ES114 Probability & Statistic
Data Narrative

Archit Dhakar 22110031

***OVERVIEW*—This data narrative is to visualize the data from a book library given in the format of .csv . Using different libraries and functions of python we have to form questions or create a hypothesis which can help the user or company to improve their judgment.**

***Index Terms*— pd: pandas , np: numpy**
matplotlib.pyplot: plt , MPR: most probable rating

I. Scientific Questions/ Hypothesis

1. Using previous ratings by users show that book written by a given author will be having very good ratings or the content will be better than previous books?
2. Find most popular categories(tags) of books, read by maximum number of users? Also provide a graph to visualize this?
3. Answer the below questions in order to help the user to choose the best books.
   (a) Find the top 10 Best Books on the basis of rating?
   (b) Top 10 Books which got higher 1 rating?
   (c) Top 10 Books read/rated maximum in data?
4. Answering below question find a promoting strategy which book library organization can use to attract more number of users to read their books.
   (a) User id of the user who have read/rated maximum Books?
   (b) Also tell the Number of books read/rated by that user.
5. For the given user check whether the ratings are fare ? also give reason how this affect the overall data analysis.

II. Details of Libraries & Functions

Libraries and functions used to visualize the data are:

1. NumPy library
2. Matplotlib library to plot the graphs using data
3. Pandas Library to create the DataFrame from the given csv files, Series and to plot the bar charts, bar graphs.
4. Scipy library
5. read_csv function to read the csv files in Python.
6. .head() and .tail() to read some important values of data
7. sorted function to sort the columns of dataframe.

III. Answer of Questions

1. Using previous average ratings on books written by an author we can predict that upcoming books will be better than previous one.

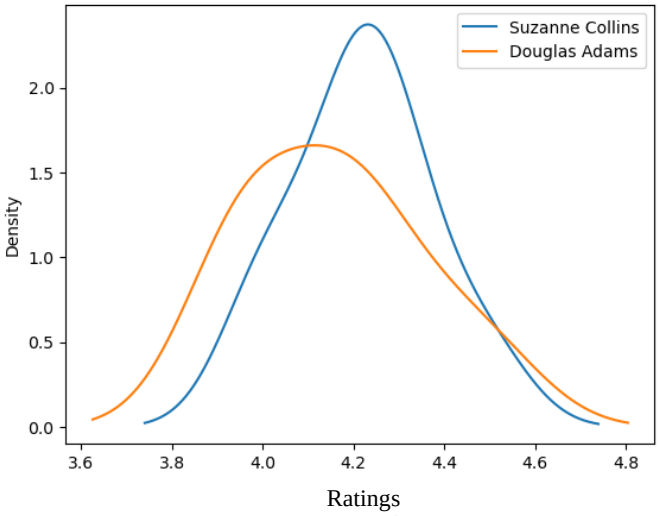

Fig1. Density distribution of average ratings by users

Here we have taken two authors Suzanne and Douglas to see the prediction using probability density for random variable.

For blue curve:
We can observe from plot that on an most of the rating is between 4.2 to 4.4 . After that the area of the curve is less than the previous version from which we can say that probability w.r.t MPR(most probable rating) is less for after part. Hence the value of ratings are decreasing for Suzanne Collins.

From this observation company can give more opportunities to those authors who are writing better books. Which we can observe from our next example.

For Orange curve:
We can observe that average rating is between 4.0 to 4.2 by seeing the MPR. Also the area of curve after MPR part is greater than previous one so the probability. Hence we can say that ratings in future will be better for Douglas Adams.

Also from whole observation a user can set a reminder to read the books of author with better content/ratings.
We can expand our analysis and make a list of such authors.

2. Here I have done analysis for 7 most popular Tags or categories of books read by maximum number of users. This data analysis can help the user to decide that which category has the interesting book collection.

User can visualize the data from the pie chart given below:

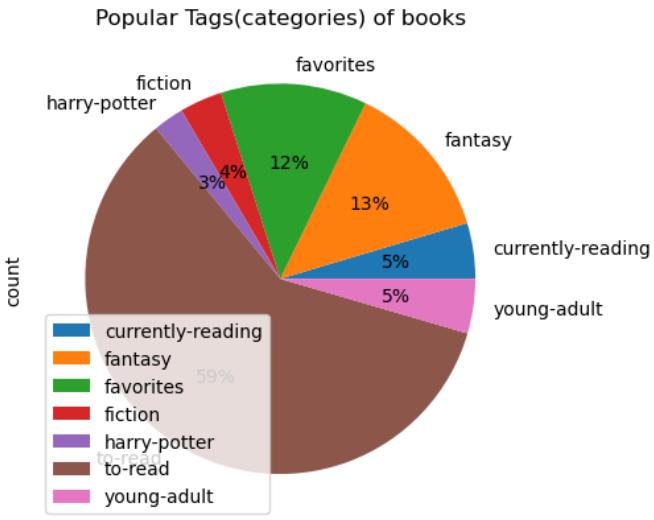

Fig2. Pie chart popular tag visualization

From the pie chart we can observe that books with to-read categories(59%) are the most popular as those are choice of many users. Also 13% of the user are interested in fantasy zone. Similarly other zones percentage can help users to choose interesting books. It can help the book library organization to bring most popular categories.

3. After doing data analysis on three parts of questions user will be able to find the best books present in the library.

1.Best rating books

	original_title	average_rating
6919	The Indispensable Calvin and Hobbes: A Calvin ...	4.73
3752	Harry Potter Collection (Harry Potter, #1-6)	4.73
6360	There's Treasure Everywhere: A Calvin and Hobb...	4.74
421	Complete Harry Potter Boxed Set	4.74
4482	It's a Magical World: A Calvin and Hobbes Coll...	4.75
8853	Mark of the Lion Trilogy	4.76
7946	NaN	4.76
861	Words of Radiance	4.77
3274	NaN	4.77
3627	The Complete Calvin and Hobbes	4.82

2.Books with highest 1 rating

	original_title	ratings_1
8	Angels & Demons	77841
51	Eclipse	83094
4	The Great Gatsby	86236
27	Lord of the Flies	92779
39	Eat, pray, love: one woman's search for everyt...	100373
55	Breaking Dawn	100994
48	New Moon (Twilight, #2)	102837
7	The Catcher in the Rye	109383
33	Fifty Shades of Grey	165455
2	Twilight	456191

3.Most readed Book/Books with higher rating count

	original_title	average_rating	ratings_count
4	The Great Gatsby	3.89	2683664
3	To Kill a Mockingbird	4.25	3198671
2	Twilight	3.57	3866839
1	Harry Potter and the Philosopher's Stone	4.44	4602479
0	The Hunger Games	4.34	4780653

Table1. To categories the books in best possible way

From the above tables user can find the best book present in the library on the basis of average ratings.
User can consider the books with higher average ratings for review to read in future.
Also to save his/her time user can ignore the books with higher rating 1 because these are the books which were not liked by users.
User can also consider the books with maximum rating count because these are the books are read by maximum number of peoples. So using these tables user can consider the best choices and remove very less rated content.

4. Using data narrative analysis we can help the daily user of that library who have read/rated maximum number of books. So we can find the user id of the user that has spent the maximum time in the book library.

```
1. User(id) who have read/rated maximum number of books
0    12874
1    30944
Name: user_id, dtype: int64

2. No of books read/rated by the user with maximum reads is 200
```

Also book librarian can use this data to promote their library by giving extra points to that user which can be used to purchase a book which normal user can not user.
This data can be helpful to attract more number of users in the book library organization.

5. We can do the analysis that whether the user is trust worthy or not but observing his/her ratings. If there is variation in ratings means he has actually read the book , otherwise he just rated every book as 5.

```
2332463    3
2332464    3
2332465    4
2332466    2
2332467    5
2332468    5
2332469    5
2332470    4
2332472    4
2332473    5
Name: rating, dtype: int64
```

From the above data we can observe that user with ID 30994 have rated books fairly because he has given every kind of ratings to different books.
This can be useful to build a function that can filter out these kind of users who are fake.
If ratings will not be fair than this can change the whole filters used for all analysis.

IV. Summary of observations

So the observations include the probability density distribution which can help to predict the future aspect of books written by an author will be good or not, which can ultimately help the user and company both. Also it include the categorization of popular tags which can help the user to choose the trending books. Also this can help that book library organization to bring the most popular books more than less populars. Categorisation of data such that user can find the best rated books easily and ignore the books which are very less rated. Also using data analysis we can find the user who have read maximum books which can be used by librarian to give some points to that user as business strategy. Also to identify whether user has given rating fairly or not.

VI. References

[1] To create bar graph using pandas.
[2] Plot pie chart using pandas.
[3] Sort the column of DataFrame in pandas
[4] To open the csv file using github url

VII. Acknowledgment

I would like to express my special thanks of gratitude to Prof. Shanmuga for his guidance and support in my Data Narrative thought building .

Date: 23-02-2023
Archit Dhakar
22110031