Milestone 1: Proposal – Computer Science Job Map

Team: *The Work in Progress Team*

Members: Jeff Chow (jchow49@gatech.edu), Jaden Lim (jlim390@gatech.edu), Yu-An Chen (euanchen@gatech.edu), Alex Chen

Idea Summary:

Simply put, for our data science project, we intend to collect data surrounding information about Computer Science or Computer Science related jobs and compile them into a visual map, highlighting the hotspots for those jobs. Our goal in the end is to hopefully be able to pinpoint where the best jobs are located based mainly on factors such as the salary, working hours, type of job and potentially working conditions or reputation as well.

Motivation and Why It Matters:

The Computer Science job market is seemingly always growing. With such a model, we can generate a rough expectation of what the future may hold for us. The central idea is that we're all going to enter the job market soon with some of us already beginning. In addition to this, many of us are specifically Computer Science oriented as well. As a result, it would be helpful to get a rough idea of what and where the Computer Science job market is like, both physically and more metaphorically to better understand the outlook for our future careers. Besides the application of creating expectation for the future, forming the job model could also enable people to pinpoint their ideal type of job and where to find it. For example, if someone wanted a cybersecurity based job, with the map, they could focus on finding where most cybersecurity jobs are located or even if it's all over the place. If they wanted to do a data science job, the same thing applies and so on. In summary, this project matters because it will allow for students like us to get a picture of where, what and how much their careers could potentially be.

Plan of Action and Data Collection:

Although we haven't officially formulated an absolute plan yet, we do have a rough idea of what we intend to do. The most important step is to figure what information to collect and how to collect that information. Currently, the data we will collect involves typical more surface level job position information. This includes things like salaries, working hours, location, type of job or position, company and more. Once we've collected enough data, we plan to compile all of them and generate a kind visual map. Ideally this will be an interactive proportional symbol map that presents which portions of the country are the "best" for Computer Science jobs with an emphasis on availability (number of positions), salary (with a possible adjustment for standard of living there), type of CS job, and of course, the location. As for how to actually execute

on this rough idea, that is currently still being discussed. Our team is fairly new to data science, so the individual steps will have to be worked out later on. Job information will likely be easy to obtain simply by scanning for information from job and career sites, applying datasets, or doing both. However, compiling and actually pulling that data will prove more challenging if done automatically, especially for job sites since each job site has varying layouts and may include or may not include information about salary and such. As for the map, such a model would likely best be done on a website, as that would make it easily accessible and interactive.

Success Measurement:

For our project, success can actually be measured relatively easily. The success is completely tied to whether the models can be created in first place and if they are accurate or a good estimate. In other words, if we can develop a proper model of the United States, label where the most jobs are and accurately to an extent (accurate meaning in comparison to the data itself and other people), then we have succeeded in our initial goal.

Datasets:

As for datasets, we are still looking into which dataset to use. We are considering using this dataset linked below:

https://www.kaggle.com/datasets/lukebarousse/data-analyst-job-postings-google-search

Since the data collected is geared toward careers in data analytics, something to consider is that using this dataset would end up narrowing the scope of our project. If we move forward with using this dataset or any similar dataset, we plan to aggregate salary data, number of job openings, etc. by state/city. In addition to this more specific data-career dataset, we have also found some others that go into a variety fields within the Computer Science sphere. But, once again, out core source of data has not been finalized. We may even rely more on raw job sites depending on the situation.

https://www.kaggle.com/search?q=programming+jobs+in%3Adatasets