Prepare for a Career in Data Scientist

*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 broadly 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 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 being a data scientist better prepare for the position, in this project, we performed an analysis of “data scientist” jobs listed on job boards and on the employment pages of major companies. We aimed 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 Collection

The job search engine that we selected 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 following pages.

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. The basic package used in R is rvest. First of all, we searched for "data scientist" position and obtained the URLs from the first 33 pages. Looping through a list of pages, we selected our target element and acquired the minimal CSS selector generated by SelectorGadget. Next, after specifying the CSS selector or XPath, "html\_node" and "html\_text()" easily extracted and read the text and tag name from HTML, which returned the corresponding information. Last, "html\_attrs()" was used to extract the attributes, which returned a list of link's URLs. We chose totally 13 common skills. For each skill, if it was required by the position, it outputted true, otherwise, it outputted false. Furthermore, for company's industry type, because the above procedure was not appropriate to be used, we employed "readLines","str\_detect" and "str\_extract" to read, detected and extracted our target information. The error handling function was "tryCatch", which returned NA when functions generate warnings or errors and then skipped to the next iteration.

Data Management

In our dataset, the positions' features collected 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. We also divided salary into maximum and minimum salaries and calculated the mean. Furthermore, we classified mainland US into four regions based on the location, which were west, midwest, south, and northeast (See appendix for more information)

After extracting positions from first 35 pages, the raw dataset consists of 990 positions. Because some companies post the same position multiple times, we removed 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 the rating of its corresponding company and 79 positions miss industry type. In terms of location, we 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. Besides, we also acquired required skills of 448 unique positions title “marketing analyst” in order to make comparison.

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

**Results**

1.Skills required by employees

Figure 1A is the bar plot in terms of the most common skills of data scientist in the US gathered from Glassdoor. 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%). It is also obvious that the number of positions that require Python, R and SQL are substantially greater than it of the rest skills Comparing with data scientist, for marketing analyst in figure 1B, Excel arrives at the first place followed by SQL and Tableau. Moreover, for Hadoop, Spark and Java (23.2%) which are highly important for data scientists, only 1.34%, 0.67% and 1.12% positions require for marketing analyst respectively. Therefore, the result suggests that Hadoop, Spark, and Java to be unique for data scientist.

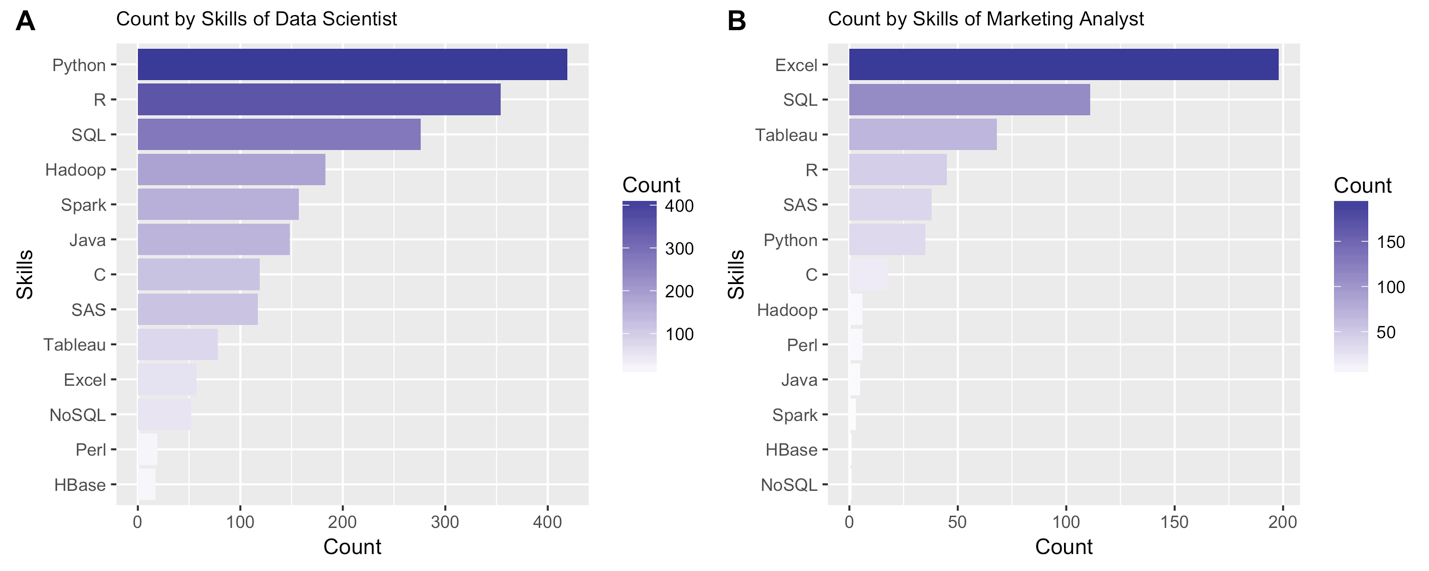
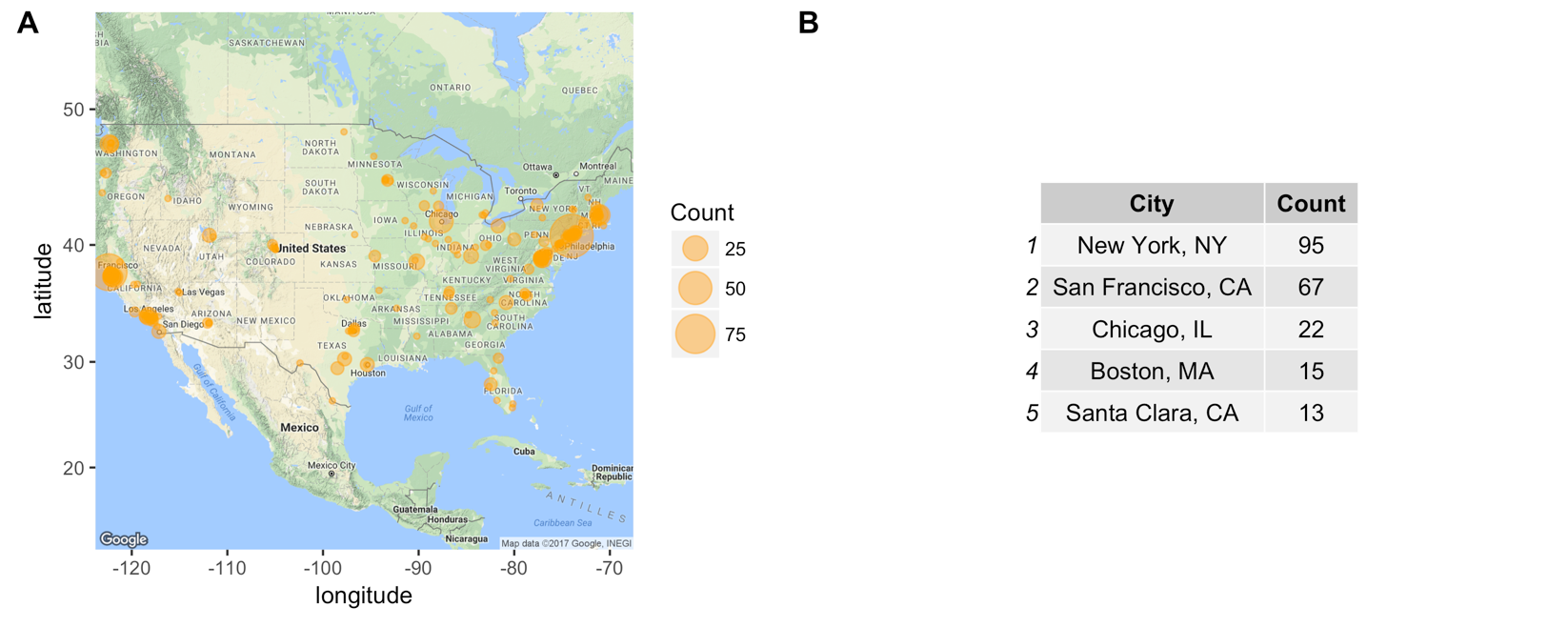


Figure 1 Bar plot of occurrences of the 13 skills of data scientist positions (A). Bar plot of occurrences of the 13 skills of statistical analyst positions (B).

2. Geographical distribution

We also geographically visualized the distribution of data scientist positions by ggmap, which is displayed in figure 2A. Here, we use orange bubbles to indicate positions and the bubble size to represent the number of the positions in that area. We can roughly observe that opportunities concentrate on west coast and northeast area. Specifically, figure 2B points out that New York, San Francisco, Chicago, Boston and Santa Clara have the most open data scientist positions.

Figure 2 Geographical visualization of distribution of data scientist positions in the US (A). Top 5 cities that employ most data scientists (B).

3. 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 scientists. 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%). It is also Noteworthy that for the rest of the industries, the number of positions provides by each individual type is only around 1/10 of it provided by IT corporations.

For each region, I listed 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 scientists while in the south, the aerospace industry rank at the top 5. Interestingly, finance falls out of top 5 only in the west region.

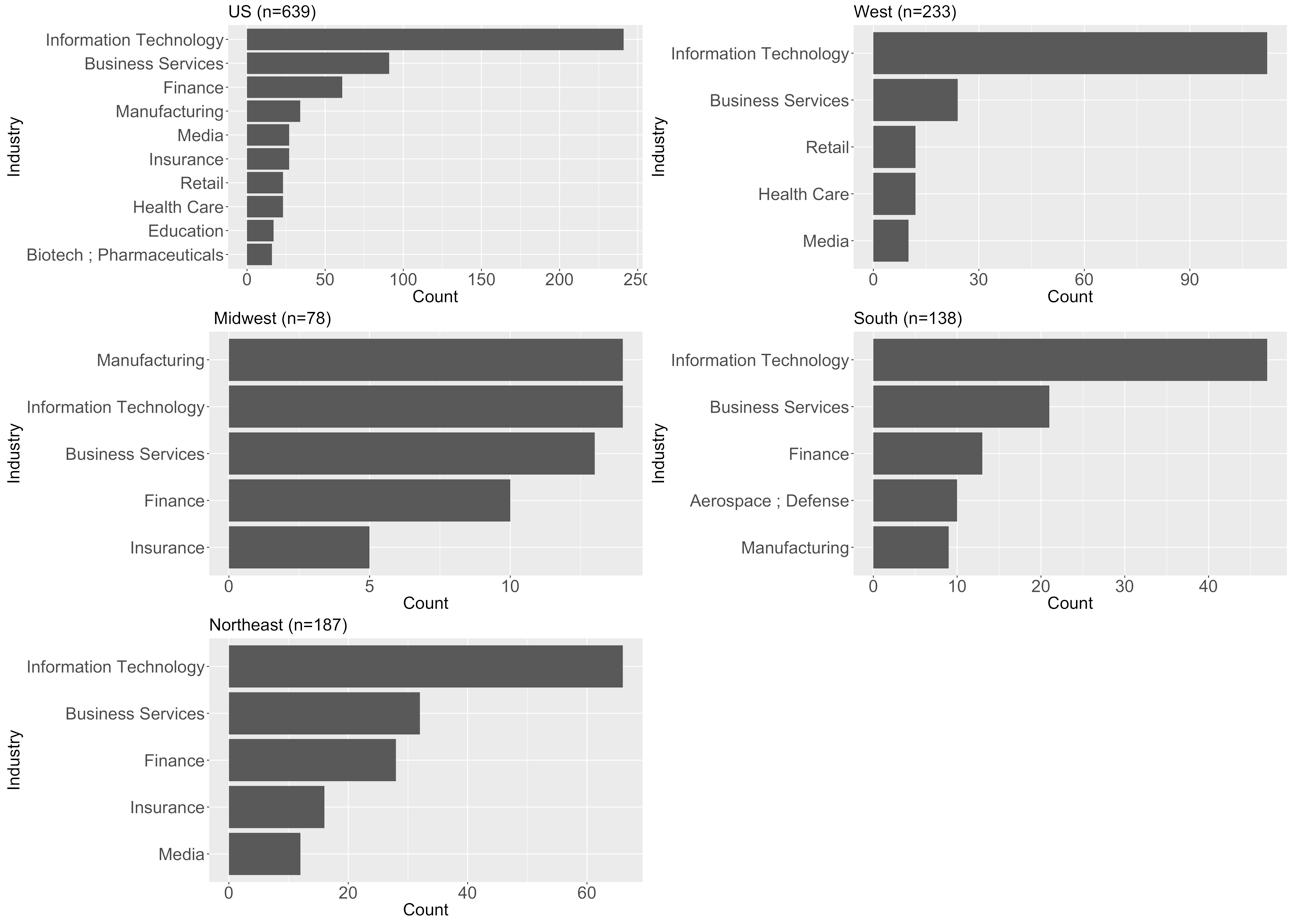


Figure3 Ranks of industries that employee most data scientists in the US and in the individual regions.

4. Factors affecting salaries

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

+ ε

Thus, a data scientists in west region in an information technology firm without the requirement of python or spark would expect to receive $120,000 per year on average (95% CI [114,125], p=0). When we hold all other variables as constant, finance firms offer $11,100 annually 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 $27,300 (95% CI [-34, -20], p=0) and $23,300(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 12,400 dollars more annually from 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 identical due to regional characteristics. We also fitted a model to analyze how the region, as well as industry, impact data scientist salaries.

Limitations did exist in this project. First of all, Glassdoor was 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 could cause selection bias. Second, as for skills, I only listed 13 widely used programming languages, some other popular skills such as machine learning, data mining and communication were not considered. Last, for my model, features like company size, company type, revenue and candidate’s degree were also correlated to salary but were excluded in this research.