

So Many Books, So Little Time

Web Scraping Good Reads



SUNANDA MISHRA

Agenda



- What is GoodReads?
- Why scrape it?
- How to scrape it?
- Preview: What to do with scraped data?
- Future studies with data

So...you like to read



goodreads Home My Books Browse ▾ Community ▾ Search books

CURRENTLY READING

Children of Blood and Bone (Legacy of O...) by Tomi Adeyemi 9 80/525 (15%) Update progress

All the Light We Cannot See by Anthony Doerr 9 43% Update progress

[View all books](#) • [Add a book](#) • [General update](#)

2018 READING CHALLENGE

2018 READING CHALLENGE 7 books completed 2 books ahead of schedule 7/10 (70%) View Challenge

UPDATES

Laura wants to read 41m

The Greatest Love Story Ever Told by Megan Mullally Want to Read Rate it: ★★★★★

At last, the full story behind Megan Mullally and Nick Offerman's epic romance, including stories, portraits, and the occasional puzzle, all telling the smoldering tale... [Continue reading](#)

[Like](#) · [Comment](#)

Write a comment...

Laura wants to read 42m

The Library Book by Susan Orlean 9 Want to Read Rate it: ★★★★★

Susan Orlean, hailed as a "national treasure" by The Washington Post and the acclaimed bestselling author

Customize

THE GOODREADS BLOG

7 Great Books Hitting Shelves Today

17 likes • 10 comments

RECOMMENDATIONS

Because you are currently reading All the Light We Cannot See:

Strong Poison (Lord Peter Wimsey, #6) by Dorothy L. Sayers ★★★★★ 4.14 avg. rating Want to Read

Why Scrape it?



- TONS of great data
- There are more than 10,500 bookclubs on GoodReads!
- Most clubs are made up of 1 person!!
- Bookclub Moderators:
 - Helpful to know how to get your club up and running
- Bookclub Members:
 - Helpful to know which bookclub to choose

What does a bookclub's page look like?



2018
Reading
Challenge

[Join Group](#)

2018 Reading Challenge

Are you ready to set your 2018 reading goal?

This is a supportive, fun group of people looking for people just like you. Track your annual reading goal here with us, and we have challenges, group reads, and other fun ways to help keep you on pace. There will never be a specific number of books to read here or pressure to read more than you can commit to. Your goal is five? Great! You think you want to read 200? Very cool!

We won't kick you out for not participating regularly, but we'll love it if ...[more](#)

category Books & Literature -> General
tags 2012, 2013, 2014, 2015, 2016, 2017, 2018, bookclub, bookriot, and buddy-reads
group type This is a public group. Anyone can join and invite others to join.
rules 1. Mark spoilers. You can do this by...[more](#)

[flag this group \(?\)](#)

CURRENTLY READING

Group Home **Events** **Invite People**
Bookshelf **Photos** **Members**
Discussions **Videos** **Polls**
Challenges

[Search](#)

MODERATORS [tools & guidelines](#)

 Kara TBR Twins	 Kadijah Michelle Group Organization
 Winter Group Reads	 Mary Pat Buddy Reads
 Ilona Welcome	 Kristin Challenges

What does a bookclub's page look like?

●

[flag this group \(?\)](#)



CURRENTLY READING



Red Queen (Red Queen, #1)
by Victoria Aveyard

Start date
August 1, 2018 **Finish date**
August 31, 2018

[view activity »](#)

DISCUSSION BOARD

[topics: all | new | unread](#)

▼ New here? Drop a line...

Showing 5 of 0 topics — 2,671 comments total

* Introduce Yourself to the Group - 2018!	By Ilona , Welcome Committee · 637 posts (637 new) · 696 views	last updated 5 hours, 56 min ago
🔒 * NEW MEMBERS: Start Here	By Kadijah Miche... , Group Organization · 3 posts (3 new) · 2032 views	last updated Dec 28, 2017 07:10AM
* 2018 Personal Challenge	By Ilona , Welcome Committee · 61 posts (61 new) · 2378 views	last updated Aug 06, 2018 07:30AM
* How do I do things on Goodreads?	By Kara , TBR Twins · 299 posts (299 new) · 5401 views	last updated 2 hours, 47 min ago

What does a bookclub's page look like?



2018 Reading Challenge > books > read

Only moderators of this group
can add books.

(showing 1-30 of 80)

SEARCH FOR BOOKS

GROUP SHELVES

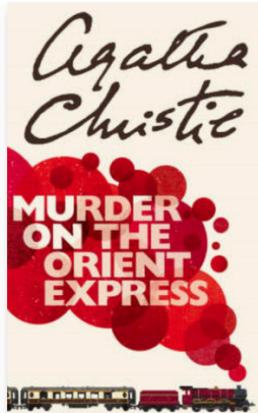
all (81)
read (80)
currently-reading (1)
to-read

group-reads (80)
upcoming (0)

view: main | covers

	title	author	my rating	shelves	date started	date finished	added by	date added	
	 Little Fires Everywhere	Ng, Celeste *	★★★★★	read, group-reads	2018/07/01	2018/07/31	Winter	2018/06/26	view activity »
	 The Shape of Water	del Toro, Guillermo	★★★★★	read, group-reads	2018/06/01	2018/06/30	Winter	2018/05/18	view activity »
	 A Study in Scarlet (Sherlock Holmes, #1)	Doyle, Arthur Conan	★★★★★	read, group-reads	2018/06/01	2018/05/30	Winter	2018/05/18	view activity »
	 Middlesex	Eugenides, Jeffrey	★★★★★	read, group-reads	2018/05/01	2018/05/31	Winter	2018/04/10	view activity »
	 Big Little Lies	Moriarty, Liane *	★★★★★	read,	2018/04/01	2018/04/30	Winter	2018/03/11	view activity »

What does a book's page look like?



Want to Read

Rate this book

Preview

Agatha Christie

MURDER ON THE ORIENT EXPRESS

Original Title Murder on the Orient Express

ISBN 0007119313 (ISBN13: 9780007119318)

Edition Language English

Series Hercule Poirot #10

Characters Samuel Edward Ratchett, Hector MacQueen, Masterman, Colonel Arbuthnot, Harriet Hubbard...more

setting Constantinople
Vinkovci (Croatia)
Yugoslavia

Murder on the Orient Express (Hercule Poirot #10)

by Agatha Christie

★★★★★ 4.16 ·  Rating details · 256,515 Ratings · 16,502 Reviews

What more can a mystery addict desire than a much-loathed murder victim found aboard the luxurious Orient Express with multiple stab wounds, thirteen likely suspects, an incomparably brilliant detective in Hercule Poirot, and the most ingenious crime ever conceived?

GET A COPY

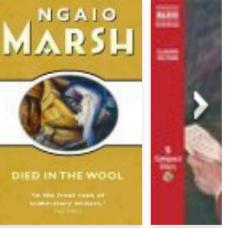
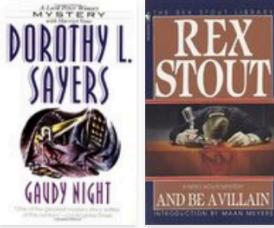
Kindle Store \$7.99 Amazon Stores ▾ Libraries

Paperback, Agatha Christie Signature Edition, 274 pages
Published June 4th 2007 by HarperCollins Publishers (first published 1934)

Share   

Recommend It | Stats | Recent Status Updates

READERS ALSO ENJOYED



GENRES

Mystery	7,951 users
Classics	3,354 users
Fiction	3,026 users
Mystery > Crime	1,624 users

[See top shelves...](#)

ABOUT AGATHA CHRISTIE



How to scrape



- 3 Main Spiders:
 - Scrape the bookclub's page
 - ✖ Group Name, Number of Members, currently reading, location, tags, group type
 - ✖ 1800 bookclubs
 - Scrape the bookshelf
 - ✖ Group name, book titles
 - Scrape the books
 - ✖ Number of ratings, what the rating is, number of pages in the book, year published

What to do with scraped data?

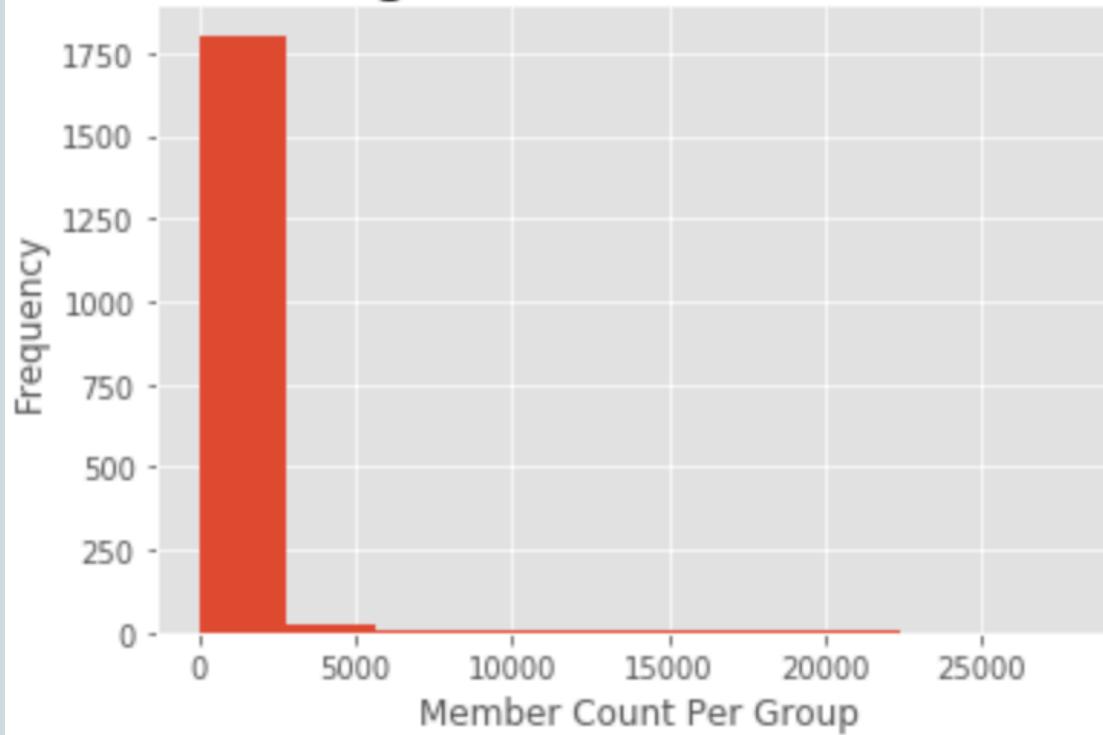


- How big are most book clubs?
- Do larger and smaller book clubs behave differently?
 - Compare average book rating
 - Compare popularity of books
 - Compare book length
 - Compare how recently published the book was

Member Count



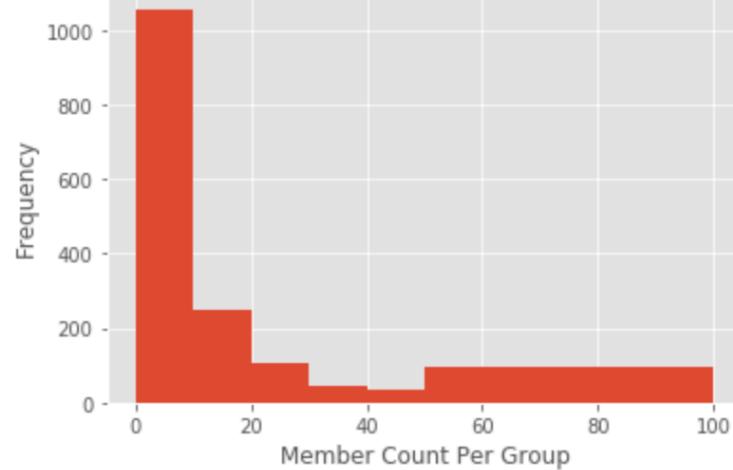
Histogram of Member Count



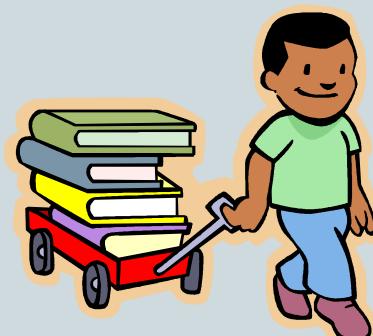
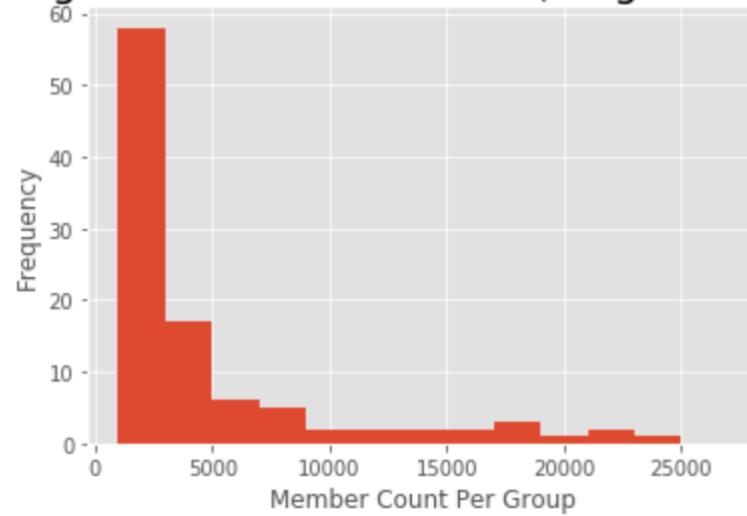
Member Count



Histogram of Member Count (Small Bookclubs)



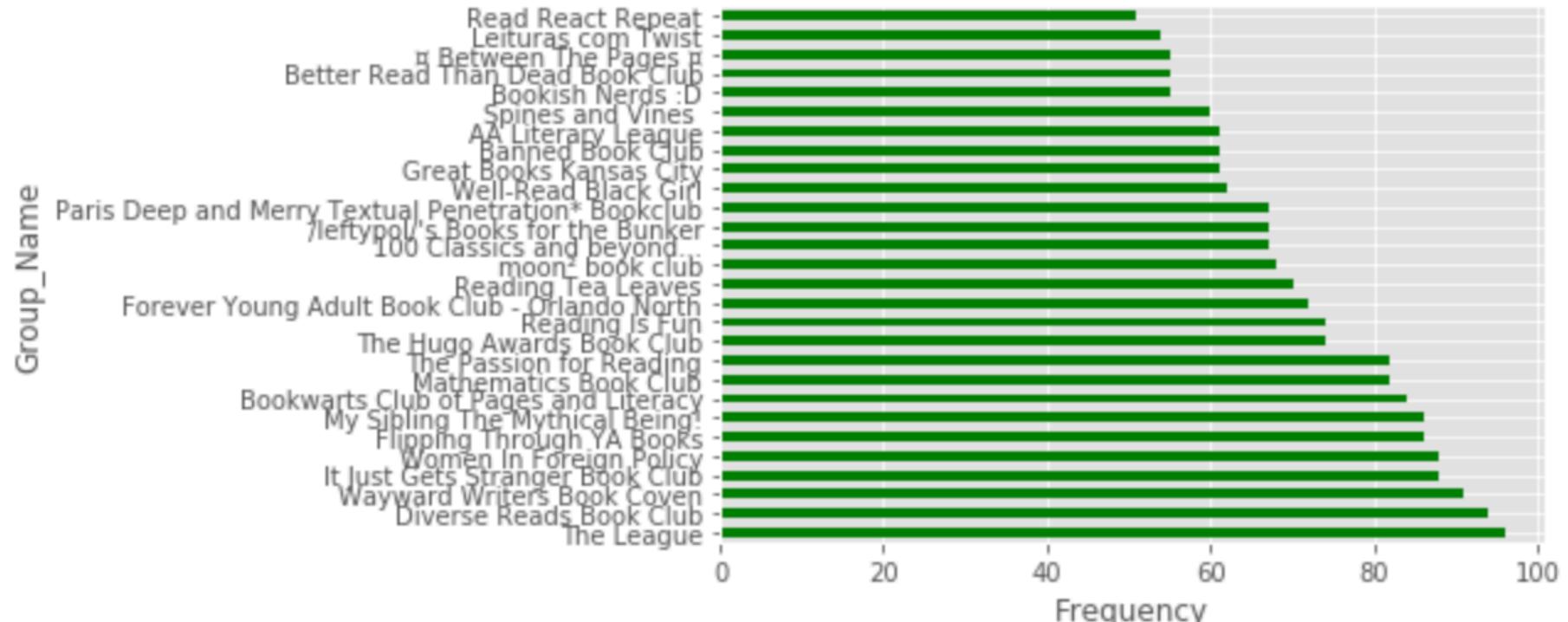
Histogram of Member Count (Large Bookclubs)



Examining Book Club Behavior



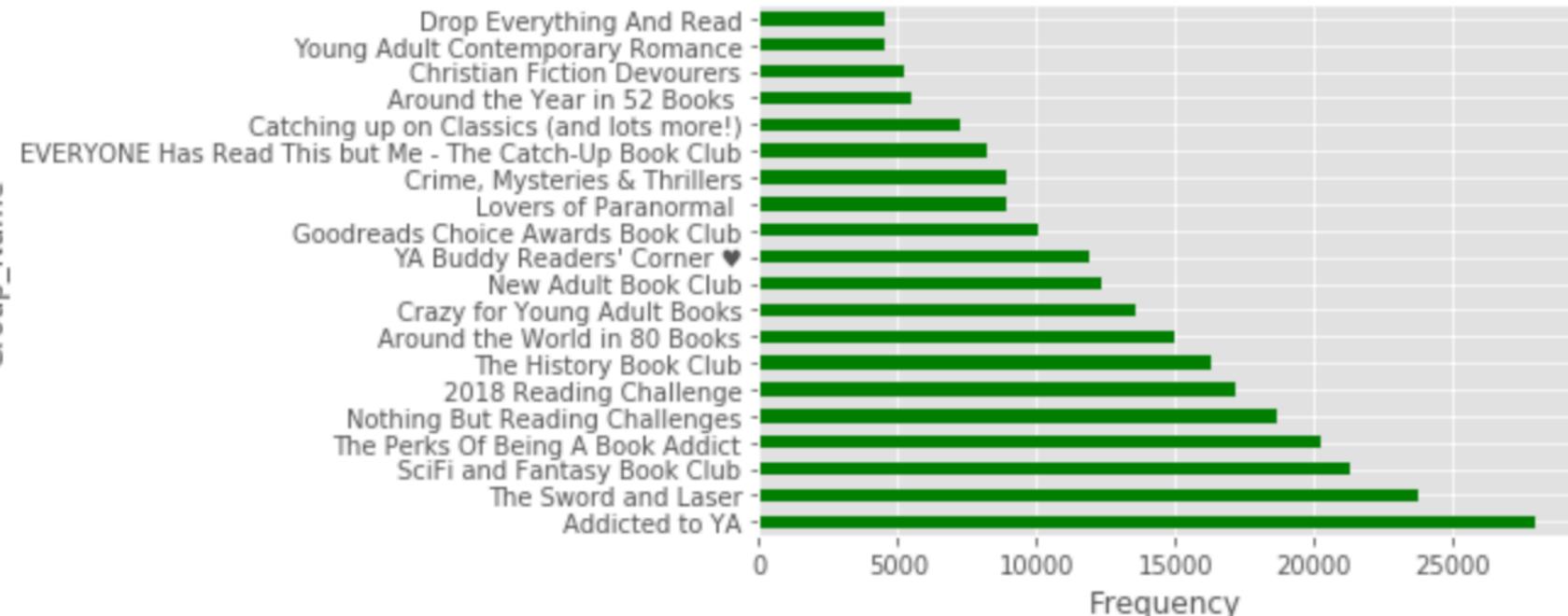
Histogram of Member Count (Small Bookclubs)



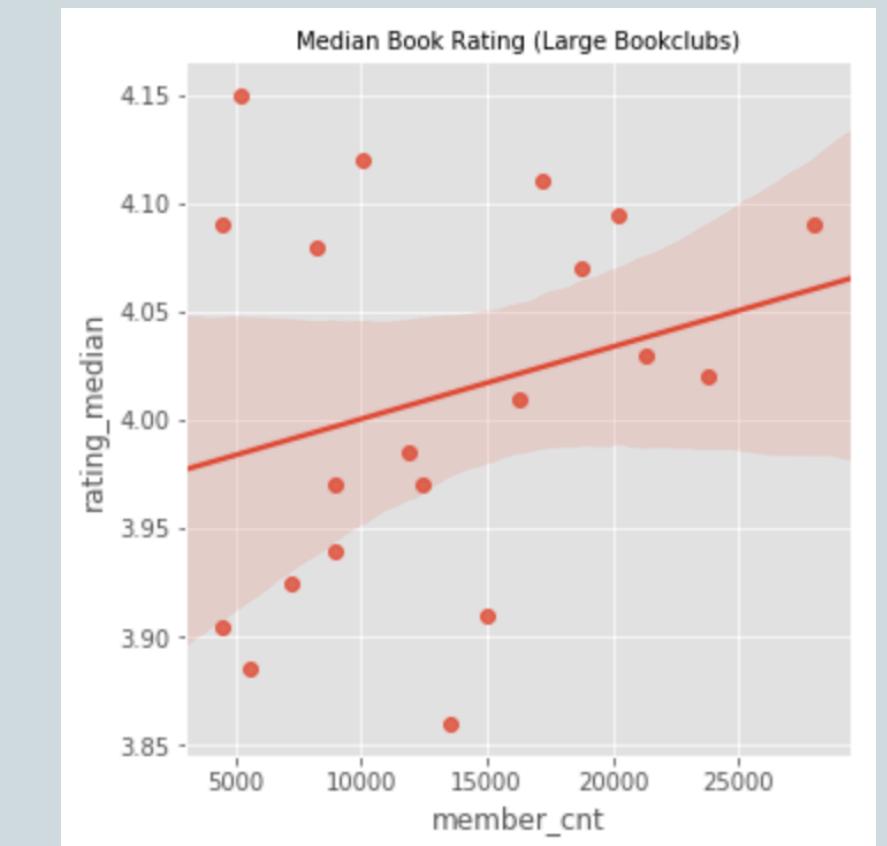
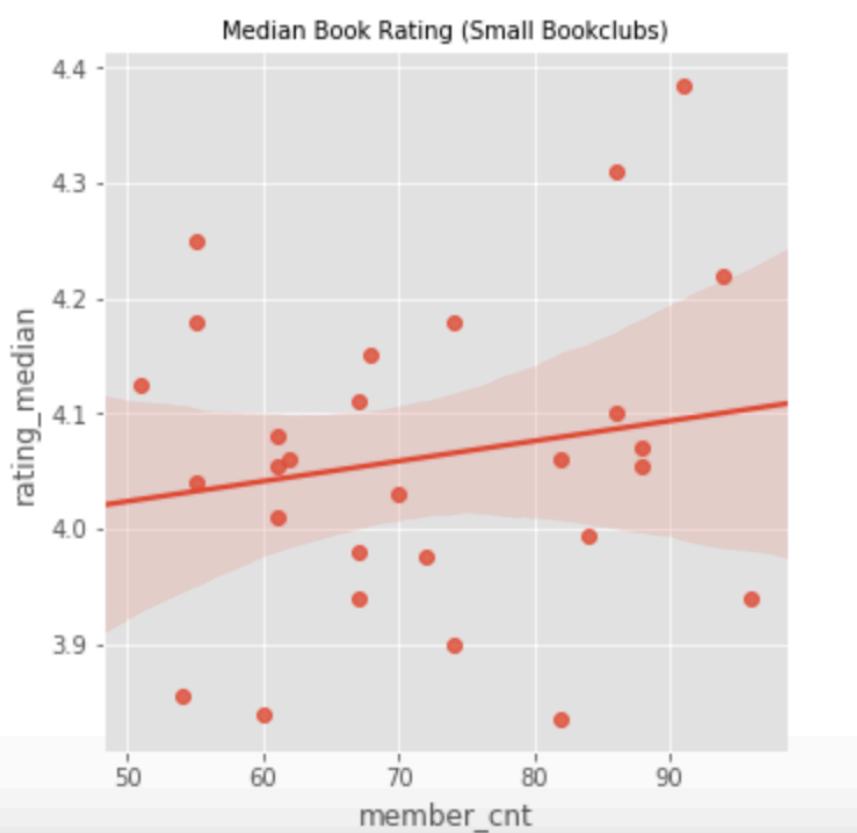
Examining Book Club Behavior



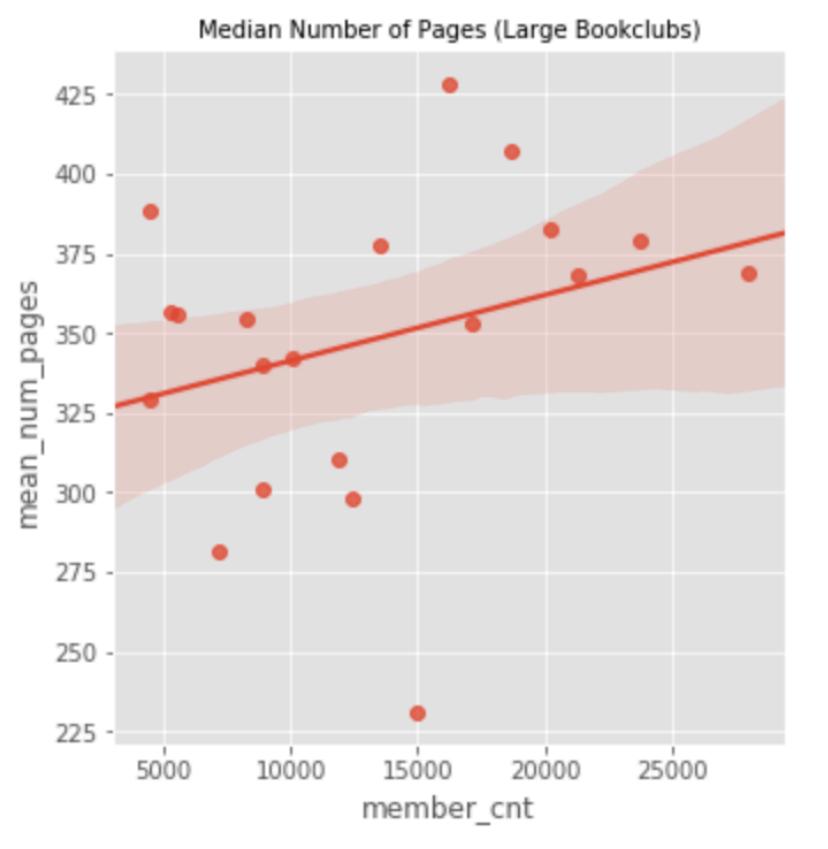
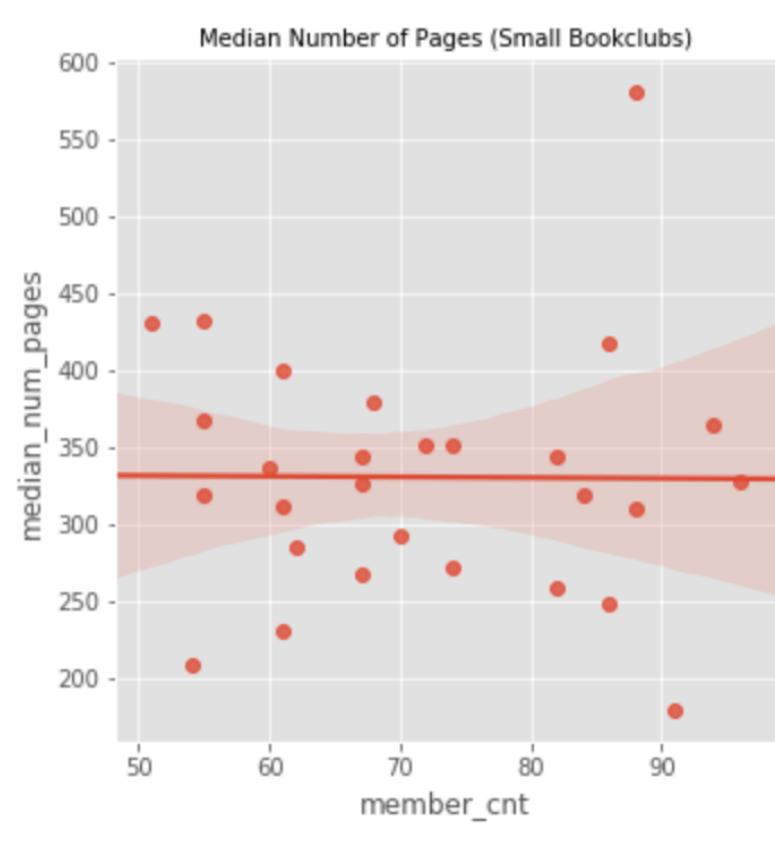
Histogram of Member Count (Large Bookclubs)



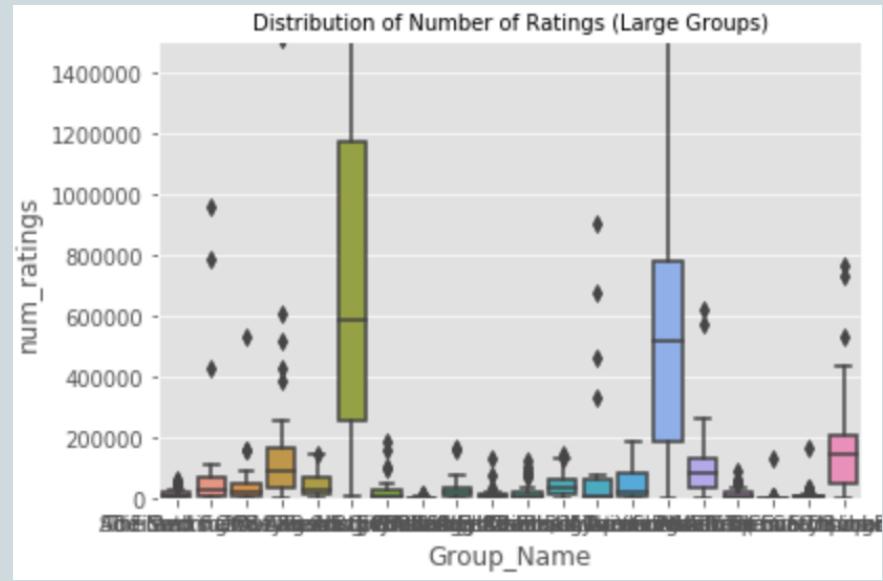
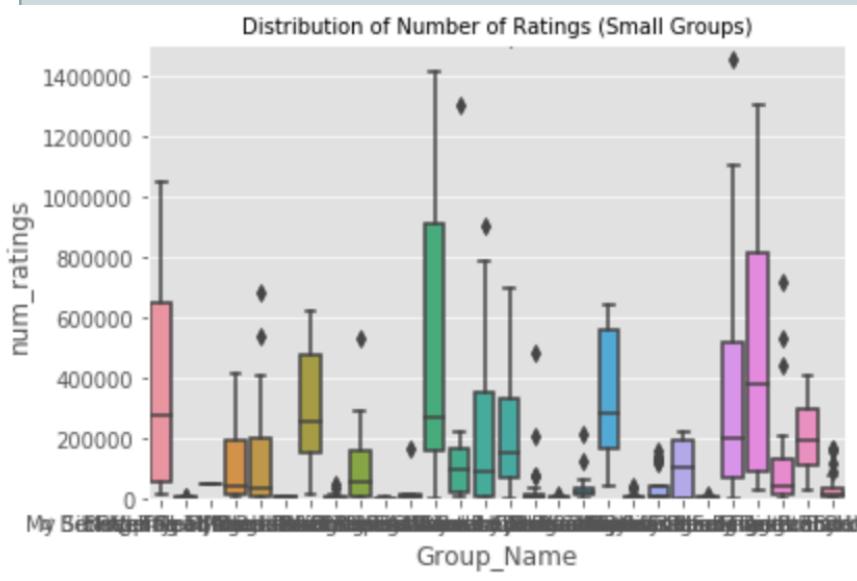
Book Club Behavior: Book Rating



Book Club Behavior: Number of Book's Pages



Book Club Behavior: Number of Ratings



Future Study Ideas



- Do certain genres draw more members into their book club?
- What makes a book club more likely to have a lot of discussions?
- What books are repeatedly selected for book club readings? What are characteristics of these books?