Cody Gunter
Sabrina Alves
Nicklaus Austin

Project 3 Report

The Data:

Our group used four csv files from two different data sources. Our first was, "Where it pays to attend college" by the Wall Street Journal, found on Kaggle. This source contained three separate CSV files: 1. A list of 50 majors with the career median salary of each. 2. 250 schools were listed with the distinction of school type given (Liberal Arts, Engineering, State, and Ivy League), and also the career median salary. 3. Our third had over 300 schools with the located region in the US and the same salary information. The second source was from IPEDS (Integrated Postsecondary Education Data System) 2020 "Institutional Characteristics". This data contains about 6,000 schools and is provided by the National Center for Education Statistics (NCES). NCES is a government agency that makes basic data on US colleges publicly available in a CSV format.

The IPEDS data points were originally foreign keys, with the primary key stored in an excel workbook on their website. We converted the columns we had an interest in using Jupyter Notebook and Pandas. The following are the converted columns (with their original columns names): OPEFLAG, SECTOR, HLOFFER, DEGGRANT, HBCU, HOSPITAL, MEDICAL, TRIBAL, LOCALE, OPENPUBL, INSTACAT, C18IPUG, C18UGPRF, C18SZSET, C18ENPRF, CCBASIC. To convert the values, we converted the worksheet with the primary keys into a csv (this had three columns of relevance: the column that the values are used for, the number used, and the string value paired to that integer) and used .replace to place the string values where the integer values were. Below, the code for one of the replacements is displayed:

```
#INSTCAT
mask = values.varname == "INSTCAT"
instacat_val = values.loc[mask]

code_val_rep = [int(x) for x in instacat_val.codevalue]
label_val = list(instacat_val.valuelabel)

df15 = df14.copy()

df15['INSTCAT'] = df15['INSTCAT'].replace(code_val_rep,label_val)
df15.head()
```

Thankfully, the dictionary workbook provided allowed us to .loc the column name (which is a universal “ID” in all of their reports), which returned all of the strings needed for replacement. The lists were made– the number used needed to be converted into integers to match with the foreign keys– and the values were replaced.

The columns were chosen for their relevance to the college search and independence from other variables. A number of the columns reiterated the same information a few different times, so we tried to grab a sampling of information to give the best overview possible to viewers.

With the WSJ data we merged the region and college type CSV files, and then merged the new dataframe with the IPEDs CSV. This new CSV from the merge contained the school type, region, latitude and longitude, as well as the salary breakdown. The problem with the new CSV file was the fact only 92 schools were included in all 3 of the previous files when joined on the name of the institution.

Our Inspiration and Research Questions:

Our inspiration was to give college-ready users a quick introduction to colleges that fit their ideal mold.

Questions we wanted to answer were: What majors have the greatest starting and career salaries? Which regions receive the greatest average salaries?

What colleges are available in these regions? At various locales? Offering medical degrees? With certain student populations?

Our Design Decisions:

We recognized that our audience is primarily made up of teenagers who likely spend the majority of their time in front of screens. For this reason, we wanted to have our website in “dark mode” as much as we can accomplish it with the time constraints we were under, as dark mode is generally better for longer periods of engagement, according to WPCommerz.

Additionally, we wanted easy to read fonts for our users. With these two requirements in mind, we decided to use the LUX theme from Bootswatch. It came ready with a dark primary color, which was used throughout our site and mimicked dark mode when there was text present.

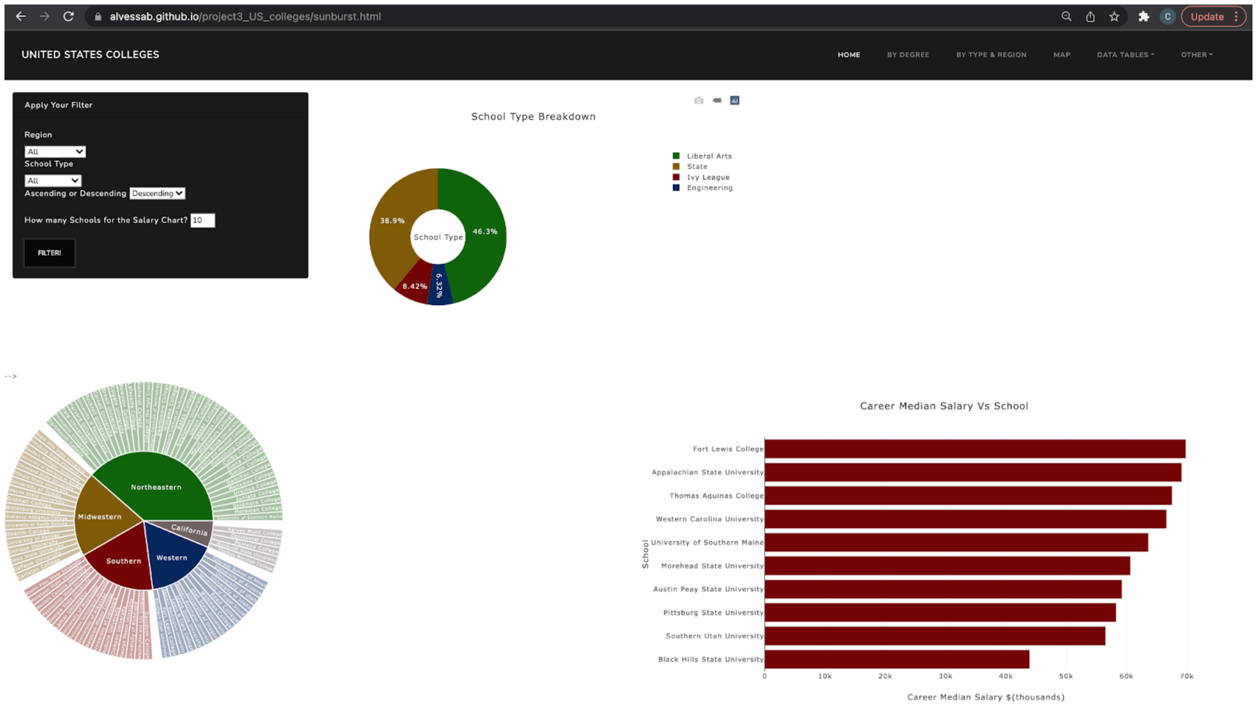
Since we did not have the need to have any recognizable branding, we used the following color palette for our visualizations and as accents. These colors are among the most recognized “academia” or university colors, from the various sources we consulted.

Some pages will have small graduation caps as a tie in with the color palette and for the customization it best offers. These icons come from Font Awesome, a sister company to Bootswatch and Bootstrap.

Finally, our index contains all of the visualization and filter pages embedded on the screen using the <iframe> html tag. We wanted the student to be able to consult all of the visualizations simultaneously. For example, if the student wanted to reference degrees they were most interested in due to their salary (see Salary vs Degrees) while they were clicking on links provided by the map, then they would be able to do so from the same window in their browser. Additionally, if the student wanted to consult the median salary of the schools in a certain region as they consulted the map, they would simply have to scroll up the page to view Salary vs Type and Region. They can travel to each of the dashboards with the link in the card or with the nav bar if they would like to have the full page.

Our Visualizations:

Our website had 2 interactive dashboards: One page with three plot.ly visualizations including a bar chart, sunburst chart and donut chart.

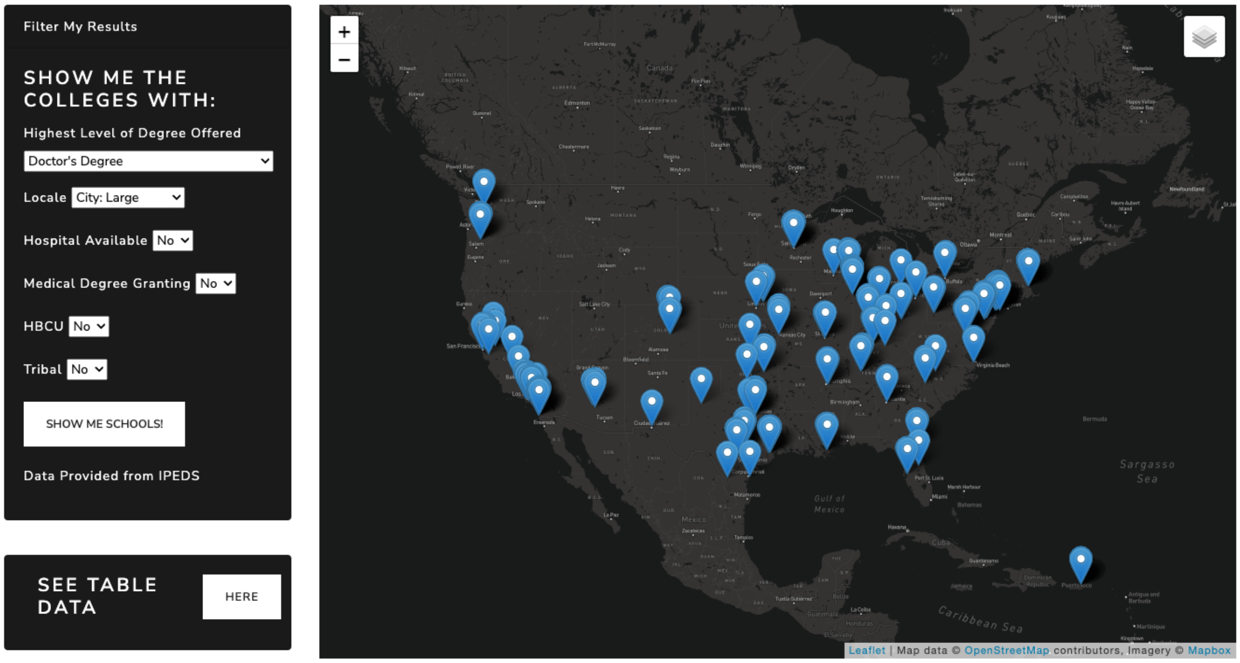


All three visualizations responded to a card with four filters listed. Filters filtered the region, school type, order of the bar chart, and a choice of the number of schools represented in the bar chart with a max of 20 schools for purposes of keeping the graph readable.

The horizontal bar chart displays the career median salary of the schools selected within the filtered set. All four filters directly affect the bar chart display. The sunburst chart displayed the five regions (Northeast, Midwest, Southern, Western, California) with every school included in each region as the outside branch. Only the region and school type filter affect the display of the sunburst chart. To show the breakdown of school type within a particular region, the donut chart responds to the region filter showing the percentage and count of each school type in the region.

Map:

Because our data is location based, we plotted the latitude and longitude coordinates provided by IPEDS on a leaflet map with map box base layers. The markers for the colleges have the following information: name of the school, link to the main page of their website, the address of the school, the degree level offered (to make sure that users are locating the correct options for themselves), and the Carnegie Size and Setting Classification. This last data point gives some information on the size of the school and the makeup of the student body– whether it's mostly undergraduate, graduate/professional, etc.



There are 6 filters that simultaneously filter the data on every load of the screen. The number of filters was necessary because the roughly 6,000 data points overloaded the map and caused a multi-second delay in load time. The presets give one of the highest counts of the markers on the map, so the user doesn't feel as though the map lacks information when they first visit it.

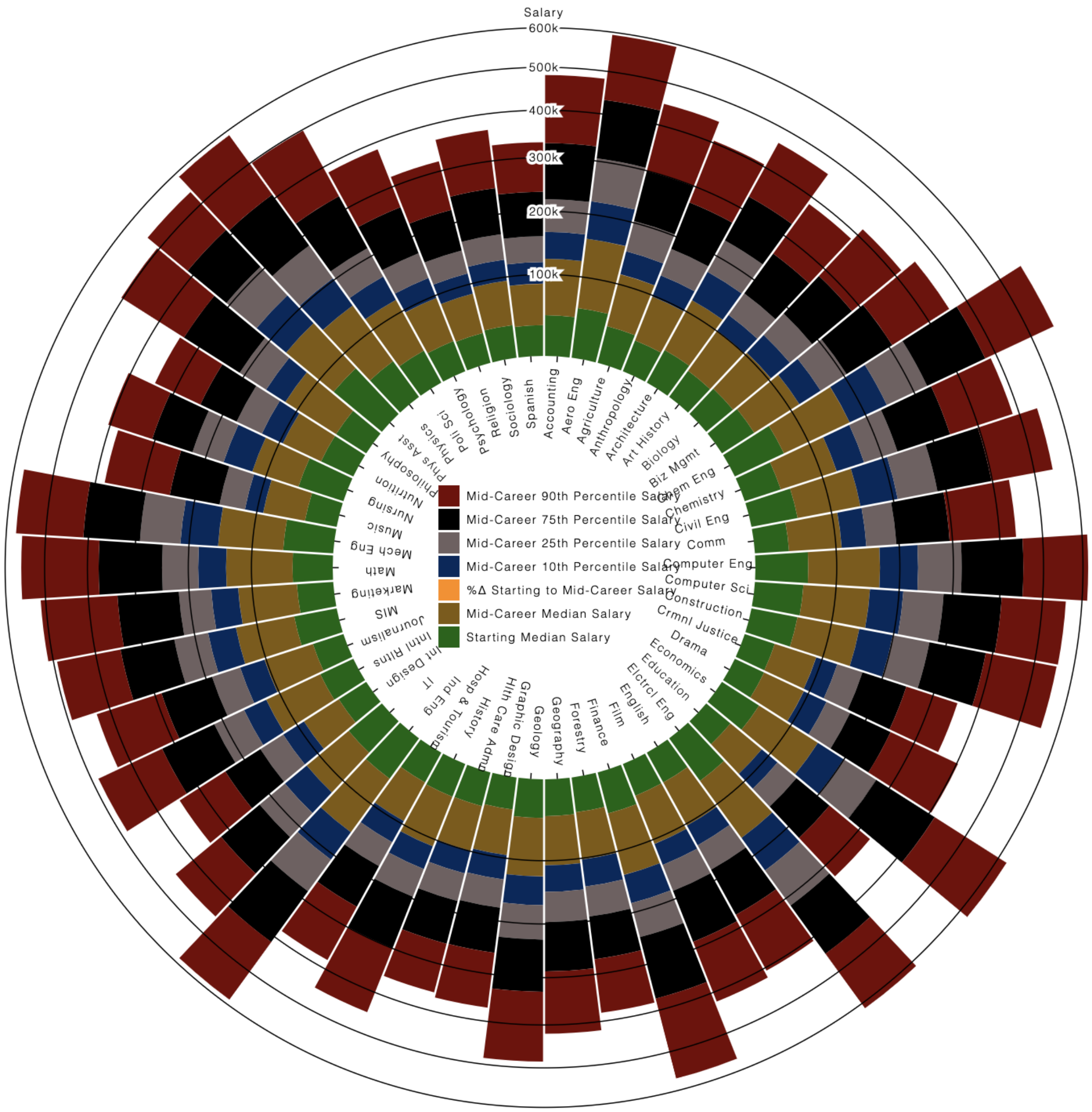
The last 2 filters are important classifications as these classifications can often be "all or nothing"; their classification as a tribal or HBCU school can determine whether or not the student would be willing/able to attend, due to the exclusivity of some of the programs. These filters are yes/no.

Given the specific research done in the search for a medical program, the "Medical Degree Granting" filter was added to distinguish these schools from the rest. This filter is a yes/no option.

Hospital availability is the one nod to students who may need to be close to medical care during the college years, since the IPEDS data does not include any other information regarding this topic. Medical care is often a hard-to-reach resource for college students, and that is often due to availability. IPEDS records when there is a university owned hospital on "campus". This is a yes/no filter.

Locale and Highest Level of Degree Offered are the "personal preference" filters. Locale contains various categories of the local population, such as city: small-large. Additionally, the highest degree offered contains a variety of filters from "Award of less than one academic year" to "doctor's degree". This filter was chosen for some of its overlap of other categories, such as Size and Setting (which is less categorical and more of a description).

Radial Stacked Bar Chart:



This radial stacked bar chart displays the median salaries at each step given. These steps include starting salary and mid-career at the 10th percentile, 25th percentile, 50th percentile , 75th percentile, and 90th percentile. Each section of the strip represents the salary for the classification given in the legend. The 90th percentile salary for a chemical engineer is not north of $600,000, but rather the cumulative total of the six classifications is north of $600,000. In reality the 90th percentile salary of chemical engineers is closer to $200,000.

Conclusions from our Data:

Looking at the regional and type analysis, 85% of our schools were either state or liberal arts colleges. Not surprisingly, Ivy league universities on average earned the highest career median salary of all the school types. For what it is worth, State college graduates earned the lowest career median salary of all the school types. State colleges did dominate the regions with the exception of the Northeastern region where state colleges comprise just 10 percent, as Liberal Arts and Ivy League types are more frequent.

When it comes to the majors that pay off the most, it's the engineering field, math, and the financial sector. This is in line with the school type analysis where engineering school graduates made just less than those from Ivy League schools, where those that graduate as an Ivy Leaguer are proportionally more likely to major in STEM and find great success after graduation. The majors that see the least amount of success when it comes to making money are those in the language/communication realm, health and wellness, and the arts.

Future Work and Our Limitations:

Future work could include incorporating census data looking at the median household income and comparing that to the salary of each college to get a more comprehensive representation. With additional resources, an analysis of the most frequent majors at individual schools or school types could help us determine why State and Liberal Arts college graduates do earn less than Ivy League and engineering schools.

Limitations included not having a larger list of schools in the region and type CSV, just having those 92 didn't allow us to have much to compare and contrast. This was evident when you compound the fact that the vast majority of the dataset was either State or Liberal Arts colleges.