Can GoodReads Reader Ratings Predict an Author's Future Success?

Jill E. Andersen
Western Governors University

Table of Contents

Project Overview 3
A. Project Highlights 3
A1. Research Question 3
A2. Project Scope 3
A3a. Solution Overview Tools 3
A3b. Solution Overview Methodologies: 3
Project Plan 3
B. Project Execution 3
B1. Project Plan 3
B2. Project Planning Methodology 3
B3. Project Timeline and milestones 4
Methodology 5
C. Data Collection Process 5
C1. Advantages and Limitations of Data Set 5
D. Data Extraction and Preparation Processes 5
E. Data Analysis Process 5
E1. Data Analysis Methods 5
E2. Advantages and Limitations of Tools/Techniques 6
E3. Application of Analytical Methods 6
Results 6
F. Project Success 6
F1. Statistical Significance 6
F2. Practical Significance 6
F3. Overall Success 6
G. Key Takeaways 7
G1. Summary of Conclusions 7
G2. Effective Storytelling 7
G3. Findings-based Recommendations 7
H. Panopto Presentation 7
Appendices 7
I. Evidence of Completion 7
Sources 7

Project Overview

A. Project Highlights

A1. Research Question

Can Goodreads Reader Ratings Predict Author's Future Success? Will an author's past books with high average ratings increase the likelihood that future books will also receive high average ratings?

A2. Project Scope

The solution for this project will include a R application to analyze the Goodread csv file to determine if there are factors in the given dataset that have a high correlation to the average ratings given by readers. Specifically, if the author is a good determinate for a publisher or an author who self-publishes to use to decide to publish additional books.

A3a. Solution Overview Tools

RStudio was used to create an R markdown file with R scripts to run the actual code on the "GoodReads.csv" file (see Appendices).

A3b. Solution Overview Methodologies:

The data analytical method used was descriptive, using a univariate analysis of average ratings against author, ratings count and publisher.

Project Plan

B. Project Execution

B1. Project Plan

During additional study of the data, it became apparent that completing Goal 1 would not be possible. (see Data Understanding under B2)

Goal 1:
This goal is to analyze if book titles by authors who have at least one book title with an average rating of 5 also have additional books who have achieved an average rating between 4 to 5.

B2. Project Planning Methodology

I used CRISP-DM to manage this project. This project management process includes six steps: 1. Business Understanding, 2. Data Understanding, 3. Data Preparation, 4. Modeling, 5. Evaluation, 6. Deployment.

Business Understanding: Large publishing companies to small publishing companies to self-publishers invest time, number of employee & money into any new book. Therefore, having any additional important factors to determine whether publishing a new book is worth the investment.

Data Understanding: It was originally thought that it would be appropriate to create subcategories for each author who had more than one book title on the list. Of these

authors I would analyze how often an author who had a book title that had achieved a rating of 5, had also written a subsequent book with a rating of at least 4.

Problems Encountered:

1. I discovered this problem when I sorted the dataset by the average rating (avgRating). The only books that scored a 5 average rating only had 1 to 2 reviews, (ratingsCount). I had planned to remove from the analysis any book titles that had less than 100 reviews. The average rating with a minimum of 100 reviews was 4.82. (See Addendum 2a)

2. When I sorted the dataset by author, I noticed that many authors who had more than one title had authored some titles alone, and other tittle were co-authored. Of the titles co-authored it was not always the same co-author who was listed as an author. This created a need for additional subcategories, which left the question how would these titles be grouped? Would I group them in as all having been written by the original author or would each combination of author and co-authors count as another grouping? It would be impossible to determine which author or co-author was the cause for subsequent titles average rating to increase or decrease. I don't believe either grouping option would be a true representation of the facts and therefore, could not be used to determine if the hypothesis as being true or false. (See Addendum 3 GoodReads.csv)

Data Preparation: I formatted the datatype from 'general type' to 'date type' in the date columns. I will exclude any data record that does not include, reader rating, published by date or author. I also excluded any title that less than 100 reviews (ratingsCount).

Modeling: I created one scatterplot to visualize the correlation between averageRating and ratingsCount. I created two histogram charts. One to display the number of books by averageRating and the second chart to show the number of ratingsCount. (See Addendum 1, the R markdown file)

Evaluation: (See Addendum 1, the R markdown file)

Deployment: I was not able to perform analysis using the author, therefore it cannot be deployed. It was not found that reader ratings are significantly affected by the number of reader reviews (ratingsCount). It is not to recommended this dataset be implementation for use in making future business decisions.

B3. Project Timeline and milestones

Milestone	Projected Duration	Actual Duration
Clean data for appropriate data types in each column & appropriate data ranges in each row.	1 day	2 hrs.

Code application to analyze correlation factor & produce graphs	2-3 days	1 day
Write up analysis report on findings. Include whether or not recommended for implementation.	2-3 days	1 hr.

Methodology

C. Data Collection Process

Discuss these elements; offer examples.

- The data collection was already completed during Task2. It was a prepared csv file and did not require any additional data collection beyond downloading the csv file to my computer.
- Obstacles to data collection: There were no obstacles encountered during data collection.
- Unplanned data governance handling: This was a public dataset and did not include any confidential or proprietary information.

C1. Advantages and Limitations of Data Set

The advantage to this data set was that it was already in Second Normal Form. Each record is unique, and each cell contains a single value. (Although, I would argue that co-author should be its own column rather than combined with author). The bookID column functions as the primary key.

The limitation of this data set is that there is only data column that can easily compared to averageRating, ratingsCount. Since ratingsCount had a low correlation .038 (See Addendum 1) to averageRating. This dataset did not provide any useful info for analysis to future business decision for publishing houses. If a data set also include book sales, genre and book format (e-reader, paper copy, cd or audiobook) there would be additional data to determine what factors affect the average rating.

D. Data Extraction and Preparation Processes

Microsoft Excel was used to read the csv file. I formatted the data has a table to allow me to easily sort and filter data.

E. Data Analysis Process

E1. Data Analysis Methods

The analytic method used was inferential statistics. This was the most appropriate method because this is a significantly large, dataset (8,403 records). The analysis of this dataset can be used to assume this dataset is significantly relevant to the population as a whole.

E2. Advantages and Limitations of Tools/Techniques

The two tools used were Microsoft Excel and R. The data was already provided in an Excel compatible format. Excel was the easiest to quickly sort and filter data for the initial data understanding.

R's simple code is robust enough to create graphs and run correlation tests. It also is capable of creating a markdown file to include both graphs and evaluation of the statistical analysis performed in one file, which outputs to Word.

E3. Application of Analytical Methods

The analytical methods tools were executed in the following steps:

- Formatted all data columns in GoodReads.csv to the appropriate datatype represented in the column.
- Formatted the data as a table with headers.
- Filtered the data by ratingsCount and removed any title that 99 or fewer listed.
- Sorted the data by ratingsCount, highest to lowest. Captured a screenshot of the top 20 titles listed. (See addendum 2c)
- Sorted the data by averageRating, highest to lowest. Captured a screenshot of the top 20 titles listed. (See addendum 2a)
- Sorted the data by averageRating, lowest to highest. Captured a screenshot of the top 20 titles listed. (See addendum 2b)
- Uploaded GoodReads.csv file into RStudio.
- Created RMarkdown file. Created charts and ran analysis as listed in B2 under DataModeling.

Results

F. Project Success

F1. Statistical Significance

A thorough evaluation of the statistical significance of the analysis is provided in the R Markdown file (See Addendum 1)

F2. Practical Significance

There was no data in this dataset that was relevant in determining why a reader liked or disliked a book title. Neither is there any determining information as to why readers chose to rate the book title or chose not to rate it.

F3. Overall Success

The project was a partial success, in that the tools used provided adequate analysis methods for averageRatings and ratingsCount. The project was not successful in being able to analysis what affect the Author had on the averageRatings or on the ratingsCount.

G. Key Takeaways

G1. Summary of Conclusions

Assuming that the rating a reader gives a title is an indicator of book sales, there was no statistically significant data in this dataset that would allow us to draw the conclusion that it would impact book sales.

G2. Effective Storytelling

The histogram of the averageRatings and the histogram of the ratingsCount provided a breakdown of the confidence intervals for book titles studied by GoodReads. This provides a good narrative that most titles received between 100 to 5,000 ratings (ratingsCount). The book title with the most reviews had 4,597,666 reviews. Most book titles received an average rating (averageRating) between 3.5 to 4.5. The highest rated title with a minimum of 100 reviews received a rating of 4.82.

G3. Findings-based Recommendations

My recommendation is not to continue to use this dataset or newer such datasets from GoodReads. My recommendation is to find or create a data set that included additional data points such as book sales, genre and book format (e-reader, paper copy, cd or audiobook) there would be additional) to determine what factors do affect the average rating.

H. Panopto Presentation

Appendices

I. Evidence of Completion

- Addendum 1 - R Markdown file
- Addendum 2a – Top 20 Highest averageRating
- Addendum 2b – Top 20 Lowest averageRating
- Addendum 2c – Top 20 Highest ratingsCount
- Addendum 3 – GoodReads.csv

Sources

Kiniulus, M. (2019, December 6). *29 Book Sales Statistics Based on Real Numbers and Studies*. Retrieved from markinblog.com: https://www.markinblog.com/book-sales-statistics/

Neary, L. (2015, Septber 19). *When It Comes To Book Sales, What Counts As Success Might Surprise You*. Retrieved from NPR 50 Hear Every Voice: https://www.npr.org/2015/09/19/441459103/when-it-comes-to-book-sales-what-counts-as-success-might-surprise-you

Prakash, Y. (2021, April 12). *26 Datasets For Your Data Science Project*. Retrieved from kaggle.com: https://towardsdatascience.com/26-datasets-for-your-data-science-projects-658601590a4c

Soumik. (2020). *Goodreads - books*. Retrieved from kaggle.com:

https://www.kaggle.com/jealousleopard/goodreadsbooks

Watson, A. (2020, November 10). *U.S. Book Industry - Statistics & Facts*. Retrieved from Statista:

https://www.statista.com/topics/1177/book-market/#topicHeader__wrapper

D195 Capstone - Can GoodReads Reader Ratings Predict Future Author Success?

Jill Andersen

StudentID 001374500

8/17/2021

```
library(ggplot2)
library(dplyr)
library(gridExtra)
library(tidyr)
library(reshape2)
library(GGally)
library(knitr)
```

Introduction

This report examines a data set with 11,127 observations with 12 variables.

Column headings

```
##  [1] "bookID"          "title"           "authors"                                "averageRating"
##  [5] "isbn"            "isbn13"          "languageCode"     "numPages"
##  [9] "ratingsCount"    "textReviewsCount" "publicationDate"  "publisher"
```

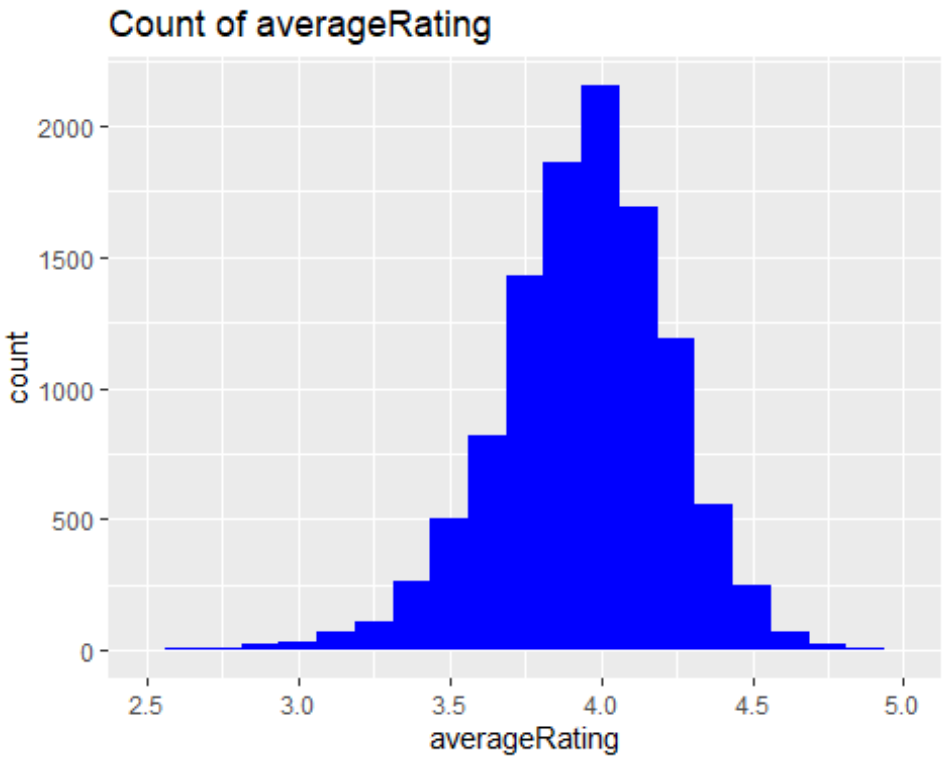
Column headings with a few observations for reference of data types

```
##   bookID                                                                  title
## 1  45531                       Montaillou  village occitan de 1294 Ã  1324
## 2  31373           In Pursuit of the Proper Sinner (Inspector Lynley  #10)
## 3  38568                                          A Quick Bite (Argeneau #1)
## 4  41864 American Genesis: Captain John Smith and the Founding of Virginia
## 5  14142                                                  The Art of Loving
## 6  43940                                                    Object Thinking
##                                        authors averageRating       isbn
## 1 Emmanuel Le Roy Ladurie/Emmanuel Le Roy-Ladurie          3.96 2070323285
## 2                                Elizabeth  George          4.10  553575104
## 3                                     Lynsay Sands          3.91   60773758
## 4                                Alden T. Vaughan          3.43  673393550
## 5         Erich Fromm/Peter D. Kramer/Rainer Funk          4.04   61129739
## 6                                   David      West          3.99  735619654
##     isbn13 languageCode numPages ratingsCount textReviewsCount
publicationDate
## 1 9.78e+12          fre      640           15                2
6/31/2982
## 2 9.78e+12          eng      718        10608              295
```

```
11/31/2000
## 3 9.78e+12          eng      360        35275             1370
3/31/2020
## 4 9.78e+12          eng      224           23                2
8/17/2019
## 5 9.78e+12          eng      192        38148             1310
8/6/2019
## 6 9.78e+12          eng      334          155               21
7/23/2019
##                       publisher
## 1                   Folio histoire
## 2                     Bantam Books
## 3                            Avon
## 4                         Pearson
## 5 Harper Perennial Modern Classics
## 6                   Microsoft Press
```

Review of averageRatings

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.770   3.960   3.934   4.135   5.000
```



The majority titles earned ratings between 3.5 and 4.5. The only titles that received a 5 rating had less than 100 ratings, so I did not include them in the study.

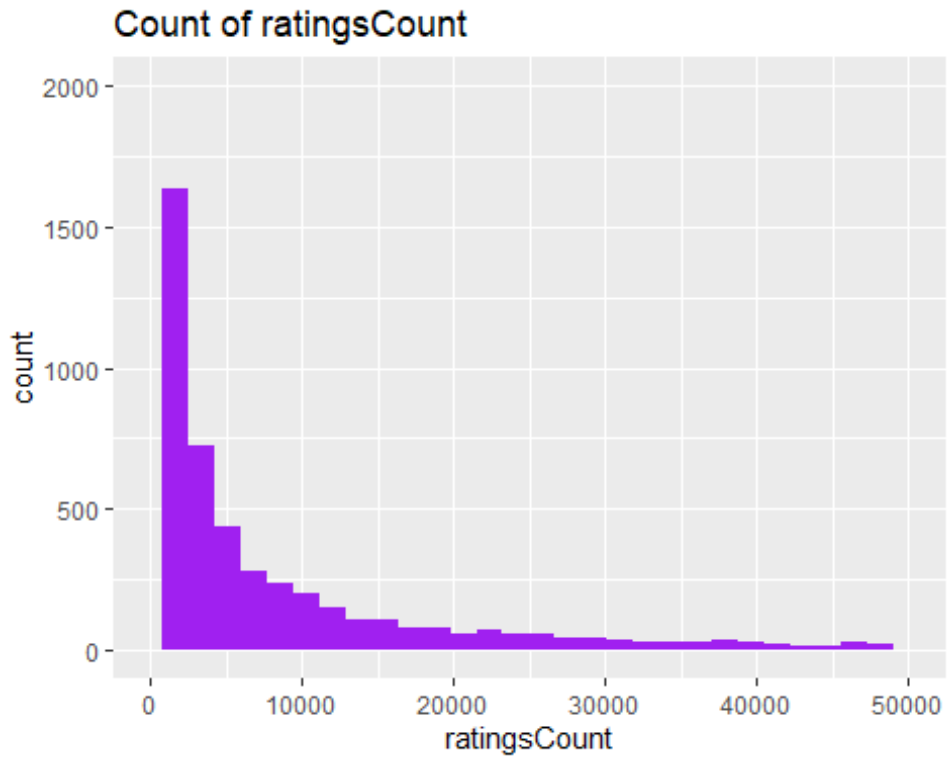
The highest rated title with a minimum of 100 ratings is 'The Complete Calvin and Hobbes', with a rating of 4.82. It had a ratings count of 32,213. Calvin and Hobbes is so popular that there are 6 Calvin and Hobbes titles in the top 20 highest rated titles.

The lowest rated title with a minimum of 100 ratings is 'Citizen Girl' by Emma McLaughlin/Nicola Kraus with a rating of 2.4. It had a ratings count of 5,412.

Addendum 2 is a Word document with various charts. One of these charts is of the top 20 highest rated titles and the top 21 lowest rated titles. (I had to include the 21st lowest rated book because it was co-authored by Oprah Winfrey who is famous for her top book recommendations.)

Review of ratingsCount

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     104     745   17936    4994 4597666
```



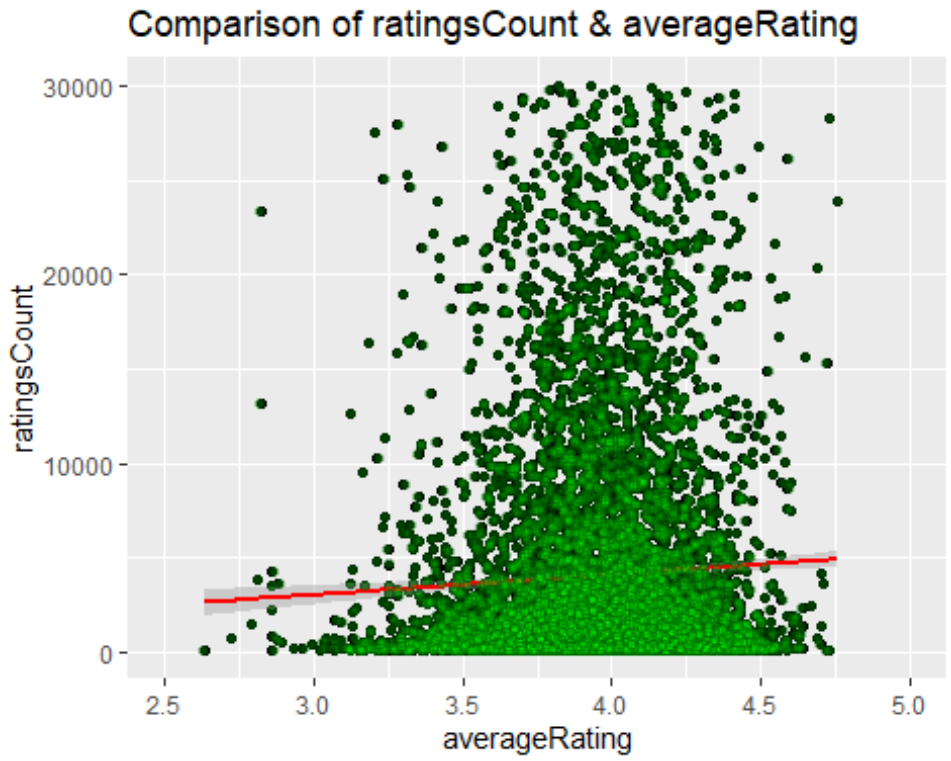
As stated in the review of averageRating, I excluded titles with less than 100 ratings as outliers since their average ratings tended to be much higher or lower than 1st and 3rd quartiles.

The majority of titles received between 100 and 5,000 ratings. The highest rated title Twilight (#1 of the series) with 4,597,666 ratings. I included a list of the top 20 titles with the highest number of ratings.

Comparison of Average Rating or Ratings Count

```
##
##  Pearson's product-moment correlation
##
## data:  GR$averageRating and GR$ratingsCount
## t = 4.0326, df = 11125, p-value = 5.552e-05
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01963810 0.05674575
## sample estimates:
##       cor
## 0.0382051
```



There is surprisingly a low correlation of .038 correlation between the number of ratings a title received compared with the average rating is earns.

As listed above the highest average rated title earned 4.82 and received 32,213 ratings. The title that received the highest number of ratings 4,597,666 received a rating of 3.59.

Review of Authors

There are 11,127 titles in this dataset. There are nearly as many authors as titles, especially if you count each time an author co-authored with a variety of others for some of their books. There are too many subsets of authors and co-author pairings to have sufficient number of groupings with more than 1 or 2 titles to determine if averageRatings has a high or low correlation to an author.

Addendum 2a
Top 20 books with the highest averageRating (with a minimum of 100 ratings)

title	authors	averageRating	ratingsCount	publisher
The Complete Calvin and Hobbes	Bill Watterson	4.82	32213	Andrews McMeel Publishing
Harry Potter Boxed Set Books 1-5 (Harry Potter #1-5)	J.K. Rowling/ Mary GrandPrÃ©	4.78	41428	Scholastic
It's a Magical World (Calvin and Hobbes #11)	Bill Watterson	4.76	23875	Andrews McMeel Publishing
Harry Potter Collection (Harry Potter #1-6)	J.K. Rowling	4.73	28242	Scholastic
Early Color	Saul Leiter/Martin Harrison	4.73	144	Steidl
Homicidal Psycho Jungle Cat (Calvin and Hobbes #9)	Bill Watterson	4.72	15365	Andrews McMeel Publishing
Elliott Erwitt: Snaps	Murray Sayle/Charles Flowers/Elliott Erwitt	4.72	102	Phaidon Press
Calvin and Hobbes: Sunday Pages 1985-1995: An Exhibition Catalogue	Bill Watterson	4.71	3613	Andrews McMeel Publishing
Study Bible: NIV	Anonymous	4.7	4166	Zondervan Publishing House
The Complete Aubrey/Maturin Novels (5 Volumes)	Patrick O'Brian	4.7	1338	W. W. Norton Company
The Price of the Ticket: Collected Nonfiction 1948-1985	James Baldwin	4.7	404	St. Martin's Press
The Days Are Just Packed	Bill Watterson	4.69	20308	Andrews McMeel Publishing
The Sibley Field Guide to Birds of Western North America	David Allen Sibley	4.69	730	Alfred A. Knopf
The Life and Times of Scrooge McDuck	Don Rosa	4.67	2467	Gemstone Publishing
The Absolute Sandman Volume One	Neil Gaiman/Mike Dringenberg/Chris Bachalo/Michael Zulli/Kelly Jones/Charles Vess/Colleen Doran/Malcolm Jones III/Steve Parkhouse/Daniel Vozzo/Lee Loughridge/Steve Oliff/Todd Klein/Dave McKean/Sam Kieth	4.65	15640	Vertigo
The Shawshank Redemption: The Shooting Script	Frank Darabont/Stephen King	4.64	2406	Newmarket Press
The New Annotated Sherlock Holmes: The Complete Short Stories	Arthur Conan Doyle/Leslie S. Klinger	4.64	1411	W. W. Norton & Company
The Gospel According to Luke	Anonymous/Thomas Cahill	4.64	169	Grove Press
The Calvin and Hobbes Tenth Anniversary Book	Bill Watterson	4.63	49122	Andrews McMeel Publishing
The Collected Autobiographies of Maya Angelou	Maya Angelou	4.63	991	Modern Library

Addendum 2b

Top 21 books with the lowest averageRating (with a minimum of 100 ratings)

Citizen Girl	Emma McLaughlin/Nicola Kraus	2.4	5415	Washington Square Press
The Governess; or The Little Female Academy	Sarah Fielding/Candace Ward	2.63	132	Broadview Press Inc
Alentejo Blue	Monica Ali	2.72	788	Scribner Book Company
Yellow Dog	Martin Amis	2.79	1449	Vintage
You Don't Love Me Yet	Jonathan Lethem	2.81	3854	Doubleday Books
Four Blondes	Candace Bushnell	2.82	23409	Grove Press
Lost	Gregory Maguire/Douglas Smith	2.82	13152	William Morrow Paperbacks
The Thomas Berryman Number	James Patterson	2.86	4320	Grand Central Publishing
Up in the Air	Walter Kirn	2.86	3504	Anchor
Lair of the White Worm	Bram Stoker	2.86	2276	Deodand Publishing
The Diagnosis	Alan Lightman	2.86	898	Vintage
Desire and Duty: A Sequel to Jane Austen's Pride and Prejudice	Ted Bader/Marilyn Bader	2.86	114	Revive Publishing
Le Divorce	Diane Johnson	2.88	3602	Plume
Checkpoint	Nicholson Baker	2.88	654	Vintage
Queen of the Underworld	Gail Godwin	2.89	541	Ballantine Books
Secret Identity (Lost #2)	Catherine Hapka	2.93	195	Voice
The Doctor's House	Ann Beattie	2.96	164	Scribner
Dinner with Anna Karenina	Gloria Goldreich	2.99	411	Mira Books
Closing Time	Joseph Heller	3.02	106	Simon & Schuster (Trade Division)
The Diviners	Rick Moody	3.03	508	Back Bay Books
In the Kitchen with Rosie: Oprah's Favorite Recipes	Rosie Daley/Oprah Winfrey	3.04	767	Knopf

Addendum 2c

Top 20 books with the highest ratingsCount

title	authors	averageRating	ratingsCount	publisher
Twilight (Twilight #1)	Stephenie Meyer	3.59	4597666	Little Brown and Company
The Hobbit or There and Back Again	J.R.R. Tolkien	4.27	2530894	Houghton Mifflin
The Catcher in the Rye	J.D. Salinger	3.8	2457092	Back Bay Books
Angels & Demons (Robert Langdon #1)	Dan Brown	3.89	2418736	Pocket Books
Harry Potter and the Prisoner of Azkaban (Harry Potter #3)	J.K. Rowling/ Mary GrandPrÃ©	4.56	2339585	Scholastic Inc.
Harry Potter and the Chamber of Secrets (Harry Potter #2)	J.K. Rowling/Mary GrandPrÃ©	4.42	2293963	Arthur A. Levine Books / Scholastic Inc.
Harry Potter and the Order of the Phoenix (Harry Potter #5)	J.K. Rowling/ Mary GrandPrÃ©	4.49	2153167	Scholastic Inc.
The Fellowship of the Ring (The Lord of the Rings #1)	J.R.R. Tolkien	4.36	2128944	Houghton Mifflin Harcourt
Animal Farm	George Orwell/Boris Grabnar/Peter Å kerl	3.93	2111750	NAL
Harry Potter and the Half-Blood Prince (Harry Potter #6)	J.K. Rowling/ Mary GrandPrÃ©	4.57	2095690	Scholastic Inc.
Lord of the Flies	William Golding	3.68	2036679	Penguin Books
Romeo and Juliet	William Shakespeare/Paul Werstine/Barbara A. Mowat	3.74	1893917	Simon Schuster
The Lightning Thief (Percy Jackson and the Olympians #1)	Rick Riordan	4.25	1766725	Disney Hyperion Books
Of Mice and Men	John Steinbeck	3.87	1755253	Penguin Books
The Da Vinci Code (Robert Langdon #2)	Dan Brown	3.84	1679706	Anchor
The Alchemist	Paulo Coelho/Alan R. Clarke/Ã–zdemir Ä°nce	3.86	1631221	HarperCollins
The Giver (The Giver #1)	Lois Lowry	4.13	1585589	Ember
The Book Thief	Markus Zusak/Cao XuÃ¢n Viá»‡t KhÆ°Æ¡ng	4.37	1516367	Alfred A. Knopf
Little Women	Louisa May Alcott	4.07	1479727	Signet Classics