Final Project Report

Project Link: https://are-ivies-worth-it.netlify.app/
M1 Link: M1
M2 Link: M2

Project Goals and Motivation

The goal of our project is to compare the Ivy League schools with the other top 100 universities in the United States. We aggregated data such as admission rate for each school, the average cost, the median amount of debt accrued by graduates, median earnings of graduated students, and so on for both the 8 Ivy League schools as well as the next top 100. We thought that it would be interesting to look at these factors as well as a variety of others from the dataset in order to decide if the level of education received at the Ivy Leagues – institutions known for their prestige – is worth the hype. We will also be looking at factors that signify diversity like the race makeup of undergraduate students and the retention rates of first generation students.

The motivation for our project actually came from assessing Cornell's famous "any person, any study" motto. At first, we wanted to dive deeper into the diversity makeup of both Cornell as a whole as well as the demographics per major. However, we soon realized that the data for this would most likely be hard to find as universities tend not to make this sort of data public. Thus, we decided to expand the scope of our project to include all ivy league schools and then thought a comparison to other schools across the United States would put that data into perspective. Similar to our original thought process, we simply want to provide the data to make comparisons among the best group of universities and see where/if any major discrepancies are found. The user will be able to explore this data with the guide of our visualizations and make decisions for themselves.

Intended Audience and Use Cases

The intended audience for this project are prospective students as well as their guardians, parents, family, etc who want to learn more about the top universities in the United States. These students are in the process of, or soon will be, applying to college and want a holistic overview of the top 100 universities in the United States as well as the data associated with each college. Instead of having to look up each university one by one, this news article will give the user the opportunity to explore all of them at once in the same place.

Imagine a student who is in the top 10% of their respective high school graduation class: they know they have what it takes to get into a prestigious university. However, they have no idea where to begin, or which colleges to explore. The overview presented in this editorial will

provide standard information like the cost, admission rate, median income upon graduation, and average debt accrued so that the user can have a solid starting point for their application endeavors. In addition, users could also be students who have recently been admitted to several colleges, and want to compare their options. Either way, the tool can be heavily used by prospective students in aiding them to be knowledgeable about the top universities: not just the well-known Ivy League but perhaps lesser known schools that can provide an education of a similar caliber.

Related Material

An example of a related tool that attracts the same audience of prospective students is the online website Niche.com. The website takes information from various different colleges and institutions and ranks them in order as well as gives them "grades" for particular subsections such as diversity, sports, food & resources, etc. On this website, you are able to view the rankings of schools and universities as a whole as well as individual schools and their respective data. Each school will also have an "Overall Niche Grade" such as an A, B, etc that captures how good the school is according to the Niche methodology. We wanted to follow some of the same ideology behind this website by offering users a view of the top schools and their respective statistics. However, we wanted to leave out the "grading" part and allow users to make these educated analyses on their own. Since we have already narrowed down the pool of universities to only the top 100 in the United States, we figured it would be best for users to make their own assumptions and rank universities according to their own personal wants, needs and goals - whether that be low cost and debt or guaranteed high earnings after graduation.

Data Sources

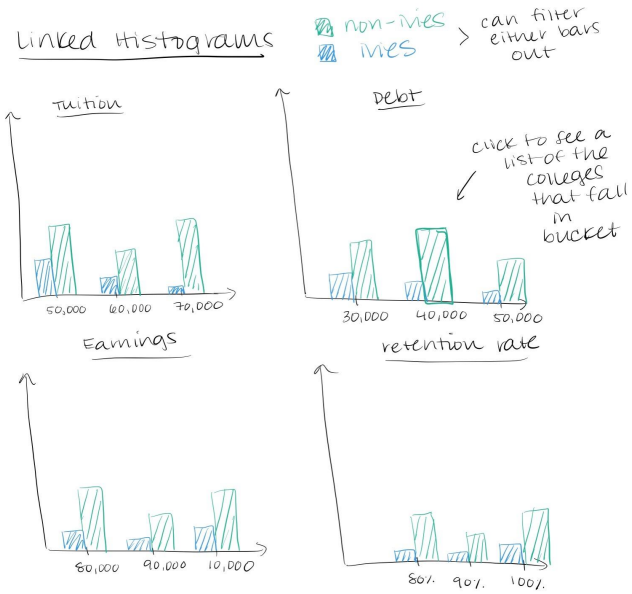
The data comes from the U.S. Department of Education College Scorecard. It includes Institution level data files from 2023 including variables regarding student completion, debt, repayment, earnings and more. To prepare the data, we used a python notebook to select only the institutions in the US. News and World Report Best National University Rankings top 100. We scraped the names off their webpage and then compiled two CSVs, one that includes data of only the Ivies and one that contains all the rest of the top 100 universities (so a total of 92 schools in this dataset). These prepared data files can be found here:
https://github.com/DGR40/INFO-4310-Final

Design Iterations

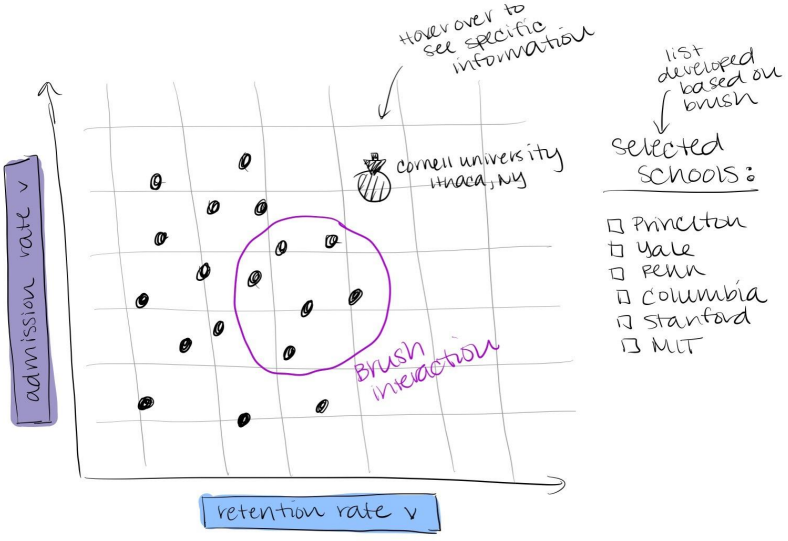
Preliminary Designs:

In Milestone 2, we designed what ended up being our preliminary sketches for the design of the visualizations that we would include in our news editorial. The first visualization that we

sketched was a set of four histograms that visually are the exact same but differ in the independent variable along the x axis. These histograms would be color coded for Ivy League schools versus non-Ivy League schools so that the distribution of each of the two separate groups could be shown simultaneously. In terms of interactions, the user is able to display only the Ivy League schools or the non-Ivy league schools and have the other bars momentarily disappear, making it easier for the user to gather the respective distribution of each group. In addition, the user can hover over any given bar and see a list of all of the schools that fall into that specific bucket. We thought this would allow the user to perhaps pick a range of values and subsequently view all of the schools that meet this criteria.



The second visualization in our preliminary sketches was the interactive scatter plot. This scatterplot would plot the four variables listed above in the histograms, with the ability to interactively change the axis. In addition, other interactions include a brush interaction to select a specific group of points and hovering over a circle to view more specific information on a school.



Intermediate Designs & Design Process

After the initial sketches, we began the implementation phase of our project - attempting to actually build what we initially sketched. The histograms ended up being fairly similar to our initial design, with two substantial changes. In our initial design, we placed the Ivy League and the non-Ivy League bars next to each other. However, whilst implementing, we realized that with bars next to each other, it is hard to see the fluid distribution of each group. Therefore, we decided to put the Ivy League bars *on top* of the Non-Ivy League bars since the Ivy League dataset consists of only 8 schools. This way, you are able to see the large distribution of Non-Ivy League schools, a set of 92, with the Ivy League school distribution placed on top. We believed that this design decision would also allow the user to make comparisons between how many Ivy League schools fell in a specific bucket versus Non-Ivy League schools. Secondly, in our final design iteration, we decided to add in a fifth variable to the histogram diagrams: retention rate. At first, we overlooked this variable as not being critical to our analysis; however, once we plotted the variable on the histogram we found an interesting trend in the distribution and thus thought it would be worthwhile to show to the reader of our editorial as well.

The evolution of our design for the second visualization ended up looking significantly different from our original layout. Instead of showing a list of schools that would come from brushing over a specific subset of data points, we decided instead to implement a “college comparer”. With this comparer, the user is able to hover over any circle on the scatterplot and view its corresponding data in a panel on the right hand side next to the visualization. If the user clicks on a specific dot, it will be locked in as one of the colleges being compared. Next, the user can continue to hover over more dots to choose a second school; once again, a click interaction will lock it in as the second school. We also give a numerical difference between the values for each school. Ultimately, since our tool is all about making comparisons, we wanted to make it easy for the user to objectively compare the statistics of two separate schools. The brush tool we

had originally proposed did not seem to afford this comparison as well as the final design we implemented.

We also had various ideas for how we should incorporate diversity data into our project as well. At first, we did not include this in our initial designs at all. However, as we began to think about our project and its goals, we figured that diversity was a key factor in determining the richness of an institution. We discussed including diversity statistics in the scatterplot, but thought that might be overwhelming to have that many options for each axis. We also liked the clear link between the five variables of the histogram to the five variables of the scatterplot. Ultimately, we thought it would be best to include the diversity data as part of the detailed view that occurs when a user clicks on a specific school. We initially thought that a pie chart, as opposed to plain numbers, would be the best way to represent diversity at each school. However, as we tried to implement the pie chart, we ran into a few setbacks. At first, we wanted to implement hovering over the pie chart that would give information as to what demographic corresponded to each slice of the chart. However, we decided that hovering over each slice to see the label was making the user do too much work to discern labels of slices. We also tried to implement labels on the pie chart, but we were having difficulty implementing them as the space for the pie chart to occupy was not big enough to compensate for the labels.

Ultimately, we settled on a bar chart in order to represent this data. We feel as if a bar chart is the easiest way for users to compare the relative lengths of each bar to another school. In addition, we decided to toggle between the statistics from the 5 variables and the race demographic information. It was too much information to squeeze in the space available when we tried to show both at the same time. In addition, having only the bar chart for race demographics actually makes it even easier for the user to compare the two graphs directly in line with one another. Lastly, we used a common scale on the y-axis between every race demographic bar chart to make comparisons much easier as opposed to an ever changing y-scale that adjusts to just that one college's data. This standardizes bar sizes and makes it very easy to see the differences in percentages between different school demographics.

Final Design

Final Design

In the final iteration of our design, we implemented the evolved version of each of the visualizations: the histograms and the scatterplot. In the final version of the histograms, there are 5 individual histograms, each corresponding to one of the five variables that we explore throughout the editorial: average annual cost, median debt, median earnings, admission rate, and retention rate. The different distributions for Ivy League versus non-Ivy League schools are color coded and the bars for the Ivy League distribution are pasted on top of the bigger distribution. In addition, the user can hover over each bar in order to view a list of the schools that fall into each bar domain bucket. This is helpful because the user can begin to curate a general list of schools that they may be interested in based on the properties of their values on the histograms.

In the final version of our scatterplot, we included the five variables from the histograms as the variables on each axis. The user is able to switch the axes as they please using the arrow buttons in order to view relationships between two or the same variable. The biggest component of this scatterplot is being able to use the college comparer function. The user can hover over the points on the graph as they choose, and see the information for the highlighted school. However, once they click on the specific school it will become a part of the comparison. The user selects two schools, and the differences between the variable values will be highlighted in red and green respectively, depending if the value is lower or higher. In addition, there are delete and clear buttons for the user to remove their selections and start over with the comparison. Lastly, the user is able to toggle between the variable data and race demographic data in the side panel as well.

Feedback

The scope and design of our project really evolved a lot due to the feedback that we received throughout the entirety of the project from both the professor and the TA which was extremely helpful. In our first milestone submission, we outlined a much more narrow topic of just focusing on Cornell University and its "any person, any studio" motto as well as comparing this data to that of its other Ivy League counterparts. In our original sketch, we included a rough outline of being able to compare two colleges side by side - something that we ultimately turned back to in our final design. However, after receiving feedback on the Milestone 1 submission we realized that finding the diversity, gender, and admittance data for each specific major at Cornell would be quite challenging given that universities usually tend to keep any data that would make them look bad private. Thus, we decided to rethink our topic and broaden the scope to include the top 100 universities in the United States. This time, we would compare the subset of the Ivy Leagues with the rest of the top 100 schools. As it turned out, the dataset for this project was much easier to find and publicly available on the US Department of Education website.

Next, we also got some solid feedback from the professor during office hours, specifically related to the content that we would show in our visualization. At first, we were not sure if we should limit ourselves to just the five variables that were present in the histograms and the scatter plot axes, instead of trying to include diversity data in the visualization as well. However, when we thought about the factors that made a university one of the "top", the diversity of the student body definitely played a significant role. Moreover, we received feedback that including diversity data would be a much better use of our time rather than trying to implement other technical features as it contributes to the overall richness of the editorial piece.

Trade-offs

As with any visualization design project, there are often tradeoffs that we have to make and subsequently accept in our final design. The several listed below are not an exhaustive list, but rather ideas that came to mind when we were reflecting on our final design prototype. While some were conscious design decisions, others were limited due to the scope and lack of time to implement certain features.

- First, there is no zooming capability on the scatterplot. In our project journey, we decided to make a tradeoff between using our time to focus on including the diversity data rather than implementing the zoom feature. However, if we had more time we believe that this feature would be helpful to the user. Since the graph takes up basically the entire screen and dots are placed all over, zooming and panning would allow the user to see all of the dots more clearly without overlap (overlap is especially prevalent when the user is plotting the same variable on both axes).
- Another tradeoff is the fact that the user can only compare two schools at a time. We consciously chose to only compare two schools at once because we believed that would be the best way for users to make direct comparisons. If the user needed to compare 3 or 4 schools at once, this might be too much information at once. In addition, being able to have the information for both schools right next to the graph was useful because the user could hover over schools and click specific schools, all while having the information update right in front of them.
- Another tradeoff we made was that in our design, there is no perceptual link between the histograms and the scatterplot besides the use of the same five variables. Again, if we perhaps had a little more time, it would have been neat to implement some sort of interactivity between the two graphs. A potential idea for such an interaction that we came up with during a brainstorm session was having the specific schools of each bar that was hovered over also be highlighted in some way on the scatterplot. The user could hover over, or click on, a specific bar that would show the schools within that bar, and also highlight their location on the scatterplot.
- Lastly, another tradeoff that we discussed was in regards to showing the race demographic information and the other school statistics both at the same time, or toggle the two views. We ultimately decided on toggling the two views because it was too much information at once for the user when we attempted to squeeze both sets of data on the same information card for each school. With the toggling views, however, there is the tradeoff that you are not able to see the general overview of all the information at once. However, we believed that this tradeoff was necessary to prevent information overload, and to make it easier for the user to compare both bar charts without having other text in the way.

<u>Visual Channels/Metaphors</u>

Histograms:

- Visual channels – positioning of the bars on an aligned scale, length of the bars on an aligned scale, distinct color hue of Ivy versus non-Ivy bars
- Interactions – hover over each bar to see a list of schools that it contains

Scatterplot:

- Visual channels – x axis positioning on an aligned scale, x axis positioning on an aligned scale, size corresponding to how large the school is, color hue corresponding to whether it is Ivy or non-Ivy league
- Interactions – hover over points to see more detailed information, click on points to add that college to the comparer, click arrows to switch the axes variables, click on the "x" or "clear" to get rid of previous college comparing selections, link to toggle between race demographic information and statistics from the graph.

Implementation

In terms of the actual implementation of the project, in general our code employs heavy use of a lot of different functions that divide the code up depending on which visualization it pertains to. We have one big asynchronous function for the entirety of the script section that handles loading the data and creating the subsequent visualizations. We also included a bunch of helper functions outside of the asynchronous function to help with things like making comparisons, finding data, and keeping track of the school comparer.

To implement the histograms specifically, we created a function called "makeHistogram" that would take in the id of the svg element, the respective data attribute, and the title that should be displayed underneath the histogram. This was an easy way to create the same graph several times with peak efficiency. In addition, we implemented another function called getSchools that would loop through the data and create a list of schools that needed to be displayed for each bar on a case by case basis as the user hovered over bars. This was a fairly simple way to retrieve the data behind each bar so that it could be displayed for the user.

To implement the interactive scatterplot, I first drew the circles with a fixed x and y variable. Then, when the user clicked the arrows on either axis, I incremented/decremented the variable index and adjusted each circle's either x or y coordinates to the new scale that corresponded with the new variable. This was more efficient than redrawing the scatterplot each time an axis was changed.

To implement the college comparer, I needed a global variable to help me keep track of what part of the sequence the user was in, this variable was called numSelected. This kept track of how many circles were currently selected by the user. To show the first college, I encoded in attributes of each circle all the relevant statistics I would show. On each hover when no school is selected (clicked on), the first inspector is updated freshly. Once one school is selected, upon hover, the second inspector is updated. Finally, once two schools are selected, the compare function runs, which just takes in a type of statistic and then spits back the relative numerical difference of that statistic between the two selected schools. Additionally, I wrote a makeGraph function that makes the race distribution histogram each time a school is hovered on and sets its display to none. Then, when the user clicks the "View Race Demographics" button on the card of a selected school, I simply hide the stats table and show the histogram. Lastly, I added delete

buttons that clear both schools and delete individual schools (individual schools must be deleted in sequential order, second selected school deleted first, then user can delete the first).

Contributions

Isabelle

- Preliminary brainstorm/ idea generation meeting
- Attended feedback session (in class) for Milestone 1
- Sketching of preliminary ideas
- Contributed to overall layout/flow/styling of the editorial
- Implemented the bar chart & its associated interactivity (displaying college names when hovering over bars)
- Wrote text in between visualizations for editorial
- Worked on project presentation for Demo Days in class
- Contributed and edited the final report
- Final look over project - make sure it is good to submit

Danny

- Preliminary brainstorm/ idea generation meeting
- Sketching of preliminary ideas
- Contributed to overall layout/flow/styling of the editorial
- Implemented the scatterplot & its associated interactivity (switchable axis, hover over points, college comparer, reset button)
- Worked on project presentation for Demo Days in class
- Made sure the styling was consistent across the report (especially with styling of the side panels for information)
- Contributed and edited the final report
- Final look over project - make sure it is good to submit