CPSC 4030 Project Report – Group 11

Jackson Van Hyning

**Overview and Motivation**

It is no secret that in the modern day, the process of searching for a job is tedious, time-consuming, and often frustrating. The increasing usage of AI screening algorithms by employers has made it more difficult for an application to even be reviewed by a real person. On top of this, once a person actually secures a job at a company, it is increasingly common for them to face an unhealthy work-life balance, unreasonable deadlines and expectations, and generally poor working conditions. Because of this, it is important for a person to utilize every tool available to assist them when looking for jobs. While many services, such as Glassdoor, have been created for this exact purpose, it can be beneficial to look at the statistics of job positions in a field of interest more in-depth in order to get a better understanding of the environment of that industry. While I personally am not interested in pursuing a career in Data Analytics, examining the trends shown in statistics from job listings for this field is not only relevant to this class, but serves as an example of how doing so can be helpful for any field that someone is interested in working in.

**Related Work**

I have used Glassdoor as a tool for job searching in the past, and thus I was interested to see how data scraped from this website could be compiled and used to visualize trends in a field, and how doing so can provide information about the landscape of jobs in a particular field that Glassdoor itself cannot.

**Questions**

In the process of searching for a job, determining which of the many job listings you find are actually viable places of employment is a complicated process that requires the consideration of many factors. While I can not hope to entirely encapsulate this dilemma within the scope of this project, some questions I hope to shed light on through the visualizations I implement are:

* How do factors such as company’s size and sector contribute to its rating?
* Are the highest-rated companies to work at centralized in a certain area?
* Are there visible trends between job title and salary estimate?

**Data**

The data was sourced from the “Data Analyst Jobs” dataset posted by user ‘Larxel’ on Kaggle. While this data was scraped from job listings on Glassdoor.com, the data has already been compiled into a CSV and thus no further scraping was required. The data cleanup process involved the following:

* Excluding the ‘Job Description’, ‘Competitors’, and ‘Easy Apply’ attributes as their values were either inconsistently present or not relevant for answering any of the questions posed.
* Utilized a Python script for most of the cleanup. This script will be present in the project directory (*cleanup.py*):
  + Modified column titles to better describe the attributes
  + Change ‘Salary Estimate’ formatting from (*X*K – *Y*K, Glassdoor est.) to an average of the salary range as a double (e.g. 37K – 66K becomes $51,500.00)
  + Removed city from ‘Location’ (e.g. ‘Los Angeles, CA’ becomes ‘CA’)
  + Removed state from ‘Headquarters) (e.g. ‘Los Angeles, CA’ becomes ‘Los Angeles)
  + Calculated a new attribute ‘Overall Score’ using a weighted sum of ‘Salary Estimate’ and ‘Company Rating’, with weights of 0.7 and 0.3 respectively, with the Overall Score being a range from 0 - 10. This attribute can be used to more easily show correlations between the desirable attributes of a company (high ‘Salary Estimate’ and high ‘Company Rating’) and other attributes. While this attribute could be considered to be somewhat redundant with the ‘Company Rating’ attribute already present, this accounts for situations in which a company might have a good working culture but offer relatively poor pay, which could dissuade someone from wanting to seek employment there.
    - Formula = \* 10
    - Note: For items in which a ‘Company Rating’ is not present, the above calculation is performed with ‘Salary Estimate’ alone instead
  + Removed a redundant value of ‘Company Rating’ present in the ‘Company Name’ column
  + Standardized formatting of the ‘Size’ and ‘Revenue’ attributes (actual values are unchanged)

**Exploratory Data Analysis**

One of the first visualizations I considered was a scatter plot that showed the correlation between the number of employees at a company and the aforementioned computer ‘quality score’, aiming to answer the question “How do factors such as company’s size and sector contribute to its rating?”. However, after implementing this, it became clear that this was a suboptimal choice for two reasons. One, most companies had an employee count that ranged from 100-1,000, however because the maximum count present in the data was 10,000, it was hard to plot this in a way that was consistent and easy to interpret. Additionally, while there was a trend of companies with more employees having a shorter range of quality scores that were generally higher, the data as a whole did not show a clear enough correlation between the two that it felt meaningful to include on the website.

**Design Evolution**

**Implementation**

**Evaluation**