Book Rating Prediction Model

A-21 Python Project | DSTI

Niraj Chandrashekhar Deobhankar
niraj-chandrashekhar.deobhankar@edu.dsti.institute

Contents

Table of Figures

Introduction

Nowadays with so many books available, it can be hard to select the best ones to read. The dataset provided is a curation of Goodreads books based on real user information. It can be used for many tasks like predicting a book's rating or recommending new books. Here is an attempt to predicting a book's rating using two different algorithms.

Available data

1. **bookID:** A unique identification number for each book.
2. **title:** The name under which the book was published.
3. **authors:** The names of the authors of the book. Multiple authors are delimited by "/".
4. **average_rating:** The average rating of the book received in total.
5. **isbn:** Another unique number to identify the book, known as the International Standard Book Number.
6. **isbn13:** A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.
7. **language_code:** Indicates the primary language of the book. For instance, "eng" is standard for English.
8. **num_pages:** The number of pages the book contains.
9. **ratings_count:** The total number of ratings the book received.
10. **text_reviews_count:** The total number of written text reviews the book received.
11. **publication_date:** The date the book was published.
12. **publisher:** The name of the book publisher.

Objective

Using the provided dataset, we train a model that predicts a book's rating.

1. Data Examination and cleaning

a. Observations on data

```
100*df.isna().sum()/len(df)
```

```
bookID                0.000000
title                 0.000000
authors               0.000000
average_rating        0.000000
isbn                  0.000000
isbn13                0.000000
language_code         0.000000
  num_pages           0.000000
ratings_count         0.000000
text_reviews_count    0.000000
publication_date      0.000000
publisher             0.000000
Unnamed: 12          99.964051
dtype: float64
```

Figure 1-1 NA values in the data

When the data was imported, an unwanted column called unnamed:12 appeared that had more than 99% values.

b. Format of the data

Also, it was observed that, except for bookID, ratings_count, text_reviews_count all other columns have object as data type.

```
df.dtypes
```

```
bookID                int64
title                object
authors              object
average_rating       object
isbn                 object
isbn13               object
language_code        object
  num_pages          object
ratings_count         int64
text_reviews_count    int64
publication_date     object
publisher            object
Unnamed: 12          object
dtype: object
```

Figure 1-2 Initial data types

c. Data Cleaning

i. Four shifted columns

A cursory search on language_code to know the variety of languages of the books revealed that four rows were shifted to the right by one column.

```
df.loc[df["language_code"].isin(["9.78085E+12", "9.78156E+12", "9.78159E+12", "9.78067E+12"])]
```

	bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	pu
3348	12224	Streetcar Suburbs: The Process of Growth in Bo...	Sam Bass Warner	Jr./Sam B. Warner	3.58	674842111	9.78067E+12	en-US	236	61	6	4/2
4702	16914	The Tolkien Fan's Medieval Reader	David E. Smith (Turgon of TheOneRing.net	one of the founding members of this Tolkien w...	3.58	1593600119	9.78159E+12	eng	400	26	4	
5877	22128	Patriots (The Coming Collapse)	James Wesley	Rawles	3.63	156384155X	9.78156E+12	eng	342	38	4	1/1
8979	34889	Brown's Star Atlas: Showing All The Bright Sta...	Brown	Son & Ferguson	0	851742718	9.78085E+12	eng	49	0	0	

Figure 1-3 Columns that were shifted to the right

ii. Deleted columns

Once they were adjusted, the unwanted column unnamed:12 had 100% all NA values and hence it was deleted.

d. Data formatting

To use dates effectively, three different columns were formed that reflected day, month and year so that they can be used later for data analysis and modelling.

```
df[['month', 'date', 'year']] = df[['month', 'date', 'year']].astype(int)
```

```
df.head()
```

uthors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher	month	date	year
J.K. Rowling	4.57	439785960	9.78044E+12	eng	652	2095690	27591	09/16/2006	Scholastic Inc.	9	16	2006
J.K. Rowling	4.49	439358078	9.78044E+12	eng	870	2153167	29221	09/01/2004	Scholastic Inc.	9	1	2004
J.K. Rowling	4.42	439554896	9.78044E+12	eng	352	6333	244	11/01/2003	Scholastic	11	1	2003
J.K. Rowling	4.56	043965548X	9.78044E+12	eng	435	2339585	36325	05/01/2004	Scholastic Inc.	5	1	2004

Figure 1-4 Creation of three columns month, date and year

Then, format of publication date was converted into date and time format and age of the book is calculated in years from 'today'.

```
df['publication_date'] = pd.to_datetime(df['publication_date'], errors='coerce', format='%m/%d/%Y')
```

Now, let's calculate the time lapsed since the publication of the book as Age of the book.

```
now = pd.to_datetime('now')
```

```
C:\Users\ndeob\anaconda3\envs\class_project\lib\site-packages\pandas\core\arrays\datetimes.py:2199: FutureWarning: The parsing
of 'now' in pd.to_datetime without `utc=True` is deprecated. In a future version, this will match Timestamp('now') and Timestam
p.now()
  result, tz_parsed = tslib.array_to_datetime(
```

```
df['Age_of_book'] = (now.year - df['publication_date'].dt.year) - ((now.month - df['publication_date'].dt.month) < 0)
```

```
df.head()
```

e_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher	month	date	year	Age_of_book
4.57	439785960	9.78044E+12	eng	652	2095690	27591	2006-09-16	Scholastic Inc.	9	16	2006	15.0
4.49	439358078	9.78044E+12	eng	870	2153167	29221	2004-09-01	Scholastic Inc.	9	1	2004	17.0

Figure 1-5 Two columns that were deleted due to NA values

i. Removing columns

Just to be sure of the data, when NA values were checked once again. It was found that two of rows out of 11127 had NA values for publication dates. They were deleted.

e. Clean data source

A cleaner data source was then exported in CSV format for further analysis.

2. Data Analysis

a. Cursory understanding of data

The clean CSV file was imported and the data in it was examined. Some of the relevant finding are as below.

```
df.describe()
```

	average_rating	isbn13	num_pages	ratings_count	text_reviews_count	month	date	year	Age_of_book
count	11125.000000	1.112500e+04	11125.000000	1.112500e+04	11125.000000	11125.000000	11125.000000	11125.000000	11125.000000
mean	3.933613	9.759884e+12	336.315326	1.793868e+04	541.925213	6.546427	11.258427	2000.169169	21.318652
std	0.352473	4.429361e+11	241.104641	1.124894e+05	2576.402036	3.413982	10.279753	8.247779	8.271156
min	0.000000	8.987060e+09	0.000000	0.000000e+00	0.000000	1.000000	1.000000	1900.000000	2.000000
25%	3.770000	9.780350e+12	192.000000	1.040000e+02	9.000000	4.000000	1.000000	1998.000000	16.000000
50%	3.960000	9.780590e+12	299.000000	7.450000e+02	46.000000	7.000000	8.000000	2003.000000	19.000000
75%	4.140000	9.780870e+12	416.000000	4.991000e+03	237.000000	9.000000	20.000000	2005.000000	23.000000
max	5.000000	9.790010e+12	6576.000000	4.597666e+06	94265.000000	12.000000	31.000000	2020.000000	122.000000

Figure 2-1 Cursory glance at cleaned data

i. Average rating

On an average, books rating is about 3.9. Also, only 25 percentile of the books have ratings lower than 3.77 - a fact reflected in the standard deviation as well.

ii. Number of pages

Roughly we could say that number of pages are increasing with the increasing book ratings. But we need to verify the trend with more scientific means.

iii. Ratings count

It seems that less than 25 percentile books have about 100 or lesser ratings count. In general, that could be considered as good aspect of the data. Since number of ratings can affect the quality of average ratings for any book.

iv. Text Review counts

Hypothetically, the written text reviews a book received can also be measure of confidence on the quality of ratings. Here first 25 percentiles of the books have 9 or lesser text_review_counts whereas last 25 percentile of books have more than 237 text_count_reviews.

v. Month, Date and Year of Publication

This analysis may not be very important, but an interesting observation is about the range of publication date. There are some books that were published in the year 1900 whereas the latest year of publication is 2020.

vi. Age of the book

Average age of the book is 21 years. But at the same time, it is interesting to note that for first 25 percentile have age 16 years and the age of the first half of books is 19 years.

b. Average Ratings

Since we are attempting to predict average ratings, it would be very helpful to explore more data regarding average ratings.

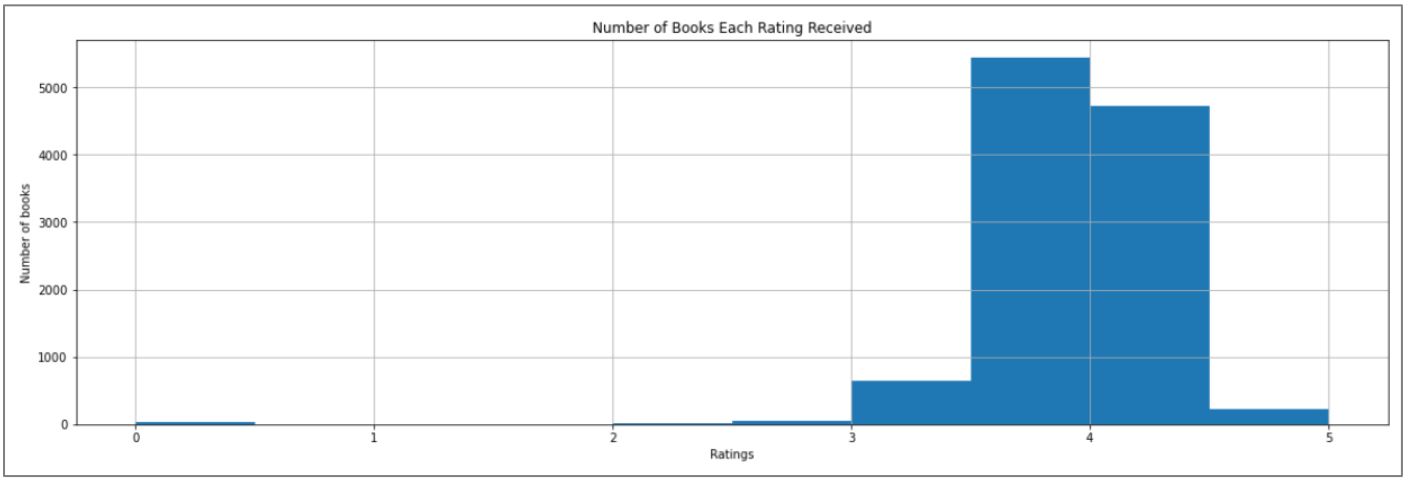

Figure 2-2 Distribution of books with their ratings

As confirmed earlier, about 50 percentile books have a rating close 4.0.

i. Average ratings per month

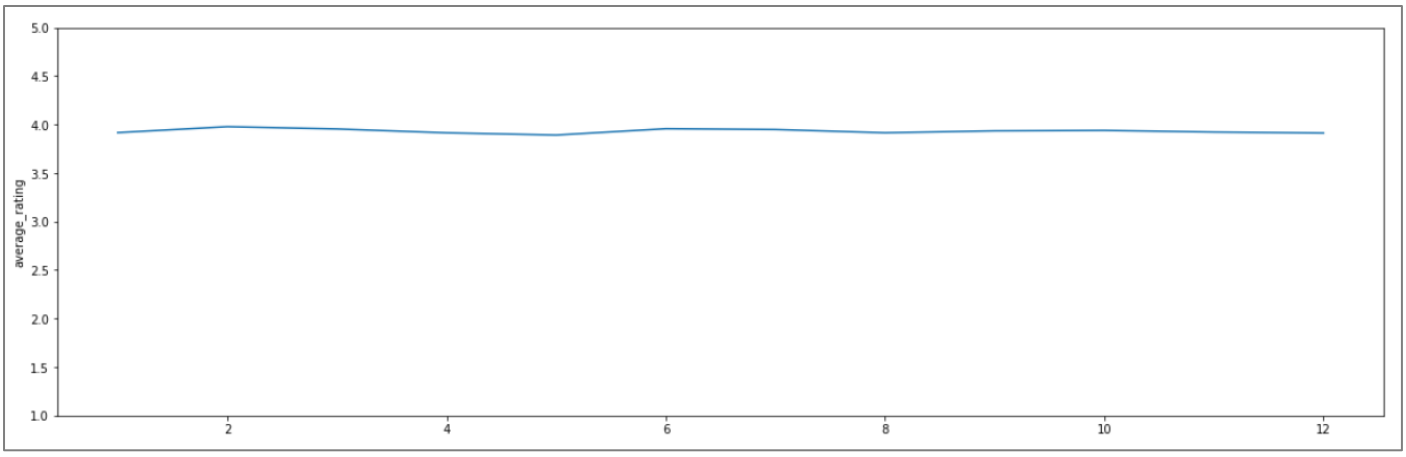

Figure 2-3 Average rating per month

Average rating per month hardly varies with months showing that month of publication has no influence on it.

ii. Average Ratings for top languages

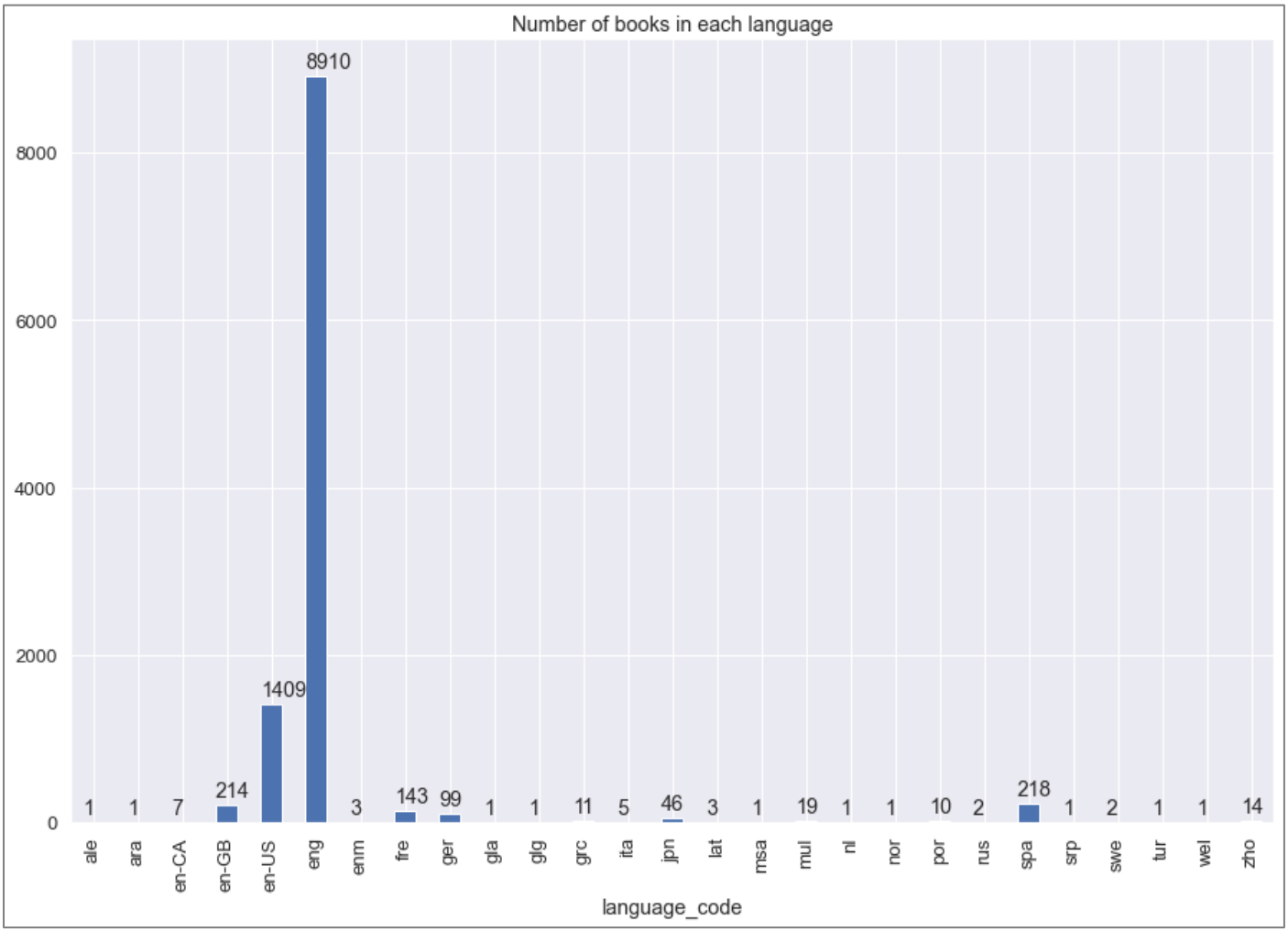

Figure 2-4 Number of books for each language

It is clear that almost all books are from English language. Also, the other peaks also belong to American and British English. The fourth language in which most books are published in the list is Spanish followed by French and German.

iii. Highly rated authors

Last quartile for average ratings starts at 4.14. So, any author having number of books can be called prolific. Let's explore which authors have highest number of books with more than that 4.14 rating.

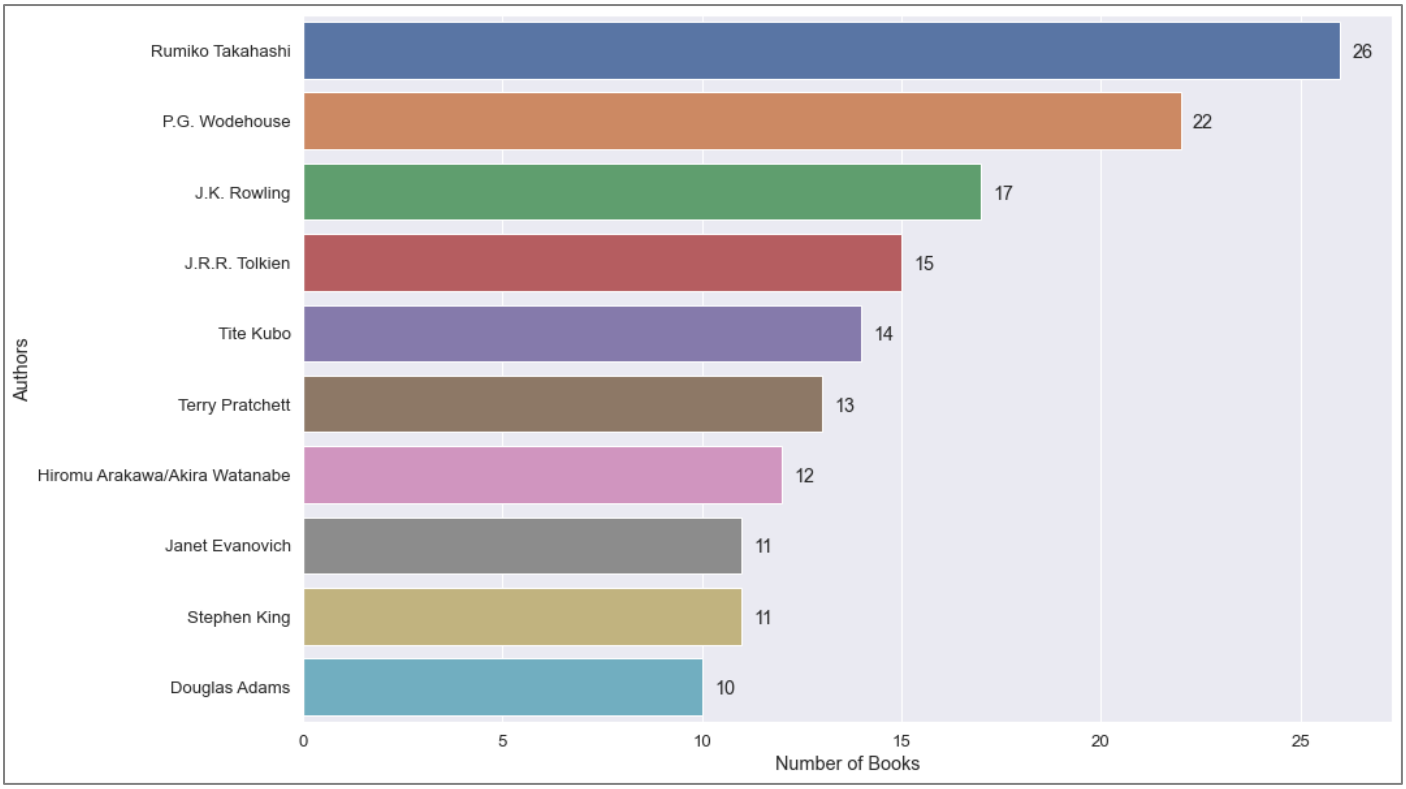

Figure 2-5 Top 10 highly rated authors

Rumiko Takahashi is the highest rated author since he or she has 26 books with average ratings in the last quartile.

c. Ratings count

i. Top 10 books with highest number of rating counts

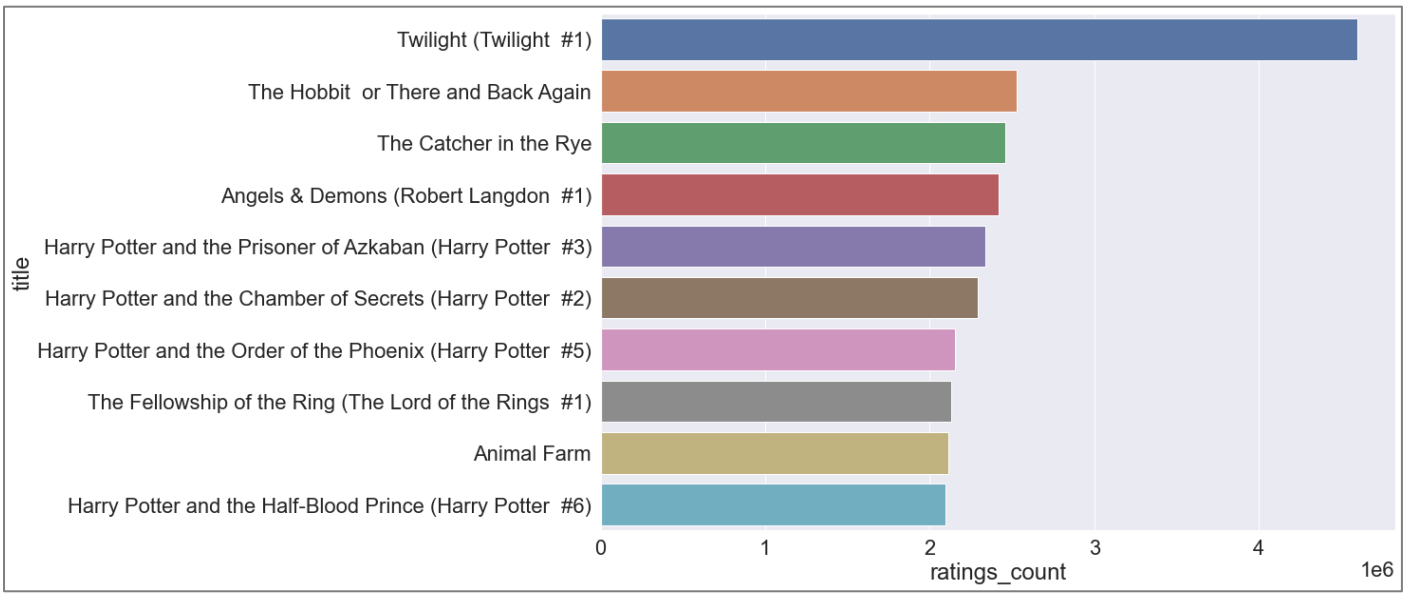

Figure 2-6 Top 10 books with highest number of rating counts

From the given data, the book Twilight has the highest number of ratings whereas rest of the 9 books have almost half of that.

ii. Top 10 books with highest text review counts

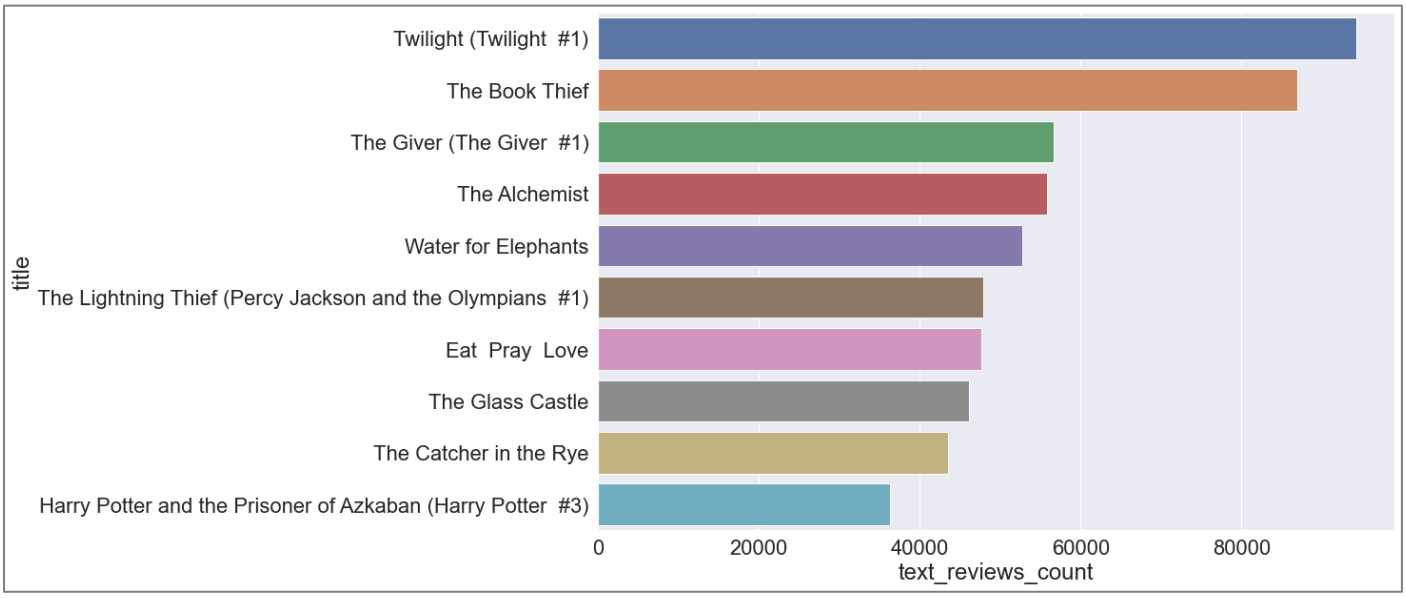

Figure 2-7 Top 10 books with highest text review counts

Although text review counts for the book Twilight are also highest, rest of the top books are not mentioned in the previous list.

d. Authors

i. Top 10 authors

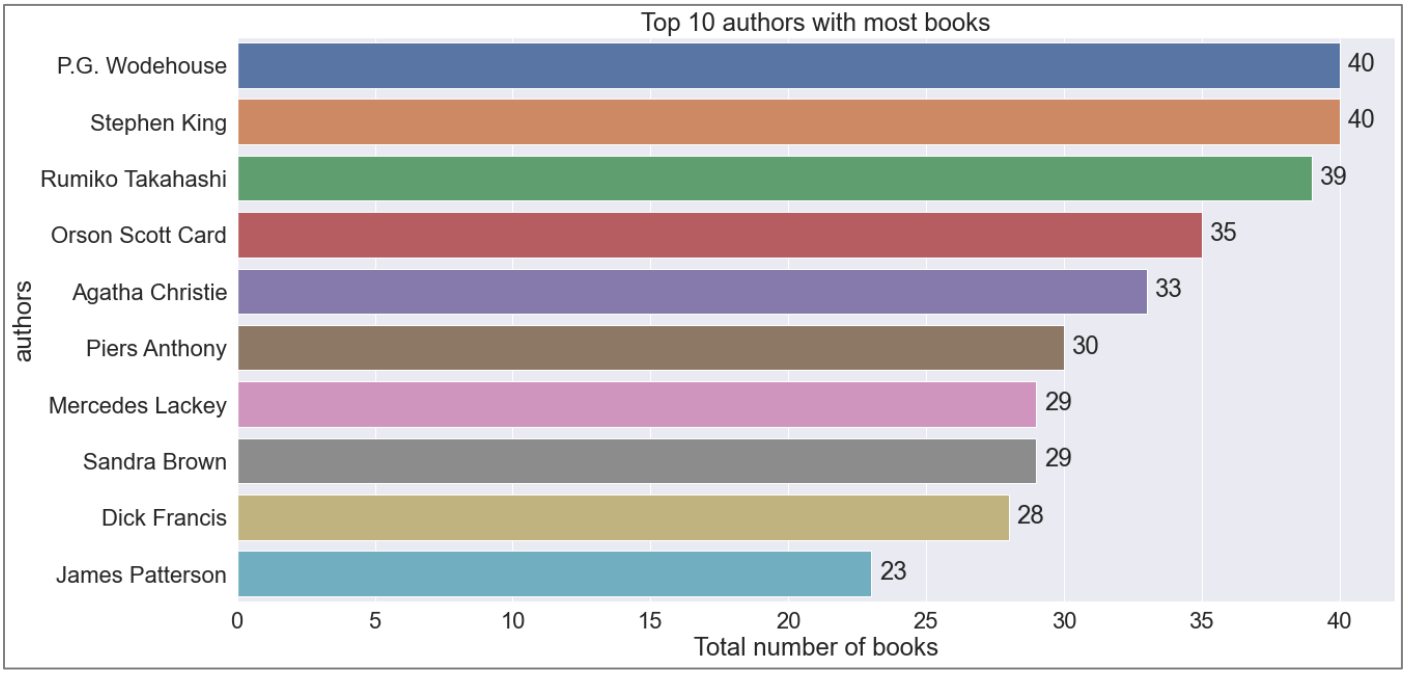

Figure 2-8 Top 10 authors

P.G. Woodhouse and Stephen King has the greatest number of books in the list closely followed by Rumiko Takahashi.

e. Study Corelation among variables

Just to have an idea if there is possibly any corelation between in the given dataframe, we may try to plot all variables that are available in integers as shown.

```
df.corr(method='pearson')
```

	average_rating	isbn13	num_pages	ratings_count	text_reviews_count	month	date	year	Age_of_book
average_rating	1.000000	-0.002015	0.150763	0.038209	0.033740	0.023233	-0.000771	-0.028790	0.027343
isbn13	-0.002015	1.000000	-0.009836	0.005492	0.008150	-0.010103	0.005273	-0.000343	0.000872
num_pages	0.150763	-0.009836	1.000000	0.034387	0.037043	0.025479	0.021943	-0.018956	0.016957
ratings_count	0.038209	0.005492	0.034387	1.000000	0.865978	-0.015694	-0.001674	0.044554	-0.043576
text_reviews_count	0.033740	0.008150	0.037043	0.865978	1.000000	-0.024909	0.010930	0.066896	-0.065358
month	0.023233	-0.010103	0.025479	-0.015694	-0.024909	1.000000	0.040779	0.022746	-0.075619
date	-0.000771	0.005273	0.021943	-0.001674	0.010930	0.040779	1.000000	0.056704	-0.058682
year	-0.028790	-0.000343	-0.018956	0.044554	0.066896	0.022746	0.056704	1.000000	-0.998173
Age_of_book	0.027343	0.000872	0.016957	-0.043576	-0.065358	-0.075619	-0.058682	-0.998173	1.000000

Figure 2-9 Corelation among numerical variables

As such it seems that there is hardly any mentionable corelation of any variable with the book's ratings. So, it is important to study the individual relation.

i. Relation between number of pages and average ratings

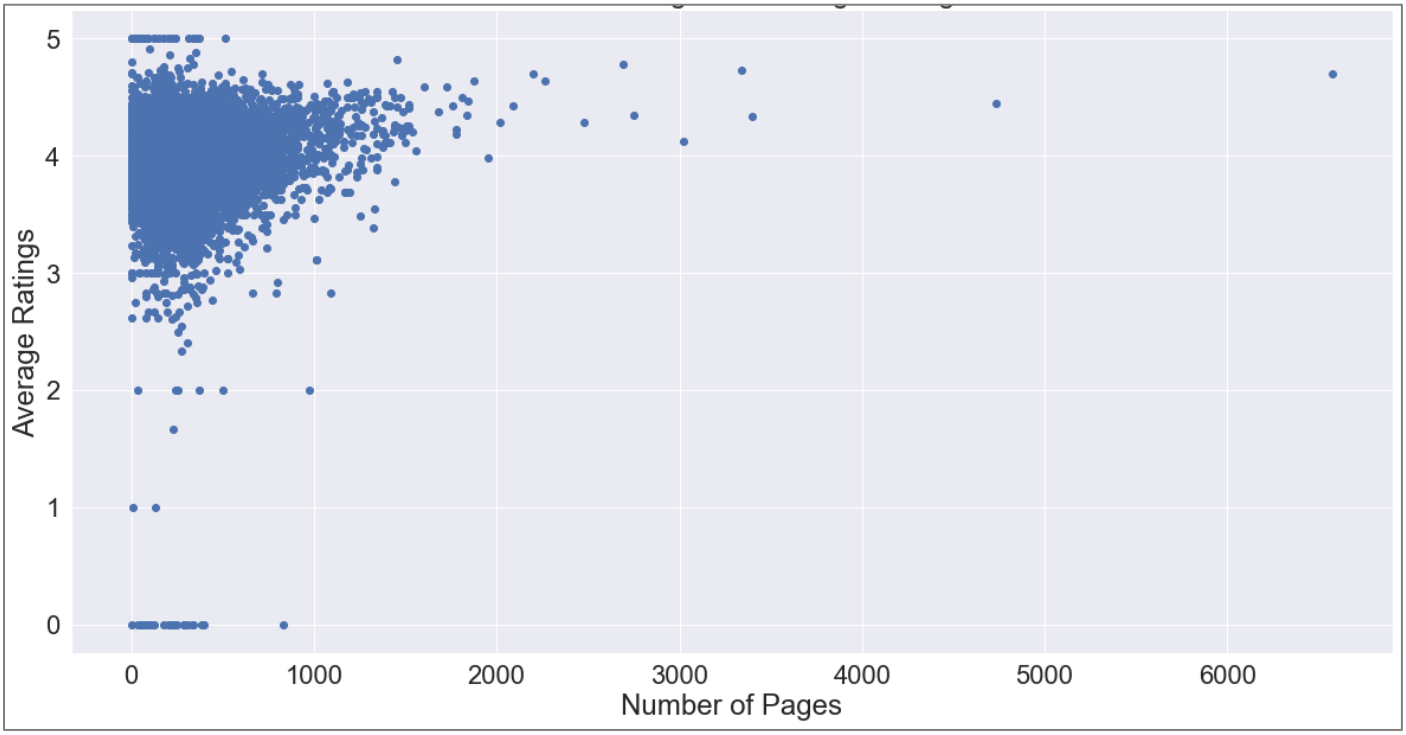

Figure 2-10 Relation between number of pages and average ratings

From the chart above it is clear that there is hardly any corelation between the number of pages and ratings. However, is is interesting to see that above roughly 1600 pages, the book always has got really good ratings.

It could happen that such long books are read by only few people and very few people provided any feedback on such books.

ii. Number of pages vs Ratings Count

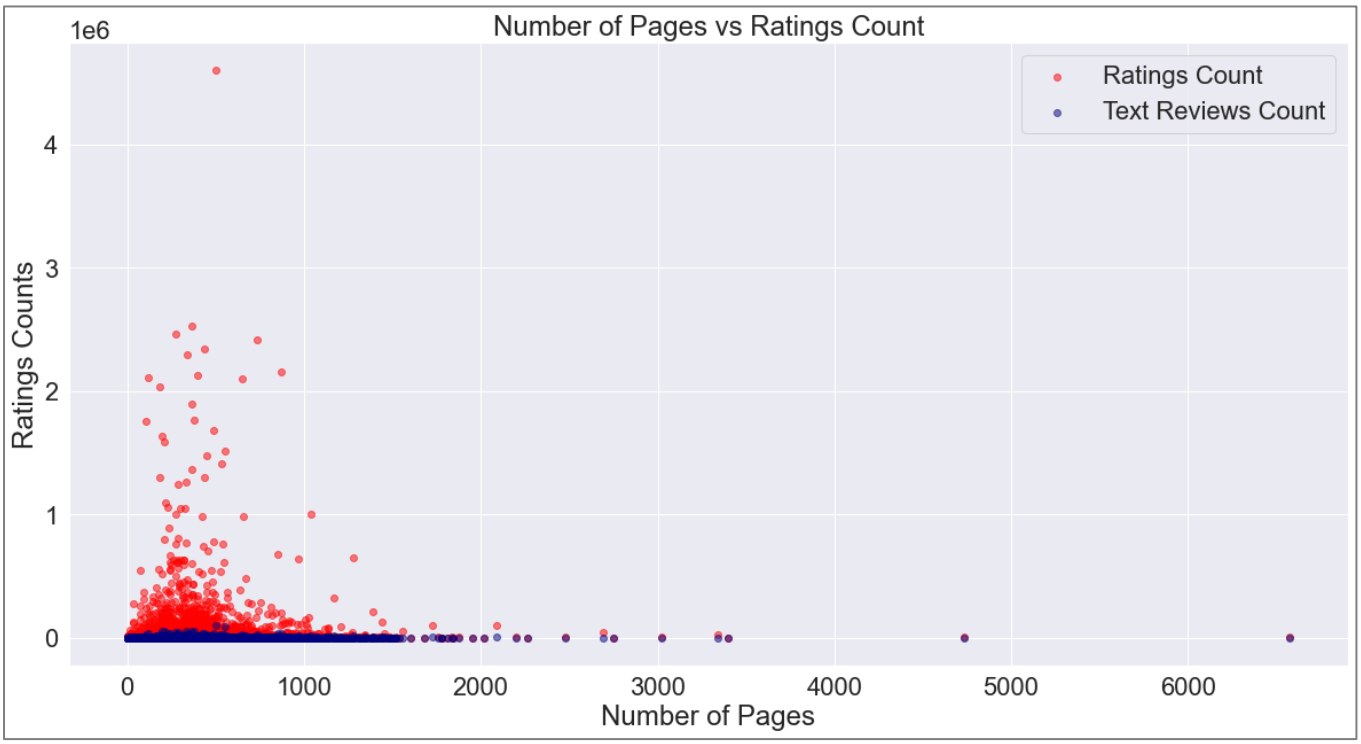

Figure 2-11 Relation between number of pages and ratings count

As expected, it is clear that number of ratings, even in the form of texts is rare for books with pages more than 2000 from this graph. To explore more of these things lets chart the outliers for number of pages.

iii. Relations between Average ratings and Ratings count and Text Ratings count

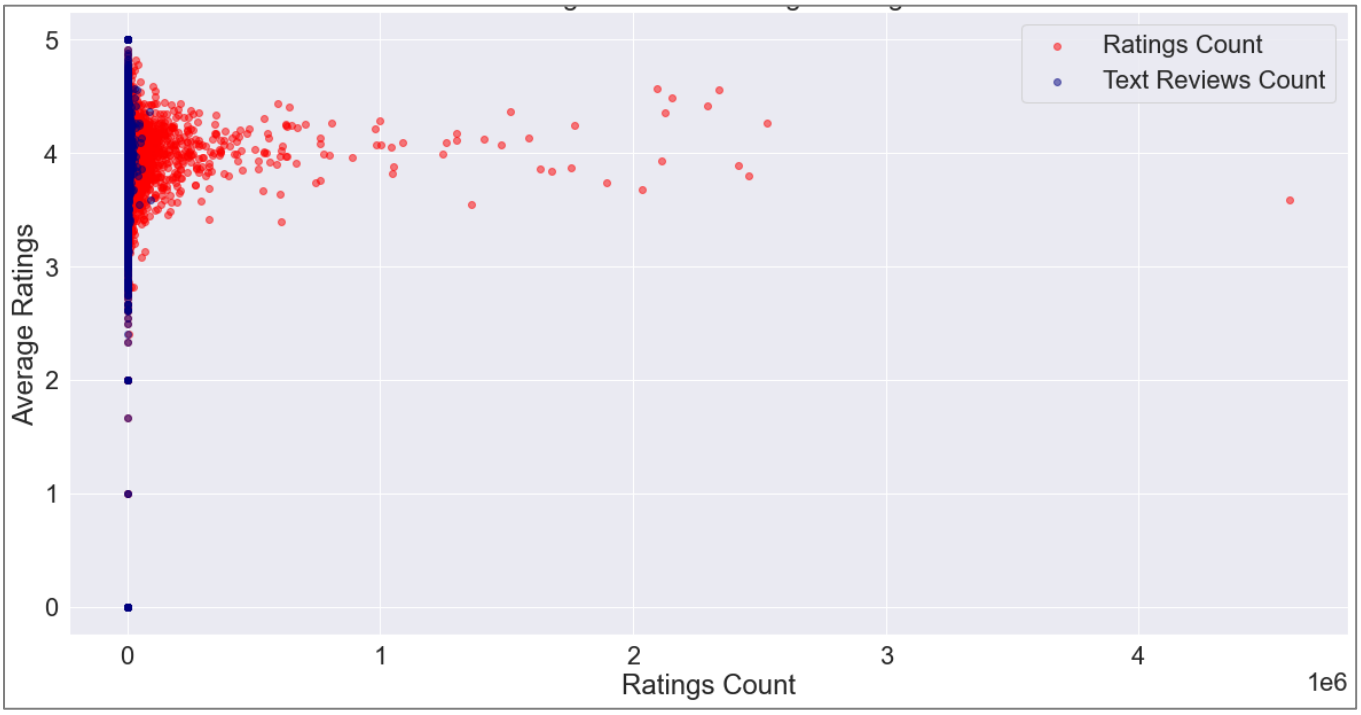

Figure 2-12 Relation between Average Ratings and Rating count and Text rating count

From the above chart, it is clear that as number of counts are increasing, we may see that trend seems a bit clearer. So, the books with lesser ratings counts have no clarity for the trend for average ratings.

f. Outliers

Let's explore if there are any outliers among the important variables.

i. Outliers for average ratings in each month

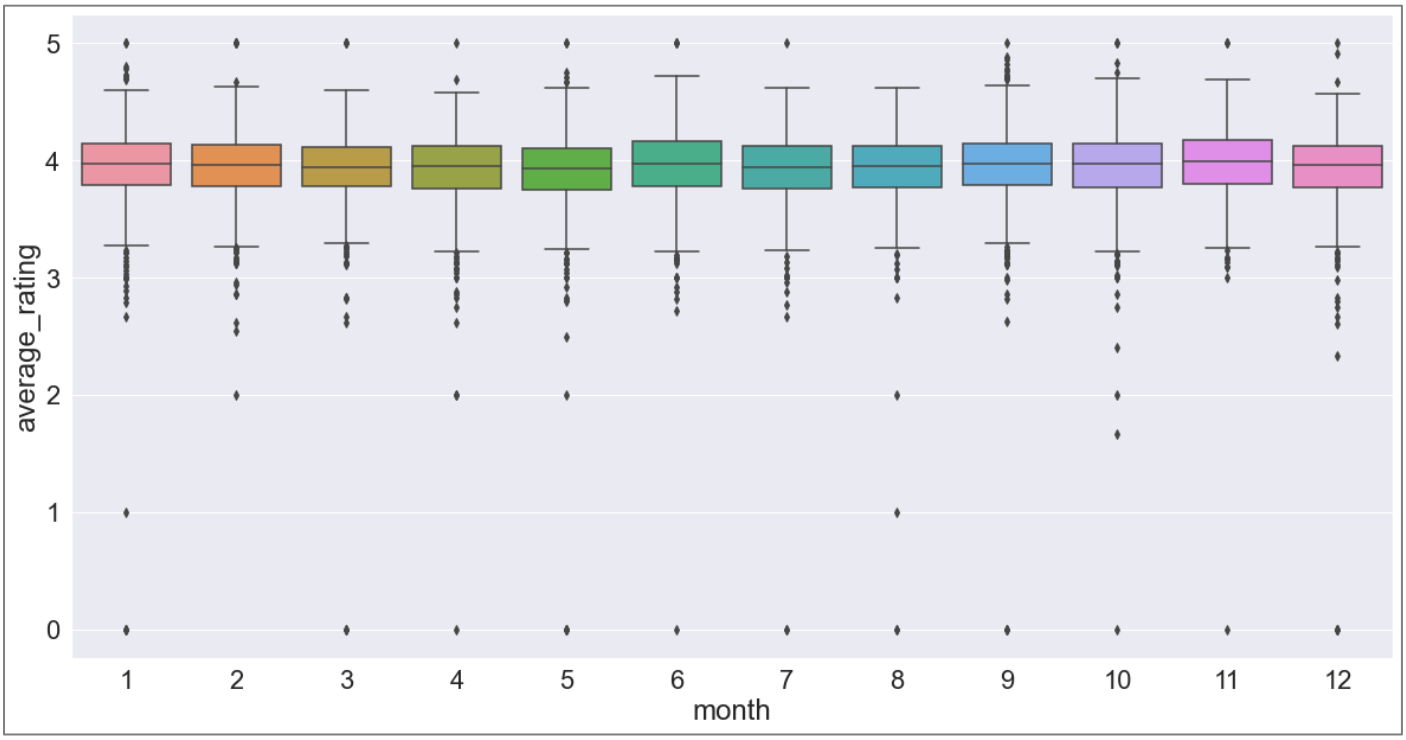

Figure 2-13 Outliers for average ratings in each month

Interestingly, there is no book in February that has an average rating lesser than 2. As expected almost months have an average rating of 4 and a few outliers lie on both sides.

ii. Ratings Count

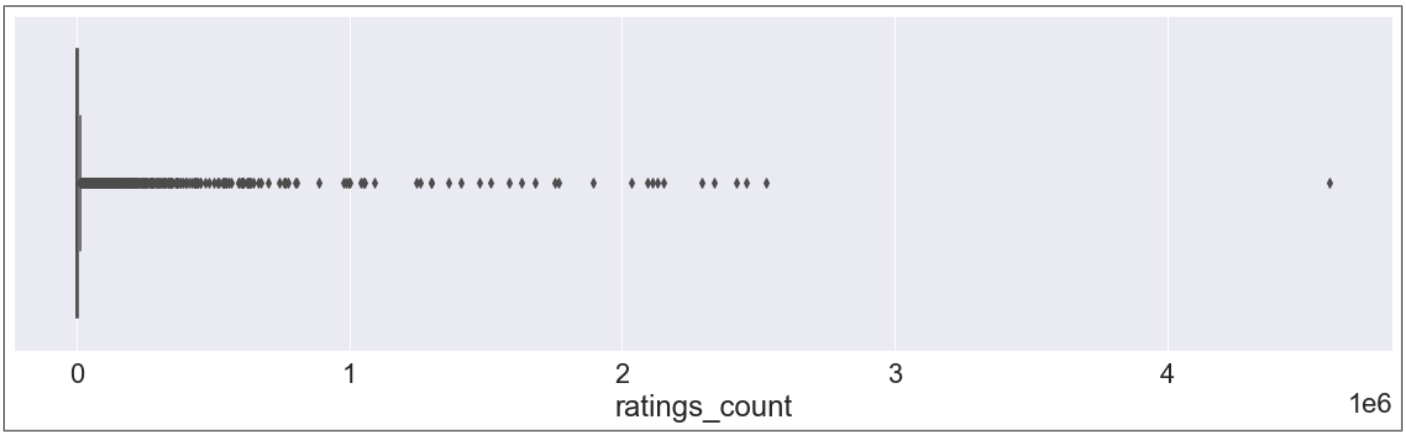

Figure 2-14 Outliers for Ratings count

iii. Text Review Count

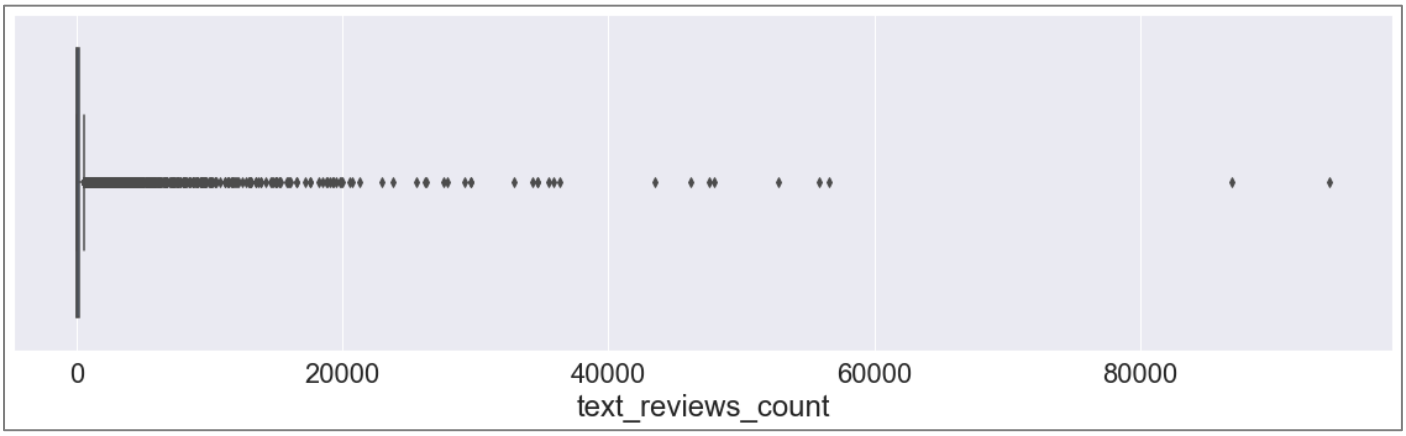

Figure 2-15 Outliers for Text review counts

3. Model Building

Once we understand that the corelation among different variables is a weak, we try two algorithms for predicting average ratings for the given data.

a. Feature Engineering

For meaningful result, we prune the data even further.

i. Delete unwanted columns.

ISBN and ISBN13 are assigned randomly. So, they may not actually influence the average ratings. Also, the essence of publication date is taken into date, month, year, and age of the book. So, these three columns are deleted.

ii. Create dummies for language_code

Languages are in the form of categorical value. To use them in the actual analysis, we can create dummies for them.

iii. Data Model Creation

The model data with 11125 rows are divided into two parts, training data and test data. As per convention, number of rows in the training data is 80% or 8900 while number of rows in the test data is 20% or 2225.

b. Model Selection

Before actually training the model with any method, let's try to understand the quality of model for Ordinary Least Squared Linear Regression.

i. Coefficient:

The coefficient term tells the change in Y for a unit change in X. In this case, it is far off from 1 in most of the cases except for a few languages. It can be coincidence.

ii. Standard error of parameters:

Standard error is also called the standard deviation. Standard error shows the sampling variability of these parameters. As expected, all categorised parameters are meaningless.

iii. t-statistics and p-value:

They help us to know whether null hypothesis is true or not. From almost all given p-values we fail to reject the null hypothesis.

iv. R – squared value:

R2 is the coefficient of determination that tells us that how much percentage variation independent variable can be explained by independent variable. Unfortunately, only 4.9% variation can be explained by given parameters.

v. Adj. R-squared:

This is the modified version of R-squared which is adjusted for the number of variables in the regression. It increases only when an additional variable adds to the explanatory power to the regression. Since, there is hardly change in this value, it is again clear that it is hard to predict ratings.

vi. Prob(F-Statistic):

This tells the overall significance of the regression. As per the above results, probability is close to zero. This implies that overall the regressions is meaningful.

vii. Omnibus / Prob(Omnibus):

One of the assumptions of OLS is that the errors are normally distributed. We hope to see the value for Omnibus close to zero which would indicate normalcy. Prob(Omnibus) is supposed to be close to the 1 in order for it to satisfy the OLS assumption. However, the values from the test indicate that the data is not normally distributed at all.

Looking at all constraints, we select two regression models: linear and random forest.

c. Linear Regression

The first model we try here is linear regression. After modelling the data, we find that following.

i. Comparison of actual ratings vs predicted ratings: Linear Regression – 10 Examples

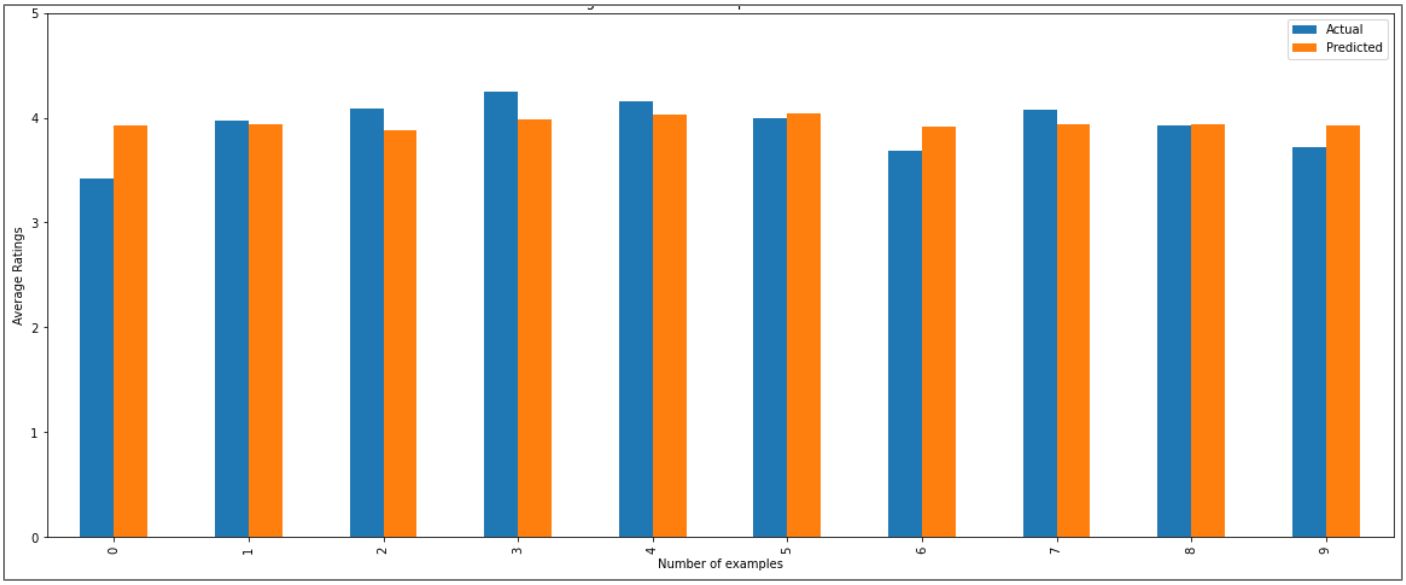

Figure 3-1 Comparison of actual ratings vs predicted ratings: Linear Regression – 10 examples

From the figure above it seems that the model cannot predict the ratings precisely but could give some indication. However, this chart is for only 10 examples. So, it would be interesting to find such comparison for all.

ii. Comparison of actual ratings vs predicted ratings: Linear Regression

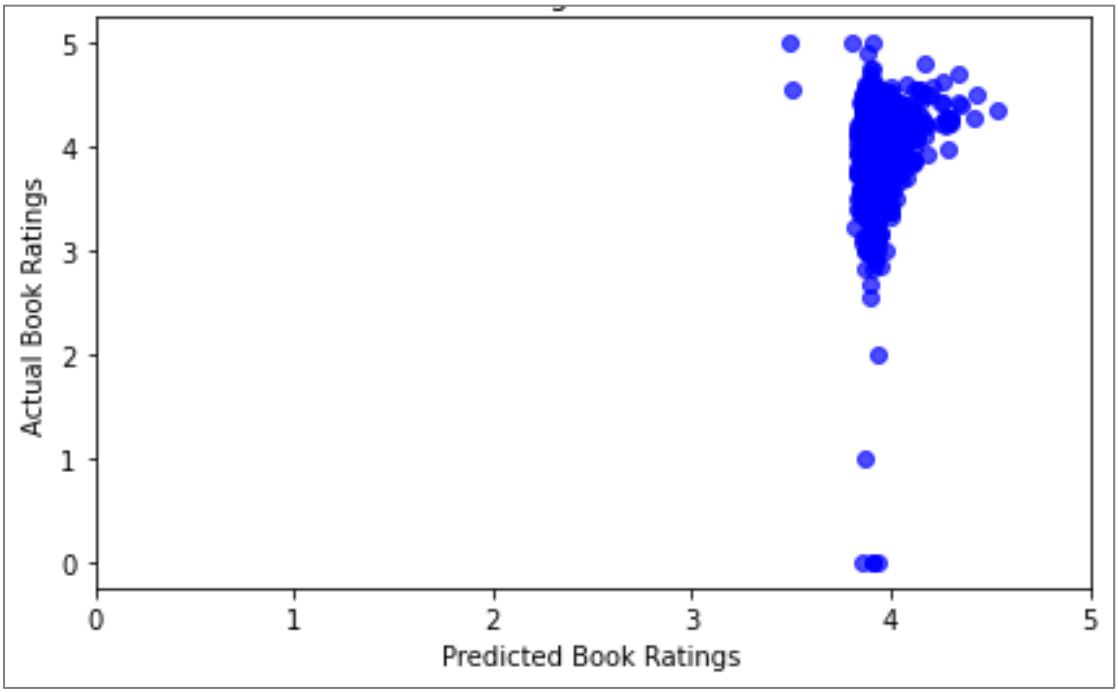

Figure 3-2 Comparison of actual ratings vs predicted ratings: Linear Regression

Visually, it seems that the model works okay for all books that may have ratings at around 4. For any rating below that, it is terrible model.

d. Random Forest

Then we try to random forest algorithm for modelling.

i. Comparison of actual ratings vs predicted ratings: Random Forest – 10 examples

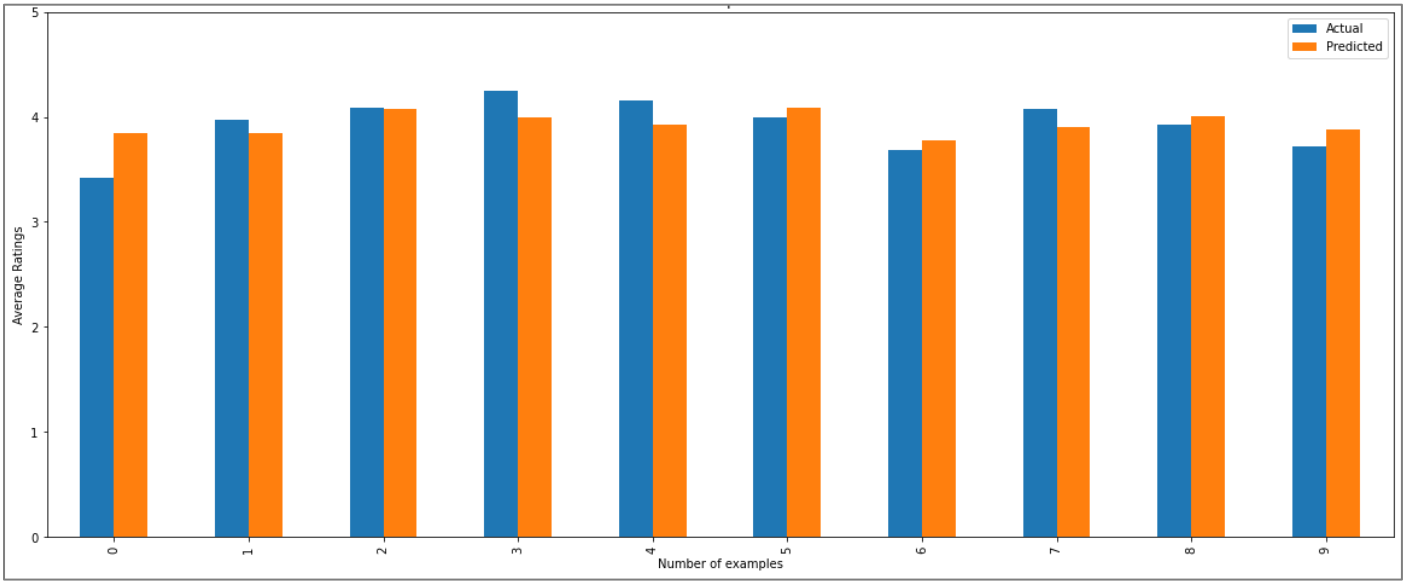

Figure 3-3 Comparison of actual ratings vs predicted ratings: Random Forest – 10 examples

ii. Comparison of actual ratings vs predicted ratings: Random Forest

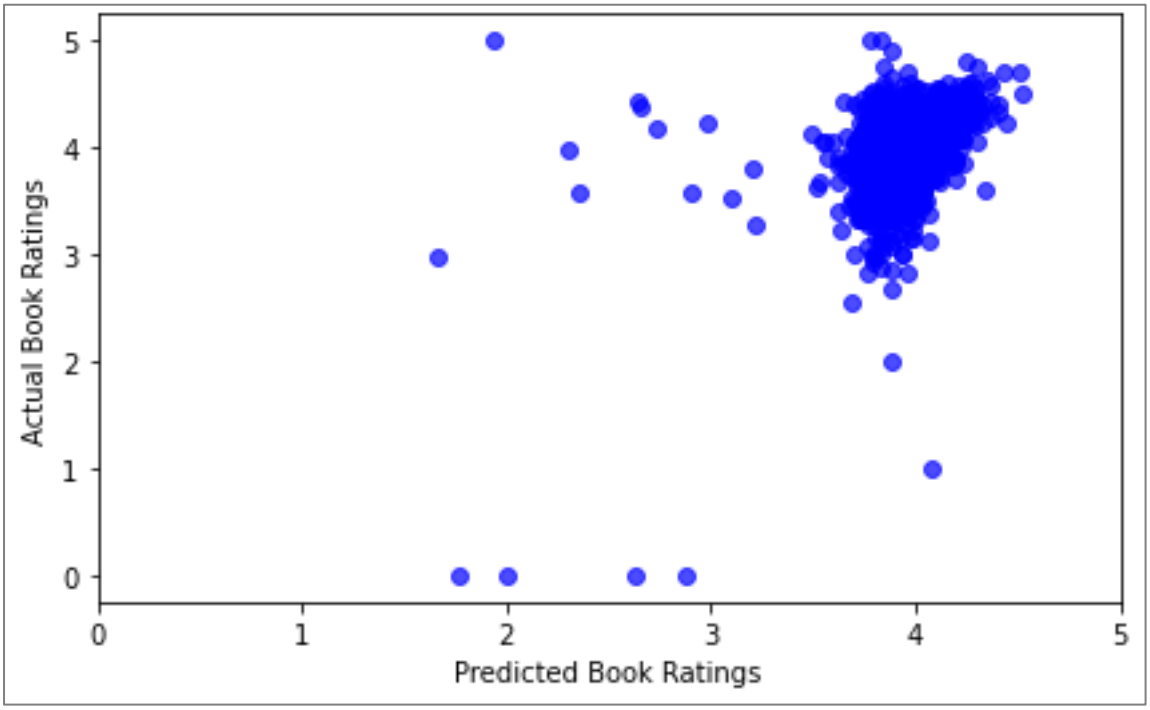

Figure 3-4 Comparison of actual ratings vs predicted ratings: Random Forest

e. Linear Regression: Lighter data

In the previous Linear Regression model, we used some of the categorical variables. Also, we did not convert the average ratings into classes. Here, we attempt a simplistic model without categorical variables and target in the form of classes.

So, we take only 'average_rating', 'num_pages', 'ratings_count', 'text_reviews_count', 'month', 'date', 'year', 'Age_of_book' as parameters. Further, we round off average ratings to integers as follows.

```
df_model1.head()
```

bookID	average_rating	num_pages	ratings_count	text_reviews_count	month	date	year	Age_of_book
1	5	652	2095690	27591	9	16	2006	15.0
2	4	870	2153167	29221	9	1	2004	17.0
4	4	352	6333	244	11	1	2003	18.0
5	5	435	2339585	36325	5	1	2004	18.0
8	5	2690	41428	164	9	13	2004	17.0

Figure 3-5 Lighter model for linear regression

i. Comparison of actual ratings vs predicted ratings: Linear Regression – Lighter -10 examples

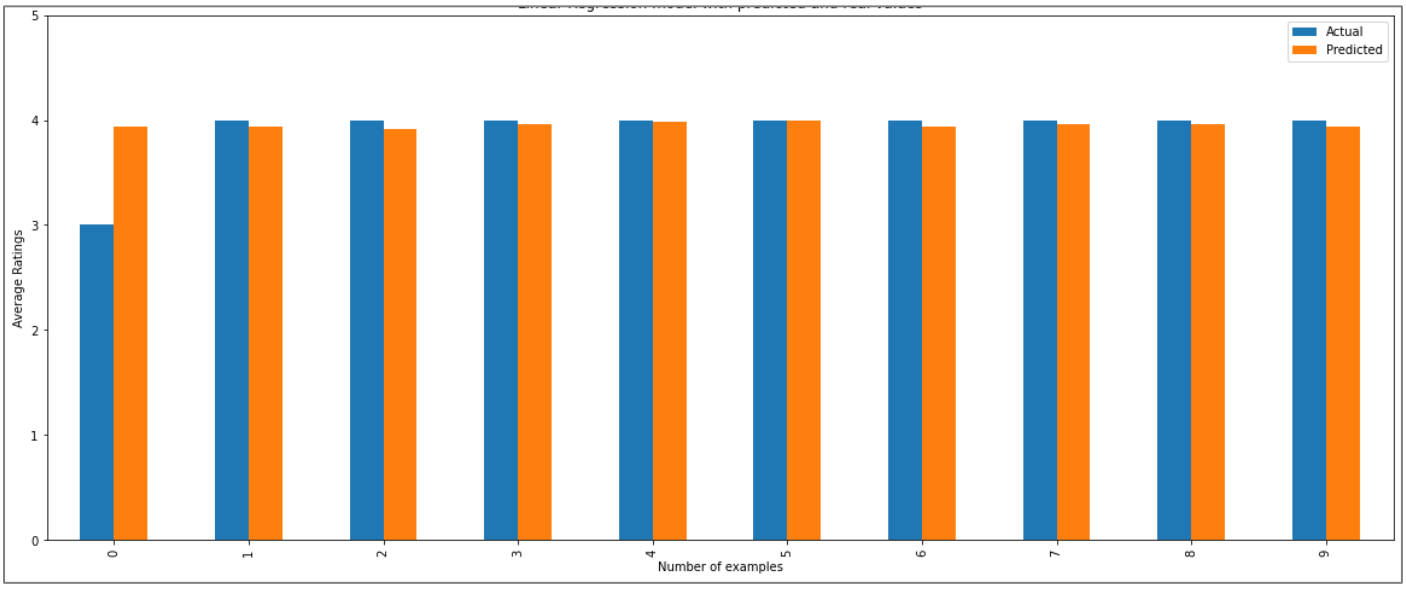

Figure 3-6 Comparison of actual ratings vs predicted ratings: Linear Regression – Lighter - 10 examples

ii. Comparison of actual ratings vs predicted ratings: Linear Regression - Lighter

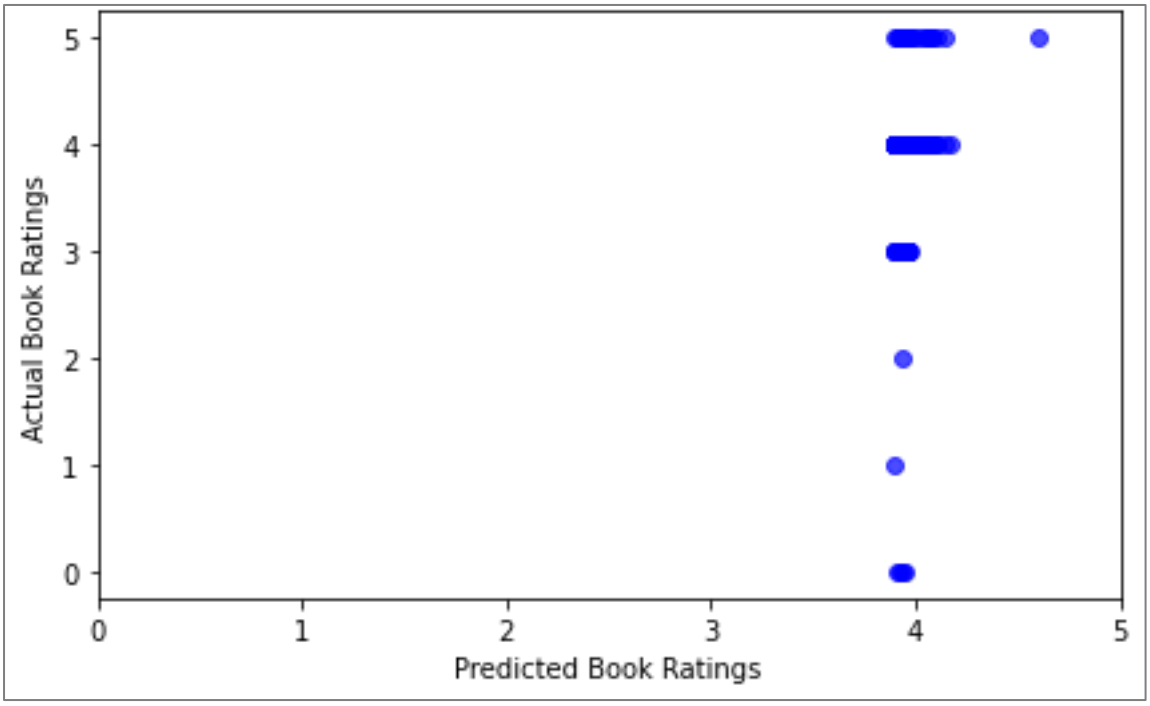

Figure 3-7 Comparison of actual ratings vs predicted ratings: Linear Regression – Lighter

f. Evaluation Matrix

For regression models, results are mainly evaluated with Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared (R2). The following table summarises the results.

	Linear Regression		Random Forest Regression		Linear Regression – Lighter data	
	Train	Test	Train	Test	Train	Test
Mean Absolute Error (MAE)	0.230	0.220	0.080	0.211	0.144	0.137
Mean Squared Error (MSE)	0.121	0.109	0.015	0.094	0.121	0.106
Root Mean Squared Error (RMSE)	0.048	0.039	0.884	0.165	0.006	0.018
R-squared (R2)	0.348	0.330	0.121	0.307	0.348	0.326

i. Mean Absolute Error (MAE)

In statistics, mean absolute error (MAE) is an arithmetic average of the absolute errors.

ii. Mean Squared Error (MSE)

Mean Squared Error (MSE) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and the actual value.

iii. Root Mean Squared Error (RMSE)

Root Mean Squared Error is a common way of measuring the quality of the fit of the model.

When all these three values are closer to 0 rather than 1, one can say that the model fits the data well.

- From the table, one can observe that MAE and MSE for random forest model is lower than the rest of the two models. Also, making the data lighter doesn't make much difference.
- However, RMSE for random forest is higher than that for the two linear regression model. One of explaining this difference is as the amount of data increases MAE has lower values than RMSE.
- The difference between MAE, MSE and RMSE values for train and test data for linear regression models is smaller compared to that for random forest regression model.
- There is a significant difference between MAE and RSME values for random forest regression model.

iv. R-squared (R2)

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

- For all three models, one can say that we have low positive corelation.
- While for both linear regression models the R2 value for training data is higher than the test data, for random forest model it is not the case which is a bit surprising.