Data Science Project

Feiyang Zheng

Introduction

Data scientists are currently in high demand. Based on the estimation of McKinsey, "US will be facing a shortage of 140000 to 190000 data scientists by 2018"(1). This is because of the wide use of data. The Economist magazine even claims that "The World's most valuable resource is no longer oil, but data"(2). Because of Internet and smartphone, data, a collection of information, become so abundant in our world. No matter when we are watching TV, accepting health service or even driving a car, almost all the activities that we have can be collected by the electronic device and represented in digital format. Therefore, Data scientists, someone who collects interpret, and transfer complex data in a more meaningful manner after complex analysis, play a huge role in helping the companies make crucial decisions.

The abundance of data benefits many types of industry. For example, to improve the efficiency of full order, Amazon links with manufacturers. By tracking their inventory, Amazon would be able to select warehouse closest to the vendor and/or customer, which reduces shipping cost by 10% to 40% (3). Customers also take advantages of the harness of data. By collecting more data, companies have a better scope about how to improve their products or service to attract more customers, which results in better products at a lower price.

In order to help candidates who are interested in data scientist position better prepare for the position, in this project, we perform an analysis of “data scientist” jobs listed on job boards and on the employment pages of major companies. We aim to find the most common skills that employers look for and the types of companies that employ the most data scientists. Besides, we also build a linear regression model to observe the difference in salaries between distinct regions and diverse industries.

Method

Data source

The job search engine that we select is Glassdoor, which is a widely used recruitment tool for job seekers to search potential employees. Glassdoor provides fresh data, which means that the job positions are always updated to the latest date. Another reason to select Glassdoor is that it not only contains detailed information about a position but also provides relevant information about the corresponding company such as salary information, company reviews, and company type. Last, the URLs of different pages in Glassdoor have a common pattern. To be more specific, after we modify the first page, the list of URLs could bring us to all the position pages.

Web Scraping

Web Scraping is the first and one of the most important steps in this project for the purpose of gathering data. The main tool here is SelectorGadget, an open source Chrome extension that makes CSS selector generation and discovery on complicated sites. The basic package used in R is rvest, which extracts attributes, text and tag name from HTML.

1. Search "Data Scientist" position on Glassdoor website and obtained the URLs from each page.
2. Select each position using SelectorGadget and collect "XPath". Within each position, extract pieces out of HTML documents that we are interested in. "html_node" identifies nodes on the web page and return the HTML element. "html_text" presents the information in the text format. We also use this procedure to extract a list of links to positions from URLs by "html_attr".
3. Create a list of 13 most popular skills of data scientist. These technique skills include Python, R, SAS, SQL, Java, Tableau, Spark, C, Perl, Excel, Hadoop, NoSQL and HBase.
4. Use "readline" and the same procedure in step 3 looping through the list of the linking URLs, which returns location, company’s name as well as industry type. For skills, "grepl" is used. If the "job description" section contains the skill, it outputs true, otherwise, it outputs false. The error handling function is "tryCatch". It returns NA when functions generate warnings or errors and skip to the next iteration.

Data

In our dataset, the positions' features we collect are company’s name, location, industry, rating, maximum and minimum salary. There are 13 technique skills including Python, R, SAS, SQL, Java, Tableau, Spark, C, Perl, Excel, Hadoop, NoSQL and HBase. I also divide salary into maximum and minimum salaries and calculate the mean.

After extracting positions from first 35 pages, we obtain 990 positions. Because some companies post the same position multiple times, we remove 173 duplicated jobs. For the 817 unique positions, 102 of these miss company’ names while 78 positions miss maximum or minimum salary information. 41 positions do not contain rating of its corresponding company and 79 positions miss industry type. In terms of location, I also regard location labeled "remote" and "United States" as missing value besides NA (21). Moreover, there are 13 jobs lacking any of the 13 skills.

Overall, I exclude jobs that are absent of information of location, salary, industry and skills and end up with 639 non-missing positions.

Results

Skills required by employees

Figure 1 is the bar plot in terms of the most popular skills of data scientist gathered from Glassdoor in the US. The top five most common skills are Python (65.6%), R (55.4%), SQL (43.2%) followed by Hadoop (26.8%) and Spark (24.6%). However, only 2.7% companies require a candidate have knowledge of HBase. As we can observe from the plot, the number of positions that require Python, R and SQL are substantially greater than the rest of the skills.

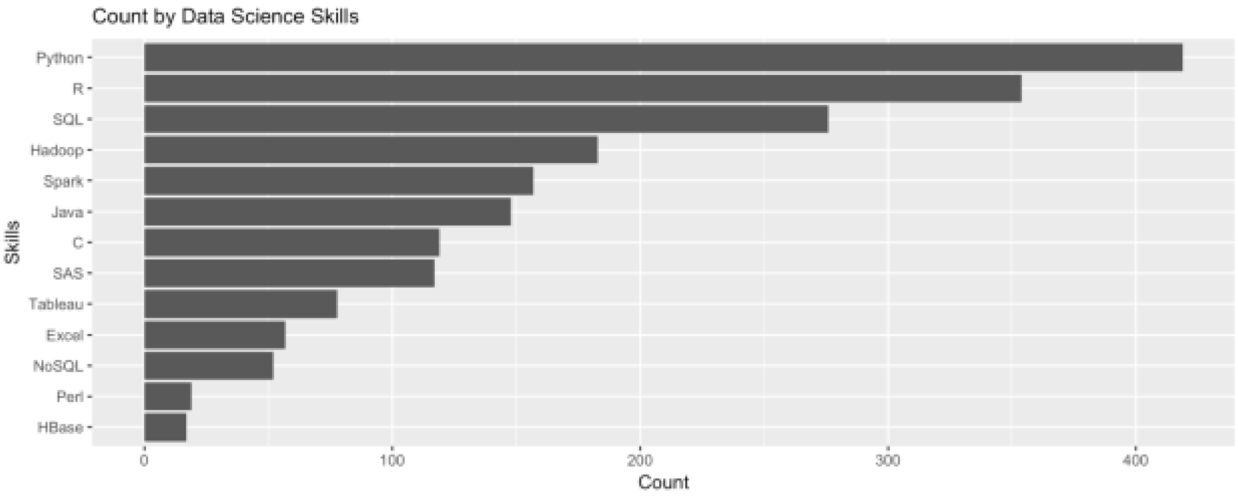

Figure 1 The bar plot of the count of each 13 skills of data scientist position from Glassdoor.

Geographical distribution

We also geographically visualize the distribution of data scientist positions by ggmap, which is displayed in figure 2. Here, we use orange spots to indicate positions and the brighter the sport, the more positions are located in that area. Based on figure2A, we can roughly observe that west coast and northeast area especially San Francisco and New York city are most likely to hire data scientists.

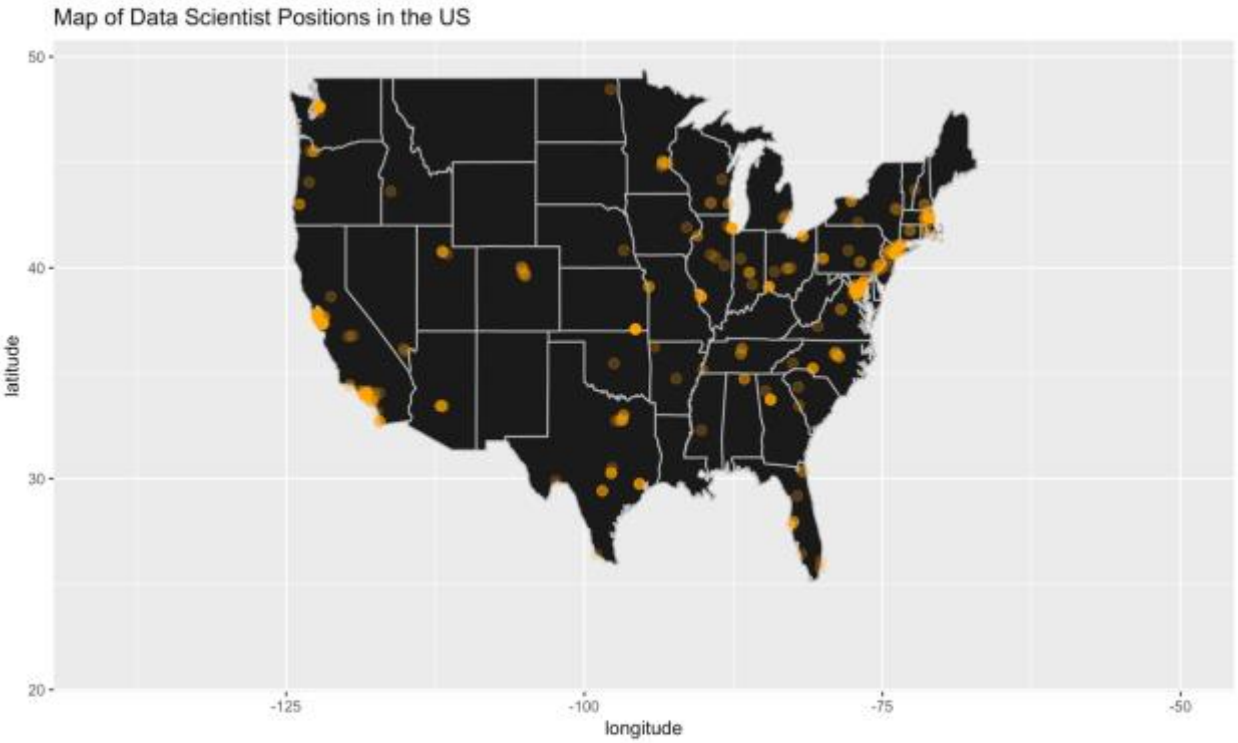

Figure 2 The distribution of data scientist positions in the US displayed on the map.

Industries that employ data scientist

Our dataset consists of 25 industry types in a consistent manner (See appendix). Figure 3 demonstrates a summary of top ten industries that hire most data scientist. Obviously, most of the positions are posted by information technology firms (37.7%) followed by business service companies (14.2%). Finance company arrives in the third place (9.5%) and for the rest of the industries, the number of positions is only around 1/10 of it provided by IT corporations.

I also narrow down the geographic region by classifying the mainland US into four separate areas, which are west, midwest, south, and northeast (See appendix for more information). For each region, I list top 5 industries which employ most data scientists. Similarly, information technology is still the industry that recruits most data scientists. However, in the midwest, the manufacturing industry is also in the first place with the same frequency of IT industry. In the west, more retail and healthcare companies need data scientist while in south, the aerospace industry is in the top 5. Interestingly, finance falls out of top five only in the west region.

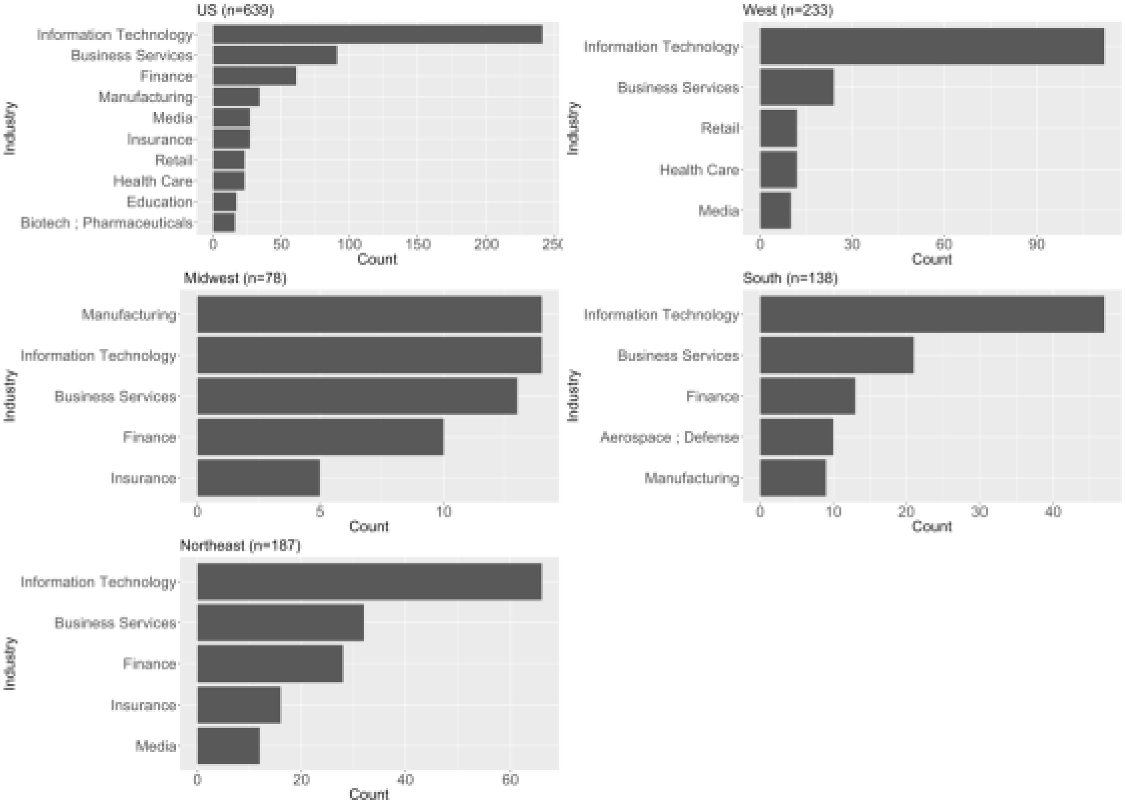

Figure3 Top 10 industries that employee most data scientists in the US and in separate regions

Model

For the purpose of understanding how industry and region affect the salary of a position, a linear regression model is fitted. According to the results obtained from the ranking of industry, I select IT, business service and finance as my primary interest and categorize all the rest as "others". After model selection using stepwise AIC method, my final model is

$$Salary = \beta_0 + \beta_1 Industry + \beta_2 Region + \beta_3 Python + \beta_4 Spark$$

Thus, the salary of data scientists in west region in an information technology firm without the requirement of python or spark would be $120k per year on average (95% CI [114,125], p=0). When we hold all other variables as constant, finance firms offer $11.1k more to data scientist comparing with IT companies on average (95% CI [3.5, 18.7], p=0.005). As for location, average annual salaries in midwest and south are $27.3 (95% CI [-34, -20], p=0) and $23.3(95% CI [-28, -17], p=0) lower than it in west respectively holding all other variables as constant. Additionally, we would expect an average of 12.4k dollars more per year for corporations that are located in the west than these in the northeast (95% CI [-19, -7.2], p=0).

Discussion

In conclusion, based on our research, Python, R, SQL, Hadoop, and spark are the top five most commonly required skills to data scientist in the job market while Perl and HBase become the least important skills. Positions are basically concentrated in west coast and northeast especially in several typical cities. As for industry, unsurprisingly, information technology firms offer most opportunities followed by business service and finance in the US. However, as we narrowed down to regions, we found that not all regions are the identical due to regional characteristics. We also fitted a model to analyze how the region, as well as industry, affect data scientist salary.

Limitations do exist in this project. First of all, Glassdoor is the only job search engine that we used to collect data. Since employees pay the job hunting website to list jobs that they are seeking to fill, the unique source can cause selection bias. Section, as for skills, I only list 13 widely used programming languages, some other popular skills such as machine learning, data mining and communication are not considered. Last, for my model, features like company size, company type, revenue and candidate's degree are also correlated to salary.

Need citation and appendix