Data Narrative

Lavanya (22110130)
B'Tech'22 Computer Science and Engineering
Probability, Statistics and Data Visualization

I. Overview of the Dataset

The Dataset contains six million ratings for ten thousand most popular (with most ratings) books. It also contains:

- Books marked to read by the users
- Books metadata
  1) number of books published
  2)Authors
  3)publication year
  4)title of the book
  5) Language
  6) ratings
- tags/ genres/ shelves

II. Scientific Questions/ Hypotheses

A. Are the top 10 most preferred books to read by the users also the top 10 books having the highest average rating?

B. What are the top 15 book publishing years,i.e, years with the maximum number of books published?

C.Who are the top 10 authors?(determined by the number of books written by them having average rating greater than 4)

D. What is the probability that the top 10 most popular books (determined by the number of ratings) also have the highest number of 5 ratings?

E. It is highly probable that a random book having average rating greater than 4 published in the years 2001 to 2015 was published in 2012

III. Details of Libraries and functions

Libraries used:

- Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.(I)
- Matplotlib: Matplotlib is a visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays.(II)
- Numpy: NumPy is a Python library used for working with arrays.It also has functions for working in the domain of linear algebra, fourier transform, and matrices.(III)

Functions used :

- Read_csv: used to load a csv file into a dataframe
- Value_counts: used to get a series containing count of unique values
- Index: returns the index information of a dataframe
- Isin:checks if the dataframe contains the specified values
- Plot:used to visualize a dataframe by plotting different columns into a graph.
- Set_xticklabels(IV):is used to set the x tick labels with list of string labels.
- Sort_values: sorts the dataframe by the specified label
- Set_xlabel: to provide label to the x axis
- Set_ylabel: to provide label to the y axis

- Len: used to return the number of rows in a dataframe

IV. Answers to the questions(with appropriate illustrations):

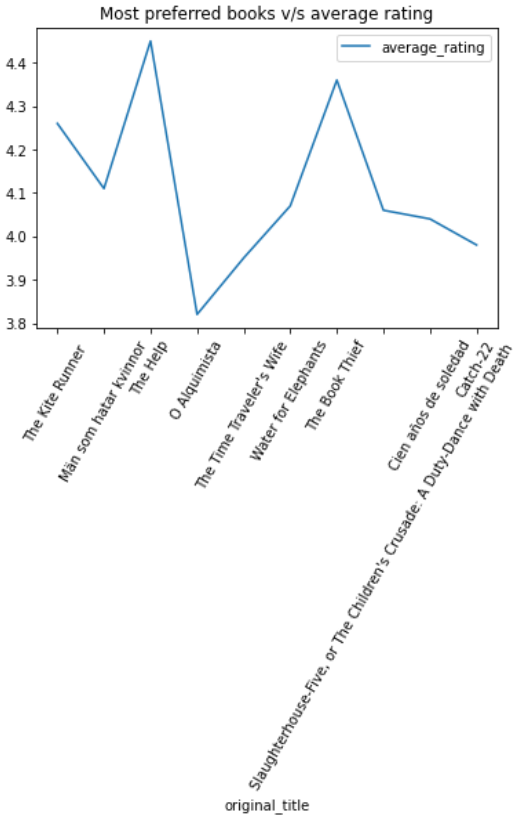
A. After Extracting the top 10 most preferred books to read by the users and the top 10 books having the highest average rating from the dataframes, it is found that there is no book common in both the lists. Thus, the answer is no.

Top 10 most preferred books and Top 10 books having the highest average rating:

Sr. no.	Book title	Average rating
1.	The Kite runner	4.26
2.	Nineteen eighty four	4.14
3.	O Alquimista	3.82
4.	A Game of thrones	4.45
5.	Life of pi	3.88
6.	The book thief	4.36
7.	Slaughterhouse-Five, or The Children's Crusade	4.06
8.	Catch-22	3.98
9.	Miss Peregrine's Home for Peculiar Children	3.89
10.	NaN	4.31

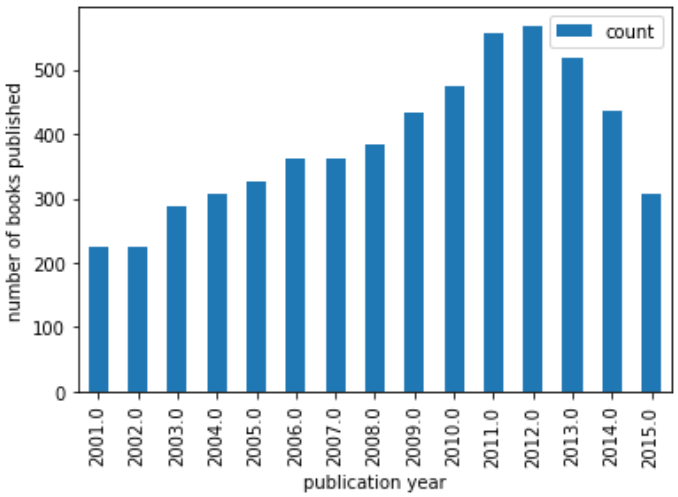
Sr. No.	Book title	Average Rating
1.	The Complete Calvin and Hobbes	4.82
2.	NaN	4.77
3.	Words of Radiance	4.77
4.	Mark of The Lion Trilogy	4.76
5.	NaN	4.76
6.	It's a Magical world: The Calvin and Hobbes Collection	4.75
7.	There's Treasure everywhere: A Calvin and Hobbes collection	4.74
8.	Complete Harry Potter boxed set	4.74
9.	Harry Potter Collection (Harry Potter #1-6)	4.73
10.	The indispensable Calvin and Hobbes: A Calvin and Hobbes collection	4.73

The following graph helps visualize how most preferred books have a range of varying average rating, instead of the expected high values of average rating.



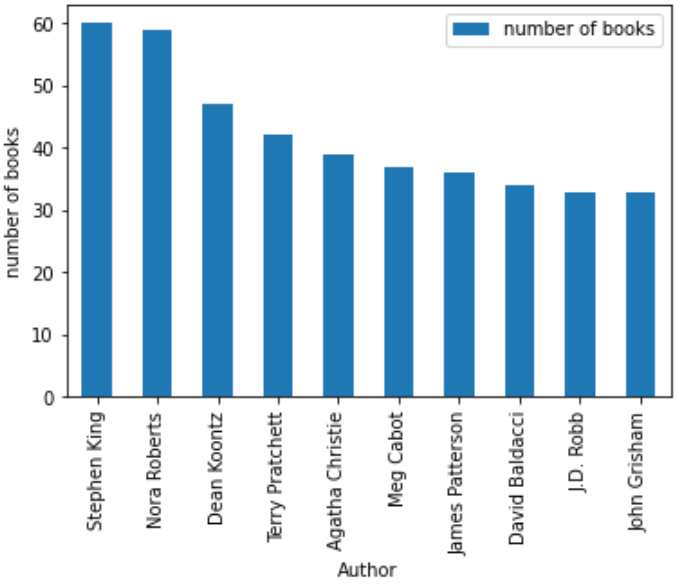
B. The top fifteen book publishing years are 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011,2012, 2013, 2014, 2015.

The following graph helps visualize the trend of the number of books published over the highest book publishing fifteen years.



C. The top 10 authors are Stephen King, Nora Roberts, Dean Koontz, Terry Pratchett, Agatha Christie, Meg Cabot, James Patterson, David Balddacci, J.D. Robb, John Grisham.

The following graph shows the most famous authors and compares the number of books written by them having average rating greater than 4.



D. Top 10 Most popular books

	ratings_count	book title
0	4780653	The Hunger Games
1	4602479	Harry Potter and the Philosopher's Stone
2	3866839	Twilight
3	3198671	To Kill a Mockingbird
4	2683664	The Great Gatsby
5	2346404	The Fault in Our Stars
6	2071616	The Hobbit or There and Back Again
7	2044241	The Catcher in the Eye
8	2001311	Angels & Demons
9	2035490	Pride and Prejudice

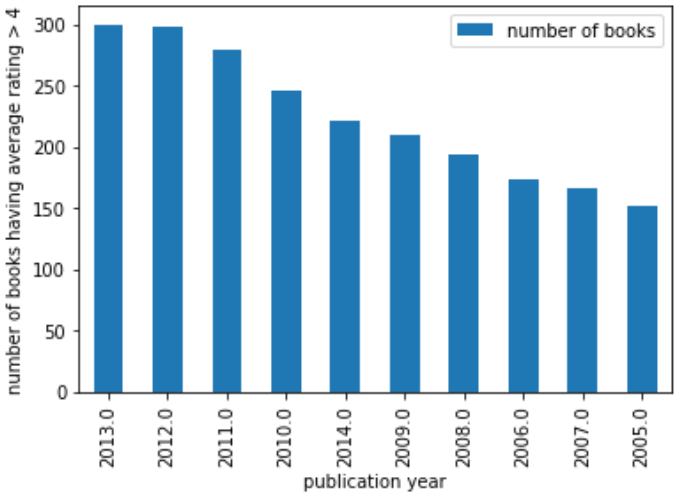
. Top 10 books with maximum number of 5 ratings

	ratings_count	book title
1	4602479	Harry Potter and the Philosopher's Stone
2	4780653	The Hunger Games
3	3198671	To Kill a Mockingbird
4	3866839	Twilight
5	1746574	Harry Potter and the Deathly Hallows
6	2346404	The Fault in Our Stars
7	1832823	Harry Potter and the Prisoner of Azkaban
8	1753043	Harry Potter and the Goblet of Fire
9	1678823	Harry Potter and the Half-Blood Prince
10	2035490	Pride and Prejudice

From both the tables, it is clear that six out of the ten most popular books are in the list of top 10 books with the highest number of five ratings. Thus the probability is 0.6.

E. The hypothesis is not correct, since only 298 books were published in 2012, out of the 2242 books having high average ratings published from the year 2001 to 2015.

The following graph shows the number of books having an average rating greater than 4 published in the years from 2001 to 2015.



V. Summary of Observations:

It is observed that most preferred books, i.e, the books that maximum users want to read have no relation with the ratings of the books. Thus, a user does not necessarily prefer reading a book with a high rating, rather it depends on personal choice. Also, a maximum number of high rated books have been published in the year 2012. It is also observed that the most popular books are not necessarily those with a high number of 5 ratings.

VI. References:

(I) "Pandas Introduction," n.d. https://www.w3schools.com/python/pandas/pandas_intro.asp.

(II) GeeksforGeeks. "Python Introduction to Matplotlib," May 14, 2018. https://www.geeksforgeeks.org/python-introduction-matplotlib/.

(III) "Introduction to NumPy," n.d. https://www.w3schools.com/python/numpy/numpy_intro.asp.

(IV) GeeksforGeeks. "Matplotlib.Axes.Axes.Set Xticklabels in Python," July 1, 2022. https://www.geeksforgeeks.org/matplotlib-axes-axes-set_xticklabels-in-python/.

VII. Acknowledgements:

Professor Shanmuganathan Raman, department of Computer Science and engineering and Electrical Engineering, Indian Institute of technology Gandhingar.