DATA NARRATIVE

Rutvi shah

Material Science Engineering

IIT Gandhinagar

Abstract—

The project "data narrative" helps us to analyze and visualize data using the figures which we can make using python or any other language.in this we have given a data set of six million ratings for ten thousand most popular (with most ratings) books which is to big to be analyze manually so we need to do data narrative and do the hypothesis

I. Overview of the Dataset

The given dataset contains six million ratings for ten thousand most popular (with most ratings) books**.** The dataset contains many csv files like rating.csv, to_read.csv, tags.csv, books.csv and books_tags.csv.

to_read.csv provides IDs of the books marked "to read" by each user, as user_id,book_id pairs, sorted by time.
books.csv has metadata for each book (goodreads IDs, authors, title, average rating, etc.). The metadata have been extracted from goodreads XML files, available in books_xml.
book_tags.csv contains tags/shelves/genres assigned by users to books. Tags in this file are represented by their IDs.
ratings.csv contains ratings sorted by time.
books_tags.csv provide us with the book_id with their book tags i.e.genre of the book.

II. Scientific Questions/Hypotheses

Below are the questions according to which data is being analyzed from the data set given.

.

A. Find the top 10 authors who wrote the highest number of books and show them in the bar graph.

B. Now, find the average rating of the books written by these top 10 authors.

C. plot the graph showing the number of books published in years from

D. calculate the percentage of each rating the book got.and show them in the pie chart along with the percentage.

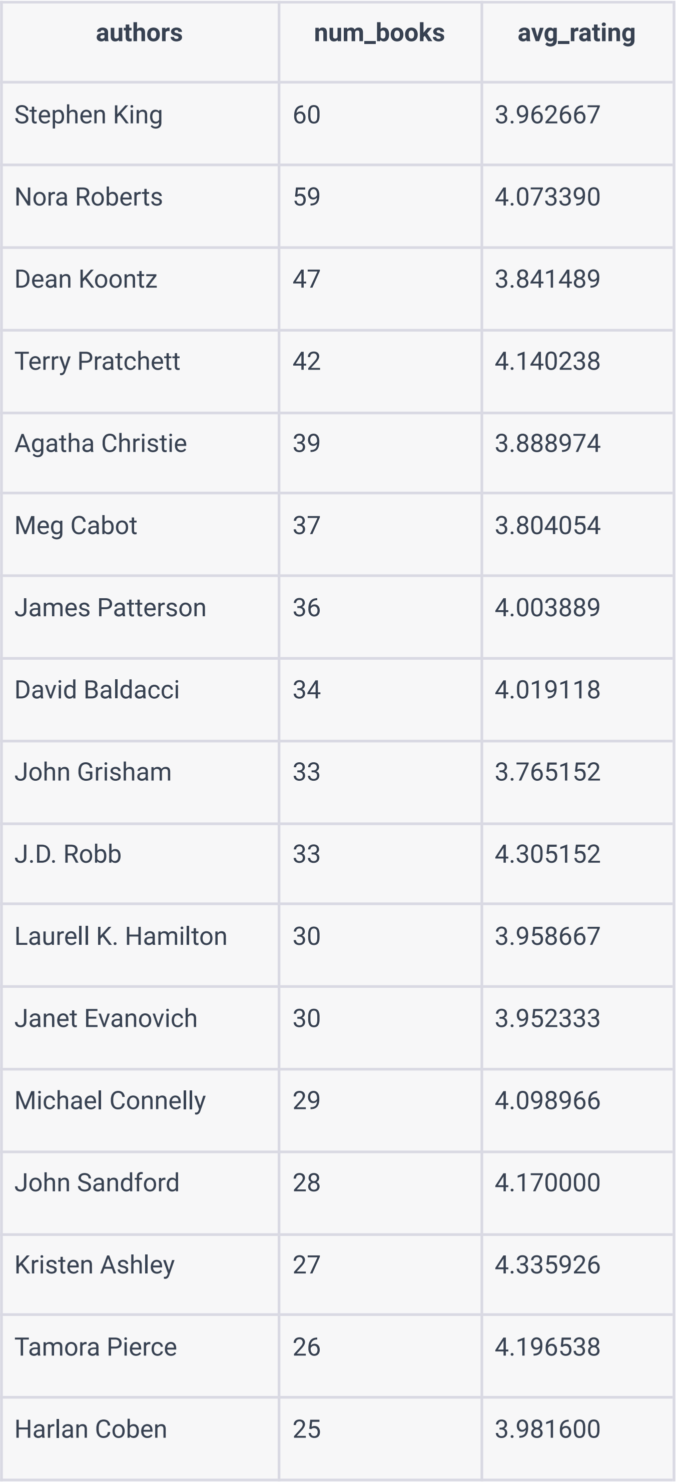
E. find the data representing the top 10 genres in which books are mostly written.

III. Details of Libraries and Functions

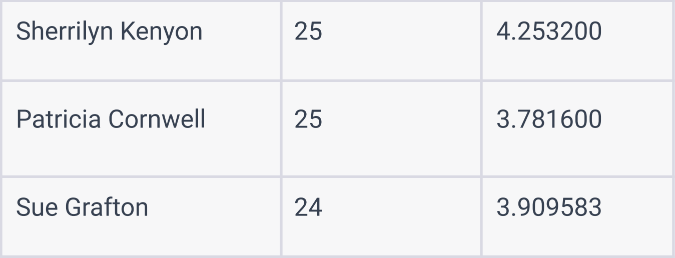
I used libraries like matplotlib, pandas and numpy.

IV. Answers to the Questions (with Appropriate Illustrations)

A.
The data below shows the top 20 authors whose average rating of the books are mostly hgh that is above 4.



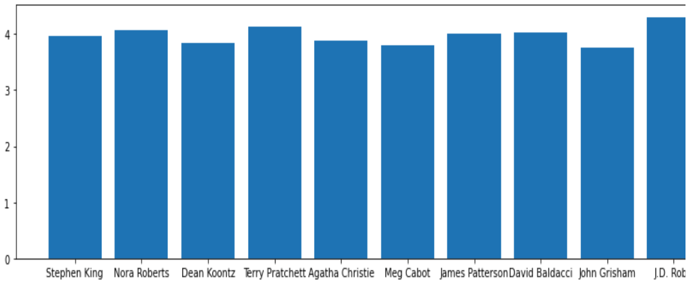
authors	num_books	avg_rating
Stephen King	60	3.962667
Nora Roberts	59	4.073390
Dean Koontz	47	3.841489
Terry Pratchett	42	4.140238
Agatha Christie	39	3.888974
Meg Cabot	37	3.804054
James Patterson	36	4.003889
David Baldacci	34	4.019118
John Grisham	33	3.765152
J.D. Robb	33	4.305152
Laurell K. Hamilton	30	3.958667
Janet Evanovich	30	3.952333
Michael Connelly	29	4.098966
John Sandford	28	4.170000
Kristen Ashley	27	4.335926
Tamora Pierce	26	4.196538
Harlan Coben	25	3.981600



Sherrilyn Kenyon	25	4.253200
Patricia Cornwell	25	3.781600
Sue Grafton	24	3.909583

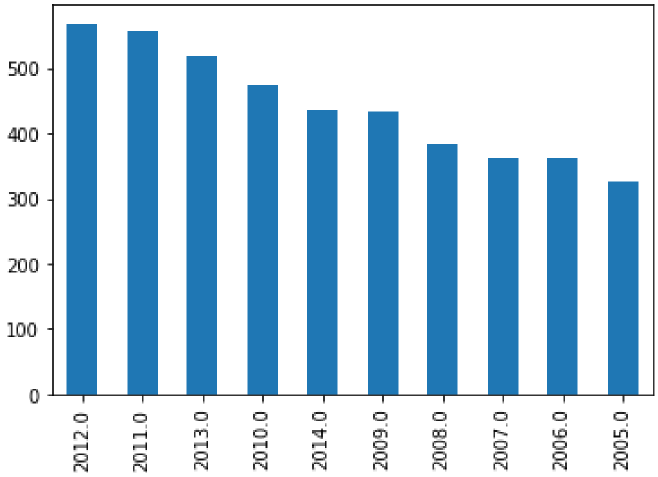
B.

The bar graph below concludes that the authors who wrote the most number of books are averagely highly rated. These hypotheses conclude that the readers changed their opinions of writings after reading many books from the same authors and these results in the high rating of their books.



C.

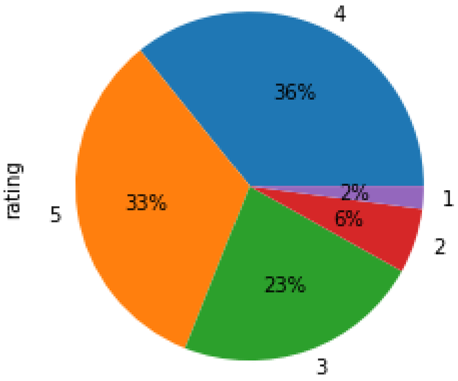
The graph given below showcases that the most number of books are mainly written in 20s .These conclude that the writing has been praised mostly after 20s and that's why most of the books are written after that time.



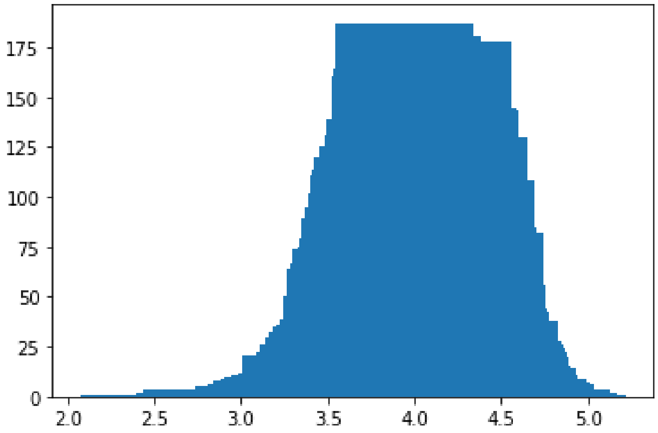
:

D.

The below pie chart represents the percentage of each rating people give to the books.we can observe that most of the people generally rate higher that represents that no matter what authors are being respected for the writings.and mostly there are good authors.

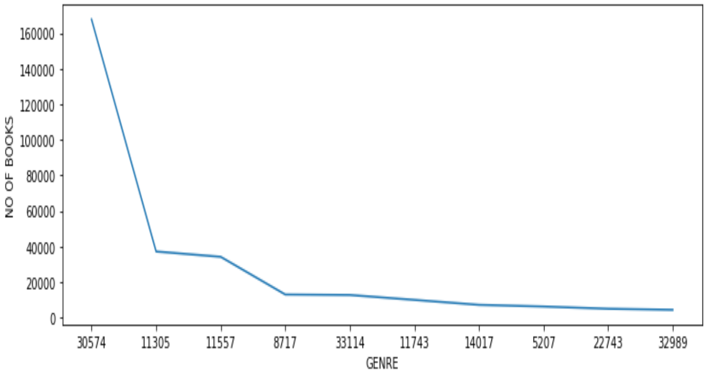


From the bar graph it can be analyzed easily that the most number of ratings lie between 3.5 to 4.5.



E.

The below graph shows us the top 10 genres(book_tags) in which most books are generally written by the top authors. From these analyses we can easily select a good genre to read a good book written by the top authors.



V. Summary of the Observations

The word “data” is plural, not singular. The subscript for the from all the above writings we can observe that most of the people like the genres written by the top authors.and that's why also tend to give higher ratings to the books written by these top authors.

Most of the books are written after the 20s which we can say by analyzing the data and observing that most of the books are written after that only.

VI. Acknowledgements

I sincerely thank Prof.Shanmuga and all the other teaching assistants for guiding me and helping me in this assignment.

This project helps me to know more about data analysis and how to narrate data from dataset.

VII. References

[1]McKinney, Wes, and P. D. Team. "PandasPowerful python data analysis toolkit." Pandas—
Powerful Python Data Analysis Toolkit 1625 (2015]

[2]. Hunt, John, and John Hunt. "Graphing with Matplotlib pyplot." Advanced Guide to Python 3 Programming (2019): 43-65.