Hnin Nandar Win
Antoine Teillac
Emilie Schmitter

Python Project - Goodreads Analysis

I. Cleaning the database:

After separating the fields by comma, we found out that there was an error line 3350 in the Excel file. The author field consists of 2 people, separated by a comma. After correction from the file opened in Excel, we resumed cleaning the database.

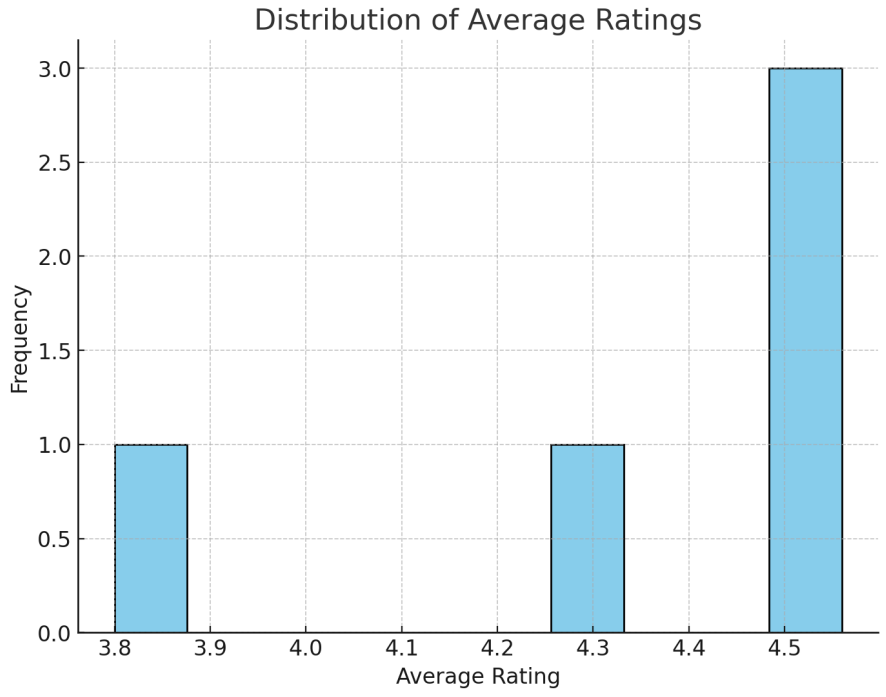
We noticed another error for a column named “Unnamed: 12”, which was probably a column created by a separator included in one of the fields. We will also rename the column ' num_pages' to 'num_pages'.

The rows and columns where the average_rating is “NaN”, are not usable for the rest of the analysis.

II. Exploratory Data Analysis

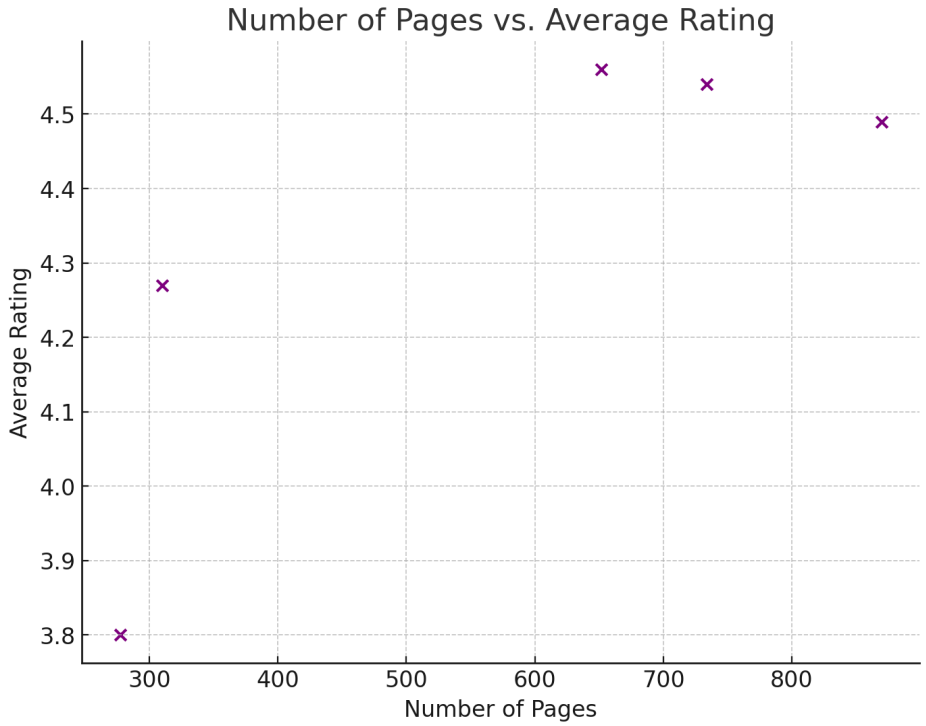
After cleaning the database and before training our model, we did an analysis to highlight important factors which could help us ameliorate our data predictions for future books reviews. We tried to comprehend what could actually influence a book’s rating depending on multiple characteristics.

1) **Distribution of Average Ratings**:



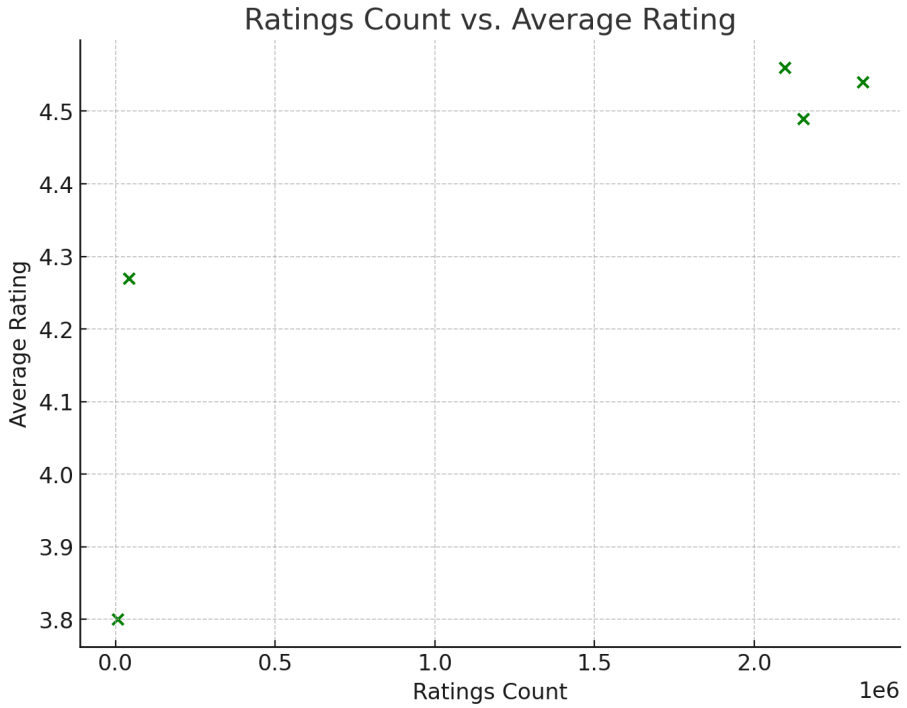
We could see that the majority of the books in the dataset have an average rating between 4.0 and 4.5, indicating that most books are rated highly by readers.

2) **Number of Pages vs. Average Rating**:



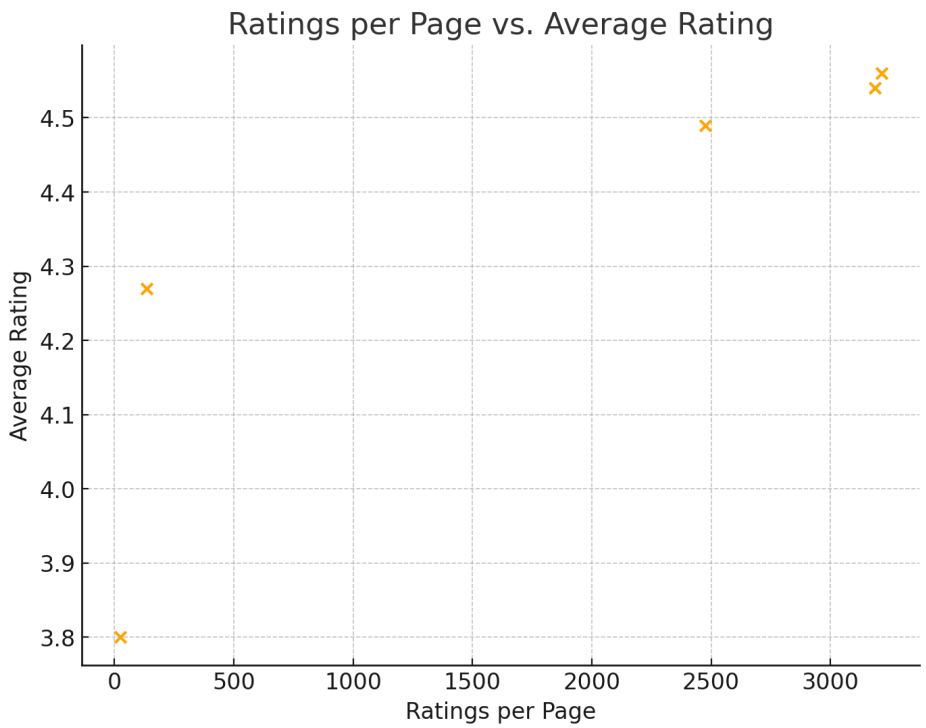
There is no clear relationship between the number of pages and the average rating, though it appears that books with fewer pages tend to sometimes receive low ratings.

3) **Ratings Count vs. Average Rating**:



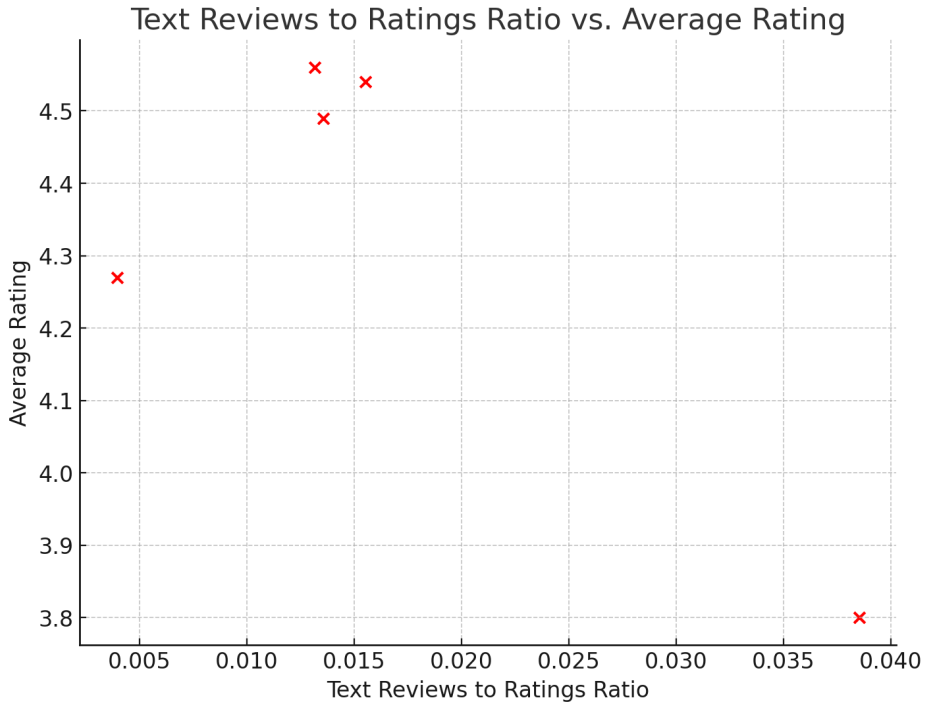
We found that popular books (those with higher ratings counts) tend to have higher average ratings, but there is still variation, suggesting that a large number of ratings does not always guarantee high average ratings.

4) **Ratings per Page vs. Average Rating**:



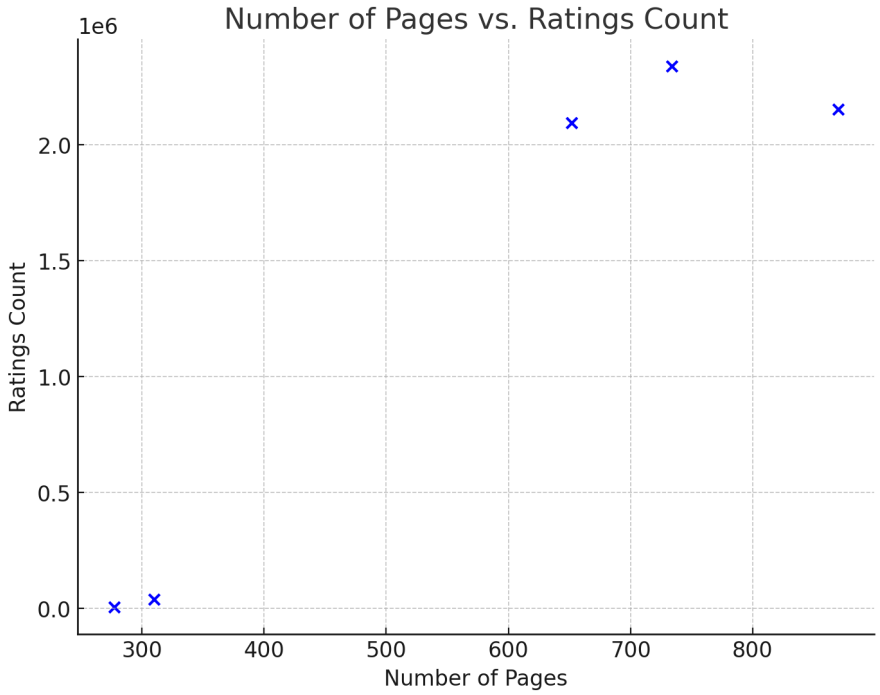
We noticed that there isn't a strong visible trend between the number of ratings per page and the average rating, indicating that the rating density does not significantly influence the overall rating.

5) **Text Reviews to Ratings Ratio vs. Average Rating**:



A higher text reviews to ratings ratio might correlate with slightly higher average ratings, suggesting that books with more engaged readers (who leave text reviews) may be rated more favorably.

6) **Number of Pages vs. Ratings Count**:



There seems to be a weak positive correlation between the number of pages and the ratings count, indicating that longer books tend to get more ratings, though the relationship isn't very strong.