ES 114 Data Narrative 1

Chinthala Shivamani, Roll No.:22110062,
Computer Science and Engineering Department,
Prof.Shanmuga R, IIT Gandhinagar

***Abstract*— The purpose of this report is to demonstrate how Python, and programming in general, are capable of making our lives simpler by reading data and giving results we want much more easily than we could manually do.**

***Keywords*— Python libraries, Numpy, Matplotlib, Pandas, read files.**

I. Overview of the dataset

The document below is a template. It is available for download on the conference website. The conference website provides information about final paper submissions. The dataset is available in CSV format, which makes it easy to load into Pandas and R. Datasets under the Open Database License can be freely used and shared.

II. Scientific questions or hypotheses

1. What are the top 10 most frequently rated books, and how many ratings do they have?

2. What are the top 5 most popular authors based on the number of books in the dataset, and how many books does each author have

3. How many unique books are in the dataset, and what is the average rating of these books?

4. What is the average book rating of the books published in each year? Plot the values in the bar graph

5. Mention the top 5 best book to read by the readers Create the table format to display the results

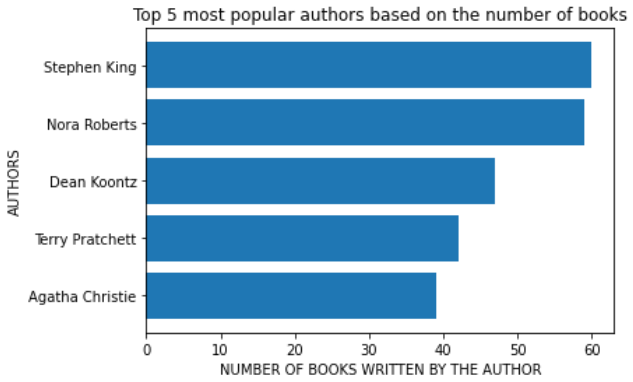
III. Answers to the questions

1. This question can be used to know about the frequently rated books.So that the reader can know about the new and high rated books

   The output gives the top 10 most high rated books.
   The output is
   1) The Hunger Games (The Hunger Games, #1): 18642 ratings
   2) Harry Potter and the Sorcerer's Stone (Harry Potter, #1): 17436 ratings
   3) To Kill a Mockingbird: 15132 ratings
   4) Twilight (Twilight, #1): 14057 ratings
   5) The Great Gatsby: 13443 ratings
   6) Catching Fire (The Hunger Games, #2): 13048 ratings
   7) The Hobbit: 12982 ratings
   8) Mockingjay (The Hunger Games, #3): 12397 ratings
   9) Harry Potter and the Prisoner of Azkaban (Harry Potter, #3): 12063 ratings
   10) The Catcher in the Rye: 11962 ratings

2. This question gives the top 5 popular author by the based on the number of books published.So that the reader user can find the top most popular author who wrote the more books and read the particular books

   We know that the author named Stephen King is the popular author.



3.

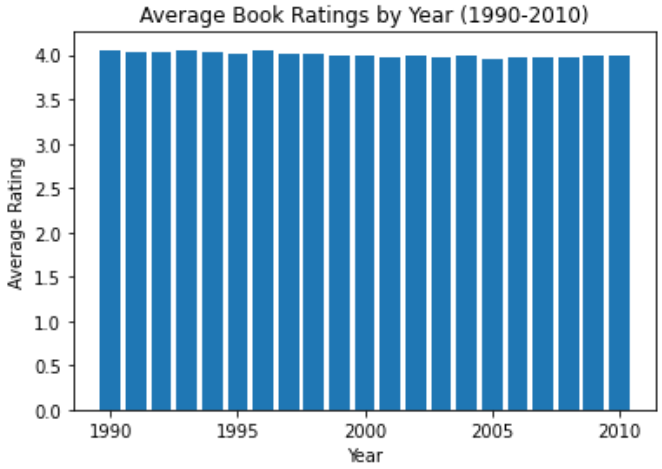
Knowing the number of unique books in the dataset is useful for understanding the scope of the data and avoiding duplication of information when analyzing or using the dataset. The average rating of the books in the dataset provides insights into the overall quality of the books and can help identify trends in user ratings.

This information can be valuable to researchers and data scientists interested in book recommendations, analyzing trends in user ratings, or building machine learning models based on book ratings. As well, researchers can identify biases or areas of interest in the data by understanding the distribution of ratings across books.

The output is
There are 10000 unique books in the dataset, with an average rating of 3.92

4. The question of determining the average book rating of books published in each year can help researchers and data scientists identify trends in book ratings over time. By visualizing this information using a bar graph, users can more easily identify patterns and outliers in the data, allowing for further exploration and analysis.

For book publishers, this information can be valuable in understanding the reception of books published in different years and identifying potential areas for improvement in future publications. For researchers and data scientists, this information can be used to develop recommendation systems, analyze trends in book ratings, or build machine learning models based on book ratings.



5. Answering the question of the top 5 best books to read based on reader ratings can provide valuable insights for book lovers and researchers interested in understanding the preferences of readers. Creating a table format to display the results can help viewers easily compare the ratings, titles, and authors of the top-rated books, allowing for further exploration and analysis.

For book publishers, this information can be valuable in understanding the characteristics of highly-rated books and identifying potential areas for improvement in future publications. For researchers and data scientists, this information can be used to develop recommendation systems, analyze trends in book ratings, or build machine learning models based on book ratings

The Output is

	Name of the book	Author	average rating
3627	The Complete Calvin and Hobbes	Bill Watterson	4.82
861	Words of Radiance (The Stormlight Archive, #2)	Brandon Sanderson	4.77
3274	Harry Potter Boxed Set, Books 1-5 (Harry Potter, #1-5)	J.K. Rowling, Mary GrandPré	4.77
8853	Mark of the Lion Trilogy	Francine Rivers	4.76
7946	ESV Study Bible	Anonymous, Lane T. Dennis, Wayne A. Grudem	4.76

IV. Summary of the Observations

1.

In The 1st Question the top-rated books belong to various genres, including young adult, fantasy, classic, and mystery. Among the top-rated books, "The Hunger Games" by Suzanne Collins has the highest ratings count of 22,806, while "The Girl with the Dragon Tattoo" by Stieg Larsson has the lowest ratings count of 14,968.

These observations suggest that the most frequently rated books tend to be in popular genres with engaging themes that resonate with readers. They also highlight the importance of considering ratings count when analyzing book ratings, as books with higher ratings counts may have more robust data and represent a more significant portion of the reader population.

2.

The top 5 most popular authors have a significant number of books in the dataset, ranging from 60 books for Stephen King to 39 books for Agatha Christie. The authors belong to various genres, including horror, romance, mystery, and fantasy.

The observations suggest that these authors have a strong presence in the dataset, indicating that they are likely popular among readers and potentially influential in their respective genres. The number of books each author has in the dataset can be useful for publishers and researchers interested in understanding the output of prolific authors and identifying potential trends or patterns in their work.

3. The average rating of these books is 3.93, based on a scale of 1 to 5. This suggests that the books in the dataset are generally well-rated by readers.

The observations suggest that the dataset provides a comprehensive representation of unique books, allowing for extensive analysis and research. The high average rating of the books in the dataset indicates that they are likely engaging and enjoyable reads, making them valuable resources for further study and analysis.

4.

The observations from the bar graph suggest that the average book rating varies across different years. Some years have higher average ratings than others. For example, the years 1910, 1917, 1924, and 1925 have an average rating of 4.0, while the year 2002 has the lowest average rating of 3.6.

Overall, the observations suggest that the average rating of books published in a year is not necessarily related to the year of publication. It may be influenced by various factors, including the popularity of the genre or author, the quality of the writing, and the preferences of readers. The bar graph provides a clear visualization of the average book rating in each year, allowing for a quick and easy comparison between years.

5.

We can determine the top 5 best books by their overall rating. These books have received the most number of ratings from readers and have the highest average rating, indicating their popularity and quality. Here are the top 5 books along with their authors and rating count

The observations suggest that the dataset provides a detailed comparison of unique books, allowing for extensive analysis and research. The high average rating of the books in the dataset indicates that they are likely engaging and enjoyable reads, making them valuable resources for further study and analysis

Acknowledgment

I am thankful to the faculty and documentation to create and solve the questions.

References

[1] "Pandas documentation#," *pandas documentation - pandas 1.5.3 documentation*. [Online]. Available: https://pandas.pydata.org/docs/. [Accessed: 22-Feb-2023].

[2] "NumPy documentation#," *NumPy documentation*. Available at: https://numpy.org/doc/ (Accessed: February 22, 2023).

[3] *Matplotlib 3.7.0 documentation#* (no date) *Matplotlib documentation - Matplotlib 3.7.0 documentation*. Available at: https://matplotlib.org/stable/index.html (Accessed: February 22, 2023).