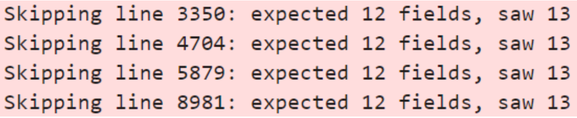
While the project, we encountered some data consistency errors.
When extracting the file by panda, we noticed that there were errors and 4 lines were missing.

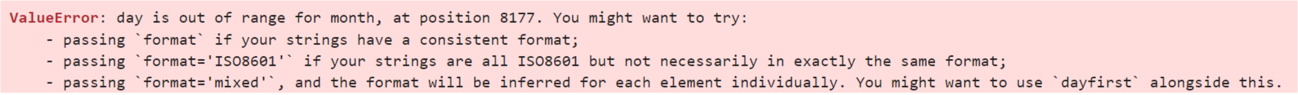


```
Skipping line 3350: expected 12 fields, saw 13
Skipping line 4704: expected 12 fields, saw 13
Skipping line 5879: expected 12 fields, saw 13
Skipping line 8981: expected 12 fields, saw 13
```

After looking at the lines one by one, and comparing the information on the Gooreads website. We concluded that there were data writing errors. There are excess commas that have been removed. This had the effect of shifting the columns and creating new ones. For example:

12224,Streetcar Suburbs: The Process of Growth in Boston 1870-1900,Sam Bass Warner, Jr./Sam B. Warner,3.58,0674842111,9780674842113,en-US,236,61,6,4/20/2004,Harvard University Press
Becomes : 12224,Streetcar Suburbs: The Process of Growth in Boston 1870-1900,Sam Bass Warner-Jr./Sam B. Warner,3.58,0674842111,9780674842113,en-US,236,61,6,4/20/2004,Harvard University Press

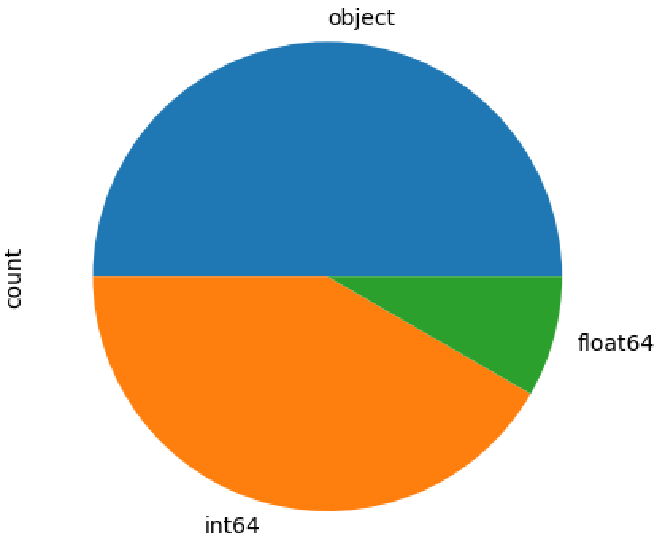
We changed the format of the dates to be able to work on them and new errors appeared which were modified in the new file.

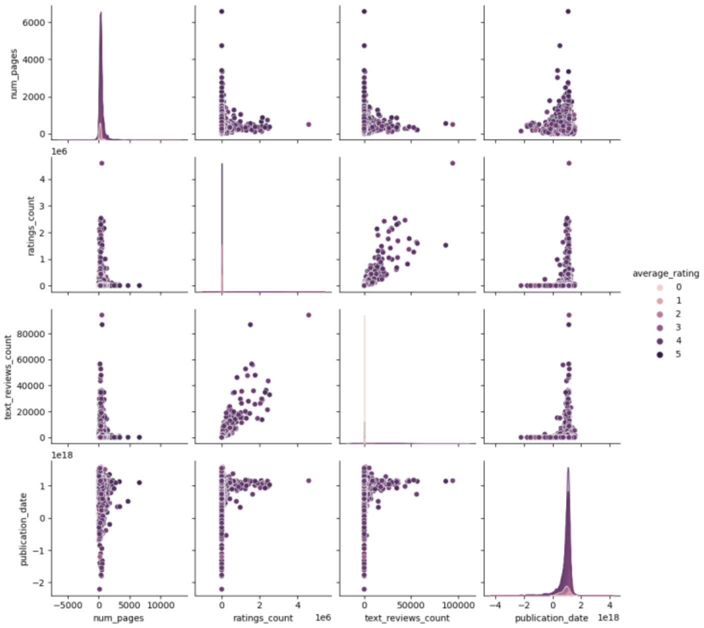


```
ValueError: day is out of range for month, at position 8177. You might want to try:
    - passing `format` if your strings have a consistent format;
    - passing `format='ISO8601'` if your strings are all ISO8601 but not necessarily in exactly the same format;
    - passing `format='mixed'`, and the format will be inferred for each element individually. You might want to use `dayfirst` alongside this.
```

31373,In Pursuit of the Proper Sinner (Inspector Lynley #10),Elizabeth George,4.10,0553575104,9780553575101,eng,718,10608,295,11/31/2000,Bantam Books. H
Here we see that the date is incorrect, November 31 does not exist and we have modified it to November 30 as on the Gooreads website.

Thanks to this manual control we manage to reload the file with all the initial lines.

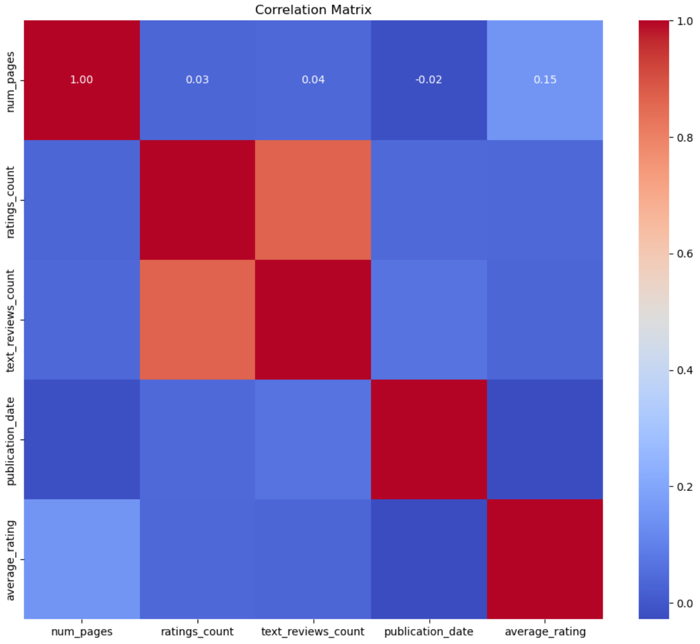




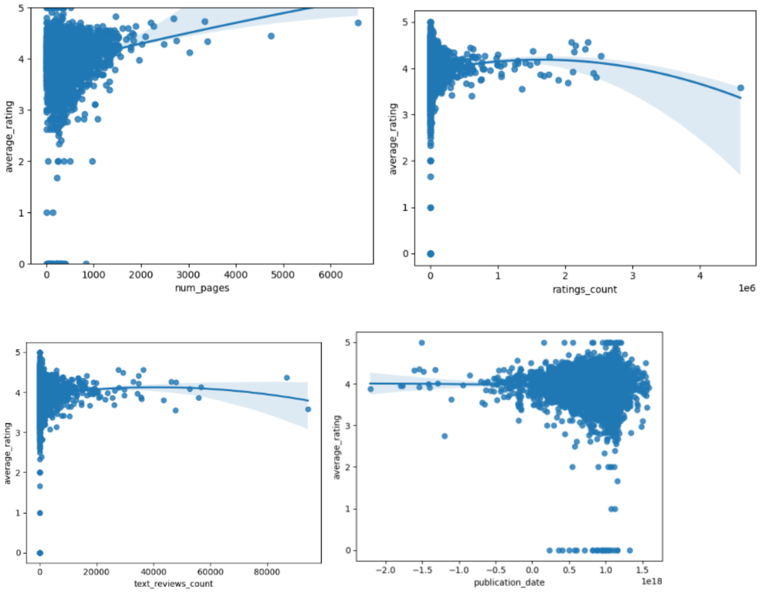
We can see that text_reviews_count/rating_count are highly correlated because they grow together. Ratings_count/num_pages are decorrelated because we can see some L on the graph that is the sign that the variables evolve independently. Num_pages/rating_count and num_pages/text_reviews_count are weakly correlated because they do not do a perfect L.

We are not able to see on this graph that a couple of columns are correlated with average_rating.

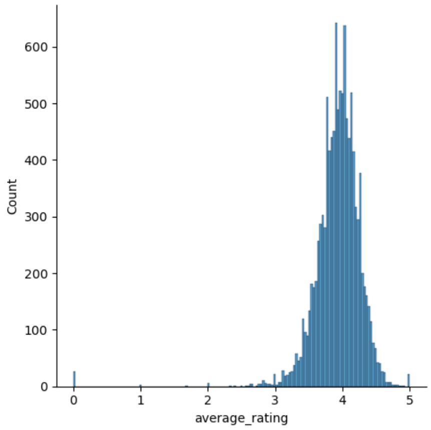
Obviously the number of ratings given is lower for books that were published before the creation of the website. The prediction will probably be less reliable for old publication dates.



The most interesting line/column is average_rating. The best correlation is with num_pages, then ratings_count, text_reviews_count. Publication_date is decorrelated from average_count but is correlated with text_reviews_count. Using both publication_date and text_reviews_count may help an algorithm to better predict average_rating.



We can see on this graph what the correlation with average_rating looks like.



Analysis of the prediction models :

For the first test we used linear regression to predict the average rating based on quantitative variables such as page count and ratings. The mean square error was equal to 0.06440275632143355.

For the second test we used Gradient Boosting Regressor to predict the average rating built on the other prediction models. The mean square error was equal to 0.06036940510948817.

For the third model we used Random Forest Regressor which is a meta estimator that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The mean square error was equal to 0.05859136699392239.

Based on the mean square error, the Random Forest Regressor is the best model for our prediction model for books ranking.