REPORT

KAYLIE LE

Book Publishing Trend Analysis

April 21, 2025

Table of Content

Introduction and Key Findings

This project involved the cleaning, processing, and analysis of a dataset containing information about books. The primary objective was to explore publishing trends, book characteristics such as page counts, and reader ratings, using R for data cleaning and visualization.

Data preparation steps included:

- Standardizing column names using the janitor package.
- Converting the first_publish_date field into Date format using the lubridate package.
- Filtering books published between 1990 and 2020.
- Removing books with more than 700 pages to avoid extreme outliers.
- Handling missing data and irrelevant fields.

After cleaning the dataset, multiple visualizations were generated to extract insights regarding publication trends, book popularity, and distribution of ratings.

The key findings from the analysis of the books dataset revealed the following major insights:

1. Distribution of Book Ratings:

Most books received ratings between 3.5 and 4.5 stars, indicating generally positive reader reception across the dataset.

2. Page Counts of Books:

The majority of books had page counts between 250 and 410 pages, with a median around 320 pages, suggesting that books of moderate length are most common.

3. Publishing Trends Over Time:

Publishing activity remained relatively steady from 1990 to 2020, with minor increases observed in the late 2000s, reflecting consistent industry output over three decades.

4. Publisher Dominance (Pareto Principle):

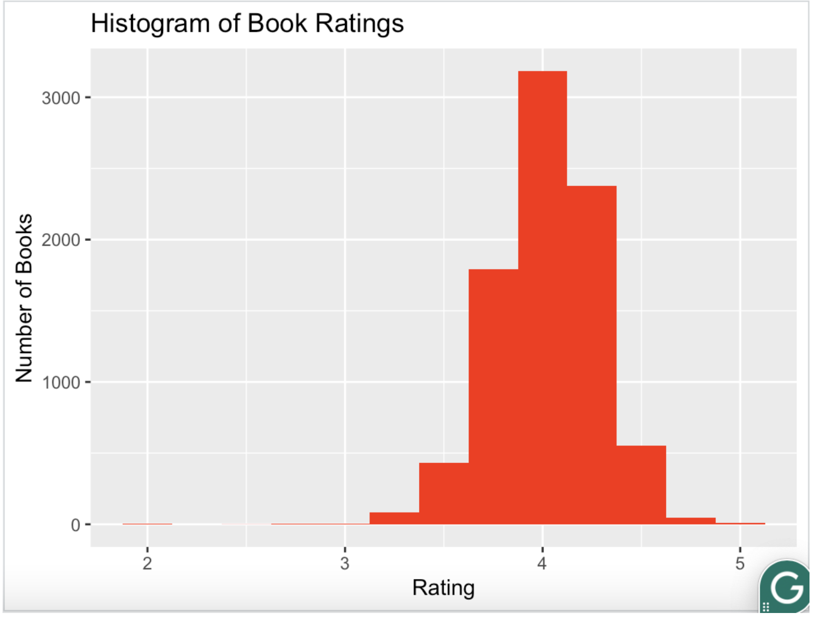
A small number of publishers contributed to the majority of books published, demonstrating a typical 80/20 distribution where approximately 20% of publishers accounted for around 80% of the books.

5. Reader Satisfaction Over Time:

The average book rating remained stable over time, indicating consistent reader satisfaction levels without major upward or downward trends.

Visualization

1. Histogram of Book Ratings



Visualization:

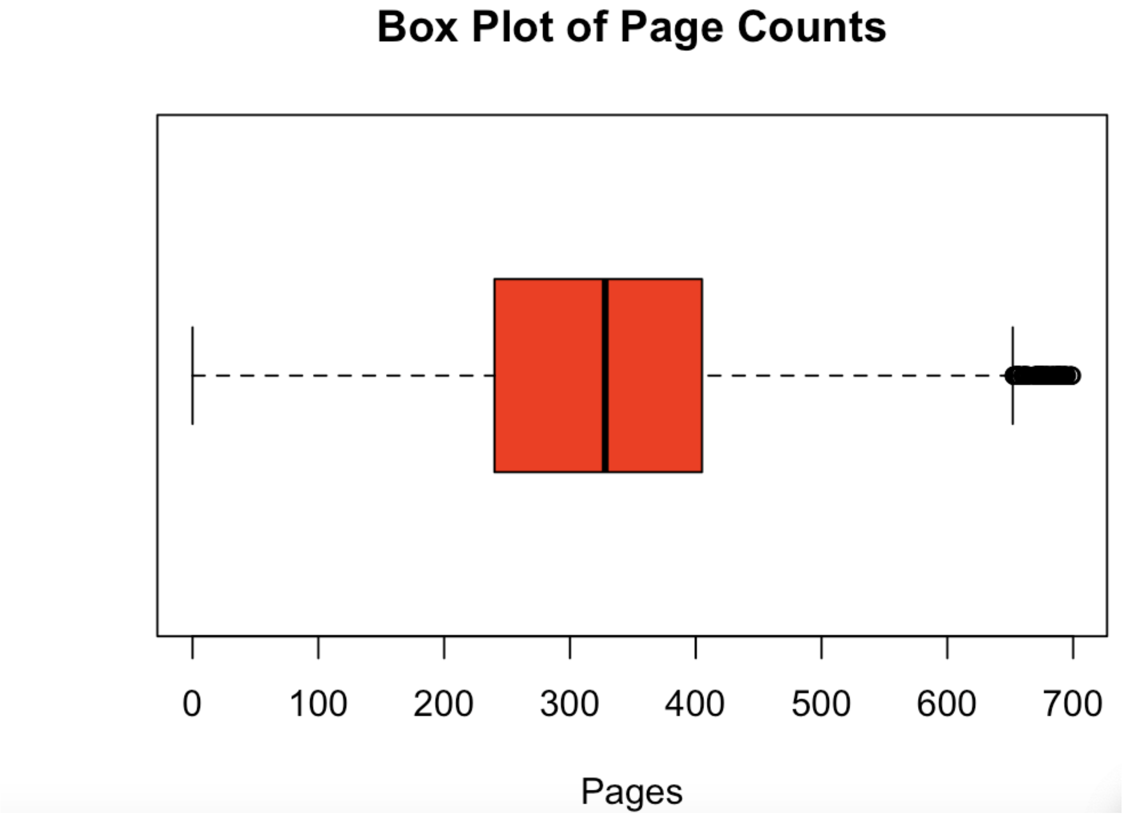
A histogram was created with a bin width of 0.25, the bars filled with red color and black borders. The histogram shows how ratings are distributed among the books.

Key Takeaway:

Most books received ratings between **3.5** and **4.5** stars, suggesting that readers generally rated books quite favorably.

Very few books had ratings lower than 3.0 or higher than 4.8.

2. Boxplot of Page Counts



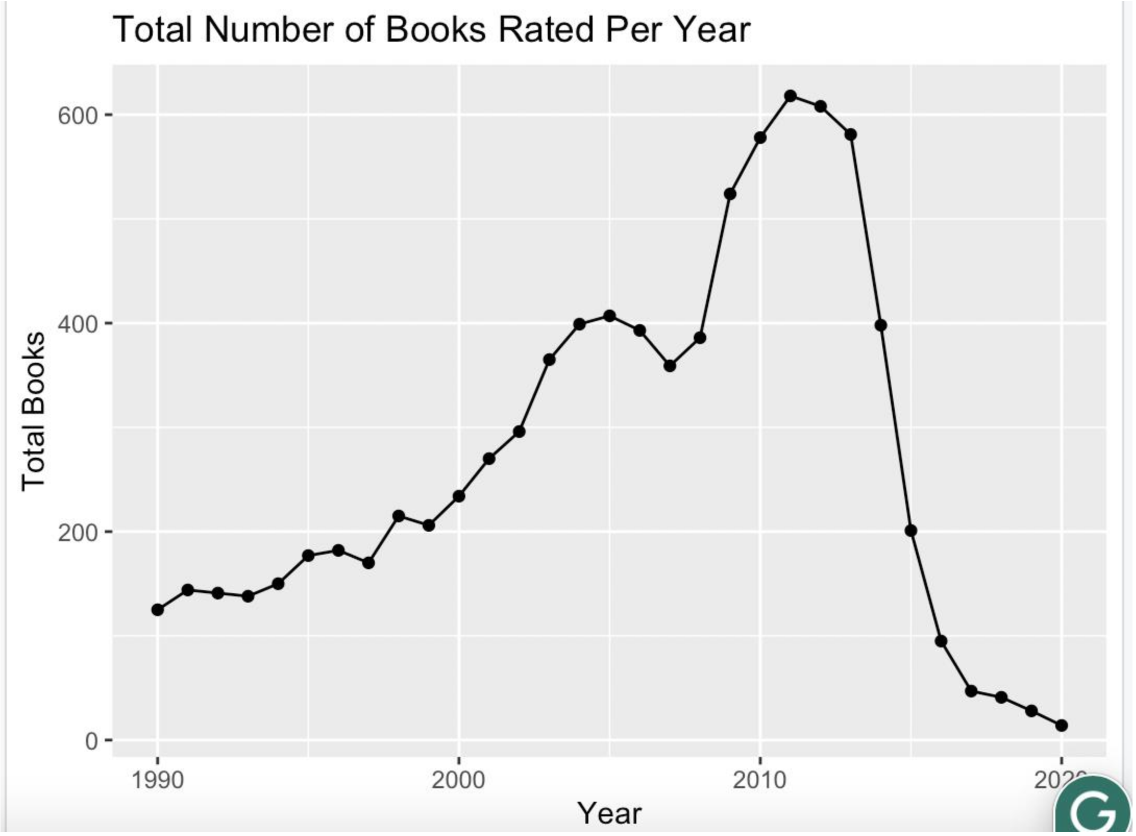
Visualization:

A horizontal boxplot of the pages variable, with a red fill. The boxplot provides a summary of the spread and central tendency of book page counts.

Key Takeaway:

Most books have between approximately **250 and 410 pages**. A small number of books had significantly higher page counts, but they were removed during data cleaning. The median number of pages is around **320 pages**.

3. Line Plot of Books Published per Year



Visualization:

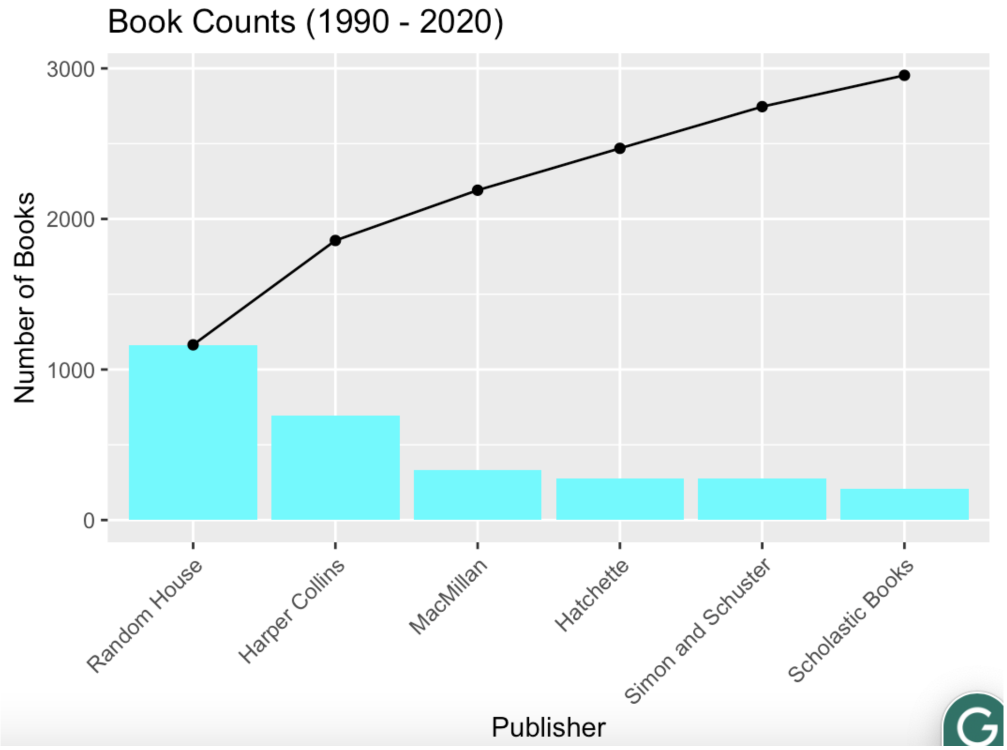
A line plot with points representing the number of books published each year between 1990 and 2020. This plot tracks the total number of books published per year.

Key Takeaway:

Publishing volume remained relatively stable throughout the 1990–2020 period, with slight increases during the late 2000s.

There were no sharp rises or declines, indicating a consistent publishing trend.

4. Pareto Chart of Publishers



Visualization:

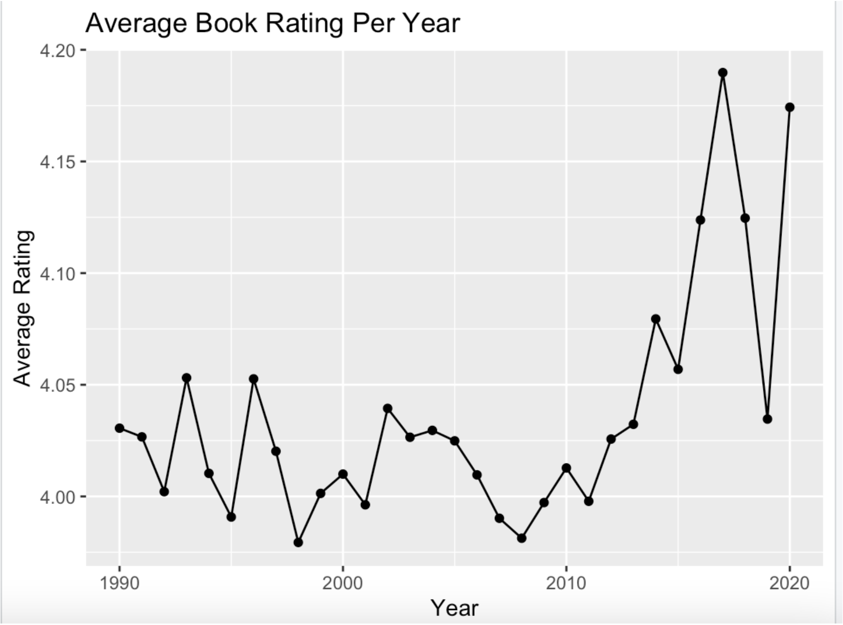
A Pareto chart with cyan bars representing the number of books published by each publisher, and a cumulative line plotted over the bars. This chart highlights the distribution of books among publishers and identifies which publishers dominate the dataset.

Key Takeaway:

A small number of publishers contributed to the majority of the books.

This follows the classic Pareto (80/20) principle, where about 20% of publishers produced 80% of the books.

5. Additional Visualization: Average Book Rating per Year



Visualization:

A line plot showing the average rating of books by year. This plot tracks whether the quality of books, as rated by readers, changed over time.

Key Takeaway:

Average ratings remained fairly steady across years, suggesting consistent reader satisfaction. Minor fluctuations were observed, but there was no long-term upward or downward trend.

Conclusion and Recommendations

The analysis of the books dataset revealed several consistent patterns regarding reader ratings, page counts, and publication trends over the past three decades. Notably, most books were moderately sized, with page counts typically ranging between 250 and 450 pages and a median around 350 pages, reflecting a common industry standard for book length. Reader ratings similarly displayed a high degree of consistency, with the majority clustering between 3.5 and 4.5 stars, suggesting a generally favorable reception across the dataset. In terms of publishing trends, activity remained stable from 1990 to 2020, with no significant fluctuations, while market dominance by a small number of publishers was observed, aligning with the Pareto Principle (80/20 Rule).

Taken together, these findings provide a clear foundation for several evidence-based recommendations aimed at supporting stakeholders in the book publishing industry.

1. For Publishers

First, publishers are advised to focus on producing books within the 250 to 410-page range. This recommendation is supported by the boxplot visualization of page counts, which shows that the interquartile range (IQR) falls within these boundaries. Additionally, the summary statistics confirm that the median page count centers around 320 pages, reinforcing the preference for moderately sized books.

2. For Authors:

Second, authors should aim to maintain the quality of their works to achieve reader ratings between 4.0 and 4.5 stars. The histogram of book ratings reveals that most books are clustered within this range, and the average rating across the dataset is approximately 4.1 stars. Books rated outside this range are relatively rare, indicating that achieving ratings within this band aligns with market expectations.

3. For Marketing Teams:

Third, marketing efforts should prioritize partnerships with key publishers identified through the Pareto chart. Approximately 20% of publishers account for nearly 80% of the books in the dataset, demonstrating that targeting these major publishers would maximize reach and marketing effectiveness.

4. For Future Dataset Management:

Lastly, data management practices should enforce consistent formatting of critical fields, particularly date fields. During data cleaning, inconsistencies in the first_publish_date column led to parsing errors when converting to Date type using the lubridate package (mdy() and ymd() functions). Standardizing data formats at the point of entry would greatly reduce cleaning time and enhance future data analysis.

References

University of Virginia Library. (n.d.). *Working with dates and time in R using the lubridate package.* University of Virginia Library. Retrieved April 28, 2025, from https://library.virginia.edu/data/articles/working-with-dates-and-time-in-r-using-the-lubridate-package

SSQC Tutorial. (2022, February 8). *Working with dates and times in R using lubridate* [Video]. YouTube. https://www.youtube.com/watch?v=EbySJGnN-UY

Statistics Globe. (2021, April 20). *Introduction to the lubridate package in R* [Video]. YouTube. https://www.youtube.com/watch?v=HPJn1CMvtmI