Data Narrative

Hitesh Kumar,
Roll Number - 2110098,
Branch- Computer Science and Engineering,
Professor - Shanmuga Nathan Raman, IITGN.

Aim:

This report aims to explore the Goodbooks-10k dataset, analyse it, and make some questions or develop some hypotheses and find the answers to them or try to prove or disprove them. This is for developing better understanding in analysing data and getting familiar with the python libraries and functions.

Overview:

The dataset 'Goodbooks-10k' provides information about six million ratings on the top ten thousand books, their authors, publishing years, genres and reviews. It also contains the books ID's in to_read list. We have made a data narrative on the dataset of books provided. This report has answered 5 scientific questions using the programming language python and some of its libraries like numpy, pandas and matplotlib with proper illustrations. We have also tried explaining the written code attached in the zip file.

Theory:

1. Python: Python is a high-level, general-purpose programming language. Its design philosophy emphasises code readability with the use of significant indentation. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming.[1]

2. Libraries:

a) Numpy: It is a software library primarily used to make and analyse arrays and plots and perform different operations.

b) Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. [2]

c) Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualisations in Python. Matplotlib makes easy things easy and hard things possible.[3]

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

Questions and their Answers:

Q1: What is the probability that J.K. Rowling's book has rating 4 or more?

Ans: To find the solution to this question, we have written a code which first encounters the average rating of each book and then counts the number of books having an average rating of 4 or more than that. We have also plotted the histogram for these books, as shown below in Fig. 1.

From the output, we can see that the most of the J.K. Rowling's books are highly rated, that is why the probability turns out to be higher, i.e. 0.6547.

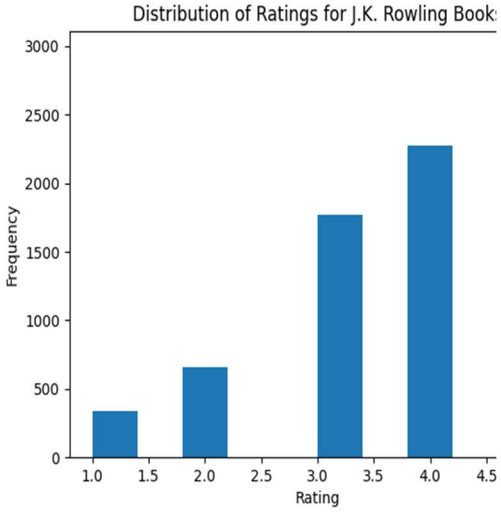

Fig. 1

Q2: Readers tend to read 21st-century-fiction books more than historical-non-fiction books. (Hypothesis)

Ans: (We are taking the assumption that each reader has given ratings for the books he has read) For this question, we have followed the approach that first, we have found the tag_id of the genres given in tag_name in the tag.csv file. We compared it with the book_tags.csv file and found the goodreads_book_id; then we have seen the number of counts of ratings given to books of each genre, i.e. of 21st-century-fiction and historical-non-fiction from the file book_ratings.csv. Moreover, we have also plotted the bar graph, as shown in Fig.2, for the number of ratings for the books of both genres.

From the output, we have found that our hypothesis is somewhere wrong, as the number of ratings given to historical-non-fiction books is far more than the number of ratings given to 21st-century-fiction books. However, we should also keep in mind that it is not necessary that all readers rate the books that they have read.

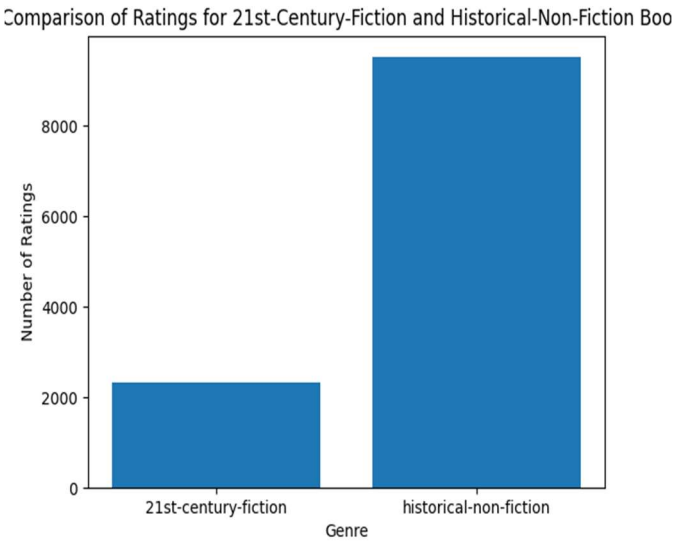

Fig. 2

Q3: Find the book having highest average rating, print its author name then calculate the standard deviation of the average rating of all of his books.

Ans: This is an easier question. For this question, at first we have found the book which have the highest average rating, then we have made a separate series of the author's books who wrote the particular highest rated book. Then we

have calculated the standard deviation of the data using the .std() function of pandas. We have also attached the code snippet for better understanding in Fig.3.

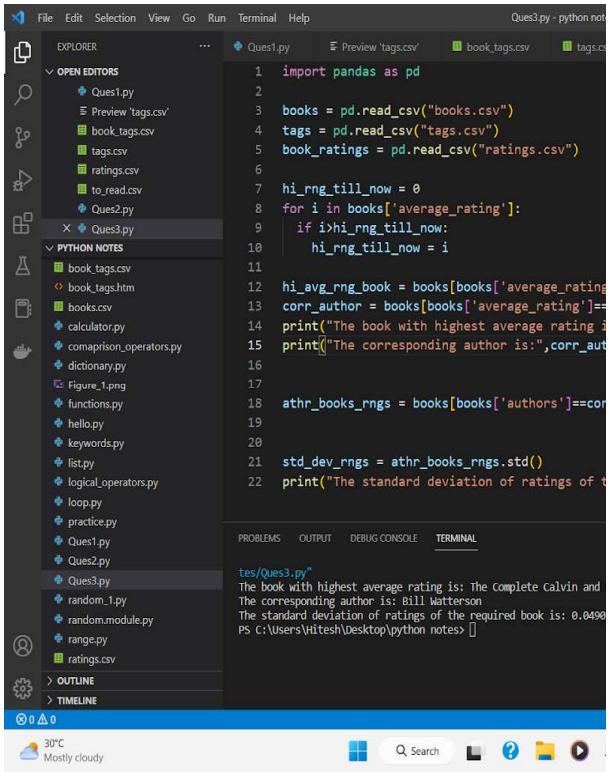

Fig.3

The output of this code is shown in Fig.4.

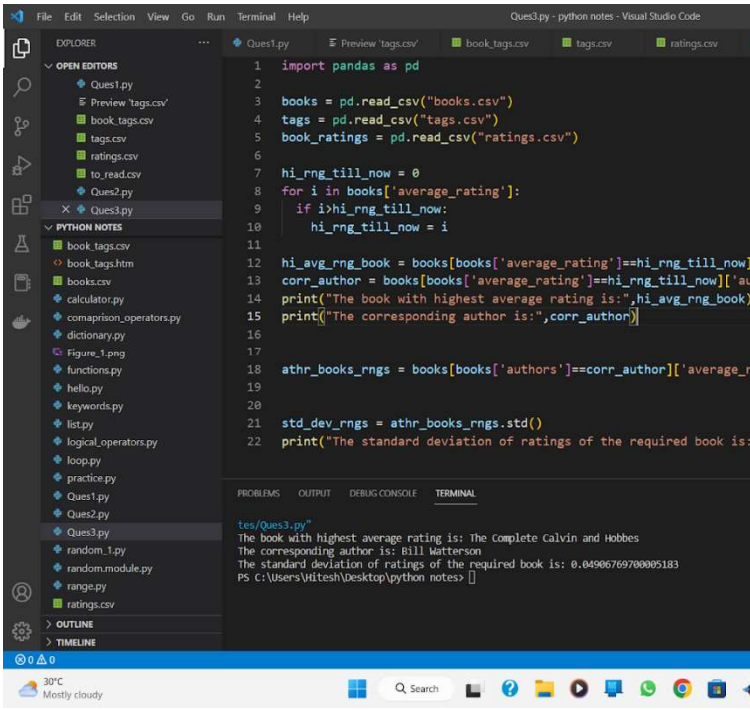

Fig. 4

Q4: The highly rated books are more likely to be popular (i.e. have more ratings) than the books that are less rated. (Hypothesis)

Ans: We have calculated the mean of ratings of every book from the ratings.csv file and also calculated how many readers have given that rating. Then we plotted the scatter plot (as shown in Fig.5) using the calculated data.

According to the output scatter plot, we can conclude that our hypothesis is somewhere correct, and the highly rated books are more popular among the readers since they have rated them in large numbers.

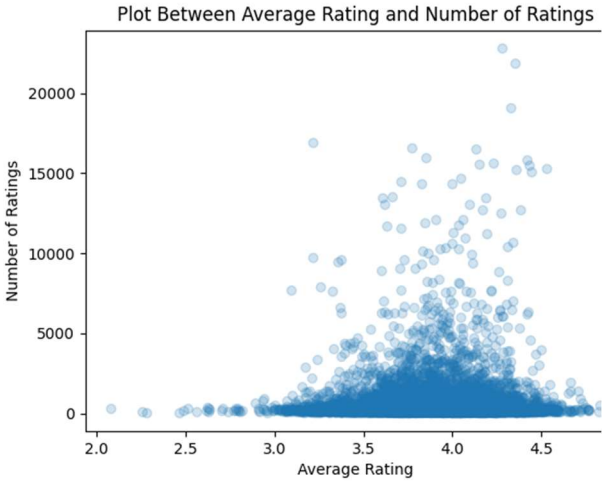

Fig.5

Q5: Recently published books are more popular and read by more readers than those published earlier. (Hypothesis)

Ans: For this question, we have at first changed the format of the publishing year from 2000.0 to 2000; then we made a new column in books dataframe named pb_yr containing the formatted publishing year, and then we merged the 'book_id' column, 'pb_yr' column and book_ratings in book_rng_and_pbyr and plotted the respective graph as shown in Fig.6.

From the output plot, we can conclude that the recently published books are more popular than the other books and there is high peak after about 1950

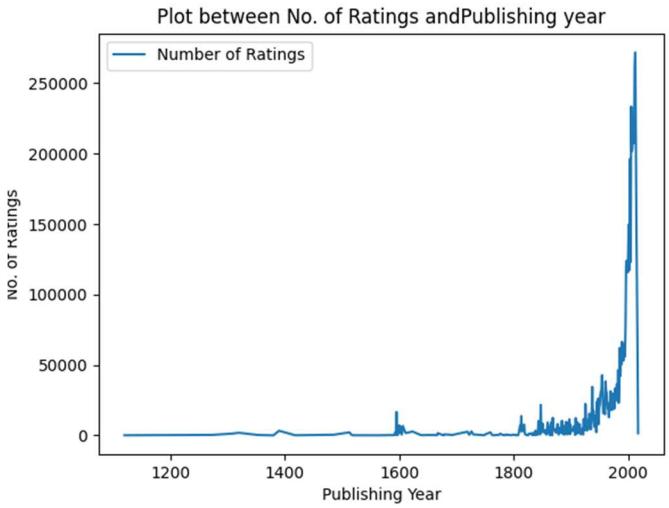

Fig.6

These were the five questions/hypotheses based on the dataset and their answers. We hope that readers will understand these concepts and the approach very well.

Unanswerable Questions:

Some questions cannot be answered from the data given in the dataset provided, such as-

- Questions related to the pages in a book.
- Questions related to the books read by most of the readers.
- Questions related to the genre are not so easily answered because tags.csv file contains the genre name in a different format, like "03-informational", which contains some more useless characters. Not only this, the non-needed characters are different characters of different length in each genre name which

makes it even more difficult to use a code in it.

Acknowledgements:

I want to thank my teachers and batchmates who helped me a lot in understanding the question and actually what do we have to write in a Data Narrative in a better way, how to cite in Chicago style.

[1] Wikipedia. 2023. "Python (programming language)." Wikimedia Foundation. Last modified February 21, 2023. https://en.wikipedia.org/wiki/Python_(programming_language)

[2] Wikipedia. 2023. "Pandas (software)." Wikimedia Foundation. Last modified February 5, 2023. https://en.wikipedia.org/wiki/Pandas_(software)

[3] Hunter, J.D.,"Matplotlib."matplotlib, Feb 13, 2023. https://matplotlib.org/