Barriers to College: INFO 4310 Final Report

Melissa Avila, Amanda Chen, Crystal Liu, Benjamin Stevens

Project Goals and Motivation

The main purpose of our visualization is to explore the environment, reasons, and consequences surrounding barriers to college attainment and achievement (and the lack thereof).

In developing this visualization, we sought to explore a few guiding questions:

- Does going to an elite small liberal arts college yield a higher salary on average?
- How might students of color be disadvantaged from their peers in admissions?

As we mentioned in our design document, our primary goals in developing this visualization were to showcase how institutions of higher education differ in their barriers and outcomes for varying demographic groups. We intended to:

- Help readers explore trends in the economics of school choice (endowment, levels of financial aid, expected loan amounts, etc.)
- Allow applicants to identify schools which meet student search criteria across disparate variables (measures of racial diversity, gender distribution, school size)
- Demonstrate systemic advantages and disadvantages afforded to demographic groups

Intended Audience & Insights

One part of the intended audience is prospective college students and their families looking to make an informed decision during the college application process so that they can maximize their values, whether it be career success or student diversity.

Our visualization also intends to address influencers and notable people within the education system such as university trustees, those working under the U.S. Department of Education, and other related players.

One of the key questions we hope to shed light on is: how are certain demographics disadvantaged within the U.S. education system?

In exploring our visualizations, users may come to understand the true extent to which certain demographics are disadvantaged in the school system (namely, black and hispanic students). Users for example may discover that the only two non-majority white schools with a larger than average endowment are University of Florida and Howard University but that, even with their large endowment, the endowments of these schools pale in comparison to those of majority white schools.

Data Sources

Data was primarily obtained from the College Scorecard Data, collected and distributed by the U.S. Department of Education. Data is provided through federal reporting from institutions, data on federal financial aid, and tax information. While most of the dataset had a focus on data concerning Title IV recipients and student who receive federal grants and loans, there is a ton of information we explored regarding diversity, income, and gender distributions on enrollment, post-graduation income, academics, and tuition costs.

Another dataset we used is the Integrated Postsecondary Education Data System, managed by the Nation Center for Education Statistics (NCES). This database allows for filtering by specific schools and variables, similar to the ones described in the previous data set.

Lastly, because of sheer number of schools, we wanted to provide a way for the user to view the data in a less congested fashion. To drastically trim the data, we gave the user to options to view only “Nationally Ranked Schools” using data from this dataset which lists nation and global 2016 school rankings published by the Times Higher Education. From this dataset we used rankings related to quality of education, faculty quality, job employment options, and general national rankings. By using these rankings we were able to allow the user to filter the data in a relevant and interesting manner.

Design Changes from Conception

From when we started this project our biggest design change has, of course, been the topic change. We originally began planning for the project around the topic of comparing Ubers to Taxis and Lyfts and exploring whether it’s possible to earn a living wage from being a driver. However, due to a lack of data and conflicting data sources, we changed to exploring the topic of barriers to education. Originally, we had grand plans for exploring this topic surrounding comparing intersecting variables such as where different demographics fall into different income brackets within each school (to develop more supportive visualizations showing how some demographics are more concentrated in particular income brackets which could consequently affect their educational opportunities); however, due to an overall lack of data with intersectional variables, we were unable to delve as deep into the topic as we had hoped (because of this, we had to cut out our original tree branching diagram).

Implementation

During the implementation of our visualizations, in terms of external libraries, we primarily depended on using D3 and JQuery for the coding aspects and Bootstrap for some HTML structuring. For the purpose of gathering university thumbnails in the final visualization, we used the Clearbit API. Lastly, for efficient data sorting/filtering, we utilized the external library lodash.

Implementation was divided into four major sections each handled by one team member:

- The scatterplot brushing and output
- Scatterplot school details after brushing
- The bubble chart exploring family income by demographics
- The college filtering tool

Project/Video Description

In the first visualization we combine both exploratory freedom and narrative structure by giving users the options to either set the scatterplot axes manually (and explore different factors related to academic diversity and success), or to click pre-defined figures associated with explanatory descriptions. We see the former demonstrated first as the user explores the different axes available for the scatterplot. Each of the 2 axes for each scatterplot can be manipulated.

The video next shows the user exploring the “Show only Nationally Ranked Schools” checkbox. This box toggles between showing every university in the dataset and showing only universities that are nationally ranked by US News. We have the latter implemented as the default because more information is available for those schools and we wanted a way to reduce the confusion caused by the entire dataset while also showing relevant and non-arbitrary information. Since there are multiple variables for which sections of the dataset is missing data, when the user deselects the toggle, a warning message pops up, notifying the user the some of the data may be inaccurate/missing/incomplete. If the user dismisses the notification, it will not appear in future toggles of the checkbox.

The user then clicks on the figure buttons underneath the scatterplots. When pressed, each of these buttons, sets the graph to predefined axes then displays narrative paragraphs associated with and describing them. If the descriptive paragraphs reference other figures within them, then they are hyperlinked to their respective figures.

When points on the scatterplots are brushed, the user can then see details regarding the brushed points (whether it’s a single point or an aggregate) by clicking the “Show Brushed Data” button that appears following brushing the data.

The second visualization is an interactive bubble chart comparing average family income of students at each university to the average first year completion rate by ethnicity. The orange and purple dots represent public and private universities, respectively. Choosing a different ethnicity filter at the top animates and changes the dots accordingly. The average family incomes are binned by the three income brackets: low (0-30k), medium (30k-70k), high (70k+) based on College Scorecard’s income brackets. The leftmost value on the x-axis is N/A, where gray dots lie, representing universities that do not have data on that group of students’ completion rate. The video runs through the above description and goes through the ethnicity filters.

Finally, the video filters on a few variables in the college search section. Only a limited number of variables are shown due to space reasons.

Technical and Design Challenges

The data

The data and visualizations were fraught with more challenges than originally anticipated. The data itself was incomplete as most of the values published in the College Scorecard and NCES datasets were either null or blank. Consequently, we were left without many variables we thought we would be able to use to support the article and narrative. There was also a large absence of intersecting variables and columns (for example, the percentage of different ethnicities that fell into different income brackets) that forced us to throw out our original branching-tree idea and made it difficult to support our assertion that minority students are underrepresented in college due to lower income. Moreover, despite the inclusion of a data dictionary, the College Scorecard variables were often vague and incomplete. For example, were were unsure as to whether the “College Completion” percentages were inclusive of school transfer rates.

The Scatter Plot Visualization

The brushing feature for the scatter plot visualization led to a lot of issues and challenges, some of which were ultimately left unresolved. The main difficulty with the scatterplots was having the opposite scatter plot highlight the respective points brushed in the original scatterplot. In the development of the brush feature led to many event conflicts with hovering (which was ultimately removed as no valid solution could be found to resolve the problem) and scatterplot updates. As a result, the code for the scatter plots is slightly convoluted and plagued by nested functions.

Brushing Panel with Additional Statistics

The main challenge with the brushing panel was processing the average data statistics and choosing which data to show out of the multitude of information we had available. To accomplish this, the code averages, rounds, and formats the data before displaying it on the panel.

The Bubble Chart

There were not that many issues with implementing the bubble chart. The hardest technical part was figuring out a way to filter out null data or represent it well in the chart so that it does not create misleading observations. This was difficult because there would be some columns in the dataset, such as certain ethnicity completion rates, that were not null, while others were. As a result, I could not preprocess the dataset by removing null points because it's possible that another filter would cause the same datapoint to not have a null value. Therefore, I decided to create a column in the x axis dedicated to null values, which would turn gray if null. Other than that, I ran into design issues from deciding what to compare between x and y axes, and the filters. A lot of time was spent tweaking the income brackets to find the most visually significant correlations, and testing different variables to find trends.

College Search Filtering

As stated above, we initially wanted a personalized branching tree approach to make our visualization more engaging. The data didn't support this, so we opted to include a simple filter-by-variable search. There were some formatting challenges, and working with different types of variable proved difficult, so some had to be cut.

Design Trade-Offs

Our biggest design trade-offs consisted of the manners in which we would display our visualization. In the scatter plots, we chose to focus on comparability over details and aggregated data as we plot our data points by university rather than showing overall averages and proportions for different overall factors. Moreover, in our Scatterplot we decided to give the user access to all the drop-down variables rather than a select few to emphasize exploration at the expense of overwhelming the user with data.

In the scatter plot details, because of time limitations, we opted to display scatter plot details in an aggregate form given statistics rather than detailed individual graphs.

For the bubble chart visualization, we decided to include smaller minorities such as American Indian and Pacific Islander rather than limiting our data to the largest minorities because we wanted to emphasize just how underrepresented those smaller minorities are in the majority of colleges (evidenced by their relative lack of data).

Changes or Features we Wish We Could Add

Some of the features we wish we could have added given that we had more time/data

- Hovering on the scatter plot visualization brings up a tooltip with school-specific information
- A branching tree diagram that provided the user with school recommendations based on their demographic characteristics
- More complex diagrams comparing the best schools with the worst schools and the factors that may have contributed to their ratings
- More exploration into progressive vs non-progressive universities
- After data is brushed having more complex visualizations appear

Team Contributions

1. Melissa Avila
   a. Scatter plot + Brushing Capability + Scatter plot Narrative features/figures- 12 hours
   b. CSS and HTML general structure/design - 1.5 hours
   c. Misc narrative elements + documentation- 3.5 hours
   d. Final Report - 1.5 hr
2. Crystal Liu
   a. Bubble chart - 10 hours
3. Amanda Chen
   a. Brushing data detail panel - 8 hours
4. Ben Stevens
   a. Attempt at branching tree vis - 2 hours
   b. Achievement gap research - 3 hours
   c. Minor overall aesthetic and bug fixes - 1 hour
   d. College filtering - 8 hours
   e. Video prep and recording - 2 hours

* This visualization can be used in future interactions of the class