Project Title: "Indeed, We Want to Be Data Scientists"
Team Members: Clay Selleck, Kyna Thorberg, Jodi Heen
Project Description: Using public data sets, we will identify major trends in data science job postings.

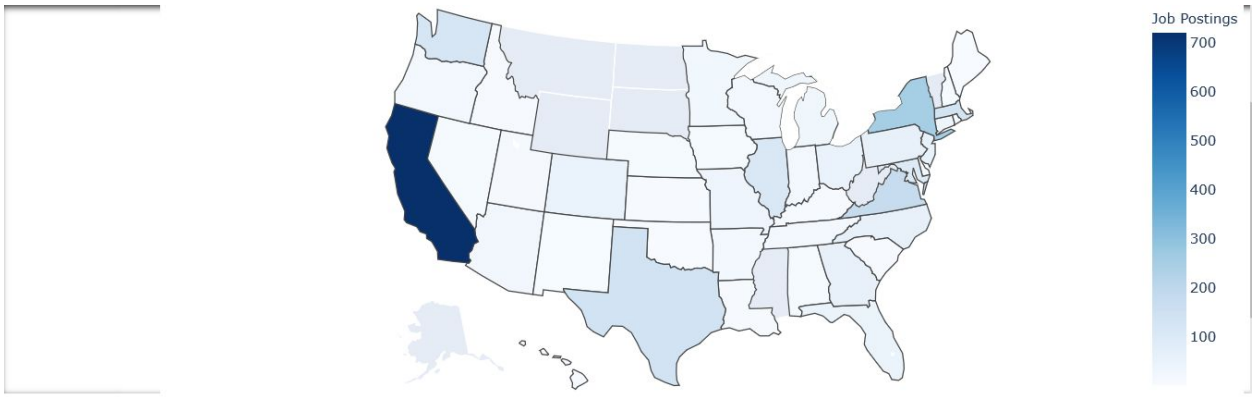
Question One: How many job postings are there for "data scientists"?

Our big question: "can we find data that validates if data science jobs are growing in the marketplace?" Our search was inspired by the idea of using the Indeed job board website to find live or recent data through their API. We were not able to get authorized for this, however, but did identify public data sets that had collected thousands of job postings that included "data science" jobs. In reflecting on our question, we never specified how many jobs would adequately answer it. Once we found two data sets that totaled around 15,000 rows of data, however, we were satisfied that there were enough job posting data to start analyzing it for more specific findings related to job locations, keyword skills in the posting, and the industries who were hiring.

Notes for future study and research

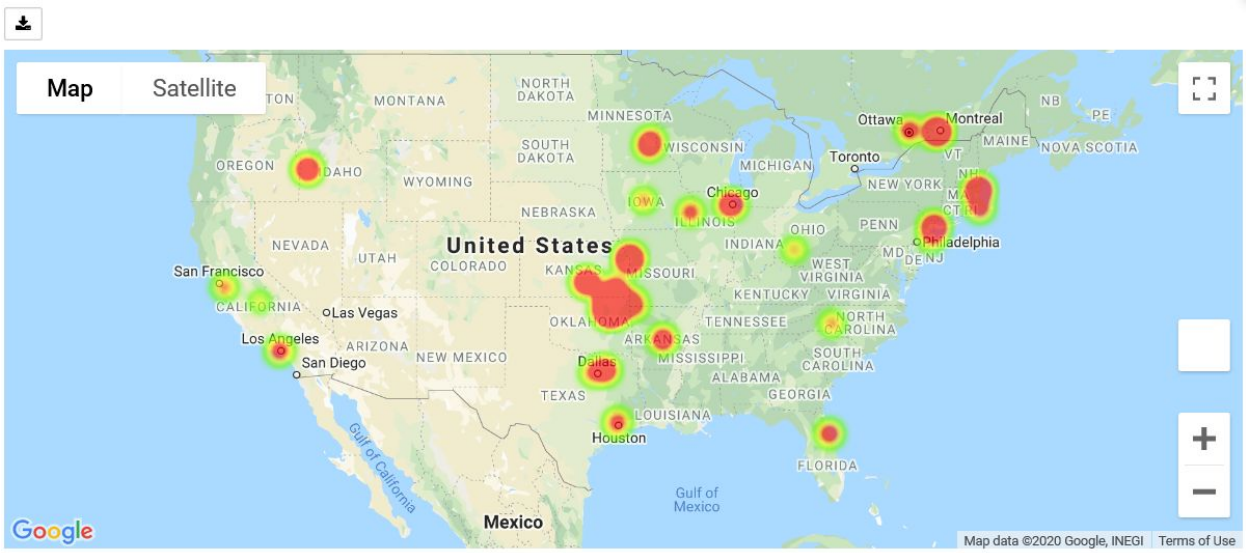
If we were to revisit this question or conduct further research, we would likely try to find some comparison data, to see if this amount of job postings for a period of time was unique and if we could center our search on a specific area (such as the Greater St. Louis area) and see how many "data scientist" jobs there were compared to all results for that area.

We also did not necessarily identify a standard hypothesis regarding where jobs might be located. We operated out of the general understanding that "tech jobs" are found in the highest volumes in California and New York (the "Silicon Valley" and "Silicon Alley", respectively). When visualizing these using Matplotlib, we found that our choropleth map (below, with a blue color range) that was generated using a data set of only "Data Scientist" jobs, did match these assumptions.



We learned that sometimes a data set can be limited, however, when we used an alternate data set that had a greater variety in the types of jobs posted. The heat map (below) that was generated showed hot spots of job opportunities that did include job openings in San Francisco and New York, but also a high volume in the center of the United States. This was a good reminder to be aware to investigate how a

data set was produced and understand that its contents needs to be verified with other similar sets of data.



Question Two: Have those increased over time?

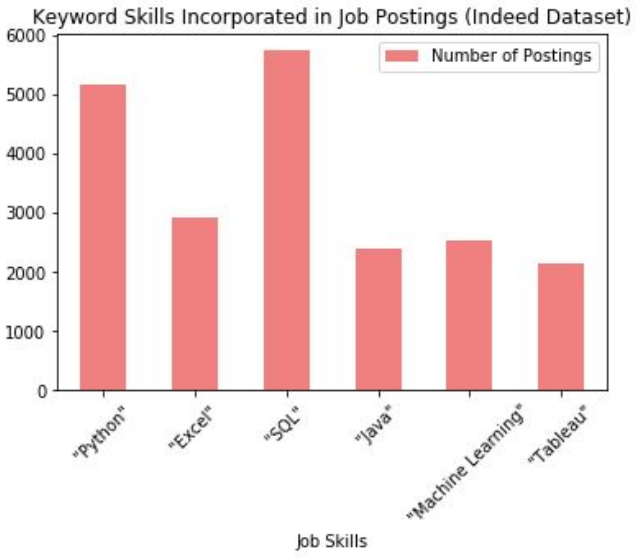
This was a question we regret we were not able to answer in the scope we had hoped for. Initially, we had aimed to find ten years' worth of job listings or data from the Bureau of Labor Statistics. One of our data sets did capture when a job opening was posted, however. These entries were captured into a line graph (left, entitled "Job Opportunities Over Time (Feb - Sept 2019)") that served as more of a snapshot of how many data scientist roles might be posted at one time. This data was dependent on how avidly and consistently the researcher was collecting the job postings and adding them to the data set, however, not necessarily on how many job opportunities were altogether available.

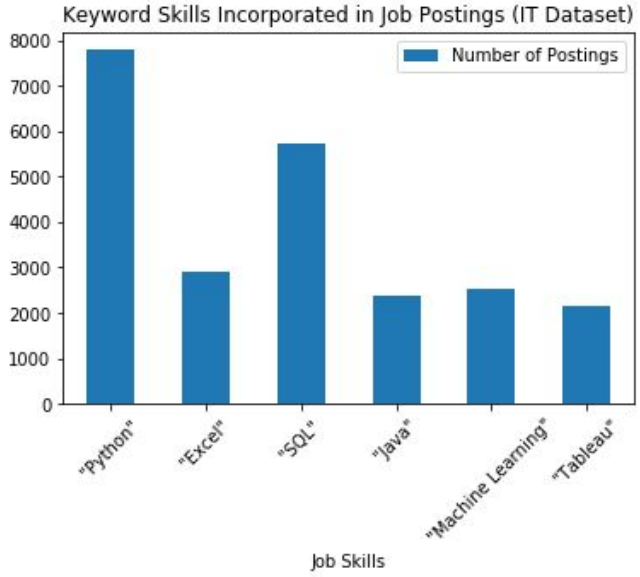
Notes for Future Study and Research:

Wrangling and cleaning the data was not only important for creating our project findings in the short term, but it gave insight for the future as we collect data. We saw that helpful documentation and streamlined, standardized, and clean data is much more usable in sharing insights to others, else they are left to try and make more decisions when trying to gain insights. Although this was a lesson we understood in theory before, we understand it in practice now.

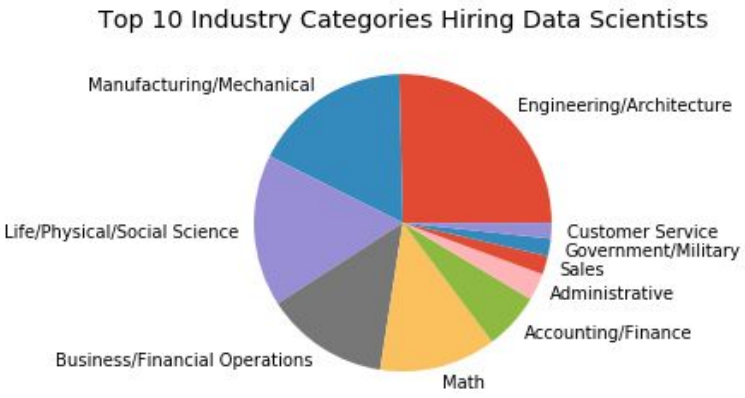
Question Three: What skills are sought out?

The primary, pragmatic motivation for this question was how this data could help us as prospective data scientists acquire jobs in the future. Using our common understanding of job postings and their inclusion of desired skills, often with keywords, we wanted to see if there were any trends in what skills were commonly sought out on these listings. Using a list of skills that will be covered in our course curriculum (with the exception of "Java"), we were able to find the number of times a specific skill keyword was included in a job description. The findings are in the two sets of bar charts (below) and show comparable findings. Since Python and MySQL are both covered in our course curriculum, we not only expected to see them in high numbers but were also very relieved to see that our current investment looks like it will be sought out by recruiters and employers looking to hire data scientists in the coming months.





This initial question also led us to the question: "what industries and companies are hiring data scientists?" in order to see if these were varied or if they were mostly within the technology sector.



Breaking down the industries that had job postings in our data set showed that this result varied (see left, a pie chart entitled, "Top 10 Industry Categories Hiring Data Scientists").

Notes for Future Research and Study:
Another chart (right, entitled "Top 10 Companies Hiring Data Scientists") showed that the top companies who had contributed job postings were mostly recruiting companies (namely, CyberCoders and Robert Half Technology). If we were to do further research, we would likely try to create a comparison chart that excluded recruiters.

As we all concluded, however, we want to be data scientists and we would take a role posted by a recruiter or one posted by a company if it was the right opportunity!