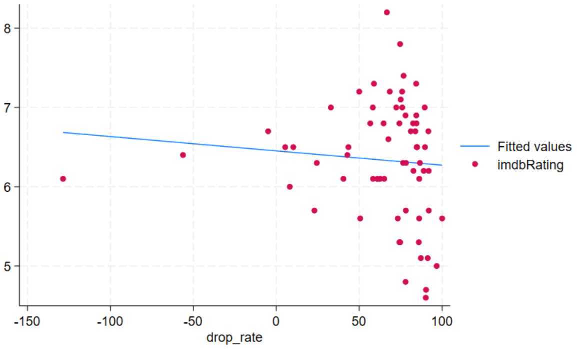


Figure 1: Scatterplot IMDb rating & drop rate

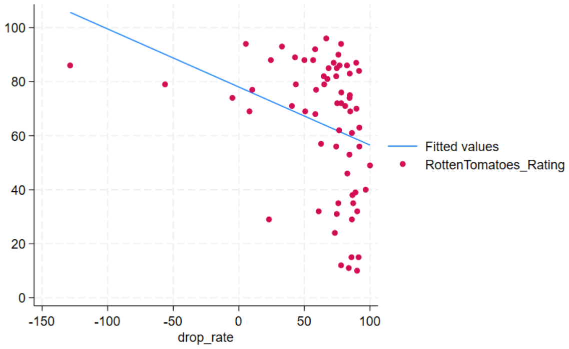


Figure 2: Scatterplot Rotten Tomatoes rating & drop rate

5. Discussion

This study is intended to investigate the relationship between the decline in Google Trends search interest during the initial 21 days post-release of a film and its audience reception, as quantified by IMDb and Rotten Tomatoes ratings. The results indicated that the correlation could have been stronger than thought, although a relationship exists between these variables. The drop rate was not statistically significant for IMDb ratings, yet it correlated significantly with Rotten Tomatoes ratings. The difference in these results suggests that various rating platforms show unique aspects of audience engagement, requiring taking into account platform-specific details in further analyses.

The data collection combines information from various sources, including IMDb, Box Office Mojo, and Google Trends. We utilized R packages, including rvest, httr, and gtrendsR, to extract and summarize metadata such as ratings, runtime, box office revenue, and daily search interest for 65 films released in 2022. This approach helped the development of a comprehensive dataset, but unfortunately some problems appeared. Lack of values, inefficient entries, and formatting issues needed thorough cleaning and filtering. Also, because the output did not reflect absolute search volumes, Google Trends' use of normalized and relative search values limited the depth of the analysis and made it more challenging to interpret search interest.

The challenges continued during the analysis stage, where statistical models demonstrated limited explanatory capacity. The regression models accounted for only a limited portion of the variance in movie ratings, with IMDb ratings demonstrating weaker correlations with the drop rate than Rotten Tomatoes ratings. Outliers complicated the analysis by distorting trends in scatter plots and diminishing the reliability of regression coefficients. These limitations highlight the intrinsic complexities of secondary data, especially when sourced from varied and unstructured origins.

Future studies should focus on overcoming these limitations by expanding data collection parameters and enhancing analytical techniques. Integrating supplementary variables, including social media sentiment, critical evaluations, and demographic data, may yield a more comprehensive insight into audience engagement. Also, expanding the analysis to encompass multiple years and global search trends could improve the generalizability of the results. Advanced methodologies, including machine learning algorithms, may be utilized to discern patterns and interactions that remain obscured in linear regression models. Future studies can expand upon the foundation established by this research, providing more accurate insights into the intricate relationship between digital engagement and audience reception.

Nevertheless, our study possesses limitations that must be recognized to contextualize our findings. Although innovative, dependence on Google Trends data creates a reliance on the accuracy and representativeness of this information, which is not consistently transparent. Moreover, our emphasis on data from a single year constrains the applicability of our findings to various timeframes or filmic trends. Subsequent research may rectify these limitations by utilizing a more extensive dataset over several years and investigating the influence of diverse cultural and social contexts on the relationship between search interest and movie ratings. Although using Google Trends as a data source gives researchers unique insights into search interest patterns, some significant issues must be resolved. The normalized and indexed data represent relative search interest rather than absolute figures, potentially obscuring the actual search volume and complicating comparisons across different periods or regions. Moreover, Google Trends uses a sampled dataset rather than the complete search population. This may result in gaps, especially for uncommon or low-frequency terms, which could go through biases due to geographic or demographic influences, such as differing levels of internet accessibility. Issues with search queries, such as misspellings and overlap, could result in gathering unrelated information. Historical data is limited for particular queries, and changes in Google's algorithms or data collection methods may lead to temporal discrepancies.

Moreover, analyzing or mimicking results is difficult due to Google Trends' need for more transparency regarding its sampling and normalization methods. API limitations, such as rate restrictions and prohibitions on concurrent term comparisons, complicate extensive research endeavors. Temporal and regional variations, including disparate trends across locations or inadequacies in data resolution, may constrain analysis. Moreover, increases in search interest may only sometimes correlate with positive engagement, as they can arise from negative publicity or unrelated events, raising concerns about the misinterpretation of trends. The lack of temporal clarity and the broad categorization of subjects can obscure detailed insights, limiting the depth of analysis.

We propose multiple directions for future research. A promising field is the incorporation of machine learning models to improve the predictive precision of consumer interest metrics. These models could integrate supplementary variables, including social media sentiment, critical reviews, and demographic data, to generate more sophisticated and precise predictions of cinematic success. Furthermore, long term studies could investigate the evolution of these relationships, providing insights into consumer behavior and the dynamics of digital engagement. By broadening the scope of research in these manners, academics and industry professionals can enhance their comprehension and utilization of the intricate relationship between digital content engagement and consumer behavior.

6. Conclusion

This study clarifies the correlation between digital engagement metrics, particularly Google Trends search interest, and audience reception as indicated by IMDb and Rotten Tomatoes ratings. We analyzed the decline in search interest during the initial 21 days following the release to identify patterns connecting ongoing digital engagement to audience approval. The study reveals a weak correlation between IMDb ratings and a stronger correlation with Rotten Tomatoes ratings, highlighting the potential of digital metrics as predictive instruments for understanding consumer behavior. These findings enhance the existing literature examining the relationship between online activity and cultural products such as films.

The research's limitations, however, highlight the necessity of carefully interpreting the findings. The reliance on normalized Google Trends data, the presence of outliers in the dataset, and the limitations of utilizing data from a single year negatively impacted the accuracy of the findings. Moreover, variations in audience behaviors on platforms such as IMDb and Rotten Tomatoes indicate that a uniform method for analyzing digital engagement needs to be revised. These complexities necessitate more advanced methodologies and datasets to accurately reflect the multifaceted nature of consumer interactions with digital content.

This study establishes a foundation for future investigation into digital engagement and media analytics convergence. Extending the parameters to encompass multiple years, wider geographic contexts, and supplementary variables such as social media sentiment and critical reception will improve predictive precision. Advanced methodologies, including machine learning, can uncover concealed patterns and enhance our comprehension of audience behaviors. As the digital landscape progresses, utilizing these insights will be crucial for industry professionals aiming to synchronize marketing strategies with audience preferences and for researchers striving to clarify the intricacies of digital consumer culture.

References

Box Office Mojo. (n.d.). *Yearly box office results*. https://www.boxofficemojo.com (November 16, 2024)

Cebrián, Eduardo, and Josep Domenech. 2023. "Is Google Trends a Quality Data Source?" *Applied Economics Letters* 30(6): 811–15. doi:10.1080/13504851.2021.2023088 (November 8, 2024)

Demir, Deniz, Olga Kapralova, and Hongze Lai. 2012. "Predicting IMDB Movie Ratings Using Google Trends." https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=fb53e9605997374f178359d3e1e86008dac6c28a (November 1, 2024).

Fritzsch, Benjamin, Kai Wenger, Philipp Sibbertsen, and Georg Ullmann. 2020. "Can Google Trends Improve Sales Forecasts on a Product Level?" *Applied Economics Letters* 27(17): 1409–14. doi:10.1080/13504851.2019.1686110 (November 8, 2024).

Google. Google News Initiative. 2024. *Google Trends*. https://newsinitiative.withgoogle.com/resources/trainings/advanced-google-trends/ (November 2, 2024).

Hand, Chris, and Guy Judge. 2012. "Searching for the Picture: Forecasting UK Cinema Admissions Using Google Trends Data." *Applied Economics Letters* 19(11): 1051–55. doi:10.1080/13504851.2011.613744 (November 4, 2024).

Jun, Seung-Pyo, Hyoung Sun Yoo, and San Choi. 2018. "Ten Years of Research Change Using Google Trends: From the Perspective of Big Data Utilizations and Applications." *Technological Forecasting and Social Change* 130: 69–87. doi:10.1016/j.techfore.2017.11.009 (November 4, 2024).

Massicotte, Pierre, and Dirk Eddelbuettel. 2022. gtrendsR: Perform and Display Google Trends Queries. R package version 1.5.1. https://CRAN.R-project.org/package=gtrendsR (December 3, 2024).

OMDb API n.d.. *The Open Movie Database*. from https://www.omdbapi.com (November 17, 2024)

OpenAI. 2024. *ChatGPT*. https://openai.com/chatgpt (November 30, 2024)
Shukla, Anuja, Aditya Yadav, and Shiv Kumar Sharma. 2022. "Predicting Movie Ticket Sales Using Google Trends: Implication of Big Data Analytics." *IUP Journal of Management Research* 21(1).https://openurl.ebsco.com/EPDB%3Agcd%3A11%3A23114539/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A156653694&crl=c&link_origin=scholar.google.com (November 4, 2024).

Silva, Emmanuel Sirimal, and Dag Øivind Madsen. 2022. "Google Trends." In *Encyclopedia of Tourism Management and Marketing*, ed. Dimitrios Buhalis. Edward Elgar Publishing, 446–47. doi:10.4337/9781800377486.google.trends (November 2, 2024).

Wickham, Hadley, and Davis Vaughan. 2024. tidyr: Tidy Messy Data. R package version 1.3.1. https://CRAN.R-project.org/package=tidyr (December 3, 2024).

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. readr: Read Rectangular Text Data. R package version 2.1.5. https://CRAN.R-project.org/package=readr (December 3, 2024).

Wickham, H. 2024. rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.4, https://github.com/tidyverse/rvest, https://rvest.tidyverse.org/ (November 5, 2024).

Wooldridge, Jeffrey M. 2013. *Introductory econometrics: a modern approach*. 5th ed. Mason, OH: South-Western Cengage Learning (November 5, 2024).

Attachment: Confirmation of Authorship

We hereby certify under oath that we have written this work independently and without unauthorized assistance, have not yet submitted it in whole or in part as an examination, and have not used any resources other than those specified. All parts of the work that were taken from other sources in wording or meaning are identified by stating their origin. This also applies to drawings, sketches, pictorial representations and the like, as well as to sources from the Internet. In the event of a violation, the seminar is deemed to have failed. We are aware that plagiarism is serious academic misconduct, which can be further sanctioned if repeated.

We have used ChatGPT for the purpose of proofreading this paper.

Dallas, December 3 2024