Capstone Project Milestone Report

Jihye Ko

# Introduction

**Importance**

These days, the importance of mental health is being raised. It is estimated that one in four or five of people would experience some mental health problems in lifetime. For this reason, it is very important to recognize signs of mental health problems and provide help which people can easily access to. One of the reasons which cause mental issue is stress. People are stressed out in many different reasons, such as school, works, relationships and so on. Especially, many workers in companies suffer from mental health in some way. Taking care of employees is one of responsibilities that the company should have. By solving this problem, the company can predict how effectively their mental health benefits are used by employees.

**Question**

Can we predict the employee’s willingness of seeking help of mental health problem in companies based on the size of the company, knowledge of wellness probram and care options that the company supply, and protection of anonymity of employee’s mental health issue?

# Description of Data

**Dataset**

The dataset used for this project was retrieved from [kaggle.com](https://www.kaggle.com/osmi/mental-health-in-tech-survey)

**Description of Data**

The data that I am using is survey type. This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. This dataset contains 26 variables which are related to mental and physical health and their accessibility about health services in their company, especially in tech companies in this case. The varialbes as follows:

Description of variables

Age  
Gender  
Country  
State: If you live in the United States, which state or territory do you live in?  
Self\_employed: Are you self-employed?  
Family\_history: Do you have a family history of mental illness?  
Treatment: Have you sought treatment for a mental health condition?  
Work\_interfere: If you have a mental health condition, do you feel that it interferes with your work?  
No\_employees: How many employees does your company or organization have?  
Remote\_work: Do you work remotely (outside of an office) at least 50% of the time?  
Tech\_company: Is your employer provide mental health benefits?  
Benefits: Does your employer provide mental health benefits?  
Care\_options: Do you know the options for mental health care your employer provides?  
Wellness\_program: Has your employer ever discussed mental health as part of an employee wellness program?  
Seek\_help: Does your employer provide resources to learn more about mentlah health issue and how to seek help?  
Anonymity: Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?  
Leave: How easy is it for you to take medical leave for a mental health condition?  
Mental\_health\_consequence: Do you think that discussing a mental health issue with your employer would have negative consequences?  
Phys\_health\_consequence: Do you think that discussing a physical health issue with your employer would have negative consequences?  
Coworkers: Would you be willing to discuss a mental health issue with your coworkers?  
Supervisor: Would you be willing to discuss a mental health issue with your direct supervisor(s)?  
Mental\_health\_interview: Would you bring up a mental health issue with a potential employer in an interview?  
Phys\_health\_interview: Would you bring up a physical health issue with a potential employer in an interview?  
Mental\_vs\_physical: Do you feel that your employer takes ental health as seriously as physical health?  
Obs\_consequence: Have you heard of or observed negative consequences for coworkers with mental health conditions in your workplace?

# Lists of libraries

* library(rattle)
* library(rpart.plot)
* library(RColorBrewer)
* library(rpart)
* library(tidyr)
* library(dplyr)
* library(ggplot2)
* library(plyr)
* library(corrplot)

# Analysis and Cleaning of variables in the Dataset

**Load the data from csv file**

## Set this to your current path of datasets  
setwd("C:\\");  
## Loading data  
raw = read.csv(file='raw.csv',header=TRUE,sep=',')  
summary(raw)

## Age Gender Country state   
## Min. :-1.726e+03 F :251 United States :751 CA :138   
## 1st Qu.: 2.700e+01 M :995 United Kingdom:185 WA : 70   
## Median : 3.100e+01 OTHER: 13 Canada : 72 NY : 57   
## Mean : 7.943e+07 Germany : 45 TN : 45   
## 3rd Qu.: 3.600e+01 Ireland : 27 TX : 44   
## Max. : 1.000e+11 Netherlands : 27 (Other):390   
## (Other) :152 NA's :515   
## self\_employed family\_history treatment work\_interfere  
## No :1095 No :767 No :622 Never :213   
## Yes : 146 Yes:492 Yes:637 Often :144   
## NA's: 18 Rarely :173   
## Sometimes:465   
## NA's :264   
##   
##   
## no\_employees remote\_work tech\_company benefits   
## :451 No :883 No : 228 Don't know:408   
## 06월 25일 : 1 Yes:376 Yes:1031 No :374   
## 100-500 :176 Yes :477   
## 26-100 :289   
## 500-1000 : 60   
## More than 1000:282   
##   
## care\_options wellness\_program seek\_help anonymity   
## No :501 Don't know:188 Don't know:363 Don't know:819   
## Not sure:314 No :842 No :646 No : 65   
## Yes :444 Yes :229 Yes :250 Yes :375   
##   
##   
##   
##   
## leave mental\_health\_consequence  
## Don't know :563 Maybe:477   
## Somewhat difficult:126 No :490   
## Somewhat easy :266 Yes :292   
## Very difficult : 98   
## Very easy :206   
##   
##   
## phys\_health\_consequence coworkers supervisor   
## Maybe:273 No :260 No :393   
## No :925 Some of them:774 Some of them:350   
## Yes : 61 Yes :225 Yes :516   
##   
##   
##   
##   
## mental\_health\_interview phys\_health\_interview mental\_vs\_physical  
## Maybe: 207 Maybe:557 Don't know:576   
## No :1008 No :500 No :340   
## Yes : 44 Yes :202 Yes :343   
##   
##   
##   
##   
## obs\_consequence  
## No :1075   
## Yes: 184   
##   
##   
##   
##   
##   
## comments   
## \* Small family business - YMMV. : 5   
## - : 1   
## : 1   
## (yes but the situation was unusual and involved a change in leadership at a very high level in the organization as well as an extended leave of absence) : 1   
## A close family member of mine struggles with mental health so I try not to stigmatize it. My employers/coworkers also seem compassionate toward any kind of health or family needs.: 1   
## (Other) : 155   
## NA's :1095

**Dataset**

1259 observations and 26 variables

**Data Wrangling**

1. Variable with missing values

The variable which had missing value was *no\_employees*, in other words, the size of the company. Since imputation was needed for dealing with missing values, the distribution of that varialbe was first analyzed. If it was normally distributed, then the mean can be used for missing values. If not, the median can be used for them.

Since the data of *no\_employees* were not normally distributed, the missing values were replaced by median (“100-500”). Also, the data of *no\_employees* were catagorized into two sizes: Medium or Large.

## Categorizing company into MEDIUM or LARGE (2 categories only)  
raw$no\_employees = plyr::revalue(raw$no\_employees, c("06월 25일"="MEDIUM", "100-500"="MEDIUM" , "26-100" = "MEDIUM" , "500-1000"="LARGE" , "More than 1000" = "LARGE"))  
## All unknown size companies default to MEDIUM  
raw$no\_employees[raw$no\_employees == ''] <- "MEDIUM"

1. Variable with variety answers

One of the variables, *Country*, had about more than 100 answers which made the further analysis difficult. Therefore, countires were catagorized with continents (South America, North America, Africa, Oceania, Asia and Europe).

## Categorizing countries into continents (6 categories)  
raw$Country = plyr::revalue(raw$Country, c("Mexico" = "North America", "Canada"="North America", "United States"="North America" , "United Kingdom" = "North America", "Costa Rica" = "North America", "Bahamas, The" = "North America"))  
raw$Country = plyr::revalue(raw$Country, c("Thailand" = "Asia", "Singapore" = "Asia", "Israel" = "Asia", "Japan"="Asia", "India" = "Asia", "China"="Asia", "Philippines" = "Asia"))  
raw$Country = plyr::revalue(raw$Country, c("Czech Republic" = "Europe", "Denmark" = "Europe", "Hungary" = "Europe", "Bosnia and Herzegovina"="Europe", "Greece" = "Europe", "France" = "Europe", "United Kingdom = Europe", "Portugal" = "Europe", "Switzerland" = "Europe", "Georgia" = "Europe", "Moldova" = "Europe", "Poland" = "Europe", "Austria" = "Europe", "Germany" = "Europe", "Slovenia" = "Europe", "Russia" = "Europe", "Ireland" = "Europe", "Italy" = "Europe", "Bulgaria" = "Europe", "Sweden" = "Europe", "Latvia" = "Europe", "Romania"="Europe", "Belgium" = "Europe", "Spain"="Europe", "Finland"="Europe", "Netherlands"="Europe", "Croatia" = "Europe", "Norway"="Europe"))  
raw$Country = plyr::revalue(raw$Country, c("Brazil"="South America", "Colombia" = "South America", "Uruguay"="South America"))  
raw$Country = plyr::revalue(raw$Country, c("Australia" = "Oceania", "New Zealand" = "Oceania"))  
raw$Country = plyr::revalue(raw$Country, c("Nigeria" = "Africa", "Zimbabwe"="Africa", "South Africa" = "Africa"))  
summary(raw$Country)

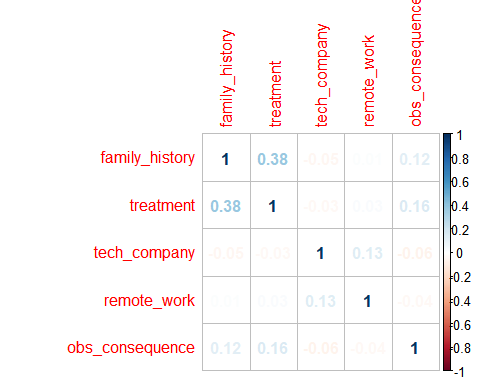
## Oceania Europe North America South America Asia   
## 29 177 1013 9 23   
## Africa   
## 8

# Correlation among variables

**Variables with two answers (Yes/No)**

Some variables (i.e. family\_history, treatment, tech\_company, remote\_work, obs\_consequence) had either ‘Yes’ or ‘No’ answers. To observe the relationship among variables, corrPlot was used. By default, corrPlot needs numeric values not string. Therefore, changing string to numeric values was necessary steps. Then, correlations between variables were plotted with correlation coefficients.

## Changing string to numeric values  
raw$family\_history = plyr::revalue(raw$family\_history, c("Yes" = 1, "No" = 3))  
raw$treatment = plyr::revalue(raw$treatment, c("Yes" = 1, "No" = 3))  
raw$tech\_company = plyr::revalue(raw$tech\_company, c("Yes" = 1, "No" =3))  
raw$remote\_work = plyr::revalue(raw$remote\_work, c("Yes" = 1, "No" = 3))  
raw$obs\_consequence = plyr::revalue(raw$obs\_consequence, c("Yes" = 1, "No" = 3))  
  
raw$family\_history <- as.numeric(as.character(raw$family\_history))  
raw$treatment <- as.numeric(as.character(raw$treatment))  
raw$tech\_company <- as.numeric(as.character(raw$tech\_company))  
raw$remote\_work <- as.numeric(as.character(raw$remote\_work))  
raw$obs\_consequence <- as.numeric(as.character(raw$obs\_consequence))  
  
## corrPlot to examine the relationships among variables  
MentalHealth\_Data\_1 <- subset(raw, select = c(family\_history, treatment, tech\_company, remote\_work, obs\_consequence))  
N <- cor(MentalHealth\_Data\_1)  
corrplot(N, method = "number")



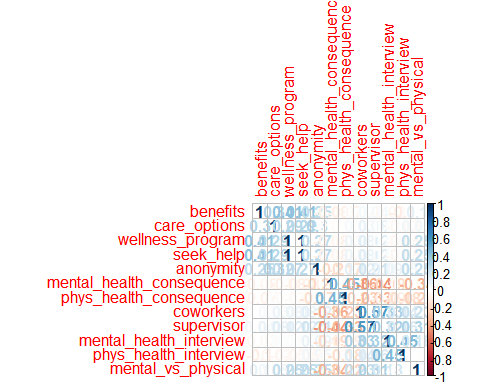
*Conclusion 1:*

From this corrPlot, there was no strong correlation between varaibles which had two answers.

**Variables with three answers (Yes/Don’t know/No)**

Other variables (i.e. benefits, care\_options, wellness\_program, seek\_hlep, anonymity, mental\_health\_consequence, phys\_health\_consequence, coworkers, supervisor, mental\_health\_interview, phys\_health\_interview, mental\_vs\_physical) had three answers as follows: ‘Yes’, ‘Don’t Know’, and ‘No’. Just like above, strings were changed to numbers and correlation between variables were examined.

## Changing string to numeric values  
raw$benefits = plyr::revalue(raw$benefits, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
raw$care\_options = plyr::revalue(raw$care\_options, c("Yes" = 1, "Not sure" = 2, "No" = 3))  
raw$wellness\_program = plyr::revalue(raw$wellness\_program, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
raw$seek\_help = plyr::revalue(raw$wellness\_program, c("Yes" = 1, "Don't know" = 2, "No" =3))  
raw$anonymity = plyr::revalue(raw$anonymity, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
raw$mental\_health\_consequence = plyr::revalue(raw$mental\_health\_consequence, c("Yes" = 1, "Maybe" = 3, "No" =3))  
raw$phys\_health\_consequence = plyr::revalue(raw$phys\_health\_consequence, c("Yes" = 1, "Maybe" = 2, "No" = 3))  
raw$coworkers = plyr::revalue(raw$coworkers, c("Yes" = 1, "Some of them" = 2, "No" =3))  
raw$supervisor = plyr::revalue(raw$supervisor, c("Yes" = 1, "Some of them" = 2, "No" = 3))  
raw$mental\_health\_interview = plyr::revalue(raw$mental\_health\_interview, c("Yes" = 1, "Maybe" = 2, "No" =3))  
raw$phys\_health\_interview = plyr::revalue(raw$phys\_health\_interview, c("Yes" = 1, "Maybe" = 2, "No" =3))  
raw$mental\_vs\_physical = plyr::revalue(raw$mental\_vs\_physical, c("Yes" = 1, "Don't know" = 2, "No" = 3))  
  
raw$benefits <- as.numeric(as.character(raw$benefits))  
raw$care\_options <- as.numeric(as.character(raw$care\_options))  
raw$wellness\_program <- as.numeric(as.character(raw$wellness\_program))  
raw$seek\_help <- as.numeric(as.character(raw$seek\_help))  
raw$anonymity <- as.numeric(as.character(raw$anonymity))  
raw$mental\_health\_consequence <- as.numeric(as.character(raw$mental\_health\_consequence))  
raw$phys\_health\_consequence <- as.numeric(as.character(raw$phys\_health\_consequence))  
raw$coworkers <- as.numeric(as.character(raw$coworkers))  
raw$supervisor <- as.numeric(as.character(raw$supervisor))  
raw$mental\_health\_interview <- as.numeric(as.character(raw$mental\_health\_interview))  
raw$phys\_health\_interview <- as.numeric(as.character(raw$phys\_health\_interview))  
raw$mental\_vs\_physical <- as.numeric(as.character(raw$mental\_vs\_physical))  
  
## corrPlot to examine the relationships among variables  
MentalHealth\_Data\_2 <- subset(raw, select = c(benefits, care\_options, wellness\_program, seek\_help, anonymity, mental\_health\_consequence, phys\_health\_consequence, coworkers,supervisor, mental\_health\_interview, phys\_health\_interview, mental\_vs\_physical))  
M <- cor(MentalHealth\_Data\_2)  
corrplot(M, method = "number")



*Conclusion 2:*

From this corrPlot, there was strong relationship between *wellness\_program* and *seek\_help* whose correlation coefficient was 1. Since *seek\_help* had a positive correlation (not significant) with other variables (i.e. *benefits*), this variable would be proper for a dependent variable.

# Preliminary Exploration

* Most of variables did not show strong correlation between them, except *seek\_help* and *wellness\_program*.
* Especially, *seek\_help* had a positive relation with other varialbe, such as *benefits*. Therefore, further analysis should be conducted with *seek\_help* as a dependent variable.