Overall Conclusion:

This project was an iterative process to steadily improve the quality of the data imported, processed, and analyzed. There was a discovery process involved in scraping job websites and as to which processes worked and which didn't work. New tools were discovered such as bulkdachecker and the library boilerpipeR. Techniques were implemented to better capture additional keywords. A database process was setup that is flexible to handle many-to-many relationships and setup a robust database structure five tables and a database process that could be run repeatedly and eliminate any duplicates during the load process. Also the database was setup for web access using amazon web services.

Extensive graphs were provided to evaluate the data. Overall the top skills were communication tops the list, followed by Excel, JAVA, SQL, machine learning, and Python. For data scientists Machine learning, Excel, Python, statistics, Hadoop, R, and communication were the top skills found. Job postings were analyzed for top cities (Seattle, Chicago, and Plano,Texas) and states (Washington, California, and Texas) offering data science jobs. There was an interactive graph to view the most popular data science skills by a user selected city using the R datatable function. Predictive analytics (ctree, predict, and random forest) were done to show what role you are bested suited for based on your skills (What am I ?). Techniques were shown to improve the accuracy of the models.

And the group learned to work together as a data science team, sharing technologies and techniques learned along the way, expanding skills in collaboration, planning, and delivering on time within a structure as directed by the project manager.