ds_writeup

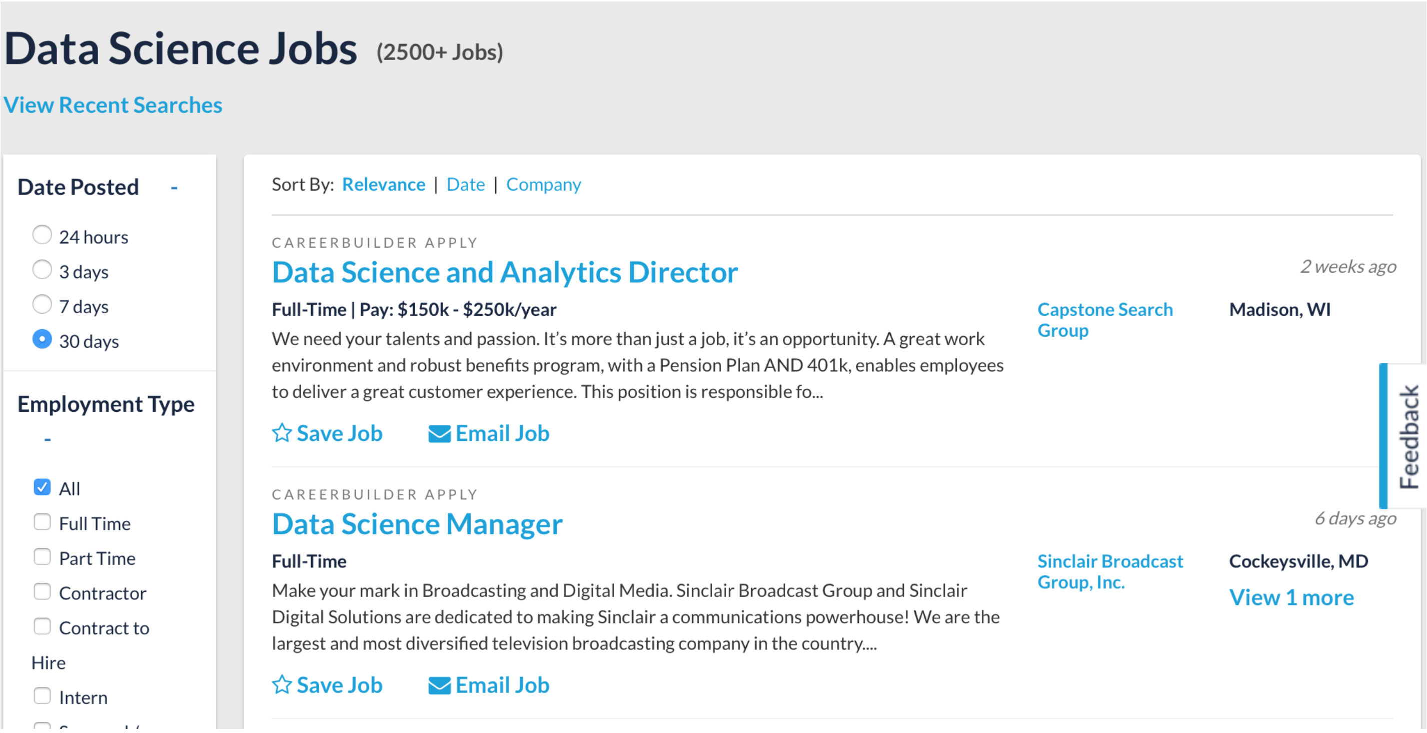
Yeya Zheng

Introduction

This project performs an analysis of “data scientist” jobs listed on some major job boards including careerbuilder and stackoverflow, and on the employment pages of some major pharmaceutical companies, such as Merck,Roche,and Sanofi. 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 jobs, also identify the types of companies that employ the most data scientists.

Data description

I used rvest package to scrape job listings from some major job boards including careerbuilder and stackoverflow, and on the employment pages of some major pharmaceutical companies, such as Merck,Roche,and Sanofi.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. The layout of the website looks like below. After putting “data science” as my searching criteria, the website directs me to the first page of my searching result.The page lists job title, company name, company location and job link to a more detailed job description.



Data Science Jobs (2500+ Jobs)

View Recent Searches

Date Posted -

○ 24 hours
○ 3 days
○ 7 days
● 30 days

Employment Type -

☑ All
□ Full Time
□ Part Time
□ Contractor
□ Contract to Hire
□ Intern

Sort By: Relevance | Date | Company

CAREERBUILDER APPLY

Data Science and Analytics Director

2 weeks ago

Full-Time | Pay: $150k - $250k/year

Capstone Search Group

Madison, WI

We need your talents and passion. It’s more than just a job, it’s an opportunity. A great work environment and robust benefits program, with a Pension Plan AND 401k, enables employees to deliver a great customer experience. This position is responsible fo...

☆ Save Job ✉ Email Job

CAREERBUILDER APPLY

Data Science Manager

6 days ago

Full-Time

Sinclair Broadcast Group, Inc.

Cockeysville, MD

View 1 more

Make your mark in Broadcasting and Digital Media. Sinclair Broadcast Group and Sinclair Digital Solutions are dedicated to making Sinclair a communications powerhouse! We are the largest and most diversified television broadcasting company in the country....

☆ Save Job ✉ Email Job

I put the url “https://www.careerbuilder.com/jobs-data-science (https://www.careerbuilder.com/jobs-data-science)” in the read_html function, and used the selectgadget tool to generate CSS selector for the element I need, including job titles, company name, location, and job links. Then I used html_nodes and html attributes to find the nodes/attributes that match the selector, and used html_text to extract the information. I repeated the same process to scrpe job listings for the next 6 pages. Then I looped through all the individual job link and scraped the corresponding job description as well as a label which described the job type (full time/contract/temporary/part-time) and industry classification. I scraped 175 data science related job from CareerBuilder.I also put “data” as my searching keyword and repeated the same process, I scraped 300 job listings related to data from CareerBuilder.Thus, I scraped 475 job listings from CareerBuilder in total.

I also scraped jobs from Stackoverflow, which 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.I use “data science” and “analyst” as my searching keywords. Using the same procedure I described before, I scrape 323 data science/analyst related jobs.

Since I noticed that CareerBuilder and Stackoverflow have fewer listings from pharmaceutical industry as compared with the technology industry and financial industry, this may result in an undersample of data science jobs from the pharmaceutical industry. Therefore, I specifically scraped the data science related jobs from the webpages of Merck, Sanofi, and Roche, which are three of the top pharmaceutical companies worldwide. I scraped 15 data science related jobs from each company.

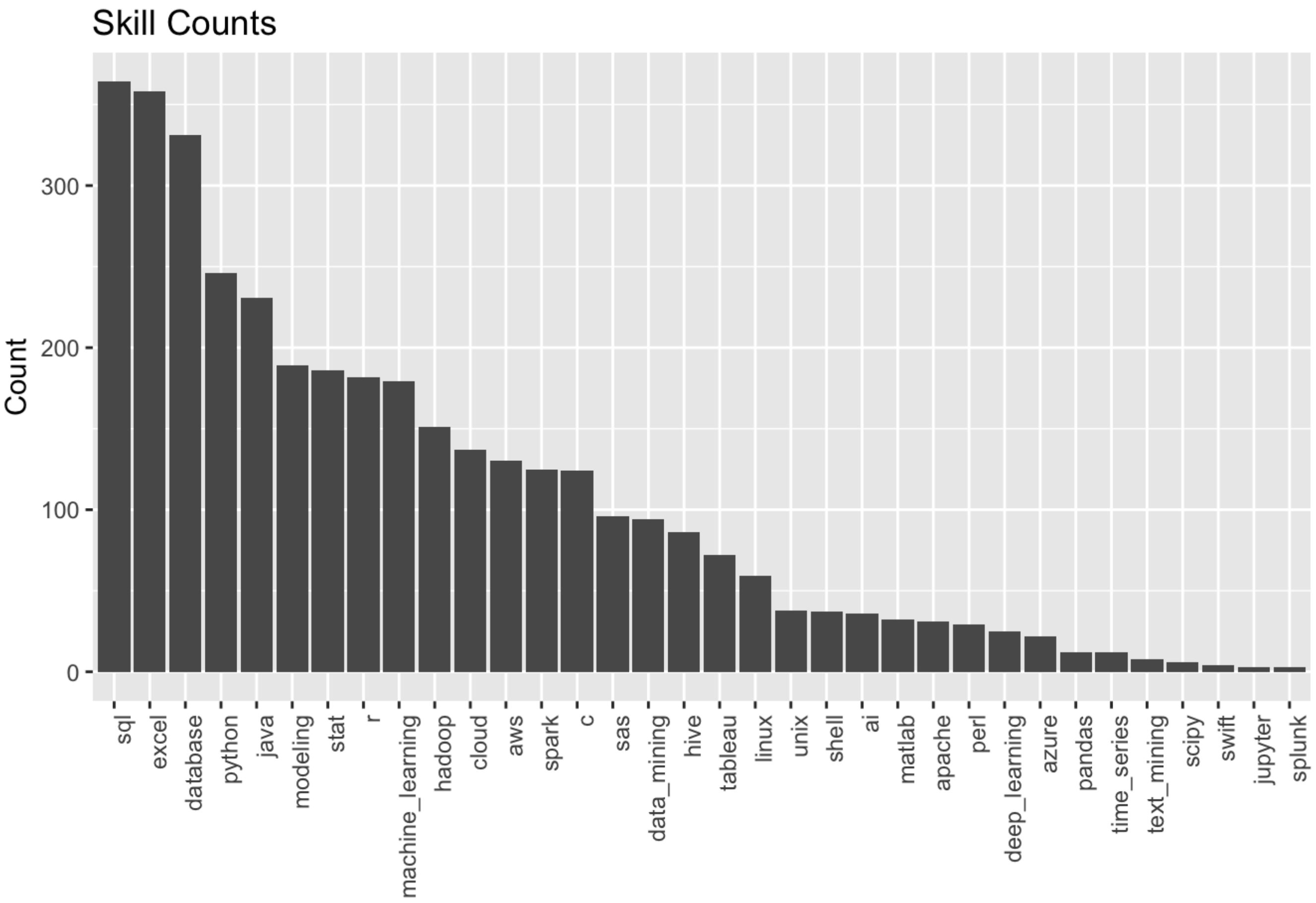
Data Processing and Exploratory Analysis

I firstly compiled the scraped data from CareerBuilder and StackOverflow into one dataset named full_dat in order to manipulate the industry categories. The industry categories are not listed in a consistent way and there are more than 300 levels for industry categories. Some example industry categories before processing are listed below:

- 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

I did some text manipulation to group similar industry categories together. I firstly removed all the special characters in industry categories and converted them into lower cases. Then I used str_detect to detect keywords in industry category listings. If there are ‘care’,‘biotech’,“lifescience”,“pharm” detected in the category listings, then they are reassigned the “Health Care” as their industry category listing; if there are “education”,“academia”,“academic”,“research” detected in the category listings, then they are reassigned the “Education” as the industry category listing; if there are ‘tech’,‘software’,‘cloud’,‘internet’,‘cyber’,‘it’,‘web’,‘saas’,‘speech’,‘computing’,‘artificial intelligence’ ,“fraud” detected, then they are reassigned the “Information Technology” as the industry category listing… There are more than 40 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, I manually checked them one-by-one and input their industry categories. However, in the end, there are still some companies with missing industry categories, due to time limitation, I removed those job listings from future analysis. After grouping, there are 16 industry categories, including Business Services, Health Care, Retail,Finance,Information Technology,Manufacturing, Education,Aerospace & Defense,Oil, Gas, Energy & Utilities, Media Telecommunications,Tourism,Non-Profit,Government,Agriculture & Forestry, and Real Estate. After removing job listings with missing industry information, the duplicated ones, and appending the job listings from Merck, Sanofi, and Roche, I have 783 jobs in total.

Then I extracted job skills required for each job. In order to do that, I removed all the special character and converted all the capitalized words into lower-case ones in the job scriptions I scraped from each individual job link. Then I created a list of job skills that are most commonly used in Data Scientist Job, including hadoop, python, sql, r, spar,sas,aws,excel,azure,java,tableau, deep learning, machine learning, AI,statistics,pandas, scipy, c,perl,text mining, matlab,hive,splunk,database,modeling, data mining, time series, jupyter, shell, unix, linex, cloud, apache, and swift. Then I used str_detect from stringr package to find a pattern match in each job description, I counted each skill once if a match is found. By doing this iterating over all jobs, I obtained a total count of each skill among all jobs.The following barplot shows the counts for each skill in descending order for all the job listings. Surprisingly, SQL turns out to be the skill with most counts, following by excel, database, python, java, and so on.



The following figures show the skill counts broken down by websites. This time, y axis represents proportion of skill counts, which is computed using counts for a specific skill divided by the total # of job listings from a specific website, in this way, I can adjust for the imbalance in websites' listing volumns. Figure a) shows the skill proportion for jobs listed on CareerBuilder, figure b) shows the skill proportion for jobs listed on StackOverflow, and figure c) shows the skill proportion for jobs listed on pharmaceutical company websites (Merck, Sanofi, and Roche). The skill rankings vary a lot across websites. For CareerBuilder, excel seems to be the skill with most counts, following by database, sql, modeling, statistics, python and R. For StackOverflow, sql seems to be the skill with most counts, following by Java, database, python, excel, amazon web service, and c. For major pharmaceutical websites, excel seems to be the skill with most counts, following by c, spark, database, and r. This makes sense to some extent, since the majority job listings I scraped from CareerBuilder I used keyword “data”. Many of them may be entry-level analyst positions which have little technical focus. For StackOverflow, it has more technical jobs for developers, therefore it requires more skills in java and SQL. However, it still looks problematic that R is not listed as the top 3 skills in any of the website.

Figure a) CareerBuilder Skill Proportion

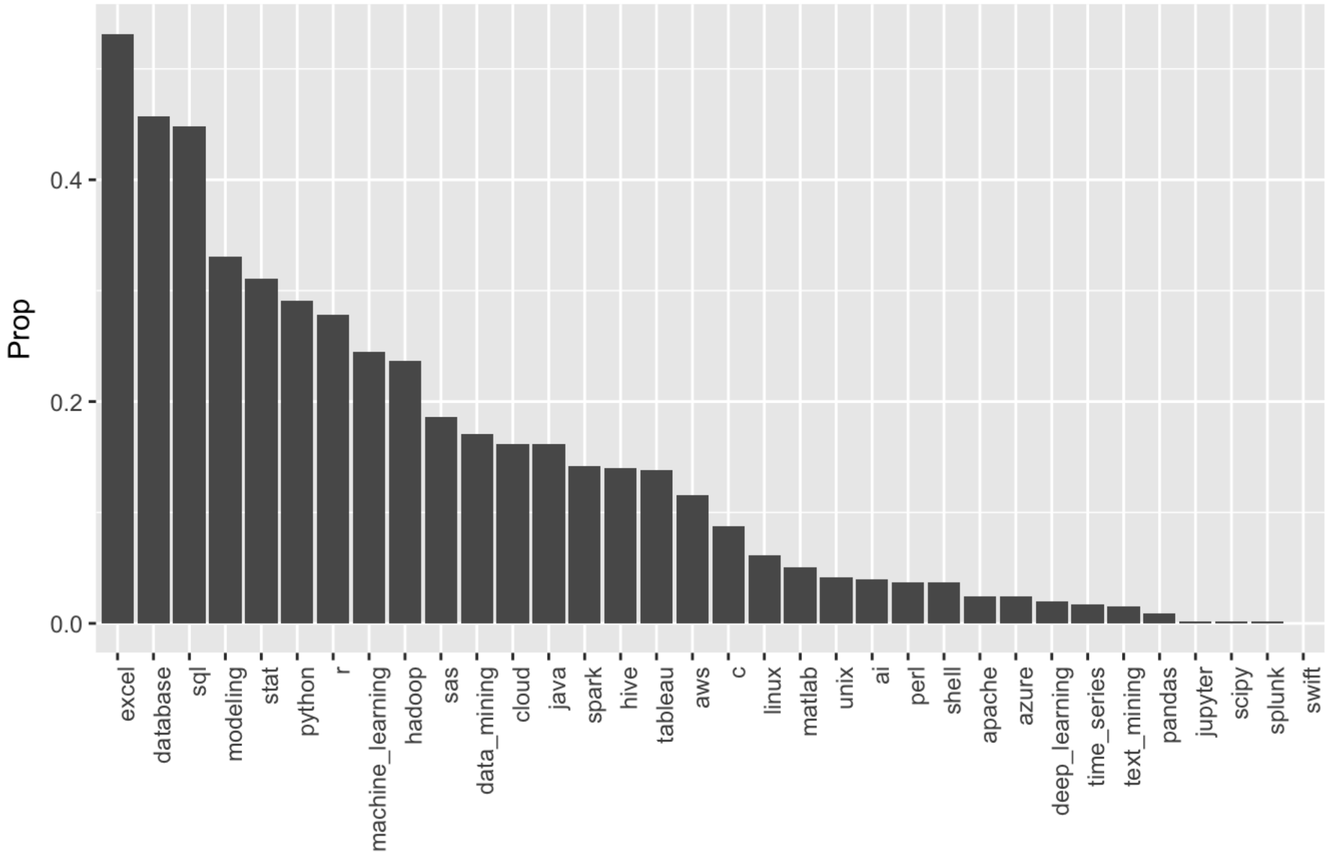
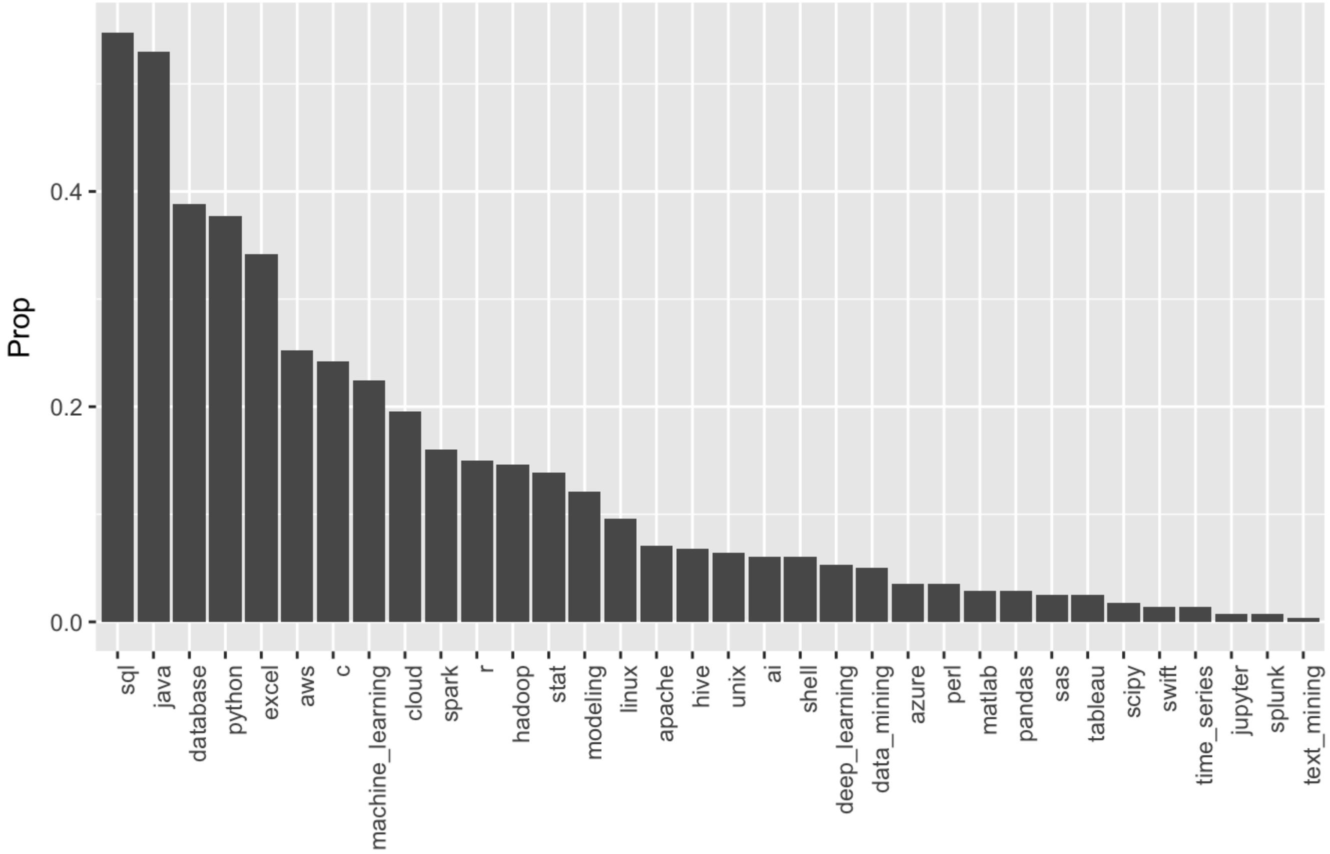
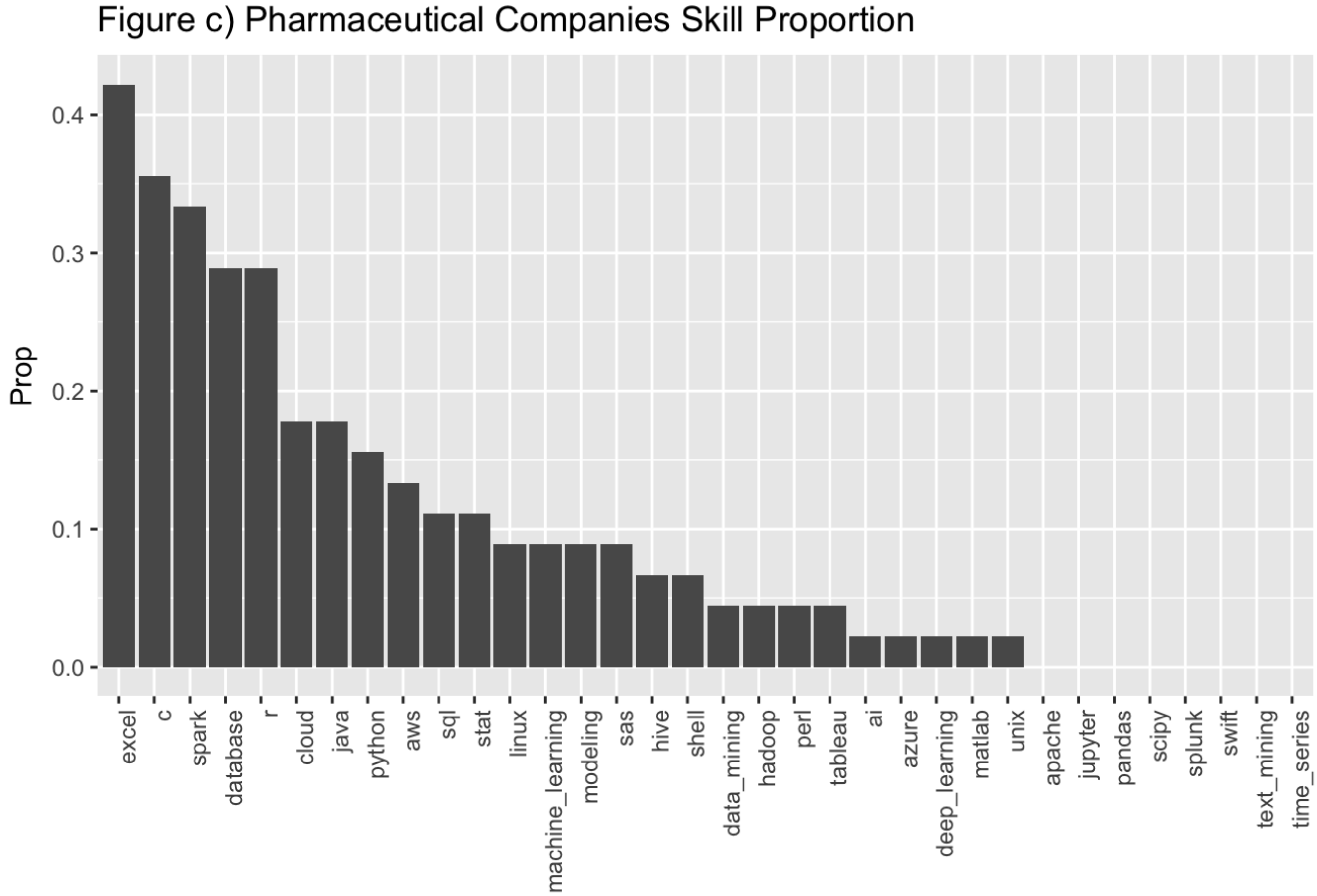
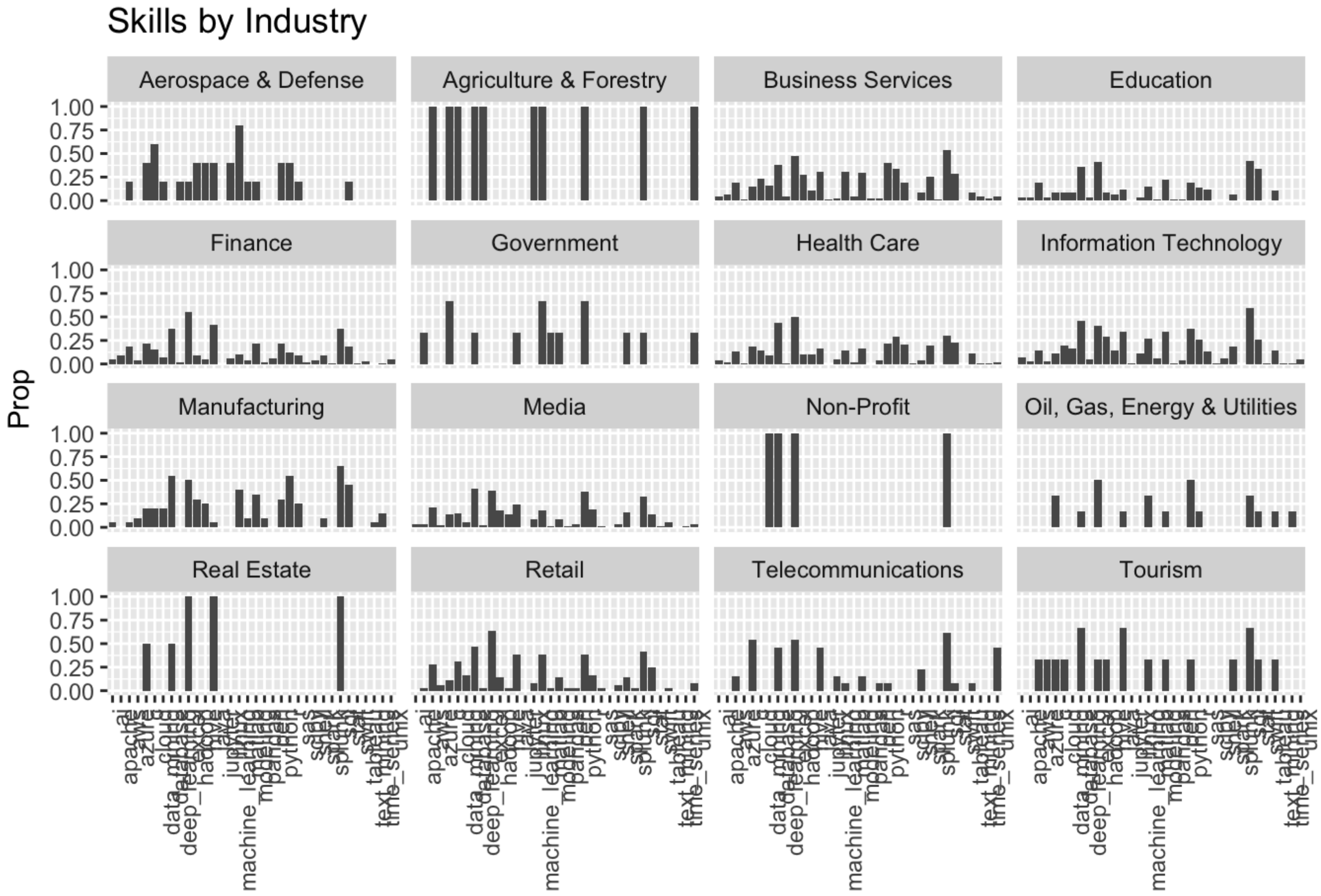


Figure b) Stackoverflow Skill Proportion





The following figures show the skill counts broken down by industries. Y axis still represents proportion of skill counts. The ranking of skills also vary a lot across industries. However, for industries other than finance, information technology, and health care, the ranking may be biased since we don’t have enough sample size. Table 1 shows the number of job listings broken down by industries, we can tell that non-profit, agriculture& forestry, government, real estate, tourism, and aerospace & defense industries all have no more than 5 job listings.



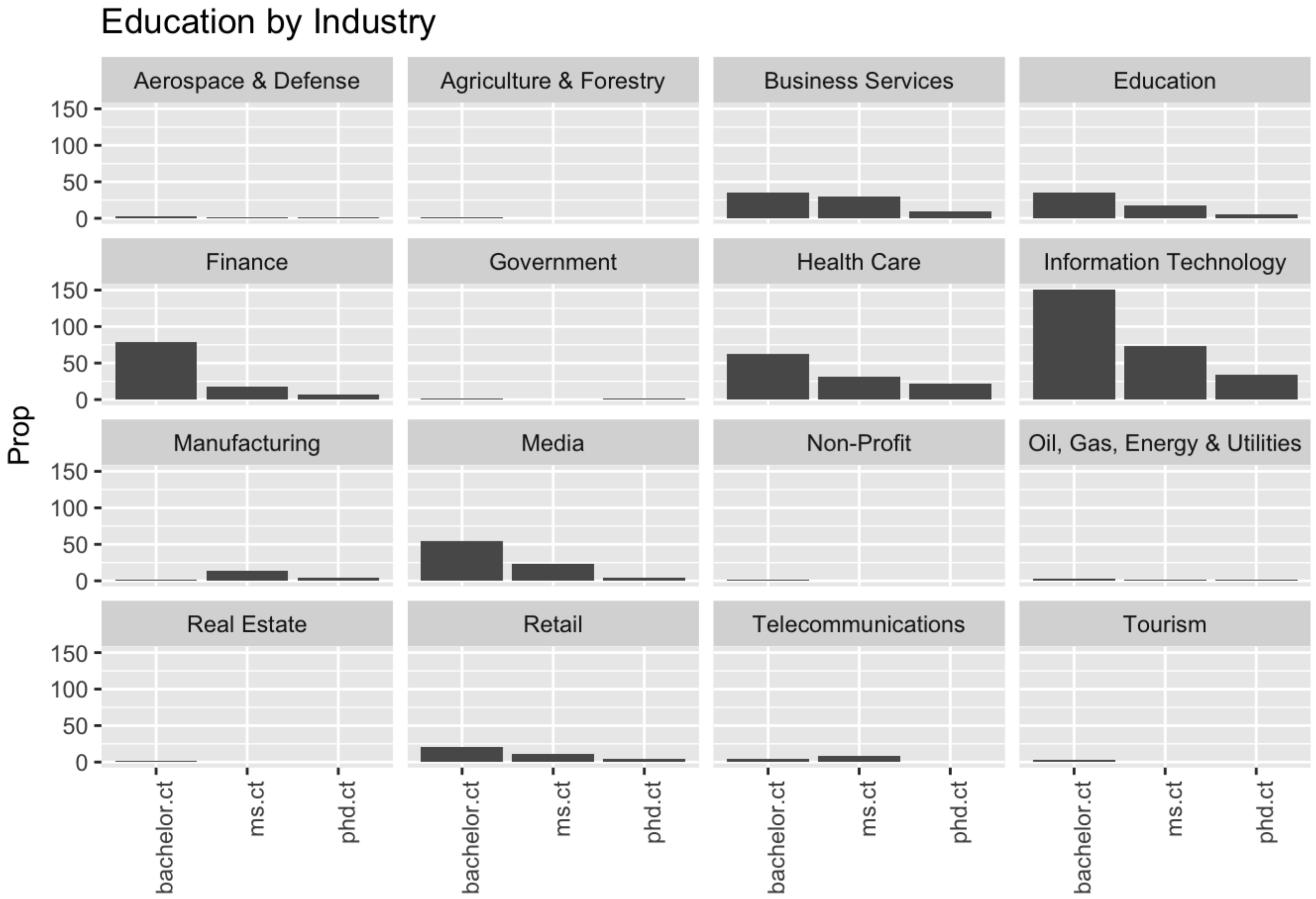
```
## Table 1. Number of Job Listings by Industry
```

```
##                   industry_clean Job_counts
## 1            Aerospace & Defense          5
## 2         Agriculture & Forestry          1
## 3              Business Services         74
## 4                      Education         59
## 5                        Finance        104
## 6                     Government          3
## 7                    Health Care        115
## 8         Information Technology        259
## 9                  Manufacturing         20
## 10                         Media         82
## 11                    Non-Profit          1
## 12 Oil, Gas, Energy & Utilities          6
## 13                   Real Estate          2
## 14                        Retail         36
## 15            Telecommunications         13
## 16                       Tourism          3
```

I also extracted education requirements for each job.I used the processed job descriptions (removed all special character and converted to lower-case) to find match with some degree related keywords. For each job, if a match is found with “master”,“masters” or “ms”, then I counted master degree once; if a match is found with “”doctor“,”doctoral" or “phd”, then I counted phd degree once; and if a match is found with “bachelor”,“bachelors”,“undergrad”,or “undergraduate”, I 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 I assumed the education requirement for this job is master and recorded “master” as the highest education requirement for this job. By doing this iterating over all jobs, I obtained the highest education requirement among all jobs.Table 2 shows the distribution of required education background among all jobs. There are 460 jobs required only bachelor degrees, 229 jobs required master degrees, and only 94 jobs required phd degrees. The following figure shows the education requirement broken down by industries. It looks like that the pattern we observed for all jobs still holds for each industry: bachelor degree are enough for most of the jobs, fewer jobs require master degree, and even fewer ones require phd degree. However, if comparing number of jobs required phd degree across industries, I found that information technology industry posts most jobs required phd degree, following by health care industry. Finance, media, and business service industries have less strict degree requirement, which makes intuitive sense for me.

```
## Table 2. Count of Highest Degree Required by Jobs
```

```
##
## bachelor       ms      phd
##      460      229       94
```



If we only focus on the finance industry, information technology industry, and health care industry, and perform a chi-square test to examine whether the degree requirements are the same across industries, the chi-square statistics is 15.1834355 and the P-value is 0.0043355, which suggests evidence that the degree requirements are not the same across these three industries.