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
Project Website: http://sdallred.github.io

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.

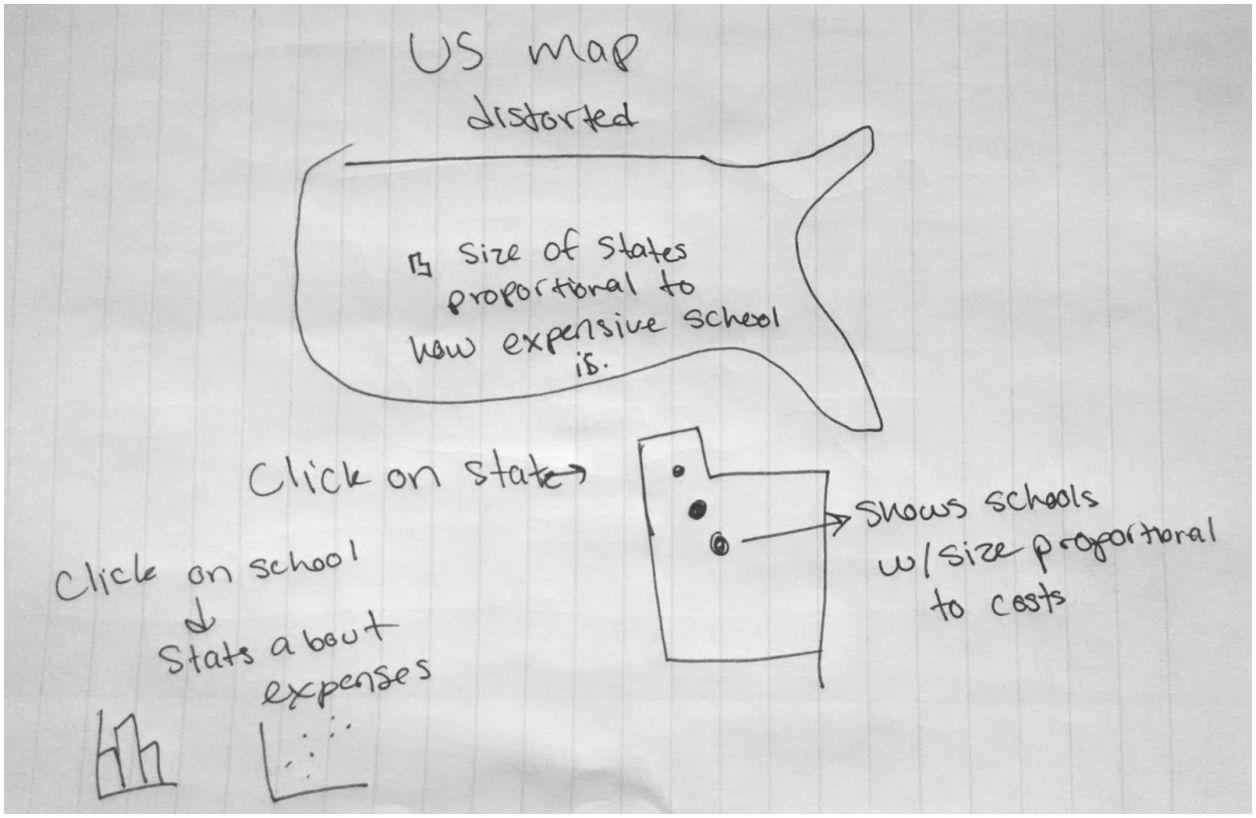
We plan on using basic information about each school, including location, URLs for the Institution's Homepage, whether it is a main versus a branch campus, if it is public, private nonprofit, or private for-profit, if the school has a specialized mission or religious affiliation, if there is a distance-education-only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

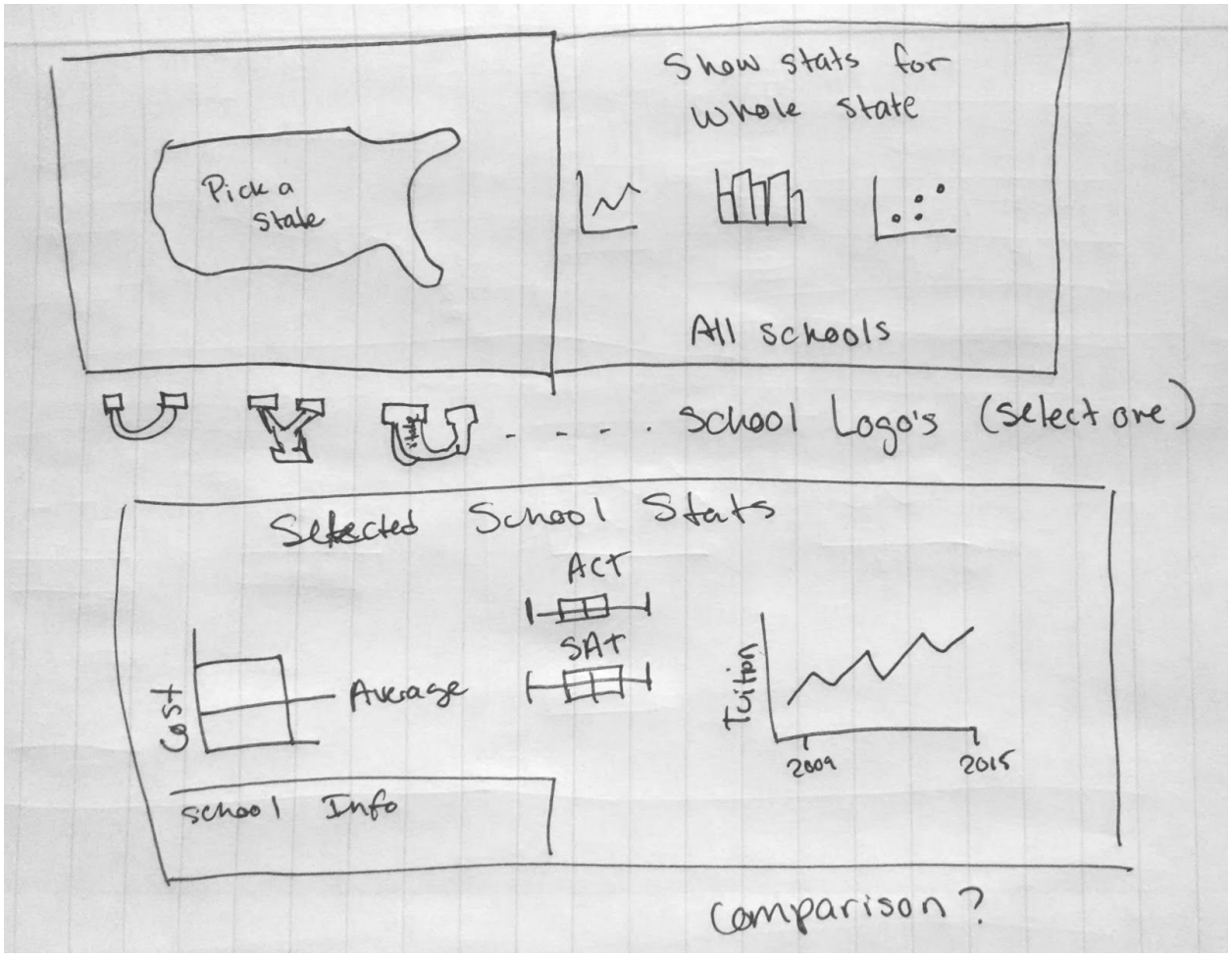
For data analysis purposes we will be using data on cost, financial aid, and projected earnings. In the cost category we will be using the average cost of attendance, tuition and fees, and possibly average net price by income level. For financial aid we will be using percent of undergraduates receiving 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.

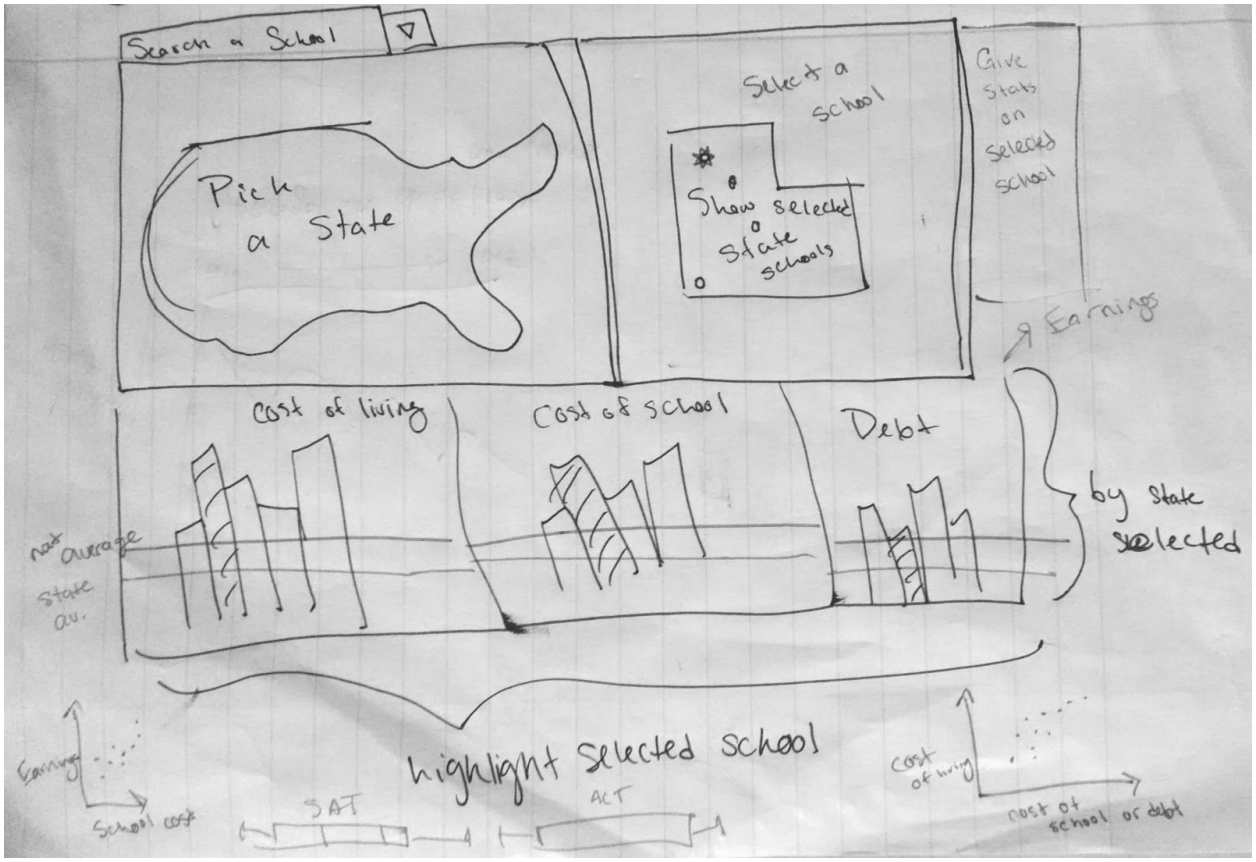
Visualization Design



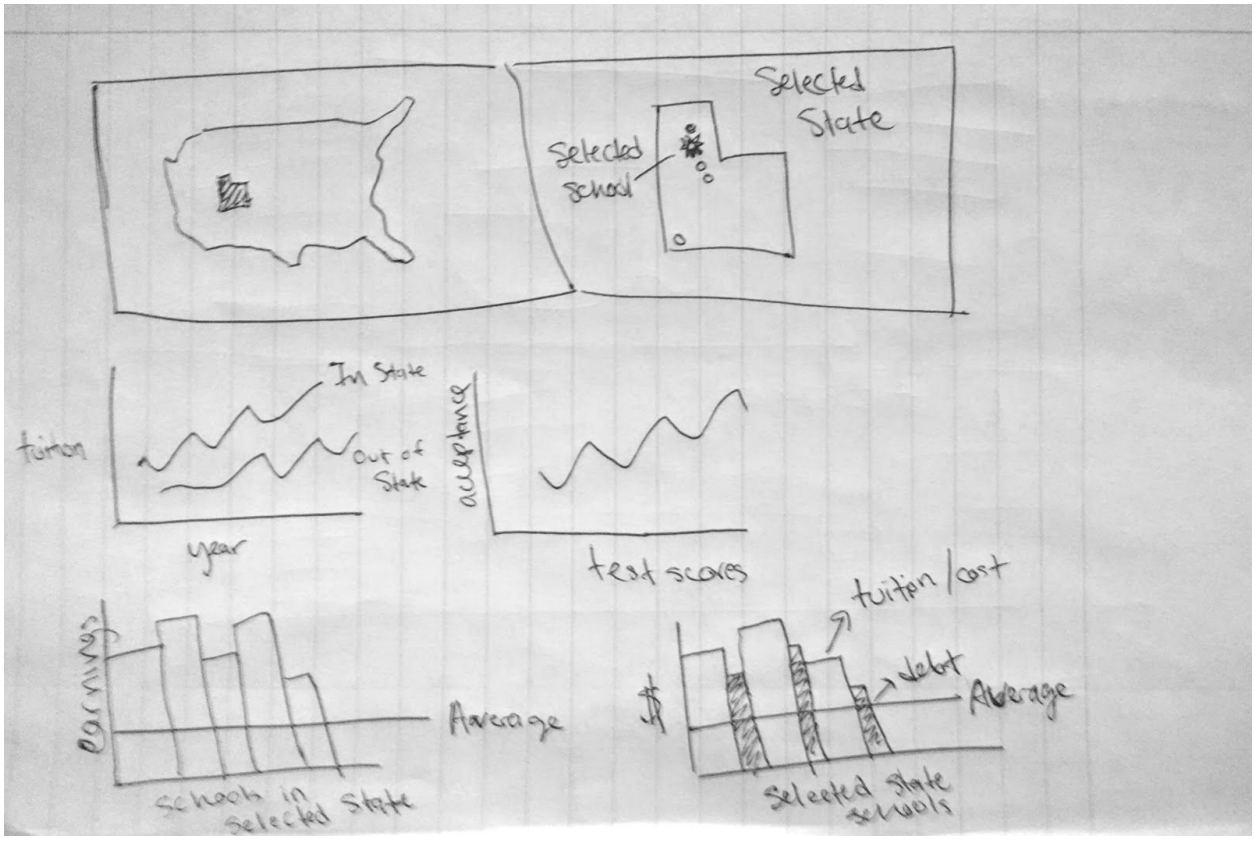
This first sketch was one of our original ideas. Due to the nature of our data we knew that the best way we could organize all of the schools in the US was by using a map, because there are not real links between schools in our data. This design uses the average cost of all the schools in each state 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.



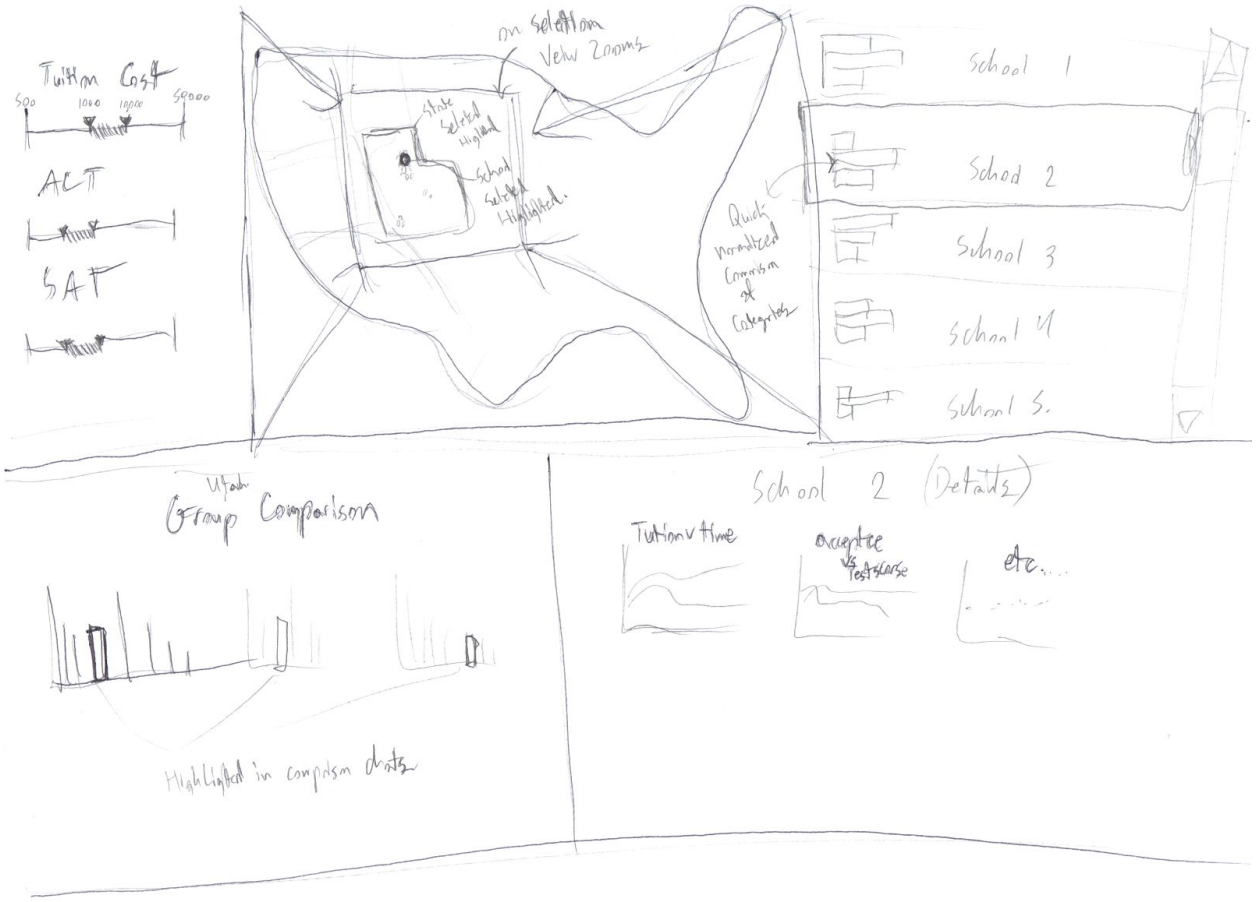
This design still uses a map like the first one, but does not use distortion and area to show the data. Instead, when you select a state you see statistics next to that window about the education in that state as a whole. Underneath is listed the logos of all the schools in that state, and if you click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

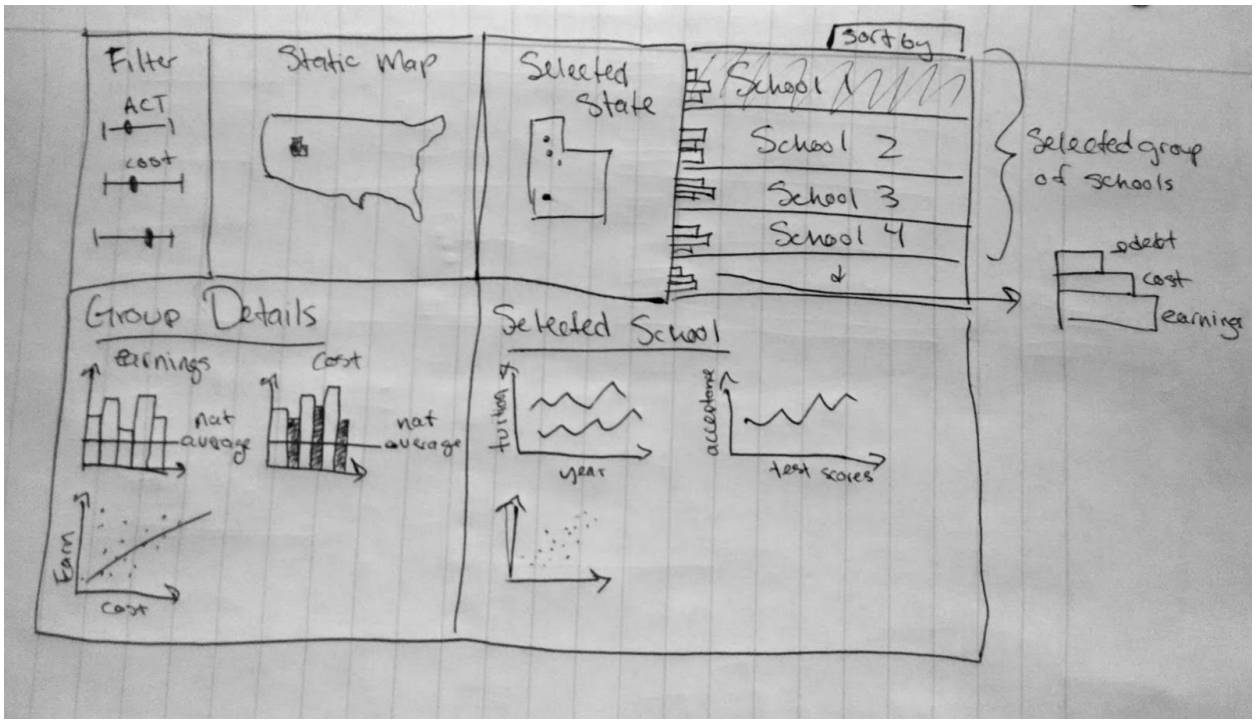


This design sketch shows some of the major pieces that we wish to have in our visualization. In the top left are dynamic query filters which allow the user to filter the data into a subset that they are looking for. Examples are filtering by tuition cost or test scores. In the center is a map which gives 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

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.

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.

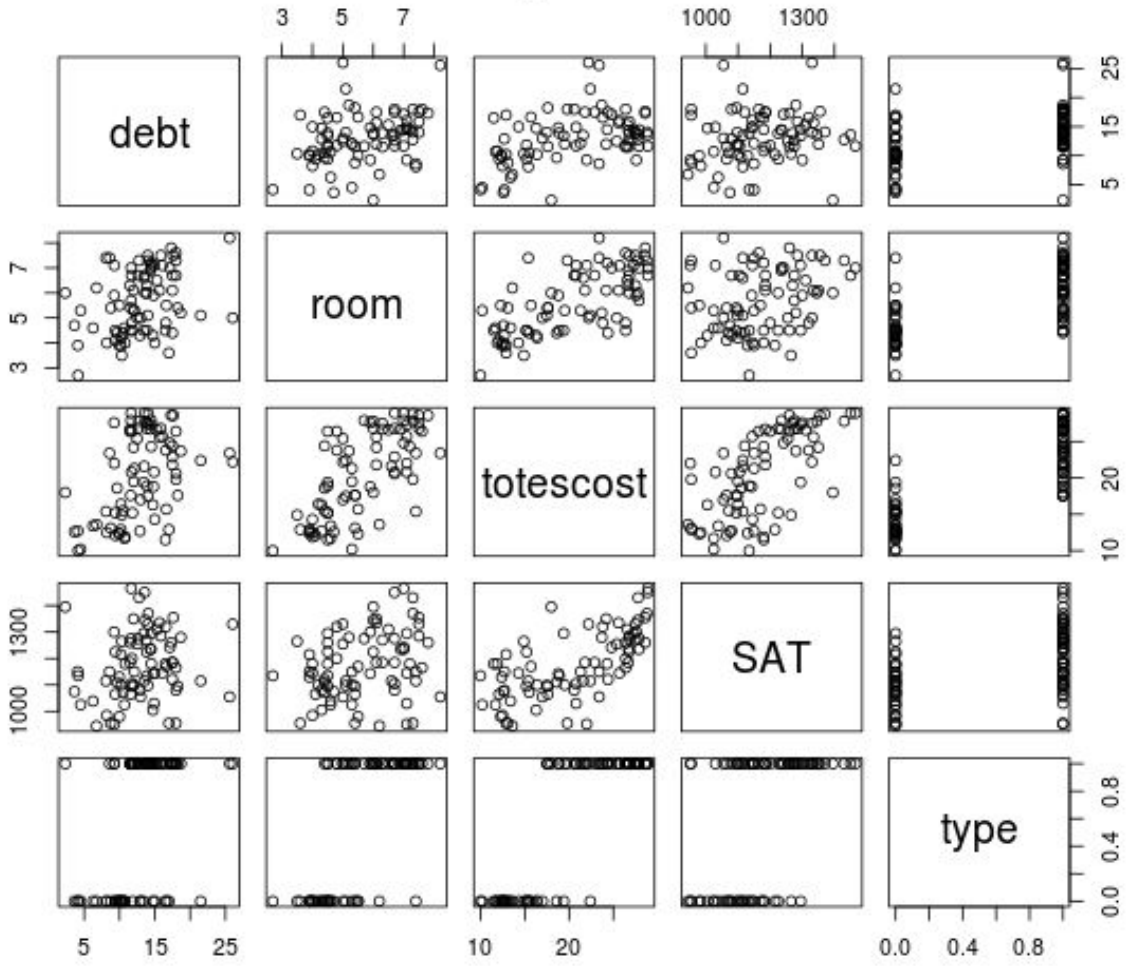
Because our project is using so much data, and because the files were so big, we spent a lot of time cutting the files down. The first step in this process was reading through the entire data dictionary to decide which parts of the data were worth keeping and marking the most important fields. This took 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 use a data dictionary to keep track of the 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. For example, in the graph below we plotted the in-state and out-of-state tuition of a group of school against each other, and found an interesting correlation. The schools that form a straight line have very similar prices for in-state and out-of-state, and these schools are usually private schools. The bunch of schools that show a significantly higher out-of-state tuition were public.

Figure 1



This graph shows the correlations between different factors, the more linear the grouping of points, the higher the correlation. We found that cost correlated with a lot of factors.

Although we found a lot of cool trends in the data, we really wanted our visualization to be concise and clear. MAny of the graphs would confuse readers and it would be difficult to explain their significance. So instead we chose to limit our graphs to comparisons between schools with data that would be pertinent and interesting to the users of our visualization.

Major Design Changes

- 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.

- To remove confusion and make our visualization easier to use, we decided not to include graphs in the section about the selected school. Instead we display important information about that school. Making the graphs would have taken our visualization off on an unrelated tangent, because we would be comparing how prices have changed for that school over the years, when the user will be using it to find recent data.

- Instead of using many small graphs next to each other, we chose the datasets we wanted to plot and made one large dynamic graph that allowed the user to select what they wanted to see, this saved us a lot of space and made the whole visualization a lot cleaner.

- We chose not to make the graph selectable, because we felt having the filters was thorough enough and that it would confuse the user to be able to select schools in so many places, instead we used it as an informational tool to show where the schools were that they were looking for.

Filter Design

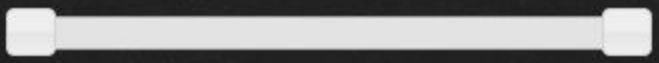
The filters are used to allow the user to narrow down what school they want to look at by having them choose their preferences. The filters we chose are items that someone would most commonly look at when trying to select a school. These include: SAT scores the school accepts, ACT scores the school accepts, what the tuition is for the school in and out of the state, the population size of the school, and the type of school that it is. For the SAT scores, ACT scores, and the tuition, we decided that a 2 ended slider would work best as a filter for these. The population size and school type, we decided that checklist would work better for these since their wasn't as much to choose from in these list.

Filters

Avg Sat Admitted: 0 - 2400

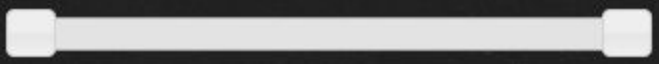
☐ Remove schools with no SAT score data

Median of Act Admitted: 0 - 36



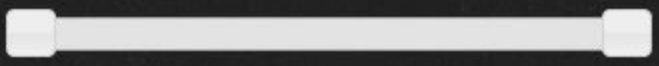
☐ Remove schools with no ACT score data

Tuition Cost In-State: 0 - 55,000



☐ Remove schools with no in-state tuition cost

Tuition Cost Out-of-State: 0 - 55,000



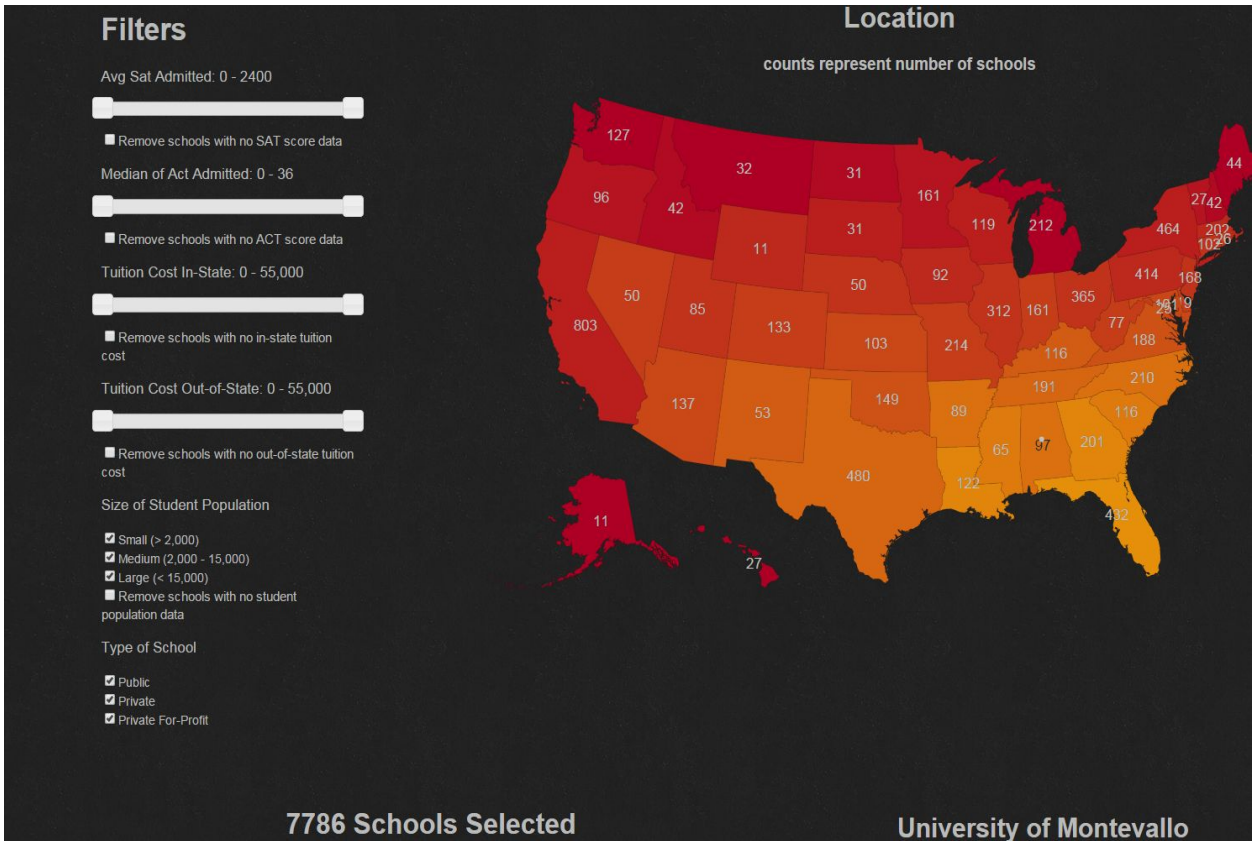
☐ Remove schools with no out-of-state tuition cost

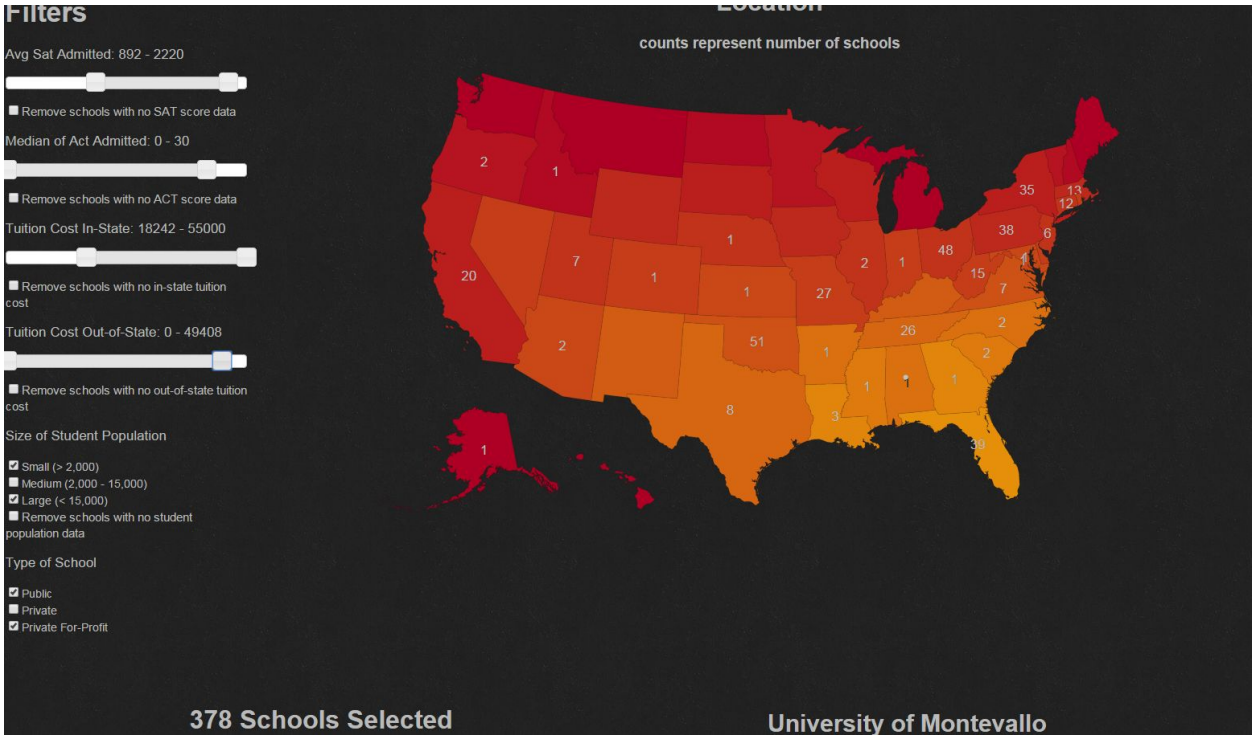
Size of Student Population

☑ Small (> 2,000)
☑ Medium (2,000 - 15,000)
☑ Large (< 15,000)
☐ Remove schools with no student population data

Type of School

☑ Public
☑ Private
☑ Private For-Profit


Filters
Avg Sat Admitted: 0 - 2400
Remove schools with no SAT score data
Median of Act Admitted: 0 - 36
Remove schools with no ACT score data
Tuition Cost In-State: 0 - 55,000
Remove schools with no in-state tuition cost
Tuition Cost Out-of-State: 0 - 55,000
Remove schools with no out-of-state tuition cost
Size of Student Population
Small (> 2,000)
Medium (2,000 - 15,000)
Large (< 15,000)
Remove schools with no student population data
Type of School
Public
Private
Private For-Profit
Location
counts represent number of schools
127
32
31
161
44
96
42
31
119
212
464
11
50
92
414
168
50
85
133
103
214
312
161
365
803
116
77
188
210
137
53
149
89
191
116
65
97
201
480
122
432
11
27
7786 Schools Selected
University of Montevallo


Filters
Avg Sat Admitted: 892 - 2220
Remove schools with no SAT score data
Median of Act Admitted: 0 - 30
Remove schools with no ACT score data
Tuition Cost In-State: 18242 - 55000
Remove schools with no in-state tuition cost
Tuition Cost Out-of-State: 0 - 49408
Remove schools with no out-of-state tuition cost
Size of Student Population
Small (> 2,000)
Medium (2,000 - 15,000)
Large (< 15,000)
Remove schools with no student population data
Type of School
Public
Private
Private For-Profit
counts represent number of schools
2
1
35
13
12
38
6
1
48
20
7
1
1
2
1
15
7
27
2
26
2
51
1
2
1
1
1
8
3
1
39
378 Schools Selected
University of Montevallo

As you can see when the filters are changed the number of schools on the map changes.

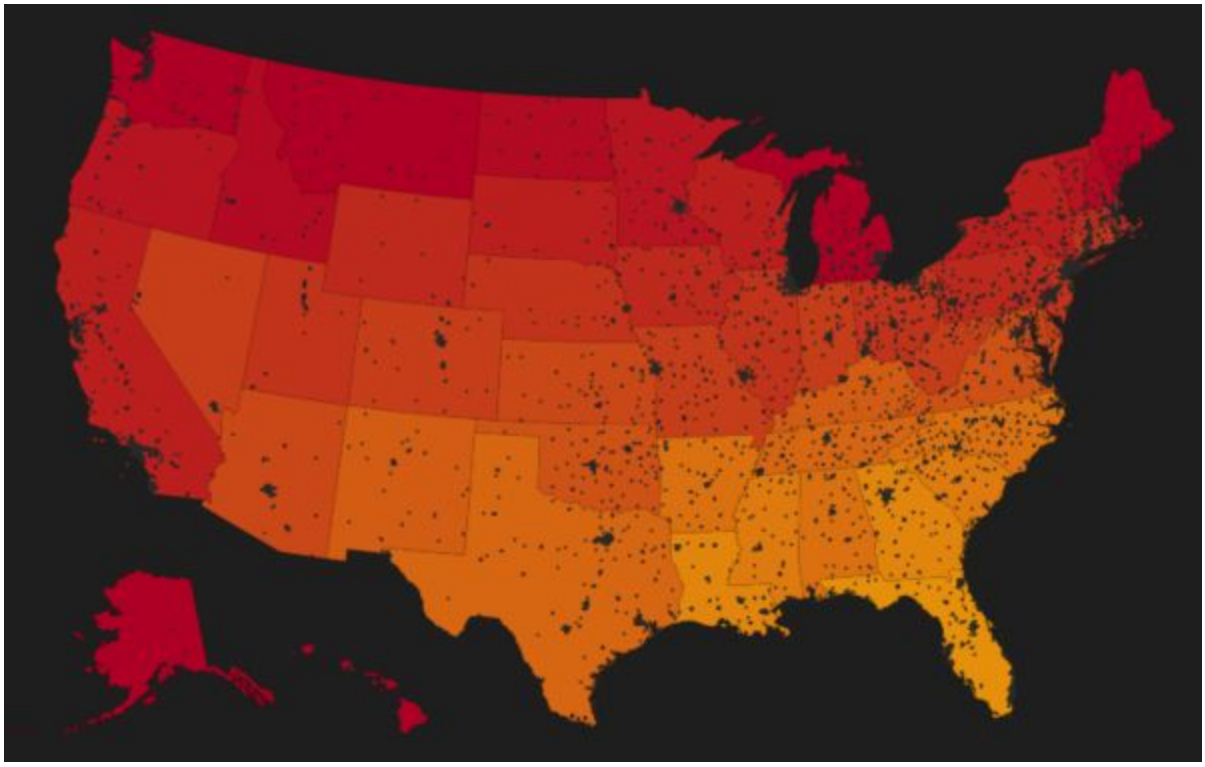
We ran into quite a few problems with the filters. The first one was the SAT and ACT scores. The data that we used had an SAT average for each school, which was perfect for the filter. The data unfortunately did not have an average for the ACT scores though. So we had to resort to using the median score for each school instead for the ACT scores.

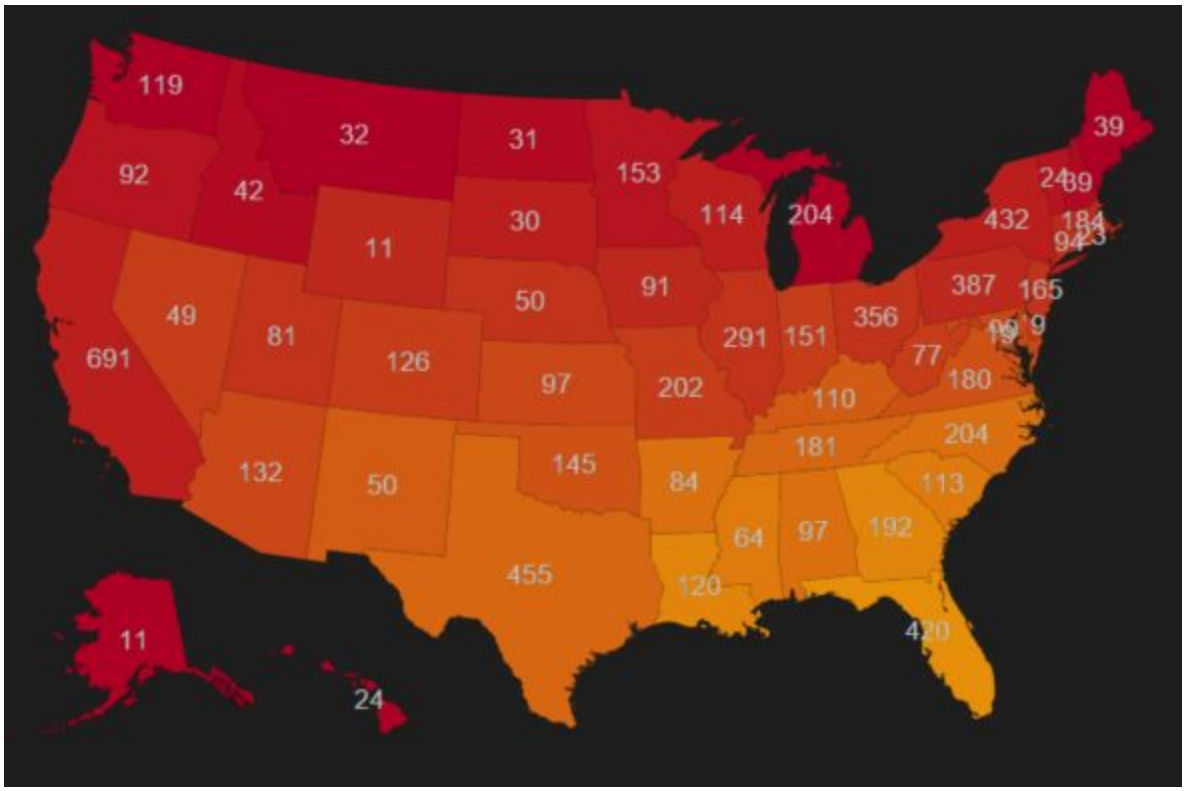
Another problem we ran into was that each of the different types of data, except for the school type, for the filters had null values. At first we decided that we would just leave them be and not filter them out, but then we realized that that could make it hard to filter down to a small list of schools. So we fixed this by adding a checkbox to each filter, besides the school type filter, that if checked will remove the schools that have a null value for that type of data from what is being presented.

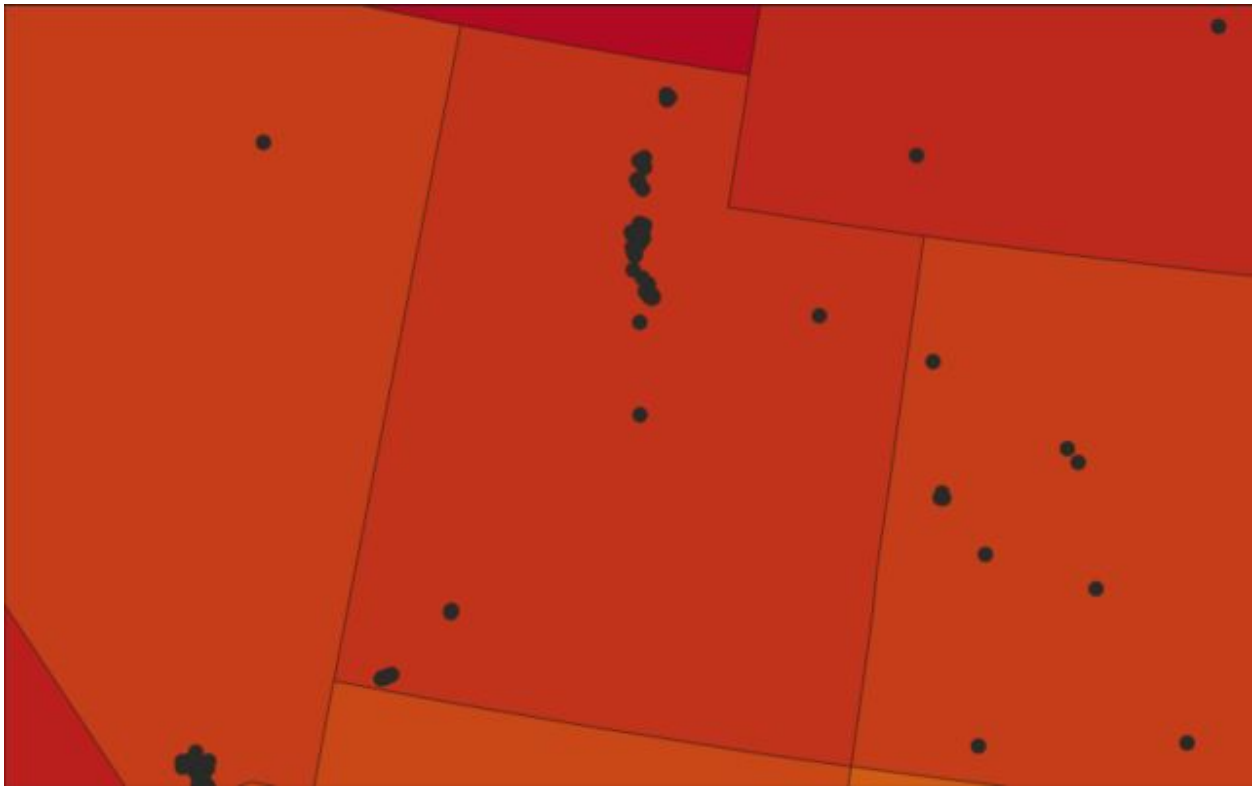
Map Design

The goal of the map is to provide the user with context in multiple ways. The first is to provide a locational reference for the schools. This way the user can narrow their search by location, targeting schools that are closer to their location, or at locations that they wish to explore.

This was done by placing the schools as dots on the map as shown below. The problem that we had with just showing the dots was that when viewed from afar it was hard to get a feel for how many schools were shown since they overlap. To solve this we used the idea of contextual zoom. On a high level you get a number of how many schools are in the state. Then when you zoom in you then get the locational dots for the schools.



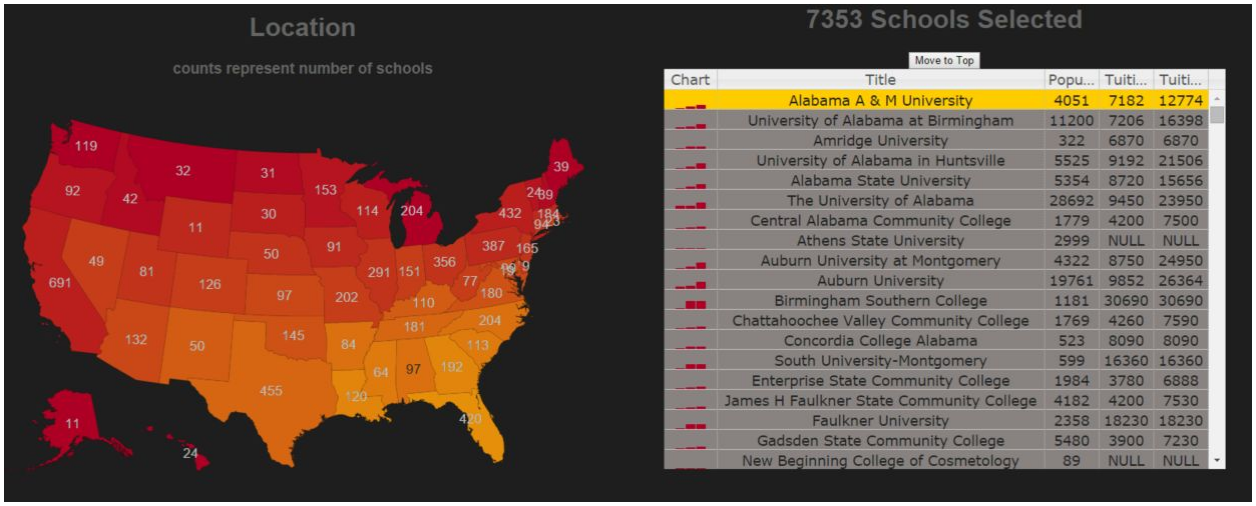

119
32
31
39
92
42
153
24
89
30
114
204
432
184
11
94
3
91
387
165
50
49
356
81
291
151
77
691
126
180
97
202
110
204
181
132
145
50
84
113
64
97
192
455
120
11
420
24

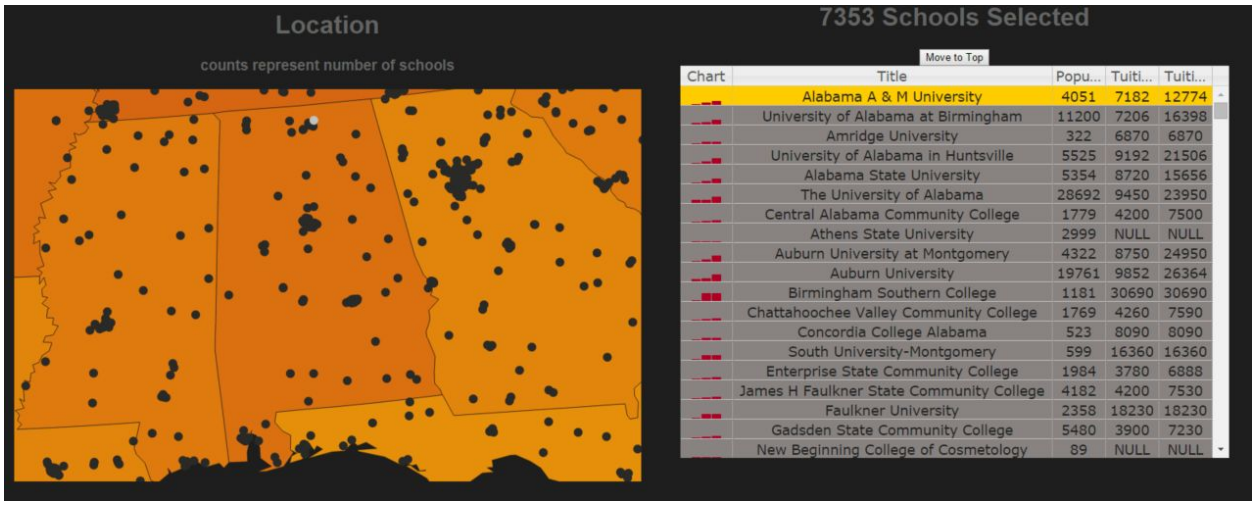


This allows the user to get a high-level view of the distribution of the schools that are selected, and then get a closer more detailed topological view of the distribution.

Selecting a School on Map

Because of the overlap of the school's locator dots we thought that it would would be infeasible to select the schools from the map. Instead to select one school to view in detail the user double clicks on the school in the Selected Schools list. When they do that the state that the school is in and the schools dot are both highlighted. This allows to use to view the location of the selected school, and is shown in the images below





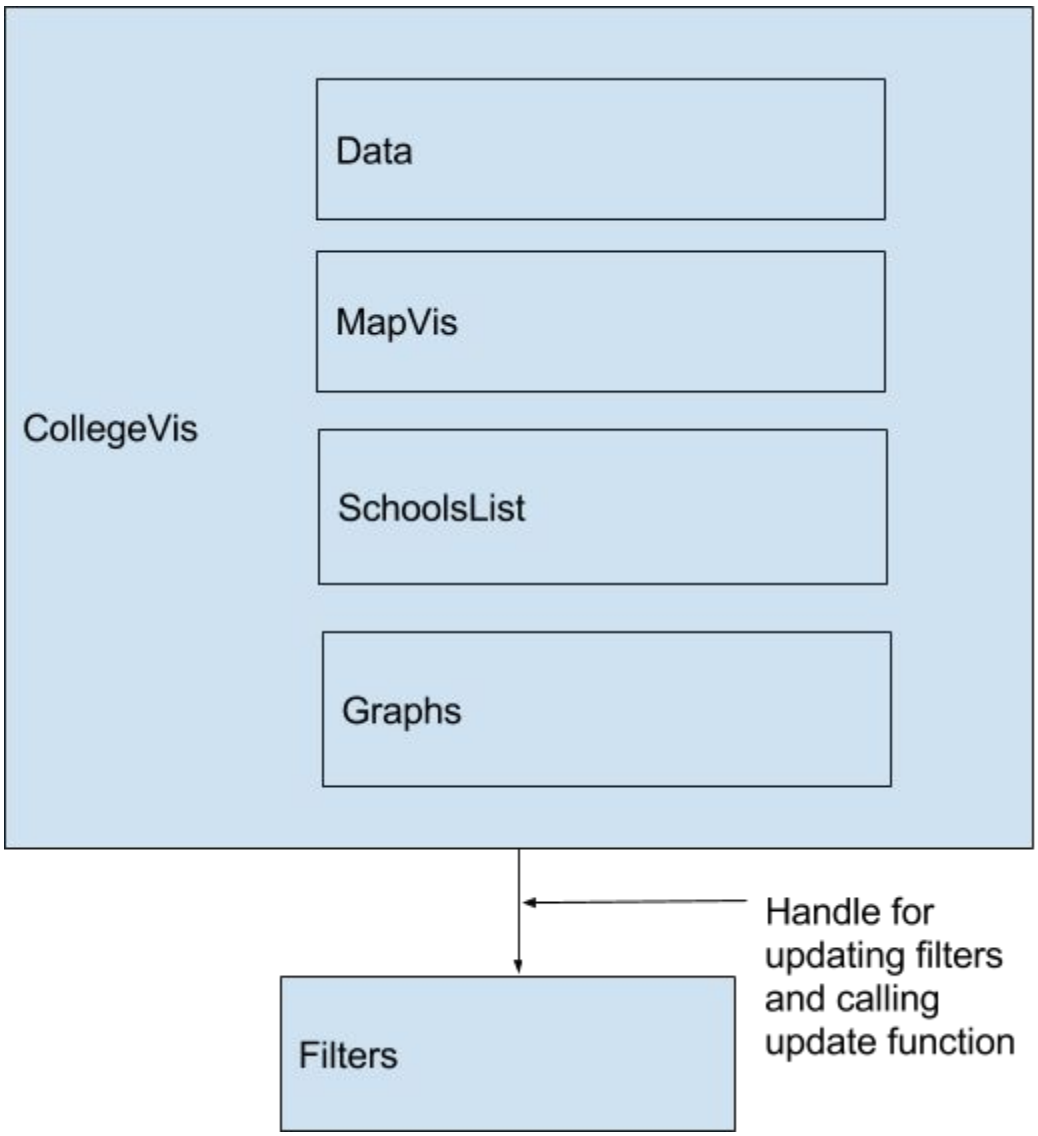
Development Notes

When first developing the map we were trying to make it so that the user could navigate freely across the US, looking at the dots of the schools. Because the number of schools was so high the panning speed was very slow. Since that was intractable we looked for another example. We found one here that did a fixed zoom/pan. When a country was clicked on it zoomed into that country. Using that as an example we modified it to work with the United States.

Next was doing the counts. We had some difficulty until we found the companion id file for the US.json file. We used that to identify the states. Once we had that it was a simple matter to count the schools in each state and display that number in the center of the state. Then all we had to do was integrate the hide and show functionality into the clicks.

Framework

When developing the map and the Schools List we discovered that just like homework 4, which had many linked graphs, this also was going to have many linked controls and graphs. Chris Garrett built the backbone framework for the page. It is started off by the loading of the data and the creation of a CollegeVis object. This object holds the data and all of the graphs. Its job is to build and update all of the maps, lists, and graphs. A handle for the CollegeVis object is given to the code that handles the listeners for the filters, and to the graph objects so that when their events happen they can also call the main update function.



Selected Schools List

This list's purpose is to show the user what schools are currently selected by their filters. They are mainly listed by name. Chris Garrett had the idea to display a small graph showing how their schools chosen statistics rank relative to the total selected group. He found a library SparkLines which did this nicely. The data is scaled by finding the max in each category and then displaying the percent of the school's value of that max.

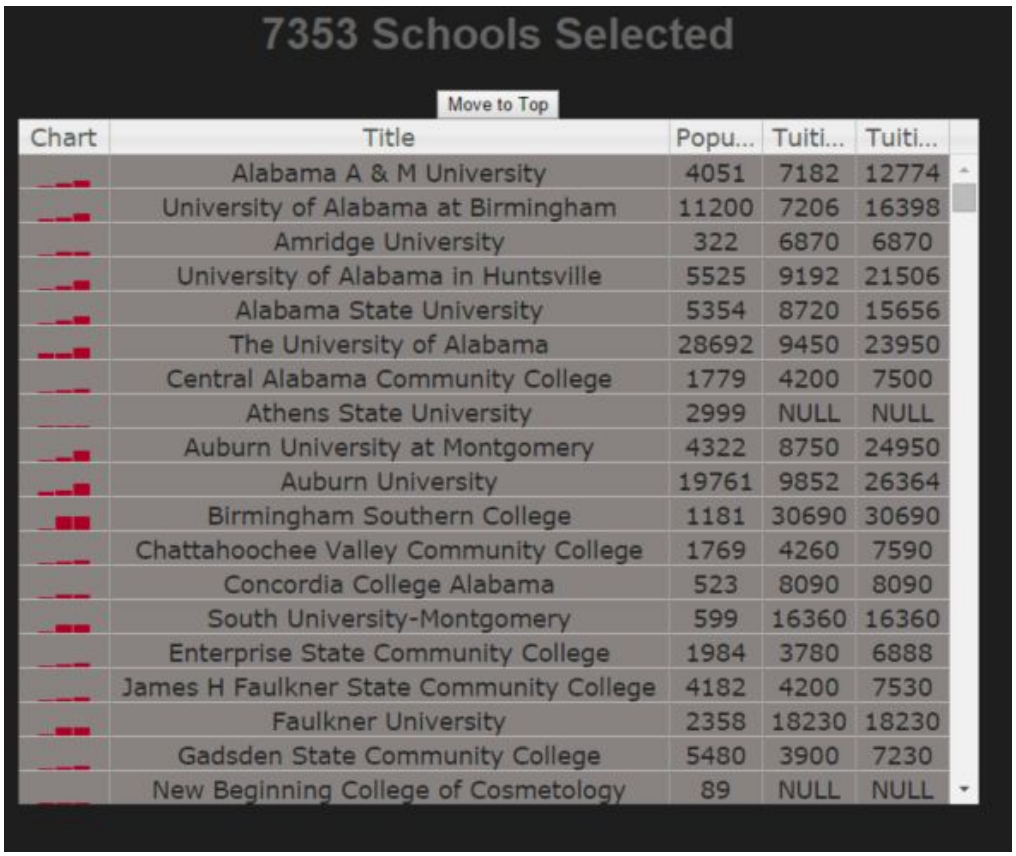
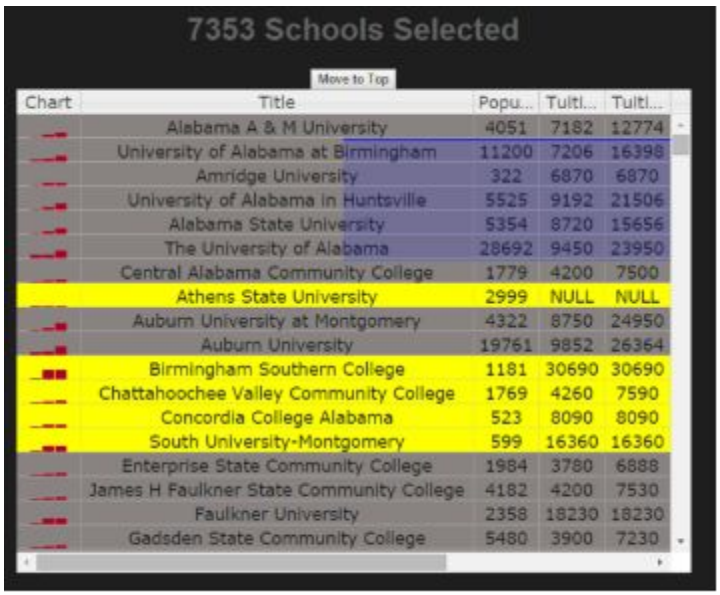


Chart	Title	Popu...	Tuiti...	Tuiti...
	Alabama A & M University	4051	7182	12774
	University of Alabama at Birmingham	11200	7206	16398
	Amridge University	322	6870	6870
	University of Alabama in Huntsville	5525	9192	21506
	Alabama State University	5354	8720	15656
	The University of Alabama	28692	9450	23950
	Central Alabama Community College	1779	4200	7500
	Athens State University	2999	NULL	NULL
	Auburn University at Montgomery	4322	8750	24950
	Auburn University	19761	9852	26364
	Birmingham Southern College	1181	30690	30690
	Chattahoochee Valley Community College	1769	4260	7590
	Concordia College Alabama	523	8090	8090
	South University-Montgomery	599	16360	16360
	Enterprise State Community College	1984	3780	6888
	James H Faulkner State Community College	4182	4200	7530
	Faulkner University	2358	18230	18230
	Gadsden State Community College	5480	3900	7230
	New Beginning College of Cosmetology	89	NULL	NULL

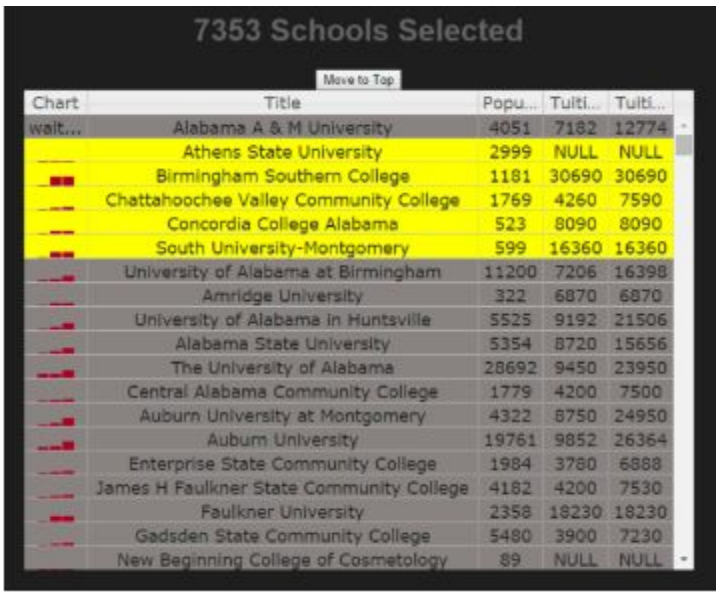
The problem that he encountered is that because the list can contain several thousands of schools creating and displaying all those graphs bogged down the page tremendously. In searching for a solution he found another library SlickGrid which enabled asynchronous rendering so that only those currently seen were rendered. That improved the performance of the page to be more reasonable.

This library also allowed for dragging and dropping of the lists to reorder it. This was discovered to be necessary. The Group Details Graph which displays the statistics

for the entire group can only display a finite number of schools at a time, because of the width of the graph. So to keep things simple only the top 100 or so schools are displayed on the graph. This meant that is was necessary to allow the user to reorder the list to allow them to fully explore the data.



Dragging Schools

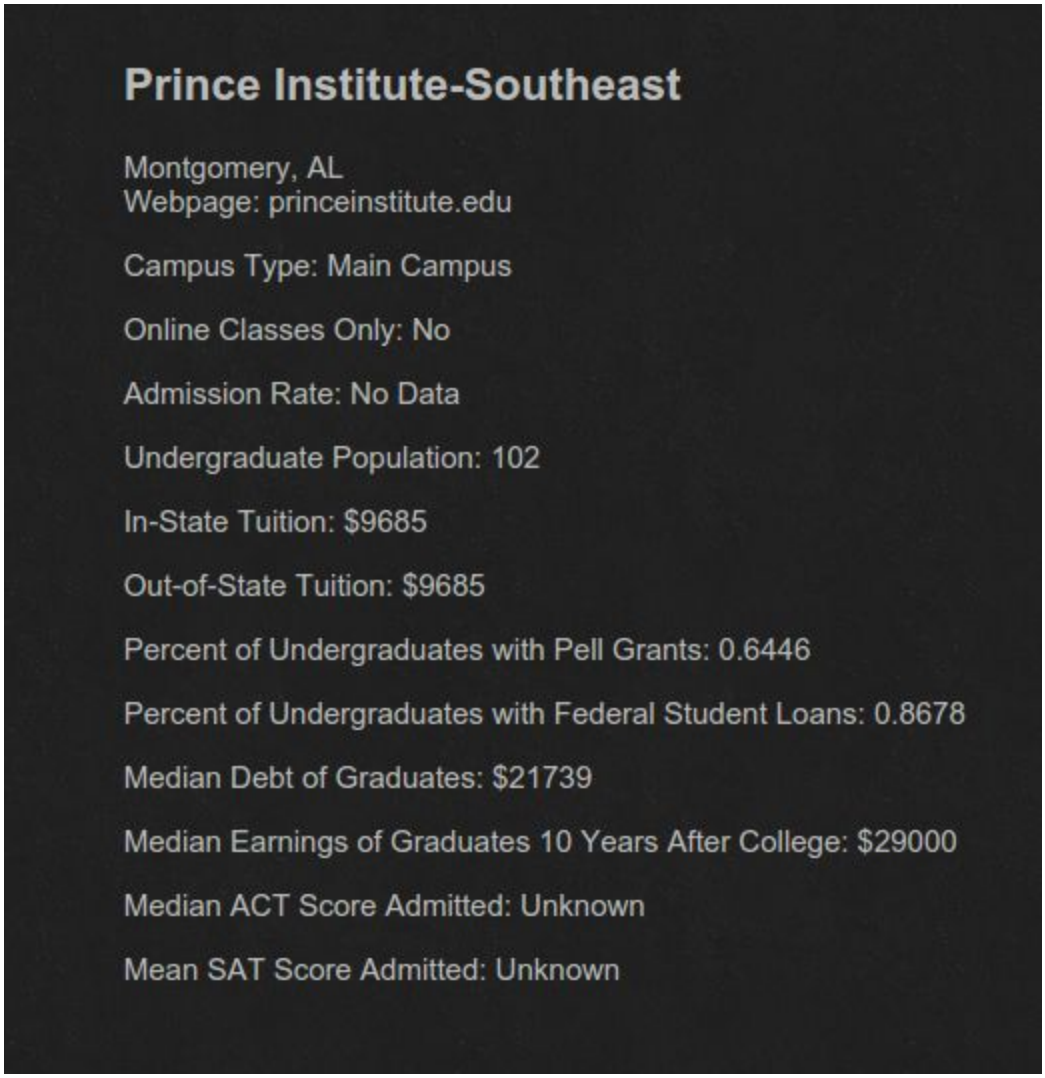


Schools Dropped

Because the list can be very long dragging schools to the top does not always work. For this reason we put in the “Move to Top” button. This moves selected schools to the top of the list. This ensures that in the Group Graph they are the first schools displayed on the left.

Selected School

When a school is selected its name and information are displayed to the right of the list of selected schools. The city, state, and webpage of the institution are displayed first, and then details about what kind of school it is. We display whether it is a main or satellite campus, where the the school is online only, and if it is a men's or women's only college.



We then give data about the school relating to its cost, size, admission rates, and admitted test scores. When data is null or protected we write that it is unknown, or that there is no data, depending on the situation.

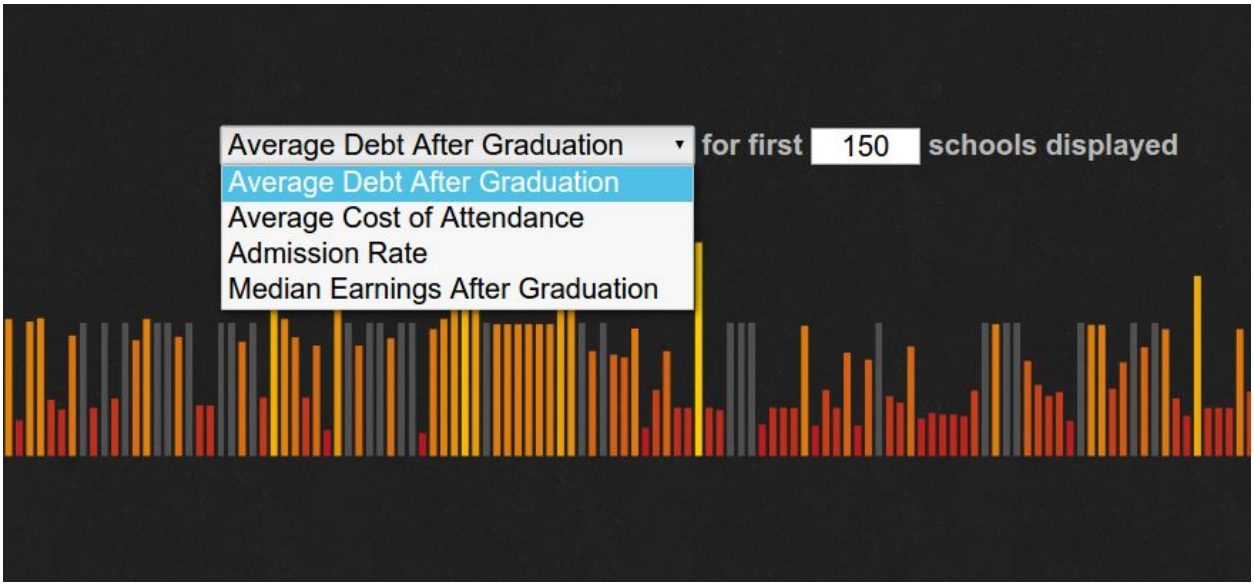
Group Graph

Sierra created individual graphs for the data, but these were cumbersome and took up a lot of space. She was unable to get them to align properly and they required a lot of scrolling to see all of them. Chris Garrett helped refactor her code using his from homework 2. He discovered that the data that we were comparing for the yScale was not in floats but was still a string. This caused the max function to pick the wrong amount as the max value of the data. After figuring this out and fixing it he used what he did in his homework 2 to add tooltips to the graph.

We noted that schools that did not have any data for the selected attribute did not show up in the graph (they had been labeled a value of 0). To make them visible we colored them a grey color with a 0.5 opacity and gave them a height of half the graph. This allowed them to be visible.

We also wanted the user to be able to view the data of different attributes. Chris Garrett created a dropdown that allows the user to select an attribute to view. The Group Graph is then updated to show the data relative to that attribute.

The four data attributes we chose to display were the average debt of students after they graduate, the average cost of attending the school, the admission rate of the school, and the median earnings of graduated students 10 years after enrollment.



Evaluation

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?

In the course of building this visualization we found that the data held so many interesting trends that we could not possibly display them all, but rather left it up to the user to find the trends themselves, based on their interest in the data. We were able to create a way to find out information about any kind of school we wished, and learned about new schools we had never heard of when we saw interesting outliers in the data.

Because of the breadth of the data it was important to keep a narrow scope for our visualization. We accomplished this by only focusing on a few of the data attributes. We feel like the filtering controls do their job adequately. There can always be improvements, but the simplicity of the controls are well balanced for their chosen attributes.

The map was somewhat of a compromise. We could either display fine detail or have a quick and responsive display. We could not have both. By using semantic zooming we alleviated some of the complexity by only showing it at the user's request. This allowed the user to get a feel for the locational aspect of the data, but also allowed them to get more detailed locational data on request. We experimented with having the school dots have hover text to show their identities but we were never able to get that working smoothly. We feel that is the only thing that the map interface is lacking, or what would make the map interface more complete. Adding the ability to see what dot represents what school, and then be able to click on them. That would greatly enhance the user's ability to search by location. Unfortunately we were unable to get this functionality working.

The next aspect is the list of schools. This allows the user to see the schools captured by their filter, but the size of the list makes it hard to work with. The sorting and reordering mechanisms do help the user to explore the data. But a full featured search mechanism might have been more useful.

Another tool for debate is the small graph that is included with the school in the list. This graph displays not the numerical values of the data points, but the ranking of those points relative to the group. The idea is that the user can then form an idea of a statistical outline that they are looking for. If they had this then could then quickly be able to scan through the list and be able to tell which schools that they wish to

investigate further. If this tool accomplished this goal I do not know. Further usage and experimentation seems to be needed.

The group bar graph on the bottom is a very interesting, informative tool. Because of the size of the data set it is hard to use it to its full potential. The ability to change the value of the category allows the user to compare a group of schools with the given attributes. The user can then be able to make trade-off decisions to decide which school would be best.

Another use of the graph might be to explore trends in the attributes and how they correlate. If this graph was done as a scatterplot, with the ability to edit the attributes that govern either axis, the correlation between the different attributes would probably be easier. But that is slightly outside the goal of this tool. It is meant to help the user to make decisions on their choice of school, not to find higher level correlations between a group of schools attributes.

The detailed information about the school is a fairly straightforward thing. It presents the data in a clean and straightforward manner. Comparing the given attributes against some median might have helped, but I feel that the group graph does that job better. It helps the user to compare the given attributes with group of schools they are deciding between. If we were to display those values (say the max and min of an attribute of the group) the computational cost might have slowed down the page visibly.

As a whole our visualization works as we intended it to, and we are happy with the result. Even though it looks a bit different than our original sketches, we feel that the current design is better than we first envisioned it.