DATA NARRATIVE 1

ES 114 PROBABILITY,STATISTICS AND DATA VISUALIZATION

Purva Kaushalbhai Shah
Chemical Engineering
Indian Institute of Technology Ganghinagar
Ganghinagar,India
purva.shah@iitgn.ac.in

***Abstract*—This document is the report on Data Narrative. The data given in form of 5 csv file containing the data related to the books was analyzed. In total 10,000 books were analyzed. The data contains the tag, tag_id, user_id, title, author, count, ratings, image etc. The data is interconnected between all the files.**

I. INTRODUCTION

In this task of Data Narrative, we are analyzing the data present in csv files by using few visuals and databases. The database is about books so, in order to understand the database, the basic questions like The language in which the book is written, most frequently used rating, the authors publishing more books, the probability of getting type of books people prefer, the users who read most number of books, the id of book which has maximum rating etc.

II. OVERVIEW

In this task, we are reading the csv files using pandas, importing the data and reading. The data is provided in Github[1] in the form of 5 csv files. Every csv file can be inter connected as they have atleast 1 column in common. The file consist of 10,000 books data including the rating, title, year of publication, author name, average rating, number of people reading the books, tags, etc.

III. QUESTIONS / HYPOTHESIS

A. Top 7 languages in which books are writtern**.**

B. Top 10 authors who published maximum number of books.

C. Which rating is given by most of people?What is the probabilty of that rating to occur?

D. Pie chart to show top 10 type of books people have. What is the probabilty that a person picks a book of to-read or to-buy.

E. Top 21 user id who read maximum number of books. Probability that a person chosen from these 21 has read same number as other one given that they have read books less than 100

F. Showing the id and author of books having the highest avg.rating. Printing the name of the authors of these books

IV. SOFTWARE / LIBRARIES USED

A. Google colab

A platform provided by google where wee can code. In this tast google collab was used to read the uploaded file, code and write some texts in between. The code in google collab is written in python language.

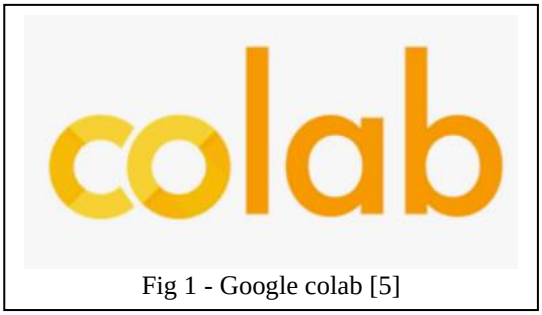

Fig 1 - Google colab [5]

B. Pandas [3]

Pandas is a software which provides us with various inbuilt function.In this task, we are using pandas to read the csv file and operate on it. The csv files are viewed are operated like we operate the Dataframes.

1. Functions used:
   a. pd.read_csv(): to read the csv file
   b. pd.head(): to return few rows of the dataframe from the top
   c. groupby() : to group the identical values and separate out the data
   d. sort_values(): it is used to sort the data using a specific label
   e. count(): used to count the occurrence of any data

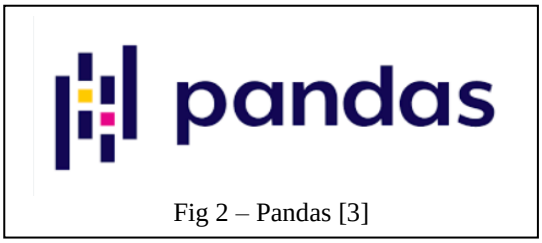

Fig 2 – Pandas [3]

C. Matplotlib[2]

Matplotlib is the software used to plot the graphs. In the task of data narrative, it was used to plot the various type of graphs like bar graph, pie chart etc.

1. Functions used:

a. figure(): to open a panel where graph can be plotted
b. plot(): to plot the given graph
c. xlabel(): to label x axis
d. ylabel(): to label y axis
e. bar(): to plot the bar graph
f. pie(): to plot the pie chart
g. show(): to show the plotted graph

Fig 3 – Matlplotlib [2]

V. ANSWERS

A. The top 7 launguges in which books are written are:

1. eng
2. en – US
3. en-GB
4. ara
5. en-CA
6. fre
7. ind

To reach to this conclusion I have counted the number of people who read book in a particular language and then sorted this count in descending order. To print the top 7 I used the head() function.

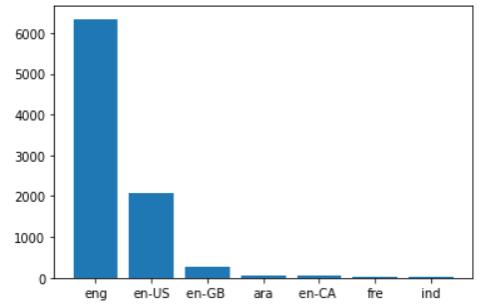

Fig 4 – Graph of number of readers vs launguges

B. The authors who published the maximum number of books are:

1. Stephen King
2. Nora Roberts
3. Dean Koontz
4. Terry Pratchett
5. Agatha Christie
6. Meg Cabot
7. James Patterson
8. David Baldacci
9. John Grisham
10. J.D. Robb

To reach to this conclusion I first counted the books published by each author and just like previous sorted in descending order and then printed the top 10 authors.

C. The rating that most people gave is 4. It is quite evident from the pie chart given below (Fig 5).

The probability of rating 4 to occur when a random rating is chosen is 0.35790605137238835

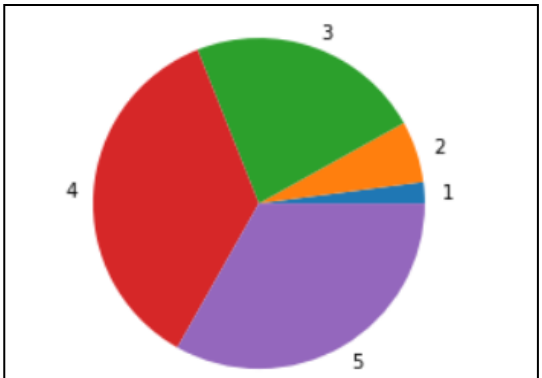

Fig 5- Pie chart showing how many people gave these ratings

To reach to this conclusion I have counted the number of people who gave the rating and the I plotted the pie chart using matplotlib.

D. The pie chart showind top 10 type of books is:

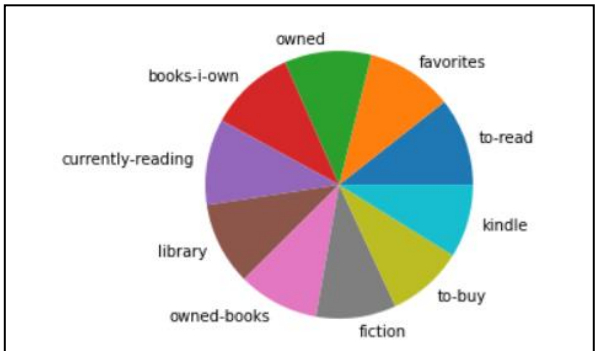

Fig 6 – pie chart showing top 10 type of books preferred by the people

The probability that a person picks a book of to-read or to-buy is 0.018676624866363373

To reach to this conclusion I first found the data and then plotted it in pie chat.

For the probability I used the basic formulas.

E. The top 21 user id who read maximum number of books are:

38457
28259
38076
44530
46000
46555
34162
34487
39174
24784
40362
47363
44389
48339
25952
11924
17716
36428
19246
46331
31631

probability of person who has read less than 100 books and has read the same number of books as the other one is: 0.8571428571428571

For the probability I used the formula of conditional probability and for the data I counted the books read by the author.

F. The top 5 books with highest ratings,their author and id are:

Book id	ratings	Author
3627	4.82	Bill Watterson
3274	4.77	Brandon Sanderson
861	4.77	J. K. Rowling, Mary Grandprao
8853	4.76	Lane T. Dennis, Wayne A. Grudem, Anonymous
7946	4.76	Francine Rivers

To get this data I first sorted the avg. ratings in descending order and then found the book id and author.

VI. OBSERVATION

From the above answered ques we ca see that book written english is the most commonly read. Author Stephen king has written 60 books till now. Users genrally rate the book with rating 4. And the probability of user rating a book 4 is 0.35. People mostly prefer the to-read books. The user id 38457 read the most number of books i.e. 117. The book written by Bill Waterson having id 3627 has the highest avg rating 4.82. This means it in one of the best book out of 10,000 books.

VII. UNANSWERED QUESTIIONS

There are no unanswered questions asked by me.

VIII. ACKNOLEDGEMENT

I would like to thank to prof. Shanmuga for providing me with the data to do the analysis. The links provided were very helpful.

IX. REFERENCES

[1] Zając, Zygmunt. “Zygmuntz/Goodbooks-10k.” *GitHub*, 2 Apr. 2021, github.com/zygmuntz/goodbooks-10k.

[2] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007

[3] Pandas. “Python Data Analysis Library — Pandas: Python Data Analysis Library.” *Pydata.org*, 2018, pandas.pydata.org/..

[4] Python. “Welcome to Python.org.” *Python.org*, Python.org, 29 May 2019, www.python.org/.

[5] https://colab.research.google.com/