Natalia Uruska and Alison Reikher

IST 652- Scripting for Data Analytics: Final Project Report

March 15, 2022

Introduction

In recent years, there has been controversy surrounding certain types of books. Many well-known titles have been disappearing off library shelves at rapid rates, especially those deemed inappropriate to children and teens. This includes books that touch upon subjects such as violence, sexuality, abuse, and racism. There are also many banned books surrounding support for extreme political parties, such as communism or fascism. In some states, librarians could even potentially face criminal charges for providing access to certain challenged books (Sye, 2022). Book banning efforts are rapidly spreading across the United States and will likely continue to do so in the coming months and years.

While some books may be banned from certain libraries and schools across the country and the world, it doesn't stop retailers like Amazon and Barnes & Noble from selling them. Since these books are widely discussed, more and more people are purchasing and reading them for themselves.

This report explores the differences between Amazon and Barnes & Noble readers regarding banned books. Banned books are often discussed in the context of whether people should be able to read them or not, but are they worth reading, to begin with? To gain insight into the quality of literature, the ratings and reviews are analyzed in different ways. The goal is to gain a better understanding of the books and the readers who use Amazon and Barnes & Noble.

About the Data

The dataset being evaluated for this project contains information on a variety of 20 banned books, which were selected at random from Davenport University's Banned Books Week Library Guide[1]. The list of chosen books and respective authors is as follows:

Title	Author
The Absolutely True Diary of a Part-Time Indian	Sherman Alexie
Captain Underpants (book 1)	Dav Pilkey
Thirteen Reasons Why	Jay Asher
Looking for Alaska	John Green
George	Alex Gino
And Tango Makes Three	Justin Richardson and Peter Parnell
Drama	Raina Telgemeier
Fifty Shades of Grey	E. L. James
Internet Girls (series)	Lauren Myracle
The Bluest Eye	Toni Morrison
The Kite Runner	Khaled Hosseini
Hunger Games (Trilogy)	Suzanne Collins
I Am Jazz	Jazz Jennings and Jessica Herthel
The Perks of Being a Wallflower	Stephen Chbosky
To Kill a Mockingbird	Harper Lee

[1] https://davenport.libguides.com/bannedbooks/top10

Bone (series)	Jeff Smith
The Glass Castle	Jeannette Walls
Two Boys Kissing	David Levithan
A Day in the Life of Marlon Bundo	Jill Twiss
Sex is a Funny Word	Cory Silverberg

In addition to the title and author, other information was also collected from the respective books' pages on Amazon, and on Barnes & Noble's websites[2]. These fields and their corresponding descriptions can be found in the table below:

Fields	**Description**
Title	Title of book
Author	Author of book
Genre	The book's genre
Year Published	The year the book was published
Average Rating	The book's average rating (on scale 1-5)
# of 5, 4, 3, 2, and 1 stars	The number of corresponding stars for each book (each number is in unique field)

[2] https://www.amazon.com/books-used-books-textbooks/b?ie=UTF8&node=283155
https://www.barnesandnoble.com/

5, 4, 3, 2, and 1 star comments	Excerpt of comment from corresponding star review for each book (each comment is in unique field)
Reason for ban	The reason the corresponding book was banned
Movie or TV adaptation	True or False based off it the book was adapted for film or television

As a crucial first step to working with the dataset, preprocessing is required to ensure the data is ready for exploratory analysis. The data was initially collected in Excel (CSV file) and then split into two separate files, one corresponding to Amazon, and the second to Barnes & Noble. After importing the necessary libraries in Jupyter Notebook (pandas, matplotlib, seaborn, numpy, csv, nltk, and wordcloud), the two CSV files were read in and saved as dataframes. A column was later added to each of the two dataframes specifying whether each book was also based on a movie or TV show. Both dataframes were inspected to ensure there are no missing values. Missing entries would be replaced by an empty string as to not interfere with the various methods of analysis. The structure of each dataframe was then examined, and both have 20 rows and 17 columns.

Methods of Analysis

Research Question 1: Do users rate books differently on the different websites?

The first step in the analysis is to examine the difference between the average ratings of Amazon and Barnes & Noble books. It was discovered that they are incredibly similar. The average rating for the banned books on Amazon was 4.64 out of 5, and the one on Barnes &

Noble was 4.6 out of 5. This shows that audiences on both sites felt similar about the books overall and responded positively towards them.

Research Question 2: Are there any predominant book genres in the dataset?

As mentioned previously, the books for this analysis were chosen at random. It is interesting to know if any book genres are predominantly occurring in the data set. This finding was interesting, as there were 6 books out of 20 under the 'Young Adult Fiction' genre, 3 under "Children's Literature", 3 under "Graphic Novel", and then 1 book in each remaining genre. This indicates that many of the banned books within the dataset are meant for a younger audience.

Research Question 3: What are some strong correlations and how do they compare to each other in the two data sets?

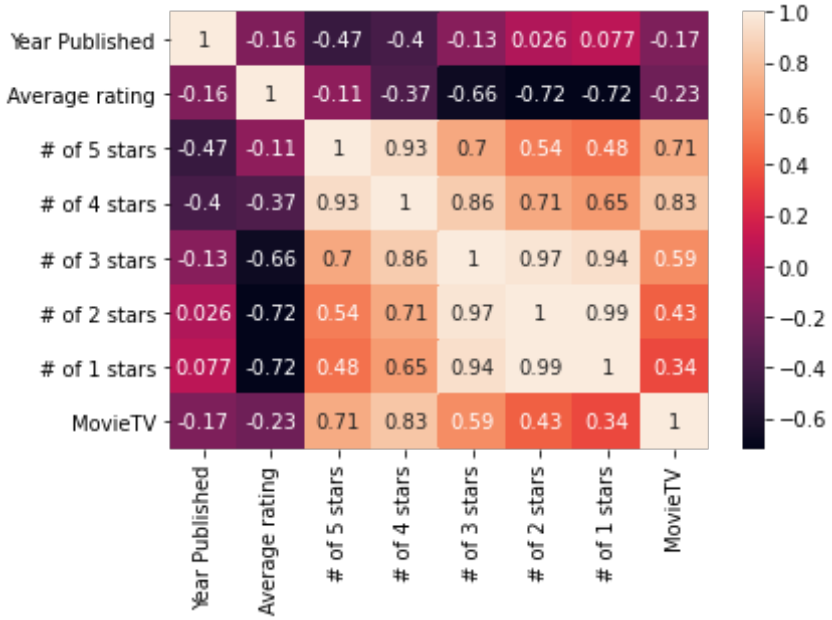

Figure 1

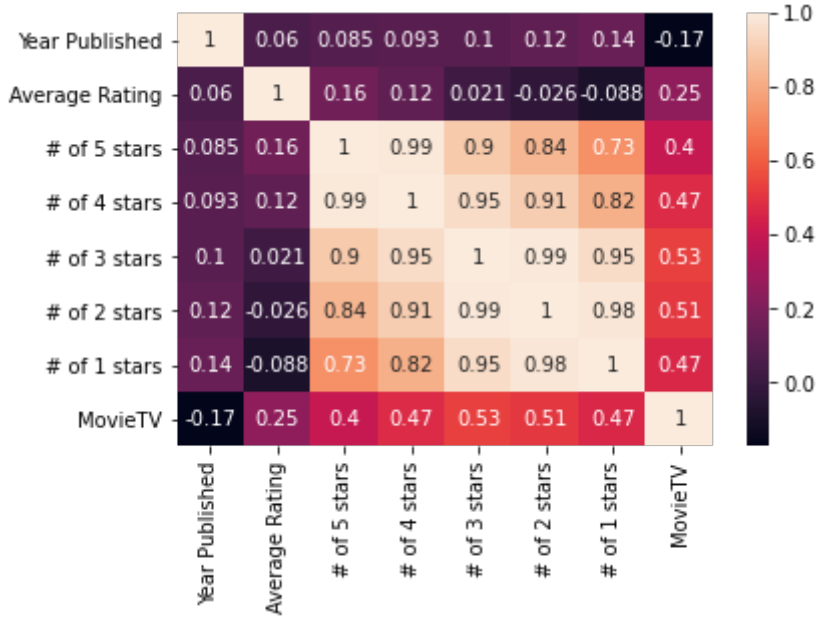

Figure 2

As a next step in the analysis, a correlation matrix was created to find any significant correlations among the variables in the datasets. Figure 1 shows the correlation matrix and heat map for the Amazon dataframe and Figure 2 shows the correlation matrix and heat map for the Barnes & Noble dataframe. One discovered finding was that all books which were based on movies or tv shows had stronger positive correlations with their respective ratings on Amazon than the ratings on Barnes & Noble. For Amazon, MovieTV had a 0.83 correlation coefficient with number of 4 stars and a 0.71 coefficient with number of 5 stars. The highest correlation coefficients with Barnes & Noble for MovieTV had a correlation coefficient of 0.53 with number of 3 stars and a 0.51 correlation coefficient with number of 2 stars. This is possibly because Barnes & Noble consumers are usually avid readers who probably don't consider if the book is based on a movie before purchase, whereas Amazon consumers typically shop for other items on

Amazon. There were also strong positive correlations between some of the neighboring star ratings, such as the number of 4 stars and number of 5 stars having a correlation coefficient of 0.93 for Amazon and 0.99 for Barnes & Noble, which was to be expected. Neither correlation matrix showed a significant correlation between Year Published and any of the other variables. This indicates that the year had almost no effect on the other variables.

Research Question 4: How do the reviews for Amazon compare to the reviews for Barnes & Noble?

This research question can be analyzed in two different ways. The first form of analysis is through word clouds and word frequency analysis. A dictionary was created for each dataframe that contained the words for the reviews from their respective star reviews with the value being the most frequently used words. The stopwords were removed from the comments to get only the relevant terms.

Amazon	Barnes & Noble
5 Stars:	5 Stars:
book : 25	book : 19
read : 17	read : 9
story : 9	make : 7
think : 9	it. : 7
it. : 9	many : 6
even : 9	would : 5
reading : 7	one : 5
like : 7	story : 5
would : 7	young : 4
books : 7	like : 4
4 Stars:	4 Stars:
book : 30	book : 24
read : 13	read : 13
young : 11	book. : 8
would : 11	really : 8
story : 11	good : 7

book. : 9 little : 6 much : 6 think : 6 life : 5	liked : 7 would : 6 people : 5 like : 5 reading : 4
3 Stars: book : 22 one : 13 girls : 9 read : 8 first : 8 see : 7 story : 7 think : 6 nothing : 6 school : 5	3 Stars: one : 7 kids : 6 book, : 5 first : 5 reading : 4 girl : 4 school : 4 like : 4 book : 3 would : 3
2 Stars: book : 26 read : 9 kids : 9 book. : 8 like : 7 good : 7 children : 5 love : 5 think : 5 would : 4	2 Stars: book : 15 read : 6 like : 5 page : 4 good : 3 native : 3 however, : 3 teen : 3 reading : 3 took : 3
1 Stars: book : 25 would : 9 kids : 8 reading : 6 book. : 5 teach : 5 like : 4 young : 4 character : 4 could : 4	1 Stars: book : 21 book. : 8 would : 8 reading : 7 things : 6 get : 6 even : 6 read : 6 many : 5 like : 5

Consistently among all the comments, the words book and read are used most frequently. Some of the words used often describe the target audience like kids, young, or girls. Depending on the rating, it would most likely have to do with how appropriate that book is for that demographic.

While word frequency can give some insight, there are still a lot of other keywords that are not included in these top 10 lists. Word clouds feature all the important words that stick out and can give more insight.

5 Star Amazon comments

5 Star Barnes & Noble comments

Looking at the 5 star word clouds, both have positive words like love and good. They also both describe the audience with words like teen, young, junior, and girl. There aren't many differences between these word clouds.

3 Star Amazon comments

3 Star Barnes & Noble comments

The 3 star word clouds aren't that much different from the 5 star. Again, we see the word book prominently featured. The biggest difference is that now the word girl is featured more prominently specifically in the Amazon word cloud. The Amazon word cloud largely focuses on the audience the book is for while the Barnes & Noble word cloud focuses more on the book itself. Words like girl, boy, banned, behavior, and think suggest a judgment as a whole of who the primary audience is or should be for the Amazon books. Words like story, adventure, love,

pretty, time, work, understand, question, and storytelling suggest a judgment of the content of the book itself from the Barnes & Noble reviews.

1 Star Amazon comments

1 Star Barnes & Noble comments

The 1 star comments use more negative words as expected. The word book is still featured prominently. Like the 3 star word clouds, Amazon comments are more criticizing or discussing the audience for the books while Barnes & Noble comments focus more on the book itself. The Amazon word cloud uses words like girl, kid, inappropriate, recommend, sexual, experience, dirty, and grade. These words suggest a discussion about who this book is or isn't for. Many of the reasons why the books were banned revolve around the discussion of books being inappropriate for certain ages. Topics around sex and sexuality are primary reasons for books being banned and these comments reflect that sentiment. On the other hand, Barnes & Noble word cloud used words like reading, written, recommend, review, sad, boring, depressing, and bad. These words clearly describe the book itself. The Barnes & Noble reviewers are taking the audience out of the equation and reviewing the book for what it is.

The second way to analyze the reviews is through Sentiment Analysis. Since the comments were associated with a certain star rating, Sentiment Analysis can be used to analyze how positive, neutral, or negative the comment is. The expectation is that 5 star comments have more positive sentiment, 3 star comments have more neutral sentiment, and 1 star comments have a more negative sentiment.

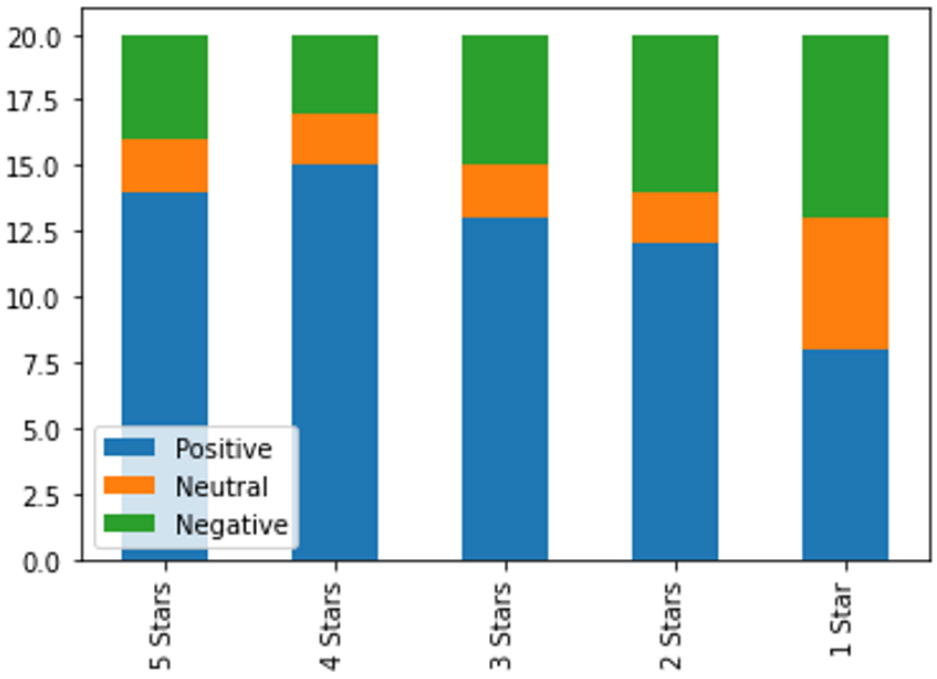

Figure 3

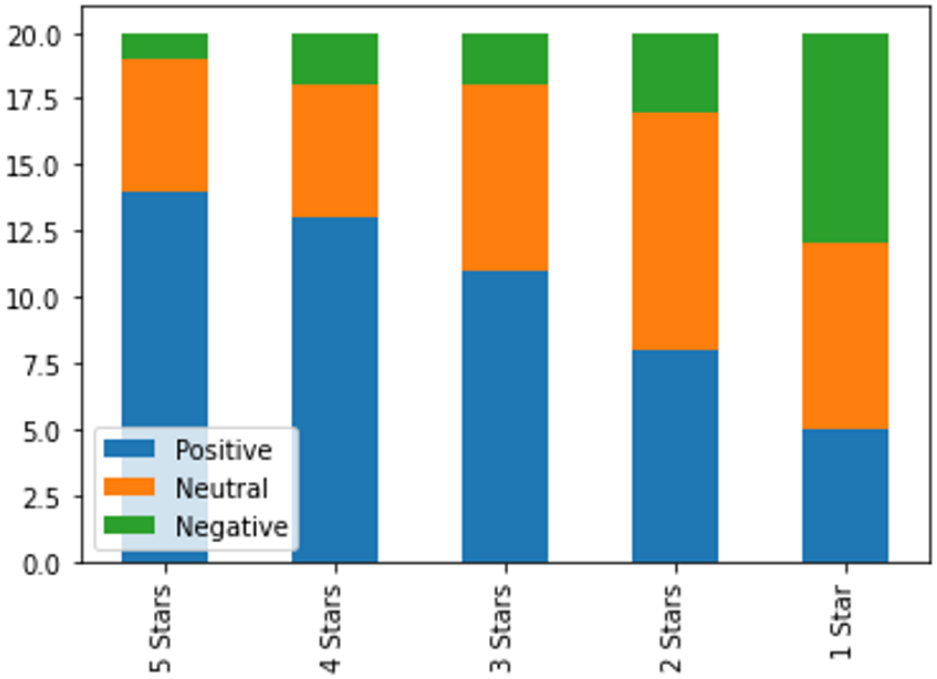

Figure 4

Figure 3 shows the stacked results for the Amazon dataframe and Figure 4 shows the stacked results for the Barnes & Noble dataframe. Looking at the Amazon dataframe, the sentiment was often either positive or negative. The number of neutral comments were slim. While the positivity levels decrease over the stars, it is surprising that there are more positive reviews for 4 star than 5 star. The positive sentiment is consistently higher than any other sentiment among all the five stars. The 5 star distribution starts with 14 positive, 2 neutral, and 4 negative comments. The 3 star distribution has 13 positive, 2 neutral, and 5 negative comments. The 1 star distribution is the only place with the evenest distribution of the sentiment with 8 positive, 5 neutral, and 7 negative comments. Figure 3 demonstrates that Amazon reviewers tend to use either positive or negative words in their reviews with a tendency more towards the positive.

On the other hand, Barnes & Noble results are more of what was expected. The number of positive comments slowly decreases over each rating while the number of negative comments increases. The neutral comments remain consistent throughout the distribution. The 5 star distribution starts with 14 positive, 5 neutral, and 1 negative comment. Then the 3 star distribution has 11 positive, 7 neutral, and 2 negative comments. The 1 star distribution has 5 positive, 7 neutral, and 8 negative comments. Figure 4 demonstrates that Barnes & Noble reviewers are more critical of the books and tend to lean more towards neutral language.

Description of Program

As a first step in creating the program, the necessary libraries were imported. The data was loaded from CSV files into two separate pandas dataframes in Jupyter Notebook, and

initially explored for any anomalies, such as missing data. After performing thorough cleaning on the data, the two dataframes (Amazon and Barnes & Noble) allowed for a more manageable and clear-cut analysis. Exploratory data analysis was performed to compare average book ratings in the two dataframes. The frequency of the different genres across the books was also determined using a simple for loop. To determine our strongest correlations, two correlation matrices were created using .corr() function and plotted using a heatmap to pinpoint the most significant ones. The lambda function was used to remove all the stopwords from the review data. The nltk library was imported for sentiment analysis, which was performed on each of the dataframes to determine the sentiment of each star's comment (positive, neutral, and negative). A histogram was created using the plt.bar() function to visualize the frequency of the sentiments. A stacked bar plot was also created using the plot.bar() function to easily visualize the overall sentiment within each star comment. Lastly, the wordcloud library was used to create word clouds to visualize the most occurring words included in each of the sentiments, and the reasons for ban. A count of most frequently occurring words per sentiment was created using a counter() function and a for loop.

Conclusion

As previously discussed, books can be banned and challenged for a variety of reasons. The results show there are many different aspects to how a book is rated based on which website the user is purchasing books from even though they have similar average ratings. Amazon raters often took into account outside factors when rating and reviewing the book. They were much more likely to give the book a 5 or 4 star rating if the book was adapted for film or television. They also tended to include discussions about why the book should or shouldn't be banned.

Since the typical reason regarded how appropriate the book was for the intended audience it was written for, the reviews focused more on that rather than the quality of literature itself. They also tended to be more positive in their reviews even when giving a 1 star review.

Barnes & Noble readers tended to review the book from a more literary perspective. The reviewers were unaffected if the book was adapted for film or television. If anything, they were more likely to give a lower rating for that. The reviews tended to discuss more of the story itself rather than who the story was written for. The Barnes & Noble reviewers used more critical and neutral language than Amazon reviewers.

One important note to highlight is that there are some limitations to working with this dataset. The 20 books were randomly selected, and the corresponding data were manually collected. There is a chance that some of the data could have been the result of bias. For further exploration, a larger sample size would need to be collected.

Overall, the two dataframes have both similarities and differences. Both have their merit for their respective audience. The average consumer who is curious about a banned book that might have been adapted for film or television will go on Amazon while the avid reader will go on Barnes & Noble. Both platforms continue to support access to literature for those who would not be able to access it otherwise.

Team Breakout

The data for this project was manually collected and analyzed by Natalia Uruska and Alison Reikher. The use of the data relating to banned and challenged books was mutually agreed upon by both team members. Alison and Natalia came up with all the research questions together, and

both contributed equally with ideas on how to analyze all questions. The data collection task was split among the team members with each collecting data for 10 books. The code for the exploratory data analysis and the research questions was also done by both team members, each taking turns and coming up with different ideas on how to take an initial look into the data and perform any necessary cleaning based on the questions which have been created. Lastly, the code for each question was reviewed by both team members and minor changes were made, run, and debugged. Both team members approved of the final code which was executed successfully.

Data Sources

- Ideas for banned books were obtained from Davenport University's Banned Books Week Library Guide: https://davenport.libguides.com/bannedbooks/top10
- The banned book ratings and reviews from Amazon and Barnes & Noble were manually obtained by performing book searches via the following URLs:
  https://www.amazon.com/books-used-books-textbooks/b?ie=UTF8&node=283155
  https://www.barnesandnoble.com/

Word Cited

Sye, D. (2022, March 8). *Beyond book banning: Efforts to criminally charge librarians*. Intellectual Freedom Blog. Retrieved March 13, 2022, from https://www.oif.ala.org/oif/beyond-book-banning-efforts-to-criminally-charge-librarians/