# The Evolution of Gender Representation in Literature: A Statistical Analysis of Goodreads Data

Ayesha A. Jan (3122167)

Berlin School of Technology (BST)

BSc Computer Science

Statistics (BST-BCS-12a)

Dr. Kristian Rother

## Abstract

In today's literary landscape, gender representation plays a pivotal role in the publishing industry, influencing author visibility and book popularity. This report explores gender dynamics in literature through Goodreads metadata, analyzing over 20 000 books to uncover patterns in author demographics, genre preferences, and book ratings.

The findings reveal a nearly balanced gender distribution, despite previous centuries being dominated by male authors, and female authors surpassing them only in the 21st century. Genre analysis shows that female authors dominate Romance and Young Adult, while male authors lead in Science Fiction and Classics.

Regression analysis indicates that author ratings strongly predict book ratings, while time series analysis highlights the growing influence of self-publishing platforms on female authors' rise. These insights shed light on the evolving landscape of literature, reflecting broader societal shifts toward greater gender inclusivity.


**Keywords**: Gender representation, Literature, Goodreads analysis, Author demographics, Book ratings

## Introduction

Gender representation in literature has long been a subject of discussion, shaping the landscape of publishing and influencing author recognition, book popularity, and reader engagement. The literary world, like many creative industries, has historically been male-dominated, with female authors often facing barriers to visibility and critical acclaim.

However, shifts in societal attitudes and the rise of digital platforms, such as Goodreads, have provided new opportunities to analyze and understand gender dynamics in literature.

Goodreads is a popular social networking platform for readers, where they can discover, rate, and review books, and for authors to promote their books, track ratings and reviews, and engage with their audience.

It also offers a rich dataset containing information on author demographics, book genres, ratings, and reader reviews. By examining these factors, valuable insights can be gained into gender disparities in literature, the extent of genre specialization among male and female authors, and whether reader reception differs based on an author's gender, the genre of the book, or the year it was published.

Understanding the role of gender in literature is essential for recognizing historical trends, addressing biases, and fostering a more inclusive publishing industry.

This statistical report explores the key patterns in literary publishing, focusing on how gender influences book ratings, genre dominance, and reader preferences. Specifically, this paper aims to answer the following research questions:

1. Is there a gender disparity in the number of books published by male and female authors?
2. Do male and female authors prefer different genres, and if so, which genres?
3. How do gender, genre, and publication date influence the popularity of books?

# Methodology

## Data

The dataset that has been analysed in this data report consists of author and book information that was web-scraped from Goodreads' most popular lists and contains books dating from 720 BC to 2019 AD. This dataset consists of a total of 22 891 observations, wherein the information of each book has the following 20 variables:

1. author_average_rating: The average is calculated as the sum of all of the author's book ratings divided by the total number of ratings.
2. author_gender: Whether the author is referred to as a woman or a man.
3. author_genres: The genres of the books written by the author.
4. author_id: A unique ID for each author.
5. author_name: The name(s) of the author(s) of the book.
6. author_page_url: The link to the author's Goodreads page.
7. author_rating_count: The total amount of ratings (1-5 stars) the author has received on all their books.
8. author_review_count: The total amount of reviews the author has received on all their books.
9. birthplace: The country the author was born in.
10. book_average_rating: The average of the book rating is calculated as the sum of all of the ratings from 1 to 5 divided by the total number of ratings.
11. book_fullurl: The link to the book on Goodreads.
12. book_id: The unique ID of each book on Goodreads.
13. book_title: The title of the book.
14. genre_1: The main genre of the book.
15. genre_2: The secondary genre of the book.
16. num_ratings: The total number of ratings the book contains.

17. num_reviews: The total number of reviews of the book.
18. pages: The number of pages each book contains.
19. publish_date: When the book was published.
20. score: calculated based on a combination of the number of votes a book receives on a list and its overall rating.

The dataset was uploaded onto Kaggle by Ben Rosen in 2019 and can be found here: https://www.kaggle.com/datasets/brosen255/goodreads-books

### Methods

In this study, a combination of data processing, statistical analysis, and visualization techniques was utilized to examine the evolution of gender representation in literature using Goodreads data. Beginning with descriptive analyses, categorical and numerical variables were explored to assess their distribution, central tendency, and variability. Correlation analysis and regression modeling were used to investigate relationships between author ratings, book ratings, and other relevant factors. Additionally, time series analysis was performed to identify trends in gender representation over time. Data visualization techniques, including boxplots, bar graphs, histograms, and line graphs, were used to illustrate the findings.

## Results

### Categorical Columns

After analysing and evaluating the categorical columns the results are as follows:

1. author_gender: This column consists of 2 unique values (male and female) with the most frequent value being male (12 201) which is slightly more than female (10 491). Therefore the distribution is nearly balanced. There are 2 distinct groups in this column (male and female).
2. author_genres: This column has 1 803 unique values, with the most frequent genres being Romance (2 595 books) and Young Adult (2 511 books). The distribution is highly diverse, but Romance and Young Adult are dominant. There are major subgroups of Romance, Young Adult, and Fiction authors.
3. author_name: This column consists of 12 156 unique values, with some authors appearing multiple times. The most frequent values are Sara Gruen, Margaret Mitchell, Seth Grahame-Smith, Lenore Appelhans, and Barbara Kingsolver, each appearing 6 times. The distribution is mostly unique, but a few authors have multiple books. The major subgroup consists of highly prolific authors.

4. author_page_url: This column has 12 156 unique values, with some authors having multiple books linked to the same page. The most frequent values are URLs for the same authors mentioned above. The distribution is mostly unique, but some URLs appear multiple times. The major subgroup consists of authors with multiple books.
5. birthplace: Consists of 435 unique values. The most frequent values are \n (4 008 times) indicating a null/missing value, United States (11 471 times), and United Kingdom (2 056 times). The distribution is mainly dominated by U.S. authors. The major subgroup is authors from the U.S. and United Kingdom.
6. book_fullurl: This column consists of 16 830 unique values, with some books appearing multiple times under different editions. The most frequent values are Dracula, Water for Elephants, The Road, Dualed, and The Outsiders, each appearing 6 times. The distribution is mostly unique, but some books have multiple records. The major subgroup consists of popular books with multiple editions.
7. book_id: This column has 16 830 unique values, with the most frequent values being 17245, 12144569, and 6288, appearing 6 times. The distribution is mostly unique, but some books appear multiple times. The major subgroup consists of books with multiple editions or versions.
8. book_title: Consists of 16 350 unique values, with the most frequent title being Forbidden (16 times). The distribution is mostly unique, but some book titles are repeated, possibly due to multiple editions or common names. The major subgroup consists of books with commonly used titles.
9. genre_1: This column consists of 124 unique values, with the most frequent genres being Fantasy (3 397 books), Romance (3 397 books), and Young Adult (2 497 books). The distribution is heavily concentrated around Fantasy, Romance, and Young Adult. The major subgroup consists of books in these dominant genres.
10. genre_2: This column consists of 157 unique values, with the most frequent values being Romance (3 042 books) and Fiction (2 930 books). The distribution follows a similar pattern to primary genres, reinforcing the dominance of Romance, Fantasy, and Young Adult. The major subgroup shows that some books have cross-genre appeal, leading to overlap in the Romance, Fiction, and Fantasy categories.

**Final Insights:**

Most of the columns contain mostly unique values, except certain ones like genres and gender. The dataset includes a relatively balanced mix of male and female authors. There is a concentration around a few major genres, with Fantasy, Romance, and Young Adult dominating. A high number of books were published in the 2010s,

suggesting the dataset may focus more on recent literature. Some authors and books appear multiple times, possibly due to multiple editions or popular votes.

## Numerical Columns

After analysing and evaluating the numerical columns the results are as follows:

1. author_average_rating: The mean is 3.9 with a standard deviation(std) of 0.2, indicating low variability in author ratings. The range is from 1.8 to 5.0, but 75% of authors have ratings above 3.8, meaning most authors receive generally high ratings. The distribution is slightly right-skewed, as the median (4.0) is slightly above the mean. Possible outliers include authors with ratings close to 1.8.
2. author_id: The mean is 3.2 million, with a std of 3.9 million, indicating a large spread in author IDs. The range is from 4 to 18.8 million, showing that author IDs vary significantly. The distribution is right-skewed, with a long tail of authors having very high IDs.
3. author_rating_count: The mean is 172 031.9, but the std is 654 690.2, suggesting a huge variation in author popularity. The range is from 6 to 21 million, meaning some authors have millions of ratings, while others have very few. The distribution is highly right-skewed, as the median (24 635) is much lower than the mean. Outliers include authors with millions of ratings.
4. author_review_count: The mean is 9 369.8 reviews, with a std of 24 949.8, indicating high dispersion. The range is from 0 to 516 745, meaning some authors receive no reviews, while others receive hundreds of thousands. The distribution is heavily right-skewed, as the median (2 273) is much lower than the mean. Outliers include authors with over 100 000 reviews.
5. book_average_rating: The mean rating is 4, with a std of 0.3, indicating low dispersion. The ratings range from 0.0 to 5.0, though most books fall between 3.8 and 4.1 (interquartile range). The median rating (4.0) is close to the mean, suggesting a slightly right-skewed distribution. Possible outliers include books with extremely low ratings, particularly those close to 0.0.
6. num_ratings: The mean is 46 683.5, but the std is 180 069.8, indicating a huge variation in book popularity. The range is from 0 to 3.8 million, meaning some books are barely rated, while others have millions of ratings. The distribution is heavily right-skewed, as the median (4 403) is much lower than the mean. Outliers include books with millions of ratings.

7. num_reviews: The mean is 2 324.8 reviews, with a std of 6 837.5, indicating high dispersion. The range is from 0 to 147 696. The distribution is right-skewed, as the median (384) is much lower than the mean. Outliers include books with over 70 000 reviews.
8. score: The mean is 3 893, but the std is 11 022, suggesting large differences in book rankings. The range is from 55 to 598 270, meaning some books are barely ranked, while others are extremely popular. The distribution is right-skewed, as the median (1 727) is much lower than the mean. Outliers include books with scores above 250 000.
9. publish_date: The mean is 1990.9, with a std of 130, indicating a wide range of publication years. The range is from 720 BC to 2019 AD, meaning some ancient books are included. The distribution is bimodal, with one peak in recent years (2000s) and another peak in classical literature (pre-1900s). Outliers include books before 0 AD. Major subgroups include modern books (2000-2019) which dominate the dataset, and classic books (pre-1900) which form a smaller but significant group.

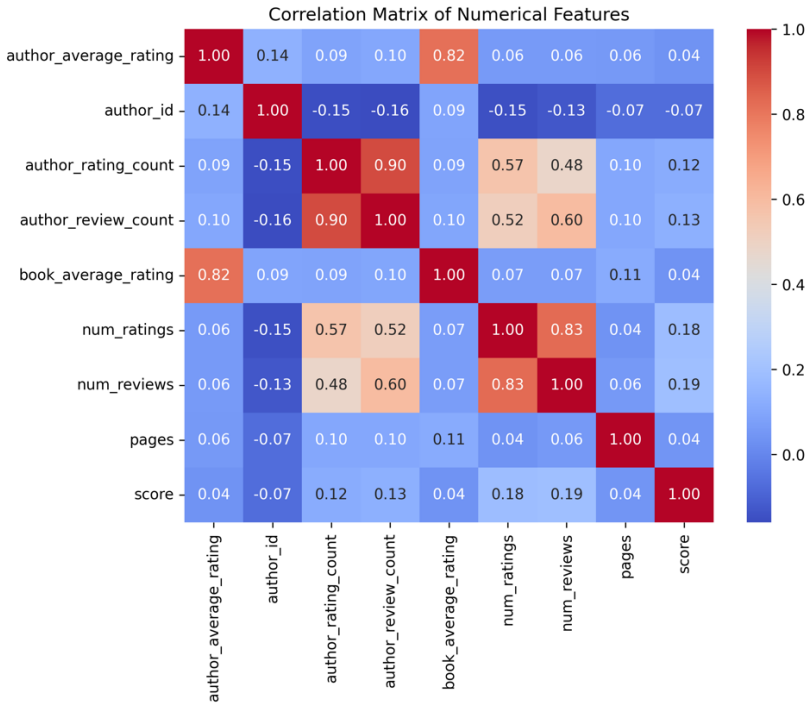
   There are missing values that were later removed when creating the visualizations. This reduced the dataset from 22 891 rows to 22 294 rows.
10. pages: The mean is 333 pages, with a std of 219 pages, indicating moderate variability in book lengths. The range is from 0 to 6 680 pages, meaning some entries have missing values(0) while others are exceptionally long. The distribution is right-skewed, as the median (316 pages) is slightly lower than the mean. Outliers include books with over 1 000 pages, especially the maximum of 6 680 pages, which is extremely rare. Major subgroups include: Short books (under 250 pages) which are likely novellas or shorter works, standard-length books (250-400 pages) which form the majority, and long books (over 500 pages), including epic novels and academic texts.

**Final Insights:**

The numerical columns show a wide variation in values, with some indicating a high degree of dispersion. Author ratings are generally high, with most ratings above 3.8, and book ratings follow a similar trend. Author and book popularity varies greatly, as seen in the large standard deviations for ratings, reviews, and scores, suggesting that while some authors and books are extremely popular, many others have lower engagement. The dataset includes a mix of books spanning different eras, with a notable concentration in the 2000s and 2010s. Outliers appear in some columns, particularly in the number of ratings, reviews, and pages, reflecting a few exceptionally popular books or authors.

## Relationships Between Columns

### Correlation:




The correlation analysis reveals strong relationships among key numerical features in the dataset. A high correlation (0.9) exists between an author's total rating count and their review count, as well as between the number of ratings and reviews for books (0.8), suggesting that more popular books and authors attract both ratings and reviews. Additionally, an author's average rating is strongly correlated (0.8) with their books' average ratings, indicating that well-regarded authors consistently produce highly rated books. In contrast, book length shows little to no correlation with ratings, reviews, or overall score, suggesting that it does not significantly influence reception.

### Regression:

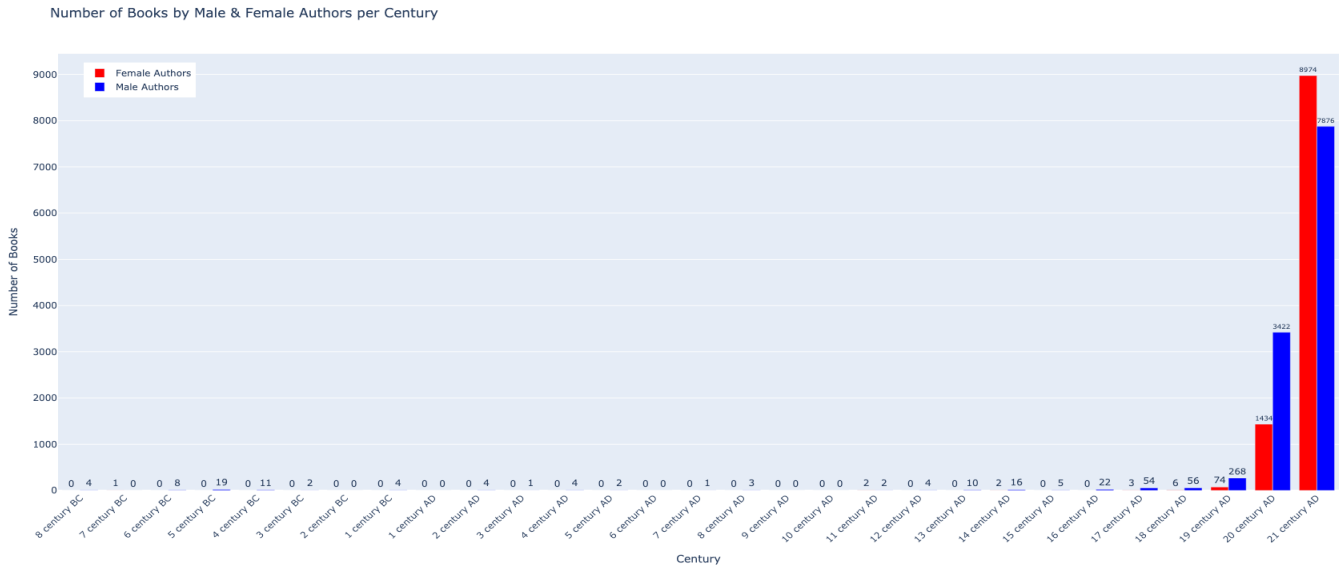
A linear regression analysis (Ordinary Least Squares regression model) further examines the relationship between author_average_rating and book_average_rating, explaining 66.5% of the variance in book ratings. The regression coefficient (0.99) indicates a near one-to-one relationship. The relationship is highly statistically significant (p-value = 0.0), confirming that author ratings serve as a strong predictor of book ratings.

The resulting linear fit equation is: book_average_rating = 0.048 + 0.99 × author_average_rating.

**Conditional Probability Analysis:**

To further examine relationships in the dataset, the conditional probabilities for key variables have been calculated. The probability that a book receives a high rating (≥4.5) given that the author's average rating is also high (≥4.5) is 0.8. This suggests that books from highly rated authors are very likely to receive strong ratings themselves. Additionally, the relationship between book ratings and engagement was examined. The probability that a book has more than 1 000 reviews, given that its rating is above 4.0, is 0.4. This indicates that while higher ratings may increase review activity, other factors such as book popularity, marketing, and reader interest also contribute significantly to review counts.

## Research Question 1: Is there a gender disparity in the number of books published by male and female authors?



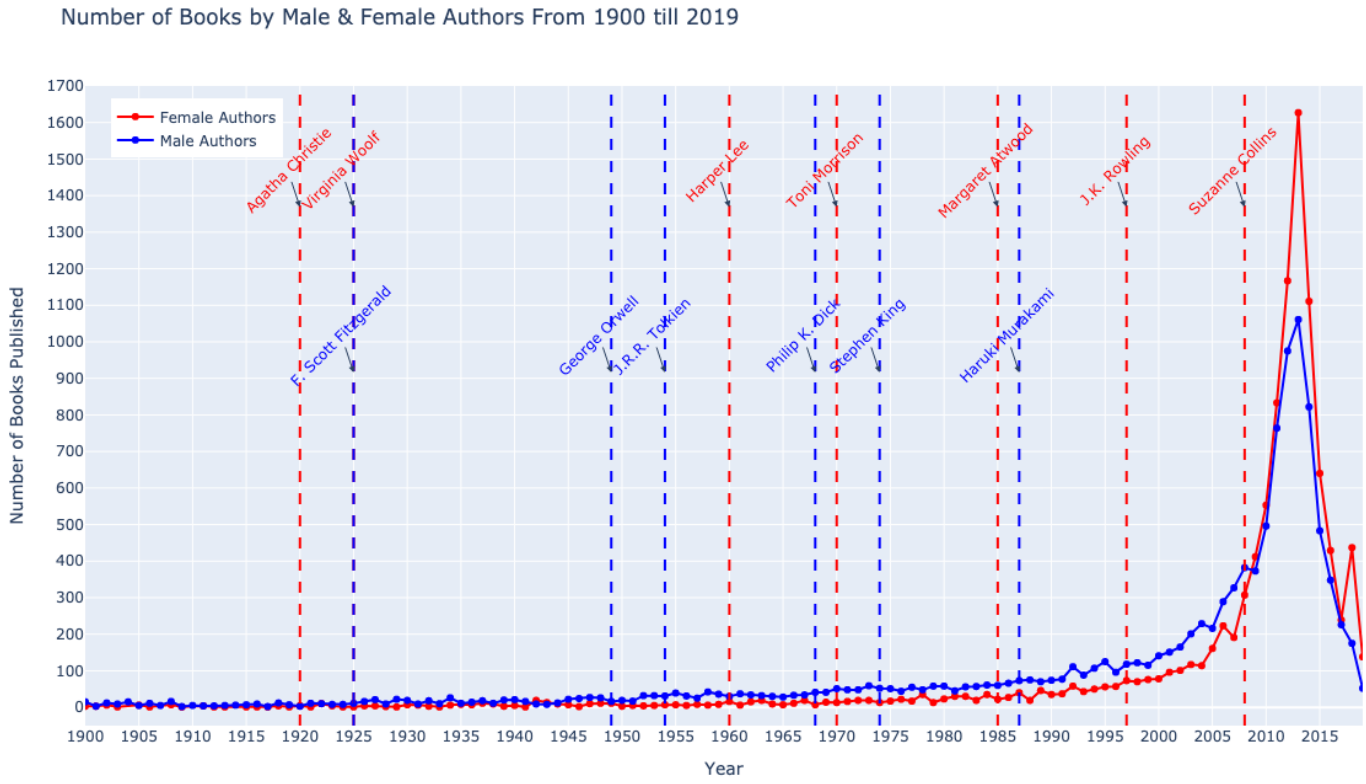

To answer this research question, a double bar graph has been constructed to illustrate the number of books by male and female per century beginning from 720 BC to 2019 AD. Based on this, there is a gender disparity between the number of books published by male and female authors. The chart shows that male authors consistently published more books than female authors across all centuries, until the 21st century.

The disparity is especially pronounced in earlier centuries, where female authors contributed very few books. In

the 20th century, the number of books by female authors had increased significantly but still remained lower than the number of books by male authors.

However, based on the graph, the gender disparity reverses in the 21st century. Female authors have published more books than male authors in the 21st century, with 8974 books by female authors vs. 7876 by male authors.

The trend shows that while male authors consistently dominated book publishing in earlier centuries, female authors have significantly closed the gap in recent centuries, and even surpassed male authors in the 21st century. This graph shows a really interesting shift in gender representation in literature over time.




Analysing this line plot allows one to more clearly view the gender disparity. The plot represents the number of books published by male and female authors from 1900 to 2019. This graph also highlights notable authors of each gender, such as Agatha Christie, Virginia Woolf, Harper Lee, and Toni Morrison (female), alongside George Orwell, J.R.R. Tolkien, Stephen King, and Haruki Murakami (male). The lines show when those authors were active or published notable works.

In the early 20th century, male authors consistently published more books than female authors from 1900 to about 1980, and the gap was quite wide during this time.

Despite that, the gap starts closing gradually as female authors such as Margaret Atwood and Toni Morrison start making a bigger mark. The lines for female authors and male authors, beginning from 2009 to 2011,

converged after which female authors overtook the male authors by a huge jump. This is especially prevalent in 2013 when female authors published 1 627 books as compared to male authors' 1 061 books.

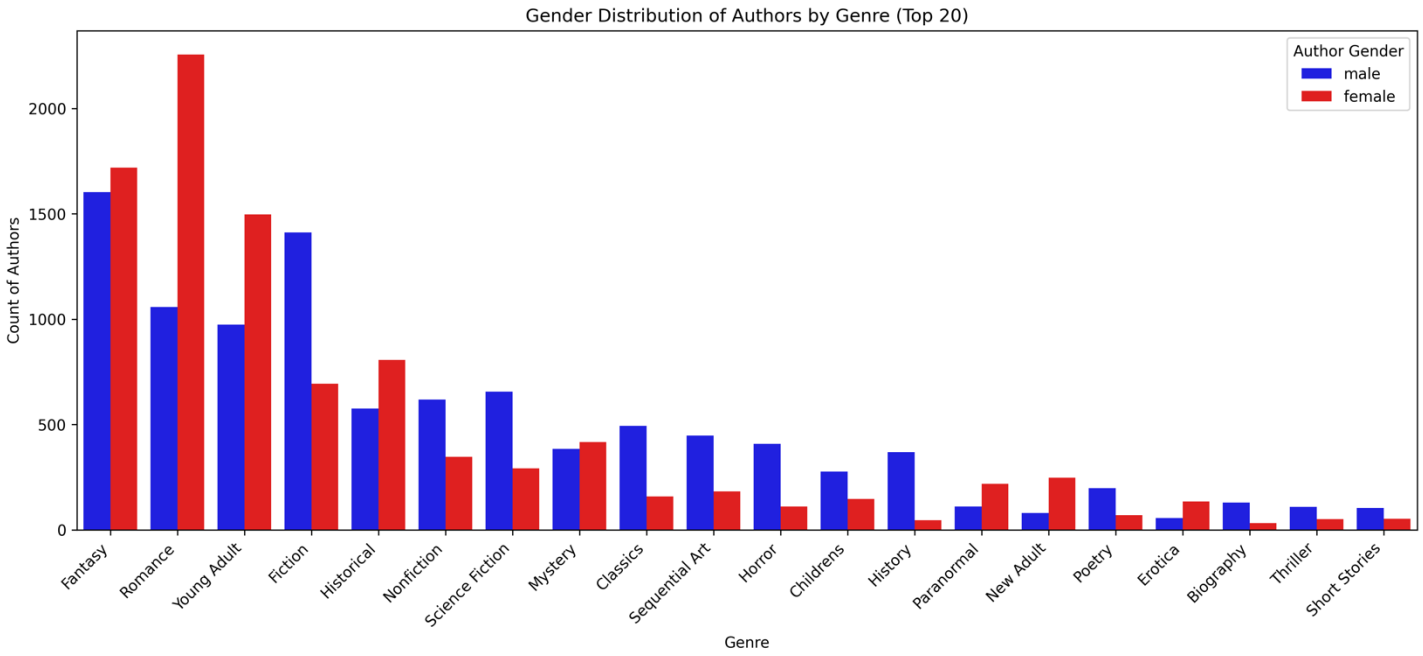
This peak could be attributed to hugely successful authors like J.K. Rowling and Suzanne Collins, as well as the rise of the Young Adult genre. The interesting switch in 2013 will be analysed later on in this report.

The data highlights a significant transformation in the publishing industry. While male authors historically dominated book publishing, the 21st century marks a turning point where female authors have not only closed the gap but surpassed their male counterparts. This shift could be attributed to broader societal changes, including increased gender equality, greater access to education, and evolving perceptions of women's roles in literature.

To further visualize this trend, an animated line graph was created showing the number of books published by male and female authors from 1900 to 2019. Watch the full video here:

https://drive.google.com/file/d/1mjS_75QK7TKku5e-nHke12IRJ69w3M0e/view?usp=sharing

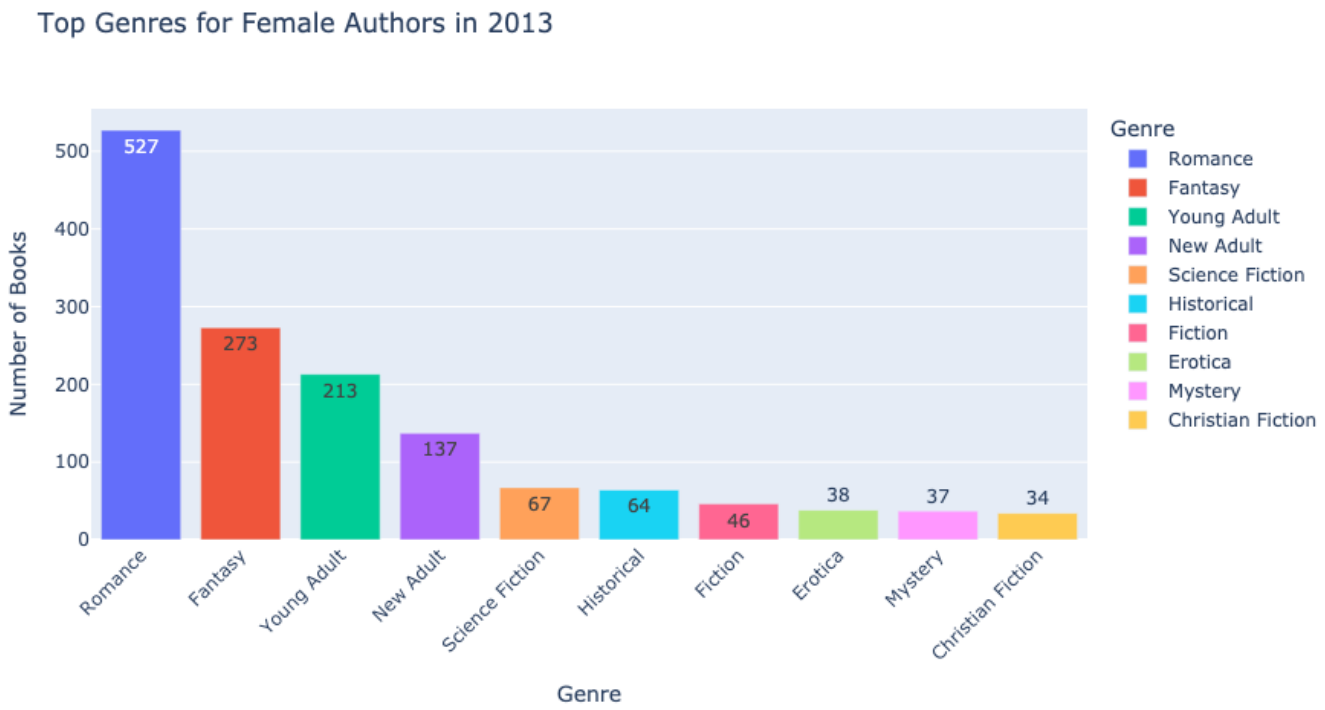
## Research Question 2: Do male and female authors prefer different genres, and if so, which genres?




This double bar graph provides valuable insights into gender representation across different literary genres and showcases clear gender dominance in certain genres. Female-dominated genres include: Romance and Young Adult, which is a reflection of broader industry trends where these genres cater to audiences that are women, as can be seen from a 2017 report from the Romance Writers of America which found that 82% of romance readers are women (Romance Writers of America, 2017), as well as New Adult and Paranormal, which could be attribute to the 2010s boom of Young Adult Paranormal books due to series like Twilight by Stephanie Meyer (Crawford, 2014).

In contrast, male authors dominate Science Fiction, Fiction, and Classics, which could indicate a lingering effect of past publishing trends where these fields were traditionally associated with male writers.

Fantasy, while balanced, leans slightly towards male authors. Nevertheless, the landscape is evolving. The article, Women in Literature: The Impact of Feminism on Fantasy Literature, 1950–1990, explores the feminist movements of the 1960s and 1970s and how they significantly impacted fantasy literature, leading to an increase in strong female characters and more female authors entering the field. Authors like Ursula K. Le Guin and Marion Zimmer Bradley played pivotal roles in this transformation (Dassler, 2021).



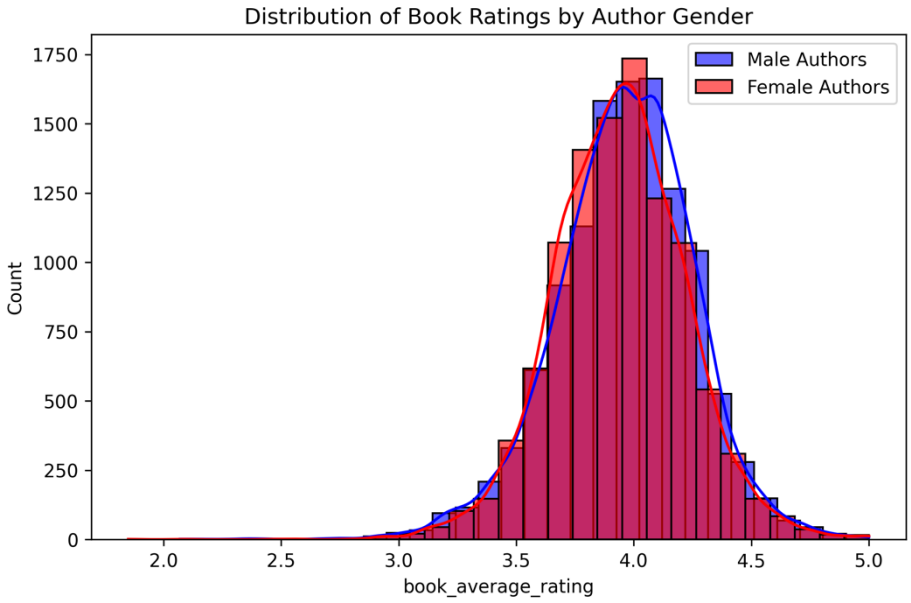

The year 2013 has been specifically analysed to understand the interesting flip that occurred during this year as women authors surpassed male authors by a lot. The most published genre by female authors was Romance (527 books), almost double the next genre. Fantasy (273 books) and Young Adult (213 books) were the second and third most popular genres. This shows that female authors dominated genres that were booming in popularity at the time, particularly in commercial fiction markets like Romance, Young Adult, and Fantasy, as is evident in an article by Meghan Lewit from The Atlantic (Lewit, 2012).

Romance has long been a female-dominated genre, but the self-publishing boom (via platforms like Amazon Kindle Direct Publishing) amplified the voices of female authors, making it easier for them to publish independently (Flood, 2015).

Genres like Romance, Young Adult, and New Adult flourished in the self-publishing scene, where women were the primary creators and primary audience. A 2013 study on young adult female readership determined that readers frequently compared their own lives to those of the characters, indicating a demand for relatable, diverse narratives. (Suico, 2013)

The popularity of series like The Hunger Games, which features a strong female protagonist, reflects this shift in reader preferences toward more diverse and female-centric stories (Elsdon, 2020).

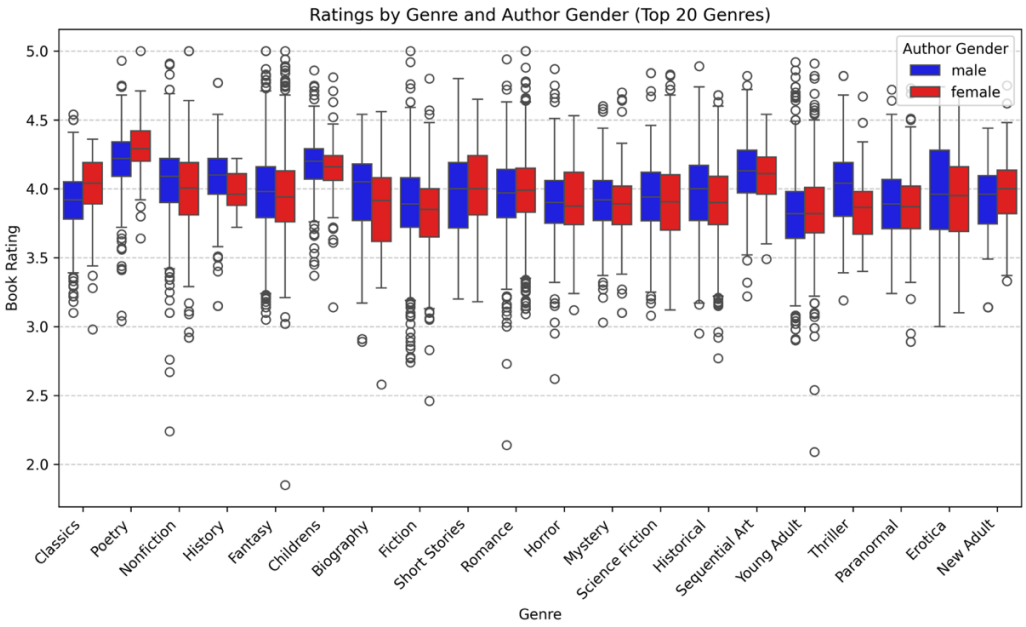
## Research Question 3: Do gender, genre, and publication date influence the popularity of books?




The histogram displaying the distribution of book ratings by author gender shows that both male and female authors receive similar ratings, with the majority of ratings clustering around 4.0. The curves indicate that both distributions follow a normal bell-shaped pattern, meaning that most books are rated within a consistent range.

Moreover, female authors appear to have a slightly higher frequency of ratings above 4.0, while male authors show more variation and a higher frequency of ratings between 3.5 and 4.0.

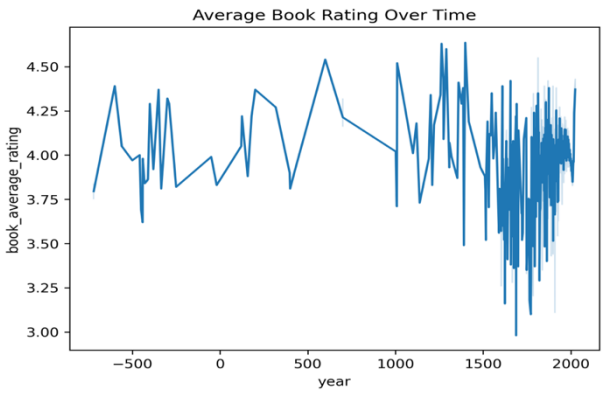
This suggests that books by female authors may be less polarizing in reader reception. Despite these minor differences, the overall similarity between the distributions highlights that readers rate books by male and female authors in largely comparable ways, implying that gender does not play a significant role in how books are evaluated.




The box plot compares book ratings across 20 genres based on author gender. Overall, median ratings range mostly between 3.5 and 4.5. Both male and female authors receive similar ratings, though in some genres, like Poetry and Classics, female authors have slightly higher medians, while in others, like Thriller and Science Fiction, male authors do. However, these differences are minor and do not suggest a consistent gender-based advantage.

Genres like Classics and Children's tend to have higher median ratings, while Horror and Erotica show more variability and lower medians. Science Fiction and Fantasy exhibit wider distributions, indicating diverse reader opinions. Outliers are common in several genres, particularly Fiction and Young Adult. While some genres consistently receive higher ratings, overall differences are moderate, suggesting that genre influences ratings but not in an extreme or uniform way.

Nevertheless, a one-way ANOVA test conducted across genres (regardless of gender) yielded an F-statistic of 21.0 with a p-value of 0.0, indicating that differences in average book ratings across genres are statistically significant. This suggests that while gender may not strongly influence ratings within a genre, the genre itself does slightly affect book ratings, with some genres consistently receiving higher or lower ratings on average.



To explore whether book ratings have changed over time, the average book rating by publication year was plotted, with the negative values indicating BC. The graph shows a time-dependent trend in book ratings, with older books generally receiving higher and more stable ratings, likely due to survivorship bias and historical significance.

In contrast, modern books exhibit greater variability, possibly reflecting the increasing volume of published works, more diverse reader opinions, and the rise of online reviews. These factors have contributed to harsher criticism and wider rating fluctuations in recent years (Chevalier & Mayzlin, 2006). This suggests that as time progresses, books are judged more critically, making ratings more dynamic in the modern era. Therefore, this implies that the publication date influences the popularity and ratings of the book.

## Discussions And Conclusions

The analysis of gender disparity in book publishing reveals a significant shift over time. Historically, male authors consistently published more books than female authors, with the disparity being most pronounced in earlier centuries. This trend aligns with broader historical constraints on women's participation in the public sphere, especially in academic and creative fields like literature.

However, the 21st century marks a notable turning point where female authors not only closed the gap but surpassed their male counterparts in terms of books published. This shift can be attributed to several factors, including increased access to education for women, growing gender equality, and the rise of self-publishing platforms.

Comparing the trends in publishing genres, female authors have long dominated genres such as Romance and Young Adult, which cater primarily to female audiences. The rise of successful series like Twilight and The Hunger Games further emphasizes the popularity of female-authored books in these genres. In contrast, male authors dominate traditionally male-associated genres like Science Fiction and Fiction, although female authors are beginning to make significant inroads, especially in Fantasy.

The analysis of book ratings also suggests that gender does not play a significant role in how books are perceived by readers. Both male and female authors received similar ratings, with no substantial bias toward one gender. In contrast, genre and publication date appear to have a measurable impact on book ratings.

While this study sheds light on important trends, it has limitations. The dataset, though extensive, may not fully represent the global literary landscape, and the focus on Goodreads data introduces potential biases. Future research could explore additional datasets, incorporate intersectional factors like race and ethnicity, and

investigate the long-term impact of self-publishing on the industry. Given that female authors have gained significant traction through platforms like Amazon Kindle Direct Publishing, it would be interesting to investigate the specific impact of these platforms on the publishing landscape and how they might further reshape gender dynamics in the future.

In conclusion, the findings reflect broader societal shifts toward greater gender inclusivity in literature. As the publishing industry continues to evolve, these trends emphasize the importance of fostering diversity and supporting authors from all backgrounds.

## References

Romance Writers of America. (2017). *The romance book buyer 2017: A study by NPD Book for Romance Writers of America*. https://www.rwa.org/the-romance-genre

Crawford, J. (2014). *The Twilight of the Gothic?: Vampire fiction and the rise of the paranormal romance, 1991-2012.* University of Wales Press.

Dassler, J. (2021). Women in literature: The impact of feminism on fantasy literature, 1950–1990. *International Social Science Review, 97*(4). https://issr.ungjournals.org/articles/131

Elsdon, L. (2020). The rise of strong female characters in YA Fantasy. *YA Hotline, (112).* https://ojs.library.dal.ca/YAHS/article/view/10294

Flood, A. (2015). Self-publishing lets women break book industry's glass ceiling, survey finds. *The Guardian*. https://www.theguardian.com/books/2015/mar/06/self-publishing-lets-women-break-book-industrys-glass-ceiling-survey-finds

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345

Lewit, M. (2012). Why do female authors dominate young-adult fiction? *The Atlantic*. https://www.theatlantic.com/entertainment/archive/2012/08/why-do-female-authors-dominate-young-adult-fiction/260829/

Rosen, B. (2019). *Goodreads books/author data* [Dataset]. Kaggle. https://www.kaggle.com/datasets/brosen255/goodreads-books

Suico, T. G. (2013). *Privileged high school girls' responses to depictions of femininity in popular young adult literature*. *Dissertation Abstracts International, 75*(2). https://hdl.handle.net/2144/11058