



People Still *Read*?


Understanding the current Reader Behavior Landscape of
Today to Influence Publishing of Tomorrow

Mallory Banks


Metis Module 3 Business Fundamentals

Goals, Impacts, and Solution Path


Goals


 The Goal of this Analysis is to understand the landscape of reading in the modern Era by analyzing one of the most popular reading social networks today as well as processing a survey on reading habits of different demographics


Impacts


 The Impact this can have for the client is that they will be able to create a new concept of what the “traditional” reader looks like based on the datasets explored. Additionally, they may be able to increase traditional book sales based on devising new and more applicable marketing campaigns that identify and galvanize readers that may have been underserved to this point.


Solution Path:

 **Data Acquisition** from GoodReads (Dataset procured from now defunct GoodReads API) and Dataset of Reading Habits Survey 2020 from Pew Research Center

 **EDA:** Manipulate/concat massive Book Dataset, divide into Highly Rates and Lowly rated, and re-sample Survey data on racial categories


 **Insight Gathering:** Tagged “High Activity” Readers to Understand Demos of Readers who are reading above national Average of 11 books per year or the median of the dataset (6 books per year).


 **Insight Gathering:** Looked at attributes of Highly Rated Books and Lower Rated Books to Understand Differences

 **Future Steps:** Classification Research needed to understand attributes and groupings of book types leading to higher reviews


Assumptions and Risks Impacting this Project


Risks

 Due to the limited nature of research done on Literacy in the United States, the responses of Reading habits surveys are skewed White and Upper Class. Some of this is mitigated with slight upsampling, but the fact remains that survey data will be biased

 Because Good Reads retired it's API once it was acquired by Amazon, we aren't able to go back and easily collect even more data leaving the dataset limited only the years of publication through to ~2013 (and even this is a little shakey). This analysis acts as a proof of concept for increased funding in literacy and reading research

Assumptions

 Because of the Risks as previously states, we are assuming that Good Reads and its users are a good representation of public POV on Books themselves based on their ratings. We know that these things can change quickly and more so, that there are many other places that readers are entering to make their voices known (booktok, bookstagram, startups like Story Graph, and Amazon Reviews)

 Additionally, we are assuming that the survey data collected by Pew in 2020, though self reported, is good faith answered by respondents and a good proxy for a wider range of different types of readers

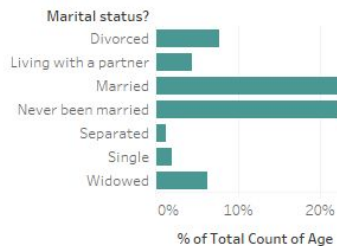


Who is Reading The Most These Days?

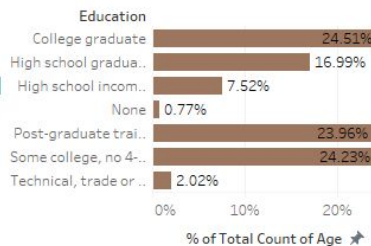
Across the Board, Women are reading the most. Generally, Higher Income household read at highest rates, as well as more highly educated communities

Highly Active Reader Demographics

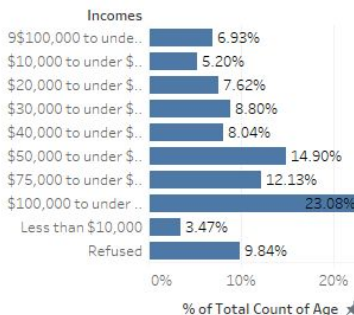
Marital Status of Highly Active Readers



Education Distribution of Highly Active Readers



Income Distribution of Highly Active Readers



Racial and Gender Distribution of Highly Active Readers

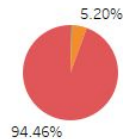
Race	Sex	
African-American	Male	3.68%
Mixed race	Female	0.90%
	Male	1.11%
Native American/ American...	Female	0.35%
	Male	0.56%
Other	Female	0.76%
	Male	0.63%
Refused	Female	0.56%
	Male	1.00%



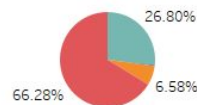
How Are Readers Reading Across Activity Levels?

While we do see that many Highly Active Readers read across many different mediums, Low activity readers are those who mainly stick with traditional publishing - is there room in publishing to bring new types of reading behavior to “Low Activity” Readers

Trad High Activity Readers

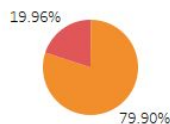


Traditional Low Activity Readers

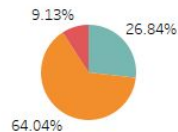


Legend
Null
No
Yes

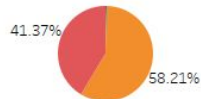
Audio Book High Activity Readers



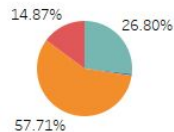
Audio Book Low Activity Readers



E Book High Activity Readers



E Book Low Activity Readers

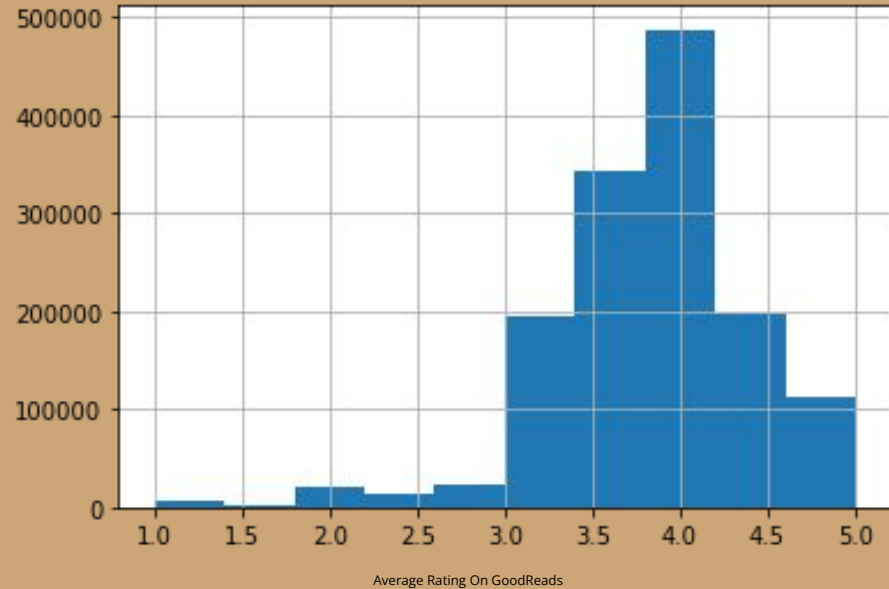




Do Folks Care About What They Are Reading? Do they take to the internet to complain or share the love?

Yes! Though, many books rated on GoodReads are those that are Highest Rated. Contrary to the idea that most people online on want to complain! Is there in house survey research that could work to continue to fill out this story?

Histogram of Avg. Rating of 1.2M+ Books on GoodReads

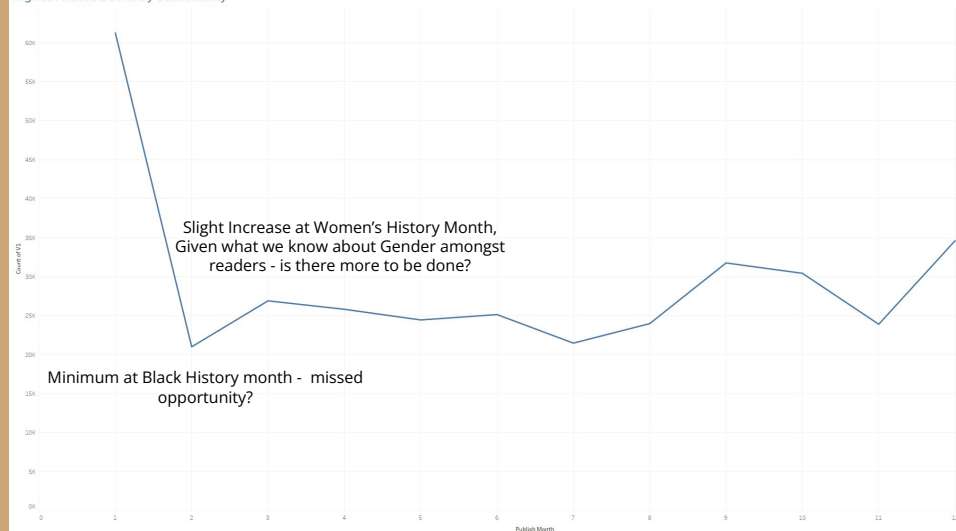




When do the Best Books Come Out?

When Subsetting on Books Rated higher than the mean, we see that these books are Most Likely to be released in the winter season. Books Released in the Winter Season

Highest Rated Books by Seasonality

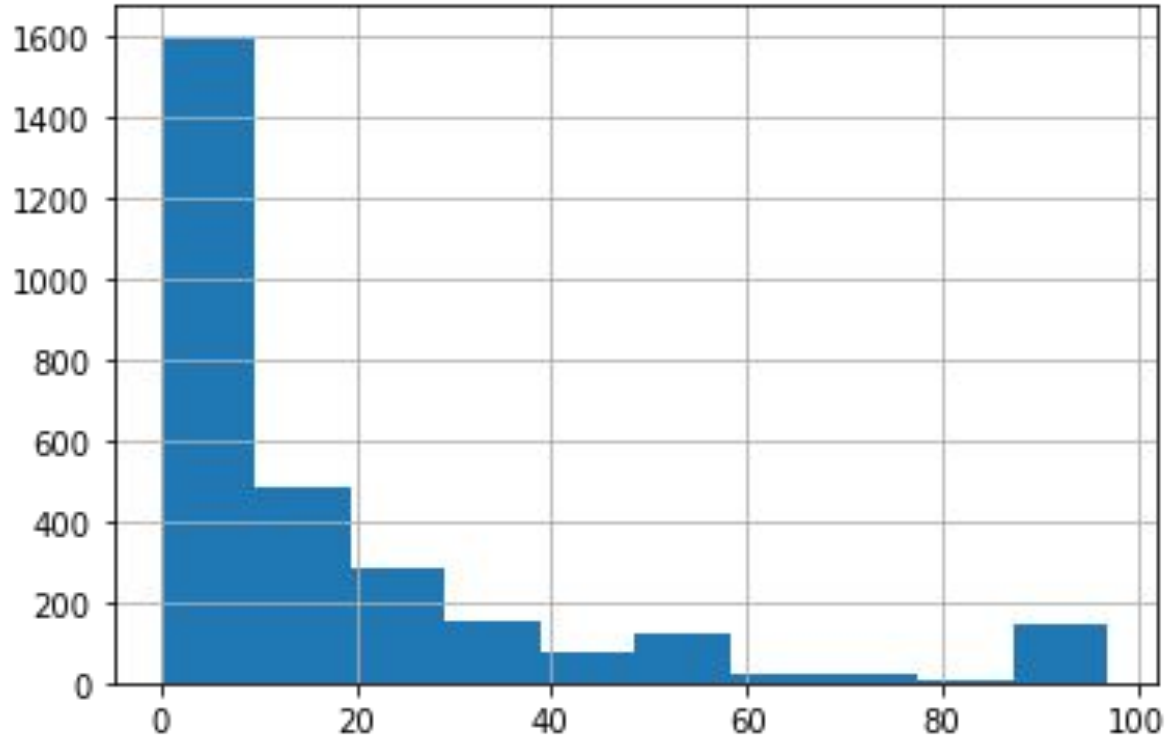


Conclusions and Future Research Needed

- Across the board, and perhaps contrary to some popular belief, the female demographic is most likely to be highly active readers
- The Max of Books Read of Respondents is 97 in a single year
- Many Readers use GoodReads as less of a “Complaint” mechanism and more of a “Social Network” mechanism where they discuss the things that mean the most to them, thus a mean average of 3.83 rating (Higher than one might think)
- Is it that the highest rated books are released during the holidays? Or are more books writ large during the holidays
- Further Data Science Work needed to scrape in Genre data, prime candidate for classification work

Appendix

Distribution of Survey Q “How Many Books Read in the last Year” by Respondents



Racial Distribution of Reading Habits Respondents

