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)

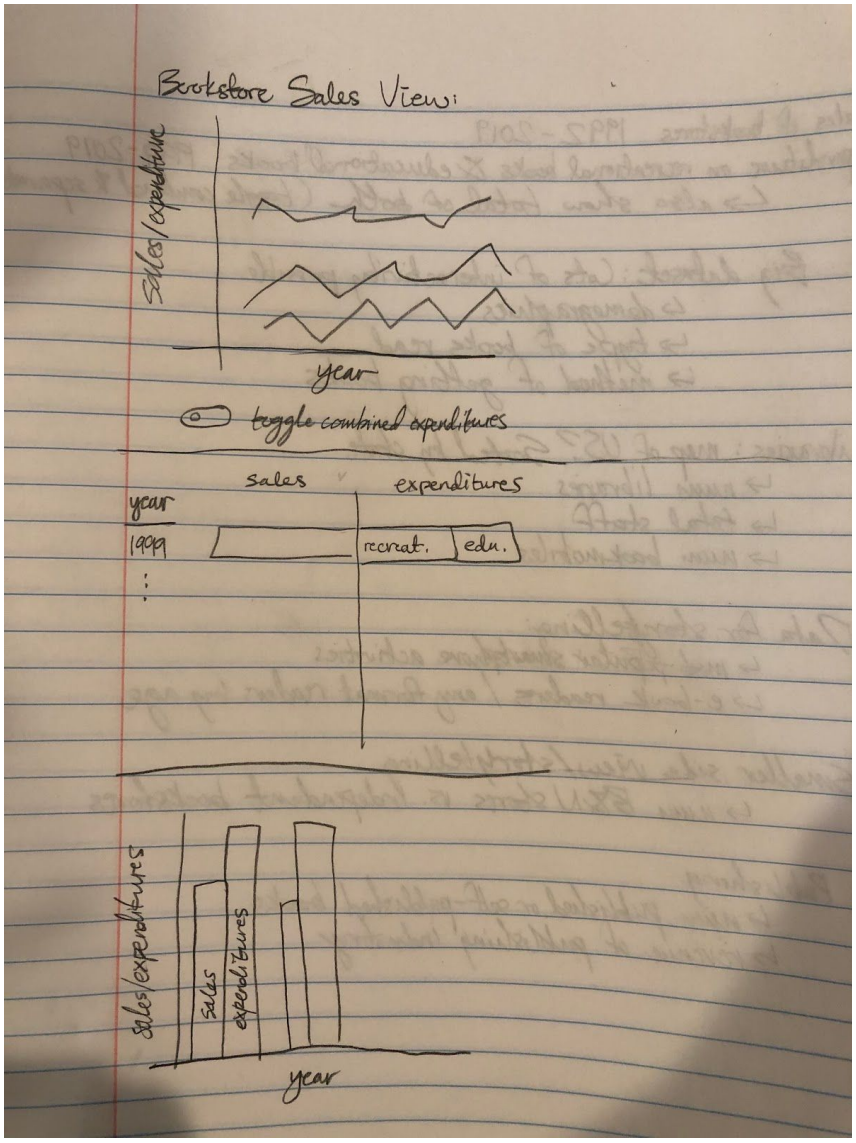
Big dataset: Lots of interactivity possible
↳ demographics
↳ type of books read
↳ method of getting books

Libraries: map of US? Sorted by state
↳ num libraries
↳ total staff
↳ num bookmobiles

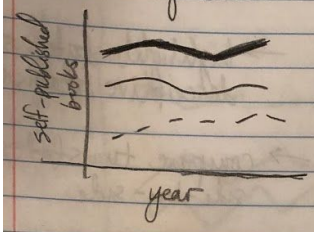
Data for storytelling:
↳ most popular smartphone activities
↳ e-book readers / any format readers by age

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:
sales/expenditure
year
toggle combined expenditures
sales
expenditures
year
1999
recreat.
edu.
sales/expenditures
sales
expenditures
year

Publishing view:



— xlibris
-- createspace
...
▬ total combined

Storytelling blurb on highlight point: publishing industry revenue

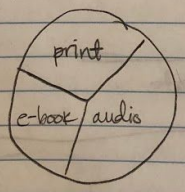
Storytelling box to side: overview of publishing industry

Big dataset: synchronized highlighting!

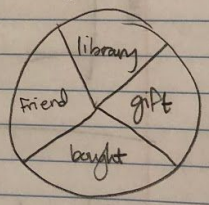
Demographics Table

Age	Race	Income	Educ	Marital	Books Read

% Read Book in Format

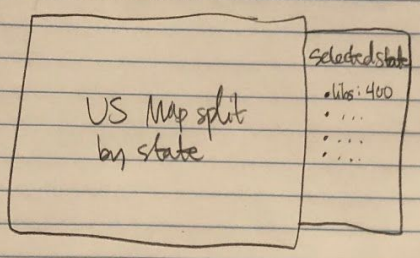


% got book from



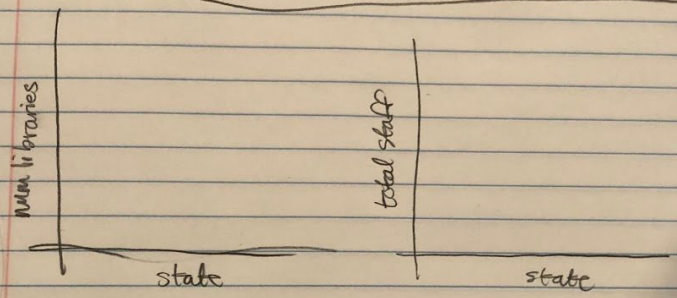
Overall statistics/storytelling for age group, etc.

Libraries view:

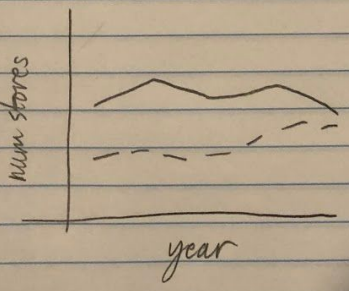


→ highlight state, show descriptive statistics

→ compare two states side-by-side

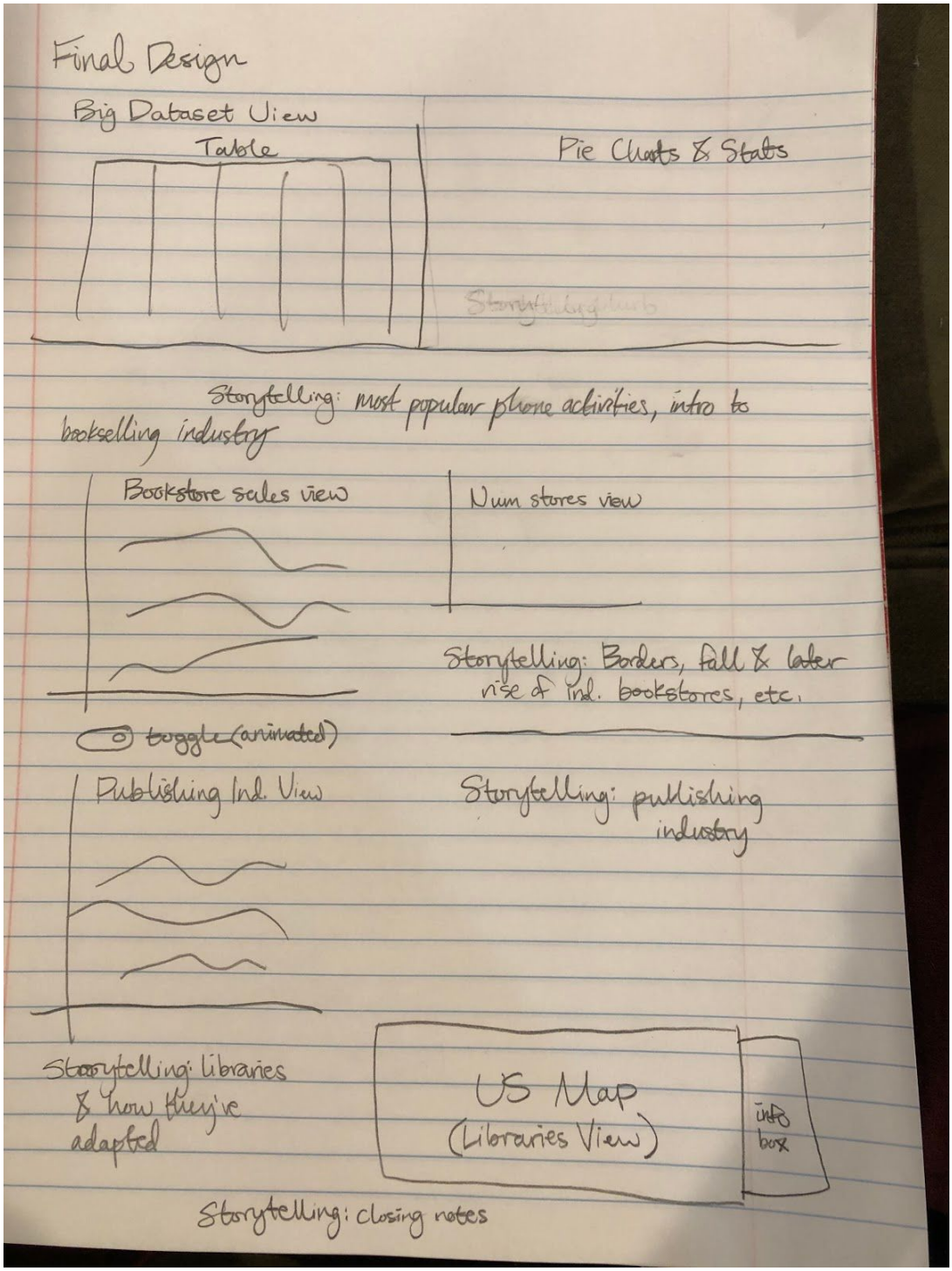


Side note: num stores



—— B&N stores

- - - independent bookstores



- 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.

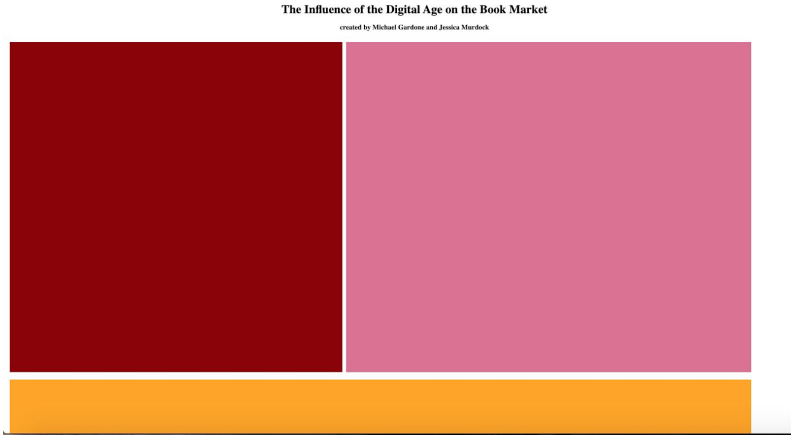
- Nov 1st-7th
    - All data wrangled
    - Brainstormed storytelling points
    - Basic views
- Nov 8th - 14th
    - More views
    - Create an easy way to add and remove from the story
- Nov 15th - 21st - Project Check In
    - Storytelling is in place for one view/category
    - All views in
- Nov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

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

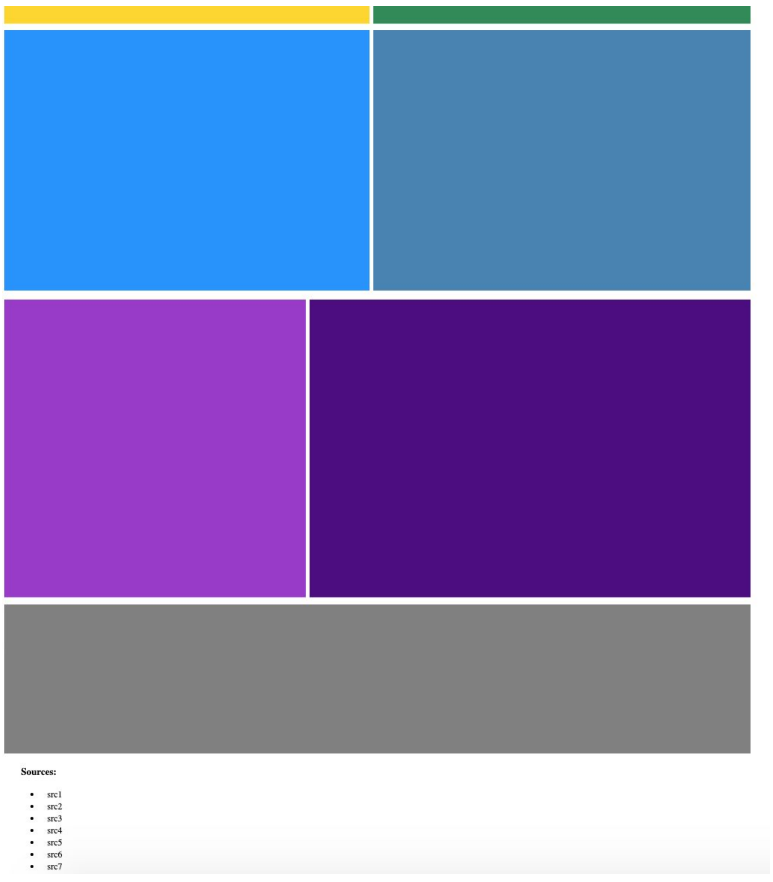
Part 2: Updates from Proposal to Milestone

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:



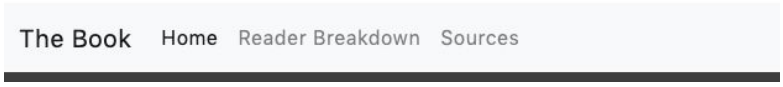
...



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:



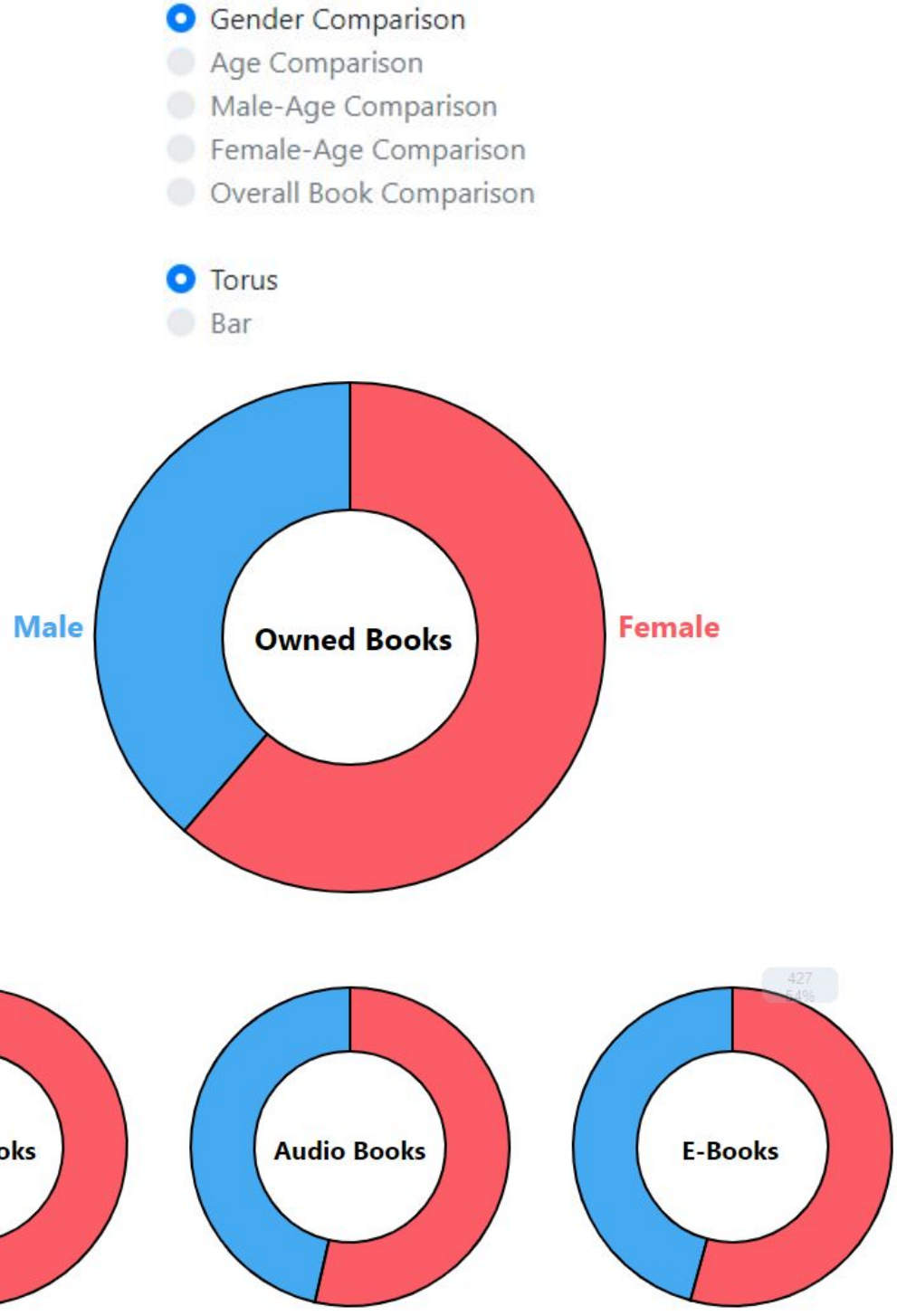
Continuing with that, by not including the data that was incomplete we were able to clean up the view of the Reader Breakdown page and we were able to show the two related views close together/side-by-side. While they are not linked like in the line charts (discussed further below), we are able to show 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

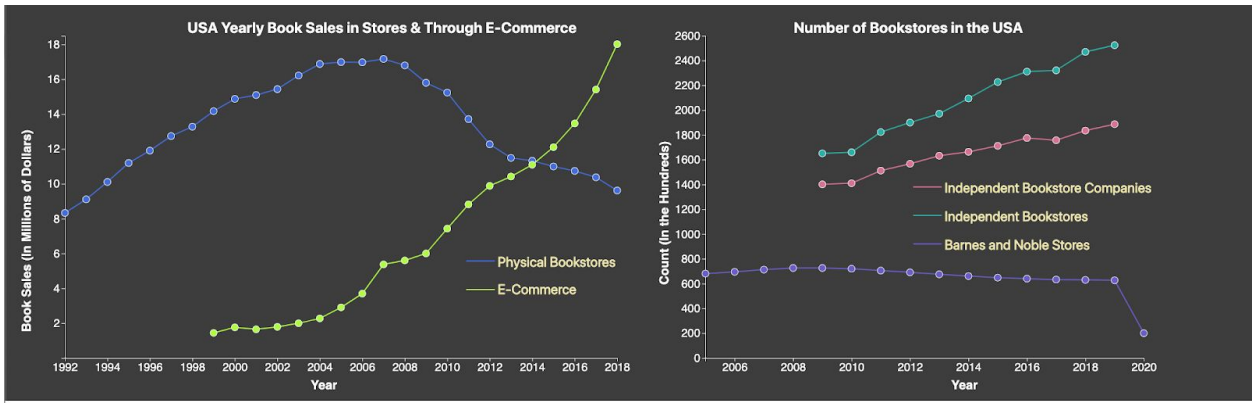
details. To improve readability, we are also going to offer a bar chart version of the graphs that can be swapped between so people are really interested in seeing the data, in addition to the tooltip popup, can see it in a different way that is potentially easier to see the differences.



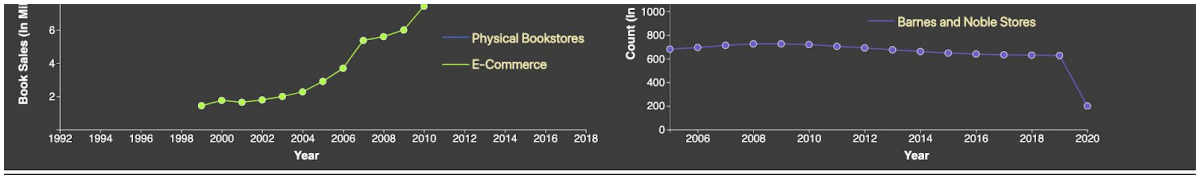
The three line charts from our original design were made to match each other to keep a consistent design, using our css file. Once we got started, we realized that we hadn't considered how we were going to

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.

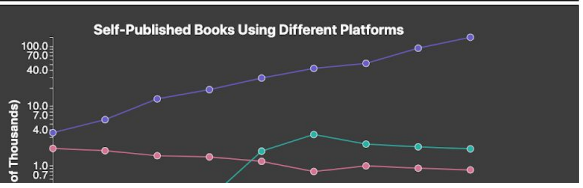
Of the other two charts, one depicts the annual sales of books through stores vs. e-commerce and the other depicts the number of independent bookstores, companies and Barnes & Noble locations. Once these were up, we were surprised at first to notice that independent bookstores have actually been rising 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.

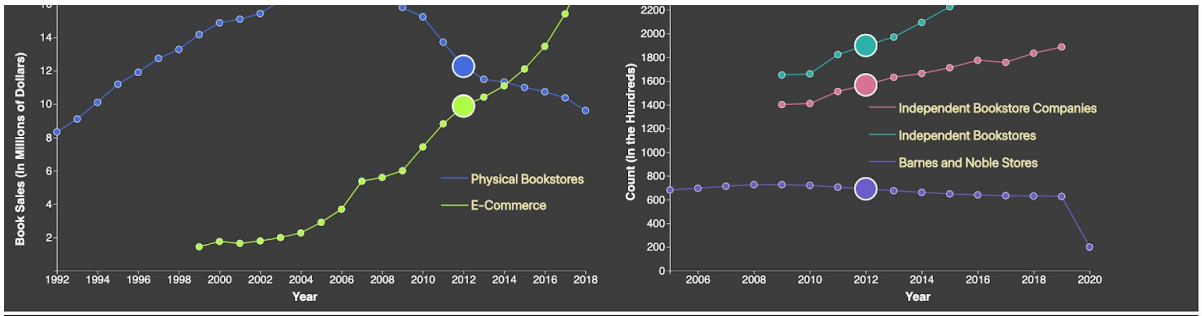


With all three graphs in place, we tweaked the original design to better fit what we were seeing: the number of stores graph and self-publishing graphs covered fewer years, so we shortened them and stacked them one on top of the other to make the page more compact and cleaner. We also realized that any 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.



Hover Over a Point to See More!





2012

placeholder text