Vincent Jiang
Bahaa Korb
Akaash Mahinth

INFO 3300 P1 — Final Report

Charts and Visualization

Chart 1

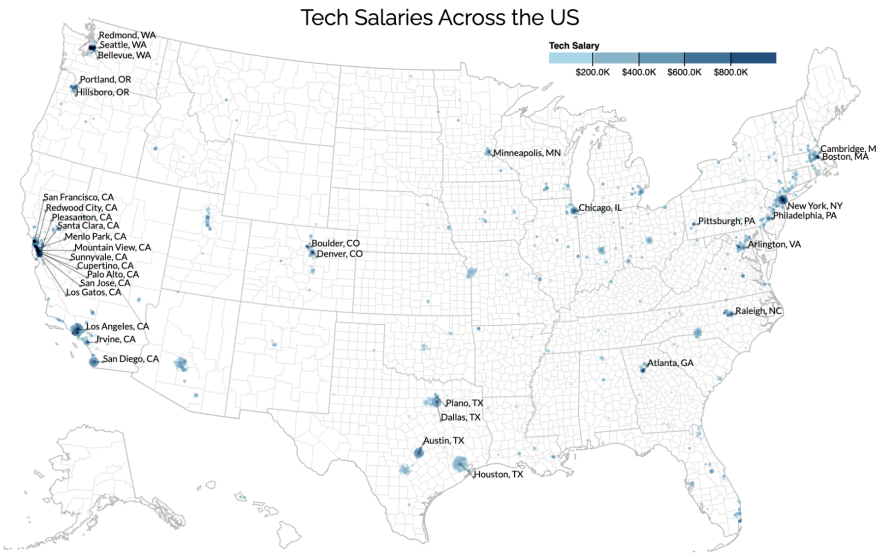
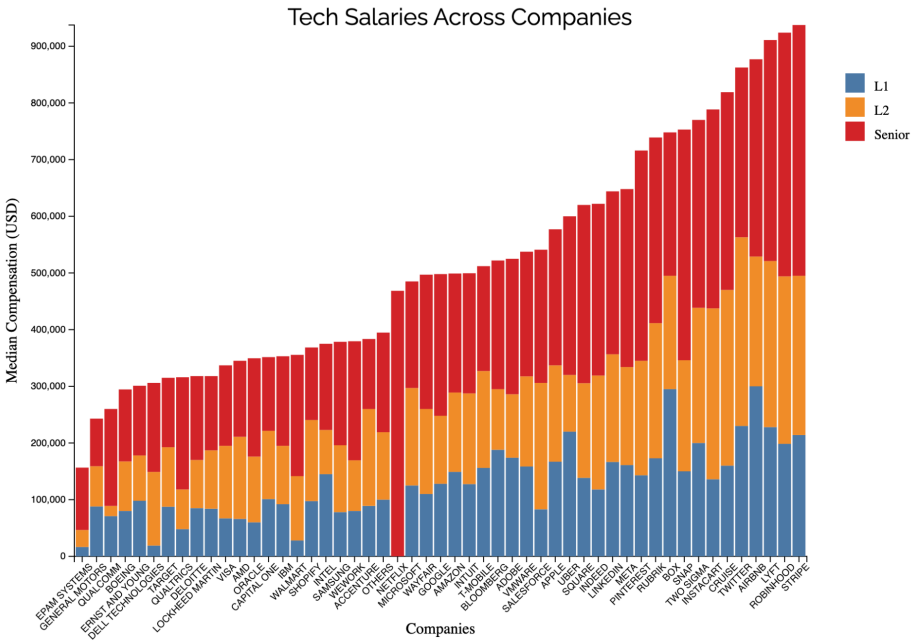
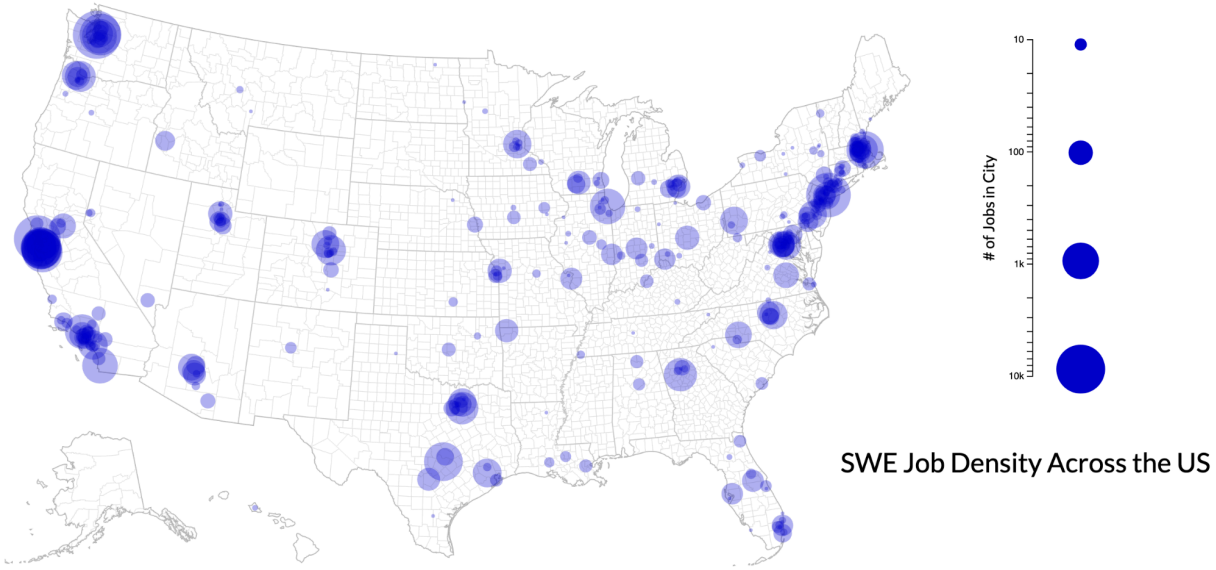


Chart 2



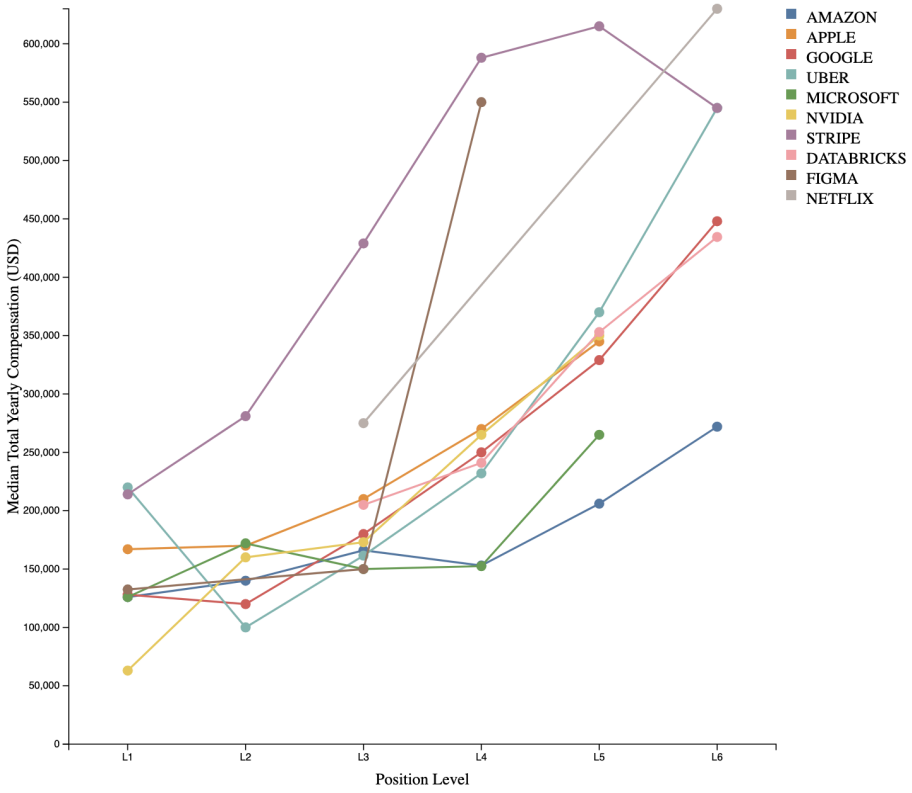
Vincent Jiang
Bahaa Korb
Akaash Mahinth

Chart 3



SWE Job Density Across the US

Chart 4



Vincent Jiang
Bahaa Korb
Akaash Mahinth

Dataset Parsing and Integration

Our data came from multiple sources. Our main source was a kaggle data set which contained statistics like pay, company, location, title, level, etc. for tens of thousands of software engineers (SWEs) across the world. This was supplemented by a couple of data sets containing information on US cities, including longitude/latitude, county, state, population median income, FIPS code, etc. The combination of these two data sets allowed us to take city data from the SWE information and further wring out macroscopic information at the county or state level.

We also used a topoJSON data set provided by D3 to create our maps. This data set contains the state and county borders in svg applicable formats, and is how we were able to make county specific graphics. There are three main variables taken from the SWE data set. We wanted to investigate the relationship between density of jobs, density of junior to senior roles, and expected pay all in the US. The first thing we did and the only major filter that occurred across both map graphs was a filtering of jobs that either existed outside the US, or couldn't be found in our data on cities in the US. These data entries were discarded, and only those with data correlating to our search were kept.

For the bar graph, the preprocessing involved handling the discrepancies between different companies and the leveling system, so a standardization was performed in order to standardize all the companies on one common scale. There also were so many companies that some were aggregated together into another section and we took the top 50 companies based on the number of datasets present for a certain company in order to give us a better representation of data. We also needed to handle some common abbreviations for companies such as MSFT for Microsoft, AMZN for AWS for amazon and other common abbreviations for big companies to standardize these into one company too rather than being considered different companies. Any intern data was removed as we cared mainly about full time engineers. We also filtered out any data that didn't follow the "Lx" format for levels where x is an integer in 1,2,3. An odd case is where all of our netflix data was only senior engineers so we had to specifically handle all netflix data in its own case. We processed the data so that we used medians of the salaries at each level to eliminate any outliers meaning we might give up on some larger data points, but this also allows us to see the bigger picture. A similar parsing was done for chart 4, as it also needed to take into account company abbreviations. Additionally, we only wanted positions that explicitly had level be a string of format "Lx" format, and didn't consider levels = "INTERN."

Vincent Jiang
Bahaa Korb
Akaash Mahinth

Design Rationale

Chart 1

Various steps were taken to iterate on the design to reach the final data visualization. Firstly, we wanted to portray location and geographic information of tech jobs in the US. We thought of potentially doing a choropleth map and took inspiration from the geoJSON lecture we had in class. We thought it best to represent the salary difference in terms of the saturation of the colors, but we had a big problem. Many of the data points had the exact same city location on the US map, some with lower salaries (low saturation) and others with high salaries (high saturation). With no opacity, salaries are covered on top of each other depending on how they are drawn. The information presented will therefore not be too useful. One approach is to instead take the median salary at a specific location, and then draw just one point per city. However, this isn't too useful since we won't be able to see a holistic picture of what cities have a high number of very high salaries, etc. Additionally, it made intuitive sense that each point represented a tech job, not just a summary of tech jobs at a location. Therefore, I did the "jittering" method by adding a random shift (depending on the area of the city) to points that were in the same city. Along with opacity, this generated a more wholistic picture of tech salaries across the US. After this step, I noticed that cities with many tech jobs would have a cluster of points. To make it easier to distinguish which cities are which, I added labels to point to each cluster for the top most populous point clusters. **To this end, the marks on this graphic are the circles, each representing a salary data point. The visual channels in this graphic are the horizontal and vertical position of the dots, indicating the location the data comes from, and the color of the data point, indicating the salary according to the scale.**

Chart 2

A lot of decisions had to be made and I iterated on a lot of sketch ideas before going with the final one. First I'll describe the marks and channels. The **marks** are the **bars** which represent the companies and levels. The **channels** are the **aligned and unaligned positions** since as we go up on the y-axis we can see that the salary increases, but we also see that another channel is **color hue** as the color changes as we transition from one level to the next. I wanted the different colors to be in a way where the transition is pretty visible to someone (assuming that they might not have any color blind deficiencies as discussed in lecture) so we had made the colors blue, orange, green to try and differentiate them somewhat with color blindness. Aside from the channels and marks I decided to go with a bar chart since this allows the easy color transitions while still illustrating on somewhat of an aligned scale the difference in salaries between companies. As for some design decisions I had to make along the way, the first big one was which companies I wanted to include vs aggregate, I decided the best way to go about this was using the companies that had the most data points as this can best tell the story as we have more data points to represent for these companies (Think law of large numbers, even though this is technically not entirely independent). The next thing I had to decide is whether I wanted the channel to color saturation or color hue, and I felt like color saturation didn't fit in really well when I went with that approach since the transitions weren't as

entirely clear so it made sense to go with color hue instead. I decided to add a legend due to this as it wasn't entirely clear what color corresponded to what level and I didn't want the viewer to just assume that the colors were in L1 -> L2 -> L3 order. Finally, I decided to go with medians to avoid outliers affecting/skewing the data too much. **To summarize, I mapped companies to the x-axis for the top 50 and aggregated the rest, salaries to y-axis, different levels to color hues.**

Chart 3

I used a map of the US with circles to inform the user on the Job density in different parts of the US. **The marks on graph 3 are the circles, which each represent a certain city. The channels used are the vertical and horizontal position and radius/area of the circles, giving us information on the location the data point represents and the population at that city.** The first design decision that I had to make was the max and min size for the circles in the graph. I settled on 0 to 25 pixels, as I felt this gave the best feel for large vs small cities while not confuscating where the circles were centered on the map. I further realized that it would not be possible to give a good range of circle sizes while using a linear scale, as there were many different cities at the low end, but very few cities near the maximum value. I decided to use a log scale for this reason. I chose to make opacity low so that it was clear when a location had multiple cities in near proximity that all had high job density. When the opacity was high it would hide data points that closely overlapped, and I didn't want this to happen as it removed information from the graphic. Finally the color was chosen to match a theme of blue that exists in both map graphics, and I chose a darker blue so that it contrasted well with the white background of the map. **In summary, I mapped populations of SWEs to their physical locations in the US, and adjusted the radii of the data points to illustrate the size of the industry in that location.**

Chart 4

We wanted to further illustrate the salary progression between levels and the difference between popular big tech companies in that regard. While chart 2 displayed 3 positions/levels (L1, L2, Senior) between companies and mostly compares companies, we wanted chart 4 to present the increase in salary between levels. Thus, it made sense to graph a line chart with salary (y-axis) with the position/level with increasing seniority (x-axis). **In this chart the marks are the lines on the graph. The visual channels used are the color of the lines, indicating the company the line represents, and the (x, y) positions in the graph representing the level of engineer and pay, respectively.** Some design decisions included how many levels to add. For simplicity L1 - L6 were added. Additionally, to make comparing different lines/company progression easier, I added a legend to match each company with a specific color. To not overcrowd, I decided to generate lines for only the 10 companies. Additionally, I added a point similar to a scatter plot, at each level and the corresponding salary to make it more obvious that

there indeed exists such levels at a certain company. In other words, sometimes companies may skip levels represented by no circular point at that level.

Visualization Story

There were three stories we wanted to tell through these graphs, the first being where in the US are the best paying SWE jobs, and where in the US are the most SWE jobs and which companies had the highest pays for SWE jobs. One of the main goals we aimed to understand using the map graphics was seeing if there were hidden gems that stood out in terms of low jobs/high pay (relatively unknown places to work), or if there were particularly bad places to be in terms of job opportunity and expected salary. Of course many of the expected outcomes held true, such as high amounts of jobs on the west coast, particularly in cities like LA, San Francisco, and Seattle. Similarly cities on the east coast typically held in high regard like NYC, Boston, and the DC area also show high job counts and high job pays. Some surprising results we found was that there was some density of tech jobs that we didn't expect such in Georgia, inside of Colorado, Chicago, Detroit and other states and cities we didn't entirely expect. This told us that there is a spread of tech jobs throughout the US outside of New York, California and Washington. We wanted to convey this to our viewer and show that the density of tech jobs is spreading sporadically throughout the US with competitive pay in some of the biggest top companies.

Team Contribution Breakdown

- **Akaash Mahinth: I** put together graph 3 which displays number of jobs vs city, and wrote most of the section on data description, preprocessing, and variables used. I also wrote the sections describing the story that is being told by graphs 1 and 3 and the data we found interesting there. Finally I wrote the design rationale for graph 3 and most of the visualization story. I would say I spent about 4-5 hours a week for the last 2.5 weeks on this project. Creating graphic 3 took the most time.
- **Bahaa Kotb:** I put together graph 2 which displays the companies, their associated levels distribution and their associated pays (and the designs) and worked on writing up the report in describing the data, the story, and a design overview rationale all for chart 2. Worked on polishing the visualizations up as well as telling the story on the actual page too. Out of all of these, developing chart 2 took the most time. I want to say I spent time everyday since the project came out until the deadline. I worked to incorporate the feedback we got from the demos on Friday's lecture into our charts and added the story to our visualizations on the actual website.
- **Vincent Jiang:** I put together charts 1 and 4 that display the tech salaries across the US and the salary progression across levels/positions for popular companies, respectively. I did a lot of the work in researching geoJSON and topoJSON in order to accurately chart

Vincent Jiang
Bahaa Korb
Akaash Mahinth

the US map, its states and counties. I also worked on polishing the code after receiving suggestions from our final presentation in class. I've spent time progressively for each MS and have thoroughly planned the design of my charts. Chart 4 is a suggestion made by a TA after MS 1, and I've worked to incorporate the feedback that we got to incorporate it to tell a separate aspect and story of the data that all 4 charts share.