# Goodreads Book Analysis & Collaborative Recommendation Engine

By Langhan Dee

## The Problem:

A good book can transport you into another world or give you deep insight into the one around you. Hundreds of thousands of books are published in the US each year, more than any one person can read. With so many books available, how can you find one you'll like (or better yet, love)? In my experience, the best recommendations come from people who know me well (friends) or people who know books well (professional booksellers). When there isn't time to visit a local independent bookstore, book reviews can be found online. One of the largest platforms for the book-loving community is GoodReads.com. Members can keep track of what they've read, leave reviews and ratings, and make lists of books they want to read. In this project, I will explore GoodReads user's ratings to examine trends and build several recommendation models to predict which books a user will love.

## The Client:

Clients who may be interested in this project's findings are readers and booksellers. With so many books out there and so little time, these recommendation models could help a reader narrow down their list to books they're likely to love. It would also help booksellers make suggestions to customers. Exploration of emerging topics and changing the popularity of genres could increase sales by giving publishers and booksellers insights into which books to promote or add on to a sale.

## The Data:

### I. Summary

The data used for this project originally comes from the GoodReads.com website. It was collected and made available by Zygmunt Zając on GitHub (https://github.com/zygmuntz/goodbooks-10k). The user and item data was downloaded as CSV files.

**books.csv -** There are 10,000 books in this dataset. Descriptions of the columns are below. It's noted if the column is missing entries.

Columns:

- **books_count**: the number of editions for a given work.
- **authors**: the authors(s) name.
- **original_publication_year**: (9979 non-null)
- **original_title**: (9415 non-null)
- **title**: similar to *original_title* and includes the order it appears in a books series.
- **language_code**: unclear if this is the language of the original edition or post popular edition. (8916 non-null, object)
- **average_rating**: the mean rating out of 1-5 stars.
- **ratings_count**: presumably, the number of unique users who have rated this book. It is lower than *work_ratings_count*.
- **work_ratings_count**: number of total ratings a book received, may include multiple ratings per user?
- **work_text_reviews_count**: number of written reviews, which is different than numerical ratings (number of stars).
- **ratings_1**: number of 1-star ratings.
- **ratings_2**: number of 2-star ratings.
- **ratings_3**: number of 3-star ratings.
- **ratings_4**: number of 4-star ratings.
- **ratings_5**: number of 5-star ratings.

**ratings.csv** - On a scale of 1-5, how did the user rate a book? The list contains 5,976,479 book-user pairs and 53,424 unique users.

Columns: user_id, book_id, rating (1, 2, 3, 4, or 5).

**to_read.csv** - Each entry represent a book that a user wants to read. There are 912,705 pair-wise entries.

Columns: user_id, book_id.

**tags.csv** - A list of user-created tags. These vary widely from genres to “16th-century” to “30-books-to-read-before-30”. There are 534,252 unique tag names.

Columns: tag_id, tag_name.

**book_tags.csv** - The number of times each tag was given to a specific book. The list contains 999,912 book-tag pairs.

Columns: goodreads_book_id, tag_id, count.

### II. Cleaning and Consolidating the Data

With 534,252 unique tags included, additional cleaning steps were needed to sift through and extract a meaningful subset that could be used as genre categories. GoodReads has a standard set of tags that can be used, however they aren't limited. Users can create custom tags, as well. Because of the open ended nature of this, there can be many variations of the same theme, ex. "owned", "books-i-own", "owned-books", "i-own" are all separate tags. This makes comprehensive analysis tricky. To compile a set of meaningful, common genres, we need to narrow down the tags list.

# Exploratory Data Analysis:

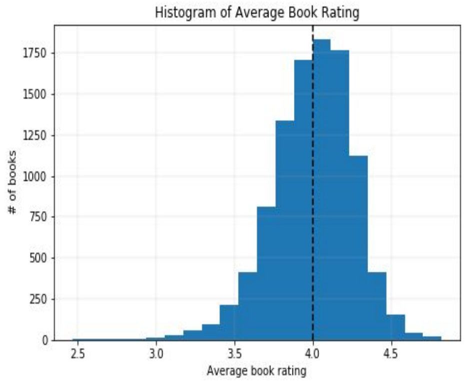
## Executive Summary:

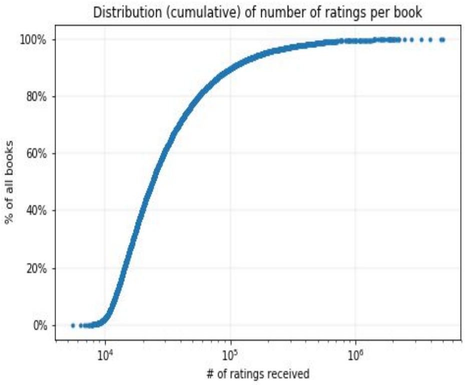
1. The average rating per book is **4.00219**.
2. Most books in this dataset (~90%) have been rated between 10,000 and 100,000 times, with a **mean of 59,687 ratings per book**.
3. Overall, the numerical metadata we have in this dataset (such as the date a book was published, number of ratings, or the number of editions) doesn't tell us much about how highly it will be rated. This implies there are other factors, such as the quality of writing, that have more impact on ratings.
4. Of the numeric variables, the presence of 2-star ratings (even more than the presence of 1-star ratings) seem to be an indicator that a book will have a lower overall rating.
5. **Young adult series are popular** -- think The Hunger Games, Harry Potter, Twilight, Divergent, and the Hobbit. Specifically, the first book in this type of series receive the most ratings.
6. The first book of the **Twilight** series was rated by nearly 400,000 GoodReads users. It's the 3rd most rated book on this dataset. However, you have to go all the way down the list to the 46th most rated book ("Eat, Pray, Love") before find one that's rated lower than Twilight's 3.57 average. This prompts the question, why have so many people read it if it wasn't liked by the majority of them?
7. **Calvin and Hobbes** and **Harry Potter** top the charts for most beloved books.
8. Looking at books with a high variance in ratings, it isn't surprising to find that **religion and sexuality are controversial subjects.** The list of the top 15 books we don't agree on include **Fifty Shades of Grey, Twilight, The Book of Mormon, The Qur'an,** and **The Holy Bible.**
9. **When ratings were averaged per user, only 20% of users had an average less than 3.5 stars.** This could suggest a couple things: Goodreads users like reading and have a positive bias towards books in general; or people are more likely to take the time to rate something they like.

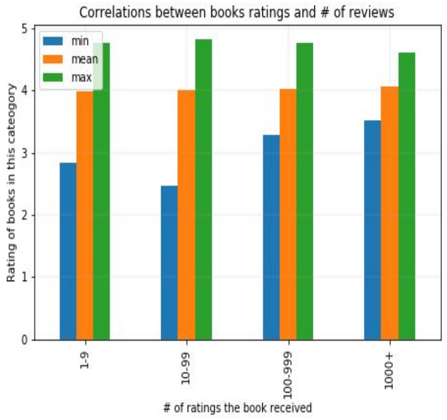
10. **80% of users only vary their ratings by 1 point.** For example, a user might always give books 4-star or 5-star ratings. Another might only give out 1-star ratings.
11. Some people only give out 5-star ratings, some only give 1-star ratings, but **no user in this data set gave only 2-star ratings.**
12. The top 3 most used tags are "to-read", "favorites", and "owned".
13. The most popular genres by volume are **Fantasy, Romance, Mystery and Young Adult**, each were tagged on 3,600-4,300 books (recall that there are 10,000 books total in the dataset).
14. Genres that have a positive correlation with book ratings are: 'childrens' 'christian' 'fantasy' 'graphic-novels' 'paranormal'. On the other end, 'chick-lit' and 'novels' tend to receive lower ratings.

## 1. Distribution of Ratings (1-5 Stars)

- The distribution of average rating per book is left-skewed with a mean of **4.00219**.
- Most books in this dataset (~90%) have been rated between 10,000 and 100,000 times, with a **mean of 59,687 ratings per book**.
- When a book is reviewed more often, the min, max, and mean ratings converge to the mean, 4.0012. Which is to be expected, statistically.
- The min and max ratings had the largest spread for books that were rated 10-99 times.







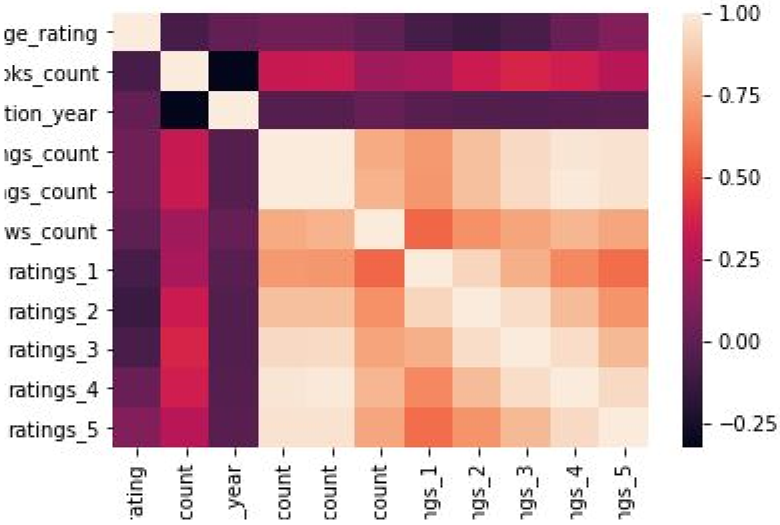
## 2. Correlation Among Numerical Features

None of the numeric features in this dataset have a strong correlation with mean book rating.¶

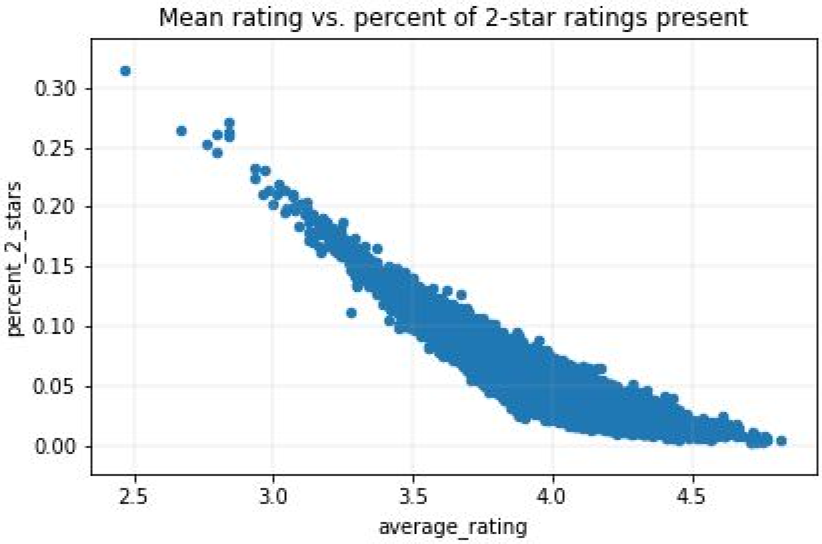
- A user is more likely to take the time to write a review if they're rating the book with 4 stars. They're least likely to leave a review with a 1-star rating.
- **The strongest indicator of a low average rating is the presence of 2-star ratings.** Mathematically, we might expect that a 1-star rating would bring the average down more.

However, 2-star ratings are less common than 1-star ratings, which might make them more significant. With a 0.116 correlation, it's still a fairly weak predictor.

- Unsurprisingly, 5-star ratings have the next strongest correlation to average rating.
- The older the book, the more editions it will have.
- Having more editions of a book has a slightly negative correlation with average rating.



- We saw in the correlation matrix that a book's average rating is negatively correlated with the number of 2-star ratings it received. 2-stars is the least common rating, which may make them a telling measure of how good a book is.

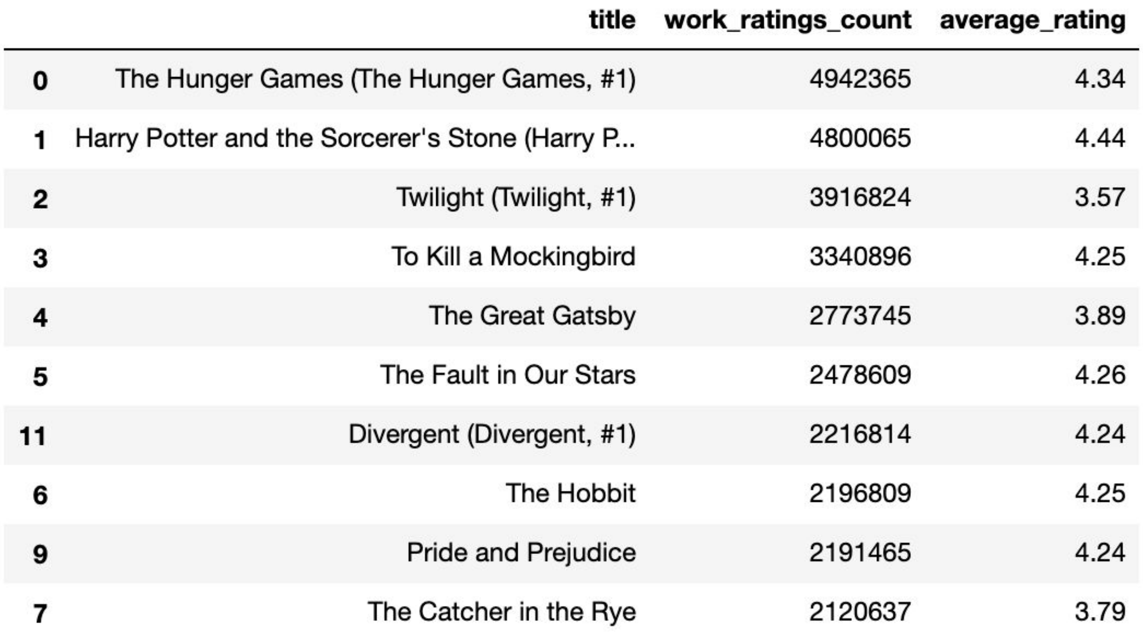


## 3. Most commonly read books: Young adult fantasy series

- The books with the most ratings (and we assume, the most widely read) are the first book in young adult series (The Hunger Games, Harry Potter, Twilight, Divergent, and the Hobbit).

- Books assigned in school. Many books on this list are also banned books or frequently challenged books: Harry Potter, To Kill a Mockingbird, The Great Gatsby, The Catcher in the Rye.

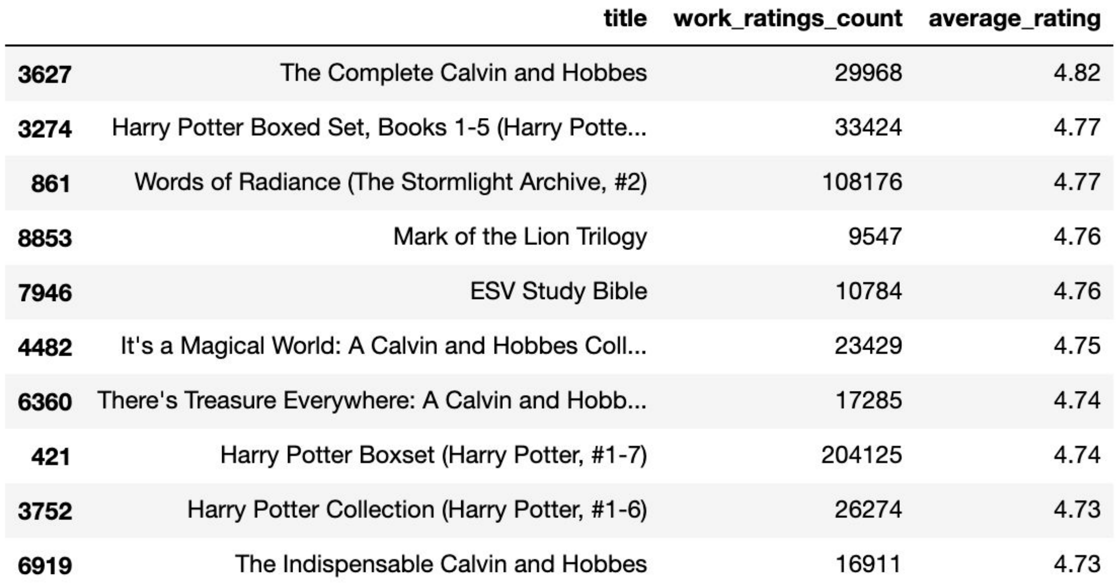
- **A lot of people have read Twilight. Most didn't think it was very good.** Vampires, werewolves, and moody teens have been popular in recent years and the first book of the Twilight series was rated by nearly 400,000 GoodReads users. It's the 3rd most rated book on this dataset. However, you have to go all the way down the list to the 46th most rated book ("Eat, Pray, Love") before find one that's rated lower than Twilight's 3.57 average.



| | title | work_ratings_count | average_rating |
|---|---|---|---|
| **0** | The Hunger Games (The Hunger Games, #1) | 4942365 | 4.34 |
| **1** | Harry Potter and the Sorcerer's Stone (Harry P... | 4800065 | 4.44 |
| **2** | Twilight (Twilight, #1) | 3916824 | 3.57 |
| **3** | To Kill a Mockingbird | 3340896 | 4.25 |
| **4** | The Great Gatsby | 2773745 | 3.89 |
| **5** | The Fault in Our Stars | 2478609 | 4.26 |
| **11** | Divergent (Divergent, #1) | 2216814 | 4.24 |
| **6** | The Hobbit | 2196809 | 4.25 |
| **9** | Pride and Prejudice | 2191465 | 4.24 |
| **7** | The Catcher in the Rye | 2120637 | 3.79 |

## 4. Highest rated books

- **Calvin and Hobbes and Harry Potter** top the charts for most beloved books. Although they have very different story lines (one is a humorous comic strip and the other is an epic fantasy adventure story), the protagonist of both books series is a young boy. Many people started reading these books as children and continued to enjoy them as adults. The ageless appeal of these works seem to translate into chart-topping ratings.



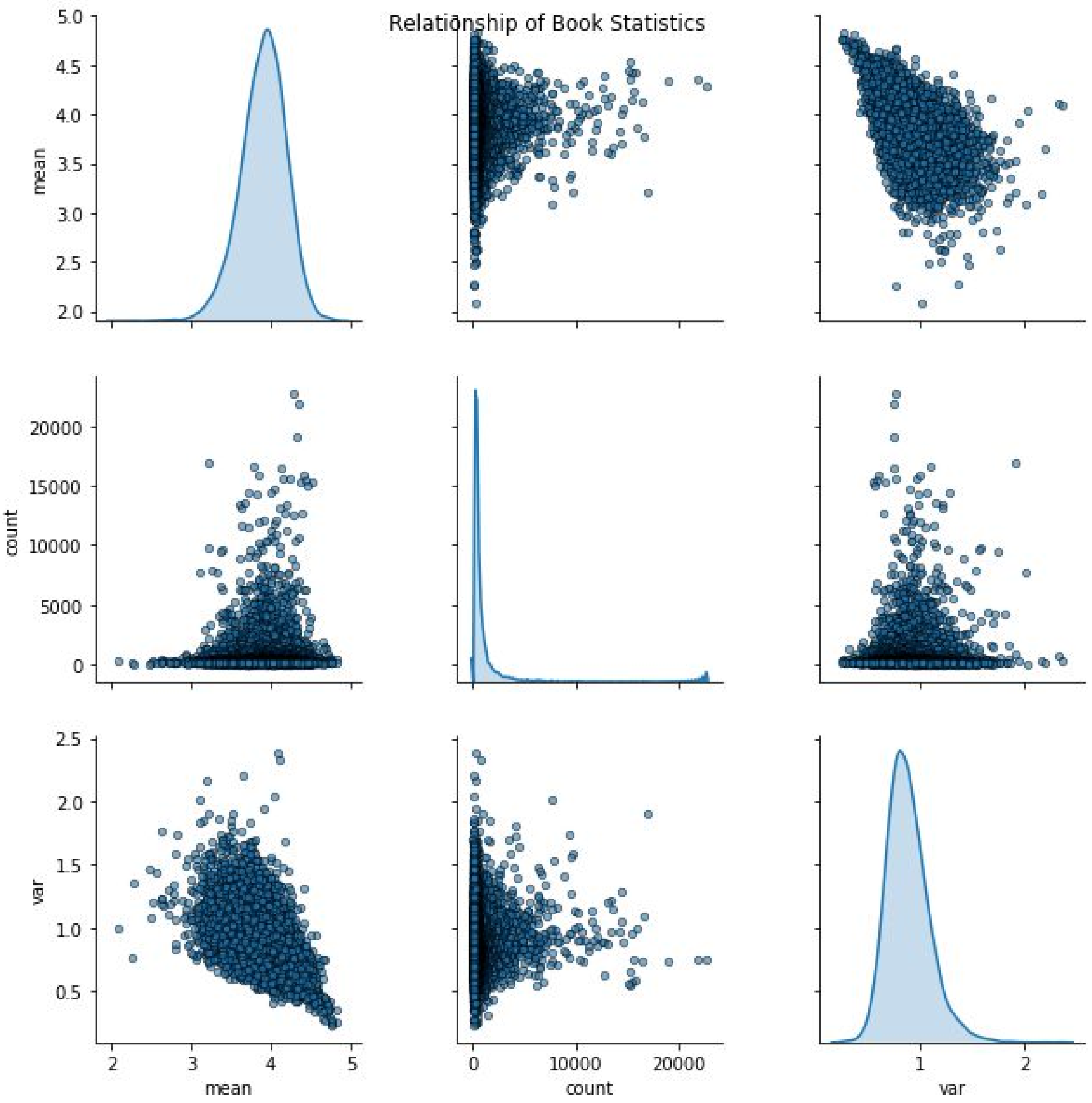
| | title | work_ratings_count | average_rating |
|---|---|---|---|
| 3627 | The Complete Calvin and Hobbes | 29968 | 4.82 |
| 3274 | Harry Potter Boxed Set, Books 1-5 (Harry Potte... | 33424 | 4.77 |
| 861 | Words of Radiance (The Stormlight Archive, #2) | 108176 | 4.77 |
| 8853 | Mark of the Lion Trilogy | 9547 | 4.76 |
| 7946 | ESV Study Bible | 10784 | 4.76 |
| 4482 | It's a Magical World: A Calvin and Hobbes Coll... | 23429 | 4.75 |
| 6360 | There's Treasure Everywhere: A Calvin and Hobb... | 17285 | 4.74 |
| 421 | Harry Potter Boxset (Harry Potter, #1-7) | 204125 | 4.74 |
| 3752 | Harry Potter Collection (Harry Potter, #1-6) | 26274 | 4.73 |
| 6919 | The Indispensable Calvin and Hobbes | 16911 | 4.73 |

- **Highest rate books after Calvin and Hobbes and Harry Potter: Christian Literature and Fantasy Novels¶**
  - Words of Radiance = epic fantasy novel
  - Mark of the Lion Trilogy = historical fiction with a christian heroine
  - A Court of Mist and Fury = fantasy romance
  - The Revenge of the Baby-Sat = another Calvin and Hobbes compilation
  - The Absolute Sandman = fantasy
  - The Way of Kings, Part 1 = fantasy

## 5. Book Ratings: Mean, Count, and Variance

- **Variance vs. mean** - It makes sense that books with a mean rating close to the maximum of 5-stars will always have a low variance. Mathematically, the only way it will achieve a high mean is if most of it's ratings are around 5-stars.

- **The most compelling or thought-provoking books are likely to have a high variance as well as a relatively good mean rating.**

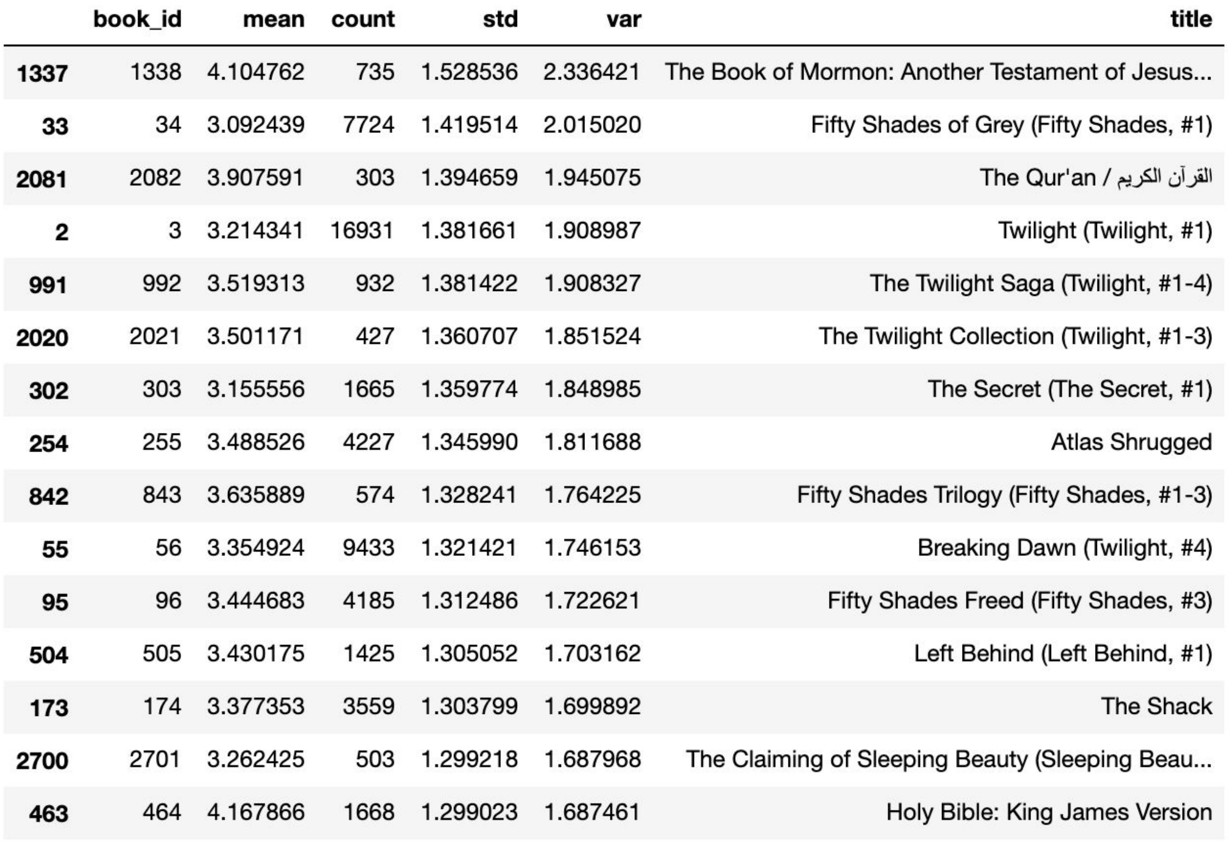


## 6. Religion and Sexuality are controversial

The most polarizing ideas tend to have the strongest followings, because they evoke powerful emotions, whether of passion or hatred. When building book recommendations, these highly contested books may be the most helpful indicators for future suggestions. And what's divided people more over the centuries than religion?

Below is a list of 15 books that a significant number of people have rated (count > 300) with wide spread of high and low ratings (VAR > 1.5).

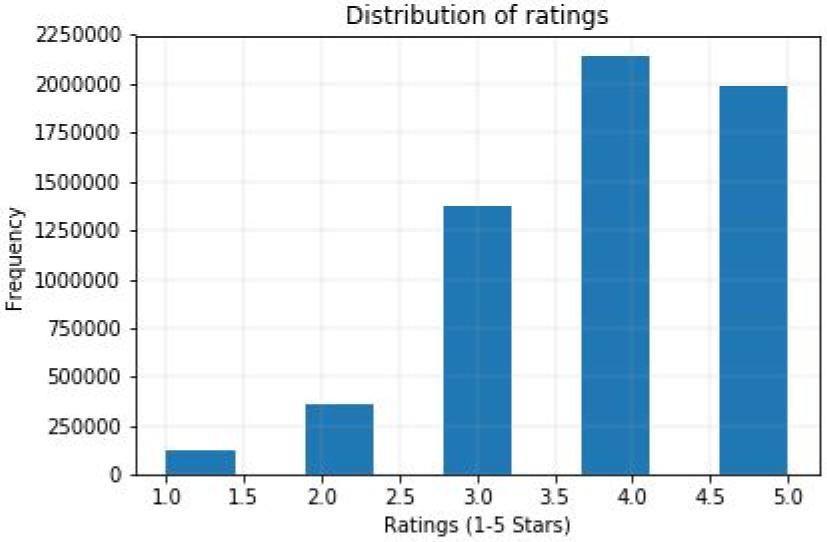
- ***The Book of Mormon***, ***The Qur'an***, and ***The Holy Bible*** are all in the top 15 most controversial books in our dataset. Though not holy texts, most other books on the list have a foundation in religion.
- ***The Secret*** is a self help book teaching positive-thinking to directly affect one's life.
- ***Atlas Shruged*** is a novel with controversial commentary on government and theology. Ayn Rand's other novel, ***Fountainhead*** also made the list when it was narrowed down to books that were rated at least 1000 times.
- ***The Shack*** is a Christian based novel with an unconventional interpretations of God and the Holy Trinity.
- ***Left Behind*** is a novel that takes place during the Rapture and the weeks that follow.
- The ***Twilight*** series is a love saga about vampires and werewolves. On the surface that doesn't seem controversial, however it's themes include prejudice, Mormonism, teen sex vs. abstinence, and feminism. The book is both widely loved and widely hated.
- Last to be discussed, but definitely not least on the list is the ***Fifty Shades of Grey*** series. This widely-read erotic romance trilogy has been accused of romanticizing abusive relationships and misrepresenting BDSM practices.



| | book_id | mean | count | std | var | title |
|---|---|---|---|---|---|---|
| **1337** | 1338 | 4.104762 | 735 | 1.528536 | 2.336421 | The Book of Mormon: Another Testament of Jesus... |
| **33** | 34 | 3.092439 | 7724 | 1.419514 | 2.015020 | Fifty Shades of Grey (Fifty Shades, #1) |
| **2081** | 2082 | 3.907591 | 303 | 1.394659 | 1.945075 | The Qur'an / القرآن الكريم |
| **2** | 3 | 3.214341 | 16931 | 1.381661 | 1.908987 | Twilight (Twilight, #1) |
| **991** | 992 | 3.519313 | 932 | 1.381422 | 1.908327 | The Twilight Saga (Twilight, #1-4) |
| **2020** | 2021 | 3.501171 | 427 | 1.360707 | 1.851524 | The Twilight Collection (Twilight, #1-3) |
| **302** | 303 | 3.155556 | 1665 | 1.359774 | 1.848985 | The Secret (The Secret, #1) |
| **254** | 255 | 3.488526 | 4227 | 1.345990 | 1.811688 | Atlas Shrugged |
| **842** | 843 | 3.635889 | 574 | 1.328241 | 1.764225 | Fifty Shades Trilogy (Fifty Shades, #1-3) |
| **55** | 56 | 3.354924 | 9433 | 1.321421 | 1.746153 | Breaking Dawn (Twilight, #4) |
| **95** | 96 | 3.444683 | 4185 | 1.312486 | 1.722621 | Fifty Shades Freed (Fifty Shades, #3) |
| **504** | 505 | 3.430175 | 1425 | 1.305052 | 1.703162 | Left Behind (Left Behind, #1) |
| **173** | 174 | 3.377353 | 3559 | 1.303799 | 1.699892 | The Shack |
| **2700** | 2701 | 3.262425 | 503 | 1.299218 | 1.687968 | The Claiming of Sleeping Beauty (Sleeping Beau... |
| **463** | 464 | 4.167866 | 1668 | 1.299023 | 1.687461 | Holy Bible: King James Version |

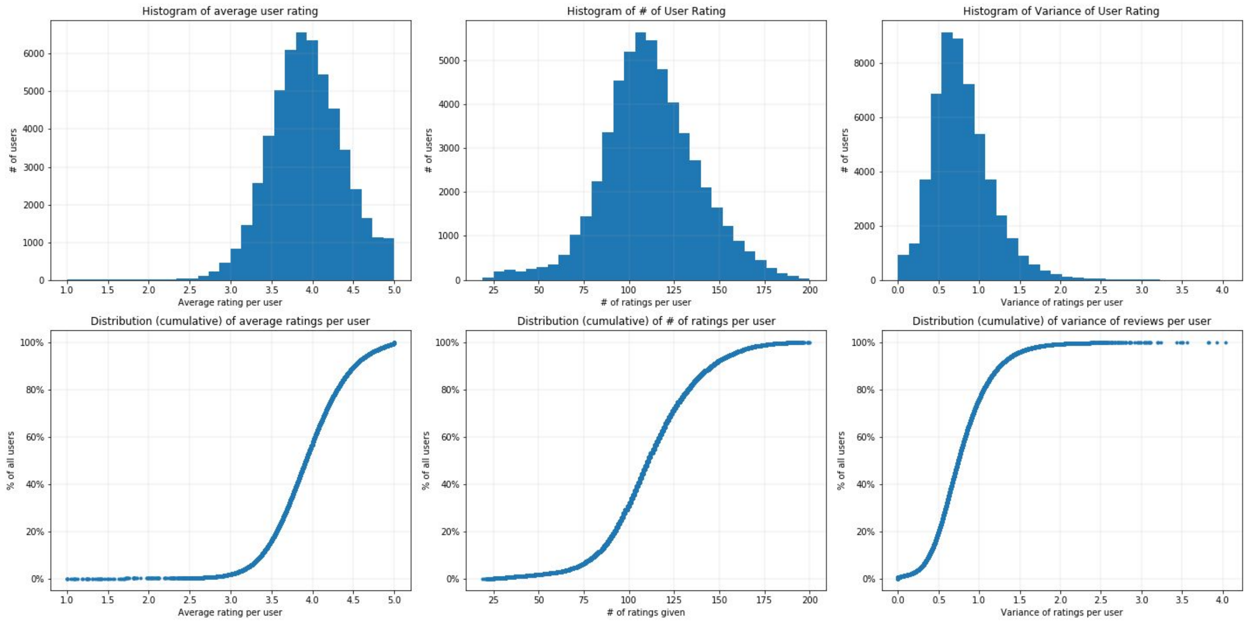
## 7. Distribution of Raw Ratings

- A huge portion of the ratings are 4- and 5-stars.
- The mean rating overall (not mean rating per book) is 3.9199



## 8. User rating stats - They generally like what they read

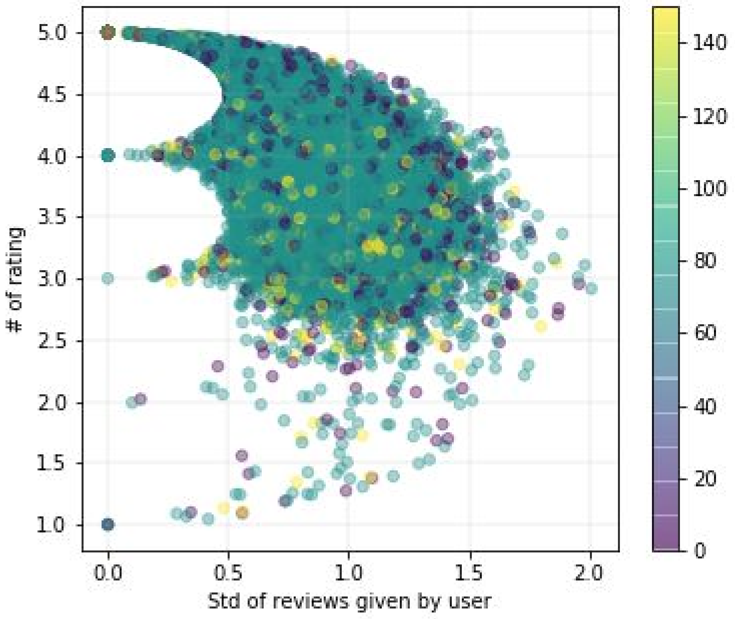
- **Only 20% of users have an average rating that's less than 3.5 stars.** This could suggest a couple things: Goodreads users like reading and have a positive bias towards books in general; people are more likely to take the time to rate something they like; or
- **80% of users only vary their ratings by 1 stars.** For example, a user might always give books 4-star or 5-star ratings. Another might only give out 2-star or 3-star ratings.



## 9. Do low raters vary their ratings more than high raters?

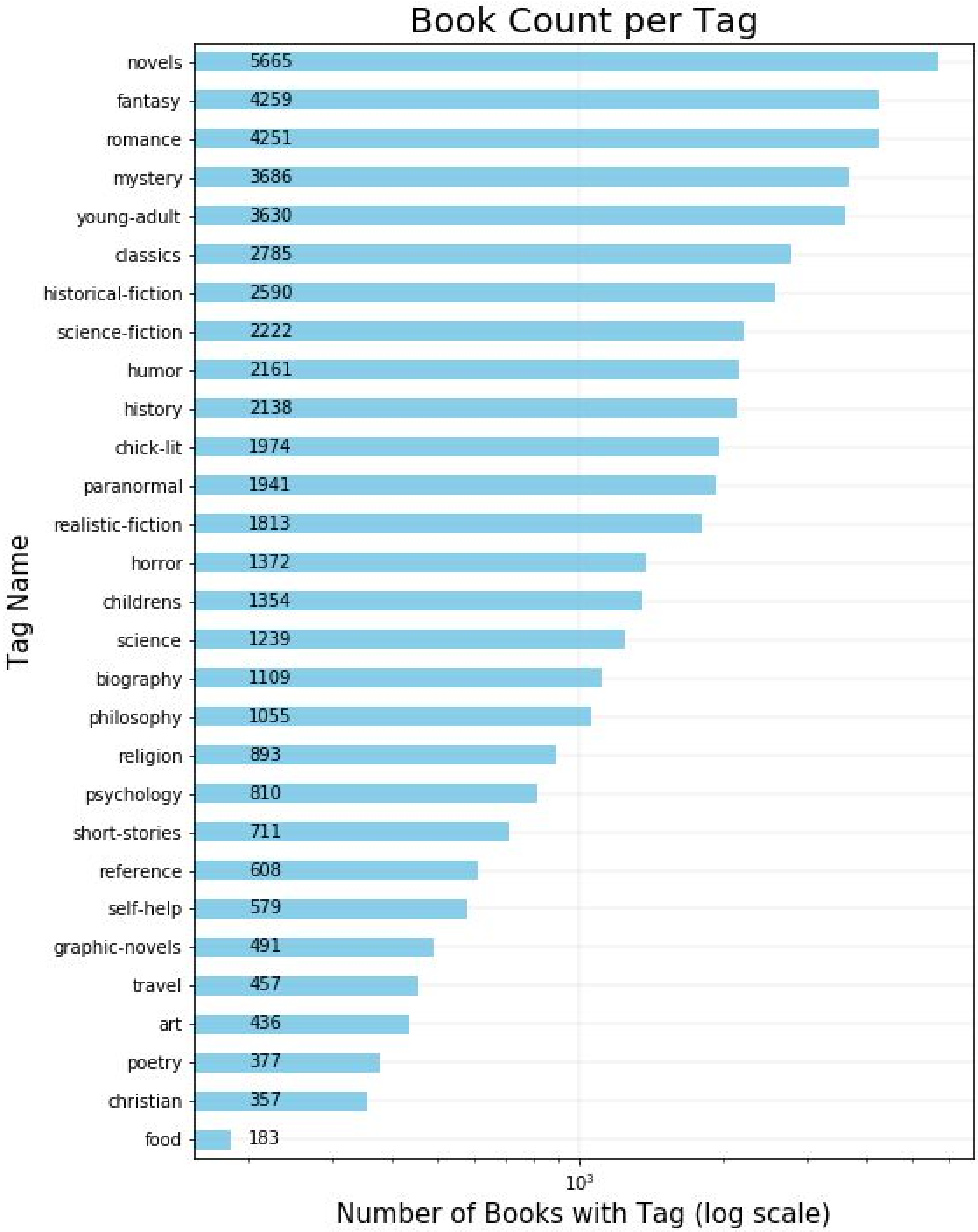
- Again, we see that most users give an average rating of 3.5-5 stars, and don't vary their ratings by much more than 1 point.
- The interesting areas are along the y-axis where variance = 0. These are users who rate all books the same way. There's a high concentration of people who give every book 4-stars or every book 5-stars.
- The group who only ever give out 1-star is smaller. It's rare for a user to give all books 3-stars, and there isn't a single user in our sample who only gives 2-star ratings.
- The pattern of how high & how varied ratings are doesn't seem to be affected by how many books a user has rated.

Correlation between std and mean ratings(color = # of ratings)



## 10. Most commonly used tags and genres

- The top 3 most used tags are "to-read", "favorites", and "owned".
- Fantasy, Romance, Mystery and Young Adult genres were each tagged on 3,600-4,300 books.
- Genres are rated differently:
  - Slightly poorer performing genres: 'chick-lit' 'novels'
  - Slightly better performing genres: 'childrens' 'christian' 'fantasy' 'graphic-novels' 'paranormal'



# Machine Learning Process:

Two types of recommendation engines were built. The first is a **Simple Genre Recommender** and a **Collaborative Filtering Recommender**.

## I. Simple Recommendation Engine by Genre

When finding a book for a friend, the first question is: "What kind of books do you like to read?" With their favorite genre in mind, I can start the search for my favorite kind of gifts to give...books!

That's the premise of this Simple Book Recommender which filters the books by the specified genre and returns the top N books with the highest weighted average rating.

### Feature Engineering

Genre features had to be extracted from the extensive tag dataset. I started with the most commonly suggested genres on the Goodreads website. Several more genres were added from my own subject matter knowledge. A simple algorithm was run in order to reduce the genre list as much as possible while making sure that each book was tagged by at least one of the remaining genres. Each book was given a percentage rating depending on how many times it was tagged per genre.

### Example Recommendations

| The top 5 rated biography books are: | | | |
|---|---|---|---|
| title \ | authors | biography % | wt_avg_rating |
| 143 Unbroken: A World War II Story of Survival, Re... | Laura Hillenbrand | 0.451054 | 4.359635 |
| 982 Between the World and Me | Ta-Nehisi Coates | 0.45961 | 4.254002 |
| 851 When Breath Becomes Air | Paul Kalanithi, Abraham Verghese | 0.595614 | 4.235827 |
| 3884 Born a Crime: Stories From a South African Chi... | Trevor Noah | 0.610811 | 4.232601 |
| 690 On Writing: A Memoir of the Craft | Stephen King | 0.531602 | 4.223177 |

## II. Collaborative Filtering Recommendation Engine

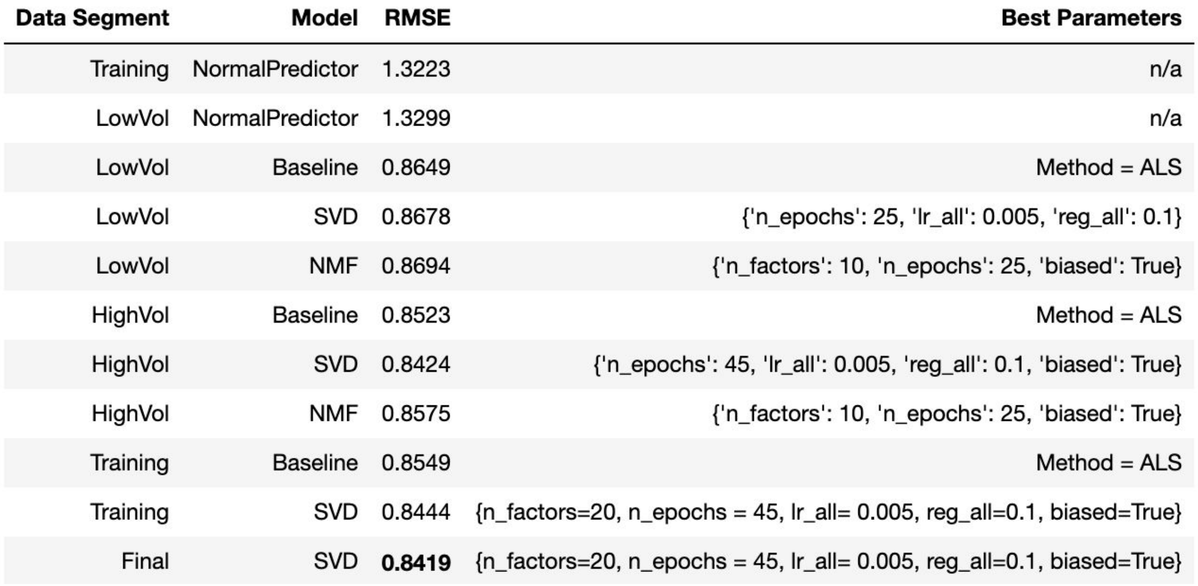
Who better to ask for a recommendation than from someone who has the same taste in books? Collaborative filtering engines use this idea to create personalized recommendations. These models compare the similarity in users ratings in order to predict other books they'll like.

### Model Evaluation

A collaborative filtering recommender module was built with user (rows) by book (columns) rating matrix. The dataset has a total of 53,424 users, 10,000 books, and 5,976,479 ratings. The user-by-book matrix is very sparse (1.12% non-empty). Therefore, matrix factorization using ALS (alternative least square) was used to complete the matrix and generate recommendations.

Several algorithms were tested: **BaselineOnly, SVD** (the famous Netflix prize winner), and **NMF** (Non-negative Matrix Factorization). The **NormalPredictor** was also used to give random predictions to set a baseline. The best score for each algorithm was compiled after tuning parameters.

**Metrics chosen for evaluating and optimizing the ‘goodness’ of the algorithms:** a) measure prediction accuracy: RMSE(root mean squared error), and MAE (Mean Absolute Error).



| Data Segment | Model | RMSE | Best Parameters |
|---|---|---|---|
| Training | NormalPredictor | 1.3223 | n/a |
| LowVol | NormalPredictor | 1.3299 | n/a |
| LowVol | Baseline | 0.8649 | Method = ALS |
| LowVol | SVD | 0.8678 | {'n_epochs': 25, 'lr_all': 0.005, 'reg_all': 0.1} |
| LowVol | NMF | 0.8694 | {'n_factors': 10, 'n_epochs': 25, 'biased': True} |
| HighVol | Baseline | 0.8523 | Method = ALS |
| HighVol | SVD | 0.8424 | {'n_epochs': 45, 'lr_all': 0.005, 'reg_all': 0.1, 'biased': True} |
| HighVol | NMF | 0.8575 | {'n_factors': 10, 'n_epochs': 25, 'biased': True} |
| Training | Baseline | 0.8549 | Method = ALS |
| Training | SVD | 0.8444 | {n_factors=20, n_epochs = 45, lr_all= 0.005, reg_all=0.1, biased=True} |
| Final | SVD | **0.8419** | {n_factors=20, n_epochs = 45, lr_all= 0.005, reg_all=0.1, biased=True} |

- The **BaselineOnly** algorithm performed best on **low volume users** with RMSE = 0.8649.
- The **SVD** algorithm performed best on **high volume users** with RMSE = 0.8424.

## Final Predictions

A validation set (10% of the total ratings) was held out while testing and tuning the data. Finally, the SVD model was fit to the 90% of the data previously used for training and testing the various algorithms. **Recommendation predictions made on validation set had a root mean squared error of 0.8419.** This was less error than seen with previous testing on smaller segments of the data. It appears that having more a robust set of users to collaborate was beneficial.

## Example Recommendations

Examining some examples of the top 10 book recommendations per user, we see books that fall within fairly specific genres, along with a few books that branch out into others. The recommendations tell an interesting story when seen together, and appear to be highly tailored for the user.

For UserID 12345: Title - Author - Estimated Rating

1. [("Ender's Game (Ender's Saga, #1)", 'Orson Scott Card', 4.049455231926951),
2. ('Wintersmith (Discworld, #35; Tiffany Aching, #3)', 'Terry Pratchett', 3.9644772940049595),
3. ('His Dark Materials (His Dark Materials #1-3)','Philip Pullman', 3.9406634522650252),
4. ('John Adams', 'David McCullough', 3.918721387055269),
5. ("Ptolemy's Gate (Bartimaeus, #3)", 'Jonathan Stroud', 3.8920036141586136),
6. ('Nation', 'Terry Pratchett', 3.7975646548868154),
7. ('High Fidelity', 'Nick Hornby', 3.7596392267362155),
8. ('Cod: A Biography of the Fish that Changed the World', 'Mark Kurlansky', 3.7376651915868053),
9. ('One Hundred Years of Solitude', 'Gabriel García Márquez, Gregory Rabassa', 3.713380647258924),
10. ('The Lost City of Z: A Tale of Deadly Obsession in the Amazon', 'David Grann', 3.572136295499298)]

## Future Considerations

The recommendation engine could be expanded into a hybrid model that incorporates genre, keyword, whether the book has won awards, and information about the authors. It could be improved further by talking with booksellers (subject matter experts) about their process for getting the right books into the right reader's hands.

With additional input and development, I intend to build an interactive book recommendation engine app that booksellers and individuals can use to find new, beloved books that go beyond what they may have thought of on their own.

## Links to Code

Exploratory Data Analysis:
https://github.com/lwdee/Springboard/blob/master/Book_Recommendations/Book_Recommender_EDA.ipynb

Collaborative filtering recommender:
https://github.com/lwdee/Springboard/blob/master/Book_Recommendations/Book_Recommender_Collaborative.ipynb

Simple Recommender by Genre:
https://github.com/lwdee/Springboard/blob/master/Book_Recommendations/Simple_Recommender_with_Genres.ipynb

## Acknowledgements

Many thanks go to the staff at Springboard, especially to my mentor Max Sop.