615 Midterm Project: BU Healthy Minds Study

Sijia Li

December 6, 2016

Background

Healthy Minds Study is an annual web-based survey study examining mental health, service utilization, and related issues among BU students.

My role in the project is to work with the team to cleaned and explore data from the survey, in order to find relationship between variables such as drug utilization, anxiety and depression score, and compare the situation between target groups seperated by gender, race and year of school.

Demographics Data Cleaning

The data clients handed over to us are coded as binary data. The first part of this project is to create a demographic data frame with readable text information in it instead of binary data.

```
load("hms.rda")

# 1. academic status
educ <- hms[41:50]
educ$aca_status[educ$degree_bach == 1] <- "undergraduate"
educ$aca_status[rowSums(educ[2:5], na.rm = TRUE) >= 1] <- "graduate"

other <- subset(educ, degree_other == 1 & !is.na(degree_other_text))
other_text <- unique(other$degree_other_text)
other_text_indicator <- c(T, T, T, F, F, F, T, T, T, T, T, T, T, T, T,
                          T, F, F, F, T, T, T, F, F, T, F, F, F, T, T, F)
grad <- other_text[other_text_indicator]
educ$aca_status[educ$degree_other_text %in% grad] <- "graduate"
educ$aca_status[is.na(educ$aca_status) & educ$degree_other == 1 &
                  !is.na(educ$degree_other_text)] <- "other"

# 2. gender
gender <- hms$gender
gender[gender == 1] <- "male"
gender[gender == 2] <- "female"
gender[gender == 3 | gender == 4] <- "trans gender"
gender[gender == 5 | gender == 6] <- "other"


# 3. citizenship
citizenship <- hms$citizen
citizenship[citizenship == 1] <- "domestic"
citizenship[citizenship == 0] <- "international"

# 4. field of study
df_field <- hms[53:73]
fieldls <- c(
  'Humanities (history, languages, philosophy, etc.) ',
```

```
  'Natural sciences or mathematics ',
  'Social sciences (economics, psychology, etc.) ',
  'Architecture or urban planning ',
  'Art and design ',
  'Business ',
  'Dentistry ',
  'Education ',
  'Engineering ',
  'Law',
  'Medicine',
  'Music, theatre, or dance ',
  'Nursing ',
  'Pharmacy ',
  'Pre-professional',
  'Public health ',
  'Public policy ',
  'Social Work',
  'Undecided',
  'Other'
)
tmp1 <- apply(df_field[,1:20], 1, function(x) which(!is.na(x)))
tmp2 <- lapply(tmp1, function(x) x[1])
pos <- unlist(tmp2)
df_field$field <- fieldls[pos]


demographics <- data.frame(citizenship, gender, educ$aca_status, df_field$field)
names(demographics) <- c("citizenship", "gender", "academic_status", "field")
```

Now we have the Demographics data that will be used by the team throughout the project.

```
head(demographics)
```

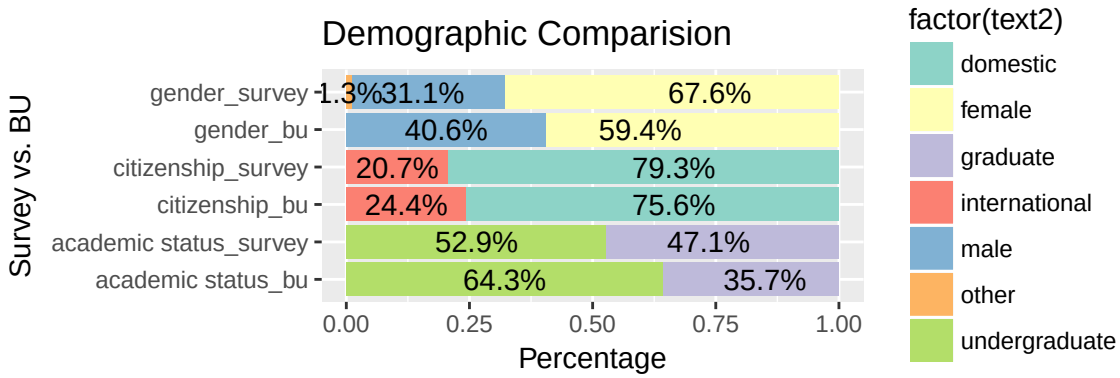
```
##     citizenship gender academic_status            field
## 1      domestic female   undergraduate Pre-professional
## 2 international female        graduate  Art and design
## 3      domestic female   undergraduate  Art and design
## 4      domestic female        graduate   Public health
## 5      domestic female   undergraduate     Engineering
## 6 international female   undergraduate   Public health
```

Profile

The next step is to build a profile of how representitive this survey data is. We compare three demographic features with the whole BU population. The percentages are from the BU official website.

```
library(ggplot2)
bu <- c(0.594, 0.406, 0.756, 0.244, 0.643, 0.357, 0.013)
survey <- c(0.676, 0.311, 0.793, 0.207, 0.529, 0.471)
text1 <- c("gender", "gender", "citizenship", "citizenship", "academic status", "academic status")
text2 <- c("female", "male", "domestic", "international", "undergraduate", "graduate")

df2 <- data.frame(c(text1, "gender", text1), c(text2, "other", text2),
                  c(rep("bu", 6), rep("survey", 7)), c(bu, survey))
```

```
names(df2) <- c("text1", "text2", "source", "data")
df3 <- df2
df3$source <- paste(df3$text1, df2$source, sep="_")
df3$pos <- c(0.5970, 0.2030, 0.665, 0.1220, 0.3215, 0.85,
             0.0065, 0.7380, 0.1555, 0.665, 0.1035, 0.2645, 0.7355)
ggplot(df3, aes(x = source, y = data, fill = factor(text2))) + geom_bar(stat = "identity") +
  geom_text(aes(y = pos, label = paste0(data*100, "%"))) +
  coord_flip() + scale_fill_brewer(palette = "Set3") +
  labs(x = "Survey vs. BU", y = "Percentage", title = "Demographic Comparision")
```



We can see that the survey data is a rather representative sample of the BU population, though we seem to have more female and undergraduate participants.

Mental Status

```
## mental module
mental <- data.frame(demographics, hms[117:204])

## drug use
mental_drug <- subset(mental, is.na(mental$drug_none) &
                        rowSums(mental[, c(77:83)], na.rm = TRUE) > 0)
colSums(mental_drug[c(77:83)], na.rm = TRUE)
```

```
##   drug_coc   drug_her   drug_met  drug_stim   drug_ect drug_other
##         32          0          2         44         10         22
##  drug_none
##          0
```

```
# drug_mar   drug_coc   drug_her   drug_met  drug_stim   drug_ect drug_other
# 372         32          0          2         44         10         22

# the logic here is, someone using any kind of drug will be categorized as "drug user"
mental$dataset <- "non drug user"
mental$dataset[is.na(mental$drug_none) &
                 rowSums(mental[, c(77:83)], na.rm = TRUE) > 0] <- "drug user"

table(mental$dataset)
```

```
##
##     drug user non drug user
##            82          2469
```

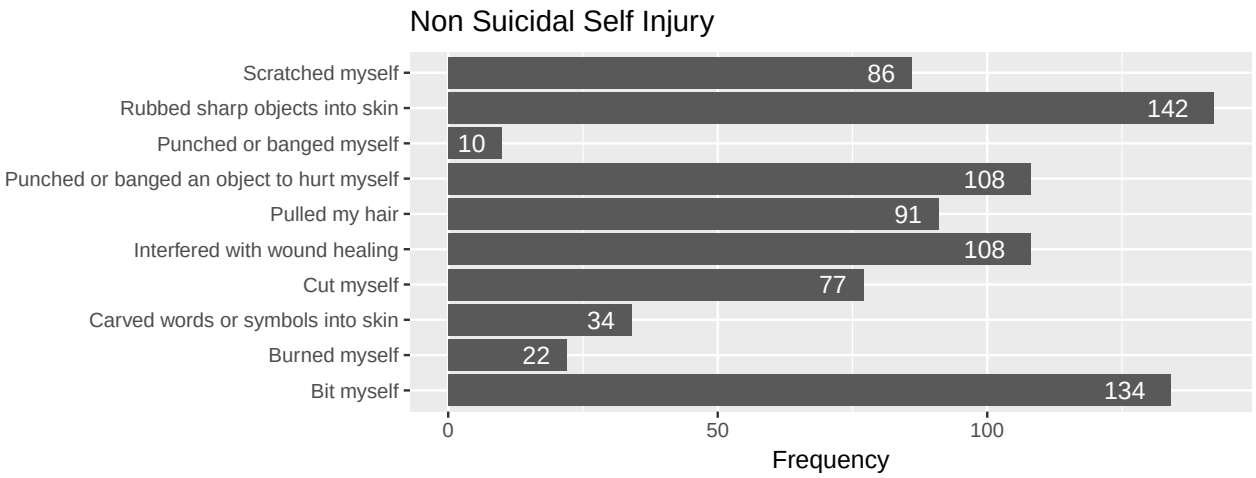
```r
# 5. Non-suicidal self-injury
suic <- mental[c(47:60)]

nrow(subset(suic, rowSums(suic[,c(1:11,13)], na.rm = TRUE) >= 1 )) # all suic data with values
```

```
## [1] 2077
```

```r
nrow(subset(suic, sib_none == 1))  #1692 out of 2077, 81.46% are none suicidal
```

```
## [1] 1692
```

```r
suic_vec <- names(na.omit(unlist(suic[,1:10])))
suic_vec <- gsub("[[:digit:]]","",suic_vec)
suic_table <- as.data.frame(table(suic_vec))
suic_table[1] <- c("Cut myself", "Burned myself",
                   "Punched or banged myself",
                   "Scratched myself",
                   "Pulled my hair","Bit myself",
                   "Interfered with wound healing",
                   "Carved words or symbols into skin",
                   "Rubbed sharp objects into skin",
                   "Punched or banged an object to hurt myself")
ggplot(data = suic_table, aes(x = suic_vec, y = Freq)) +
  geom_bar(stat="identity") + coord_flip() +
  ggtitle("Non Suicidal Self Injury") +
  geom_text(aes(label = suic_table$Freq), hjust=1.6, color="white", size=4) +
  labs(x = "", y = "Frequency")
```



Drug Use and Binge Drinking

Drug Use Distribution

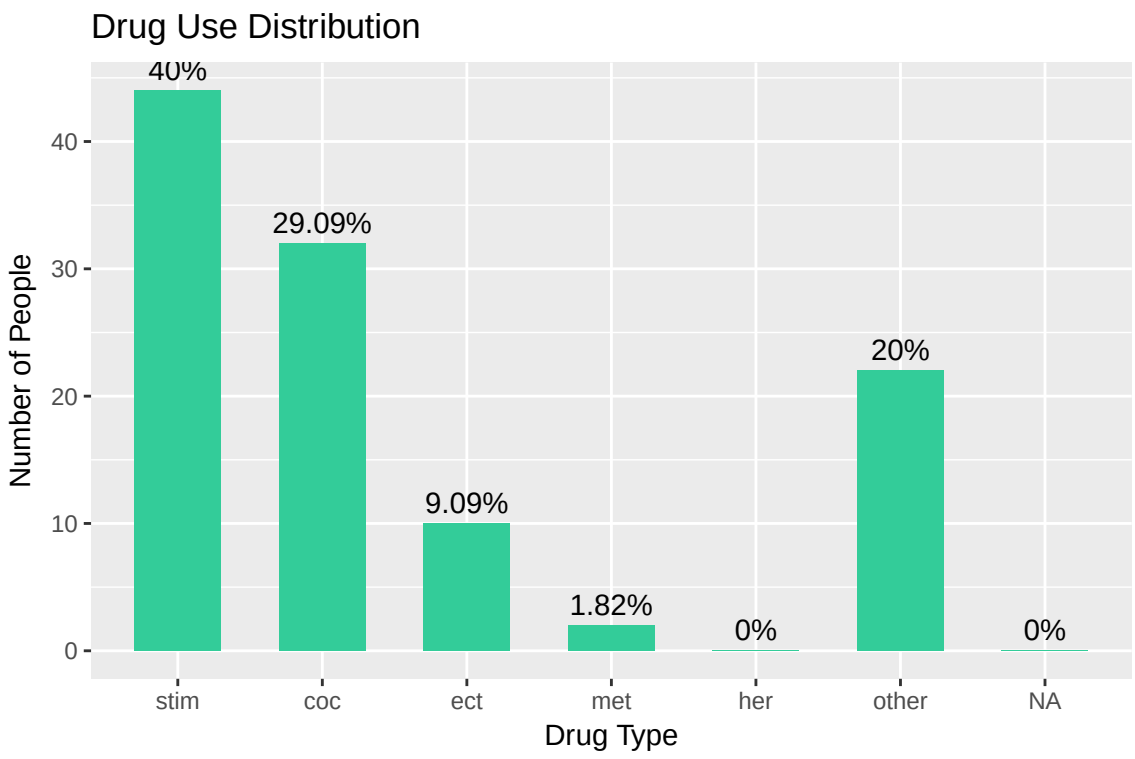
```r
mental <- data.frame(demographics, hms[117:204])

## drug use
mental_drug <- subset(mental, is.na(mental$drug_none) &
                        rowSums(mental[, c(77:83)], na.rm = TRUE) > 0)
colSums(mental_drug[c(77:83)], na.rm = TRUE)
```

```
##   drug_coc   drug_her   drug_met  drug_stim   drug_ect drug_other
##         32          0          2         44         10         22
##  drug_none
##          0
```

```
# the logic here is, someone using any kind of drug will be categorized as "drug user"
mental$dataset <- "non drug user"
mental$dataset[is.na(mental$drug_none) &
                 rowSums(mental[, c(77:83)], na.rm = TRUE) > 0] <- "drug user"

table(mental$dataset)
```

```
##
##     drug user non drug user
##            82          2469
```

```
d <- data.frame(colSums(mental_drug[c(77:83)], na.rm = TRUE))
d <- mutate(d, drug = gsub("drug_","", rownames(d)))
names(d) <- c("num", "drug.type")
d$perc <- paste0(as.character(round(100 * d$num/sum(d$num), 2)), "%")

d$drug.type <- factor(d$drug.type,
                      levels = c("mar", "stim", "coc", "ect", "met", "her", "other"))
ggplot(d, aes(x = drug.type, y = num)) +
  geom_bar(stat = "identity", fill = "#33CC99", width = 0.6) +
  labs(x = "Drug Type", y = "Number of People", title = "Drug Use Distribution") +
  geom_text(aes(label = perc), color = "black", vjust = -0.5)
```



Binge Drinking

```
binge <- hms[183:186]
binge <- subset(binge, !is.na(binge$alc_any))
binge.sum <- data.frame(binge$alc_any, rowSums(binge[2:4], na.rm = TRUE))
names(binge.sum) = c("drink.or.not", "times")
sum(binge.sum$binge.alc_any)
```

```
## [1] 0
```

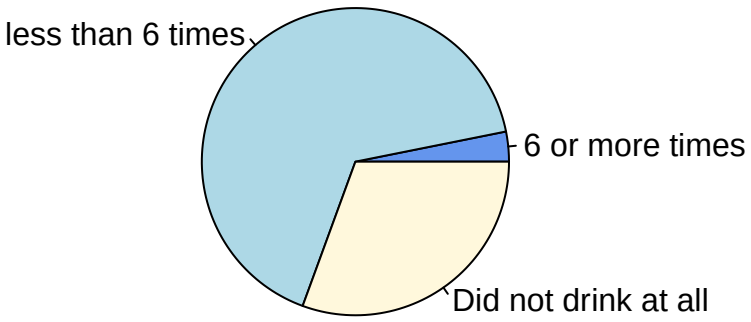
```
#1521
length(binge.sum$binge.alc_any)
```

```
## [1] 0
```

```
#2191
sum(binge.sum$rowSums.binge.2.4...na.rm...TRUE. >= 5)
```

```
## [1] 0
```

```
#69

pie(c(69, 1452, 670), labels = c("6 or more times", "less than 6 times", "Did not drink at all"),
    col = c("cornflowerblue", "lightblue", "cornsilk"),
    main = "times of participant drinking alcohol over past 2 weeks")
```

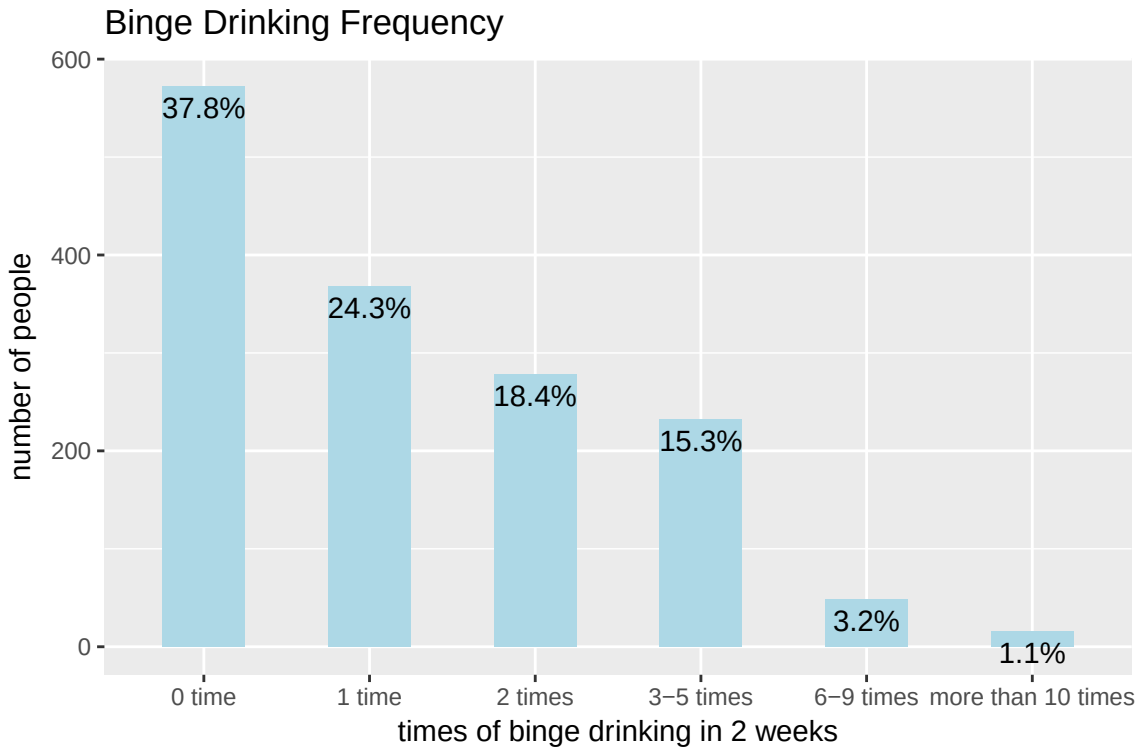



```
t <- binge.sum$rowSums.binge.2.4...na.rm...TRUE.
t <- t[t != 0 & t != 7]
t[t == 1] <- "0 time"
t[t == 2] <- "1 time"
t[t == 3] <- "2 times"
t[t == 4] <- "3-5 times"
t[t == 5] <- "6-9 times"
t[t == 6] <- "more than 10 times"
table(t)
```

```
## < table of extent 0 >
```

```
# 0 time       1 time       2 times    3-5 times     6-9 times   more than 10 times
# 572          368          278        232           48          16
freq <- c(572, 368, 278, 232, 48, 16)
perc <- sapply(freq, function(x) x/1514)
```

```
perc.t <- paste0(round(perc*100,1),"%")
text <- c("0 time","1 time","2 times","3-5 times","6-9 times","more than 10 times")
df <- data.frame(text, freq, perc.t)

g <- ggplot(df, aes(x = text, y = freq))
g + geom_bar(width=.5, fill = "lightblue", stat = "identity") + geom_text(aes(label = perc.t,vjust=1.5)
  labs(x = "times of binge drinking in 2 weeks", y = "number of people") + ggtitle("Binge Drinking Frequ
```



Summary

This report include all the things that have been approved and used in the final presentation. I'm taking this report, this opportunity as a reflection of all the code I've been writing and all the exploratory data analysis that I've been doing. I can definitely see the progress I'm making from day one when I started cleaning demographics information. This report will also be updated during winter break - I'm gonna use this to practice using sweave and latex. Happy Holidays!