University of Exeter coursework header sheet

Extension
211643

BEMM457J Topics in Business Analytics (A, TERM2 2021/2)

1073345

Coursework: Individual coursework
Submission Deadline: Mon 2nd May 2022 12:00
Personal tutor: Dr Frank Donkor

710078558

Marker name: N/A
Word count: 3550

By submitting coursework you declare that you understand and consent to the University policies regarding plagiarism and mitigation (these can be seen online at www.exeter.ac.uk/plagiarism, and www.exeter.ac.uk/mitigation respectively), and that you have read your school's rules for submission of written coursework, for example rules on maximum and minimum number of words. Indicative/first marks are provisional only.

Goodreads Books Data Analysis

Table of Contents

1.Introduction:

This report is a mini analytics report on the data collected from a social cataloguing website called Goodreads. Goodreads is an American website, which is currently subsidized by Amazon. It is a social network site, like Facebook, but for book enthusiast where users can sign up and register to access its vast book collection it has to offer. Goodreads was founded in December 2006 by Otis Chandler and Elizabeth Khuri Chandler. When started it had only 650,000 user and 10,000,000 books added. The cite currently has 90 million subscribers with over 395 million books catalogued. (Wikipedia Contributors, 2019).

Due to its large and diverse user base, the site generates lots of interesting data. Goodreads is usually used by its users for browsing, reading or reviewing books from its catalogue. (Thelwall & Kousha, 2016).It is also used for user to create their own groups of book suggestions, surveys, polls, blogs, discussions and track their reading activity. The data will be a coalition of data related to books as well as the data which reader/user of Goodreads website generated.

2. Method & Approach:

This report, will list the findings of an Exploratory data analysis on data generated by the Goodreads website to establish an working idea of how different fields like the ratings of a book in the dataset change with respect to other attributes in the dataset. I will also analyse the trends of how a book performs over time and why a particular book is being rated the best over other books.

The main purpose of this analysis is to establish trends among the many attributes which drive reader to choose a particular book. Some of the key area we will be covering are:

- Which book are more popular? Who are the most popular authors?
- Why does a book name appear multiple times in the Goodreads website? What does it signify?
- How are books distributed across languages in Goodreads website?
-
- If and how the ratings given by users on Goodreads is related to the books performance
- Is there an unseen relation between number of pages of a particular book and how well it is received?
- Do books with more reviews have better Ratings?

3.Data

Data wrangling is creating, filling, maintaining, and governing a curated data lake. It involves seven steps, namely Data procurement, Vetting Data for Licensing and Legal use, obtaining data, describing data, grooming data, Provisioning data and Preserving data **(Terrizzano et al., 2015)**

The Data set used to draw my findings is put together and Legal available to use in a csv files with Schematic and semantic metadata. The data set used for this particular report is obtained from Kaggle

and named Goodreads-books. It was originally scrapped directly from Goodreads API. in 2019. I choose this dataset because it has a well curated attributes of a variety of books listed on the Goodreads website as well as Goodreads user generated data.

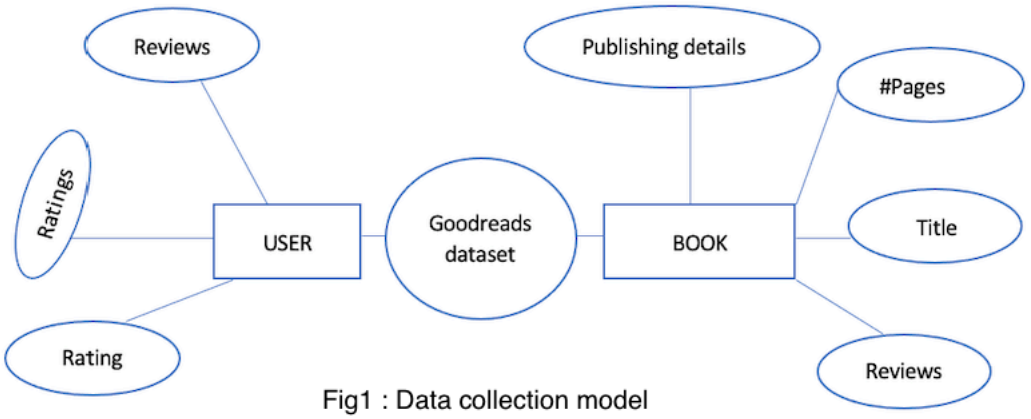


Fig1 : Data collection model

The books dataset contains 12 columns data columns 11123 rows. The classifying columns were namely: bookID , Title, authors, average_rating, isbn, isbn13, language code, ratings_count, num_pages and text review counts. This database was chosen because the size of the database and the different variable it describes, gives us more precise results.

Dataset information using python:

```
[232]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11121 entries, 0 to 11122
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype         
---  ------              --------------  -----         
 0   bookID              11121 non-null  int64         
 1   title               11121 non-null  object        
 2   authors             11121 non-null  object        
 3   average_rating      11121 non-null  float64       
 4   isbn                11121 non-null  object        
 5   isbn13              11121 non-null  int64         
 6   language_code       11121 non-null  object        
 7     num_pages         11121 non-null  int64         
 8   ratings_count       11121 non-null  int64         
 9   text_reviews_count  11121 non-null  int64         
 10  publication_date    11121 non-null  datetime64[ns]
 11  publisher           11121 non-null  object        
 12  Ratings_Dist        11121 non-null  object        
dtypes: datetime64[ns](1), float64(1), int64(5), object(6)
memory usage: 1.4+ MB
```

Columns Description:

- **bookID** is a unique ID for each book/series
- **title** gives us the name of the books
- **authors** contains name of the author/authors.
- **average_rating** the average rating of the books(these ratings are given by the reader)
- **ISBN** International Standard Book Number, unique book id across the globe
- **ISBN 13** 13 digits new format of ISBN
- **language_code** Contains which language the books are written in
- **Num_pages** counts of the number of pages per the book
- **Ratings_count** gives the number of ratings for the book given by the readers/users
- **text_reviews_count** Has the count of reviews left by users

Statistical description of the database:

	bookID	average_rating	isbn13	num_pages	ratings_count	text_reviews_count
count	11121.000000	11121.000000	1.112100e+04	11121.000000	1.112100e+04	11121.000000
mean	21307.774301	3.934058	9.759876e+12	336.343944	1.794512e+04	542.118874
std	13093.542472	0.350513	4.430156e+11	241.129968	1.125091e+05	2576.845134
min	1.000000	0.000000	8.987060e+09	0.000000	0.000000e+00	0.000000
25%	10270.000000	3.770000	9.780345e+12	192.000000	1.040000e+02	9.000000
50%	20264.000000	3.960000	9.780582e+12	299.000000	7.450000e+02	47.000000
75%	32104.000000	4.140000	9.780872e+12	416.000000	4.996000e+03	238.000000
max	45641.000000	5.000000	9.790008e+12	6576.000000	4.597666e+06	94265.000000

The dataset has been analysed using Python in Jupyter notebook. Data was cleaned and had no null values and I dropped all duplicated data before beginning my analysis. Extracted month and Year information from publication_date using python datetime module.

4.Exploratory Data Analysis (EDA)

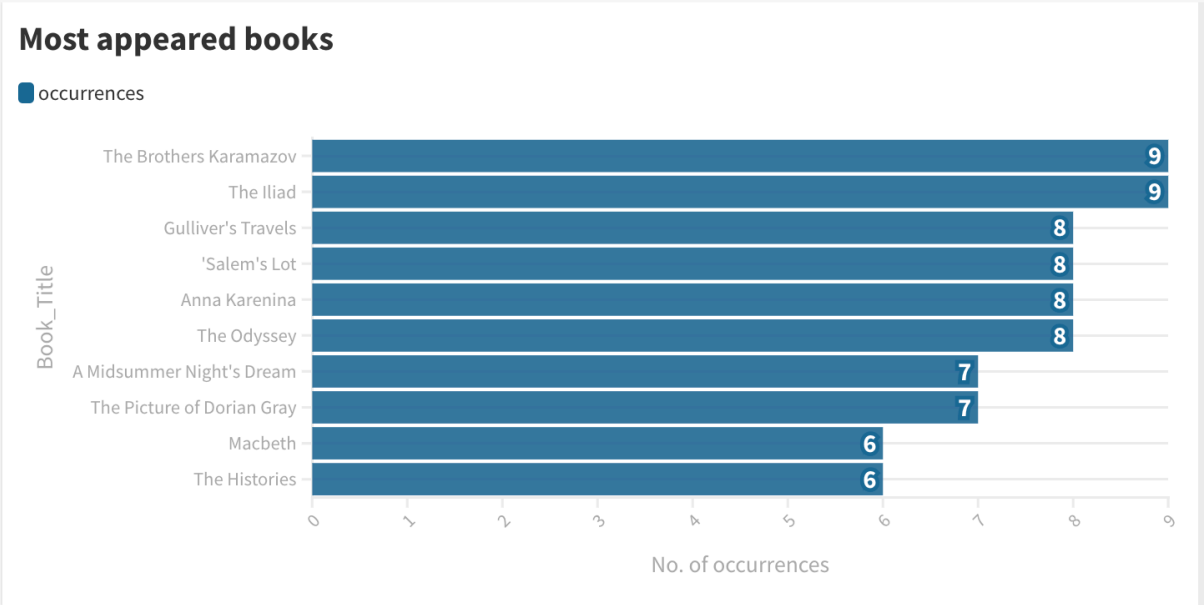
For this report I am going to preform exploratory data analysis to investigate the data and discover patterns and check assumptions of the Goodreads Books data by using statistics and visualization.

4.1 Book title versus number of occurrences

Why does a book name appear multiple times in the Goodreads website? What does it signify?

I have plotted a graph with top 10 book occurrences within the Goodreads dataset to. These occurrences are the number of times a particular book has been published. i.e., it gives us the number of editions of the book present in the Goodreads Catalogue. It is clearly shown by the bar graph that ‘The Brothers Karamazov and ‘the Iliad ‘have been published 9 different time each i.e., they are 9 editions of each of these books in the data set. Another observation is that if we were to classify the

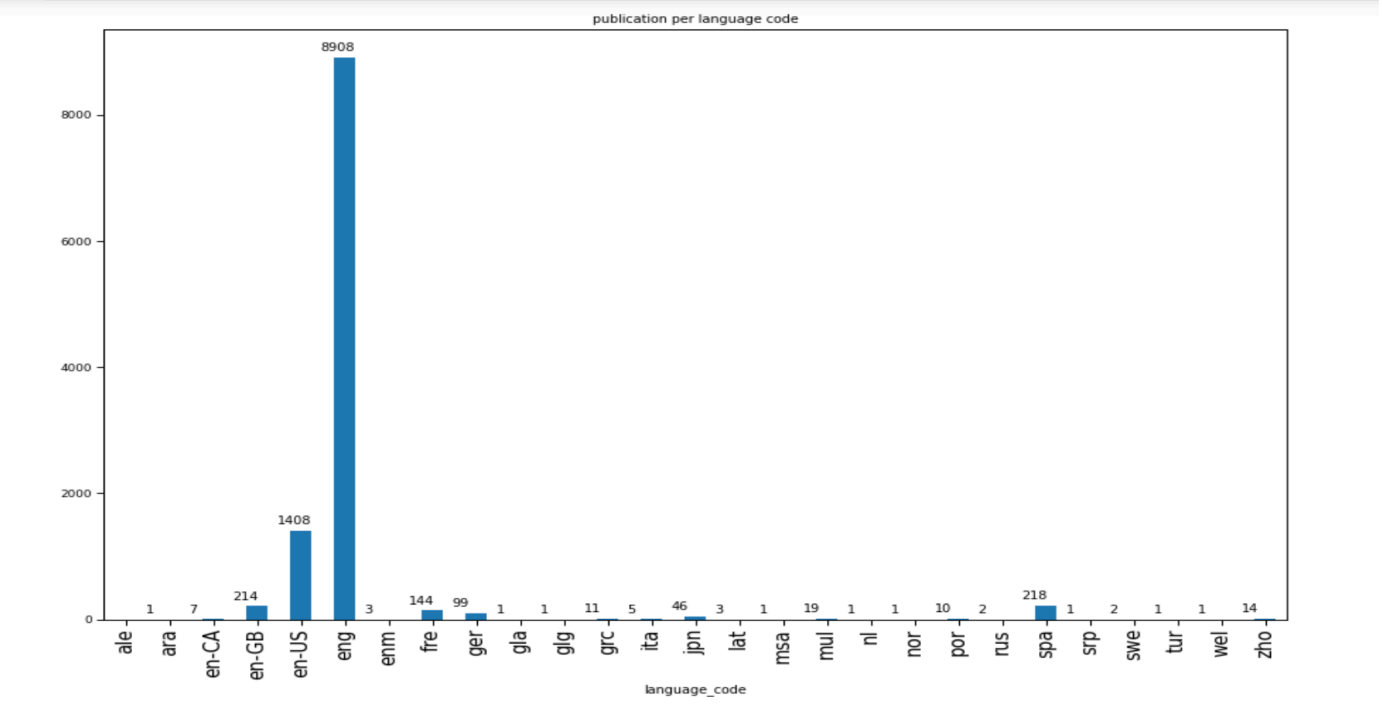
below books, most of them would be categorized as classics. Hence, we can also say there is a probability a classic book will be available in more than one edition in Goodreads website.



4.2 Books and language

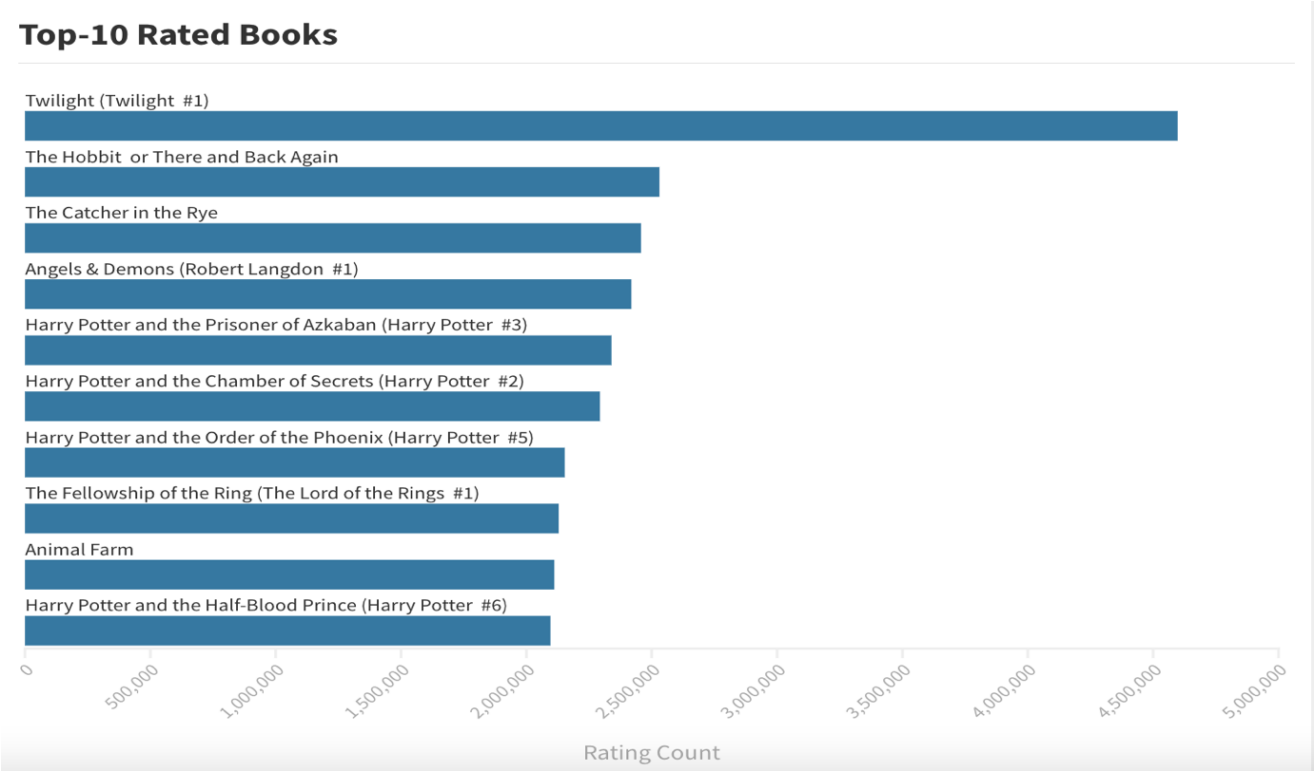
How are books distributed across languages in Goodreads website?

The bar graph below is plotted to visualize number of publications against each language code present in the dataset of Goodreads. The most books published are assigned language code ‘eng’. If we combine all the language codes which represent English, it is observed 10,537 of the 11,127 books are published in English. After English, maximum book in this dataset are in Spanish i.e., 218.



4.3 Analysis of top 10 rated books

The chart below is plotted between the highest numerical value of rating count column and its consecutive book title. The chart below shows the top 10 book titles in the data set which have the highest rating counts, i.e., highest number of ratings a book has received. So, it can be observed that the book name 'Twilight' has the highest ratings count, i.e., 4597666. It can be inferred that 4597666 users of Goodreads have rate the book 'Twilight'. It is also due to be noted that most of the books in the top 10 rated books dataset are parts of a series except for 'The Hobbit or There and Back Again', 'The catcher in the Rey' and 'Animal Farm'. So, form this observation I can draw that people usually prefer reading and reviewing a series more than a one - off book. However, it is also to be noted that count of rating for the harry potter series in the top 10 list fell from the series book # 3 to book # 6, indicating not everyone who read and rated Harry Potter book #3 has rated Harry Potter book#6

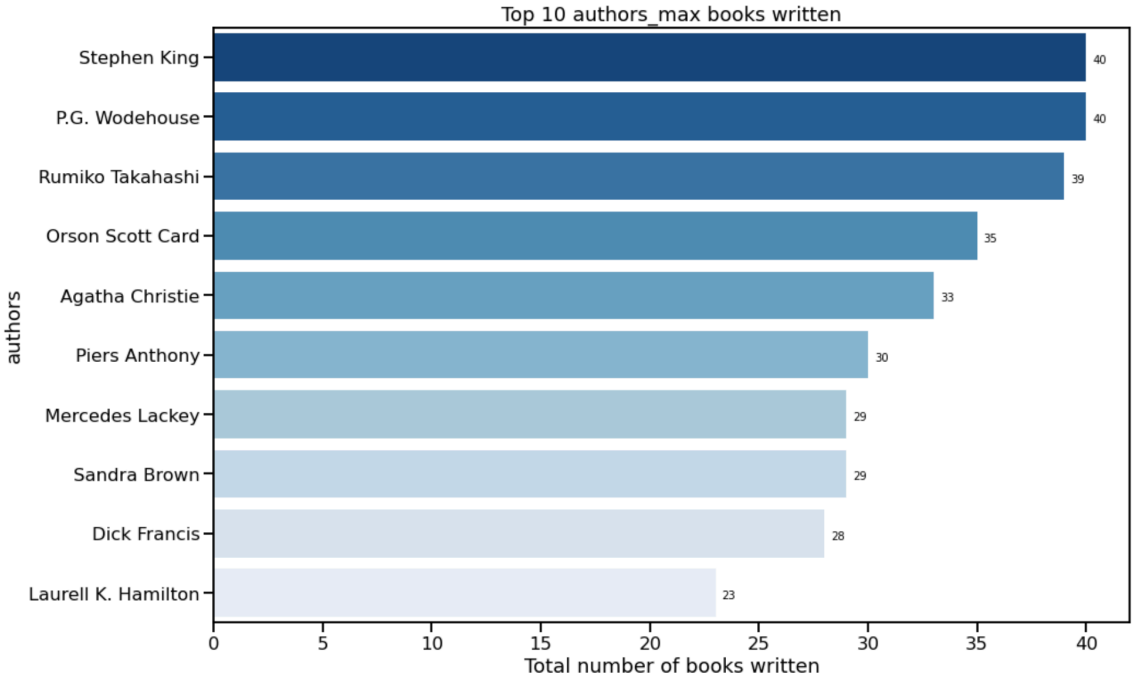


The sun burst visualizes top 10 rated books and their respective authors and publishing house. It is evident that scholastic Inc. publishing house has published almost $2/5^{th}$ of the top 10 rated books, with Houghton Mifflin publishing $1/5^{th}$ of the top 10 rated books in this data set.

4.4 Authors versus number of publications

Authors with the greatest number of published books

The bar plot below is plotted between authors and total number of books which are present in the dataset that are written by them. The plot show that Stephen King and PG. Wodehouse have the greatest number of datapoints in book titles column in the data set, i.e., 40 each.



But give that the older books have been known to be published in different editions sometimes, authors Stephen King and PG. Wodehouse probably have few of their books published in multiple editions.

```
In [67]: sample2 = df[df['authors']=='P.G. Wodehouse']
         sample2['title'].nunique()
Out[67]: 39
```

```
In [60]: sample1 = df[df['authors']=='Stephen King']
         sample1['title'].nunique()
Out[60]: 30
```

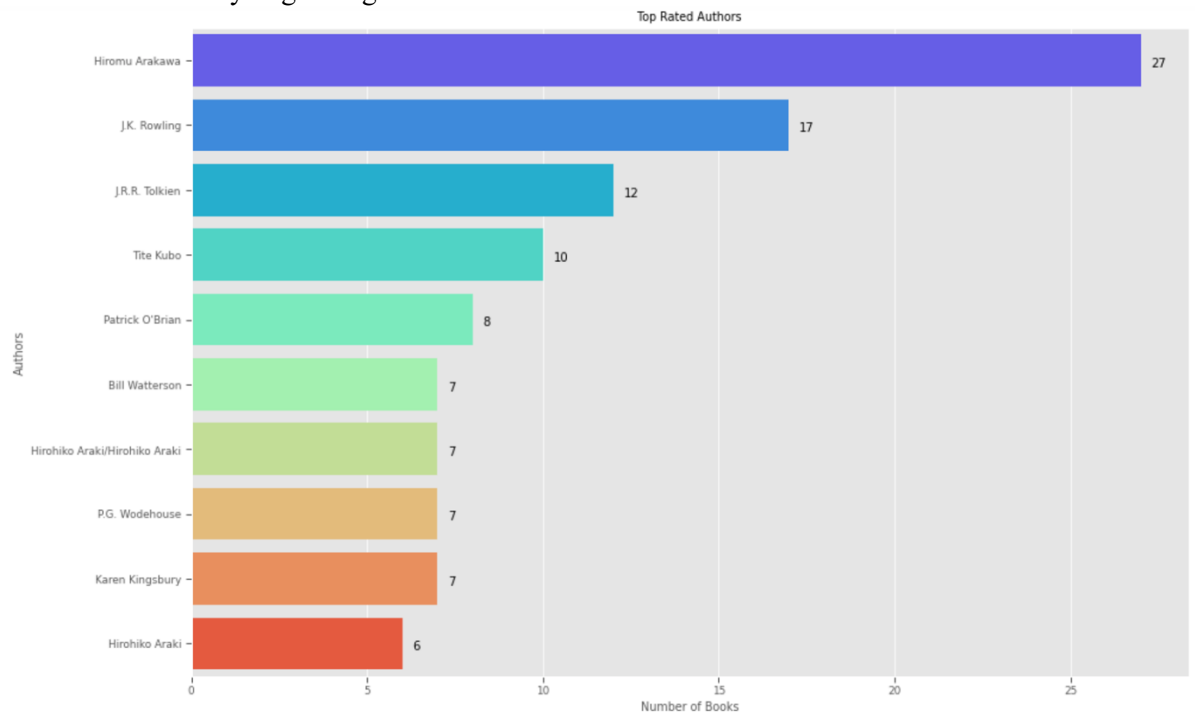
So, filtering the titles of authors using nunique () function in Python, it returns that Stephen King has only written 30 books and the rest 0 books are different editions of the books he previously wrote. Using the same nunique () function P.G. Wodehouse has 39 unique book title and only one edition published. Hence seen that P.G. Wodehouse had written more books than Stephen King.

So, It can be assumed that the authors in 'the top 10 authors with max books' list would have be writing for a long time, or wrote classics hence have their books republished or are just active writers

4.5 Authors versus Rating

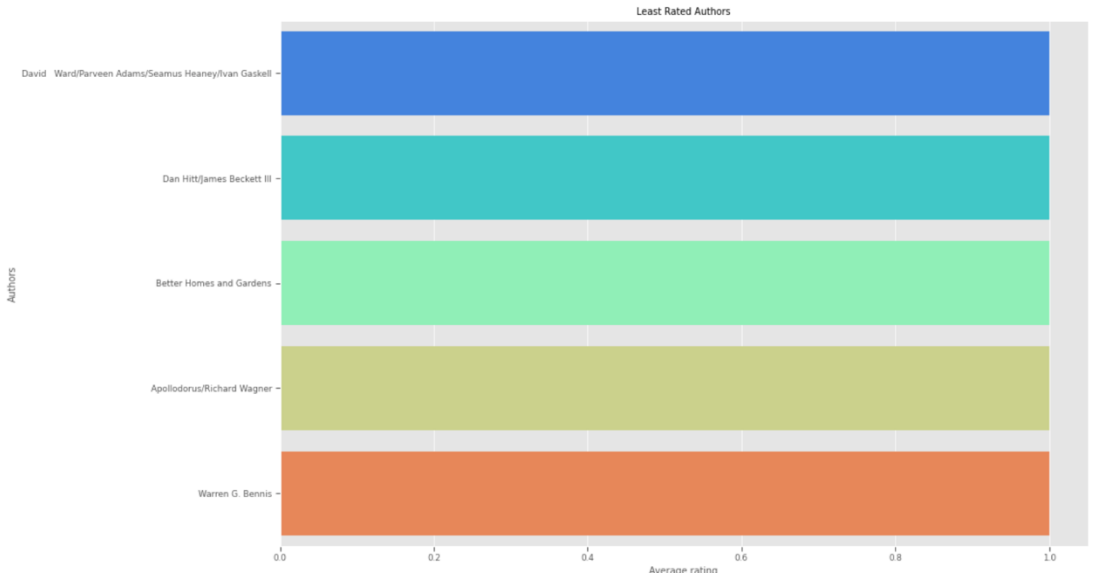
Who are the top 10 highly rated authors?

The graph is plotted between authors and their titles of their books for all value of average rating greater than and equal to 4.3 rating. So the graph show how author Hiromu Arakawa has highest number of highly rated book in Goodreads catalogue , with J K Rowling in second place but author Hiromu Arakawa is ahead by huge margin of 10 books.



Who wrote the least rated books on Goodreads?

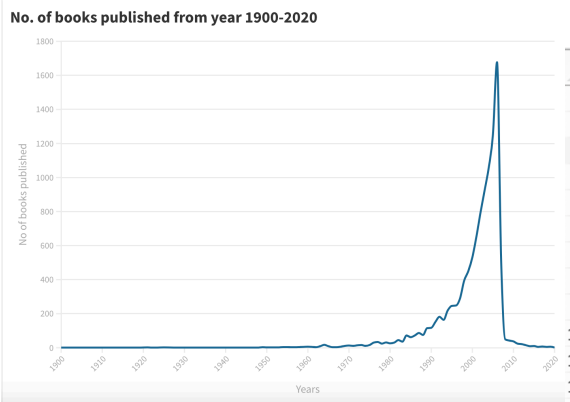
The graph below is a bar plot between authors and their titles of their books for all value of average rating less than and equal to 2 rating. Like the graph suggests with better homes and gardens, these probably are books in lesser popular genre.



4.6 Trends of Books published with respect to time

Which year were maximum number of books published?

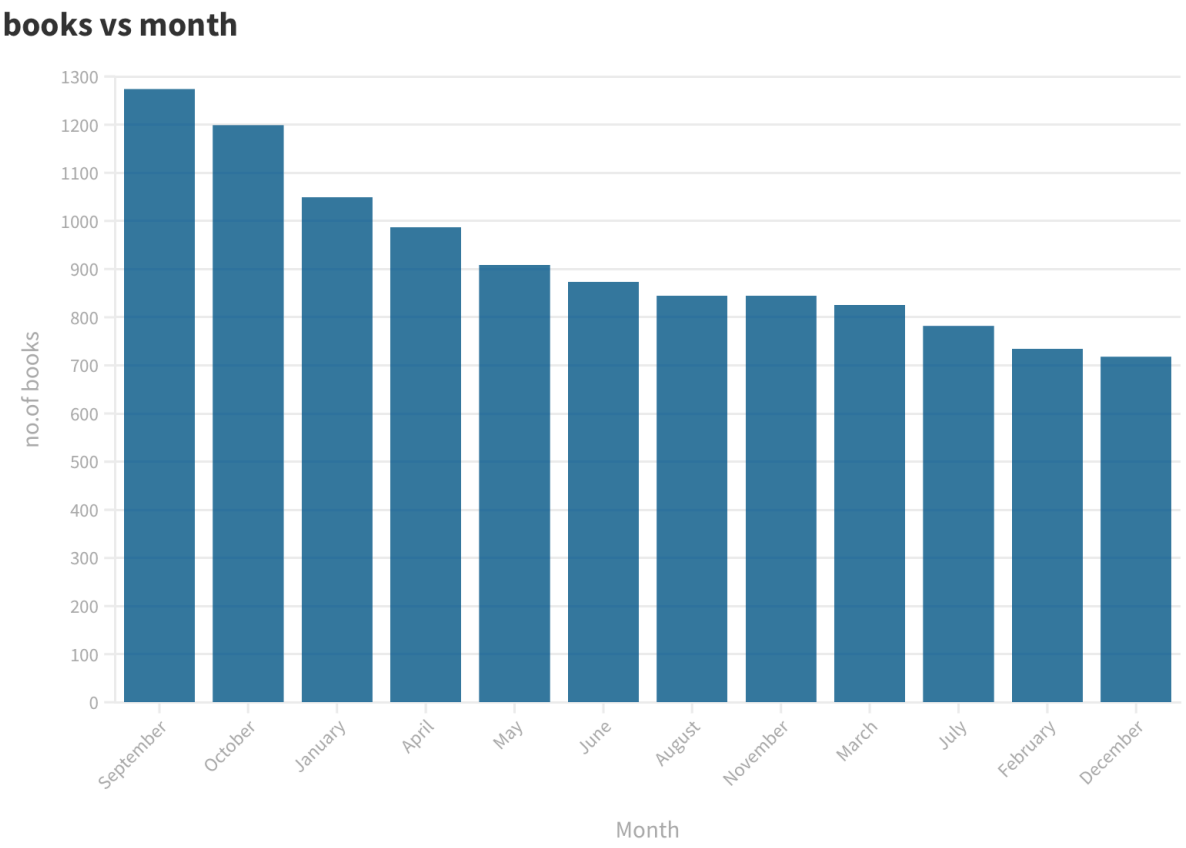
I have grouped the all the books in the dataset against the year in which they were published in. Then I have taken the result and I used a line chart to plot it against years. It clearly shows the years 2000 to 2010 saw the sharp raise and fall of books published. The table shows the top 10 years with maximum number of book publication.
2006 has so far been the year with maximum number of books published i.e,1677.



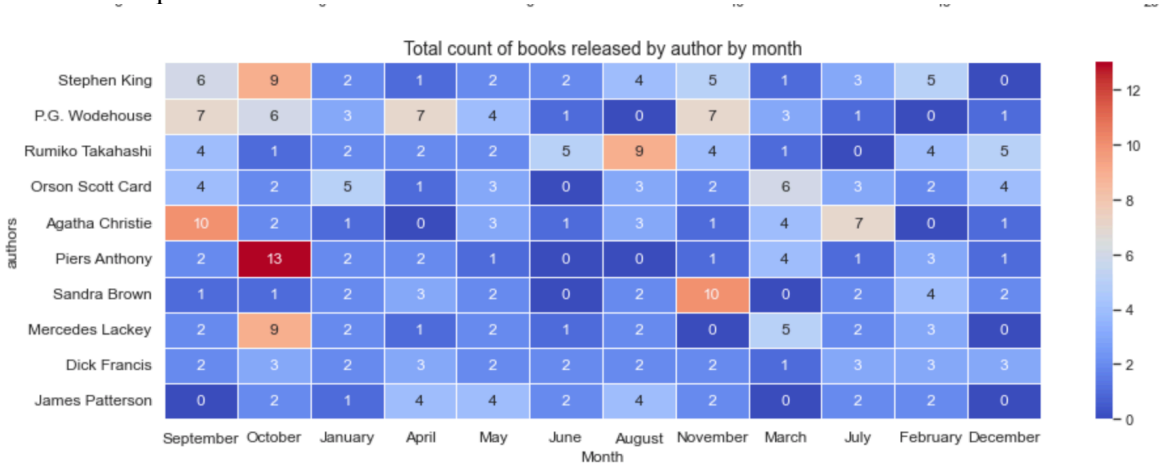
	A	B
1	Year	books released
2	2006	1677
3	2005	1251
4	2004	1060
5	2003	927
6	2002	797
7	2001	655
8	2000	531
9	2007	513
10	1999	449
11	1998	394
12	1997	290

Which month is a new book likely to come out?

The below graph categorizes all the published books in the data set with respect to months.
We can see that September is the most popular month to be publishing a book. The least popular months seems to be December which is closely followed by February

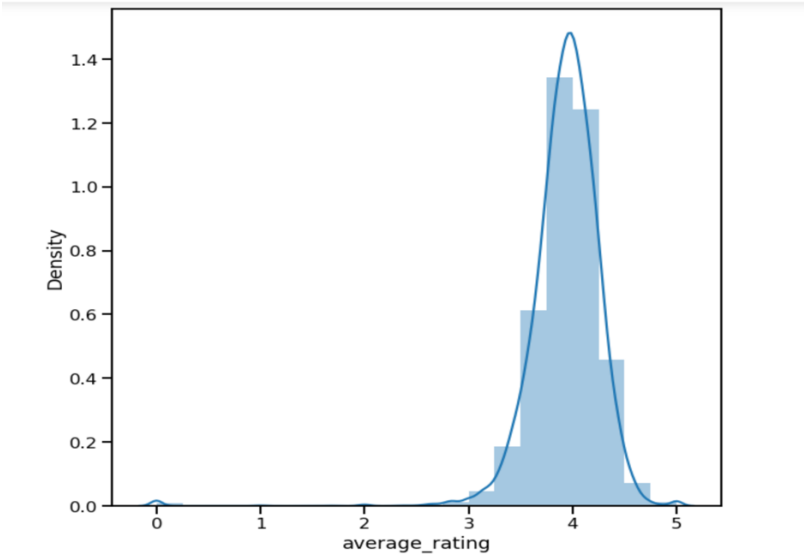


The heatmap which is plotted against authors with maximum no of published books against months they published corelates to the above graph. We can see most of the authors books being published in the months of September

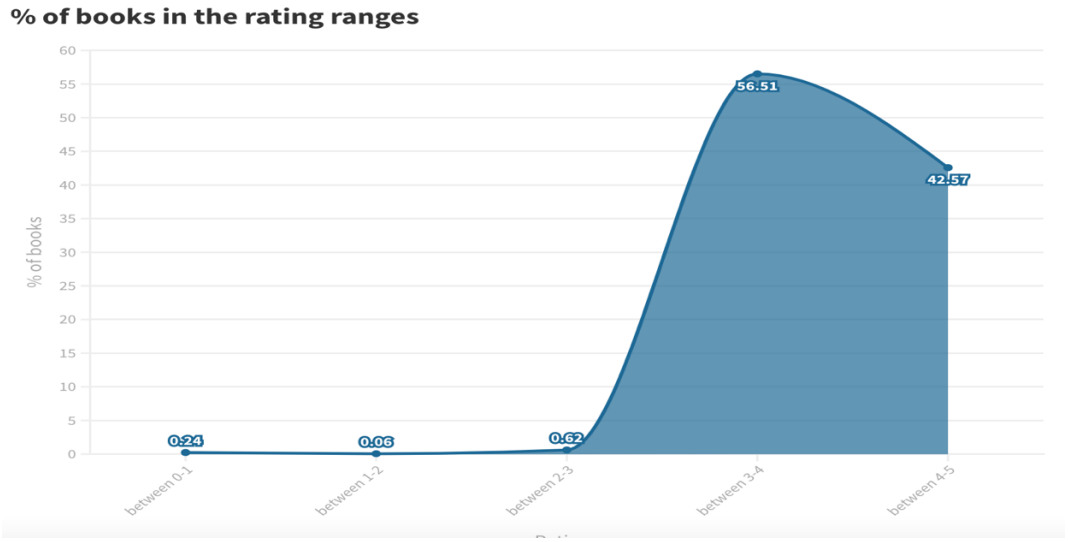


4.7 Rating trends

What is the rating distribution for the books?
To find the average rating of most books we are deploying a density plot across average ratings column. From the graph below we can derive that most of the book in the data set are rated between 3 .6 and 4.3(high) and very few books are rated between 1and 2(low) and even fewer books were rated 5.

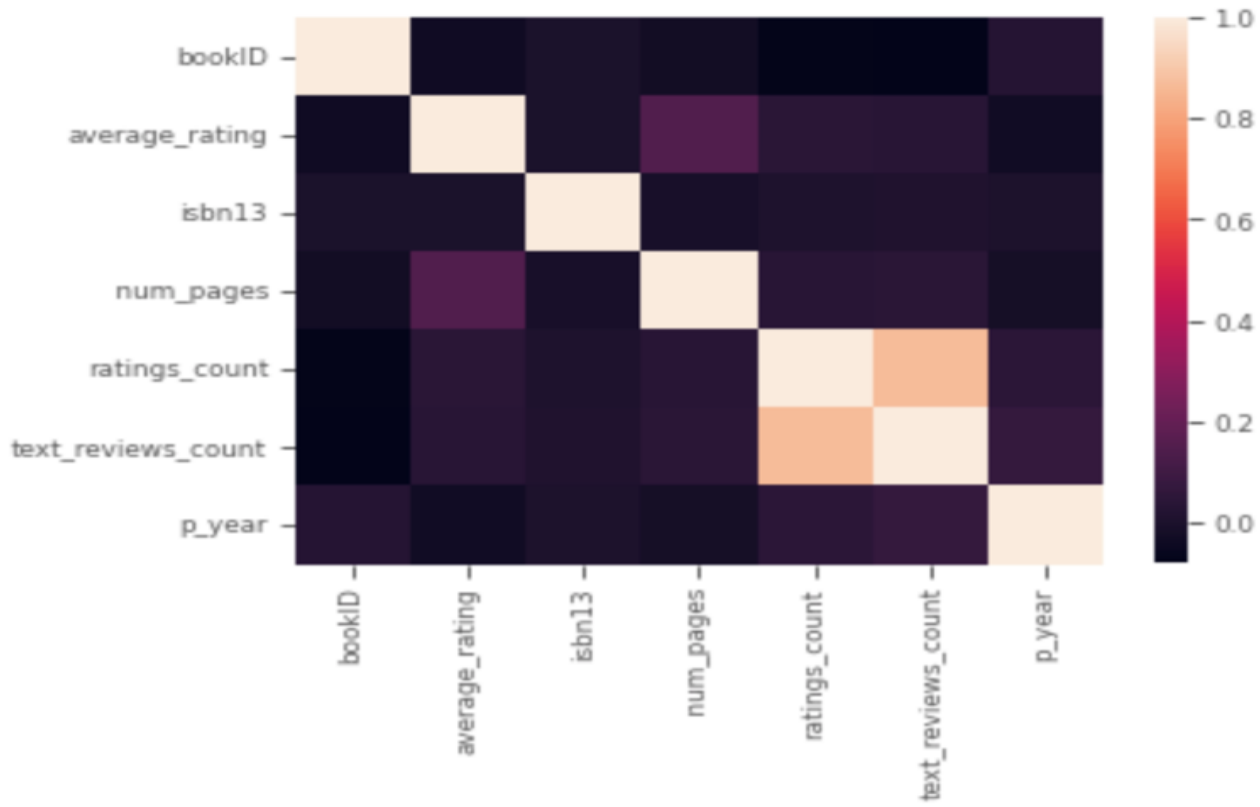


As the graphs show we categorized books rated between 0-1 ratings,1-2 ratings,2-3 ratings,3-4 ratings and 4-5 ratings. And they are 0.24% of total books in the range rating 0-1, 0.06% of books are rated between 1-2, 0.62% of books rated between 2-3 range and 56.51% of the books have the rating 3 to 4 and 42.57 % of the books fall under 4 to 5 rating categories. The numerical rating of these ranges is higher compared to the percentage. It can be observed that a user leaves an at least 2- star rating for even their least favored book. Maximum number of ratings lie between 3-4 lead us to believe a majority of the readers rate the book they enjoyed the most between 3-4 stars.



4.8 Rating counts versus Text review counts.

I have generated a heat map using matplotlib in python to see correlation between the numerical values of the dataset. From the plot below we can see that only 'ratings count' column and text review count are highly corelated positively. Hence books with higher rating count will have higher text review counts and books with few numbers of ratings will have low number of text count.

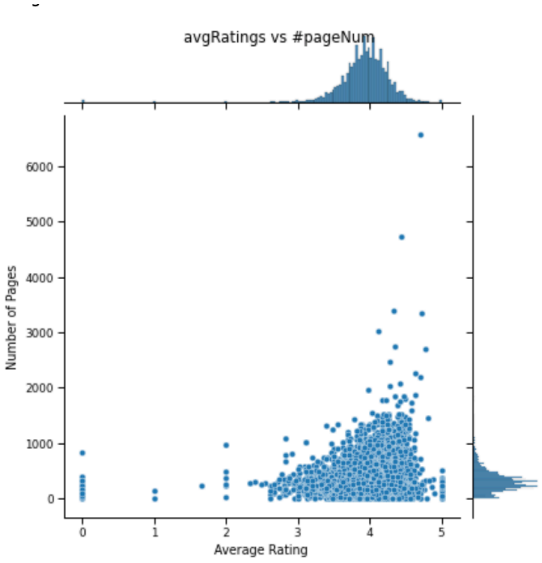


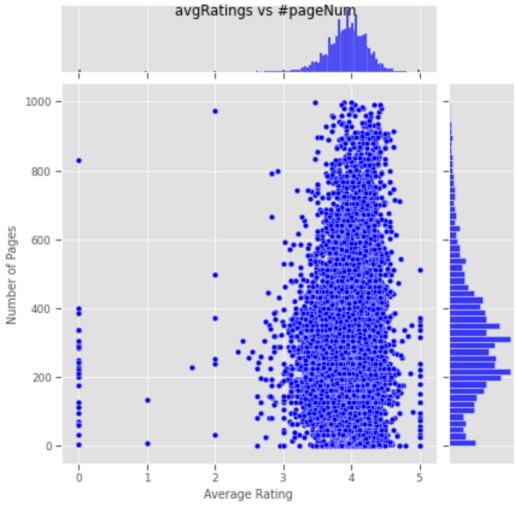
4.9 Relationship between Number of pages and Ratings?

From above heat map we can see a noticeable correlation between average ratings and number of pages of a book. Let's explore more with what is the relation between average ratings and number of pages.

The plot below between average rating and number of pages, has a lot of outliers for number of pages greater than 1000 but the density of the points lie below 1000 signifying that most books are under 1000 pages .A second more closer scaled graph is taken and we get to observe that maximum number of rating are given for books having pages under 400 and books with pages between 200 and 400 have more ratings and it can clearly be observed that ratings decrease as the volume of the book increases.

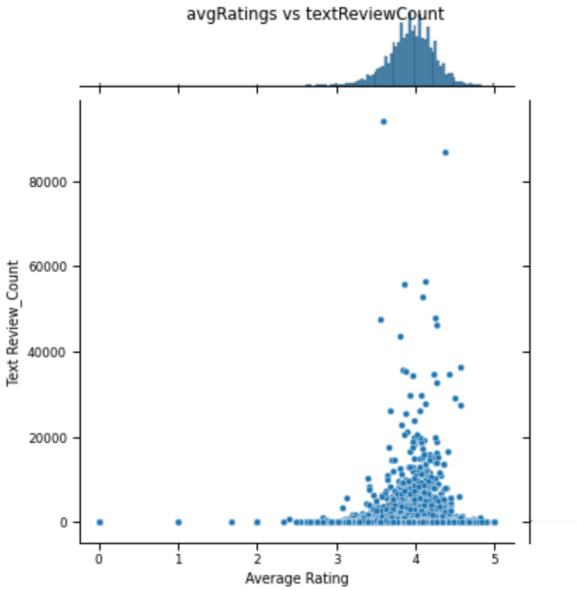
We can derive on this observation to believe readers of Goodreads website prefer smaller volume books. Or for a book to have higher probability to get more rating, it should have fewer pages.

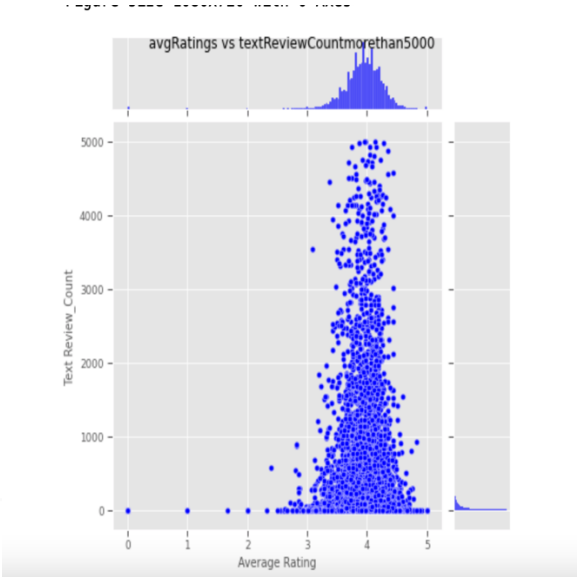




4.10 Relation between Average Ratings and Text Review Count

We have used a joint plot function to plot a scatterplot of points correlation of average rating and text review count columns in the dataset.
The plot show that maximum number of ratings for the books in the dataset lie in the 3-4 range with points leaning more towards 4. The is a cluster of dense points plotted against average rating 4 and under text review count 20000. This signifies most books received around 4 rating and less than 20,000 text reviews.

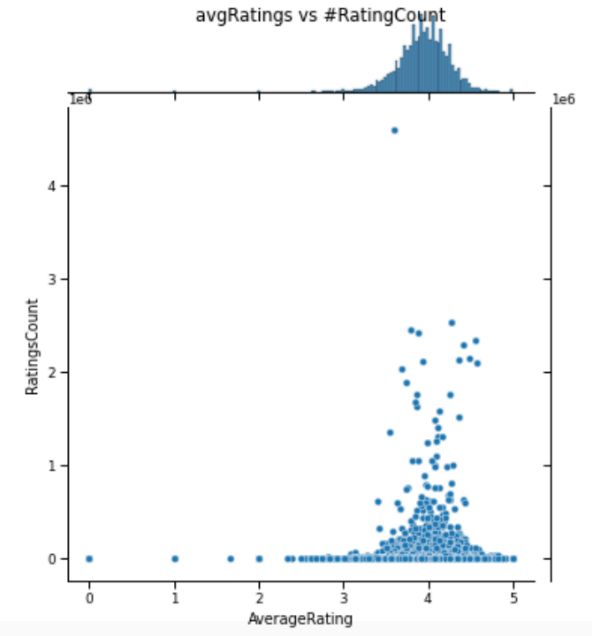


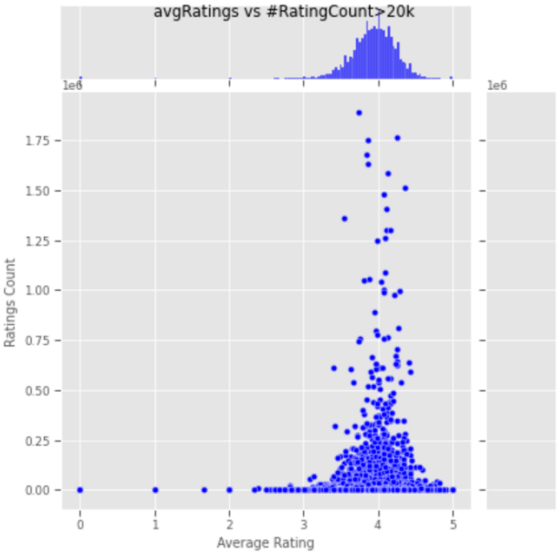


From the second plot is plotted for avg ratings and text review counts under 5000 gives a better look at the clustered points. We can observe that the points are more populated under 1000, which is counter intuitive. We can only conclusively derive that book with reviews have good ratings.

4.11 Is there a relationship between ratings and ratings count?

We are plotting a graph between the average ratings and rating count. From the first graph we can see that the most counts of rating instance is in the 3-4 average rating range, which seem likely give that maximum reader are reading the books between 3 to 4 rated. But the plot also shows a noticeable outliers. To explore this I have plotted another graph with rating count >20,00000 . There is a slight corelation between the average rating and rating count to the extent that as the number of ratings increase for books of better are average rating i.e. books with average rating of 3-4 received some of the highest number of ratings.

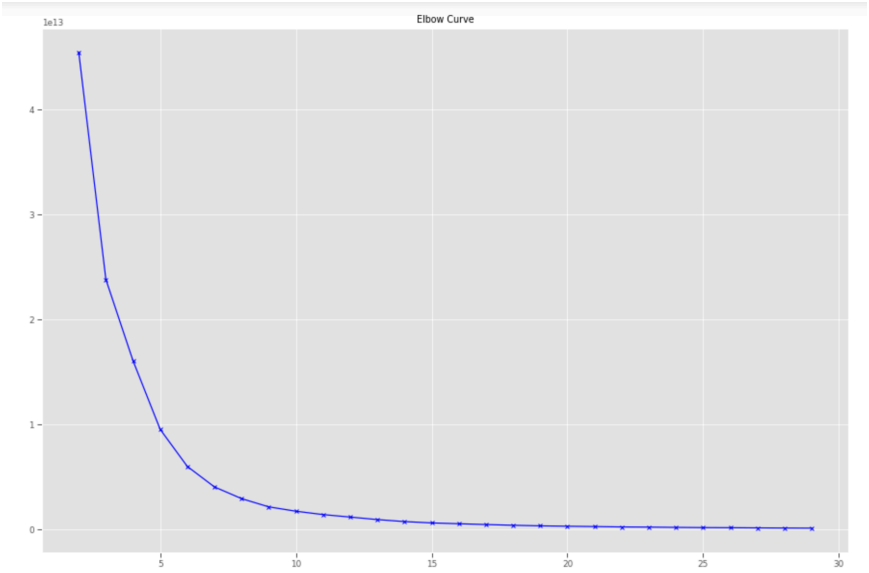




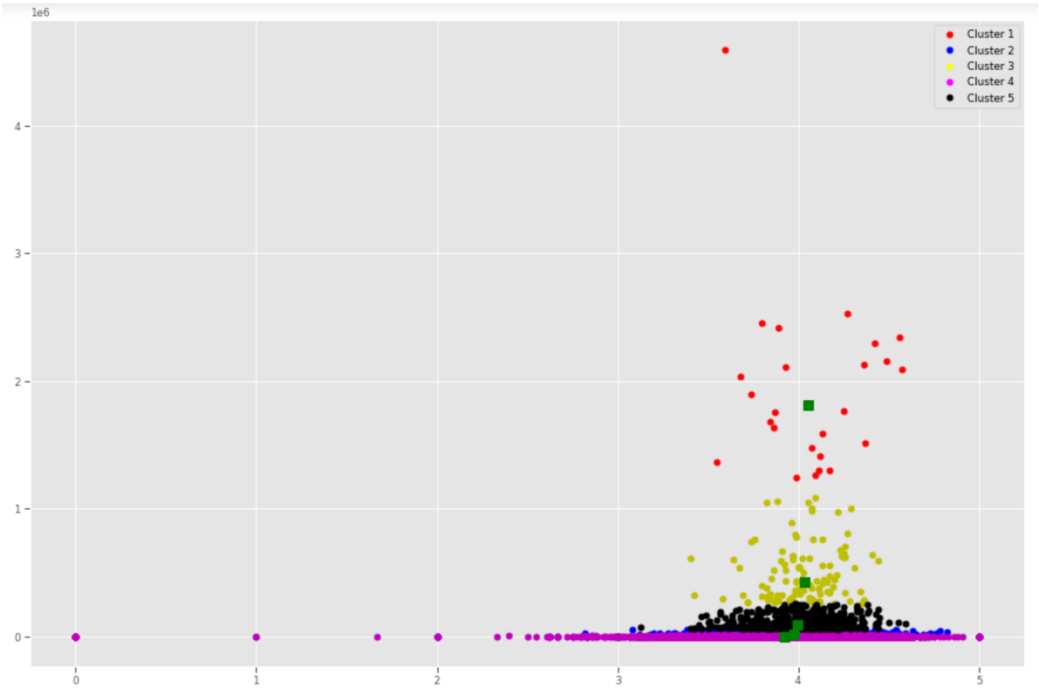
5. Modelling: Using K- mean's clustering

K- means clustering is used to split and group observations into k clusters based on patterns with each observation belong to nearest mean's cluster. It used to find groups among unlabelled datapoint. (Analytics Vidhya, 2019)

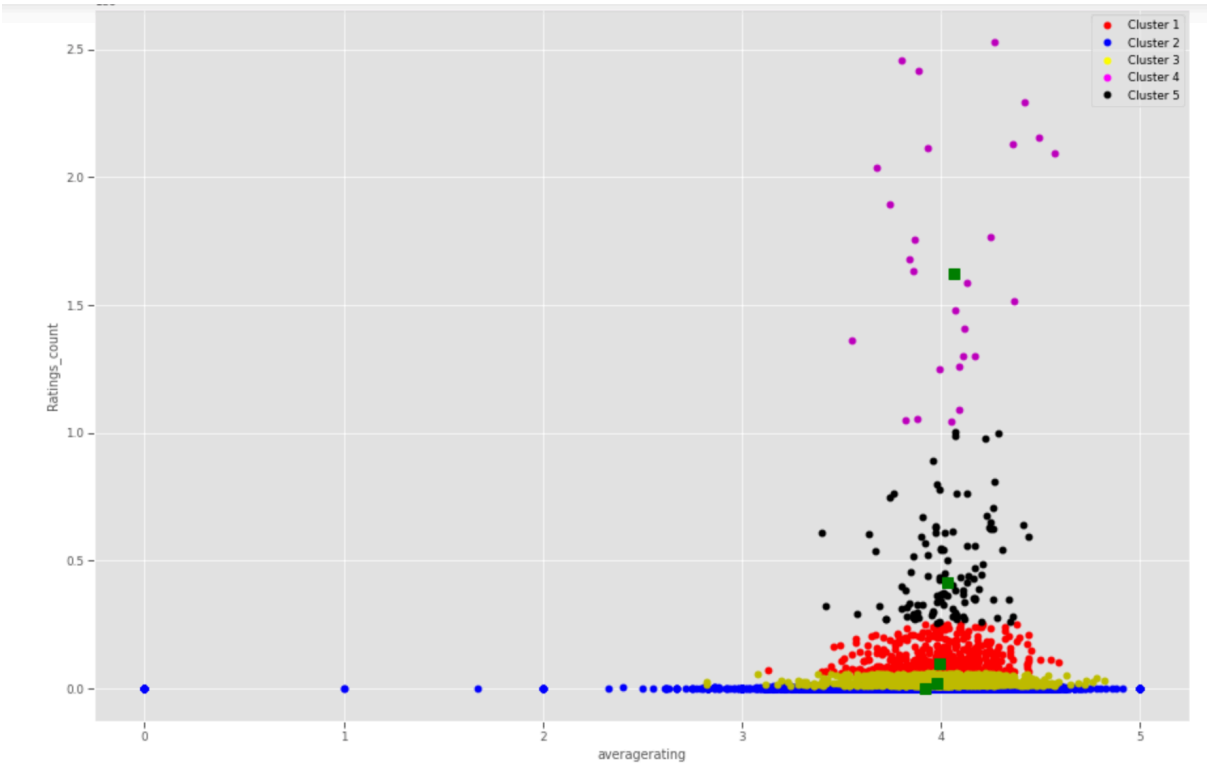
Grouping observation between rating count and average rating value can help establish an relationship between the two. In a dataset with certain number of observations, we can plot a graph, called elbow curve, whose x-axis represents the number of clusters possible.



The above elbow graph show they should be 5 cluster, since the 'elbow' of the graph is 5.



The outliers in the graph are skewing all the clustering's in the graph above. So, I have eliminate the outliers by setting and upper limit to the ratings count value.



The plot shows that as the count increases, ratings will end up in the cluster above and as rating count decreases, the average rate tends to become sparser.

6. Data Ethics:

Data ethics are very important quality a person handling any kind of data must have. Data ethics refers to the moral responsibilities attached to gathering, protecting, and using personally Identifiable information and how it affects the individuals (Cote, 2021).

To put this report together , I have been handling data scrapped from a website called Goodreads, which is made available to the public by Kaggle. All though I only handled information relating to books and didn't handle any individuals personal information, the data and analysis of the data used for this report, was handled with adherence to The UK Government Data Ethics Framework. The three main principles of this framework are Accountability, Fairness and Transparency

Accountability:

- I have made minor change to the dataset while exploring it to draw insights, and all of them were ethical. I have not made any permanent changes to the dataset obtained from the Kaggle. I have fully complied to data ethics while exploring the data.

Fairness:

- The dataset I have use to conduct my analysis is Licensed under **CC0: Public Domain**, which allowed me access to the dataset with no copy write restriction and is available to the public to copy, modify, and share it.

- The data is scraped by the original owner from the Goodreads API, which is open to the public

- Confidential Information of users was never shared or even handled by me for this project. Even the name of the users who gave reviews was not shared, every personally identifiable information has been already anonymised.

- I have not share the information about this data to anyone else.

Transparency:

- The data was handled with integrity and was never manipulated with intention of generating a favourable outcome.
- I didn't need to fill in any Null values, however I did drop duplicate values from the original dataset to keep the integrity of my outcomes.

- I have enough evidence to prove that my data manipulation did not skew the outcome of my results in any way or break the integrity of the original data. I have only extracted data from an already existing column for the easy of writing a code. For example I have extracted year from publishing date and put it into a separate column for ease of reading data which never compromised the integrity of the dataset.

7. Conclusion:

After analysing and visualizing the data of the Goodreads website, I have drawn upon the conclusion that 'The Brothers Karamazov' and 'The Iliad' have been re-published as different edition 9 other times. Hence, they are the most occurring book titles in the database.

After visualizing the data, I can state that 95% of the books available on Goodreads catalogue are written in English. Even though books from The Harry Potter series make up 40% of the 'top 10 rated' books category, the book titled 'Twilight' is the top-rated book with 45000000 ratings

Maximum number of books written are by the author P.G. Wodehouse, and the top-rated author is Hiromu Arakawa. And

In the year 2006, 1677 books were published, i.e., maximum number of books published in a year across all years and throughout the years month of September sees
maximum publications and December sees the least.

For all the books available on Good reasd56.51% of books available on Goodreads have a 3-4 rating.

I couldn't conclusively draw the relation between average ratings of a book and number of text reviews it had. I can only derive that book with good reviews have good ratings. I could, however, conclusively

say that readers of Goodreads website prefer smaller volume books. So, for a book to have higher probability to get more rating, it should have fewer pages

There is a only a slight corelation between the average rating and rating count to the extent that as the number of ratings increase for books of better are average rating. And we also can draw that though reviews matter there is no casual effect on Ratings.

I used the K-means clustering to deduce that as the count of ratings increase average ratings more together. I.e. fall in the ranges of rating ,giving a more definitive rating range to the users. So the more the rating and review count the more accurate the average ratings are for a book

References

Analytics Vidhya. (2019, August 19). *The Most Comprehensive Guide to K-Means Clustering You'll Ever Need*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Cote, C. (2021, March 16). *5 Principles of Data Ethics for Business*. Business Insights - Blog. https://online.hbs.edu/blog/post/data-ethics

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Flourish | Data Visualization & Storytelling. (n.d.). Flourish. https://flourish.studio/

Soumik/. (2019). Goodreads- books. [Dataset] Retrieved from https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks/code.

Terrizzano, I., Schwarz, P., Roth, M., & Colino, J. (2015). *Data Wrangling: The Challenging Journey from the Wild to the Lake*.

Thelwall, M., & Kousha, K. (2016). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, *68*(4), 972–983. https://doi.org/10.1002/asi.23733

Wang, K., Liu, X., & Han, Y. (2019). Exploring Goodreads reviews for book impact assessment. *Journal of Informetrics*, *13*(3), 874–886. https://doi.org/10.1016/j.joi.2019.07.003

Wikipedia Contributors. (2019, April 29). *Goodreads*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Goodreads