Data Narrative Part I

Om Gupta
Chemical Engineering
IIT Gandhinagar
Gandhinagar,India
om.gupta@iitgn.ac.in

I. OVERVIEW OF DATASET

The dataset is a collection of ratings for 10,000 books by 53,424 users. The ratings are on a scale of 1 to 5 and were obtained from the Goodreads website. The dataset is available in CSV format and can be downloaded from the GitHub repository you provided.

The dataset is commonly used for recommendation system research and evaluation.

II. DETAILS OF LIBARIES AND FUNCTIONS

A. *PANDAS*

Pandas is a popular python library used for the data manipulation and analysis. It is a open source library. It provides easy-to-use data structures and data analysis tools for working with structured data. Some of the key features are-

DataFrame : A two-dimensional table-like data structure. It allows you to store and manipulate data with rows and columns, and provides a variety of functions for data filtering, selection, aggregation, and transformation.

Series: A one-dimensional array-like object that can hold any data type, including integers, floats, and strings. It is similar to a column in a DataFrame.

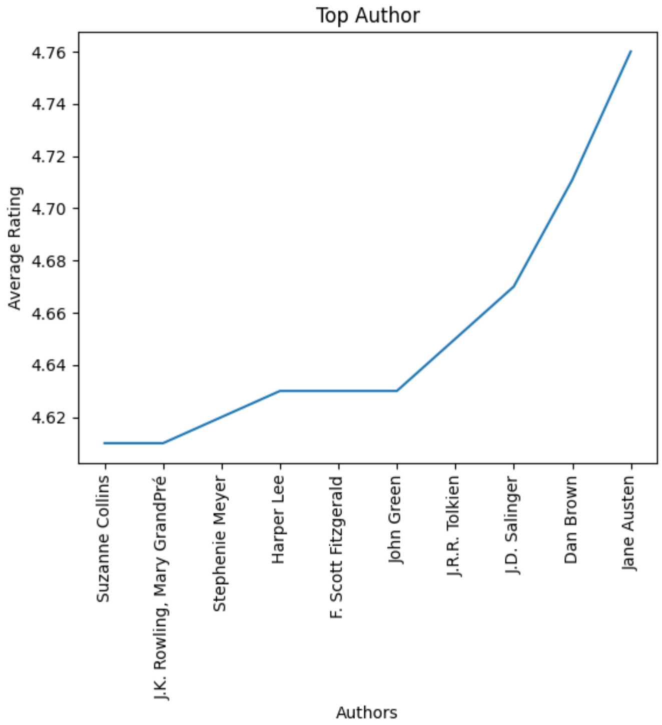
B. *MATPLOTLIB*

Matplotlib is a popular open-source data visualization library for Python. It provides a wide range of customizable 2D and 3D plots, including line plots, scatter plots, bar plots, histograms, and many others.

III. SCIENTIFIC QUESTIONS/HYPOTHESIS

A. Finding the top Authors among the given dataset considering all their books ratings.

Every author had published many books and ratings of books varries unpredictably, therefore we can't judge the authors by their one book. So, averaging the ratings given by the users to all their written books helps to get a proper idea.

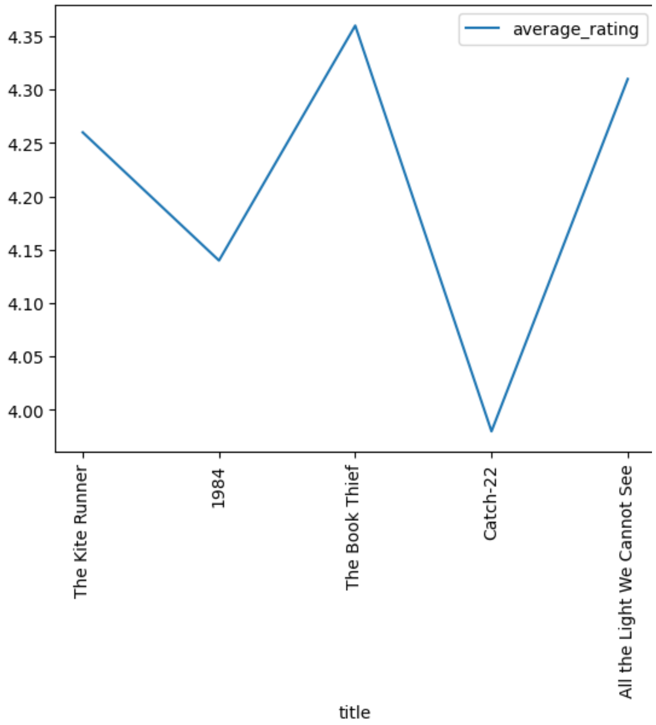


Observation- Author Jane Austen has the highest rating among many authors and could be called as an best author.
Unanswerable Question- The plot doesn't consider the no. of books particular author had written, if it is considered then observation could be changed.

Identify applicable funding agency here. If none, delete this text box.

B. Calculating the rating of the books of most readed books by users.

The most readed books are the most popular books but some of them are not good rated.



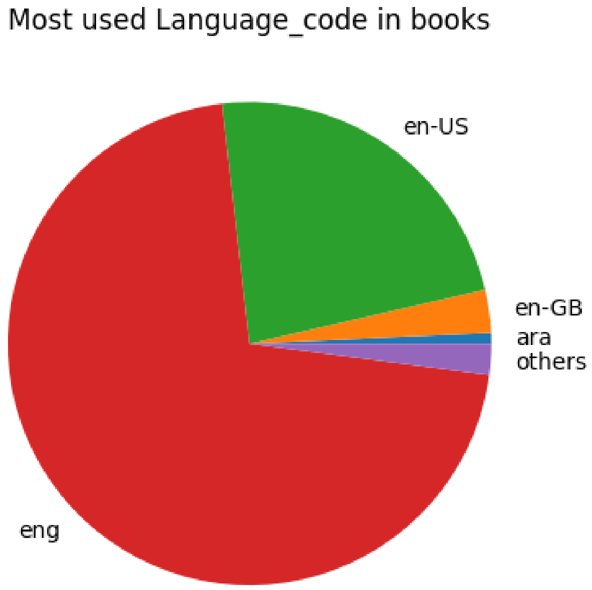
Observation- Book among the top 5 most popular book has rating less than 4.

C. Finding out the probability that in a particular year how many books are among the list of top 100 rated books.

Favourable outcome = No. of books published in that year in the top 100 rated books
Total no. of outcome = Books published in the that year
Probability=Favourable outcome/Total no. of outcome
Example- for year 2000, Probability= 0.019138755980861243

D. Categoriesing the Languages used in writing the books

The languages most used



IV.REFERENCES

1. https://github.com/zygmuntz/goodbooks-10k
2. https://pandas.pydata.org/docs/index.html
3. https://matplotlib.org/stable/index.html