INFO-H 517 Visual, Design and Analysis

Analysis of Unemployment, Cost of Living and Number of Jobs in Data Science across US

Group 6: Screaming Falcons
Haseeb Khan, Nathan Niese and Wilson Rodden

December 7th, 2020

Introduction

Since many people in our class are preparing to graduate, we wondered what state would be best suited for our careers. There are many things that go into choosing the best state for employment including job availability, cost of living, and stability of jobs. Our goal in this project is to create visualizations that give insight into taking the first steps into our professional careers.

Data Sources

To create a visualisation that would give insight for stability of jobs, unemployment and the cost of living for each state of the United States we gathered data from 3 sources. Our job opening, salary range, and employer rating data was scraped from Glassdoor. This was done through a proprietary scraping algorithm that parsed the data from currently open positions for the first 30 pages of a job search url, allowing us a sample of the first 900 jobs from each state. The total job counts for each state was provided by a separate piece of data on the job search url since we could only get the first 900 jobs (Glassdoor only shows the first 30 pages with 30 jobs per page).

For the cost of living data we used data published by the World Population Review. The data broke down cost of living by grocery cost, utility cost, transportation cost, and miscellaneous costs all aggregated in the total cost index which was used to shade our choropleth under the "Overall Cost" filter.

Unemployment data of course was pulled from the Bureau of Labor Statistics. At first we pulled a report for September 2019-September 2020 that had a large variety of data including labor force, raw unemployment population, and employment rates by percent. Unfortunately the lockdown made the data collection for the last 12 months incomplete. After some feedback we adjusted it to an annual employment rate from 2008-2020 that provided more 1-dimensional data restricted to simply the unemployment rates.

The data for our state recommendation algorithm was generated by aggregating the percentiles of this data, calculated through their z-scores across the sampled population, and run through a cosine similarity calculation that compares vectors that live on the same plane in multidimensional space. For anyone who remembers Soh Cah Toa from geometry, it's a similar calculation in more than 2 dimensions.

Methodology

As we mentioned earlier, the goal of this data visualization is to help our colleagues put their best foot forward in beginning their careers in data science. The visualization should be simple enough to be understood by someone unfamiliar with the data but complex enough to provide actionable insights.

In order to ensure this we ran A/B testing with 12+ users, 6 of whom are familiar with data science as a practice and 6 users who aren't. This resulted in several adjustments throughout the process to make the data easily digestible such as simplifying the line graph in the bottom right of our design from what was originally a dual graph then progressing to the faded line backdrop for easy comparison to the total state population among other adjustments. The results of the A/B testing helped us improve

readability and interpretability of the data so our users can answer specific questions about an incredibly broad dataset. We will discuss this further in the next section.

To ensure clean version control and collaboration we leveraged GitHub to store our data and code. The design heavily leveraged D3, JavaScript, and HTML. The data scraping was done through Python using Beautiful Soup and common Python packages (Pandas, Numpy, etc.).

Design Process

To begin designing the visualisation, some initial ideas were penned which answered basic questions. How do we illustrate our data across all states for easy comparison, with raw data or percentiles? We have a lot of data here and could make dozens of views for our choropleth. What will be our main focal point for the visualization? Should we allow the user to change the main portion or make the surrounding data customizable? What data needs to be fit into the primary choropleth and what can be represented by more traditional graphs?

Our first design focused on the choropleth as the center of our design to represent percentiles of our data points across the country. We then brainstormed numerous ways to display the data and make it interactive using techniques like hover tooltips, dual graphs, a percentile slider to bring certain states into focus depending on the view, among others. The initial design focused on whatever state was selected. After some critiquing it seemed like the visualization needed to give the user easier access to comparing the data at scale.

We then came up with the idea to use a card based design to give structure and granularity to the state level data above the choropleth. To display information crucial to the user we used the top outer cards to display important data that did not need large-scale comparison (most of which are already represented in the choropleth design) such as total open jobs, expected salary ranges in the selected state, the cost index, and unemployment rate. We then use the right graphs to compare related data in more detailed and precise ways from the data shown in the choropleth. The next step was to introduce a reverse interaction in case the user wanted to see the cost index graphs represented state-wise as a choropleth, allowing the user to compare insights they care about across all states.

Given our graphs we found plenty of interesting information including some of the ideal areas to work for data scientists, including Indiana which compares favorably to other states. One of our first questions is which state is the cheapest to live while still providing data scientists strong salaries? Our choropleth shows us that there are several states hiring data scientists with high salaries where cost of living is extremely low such as Kansas, Missouri, Arkansas, and Oklahoma.

While these are good leads, it looks like the total number of opportunities in these states are low with only 40 to 45 open positions in Kansas, Arkansas, and Oklahoma as well as Missouri with 179 jobs however Missouri only lands in the 39th percentile for total open jobs. From here we ask what states have a higher number of opportunities, even if it means a slightly higher cost of living? We found that states like Georgia and Washington have higher numbers of open jobs with higher salary ranges compared to the cost of living in the area.

Now that we have found 2 options that we like, what are some similar states? If we stick with the initial design and we want to find states that are similar to Georgia and Washington we would have to click through 144 different combinations of views, which seems excessive. Instead we ran a machine learning technique called cosine similarity to compare the percentile data we calculated for each state. The result was a series of recommendations for the top 3 most similar states whenever the user selects a state. Now we know that if we like Georgia and Washington we should look at Indiana, Ohio, Texas, Minnesota, Colorado, and Iowa as well. It looks like one of the places we should be looking at is in our own backyard! With this version we go from having 2 to having 8 strong options as to where to start our career. Now how do we drill down further into what it will be like to live in these places?

We needed a way to see what it's like to live in these spaces. To do this, we take apart the cost index information provided by World Population Review. Since, the cost of living index values were fixed and were not temporal, so designing a bar graph to display that information seemed plausible. Do you like to cook? Click the grocery cost bar to see Texas has low grocery costs and everyone knows they're well known for quality produce and barbecue. Do you not have a car? Click the transportation costs bar to see Washington has high transportation costs that you would be able to ignore since you won't be driving, pick an apartment close to your office and pocket the money you just saved on driving. Do you plan on setting up a gaming rig in your apartment with dual monitors or will you have an awesome programming setup whose GPU will sap electricity? Click on the utility cost bar to see those costs are low in Washington, Georgia, Ohio, and Colorado so you can run your side hustle cheaply streaming on Twitch nights and weekends.

Last, we need to understand how stable employment is in the area. If unemployment is high that means we have more competition for these open jobs. How are we going to compare the states we're interested in against the rest of the country? By setting up a multi line graph we can compare every other state to our selected state. Washington, Georgia, Ohio, and Colorado all have relatively low unemployment rates; it looks like Colorado was the fastest to recover from the 2009 financial collapse. If we're concerned with the recovery rate, Colorado may be the fastest to get back up to speed once the economy begins to return to normal.

The chart interactions between the choropleth and the bar graph provide powerful insight to imagine what it would be like to actually live in these states broken down by cost. Before we leveraged this tool we had to go off of word of mouth. Where do we go to start our career as data scientists? You only hear about California and New York to work in tech or financial services. While those areas and industries are great, they can be exhausting. Now we know that California and New York aren't even on the list of states for us to look further into. Amazon is hiring hundreds of data scientists in their Seattle office to expand their AI team. We could work for Apple in Texas where there won't be as much competition as California to land a job at a big tech company. We can move to Colorado if we want to be in an area that will bounce back from financial issues faster. Georgia has a research institute with a great salary range hiring data scientists in an area where the cost of living is low and there's still plenty to do in the area. We went from having no frame of reference as to where to start our careers to having 4 extremely strong candidates.

Questions posed from the dataset and explanation of the insights gathered

We formulated many questions from the datasets we used to create the visualisation and gathered facts from visually inspecting and interacting with the graphs that were created. Some of the questions and insights gathered related to those questions are mentioned as follows:

1. Which states have been hit hardest by unemployment during COVID-19?

It is evident from the choropleth that Hawaii, California and Nevada are the states where the unemployment rate has skyrocketed this year. Following these three states are Illinois, New York and Massachusetts which also exhibit an alarming spike in unemployment. This tells us that these states are still deep in the recovery phase and have quite some time before returning to the pre-COVID economy. Planning to move to these previously affluent areas might involve some risk.

2. What are the states where the unemployment rate has been the most resilient?

Although the pandemic is still not finished and the number of COVID-19 cases increase everyday there are a few states whose unemployment rates have weathered the storm. Nebraska, Vermont and Utah have the lowest unemployment rates as of now which is a positive sign for anyone planning to make a move to these more rural states. However it appears Utah is the best option of these states given the higher salaries and reasonable salaries. This insight ultimately reduces the risk factor involved in planning the first career move after graduation.

3. What are the states that sound appealing at first but should be further investigated?

States on the west coast, especially California have a reputation for well paying jobs as well as very high overall cost of living. States on the east coast such as New York, Massachusetts, Connecticut, New Jersey and Maryland aren't slouches either when it comes to cost of living. As new graduates we will have to evaluate the salary ranges for these states to see if living in these states is feasible. Where would some cheaper, more thrifty options be?

Mississippi, Missouri, Arkansas, New Mexico, and Oklohoma seem to have very low overall cost but have low total job opportunities. If we were to move to one of these areas we would only have a handful of options for places to work, likely in a remote position. It could be a great way to pay off student loans but if we want to make a career move, we will likely move states as well.

4. Which states offer the most data science jobs and who is the top employer of each state?

It is clearly evident from the visualisation that the highest number of jobs are in the state of New York with Facebook as the top employer of the state. We can also see that states on the east coast in

general have more data science job openings than any other states. Although California has a lot of job openings in data science but cost of living in this state is very high as compared to other states and should be kept in mind while starting a career move to this state. If we were to consider the west coast it may make more sense to look towards Washington state where Amazon is hiring hundreds of data scientists in 2021.

5. What states are most similar to New York?

We've all heard stories of New York and how exciting but expensive the big city life can be. Now that we know that New York has the highest number of data science job openings we can also see from the visualisation that California, New Jersey and Massachusetts are most similar to New York across all of our data points. If we wanted a slightly more budget friendly version of New York we may consider New Jersey!

6. Where can we find a high number of data science positions matched with a high salary range and low cost index?

This is the most important question for which we have designed this visualization and its answer would totally depend on the user's preferences. Factors such as cost of living and macroeconomic variables involved and the most important thing to keep in mind when making a move to this state, Would I be able to save money in this state if I move?

Further Improvements

If we were to take this concept to the next level we would want to accomplish three goals: simplified broad cost comparisons, the ability to select multiple states at once for comparison, and a filter allowing our new grads to filter by jobs other than just data science.

The simplified broad cost comparisons would be an adjusted chart to what we have now in the bar chart made to display each state as a medium sized circle with a low opacity, similar to the line graph, with the selected state as a solid color (opacity of 1). This would be a more simple interaction than the reverse indexing we have now and would match the display of the lower graph simultaneously.

Being able to select multiple states at once for comparison would give more context to the data. Rather than all other data being present with low opacity and only one state remaining solid, we could implement an outlining system to outline selected states with different colors. There would have to be a maximum number of possible selections or else the graph would become too chaotic, but this would let users compare several states. 4 could be a reasonable number given the fact that we have 3 recommended "similar states" for each selection.

The ability to filter by jobs other than data science was mostly an issue of resources in scraping jobs. To scrape just data science from Glassdoor for all 50 states takes roughly 3-4 hours. To scrape for all

majors would increase this time exponentially, however it could be done. This feature would broaden the useability and impact of this tool beyond the narrow scope of what it can do now.

These would be the 3 major design adjustments that would be next on the proverbial to-do list with the highest impact. There are a series of other smaller changes that we would look into but these larger adjustments would drive value for the entirety of our school rather than just the data science program.

Source Code and Datasets

The source code for this visualisation is readable, minimal and reproducible. The source code is hosted on Github and it can be found on the link mentioned as follows:

Link to the source code

References:

[1] Mike Bostock, https://d3js.org/

[2] https://worldpopulationreview.com/state-rankings/cost-of-living-index-by-state

[3]https://www.bls.gov/bls/news-release/laus.htm#2019

[4] https://www.glassdoor.com/

[5] https://gist.github.com/mbostock/4090848#file-index-html

[6] https://bl.ocks.org/duspviz-mit/9b6dce37101c30ab80d0bf378fe5e583