BAIS3250 Data Wrangling Project Report

IMDb Movies and Goodreads Books Ratings

1. Introduction

Online reviews play a big role in how people make decisions, whether they are choosing what book to read next or deciding which movie to watch. But how well do these reviews actually reflect success or quality? I wanted to explore how audience ratings for books compare to the ratings of their movie adaptations. Does a highly rated book usually lead to a well-rated film? Or are there times when a book with average reviews turns into a movie that audiences really enjoy? I was also curious if other factors influence how these adaptations are received. For example, does the length of a book or the emotional tone of a movie review help explain differences in audience reactions? Can things like audience sentiment or review patterns help predict movie success?

For this project, I used data from Goodreads and web scraped top user reviews from IMDb. By combining these two sources, I was able to explore the connection between books and their movie adaptations and analyze what drives strong reception across both platforms.

2. Data

This project uses two main sources of real-world data: a Goodreads dataset that includes ratings and details for a variety of books, and review data scraped directly from IMDb[1] for the corresponding movie adaptations. These two sources provided both quantitative metrics like star ratings, vote counts, and page numbers, and qualitative insights from user reviews

2.1 Goodreads Books

The first dataset I used came from Goodreads and was downloaded from Kaggle. It contains information about over 40,000 books that were later adapted into films. For each book, the dataset includes the *title*, *author*, *average user rating*, *isbn*, *language*, *number of ratings*, *number of pages*, *publication date*, and *publisher*. The data was downloaded and saved as the file Goodreadsbooks.csv.

[1] https://www.imdb.com/list/ls026212430/

I used this dataset to understand how readers felt about each book and to look for trends in user ratings, length, and popularity. I also used it to compare the books to their movie versions. Most of the cleaning for this dataset involved removing missing or duplicate values, deleting unnecessary columns, standardizing the titles so they match IMDb movie titles, and formatting the numeric and text fields for analysis.

This data was imported and cleaned in the notebook titled esolie_bais3250project_integration.ipynb. It provides the foundation for identifying book traits and joining with the scraped movie data.

2.2 IMDb Movies

The second dataset was created by scraping IMDb to collect data on the movie adaptations of the books in the Goodreads dataset. For all 434 movies, I gathered the *title*, *release year*, *MPAA rating*, *runtime*, *average user rating*, *total vote count*, and the *top user review text*. The *IMDb ID* was scraped to uniquely identify and scrape each film.

To collect this data, I used Python with the selenium library and webdriver_manager to automate browser interaction and extract movie details directly from IMDb. I also used the re and requests libraries to handle URL formatting and support the scraping workflow. The entire process was completed in my notebook esolie_bais3250project_webscraping.ipynb, and the final dataset was exported as esolie_bais3250_project_webscrape_imdb.csv.

The dataset did require some cleaning, like reformatting runtimes, removing missing values, and dropping duplicates. This IMDb movie data added a review-focused and film-specific layer to the project, which helped me compare how books and movies are received by audiences and analyze patterns in how people describe their reactions in reviews

2.3 Combining IMDb Movies and Books

After cleaning both the Goodreads and IMDb datasets, I merged them to create one final integrated dataset. To make the merge work, I created a new column in each dataset called Cleaned_Title, where I removed extra formatting from the book and movie titles. For IMDb titles, I stripped out punctuation and numbering. For Goodreads titles, I cut off anything

that came after a forward slash. This helped standardize the titles and made it easier to match books to their corresponding movies.

Once the titles were cleaned, I used an inner merge on the Cleaned_Title column to combine the two datasets. This step was completed in the notebook esolie_bais3250project_integration.ipynb. After the merge, I dropped unnecessary columns like the original title fields and extra identifiers, renamed the remaining variables for clarity, and reordered the columns to make the dataset easier to read. I also went through the dataset to make sure all fields had the correct data types. I formatted numeric values like ratings and vote counts, and removed any duplicates so that each row represented one unique book and movie pair.

The final dataset contains 79 rows and 14 columns. I saved the cleaned and merged file as esolie_bais3250_project_integrated.csv. A description of each variable is provided in Table 1.

Table 1 Data Dictionary

Field Name	***Data Type***	***Description***
title	String	Title of the book and its movie adaptation (standardized for merge).
imdb_id	String	Unique identifier for the movie on IMDb.
release_year_movie	Integer	Release year of the movie.
movie_rating	String	MPAA rating of the movie (e.g., PG, PG-13, R).
duration_movie	String	Duration of the movie in hours and minutes (e.g., 2h 5m).
average_star_rating_movie	Float	Average user star rating of the movie on IMDb (scale of 1–10).
vote_count_movie	Integer	Total number of votes received on IMDb.
review_text	String	Scraped user reviews from IMDb.
book_authors	String	Name(s) of the author(s) of the book.
average_rating_book	Float	Average Goodreads rating of the book (scale of 1–5).
num_pages_book	Integer	Total number of pages in the book.
ratings_count_book	Integer	Total number of user ratings the book received on Goodreads.
publication_date_book	String	Date the book was originally published (in MM/DD/YY format).
publisher_book	String	Publisher of the book.

3. Analysis

3.1 Book Ratings and Movie Adaptation Reception

I explored whether books that are rated more highly tend to lead to better-received movie adaptations. I started by looking at the average Goodreads rating for each book in my dataset and created a histogram to visualize the distribution. This chart appears on the left side of Figure 1. The ratings were mostly clustered around 4.0, which was not too surprising since users on Goodreads are generally more positive when leaving book reviews. There were only a few books rated below 3.5 or above 4.5.

also examined how movies were rated overall by creating a histogram of average IMDb star ratings for the films in the dataset, shown on the right side of Figure 1. Compared to the books, movie ratings were more spread out. Most fell between 6 and 8 stars, with a few outliers on either end. The distribution was less symmetrical, which suggested that while many adaptations were received fairly well, audience responses to movies varied more than their responses to the books.

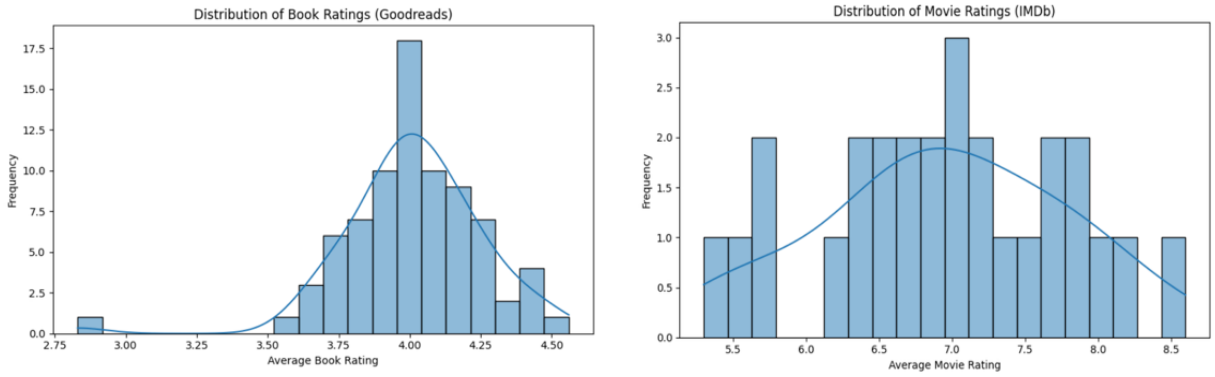

Figure 1: Rating Scores Distributions

To compare the ratings directly, I plotted the average book rating against the average IMDb rating and ran a Pearson correlation test. As shown in Figure 2, the result was a moderately positive correlation of 0.537 with a p-value well below 0.001. This means there was a statistically significant relationship between how well a book was received and how its movie performed with audiences. In general, if a book was well-liked by readers, its movie adaptation had a better chance of being well received too.

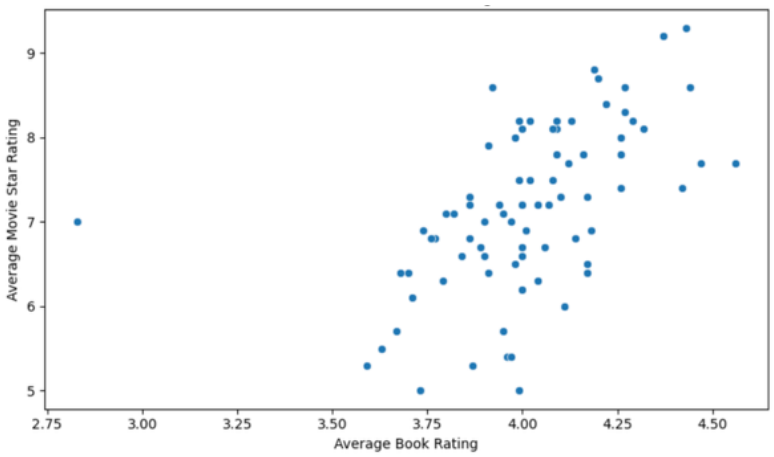

Figure 2: Book vs. Movie Ratings

To take it a step further, I tested whether movie ratings could be predicted using book-level features such as average rating, number of pages, and rating count. I ran four different regression models: Linear Regression, Ridge, Lasso, and Random Forest, using an 80/20 train-test split. Each model was evaluated using Mean Squared Error (MSE) and R-squared (R^2).

Model Results:

- Linear Regression: MSE = 0.878, R^2 = 0.228
- Ridge Regression: MSE = 0.920, R^2 = 0.191
- Lasso Regression: MSE = 1.235, R^2 = -0.086
- Random Forest: MSE = 1.143, R^2 = -0.005

Linear Regression performed the best, but even that model only had an R^2 of 0.228, meaning it explained very little of the variation in movie ratings. The negative R^2 values for Lasso and Random Forest suggest those models performed worse than a simple average predictor. While book features like rating, length, and popularity added a little insight, they were not strong enough on their own to predict how well a movie adaptation would be received.

3.2 MPAA Ratings and Audience Reception

I looked into whether a movie's MPAA rating had any effect on how well it was received by audiences. I focused on the three most common categories in the dataset: PG, PG-13, and R, and excluded any movies that were unrated or marked NC-17. As seen in Figure 3, I created a count

plot showing the number of movies in each category. Most movies were rated either PG-13 or R, while only a few were PG.

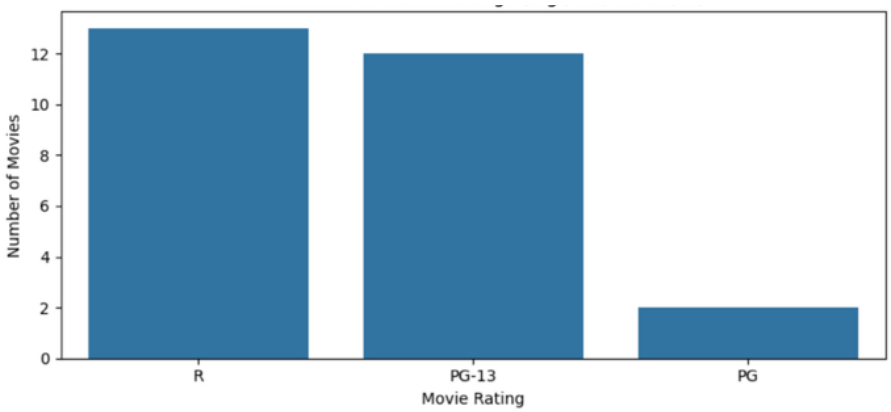

Figure 3: Distribution of Movie Ratings (e.g., PG, PG-13, R)

Next, I created a boxplot comparing IMDb ratings across these MPAA categories. This boxplot is shown in Figure 4. PG rated movies had the highest median rating and the smallest spread, suggesting that audiences rated them consistently and positively. PG-13 movies had more variation and slightly lower scores, while R rated movies showed the widest range but still performed reasonably well. I also ran a Spearman correlation to test for a relationship between MPAA rating and IMDb score. The correlation was 0.124 with a p-value of 0.3155, which was very weak and not statistically significant. Even though the test did not support a strong relationship, the visual trends suggested that movies with less restrictive ratings might be received a bit more positively.

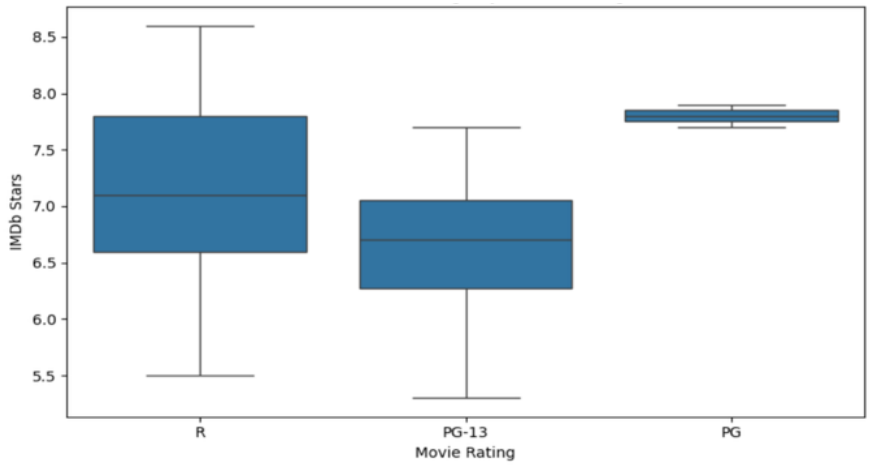

Figure 4: IMDb Star Ratings by MPAA Movie Rating

3.3 Sentiment of IMDb Reviews

For this section, I wanted to see if the way people describe movies in their IMDb reviews had any connection to the overall star rating those movies received. To do that, I used sentiment analysis to calculate a polarity score for each review. I then looked at the distribution of those scores to see how sentiment tends to show up in this dataset. As shown in Figure 5, most polarity scores fell between 0.1 and 0.3, meaning the average review leaned slightly positive but not overly emotional.

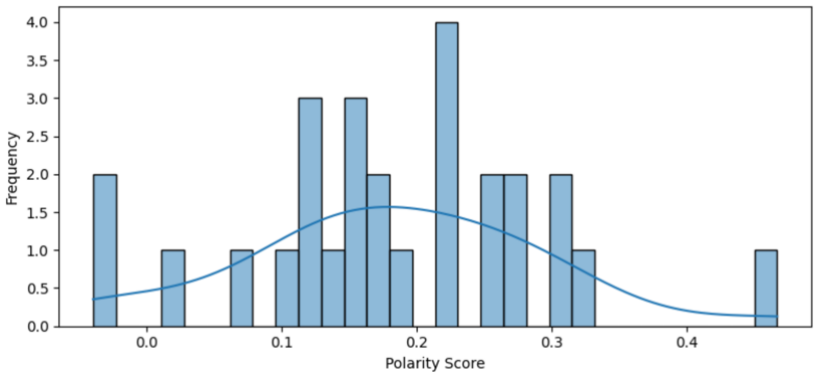

Figure 5: Distribution of Review Sentiment Polarity

Then I compared the polarity scores to the average IMDb rating to see if more positive reviews led to higher ratings. The correlation was 0.244, with a p-value of 0.045, which is still a pretty weak relationship but at least statistically significant. The scatterplot in Figure 6 shows a small upward trend, suggesting that more positively worded reviews might be tied to better ratings.

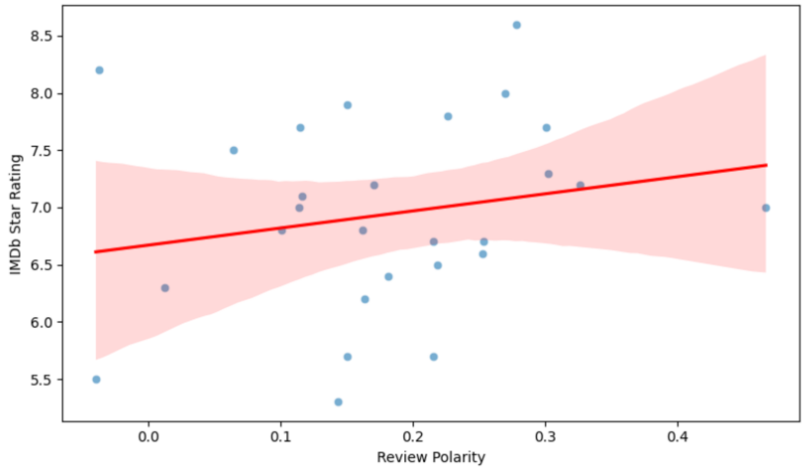

Figure 6: IMDb Rating vs. Review Sentiment

To better understand how people described the movies, I created word clouds based on the sentiment scores. Figure 7 shows the most common words from positive reviews. Words like "people," "good," "audience," and "life" came up frequently, which makes sense for reviews that felt meaningful or uplifting. Figure 8 shows the most common words from negative reviews, including terms like "horror," "think," "childhood," and "weird," reflecting a tone that leaned more critical or nostalgic. Although sentiment scores did not strongly predict audience ratings, they offered valuable insight into how people describe the movies they connected with or did not

Figure 7: Word Cloud of Positive IMDb Review Top Words

Figure 8: Word Cloud of Negative IMDb Review Top Words

4. Conclusion

In this project, I looked at how books and their movie adaptations are received by audiences. Using data from Goodreads and IMDb, I explored three main questions: whether well-rated books lead to well-rated movies, if MPAA ratings affect how a film is perceived, and whether the sentiment of a review can help explain a movie's overall reception.

1. *To what extent do audience ratings of movie adaptations reflect the reception of their original books?*

   There was a moderate positive correlation between Goodreads book ratings and IMDb star ratings, meaning that books rated more highly tend to have movies that were also well received. However, when I tried to predict movie ratings using book features, the results were limited. Even the best performing model, linear regression, only explained a small portion of the variation. While strong book ratings help, they are not enough on their own to predict movie success.

2. *Are certain MPAA ratings (PG, PG-13, R) associated with higher or lower average audience scores for book-to-movie adaptations?*

   I compared IMDb ratings across PG, PG-13, and R rated movies. PG rated movies had the highest median score and the tightest range, while R rated movies had the most variation. A Spearman correlation showed a very weak and statistically insignificant relationship. Although the data did not show anything strong, the visual trends suggested that less restrictive ratings might be connected to slightly better reception.

3. *How does review sentiment polarity relate to IMDb star rating between positively and negatively received movie adaptations?*

   The correlation between review sentiment and IMDb rating was weak but statistically significant. More positive wording was slightly linked to higher ratings. Word clouds showed that users who liked the movies used more reflective language, while negative reviews leaned nostalgic or critical. Though sentiment analysis is not a strong predictor on its own, it added useful context to how audiences describe their reactions.

This project has a few limitations, including the small number of book-to-movie pairs, limited review data, and missing details that could have added more context to the analysis. While I was able to analyze audience ratings and review sentiment, those factors alone did not strongly predict how well a movie adaptation would be received. Future work could include expanding the dataset to cover more adaptations, adding critic reviews and box office data, or analyzing social media conversations to capture a broader picture of audience response.