Socioeconomic Deprivation Indices as Compared to High School Quality

VisualizEd

CS5764: Information Visualization
Virginia Tech
Fall 2023

By: Alexandra Thompson, Ayush Roy,

Jackson Livanec, and Sandhya Vinukonda

Table of Contents

Teams, Topics, and Charter

Team Name

VisualizEd (pronounced as Visualized)

Team Members

- Alexandra Thompson, alexthompson06@vt.edu
- Ayush Roy, ayushroy24@vt.edu
- Jackson Livanec, jlivanec@vt.edu
- Sandhya Vinukonda, sandhyav@vt.edu

Short Description

This project is inspired by Alexandra's current CS PhD research on socioeconomic diversity in introductory computer science courses (CS1). Her prior research asks whether the socioeconomic neighborhood status of a student has any correlation with their CS1 final grades. She used neighborhood socioeconomic indices (which are primarily used in the medical field) to best estimate a student's individual socioeconomic status. The primary socioeconomic index I used is the Area Deprivation Index (ADI) created by the University of Wisconsin-Madison's Center for Health Disparities, but I also supplemented this with the Social Deprivation Index (SDI) by the Robert Graham Center in the full paper.

(Video on prior research: https://youtu.be/_e9x3tLqu4c?si=Qa9iWZpF3NKYsn0j)

The primary goal of this project asks: **Does regional socioeconomic deprivation indices correlate to the quality of education students received at high schools in those regions?**

By comparing primarily medically-used socioeconomic deprivation indices (see ADI & SDI) to high school quality indicators found on the National Center for Education Statistics website and Virginia School Quality Profiles. Some considerations for determining a "high school's quality" could be indicators like chronic absenteeism, dropout rate, graduation, AP course offerings, spending per pupil, and school wide performance on state standardized testing. This work would be able to pave the way for researchers in education to use socioeconomic deprivation indices in their future work.

Communication & Collaboration Plan

Team communication channels and policies:

- We will use Discord as our primary platform of communication and email as our secondary. We will use Discord Threads and Google Document notes to organize task communication. We will also implement a policy of a maximum of 24-hour response time to any communication within the team. This is to ensure reliable communication.

How often we will meet and when: Weekly, Monday after class (Wednesday after class when necessary)

Methods of meeting: In-Person (Virtually, as needed)

Inter-team disputes:

- Fight to The Death (just kidding)
- Call team meeting, talk about dispute with entire team
    - Communication ceases between disputing team members until emergency team meetings

Roles and Responsibilities

- Alexandra

  Supreme Product Captain: She manages the team, communication and the overall big-picture vision of the project. She also will lead collection of the data from resources. Her secondary role is to help support the team in all phases whenever necessary.
  Alexandra will specifically lead:
    - Phase 1 **Teams, Topics, and Charter**
    - Phase 2 **Abstractions**
- Ayush

  Boring Backend Beast: He will be in charge of any code that is traditionally backend (database management, interactions with the data, organization of code). He will help support the team by leading the coding phase of development. His secondary role is to help support the team in all phases whenever necessary.
  Ayush will specifically lead:
    - Phase 5 **Build**
- Jackson

  Analytics Wizard: He will be in charge of the analysis aspect of the code and grand-scale ideas concerning the look of the final product. He will work closely with the *Boring Backend Beast*™ to create these changes on the backend and will lead evaluation of the tool. His secondary role is to help support the team in all phases whenever necessary.
  Jackson will specifically lead:
    - Phase 3 **Tableau Prototype**
    - Phase 6 **Evaluate**
- Sandhya

  Frontend Superwoman: She will be in charge of any code that is traditionally frontend (design, color scheme, layout, organization of code). She will support the team by leading the design aspect of the final product. Her secondary role is to help support the team in all phases whenever necessary.
  Sandhya will specifically lead:
    - Phase 4 **Design**
    - Phase 7 **Final Presentation**

Abstractions

Domain Situation

Problem description:

- Difficult to collect socioeconomic data
- Not well studied
- Can socioeconomic indices be used as an approximation of a student's socioeconomic status? Or as an approximation of a high school's socioeconomic status?

Users:

- Educational researchers
- (secondary) High school and college teachers
- (primary) Potential policy makers on public educational funding
- (secondary) Potentially parents looking where to send their kids to school

Existing solutions:

- Collecting data themselves (hard to obtain!)
- Not including socioeconomic information

Data Abstraction

Area Deprivation Index (ADI) , Center for Health Disparities Research at the University of Wisconsin, https://www.neighborhoodatlas.medicine.wisc.edu/, public data condensed to an index.

- Dataset type: Table (.csv)
- Table 1: ADIs (FIPS version) across Virginia
  Data attributes, types, meanings:
    Census Block FIPS (categorical): The unique census block area that the neighborhood represents. Data Size: BigInt
    ADI_National (Quantitative, sequential): The ADI national percentile of the neighborhood. Data Size: Float
    ADI_Staterank (Quantitative, sequential): The ADI statewide percentage of the neighborhood. Data Size: Float
- Data size: 5963 rows x 3 columns (For Virginia)
- Linking to other sets: The FIPS will be able to be linked to other geographical representations (addresses, zip codes, etc)
- Important derived data tasks: Not within this dataset

Social Deprivation Index (SDI), Robert Graham Center, https://www.graham-center.org/maps-data-tools/social-deprivation-index.html, public data condensed to an index

- Dataset type: Tables (.csv)
- Table 1: Census Tract Table

Data attributes, types, meanings:

Census Tract FIPS (categorical): The unique census tract that the neighborhood represents . Data Size: BigInt

SDI_National (Quantitative, sequential): The SDI national percentile of the neighborhood. Data Size: Float

Census Tract Population (Quantitative, sequential): The number of people who live in that census tract. Data Size: Int

- Data size: 73056 rows x 3 columns (For USA)
- Linking to other sets: The FIPS will be able to be linked to other geographical representations (addresses, zip codes, etc).
- Important derived data tasks: Not within this dataset

- Table 2: County Table
  - Data attributes, types, meanings:

    County FIPS (categorical): The unique census tract that the neighborhood represents. Data Size: BigInt

    SDI_National (Quantitative, sequential): The SDI national percentile of the neighborhood. Data Size: Float

    County Population (Quantitative, sequential): The number of people who live in that census tract. Data Size: BigInt
  - Data size: 3143 rows x 3 columns (For USA)
  - Linking to other sets: The FIPS will be able to be linked to other geographical representations (addresses, zip codes, etc)
  - Important derived data tasks: Not within this dataset

National Center for Education Statistics, https://nces.ed.gov/ccd/schoolsearch, This data is collected annually directly from State Education Agencies (SEAs).

- Dataset type: Table (.csv)
- Table 1:

  Data attributes, types, meanings:

Data attributes	Data Types	Meaning	Size
NCES School ID	Categorical	A unique identifier assigned to each school by NCES (National Centre for Educational Statistics)	BIGINT
Low Grade	ordinal, ordered	The lowest grade/class number present in the school	VARCHAR
High Grade	ordinal, ordered	The highest class number present in the school	VARCHAR

School Name	Categorical	Name of School	
District	categorical	The name of the district the school is present in	VARCHAR
County Name	categorical	The name of the county the school is present in	VARCHAR
Street Address	Categorical	The name of the street the school is present in	VARCHAR
City	categorical	The name of the city the school is present in	VARCHAR
State	categorical	The name of the state the school is present in	VARCHAR
ZIP	categorical	The zip code in which the school resides in	INT
Locale	categorical	The name of the locale the school is present in	VARCHAR
Charter	categorical	Answers if the school is charter or not. Only answers are yes or no	VARCHAR
Magnet	categorical	Answers if the school is Magnet or not. Only answers are yes or no	VARCHAR
Title 1 School	categorical	Answers if the school is tier 1 or not.	VARCHAR
Students	Quantitative, sequential	The total number of students.	VARCHAR
Teachers	Quantitative, sequential	The total number of teachers.	float

Student:Teacher Ratio	Quantitative, sequential	The ratio between students and teachers, derived from above	Float
Type	categorical	the type of program that is focused on like regular, vocational (trade school or technical education) or special education school.	VARCHAR
Status	categorical	Indicates the current status of the school. Whether it's open or is going to open in future.	VARCHAR

- Data size: 2140 rows X 29 columns
- Linking to other sets: The FIPS will be able to be linked to other geographical representations (addresses, zip codes, etc)
- Important derived data tasks:
    - Combining aspects of the address
    - Filtering out closed schools

Virginia Dept. Education quality profile**,**
https://schoolquality.virginia.gov/#:~:text=Virginia's%20School%20Quality%20Profiles%20provide,parents%20and%20the%20general%20public.
Virginia's School Quality Profiles provide information about student achievement, college and career readiness, program completion, school safety, teacher quality and other topics of interest to parents and the general public.

- Dataset type: Table (.csv)

Table 1: Chronic absenteeism

Data attributes	type/s	meaning	Size
Year	categorical	information on year	Int
Division	categorical	information on county	Varchar

School	categorical	information on school	Varchar
Indicator	categorical	indicates category:Chronic absenteeism	Varchar
Description	categorical	information on school session	Varchar
Number of Students Missing 10% or More of the Days Enrolled	quantitative	indicates Number of Students Enrolled for Half the Year or more	Int
Number of Students Enrolled for Half the Year or More	quantitative	indicates Number of Students Enrolled for Half the Year or More	Int
Chronic Absenteeism Rate	quantitative	percentage of students who are habitually or persistently absent from school.	float

- Data size:1823 rows, 8 columns
- Linking to other sets: combination of division and school could be used link to other datasets.
- Important derived data tasks:
    - Combining division and school as key
    - Filtering out closed schools.

Table 2: Dropout rate

Data attributes	type/s	meaning	Size
Year	categorical	information on year	Int
Division	categorical	information on county	Varchar
School	categorical	information on school	Varchar

Indicator	categorical	indicates category:Dropout	Varchar
Description	categorical	information on school session	Varchar
Dropouts	quantitative	Number of dropouts	Int
Students in the Cohort	quantitative	Number of students in Cohort	Int
Cohort Dropout Rate	quantitative	Rate of cohort dropouts	float

- Data size:334 rows, 8 columns
- Linking to other sets: combination of division and school could be used link to other datasets .
- Important derived data tasks:
    - Combining division and school as key
    - Filtering out closed schools

Table 3: Graduation and completion

Data attributes	type/s	meaning	Size
Year	categorical	information on year	Int
Division	categorical	information on county	Varchar
School	categorical	information on school	Varchar
Indicator	categorical	indicates category:Dropout	Varchar
Description	categorical	information on school session	Varchar
Diplomas	quantitative	Information on number of people graduating with diplomas	Int
GEDs	quantitative	No. of GEDs	Int
Certificates of	quantitative	Indicated number of	Int

Completion		people attaining Certificate of Completion	
Still Enrolled	quantitative	Indicates number of people who are still enrolled	Int
Graduation Completion Index	quantitative	Rate of students graduated	Float

- Data size:334 rows, 10 columns
- Linking to other sets: combination of division and school could be used link to other datasets .
- Important derived data tasks:
    - Combining division and school as key
    - Filtering out closed schools

Table 4: Per pupil spending

Data attributes	type/s	meaning	Size
Year	categorical	information on year	Int
Level Code	categorical	Level Code	Varchar
Division	categorical	information on county	Varchar
School	categorical	information on school	Varchar
End-of-Year Average Daily Membership	quantitative	Indicates End-of-Year Average Daily Membership	float
School-Level Expenditures Per-Pupil Federal	quantative	Information on School-Level Expenditures Per-Pupil federal	Int

School-Level Expenditures Per-Pupil State	quantative	Information on School-Level Expenditures Per-Pupil federal	Int
School-Level Expenditures Per-Pupil Subtotal	quantative	Subtotal of School level expenditures per pupil	Int
Division-Level Expenditures Per-Pupil Federal	quantative	Information on Division-Level Expenditures Per-Pupil Federal	Int
Division-Level Expenditures Per-Pupil State	quantitative	Information on Division-Level Expenditures Per-Pupil State	Int
Division-Level Expenditures Per-Pupil Subtotal	quantitative	Subtotal of division level expenditures per pupil	Int
Total Per-Pupil Expenditures	quantitative	Total per pupil expenditure	BIGINT
Per-Pupil Exclusions	quantitative	Total per pupil Exclusions	BIGINT
Total Expenditures	quantitative	Total expenditure	BIGINT
Per-Pupil Federal Funds	quantitative	Total Federal funds per pupil	BIGINT
Per-Pupil State Funds	quantitative	Total state funds per pupil	BIGINT

- Data size:1811 rows, 16 columns
- Linking to other sets: combination of division and school could be used link to other datasets .
- Important derived data tasks:
    - Combining division and school as key
    - Filtering out closed schools

Table 5: Teacher quality

Data attributes	type/s	meaning	Size
Year	categorical	information on year	Int
Level Code	categorical	refers to code used to categorize and classify schools based on various criteria	Varchar
Division	categorical	information on county	Varchar
School	categorical	information on school	Varchar
Poverty Level	categorical	Level of poverty	Varchar
Title1 Code	categorical	specific code or identification assigned to schools that qualify for Title I funding	Varchar
Percent of Inexperienced Teachers	quantitative	Percentage of inexperienced teachers	float
Percent of Out-of-Field Teachers	quantitative	Percent of Out-of-Field Teachers	float
Percent of Out-of-Field and Inexperienced Teachers	quantitative	Percent of Out-of-Field Teachers	float

- Data size:3532 rows, 9 columns
- Linking to other sets: combination of division and school could be used link to other datasets .
- Important derived data tasks:
    - Combining division and school as key
    - Filtering out closed schools

Linking to other sets

- The ADI data set will be linked using FIPS code after we filter out data representing Virginia with SDI and NCES data set.
- Using the division and school, the above dataset will be linked to the Virginia Dept. Education quality profile data set.
- This key(division and school) will be used to link to the tables in Virginia Dept. Education quality profile data set **,**
- final dataset will be created with the listed attributes, enabling us to analyze the data and draw conclusions from it.

Task Abstractions

Domain questions

- 1) Users want to know where areas of low and high socioeconomics is.
- 2) Users want to know where schools with low/high funding, low/high academic achievement, low/high graduation rates, and where schools with other quality factors are located.
- 3) Users want to know how the socioeconomic status of an area compares to school quality measures
- 4) exceptions

Structuring, grouping, breakdown of the task space

- Should start with 1, then 2, then 3, then 4

- Let **D** represent the joined ADI <- SDI dataset and **Q** represent the joined educational quality dataset.

Q - These tasks rely on the education quality measures joined into a single dataset broken out by school / location

- Search specific schools (Target known, location known/unknown)
- See how two schools compare (identify)
- Derive a new metric/index

Questions

EQP :

1. Compare the school-level expenditure per pupil Federal and State. Same with division level expenditures. - Easy level question
2. See if there is any correlations between poverty level and chronic absenteeism. - Easy level question
3. See if there is a correlation between the total expenditure of the school and poverty level. - Easy level question
4. Is there any correlation between chronic absenteeism of each division wrt to each poverty level? - Easy level question
5. Compare the poverty levels of each division using stacked bar graph - Easy level question

6. Separate the poverty levels to show in a heatmap which county/division has the highest/lowest level of chronic absenteeism. - Easy level question
7. Is there a correlation between Cohort Dropout and division-level expenditure ? Is it different for different poverty levels, if so in what way and can we draw any conclusions from this ? - Medium level question
8. Is there any relation between the graduation completion index and division level expenditure per-pupil? - Medium level question
9. To explore the potential correlation between dropout rates and the percentages of Out-of-Field and inexperienced Teachers across different poverty levels especially considering the continuous nature of these variables? - Medium questions

D - These tasks rely on the ADI and SDI datasets joined into a single dataset broken out by location

- Does the spatial representation of socioeconomic status indicate any particular patterns throughout Virginia? - Hard level question
- What do these patterns reveal about the different geographical regions of virginia (e.g. mountains, tidewater, etc.) - Hard level question

Q + D - These tasks rely on the two previous datasets joined with one another to yield a single dataset containing both the SES measures and educational quality measures broken out by location

- Are there pockets in Virginia with similar correlational trends? What attributes of these pockets make them homogenous?
- What factors informing the socioeconomic index are the best indicators of educational quality? Why is this the case?
- What can be done with this information to maximize educational quality?

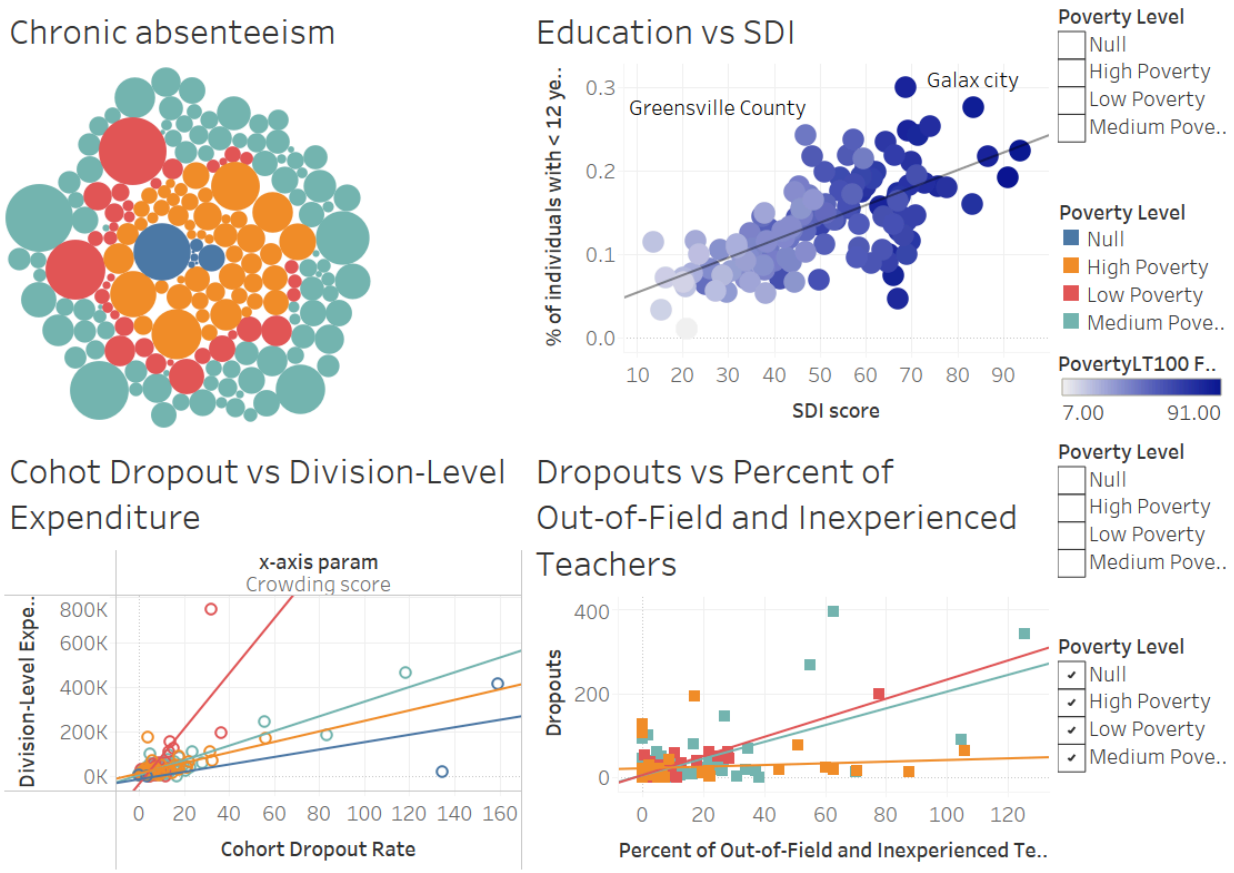
Any needed definitions of task domain concepts and terms

- SES: Socioeconomic Status - The idea of how well-off a person or group of people is based on social or economic factors. Some of these factors include: household income, household resources,
- ADI: Area Deprivation Index - A percentile between 1 and 100 that determines how socioeconomically advantaged/disadvantaged a census block group (neighborhood) is based on 17 socioeconomic indicators.
- SDI: Social Deprivation Index - A percentile between 1 and 100 that determines how socioeconomically advantaged/disadvantaged a census tract or county is based on 7 socioeconomic indicators.
- Chronic Absenteeism - Students who continuously skip or are absent from class. Does not include excused absences.

Tableau Prototype

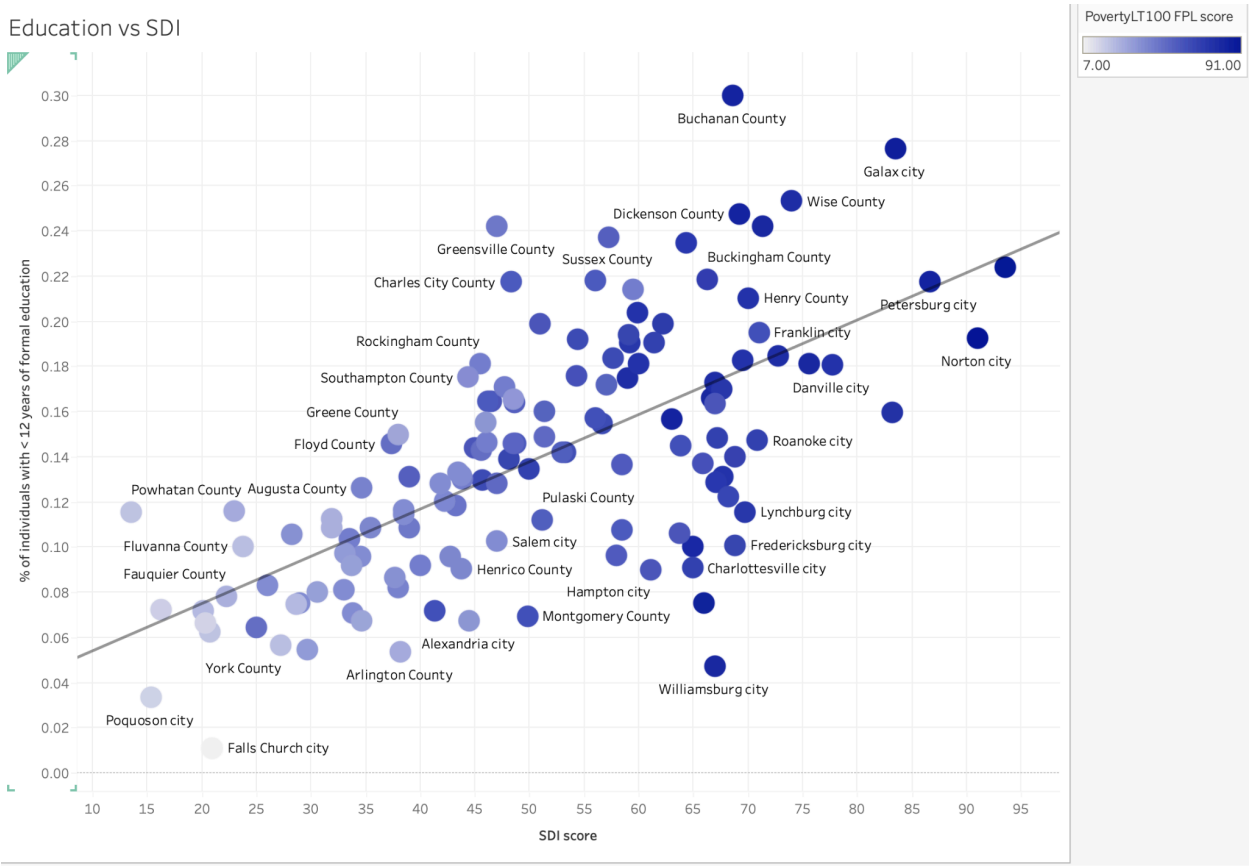
Tasks from Abstractions that have been answered:

Dashboard



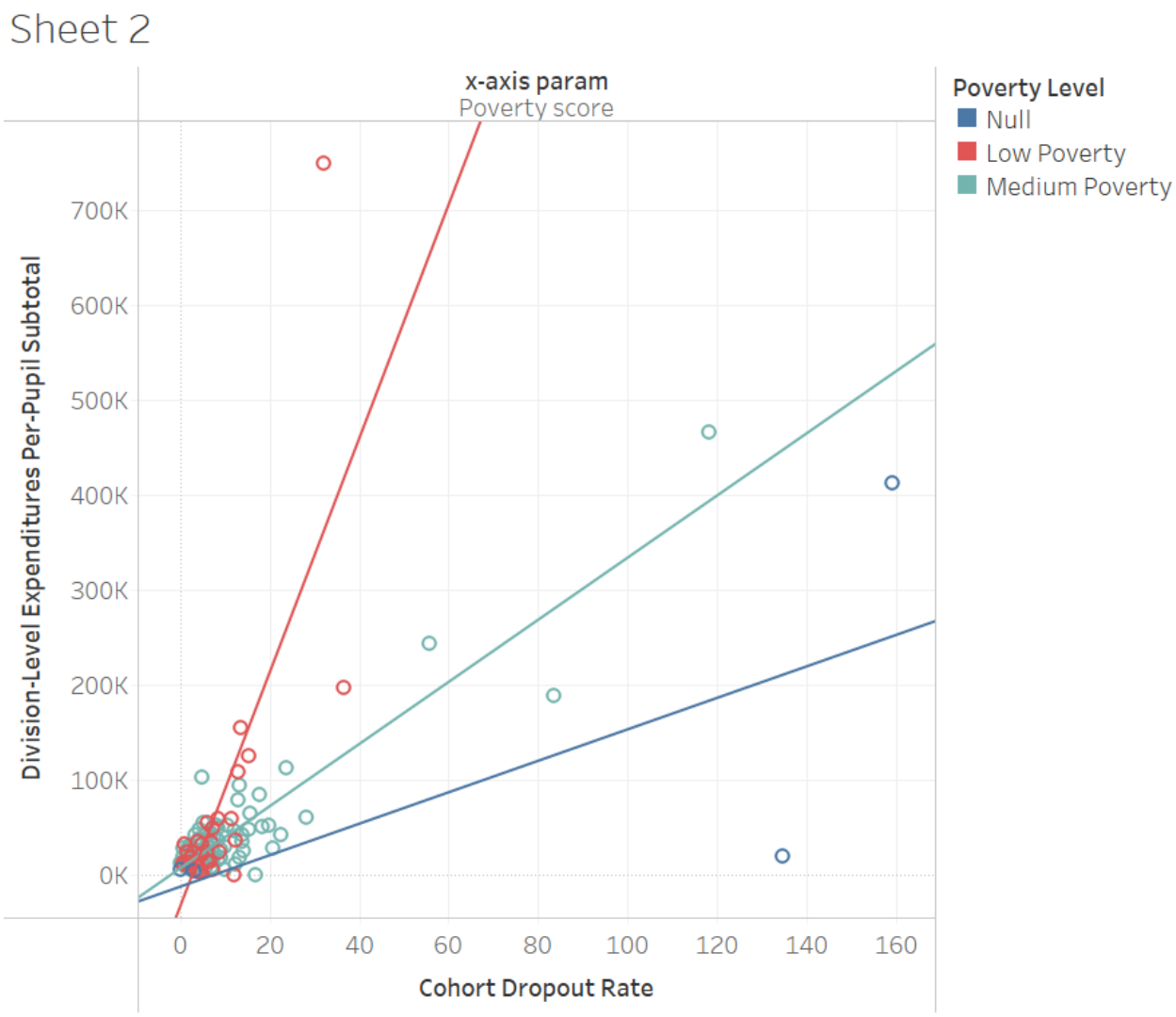
Q1. How does the SDI score effect education in different counties while taking into account varying poverty levels in each county, and what is the most effective way to represent this comparison visually? Are they any outliers, if so, when and what?

Ans:As we wanted to show correlation between sdi score and accessibility or affordability of education over counties I went with a scatter plot (which I think is best for showing correlations) .we plotted x-axis as sdi score and y axis as Education_LT12years_score(the percentage of individuals with less than 12 years of formal education)and poverty as a channel (color) to find some correlation in the given graph we can see a positive trend line which shows that there is a correlation.



Justification :

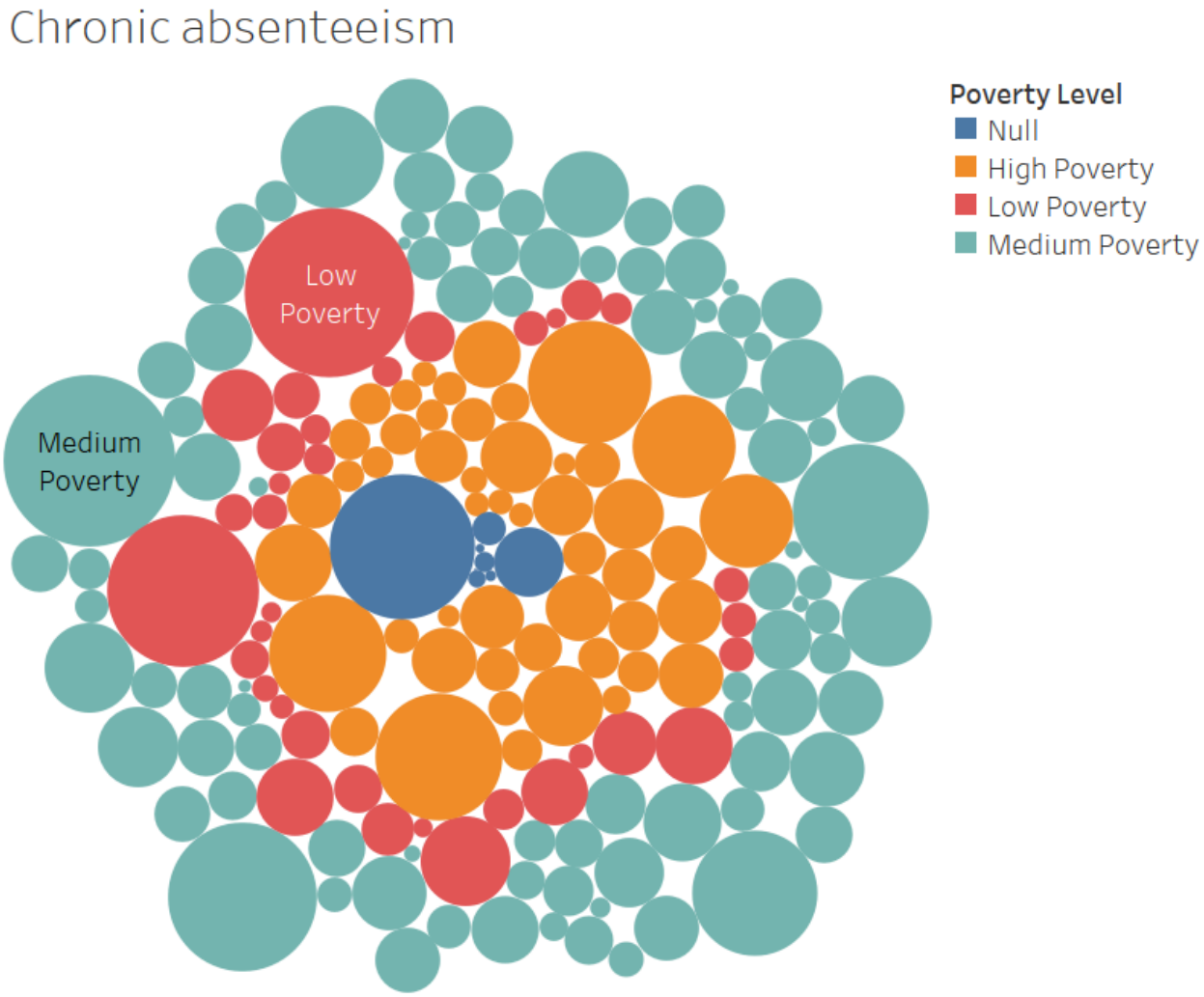
Q2. How does the Chronic Absenteeism rate change over time within different divisions while taking into account varying poverty levels in each division, and what is the most effective way to represent this temporal comparison visually? Are they any outliers, if so, when and what?



X-axis param (Parameters) and sum of Cohort Dropout Rate vs. sum of Division-Level Expenditures Per-Pupil Subtotal. Color shows details about Poverty Level. Details are shown for Division. The view is filtered on Poverty Level, which keeps Null, Low Poverty and Medium Poverty.

Justification : The question wants to see a trend in the cohort dropout rate and division-level-expenditure. To look at outliers, one of the best ways of visualization is scatter plot. Different colors are used for different levels of poverty whereas the size of each data point reminds the same.

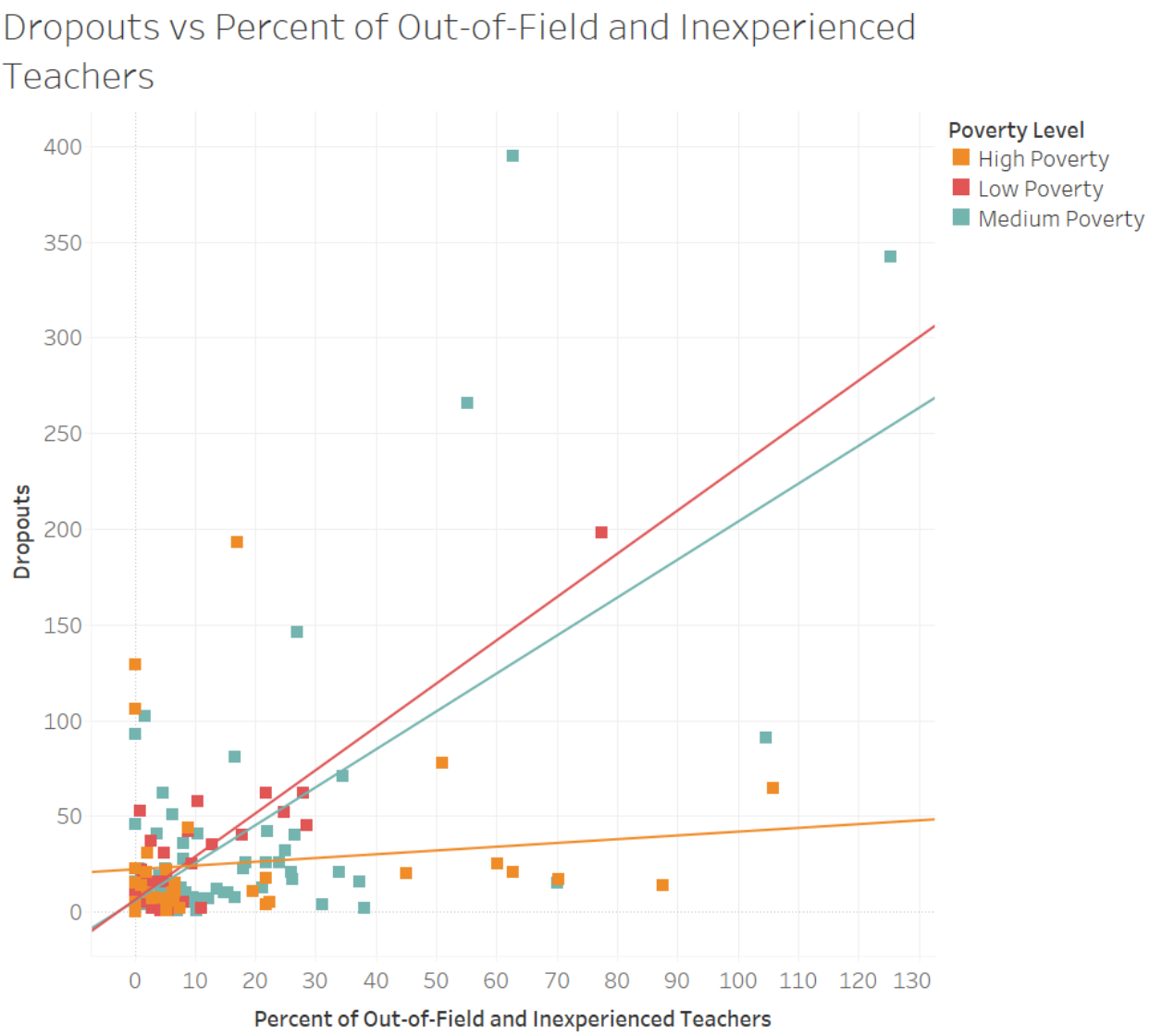
Q3. How does the Chronic Absenteeism rate vary across divisions while considering the different levels of poverty within each division and also as a whole, and what is the most effective way to visualize this comparison?



Poverty Level and Division. Color shows details about Poverty Level. Size shows sum of Chronic Absenteeism Rate. The marks are labeled by Poverty Level and Division. The view is filtered on Poverty Level, which keeps Null, High Poverty, Low Poverty and Medium Poverty.

Justification :
The question required us to compare density of chronic Absenteeism as a whole as well as parts with respect to poverty, hence we choose packed bubbles. This chart was also made dynamic so that we can compare different sections with just the click of a button.
The colors represent different levels of poverty : high, low and medium; making it easier for the user to easily identify. Also we used chronic Absenteeism as size of the bubbles to distinguish and put focus on counties with high chronic Absenteeism rates.

Q4. Analyze whether there exists a correlation between dropout rates and the "percentage of Out-of-Field and Inexperienced teachers" for students within each poverty level. Additionally, explore variations in this relationship across different poverty levels to discern any distinct patterns.



Sum of Percent of Out-of-Field and Inexperienced Teachers vs. sum of Dropouts. Color shows details about Poverty Level. Details are shown for Division. The view is filtered on Poverty Level, which keeps High Poverty, Low Poverty and Medium Poverty.

Justification : The question asks us to draw conclusions about the relationship between different data points. Since the data points are continuous, scatter plots are the best way to visualize them. Using scatter plots helps in seeing all the outliers and the relation between them better.

Limitations of Tableau :

In our dashboard we have used the following visual techniques :

1.Scatter plots:

1.In a large dataset, points can overlap, making it hard to distinguish individual data points and interpret density.

2.Customizing aspects of the scatter plot like scales, axes, or adding annotations is limited to Tableau's built-in functionalities.

3.Scatter plots in Tableau are primarily two-dimensional, analyzing more than two variables can be challenging and may require color, size, or shape encoding which can lead to clutter.

Improving Scatter Plots using Observable:

1.Using techniques like jittering, transparency, or interactive brushing to address overlap of points.Implementing zooming and panning could allow us to explore dense regions of the plot.

2. Creating a Multi-Dimensional visualisation where we can use interactive elements to allow users to explore relationships between more than two variables dynamically.

3.Customize every aspect of the plot, including scales, axes, annotations, and tooltips, to a great extent.

2.Packed bubbles

1.Dealing with numerous varied-sized bubbles, it can lead to overlap and clutter, making it difficult to interpret individual bubble values accurately.

2.It is challenging to extract precise quantitative values or compare sizes of bubbles.

3.Interactivity is limited to tooltips and highlight actions, and customization is constrained by the available settings in Tableau.

Improving Packed Bubbles using Observable:

Custom Interaction:
1.Use of JavaScript to create custom interactions, such as click, hover, or drag, to allow users to explore individual bubbles more deeply or to declutter the visualization could help.

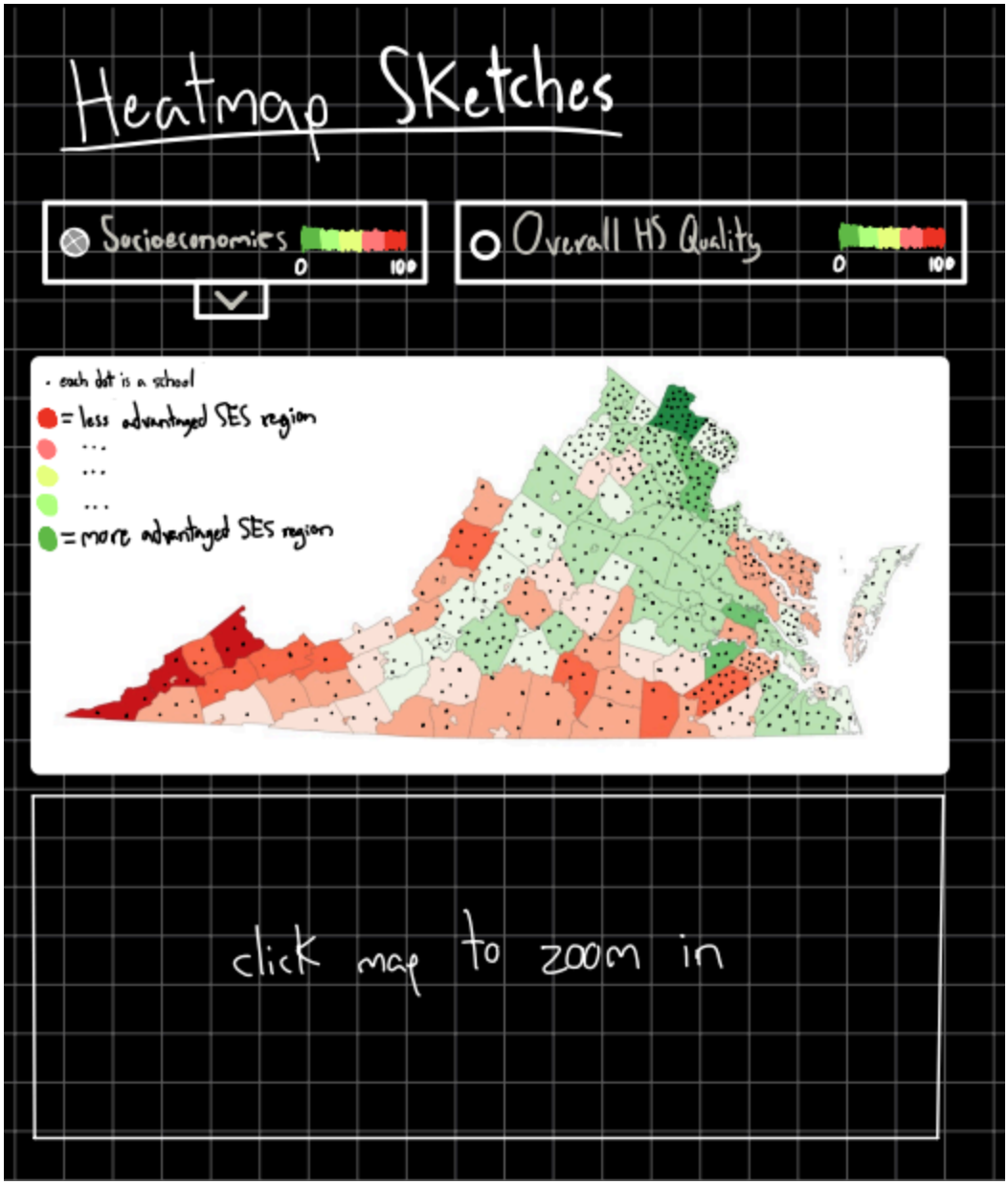
2.Implementing dynamic resizing and filtering options, allowing users to focus on specific subsets of the data could make it more insightful.

Also we plan to make a interactive heatmap adding all the visualizations with zoom and pan we tried this in tableau but the representation was not clear enough due to lack of customizations.

Design

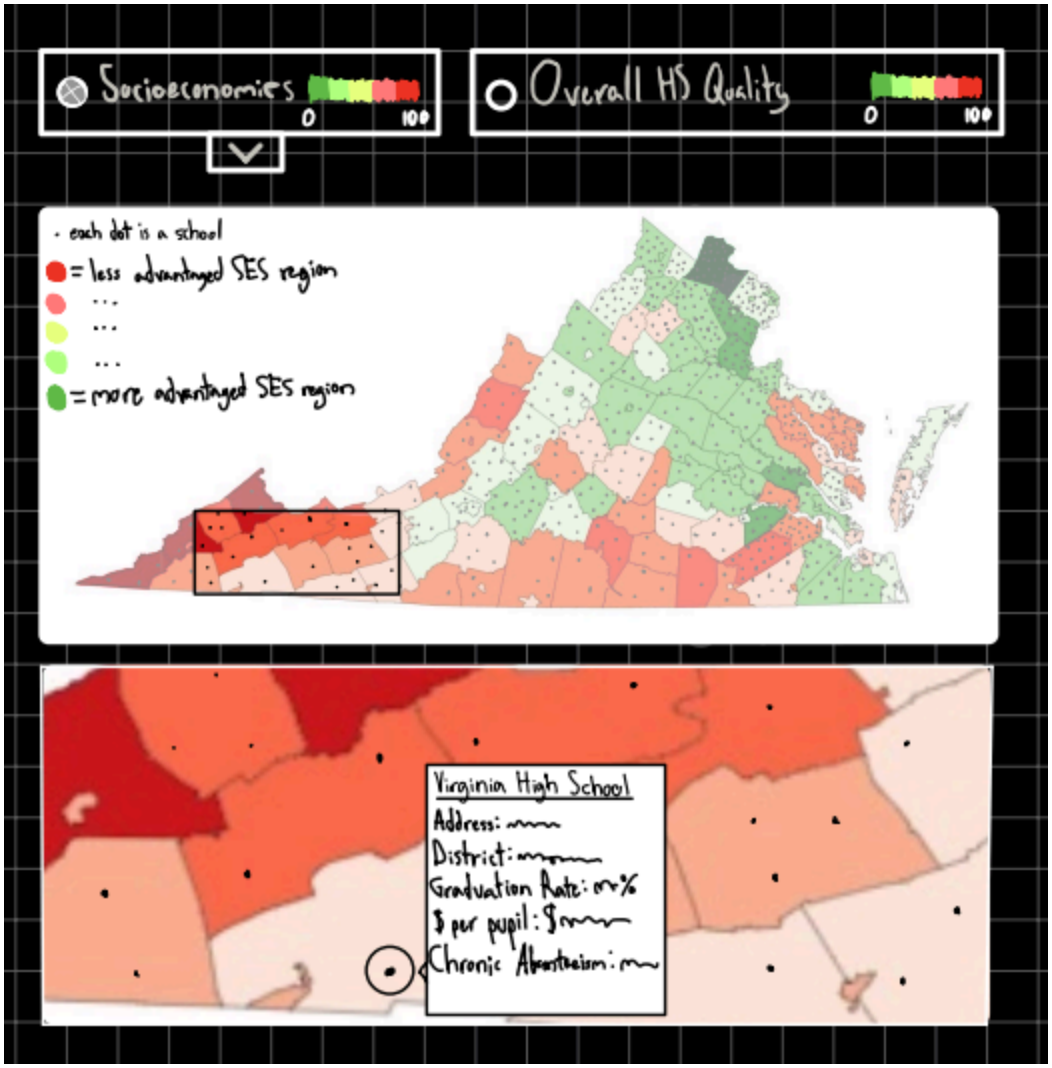
Tableau is unable to adequately visualize two datasets spatially at distinct levels of granularity simultaneously, which is a core requirement of analyzing correlation between our two datasets (socioeconomic deprivation and education quality). In addition, we want to be able to adjust the factors contributing to these index calculations on-the-fly, dynamically updating as different features are toggled on and off. The re-weighting of features to inform a calculated metric and ability to toggle certain indicators on and off is another complication that transcends the capabilities of Tableau.

Our design sketches below seek to address these deficiencies in Tableau reporting and produce a custom JS visualization that better answers our main questions: how does SDI correlate to EQP spatially? Are there regions with higher and lower correlation? Do these regions follow a pattern? What factors contribute the most to these metrics?



- Socioeconomic data populates the chart at the county level
- Selecting ‘socioeconomics’ applies a gradient to counties according to SDI
    - Low SDI -> Red
    - High SDI -> Green
- Small circles on the map represent schools
    - Plotted by longitude and latitude to achieve precise location within county
    - EQP Gradient is off by default to reduce noise

These design choices were chosen to maximize the information and ‘scent’ available to the user. The idea being conveyed should be very clear upon initial engagement with the design (gradients). Multiple marks imply multiple types of data being shown (e.g. county vs school).

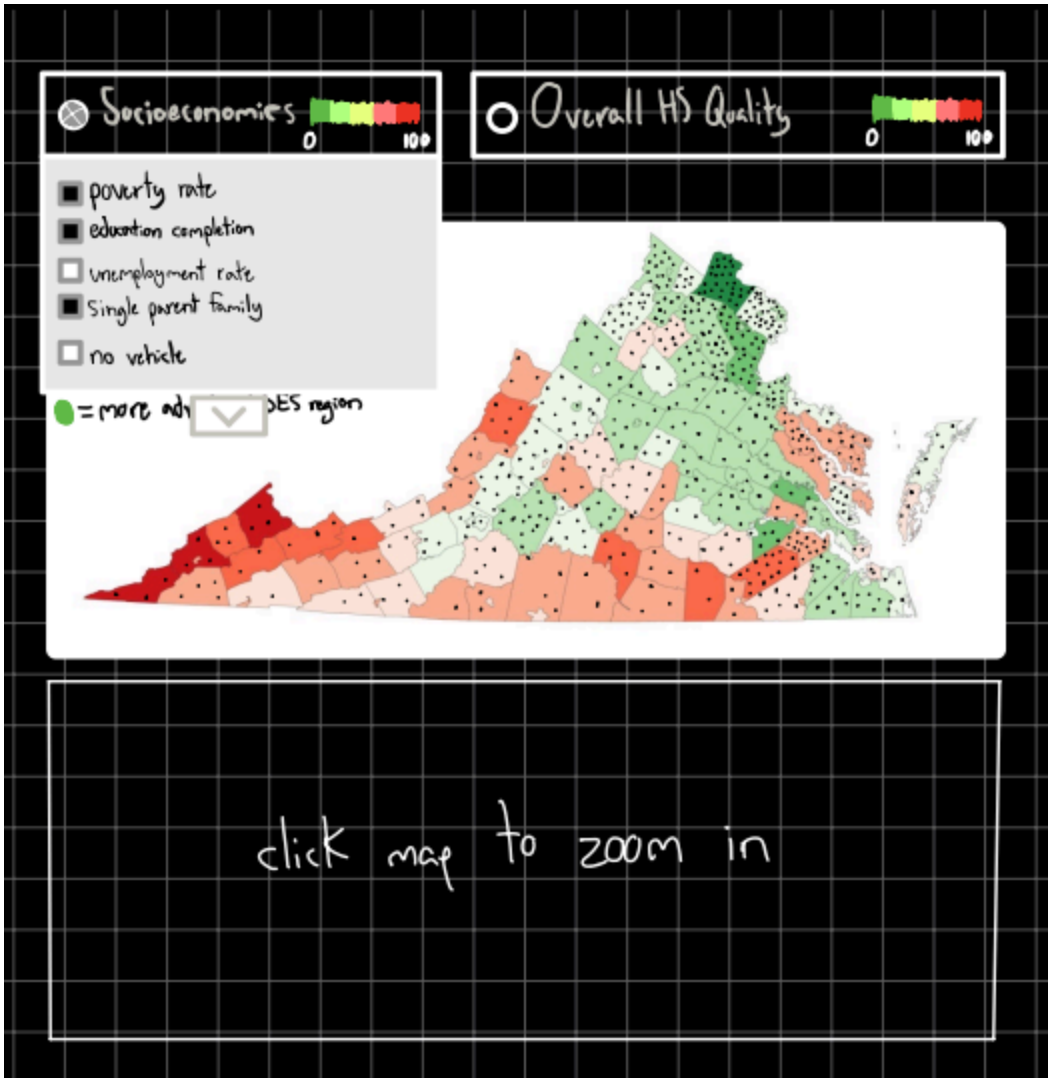


- The map zooms with scroll and double-click
- The school radius has a hitbox which dynamically increases size as zoom ratio increases to maximize visibility and decrease precision needed for mouseover
- Mousing over hitbox reveals school name and its associated education quality markers

Zooming allows the user to analyze the data with increased precision for a certain region. The school markers will be small enough to not clutter the map in denser regions, necessitating a hitbox radius that will ease mouseover interaction. We want to show detailed data per school, which can be revealed by these mouseover interactions.

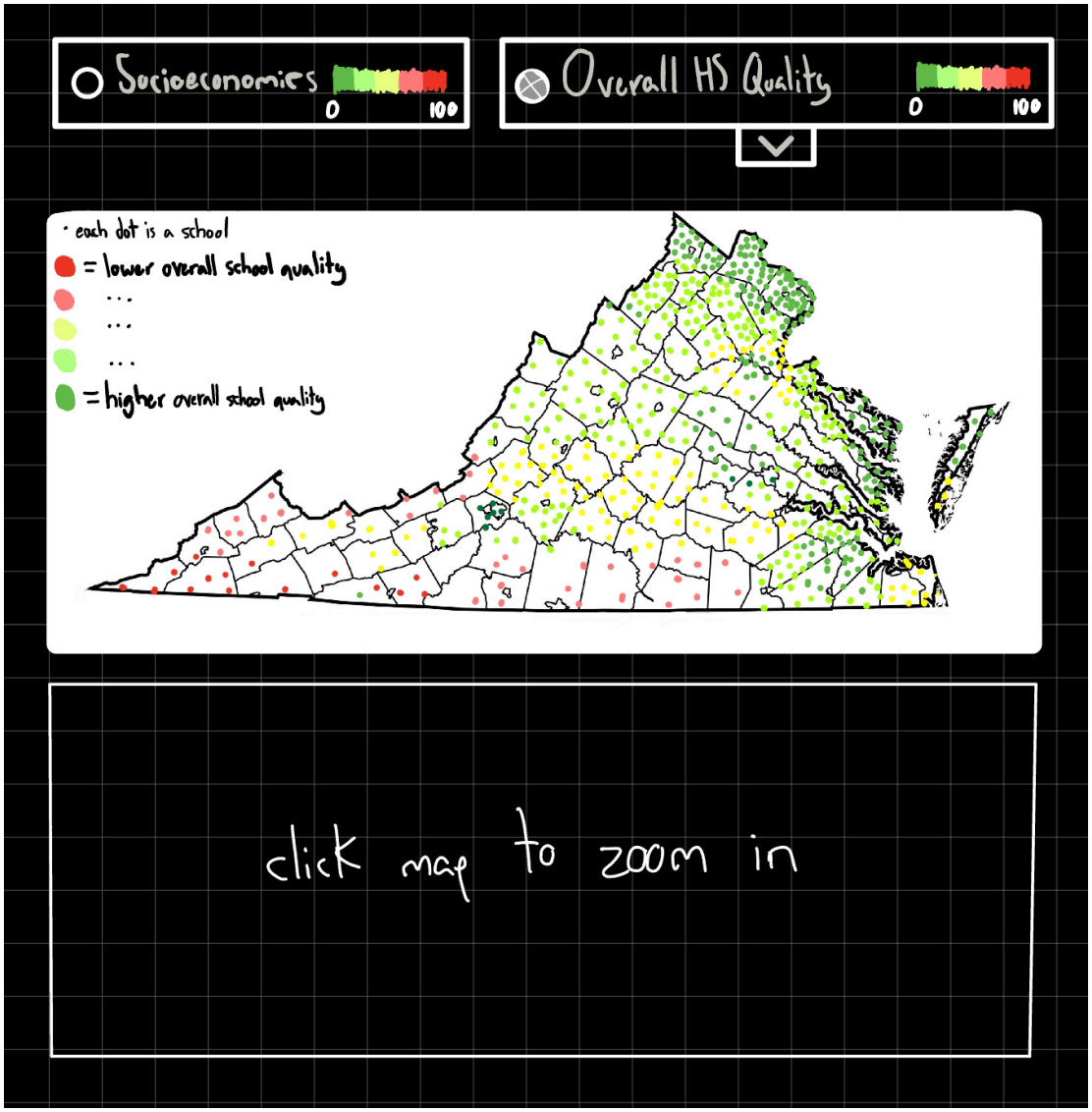
- Filters located below the chart will determine which schools are visible on the map
- These sliders will control relevant educational quality metrics
- Set to minimum, maximum range by default

We want the visualization tool to be an effective tool for multiple end-users. Providing filters allows parents and educators to reveal information about schools and counties within chosen ranges.



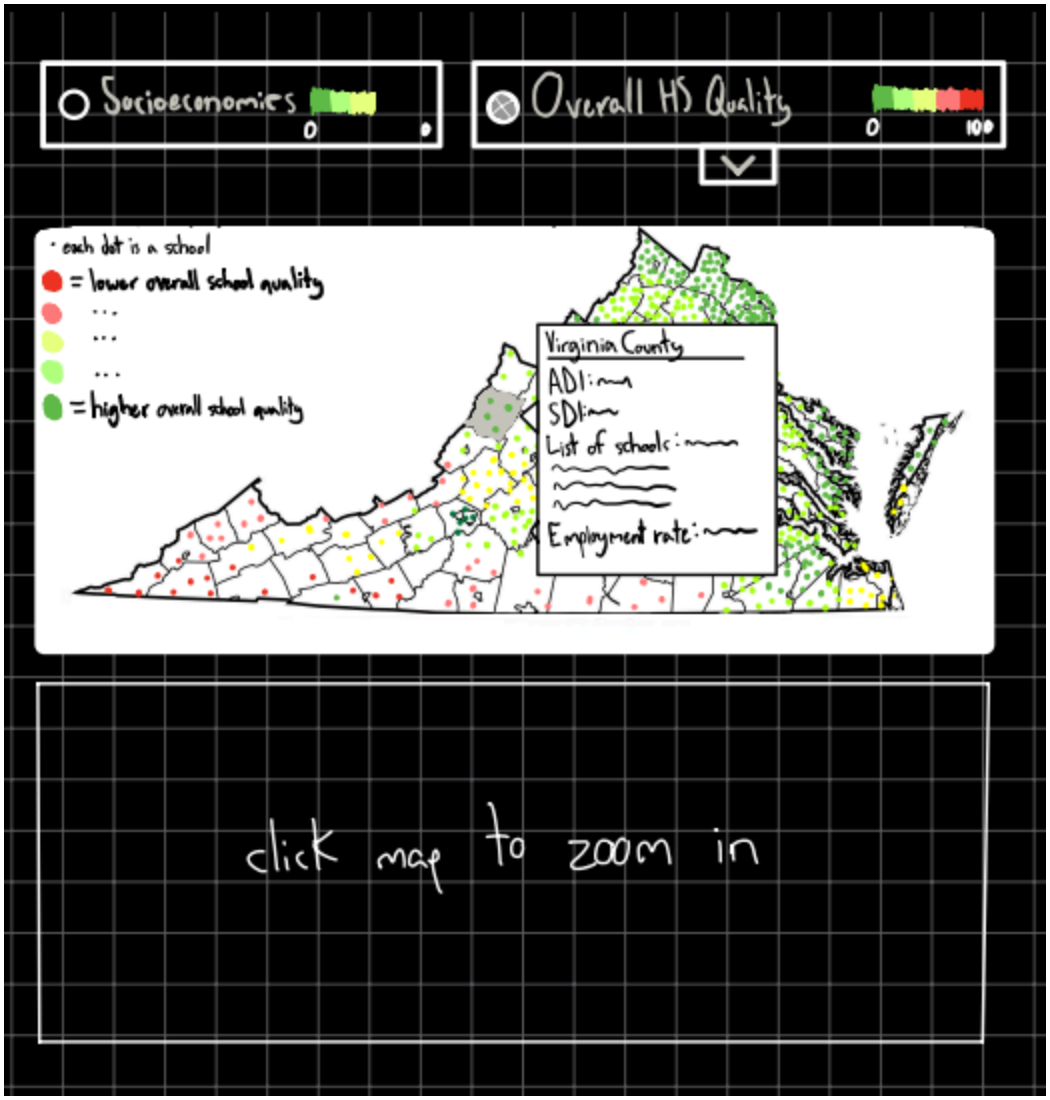
- For each of the series, a dropdown menu will appear revealing the underlying metrics contributing to the gradient
- [1, all] features may be toggled
- These metrics are scaled appropriately in the backend and will linearly combine to form the overall index which informs the color gradient

Our team decided to add the ability to control which features contribute to the overall gradient calculation in order to examine trends based on different features. By default, all features are selected, but as few as one feature may be selected in order to show the distribution of different factors contributing to the overall socioeconomic deprivation of a county. This allows the user to examine trends based on specific relevant metrics.



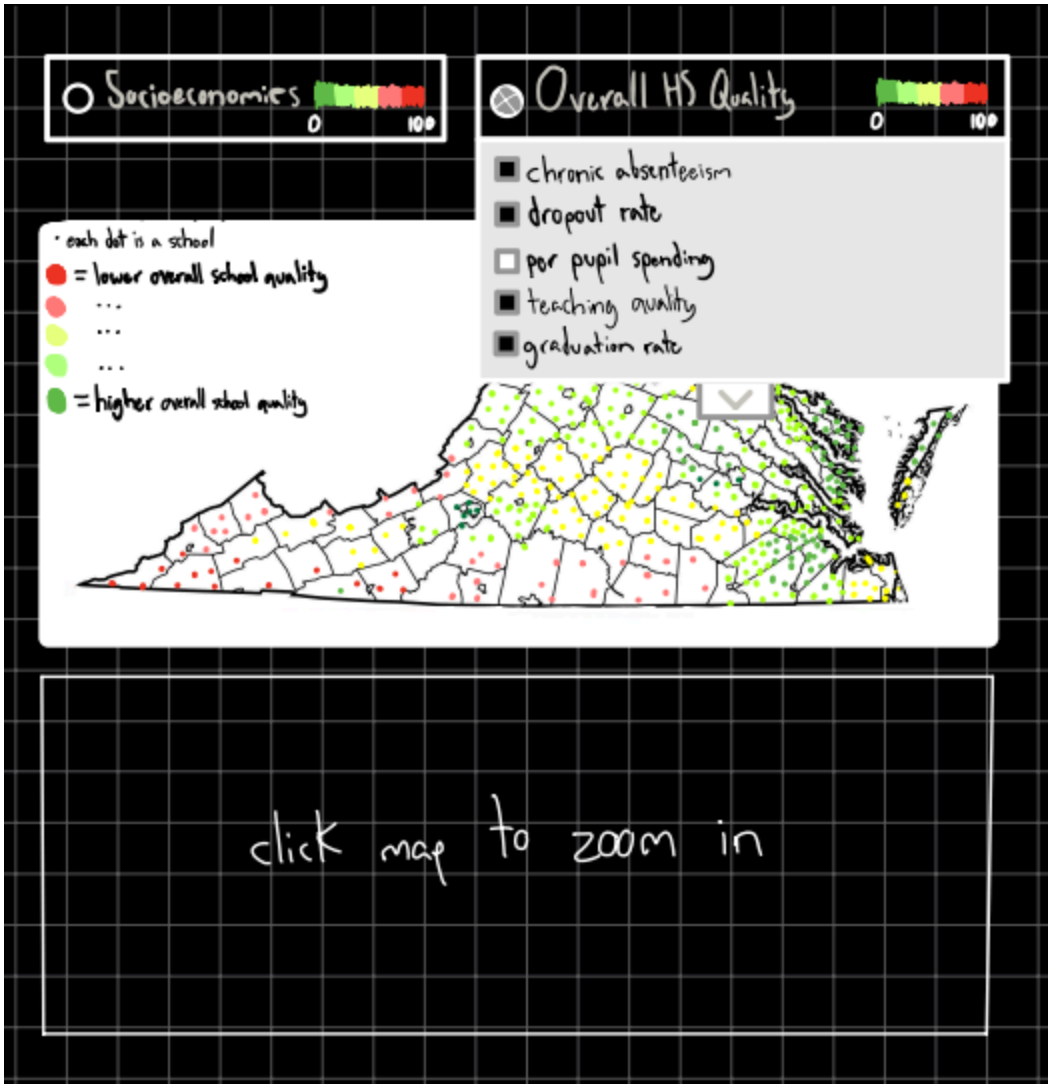
- High school quality data populates the chart by longitude and latitude
- Selecting ‘Overall HS Quality’ applies a gradient to schools according to EQP
    - Low quality -> Red
    - High quality -> Green
- Small circles on the map represent schools
- SDI gradient is toggled off in this example to reduce noise

We decided to represent schools as points on the map in order to signal to the user that these data points are distinct from county data. The data itself is also better represented in its most granular form instead of being aggregated by county (which was our first approach). Applying a gradient to these points based on education quality allows the user to examine trends in educational quality and form insights based on viewing data spatially at school level.



- Mousing over a county reveals county name and relevant SDI metrics
- The schools contained within each county are aggregated and displayed within this view

We want to be able to communicate all of the data simultaneously, but the dataset is simply too large to do this effectively. Our compromise is to reveal aggregated school data (by county) alongside SDI data when mousing over each county. This reveals the name of the county, the list of schools within the county, as well as the full list of toggled fields and their aggregated values.



- For each of the series, a dropdown menu will appear revealing the underlying metrics contributing to the gradient
- [1, all] features may be toggled
- These metrics are scaled appropriately in the backend and will linearly combine to form the overall index which informs the color gradient

Our team decided to add the ability to control which features contribute to the overall gradient calculation in order to examine trends based on different features. By default, all features are selected, but as few as one feature may be selected in order to show the distribution of different factors contributing to the educational quality of each school. This allows the user to examine trends based on specific relevant metrics.

For other considerations, please see: Appendix A

Build

Final Observable Notebook Link: https://observablehq.com/d/0a544c0eb9247338

During Phase 4 of our design project, we implemented several notable modifications and successfully incorporated specific design elements while retaining key features from our original plan. Here is an overview of the changes and elements present in our final design:

Modifications from Phase 4:

1. **No School Color Coding:** Originally, our plan included the color-coding of schools based on their overall quality. However, due to the difficulty of trying to compare school qualities (how would dropout rate compare to per-pupil spending), we chose not to create an index that would result in schools being colored by overall quality. Instead, we opted to focus more on sliders that would filter the schools by their quality.

2. **County Color Coding:** In lieu of school color-coding, we introduced a feature where counties were color-coded using a range of colors, predominantly red and blue. This adjustment was made with consideration for individuals with color blindness, ensuring that index differences were accessible to a broader audience.

3. **Dynamic Zoom-In Feature:** We departed from the initial plan, which involved displaying a separate zoomed-up section. Instead, we integrated a dynamic zoom-in feature. This enhancement reduces the screen space required and offers users the flexibility to zoom in and explore specific map regions seamlessly.

4. **Removed the Original Dropdown Selection:** We eliminated the original dropdown selection feature and replaced them with checkboxes, streamlining the interface for a more user-friendly experience.

Continued Elements from Phase 4:

1. **Base Map:** We retained Virginia's map as the foundational component of our design.

2. **School Representation:** Each school continued to be visually represented on the map as a dot, facilitating easy identification and location.

3. **Popup Tooltip Boxes:** For both counties and schools, we maintained the use of popup tooltip boxes. These tooltips provided crucial information, including the name of the school or county and a unique identification number associated with each dot.

4. **Filter Implementation:** Filters remained an integral part of our design. Initially, we included filters for dropout rate, employment rate, and household income. In the final

version, we expanded the range of filters to encompass dropout rate, graduation rate, chronic absenteeism rate, and the percentage of inexperienced teachers. Additionally, we adjusted the filter range for total per-pupil expenditures, spanning from $7,000 to $88,000.

5. **Additional Socioeconomic Factors:** We introduced checkboxes that allowed users to filter data based on various socioeconomic factors, such as the "percentage of the population living below 100% of the Federal Poverty Level" and more. This enhancement added depth and flexibility to the filtering options.

These modifications and the inclusion of various features were implemented to enhance the overall user experience, improve data accessibility, and cater to a broader range of user preferences and needs.

While we were building our final prototype, we needed a way to recalculate the socioeconomic index based on the socioeconomic factors that are currently selected. Given feature vectors $\bar{x_1}, \bar{x_2}, \cdots, \bar{x_7}$ we need a resulting vector $\hat{y}$ which represents the calculated socioeconomic deprivation for each county. For each linear combination of variables $\bar{x_i}$ representing which feature vectors the user has selected, we need to generate corresponding weights β_i such that $\sum_{i_0}^{i} \beta_i \bar{x_i} = \hat{y} \; \forall i \in [AUT(\{1, 2, \dots, 7\})]$. Therefore, for as many or as few variables with any permutation of features selected, we need weights that will accurately approximate the overall expected socioeconomic deprivation score. We generated weights for each permutation and stored them in a matrix for retrieval. Using this method, we are able to efficiently and accurately recalculate the socioeconomic deprivation given a subset of features in order to maximize the accuracy and utility of the visualization.

Here are some of the core implementation methods we used to create our design:

- Since our data covered Virginia, we retrieved a TopoJSON file that mapped the counties in Virginia from the United States Census Bureau Cartographic Boundaries website.

- Because we wanted a two-tailed slider as opposed to the built-in Observerable one-dot input slider, we installed a package that allowed us to use a range slider for our school filtering options.

- To mass submit our School Filters and Socioeconomic Factors, we used a submittable form template.

- To create a legend for our visualization, we imported D3's Color legend.

- To zoom and pan around the map, we used D3's Zoom.

- To add pop-up boxes on our map when a school or county is clicked, we used tooltips.

- For our overall observable notebook and ease of access when jumping to sections of the notebook, we added a table of contents.

Our visualization is interesting as it allows the user to make comparisons between socioeconomic statuses of counties in Virginia and public school quality factors. It helps lead the user to answer questions concerning what school quality factors are correlated with regional socioeconomic advantage, what socioeconomic factors influence school quality factors, and where advantaged/disadvantaged schools are located.

This visualization is superior to our Tableau dashboard because it is able to accurately compare our two different datasets by giving each school a regional socioeconomic status and further investigate each school's quality of education. Our Tableau dashboard did not help us understand Virginia's socioeconomic characteristics or where schools were located across Virginia due to the lack of geographic information the Tableau dashboard provided. Additionally, Tableau's lack of interactivity did not allow the user to pick-and-choose what factors are important to their use case and investigation of this issue. From our final visualization, we are able to determine and analyze closer pockets of Virginia based on geographical location. The visualization can provide socioeconomic and educational context to all areas in Virginia in an accessible way.

Evaluate

Benchmark Tasks and Target Metrics

For the following tasks, we have identified them as Benchmark Tasks 1-7 (BT1, BT2, …, BT7), with Benchmark Task 7 being open-ended.

Benchmark Task 1

> Identify the county in the highest socioeconomic deprivation percentile. In other words, identify the most socioeconomically deprived county.
>
> Expected result: Either Emporia City or Norton City
> Expected time (seconds): 15
> Task difficulty: Hard
> Expected accuracy: 50%

Benchmark Task 2

> How many schools have a dropout rate of 70% or more?
>
> Expected result: 5
> Expected time (seconds): 45
> Task difficulty: Medium
> Expected accuracy: 80%

Benchmark Task 3

> What is the name of one of those schools that have a dropout rate of 70% or more?
>
> Expected result: {William Flemming, Meadowbrook, John Randolph Tucker, Gar-Field, Justice}
> Expected time (seconds): 5
> Task difficulty: Easy (given last answer is correct)
> Expected accuracy: 70%

Benchmark Task 4

> Which county has the highest percentage of adults with less than 12 years of education?
>
> Expected result: Buchanan County (30%)
> Expected time (seconds): 75
> Task difficulty: Medium
> Expected accuracy: 70%

Benchmark Task 5

Do you notice any regions of Virginia with consistently high deprivation? If so, where?

Expected result: Southern Virginia, Appalachian
Expected time (seconds): 15
Task difficulty: Easy
Expected accuracy: 90%

Benchmark Task 6

Do you notice any regions of Virginia with consistently low deprivation? If so, where?

Expected result: Northern Virginia
Expected time (seconds): 15
Task difficulty: Easy
Expected accuracy: 90%

Benchmark Task 7

Do you see any correlation between these regions and any educational quality measures? Explain.

Expected result: Open-ended result.
Expected time (seconds): 120
Task difficulty: Hard

Participants in Our Study

Our participants consisted of classmates in CS5764 and one of Jackson's friends who works in the data science field. We did not collect demographic information about our participants.

Measures and Observations from Our Study

	BT1	BT2	BT3	BT4	BT5	BT6
Participant 1	emporia	5	meadowbrook high	buchanan	southwest virginia	northern virginia
Participant 2	Powhatan county	5	gar-field high school	buchanan county	buchanan county	nova
Participant 3	emporia city	5	justice	wise county	the parts that no one lives in -	affluent counties - nova, parts of

					south and west	richmond
Participant 4	emporia city	all of them	grayson county high school	doesn't know	southern va	appalachia
Participant 5	emporia city	5	roanoke academy of math	buchanan county	southeast	nova
Participant 6	Emporia City	50-100	rocky run elementary	wise county	southern	northern va
Participant 7	wise county	5	ruffner	lunenburg county	southwest	nova
Participant 8	wise county	5	william fleming	buchanan county	southwest	nova

Observations:

Benchmark task 4
This task had a lower than expected accuracy. The question can be answered by toggling the appropriate checkbox for the socioeconomic indicator and using the reweighted gradient to determine the county extrema. The challenge on this question may be twofold:

- It was observed that some users expected the SDI number itself to be reweighted as the indicators changed which skewed the accuracy of the users under this impression.

- Some users didn't notice that the metric itself was available via details-on-demand.

- The gradient was difficult to distinguish when most of the counties were skewed towards the extreme end of the scale.

Benchmark task 3
This task had a remarkably lower than expected accuracy. This task can be answered by using the filters located below the figure to view the updated schools. Details on demand will reveal the school name. This accuracy was lower than expected for two main reasons:

- Correctness on this task could only be equal to or lesser than the accuracy of the previous task due to their sequential dependence.

- It was observed that some users didn't understand that when the schools become transparent, they are excluded from the constraints of the filter selection.We thought this would be a helpful feature for determining which schools were excluded from the filter selection.

Benchmark task 6

This task was open ended and yielded mixed results. Some users modified the sliders in order to determine threshold values that roughly bisected the data. For example, when adjusting per-pupil spending to the median value, almost all of the schools outside the NOVA area would be filtered out. This led users to the understanding that educational quality metrics were consistently higher in areas of low socioeconomic deprivation. Other users thought less critically about the data and commented on the density of schools in the different regions.

Analysis of Our Study

From assignment handout: Show analysis results and report a final degree-of-success verdict on each task, with hypothesized explanations.

	BT1	BT2	BT3	BT4	BT5	BT6
Participant 1	Correct	Correct	Correct	Correct	Correct	Correct
Participant 2	Incorrect	Correct	Correct	Correct	Correct	Correct
Participant 3	Correct	Correct	Correct	Incorrect	Correct	Correct
Participant 4	Correct	Incorrect	Incorrect	Incorrect	Correct	Correct
Participant 5	Correct	Correct	Incorrect	Correct	Correct	Correct
Participant 6	Correct	Incorrect	Incorrect	Incorrect	Correct	Correct
Participant 7	Incorrect	Correct	Incorrect	Incorrect	Correct	Correct
Participant 8	Incorrect	Correct	Correct	Correct	Correct	Correct
ACCURACY	**62.5%**	**75.0%**	**50.0%**	**50.0%**	**100%**	**100%**
Difference from predicted accuracy	+12.5%	-05.0%	-20.0%	-20.0%	+10.0%	+10.0%

Analysis Results:

Benchmark Task 1 (BT1):

- Degree of Success: 5 out of 8 participants (62.5%) answered correctly.
- Hypothesized Explanations:
    - The task complexity was not in the action required but in the small size of the county, making it easy for users to overlook.

Benchmark Task 2 (BT2):

- Degree of Success: 6 out of 8 participants (75%) answered correctly.
- Hypothesized Explanations:
    - Considered moderately challenging, success depended on whether users utilized the zoom option to count all the schools accurately.

Benchmark Task 3 (BT3):

- Degree of Success: 4 out of 8 participants (50%) answered correctly.
- Hypothesized Explanations:
    - Despite the question not being highly challenging, the lower success rate was due to a misunderstanding. Some users misinterpreted white schools as being part of the category, leading to errors in counting.

Benchmark Task 4 (BT4):

- Degree of Success: 4 out of 8 participants (50%) answered correctly.
- Hypothesized Explanations:
    - The task's moderate difficulty level, combined with subtle color differences among many similar options, led to a lower success rate. Participants may have overlooked details in the hustle.

Benchmark Task 5 (BT5):

- Degree of Success: 8 out of 8 participants (100%) answered correctly.
- Hypothesized Explanations:
    - The task was straightforward, and all participants effectively understood and responded correctly.

Benchmark Task 6 (BT6):

- Degree of Success: 8 out of 8 participants (100%) answered correctly.
- Hypothesized Explanations:
    - Similar to BT5, Task BT6 was straightforward, with all participants demonstrating a clear understanding and effective response.

Final Degree-of-Success Verdict:

- BT1: Low to Moderate Success
- BT2: Moderate Success
- BT3: Low to Moderate Success
- BT4: Low to Moderate Success

- BT5: High Success
- BT6: High Success

Overall Observations:

- Participants generally performed well on tasks BT1, BT5 and BT6.
- Tasks BT2, BT3, and BT4 had a lower success rate, with participants encountering challenges.

Current Problems and Potential Solutions

Problems

- Users had to scroll down to set the filters and hit the submit button to view the changes on the map which led to users taking more time in their analysis as it did not allow side by side comparison. This extra time spent scrolling and clicking submit made it harder for users to efficiently analyze the data.
- Users were not able to distinguish areas based on a single metric. The inability to filter and highlight specific metrics made it challenging for users to spot patterns and trends within a certain measure.
- Users found it difficult to distinguish between filtered schools and non-filtered schools. The map did not clearly differentiate between schools that met the filtered criteria versus those that did not. This made it hard for users to see the schools they had filtered for.
- Users could not effectively analyze regional differences of Virginia with the color hue presented. This hindered users' ability to effectively analyze regional differences.

Design solutions

- Update filter sliders to be dynamic without a submit button and combining it with the heatmap svg. This allows users to see the map update in real-time as they adjust filters, without needing to hit submit. Integrating the heatmap directly makes the changes more visible.
- Introduce gradient for regions which is toggleable and is based on a single metric. Adding a gradient color overlay lets users easily distinguish regions based on the level of a chosen metric, since schools will be shaded along a color range. Making this toggleable gives users control over when to enable gradient versus basic colors.
- Remove white outline around school markers(circles) which are excluded from filter selection in order to further signal that they were not included in the selection.

Eliminating the white rings around unselected schools cleans up the visual clutter and makes it more obvious which schools are filtered out.

- Use a better color hue for heatmap to distinguish the regions better. The previous heatmap colors did not allow for easy differentiation of social deprivation levels indexes across regions. Choosing colors that have sufficient contrast will help users better interpret heatmap patterns.

Final Presentation

Links

Video presentation link:
https://drive.google.com/file/d/1V1uq2e22fhDyJn9i6k0BY3ehZmbtu-wF/view?usp=drive_link
Final Observable link: https://observablehq.com/d/0a544c0eb9247338

Presentation Notes

Final presentation breakdown:

- **Problem we're trying to solve (3 minutes)**
    - There is little-to-no research that heavily investigates socioeconomic status and computing education.
    - Really inspired by this: https://www.neighborhoodatlas.medicine.wisc.edu/mapping
        - The medical field is using socioeconomic information in a really interesting way. Why can't we look at its relation with education.
    - Briefly touch on key takeaways.
- **What we created (3 minutes)**
    - Explains what was created and why it was created with this in mind
    - Why is having a changing socioeconomic status important?
    - Why is having sliders for the schools important?
    - Scenarios to answer questions on what we created:
    - Show important user scenarios and the insights that can be taken away
        - Caring just about the spending per-pupil in regards to financial aspects of socioeconomic status?
        - Seeing where students skip school often in relation to having a car?
        - etc, etc, etc (specific scenarios need to be developed further and have the exact settings for the visualization determined before presenting)
- **Key takeaways, values, limitations of this tool, and future use (6 minutes)**
    - **Key takeaways (4 minutes):**
        - Researchers can better draw conclusions on the relationship between socioeconomic status and school qualities in Virginia.
        - This tool sets out to use socioeconomic indices in a new context.
        - The data shows that students' regional socioeconomic status is tied to school quality factors in certain capacities.
    - **Limitations & Future use (2 minutes):**

      Limitations:
        - Clumping vocational, special education, and regular public schools.
        - Clumping elementary, middle, high schools.
        - (add limitations from user study)

      Future use:
        - Improve selectors feature.
        - Give more written statistics as opposed to the purely visual representation upon selection
        - Give users a better sense of what the school slider quality number range looks like.
        - (Add more suggestions from user study)

Credits

Alexandra:

Phase 1 Contributions:

- Proposed the project based on data related to current research.
- Wrote the description for the project.
- Organized first team meeting where a lot of the work of this section was done.

Phase 2 Contributions:

- Laid out guidelines for domain situation.
- Helped team members navigate datasets.
- Outlined task abstractions and helped brainstormed somekeep questions.
- Listed task domain concepts and terms that others may not be familiar with.

Phase 3 Contributions:

- Limited contributions for this phase.

Phase 4 Contributions:

- Created all hand-drawn sketches with the support and feedback of my team.

Phase 5 Contributions:

- Created main structure for map of Virginia.
- Helped integrated school points onto maps.
- Integrated initial zooming and panning functionalities.
- Created frontend aspects for checkboxes and two-tailed sliders.
- Bugfixes + QA.
- Quality of life improvements such as: adding a legend to the map, sectioning the Observable notebook, adding a table of contents to the notebook, adding descriptions to the notebook that give additional context to the visualization.

Phase 6 Contributions:

- Found solutions for benchmark tasks outlined by Jackson.
- Created tables and accuracy results.

Phase 7 Contributions:

- Created initial presentation notes.
- Opens the presentation with the problem at hand and provides context to issue.
- Presents the conclusion of the presentation in place of Jackson in-class.

Ayush:

Phase 1 Contributions:

Discussed collaboration,and distributions of tasks.

Phase 2 Contributions:

1. Added to task abstraction questions
2. Worked on the data abstraction task
3. Created table with information on each data attribute and dimension of data set
4. Figuring out how to join all the different datasets
5. Figuring out the derived data tasks

Phase 3 Contributions:

1. Did data manipulation tasks:Joined ADI-SDI datasets mapped them with counties

2.Made tableau prototype for the same.
3.Combined all the tableau visualizations into a dashboard.
4.Worked on documentation-Limitations of Tableau.
5.Gave presentation

Phase 4 Contributions:

1.Gave feedback on the sketches

Phase 5 Contributions:

1.Joined all the datasets into one (ADI+SDI+NCES).
2.Loaded the data set on to observable
3.Implemented :Aggregation and filtering of data using checkboxes, filters.
4.Joining datasets for a detailed view when selecting county
5.Implementation of school quality filters(highlighting schools based on the filters)
6.Bug Fixes

Phase 6 Contributions:

1.Assisted in Evaluation running user studies in class.
2.Worked on Measures and Observations from Our Study,Current Problems and Potential Solutions

Phase 7 Contributions:

1.Explaining What we created in our presentation

Jackson:

Phase 1 Contributions:

- Collaborated on roles + responsibilities planning
- Collaborated on communication + collaboration plan
- MOST IMPORTANTLY came up with team name 😎

Phase 2 Contributions:

- Came up with task abstractions - challenging questions and dataset mapping

Phase 3 Contributions:

- Data joining, transformation, mapping, metric re-aliasing, technical support
- gave presentation

Phase 4 Contributions:

- Gave feedback on sketches
- Wrote description for each one

Phase 5 Contributions:

- Wrote helper functions that could be used generally to reference the data
- Created a matrix of feature weights for dynamic re-weighing of SDI linear regression
- Implemented county gradient
- Implemented county labels
- Bug fixes + QA

Phase 6 Contributions:

- Created benchmark tasks and target
- Performed all tests on users
- Analysis of results

Phase 7 Contributions:

- Limitations and next steps section of presentation

Sandhya:

Phase 1 Contributions: Discussed collaboration and distributions of tasks.

Phase 2 Contributions: Worked on the writing of the abstraction and adding points to the project proposal.

Phase 3 Contributions:

Worked on two tableau prototypes that answer the questions asked in the abstraction.

Phase 4 Contributions: Provided feedback on the visualization design

Made points about the setbacks in the design during the group meeting.

Phase 5 Contributions: Helped in building the project. Added the legend to the datapoints and counties.

Found the latitude and longitude of the addresses of the schools using google sheets add on geostat.

Make the datapoints visible on the map of Virginia on the D3 observable.

Phase 6 Contributions:

Helped in the survey and collection of datapoints.

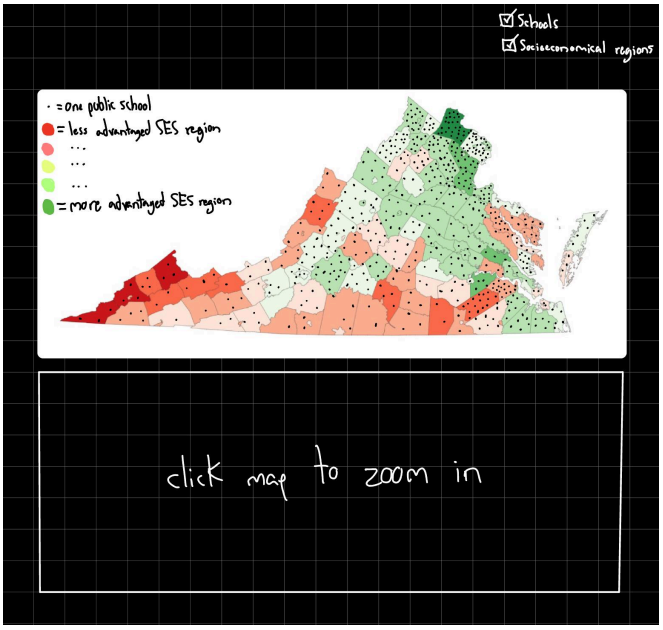
Helped in making an analysis of the datapoints.

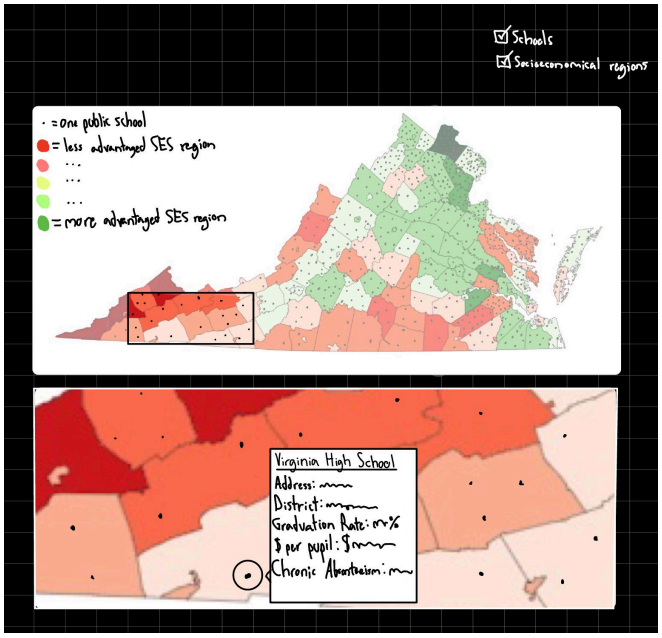
Phase 7 Contributions:

Found important insights in the presentation of the D3 visualization that can be included in the presentation.

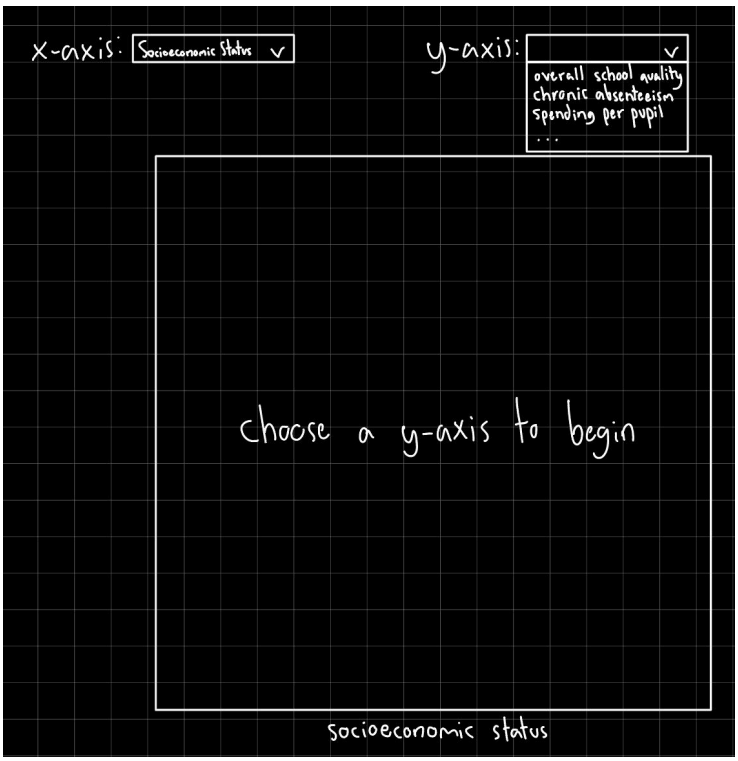
Appendix A: Other Sketches

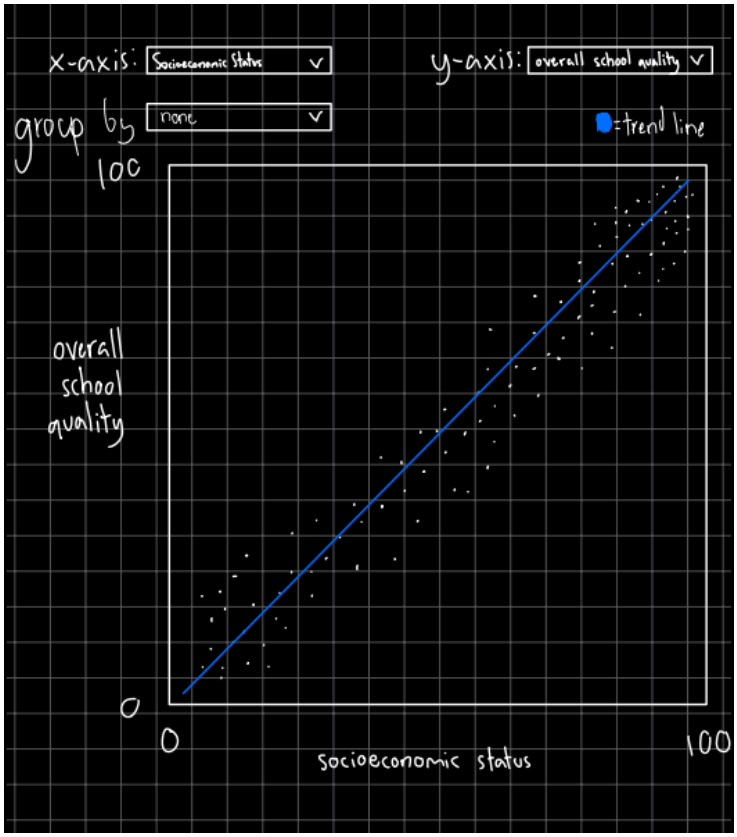
Earlier Sketches of Heatmap Visualization

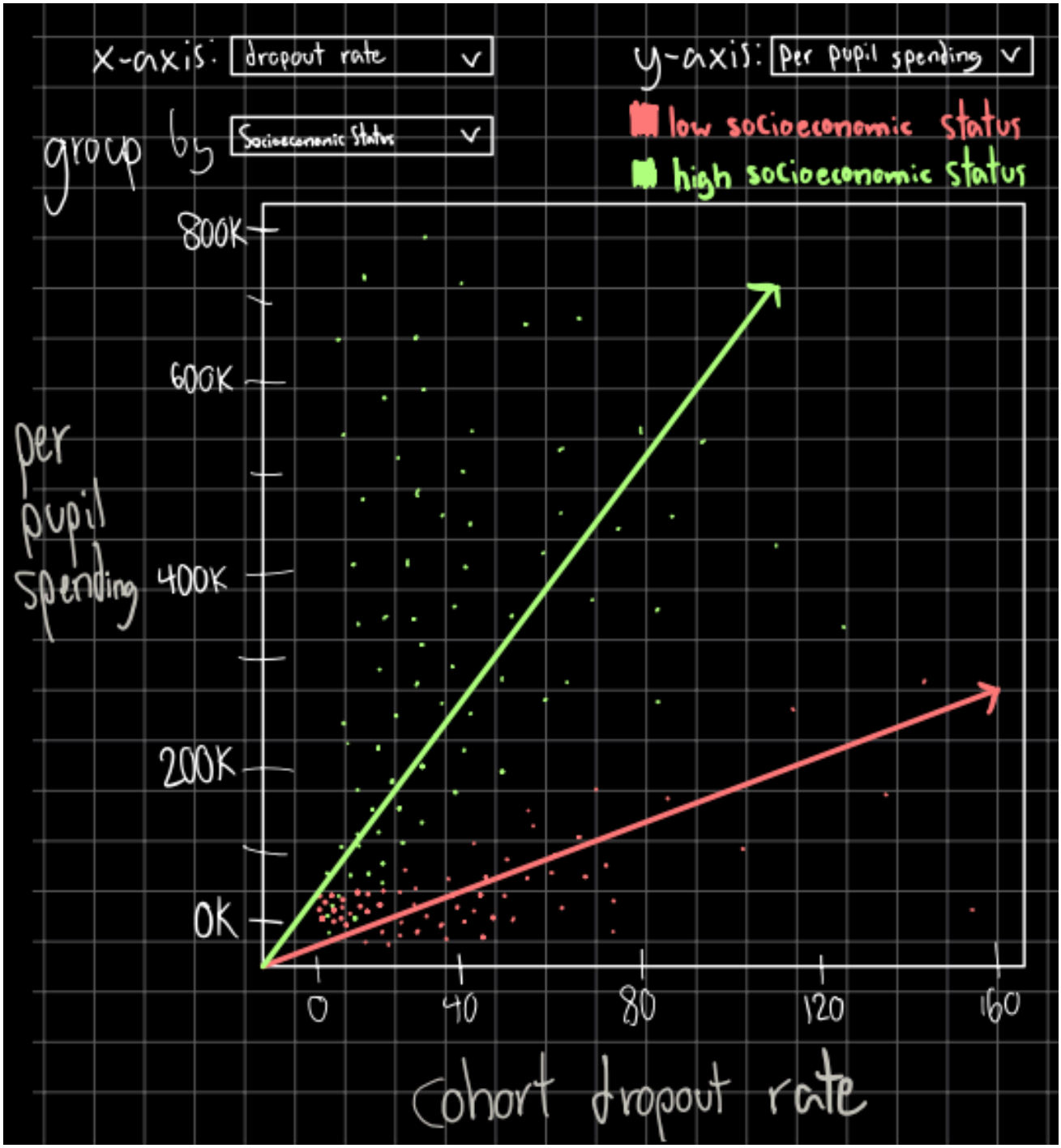




Scatterplot Visualization Brainstorming






x-axis: dropout rate
y-axis: per pupil spending
group by Socioeconomic Status
low socioeconomic status
high socioeconomic status
800K
600K
400K
200K
0K
per pupil spending
0
40
80
120
160
cohort dropout rate