NATHASYA GUNAWAN | FINAL PROJECT

01

book recommendation system.

NATHASYA GUNAWAN | JCDS-09 JKT

→

02

HI, I'M NATHASYA GUNAWAN!

BACHELOR OF ARTS IN MEDIA, COMMUNICATION, AND CULTURAL STUDIES - NEWCASTLE UNIVERSITY

SEO CONTENT WRITER & DIGITAL STRATEGIC PLANNER

Timeline

03

BUSINESS PROBLEM	DATASET	EDA	MACHINE LEARNING	EVALUATION

Business Problem

04

Recommendation System in Book Retail

Big data is now being utilized at a level that we could have never previously imagined, but the important part still remains on how we apply the data in a business context, and how we make the most out of it.

For online book retailers, product ratings can play a huge role for making sound business decisions. As the data on product ratings continue to grow over time, companies can take advantage of this information and enhance customer experiences.

Business Problem

05

Recommendation System in Book Retail

Can book recommendation system help book retail in giving out the best recommendation for their customer?

DATASET

The dataset contains about 1 million ratings across 10000 different books. In most cases, there are at least 10 books rated by each user and the rating lies between 0 and 5.

GOODBOOKS-10K

https://www.kaggle.com/zygmunt/goodbooks-10k

DATASET

- books.csv - metadata for each book (goodreads IDs, authors, title, average rating, etc.) - (1000 x 23)
- book_tags.csv - contains tags/shelves/genres assigned by users to books. Tags in this file are represented by their IDs. - (999912 x 3)
- tags.csv - translates tag IDs to names. - (34252 x 2)
- rating.csv - contains ratings - (5976479 x 3)

DATA FEATURES

07

TITLE

The
name under which the book was published.

AUTHORS

Names
of the authors of the book.
Multiple authors are delimited with a comma (,).

AVERAGE_RATING

The
average rating of the book received in total.

DATA FEATURES

08

RATINGS_COUNT

Total
number of ratings the book received.

USER_ID

A
unique Identification number for each user.

BOOK_ID

A
unique Identification number for each book.

DATA FEATURES

09

RATINGS

Ratings given to a book by a user.

TAG_NAME

Book genres assigned to each book

BOOKS_COUNT

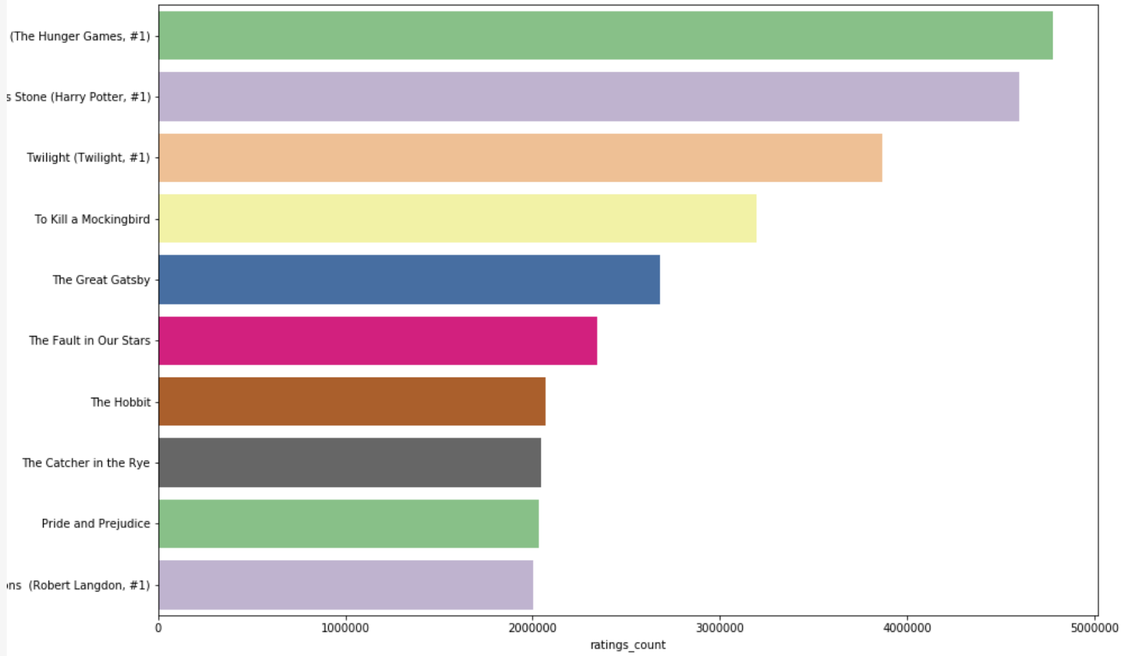
Number
of editions a book have.

EXPLORATORY DATA ANALYSIS

10

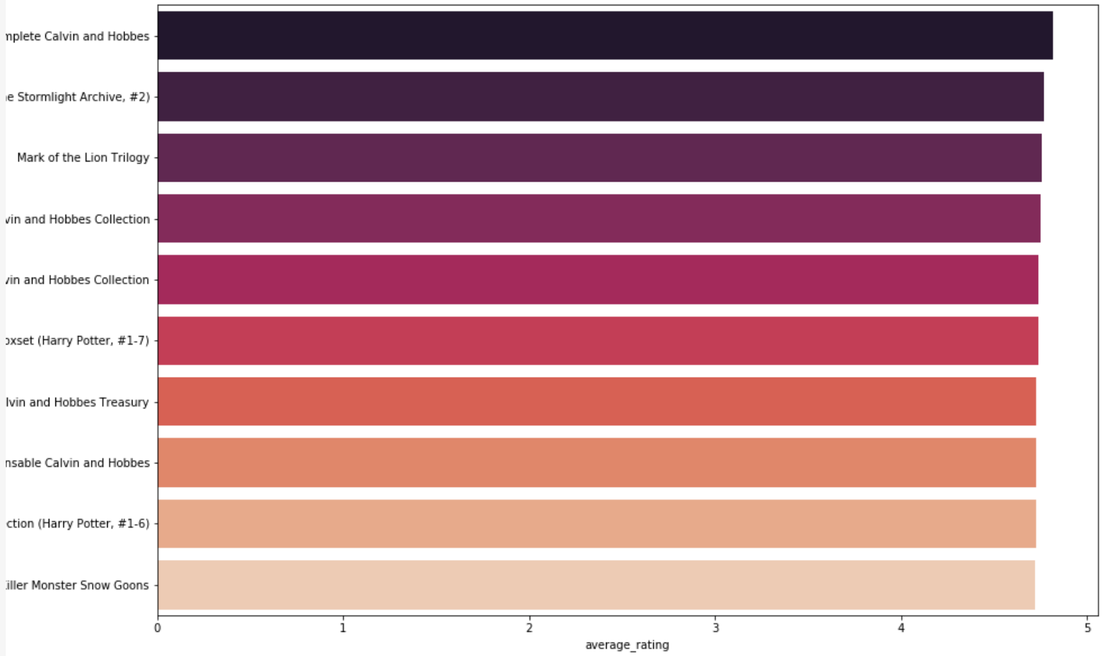
→

Most Rated Books



Books with the highest amount of ratings given on Goodreads

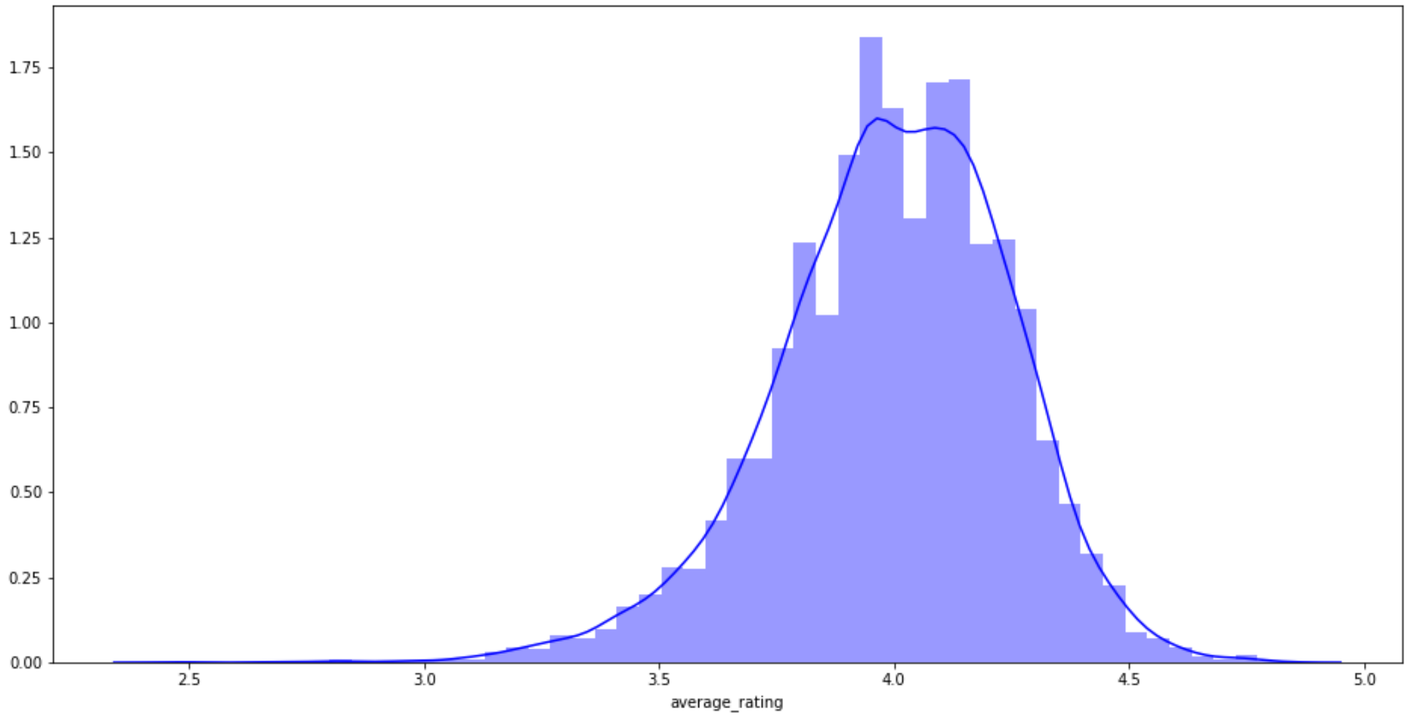
Top Rated Books



Books with the highest ratings on Goodreads

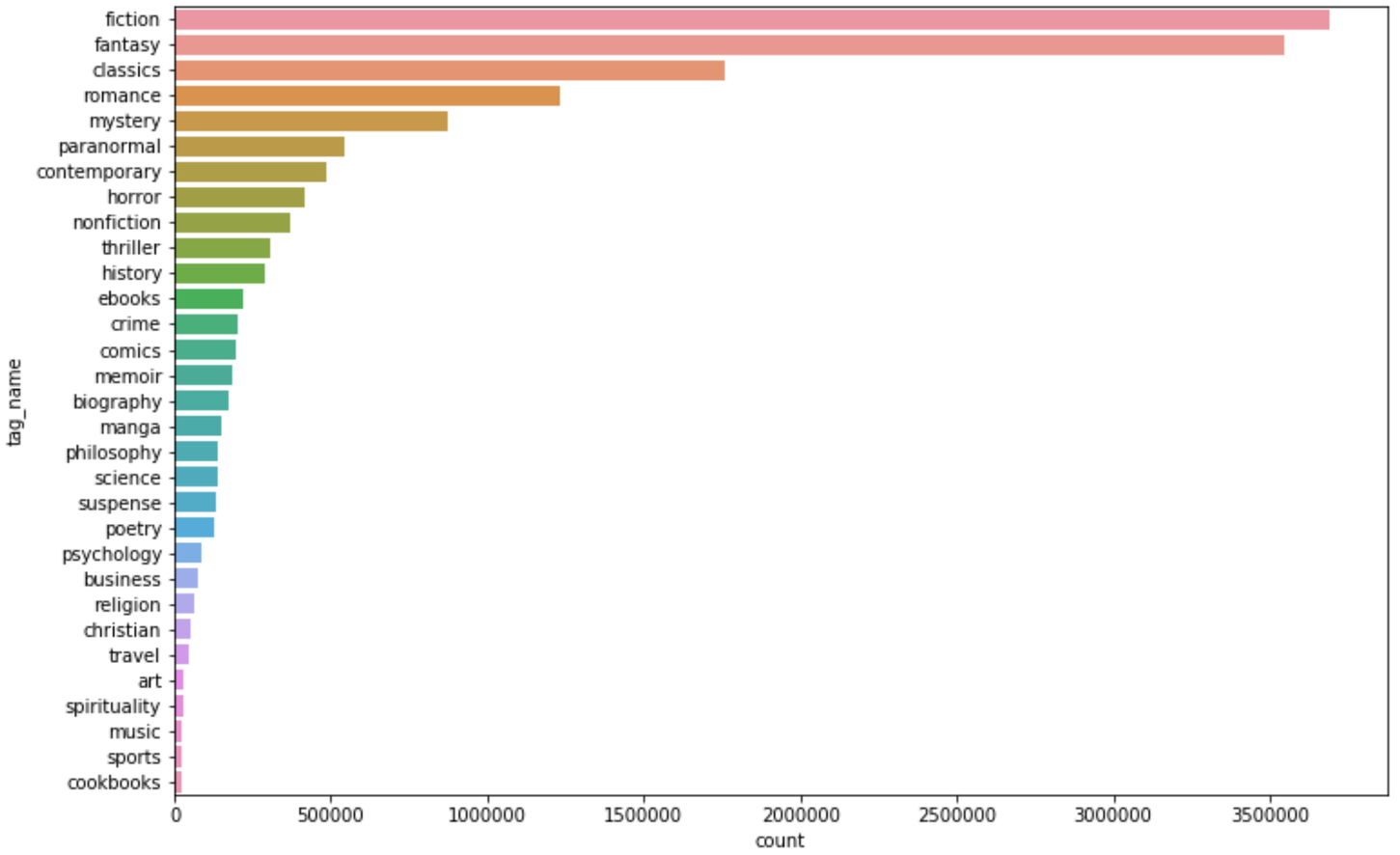
→

Book Rating Distribution



Most people tend to give quite positive ratings to books. Most of the ratings are in the 3-5 range, while very few ratings are in the 1-2 range.

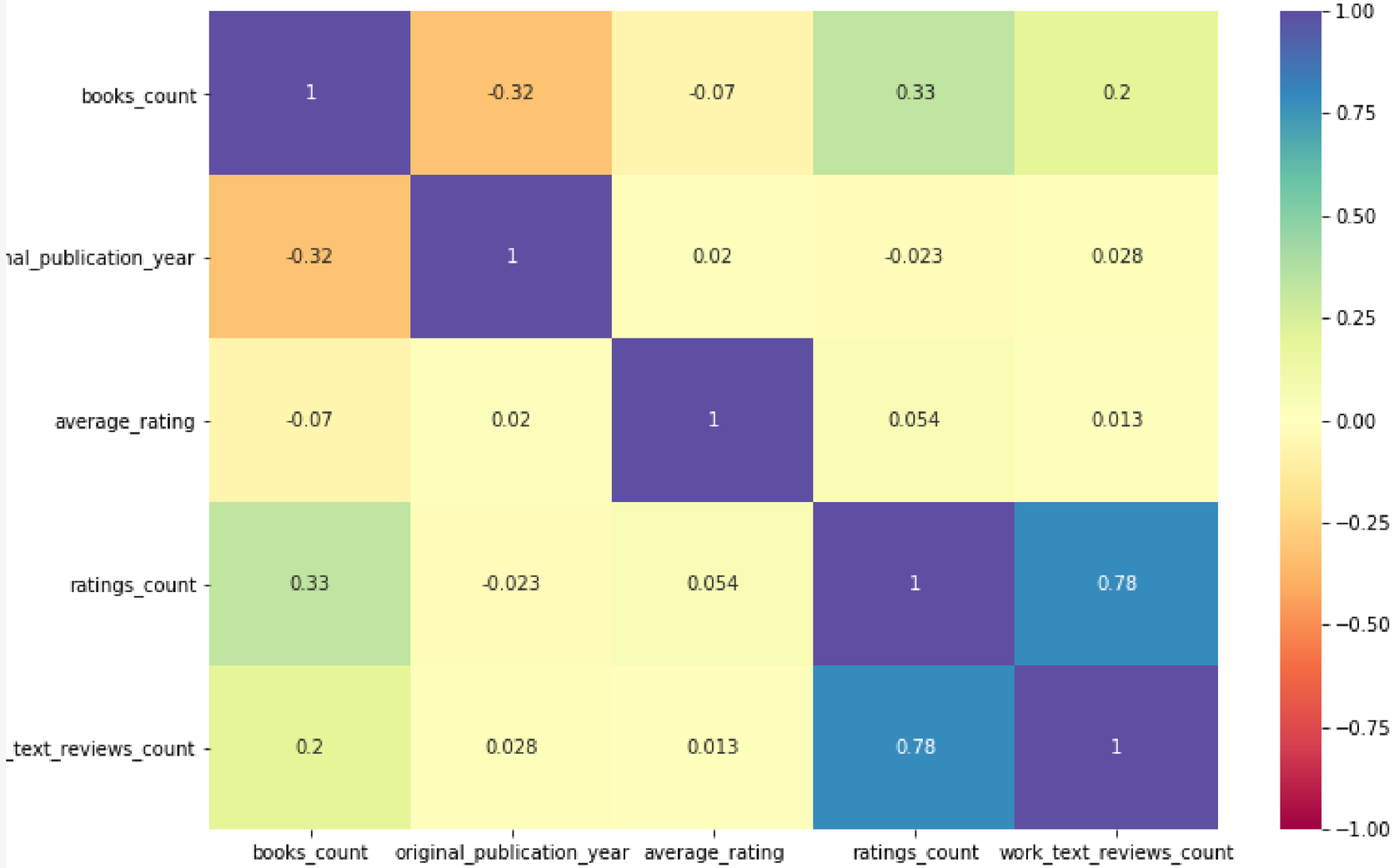
Book Genres



In this data there is a high number of Fiction books, followed by Fantasy and Classics. Interestingly there is a low number of music, sports, and cookbooks in this data.

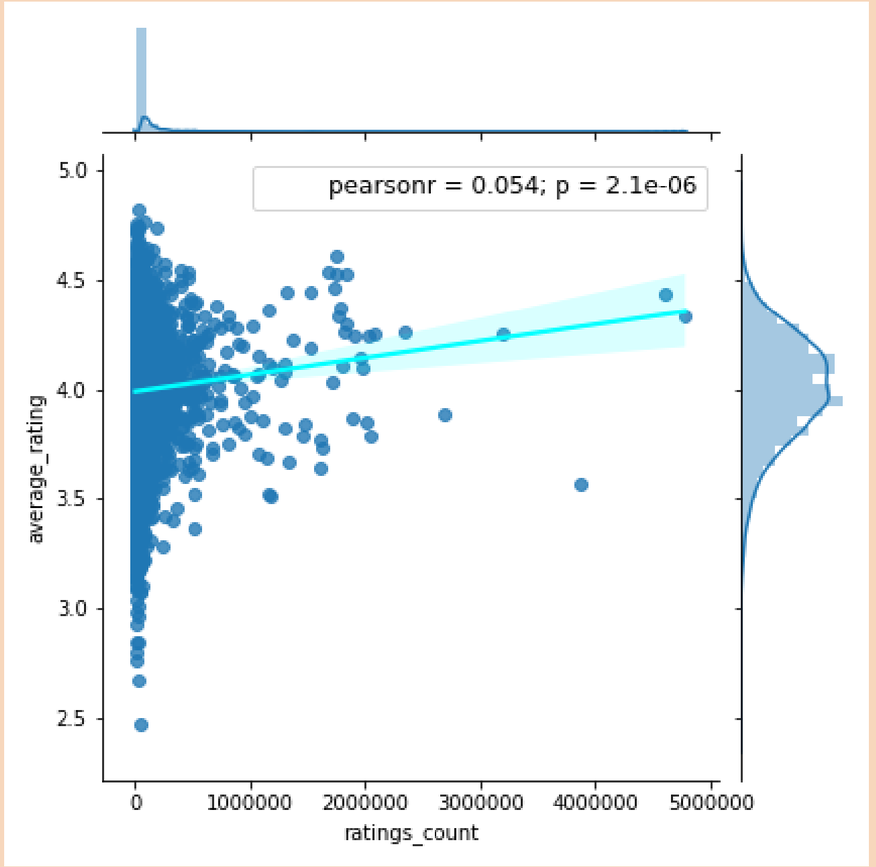
→

Correlation Heatmap



→

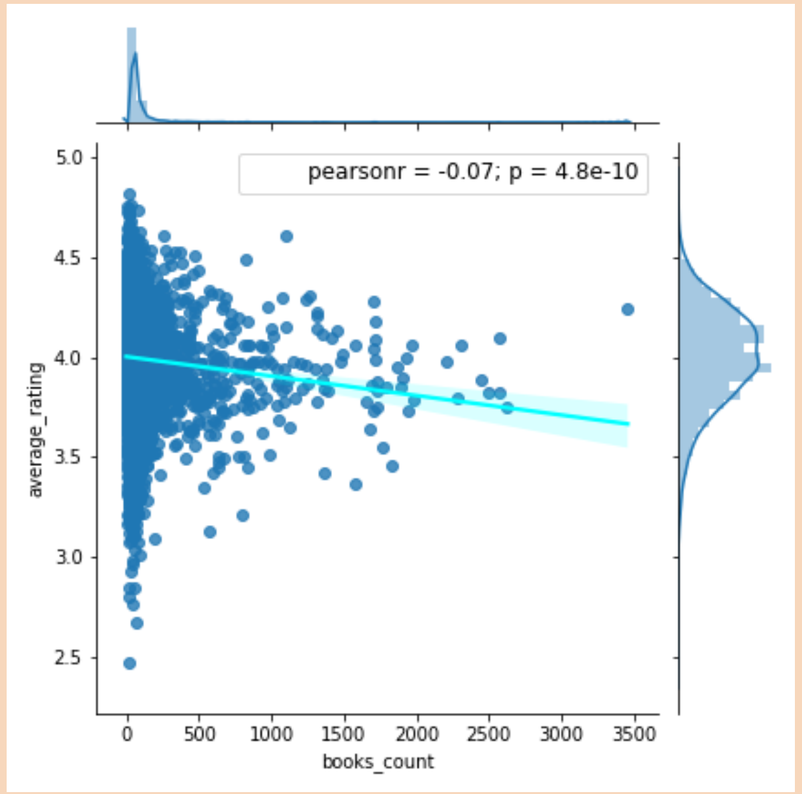
RELATIONSHIP BETWEEN THE NUMBER OF RATINGS & THE AVERAGE RATINGS



Theoretically, it might be that the popularity of a book (in terms of the number of ratings it receives) is associated with the average rating it receives, such that once a book is becoming popular it gets better ratings. However, our data shows that this is true only to a very small extent. The correlation between these variables is only 0.054.

→

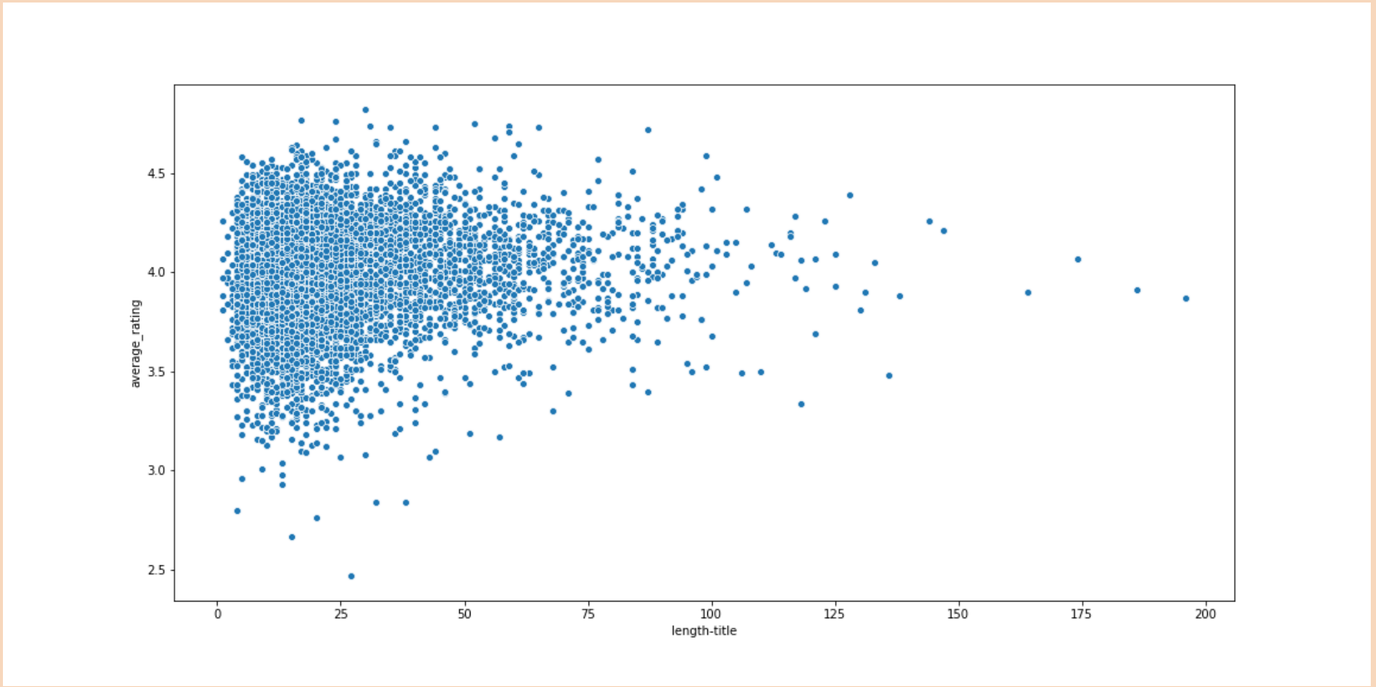
RELATIONSHIP BETWEEN THE NUMBER OF EDITIONS & THE AVERAGE RATINGS



The dataset contains information about how many editions of a book are available in books_count. These can either be different editions in the same language or also translations of the book into different languages. So one might assume, that the better the book is the more editions should be available. In fact, data show exactly the opposite pattern: The more editions a book has the lower is the average rating. The causal direction of this association is of course unclear here.

→

DOES TITLE LENGTH AFFECTS RATING?



So, the highly rated books have rather short titles. The graph shows that a straight line can be plotted but very approximately to say that as the length of title increases, the rating remains constant (at around 4).

→

MACHINE LEARNING

CONTENT-BASED FILTERING

This approach utilizes the characteristics of an item to find items with similar properties. Those characteristics are the keywords of an item. For this particular machine learning, I utilize the feature Title, Authors, and Tag_Name (Genre) to give out recommendations.

Get recommendation for title:
Romeo and Juliet

CountVectorizer	TfidfVectorizer
Much Ado About Nothing	Much Ado About Nothing
The Taming of the Shrew	The Merchant of Venice
As You Like It	Measure for Measure
Twelfth Night	The Taming of the Shrew
Measure for Measure	As You Like It
Hamlet	Twelfth Night
The Merchant of Venice	Hamlet
Anne of the Island (Anne of Green Gables, #3)	A Midsummer Night's Dream
Women in Love (Brangwen Family, #2)	Absolute Fear (New Orleans, #4)
The Blue Castle	Between, Georgia

TFIDFVECTORIZER & COUNTVECTORIZER

Get recommendation for title:
Romeo and Juliet

```
Recommendation for 'Romeo and Juliet':

1:'Hamlet', with distance: 0.5446055579484093
2:'Macbeth', with distance: 0.5499705087391189
3:'The Great Gatsby', with distance: 0.5505435126812175
4:'To Kill a Mockingbird', with distance: 0.5587650419497711
5:'The Adventures of Huckleberry Finn', with distance: 0.567469425854512
6:'Pride and Prejudice', with distance: 0.5695462000464738
7:'Lord of the Flies', with distance: 0.5737375277153554
8:'Of Mice and Men', with distance: 0.5827465175266584
9:'Animal Farm', with distance: 0.5878748701562182
10:'Little Women (Little Women, #1)', with distance: 0.5890437006003466
```

COLLABORATIVE FILTERING

The collaborative filter recommender systems are based on interactions between users and items. Instead of focusing on the characteristics of an item, the system compares similar actions made by other users.

NEAREST NEIGHBORS

For this machine, I use Nearest Neighbors by utilizing ratings to find the nearest title to the title being input by the user.

→

EVALUATION

Both model seems to give out reasonable recommendations that will help customers who are looking for their next read. However each models come with their own weaknesses.

Content-Based Filtering

To get a better recommendation using content-based filtering, there needs to be data on either book descriptions or written text review. By utilizing these two data, we may be able to get a better recommendations.

Collaborative Filtering

As collaborative filtering rely on ratings given by the users, the data may be bias as popular books may receive more ratings.

THANK YOU!