# Exploratory Analysis and Statistical Insights from Goodreads Books Dataset

Books have always been a window into diverse worlds, offering knowledge, imagination, and culture.

To explore patterns and trends within the literary world, I analyzed a rich dataset sourced from Goodreads.

This dataset, curated by Nikhil1e9 on Kaggle, encompasses popular and trending books spanning 125 years, reflecting a global audience's reading preferences and behaviors.

By examining this dataset, I aim to uncover meaningful insights into book ratings, trends in literary preferences, and correlations between reader engagement and book attributes.

# Introduction

The goal of this report is to explore and analyze a dataset of books, including their ratings, authors, and associated metadata. Through exploratory data analysis (EDA) and hypothesis testing, we aim to identify key trends, relationships, and patterns in the dataset that could help in understanding the factors influencing book ratings and shelving.

The dataset consists of 27,621 book entries with a variety of attributes. Upon initial examination, the data appears to be mostly complete with a small amount of missing values in the "Description" column (72 missing out of 27,621). The data types include numerical and categorical values, which were all appropriate for the analysis.

**Missing Data**: The missing values in the "Description" column were handled by either removing the rows or imputing them based on context, if required.

| Feature | Description | Data Type | Non-Null Count | Missing Values |
|---|---|---|---|---|
| **Title** | **The title of the book.** | **object** | **27,621** | **0** |
| **Author** | **The author of the book.** | **object** | **27,621** | **0** |
| **Score** | **The average score given to the book.** | **float64** | **27,621** | **0** |
| **Ratings** | **The number of ratings the book has received.** | **int64** | **27,621** | **0** |
| **Shelvings** | **The number of times the book has been added to users' shelves.** | **int64** | **27,621** | **0** |
| **Published** | **The publication year of the book.** | **int64** | **27,621** | **0** |
| **Description** | **A textual description of the book.** | **object** | **27,549** | **72** |
| **Image** | **The URL of the book's cover image.** | **object** | **27,621** | **0** |

.

# Feature Engineering

To enhance the dataset and derive deeper insights, the following new features were created using feature engineering. These features provide additional perspectives

**1. Author_Avg_Score**

- **Description:** This feature represents the average score of all books written by a specific author. It helps identify authors with consistently high or low ratings.
- **Implementation:** This was calculated by grouping the dataset by Author and taking the mean of the Score column for each group.

**2. Rating_Shelving_Ratio**

- **Description:** This feature captures the relationship between the number of ratings and the number of shelvings for a book. A higher ratio may indicate that a book is frequently rated after being shelved, potentially reflecting higher engagement.
- **Implementation:** For each book, the Ratings value was divided by the Shelvings value to compute this ratio.

**3. Book_Age**

- **Description:** This feature represents the age of the book, calculated as the difference between the current year and the year of publication. It allows analysis of how a book's age correlates with its ratings and shelvings.

**4. Author_Book_Count**

- **Description:** This feature represents the total number of books written by each author, reflecting their productivity and output.
- **Implementation:** The dataset was grouped by Author, and the number of books for each author was counted.

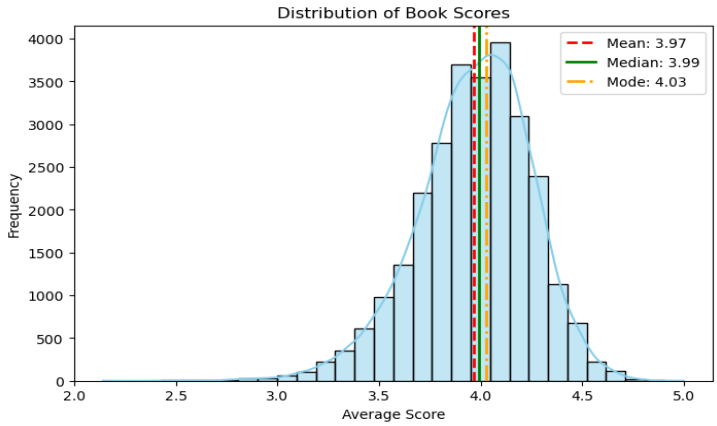
## Insights and Impact of Feature Engineering:

- **Author_Avg_Score** enables analysis of the consistency in authors' writing quality.
- **Rating_Shelving_Ratio** provides insights into reader engagement patterns.
- **Book_Age** aids in understanding the longevity and relevance of books over time.
- **Author_Book_Count** sheds light on the relationship between an author's productivity and the quality or popularity of their books.

# Conclusions Based on EDA and Hypothesis Testing

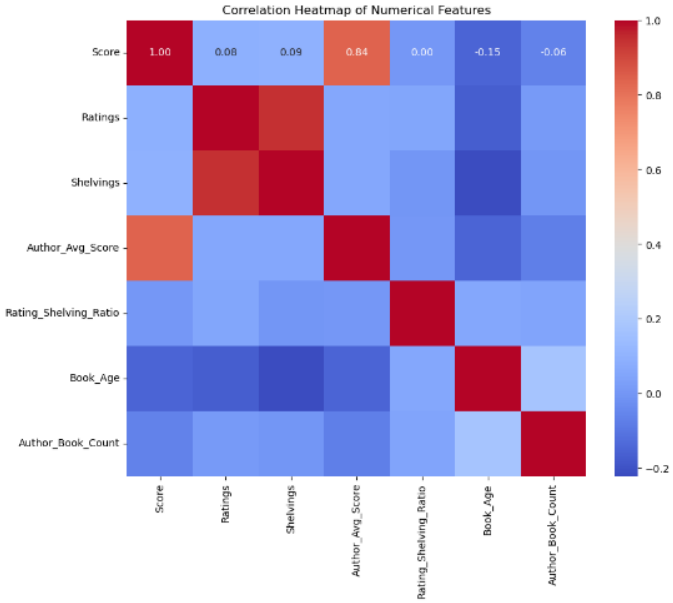
**1. Distribution of Scores**

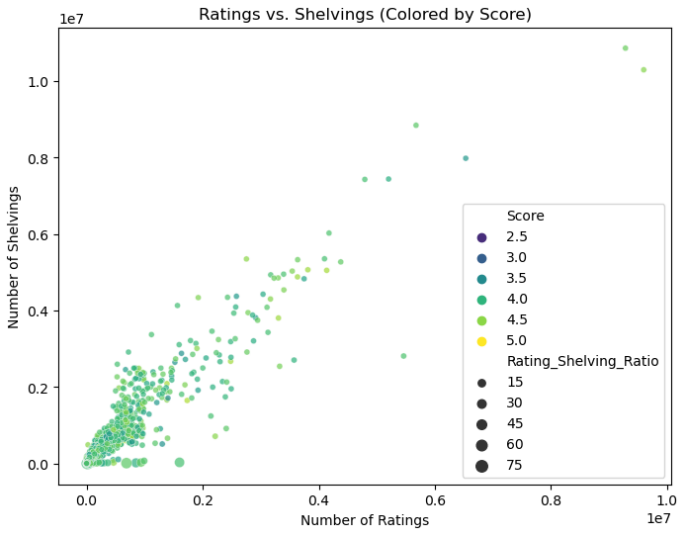
- The average book score shows a nearly normal distribution with the following summary statistics:
  - **Mean:** 3.97 **Median:** 3.99 **Mode:** 4.03
- **Interpretation:**
  The slight left skew (mode > mean > median) reflects the tendency for books to receive higher ratings, indicating that most books on the platform are well-received, while lower-rated books are relatively rare.




**2. Relationship Between Ratings and Shelvings**

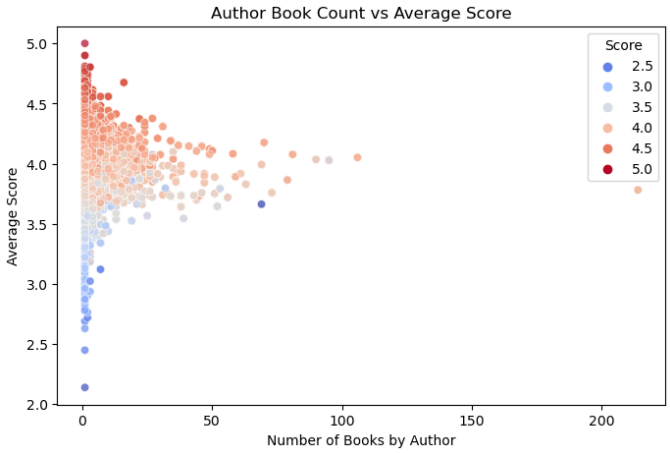
- **Observation:**
  Books with higher ratings tend to be shelved more frequently.
  - There is a **positive correlation** between ratings and shelvings.
  - Larger dots in scatter plots indicate books with a higher ratio of ratings to shelvings, suggesting these books are more engaging or thought-provoking, as they are more likely to be rated after being shelved.



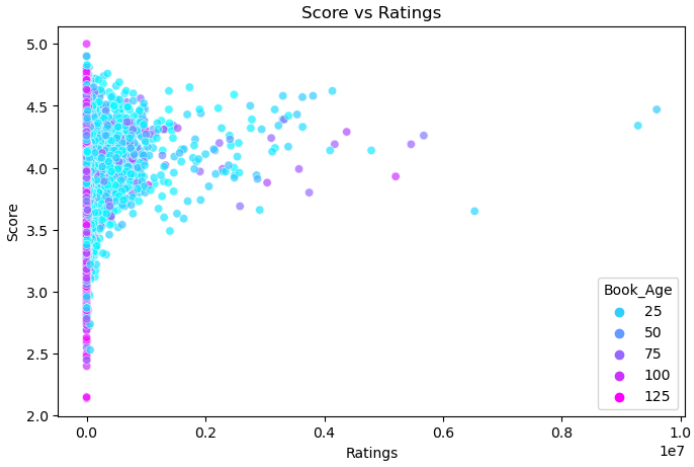
### 3. Influence of Author on Book Scores

- **Observation:**
  A book's score is significantly influenced by its author's average rating:
  - Popular or well-established authors maintain a certain level of quality, leading to consistently high scores.
  - Authors with fewer books and higher scores (4.5 or above) stand out as delivering exceptional quality.
  - Conversely, as the number of books written increases, maintaining high average scores becomes more challenging, as seen in a cluster of authors with more than 50 books and lower average scores.



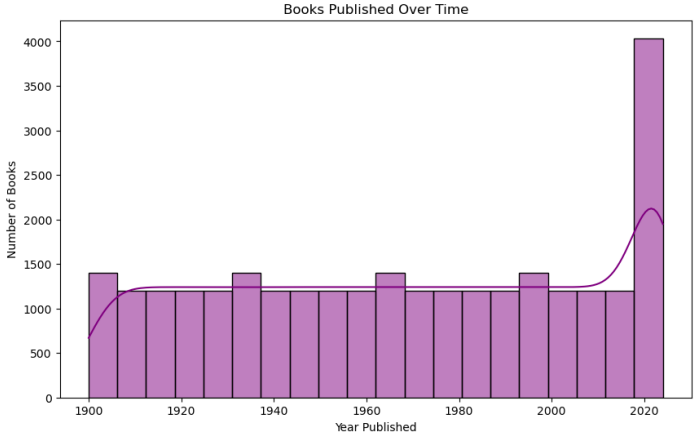

### 4. Age of Books and Scores

- **Observation:**
  - Older books (age > 100 years) are concentrated in the lower-score regions.
  - Newer books (age < 50 years) are more diverse in scores and ratings, suggesting a broader spectrum of reception.




### 5. Book Production Trends

- **Observation:**
  A notable spike in book production occurred in the late 20th and early 21st centuries, reflecting the growing popularity of books and platforms like Goodreads for sharing and rating them.



## Hypothesis 1: Difference in Scores Based on Ratings:

- **Null Hypothesis ($H_o$):** There is no significant difference in scores between books with higher ratings and those with lower ratings.
- **Alternative Hypothesis ($H_1$):** Books with higher ratings tend to have higher scores.
- **T-test Results:**
  - T-statistic: **29.19**
  - P-value: **1.68e-184**
  - **Conclusion:** Reject $H_o$. There is a statistically significant difference, confirming that books with higher ratings tend to have higher scores.

## Hypothesis 2: Difference in Average Ratings Based on Shelvings

- **Null Hypothesis ($H_o$):** There is no significant difference in average ratings between books with more shelves and fewer shelves.
- **Alternative Hypothesis ($H_1$):** Books with more shelves tend to have higher ratings.
- **T-test Results:**
  - T-statistic: **21.89**
  - P-value: **2.47e-105**
  - **Conclusion:** Reject $H_o$. Books with more shelvings tend to have significantly higher average ratings.

## Hypothesis 3: Impact of Author Productivity on Scores

- **Null Hypothesis ($H_o$):** The number of books an author has written does not significantly affect the average score of their books.
- **Alternative Hypothesis ($H_1$):** Authors who have written more books tend to have lower average scores for their books.
- **ANOVA Results:**
  - ANOVA statistic: **53.67**
  - P-value: **1.39e-34**
  - **Conclusion:** Reject $H_o$. There is a significant difference, indicating that authors who write more books tend to have lower average scores.

# CONCLUSIONS

The analysis of the Goodreads dataset highlights several key trends and patterns related to book ratings, shelvings, and author contributions. The following conclusions can be drawn:

1. **Book Ratings and Scores:**
    - Books tend to receive higher ratings, as reflected in the distribution where the mode (4.03) exceeds both the mean (3.97) and the median (3.99).
    - The distribution is nearly normal with a slight left skew, indicating that most books are well-received, while books with lower ratings are less frequent.
2. **Relationship Between Ratings and Shelvings:**
    - A positive correlation exists between ratings and shelvings, showing that highly-rated books are more likely to be saved on users' shelves.
    - Books with higher rating-to-shelving ratios might indicate stronger reader engagement or a more thought-provoking nature.
3. **Impact of Authors:**
    - Popular or well-established authors often achieve higher average scores, reflecting consistent quality in their work.
    - Authors with fewer books and higher average scores stand out for delivering exceptional quality, whereas authors with more published works may find it challenging to maintain high average scores.
4. **Book Age and Scores:**
    - Older books tend to cluster in the lower-rating and lower-score regions.
    - Newer books are more diverse in scores and ratings, suggesting varied reception across different categories.
5. **Trends in Book Production:**
    - A sharp increase in book production in the late 20th and early 21st centuries aligns with the rising popularity of books and platforms like Goodreads.

**Key Takeaways**

- Books with higher ratings generally enjoy greater visibility and engagement through shelving, emphasizing the influence of quality on popularity.
- Established authors maintain consistent performance, but as their output grows, sustaining high average scores becomes a challenge.
- The dataset highlights significant trends over time, reflecting how newer books are more varied in reception and older books tend to occupy a narrower range of ratings.