



DATA ANALYSIS ON AMAZON BOOKS

What's the relation between factors such as genre, author, and user engagement with reviews, ratings and reception of books on Amazon?

B. Nil Atabey Julio Zenelaj



TABLE OF CONTENTS



Introduction



The Dataset



Pipeline



Analyses



Conclusion



01 INTRODUCTION



AMAZON – THE LARGEST ONLINE RETAILER

- Amazon is the largest online retailer in the world (Zufelt, 2024).
- Amazon.com implemented online reviews in 1995 (Ante, 2009)
- There are good alternatives for online reviews on books:
 - Goodreads
 - websites of retailers (Barnes & Noble)



(Brands of The World, 2018)

Customer Reviews

★★★★★ (9,305)

4.8 out of 5 stars



[See all 9,305 customer reviews](#)

Thanks J.K. Rowling for writing such a great story.

"judicita"

Great plot, fun story, great characters, it was really just a well rounded book.

ShortyThgAddict

Amazon Books - 2014

Review/Summary

ProfileName

Review/Score

Most Helpful Customer Reviews

209 of 227 people found the following review helpful

★★★★★ **Three Harry Potter Books in Three Days!**

By [Don Halpern](#) on July 1, 2000

Format: Hardcover

An adult friend (age 49) loaned me three Harry Potter books for the summer. Wednesday evening I began the first book and I finished Saturday morning. I am writing this review before I order the fourth Potter book. Will my friend be surprised to get 4 books back vividly presented in a cast of almost believable characters attending a school we all wish we could attend. Classes like "Defense Against the Dark Arts", "Transfiguration", "Arithmancy" and "Care of Magical Creatures" are written as if the author actually attended them minute of class. More than can be said for most of the classes I have attended. Each book in the series encompasses one year of the Potter books are written in a way that can charm any age reader. I am 64.

Review/Text

4 Comments | Was this review helpful to you?

Yes

No

Review/Helpfulness

93 of 101 people found the following review helpful

★★★★★ **Ages 9-12? Hah!**

By [A Customer](#) on November 20, 1999

Format: Paperback

Harry Potter and the Sorcerer's Stone is one of those rare children's books that seems to be utterly wasted on children. The plot is likeable, and it's a good quick read for those older than the specified ages. I'm 18, and I finished it in a few hours, then handed it to my 39-year-old sister. After she finished it, she agreed that we needed to get the rest of the series. In a family that regularly reads Shakespeare, the



02 THE DATASET



OUR DATASET

Books.csv

Book Details
Title
Description
Authors
Publisher
PublishedDate
Categories
RatingsCount

~200K books and ~3M reviews,
Contains 1996 - 2014

(Bekheet, 2022)

Reviews

Reviews
ID
Title
Price
User_ID
ProfileName
Review/Score
Review/Summary
Review/Text
Time/Day
Time/Month
Time/Year
Upvotes
Downvotes

Reviews.csv



03 **PIPELINE**





Loading the data

```
library(<insertimports>)\n\nreviews <- fread("<insertpath>/reviews.csv")\nbooks <- fread("<insertpath>/books.csv")
```



Data wrangling

E.g. ['Fiction'] -> Fiction

```
all$categories <- gsub("[[:punct:]]", "", all$categories)
```



Data analysis

```
names(reviews) # check the dataset\nreview_c <- table(reviews$Title) # # of reviews per book\nmost_reviewed <- sort(review_c, decreasing = TRUE) # sort
```



Visualization

```
ggplot(top_authors_table, aes(x = reorder(Author, Book_Count), y = Book_Count))\n  geom_bar(stat = "identity") +\n  coord_flip() +\n  labs(title = "Number of Books Written by Authors",\n       x = "Name Of Author",\n       y = "Number of Books Published")
```

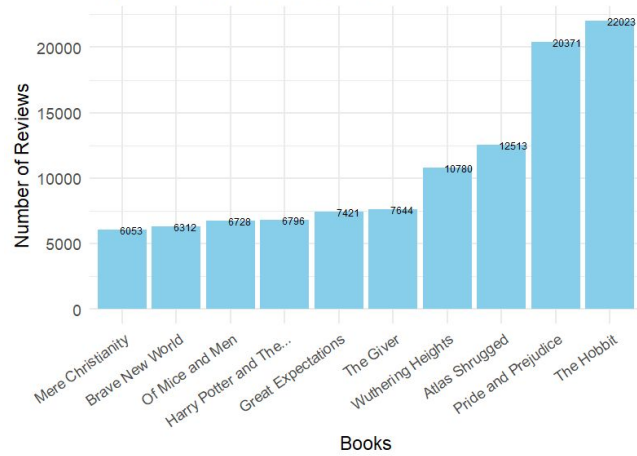


04 **ANALYSES**

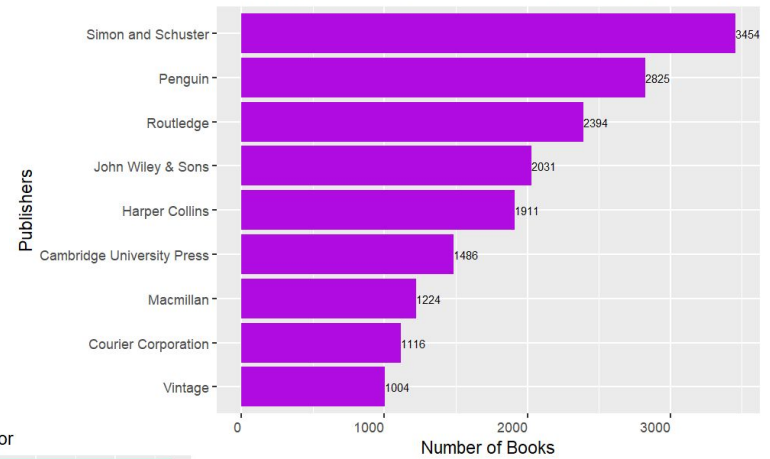


BOOKS

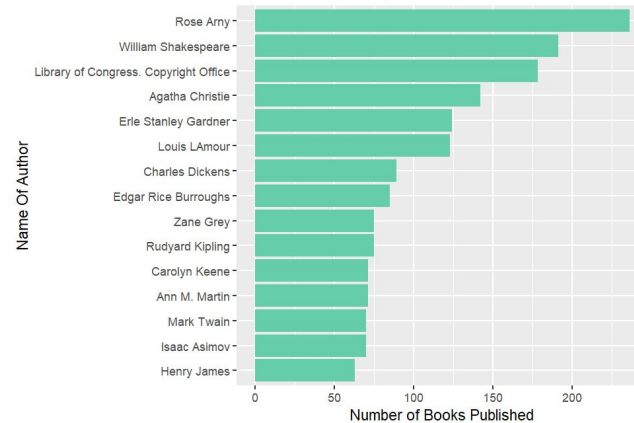
Top 10 Most Reviewed Books



Publishers with More Than 1000 Books On the Market

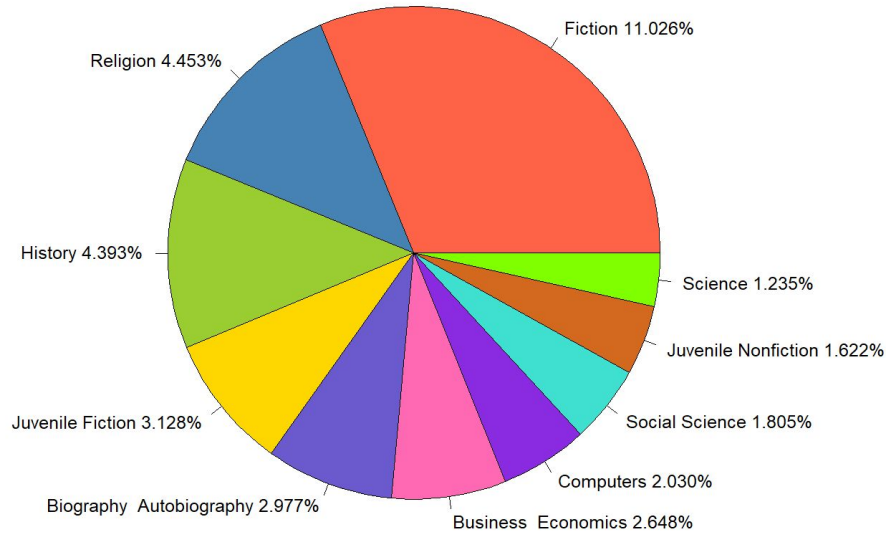


Number of Books per Author



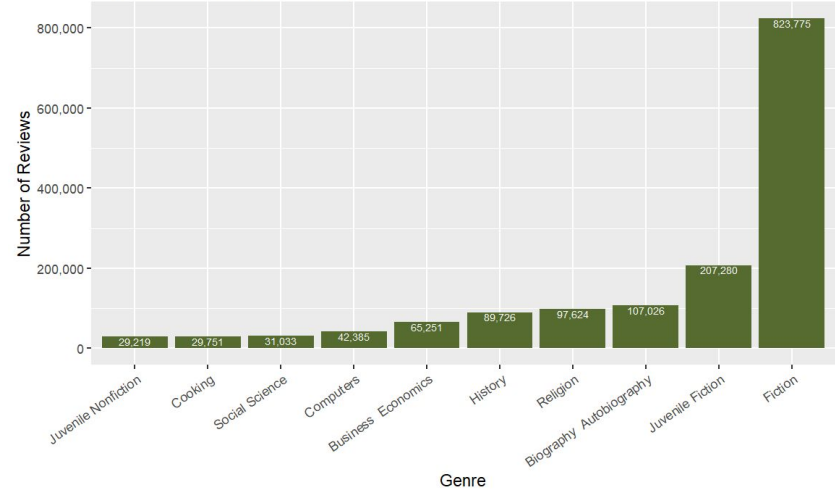
BOOKS

Distribution of Books in Market based on The Top 10 Genres



The remaining 65% is made up of all other categories

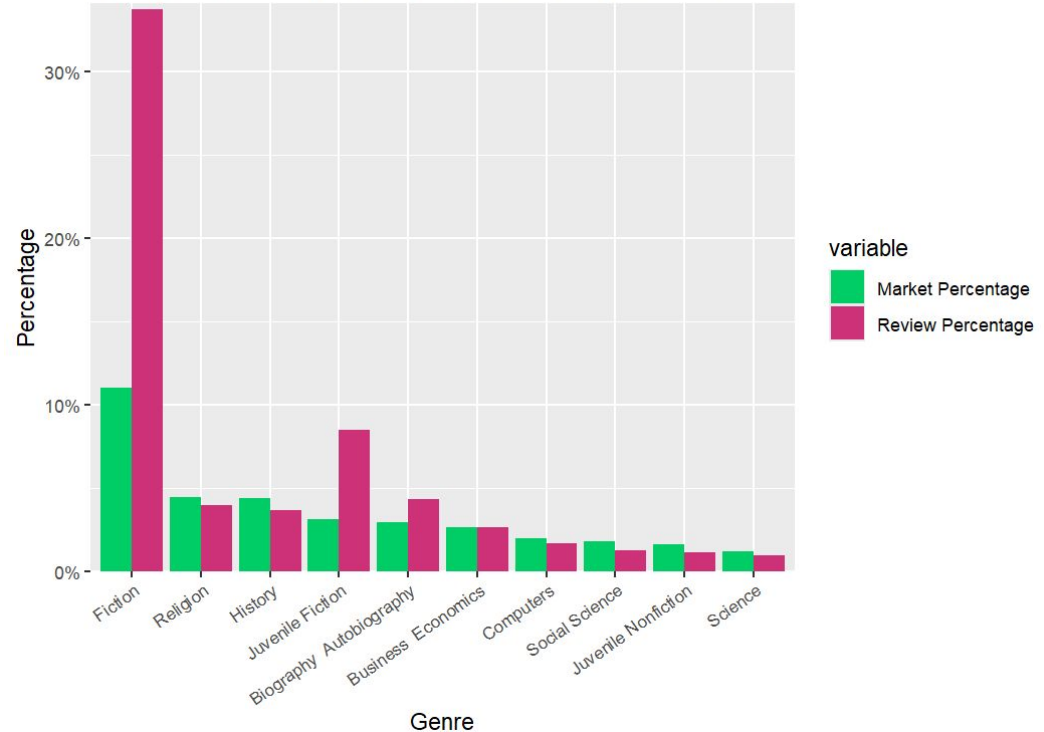
Reviews of the Top 10 Genres



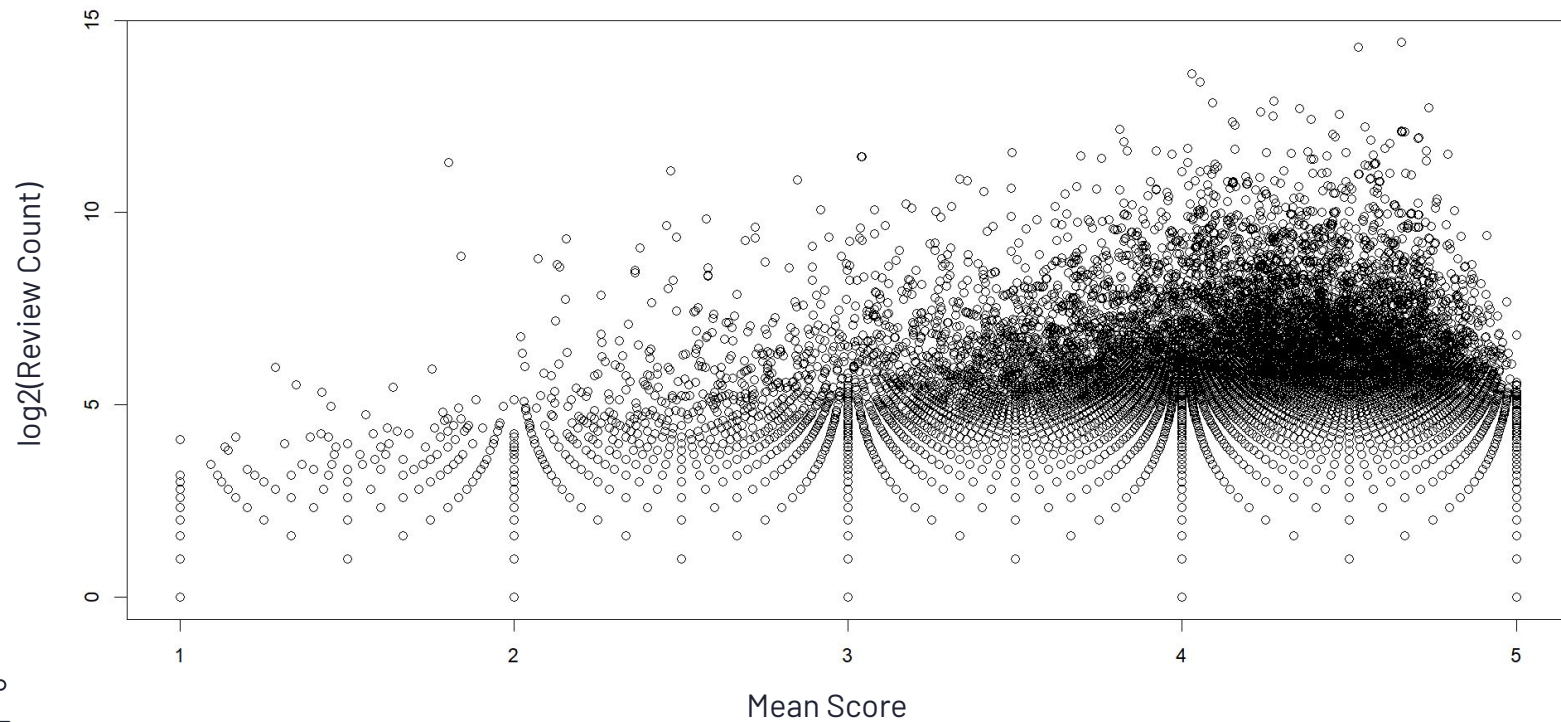
REVIEWS

- Fiction makes up ~12% of the books sold while being the ~35% of the reviews
- Fiction and Juvenile Fiction stand out as the most popular genres overall
- This could suggest that fiction books generate more reader engagement and discussion

Comparison of Top Genres: Market vs Review Percentage

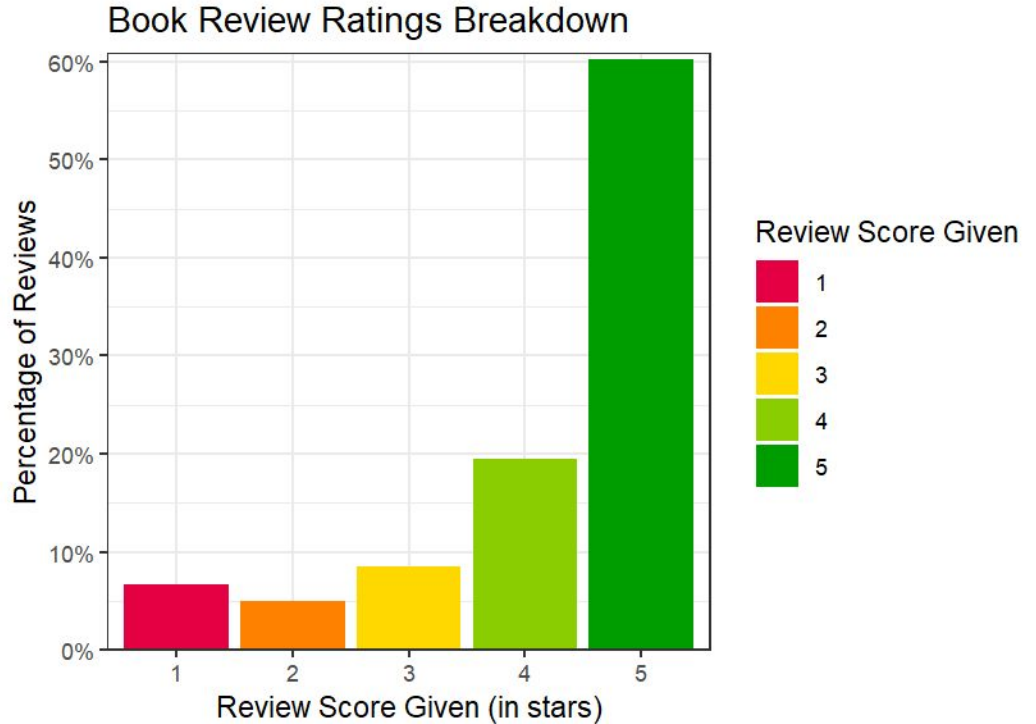


REVIEWS



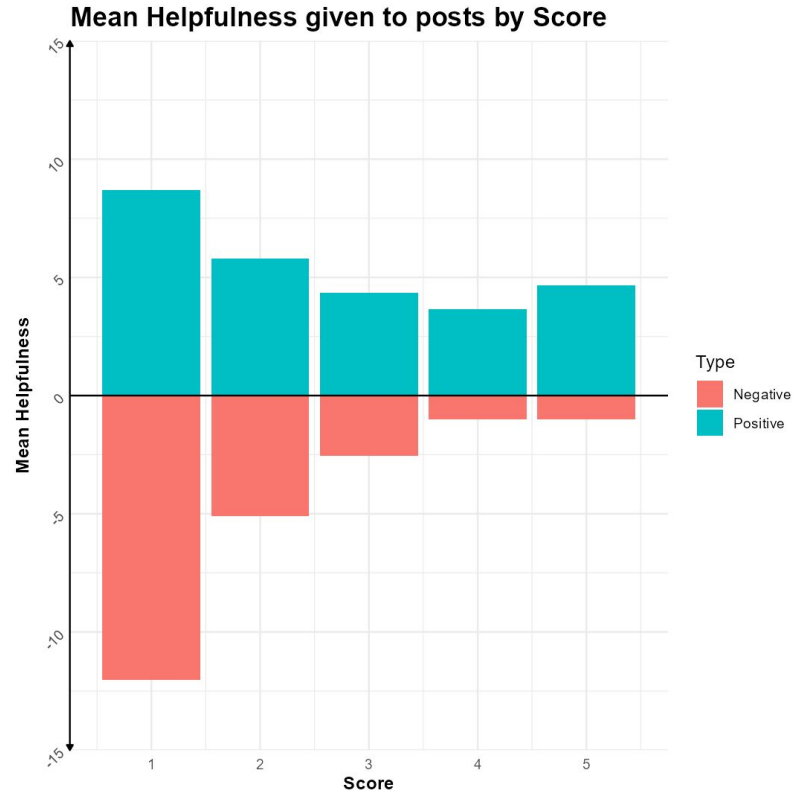
RATINGS/SCORES

- 5 star ratings make up more than half of all ratings
- This could suggest that users are more likely to rate a book they think highly of



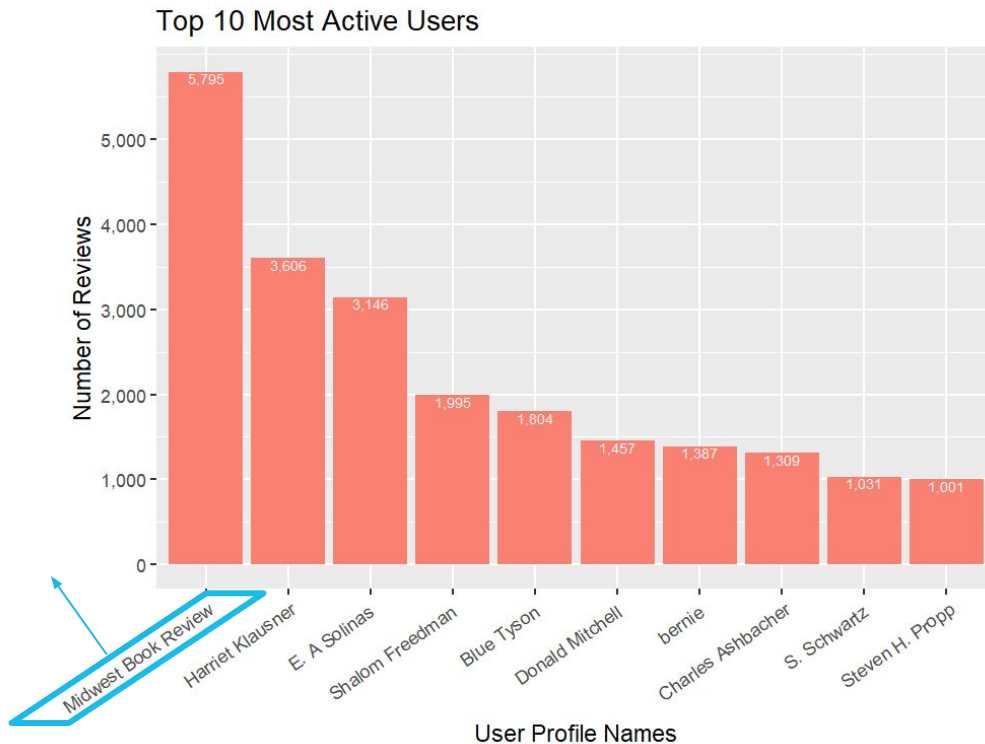
RATINGS/SCORES

- On the contrary, users *interact* with negative reviews more than positive reviews
- This could suggest that controversy heightens interest/engagement among users



NETWORK

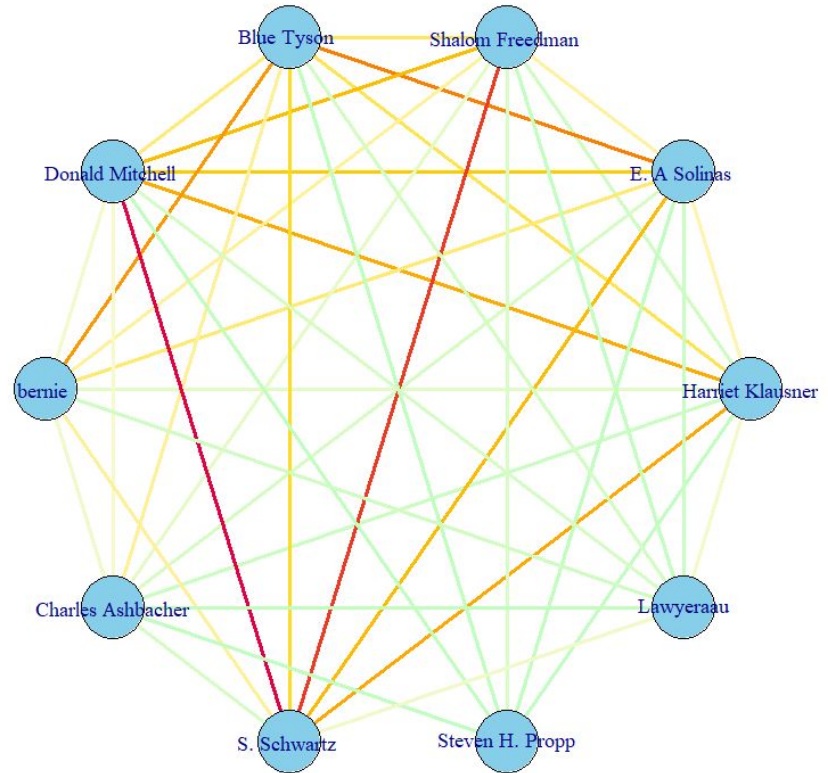
- “Midwest Book Review” is seen as the most active user, but why?
- Causes disruption of the data
- Shared accounts are not prohibited per Amazon guidelines



Network of Top 10 Users and Shared Book Reviews

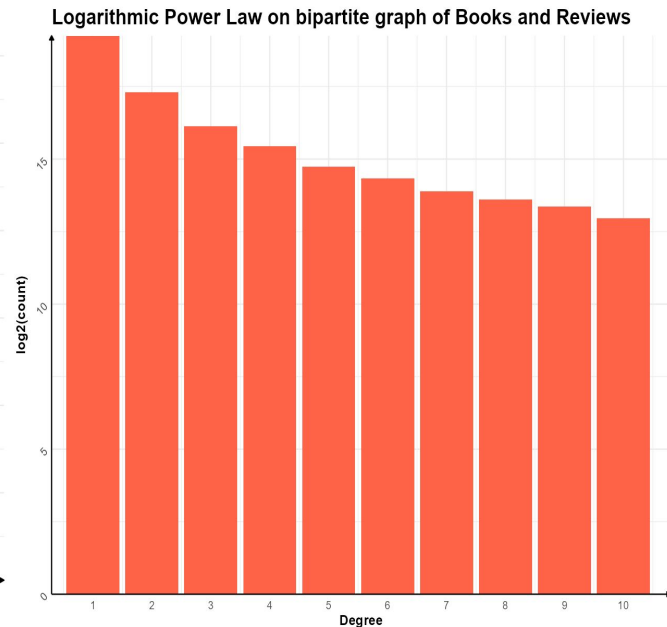
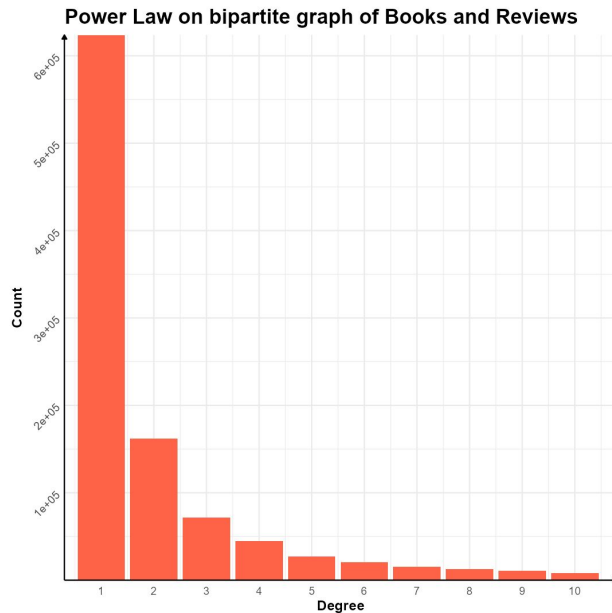
NETWORK

- Closer to **red** indicates more reviewed books in common
- This could be proof of concept for a friend recommendation system based on books read in common



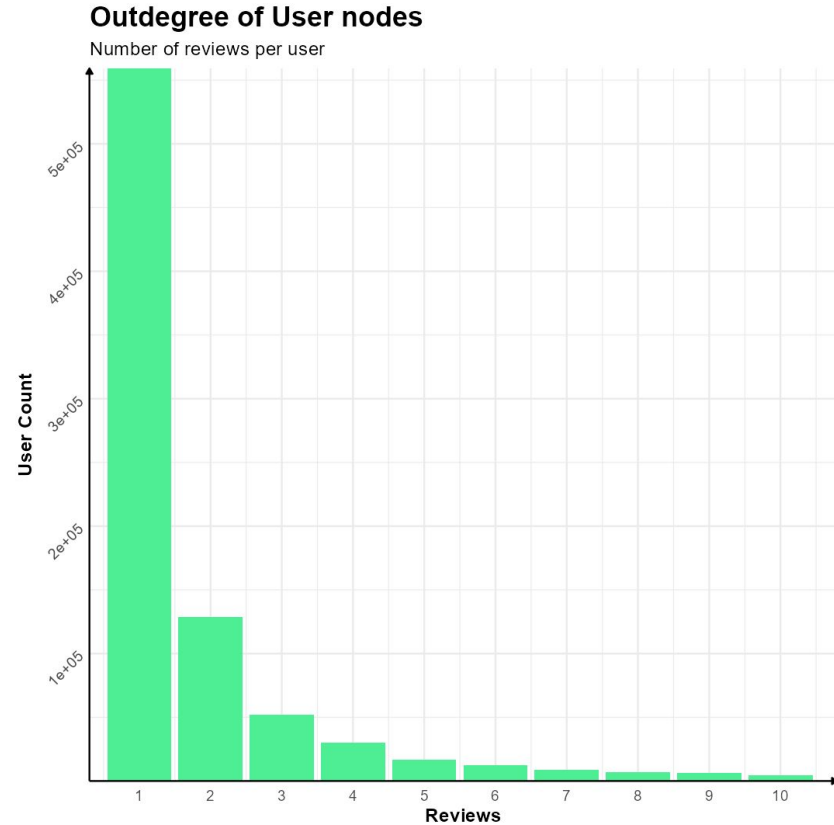
POWER LAW

- Plotting the distribution of connections per node shows the emergence of **Power Law**
- On the logarithmic scale we can see a quasi-linear distribution

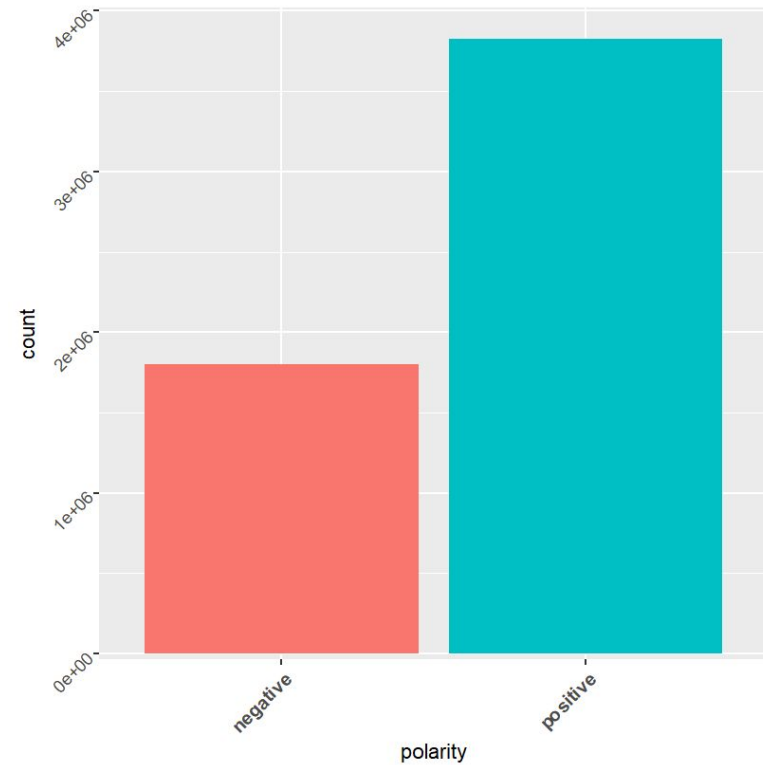
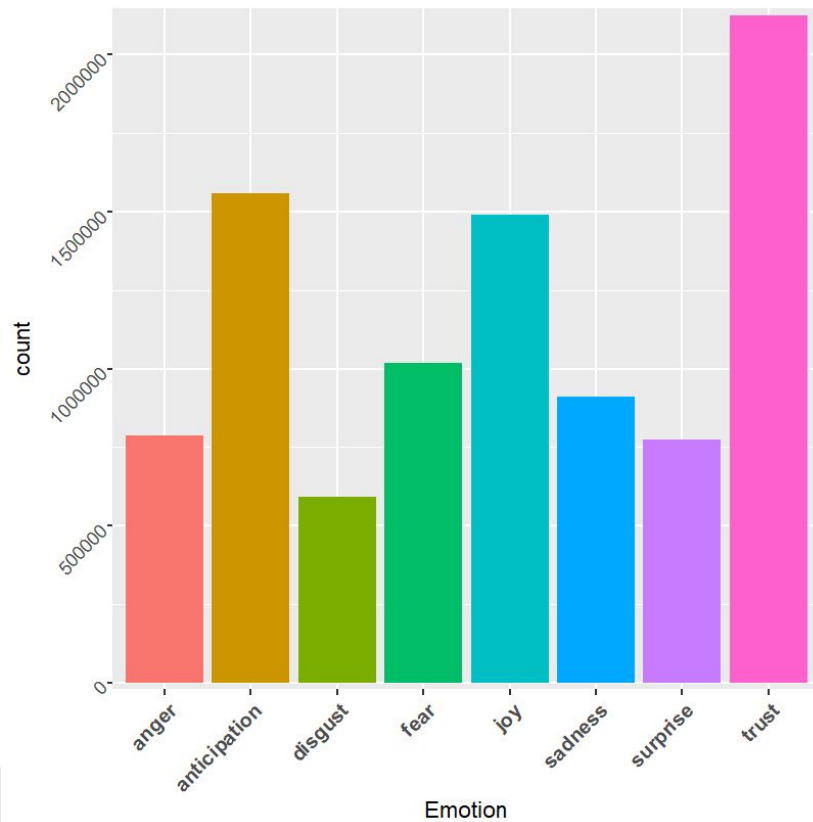


POWER LAW

- Power Law is present also in the number of reviews per user, which is the outdegree of user nodes



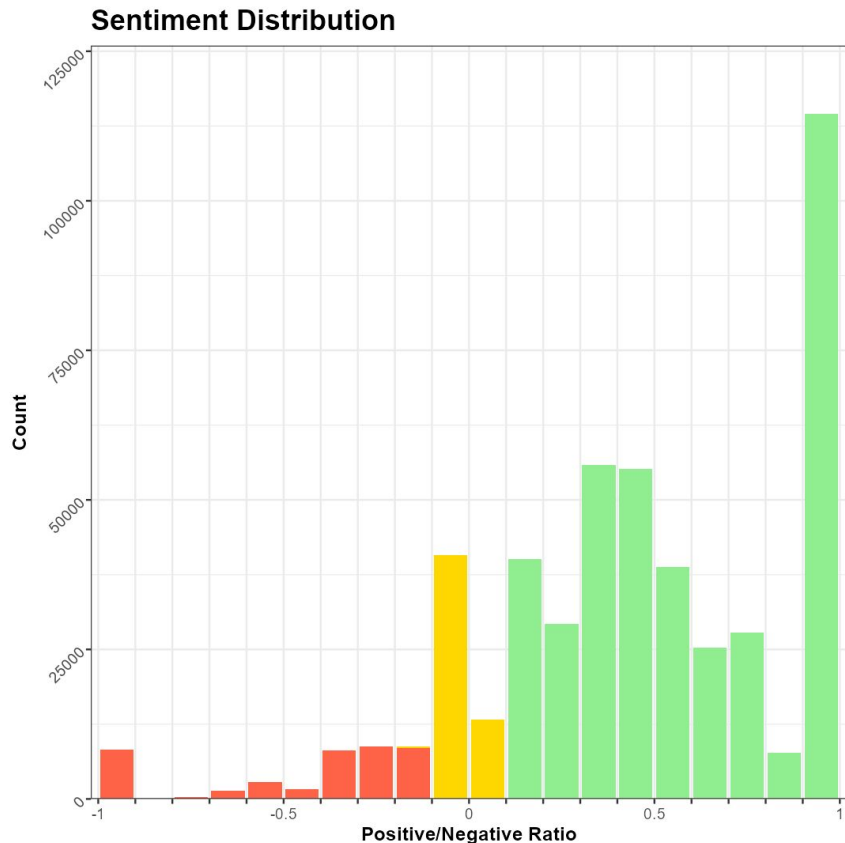
SENTIMENT ANALYSIS



SENTIMENT ANALYSIS

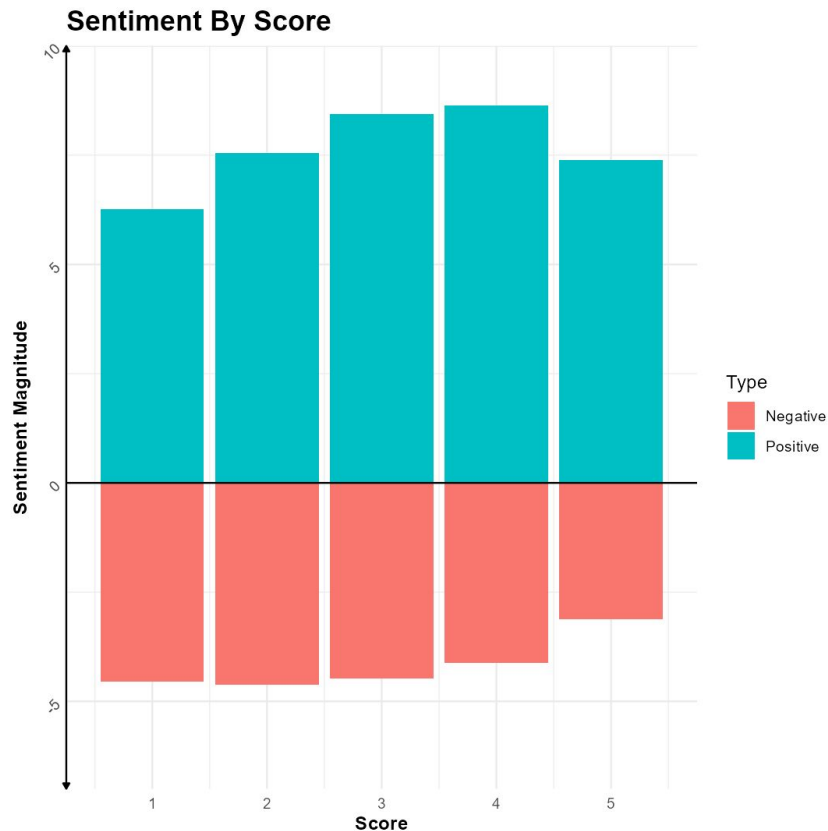
Sample: 500K reviews

- A general trend of positive sentiment in the reviews
- The graph supports the point that users are more likely to rate a book they think highly of



SENTIMENT ANALYSIS

- Reviews can mention positive and negative aspects of a book
- Users' sentiment isn't exclusively correlated with the rating they give based on the consistency in sentiment magnitude






05 **CONCLUSION**





CONCLUSION

- Users leave reviews when they think highly of a book
 - Despite this, negative reviews tend to get more interaction by other users on the platform
 - Limitations of the dataset and the analysis are present
 - Noise in data will always be present in human data because humans are unpredictable
 - Other limitations such as memory issues
 - Possibilities are endless
 - Friend recommendation, sentiment filtering and other features can be implemented
- 

THANK YOU

All materials are publicly available on GitHub.
Dataset available on Kaggle.com





SOURCES



“Amazon Books.” *Amazon Books | Brands of the World*, 10 June 2018,
www.brandsoftheworld.com/logo/amazon-books.

Ante, Spencer E. “Amazon: Turning Consumer Opinions into Gold.” *Bloomberg.Com*, Bloomberg, 15 Oct. 2009,
www.bloomberg.com/news/articles/2009-10-15/amazon-turning-consumer-opinions-into-gold.

Bekheet, Mohamed. “Amazon Books Reviews.” *Kaggle*, 13 Sept. 2022,
www.kaggle.com/datasets/mohamedbakheta/amazon-books-reviews/data.

Zufelt, Megan. “10 Largest Book Sellers Online in the USA.” *American Print & Bindery*, American Print & Bindery, 22 Feb. 2024,
americanprintandbindery.com/blogs/print-bind-closer-look/10-largest-book-sellers-online-in-the-usa.

