DATA NARRATIVE(ES114)

Aditya Kumar, *22110015*

Data Analysis

Abstract—The objective of this study was to perform a data analysis task on a given dataset containing information about books, authors, ratings, tags, and user-to-read lists. The dataset was cleaned, processed and analysed to extract meaningful insights and observations. The analysis includes tasks such as merging datasets, filtering data, grouping, aggregating data, and visualizing data using graphs. The results obtained were used to draw conclusions and make recommendations for future studies.

I. Introduction

The dataset used in this study dataset contained information about books, authors, ratings, tags, and user-to-read lists. The data analysing tasks were performed to gain insights into the popularity of book, authors, and tags, and to identify trends and patterns in the ratings and user preferences.

II. Overview of the Dataset

The dataset contains several files:

1. books.csv: a dataset containing information about books including the book_id, goodreads_book_id, best_book_id, work_id, books_count, isbn, isbn13, authors, original_publication_year, original_title, title language_code, average_rating, ratings_count, work_ratings_count,work_text_review_counts, ratings_1, ratings_2, ratings_3, ratings_4, ratings_5, image_url, small_image_url. This dataset provides information about books and their authors, as well as their ratings and popularity based on the number of ratings.

2. book_tags.csv: a dataset containing information about book tags, including the book ID, tag ID, and count (how many times the tag was assigned to the book). The data provides information about book tags, which are labels assigned to books based on their content or theme. The tags are assigned by users and their count denotes how frequently that tag has been assigned to the book.

3. tags.csv: a dataset containing the tag ID and corresponding tag name. The data contains information about the tags assigned to books in the **book_tags.csv** file. It maps the tag ID to the corresponding tag name.

4. to_read.csv: a dataset containing information about books that users want to read, including the book ID and user ID.

5. ratings.csv: contains a dataset of user ratings of books on a scale of 1 to 5. Each row in the dataset represents a single user rating of a book and contains the following columns: user_id, book_id, rating.

III. Scientific Questions/Hypotheses

1. What rating does books with maximum “to-read” tag?
2. What are top 50 most popular genres(tags) in our given data?
3. What are the amount of books required to serve the demand of the user base(book demand is based on number of user who want read the particular book(in their to-read))? If book_count(books available is less than required), what books are they?
4.

IV. Details of Libraries and Functions

- pandas: Used for data manipulation and analysis, providing data structures for efficiently storing and querying large datasets.

Functions used from pandas:

- read_csv(): Used for reading data from a CSV file and returning a pandas DataFrame.
- groupby(): Used to group the data in the DataFrame based on a specified column(s).
- nunique(): Used to count the number of unique values in a pandas Series or DataFrame.
- reset_index(): Used to reset the index of a DataFrame.
- merge(): Used to combine two DataFrames based on a specified column(s).

- matplotlib: Used for data visualization and creating charts and plots.

Functions used from matplotlib:

- pyplot.bar(): Used to create a bar plot.
- pyplot.xticks(): Used to set the x-axis tick labels.
- pyplot.xlabel(): Used to set the x-axis label.
- pyplot.ylabel(): Used to set the y-axis label.
- pyplot.show(): Used to display the plot.

- numpy: Used for numerical computations and mathematical operations.

Functions used from numpy:

- nanmean(): Used to calculate the mean of a numpy array, ignoring any NaN values.

V. Answers to the Questions

1. What rating does books with maximum “to-read” tag?

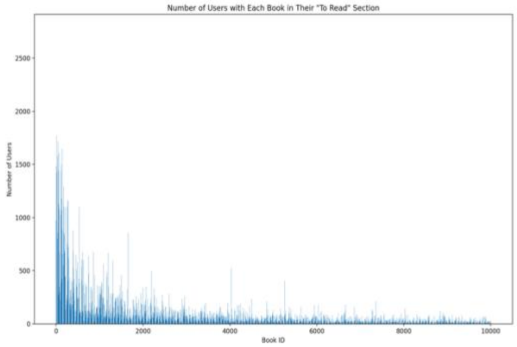


Fig 1. Book_id v/s Number of user who has the book in there to-read list.

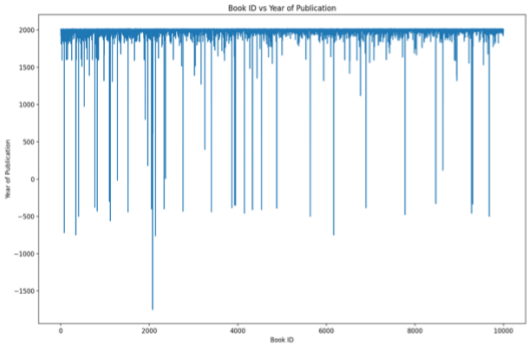


Fig 2. Book_id v/s Year of Publication

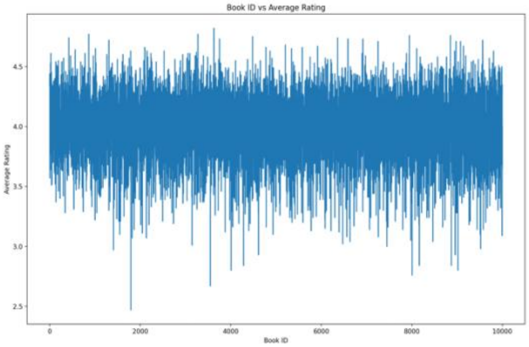


Fig 3. Book_id v/s Average Rating

From Fig1 and Fig 3, I found a relationship between the books which has highest demand among users(books with maximum to-read tag by user) and their rating, the books with high demand among user tend to have higher rating. Fig 1 book_id less than 2500 are in relatively higher demand but Fig 3 shows a relatively uniform graph in terms of average ratings as most of the books rating lies in range 4.0 to 4.5.

From Fig 1 and Fig 2, I found that very less books with a year of publication before 1700 were in high demand by users. In Fig 1 book_id less than 2500 are in relatively higher demand but in Fig 2 very less books with publication year before 1700 are in that range.

2. What are top 50 most popular genres(tags) in our given data?

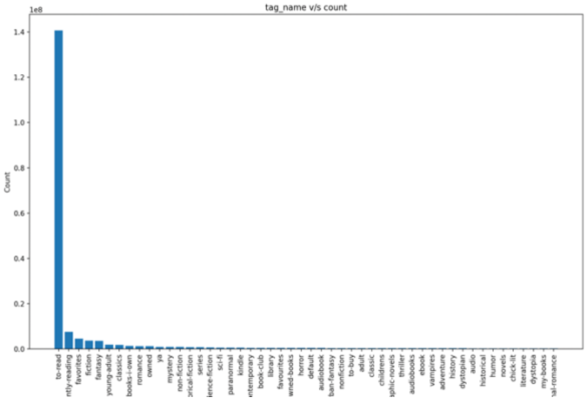


Fig 4. tag_name v/s No of tag count for each book(Top 50 books are considered)

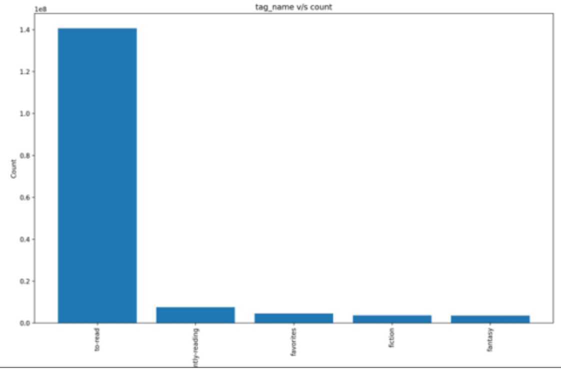
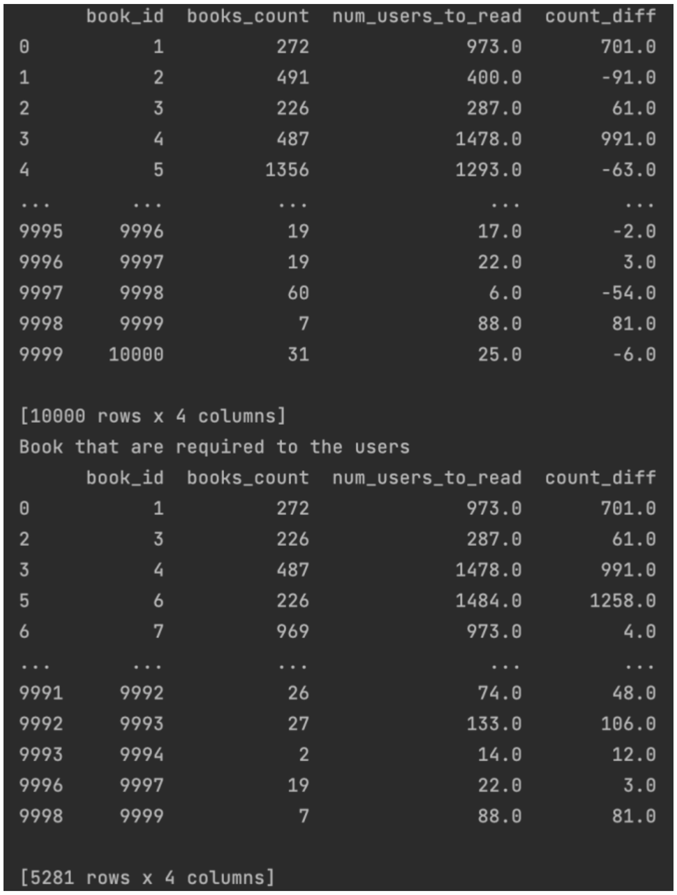


Fig 5. tag_name v/s No of tag count for each book(Top 5 books are considered)

From fig 4, we can observe the most popular tags(genres) for Top 50 tags. As we can observe “to-read” tag has the maximum tag count among all the tags.

From fig 5, we can observe the most popular tags(genres) for Top 5 tags. As we can observe “to-read” tag has a dominant difference among other 4 tags. The “to-read” might even be overshadowing the sum of all other tags given.

3. What are the amount of books required to serve the demand of the user base(book demand is based on number of user who want read the particular book(in their to-read))? If book_count(books available is less than required), what books are they?

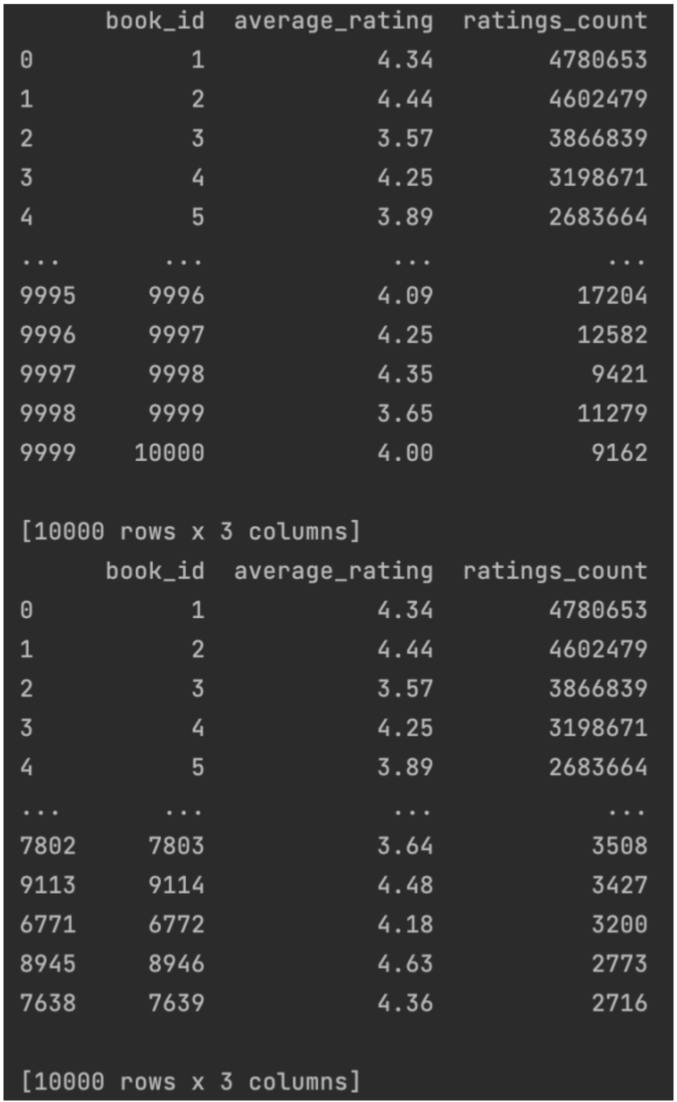


```
      book_id  books_count  num_users_to_read  count_diff
0           1          272              973.0       701.0
1           2          491              400.0       -91.0
2           3          226              287.0        61.0
3           4          487             1478.0       991.0
4           5         1356             1293.0       -63.0
...       ...          ...                ...         ...
9995     9996           19               17.0        -2.0
9996     9997           19               22.0         3.0
9997     9998           60                6.0       -54.0
9998     9999            7               88.0        81.0
9999    10000           31               25.0        -6.0

[10000 rows x 4 columns]
Book that are required to the users
      book_id  books_count  num_users_to_read  count_diff
0           1          272              973.0       701.0
2           3          226              287.0        61.0
3           4          487             1478.0       991.0
5           6          226             1484.0      1258.0
6           7          969              973.0         4.0
...       ...          ...                ...         ...
9991     9992           26               74.0        48.0
9992     9993           27              133.0       106.0
9993     9994            2               14.0        12.0
9996     9997           19               22.0         3.0
9998     9999            7               88.0        81.0

[5281 rows x 4 columns]
```

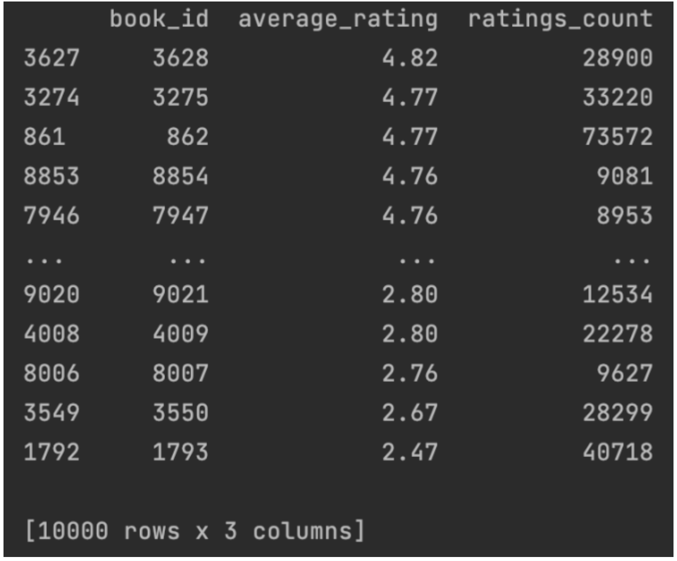
Fig 6. Table for sample

From the file “books.csv”, I created a dataframe with book_id, and book_counts(which are available). From file “to_read.csv”, I created another dataframe with book_id, and num_user_to_read(Number of user who want to read that book). Then I merged both my dataframe with respect to book_id and from that merged dataframe (merged_df), I find the difference between the num_user_to_read and book_counts to find how many books are required to meet the needs of the users.

4. Find whether books with good rating have more reviews or not ?



```
      book_id  average_rating  ratings_count
0           1            4.34        4780653
1           2            4.44        4602479
2           3            3.57        3866839
3           4            4.25        3198671
4           5            3.89        2683664
...       ...             ...            ...
9995     9996            4.09          17204
9996     9997            4.25          12582
9997     9998            4.35           9421
9998     9999            3.65          11279
9999    10000            4.00           9162

[10000 rows x 3 columns]
      book_id  average_rating  ratings_count
0           1            4.34        4780653
1           2            4.44        4602479
2           3            3.57        3866839
3           4            4.25        3198671
4           5            3.89        2683664
...       ...             ...            ...
7802     7803            3.64           3508
9113     9114            4.48           3427
6771     6772            4.18           3200
8945     8946            4.63           2773
7638     7639            4.36           2716

[10000 rows x 3 columns]
```

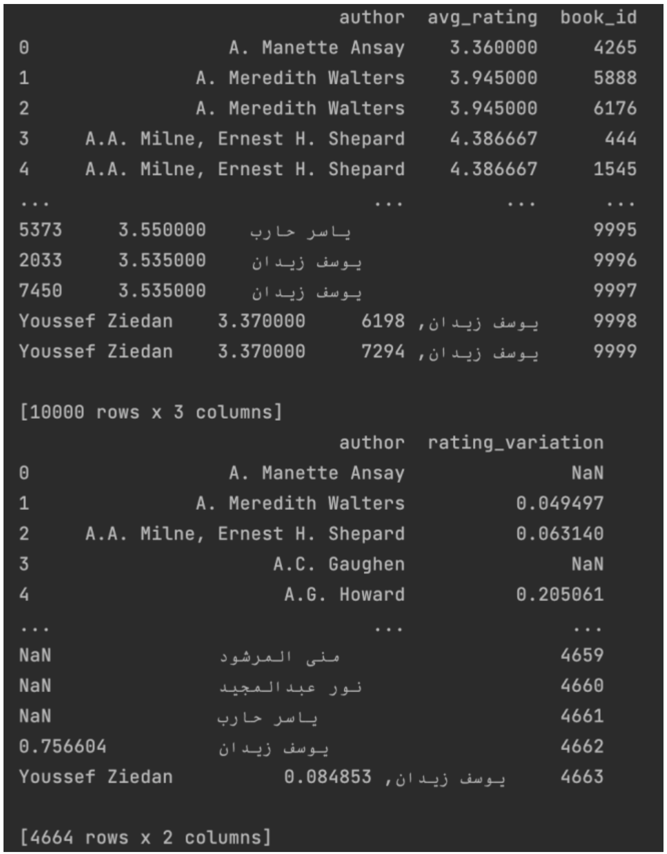


```
      book_id  average_rating  ratings_count
3627     3628            4.82          28900
3274     3275            4.77          33220
861       862            4.77          73572
8853     8854            4.76           9081
7946     7947            4.76           8953
...       ...             ...            ...
9020     9021            2.80          12534
4008     4009            2.80          22278
8006     8007            2.76           9627
3549     3550            2.67          28299
1792     1793            2.47          40718

[10000 rows x 3 columns]
```

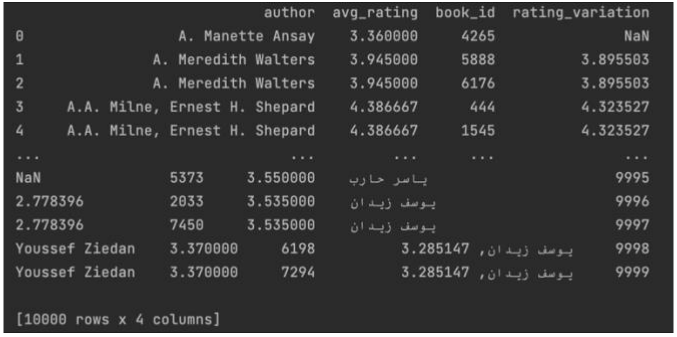
From the above data, I tried to observe relationship between number of reviews of a book and its average rating. My statement was, “If books with good rating have more reviews or not”.

After analysing the data, I found that it might be true for some books but not for most of them. So, it can’t be concluded that books with more reviews will have better rating.

5. Find the rating variation for all the authors with respect to their individual books average rating.



```
                          author  avg_rating  book_id
0                A. Manette Ansay    3.360000     4265
1              A. Meredith Walters    3.945000     5888
2              A. Meredith Walters    3.945000     6176
3     A.A. Milne, Ernest H. Shepard    4.386667      444
4     A.A. Milne, Ernest H. Shepard    4.386667     1545
...                            ...         ...      ...
5373      3.550000     ياسر حارب                 9995
2033      3.535000     يوسف زيدان                9996
7450      3.535000     يوسف زيدان                9997
Youssef Ziedan     3.370000     6198 ,يوسف زيدان     9998
Youssef Ziedan     3.370000     7294 ,يوسف زيدان     9999

[10000 rows x 3 columns]
                          author  rating_variation
0                A. Manette Ansay               NaN
1              A. Meredith Walters          0.049497
2     A.A. Milne, Ernest H. Shepard          0.063140
3                    A.C. Gaughen               NaN
4                     A.G. Howard          0.205061
...                            ...               ...
NaN              منى المرشود                  4659
NaN              نور عبدالمجيد                4660
NaN              ياسر حارب                    4661
0.756604         يوسف زيدان                   4662
Youssef Ziedan          0.084853 ,يوسف زيدان     4663

[4664 rows x 2 columns]
```



```
                          author  avg_rating  book_id  rating_variation
0                A. Manette Ansay    3.360000     4265               NaN
1              A. Meredith Walters    3.945000     5888          3.895503
2              A. Meredith Walters    3.945000     6176          3.895503
3     A.A. Milne, Ernest H. Shepard    4.386667      444          4.323527
4     A.A. Milne, Ernest H. Shepard    4.386667     1545          4.323527
...                            ...         ...      ...               ...
NaN              5373      3.550000  ياسر حارب                       9995
2.778396         2033      3.535000  يوسف زيدان                      9996
2.778396         7450      3.535000  يوسف زيدان                      9997
Youssef Ziedan   3.370000      6198         3.285147 ,يوسف زيدان      9998
Youssef Ziedan   3.370000      7294         3.285147 ,يوسف زيدان      9999

[10000 rows x 4 columns]
```

From the above analysis, I tried to find the rating variation of the each author for their books with respect to their own average rating.
If rating variation for an author is high, that means that the author has both good and bad rated books.

6. From the current reading list of our input user, recommend him other book with the help of other user book read history?

I made list of our users read books and check our input user reading history with other users and find book which is most common among other user. The other user also have read same books from our current user history.

VI. Acknowledgement

- Most of the books have a rating of around 3.5-4.5 and the distribution of ratings is slightly skewed towards the higher end.
- Some authors have a wider variation in the ratings of their books compared to others.
- The most popular tags across all books include "to-read", "favorites", and "fiction".
- The number of users who want to read a book is usually less than the total number of ratings given to that book.
- Some books have a higher number of users who want to read them compared to their total book count.
- The ratings given to a book are positively correlated with the number of ratings given to that book.

Overall, the dataset provides insights into the popularity and ratings of different books and authors, as well as the tags associated with them. It can be useful for various applications such as recommending books to users, analyzing book trends, and understanding user preferences.

VII. References

1. How to plot bar graph from dataframe.
2. How to sort data in a dataframe
3. How to merge two dataframes.
4. How to take mean of data in dataframe.
5. How to group data from a dataframe with some columns of the dataframe.

VIII. Acknowledgement

I would like to express my special thanks of gratitude to Prof. Shanmuga for his guidance and support in my Data Narrative analysis by reviewing and helping me improve the questions.