650 Final Project

Data Preprocessing

Yubo SHAO

2021-11-18

```
library(NHANES)
raw.data <- force(NHANES) #import dateset
dim(raw.data)  # raw data has 10000 observations and 76 variables
## [1] 10000    76
data2 <- unique(raw.data)  # delete those duplicate observations
dim(data2)  # 7832 observations left
## [1] 7832   76

## still exist some people who have the same ID but different covariates
## one single person can have at most 5 observations in the dataset
range(table(data2$ID))
## [1] 1 5
```

Find out why people with the same ID can have different covariates:

```
longitudinal.variable <- c(NULL)  # a vector that contains the name of repeated measurement variable
for (i in unique(data2$ID)) { #63149
  temp.data <- data2[which(data2$ID == i),]
  if (nrow(temp.data) > 1){
    for (j in 1:dim(data2)[2]) {
      if (nrow(unique(temp.data[,j])) >1){
        longitudinal.variable <- c(longitudinal.variable, colnames(temp.data)[j])
      }
    }
  }
}
head(longitudinal.variable, n = 3)
## [1] "PhysActiveDays" "PhysActiveDays" "PhysActiveDays"
unique(longitudinal.variable)  # it can be seen that only "PhysActiveDays" has repeated measurement
## [1] "PhysActiveDays"
```

The defination of "PhysActiveDays" is: Number of days in a typical week that participant does moderate or vigorous-intensity activity. (Ranged from 1 to 7, but also exist missing data, i.e. "NA")

```
unique(data2$PhysActiveDays)
## [1] NA  5  7  1  2  3  4  6
max.repeated.ID <-
  names(which(table(data2$ID) == max(table(data2$ID))))
max.repeated.ID.selected.data <-
  data2[which(data2$ID %in% as.double(max.repeated.ID)),
```

```
        c("ID", "SurveyYr", "PhysActive", "PhysActiveDays")]
print(max.repeated.ID.selected.data)
## # A tibble: 15 x 4
##       ID SurveyYr PhysActive PhysActiveDays
##    <int> <fct>    <fct>               <int>
##  1 63149 2011_12  Yes                    NA
##  2 63149 2011_12  Yes                     5
##  3 63149 2011_12  Yes                     1
##  4 63149 2011_12  Yes                     3
##  5 63149 2011_12  Yes                     2
##  6 63297 2011_12  Yes                    NA
##  7 63297 2011_12  Yes                     6
##  8 63297 2011_12  Yes                     3
##  9 63297 2011_12  Yes                     2
## 10 63297 2011_12  Yes                     4
## 11 67118 2011_12  No                      5
## 12 67118 2011_12  No                      3
## 13 67118 2011_12  No                      4
## 14 67118 2011_12  No                     NA
## 15 67118 2011_12  No                      1
```

- It can be seen that the same person could have different reported "PhysActiveDays", but the "SurveyYr" is the same.
- I cannot figure out why this occurs, one possible reason might be: the "PhysActiveDays" is measured each week, but we donnot have any week-related variables in our dataset.
- Another variable "PhysActive" has a similar defination. It is a binary variable, and we can use this one as alternative.

We delete "PhysActiveDays", and then delete the duplicate observations.

```
data3 <- data2 %>% select(-c("PhysActiveDays"))
data4 <- unique(data3)
dim(data4)  # 6779 observations left
```

```
## [1] 6779   75
```

```
length(unique(data4$ID)) # each ID is unique
```

```
## [1] 6779
```

Define the study population

From the figures below, you may also think that we need to define our study population first. Many variables were only meansured among specific population (e.g."Alcohol"-related variables were only reported by people aged > 18), and if we donnot define the study population, there will be a great number of missing data.

Choose potenial confounders

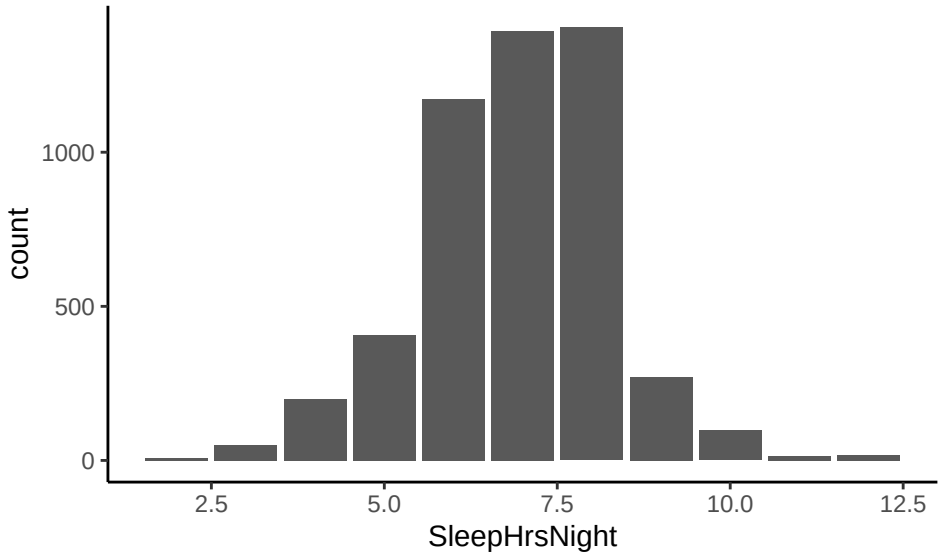
Before conducting formal analysis, we delete those variables that cannot be a potential predictor of sleep time

`Below, I mention those variables that may need to be discussed further`

- (1) SleepHrsNight: (Outcome) Too many missing data.

  Generally, it not easy to do data imputation on outcome. So we just delete those observations with missing Y.

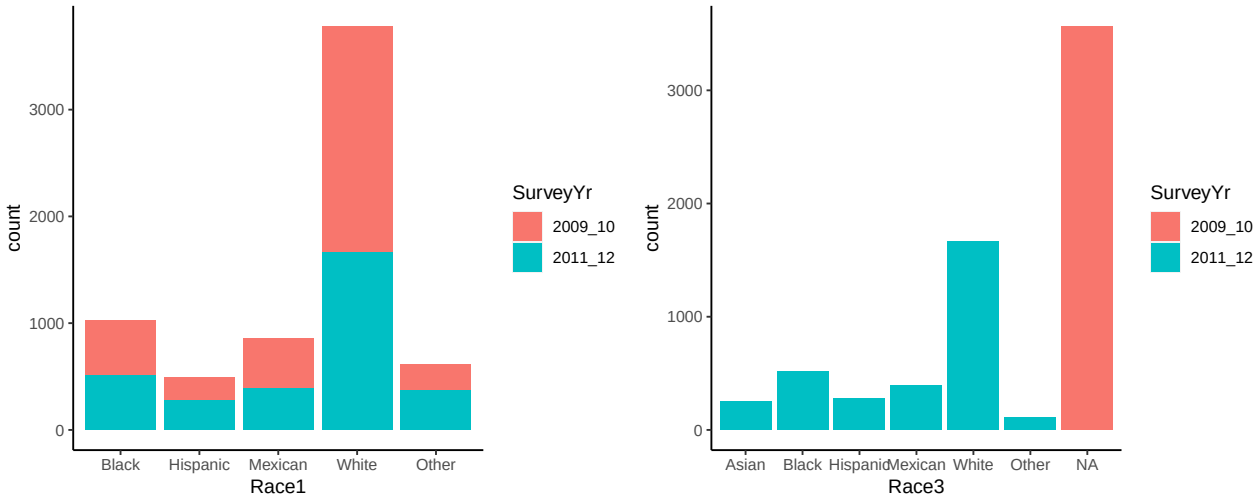
```
sum(is.na(data4$SleepHrsNight))  #1744 missing value
## [1] 1744
SleepHrsNight.plot <- ggplot() + geom_bar(aes(SleepHrsNight), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
SleepHrsNight.plot
```




```
## I don't delete those observations with missing outcome here,
## but you can do this yourself and see what will change.

# data4 <- data4[which(!is.na(data4$SleepHrsNight)),]
```

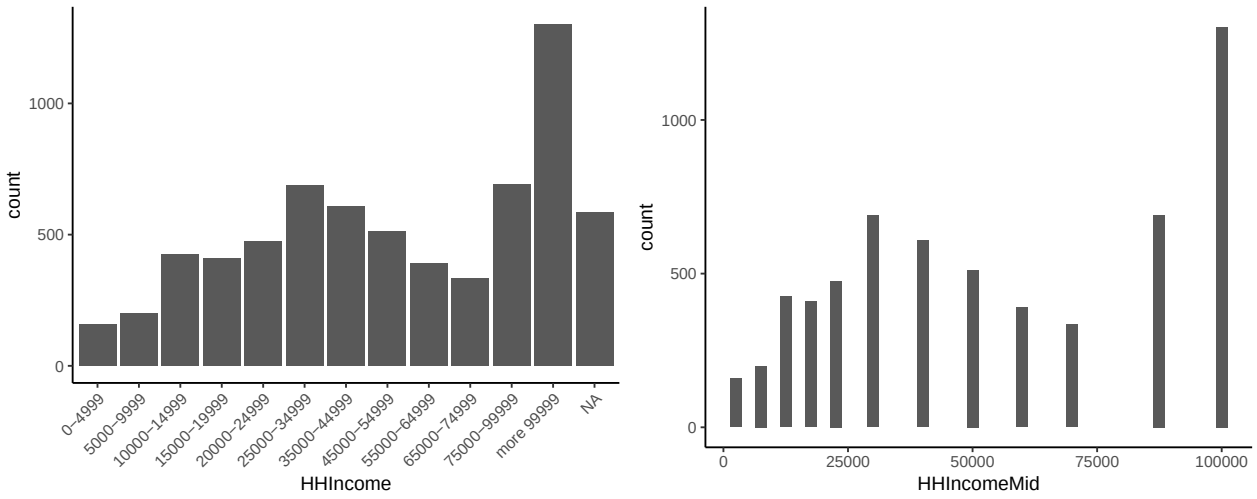
- (2) "Race": There're 2 race variables in our dataset. "Race3" has many missing value, because it is not avaliable for 2009-2010.

```
race1.plot <- ggplot() + geom_bar(aes(Race1, fill = SurveyYr), data4) +
  theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
race3.plot <- ggplot() + geom_bar(aes(Race3, fill = SurveyYr), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
grid.arrange(race1.plot, race3.plot, nrow = 1)
```



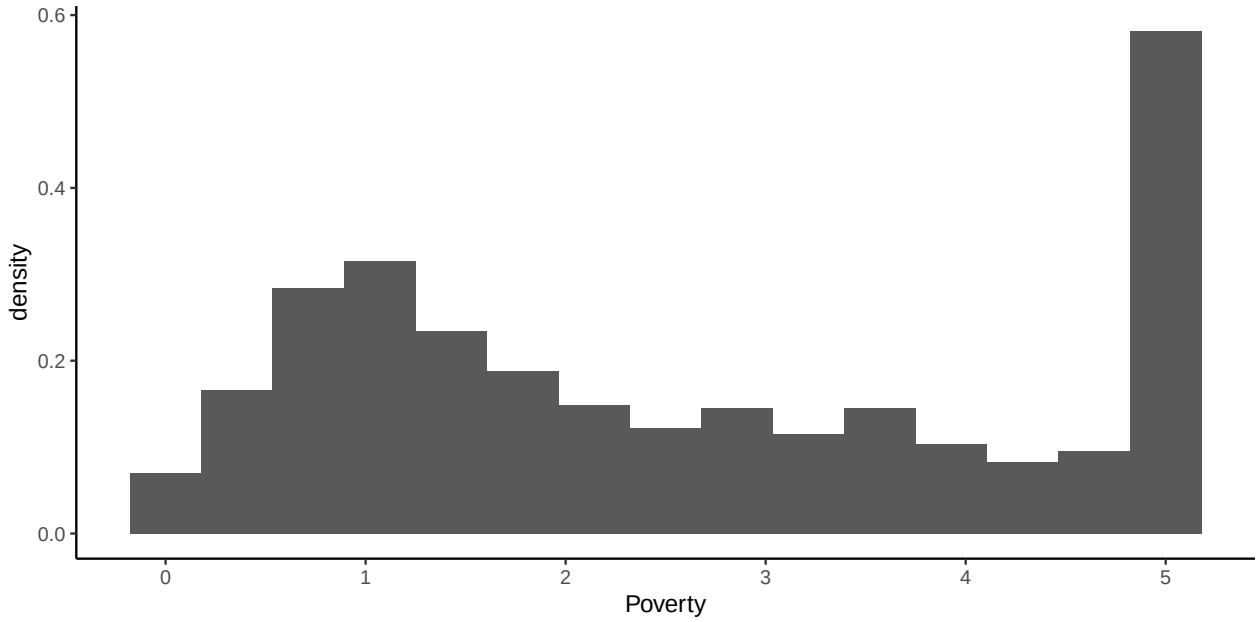
- (3) "Income": We have 3 income-related variables: HHIncome, HHIncomeMid and Poverty. "HHIncome" and "HHIncomeMid" are categorical variables, and have a similar definition (and thus have a similar distribution).

```
HHIncome.plot <- ggplot() + geom_bar(aes(HHIncome), data4) +
  theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"), axis.text.x = element_text(angle = 45, hjust = 1))
HHIncomeMid.plot <- ggplot() + geom_bar(aes(HHIncomeMid), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
grid.arrange(HHIncome.plot, HHIncomeMid.plot, nrow = 1)
```



"Poverty" is a continuous variable, which equals the ratio of family income to poverty guidelines. The figure looks a little bit weird, but about 17.98% people reported that their "poverty ratio" equals 5 (the largest value). OUR PEOPLE ARE RICH!

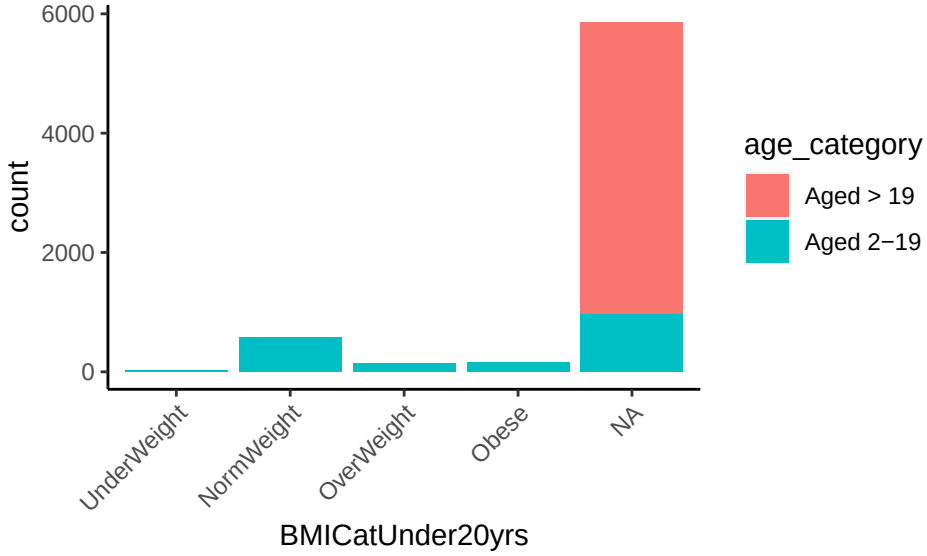
```
Poverty.plot <- ggplot() + geom_histogram(aes(Poverty, y = ..density..), data4, bins = 15) +
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
Poverty.plot
```



- (4) "Weight", "Height" and "BMI": similar measurements, and we could only keep "BMI".

  3 variables measure "BMI": expect the original "BMI", we also have a variable called "BMICatUnder20yrs", which is only reported by people aged 2-19years. (From the figure, we should not use "BMICatUnder20yrs")

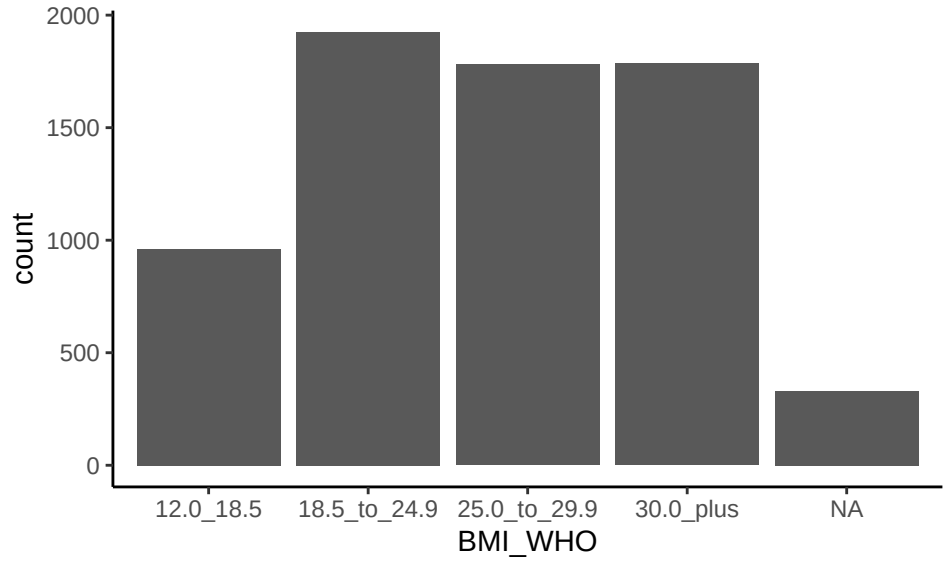
```
data4$age_category[data4$Age >= 2 & data4$Age <= 19] <- "Aged 2-19"
data4$age_category[data4$Age < 2 | data4$Age > 19] <- "Aged > 19"
BMICatUnder20yrs3.plot <- ggplot() + geom_bar(aes(BMICatUnder20yrs, fill = age_category), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"), axis.text.x = element_text(angle = 45, hjust = 1))
BMICatUnder20yrs3.plot
```



"BMI_WHO" is another BMI-related variable. It is a categorical variable, and defined by WHO.

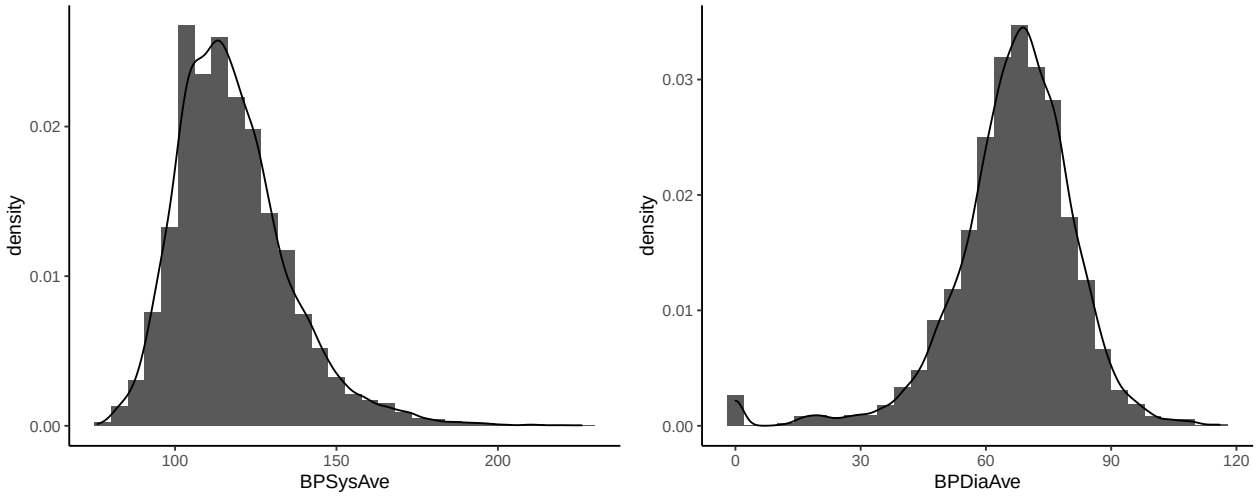
```
BMI_WHO.plot <- ggplot() + geom_bar(aes(BMI_WHO), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```

```
BMI_WHO.plot
```



- (5) Blood pressures: Both SBP & DBP have at most 3 readings, and we should use the average BP if necessary.

```
BPSysAve.plot <- ggplot() + geom_histogram(aes(BPSysAve, y = ..density..), data4, bins = 30) +
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black")) + geom_density(aes(BPSysAve), data4)
BPDiaAve.plot <- ggplot() + geom_histogram(aes(BPDiaAve, y = ..density..), data4, bins = 30) +
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black")) + geom_density(aes(BPDiaAve), data4)
grid.arrange(BPSysAve.plot, BPDiaAve.plot, nrow = 1)
```



From the right figure, we can find that some people reported a "0" diastolic blood pressure. Those should not be "ourliers", but "wrong data". If we want to use diastolic blood pressure, we need to delete those observations first.

```
length(which(data4$BPDiaAve == 0)) # 59 people reported a "0" diastolic blood pressure
```

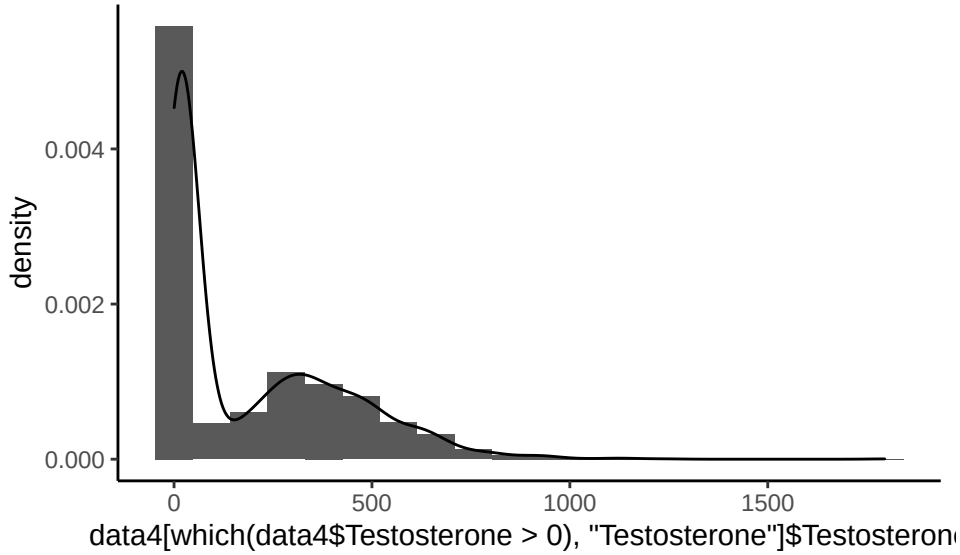
```
## [1] 59
```

- (6) Testosterone: I don't know what is the function of "testoterone", but I have simply googled it,

and I found it could be a strong predictor of sleep time. (you can check the article here: *"The relationship between sleep disorders and testosterone in men"*)

According to *MedicalNewsToday*: Normal testosterone levels in men are around 280 to 1,100 nanograms per deciliter (ng/dL).Women secrete much lower amounts, with normal levels considered to be between 15 and 70 ng/dL.

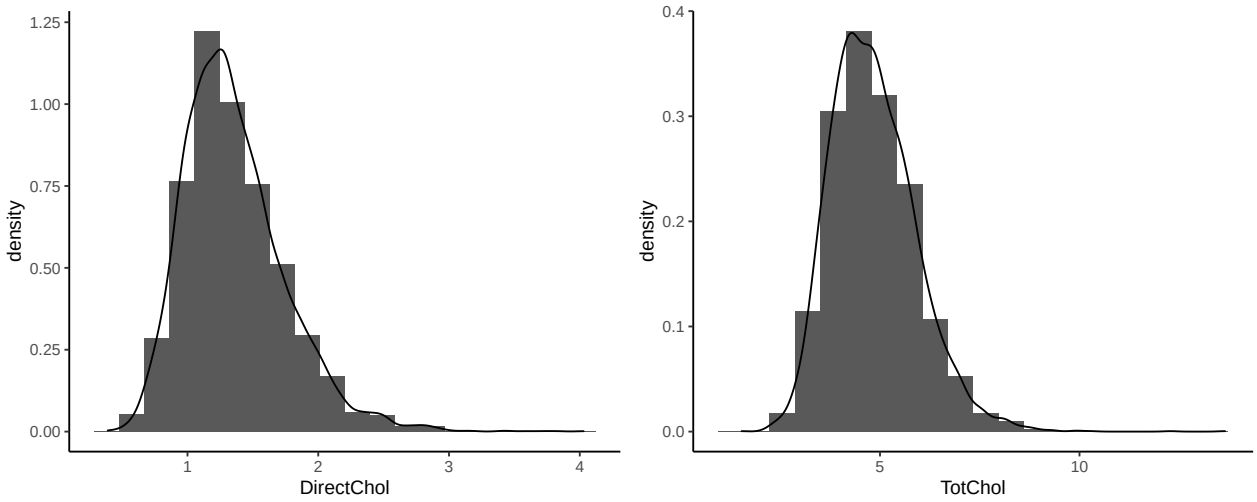
```
BPSysAve.plot <- ggplot() +
  geom_histogram(aes(data4[which(data4$Testosterone > 0),"Testosterone"]$Testosterone, y = ..density..)
  geom_density(aes(data4[which(data4$Testosterone > 0),"Testosterone"]$Testosterone)) +
  theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
BPSysAve.plot
```




```
# the minimum Testosterone is 0.25ng/dL
min(data4[which(is.na(data4$Testosterone) == FALSE),"Testosterone"]$Testosterone)
## [1] 0.25
# but we have too many missing value
sum(is.na(data4$Testosterone))
## [1] 4218
```

- (7) Direct HDL & Total HDL could also be important confounders, you can find the article here: *"Associations of Usual Sleep Duration with Serum Lipid and Lipoprotein Levels"*

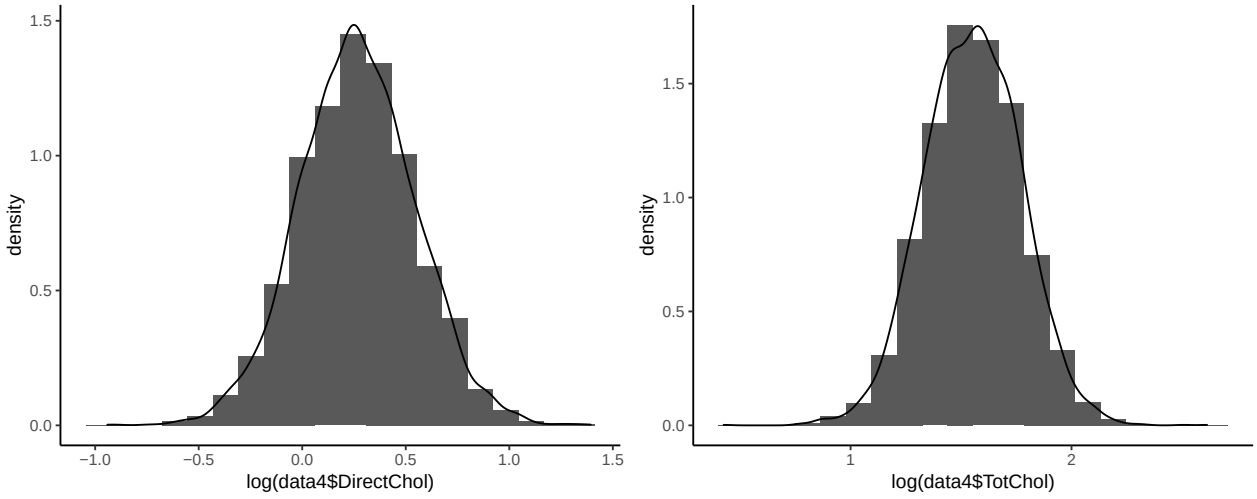
```
DirectChol.plot <- ggplot() + geom_histogram(aes(DirectChol, y = ..density..), data4, bins = 20) +
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black")) + geom_density(aes(DirectChol), data4)
TotChol.plot <- ggplot() + geom_histogram(aes(TotChol, y = ..density..), data4, bins = 20) +
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black")) + geom_density(aes(TotChol), data4)
grid.arrange(DirectChol.plot, TotChol.plot, nrow = 1)
```



It can be seen that both "Direct HDL" and "Total HDL" are right skewed. So we need to transform HDL if we want to use this two variables.

If we do a log-transformation of HDL, it looks much better:

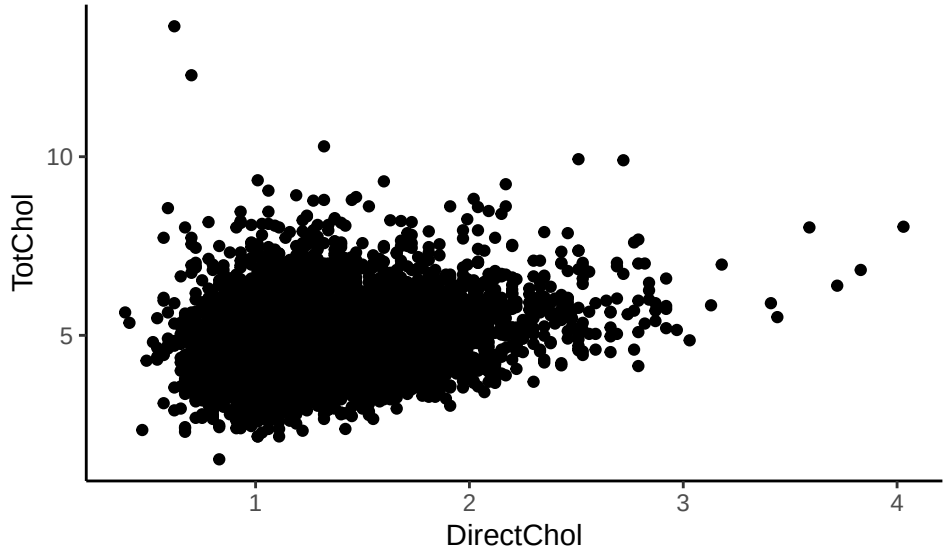
```
log.DirectChol.plot <- ggplot() + geom_histogram(aes(log(data4$DirectChol), y = ..density..), bins = 20)
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black")) + geom_density(aes(log(data4$DirectChol)))
log.TotChol.plot <- ggplot() + geom_histogram(aes(log(data4$TotChol), y = ..density..), bins = 20) +
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black")) + geom_density(aes(log(data4$TotChol)))
grid.arrange(log.DirectChol.plot, log.TotChol.plot, nrow = 1)
```



And we cannot use both, because they are high-correlated with each other.

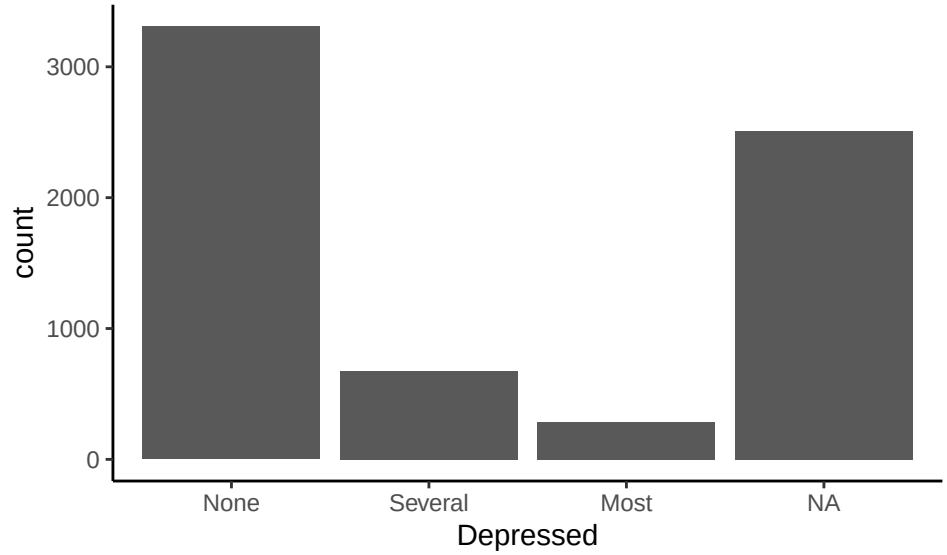
```
cor.test(data4$DirectChol, data4$TotChol)
##
##  Pearson's product-moment correlation
##
## data:  data4$DirectChol and data4$TotChol
## t = 15.792, df = 5594, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1813721 0.2315400
```

```
## sample estimates:
##       cor
## 0.2065919
scatter.HDL <- ggplot() + geom_point(aes(DirectChol, TotChol), data4) +
      theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
scatter.HDL
```



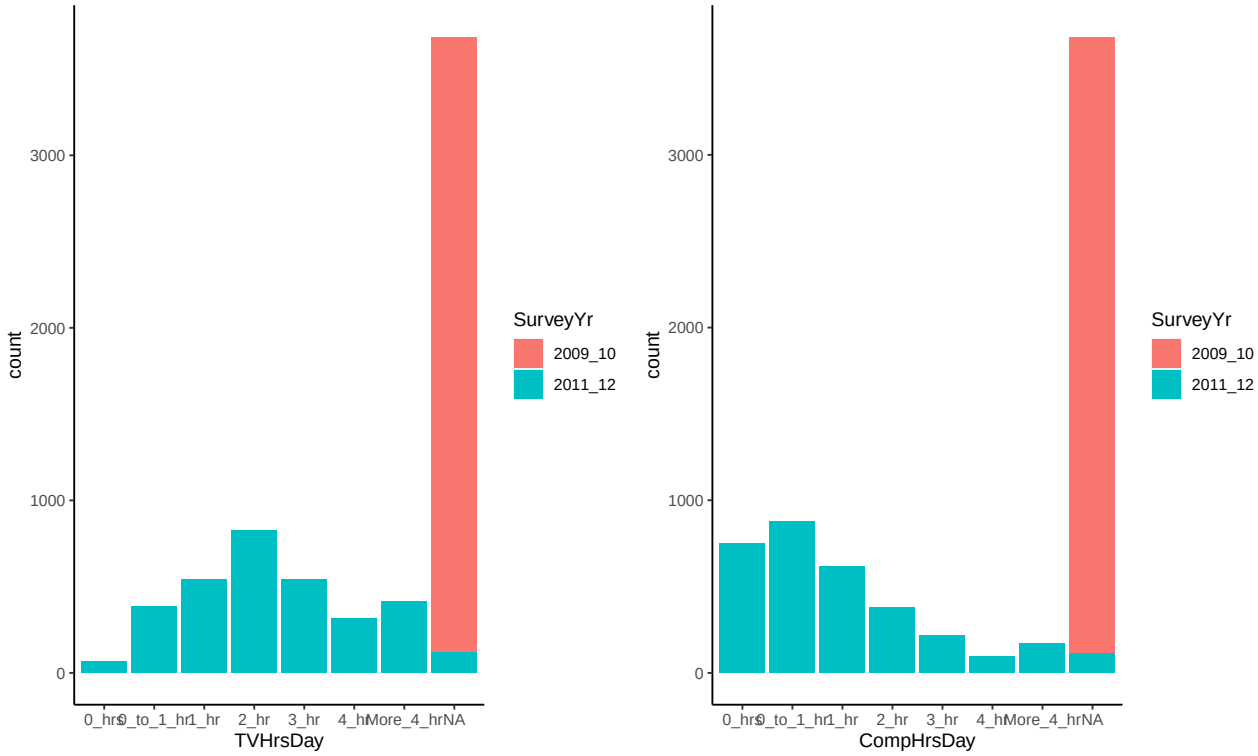
- (8) Depressed: too many missing data.

```
sum(is.na(data4$Depressed)) # count of missing data
## [1] 2507
Depressed.plot <- ggplot() + geom_bar(aes(Depressed), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
Depressed.plot
```



- (9) TVHrsDay & CompHrsDay: I delete them, because it's hard to find articles, and there're too many missing value (Not available 2009-2010).

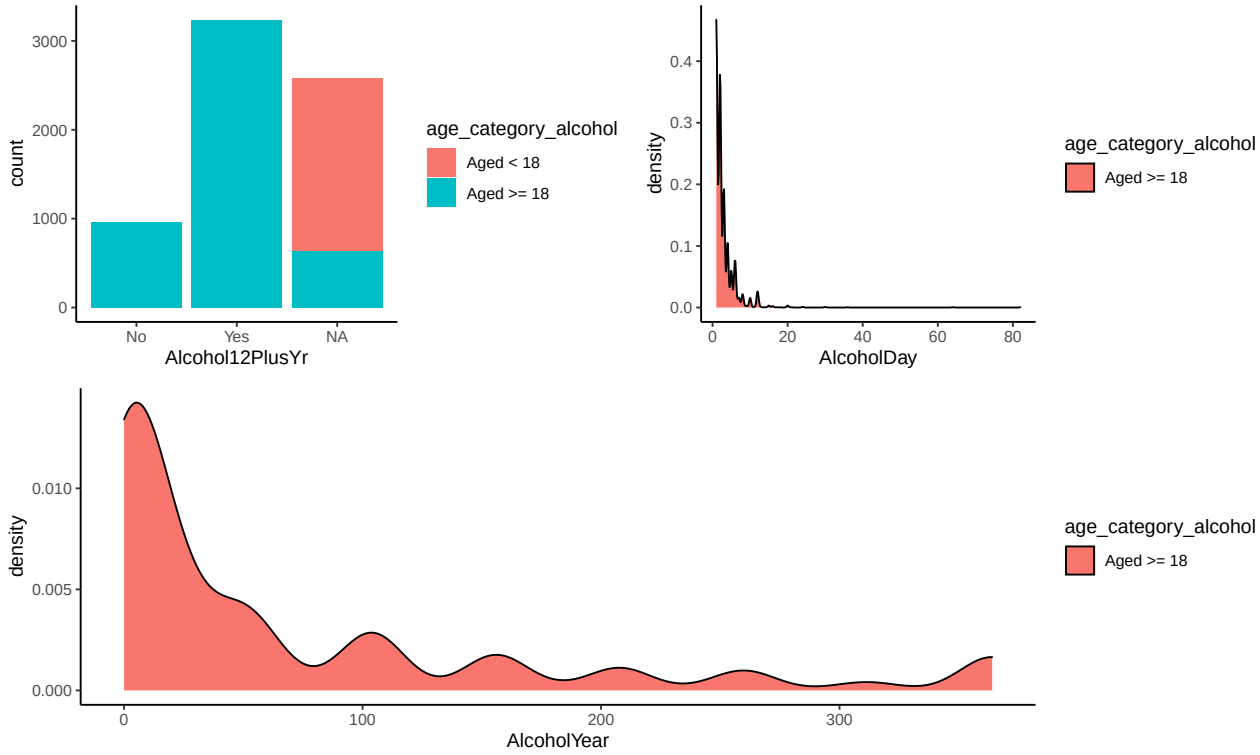
```
## TVHrsDay: Number of hours per day on average participant watched TV
sum(is.na(data4$TVHrsDay))
## [1] 3686
## CompHrsDay:  Number of hours per day on average participant used a computer or gaming device
sum(is.na(data4$CompHrsDay))
## [1] 3683

TVHrsDay.plot <- ggplot() + geom_bar(aes(TVHrsDay, fill = SurveyYr), data4) +
  theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
CompHrsDay.plot <- ggplot() + geom_bar(aes(CompHrsDay, fill = SurveyYr), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
grid.arrange(TVHrsDay.plot, CompHrsDay.plot, nrow = 1)
```



- (10) Alcohol-related: Alcohol12PlusYr & AlcoholDay & AlcoholYear. Reported for participants 18 years or older.

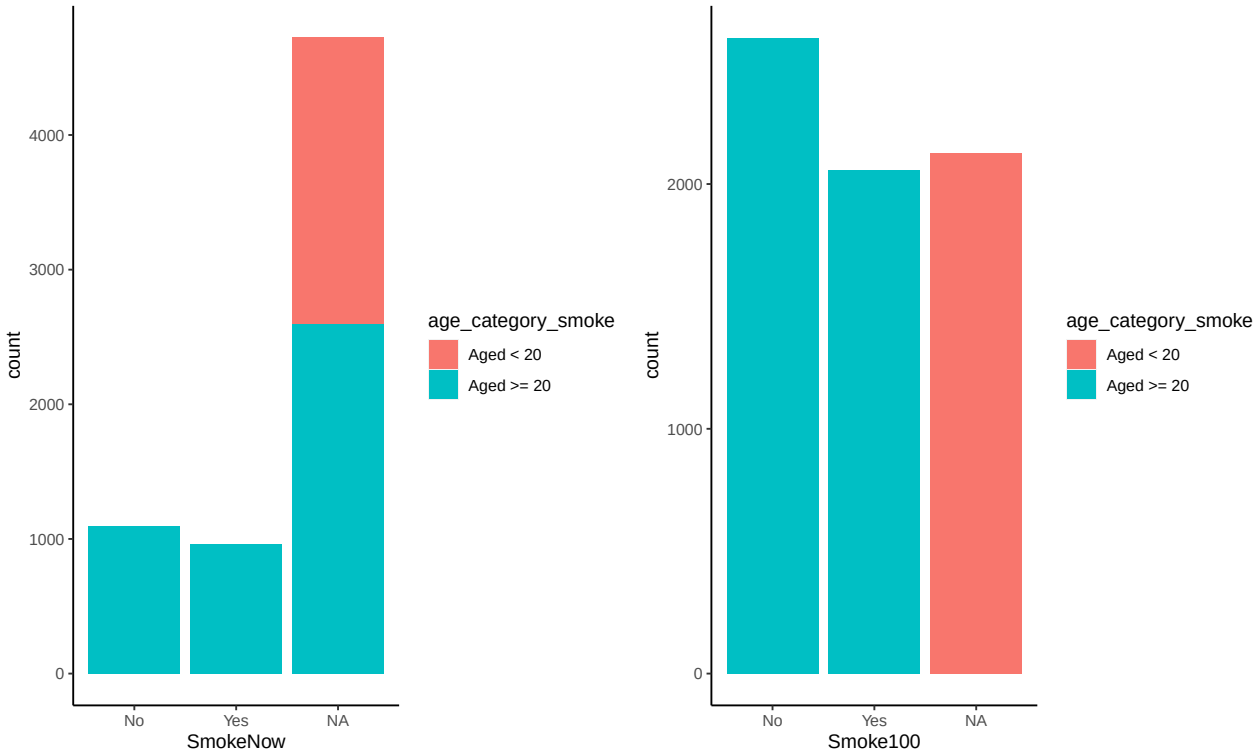
```
data4$age_category_alcohol[data4$Age >= 18] <- "Aged >= 18"
data4$age_category_alcohol[data4$Age < 18] <- "Aged < 18"

## Participant has consumed at least 12 drinks of any type of alcoholic beverage in any one year
Alcohol12PlusYr.plot <- ggplot() + geom_bar(aes(Alcohol12PlusYr, fill = age_category_alcohol), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

## Average number of drinks consumed on days that participant drank alcoholic beverages.
AlcoholDay.plot <- ggplot() + geom_density(aes(AlcoholDay, fill = age_category_alcohol), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
```

```
        axis.line = element_line(colour = "black"))

## Estimated number of days over the past year that participant drank alcoholic beverages.
AlcoholYear.plot <- ggplot() + geom_density(aes(AlcoholYear, fill = age_category_alcohol), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

lay <- rbind(c(1, 2), c(3, 3))
grid.arrange(Alcohol12PlusYr.plot, AlcoholDay.plot, AlcoholYear.plot, layout_matrix = lay)
```



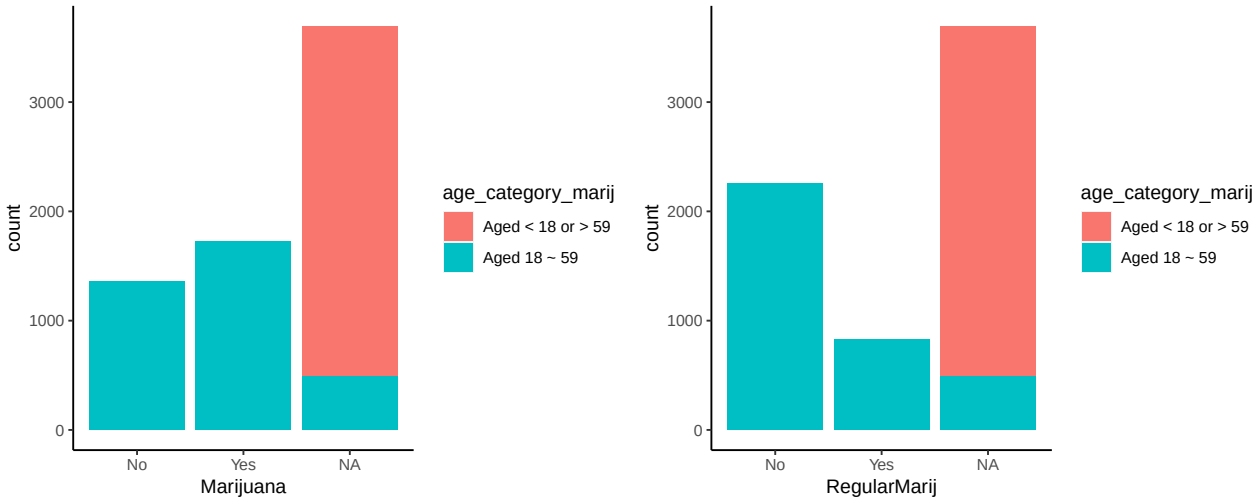
- (11) Smoke-related: SmokeNow & Smoke100. Reported for participants aged 20 years or older.(Yes/No)

```
data4$age_category_smoke[data4$Age >= 20] <- "Aged >= 20"
data4$age_category_smoke[data4$Age < 20] <- "Aged < 20"

## Study participant currently smokes cigarettes regularly
SmokeNow.plot <- ggplot() + geom_bar(aes(SmokeNow, fill = age_category_smoke), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

## Study participant has smoked at least 100 cigarettes in their entire life
Smoke100.plot <- ggplot() + geom_bar(aes(Smoke100, fill = age_category_smoke), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

grid.arrange(SmokeNow.plot, Smoke100.plot, nrow = 1)
```

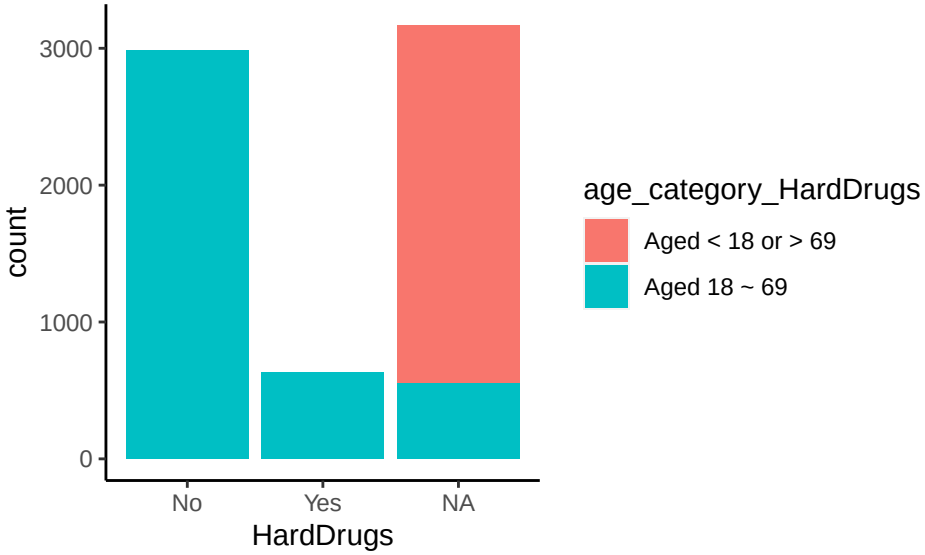


- (12) Drugs-related: Marijuana & RegularMarij & HardDrugs

```
data4$age_category_marij[data4$Age >= 18 & data4$Age <= 59] <- "Aged 18 ~ 59"
data4$age_category_marij[data4$Age < 18 | data4$Age > 59] <- "Aged < 18 or > 59"
##  Participant has tried marijuana.
Marijuana.plot <- ggplot() + geom_bar(aes(Marijuana, fill = age_category_marij), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

## Participant has been/is a regular marijuana user (used at least once a month for a year).
RegularMarij.plot <- ggplot() + geom_bar(aes(RegularMarij, fill = age_category_marij), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
grid.arrange(Marijuana.plot, RegularMarij.plot, nrow = 1)
```




```
data4$age_category_HardDrugs[data4$Age >= 18 & data4$Age <= 69] <- "Aged 18 ~ 69"
data4$age_category_HardDrugs[data4$Age < 18 | data4$Age > 69] <- "Aged < 18 or > 69"
## Participant has tried cocaine, crack cocaine, heroin or methamphetamine
HardDrugs.plot <- ggplot() + geom_bar(aes(HardDrugs, fill = age_category_HardDrugs), data4) +
    theme(panel.grid = element_blank(), panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
HardDrugs.plot
```



- (13) Sex-related: I delete them :)

```
data.clean <-data4 %>% select(c("SurveyYr", "SleepHrsNight", "Gender", "Age", "AgeDecade", "Race1",
                          "MaritalStatus", "HHIncome","HHIncomeMid", "Poverty", "BMI", "BMI_WHO",
                          "BPSysAve", "BPDiaAve", "Testosterone","DirectChol", "TotChol",
                          "Diabetes", "HealthGen", "Depressed", "nBabies", "PhysActive",
                          "Alcohol12PlusYr", "AlcoholDay", "AlcoholYear", "SmokeNow", "Smoke100",
                          "Marijuana", "RegularMarij", "HardDrugs"))

# write.csv(data.clean, file = "/Users/shaoyubo/Desktop/UMich/Course/Fall 2021/BIOSTAT 650/Final_Projec

# save(data.clean, file="/Users/shaoyubo/Desktop/UMich/Course/Fall 2021/BIOSTAT 650/Final_Project/Final_
```