ES114 Probability, Statistics and Data Visualization

Data Narrative

Astitva Aryan, Mechanical Engineering,
B. Tech'22, Roll Number-22110041,
Indian Institute of Technology
Gandhinagar, astitva.aryan@iitgn.ac.in

***Abstract*— This report is a Data Narrative of the Dataset. It significantly contains scientific questions/ hypotheses on the Dataset, graph and answers of these questions and summary of the observations on the graph.**

I. Overview of the Dataset

The Dataset contains 5 .csv files of Dataframes and I have used books.csv file for this report. It contains book id, user id, goodreads book id, best book id, work id, books count, isbn, isbn13, author of books, publication year, language code, titles of books, average ratings, ratings count, text reviews count and images URL. It has 10,000 samples of Data with 10,000 rows and 23 columns.

II. Scientific Questions/Hypotheses

Question 1: What is the average rating of high rated books by year?
Question 2: What is the Probability Mass Function (PMF) of average rating of books in 21st century?
Question 3: What is the number of books by language?
Question 4: What are the top 10 authors with most number of books of high average rating (>4)?
Question 5: What are the top 10 most rated books?

III. Details of Libraries and Functions

Libraries used:

1) Pandas library is used to analyse Dataset.

2) Matplotlib library is used to create visual representations for the Dataset like bar graphs, pie charts, etc.

Functions used:

1) .groupby() function is used to separate identical data into groups for further analysis.

2) plt.scatter() function is used to draw a scatter plot.

3) plt.xlabel() function is used to set the label for x-axis.

4) plt.ylabel() function is used to set the label for y-axis.

5) plt.title() function is used to give the title for the graph.

6) plt.xticks() function is used to set the labels of the x-axis.

7) plt.show() function is used to show the plot.

8) plt.stackplot() function is used to plot several Datasets on top of one another rather than overlapping with one another.

9) value_counts() function returns a Series containing counts of unique values.

10) sort_index() function is used to sort Series by index labels.

11) sum() function is used to add all values.

12) head() function is used to pick top values of the column.

13) plt.bar() function is used to draw the bar plot.

14) count() function is used to count the number of non-NA/null observations across the given axis.

15) reset_index() function is used to reset the index back to the default.

16) sort_values() function is used to sort by the values along either axis.

17) set_index() function is used to set the Dataframe index.

18) plt.stem() function is used to draw the stem plot.

IV. Answers to the Questions

1.

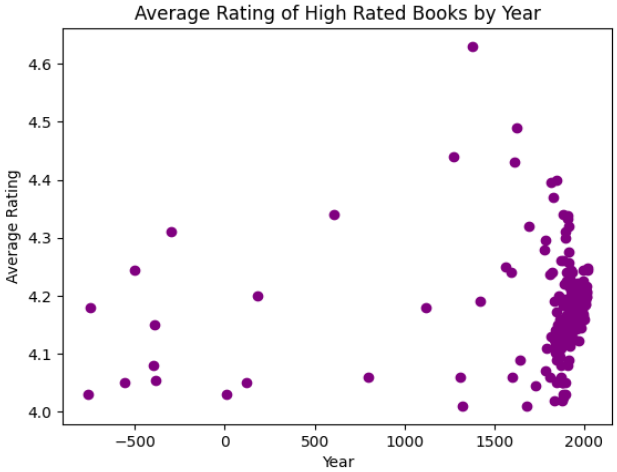

Fig 1 – Between 1000 BCE and 2022 AD
*Negative sign in years mean years are in BCE.

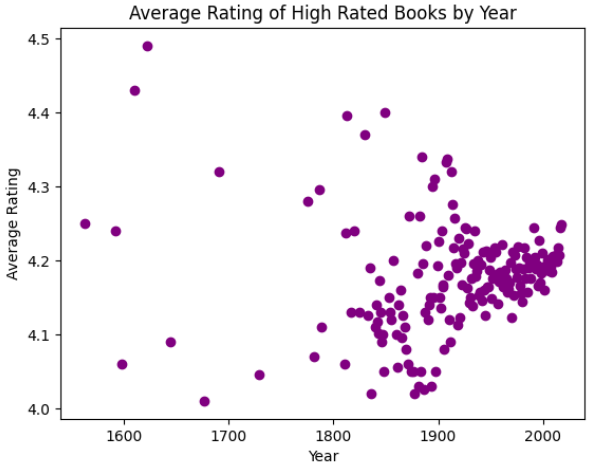

Fig 2 - Zoomed between 1500 and 2022 AD

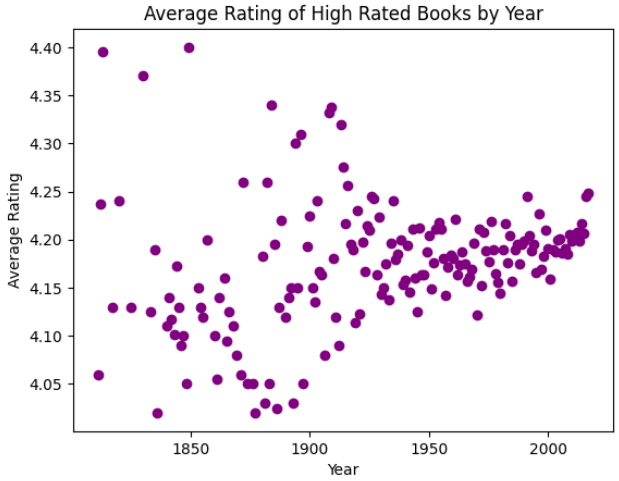

Fig 3 – Zoomed between 1800 and 2022 AD

The above figures show the average rating of high rated books by year. We can see from the Fig 1 that most of the high rated (>4) books are published in recent centuries and the scatter plot of Fig 1 does not helps us to identify the plots of recent centuries because of congestion. So, we stretched out the graph in Fig 2 and Fig 3 to clearly see the plot of average rating vs year for high rated books of recent centuries.

2.

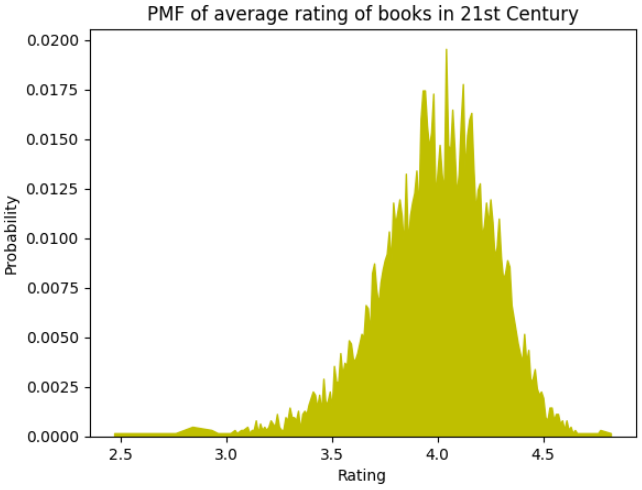

Fig 4

Fig 4 shows the PMF of average rating of books in 21st century. We can observe from the graph that it looks like a Beta distribution curve. We can also see the highest probability in the graph < 0.02 and the average rating of the books in 21st century is between 2.4 and 5.

3.

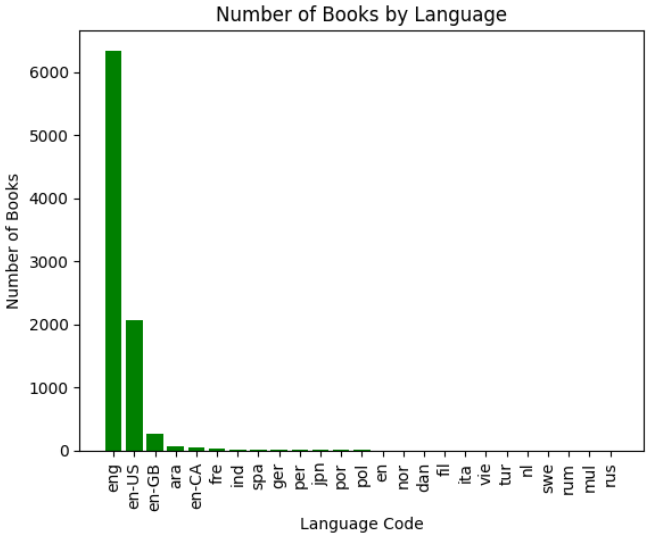

Fig 5

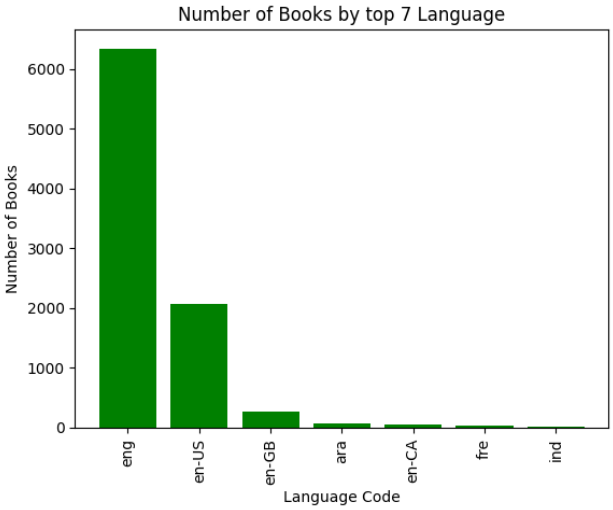

Fig 6 – Bar graph of top 7 language

The above figures show the number of books by language. In Fig 5 we can see that except for top 3 languages, graph for other languages is not either clearly visible or is very small for the human eye. After seeing Fig 6 we can say that most of the books are published in different dialects of English, Arabic, French and Indonesian.

4.

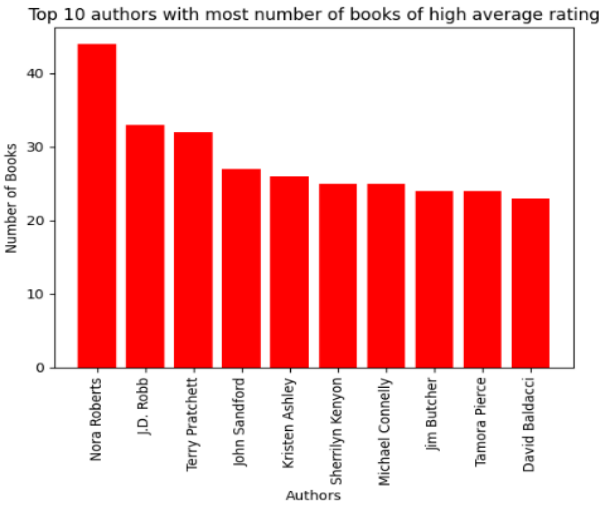

Fig 7

The above graph shows the top 10 authors with most numbers of books of high average rating. It also tells that these are the most famous authors whose work is loved by majority of the people. We can also the author with the highest number of books of high average rating is Nora Roberts.

5.

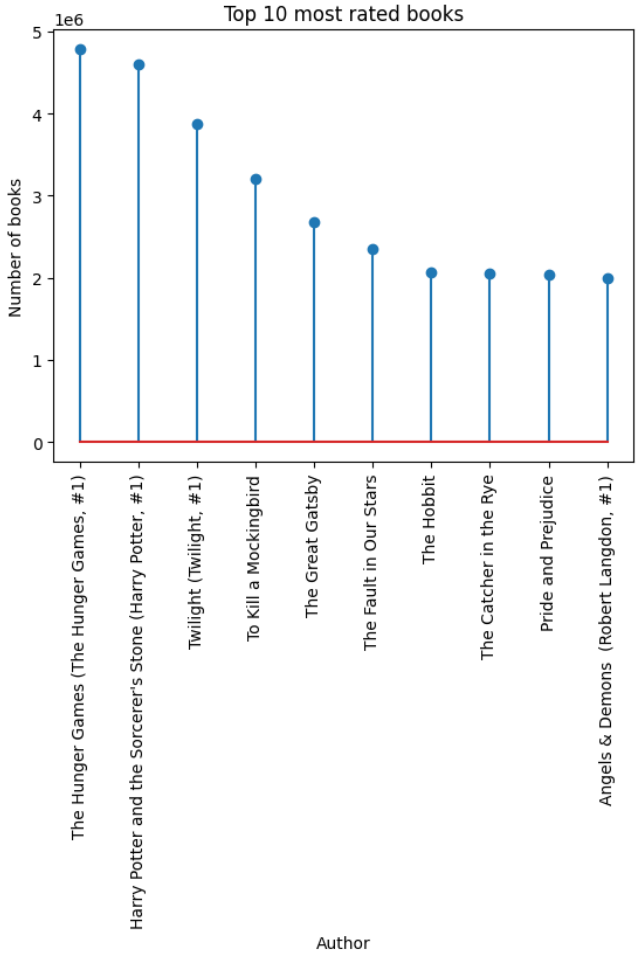

Fig 8

*#1 means the duplicate version of the original book

The above figure shows the top 10 most rated books. We can analyze that this doesn't necessarily means the book is good, it may be bad. Also, after observing this graph we can say that the highest rated book is The Hunger Games.

V. Summary of the Observations

We summarize that the first figure shows the average rating of high-rated books by year, and it indicates that most high-rated books are published in recent centuries. The scatter plot in Figure 1 is congested, so Figures 2 and 3 are stretched out to clearly see the plot of average rating vs year for high-rated books of recent centuries.

Figure 4 shows the probability mass function (PMF) of the average rating of books in the 21st century. The graph looks like a Beta distribution curve, with the highest probability being less than 0.02. The average rating of books in the 21st century is between 2.4 and 5.

Figures 5 and 6 show the number of books by language. Most books are published in different dialects of English, Arabic, French, and Indonesian. The graph for other languages is either not clearly visible or very small for the human eye.

The top 10 authors with the most number of books of high average rating are the famous authors whose work is loved majority of the people.

Finally, the top 10 most rated books are shown in the last figure. It is important to note that high ratings do not necessarily mean a book is good, as it may be subject to biases or preferences.

We can also say that we can plot different types of graphs for different conditions to visualize the Data conveniently.

VI. Unanswerable Questions

Question 1: Why there were less high rated books before 1500 AD?

Question 2: Why majority of the books are in different dialects of English only?

VII. References

1) 'Pandas documentation,' accessed 23 Jan, 2023, https://pandas.pydata.org/docs/

2) 'Matplotlib documentation,' accessed 23 Jan, 2023, https://matplotlib.org/stable/index.html

3) 'goodbooks-10k,' accessed 23 Jan, 2023, https://github.com/zygmuntz/goodbooks-10k

4) 'GeeksforGeeks,' accessed 23 Jan, 2023, https://www.geeksforgeeks.org/

VIII. Acknowledgements

I would like to express my sincere gratitude to Professor Shanmuganathan Raman for their valuable guidance. Their expert advice and feedback helped me in developing my skills and improving the quality of this report.