

DATA NARRATIVE

#1 Vipul Sunil Patil (22110189)
Btech First Year Student (CSE),
IIT Gandhinagar,
Gandhinagar, Gujrat, India.
vipul.patil@iitgn.ac.in

I. OVERVIEW OF THE DATASET

The dataset consists of 5 file in side every file we get an big data that we had to analyze and make questions of The dataset contains in total 27 variables, including user_id, book_id, rating, goodreads_book_id, tag_id, count, goodreads_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_reviews_count, ratings_1, ratings_2, ratings_3, ratings_4, ratings_5, image_url, small_image_url, tag_id and tag_name.

Strings, integers, and floats are all present in this data. However, although being large, this data is insufficient in certain ways. Certain things are not adequately described. Several missing information had needed to be filled in, and duplicate responses had to be eliminated. Data was usable after all, being easily transformed, calculated according to our need; this data was ready to create questions.

In my situation, I make use of data visualization and statistics to make sense of the data. It also helps us understand what rating count, picture url, publication of books, year of release, and many other things. On the other hand, exploring this data and seeing patterns and trends explains to me how books are rated and what additional criteria we use.

Statistical analysis and charts came to provide numerous conclusions in front of me, but this data did not support some of them. With this information, we were able to provide several answers. Below questions and their answers will explain the data much closer and will make it easy to understand the data.

II. SCIENTIFIC QUESTIONS/HYPOTHESES

1. 'Most of the popular authors prefer to write in the English language' -find whether the statement is True or False.

2. Check Which language is the most extensively used and widely accepted language for book writing?

3. Determine the user group that is or is not Fascinated for reading books. Find the overall rating each user has given a book?

4. Locate people who had given books an overall rating of 0-1, 1-2, 2-3, 3-4, or 4-5 in question number 3. Also, the number of users with ratings that fall within 0-1, 1-2, 3-4, and 4-5.

5. Determining relation of user average ratings and best book rankings of a particular book? Find whether rating on the main criteria for book ranking.

III. DETAILS OF LIBRARIES AND FUNCTIONS

Labraries Used in the Code are:

1. Pandas can be defined as a Python package that offers user-friendly data structures and tools for data analysis. It is constructed on top of NumPy, a scientific computing library for Python, and presents an effective and simple environment for data manipulation and analysis to Python developers. Pandas includes two primary data structures, namely the Series and DataFrame objects. The Series is a one-dimensional array-like object that can contain any data type, whereas the DataFrame is a two-dimensional table-like data structure with rows and columns capable of holding various data types.[1]
2. Matplotlib is a versatile and easy-to-use visualization library in Python that offers powerful tools for creating different types of plots such as line, bar, scatter, histogram, and many more. It is based on NumPy arrays and has seamless integration with the broader SciPy ecosystem.[2]

This are the Fuctions used in code.

1. For making dataframe we Dataframe().
2. To sum the values we use sum().
3. To group data as we need to access we use groupby().
4. reset_index() method to reset the index of the dataframe.
5. using the plot() we can create plot of graph and specify kind using kind.
6. We show the chart using the show().
7. sort_values() is used to sort values ascending or descending accordingly.
8. scatter() helps us in plotting scatter plots.
9. For the bar plot we use barh() where h means horizontal.
10. We set the title for the chart using the title() and x and y labels using xlabel() and ylabel() respectively.

IV. ANSWERS TO THE QUESTIONS (WITH APPROPRIATE ILLUSTRATIONS)

1. For this question we need to first find the author of the books which are the maximum readed (assuming those who had read had given rating) to get a popular author.

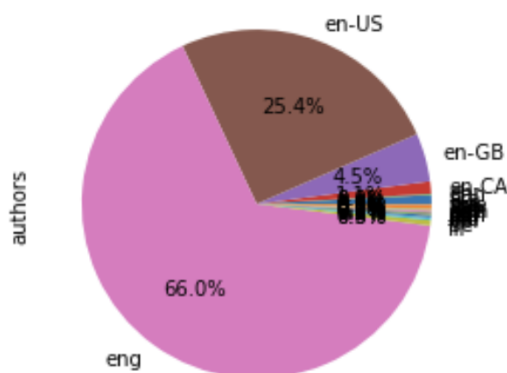
```
groups=df.groupby(['authors', 'language_code'])
rating=groups['ratings_count'].mean()
```

Further we need to find languages that authors prefer to write their books

```
lang_counts = top_lang.groupby('language_code')['authors'].count()
```

and then by plotting a pie graph between popular authors and the language they use will clearly show that it is true how the Most of the popular authors prefer to write in the English language.

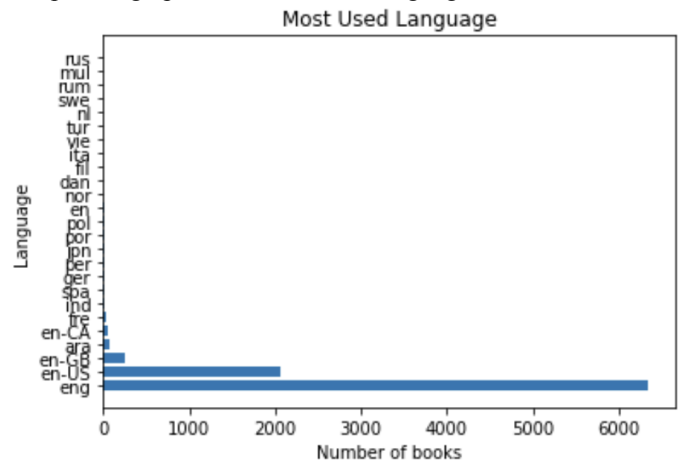
Language Prefered by Top 5000 Authors



2. To check the dominance of the a particular language in the market of readers and authors we need to plot a the most used language graph. For this we need to count the number of written in different languages

```
lang_counts = df['language_code'].value_counts()
```

and plot a graph that shows all languages and their count.



3. For this first we need to count the number of books readed by individual users,

```
books_count=df.groupby('user_id')['book_id'].count()
```

Next we need to find the rating given so we calculate its mean value to get the mean rating that the user had given to those books.

```
mean_ratings_of_users=df.groupby('user_id')['rating'].mean()
```

We need a new data frame for analyzing dedicated book readers. This is the data of some users who had read maximum books and minimum books with their mean ratings.

	book_counts	mean_ratings
user_id		
30944	200	4.210000
12874	200	3.450000
52036	199	3.442211
12381	199	3.427136
28158	199	3.939698
...
32128	21	4.000000
40753	21	3.428571
51725	21	3.523810
43675	20	4.150000
34590	19	4.473684
[53424 rows x 2 columns]		

4. For this question, we can consider people who have given books ratings of 0-1, 1-2, 3, 4, or 4-5.

```
rating_01=result[(result['mean_ratings']>=0) & (result['mean_ratings']<=1)]
rating_12=result[(result['mean_ratings']>=1) & (result['mean_ratings']<=2)]
rating_23=result[(result['mean_ratings']>=2) & (result['mean_ratings']<=3)]
rating_34=result[(result['mean_ratings']>=3) & (result['mean_ratings']<=4)]
rating_45=result[(result['mean_ratings']>=4) & (result['mean_ratings']<=5)]
```

From this we will get the following data table.

	book_counts	mean_ratings
user_id		
47953	65	1.0
48515	94	1.0
49679	63	1.0

	book_counts	mean_ratings
user_id		
3	91	1.736264
7871	98	1.765306
11793	92	1.717391

	book_counts	mean_ratings
user_id		
53	49	2.816327
68	51	2.980392
80	64	2.656250

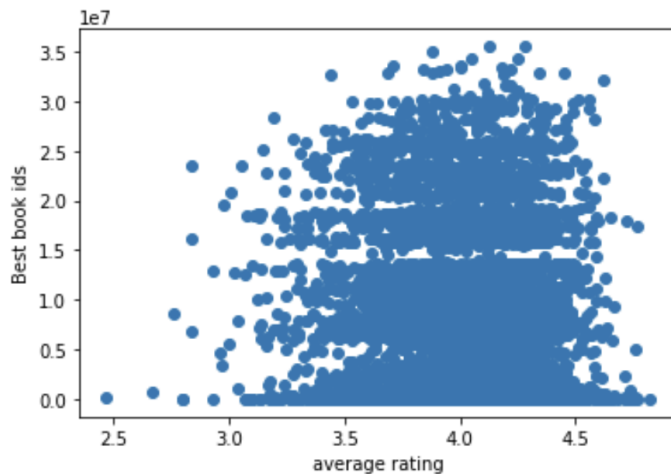
	book_counts	mean_ratings
user_id		
1	117	3.589744
4	134	3.768657
7	155	3.819355

	book_counts	mean_ratings
user_id		
2	65	4.415385
5	100	4.040000
6	90	4.322222

5. Here we need to plot a graph between average rating and best book ranking which has been provided in the data itself. Plotting a scatter plot will be best suitable for this as we need to see in which criteria are the books fitting well.

```
plt.scatter(df['average_rating'], df['best_book_id'])
```

No Rating is not the main criteria for book ranking.



The Data shows how book ranking is made, and only rating is not only a parameter of it; it also comprises other aspects. We can see that there are many underrated books which were not at the highs in terms of best book ranking, but the people who read them give them the best rating.

V. SUMMARY OF THE OBSERVATIONS

1. Popular authors' preferred languages might reveal useful information about their writing style, target audience, and cultural background. Many applications are possible for this information.

First off, understanding an author's preferred language can help in translation and localization, ensuring that the meaning and tone of the original work are preserved in other languages if that author's work is well-unknown in several other nations.

Second, if you are advertising a book or a product to a particular author's audience, you can better personalise your marketing plan to appeal to that target market by learning their favoured language.

2. The Data demonstrates how book rankings are created, and it includes more than just ratings as one of its parameters. We can see that there are many underappreciated books that receive the highest ratings from readers while not being at the top of the best book rankings.

Therefore we can conclude that there are many books which have high average ratings but still are not a part of the best books (upper right part shows this). On the other hand, best books with low average ratings also exist (bottom left part shows this).

3. This information allows us to identify the users who read books the most. This data will be useful to give suggestions for those people who read books for a maximum amount of time as they won't ignore it because they want more and more books for reading. On the other hand, we don't need to recommend books to those who just read books for a very minimal amount of time because we can make other recommendations to them.

As a result, the advertising sector may use this data more effectively to display ads.

4. This information enables us to group users according to their ratings and identify those who provide higher and lower ratings. We can categorise people based on their preferences from this. On the other hand, we can ask users who have given books low ratings what should be done to enhance the quality of the books, and what were the weak points of the books, and what books they would be interested in reading in order to raise the ratings of the books.

5. Through the overview of this data we come to the conclusion that English will be the most popular language and will have its dominance for more upcoming years. This could be useful for upcoming writers as they can see that English work has demand in the market.

As English demand has increased it replies that there will also be an increase in the number of translators. So we can predict that there will be an increase in English content and vice versa the Translators.

VI. UNANSWERABLE QUESTIONS, IF ANY

1. Which books were popular during a specific time interval (Year)? Through this can we determine Famous Genres by Year?
2. How do readers rate a particular book and on what basis? Is it linked according to the book author, genre, and year of publishing?

SUMMARY OF THE OBSERVATION THAT COULD BE CREATED BY UNANSWERABLE QUESTIONS


1. This data could help us gain a better understanding of how people change their preferences with respect to time and according to circumstances. It could also give us a clear perspective of how Genres affect people's thoughts. On the other hand, it would have proves how things around us play a vital role in our life, knowingly or unknowingly. This would also provide insight of how cultural preferences shift as new technologies emerge. It could also help us understand how artists and creatioers adjust their approaches to meet consumer demand. Similarly, it would have help us in predicting future trends that how the public will change its interest in different fields with respect to time and change in particular genre industries.

VIII. REFERENCES

[1]The pandas development team. 2023. *Pandas-Dev/Pandas: Pandas*. Zenodo.

[2]"Matplotlib Tutorial." 2021. GeeksforGeeks. February 8, 2021. <https://www.geeksforgeeks.org/matplotlib-tutorial/>.

This are some webs I used in my code:

1. "Python  Sort Grouped Pandas Dataframe by Group Size." n.d. Tutorialspoint.com. Accessed February 23, 2023.
2. "How to Sort Data by Column in a CSV File in Python?" 2021. GeeksforGeeks. June 3, 2021. <https://www.geeksforgeeks.org/how-to-sort-data-by-column-in-a-csv-file-in-python/>.
3. "Difference between '&' and 'and' in Pandas." n.d. Stack Overflow. Accessed February 23, 2023. <https://stackoverflow.com/questions/54315627/difference-between-and-and-in-pandas>.
4. "Range Filtering with BETWEEN." n.d. Peachpit.com. Accessed February 23, 2023.

<https://www.peachpit.com/articles/article.aspx?p=1276352&seqNum=8>.

5. "Pandas DataFrame Count() Method." n.d. W3schools.com. Accessed February 23, 2023. https://www.w3schools.com/python/pandas/ref_df_count.asp.

IX. ACKNOWLEDGEMENTS

My sincere gratitude goes out to Sir Shanmuga R, our course lecturer for ES 114 Probability, Statistics, and Data Visualization, as well as to all of the TAs who helped me out whenever I needed assistance with this assignment.

First and foremost, I want to express my gratitude to all the online resources that provided me with the necessary resources to finish this task. Also, I want to thank all of the organisations that have helped me by lending me their knowledge and insight in a variety of sectors through a variety of additional channels. Finally, I want to express my gratitude for the chance you offered me to ask questions, learn how to interpret facts, and acquire this incredible skill.