**Books**

BY RUQAYYAH MUSE

**Overview:**

Books are a 25-billion-dollar industry in the United States alone\*, and there is an entire community dedicated to reading, reviewing, and promoting books. This project looks at a dataset of 11,127 books and uses this data to analyze the trends between ratings, reviews, publication years, and authors.

[Dataset](https://www.kaggle.com/jealousleopard/goodreadsbooks): There were 11,127 entries and 12 attributes. Out of the 12 attributes, only 7 were used.

The Attributes:

* BookID: the dataset came with an id for the books such as **1**
* Title: the title of the book such as **Harry Potter**
* Authors: the authors of the book (included co-authors) such as **J.K. Rowling**
* Average Rating: the average rating of the book (ratings being from 1 to 5) such as **4.45**
* ISBN: the commercial identifier of the book such as **439785960**
* ISBN 13: a 13-digit version of an ISBN such as **9.78044E+12**
* Language Code: a string determining the language of the book such as **eng**
* Number of Pages: the number of pages of a book such as **656**
* Ratings Count: the number of ratings of a book such as **2095690**
* Text Review Count: the number of text reviews of a book such as **27591**
* Publication Date: the month, day, and year a book was published such as **9/16/2006**
* Publisher: the publisher of the book such as **Scholastic Inc.**

**Cleaning the Data:**

* Some entries caused errors as I tried to load the file because they were miswritten, so I had to clean that out using Excel before being able to load and clean the data in Jupyter Lab.
* There was an unnamed column full of null values that wasn’t a part of the original dataset, so I just deleted that.
* Co-authors were included in the authors column and separated with ‘/’, so I decided to remove all the co-authors except for the first one listed.
* The values in **average\_rating**, **num\_pages**, **ratings\_count**, and **text\_reviews\_count** were all strings, so I changed them to numerical values using down casting.
* Finally, entries with an average rating of 0 were deleted. I then deleted entries with rating counts of 0, because a book can’t have an average rating if there are no ratings at all.

Once the data was cleaned, I was left with 11,044 entries and 12 attributes.

**Calculating Statistics:**

I calculated the mean, median, variance, and standard deviation of the average rating, the number of pages, the ratings count, and the text reviews count.

For the **average rating:**

* Mean: 3.943 Median: 3.96 Variance: 0.087 Standard Dev: 0.295
* The mean and median are pretty close, and the variance/standard dev is really small which suggests that the data is evenly spread out.

For the **number of pages:**

* Mean: 336.8 Median: 300 Variance: 58077.3 Standard Dev: 241
* The mean and median are not that close, and the variance/standard dev is large which suggests that the data is not as uniform.

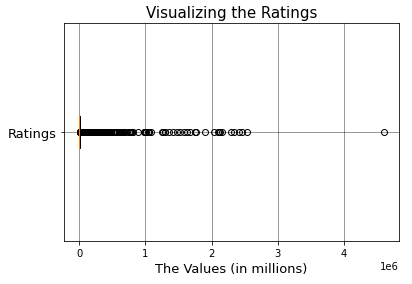
For the **ratings count:**

* Mean: 18071.2 Median: 765.5 Variance: 12744279295.5 Standard Dev: 112890.6
* The mean and median are not close, with the mean being 10,000 and some more than the median, and the variance/standard dev is really huge suggesting that there are very big outliers.

For the **text reviews count:**

* Mean: 543.9 Median: 48 Variance: 6684345.4 Standard Dev: 2585.41
* The mean and median are not close, and the variance/standard dev is really large which suggests that the data has some big outliers and is not that evenly spaced.

Visualizing the Ratings Count and Text Reviews Count Statistics:

Chart

Description automatically generated

Ratings has a lot of huge outliers (the x-axis is in millions), with the max being 4,597,666 so you can understand why the mean and variance is so huge.

Text Reviews also has a fair number of outliers far from its median, which again explains why the mean is so large compared to the median, as well as the huge variance.

**Visualizing the Data:**

Chart, scatter chart

Description automatically generatedCorrelation Between Rating & Page Length:

There doesn’t seem to be a definitive correlation between the two. As you can see, most books are between 0 and 1000 pages, and the average rating seems to be between 3 and about 4.6. The color is chosen based on the number of Text Reviews, and you can see that the majority is solidly purple, which is a range of about 10000 and less for text reviews.

Chart, bar chart

Description automatically generatedMost Common Authors & Popular Books:

Using a dictionary, I recorded the frequency of each author in the dataset and found the top ten that occurred. Shakespeare was the most common with 88 entries.

Chart, bar chart

Description automatically generatedThen, I found the most popular books by rating and graphed that. I couldn’t just sort the average-ratings column and take the top 10 titles, because that wasn’t very representative as the results could have small number of ratings but a high average rating, which could easily skew the result. Instead, I added another column titled **popularity** that was determined by dividing the number of ratings by the average rating. That way, I could more accurately find the most popular books.

Total Ratings vs Total Reviews:

Initially, it was hard to see the overlap between the ratings and review because the number of ratings was so much larger than the number of reviews. As a result, I sorted the day based on rating counts, and then split the rating counts and text review counts into three groups so it will be clearer when graphed.

Chart, histogram

Description automatically generatedThe first graph (marked P1) has the first half of the data, with **blue representing ratings**, and **yellow representing text reviews**. The reviews start off taller than the ratings because there are books that have zero reviews, but they have at least one rating. So the graph for ratings doesn’t start at zero like the reviews do. As the ratings graph overlaps the reviews graph, you can see that it’s consistently smaller than the ratings, which suggests that books generally have a lot of ratings, but not as many text reviews.

Chart, histogram

Description automatically generatedText, letter

Description automatically generated



The second graph (P2) also has the ratings and reviews starting at different spots, because the books in this second splice of data are starting to have much larger numbers of ratings, but they have small numbers of text reviews in comparison, which is illustrated by the book data on the right. You can see that the rating\_count is 766 while the text\_reviews\_count is 47, so there is a huge gap between the two as shown in the graph.

The third graph (P3) follows that same trend. In the source code, I also included a graph with all the data together to show how it was almost impossible to see it originally.

**Project Summary:**

* Is there any correlation between the average ratings and page lengths? **Possibly.**
* Which author appears the most frequently in the data? **William Shakespeare.**
* Which book is the most popular? **Twilight by Stephanie Meyer.**
* How do the number of ratings and reviews compare? **Reviews are generally less compared to ratings.**

Future projects could look at a larger dataset of books with an attribute for genres. I think that within genres there are specific trends and tendencies, and it would be interesting to analyze those as well.