ICS-321

Data Warehousing and Data Mining

LAB-6

2020BCS0082

ADITYA SANJAY SHITALE

**Step 1: Initial Exploratory Analysis**

**• Import the given data into R and save it as dataframe**

**• Familiarize with the dataset.**

**• Check the class, number of rows and columns and summary statistics for all the columns of the data frame.**

library(dplyr)

df = read.csv("Sampledata\_L6.csv")

ncol(df)

nrow(df)

class(df)

sapply(df, class)

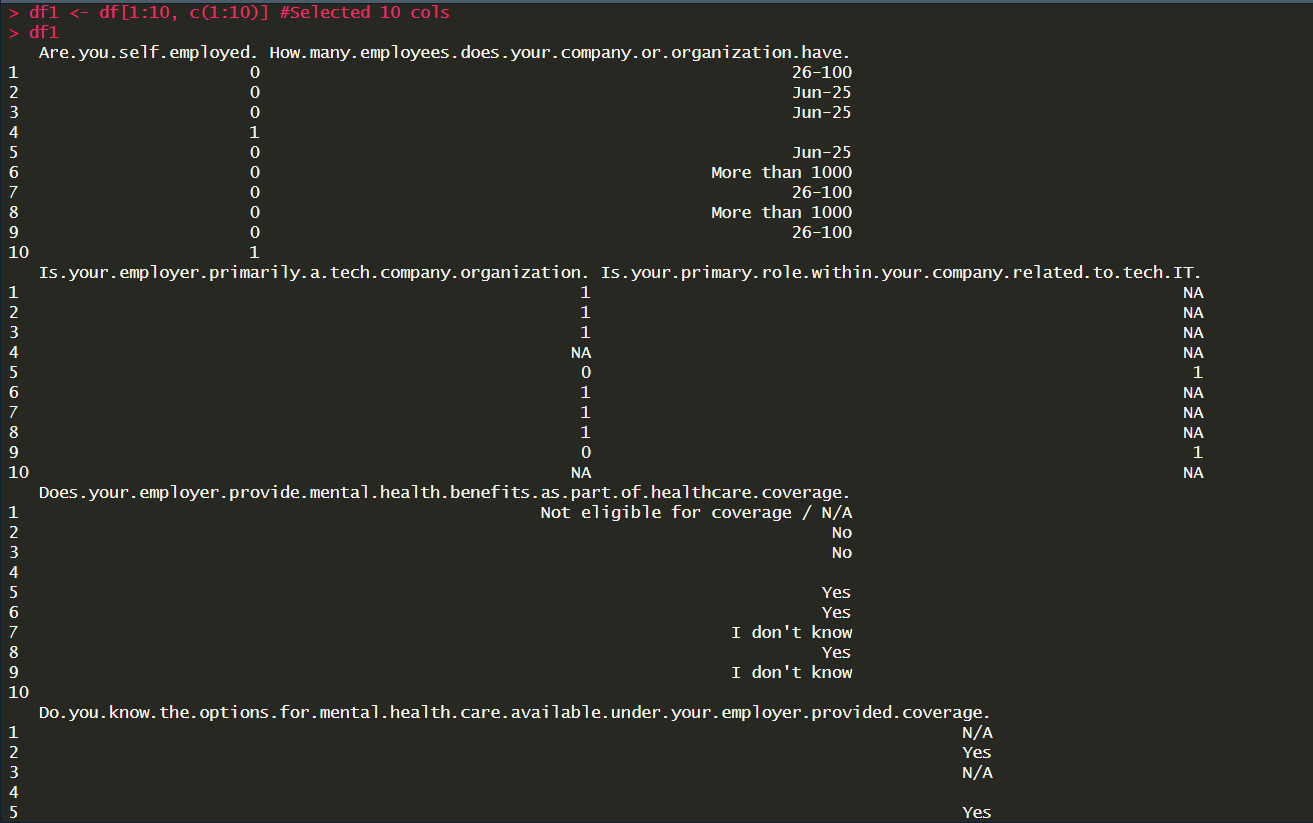
****

Step 2: Error Correction

• Create a new dataframe using the first ten columns of dataset and perform the below given queries.

df1 <- df[1:10, c(1:10)] #Selected 10 cols

df1

****

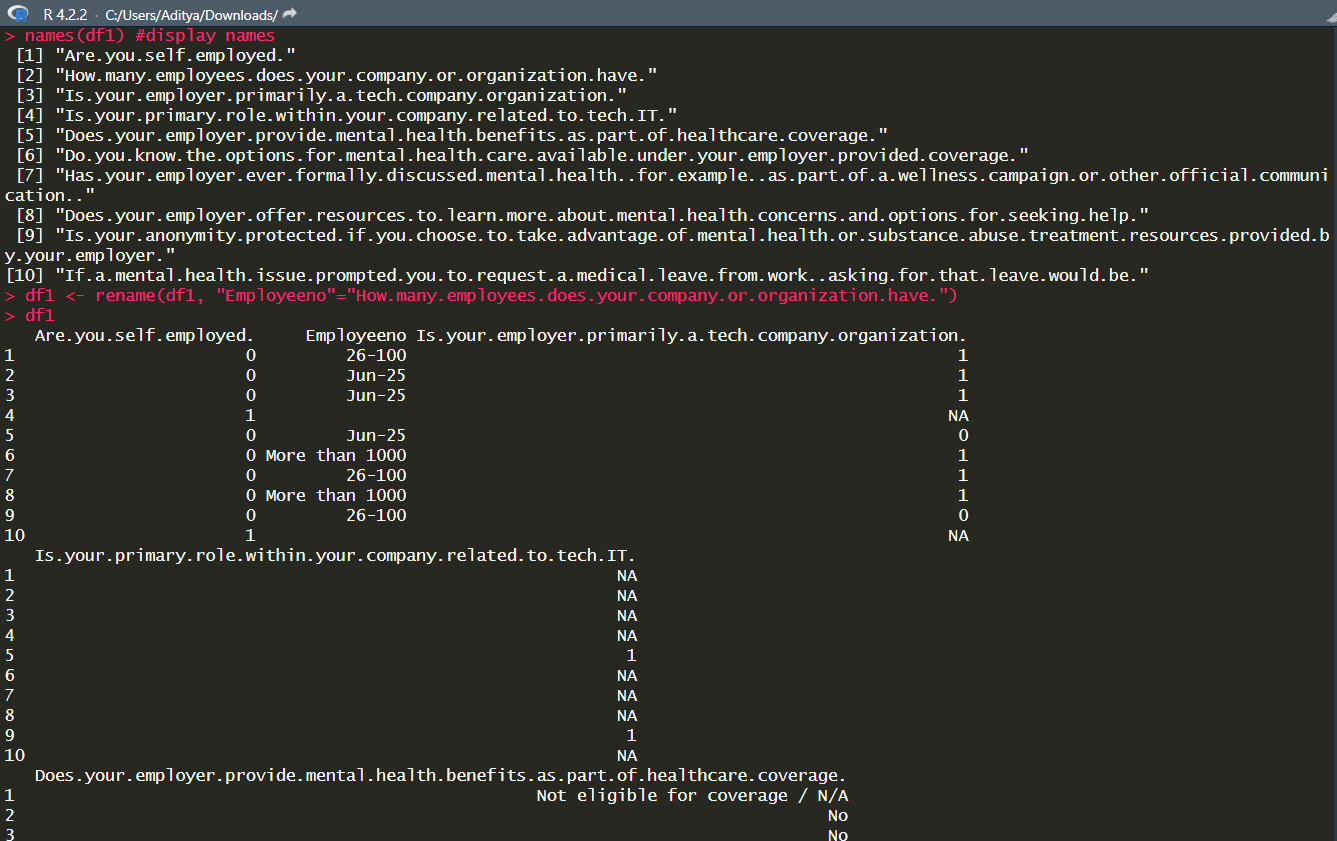
• View all variable labels with the names() function.

• Modify the variable label with dplyr’s rename().(“How many employees does your company or organization have?” can be changed to Employeeno)

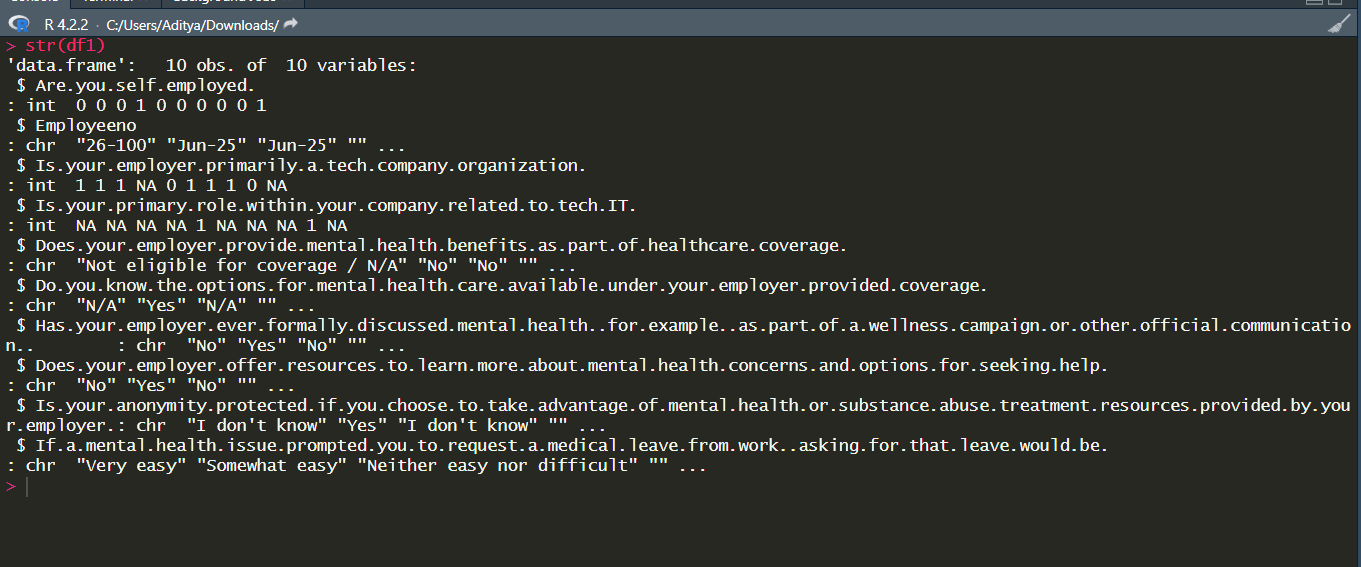
names(df1) #display names

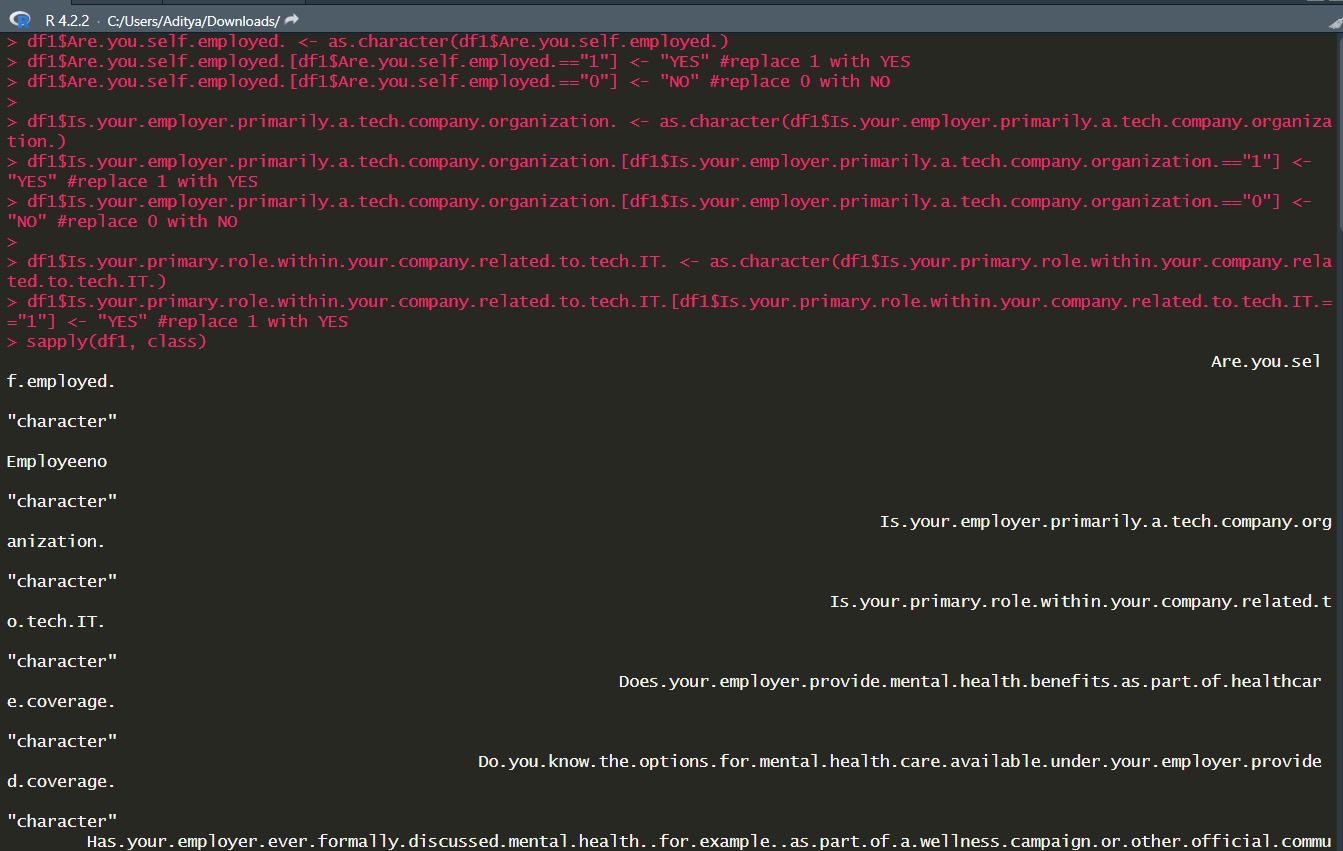
df1 <- rename(df1, "Employeeno"="How.many.employees.does.your.company.or.organization.have.")

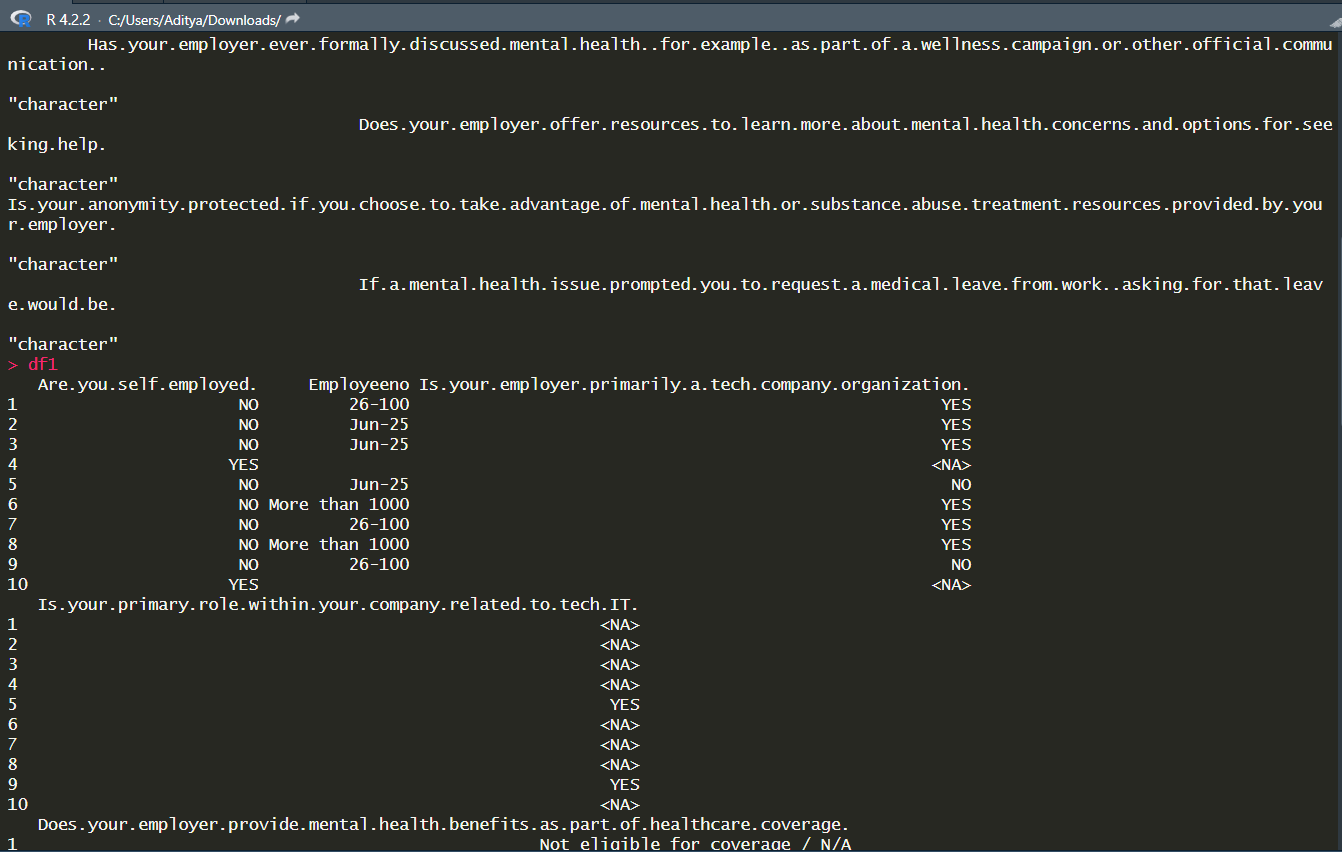
df1



• Determine the Faulty data types by the str() function and use the function such as as.character(),as.numeric(),as.integer(),as.logical(), as.factor() to change the data type .Types of data



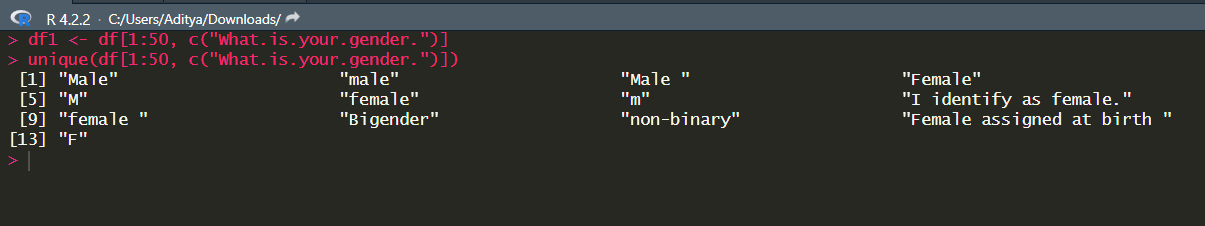




• Check the String inconsistencies of the column “What is your gender?” using unique() function and correct it to either “Male” or “Female” using function gsub().

df1 <- df[1:50, c("What.is.your.gender.")]

unique(df[1:50, c("What.is.your.gender.")])



df1 <- gsub("I identify as female", "Female", df1)

df1 <- gsub("Bigender", "Male", df1)

df1 <- gsub("non-binary", "Male", df1)

df1 <- gsub("Female assigned at birth", "Female", df1)

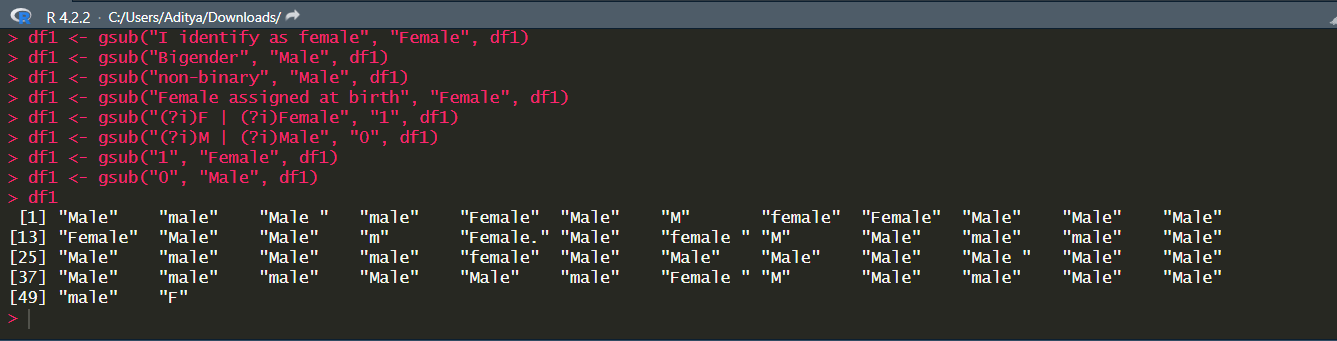
df1 <- gsub("(?i)F | (?i)Female", "1", df1)

df1 <- gsub("(?i)M | (?i)Male", "0", df1)

df1 <- gsub("1", "Female", df1)

df1 <- gsub("0", "Male", df1)

df1

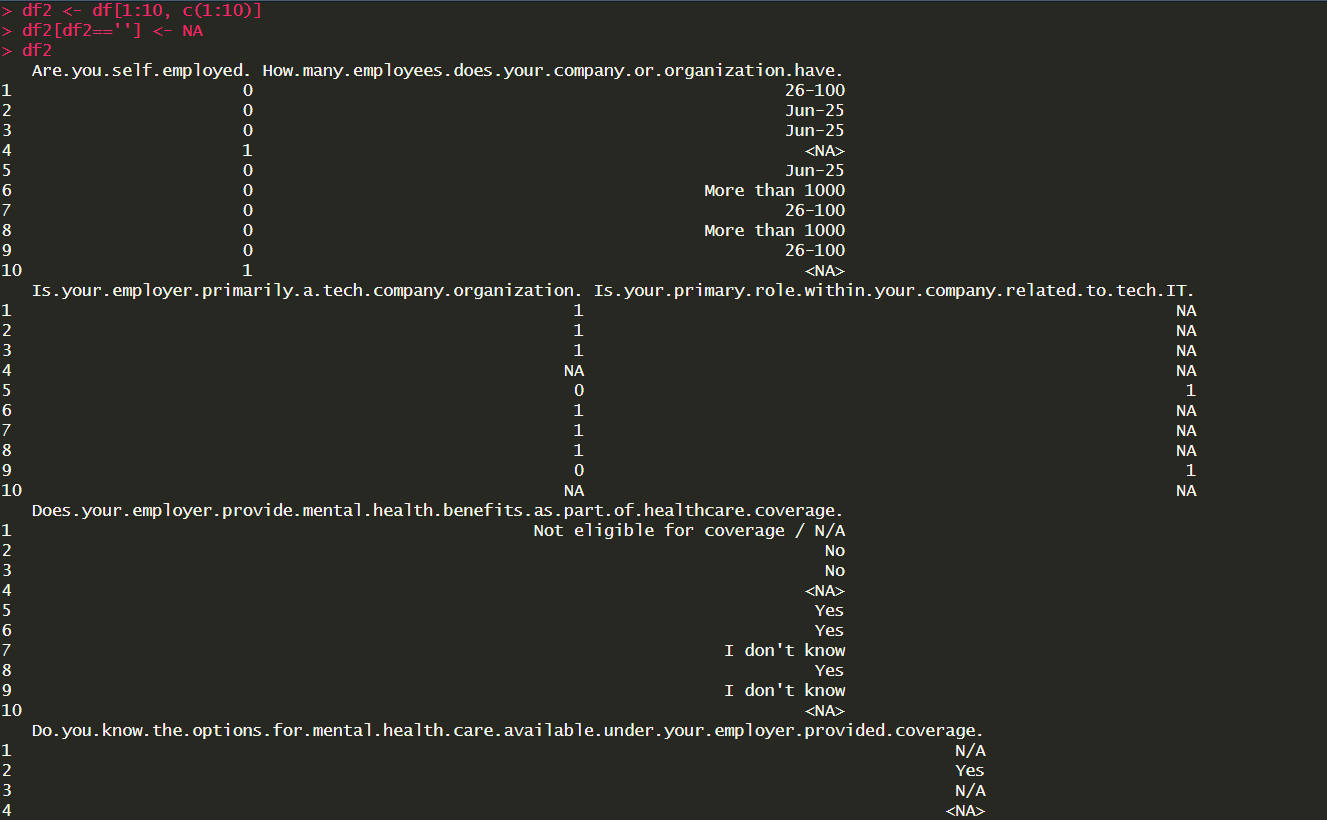


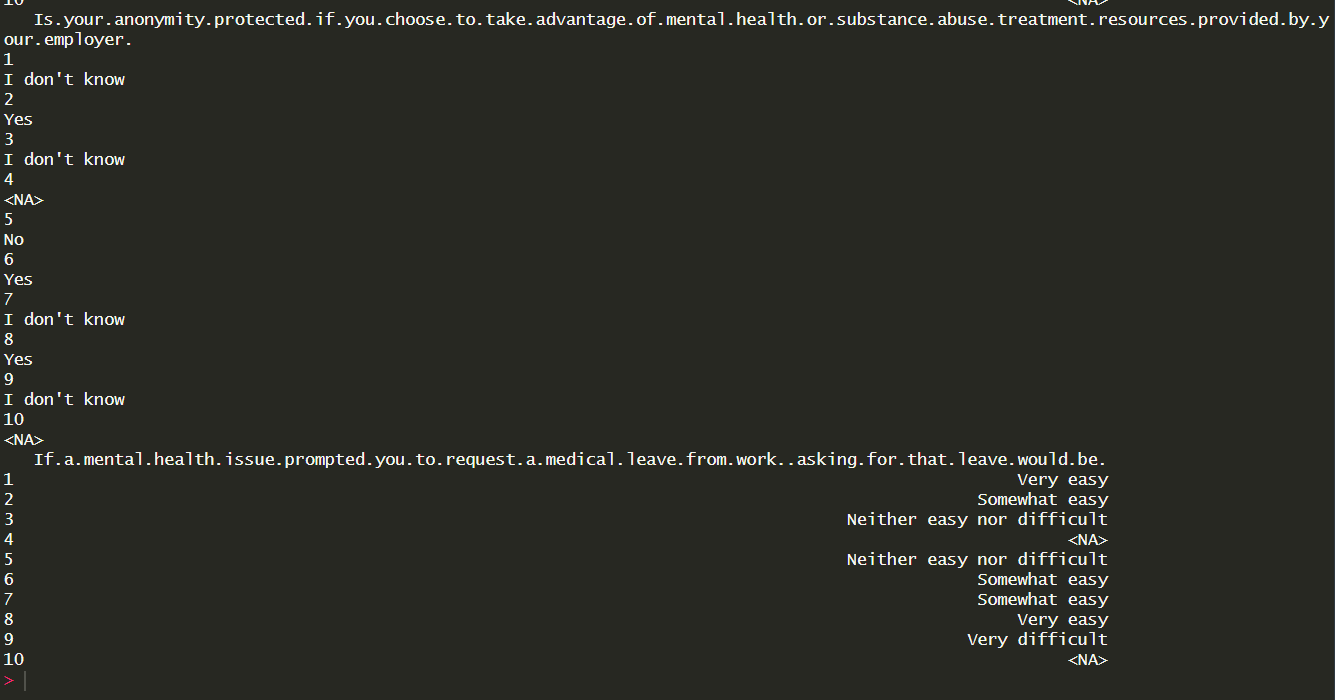
• Check for missing values in the entire dataframe.

df2 <- df[1:10, c(1:10)]

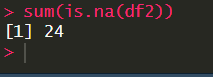
df2[df2==''] <- NA

df2



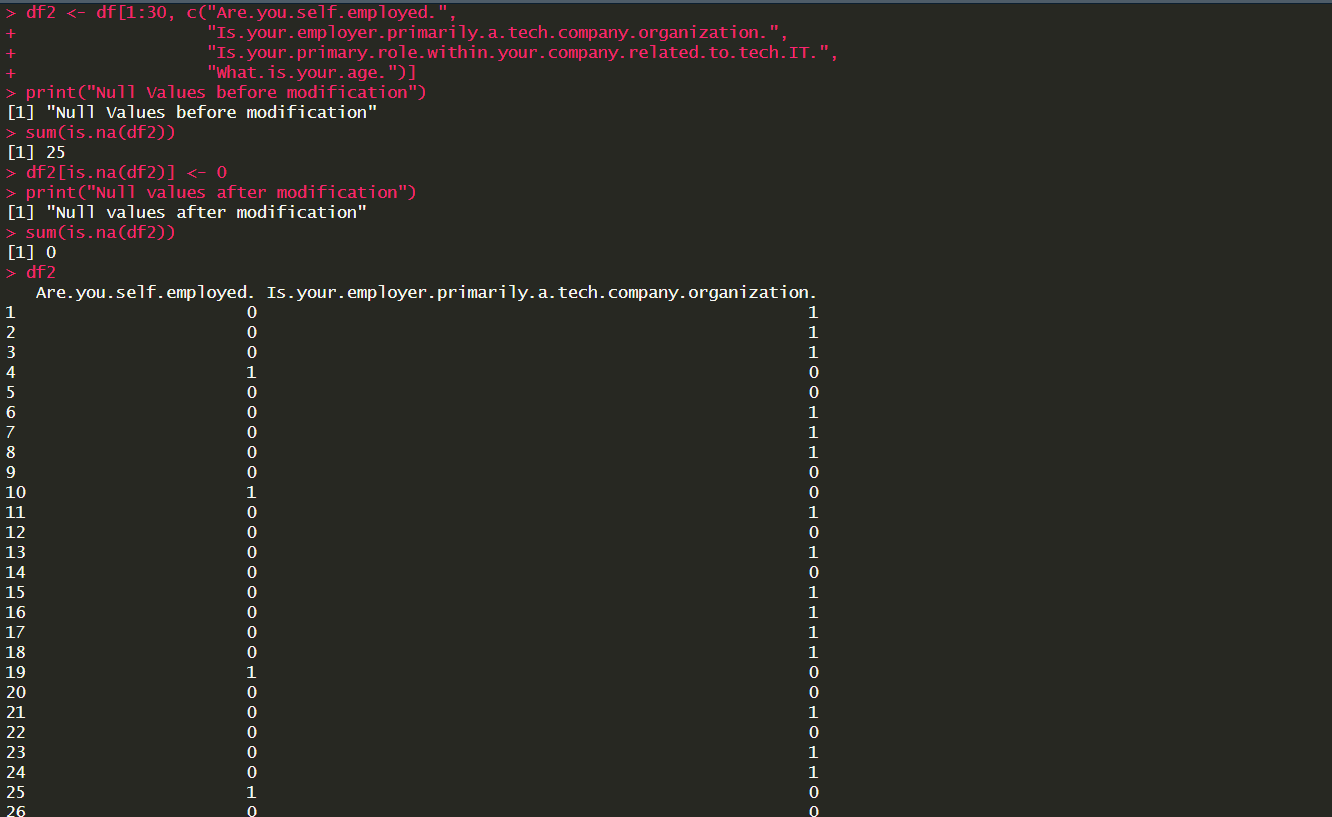


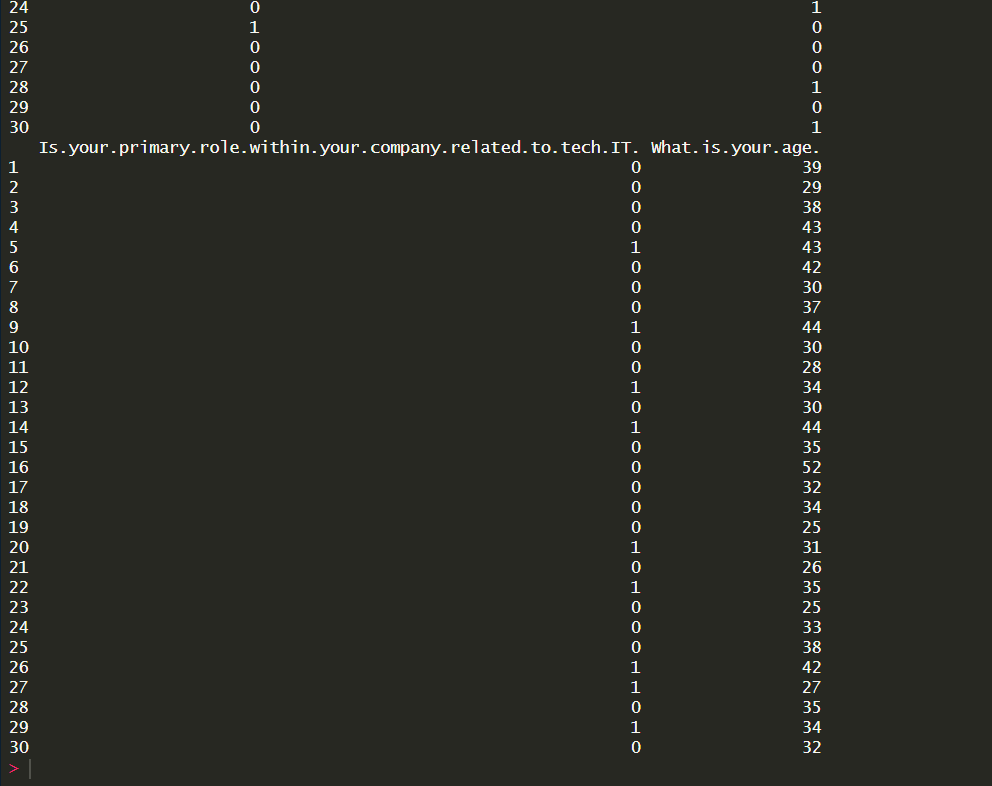
• Check for the total number of missing values in the entire dataframe.



• Eliminate missing values completely from the entire dataframe.

• Replace the NA’s in the entire dataframe with ‘0’s.





**Step 3: Visual Analysis**

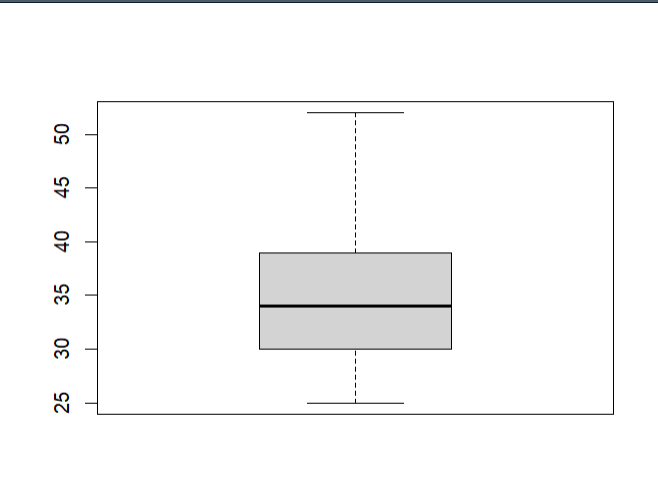
**• Using package “plyr” plot the Histogram and the Boxplot of age, the number of diagnosed mental illnesses each respondent has, and the number of believed mental illnesses each respondent has.**

age <- df2$What.is.your.age.

hist(age, col="grey")

boxplot(df2$What.is.your.age.)

****

****