Mental Health in Tech

Prashanth Paisa

2022-10-16

## Library Packages used here

library(readxl)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.5 ✔ purrr 0.3.4  
## ✔ tibble 3.1.6 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(dplyr)  
library(tidyr)  
library(stringr)  
library(ggplot2)

#Reading source file having data in two separate sheets

#Loading individual datasheets  
  
df\_mental\_datasheet\_1 <- read\_excel("survey.xlsx", sheet = "Sheet1",   
 col\_types = c("date", "numeric", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text"))  
  
df\_mental\_datasheet\_2 <- read\_excel("survey.xlsx", sheet = "Sheet 2",   
 col\_types = c("date", "numeric", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text", "text",   
 "text", "text", "text", "text"))

## Warning: Coercing numeric to date in A5 / R5C1

#Adding two datasheet into one  
  
df\_mental <- merge(df\_mental\_datasheet\_1, df\_mental\_datasheet\_2, all.x= TRUE, all.y = TRUE)

#Observing combined Mental Heath data  
view(df\_mental)

#Handling missing values  
  
df\_mental$Timestamp <- as.character(df\_mental$Timestamp)   
df\_mental$Timestamp[is.na(df\_mental$Timestamp)] <- " "

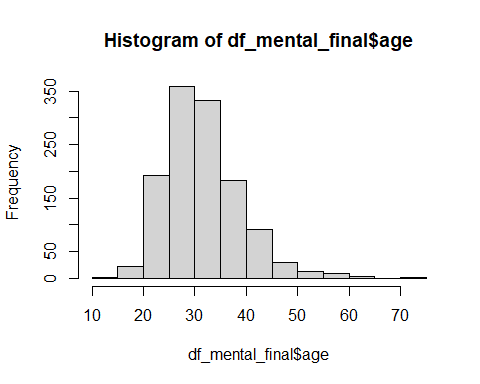
#Finding and removing missing values  
  
df\_mental\_final <- df\_mental %>% drop\_na()

#Converting the columns names format into lowercase  
  
names(df\_mental\_final)[1]<-"timestamp"  
names(df\_mental\_final)[2]<-"age"  
names(df\_mental\_final)[3]<-"gender"  
names(df\_mental\_final)[4]<-"country"

#Plotting histogram of Age distribution of ages between 10 and 75  
  
df\_mental\_final <-   
 df\_mental\_final[-c(which(df\_mental\_final$age > 75 | df\_mental\_final$age < 10)), ]  
  
summary(df\_mental\_final$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 11.00 27.00 31.00 32.08 36.00 72.00

hist(df\_mental\_final$age)



#Fixing Gender Data by converting all into Male, Female and Others  
  
df\_mental\_final$gender <- tolower(df\_mental\_final$gender)  
col\_gender <- as.vector(df\_mental\_final$gender)  
  
Female <-   
 c('female', 'cis female', 'f', 'woman', 'femake', 'female ', 'cis-female/femme', 'female (cis)', 'femail')  
  
Male <-   
 c('m', 'male', 'male-ish', 'maile', 'cis male', 'mal', 'male (cis)', 'make', 'male ', 'man', 'msle', 'mail', 'malr', 'cis man','Mle')   
  
Others <-   
 c('queer/she/they', 'non-binary', 'nah', 'enby', 'fluid', 'genderqueer', 'androgyne', 'agender', 'guy (-ish) ^\_^', 'male leaning androgynous', 'neuter', 'queer', 'ostensibly male, unsure what that really means','a little about you','p','all', 'trans-female', 'trans woman', 'female (trans)','something kinda male?', '1','2')  
  
col\_gender <- sapply(as.vector(col\_gender), function(x) if(x %in% Male) "Male" else x)  
col\_gender <- sapply(as.vector(col\_gender), function(x) if(x %in% Female) "Female" else x)  
col\_gender <- sapply(as.vector(col\_gender), function(x) if(x %in% Others) "Others" else x)  
  
df\_mental\_final$gender <- col\_gender

#Removing Timestamp  
  
df\_mental\_final <- df\_mental\_final[,-1]

#Replacing country data "US" with "United States"  
  
df\_mental\_final$country[df\_mental\_final$country=="US"]<-"United States"

#Finding incorrect values of number of employees  
unique(df\_mental\_final$no\_employees)

## [1] "26-100" "500-1000" "100-500" "44372"   
## [5] "More than 1000" "44201"

table(df\_mental\_final$no\_employees)

##   
## 100-500 26-100 44201 44372 500-1000   
## 174 284 155 285 59   
## More than 1000   
## 280

#Converting treatment into Yes, No and NA  
  
table(df\_mental\_final$treatment)

##   
## - N NA No Y Yes   
## 2 2 1 611 1 620

df\_mental\_final$treatment[df\_mental\_final$treatment=="N"]<-"No"  
df\_mental\_final$treatment[df\_mental\_final$treatment=="-"]<-"NA"  
df\_mental\_final$treatment[df\_mental\_final$treatment=="Y"]<-"Yes"

#Replacing work\_interfere data to 0 to Never  
  
table(df\_mental\_final$work\_interfere)

##   
## 0 NA Never Often Rarely Sometimes   
## 2 255 210 137 173 460

df\_mental\_final$work\_interfere[df\_mental\_final$work\_interfere=="0"]<-"Never"

#Fixing and converting "-" to No  
  
table(df\_mental\_final$remote\_work)

##   
## - No Yes   
## 2 870 365

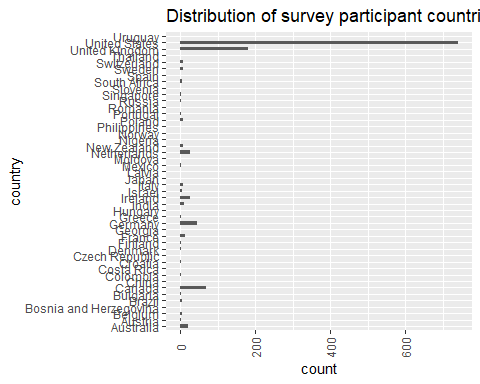
df\_mental\_final$remote\_work[df\_mental\_final$remote\_work=="-"]<-"No"

#Converting phys\_health\_consequence data to No from N and 2  
  
table(df\_mental\_final$phys\_health\_consequence)

##   
## 2 Maybe N No Yes   
## 5 272 3 899 58

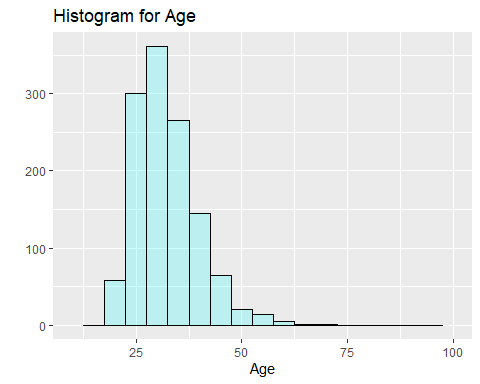
df\_mental\_final$phys\_health\_consequence[df\_mental\_final$phys\_health\_consequence=="N"]<-"No"  
df\_mental\_final$phys\_health\_consequence[df\_mental\_final$phys\_health\_consequence=="2"]<-"No"

#Barplot of surveying contries  
  
data\_country <- ggplot(df\_mental\_final, aes(country))  
data\_country + geom\_bar(width = 0.5) +   
 coord\_flip() +  
 theme(axis.text.x = element\_text(angle=90, hjust=0.5, vjust=0.5)) +  
 labs(title="Distribution of survey participant countries")



#Histogram plot of Age distribution  
  
qplot(df\_mental\_final$age,  
 geom="histogram",  
 binwidth = 5,   
 main = "Histogram for Age",   
 xlab = "Age",   
 fill=I("cyan"),   
 col=I("black"),   
 alpha=I(.2),  
 xlim=c(10,100))

## Warning: Removed 2 rows containing missing values (geom\_bar).



#Saving the processed data into cleaned\_mental\_dataset.csv  
  
write.csv(df\_mental\_final, "cleaned\_mental\_dataset.csv")