Roy Bastien
Priyanka Parekh
Last updated 12/4/15

Process Book

Overview:

Our initial plan as of October 23rd:

The initial idea behind our project was visualizing the predicted salary for a job based on keywords in the job description. From our experiences, the majority of the time, job listings are not accompanied by estimated salaries, forcing the job seeker to visit external resources or ask others for information regarding their future salary. Additionally, individuals often times may consider how salaries for certain skills vary across regions or how developing certain skills will benefit them financially. Our tool will about for users to enter a skill, for example "Java" and then be able to see what the associated salaries are for that skill. Our decision to pursue this project is based on our personal frustration when searching for jobs and not being able to immediately garner the salary associated with said job. The initial problem, of predicting salaries based on keywords will be addressed as a part of our project for our machine learning class. We will use the Glassdoor API to collect job descriptions and parse these descriptions for keywords. And then we will create a training set based on salary data that we will also pull from Glassdoor and use that training set to predict future salaries based on provided keywords. While the work will be doing in ML is very interesting, in our opinion a key extension of this project would be enabling users to query our data set and interact with it. That is why we want to visualize this particular data set as our project for Visualization.

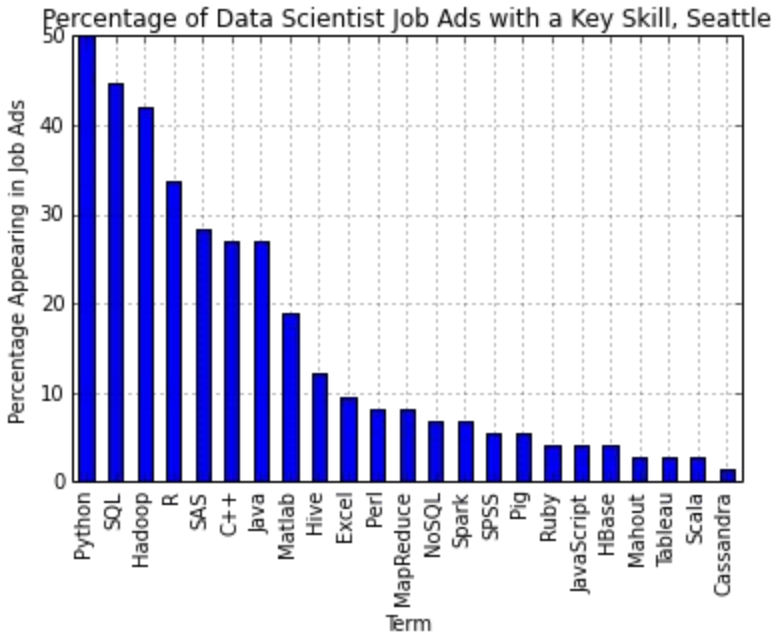
Plan as of November 13th:

After interacting with the Glassdoor API over the past several weeks and realizing its features and limitations we slightly refined our initial goal. We took a step back from depending on our learned data set from our machine learning class and rather decided to mainly utilize the data that is accessible from the Glassdoor API. Additionally, our initial view that we decided to develop was a view that allows users to enter a job title and then users are able to see the job titles users have subsequent to the entered job as well as salaries associated with these job titles as well as the 15 cities in America with the most jobs associated with the entered job title. Currently a user can center any job title and see the job progression, as is highlighted in a dropdown menu (which can be seen in

accompanying images below). Our plan it to populate this dropdown menu with other options, such as “Job Skills” which was what we talked about earlier.

Related Work:

Our initial inspiration for this project was this project: https://jessesw.com/Data-Science-Skills/.
In this project Indeed job posting data was scraped to find out which skills are most sought after when hiring a Data Scientist. Although the web scraping component of this project and the subsequent data that is presented is very interesting, the visualizations are static and rather basic. We want to build on the premise of this post and extend it to provide even more insights and allow for interactions with the data set. Below is an example of the type of data representation used in his project:



Additionally the current look of our project is modeled after an assignment from class. Keeping in mind the effectiveness of bar charts in communicating messages about data we wanted to use them whenever possible. As has been taught in class we try to make use of effective color schemes and additionally use color to signify magnitude within our bar charts. We make a use of a map as well in order to properly represent location data with magnitude representing the prevalence of jobs.

Questions:

Below are some questions that we hope our project is able to address or already addresses.

What is the next job for people in my current job?
What is the pay associated with that next job?
These two questions are addressed in our current visualization. Users are able to figure out answers to these question with a simple search. They are also able to see where the most job are associated with a certain job title.

What job titles are associated with a particular skill that I have?
Where are the most jobs associated with that job title?
How much do jobs associated with a skill that I have pay?
These three questions are the aim of the next step in our project. On top of the “Job Progression” dropdown we will allow for a “Job Skill” dropdown where users can find salaries and job titles associated with a particular skillset.

At a high level we hope users would be able to use our tool to gain insights as to that types of jobs they could get with certain skills, what a typical career trajectory could look like, the expected salaries of various jobs, and where these jobs are actually located.

Data:
We will be utilizing the Glassdoor API to access their data set. Use of their API simply required the creation of a Glassdoor account and a short approval process which we have already gone through. We will be able to make API calls to see job descriptions as well as salaries. Here is a link to the API overview: http://www.glassdoor.com/developer/index.htm.

We were able to fully understand how to use the API to get all of the data that we need to complete this project. Through various API calls which are constructed based on user input and our decisions regarding the types of data we want to visualize we are able to get any and all data that is relevant.

Exploratory Data Analysis:
Our initial plan as of October 23rd:
Because we want to visualize the "learned" data set we do expect a substantial amount of data processing however that will be completed as a part of our machine learning course. Through the use of the Glassdoor API, and web scraping (if the accessible data is not enough), we should be able to

garner enough data to predict a salary based on a feature(skill). The results from our work in for our machine learning course will serve as our input data set for this course.
In the unlikely event that our ML project is unsuccessful we will simply use existing data on Glassdoor to manually associate skills and salaries and visualize this aggregated data set. In this scenario the data acquisition may be time-intensive but we do not believe that data cleaning will be a major concern.

Plan as of November 13th:
Our current design and data usage is simply making use of the data that is gathered by making API calls based on user input into our search field. When we do have the results from our ML project we can also utilize that data within our project, by simply parsing that data set rather than the Glassdoor dataset but that would simply be a couple line code change. We do not want to limit ourselves to the dataset that is produced from our ML project, our current approach allows for the visualization of much more data allowing users a lot more flexibility and functionality. When we allows users to view a category of data one of the options could be viewing a learned dataset and in that case we will reference our ML dataset.

Design Evolution:
Our initial design sheets using the “Five Design-Sheet Methodology” can be found below followed by an implementation discussion with accompanying screenshots of our current design. We arrived on the use of bar charts and a map early on, we just have made slight refinement in the actual content we want to display in our different views and in our different visual elements.

As of mid November we are able to populate our bar charts and maps with data, additionally we have provided transitions for the bars in the graph. In the coming weeks we will extend the views to include information about different aspects of the Glassdoor data, the specifics of this is discussed elsewhere. We will also add further interactivity and other necessary linking between views of our visualization.

Sheet 1

How to search - dropdown w/criteria. search by:

- job title
- company
- state
- city
- keyword
- industry

Display Results: location on map
top company names
jobs available on Glassdoor
job progression -
what jobs did they get next
salaries of job progression.
jobs within a certain radius
available jobs within a company
salaries of available jobs

Salaries - national averages
industry averages
title averages

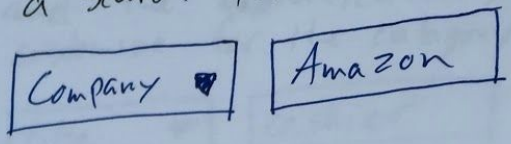
Title: Employment & Salary Visualization
Authors: Roy Bastien & Priyanka Parekh
Date: 10/22/15
Sheet: 1

Sheet 2

Layout

Operations

User selects catagory from drop box (ie Company, State, industry) and enters a search term.



Focus

Should the visivalizations change based on search catagory? If a state is selected the map should only display that state. If a city is selected, there may not be a map needed.

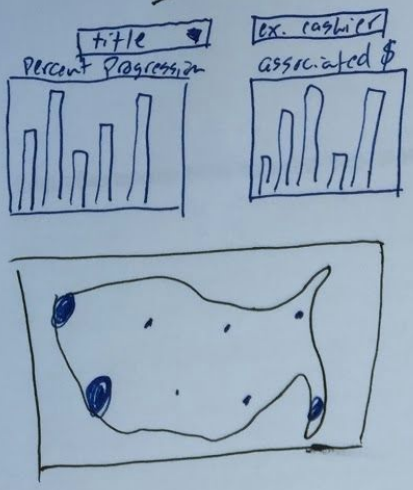
Discussion

A dynamic initial visual layout that changes for each search catagory will be extra work to implement but will add an additional layer of interaction with the user. Map and bargraphs will be the main display mediums.

Title: Employment & Salary Visualization.
Authors: Roy Bastien & Priyanka Parekh
Date: 10/20/15
Sheet: 2

Sheet 3

Layout



Operations

User selects search catagory from drop menu. The text box displays an example from the catagory and the visualizations customize for the catagory

title ▼ | cashier

Focus

Within the map, dots change size based on available positions across the country. Progression shows the jobs the title gets next along with associated salaries in a seperate(?) bar chart.

Discussion

Data should be populated in the charts before the user searches for a term. Display data should match the example given in the text field. Visualizations will change based on search catagory.

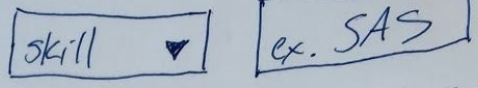
Title: Employment & Salary Visualization
Authors: Ray Bashien & Priyanka Parekh
Date: 10/22/15
Sheet: 3

Sheet 4

Layout

Operations

user selects the skill from a dropdown menu and enters text into the search field.



The display examples go to Skill with data and predicted job titles.

Focus

Discussion

Our algorithm will search job descriptions with the keyword and return the 10 highest paying titles with the entered keyword along with the salaries to display. Response time will be a challenge to effectively engage the user.

Title: Employment & Salary Visualization
Authors: Roy Bastien & Priyanka Parekh
Date: 10/22/15
Sheet: 4

Sheet 5

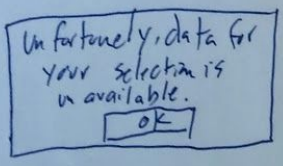
Layout

Operations

The final user inputs will consist of just the dropdown menu and a text field. If data is available for the selection, it will be displayed in the plots. Initial data will be displayed in the plots and selections will remain in the plots if user data is unavailable.

Focus

If data is unavialable for a given input, an alert box will be displayed in the center of the screen with an "ok" selection for the user to press to dismiss.



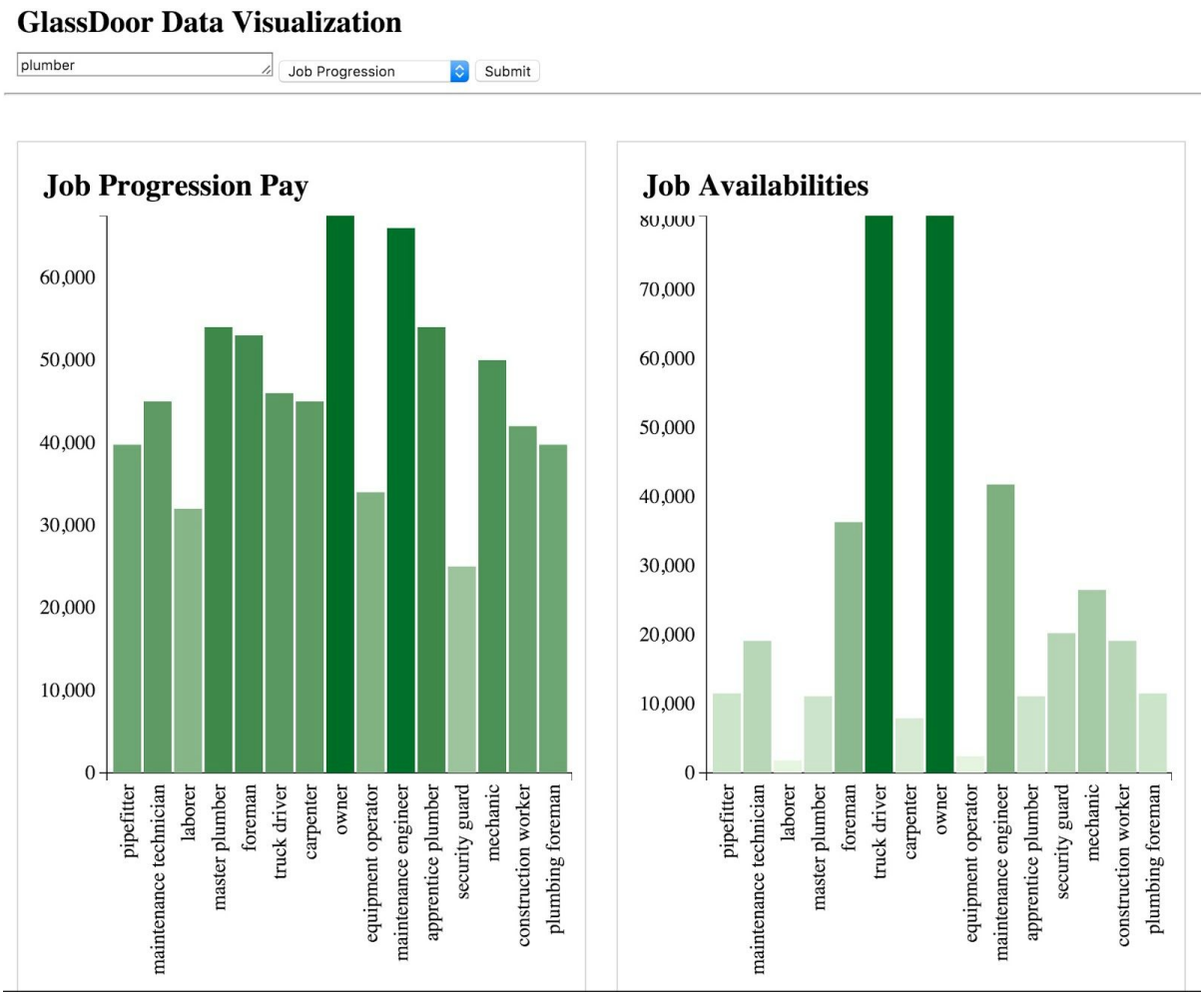
Detail

D3.js will be the software used for creating the majority of the necessary choices for visualizing the data. We will need to create a novel algorith for predicting the industry and job based on skill input. We will use a JSON format to interface between the model, view, and controller.

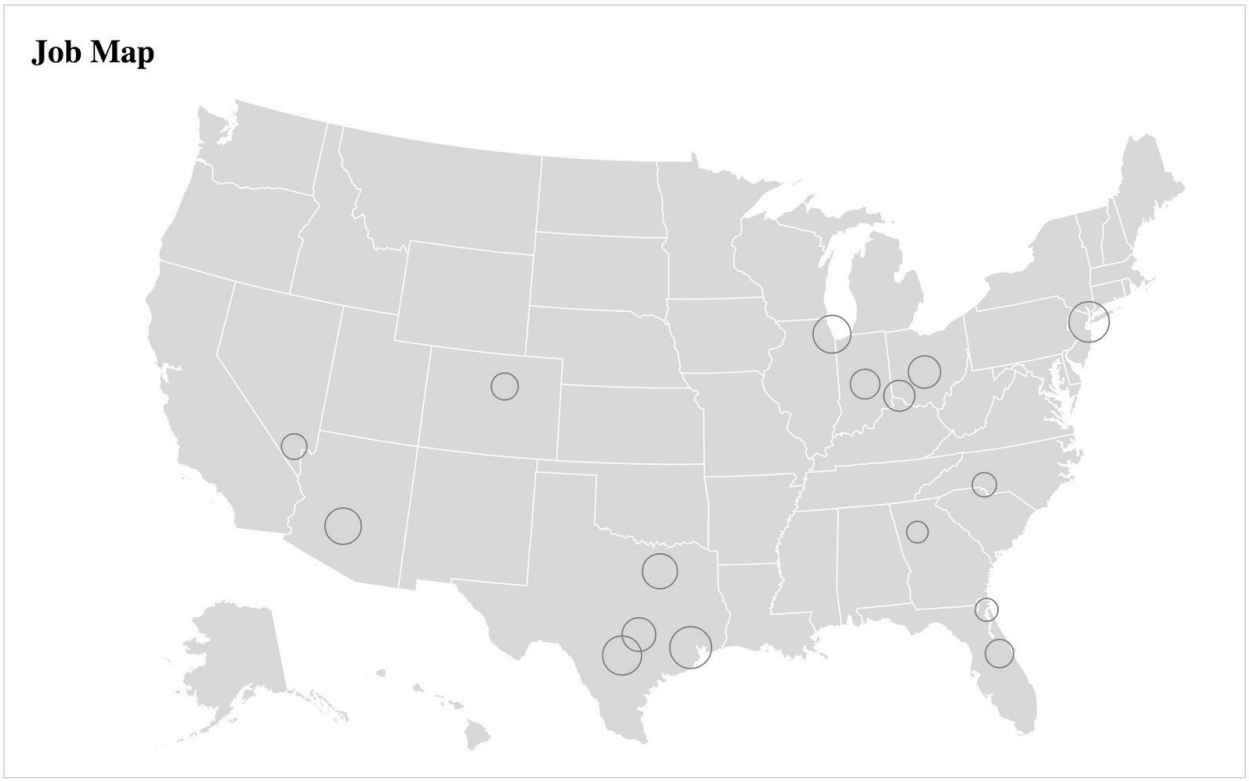
Title: Employment & Salary Visualization
Authors: Ray Bostick & Priyanka Parekh
Date: 10/22/15
Sheet: 5

Implementation:

In its current state, our implementation consists of three main subcomponents. First is the main form of user interaction made up of a textbox, a dropdown menu and a submit button. The user can enter a job title into the textarea, choose a search category from the dropdown menu, and press the submit button to submit a search. The user can also press enter while in textarea to submit a search. Once submitted, the search queries the GlassDoor API via an XMLHttpRequest object. The XMLHttpRequest requires a custom URL upon which it receives a response via a "GET" request. The response is parsed to a JSON object for visualization by our second main subcomponent consisting of a pair of bar graphs. Currently, the tool only supports the "Job Progression" category of search and the associated bar graphs each show the 15 positions that an employee is most likely to have following the queried job title. One bar graph shows the pay associated with these 15 jobs while the other shows the number of positions available nationwide for each of the 15 job titles. For added visual effect, transitions were added for the bars in the bar graphs. The search capability and the bar graphs are shown below.

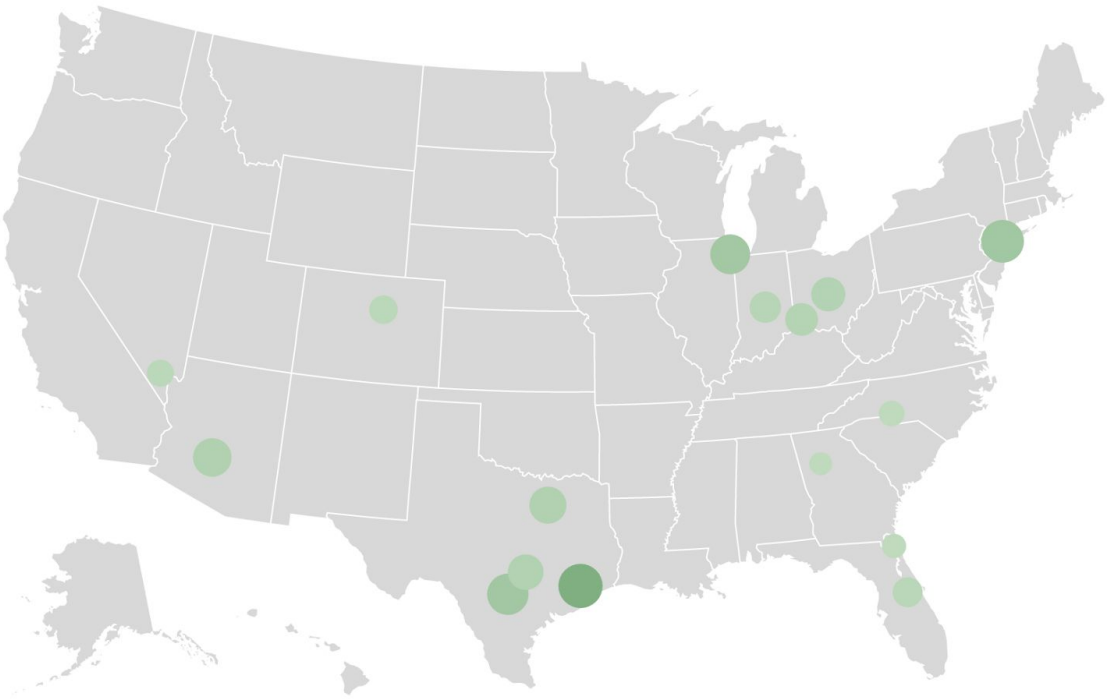


The final subcomponent implemented so far is a map of the US. This map displays circles around the 15 US cities with the most job openings for the searched job title. The circles are size relevant with the largest circle representing the city with the most jobs and the size of the circle declining with the number of jobs in the other cities. The map is shown below.



We went on to be able to update the map with colors circles based on the prevalence of jobs in differents areas. This added feature can be seen below with the accompanying title that depicts the data that is being shown in the map:

Cashier Job Map



Feedback:

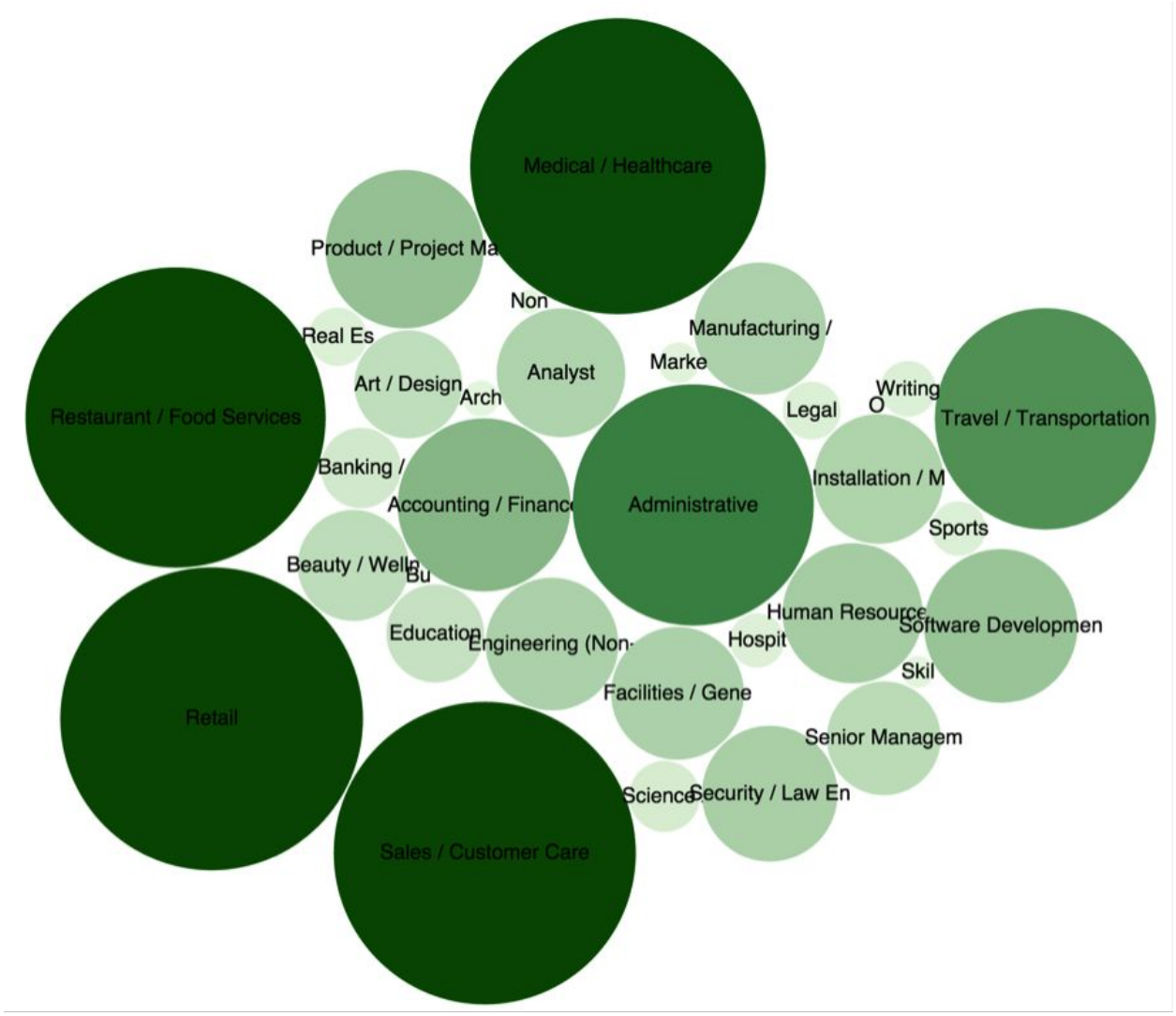
There were numerous suggestions for possible improvements or deviations given to us during our feedback review session with our assigned TA, Zinnia Mukherjee. The main takeaways from our session are outlined below:

- Add a visualizing that addresses how different industries compare
    - Based on salary
    - Based on job availability
- Be able to click on the map and see some sort of meaning information, such as the highest paying jobs in a certain location
- Move away from solely relying on bar charts. Incorporate a bubble diagram where volume corresponds to prevalence.

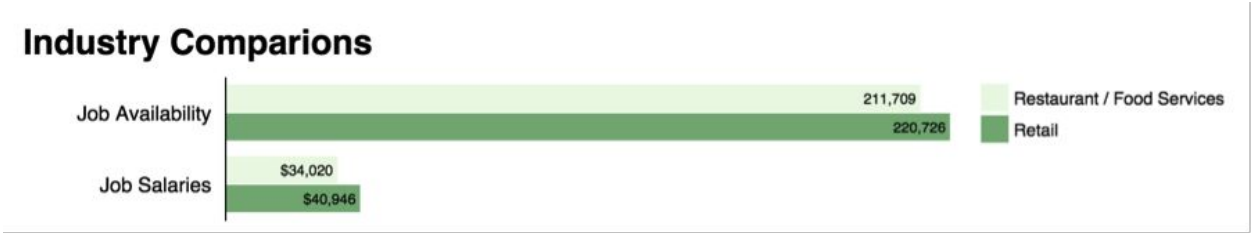
We took this feedback to heart and included all of the suggested ideas in our design moving forward. We added an additional div that contains a bubble diagram which shows each of the 32 industries Glassdoor contains data for and where volume corresponds to job availability in each of those

industries. Users are able to click on an industry and get information about associated locations, salaries, and job titles. Additionally, we added the ability to interact with our map and glean even more information about job salaries in a given area. We also added the ability to compare job salary and job availability information between two industries at a time via the introduction of another set of bar charts.

Our bubble chart:



Our comparison chart:



The interactivity that was suggested, and that we proceeded to include is better showcased in our screencast, which is embedded within our website.

Evaluation:

We were able to answer certains questions we were not planning on being able to answer based on the inclusion of feedback from our TA, such as how different industries compare in terms job availability and job salaries across industries. We took a slight deviation from visualizing skill based data to rather visualization industry, job title, and location based data based on the data set that was accessible using the Glassdoor API. Using the API, had its advantages and its disadvantages. At a high level it was great to be able to have free, somewhat convenient access to their data set. However accessing the data we wanted for our visualization was not always trivial because of the way that the API was configured. Additionally, we found when implementing our industry comparison chart that we were

We are able to gain interesting insights and were able to see surprising trends with the use of our visualization tool. One trend that found particularly interesting was the fact that in NYC the majority of the jobs with the greatest availability were jobs surrounding Finance, something that was not expected however, was that alongside typical Finance related jobs, there was also a high job number of barista jobs available in NYC as well. The same phenomenon could be occurred with the prevalence of Software Engineering related jobs in Seattle and the high availability of barista jobs there as well.

A limitation of the API is that when attempting to do industry comparisons we are often met with “access denied” errors because of what we believed to be unpublished request rate limits.

Another problem we were ran into was making the content of the industry comparison tool change based on the selected industries, going forward this is something we would definitely improve.