Skills, Industries and Locations for Data Scientist Jobs

Yue Cao

1 Introduction

As the modern society becomes more driven by digital activities and networks, data has become increasingly important in many industries. Besides software and Internet companies, some traditional industries, such as manufacturing and retail, are more dependent on high-quality data analysis. The demand for data scientists is rising accordingly.

How do we define a data scientist job? In general, Data Science is a field which studies how information is gathered, what it conveys, and how that information can be converted into a valuable resource to generate meaningful insights for the betterment of a business as a whole[1]. Thus data scientists are the people who conduct these kinds of tasks in various industries.

To learn more about the requirement for data scientists, we perform a web-scraping process to extract information about data scientist jobs from three main online job boards, Dice, Glassdoor and Stackoverflow. We use barplots and tables to show that the most popular skills that employers preferred for data scientists are SQL, Python and R. The most unique skill for a data scientist is to address questions about a given business situation, find the root causes of the problems and give proper solutions.[2] The city with the most data scientist job openings is New York, NY. The industry with the highest demand for data scientists is Information Technology. Through a cross-validated hierarchical clustering on industries based on the skills needs for data scientists, when the number of clusters is four, Information Technology and Business Services form two individual clusters; Finance and Education are grouped into the third cluster; Insurance, Health Care, Biotech & Pharmaceuticals and other three industries become the fourth cluster.

2 Methods

2.1 Data Collection

There are many popular online job boards which we can search for data scientists jobs. We first tried Indeed.com and Monster.com. However, these two websites seem to have instable HTML nodes and varying structure of subpage when following the link related to a job title, which causes much trouble when we do web scraping process. We later noticed that Stackoverflow.com, Dice.com and Glassdoor.com have neat and stable web structures for web scraping. Especially, Dice returns more than 38,000 jobs when we search for data science jobs, and Glassdoor contains pre-classified industry information. Thus, we use these three jobs boards as the source of data science job data.

We mainly use `rvest` package by Hadley Wickham[3] to scrape information from online job boards. To know the HTML nodes related to specific parts of a webpage, such as job title, company name or job description, we need to use a Chrome extension, SelectorGadget. It is a web tool which returns the corresponding CSS or XPath when we click a specific part on a webpage. Taking Dice as an example, the main steps are as follows (see supplemental code section 2):

- According to the pattern of the page number in the URL, obtain the URLs for the first 1~40 pages of the search result for data scientist jobs. For Stackoverflow, we only have eight pages of search result, therefore the jobs scaped from Stackoverflow is much less than Dice or Glassdoor.
- Use SelectorGadget to obtain the URLs of subpages corresponding to each job title in the search result page. The URLs are created by pasting "www.dice.com/" and the partial URLs extracted from the attribute "href" of HTML node ".dice-btn-link.loggedInVisited". In each subpage, company name, location, job title are extracted from the HTML nodes ".dice-btn-link span", ".location span", "#jt", respectively. Job decription is obtained from the HTML node "#jobdescSec".
- Create a keywords list: Hadoop, Spark, R, SAS, Stata, Java, Perl, Python, SQL, NoSQL, Tableau, Excel, Machine Learning, Amazon Web Service, which are skills commonly asked by employers looking for data scientists[4]. A keyword search within the job description is then performed on each subpage. If the search result of a keyword is `TRUE` for a job, it means this keyword is mentioned in the job description and we regard it as a required skill for the job.
- Extract industry information by reading the HTML source file of Glassdoor. Since there is no such listed information on Stackoverflow and Dice, in this study, we only use industry information from Glassdoor.
- Create a data frame to combine all the job information together, with variables job title, company name, industry, location and search result (True or False) for each skill keyword. For data frame from Glassdoor, we have an extra variable, industry type. Finally, we obtained data sets with 498 observations from Dice, 470 observations from Glassdoor, and 107 observations from Stackoverflow.

2.2 Data Cleaning

In the data cleaning procedure (see supplemental code section 4), we first remove all the duplicates in the data sets and remove the rows which contain missing values. Reading from HTML source page may produce some HTML language which needs to be transferred into plain English. For example, we substitute the string "&" in the industry fields to "&". In each data set of the three job boards, we summarize the frequency (percent) of each skill by counting the numbers of `TRUE` divided by the total number of jobs. After data cleaning, we obtained data of 101 jobs from Stackoverflow, 498 jobs from Dice, and 444 jobs from Glassdoor.

2.3 Exploratory Analysis and Statistical Modeling

We create barplots and tables to find out the most common skills that employers looking for, the difference of job skill frequency of the three job boards, and the location as well as the industry with the most abundant job openings (see supplemental code section 5).

We randomly divide the data set from Glassdoor into half, and perform a cross-validated hierarchical clustering on industry types, based on the skill needs for data scientist. We employ scaled dataset, Euclidean distance matrix and complete linkage in the hierarchical clustering process (see supplemental code section 6).

2.4 Reproducibility

The analyses performed in this report can be reproduced through the R markdown file "final code.Rmd". To get the same results presented in this report, the data sets "stack_fulldata.csv", "dice_fulldata.csv" and "glass_fulldata.csv" need to be loaded into the above R markdown file.

3 Results

3.1 Skill Frequency for Data Scientists Jobs on Stackoverflow, Dice and Glassdoor

We first add a "source" variable to each data sets of the three jobs boards, indicating the data source website. A barplot (Figure 1) is created to compare skill frequency for data scientists jobs on Stackoverflow, Dice and Glassdoor.

Figure 1. Compare Skill Frequency for Data Scientists Jobs on Dice (N=498), Glassdoor and (N=444) Stackoverflow (N=101)

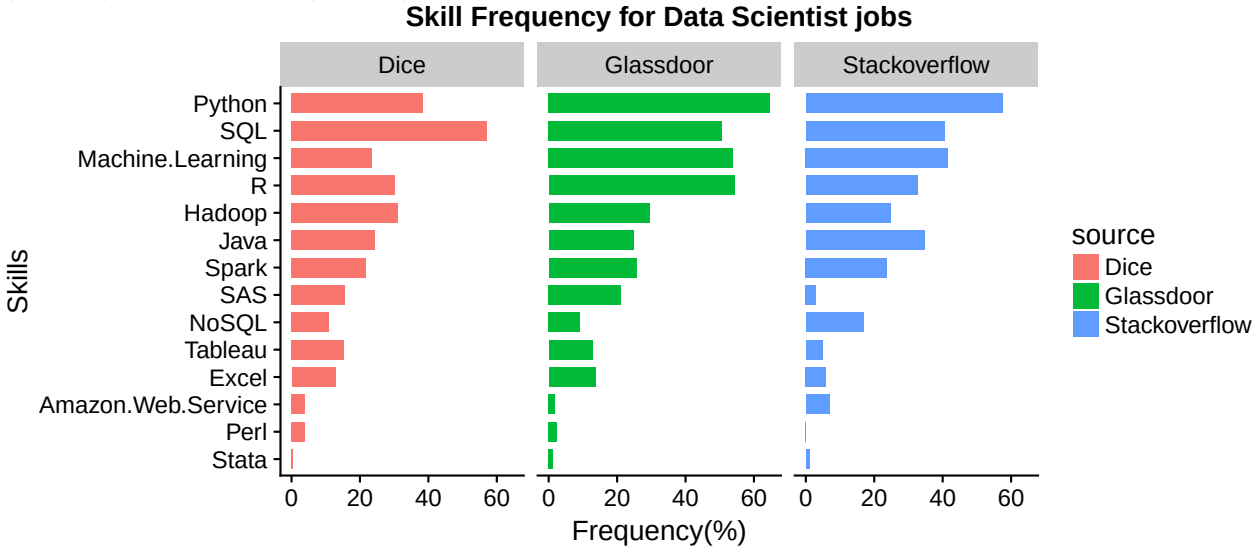


Figure 1 shows that Python is the most popular skill required by employers on Glassdoor and Stackoverflow, while SQL is the most common skill on Dice. More specifically, 64.64% of the data scientists jobs on Glassdoor required Python, and 57.43%, 57.03% of jobs on Stackoverflow and Dice require Python and SQL, respectively. Machine Learning, R, Java and Hadoop are also popular requirements for data scientists jobs. If we combine the data from the three job boards together (see Table A1 in Appendix), the most common skills that employers look for are (in the descending order of) SQL, Python, R, Machine Learning and Hadoop. Among these popular skills, SQL and Python are needed for more than 50% of the jobs in the combined data from the three job boards.

3.2 Number of Data Scientists Jobs Varies by Location and Industry

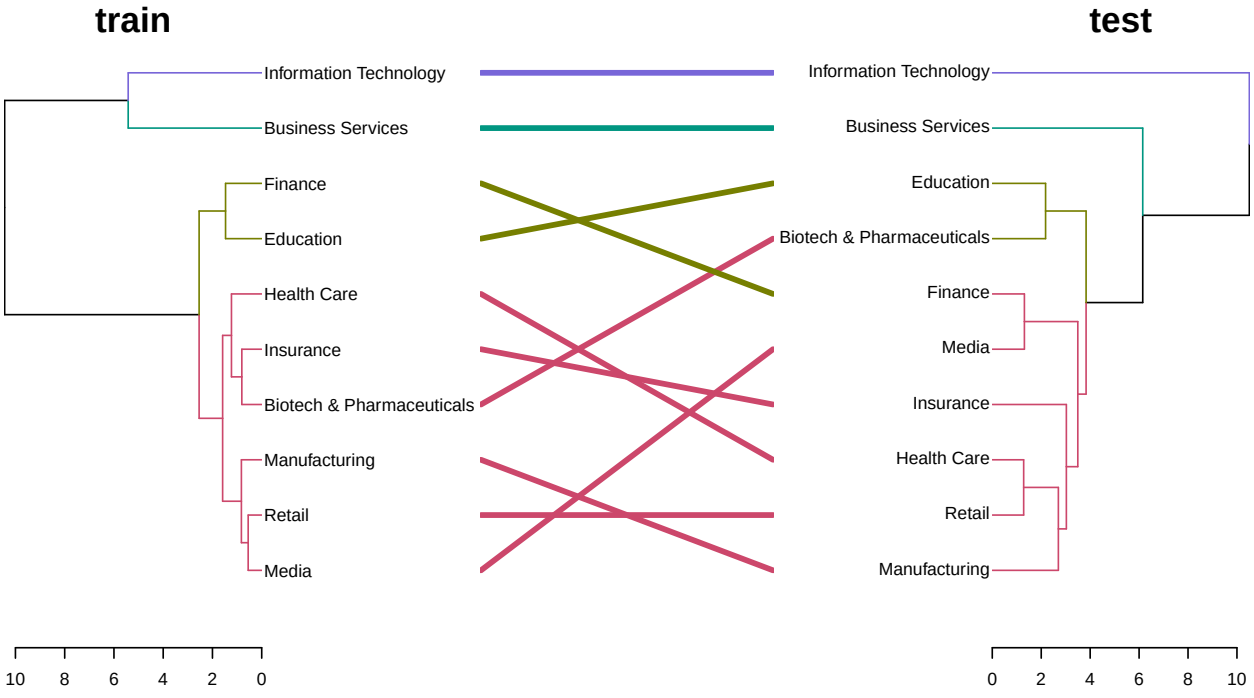
We use the combined data set from the three job boards to find out which locations have the most abundant data scientists jobs. In Figure A2 in Appendix, among all the 1043 jobs, New York, NY has the largest number of data scientists jobs. Other cities, such as San Francisco, CA and Chicago, IL also have many data scientists positions open. In detail, there are 81, 69 and 56 data scientists jobs in the location of New York, NY, San Francisco, CA and Chicago, IL, respectively.

We also draw a barplot to explore the relationship between the number of data scientists jobs and various types of industries. As we only have industry information from Glassdoor, Figure A2 in Appendix is created from the data set of Glassdoor with 444 observations, presenting the top 10 industries which employ most data scientists. Companies in the industry Information Technology employ much more data scientists than other industries. Business Services industry is also in a high demand for data scientists. Specifically, there are 149 data scientists jobs from Information Technology and 77 jobs from Business Services in total.

3.3 Hierarchical Clustering on Industries

Do industries have similarities based on their skill needs for data scientists? To answer this question, we first summarize the number of each skill grouped by industry, and then conduct a cross-validated hierarchical clustering on the industry type. Considering some industry have few observations, we only analyze the industries with more than ten observations. We set the number of clusters to be four, and use different colors to indicate different clusters. As shown in Figure 2, the two dendrograms from the train and test data set are very similar, with a cophenetic correlation of 0.82. Base on skill needs for data scientists, Information Technology and Business Services form two individual clusters; Finance and Education are grouped into the third cluster; Insurance, Health Care, Biotech & Pharmaceuticals and other three industries become the fourth cluster. Figure 2 indicates that some industries have similar requirements for data scientists, which is meaningful when we look into the relationship between industries looking for data scientists.

Figure 2. Cross-Validated Hierarchical Clustering for Industry Types Based on Skill Needs (Glassdoor, N=444). The number of clusters = 4. The same color indicates the same cluster.



4 Conclusion

In summary, the top three most popular skills that employers preferred for data scientists are SQL, Python and R. The city with the most data scientist job openings is New York, NY. The industry with the highest demand for data scientists is Information Technology. Some industries show similarity in terms of the skills needs for data scientists through the cross-validated hierarchical clustering. Information Technology and Business Services form two individual clusters; Finance and Education are grouped into the third cluster; Insurance, Health Care, Biotech & Pharmaceuticals and other three industries become the fourth cluster. The dendrograms of the train and test data are of high correlation, which means the results of hierarchical clustering are reasonable.

Due to the various description words of soft skills for data scientists, we did not perform a keyword search for soft skills from the online job boards. However, Data Science is a lot more than simply statistics or programming. The soft skills required of a data science job are as important, or even more important, than the hard skills we analyze in this report. As reviewed in Paramita Ghosh's article[2], we may conclude that the most unique skill for a data scientist is to address questions about a given business situation, find the root causes of the problems and give proper solutions.

5 Discussion

This study has several limitations. Firstly, the numbers of jobs from each job board are quite different, which may cause some extent of selection bias to the outcome of our analysis. Secondly, in the data collection procedure, we consider a skill mentioned in a job description as a required skill by the employer. However, if the jobs description says "no need for Python experience", for example, it would cause an opposite conclusion for the skill requirement of Python. Thirdly, As we only obtained industry information from Glassdoor, the sample size for jobs with industry types is relatively small, and several industries were abandoned in the hierarchical clustering due to rare observations. If we could assign the same industry types of Glassdoor to jobs from Dice and Stackoverflow, there would be a larger sample size and more industry types when we perform the hierarchical clustering.

References

[1] Shiv Shet (Apr. 29, 2016). An Introduction to Data Science. (https://dzone.com/articles/an-introduction-to-data-science, Accessed Oct. 9, 2017)
[2] Paramita Ghosh (July 12, 2016). A Comprehensive Review of Skills Required for Data Scientist Jobs. (http://www.dataversity.net/comprehensive-review-skills-required-data-scientist-jobs/, Accessed Oct 12, 2017)
[3] Hadley Wickham (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest
[4] steve-liang. (Mar 9, 2017) Scrape Data Scientist's Skills from Indeed.com. (https://github.com/steve-liang/DSJobSkill, Accessed Sep. 16, 2017)
[5] Hadley Wickham (2016). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. https://CRAN.R-project.org/package=stringr [6] Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr
[7] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
[8] Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.14.
[9] Tal Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics. DOI: 10.1093/bioinformatics/btv428
[10] David B. Dahl (2016). xtable: Export Tables to LaTeX or HTML. R package version 1.8-2. https://CRAN.R-project.org/package=xtable
[11] Claus O. Wilke (2017). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 0.8.0. https://CRAN.R-project.org/package=cowplot
[12] Stephen Cristiano, jobs_scrape.R, (https://slack-files.com/T6TTWE3G8-F75420VQD-7f4696035e, Accessed Sep. 23 2017)

Appendix

Table A1. Top Five Required skills in Combined Data from Stackoverflow, Dice and Glassdoor (N=1043)

	skill	count	percent
1	SQL	550	52.73%
2	Python	537	51.49%
3	R	425	40.75%
4	Machine.Learning	398	38.16%
5	Hadoop	311	29.82%

Figure A2. Top 10 Locations Employing the Most Data Scientists (N=1043);

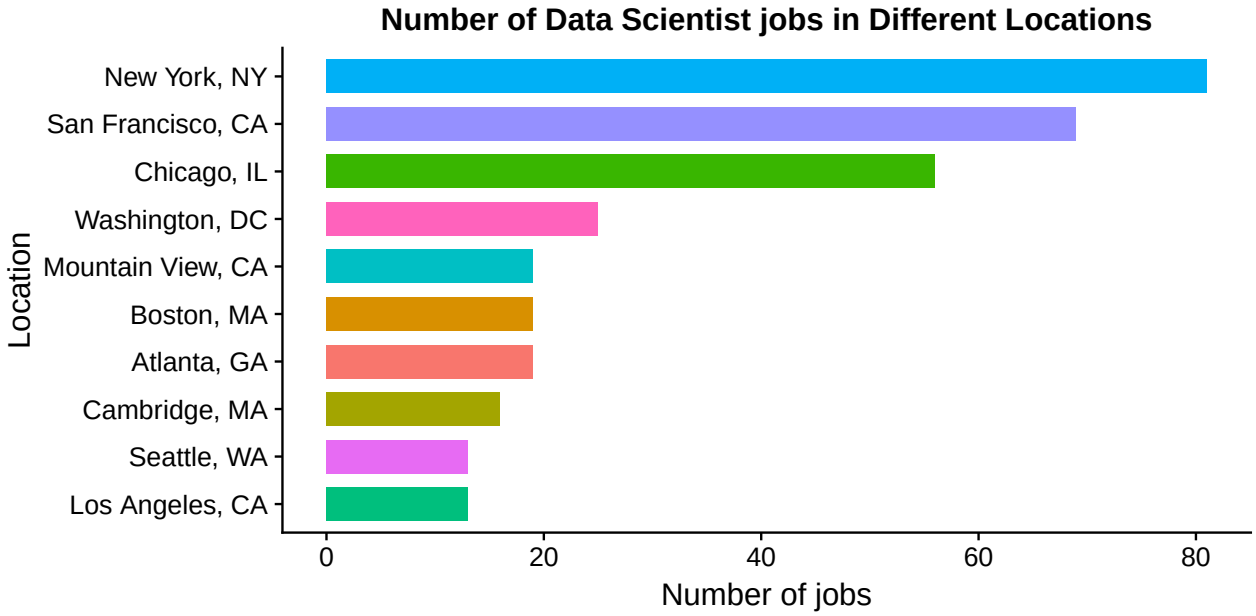


Figure A3. Top 10 Industries Employing the Most Data Scientists (Glassdoor, N=444)