PROJECT 4: COMPREHENSIVE REPORT ON BOOKS DATASET ANALYSIS

Overview

The dataset contains information about various books, including details such as the book's title, author, average rating, number of pages, language, and publisher. The column names are:

1. **bookID:** A unique identifier for each book.
2. **title:** The title of the book.
3. **authors:** The name(s) of the author(s) who wrote the book. Multiple authors are separated by a slash.
4. **average_rating:** The average rating of the book.
5. **isbn:** The International Standard Book Number, a unique identifier for books.
6. **isbn13:** A 13-character ISBN to identify the book.
7. **language_code:** The language in which the book is written.
8. **num_pages:** The number of pages in the book.
9. **ratings_count:** The number of unique users who have rated the book.
10. **text_reviews_count:** The number of text reviews the book has received.
11. **publication_date:** The publication date of the book.
12. **publisher:** The publisher of the book.

The goal of this analysis is to explore the dataset, identify key characteristics and patterns, and present insights into the Books dataset.

Before beginning your exploratory data analysis, develop 3–4 data questions.

1. What is the distribution of average ratings for the books in the dataset?
2. Which authors have the highest number of books?
3. How does the number of pages correlate with the average rating of a book?
4. What are the most common languages in which the books are written?

Part I - Exploring

1. Data Cleaning
   - Renamed columns to more meaningful names.
   - Converted 'publication_date' to datetime format.
   - Handled missing values in 'publication_date'.

In summary, the data-cleaning steps have been completed

- Columns have been renamed for easier handling.
- No missing values were found except for 2 rows with invalid 'publication_date', which have been removed.
- Data types have been corrected; specifically, 'publication_date' is now of datetime type.

2. Descriptive Statistics

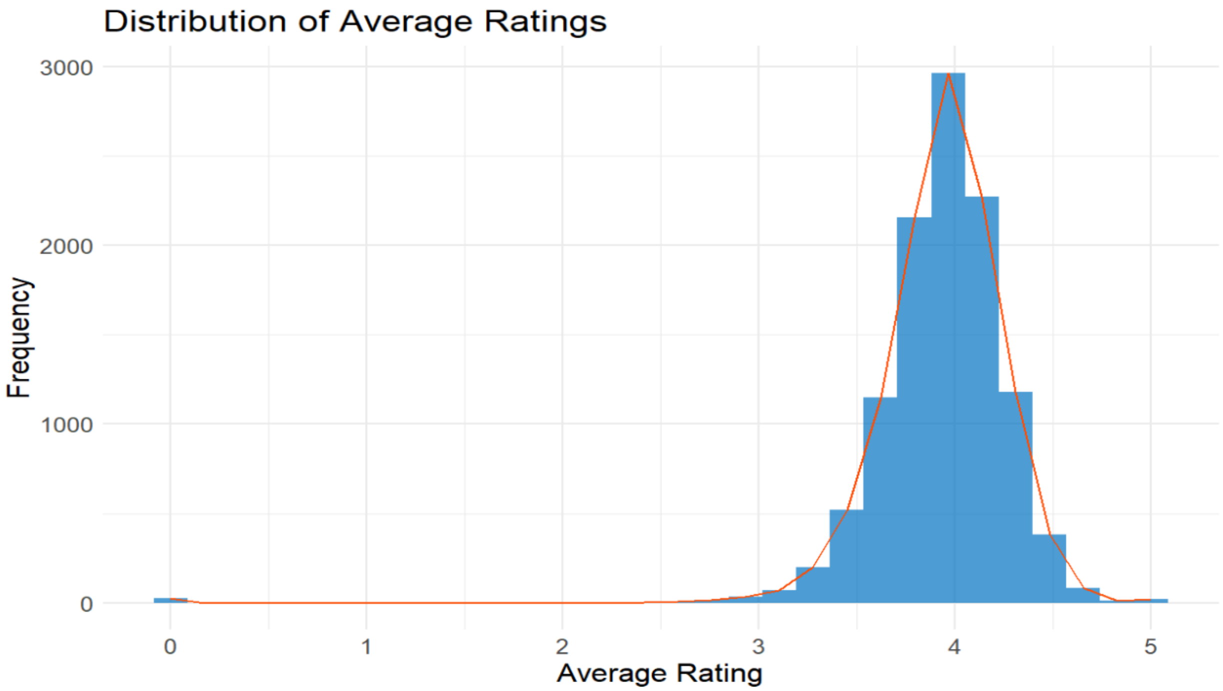
   The descriptive statistics for the interesting variables are as follows:

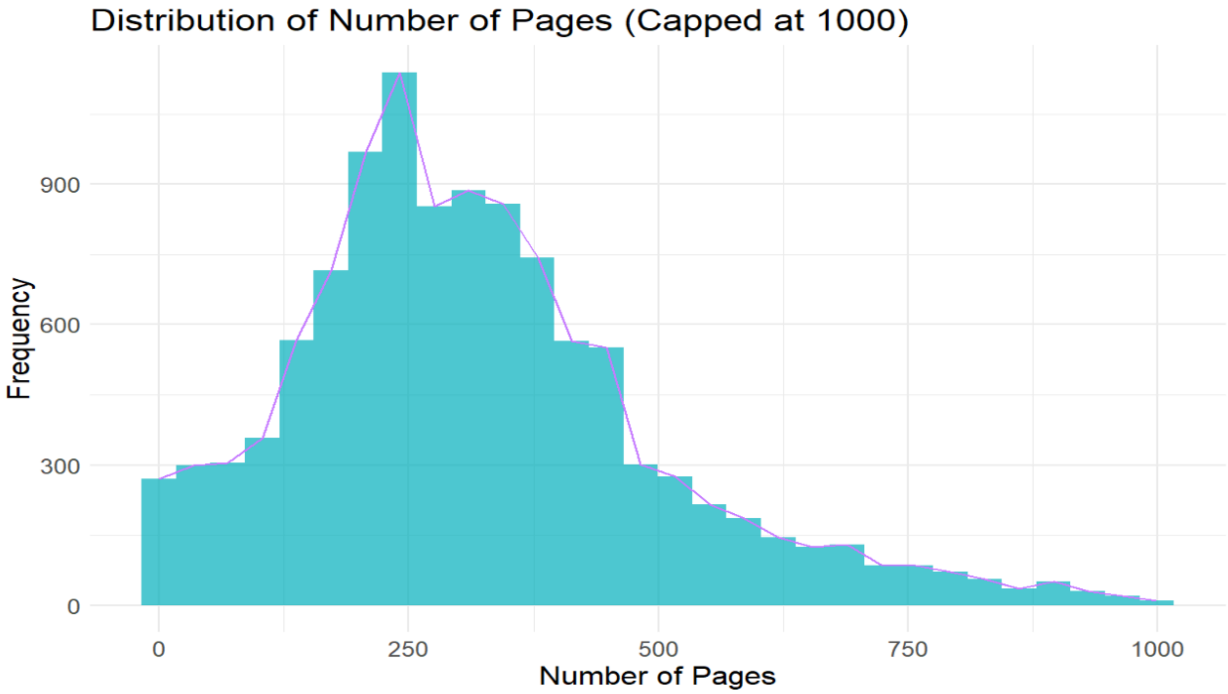
   - **average_rating**
     - Count: 11,121

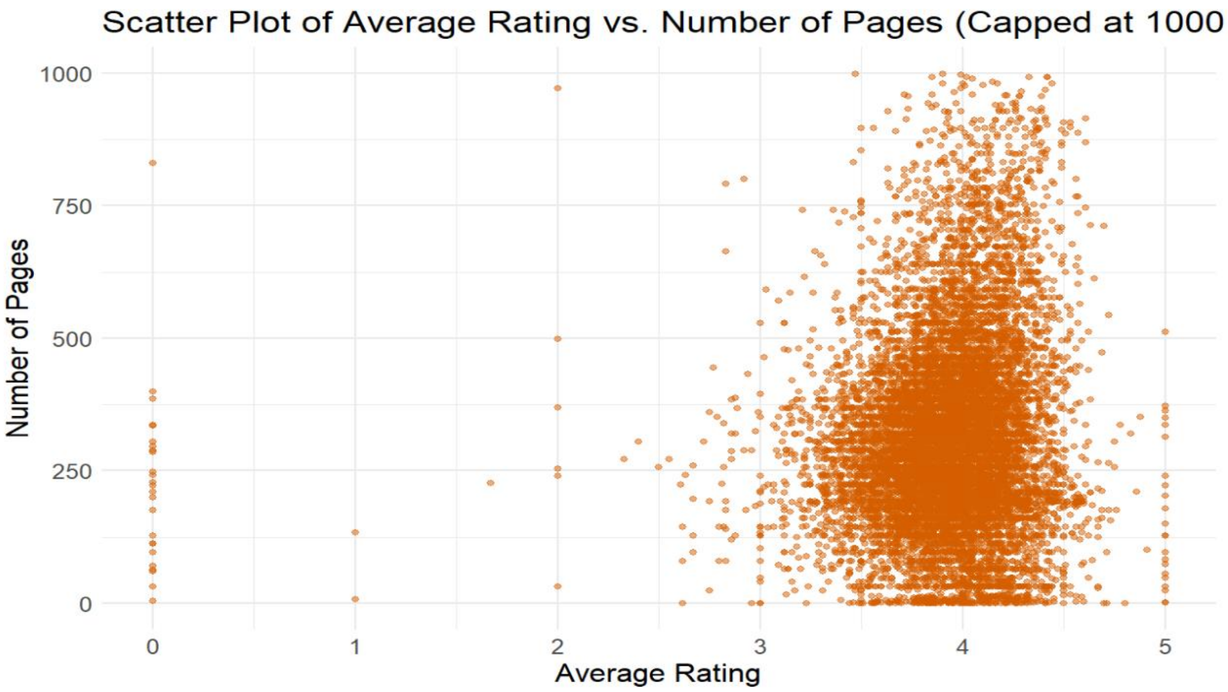
    - Mean: 3.93
    - Standard Deviation: 0.35
    - Min: 0.00
    - 25th Percentile: 3.77
    - Median: 3.96
    - 75th Percentile: 4.14
    - Max: 5.00
- num_pages
    - Count: 11,121
    - Mean: 336.34
    - Standard Deviation: 241.13
    - Min: 0
    - 25th Percentile: 192
    - Median: 299
    - 75th Percentile: 416
    - Max: 6,576
- ratings_count
    - Count: 11,121
    - Mean: 17,945.12
    - Standard Deviation: 112,509.10
    - Min: 0
    - 25th Percentile: 104
    - Median: 745
    - 75th Percentile: 4,996
    - Max: 4,597,666
- text_reviews_count
    - Count: 11,121
    - Mean: 542.12
    - Standard Deviation: 2,576.85
    - Min: 0
    - 25th Percentile: 9
    - Median: 47
    - 75th Percentile: 238
    - Max: 94,265

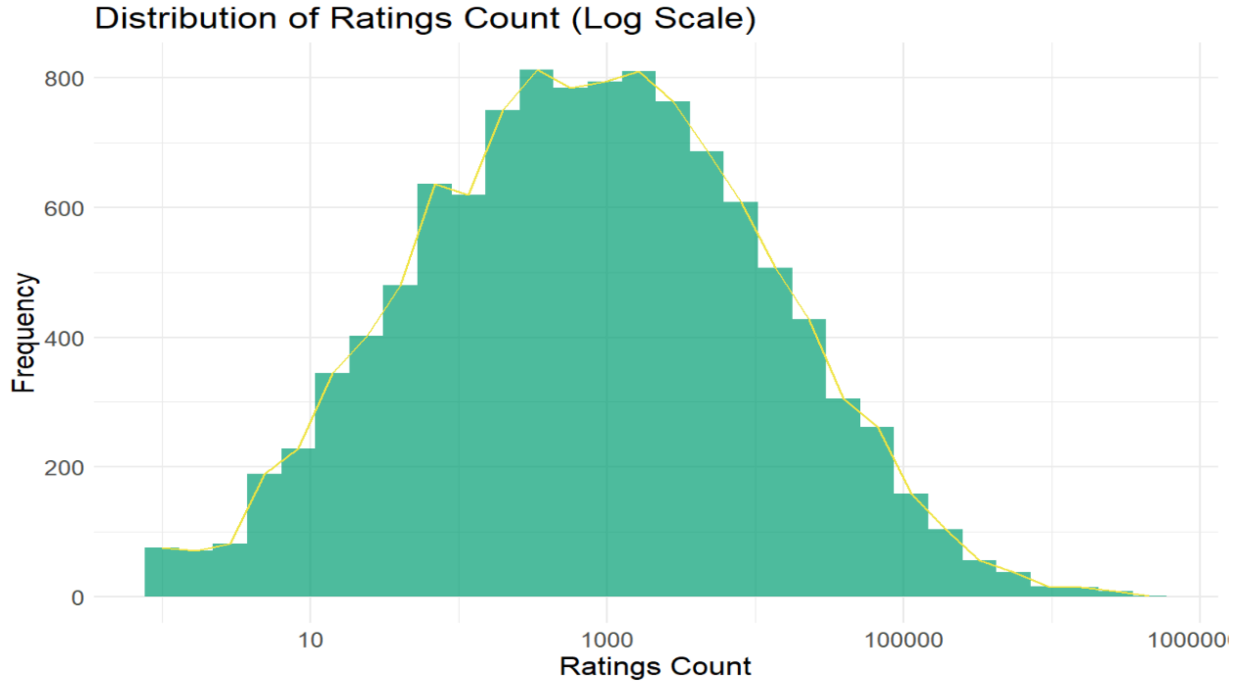
3. Data Visualization

   Plotted distributions and relationships for the aforementioned key variables.


Distribution of Average Ratings
Frequency
3000
2000
1000
0
0
1
2
3
4
5
Average Rating


Distribution of Number of Pages (Capped at 1000)
Frequency
900
600
300
0
0
250
500
750
1000
Number of Pages





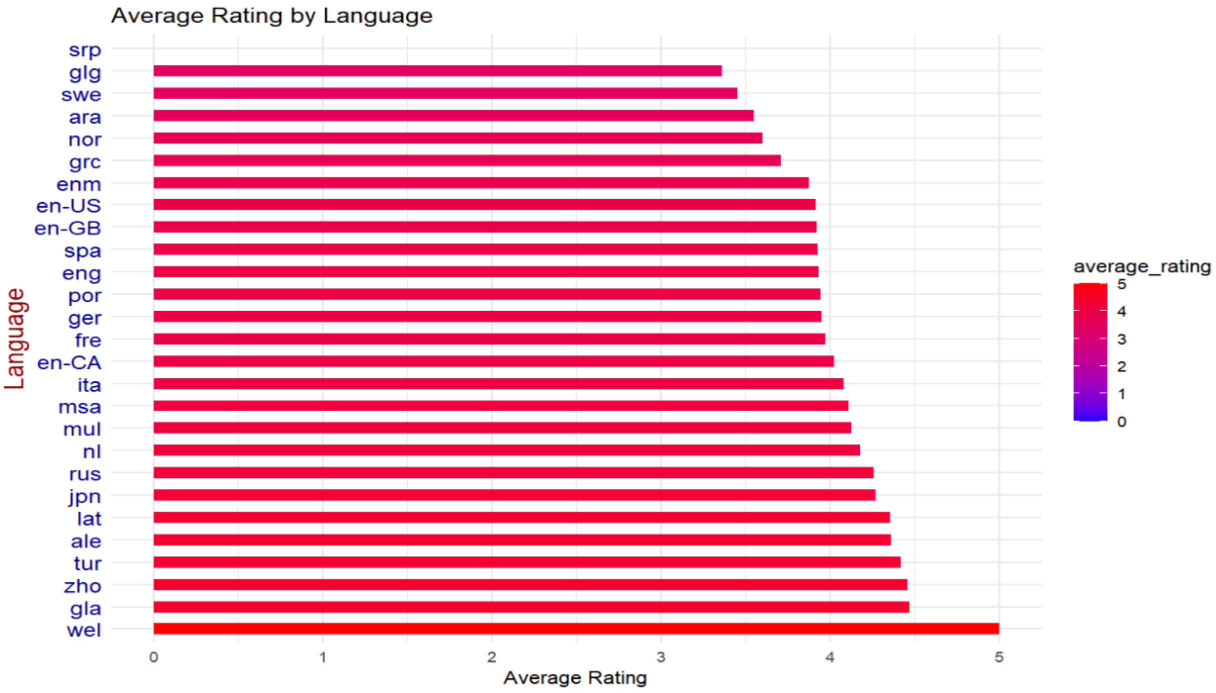
The Following Are The Visualizations For The Interesting Variables:

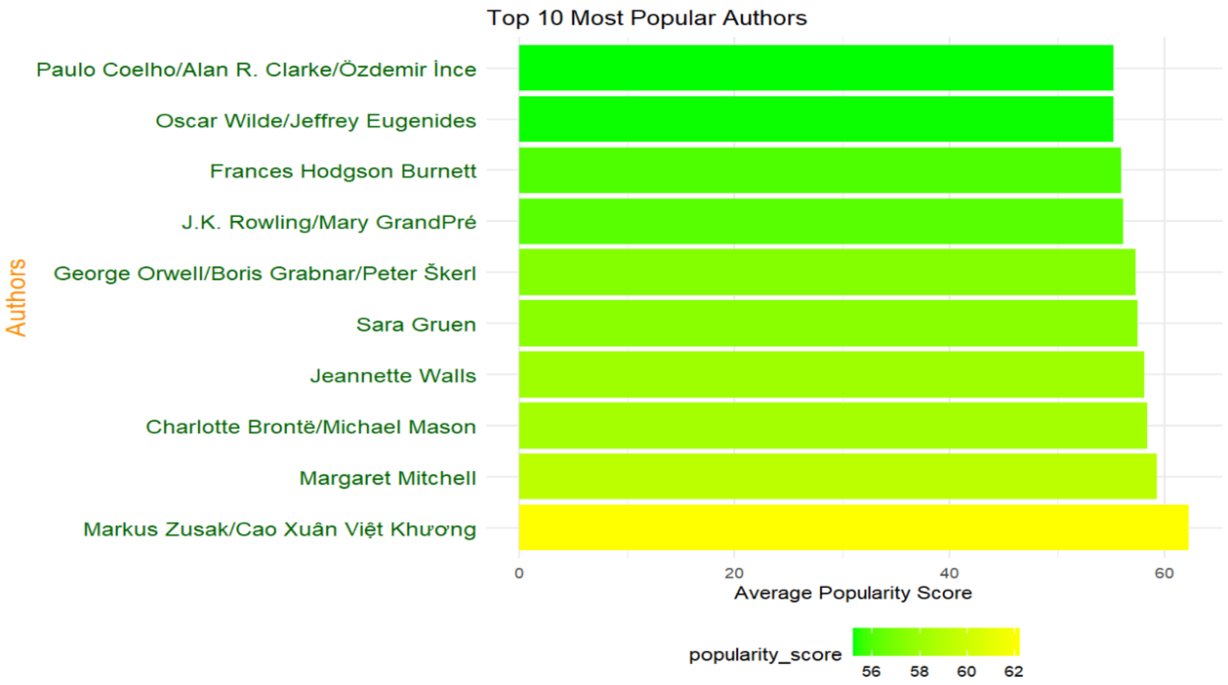
1. Distribution of Average Ratings: The histogram shows that most books have an average rating between 3.5 and 4.5, with the distribution slightly skewed towards higher ratings.

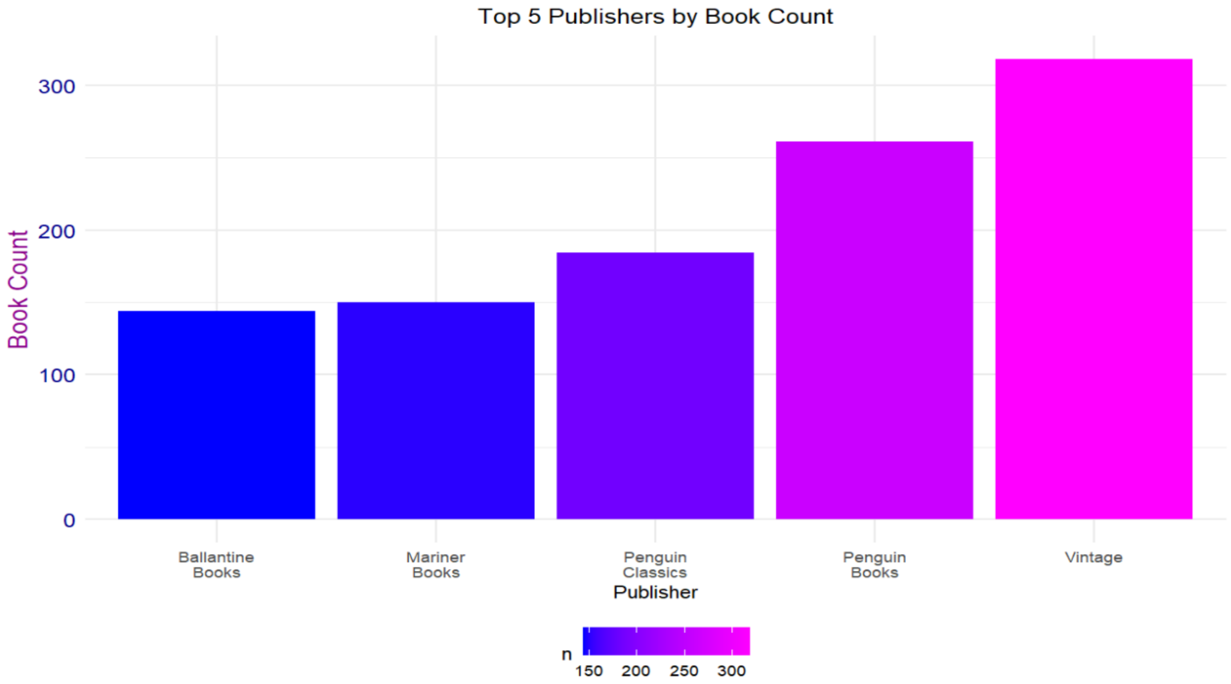
2. Distribution of Number of Pages (Capped at 1000): Most books have between 200 and 400 pages. The distribution shows a long tail, indicating that very few books have many pages.
3. Scatter Plot of Average Rating vs. Number of Pages (Capped at 1000): There doesn't seem to be a strong correlation between the number of pages and the average rating. Books with varying lengths have received both high and low ratings.
4. Distribution of Ratings Count (Log Scale): The number of ratings varies widely, so a logarithmic scale was used. Most books have received fewer than 1,000 ratings, but a few popular titles have received hundreds of thousands or even millions of ratings.

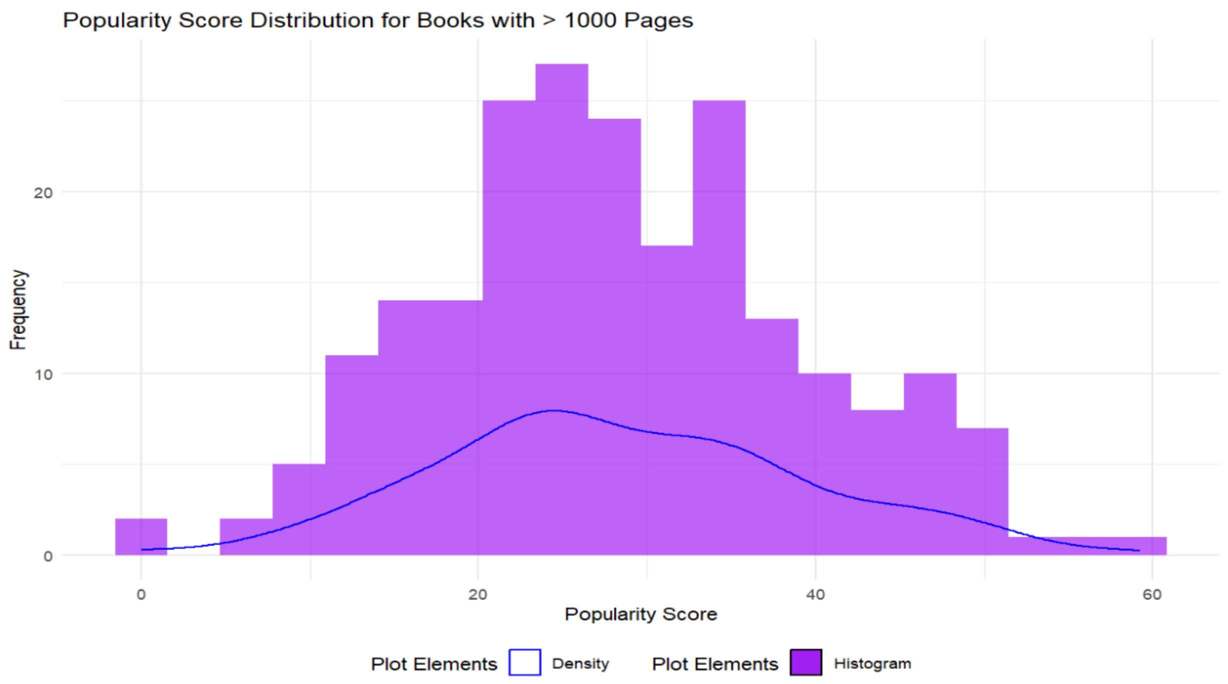
Part II - Expanding

1. Feature Engineering
- Created a 'popularity_score' variable that combines 'average_rating' and 'ratings_count'.
- Developed a 'reviews_per_rating' variable to measure the number of reviews per rating.
2. Data Summarization
- Grouped data by 'language' to summarize average ratings and popularity scores.
- Identified the top 10 most popular authors based on 'popularity_score'.
- Counted the number of books by each publisher.
3. Advanced Visualizations
- Visualized the average ratings by language.
- Showcased the top 10 most popular authors.
- Displayed the top 5 publishers by the number of books.
- Analysed the popularity score distribution for books with more than 1,000 pages.




Top 10 Most Popular Authors
Authors
Paulo Coelho/Alan R. Clarke/Özdemir İnce
Oscar Wilde/Jeffrey Eugenides
Frances Hodgson Burnett
J.K. Rowling/Mary GrandPré
George Orwell/Boris Grabnar/Peter Škerl
Sara Gruen
Jeannette Walls
Charlotte Brontë/Michael Mason
Margaret Mitchell
Markus Zusak/Cao Xuân Việt Khương
0
20
40
60
Average Popularity Score
popularity_score
56
58
60
62


Top 5 Publishers by Book Count
Book Count
300
200
100
0
Ballantine Books
Mariner Books
Penguin Classics
Penguin Books
Vintage
Publisher
n
150
200
250
300



The new data frames provide different views of the data

1. **Language Summary:** This summarizes the average 'popularity_score' and 'average_rating' for books grouped by language. For instance, books written in 'eng' (English) have an average popularity score of approximately 27.13 and an average rating of approximately 3.93.
2. **Top 10 Most Popular Authors:** This ranks authors based on their average 'popularity_score'. Markus Zusak/Cao Xuân Việt Khương tops the list with a score of approximately 62.19.
3. **Publisher Count:** This counts the number of books published by each publisher. 'Vintage' is the leading publisher with 318 books, followed by 'Penguin Books' with 261 books.
4. **Large Books Summary:** This provides summary statistics for books with more than 1000 pages. The average rating for these books is approximately 4.21, and their average popularity score is around 28.54.

Part III - Communicating

1. Key Insights

- Most books are in English and have an average rating of around 3.5 to 4.0.
- Certain authors and publishers dominate in terms of popularity and the number of books published.
- Books with more than 1000 pages tend to have higher popularity scores.

2. Future Exploration

- Investigate how book characteristics have evolved.
- Analyze book popularity by genre.
- Study the career trajectory of authors.
- Examine the influence of localization on book popularity.
- Conduct sentiment analysis on text reviews.

This report provides a structured overview of the dataset and the insights gleaned from the analysis. Further exploration based on the outlined questions could provide a more in-depth understanding and valuable findings. I have the following insights about the Dataset:

1. **Language and Ratings:** Most books in the dataset are in English and have an average rating of around 3.5 to 4.0. However, the dataset does contain books in multiple languages, providing a diverse range of literature.
2. **Popular Authors:** The dataset contains works from a wide range of authors, some of whom have achieved high popularity scores. This score is a combination of the average rating and the log of the number of ratings, offering a more comprehensive view of an author's popularity.
3. **Publishers:** Certain publishers dominate the dataset in terms of the number of books published. Publishers like 'Vintage' and 'Penguin Books' have a significant number of titles, indicating their prominence in the publishing world.
4. **Book Characteristics:** The number of pages in a book doesn't seem to have a strong correlation with its average rating. However, books with more than 1000 pages tend to have a higher popularity score, indicating they are well-received and widely read.
5. **Review Behavior:** Not all readers who rate a book leave a text review. The ratio of text reviews to ratings varies, and this could be an interesting metric for understanding reader engagement.

Observations or Follow-up Questions for Future Exploration

1. **Temporal Trends:** How have average ratings or popularity scores changed over time? Are there noticeable trends related to the publication date?
2. **Genre Analysis:** The dataset doesn't contain genre information. Could external data be incorporated to analyze popularity or ratings by genre?
3. **Author's Career Trajectory:** How do ratings and popularity scores evolve throughout an author's career? Are there patterns to an author's "peak" years?
4. **Localization:** Are there specific publishers or authors who are more popular in certain languages or regions?
5. **Review Analysis:** Could sentiment analysis on text reviews provide more insights into reader satisfaction, beyond numerical ratings?

These questions and observations could serve as the basis for more in-depth analyses and could yield valuable insights into reader preferences and behaviours.

Citation(s):

Goodreads Books Dataset. (n.d.). Kaggle. Retrieved October 22, 2023, from https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks/