Data Preparation

Xinghan Liu

2023/06

#install required packages  
if (!require("readxl")) install.packages("readxl")

## Loading required package: readxl

library(readxl)  
if (!require("openxlsx")) install.packages("openxlsx")

## Loading required package: openxlsx

library(openxlsx)  
library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

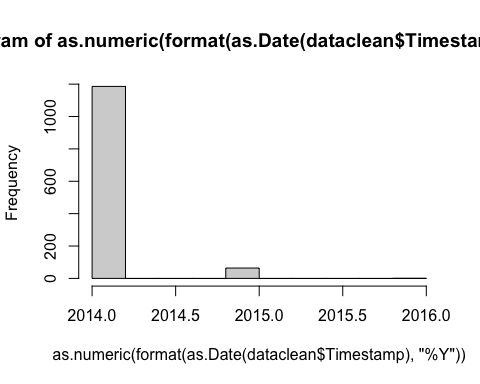
library(ggplot2)

# Read the first sheet  
sheet1\_data <- read\_excel('/Users/hankhan688/Documents/survey.xlsx',sheet ="Sheet1")  
# Read the second sheet  
sheet2\_data <- read\_excel('/Users/hankhan688/Documents/survey.xlsx', sheet = "Sheet 2")  
#I firstly realized the data of time was not in its orginial format, I need to convert them back  
time1<-openxlsx::convertToDateTime(sheet2\_data$Timestamp)  
sheet2\_data$Timestamp<- time1  
#combine two data into one  
dataall<-rbind(sheet1\_data,sheet2\_data)

#Cleaning data  
dataclean<-dataall  
dataclean<-subset(dataclean, select = c(Timestamp,Age,Gender,Country,state,self\_employed,  
 family\_history,treatment,work\_interfere))  
#Clean the Year-------------------------------------------------------  
year\_freq <- dataclean %>%  
 # Extract the year from the "Timestamp" column  
 mutate(year = format(as.Date(Timestamp), "%Y")) %>%  
 # Count the occurrences of the year  
 count(year)  
# Print the frequency table, found there are error data of 1905.  
print(year\_freq)

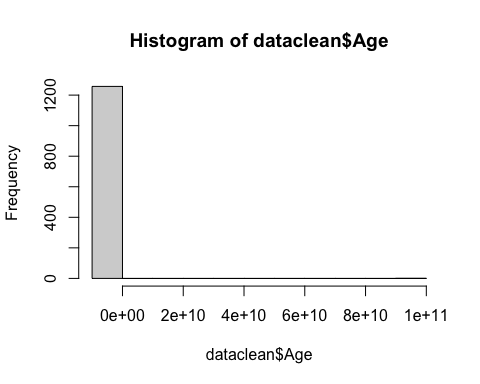
## # A tibble: 5 × 2  
## year n  
## <chr> <int>  
## 1 1905 5  
## 2 2014 1186  
## 3 2015 64  
## 4 2016 1  
## 5 <NA> 3

# from the table we need to remove the year which is 1905  
dataclean <- dataclean %>%  
 # Extract the year from the "Timestamp" column   
 mutate(year = format(as.Date(Timestamp), "%Y")) %>%  
 mutate(year = as.numeric(year))   
#Let date to be NA for error entry  
dataclean$Timestamp[dataclean$year == 1905] <- NA  
hist(as.numeric(format(as.Date(dataclean$Timestamp), "%Y")))

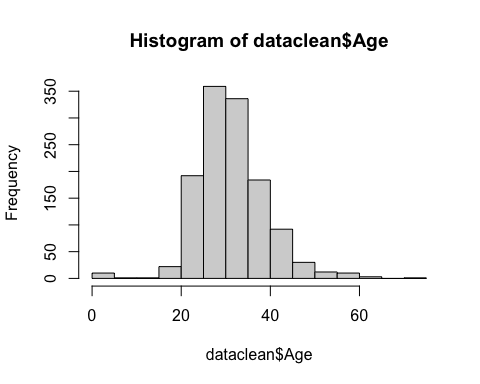


#remove the year variable  
dataclean<-subset(dataclean, select = -year)

#Clean the Age-------------------------------------------------------  
hist(dataclean$Age)

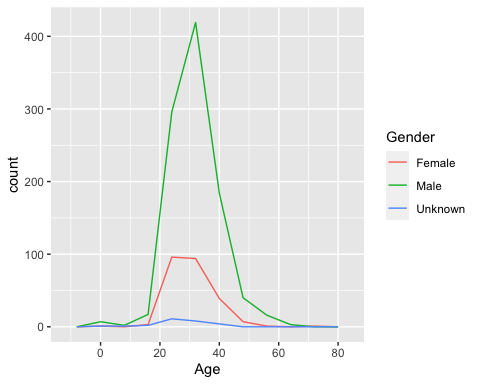


dataclean$Age[dataclean$Age<0] <-NA  
dataclean$Age[dataclean$Age>72] <-NA  
hist(dataclean$Age)



#Clean the gender  
dataclean$Gender <- ifelse(grepl("^[Mm]", dataclean$Gender), "Male",  
 ifelse(grepl("^[Ff]", dataclean$Gender), "Female", "Unknown"))  
ggplot(dataclean, aes(x = Age, colour = Gender, fill = Gender)) +  
 geom\_freqpoly(bins=10)

## Warning: Removed 6 rows containing non-finite values (`stat\_bin()`).



#Clean the country and state let different writing to be the same---------------------------  
table(dataclean$Country)

##   
## Australia Austria Bahamas, The   
## 21 3 1   
## Belgium Bosnia and Herzegovina Brazil   
## 6 1 6   
## Bulgaria Canada China   
## 4 72 1   
## Colombia Costa Rica Croatia   
## 2 1 2   
## Czech Republic Denmark Finland   
## 1 2 3   
## France Georgia Germany   
## 13 1 45   
## Greece Hungary India   
## 2 1 10   
## Ireland Israel Italy   
## 27 5 7   
## Japan Latvia Mexico   
## 1 1 3   
## Moldova Netherlands New Zealand   
## 1 27 8   
## Nigeria Norway Philippines   
## 1 1 1   
## Poland Portugal Romania   
## 7 2 1   
## Russia Singapore Slovenia   
## 3 4 1   
## South Africa Spain Sweden   
## 6 1 7   
## Switzerland Thailand UK   
## 7 1 2   
## United Kingdom United States Uruguay   
## 183 745 1   
## US Zimbabwe   
## 5 1

dataclean$Country <- ifelse(dataclean$Country == "UK", "United Kingdom",  
 ifelse(dataclean$Country == "US", "United States", dataclean$Country))  
table(dataclean$state)

##   
## AL AZ CA California CO CT DC   
## 8 7 137 1 9 4 4   
## FL GA IA ID IL IN KS   
## 15 12 4 1 29 27 3   
## KY LA MA MD ME MI MN   
## 5 1 20 8 1 22 21   
## MO MS NA NC NE New York NH   
## 12 1 514 14 2 1 3   
## NJ NM NV NY OH OK OR   
## 6 1 3 56 30 6 29   
## PA RI SC SD Texas TN TX   
## 29 1 5 3 1 45 43   
## UT VA VT WA WI WV WY   
## 11 14 3 70 12 1 2

dataclean$state <- ifelse(dataclean$state== "California", "CA",  
 ifelse(dataclean$state == "New York", "NY",  
 ifelse(dataclean$state == "Texas","TX", dataclean$state)))  
#Clean the treatment error term  
dataclean$treatment <- ifelse(dataclean$treatment== "-", NA,  
 ifelse(dataclean$treatment == "N", "NA",  
 ifelse(dataclean$treatment== "Y","Yes",dataclean$treatment)))  
#Clean the work\_interfere error term  
dataclean$work\_interfere <- ifelse(dataclean$work\_interfere== "0", NA ,dataclean$work\_interfere)  
#save the cleaned data  
write.csv(dataclean, "/Users/hankhan688/Documents/dataclean.csv")