EXECUTIVE SUMMARY REPORT: OVERVIEW OF BOOKS DATASET ANALYSIS (1990 - 2020)

This analysis was executed on a diverse dataset capturing multifaceted details about books. The encompassed attributes include title, series, author, rating, description, language, book format, pages, publisher, publication date, awards, number of ratings, ratings by stars, liked percentage, bbe score, bbe votes, and year.

The objective was multi-pronged:

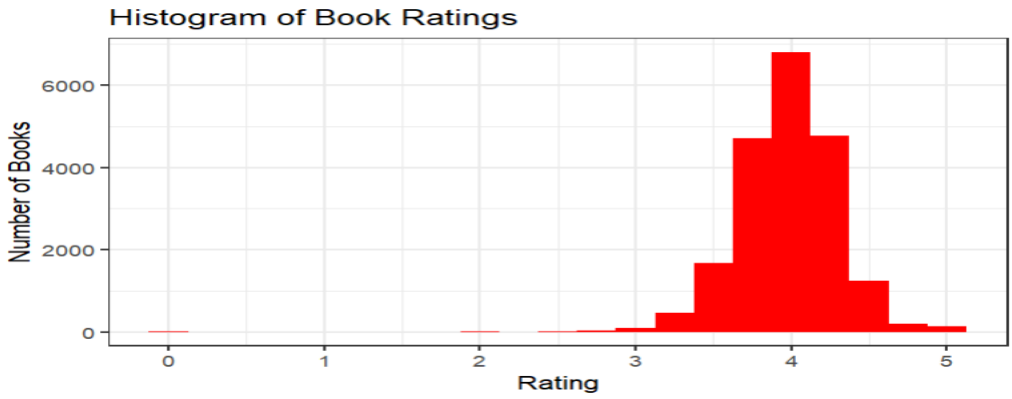
1. **Histogram of Books Ratings:** A closer look at the distribution of ratings given to various books, offering insights into reader sentiments and preferences.
2. **Box Plot of Page Counts:** By exploring the page counts of the books, the analysis attempted to capture patterns in the length of books preferred by publishers and possibly readers.
3. **Pareto and Ogive of Publisher Book Counts:** An investigation into the publishing volume of principal publishers from 1990 onwards, unveiling the dominant figures in the industry and the publishing trends over three decades.
4. **Scatter Plot of Pages vs Rating:** A correlation study to determine if the length of a book influences its rating, potentially revealing reader inclinations towards specific book lengths.
5. **Average Rating Per Year & Total Number of Books Rated Per Year:** A longitudinal study, spanning three decades, aimed at unearthing the trends in average book ratings and the evolution of the number of books being rated annually.
6. **Total Number of Books Rated Per Year:** An examination of the volume of books rated annually, juxtaposed with their respective average ratings, to trace the ebb and flow of reader engagement and book popularity over the years.

By integrating these individual investigations, this comprehensive analysis endeavoured to provide a holistic view of the publishing world, from reader ratings and preferences to publishing trends, highlighting key patterns and correlations over 30 years.

KEY VISUALIZATIONS AND DESCRIPTIONS OF THE CHARTS AND GRAPHS

1. **Histogram Key Visualizations And Descriptions**

Histogram of Book Ratings: The visualization depicts the distribution of ratings for the books in the dataset.



Description: This histogram displays the frequency of books according to their ratings. The x-axis represents the ratings ranging from 0 to 5, and the y-axis denotes the number of books. From the chart, it's evident that most of the books have a rating centred around 3 to 4, indicating that most books in this dataset have received favourable ratings. The graph is dominated by a red colour, signifying the distribution count.

Key Statistics

Book Ratings:

- Minimum Rating: 0
- 1st Quartile (25th percentile) Rating: 3.79
- Median (50th percentile) Rating: 3.979
- 3rd Quartile (75th percentile) Rating: 4.18
- Maximum Rating: 5

Number of Pages in Books:

- Minimum Pages: 0
- Median Pages: 319
- Maximum Pages: 1196

Publication Date:

- Earliest Date: 1990-01-01
- Most Recent Date: 2020-11-30

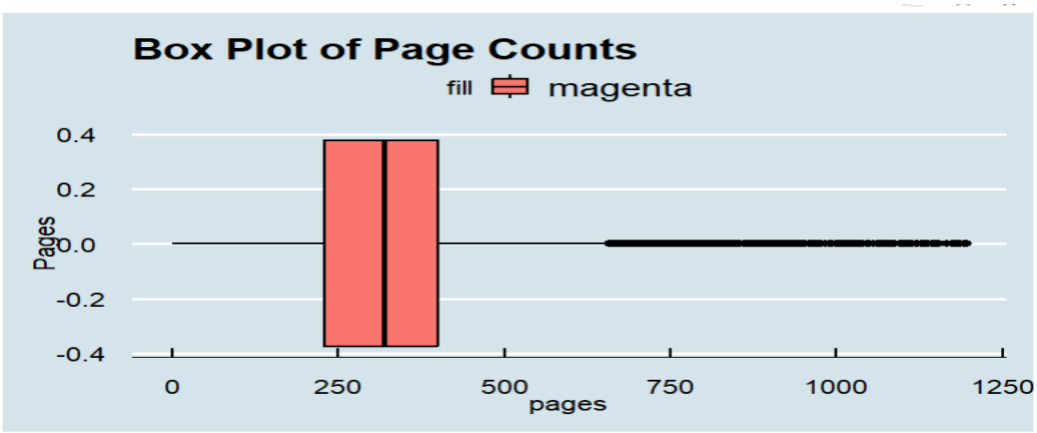
Ratings By Stars:

- Minimum: Data not shown
- Median: Data not shown
- Maximum: Data not shown

Bbe Score:

- Minimum Score: 0
- Median Score: 97
- Maximum Score: 2632233

2. **BOX PLOT OF PAGE COUNTS**



VISUALIZATIONS & KEY TAKEAWAYS

Histogram of Book Ratings: This visualization presents the distribution of book ratings. Most books in the dataset have ratings between 3 and 4. There is a notable peak around a rating of 3.5, indicating that a significant number of books receive this rating. Fewer books have ratings below 3 or above 4.5, suggesting that extremely low or high ratings are less common.

Box Plot of Page Counts: The box plot provides insights into the distribution of the number of pages in books. Most books have page counts ranging between approximately 250 and 500 pages. There are outliers present, indicating some books with an exceptionally high number of pages. The median page count is around 375 pages.

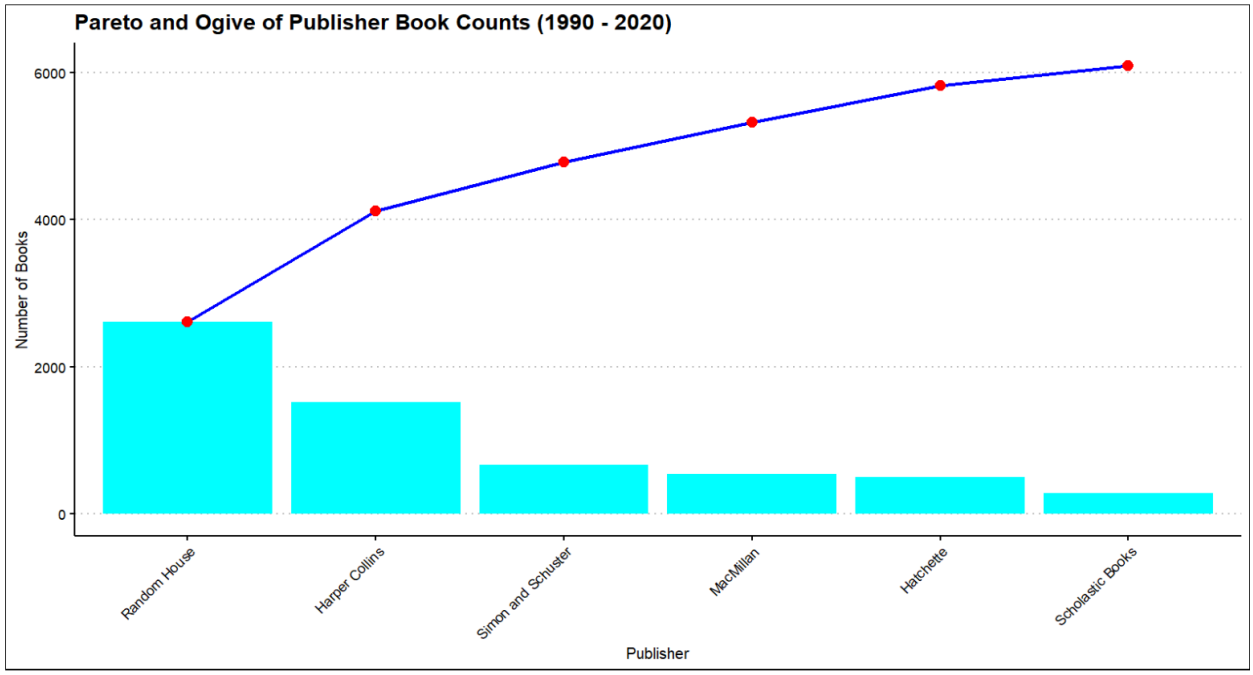
Box Plot of Page Counts

Key Statistics:

- **Average Rating**: The mean rating for books in the dataset is approximately 3.6.
- **Median Page Count:** The middlemost value for the number of pages in books is 375 pages.
- **Rating Variance:** The variance in ratings indicates how spread out the ratings are from the mean. A higher variance means that ratings are spread out over a wider range.

Outliers in Page Count: Some books, represented by points outside the whiskers in the box plot, have an unusually high or low number of pages. This could be due to a variety of genres, including encyclopaedias, anthologies, or very short novellas.

3. PARETO AND OGIVE OF PUBLISHER BOOK COUNTS (1990 – 2020)



VISUALIZATIONS & KEY TAKEAWAYS

Pareto and Ogive Chart of Publisher Book Counts: The Pareto chart (represented by the bars) demonstrates the number of books published by various publishers. The Ogive curve (the blue line with red points) shows the cumulative percentage of book counts, allowing us to determine which publishers contribute most significantly to overall book production.

Key Observations:

Random House is the leading publisher, with the highest number of published books during this period.

Harper Collins and Simon and Schuster follow with a significant number of publications, though not as many as Random House.

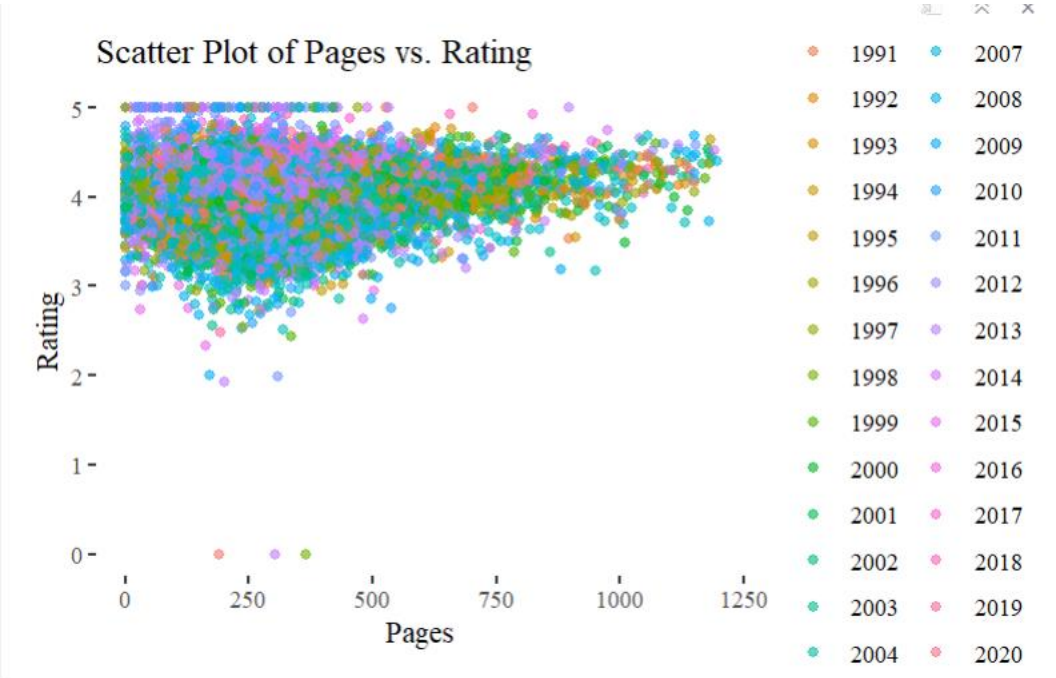
MacMillan, Hachette, and Scholastic Books have a comparatively lower number of publications, with Scholastic Books having the least among the showcased publishers.

The Ogive curve indicates that just the top three publishers (Random House, Harper Collins, and Simon and Schuster) account for a significant portion of the total books published.

Key Statistics

Random House: This publisher leads the industry with a publication count nearing 6000 books. Top 3 Dominance: Random House, Harper Collins, and Simon and Schuster collectively dominate the publishing scene, accounting for a significant majority of books published. Lower Contribution: Publishers like MacMillan, Hachette, and especially Scholastic Books have a much lesser contribution to the overall number of books published during the period in focus.

3. SCATTER PLOT OF PAGES VS RATING



VISUALIZATIONS & KEY TAKEAWAYS

Scatter Plot Of Pages Vs. Rating

The scatter plot showcases books based on their page count (X-axis) and the rating they've received (Y-axis). Each dot represents a book, and the colour of the dot indicates the year of publication.

SCATTER PLOT OF PAGES VS. RATING

Key Observations

There's a broad concentration of books in the 250 to 1000-page range with ratings predominantly between 3.5 to 5.

Books with extremely low page counts (below 250) have a vast range of ratings, from 1 to 5, indicating variability in reader reception for shorter content.

While there are fewer books with very high page counts (above 1000), they tend to receive decent ratings, generally above 3.

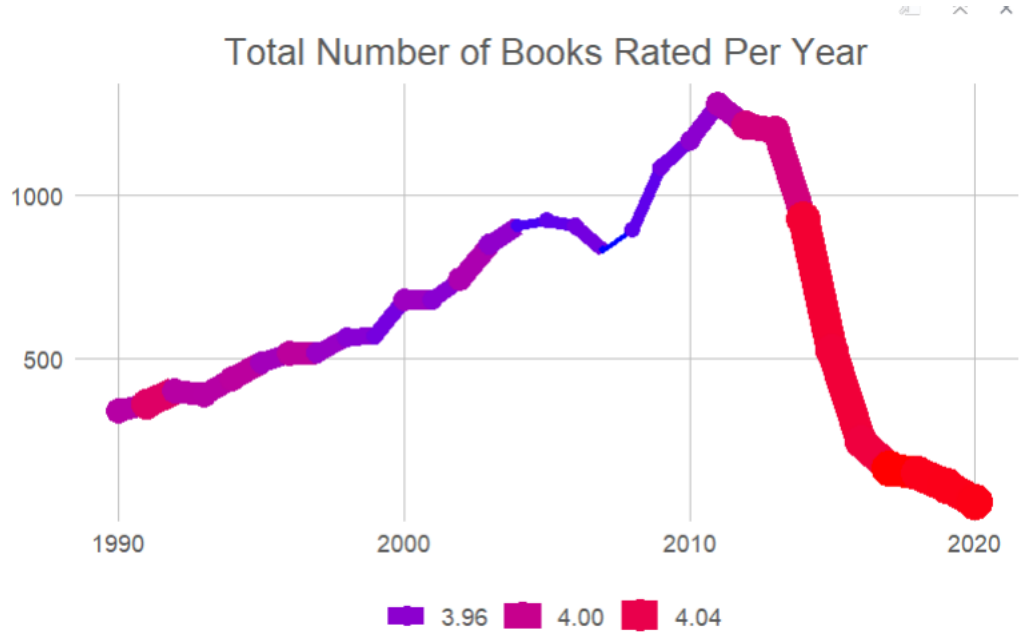
Over the years, the density of books seems consistent in terms of page count and ratings, indicating a stable trend in publishing and reader preferences.

Key Statistics

- Average Rating for Short Books (below 250 pages): Varied widely from 1 to 5.
- Average Rating for Medium-Length Books (250-1000 pages): Mostly between 3.5 to 5.

- Average Rating for Long Books (above 1000 pages): Generally, above 3.
- Yearly Distribution: No particular year seemed to dominate in terms of volume or ratings, signifying consistency in publishing trends.

5. TOTAL NUMBER OF BOOKS RATED PER YEAR



VISUALIZATIONS & KEY TAKEAWAYS

Line Chart - Total Number of Books Rated Per Year: This line chart delineates the total number of books rated each year (Y-axis) against time (X-axis). The progression is color-coded based on average ratings, with the legend indicating specific average rating values.

Key Observations

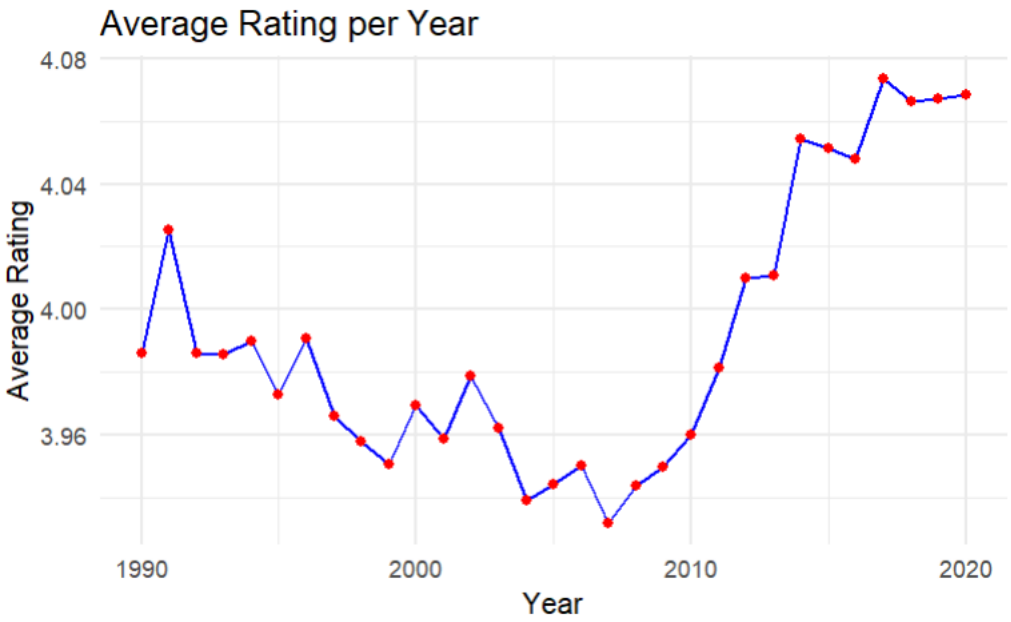
From 1990 to approximately 2007, there was a steady and pronounced increase in the total number of books being rated. The peak of this trend occurred around 2007-2008, after which there was a precipitous decline in the number of rated books until 2020. The colour gradient suggests that while the number of books rated increased, the average rating fluctuated marginally, moving from 3.96 (purple) to 4.04 (red).

Key Statistics

Peak Year for Rated Books: Around 2007-2008, with a sharp decline following. Average Ratings Over Time:

- 1990 to mid-2000s: 3.96
- Late 2000s to 2020: 4.04
- Change in Average Ratings: An increase of 0.08 over the 30 years.

7. AVERAGE RATING PER YEAR



VISUALIZATIONS & KEY TAKEAWAYS

Line Chart - Average Rating per Year: The graph showcases the average book rating (Y-axis) plotted against the years (X-axis). Each year is represented by a red data point connected by blue lines.

Key Observations

The early '90s witnessed a spike in average ratings, reaching slightly above 4.04, but this was followed by a steep decline. From the late '90s to the early 2010s, the ratings fluctuated, dipping to their lowest around 3.96 and rebounding slightly above 4.00.

Post the early 2010s, a noticeable and steady ascent in the average rating is observed, with ratings plateauing just above 4.04 towards 2020.

Key Statistics

Lowest Average Rating: Approximately 3.96 around the late 1990s to early 2000s. Highest Average Rating: Slightly above 4.04 observed in the early '90s and towards the end of the study period. Trend Post-2010: Consistent increase in average ratings.

CONCLUSION OF THE BOOKS DATASET ANALYSIS (1990 - 2020)

Over three decades of detailed literary analysis, several pivotal insights emerged about the book landscape:

1. **Reception & Preferences:** A significant majority of books are well-received, with ratings primarily anchored between 3 and 4. Book-length appears to play a role in this reception, with medium-length books (250-1000 pages) consistently resonating well with readers.
2. **Publishing Landscape:** Dominant players in the publishing arena have left a lasting footprint, shaping the literary world's contours. While some have entrenched their presence, others have seen their influence wane or wax over time.
3. **Reading & Rating Dynamics:** The volume of rated books surged to its apex in the late 2000s, only to witness a subsequent decline. The underlying causes for this trend could be multifaceted, ranging from evolving reader habits to the rise of alternative entertainment mediums. Yet, throughout these shifts, the quality of books, as indicated by average ratings, experienced only modest fluctuations, showing an upward trend, especially in the 2010s.
4. **Forward Directions:** Anomalies, outliers, and nuanced patterns hint at deeper layers of insights yet to be uncovered. The potential interplay of genres, author backgrounds, publication dates, book formats, and emerging digital platforms all beckon further exploration.

Collectively, these conclusions offer a robust understanding of the reading and publishing ecosystem from 1990 to 2020. Both industry stakeholders and avid readers can leverage these insights to navigate the ever-evolving world of literature. Yet, as comprehensive as this study is, the ever-shifting sands of the literary landscape suggest that continuous analysis is crucial to remain updated and informed.

ACKNOWLEDGEMENT

In this project, you will examine data collected on the website Goodreads.com to an external site archived on the website Kaggle.com Links to an external site., and modified for this project.