# Project Report

GitHub link **here.**

As we know, predicting the rating of a book can be approached as a regression task, as it involves predicting a continuous numeric value, namely the average rating. Therefore, utilizing a supervised learning method is suitable, as it enables us to predict a continuous outcome variable (the book's rating) based on one or multiple input variables (including titles, authors, page count, rating count, publishers, and other factors).

Our work will be declined in five steps:

1. Data Exploration

2. Data Visualization

3. Feature Engineering

4. Data Modeling

5. Conclusion

## 1. Data Exploration

Our dataset contains 11123 rows and 12 columns such as bookID, title, authors, average_rating, isbn, isbn13, language_code, num_pages, ratings_count, text_reviews_count, publication_date, publisher. There are no null values and no duplicated values.

Quantitative variables:

- Average_rating : The average rating of the book received in total
- num_pages: The number of pages the book contains
- ratings_count: The total number of ratings the book received
- text_reviews_count: The total number of written text reviews the book received

Qualitative variables:

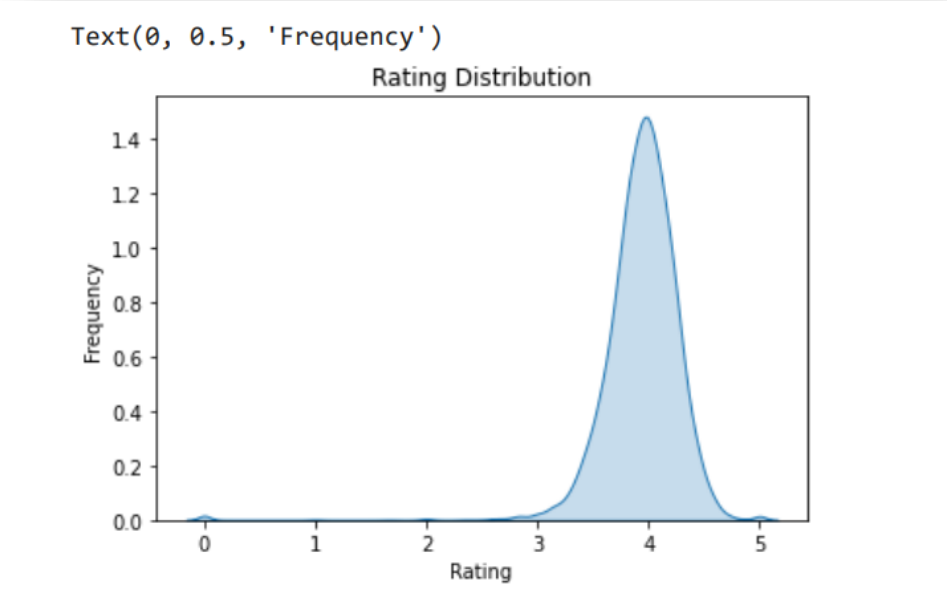
- title: The name under which the book was published.
- authors: The names of the authors of the book. Multiple authors are delimited by “/”.
- language_code: Indicates the primary language of the book. publication_date: The date the book was published.
- publisher: The name of the book publisher.

We also have some other variables for books identification: (bookID, isbn, isbn13).

We notice immediately that the target feature average_rating.

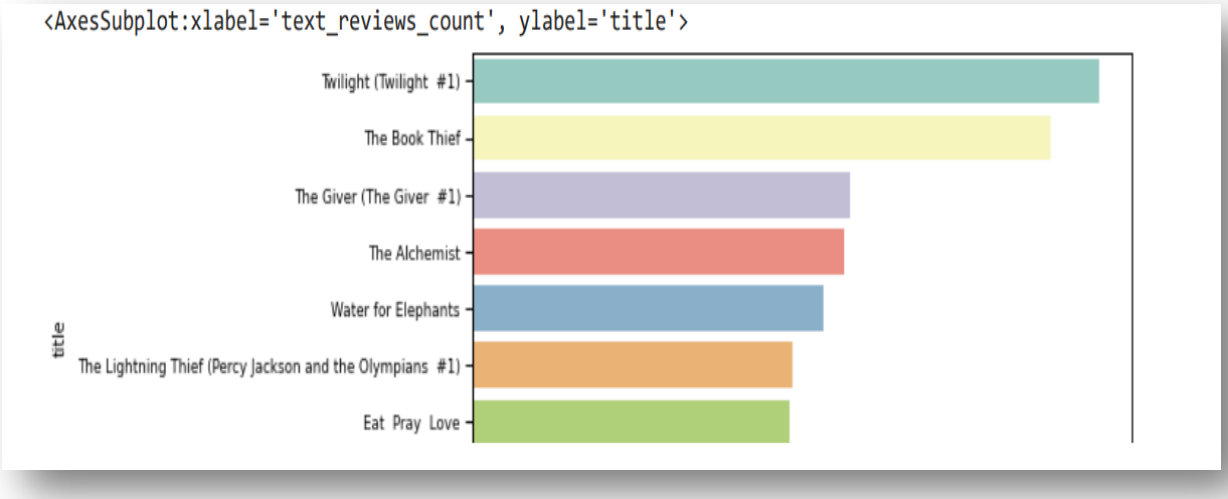
## 2. Data Visualization

- Average ratings distribution



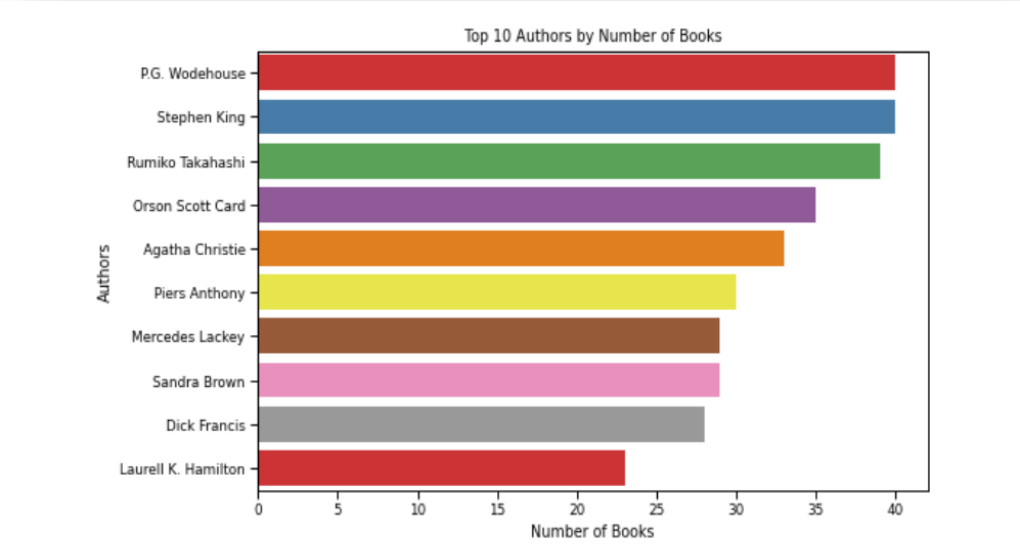

*This plot provide a visual representation of how the average ratings are distributed. The KDE (Kernel Density Estimation) plot helps us understand the shape and concentration of data points along the rating scale. The x-axis represents different rating values and the y-axis represents the frequency of those ratings.*

- Books with more written text reviews



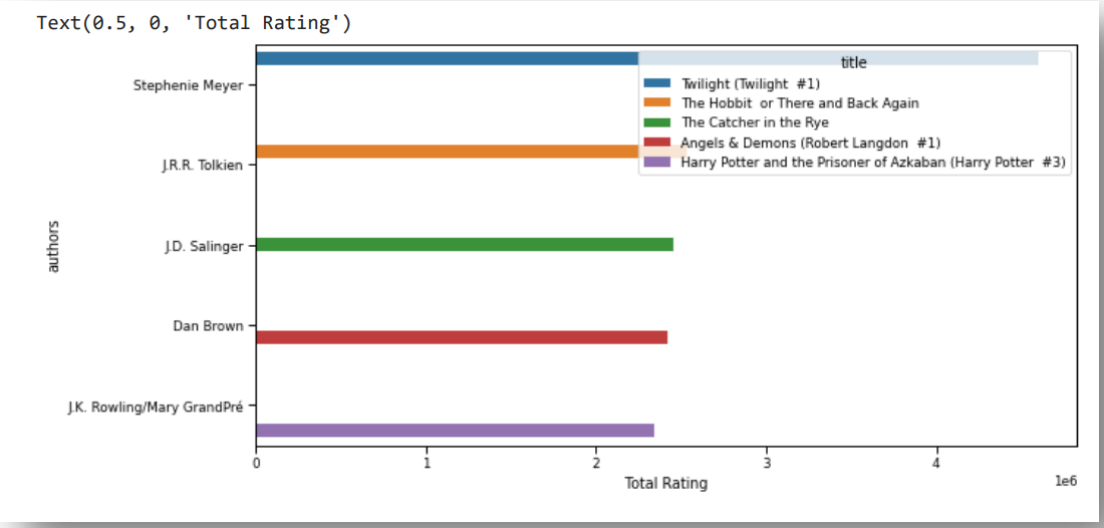

*This plot gives us a quick overview of which books are generating the most text reviews, which could be an indicator of their popularity or engagement among readers. We visualize the top 10 books with the highest number of text reviews. Each bar in the plot represents a book, and its length indicates the number of text reviews that particular book has received. The longer the bar, the more text reviews the book has garnered.*

- Top 10 authors in our dataset



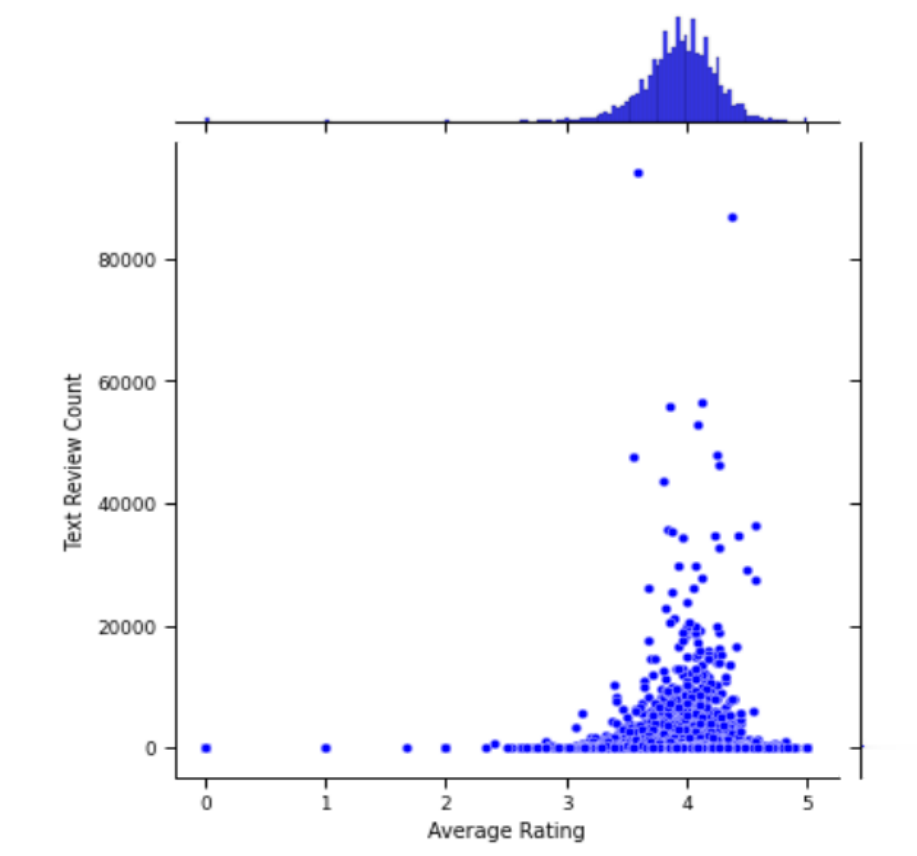

*This visualization allows us to quickly identify which authors have contributed the most books. It provide insights into which authors are particularly prolific and may be of interest to readers. By focusing on the top 10 authors, we can understand the distribution of book counts across these influential authors. This information could be valuable for various decision-making processes, such as marketing strategies, author engagement, and content recommendations.*

- The authors whose books have received the highest ratings count




*Here a meaningful plot that highlights the top authors based on the total ratings count for their books. With this visualization, we can better understand which authors have successfully captured readers' interest, as well as the specific books that have contributed significantly to their high ratings counts. Overall, this plot provides a comprehensive overview of the authors who have achieved the highest levels of engagement and recognition from readers.*

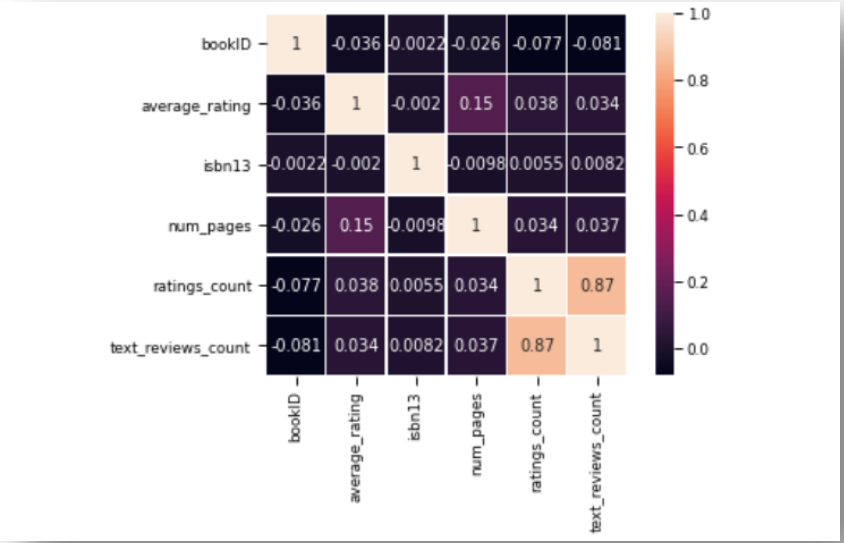
- Distribution between Rating and Text Reviews




*We are observing a notable presence of outliers. It's evident that there isn't a significant correlation between the average rating and the book's page count. This observation is further supported by the heatmap, which reveals a correlation coefficient of 0.15.*

*To address this, we'll be excluding the outliers specifically those with a page count >= 1500.*

- Relationship existing between the selected variables of our dataset



*This insightful heatmap plot provide a visual representation of the correlations between different numerical variables in our dataset. We see a high correlation between the ratings_count and the text_reviews_count around 87%.*

## 3. Feature Engineering

In this step we defining training variable feature, evaluating Other Variables and Handling Outliers (please Refer to the notebook for this step).

## 4. Data Modeling

Based on the feature engineering, we decided to keep the following predictor variables: num_pages , ratings_count, text_reviews_count, normalized_age, language. The variable to be predicted is average_rating.

Prior to the modeling, we split our data into two subsets; a 20% subset for test and the remaining 80% for training.

We carried out 04 different models, in order to compare them and evaluate which one is the most optimal. The 04 models being: Linear regression, Decision Tree, Random Forest and XGBoost model.

## 5. Conclusion

In wrapping up our Book Ratings Prediction project and summarizing the insights we've gathered, we find ourselves favoring the Random Forest and XGBoost models due to their lower RMSE values. These models have shown promise in accurately predicting book ratings, marking a notable stride in our analytical journey. As we conclude this endeavor, we stand poised to leverage these insights for informed decision-making and further advancements in the realm of book rating predictions.