Mary Shoop and Charlie Elliot
CSC362 Final Project

Data Write-Up

For our project, we are using data from three sources. First, we have our salary potential dataset (link to data). The items in this dataset are colleges, and the attributes provide information on the average salary for students of these colleges after graduation. The datasets contain data for the 25 colleges with the highest mid-career pay for each state (not all states reach 25—for example, Alaska only has two). For each college, there are values for the item's rank within the state (1-25), the name of the school, the state that the college is located, the average early career pay for graduates, the average mid career pay for graduates, the percent of alumni who think they are making the world a better place, and the percent of student body in STEM.

Secondly, we have the tuition cost dataset (link to data). Again, the items in this dataset are colleges, but there are many more items than the salary potential dataset (936 vs. 2,900). The attributes in this dataset provide information about the cost of tuition at each respective college. The attributes include the college name, the state the college is located in (and its abbreviation), the type of the college (public, private, or for profit), the length of the degree programs (2 or 4 years), the cost of room and board, the cost of in state tuition, out of state tuition, and then a column for the total cost of attendance for in state students and out of state tuitions (calculated by tuition + room/board).

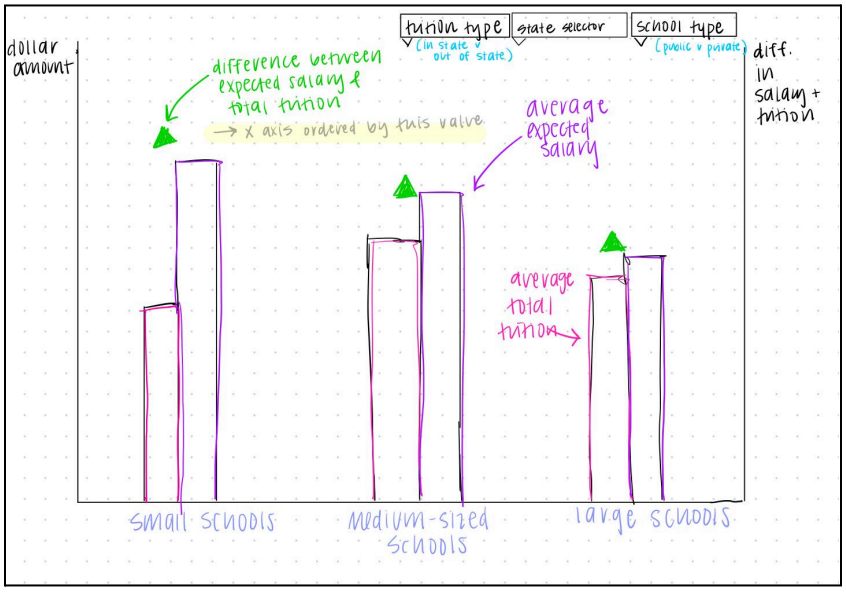
Thirdly, we have data on the undergraduate enrollment levels for colleges in the US. This dataset was downloaded from Integrated Postsecondary Education Data System (IPEDS) as a csv file. In this dataset, each item is a college in the IPEDS. The only attributes are the name of the school and the total student population for the respective college as of 2020 (the year in which the other two data sources are from as well).

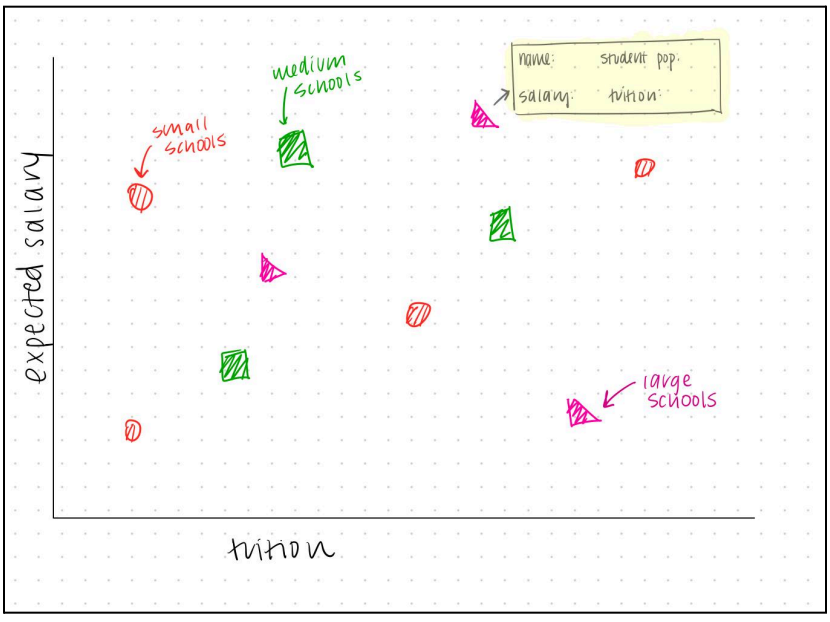
In our final project, we need the following attributes: college name, state the college is located in, in-state tuition, out of state total tuition, start of career expected salary, type of school (public v private), and undergraduate enrollment number. From these attributes, we will derive 3 additional variables. First, we will derive the difference between in-state tuition and expected salary. Second, we will derive the difference between out of state tuition and expected salary. Third, we will derive a 3-level ordinal variable for the size of the school. Using the enrollment variable from the IPEDS data, we assigned schools with less than 5,000 students to small, those with between 5 and 15 thousand students to medium, and those with more than 15,000 students to large.

To get all of these attributes together in one dataset, we used a python script to join the three respective datasets. We performed a left join on the salary potential dataset, because that was the smallest dataset. We ran into issues with the join when the names of state colleges with multiple locations used mismatched icons across the dataset. For example, one dataset would refer to UCLA as University of California–Los Angeles while another would refer to it as

University of California: Los Angeles. This was alleviated by sorting the schools alphabetically and manualing changing the dash to a colon for schools with multiple campuses before joining the data together. In the event that a school did not have data in either salary, tuition, or enrollment, that college was dropped from the dataset. Our cleaning processes yielded one dataset with 796 observations. Each item in the final dataset is a college. Each item has 7 attributes, in addition to the name of the college: state, school type, in-state total cost, out-of-state total cost, early career salary, and total enrollment count.

Project Sketches

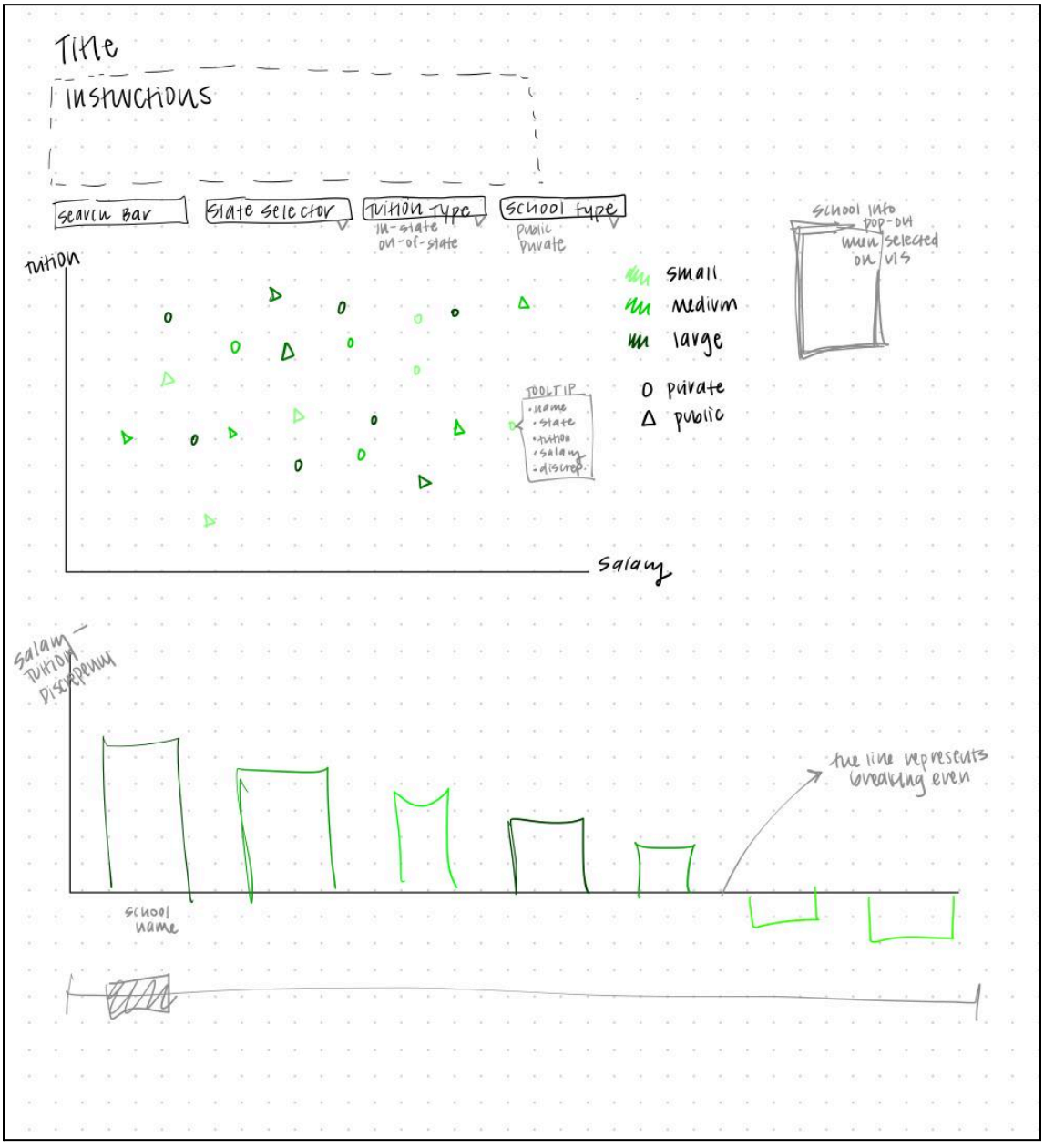




Iterations

After feedback and preliminary task analyses, we realized that the scatter plot was the idiom we were relying on for most of the tasks, not the bar charts. Thus, we decided to make the scatter plot the primary idiom in our vis, and take out the bar charts displaying the count of schools based on their size. Instead, the scatterplot will now be the vis that users can filter and interact with through search bars and drop down selectors.

After thinking about the order task, we decided to create a scrolling bar chart that explicitly ordered the items by their salary-tuition discrepancies. This makes it easier for users to understand the range of this attribute across the dataset. This bar chart will still be interactive, as it will respond to the search bar and drop down selector inputs.



Task & Accessibility Analyses

Comparison Task

The first task users should be able to accomplish with this visualization is comparing the salary-tuition discrepancy between 2 colleges. For instance, a user should be able to easily compare the salary-tuition discrepancy between Davidson College and Duke University.

To do this, a user could search for Davidson College and select the highlighted point on the scatterplot. When the point is selected, the information on Davidson will appear on a side panel. Then, a user can return to the search bar and input Duke University, and select the point for Duke, allowing them to see the information on Duke and Davidson side by side. Further, to better see the two schools compared on the scatterplot, a user can use the state selector and filter the points to only see colleges in North Carolina. Since there are fewer data points after this filter, a user can hover over each point (triggering the tooltip display) to locate the points for Davidson and Duke. When these points are selected, the symbol will be highlighted on the visualization, allowing the user to more easily compare the location of the point on the plot. Comparing the location on the 2 axes allows for a comparison of the salary-tuition discrepancy between 2 colleges—points closer to the top left corner of the graph represent the best salary-tution discrepancy (high salary, low tuition) and points in the bottom right corner represent the inverse of this (high tuition, low salary).

I. Accessibility

To support accessibility, this visualization uses persistent highlights and descriptive tooltips to help users compare colleges without relying on memory or precise cursor control. When a college is selected, its point remains visibly marked on the scatterplot and its details are displayed in a side panel, allowing for clear, side-by-side comparisons. Dropdown filters reduce visual clutter, and all data is accessible via text, supporting screen reader navigation and users with cognitive or visual impairments.

Range Task

The second task supported by this vis is observing the range of early career salaries across small colleges. To accomplish this task, a user can use the ‘School Size’ selector, and select the ‘Small’ category. This will filter the scatter plot to show only small schools (those with less than 5,000 students). From here, a user can observe the location of these points along the X axis, which displays the average early career salary from these colleges.

A user can see that most of the schools are concentrated within 40-60 thousand dollars/year as an early career salary. They can see that there are relatively few small schools that have average salaries below $40k, and very few that have average salaries above $80k.

II. Accessibility

The filtered view enhances accessibility by simplifying the interface and reducing the number of visual elements, which benefits users with cognitive disabilities or visual processing

difficulties. The axis labels are clearly defined, enabling screen reader users to navigate and interpret the data more easily. Additionally, tooltips provide detailed textual information about each college—including name and salary—so users do not need to rely solely on visual positioning. These design choices help ensure that the visualization supports inclusive interaction for a wide range of users.

Order Task

The third task that can be completed with this vis involves ordering a selection of schools by their salary-tuition discrepancy. An example of this would be ranking large schools in North Carolina from most profitable to least profitable. In order to complete this request, users can filter schools with 15k+ undergrads by selecting "Large" from the "School Type" filter. From there, they can use the "State" filter to only show schools in North Carolina. At this point, the bar chart will have filtered and ordered the remaining schools by the discrepancy allowing the user to easily see the rank.

III. Accessibility

In order to use the required filters, users can tab to the correct dropdown and use the arrow keys to select the necessary attributes. Each bar on the bar chart is labelled with the name of the school, giving even those who can't see the vis an understanding of the order. While it is not possible to tab from bar to bar as there can be up to 796 at a time, tabbing through the filters or searches allows users to easily narrow down to schools to view different orders and rankings. The bar chart also displays tooltips when hovering over bars with exact values in the event that two bars are hard to differentiate spatially.

Pilot Testing

To pilot test our visualization, we worked with a student unfamiliar with our project. We asked her to walk us through her completion of the three tasks outlined above so we could observe the functionality of our vis. The script we used with the user as well as a report of their process and feedback is provided below.

Script

Speaker: We are evaluating our visualization and are asking you, the participant, to complete some tasks using the visualization and then provide feedback about the visualization and experience. As a reminder, we are evaluating the visualization, not you as a participant, so you don't need to worry about being "right" as you complete these tasks. There are three tasks, followed by a brief feedback session. The whole pilot session should take under 5 minutes. Do you consent to participate?

Speaker: Thank you for agreeing to participate. We will start with the three tasks. Please 'think aloud' as you complete the task, meaning voice what you are thinking as you work through the task.

Your first task is: compare the salary-tuition discrepancy between Davidson College and Duke University.You may answer clarifying questions, such as "what do you mean by x?" or "are you looking for an exact number or an approximation?"
[Pause for user to complete first task]

Your second task is: observe the range of early career salaries across small colleges.
[Pause for user to complete second task]

Your third task is: ordering and ranking the salary-tuition discrepancy of medium sized universities in your home state.
[Pause for user to complete third task]

Speaker: "That is the end of the third task. For this last bit, we welcome any feedback you may have about the visualization or about your process for completing the tasks."
[Allow participant to speak first, then informal discussion]

Task 1

For the first task, our user was asked to compare the salary-tuition discrepancy between Davidson College and Duke University. While we expected this task to be completed with the search bar functionality, our user first tried to locate the points for each college on the scatterplot by hovering over many different points to see the information on the tooltips. When the user realized that she would not be able to locate Duke and Davidson from the main plot, she moved to the state selector and filtered to only display North Carolina schools and then to small schools to find Davidson. After this filtering process, the user realized she used the search bar to locate the schools. She searched for Duke, clicked on the highlighted point in the vis, and then had the two school's information displayed side-by-side.

She was able to get information on each school's tuition and salary easily from this process, but the task was asking for the user to compare the *discrepancy* between these values, which was not listed directly as a value in the tooltip. We expected that the comparison of the two points on the scatterplot would allow our user to compare the discrepancies, but because of how our user used the filtering based on school size, Davidson and Duke were selected from different views of the vis, so the user wasn't comparing the two on the same plot at the same time.

Task 2

The next task was for our user to observe the range of early career salaries across small colleges. The user went through the appropriate filters and selected the small schools filter. She then had all of the small schools displayed in the scatterplot. She was able to talk about the distribution of early career salaries rather easily after this process. She explained that most small schools were concentrated between 40-60k/year, and identified outliers as well. She also used the line to describe that for most of the small schools, students would be profiting (salary > tuition).

Task 3

For the final task, the user was asked to order and rank the salary-tuition discrepancy of medium sized universities in your home state. She easily used the selectors to filter the data by state and size. Then, she attempted to complete the task using the scatterplot. She was trying to figure out the appropriate way to complete the task with the scatterplot, but was expressing difficulty. Then, she scrolled down to the bar chart. As soon as she saw this vis, she was easily able to explain that this idiom was ordered by the discrepancies of the filtered data. She said that she had not even thought to scroll to the lower idiom at first, as there were no indicators in the interface that drew attention to the bar chart that was below the scatterplot.

Feedback from the Participant

Overall, the user expressed that she wished there were more instructions on the interface that directed her around each visualization. She felt like key elements of the tool—the search bar and the bar chart—were not catching her attention at first, and she had trouble completing the tasks before she located each of them. Additionally, she suggested that the annotations on the scatterplot, as well as the title for the bar chart should be bigger in size, and that the scrollability of the bar chart, as well as the presence of the bar chart at all, should be highlighted more.

List of Changes

1. List the value of the salary-tution discrepancy in the tooltip
2. Add commas in large numerical values to improve readability of monetary values.
3. Increase size of the bar chart title, and provide reference to it higher on the interface page.
4. Increase the size of the scatterplot annotations.
5. Increase the size of the search bar.

Results From Final Vis User Tests

User Test #1

Task 1

We ran our first user test with an Animal Behavior major. Upon starting the visualization, she read the description at the top of the page and went from there to the search bar. It was a straightforward search of both Davidson and Duke, with the highlighting of the symbols making the process easier. However, she did not click on the symbols to compare, but rather hovered over one tooltip before searching for the next school. This suggests that it isn't as intuitive as we

thought to click on the symbols, and we should clarify that in the directions. Another issue that arose was the bar chart not updating with the search bar. While it didn't inhibit her ability to locate the two schools on the scatter plot, she mentioned that it may be easier to compare the two if there are multiple ways of visualizing their differences.

Task 2

For the second task, the user clearly understood where to locate specific filters, what they did, and how they worked. She first clicked on the state filter and returned to "All" before jumping to the school type filter and selecting "Small" When shown with the resulting scatter plot, she hovered her mouse over the symbol on the far right and far left, correctly identifying Harvey Mudd as the small school with the highest average salary and Mississippi Valley as the small school with the lowest average salary. As with the first task, she did not click on the symbols to get a better observation of the range between the small schools, only using hover tooltips and remembering the data that popped up.

Task 3

The third task began as the second task did. The user easily located the state filter and selected Massachusetts. Once filtered, she scrolled down to observe the bar chart and was able to correctly list schools in order of salary-tuition discrepancy. Initially she forgot to switch the school type filter from small to medium, but stated later that she did not hear that part of the task. This misunderstanding brought about an interesting development, as there are enough small schools in Massachusetts to prompt scrolling. When met with this opportunity, the user attempted to drag two fingers across the trackpad, click the chart and drag two fingers across the trackpad, click and use the arrow keys, and click and drag on the chart itself. The grey scroll bar on the overall view was very wide in the filtered scenario and did not appear to be a scroll bar, nor was it labelled as one. This made it difficult to figure out how to navigate the bar chart and complete an altered version of the task, and the user recommended more instructions surrounding the use of the bar chart.

Summary

Overall, the user found navigation about the site to be straightforward after reading the instructions upon opening the site. Despite not clicking on the symbols for comparison, there were no complaints about not being able to view both schools' information at the same time. The user thought the upper half of the visualization was well labelled, but that the bar chart instruction could be improved.

User Test #2

Task 1

Our second user test was with a political science major with a minor in Hispanic studies. Her first task was to compare the salary-tuition discrepancy between Davidson College and Duke

University. She began by filtering the data for North Carolina schools using the state selector. Then, she scrolled down to the bar chart to look for the names of the school, however she did not realize that she could scroll to find them. Instead, she returned to the top of the vis and began using the search bar. She then was able to locate both schools on the scatterplot. However, she did not immediately realize that she could click on the point to have the information displayed on the side, so she tried relying on the tooltips to compare each school's salary discrepancy.

Task 2

For the second task, our user was asked to observe the range of early career salaries across small schools. The user easily located and used the school size filter so that only small schools were displayed on the scatterplot. She then identified which axis was displaying the item's salary prediction, as opposed to their tuition. She was easily able to identify outliers in the distribution as well as the concentrations in the data. She discussed the distribution by saying things like "Harvey Mudd is an outlier here" and "it looks like the average salary is around 48k for these schools".

Task 3

For the third task, the user was asked to order and rank the salary-tuition discrepancy of medium sized universities in her home state. She easily located and used the size filter, selecting 'Medium'. Then, she also easily found and used the state selector to filter for Rhode Island schools. Only three schools were displayed for Rhode Island, so the scatterplot visualization was rather sparse. Because of this, the user scrolled to the bar chart vis. Once there, the user was easily able to see the ranking of salary tuition discrepancies for the 3 medium schools in Rhode Island.

Summary

Overall, this user was able to complete the tasks efficiently. She highlighted two things in her feedback: first, that she wasn't aware that clicking on the scatterplots would display the data, and second, she was not aware that she could scroll over on the bar chart. Even though this information is displayed in the instructions sections at the top of the page, this user test showed that these features should be highlighted more explicitly for users.

User Test #3

Task 1

Our third user was a political science major with a minor in data science. His first task was to compare the salary-tuition discrepancy between Duke and Davidson. Upon hearing the wording of this question, he saw the title of the bar chart was similar and focused on that vis. He was able to locate the search bar and seemed to expect a dropdown as he typed the names in, however had no problems understanding to press return to search. After searching, he scrolled back down to the bar chart expecting to see Davidson, however the bar chart didn't update with

the search. After clicking around the bar chart, he returned to the scatter plot to complete the task, and was able to visualize the difference on the graph and with the tooltips. He did not know to click the symbols for more in depth comparison

Task 2

On the second task, the user had to observe the range of early career salaries at small colleges. While he was easily able to filter to small schools on the school type filter, he was not aware that he had to press 'esc' to undo his previous searches indicating that he did not read the blurb at the top of the site. The faded light green symbols were hard to see and he was not able to see the trends as well as possible. We eventually told him to read the blurb at which point he pressed esc and was easily able to identify Harvey Mudd and Mississippi Valley as the highest and lowest, respectively.

Task 3

The third task required ranking the medium schools from a specific state in order by salary-tuition discrepancy. The user easily navigated to New York and medium on the respective filters. He initially was looking at the scatterplot saying the spread looks good, however upon clarifying that we were asking him to rank the schools he knew to access the bar chart. There were only 5 schools so scrolling wasn't necessary but scrolling the entire page was needed to see schools on the far right.

Summary

Overall there were a couple important issues emphasized by this run. In the first task, the user requested that the bar chart respond to the search bar for alternate methods of comparison. He wanted to be able to select and look at the bars for Davidson and Duke next to each other. Another issue arose when the user did not read the blurb. He said that it looked too long and lost his interest. Given the important information within the blurb, we have to either emphasize its importance, make it more concise, or clarify navigation instructions elsewhere in case the blurb is ignored. Other issues include mentioning that scatter points can be clicked and fitting the barchart within the page boundaries.

Link to Visualization:
https://maryelizabethshoop.github.io/DataVisFinal/index.html