Analysis of Data Scientists Jobs in the United States

Shulin Qing

October 9, 2017

Introduction

Big data are becoming ubiquitous in 21th century. Different from traditional data collected in past centuries, big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or old-fashioned data procession applications [1]. People are trying hard to unearth patterns from big data and are enthusiastic to predict the future trend. Data scientists are the names of this group of people. They have sophisticated discipline of statistics and mathematics and rigorous training in computer science, and blend these disciplines into a harmonious and united entity. New insights and new forms of value can be extracted from a large scale of data through the insights and knowledge of data scientists. Take the recommender system as an example of the application of data science to our daily life. A great number of companies have fervidly used recommender system to promote their products according to their online customer's interests and relevant information [2]. When we browse Youtube, we may often notice that some video suggestions are showed up based on our previous watching history. The recommender system summarizes and explores information from customer's search results and uses algorithm and statistical models to predict preferences of customers. Nowadays, data scientistis jobs has been becoming one of the most popular jobs to pursue. As noted by the blogger, Christopher Watkins, from Udacity, data scientists can be seen as a hybrid of data hacker, analyst, communicator, and trusted adviser [3]. The promising perspectives attract an increasing number of people to become a member of the data scientist team. What they pay attention most are the most common skills that employers look for and the most unique skills that would make employers impressive. Moreover, understanding what types of companies employ the most data scientists could help job seekers get an overview of their future job industries and get a sense of whether the jobs would match their life goals. In this data analysis project, we performed data analysis of data scientistis jobs listed on

"Glassdoor", one of most popular job search websites, and the analysis results would be used as data evidence to answer the above questions.

Method

For this project, we explore data science related jobs in the United States posted on glassdoor.com which is a job aggregator that updates multiple times daily. We conducted webscraping the job links and page links by using "rvest" R package and Chrome SelectorGadget extension to gather information from multiple glassdoor's pages and then used dplyr,stringr,tidyr and ggplot2 R packages to clean data and make visualization tables and graphs.

Data Collection and Data Management

We searched for the keyword "data science" on glassdoor.com. Each page of job results have 30 job postings and our data were collected from 30 pages. All of the information on a web page is coded with HTML tags. HTML is the coding that tells internet browser how to display a given page's contents upon accessing it [4].First, we looped through each job listing on a page. In order to acquire the CSS seclectors for webpage element, for example, the job link, the Chrome SelectorGadget extension was used. By clicking on the page element that we would like the selector to match, a minimal CSS selector for the page element would be generated in a box in the bottom right of the website. Using the functions html_nodes() and html_text() can easily extract and then read pieces of information out of HTML documents. For each job link, we wanted the URL of the page the link goes to. Hence, html_attrs() was used to extract the attributes, text and tag name from html and selecting one of the attributes, "href", specifies the link's destination. Through each data science job listing, we collected information about company names, locations, job titles, and evaluated if each job description contains a keyword in a specified data science skill set. The specified data science skills were chosen from websites including "Forbes", "Udacity" and "mastersindatascience.org" and they are Python,R,SAS,SQL,Java, Tableau, C, Perl, Excel, MATLAB,and HIVE. The Chrome SelectorGadget extension is not effective on getting a useful CSS selector for company industries and company size, because glassdoor website is not well structured. We instead used function readLines() to read all text lines from a job link connection and then used str_detect and str_extract to detect and extract keyword related to industries and company size. Then, by following the method above, we looped across 30 pages and collected raw data for data scientist jobs.

The raw data set contains 921 observations. After the duplicated observations were removed from our raw data set, data cleaning was performed. We recoded each skill originally in True/False to 1/0 for our later data analysis. Then we identified total number of missing cases from each variable and found that they are all less than 10%. Thus, our statistical analysis would be not likely to be biased by the missing data. After missing data were removed, the total number of observations in the cleaned data set is 531.

For our data analysis purposes, we assigned each industry in our data to an industry category on the basis of the information given by "United States Census Bureau" [5]. Therefore,63 industry types in the original data set were collapsed into 16 industry categories.

Exploratory Data Analysis

Exploratory data analysis was used to summarize the most common and most unique data science skills that employers look for, and to investigate the attributes of companies that demand for data scientists. The occurences of skills and job locations were ranked and visualized by using bar charts from R package ggplot2. The industry category combined with company size for each company is visualized by using stacked bar chart, which helps us to learn the relationship between company size and industry.

Statistical Modeling

To predict the industry that a person with a specific computing skill would find a data science job in, we used conditional inference trees (ctree) modeling. The ctree modeling uses recursive binary partitioning and develops a easy way to visualize decision rules for predicting a categorical outcome, industry categories [6].

Reproducibility

All analyses conducted in this project can be reproduced in the R markdown file Exploratory-DataAnalysis.Rmd. To reproduce the exact results presented in this data analysis report, the cleaned data set "glassdoor_dataset_final.csv" must be used as the data webscrapped from glassdoor.com changes over time.

Results:

As shown in Figure 1, the top five skills that data scientist jobs look for are Python, R, SQL, Java and SAS. Among the 531 job listings, 301 (56.7%) jobs require the Python skill, 278 (52.4%) jobs require the R skill, 239 (45.1%) jobs require the SQL skill, 129 (24.3%) jobs require Java skill and 127 (23.9%) jobs require the SAS skill. The three least popular or most unique skills that employers look for are Excel, MATLAB and Perl. Among the 531 job listings, 60 (11.3%) jobs require the Excel skill, 57 (10.7%) jobs require the MATLAB skill and 20 (3.8%) jobs require Perl skill.

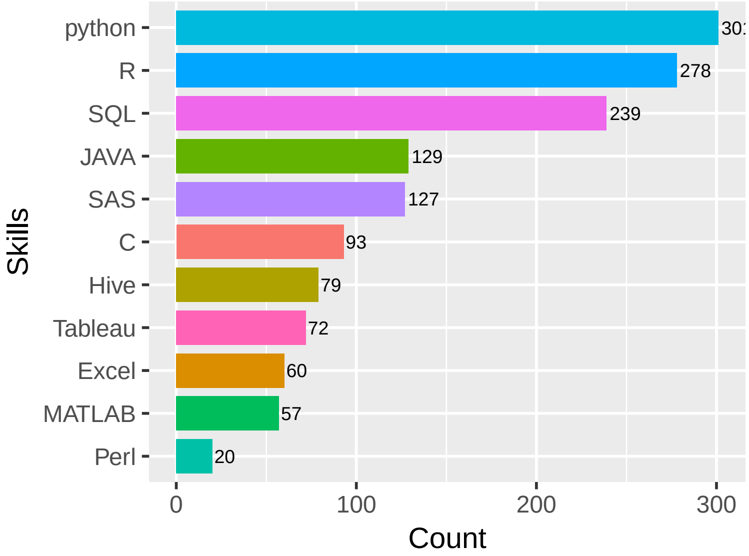

Figure 1: Required Skills in Job Listings

The stacked bar chart of industry job categories is shown in Figure 2. The bars are stacked to different colors which represents different company sizes. Each colored rectangle represents a combination of industry and company size. As shown in the bar chart, the top four industries to find a job in data science are "Professional, Scientific and Technical Services", "Information" and "Manufacturing". As defined by the United States Census Burea, professional, the industry "scientific and technical Services" includes industry subcategories: legal services, accounting, engineering related services, computer systems design, and scientific and technical consulting services, etc. This industry provides 133 (25.0%) data science related jobs. Information industry includes publishing, motion picture, radio and television broadcasting, and telecommunication services, etc. It provides 131 (24.7%) data science related jobs. Manufacturing industry provides 68 (12.8%) data science related jobs. In professional,scientific and technical services industry, most companies that demand for data scientists are large companies that have more than 250 empoyees. Same patterns can also be seen in other industries that have observations more than 10, except administrative and support industry. According to international standards, greater than or equal to 250 employees is taken as an indicator that it is a large business and less than 50 employees is taken as an indicator that it is a small business [7]. Too few observations (less than 10) in the industries may bias our analysis so we exclude them from our discussion. However, in administrative and support industry, more than a half of companies that are in need of data scientists are small companies as shown by the red color.

Figure 3 suggests that the top four cities that have the most number of data science jobs are New York (NY), San Francisco (CA), Chicago (IL) and Cambridge (MA). These cities are nationwide job centers which emcompass lots of tech companies. The prediction of industry categories from data science skills, Python, R, SQL and SAS, is shown in Figure 4. The data is divided into two categories according to whether a person knows Java or not (for example: Figure 4a). For each condition or node, the predicted job industry bar graph is plotted. Comparing the two bar graphs, we could see that equipped with Java skill, a person is predicted to be more likely to find a job in professional, scientific & technical services industry and information industry, but less likely to find a job in manufacturing industry, and finance

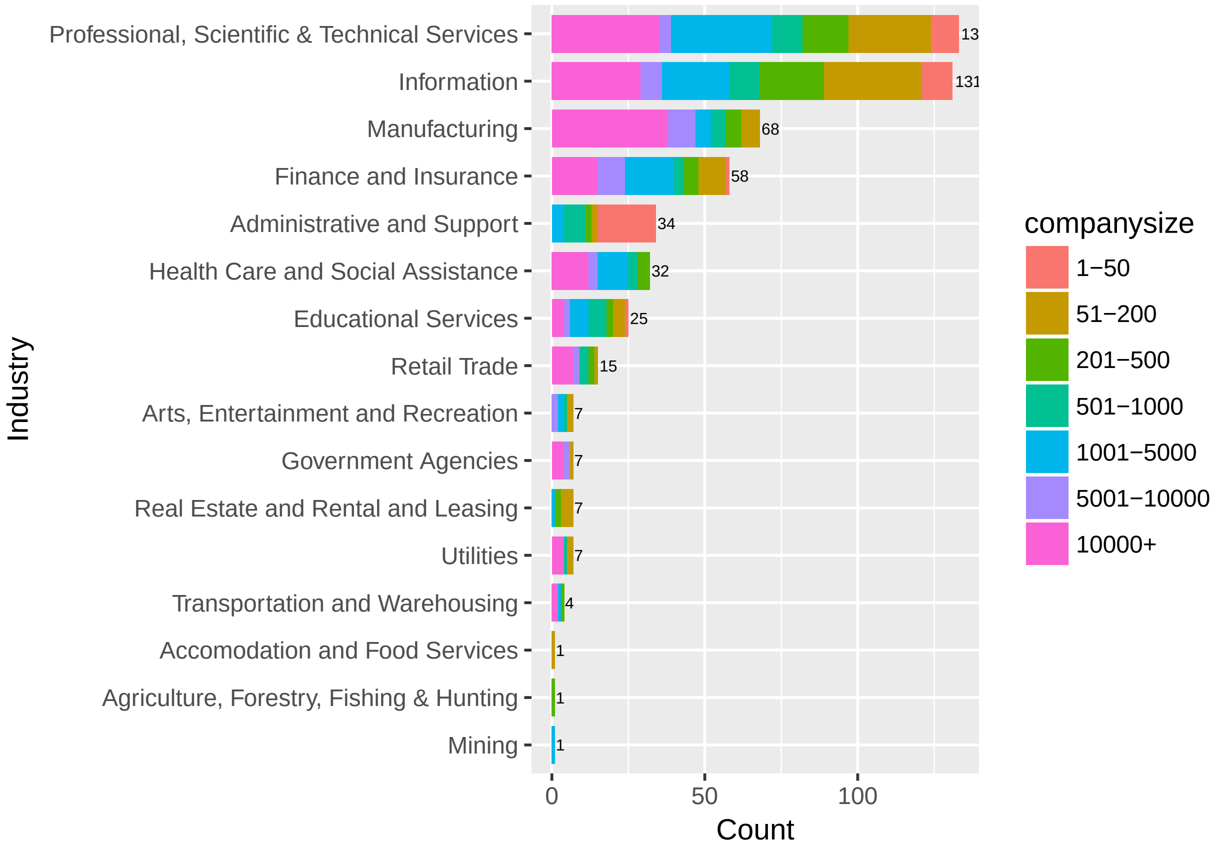

Figure 2: Industries that Need Data Scientists

and insurance industry. This may provides information to job seekers that manufacturing and finance and insurance industries depends less heavily on Java. In contrast, in Figure 4b, we could find that a person who knows SAS is predicted to be more likely to find a job in finance and insurance industry, health Care and social assistance industry, and professional, scientific & technical services industry, but less likely to find a job in information industry and manufacturing industry. This may be due to the fact that in the health care industry, SAS is used because it is what the FDA uses and prefers, and in finance and insurance industry, SAS Financial Management is widely used to analyze financial data. Hence, for job seekers who want to find jobs in finance and insurance industry or health care industry, learning SAS may improves their possibilities.

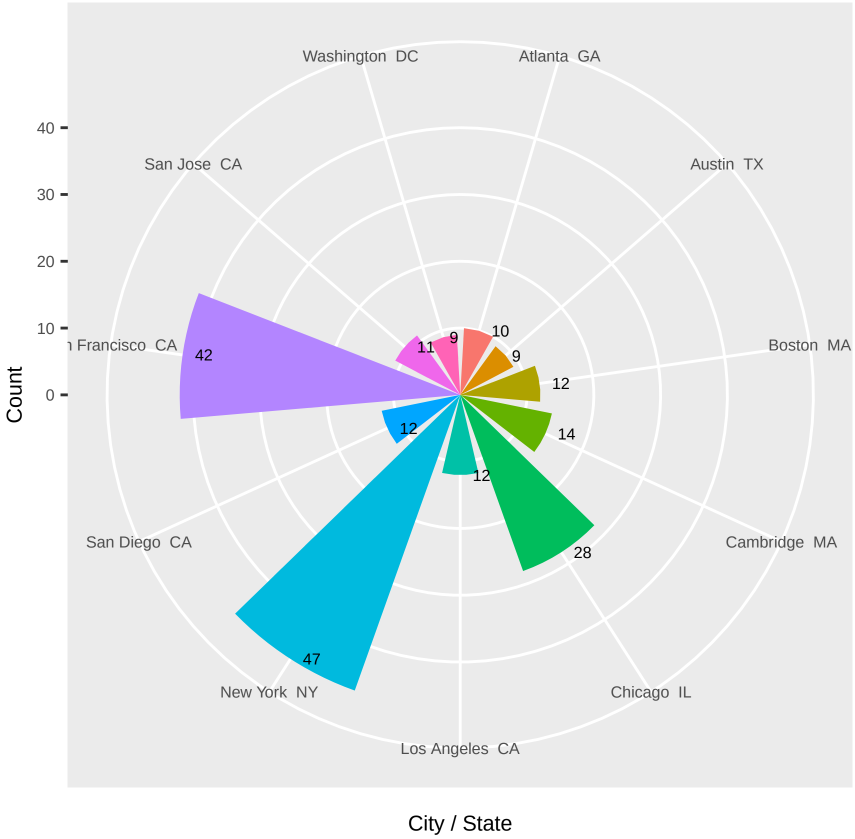

Figure 3: Top 10 Cities with the Most Job Listing

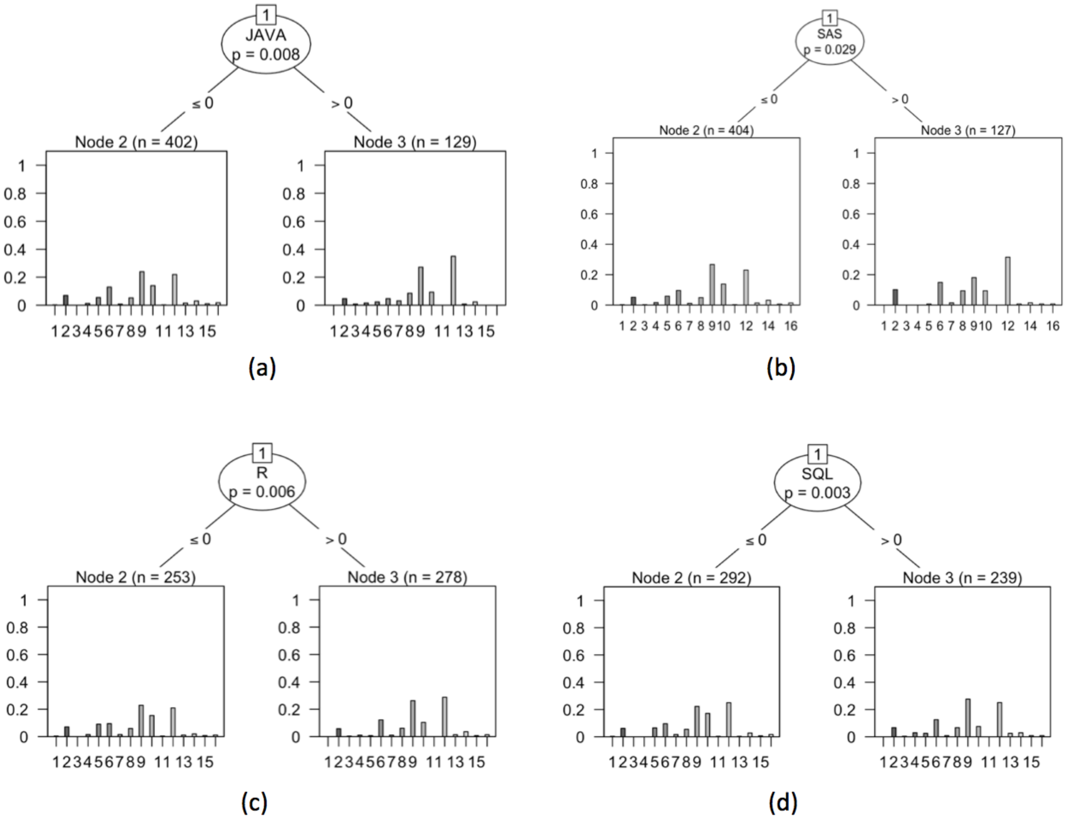

Figure 4: Conditional Inference Trees (1)Accomodation and Food Services (2)Administrative and Support (3)Agriculture, Forestry, Fishing & Hunting (4)Arts, Entertainment and Recreation (5)Educational Services (6)Finance and Insurance (7)Government Agencies (8)Health Care and Social Assistance (9)Information (10) Manufacturing (11)Mining (12)Professional, Scientific & Technical (13)Real Estate and Rental and Leasing (14)Retail Trade Services (15)Transportation and Warehousing (16)Utilities

Discussion:

References

1. Press, Gill “12 Big Data Definitions: What’s Yours?”, Forbes Page. URL: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#199010c513ae. Accessed 10/10/2017.

2. Analytics Vidhya “13 Amazing Applications / Uses of Data Science Today” Page. URL: https://www.analyticsvidhya.com/blog/2015/09/applications-data-science/. Accessed 10/10/2017.

3. Watkins, Christopher “Hottest Jobs in 2016 #3 Data Scientist”, Udacity Page. URL: https://blog.udacity.com/2016/01/hottest-jobs-in-2016-3-data-scientist.html. Accessed 10/10/2017.

4. Salmon, Michael “Web Scraping Job Postings from Indeed”, Medium Page. URL: https://medium.com/@msalmon00/web-scraping-job-postings-from-indeed-96bd588dcb4. Accessed 10/10/2017.

5. United States Census Bureau, “Annual Capital Expenditures Survey (ACES)” Page. URL: https://www.census.gov/programs-surveys/aces/information/iccl.html. Accessed 10/10/2017.

6. Quik-R “Tree-Based Models” Page. URL: http://www.statmethods.net/advstats/cart.html. Accessed 10/10/2017.

7. Support Site, “What is the definition of Small, Medium and Large business within the Observatory?” Page. URL: http://support.spikescavell.com/faqs/what-is-the-definition-of-small-medium-and-large-business-within-the-observatory/