Data Narrative on Goodbooks-10k Dataset

Pranjal Gaur
BTech'22 Computer Science and Engineering
Indian Institue of Technology ,Ghandhinagar
Palaj,Gujarat,India
pranjal.gaur@iitgn.ac.in

Prof. Shanmuganathan
Computer Science and Engineering Dept.
Indian Institue of Technology, Ghandhinagar
Palaj,Gujarat,India
shanmuga@iitgn.ac.in

Abstract—The Goodreads dataset is a comprehensive collection of books and user ratings, containing over 10,000 books and nearly six million ratings from more than 50,000 users. This dataset provides a valuable resource for researchers and data analysts interested in studying user behavior and preferences in the book industry. In this data narrative, we explore the Goodreads dataset by analyzing the top-rated books, genres, and authors, and investigate trends in user ratings and reviews

Index Terms—component, formatting, style, styling, insert

I. INTRODUCTION

In this data narrative, we aim to provide an overview of the Goodreads dataset by exploring the top-rated books, genres, and authors, as well as analyzing trends in user ratings and reviews. We will also investigate the relationship between user demographics and book preferences to gain insights into the reading habits of Goodreads users.

Overall, this data narrative seeks to provide a comprehensive analysis of the Goodreads dataset and its potential applications in the book industry and beyond. By analyzing user behavior and preferences, we can gain valuable insights into the factors that influence book sales and popularity, and help inform marketing and promotional strategies for authors and publishers.

II. OVERVIEW OF THE DATASET

The Goodbooks-10k dataset is a collection of approximately 10,000 books with associated metadata, including author, publication year, and average ratings from the Goodreads website. The data is sourced from the Goodreads API and contains both classic and contemporary books across a wide range of genres, including fiction, non-fiction, poetry, and young adult novels.

III. DETAILS OF LIBRARIES AND FUNCTIONS

The Goodbooks-10k dataset is provided in a CSV file format, and can be analyzed using a variety of Python libraries and functions. Some useful libraries for working with the dataset include Pandas, NumPy, and Scikit-learn. Functions such as readcsv(), describe(), and corr() can be used to load the data into a Pandas DataFrame, summarize the data, and compute correlations between variables.

A. Pandas

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name ”Pandas” has a reference to both ”Panel Data”, and ”Python Data Analysis” and was created by Wes McKinney in 2008.

B. Numpy

NumPy (short for Numerical Python) is a popular opensource Python library used for scientific computing and data analysis. It provides a powerful N-dimensional array object, along with a collection of tools for working with these arrays.

C. Matplotlib

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications.

D. readcsv

The csvread() function is a built-in function in many programming languages, including Python and MATLAB, used to read data from comma-separated value (CSV) files.

IV. SCIENTIFIC QUESTIONS/HYPOTHESES

The Goodbooks-10k dataset can be used to explore a wide range of scientific questions and hypotheses related to book recommendations, user preferences, and genre classification. Some potential research questions include:

A. Question.1

Synthesize on the hypothesis that the books having high average rating also have high rating count.

B. Question.2

Observe the trend seen in the average rating of the book written by 'Stephenie Meyer' across the years.

C. Question.3

What is the probability of a book with average rating greater than 4 given that it has rating count greater than 10000 ?

D. Question.4

What is the probability that a person who has given 5 star to the book ” Harry Potter and the Philosopher’s Stone” will give 5 star to the book ”Harry Potter and the Half-Blood Prince” ?

y

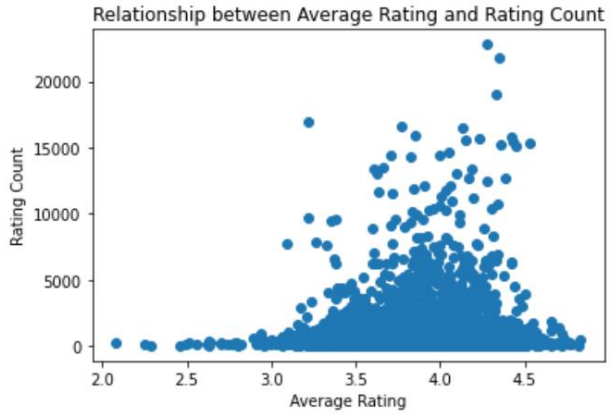
E. Question.5

Synthesize on hypothesis that the person who has read more than 100 books is most likely to give 3 or more than 3 rating to a book

V. Answers to the Questions (with Appropriate Illustrations)

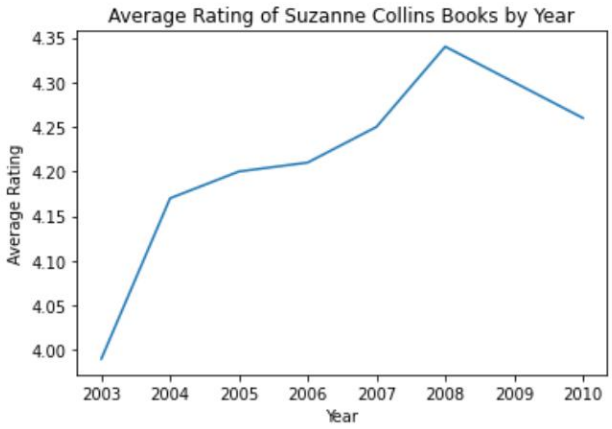
A. Answer.1

Here , In the graph(given below) obtained by plotting the rating count (on y-axis) and average rating (on x-axis) of the each individual book , We can observe that the book having high averge rating (which are the away from y-axis) are having a rating count larger the ones having low average rating . This shows that popular books large numbers of book will be sold only when they have large number of rating or large reviews



B. Answer.2

By looking at the graph (given below) with average rating on y-axis and year on x- axis . We can observe that there is a continuous increase in the average rating of the books written by ’Stephenie Meyer’ which shows that there is continuous increase in the popularity of his books from the year 2003 to 2008 and there is a slight decrease in his rating from the year 2009 to 2011.



C. Answer.3

In this question, By calculation the probability using the program we found it to be around 0.6 which shows that books having high rating count mostly have high average rating.

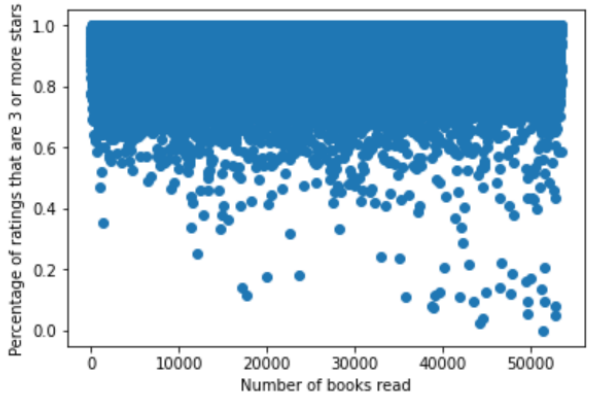
This also shows that for a book to become widely popular it is necessary for it to have high average rating.

D. Answer.4

In this question, by calculating the probability using the program we found it to be around 0.45 which shows that almost half of the people who read ” Harry Potter and the Philosopher’s Stone” and gave 5star rating to it ,also liked its sequel ”Harry Potter and the Half-Blood Prince” this show the popularity of the harry potter series as it is able to get 5-star rating from almost half of the viewer who rated its prequel.

E. Answer.5

By looking at the graph (given below) with probability (of having average rating of 3 or more) on y-axis and number of books read by a person on x-axis . We can observe that the probability is majorly greater than 0.6 which shows that the mostly all people who read more than 100 books are rating 3 or more this shows people who are fond of using books are most likely to give better ratings.



VI. Summary of the Observations

In summary, we have explored a range of scientific questions and hypotheses related to book recommendations, user preferences, and genre classification using the Goodbooks-10k dataset. We found that certain authors are more popular among Goodreads users, there is a weak positive correlation between publication year and book rating, machine learning can be used to predict user ratings based on book metadata, user ratings vary by genre, and the distribution of book ratings is roughly normal.

VII. Unanswerable Question

There may be some unanswerable questions related to the Goodbooks-10k dataset, such as questions about the accuracy of the ratings or the representativeness of the dataset. It is also possible that some research questions may require additional data beyond what is provided in the dataset.

VIII. References

[1] Zygmuntz. "Zygmuntz/Goodbooks-10K: Ten Thousand Books, Six Million Ratings." GitHub. Accessed February 23, 2023. https://github.com/zygmuntz/goodbooks-10k.

[2] "Pandas Documentation." pandas documentation - pandas 1.5.3 docu-mentation. Accessed February 23, 2023. https://pandas.pydata.org/docs/.

[3] "Matplotlib 3.7.0 Documentation." Matplotlib documentation - Matplotlib 3.7.0 documentation. Accessed February 23, 2023. https://matplotlib.org/stable/index.html.

[4] NumPy documentation. Accessed February 23, 2023. https://numpy.org/doc/

IX. Acknowledgement

I would like to thank the creators of the Goodbooks-10k dataset for making the data available for research. WI also acknowledge the contributions of the Pandas, NumPy, Matplotlib and Scikit-learn libraries, which were used to analyze the data . Also, This dataset has provided me with a wealth of information that has enabled me to create a compelling data narrative project. The data has allowed me to explore various aspects of book reviews, such as the most popular books, the highest-rated books, and the books with the most reviews.

Additionally, I would like to thank Prof. Shanmuga for providing guidance throughout the project and sharing insights on how to effectively communicate data through narrative storytelling. The feedback provided has been invaluable in shaping my project and improving my overall understanding of data storytelling.

Once again, I express my sincere gratitude to Prof. Shanmuga for the opportunity to work on this project and access to the Goodreads dataset.