## Data Report Markdown

Group 5

* Ong Heng Kiat (17205923/1)
* Gan Joo Han (17206504/1)
* Mohammad Afif Bin Mohammad Fauzi (17197449/2)
* Nurul Asyiqin binti Mohd Harmizi (17065343/2)
* Muhammad Iqmal Hafiy Bin Hamdi (17077341/2)

### Introduction

Rapidly rising unemployment and an extremely low demand for labour are fundamental labour market problems currently facing all countries. Previous research has shown that the unemployment experience increases the risk of suffering from negative psychological outcomes, although, the effect appears to differ depending on life situation and individual characteristics. This heterogeneous effect of unemployment on mental health raises the question whether the labour market context, in the form of the local unemployment rate or vacancy rate, could mediate the effect in a similar manner. Such an effect of labour market conditions on the effect of unemployment on mental health, where the context affects the health effect, could be called, using Greenland and Morgenstern’s term, ecological effect modification.

### Problem Statement

What are the factors that will affect the employment rate? How much do these factors affect it? Why these factors can affect the employment rate so much?

### Data Acquisition

For this assignment, we will be implementing all the knowledge that we have learned to produce valuable insights using R, from data acquisition to building a simple web app using Shiny.

The data set we choose to be used for our question is a data set that is obtained through a paid research survey to explore the linkage between the mental illness and unemployment rate.

The data set has a sample size of 354 and 31 variables. Beside mental illness there are also lots of other useful features that may be the factors affecting employment rate such as the **education level**, **gender**, **age**, **body condition** and so on.

library("openxlsx")

## Warning: package 'openxlsx' was built under R version 3.6.3

#Set working directory  
setwd("C:/Users/user/Documents/R/Employment\_Rate\_Analysis\_WebApp")

However, there is a fly in the ointment which the data set is not a cleaned data set. It has duplicate observations, syntax error, not standardized format, missing records and so on which require some time to clean it. The data set is in the format of xlsx excel file.

### Data Extraction and Cleaning

We will extract several useful features like **salary**, **age**, **position**, **education level** and so on from the data set acquired into a form that is easier for further processing and analysis.

#Read the data sets into Rstudio  
data <- read.xlsx("data/raw\_dataset.xlsx")  
  
#Analysis the data  
View(data)  
class(data) #check type of class

## [1] "data.frame"

dim(data) #check row & column

## [1] 355 29

str(data) #Check the structure of the data sets

## 'data.frame': 355 obs. of 29 variables:  
## $ X1 : chr "1" "2" "3" "4" ...  
## $ Employed : num 1 0 1 1 0 1 1 1 1 1 ...  
## $ Mental.Illness : num 0 0 1 0 0 1 0 0 1 0 ...  
## $ Education : chr "Undergraduate" "Highschool" "Phd" "Undergraduate" ...  
## $ Have.Computer : chr "1" "0" "1" "1" ...  
## $ Days.Hopitalized.for.Mental.illeness : num 0 0 0 0 0 35 0 0 0 0 ...  
## $ Legally.Disabled : chr "0" "0" "0" "0" ...  
## $ Internet.Access : chr "1" "1" "1" "1" ...  
## $ Live.With.Parents : chr "0" "0" "0" "0" ...  
## $ Gap.In.Resume..Months. : num 36 24 1 0 11 33 0 0 0 0 ...  
## $ Total.Annual.Income..USD. : num 42 35 22 100 0 32 0 1 11 73 ...  
## $ Study : num 0 1 1 1 1 1 1 1 1 1 ...  
## $ Annual.income.from.social.welfare.programs..USD.: num 0 0 0 0 0 30 0 0 0 0 ...  
## $ Received.Food.Stamps : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Section.8.Housing : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ Frequency.Hospitalized : num 0 0 0 0 0 4 0 0 0 0 ...  
## $ Lack.of.concentration : num 0 1 1 0 0 1 0 0 1 1 ...  
## $ Anxiety : chr "0" "1" "1" "0" ...  
## $ Depression : num 0 1 1 0 0 1 0 0 1 1 ...  
## $ Obsessive.thinking : num 0 1 0 0 0 1 0 0 0 0 ...  
## $ Mood.swings : num 0 0 0 0 0 1 0 0 0 0 ...  
## $ Panic.attacks : num 0 1 1 0 0 1 0 0 1 0 ...  
## $ Compulsive.behavior : num 0 0 0 0 0 1 0 0 1 0 ...  
## $ Tiredness : chr "0" "0" "1" "0" ...  
## $ Age : chr "18-29" "30-44" "18-29" "30-44" ...  
## $ Gender : chr "Female" "Male" "Male" "Male" ...  
## $ Household.Income : chr "$25,000-$49,999" "$25,000-$49,999" "$50,000-$74,999" "$150,000-$174,999" ...  
## $ Region : chr "Pacific" "Mountain" "East South Central" "Pacific" ...  
## $ Device.Type : chr "iOS Phone / Tablet" "Android Phone / Tablet" "MacOS Desktop / Laptop" "MacOS Desktop / Laptop" ...

summary(data)

## X1 Employed Mental.Illness Education   
## Length:355 Min. :0.0000 Min. :0.0000 Length:355   
## Class :character 1st Qu.:0.0000 1st Qu.:0.0000 Class :character   
## Mode :character Median :1.0000 Median :0.0000 Mode :character   
## Mean :0.6845 Mean :0.2394   
## 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000   
## Have.Computer Days.Hopitalized.for.Mental.illeness Legally.Disabled   
## Length:355 Min. : 0.000 Length:355   
## Class :character 1st Qu.: 0.000 Class :character   
## Mode :character Median : 0.000 Mode :character   
## Mean : 2.769   
## 3rd Qu.: 0.000   
## Max. :100.000   
## Internet.Access Live.With.Parents Gap.In.Resume..Months.  
## Length:355 Length:355 Min. : 0.000   
## Class :character Class :character 1st Qu.: 0.000   
## Mode :character Mode :character Median : 0.000   
## Mean : 8.349   
## 3rd Qu.: 3.500   
## Max. :100.000   
## Total.Annual.Income..USD. Study   
## Min. : 0.00 Min. :0.0000   
## 1st Qu.: 14.00 1st Qu.:1.0000   
## Median : 31.00 Median :1.0000   
## Mean : 38.23 Mean :0.8873   
## 3rd Qu.: 55.00 3rd Qu.:1.0000   
## Max. :100.00 Max. :1.0000   
## Annual.income.from.social.welfare.programs..USD. Received.Food.Stamps  
## Min. : 0.00 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.00 Median :0.00000   
## Mean : 3.38 Mean :0.06479   
## 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. :100.00 Max. :1.00000   
## Section.8.Housing Frequency.Hospitalized Lack.of.concentration  
## Min. :0.00000 Min. : 0.000 Min. :0.0000   
## 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.0000   
## Median :0.00000 Median : 0.000 Median :0.0000   
## Mean :0.01972 Mean : 1.141 Mean :0.1521   
## 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:0.0000   
## Max. :1.00000 Max. :100.000 Max. :1.0000   
## Anxiety Depression Obsessive.thinking Mood.swings   
## Length:355 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## Class :character 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Mode :character Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.2535 Mean :0.1296 Mean :0.1211   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Panic.attacks Compulsive.behavior Tiredness Age   
## Min. :0.0000 Min. :0.00000 Length:355 Length:355   
## 1st Qu.:0.0000 1st Qu.:0.00000 Class :character Class :character   
## Median :0.0000 Median :0.00000 Mode :character Mode :character   
## Mean :0.1493 Mean :0.08451   
## 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.00000   
## Gender Household.Income Region Device.Type   
## Length:355 Length:355 Length:355 Length:355   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##

Then, we transform the data into a more beautiful data in such a way that it would be easier to be managed and also analyzed. This process includes detecting and correcting or removing the unwanted records such as duplicate records and inaccurate records.

###Cleaning process  
#Convert Data Type   
data$NA.<- NULL #remove entire column 1  
data$Employed <- as.logical(as.integer(as.character(data$Employed)))  
data$Mental.Illness <- as.logical(as.integer(as.character(data$Mental.Illness)))  
data$Education <- as.character(data$Education)  
data$Have.Computer <- as.logical(as.integer(as.character(data$Have.Computer)))  
data$Legally.Disabled <- as.logical(as.integer(as.character(data$Legally.Disabled)))  
data$Internet.Access <- as.logical(as.integer(as.character(data$Internet.Access)))  
data$Live.With.Parents <- as.logical(as.integer(as.character(data$Live.With.Parents)))  
data$Study <- as.logical(as.integer(as.character(data$Study)))  
data$Received.Food.Stamps <- as.logical(data$Received.Food.Stamps)  
data$Section.8.Housing <- as.logical(data$Section.8.Housing)  
data$Lack.of.concentration <- as.logical(data$Lack.of.concentration)  
data$Anxiety <- as.logical(as.integer(as.character(data$Anxiety)))  
data$Depression <- as.logical(data$Depression)  
data$Obsessive.thinking <- as.logical(data$Obsessive.thinking)  
data$Mood.swings <- as.logical(data$Mood.swings)  
data$Panic.attacks <- as.logical(data$Panic.attacks)  
data$Compulsive.behavior <- as.logical(data$Compulsive.behavior)  
data$Tiredness <- as.logical(as.integer(as.character(data$Tiredness)))  
  
#convert data type  
data$Age <- as.character(data$Age)  
data$Gender <- as.character(data$Gender)  
data$Household.Income <- as.character(data$Household.Income)  
data$Region <- as.character(data$Region)  
data$Device.Type <- as.character(data$Device.Type)  
  
# Remove duplicate data rows  
data <- unique(data)  
  
#Make all lowercase in variable Education  
data$Education <- tolower(data$Education)

Next, we will remove all the rows that have the null values (which is denoted by “na”), so that the data we get more refined and cleaned. We also remove any columns that is unnecessary to what we’re looking for.

###Replacing Na / Null value  
  
#Check number of null data before removing  
sum(is.na(data))

## [1] 10

#Method 1  
#Replace all null value with Mode by referring to summary()  
summary(data)

## X1 Employed Mental.Illness Education   
## Length:354 Mode :logical Mode :logical Length:354   
## Class :character FALSE:112 FALSE:270 Class :character   
## Mode :character TRUE :242 TRUE :84 Mode :character   
##   
##   
##   
## Have.Computer Days.Hopitalized.for.Mental.illeness Legally.Disabled  
## Mode :logical Min. : 0.000 Mode :logical   
## FALSE:44 1st Qu.: 0.000 FALSE:315   
## TRUE :309 Median : 0.000 TRUE :36   
## NA's :1 Mean : 2.777 NA's :3   
## 3rd Qu.: 0.000   
## Max. :100.000   
## Internet.Access Live.With.Parents Gap.In.Resume..Months.  
## Mode :logical Mode :logical Min. : 0.000   
## FALSE:12 FALSE:314 1st Qu.: 0.000   
## TRUE :341 TRUE :38 Median : 0.000   
## NA's :1 NA's :2 Mean : 8.373   
## 3rd Qu.: 3.750   
## Max. :100.000   
## Total.Annual.Income..USD. Study   
## Min. : 0.00 Mode :logical   
## 1st Qu.: 14.00 FALSE:40   
## Median : 31.50 TRUE :314   
## Mean : 38.27   
## 3rd Qu.: 55.00   
## Max. :100.00   
## Annual.income.from.social.welfare.programs..USD. Received.Food.Stamps  
## Min. : 0.00 Mode :logical   
## 1st Qu.: 0.00 FALSE:331   
## Median : 0.00 TRUE :23   
## Mean : 3.39   
## 3rd Qu.: 0.00   
## Max. :100.00   
## Section.8.Housing Frequency.Hospitalized Lack.of.concentration Anxiety   
## Mode :logical Min. : 0.000 Mode :logical Mode :logical   
## FALSE:347 1st Qu.: 0.000 FALSE:300 FALSE:249   
## TRUE :7 Median : 0.000 TRUE :54 TRUE :104   
## Mean : 1.144 NA's :1   
## 3rd Qu.: 0.000   
## Max. :100.000   
## Depression Obsessive.thinking Mood.swings Panic.attacks   
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:265 FALSE:309 FALSE:312 FALSE:302   
## TRUE :89 TRUE :45 TRUE :42 TRUE :52   
##   
##   
##   
## Compulsive.behavior Tiredness Age Gender   
## Mode :logical Mode :logical Length:354 Length:354   
## FALSE:324 FALSE:248 Class :character Class :character   
## TRUE :30 TRUE :105 Mode :character Mode :character   
## NA's :1   
##   
##   
## Household.Income Region Device.Type   
## Length:354 Length:354 Length:354   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

data$Have.Computer[is.na(data$Have.Computer)] <- TRUE  
data$Legally.Disabled[is.na(data$Legally.Disabled)] <- FALSE  
data$Internet.Access[is.na(data$Internet.Access)] <- TRUE  
data$Live.With.Parents[is.na(data$Live.With.Parents)] <- FALSE  
data$Anxiety[is.na(data$Anxiety)]<-FALSE  
data$Tiredness[is.na(data$Tiredness)] <- FALSE  
  
#Method 2  
#Simply delete column with null value   
#data <- na.omit(data)   
  
#Check number of null data after removing  
sum(is.na(data))

## [1] 1

#Analysis the data  
str(data)

## 'data.frame': 354 obs. of 29 variables:  
## $ X1 : chr "1" "2" "3" "4" ...  
## $ Employed : logi TRUE FALSE TRUE TRUE FALSE TRUE ...  
## $ Mental.Illness : logi FALSE FALSE TRUE FALSE FALSE TRUE ...  
## $ Education : chr "undergraduate" "highschool" "phd" "undergraduate" ...  
## $ Have.Computer : logi TRUE FALSE TRUE TRUE TRUE TRUE ...  
## $ Days.Hopitalized.for.Mental.illeness : num 0 0 0 0 0 35 0 0 0 0 ...  
## $ Legally.Disabled : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ Internet.Access : logi TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ Live.With.Parents : logi FALSE FALSE FALSE FALSE TRUE FALSE ...  
## $ Gap.In.Resume..Months. : num 36 24 1 0 11 33 0 0 0 0 ...  
## $ Total.Annual.Income..USD. : num 42 35 22 100 0 32 0 1 11 73 ...  
## $ Study : logi FALSE TRUE TRUE TRUE TRUE TRUE ...  
## $ Annual.income.from.social.welfare.programs..USD.: num 0 0 0 0 0 30 0 0 0 0 ...  
## $ Received.Food.Stamps : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Section.8.Housing : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Frequency.Hospitalized : num 0 0 0 0 0 4 0 0 0 0 ...  
## $ Lack.of.concentration : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Anxiety : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Depression : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Obsessive.thinking : logi FALSE TRUE FALSE FALSE FALSE TRUE ...  
## $ Mood.swings : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ Panic.attacks : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Compulsive.behavior : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ Tiredness : logi FALSE FALSE TRUE FALSE FALSE TRUE ...  
## $ Age : chr "18-29" "30-44" "18-29" "30-44" ...  
## $ Gender : chr "Female" "Male" "Male" "Male" ...  
## $ Household.Income : chr "$25,000-$49,999" "$25,000-$49,999" "$50,000-$74,999" "$150,000-$174,999" ...  
## $ Region : chr "Pacific" "Mountain" "East South Central" "Pacific" ...  
## $ Device.Type : chr "iOS Phone / Tablet" "Android Phone / Tablet" "MacOS Desktop / Laptop" "MacOS Desktop / Laptop" ...

dim(data)

## [1] 354 29

summary(data)

## X1 Employed Mental.Illness Education   
## Length:354 Mode :logical Mode :logical Length:354   
## Class :character FALSE:112 FALSE:270 Class :character   
## Mode :character TRUE :242 TRUE :84 Mode :character   
##   
##   
##   
## Have.Computer Days.Hopitalized.for.Mental.illeness Legally.Disabled  
## Mode :logical Min. : 0.000 Mode :logical   
## FALSE:44 1st Qu.: 0.000 FALSE:318   
## TRUE :310 Median : 0.000 TRUE :36   
## Mean : 2.777   
## 3rd Qu.: 0.000   
## Max. :100.000   
## Internet.Access Live.With.Parents Gap.In.Resume..Months.  
## Mode :logical Mode :logical Min. : 0.000   
## FALSE:12 FALSE:316 1st Qu.: 0.000   
## TRUE :342 TRUE :38 Median : 0.000   
## Mean : 8.373   
## 3rd Qu.: 3.750   
## Max. :100.000   
## Total.Annual.Income..USD. Study   
## Min. : 0.00 Mode :logical   
## 1st Qu.: 14.00 FALSE:40   
## Median : 31.50 TRUE :314   
## Mean : 38.27   
## 3rd Qu.: 55.00   
## Max. :100.00   
## Annual.income.from.social.welfare.programs..USD. Received.Food.Stamps  
## Min. : 0.00 Mode :logical   
## 1st Qu.: 0.00 FALSE:331   
## Median : 0.00 TRUE :23   
## Mean : 3.39   
## 3rd Qu.: 0.00   
## Max. :100.00   
## Section.8.Housing Frequency.Hospitalized Lack.of.concentration Anxiety   
## Mode :logical Min. : 0.000 Mode :logical Mode :logical   
## FALSE:347 1st Qu.: 0.000 FALSE:300 FALSE:250   
## TRUE :7 Median : 0.000 TRUE :54 TRUE :104   
## Mean : 1.144   
## 3rd Qu.: 0.000   
## Max. :100.000   
## Depression Obsessive.thinking Mood.swings Panic.attacks   
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:265 FALSE:309 FALSE:312 FALSE:302   
## TRUE :89 TRUE :45 TRUE :42 TRUE :52   
##   
##   
##   
## Compulsive.behavior Tiredness Age Gender   
## Mode :logical Mode :logical Length:354 Length:354   
## FALSE:324 FALSE:249 Class :character Class :character   
## TRUE :30 TRUE :105 Mode :character Mode :character   
##   
##   
##   
## Household.Income Region Device.Type   
## Length:354 Length:354 Length:354   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

#List the name of column  
colnames(data)

## [1] "X1"   
## [2] "Employed"   
## [3] "Mental.Illness"   
## [4] "Education"   
## [5] "Have.Computer"   
## [6] "Days.Hopitalized.for.Mental.illeness"   
## [7] "Legally.Disabled"   
## [8] "Internet.Access"   
## [9] "Live.With.Parents"   
## [10] "Gap.In.Resume..Months."   
## [11] "Total.Annual.Income..USD."   
## [12] "Study"   
## [13] "Annual.income.from.social.welfare.programs..USD."  
## [14] "Received.Food.Stamps"   
## [15] "Section.8.Housing"   
## [16] "Frequency.Hospitalized"   
## [17] "Lack.of.concentration"   
## [18] "Anxiety"   
## [19] "Depression"   
## [20] "Obsessive.thinking"   
## [21] "Mood.swings"   
## [22] "Panic.attacks"   
## [23] "Compulsive.behavior"   
## [24] "Tiredness"   
## [25] "Age"   
## [26] "Gender"   
## [27] "Household.Income"   
## [28] "Region"   
## [29] "Device.Type"

Finally, we renamed the columns so that it is readable for users to understand.

#Renaming column  
  
names(data)[names(data) == "Mental.Illness" ] <- "Mental Illness"  
names(data)[names(data) == "Education" ] <- "Education"  
names(data)[names(data) == "Have.Computer" ] <- "Own Computer"  
names(data)[names(data) == "I.have.been.hospitalized.before.for.my.mental.illness" ] <- "Hospitalized"  
names(data)[names(data) == "Days.Hopitalized.for.Mental.illeness" ] <- "Days Hospitalized for Mental Illness"  
names(data)[names(data) == "Legally.Disabled" ] <- "Disabled"  
names(data)[names(data) == "Internet.Access" ] <- "Internet Access"  
names(data)[names(data) == "Live.With.Parents" ] <- "Live With Parents"  
names(data)[names(data) == "I.have.a.gap.in.my.resume" ] <- "GapInResume"  
names(data)[names(data) == "Gap.In.Resume..Months." ] <- "Gaps In Resume(in a Month)"  
names(data)[names(data) == "Total.Annual.Income..USD." ] <- "Annual Income (USD)"  
names(data)[names(data) == "Annual.income.from.social.welfare.programs" ] <- "Annual Income from Social Welfare programs (USD)"  
names(data)[names(data) == "Received.Food.Stamps" ] <- "Receive Food Stamps"  
names(data)[names(data) == "Section.8.Housing" ] <- "Section 8 Housing"  
names(data)[names(data) == "Frequency.Hospitalized" ] <- "Frequency Hospitalized"  
names(data)[names(data) == "Lack.of.concentration" ] <- "Lack of Concentration"  
names(data)[names(data) == "Obsessive.thinking" ] <- "Obsessive Thinking"  
names(data)[names(data) == "Mood.swings" ] <- "Mood Swings"  
names(data)[names(data) == "Panic.attacks" ] <- "Panic Attacks"  
names(data)[names(data) == "Compulsive.behavior" ] <- "Compulsive Behavior"  
names(data)[names(data) == "Household.Income" ] <- "Household Income"  
names(data)[names(data) == "Device.Type" ] <- "Device Type"  
  
#Write new datasets (after cleaning) to new xlsx file  
write.xlsx(data,"data/CleanedData.xlsx",sheetName = "Sheet1" , col.names = TRUE, row.names = TRUE, append = FALSE, showNA = TRUE)

### Data Analysis

Now, we can visualize and analyze the data with various tools like R, and develop an algorithm to track the trends and the relationships between the features of the data.

#read the file  
mydata <- read.xlsx("data/CleanedData.xlsx")  
print(str(mydata))

## 'data.frame': 354 obs. of 30 variables:  
## $ : chr "1" "2" "3" "4" ...  
## $ X1 : chr "1" "2" "3" "4" ...  
## $ Employed : logi TRUE FALSE TRUE TRUE FALSE TRUE ...  
## $ Mental.Illness : logi FALSE FALSE TRUE FALSE FALSE TRUE ...  
## $ Education : chr "undergraduate" "highschool" "phd" "undergraduate" ...  
## $ Own.Computer : logi TRUE FALSE TRUE TRUE TRUE TRUE ...  
## $ Days.Hospitalized.for.Mental.Illness : num 0 0 0 0 0 35 0 0 0 0 ...  
## $ Disabled : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ Internet.Access : logi TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ Live.With.Parents : logi FALSE FALSE FALSE FALSE TRUE FALSE ...  
## $ Gaps.In.Resume(in.a.Month) : num 36 24 1 0 11 33 0 0 0 0 ...  
## $ Annual.Income.(USD) : num 42 35 22 100 0 32 0 1 11 73 ...  
## $ Study : logi FALSE TRUE TRUE TRUE TRUE TRUE ...  
## $ Annual.income.from.social.welfare.programs..USD.: num 0 0 0 0 0 30 0 0 0 0 ...  
## $ Receive.Food.Stamps : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Section.8.Housing : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Frequency.Hospitalized : num 0 0 0 0 0 4 0 0 0 0 ...  
## $ Lack.of.Concentration : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Anxiety : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Depression : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Obsessive.Thinking : logi FALSE TRUE FALSE FALSE FALSE TRUE ...  
## $ Mood.Swings : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ Panic.Attacks : logi FALSE TRUE TRUE FALSE FALSE TRUE ...  
## $ Compulsive.Behavior : logi FALSE FALSE FALSE FALSE FALSE TRUE ...  
## $ Tiredness : logi FALSE FALSE TRUE FALSE FALSE TRUE ...  
## $ Age : chr "18-29" "30-44" "18-29" "30-44" ...  
## $ Gender : chr "Female" "Male" "Male" "Male" ...  
## $ Household.Income : chr "$25,000-$49,999" "$25,000-$49,999" "$50,000-$74,999" "$150,000-$174,999" ...  
## $ Region : chr "Pacific" "Mountain" "East South Central" "Pacific" ...  
## $ Device.Type : chr "iOS Phone / Tablet" "Android Phone / Tablet" "MacOS Desktop / Laptop" "MacOS Desktop / Laptop" ...  
## NULL

#change the type into factor type  
mydata$Employed <- as.factor(mydata$Employed)  
mydata$Mental.Illness <- as.factor(mydata$Mental.Illness)  
mydata$Own.Computer <- as.factor(mydata$Own.Computer)  
mydata$Disabled <- as.factor(mydata$Disabled)  
mydata$Internet.Access <- as.factor(mydata$Internet.Access)  
mydata$Live.With.Parents <- as.factor(mydata$Live.With.Parents)  
mydata$Study <- as.factor(mydata$Study)  
mydata$Receive.Food.Stamps <- as.factor(mydata$Receive.Food.Stamps)  
mydata$Section.8.Housing <- as.factor(mydata$Section.8.Housing)  
mydata$Lack.of.Concentration <- as.factor(mydata$Lack.of.Concentration)  
mydata$Anxiety <- as.factor(mydata$Anxiety)  
mydata$Depression <- as.factor(mydata$Depression)  
mydata$Obsessive.Thinking <- as.factor(mydata$Obsessive.Thinking)  
mydata$Mood.Swings <- as.factor(mydata$Mood.Swings)  
mydata$Panic.Attacks <- as.factor(mydata$Panic.Attacks)  
mydata$Compulsive.Behavior <- as.factor(mydata$Compulsive.Behavior)  
mydata$Tiredness <- as.factor(mydata$Tiredness)  
mydata$Gender <- as.factor(mydata$Gender)  
mydata$Education<- as.factor(mydata$Education)  
mydata$Age<- as.factor(mydata$Age)  
mydata$Household.Income<- as.factor(mydata$Household.Income)  
mydata$Device.Type<- as.factor(mydata$Device.Type)  
print(str(mydata))

## 'data.frame': 354 obs. of 30 variables:  
## $ : chr "1" "2" "3" "4" ...  
## $ X1 : chr "1" "2" "3" "4" ...  
## $ Employed : Factor w/ 2 levels "FALSE","TRUE": 2 1 2 2 1 2 2 2 2 2 ...  
## $ Mental.Illness : Factor w/ 2 levels "FALSE","TRUE": 1 1 2 1 1 2 1 1 2 1 ...  
## $ Education : Factor w/ 4 levels "highschool","masters",..: 4 1 3 4 4 4 1 4 4 4 ...  
## $ Own.Computer : Factor w/ 2 levels "FALSE","TRUE": 2 1 2 2 2 2 2 2 2 2 ...  
## $ Days.Hospitalized.for.Mental.Illness : num 0 0 0 0 0 35 0 0 0 0 ...  
## $ Disabled : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 2 1 1 1 1 ...  
## $ Internet.Access : Factor w/ 2 levels "FALSE","TRUE": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Live.With.Parents : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 2 1 2 1 2 1 ...  
## $ Gaps.In.Resume(in.a.Month) : num 36 24 1 0 11 33 0 0 0 0 ...  
## $ Annual.Income.(USD) : num 42 35 22 100 0 32 0 1 11 73 ...  
## $ Study : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 2 2 2 2 2 2 2 ...  
## $ Annual.income.from.social.welfare.programs..USD.: num 0 0 0 0 0 30 0 0 0 0 ...  
## $ Receive.Food.Stamps : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Section.8.Housing : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Frequency.Hospitalized : num 0 0 0 0 0 4 0 0 0 0 ...  
## $ Lack.of.Concentration : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 1 1 2 1 1 2 2 ...  
## $ Anxiety : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 1 1 2 1 1 2 2 ...  
## $ Depression : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 1 1 2 1 1 2 2 ...  
## $ Obsessive.Thinking : Factor w/ 2 levels "FALSE","TRUE": 1 2 1 1 1 2 1 1 1 1 ...  
## $ Mood.Swings : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 2 1 1 1 1 ...  
## $ Panic.Attacks : Factor w/ 2 levels "FALSE","TRUE": 1 2 2 1 1 2 1 1 2 1 ...  
## $ Compulsive.Behavior : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 2 1 1 2 1 ...  
## $ Tiredness : Factor w/ 2 levels "FALSE","TRUE": 1 1 2 1 1 2 1 1 2 2 ...  
## $ Age : Factor w/ 4 levels "> 60","18-29",..: 2 3 2 3 3 3 3 2 2 3 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 2 2 2 2 ...  
## $ Household.Income : Factor w/ 11 levels "$0-$9,999","$10,000-$24,999",..: 8 8 9 5 8 8 1 3 3 9 ...  
## $ Region : chr "Pacific" "Mountain" "East South Central" "Pacific" ...  
## $ Device.Type : Factor w/ 5 levels "Android Phone / Tablet",..: 2 1 3 3 5 2 1 5 5 2 ...  
## NULL

From the code above, the **as.factor()** in R is used to convert the data type of a variable to a factor/categorical variable. Now, we want to check if there is null values in the data frame.

nrow(mydata[is.na(mydata$Study)|is.na(mydata$Depression)] )

## [1] 354

mydata[is.na(mydata$Study)|is.na(mydata$Depression),]

## [1]   
## [2] X1   
## [3] Employed   
## [4] Mental.Illness   
## [5] Education   
## [6] Own.Computer   
## [7] Days.Hospitalized.for.Mental.Illness   
## [8] Disabled   
## [9] Internet.Access   
## [10] Live.With.Parents   
## [11] Gaps.In.Resume(in.a.Month)   
## [12] Annual.Income.(USD)   
## [13] Study   
## [14] Annual.income.from.social.welfare.programs..USD.  
## [15] Receive.Food.Stamps   
## [16] Section.8.Housing   
## [17] Frequency.Hospitalized   
## [18] Lack.of.Concentration   
## [19] Anxiety   
## [20] Depression   
## [21] Obsessive.Thinking   
## [22] Mood.Swings   
## [23] Panic.Attacks   
## [24] Compulsive.Behavior   
## [25] Tiredness   
## [26] Age   
## [27] Gender   
## [28] Household.Income   
## [29] Region   
## [30] Device.Type   
## <0 rows> (or 0-length row.names)

nrow(mydata)

## [1] 354

xtabs(~ Employed + Study, data=mydata)

## Study  
## Employed FALSE TRUE  
## FALSE 13 99  
## TRUE 27 215

unique(mydata$Education)

## [1] undergraduate highschool phd masters   
## Levels: highschool masters phd undergraduate

unique(mydata$Employed)

## [1] TRUE FALSE  
## Levels: FALSE TRUE

We will be split the data obtained to 2 parts:

* the training set, to train the model
* the test set, to test the model

We will be using the regression model to do prediction and produce values for analysis. We will use the **glm()** command to run logistic regression with **train\_set** table,and after the comma, we specify that the distribution is binomial.

We can access the model output using **summary()**.

After we run the code, we can see the summary of the **logistic\_model**, the **train\_matrix**, **train\_accuracy**, **test\_matrix** and **test\_accuracy**.

If we run the this chunk few times, we noticed that the values of train\_accuracy and test\_accuracy differs on each run. It is normal because the initial weight is randomly generated.

To get the same accuracy on each run, we use **set.seed()** function to set the starting number used to generate a sequence of random numbers. It ensures that we can get the same result if you start with that same seed each time you run the same process.

### Data Visualization using Shiny Web App

Using the web app, we can see many graphs and charts appeared, show us the correlation between all the factors that can be connected to the unemployment rate.

The link for the app :

<https://onghengkiat.shinyapps.io/employment_rate_analysis_webapp/>

After you open the link, you can see the Data Explorer For Employment Rate. We can do some correlation between the data we seek based on the graphs, table, bar and prediction model.

Here are some of the data obtained from observation.



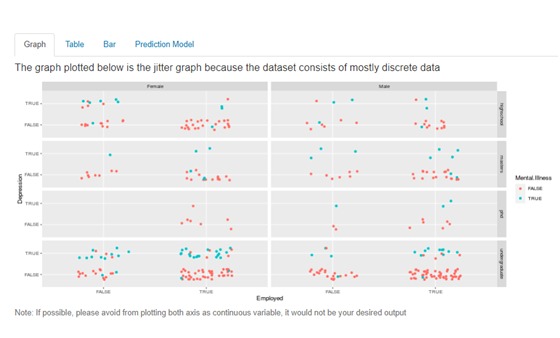
From the image above, we are trying to look for linkage between age and mental illness. There is an increase of mental illness from age 18 to 44 years old, and the bars began to drop at the age of 45 to more than age 60. About 29.63% of people in age 18-29 suffered with mental illness. It is due to the fact that this age group has started to further their studies and looking for the any job vacancies, ready to be employed. It showed the highest peak of mental illness at the age between 30 to 44 years old. It is because of the work pressures and mid-life crisis.



Now, we checked on the correlation between the employment and anxiety, separated by education and using only 100 samples. Please be noted that the ‘FALSE’ part is for people who are currently in their studies while the ‘TRUE’ part represents people that already completed their studies.

We can see that most of the Completed undergraduate students and Completed PhD students that already employed to their respective jobs said that anxiety are happened in their workplace. It is because that people cannot foresee all the obstacles they might encounter in their work, which affects attention and concentration.

We can see from scatter plots at the graph charts below, with all the samples, people that is employed and suffered mental illness is the more likely to be the completed undergraduate students. It is found that women is likely to get mental health problems such as depression than men.



### Conclusion

As we all know, mental health problems has become such a serious issue in 21st century, especially among youngsters and adults. We demonstrate that data science can be applied to find the factors that affect the employment rate . From the analysis and factors that we obtained, we can find the linkage between the employment rate and the mental illness. Although the prediction might be not accurate, we can still fix the problem with other approaches. The model need be trained with more data so that we can make the better accuracy for us to predict the outcome for data scientists to produce data insights, but there is much more to be done now and in the future.