Project Milestone 2

Yuning Wu, Justin Yang, David Young

3/13/2022

Project Description

Through this visualization project, we have come up with a set of seven questions for which we identified potential points of interests and areas of use for our targeted audience. These are:

1. State vs salary: which states have the highest paying jobs

2. Industry vs salary: which industries (media, technology, finance for example) have the highest paying jobs

3. Sector vs salary: which sectors (healthcare, government, education) have the highest paying jobs

4. Job title vs salary: which positions earn the most money

5. Company rating vs average salary by company: do higher rated companies on glassdoor pay their workers more

6. Cost of living vs salary: what is the most optimal job in terms of net earnings

7. Comparisons with other similar jobs

Data

The data we used was scrapped from glassdoor, a review aggregation website of companies. The website allows users to anonymously rate companies, submit salaries and apply for jobs, all important for this project. (Source: wikipedia) The initial kaggle dataset is pre-cleaned and filtered, already containing information about salary, job descriptions and company ratings. After filtering out unwanted variables, removing those with missing data and mutating our columns, our newly processed data contains the Company Name, the

city and state the job is located in, Rating, Industry, Sector, Job Title and two salary related columns: one with Average reported salary ranges and one with the calculated Average Salary based on the Salary Estimate. (the latter is still to be implemented)

Additionally, we appended a column containing the cost of living index, obtaining the numbers from a secondary calculated dataset from AdvisorSmith. While appending, a major hurdle occurs as AdvisorSmith only accounts for the major cities, not the suburbs found within their respective metropolitan areas. This is most prominently seen for jobs located in the San Francisco Bay Area, as we had to discard jobs located in cities such as Palo Alto, Foster City and Santa Clara. Ultimately, this resulted in around 44% of the initial downloaded dataset being discarded, something that needs to be rectified in the near future.

Sketches of data visualization

Below are the 3 sketches for the App. The main factors that we are looking to visualize are salary and cost of living.

Data Analyst/Scientist Jobs

Slider
salary range

Check box Input
State location

Select Input
Company

- Scatter plot company salary to rating
- Points labeled by company

Geo visualization (static)
- US Map w/ color of state describing
  - > Cost of living

Data Table of company competitors

Uses company as Input

Data table
- Will consist of jobs matching reactive input

Data Analyst/Scientist Jobs

Select Input
Sector/Industry

Slider Input
Salary range

Slider Input
company size

Scatter plot
- Cost of living vs. salary
- colored by state

Geo visualization (static)
- US Map w/ color of state describing
  - > Cost of living
  - > availability
  - > Avg Company rating

Data table
- Will consist of jobs matching reactive input