

**IST 652 SCRIPTING FOR DATA ANALYSIS**

**Final Project Report**

**By John Jefferson**

**September 2022**

INTRODUCTION / SCOPE

Goodreads the world’s largest site for book readers and has over 120 million users worldwide. Goodreads was founded in 2006 and has been a subsidiary of Amazon since March of 2013. The website allows its users to search for books, annotations and quotes using its databases. Goodreads also allows its users the ability to review books and have forums discussing about the books. For this project I took a dataset of Goodreads books and users review data. The dataset I found is on Kaggle ([*https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks?resource=download*](https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks?resource=download)) and it uses the Goodreads API. The goal of this project is for me to use the dataset to figure out what affects the ratings of the books the most.

books.csv is the dataset I’m working with has 12 columns and it has book data from Goodreads. It is 1.56 MB. 

Here’s a summary of what each column represents in the books.csv dataset before I talk about the questions in which I’m trying to answer in my project.

METADATA

* **TITLE-** The name under which the book was published
* **AUTHORS-** Names of the authors of the book
* **AVERAGE\_RATING-** The average rating of the book received in total.
* **ISBN-** Unique number to identify a book (International Standard Book Number)
* **ISBN13-** 13-digit ISBN instead of the standard 11 digit one
* **LANGUAGE\_CODE-** Primary language of the book.
* **NUM\_PAGES-** Number of pages in a book
* **RATING\_COUNT-** Total number of rating a book received.
* **PUBLICATION\_DATE-** Date when the book was first published
* **PUBLISHER-** Publisher of the book

BUSINESS QUESTIONS

For this project I wanted to take a good look at the book.csv dataset and try to determine what drives the ratings that the Goodreads users give to each book. With this dataset I have the author, book language, number of pages, number of ratings, publication date, and publisher of the book to look at. The goal of this project is to see which column has the greatest impact on a book’s rating. Here are the questions the I decided to purse with this project.

* Which authors have the highest rated books?
* Does Language affect the rating of a book?
* Does the number of ratings affect the rating of a book?
* Does Publication date affect the rating of a book?
* Which Publishers have the highest ratings on their books?
* Does the amount of reviews a book get after the rating of a book?

DATA IMPORTATION

For this project I used the programming language called python to load in the data. The code was run in Juptyer notebook. I imported the data into python using pandas and then put it into a pandas dataframe.

Graphical user interface, application

Description automatically generated

DATA CLEANING AND PREPARATION

First, I dropped an unused column after making a pandas dataframe. Then I checked for NAs to see if anything was missing within the dataset. Then I took the publisher date column and made a number so I could run analysis with it.

A picture containing text, screenshot, plant, aquatic bird

Description automatically generated

DATA ANALYSIS / PLOTS

After cleaned the data I wanted to do so data exploration. The first thing that I started to do was to start pulling apart some of the data to get a better look at what I needed to create some plots. Because the dataset that I was working with was so big I had to make some changes into order for my data to fit nicely within each plot. I decided new tables and data frames to help reflect what I needed to create each plot. Below are the plots I created to answer the business questions that I mentioned before.

DOES LANGUAGE AFFECT THE RATING OF A BOOK?

I wasn’t sure what I would find when I was first making this plot. Initial I had some thoughts that because Goodreads was an American site that English books would have the highest ratings but that also means that American readers could critique them the hardest as well. When looking at the plot below I can see that English (outside the US, Canada, and Great Britain), Greek and Welsh books have the highest average ratings. I wasn’t really expecting to see Greek or Welsh so that was a surprise.

Chart, bar chart

Description automatically generated

WHAT AUTHORS HAVE THE HIGHEST RATED BOOKS?

This question was the one I was most interested in finding out because I want to know which Authors are being well received on Goodreads. As you can see from the plot below that a few authors do have perfect average ratings. A book written by Aristophanes/F.W Hall and W.M Geldart in the most left dot. I assumed that it was authors with one book and few reviews. This means it’s hard to get a true sense of which Authors on Goodreads are truly being well received.

A picture containing shape

Description automatically generated

WHICH PUBLISHERS HAVE THE HIGHEST RATINGS ON THEIR BOOKS?

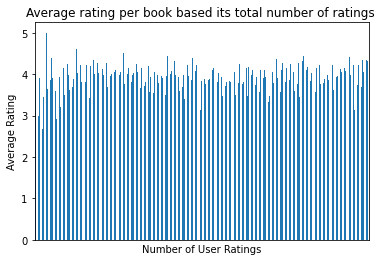
Like with Authors, I wanted to see which Publishers were the most liked on Goodreads. It seems like this business question essentially had the same fate as the one before. I assumed that it was publishers that had published one book and it got a few reviews. This means it’s hard to get a true sense of which Publishers on Goodreads are truly being well received. One of the Publishers with a perfect average rating is Oxford University Press USA.

A picture containing chart

Description automatically generated

DOES THE TOTAL NUMBER OF RATINGS AFFECT THE BOOKS AVERAGE RATING?

With this plot, I wanted to see whether books were fewer or greater number of ratings would end up with lower or highest average rating scores. Initial I thought that books with more reviews would have higher average ratings but then I thought that some books might get review bombed and that would cause that to not be the case. As you can see from the plot below that it seems like there is a small difference between the average rating of a book based on the total number of ratings it has. Although I do see a small u-shaped curve indicating that the lower and higher number of ratings a book receives does slighting correlate into having a higher average rating. Also, strangely enough a book got a perfect score near the left side of the plot. The plot below is arranged so that the books with the smallest total number of ratings is to the left and books with the largest total number of ratings is to the right.



DOES THE AMOUNT OF REVIEWS A BOOK GET AFTER THE RATING OF A BOOK?

Just like the plot before I decided to arrange the text reviews a book received from smallest to largest. The point of this plot was to see if book with more or less reviews had any effect on the average rating. Looking at the plot below I can say that overall, there isn’t much correlation between the too. It seems like it is all over the place with no real overall trend.

Chart, bar chart

Description automatically generated

DOES PUBLICATION DATE AFFECT THE RATING OF A BOOK?

For this plot I had to get the publication dates to an integer. Then, like the last two plots, I decided to arrange it from oldest publication date to the newest publication date. With this plot I wanted to see whether older or newer books were better received by users on Goodreads. I honestly didn’t have a clue on how this plot would look and it’s safe to say that based on how it looks that publication date has very little impact on a book’s average rating. It does seem that a few years didn’t do very well but overall, not a lot of correlation.

Chart, bar chart

Description automatically generated

CONCLUSION

The goal of this project was to see what drives the ratings that the Goodreads users give to each book. I was very interested in seeing how Goodreads user base rated its boos. Going into this project I thought that the total number of reviews or a book’s language would have the greatest image on its average rating.

After my date exploration I discovered that Publication Date, Text Reviews didn’t have much impact on a book’s average rating. This is probably due to the sheer mass of books that were in the database which was going to even out the data.

With the total amount of ratings, I found a small difference between the average rating of a book based on the total number of ratings it has. Books with lower and higher number of ratings did seem to have a higher average rating than books that were closer to the middle.

I was able to see which Authors and Publishers had the highest average reviews. I did come to the conclusion that it was most likely authors and publishers that had one highly rated book in the database based on the plots that I had. It would be cool to see if there is a different approach to gathering insight on how authors or publishers are being received on Goodreads.

I also think it would have been cool to look at some descriptive statistics to see exactly what impact each column had on the average rating.