Once Upon a Vis

Abby Carr   Anika Das   Riya Gurnani   Veronica Aguiar Ortega

Northeastern University

ABSTRACT

As more of our time is spent in the digital world, streaming services and social networking apps have almost entirely replaced books. Despite the availability of e-readers such as Kindles, reading has become less and less common. One way this shift can be reversed is by giving people resources to find new books to read. The general area of interest for our project is the ratings and rankings of books, and how they compare to one another. We intend to address the comparison of popular books and allow book readers to get a sense of similarities and differences between books. This will potentially give them new books to read based on ones they have enjoyed in the past. We believe this visualization will address an important topic as reading improves one's focus, memory, empathy, and communication skills.
https://github.com/DS4200-S22/final-project-once-upon-a-vis

1 INTRODUCTION

Our project topic revolves around popular books and how they have been rated over a span of years. We hope that the visualizations that are a part of this project will serve as a resource to both those who enjoy reading, as well as those who read less often. Our end users should be able to learn from our visualizations and ideally find books that they want to read. Along with this, we intend for this website and accompanying visualizations to be useful, interesting, and relevant to those who are interested in seeing how certain books have been rated and reviewed over a period of time. The major resource available for book lovers currently is Goodreads. While we appreciate and value all of the resources present on their website and in the Goodreads community, we noticed a lack of visualizations on their website. We think visualizations could add a lot to the way people view books and would be of great interest to those who already know these books. Our motivation for this project is to provide all readers, of all levels, an easy way to compare books based on their reviews and ratings. This pursuit is impactful and important as few resources like this exist, and it will hopefully open the world of books to more visual people.

2 RELATED WORK

This topic has been researched plenty of times in the past, especially more recently due to the ability to collect large quantities of data with the increased dependence on the internet for the literary industry. Many previous projects have collected data on popular books over several years to better understand trends or answer questions about the topic. Our first related project reviews how the internet has changed the way that books are consumed, specifically science novels. It tracked data on the rapid increase in users utilizing the internet to not only purchase books, but also to find reviews and similar books. It delves into more factors of public consumption data and how this can help understand the different types of current popular science interests and values. While our project will be focusing on a wider variety of books, not just science ones, the visualizations in their project are very informative for how we can work on ours. They stress how much visualizations can show about a topic. We can consider expanding our visualization to include book popularity over a span of time to see the general trends in genre and other categories. It is also important to consider the attribute of internet usage over time, as they have done since this is where we are collecting data from. This does not cover the popularity of print sales, necessarily, which are not as prevalent now [1].

Another project collected data from Amazon's 50 best books during the years 2009-2019. It then tracked various attributes about those books, such as genre and author, to create informative visualizations and data findings. Our project is similar to this one, except we are planning to go more in depth than just learning about the characteristics of popular books. We are also going to be considering other factors such as user reviews and what a sentiment analysis of those can show us. This project shows us the types of visualizations we can create from the data and how we can use this to discover new findings. For example, instead of just creating a visualization with author frequency in this list over the years to demonstrate popularity, it delves deeper by also including the count of their unique books and appearances. This creates more informative visualizations beyond the pie charts and bar graphs they began with [2].

Another project analyzes user reviews on Goodreads to understand how they define and discuss what they consider to be "classic" literature. It also goes into the backgrounds of these books to find patterns with genre and history of the books, while considering how the users review them. Most of the "classics" were considered books that were taught in school, but this does not guarantee a high rating from users. This paper using user reviews is similar to the direction we wish to go for our project; however, they do not run a sentiment analysis of these reviews. As this paper shows, quantifying user reviews and comparing them over time with a book's rating and other characteristics can help to discover a lot of information. This project discovered trends in contemporary literature and discussed how we can utilize this new data in the future. We can use the computational methods they discussed to help gather and organize our data to create our visualizations [3].

Another related work analyzed if book reading behavior on Goodreads can predict Amazon best sellers. The project began trying to find correlations between reading behavior and bestselling books on Amazon to create a predictive model. While the scope of that is different from our project, they collected interesting data and found relationships for visualizations that are helpful to inform our work. They track

E-mails: [carr.ab | das.ani | gurnani.r | aguiarortega.ve] @northeastern.edu

various features including number of status updates from the readers and the average sentiments from the status posts. They create multiple grouped bar charts to depict the correlation between a bestseller and that book having multiple status updates. We can see how they found this relationship through visualizations, which is helpful for our project. We also see how they utilized the sentiment analysis from the posts [4].

Our final related research paper compares and contrasts various features of printed books versus electronic books. A theoretical model is constructed to study their coexistence, and the ideal price for each is found to maximize the profits for the companies involved (the publisher, printed book retail store, and electronic book retail store). This related work delves deep into economics and the dual-channel supply chain with regards to publishing. Our project will not have an economic or financial focus. However, this previously accomplished work includes a model which compares printed books to electronic books. We believe this could be an interesting attribute to look at for all the books that end up being in our data set. Perhaps there will be biases towards either printed or electronic books. We could also take into consideration the retail price for the books in our data set and see if there is any correlation between retail price and book popularity/ reviews. This previous effort and research done in our area of interest has given us a few more ideas of attributes to consider when it comes to the books and novels, we will base our project on. We can extend the ideas put forth and carried out in this paper to our use case, and ideally create interesting, informative visualizations from our data [5].

3 USE CASE

Our project aims to create a visualization depicting a set of books, and how their popularity has changed over time. We also want to include filters such as author, genre, length, etc. that could give users some flexibility in finding specific information in our data set. We imagine that our visualization(s) can be used by book lovers and amateur readers alike. A specific scenario in which our visualization tool could be used is by a young adult who is looking for a new book to read. They will be able to visit our website that hosts our visualization and see line charts and graphs which represent popularity/ reviews of the books in our data set. They will also be able to view similar books to ones in the data, whether this similarity be in author, genre, rating, or another attribute.

Another use case would be a reader who is interested in knowing how some of the most popular books are ranked in terms of sales. As we were conducting our research for this project, we learned that publishing houses rarely, if ever, make sales data available. One of the few sources that does collect book sales data is Bookscan, a private data provider that allows access to different datasets, but a single subscription costs 2,500 dollars a year. Through our visualization, users will be able to see and compare cumulative book sales of books such as the Harry Potter series, The Boy in the Striped Pajamas, and classics like To Kill A Mockingbird. The visualization will help users quantify and put into perspective the success and influence of some of their favorite books.

4 DATA

Our collected data is stored on GitHub and will be updated as our cleaning and collection moves forward. In 'top100_goodreads.csv' we hold the bulk of the data columns we plan to use; The columns collected come from two different sources. Nielsen Bookscan is a book tracking service centered in the UK that collects book transaction data. In the article from the Guardian, Neilsen provided a snippet of their data in the form of the top 100 books of all time through the period 1998-2010. Data includes categorical data about the books such as Author and Publisher, but also provides interesting data surrounding the value of the book in total sales and in general monetary worth [6]. Alongside the data from Neilson, we use Goodreads data scraped using the scraper utilized in Antoniak's paper, "The Goodreads 'Classics': A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism." With the use of this tool, we added Goodreads specific ratings, genres, and shelf data to the table we had originally sourced [4]. Still using the Goodreads scraper, we have also collected up to 300 reviews and their metadata for each of our top books. These are stored in the 'reviews' folder in the repository, currently individually stored in their own Json files.

Our data sources span an interesting range of quantitative data, such as ratings and monetary value, as well as our qualitative review and genre data. Although the data includes much metadata on the books, our main goal relies heavily on the sentiment of book reviews and focuses on the way in which we review books. It will be very important to continue our data processing on the reviews, as this data will provide the bulk of the uniqueness our data can provide to users. Nielson's data is very focused on tracking the economic movement of books but is focused on the UK. Sale data for books is hard to come about, and this source provides a reliable and well-known source of numeric and base data for us to build on. Although we rely on the top 100 books of all time, we can easily draw on the additional pages of popular books also available on the Guardian website. The data may be biased towards UK-sales but can still provide an important base performance of the books. On the other hand, Goodreads data is incredibly subjective; this subjectivity, however, is the point of the Goodreads data. Numerical ratings alongside reviews and top shelves lean into the way the audience feels about books and is meant to challenge this project to show book ratings in a different way. The sentiment analysis to be performed on reviews is meant to home in on the way the public feels about a particular book. Lastly, collecting the data gives way to bias and difficulties. Our base 100 novels had to be reduced because different versions of the same Harry Potter novels being included in the top 100 and books that did not appear in our Goodreads search were left behind as the generalization of our methods stop some specific books from potentially being recognized by the Goodreads website. In addition, when scraping reviews, ads and malfunction kept the full 300 reviews to be collected per book from fully being collected. We have overcome data collection and extraction errors to ensure all our data is fully available, but difficulties in coding errors when increasing the amount of data, we collect may be frustrating to work through. In addition, there is a possibility that the sentiment analysis tools we test will not cater to the review data we have, and it may take extra time and lower expectations to fulfill the sentiment analysis for book reviews.

5 TASK ANALYSIS

Task ID #	Domain Task	Analyze Task (high-level)	Search Task (mid-level)	Analytic Task (low-level 'query')
1	I want to know which genre(s) of books tend to get the highest ratings.	Consume → Discover	Browse	Summarize
2	I want to see what the general public rates HP book	Consume → Discover	Locate	Identify
3	I want to investigate which book in a series is the most popular	Consume → Discover	Locate	Compare
4	I want to compare two books based off of their statistics and public opinion	Consume → Discover	Lookup	Compare

The primary consumer of our visualization will be readers who are interested in the ratings and reviews of books. A possible secondary consumer could be people who work in the publishing industry since our visualization can help them see and understand the current trends. Our visualization will be primarily developed for Discover consumption, as its purpose will be to allow users to discover new information about their favorite books, or books they have yet to read. The information they will be able to discover includes the average rating of books as well as the sentiments of reviews for these books. We plan to include access to information based on quantitative data such as the length of the book or average rating, categorical data such as genres, and derived statistics such as the sentiment analysis of reviews for the book. A user will be able to discover trends between different types of books and characteristics about them through general explorations. An example aspect of our visualization is enabling the discovery of which genres are the most popular among the Top 100 books, and overall patterns in popular literature material. One of our visualization views will provide a lookup functionality which will allow a user to select up to two books they wish to see the individual information of and will then present a working screen of the statistics of each chosen book. In this way, our visualization presents a large-scale exploration through overarching information and trends for all books while also catering to individual interest by allowing selections of individual books. Secondarily, the Discovery application of our visualization could be informally used as a Present application for small groups such as book clubs who may wish to talk about information we provide.

6 IMPLEMENTATION PLAN & PRELIMINARY WORK REFERENCES

The visualization tool we plan to create will be built using the capabilities of the D3 library. We plan to utilize colors and spatial elements to represent books and their sales and hope to explore book differences in reviews with a colored comparison of books.
To build this visualization, we will need to use web scraping capabilities such as BeautifulSoup and a web driver to collect data. We will need to leverage common Python tools such as NumPy and pandas alongside file readers to collect and distribute the data we need. Our project requires sentiment analysis tools and will use one-hot encodings to work with large amounts of categorical data in each book.

We will have several visual encodings to implement for our final project:

- We will have two bubble charts to show the number of five star and number of one star reviews for each book title in our dataset. The size of the shape (star for five star and thumbs-down for one star) will be dictated by the number of corresponding reviews (the more reviews, the larger the shape). Each bubble chart will also have a unique shape/symbol: gold star for 5-star reviews, and a red thumbs-down for 1-star reviews
- We will also have a percent-tone bar chart for a single book that a user can select. This will show the number of each type of reviews for the title (one, two, three, four, and five stars) based on the size of the bar. It will also have three colored portions in each bar (green, gray, and red) which represents the percent of those reviews that are positive, neutral, and negative sentiments.
    - There will also be a circular ring above this chart, where the percent of it colored will be the percent of reviews that are subjective.
- Another graph will be a line graph to represent the number of book sales by year. There will be points plotted as well for each book so the user can see more detailed information. The location of the line will represent the amount of money made that year. The location of the points represents the money made by that book in that year.
- We will have another bubble chart for the genres of the books in our dataset. Similarly, the size of the

bubble corresponds with the number of books that are classified as that genre.

We will also have several interactive components:

- There will be linking in our first bubble charts through the book selected and the highlighted star and highlighted thumbs-down.
- There will be brushing and linking between the points highlighted in the sales graph and the genres highlighted.
- The user will be able to select a book from a drop down menu and this will filter a book visualization to show information for the selected book.
- For the reviews graph, the column can be highlighted. Additionally, the green, gray, and red colored sections can be hovered over to view exact numbers.

The required items that we deem necessary for our final project are:

- The brushing and linking between the line graph representing total book sales per year and the genre bubble chart.
- Sentiment analysis reviews bar chart with the percentages of each category colored and with highlighting incorporated

Some items that are not required, but would be great to have include:

- Creating two separate pages
- Incorporating transitions when clicking on the book cover to the reviews graph
- Most, if not all hover interactions implemented after brushing and linking
- Linking between the 1-star and 5-star bubble charts and the selected book
- Shapes for the 1-star and 5-star bubble charts (for now, we will be doing circles)

7 VISUALIZATION DESIGN

The final visualization tool will be composed of two groups of symbols. Stars will be on the left representing the top 5 star reviewed books, while negative, thumbs down emojis will be on the right representing the books with the most 1-star reviews. Users will have the option to return to the entire bookshelf which contains the top 100 books dataset from Goodreads. If a user hovers over a star or a thumbs-down, the title of the book as well as the number of the 1 or 5 star reviews will be shown. We chose this visualization because it allows the user to compare certain statistics about the books, specifically the quantity of these ratings. The second section of the visualization will be a large book (specifically the book they choose in the drop-down menu to the left of the book). In this portion of the visualization, the user can hover over the different portions of the cover to see different information about the book such as the author, books average selling price, etc. They can also click on the sticker on the cover of the book which will change the page to more detailed book review statistics. We chose this design because it resembles what a real book might be like. The next portion of the visualization is a stacked bar graph that shows the reviews of the book divided by sentiment (green for positive, grey for neutral, and red for negative). If the user hovers over the bars they can see the exact information. Another part of the visualization is a line graph that depicts the total number of sales of each book each year since they have been published. We chose to implement this visualization because it allows users to view the sales trend line and compare different books through points below the line on the graph. The final portion of our visualization is a bubble chart of the genres of the top 100 books. The size of the bubbles corresponds to how many books of that genre are in our dataset. Users can also hover over a genre bubble to see the exact count. We thought this visualization would be a good choice because it provides a general overview of what kind of books are in our dataset.

REFERENCES

[1] Y. Wang and Q. Zhong, "Consumption Behavior of Popular Science Books Based on Big Data," 2020 *International Conference on Big Data Economy and Information Management (BDEIM),* 2020, pp. 79-82, doi: 10.1109/BDEIM52318.2020.00027.

[2] Kharwal, A. (2020, Nov 30). *Amazon Bestselling Books Analysis with Python*. Retrieved from The Clever Programmer: https://thecleverprogrammer.com/2020/11/30/amazon bestselling-books-analysis-with-python/

[3] Maity, S. K., Panigrahi, A., & Mukherjee, A. (2017). Book Reading Behavior on Goodreads Can Predict the Amazon Best Sellers. *International Conference on Advances in Social Networks Analysis and Mining* (pp. 451-454). IEEE/ACM, doi: 10.1145/3110025.3110138.

[4] Walsh, M., & Antoniak, M. (2021). The Goodreads "Classics": A Computational Study of Readesr, Amazon, and Crowdsourced Amateur Criticism. *Journal of Cultural Analytics*, 243-287, doi: 10.22148/001c.22221.

[5] Y. Cheng, C. Dong and R. Liu, "The coexistence of printed book and electronic book in a book supply chain," 2017 *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM),* 2017, pp. 1421-1425, doi: 10.1109/IEEM.2017.8290127.

[6] *Top-selling 100 books of all time*. (n.d.). Retrieved from The Guardian: https://www.theguardian.com/news/datablog/2011/jan/01/top-100-books-of-all-time

APPENDIX A: GROUP CHARTER

GROUP PURPOSE

The reason for our group's formation is our shared interest in books, reading, and comparing book ratings to one another. We all like the idea of using qualitative data to quantify various attributes and the ratings of books. It would also be interesting to learn more about books and how they are reviewed and rated.

GROUP GOALS

Our goals as a group are to learn more about how the different aspects of books are rated and use data from some books to create a meaningful visualization. We want to find data on our favorite books, books we have heard about, as well as titles that are new to us. Using high quality, well collected and documented data, we hope to learn what the public thinks about this group of books, and convey that information in high quality visualizations and a thorough and interesting website, explanation, and presentation. We have a mutual understanding that we will all work to the best of our abilities and supply a high level of/ peak performance when it comes to this project. We are collectively aiming for an A, and will work smart, as well as hard when it comes to dividing up and completing the various portions of this project.

GROUP MEMBER ROLES/RESPONSIBILITIES

Each of our group members will share the same general responsibilities. These responsibilities include being responsive to one another, texting back timely, sharing availability, planning, and attending meetings, and completing our assigned portions of the work. Our group members will also have unique roles and responsibilities when it comes to splitting up the work. We plan to meet at least 2-3 times a week, and at each of our meetings we will divide up work for the next few days and report back to each other on our progress at the next meeting. For example, during our first meeting we decided that Abby and Veronica will spend the next few days gathering and cleaning quality data, Anika will write up the group charter based on what we discussed during the meeting, and Riya will begin working on the related works section. At our next meeting we plan to solidify our project topic and plan based on the work that we each completed. We will all work to facilitate productive meetings to use all our time in a productive way.

GROUND RULES

Our group will meet 2-3 times per week/progress milestone to share and discuss our progress since the last meeting, work together on shared responsibilities, and plan what work will need to be completed by the next meeting. Meeting times will be based on availability and change week to week due to busy and changing schedules, but will be held on mainly Tuesdays and Fridays as those are the days we are the most available. Our first meeting will be to read over and split up work, the second meeting will be to work, share our progress, and troubleshoot any issues we come across, and our third meeting will be to combine and clean up our work before submitting. We will meet primarily on zoom as it will be easier to work around groupmates schedules but will plan to meet in person if we think it would be beneficial to do so. We will discuss and consider all member's opinions. If a disagreement arises, we will consider pros and cons to each option, and listen clearly to each other's ideas, taking everyone's thoughts into consideration before making an informed final decision. We will contact each other through text messaging as well as in class to keep each other on track. We expect high participation and level of commitment from all members as we hope for a good grade on this project.

POTENTIAL BARRIES AND COPING STRATEGIES

As a group, we see our biggest possible barrier being availability to meet and complete work as we all have busy schedules and will likely experience super-busy weeks. It also may be hard to meet as we all may be in different locations. We plan to handle this potential barrier by communicating often and effectively. If needed, we will set up smaller meetings and communicate responsibilities through texting or online. We will meet on zoom to manage location barriers when we are in different places. During past group projects, we have experienced problems with members not doing their work in a timely manner, not communicating their availability/progress, doing things in a different way than agreed upon, as well as various others. If we run into problems with group dynamics, we will not be afraid to bring them up to the group during meetings. To proactively combat these potential issues, we have mutually agreed to be self-motivated, avoid stepping on each other's toes, and to bring up and discuss conflicts and issues and talk them out. If we feel that a member is not fulfilling their responsibilities, we will first address it as a group. If the issue is not resolved, we will bring the issue to the teaching staff for further escalation.

REVISITING OUR GROUP CHARTER

We, as a group, believe that we have all been abiding by our agreed-upon guidelines from the start of this project. We have all been easily reachable through text and continue to send our availability so as to find times when we are all available to meet. We have been showing up to meetings when we are available and have successfully been dividing up the work between group members. We feel comfortable with our group roles and believe that there are no major problems we need to troubleshoot at this time. However, one thing we as a group think could use some extra attention is finding times to meet. We all have busy, conflicting schedules, so finding an adequate block of time during which we are all available to meet has proven to be difficult in the past few weeks. In the coming weeks, we plan to communicate our availability for the entire week each Monday, and we will schedule 2 or 3 meetings per week at the beginning of each week. We hope this will prevent us from struggling to find time to meet close to a deadline. Each of our group members has included one positive thing they have seen other group members contribute to the project below:

Anika: the other group members have been easily reachable by text, especially right before deadlines, which is helpful in deciding who is available to make final changes and submit on Gradescope.

Abby: each member is great at completing work we split up, and no one tries to slack off or do less than others. The level of contribution makes me feel sure we'll always be able to complete our work well and on time!

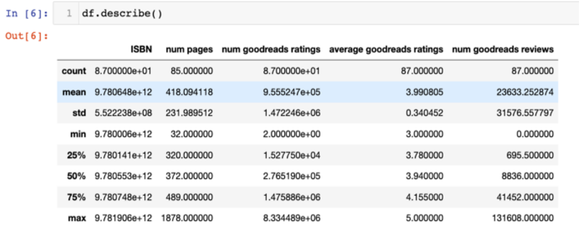
Riya: I appreciate how organized everyone's been with getting their work done on time and reaching out to each other to figure out how to split up work for an assignment and also to help each other out when needed

Veronica:
I really appreciate how collaborative and effective our team is in communicating our ideas and needs, and how that has improved the quality of this project beyond my expectations.

Appendix B: Data Exploration

Data Review

The data types present in our dataset are items and attributes describing different information about these items. See the additional subsection Data Structures for more information. After performing an initial review of our datasets, we found some interesting statistics regarding the items in our two datasets. Our data consists of 2 main data sets: a csv file containing the Top 100 Books on Goodreads and a csv containing book reviews. The top 100 Goodreads csv includes 100 book titles and data describing those books such as the author, number of pages, genre, average rating, number of ratings, and the rating distributions. Some notable statistics from this data set are that the mean Goodreads rating is 3.99/5, the mean number of pages is 418, the minimum number of ratings for a book is 200, while the maximum number of ratings for a book is 8334489.



```
In [6]: df.describe()
Out[6]:
```

	ISBN	num pages	num goodreads ratings	average goodreads ratings	num goodreads reviews
count	8.700000e+01	85.000000	8.700000e+01	87.000000	87.000000
mean	9.780648e+12	418.094118	9.555247e+05	3.990805	23633.252874
std	5.522238e+08	231.989512	1.472246e+06	0.340452	31576.557797
min	9.780006e+12	32.000000	2.000000e+00	3.000000	0.000000
25%	9.780141e+12	320.000000	1.527750e+04	3.780000	695.500000
50%	9.780553e+12	372.000000	2.765190e+05	3.940000	8836.000000
75%	9.780748e+12	489.000000	1.475886e+06	4.155000	41452.000000
max	9.781906e+12	1878.000000	8.334489e+06	5.000000	131608.000000

The book reviews file contains information such as book title, a rating out of 5 for the title, text describing a review for the book, the number of likes that review has, overall negative or positive describing the review, as well as other values. When a sentiment analysis was carried out on the dataset using VADER, some notable statistics we found in our initial review of this dataset include the mean rating being 3.7/5 and the mean Vader neutral score being -551.95.

When first reviewing our data, we found some issues with the formatting for accessibility in a csv-based visualization. In addition, some formats seemed difficult to parse, and the currency mode was not proper for our intended audience.

Problems

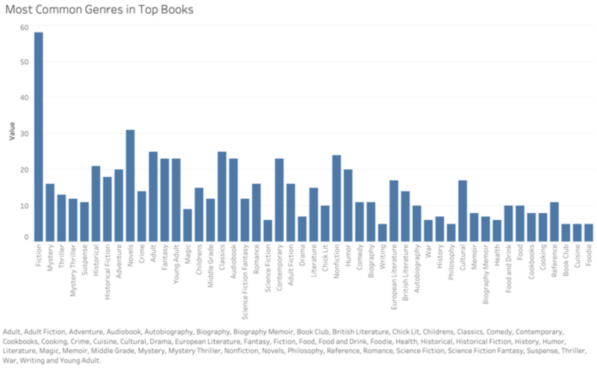
- The numerical, or rather the financial, data we used was collected in the UK and based off of UK sales data, meaning the units were taken in pounds. Our visualization is meant to be shown in class and to other university members, who are generally more familiar with the American dollars. As a solution, we removed the pounds and commas from the string, used the pound-to-dollar conversions, and formatted the amount back into a string format
- While our data included plenty of relevant information, we wanted our data to have a good key to call. To fix this, we removed the irrelevant “Index” column, and set the “Goodreads_ID” to the index for the CSV. Each Goodreads_ID is unique, combining an ID number and the Goodreads ID, and is a good fix for our situation given we will be externally storing some multi-column and hot vector encoded data in separate data tables keyed or named with the unique value as well.
- The dates included with the Publishment Date column were given in day, month abbreviation, and year format- separated by new lines. The more textual format would be harder to parse, and as such it was reformatted as %d/%m/%Y in the DataFrame. This same solution was implemented in the reviews csv.
- We had previously included a list of genres for each book as provided by Goodreads. While lists are difficult to store in DataFrames, we had originally managed this. Now, for accessibility and for the ability to parse through the genres easily, we removed the genres column and created a separate DataFrame the length of the books by the length of the genres and one-hot encoded it. This means that at every intersection between a book and a genre, the matrix holds false if that book is not considered that genre while the matrix holds true if that book is considered that genre.
- Another attribute removed for usability was the ratings distributions. Originally there was a dictionary inside the cell that held the number of ratings of x out of 5 stars for the given book, for 1-to-5-star ratings. To do this we created a new csv that created columns for 1 to 5 stars where each cell is the number of that type of review the book of that row had received.
- The last removal we made was with the shelves of each book on Goodreads. For accessibility and for the ability to parse through the shelves easily, we removed the shelves column and created a separate DataFrame the length of the books by the length of the shelves. In each cell is held the number of times a given book has been included in a shelf of that name. Multiple shelves of the same name exist and may be owned by many, many different users.
- Originally the Flair sentiment analysis was read by keeping together the sentiment and the scoring in a text object. To make this data easier to control, we split the sentiment and its score into two columns that handle the sentiment and the score separately.

Insights

Our greatest finding, especially with the datasets that deal with subjective data, such as reviews, genres, and the shelves datasets, is that there exists a lot of ambivalence within the data. Some of the best performing books in terms of financial success are also some of the worst performing in terms of reviews. A great example is the Harry Potter and The Philosopher’s Stone, a financial and public success, yet it is ranked 3rd in cumulative 1-star (negative) reviews, despite it simultaneously having the most 5-star (positive) reviews. This speaks volumes about how public opinion and perception of pieces of literature view can vary greatly and is ever changing. Another insight we found is that the majority of the Top 100 books aren’t part of a series. However, the top selling series of all time is Harry Potter, followed by Robert Langdom (The Da Vinci Code) and Twilight. This insight also corroborates with the most popular genres of the Top 100 books, Fiction, Mystery, and Thriller.

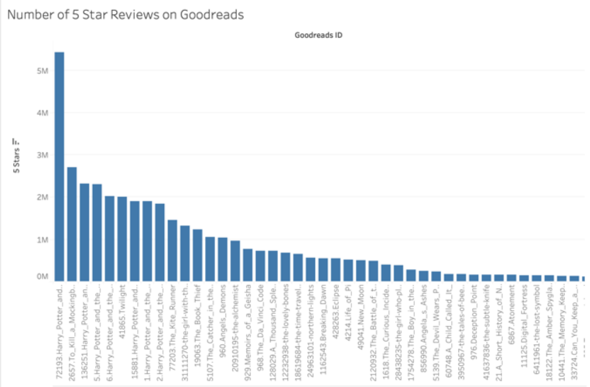
Now to delve into some other sections of our data, such as top genres. This dataset lists out the genres each book in the Top 100 list is classified in. Since a book can be classified under multiple genres, there are several columns for each genre (150+). From this, we are able to see what are considered the more popular genres based on how many books are classified under a genre. For example, "Fiction" is the most popular genre based on our data, since 58 of the books are listed as fiction. Some other popular genres include "Novels" with a count of 31 books, "Classics" with a count of 25 books and "Nonfiction" with a count of 24 books. The other end of the spectrum can be considered as well with many genres only have one book listed, such as "Action," "Young Adult Romance," and "Time Travel." Another one of our sections of data, top shelves, contains what user shelves each book is in and for how many users. There is several listed shelves here (2000+), however certain trends can be seen by choosing to view the data from certain shelves and seeing which shelf titles tend to be the most popular for these books. For example, we can see many of the Top 100 books are under shelves titled "Fiction" and "Favorites."

SCREENSHOTS

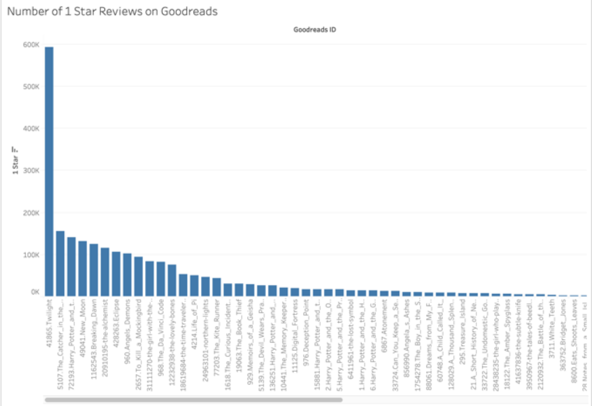


This visualization explores the genre subset of our data. Specifically, it records how many books within the Top 100 Goodreads list are classified under each genre. One book can be listed under multiple genres. To make the visualization cleaner, only the genres with at least five books under it are recorded. The bars for each genre depict the trend to showcase which genres had the most books in the list, such as "Fiction," indicating this to be a popular genre.

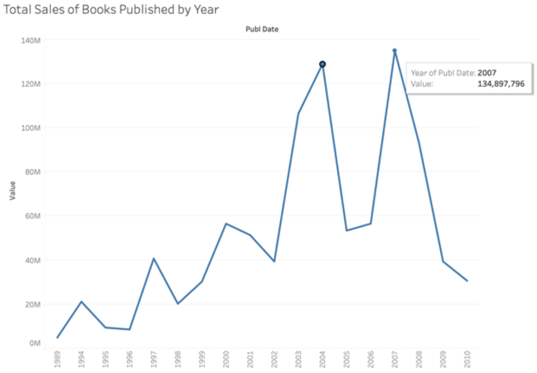
The next two visualizations encapsulate how the Top 100 books performed in terms of reviews on Goodreads.



Specifically, it displays the cumulative number of 5-stars reviews for all the Top 100 books. The leading book in this measure of success is "Harry Potter and The Philosopher's Stone," which is indicative of the series' performance, as 4/5 of the Top 5 in this visualization are part of the "Harry Potter" Series. However, the second book with the most 5-star reviews is "To Kill a Mockingbird."

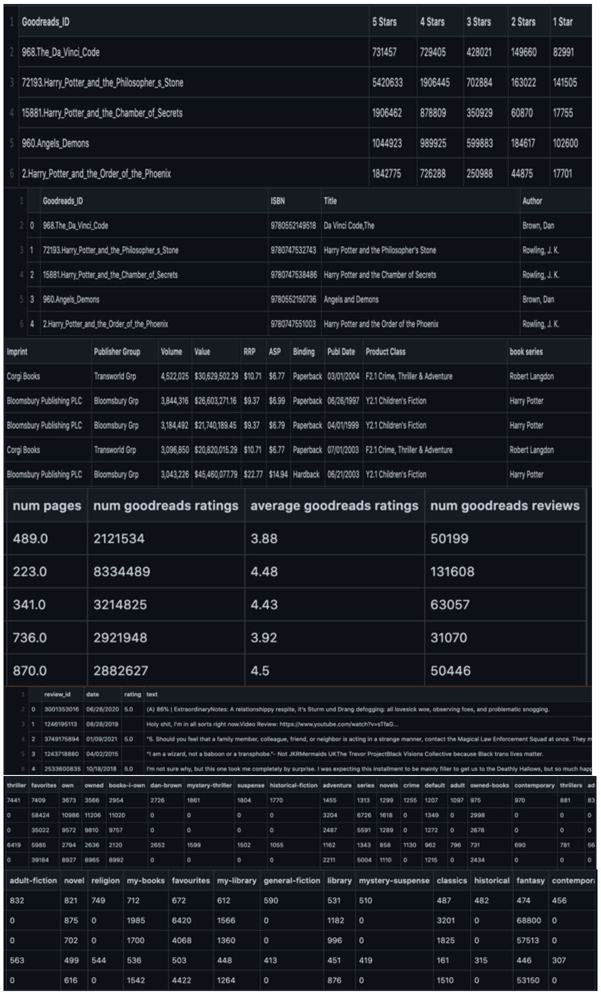


Next, this visualization displays which the cumulative number of 1-star reviews for all the Top 100 books. The leading book in this measure is "Twilight." The "Twilight" series is another series that dominates review performance, with 3/5 of the Top 5 in this visualization being a part of Meyer's series. Interestingly, "Harry Potter and the Philosopher's Stone," the book with the most 5-star reviews out of all the Top 100 books, is also the 3rd book with the most 1-star reviews, showing the ambivalence of the public's opinion on pieces of literature.



This visualization aggregates the total lifetime sales of all the Top 100 books, grouped by the year that the books were published. 2007 was the year that the most financially successful books were published, with over 134 million dollars in lifetime sales. Literature published on this year include "Harry Potter and the Deathly Hallows," "Twilight," and "The Boy in the Striped Pajamas." In 2008, although there was only a small difference in the number of books published, lifetime sales for those books took a sharp decline.

DATA SNIPPET



Goodreads_ID	5 Stars	4 Stars	3 Stars	2 Stars	1 Star
968.The_Da_Vinci_Code	731457	729405	428021	149660	82991
72193.Harry_Potter_and_the_Philosopher_s_Stone	5420633	1906445	702884	163022	141505
15881.Harry_Potter_and_the_Chamber_of_Secrets	1906462	878809	350929	60870	17755
960.Angels_Demons	1044923	989925	599883	184617	102600
2.Harry_Potter_and_the_Order_of_the_Phoenix	1842775	726288	250988	44875	17701

	Goodreads_ID	ISBN	Title	Author
0	968.The_Da_Vinci_Code	9780552149518	Da Vinci Code,The	Brown, Dan
1	72193.Harry_Potter_and_the_Philosopher_s_Stone	9780747532743	Harry Potter and the Philosopher's Stone	Rowling, J. K.
2	15881.Harry_Potter_and_the_Chamber_of_Secrets	9780747538486	Harry Potter and the Chamber of Secrets	Rowling, J. K.
3	960.Angels_Demons	9780552150736	Angels and Demons	Brown, Dan
4	2.Harry_Potter_and_the_Order_of_the_Phoenix	9780747551003	Harry Potter and the Order of the Phoenix	Rowling, J. K.

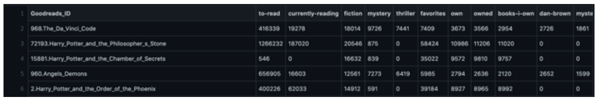
Imprint	Publisher Group	Volume	Value	RRP	ASP	Binding	Publ Date	Product Class	book series
Corgi Books	Transworld Grp	4,522,025	$30,629,502.29	$10.71	$6.77	Paperback	03/01/2004	F2.1 Crime, Thriller & Adventure	Robert Langdon
Bloomsbury Publishing PLC	Bloomsbury Grp	3,844,316	$26,603,271.16	$9.37	$6.99	Paperback	06/26/1997	Y2.1 Children's Fiction	Harry Potter
Bloomsbury Publishing PLC	Bloomsbury Grp	3,184,492	$21,740,189.45	$9.37	$6.79	Paperback	04/01/1999	Y2.1 Children's Fiction	Harry Potter
Corgi Books	Transworld Grp	3,096,850	$20,820,015.29	$10.71	$6.77	Paperback	07/01/2003	F2.1 Crime, Thriller & Adventure	Robert Langdon
Bloomsbury Publishing PLC	Bloomsbury Grp	3,043,226	$45,460,077.79	$22.77	$14.94	Hardback	06/21/2003	Y2.1 Children's Fiction	Harry Potter

num pages	num goodreads ratings	average goodreads ratings	num goodreads reviews
489.0	2121534	3.88	50199
223.0	8334489	4.48	131608
341.0	3214825	4.43	63057
736.0	2921948	3.92	31070
870.0	2882627	4.5	50446

	review_id	date	rating	text
0	3001353016	06/28/2020	5.0	(A) 86% | ExtraordinaryNotes: A relationshippy respite, it's Sturm und Drang defogging: all lovesick woe, observing foes, and problematic snogging.
1	1246195113	08/28/2019		Holy shit, I'm in all sorts right now.Video Review: https://www.youtube.com/watch?v=sTfaG...
2	3749175894	01/09/2021	5.0	"5. Should you feel that a family member, colleague, friend, or neighbor is acting in a strange manner, contact the Magical Law Enforcement Squad at once. They m
3	1243718880	04/02/2015		"I am a wizard, not a baboon or a transphobe."- Not JKRMermaids UKThe Trevor ProjectBlack Visions Collective because Black trans lives matter.
4	2533600835	10/18/2018	5.0	I'm not sure why, but this one took me completely by surprise. I was expecting this installment to be mainly filler to get us to the Deathly Hallows, but so much happ

thriller	favorites	own	owned	books-i-own	dan-brown	mystery-thriller	suspense	historical-fiction	adventure	series	novels	crime	default	adult	owned-books	contemporary	thrillers	ad
7441	7409	3673	3566	2954	2726	1861	1804	1770	1455	1313	1299	1255	1207	1097	975	970	881	83
0	58424	10986	11206	11020	0	0	0	0	3204	6726	1618	0	1349	0	2998	0	0	0
0	35022	9572	9810	9757	0	0	0	0	2487	5591	1289	0	1272	0	2678	0	0	0
6419	5985	2794	2636	2120	2652	1599	1502	1055	1162	1343	858	1130	962	796	731	690	781	56
0	39184	8927	8965	8992	0	0	0	0	2211	5004	1110	0	1215	0	2434	0	0	0

adult-fiction	novel	religion	my-books	favourites	my-library	general-fiction	library	mystery-suspense	classics	historical	fantasy	contempora
832	821	749	712	672	612	590	531	510	487	482	474	456
0	875	0	1985	6420	1566	0	1182	0	3201	0	68800	0
0	702	0	1700	4068	1360	0	996	0	1825	0	57513	0
563	499	544	536	503	448	413	451	419	161	315	446	307
0	616	0	1542	4422	1264	0	876	0	1510	0	53150	0



Goodreads_ID	to-read	currently-reading	fiction	mystery	thriller	favorites	own	owned	books-i-own	dan-brown	myste
968.The_Da_Vinci_Code	416339	19278	18014	9726	7441	7409	3673	3566	2954	2726	1861
72193.Harry_Potter_and_the_Philosopher_s_Stone	1266232	187020	20546	875	0	58424	10986	11206	11020	0	0
15881.Harry_Potter_and_the_Chamber_of_Secrets	546	0	16632	839	0	35022	9572	9810	9757	0	0
960.Angels_Demons	656905	16603	12561	7273	6419	5985	2794	2636	2120	2652	1599
2.Harry_Potter_and_the_Order_of_the_Phoenix	400226	62033	14912	591	0	39184	8927	8965	8992	0	0

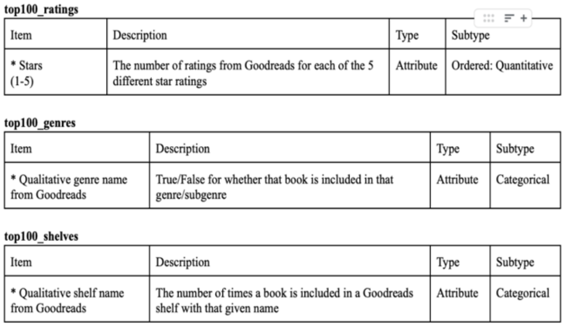
DATA STRUCTURE

top100_goodreads_cleaned

Item	Description	Type	Subtype
Goodreads_ID	The ID number concatenated with "." and the Title of the given book	Attribute	Ordered: Quantitative
ISBN	The International Standard Book Number for the given book	Attribute	Ordered: Quantitative
Title	The Title of the book as listed in the top100 list	Attribute	Categorical
Author	The Title of the book as listed in the top100 list	Attribute	Categorical
Imprint	The printer of the book within the publishing group as listed in the top100 list	Attribute	Categorical
Publisher Group	The Publisher of the book as listed in the top100 list	Attribute	Categorical
Volume	The amount of books sold as listed in the top100 list	Attribute	Ordered: Quantitative
Value	The total dollar amount of books sold as listed in the top100 list	Attribute	Ordered: Quantitative
RRP	The recommended retail price of books sold as listed in the top100 list	Attribute	Ordered: Quantitative
ASP	The average selling price of books sold as listed in the top100 list	Attribute	Ordered: Quantitative
Publ Date	The Publish Date of the book as listed in the top100 list	Attribute	Ordered: Ordinal, cyclical
Product Class	The Product Class of the book as listed in the top100 list	Attribute	Categorical
book series	The numbered book in a series, if in a series as on Goodreads	Attribute	Ordered: Ordinal
num pages	The number of pages in the book as listed on Goodreads	Attribute	Ordered: Quantitative
num goodreads ratings	The number of ratings the book received on Goodreads	Attribute	Ordered: Quantitative
num goodreads reviews	The number of reviews the book received on Goodreads	Attribute	Ordered: Quantitative
average goodreads ratings	The average rating the book received on Goodreads	Attribute	Ordered: Quantitative

In reviews_csvs (for each csv…)

Item	Description	Type	Subtype
review_id	The ID a review is given by Goodreads	Attribute	Ordered: Quantitative
date	The date the review was made on Goodreads	Attribute	Ordered: Ordinal, cyclical
rating	The rating out of 5	Attribute	Ordered: Ordinal
text	The textual component of the review	Attribute	Categorical
num_likes	The number of likes the given review has	Attribute	Ordered: Quantitative
Vader_score	"summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is accurate."	Attribute	Ordered: Quantitative
Vader_neg	pos, neu, and neg scores are ratios for proportions of text that fall in each category (so these should all add up to be 1... or close to it with float operation)	Attribute	Ordered: Quantitative
Vader_neu		Attribute	Ordered: Quantitative
Vader_pos		Attribute	Ordered: Quantitative
TextBlob_sub	The amount of subjectivity in the Text [0,1]	Attribute	Ordered: Quantitative
Flair_sent	The sentiment Flair predicted [POSITIVE, NEGATIVE]	Attribute	Ordered: Ordinal
Flair_conf	The confidence of Flair in its classification of the Text Sentiment as a probability. [0,1]	Attribute	Ordered: Quantitative



top100_ratings

Item	Description	Type	Subtype
* Stars (1-5)	The number of ratings from Goodreads for each of the 5 different star ratings	Attribute	Ordered: Quantitative

top100_genres

Item	Description	Type	Subtype
* Qualitative genre name from Goodreads	True/False for whether that book is included in that genre/subgenre	Attribute	Categorical

top100_shelves

Item	Description	Type	Subtype
* Qualitative shelf name from Goodreads	The number of times a book is included in a Goodreads shelf with that given name	Attribute	Categorical

APPENDIX C: INTERVIEW

END USER PERSONAS

1) Melanie King : Book Club Host

   Melanie is a young adult in her early 20s who hosts a monthly book club. She has been struggling to manage the book club in an interesting way, since most of the members are busy and have been complaining about the lack of connection between some of the books chosen by members. Melanie is a college student studying Biology, and although she is familiar with general technology and technical readings, she hasn't spent too much time with a specialized visualization before. Melanie could be helped by the visualization to help her provide a more robust tool to provoke discussion about the different novels her book club is reading. Currently, she has been presenting the main genres and themes of novels to her book club to discuss differences, but it's difficult to derive much about the differences in books just from the few pieces she googles. Melanie thinks that having a on screen comparison of books, or a view of the trends in the different books her club reads could give them a centralized source to make discussion based on.

2) Larry Palmer : Harry Potter Fanatic

   Larry is an adult male in his 30s that has been a lifelong fan of Harry Potter and the Wizarding World. Although he enjoys all aspects of the series, he especially likes to see how the series as a whole and the individual books compare to other successful literature pieces. He also likes to be kept updated on the public opinion and discourse around the books. However, Larry currently performs these searches manually, which is extremely tedious and often hard to see the "big picture". Larry was an English major in university and is now a freelance writer, so his exposure to visualization technology is limited. A visualization for Larry would facilitate the process of mentally grasping how successful the Harry Potter series has been on a global scale, and to quantify the cultural impact the books had. The visualization would also reveal how people are rating the books on Goodreads, his favorite online book community, and how these reviews may vary book-to-book intra-series or book-to-book with other successful novels.

INTERVIEW SCRIPT

3) What is the intended purpose of this visualization?
   a) What do you feel is the most important information that the visualization should depict?
   b) Why do you want to create this new visualization?
4) Who is the target audience?
   a) Should the visualizations differ based on audience?
5) Do you currently have a visualization for this purpose? Or does one exist?
   a) What problems are you facing with the current method?
   b) Is there specific information you feel they are not portraying? How do you feel this could be improved upon?
6) Are there specific things you would like to showcase within the visualization?
7) How should a user be able to interact with the visualization?
   a) What do you hope they can learn from the visualization?

INTERVIEW NOTES

Melanie King

1. Comparison of books & their stats for her book club
2. Snapshot of information for one book as well as larger trends
   a. Good for book clubs to find more specific information
3. Can't find a good book review that isn't just huge blocks of text so having a visual display would be very helpful
4. Target audience – book club to give access to same information
5. Don't think the visualizations should differ based on audience
   a. Be able to see more information if on a bigger screen
6. Haven't found a current visualization
   a. Mainly find readings & text reviews to compare
7. Hard to connect different books from the current method of reading information on Goodreads
   a. & hard to remember so would be easier if there was a visualization
8. Nice to see more about other people's reviews on books beyond her book club
9. Be able to see other people's opinions on books
   a. & the book specific information
10. Users should be able to select books
    a. But also see those books in overall trends
    b. Highlight them in the overall trends
11. Hoping people can compare & contrast books to create a lively book club discussion

Larry Palmer

12. Comparing how HP books stack up against each other, knowing public opinion of books over time (reviews)
13. Reviews are important! Info on how books are being reviewed, criticized, praised, GLOBALLY, no country or age bias
14. Looking at reviews manually takes a long time, hard to grasp large scale opinions of books
15. Target Audience: Themself, for personal pleasure and general harry potter community
    a. Should not vary too much for different audiences, but should be able to look at same info for different series
16. Has not found or used a visualization for this purpose, relies on Goodreads

   a. Very tedious process to go through all Goodreads reviews, hard to remember overall sentiments of reviews bc there is so much to read

17. Enjoys comparison a lot, would like information about other books so she can compare
18. Highlight historical success of the series historically
19. Ability to zoom out and see bigger picture while also zooming in on a specific book and seeing details on that book
20. Cultural impact and success of HP in literature and globally

INTERVIEW RESULTS

Melanie King

What is the intended purpose of this visualization?
MK: To show my book club a comparison of books, and also to give us interesting statistics to refer back to as we discuss books we've read and wat to read as a club.

What do you feel is the most important information that the visualization should depict?
MK: I think giving a good chunk of knowledge for an individual or pair of books is the most important. I want to see snapshots of each book information easily.

Why do you want to create this new visualization?
MK: I can't find a concrete book overview that isn't pages and pages of text. I want something way easier to read.

Who is the target audience?
MK: My book club! We want each to have access to the same information.

Should the visualizations differ based on the audience?
MK: I don't think so. It'd be nice if people could look on their phones but then I could pull up a computer screen or it on the TV and have it hold more information because it's bigger.

Do you currently have a visualization for this purpose? Or does one exist?
MK: I haven't found one. I mostly just read off a few pieces of information about the book on Goodreads or something when we talk about the higher-level overview of books.

What problems are you facing with the current method?
MK: No one remembers the couple of facts I say out, we need something we can look back at. Plus, I'm not good at connecting the different books we read, and some of my members have been complaining.

Is there specific information you feel they are not portraying? How do you feel this could be improved upon?
MK: I don't know what the information is but It'd be nice to see what other people think of a given book so we could agree/disagree with the general opinion.

Are there specific things you would like to showcase within the visualization?
MK: I think what I said before, probably some opinions of the books would be cool.

How should a user be able to interact with the visualization?
MK: Selecting books but also seeing overall trends. Maybe if you select a book you can see it in the overall trends? I'm not sure how that works though.

What do you hope they can learn from the visualization?
MK: To compare and contrast books! Making discussion is important to make book club a success.

Larry Palmer

What is the intended purpose of this visualization?
LP: I would like to compare how books (especially Harry Potter books) stack up against one another and find out the general public opinion of these books over time through their reviews.

What do you feel is the most important information that the visualization should depict?
LP: I think being able to visually see reviews is important! Also information on how books are being reviewed, criticized, and praised.

Why do you want to create this new visualization?
LP: The creation of this new visualization is important because looking at reviews for books manually/ one by one takes a very long time. Furthermore, it is hard to grasp large-scale opinions of books through singular reviews.

Who is the target audience?
LP: I am the target audience as I will use this visualization for my own enjoyment and pleasure. However, the general Harry Potter community, as well as all book lovers and readers are also included in the target audience.

Should the visualizations differ based on audience?
LP: The visualization should not vary too much for different audiences, but I hope that users will be able to look at the same information for different books and different series.

Do you currently have a visualization for this purpose? Or does one exist?
LP: I have yet to find a successful visualization for this purpose. I mostly rely on Goodreads for my book review information.

What problems are you facing with the current method?
LP: It is a very tedious process to go through all the Goodreads reviews. It's especially hard to remember the overall sentiments and feelings of reviews because there are so many to read.

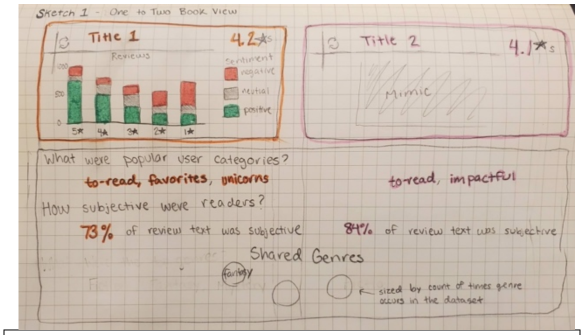
Is there specific information you feel they are not portraying? How do you feel this could be improved upon?
LP: I feel like the way Goodreads is set up, there isn't a great way to compare book ratings and reviews. I enjoy comparing ratings and reviews a lot, so I would like to see information about other books at the same time so I can compare them.

Are there specific things you would like to showcase within the visualization?
LP: I would like to see highlights of the historical success of the series in the past and how it has changed over time.

How should a user be able to interact with the visualization?
LP: A user should be able to zoom in and out to see the bigger picture as well as see information and details on a specific book when they zoom in.

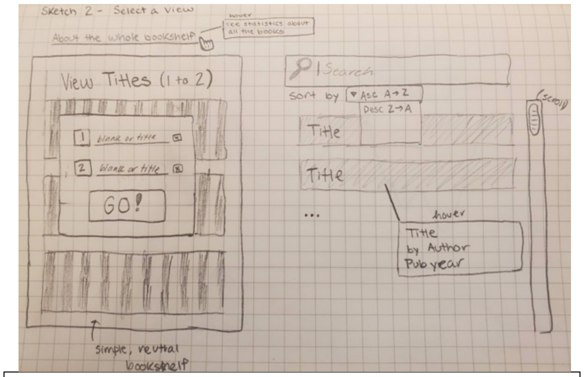
What do you hope they can learn from the visualization?
LP: I hope that users can learn about the cultural impact and success of books and series in literature and globally. Especially Harry Potter! I'm a huge Harry Potter fan if that isn't clear!

APPENDIX D: DESIGN SKETCHES

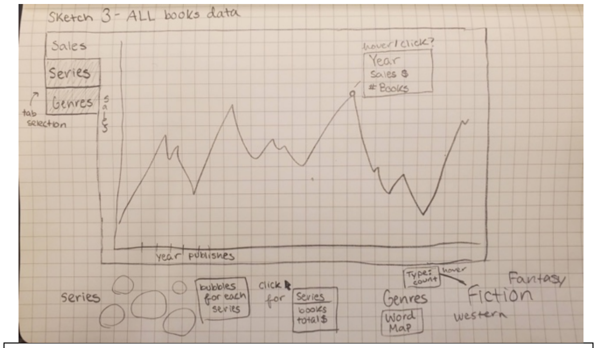

Sketch 1 - One to Two Book View
Title 1
4.2★s
Reviews
1000
500
0
5★
4★
3★
2★
1★
Sentiment
negative
neutral
positive
Title 2
4.1★s
Mimic
What were popular user categories?
to-read, favorites, unicorns
to-read, impactful
How subjective were readers?
73% of review text was subjective
84% of review text was subjective
Shared Genres
fantasy
sized by count of times genre occurs in the dataset

Favorite

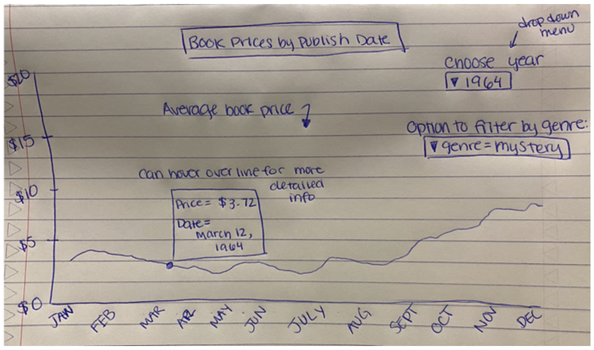
Abby Carr
I decided to include rating counts and their sentiments within a separate bar chart for each book since for each book we would be more concerned with seeing a distribution comparison rather than a circle comparison. The percentages and top shelves (user categories) are indicated by the color associated with a given book to distinguish them, and the subjective percentage may also be distinguished by a ring filled with XX% of that color for more visual appeal. Ranking genres in circles sized by commonality gives users an easier time distinguishing and reading different genres as opposed to a list. They may also be lightened/colored directly correlated to size for the same purpose.
This view addresses the public opinion of a book as well as book to book comparison by providing a view of ratings and subjective rating of the given book(s).



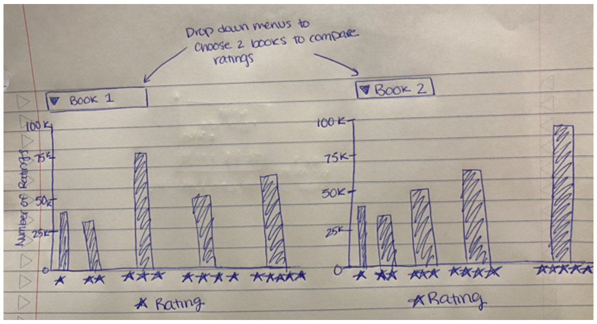
Abby Carr
This view focuses on the visual aspects for navigation. A subtle and clean bookcase visual behind the panel of the selected books helps users tie together our separate sections of choosing single books from the shelf and choosing to look at the entire bookshelf. The use of matching panels for each book in the scrolling search & sort section give users the clean reading of the books, with a hover or potentially another panel that populates with a few more pieces of the book's information to provide more information while browsing.
This view was not originally included in our Task Analysis but would easily fall into a Search and Discover task where users find the type of information they want to see.



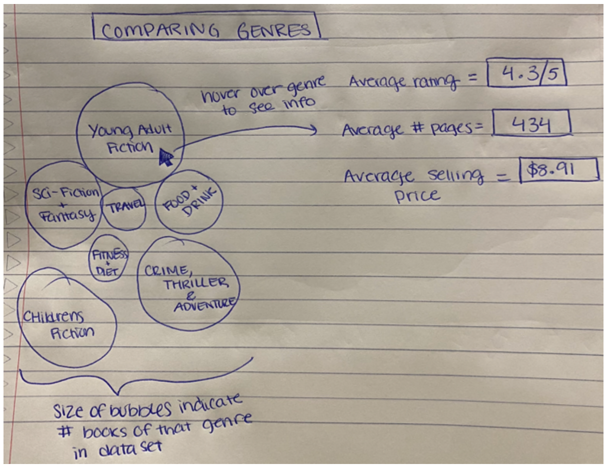
Abby Carr
This view is meant to highlight some of the big-picture information included in our data. The Sales tab uses a line to chart sales since the goal in this visual is to see the trend of sales for books published in each year. Looking at the series, bubbles colored for each series and sized by sales give users a good intuition into the success of different book series by having color discern categories while size shows the levels of success a book series had. Genres provide a broad idea of how common a genre is by having the size of that word in the world map correlates to the size/count of occurrences.
This view covers the popularity of series and the 'public opinion' tasks. We have a view comparing the sales of series and more information on sales, an indicator of the level of popularity and public mania a book generated.



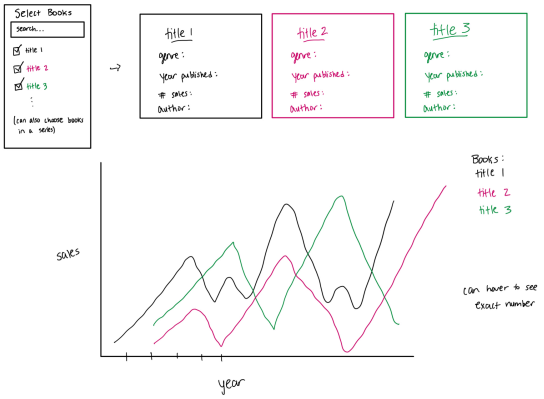
Anika Das
This sketch displays the average book prices for the books in our data set by the date they were published. Users can choose which year they would like to see data for using the drop-down menu. They also have the option to filter books by genre using the second drop-down menu. The main mark used in this visualization is the line that represents the average book price over time. The channels used are horizontal and vertical position of the line which conveys information regarding the date and average price of books for that publish date. I chose these marks and channels because I thought a line graph would be a good way to visualize average price of books over time. This visualization does not address a specific domain task mentioned in our task table, but it easily answers questions regarding how the average price of books (of a certain genre) changed over time.



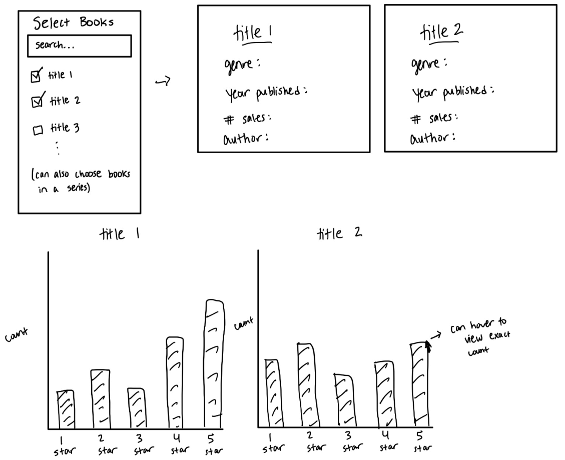
Anika Das
This sketch compares the ratings of 2 books that can be chosen by the user using the 2 drop-down menus. The marks I used in this visualization are lines, which make up the bars, and area 2D, both height and width, because when used in a bar graph, it represents magnitude. In this case it represents how many 1-star, 2-star, 3-star, 4-star, and 5-star ratings a book has. The channels I chose to use are position (both horizonal and vertical) because where the bar is placed represents which rating it represents. The length of the marks (specifically height in this case) represent how many ratings of a certain star level each of the books got. This visualization addresses the task from our task table regarding comparing two books based off of their statistics. In this case, the specific statistic is rating, but that can easily be changed if we want to implement the option of comparing different statistics.



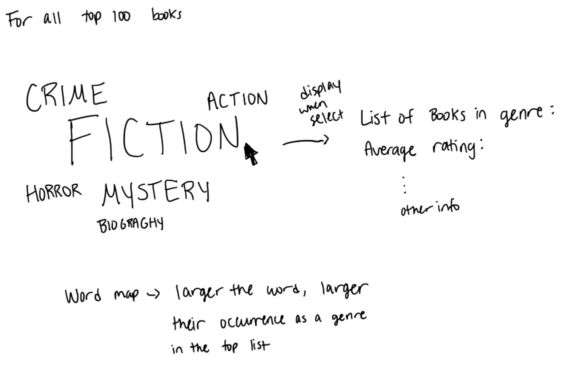
Anika Das
This visualization shows different genres of books in our dataset. Users can hover over the different bubbles, each representing a genre, and the information on the right side of the visualization will update with the specific information (such as average rating, average number, of pages, average selling price, etc.) of books of that genre. The size of the genre bubbles also indicates how many books of that genre we have in the data set so users can get a sense of how common books of that genre are. The marks I chose to use are areas of the genre bubbles because it is an easy way to convey information about numbers, however it is general and not specific. The main channel used is the area of the bubbles, representing again the number of books of that genre. The statistics on the right side are a simple way to convey information but can be made into different types of visualizations if we want. The task from our task table that this visualization addresses is knowing which genre(s) of books tend to get the highest rating because it allows users to easily see basic information (such as rating) about a genre by simply moving their cursor.



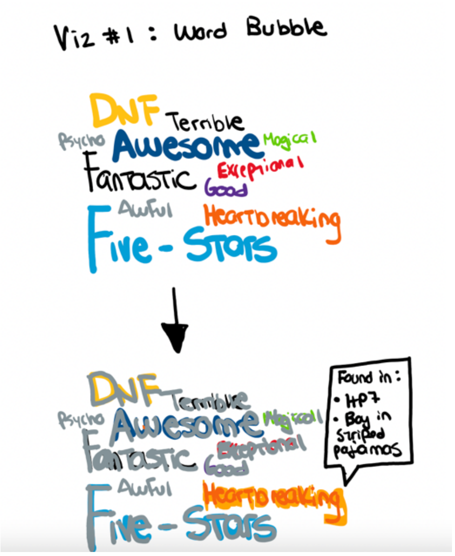
Riya Gurnani
This visualization sketch shows a view where users can select book titles, which would then update the graph. More detailed information about each selected title would be displayed at the top, including the genre of the book, the year it was published, its number of overall sales, and its author. The graph displays the trend of each title's number of sales each year since it was published. The user would also be able to hover over each point to see the exact number for each year. The marks in the graph are the trend lines for each title. The channels are color to differentiate between the titles and both horizontal and vertical position. I chose these because I felt a trend line was the best way to represent book sales over the years, and the ability to have multiple titles together makes it easier to compare. This addresses the task of comparing books based on their statistics (sales).



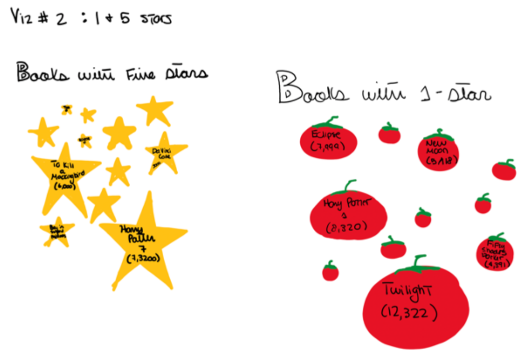
Riya Gurnani
This visualization sketch is similar to the previous one in that there is a panel where the user can select book titles, which updates the visualization. Again, the user can see more information on their selected titles, including its genre, year it was published, its overall sales, and its author. Then, it will update the bar graphs at the bottom so there is one for each title. This depicts the number of one star, two start, three star, four star, and five star reviews for the title and places them next to each other for easier comparison. The user can also hover over the bar to view the exact number. The marks used are can be considered as lines or area, depending on how the bars are viewed. The channels are the size of the bars and both horizontal and vertical position. Due to the categorical nature of the ratings, bar graphs with these marks and channels were the best representation, I felt. This addresses the task of wanting to compare books based on their statistics and public opinion (ratings). It would probably be best to limit how many books a user can indicate here otherwise it would become too complicated.



Riya Gurnani
This visualization sketch represents a word map/cloud of the different genres for the top 100 books data. The larger the word means that larger the count of books under that genre is. The user can also select a word (genre) to view either the count of books classified with that genre and / or the list of books classified with that genre. The marks could be considered points, with each word being one. The channels would be the size of each word since that represents how large that genre's book count is. I choose this because it's visually appealing and different from the more traditional graph, while still relaying pertinent information about which genres tend to be more popular. This addresses the task of wanting to know which genres tend to get the higher ratings. Since it would be pulling from the top 100 books dataset, we know these books have higher ratings so seem which genres are seen more in this list addresses this question.



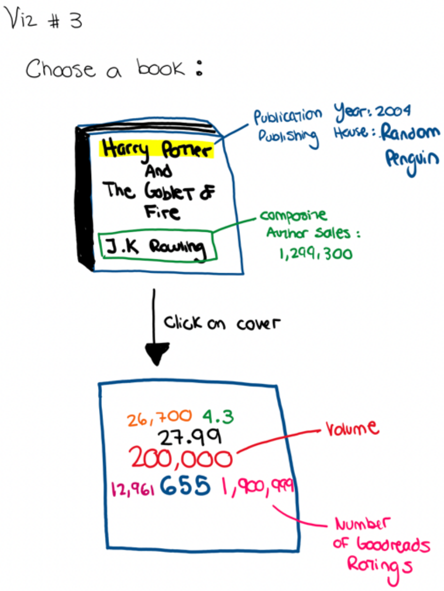
Veronica Aguiar
This visualization uses the concept of a word cloud to represent sentiment analysis results. Since book reviews often feature the same words, phrases or adjectives, visually encoding how often they're used and by what books is a creative way to communicate the overall sentiment. It's also interesting to see how common descriptive words change depending on the review level (1 star vs. 5 star).



Favorite

Veronica Aguiar
This visualization is based off a bubble graph, but instead of circles it uses common symbols to represent the average ratings for each book on Goodreads. The

visualization has two views: Five-Star Reviews and 1-Star Reviews. For the 5-Star Reviews, the books are represented by gold stars, a common symbol for success. The 1-Star Reviews are represented by tomatoes, a symbol made famous by the movie review website Rotten Tomatoes to represent a bad movie. Each figure would show the title of the book and the number of reviews in this category. Like a bubble graph, the size of the bubble is indicatory of the number of reviews.


Viz # 3
Choose a book :
Harry Potter And The Goblet of Fire
J.K Rowling
Publication Year: 2004
Publishing House: Random Penguin
composite Author Sales : 1,299,300
Click on cover
26,700 4.3
27.99
200,000
12,961 655 1,900,999
Volume
Number of Goodreads Ratings

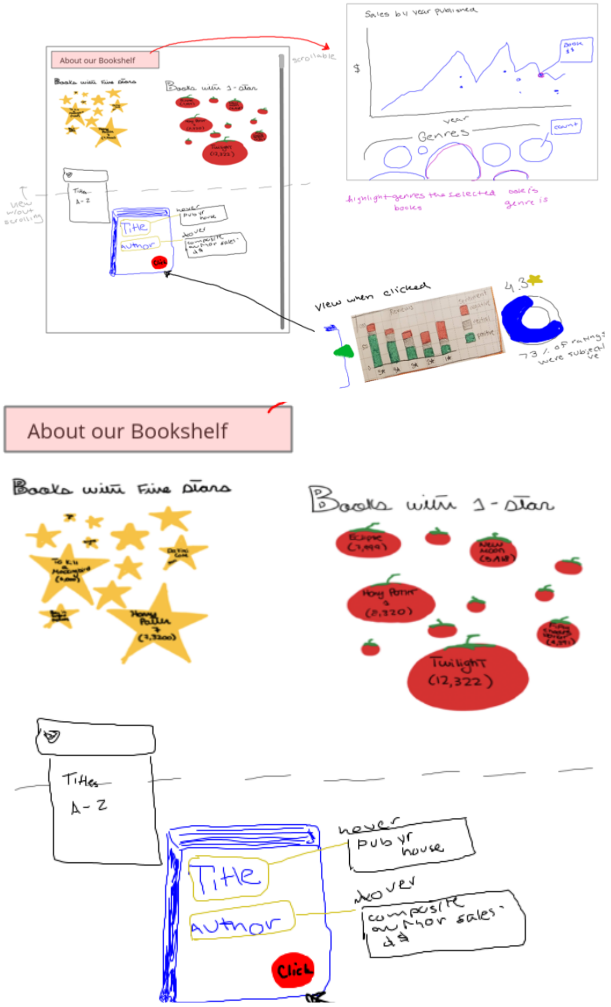
Favorite

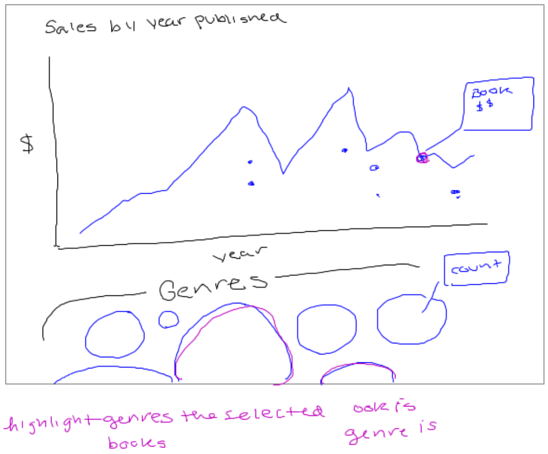
The third visualization uses a literal book representation to show facts and figures about each book. The cover of the digital book is interactive and can show the title, author, composite author sales (calculated), year of publication, and publishing house. After clicking on the cover, a number cloud shows, which represents different numbers related to the book, such as volume, total sales, number of Goodreads ratings, average ratings, and more.

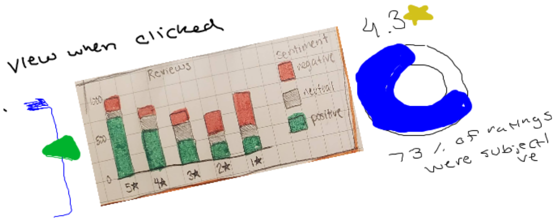
For our favorites, we chose a drawing of reviews and their positive, negative, and neutral tones, a drawing showing top counts of 5- and 1-star reviews, and a drawing of a book that focused on singular book statistics. Our first selected visual was chosen to **analyze** further the overall sentiment of the reviews made by readers on Goodreads. By displaying what proportion of the reviews at each level are positive, negative, or neutral, we can **discover** how sentiment analysis changes for every rating level and compare decreases or increases in different tones. The data types we're working with here are ordinal quantitative attributes.

The drawing showing top counts of 5-star and 1-star displays the magnitude of the popularity (or unpopularity) of each book based on Goodreads reviews and includes the numeric the total count. We used common symbols like gold stars and tomatoes to represent the performance of each book, and the size of each symbol is indicative of the number of reviews, which also identify **trends** in the number of ratings for different books. The data types in this drawing are quantitative attributes. Lastly, our interactive book drawing displays data for each individual book, such as title, date published, publishing house, and more on the cover. When clicked, the figure should change to reveal different information. We thought this drawing would be very creative and helpful in organically **presenting** the data we're working with in an **enjoyable** way. Additionally, the ability to see data book per book is also beneficial to the end user. The data types used here are attributes.

Final Hand-Drawn Sketch

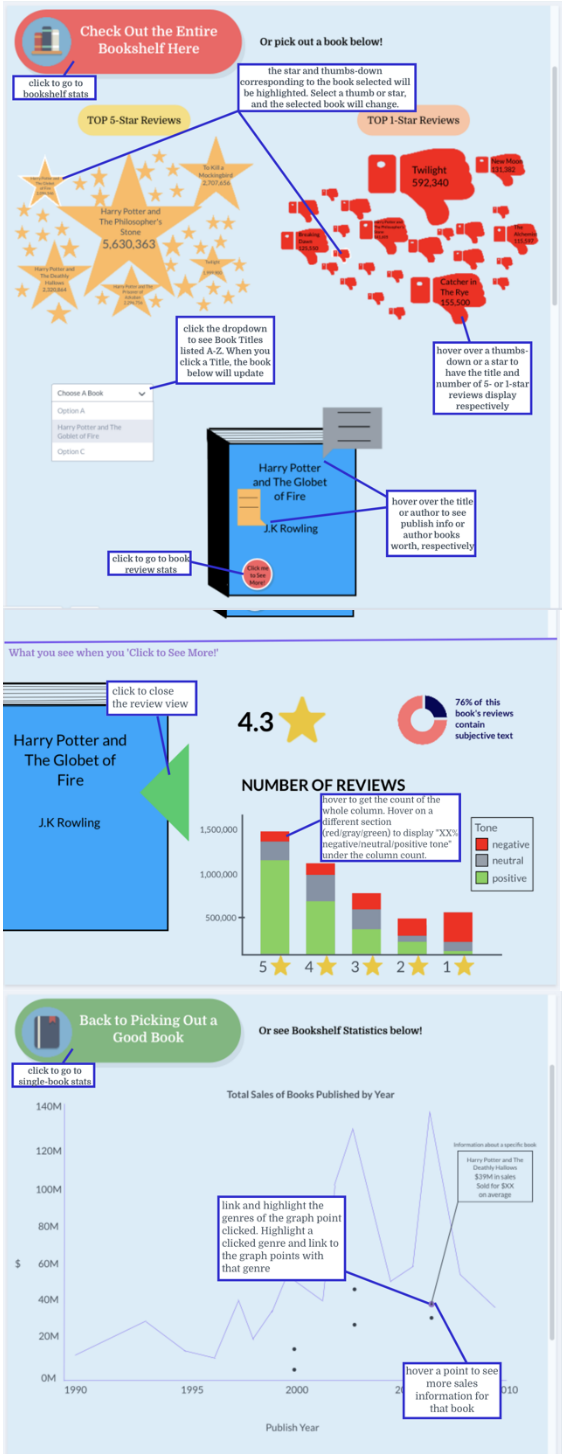






Our Final Sketch has 2 main pages. The home page has a fun interactive graphic that lets you hover over and select star or thumbs-down images that have the top number of 5- and 1-star reviews. The use of size as the visual encoding makes the page less visually intimidating for users who may not be familiar with visual and interactive webpages. Below is a static display of the currently selected book, this is important as a user should be able to read and get the text-based information before they can look at reviews or other moving information. When the user ‘opens’ or clicks on a book sticker, they can see a chart with the number of each star reviews. These are encoded by size as bars, which is an easy way for users to see the distribution of the ratings better than in a pie chart or graph since a bar chart is also positional/ordinal-friendly, meaning we can line up the bars from most to least stars. The color encoding for the tone of reviews was chosen to coordinate with common associations users will already have- green as good, and red as negative or bad. This section also has a static ring chart indicating the average sentimentality of the books’ reviews. This is encoded by size so that the ring bar looks more filled if a larger percentage of reviews are subjective which is easy for a user to see without having to support more interactions than necessary. The second bookshelf page holds a graph with a trendline and corresponding points as well as a bubble graph that holds genres. Like the stars, our genres are encoded by size, once again for the ease of viewers to understand the visualization. We decided to add sales as a line graph to show change over time without a visualization that constantly cycles or changes and adding in the locational dots gives the users a very easy way to see the number of bullet points that summed to the trend line overall. We found that this simple approach was more readable than other moving options. Not using different colors as an encoding was purposeful, since we need to use highlighting for brushing and linking, and the number of categories we have for each type of label for a book is too large to be useful for making an intuitive visual.

APPENDIX E: DIGITAL SKETCH



Check Out the Entire Bookshelf Here
Or pick out a book below!
TOP 5-Star Reviews
TOP 1-Star Reviews
Harry Potter and The Philosopher's Stone 5,630,363
To Kill a Mockingbird 2,707,656
Harry Potter and The Deathly Hallows 2,320,864
Harry Potter and The Prisoner of Azkaban 2,296,756
Twilight 1,999,900
Twilight 592,340
New Moon 131,382
Breaking Dawn 125,550
The Alchemist 115,597
Catcher in The Rye 155,500
Choose A Book
Option A
Harry Potter and The Goblet of Fire
Option C
Harry Potter and The Globet of Fire
J.K Rowling
Click me to See More!
What you see when you 'Click to See More!'
Harry Potter and The Globet of Fire
J.K Rowling
4.3
76% of this book's reviews contain subjective text
NUMBER OF REVIEWS
1,500,000
1,000,000
500,000
5
4
3
2
1
Tone
negative
neutral
positive
Back to Picking Out a Good Book
Or see Bookshelf Statistics below!
Total Sales of Books Published by Year
140M
120M
100M
80M
60M
40M
20M
0M
$
1990
1995
2000
2010
Publish Year
Information about a specific book
Harry Potter and The Deathly Hallows $39M in sales Sold for $XX on average

click to go to bookshelf stats

the star and thumbs-down corresponding to the book selected will be highlighted. Select a thumb or star, and the selected book will change.

click the dropdown to see Book Titles listed A-Z. When you click a Title, the book below will update

hover over a thumbs-down or a star to have the title and number of 5- or 1-star reviews display respectively

hover over the title or author to see publish info or author books worth, respectively

click to go to book review stats

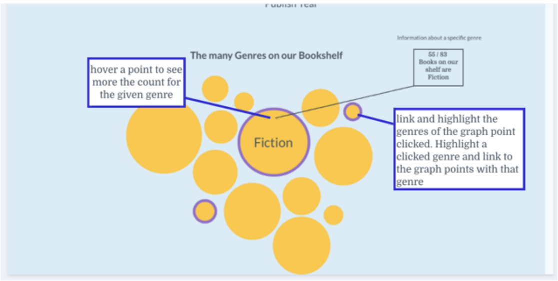
click to close the review view

hover to get the count of the whole column. Hover on a different section (red/gray/green) to display "XX% negative/neutral/positive tone" under the column count.

click to go to single-book stats

link and highlight the genres of the graph point clicked. Highlight a clicked genre and link to the graph points with that genre

hover a point to see more sales information for that book



Publish Year
The many Genres on our Bookshelf
Information about a specific genre
55 / 83 Books on our shelf are Fiction
Fiction

hover a point to see more the count for the given genre

link and highlight the genres of the graph point clicked. Highlight a clicked genre and link to the graph points with that genre

The user would first see the first image as the front page of this entire multi-view visualization. There they would have the option to view the entire bookshelf (which is our top 100 books dataset). They can also view a bubble graph of sorts. This depicts the books in the bookshelf and the size of its shape (star or thumbs-down) is larger based on the number of corresponding star (five or one) reviews that book received. The user can hover over a star or thumbs-down in order to view the title of the book and the number of its five star or one star reviews, respectively. This allows the user to compare certain statistics about the books, namely the quantity of these ratings. Below, they will see a book where they can change this visualization based on the drop down menu to the left of it. Based on the book selected, this book visualization will update. Here, the user can hover over the title to see publishing information or hover over the author to see the author books worth. They can also click on “click me to see more” on the bottom of the book to change the page to see more detailed book review statistics. Now, the next image shows what is shown when that button is clicked. Here, the user can see a graph to show the number of each of the star ratings they have received. The bar graph is also stacked to show how many of each of those reviews had a positive, neutral, or negative sentiment. The user can hover to see the exact information and then choose to go back to the first page. We have another visualization to depict the total number of sales of each book each year since they have been published. This allows the user to view the sales trend line and compare different books through the points on the graph. The last visualization we have is a bubble chart for the different genres of the top 100 books. The bigger the bubble, the larger the count of books with that genre is. The user can also hover over a genre bubble to see the exact count. This allows the user to view which genres tend to get the highest ratings.

This overall description demonstrates several domain tasks for users. We allow them to view the statistics for multiple books and compare them with each other. They also have the ability to view detailed information about a book they select to understand the different numbers of ratings it receives and the sentiment of them. They can also view the number of sales for each book and the overall trend for all of the books in the dataset. Additionally, they can see which genres tend to receive the highest ratings. Based on our original domain tasks, we changed the one where the user would want to see the how the general public rates a “Harry Potter” book and generalized it to any book title in our selection. We also omitted the task of identifying the most popular book in a series as we felt the other tasks of allowing the user to choose books was more pertinent. This task was replaced with a new one allowing the user to view and compare the sales numbers for the books.

APPENDIX F: USABILITY TESTING

PREPARATION

Our visualization uses data scraped from Goodreads about the top 100 books. The data showcases several characteristics of these books including their genres, their publishing year, count of ratings, and reviews from users for the books. The intended use of our visualization tool is to help users learn more facts about their favorite books and compare them to see which characteristics might lead to books becoming more popular. Our visualization can also be used by users to see trends in book sales over a span of years, whether that be for all 100 of the books, or for a singular book.

This leads to specific tasks for a test subject to complete with our visualization tool. One task is that we want a test subject to view which genres tend to have the most popular books. They can do this by viewing our genre count bubble chart and hovering over them to see the genres and the count of books from our dataset with that genre. We want to ensure that this bubble chart is not only easy to use but easy for users to understand its purpose and intent. We also want to understand what more could be added to the chart, if necessary, for its purpose to be made clearer to users since visually the chart is our most simple.

A second specific task that we want a user to be able to complete using our visualization tool is to see how user ratings compare for two different books. They will be able to do this by choosing a book from a drop-down menu and viewing the rating data in a bar graph below. The bar graph will display how many 5-star, 4-star, 3-star, 2-star, and 1-star ratings that book received. We want to test the usability of this portion of our visualization tool because ratings are one of the easiest ways to classify a book, and we hope that the way we visualized this data is easily understandable by users. The outcome that we hope for is for users to be able to easily choose which book they are interested in, and quickly see the distribution of ratings for that book. We are looking for users to quickly understand how much a book is liked or disliked by audiences. The feedback we want to hear back is how easy it was to switch books, as well as if they can recall any of the rating information after using the tool.

The third usability task is for the user to use the line graph in order to name two books published after 2004 and tell us which of the two books generated more sales. The use of this task is to see how easily a user can navigate through our visualization. The first step is seeing how recognizable our graph is to a user. Can they see that this is the line graph that has sale data by the year a book was published? The second part is being able to tell how readable the hover functions are. Although we know where to look, it's relevant for us to be able to tell if a first-time user can draw their eyes to the hover spot to find a book title. Lastly, the task serves as a sanity check to make sure that readers can see the years and find the relevant information we have planned to include in the graph.

RESULTS

According to the group that conducted our usability testing, there were a few small issues with each of our visualizations, but there were no major issues. The users gave us good feedback regarding each of our visualizations as well as each of our specific tasks that we asked them to take a look at. They liked the uniqueness and creativity behind our visualization ideas. In general, they pointed out that we needed to make the description of the visualizations a bit more clear and fix a few technical issues.

For the first task, we asked our users to attempt viewing which genres tend to have the most popular books using the bubble chart and hovering over them to see the genres and the count of books from our dataset with that genre. They mentioned that this visualization looks good and it is very creative. They were successfully able to see which genres had the most popular books. One thing they mentioned that we could improve upon was adding an explanation regarding what the count means for the tool tip. They also suggested that we move the tool tip to the side of the visualization since the visualization is so large a user cannot see the tool tip with the information unless they scroll down below the visualization, which is a little inconvenient for the user. We believe that the test results for this task indicate that some small changes to our visualization design are necessary.

For the second task, we asked our users to attempt to see how user ratings compare for two different books. They should be able to do this by choosing a book from a drop-down menu and viewing the rating data in a bar graph below. The feedback we got from them was the following, "We like the drop-down feature and the way it is set up. To improve this visualization, we would add x and y labels to make it clearer what you are measuring. Some of the books on the drop down don't produce bars, so make sure to fix that issue." This feedback was very helpful and highlighted a somewhat large issue with this visualization. We believe that based on their feedback, no major changes to our design are necessary, but we will have to make a few tweaks to this portion of the visualization.

For the third task, we wanted our users to use the line graph in order to name two books published after 2004 and see which of the two books generated more sales. Our users told us that this visualization is unique and clean. They were able to ascertain metrics regarding book sales, meeting our expectations for that outcome. However, they mentioned that we should make it clearer what the line vs. dots represent in the visualization. They also noted that some of the book titles have ", The" after the rest of the book title, rather than before it. Based on their feedback on this task, we believe that some little changes to our design is necessary. In general, we were successful with our usability testing as we got a lot of good feedback and specific suggestions on how to improve our visualization. We will use this feedback to make modifications before sharing our visualization tool with others.

Based on our usability testing, there are some modifications we will make to our visualization. Regarding the first task, we will add a bit of an explanation regarding the definition of book counts for our bubble visualization. We will also move the tool tip from below the visualization to overlay on top of the bubbles so that it is more clearly visible for our users. Based on the outcome of the usability testing for the second task, we will slightly modify the book rating bar chart. We will add x and y axis labels to make it more clear for the user what we are measuring for the books, and we will also be sure to fix the code so that the data shows up in bar format for each of the books included in the drop-down menu. For the third task, we will make a few modifications to the sales line chart with dots. Specifically, we will add an explanation above the visualization describing what the difference is between the line and dots and what they represent. We will also work to change the titles so that the “The” appears before the rest of the title, rather than after it. We will make this change in order to make our visualization cleaner and make more sense to users.