Data Narratives-1 (February 2023)

Birudugadda Srivibhav -22110050, Computer Science and Technology Department, Prof. Shanmuga R, IIT Gandhinagar

***Abstract*-This datanarratives project analyses the Goodreads' Goodbooks-10k dataset, which contains information on 10,000 books. The analysis focuses on book ratings, publication years, and author popularity. Visualisations are used to highlight trends and patterns within the data. The project aims to provide insights into the reading habits and preferences of Goodreads users.**

I. Aim

To study and write a data narrative on good books-10k dataset.

II. Overview of the dataset

The good books-10k dataset:

This dataset contains six million ratings for ten thousand popular (with most rating books)

Here's an overview of the dataset.

1. **books csv:** This is a CSV file containing book data. A unique ID represents each book, and the data includes the book's title, author, average rating on Goodreads, language_code etc.
2. **ratings csv:** This a CSV file containing rating data. Each rating includes the user ID, book ID, and the rating the user gives on a scale from 1 to 5.
3. **book_tags csv:** This file contains information about tags that users on Goodreads have applied to the books in the dataset. It includes the following columns: goodreads_book_id, tag_id, and count.
4. **tags csv:** This file contains information about the tags themselves rather than those applied to books. It includes the columns tag_id and tag_name.
5. **to_read csv:** This file contains information about books that users on Goodreads have added to their "to-read" lists but have not yet read or rated. It includes the columns user_id and books_id.

III. scientific questions/hypothesis

1. **Question:** Find the correlation between the number of reviews and the number of ratings?
   Hypothesis: Books with more reviews tend to have higher average ratings.
2. **Question:** Which language book has the highest average rating.?
   Hypothesis: The language book with the highest average rating may not be the most popular.
3. **Question:** Do the books with more ratings tend to go to the top of to_ read section?
4. **Question:** Find the top 50 popular book titles and their publication years. In which decades most the popular books are published ?
5. **Question:** Find the top 10 years in which more number of books are published?

IV. details of libraries and functions

The libraries and modules used are:

[1] Numpy library - A Python library for scientific computing that provides tools for working with arrays and matrices, as well as mathematical functions and linear algebra operations.

[2] Pandas library- A Python library used for data manipulation and analysis. It provides data structures for efficiently storing and accessing large datasets and tools for cleaning, transforming, and analysing data.

[3] Matplotlib library- A Python library used for creating static, animated, and interactive visualisations in Python. It provides various plotting options and customisation tools to create publication-quality figures.

[4] Requests library- A Python library which is widely used HTTP library that allows Python programs to send HTTP requests to web pages and receive response data from them. It interacts with web services and APIs to retrieve, upload, and update data.

[5] io module - It provides a way to handle streams of data. The module contains several classes that allow you to work with streams of different types, such as text or binary data. It also provides a set of functions for working with file-like objects.

[6] Seaborn library- A Python library based on Matplotlib that provides additional visualisation tools, particularly for statistical data analysis. It provides a high-level interface for creating complex plots with minimal code.

THE FUNCTIONS USED ARE:

pd.merge - A function in Pandas that combines two or more data frames based on a common column.

sort_values - A function in Pandas that sorts a dataframe by one or more columns.

set_index - A function in Pandas used to set one or more columns as the index of a dataframe.

value_counts - A function in Pandas used to count the occurrence of unique values in a column.

drop_duplicates - A function in Pandas removes duplicate rows from a dataframe.

isin - A function in Pandas that checks if a value is in a list or array.

plt.bar - A function in Matplotlib used to create a bar chart.

plt.show - A function in Matplotlib used to display a plot.

sns.barplot - A function in Seaborn used to create a bar chart with additional styling options.

np.arange - A function in Numpy creates a range of numbers with a specified start, stop, and step.

pd.DataFrame - A class in Pandas used to create a dataframe from a dictionary or array of data.

pd.read_csv - A function in Pandas used to read a CSV file into a dataframe.

pd.groupby - Pandas groupby is used for grouping the data according to the categories and apply a function to the categories. It also helps to aggregate data efficiently. Pandas dataframe. groupby() function is used to split the data into groups based on some criteria.

sns.scatterplot - The scatterplot function in seaborn is used to create a scatter plot that displays the relationship between two variables. It is useful for identifying data patterns and trends and can help reveal any correlations between the two variables.

requests.get - The get() function is one of the most commonly used functions in the requests module. It sends a GET request to the specified URL and returns a response object.

V. ANSWERS TO THE QUESTIONS

Question1: Find the correlation between the number of reviews and the number of ratings?
Hypothesis: Books with more reviews tend to have higher average ratings.

Answer:

In the analysis, I used the pandas library in Python to find the correlation between the number of reviews and the number of ratings. Specifically, I have used the corr() method to calculate the correlation coefficient between the two variables. The correlation coefficient is a statistical measure that indicates the strength and direction of the linear relationship between two variables.

I examined the correlation between the number of reviews and ratings on Goodreads. I used the books dataset containing information about various Goodreads books, including the number of reviews and ratings for each book.

First, I plotted a scatter plot using Seaborn library to visualise the relationship between the number of reviews and the number of ratings. The scatter plot clearly shows that there is a positive linear relationship between the two variables, where books with higher number of reviews tend to have higher number of ratings as well.

Later, I loaded the books dataset into a Pandas dataframe and then used the Pandas corr() function to calculate the correlation coefficient between the number of reviews and ratings. The resulting correlation coefficient was 0.807, which indicates a strong positive correlation between the two variables.

The strong positive correlation between the number of reviews and the number of ratings indicates that readers who are more likely to rate a book are also more likely to review it. This suggests that the number of reviews and ratings can be used to indicate a book's popularity and overall quality. Additionally, this finding can be useful for authors, publishers, and marketers who are interested in promoting their books and understanding their audience.

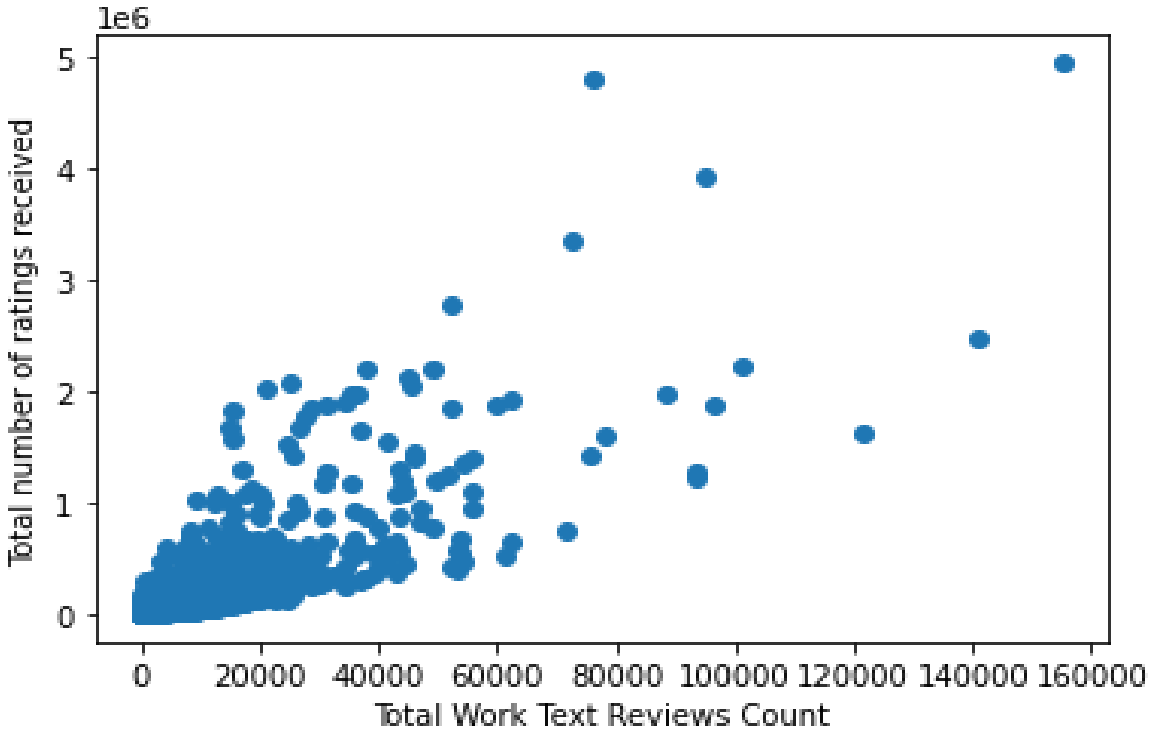

Fig-1: This figure shows the scatterplot between the number of reviews to that number of ratings.

To further support my hypothesis, let us plot a scatterplot between the number of reviews and average rating .

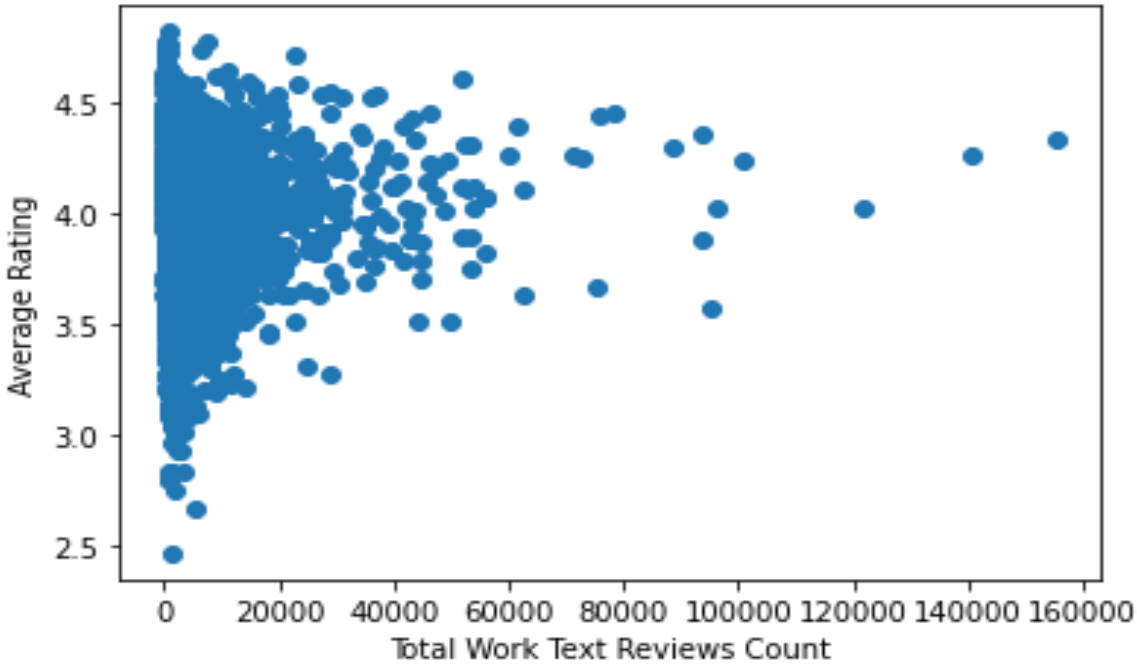

Fig-2: This figure shows the scatterplot between the number of reviews and average rating.

The graph shows that Books with more reviews need not have a high average rating.
Also the correlation coefficient using corr() function is obtained to be 0.007481118668792918, which is indeed very lower. This indicates a very weak correlation between the variables.

Question2: Which language book has the highest average rating.?
Hypothesis: The language book with the highest average rating may not be the most popular.
Answer:
To calculate the average rating for each language book, we first needed to extract the relevant columns from the books dataset, which included the book ID, language code, and the book's average rating. We then used the Pandas library to group the data by language code and calculate the mean rating for each language.

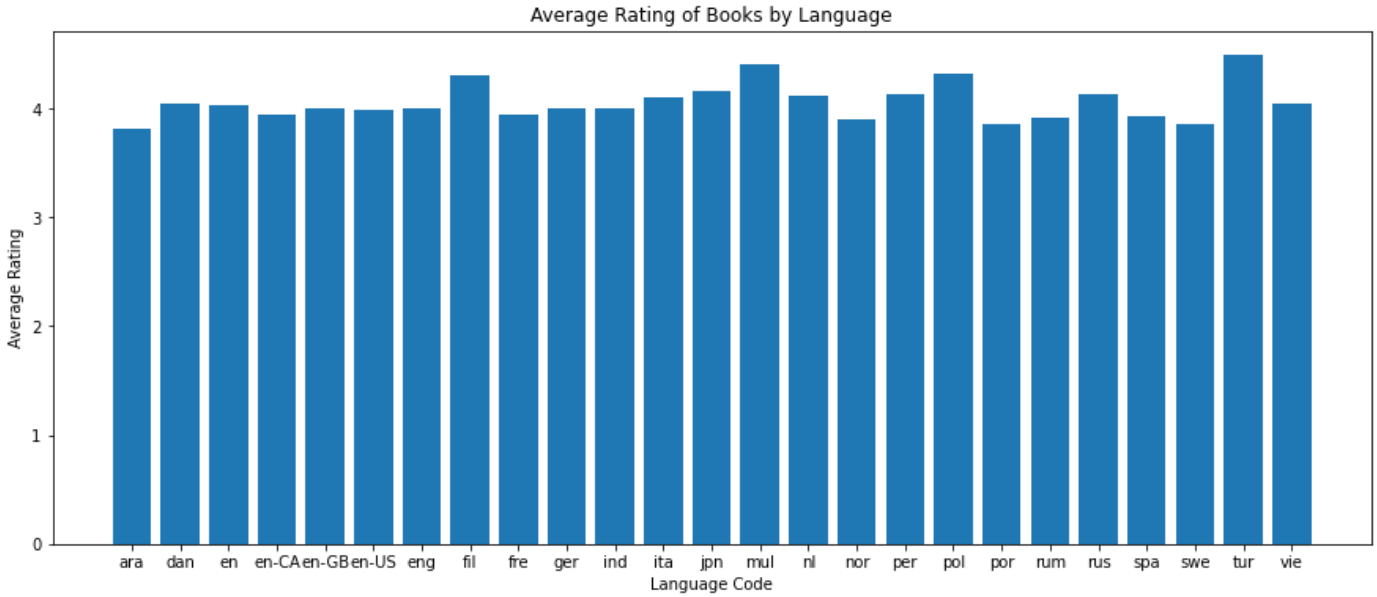

Fig-3: This figure shows the barplot between the language code and average rating.

The resulting dataset showed the average rating for each language, allowing us to compare how books in different languages were rated on average. It is important to note that this analysis only looked at books with ratings and did not consider other factors that could impact the rating of a book, such as genre or publication date.

Our overall analysis showed that turkish language have high average rating.

But now the question arises: Are the turkish language books also popular?

To find that, plot a scatterplot between the number of ratings by average rating for each language. The graph is as shown below.

Again, the turkish language books have a high average rating of about 4.4, but more number of ratings are for english language books. That implies we are getting an average rating of about 4.1 for english language books for a much more larger count of ratings. This proves that the english language has more popularity in language books that compared to turkish language books.

Hence my hypothesis that, the language book with the highest average rating may not be the most popular is correct.

We must consider both, The average rating and also the total number of ratings given by users to determine the popularity of a book.

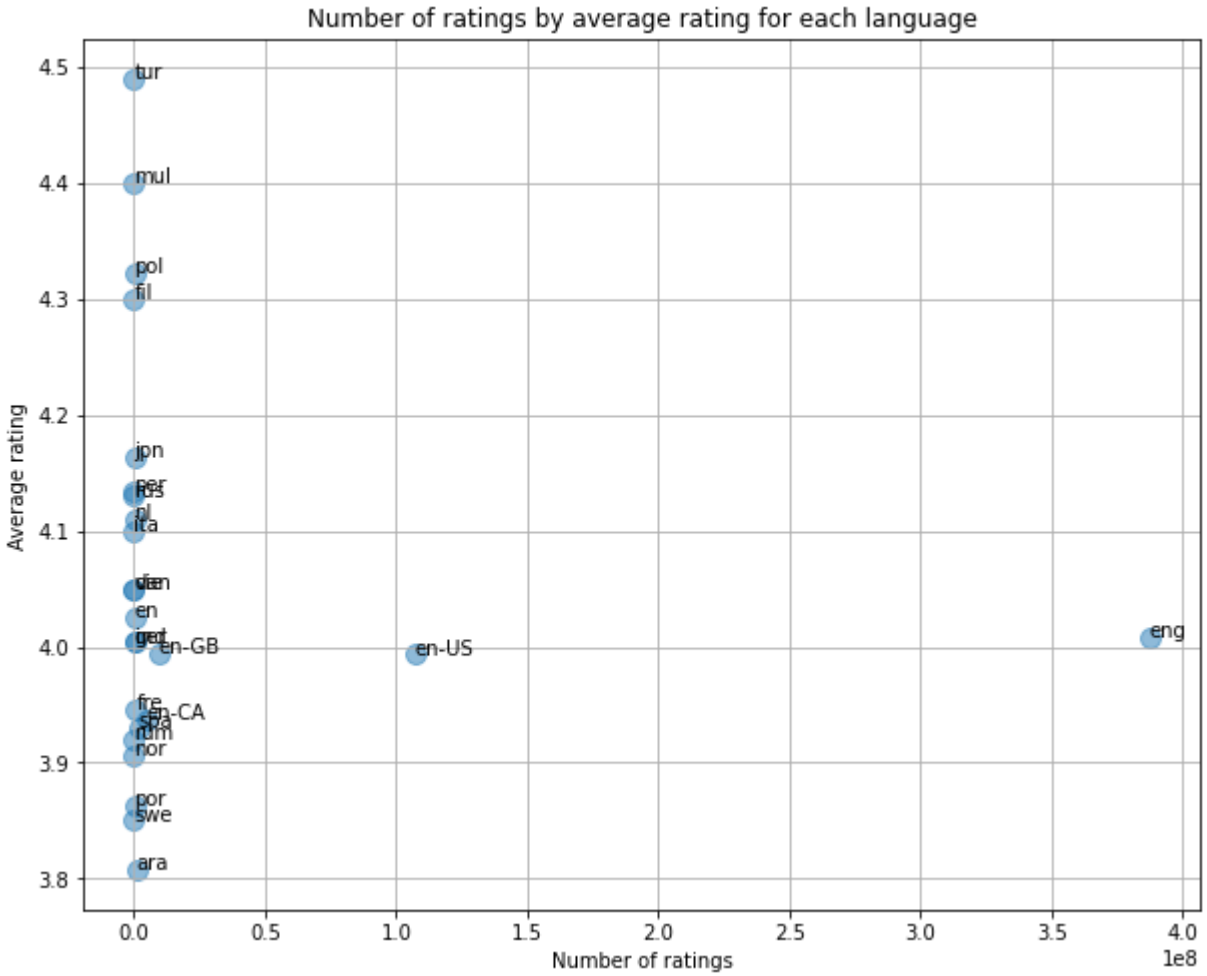

Fig-4: This figure shows the scatterplot between the number of ratings by average rating for each language.

Question3: Do the books with more ratings tend to go to the top of to_ read section?
Answer:
To answer the question of whether the top books in the "to-read" list are the same high-rated books in the "books" CSV or not, we can use the data from both CSV files and perform some data analysis.

Let us do the analysis for the top 500 books in to_read section and the top 500 high rated books in books dataset.

The analysis aims to investigate whether the top 500 books in the "to-read" list are also high-rated in the "books" CSV. The analysis involved comparing the book IDs of the top 500 books based on "to-read" counts in the "to-read" CSV with the book IDs of the top 500 books based on rating counts in the "books" CSV. If there were any common book IDs between the two sets, it would suggest that the top 500 books in the "to-read" list are also high-rated books in the "books" CSV. The analysis was performed using Python code, and the results showed whether the top 500 books in the "to-read" list were also high-rated in the "books" CSV. This analysis could be useful for understanding the relationship between popular books based on "to-read" counts and popular books based on rating counts.

First, I extracted the top 500 books from the "to-read" CSV based on the number of "to-read" counts. I then extracted these top 500 books' book IDs and searched for them in the "books" CSV.

Next, I extracted the top 500 books from the "books" CSV based on the number of ratings. I have compared the book IDs of the top 500 books in the "to-read" CSV and the "books" CSV. If there are any common book IDs between the two sets, it means that the top 500 books in the "to-read" list are also high-rated books in the "books" CSV.

I finally concluded that there are 303 books in common between the top 500 books in to_read and the top 500 books based on the rating count.

Question4: Find the top 50 popular book titles and their publication years. In which decades most of the popular books are published?
Answer:
To find the years of publication for top 50 popular books in the Goodbooks-10k dataset, I used the books.csv file along with the ratings.csv file.

I have first loaded the books.csv and ratings.csv files into separate Pandas data frames. I then used the group-by method on the rating data frame to count each book's ratings, sort the resulting counts in descending order, and select the IDs of the top 50 most highly rated books.

Next, I have selected the book data frame rows corresponding to these top-rated books using the ”isin” method to check if the book IDs are in the most_rated_book_ids list. Finally, I displayed the title and publication year columns for these books.

Now I have plotted a bar chart to visualise the same. I created the bar chart using the bar method from matplotlib, with the x-axis representing the publication year and the y-axis representing the number of books published in that year.

Finally, I added labels to the axes and a title to the plot using the xlabel, ylabel, title methods and displayed the plot using the show method.

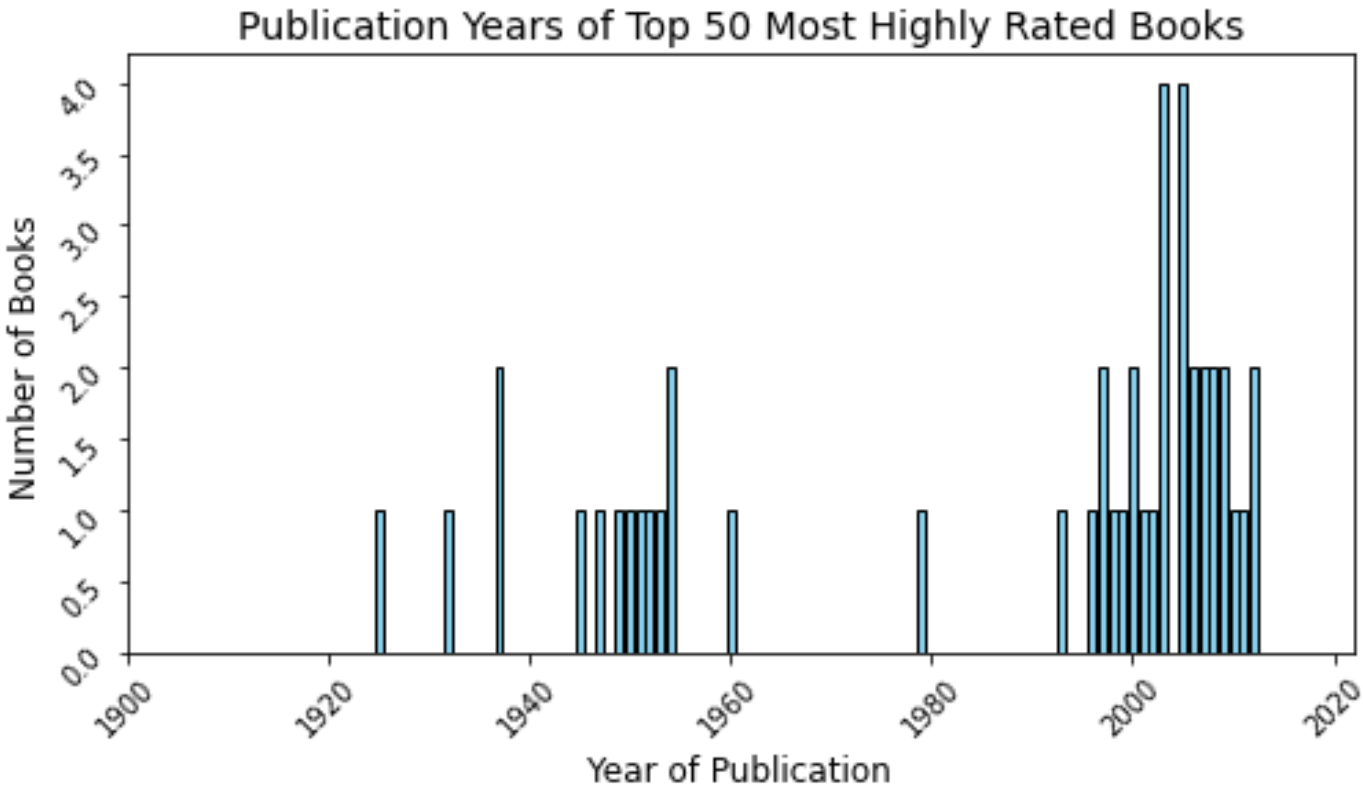

Fig-5: This figure shows the publication Years of Top 50 Most Highly Rated Books.

From the graph, we can clearly see that more bars are concentrated in the 1945-1955 and 1990-2010 regions.
Hence, we can conclude that In the decades 1945-1955, 1990-2000,2000-2010, more popular books were published. Also, 2003 and 2004 each had four highly rated books published, while some others had only one or two. This could be due to factors such as publishing trends, popular genres, or influential authors.

Question5: Find the top 10 years in which more number of books are published?
Answer:
To achieve the above task, I first wrote a code that groups the data in the "books" dataframe by original publication year using groupby function. Then I summed the number of books published each year using the sum function on the books_count column.

The resulting series is then sorted in descending order to get the top 10 years with the most published books. Next, The code filters the books dataframe to include only books published between 2000 and 2020 and groups the resulting dataframe by original publication year. The books_count column is then summed for each year using the sum function, and the resulting series is used to create a pie chart using the pie function from the matplotlib library. The pie chart shows the distribution of the number of books published between 2000 and 2020 for each year.

Also the top ten years in which more number of books are published are:
Top 10 years with the most number of books published:
2011.0: 19925 books
2012.0: 19057 books
2009.0: 18235 books
2010.0: 17685 books
2006.0: 17073 books
2008.0: 16974 books
2013.0: 16481 books
2007.0: 16297 books
2005.0: 16254 books
2003.0: 14244 books

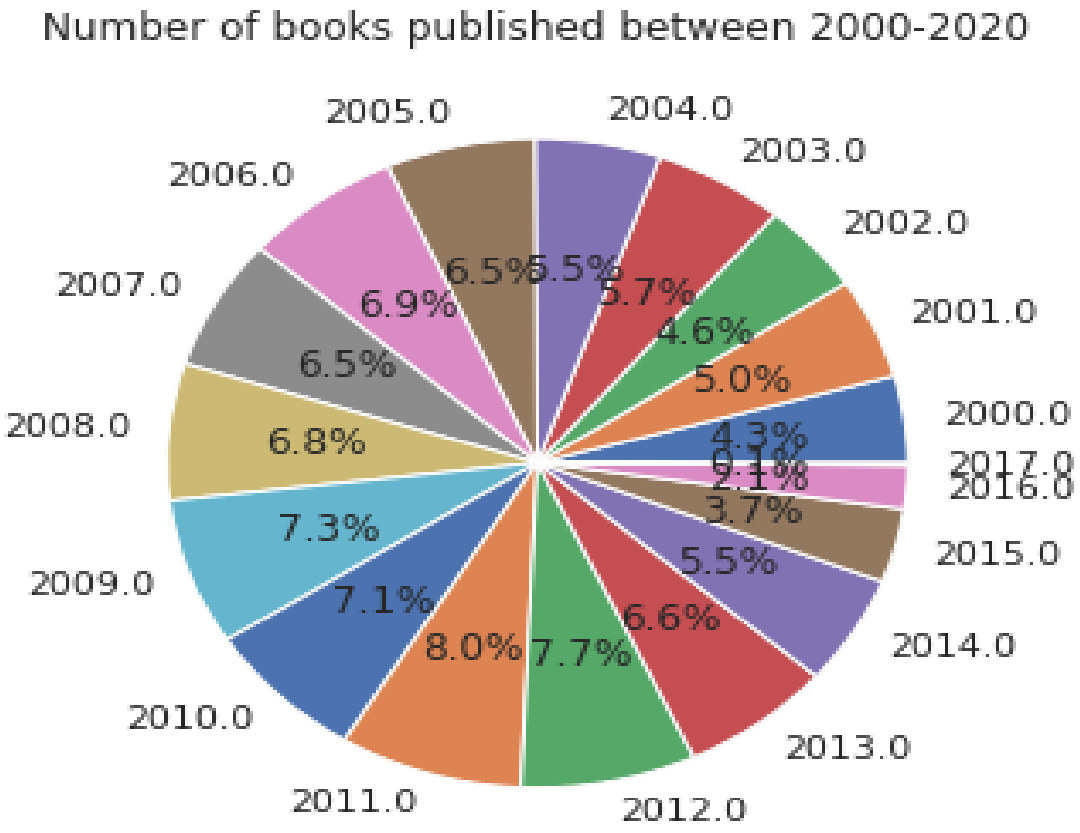

Fig-6: This figure shows a pie chart for the number of books published between 2000-2022

VI. SUMMARY AND OBSERVATIONS

1. Question1:

We have established the corelationship between the number of reviews and the number of ratings, whose value turns out to be 0.8070090183152889
The strong positive correlation between the number of reviews and the number of ratings indicates that readers who are more likely to rate a book are also more likely to review it. This suggests that the number of reviews and ratings can be used to indicate a book's popularity and overall quality. Additionally, this finding can be useful for authors, publishers, and marketers who are interested in promoting their books and understanding their audience.

2. Question2:

First, we compared how books in different languages were rated on an average. Later, we found out that using average rating alone is not a correct way to find the popularity of a language book. Then we used the scatterplot between the number of ratings by average rating for each language. As english language has a very large number of ratings compared to turkish language and a comparable average rating as that of turkish language, we say english is more popular.\

3. Question3:

The analysis aims to investigate whether the top 500 books in the "to-read" list are also high-rated in the "books" CSV. The analysis involved comparing the book IDs of the top 500 books based on "to-read" counts in the "to-read" CSV with the book IDs of the top 500 books based on rating counts in the "books" CSV.
I finally concluded that there are 303 books in common between the top 500 books in to_read and the top 500 books based on the rating count.

4. Question4:

The graph shows that more bars are concentrated in the 1945-1955 and 1990-2010 regions.
Hence, we can conclude that In the decades 1945-1955, 1990-2000,2000-2010, more popular books were published. Also, 2003 and 2004 had four highly rated books published, while others had only one or two. This could be due to factors such as publishing trends, popular genres, or influential authors.

5. Question5:

The aim of the above task is to find the top 10 years in which more number of books were published. I was able to conclude that, as in the question4, where more popular books were published in the 2005-2022 range, I have also found out the maximum number of books were published in 2005-2015.

VII. REFERENCES

1. NumPy: Travis E, Oliphant. "A guide to NumPy", Proceedings of the 7th Python in Science Conference, vol. 1, 2008, pp. 1-7.
2. Pandas: Wes McKinney. "Data structures for statistical computing in Python", Proceedings of the 9th Python in Science Conference, vol. 445, 2010, pp. 56-61.
3. Matplotlib: J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
4. Requests: K. W. Chen, T. A. Kelley, and R. H. W. Hoppe, “Requests: HTTP for Humans™,” 2018. [Online]. Available: http://python-requests.org.
5. io: Python Software Foundation. "io — Core tools for working with streams", Python 3.10.0 documentation, 2021. [Online]. Available: https://docs.python.org/3/library/io.html.
6. J. Waskom, M. Botvinnik, O. Hobson et al., “seaborn: v0.11.2 (November 2021),” Zenodo, 08-Dec-2021. [Online]. Available: https://doi.org/10.5281/zenodo.5737980 . [Accessed: 23-Feb-2023].

VIII. ACKNOWLEDGEMENT

I want to sincerely thank Professor Shanmuga R for his important advice and assistance for the completion of the project. My knowledge and skills have benefited greatly from his experience and commitment. I am grateful for his never-ending inspiration and enthusiasm, which have helped me to flourish in my academic endeavours. I will always be beholden to him for his mentorship, and I sincerely appreciate all the time and effort he has put into me.