

Analyze Mental health disorder in Tech Companies using R

Chitramoy Mukherjee

2023-02-16

Introduction :

In the document we are going to analyze, How does employees mental health affects employees performance and overall company productivity in tech companies.

Mental health affects your emotional, psychological and social well-being. Mental health is a key factor to determine the productivity of the employee as well as the total performance of the company. If someone is not mentally fit, he can't produce the expected output what he is capable of and it also impacts his co-workers performance and impacts the work environment.

This sort of analysis helps the employer to identify and support an individual who may be experiencing a mental health or substance use concern or crisis and connect them with the appropriate employee resources. This sort of analysis helps to increase mental health awareness in employee and co-workers. As a whole this sort of analysis helps employer to increase productivity and make sure they are providing a healthy work environment and takes care of it's employees mental health. This allows employer to recognize the signs of someone who maybe struggling and teaches them the skills to know when to reach out and what resources are available. Organizations that incorporate mental health awareness help to create a healthy and productive work environment that reduces the stigma associated with mental illness, increases the organizations mental health literacy and teaches the skills to safely and responsibly respond to a co-workers mental health concern.

This topic is relevant to data science as we can analyze and identify the factors/variables that impacts the mental health and justify the relations between variables which is closely related to determine the mental health of employees. We can create a model and feed data into it to identify the employees mental health in the company and provide directions to them to overcome the situation.

Research questions :

1. no_employees and tech_company are statistically significant predictors of mental health condition?
2. Does family_history have any relation with mentalhealth?
3. How does the frequency of mental health illness differ by workplace type (i.e., tech vs. non-tech) and size?
4. How easy is it for you to take medical leave for a mental health condition?
5. Is there any relationship between the responders location and the occurrence of mental illness?
6. Would you bring up a mental health issue with a potential employer in an interview?
7. Is the number of employees in the work place a factor of causing mental illness?
8. Which variable has a correlated relationship with the occurrence of mental illness?
9. Would you be willing to discuss a mental health issue with your coworkers?
10. Does the care_options in tech companies varies based on country?

Approach :

- First step for any analysis in data science is to view the data in the dataset and understand the structure of the imported data.
- Understand each attributes in the dataset.
- Data cleansing will be the next step and Null/NA data handling in the source dataset to make it ready for analysis.
- Outlier Analysis and Treatment. Aggregation functions are very useful for understanding the data and present its summarized picture.
- Identify the dependent and independent variable from the dataset.
- perform plotting using the differnt library functions.
- Compare and contrast predictive models using simple linear, multiple linear, and polynomial regression methods.
- Evaluate a model for overfitting and underfitting conditions and tune its performance using regularization and grid search.
- Create training and test dataset from source data to perform accuracy testing of the model.

Approach to address the problem :

- Prepare data for analysis by handling missing values, formatting and normalizing data, binning, and turning categorical values into numeric values.
- Examine data using descriptive statistics, data grouping, analysis of variance (ANOVA), and correlation statistics.
- Compare and contrast predictive models using simple linear, multiple linear, and polynomial regression methods.
- Evaluate a model for overfitting and underfitting conditions and tune its performance using regularization and grid search.

Datasets and Source of Data :

1. 3 datasets will be analyzed as a pert of this analysis which sourced from one dataset and splitted based on country.
 - survey_USA
 - survey_UK
 - survey_Others
2. Data sourced from : KAGGLE.

This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace.

no_employees column have date for some of the rows which is data impurities and needs to be cleaned.

```
setwd("C:/Users/14024/Desktop/dsc520-fork-chitro")

## Load the `2014 data' to
survey_USA <- read.csv("Final project/survey_USA.csv")

## Load the `2016 data' to
survey_UK <- read.csv("Final project/survey_UK.csv")

## Load the `Others data' to
survey_others <- read.csv("Final project/survey_others.csv")
```

Understanding dataset columns:

- Timestamp :
- Age : Age of the employee.
- Gender : Male or Female employee.
- Country : Employee working in which 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 primarily a tech company/organization?
- 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 mental health issues 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?
- mentalhealthconsequence: Do you think that discussing a mental health issue with your employer would have * negative consequences?
- physhealthconsequence: 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?
- physhealthinterview: Would you bring up a physical health issue with a potential employer in an interview?
- mentalvsphysical: Do you feel that your employer takes mental 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?
- comments: Any additional notes or comments

Required Packages :

```
library("ggplot2")# plot
library("plyr")# data cleanning
library("dplyr")# data cleanning
```

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

## The following objects are masked from 'package:plyr':
##
##   arrange, count, desc, failwith, id, mutate, rename, summarise,
##   summarize

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

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

```
library("magrittr")  
library("foreign")  
library("coefplot")  
library("caTools")  
library("rlang")
```

```
##  
## Attaching package: 'rlang'
```

```
## The following object is masked from 'package:magrittr':  
##  
## set_names
```

```
library("lazyeval")
```

```
##  
## Attaching package: 'lazyeval'
```

```
## The following objects are masked from 'package:rlang':  
##  
## as_name, call_modify, call_standardise, expr_label, expr_text,  
## f_env, f_env<-, f_label, f_lhs, f_lhs<-, f_rhs, f_rhs<-, f_text,  
## is_atomic, is_call, is_formula, is_lang, is_pairlist, missing_arg
```

```
library("gridExtra")
```

```
##  
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':  
##  
## combine
```

```
library("tidyr")
```

```
##  
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':  
##  
## extract
```

```
library("readr")  
library("purrr")
```

```
##
## Attaching package: 'purrr'

## The following objects are masked from 'package:lazyeval':
##
##   is_atomic, is_formula

## The following objects are masked from 'package:rlang':
##
##   %@%, flatten, flatten_chr, flatten_dbl, flatten_int, flatten_lgl,
##   flatten_raw, invoke, splice

## The following object is masked from 'package:magrittr':
##
##   set_names

## The following object is masked from 'package:plyr':
##
##   compact
```

```
library("tibble")
library("stringr")
library("forcats")
library("rworldmap")
```

```
## Loading required package: sp
```

```
## ### Welcome to rworldmap ###
```

```
## For a short introduction type : vignette('rworldmap')
```

```
##library('maps')# world map
##library("corrplot")# corrplot
##library("fiftystater")# us map
```

Plots :

- histogram
- boxplot
- Points plot
- bar plot
- line plot
- Points and line plot

Questions for future steps :

- Need to join all 3 datasets to analyze/answer questions across different countries.
- What level of data munging required to develop a model.
- What should be the split percentage of test and training data for the accuracy testing of the model.

Import and clean data :

- Downloaded the data from kaggle and kept in local machine. Split the 2014 into 3 different dataset based on the country value.
- Use read.csv function to read and import the data into different variables.
- Check which all columns have 'NA' and count of such rows in each dataset.
- As this data entered in the survey by employee and some of the columns data are not consistent which will be useful for analysis, make them consistent.
- Remove missing values Sometimes, people tend to not put a value while filling their survey. Since there is an abundance of data instances, it is not too much of a worry to drop them from our model.
- Detect anomalies (outliers) in numeric attributes, in this dataset age is only numeric variable. Some of the columns data is invalid which need to filter out before making the dataset final.

```
setwd("C:/Users/14024/Desktop/dsc520-fork-chitro")

## Load the `USA data' to
survey_USA <- read.csv("Final project/survey_USA.csv")
#head(survey_2014)
## Load the `UK data' to
survey_UK <- read.csv("Final project/survey_UK.csv")
#head(survey_UK)
## Load the `Others data' to
survey_Others <- read.csv("Final project/survey_Others.csv")
#head(survey_Others)

## Display internal structure of survey_USA and Columns contains 'NA' value
str(survey_USA)
```

```
## 'data.frame': 751 obs. of 27 variables:
## $ Timestamp : chr "8/27/2014 11:29" "8/27/2014 11:29" "8/27/2014 11:30" "8/27/2014 11:30" ...
## $ Age : int 37 44 31 33 35 42 31 42 36 29 ...
## $ Gender : chr "Female" "M" "Male" "Male" ...
## $ Country : chr "United States" "United States" "United States" "United States" ...
## $ state : chr "IL" "IN" "TX" "TN" ...
## $ self_employed : chr NA NA NA NA ...
## $ family_history : chr "No" "No" "No" "Yes" ...
## $ treatment : chr "Yes" "No" "No" "No" ...
## $ work_interfere : chr "Often" "Rarely" "Never" "Sometimes" ...
## $ no_employees : chr "25-Jun" "More than 1000" "100-500" "25-Jun" ...
## $ remote_work : chr "No" "No" "Yes" "No" ...
## $ tech_company : chr "Yes" "No" "Yes" "Yes" ...
## $ benefits : chr "Yes" "Don't know" "Yes" "Yes" ...
## $ care_options : chr "Not sure" "No" "No" "Not sure" ...
## $ wellness_program : chr "No" "Don't know" "Don't know" "No" ...
## $ seek_help : chr "Yes" "Don't know" "Don't know" "Don't know" ...
## $ anonymity : chr "Yes" "Don't know" "Don't know" "Don't know" ...
## $ leave : chr "Somewhat easy" "Don't know" "Don't know" "Don't know" ...
## $ mental_health_consequence : chr "No" "Maybe" "No" "No" ...
## $ phys_health_consequence : chr "No" "No" "No" "No" ...
## $ coworkers : chr "Some of them" "No" "Some of them" "Yes" ...
## $ supervisor : chr "Yes" "No" "Yes" "Yes" ...
## $ mental_health_interview : chr "No" "No" "Yes" "No" ...
## $ phys_health_interview : chr "Maybe" "No" "Yes" "Maybe" ...
```

```
## $ mental_vs_physical      : chr  "Yes" "Don't know" "Don't know" "Don't know" ...
## $ obs_consequence         : chr  "No" "No" "No" "No" ...
## $ comments                 : chr  NA NA NA NA ...
```

```
sapply(survey_USA, function(x) sum(is.na(x)))
```

```
##          Timestamp          Age          Gender
##          0              0              0
##          Country          state          self_employed
##          0              11              11
##          family_history    treatment    work_interfere
##          0              0              144
##          no_employees      remote_work  tech_company
##          0              0              0
##          benefits          care_options  wellness_program
##          0              0              0
##          seek_help          anonymity    leave
##          0              0              0
## mental_health_consequence  phys_health_consequence  coworkers
##          0              0              0
##          supervisor        mental_health_interview  phys_health_interview
##          0              0              0
##          mental_vs_physical  obs_consequence          comments
##          0              0              647
```

```
## Display internal structure of survey_UK and Columns contains 'NA' value
str(survey_UK)
```

```
## 'data.frame':   185 obs. of  27 variables:
## $ Timestamp      : chr  "8/27/2014 11:29" "8/27/2014 11:34" "8/27/2014 11:38" "8/27/2014 ...
## $ Age            : int  31 23 37 32 30 24 28 27 38 19 ...
## $ Gender         : chr  "Male" "Male" "Male" "Male" ...
## $ Country        : chr  "United Kingdom" "United Kingdom" "United Kingdom" "United Kingdom ...
## $ state          : logi  NA NA NA NA NA NA ...
## $ self_employed  : chr  NA NA "No" "No" ...
## $ family_history  : chr  "Yes" "No" "No" "No" ...
## $ treatment      : chr  "Yes" "Yes" "No" "No" ...
## $ work_interfere  : chr  "Often" "Sometimes" "Sometimes" "Never" ...
## $ no_employees   : chr  "26-100" "26-100" "25-Jun" "25-Jun" ...
## $ remote_work    : chr  "No" "Yes" "No" "Yes" ...
## $ tech_company   : chr  "Yes" "Yes" "Yes" "Yes" ...
## $ benefits       : chr  "No" "Don't know" "No" "No" ...
## $ care_options   : chr  "Yes" "No" "No" "No" ...
## $ wellness_program : chr  "No" "Don't know" "No" "No" ...
## $ seek_help      : chr  "No" "Don't know" "No" "No" ...
## $ anonymity      : chr  "No" "Don't know" "Don't know" "Don't know" ...
## $ leave          : chr  "Somewhat difficult" "Very easy" "Very difficult" "Don't know" ...
## $ mental_health_consequence : chr  "Yes" "Maybe" "Yes" "Yes" ...
## $ phys_health_consequence : chr  "Yes" "No" "Maybe" "Yes" ...
## $ coworkers       : chr  "Some of them" "Some of them" "Some of them" "Some of them" ...
## $ supervisor     : chr  "No" "No" "No" "Some of them" ...
## $ mental_health_interview : chr  "Maybe" "Maybe" "No" "No" ...
## $ phys_health_interview : chr  "Maybe" "Maybe" "Maybe" "Maybe" ...
```

```
## $ mental_vs_physical      : chr  "No" "No" "No" "No" ...
## $ obs_consequence         : chr  "Yes" "No" "No" "No" ...
## $ comments                 : chr  NA "My company does provide healthcare but not to me as I'm on a
```

```
sapply(survey_UK, function(x) sum(is.na(x)))
```

```
##          Timestamp          Age          Gender
##          0              0              0
##          Country          state          self_employed
##          0              185              2
##          family_history    treatment    work_interfere
##          0              0              44
##          no_employees      remote_work    tech_company
##          0              0              0
##          benefits          care_options    wellness_program
##          0              0              0
##          seek_help          anonymity      leave
##          0              0              0
## mental_health_consequence  phys_health_consequence  coworkers
##          0              0              0
##          supervisor    mental_health_interview    phys_health_interview
##          0              0              0
##          mental_vs_physical    obs_consequence    comments
##          0              0              163
```

```
## Display internal structure of survey_Others and Columns contains 'NA' value
str(survey_Others)
```

```
## 'data.frame':   323 obs. of  27 variables:
## $ Timestamp      : chr  "8/27/2014 11:29" "8/27/2014 11:32" "8/27/2014 11:32" "8/27/2014 ...
## $ Age            : num  32 39 23 29 27 36 29 38 27 18 ...
## $ Gender         : chr  "Male" "M" "Male" "male" ...
## $ Country        : chr  "Canada" "Canada" "Canada" "Bulgaria" ...
## $ state          : chr  NA NA NA NA ...
## $ self_employed  : chr  NA NA NA NA ...
## $ family_history  : chr  "No" "No" "No" "No" ...
## $ treatment      : chr  "No" "No" "No" "No" ...
## $ work_interfere  : chr  "Rarely" "Never" "Never" "Never" ...
## $ no_employees    : chr  "25-Jun" "5-Jan" "26-100" "100-500" ...
## $ remote_work     : chr  "No" "Yes" "No" "Yes" ...
## $ tech_company    : chr  "Yes" "Yes" "Yes" "Yes" ...
## $ benefits        : chr  "No" "No" "Don't know" "Don't know" ...
## $ care_options    : chr  "No" "Yes" "No" "Not sure" ...
## $ wellness_program : chr  "No" "No" "Don't know" "No" ...
## $ seek_help       : chr  "No" "No" "Don't know" "No" ...
## $ anonymity       : chr  "Don't know" "Yes" "Don't know" "Don't know" ...
## $ leave           : chr  "Somewhat difficult" "Don't know" "Don't know" "Don't know" ...
## $ mental_health_consequence: chr  "No" "No" "No" "No" ...
## $ phys_health_consequence : chr  "No" "No" "No" "No" ...
## $ coworkers       : chr  "Yes" "No" "Yes" "Yes" ...
## $ supervisor      : chr  "Yes" "No" "Yes" "Yes" ...
## $ mental_health_interview : chr  "Yes" "No" "Maybe" "Yes" ...
## $ phys_health_interview  : chr  "Yes" "No" "Maybe" "Yes" ...
```



```
## $ mental_vs_physical      : chr  "No" "No" "Yes" "Don't know" ...
## $ obs_consequence         : chr  "No" "No" "No" "No" ...
## $ comments                 : chr  NA NA NA NA ...
```

```
sapply(survey_Others, function(x) sum(is.na(x)))
```

```
##           Timestamp           Age           Gender
##           0             0             0
##           Country           state       self_employed
##           0             319             5
##           family_history     treatment   work_interfere
##           0             0             76
##           no_employees       remote_work tech_company
##           0             0             0
##           benefits          care_options wellness_program
##           0             0             0
##           seek_help          anonymity   leave
##           0             0             0
## mental_health_consequence   phys_health_consequence   coworkers
##           0             0             0
##           supervisor        mental_health_interview   phys_health_interview
##           0             0             0
##           mental_vs_physical   obs_consequence         comments
##           0             0             285
```

```
## checking the unique values in the Gender column in each dataset
unique(survey_USA$Gender)
```

```
## [1] "Female"      "M"           "Male"        "female"
## [5] "male"        "Male-ish"    "maile"       "Trans-female"
## [9] "Cis Female"  "F"          "Cis Male"    "m"
## [13] "f"          "Male (CIS)"  "queer/she/they" "non-binary"
## [17] "Femake"     "woman"       "Make"        "Nah"
## [21] "Genderqueer" "Female "     "Woman"       "cis-female/femme"
## [25] "Male "      "Trans woman" "Man"         "msle"
## [29] "Female (trans)" "Female (cis)" "Mail"        "cis male"
## [33] "p"          "femail"
```

```
unique(survey_UK$Gender)
```

```
## [1] "Male"
## [2] "male"
## [3] "M"
## [4] "Woman"
## [5] "Female"
## [6] "Enby"
## [7] "Androgyne"
## [8] "female"
## [9] "Agender"
## [10] "f"
## [11] "m"
## [12] "Neuter"
```

```
## [13] "F"
## [14] "Male "
## [15] "Cis Man"
## [16] "ostensibly male, unsure what that really means"
```

```
unique(survey_Others$Gender)
```

```
## [1] "Male"           "M"
## [3] "male"           "m"
## [5] "Female"         "something kinda male?"
## [7] "F"              "female"
## [9] "f"              "Mal"
## [11] "All"            "fluid"
## [13] "Guy (-ish) ^_^" "male leaning androgynous"
## [15] "Man"            "queer"
## [17] "A little about you" "Malr"
## [19] "Male "
```

```
## converting the column values to lowercase
```

```
survey_USA$Gender <- tolower(survey_USA$Gender)
survey_UK$Gender <- tolower(survey_UK$Gender)
survey_Others$Gender <- tolower(survey_Others$Gender)
```

```
## defining lists on the basis of unique values present in the dataset
```

```
male <- c('male-ish', 'cis male', 'male (cis)', 'make', 'mail',
          'ostensibly male, unsure what that really means', 'm', 'maile',
          'male ', 'msle', 'mal', 'man', 'malr', 'cis man', 'male')
female <- c('female', 'female ', 'female (cis)', 'woman', 'cis female',
            'cis-female/femme', 'femail', 'f', 'femake')
others <- c('p', 'nah', 'all', 'a little about you', 'genderqueer',
            'non-binary', 'trans woman', 'androgynous', 'neuter', 'trans-female',
            'agender', 'female (trans)', 'something kinda male?', 'enby', 'guy (-ish) ^_^',
            'queer', 'queer/she/they', 'male leaning androgynous', 'fluid', 'genderqueer',
            'non-binary', 'trans woman', 'androgynous', 'neuter', 'trans-female',
            'agender', 'female (trans)', 'something kinda male?', 'enby',
            'guy (-ish) ^_^', 'queer', 'queer/she/they', 'male leaning androgynous', 'fluid')
```

```
## replace the values in the data set so that the categories are consistent
```

```
survey_USA <- survey_USA %>%
  mutate(Gender = replace(Gender, which(Gender %in% male), 'Male'),
         Gender = replace(Gender, which(Gender %in% female), 'Female'),
         Gender = replace(Gender, which(Gender %in% others), 'Others'))
survey_UK <- survey_UK %>%
  mutate(Gender = replace(Gender, which(Gender %in% male), 'Male'),
         Gender = replace(Gender, which(Gender %in% female), 'Female'),
         Gender = replace(Gender, which(Gender %in% others), 'Others'))
survey_Others <- survey_Others %>%
  mutate(Gender = replace(Gender, which(Gender %in% male), 'Male'),
         Gender = replace(Gender, which(Gender %in% female), 'Female'),
         Gender = replace(Gender, which(Gender %in% others), 'Others'))
```

```
## checking if all values are covered and category is consistent
```

```
unique(survey_USA$Gender)
```

```
## [1] "Female" "Male" "Others"
```

```
unique(survey_UK$Gender)
```

```
## [1] "Male" "Female" "Others"
```

```
unique(survey_Others$Gender)
```

```
## [1] "Male" "Female" "Others"
```

```
## checking the summary of the Age column
```

```
summary(survey_USA$Age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -29.00  27.50   32.00   33.33  37.50  329.00
```

```
summary(survey_UK$Age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1726.00  26.00   31.00   21.44  34.00   55.00
```

```
summary(survey_Others$Age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 8.000e+00 2.600e+01 2.900e+01 3.096e+08 3.400e+01 1.000e+11
```

```
## age have some negative and more than 100 values which needs to be cleaned as age -ve is  
## invalid and more than can be considered as invalid  
## keeping rows with age between 18 and 100
```

```
survey_USA_df <- survey_USA %>% filter(Age >= 18 & Age < 100)
```

```
survey_UK_df <- survey_UK %>% filter(Age >= 18 & Age < 100)
```

```
survey_Others_df <- survey_Others %>% filter(Age >= 18 & Age < 100)
```

```
survey_USA_final_df <- survey_USA_df %>% select(-c(Timestamp,comments))
```

```
survey_UK_final_df <- survey_UK_df %>% select(-c(Timestamp,comments))
```

```
survey_Others_final_df <- survey_Others_df %>% select(-c(Timestamp,comments))
```

What does the final data set look like :

```
survey_data <- rbind(survey_USA_final_df, survey_UK_final_df, survey_Others_final_df)  
survey_data %>% dim
```

```
## [1] 1251 25
```

```
summary(survey_data)
```

```
##      Age      Gender      Country      state
## Min.   :18.00  Length:1251    Length:1251    Length:1251
## 1st Qu.:27.00  Class :character  Class :character  Class :character
## Median :31.00  Mode  :character  Mode  :character  Mode  :character
## Mean   :32.08
## 3rd Qu.:36.00
## Max.   :72.00
## self_employed  family_history  treatment  work_interfere
## Length:1251    Length:1251    Length:1251    Length:1251
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
## no_employees  remote_work  tech_company  benefits
## Length:1251    Length:1251    Length:1251    Length:1251
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
## care_options  wellness_program  seek_help  anonymity
## Length:1251    Length:1251    Length:1251    Length:1251
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
## leave  mental_health_consequence  phys_health_consequence
## Length:1251    Length:1251    Length:1251
## Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character
##
##
## coworkers  supervisor  mental_health_interview
## Length:1251    Length:1251    Length:1251
## Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character
##
##
## phys_health_interview  mental_vs_physical  obs_consequence
## Length:1251    Length:1251    Length:1251
## Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character
##
##
##
```

```
summarise(survey_data, min_age = min(Age))
```

```
##   min_age  
## 1      18
```

```
summarise(survey_data, max_age = max(Age))
```

```
##   max_age  
## 1      72
```

```
summarise(survey_data, med = median(Age))
```

```
##   med  
## 1   31
```

```
state_freq <- as.data.frame(table(survey_data$state))  
colnames(state_freq) <- c("state", "freq")  
as_tibble(state_freq)
```

```
## # A tibble: 45 x 2  
##   state freq  
##   <fct> <int>  
## 1 AL      7  
## 2 AZ      7  
## 3 CA     138  
## 4 CO      9  
## 5 CT      4  
## 6 DC      4  
## 7 FL     15  
## 8 GA     12  
## 9 IA      4  
## 10 ID     1  
## # ... with 35 more rows
```

What information is not self-evident :

- This data is entered by employee during survey, so some of the columns such as family_history, mentalhealthconsequence and physhealthconsequence might not be self evident as some of the employees might think they might have some -ve consequences for providing this info in survey.

Questions for future steps :

- Binary Variables are treatment, family_history and Tech_company in the dataset and No_employees is categorical variable. Need to use this for linear and logistic model design.
- Based on which column we should group the data from each dataset which will provide relevant answer to the research questions.
- Identify the variables from dataset which have linear relationship between them.
- Identify the variables which will have binomial output and can be used for logistic regression.

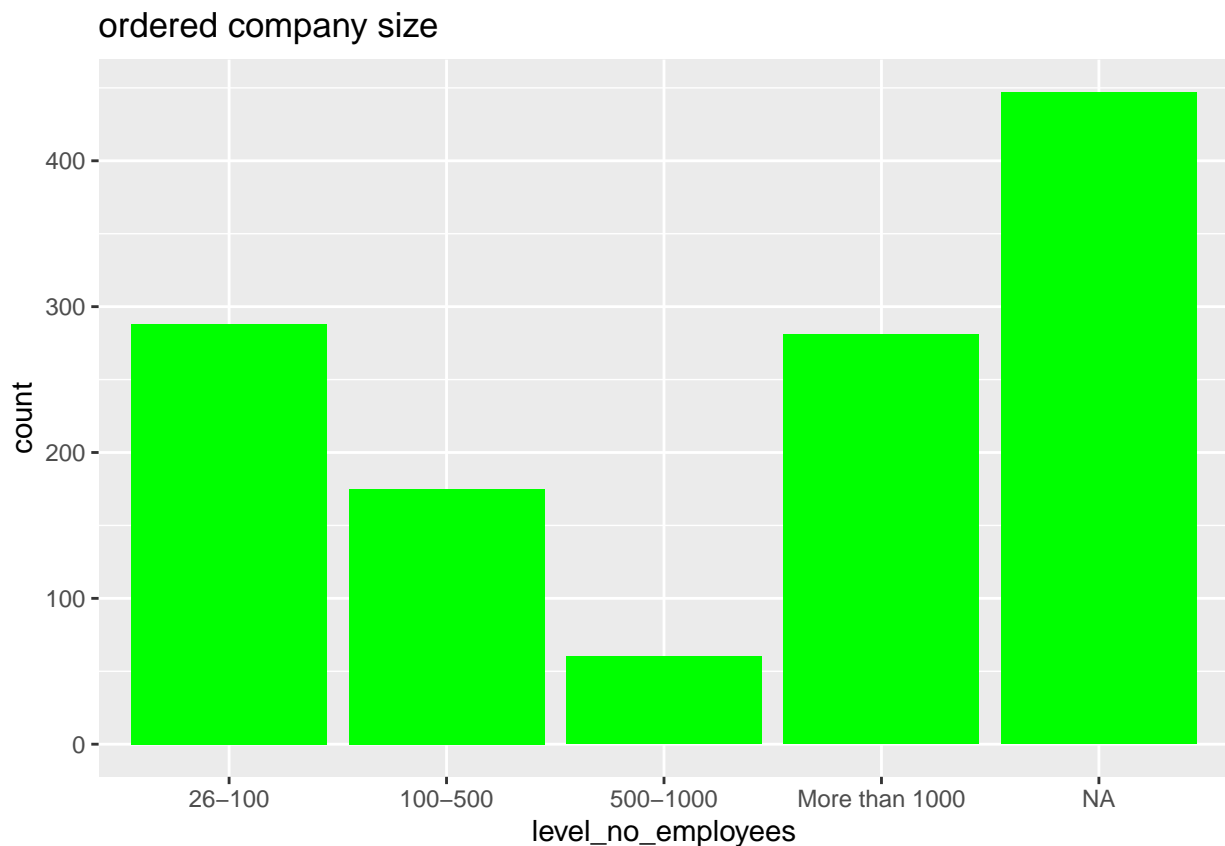
Different ways to look at data and slicing and dicing of data with visualization :

- Visualize based of no_employees in company
- Visualize based on gender
- Visualize based on age

```
## view the specific number of company in each size range
survey_data$no_employees %>% table
```

```
## .
##      100-500      25-Jun      26-100      5-Jan      500-1000
##      175      289      288      158      60
## More than 1000
##      281
```

```
## ordered no_employees level
level_no_employees <- factor(survey_data$no_employees, levels =
  c("1-5", "6-25", "26-100", "100-500", "500-1000", "More than 1000"))
## view the company distribution based on company size
survey_data %>% ggplot(aes(x=level_no_employees))+
  geom_bar(fill = "green") + ggtitle("ordered company size")
```



```
gender_diversity <- survey_data %>%
  group_by(Gender) %>%
```

```
dplyr::summarize(count = n())
gender_diversity
```

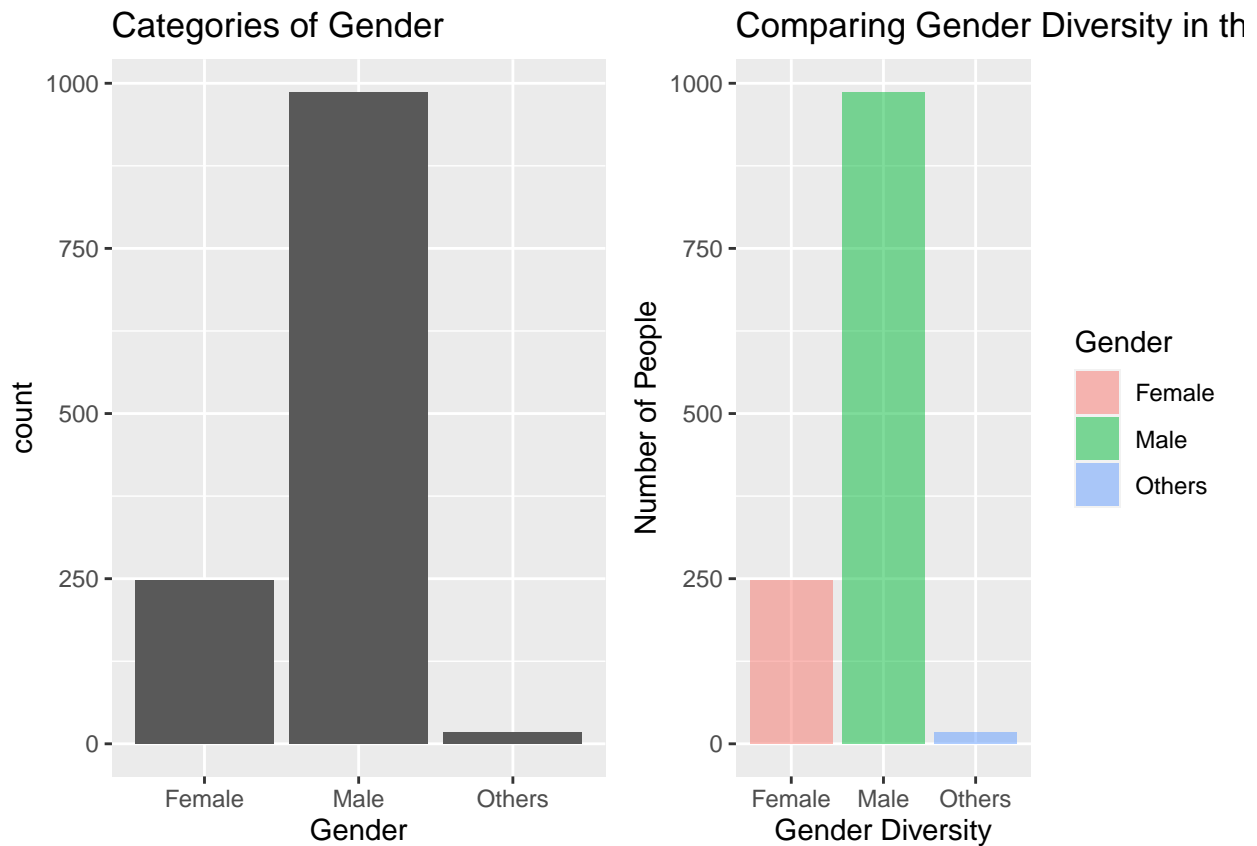
```
## # A tibble: 3 x 2
##   Gender count
##   <chr>   <int>
## 1 Female   247
## 2 Male    987
## 3 Others    17
```

```
## Visualize the number of subjects in each gender type
```

```
G1 <- survey_data %>% ggplot(aes(x=Gender)) +
  geom_bar() + ggtitle("Categories of Gender")
```

```
G2 <- ggplot(gender_diversity, aes(x = Gender, y = count, fill = Gender)) +
  geom_bar(stat = "identity", alpha = 0.5) +
  xlab("Gender Diversity") +
  ylab("Number of People") +
  ggtitle("Comparing Gender Diversity in the 2014 Mental Health in Tech Survey")
```

```
grid.arrange(G1, G2, nrow = 1)
```



```
#Age distribution
```

```
g1 <- survey_data %>% ggplot(aes(x=Age)) +
```

```

    geom_histogram() + ggtitle("Distribution of Age")

# Histogram of Age
g2 <- ggplot(survey_data, aes(x=Age)) + geom_histogram(aes(y=..density..), fill="green") + geom_density(col="green",
    title="Distribution of Transformed Age")

# Age categorization#
survey_data$Age <- cut(survey_data$Age, breaks = c(0, 16, 34, 60, 75),
    labels = c('Fresh', 'Junior', 'Senior', 'Super'))

# Verify Age group
survey_data$Age %>% table

```

```

## .
## Fresh Junior Senior Super
##      0      863      384      4

```

```

# Group by Age Group and count each group
age_group <- survey_data %>%
  group_by(Age) %>%
  dplyr::summarize(count = n())
age_group

```

```

## # A tibble: 3 x 2
##   Age      count
##   <fct>   <int>
## 1 Junior     863
## 2 Senior     384
## 3 Super        4

```

```

g3 <- ggplot(age_group, aes(x = Age, y = count, fill = Age)) +
  geom_bar(stat = "identity", alpha = 0.5) +
  xlab("Age Group") +
  ylab("Number of People") +
  ggtitle("Comparing Age Group in the 2014 Mental Health in Tech Survey")

grid.arrange(g1, g2, g3, nrow = 1)

```

```

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```

```

## Warning: The dot-dot notation ('..density..') was deprecated in ggplot2 3.4.0.
## i Please use 'after_stat(density)' instead.

```

```

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```




summarize your data to answer key questions

First, re-value the variables of interest

```
survey_data$no_employees <- as.factor(revalue(survey_data$no_employees,
  c("1-5"="1", "6-25"="2", "26-100"="3", "100-500"="4",
    "500-1000"="5", "More than 1000"="6")))
```

The following 'from' values were not present in 'x': 1-5, 6-25

```
survey_data$treatment <- as.numeric(revalue(survey_data$treatment,
  c("No"="0", "Yes"="1")))
```

Get a collection of tables for the responses

```
questions_only <- survey_data[,5:24]
```

```
response_tables <-
```

```
  lapply(questions_only, function(Response) as.data.frame(table(Response)))
```

Add names of questions for each for plotting purposes

```
plot_tables <-
```

```
  lapply(seq_along(response_tables),
    function(x) mutate(response_tables[[x]], name = names(response_tables)[x]))
```

```
all_questions <- bind_rows(plot_tables)
```

```
all_questions
```

```
##           Response Freq           name
```

## 1	No	1091	self_employed
## 2	Yes	142	self_employed
## 3	No	762	family_history
## 4	Yes	489	family_history
## 5	0	619	treatment
## 6	1	632	treatment
## 7	Never	212	work_interfere
## 8	Often	140	work_interfere
## 9	Rarely	173	work_interfere
## 10	Sometimes	464	work_interfere
## 11	25-Jun	289	no_employees
## 12	3	288	no_employees
## 13	4	175	no_employees
## 14	5	60	no_employees
## 15	5-Jan	158	no_employees
## 16	6	281	no_employees
## 17	No	880	remote_work
## 18	Yes	371	remote_work
## 19	No	226	tech_company
## 20	Yes	1025	tech_company
## 21	Don't know	407	benefits
## 22	No	371	benefits
## 23	Yes	473	benefits
## 24	No	499	care_options
## 25	Not sure	313	care_options
## 26	Yes	439	care_options
## 27	Don't know	187	wellness_program
## 28	No	837	wellness_program
## 29	Yes	227	wellness_program
## 30	Don't know	363	seek_help
## 31	No	641	seek_help
## 32	Yes	247	seek_help
## 33	Don't know	815	anonymity
## 34	No	64	anonymity
## 35	Yes	372	anonymity
## 36	Don't know	561	leave
## 37	Somewhat difficult	125	leave
## 38	Somewhat easy	265	leave
## 39	Very difficult	97	leave
## 40	Very easy	203	leave
## 41	Maybe	476	mental_health_consequence
## 42	No	487	mental_health_consequence
## 43	Yes	288	mental_health_consequence
## 44	Maybe	273	phys_health_consequence
## 45	No	920	phys_health_consequence
## 46	Yes	58	phys_health_consequence
## 47	No	258	coworkers
## 48	Some of them	771	coworkers
## 49	Yes	222	coworkers
## 50	No	390	supervisor
## 51	Some of them	349	supervisor
## 52	Yes	512	supervisor
## 53	Maybe	207	mental_health_interview
## 54	No	1003	mental_health_interview

```
## 55          Yes  41  mental_health_interview
## 56        Maybe 555  phys_health_interview
## 57          No 496  phys_health_interview
## 58          Yes 200  phys_health_interview
## 59    Don't know 574  mental_vs_physical
## 60          No 338  mental_vs_physical
## 61          Yes 339  mental_vs_physical
```

Group by family_history and count each group

```
fh_count <- survey_data %>%
  group_by(family_history) %>%
  dplyr::summarize(count = n(), proportion = n()/nrow(survey_data))
fh_count
```

```
## # A tibble: 2 x 3
##   family_history count proportion
##   <chr>          <int>     <dbl>
## 1 No             762     0.609
## 2 Yes            489     0.391
```

Group by treatment and Gender and count each group

```
treat_count <- survey_data %>%
  group_by(treatment, Gender) %>%
  dplyr::summarize(count = n(), proportion = n()/nrow(survey_data))
```

'summarise()' has grouped output by 'treatment'. You can override using the
'.groups' argument.

```
treat_count
```

```
## # A tibble: 6 x 4
## # Groups:   treatment [2]
##   treatment Gender count proportion
##   <dbl> <chr> <int>     <dbl>
## 1      0 Female   77     0.0616
## 2      0 Male   538     0.430
## 3      0 Others    4     0.00320
## 4      1 Female  170     0.136
## 5      1 Male   449     0.359
## 6      1 Others   13     0.0104
```

care_options available based on country

```
care_count <- survey_data %>% filter(care_options == "Yes") %>%
  group_by(Country) %>%
  dplyr::summarize(count = n(), proportion = n()/nrow(survey_data))
care_count
```

```
## # A tibble: 30 x 3
```

```
##      Country      count proportion
##      <chr>        <int>      <dbl>
##  1 Australia         9    0.00719
##  2 Austria           1    0.000799
##  3 Belgium           1    0.000799
##  4 Brazil            2    0.00160
##  5 Canada           28    0.0224
##  6 China             1    0.000799
##  7 Colombia          1    0.000799
##  8 Costa Rica        1    0.000799
##  9 Croatia           1    0.000799
## 10 Czech Republic    1    0.000799
## # ... with 20 more rows
```

```
# Group by work_interfere response and count each group
```

```
survey_data$work_interfere <- factor(survey_data$work_interfere, ordered = TRUE,
                                     levels = c('NA', 'Never', 'Rarely', 'Sometimes', 'Often'))

wi_count <- survey_data %>%
  group_by(work_interfere) %>%
  dplyr::summarize(count = n(), proportion = n()/nrow(survey_data)) %>%
  arrange(work_interfere)
wi_count
```

```
## # A tibble: 5 x 3
##   work_interfere count proportion
##   <ord>          <int>      <dbl>
## 1 Never         212      0.169
## 2 Rarely        173      0.138
## 3 Sometimes     464      0.371
## 4 Often         140      0.112
## 5 <NA>          262      0.209
```

```
# Group by Country and count each group
```

```
country_count <- survey_data %>%
  group_by(Country) %>%
  dplyr::summarize(count = n())
country_count
```

```
## # A tibble: 46 x 2
##   Country      count
##   <chr>        <int>
## 1 Australia         21
## 2 Austria           3
## 3 Belgium           6
## 4 Bosnia and Herzegovina 1
## 5 Brazil            6
## 6 Bulgaria          4
## 7 Canada           72
## 8 China             1
## 9 Colombia          2
## 10 Costa Rica        1
## # ... with 36 more rows
```

```

mental_health_interview <- survey_data %>%
  group_by(mental_health_interview) %>%
  summarise(n = n()) %>%
  mutate(percentage = signif(100 * ( n /sum(n)),2))
mental_health_interview

```

```

## # A tibble: 3 x 3
##   mental_health_interview      n percentage
##   <chr>                  <int>      <dbl>
## 1 Maybe                  207         17
## 2 No                    1003         80
## 3 Yes                     41         3.3

```

```

coworkers <- survey_data %>%
  group_by(coworkers) %>%
  summarise(n = n()) %>%
  mutate(percentage = signif(100 * ( n /sum(n)),2))
coworkers

```

```

## # A tibble: 3 x 3
##   coworkers      n percentage
##   <chr>        <int>      <dbl>
## 1 No          258         21
## 2 Some of them 771         62
## 3 Yes         222         18

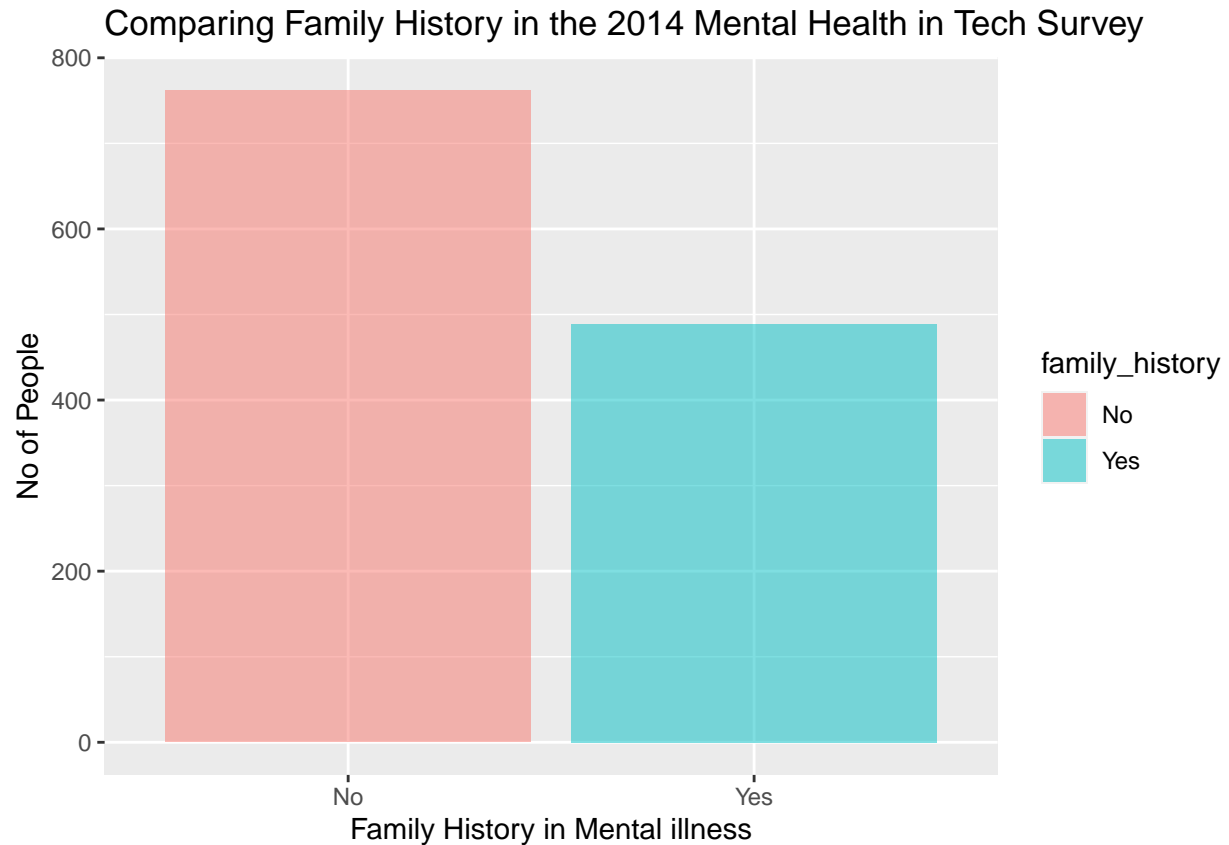
```

Plotting to illustrate the finding of key questions :

```

# Does family_history have any relation with mentalhealth
ggplot(fh_count, aes(x = family_history, y = count, fill = family_history)) +
  geom_bar(stat = "identity", alpha = 0.5) +
  xlab("Family History in Mental illness") +
  ylab("No of People") +
  ggtitle("Comparing Family History in the 2014 Mental Health in Tech Survey")

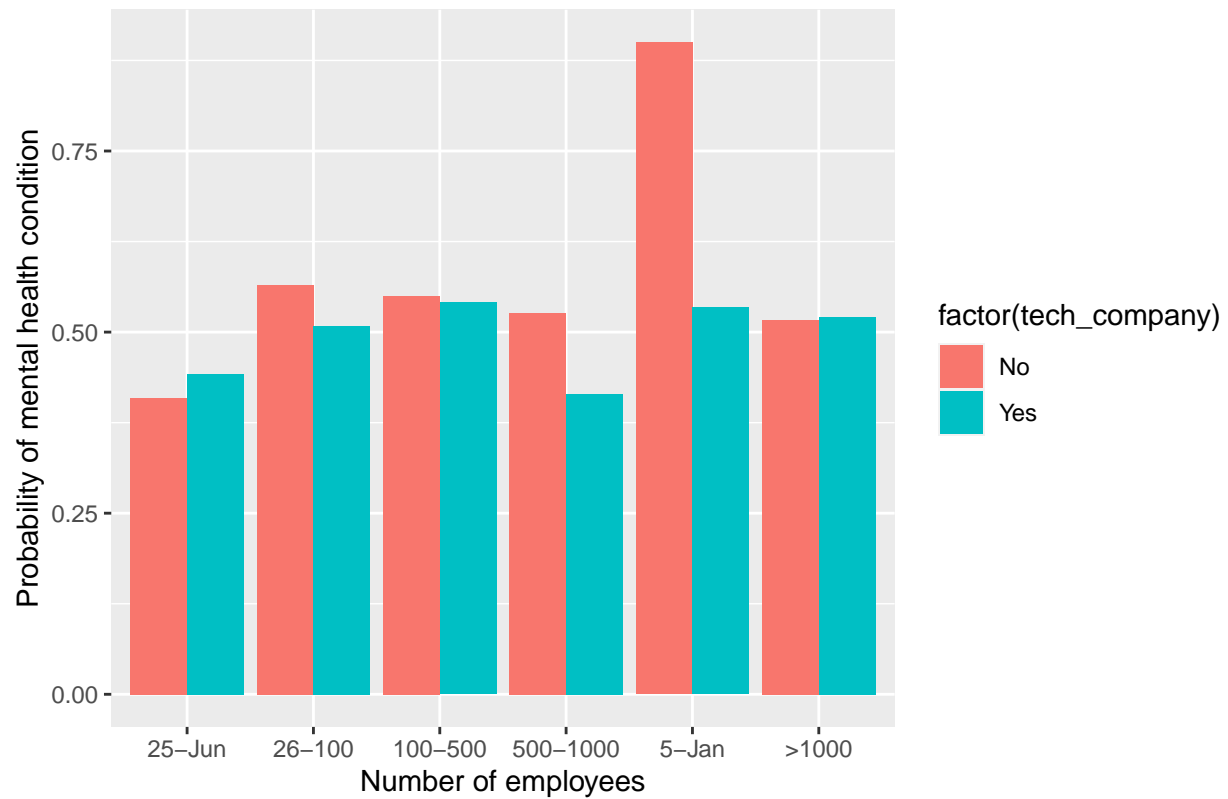
```



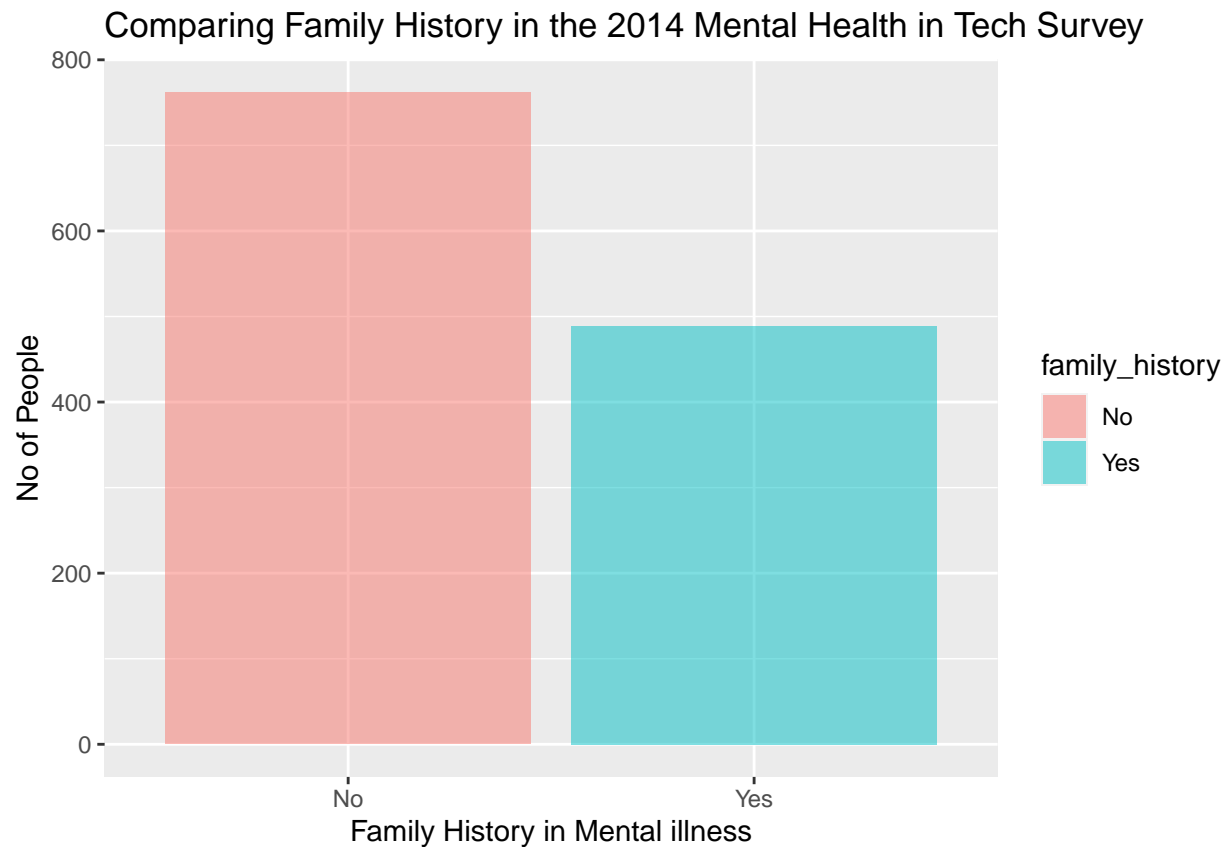
```
# Use a bar graph to graph `treatment` (treated mental health condition)
# over `no_employees` (number of employees) and `tech_company` (company type)
ggplot(survey_data,aes(x=no_employees,y=treatment, fill=factor(tech_company)), color=factor(vs)) +
  stat_summary(fun.y=mean,position=position_dodge(),geom="bar") +
  labs(x = "Number of employees", y = "Probability of mental health condition",
       title = "Probability of mental health illness by workplace type and size") +
  scale_x_discrete(labels=c("1" = "1-5", "2" = "6-25", "3" = "26-100", "4"="100-500",
                           "5"="500-1000", "6"=">1000"))
```

```
## Warning: The 'fun.y' argument of 'stat_summary()' is deprecated as of ggplot2 3.3.0.
## i Please use the 'fun' argument instead.
```

Probability of mental health illness by workplace type and size

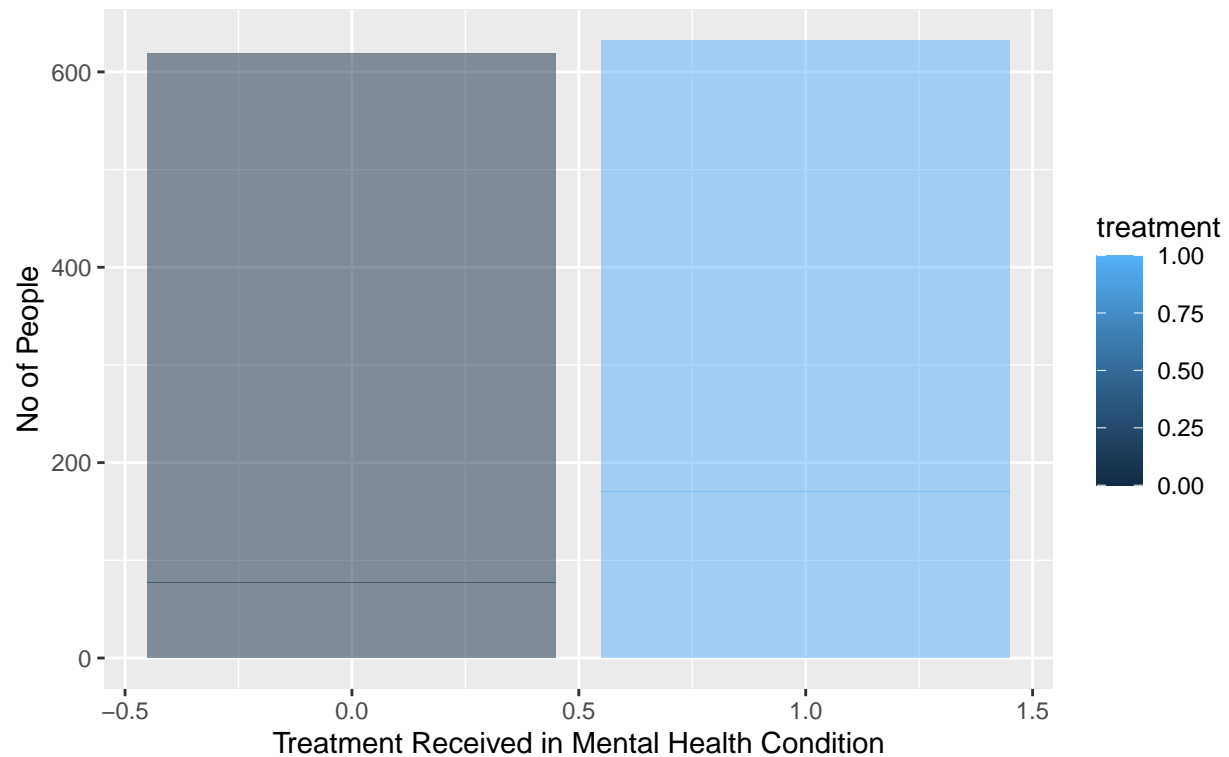


```
ggplot(fh_count, aes(x = family_history, y = count, fill = family_history)) +
  geom_bar(stat = "identity", alpha = 0.5) +
  xlab("Family History in Mental illness") +
  ylab("No of People") +
  ggtitle("Comparing Family History in the 2014 Mental Health in Tech Survey")
```



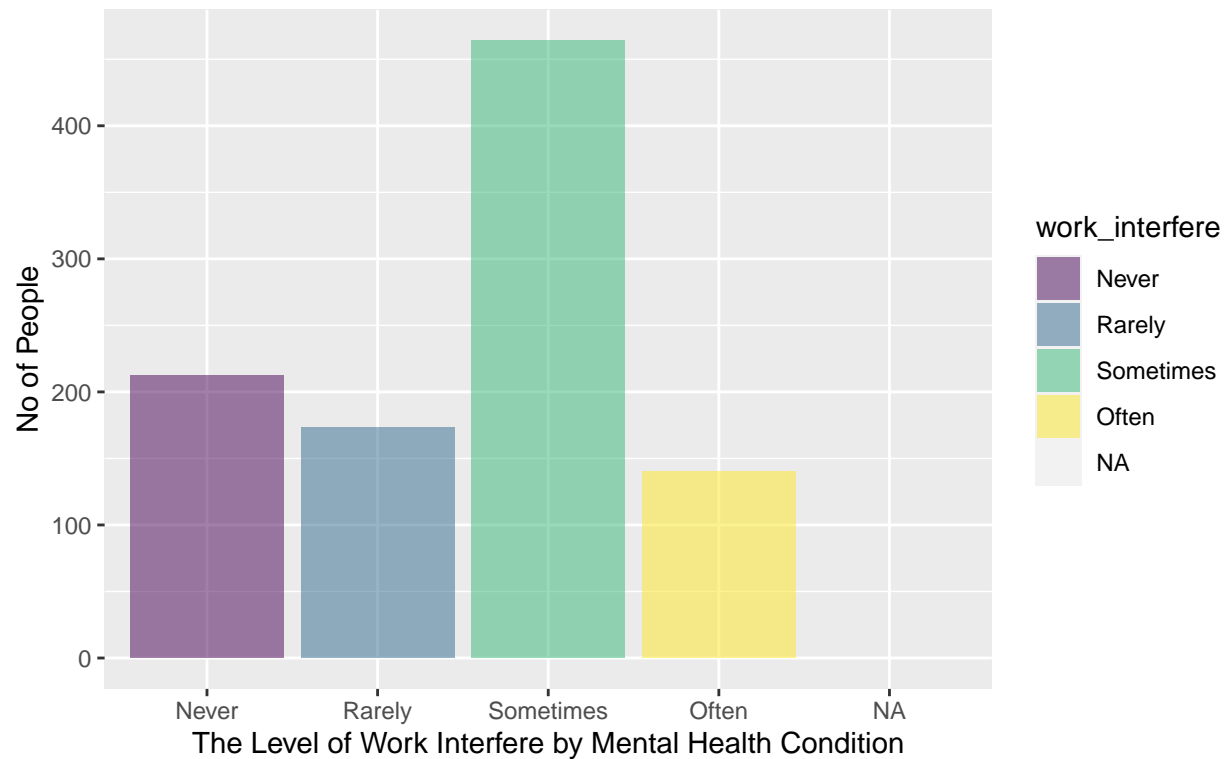
```
# Visualize the number of subjects
ggplot(treat_count, aes(x = treatment, y = count, fill = treatment)) +
  geom_bar(stat = "identity", alpha = 0.5) +
  xlab("Treatment Received in Mental Health Condition") +
  ylab("No of People") +
  ggtitle("Comparing the Treatment Received in the 2014 Mental Health in
          Tech Survey")
```


Comparing the Treatment Received in the 2014 Mental Health in Tech Survey



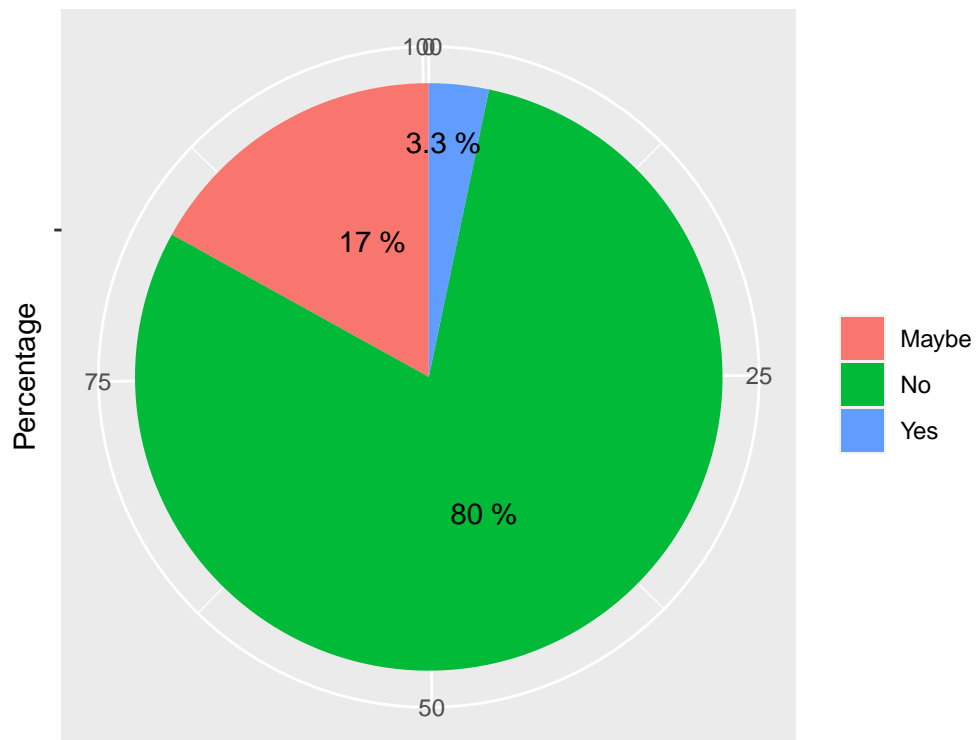
```
# Visualize the number of subjects
ggplot(wi_count, aes(x = work_interfere, y = count, fill = work_interfere)) +
  geom_bar(stat = "identity", alpha = 0.5) +
  xlab("The Level of Work Interfere by Mental Health Condition") +
  ylab("No of People") +
  ggtitle("Comparing the Level of Work Interfere in the 2014 Mental Health in
    Tech Survey")
```

Comparing the Level of Work Interfere in the 2014 Mental Health in Tech Survey

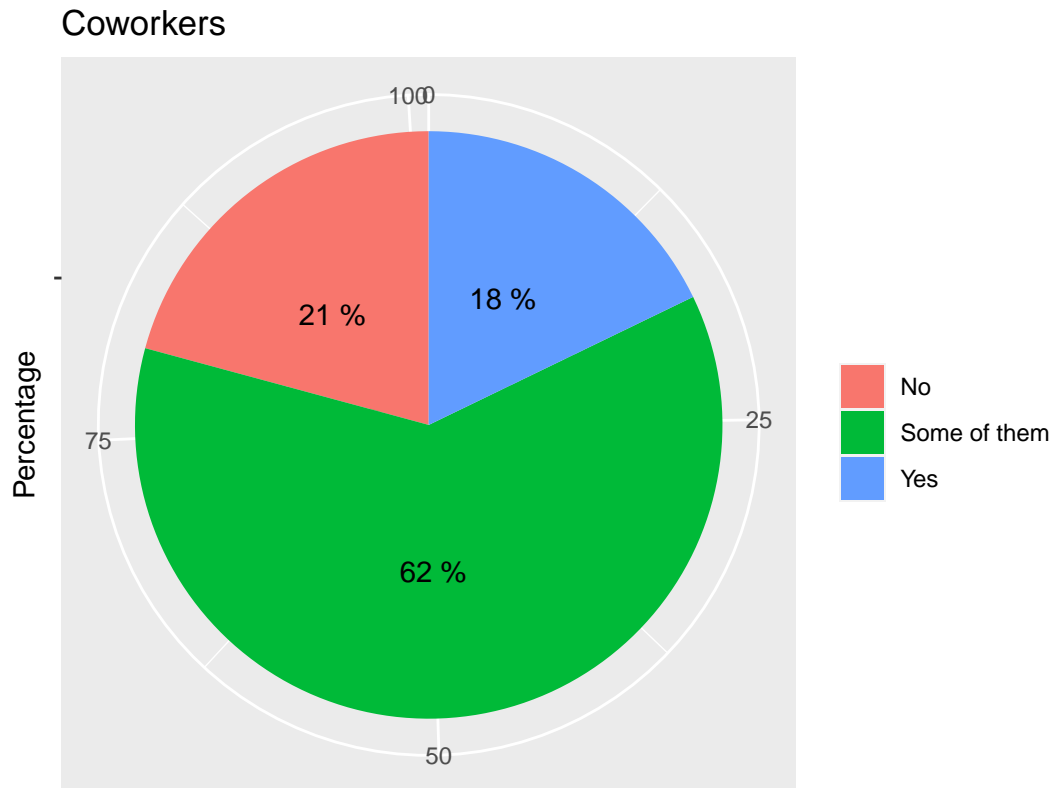


```
ggplot(data = mental_health_interview) +
  geom_bar(mapping = aes(x = "", y = percentage, fill = mental_health_interview), stat = "identity", v
  geom_text ( aes(x = c(1,1,1.3), y = c(94,44,1),label = paste(percentage,"%")),size = 4) +
  ggtitle("Mental Health Interview") +
  labs(x = "Percentage",y = "") +
  guides(fill = guide_legend(title = NULL)) +
  coord_polar("y")
```

Mental Health Interview



```
ggplot(data = coworkers) +
  geom_bar(mapping = aes(x = "", y = percentage, fill = coworkers), stat = "identity", width = 1) +
  geom_text ( aes(x = c(1,1,1), y = c(89.5,50,8.5),label = paste(percentage,"%")),size = 4) +
  ggtitle("Coworkers") +
  labs(x = "Percentage",y = "") +
  guides(fill = guide_legend(title = NULL)) +
  coord_polar("y")
```



```
## Few key questions are answered based on the plotting done above and for rest of the
## questions we need to figure out the correlation between those variables and have to design
## linear model or logistic regression model.

## Next step we will plan to build a model (Generalized Liner model) and split the input data
## into train and test data and test the model. Will calculate the accuracy of the model based
## on the input data.
```