Link: Process Book CS171

Process Book - CS 171

Team Members:

Anchal Bhardwaj abhardwaj@college.harvard.edu
Chloe Manilay chloemanilay@college.harvard.edu
Somto Unini sunini@college.harvard.edu
Oakley Browning oborwning@college.harvard.edu

Team Agreement:

- Roles and Responsibilities:
    - Point people:
        - Anchal - statistics, data visualization
        - Chloe - front-end design
        - Somto - data storytelling, big picture project planning
        - Oakley - data preparation and modelling
    - All code should be properly documented and all members will contribute to the technical aspect of the project. Point people are meant to check in on each task.
- Communication protocols
    - iMessage group chat as primary communication.
    - 24hrs response time, cancelling meeting attendance 24hrs in advance unless it's an emergency.
- Working arrangement
    - Primary worktime will be during the lab so attendance is expected.
- Work allocation
    - Point people for each of the separate tasks, work should be evenly split.
    - Group meeting to delegate tasks, async for actual work time.
    - All design decisions will be agreed upon unanimously by all team members.
- Accountability
    - Have a running progress check in that we can start each weekly meeting.
    - Checking in via text to make sure work is being completed.

Signatures:
Anchal Bhardwaj, 09/08/2025
Chloe Manilay, 09/08/2025
Somto Unini, 09/08/2025
Oakley Browning 09/19/2025

# Project Proposal:

1. Basic Information
    - **Project Title:** Who Has the Right to Read?
    - **Team Members:** Somto Unini, Anchal Bhardwaj, Chloe Manilay, Oakley Browning
    - **Team Name:** The Literacy Liaisons
2. Abstract

    Our goal is to explore the literacy landscape in the United States, especially considering the decreasing literacy rates in recent years. We want to study the factors behind this trend, any connections we can find to socioeconomic status or location, and the implications of decreasing literacy rates. For example, we can use public school funding data and early achievement scores, then overlap with the literacy gap map to understand how schooling can impact literacy. Another trend is among older crowds and how hobby reading in even higher socioeconomic regions has declined due to the digital age. Other trends can be observed by comparing literacy rates between different gender groups, among genres, and how accessibility to books plays a role.The types of visualizations we plan on creating are the literacy gap map to show literacy rates across the country, maps of library closures, maps of library social programs, documented censorship attempts over the years, and the different genres that people are consuming.

3. Background and Motivation

    Our team's shared interest in reading initially drew us to the question of why fewer people seem to be engaging with books today—particularly as social media trends highlight that younger children are not reading at the same levels as in previous years. One team member, pursuing a secondary in Education and having completed coursework on reading in schools and banned books, was eager to explore these issues further through a sociological lens. Collectively, we are also deeply interested in the socioeconomic discrepancies that shape literacy outcomes, especially given how strongly literacy rates predict future success. For some of us, this curiosity is personal, having grown up in a state where literacy rates varied dramatically, sparking a desire to understand the root causes of these inequities.

    Our project was inspired by a combination of course materials, prior work, and outside sources. The *Exploration of Linguistic Diversity* project from the course site introduced us to the ways language and identity intersect in literature, sparking our interest in literacy as a socio-cultural issue. One team member had also interviewed a transgender author who writes children's books that

thoughtfully introduce non-binary characters and themes, providing young readers with opportunities to understand and embrace diverse identities. In addition, the author discussed their resistance to book bans and censorship in states such as Texas, Louisiana, and Tennessee, which motivated us to further examine how representation in literature connects to literacy rates. We were also influenced by outside sources, such as a *New York Times* article on men and fiction, which led us to consider niche socio-cultural questions around male literacy, performative reading, and the challenges of sustaining reading habits in the age of AI. Together, these works shaped our desire to investigate how identity, culture, and access influence literacy outcomes today.

4. Data
    - In terms of collection/preprocessing plans, we will sort through data, clean it and organize it into different categories, drop irrelevant factors/rows if needed, and then use our cleaned data to experiment with visualizations. We'll have a copy of our master sheet and pull from there so that we always have the original data to return to in case we want to try different visualization approaches with different combinations/groups of data.
    - List of data sources:
        1. UNESCO Institute for Statistics (UIS)
            a. Most comprehensive global literacy data
            b. Covers adult literacy rates (15+ years) and youth literacy rates (15-24 years)
            c. Data available by country, region, and demographic breakdowns
            d. Historical data going back several decades
            e. Available through UNESCO's data portal
        2. World Bank Open Data
            a. Literacy rate indicators for adults and youth
            b. Easy-to-download CSV format
            c. API access available
            d. Includes metadata and methodology information
        3. Early Achievement Reading tests (+ correlation?)
            a. Searchable by district, demographic, grade level, etc.
        4. American Library Association
            a. Documents book ban or censorship attempts in 2024
        5. Whole Person Librarianship
            a. Map of social program adoption in public libraries

6. Public School Financing, Per Pupil Expenditure- can compare with the literacy rates to see how socioeconomic status impacts reading
7. New York Times Bestseller Lists
    a. Weekly lists across multiple categories (fiction, nonfiction, children's, etc.) Historical data back to 1931

Proposal Brainstorm

5. Provide a paragraph describing your project's main goals, the data you will use, and the type of visualization(s) you plan to create.
    i. Main goals
        1. Explore the literacy landscape in the US (factors, causes, trends, etc.)
        2. Results and implications of decreasing literacy
        3. (banned books, library closures, etc)
        4. Understanding the dichotomy between bad schooling leading to lower literacy rates, and hobby reading in higher socioeconomic regions declining due to the digital era.
        5. Understanding digital impact on literacy
        6. Comparison with school funding
        7. Comparing literacy rates between diff groups during the pandemic
        8. How much time is spent on the screen
        9. Gender gaps in literacy and hobby reading.
        10. Genre based distribution (what are we reading through time and who is reading it)
        11. Reading accessibility
        12. Rise in reading technology vs. actual literacy rates (number of published books vs sales).
    ii. Types of visualizations
        1. Literacy gap map, overlapped with socioeconomic data/school district map.
        2. Maps of library closures, library community programs
        3. Censorship map
        4. Library deserts
        5. Comparing top 100 books per year and break it down (think spotify wrapped) using New York Times Bestseller Lists

# Milestone 4: Project Map:

1. Who is your audience? Come up with **at least three** options and pick one target audience.

    1.1. Our target audience could be parents and educational stakeholders, general US consumers, or policy makers and community leaders - we chose **US consumers** as our target consumer.

2. Describe your target audience in more detail. What do they know? What are their interests? What visualization literacy do they have? At what level of detail will you present information to them?

    2.1. **Our target audience is US consumers aged 18-65 who have diverse relationships with reading and literacy.** This includes parents concerned about their children's education, adults who may have lapsed in their reading habits, avid readers worried about library closures, and digital natives navigating new forms of content consumption. They possess moderate to high visualization literacy from consuming news media, social platforms, and workplace dashboards, so they can interpret standard charts, maps, and interactive elements without extensive guidance.

    2.2. **What they know:** They understand the US public education system (K-12 grade levels, standardized testing, local school funding), are familiar with library services (checkout systems, programming, inter-library loans), and recognize major cultural shifts like the rise of audiobooks, e-readers, and social media. They're aware of ongoing debates about book banning, screen time concerns, and "kids these days don't read" narratives, but may lack deeper data to understand these trends.

    2.3. **What interests them:** Personal connections to the data - how their own community compares, whether their reading habits are "normal," what factors affect their children's education, and validation or challenge of their assumptions about literacy decline. They want actionable insights: should they be concerned about their local library's budget cuts? Is audiobook listening "real" reading? How does their social media use affect their attention span?

    2.4. **Information presentation levels:** We'll use a layered approach - headline statistics and clear visual patterns for quick consumption, with drill-down capabilities for those wanting deeper analysis. We'll provide multiple entry points: geographic (explore your state/region), demographic (compare your age group), and behavioral (analyze your reading format preferences). Technical literacy concepts will be explained through analogy and context rather than jargon, ensuring accessibility while maintaining analytical rigor for those who want it.

3. What questions about your data will be interesting for your audience? Come up with a list of interesting questions that your audience may have about your data. The more, the better, but your team should come up with **at least ten questions**.
    3.1. Are we more literate compared to a decade ago?
    3.2. Funding into the American public library system over time?
    3.3. How does socioeconomic status impact public library funding?
    3.4. Test scores for American students over time?
    3.5. How many books are people reading per year on average?
    3.6. Are there deserts where there are no bookstores/libraries available?
    3.7. Do men and women have different reading patterns?
    3.8. Do men and women like different book genres?
    3.9. What genres are most popular in the US?
    3.10. How much time is spent on the screen compared to reading for students?
    3.11. What age do students learn to read now compared to a decade ago?
    3.12. How many books are banned per year?
    3.13. What types of books are banned per year?
    3.14. What might be causing the decline of reading rates for American men?
    3.15. How is AI impacting the literacy rates of elementary students in American schools?
    3.16. What are the most popular types of books consumed by each generation?
    3.17. How do library hours/services correlate with community employment patterns (e.g., do working-class areas have libraries open when people can actually visit)?
    3.18. What's the relationship between local bookstore closures and community reading rates?
    3.19. How has the cost of books changed relative to median income over time?
    3.20. Do areas with higher internet costs show different digital vs. physical reading patterns?
    3.21. How do literacy rates correlate with local economic indicators (unemployment, median income, housing costs)?
    3.22. Are audiobook listeners considered "readers" in literacy statistics, and how does this affect the data?
    3.23. How do social media usage patterns correlate with book reading habits by age group?
    3.24. What's the relationship between smartphone ownership and reading frequency?
    3.25. Do students who use e-readers vs. physical books show different comprehension scores?
    3.26. How has the rise of short-form content (TikTok, Instagram) affected attention spans for longer reading?
    3.27. How do literacy rates vary between homeschooled, private school, and public school students?

3.28. What's the impact of school librarian presence on student literacy outcomes?
3.29. Do states with later school start times show different literacy development patterns?
3.30. How do summer reading program availability and participation affect literacy retention?
3.31. What role do parent literacy levels play in child literacy outcomes across different income brackets?
3.32. How do reading habits differ between urban, suburban, and rural communities?
3.33. Do communities with higher rates of English as a second language show different literacy patterns?
3.34. How do reading club/book club participation rates vary geographically and demographically?
3.35. What's the relationship between local news consumption and book reading habits?
3.36. How do religious communities' reading patterns differ from secular communities?
3.37. Are graphic novels and manga helping or hindering traditional literacy development?
3.38. How has the popularity of self-published books changed reading habits?
3.39. Do romance and fantasy readers (predominantly female genres) show different literacy metrics than typical "literary fiction" readers?
3.40. How do book banning attempts correlate with local political voting patterns?
3.41. At what life stages do people typically stop/start reading regularly?
3.42. How do reading habits change after major life events (parenthood, retirement, job loss)?
3.43. Do grandparents' reading habits influence grandchildren's literacy more than parents' habits?
3.44. How has the "summer slide" (literacy loss during summer break) changed over time?
3.45. Is there a correlation between reading frequency and reported mental health outcomes?
3.46. How do reading habits correlate with sleep quality and screen time before bed?
3.47. Do people who read regularly show different healthcare utilization patterns?

4. What data do you have or plan to obtain? Briefly describe the data you envision to use and the respective data types (categorical, ordinal, or quantitative) in your process book. It’s OK if you are unsure about the data type for some attributes - you can simply describe them (e.g., geographic location).

4.1. Core Literacy & Education Data

4.1.1. Literacy rates by state/region (quantitative) - percentage of adults reading at proficient levels
4.1.2. Student reading test scores (quantitative) - standardized assessment scores over time
4.1.3. Age when students learn to read (ordinal/quantitative) - grade level or age milestones
4.1.4. Books read per person annually (quantitative) - average number from surveys like Pew Research
4.1.5. Reading time vs. screen time for students (quantitative) - hours per day in each activity

4.2. Infrastructure & Access Data
4.2.1. Public library locations (geographic coordinates, categorical by type)
4.2.2. Library funding over time (quantitative) - budget amounts by location and year
4.2.3. Library closures (geographic, temporal) - dates and locations of closures
4.2.4. Bookstore locations (geographic coordinates, categorical by independent/chain)
4.2.5. Library social programs (categorical) - types of programs offered, frequency

4.3. Socioeconomic & Demographic Data
4.3.1. Median household income by region (quantitative)
4.3.2. School funding per district (quantitative) - dollars per student
4.3.3. Gender (categorical) - for analyzing reading pattern differences
4.3.4. Age/generation (ordinal) - Gen Z, Millennial, Gen X, Boomer categories
4.3.5. Geographic location (categorical/geographic) - urban/suburban/rural classifications, state/county identifiers
4.4. Content & Preference Data

4.4.1. Book genres (categorical) - fiction, non-fiction, romance, mystery, etc.
4.4.2. Reading format preferences (categorical) - physical books, e-readers, audiobooks
4.4.3. Banned books data (categorical, quantitative) - titles, reasons for banning, frequency by location
4.4.4. Popular book titles/authors (categorical) - bestseller lists, library circulation data

4.5. Technology & Social Media Data

4.5.1. Social media platform usage (categorical, quantitative) - platform types, hours of use

4.5.2. E-reader adoption rates (quantitative) - ownership percentages over time

4.5.3. Digital vs. physical book sales (quantitative) - revenue/unit sales by format

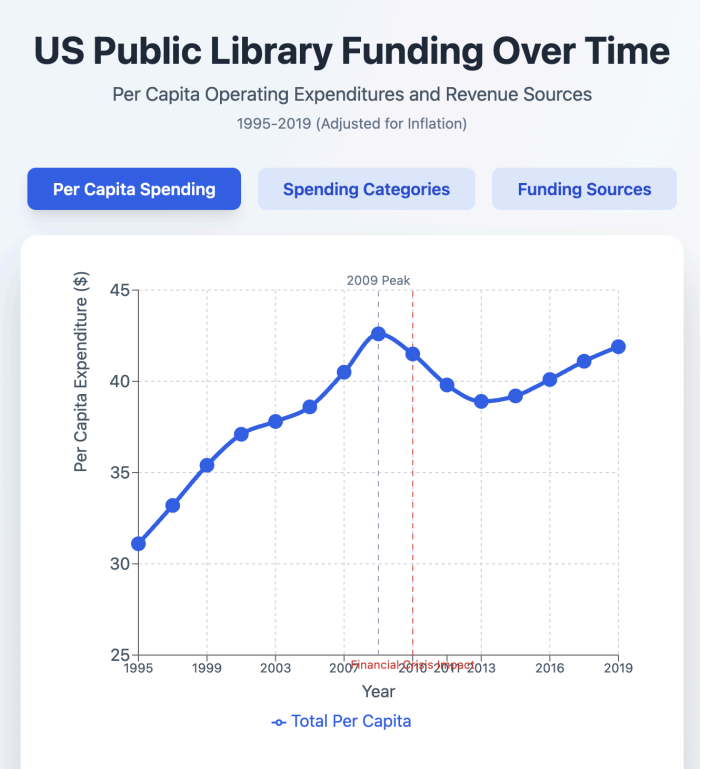
4.6. Most of this data would come from sources like the American Library Association, Pew Research Center, Department of Education, Census Bureau, and book industry organizations. Some attributes like "geographic location" serve multiple purposes - as both categorical groupings and spatial coordinates for mapping visualizations.

Chloe Manilay Screenshots:

1. Standardized test scores over time -> Standardized test scores for American elementary school students over time?


2. What types of books are banned per year?

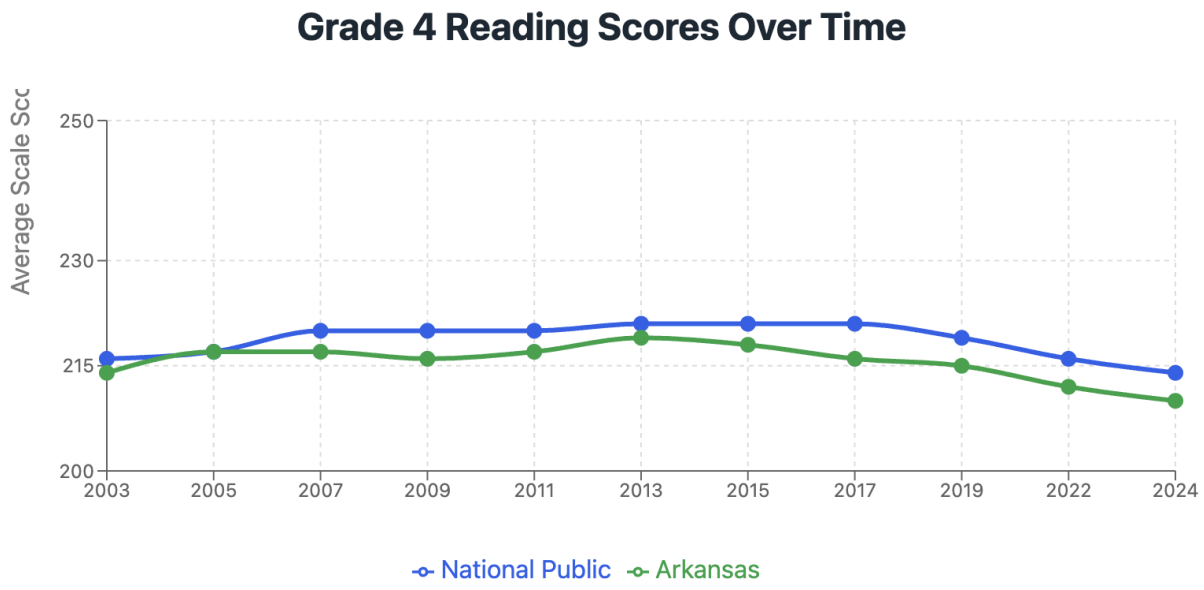
3. Funding into the American public library system over time?



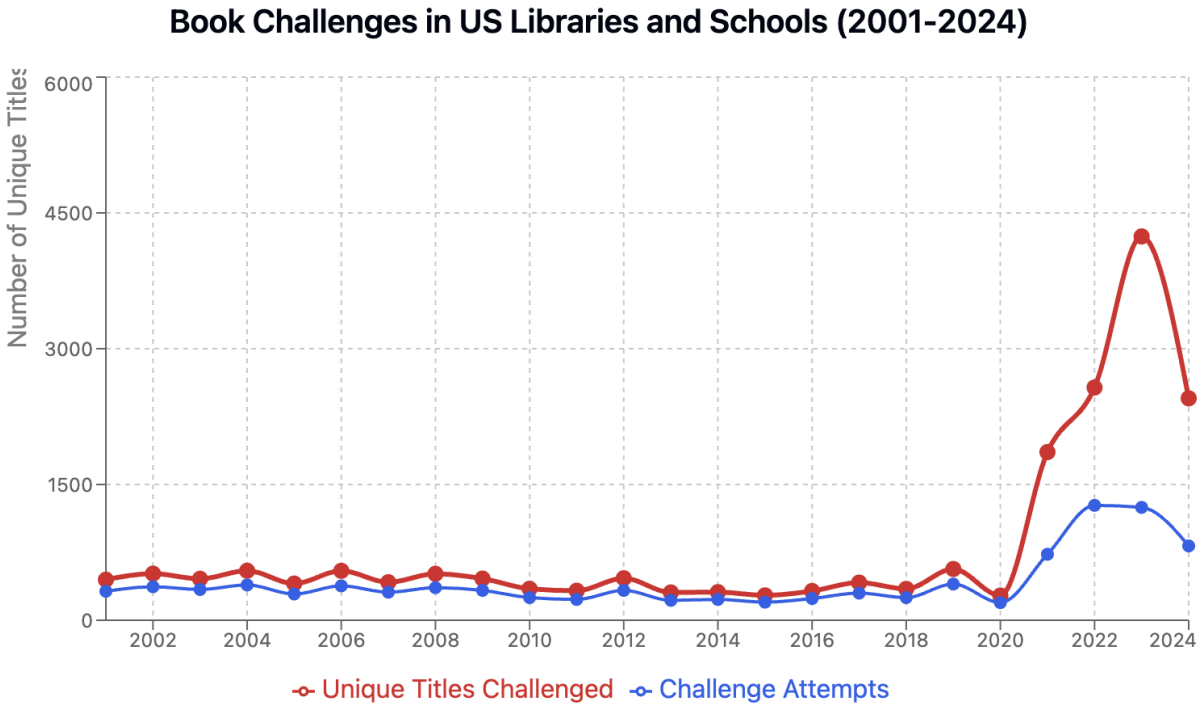

Chloe Manilay reflection: The questions I explored in the above visualizations were relatively straightforward and completely quantitative, offering insight into the most general/obvious trends but setting the stage for more insightful data exploration - which can branch off of the most basic questions (essentially what Claude did for me in the little buttons you can explore above the graph that break down the dataset in different ways). In my opinion, it is difficult/risky to draw honest and significant qualitative conclusions without a solid quantitative data foundation to back up those insights. By starting with the biggest and most easily answerable problems, we can delve into more unique ones (ex. Overall per capita spending patterns -> patterns of overall funding sources, book bans overall -> book bans per topic, and elementary test scores -> elementary test scores centered before/after covid)

Anchal Bhardwaj Screenshots:

1. How does the reading scores for fourth grade students in Arkansas differ from the National Public reading score from 2003 to 2024?

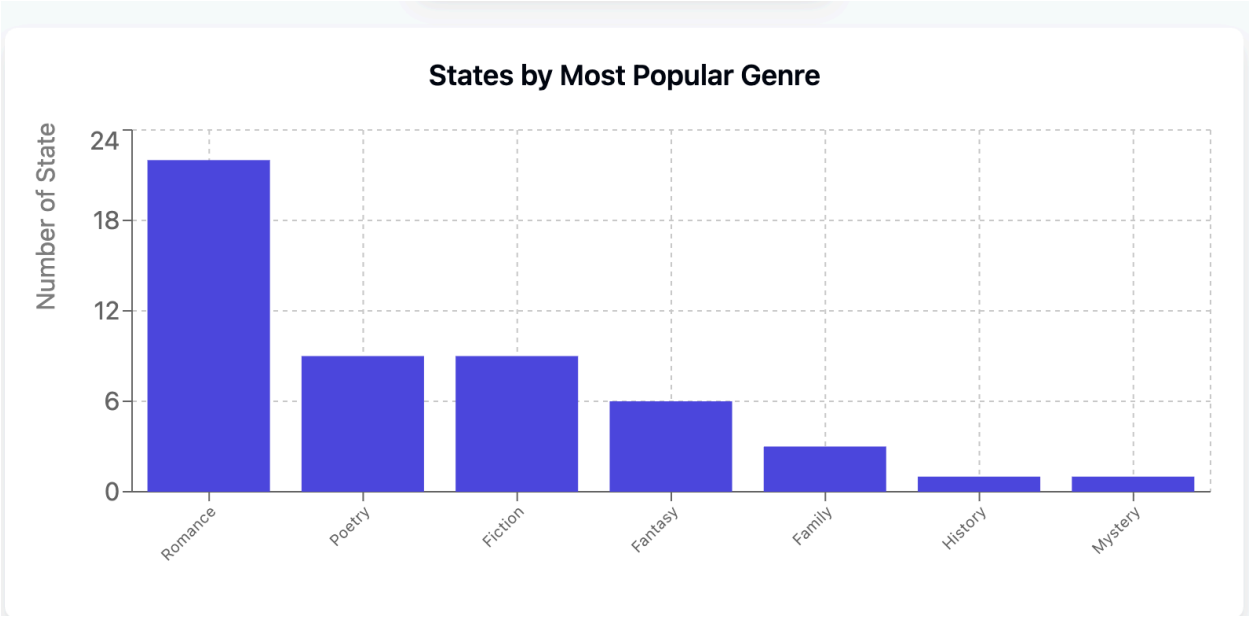


2. How many books are banned and/or challenged per year in US Libraries and Schools from 2001 to 2024?




**Note:** Data shows dramatic surge beginning in 2021. The 2023 peak represents the highest number ever documented by ALA.

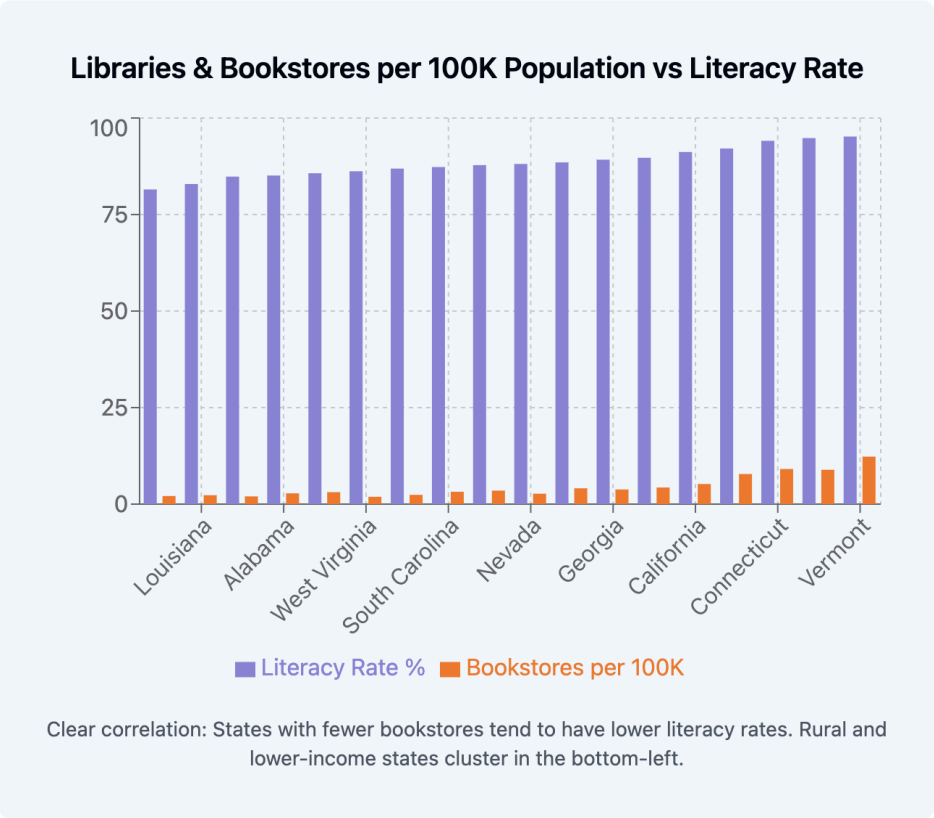
3. What genres are most popular in the US per state?

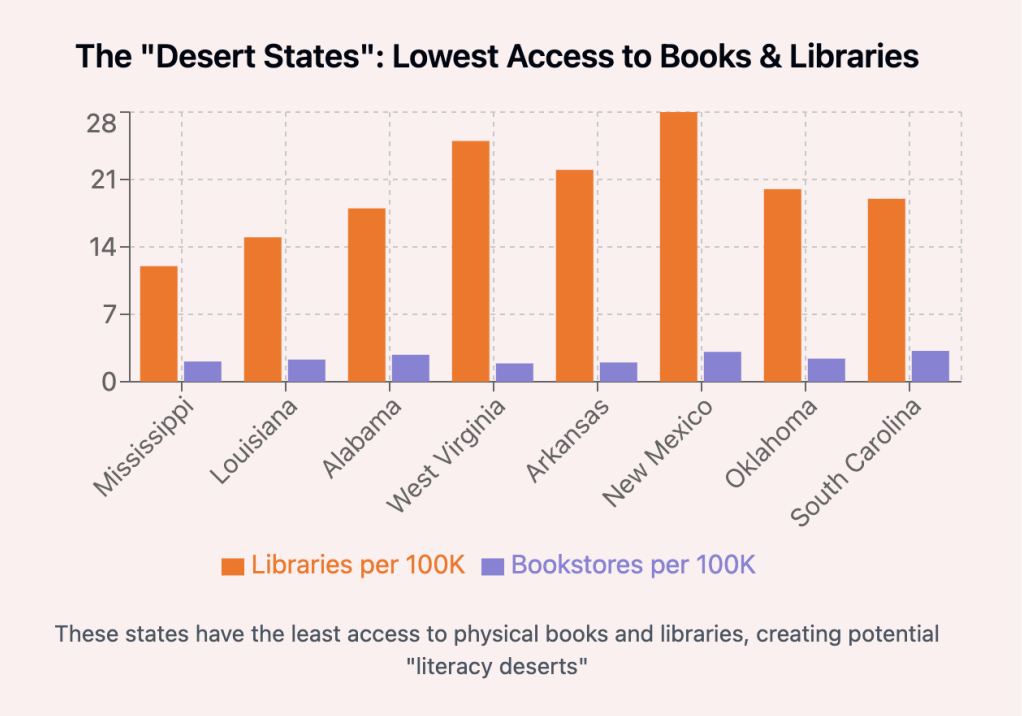


Anchal's Reflection: The questions addressed in my sketches are far more specific than the ones that we originally brainstormed - especially when it comes to the location, data being compared, and the timeframe. For example, rather than the questions being about books being banned generally over time, the question is about books in US libraries being challenged versus challenge attempts from a time period of 2001 to 2024. I find that these more specific questions are more helpful in making claims about the data because the claims are more specific as well. I ended up staying with the original question for the genre popularity per state because it didn't require additional specificity but in the future visualizations, I could include data on most popular genres per gender.

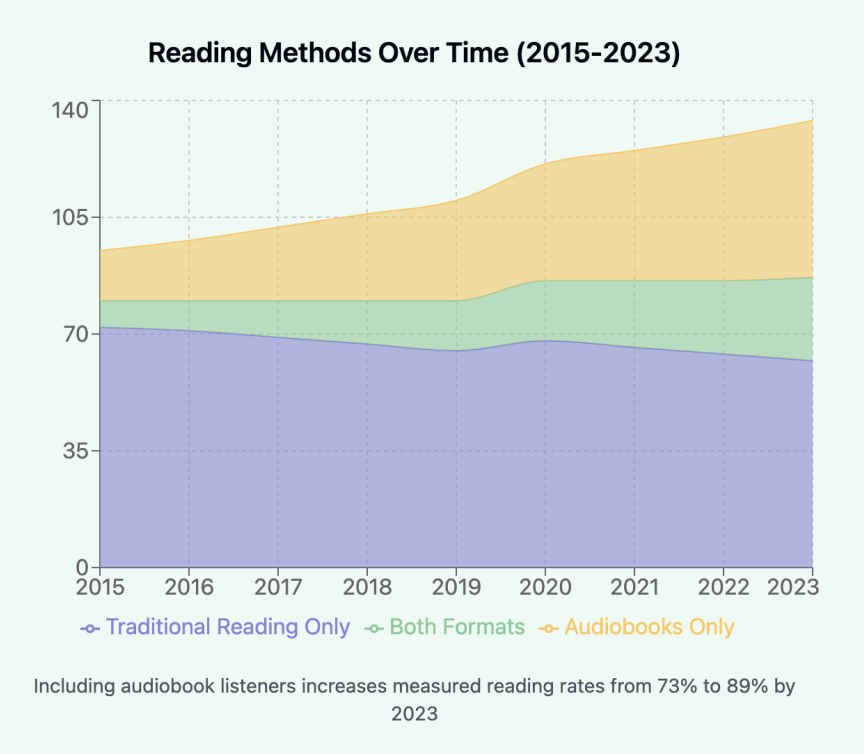
Somto Unini Screenshots:

1. Are there book and library deserts in America? → Examining access to books and libraries across different states and income levels



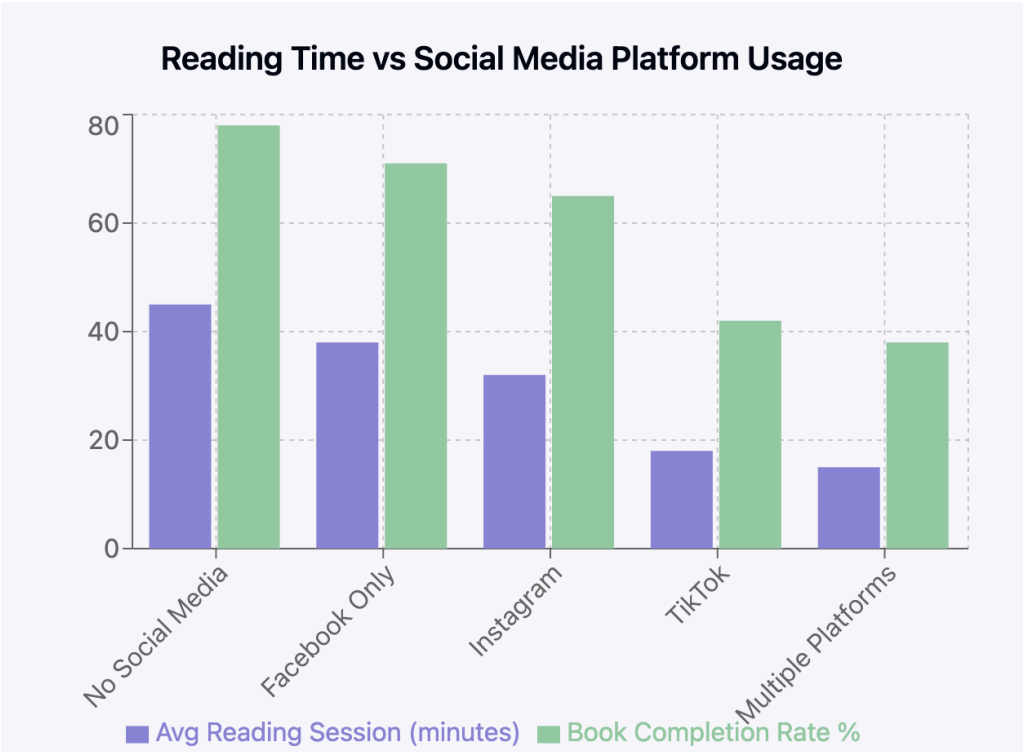



2. Do audiobooks count as "reading" in literacy statistics? → How including audiobook listeners changes our understanding of American reading habits

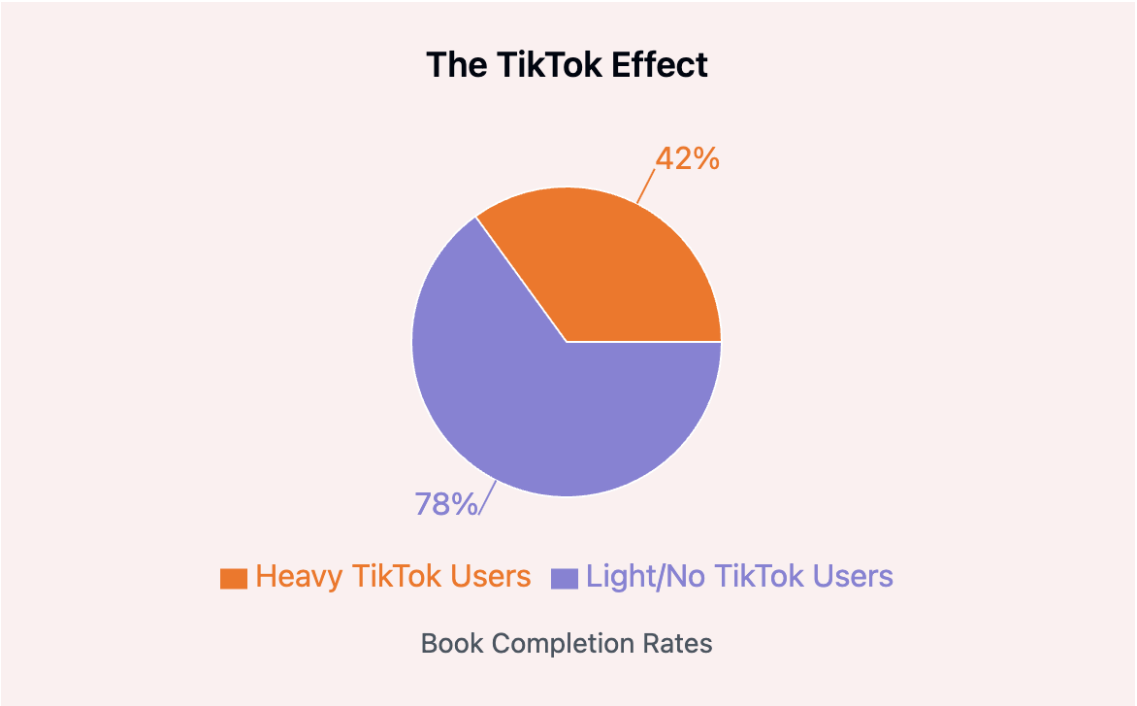




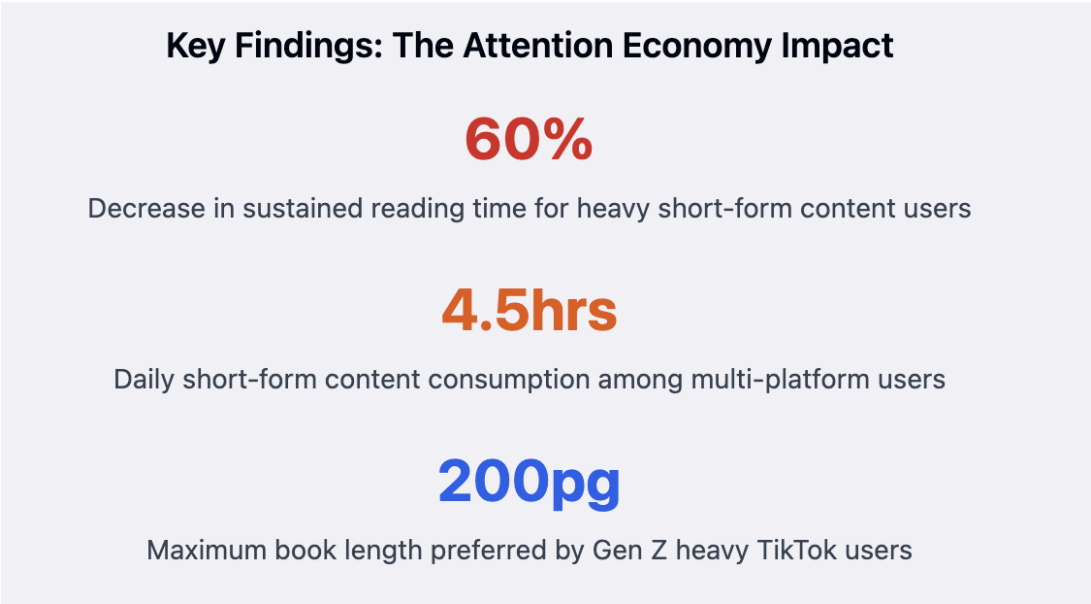
3. How does short-form content affect reading habits? → The relationship between TikTok/Instagram usage and sustained reading ability




Heavy TikTok users average just 18 minutes per reading session vs 45 minutes for non-users



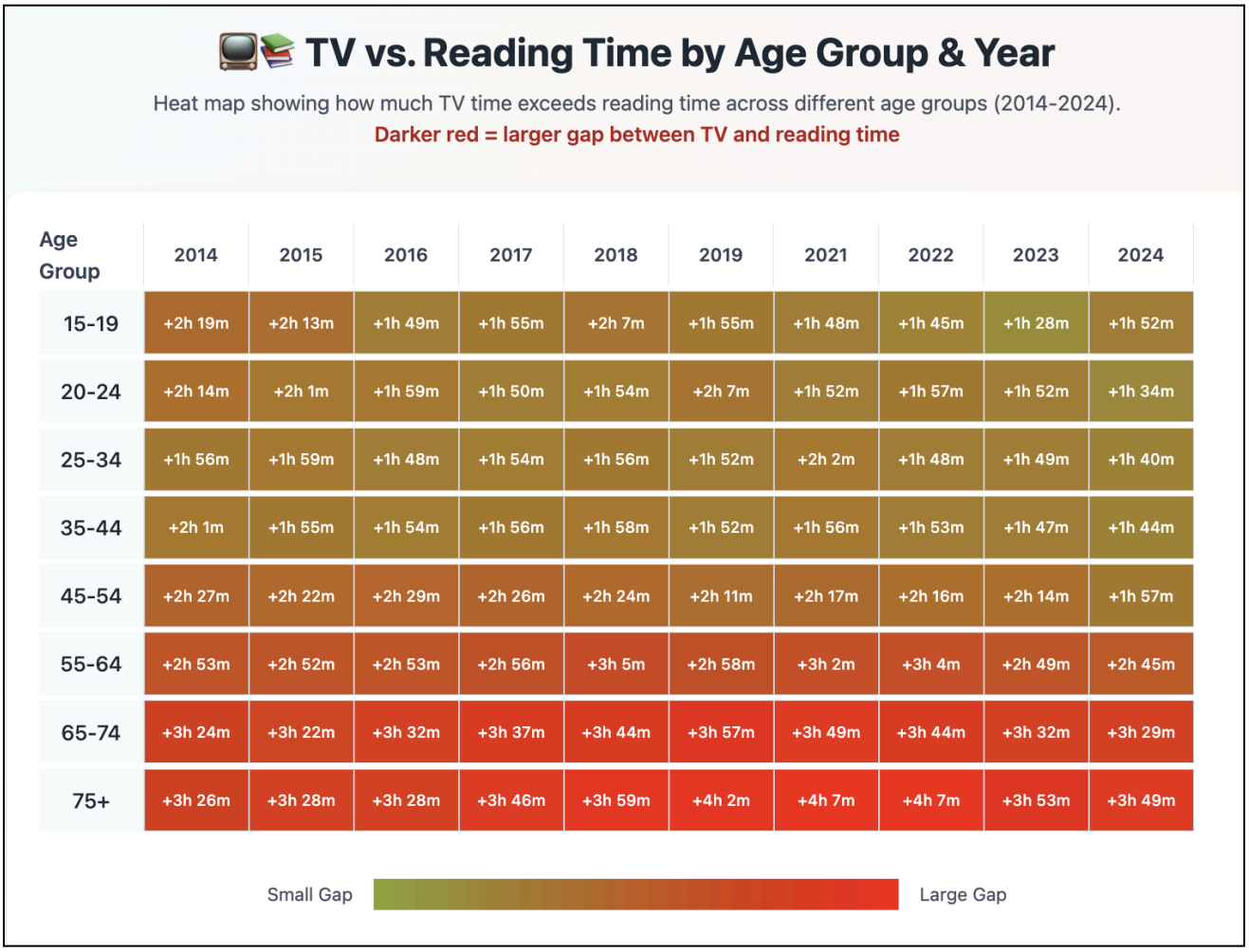

Book Completion Rates




Somto's Reflection: Compared to the team's original 16 questions (before we added many more), I chose to visualize questions that were more analytically complex and socially relevant. While questions like "How many books are people reading per year?" and "What genres are most popular?" are important for establishing baselines, I wanted to explore questions that challenge assumptions and uncover systemic issues. For instance, Question 6 (book/library deserts) highlights geographic inequality, Question 22 (audiobooks in literacy stats) shows how measurement methods shape perceptions of a "literacy crisis," and Question 26 (social media's impact on attention spans) addresses a pressing, lived concern. I see these as "better" not because the original questions lack value, but because they tell more compelling stories (revealing hidden patterns, questioning conventional wisdom, tying literacy to broader social issues like inequality and technology). I still think the strongest approach would combine both: using baseline questions for context and the more analytical ones to engage the audience with deeper insights.

## Oakley's Visualizations:

How Does Time Spent Watching TV Compare to Reading Across Age Groups?



### 📺📚 TV vs. Reading Time by Age Group & Year

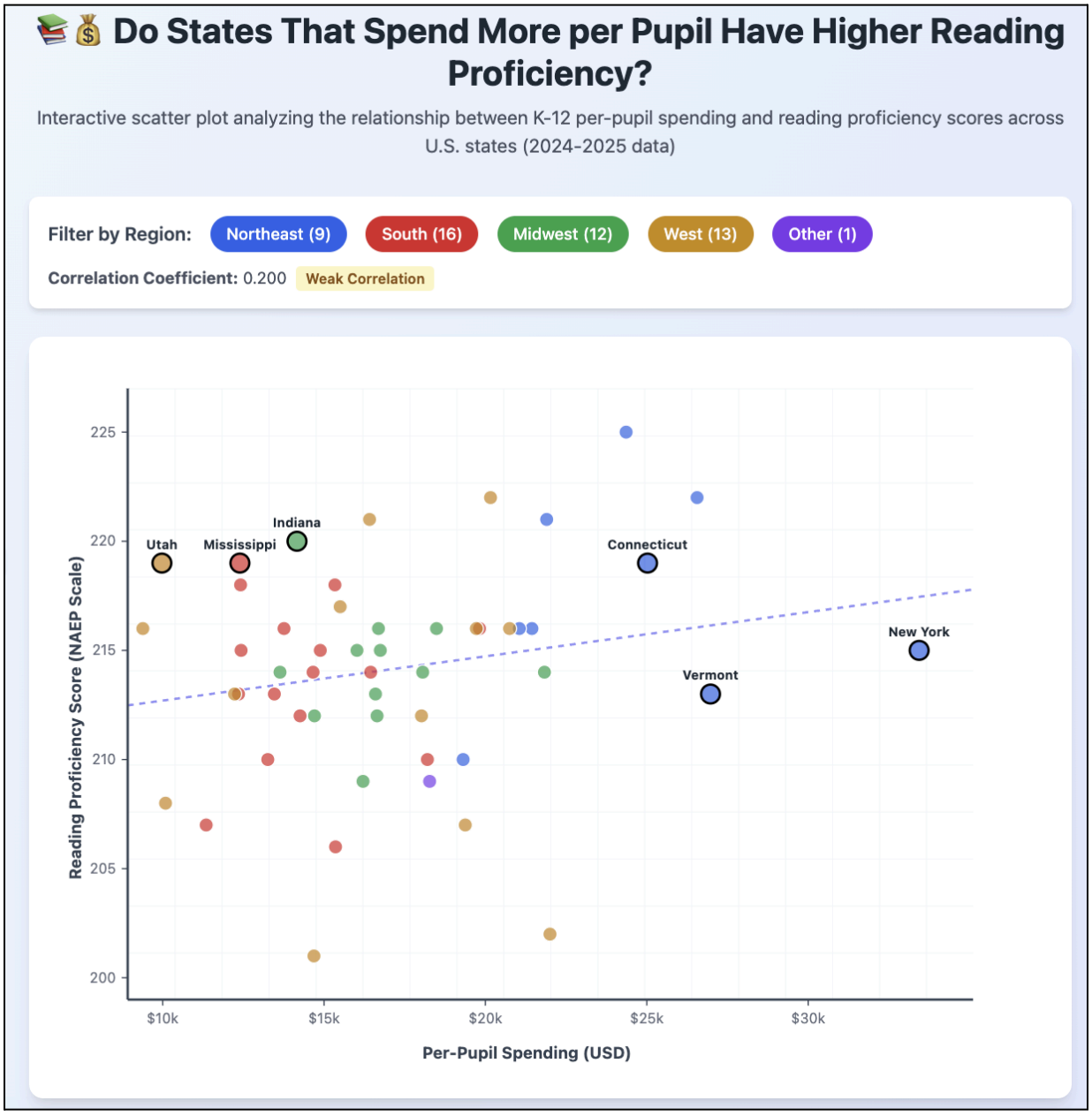
Heat map showing how much TV time exceeds reading time across different age groups (2014-2024).

**Darker red = larger gap between TV and reading time**

| Age Group | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2021 | 2022 | 2023 | 2024 |
|---|---|---|---|---|---|---|---|---|---|---|
| 15-19 | +2h 19m | +2h 13m | +1h 49m | +1h 55m | +2h 7m | +1h 55m | +1h 48m | +1h 45m | +1h 28m | +1h 52m |
| 20-24 | +2h 14m | +2h 1m | +1h 59m | +1h 50m | +1h 54m | +2h 7m | +1h 52m | +1h 57m | +1h 52m | +1h 34m |
| 25-34 | +1h 56m | +1h 59m | +1h 48m | +1h 54m | +1h 56m | +1h 52m | +2h 2m | +1h 48m | +1h 49m | +1h 40m |
| 35-44 | +2h 1m | +1h 55m | +1h 54m | +1h 56m | +1h 58m | +1h 52m | +1h 56m | +1h 53m | +1h 47m | +1h 44m |
| 45-54 | +2h 27m | +2h 22m | +2h 29m | +2h 26m | +2h 24m | +2h 11m | +2h 17m | +2h 16m | +2h 14m | +1h 57m |
| 55-64 | +2h 53m | +2h 52m | +2h 53m | +2h 56m | +3h 5m | +2h 58m | +3h 2m | +3h 4m | +2h 49m | +2h 45m |
| 65-74 | +3h 24m | +3h 22m | +3h 32m | +3h 37m | +3h 44m | +3h 57m | +3h 49m | +3h 44m | +3h 32m | +3h 29m |
| 75+ | +3h 26m | +3h 28m | +3h 28m | +3h 46m | +3h 59m | +4h 2m | +4h 7m | +4h 7m | +3h 53m | +3h 49m |

Small Gap — Large Gap

Do States That Spend More per Pupil Have Higher Reading Proficiency?



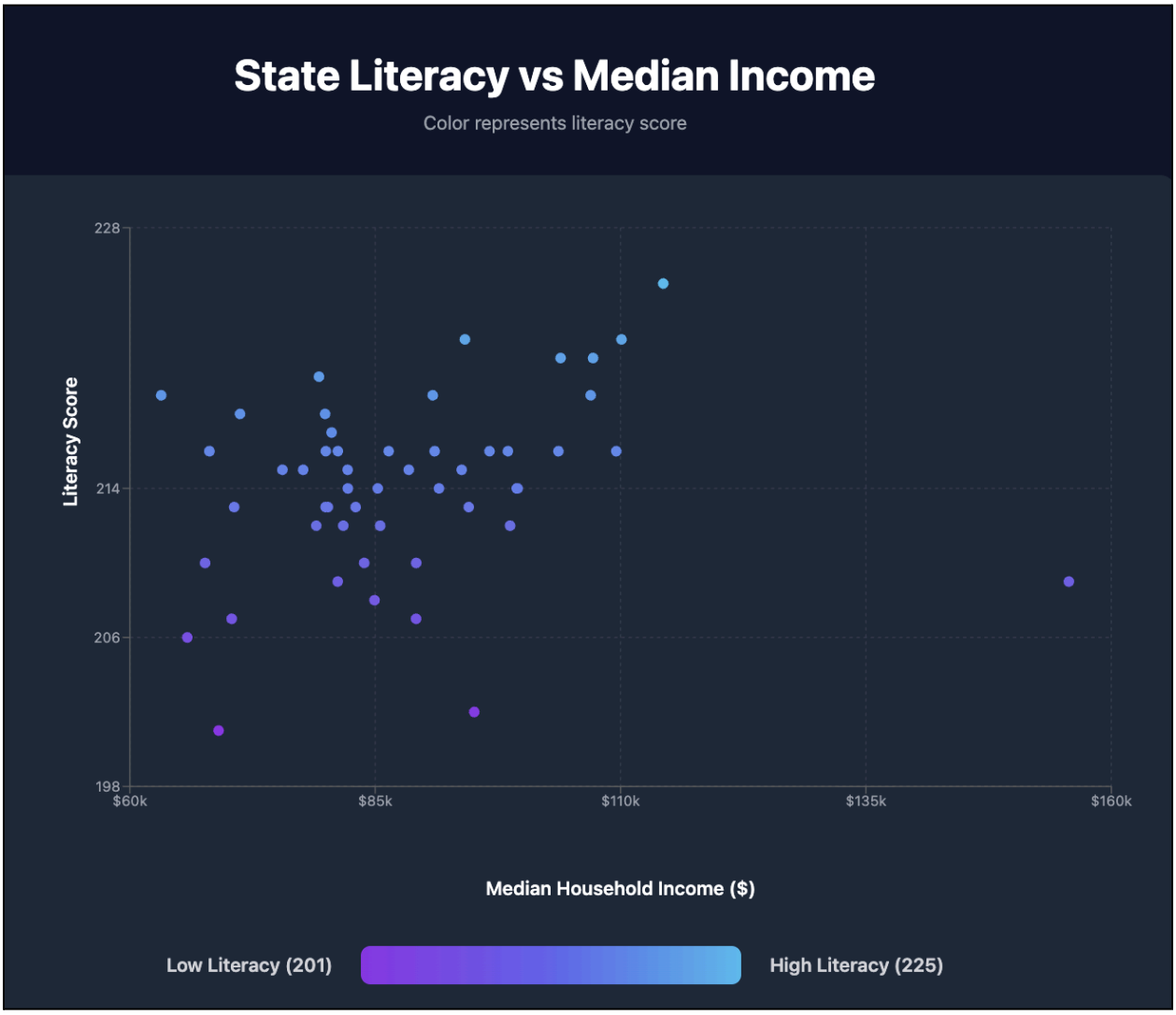
# 📚💰 Do States That Spend More per Pupil Have Higher Reading Proficiency?

Interactive scatter plot analyzing the relationship between K-12 per-pupil spending and reading proficiency scores across U.S. states (2024-2025 data)

**Filter by Region:** Northeast (9) | South (16) | Midwest (12) | West (13) | Other (1)

**Correlation Coefficient:** -0.077 Very Weak Correlation

Does Higher Household Income Translate to Higher Literacy?




Reflection: My work builds on our team's initial questions by extending them into correlational analysis. One partner examined the broad tradeoff between screen time and reading, and I expanded this by introducing age demographics, showing that the imbalance is sharpest among certain cohorts, particularly with television use. I also moved beyond descriptive trends such as book bans to investigate systemic factors, developing visualizations that connect school funding with literacy proficiency and state-level household income with literacy scores. My approach shifts the focus from basic observations to measurable relationships between demographics, resources, and outcomes. I personally prefer my questions because they seek to uncover relationships rather than simply describe patterns. Hopefully these findings will spark meaningful conversations within our team about what we are aiming to demonstrate with this project.

# Milestone 5: Data

Datasets Chosen

1. PEN America Index of School Book Bans (2021-2022, 2022-2023, 2023-2024, and 2024-2025)
    a. File format/data restrictions:
        i. File Type: CSV (Comma-Separated Values)
        ii. Total Records: 6,719 book ban instances
        iii. Total Fields: 10 columns
        iv. Data Restrictions:
            1. No exact duplicate rows found
            2. All records have complete core data (Title, Author, State, District, Date, Ban Status)
            3. Optional fields have high missing rates (expected for bibliographic data)
            4. Each row represents a single book ban/challenge instance
    b. Data inventory (list of items, variables, data types, value ranges, how it was collected, missing patterns:
        **i. Value Ranges**
        ii. States (22 locations):
            1. Arizona, Colorado, Florida, Georgia, Indiana, Iowa, Kansas, Maryland, Minnesota, Missouri, Nation (DoD schools), New Hampshire, North Carolina, Oregon, Pennsylvania, South Carolina, Tennessee, Texas, Utah, Virginia, Wisconsin, Wyoming
            2. Ban Status Categories (4 types):
                1. Banned (67.8%)
                2. Banned by Restriction (18.1%)
                3. Banned Pending Investigation (14.0%)
                4. Banned - Professionally Weeded (0.1%)

    Date Formats (18 variations): Various formats including "Month YYYY", "YYYY-YYYY", "Month YYYY - Month YYYY", and "24-25" style
    Data Collection Context: This dataset tracks school book challenges and removals from July 2024 through June 2025, documenting instances where books have been banned, restricted, or challenged in school libraries and curricula across the United States and Department of Defense Educational Activity schools.

| Column | Type | Completeness | Unique Values | Notes |
|---|---|---|---|---|
| **Title** | String | 100% (6,719/6,719) | 3,788 | Book titles |
| **Author** | String | 100% (6,719/6,719) | 2,307 | Primary author names |
| **Secondary Author(s)** | String | 7.4% (495/6,719) | 267 | Co-authors (92.6% missing) |
| **Illustrator(s)** | String | 6.2% (417/6,719) | 243 | Book illustrators (93.8% missing) |
| **Translator(s)** | String | 1.4% (94/6,719) | 38 | Translators (98.6% missing) |
| **Series** | String | 32.8% (2,205/6,719) | 727 | Book series names (67.2% missing) |
| **State** | String | 100% (6,719/6,719) | 22 | US states + "Nation" |
| **District** | String | 100% (6,719/6,719) | 85 | School districts |
| **Date of Challenge/Removal** | String | 100% (6,719/6,719) | 18 | Challenge/removal dates |
| **Ban Status** | String | 100% (6,719/6,719) | 4 | Type of ban |

c. Explore data (AI, Excel), with summaries/distribution/outlier detection:
Summary Statistics:

Total Ban Instances: 6,719
Unique Books Affected: 3,788
Authors Impacted: 2,307
States/Locations: 22
School Districts: 85
Time Period: July 2024 - June 2025
Data Quality: High (100% complete core fields, no duplicates)

| Title | |
|---|---|
| Clockwork Orange | 23 |
| Breathless (JN) | 20 |
| Sold | 20 |
| Last Night at the Telegraph Club | 19 |
| A Court of Mist and Fury | 18 |
| ... | ... |
| Dr. Xargle's Book of Earthlets | 1 |
| Boys I Know | 1 |
| Blood Lad, Vol. 1 | 1 |
| Batman: The Dark Knight Strikes Again | 1 |
| Cantarella, Vol. 1 | 1 |

3788 rows × 1 columns

| State | |
|---|---|
| Florida | 2304 |
| Texas | 1781 |
| Tennessee | 1622 |
| Nation | 590 |
| Iowa | 113 |
| Virginia | 97 |
| Pennsylvania | 73 |
| Georgia | 43 |
| Utah | 26 |
| Colorado | 19 |
| Minnesota | 16 |
| Wyoming | 8 |
| Maryland | 6 |
| South Carolina | 5 |

| Ban Status | |
|---|---|
| Banned | 4555 |
| Banned by Restriction | 1215 |
| Banned Pending Investigation | 939 |
| Banned - Professionally Weeded | 10 |

| District | |
|---|---|
| North East Independent School District | 752 |
| Hillsborough County Public Schools | 608 |
| Department of Defense Educational Activitity | 590 |
| Monroe County Schools | 564 |
| Katy Independent School District | 513 |
| ... | ... |
| Canyon Independent School District | 1 |
| Highland Park Independent School District | 1 |
| Lynchburg City Schools | 1 |
| Germantown Schools | 1 |
| Platteville School District | 1 |

| Author | |
|---|---|
| King, Stephen | 206 |
| Hopkins, Ellen | 167 |
| Maas, Sarah J. | 162 |
| Picoult, Jodi | 62 |
| Matsui, Yūsei | 54 |
| ... | ... |
| Carter, Pamela J. | 1 |
| McGraw Hill | 1 |
| Hamid, Mohsin | 1 |
| No Further Information | 1 |
| Watkins, Steve | 1 |

Summary Statistics for Numerical Columns:

| | Title | Author | Secondary Author(s) | Illustrator(s) | Translator(s) | Series | State | District | Date of Challenge/Removal | Ban Status |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 6719 | 6719 | 493 | 417 | 94 | 2205 | 6719 | 6719 | 6719 | 6719 |
| unique | 3788 | 2307 | 266 | 243 | 38 | 727 | 22 | 85 | 18 | 4 |
| top | Clockwork Orange | King, Stephen | Cast, Kristin | No Further Information Available | No Further Information Available | A Court of Thorns and Roses | Florida | North East Independent School District | June 2025 | Banned |
| freq | 23 | 206 | 47 | 21 | 21 | 75 | 2304 | 752 | 1603 | 4555 |

d. Document missing values/inconsistencies/duplicates/transformation needs:

i. Missing Values Assessment
    1. Expected Missing Data (Bibliographic): Overall no issues since a margin of error is expected
        a. Secondary Author(s): 92.6% missing - Normal for single-author books
        b. Illustrator(s): 93.8% missing - Normal for text-only books
        c. Translator(s): 98.6% missing - Normal for English-original works
        d. Series: 67.2% missing - Normal for standalone books
    2. Date Format Variations (18 different formats): Data standardization is critical
        a. Categorical Encoding needed for ban status
        b. Geographic Standardization: add state codes (2 letter)
        c. Author Name Normalization: maybe need to split into first name last name fields

e. Clean data: Implemented recommended data cleaning changes, downloadable cleaned data is here in csv format: https://drive.google.com/file/d/1lKqMgC3IwKWdJDd8ZXEhHH2YaOStCV5U/view?usp=sharing

## 2. NCES: U.S. PIAAC Skills Map: State and County Indicators of Adult Literacy and Numeracy

a. File format/data restrictions
    i. Data is an interactive map that is available in .xlsx format; creates separate sheets for county, state, and nation level info
    ii. Info is available for public use
b. Data inventory (list of items, variables, data types, value ranges, how it was collected, missing patterns:
    i. County: 34,562 items, 92 variables
    ii. State: 561 items, 77 variables
    iii. Nation: 11 items, 67 variables
    iv. State and county estimates are based on the combined PIAAC data collected in 2012, 2014, and 2017 and the data from the American Community Survey (ACS); estimates are modelled using small area estimation (SAE)
    v. Variable List

| Variable | Description | Data Type | Typical Range / Units |
|---|---|---|---|
| FIPS_code | Federal Information Processing Standard code | Identifier | 2–5 digit string |

| | | | |
|---|---|---|---|
| Country | Country name (only “United States”) | Categorical | Text |
| State | State name | Categorical | 50 states + D.C. |
| County | County name | Categorical | ~3,000 U.S. counties |
| grpName | Group name (domain grouping) | Categorical | Text |
| Lit_P1 | Literacy proportion ≤ Level 1 | Quantitative | 0–1 (proportion) |
| Lit_P1_CI_L, Lit_P1_CI_U | Lower/upper credible interval bounds for Lit_P1 | Quantitative | 0–1 |
| Lit_P1_CV | Coefficient of variation for Lit_P1 | Quantitative | 0–1 (usually 0.05–0.3) |
| Lit_P1_indicator | Out-of-bound indicator for Lit_P1 | Categorical | {1,2,3} |
| Lit_P1_CI_L_indicator, Lit_P1_CI_U_indicator, Lit_P1_CV_indicator | Indicators for CI and CV quality | Categorical | {1,2,3} |
| Lit_P2 | Literacy proportion = Level 2 | Quantitative | 0–1 (proportion) |
| Lit_P2_CI_L, Lit_P2_CI_U | Interval bounds for Lit_P2 | Quantitative | 0–1 |
| Lit_P2_CV | Coefficient of variation for Lit_P2 | Quantitative | 0–1 |
| Lit_P2_indicator, Lit_P2_CI_L_indicator, Lit_P2_CI_U_indicator, Lit_P2_CV_indicator | Data quality indicators | Categorical | {1,2,3} |
| Lit_P3 | Literacy proportion ≥ Level 3 | Quantitative | 0–1 (proportion) |
| Lit_P3_CI_L, Lit_P3_CI_U | Interval bounds for Lit_P3 | Quantitative | 0–1 |

| | | | |
|---|---|---|---|
| Lit_P3_CV | Coefficient of variation for Lit_P3 | Quantitative | 0–1 |
| Lit_P3_indicator, Lit_P3_CI_L_indicator, Lit_P3_CI_U_indicator, Lit_P3_CV_indicator | Data quality flags for Lit_P3 | Categorical | {1,2,3} |
| Lit_A | Literacy average score | Quantitative | 180–320 (PIAAC scale) |
| Lit_A_CI_L, Lit_A_CI_U | Interval bounds for Lit_A | Quantitative | 180–320 |
| Lit_A_CV | Coefficient of variation for Lit_A | Quantitative | 0–0.3 |
| Num_P1 | Numeracy proportion ≤ Level 1 | Quantitative | 0–1 |
| Num_P1_CI_L, Num_P1_CI_U, Num_P1_CV | Interval bounds / CV for Num_P1 | Quantitative | 0–1 |
| Num_P1_indicator, Num_P1_CI_L_indicator, Num_P1_CI_U_indicator, Num_P1_CV_indicator | Indicators for Num_P1 quality | Categorical | {1,2,3} |
| Num_P2 | Numeracy proportion = Level 2 | Quantitative | 0–1 |
| Num_P2_CI_L, Num_P2_CI_U, Num_P2_CV | Interval bounds / CV for Num_P2 | Quantitative | 0–1 |
| Num_P2_indicator, Num_P2_CI_L_indicator, Num_P2_CI_U_indicator, Num_P2_CV_indicator | Indicators for Num_P2 quality | Categorical | {1,2,3} |
| Num_P3 | Numeracy proportion ≥ Level 3 | Quantitative | 0–1 |

| | | | |
|---|---|---|---|
| Num_P3_CI_L, Num_P3_CI_U, Num_P3_CV | Interval bounds / CV for Num_P3 | Quantitative | 0–1 |
| Num_P3_indicator, Num_P3_CI_L_indicator, Num_P3_CI_U_indicator, Num_P3_CV_indicator | Indicators for Num_P3 quality | Categorical | {1,2,3} |
| Num_A | Numeracy average score | Quantitative | 170–310 |
| Num_A_CI_L, Num_A_CI_U, Num_A_CV | Interval bounds / CV for Num_A | Quantitative | 170–310 |
| POP | Total population | Quantitative | $10^3$–$10^7$ (count) |
| POP_DOMAIN | Population by group | Quantitative | $10^3$–$10^7$ (count) |
| Male, Female | Gender proportions | Quantitative | 0–1 |
| White, Black, Hispanic, Asian, AIAN, NHPI, Other_race | Race proportions | Quantitative | 0–1 |
| Less_HS, HS, More_HS | Education attainment proportions | Quantitative | 0–1 |
| Eng_not_well | English proficiency (not well / at all) | Quantitative | 0–0.15 |
| FB_after2010, FB_1990_2009, FB_before1990, FB | Foreign-born proportions | Quantitative | 0–1 |
| Poverty_100, Poverty_150 | Below 100% / 150% poverty level | Quantitative | 0–1 |
| SNAP | Households receiving SNAP | Quantitative | 0–0.35 |
| Employed, Unemployed, Not_in_labor | Employment status proportions | Quantitative | 0–1 |

| | | | |
|---|---|---|---|
| OCC_Manage, OCC_Service, OCC_Sales, OCC_Natural, OCC_Military, OCC_Prod | Occupational category proportions | Quantitative | 0–1 |
| No_Insurance | Proportion without health insurance | Quantitative | 0–0.25 |

vi. Missing Data Patterns

1. County
    a. Severe missingness: Nearly all literacy and numeracy indicator variables (e.g., Lit_P1_indicator, Num_P3_indicator, confidence intervals, CVs) are >99% missing.
    b. Demographic variables (POP, Male, Female, race breakdowns, education, employment, poverty, etc.) are ~91% missing.
    c. Only a few fields such as POP_DOMAIN have almost no missing values.
2. State
    a. Similar pattern, but less extreme: many literacy/numeracy indicators are 90–99% missing.
    b. Some demographic and education variables (Less_HS, More_HS) are ~36% missing, suggesting partial data availability for some states.
3. Nation
    a. Most variables (FB_1990_2009, Male, White, OCC_*, POP, etc.) are 90.9% missing, implying data only for one or a few rows (perhaps one per indicator or year).
    b. Education variables (Less_HS, More_HS) are ~36% missing.

c. Explore data (AI, Excel), with summaries/distribution/outlier detection:

i. Quick summary statistics

| Variable | Mean | Min | Max | Notes |
|---|---|---|---|---|
| Lit_P1 | 0.21 | 0.02 | 0.71 | Basic literacy proficiency rate |
| Lit_P2 | 0.33 | 0.09 | 0.52 | Intermediate proficiency |
| Lit_P3 | 0.37 | 0.14 | 0.61 | Advanced proficiency |
| Num_P1 | 0.19 | 0.02 | 0.64 | Basic numeracy |
| No_Insurance | 0.12 | 0 | 0.4 | Share of uninsured adults |

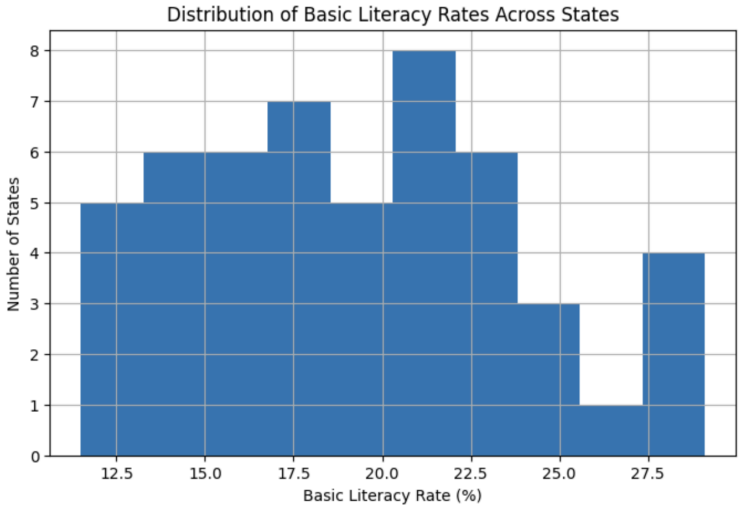
ii. Outliers snippet (states with extreme literacy proportions)


```
{'Lit_P1': ['District of Columbia',
  'Louisiana',
  'Mississippi',
  'Alabama',
  'Arizona',
  'Arkansas',
  'California',
  'Colorado',
  'Connecticut',
  'Delaware',
  'District of Columbia',
  'Florida',
  'Georgia',
  'Hawaii',
  'Illinois',
  'Kentucky',
  'Louisiana',
  'Maryland',
  'Massachusetts',
  'Mississippi',
  'Nevada',
  'New Jersey',
  'New Mexico',
  'New York',
  'North Carolina',
  'Rhode Island',
  'South Carolina',
  'Tennessee',
  'Texas',
  'Virginia',
  'West Virginia'],
 'Lit_P2': ['Mississippi',
```


iii. Post-cleaning table snippet

| | FIPS_code | State | County | Lit_Basic | Lit_Advanced | Lit_ProficientPlus | Lit_P1 | Lit_P2 | Lit_P3 |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 01001 | Alabama | Autauga County | 20.500202 | 41.099842 | 79.499793 | 20.500202 | 38.399951 | 41.099842 |
| **1** | 01003 | Alabama | Baldwin County | 16.100044 | 49.299945 | 83.999951 | 16.100044 | 34.700006 | 49.299945 |
| **2** | 01005 | Alabama | Barbour County | 39.400858 | 20.099844 | 60.499164 | 39.400858 | 40.399321 | 20.099844 |
| **3** | 01007 | Alabama | Bibb County | 26.900682 | 27.699743 | 73.099323 | 26.900682 | 45.399579 | 27.699743 |
| **4** | 01009 | Alabama | Blount County | 24.800234 | 38.199858 | 75.199761 | 24.800234 | 36.999903 | 38.199858 |

iv. Post-cleaning viz snippet



d. Document missing values/inconsistencies/duplicates/transformation needs:
    i.

| Issue Type | Example(s) | Action Needed |
|---|---|---|
| Missing values | Literacy/numeracy confidence interval "_indicator" columns missing 90–99% | Drop or ignore these columns (too sparse) |
| Partial coverage | Less_HS, More_HS missing ~36% | Consider imputing using mean or removing rows if small sample |
| Duplicates | None detected | No action needed |
| Inconsistent data | Literacy & numeracy columns likely proportions (0–1) | Verify consistency — no conversion needed |
| Transformation | Population and demographic columns only valid for 51 states | Filter down to those rows for visualization |

e. Clean data
    i. To prepare the dataset for analysis, I focused exclusively on literacy-related variables and removed all numeracy fields. I dropped columns with excessive missingness (particularly those ending in _indicator or _CV) while retaining key literacy measures (Lit_P1, Lit_P2, Lit_P3) and socioeconomic variables such as education level, poverty, and insurance coverage. Literacy values were converted from proportions to percentages for easier interpretation. I filtered the dataset to include only overall state-level observations (excluding subgroup breakdowns) and standardized geographic identifiers (FIPS and state names) for consistency. Missing socioeconomic data were imputed using median values, while counties with sparse literacy data were flagged rather than removed to preserve geographic completeness.

3. SFID: District Cost Database

a. District-level dataset of K-12 spending adequacy in roughly 12,000 public school districts in each year between 2009 and 2022. Downloadable in .xlsx form.
b. Total observations: 171,004, Unique districts: 12,263, Variables: 18. The data is published by the School Finance Indicators Database, which is compiled by the SFID and updated by the Albert Shanker Institute, University of Miami, and Rutgers University. Data originates from various sources within the US Census.

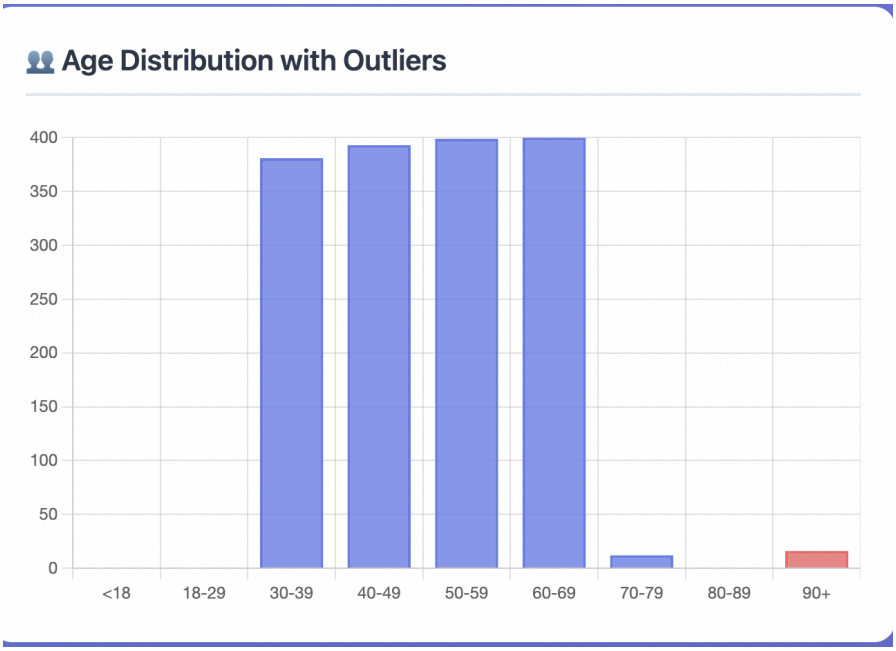
| | Variable | Data Type | Range / Example | % Missing |
|---|---|---|---|---|
| 0 | year | Temporal | 2009 to 2022 | 0.00 |
| 1 | district | Categorical | ALBERTVILLE CITY | 0.00 |
| 2 | state | Categorical / Geographic | AL | 0.00 |
| 3 | spending per pupil | Quantitative (ratio) | 251 to 83114 | 0.00 |
| 4 | required S/P | Quantitative (ratio) | 3802 to 98479 | 0.00 |
| 5 | funding gap | Quantitative (interval) | -78683 to 62435 | 0.00 |
| 6 | SDs from Avg US Test Scores | Quantitative (interval) | -3.7977975 to 1.6575414 | 33.39 |
| 7 | students enrolled | Quantitative (count) | 100 to 1014020 | 0.00 |
| 8 | poverty rate | Quantitative (proportion) | 0.0 to 1.0 | 0.00 |
| 9 | Special Ed | Quantitative (proportion) | 0.0 to 1.379532 | 3.43 |
| 10 | English learners | Quantitative (proportion) | 0.0 to 0.9567404 | 0.75 |
| 11 | amind | Quantitative (proportion) | 0.0 to 1.0 | 0.00 |
| 12 | asian | Quantitative (proportion) | 0.0 to 0.7474986 | 0.00 |
| 13 | black | Quantitative (proportion) | 0.0 to 1.0 | 0.00 |
| 14 | hisp | Quantitative (proportion) | 0.0 to 1.0 | 0.00 |
| 15 | multi | Quantitative (proportion) | 0.0 to 0.9902912 | 11.50 |
| 16 | pac | Quantitative (proportion) | 0.0 to 0.2674961 | 12.97 |
| 17 | white | Quantitative (proportion) | 0.0 to 1.0 | 0.00 |

c. I explored the dataset in Colab using pandas.describe(), .isna(), and boxplots. Spending per pupil ranged from roughly $250 to $83k, required spending from $3.8k to $98k, and funding gaps from −78k to +62k. The test-score variable ranged from about −3.8 to +1.7 standard deviations, showing expected variation in academic success. Poverty rates averaged 17%, with special education and English learner shares around 14% and 5%. Outliers appeared mainly in very large urban districts exceeding one million students.
d. Missing data was limited, mostly in test-score and demographic variables.
e. To clean my data, numeric values were converted to floats, blanks replaced with NaN, and proportions rounded to four decimals. After removing unnecessary fields (year, leaid, state_name), the dataset contained about 13,000 unique districts and 17 clean variables.

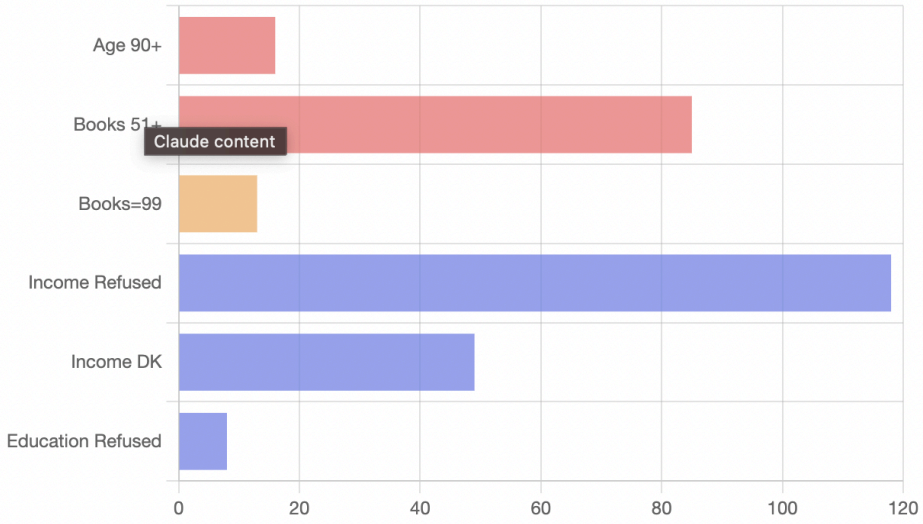
4. Pew Research Center Data on Reading Habits/Reasons for Reading/Mediums
    a. File format/data restrictions:
        i. Data formats available include SPSS, CSV, Free to use for research, analysis, and publication, files are open to public use
    b. Data inventory (list of items, variables, data types, value ranges, how it was collected, missing patterns:
        i. Demographics
            1. age or AGE: Age in years or categories
            2. sex or GENDER: Gender (1=Male, 2=Female)
            3. educ or EDUCATION: Education level (categorical)
            4. income or INCOME: Household income brackets
            5. race or RACETHN: Race/ethnicity
            6. hisp: Hispanic ethnicity
            7. region: Geographic region
            8. metro: Metropolitan status (urban/suburban/rural)
        ii. Reading Habits Variables
            1. book*: Book reading frequency/format
            2. ebook*: E-book reading
            3. print*: Print book reading
            4. audio*: Audiobook listening
            5. read*: General reading frequency
        iii. Reading Reasons/Purposes
            1. Reading for pleasure
            2. Reading for work/school
            3. Reading for research
            4. Reading to keep up with current events
        iv. Library Usage
            1. Library visits
            2. Library card ownership
            3. Borrowing patterns
            4. E-book borrowing from libraries
        v. Technology/Device Ownership
            1. Smartphone ownership
            2. E-reader ownership
            3. Tablet ownership
            4. Computer access
        vi. Weight Variables
            1. weight or WEIGHT: Survey weight for analysis
            2. May have multiple weight variables for different subsamples
    c. Explore data (AI, Excel), with summaries/distribution/outlier detection: I explored the csv dataset on Libraries in 2016. Initially I tried python with cursor but it didn't work so I used Claude here.
        i. Key Statistics

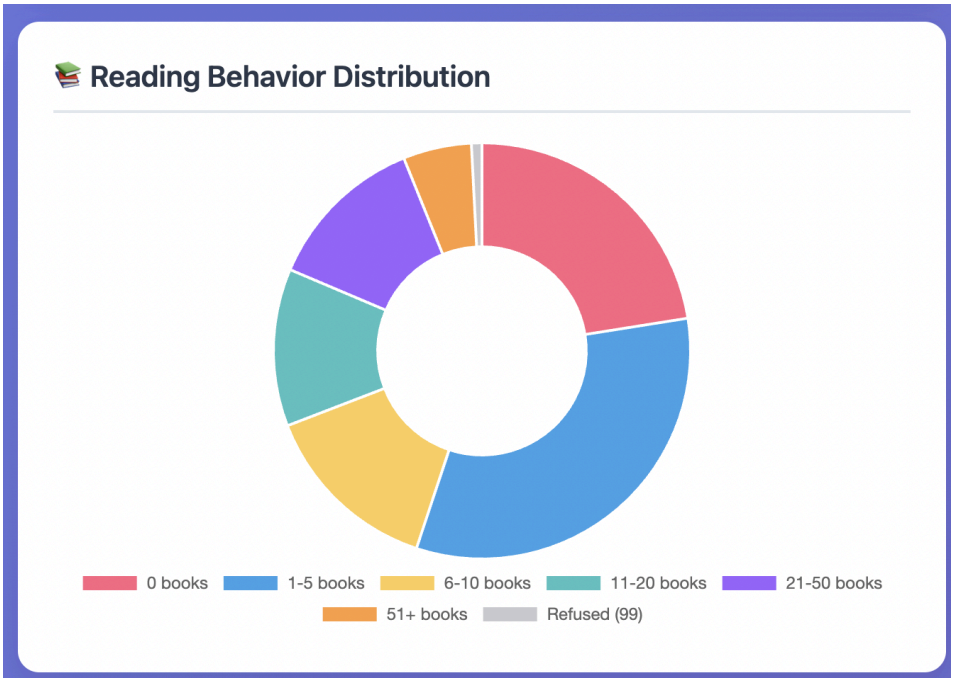
1. Total Records: 1,601
2. Mean Age: ~50 years
3. Median Books Read: 5 per year
4. Library Users: 7%
5. E-book Awareness: 47%
6. Smartphone Adoption: 65%

ii.




## ⚠️ Outlier Summary Across Variables







d. Document missing values/inconsistencies/duplicates/transformation needs:
    i. Missing Values: 47% average missing (mostly due to skip logic)
        1. 70-90% missing in follow-up questions (q11a-d, race3m2-4)
        2. 12% refusal rate for income
    ii. No duplicates found, but inconsistencies include;
        1. Mixed response scales (libvisit uses 1/3 instead of 1/2)
        2. String/numeric type mismatches
        3. Redundant weight variables
        4. Inconsistent special code usage

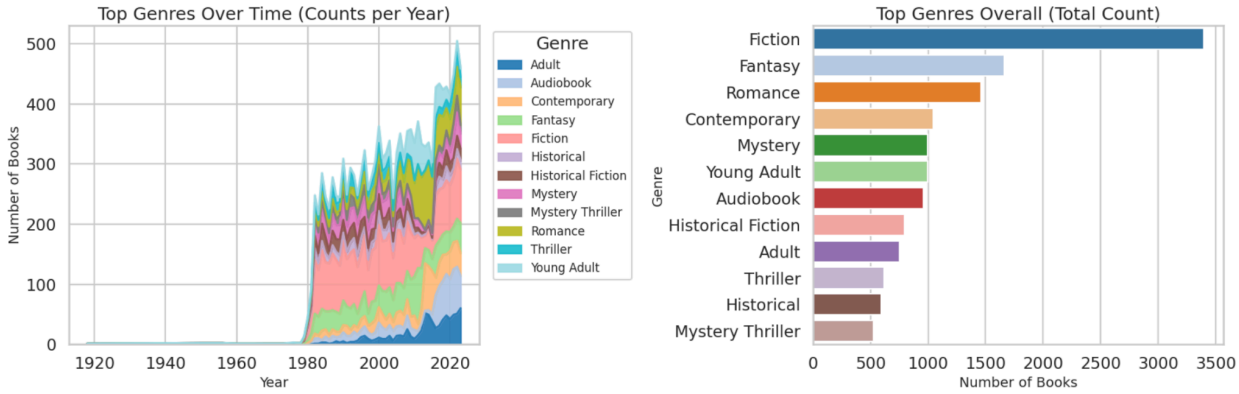
e. Clean data: I didn't remove any variables because I found them all to be potentially relevant, especially when used in conjunction with other variables in the same dataset (ex. Access to computers/demographics, amt of library card holders/region, etc).

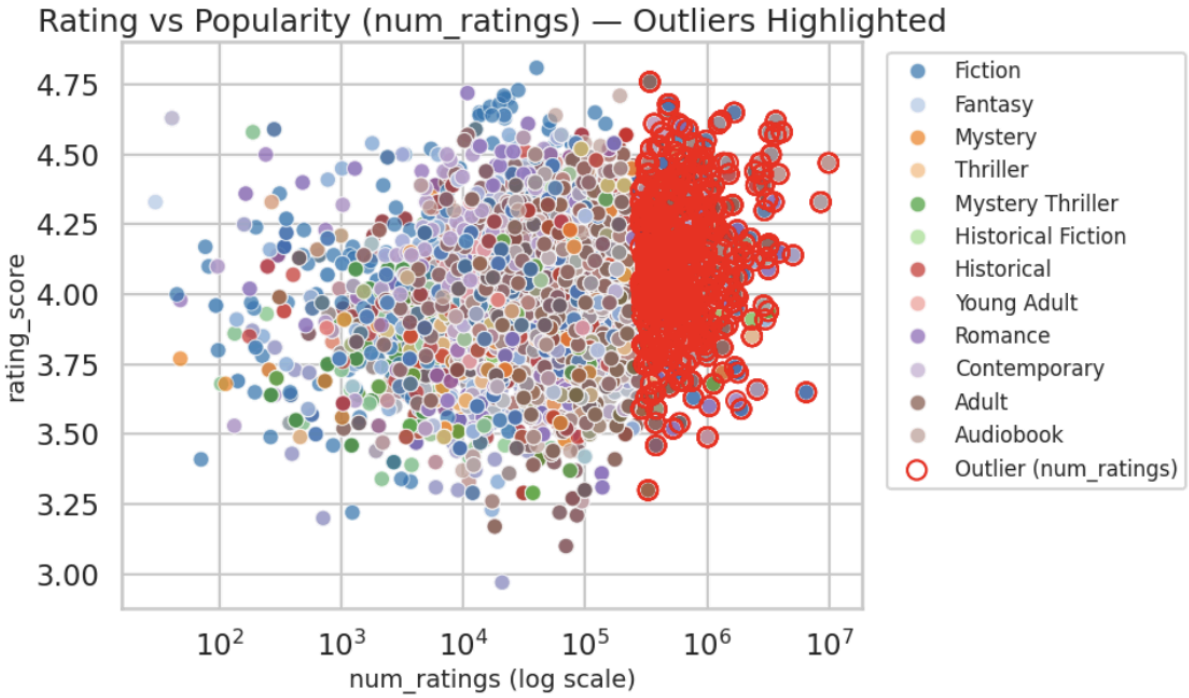
## 5. Kaggle: Top Goodreads Books Collection (1980-2023)

a. File format/data restrictions:
   i. Download in a CSV file
   ii. Open source - no usage restrictions - collected by scraping good reads collections.

b. Data inventory (list of items, variables, data types, value ranges, how it was collected, missing patterns:
   i. ISBN (quantitative): ISBN codes of the books.
   ii. Title: The title of the book, serving as the primary identifier for each literary work.
   iii. Series (categorical): Indicates whether the book is part of a series, with details on the series name.
   iv. Release Number (ordinal): Specifies the position of the book within a series, offering insights into its chronological order.
   v. Publisher (categorical): Publishers of the books.
   vi. Language (categorical): Language in which the books are written.
   vii. Author (categorical): The name of the author or authors associated with the book.
   viii. Description: A textual overview of the book's plot, providing a glimpse into its narrative.
   ix. Num Pages (quantitative): The total number of pages in the book, offering insights into its length.
   x. Format (categorical): Specifies the physical or digital format of the book, such as paperback, hardcover, or e-book.
   xi. Genres (categorical): the literary genres associated with the book, offering insights into its thematic categorization.
   xii. Publication Date (ordinal): The date when the book was first published, providing historical context to its release.
   xiii. Rating (ordinal): The average rating of the book on Goodreads, reflecting reader assessments.
   xiv. Number of Voters (quantitative): The count of votes or ratings received by the book on Goodreads, indicating reader engagement.
   xv. Current Readers (quantitative): The current number of readers of the book.
   xvi. Want to Read (quantitative): The number of people interested in reading the book.
   xvii. Price: The price of the book.
   xviii. URL: The URL of the book.

c. Explore data (AI, Excel), with summaries/distribution/outlier detection:




d. Document missing values/inconsistencies/duplicates/transformation needs:
    i. 14% of data missing for isbn, 48% missing series_title - optional value, 49% missing series_release_number, 3% missing publisher, 1% missing language, 2% missing description, 2% missing num_pages.
    ii. Goodreads data is global - there is no restriction on the user origin country - can't determine favorite genres by gender or age.
e. Clean data
    i. Steps taken - got rid of series title, release number, language (all was already English), publisher.
    ii. Split genre to see what "top" genre was for each book - put into separate column called top_genre.

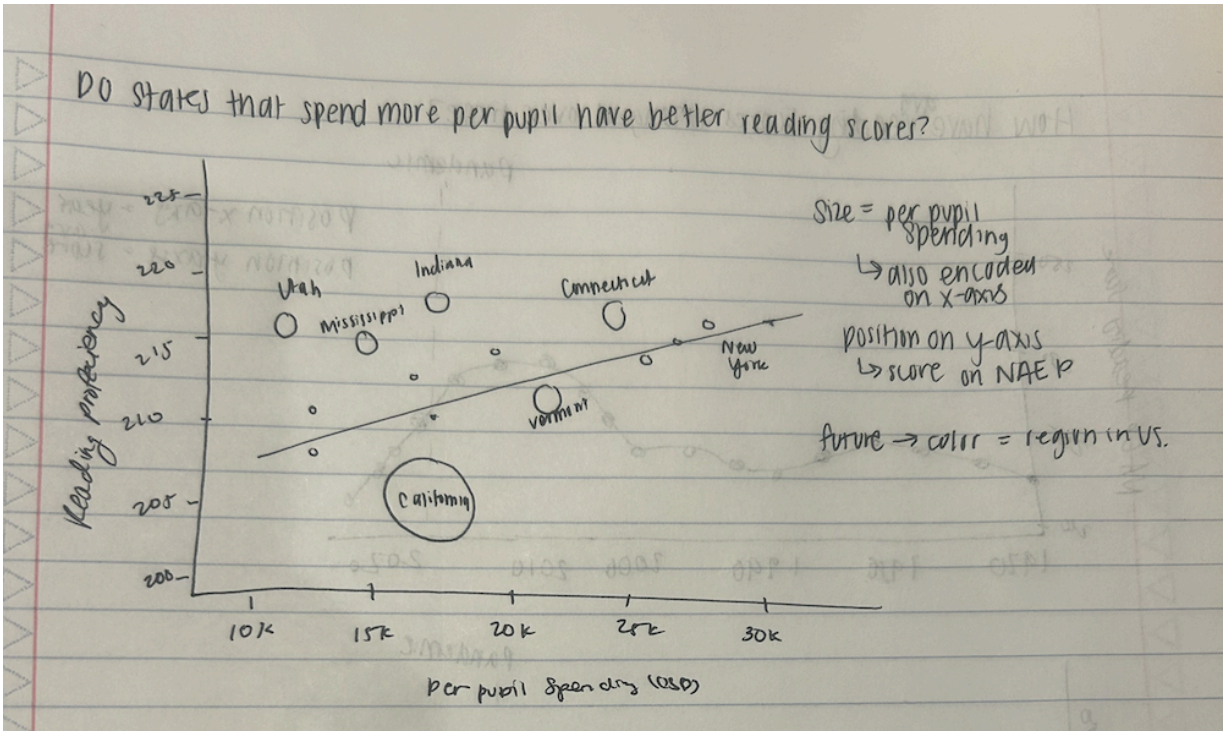
Reflection: For our project about analyzing literacy and reading culture in America, we chose these datasets to show different avenues of the issue. The PEN America Index of Book Bans is a core data set that shows over 6,000 ban instances across 22 states and shows insights about

what books are banned and why. This can be visualized to show the increase of bans over time and what kinds of books are being censored over time. Pairing this with the NCES PIAAC Skills Map will help us understand adult literacy and numeracy levels over states and counties - we can use this to see whether socioeconomic status or demographics impacts the literacy rates across counties.

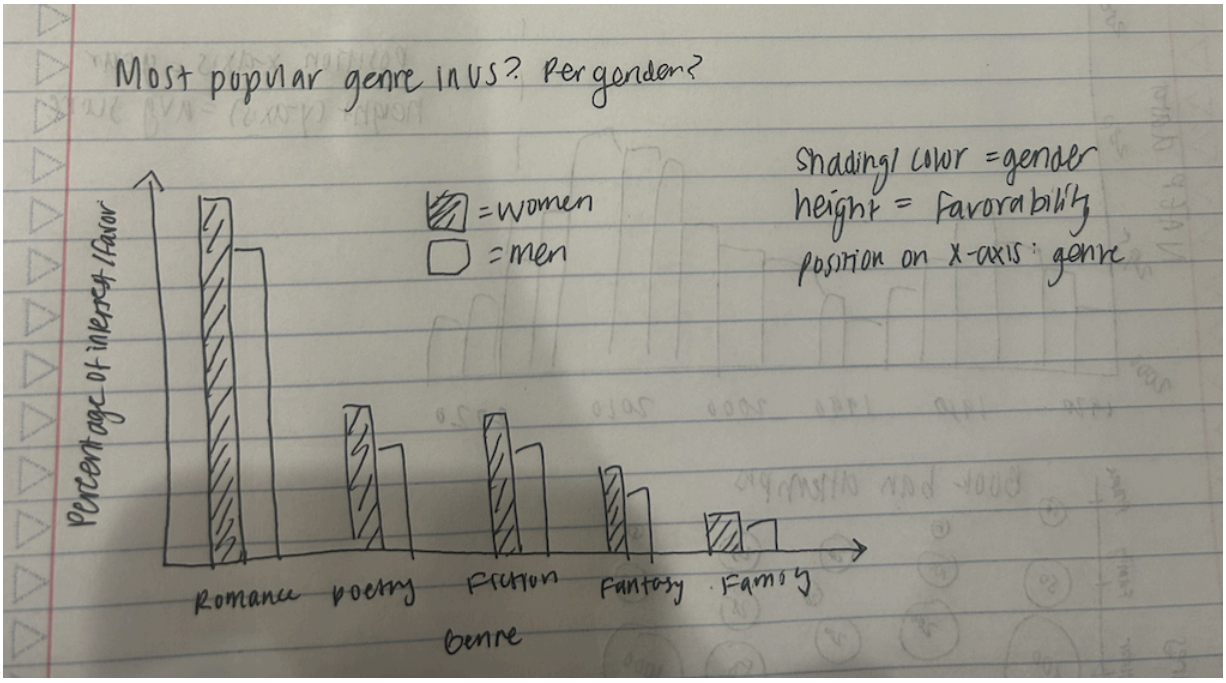
The SFID District Cost Database gives us a visualization on the spending across all 12,000 districts to understand the funding disparities around literacy. Lastly, we have the Pew Research Center survey on behaviors and library usage as well as the Kaggle collection of Goodreads reviews from 1980 to 2023. This shows us how reading culture has evolved over time: which genres are popular amongst which age groups, time periods, and genders. We will have to deal with the temporal misalignment of our data (some are worldwide, some only focus on the late 2010s while some are from the 1980s). We will also have to make sure that we don't align correlation and causation while we make our visuals, especially when overlapping socioeconomic data and literacy rates.

# Milestone 6: Sketch

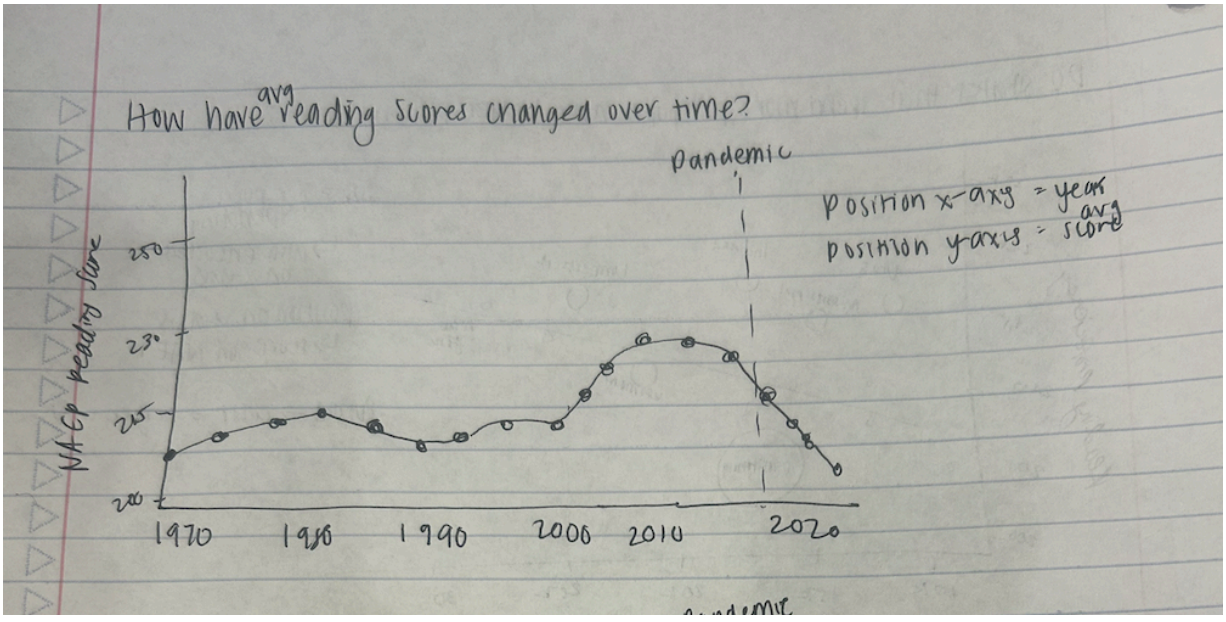
## Anchal's Sketches:



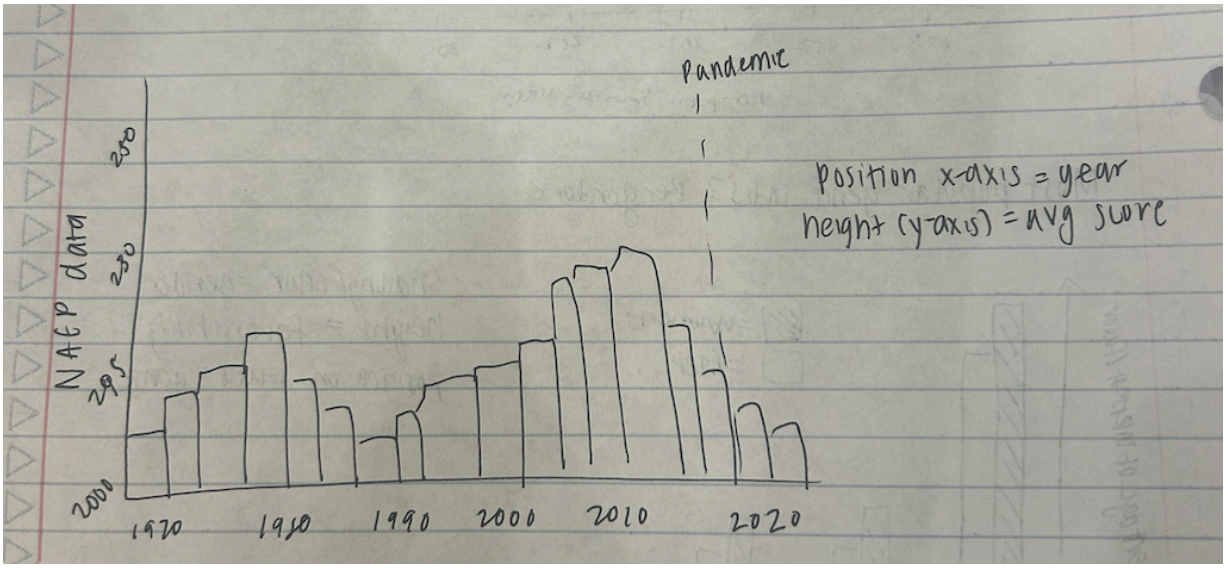

Do states that spend more per pupil have better reading scores?
Sketch ID: 1
Question ID:



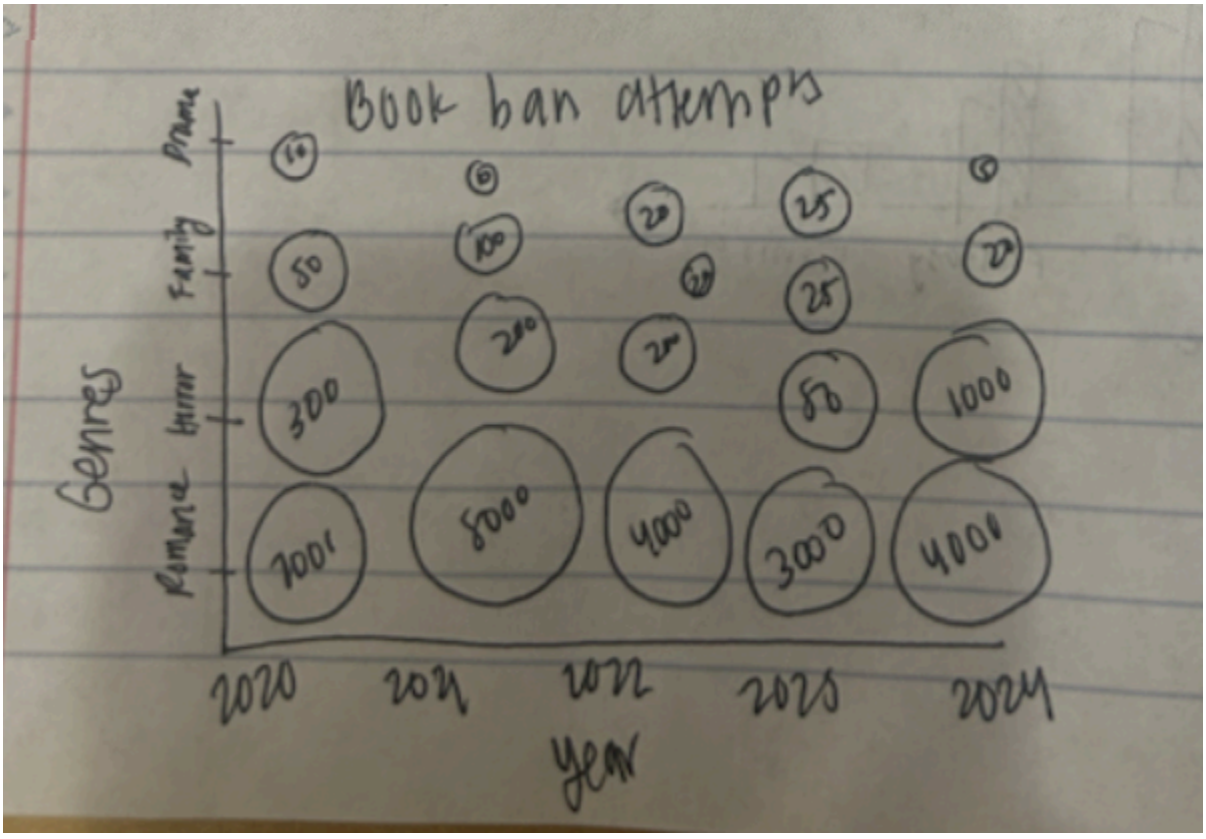

What are the most popular genres in the US per gender?
Sketch ID: 2
Question ID:




How have average NAEP reading scores changed over time in the US?
Sketch ID: 3
Question ID:




How have average NAEP reading scores changed over time in the US?
Sketch ID: 4
Question ID: 3



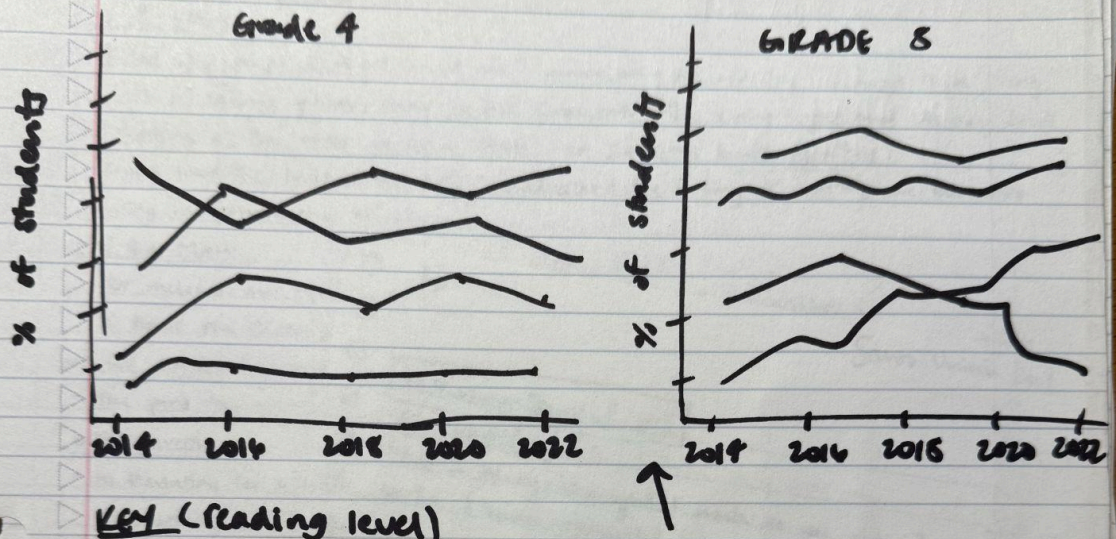

How many and what types of books are attempted to be banned from 2020 to 2024?
Sketch ID: 5

Question ID:

## Somto’s Sketches:

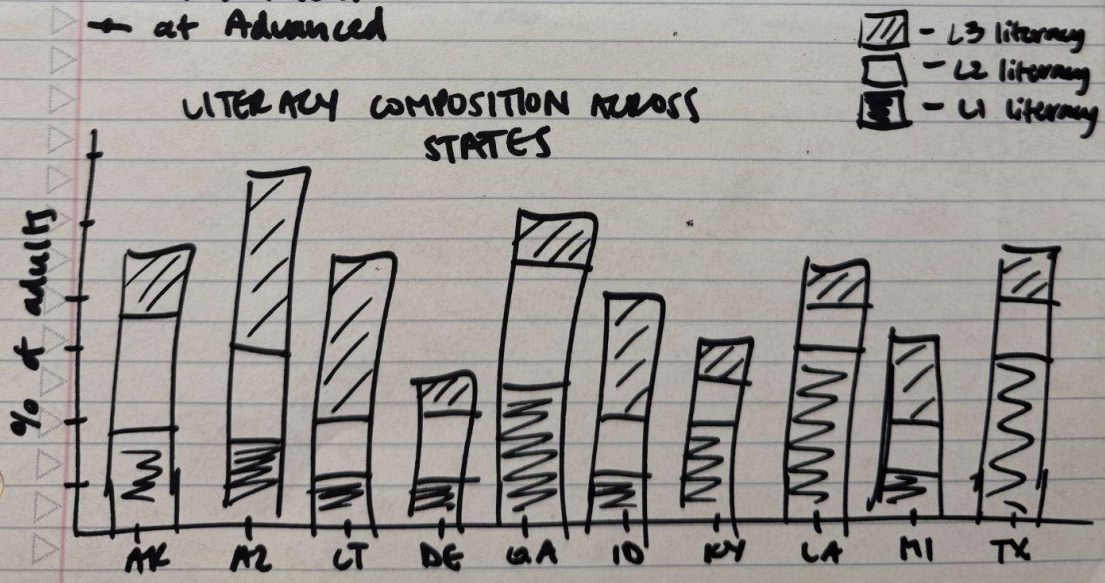
Are we more literate compared to a decade ago?
(Top) Sketch ID: 6
(Bottom) Sketch ID: 7
Question ID:

Q: ARE WE MORE LITERATE COMPARED TO A DECADE AGO?



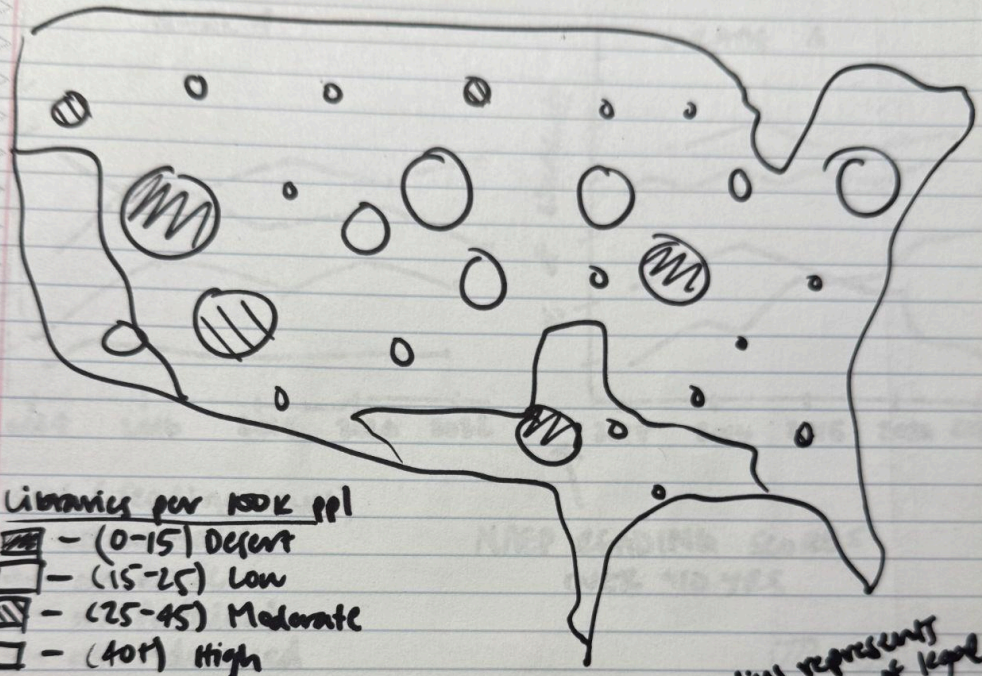

NAEP READING SCORES OVER ~10 YRS

KEY (reading level)
- below Basic
- at Basic
- at Proficient
- at Advanced

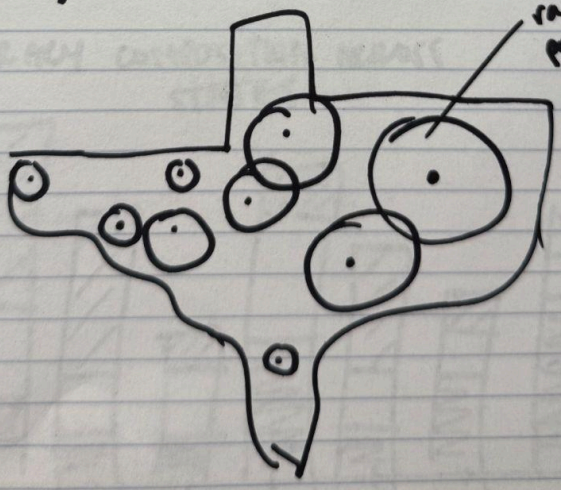


Are there deserts where there are no bookstores/libraries available?
Sketch ID: 8
Question ID:

Q: ARE THERE DESERTS WHERE THERE ARE NO LIBRARIES / BOOKSTORES AVAILABLE?

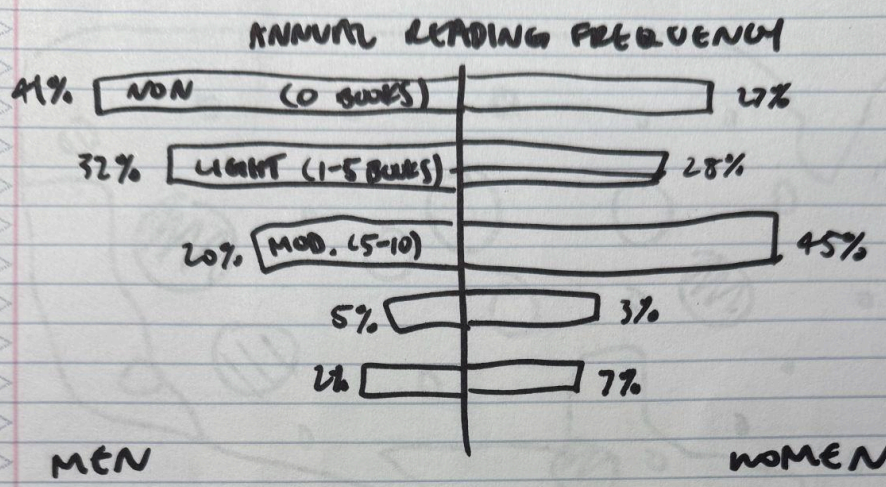


Libraries per 100K ppl
- (0-15) Desert
- (15-25) Low
- (25-45) Moderate
- (40+) High

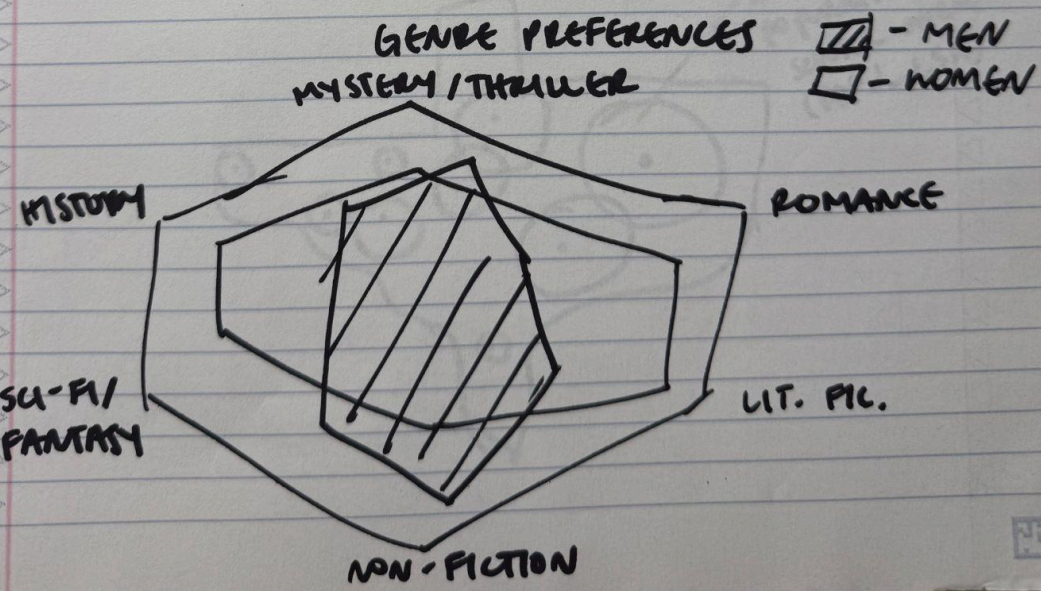


Do men and women have different reading patterns and preferences?
(Top) Sketch ID: 9
(Bottom) Sketch ID: 10
Question ID:

Q: DO MEN & WOMEN HAVE DIFFERENT READING PATTERNS & PREFERENCES?

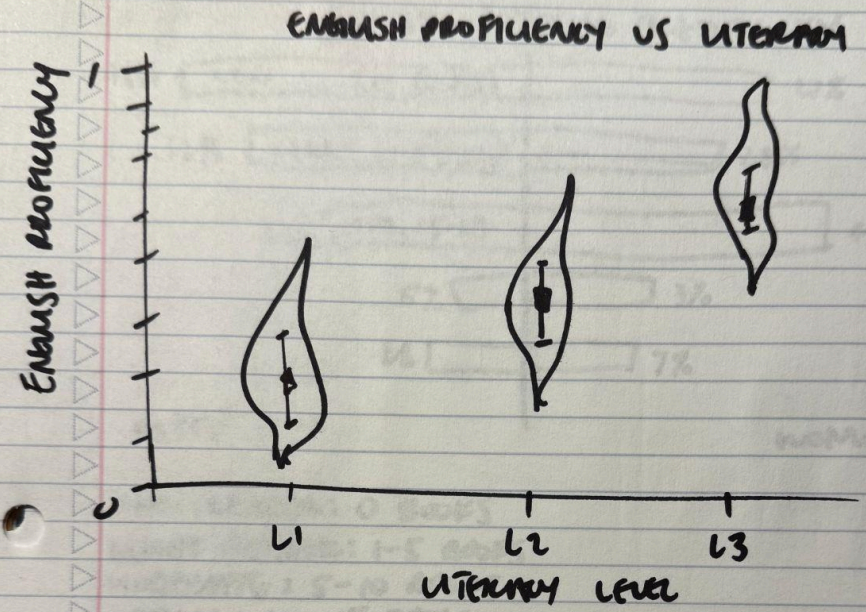


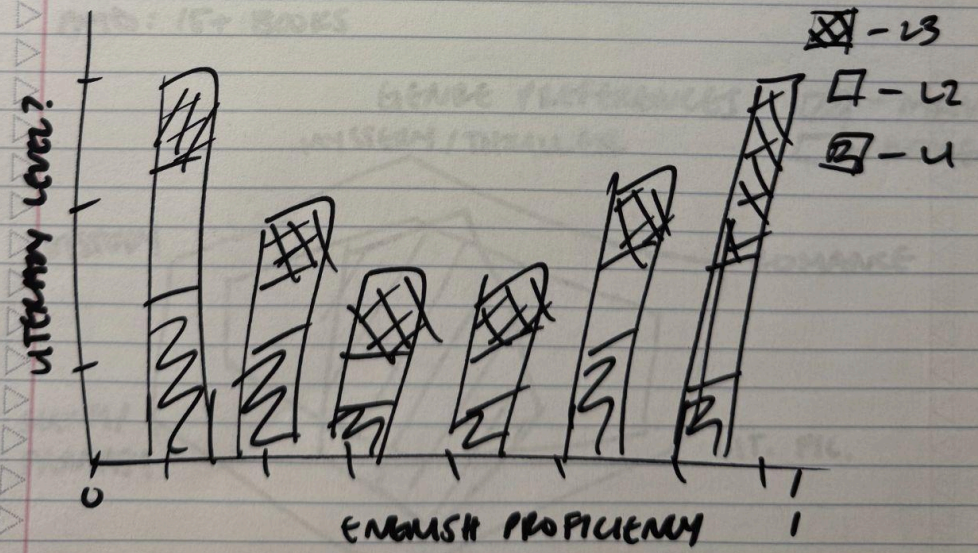

NON-READER: 0 BOOKS
LIGHT READER: 1-5 BOOKS
MODERATE: 5-10 BOOKS
HEAVY: 10-15 BOOKS
AVID: 15+ BOOKS



Do communities with higher rates of English as a second language (ESL) show different literacy patterns?
(Top) Sketch ID: 11
(Bottom) Sketch ID: 12
Question ID:

Q: DO COMMUNITIES w/ HIGHER RATES OF ESL SHOW DIFFERENT LITERACY PATTERNS?



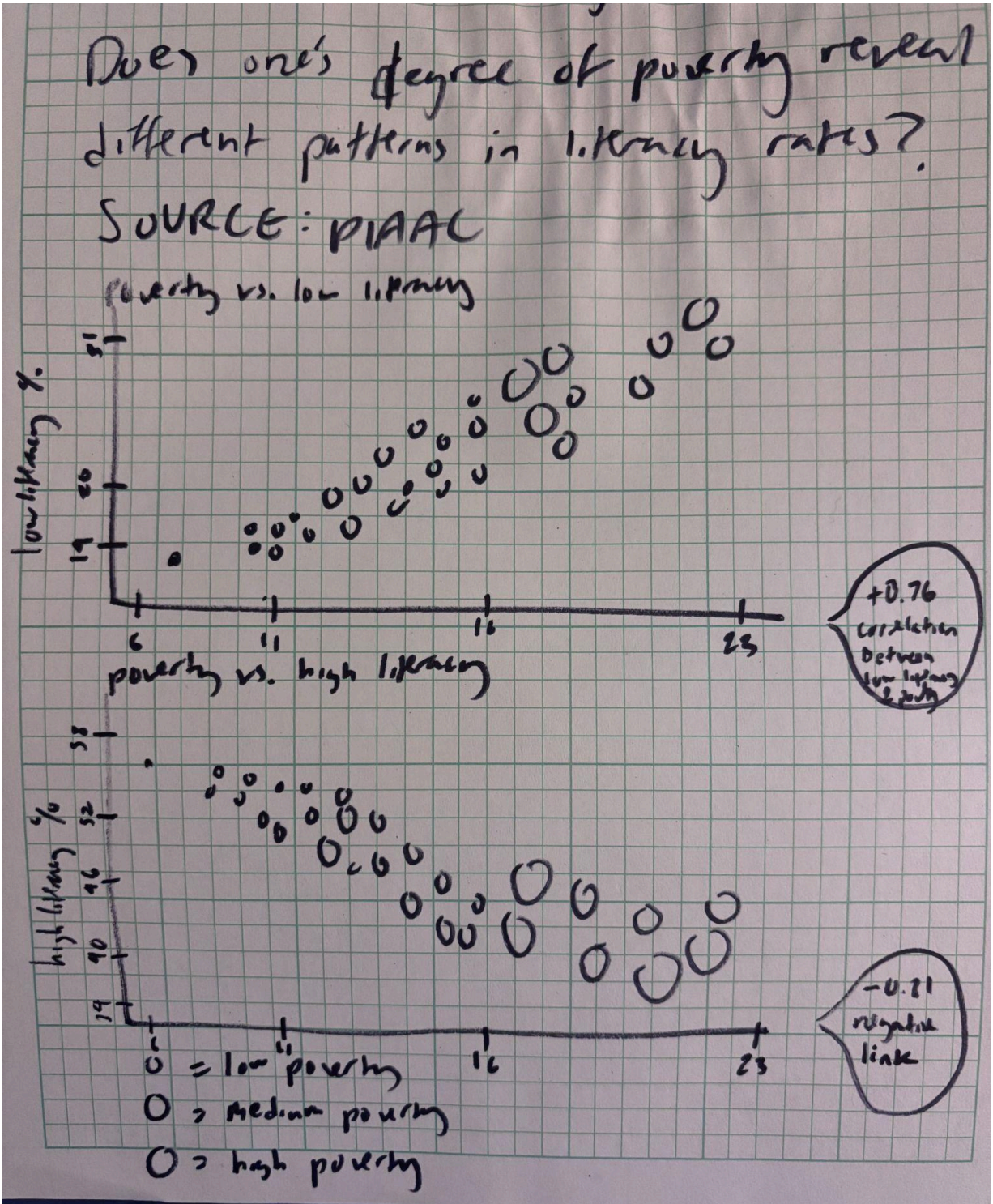



## Chloe’s Sketches:

Does one’s degree of poverty reveal different patterns in literacy rates?
Sketch ID: 13
Question ID:

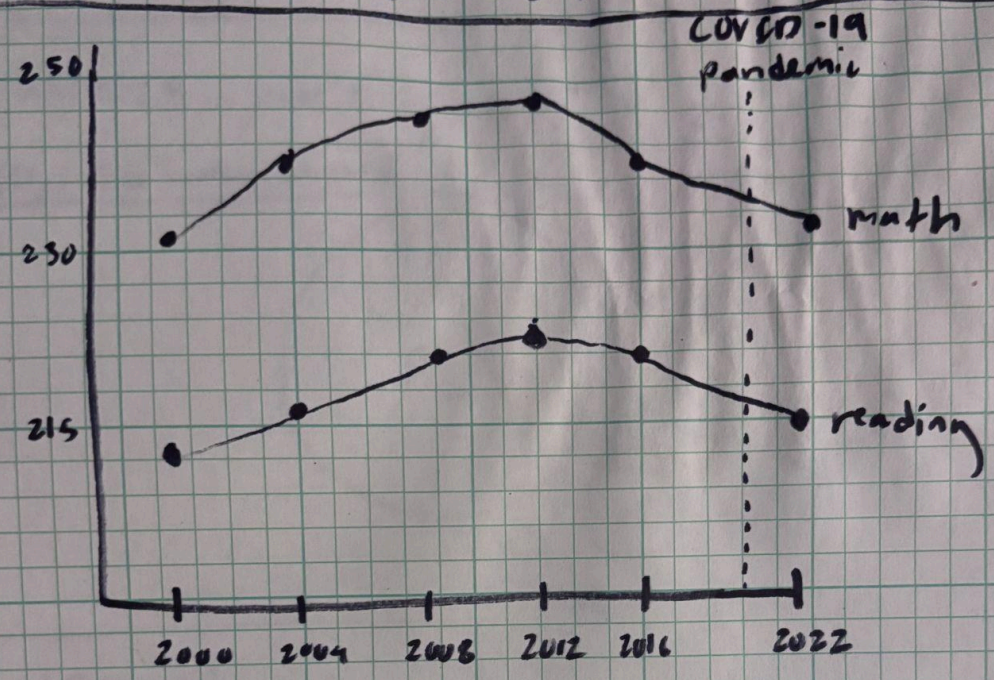
Does one's degree of poverty reveal different patterns in literacy rates?

SOURCE: PIAAC



How have NAEP Standardized testing scores in the U.S. changed over time, narrowed to the years surrounding the COVID-19 pandemic?
Sketch ID: 14
Question ID:

# U.S. NAEP standardized test scores over time




● = average test score at that year

note = missing 2020 data because of COVID

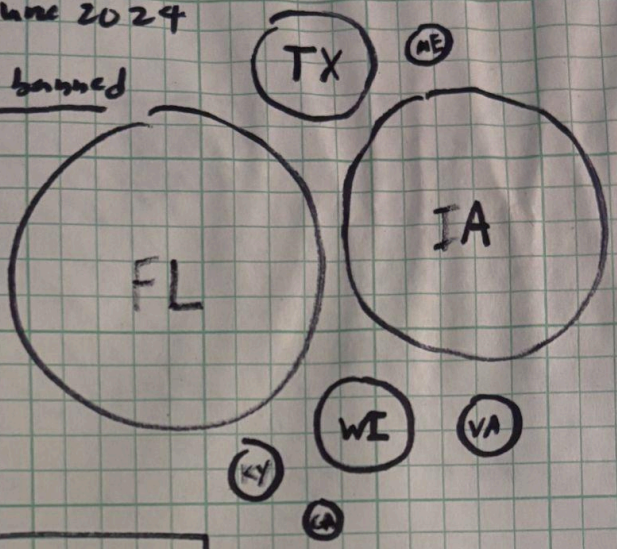
SOURCE: PEN

What is the number of books successfully banned by state?
Sketch ID: 15
Question ID:

# School book bans by state

july 2023 – june 2024

10,096 books banned




## Legend:

Florida = 4561 bans

Iowa = 3671 bans

Texas = 538 bans

Wisconsin = 908 bans

Virginia = 121 bans

Kentucky = 100 bans

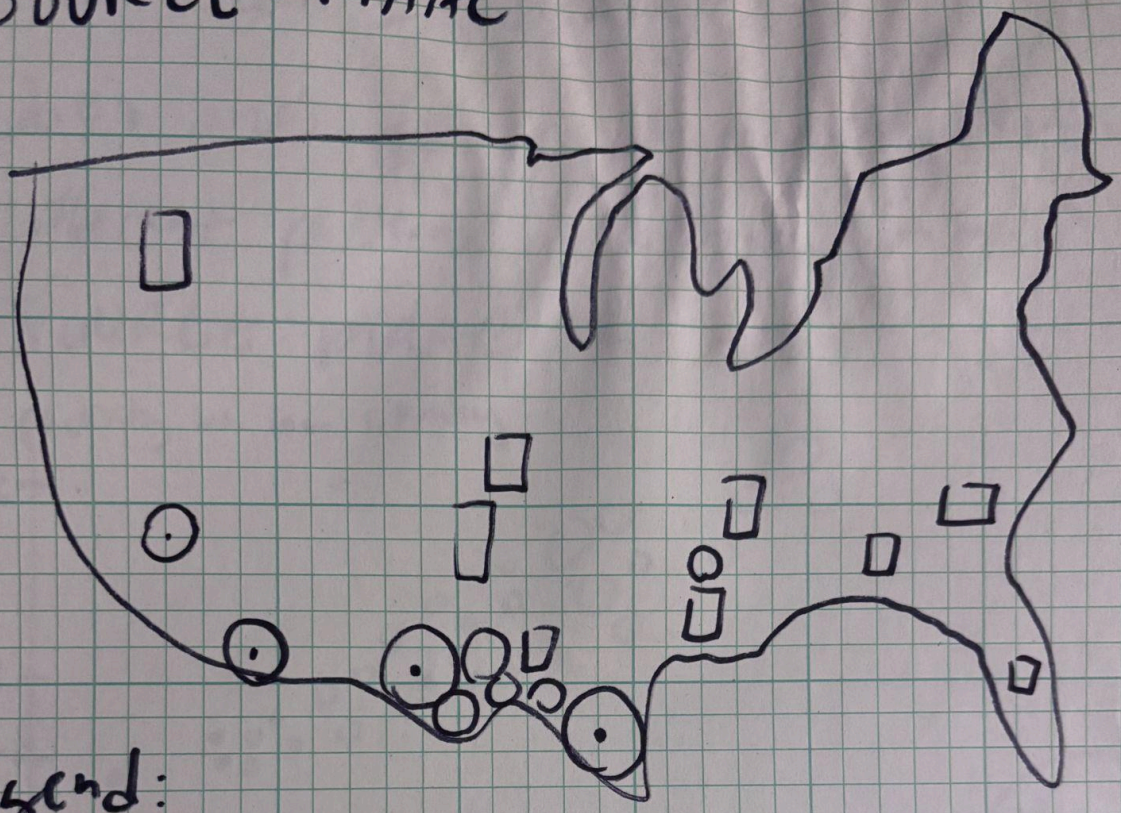
Georgia = 83 bans

Maine = 83 bans

SOURCE: PEN

Which areas of the country have the highest percentage of overall population below Level 1 literacy?
Sketch ID: 16
Question ID:

SOURCE: PIAAC



legend:

□ = 46-50% at/below level 1 literacy

○ = 50-55% at/below level 1 literacy

⊙ = >55% at/below level 1 literacy

Which areas of the country have the highest overall percentage of overall population below/at level 1 literacy?

Which states have the highest overall average literacy rates for youth/young adults? (ages 16-24)
Sketch ID: 17
Question ID:

Average literacy scores for ages 16-24

SOURCE: PIAAC

legend

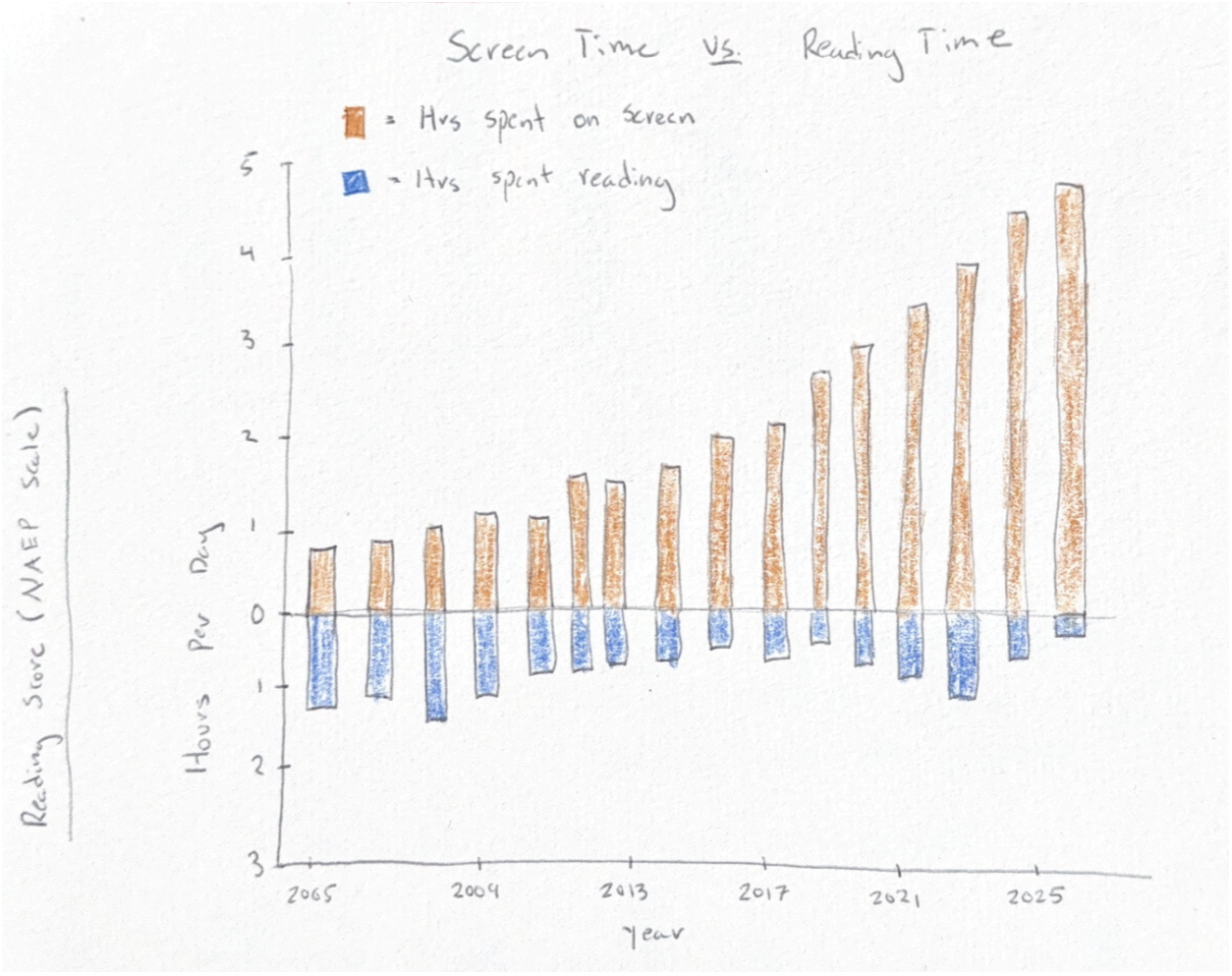
= high 280-350

= med 250-280

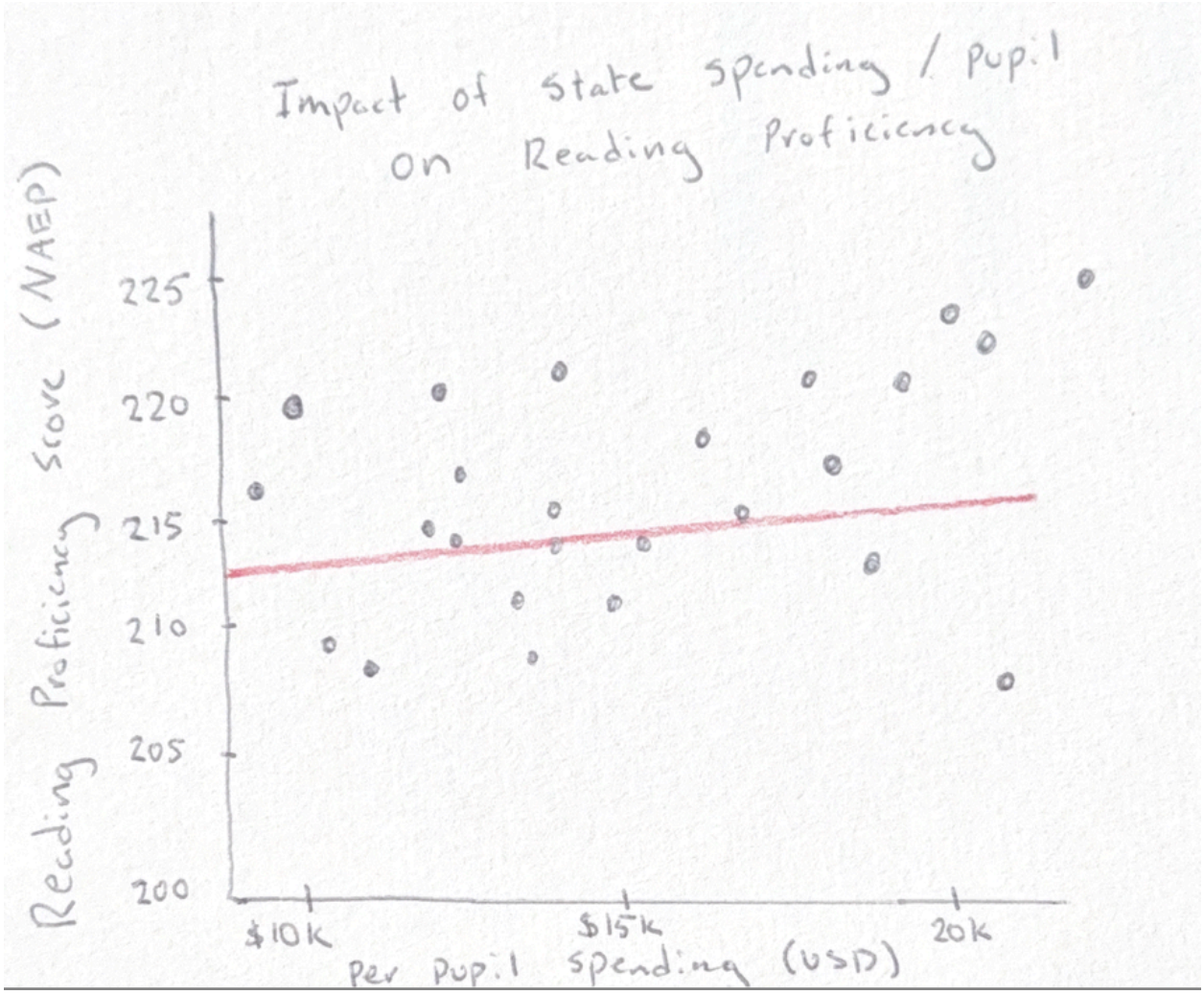
= low 200-250

Oakley Browning:

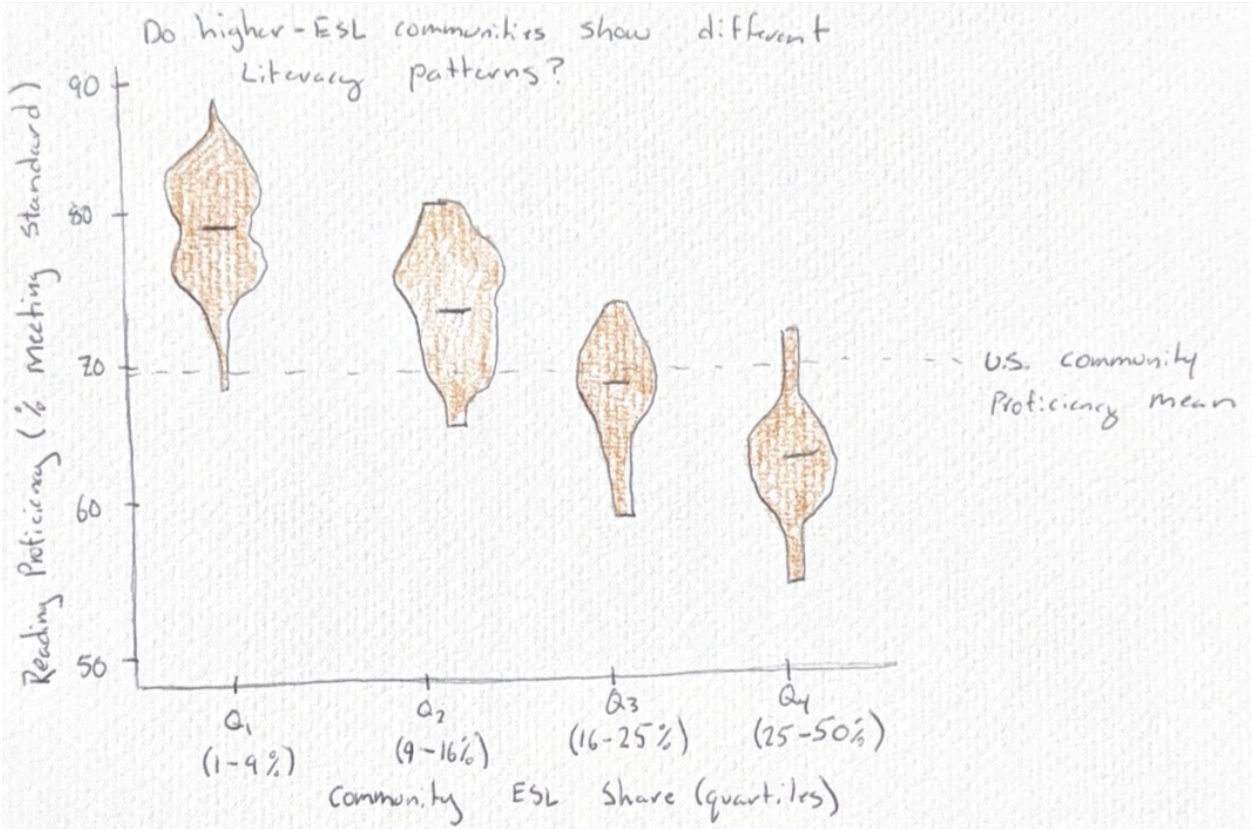
To what extent has digital consumption displaced reading since 2005?
Sketch ID: 18
Question ID: 6



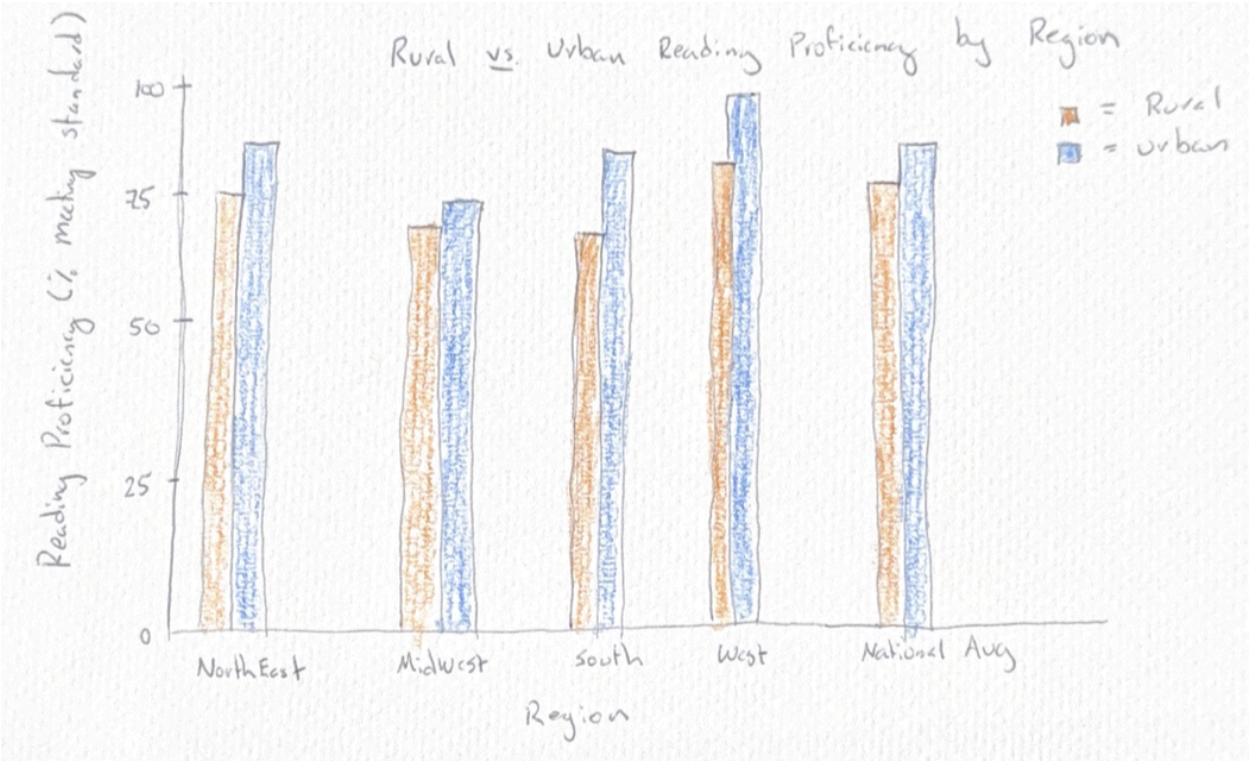
How does state spending per pupil relate to student reading proficiency?
Sketch ID: 19
Question ID: 2



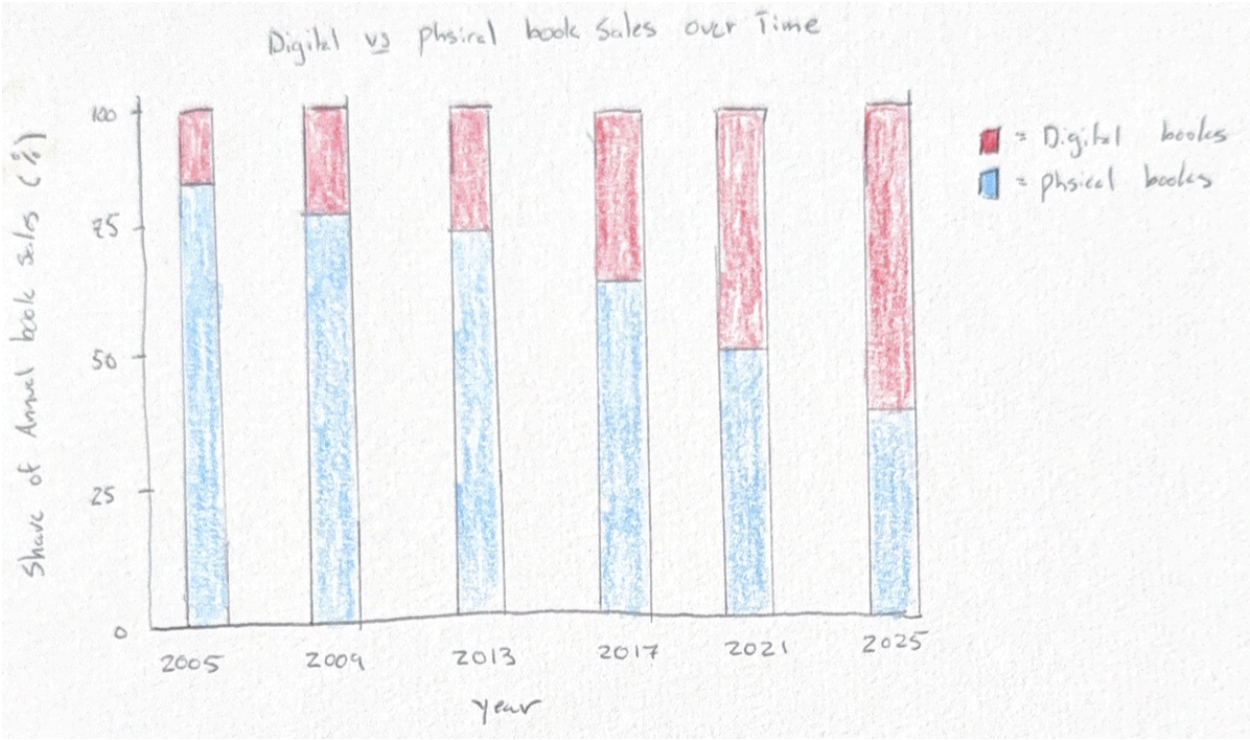
Do communities with higher rates of English learners show lower literacy proficiency?
Sketch ID: 20
Question ID: 4



How does reading proficiency differ between rural and urban communities?
Sketch ID: 21
Question ID: 4, 5



How have digital and physical book sales changed relative to each other over time?
Sketch ID: 22
Question ID: 6



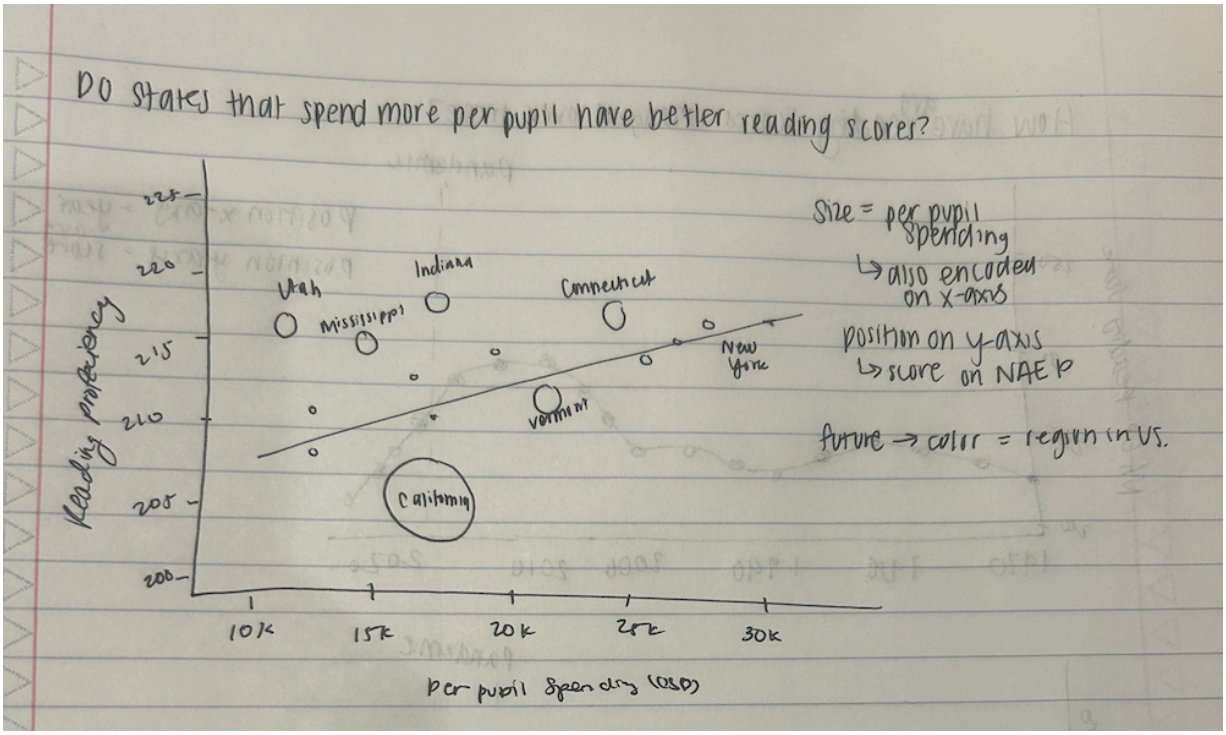

## Milestone 7: Decide

| Sketch ID | Question Group ID | Author | Votes |
|---|---|---|---|
| 3, 4, 6, 7, 14 | 1 | AB, SU, CM | Sketch 3: 1<br>Sketch 6: 2 |
| 1, 19 | 2 | AB, OB | Sketch 1: 3 |
| 5, 8, 15 | 3 | AB, SU, CM | III<br>Sketch 5: 1<br>Sketch 8: 3<br>Sketch 15: 1 |
| 2, 9, 10, 11, 12, 21, 20, 13 | 4 | AB, SU, OB, CM | Sketch 9: 1<br>Sketch 20: 1<br>Sketch 13: 3 |
| 8, 16, 17, 21 | 5 | SU, CM, OB | II<br>Sketch 16: 1<br>Sketch 21: 1 |

| | | | |
|---|---|---|---|
| 18, 22 | 6 | OB | Sketch 18: 2 |

## Selected Top 5 Visualizations:

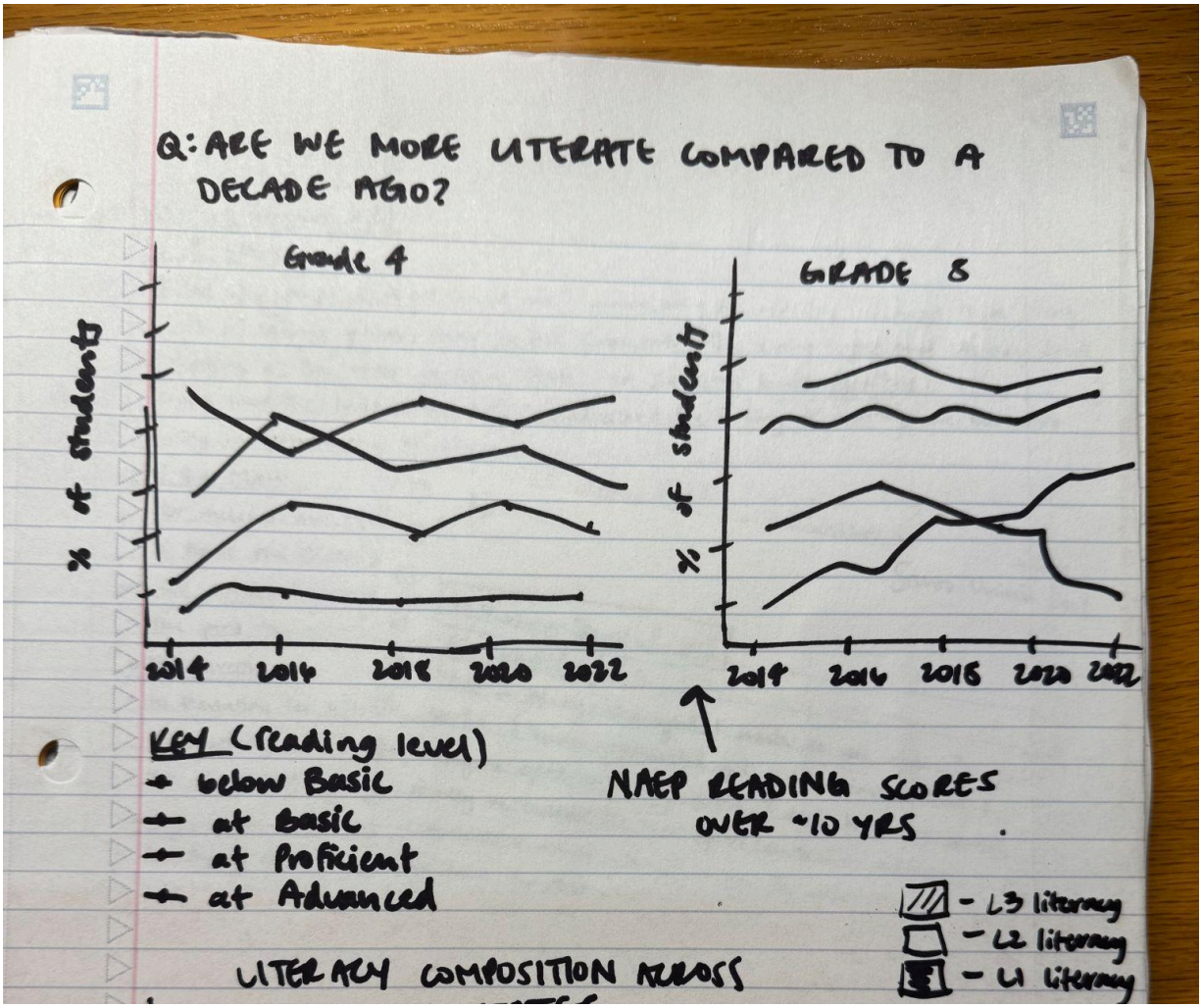
### Sketch #1




Do states that spend more per pupil have better reading scores?
Sketch ID: 1
Question ID: 2
Author: AB

### Sketch #6

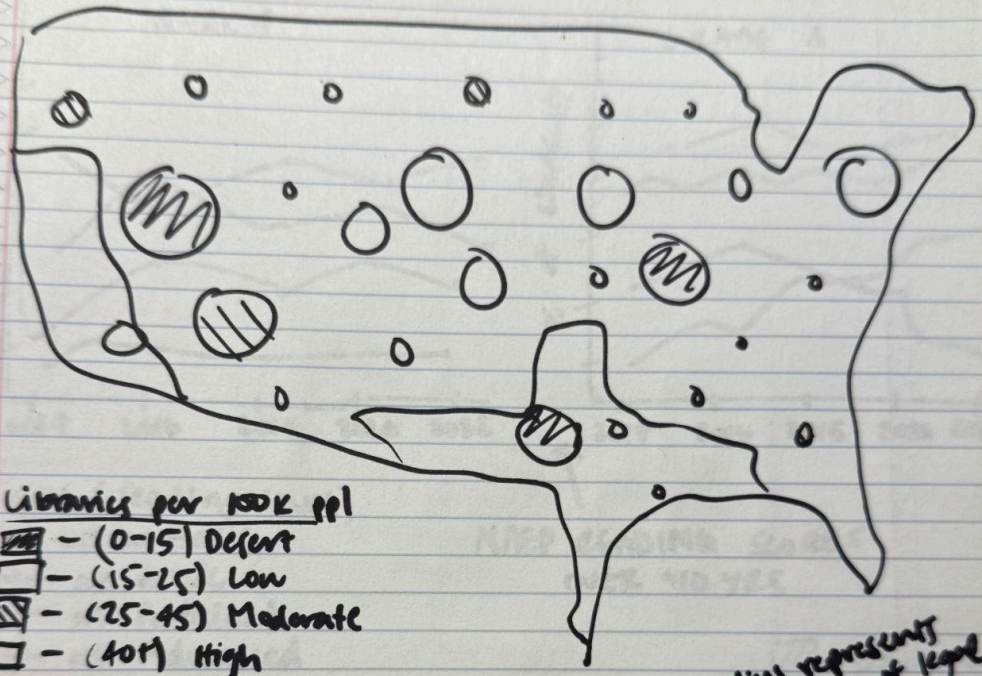
Are we more literate compared to a decade ago?
(Top) Sketch ID: 6
Question ID: 1
Author: SU



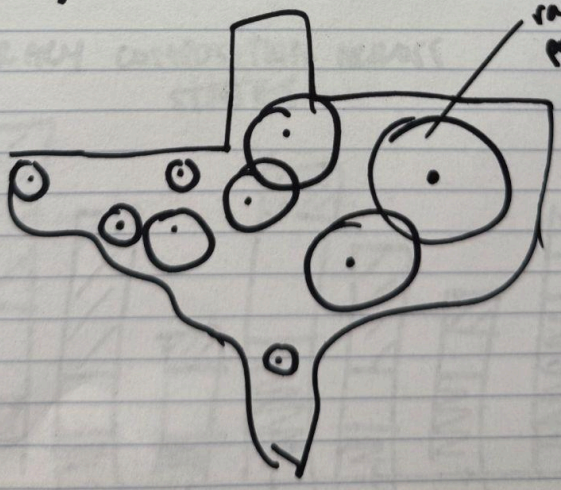

Sketch #8

| |
|---|
| Are there deserts where there are no bookstores/libraries available?<br>Sketch ID: 8<br>Question ID: 3<br>Author: SU |

Q: ARE THERE DESERTS WHERE THERE ARE NO LIBRARIES / BOOKSTORES AVAILABLE?

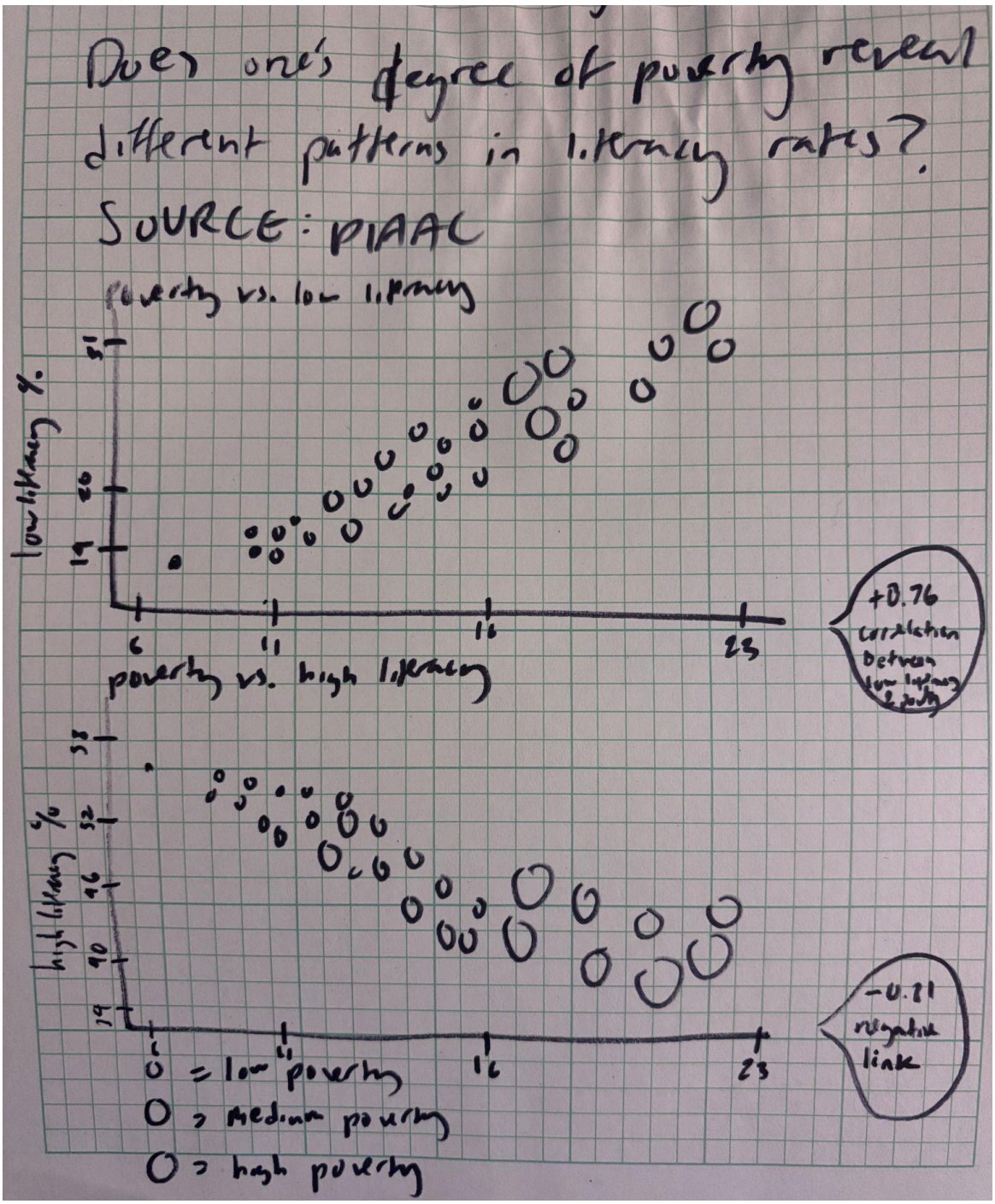


Libraries per 100K ppl
- (0-15) Desert
- (15-25) Low
- (25-45) Moderate
- (40+) High



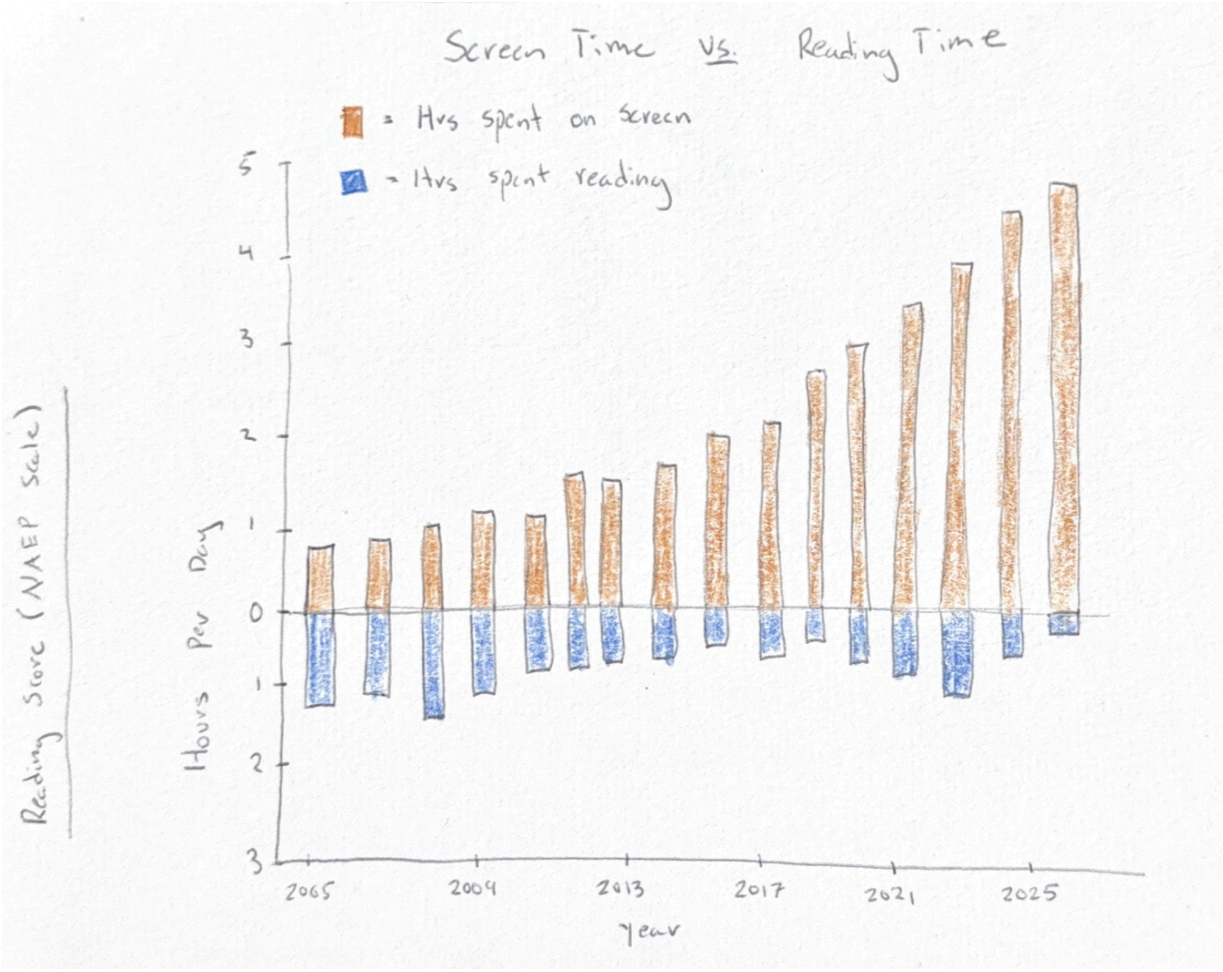
Sketch #13

Does one's degree of poverty reveal different patterns in literacy rates?
Sketch ID: 13
Question ID: 4
Author: CM



Sketch #18

To what extent has digital consumption displaced reading since 2005?
Sketch ID: 18
Question ID: 6
Author: OB




**Questions Grouped by ID:**

1. General literacy
    a. How have NAEP Standardized testing scores in the U.S. changed over time, narrowed to the years surrounding the COVID-19 pandemic?
    b. How have average NAEP reading scores changed over time in the US?
    c. Are we more literate compared to a decade ago?
2. Spending

    a. How does state spending per pupil relate to student reading proficiency?
    b. Do states that spend more per pupil have better reading scores?
3. Restrictions and deficits
    a. How many and what types of books are attempted to be banned from 2020 to 2024?
    b. What is the number of books successfully banned by the state?
    c. Are there deserts where there are no bookstores/libraries available?
4. Demographics
    a. Do communities with higher rates of English as a second language (ESL) show different literacy patterns?
    b. Do communities with higher rates of English learners show lower literacy proficiency?
    c. Does one's degree of poverty reveal different patterns in literacy rates?
    d. What are the most popular genres in the US per gender?
    e. Do men and women have different reading patterns and preferences?
5. Geographics
    a. Which areas of the country have the highest percentage of overall population below Level 1 literacy?
    b. Which states have the highest overall average literacy rates for youth/young adults? (ages 16-24)
    c. How does reading proficiency differ between rural and urban communities?
6. Digital Consumption
    a. To what extent has digital consumption displaced reading since 2005?
    b. How have digital and physical book sales changed relative to each other over time?

### Reflection:

Our group began by listing every question we had utilized to sketch visualizations. We then sorted them into themes (i.e. book restrictions or digital consumption) since many overlapped. Once we assigned IDs to all of the sketches, we then sorted them by the question(s) they sought to answer. The five sketches with the most votes each represented a different question category. This was purely coincidental, but will likely be helpful in representing the broad complications of literacy rates. Some themes we explore include state comparisons, trends over time, and socioeconomic patterns. Each selected sketch uses a different visualization type, adding to the overall interest and variety of our project.

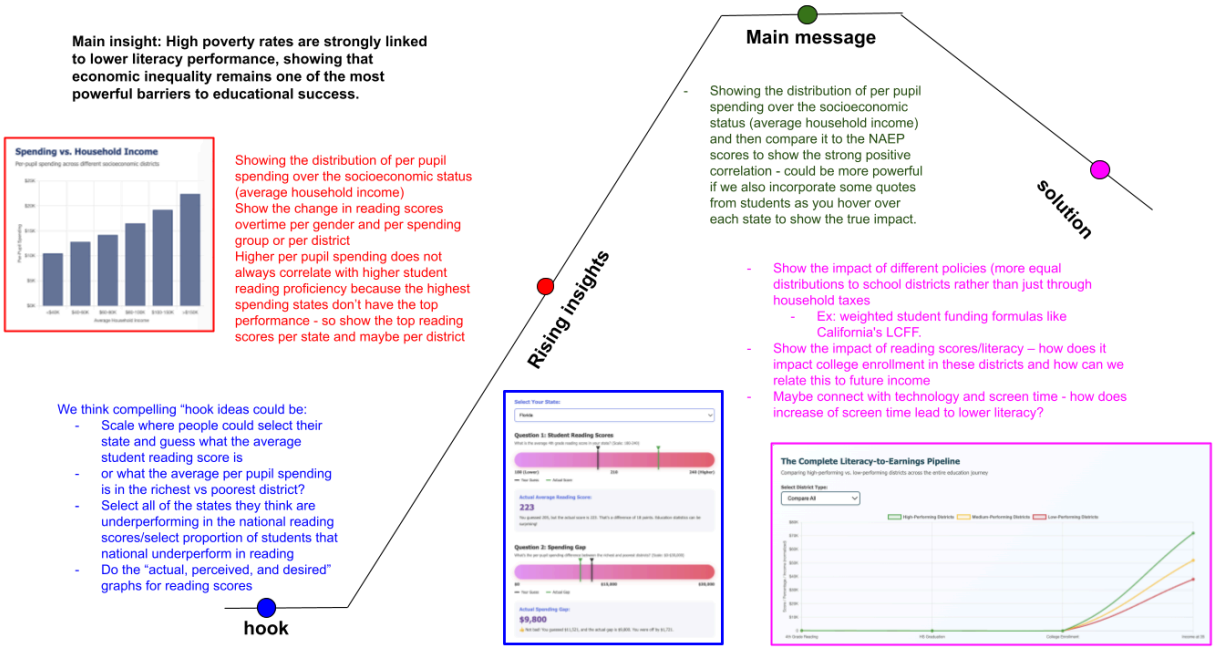
## Milestone 8 | Storyboard

List of Insights:

1. **Low literacy rates and high poverty rates are positively correlated.- Chloe**
    a. The same goes for the reverse = poverty and high literacy rates are negatively correlated. This is interesting because it pulls on two quantitative measures yet reveals a significant societal insight
2. **Below-proficient literacy levels are concentrated in the southern half of the country - Chloe**
    a. We could look into a lot of different regional and political factors if we investigated possible causes of this. This could point to demographic differences, or differences in education systems, etc.
3. **NAEP scores in ELA were already declining before the COVID pandemic hit - Chloe**
    a. This kind of surprised me because it would make sense if NAEP scores peaked right before COVID hit but instead they were already on a slight decline before 2020 and peaked a few years before then, as seen in Sketch 3. This could point to the patterns we see in screen usage increasing as reading time decreased, and other insights from our other graphs.
4. **Higher urbanization is positively correlated with higher literacy rates - Anchal**
    a. This isn't a surprising insight because in rural communities, there are on average households with lower socioeconomic status and more limited community resources which explains why urban students often score higher on literacy exams.
5. **Higher per pupil spending does not always correlate with higher student reading proficiency because the highest spending states don't have the top performance. - Anchal**
    a. This could be an interesting insight because it shows that the use of the money (targeted, effective allocation) might be more important than just spending more money. We could look at the different factors that go into higher student reading performance beyond just district-by-district spending.
6. **Higher poverty rates are strongly associated with lower literacy performance – Oakley**
    a. This wasn't surprising, as economic hardship is closely tied to educational disadvantage. Districts with higher poverty show significantly lower standardized reading outcomes, as indicated by a strong negative correlation ($r = -0.66$). This suggests that economic disadvantage remains one of the most powerful predictors of literacy gaps. The relationship affirms how limited resources, instability, and fewer educational supports directly constrain students' reading development.
7. **Districts with larger English learner populations tend to have lower average literacy scores, but not as drastically as expected – Oakley**
    a. This was somewhat surprising because, although language barriers often reduce test performance, the relationship ($r = -0.31$) was weaker than expected. It suggests that districts with higher ELL populations may be mitigating these effects through bilingual programs or targeted reading interventions.

We chose: High poverty rates are strongly linked to lower literacy performance, showing that economic inequality remains one of the most powerful barriers to educational success. We chose this statement because it allows us to look at different measures of poverty (per pupil spending, household income, school funding, etc) and then look at a measurable outcome that we can easily find through the NAEP data.

Storyboard:




# Milestone 9 | Prototype I

Drive: CS 171 Final Project

Team Member Documentation:


**Somto Unini**

**Anchal Bhardwaj**

**Oakley Browning**

**Chloe Manilay**

Figma Link:
https://www.figma.com/design/KIyidrrTAfzIICncAQGo5o/CS171-Draft-1?node-id=0-1&t=a4sea52l7PbC2OW1-1

https://www.figma.com/design/KIyidrrTAfzIICncAQGo5o/CS171-Draft-1?node-id=0-1&m=dev&t=a4sea52l7PbC2OW1-1

Requirements:

Your Prototype V1 submission must include:

1. **Team member documentation**: List names of all students who actively worked on this prototype submission in your process book
2. **Two functional D3 visualizations**: At least partially implemented with data loading and basic visualization rendering (interactive filtering not required yet)
3. **Additional visualization drafts**: Detailed designs for 2-3 more visualizations in your process book that you haven't implemented yet
4. **Website structure**: Rough webpage design and layout implemented (placeholders for visualizations, text, and images are acceptable)
5. **Clear storytelling**: Your narrative and message should be evident in the current structure
    a. Ideas:
        i. Contrasting what you think your state's literacy rate vs. what it actually is
        ii. Follow a student through their life based on socioeconomic status
        iii. Propose policy implementations
        iv. Impact of literacy scores on higher education (would need data)
        v. Bookstore/library desert map
6. **Innovative view design**: First design of your innovative visualization approach
7. **Updated process book**: All previous milestones plus new prototype documentation

**Submission**

- **Implementation code**: All current code in its present state
- **Process book**: Updated PDF version including innovative visualization and interaction designs
- **Package format**: Submit everything in a single zip file

**Narrative**:

- Largely follows the storyboard from earlier milestone

Users can select a few factors for a “choose your own story” - kind of like a “choose your player” view in an older videogame. The factors are state, name, district, socioeconomic status, and grade from 1-8.

From there, a student avatar is generated and then a book shelf will pop up next to the avatar. The books will each have a different visualization - and users will be guided to press on each book in order.

We will first start with a visualization where the user can select the “predicted” NAEP score for their student, the score it should be and then reveal the actual score → then transitioning to a visualization that allows the user to see how socioeconomic status and state can change the NAEP reading scores.

The next visualization will transition into literacy scores and per pupil spending - where the user will be able to see on average, how much schools are spending on their student and how that spending impacts their reading scores - this is all working towards the main message of how socioeconomic status impacts reading scores, and overall lifetime success.

We can then work towards visualizations that show how many libraries on average are available near the student, library programs, etc. Essentially, working through the book desert data.

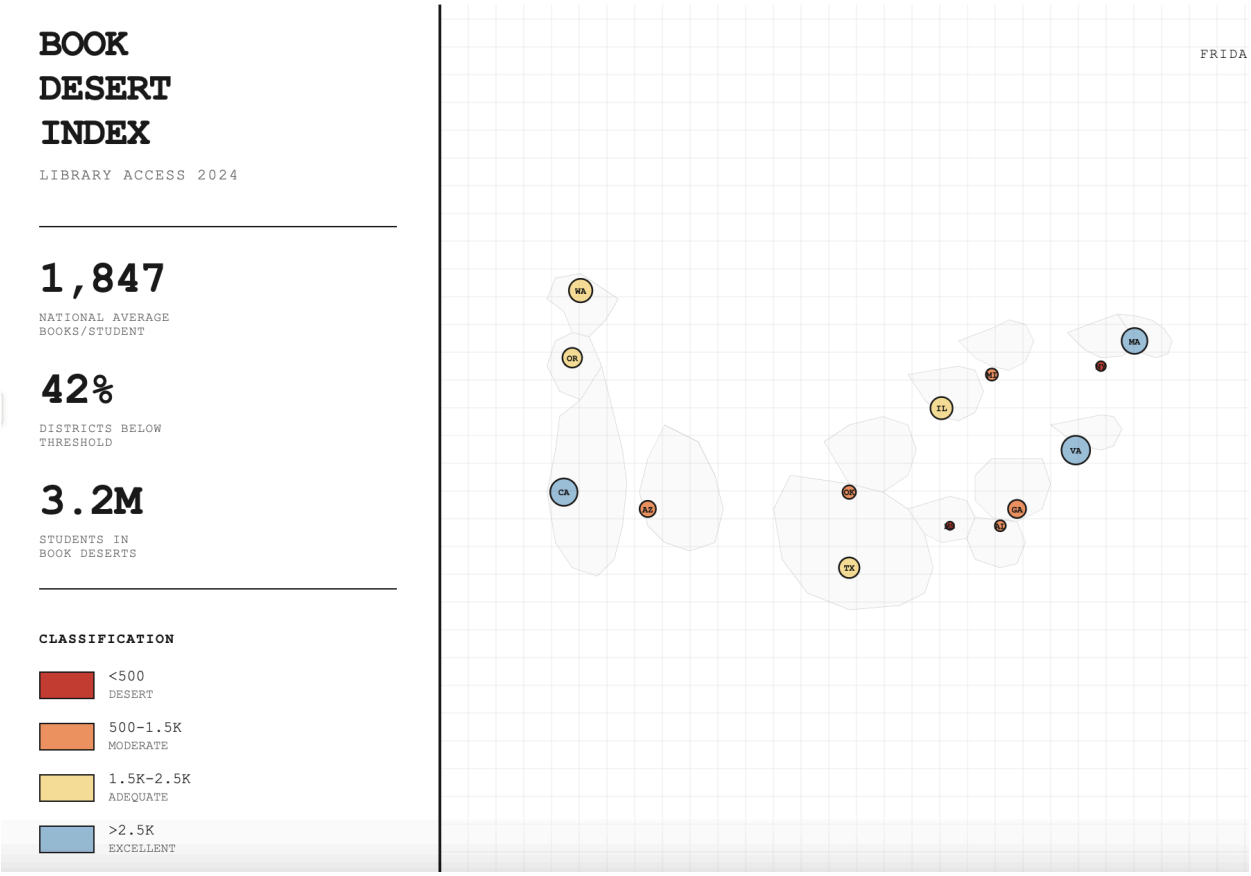
Lastly, we can show how literacy rates impact college outcomes/graduation rates. We can show what the average graduation rate would be for a student like the one the user generated would be and how that can impact future income/socioeconomic status.

## Visualizations + Mock Ups

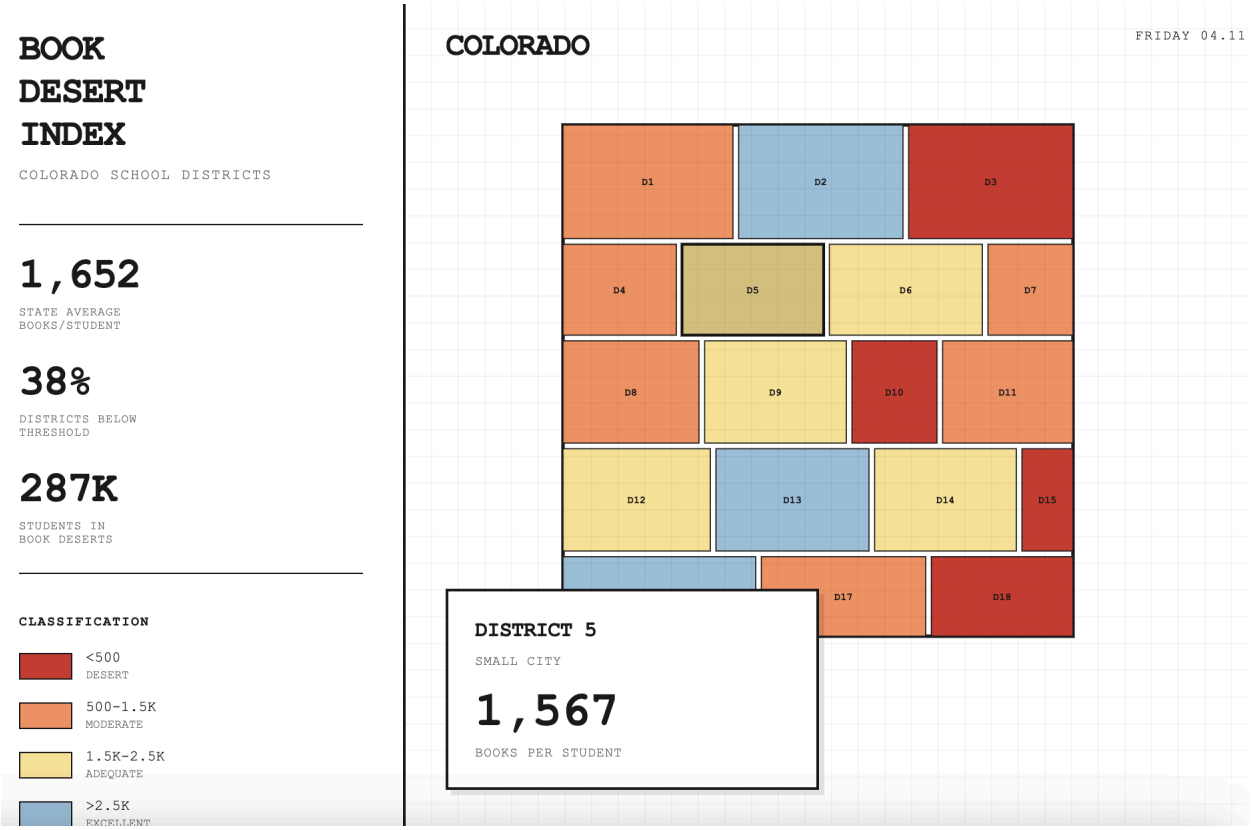
1. NAEP Score: Predicted vs Actual - **implemented**
2. Literacy Scores per state overall - **implemented**
3. Per Pupil Spending: Predicted vs Actual
4. Per Pupil Spending Overall

5. Libraries on average available per student - book desert
6. Outcomes - how does literacy impact college outcomes/graduation

**Mock Up - Libraries on average available per student - book desert**




But we can have actual state shapes - and then overlay them with bubbles showing the nearest cities and how available books are - under 500 would be categorized as a desert.



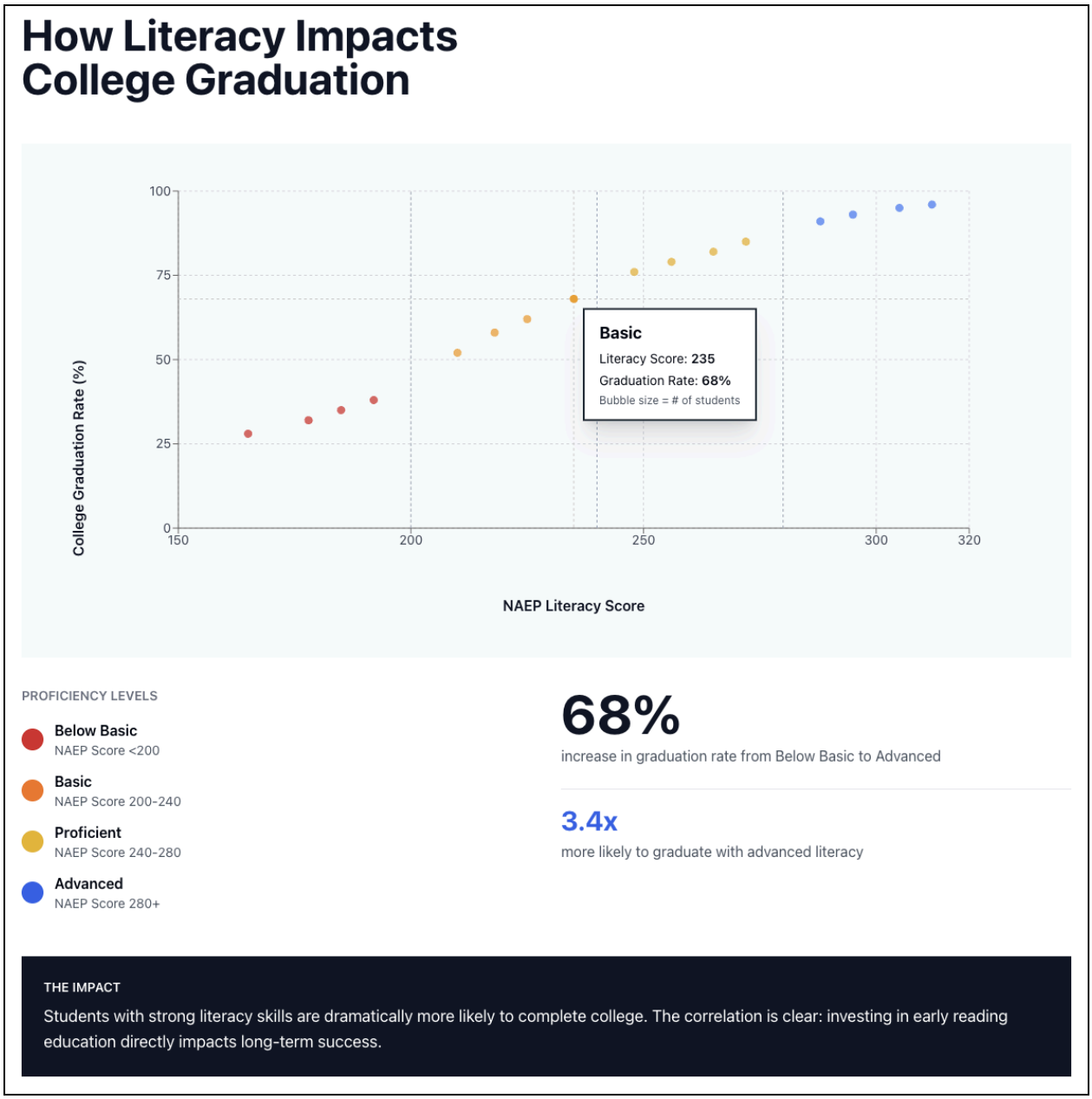

Would most likely do a district level mapping where you can see the numbers per district in each state (mocked up by Claude using synthetic data for Colorado). Color-coded by book access levels - each district filled with its classification color. District labels (D1, D2, etc.) centered in each area.

Hover interactions reveal:

- District name
- Type (Rural, Urban, College Town, etc.)
- Exact books per student count - per pupil

**Mock Up - Outcomes - how does literacy impact college outcomes/graduation**

Created a scatter plot visualization using synthetic data to demonstrate how literacy scores correlate with college graduation rates. Each bubble represents a student cohort at a specific NAEP literacy level, with bubble size indicating population and color showing proficiency classification (red = Below Basic, orange = Basic, yellow = Proficient, blue = Advanced). The x-axis shows literacy scores (150-320), the y-axis shows graduation rates (0-100%). Reference lines at 200, 240, and 280 mark proficiency thresholds.




# Milestone 10 | Prototype II

## Requirements

Your Prototype V2 submission must include:

1. **Team member documentation**: List names of all students who actively worked on this prototype submission in your process book.
2. **Complete data pipeline**: Data scraping and cleaning finished using your real datasets.
3. **Functional D3 visualizations:** All visualizations should be implemented with interactive functionalities. While some visualizations may still need refinement, there should be an initial **functional** version for all planned visualizations.
4. **Website structure:** Your webpage layout must be fully implemented and your design vision should be clear.
5. **Clear storytelling:** Your narrative and message should be evident in the current structure.
6. **Updated process book**: All previous milestones plus new prototype documentation
7.

## Milestone 11

Things to fix:

Viz5 - Financial

- ☑ ~~Doesn't use avatar info~~
- ☑ ~~Remove literacy pathway viz and make other two bigger~~
- ☑ ~~Needs context pop up~~
- ☑ ~~Viz shows up twice not sure why~~

Viz4 - Outcomes

- ☑ ~~Everyone is advanced proficiency in the NAEP visualization → redo scale and get TF opinion~~
- ☑ ~~Doesn't use avatar info~~
- ☑ ~~Use percentiles instead? Use a gradient? —> ask Chloe~~
- ☑ ~~Pop up that displays info about data point is too far from cursor~~
- ☑ ~~Change title (currently says "High School Outcomes") → this now shows college grad predictions~~
- ☑ ~~About this prediction section super uggo and needs to match aesthetic~~
- ☑ ~~Viz shows up twice not sure why~~
- ☐ Sanity check on the predictions → are the odds crazy?
- ☑ ~~Needs context pop up~~

Viz3 - Library

- ☑ ~~Highlight/pin avatar state~~
- ☑ ~~State info appears too far from the cursor, needs to be moved closer~~
- ☑ ~~Needs context pop up~~

Viz1 - Prediction

- ☑ ~~For predicted naep scores, check for API~~
- ☑ ~~Text for predicted scores overlaps/hard to read~~
- ☑ ~~Doesn't use avatar info~~
- ☑ ~~Feels kinda empty at the top; maybe do something about this~~
- ☑ ~~Pop up button and X out button don't click well~~
- ☑ ~~Needs context pop up~~

Viz2 - Reading Scores Map

- ☑ ~~Reading scores map, check for API~~
- ☑ ~~Doesn't use avatar info~~
- ☑ ~~Change state highlight color~~
- ☑ ~~State info appears too far from the cursor, needs to be moved closer~~
- ☑ ~~Needs context pop up~~

General

- ☑ ~~Update table of contents and make them clickable - **Chloe**~~
  - ☑ ~~Also make sure just 5 chapters and match chapter titles to viz titles on desk.html~~
- ☑ ~~Typing new letter changes the character on the home screen~~
- ☐ Little floating stars, dots, pins to label clickable items - **Chloe**
- ☑ ~~What's next page / Now what page - **Chloe**~~
- ☐ Should magnifying glass be clickable? → this could be what returns you to table of contents bc it would match investigate.html theme; open book could then become library desert map

Last Steps/Low Priority

- ☐ Final step: readme
  - ☐ What each visualization is
  - ☐ How to use/walkthrough example
- ☐ Put prototype 2 on git

<u>Context Pop Ups</u>

1. Viz1 - Predict your reading score

   **Let's Talk About Scores**

   Before we explore the data, let's start with a question: What do you think your reading score would be? Make a prediction on the scale below. Then we'll reveal your actual NAEP score and see how close you were.

This simple exercise reveals something important: the gap between what you predict and your actual score often reflects how invisible educational inequality really is. We tend to attribute reading achievement to individual effort, but location, funding, and resources play a massive role.

*Ready to make your prediction?*

2. Viz2 - Reading scores map

   **The Geography of Literacy**

   Where you grow up shouldn't determine your ability to read—but in America, it often does. This map shows how NAEP reading scores vary dramatically by state, with some regions consistently outperforming others by margins that represent years of learning. These gaps reflect systemic differences in school funding, teacher quality, and access to books.

   *Explore your state's scores across grades and years to see the full picture.*

3. Viz3 - Library deserts map

   **Where Are America's Library Deserts?**

   In some states, students have access to 20+ books per person through public libraries. In others, that number drops below 5. This visualization maps "library deserts"—regions where access to books and literary spaces is critically limited. These gaps affect millions of K-12 students and reflect systemic underinvestment in public infrastructure that supports literacy.

   *Explore your state's book access category and see how it compares nationally.*

4. Viz4 - Outcomes graph

   **The Long Shadow of Early Literacy**

   A student who reads proficiently in 4th grade is 4x more likely to graduate college than one who doesn't. But even among strong readers, outcomes diverge dramatically based on income and geography. This visualization predicts your college graduation likelihood using factors you encountered years ago: your reading score, your state's inequality, and your socioeconomic background.

   The scatter plot reveals an uncomfortable truth: literacy matters, but wealth and location often matter more.

   *See where you fall in the graduation prediction model.*

5. Viz5 - Pupil spending

**What's Your District Spending on You?**

Behind every reading score is a budget decision. This personalized visualization reveals how much money is invested in your education—and how that investment compares to students across your state and nation. You'll discover whether your district is above or below average, and see the broader truth: funding gaps often mirror literacy gaps, with profound implications for academic success.

*Explore your funding story and the patterns that shape educational opportunity.*

What can we do? Section:

# Literacy Organizations (US-Focused)

**National Organizations:**

- **Reading Partners** - www.readingpartners.org - Provides one-on-one tutoring in low-income communities
- **First Book** - www.firstbook.org - Distributes books and resources to 6.5 million kids annually in low-income areas
- **ProLiteracy** - www.proliteracy.org - Focuses on adult literacy education nationwide
- **Reading is Fundamental (RIF)** - www.rif.org - Reaches 91% of US elementary schools, serving 24 million children
- **Reach Out and Read** - www.reachoutandread.org - Provides books through pediatric checkups, reaching 4.4 million children (75% from low-income families)
- **Ferst Readers** - www.ferstreaders.org - Mails monthly books to children ages 0-5 in low-income communities across 11 states
- **Book Trust** - www.booktrust.org - Empowers kids from low-income families to choose their own books
- **Children's Literacy Initiative** - www.cli.org - Works with Pre-K through 5th grade teachers in under-resourced schools
- **Jumpstart** - www.jstart.org - Provides language and literacy programs for preschoolers in low-income neighborhoods
- **National Center for Families Learning** - www.familieslearning.org - Eradicates poverty through family education solutions
- **Imagination Library** - imaginationlibrary.com - Dolly Parton's program mailing free books to children birth to age 5

# Philanthropic Organizations Supporting Public Libraries

**Major Foundations:**

- **Carnegie Corporation of New York** - www.carnegie.org - Historic library funder, recently awarded $9 million to library systems nationwide
- **Knight Foundation** - www.knightfoundation.org/topics/libraries - Funds library innovation and community adaptation
- **Mellon Foundation** - Provides major grants for library support and operations
- **W.K. Kellogg Foundation** - www.wkkf.org - Supports libraries serving vulnerable children and communities
- **Laura Bush Foundation for America's Libraries** - www.laurabushfoundation.org - Funds school library collections
- **Wish You Well Foundation** - www.wishyouwellfoundation.org - Supports adult and family literacy programs

**Library Advocacy Organizations:**

- **EveryLibrary** - www.everylibrary.org - 501(c)(4) supporting library funding campaigns nationwide
- **EveryLibrary Institute** - www.everylibraryinstitute.org - 501(c)(3) think tank focused on library policy and advocacy
- **American Library Association (ALA)** - www.ala.org - Major lobbying force for libraries with various grant programs
- **Little Free Library** - littlefreelibrary.org - Grassroots book-sharing movement

These organizations accept donations, offer volunteer opportunities, and provide various ways to support literacy and library access in underserved communities.