The Next Great Reading Adventure

CREATING A RECOMMENDATION SYSTEM FOR A BOOK REVIEW WEBSITE

An exercise in content-based filtering

Rae'e Yamin

Why?

- Marketing:
  - Book vendors normally have to rely on customers' preferences, as well as Co-oping, in order to market books.
  - However, such methods are used as more of a wide net rather than personal tailoring to a customer's tastes.
  - With online retailers creating diverse ranking systems based on user reviews, there are now methods to know how popular a book is and how to cluster similar books.
  - With a more accurate account of a reader's purchases and review history, targeted marketing is more possible now than ever before.
- Fun:
  - There are many who would simply enjoy knowing what book they should read next.

Data

Data was collected from a Kaggle project that included a large dataset from Goodreads.com API. The link to the CSV file for the dataset can be found below:

Goodreads.com CSV

The data was collected into 10 columns:

- bookID
- Title
- Authors
- Average Rating
- ISBN
- ISBN13
- Language Code
- Number of pages
- Ratings Count
- Text Review Count

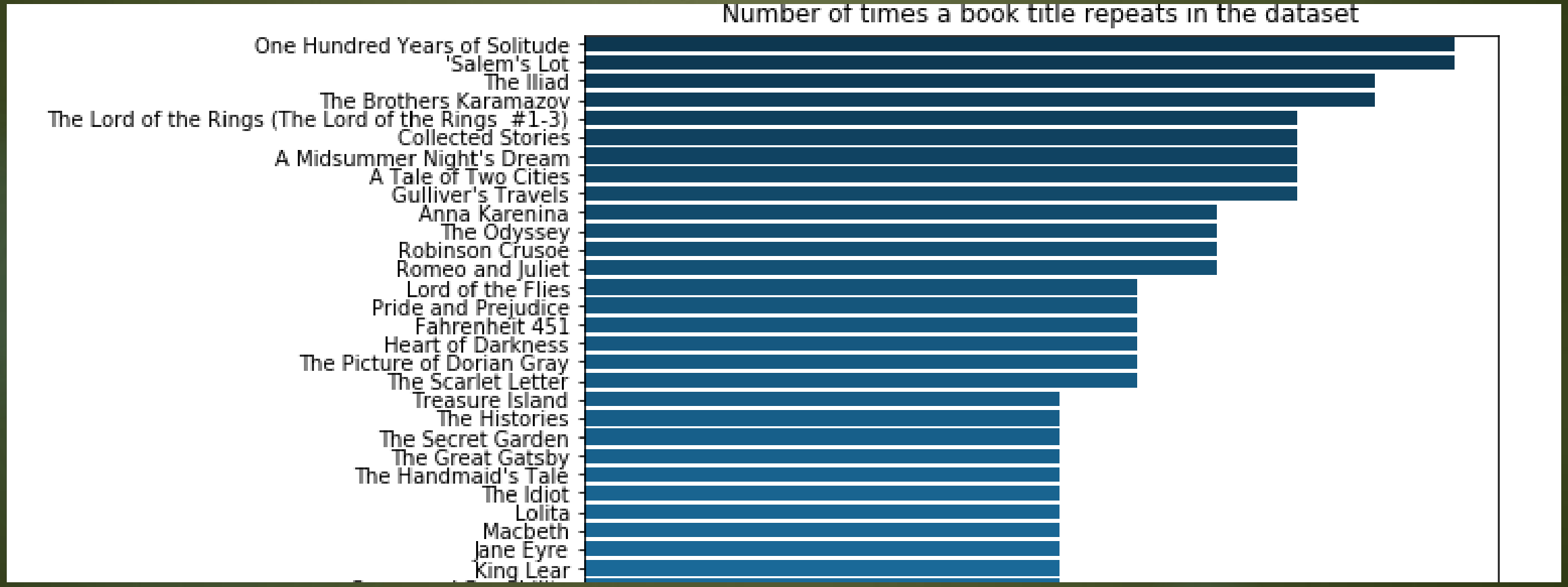
Data Cleaning

Two points to clean up:

- Problem 1: Misaligned columns
  - 6 rows contained data outside of the rest of dataframe of the data, and could not be read into the pandas dataframe.
  - *Solution*: Skip these 6 rows.
- Problem 2: Multiple authors
  - *Solution:* First author was designated as the primary author, while the ones that followed were secondary.

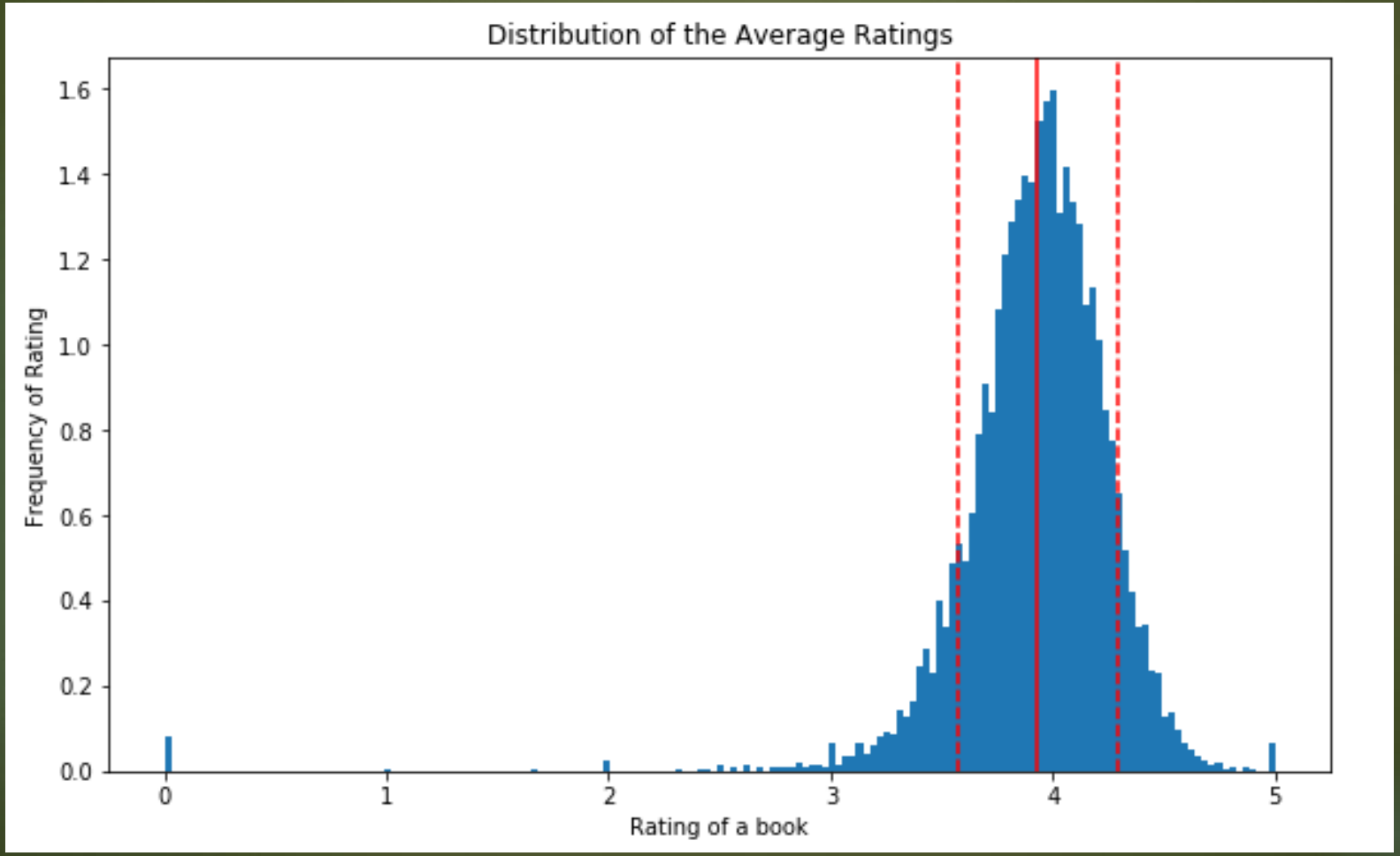
EDA

Exploring number of repeated titles in the dataset:



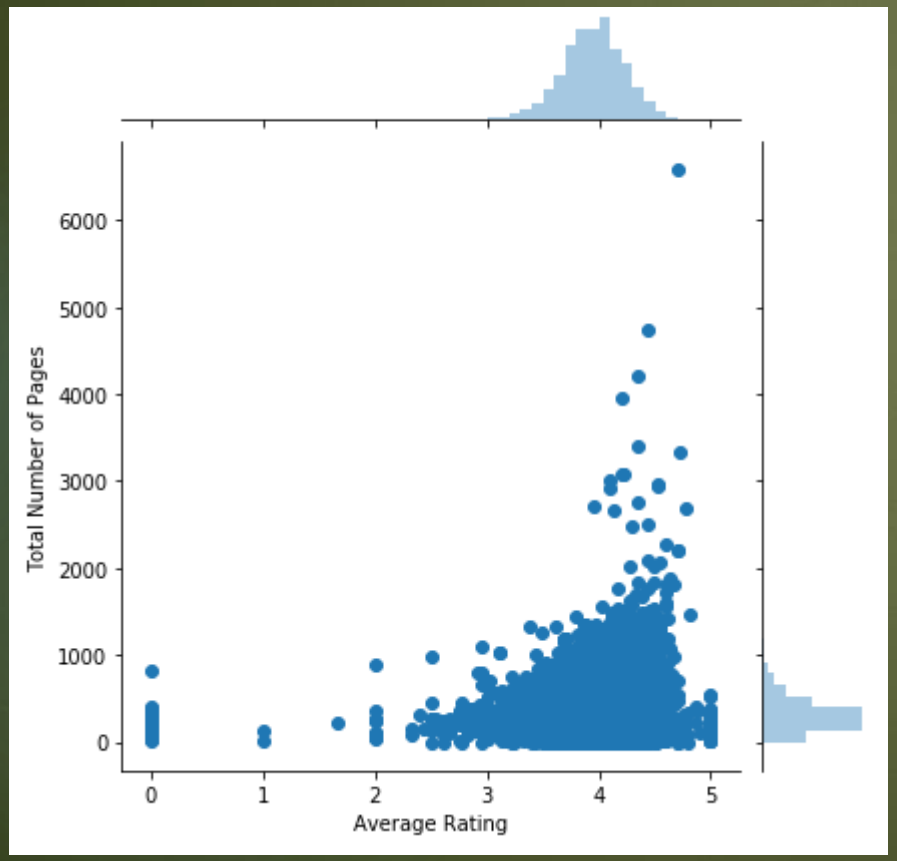
EDA

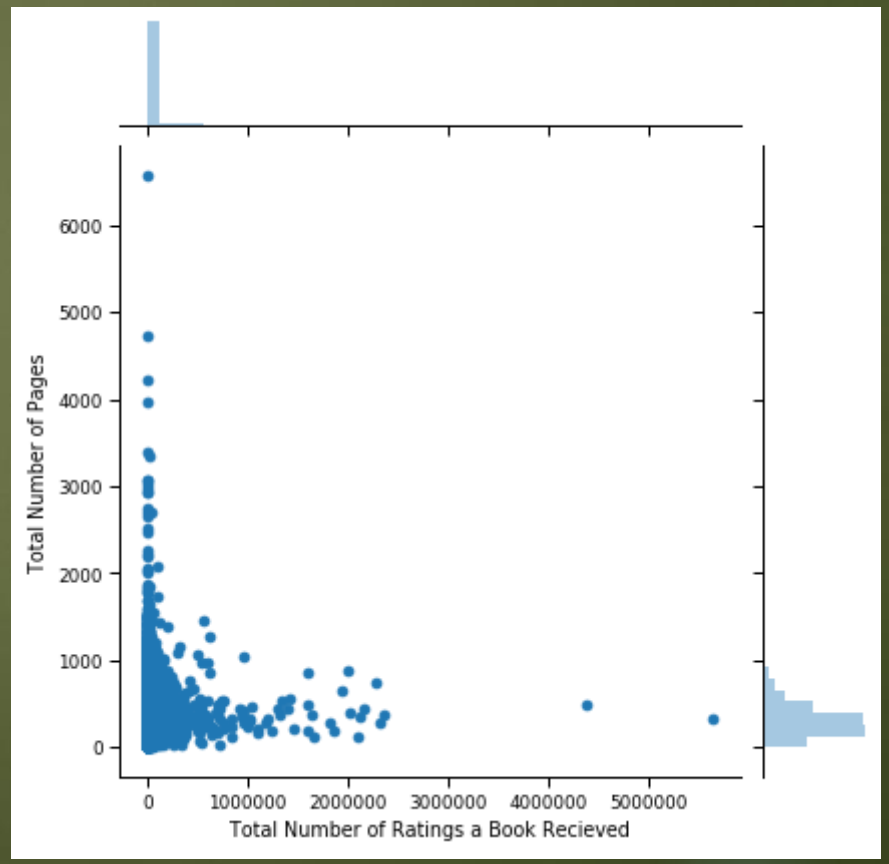
Distribution of average ratings by total count:



EDA

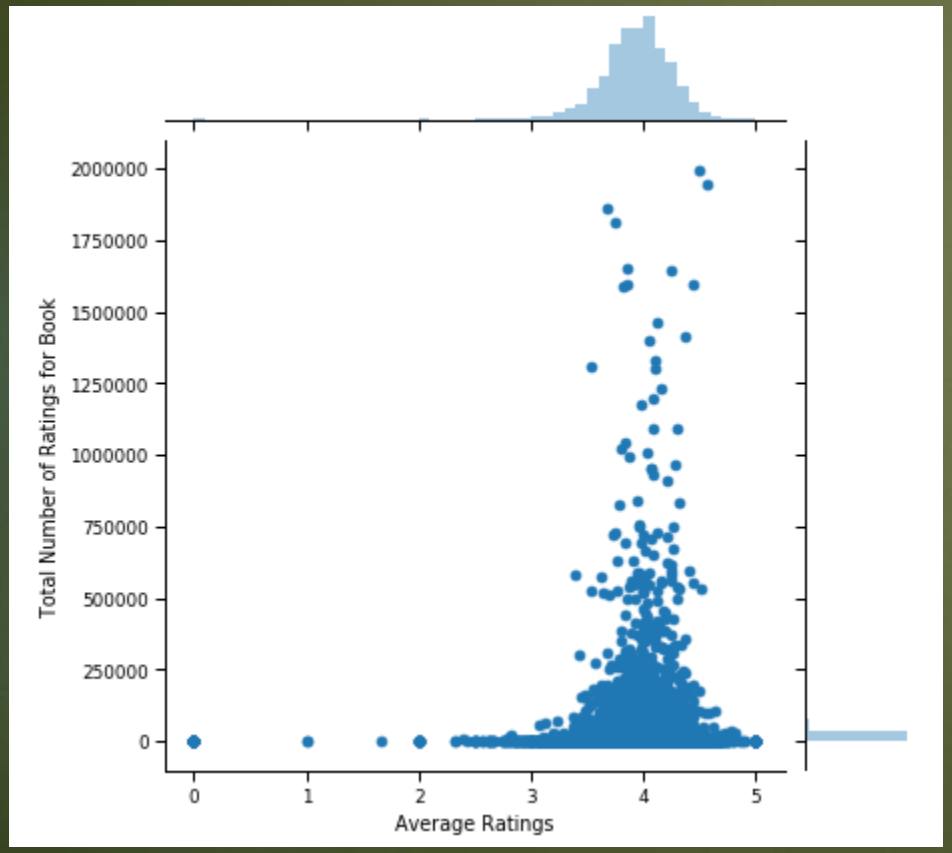
Relationship between average rating and number of pages:

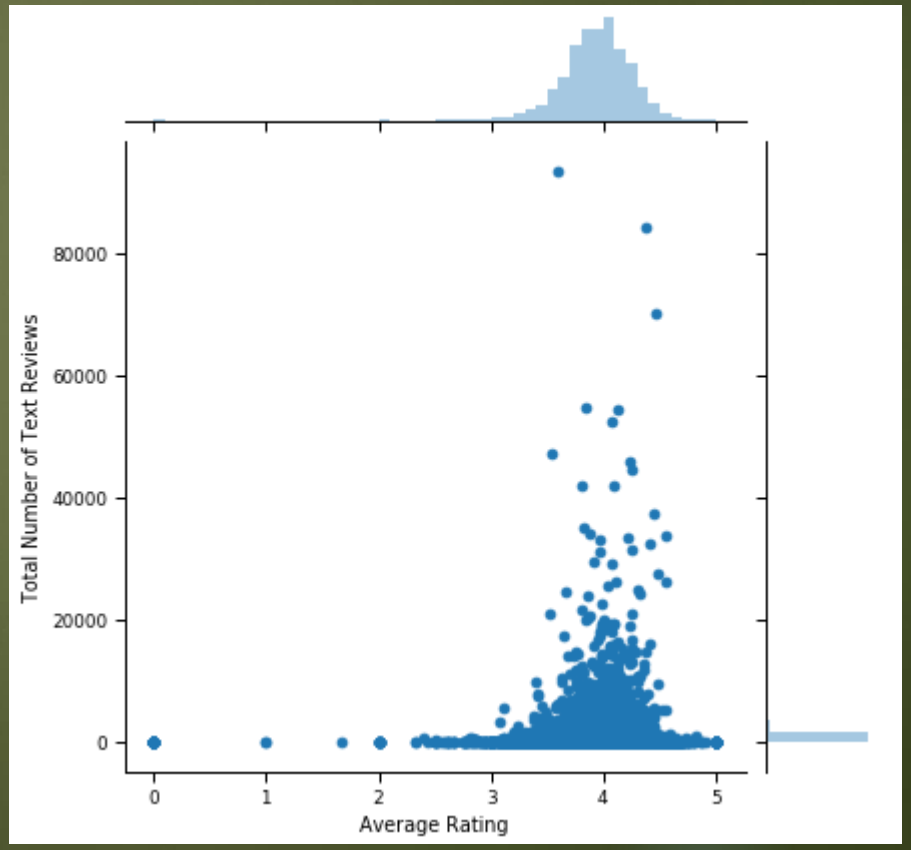




EDA

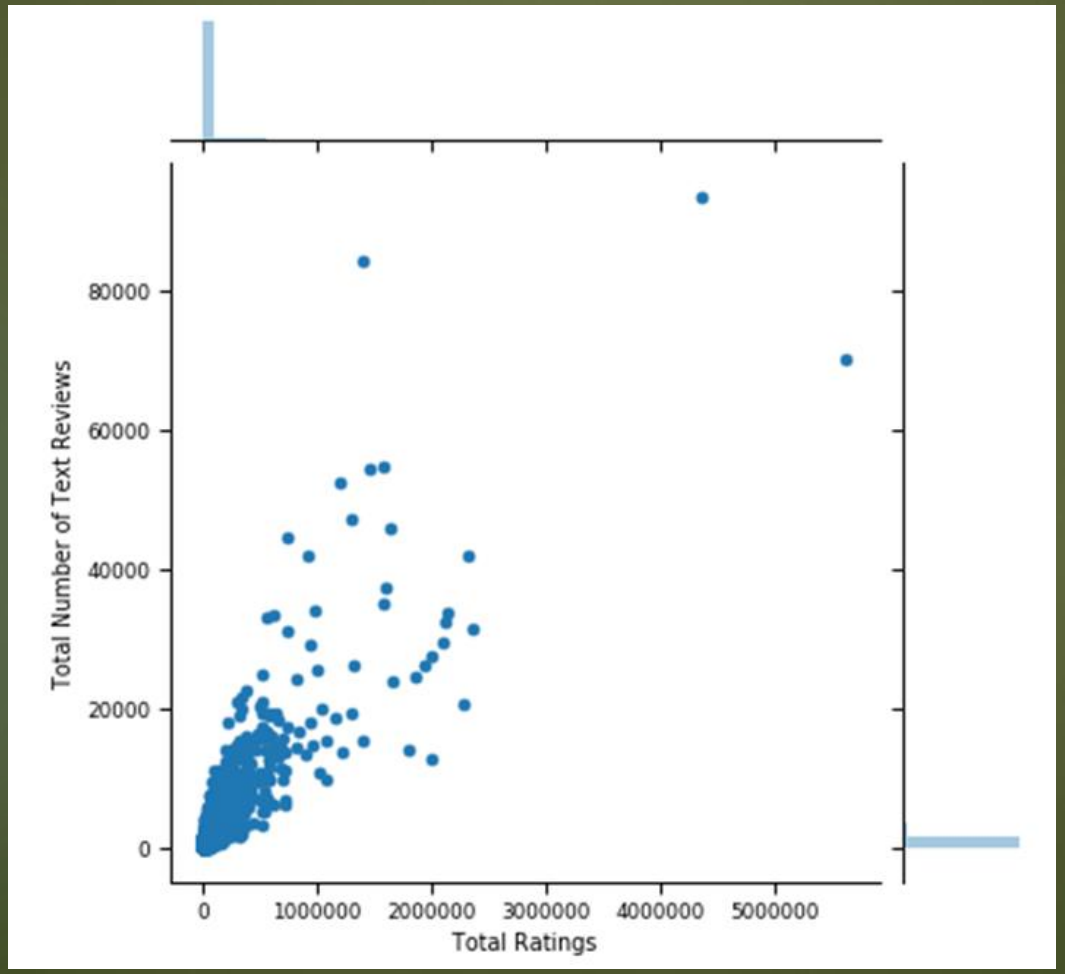
Average rating vs total reviews and total text reviews:





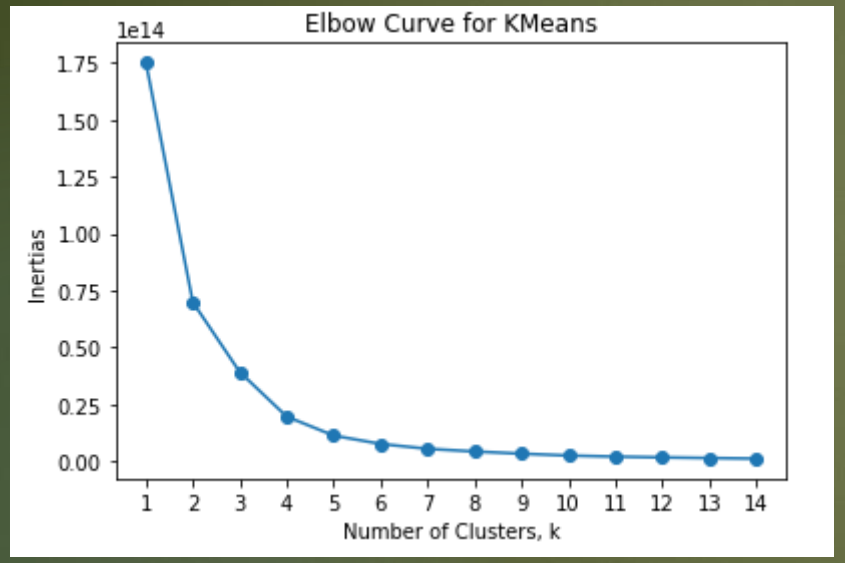
EDA

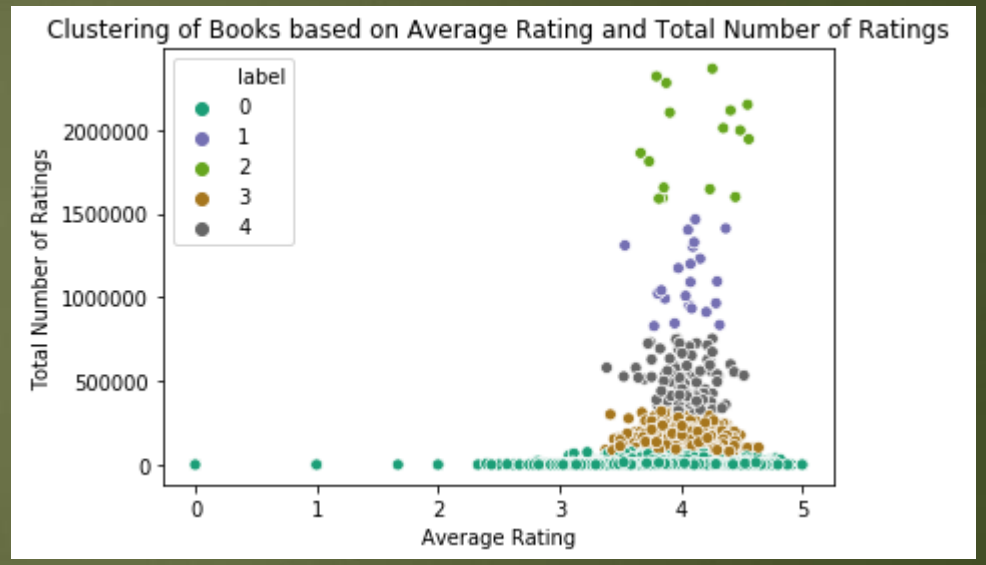
Positive correlation between total ratings and total text reviews



In-Depth Analysis

- In order to create labels, Kmeans clustering was performed.

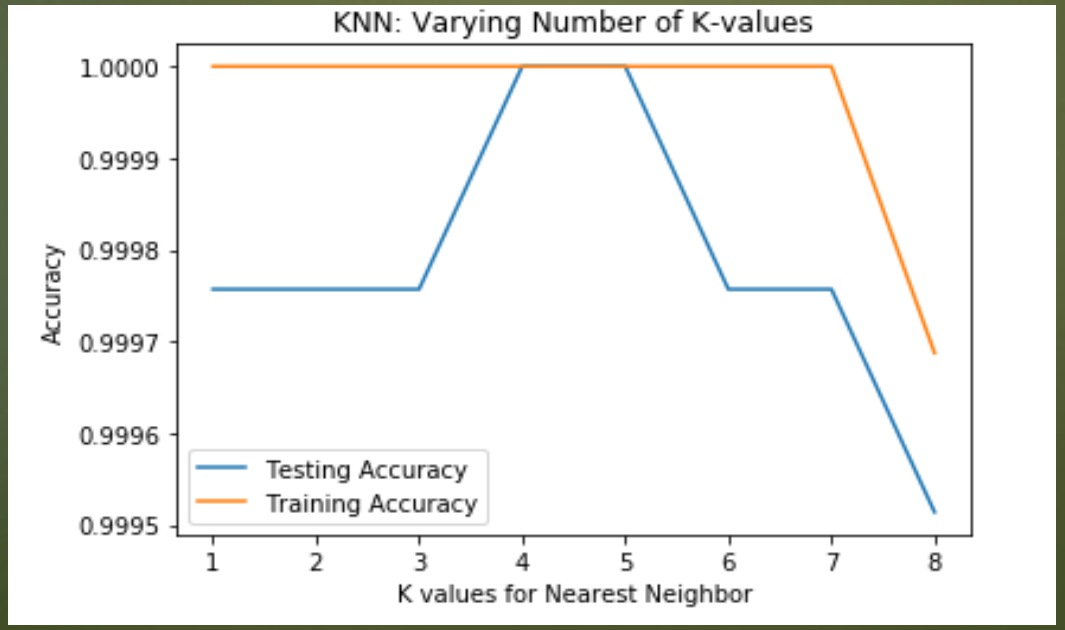




- From the bend in the elbow curve graph, it was determined that there were 5 clusters to the dataset.
- The data was then labeled into the 5 categories of popularity they fell into. Graph above shows grouping after outlier removal.

In-Depth Analysis

- After clustering confirmed organizational labeling based on relative popularity of a book, a recommendation system was created based on KNearestNeighbor Algorithm.
- After repeated testing, a neighbor value of 5 showed consistent training and testing accuracy.



Book Recommender

- Created function that takes in either the book title or bookID.
- If title or partial title not in dataset, will return error message.
- If title has single match in dataset, returns the top 5 closest books in comparison according to KNearestNeighbors.
- If title has multiple matches, returns possible titles with bookIDs for each so bookID can be used instead to complete search.

Book Recommender

- Partial title →

```
# Testing with partial title of full book
Books_Recommended("dark crystal")
```

```
The top 5 books similiar to "dark crystal" are (in order):
1. Tyler's Ultimate: Brilliant Simple Food to Make Any Time
2. Classical Drawing Atelier: A Contemporary Guide to Traditional Studio Practice
3. The Heritage of Shannara (Heritage of Shannara  #1-4)
4. Young Warriors: Stories of Strength
5. Silver Bullet
```

- Multiple titles →

```
# Testing with partial title of a book with multiple entries
Books_Recommended("harry potter and the half")
```

```
There are too many books with a similar title.
Please set ID based on Book ID below:
                                                 title  bookID
 Harry Potter and the Half-Blood Prince (Harry ...       1
 Harry Potter and the Half-Blood Prince (Harry ...    2005
```

- BookID →

```
# Testing with bookID as
Books_Recommended(ID=1)
```

```
The top 5 books similiar to " Harry Potter and the Half-Blood Prince (Harry ..." are (in order):
1. Harry Potter and the Order of the Phoenix (Harry Potter  #5)
2. The Fellowship of the Ring (The Lord of the Rings  #1)
3. Lord of the Flies
4. Romeo and Juliet
5. Animal Farm
```

Future Improvements

- **User Data.** More user data from individual accounts could help build a more collaborative recommendation system that tailors more specifically to the user, rather than general popularity of a title.

- **Authorship.** Further analysis of user history may also mean finding a way to incorporate authors into the recommendation system as well, since users may read multiple titles by the same author.

- **More updated data.** The dataset was collected in 2014 and more recent data could improve overall accuracy for this kind of recommendation system.

Thank You!

QUESTIONS?

Rae'e Yamin