Capstone Project 2

1. Problem statement

The goal of this project is to build a book recommendation system for users, based on the **Goodreads** data set, which can be found here. I additionally included via Goodreads API two columns, relative to the book description and the book popular shelves.

Goodreads is a social cataloging website that allows individuals to freely search its database of books, annotations, and reviews. Users can sign up and register books to generate library catalogs and reading lists. They can also create their own groups of book suggestions, surveys, polls, blogs, and discussions. The company is owned by the online retailer Amazon.

The problem is very relevant for Amazon or any other related company since it can potentially increase its profits as it gives recommendations of interest to their clients. Some of the key statistics about recommender systems are the following:

- At Netflix, 2/3 of the movies watched are recommended.
- At Google, news recommendations improved click-through rate (CTR) by 38%.
- For Amazon, 35% of sales come from recommendations.

The way I intend to solve the problem is to first build a collaborative-filtering system, and if I later obtain text data (book reviews) I can possibly incorporate that into the algorithm and therefore construct a hybrid approach: collaborative-filtering + content-based filtering.

Deliverables: code, a written report, and a slide deck.

2. Data inspection and Cleaning

2.1) Tags

book_tags.csv contains tags/shelves/genres assigned by users to books. Tags in this file are represented by their IDs. They are sorted by goodreads_book_id ascending and count descending. ***tags.csv***, on the other hand, translates tag IDs to names. I then naturally joined these two tables, creating ***tags_df.*** There were 6 negative values in the count column which I did not understand, so I discarded them.

2.2) Ratings

ratings.csv contains almost 6 million book ratings by users. Ratings can go from one to five stars, where:

- **5 stars**: "it was amazing"
- **4 stars**: "really liked it"
- **3 stars**: "liked it"
- **2 stars**: "it was ok"
- **1 star**: "did not like it"

There was everything ok with the data set.

2.3) To read

to_read.csv provides IDs of the books marked "to read" by each user. There was everything ok with the data set.

2.4) Books

books.csv has metadata for each book (goodreads IDs, authors, title, average rating, etc.).

goodreads IDs:

Each book may have many editions. goodreads_book_id and best_book_id generally point to the most popular edition of a given book, while goodreads work_id refers to the book in the abstract sense.

It's possible to use the goodreads book and work IDs to create URLs as follows:

- https://www.goodreads.com/book/show/17397466
- https://www.goodreads.com/work/editions/24219959

Note that book_id in *ratings.csv* and *to_read.csv* maps to work_id, not to goodreads_book_id, meaning that ratings for different editions are aggregated. 'books_count' is the number of editions for a given work. I changed this name, for clarity, to book_editions.

book_id, goodreads_book_id, best_book_id, authors, average_rating and **work_id** seem ok. There were some rows where the **ISBN** and **isbn13** number are missing, but for now we will leave it like that and return later if necessary. The same happened with the columns '**original publication year**', '**language code**' and '**original title**', but since there are not that many I will leave it like that for now and return later if necessary.

When comparing the title and author columns on these cases we see that there is nothing wrong with the data. They just wrote a book with the same title than other author did, and not twice.

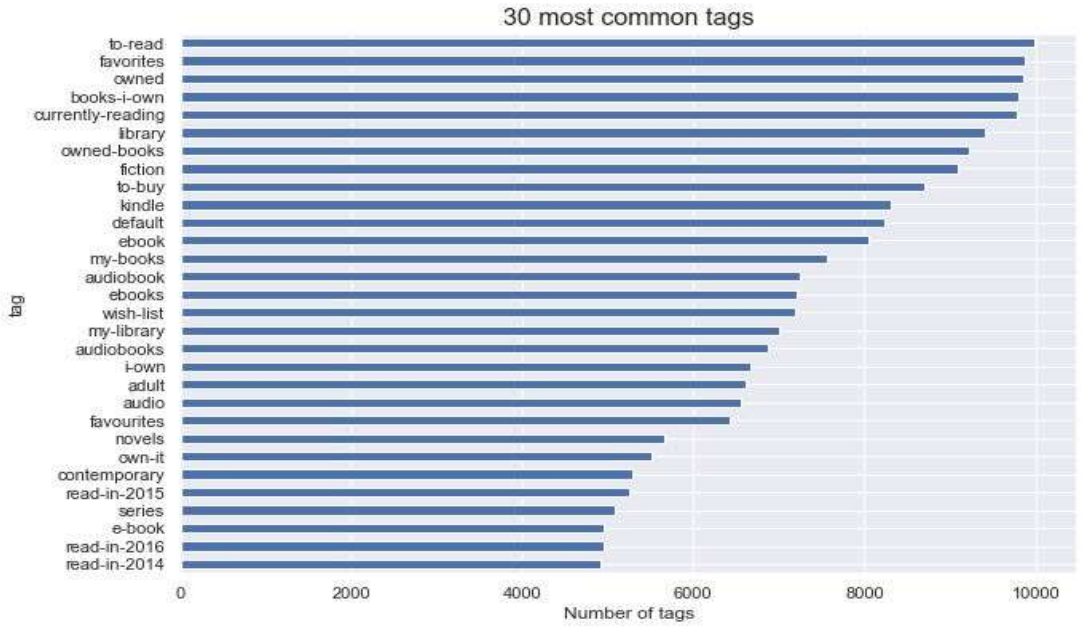
It is the column **work_ratings_count** that corresponds to the total number of ratings per row/book.

- The columns representing the **ratings** from 1 to 5 seem ok, as **work_text_reviews_count** do.
- Finally, the last two columns - **image_url** and **small_image_url** - represent the image of the cover of the book.

3. Exploratory Data Analysis

3.1. Tags

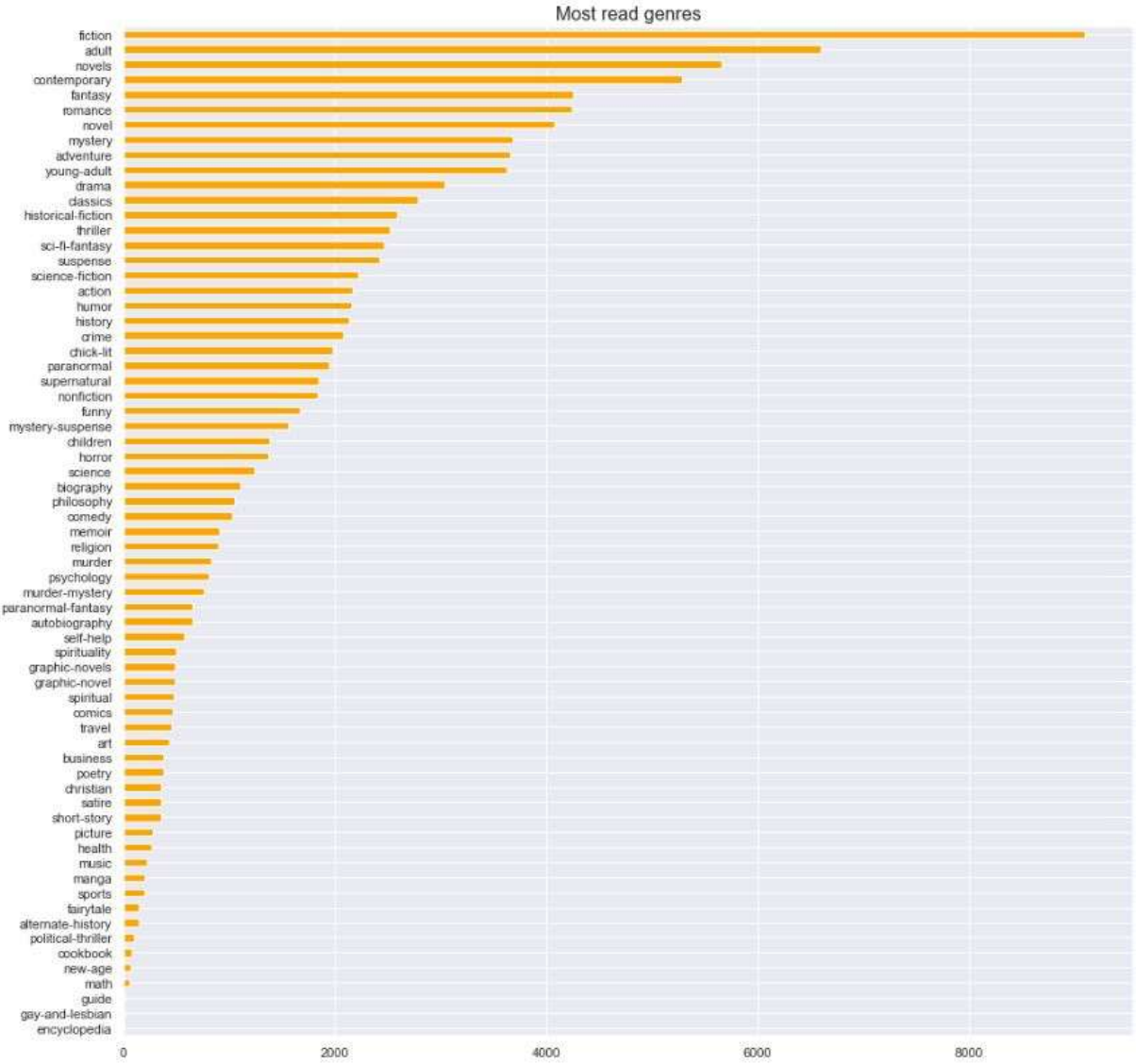
- What are the most common tags?



Although "to-read" appears at the top of the list, some tags that have the same meaning easily surpass it when aggregated. For example, "owned", "books-i-own", owned-books" and "my-books" means the same.

- What are the most common genres?

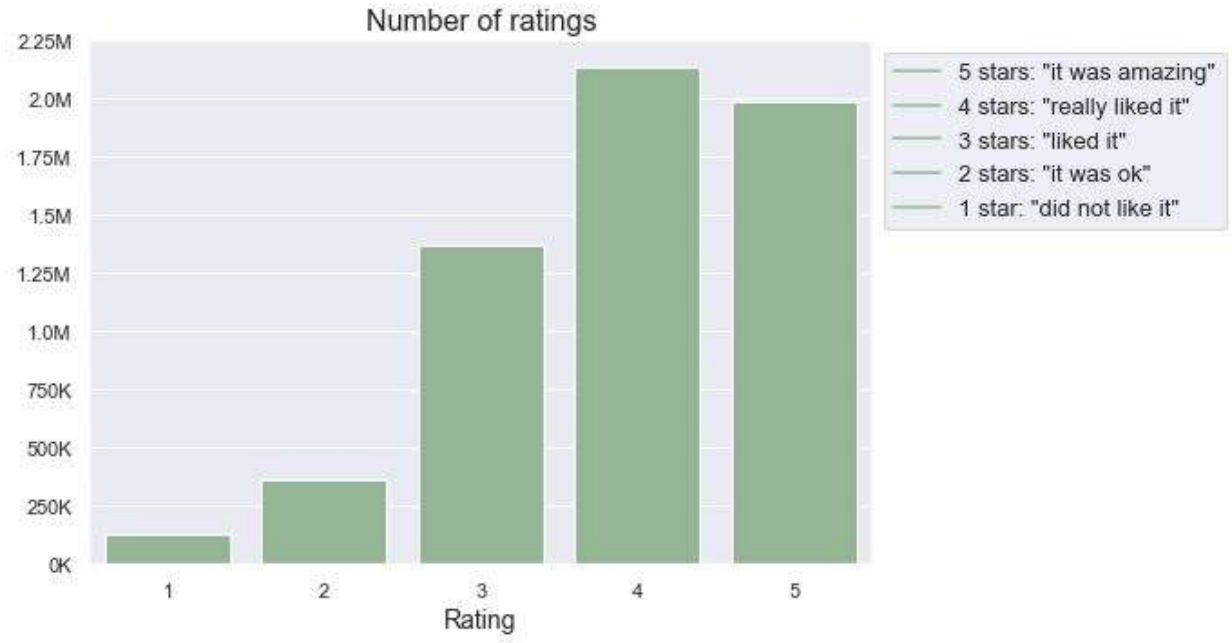
Let us take a look at the main genres. To do that I picked a list with genres from the web and added some relevant names. I will base the analysis on the previously built tag_df.



It seems that fiction books, mainly of the genre "adult", "novel"/"romance", "contemporary" and "fantasy" are amongst the most read genres by users.

3.2. Ratings

- How are ratings distributed? Do people tend to give higher or lower ratings?



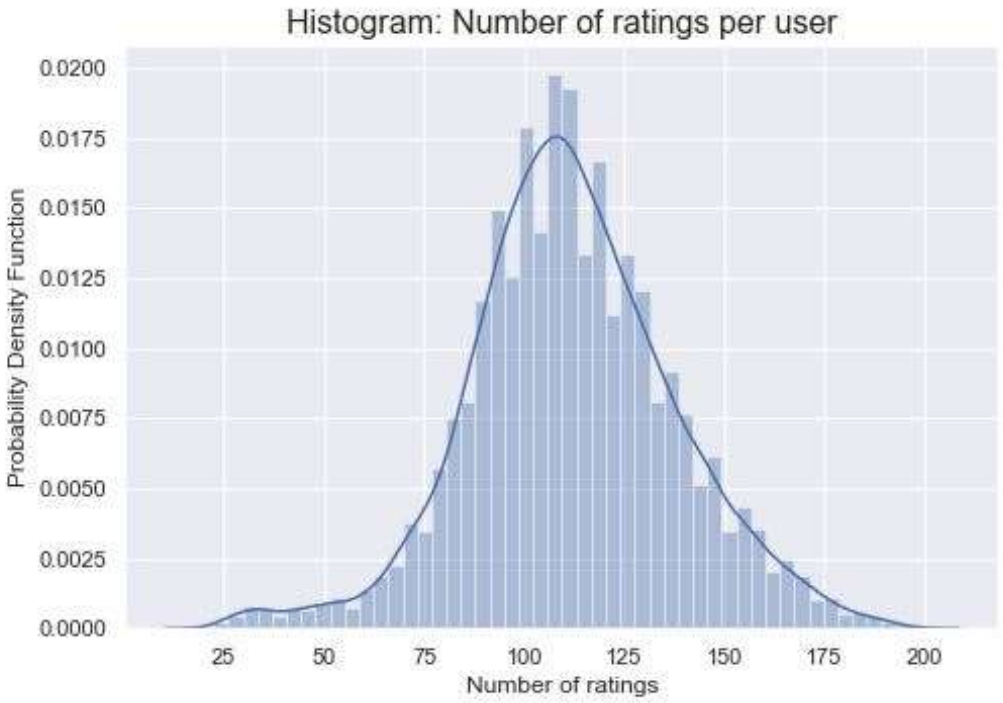
Regarding our users data set, we see that **almost 70%** of the ratings are classified as either **"it was amazing"** (5 stars) or **"really liked it"** (4 stars). We can say that in general users give good ratings. This is not surprising since this is a data set for only popular books and, among other reasons, the users are the ones who choose the books they are going to read and the topics they are going to pursue, according to their interests. If you add a **"liked it"** rating (3 stars), the percentage goes up to **91%.**

The average rating is in fact 3.9, nearly close to a "really liked it" evaluation, which is very close to the mean of all books (596873216) data frame as we've seen above (4.002191000000001). The median is 4.

- Does the distribution of the number of ratings given follow a normal distribution?

There are a total of 53424 different users in the data set. The Kernel Density Estimate which can be seen in the plot below does not suggest that, but we will check with a

normality test: D'Agostino's K-squared test. The null hypothesis assumes normality of the distribution.



D'Agostino's K-squared test: statistic=460.696; p-value=0.000

Since the p-value is practically zero, we have the statistical evidence to **reject** the hypothesis that the number of ratings given follows a normal distribution. Looking at the distribution, we see that it is somewhat **left-skewed**.

- Can we have a 95% confidence interval for the mean of the ratings a user gives?

I'll assume that the data is representative of the population. The diversity in book genre is pretty widespread and users that rate the most popular books and have this many ratings given on average is good enough.

We can make for now some inferences about the population of Goodreads users using the data in 'ratings.csv' (6 million ratings) - it may be useful when making recommendations later.

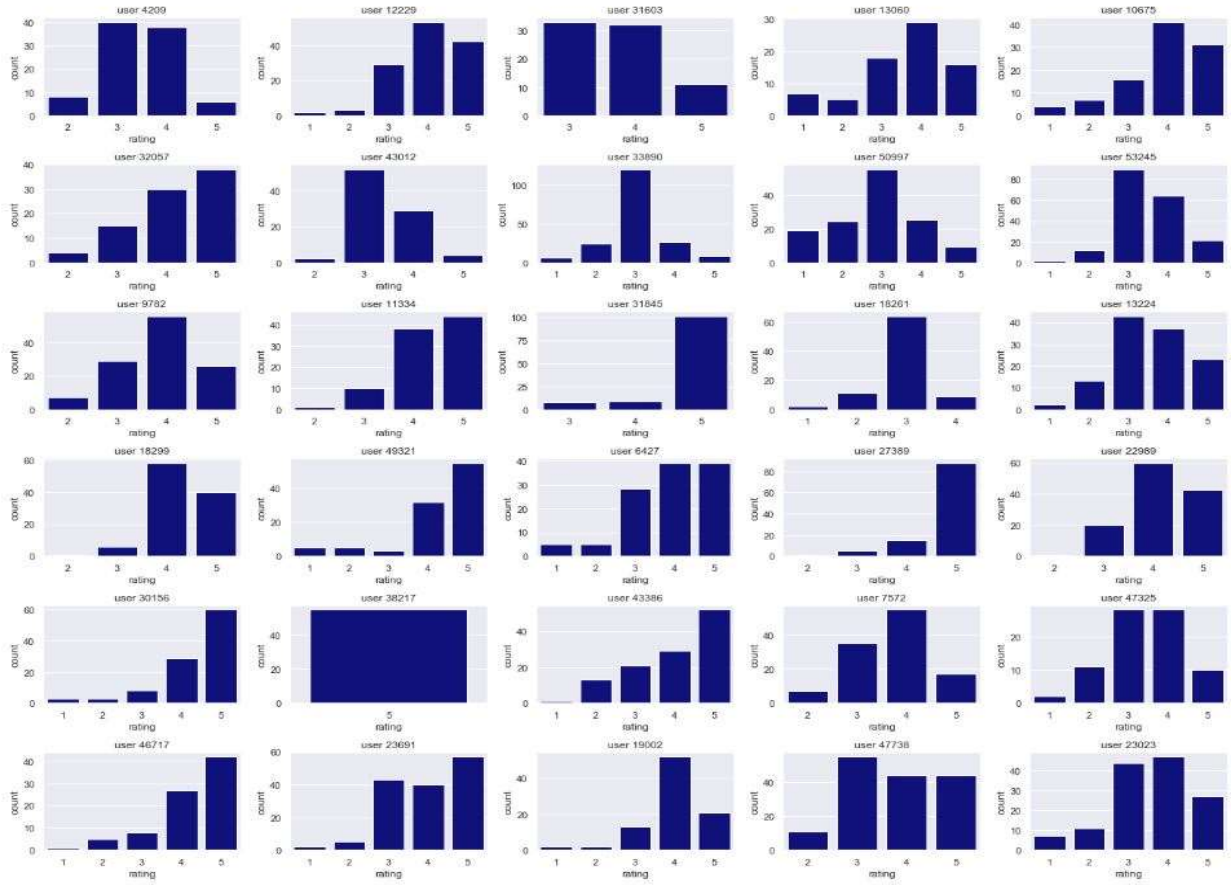
The conditions for assuming the Central Limit Theorem implications are met here, since:

- the sample is more or less symmetric and bigger than 30, the value normally assumed for the validity of the Central Limit Theorem.
- there was no replacement when retrieving the sample, but the observations are considered to be independent, since they constitute less than 10% of the population size (the 10% rule).
- the sample was taken randomly (I assumed that it was).

I created a bootstrap confidence interval for the mean between [111.64700883, 112.09271395]. This is a pretty close range. If we repeated measurements over and over again, 95% of the observed values for the mean of the number of ratings given by users would be very close to 112.

- Do distributions of ratings among users vary significantly?

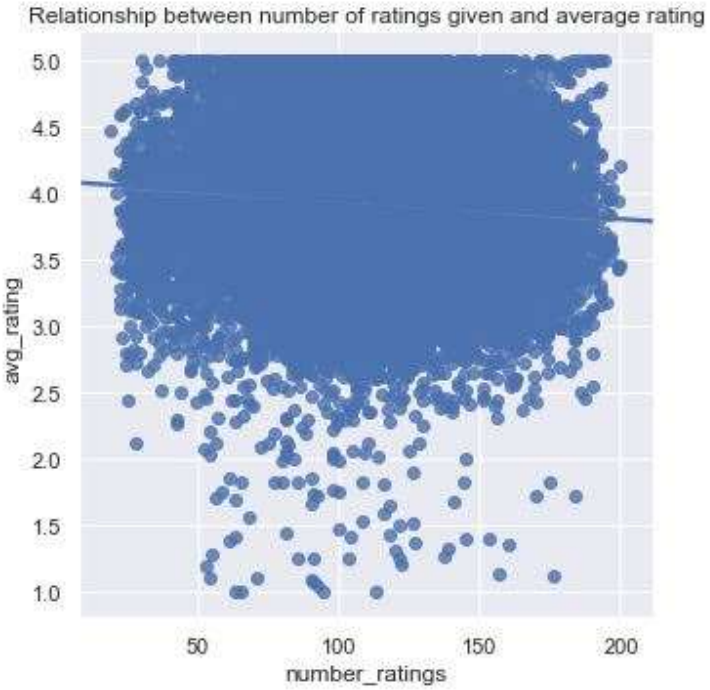
For a sample of 30 users:



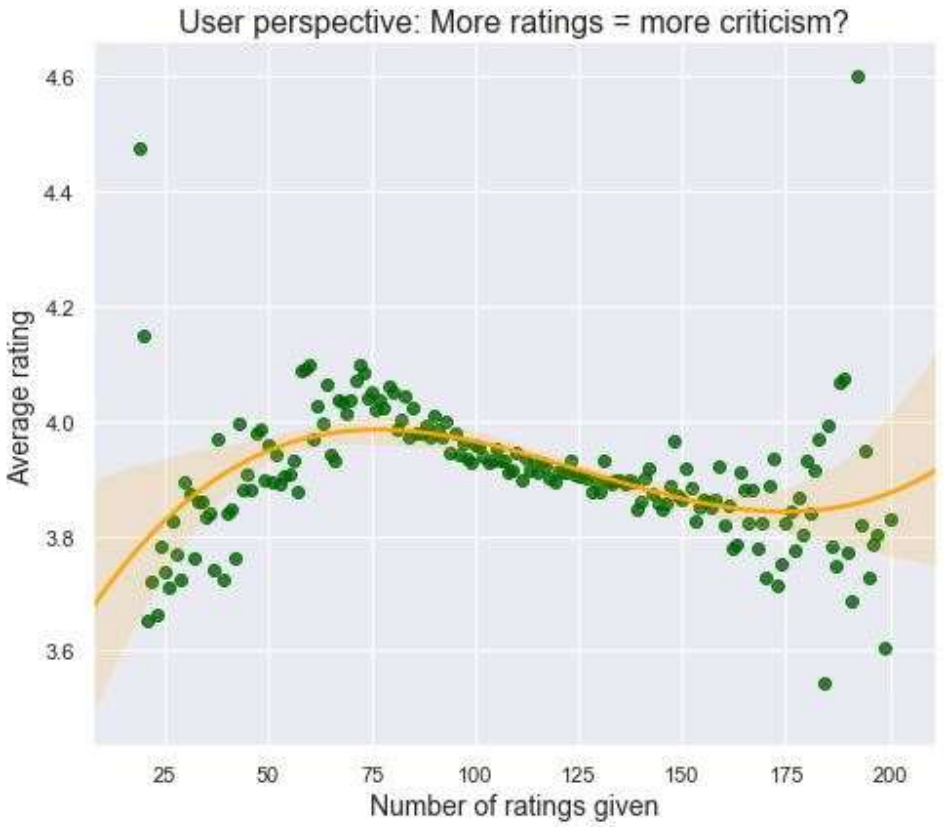
It is straightforward to see that the distributions of ratings vary significantly among users. We can see right away that the average rating by users are also significantly different.

- How does the AVERAGE rating given by a user evolves as the number of his/her number of given ratings increases, on average? Do people tend to become more critic and give lower ratings?

I first created a new data frame with user_id, number of ratings given and average rating. When considering all the points, the scatter plot looks like this:



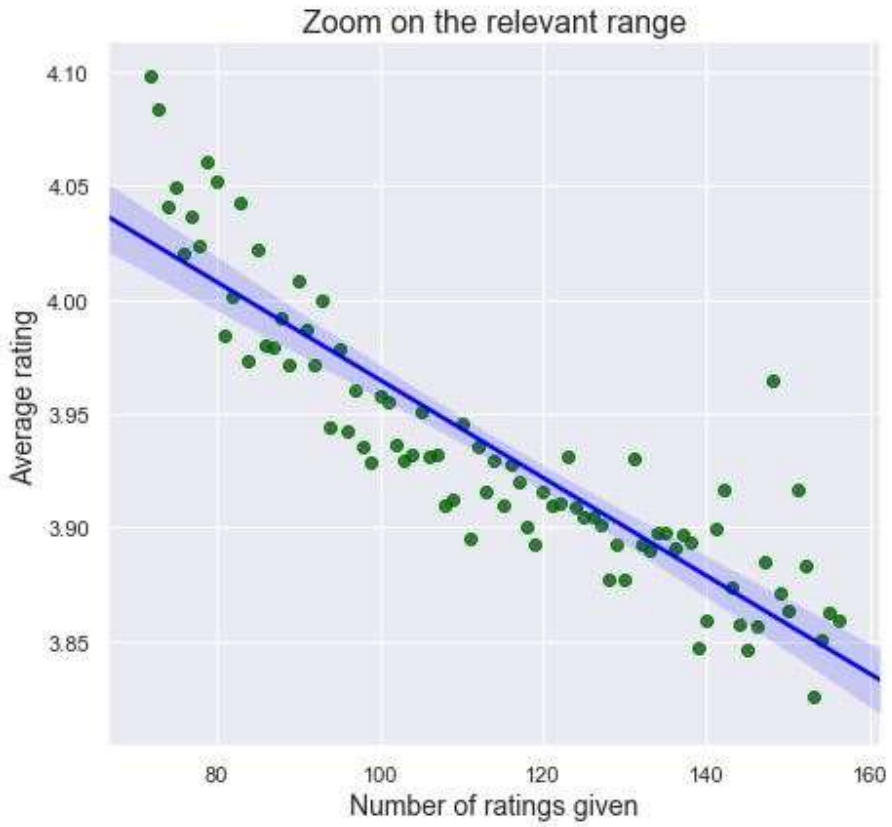
In order to properly see the relationship I am going to agglomerate the users with the same number of given ratings and compute the average of their average rating given. Otherwise I would have almost 6 million points on the scatter plot as seen above. In any case, it's a negative relationship.



Each point encompasses a different number of users. As we have seen in the previous histogram, most of the data reside between 75 and 150 ratings. In fact, almost 90%, as computed below. Curiously, we see here that up until 75 ratings given by a "user" the average rating tends to increase and probably the enthusiasm for books and reading as well. After that, and until 150, which is exactly the range where most of the data is, the average rating linearly decreases as the number of ratings increase. It then spreads across the range between 3.5 and 4.6 without a clear tendency.

This may suggest that, on average, the user tends to become more critic and give lower ratings as it reads more books.

Taking now a closer look at the relevant range, we see more clearly this negative relationship:

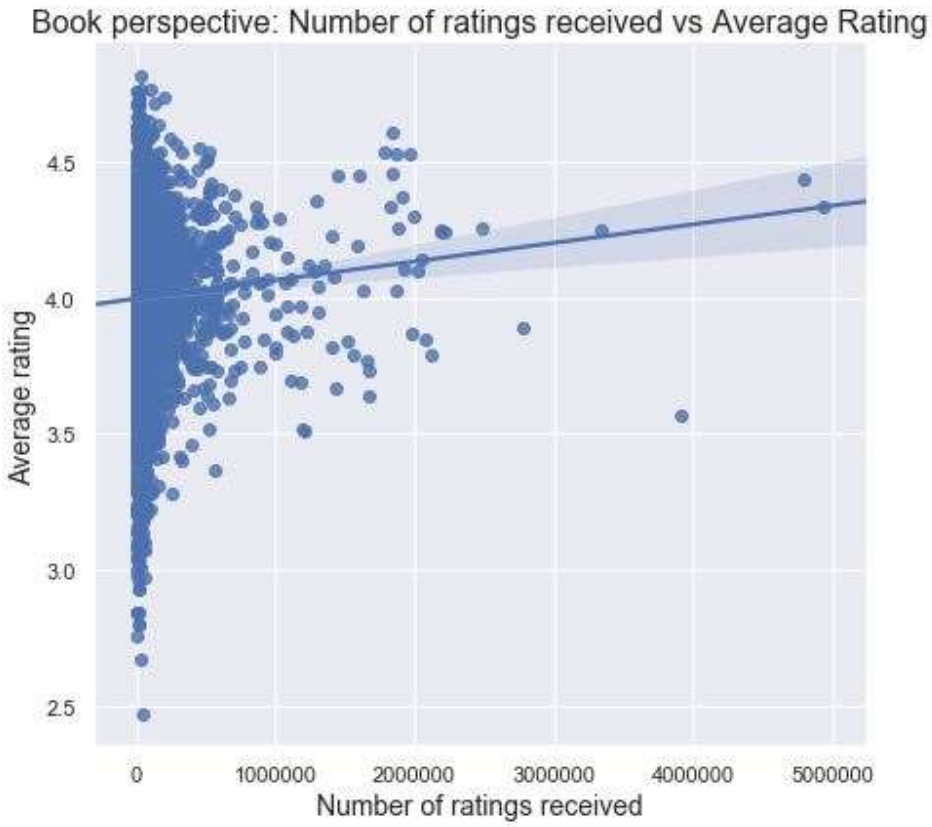


There is a very strong negative correlation (-0.89) between the number of ratings a user has given and the average rating, suggesting, again, that the user tends to become less enthusiastic, on average, as he/she reads more books.

- How does the average rating of a book relate to the number of ratings it has received?

Now from the book perspective, we now see a slight positive relationship between the number of ratings (or fame) a book has received and its average rating. We see indeed a small positive correlation coefficient equal to 4.5%. Can we be sure this correlation is not due by chance? To know that I did a test of correlation, where

- **Null hypothesis**: "the two variables are completely uncorrelated"
- **Alternate hypothesis**: "the two variables are correlated"
- **Test statistic**: "pearson correlation coefficient"



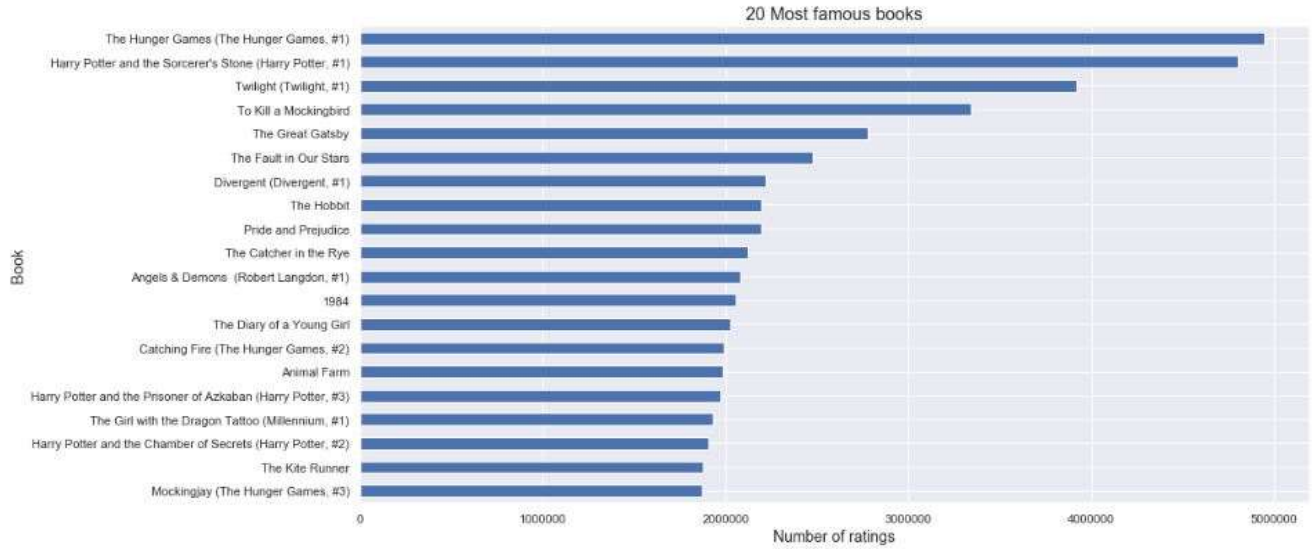
With a p-value equal to 0.0053, we reject the hypothesis that the two variables are not correlated, considering a level of significance of 1%. As the number of ratings a book gets increases, the higher the chance that it will have a higher average rating. In any case, there is too much noise and the correlation is not clear.

- How does the distribution of ratings given by a user evolve as the number of ratings increases, on average?

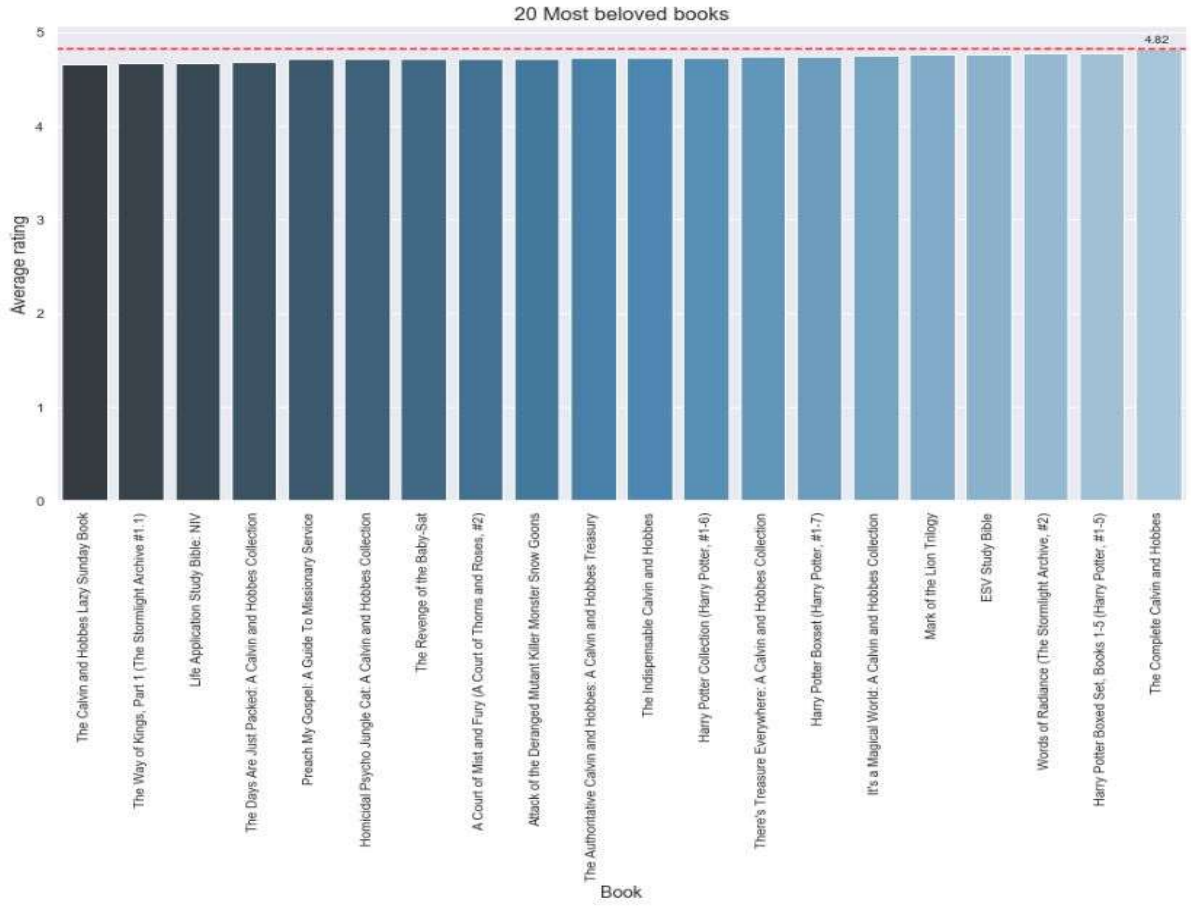
To do this I built a plot using bokeh (please see the jupyter notebook) where we can clearly see that there is a tendency on the distribution to become rightly-skewed as the number of given ratings increase.

3.3. Books

- What are the most famous books?



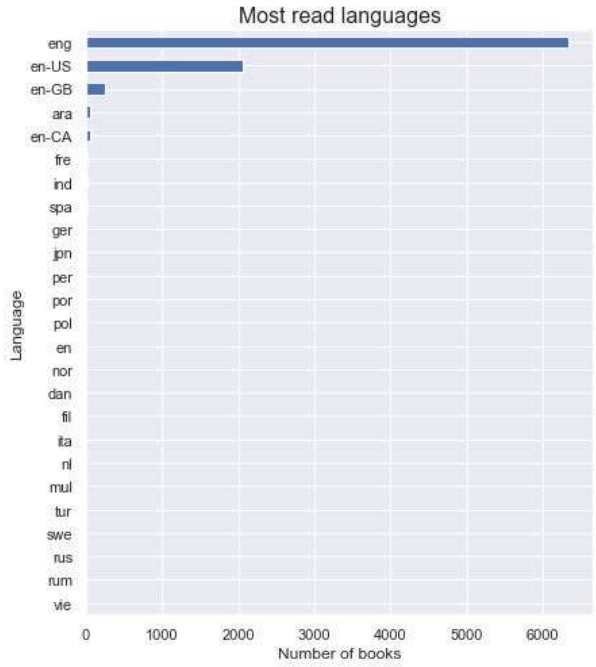
- What are the most beloved books?

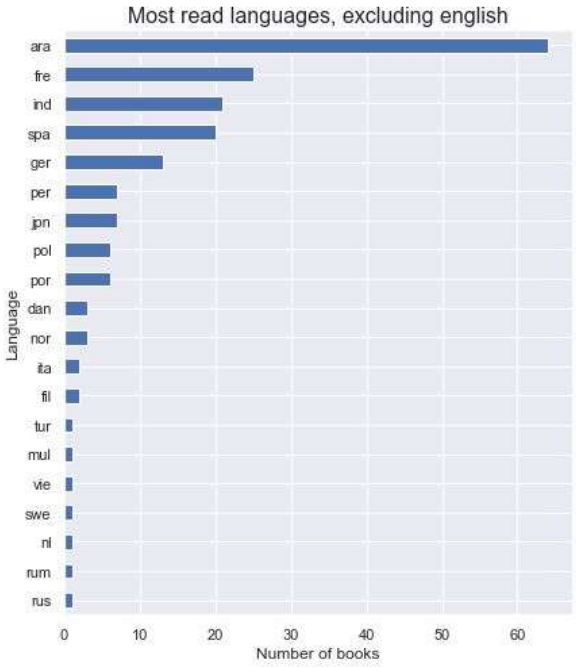


By famous, I meant the books with the highest number of given ratings. By beloved, the books with the highest average ratings

- What are the most read languages?

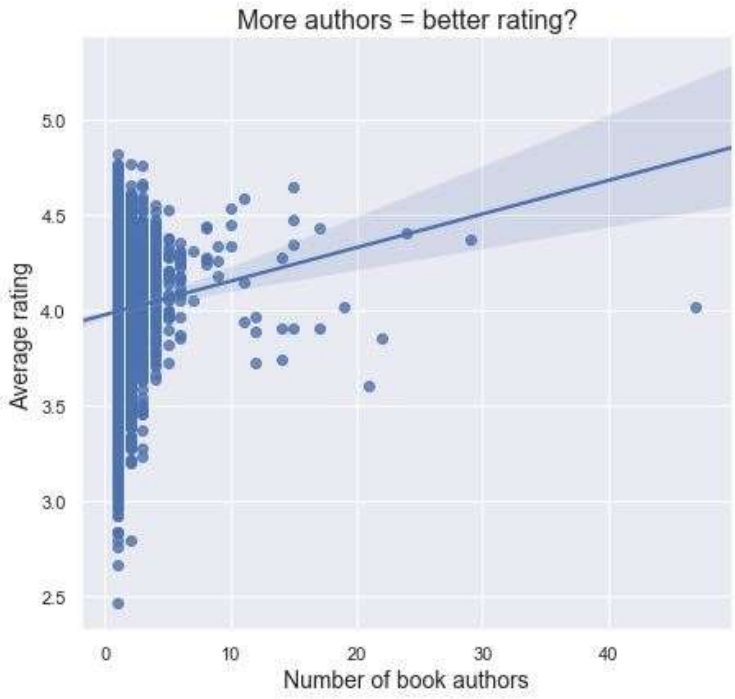
English is by far the most read language, and almost the only one in this data set, in relative terms.



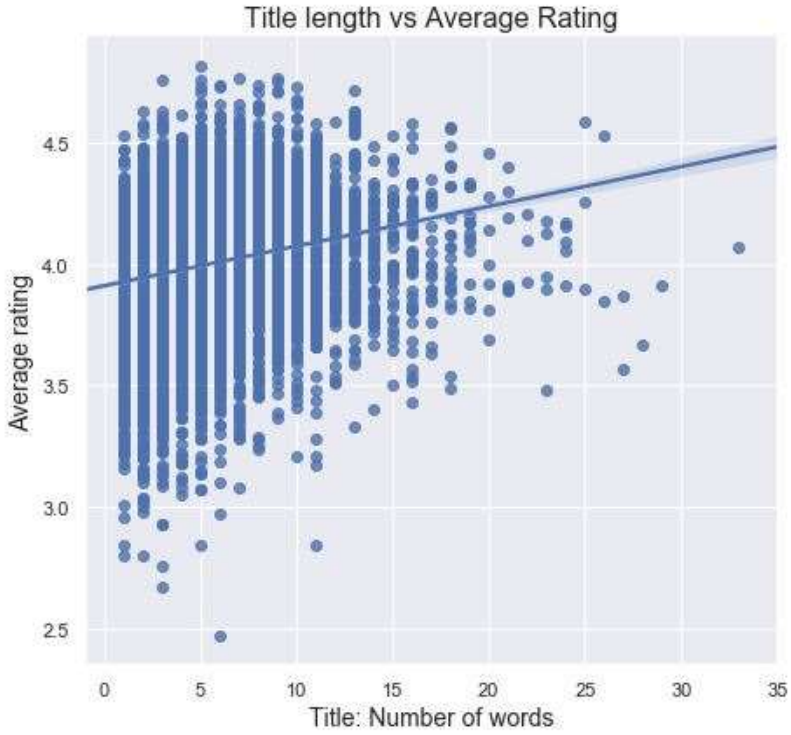


- Is having more authors better?

Here the correlation is slightly more positive (7.5%): as the number of authors increases, so does the average rating. There is great variability, so this result should not be taken with too much weight.



- Does the length of the title influence the average rating?



There is also a small positive correlation (21%) between the length of the title and the average rating.

4) Recommendation System