College Data Analysis

Team Members:
Christopher Garrett u0933002 christopher.d.garrett.9@gmail.com
Sierra Allred u0740878 sierradorten@gmail.com
Christopher Golling u0708780 chris.golling@verizon.net

Project Repository: https://github.com/SDAllred/dataviscourse-pr-collegedata.git

Background and Motivation

The data we are using for our project is newly released and has not been explored thoroughly. We thought it would be interesting to see if we can find important trends in the data that could be useful to students searching for prospectives schools in the Unites States.

Project Objectives

With this project, we would like be able to give students the opportunity to compare and contrast prospective universities. We would like to do cost analysis on each school, and see how the tuition, average total cost, average debt, and earning potential affect each other. We want to see what trends our visual analysis can show us that will be useful to undergraduates looking for a cost effective school. We would also like to give students the ability to pinpoint schools by letting them search for which schools commonly accept their standardized test scores or which schools are in their price range. The data provides us with information from many different years, so we would like to see how prices and debt have changed for individual schools and across the nation in the past 10 years.

Data

We are using the data provided by College Scorecard under the U.S. Department of Education.

https://collegescorecard.ed.gov/data/

They have easily available downloads for their extensive data collection.

Data Processing

The data we have chosen is very neatly packed in well documented csv files. There will not be very much data cleanup, unless there are missing data points for some schools or years, which is expected. Because there is a large amount of data, we will need to spend plenty of time reading through the provided data dictionary to make sure that we do not miss potentially interesting records to include in our visualization.

Many fields, such as the price of going to a certain university, are broken down into family income brackets, so if we want to use an average of that data we will need to figure out a way to wrangle that data so it does not slow down our visualization. We may need to manually cut down the csv files to get rid of fields that are irrelevent to our project, so that we do not take up excessive amounts of space in our repository.

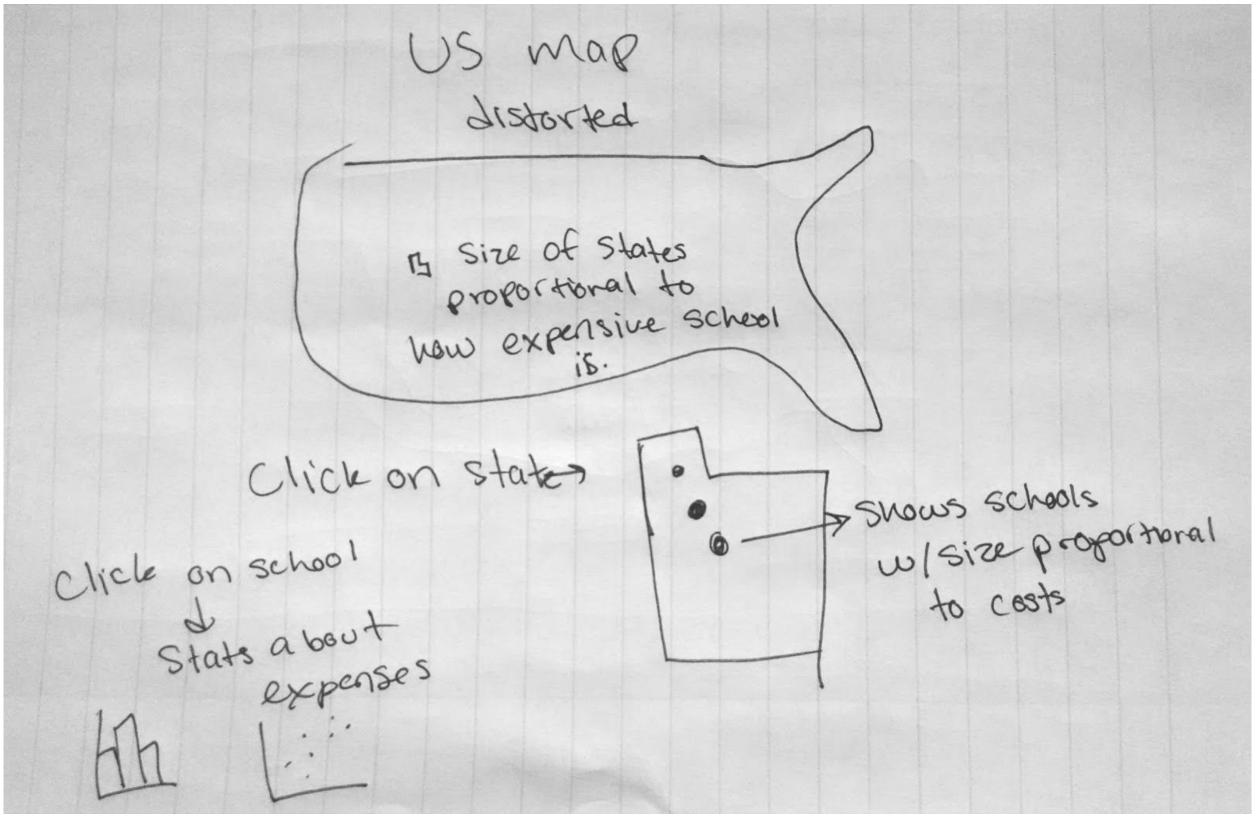
We plan on using basic information about each school, including location, URLs for the Institution's Homepage, whether it is a main versus a branch campus, if it is public, private nonprofit, or private for-profit, if the school has a specialized mission or religious affiliation, if there is a distance-education-only indicator, and possibly school revenue and expenditures if that dataset is complete.

For searching purposes, we plan on using student population and acceptance rates as well as the average ACT and SAT scores of each school, including the 25th and 75th percentile scores. We do not plan on doing this for each individual section of the test unless we find later that it would add some benefit.

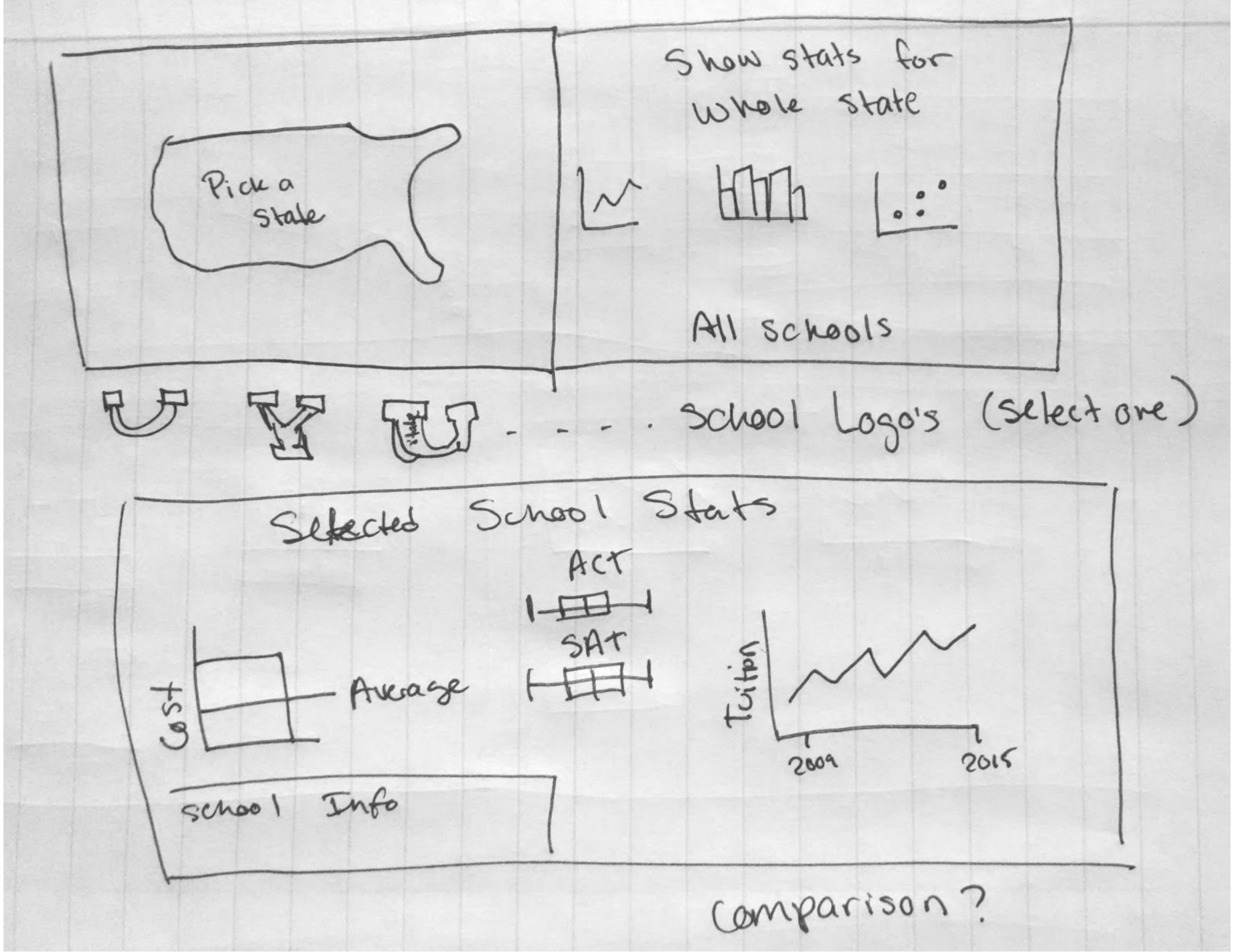
For data analysis purposes we will be using data on cost, financial aid, and projected earnings. In the cost category we will be using the average cost of attendance, tuition and fees, and possibly average net price by income level. For financial aid we will be using percent of undergraduates receiving federal loans, cumulative median debt disaggregated by student subgroups, and the typical monthly loan payments of graduates. Finally, we would like to use the average and median earnings of college graduates disaggregated by student subgroups, as well as the share of former students earning over $25,000 (typical income of a high school graduate).

Using this data, we expect to see strong correlation between the cost, financial aid, and cumulative median debt of each school. We are interested to see what the correlation is between potential earnings of students and the average cost and debt for those students. If this is significant it would make a big impact on what schools are considered the most cost effective.

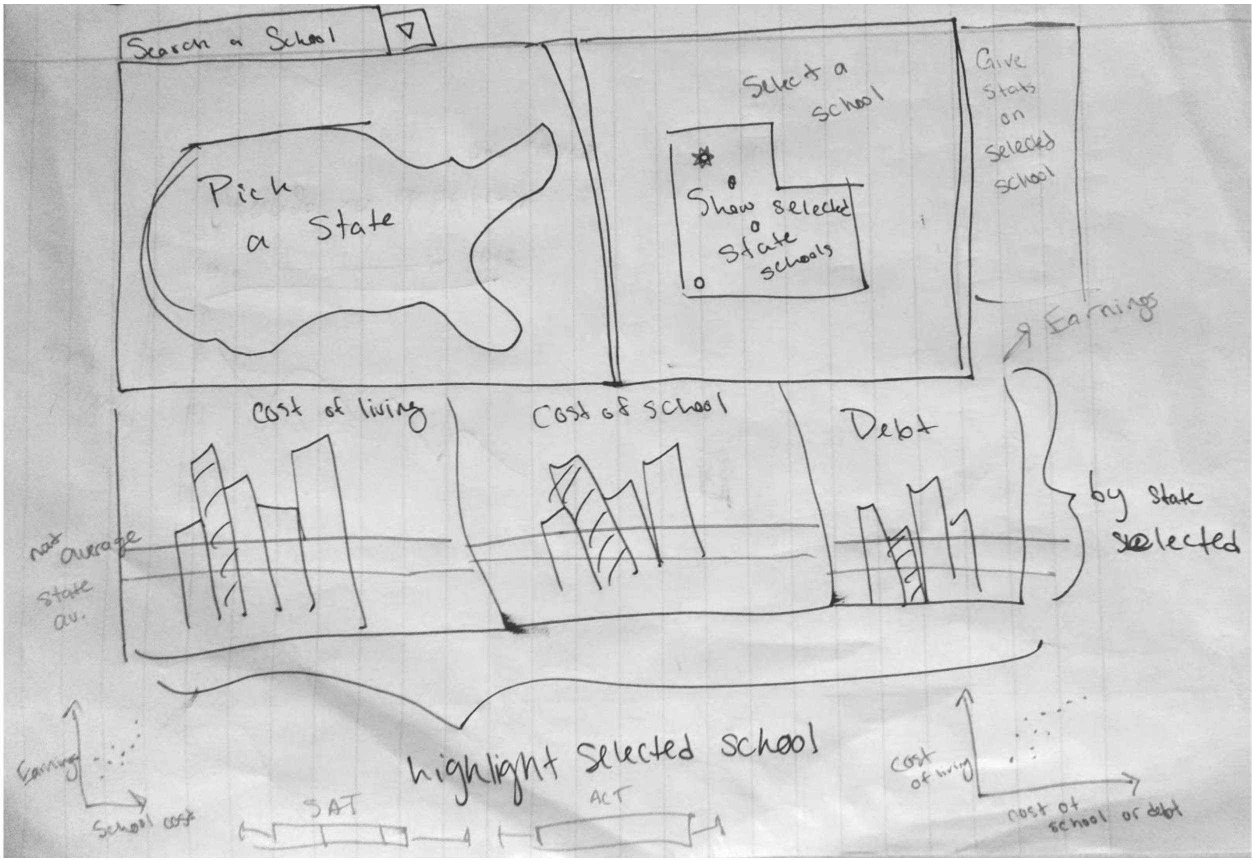
Visualization Design



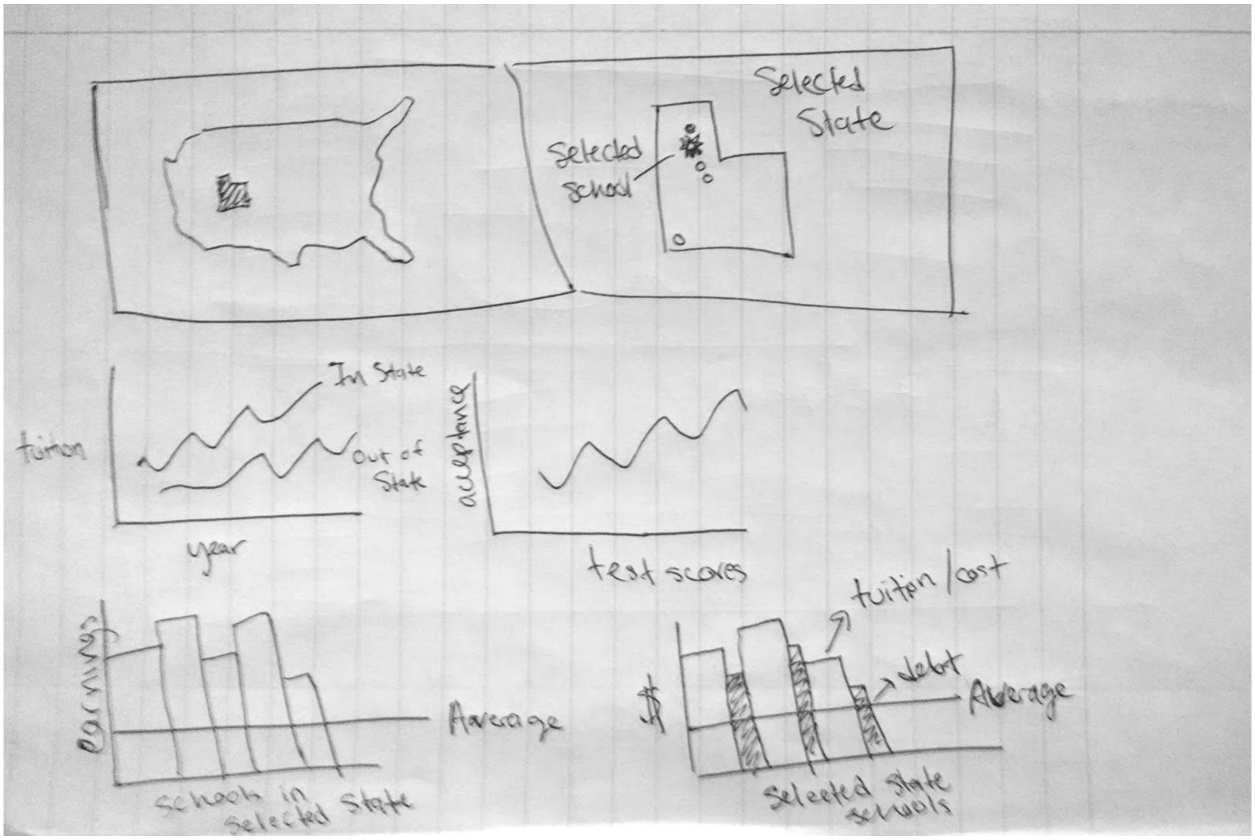
This first sketch was one of our original ideas. Due to the nature of our data we knew that the best way we could organize all of the schools in the US was by using a map, because there are not real links between schools in our data. This design uses the average cost of all the schools in each state to show which states are the most expensive to attend school in. It does this by distorting the map so that the area of each state is proportional to how expensive it is to go to a University there. If you select a state it is shown in a different window, and all of the schools in that state are represented by icons that are also proportional to the average cost of attending that school. If you select a school you are shown bar graphs, scatter plots, and histograms detailing individual information about that school.


Pick a State
Show stats for whole state
All schools
. . . . school Logo's (select one)
~~Selected~~ School Stats
ACT
SAT
Cost
Average
Tuition
2009
2015
school Info
Comparison?

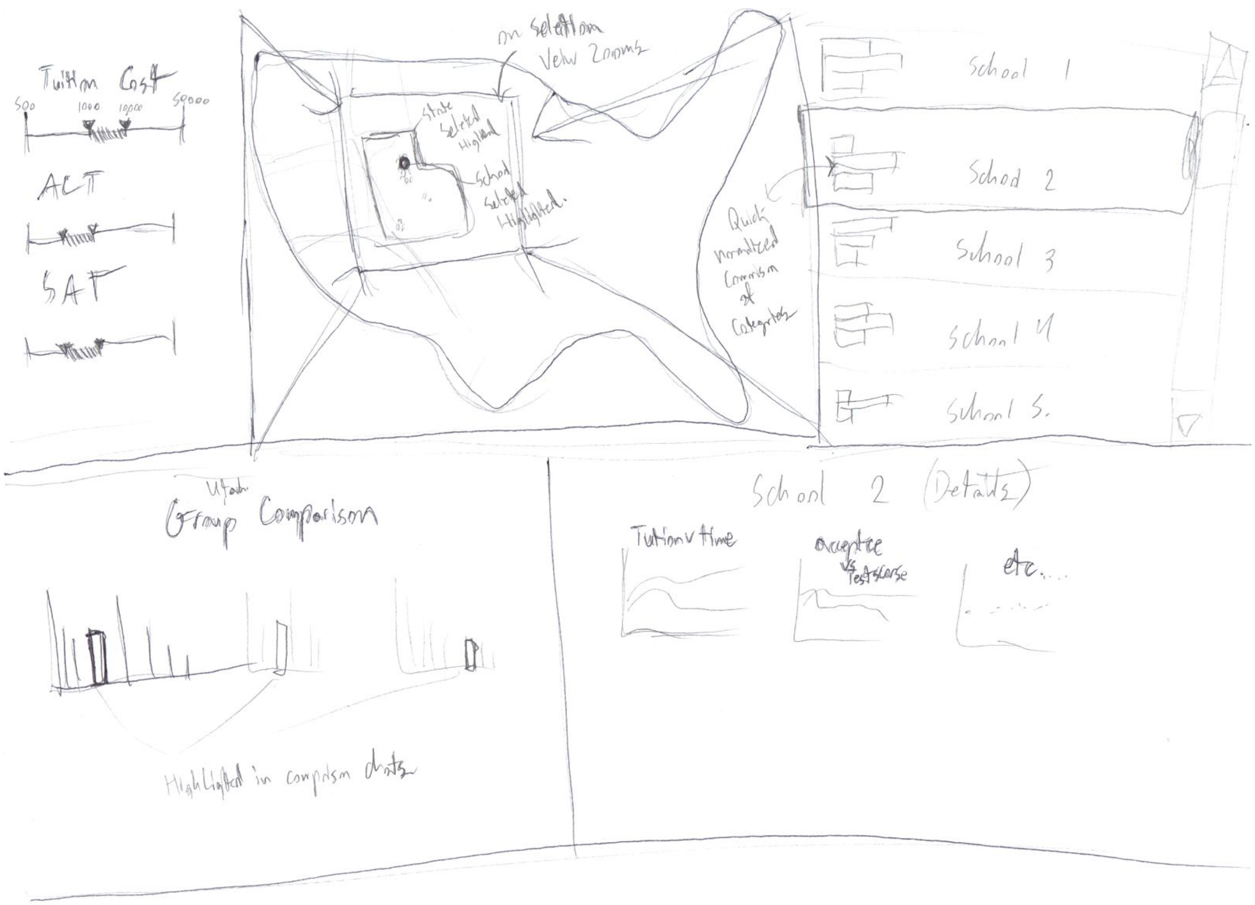
This design still uses a map like the first one, but does not use distortion and area to show the data. Instead, when you select a state you see statistics next to that window about the education in that state as a whole. Underneath is listed the logos of all the schools in that state, and if you click on each logo you will see statistics and graphs about that individual school, as well as information like the school website and address.



Our third design mixes aspects from the first two, and gets rid of some aspects. When you select a state on the map, you are shown that state in an adjacent window, with all of the schools in that state mapped onto it. Underneath you will see a comparison of all of that state's schools by debt, cost, earnings, and other factors. If you click on a school, the bar in each chart representing that school is highlighted. The lines through the bar charts represent the state and national averages for that category. There is also a search bar at the top that will direct to to the state of the searched school.



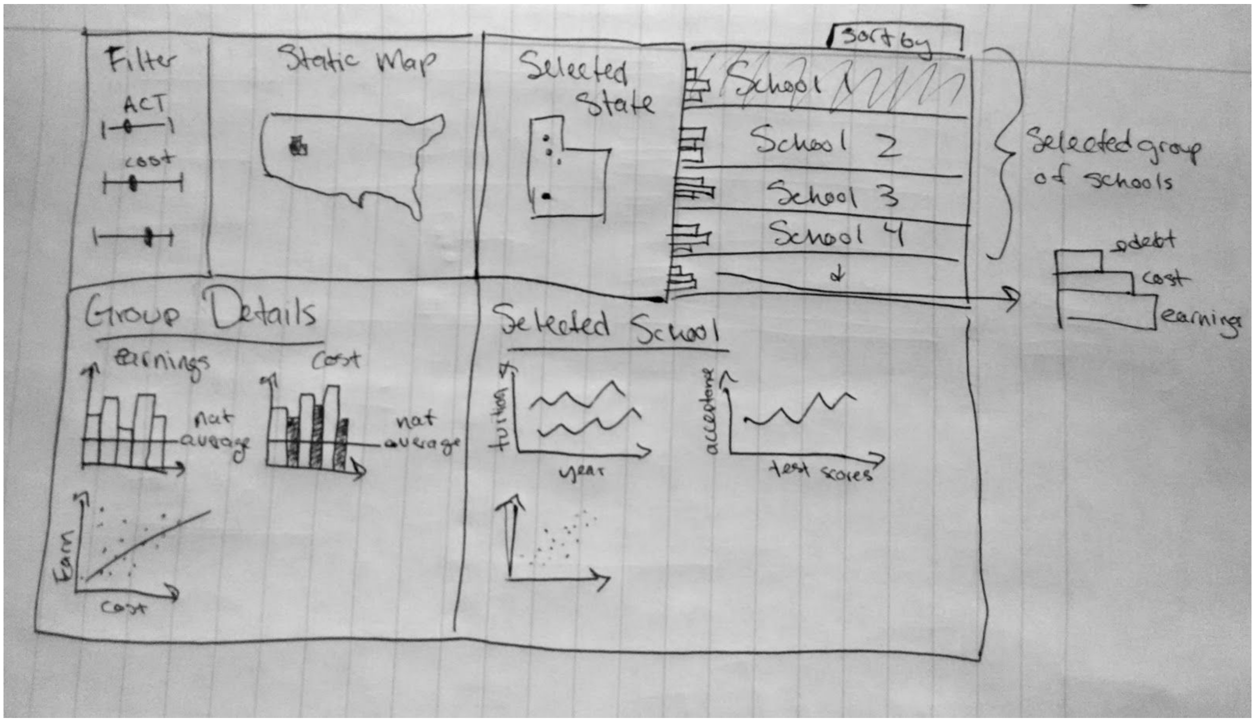
This sketch spends some time outlining exactly what kinds of graphs we wanted to use. We want to have a graph that shows the change in public and private tuition over the years for each school. We also wanted to graph the acceptance rates against test scores, to give an idea of what scores get in the most. We would also like to graph multiple schools against each other for comparison. Specifically we would like to compare the earnings made per year on average by students from each university. We also had the idea to do a paired bar graph with the tuition and average debt of students from each university, so we could see the difference between both. All of our bar graphs would also include national averages to compare against.



This design sketch shows some of the major pieces that we wish to have in our visualization. In the top left are dynamic query filters which allow the user to filter the data into a subset that they are looking for. Examples are filtering by tuition cost or test scores. In the center is a map which gives locational reference to the schools that are selected by the filter. In the map frame the user can further filter the results according to location, choosing an area of the country to bound their search by. It would also zoom the map to clip and center on the current selected area (state) or school. In the upper right there is a list of the current schools that are selected. They are shown with their name and logo. Next to each school name is shown a normalized bar chart of the information that the user is looking for. Each bar is a category (like tuition, or debt) and is normalized by the total selection. So the school in the group with the highest tuition would have the largest bar for that category and the school with the lowest tuition likewise. This setup allows us to put differently scaled categories in the same clustered bar chart.

The lower two frames have more detailed information. The frame on the lower left holds detailed comparison charts for the group. From this frame the user can tell the exact amounts for schools for the different categories and compare them. In the lower right detailed information about the specific school selected is shown. Here they would be able to see information like the changes in tuition costs over the years.

As also shown the selected school would be highlighted in the other views to link them together and allow the user to better track the current school and compare it.



This design sketch is a more detailed version of the one before, however it still shows a different view for the selected state, which we have decided to get rid of in favor of adding a zoom feature to the whole U.S. map. We also added a “sort by” dropdown menu to sort the selected group of schools.

There will be multiple ways to select groups of schools. Selecting a state will show all of the schools in that state, and selecting a particular school on the map will show only that school. One of the schools in each list will be selected and its individual statistics will be show in the bottom right view. The user may change this selection. Another way to select a group of schools will be to use the filter at the top left, which will return a group of schools that meets the criteria filtered by. We will also give the user another option to select whichever schools they like by adding an “Add School” button to the top left by the list. This will allow them to choose any schools they wish to compare.

The cumulative group statistics will use dynamic transformations and will change to fit any group selected. If no group is selected, it will show national averages.

There will be a group limit, somewhere around 10 schools, that you can use to compare at a time.

Must-Have Features

- An efficient and easy to use Map for navigation.
- Easy and thorough comparison of schools.
- It needs to be easy to select groups of schools
- The data needs to be clear and easy to understand.

Optional Features

- School logos
- A Smaller breakdown of test scores and earnings by subgroups
- A link to a page that describes our data and the importance of it, as well as what about the data could be misleading.
- Cool animations for the zoom and the change in graphs

Project Schedule

Week	Date	Deadlines		
		Charts: Sierra Allred	Dynamic Query: Chris Golling	Maps: Chris Garrett
1	Oct 30	Basic Bar charts	Filters chosen, data setup	basic map, school counts placed
2	Nov 6	Full charts started	javascript for queries	map zoom/state selection
3	Nov 13	All charts done	Query filters done	Selected schools list started
4	Nov 27	Integration (selected schools list)	integration (selected schools list)	integration (selected schools list)
5	Dec 4	Project Due, polish selection, hover tips. etc		

- Overview and Motivation: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.
- Related Work: Anything that inspired you, such as a paper, a web site, visualizations we discussed in class, etc.
- Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?
- Data: Source, scraping method, cleanup, etc.
- Exploratory Data Analysis: What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?
- Design Evolution: What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?
- Implementation: Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.
- Evaluation: What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

Data Collection

The data we used came organized in a big set of CSV files organized by year. There were also additional files that contained subsets of the data. The biggest problem we had was that the data file were so large that our file reader could not open all of on file, so we could only access the first 75% of each file. The data on debt was in the last part of the files, so we had to search through the data subsets to find those field. First we processed the location data so that the map could be built, and then we worked on cutting and organizing the data.

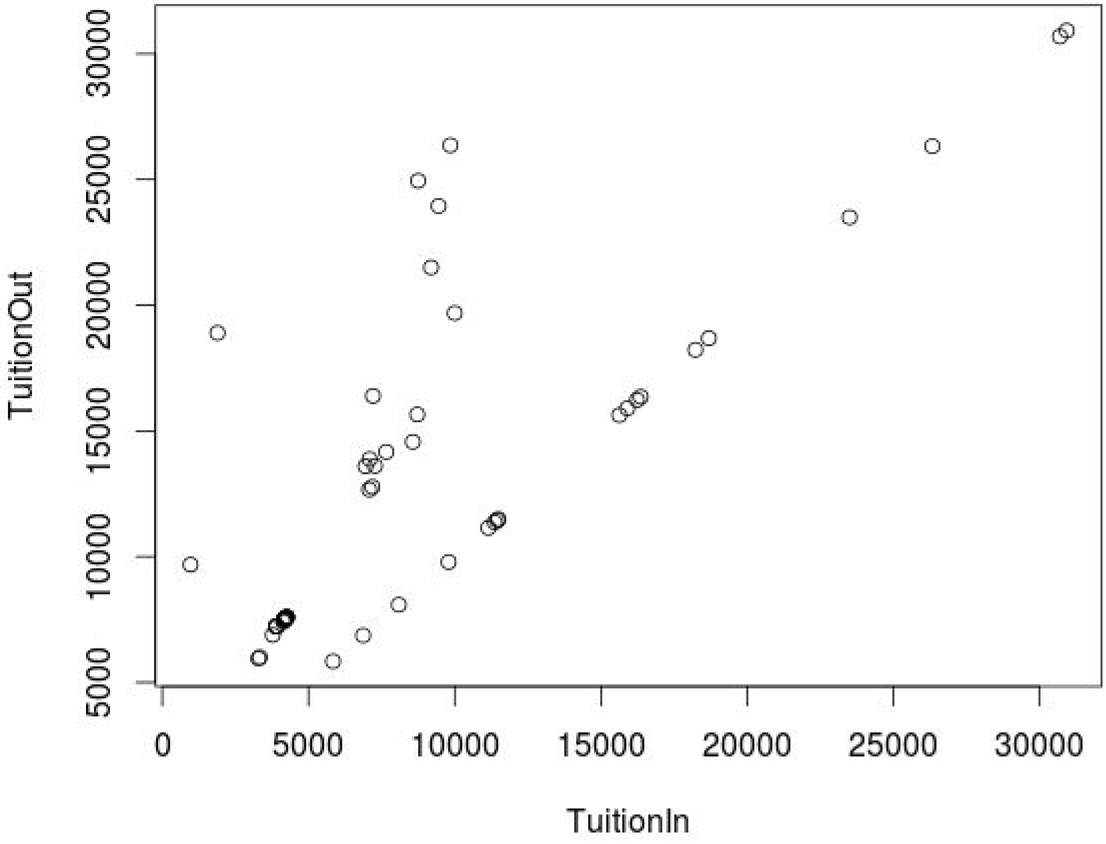
Because our project is using so much data, and because the files were so big, we spent a lot of time cutting the files down. The first step in this process was reading through the entire data dictionary to decide which parts of the data were worth keeping and marking the most important fields. This took the most time, but also helped us discover data that we did not know was available because it was not described in the overview of the data. Next we went through each document by hand and deleted the columns that were not useful to us. We then created our own small data dictionary to keep track of the most important field names.

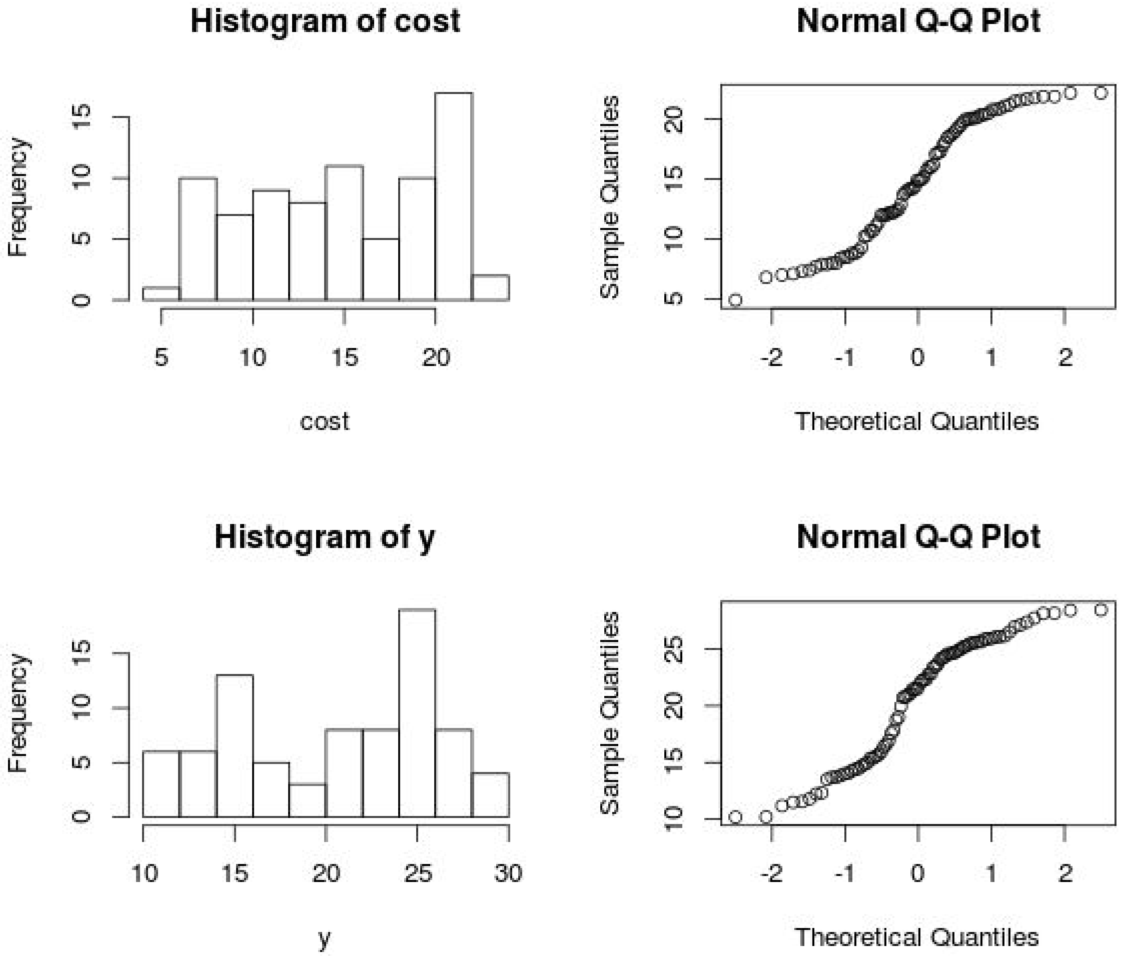
Data Analysis

After the data collection was completed we spent some time graphing subsets of the data fields to see if anything interesting existed that would be useful in the graphs portion of our visualization. To do this quickly we used R and spent some time plotting histograms, Q-Q plots, line graphs, and scatterplot matrices. We also looked at the correlation coefficient between many different data types (i.e. average debt, cost, SAT scores, type of school) to find a good fit.

We found a lot of interesting things, so it was hard to choose which graphs would be the most interesting and informational for our audience.

[Talk about the correlations found, what we chose to use, and why]


TuitionOut
5000
10000
15000
20000
25000
30000
TuitionIn
0
5000
10000
15000
20000
25000
30000



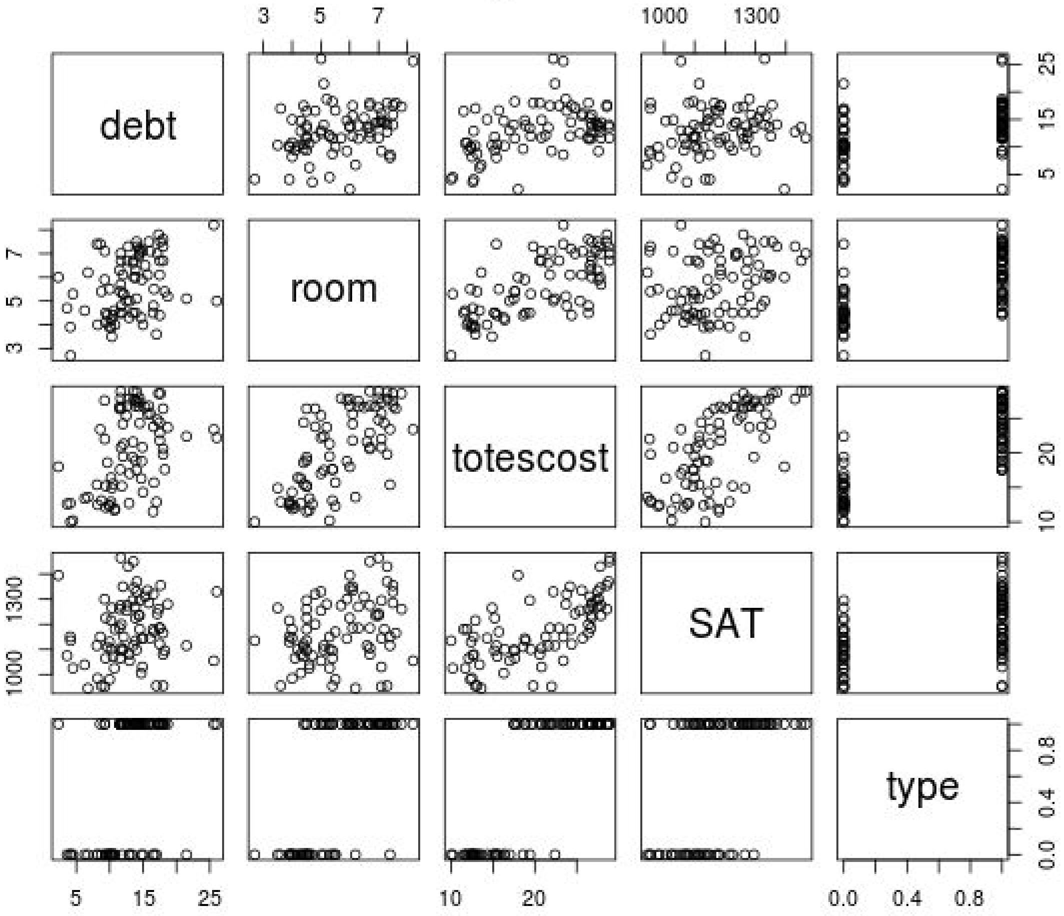
Histogram of cost

Normal Q-Q Plot

Histogram of y

Normal Q-Q Plot

Figure 1



Design

(Things we change about our design)

We decided when we started that it would be difficult to have the map, search bar, and list of schools right next to each other, so we decided to move the list of the schools down next to the graphs.

Peer Feedback

Peer feedback was very helpful in focusing the design of our project. Listed below are some of the questions we had to consider and what we did about them.

What are you doing on top of what the college scorecard already does?
We are showing the data in a very different way. College Scorecard focuses on statistics for individual schools and there is almost no comparison among schools. We will really be focusing on the comparison and analysis of the data. We are incorporating a map to visually represent all of the schools in a beautiful way, which College scorecard does not do.

Would you consider the use of tooltips and popovers rather than link to new page?
We will definitely be using tooltips and popovers in our design to explain small things about the graphs and search bar, however there still will be a link to another page to describe our data and the discrepancies therein so that we are honest about the claims we make in the visualization.

Currently it looks like there are 15 charts in your visualization, it seems like a lot to look at.
There will be a maximum of 8 charts. We found this necessary to present the data in a complete manner.

How are you going to handle scale in terms of number of school in a state?
This will be done with a zoom technique. You will not see all the schools in a state until you zoom in on it.

What if you want to compare schools in different states?
The selection process will allow you to select any combination of schools, so you can compare schools in different states.

There is data you're not including from the college scorecard dataset. Was that the intention of your visualization?

Yes, there is way too much data for us to implement in one project. We chose to focus on cost analysis and admittance.

What are you going for null data?

Throughout the visualization there will be small notes when data is missing for a field. We plan to handle this elegantly, while being clear that the data is unavailable.

We were critiqued by Priyanka Parekh's group. The feedback they gave us was very fair and helpful.