Data Narrative Report

Name: Mohit Maurya
Discipline:B-tech M-tech Electrical Engineering

Roll number: 22110145

I. Overview of the Dataset

The dataset contains the rating of ten thousand most popular books. In the given dataset, we have been provided with five useful excel files, namely book_tags, tags, to_read, ratings, and books. These files contain data like the Name of the book, name of the author, the ID of the book, user ID, the publication year of the book, goodbooks id, average rating of the book, image URL, number of the review for a particular book, tag/shelves/genres, etc.

II. Scientific Question / Hypotheses

1) What is the reader's preference, do they prefer recently published books or older ones?
2) What is the average rating range most users give to the books they read?
3) Who are the top 20 authors whose books are reviewed the most?
4) In which language do most of the books are published?
5) In which language do most of the books are published?

III. Details of Library and Functions

To answer the above scientific question, I have used three libraries: Pandas, Matplotlib, and Numpy.

1) The function used in library Pandas are

- pd.read_csv( ):
  The use of this function is to read the excel file, which is in .csv format; and return the dataframe of the information in that excel file
- pd.DataFrame( ):
  This function is used to convert dictionaries, lists, 2-D numpy, or series into dataframes
- df.pivot_table( ):
  This function takes input like an index and a mathematical function like sum,count, mean, etc., and returns a dataframe whose index is the same as the specified index. That mathematical function is applied to the rest of the columns.
- .head( ):
  This function when used without specifying input, give first five rows of the dataframe; when information which is a number(n), is specified, it provides the first n row in the dataframe.
- .sort_values( ):
  This function, when provided with the columns of the dataframe it sorts the columns in ascending order along with the dataframe. If ascending =True is passed in the input, and if False is passed instead of True dataframe is sorted in descending order
- .value_count( ):
  This function returns the series of frequency/count of the unique values.
- .index:
  It gives the index of the dataframe as an array

2) The function in matplotlib are

- Plt.bar( ):
  It plots the bar graph of the two values given as input. Also, it takes other inputs as color to change the color of the bar.
- .plot( ):
- This function also plots the graph of need by passing the name of the graph which is to be plotted like hist refer to histogram or barh for horizontal bar graph. also it takes figsize which takes turple for making graph more bigger and color to color the graph.
- Plt.xtlabel( ):
- Use to label the x axis
- Plt.ylabel( ):
- Use to label y axis

- Plt.xticks( ):
- Use to mark the points on the x axis

3) The Function in Numpy are

- Np.where( ):
- It provide the array of the values which satisfy the condition sepecified in as an input.

4) Some general function used are

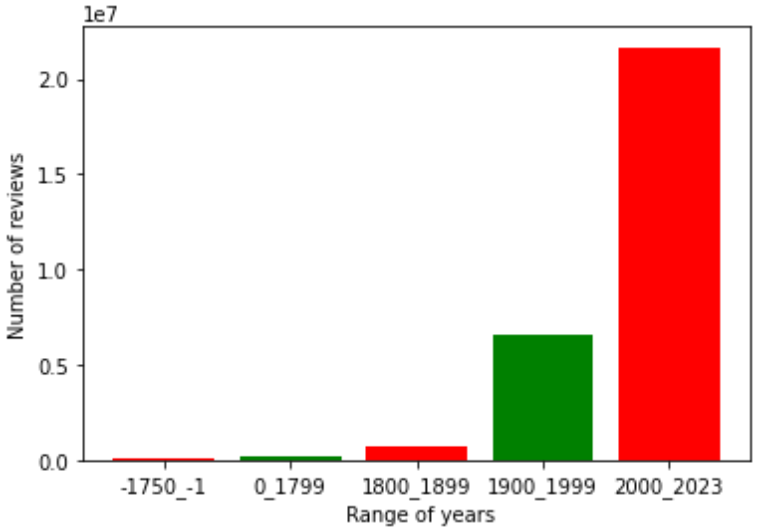
- Sum( ):
- Provide the sum of the input.

IV. Answer to the Questions.

1) What is the reader's preference, do they prefer recently published books or older ones?

I have ploted the graph of total number of reviews on the book published in certain time interval which the time interval here I took the time interval as from 1750 BC to 0 BC ,year 0 to year 1800, year 1800 to 1900, year 1900 to 2000 and year 2000 to 2023.

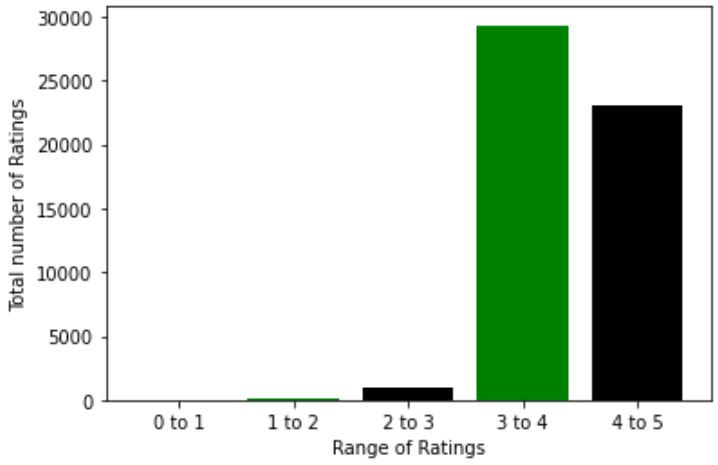
I found that the bar corresponding to year 2000 to 2023 is high when compared with others which means that the number of reviews to book published after the year 2000 is more. This mean that reader generally prefer to read the recent publisher book.



2) What is the average rating range most users give to the books they read?

To answer this question I first seperated the ratings in ranges 0 to 1 , 1 to 2 , 2 to 3, 3 to 4, 4 to 5 and calculated the average rating of ratings given to different books by same user and did this for all user.

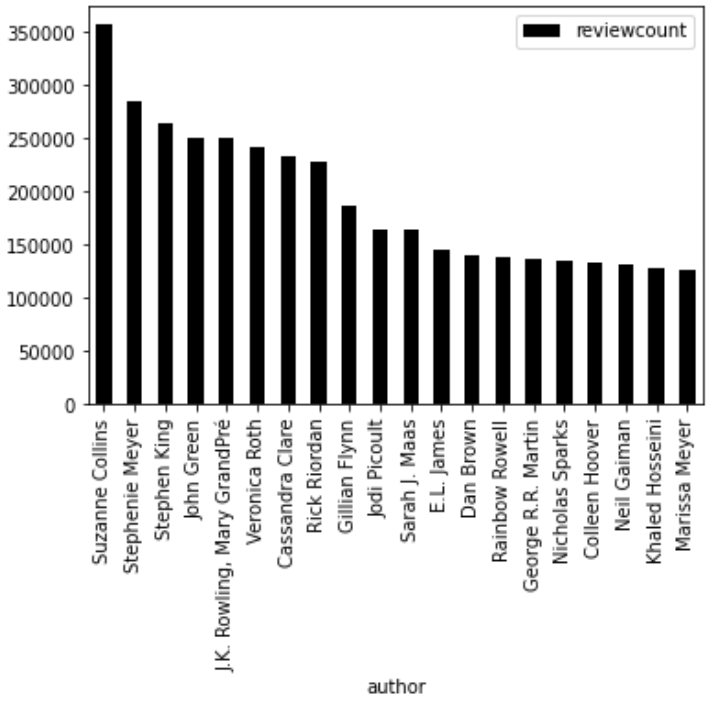
After this I ploted the graph between a range of ratings with their occurrence in the dataframe, and I found that the majority of the reader give a rating to a book in between the range of 3 to 4 and very few reader give rating to the book the in the range 0 to 1,1 to 2 and 2 to 3.



3) Who are the top 20 authors whose books are reviewed the most?

For this question I calculated the total review on all the books of a particular author and the ploted against the name of the author.

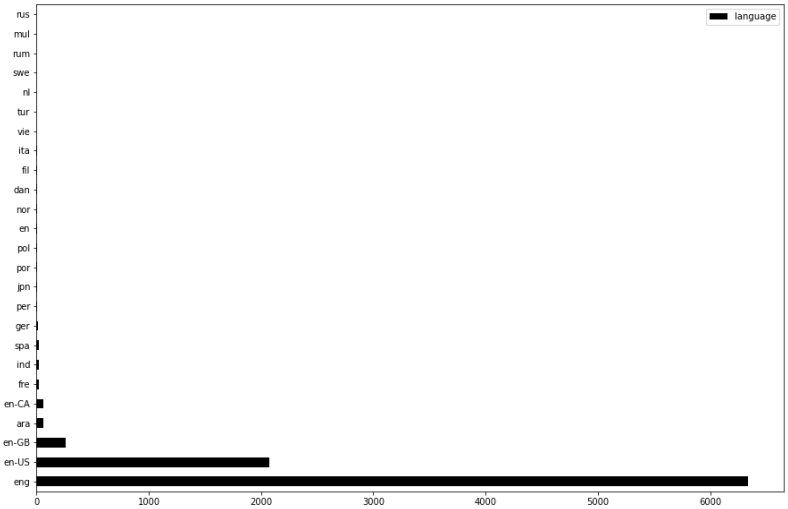
From the graph we can conclude that the author with majority of review is Suzanne collins and the second author on ths list is Stephen kings and so on.



4) In which language do most of the books are published?

In this question I ploted the bar graph between different languages present in the dataset against the number of book published in that language.

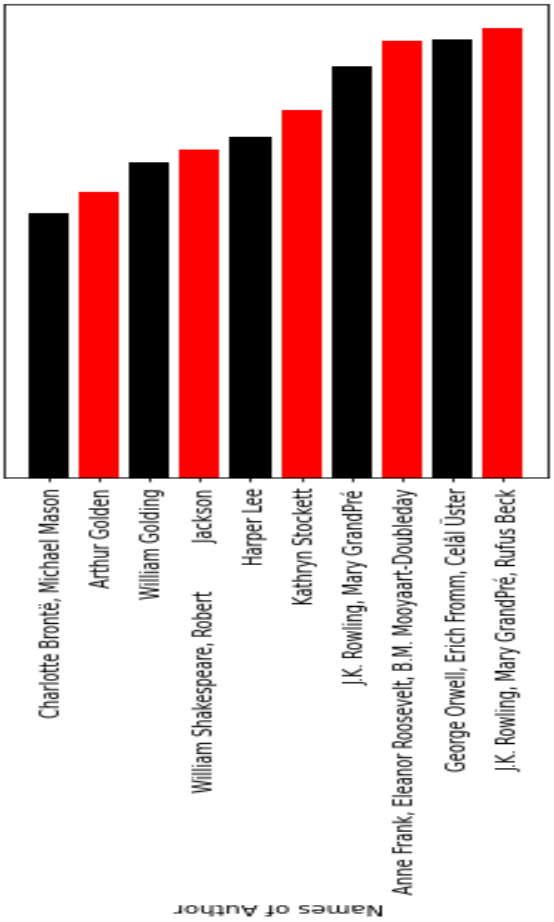
I found that most of the book is published in the English language more than 6000 books among 10000 books present and the second most common language was US English.



5) Who are the top 10 authors which are most succesfull?

For this question I calculate the average rating on the books written by this author and the total rating count on all his books and the multiplied both in order to get the author which have high average rating and high rating count .

I found that J.k Rowling,Mary GrandPre,Rufus Beck have the highest value of product of average rating and high rating count. Thus we can conclude that J.k Rowling,Mary GrandPre,Rufus Beck is among the most successful author followed by George Orwell, Erich Fromm, Celal Uster and many other.



V. SUMMARY OF THE OBSERVATIONS

The summary of the observation is that there are many book with rating ranging from 0 to 5. Majority of the books are published in English language. The majority of reader like to read books which are recently published. There are different authors with most review on their books and different authors with high average rating and rating count on their books. There can be many conclusion drawn via further mining the dataset.

References

1) "Pandas tutorial," GeeksforGeeks, 29-Feb-2020. [Online]. Available: https://www.geeksforgeeks.org/pandas-tutorial/. [Accessed: 23-Feb-2023].

2) "Pandas documentation#," pandas documentation - pandas 1.5.3 documentation. [Online]. Available: https://pandas.pydata.org/docs/. [Accessed: 23-Feb-2023].

3) "numpy doc," NumPy documentation. [Online]. Available: https://numpy.org/doc/. [Accessed: 23-Feb-2023].

4) "Matplotlib 3.7.0 documentation#," Matplotlib documentation - Matplotlib 3.7.0 documentation. [Online]. Available: https://matplotlib.org/stable/index.html. [Accessed: 23-Feb-2023].

Acknowledgment

1) Hari Balaji
2) Shubham more
3) Teaching assistance
4) Piysuh singh