Understanding the Current Job Market for Data Scientists

Jing Li

Introduction

It is known to all that the job market has been changing sharply over time in US. Back in 2000s, Healthcare and Biotechnology related occupations were of greatest interest, while in 2010s, a large portion of college students started to major in Computer Science due to the rapid growth of technical industry (U.S. Bureau of Labor Statistics, 2017). However, as the capacity of data increases dramatically especially for large companies, the demand for making use of the data and ultimately drawing business insights from analysis of the data grows as well (Strauss, 2017). And this directly causes the increasing demand for Data Science Expert who can tell a story from the data based on the application of Computer Science and Statistics. According to Glassdoor 2017 job market report, Data Scientist is taking over Software Developer as the best job in US based on the salary and job satisfaction. Indeed, the candidate pool will also grow with the demand and therefore knowing the recruiment criteria beforehand is essential in becoming a successful candidate.

This report aims at analyzing the current job market for full time Data Scientist positions; specifically the analysis will be focusing on skill requirement (programming and soft skills), geographical trend as well as industry preference. The hypotheses are as followed:

1. More companies now require R as the primary language to process data instead of Python or Perl (Muenchen, 2009).
2. The combination of statistical programming skills as well as teamwork makes the required skill sets for Data Scientist unique out of other technical positions (Press, 2014).
3. IT industry would be more likely to hire Data Scientists and West Coast prefers to hire data-related talents due to the large quantity of Technical Companies (pwc, 2017).

Data will be obtained from Monster.com through web scraping (details will be covered in Approach Section). Wordcloud, barplot and map will be used for exploratory analysis and Marascuilo Procedure will be used for testing. In general, technical firms are the major employers for Data Scientist but the locations of the companies do not seem to have significant impacts. In terms of pro gramming skills, employers still prefer Python slightly over R but this trend differs based on industry.

Approach

Data Mining

As mentioned previously, Monster will be used as primary sources to obtain the required data. The detailed procedures are as followed:

1. Search "Data Scientist" in Monster and only selects results for full time positions. Save the urls (40 pages in total and each page has 40 job posts) for each page.
2. Use Selector Gadget, a Google Chrome application, to obtain the html xpath for job title, company name, geographical location, industry as well as job description.
3. Use Rvest package to extract information from the html xpath and save it as data frame in r.
4. Since Monster's job posts may have different formats, data can only be extracted from the default format. In order to solve this problem, observations that have missing industry and skill sets information will be automatically deleted. Meanwhile, duplication will be removed either in order to avoid bias. The final data set contains 481 observations and 12 variables (company name, job title, location, industry and tags for R, Python, Perl, Hadoop, SQL, Perl, Phd Degree and Java).

Exploratory Analysis

Wordcloud, barplot and map are used for visulization. The job description is used for generating the wordcloud and the purpose of this approach is to check soft skills requirement for Data Scientist other than technical backgrounds. Industry map and barplot are generated using Tableau in order to examine the difference in job openings and requirement by location and by industry.

Statistical Testing

Marascuilo Procedure is used for testing if the occurance of each skill within each industry truly differs from each other. The null hypothesis is that the occurance of skill A does not differ significantly comparing industry B with industry C. The strength of using this approach is that it can measure the critical value for all pairs and then decide which pairs are significantly different regardless of the sample size difference (Wagh, 2016).

1. Calculate the absolute difference $(p_i - p_j)$ for all possible pair combinations. This is the test statistics for Marascuilo Procedure.
2. Compute the critical value using $r_{ij} = \sqrt{\chi^2_{1-\alpha,k-1}}\sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}$
3. Compare the test statistics with the critical value. The null hypothesis is rejected when the test statustics exceeds the critical values.

Result

Exploratory Analysis

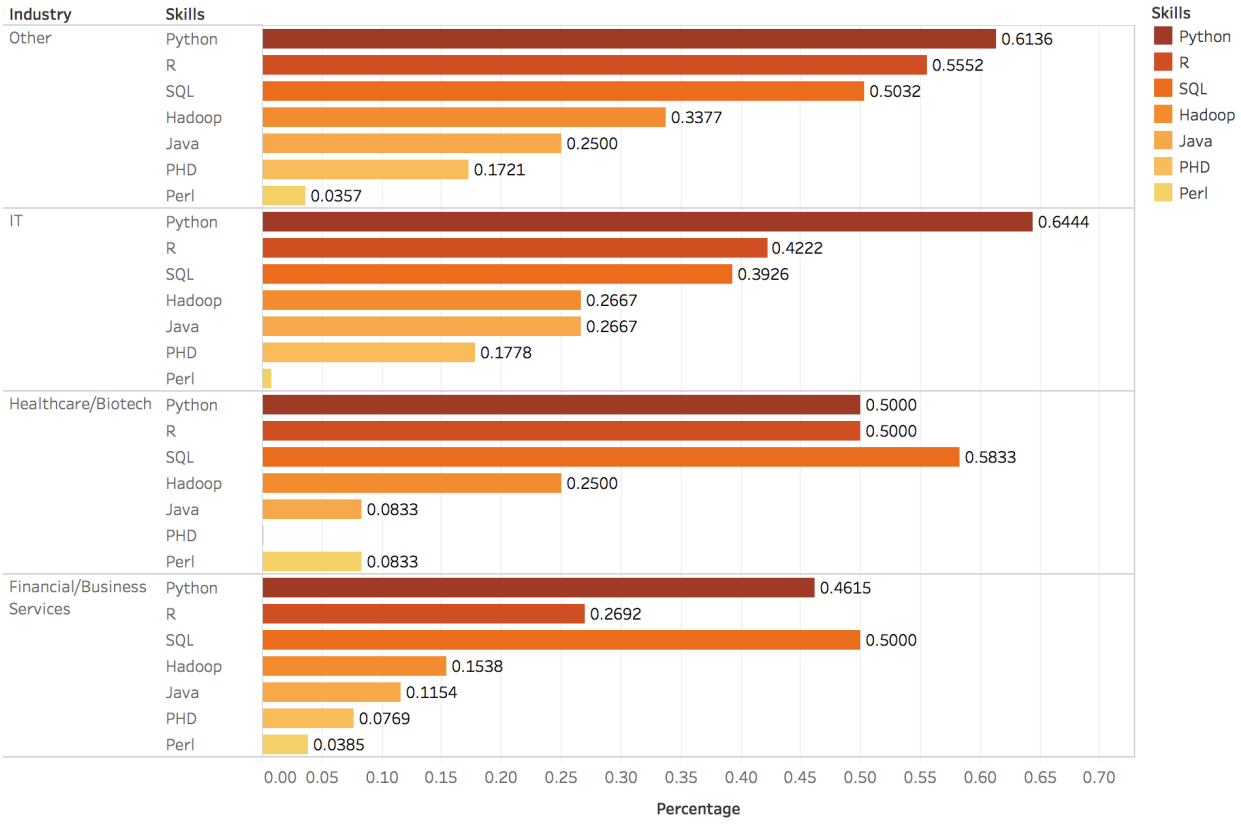

Figure 1: Bar Chart Based on Industry and Skills

Figure 2 in the appendix shows a map based on industry and the number of positions; larger circles indicate more openings within that area. In order for better presentation, only 3 types of industries are considered:

Financial/Business Services, Healthcare/Biotech as well as IT. The remaining industries are categorized into "other". As expected, technical companies are more likely to hire Data Scientist compared to other industries. However, based on Figure 2, technical companies from West Coast and East Coast do not differ significantly in hiring Data Scientist.

In order to dig more into the relationship of industry and skill requirement, the specific soft and technical backgrounds are being discussed in here. The wordcloud in the appendix is generated from the job description. As expected, technical backgrounds such as machine learning, statistics, modeling as well as analytics are the primary skill sets required for Data Scientist positions. Meanwhile, recruiters would also require candidates to have relevant experience in solving analytical problems and ultimately to provide business insights from the data. However, other than technical skills, the abilities to collaborate with the team, to support other colleages and to work with customers are necessary as well.

Table 1 in the appendix further quantifies the requirement for technical backgrounds. Approximately 61.1% of Data Scientist jobs posted on Monster require Python background while only 50.1% of positions require R, which contradicts with our hypothesis that companies now have preference for R over Python. Interestingly, only 15.4% of jobs require Phd degrees and less than 3% of jobs ask for old-school programming languages such as Perl.

Combining the above information, the bar chart from Figure 1 further examines the relationship of industry and skill requirement. Based on the plot, Python is still the primary language especially for IT industry. Interestingly, Healthcare/Biotech and Financial/Business industries value more about SQL over Python and R. Perl is most popular among Healthcare/Biotech industries but the percentage is still significantly lower; only 8.3% of Healthcare/Biotech companies require Perl. Surprisingly, Phd degrees are not a requirement at all for Healthcare/Biotech companies (this may happen due to small sample sizes).

Statistical Testing

Table 2. Result from Marascuilo Procedure

Skills	$Other - IT$	$Other - Healthcare$	$Other - Finance$	$IT - Healthcare$	$IT - Finance$	$Healthcare - Finance$
Python	0.098(0.143)	0.02(0.411)	0.251(0.256)	0.078(0.421)	0.153(0.271)	0.231(0.471)
R	0.133(0.143)	0.055(0.411)	**0.286(0.256)**	0.078(0.421)	0.153(0.271)	0.231(0.471)
SQL	0.111(0.142)	0.08(0.406)	0.003(0.285)	0.191(0.415)	0.107(0.298)	0.083(0.483)
Hadoop	0.071(0.13)	0.088(0.357)	0.184(0.212)	0.017(0.365)	0.113(0.225)	0.096(0.402)
Java	0.017(0.127)	0.167(0.233)	0.135(0.188)	0.183(0.247)	0.151(0.205)	0.032(0.284)
Phd	0.006(0.11)	**0.172(0.06)**	0.095(0.158)	**0.178(0.092)**	0.101(0.173)	0.077(0.146)
Perl	**0.036(0.03)**	0.048(0.225)	0.003(0.11)	0.083(0.223)	0.038(0.105)	0.045(0.247)

The test statistics and critical values (in the parenthesis) from Marascuilo Procedure are attached in Table 2. Note that the null hypothesis is rejected when the test statistics exceeds the critical value. Based on Table 2, 4 comparisons are statistically different: R occurance comparing other with Financial/Business industry, Phd Degree occurance comparing other with Healthcare/Biotech industry and IT with Healthcare/Biotech industry, and Perl occurance comparing other with IT industry.

Summary

As expected, Technical industry is the main employer for Data Scientist. However, in terms of geographical preference, West Coast does not seem to have more openings; in fact, East Coast together with some states of Midwest have more positions for Data Scientist compared to West Coast, which contradicts with our hypothesis. Interestingly, financial and business services seem to have more openings for Data Scientist compared to Healthcare/Biotech and others.

In terms of the most common skills, Python is still the primary language required for Data Scientist but R still has the tendency to join the mainstream. Additionally, the breakdown of skill requirement differs based on industry; for Healthcare/Biotech and Financial/Business companies, SQL is actually the most common skill. The result from

Marascuilo Procedure further indicates that certain skill requirement does differ based on industry. For example, the occurance of R is statistically different among other and Financial/Business industry. In other words, R is not necessarily required for candidates applying Data Scientist positions in Financial area.

Other than technical background, successful candidates should also have abilites to work with team and to support colleages. To be more specific, the major role of Data Scientists is to provide insight based on data analysis and this process would generally require collaboration and communication with other teams. Therefore, it is the combination of programming skills and the ability of collaboration makes the required skill set for Data Scientist unique among other technical positions.

Appendix

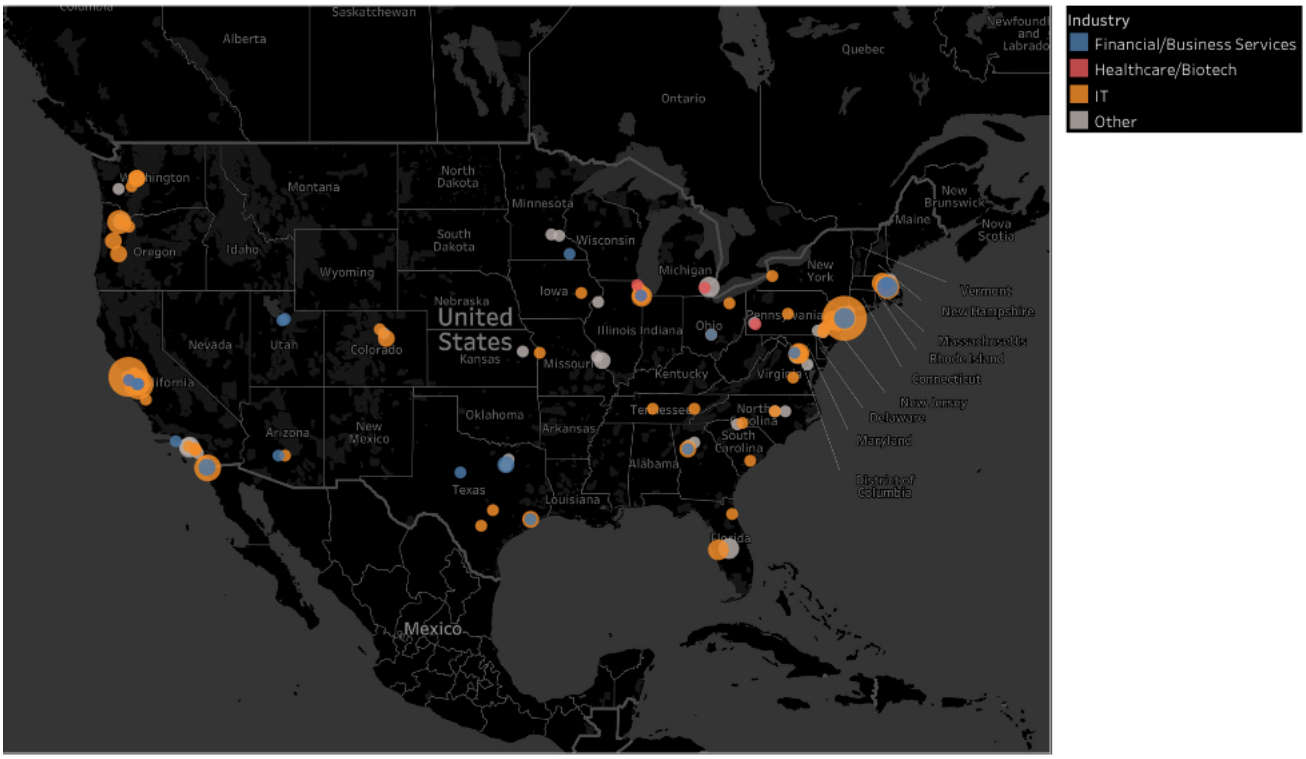

Figure 2: Map Based on Industry

Table 1. Technical Skills Occurance

Python	*R*	*SQL*	*Hadoop*	*Java*	*Phd Degree*	*Perl*
0.6103093	0.5010309	0.4701031	0.3030928	0.2453608	0.1670103	0.028866

Reference

U.S. Bureau of Labor Statistics: https://www.bls.gov/emp/ep_table_104.htm

Strauss, K. (SEP 21, 2017). Becoming A Data Scientist: The Skills That Can Make You The Most Money. Forbes. https://www.forbes.com/sites/karstenstrauss/2017/09/21/becoming-a-data-scientist-the-skills-that-can-make-you-the-most-money/#3b8d6cfc634f

Glassdoor Report: https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

PWC Report: https://www.pwc.com/us/en/publications/data-science-and-analytics.html

Muenchen, R.A. (2009) The Popularity of Data Science Software. http://r4stats.com/articles/popularity/

Press, G. (FEB 11, 2014). Big Data Debates: Individuals Vs. Teams. Forbes. https://www.forbes.com/sites/gilpress/2014/02/11/big-data-debates-individuals-vs-teams/#63b247a17e25

Coordinate Data was obtained from: https://www.gaslampmedia.com/wp-content/uploads/2013/08/zip_codes_states.csv

Sunanda T Wagh, Naser Ahmed Razvi. Marascuilo method of multiple comparisons (an analytical study of caesarean section delivery). International Journal of Contemporary Medical Research 2016;3(4):1137-1140.

Personal Communication: Stephen Cristiano

Personal Communication: Shannon Wongvibulsin