Process Book
Lisa Yao and Morgan Paull
CS171
Project 3

Topic and Motivation

California consistently ranks far below the national average, and often in the lowest third of the country in its K-12 education. We were both educated in California in what we believe to be quality school systems. We created this visualization to explore the regional academic performance in our state and to determine what factors affect academic peformance in California.

We have gathered county-level data on educational outcome in California schools in the form of test scores and dropout rates, as well as a predefined aggregate identifier of educational outcome defined by the state (academic performance index or API). We will correlate these data with other variables such as average income, truancy and discipline rates, property tax values, and teaching staff education levels, and seek to reveal important patterns in the data on educational outcome in California.

Techniques/features and justifications

The county-level map of California will be a choropleth encoding any one of the variables in our dataset, with the variables and coloration divided into quartiles. The coloration of the choropleth was selected to be blue, to allow easy highlighting of selected areas in red for brushing and linking. On mouseover of the counties, datapoints in the other parts of the display highlight, and on selection of data points in other parts of the display, counties on the choropleth highlight.

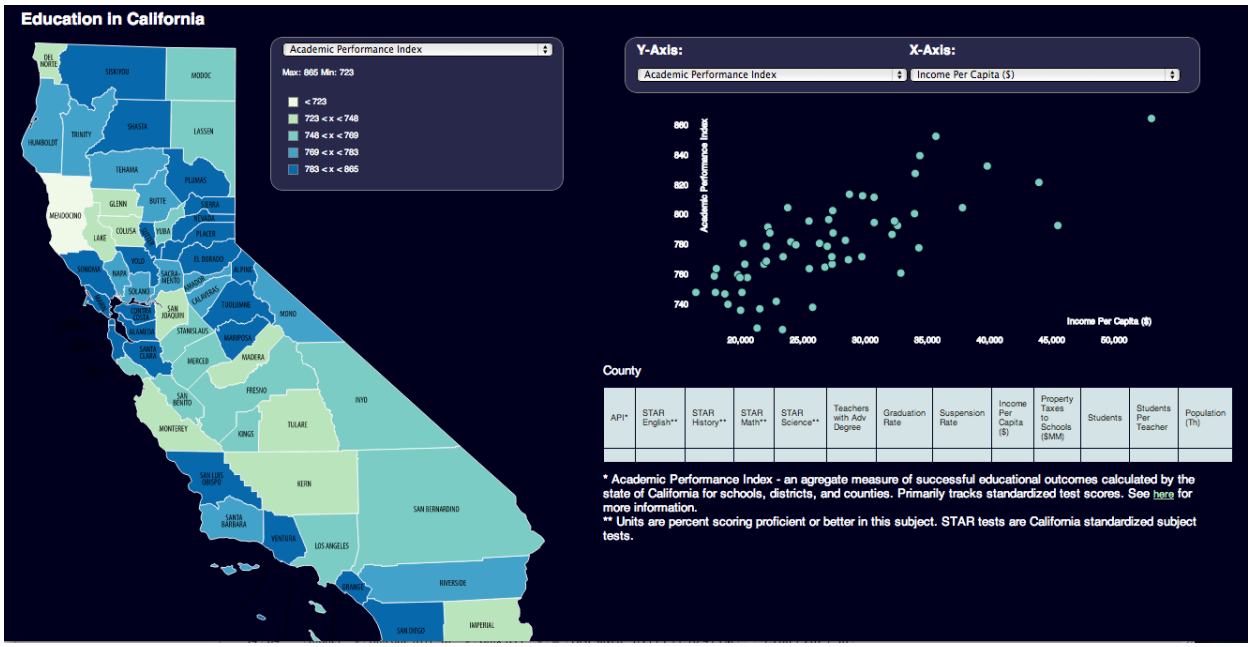
The scatterplot is the central element of data display for our project, allowing easy visual interpretation of the relationship between two variables. Our primary interest is to convey the factors which correlated with successful educational outcomes by county in California, and so the scatterplot as a means to convey correlations is a key part of our message. To allow the user control over the display the variables associated with the X and Y axes of the scatter plot will be mutable, and selectable by the user.

The bar charts allow the user to compare about 4 variables across all of the counties. The magnitude of the variable will be reflected in the bar. The 4 variables can be selected by the user by a checkbox on the right.

All elements of the display will be linked with highlighting and brushing. Users can select a county on the map and see the data associated with that county highlighted on all the other displays of the page, or can select clusters of points using brushing on the scatter plot to see if certain

points geographically cluster on the map. So the user can brush a cluster of dots (representing counties) on the scatterplot, and the corresponding counties will be highlighted in the bar charts and on the map. Conversely, one county may be selected on the map, and that corresponding dot will be highlighted in the scatterplot and the bar charts.

The table will be displayed below the bar charts, and will act as a somewhat more secondary element of the visualization, to allow the user to see the actual numerical values of any information desired. When no elements of the display are selected, the entire dataset of county-level data is displayed. When some subset of the counties are selected, only those rows of the table devoted to those selected counties is displayed, to allow the user to more easily access and examine the specific quantitative values associated with that row.



Design Principles and Justifications

Principles

- Below, we enumerate a number of design principles that have ben spoken of in class and how we have adhered to them in our visualization. There are many so we focus on the most salient ones.

Colors

- For our Choropleth map, we used Color Brewer (5) to find a set of map colors appealing to the eye and amenable to those with color blindness. The colors we chose were "sequential colors" since our data is gradential. We chose this color scheme because it is "calm", doesn't create alarment (we may have used a

warmer, louder color to visualize crime rates), and reflected the emotional neutrality of our subject, education. No color is too loud or contrasting as to make the area seem disproportionate (although light colors tend to do this in general).
    - We chose a dark navy blue background to make the colors pop out
    - We chose the color red for highlighting because it stood out really well. At first we tested complementary colors, but those were actually (surprisingly) less "popping".
    - We also used RGBA in our CSS styling, which allows you to set an alpha value and is very visually appealing. We also used intro.js to adjust transparency of highlighting boxes as well as select menu backgrounds.
- <u>Tufts Integrity Principles</u>
    - Thorough labeling and scales
        - We made sure to use detailed labeling, especially by making a dynamic legend and axes on both the map and the scatter plot. We scaled the scatterplot axes using the D3 function: .scale.linear() as exemplified in the source code. This also helped us maintain graphical integrity. Further, we chose to use a California SVG map with labeled counties to make it easy for users to identify counties.
    - Show data variation, not design variation
        - To do this, we kept colors in the scatter plot a subset of the colors used in the map. We also tried to keep everything organized and used neutral background colors in the visualization and selectmenus.
    - Avoid chart junk
        - We decided not to add tick-marks, created legend from scratch and tried to minimize text in it (fewer distractions)
    - Layer information
        - The choropleth map shows both location as well as a scaled data figure. Dropdowns menus allow you to use the same space to visualize different data

  <u>Graphic Design Principles</u>
    - Contrast
        - We created contrast by using the color red for anything we wanted to highlight and make pop out. We created gray boxes to make select menus pop out
    - Repetition
        - We repeated many colors in the map, scatterplot, and lower text and created a table that is dynamically updating. This makes our data less distracting to users.
    - Alignment
        - Scatterplot and table are on the same side because both change

based on hovering over a county in the map. You can hover on the left and view on the right.

Data Collection and Cleaning

The data used for this project was obtained from the California Department of Education website (4) using the python web scraper included in our submission, education_scraper.py. One exception to this was the value for the total number of teachers in one county, Plumas County. We observed that the student-teacher ratio was an extreme outlier with 17 teachers for 2,400 students. We speculated that this was a result of a reporting or record keeping error. We contacted the Plumas County school district human resources officer Kimberly Retallack to confirm if our number was accurate. She corrected the information we received from the California Department of Education to 140. This was the only outlier removed from our data, and the only one obviously falling outside a different range than the rest of the data. We took a risk in replacing this value. We risked most of all that some other county also had reporting errors which would have made them stand out when they currently do not, for example. But the fact remains that a student-teacher ratio of 1:141 is highly improbable (and in fact was demonstrated to be false upon further inspection. However we acknowledge that this demonstrates, however troubling, that some of the other data from the California Department of Education website may also be erroneous and this might negatively affect the validity of our findings.

In addition to scraping the California Department of Education website (4), our economic data for each county (income tax, income tax levied to schools, income per capita, and other metrics, not all of which were used) came from Wikipedia and the Bureau of Economics in California (4).

Storytelling

The main message we sought to convey, what was that main driver of educational success, or at minimum the strongest correlate in our dataset, is per capita income of the county. A closely related and non-independent factor, total income levied to schools per student, also correlated strongly with academic performance index or API.

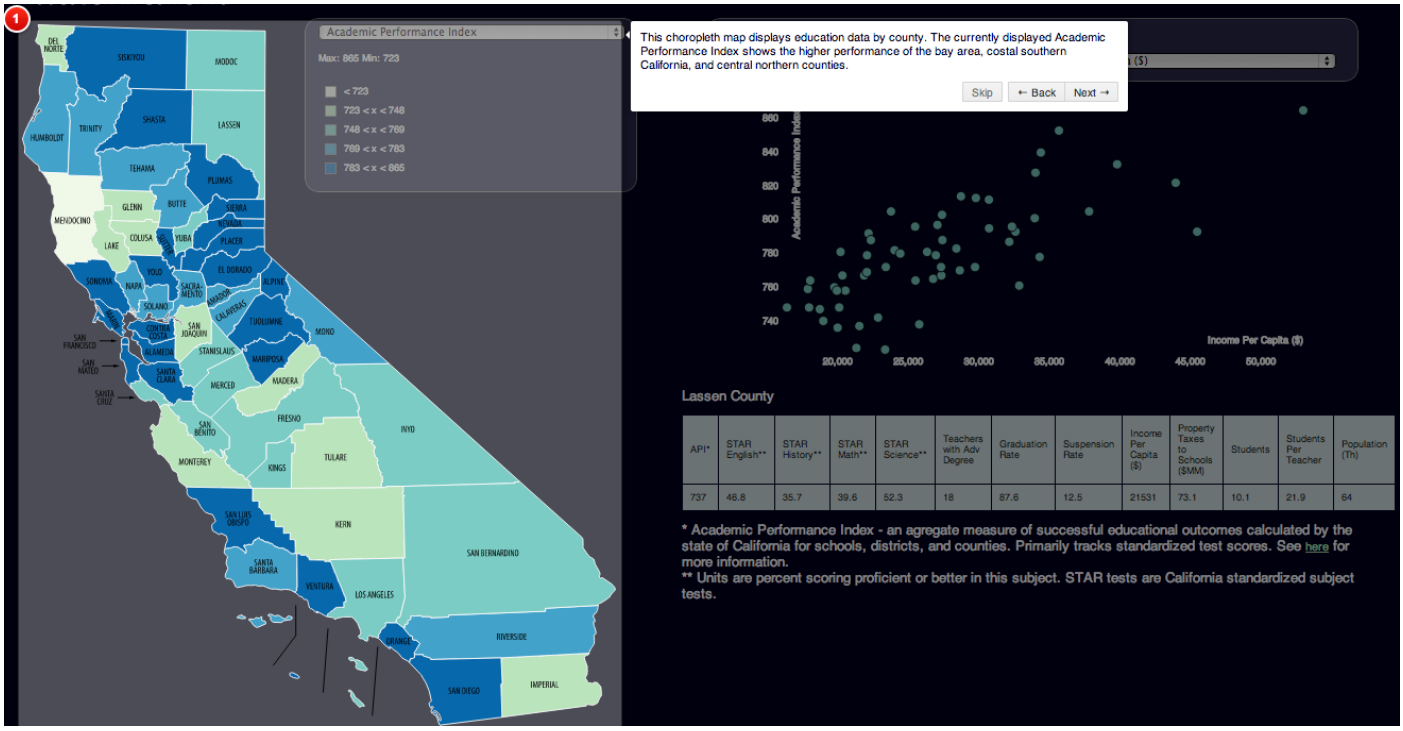
In order to put this on immediate display, the scatterplot is set to default to map API on the y axis and income per capita on the x axis. This is highlighted in the first intro.js narration through the page, the one that runs immediately upon loading the page.

A secondary message we hoped to convey was that, at least to us counterintuitively, essentially nothing else correlated with academic outcome in California at the county level. Of particular interest, based on common assumptions made about educational outcomes, is that neither student-teacher ratio nor the number of teachers with advanced degrees correlated with academic success for students on average.

The final insight is gleaned from the geographic data - high performance in academics clusters geographically. Three regions - the San Francisco Bay area, the central regions east of Sacramento, and coastal southern California account for the lion's share of high-performing academic counties. In agreement with our first insight, these regions also account for ALL of the highest per-capita income counties. A further insight, derivative of this, is that those counties which manage to achieve high performance academically and yet do not come from one of the clustered high-income regions are mostly more northern, alpine regions of California (and usually still have relatively high-to-moderate incomes).

The choropleth map is set to default to display API to highlight this geographic clustering of academic outcomes, and this clustering is highlighted in the intro.js walkthrough for this insight.

For an interactive summary of our findings and policy conclusions, please refer to the bottom of our visualization!



Storytelling with intro.js

In order to facilitate users reaching our same insights, as well as setting them up to explore more on their own, we used a series of intro.js walkthroughs to highlight the views and points of data that we found interesting. These views and walkthroughs are linked at the bottom of the page in the text explaining each of our insights, and always contain a suggestion for the user to explore or try other variables themselves. Furthermore, we used an intro.js walkthrough at the

first point of loading on the page to highlight features and explain what the visualization can do as the user first arrives.

Day-by-day Process

4/16/13 - Meeting with Chelsea

We met with Chelsea to discuss our past project and plans for moving forward with our new project. We first explored the possibility of moving forward with our existing force-directed network project and adapting it to data scraped from PubMed. However, we decided that we could expand both the scope and storytelling aspects of our project if we pursued a new dataset and visualization techniques. Thus, we decided to start over with our project and visualize data on education in California by county.

Chelsea provided us with examples of good visualizations with brushing and linking, storytelling, and complexity.

Our main focus for Project 3 is to integrate *brushing and linking*, add d*epth and complexity*, improve *storytelling*, and acquire *more data*.

4/17/13 - Scrape, sketch, brainstorm

We first decided on what data we would collect (listed above). Then we proceeded to collect all of our data from our onlines sources. We used a scraper to obtain data from the California Department of Education Website. We directly downloaded economic data such as income per capita and property tax rates from the Board of Equalization website and Wikipedia.

We also sketched our proposed visualization (detailed above) and searched for potential source codes, such as a d3 map of California.

4/25/13 - Build scatterplot, scrape, clean data

We finished scraping data from the California Department of Education Website, cleaned up and reorganized the data, and imported it as a JSON object. Simultaneously we implemented a rough scatterplot feature and began to debug it. We continued looking for a good svg map of California with the county names on the image.

4/27/13 - Debug scatterplot, implement chloropleth map

We finished debugging the scatterplot, and implemented in a separate file from the scatterplot a chloropleth map of California, with the county names displayed and colorations based on quartiles of each available dataset. Problems still arising due to mixture of path and polyline objects in the California map svg, causing the data for certain counties to be mapped onto the region for a different county.

4/28/13 - Debug chloropleth, combine scatterplot and chloropleth, implement linking and partially implement brushing

We debugged the data to county mapping problem, and implemented linking from the map to the scatterplot. We also implemented brushing on the scatterplot but have not yet successfully implemented linking off of the brushing. In addition, the map and the scatterplot were combined into a single layout, and issues with having two svgs on the same page (issues that were encountered in our project II as well) were debugged with considerable effort and help.

5/2/13 - Debug brushing and linking, clean up entire display, implement intro.js and add text for storytelling

We debugged the brushing features, and cleaned up the display of the entire page, changing the style and layout. We added intro.js features to facilitate explaining the layout and controls of the page, and to aid in our storytelling. Furthermore, we added a section of text below the visualization explaining our data and exploring our results and insights. Finally, we completed the explanatory video and organized our project for submission.

Bibliography

References

1) Scatterplot:
   - Mike Bostock, May 2013, “Scatterplot”, April 2013 accessed, http://bl.ocks.org/mbostock/3887118)

2) SVG Map:
   - Wikipedia, June 2010, “California County Map”, April 2013 accessed, http://commons.wikimedia.org/wiki/File:California_county_map_(labeled).svg

3) Brushing:
   - Author NA, 2013, “Brushing”, April 2013 accessed, http://static.cybercommons.org/js/d3/examples/brush/brush.html

4) Data:
   - Department of Education, accessed April 2013, http://dq.cde.ca.gov/dataquest
   - Wikipedia, accessed April 2013,
   - http://en.wikipedia.org/wiki/California_locations_by_per_capita_income
   - Bureau of Economics, accessed April 2013, http://www.boe.ca.gov/annual/pdf/2011/2010-11_statistical_appendix.pdf

5) Color Brewer:
   - Color Brewer, http://colorbrewer2.org/

Initial Plan:

Data + Sources

We obtained data for each of the 58 counties in California for:

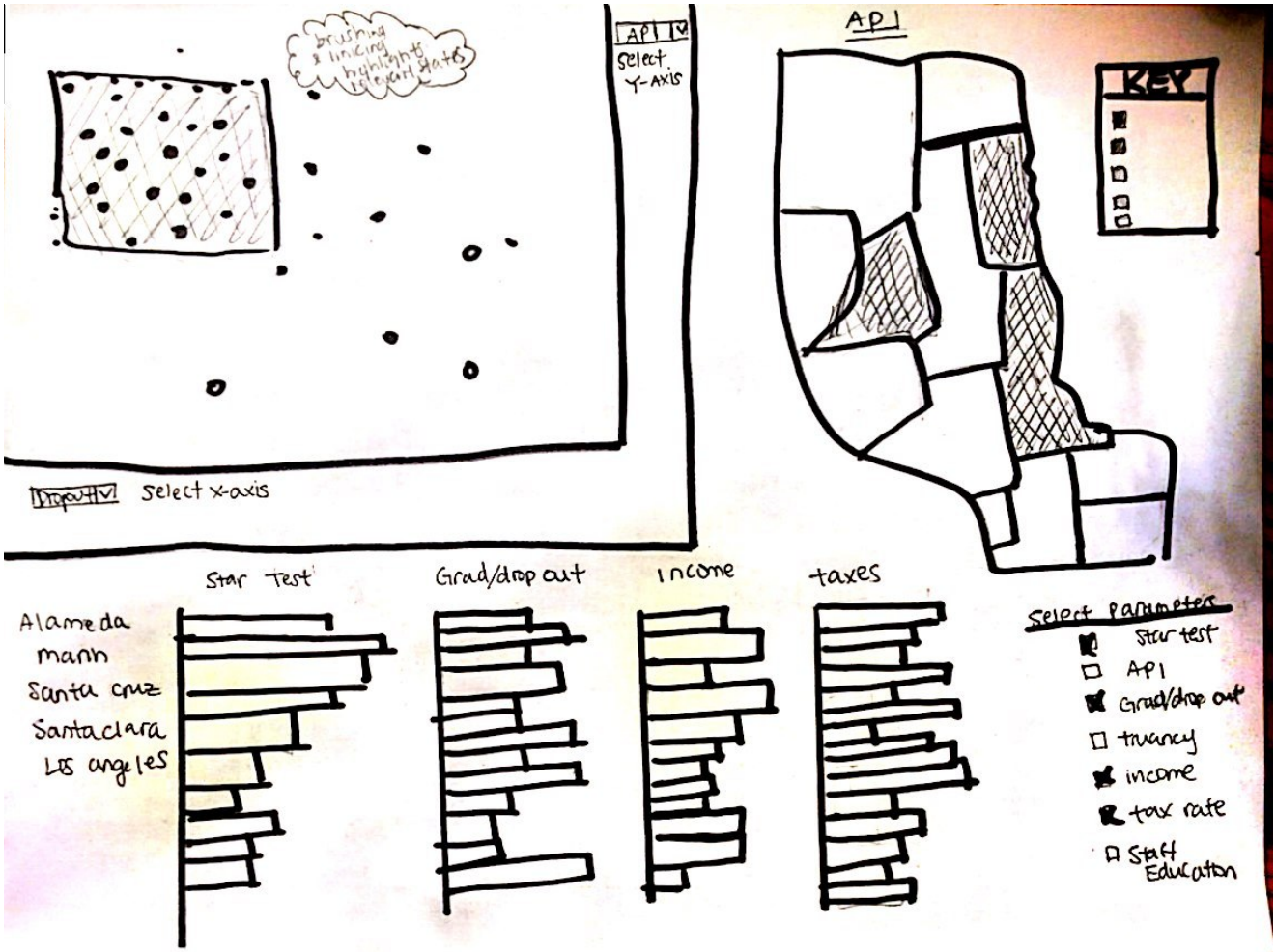
- STAR test scores

- Academic Performance Index
- Graduation and Drop-out rate
- Truancy rates
- Income per capita
- Property tax rates (directly funds education)
- Staff education level
- Number of students in each county

Visualization Plans

We plan to visualize this data on a variety of different dimensions (more details below the sketch).

- Scatterplot with brushing and linking to the map
    - Highlight relevant states
- Choropleth map showing API in gradient color
- Set of about 4 bar graphs that can be used to compare about 4 parameters across all counties
    - Checkbox menu to select parameter
- Chart to show ALL numbers for a particular county
    - Select county using a drop-down menu



County	Star test	API	Grad rate	truancy	income	tax rate	staff
Alameda	1500	304	92%	58%	58k	10%	14%

Introduction

Purpose:

Defn of variables

Timeline:

Week of April 20th

- Solidify data set and clean
- Explore data in ManyEyes

Week of April 27th

- Implement basic features
- Add brushing and linking
- Refine formatting and visual aesthetics
    - Coloring, alignment

Week of May 3rd

- Resolve any bugs
- Add additional features
- Finish process book
- Ask for feedback from test users