Data Narrative: Goodreads' Most Popular Books

Sriram Srinivasan
Electrical Engineering
Indian Institute of Technology, Gandhinagar
Gandhinagar, India
sriram.srinivasan@iitgn.ac.in, Roll No. 22110258

***Abstract*— The following report aims to raise significant scientific questions based on a dataset of books from Goodreads, which contains information about the title of books, authors, ratings, year of publication and other parameters. The raised questions will be answered by using the following statistical modules in python – NumPy, Pandas and Matplotlib. All the statements will be backed by evidence using figures, tables and plots.**

I. Overview of the Dataset

The Goodreads Books dataset contains well-compiled information about the 10,000 most popular books, including the title of the books, authors, ratings, and other statistical parameters. The Polish writer and machine learning expert Zygmunt Zając was instrumental in assembling the dataset. The dataset is available in the following GitHub repository: https://github.com/zygmuntz/goodbooks-10k.

The dataset consists of the following five CSV files: book_tags, books, ratings, tags and to_read. "books.csv" contains information about the books, including their IDs, titles, authors, publication year, and ratings. "book_tags.csv" contains information about the tag IDs assigned to each book by users. "tags.csv" contains the mapping between the tag ID mentioned in the book_tags file and the tag name itself. "ratings.csv" contains information about the ratings assigned to each book by users. "to_read.csv" stores the IDs of the books marked "to read" by the users as (user_id, book_id) pairs.

II. Scientific Questions/Hypotheses

A. *Does the rating of a book have a significant impact on the number of readers adding it to their to-read basket?*

B. *Do books with extreme ratings (high and low) have a considerably lower number of ratings in comparison to the other average ratings?*

C. *What is the trend in the count of books and the number of authors?*

D. *Has there been any change in the average title length of books over the years?*

E. *Have the quality of books increased or decreased over the years? (based on mode ratings)*

III. Details of Libraries and Functions

The following libraries will be used to analyse the given dataset and answer the above questions:

A. Pandas

Pandas provides powerful data structures such as Series and DataFrame that allow for the manipulation and analysis of structured data. It has a variety of functions for reshaping, grouping, merging and plotting data that increase the efficiency and performance of data analysis tasks.

pandas.read_csv(): Reads data from a CSV file and creates a Pandas DataFrame for easy manipulation.

pandas.DataFrame.group_by(): Groups data based on one or more columns and allows for applying aggregate functions to each group.

pandas.merge(): Merges two or more Pandas DataFrames based on common columns.

B. NumPy

Numpy is a library used for numerical computing in Python. It provides a powerful array object that enables one to perform mathematical operations on large datasets. Numpy has a wide range of mathematical functions for operations such as linear algebra, Fourier transforms, and random number generation.

numpy.unique(): Finds unique values in an array and returns the count of each unique value.

numpy.polyfit(): Fits a polynomial curve to a set of data points in Numpy and return the polynomial coefficients.

C. Matplotlib

Matplotlib is a plotting library in Python that provides various functions for creating visualizations such as line charts, scatter plots, bar graphs and histograms.

matplotlib.pyplot.scatter(): Creates a scatter plot in Matplotlib with customizable parameters.

IV. Answers to the Questions (with Appropriate Illustrations)

A. Question 1: To analyse the likelihood of a reader adding a book of a given rating to their to-read basket.

Solution: The average ratings of all the books with the respective book IDs can be obtained from the books.csv file. Using the to_read.csv file, the number of readers who are interested in reading a particular can be obtained can be grouping together records with the same book ID. A scatter plot of the number of interested readers against the average rating of a book can be plotted using matplotlib.

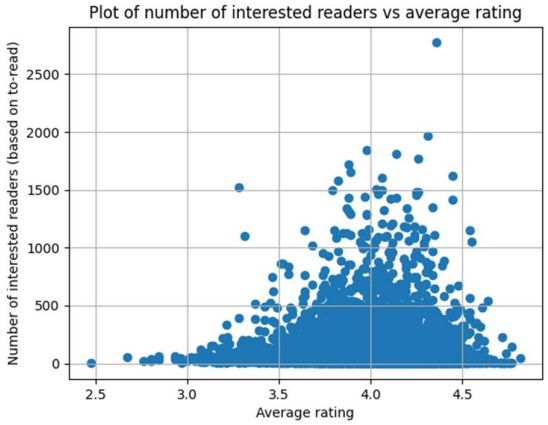

Figure 1: Number of Interested Readers vs Average Rating

B. Question 2: Do books with extreme ratings (high and low) have a considerably lower number of ratings in comparison to other ratings?

Solution: The average ratings and the number of ratings of all the books can be obtained from the books.csv file. A scatter plot of the number of ratings against the average rating of a book can be plotted using matplotlib.

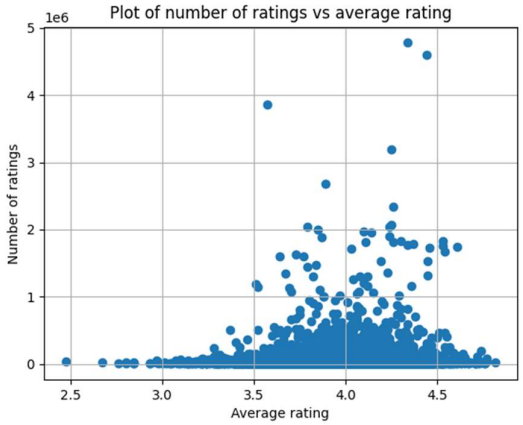

Figure 2: Number of Ratings vs Average Rating

C. Question 3: What is the trend in the count of books and the number of authors?

Solution: The authors can be obtained from the books.csv file. Using some python functions, the number of authors for each book can be obtained. Furthermore, a NumPy function which returns the unique values (number of authors) and the count of each value (number of books) can then be displayed as a table using a pandas.DataFrame object.

	Number of Authors	Number of Books
0	1	7921
1	2	1544
2	3	346
3	4	103
4	5	33
5	6	19
6	7	2
7	8	5
8	9	3
9	10	3
10	11	3
11	12	3
12	14	3
13	15	4
14	17	2
15	19	1
16	21	1
17	22	1
18	24	1
19	29	1
20	47	1

Figure 3: Correlation between the number of authors and number of books

D. Question 4: Has there been any change in the average title length of books over the years 1850 to 2000?

Solution: The titles of the books can be obtained from the books.csv file. The number of characters in the title of each book can be stored in a new column of a pandas.DataFrame object. The mean value of the title length can then be obtained after grouping together books published in the same year. A bar graph of the average title length can be plotted against the publication year using matplotlib. A trend line can also be displayed to better understand the data.

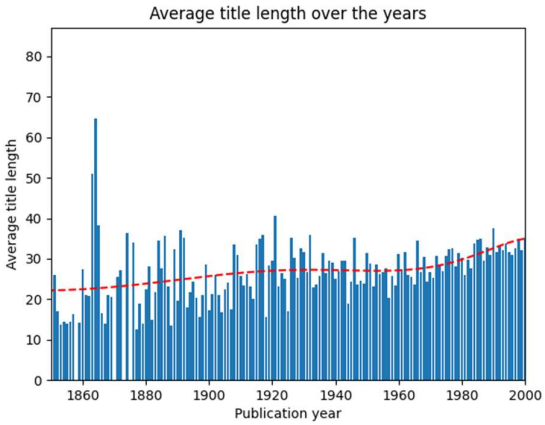

Figure 4: Average title length from 1850 to 2000

E. *Question 5: Have the quality of books increased or decreased over the 19th and 20th centuries? (based on mode ratings)*

Solution: The average ratings for each book can be obtained from the books.csv file and can be grouped together based on the publication year. To find the most common rating, the mode of all the ratings in a given year can be obtained. A line plot of the most common rating of books against a publication year can then be plotted using matplotlib.

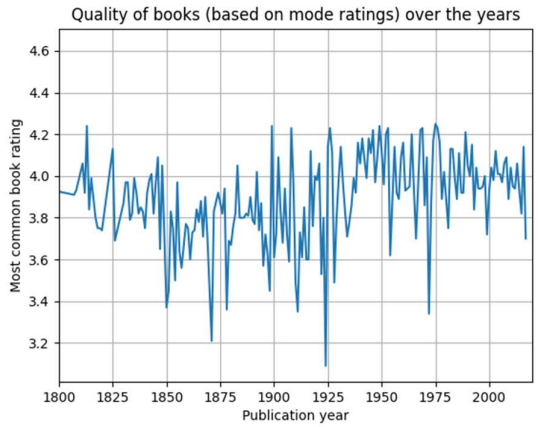

Figure 5: Quality of books over the 19th and 20th centuries

V. Summary of the Observations

A. *Number of Interested Readers vs Average Rating*

The scatter plot shows a high density in the number of readers who are interested in reading a book, with a rating of around 4. The density of the plot, however, does not steadily increase with an increase in rating. Unlike common expectations, readers are not solely interested in books having the maximum rating. Books with ratings of around 4 are the common choice of readers.

B. *Number of Ratings vs Average Rating*

From the scatter plot, it is clear that extremely high, and low-rated books have a significantly lower rating count. In the case of low-rated books, it can be attributed to the lack of interest among readers. But, the high-rated books can have two reasons for the observed trend. Firstly, the ratings might be skewed by some people who might have connections with the author or the publication. Secondly, it is possible that the views of the readers regarding the complete excellence of a book are relative. Many readers might avoid giving a high rating as they believe there is always a scope for improvement.

C. *Number of Authors vs Number of Books*

It can be observed from the table that a vast majority of books have single authors, and a reasonable number of them have two authors. The probability of finding a book authored by 6 or more authors significantly drops. In fact, it is surprising that some books were even authored by more than 20 authors!

D. *Average Title Length over the Years*

The bar graph shows the possible existence of an unsteady increase in title length over the 19th and 20th centuries. The authors might be unknowingly inclined to create longer titles in order to draw the attention of readers.

E. *Quality of Books (based on Most Common Ratings) over the Years*

From the line plot, it is clear that the quality of books has substantially improved from the 19th to the 20th century. The frequency of books with high ratings has gone up. There are two possible explanations for the same. The readers who rate the books are from the current century and might not be interested in the older books. Alternatively, the quality of books might have significantly improved due to a better understanding of readers' interests with greater reach using technology.

Acknowledgment

I would like to thank Professor Shanmuga R for giving me the opportunity to work on the dataset provided and use my data analysis skills to extract valuable information and trends from the source.

References

[1] "API Reference — Pandas 1.5.3 Documentation," n.d. https://pandas.pydata.org/docs/reference/index.html#api.

[2] "NumPy Documentation," n.d. https://numpy.org/doc/.

[3] "API Reference — Matplotlib 3.7.0 Documentation," n.d. https://matplotlib.org/stable/api/index.html#.