Project 2 – Exploring Visualization

ALY6000

Prepared By: Muhammad Umer

Presented To: Prof. John Wilder

Date: 2024/10/05

Introduction and Key Findings

Overview:

This report presents an analysis of book data collected from Goodreads.com, focusing on books published between 1990 and 2020. The dataset, originally archived on Kaggle.com and modified for this project, contains information on 52,448 books across 23 different fields. After data cleaning and filtering, our analysis focuses on a subset of this data.

Key fields of interest in this dataset include:

- Title and author of the books
- Publication date (from which we extracted the year)
- Number of pages
- Average rating and number of ratings
- Publisher information

The time period covered in our analysis spans three decades, from 1990 to 2020, allowing us to observe trends in publishing and reader preferences over a significant period. This dataset provides a rich source of information for exploring various aspects of the book industry, reader behavior, and publishing trends.

Analysis aims to examine book ratings, page counts, publishing trends, and the relationship between these variables over the specified period. We've employed various statistical techniques and data visualization methods to extract meaningful insights from this comprehensive dataset.

Key Findings:

1. **Distribution of Book Ratings:**

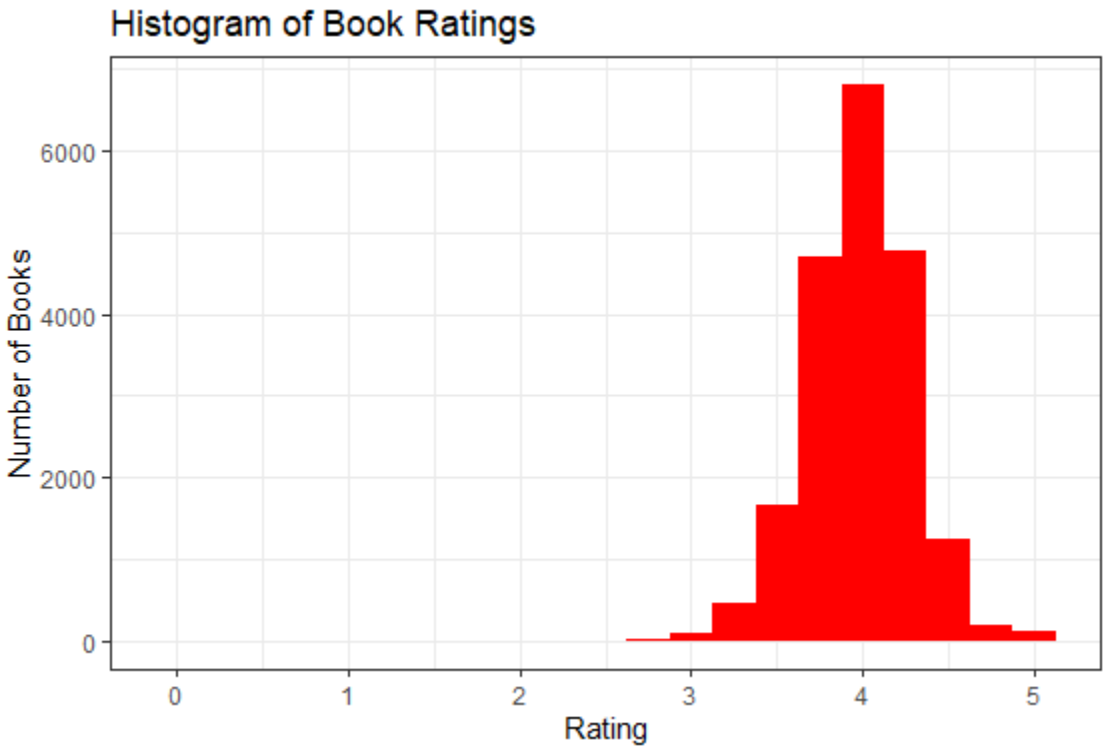

Figure 1: Histogram of book ratings

The histogram of book ratings reveals:

- The majority of books are rated between 3.5 and 4.5 stars.
- There's a noticeable peak around the 4-star rating.
- Very few books receive ratings below 3 stars or above 4.5 stars.

This distribution suggests that Goodreads users tend to rate books positively, with a preference for using whole or half-star ratings.

2. **Page Count Distribution:**

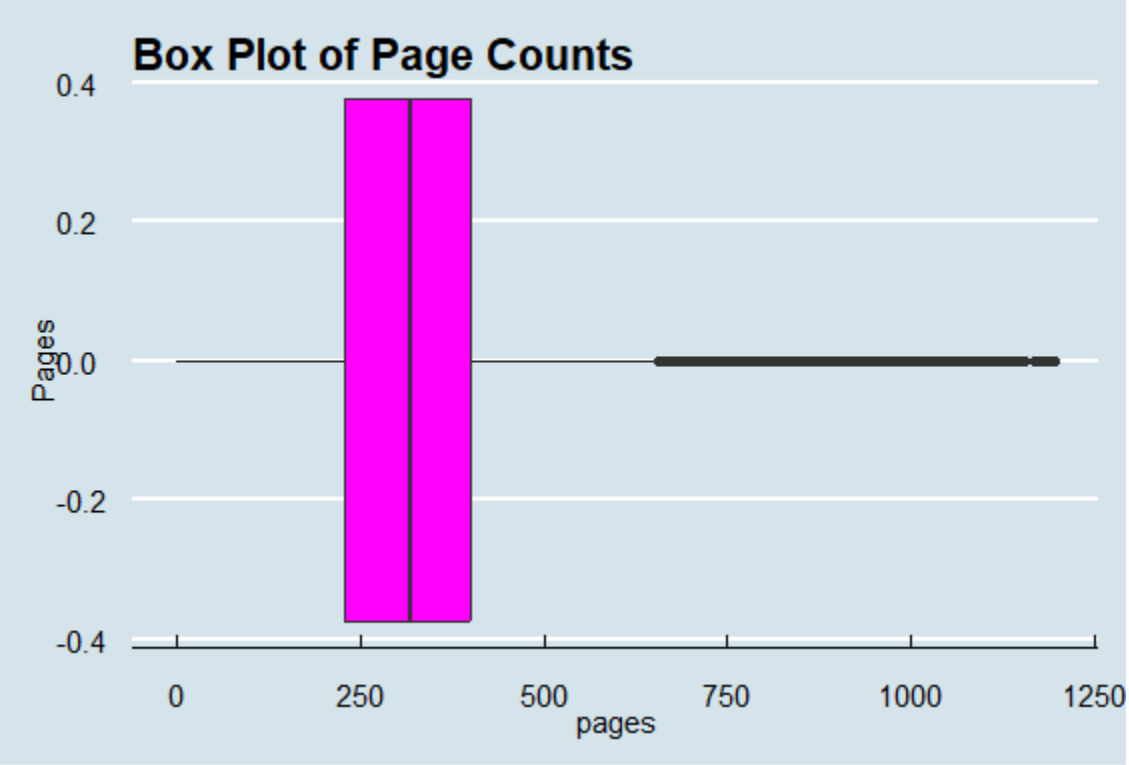

Figure 2: Box Plot of Page Counts

The box plot of page counts shows:

- The median page count is around 300-350 pages.
- There's a wide range of page counts, with some books exceeding 1000 pages.
- The distribution is right-skewed, indicating that while most books are of moderate length, there are some exceptionally long books in the dataset.

3. Top Publishers and Market Share:

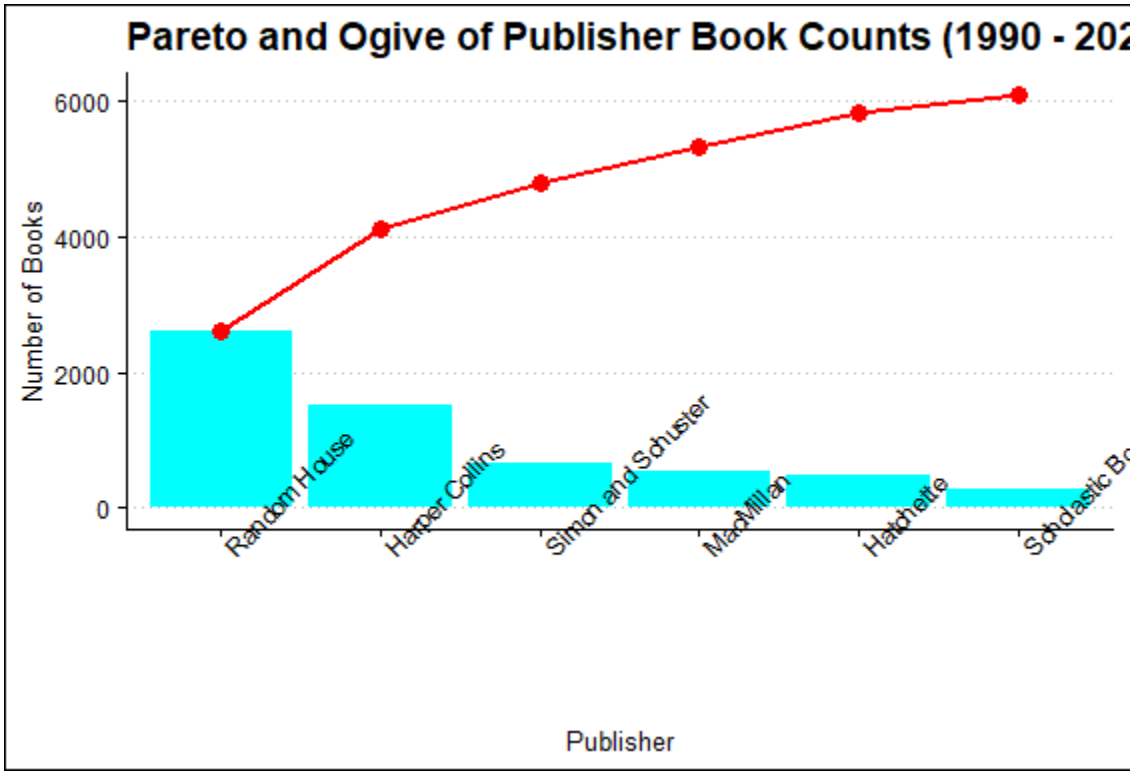

Figure 3: Pareto Chart of Publisher Book Counts

The Pareto chart of publisher book counts reveals:

- Random House and Harper Collins dominate the market, accounting for over 60% of the books in the dataset.
- The top 6 publishers (Random House, Harper Collins, Simon and Schuster, MacMillan, Hatchette, and Scholastic Books) account for approximately 80% of the books.
- There's a long tail of smaller publishers not shown in the chart.

This distribution follows the Pareto principle, where a small number of publishers are responsible for a large portion of the books.

4. Relationship Between Book Length and Rating:

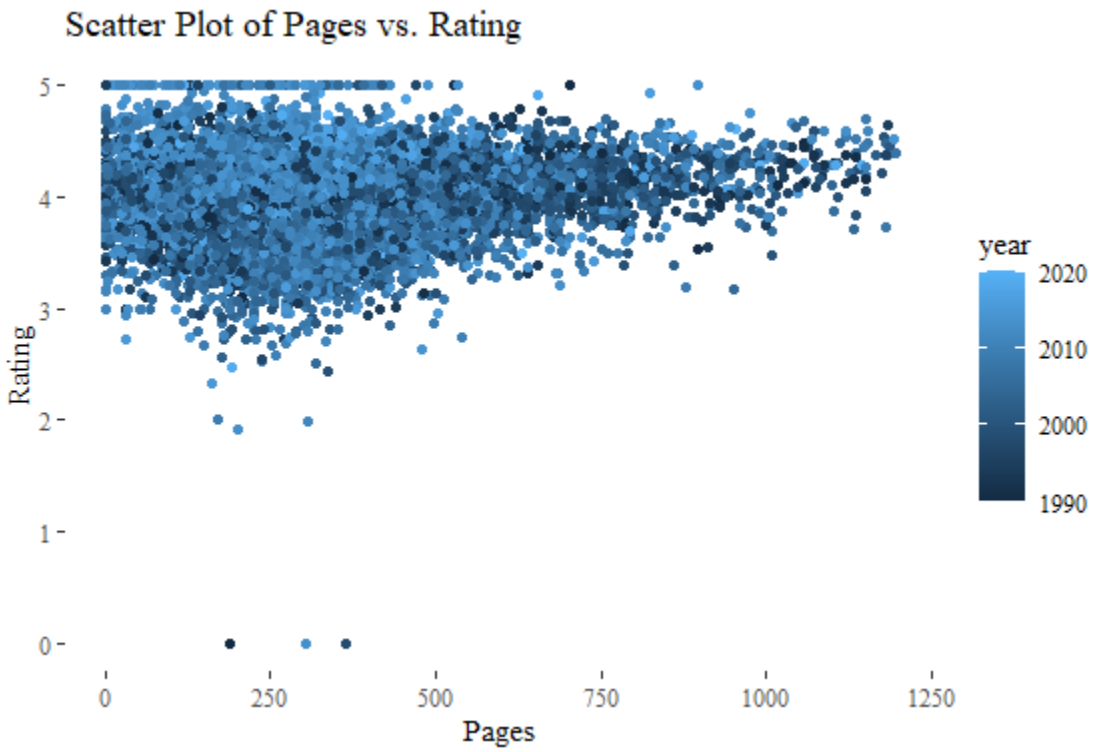

Figure 4: Scatter Plot of Pages vs. Rating

The scatter plot of pages vs. rating indicates:

- There's no strong correlation between book length and rating.
- Books of all lengths can receive high or low ratings.
- The majority of books cluster between 200-600 pages and 3.5-4.5 star ratings.
- The color gradient suggests no clear trend in ratings based on publication year.

5. Publishing Trends Over Time:

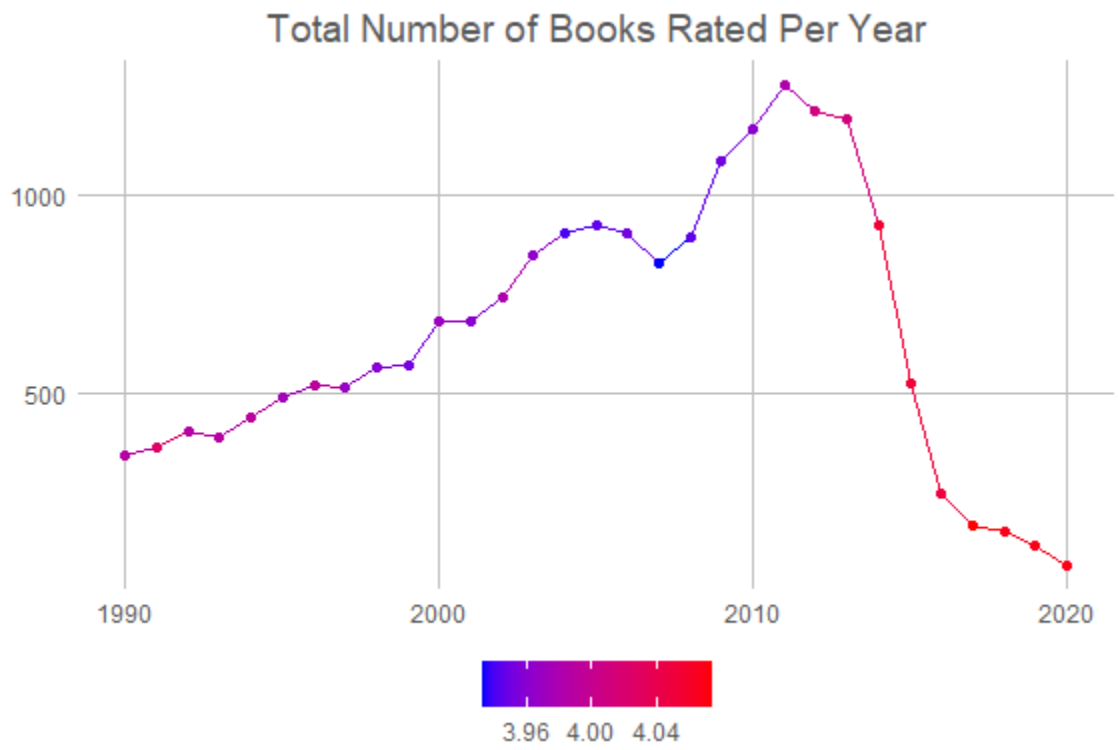

Figure 5: Line Plot of Total Books per Year

The line plot of total books per year shows:

- A general upward trend in the number of books published from 1990 to 2020.
- Some fluctuations, with notable increases in the early 2000s and 2010s.
- The color gradient indicates that average ratings have remained relatively stable over time, with slight variations.

6. Top Authors by Average Rating:

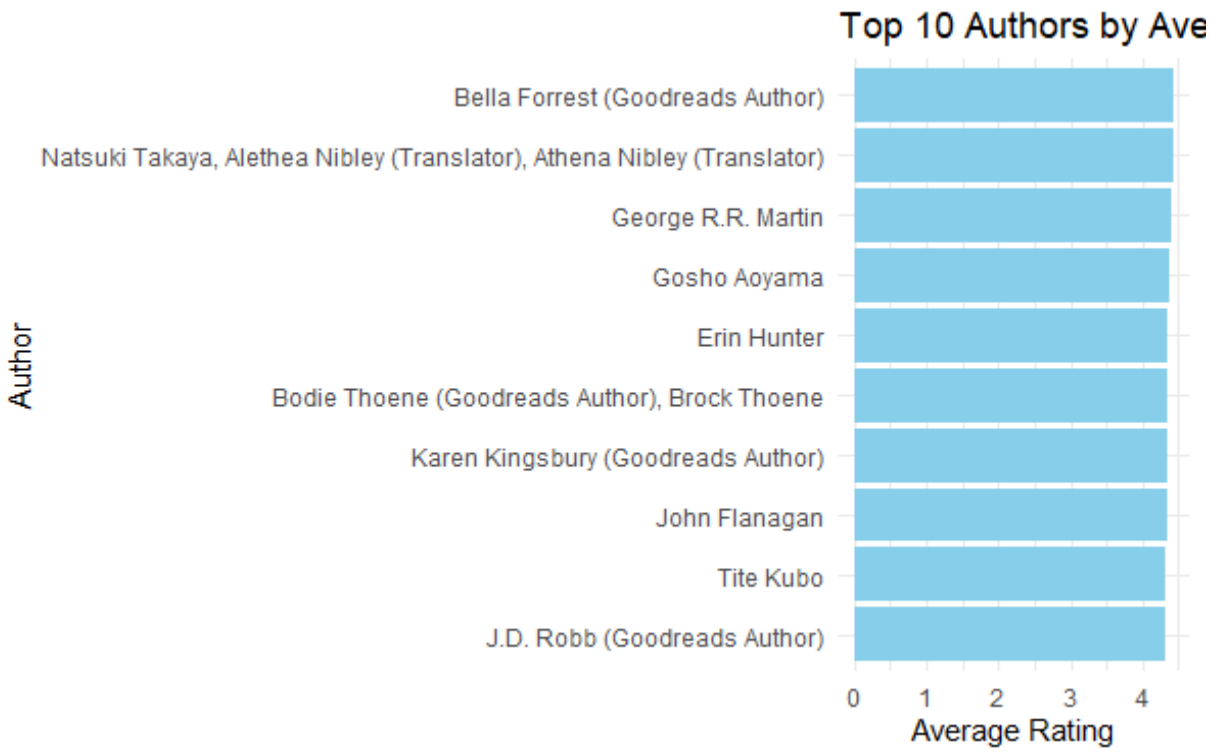

Figure 6: Top 10 Authors by Average

The bar chart of top authors reveals:

- The highest-rated authors (with at least 10 books) have average ratings between 4.2 and 4.5 stars.
- These top authors likely represent a mix of critically acclaimed and popular writers.

Conclusion and Recommendations:

Based on the analysis, we can draw the following conclusions:

1. Book ratings on Goodreads tend to be positively skewed, with most books receiving ratings between 3.5 and 4.5 stars.
2. There's a wide range of book lengths, but most books fall between 200-600 pages.
3. The publishing industry is dominated by a few large publishers, following a Pareto distribution.
4. Book length doesn't strongly correlate with ratings, suggesting that quality is more important than quantity.
5. The number of books published has generally increased from 1990 to 2020, possibly reflecting the growth of the publishing industry or increased data collection by Goodreads.

Recommendations:

1. Focus on quality over quantity, as book length doesn't significantly impact ratings.
2. Consider the 300-350 page range as a sweet spot for book length, as this is the median range.
3. For new or smaller publishers, explore niche markets or specialized genres to compete with the dominant publishers.
4. Authors should aim for consistently high-quality work, as the top-rated authors maintain average ratings above 4.2 stars.
5. Publishers should be aware of the upward trend in book publications and consider how to stand out in an increasingly crowded market.

References

Bluman, A. G. (2018). Elementary statistics: A step by step approach. McGraw-Hill Education.

Holtz, Y. The complete ggplot2 tutorial. R Graph Gallery. https://r-graph-gallery.com/ggplot2-package.html

W3Schools. R tutorial. https://www.w3schools.com/R/

Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686