# Analysis of How to 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 extensive 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 what we do, watching TV, accepting health service or even driving a car, almost all the activities that we perform can be collected by electronic devices and represented in digital format. Therefore, Data scientists, someone who collects interpret, and transfers complex data into complex data in a more meaningful manner after complex analysis, play a significant 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 and unique skills that employers look for and the types of companies that employ the most data scientists.  Finally, we also build a linear regression model to observe the difference in salaries among people with different skills, distinct regions and diverse industries in the U.S.

### Method

#### Data Collection

The job search engine that we selected is Glassdoor[6], a popular 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 sufficient 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 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[7] and the basic package used in R is rvest[8]. We searched for "data scientist" position posted before October 16, 2017, and created a list of urls of the first 33 pages. Because instead of only showing jobs that matched exactly the word we

searched, Glassdoor also displayed jobs with high similarity such as “lead data analyst”. Therefore, fuzzy matching was allowed in this research. By specifying the CSS selector or XPath of our target elements, we extracted the linking urls, rating, and salaries of each position from the main page (see appendix table 1 for specific CSS selector). Reading the job links in sequence, attributes such as company’s name and location were extracted from title and subtitle sections. Next, we detected and extracted for exact matches to the text string “sector” and 13 skills for data scientist from the job description and qualification section respectively (see supplementary code step1[9,10,23]). The 13 skills were primarily identified by internet research[4,5]. Same procedure was applied to collect company’s name, location, and skills for positions titled “data engineer” from the first 20 pages before October 22, 2017, on Glassdoor.

In our dataset, position' features collected are company’s name, location, industry, rating, maximum and minimum salary. There are 13 common 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.

After extracting positions from the first 33 pages, the raw dataset consists of 990 positions. Because some companies post the same position multiple times, we removed 173 duplicated jobs with the same location, industry rating, salary, and skills. Overall, we excluded jobs that are absent of the information including location, salary, industry, rating, and skills and ended up with 639 non-missing positions with 25 industry types (see supplementary code step2[11,12,13,14]). Besides, we also acquired 497 unique positions titled “data engineer” in order to make a comparison.

## Exploratory Analysis

An exploratory analysis was first conducted to summarize the most common skills required for data scientists. Bar plot was generated to visualize the occurrence of each skill for data scientist and data engineer. For data engineer, if the occurrence of a skill was observed to have a substantial decrease compared with it for data scientist, we performed a one sided two-proportion test testing for the difference in proportion between the two groups in order to access the unique skill for data scientist (see supplementary code step3 section1[15,16,17,18,19]). Additionally, we collected the longitude and latitude of position’ location and geographically visualized the distribution of work opportunities in a US map (see supplementary code step3 section2 &3[20,21]). We also ranked and tabulated industries that hire most data scientists.

## Model

For the purpose of understanding how features collected impact the salary of a position, a multiple linear regression model was fitted (see supplementary code step4). We classified locations into four regions, which are West, Midwest, South and Northeast. As for industry, we selected information technology, business service and finance as our primary interest and categorized all the rest as “others”. Specifically, information technology is treated as the reference group. For region, the West is treated as reference group. After model selection using stepwise AIC method, our final model with lowest AIC is

$$Annual\ Average\ Salary = \beta 0 + \beta 1 Business\ Service + \beta 2 Finance + \beta 3 Other\ Industry + \beta 4\text{Midwest} + \beta 5\text{South} + \beta 6\text{Northeast} + \beta 7\text{Python} + \beta 8\text{Spark} + \varepsilon$$

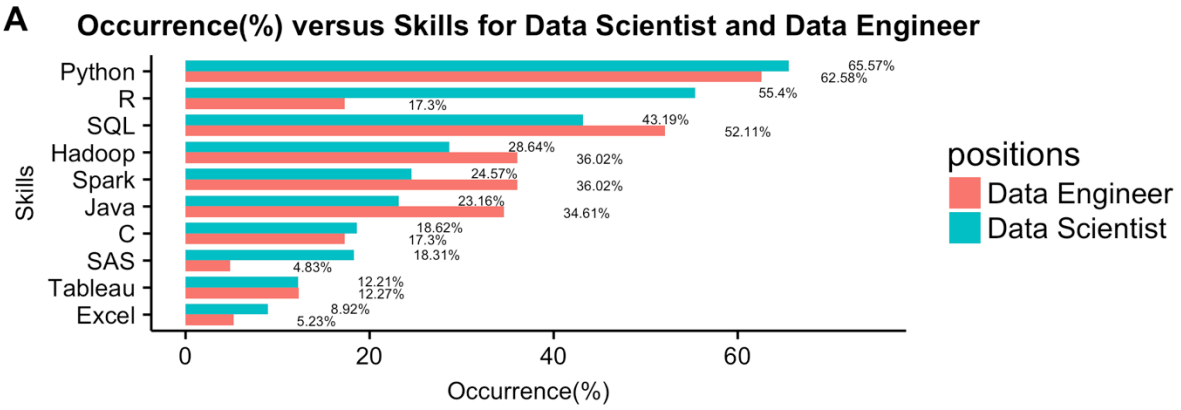
# Results

## Exploratory Analysis

Figure 1A is the bar plot in terms of the top 10 most common skills for data scientist and the corresponding count for data engineer. 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%). Compared with data scientist, for data engineer, Python (62.6%) is still in the first place followed by SQL (52.1%), Hadoop (36.0%), Spark (36.0%) and Java (34.61%) (see appendix table 2).

Although the top 5 skills show high similarity between the two groups, compared with data scientist, R and SAS show substantial decrease in occurrence for data engineer. For data scientist, of 639 positions posted, 117 positions require SAS (18.3%). However, only 24 out of 497 jobs use SAS (4.83%) tag for data engineer. In addition, R is only in the 6$^{th}$ place with count 86 (17.3%) for data engineer. The results of proportion tests are displayed in figure 1B. P values for both tests are much less than 0.05. Therefore, R and SAS are suggested to be more common for data scientist. Therefore, we might consider SAS and R to be the two unique skills for data scientist compared with data engineer.

A




B

| | Skill | p value | Proportion Difference(%) |
|---|---|---|---|
| 1 | R | 2.273529e-39 | 38.10 |
| 2 | SAS | 4.065915e-12 | 13.48 |

Figure 1. (A) Bar plot of top 10 skills with the highest occurrence for data scientist positions and the corresponding occurrence for data engineer. (B) P values of two-proportion tests to compare the proportion of R and SAS in the two groups.

Figure 2A summarizes top 5 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 arrives in the third place (9.5%) (see appendix table 3).

We could geographically visualize the distribution of data scientist positions in a map shown in figure 2B. Here, orange bubbles were used to indicate positions and the bubble size represents the number of positions in that area. We could roughly observe that opportunities concentrate on west coast and northeast area. Specifically, New York, San Francisco, Chicago, Boston and Santa Clara have the most open data scientist positions (see appendix table 4).

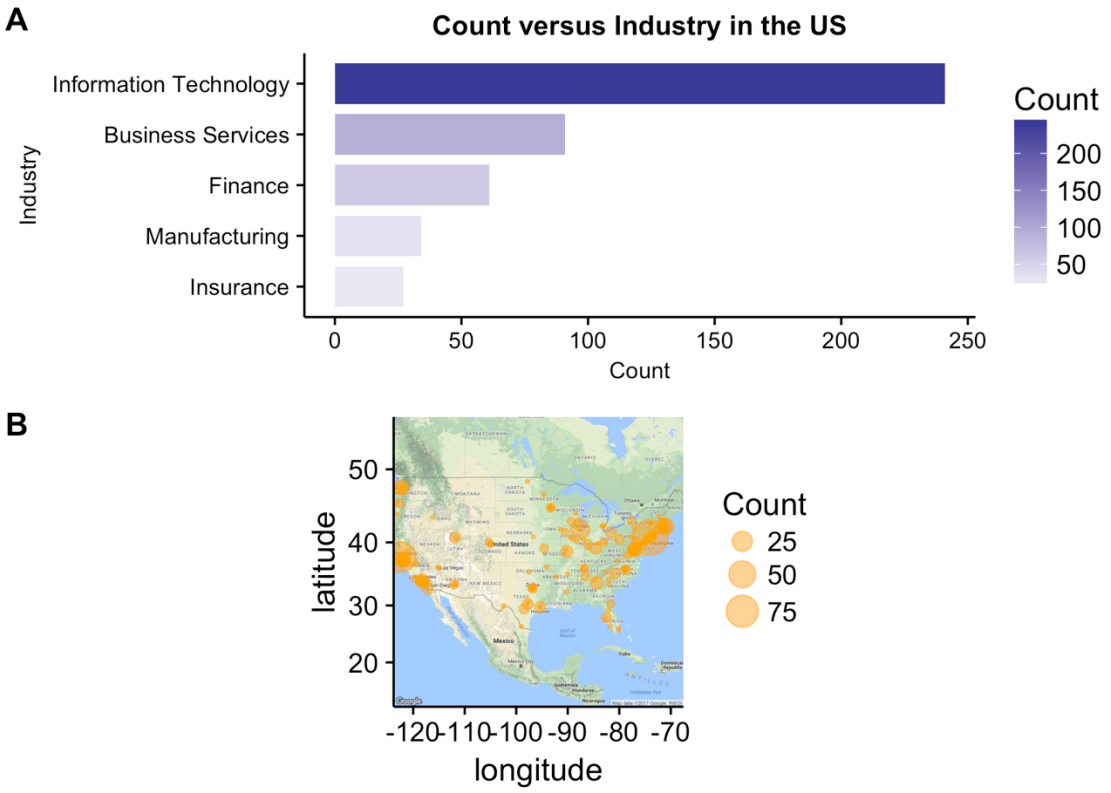



Figure 2. (A) Top 5 industries that employ most data scientists. (B) Geographical visualization of the distribution of data scientist positions in the US.

## Model

Multiple linear regression was employed to investigate how some factors would impact salary. We focus on estimates that show statistical significant here. According to the result of (see appendix table 5), holding all other variable as constant, finance firms offer $11,100 more annually compared with information technology companies on average (95% CI [$3457, $18,800]). In terms of location, people working in Midwest and South would receive $27,300 (95% CI [-$34,270, -$20,300]) and $23,300(95% CI [-$28,900, -$17,600]) lower than those working in West respectively controlling all other variable as constant. Additionally, people

would expect to earn $12,400 more on average in the West than those working in the Northeast (95% CI [$7200, $17,600]). In addition, knowing python or spark is significantly associated with an increase of annual salary. Individuals with python skill would make $7240 more than individuals without python skill (95% CI [$2493, $12,000]). Those who know spark would expect to earn $5200 more compared with people who don't (95% CI [$251, $10,000])

# Discussion

In conclusion, based on our research, Python, R, SQL, Hadoop, and spark are the top five most commonly required skills for data scientists in the job market. Compared with data engineer, SAS and R are considered to be the two unique skills for data scientists. Based on the hire map, positions concentrate in west coast and northeast, especially big cities such as New York and San Francisco. As for industry, unsurprisingly, information technology firms offer most opportunities followed by business service and finance. The linear regression model fitted could guide potential data scientists to search jobs with satisfied salary.

There are several limitations in this research. First of all, Glassdoor was the only job search engine employed to collect data in this project. Since employees pay the job hunting website to list jobs that they are seeking to fill, positions extracted might not be appropriate to represent the demand of the entire job market. Next, we did not search for a perfect match for the key word of positions when we collected data. It is also common for a job title not to be an accurate portrayal of one's actual job responsivities and activities. In this case, even though some required skills of these jobs overlap, noise can still make a difference on the result. Last, there are other features highly related to salary offered by a company such as company size, company type, revenue and candidate's degree. It might be inappropriate to exclude the impact, and might cause low accuracy of the model.

Extended research related to data scientist position could be conducted. For example, we could classify jobs into distinct levels in order to investigate how requirements vary from entry-level positions gradually to senior level positions, which will benefit potential candidates to a greater extent.

Citation

1. McKinsey Global Institute, "Big data, The next frontier for innovation, competition, and productivity." *Digital McKinsey* (2011), access October 20, 2017, https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

2. "The world's most valuable resource is no longer oil, but data." *The economist* (2017), access October 20, 2017, https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource.

3. Agarwal, Nitin. "How Big Data Analytics Can Benefit Supply Chain & Logistics Industry – Part 2." *Motifworks* (2017), access October 20, 2017, https://motifworks.com/2017/02/23/how-big-data-analytics-can-benefit-supply-chain-logistics-industry-part-2/.

4. Mcnulty, Eileen. "Top 10 Data Science Skills, and how to learn them." *Dataconmy* (2014), access October 22, 2017, http://dataconomy.com/2014/12/top-10-data-science-skills-and-how-to-learn-them/.

5. Van Loon, Ronald. "What skills do I need to become a Data Scientist?" *Simplilearn* (2017), access October 22, 2017, https://www.simplilearn.com/what-skills-do-i-need-to-become-a-data-scientist-article.

6. *Glassdoor*, accessed October 16, 2017. https://www.glassdoor.com/index.htm.

7. "SelectorGadget." Google Chrome Extension. https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjick.

8. Hadley Wickham (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest.

9. Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation.package version 0.7.4 https://CRAN.Rproject.org/package=dplyr.

10. Hadley Wickham (2017). httr: Tools for Working with URLs and HTTP. R package version 1.3.1. https://CRAN.R-project.org/package=httr.

11. Hadley Wickham (2017). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.2.0. https://CRAN.R-project.org/package=stringr.

12. Jeroen Ooms (2017). curl: A Modern and Flexible Web Client for R. R package version 2.8.1. https://CRAN.R-project.org/package=curl.

13. Mauricio Zambrano-Bigiarini. (2017) hydroTSM: Time Series Management, Analysis and Interpolation for Hydrological ModellingR package version 0.5-1. URL https://github.com/hzambran/hydroTSM. DOI:10.5281/zenodo.839864.

14. Hadley Wickham (2016). tidyr: Easily Tidy Data with `spread () ` and `gather () ` Functions. R package version 0.6.0. https://CRAN.R-project.org/package=tidyr

15. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

16. Alboukadel Kassambara (2017). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.5. https://CRAN.R-project.org/package=ggpubr.

17. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

18. Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

19. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

20. D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

21. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

22. Yihui Xie (2017). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.17. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963 Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

23. Personal Communication: Stephen Cristiano

24. Personal Communication: Shannon Wongvibulsin

# Appendix

Table1: Target Elements Collected and Their Corresponding html_node

| | Information | html_node |
|---|---|---|
| 1 | field | .jl |
| 2 | linking url | a.jobLink |
| 3 | salary | small |
| 4 | rating | .compactStars |
| 5 | company's name | .padRtSm |
| 6 | location | .subtle |
| 7 | skills | #JobDescContainer |

Table 2. Count and Occurrence of 13 Skills for Data Scientist and Data Engineer

| | Skills | Count(DS) | Count(DE) | Occurrence%(DS) | Occurrence%(DE) |
|---|---|---|---|---|---|
| 1 | Python | 419 | 311 | 65.57 | 62.58 |
| 2 | R | 354 | 86 | 55.40 | 17.30 |
| 3 | SQL | 276 | 259 | 43.19 | 52.11 |
| 4 | Hadoop | 183 | 179 | 28.64 | 36.02 |
| 5 | Spark | 157 | 179 | 24.57 | 36.02 |
| 6 | Java | 148 | 172 | 23.16 | 34.61 |
| 7 | C | 119 | 86 | 18.62 | 17.30 |
| 8 | SAS | 117 | 24 | 18.31 | 4.83 |
| 9 | Tableau | 78 | 61 | 12.21 | 12.27 |
| 10 | Excel | 57 | 26 | 8.92 | 5.23 |
| 11 | NoSQL | 52 | 91 | 8.14 | 18.31 |
| 12 | Perl | 19 | 20 | 2.97 | 4.02 |
| 13 | HBase | 17 | 37 | 2.66 | 7.44 |

Table3. Count and Frequency of Top 5 Industries that Hire Most Data Scientist

| | Industry | Count | Frequency(%) |
|---|---|---|---|
| 1 | Information Technology | 241 | 37.715 |
| 2 | Business Services | 91 | 14.241 |
| 3 | Finance | 61 | 9.546 |
| 4 | Manufacturing | 34 | 5.321 |
| 5 | Insurance | 27 | 4.225 |

Table4. Count of top 5 Cities that hire most data scientists

| | City | Count |
|---|---|---|
| *1* | New York, NY | 95 |
| *2* | San Francisco, CA | 67 |
| *3* | Chicago, IL | 22 |
| *4* | Boston, MA | 15 |
| *5* | Santa Clara, CA | 13 |

Table5. Linear regression coefficient table estimating annual mean salary (thousand USD)

| | Estimate | p value | 2.5 % | 97.5 % |
|---|---|---|---|---|
| *Intercept* | 119.671 | 0 | 114.187 | 125.154 |
| *Finance* | 11.11 | 0.005 | 3.457 | 18.762 |
| *Midwest* | -27.267 | 0 | -34.272 | -20.263 |
| *South* | -23.264 | 0 | -28.928 | -17.6 |
| *Northeast* | -12.404 | 0 | -17.632 | -7.175 |
| *Python* | 7.243 | 0.003 | 2.493 | 11.994 |
| *Spark* | 5.157 | 0.039 | 0.251 | 10.062 |