Bin Lin - Project 3

Bin Lin

2016-10-22

```
#install.packages("RMySQL")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("ggplot2")
#install.packages("DBI")

library(DBI)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library(stringr)
library(RMySQL)
```

Jose's Code

```
JA_Data <- read.csv("https://raw.githubusercontent.com/juddanderman/cuny-data-607/master/Project
3/linkedin-profiles-skills.csv", encoding="UTF-8", na.strings=c("","NA"), stringsAsFactors = FAL
SE)
JA_Data <- cbind("LinkedIn", JA_Data[ , c(10,3,4,2,5,6)], NA)
JA_Data[ , 2] <- tolower(JA_Data[ , 2])
JA_Data[ , 2] <- iconv(JA_Data[ , 2], from = "latin1", to = "UTF-8")
JA_Data <- unique(JA_Data)
JA_Data$ID <- seq.int(nrow(JA_Data))
colnames(JA_Data) <- c("Source","Skill","Title","Location","Name","School","Degree","Company","R
ecord_ID")
t(head(JA_Data, 1))
```

```
##           1
## Source    "LinkedIn"
## Skill     "talent management"
## Title     "Principal and Founder, Bersin by Deloitte"
## Location  "Oakland, California"
## Name      "Josh Bersin"
## School    "University of California, Berkeley - Walter A. Haas School of Business"
## Degree    "MBA, 1988"
## Company   NA
## Record_ID "1"
```

```
KC_Data <- read.csv("https://raw.githubusercontent.com/cunyauthor/Project3/master/API_Job.csv",
encoding="UTF-8", na.strings=c("","NA"), stringsAsFactors = FALSE)
KC_Data <- KC_Data[KC_Data[ , 1] != "count",] # Remove heading rows
KC_Data <- KC_Data[!is.na(KC_Data[ , 5]),] # Remove rows with blank skills
KC_Data <- cbind(Source = "KDnuggets+Dice", KC_Data[ , c(5,7,9)], NA, NA, NA, KC_Data[ , 8])
KC_Data[ , 2]  <- as.character(str_extract_all(KC_Data[ , 2] , "l\\=\\S+\\&c"))
KC_Data[ , 2]  <- str_replace_all(KC_Data[ , 2] , "(l\\=|\\&c)", "")
KC_Data[ , 2]  <- str_replace_all(KC_Data[ , 2] , "\\+", " ")
KC_Data$ID <- seq.int(nrow(KC_Data))
colnames(KC_Data) <- c("Source","Skill","Title","Location","Name","School","Degree","Company","R
ecord_ID")
t(head(KC_Data, 1))
```

```
##           1
## Source    "KDnuggets+Dice"
## Skill     "Owning Up To The Title"
## Title     "Sr Sitecore Web Developer"
## Location  "Milford"
## Name      NA
## School    NA
## Degree    NA
## Company   "UROOJ Corporation"
## Record_ID "1"
```

```
# I just changed the connection to the MySQL, so that the password and user name #won't be show
n.
rmysql.settingsfile<-"C:/ProgramData/MySQL/MySQL Server 5.7/my.ini"
connection <- dbConnect(RMySQL::MySQL(), default.file=rmysql.settingsfile, dbname = "assignment
2", user=NULL, password=NULL)

dbSendQuery(connection, 'CREATE SCHEMA IF NOT EXISTS Skills;')
```

```
## <MySQLResult:2,0,0>
```

```
dbSendQuery(connection, 'USE Skills;')
```

```
## <MySQLResult:132913984,0,1>
```

```
dbSendQuery(connection, 'DROP TABLE IF EXISTS tbl_LinkedIn;')
```

```
## <MySQLResult:8,0,2>
```

```
dbSendQuery(connection, 'DROP TABLE IF EXISTS tbl_KDnuggets_Dice;')
```

```
## <MySQLResult:1383096653,0,3>
```

```
dbWriteTable(connection, "tbl_LinkedIn", JA_Data, append = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```
dbSendQuery(connection, "ALTER TABLE tbl_LinkedIn
            MODIFY COLUMN Record_id MEDIUMINT NOT NULL,
            MODIFY COLUMN Source VARCHAR(25) NOT NULL,
            MODIFY COLUMN Skill VARCHAR(50) NOT NULL,
            MODIFY COLUMN Title VARCHAR(250) NULL,
            MODIFY COLUMN Location VARCHAR(50) NULL,
            MODIFY COLUMN Name VARCHAR(50) NULL,
            MODIFY COLUMN School VARCHAR(75) NULL,
            MODIFY COLUMN Degree VARCHAR(100) NULL,
            MODIFY COLUMN Company VARCHAR(50) NULL,
            ADD PRIMARY KEY (Record_id);")
```

```
## <MySQLResult:0,0,7>
```

```
dbWriteTable(connection, "tbl_KDnuggets_Dice", KC_Data, append = TRUE, row.names = FALSE)
```

```
## [1] TRUE
```

```
dbSendQuery(connection, "ALTER TABLE tbl_KDnuggets_Dice
            MODIFY COLUMN Record_id MEDIUMINT NOT NULL,
            MODIFY COLUMN Source VARCHAR(25) NOT NULL,
            MODIFY COLUMN Skill VARCHAR(50) NOT NULL,
            MODIFY COLUMN Title VARCHAR(250) NULL,
            MODIFY COLUMN Location VARCHAR(50) NULL,
            MODIFY COLUMN Name VARCHAR(50) NULL,
            MODIFY COLUMN School VARCHAR(75) NULL,
            MODIFY COLUMN Degree VARCHAR(100) NULL,
            MODIFY COLUMN Company VARCHAR(50) NULL,
            ADD PRIMARY KEY (Record_id);")
```

```
## <MySQLResult:97477184,0,11>
```

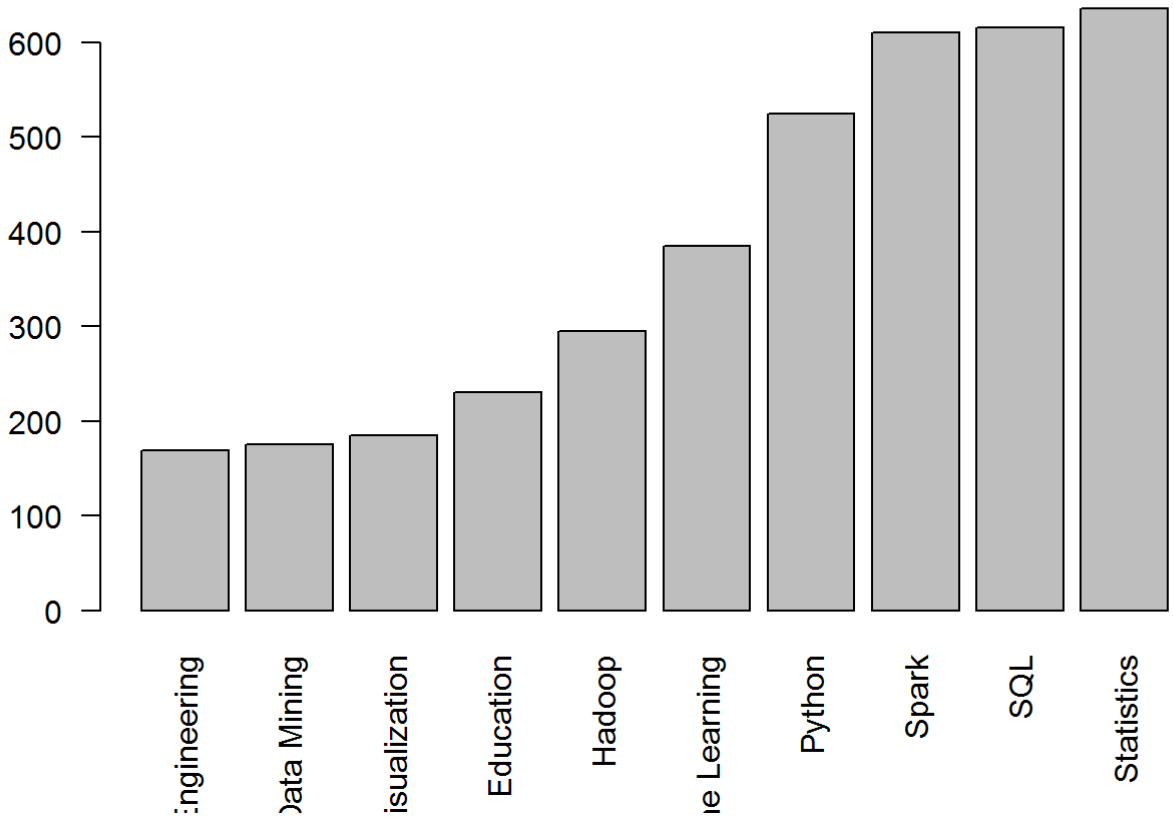
```
All_Data <- dbGetQuery(connection, "SELECT * FROM tbl_LinkedIn
                       UNION SELECT * FROM tbl_KDnuggets_Dice
                       ORDER BY Source, Skill, Title;")
```

My Code: I created a barplot for the entire dataset, without regarding to linkedin or KD Nuggets or Dice. The graph shows the top 10 skills and their frequency.

```
raw_data <- All_Data[!(is.na(All_Data$Skill) & All_Data$Skill != "character(0)"), ]
a <- which(with(raw_data, table(raw_data$Skill)) > 100)
head(a)
```

```
##         character(0) Communication Skills     Data Engineering
##                   96                  113                  169
##          Data Mining   Data Visualization            Education
##                  175                  185                  230
```

```
barplot(a[3:12], las=2)
```



I am interested in the rlationship between each company and their most desired employee skills.

```
raw_data <- All_Data%>%
  select(Company, Skill)%>%
  na.omit()
head(raw_data)
```

```
##                         Company             Skill
## 2              NORTHROP GRUMMAN Advanced Analysis
## 3                        Amazon Advanced Analysis
## 4                        Amazon Advanced Analysis
## 5                 Intellisearch Advanced Analysis
## 6 Bebee Affinity Social Network Advanced Analysis
## 7                Eliassen Group Advanced Analysis
```

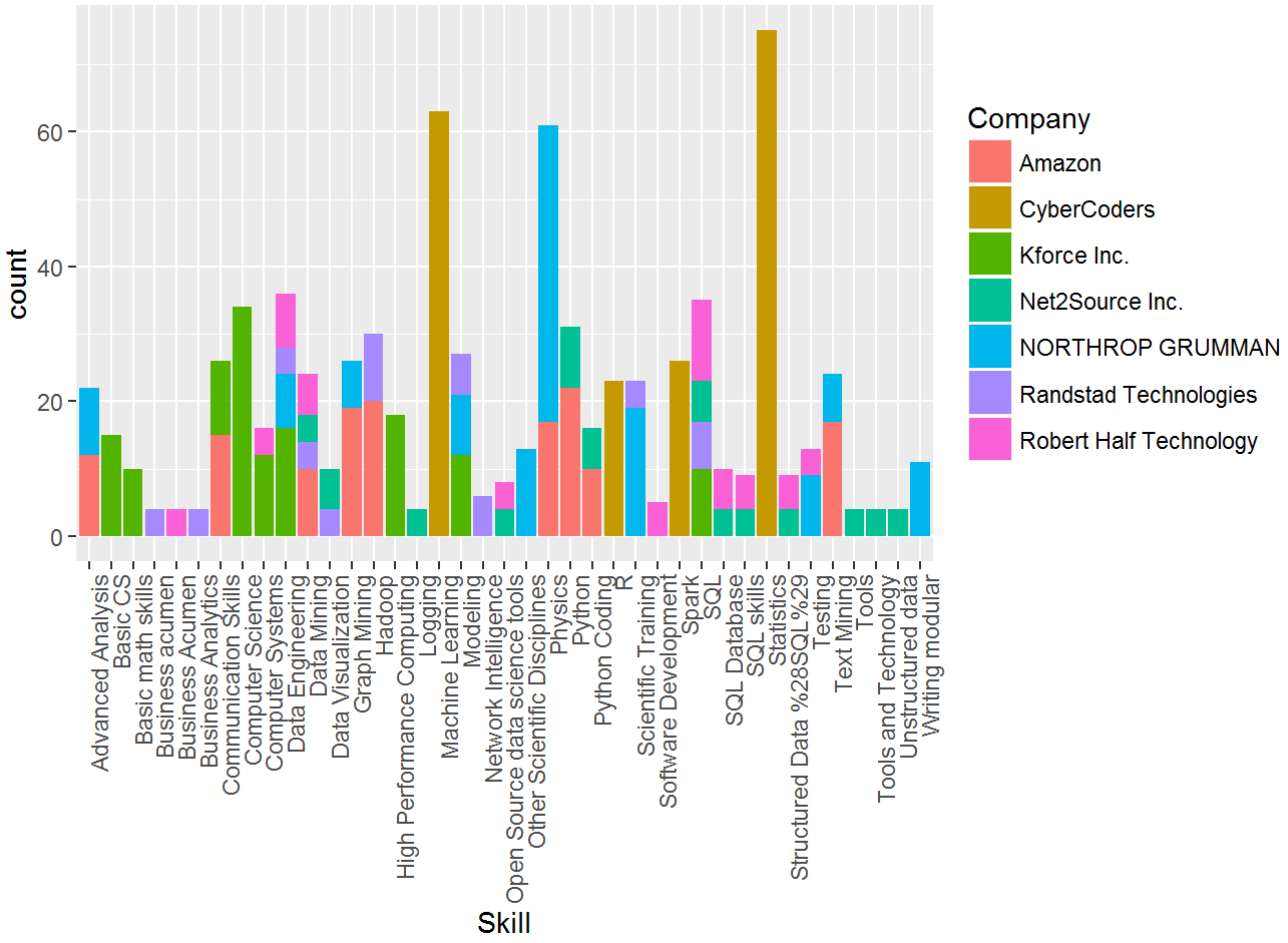
```
b <-raw_data%>%
  group_by(Company, Skill)%>%
  summarize(count = n())%>%
  mutate(total = sum(count), percentage = count/total)%>%
  filter(total > 100 & percentage > 0.03)%>%
  arrange(desc(total), desc(percentage))
head(b)
```

```
## Source: local data frame [6 x 5]
## Groups: Company [2]
##
##       Company            Skill count total percentage
##         <chr>            <chr> <int> <int>      <dbl>
## 1 CyberCoders       Statistics    75   625 0.12000000
## 2 CyberCoders Machine Learning    63   625 0.10080000
## 3 CyberCoders            Spark    26   625 0.04160000
## 4 CyberCoders                R    23   625 0.03680000
## 5      Amazon           Python    22   324 0.06790123
## 6      Amazon           Hadoop    20   324 0.06172840
```

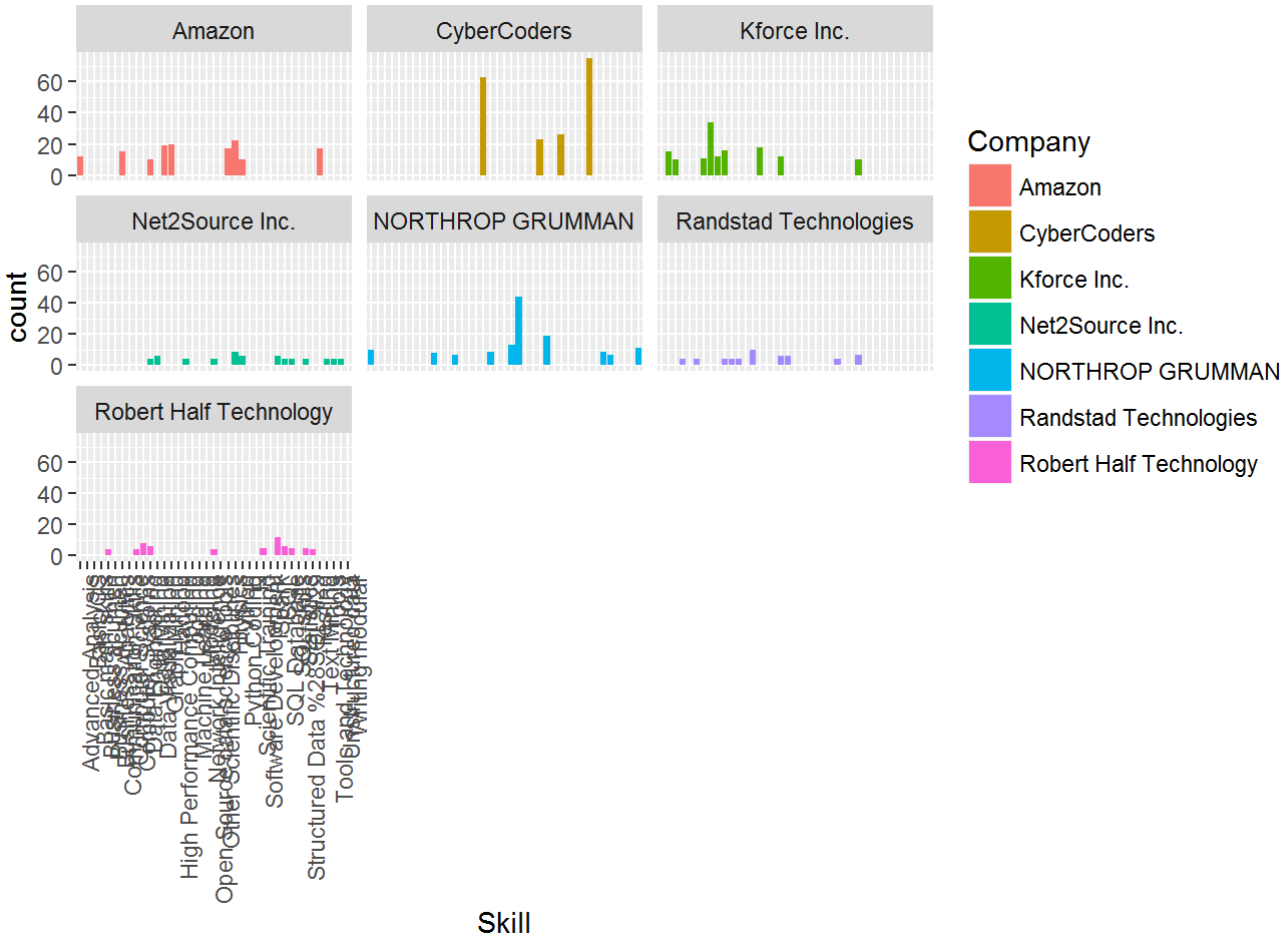
The first barplot shows different companies actually have different interest of skill sets they want candidates to have. I pick whatever companies have more than 100 positions open. However, I think the better way to do is to get the top 10 companies that are hiring. The percentage I set up is just for the purpose of getting rid off the skills that do not meet most companies' interest. The second plot looks a little messy. If someone can help me organize it, (for each barplot, the skills can line up from the most frequent to the lleast frequent), I really appreaciate. If not, maybe just take off the second barplot.

I also want to connect skills with each schools. But it is actually a bad idea, because a lot of data is missing scholl information.

```
ggplot(data = b, aes(x = Skill, y = count, fill = Company)) + geom_bar(stat="identity") +
theme(axis.text.x = element_text(angle=90, hjust = 1))
```




```
ggplot(data = b, aes(x = Skill, y = count, fill = Company)) + geom_bar(stat ="identity", positio
n = "dodge") + facet_wrap(~Company) +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Jose's Code

```
dbSendQuery(connection, 'DROP TABLE tbl_LinkedIn;')
```

```
## <MySQLResult:8,0,13>
```

```
#dbSendQuery(connection, 'DROP TABLE tbl_Friends;')
dbSendQuery(connection, 'DROP SCHEMA Skills;')
```

```
## <MySQLResult:8,0,14>
```

```
dbDisconnect(connection)
```

```
## Warning: Closing open result sets
```

```
## [1] TRUE
```