Hieu Nguyen, U92828434
Data Science Report

In the project proposal, I said that I would be analyzing a dataset of Data Science job salaries. The questions that I was trying to answer were "what data science jobs have the highest pay range? What data science jobs are the most similar in pay? How popular are certain jobs within a salary range?" The first thing that I did was clean the data. I didn't really have to do much since the data was already clean. After cleaning the data, I partitioned the data in the CSV file by putting it into a list of tuples. This is what the functions open_file(), check_csv(), and read_file() do. Once I had the data partitioned, I created a function graph() that takes in the vector of tuples and created an adjacency list. This adjacency list had what jobs were connected to what salary range. Once I had the graph created, I began writing the functions that would help me answer my questions. The first of these functions was the highest_weight function. This function found which job node had the highest weight connected to the salary range. I computed this weight by finding how many one specific job was connected to that specific salary range. For example, if there were 10 "data analyst" jobs in the range "<$40,000" and there were nodes connected to that salary range node, then the weight would be 10/20 for data analyst. The connected_to() and weight() functions are helper functions for highest_weight(). What this helped answer was "How popular are certain jobs within a salary range?" because the weights represent frequency of a job connected to a specific salary range. This function had a complexity of O(n) since it was only 2 for-loops through the graph. The most_connected() found the job node with the most connections to other salary range nodes. This answered the question "What data science job has the highest pay range?" and the function had a complexity of O(n) since it was only 2 for-loops. The last function most_similar() found 2 jobs that were the most similar in pay for a given salary range, which answered the third question. This function had a complexity of O(n^2) since there were 2 for-loops nested. With all the analysis done, I was able to answer that for this specific dataset, the job with the highest weight to the highest salary range was "Data Engineer", the job with the highest connection to other salary nodes was "Machine Learning Engineer", and the 2 jobs that were the most similar to each other in the salary range "$240,000+" are "Research Scientists" and "Machine Learning Engineer." In doing this project, I learned a lot about the diversity of data science related jobs and how connected/distributed salaries were for different jobs. I also implemented a lot of the techniques learned in class like graph representations, borrowing and ownership, BTreeMaps, sorting algorithms, iterators, and much more.