Process Book

Part 1: Project Proposal

Basic Info:

Project title: The Influence of the Digital Age on the Book Market
Team members: Jessica Murdock, Michael Gardone
Email addresses: u0973401@utah.umail.edu, u1000771@utah.umail.edu
UIDs: u0973401, u1000771
Repo: https://github.com/jemurdock/DataVisProj

Background and Motivation. How people consume media, namely books in regards to this project, has changed dramatically since the rise of the Internet and electronic alternatives. There have been rising debates about the sustainability of public libraries and brick and mortar bookstores, especially independent booksellers. We wanted to understand how these trends relate to each other. Rather than focus on just one possible influence on the bookselling industry, we want to form a more comprehensive view on the many forms books can now take, and how each is faring in the digital age.

Project Objectives. Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits.

- How have brick and mortar bookstore sales changed with the rise of ebooks and audiobooks?
- Do younger generations read more ebooks or physical books?
- How has e-commerce impacted publishing?
- Do people buy more books for recreational or educational purposes?
- How well have small or independent bookstores reacted with the changing marketplace?
- How much use do libraries see?

Data:

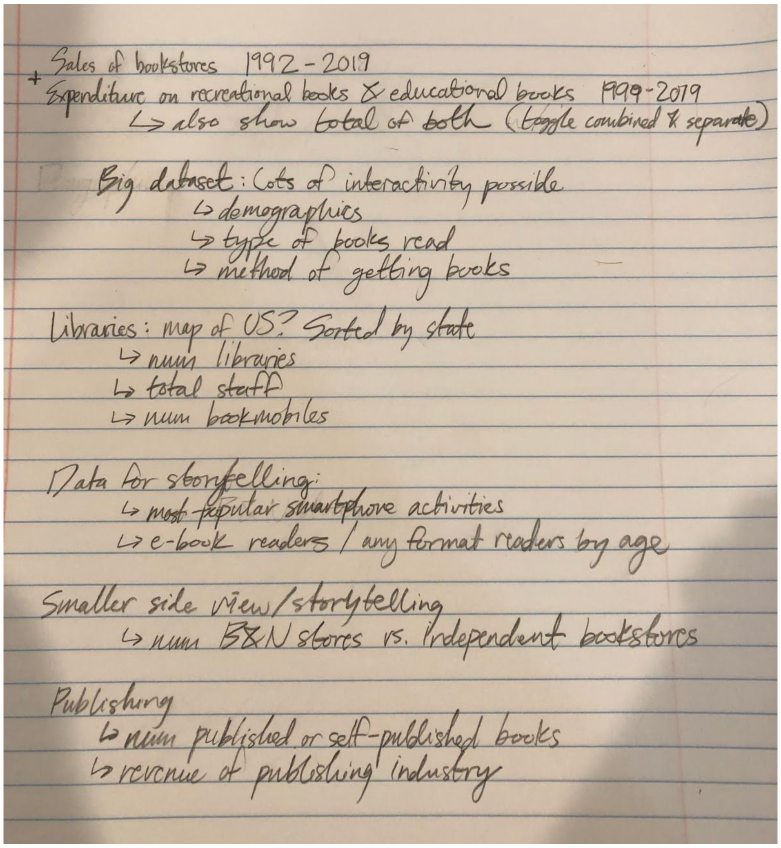
- US Census Bureau
  - https://www.census.gov/data/tables/2018/econ/arts/annual-report.html
- Pew Research
  - https://www.pewresearch.org/fact-tank/2019/09/25/one-in-five-americans-now-listen-to-audiobooks/
- Kaggle
  - https://www.kaggle.com/vipulgote4/reading-habit-dataset?select=BigML_Dataset_5f50a62795a9306aa200003e.csv
  - https://www.kaggle.com/imls/public-libraries
- Statista
  - https://www.statista.com/statistics/197710/annual-book-store-sales-in-the-us-since-1992/
  - https://www.statista.com/statistics/282808/number-of-independent-bookstores-in-the-us/
  - https://www.statista.com/statistics/271931/revenue-of-the-us-book-publishing-industry/
  - https://www.statista.com/statistics/185246/estimated-expenses-of-us-book-publishers-since-2005/
  - https://www.statista.com/statistics/249787/book-reading-population-in-the-us-by-age/

- https://www.statista.com/statistics/192861/consumer-expenditures-on-recreational-books-in-the-us/
- https://www.statista.com/statistics/192867/consumer-expenditures-on-educational-books-in-the-us-since-1999/
- https://www.statista.com/statistics/605000/createspace-number-books-published/
- https://www.statista.com/statistics/605039/blurb-number-books-published/
- https://www.statista.com/statistics/605050/xlibris-number-books-published/
- https://www.statista.com/statistics/187128/leading-us-smartphone-activities/
- https://www.statista.com/statistics/199012/number-of-barnes-noble-stores-by-type-and-year-since-2005/
- https://www.statista.com/statistics/249767/e-book-readers-in-the-us-by-age/
- https://www.statista.com/statistics/237070/frequency-of-reading-e-books-on-an-ebook-reader-in-the-united-states/
- https://www.statista.com/statistics/222737/book-reading-trends-by-number-of-read-books/

Data Processing:

We'll need to do some manual combining/pre-processing, assisted in part with Python data wrangling scripts. From this, we'll derive demographic data about who reads books, how frequently, and in what format. Other data sources will be used to retrieve sales data about the books and publishing industries. Finally, we'll derive data by state concerning public library use.

Visualization Design.



Sales of bookstores 1992-2019

+ Expenditure on recreational books & educational books 1999-2019

↳ also show total of both (toggle combined & separate)

Big dataset: Lots of interactivity possible

↳ demographics

↳ type of books read

↳ method of getting books

Libraries: map of US? Sorted by state

↳ num libraries

↳ total staff

↳ num bookmobiles

Data for storytelling:

↳ most popular smartphone activities

↳ e-book readers / any format readers by age

Smaller side view/storytelling

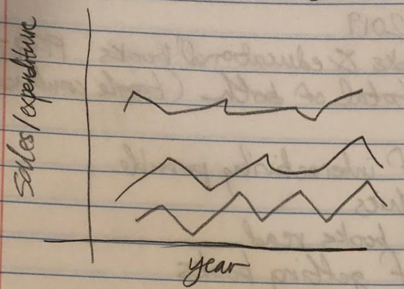
↳ num B&N stores vs. Independent bookstores

Publishing

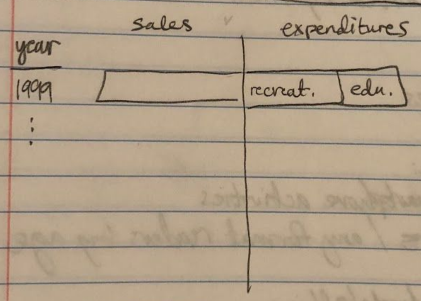
↳ num published or self-published books

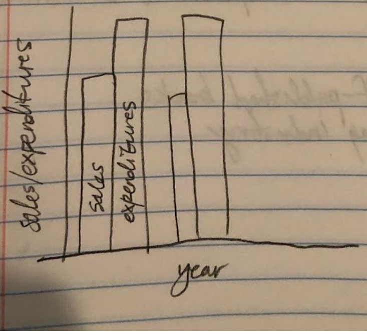
↳ revenue of publishing industry

Bookstore Sales View:

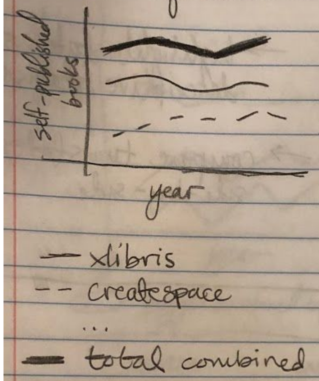


○ toggle combined expenditures





Publishing view:



Storytelling blurb on highlight point: publishing industry revenue

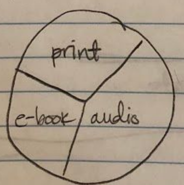
Storytelling box to side: overview of publishing industry

Big dataset: synchronized highlighting!

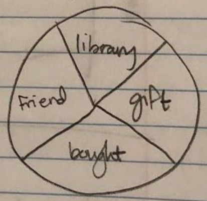
Demographics Table

Age	Race	Income	Edu	Marital	Books Read

% Read Book in Format

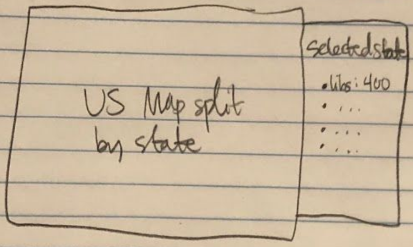


% got book from



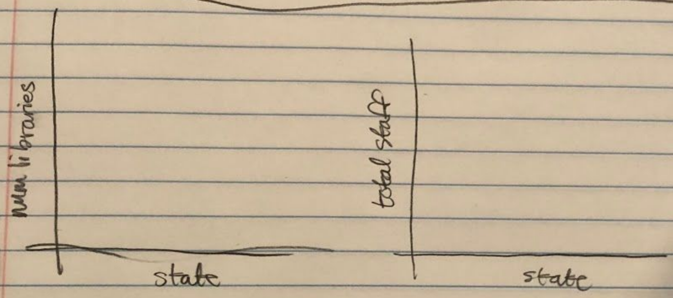
Overall statistics/storytelling for age group, etc.

Libraries view:

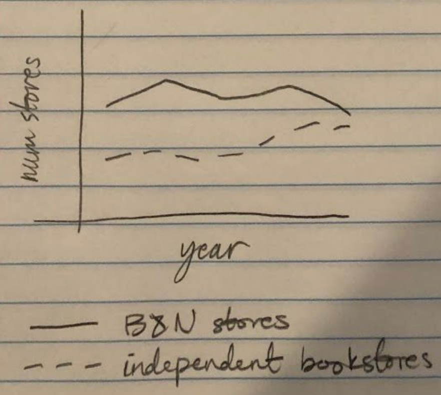


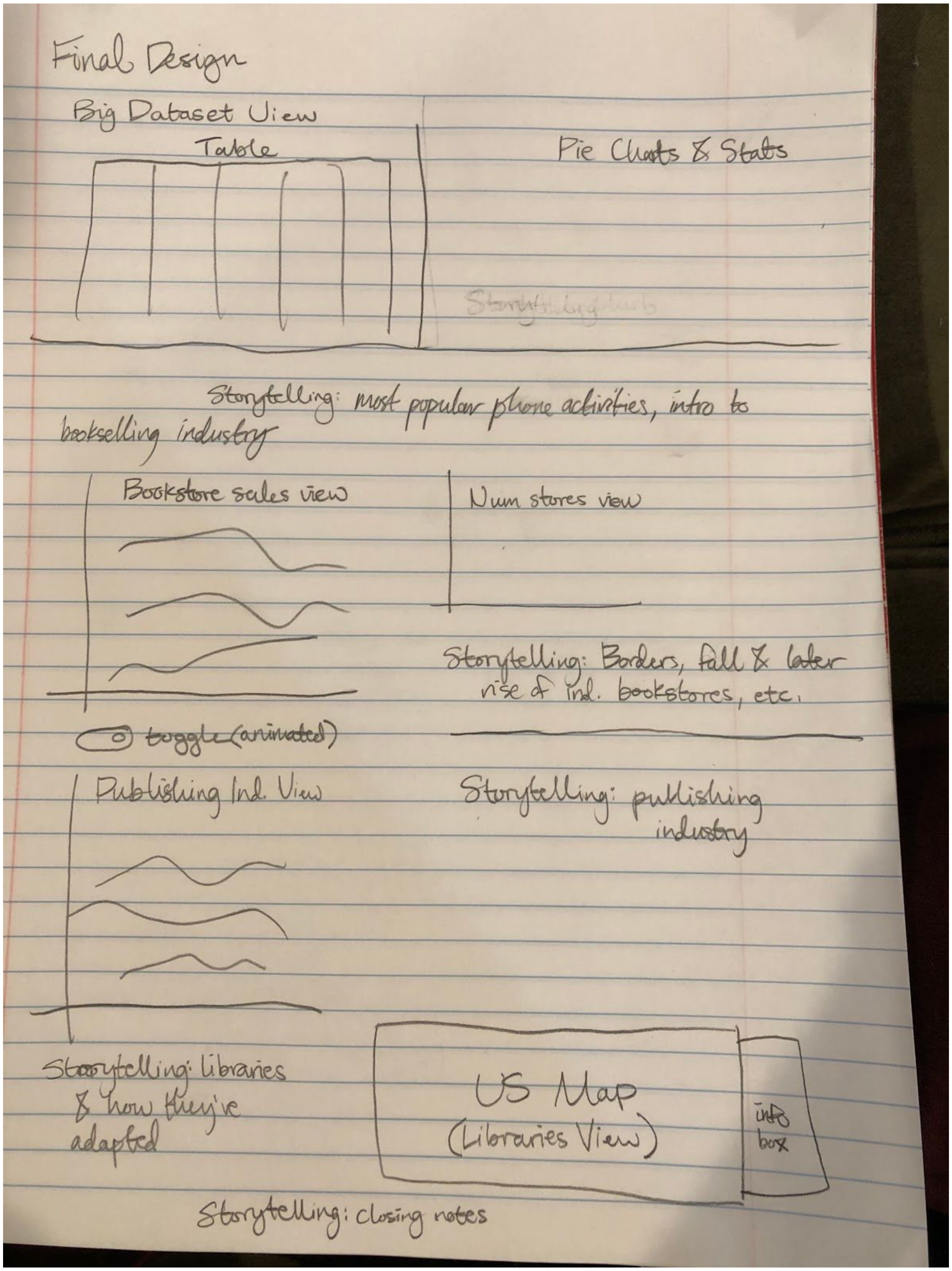
→ highlight state, show descriptive statistics

→ compare two states side-by-side



Side note: num stores




Final Design
Big Dataset View
Table
Pie Charts & Stats
Storytelling: most popular phone activities, intro to bookselling industry
Bookstore sales view
Num stores view
Storytelling: Borders, fall & later rise of ind. bookstores, etc.
toggle (animated)
Publishing Ind. View
Storytelling: publishing industry
Storytelling: libraries & how they've adapted
US Map (Libraries View)
info box
Storytelling: closing notes

- Overview of Customers - Allows us to show how consumers decide to read and obtain books and in what form. They'll let us show and see who are the usual suspects for a specific method.
    - Table & Filters - There is a huge dataset, and lots of features in it. We want to be able to easily distinguish and filter based on qualities both in our story and to allow the users to interact with it when they aren't in that view.

    - Pie charts & Stats - break down of demographic information based on the filters into a more digestible way. This data is part-of-whole, so we feel this warrants the usage of such representation.
- Bookselling Industry - Allow us to do a deep dive into seeing how stores like Borders and Barnes and Noble/publishing has changed
    - Line Charts - This is the usual way to encode timed data.
- Library View - This will let us see and talk about whether or not public libraries are useful (hint: they are).
    - Map of the US, color coded from white to some color (debating on that one) - to show the number of libraries in the state per capita.

Must-Have Features. List the features without which you would consider your project to be a failure.

- Comparison of brick and mortar bookstore sales vs ebook sales
- Demographic view for readers in the US
- Format view (how many people read print, audio, and e-books)
- Number of libraries by state view
- Publishing data view
- Highlighting, animations (mostly through the storytelling)
- Storytelling
    - Slide-show of what's happening, as per stats/events
    - Bookstores: barnes and noble dropping, independent bookstores rising
    - Popularity of e-books has plateaued
    - Reading on smartphones

Optional Features. List the features which you consider to be nice to have, but not critical.

- Revenue of book sellers (Amazon vs Barnes & Noble)

Project Schedule. Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

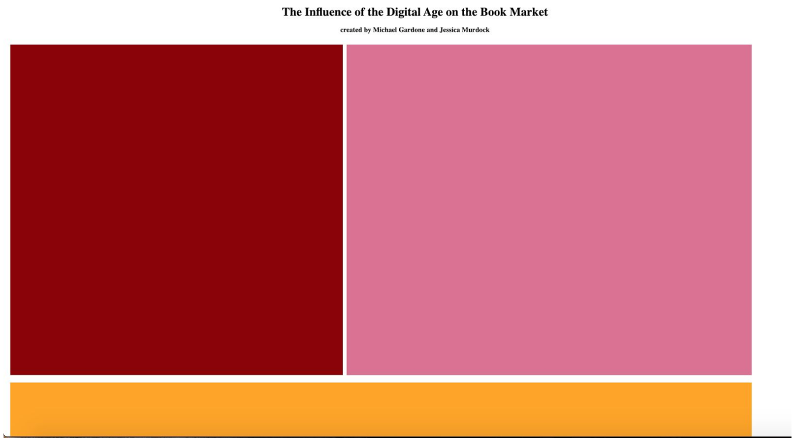
- Nov 1st-7th
    - All data wrangled
    - Brainstormed storytelling points
    - Basic views
- Nov 8th - 14th
    - More views
    - Create an easy way to add and remove from the story
- Nov 15th - 21st - Project Check In
    - Storytelling is in place for one view/category
    - All views in
- Nov 22nd - 28th
    - Storytelling is finished
    - Interactivity is finished as well

- Nov 29th - Dec 2 - Final Project Submission
    - Video

Part 2: Updates from Proposal to Milestone

Initial Steps:

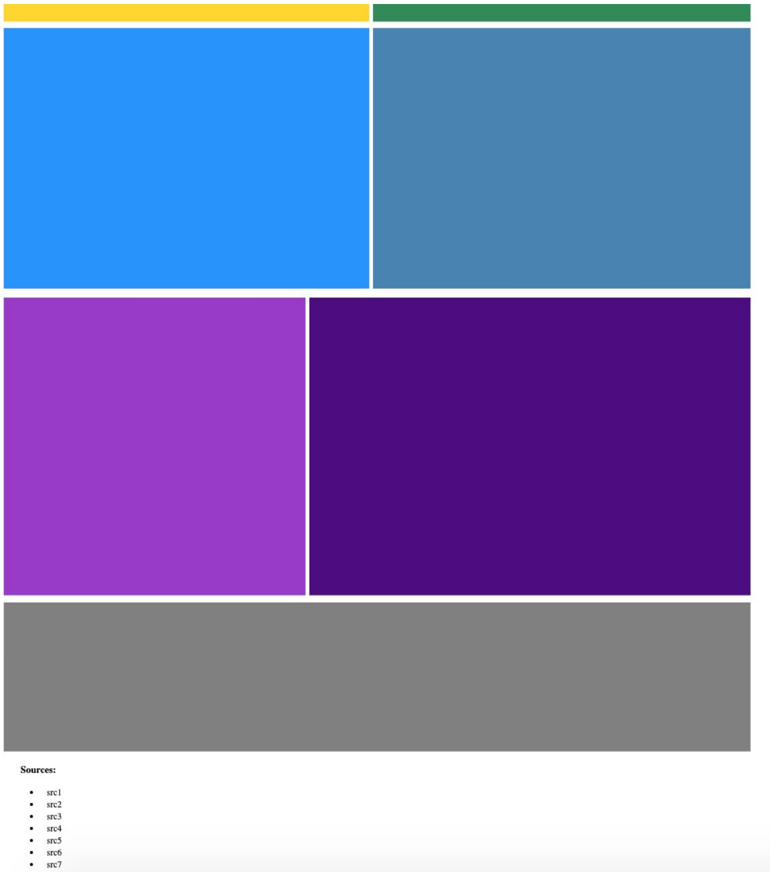
Seeing as we are using multiple data sets, many of which could not be used on their own, we did a lot of data wrangling to start, either by modifying/combining sets by hand or using scripts in Java or Python. We also started with a very simple html page with rectangles filled in where each graph or text area would be so that we could tweak the layout:



The Influence of the Digital Age on the Book Market

created by Michael Gardone and Jessica Murdock

...



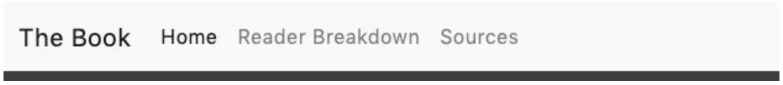
Sources:

- src1
- src2
- src3
- src4
- src5
- src6
- src7

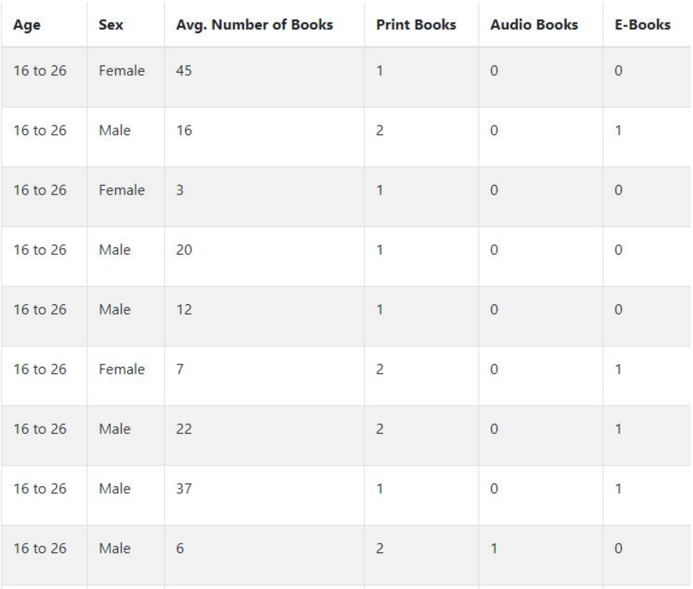
Once the data was processed, we split up elements of the project into classes to make it more manageable, and also to make it easy to avoid merge conflicts with github, since we could just work on different parts in different files.

Design Evolution and Insights:

We quickly realized that the table data wasn't quite what we thought it was after wrangling it into more manageable, more readable categories: the racial and income categories were incomplete, and other categories were very sparse. The table also wasn't going to fit easily in our original design, so we moved it to its own page and consolidated some categories to simplify the data. Our current design uses bootstrap styling and three tabs: the table and pie charts have their own tab, while the rest are on a "home" tab, and sources are listed on a third tab:



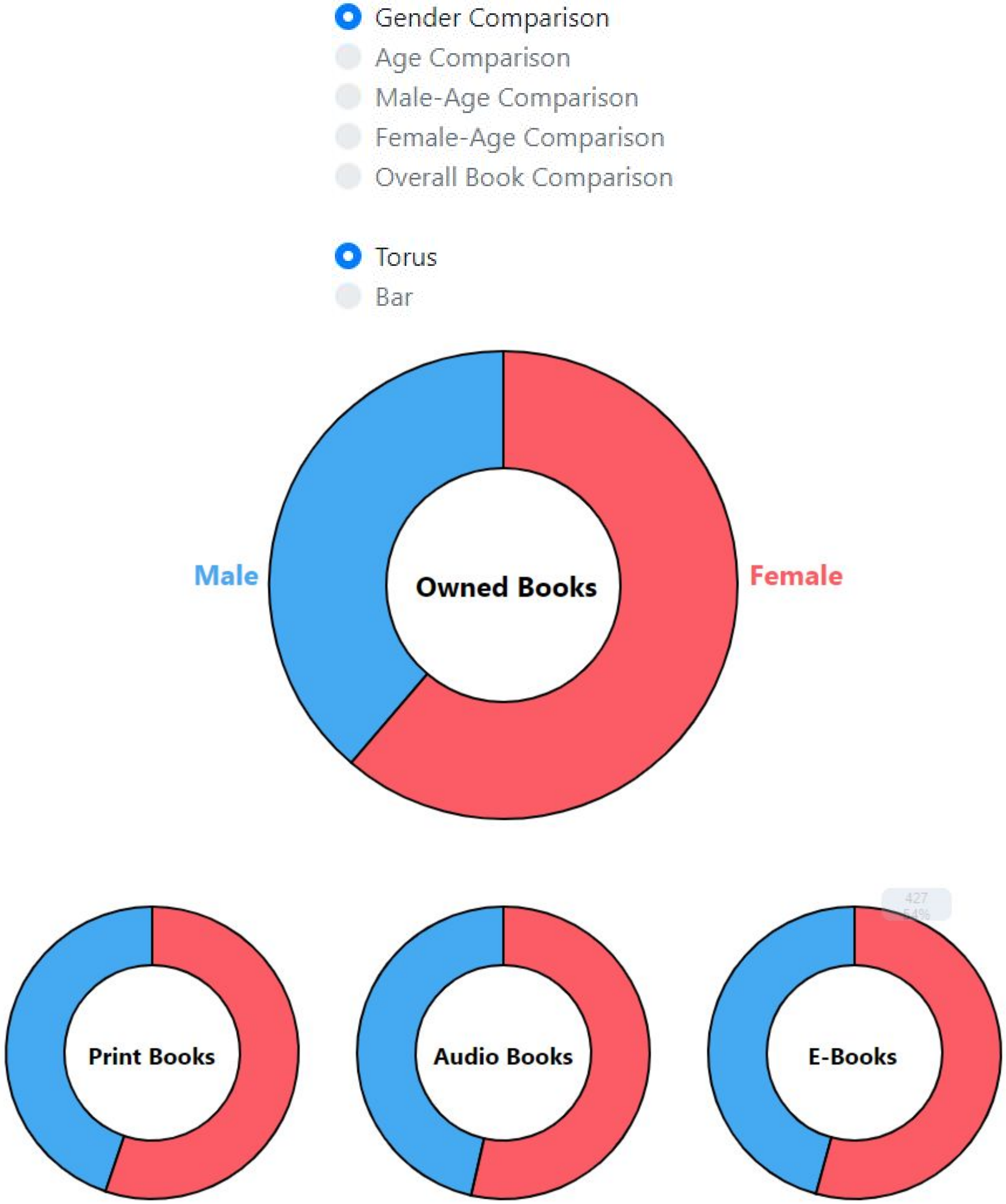
Continuing with that, by not including the data that was incomplete we were able to clean up the view of the Reader Breakdown page and we were able to show the two related views close together/side-by-side. While they are not linked like in the line charts (discussed further below), we are able to show off everything much more cleanly. The main reason for doing this is that D3 is not the best for trying to show off all the data in such a tight space, and it did not look clean at all.



Age	Sex	Avg. Number of Books	Print Books	Audio Books	E-Books
16 to 26	Female	45	1	0	0
16 to 26	Male	16	2	0	1
16 to 26	Female	3	1	0	0
16 to 26	Male	20	1	0	0
16 to 26	Male	12	1	0	0
16 to 26	Female	7	2	0	1
16 to 26	Male	22	2	0	1
16 to 26	Male	37	1	0	1
16 to 26	Male	6	2	1	0

To expand on showing data, we highlighted potential comparisons we feel are important with our data. This radial menu that shows the options is shown below, and we plan to include a bar graph version of our torus charts. We believe it's best to also look at consumer habits, not just general industry information as the consumers drive what the publisher's or retail stores do (or what the industry feels is best).

We want to also include a critique that was raised during the team sit downs a few weeks ago. Torus charts, while they are good for our data, may be difficult to read for those who are looking for the nitty-gritty details. To improve readability, we are also going to offer a bar chart version of the graphs that can be swapped between so people are really interested in seeing the data, in addition to the tooltip popup, can see it in a different way that is potentially easier to see the differences.

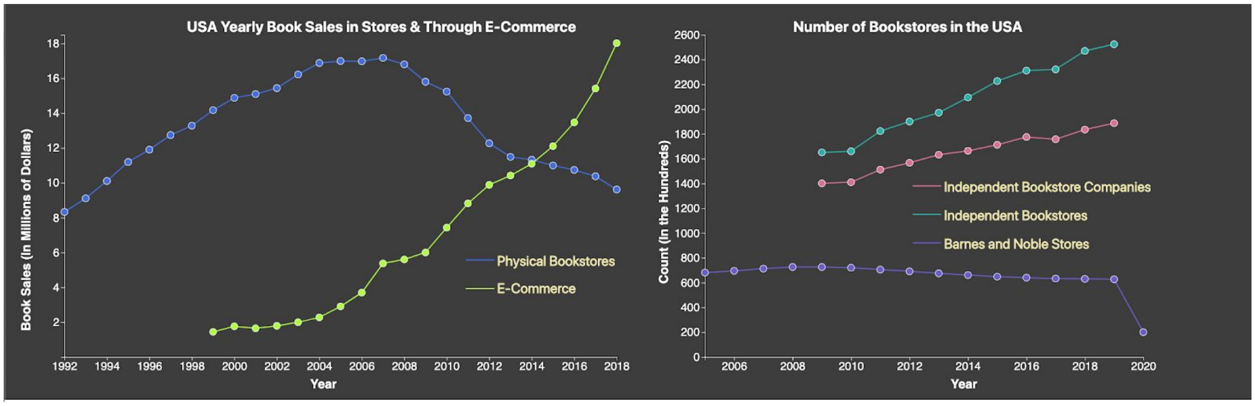


● Gender Comparison
○ Age Comparison
○ Male-Age Comparison
○ Female-Age Comparison
○ Overall Book Comparison

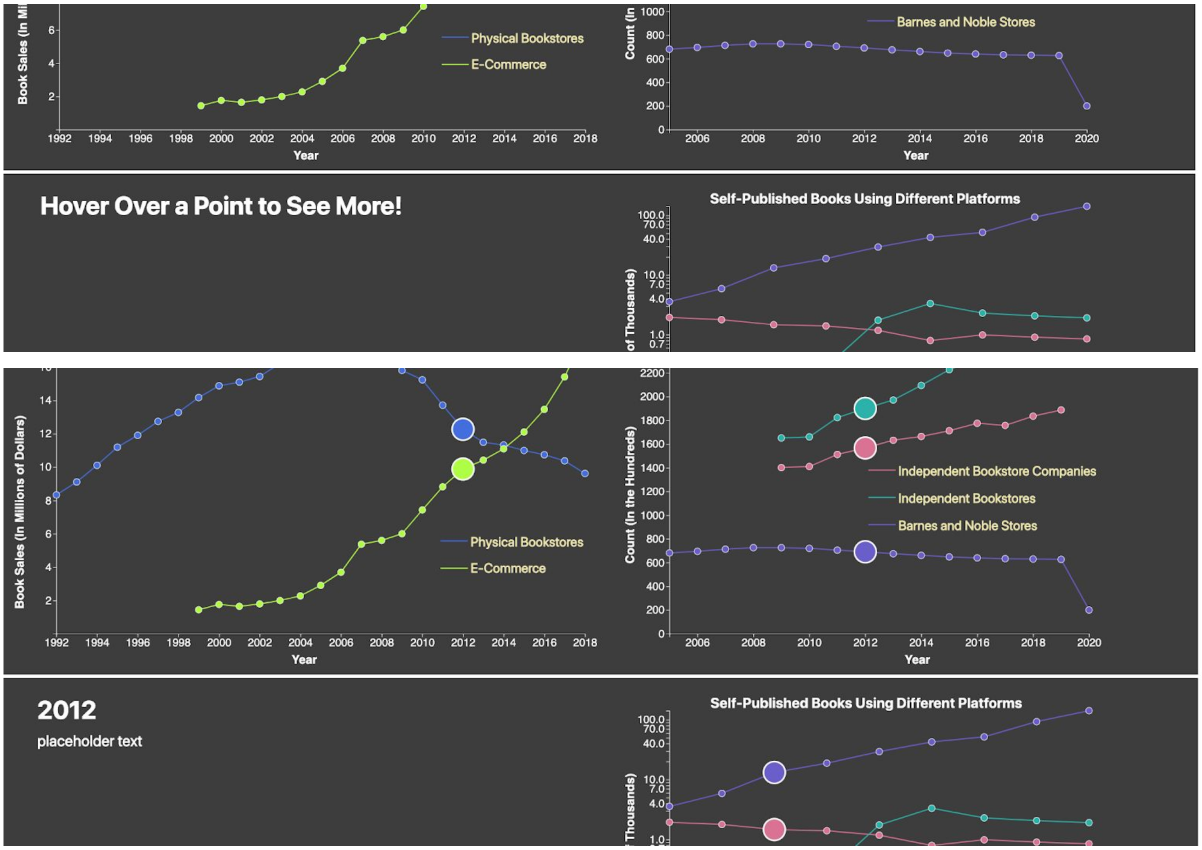
● Torus
○ Bar

The three line charts from our original design were made to match each other to keep a consistent design, using our css file. Once we got started, we realized that we hadn't considered how we were going to combine the two types of data we had for the publishing industry: we had data on the number of self-published books over time by a few self-publishing platforms, and data for the cost and revenue of the publishing industry in the US as a whole. For now, we've filled in the publishing industry graph in our original design using the self-publishing data, and are considering what to do with the rest, if anything.

Of the other two charts, one depicts the annual sales of books through stores vs. e-commerce and the other depicts the number of independent bookstores, companies and Barnes & Noble locations. Once these were up, we were surprised at first to notice that independent bookstores have actually been rising as Barnes & Noble has been falling, but we quickly found several articles on just that topic. The consensus is generally that people go to independent bookstores for different reasons than they go to bigger corporate stores or Amazon, so independent bookstores don't have to compete with Amazon, while Barnes & Noble does. People go to independent bookstores for the charm and atmosphere, for book clubs and community events, not simply to buy books. That seems to be the reason for this trend, and we will definitely want to include this insight in our storytelling. We also discovered that shortly after the pandemic started for the US, Barnes & Noble declared that they were shutting down 400 of their 627 remaining stores, which is the reason for the sharp decline at the end of the graph.



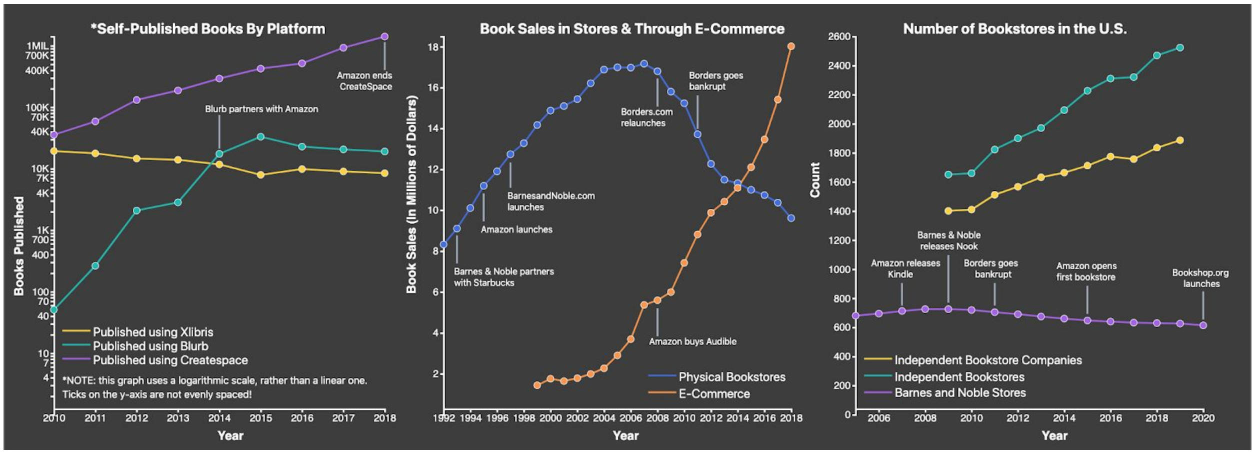
With all three graphs in place, we tweaked the original design to better fit what we were seeing: the number of stores graph and self-publishing graphs covered fewer years, so we shortened them and stacked them one on top of the other to make the page more compact and cleaner. We also realized that any storytelling we would want for a point in time for one graph, we would likely want to add information for the other two, so we converted the empty space left of the self-published graph into an info box. When the user hovers over a point, points across all three graphs are also highlighted, and the info box shows information for the year selected (for now, just placeholder text). Next, we want to add to the hovering selection so that it also shows the actual values of the data points selected, to make things more readable and more interactive.



Part 3: Updates from Milestone to Final Project

Initial Steps:

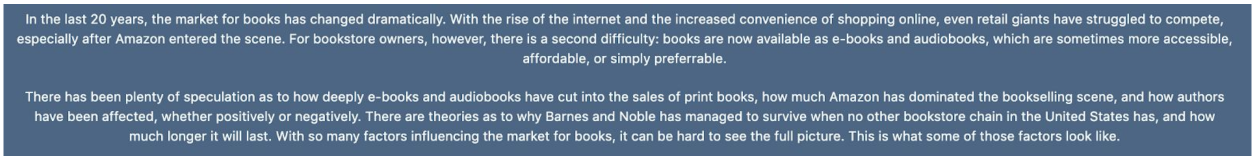
After the Milestone was submitted and we had our check-in, we started with a few cosmetic changes to the home page, updating the colors to be colorblind-friendly and agreeing on a common color palette for the whole project. We changed the hover selection so that the data point selected would be bigger than the linked ones, and shrunk the three graphs on the home page onto one line to make the page less cluttered.



We then removed the entire library map section, as it didn't seem as relevant or helpful as we initially thought, and would have added a lot more work that could have been better devoted to storytelling and interactivity.

Design Evolution and Insights:

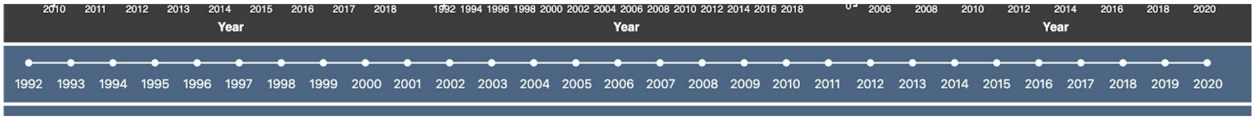
On the home page, we first decided on the text introduction to our website, giving an overview of the site and the topics we wanted to explore.



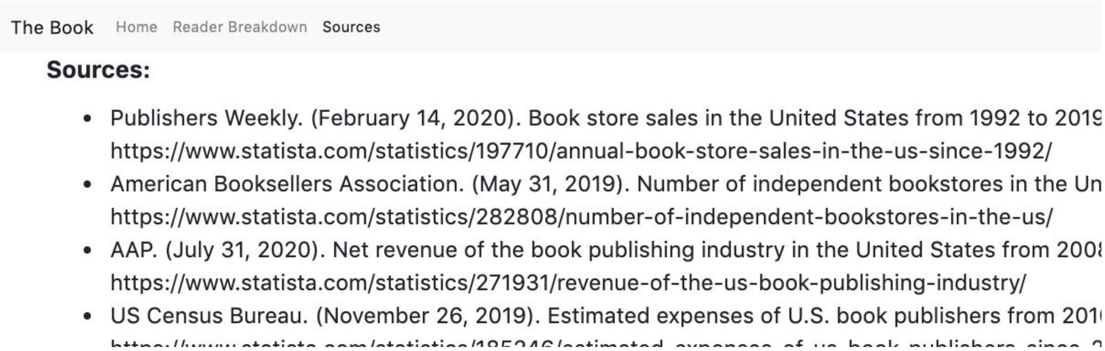
In the last 20 years, the market for books has changed dramatically. With the rise of the internet and the increased convenience of shopping online, even retail giants have struggled to compete, especially after Amazon entered the scene. For bookstore owners, however, there is a second difficulty: books are now available as e-books and audiobooks, which are sometimes more accessible, affordable, or simply preferrable.

There has been plenty of speculation as to how deeply e-books and audiobooks have cut into the sales of print books, how much Amazon has dominated the bookselling scene, and how authors have been affected, whether positively or negatively. There are theories as to why Barnes and Noble has managed to survive when no other bookstore chain in the United States has, and how much longer it will last. With so many factors influencing the market for books, it can be hard to see the full picture. This is what some of those factors look like.

We realized that a timeline would be a useful interactive feature to have so that, in addition to having the option of finding a specific data point of interest, a user could select a year of interest and see all of the relevant data points.



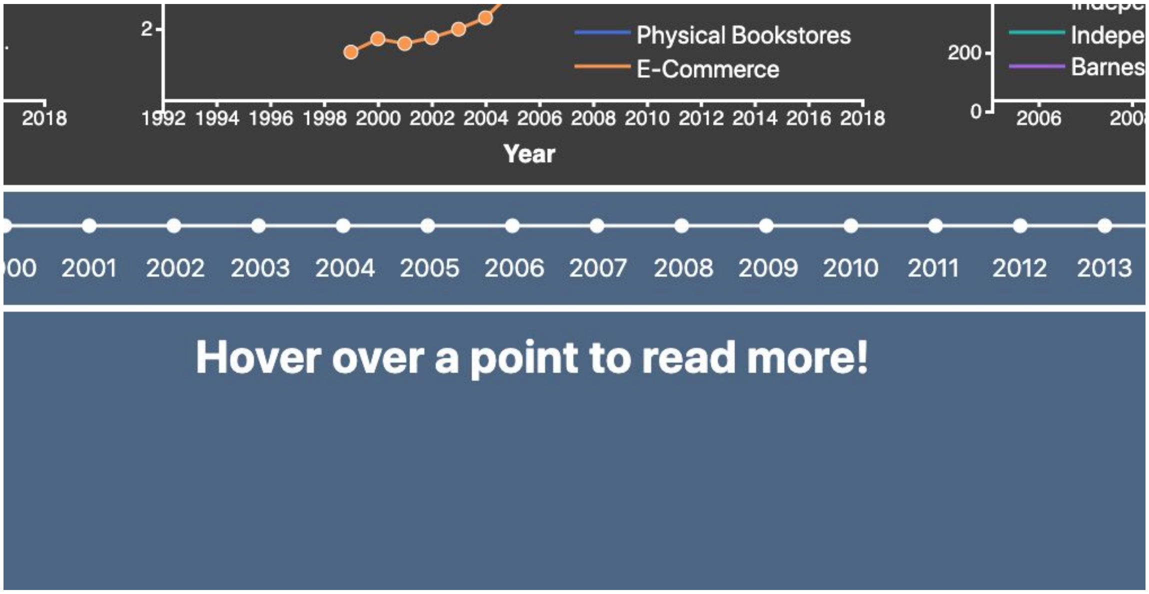
We also split the info box into three columns, one for each of the graphs, and researched the three relevant topics in more detail: all of our sources are included in the sources page. This yielded a lot of new interesting points: we discovered, for example, that Barnes & Noble was bought last year by a hedge fund that had bought the last living British bookstore chain, Waterstones, some time before. Waterstones used to be struggling immensely and was bleeding money much like Borders was right before it went bankrupt. But the CEO, James Daunt, was able to save it by making the stores as much like independent bookstores as possible. Stores were no longer kept all the same in decor and display and community event to engage a customer base: rather, stores were made unique by the individual employees, making them look and feel like indie stores. It worked so well that Waterstones has actually managed to expand again. Now, James Daunt has been put in charge of Barnes & Noble as well, so we wanted to mention this story in our storytelling info boxes.



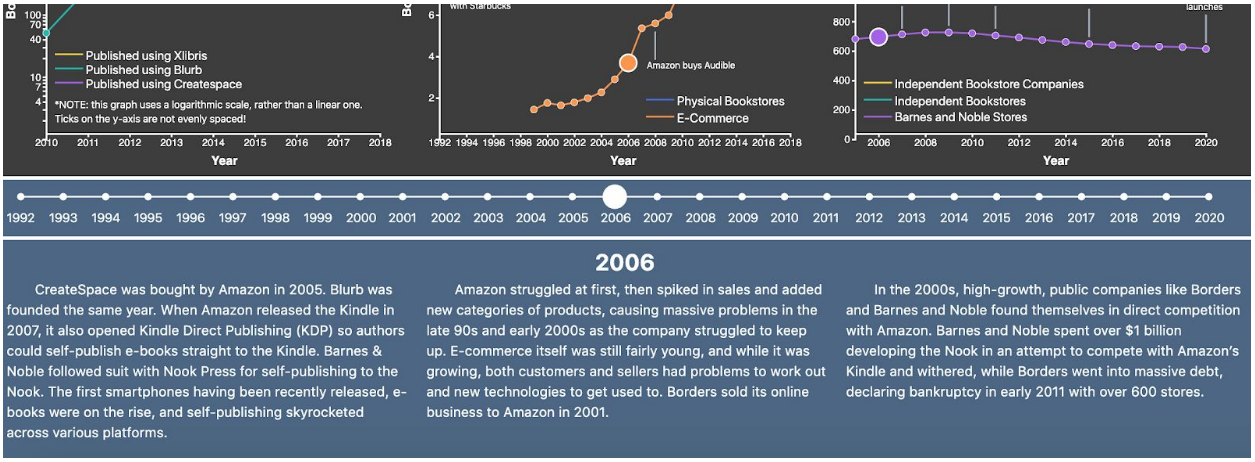
Sources:

- Publishers Weekly. (February 14, 2020). Book store sales in the United States from 1992 to 201
  https://www.statista.com/statistics/197710/annual-book-store-sales-in-the-us-since-1992/
- American Booksellers Association. (May 31, 2019). Number of independent bookstores in the Un
  https://www.statista.com/statistics/282808/number-of-independent-bookstores-in-the-us/
- AAP. (July 31, 2020). Net revenue of the book publishing industry in the United States from 200
  https://www.statista.com/statistics/271931/revenue-of-the-us-book-publishing-industry/
- US Census Bureau. (November 26, 2019). Estimated expenses of U.S. book publishers from 201
  https://www.statista.com/statistics/185246/estimated-expenses-of-us-book-publishers-since-2

Each info box changes depending on the year selected, offering relevant information that can be taken all together or individually: the info boxes are independent, so you don't need to read them in order or across columns to understand. This way, the user can choose just how much or how little they're interested in reading and it will still make sense.



Hover over a point to read more!

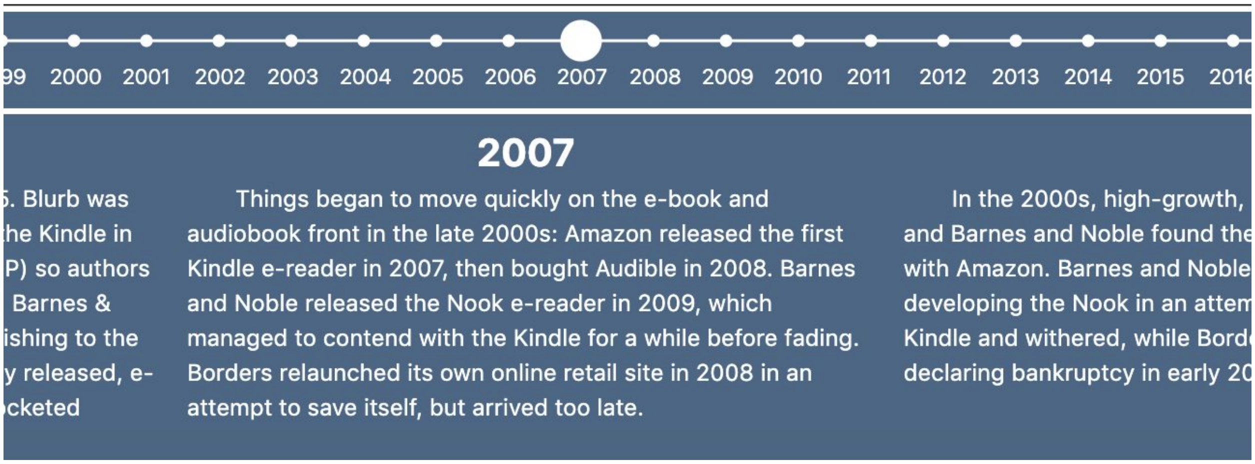


2006

CreateSpace was bought by Amazon in 2005. Blurb was founded the same year. When Amazon released the Kindle in 2007, it also opened Kindle Direct Publishing (KDP) so authors could self-publish e-books straight to the Kindle. Barnes & Noble followed suit with Nook Press for self-publishing to the Nook. The first smartphones having been recently released, e-books were on the rise, and self-publishing skyrocketed across various platforms.

Amazon struggled at first, then spiked in sales and added new categories of products, causing massive problems in the late 90s and early 2000s as the company struggled to keep up. E-commerce itself was still fairly young, and while it was growing, both customers and sellers had problems to work out and new technologies to get used to. Borders sold its online business to Amazon in 2001.

In the 2000s, high-growth, public companies like Borders and Barnes and Noble found themselves in direct competition with Amazon. Barnes and Noble spent over $1 billion developing the Nook in an attempt to compete with Amazon's Kindle and withered, while Borders went into massive debt, declaring bankruptcy in early 2011 with over 600 stores.



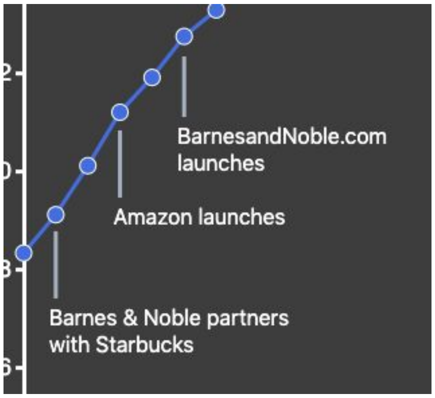
2007

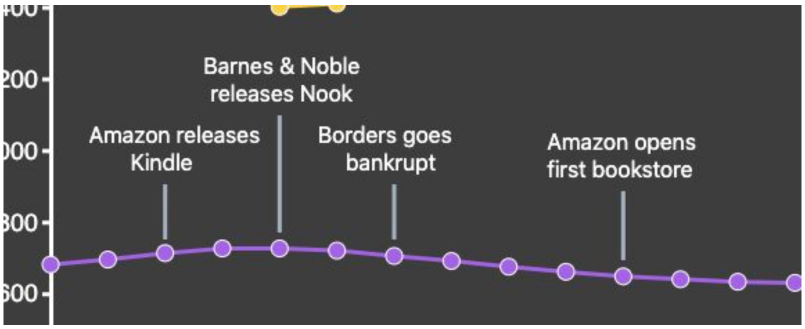
5. Blurb was
he Kindle in
P) so authors
Barnes &
ishing to the
y released, e-
cketed

Things began to move quickly on the e-book and audiobook front in the late 2000s: Amazon released the first Kindle e-reader in 2007, then bought Audible in 2008. Barnes and Noble released the Nook e-reader in 2009, which managed to contend with the Kindle for a while before fading. Borders relaunched its own online retail site in 2008 in an attempt to save itself, but arrived too late.

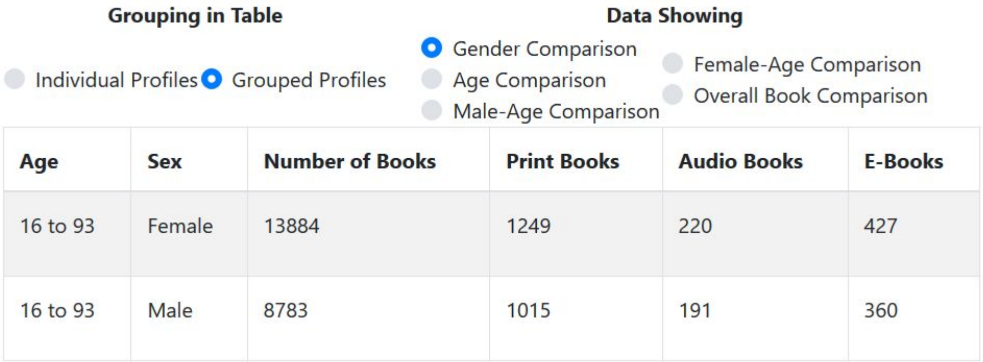
In the 2000s, high-growth,
and Barnes and Noble found the
with Amazon. Barnes and Noble
developing the Nook in an attem
Kindle and withered, while Bord
declaring bankruptcy in early 20

If the user doesn't feel like reading the info boxes at all, the simple annotations in the graphs themselves are a good quick-and-clean summary.





Moving over to the Reader Breakdown page, there have been some significant overhauls to the layout and way data is presented. First and foremost, we have expanded the selectable categories for the data to include various breakdowns that are important to observe consumer behaviors. In addition to this, a control toggle was set up for showing the grouped data



Grouping in Table

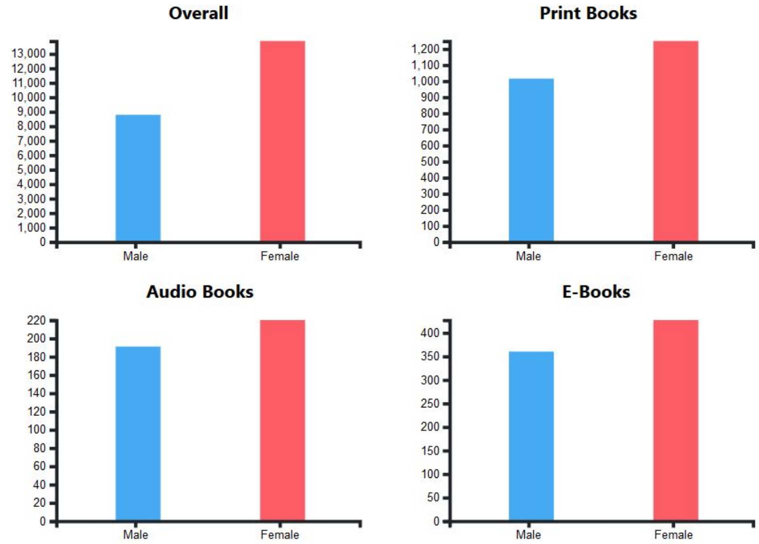
○ Individual Profiles ● Grouped Profiles

Data Showing

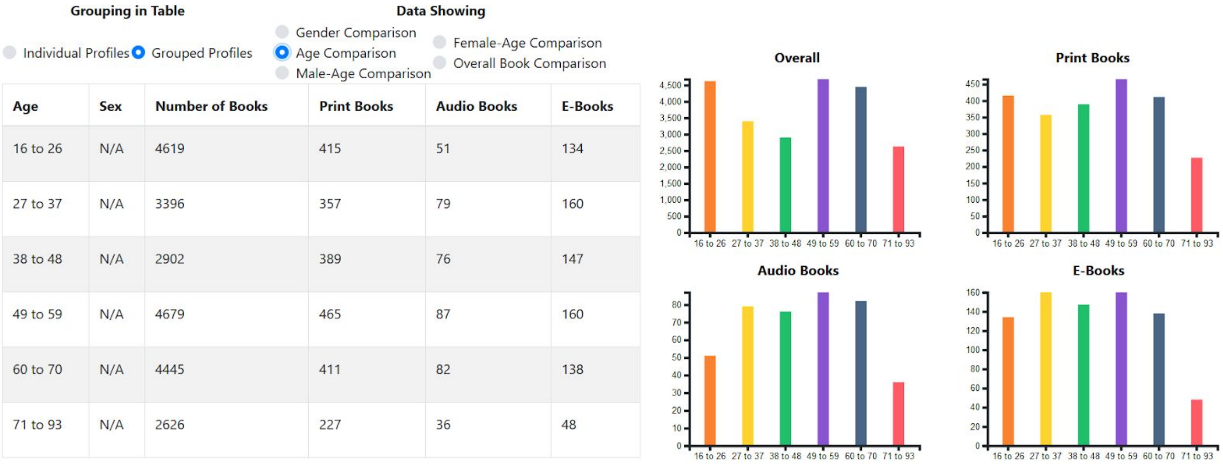
● Gender Comparison
○ Age Comparison
○ Male-Age Comparison
○ Female-Age Comparison
○ Overall Book Comparison

Age	Sex	Number of Books	Print Books	Audio Books	E-Books
16 to 93	Female	13884	1249	220	427
16 to 93	Male	8783	1015	191	360

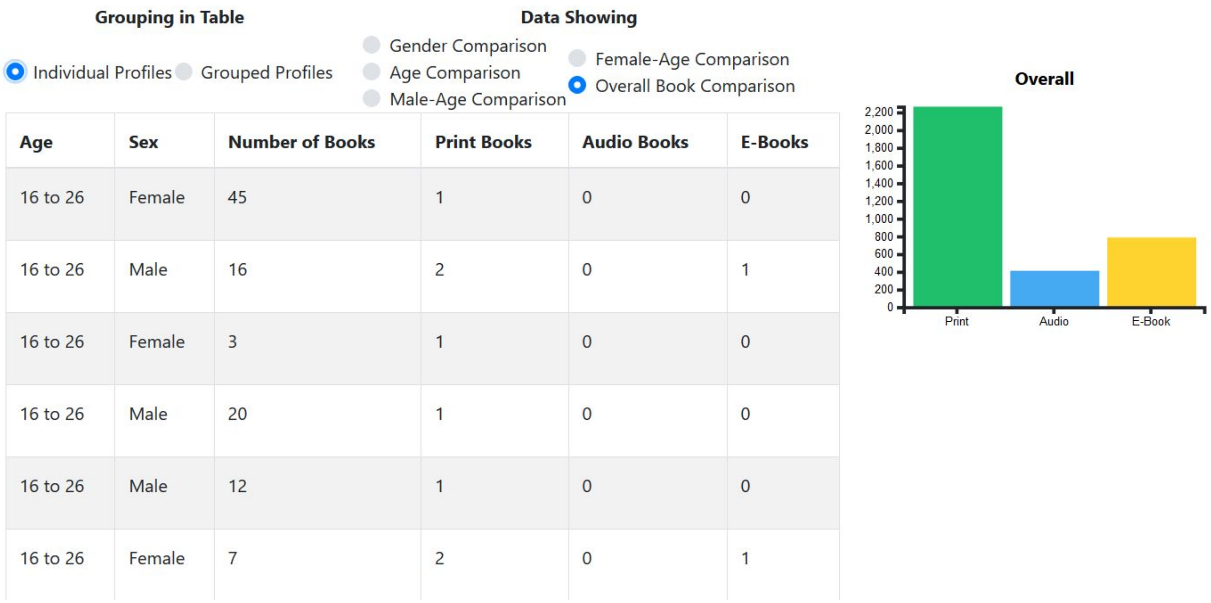
In addition to this, rather than including the tauruses, we opted to remove them. Now we have it as only the barcharts show. Part of this is mainly because attempting to set up the pie charts became a massive hassle when switching between data sets and another part because we found that having both was redundant. All data sets has four views to show off besides the “Overall Book Comparison” set, which has one graph to show off the total number of people who have consumed a print, audio, or e-book.



As well, switching which data shown will affect both the table and the graphs, which is to be expected. What switches the table view independently, however, is if "Individual Profiles" is toggled. In the screen shot just below the current one, you can see every individual correspondent record that the survey reported on.



Age	Sex	Number of Books	Print Books	Audio Books	E-Books
16 to 26	N/A	4619	415	51	134
27 to 37	N/A	3396	357	79	160
38 to 48	N/A	2902	389	76	147
49 to 59	N/A	4679	465	87	160
60 to 70	N/A	4445	411	82	138
71 to 93	N/A	2626	227	36	48



Grouping in Table

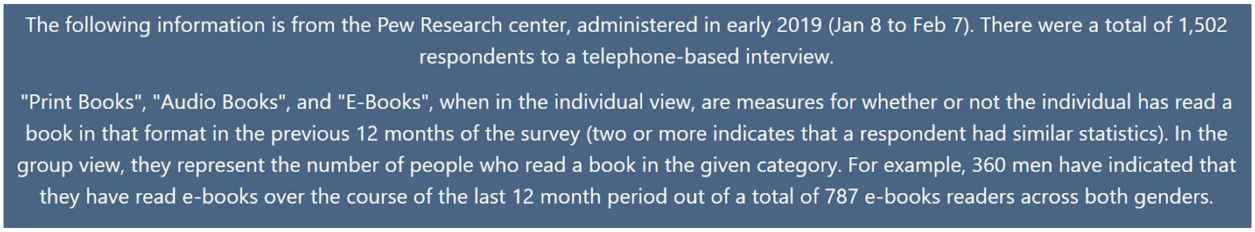
● Individual Profiles ○ Grouped Profiles

Data Showing

○ Gender Comparison
○ Age Comparison
○ Male-Age Comparison
○ Female-Age Comparison
● Overall Book Comparison

Age	Sex	Number of Books	Print Books	Audio Books	E-Books
16 to 26	Female	45	1	0	0
16 to 26	Male	16	2	0	1
16 to 26	Female	3	1	0	0
16 to 26	Male	20	1	0	0
16 to 26	Male	12	1	0	0
16 to 26	Female	7	2	0	1

We also included this header to provide a little extra information about the survey that we were using and what the two views mean as the headers in the table are easy to confuse/conflate. In addition to this, the source's information is hidden on the main website underneath a bunch of text so we wanted to put what the original survey was saying in front of it all.



The following information is from the Pew Research center, administered in early 2019 (Jan 8 to Feb 7). There were a total of 1,502 respondents to a telephone-based interview.

"Print Books", "Audio Books", and "E-Books", when in the individual view, are measures for whether or not the individual has read a book in that format in the previous 12 months of the survey (two or more indicates that a respondent had similar statistics). In the group view, they represent the number of people who read a book in the given category. For example, 360 men have indicated that they have read e-books over the course of the last 12 month period out of a total of 787 e-books readers across both genders.

The survey came with more information regarding race, income, and education, but these categories were either very broad (income ranges crossed thresholds of what is considered lower, middle, and upper class) or were incomplete (race data came up with only five options). We felt this would be disingenuous and not be useful information given how poorly these categories were created. We chose to focus on sex and age as these two were the most developed; each age bracket coincided with a period of life ("college age", "indepdent adult", "midlife", "retirement", "senior").