D214 Performance Assessment Report

Task 2: Data Analytics Report and Executive Summary

Instructor: Dr. Smith

Student Name: Nicole Haibach

Student ID: 001260374

Email: nhaibac@wgu.edu

Romance And Action Reader Trends

# Research Question

## Summarize the original real-data research question you identified

Between Action and Romance genres, which genre has the highest average rating and number of reviews across the decades?

Null hypothesis-There is no statistically significant difference between the mean rating scores (or number of reviews) of books in the Action genre and the Romance genre.

Alternate Hypothesis- There is a statistically significant difference between the mean rating scores (or number of reviews) of books in the Action genre and the Romance genre.

The research question investigates whether there is a statistically significant difference between the mean rating scores and the number of reviews for books in the Action and Romance genres, specifically examining how these metrics have evolved across decades (1980–2023). This question is justified by the increasing accessibility of reader engagement data from platforms like Goodreads, which provides a wealth of information to analyze trends in literary preferences. Understanding how these genres perform over time and differ in reader reception can offer valuable insights for publishers, authors, and marketers in making data-driven decisions about content creation and promotion strategies.

This research exists in the broader context of the evolving publishing industry, where consumer preferences are heavily influenced by cultural trends, technological advances, and the rise of digital book formats. Both Action and Romance genres have distinct, loyal readerships, but their popularity and engagement levels may fluctuate over time, reflecting changes in societal tastes and literary consumption patterns.

The hypothesis guiding this research posits that there is a statistically significant difference in the mean rating scores and number of reviews between the two genres. The null hypothesis assumes no significant difference between the group means, while the alternate hypothesis asserts a significant difference. This hypothesis will be tested through statistical methods to determine whether the observed variations are meaningful and to identify patterns in reader engagement and genre performance over the decades.

# Data Collection

## Report on your data-collection process by describing the relevant data you collected

For this project, the data was sourced from the Top Goodreads Books Collection (1980-2023) dataset available on Kaggle. This dataset includes relevant attributes such as book titles, genres, average rating scores, number of reviews, publication dates, and other metadata needed to analyze reader engagement trends across decades. The dataset was downloaded in a structured format (CSV), and additional preprocessing steps were taken to clean and prepare it for analysis. This included addressing missing values, standardizing genre labels, and organizing publication dates into decades.

The primary advantage of using this dataset was its comprehensiveness and accessibility. It provided a rich set of attributes for analysis, covering multiple aspects of reader engagement like ratings, reviews, and genres, all in a clean, pre-compiled format. This saved significant time in gathering raw data manually from various sources, allowing more focus on analysis and hypothesis testing.

One limitation of the methodology was the potential lack of representativeness. The dataset is restricted to Goodreads data, which may not fully capture global readership trends or engagement on other platforms. Goodreads users might have demographic biases (e.g., predominantly English-speaking audiences) that could affect the generalizability of the findings.

A key challenge was dealing with inconsistencies in genre classification, as some books were assigned multiple or ambiguous genres. This was addressed by standardizing genres based on dominant themes or manually validating classifications for the most frequently occurring genres. Additionally, some entries had missing values for critical attributes like ratings or reviews. These were handled by either imputing data where possible or excluding incomplete entries from the analysis to ensure data quality.

Overall, the data-collection process ensured the dataset was ready for meaningful and reliable analysis, overcoming its limitations through careful preprocessing and validation.

# Data Extraction and Preparation

## Describe your data-extraction and -preparation process

### Missing Values:

To ensure the dataset was ready for analysis, I began by checking for missing values using R's colSums(is.na()) function, which counts the number of missing values in each column. This step revealed that while some columns, such as price and genres, contained missing data, none of the columns critical to the analysis—such as rating\_score, num\_reviews, num\_ratings, and genres—had missing values. Therefore, no imputation or further handling of missing data was required for the variables used in the study. For data extraction, I used the read\_excel() function from the readxl package to load the dataset into R, as it efficiently handles Excel files and preserves formatting. For data preparation, I relied on the tidyverse library, which provides a cohesive framework for manipulating, cleaning, and transforming data, and lubridate to handle and standardize date formats.

These tools and techniques were chosen for their versatility and ease of use in handling structured datasets. One advantage of using R and its libraries is the seamless integration of data extraction, cleaning, and preparation into a single workflow, allowing for a streamlined process. However, a disadvantage is that handling complex or text-heavy columns, such as genres, often requires custom scripting or advanced regular expressions, which can be time-consuming. Despite this, the combination of tools ensured the dataset was clean, organized, and ready for statistical analysis.

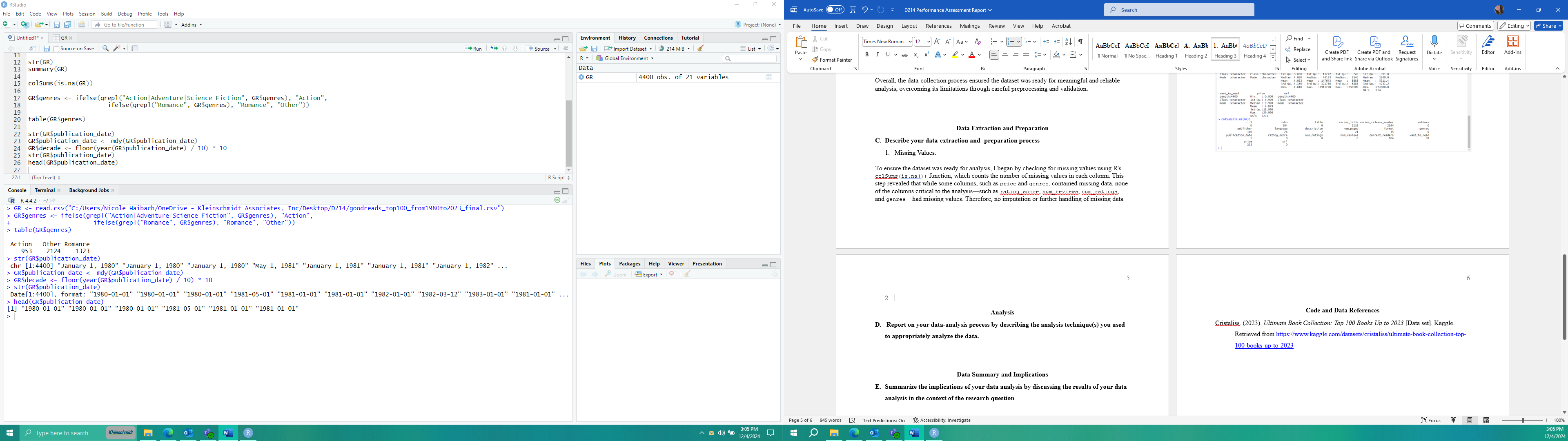
A screenshot of a computer

Description automatically generated

### Data Cleaning:

To prepare the dataset for analysis, the genres column was simplified into three categories: Action, Romance, and Other. This was achieved using the ifelse() function with grepl() to identify specific keywords in the genre descriptions. Books with "Action," "Adventure," or "Science Fiction" in their genres were categorized as Action, while those containing "Romance" were grouped as Romance. All other genres were labeled as Other. This categorization streamlined the dataset, making it easier to analyze trends across major genre groups. The table() function was then used to verify the distribution of books among the newly created categories. This approach ensures clarity in genre comparisons, though it requires careful selection of keywords to avoid misclassification.

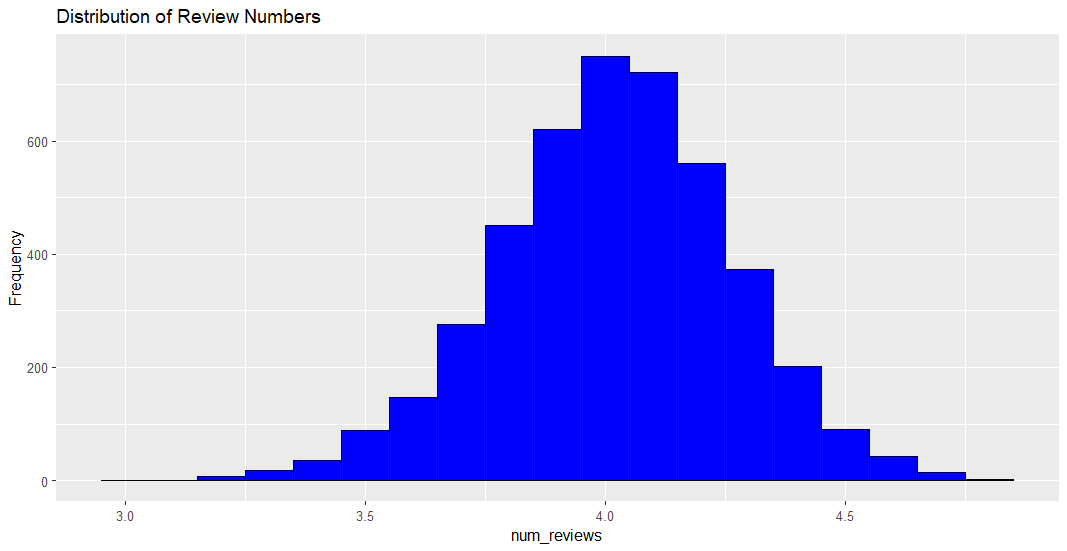
For temporal data preparation, the publication\_date column, originally in character format (e.g., "January 1, 1980"), was converted to proper date objects using the lubridate::mdy() function. This allowed for easy manipulation and analysis of the dates. A new decade column was then created by extracting the year with the year() function and rounding it down to the nearest decade using floor(). This grouping enables trend analysis over time, helping to identify patterns in book publication and popularity. The str() and head() functions were used to verify that the publication\_date column was correctly transformed and that the new decade column was accurately calculated. This method provides a systematic way to clean and organize temporal data, though it may introduce challenges if dates are in inconsistent formats. However, the use of lubridate ensures that date parsing is robust and efficient.

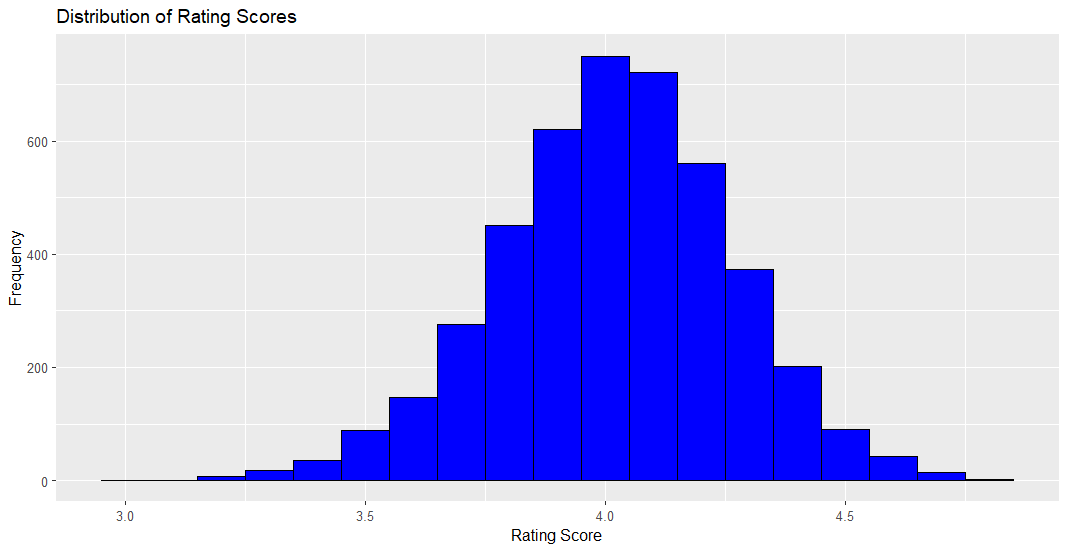


### Visualize:

For the visualization portion, histograms were created using the ggplot2 package in R to display the distributions of number of reviews and rating scores. These graphs provide a clear overview of the data, showcasing the frequency of books across different ranges of reviews and ratings. The histograms were plotted with the number of reviews and rating scores on the x-axes, and the frequency of books on the y-axes, making it easy to identify trends such as the most common ranges or any potential outliers.

The ggplot() function was chosen for its flexibility and ability to produce high-quality, customizable visualizations. The histograms effectively summarize the data and allow for quick insights into the distributions of key variables, such as the concentration of books with high ratings or review counts. One advantage of this technique is its ability to visually identify patterns and anomalies, which aids in understanding the dataset's overall structure. However, a disadvantage is that fine-tuning visualizations for large datasets can require additional customization and computational resources, especially if the dataset is complex or includes many variables. Despite this, ggplot2 remains a powerful tool for creating clear and impactful data visualizations.





# Analysis

## Report on your data-analysis process by describing the analysis technique(s) you used to appropriately analyze the data.

The dataset was analyzed to calculate the average rating scores, number of reviews, and book counts for the Action and Romance genres. The group\_by() and summarize() functions were used to compute these metrics, providing an overview of how these genres perform on Goodreads. The analysis revealed that Romance books tend to have more reviews on average, while the average ratings are similar between the genres. These descriptive statistics establish a foundation for deeper comparisons and trend analyses. Provides a clear and concise summary of key metrics, serving as a foundation for deeper analysis. Descriptive statistics alone do not account for variability or statistical significance, limiting their interpretive power.

> print(descriptive\_stats)

# A tibble: 3 × 4

genres avg\_rating avg\_reviews book\_count

<chr> <dbl> <dbl> <int>

1 Action 4.04 8615. 953

2 Other 4.00 7380. 2124

3 Romance 4.05 11505. 1323

To explore temporal trends, the dataset was grouped by genres and publication decades. The publication dates were converted into decades using the lubridate::year() function and then rounded down with floor(). The grouped data was summarized to calculate decade-wise average ratings and reviews for each genre. This aggregation provided insights into how reader preferences and engagement evolved over time, enabling a more nuanced understanding of genre performance. Enables detailed temporal analysis, uncovering long-term trends and shifts in reader preferences. Aggregation can obscure individual data points, leading to potential loss of granularity and outlier information.

> print(aggregated\_data)

# A tibble: 15 × 4

# Groups: genres [3]

genres decade avg\_rating\_decade avg\_reviews\_decade

*<chr>* *<dbl>* *<dbl>* *<dbl>*

1 Action 1980 4.02 2586.

2 Action 1990 4.05 4545.

3 Action 2000 4.02 10369.

4 Action 2010 4.07 20236.

5 Action 2020 4.09 16573.

6 Other 1980 4.01 2235.

7 Other 1990 4.00 3053.

8 Other 2000 3.96 8311.

9 Other 2010 4.03 20935.

10 Other 2020 4.04 16046.

11 Romance 1980 3.95 2627.

12 Romance 1990 3.99 3890.

13 Romance 2000 3.94 8977.

14 Romance 2010 4.11 11938.

15 Romance 2020 4.09 23979.

Two-sample t-tests were conducted to statistically compare the mean rating scores and number of reviews between Action and Romance genres. The tests showed whether the differences observed in the descriptive analysis were statistically significant. If normality assumptions were not met, a Mann-Whitney U test would have been used as a non-parametric alternative. These statistical tests added rigor to the analysis by validating whether the observed differences were meaningful. Provides robust evidence for whether differences between groups are statistically significant. Relies on assumptions (e.g., normality), which may require additional transformations or alternative tests if violated.

> print(t\_test\_rating)

Welch Two Sample t-test

data: rating\_score by genres

t = -1.3508, df = 2163.3, p-value = 0.1769

alternative hypothesis: true difference in means between group Action and group Romance is not equal to 0

95 percent confidence interval:

-0.031455120 0.005796497

sample estimates:

mean in group Action mean in group Romance

4.038168 4.050998

> print(t\_test\_reviews)

Welch Two Sample t-test

data: num\_reviews by genres

t = -3.3506, df = 2247.9, p-value = 0.0008198

alternative hypothesis: true difference in means between group Action and group Romance is not equal to 0

95 percent confidence interval:

-4561.114 -1193.222

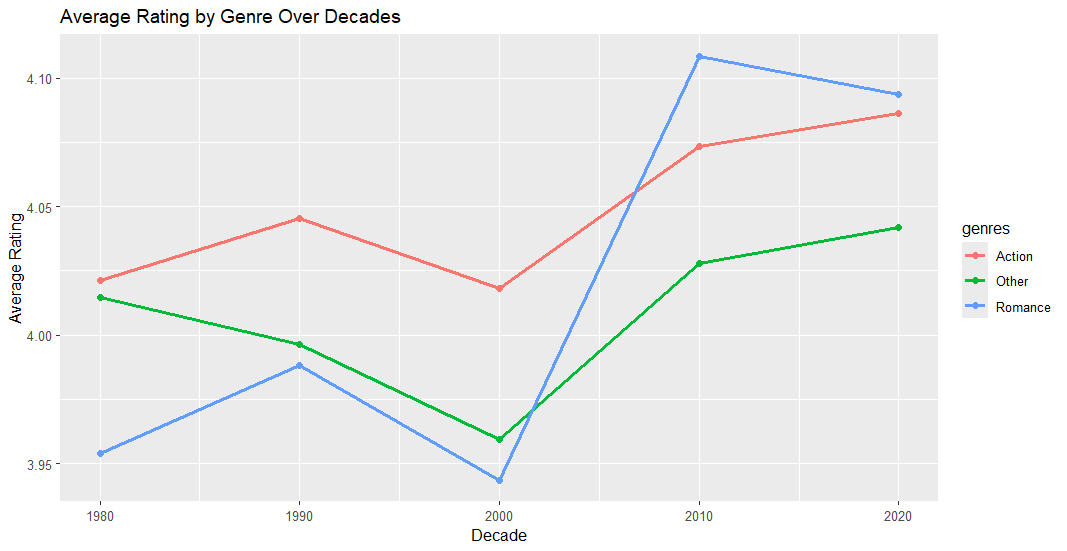
sample estimates:

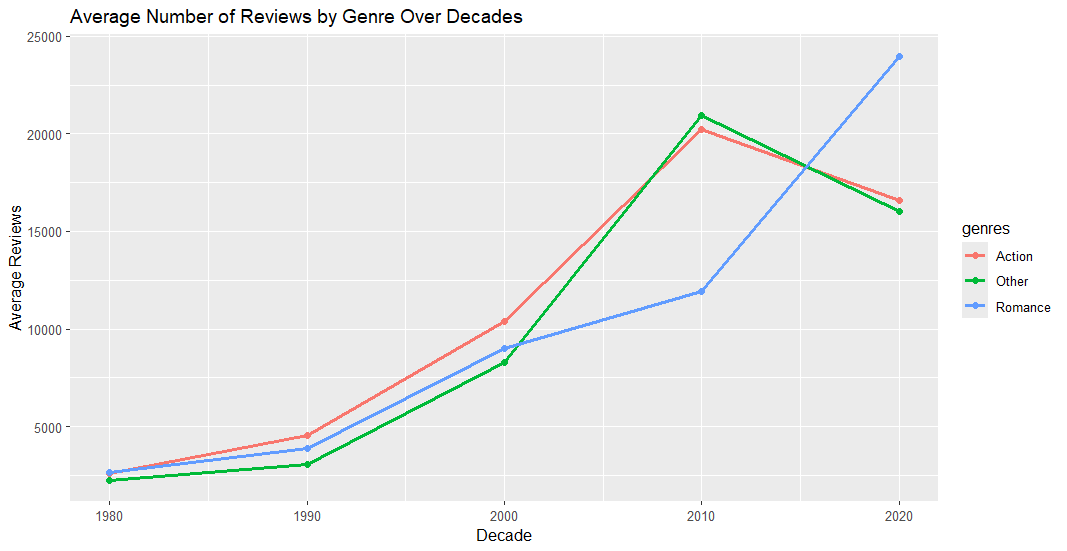
mean in group Action mean in group Romance

8627.335 11504.503

Finally, line graphs and bar charts were created to visualize the temporal trends in average ratings and reviews. These visualizations provided a clear picture of how the performance of Action and Romance genres fluctuated over decades. Line graphs highlighted trends in ratings, while bar charts emphasized differences in review counts across time. These visualizations made the results more interpretable and highlighted key patterns in reader engagement.

By combining statistical analysis and visualization, this process ensured robust and actionable insights into the performance of Action and Romance genres. Offers a clear and intuitive way to communicate trends and differences, making complex data accessible. Creating and fine-tuning visualizations can be time-intensive, especially for large datasets or multi-faceted trends





# Data Summary and Implications

## Summarize the implications of your data analysis by discussing the results of your data analysis in the context of the research question

The analysis revealed valuable insights into the performance of Action and Romance genres on Goodreads from 1980 to 2023. Romance books consistently received more reviews than Action books, especially in recent decades, reflecting a higher level of reader engagement. However, the average rating scores for both genres were similar, with Romance slightly edging out Action in the most recent decades. This suggests that while both genres are well-received by readers, Romance books have broader appeal and attract more reader feedback. These results answer the research question by highlighting that Romance outperforms Action in reader engagement while maintaining comparable quality in terms of ratings.

One limitation of the analysis is its reliance on aggregated metrics, such as average ratings and reviews, which may obscure underlying variability or outliers in the dataset. Additionally, the analysis focused only on Goodreads data, which predominantly reflects the preferences of English-speaking readers and may not fully represent global trends. These factors could limit the generalizability of the findings to other markets or cultural contexts.

Based on these results, authors and publishers should consider investing more in the Romance genre, particularly in marketing and promotions to sustain its strong engagement. For the Action genre, focusing on increasing visibility and enhancing reader interaction could help narrow the gap in reviews and reach broader audiences.

Future studies could analyze how specific subgenres (e.g., historical romance, sci-fi action) perform relative to the broader Romance and Action genres. This would provide a more detailed understanding of reader preferences within each genre.

A deeper investigation into external influences, such as the impact of media adaptations or marketing strategies on reader engagement, could provide additional context for the observed trends. Analyzing how these factors correlate with reviews and ratings would offer actionable insights for publishers.

**Code and Data References**

Cristaliss. (2023). *Ultimate Book Collection: Top 100 Books Up to 2023* [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/cristaliss/ultimate-book-collection-top-100-books-up-to-2023>