Tanner Sax

Capstone Project – Milestone Report

Introduction

Technology companies have been appearing all over the nation. They have already made huge advances in industries across the board and continue to do so. More technology is being utilized than ever before, and with this brings more data. As these companies grow, there will be more demand for employees with certain skillsets to navigate this data. There are many exciting opportunities in this field for recent graduates with analytical and problem-solving skills. Harvard Business Review coined the Data Science profession as “The sexiest job of the 21st century.”[[1]](#footnote-1) This is good news for those with an interest in technology, however it is important to take into consideration the experience and skillset required. An article released by Northeastern states that “Data scientists—who have typically earned a graduate degree, boast an advanced skill set, and are often more experienced—are considered more senior than data analysts.”[[2]](#footnote-2) For a recent graduate, it is more likely to land a role as a data analyst before progressing into a data scientist role.

This analysis discovers the most preferred programming language skills and degree field for job seekers in the northeast searching for a data analyst role. According to Business Insider, among the eleven most high-tech cities in the United States are New York, Boston, Washington D.C., and Philadelphia.[[3]](#footnote-3) Analytics India Magazines list of top programming languages was used to select eight popular languages.[[4]](#footnote-4)

A web scrape was performed on Glassdoor.com to collect information for Data Analyst job postings from these four cities.

Data

Glassdoor.com is one of the largest job listing websites in the world, and new jobs are posted daily. Job seekers can view information such as the job title, company, reviews, salary estimates, and job descriptions that contain qualifications and responsibilities.

The web scrape collects data from the first thirty pages for each city from the Glassdoor.com. This is because only the first thirty pages are viewable on the website. This lessens the sample size; however, 3,544 postings were still able to be collected. An initial analysis idea included the salary estimates for jobs. Since this data was unable to be collected, the analysis had to be changed accordingly.

A search for ‘Data Analyst’ was performed for each of four major tech cities in the northeast – including New York, Boston, Washington D.C., and Philadelphia. The data consists of the job title, company name, location, and a full job description from the Glassdoor job postings. The web scrape collected the data for each city separately and stored it in a data frame. There were several job description links that were unable to be scraped due to ads. These values were set to return NA in the data frames. The rows containing NA in the job description column were then removed using the complete.case() function.

*New\_York <- New\_York[complete.cases(New\_York), ]*

Since there were only 10 NA’s total in 3,555 results, removing them did not have a large impact on the sample.

The most important aspect of the data is the job description column. This is because the analysis requires keyword extraction to identify the most common programming language skills and degree fields. Separate columns were created that identify keyword matches through binary variables. This was done through the mutate function for each of the keywords.

*New\_York <- New\_York %>%*

*mutate(python = ifelse(grepl("python", job\_description, ignore.case = TRUE), 1, 0))*

The grepl function searches for a match and returns a logical vector. If it returns TRUE, then according to the ifelse statement, a ‘1’ will be returned. Otherwise, a ‘0’ will be returned.

The means of these columns then represent the percentage of job postings that mention the keyword. Then two separate tidy data frames were created for the language skills and degree fields with these values. To have a tidy data frame, vectors were created for the cities and languages/fields that placed them in the correct order with the means. The vectors were then combined into a tibble data frame.

*language\_df <- tibble(cities1, language, Percent\_of\_jobs\_language)*

*degree\_field\_df <- tibble(cities2, degree\_field, Percent\_of\_jobs\_degree)*

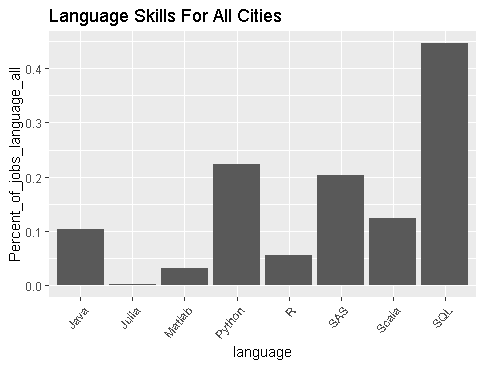
The first data frame includes the city, programming language, and percent of job postings that mention the programming language. The second data frame contains the city, degree field, and percent of job postings that mention the degree field.

Next, a data frame was created that includes all the data from all four cities. Similar keyword extraction methods were utilized to analyze the interactions of top three programming languages and the top four degree fields. Each of the top programming languages were paired with one of the top four degree fields. The job description had to contain both the programming language and the degree field to return a ‘1’. Then the means of these combinations were put into a vector. Each combination name was put into a vector as well. Then, a data frame was created of the combination names and the means.

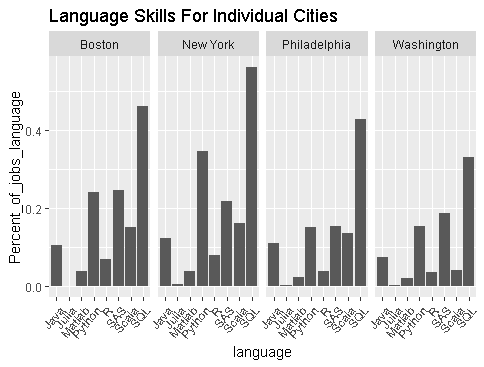
top\_df <- tibble(top\_names, Percent\_of\_jobs\_top)

Analysis

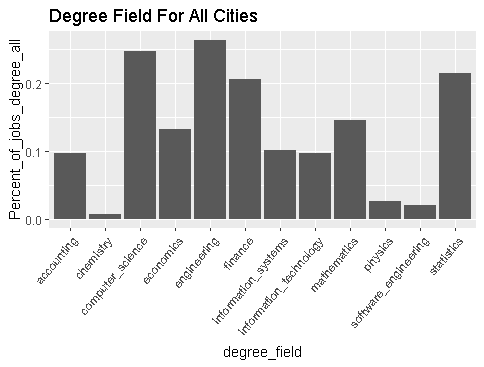
After the data was cleaned and the new appropriate data frames created, several graphs were created with the ggplot2 package. According to the graph below, SQL is the most important programming language for data analysts in all four cities, appearing in 44.6 percent of job postings. This is to be expected for a data analyst job. Julia, a very new programming language, is mentioned the least amount of times at 0.28 percent, as expected.



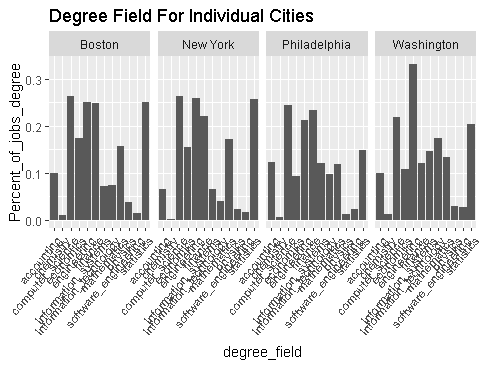
In New York, Boston, Washington D.C., and Philadelphia, SQL is mentioned in 56.22, 46.07, 33.18, and 42.91 percent of job postings, respectively. In New York, Boston, Washington D.C., and Philadelphia, Julia is mentioned in .54, 0, .23, and .34 percent of job postings, respectively. Python appears to be more important in New York than in other cities, appearing in 34.66 percent of job postings there. Scala appears to be less important in Washington D.C. than in other cities, appearing in only 4.16 percent of job postings.



Computer science, engineering, finance, and statistics appear to be among the most popular degree fields. Engineering appears in 26.3 percent of job postings.

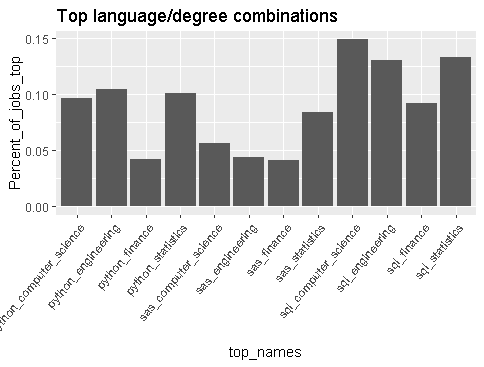


Information technology degrees appear more in the job postings in Washington D.C. than in other cities.



For data analyst jobs in the northeast the most preferred programming languages are SQL, Python, and SAS, and the preferred degree fields are computer science, engineering, finance, and statistics. It is more important to know Python for jobs in New York City than for jobs in other cities. It is less important to know Scala for jobs in Washington D.C. than in other cities.

A further analysis compares each of these top majors with the top programming languages. This suggests which combination of programming language and degree field is the most advantageous for a data analyst job. The combination of SQL and computer science is the most common. In 14.84 percent of job postings, both the keywords ‘SQL’ and ‘computer science’ appeared.



1. DavenportD.J, Thomas H., et al. “Data Scientist: The Sexiest Job of the 21st Century.” *Harvard Business Review*, 26 May 2017, hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century. [↑](#footnote-ref-1)
2. “Data Scientist vs. Data Analyst: What's the Difference?” *Cyberbullying in the Workplace*, Kevin Carvalho Https://Www.northeastern.edu/Iuhrp/Wp-Content/Uploads/2016/09/Northeastern-Iuhrp.png, 27 Nov. 2018, www.northeastern.edu/graduate/blog/data-scientist-vs-data-analyst/. [↑](#footnote-ref-2)
3. Weller, Chris. “The 11 Most High-Tech Cities in the US.” *Business Insider*, Business Insider, 22 Aug. 2017, www.businessinsider.com/most-high-tech-cities-in-the-us-2017-8. [↑](#footnote-ref-3)
4. Deoras, Srishti. “Top 10 Programming Languages For Data Scientists to Learn In 2018.” *Analytics India Magazine*, 1 Jan. 2019, www.analyticsindiamag.com/top-10-programming-languages-data-scientists-learn-2018/. [↑](#footnote-ref-4)