CS 3300 - Project 1 Report

Alice Meng (aym27), Joe Antonakakis (jma353)

1 Running Our Submission

To run our submission, start a server. If you have Python installed, simply type

```
python -m SimpleHTTPServer 8080
```

into your terminal. Then open up

```
http://localhost:8080/index.html
```

in a Google Chrome browser.

2 Data Report

2.1 Data Description

We wanted to visualize data from the healthcare sector because healthcare is a widespread topic of interest today. After some research, we decided to use two main datasets - the median salaries of doctors in hospitals across the nation and the mean readmission rate for pneumonia and for heart attack in each state. We did not find preexisting formatted data online, so we scraped the salaries from `http://www1.salary.com/Physician-Hospitalist-salary.html` and the readmission rates from `http://www.hospital-data.com/` (see section 3 for all sources used). Variables included readmission rates on a state-wide and nation-wide level, as well as readmission rates for specific hospitals in the best and worst states according to an algorithm we used. Variables also included salaries on a town and state level for Hospitalists (general physicians in hospitals).

The `state_readmission_avgs.json` contains the readmission rate data we manually copied and pasted from the hospital site [3]. The `id_to_state_.json` file contains a mapping of state IDs to states and was formatted from `https://gist.github.com/mbostock/4090846#file-us-state-names-tsv`; we needed this data to fill in the states with appropriate colors using d3. The `final_salaries.json` contains salaries of hospital doctors that was scraped from the salary site [4]. We filtered the data while we were scraping - we mainly collected only the data that we wanted to use in our visualizations (see section 4 for additional details).

These datasets were integrated via the algorithm described in section 2.2.

2.2 Mapping from Data to Visual Elements

Geographic Color Mapping

We created two U.S. maps - one to visualize data on pneumonia readmission rates and one to visualize data on heart attack readmission rates. We mapped the algorithmic score of each state to a color scale ranging from a green (#1a9641) to vanilla (#ffffbf) to red (#d7191c). The algorithmic score of a state was computed

by taking the linear of combination of the normalized [pneumonia or heart attack] median readmission rate and the normalized median salary of hospital doctors in the given state. Further details of how the score was calculated are described below.

The average readmission rate was taken directly from the hospital site [3]. We iterated through all the states and calculated the national median for readmission rates and for salaries. For the salaries, we scraped the median salaries in each town for each state. Thus, we iterated through each town, calculated the state median salary, and then iterated through each state to calculate the national median salary of hospital doctors. A normalized value simply required dividing a state's data value by the national data value. A normalized score of 1, greater than 1, and less than 1 means that the state's value is equivalent to the national median, greater than the national median, and less than the national median, respectively.

With a normalized state salary s and a normalized [pneumonia or heart attack] readmission rate r, an algorithmic score $score$ was computed using the following formula:

$$score = 0.4 * s + 0.6 * r$$

Lower scores are better because doctors are likely being paid less than the national average *and* readmission rates are likely lower than the national average. Higher scores are worse because doctors are likely being paid more than the national average even though the readmission rates are likely higher than the national average. We weighted the readmission rate slightly more than the salary because we felt that the readmission rate is more indicative of the "goodness" score than the salary is.

Bar Graph Mapping

In addition to our heat maps of the U.S., we created bar graphs for the best scoring and worst scoring states for both pneumonia readmission and heart attack readmission. The x-axes are labeled with the towns within the state, the y-axes are labeled with the algorithmic score, and the height of each bar corresponds to the algorithmic score of a specific town - in this case, the algorithmic score is computed using specific town statistics normalized to the nation's statistics. For towns with multiple hospitals, we averaged out the hospitals in the area to create one score for that town. This was done for all states. The towns listed were towns we had salary information *and* readmission information for. This was our way of directly ranking specific locales in terms of our algorithm for the two most extreme states.

For these graphs, we limited our y-axis to show values from 0.8 to 1.4 because this is where a lot of our data fell, in terms of our algorithm. By limiting the axis range, a greater discrepency was visible across towns of one state and across states in general, in terms of our algorithmic score.

2.3 Our Story

Our visualization captures the relationship between how much hospital doctors are being paid and the readmission rates for pneumonia and for heart attack. We wanted to see if there were any interesting correlations - state(s) that have doctors who are being paid less than average and have lower readmission rates and state(s) that have doctors who are being paid more than average but have higher readmission rates.

Somewhat surprisingly, New Jersey had the worst scores for both pneumonia and heart attack. New Jersey is known for attracting the "best" doctors around the country, which justifies why their median salary is much higher than the national average. However, even if the doctors perform their job well, there may be many other factors contributing to the high readmission rates. A few possible factors are stress and air quality. Other cultural and environmental factors may also contribute to the high readmission rates.

On the other hand, South Dakota and Montana had the best scores for pneumonia and for heart attack, respectively. Doctors are paid less in both states possibly due to the fact that they are remote states - the population densities are not very high. Again, possible factors for lower readmission rates may be lower

stress levels and better air quality. Since the states are not densely populated, the air quality is likely to be much better than that of an east coast city. The lower salary and the lower readmission rates yield the good scores.

3 Outside Sources

1. In order to scrape the readmission rate data, which was dynamically generated on a paginated page, we utilized `phantomJS`, which can be downloaded from `http://phantomjs.org/`.
2. We also used `d3`, `d3 geo projection`, and `d3 topojson` to help visualize our data. The corresponding code can be found at `https://github.com/mbostock/d3`, `http://d3js.org/d3.geo.projection.v0.min.js`, and `http://d3js.org/topojson.v1.min.js`, respectively. We included the cloned code for `d3` directly in our submission.
3. We scraped readmission rates for pneumonia and for heart attack from `http://www.hospital-data.com/`.
4. We scraped the median salaries of hospital doctors across the nation from `http://www1.salary.com/Physician-Hospitalist-salary.html`.
5. We wrote a formatting script to convert a .tsv file to a .json file. The .tsv file contains the mapping of state IDs to state names used by `topojson` from `https://gist.github.com/mbostock/4090846#file-us-state-names-tsv`.
6. We borrowed ideas from lecture on some of the visual elements, such as how to create the U.S. map and how to create the bar graphs.

4 Other Notes

We originally wanted to analyze the relationship between doctor salaries and readmission rates on the county level because we had access to this fine-grained data. However, in order to scrape the data, we needed to make many, many HTTP requests. We could not request faster than one request per five seconds because the site would block us from making requests for an hour. We attempted to request at a rate of one request per five seconds, but we realized that given the deadline we would not be able to scrape all counties across the nation. Sadly, we had scraped a significant amount of data (days worth of scraping) that we mostly ended up not using. To show all of our hard work scraping data, we included the files in the directory `readmission_info_folder`.