The Relationship Between Book Ratings and Popularity on Goodreads: A Statistical Analysis

Precious Ogu

BSOS326

Professor Neil Lund

05-14-2025

Goodreads, the world’s largest platform for book reviews and recommendations, influences reading habits for millions of users. In today’s digital landscape, where algorithms and online visibility shape consumer choices, understanding what drives a book’s popularity is more relevant than ever. A key question in understanding literary popularity is whether higher-rated books naturally attract more readers, or if other factors, such as cultural impact, controversy, or genre, play a more significant role.

This study investigates the relationship between a book’s average rating and its popularity, measured by the number of ratings on Goodreads. Specifically, we ask: Does a statistically significant relationship exist between a book’s average rating and its number of ratings, and what do outliers reveal about this relationship?

This question is more than academic. If early ratings primarily drive popularity, it may reinforce a "rich-get-richer" cycle, where a small number of highly rated books dominate visibility. This could make it harder for new, diverse, or independent voices to gain attention, regardless of quality. Additionally, readers may equate popularity with merit, even when other forces like marketing, controversy, or media tie-ins are driving attention. For publishers and authors, misinterpreting these dynamics could lead to decisions that prioritize hype over substance, further narrowing the kinds of stories that reach the public. Understanding this dynamic has far-reaching implications for equity, reader choice, and the future of publishing in an algorithm-driven world

Data Acquisition and Cleaning

To conduct this analysis, I collected data by web scraping Goodreads' *"Best Books Ever"* list. This section of the platform showcases books that have received high engagement from the Goodreads community, making it a useful proxy for examining the relationship between perceived quality (as reflected by average rating) and popularity (measured by number of ratings). This list is especially valuable because it includes a mix of widely known classics, contemporary bestsellers, and cult favorites, books that have stood the test of time or generated significant discussion online. By analyzing titles from this curated yet community-driven list, the study can focus on books that are both visible and culturally relevant, making the findings more applicable to today’s literary ecosystem. The two primary variables analyzed are:

1. Average Rating (avg_rating): A continuous variable (1-5 scale) representing the mean user rating for each book.
2. Number of Ratings (num_ratings): A discrete count of total ratings received, serving as a proxy for popularity.

.

Exploratory Data Analysis

Descriptive Statistics

To examine the relationship between a book's average rating and its popularity on Goodreads, I conducted an exploratory data analysis on a dataset of 200 books from the "Best Books Ever" list. This list provides a curated yet community-driven selection of titles that are both culturally relevant and widely visible, making it an appropriate sample for investigating how perceived quality relates to reader engagement.

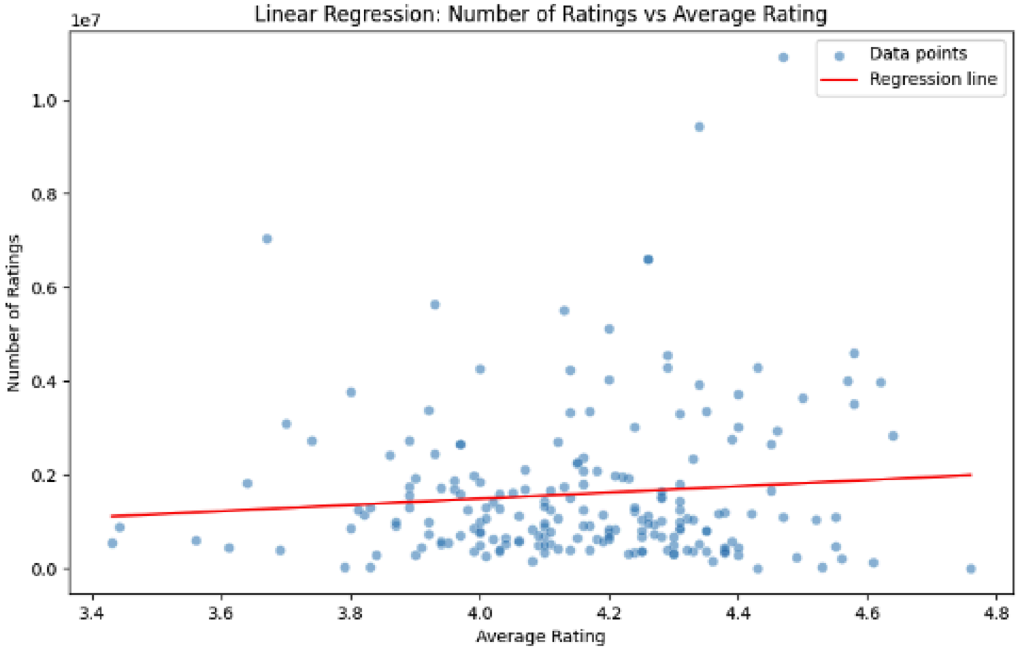
I calculated that the average rating (avg_rating) for the books ranged from 3.43 to 4.76, with a mean of approximately 4.16 and a relatively small standard deviation of 0.23. This indicates that most books maintain generally high ratings, reflecting a positive reception across the board. In contrast, the number of ratings (num_ratings), serving as a proxy for popularity, showed substantial variation. The least-rated book had just over 1,200 ratings, while the most popular book, *Harry Potter and the Sorcerer's Stone*, had over 10.9 million ratings. The mean number of ratings was around 1.59 million, with a large standard deviation indicating a heavily right-skewed distribution. This wide range underscores the vast disparity in reader engagement among even highly regarded books.

Top Books by Rating and Popularity

Furthermore, examining the extremes in the data revealed interesting insights. The highest-rated books, including *The Addiction Manifesto* (4.76), *A Court of Mist and Fury* (4.64), and various *Harry Potter* volumes (above 4.5), showcase titles that are critically well-received by readers. However, these do not always correspond to the most popular books in terms of sheer number of ratings. The most popular books, such as *Harry Potter and the Sorcerer's Stone*, *The Hunger Games*, *Twilight*, and *To Kill a Mockingbird*, amassed millions of ratings, establishing widespread visibility and engagement despite some having slightly lower average ratings.

Regression Analysis

A linear regression was then performed to quantify the relationship between average rating and number of ratings. The results indicated no statistically significant relationship (p = 0.183), with the model explaining less than 1% of the variance in popularity (R² = 0.009). While the positive coefficient suggested a trend where higher-rated books might attract more ratings, this effect was not strong or reliable enough to confirm a direct link. This lack of significance points to other influential factors beyond average rating in driving book popularity.



The regression model suggests a weak trend: higher-rated books may get more ratings, but many highly-rated books have few ratings, and some lower-rated books are very popular

Patterns and Implications

The EDA highlights a critical pattern: high average ratings alone do not guarantee a book's popularity on Goodreads. Popularity, measured by engagement in the form of number of ratings, is influenced by multiple complex factors, such as author fame, genre appeal, media exposure, and cultural relevance. Some highly rated books have niche followings without broad exposure, while others gain massive popularity despite moderate ratings. This disparity reveals the challenges faced by new, diverse, or independent voices, who often struggle to gain visibility and reader engagement regardless of the quality of their work. Without the backing of established marketing, mainstream media, or a large existing fan base, these authors may remain overlooked, making it harder for their stories to reach wider audiences. Consequently, the dynamics of popularity on platforms like Goodreads can perpetuate existing inequalities in literary exposure, privileging well-known authors and popular genres while marginalizing emerging voices. This pattern is especially important to recognize because it shapes not only which books become widely read but also which cultural narratives dominate public discourse. For readers, understanding these influences is vital to navigating digital book spaces critically, encouraging them to look beyond popularity metrics and discover hidden gems. For publishers, authors, and platform designers, these insights highlight the need to create more inclusive recommendation systems and outreach strategies that support diverse storytelling and equitable visibility in an increasingly algorithm-driven literary ecosystem.

Conclusion

This analysis reveals that while books on Goodreads generally maintain high average ratings, these ratings alone do not reliably predict a book's popularity as measured by the number of ratings. The weak and statistically insignificant relationship between average rating and reader engagement suggests that other factors, such as cultural impact, genre popularity, media adaptations, and author recognition, play a more decisive role in driving widespread readership. These findings underscore the complexity behind literary popularity in today's digital landscape and highlight the challenges faced by emerging or niche authors in gaining visibility. Recognizing these dynamics is crucial for readers, publishers, and platform designers who seek to promote diverse voices and ensure that quality storytelling reaches broad audiences beyond just the most visible or heavily marketed titles