ES 114 - Data Narrative - I

Hrriday Ruparel
22110099
Freshman, Computer Science and Engineering
Indian Institute of Technology
Gandhinagar, India
hrriday.ruparel@iitgn.ac.in

***Abstract*—This is an inferential and statistical data analysis report on a dataset containing 10,000 books and their reviews presented on *Goodreads* website. The data analysis explores the data initially in a time-domain, then from a statistical standpoint and draws some conclusions based on the inferences and observations made using data analysis and data visualization tools, supported by Python.**

I. Overview of Dataset

The dataset used in this data narrative is presented on a GitHub page. The dataset includes 6 million ratings and reviews of 10,000 books extracted from *Goodreads* website. *Goodreads* is a social book cataloguing website that contains ratings, quotes and reviews of a plethora of books.

The dataset on the GitHub page mainly contains 5 important files relevant for our discussion.

1. **books.csv** – This comma-separated-values (.csv) file contains important details like year of publication, ratings, number of ratings, language, book IDs, etc. of 10,000 books.
2. **ratings.csv** - This .csv file consists of ratings of books sorted by time and book IDs.
3. **to_read**.**csv** – This .csv file consists of book IDs marked under 'to read' by users and is additionally sorted by time.
4. **book_tags.csv** – This .csv file consists of user-defined tags/genres for books and along with book IDs.
5. **tags.csv** – This .csv file consists of tag IDs of books translated from tag names.

The files that have been predominantly used for this data narrative include **books.csv**, **to_read**.**csv**, **book_tags.csv** and **tags.csv**.

II. Scientific Questions and Hypotheses

Some useful scientific questions and hypotheses for the data analyses are as follows:

1. *Determine how the number of books published over the years has changed. Also, state the inferences drawn by plotting number of books published against the years. Use simple statistical measures to refine the plot.*
2. *Work on the refined data from the above question to draw conclusions from the average ratings of the books. Does the average ratings of the books follow a particular trend? If so, display the trend.*
3. *If you were to suggest a book to someone to read, based on what you have found out from the dataset, what would you recommend? Is it the books assigned top priority in the 'best books IDs' column of the* ***books.csv*** *file or something else?*
4. *Compare the result of the previous question with the inferences drawn from the* ***to_read.csv*** *file. Does the choice suggested by you match the choice suggested by the entire community of reviewers?*
5. *If not a book, what would be the genre of the book that you would suggest someone else to read? Does the major genres of the books match the genre of the book suggested by you?*

III. Details of Python Libraries and Functions

The data narrative has used the various data analysis and data visualization tools supported by Python. These include:

1. **NumPy** – NumPy, also known as Numeric Python, is a Python library provides tools to deal with arrays and large datasets. This library is imported into Python as **np**.
2. **Pandas** – Pandas is Python library that allows the user to create table-like structures, organise data, perform operations on the data and compute various important statistical measures. Pandas has imported as **pd**.
3. **Matplotlib** – Matplotlib is a python library that provides the facility to visualize the data by providing tools for plotting graphs. The **pyplot** functionality of this library has been imported as **plt**.

Each of these libraries have been extensively used for answering scientific questions stated in the previous section. Some of the notable functions of these libraries used for data analysis and visualization are:

1. **np.log10()b –** This function applies logarithm of base 10 to all the entries of an array.
2. **np.array()** – This function helps to create an array.
3. **np.add()** – This function helps to perform element-wise addition of two or more arrays.
4. **pd**.**DataFrame()** – This function helps to create a new DataFrame.
5. **pd.read_csv()** – This function helps convert .csv files to Pandas DataFrame.
6. **pd**.**DataFrame.drop()** – This funtion helps to remove rows or columns.
7. **pd**.**DataFrame.assign()** – This function helps to add new columns to existing DataFrame.'
8. **pd**.**DataFrame.sort_values()** – This function helps to sort the values of a DataFrame on the basis of the arguemnt passed.
9. **pd**.**DataFrame.describe()** – This function computes various statistical measures of a DataFrame.
10. **pd.rename()** – This function helps to rename the column or row labels of a DataFrame.

11. **plt.plot()** – This function helps to create plots of variousb types like bar, hist, scatter, pie, density, etc.
12. **plt.show()** – This function is used to view the plotted graph on the the figure window.
13. **plt.subplots()** – This function helps to work with subplots on a single figure. It provides tools for customisation of subplots and their layouts.
14. **plt.title()** – This function sets the title for the graphs.
15. **plt.xlabel()** & **plt.ylabel()** – These function help to label the x-axis and y-axis respectively.
16. **plt.xticks()** & **plt.yticks()** – These functions help to customise the ticks on the x-axis and y-axis respectively.
17. **plt.xlim()** & **plt.ylim()** – These functions help to set the limit/range of values plotted on the x-axis and y-axis respectively.
18. **plt.grid() –** This function displays the grid on the graphs.
19. **plt.legend()** – This function displays trh legend on the graph.

These are the major functions and methods used from the the imported Python libraries.

IV. ANSWERS TO SCIENTIFIC QUESTIONS AND HYPOTHESES

For the first question, the aim was to plot the number of books published in a particular year and draw inferences from it. In order to so, I firstly plotted the number of books published in a century. This was done to get a complete overview over the dataset from a time domain. The following plot was obtained.

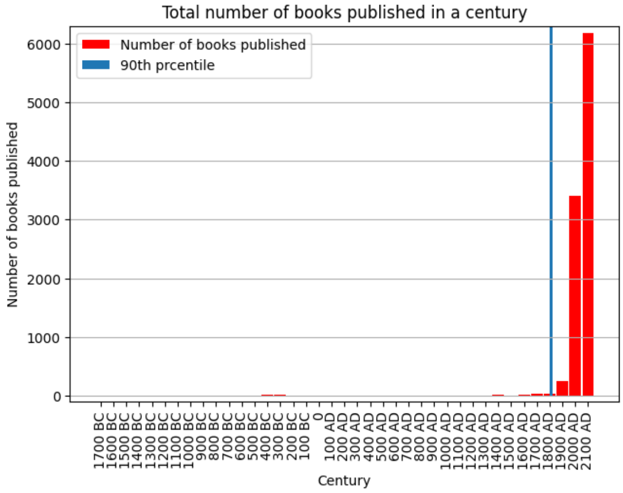

Fig. 1a – Bar graph depicting the number of books published in a century. It also shows the 90th percentile data point in terms of years.

Fig. 1a depicts that the majority of the books were published in 19th, 20th and 21st century and it increased monotonically. From Fig. 1a, it is also evident that for 90% of the timeline, the number of books published were not significant. Here, the 90th percentile data point was calculated using the equation:

$$Percentile = \frac{nth\ data\ point}{Nth\ data\ point} * 100, where\ n < N$$

This leads to an important observation that any dataset may have some extreme values called outliers that do not match the general trend. In order to shun away these outliers in the time-domain, I excluded the outliers in the discussion and the considered only the part of the graph beyond the 90th percentile data point.

This led to a more informative, detailed and predictable graph as follows:

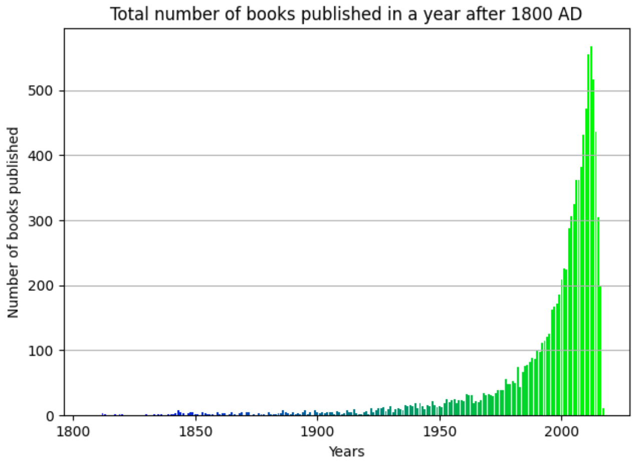

Fig. 1b – Bar graph depicting the number of books published in a year after 1800 AD.

Fig. 1b is obtained by simply using the 90th percentile statistical measure to zoom into the finer details of the graph and interpret it in much better way. The graph shows that a maximum of 567 books were published in the year 2012. Not only that, the graph seems to take shape of a beautiful distribution by the name Beta Distribution, but not necessarily exact in the sense.

With sufficient number of data points still available after plotting Fig. 1b, I decided to use the only the data points used in Fig. 1b for further discussions as these data points were more recent and, hence, relevant. In this way, I refined the data by using the 90th percentile data point.

For the second question, the aim was to plot the average ratings of books readily provided in the data set against the so-called 'best book IDs' and draw inferences from it. This was included as a point for discussion because whenever someone mentions about reviews and ratings of books, the human mind easily diverts to the numbers that may present false or incomplete information.

The initial thought or hypothesis was that the graph (scatter plot) would show a downward trend with increasing 'best book IDs' as one would expect the average rating to drop accordingly. But, the graph turned out to behave weirdly.

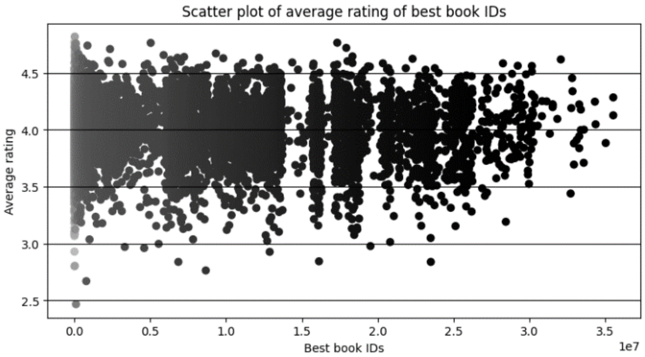

Fig. 2a – Scatter plot of average rating of books against best book IDs

In Fig. 2a, the colour coding has been done in such way that the light grey colours represent the points that must be on the higher side of the average rating and the dark black-coloured points represent the points that must be on the lower side of the average rating. But, the graph poses many anomalies as some of the dark coloured points (the books that come lower in the best books IDs) have average ratings greater than the ones in lighter grey (the books that come higher in the best books IDs). Also, the plot seems *discontinuous* for certain range of points. It also seems that the plot progressively becomes more and more *separate* and *distinctively visible* for higher values of best book IDs.

In order to have a closer look at the graph, a graph consisting of density variation against average ratings of books was plotted and it produced the following result.

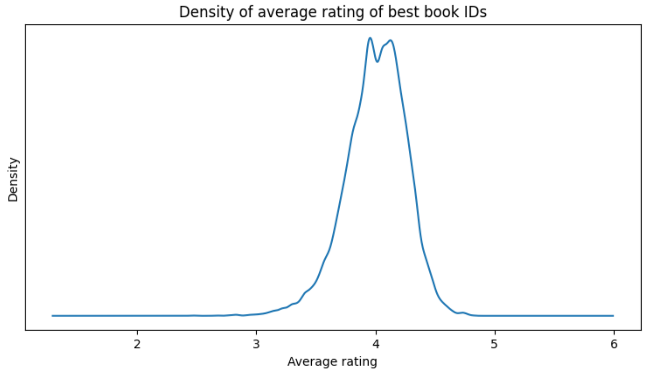

Fig. 2b – Plot depicting density variation against average ratings

The Fig. 2b almost resembles a Normal (Gaussian) Distribution and shows that the mode and coincidentally the median of the distribution is approximately 4. Hence, the average rating of majority of the data is 4.

On even closer analysis, it was found out that the average ratings were nothing but the weighted sum of the individual ratings given by the equation below:

$$W = \frac{1 * r_1 + 2 * r_2 + 3 * r_3 + 4 * r_4 + 5 * r_5}{r_1 + r_2 + r_3 + r_4 + r_5}$$

Where W is the weighted sum of the ratings and r_1, r_2, r_3, r_4, r_5 denote the respective ratings from 1 to 5.

This way of measuring and determining best book IDs is incorrect as it only considers the quality of the measure and no the quantity of the measure. The quantity of the measure that is mentioned here refers to the total number of ratings available for a particular book. This is shown in the plots given below:

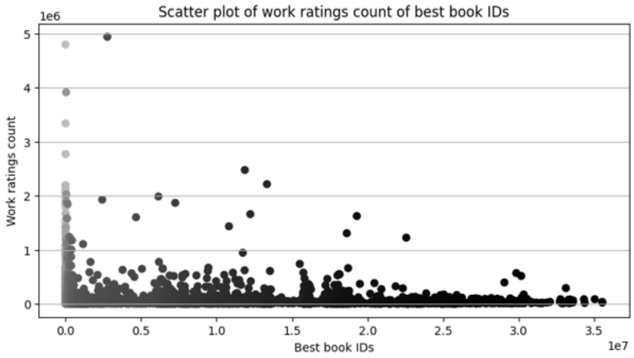

Fig. 2c – Scatter plot depicting the number of ratings (work) against best book IDs

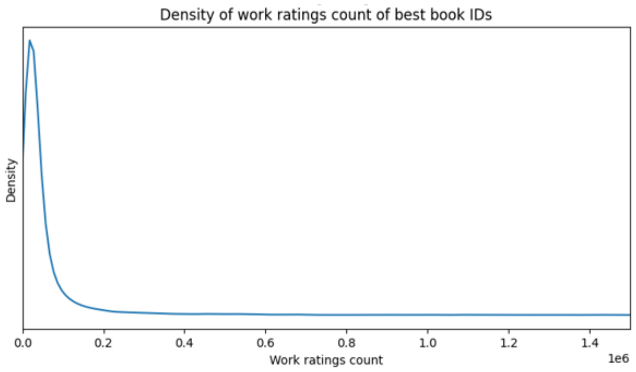

Fig. 2d – Plot depicting density variation against number of ratings (work)

Both the graphs Fig. 2c and Fig. 2d show that the so-called best books according to the data set have the most number of ratings which seems true and intuitive.

These plots do not provide the entire picture for one to derive anything useful like a trend or so. Hence, the only major takeaway from the second question is that there is a necessity for a better statistical quantity to rate a book.

For the third question, there is a necessity to develop a tool that rates the book in a better way as compared to the best book IDs and weighted average of ratings.

Since both the qualitative and quantitative aspects of the data do not reveal the complete picture, it is necessary to incorporate individual contributions and then derive the entire picture. This can be temporarily resolved by using the same concept of weighted averages, but applying it to both the already derived average ratings and slightly modified (applied logarithm of base 10) quantity of number of ratings.

$$Credit = \frac{w_1 * x_1 + w_2 * x_2}{w_1 + w_2}$$

Where *Credit* is the weighted average of x_1, x_2 (average rating of a book and number of ratings respectively) with user-defined weights w_1, w_2. The quantity *Credit* obtained was then *range-shifted* by incorporating its value in a simple sigmoid function of the form:

$$f(x) = \frac{1}{1 + e^{-x}}$$

Hence, the final *Credits* of each book was obtained and plotted to observe the change in the ratings.

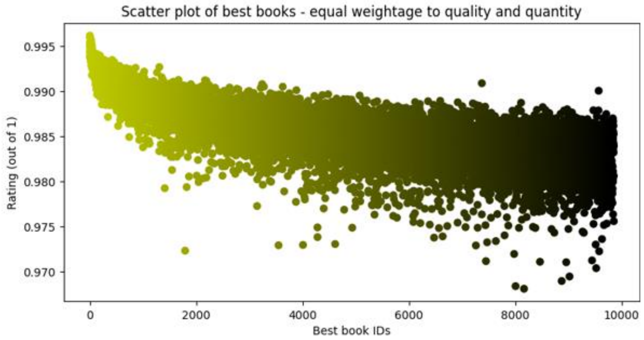

Fig. 3a – Scatter plot depicting rating (out of 1) against book IDs with w1 = 0.5 and w2 = 0.5

Logarithm of base 10 is applied on the data set consisting of number of ratings in order to bring the quantities in approximately the same range. Plots for various other values of w_1, w_2 are as follows:

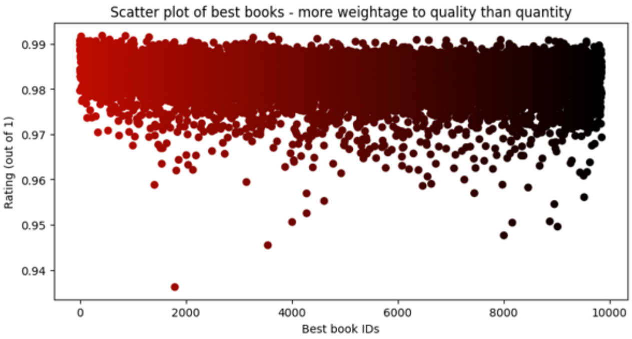

Fig. 3b – Scatter plot depicting rating (out of 1) against book IDs with w1 = 0.9 and w2 = 0.1

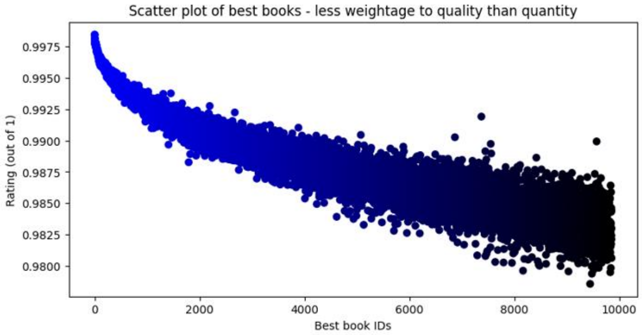

Fig. 3c – Scatter plot depicting rating (out of 1) against book IDs with w1 = 0.9 and w2 = 0.1

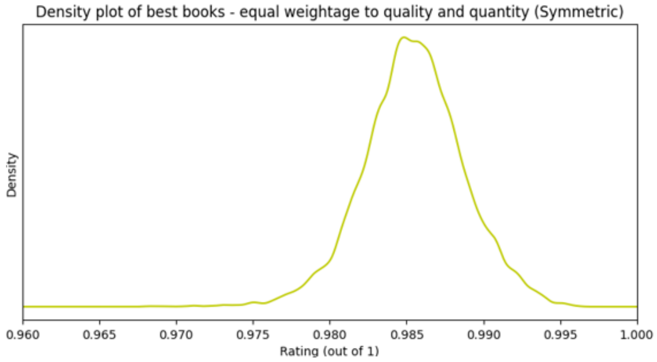

Fig. 3d –Plot depicting density variation against rating (out of 1) w1 = 0.5 and w2 = 0.5

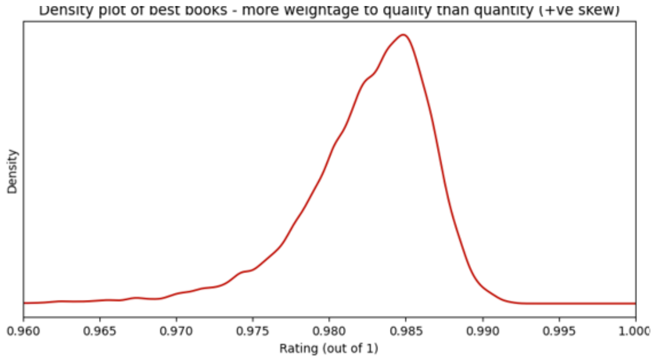

Fig. 3e –Plot depicting density variation against rating (out of 1) w1 = 0.1 and w2 = 0.9

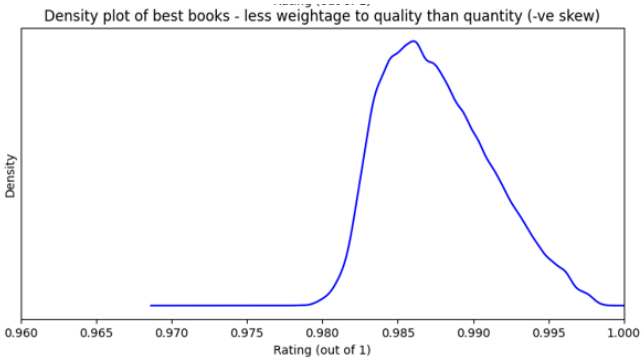

Fig. 3f –Plot depicting density variation against rating (out of 1) w1 = 0.9 and w2 = 0.1

The graphs (Fig. 3a – Fig. 3c) depict how ratings change by changing the weights and Fig. 3b and Fig. 3e fairly resemble the previous graphs Fig. 2a and Fig. 2b respectively for very high values of weights given to average ratings.

Another set of observations made from the density variation plots is that the skewness of the graph changes as we vary the weights. The graph is positively skewed in case of increasing the average ratings' weight and is negatively skewed in case of increasing the number of ratings' weight.

Taking the case for which both the weights are equal, one can conclude that the book with the best rating (according to the newly devised algorithm) would be the book named *Harry Potter and The Philosopher's Stone* which is the third best book according to the best book IDs. Hence, I would preferably recommend the book that the algorithm suggests than blindly follow the best book IDs.

For the fourth question, our aim is to compare whether *Harry Potter and The Philosopher's Stone* is the most important 'to read' book or otherwise.

For this, it is will be necessary to plot the number of times the users have referred to read the book against the book ID and the graph obtained is as follows:

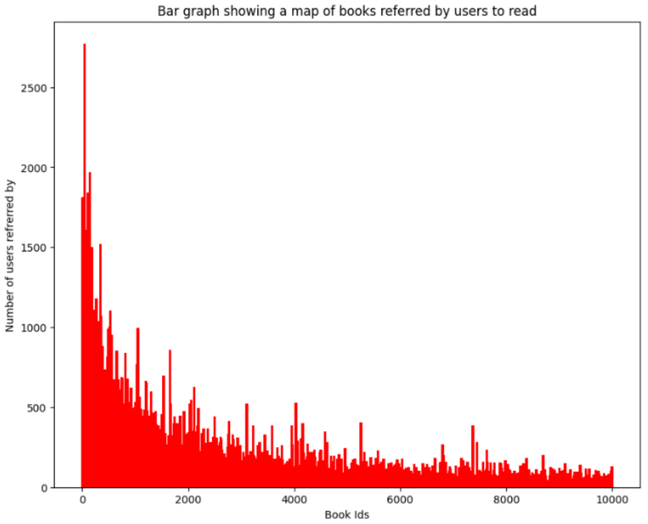

Fig. 4 – Bar graph depicting number of times a particular book has been referred to be read.

For the Fig. 4, one can conclude that the number of 'to read' suggestions decreases with increasing book IDs.

Surprisingly, in this case, the most referred to book turns out to be *The Book Thief* unlike the prediction of *Harry Potter and The Philosopher's Stone*.

For the fifth question, the main aim is to plot the graph of various types of tags and draw inferences based on the majority tags commonly used. One observation noted was that the tag 'to_read' was widely used and hence, other data points were not clearly visible while plotting the tag. Therefore, the 'to_read' tag has been omitted. Also, the for the sake of plotting, I have considered only those tags which have been used more than 500000 times.

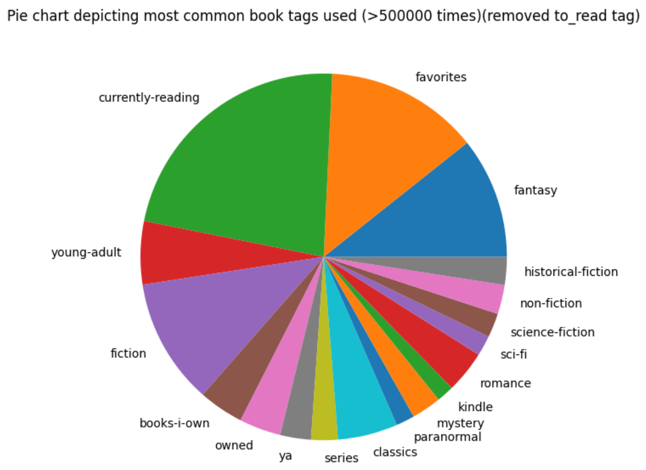

Fig. 5 – Pie chart depicting common book tags used for tags used more than 500000 times (removed the 'to read' tag)

From the Fig. 5, one can observe that one of the most common genres is Fantasy followed by Fiction. Hence, if I were to suggest someone a genre to pick, it would have been Fantasy.

Among these, the *Harry Potter and The Philosopher's Stone* book associates with 13 out of 18 tags including the most popular one – 'currently-reading'.

V. Summary of the Observations and Learnings

1. Define a context/background before beginning data analysis as it may keep us oriented towards our goal. It is essential to sometimes clean the data before using it and remove any outliers for the sake of general case.
2. It is important to analyze statistical measures and computations carefully as it may present false or incorrect information. One needs to be critical and vigilant to catch errors or flaws in a data set and not blindly follow any data.
3. While analyzing a particular data point or property of a data set, one must take into consideration the effect of all the factors on the data point.

VI. Unaswerable Questions

1. What does 'books_count' in books.csv file signify?
2. What is the difference between 'work_ratings_count' and 'ratings_count' in books.csv file?

Acknowledgment

This is to thank Professor Shanmuganathan Raman for proving this opportunity to work on a dataset and perform data analysis and visualization in order to extract some useful information.

References

1. 'NumPy documentation,' accessed Jan 23, 2023, https://numpy.org/doc/1.24/
2. 'Pandas documentation,' accessed Jan 23, 2023, https://pandas.pydata.org/docs/
3. 'Matplotlib documentation,' accessed Jan 23, 2023, https://matplotlib.org/stable/index.html
4. Youssef Hosni, '20 Pandas Functions for 80% of your Data Science Tasks,' accessed Jan 23, 2023, https://medium.com/gitconnected/20-pandas-functions-for-80-of-your-data-science-tasks-b610c8bfe63c
5. 'goodbooks-10k,' accessed Jan 23, 2023, https://github.com/zygmuntz/goodbooks-10k
6. Brendan Artley, 'Matplotlib Color Gradients', accessed Jan 23, 2023, https://medium.com/@BrendanArtley/matplotlib-color-gradients-21374910584b