CPSC 4030 Project Report – Group 11

Jackson Van Hyning

**Introductory Note**

This project was developed under unusual circumstances. Due to a complete lack of responsiveness by former group members, it became necessary for the group to be split up and this project to be worked on independently. Due to this, the project was developed from the beginning starting around the middle of the semester, and so some of the intended milestones of the project, such as creating visualizations in Tableau before implementing in D3, were not completed. Thus, there is a relative lack of information in this report about the evolution of the project over time as it was imperative to finish the project in a short span of time, as well as without help or collaboration from others.

**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 cannot 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?
* What factors of a job listing are most important to pay attention to when searching for a job?

**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 ‘Quality Score’ using a weighted sum of ‘Salary Estimate’ and ‘Company Rating’, with weights of 0.4 and 0.6 respectively, with the Quality 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**

As I explored the data, my main focus was on seeing if I could find trends that could be useful to be aware of when searching for a job, such as a correlation between job title and salary estimate, or between company size and rating (as seen in the questions asked above). Thus, I contemplated which attributes could be meaningfully compared in order to show trends such as these. One attribute that is important about any job listing is the salary, so I wanted to see which factors displayed meaningful trends between them and the salary estimate. Additionally, I wanted to see which factors were correlated with higher company ratings, as the pay of a job is important, but the quality of the work environment is important as well. While I did not create any visualizations to look at the data initially, I considered which attributes would logically have an influence on one another, and which of these could be easily and clearly visualized.

**Design Evolution**

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.

The first visualization I decided to include was one that displayed how the presence of certain keywords in the job title of a listing correlated to the salary estimate for that listing. This aims to answer the question “Are there visible trends between job title and salary estimate?”. While I originally planned to instead use full job titles, the fact that many of the titles included extra words or information such as the company name made this more difficult. Thus, I decided to filter the data based on keywords (such as “Senior” or “Financial”) to see if there was a correlation between those and the salary estimate for that listing.

Next, I wanted to create a visualization that showed the attributes that contributed to the computed “Quality Score”. Because I had just created a visualization using a categorical attribute, I wanted to compare the quality score with another numerical attribute. This led me to “Company Size”. This comparison is meaningful because it displays the difference in the opinions of employees who work at large corporations versus smaller startups (as well as in between). There are many attributes that could also be visualized to show their correlation with quality score, however I felt like this was the most meaningful one as the size of a company is going to be an important factor to consider about any job listing, in any location, in any sector.

Finally, I included a visualization showing the most common locations for job listings among the data using a map of the US. While I initially wanted to do this through a heat map, this ended up not being an effective approach because there were many status throughout the US that did not have any job listings. Therefore, I decided to use circles on the map with a size proportional to the amount of job listings for that state. Only states that had a number of job listings above a certain threshold were included, because a state with only 1 or 2 listings did not seem meaningful enough to include in the visualization. Similarly, listings in locations outside of the United States were excluded due to their relative infrequency

Afterwards, I decided to add to the first visualization. There is now the option to view the correlation between both the keywords of a job listing and the industry that job listing is in and average salary estimate. This provides more information about the trends within the dataset while still remaining relevant to this particular visualization.

**Implementation**

* Visualization 1: “Salary Estimate by Job Title Keyword”/ “Salary Estimate by Company Industry” – This is a visualization of the correlation between the presence of certain keywords in the job title of a job listing and the salary estimate for that listing. This is visualized using a bar graph, with the bars being colored on a gradient depending on the amount of listings containing that keyword. When moused over, the bars display a tooltip displaying an example of a full job title that contains that keyword, as well as the number of listings in the dataset containing the keyword and the average salary estimate in USD. There is also a dropdown menu that can be used to visualize the relationship between company industry and average salary estimate instead

A green and brown graph

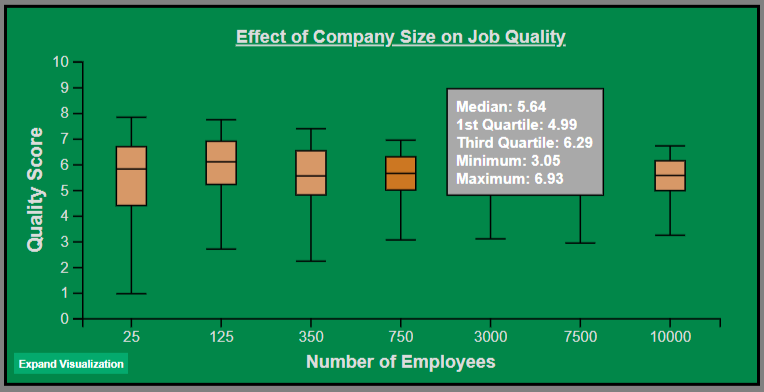
Description automatically generatedA graph showing a number of keywords

Description automatically generated

A green and brown graph

Description automatically generated

* Visualization 2: “Effect of Company Size on Job Quality” – This is a visualization between the size of the company of a job listing and the computed “Quality Score” for that listing. As previously mentioned, quality score is a measure of the desirable attributes of a position at a company – salary estimate and company rating – found using a weighted sum of the two in comparison to the minimum and maximum values of each. This is visualized using a box and whisker plot, with company size measured in number of employees and grouped into categories. Similar to visualization 1, a tooltip is displayed when a box is moused over displaying the summary statistics for that group, including median, 1st quartile, 3rd quartile, minimum, and maximum.

A chart of a company size

Description automatically generated

* A map of the united states with orange dots

  Description automatically generatedA map of the united states with many dots

  Description automatically generatedVisualization 3: “Common Locations of Data Analyst Job Listings” – This is a visualization that shows the most common locations in the US for job listings. States with a number of listings above a certain threshold are represented with a circle over the state, with a size determined by the frequency of that location among the dataset. When moused over, a tooltip showing the name of the state and number of listings for that location in the dataset is visible.

**Evaluation**

The visualizations I created using the data taught me some important lessons. The trends that were shown in the visualizations showed that there were correlations between certain attributes like I expected, but they were not always in the way I expected. One example is in visualization 2. I expected the listings from smaller companies to have a higher quality score, however what the visualization showed was that listings from larger companies have a lower median and maximum score and higher minimum score, meaning that while the best job listings from larger companies will have a lower score on average, the variance between the best and worst scores from larger companies is smaller than that of smaller companies. Another such lesson was shown in visualization 1, which was that there are certain job titles that have a noticeably higher average salary estimate that I did not expect, such as “Marketing Data Analyst” or “Data Warehouse Analyst”. I think my visualization does a fairly good job at the goal it was trying to accomplish, which is showing trends between attributes of Data Analyst job listings that can be informative when searching for a job in this field. I think it would be improved by having a larger dataset, such as including non-US locations in the third visualization. Overall, I think the visualizations do a good job of accomplishing the main goal of the project, which was to demonstrate how examining data about job listings for a field can show trends between important attributes, allowed one to inform themselves about what to pay close attention to when searching for a job.