ES 114

Probability, Statistics and Data Visualization

Data Narrative I

HARSH KUMAR KESHRI

Roll Number : 22110094
Mechanical Engineering
IIT Gandhinagar

I. OVERVIEW OF THE DATASET

The dataset I have chosen from the GitHub link is books.csv which contains 1,0000 books which also includes their titles, authors, rating publication dates and many other parts. Every row of the dataset contains different columns which include:

1. Book_id: This gives every single book in the dataset a unique position
2. Book_title: The title of the particular book
3. Book_author: The name of the author of the book
4. Book_average_rating: The average rating of the book is out of 5 stars.
5. Book ISBN: The ISBN number for the book (International Standard Book Number)
6. Book ISBN 13: The book has ISBN 13 number
7. Book Publisher: The publisher of the book
8. Book Ratings Count: The number of rating a particular book has received.
9. Book Review Count: The number of reviews a particular book has received.

I. SCIENTIFIC QUESTIONS/HYPOTHESES

According to the dataset books.csv, I have made 5 Scientific/Hypothesis questions which are as follows:

1. What is the probability of getting an English book with a rating above 4?
2. How many books were published in the 90s?
3. Which year has maximum number of books published between 1500 and 2000 and how many books are published in that particular year?
4. How many books do not have isbn or isbn13?
5. Which is the most popular book by average rating?

II. DETAILS OF LIBRARIES AND FUNCTIONS

We have used some python libraries and a lot of functions in my code to get it to run efficiently. Some of them are as given below:

1. Pandas: Pandas: Pandas is a popular library for data manipulation and analysis in Python. It provides data structures like data frames and series, along with functions for reading, writing, and manipulating data [1].

   Some of the functions of the Pandas are as given.
   - pd.read_csv(): This function is used to read data from a CSV file and return it as a data frame.
   - df.groupby(): This function is used to group data in a data frame by one or more columns.
   - df.plot(): This function is used to create various types of plots using the data in a data frame.

2. Matplotlib: Matplotlib is a plotting library for Python. It provides a wide range of functions for creating different types of plots, along with extensive customization options [2].

   Some of the functions used in the Matplotlib are:

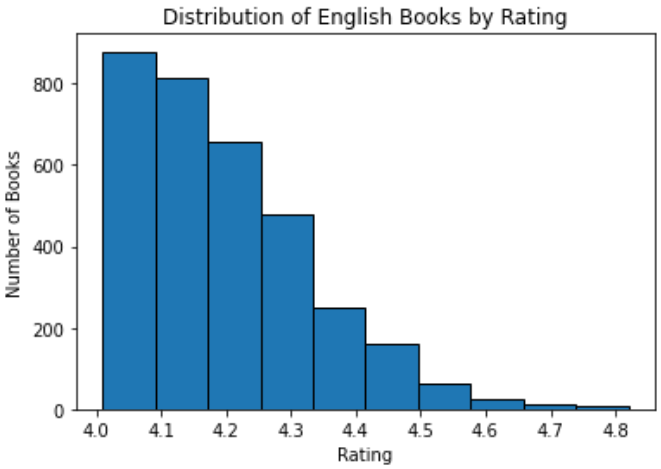
   - plt.plot(): This function is used to create a line plot using data.
   - plt.bar(): This function is used to create a bar plot using data.
   - plt.pie(): This function is used to create a pie chart using data.

.

III. Answers to the Questions (with Appropriate Illustrations)

1. What is the probability of getting an English book with a rating above 4?

- To calculate the probability of getting an English book with a rating above 4, we need to first determine the total number of English books in the dataset and also the number of English books with a rating above 4 in the given dataset.
- We can do this by filtering the dataset based on the language column and the rating column and then using the length of the resulting data frame to calculate the probabilities.
  Once we have these values, we can use the formula for probability:
  probability = number of successful outcomes / total number of possible outcomes
  In this case, the number of successful outcomes is the number of English books with a rating above 4, and the total number of possible outcomes is the total number of English books in the dataset.
  After doing all the steps we get answer to the above question to be equal to 0.5276770225516481.



2. How many books were published in the 90s?

- To determine the number of books that were published in the 90s, we need to filter the dataset based on the year of publication. We can do this by extracting the year from the publication date column and checking if it falls between 1990 and 1999.

  Once we have the filtered dataset, we can use the length of the resulting data frame to determine the number of books published in the 90s.

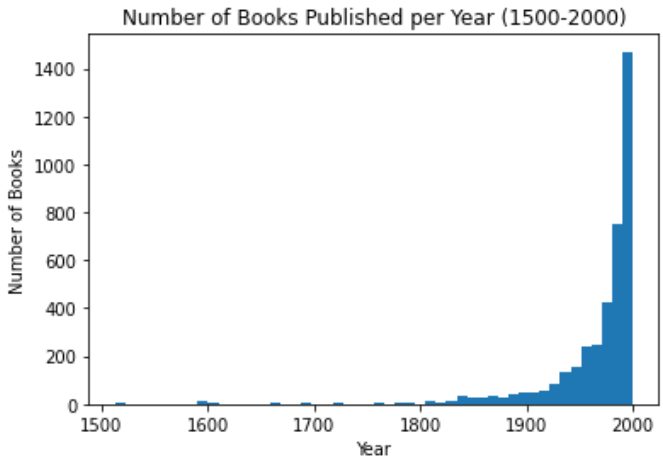
  After doing all these steps we will get the value of the question as 1360.

3. Which year has a maximum number of books published between 1500 and 2000 and how many books are published in that particular year?

- To determine the year with the maximum number of books published between 1500 and 2000, we need to filter the dataset based on the publication year and then count the number of books published in each year. We can then find the year with the highest count and the number of books published in that year.

  To illustrate this, we can use a histogram to show which year has the maximum number of books between 1500 and 2000 and its count.

  The answer to the above question is 2000.0 with 209 books in that year.



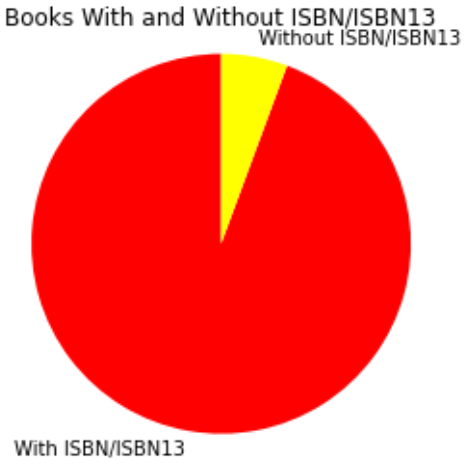
4. How many books do not have isbn or isbn13?

- To determine the number of books that do not have an ISBN or ISBN13, we need to filter the dataset based on the ISBN and ISBN13 columns and count the number of rows where both columns are missing.

  Once we have the filtered dataset, we can use the length of the resulting data frame to determine the number of books without an ISBN or ISBN13.

  To illustrate this, we can use a pie chart to show the breakdown of books by ISBN availability, with one category for books with both an ISBN and ISBN13, and one category without ISBN and ISBN13.

  After this, we get the value as 565 books without an ISBN or ISBN13.



5. Which is the most popular book by average rating?

- To determine the most popular book by average rating, we need to calculate the average rating for each book and then find the book with the highest average rating.

  So initially we need to access the column containing the rating and simultaneously we need to access the most popular book.

  We get all the details regarding that popular book.
- Book_id: 3628
- Title: The Complete Calvin and Hobbes
- Authors: Bill Watterson
- Average_rating: 4.82

IV. Summary of the Observations

1. According to the first question we came to an observation that there are many highly rated English books in the dataset, but they still make up a relatively small proportion of all English books.

2. Based on the analysis, there were around 1360 books published in the 90s. It means that in the 90s also people were drawn toward reading different kinds of books with authors publishing more and more on the basis of the requirements.

3. Based on the analysis we find that the highest number of books published between 1500 and 2000 is in 2000 which means as progress or development increases people get more inclined towards reading books to gather information and also to know the fictional characters.

4. Based on the analysis we find that there are 565 books without ISBN and ISBN13 which means after many years people start to register their books and most of them are not registering for them even after a long time.

5. Based on the analysis, the most popular book by average rating is "The Complete Calvin and Hobbes" by Bill Watterson, with an average rating of 4.82. It is worth noting that this result is based solely on the ratings listed in the dataset and may not be representative of the popularity or quality of the book in other contexts or among other audiences. Additionally, the analysis is limited to books with ratings listed in the dataset, which may not be representative of all books published. Nevertheless, this information may be useful for identifying highly-rated books within the dataset and providing insights into popular or well-received works.

Acknowledgment

I would like to express my gratitude to prof. Shanmuga to give me a chance to work on this data narrative assignment and also for their guidance and support throughout this project.

I would also like to thank all the Teaching Assistants and my friends for their encouragement and assistance throughout this assignment.

Due to this assignment I have learned a lot of new things which helped me to gain good knowledge in the course ES 114.

References

[1] Pandas Documentation
https://pandas.pydata.org/docs/

[2] Matplotlib User Guide
https://matplotlib.org/stable/users/index.html