A Holistic Look At The College Lifecycle

Joao Pedro Castro* Margaret Clark† Pratheek Mandalapu‡

ABSTRACT

The purpose of this project is to support high school students in their decision-making process for choosing a college. Select domain tasks include:

- "I want to identify which state has the cheapest out-of-state tuition."
- "I want to explore the colleges to see which ones have the largest difference between in-state tuition and salary potential."
- "I want to see which colleges have the highest diversity rates."

Although these domain tasks do not encompass the entirety of the tasks our visualizations will support, they represent a broad mix of the most relevant tasks. We want students to be able to identify the right college for themselves based on tuition costs, diversity rates, and salary potential so that they can maximize their experience across all stages of their college career.

1 INTRODUCTION

This project addresses domain tasks such as comparing two colleges for the cheapest out-of-state tuition, the largest difference between tuition and salary potential, and diversity rates across the student body.

1.0.1 Cheapest Out-Of-State Tuition

This is an important domain task as the cost of tuition is a central factor in the university decision-making process. High school students typically create a list of their preferred institutions and this visualization tool will allow them to compare their top choices based on cost.

1.0.2 Largest Difference Between Tuition and Salary Potential

Given that students come from a diverse array of socioeconomic backgrounds, different students will prioritize different factors. For example, a student from a wealthy family might not be as concerned about tuition cost as they are about salary potential. For students who are equally concerned about the cost of tuition and salary potential, this visualization tool will allow them to identify colleges that address both criteria.

1.0.3 Colleges With Highest Diversity Rates

Although diversity may not be prioritized over tuition cost or salary potential, research from College Matchpoint shows that diversity is an important factor for high school students when deciding on their university of choice [2]. Once a student has selected a set of colleges that match their tuition and salary potential criteria, this visualization tool will allow them to explore colleges' diversity rate and select the one that meets their preference. For example, some students may be more comfortable in historically black colleges while others may prefer a more racially diverse institution.

*e-mail: castro.joa@northeastern.edu
†e-mail: clark.marg@northeastern.edu
‡e-mail: mandalapu.p@northeastern.edu

2 RELATED WORK

Similar studies have been conducted to represent the cost of college attendance and the various factors that contribute to a student's decision of where they want to attend. As tuition costs increase with time and college education portfolios shift and change, so do the values of the students who seek to attend. Sobel et al. assesses the rapid growth of tuition costs for post-secondary education in the United States [5]. The author represents the given information in a line graph visualization. The y-axis provides the average cost of annual college tuition in the United States, with line points showing growth over time in the last half of the century via the x-axis. Little color is used to add further channel engagement, and the chart design is simplistic (making it easy to interpret). Data junk has been removed to focus the viewer on the pared down data. We see the value in this kind of visualization as it quickly communicates its ideas to viewers, and accurately shows the data at hand. However, this graph utilizes many numerical representations in order to prove a simple argument, and we feel this causes the lack of engagement with the viewer that we seek to obtain through our visualization tool. There could be further visual encodings which represent the data we want to show as well as allowing for interaction with the user.

In contrast, Condon et al. prefers visualizations with fewer numeric values, and represents colleges using symbols related to their attributes, displayed in organization of their geographic location [3]. Color is not an important factor, position provides the information of their general distance from each other but with no specific location information besides the state. Alternate view options show colleges by cost, by ranking, or whether they are public or private. We would like to reflect the geographic spatial aspect of this visualization, as well as its simplicity and lack of heavy numeric characters. However, it provides static views, and does not include all the factors we have determined are most important to student college decision-making. Furthermore, we feel as though the information provided by the position of the symbols is, in some cases, unnecessary to the graphs being produced, which may confuse viewers rater than inform them.

3 USE CASE

Our final use case differs slightly from our initial use case. While the scenario and target user remains the same, the data we will focus on has been modified. Rather than focusing on granular costs of a city such as food, housing, and transportation, we will instead focus on several high-level elements of a college itself. For example, when exploring the University of Arizona we want to observe the aggregate cost of tuition (in-state and out-of-state), diversity, and salary potential. Rather than compare colleges solely by cost of attendance, we want to compare colleges by other factors such as potential salary and diversity rates to give high school students a holistic understanding of the various colleges they are researching.

3.0.1 Finalized Use Case

According to College Matchpoint, when students in the United States are deciding on which college to attend, there are three major factors, among others, to consider. These are cost of attendance,

diversity of student body, and salary potential post-graduation [2]. These factors can vary widely among schools, and can offer insight into the college decision making process, the student's experience during their attendance, and post-graduation opportunities. Our visualizations would aim to highlight the full experience of attendance and allow students to weigh their decision against the quality of life they could achieve after college. This use case holds relevance right now given the uncertainty of macroeconomic conditions. Based on data from MSCI, we are likely approaching a recession, which would leads to tighter consumer discretionary spending [1]. Therefore, it is important to be aware of expected costs before committing to a college. The costs, salary potential, and diversity varies drastically between colleges. Therefore, being aware of the full picture is important when deciding which college to attend.

Our target users will be high school students and their families who are in the process of looking at potential colleges to attend. We will generate a visualization that aggregate the three main factors we identified by state. We will also provide the option to explore the map in more detail and compare two colleges on a more granular level, comparing in-state and out-state tuition costs, salary potential, and diversity metrics.

4 Data

It was very important that the data set we chose covered several attributes of college attendance. We realized through our research that finding data sets for granular costs of attendance would not be possible. Therefore, we pivoted to looking for data sets that not only aggregated costs across several colleges, but also provided information about other factors that may influence a student's decision to attend a specific college.

4.0.1 Data Source

The raw data gathered for this project comes primarily from the US Department of Education but from other sources as well, such as Priceonomics.com, TuitionTracker.org, payscale.com, and the Chronicle of Higher Education. This raw data was initially collected and condensed by Jesse Mostipak, a Developer Advocate at Baseten, who then posted 5 different Excel files to Kaggle here [4]. These files were the starting point of the final dataset used. The data is quantitative and records discrete values because values inbetween data points are not considered.

The 5 files included historical tuition costs, tuition and fees from 2019, diversity breakdowns for each college, post-college salary potential, and average net cost by income bracket. To create a single dataset that contained all the information we needed, we decided to combine data from the tuition and fees file with the diversity breakdown and post-college salary potential files, keeping only the colleges that were listed on all three files. Our final dataset covers school name, state, institution type (private vs public), degree length, room and board fees, in-state tuition, in-state total, out-of-state tuition, out-of-state total, total enrollment, early-career pay, mid-career pay, each minority group listed, and the percentage of each minority group compared to the total enrollment.

4.0.2 Biases, Ethical Considerations, and Shortcomings

Some biases and shortcomings that we found within the data include the fact that only colleges with 2-year and 4-year degree lengths were included, the diversity data is from 2014 (while tuition and fees data is from 2019) and does not specify what is meant by "unknown" race nor breaks down the "two or more races" category to capture specific biracial groups, and many colleges had to be removed since they were not listed on all three files that were used.

4.0.3 Cleaning

The three data files we ended up using came very messy so we had to do some cleaning and consolidating to achieve our desired, final dataset.

- Filtered out any colleges with N/A values across all three files
- Using the tuition and fees file as a base, we used Excel's VLOOKUP function to figure out which colleges in that file also appeared in the other two files. Any college that was not found in both of the other files was removed.
- Filtered out 2-year degree colleges to only have 4-year degree programs
- filtered out for-profit schools to just have public and not-for-profit private schools
- Filtered out America Somoa, Puerto Rico, DC, and Virgin Islands to just have schools within the 50 states.
- Used Excel's VLOOKUP, INDEX, and MATCH functions to combine all three files under one single .csv file. This allows us to have the final 612 colleges in one table with all of their attributes side-by-side.

5 Design Process

The design process of this data visualization tool involved three rough partial sketches of the final product. During the first sketch, the product features a map of the United States of America with divisions by state shown. The abbreviation of the state may be given, but these were left blank to maintain simplicity of design. Our early considerations for user interaction involved the user being able to click on a state, and view a drop down menu with all the colleges in that state that have available data. Hovering over a state also shows the average cost of tuition and average percentage of minority individuals in the given schools for both public and private, allowing users to rapidly scan the map for areas more likely to hold the attributes which are important to them. Colors will be minimal to avoid confusion, and the website will be available in English as this is the shared language of ourselves, the paper authors.

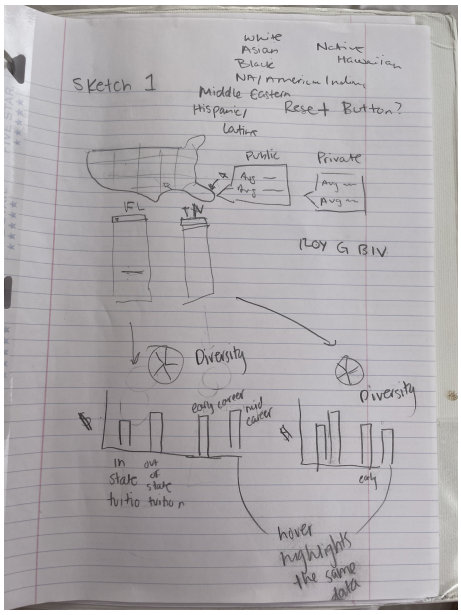

Figure 1: Rough sketch of full visualization

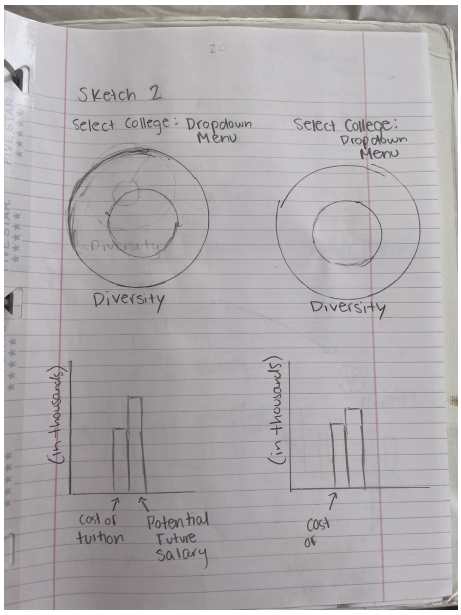

Figure 2: Rough sketch with linking and donut charts

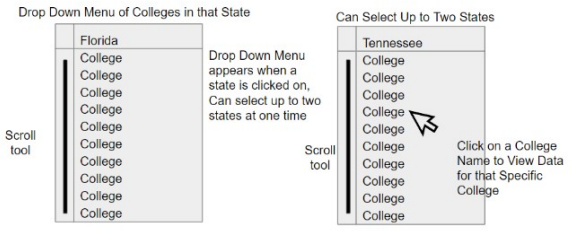

Figure 3: Initial sketch with map and interactive drop down

Upon designing these first two sketches, we assessed that not enough data was available to the user. We also wanted users to be able to compare two colleges against one another. We added a data visualization for the average diversity with divisions of racial groups in a pie chart, if students are seeking a specific student body composition. They can also view a bar chart comparing mid and early career salary prospects against the cost of attendance per year for in-state and out-of-state tuition. This allows students to assess specific colleges rather than the aggregated information of a state. It also provides an accessible visual of the data. The final subject will be minimally colored with color representations provided in an immediately available legend. Users can interact with the visualizations by hovering over the bar chart rectangles, sparking a highlight on the same measurement of the other school. For example, highlighting the out-of-state tuition bar for College A will highlight the out-of-state tuition bar for College B.

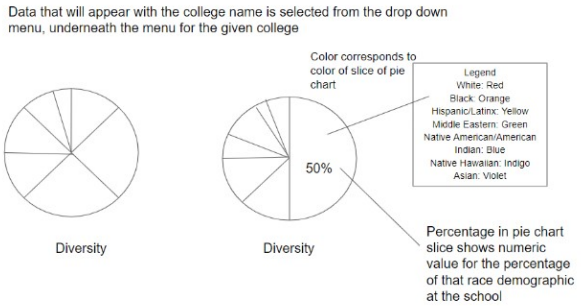

Figure 4: Initial sketch of diversity pie chart

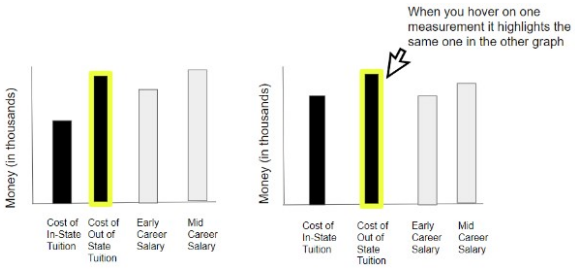

Figure 5: Initial sketch of interactive bar chart

6 Final Design

Our final design involves marks and channels to communicate the data provided through visual encoding techniques. First, the tool provides an interactive map with state borders and different colleges. Ultimately, we chose a square map instead of a realistic map because we felt the shape of states was unnecessary information that could confuse users. Position as a channel, both vertical and horizontal, is important to showing where colleges appear on the map in terms of the state they reside in. It does not communicate their specific position, but the general state only. The mouse hover function can be used to display averages of the statistics of each state in the map. This allows users to easily complete the domain task of being able to find states with low averages for tuition costs. The drop down menu will allow users to select a state to show further data about its specific diversity rates, tuition costs, and salary prospects. This fulfills the domain task of comparing colleges with the largest difference between tuition costs and salary prospects. It also allows them to see which colleges have the highest diversity rates. They can select up to two colleges to show further details, and compare

the same statistics on either side. Ideally, this would provide an easy tool for students to weigh different factors in their college decision.

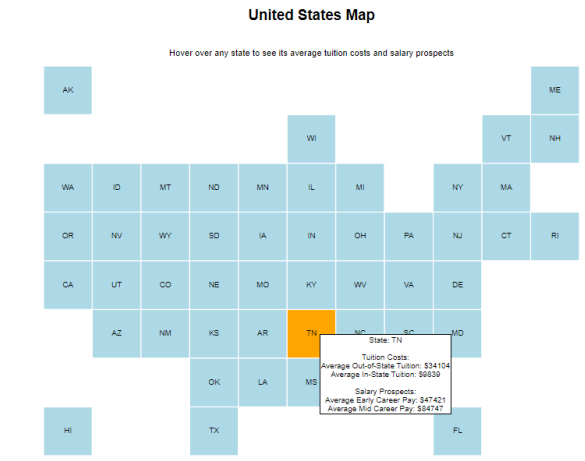

Figure 6: Final design of interactive U.S. map

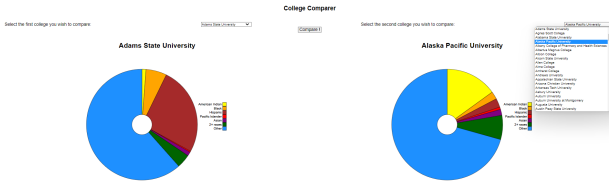

Figure 7: Final design of donut chart, dropdown

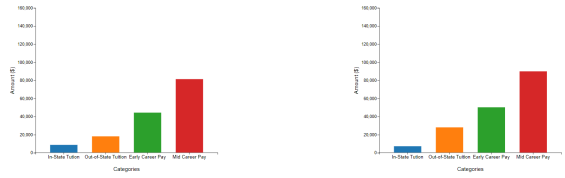

Figure 8: Final design of interactive bar chart

The second portion with additional data includes a pie chart and a bar graph. The bar graph uses the mark lines, where their length corresponds to specific data. Their position, in terms of vertical, communicates the size of the wealth for that category. The horizontal data does not display a range of data, but informs the type of measurement being shown. The visualization has an interactive functionality that allows for highlighting the same measurement in both graphs when one is hovered over and the display of the exact value of the bar, to aid in ease of comparison. Difference in color may also denote whether the data concerns tuition costs or future salary prospects. However, color does not communicate numerical data. Color is also an important channel for the pie chart, showing the different types of racial groups displayed. Area as a mark is important to the pie chart slices as well, as it denotes the percentage of that race at the given college. Tilt is also a channel involved in the pie chart. These various graphical elements allow us to communicate various aspects of the given data in a clear way to the viewer. We specifically attempted to minimize overusing marks and channels that were unnecessary to the data we hoped to represent, and focused on what was essential. This allows the viewer to have clarity on the information and speeds up their comprehension upon their initial viewing. From the rough sketches to the final sketch, we gradually added more functionality and eliminated unnecessary features to improve user experience and accessibility. The larger swath of data that is able to be visualized provides the user with more information to influence their decision about specific colleges, rather than general information for different states. Our final design thus combines all of these goals.

7 Discussion

Overall, we believe our endeavor to answer the provided domain problems was successful. Our tool turned out operational in all aspects we were seeking. It provides accessible, simple displays of key information for the college decision process. We also used a variety of strategies for visualizing information so the final result remains engaging in its legibility. In the future, however, several factors could be added to make this tool even more helpful for users. We could see the benefits of adding further information related to the cost of colleges, such as expected living costs like food, housing, and entertainment around different colleges. These can also play a significant role in the expected costs of a student's college experience and it is something we would liked to have included, but it did not fit in with the data we collected, which better represented the factors we settled on (diversity, tuition costs, and salary prospects).

8 Conclusion

To summarize this project, our team aimed to capture a holistic view of the college life cycle by examining tuition costs, diversity rates, and salary prospects. These parameters represent the college experience before, during, and after attendance, respectively. Tuition cost is an important factor to consider before committing to a college as it is a large financial commitment. Diversity relates to 'during' the college experience as it represents student life. Finally, salary prospects relate to the post-college experience and the quality of life attainable upon graduation.

The goal of our visualization tool was to support high school students and their families in the college decision-making process. Having completed this project, we believe we have created an adequate tool for our target users to examine a broad range of college statistics and use them to compare prospects.

9 Usability Test

9.1 Overview

We chose to run a controlled experiment for our usability testing. We paired with another group and asked them to perform a series of tasks. The purpose of this experiment was to measure the time and accuracy of our participants in order to evaluate how simple and accessible our visualization tool was.

9.2 Experiment

Question 1: Does Northeastern or Stanford have higher salary prospects?

- Time Taken: 21 seconds
- Accuracy: Stanford University (correct)

Question 2: Which state has the lowest out-of-state tuition cost?

- Time Taken: 39 seconds
- Accuracy: Wyoming (correct)

Question 3: Does Northeastern or New York University have higher diversity?

- Time Taken: 17 seconds
- Accuracy: New York University (correct)

9.3 Analysis

The results of our experiment reveal that the map is the most difficult element of our visualization to use. Consider questions 1 and 3, which ask the user to compare statistics between colleges. The time taken to answer is significantly lower than for question 2. Moreover, our participant was not confident in their answer to question 2. This indicates that the map is not intuitive nor able to condense information effectively for specific queries. The goal of the map was to serve task abstraction levels of discover, explore, and summarize. Upon reflection of the results, it is clear that we need additional tooltips to condense the information on the map for our users. We might see benefit from a 5-point summary of all states to the side of the map (max, min, Q1, Q2, and Q3 values) as well as utilizing a specified range of color to visually display the extreme states according to a specific factor.

We also learned that our bar charts and pie charts are the most well-designed elements of our visualization because of how well they condense college statistics. Our participants were able to quickly and confidently answer questions 1 and 3. The linking between charts is also helpful to compare statistics simultaneously between colleges and the highlighting of slices on the pie chart distinguishes diversity metrics.

ACKNOWLEDGMENTS

- Basics of D3
- Data Manipulation
- Filtering By Column
- Making a Map in D3
- Common Data Manipulation Tasks
- Bar Charts in D3
- Tooltips in D3
- Pie Charts in D3
- Kaggle Raw Dataset

10 APPENDIX

10.1 Data Abstractions

10.1.1 Rows

- Each item in the dataset represents a U.S college

10.1.2 Columns

- Column 1: University Name
  - Attribute Type: Categorical
- Column 2: State Name
  - Attribute Type: Categorical
- Column 3: State Abbreviation
  - Attribute Type: Categorical
- Column 4: Institution Type
  - Attribute Type: Categorical
- Column 5: Degree Length
  - Attribute Type: Categorical
- Columns 6-10: Room and Board Fees, In-State/Out-of-State Tuition, Total In-State/Out-of-State Cost
  - Attribute Type: Quantitative
  - Ordering Type: Sequential
- Columns 11-12: Early and Mid-Career Pay
  - Attribute Type: Quantitative
  - Ordering Type: Sequential
- Columns 13-19: Count of each Race Group's Enrollment
  - Attribute Type: Quantitative
  - Ordering Type: Sequential
- Column 20: Total Enrollment
  - Attribute Type: Quantitative
  - Ordering Type: Sequential

10.2 Task Abstractions

10.2.1 Cheapest Out-Of-State Tuition

- **High-Level: Discover**
  - The user is learning new information about out-of-state tuition costs.
- **Mid-Level: Lookup**
  - The user will be able to identify the spatial location of states but will need to lookup further information to understand tuition costs by clicking into the state to see the list of colleges in that state.
- **Low-Level: Compare**
  - The user will compare two colleges to see which one has the cheaper tuition costs.

10.2.2 Largest Difference Between Tuition and Salary Potential

- **High-Level: Discover**
  - The user is learning new information about tuition costs relative to salary potential.
- **Mid-Level: Locate**
  - The user doesn't know which colleges within which states have the largest differential between tuition and potential salary and must locate those colleges themselves.
- **Low-Level: Compare**
  - The user will compare two colleges to see which one has the largest difference between tuition cost and salary potential.

10.2.3 States With Highest Diversity Rates

- **High-Level: Discover**
  - The user is learning new information about diversity rates.
- **Mid-Level: Locate**
  - The user knows they want to find states with high diversity rates, but they don't know where they are located without exploring the map.
- **Low-Level: Compare**
  - The user is looking to identify a single target (state with high diversity).

REFERENCES

[1] Four scenarios for 2023: Navigating uncertainty.

[2] How students decide which college to attend.

[3] C. et al. A visualization model based on adjacency data. *Decision Support Systems*, 33(4):349–362.

[4] J. Mostipak. College tuition, diversity, and pay, Mar 2020.

[5] A. E. Sobel. The escalating cost of college, 2013.