Stat 470/670 Final Project: Goodreads Exploratory Data Analysis

Team members: Akhila Sakiramolla (asakiram), Riya Shetty (rishett), Mitali Tavildar (mtavilda)

Introduction:

One of the world's most influential reading sites, Goodreads offers a forum for those interested in books to discuss them. Book recommendations are driven by word-of-mouth, and Goodreads has amplified the word-of-mouth effect. As avid readers and we have always been interested in how books are being recommended on various platforms based on our reading history and interests. Our aim for this project is to analyze the book's information to understand the user's reading habits and also to understand what factors lead to the increasing popularity of certain books.

Statement of Goals:

Our goal is to get a fair understanding of the relationship between the multiple characteristics a book might have, such as the average rating of each book, the popularity of the authors over the years, and the number of languages it comes in. We explored the dataset to address some of the research questions:

1. The first intuitive question was to understand the distribution of ratings for various books. Where do the majority of the book lie, in terms of rating?
2. Is there a relationship between the number of ratings given and the ratings? This will help us understand if the popularity of a book is dependent on the value of the rating received or vice versa.
3. Does the number of pages have an impact on the ratings? The aim here is to point out any impact that number of pages has on the overall popularity of the book.
4. One of the important research questions was to observe authors' performances over time. Are they performing the same over time, with their new books?
5. We wanted to know if books can be recommended based on their ratings.

Data description:

For this, we analyzed the dataset which contains various information about the books on the website of the world's largest book archive and book proposal site GoodReads. The Goodreads dataset was taken from Kaggle and it was extracted using the Goodreads API to obtain a well-cleaned dataset that only contained features that were deemed promising. The dataset contains information on books by around 6,000 authors and in 27 different languages. It gives more information like the number of pages, the number of ratings, average ratings, publishing date, publisher, etc., The dataset has 12 columns and 11,131 rows. Below is the description of all the variables in the dataset:

1. **bookID** - A unique Identification number for each book. Dataset has 11,131 unique IDs.
2. **title** - The name under which the book was published. Dataset has 10,353 unique titles.
3. **authors** - Names of the authors of the book. Multiple authors are delimited with -. Dataset has 6,644 unique authors.
4. **average_rating** - The average rating of the book received in total.
5. **isbn** - Another unique number to identify the book, is the International Standard Book Number. Dataset has 11,127 unique isbn codes.

6. **isbn13** - A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN. Dataset has 11,128 unique isbn13 codes.
7. **language_code** - Helps understand what the primary language of the book is. For instance, eng is standard for English. Dataset has 32 unique language codes.
8. **num_pages** - Number of pages the book contains.
9. **ratings_count** - Total number of ratings the book received.
10. **text_reviews_count** - Total number of written text reviews the book received.
11. **publication_date** - Date on which the book is published.
12. **publisher** - The publisher who published the book. Dataset has 2,295 unique publishers.

Below are the descriptive statistics of the numerical variables:

Metric	average_rating	num_pages	ratings_count	text_reviews_count
Min	0	0	0	0
Median	3.96	299	745	46
Mean	3.93	336.4	17,936	541.9
Max	5	6,576	4,597,666	94,265

Approach and findings:

In the context of the dataset, we're dealing with, we chose to define popularity as the mean of average ratings as well as mean of the total number of ratings a book has. The reason we made this decision is that ratings are an excellent way to gauge the overall quality of a book, while counting the total number of ratings is a way to gauge how many people actually read a book or how popular a book is.

We first wanted to understand how the ratings of all the books are distributed, so we began by looking at the general distribution of average ratings provided by the users. We plotted the same and noticed that most average ratings are around 4. We then wanted to see if the number of ratings a book receives has any impact on the average rating it has. For this, we made a plot between average ratings and the total number of ratings, after removing some outliers. From this, we observed that there may be a potential positive relationship between these two variables. Similarly, we wanted to see if the length of a book has any impact on the rating it has. We observed that books with a page number range of 200 - 400 have the highest number of ratings given, which may be due to the fact that people prefer reading books with a moderate number of pages.

We then wanted to see how the authors have performed over the years in terms of the ratings they received and if this has any impact on the average rating of the books. This is because we assumed that the author's performance would affect a user's psychology while choosing a book. For this, we first filtered out authors who have published books for over years now and observed their general distribution of reviews over time. We observed that there is no consistent pattern over the years.

Preliminary Data Analysis:

We first started by investigating some of the variables of interest, based on the research questions we want to address. These variables are - **average_rating** (the average rating a book received), **num_pages** (the length of a book), **ratings_count** (the total number of ratings a book received), **authors,** and **publishers**.

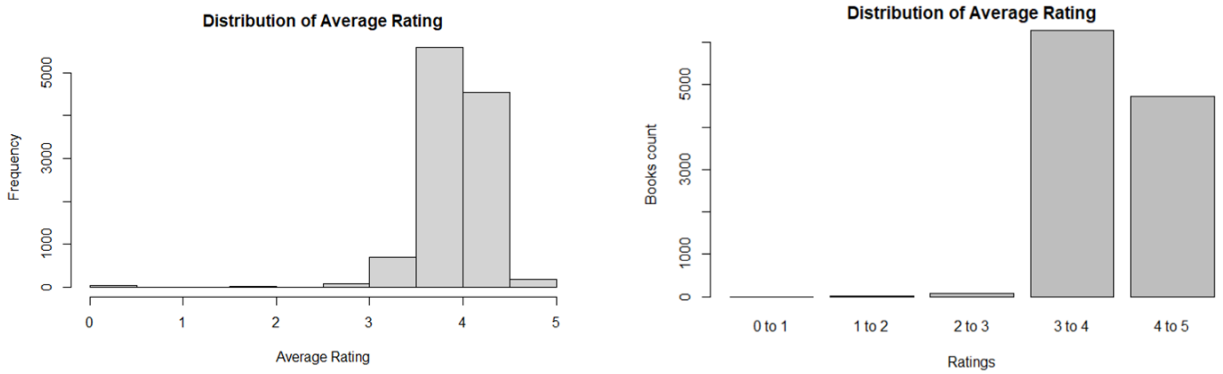

Fig 1. Distribution of Average ratings

From the above plots, we observe that majority of the ratings lie between 3 and 4. Books having scores near 5 are very rare. We also observe that there are some ratings between 0 and 1, which may mean that even though a reader doesn't like a book, they may give some rating. (**Refer to Fig 1**)

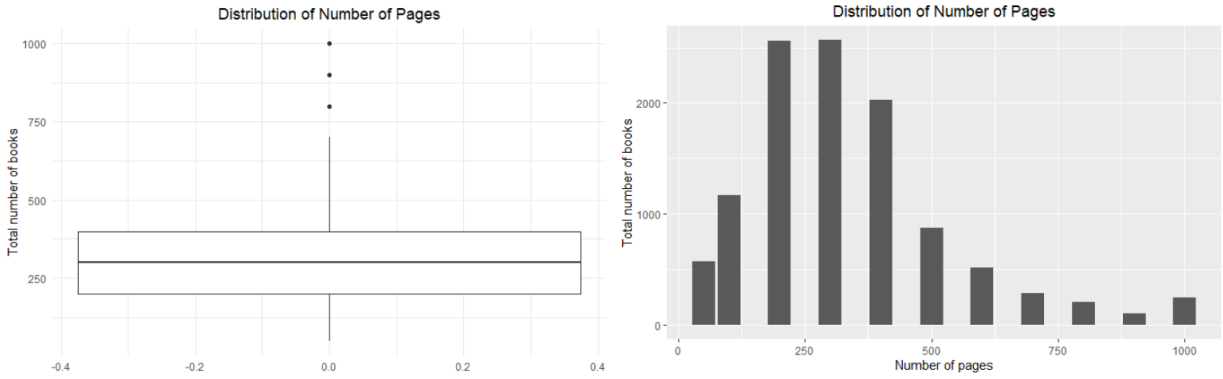

Fig 2. Distribution of Number of pages

We see that the majority number of books have a length of books between 200 to 400 pages. We also observe that there are a significant number of outliers between 750 and 1000 pages. (**Refer to Fig 2**)

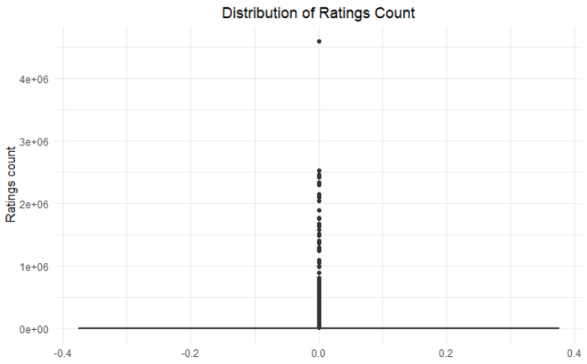

Fig 3. Distribution of Ratings counts

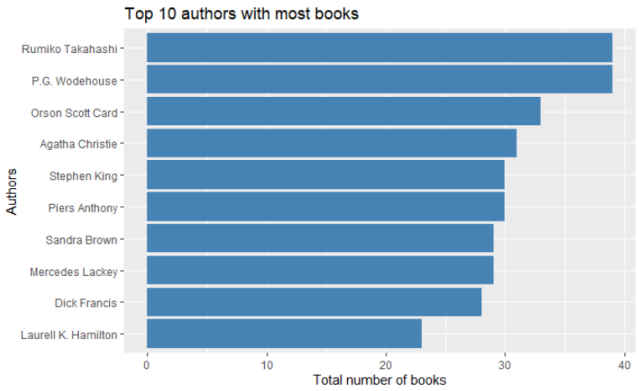

Fig 4. Top 10 authors with the most books

We observe that there are a large number of massive outliers in **ratings_count** and it will be necessary to do an outlier analysis when exploring questions related to it. (**Refer to Fig 3**). We can see that Rumiko Takahashi has the greatest number of books on the list. Several of them might be just different editions of the same book, considering that his work has been around for a long time, spanning decades. The names on the list again indicate that most of the authors have either written for decades. (**Refer to Fig 4**)

Exploratory Data Analysis:

I. Relationship between average ratings and total ratings:

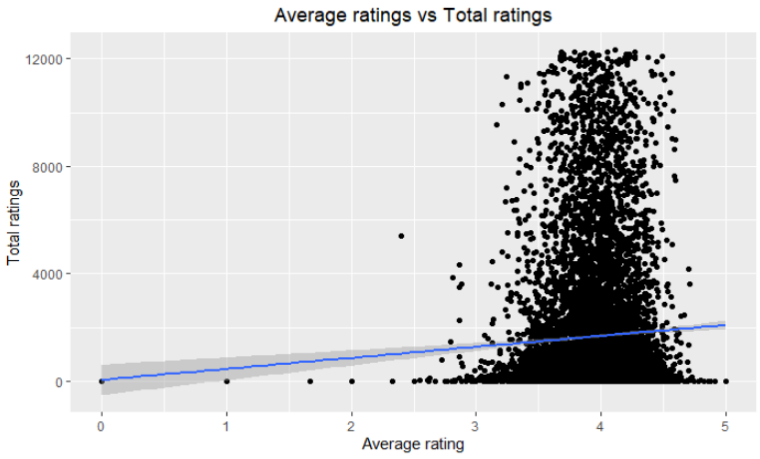

Fig 5. Relationship between average rating and total ratings

We observe that there may be a potential positive relationship between the average rating and ratings count. This answers our question if there is a relationship between the number of ratings given and the ratings? As the number of ratings increase, the rating for the book seems to move towards 4. The average rating seems to become sparse while the ratings count keeps on decreasing. There were some outliers present in total ratings, so we had to get rid of them. (**Refer to Fig 5**)

II. Relationship between average ratings and total number of pages

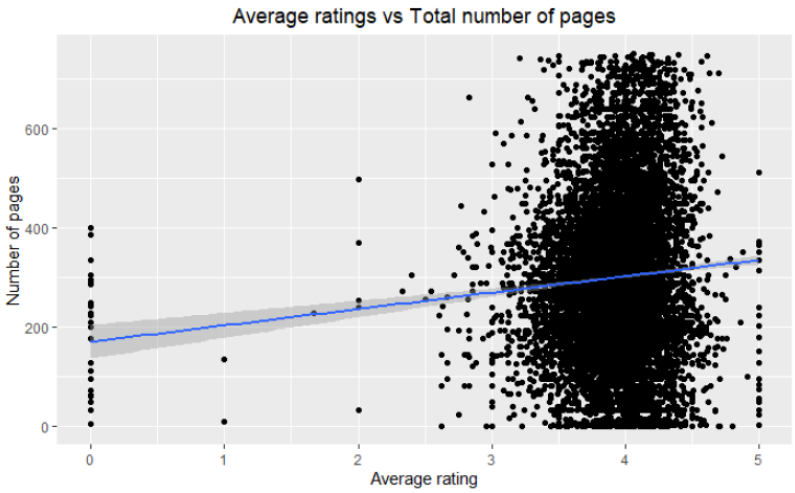

Fig 6. Relationship between average rating and total number of pages

We observe that books with the page number range of 200 – 400 have the highest rating, peaking near 250. This addresses our research question - does the number of pages have an impact on the ratings? We see that there is an impact, and our hypothesis is that may be due to the fact that people seem to prefer books with a moderate number of pages. There were some outliers present in total number of pages, so we had to get rid of them for this plot. (**Refer to Fig 6**)

III. Relationship between publication date and average ratings

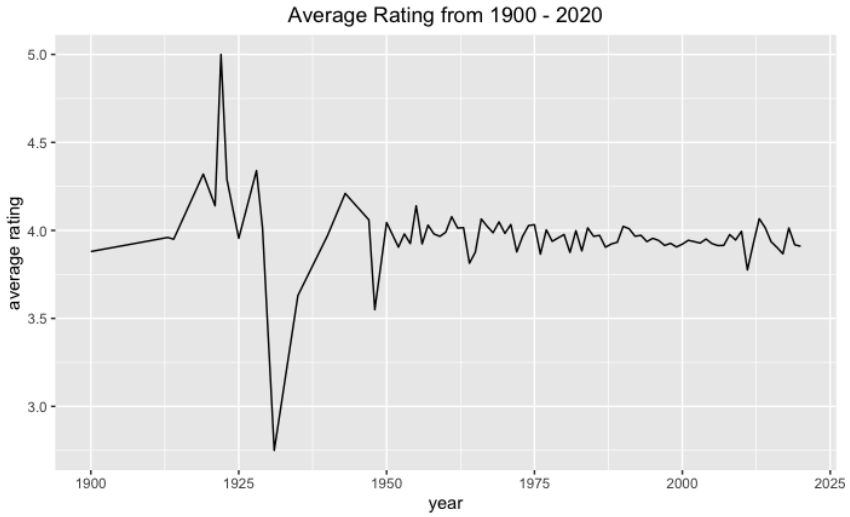

Fig 7. Relationship between average ratings and publication date

We plotted a general distribution of average ratings over the years – from 1900 to 2022, and there was a sharp dip in what is seen as early 1930s. This may be attributed to the Great Depression with lesser people with the ability to spend time reading and review books. (**Refer to Fig 7**)

IV. Relationship between publication date and total ratings

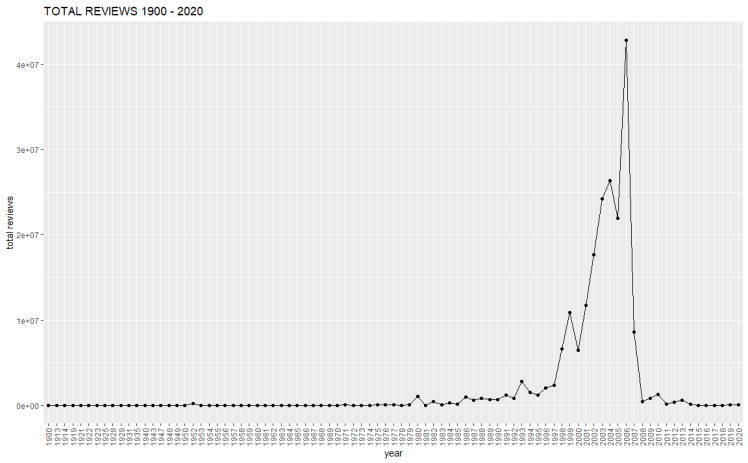

Fig 8. Relationship between total ratings and publication date

We notice that after 2000, the number of ratings were increasing and there's a very sharp dip during 2008, maybe due to the recession. It is also interesting to note that the years leading up to 2000 corresponded with the emergence of the internet in our society. It means that lay people will have more access to online rating systems and will be able to leave book reviews instead of just book critics. (**Refer to Fig 8**) The number of books published seems to be very low around the same time (1930s) with a sudden rise until 2005, and then a sharp drop which could probably be attributed to recession. The sudden dip in the previous graph with low average ratings (**Refer to Fig 7**) around 1930 could maybe be explained by the fact there were hardly any books published in that year. (**Refer to Appendix Fig 13)**

V. Authors' performance over years

author <chr>	average_rating <dbl>	count <int>
Aristophanes/F.W. Hall/W.M. Geldart	5	1
Chris Green/Chris Wright/Paul Douglas Gardner	5	1
Dennis Adler/R.L. Wilson	5	1
Elena N. Mahlow	5	1
Ian Martin/Katie Elliott	5	1
James E. Campbell	5	1
John Diamond	5	1
Julie Sylvester/David Sylvester	5	1
Keith Donohue	5	1
Laura Driscoll/Alisa Klayman-Grodsky/Eric Weiner	5	1

publisher <chr>	average_rating <dbl>	count <int>
Academica Press	5.000000	1
Boosey & Hawkes Inc	5.000000	1
Chartwell Books	5.000000	1
Courage Books	5.000000	1
Raintree	5.000000	1
Schirmer Mosel	5.000000	1
Square One Publishers	5.000000	1
T&T Clark Int'l	5.000000	1
Texas A&M University Press	5.000000	1
University Press of New England	5.000000	1

Fig 9. Authors and Publishers with high average rating

We considered most popular authors i.e., authors with high average ratings, but surprisingly, the top 10 such authors have only 1 book published. They also do not have many ratings. We also concluded that same is the case for publishers (**Refer to Fig 9**). So, for analysis of author's performance over time, we processed on the top 10 authors with most books published.

The top 10 authors with a large number of publications such as Agatha Christie, Stephen King, etc., followed a similar pattern in their performance over the years. There was a certain rise and dip seen in the

average rating (**Refer to Fig 10**), which could mean that they possibly did not perform well or did not publish any books in the period wherein we see a dip. This address our research question - Are the authors performing the same over time, with their new books? We do not notice any consistent average rating for any author.

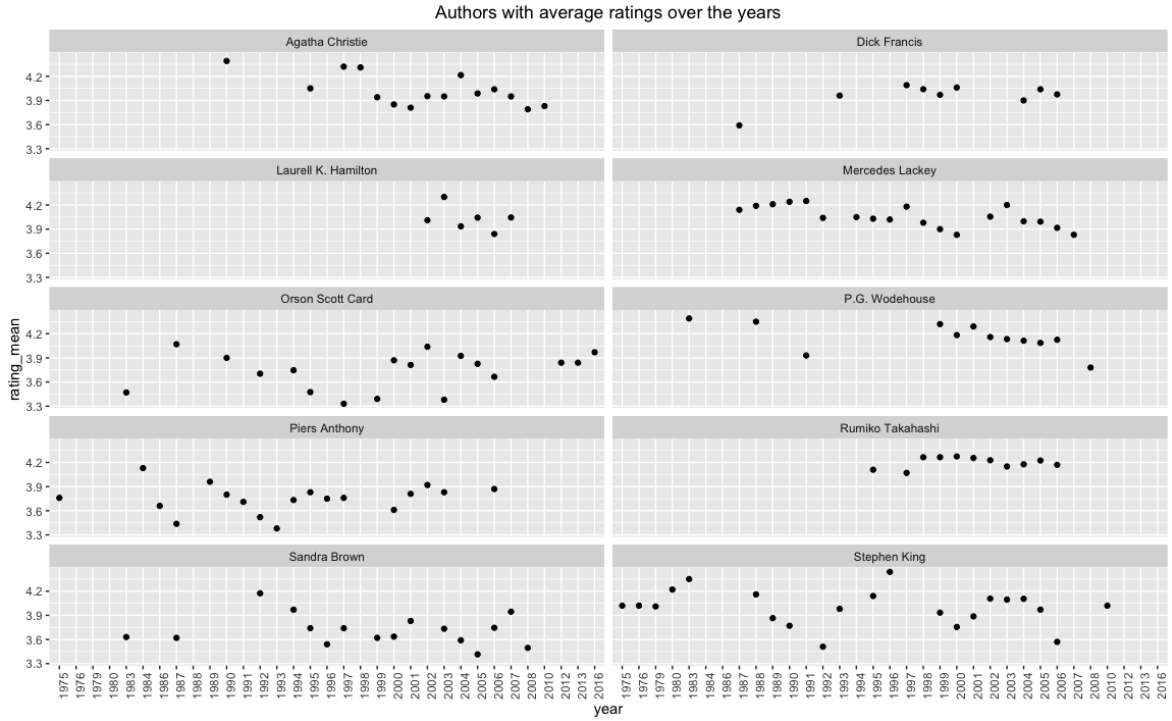

Fig 10. Average ratings of top authors over years

Correlation of the variables:

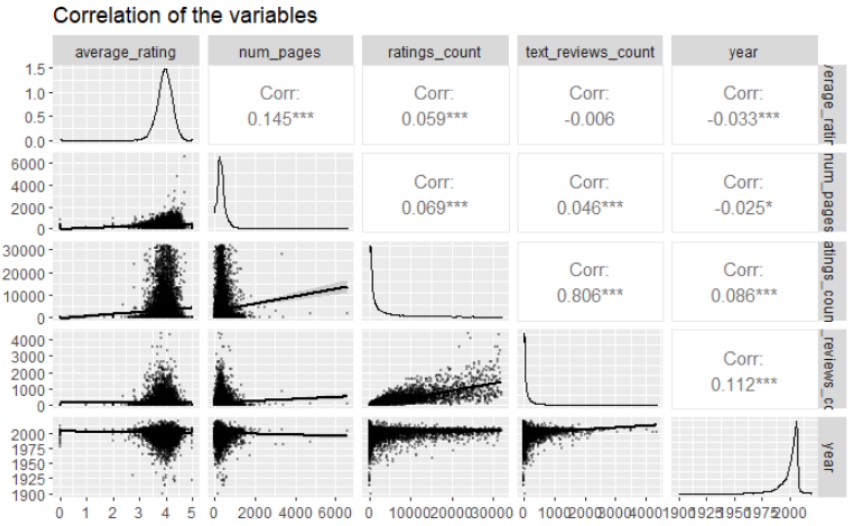

Fig 11. Correlation of the variables

We wanted to see how the variables are correlated and observed that **text_reviews_count** and **ratings_counts** have high positive correlation. We also tried to look into correlation between other variables too, but we did not get any useful insights from them.

Modeling:

We built a linear regression model with no interaction between features. The included features are **ratings_count**, **text_reviews_count, num_pages, year**. With this model, we wanted to see how all the features affected **Average Rating**, or whether they affected it at all. We have implemented the model with log values of **ratings_count** and **num_pages**. We did feature engineering with the variable **text_reviews_count** by dividing the values into **low reviews**, **medium reviews** and **high reviews** based on based on the median and 3rd quarter number. We used these features to build a linear regression model. Below is the summary of the model built:

```
lm(formula = average_rating ~ LogNumberPages + Logratings_count +
    year + text_reviews_count, data = train)
                 coef.est coef.se
(Intercept)      7.37     0.95
LogNumberPages   0.01     0.00
Logratings_count 0.03     0.00
year             0.00     0.00
text_reviews_count 0.00   0.00
---
n = 7650, k = 5
residual sd = 0.35, R-Squared = 0.03

Call:
lm(formula = average_rating ~ LogNumberPages + Logratings_count +
    year + text_reviews_count, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8011 -0.1677  0.0151  0.1982  1.2571

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.375e+00  9.520e-01   7.747 1.07e-14 ***
LogNumberPages     1.458e-02  4.266e-03   3.418 0.000633 ***
Logratings_count   2.977e-02  2.166e-03  13.744  < 2e-16 ***
year              -1.843e-03  4.758e-04  -3.874 0.000108 ***
text_reviews_count -1.419e-04  1.617e-05  -8.772  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.354 on 7645 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.03015,   Adjusted R-squared:  0.02964
F-statistic: 59.41 on 4 and 7645 DF,  p-value: < 2.2e-16
```

Fig 12. Linear Regression Model

We observe that all the **p-value** for all the variables is very low meaning that these variables are significant. But the **R-squared** value is very low meaning that the model is not very accurate. We tried to refine the model by dropping **Year** but that did not make the R-squared value any better.

Conclusion:

For this project we did not concentrate more on the modeling part as the results were not very accurate and it would need more work in terms of feature engineering and label encoding. However, we were able to generate insights from our exploratory data analysis and below are some of the important insights:

- We observed that there is a relationship between the average rating and ratings count, as number of ratings increase, rating of books seems to move towards 4 and as number decreases, the rating becomes sparse.
- We observed that readers prefer books with moderate range of pages (around 200 – 400) as highest number of ratings are given in that range.
- We also discovered that authors whose books are highly rated are Stephen King, P.G. Wodehouse, Rumiko Takahashi, Orson Scott Card, Agatha Christie, etc. considering the fact that their work has been present for a quite a while, spanning decades.
- We found that the most highly rated publishers are so due to having a limited number of highly satisfied readers.
- We also observe that the authors with highest ratings have a smaller number of total reviews, a smaller number of books published in the years and vice versa.
- The quantity of reader ratings has very less effect on the number of text reviews posted on the Goodreads website. Even if the number of reviews for a book is high, the ratings for that book may be poor.

Limitations and Future scope:

- The limitations for us were the categorical variables **authors, publishers, language_code** which have high number of categories and one-hot encoding may not be very accurate in this scenario. So, we need to explore more on how to incorporate these variables into model building for a better R-squared value.
- The models with all the categorical variables required high compute power and that created some issues for us too.
- We can answer the question - if books can be recommended based on their ratings? by checking if we can form clusters based on average ratings or number of ratings and thereby building a recommendation system that can give a list of books as recommendations based on an input book.
- We may also look at the model's future potential may be by considering attributes like genre. Furthermore, emotional analysis may be performed on the polarity of the users' text reviews.

Appendix:

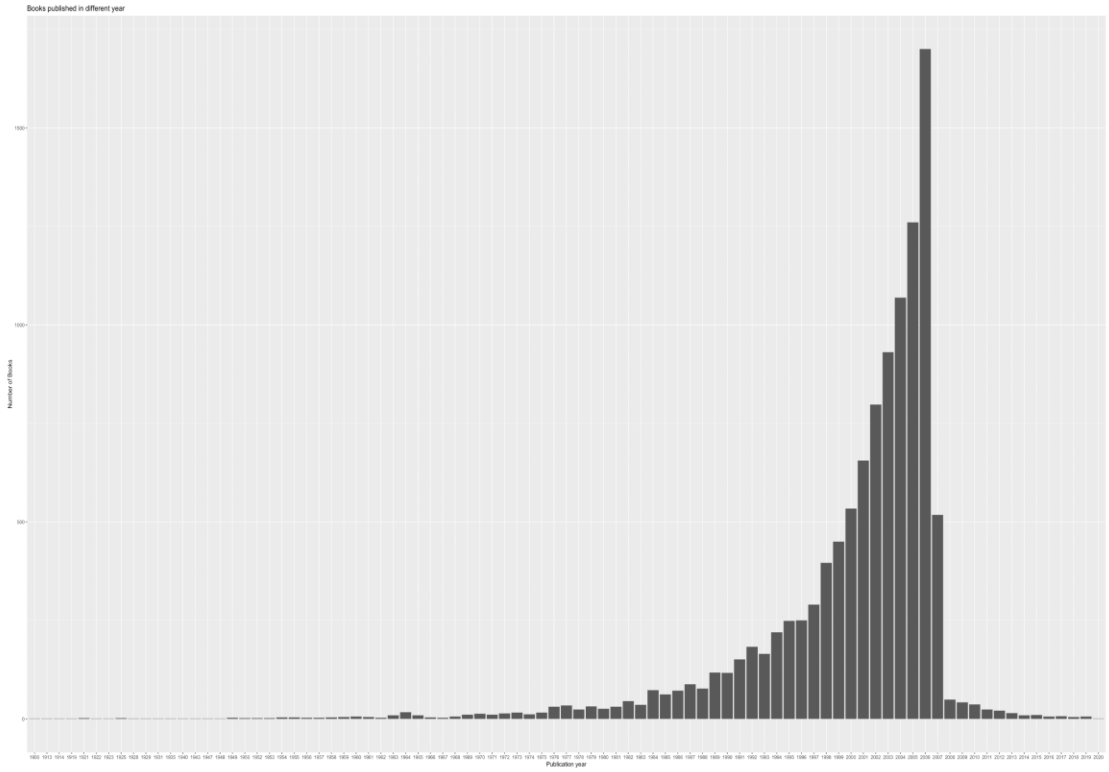

Fig13. Distribution of books published over years

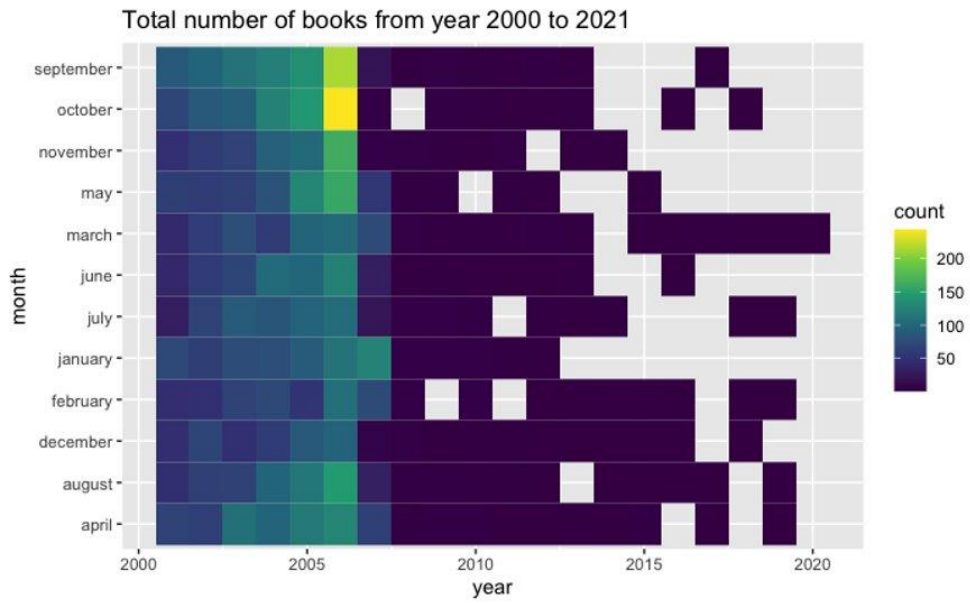

Fig1. Distribution of books published monthly

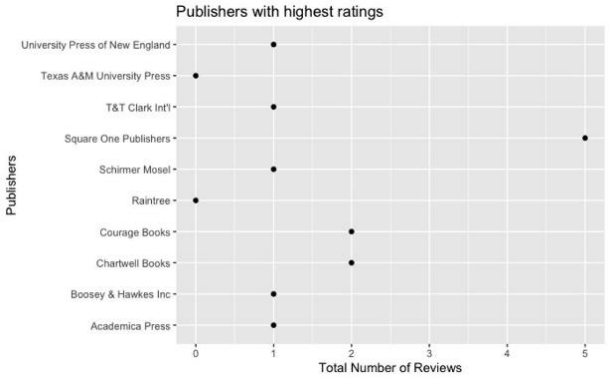

Fig15. Total number of reviews for publishers

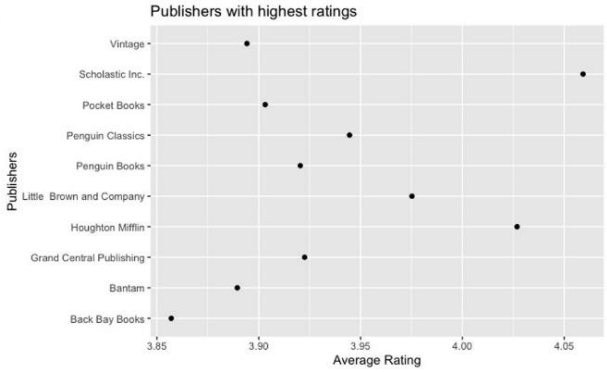

Fig16. Total average rating for top publishers