WHERE DOES IT PAY TO GO TO COLLEGE?

Oishani Ganguly, Lauren Simons, Gauri Pidatala

OVERVIEW

In this project, our group did an exploratory data analysis on the salaries of college graduates based on the type of schools they attended, the fields they majored in, and the location of their schools. We also wanted to see if there is any correlation between average SAT scores for schools and salaries earned by graduates of those schools. In particular, we wanted to see whether those high SAT cutoffs at "elite" institutions are warranted based on the kind of salaries those students end up earning.
The members of our group were thus motivated by the question, "Where does it pay to go to college?"

THE STORY AND FINDINGS

Our visualizations and interactions aim to equip the user with the necessary information to determine whether getting that Ivy League degree is worth it or whether it is just the same as attending that small state college and saving some money. Likewise, the user is able to determine whether what one majors in really makes a difference in terms of salary. And lastly, we also let the user look at whether an institute that demands a high SAT score does so because it translates into a high-paying job.

We start off by exploring starting and mid-career median salaries of 100 schools/colleges/universities and 50 majors. We limit ourselves to the undergraduate level only.

Users can choose to explore salaries by schools and school type or salaries by majors and percentage increase in salaries by majors. In both the graphs, we let users look at starting median salaries and compare them to mid-career median salaries with the option of sorting both salaries from highest to lowest and vice versa. Users have the additional option to filter by type of school and also filter by the percentage increase in salaries by major.

Findings: Ten years out, graduates of Ivy League schools earned 99% more than they did at graduation. Party school graduates saw an 85% increase in salary by the middle of their careers. Engineering school graduates did not see such high increases in pay, earning only 76% more 10 years out of school than when they joined the workforce.

Additionally, a year-long survey by PayScale, Inc. of 1.2 million people with only a bachelor's degree shows that graduates in philosophy and international relations earned 103.5% and 97.8% more respectively than what they started with. This increase was seen about 10 years post-commencement. Majors that didn't show as much salary growth include Nursing and Information Technology.

The majors with the lowest pay increases by mid-career (<50%) were Education, Interior Design, Nursing, Nutrition, and Physician Assistant. This was no surprise for the Nursing major because starting salaries for these majors were higher than most other majors and it makes sense that it would not increase dramatically over one's career. More than half the majors had a salary increase in the 50% to 70% bracket, followed by the 70% to 90% bracket. Very few majors saw more than 90% increase in their salaries by mid-career. Some of these were Economics, Marketing, and Mathematics which was expected. What was unexpected was that Philosophy fell in this bracket wherein Philosophy majors saw a 103.5% jump in their salaries by mid-career!

We then map all 100 institutes onto a map of the US to give our users a sense of where those high-cost high-return schools are and which parts of the US probably don't have the best schools if you're looking to land yourself a high-paying job. Users can filter schools by region and hover over each school to learn more statistics about them.

Findings: According to the PayScale Inc. survey, attending college in the Midwest seems to lead to the lowest salaries, both at graduation and at mid-career. Graduates of schools in the Northeast and California fared the best.

Overall, it seems that the Ivy League Schools and other schools in the Northeastern region have the highest SAT scores and earn the most money in terms of starting median salaries. Attending some of the California schools (such as CalTech and UC Berkeley) and Southern schools (such as Georgia Institute of Technology) also lead to high starting salaries. In terms of college majors, engineering seems to be the most profitable while starting out but the pay stagnates early on. So, if a student wants to make the most money right out of college, statistically speaking, it might be a good idea to study hard and ace the SAT's, major in engineering, and attend a school in the Northeast. **With that being said it's important to note that these conclusions are only based on school averages. Each person has their own future to decide and can accomplish anything they put their mind to regardless of what school you go to or what you major in!**

THE DATA

While looking for datasets to visualize, we prioritized the fact that the main entity being visualized must have multiple dimensions/attributes. Since college data and salaries after graduation are directly relevant to the members of our group, our peers, and students of the class, we decided to proceed with this idea. College data is easily available and has several dimensions that can be visualized.

For our project, we use the following 4 datasets:

1. Where It Pays to Attend College: This is a Kaggle dataset supplied by The Wall Street Journal. It contains three CSV files:
    a. salaries-by-college-type.csv: This contains salary information for 249 schools. We use the 100 schools that overlap with our SAT score information dataset below. We only use the columns '**School Name**', '**School Type**', '**Starting Median Salary**', and '**Mid-Career Median Salary**' out of 8 columns.
    b. degrees-that-pay-back.csv: This contains salary information for 50 majors. We only use the columns '**Undergraduate Major**', '**Starting Median Salary**', '**Mid-Career Median Salary**', and '**Percent change from Start to Mid-Career Salary**' out of 8 columns.
    c. salaries-by-region.csv: This contains salary information for 320 schools. We use the 100 schools that overlap with our SAT score information dataset below. We only use the columns '**School Name**', **'Region**', '**Starting Median Salary**', and '**Mid-Career Median Salary'** out of 8 columns.
2. University Statistics: This is a Kaggle JSON dataset that contains basic statistics about 311 schools in the US. We only use the **average SAT score**, **city**, and **state** columns from this dataset for the 100 schools that overlap with the above dataset. 5 SAT scores are missing.
3. US Cities: This is a dataset that contains important location information about US cities and towns. We only use the **latitude** and **longitude** columns from this dataset for the cities and towns the above 100 schools are in.

4. A topojson shape file of the map of the United States (taken from lecture notes).

Our data preprocessing involved merging datasets 1.a, 1.c, 2, and 3 to create a single dataset with 100 rows corresponding to each of the 100 overlapping schools and 10 columns for **'School Name'**, **'Region'**, '**Starting Median Salary**', '**Mid-Career Median Salary**', '**School Type**', '**Average SAT Score**', **'City'**, **'State'**, **'Latitude'**, **'Longitude'**. We rejected any school that had missing information for any of the above columns except Average SAT Score. This was one of three data files. The second data file was created simply by filtering dataset 1.b by retaining the specified columns only for all 50 original majors. The third dataset was the topojson shape file.

To keep our data specific and our goal focused, we rejected all percentiles except 50th (median) for starting and mid-career salaries.

DESIGN RATIONALE

Look and Feel

We decided to go with the Google Font "Graduate" for our headings and buttons because we are working with college data. The rest of the text is in the font 'Roboto' to make it readable. We use the colors blue and orange since they are one of the most common college colors and are color-blind-friendly. The tone of our writing and the emojis are such that they capture the attention of college-age students. Fade-in transitions are added for a pleasing, non-startling effect on content change.

Visualization 1: Interactive Dumbbell Plot

Visual Design Rationale

We decided on a dumbbell plot for our first visualization to effectively display the change in starting and mid-career median salaries using horizontal length. The marks are the dumbbells (consisting of circles and lines joining them). The channels are aligned horizontal position for the salaries, aligned vertical position for the schools or majors, aligned length of the line for the change in salary, and color of the circles for indicating starting salary or mid-career salary. Mixing the colors in the lines was a design choice to add aesthetic value and to be able to guide the user's eyes between the two types of salaries easily. The salary figures are formatted as "$x.xk" for human readability. Additionally, the salaries are included next to each circle so users can easily compare the figures and also know the exact numbers. Similarly, the percentage increases are included for each major in a light grey color so users can compare the exact figures. The color difference doesn't interfere with the other text.

A tradeoff we made while displaying the initial data is that we decided to plot all 100 schools and all 50 majors on the x-axis. While we recognize that this might be too many elements to plot and the user will need to scroll to look at all of them, it is the best default visualization that was possible. More importantly, the user is able to look at data for a single school/major and compare it to all the schools/majors if they wish to do so. We make it easy for the user to find a school/major by arranging the schools and majors in alphabetical order. The ability to filter by type of schools and by the percentage increase in salary by major is there for the user to look at specific subsets of schools/majors.

Interactive Elements Design Rationale

We experienced some minor performance issues on Chrome with the interactions being slow there. However, our visualization performs just fine on Firefox and Safari.

Interaction 1: We give users the ability to choose between viewing salaries by school type and salaries by major. By default, the visualization for salaries by school type is displayed on loading the page. The user can toggle between the two visualizations by clicking on the buttons at the top. The buttons respond to mouseover effects by changing color. Depending on which button is clicked, the visualization and related information change dynamically. The buttons and their titles clearly signal to the user to click on them to interact with the visualizations.

Interaction 2: For "Salaries by School type", the user can filter by type of schools to see how each group of schools compares to one another in terms of median salaries and their change. This filter also avoids having to view 100 schools all at once. The buttons respond similarly to the buttons in Interaction 1. The graph area changes dynamically with each applied filter. The title of the filter applied appears above the graph. Most importantly, the height of the graph (and the school names) change to be consistent with the original visualization and fit the screen better.

For "Salaries by Major", the user can filter by the percentage increase in salary to see how different majors lead to different amounts of salary increases by mid-career. This filter also avoids having to view 50 majors all at once. The buttons respond similarly to the buttons in Interaction 1. The graph area changes dynamically with each applied filter. The title of the filter applied appears above the graph. Most importantly, the height of the graph (and the majors) change to be consistent with the original visualization and fit the screen better. We selected the lower and upper limit of the percentage filters so that it roughly follows a normal distribution model where just a few majors fall in the “<50%↑” and “>90%↑” brackets and most of the majors fall under the other two brackets.

Clear instructions above the filters signal to the user to use the interactive feature.

Interaction 3: Lastly, users can apply 4 different kinds of sorting filters for the School Type graph and 2 additional sorting filters for the Majors graph from the dropdown menus. These are lowest to highest and highest to lowest for both starting and mid-career median salaries for both schools and majors. The 2 additional sorting filters for the Majors graph are lowest to highest and vice versa for percentage increase in salaries. These filters are applicable even when school-type filters or percentage increase filters are applied. They are incorporated because different users might be interested in different types of ranking of the schools/majors and their salaries and we wanted to cater to all users to the best of our abilities. We also reset the sorting filter to no filter each time a school-type/percentage increase filter is applied. The schools/majors on the y-axis and their associated dumbbells animate to swap positions so users can track the changes. The dropdown menu with “--Select sort--” clearly signals to the user to use the interactive feature.

In order to convey our findings simply to the user, we included a short summary above the graphs that change dynamically based on whether the user is looking at salaries by school type or by major. Our visualization is inspired by the fourth graph of this tableau visualization.

We decided against allowing a mouseover interaction for the dumbbells because having a tooltip appear on the graph would make the canvas cluttered. Plus, any additional information we would have provided on hover is already provided down in the map on hover. There was no new information that could have provided additional insight to the users. Hover interaction here would've been redundant.

Visualization 2: Interactive Map of the United States

Visual Design Rationale

We do an Albers projection of the map of the United States that shows state borders. The map is colored light grey so as to draw attention to the school location pins against a light background. Since locations are very specific, we use “pin” to denote locations rather than just circles where it is hard to determine the center. The stick of the pin is rooted in a specific part on the map and the circle on the top is for visibility. The marks are the specific locations (latitude and longitude) of the school cities/towns denoted by the pins (stick/line and blue circle). The channels are aligned horizontal and aligned vertical positions. The circles have a thin stroke the color in the color of the map to differentiate closely located pins. For those schools that visually overlap on the map, we added a 0.5 jitter to their latitude values.

Interactive Elements Design Rationale

Interaction 1: Users can hover their mouse over each of the circles of the pins to reveal statistics about each school. When hovered over, the circle turns from blue to orange and increases in size to indicate that that pin is currently being looked at. This change is animated so the transition is smooth. When a pin is hovered over and active, information about that school is revealed on the right of the map. The information revealed is sufficient for the user to determine the correlation between the type of school, SAT scores, and salaries with help from the two other salary graphs above it. Clear instructions above the map signal to the user to use the interactive feature.

Interaction 2: The user can also filter by region of schools to see how each group of schools compares to one another in terms of median salaries, salary increases, school type concentration in the region, and average SAT scores. This filter also avoids having to view 100 schools all at once. The buttons respond similarly to the buttons in Interaction 1 of the salary graphs. The number of pins on the map changes dynamically based on which region filter is applied. Interaction 1 (hover) still works on filtering by region. The title of the filter applied appears above the graph. Clear instructions above the filters signal to the user to use the interactive feature.

TEAM CONTRIBUTIONS AND TIME SPENT

Time Spent

- Total time spent on project: ~35 hours coding + ~1 hour doing the write-up
- Tasks that took the most time:
    - Sorting and animation functionality for visualization 1 (~3 hours)
    - Data preprocessing (~3 hours)

Team Member Contributions

1. Oishani
    a. Data preprocessing: Salaries by school type.
    b. Created Visualization 1, part 1 (salaries by school type).
    c. Wrote the rationale for Visualization 1, part 1.

2. Gauri
    a. Data preprocessing: Salaries by major.
    b. Created Visualization 1, part 2 (salaries by major).
    c. Wrote the rationale for Visualization 1, part 2.

3. Lauren
    a. Data preprocessing: Merging datasets and cleaning the resulting dataset.
    b. Created Visualisation 2 (map).
    c. Wrote the rationale for Visualization 2.