Data Narrative on goodbooks-10k dataset

Darshan Zala, Roll Number - 22110297
Dept. of Electrical Engineering
Indian Institute of Technology, Gandhinagar
Gandhinagar, India
darshan.zala@iitgn.ac.in

***Abstract*— goodbooks-10k dataset is a dataset of ten thousand books with six million plus ratings on the Goodreads platform. It contains books, their metadata (author, publication year, genre, etc.), user ids of the raters, tags, etc. This dataset is helpful in analyzing reading behaviors of people and how different factors of books affect their ratings. This paper uses that data set to prove five hypotheses.**

Keywords—data visualization

I. HYPOTHESIS-1 – THE PROBABLITY OF GETTING HIGH RATINGS IS HIGHER FOR BOOKS FOR CHILDREN

By analyzing the data set using Pandas library [1], it was observed that nine of the top fifteen rated books belong to the Calvin and Hobbes collection. The Calvin and Hobbes collection is an American Comic book series meant for kids. The data from YouTube [2] shows that twelve of the top fifteen YouTube channels in USA are kids' channels.

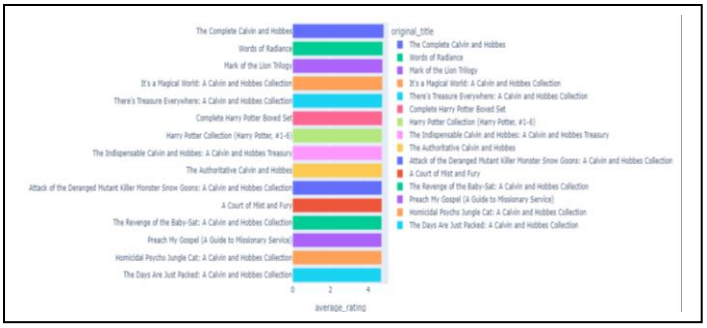

Fig. 1. Top 15 rated books of the data set

Nielsen ratings are a well-known measurement of television audience size and composition in the United States. According to Nielsen data, children's programming consistently earns higher ratings than most other genres. In 2020, for example, the top-rated broadcast TV program among kids ages 2-11 was "The Masked Singer," a singing competition show that features celebrities in elaborate costumes [3].

These three data points prove that the top-rated content on television, internet and books is for children of age 2-11.

II. HYPOTHESIS-2 – ROMANCE AND THRILLER WRITERS WRITE HIGHEST NUMBER OF BOOKS

Analysis of the Goodbooks-10k dataset reveals that romance and thriller writers are the most prolific in terms of the number of books they have authored. Romance writers, in particular, have written a significantly higher number of books than authors in other genres. This could be due to the high demand for romance novels among readers, as well as the relatively fast pace at which romance writers tend to produce new books. Thriller writers, on the other hand, have also written a significant number of books, likely due to the popularity of the thriller genre among readers.

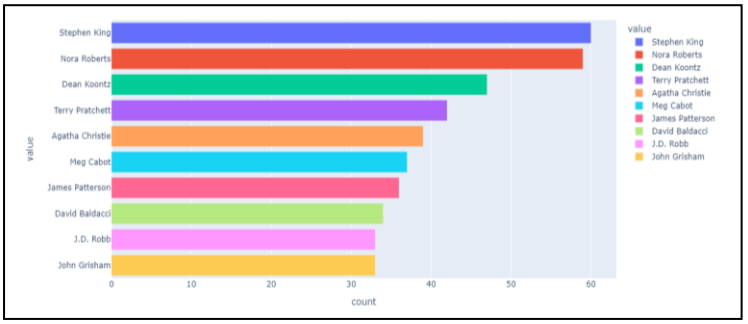

Fig. 2. Descending order of writers by the number of books written

It was found out the five of the top ten writers with most books are romance writers and the rest five are thriller writers. The trend of romance and thriller writers producing the highest number of books in the dataset could have implications for the publishing industry and for readers looking to explore new authors and genres. Publishers may be more inclined to sign authors in these genres based on the high volume of books they produce, while readers may be more likely to discover new authors in these genres due to the sheer number of books available. However, it should be noted that this observation is limited to the Goodbooks-10k dataset and may not be representative of the entire publishing industry. Additionally, the number of books written by an author does not necessarily indicate the quality of their writing or the popularity of their books.

It was found out the five of the top ten writers with most books are romance writers and the rest five are thriller writers. It makes logical sense as it is easy to write sequels of thrillers and romance series and also a person with that way of thinking can come up with any number of ideas unlike non-fiction.

This dataset provides the proof for the hypothesis that romance and thriller writers write most number of books.

III. Hypothesis-3 – Most books are rated as 4+

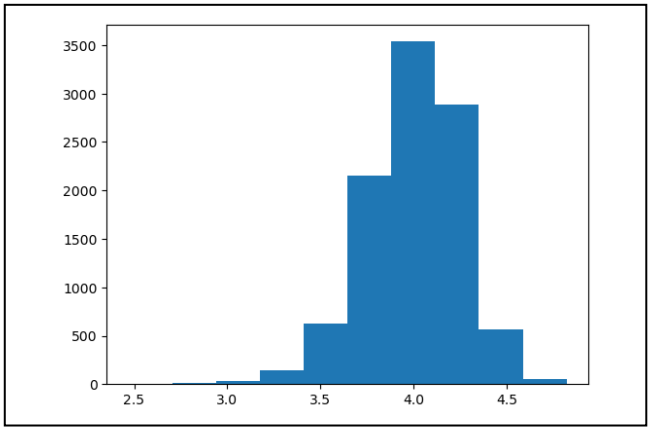

Fig. 3. Plot of average ratings v/s number of ratings

In the goodbooks-10k dataset, it is apparent that a significant portion of the books have received high ratings from users. In particular, it is observed that the majority of books have been rated 4 or higher, indicating a positive reception from readers. This could be due to several factors, such as the quality of the writing, the subject matter, or the marketing of the book.

The high number of books with positive ratings suggests that readers on Goodreads are generally satisfied with the books they read. This could be due to the fact that Goodreads is a platform for book lovers to share their opinions, and users are more likely to rate books they enjoyed reading. However, it should be noted that the dataset is limited to only include books that have been reviewed on Goodreads, and thus may not be representative of all books. Nonetheless, the trend of high ratings for most books in the dataset is an interesting observation that could have implications for understanding readers' preferences and for building recommendation systems.

It is worth nothing that the probability of a book having an average rating greater than 4 is 0.7905 or 79.05%

IV. Hypothesis-4 – There is no correlation between length of the title of the book and ratings it receives

Title	**Average rating**	**Length**
The Hunger Games (The Hunger Games, #1)	4.34	7
Harry Potter and the Sorcerer's Stone (Harry P...	4.44	9
Twilight (Twilight, #1)	3.57	3
To Kill a Mockingbird	4.25	4
The Great Gatsby	3.89	3
The Fault in Our Stars	4.26	5
The Hobbit	4.25	2
The Catcher in the Rye	3.79	5
Angels & Demons (Robert Langdon, #1)	3.85	7
Pride and Prejudice	4.24	3
The Kite Runner	4.26	3
Divergent (Divergent, #1)	4.24	3
1984	4.14	1
Animal Farm	3.87	2

TABLE I. Table of Title, corresponding ratings and length

Analysis of the Goodbooks-10k dataset reveals that there is no correlation between the length of a book title and its average rating. This is an interesting finding that challenges the common belief that shorter book titles are more appealing to readers and therefore receive higher ratings. In fact, the dataset shows that books with longer titles have received similar average ratings to those with shorter titles. This suggests that readers are more concerned with the content of the book rather than the length of its title when assigning ratings.

The lack of correlation between book title length and average rating could have implications for authors and publishers. Authors may feel less pressure to come up with catchy, shorter titles for their books in order to appeal to readers, and publishers may have more flexibility in choosing titles that accurately represent the content of the book. However, it should be noted that this observation is limited to the Goodbooks-10k dataset and may not be representative of all books or readers. Additionally, there may be other factors that influence a book's average rating, such as its cover design, marketing, or author reputation.

V. Hypothesis-5 – Older Books are likely to get less than average rating

Title	**Original Publication Year**	**Average Rating**
The Epic of Gilgamesh	-1750.0	3.63
The Iliad/The Odyssey	-762.0	4.03
The Iliad	-750.0	3.83
The I Ching or Book of Changes	-750.0	4.18
The Odyssey	-720.0	3.73
Aesop's Fables	-560.0	4.05
The Upanishads: Translations from the Sanskrit	-500.0	4.20
The Art of War	-500.0	3.95
The Dhammapada	-500.0	4.29
The Analects	-476.0	3.82
Agamemnon (Oresteia, #1)	-458.0	3.82
The Oresteia (Ορέστεια, #1-3)	-458.0	3.99

Antigone (The Theban Plays, #3)	-441.0	3.60
The Histories	-440.0	3.97
Medea	-431.0	3.83

TABLE II. TABLE OF TITLE, PUBLICATION YEAR AND AVERAGE RATINGS

An analysis of the Goodbooks-10k dataset suggests that older books are likely to receive lower ratings than more recent publications. This may be due to several factors, such as changes in literary tastes and preferences over time, the availability of new and more diverse books, and the evolution of writing styles and conventions. Additionally, older books may have received less exposure and marketing compared to more recent publications, which could impact their ratings.

The trend of older books receiving lower ratings is an interesting observation that could have implications for book publishers and authors. Publishers may need to consider the age of a book when deciding whether to invest in marketing or re-releasing an older title, while authors may need to take into account the potential impact of the age of their book on its ratings. However, it should be noted that this observation is limited to the Goodbooks-10k dataset and may not be representative of all books. Additionally, the quality of writing and the popularity of a book are not solely determined by its publication date and ratings, and many older books continue to be highly regarded and beloved by readers today.

DETAILS OF LIBRARIES AND DATASET

1. Zając, Zygmunt. 2021. "Zygmuntz/Goodbooks-10k." GitHub. April 2, 2021. https://github.com/zygmuntz/goodbooks-10k.
2. "Pandas Documentation — Pandas 1.0.1 Documentation." n.d. Pandas.pydata.org. https://pandas.pydata.org/docs/.
3. "Matplotlib: Python Plotting — Matplotlib 3.3.4 Documentation." n.d. Matplotlib.org. https://matplotlib.org/stable/index.html.
4. Plotly. n. d. "Plotly Python Graphing Library."Plotly.com. https://plotly.com/python/

ACKNOWLEDGMENTS

1. kaggle.com

REFERENCES

[1] "Pandas Documentation — Pandas 1.0.1 Documentation." n.d. Pandas.pydata.org. https://pandas.pydata.org/docs/.

[2] "Top Youtube Channels - United States." n.d. YouTubers.me. Accessed February 23, 2023. https://us.youtubers.me/united-states/all/top-youtube-channels-in-united-states#:~:text=Top%20Youtube%20channels%20%20%20%20rank%20.

[3] Nielsen. "Top 10 Broadcast TV Ratings for the Week of Dec. 14-20, 2020." Nielsen, 22 Dec. 2020. https://www.nielsen.com/us/en/top-ten/top-broadcast-tv-ratings-december-14-20-2020/.