Information Visualization Final Project

CS 4460

December 5, 2014

Stephen Okala, Jr.

Jay Patel

Pooja Kondur

Bre'Ana Deen

Milestone 7

I. Project Goals

There are many factors related to aspects of a child's life that influence their educational and life outcomes in many ways. Our goal with this project was to create a tool that parents, students, teachers, policymakers, and educators can use to gain factual insight into the factors that affect the educational and life outcomes of students in United States public schools. The general goal is that these users will be able to use the data available through our system to make informed decisions about teaching strategies, educational policies, funding decisions, and other educational decisions.

II. Possible User Questions

A few questions that users should be able to find answers to with our visualizations are:

- How does a student's employment status during their high school education affect their college attendance?
- How does a parent's involvement in their child's schooling contribute to or detract from their chance of graduating high school?
- Is there a correlation between the availability of educational resources for a student and their parents' highest levels of education?
- How do the financial resources available to a student during their high school education impact their ability to become employed after graduation?

III. Data Source

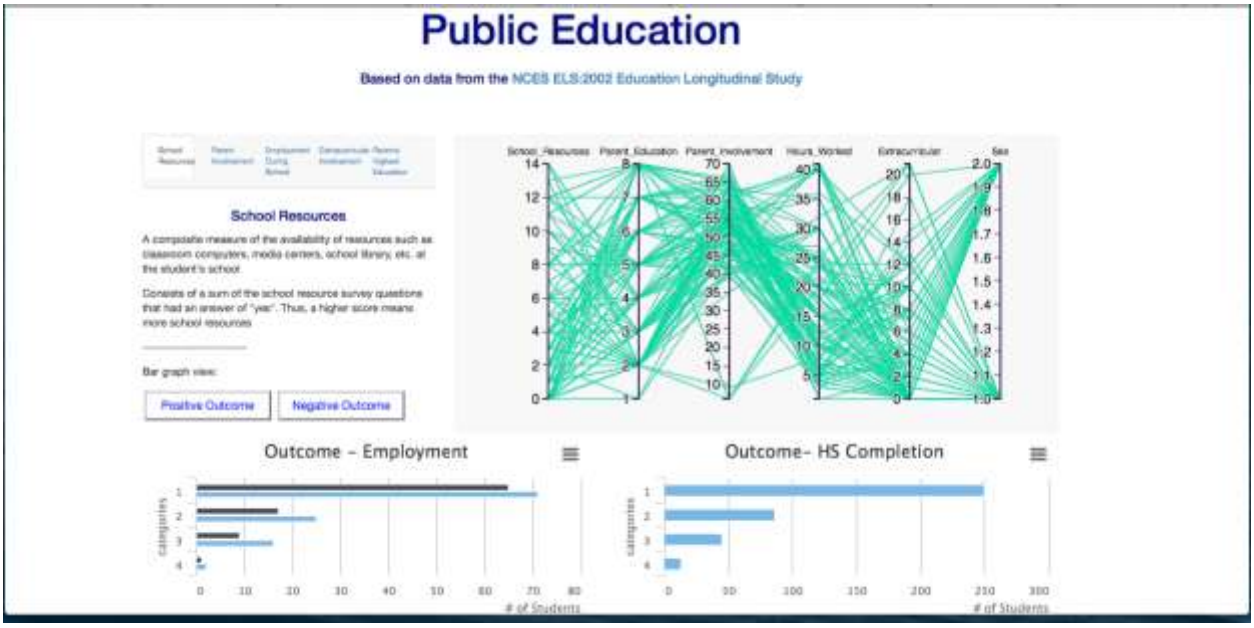
Our data comes from the Education Longitudinal Study of 2002 completed by the National Center for Education Statistics. The study collected data from students throughout their secondary and postsecondary years, focusing on understanding students' trajectories from the beginning of high school into post-secondary education, the workforce, and beyond. The participants in this study were a sample of students who were in the 10th grade in 2002 and/or in the12th grade in 2004. Students participating in the study completed follow-ups in 2004, 2006 and 2012. Data was collected for over 15,000 students in 750 schools.

Because our data was so massive, we had to decide which portion we were going to use to visualize. The biggest filtering we did on the data was to only include the students who had responded to all three follow up questionnaires. We decided to compare six different factors using parallel coordinates: the resources provided by the school, the parents' involvement in their child's schoolwork, the parents' highest level of education, the student's employment while in school, the student's involvement in extracurricular activities, and the student's sex.

The first two factors are created by aggregating the answers to multiple questions that contributed to the factors. The higher the value for those factors, the more resources were provided or the more involved the parents were. The parents' highest level of education is represented as a range, with 1 indicating no high school diploma and 8 indicating a PhD or other professional degree earned. The student's employment and extracurricular involvement are represented by hours dedicated per week. The sex is represented by a 1 for male and 2 for female.

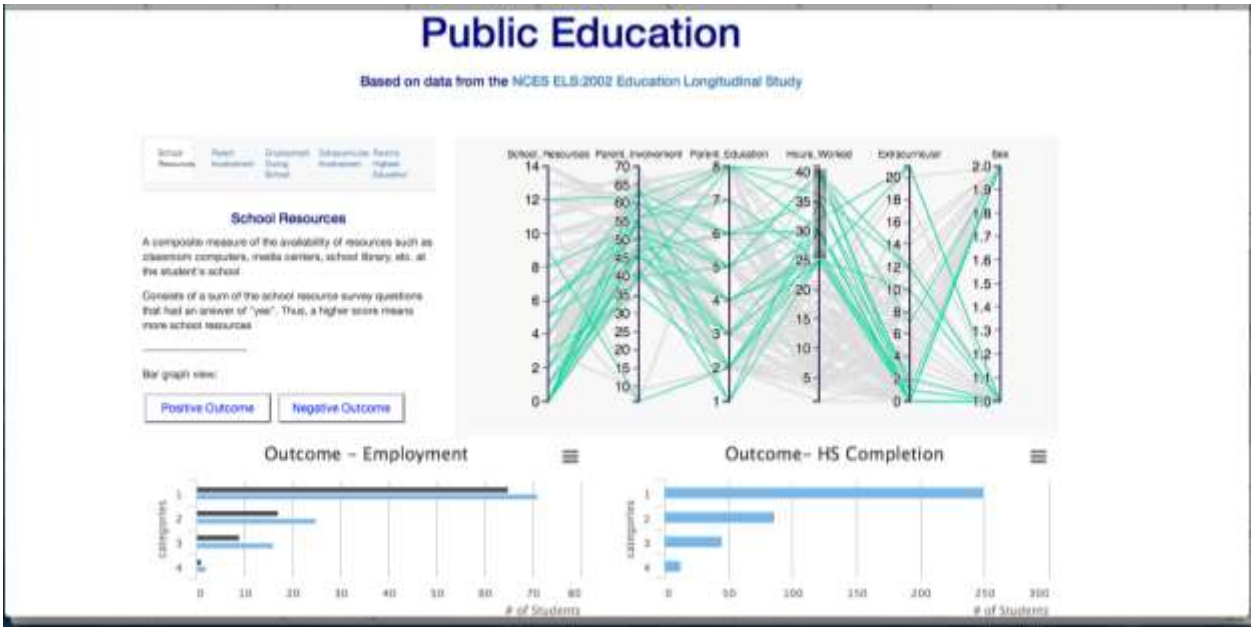
For the bar charts, we decided to look each factor and how it compared to whether the student had indicated in the follow-ups that they were employed and whether they had indicated that they graduated from high school. We divided each factor into four categories by dividing the range equally. For example, when considering parent involvement, category 1 is values 0-20, category 2 is values 21-40, category 3 is values 41-60, and category 4 is values 61+. We did a count on how many students fell into each category for each factor and outcome, and displayed those counts in the bar graphs. For the employment outcome, we display how many students answered that they are employed and how many students answered that they are not employed, and for the graduation outcome, we only display how many students answered that they graduated.

IV. Screenshots of Visualization

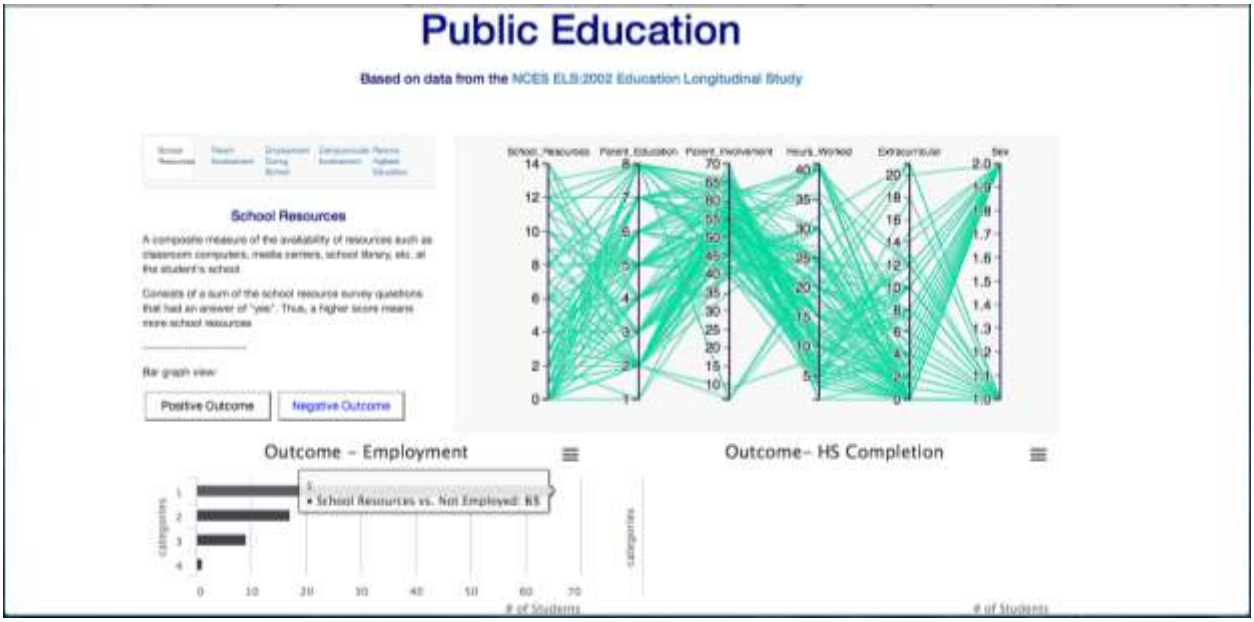


In the screenshot above, you can see the parallel coordinates visualization in the top right of the screen where each student is represented by a blue line and each

coordinate represents a factor, a tabular legend explaining each of the coordinates and their scales on the top left of the screen, a bar chart showing both the students who found employment and the students who did not find employment that are categorized by the school resources available to them on the bottom left of the screen, and a bar chart showing the students who graduated high school categorized by the school resources available to them.

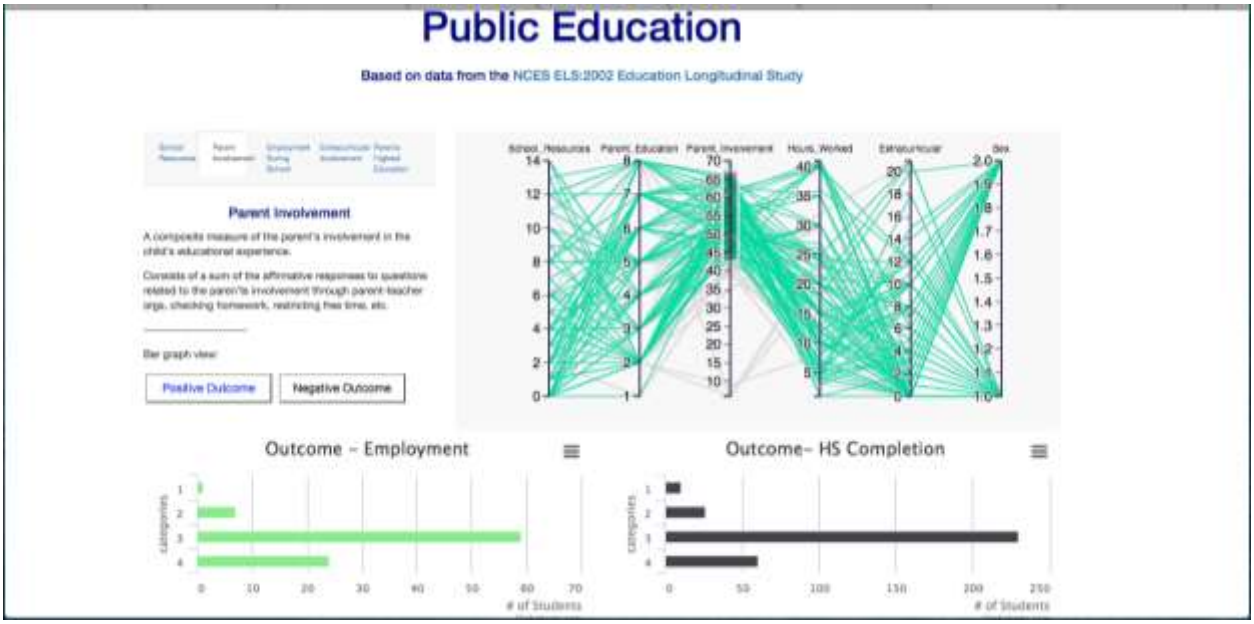


If you compare the screenshot above to the one before it, you will notice that the coordinates for parent involvement and parent education have been switched. All of the coordinates can be rearranged to allow users to quickly investigate correlations between any pair or set of different coordinates. You will also notice that there are fewer blue lines on the parallel coordinates and there are grey lines in the background. A portion of each coordinate can be highlighted to investigate only students that fall in that specific category. Multiple coordinates could have highlighted portions to further filter the students.



If you compare this screenshot to the one above, you'll notice that the bar graphs now no longer contain information about the students who said they found employment nor about the students who said they graduated high school. The Positive Outcome button in the legend was clicked, removing the positive outcome information from both graphs. If the Negative Outcome button in the legend was clicked, the information about the students who said they did *not* find employment would be removed. Each tab in the legend has these Positive/Negative Outcome buttons that will display or remove the information about these outcomes in relation to the tab's factor.

You will also notice that a tooltip has appeared, because the user has moused over one of the horizontal bars in the graph. The tooltip will display which category of students is being moused over, as well as which factor is being displayed and the value of the bar.



The screenshot above shows a new factor, Parent Involvement, being compared the outcomes on the bar graph. In this case, the user noticed that the majority of the students in categories 3 and 4, which indicate the higher levels of parent involvement, said that they were employed and that they graduated high school. The user wanted to know if higher parent involvement was also indicative of anything else, so they highlighted the higher ranges on the parent involvement coordinate in the parallel coordinates visualization. At this point, the coordinates can be reordered to look for correlations.

V. IDE Evaluation

We used three different IDEs between the four of us. Stephen and Pooja used Notepad++, Jay used Sublime Text 2, and Bre'Ana used Komodo Edit. We did not do any version control on our project, so these different text editors suited our purpose. Instead of using version control, we worked on our project using pair programming,

where multiple people worked off of the same computer, or we simultaneously worked on distinct portions of the files and merged the changes into one file.

The most useful functionality that all the IDEs had in common was that they autocomplete functions and keywords and support color coding for reserved words, functions, constants, and strings in multiple languages.

VI. Toolkits Evaluation

We used three different toolkits to develop our visualization. We began the project using only the d3.js library, but we soon decided to incorporate the highchart.js library as well. The d3.js library is a very powerful tool that allowed us to easily create complex visualizations. However, our team suffered from inexperience with javascript, and because d3.js uses javascript shorthand with selections and method chaining, an inexperience with javascript made d3.js very difficult to understand. This led us to look into using other libraries that provided some of the functionality for us.

We continued using d3.js to create the parallel coordinates portion of our visualization and we decided to use highcharts.js for the bar graphs at the bottom of the page. Highcharts.js is a very easy to use library we found that creates sophisticated and interactive charts. It was simpler and we were able to use only one file to add animations and transitions to our charts, which we wouldn't have been able to do with d3.js, based on the example code we found online.

We used Twitter Bootstrap to format the structure of the components on the HTML page. We chose to use Bootstrap because it was more efficient than formatting the page using pure HTML and CSS. The functionalities provided by the Bootstrap framework were easy to understand and implement within our project. Although we

decided to use Bootstrap after we had completed most of the implementation of the components, it wasn't very difficult to implement it into our existing functionality and enhance the structure of the components. The only issue was that the amount of space that we used on the screen was limited.

VII. Team Breakdown and Lessons Learned

Bre'Ana found our data source and she and Stephen worked to remove the portions of the questionnaires that we did not need. We worked together to decide which factors and outcomes to display and how to further filter the data because there was too much data to effectively display. Then Pooja went through the filtered data to generate the values in the bar chart.

Jay took the lead on the coding and Stephen took the lead on researching the different problems that arose while coding. However, everyone contributed to both coding and researching.

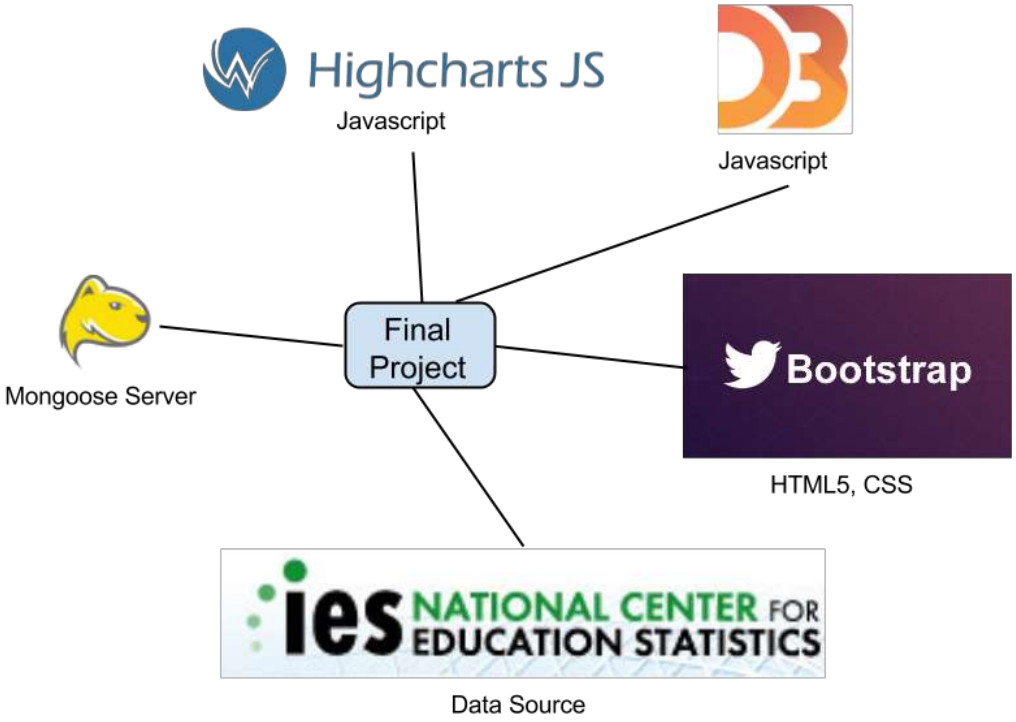
Pooja took the lead on the written portions of the project, both with the progress report and the final report, but again everyone contributed.

Bre'Ana and Stephen also worked together to edit our video, which included taking screenshots and using quicktime to record our information visualization. It also included the voice recording.

The main lesson we learned while working on this project was that someone needed to take the lead in organizing the group much sooner in the semester. Because nobody took the lead in initiating meetings, we started later on the project than we should have. We also should have evaluated the different skill sets each member had before deciding on a software to use. It would have been better to realize how difficult

using d3.js would be for all of us before we had already committed and planned our project around that library.

VIII. Software Architecture Diagram



Because we were so unfamiliar with D3, we found code that created parallel coordinates and only modified it where we needed to for our specific data and for style. We would say around 30% of the d3.js code is our own, so we imported most of it, but then we were able to modify to suit our needs and fit our dataset. We did 100% of the code in Bootstrap on our own, since it was to structure our layouts and we were able to pick the locations of where everything belonged. The code done in Highcharts was about 75% our own code; it was very simple for us to pick up quickly, so we modified it to best suit our needs and wrote more to further enhance our information visualization.