Group 19 Project Milestone 4

Yuning Wu, Justin Yang, David Young

5/8/2022

Introduction

According to the US Department of Labor Statistics, as of May 2021, the current average salary of a data scientist is $108,860, indicating a monetarily successful profession. In order to understand, a set of questions were developed, pertaining to which states have the highest paying jobs, what is the most optimal job in terms of net earnings and which industries and/or sectors have the highest paying jobs. The aim of this project is to visualize the state of this profession, providing aid in the decision making process for students thinking about becoming data scientists and the current data scientists searching for the optimal position.

Data

The data we used was scrapped from Glassdoor, a review aggregation website of companies. The initial kaggle data set is pre-cleaned and filtered, containing information about salary, job descriptions and company ratings. After filtering out unwanted variables and removing missing data values, our newly processed data contains the Company Name, the city and state the job is located in, Rating, Industry, Job Title and the Average Salary for that position.

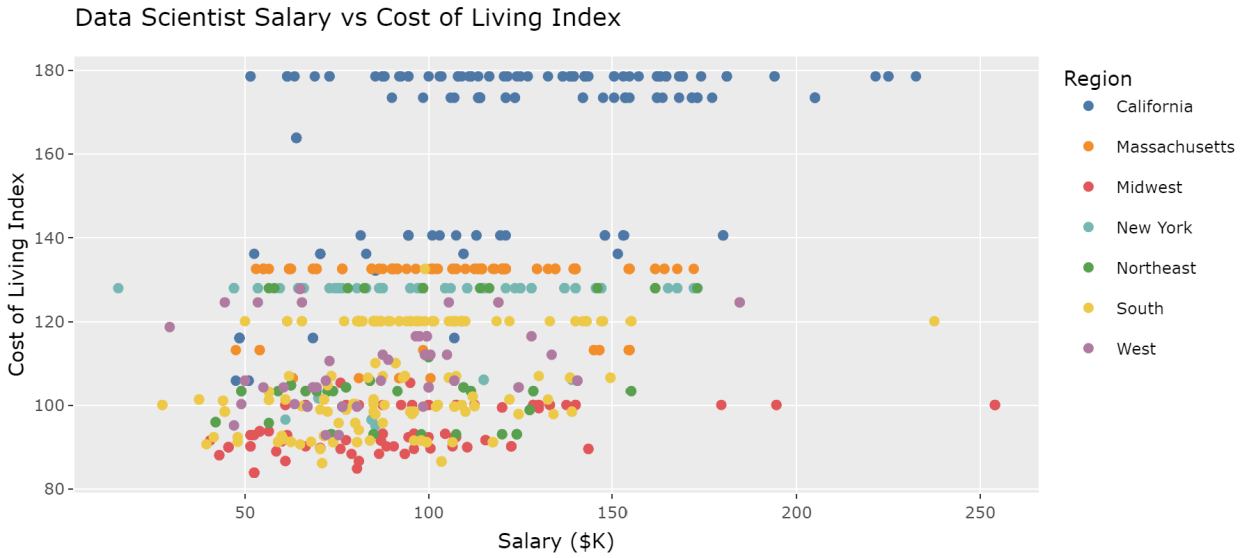
Additionally, we appended a column containing the cost of living index, obtaining the numbers from a secondary calculated data set from AdvisorSmith. The cost of living index takes the cost of living of a place and normalizes the number across the entire nation. Similar to how OPS+ works in baseball, AdvisorSmith adjusts the number on a scale where 100 represents the national average, meaning a cost of living index of 125 means that a place is 25% more expensive than an average place. As a result, this creates a standardized statistic that can be easily grasped.

While appending to our kaggle data frame, a major hurdle occurs as AdvisorSmith only accounts for the

major cities, not their suburbs. To solve this problem, we manually added a column in our salary and cost of living data sets of the metropolitan area of each unique city, allowing us to join the two.

Once we obtained our overall cleaned data set, each visualization did their own set of transformations. The first plot sorted all states by region using the rnaturalearth package. The second required grouping the data by industry and calculating the salary values based on it. For the final visualization, the process was similar through grouping the values by state.

Visualization 1: The optimal data science job



Appendix

Graph axes:

Salary in thousands of dollars

Cost of Living Index: The cost of living normalized across the entire nation, where 100 is the national average. For example, as COL index of 120 means that place is 20% more expensive than average. It is similar to how ERA+ and OPS+ work in Baseball.

Regions:

Midwest region: Iowa, Illinois, Indiana, Kansas, Michign, Minnesota, Missouri, North Dakota, Ohio, South Dakota and Wisconsin

Northeast region: Connecticut, Maine, New Hampshire, New Jersey, Pennsylvania, Rhode Island and Vermont

South region: Alabama, Arkansas, Deleware, Florida, Georgia, Kenntucky, Louisianna, Maryland, Mississippi, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia and the District of Columbia

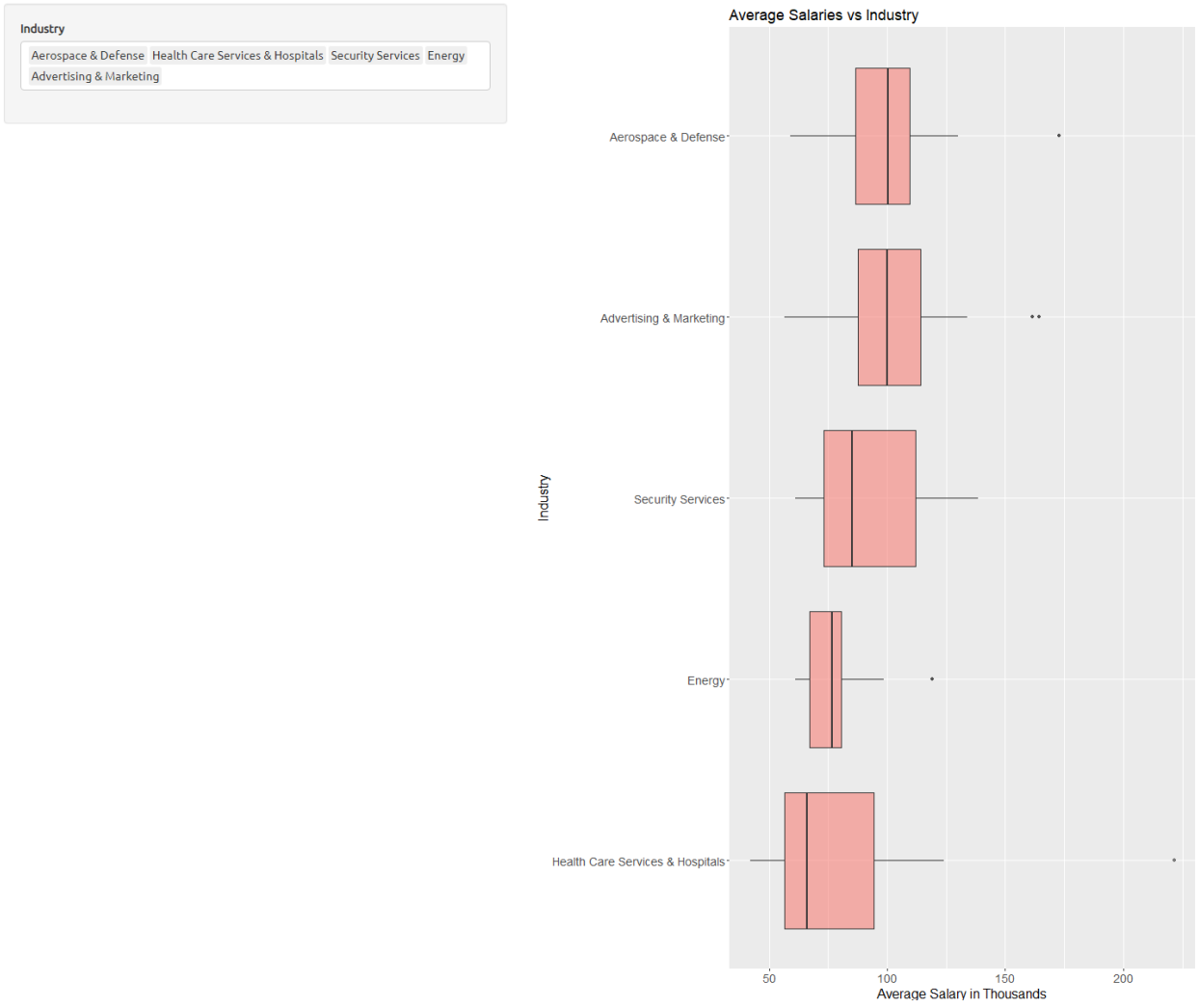
West region: Alaska, Arizona, Colorado, Hawaii, Idaho, Montana, New Mexico, Nevada, Oregon, Utah, Washington and Wyoming

The first visualization is a scatter plot that seeks to find the most optimal job in terms of salary and cost of living. It graphs the average reported salary for that specific job on the x-axis and the cost of living index on the y-axis, creating a dynamic where the ideal values are located towards the bottom right corner. Each point is colored by the region the job is based in, with the states of California, Massachusetts and New

York defined as their own region. A tooltip was added that brings up information to add more interactivity regarding the specific point upon hover. Lastly, an appendix was added below the graph, defining key terms for which the audience may not intuitively know.

From the graph, a key observation is that there is no correlation between the two variables which is slightly counter intuitive as one would expect the salary to rise in conjunction with living costs. Additionally, all the jobs based in the San Francisco Bay Area are well separated from the rest of the points at the top of the plot. Lastly, the most bottom right points are located in the Midwest, indicating that the optimal job lies within the region.

Visualization 2: What industries host the highest-paying data science jobs

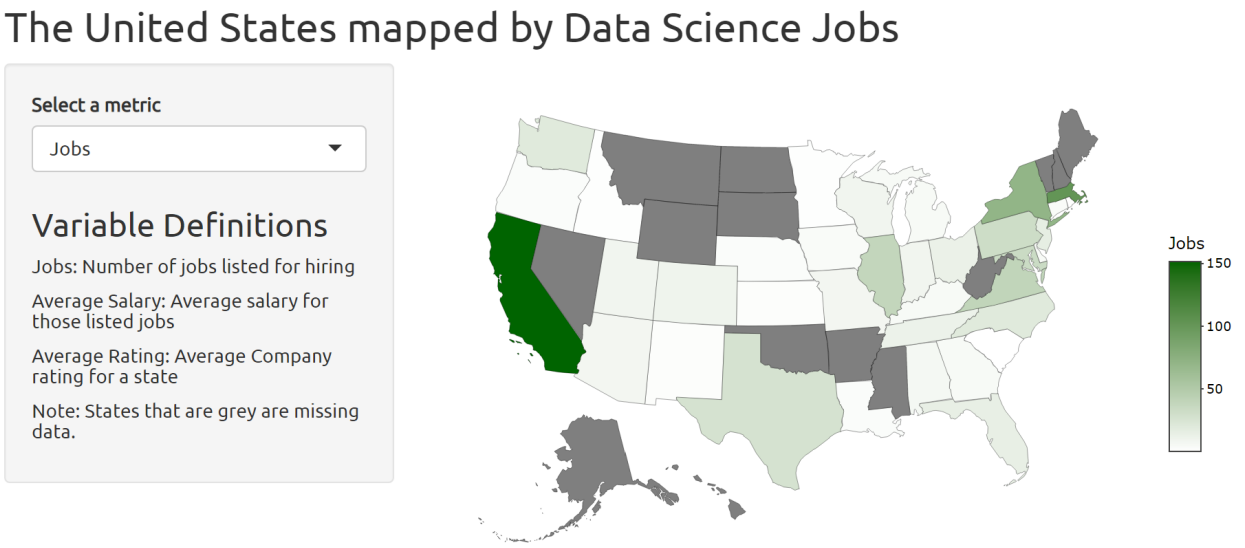


The second visualization seeks to answer the question of what industries provide the highest-paying data science jobs through a series of box plots grouped by industry. Users can select the industries in the data

set from an input box, of which the resulting box plots are ordered by descending median values.

The main finding is that data science positions in financial analytics, telecommunication services and brokerage are among the highest paying ones in terms of median salary. Contrarily, positions in food and beverage manufacturing, architectural and engineering services and social assistance industries are the lowest paying.

Visualization 3: Job Availability, Average Rating and Average Salary



This third visualization is a map of the United States that uses a continuous scale to display the average job rating, job availability and average salary. The user can select which continuous variable they want visualized, with the color of the state directly correlating to each of the metrics. When hovering the cursor of each state, the specific value for that selected metric will appear.

An important observation from the graph is that a lot of states have either none or very light color, indicating an uneven distribution of jobs to a few specific states. When selecting the other variables, pay and company rating seem to be even across the nation, emphasizing the importance of the cost of living due to similar pay from state-to-state.

Conclusion

From the three interactive plots, we were able to analyze the salary of data scientists across the United States. We conclude that working in the Midwest with a position in financial analytics, telecommunication services or brokerage is the best way to optimize personal finances when only considering cost of living, average salary and industry of employment. Oppositely, we conclude that while a majority of jobs are located in California,

living in the state with a position in social assistance or beverage manufacturing would leave someone with a minimal amount of money.

However, a shortcoming is that our data set only collected data from one job board from the year 2021. Ideally, we would utilize data from all job boards in all years including the most recent one. Nonetheless, we believe that it gives an accurate snapshot into the state of the profession and can only hope it provides the framework for those on the job hunt on what to target for their next position.

References

AdvisorSmith Solutions, Inc. (2020, June 5). U.S. Cost of Living Index by City: Downloadable Data – AdvisorSmith [Dataset]. https://advisorsmith.com/data/coli/

Data scientist salary. (2021, December 29). [Dataset]. https://www.kaggle.com/nikhilbhathi/data-scientist-salary-us-glassdoor

Data Scientists and Mathematical Science Occupations, All Other. (2021, September 28). [Dataset]. https://www.bls.gov/oes/current/oes152098.htm

"Employees Rate Their Employers, CEOS on Glassdoor | CBC News." CBCnews, CBC/Radio Canada, 29 Mar. 2013, https://www.cbc.ca/news/business/employees-rate-their-employers-ceos-on-glassdoor-1.1314945