**INFO 4310: Final Project Report**

**An Exploration Tool that Analyzes University Major, Demographic, and Expected Income**

**By: Christopher Sun (ces334), JJ Li (lbl53)**

**Render Link:** https://four310-final-ces334-lbl53.onrender.com/

## Project Goals and Motivations

When thinking about how close we were to graduating Cornell, one of the first things that came to mind was the career that awaited us when we left. It would be a whole different environment and a start of a new chapter in our lives. As we thought about this, we thought back to how impactful career prospects were when we were applying to colleges and choosing between them as prospective high schoolers. This is what led us to the idea of analyzing the different demographics and factors that can influence our careers right out of college. Money is an important aspect in everyone's lives, and discrimination is an obstacle that we all know exists and may influence this aspect. At first, we wanted to analyze these factors that influence early career income, whether it be someone's gender or race, or if their major was the determining factor, or if there wasn't a set of influencing factors at all. As we continued our research and creation of visuals, we saw that there were some influential factors, but there were many different "lessons" to be pulled out from interacting with the visualizations. For example, one thing we learned that was put in the website itself was how on the scatterplot it seemed that colleges with male majority students made more in their early career than colleges with female majority students. Of course this doesn't have to be the only factor in determining their income, which is why we allow users to explore other options to see if they can come to a different conclusion themselves.

As we thought about it more, our original goal didn't really change, as we still wanted to inform and show our users that different factors do indeed affect our income right out of college, but we then branched out to another goal. We wanted to help people explore these different factors and see for themselves whether these factors truly influenced how well you would do out of a specific college. This means we wanted to help them find a school where they think they would be perfect for, but we also wanted to help them find a school in which students could be allowed equal opportunities to succeed. Thus, our two main goals were to inform our users of the different factors that influence their career outside of college and to help them explore colleges to help them find the best fit for them.

**Intended Audience and Use Case**

Our intended audience, with these goals in mind, would mainly consist of high school students who are close to graduating and are still in the middle of their college applications. Because our other goal is just to inform people, the audience could expand to almost anyone—ranging from parents to college consultants to other curious parties—to show them how diverse and different these schools could be, but the high school student audience would fit both goals very well. These students would be able to analyze schools they are interested in to predict how well they would do in the future, from the career income analysis, and how well they would fit in, from a demographics analysis. They would then be able to compare and contrast two different schools more closely in the spider graph, to see if one is truly better than the other or if they're in the same standing. They would then finally be able to take a look at how different majors influence income and help them decide what major they would want to pursue.

One visualization that we referenced, as well as used a data source from, was from a link that was posted on Ed Discussion. This link led us towards a visualization displaying different majors and their salary scales as well as a smaller division of demographics. Although we didn't reference this visualization until we were quite deep in the project, it helped us brainstorm different visualizations that we could add on to our current analysis and therefore create a deeper analysis and exploration of the colleges.

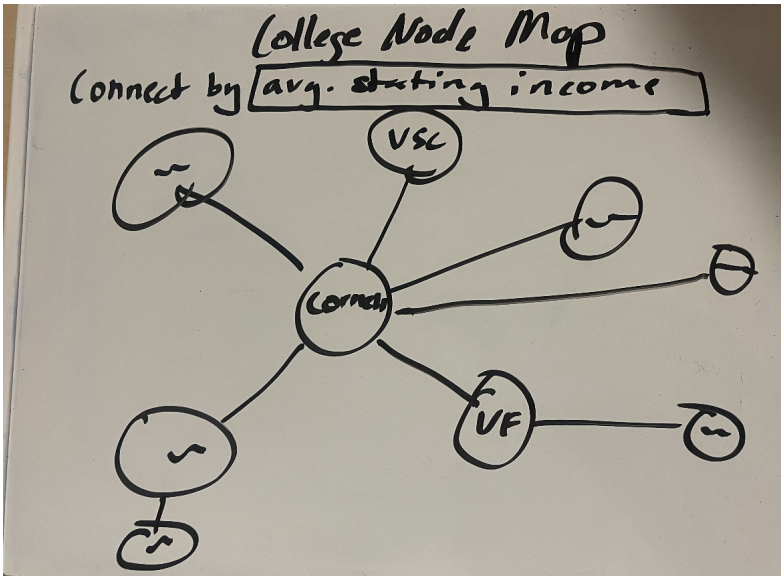
**Data Source and Processing**

We have 3 main data sources used in this project. Our first data source consists of longitude and latitude coordinates of colleges all around the United States. This was found by searching through datasets that had json or geojson data for colleges. We ended up finding a dataset from the National Center for Education Statistics that contained a lot of information, but we only needed the longitude and latitude information. In order to find our main dataset, which consisted of the demographics of colleges as well as predictive career pay, we did some Googling and managed to find the perfect dataset through Kaggle. It contained the division of demographics for both race and gender for many schools around the United States and had predictive career pay. The final dataset we found was from the visualization that inspired us, from the link from Ed Discussion. This visualization used a dataset that also had exactly what we needed. This dataset was from Kaggle as well. For the final dataset, which contained information about majors and their respective incomes, we didn't have to do much processing and could use the dataset as is. For the first and second datasets, though, there was a lot of processing to be done.

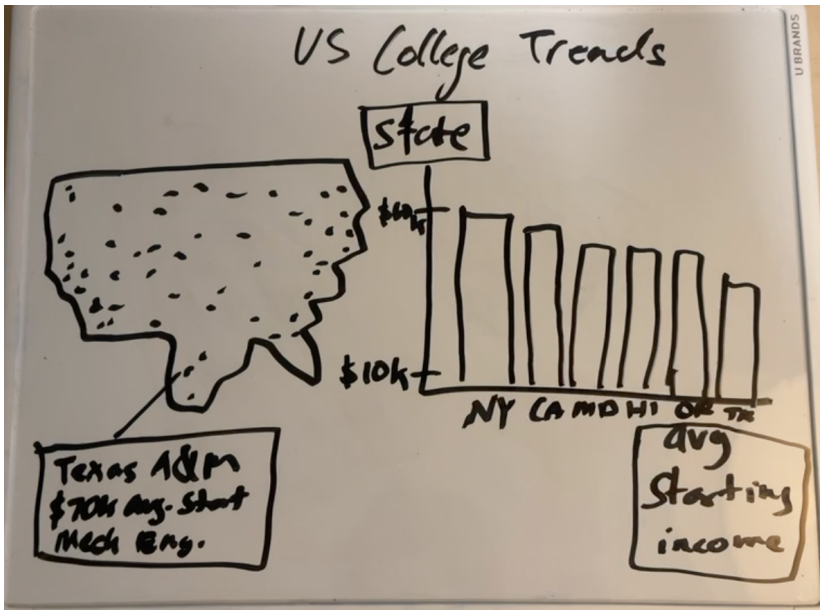
The first dataset contained a lot of information that was unnecessary for us, so we had to filter out almost everything from it except for the longitude and latitude of colleges as well as what state the college resided in. The second dataset was also problematic, as it contained several different csv files that had information we needed, but the number of colleges across them weren't consistent, meaning that there was some information left out for colleges we were displaying. In order to fix this problem, we took the first dataset and second dataset and combined the information from both. This process consisted of filtering out all colleges from the second dataset that didn't have a longitude and latitude in the first dataset. The next step was to then filter out all colleges from the second dataset that didn't have information throughout all the different csv files. This finally led us to a final filtered dataset that had coordinate information, demographics information, and predictive pay information, along with other lesser information of around 600 colleges, filtered from a thousand colleges from the location dataset and around 700 to 800 colleges in the demographic dataset. In our actual dataset folder for our project, there are a lot of "final" datasets, that resulted from a step in filtering two or more files. The steps were difficult to do from the base datasets and therefore fewer filtered out datasets were needed to create the finalFiltered.csv dataset. The other final datasets or similar name copies of datasets were used as a stepping stone for the final.

**Interaction Storyboards and Sketches**

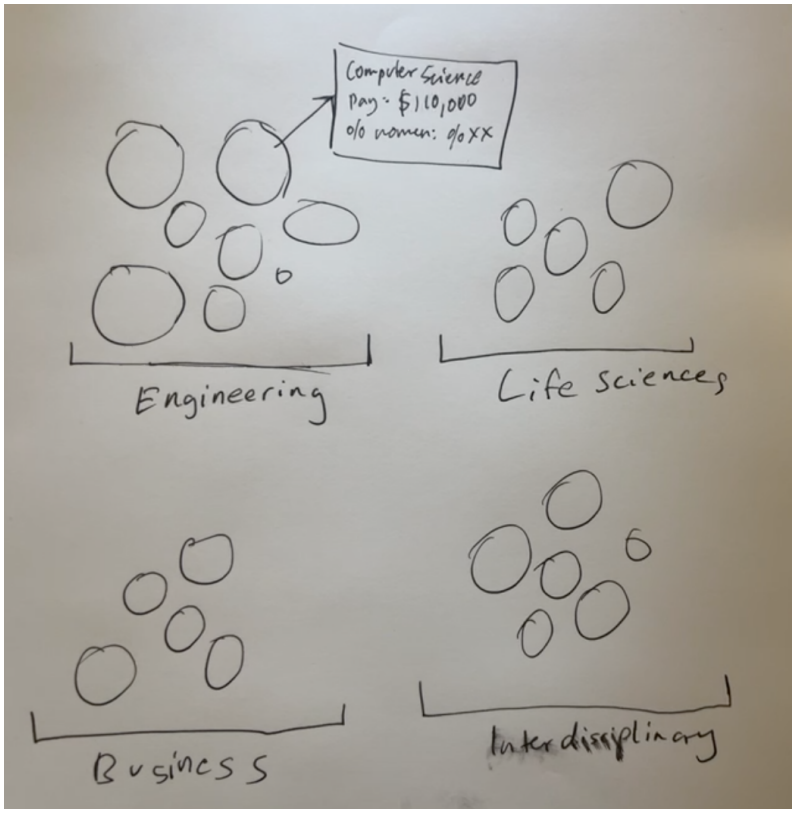
As we brainstormed and discussed the best ways to represent the data and also allow the user to explore the data to find their own trends, we thought through a few visualization designs and iterated on them. Among the different visualizations, we considered a node map, geographical map, stacked graphs, scatterplot, among others.




We first considered a node graph that would connect the different colleges based on the criteria that was selected (e.g. location, average starting income, diversity, etc). We experimented and talked through this visualization but ultimately came to the conclusions that this was not a good representation of how colleges in the US are related. Colleges are not connected by certain demographics but rather across a multitude of factors combined. Thus we forgoed this visualization.

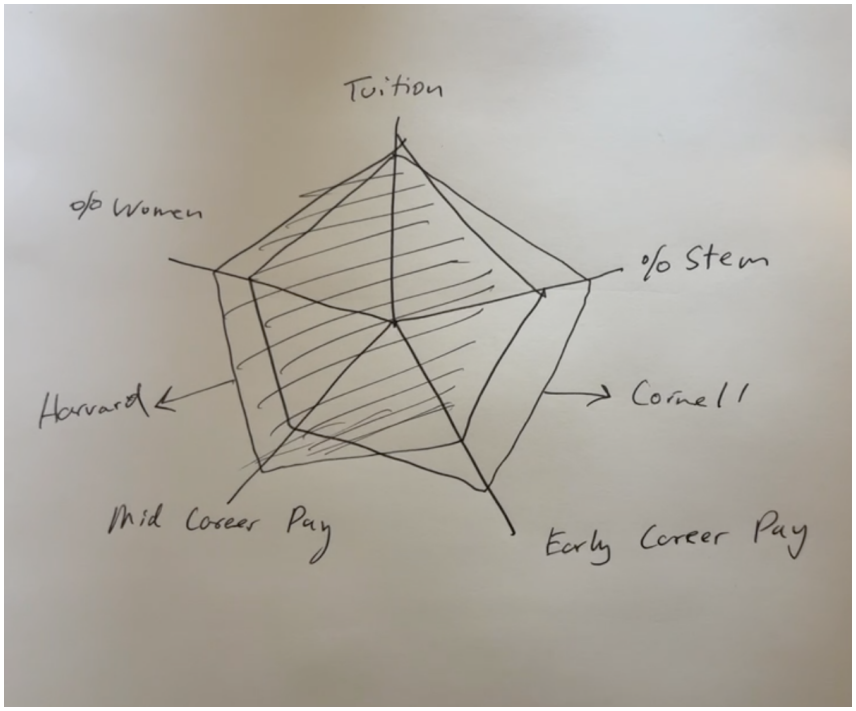


The visualization that we ended up going with was a geographical visualization as well as an interactive dynamic two-axis chart. We believed that this captured most of the ways people thought about colleges in the US and thus our interactive visualization would allow them to analyze trends holistically and from different angles. The map allows users to skim colleges across the US from a high level view and notice trends using visual cues such as color scales and legends. The chart then allows the user to break down different aspects of colleges and compare them against another aspect in depth. Using this visualization, a prospective high school student, college analyst, and parents can get a general view of trends across various colleges in the US.




After we completed the geographical visualization and two axis chart, we realized that we now were able to allow the user to explore colleges and trends based on location, tuition, demographic breakdown, and expected income. At this moment we realized another major consideration that students had to make when they decide to attend college, their major. As once

prospective college students ourselves, we had two big considerations when making the daunting choice: which college to go to and what to study/major. So we thought of different ways to showcase different majors, their associated expected incomes, and the typical demographics that participated in each major. We thought through different ways to visualize the data and tried bar charts, node maps, and even a dendrogram. We ultimately arrived at the conclusions of a categorized bubble chart by major type. We reached this conclusion from thinking through the different ways majors existed at universities; they were often separated by type/school (engineering vs business) and were not particularly connected in any other meaningful way. With the categorized bubble chart, users can target a specific category (e.g. social sciences) and explore all the different majors and associated statistics with each major within the category. We believe that this is the most effective way to show this data as most students have one or two general directions/categories that they are interested in and will likely explore within those categories.



Finally, after all of the high level visualizations we provided, we wanted to give the user a very granular way to compare colleges against each other. We wanted to provide a visual way to compare colleges against each other across their different attributes. After looking through many different types of visualizations, the one that fit the need of comparison the best turned out to be a spider graph that could overlay two colleges on top of each other across 3 or more axes. We found the spider graph to be the most effective visualization because it was able to include numerous attributes/axes and display all of them at the same time while being able to directly and visually compare one college on top of another. Such a visualization allowed the user to make as direct of a comparison as possible and eliminated the need for the user to bridge the comparison in their mind when they used other visualizations such as comparing two pie charts.

**Final Design Feedback and Tradeoffs**

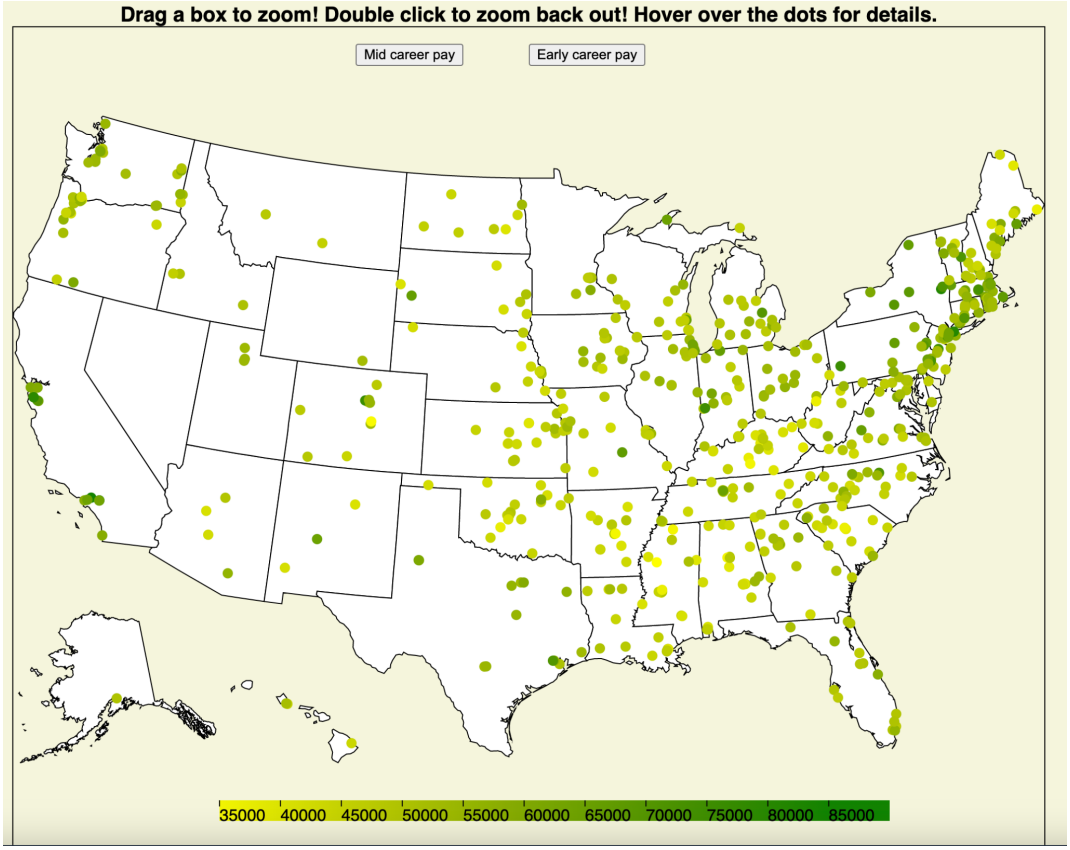
The primary feedback we received on our final design was that while the visualizations were great and there were interactions that had good depth, the overall styling and user interface needed improvement to make the webpage easier to navigate and the overall user experience better. We were able to take the feedback we received and make stylistic improvements that made the page more visually appealing and make the interactions smoothers. We implemented multiple colors such as changing the background page color to make it easier on the eyes as well as using pastel colors on our scatterplot to make the differentiation visually appealing . We implemented a scroll feature that prompted the user to scroll through the page by making the following text appear in light gray to indicate that there was more information below. We also made formatting improvements such as formatting the Early and Mid Career Pay to currency ($) format and the Stem Percentage to percent (%) format.

We tried our best to make the webpage all-encompassing and reduce the number of tradeoffs by providing both high level visualizations as well as detailed visualizations. The main tradeoff that our webpage has is that it does not include all the aspects regarding colleges. As we explored adjacent datasets to the ones we included, we realized that there were many other aspects of colleges such as the weather and if they were in a major city or not that we did not include. However, we believe that we included the most important factors of colleges that go into consideration for a prospective high school student. The other tradeoff that we made was the decision to focus more on colleges and less on majors. While there is the cluster bubble chart that allows users to analyze different majors and their respective statistics, we chose not to dive any deeper into visualizations on majors so as to not overwhelm the user with information. We could have created another spider graph or visualization to allow the user to compare directly between two different majors but chose not to.

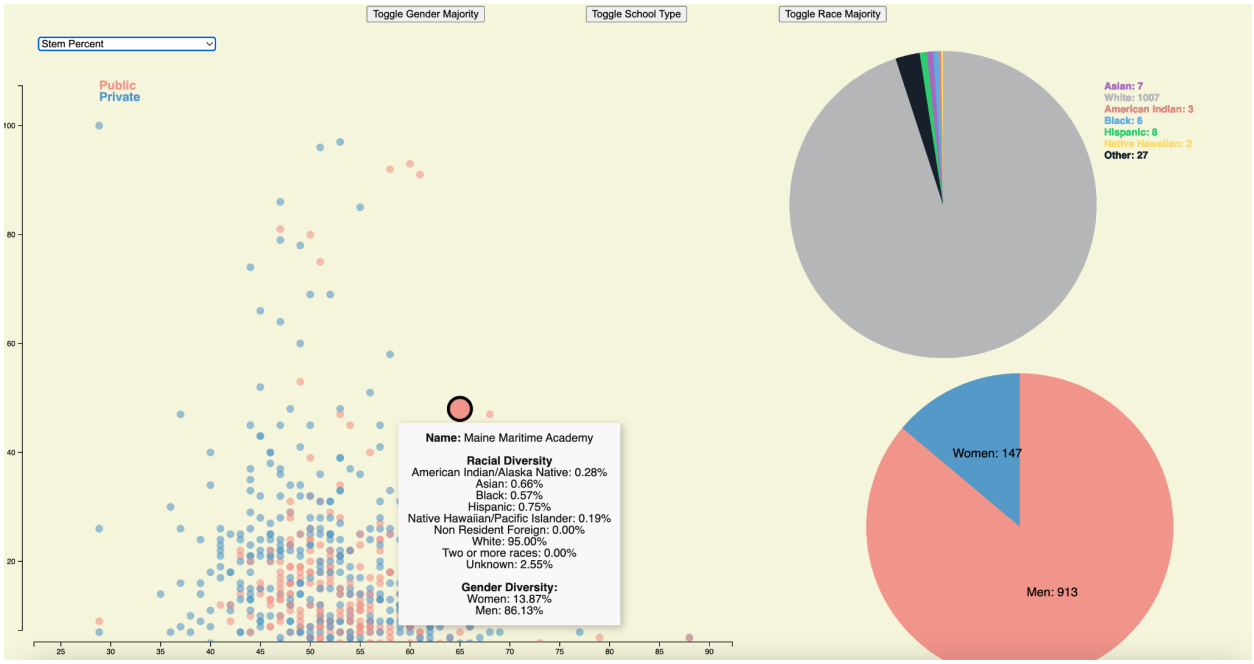
**Final Design Implementation**

The first part of our final design was the base U.S. map which is a high level overview of all the schools in our data set. In order to implement this we got the longitude and latitude of all the schools needed and turned them into coordinates, plotting them on a geoJson map of the United States to plot them. We used some of the data from our final filtered dataset to be displayed on the hover of each point and then to create the filter buttons we created a legend and a color scaling that scaled to the predicted pay for each. Since we wanted this to be our high-level, intro overview, we just wanted to display all the schools with minimum filters and

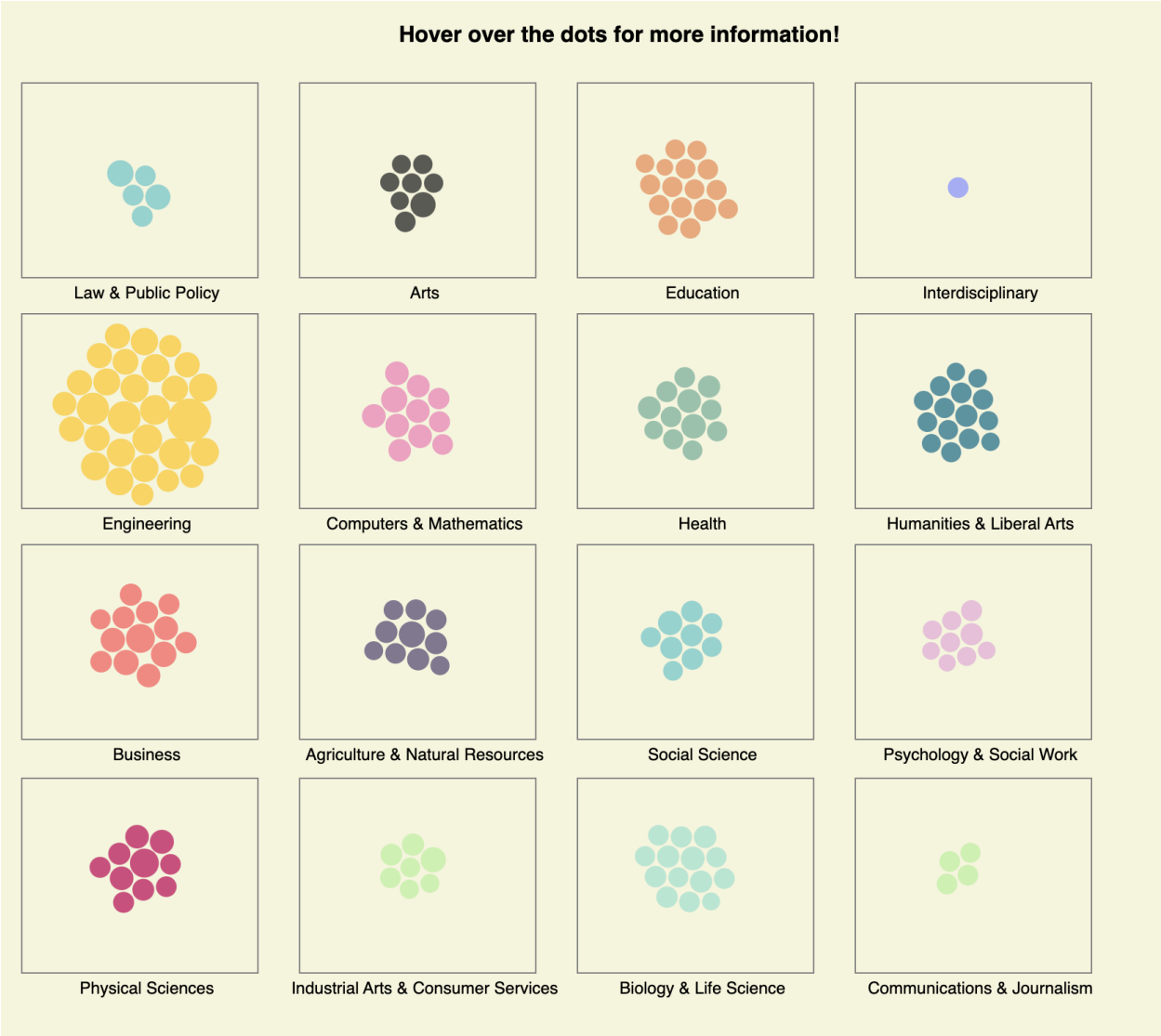
some basic information for users to explore.



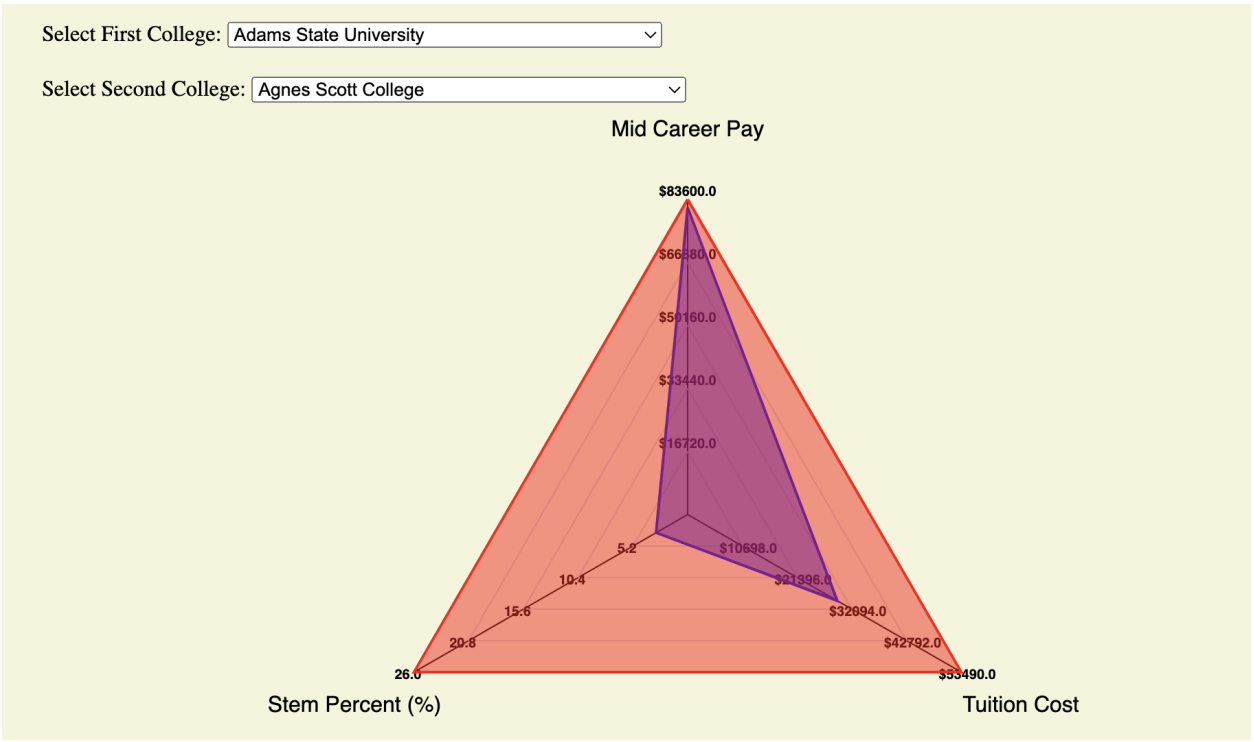

The second part of our final design was the scatterplot that went more in-depth with the data we used, being able to compare and contrast different variables as well as see the gender and race divisions of each separate school. In order to implement this two select html bars were created in order to change the x and y axis variables and therefore change the entire scatterplot. Three button filters were made to color code the graph differently, displaying a legend for each of the colors when the filter was used on the graph. On hovering of any point, two pie charts were then made, one displaying the division of gender and the other displaying the division of race. Here we get a better overview of each school and get into more details about the variables of each school and compare them.




The third part of our final design consists of a cluster graph that displays each general field and the different majors within them as well as the median salary and female percentage taking each major. We used the cluster graph from the Ed Discussion visualization link as inspiration and displayed more information about each cluster. We wanted to show some comparison of salary and major together, so the cluster in the visualization was very similar to our final design. When thinking about majors, we think about a cluster of categories that belong within a larger category, so we thought that the cluster graph was very fitting. In order to implement this, we created a bunch of circles and then scaled them to the income of each major. These circles were put close to each other in terms of the larger category they belonged to, with some space given between them. When hovering over each dot, some information will be displayed for each as well.




The final part of the final design was the spider plot. This would be the most in-depth comparison for two separate schools. Instead of a large overview users would be able to narrowly compare and contrast two schools on limited variables. Here, we scaled each scale based on the maximum value of the colleges being compared so that the triangle can fit both colleges within it. We also made some scales based on the ratio of the maximum in order to help fit these colleges. The two selection bars are html select bars which contain the names of all the colleges in our dataset and users can scroll through all of them to pick what specific colleges they're thinking of or interested in.




## Team Contributions:

## Chris:

- Creation of base U.S. map and plotting of points.
- Python Processing of datasets in order to create a singular dataset that had all information needed for colleges, such as combination of location dataset of colleges with demographics dataset of colleges in order to get one common one.
- Creation of entire scatter plot and its interactions as well as pie graphs that go with interactions
- Contributed to writing of text within website
- Created basic scroll
- Styling of website
- Contributed to final report

**JJ:**

- Creation of button filters for U.S. map, filtering both early career and medium career pay.
- Creation of hover over interaction for points in the U.S. map
- Creation of legend that accompanies button filters in U.S. map.
- Creation of the entire spider plot to compare two colleges more closely.
- Creation of cluster chart visual comparing majors and their respective incomes
- Added on to scroll
- Contributed to writing of text within website
- Contributed to the styling of the webpage including introducing colors for data points on scatterplot, formatting for different numeric values, and background page color
- Contributed to final report