Comparison of Data Science Job Postings in Two Different Websites.

Chih-Kai (Kyle) Chang

0.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[1].

1.General approach

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, we performed an analysis for *data scientist jobs* listed on two different job posting websites Glassdoor.com,and Indeed.com to identify the required skills that employers look for and look for where the types of companies that employ the most data scientists. We can categorize our approach into these parts : Data collection, Data cleaning, Exploratory (Statistical) Analysis, Result, Limitation.

2.Data collection

2.1 Selecting job posting websites

I collected data from **Glassdoor.com**, **indeed.com** by inputting **data scientist** as key word (This might cause some biases I will discuss in `5.Discussion and Limitation` section.)

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.

2.2 Scraping information from website

I collected data on September 26th and 27th and created a data set with 616 observations in *Glassdoor.com* [0] and 338 observations in *Indeed.com* [0] of these variables title, company, industry, location. I used `R.studio` to perform this analysis.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 `rvest` package in R [2] and `SelectorGadget` extension in Google Chrome [3] to scrape information from webpage or read CSS script and source code to get the nodes we want
- Copied CSS selectors, and Xpath into `html_nodes()`, a function to extract pieces out of HTML documents using XPath and css selectors in `rvest` package.[2] (See Supplemental code: ScrappingData: Chunk2)
- Cleaned data set by using `stringr` package [4] and writing a R function by myself to clean text (See Supplemental code: ScrappingData: Chunk 3,4)
- Create a data frame which contain job title, company, industry, city, state, and the link (URL) for that job.Output is in the Data folder named *glassdoor_datascientist_Info.csv* with observations 616 , and *indeed_datascientist_Info.csv* with observations 338.

2.3 Scraping required skills for data scientist

I wrote a function `ScrSkill` (See Supplemental code: ScrappingData: Chunk 4,5) to scrape the requirement for data scientist. By inputting a clean format which I created previously (*glassdoor_datascientist_Info.csv*, and *indeed_datascientist_Info.csv*) , `ScrSkill` 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` in `rvest` package [2] to read the text in whole webpage, and then cleaned the text by using `stringr`[4] to clean text. For example, in glassdoor, I got sourcecode by using `readLines` function, and found specific pattern for industry. (See Supplemental code: ScrappingData: Chunk 4)
- I output `TRUE` for skills found in the text and `FALSE` otherwise by using `sapply` and `grep` function, Finally, I created a data frame to save them. (See Supplemental code: ScrappingData: Chunk 4)

2.4 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.

3.Exploratory analysis and t-test

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

In order to get a rough idea of the occurrence rate 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.

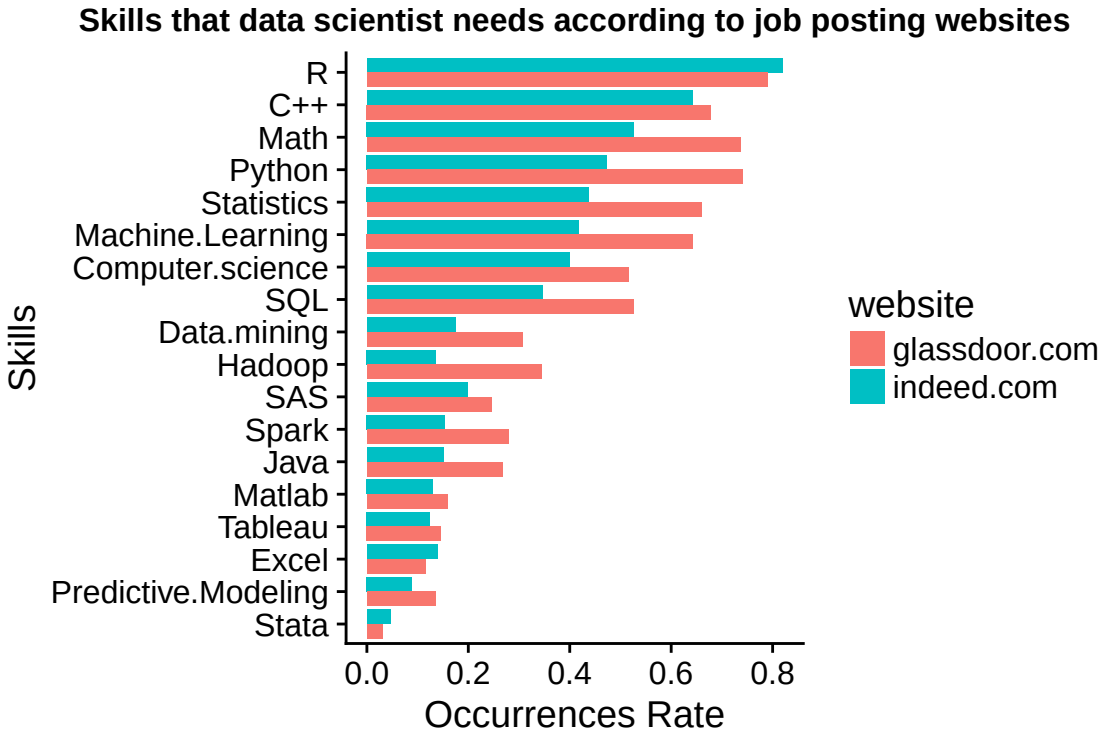


Figure 1. A bar chart lists those skills that data scientist should have, according to job posting websites. *`R` has the highest occurrence rate in both websites.*

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 Rate	Indeed	Occurrences Rate
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 2-sample t-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 t-test (See Supplemental code: Final_Report: Chunk 3) with sample size 616 in *Glassdoor* and 338 in *Indeed*. The result are as following.

Null hypothesis: the proportions of certain skills in glassdoor and indeed are the same

Alternative hypothesis: the proportions of certain skills in glassdoor and indeed are different

Skills	Difference	95% CI	p-value
R	0.0289	-0.081 , 0.0231	0.2843
Python	0.2669	0.2034 , 0.3304	1.675e-16
Math	0.2104	0.1468 , 0.274	4.936e-11

Skills	Difference	95% CI	p-value
C++	0.0366	-0.0265 , 0.0996	0.2523
Statistics	0.2212	0.1564 , 0.286	3.516e-11

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

Therefore, I conclude that `R` is the most common skills that employers look for, since `R` has the highest occurrence rate in both websites and we do not have supportive evidence to say that the requirement of programming skill in `R` is different between *Glassdoor.com* and *Indeed.com*.

In other words, `R` has the highest occurrence rate in both website and the occurrence rate does not have a significant difference between *Glassdoor* and *Indeed*. However, other skills, expect for `C++`, that employers look for might be different form website to website. Because each job website has its preference and users, it might cause sampling bias. Therefore, we have to think carefully when we scrape data form job posting website.

3.2 Unique skill for data scientist

In the part, I want to find unique skill for *"data scientist"* (There are some limitation. See **`5.Discussion and Limitation`** section)

First, I separated `Indeed.csv` data set with observations 338 and 22 variables (See Data folder) into two part . The first part is job title with exact words **data scientist** (n=254), and the other is job title which does not contain *data scientist* (n=94) such as quantitative analyst, data engineer... etc. Second, I calculated their occurrence rate differences and performed a t-test (See Supplemental code: Final_Report: Chunk 4). Last, I listed top three occurrence rate differences skills `Python`, `SQL`, and `R` respectively, and tried to find that whether there is a significant difference between "data scientist" and other job title.

*Null hypothesis: the proportions of certain skills are the same for job title with exact words **data scientist** and others job title related to data scientist*

*Alternative hypothesis: the proportions of certain skills are different for job title with exact words **data scientist** and others job title related to data scientist*

	Occr.diff	CI	p.value
R	0.155886764154481	-Inf , -0.0651	0.00261405837207071
SQL	0.175478065241845	-Inf , -0.0854	0.00076944027018411
Python	0.202193475815523	-Inf , -0.1027	0.00048852793208354

From the table above, I conclude that `Python` is the unique skill for data scientist, since it has highest difference in occurrence rate with very small p-value so that we have supportive evidence to say that the requirement of `Python` are different for job title with exact words *data scientist* and others job title related to data scientist. However, there is a limitation as I mentioned. I will discuss it in the `5.Discussion and Limitation` section.

To sum up, programming skills related to data analysis such as `Python`, `R`, and `SQL` are very unique and required for data scientist.

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

I created two data sets`loc_G`(n=616) for Glassdoor.com and `loc_I`(n=338) for Indeed.com by ising `ggmap` package[5] (See Supplemental code: Final_Report: Chunk 5) . Each of them contains these variables city

name, coordinate information, and number of job postings. 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.

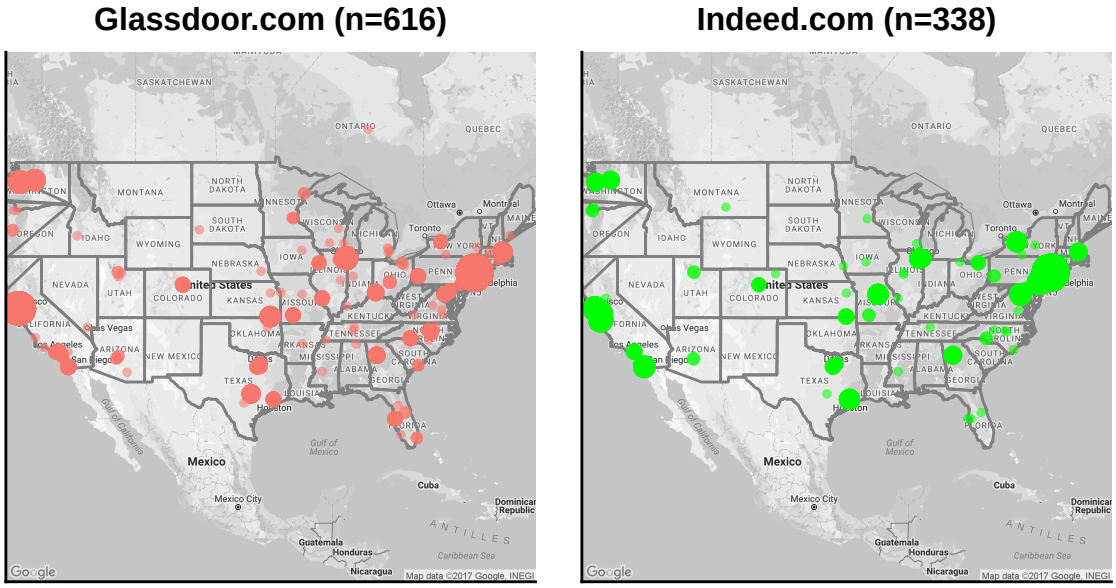


Figure 2. A map showing the distribution of data scientist job posting around U.S. in two different job posting websites. *There are bunch of job postings around coast area in United States,* ***New York (NY)*** *state and* ***California (CA)*** *state especially.*

- **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. Again, I performed 2-sample test for equality of proportions in California state, and found out that there is no significant difference between these two website with p-value = 0.7632. Furthermore, I performed binominal t-test for the job postings proportion in California state is equal to 0.25 , and we do not have significant evidence (p-value in Glassdoor= 0.3714, in Indeed= 0.3097) to say that the proportion of data scientist jobs in California (CA) is not 25%. Therefore, I concluded that one fourth of data scientist jobs are located in `California (CA)`.

- **Looking into city**

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

Glassdoor	count	Indeed	count
NewYork,NY	75	New York,NY	48
SanFrancisco,CA	58	San Jose,CA	17
Chicago,IL	21	Philadelphia,PA	16
Seattle,WA	19	Moss Landing,CA	15
Washington,DC	16	San Francisco,CA	14

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.

- **Conclusion for data scientist job postings distribution**

To sum up, 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. However, this might change from time to time ,due to renewal of job posting websites.(See `5.Discussion and Limitation`)

3.4 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%) ## 4.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 338 jobs posting listed in *Indeed*, occurrence rate of `R` is 82%. In addition, the most unique skills that employers look for are `Python`, `R`, and `SQL`, programming skill related to data analysis.

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.

With the same logic, I can easily compare two job posting websites and find important skills for data scientist.

5.Discussion and Limitation

There are some limitations to this analysis. First, since job posting websites would renew every day, our data set cannot be the same all the time and we cannot scrape "all" data science job on the website. Our analysis may be somewhat different, but the main concepts and methods do not change. Second, we defined skill set (keyword tags) by ourselves so that there would be potential for underestimation in some skills that we did not scrape. In addition, as I mentioned in exploratory analysis section, there might be a sampling bias due to the website we used.

For example, in *indeed.com* , when I input **data scientist** as searching job title, the output will contain other job titles such as quantitative analyst, data engineer. . . ect. Besides, some company might use other term as **"data scientist"**, such as quantitative scientist, and quantitative analyst. Therefore, I have not excluded them while doing analysis, since these job are very similar to "data scientist". Because of this, my analysis would have some biases, if we simply want to focus on those jobs that title has exact words *data scientist*.

Other than that, in `3.2 Unique skill for data scientist` section, I want to find unique skill for "data scientist". I separated data into two parts. The first part is those have exact word data scientist, and those who do not. However, some companies might just use different name for data scientist as I mentioned. Therefore, depend on what analysis you want to do, the result might be different.

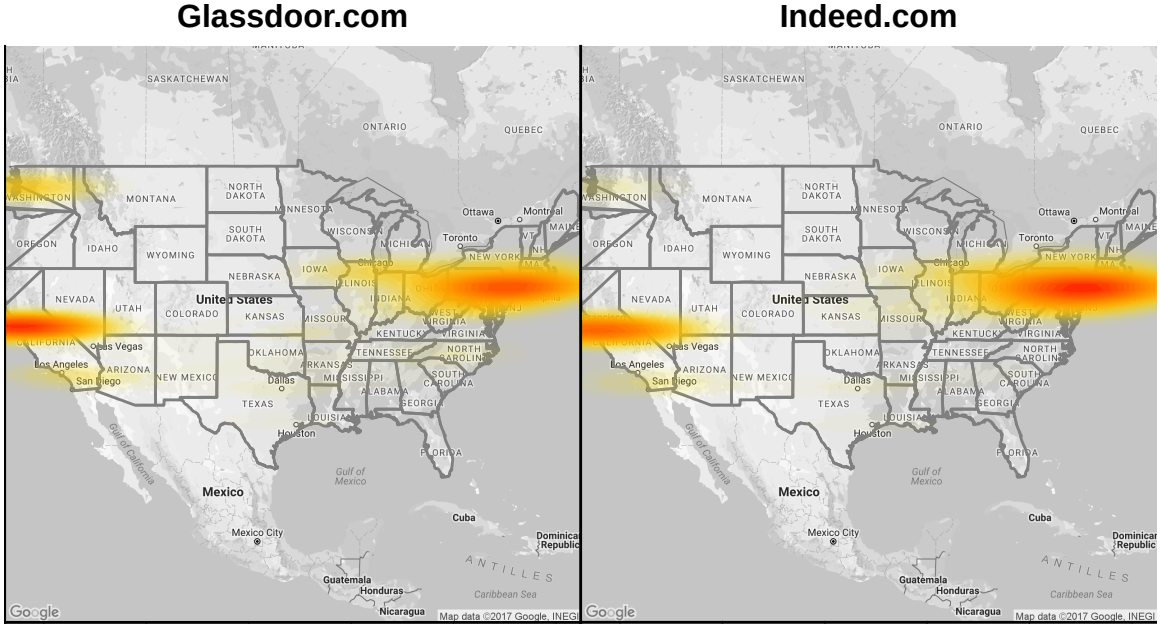
Last, I performed t-test in the analysis. Therefore, I have to make data normality assumption.

References

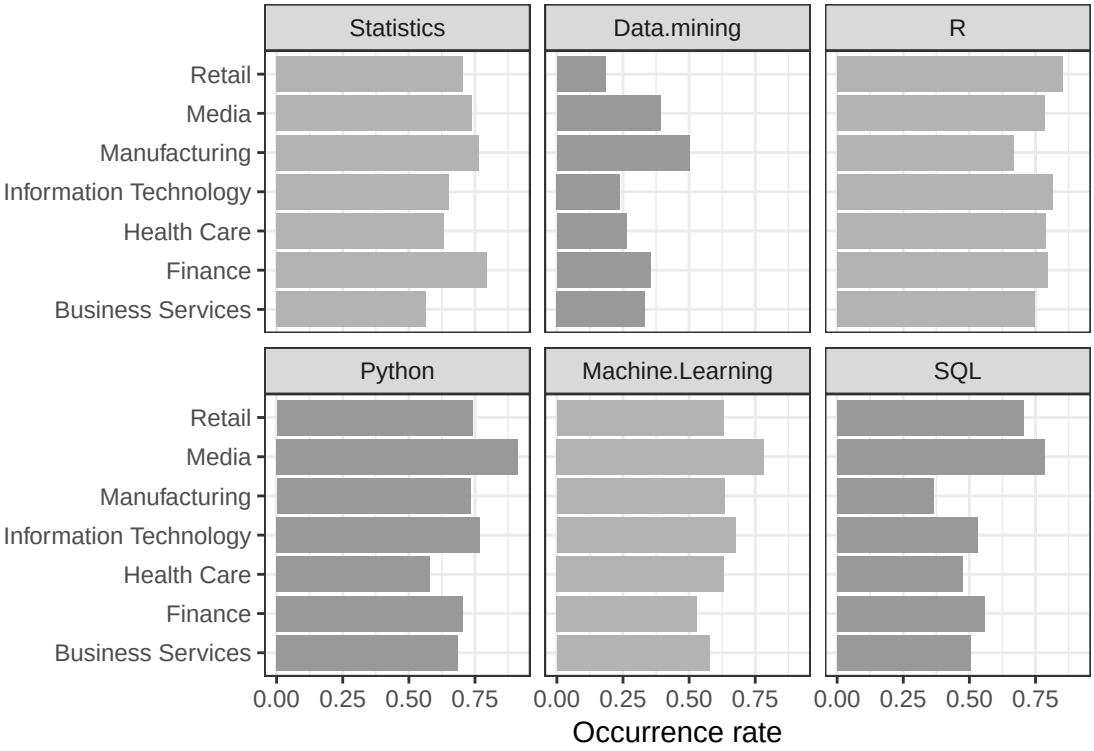
- [0] Job posting website, Glassdoor.com, and Indeed.com URL:https://www.glassdoor.com/, https://www.indeed.com/
- [1] Manyika James et al. (May 2011), "Big data: The next frontier for innovation, competition, and productivity". URL: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
- [2] Hadley Wickham [aut, cre], Easily Harvest (Scrape) Web Pages, URL: https://github.com/benjamin-ackerman/Datascience_project1/blob/master/writeup.pdf
- [3] Google extension, Select gadget, URL: https://chrome.google.com/webstore/detail/selectorgadget/mhjhnkcfbdhnjickkkdbjoemdmbfginb?hl=en
- [4] Hadley Wickham [aut, cre, cph], Simple, Consistent Wrappers for Common String Operations, URL: https://cran.r-project.org/web/packages/stringr/stringr.pdf
- [5] David Kahle [aut, cre], Hadley Wickham [aut], Spatial Visualization with ggplot2, URL: https://github.com/dkahle/ggmap
- [6] Referenced Shannon Wongvibulsin code for install.package

Appendex

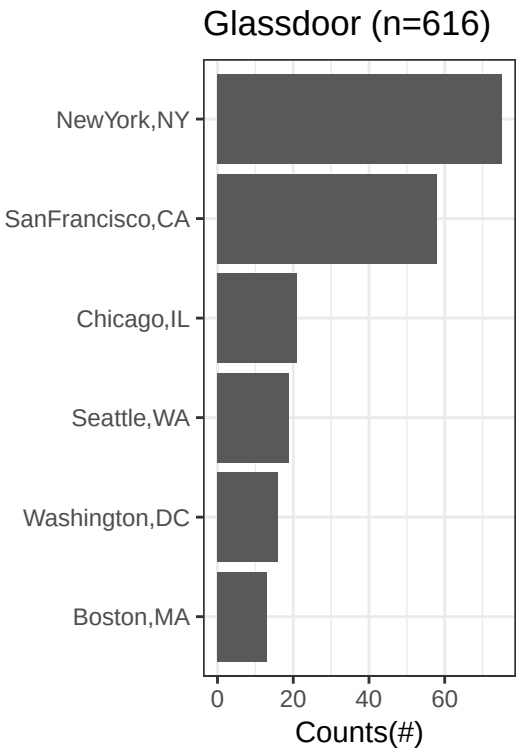
Heat map shows the distribution of data scientist jobs

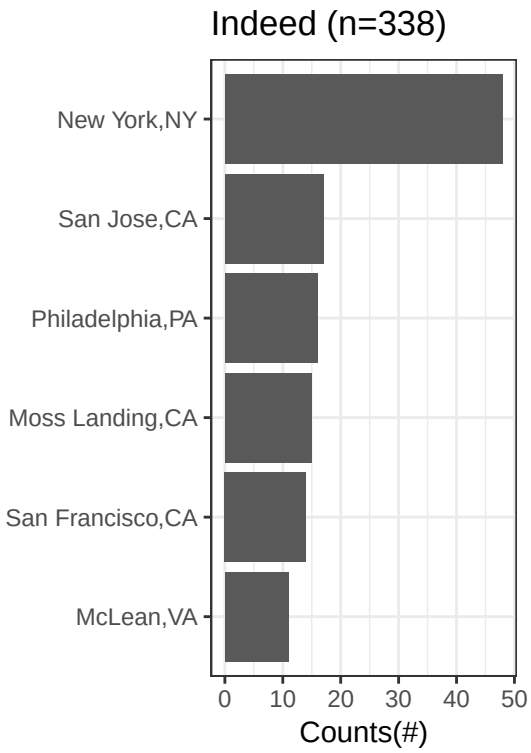


Visualized Skills that industries needed



Visualized distribution of data scientist job postings around U.S.


Glassdoor (n=616)
NewYork,NY
SanFrancisco,CA
Chicago,IL
Seattle,WA
Washington,DC
Boston,MA
0
20
40
60
Counts(#)


Indeed (n=338)
New York,NY
San Jose,CA
Philadelphia,PA
Moss Landing,CA
San Francisco,CA
McLean,VA
0
10
20
30
40
50
Counts(#)