Final project

Chih-Kai (Kyle) Chang

10/1/2017

Project choice: Option 3

Perform an analysis of "data scientist" jobs listed on job boards and on the employment pages of major companies. * What are the most common skills that employers look for? * What are the most unique skills that employers look for? * Where are the types of companies that employ the most data scientists?

Project Plan

1. Scrape data from job postings website such as **Indeed.com**, **Glassdoor.com** with search key word *Data scientist*.

- Use `rvest` package, and understand the pattern of URLs to scrape job descripiton.
- Sort search results and save them into .csv file.

2. Prcess data

- Create a dataframe that have title, company, industry, city, state, and the link of that job descripiton.
- Scrape requirement skills information from the job link.
- Get coordinate informaiton for U.S. state. (Define those skills that are commonly required or essential for data scientist jobs such as programming skill) (Define what type of programming skill: Jave, Python ect.)

3. Explratory analysis

- Find the most related skills identify common and important elements for different website by using barchart.
- Get a roguhly idea about how these data scientist jobs distribute around U.S. (ggmap)

4. Perform an proper statistical analysis and put analysis result into report.

Introduction

As technology advances and internet platform gets common nowadays, we generate a huge amount of data every day. Data scientists were not quite often on the radars a decade ago, but in recent year their popularity shows that businesses now take data seriously. There is no doubt the importance of data scientist will become significantly evident.

We live in a data-driven world, and there's no turning back. Data science is widely used in business. We might think that browsing and purchase history might not important. However, those are like cruel oil for lots of business. Data in the 21st century is like oil in the 18th century, a business report said. Amazon recommendations, Facebook page campaigns, and Netflix suggestions are all powered by data science. Most of the data sets are now so large and complex that we need tools and approaches to exploit the most of them. According to global management consulting firm McKinsey & Company, there would be 4 million big data-related jobs in the U.S, and a shortage of 140,000 to 190,000 data scientists.

As we can see from lots of business report, data scientists are in high demand and basically essential for many industries. Therefore, understanding the requirement skills for data scientist jobs is important. In this project, I performed an analysis for *data scientist jobs* listed on job posting website to identify the required skills that employers look for and look for where the types of companies that employ the most data scientists.

Data collection

Selecting job posting websites

I collected data from **Glassdoor.com**, **indeed.com**. There are a few reasons that I select these two website. First, Glassdoor as well as Indeed has been widely used and has lots of users. Besides, it is not hard for us to see their advertisement and commercial, so that those two job posting websites gain good publicity not only for their users but for employers. There is no doubt that those two websites are very popular and well-known.

Second, the URLs in the website must have certain pattern and in clean manner. Take URLs in Indeed.com as an example.(`https://www.indeed.com/jobs?q=data+scientist&l=baltimore&radius=25&start=100`)

`q=data+scientist&l=baltimore&radius=25` means that I want to search for **data scientist** jobs in **Baltimore** city within **25 miles**. In addition, `start=100` can bring us to the result of number **100 th** job posting. By changing those parameter, we can easily scrape all the data on that website. To sum up, URLs on Indeed.com and Glassdoor.com have certain pattern for us to modify so that those URLs can easily direct us to the page we want.

Scraping information from website

I used `rvest` package in R and `SelectorGadget` extension in Google Chrome to scrape information from webpage. In addition, I also checked CSS script to find certain pattern and get the information I need. My steps are as following.

- Got all URLs in that job posting website
- Used Selector Gadget or read CSS script and source code to get the nodes we want
- Copied CSS selectors, and Xpath into `html_nodes()` to present a structured format
- wrote a function and used `stringr` to clean text
- Create a data frame which contain job title, company, industry, city, state, and the link (URL) for that job.

Scraping required skills for data scientist

I wrote a function `ScrSkill` to scrape the requirement for data scientist. By inputting a clean format which I created previously, that function can read the job link and scrape certain key words that I am interested in. The steps are as following.

- Created a list of the most common and widely used skills for data scientist. That list includes statistics, computer science, math, machine learning, data mining, predictive modeling, R, Java, Python, SQL, SAS, Tableau, Excel, C++, Hadoop, Stata, Spark, matlab. (Used `\\b` before and after these words to make sure they are perfectly match)
- Wrote a for loop to run all job links URLs, and put URL into `html_text` to read the text in whole webpage, and then cleaned the text by using `stringr` to clean text. For example, in glassdoor, I used `readLines` function to get sourcecode, and found specific pattern for industry.
- Using `sapply` and `grep` function, I output `TRUE` for skills found in the text and `FALSE` otherwise. Finally, I created a data frame to save them.

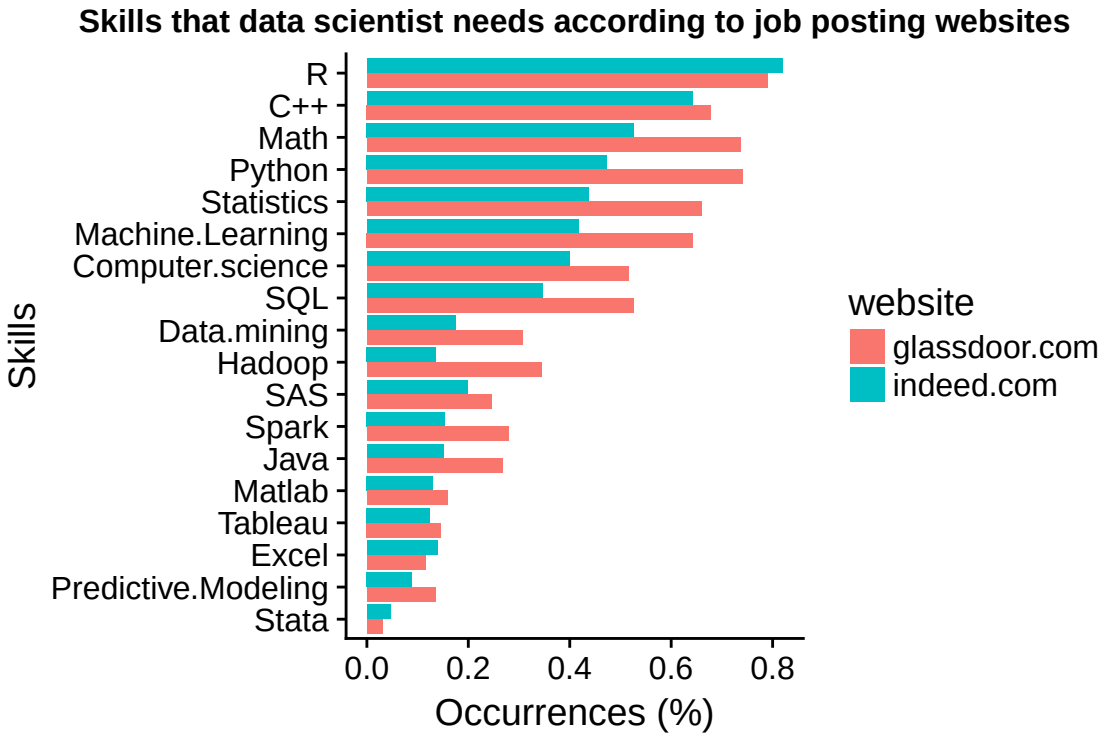
Data cleaning

Since job posting website might post the same job multiple time, I excluded the duplicate data. As for missing value `NA`, they might provide some useful information when I do exploratory analysis so I keep them for now. For example, some data might lose company location, but it does have skill requirement information. That would be helpful when I want to investigate required skills for data scientist.

Exploratory analysis

Compared required skills for data scientist in *Glassdoor.com* and *Indeed.com*

In order to get a rough idea of the occurrence for certain type of skills in these two websites, I created a bar chart. The sample size in Glassdoor.com is 616, while that in Indeed.com is 338. The result is as following.



I listed top 5 tag skills in each website. The top five required skills from the first to fifth in *Glassdoor* are `R`, `Python`, `Math`, `C++`, `Statistics` respectively, while in *Indeed* are `R`, `C++`, `Math`, `Python`, `Statistics`. `R` has the highest occurrence rate in both websites with 0.79(%) in *Glassdoor* and 0.82(%) in *Indeed*. Furthermore, `R` is followed by `Python` in *Glassdoor* with 0.74(%) occurrence rate and followed by `C++` in *indeed* with 0.64(%) occurrence rate. We can know the rest statistics from the table below.

Glassdoor	Occurrences (%)	Indeed	Occurrences (%)
R	0.7905844	R	0.8195266
Python	0.7402597	C++	0.6420118
Math	0.7370130	Math	0.5266272
C++	0.6785714	Python	0.4733728
Statistics	0.6590909	Statistics	0.4378698

- **Perform a proportional test**

The top five skills in both job posting websites are the same, but the order is somewhat different. In order to investigate whether the occurrence rate in these two website is different, I performed a proportion test. The result are as following.

Skills	p-value
R	0.2843
Python	1.675e-16
Math	4.936e-11
C++	7.374e-58
Statistics	3.516e-11

As we can see from the table, there is no significant difference for programming skills in `R`. However, other skills are very different from these two websites with very small p-value.

Therefore, I might conclude that `R` is the most common skills that employers look for. `R` has the highest occurrence rate in both website and the occurrence rate does not have a significant difference between *Glassdoor&* and *Indeed*. Other skills that employers look for are different form website to website.

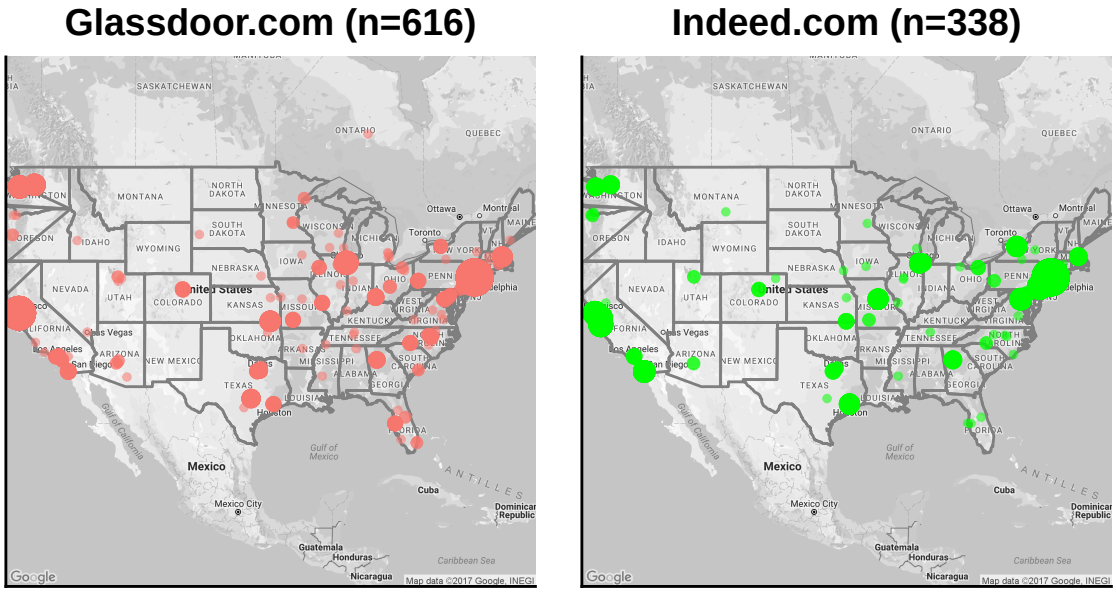
Investigated job postings location in *Glassdoor.com* and *Indeed.com*

I used `ggmap` package to get city's coordinate information and created a table.

- **Looking into state**

Glassdoor	Count	Indeed	Count
CA	160	CA	91
NY	91	NY	58
WA	32	IL	25
TX	31	PA	22
IL	30	VA	22

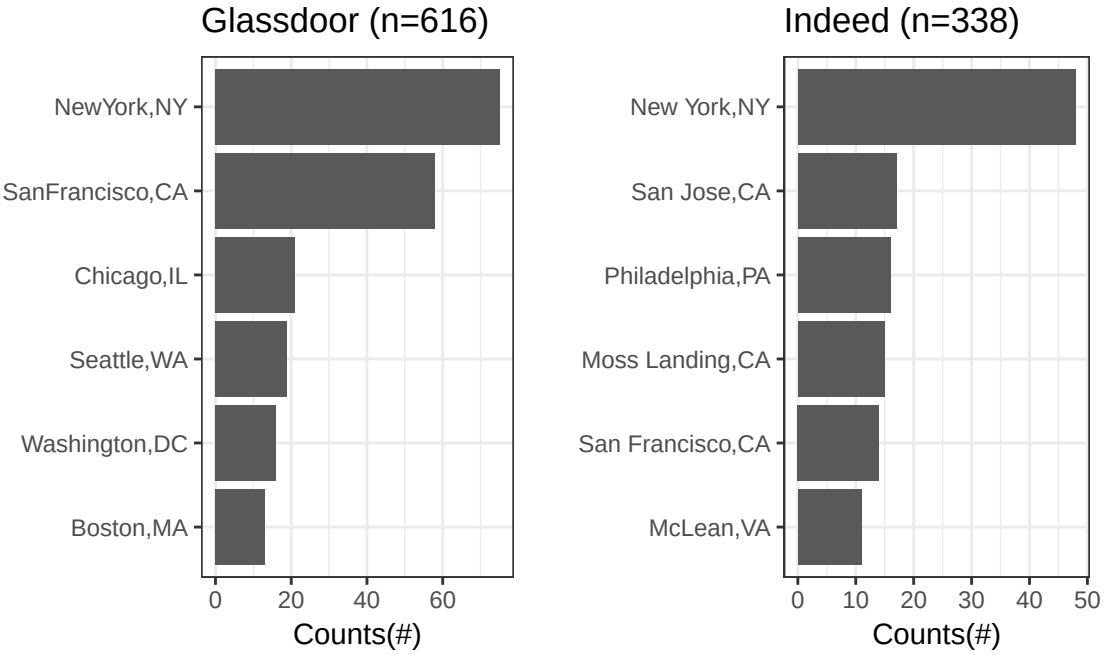
As we can see from the table above, `California (CA)` state has the most data scientists job postings in both websites with 160 in Glassdoor and 91 in Indeed. In addition, I plotted the job postings distribution around United states in both websites, so that we can get a rough idea about how these data distribute.



From those two plots, we can know that there are bunch of job postings around coast area in United States, **New York (NY)** state and **California (CA)** state especially.

- **Looking into city**

Furthermore, I investigated job postings in **city** instated of state, creating bar chart.



I found that `New York, NY` city has the most data scientist job postings in both *Glassdoor.com* and *Indeed.com*. That is followed by `San Francisco, CA` city in *Glassdoor.com* and `San Jose, CA` city in *Indeed.com*. By simply investigating cities in United States, there is no doubt that `New York, NY` city needs data scientist jobs the most.

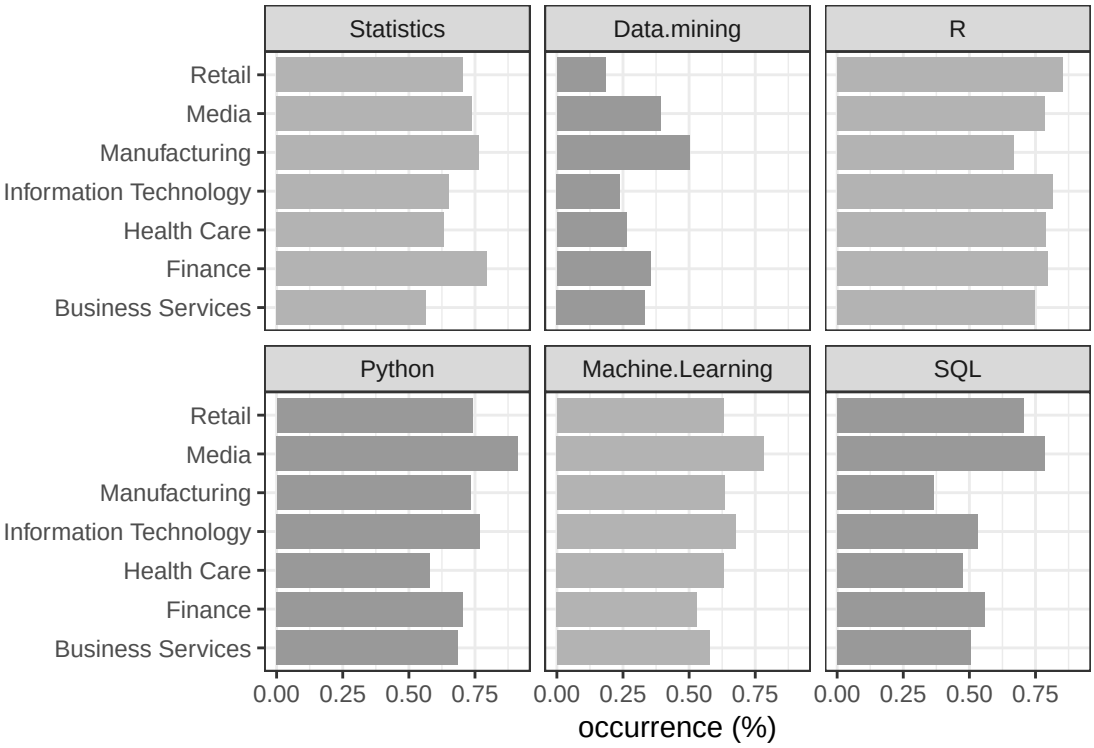
In conclusion, if we look into **state**, `California (CA)` state has the most data scientist job posting. If we look into certain **city**, `New York, NY` city needs data scientists the most.

Top 5 industries with the most job positngs on *Glassdoor.com*

According to *Glassdoor.com* with sample size 616, `information technology` industry employs the most data scientists with 222 postings, followed by `business service` industry with 99 postings, and `finance` industry with 34 postings.

Industry (n=616)	count
Information Technology	222
Business Services	99
Finance	34
Manufacturing	30
Retail	27
Media	23
Health Care	19

In conclusion, `information technology` industry needs data scientist the most, and then my next step is to investigate what certain type of skills are needed in each industry.



Top Three skills that each industry needed:

- Information technology
  - R (82%), Python (77%), Machine Learning (68%)
- Business service
  - R (74%), Python (69%), Machine Learning (58%)
- Finance
  - R (79%), Statistics (79%), Python (70%)
- Health care
  - R (79%), Machine Learning (63%), Statistics (63%)
- Manufacturing
  - Statistics (77%), Python (73%), R (66%)

Conclusion

The most common skills that employers look for is `R` programming skill. Of 616 data scientist jobs listed in *Glassdoor*, occurrence rate of `R` is 79%. Of 313 jobs posting listed in *Indeed*, occurrence rate of `R` is 82%.

Most data science jobs are near coast area in United States. `California(CA)` state has the most data science jobs. While we focus on single city, `New York, NY` has the most data science jobs.

Last, `information technology` industry has the most data scientist job posts with 222 out of 616 (36%). Besides, in information technology industry, `R (82%)`, `Python (77%)`, and `machine learning (68%)` are the top three required skills respectively.

Appendex