Analysis of American data scientist Jobs from Glassdoor shows critical skills and factors impacting salary and job satisfaction

Bingkai Wang

October 7, 2017

1 Introduction

In recent years, data scientist has become a popular job and a reasonable choice for many new graduates. The main responsibility of this job is to find, interpret, merge, visualize, model, present and communicate data sets and insights. [1] In 2002, Harvard Business Review dubbed it "The Sexiest Job of the 21st Century". [2] McKinsey & Company also projected a global excess demand of 1.5 million new data scientists. [3]

A better understanding of the job requirement can help students to be better prepared in the job market. In this project, we performed a study on data scientist jobs, including the required skills, distributions in industries and geographical locations, salaries and employee's satisfaction based on web-scraped data. Our results suggests that coding skills, especially python, and theoretical skills like statistics and machine learning are highly valued. Furthermore, counter-intuitively, employee's satisfaction of their jobs are significantly correlated with the company size, but not with their salaries.

2 Data

Job data was scraped from Glassdoor (www.glassdoor.com) with title "data scientist" using `rvest` [4] package in `R` . Glassdoor is a job search engine and outperforms others in the sense that it contains reviews of employees to the companies and positions. From this website, 990 jobs posted in September 30th, 2017 were scraped. (Glassdoor API is not accessible for students and only 1000 jobs can be viewed by normal users.) `rvest` is a web-scraping package in `R` and can pull specific elements in the webpage by specifying URL and CSS. Using this package, we scraped job description (`CSS=.desc`), company name (`CSS=.padRtSm`) and job location (`CSS=.subtitle`). For getting salary and ratings, we selected job list (`CSS=.jl`) and then extracted them from text using `stringr` [5] package. For getting industry and company size, we pulled the whole page using `R` function `readline` and extract them from text.

The raw data set contains the following features: company name, rating, salary, location (city, state), company size, industry and full text of job description. Among these variables, rating is the overall rating of all ratings of all time. Salary is preprocessed to be the mean of the salary range provided in the website. 722 out of 990 job postings contain no missing value.

For exploratory analysis, missing values are omitted. For inferential analysis, for simplicity, only complete cases are used.

To identify "unique skills" of data scientist, we also scraped 990 job postings with job title "quantitative analyst", using the same method as for "data scientist". Since both of the two jobs have responsibility on coding and modelling, we define the "unique skills" of data scientist as those appear with higher frequency in job descriptions of data scientist than quantitative analysts.

3 Methods

3.1 Identification of "skills" from job descriptions

Our first goal is to identify phrases that represent "skills". The traditional way for achieving this is to tokenize the text (partition it into a vector of words), count the frequency of each word and remove the meaningless "stop words", such as "the", "we" and "of". A set of common stop words can be find *here*. However, this method fails to work in this case for two reasons. First, the common set of stop words does not contain enough words for identifying skills in job description. For example, words like "area", "opportunity", "career" have much larger frequency than skill words like "python", "statistics", but cannot be filtered out by stop words. Furthermore, such type of words is so many to filter manually. Second, skills are often appeared as phrases rather than single words. When performing tokenization, "machine learning", "data mining" and etc will be separated into "machine", "learning", "data" and "mining", which no longer have explicit meaning as skills.

To overcome these two difficulties, we provided a method a feasible way to extract "skills" from job description. Our method is based on the following observation: in job descriptions, skills are usually listed, starting with ":" and separated by ",", "and" or "or". Hence, if we partition the text by these symbols, skill phrases will be identified as a whole and appear in high frequency since other noise words will remain in long and unique segments. For example, in the following sentence of some job description,

> "BACKGROUND/EXPERIENCE: Demonstrates proficiency in most areas of mathematical analysis methods, machine learning, statistical analysis, and predictive modelling and in-depth specialization in some areas."

we will identify "machine learning" and "statistical analysis". As we collect more job descriptions, such skill phrases will appear many times and be distinguishable. Finally we use these skills as key words to search in each job description and get the ultimate count of skill words. The final step is necessary because not all skills are listed in each job description.

As a result, we identify skill phrases with the following steps:

- **Step 1** tokenize the text of job description by separator ":", ",", ".", "and" and "or";
- **Step 2** extract the tokens with 3 or fewer words;
- **Step 3** count the frequency of each token and remove those appearing 50 times or less (58 phrases remained);

- **Step 4** filter out those irrelevant words manually, such as “religion” and “gender” (32 phrases remained);
- **Step 5** Use the remained phrases as keyword, search them in every job description and produce an skill-company matrix with a binary value indicating whether one skill is required by one company.

The above steps have theoretical guarantee. Assuming that each skill appears in each job posting independently with probability p, then by binomial distribution, the skill will remain in the final list with probability $\sum_{i=50}^{990} \binom{990}{i} p^i (1-p)^{990-i}$. When $p \geq 0.07$, the latter probability is greater than 0.99.

3.2 Statistical analysis for salary and employee’s satisfaction

Given the skill-company matrix we get, we can perform statistical analysis on salary. We want to explore which skills can have positive influence on salary, i.e. leading to increase on salary. Also, we are interested in skills that can affect the salary. The method we use is multivariate linear regression. Basically, we regress salary onto all identified skills by assuming

$$salary = \beta_0 + \beta_1 \times skill_1 + \beta_2 \times skill_2 + \cdots + \beta_k \times skill_k + \varepsilon, \quad \varepsilon \overset{i.i.d}{\sim} N(0, \sigma^2),$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^t$ is the parameter vector. (See `final_code.Rmd` for code.) Given the result of linear regression, we extract and report the significant parameters ($p < 0.1$). Since all covariates are binary-valued, then the value and standard error of significant parameter can reflect how much salary increase or decrease can one skill bring.

Apart from salary, employee’s ratings of company can be regarded as an index for their satisfaction of this job. Hence we also studied which factors can influence satisfaction from company size, location, salary and industry. Skills are not included since they don“t have much causal connection. For categorical variable such as company size, location and industry, we used multi-way ANOVA without interaction term to analyze the connection:

$$rating_{ijkl} = \mu + size_i + location_j + industry_k + \varepsilon_{ijkl}, \quad \varepsilon_{ijkl} \overset{i.i.d}{\sim} N(0, \sigma^2),$$

where $size_i$, $location_j$ and $industry_k$ are fixed effect of the i-th company size and j-th location and k-th industry on rating. The model will return whether rating is significant different across covariate groups. We inspect whether those covariates that are significant ($p < 0.01$), which indicates influence on employee’s satisfaction. For continuous variable salary, we fit a linear model

$$rating = \beta_0 + \beta_1 salary + \varepsilon, \quad \varepsilon \overset{i.i.d}{\sim} N(0, \sigma^2)$$

to determine the impact of salary of rating by inspecting the p-value of ANOVA of this linear model. (See `final_code.Rmd` for code.)

4 Results

4.1 Exploratory analysis

Exploratory analysis was performed using data sets described in section 2. The result is shown in Figure 1. For Figure 1(A), we counted the frequency of each skill across all data scientist job postings and select those with high value. For Figure 1(B), we calculated the difference of frequency of each skill between "data scientist" job postings and "quantitative analyst" job postings and regard those skills with high positive differences as unique skills for data scientist. Figure 1(C) and 1(D) show the types of companies that hire most data scientists and the distribution of this job across the United States.

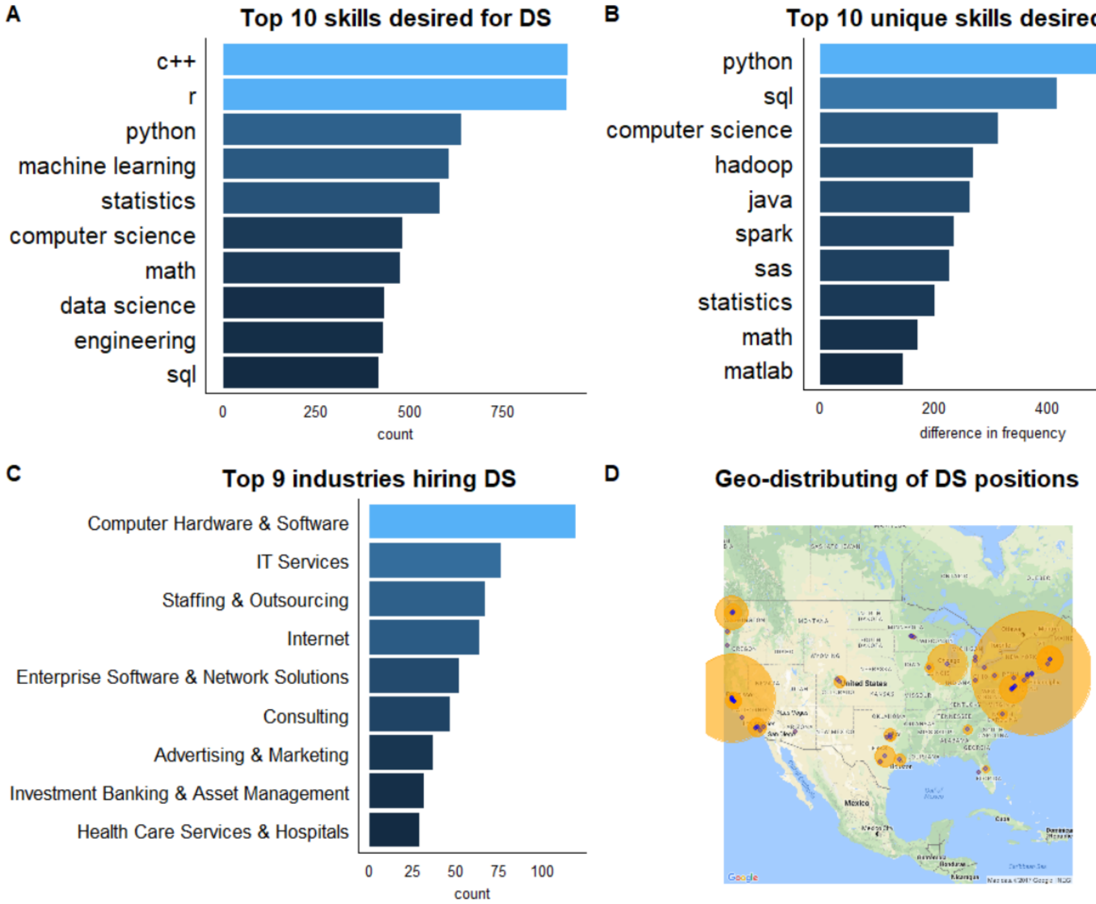

Figure 1: Visualization of exploratory analysis for Data Scientist (DS) jobs. Figure 1A shows 10 skills with highest frequency in job descriptions. Figure 1B shows 10 skills with most difference of frequency in job descriptions between data scientist and quantitative analysts. Figure 1C shows 9 industries with highest frequency in the job postings. Figure 1D shows the geo-distribution of data scientist jobs.

From Figure 1, we observe that

- Programming skills of C++ (low-level), R (statistical), Python (high-level) and SQL (database) are most welcomed by employers. Furthermore, Python and SQL are more emphasized by data scientist than quantitative analyst.

- Theoretical skills, such as machine learning, statistics, computer science and mathematics are also important, all of which are mentioned in at least half of the job descriptions.
- Big data skills like Hadoop and Spark are unique for data scientists.
- Employers of data scientists are mostly from IT, investment and health care companies.
- New York, San Francisco, Chicago, Seattle and Washington have the most open data scientist jobs.

4.2 Factors impacting salary and satisfaction

Table 1: Skills with significant ($p < 0.01$) influence on salary

Skill	Coefficient	S.E.	99% C.I.	P-value
Python	7.50	2.25	(1.71, 13.28)	0.00
SAS	-6.65	2.57	(-13.27, -0.03)	0.01
Applied Mathematics	-10.24	3.68	(-19.73, -0.76)	0.01

Table 2: Factors impacting employee's satisfaction

	Salary	State	Industry	Company Size
Mean square	0.00	0.57	0.50	4.43
P-value	0.91	0.01	0.04	0.00

Inferential analysis was performed for salary and rating by applying statistical methods described in section 3.2. For analyzing influence of skills on salary, multivariate linear regression model was implemented. For analyzing impact of salary, location, company size and industry on rating, we performed multi-way analysis of variance.

Table 1 shows skills that have significant influence on employee's salary. Significance is defined as the coefficient in linear regression is different from 0 with p-value less than 0.01. Python, SAS and applied mathematics are identified. Among them, only Python leads to a substantial increase of 7.50 in salary with standard error 2.25, which is understandable since Python has become one the most popular programming languages in recent years. On the contrary, SAS and applied mathematics affect the salary significantly. This fact might result from their decreasing popularity in current technology industries.

For employee's satisfaction, we considered four possible factors, salary, location, industry and company size, and their influence are presented in Table 2. Surprisingly, salary seems to have no impact on satisfaction, since the p-value in ANOVA is 0.91, far from being significant. Furthermore, company size is the factor that has most influence on satisfaction. This might be related to work pressure reflected by company size. Furthermore, location and industry also have an effect on explaining the variance of ratings, but not much.

5 Conclusions

In this project, we analyzed the common skills, unique skills, industries, locations, salaries and satisfaction of data scientist jobs using both exploratory and inferential analysis methods. Our analysis shows that data scientist jobs are mainly distributed in big cities and high-tech industries, and We found that coding skills, especially python, and theoretical knowledge in statistics and machine learning are highly valued in this area. We also provided some insight on finding skills that are important to increasing salary and factors that plays an role in job satisfaction. Interestingly, salary does not have much relation with satisfaction.

Beyond this project, much more meaningful analysis on job market can be performed. By collecting more data, trend of skills required, multi-job comparison and etc. can be studied. By using more complicated and meaningful models, job classification, job recommendation and etc. can be performed. Currently, many job search websites, such as linked in, has done much of them. In the future, the direction of hunting job and hiring people might be precise targeting the ideal position or the ideal candidate.

References

1. Wikipedia “Data science” Page. URL: https://en.wikipedia.org/wiki/Data_science. Accessed 10/10/2017.
2. Davenport, Thomas H.; Patil, DJ (Oct 2012), Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review.
3. Manyika James et al. (May 2011), “Big data: The next frontier for innovation, competition, and productivity”.
4. Hadley Wickham (2016). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2. https://CRAN.R-project.org/package=rvest
5. Hadley Wickham (2017). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.2.0. https://CRAN.R-project.org/package=stringr