Analysis of Data Science Skills Required by Employers

Yeya Zheng

1.Introduction

Data scientist was ranked the best job in America by Glassdoor because of its abundant career opportunities, high earning potential, and increasing demand. The McKinsey Global Institute estimates that the "United States alone will face a shortage of 140,000 to 190,000 people by 2018 with deep analytical skills". With such bright career prospect, the competition to launch a career in data science is even more fierce. Successful applicants often possess a combination of technical, communication, and analytical skills. Since there are many things to learn in this field, we are wondering what might be the most demanding skills for data science related jobs in each industry, so that we can focus on developing these skills during school and get hired into our dream positions after graduation.

This project performs an analysis of "data scientist" jobs listed on some major job boards including Careerbuilder and Stackoverflow. The objective of this project is to understand the most common skills that employers need for data science related jobs,and identify the most unique skills that employers need for data science related jobs as compared with other more traditional statistics jobs(analysts, biostatistician,staff statistician etc.), also identify the industries that employ the most data scientists.

2.Method

Job searching engine selection

We used Careerbuilder and StackOverflow as our Job searching engine. CareerBuilder is one of the largest online employment websites in the US. It is a general board with postings across a broad range of industries. Stackoverflow is a job board targeted for developers.Therefore, unlike traditional job board, such as Careerbuilder,Glassdoor, and indeed, it has more job listings from tech-company.

Getting the data

After We put "data scientist" as our searching keyword on Careerbuilder, the website directs us to the first page of our searching result.The page lists job title, company name, company location and job link to a more detailed job description. We pulled job titles, company name, location, and job links using the ".job-row" CSS selector. We then looped through all the individual job link and pulled the label which described the job type (full time/contract/temporary/part-time) and industry classification using ".small-12+ .row .columns" CSS selector, and pulled the job descriptions using the ".description" CSS selector.We repeated the same process to scrape job listings for the next 6 pages. We scraped 175 data science related job from CareerBuilder. We also put "statistics" as our searching keyword for more traditional statistics jobs, and repeated the same process, We scraped 250 job listings for "statistics" keyword, however,some of them were still "data science" related jobs. We removed the listings whose

titles contain words like "data science" or "engineer" and there were 155 listings remained for traditional statistics work.

On Stackoverflow, We used "data science" and "statistics" as our searching keywords for data science related and traditional statistics jobs respectively.We pulled job titles, company name, location, and job links using the ".-job-summary" CSS selector. We then looped through all the individual job link and pulled the industry classification using the ".-about-job" CSS selector, and pulled the job descriptions using the ".-skills-requirements , .-job-description" CSS selector. We scraped 123 data science and 36 traditional statistics related jobs.

Data Processing

We firstly compiled the scraped data from CareerBuilder and StackOverflow into one dataset in order to manipulate the industry categories. The industry categories were not listed in a consistent way and there were more than 300 levels for industry categories. We did some text manipulation to group similar industry categories together. Detailed steps for industry grouping will be mentioned in the Appendix. After grouping, there were 14 industry categories, including Business Services, Health Care, Retail,Finance,Information Technology,Manufacturing, Education,Aerospace & Defense,Oil, Gas, Energy & Utilities, Media, Telecommunications,Tourism,Government,and Agriculture & Forestry.

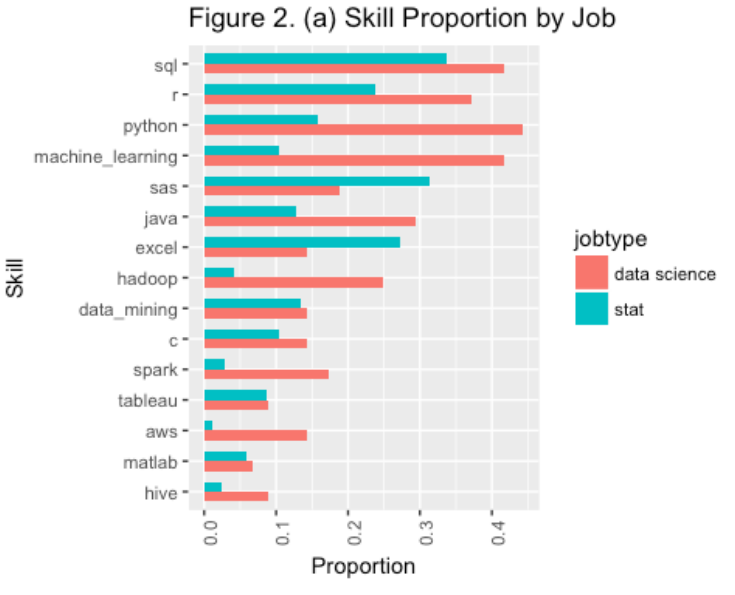
We also extracted education requirements for each job. For each job, if an exact match is found with "master","masters","ma",or "ms", then We counted master degree once; if an exact match is found with ""doctor","doctoral" or "phd", then We counted PhD. degree once; and if a match is found with "bachelor","bachelors","undergrad",or "undergraduate","bs",or "ba" We counted bachelor degree once. Sometimes, a job could contain two or more degrees, in that case, only the highest degree will be recorded. For example, if a job has 1 count for master and 1 count for bachelor, then We assumed the education requirement for this job is master and put "master" as the highest education requirement for this job. By doing this iterating over all jobs, We obtained the highest education requirement among all jobs. There were 227 jobs required only bachelor degrees, 159 jobs required master degrees, and only 79 jobs required phd degrees.

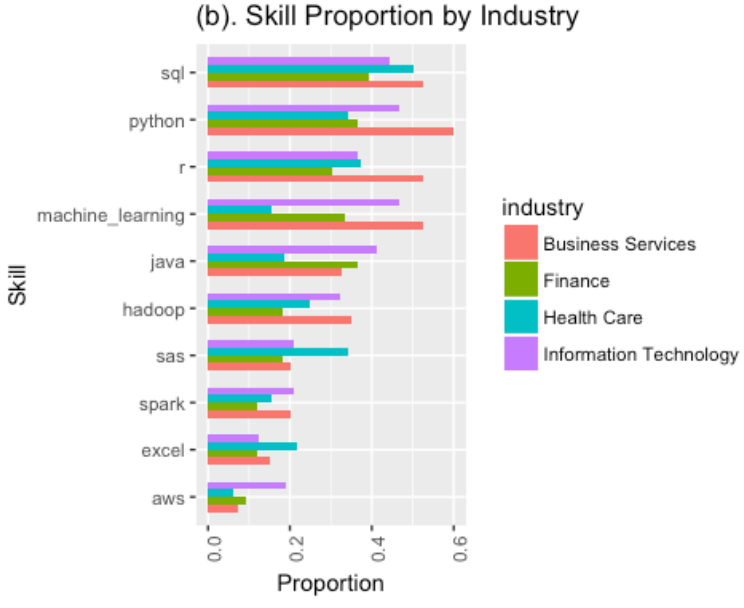
Then We extracted job skills required for each job. In order to do that, we created a list of job skills that are most commonly used in Data Scientist Job, including Hadoop, Python, SQL, R, Spark,SAS,AWS,Excel,Azure,Java,Tableau, Deep learning, Machine learning, AI,Pandas, Scipy, C,Perl,Text mining, Matlab,Hive,Splunk, Data mining, Unix/Linux, and Apache. Then we searched for pattern match in each job description, we counted each skill once if a match is found. By doing this iterating over all jobs, we obtained a total count of each skill among all jobs. After removing duplicated job listings, our final dataset contained 465 observations in total.

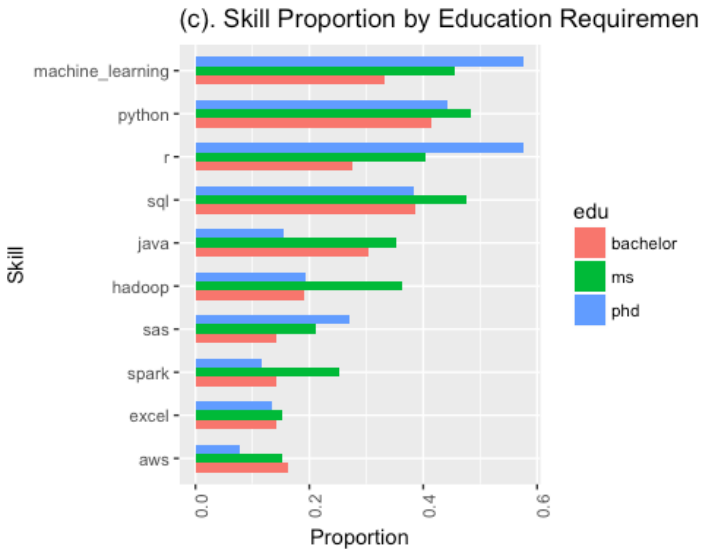
Exploratory Analysis

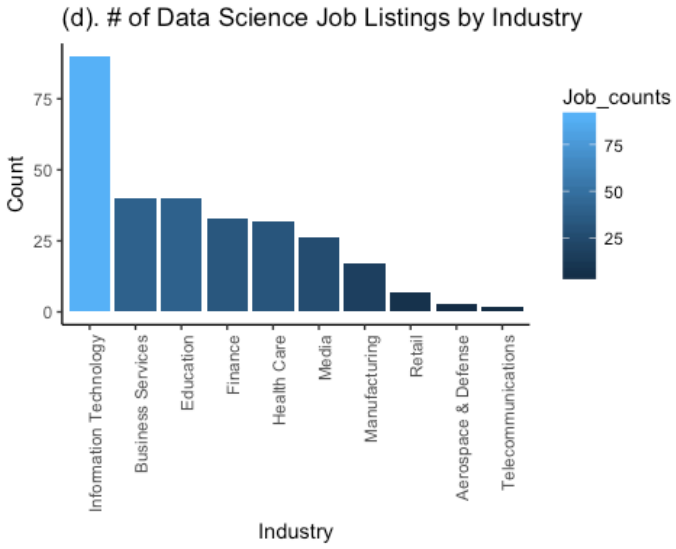
We firstly examined the geospatial distribution of the data science jobs in our analysis, and our exploratory plots are shown in Appendix Figure 1. From Figure 1. (a), it appears that we have more job listings on the east coast than on the west coast as we have many jobs clustered around NY, VA, and DC. b) shows number of data science jobs by city. The Barplot shows that New York seems to be the city with most data science job openings in our analysis, following by San Francisco, Chicago, and Atlanta.

Figure 2 . (a) shows the proportion of each skill count in descending order, broken down by jobtype (data science job/traditional statistics job). The plot only displays the top 15 skills with most counts.From the plot, we can conclude that the biggest discrepancies between data science job and traditional statistic jobs in terms of skill requirement lies in Python, machine learning, Hadoop, Spark, and Amazon Web Service(AWS).More than 40% data science job listings require Python skill while around 15% traditional statistics jobs require Python skill. More than 40% data science job listings require machine learning skill while only 10% traditional statistics jobs require it. Around 25% data science jobs require Hadoop while less than 5% traditional statistics jobs require it, around 18% data science jobs require Spark skill while around 3% traditional statistics jobs require it, around 15% data science job require AWS while less than 2% traditional statistic jobs require it. In contrast, traditional statistics jobs have more requirement in SAS and Excel as compared with data science job. Both jobs seem to have similar requirement on Tableau. (b)shows the proportion of each skill count for data science jobs, broken down by industry. We only compared industries of information technology, health care, finance and business service, as we see these are the major industries that cover most of our data science jobs.From the plot, we conclude that the Business Services industry seems to require more Python skill, R skill,and machine learning skill as compared with the other three industries. The information industry seems require Java the most, and the health care industry seems to require SAS the most. For the top ranking skill, SQL, the health care industry and the business service industry seem to have more requirement as compared with the other two industries. (c) shows the proportion of each skill count for data science jobs, broken down by requirement in educational degree. We can see that for data science jobs that require PhD. degree, they require machine learning skill, R skill, and SAS skill the most. For data science jobs that require Master degrees, they require SQL skill, Python skill, Java skill,Hadoop skill, and Spark skill the most. (d) shows number of data science job listings by industry. From the plot, we can tell that we have most job listings coming from Information technology industry, following by business service industry, education industry, finance industry, and health care industry.









Data Analysis

In order to understand the most unique skills required for data science jobs as compared with traditional statistics jobs， we conducted a two sample t-test for each skill, and use Bonferroni correction to adjust for multiplicity issue (see supplementary code chunk "ttest"). For our two sample t-test， our null hypothesis is there is no difference between data science job and traditional statistic jobs with respect to each skill. And our alternative hypothesis is that data science jobs require more as compared with traditional statistics jobs with respect to each skill. Table1. shows the results for two-sample t-test after Bonferroni correction (only displayed the significant skills). It appears that Machine learning， Python， Hadoop， AWS， Spark, Java, AI, Hive, and R all show significant differences between data science jobs and traditional statistics jobs with reported one-sided p values as 0， 0， 0， 0, 0, 0, 0.0001,0.0007,and 0.0010 respectively, which means if the null hypothesis is true, the probability of obtaining an effect at least as extreme as the one in our sample data is 0， 0， 0， 0, 0, 0, 0.0001,0.0007,and 0.0010 respectively. Therefore, we reject the null hypothesis and in favor of the alternative hypothesis that data science jobs require more Machine learning skill,Python skill, Hadoop skill, AWS skill,Spark skill,Java skill,AI skill,Hive skill, and R skill as compared with traditional statistics jobs. we also computed the p-value for the permutation test by randomly shuffling the "data science" and "stat" job labels to generate the sampling distribution of the test statistic under the null, and computed the proportion of times that the test statistic would be at least as extreme as we observed.We got 100 resamples and repeat the process for 1000 times to obtain the uncertainty of p-values. The p-value for the permutation test are shown in “P-value permutation" column, along with its standard deviation and 95% confidence interval. Using permutation tests, "Artificial Intelligence" and "Hive" became insignificant.

Table 1. Two Sample t-tests to Compare Skill Counts of Data Science Jobs and Statistics Jobs

Skill	T Statistics	P-value	P-value Permutation(SD) 95% CI
Hadoop	7.0713	0.0000 *	0.0000(0.0000) 0

Python	7.1251	0.0000 *	0.0000(0.0000) 0
Spark	5.6549	0.0000 *	0.0000(0.0000) 0
AWS	5.9641	0.0000 *	0.0000(0.0000) 0
JAVA	4.4866	0.0000 *	0.0000(0.0004) (0.0000, 0.0009)
Machine learning	8.3910	0.0000 *	0.0000(0.0000) 0
AI	3.6820	0.0001 *	0.0021(0.0044) (0.0000, 0.0108)
Hive	3.2348	0.0007 *	0.0028(0.0053) (0.0000,0.0131)
R	3.0973	0.0010 *	0.0009(0.0030) (0.0000,0.0068)

*Each test is evaluated at a significance level of 0.05/25=0.002 to adjust for multiple test issue.

To understand the association between degree requirement and skill requirement, we fitted a logistic regression with whether the job required PhD (see supplementary code chunk "logisticmodel"). degree as the outcome variable, and skill count for each job as predictors. The logistic regression model could be expressed as below:

$$logit(Pr(Y_i = 1)) = \beta_0 + \beta_1 hadoop + \beta_2 python + \beta_3 SQL + \beta_4 R + \beta_5 Spark + \beta_6 SAS + \beta_7 AWS + \beta_8 Excel + \beta_9 Azure + \beta_{10} Java + \beta_{11} Tableau + \beta_{12} MachineLearning + \beta_{13} AI + \beta_{14} Pandas + \beta_{15} Scipy + \beta_{16} Perl + \beta_{17} TextMining + \beta_{18} Matlab + \beta_{19} C + \beta_{20} SPSS + \beta_{21} Hive + \beta_{22} Splunk + \beta_{23} DataMining + \beta_{24} UnixLinux + \beta_{25} Apache$$

We choose whether the job required PhD degree as the outcome variable since we are interested in higher level data science jobs that require advanced degrees, and we want to see whether jobs that require certain skill sets are more likely to also require advanced degrees. And we choose to fit logistic regression model since the outcome variable is binary. Table 2 in appendix shows the logistic regression model output. R, Spark, and AI are the only skills that show statistical significance. We estimated the odds of a job to require PhD degree is increased by a factor of exp(1.366)=3.92 if the job requires skill in R as compared with one doesn't require skill in R (95% CI: 1.605, 9.570). We estimated the odds of a job to require PhD degree is increased by a factor of exp(3.25688)=25.97 if the job requires skill in AI as compared with one doesn't require skill in AI (95% CI: 3.87332,174.10316). And the odds of a job to require PhD degree is decreased by a factor of exp(-1.77008)=0.1703194 if the job requires skill in Spark as compared with one doesn't require skill in Spark (95% CI: 0.03517398 0.82472012). In general, jobs that require skills in R and AI are more likely to also require advanced degree, and jobs that require skills in Spark are less likely to require advanced degree. However, since our data science job sample may not be representative of the data science jobs in general, the validity of the result need to be confirmed by

further study.The 10-fold cross-validation AUC of our logistic regression model to predict whether a job require PhD degree is 0.6342, which is not very high.

2.Discussion

Based on our analysis, Python, Machine learning, SQL,R,and Java are the top 5 skills required by data science related jobs.Based on our two-sample t-test results, Machine learning， Python， Hadoop， AWS， Spark,Java,and R skills show significant differences between data science related jobs and traditional statistic jobs. Their significance remain even after Bonferroni correction.And the robustness of results are confirmed using permutation tests. Among them, Hadoop and Python skills show the most differences between data science jobs and traditional statistic jobs , therefore, we would say they are the most unique skills that employers need for data science related jobs as compared with other more traditional statistics jobs. Based on Figure 3, we would conclude that information technology seems to employ the most data scientists. Our analysis has several limitations. First of all, when exploring the industry that employ the most data scientists, our sample may not be representative of the true job market. Our initial examination showed that Information Technology seems to employ the most data scientists, however, it could be the case that since the majority of job listings on StackOverflow are from tech company, thus we might oversample jobs from this industry and under sample jobs from other industries. For future analysis, we could get a random sample of job listings from major companies within each industry, compute the proportion of data science related jobs among all jobs, and compare across industries using Chi-square test. In that way, we can draw more valid inference on this research question. Second of all, the job listings we obtained in our analysis may not be representative of the data science job listings in general. Although many jobs have the title "data scientists", however,they could mean completely different roles, some of them may actually towards traditional statistic roles. If that is the case, our comparison between data science jobs and traditional statistic jobs may be invalid. For future analysis, we should conduct additional screening on data science job listings, exclude those whose roles are towards traditional statistics jobs.

References

[1]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 [2]Hadley Wickham (2017). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.2.0. https://CRAN.R-project.org/package=stringr

[3]Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr

[4]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

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

[6]John Muschelli (NA). glassdoor: Interface to 'Glassdoor' API. R package version 0.7.6.

[7]McKinsey & Company(2011).Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

[8]Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2016). maps: Draw Geographical Maps. R package version 3.1.1. https://CRAN.R-project.org/package=maps

[9] Richard Cotton (2017). rebus: Build Regular Expressions in a Human Readable Way. R package version 0.1-3. https://CRAN.R-project.org/package=rebus

[10]Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/

[11]Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.15.1.

Appendix

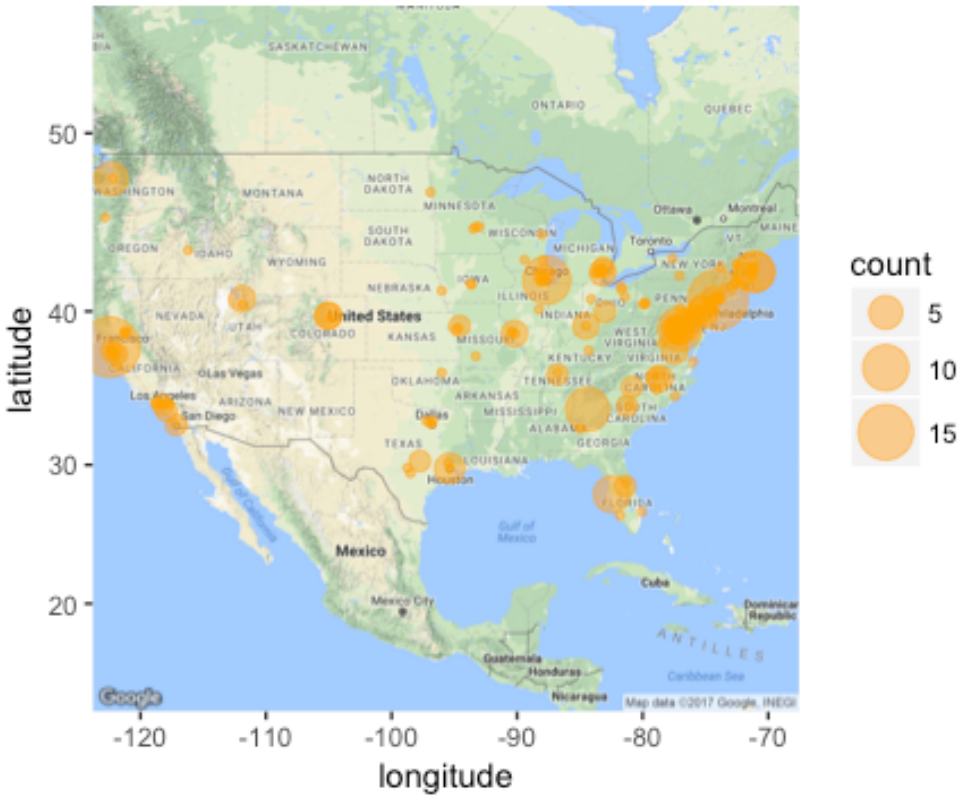
Industry cleaning

The industry categories were not listed in a consistent way and there were more than 300 levels for industry categories. Some example industry categories before processing including:

Information Technology, Engineering, Professional Services; General Business, Health Care, Information Technology;Automotive, Other;Health Care, Information Technology, Other;Finance, Information Technology, Insurance;Other Great Industries;Retail, etc.

We did some text manipulation to group similar industry categories together. We firstly removed all the special characters in industry categories and converted them into lower cases. Then we searched for exact matches to the keyword lists we created for industry grouping. If there were 'care','biotech',"lifescience","pharm" detected in the category listings, then they were reassigned the "Health Care" as their industry category listing; if there were "education","academia","academic","research" detected in the category listings, then they were reassigned the "Education" as the industry category listing; if there were 'tech','software','cloud','internet','cyber','it','web','saas','speech','computing','artificial intelligence',"fraud" detected, then they were reassigned the "Information Technology" as the industry category listing... There were more than 60 companies having industry category listed as "Other Great Industries ", "Data Analysis", "Science", which provided little information in grouping them into the right industry category. For those companies, We scraped their industry category from Glassdoor's API.Specifically, we sent company search request to Glassdoor's API and put the names of company with unknown industry as the action parameter. After grouping, there were 14 industry categories, including Business Services, Health Care, Retail,Finance,Information Technology,Manufacturing, Education,Aerospace & Defense,Oil, Gas, Energy & Utilities, Media, Telecommunications,Tourism,Government,and Agriculture & Forestry.

Figure 1.(a) Geospatial Distribution of Data Science Jobs



(b).# of Data Science Jobs by City (Top 10)

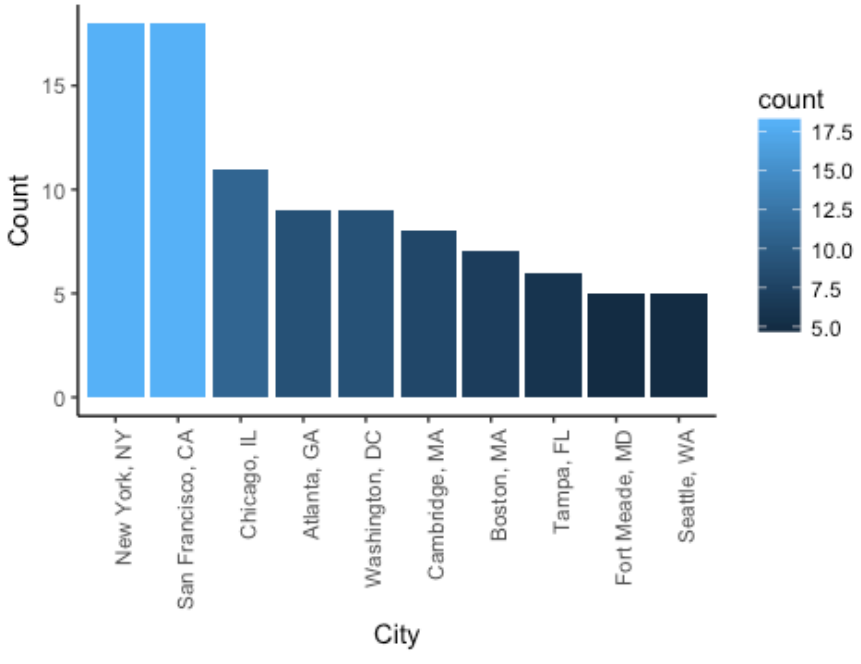


Table 2. Logistic Regression Model to Study the Association between Degree Requirements and Skill Requirements

Coefficients	Estimate	Std. Error	P-value	exp(coef) 95% CI:
R	1.36575	0.45553	0.002716 **	3.92 (1.605,9.570)
Spark	-1.77008	0.80478	0.027846 *	0.17 (0.035,0.825)
AI	3.25688	0.97080	0.000794 ***	25.97 (3.873,174.104)
Java	-0.92459	0.54147	0.087716 .	0.4 (0.137,1.146)
MachineLearning	0.73716	0.42366	0.081864 .	2.09 (0.911,4.795)
Splunk	3.59989	1.88382	0.056010 .	36.59 (0.912,1468.755)