Job Analysis

Introduction

This project mainly uses the rvest package as well as CSS selectors (which could be converted into XPath) to do the tasks the main work loop is:

- make the initial query and reading the HTML document using read_html()
- get list of nodes required for extract information wantted using CSS selectors and html_nodes():
  - preferred skills: using information of nodes CSS selector ".skill-list"
  - salary, if available: using information of nodes CSS selector ".wage"
  - full, part-time, hourly, contract, or short-term: it is included in the ".wage"
  - degree fields/subjects mentioned: using information of nodes CSS selector ".job-title a"
  - location (city, state): using information of nodes CSS selector ".location"

Note for the following 3 types of information, we need to do lots of extra work:

- required skills
- the free form text describing the position
- education level required or preferred

First, we need to get links using href attribution in the node ".job-title a" and then combined with the base URL "https://www.cybercoders.com", we obtain the full link to each job postings in details.

Second, by reading HTML in each job posting link, first find CSS selectors ".section-title" which are for the big titles such as "What You Will Be Doing", "What You Need for this Position" and then find CSS selectors ".section-data-title" which obtain the details of texts for the big titles. And the required skills are included in the section "What You Need for this Position", the free form text describing the position are included in the section What You Will Be Doing", also, the education level required or preferred are included together with the required skills.

At last, note for this project, we are interested in the search term "data analyst", "data scientist" and the job website "cybercoders.com", for other search terms in the same website, the work loop should be very similar, but for other job websites, the details should be changed correspondingly which are not discussed in this project.

Data scrape details

Following the methods described in the Approach section, we firstly obtained a data frame of raw information required. The details are in the following R codes:

```
library(rvest)
baseurl <- "https://www.cybercoders.com"
con <- read_html("https://www.cybercoders.com/search/?searchterms=data+analyst")
links1 <- con %>% html_nodes(".job-title a") %>% html_attr("href")
```

```
con2<- read_html("https://www.cybercoders.com/search/?page=2&searchterms=data%20analyst")
links2 <- con2 %>% html_nodes(".job-title a") %>% html_attr("href")

alllinks <- c(paste0(baseurl,links1),paste0(baseurl,links2))

#preferred skills
preferred_skills1<-con %>% html_nodes(".skill-list") %>% html_text()
preferred_skills2<-con2 %>% html_nodes(".skill-list") %>% html_text()
preferred_skills <- c(preferred_skills1, preferred_skills2)


#salary, if available
salary1<-con %>% html_nodes(".wage") %>% html_text()
salary2<-con2 %>% html_nodes(".wage") %>% html_text()
salary <- c(salary1,salary2)


#degree fields/subjects mentioned
title1<- con %>% html_nodes(".job-title a") %>% html_text()
title2<- con2 %>% html_nodes(".job-title a") %>% html_text()
title <- c(title1,title2)

#location (city, state)

location1<- con %>% html_nodes(".location")  %>% html_text()
location2<- con2 %>% html_nodes(".location")  %>% html_text()
location <- c(location1,location2)

dt <- NULL

for(i in 1:length(alllinks)) {

  cont <- read_html(alllinks[i])


   job_desc1 <-   cont %>%
    html_nodes(".section-title") %>%
    html_text()
  job_desc2 <-   cont %>%
    html_nodes(".section-data-title") %>%
  html_text()

  job_desc2 <- job_desc2[1:length(job_desc1)]


 #required skills
  require_skills <- job_desc2[match("What You Need for this Position",job_desc1)]

  #the free form text describing the position

  job_position <- job_desc2[match("What You Will Be Doing",job_desc1)]

  dt <- rbind(dt, c(require_skills = require_skills ,job_position = job_position))
```

```
}
dt <- data.frame(preferred_skills,salary,title,
                 location,dt)
```

As the data is raw, we show a structure as following:

```
str(dt)
```

```
## 'data.frame':    33 obs. of  6 variables:
##  $ preferred_skills: Factor w/ 24 levels "\r\n                              \r\n
##  $ salary          : Factor w/ 14 levels "Compensation Unspecified",..: 10 11 1 1 1 4 7 1 14 1 ...
##  $ title           : Factor w/ 24 levels "100% REMOTE Data Analyst - eCommerce/Digital Advertising",
##  $ location        : Factor w/ 24 levels "Benbrook, TX ",..: 23 16 4 7 15 12 9 24 21 18 ...
##  $ require_skills  : Factor w/ 26 levels "-  3-5 years of experience in data science, actuarial, sta
##  $ job_position    : Factor w/ 18 levels "-  Lead cutting edge analytics projects including:o\tSuppo
```

```
head(dt)
```

```
##
## 1                                  \r\n                           \r\n
## 2                               \r\n                           \r\n                                  \
## 3 \r\n                             \r\n                                 \r\n
## 4          \r\n                              \r\n                                 \r\n
## 5
## 6             \r\n                                 \r\n                                    \r\n
##                     salary
## 1   Full-time $80k - $110k
## 2   Full-time $80k - $120k
## 3 Compensation Unspecified
## 4 Compensation Unspecified
## 5 Compensation Unspecified
## 6  Full-time $110k - $150k
##                                                 title           location
## 1                           Data Analyst - SQL, R, Python       Seattle, WA
## 2   Senior Data Analyst - SQL, Blockchain Analysis, Python    New York, NY
## 3 100% REMOTE Data Analyst - eCommerce/Digital Advertising     Chicago, IL
## 4                                          Data Analyst       Concord, MA
## 5                                  LookML Data Analyst Mountain View, CA
## 6                  Senior Data Analyst - Pharma/Healthcare       Hoboken, NJ
##
## 1
## 2
## 3
## 4 - Interested in education, quantitative research, and statistical methods- Bachelors Degree is requ
## 5
## 6
##
## 1
## 2
## 3
## 4
## 5
## 6 You will be coding and doing data analysis within the healthcare industry. This includes predictive
```

Data cleaning

As the data is obtained in a raw format, in this section, we do some extra work to make the information scraped in a more formatted form.

First, clean preferred skills:

```
dt2 <- dt
str(dt2$preferred_skills)
```

```
##  Factor w/ 24 levels "\r\n                                    \r\n                                    \r\n
```

```
a <- gsub(" ","",dt2$preferred_skills)
a <- gsub("\r\n",";",a)
a <- strsplit(a,split=";")
dt2$preferred_skills <-  sapply(a, function(x) paste0(x[x!=""],collapse = ";"))
dt2$preferred_skills
```

```
##  [1] "SQL;R;Hive;AdHocAnalysis;Python"
##  [2] "SQL;KPI;ETL;BlockchainAnalysis;Python"
##  [3] "DataAnalyst;DigitalAdvertising;ECommerce;GoogleAnalytics;Shopify"
##  [4] "DataAnalyst;HigherEducation;SPSS;SAS;MultipleRegression"
##  [5] "DataAnalyst;Lookml;Looker;DataAnalytics"
##  [6] "R;Shinyframework;Pharmaceutical;Biostatistics;Oncology"
##  [7] "ECommerce;DigitalMarketing;Analytics"
##  [8] "Centera;Java;Security+Certification;DataMigration;Rest"
##  [9] "Clientand/orCustomer-facing;Consulting;ProductManagement;Marketing;Engineering"
## [10] "DataWarehouseAnalyst;BusinessIntelligence;Talend;PowerBI;DB2"
## [11] "Pharmaceutical;CRO;Real-worldevidence;ClaimsData;EMRData"
## [12] "Pharmaceutical;CRO;RShiny;HEOR;RWE"
## [13] "ClinicalDataAnalytics;SQL;DataVisualization;SAS;SPSS"
## [14] "DataAnalyst;SQL;Ssis;SQLServer;SSAS"
## [15] "DataWarehousing;OLAP;DataExtract/ReportingSoftware;DataQualityAssessment;DataOrganization"
## [16] "DataAnalytics;Actuarial;InsuranceSoftwareIndustry;PricingAnalysis;Statistics"
## [17] "EDI(ElectronicDataInterchange);Seeburger;SAP;JSON;XML"
## [18] "DataAnalysisTools;Web-DesignTrends;Testingmodels;UseCaseGathering;Wireframes"
## [19] "Bioinformatics;Publication;Perl;R;Python"
## [20] "Bioinformatics;Publication;Perl;R;Python"
## [21] "Bioinformatics;Publication;Perl;R;Python"
## [22] "Bioinformatics;Publication;Perl;R;Python"
## [23] "Bioinformatics;Publication;Perl;R;Python"
## [24] "SQL;IndirectLending;consumerlending;RealEstate;ComplexSQL"
## [25] "SQL;IndirectLending;consumerlending;RealEstate;ComplexSQL"
## [26] "SQL;IndirectLending;consumerlending;RealEstate;ComplexSQL"
## [27] "SQL;IndirectLending;consumerlending;RealEstate;ComplexSQL"
## [28] "SQL;IndirectLending;consumerlending;RealEstate;ComplexSQL"
## [29] "SQL;IndirectLending;consumerlending;RealEstate;ComplexSQL"
## [30] "E-Commerce;CPG;CustomerAcquisition;CRM;LTVModeling"
## [31] "PurchasingandProcurementAnalyst;Inventorycontrol/Analysis;PowerBI/PowerBW;SAPwithMMModule;Suppl
## [32] "BusinessAnalyst;SQL;Queries;HealthcareField;Data-DrivenSystems"
## [33] "Informatics;ComputationalSupport;BiomedicalDataAnalytics;PhD;Bioinformatics"
```

Then, clean preferred skills:

```
library(stringr)
str(dt2$salary)
```

```
##  Factor w/ 14 levels "Compensation Unspecified",..: 10 11 1 1 1 4 7 1 14 1 ...
```

```
dt2$fulltime <- ifelse(grepl("Full-time", dt2$salary),"Full-time","Not or Don't know")
dt2$salary <- str_trim( gsub("Full-time", "", dt2$salary))
```

And note that, as the education level required or preferred is very flexible in the required skills filed, it is too hard to find them out(sometimes, there are even no information of education levels) and for required skills and free form text describing the position, the two fields are in a flexible unstructed text format, for this website, these fields are very difficult to be formatted. However, overall, the data cleaned is in a quite better format now, it is displayed as following:

```
library(tibble)
as_tibble(dt2[,c(1,2,3,4,7,5,6)])
```

```
## # A tibble: 33 x 7
##    preferred_skills  salary  title location fulltime require_skills job_position
##    <chr>             <chr>   <fct> <fct>    <chr>    <fct>          <fct>
##  1 SQL;R;Hive;AdHoc~ $80k -~ Data~ "Seattl~ Full-ti~ "You need at ~ "You will d~
##  2 SQL;KPI;ETL;Bloc~ $80k -~ Seni~ "New Yo~ Full-ti~ "- SQL- Block~ "As a Senio~
##  3 DataAnalyst;Digi~ Compen~ 100%~ "Chicag~ Not or ~ "At least 3 y~ "- Assist S~
##  4 DataAnalyst;High~ Compen~ Data~ "Concor~ Not or ~ "- Interested~ "- Help des~
##  5 DataAnalyst;Look~ Compen~ Look~ "Mounta~ Not or ~ "- LookML- Lo~ "- Data Imp~
##  6 R;Shinyframework~ $110k ~ Seni~ "Hoboke~ Full-ti~ "MUST HAVE: -~ "You will b~
##  7 ECommerce;Digita~ $65k -~ Digi~ "Dallas~ Full-ti~ "- E-Commerce~  <NA>
##  8 Centera;Java;Sec~ Compen~ Data~ "Washin~ Not or ~ "- Centera- J~ "You will b~
##  9 Clientand/orCust~ $95k -~ Prod~ "San Jo~ Full-ti~ "At least 2+ ~ "Responsibi~
## 10 DataWarehouseAna~ Compen~ Data~ "Phoeni~ Not or ~ "- 8+ years e~  <NA>
## # ... with 23 more rows
```

Analysis

Compare with other boards

In this study, we use the board Github Jobs as a comparison, the words are:

```
library(jsonlite)
a <- fromJSON("https://jobs.github.com/positions.json?utf8=%E2%9C%93&description=data+analyst&location=
a2 <- fromJSON("https://jobs.github.com/positions.json?utf8=%E2%9C%93&description=data+scientist&locati

r <- rbind(a,a2)
library(wordcloud)
library(tidytext)
library(dplyr)
r2 <- r %>%
  unnest_tokens(word, description) %>%
  count(  word, sort = TRUE) %>% filter(nchar(word) > 6)
head(r2)
```

```
## # A tibble: 6 x 2
##   word           n
##   <chr>      <int>
## 1 experience    71
## 2 business      49
## 3 working       48
## 4 analytics     40
```

```
## 5 analysis      34
## 6 learning      33
```

```
r2 <- data.frame(r2)
wordcloud(r2[,1],r2[,2])
```

Words from this site:

```
dt2$job_position <- as.character(dt2$job_position)
r3 <- dt2 %>%
  unnest_tokens(word, job_position) %>%
  count(  word, sort = TRUE) %>% filter(nchar(word) > 6)
head(r3)
```

```
## # A tibble: 6 x 2
##   word            n
##   <chr>       <int>
## 1 business       41
## 2 analysis       29
## 3 systems        24
## 4 reporting      22
## 5 support        22
## 6 development    20
```

```
r3 <- data.frame(r3)
wordcloud(r3[,1],r3[,2])
```

So the words are consistent, as they are mainly about analysis,development,project, process, system to describe data scientists.

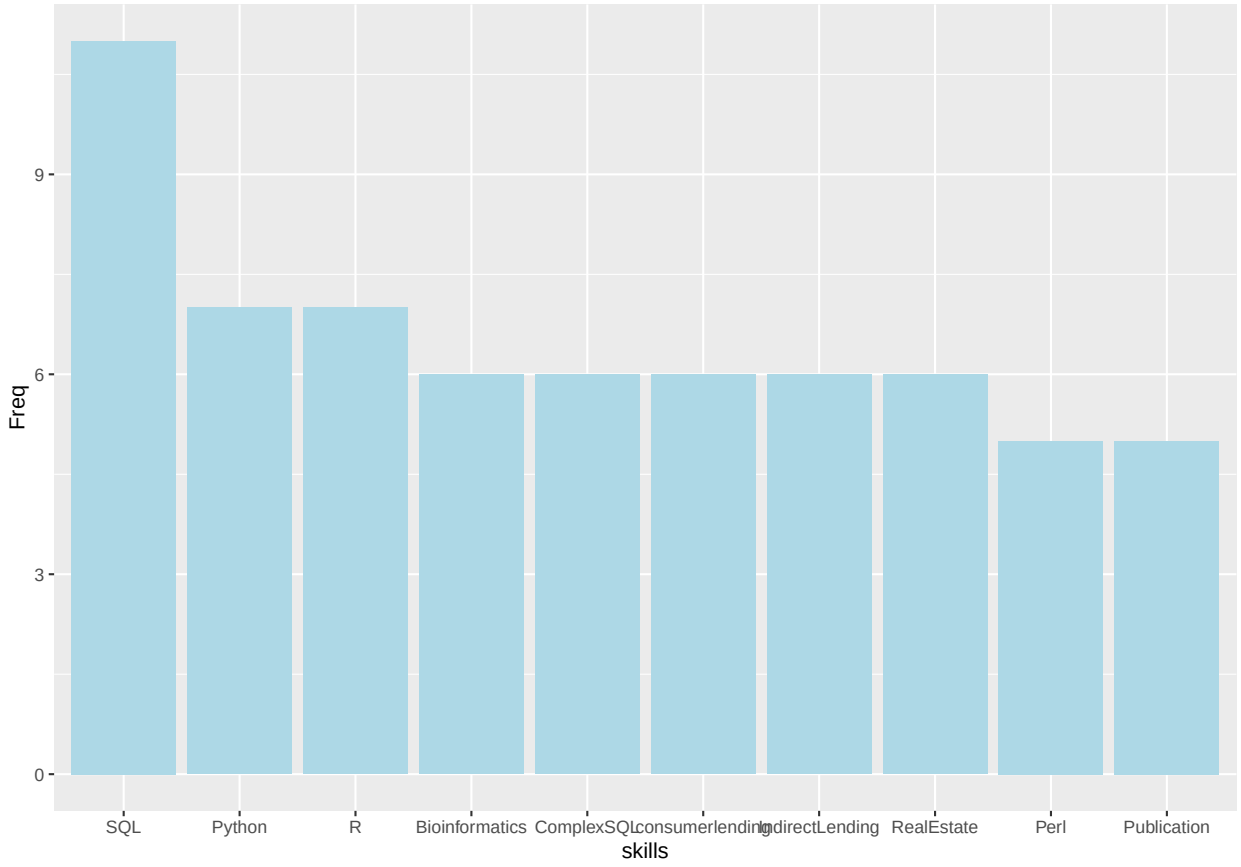
specific words to subfield

With the final cleaned data of job postings, in this section, we using the obtained data to answer an example interested questions that what are specific words to subfield of data scientists?

```
skills <- unlist(strsplit(dt2$preferred_skills, split = ";"))
df <- as.data.frame(table(skills))
library(dplyr)
library(ggplot2)
df2 <- df %>% arrange(-Freq) %>% slice(1:10)
df2
```

```
##             skills Freq
## 1              SQL   11
## 2           Python    7
## 3                R    7
## 4   Bioinformatics    6
## 5       ComplexSQL    6
## 6  consumerlending    6
## 7  IndirectLending    6
## 8       RealEstate    6
## 9             Perl    5
## 10     Publication    5
```

```
df2$skills <- factor(df2$skills, levels = df2$skills)
df2 %>% ggplot(aes(skills, Freq)) + geom_col(fill="lightblue")
```



So it can be found that for data scientist, the most frequency preferred skills are SQL, Python and R.

words related to salary levels

Now, we investigate how the words related to salary levels:

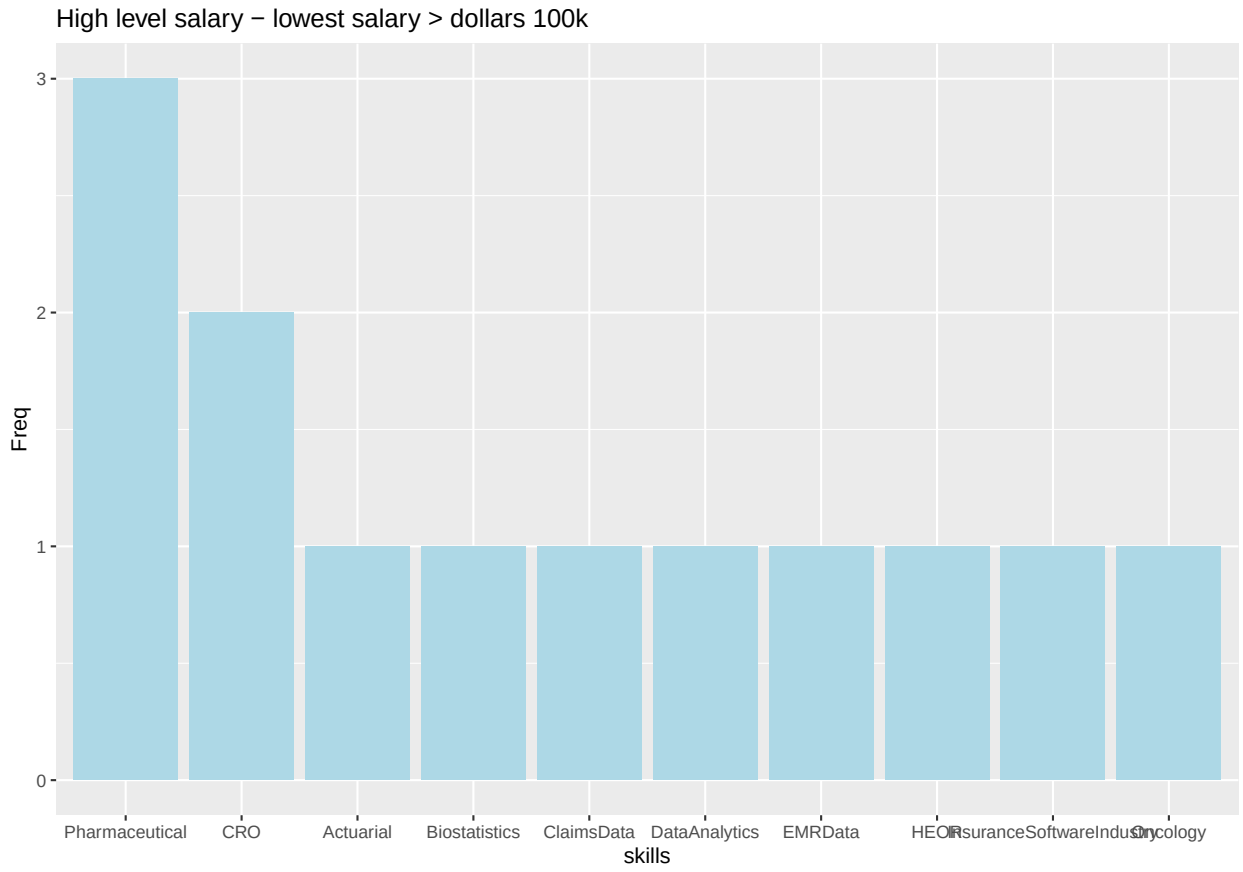
- high level salary - the lowest salary > dollars 100k
- low level salary - the lowest salary < dollars 100k"

```
skills <- unlist(strsplit(dt2$preferred_skills[dt2$salary %in% c(
                                                  "$110k - $150k" ,
                                                  "$100k - $150k" ,
                                                 "$100k - $130k" )], split = ";"))
df <- as.data.frame(table(skills))
df2 <- df %>% arrange(-Freq) %>% slice(1:10)
df2
```

```
##                    skills Freq
## 1            Pharmaceutical    3
## 2                       CRO    2
## 3                 Actuarial    1
## 4              Biostatistics    1
## 5                 ClaimsData    1
```

```
## 6              DataAnalytics    1
## 7                    EMRData    1
## 8                       HEOR    1
## 9  InsuranceSoftwareIndustry    1
## 10                  Oncology    1
```

```
df2$skills <- factor(df2$skills, levels = df2$skills)
df2 %>% ggplot(aes(skills, Freq)) + geom_col(fill="lightblue") + ggtitle("High level salary - lowest sal
```

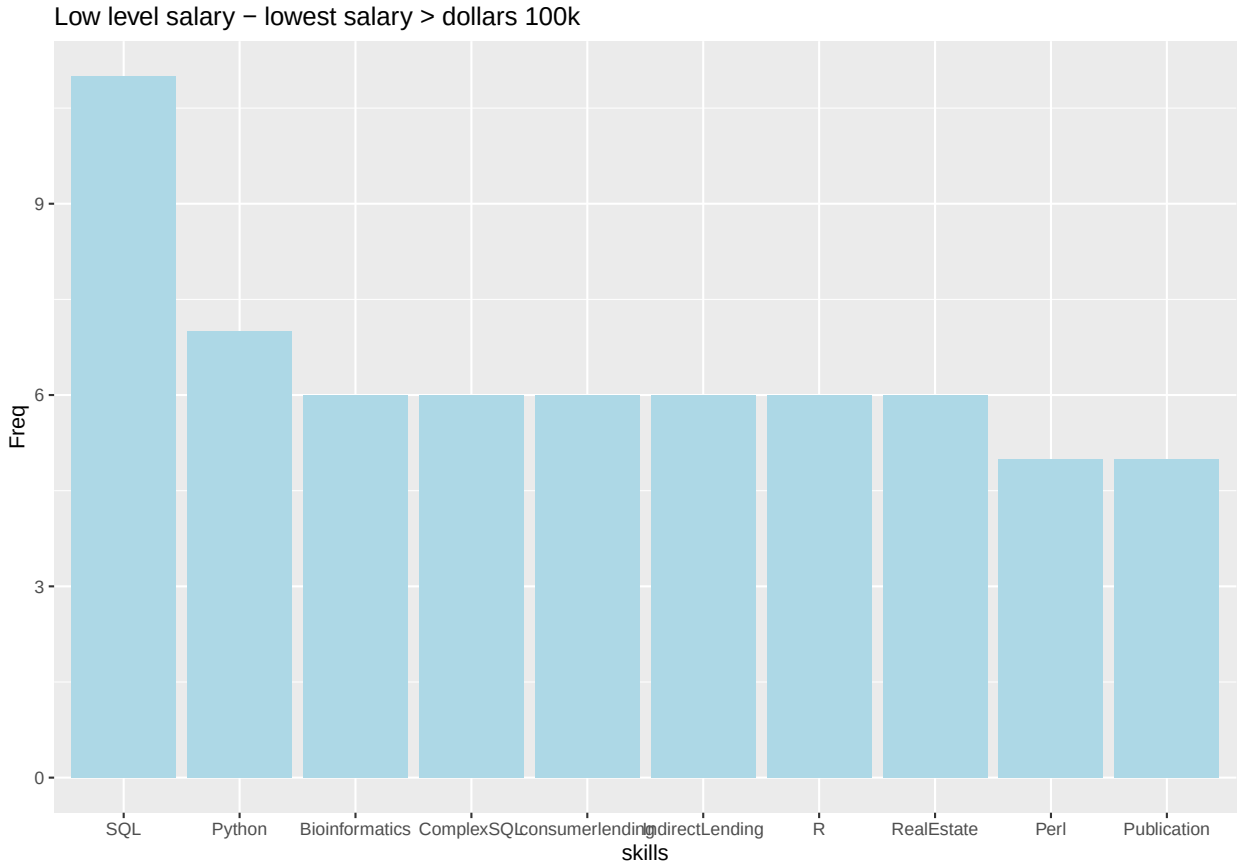



```
skills <- unlist(strsplit(dt2$preferred_skills[!dt2$salary %in% c(
                                                  "$110k - $150k" ,
                                                  "$100k - $150k" ,
                                                 "$100k - $130k" )], split = ";"))
df <- as.data.frame(table(skills))
df2 <- df %>% arrange(-Freq) %>% slice(1:10)
df2
```

```
##             skills Freq
## 1              SQL   11
## 2           Python    7
## 3   Bioinformatics    6
## 4       ComplexSQL    6
## 5  consumerlending    6
## 6  IndirectLending    6
## 7                R    6
## 8       RealEstate    6
## 9             Perl    5
```

```
## 10     Publication    5
```

```
df2$skills <- factor(df2$skills, levels = df2$skills)
df2 %>% ggplot(aes(skills, Freq)) + geom_col(fill="lightblue") + ggtitle("Low level salary - lowest sala
```



So there are clear difference between high level salary and low level one, for example, the low level salary need to know SQL, python, R and so on which might be used to do tasks in details such as data cleaning, data modeling. But the high level one such as CRO, Acturial is the type which has high theory knowledge of data science.

So that for data scientist, theory is still more expensive than programming.

A list of descriptors

Yes, at last, we can find data scientist has a list of descriptors: SQL, Python, R, Perl, analysis, development, project, process, system and so on.