# Exploring the BRFSS data

## Setup

### Load packages

```
library(ggplot2)
library(dplyr)
```

### Load data

```
load("brfss2013.RData")
```

---

---

## Part 2: Research questions

**Research quesion 1:** What is the distribution of income in the dataset? Is the distribution different for male and female?

**Research quesion 2:** Is peoples health related to their height and weight?

**Research quesion 3:** Is mental health related to physical health? How does sleeping affect mental and physical health?

---

## Part 3: Exploratory data analysis

There are 330 variables, let's select a few we are interested in.

```
health_data <- brfss2013 %>% select(genhlth, sex, income2, weight2, height3, physhlth, menthlth, wtkg3,
head(health_data)
```

```
##     genhlth    sex           income2 weight2 height3 physhlth menthlth
## 1      Fair Female Less than $75,000     250     507       30       29
## 2      Good Female   $75,000 or more     127     510        0        0
## 3      Good Female   $75,000 or more     160     504        3        2
## 4 Very good Female Less than $75,000     128     504        2        0
## 5      Good   Male Less than $50,000     265     600       10        2
## 6 Very good Female   $75,000 or more     225     503        0        0
##   wtkg3 htm4 sleptim1
## 1 11340  170       NA
## 2  5761  178        6
## 3  7257  163        9
## 4  5806  163        8
## 5 12020  183        6
## 6 10206  160        8
```

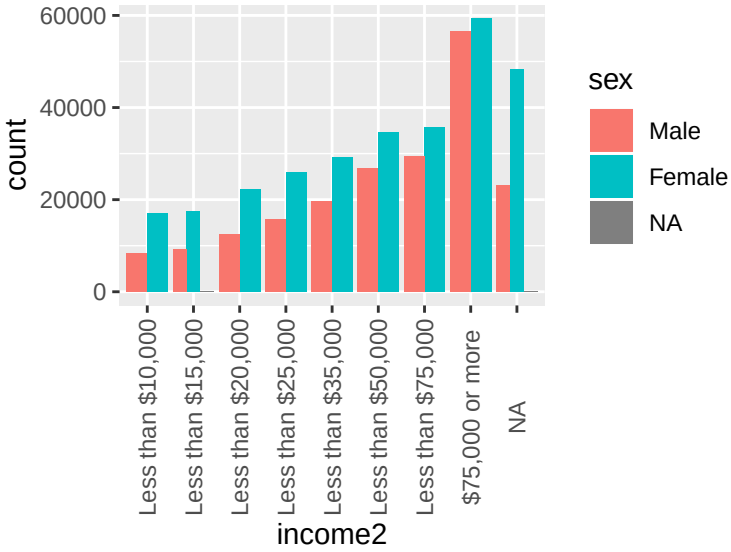
**Research quesion 1:**

The income has already be binned, so we will use barplot to see its distribution. We also make the barplot for male and female separately.

```
unique(health_data$income2)
```

```
## [1] Less than $75,000 $75,000 or more   Less than $50,000 <NA>
## [5] Less than $25,000 Less than $10,000 Less than $20,000 Less than $35,000
## [9] Less than $15,000
## 8 Levels: Less than $10,000 Less than $15,000 ... $75,000 or more
```

```
ggplot(brfss2013, aes(x=income2, fill=sex)) +
  geom_bar(position='dodge') +
  theme(axis.text.x=element_text(angle=90,hjust=0.5,vjust=0.5))
```




We can see that in the low income population, there are more females than males. The ratio of male becomes higher in high income groups. Overall the dataset consists of more female samples.

**Research quesion 2:**

First, we can compute the median of individules' weights and heights in each health group.

```
#create weight in kilogram
health_data$weight = health_data$wtkg3/100
#create height in centimeter
health_data$height = health_data$htm4

health_data %>% group_by(genhlth) %>%
  summarise(median_weight = median(weight, na.rm=TRUE), median_height = median(height, na.rm=TRUE))
```

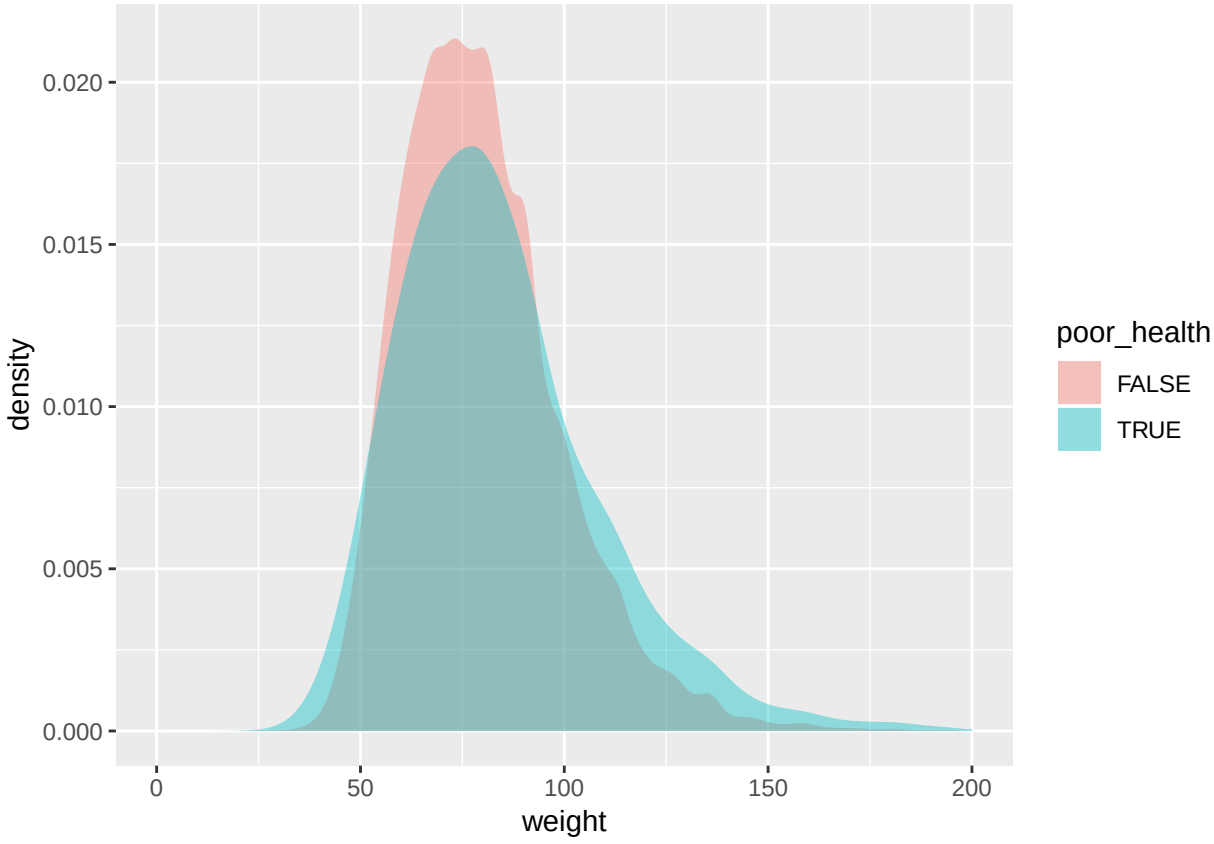
```
## # A tibble: 6 x 3
##   genhlth   median_weight median_height
##   <fct>             <dbl>         <int>
## 1 Excellent          72.6           170
## 2 Very good          77.1           170
## 3 Good               79.8           168
## 4 Fair               81.6           168
## 5 Poor               80.7           168
## 6 <NA>               75.8           168
```

It seems the is a negative correlation between health level and weight: the median of weight is higher in less healthy groups. Also there is a somewhat positive but weaker correlation between health level and height.

Let's visualize these effect using density plot of weight and height for healthy and non-healthy people. We use density plot instead of histogram so that the healthy and non-healthy group can be plotted in the same scale. To simplify the grouping, we define a new logical variable 'poor_health' which is True for genhlth == 'Poor'

```
health_data$poor_health = health_data$genhlth == 'Poor'

health_data %>%
  filter(!is.na(poor_health)) %>%
    ggplot( aes(x=weight, fill=poor_health, na.rm = TRUE)) +
    geom_density(col=NA, alpha=0.4, adjust=2) +
    xlim(0, 200)
```

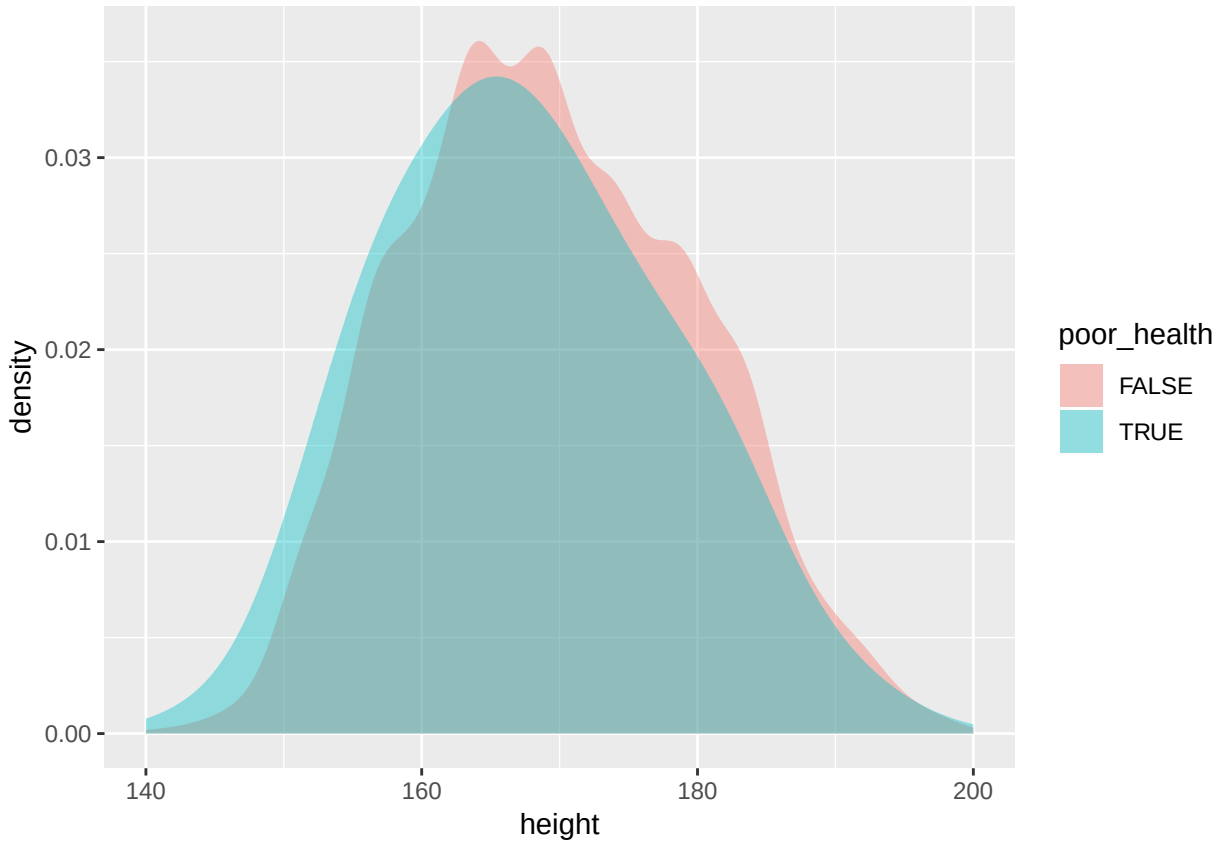
```
## Warning: Removed 20725 rows containing non-finite values (stat_density).
```




We can see the weight distribution poor-health group has a longer tail at both large and small weight region. This means being either too slim or too heavy is more likely to be unhealthy.

```
health_data %>%
  filter(!is.na(poor_health)) %>%
    ggplot(aes(x=height, fill=poor_health), na.rm = TRUE) +
    geom_density(col=NA, alpha=0.4, adjust=3) +
    xlim(140,200)
```

```
## Warning: Removed 8906 rows containing non-finite values (stat_density).
```

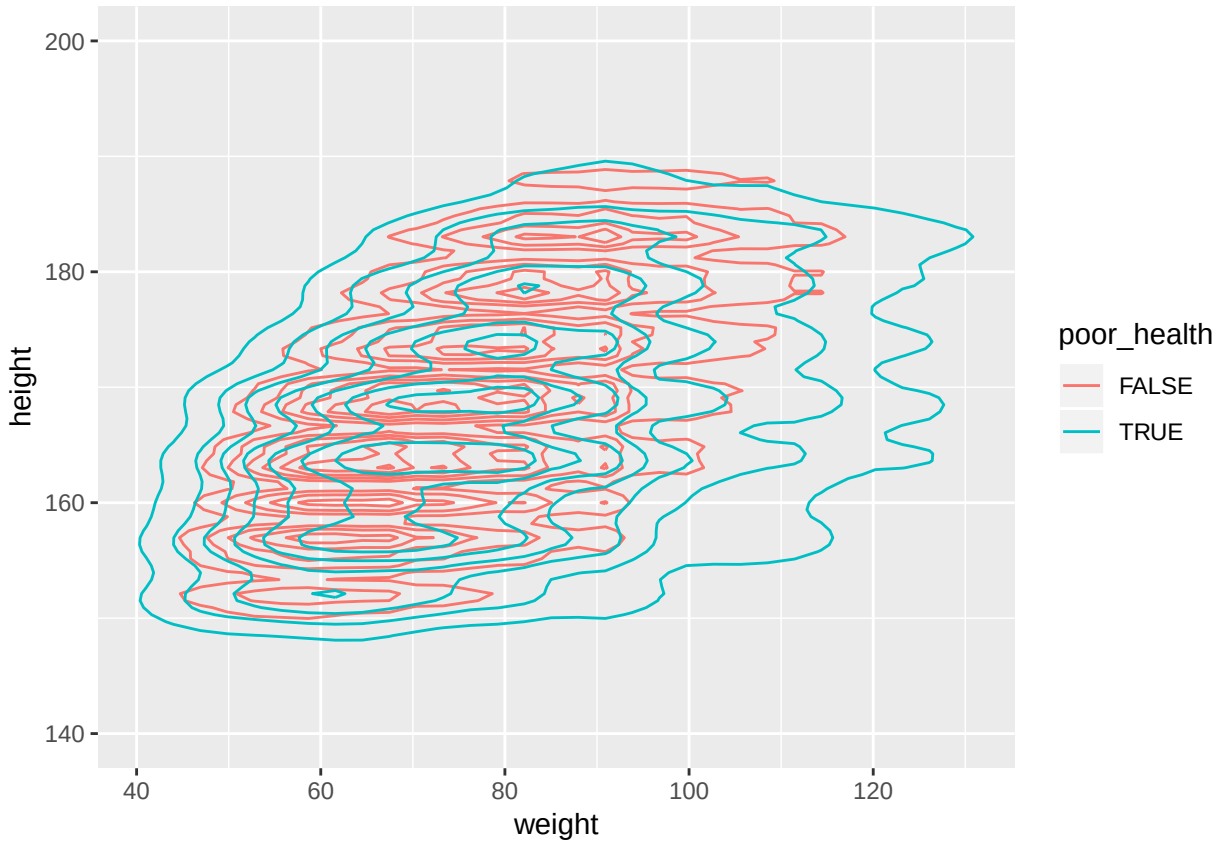



Height distribution of poor-health group has a longer tail at larger height region.

We can also see the 2D-density distribution of weight and height:

```
health_data %>%
  filter(!is.na(poor_health)) %>%
    ggplot(aes(x=weight, y=height, col=poor_health)) +
    geom_density2d() +
    ylim(140,200)
```

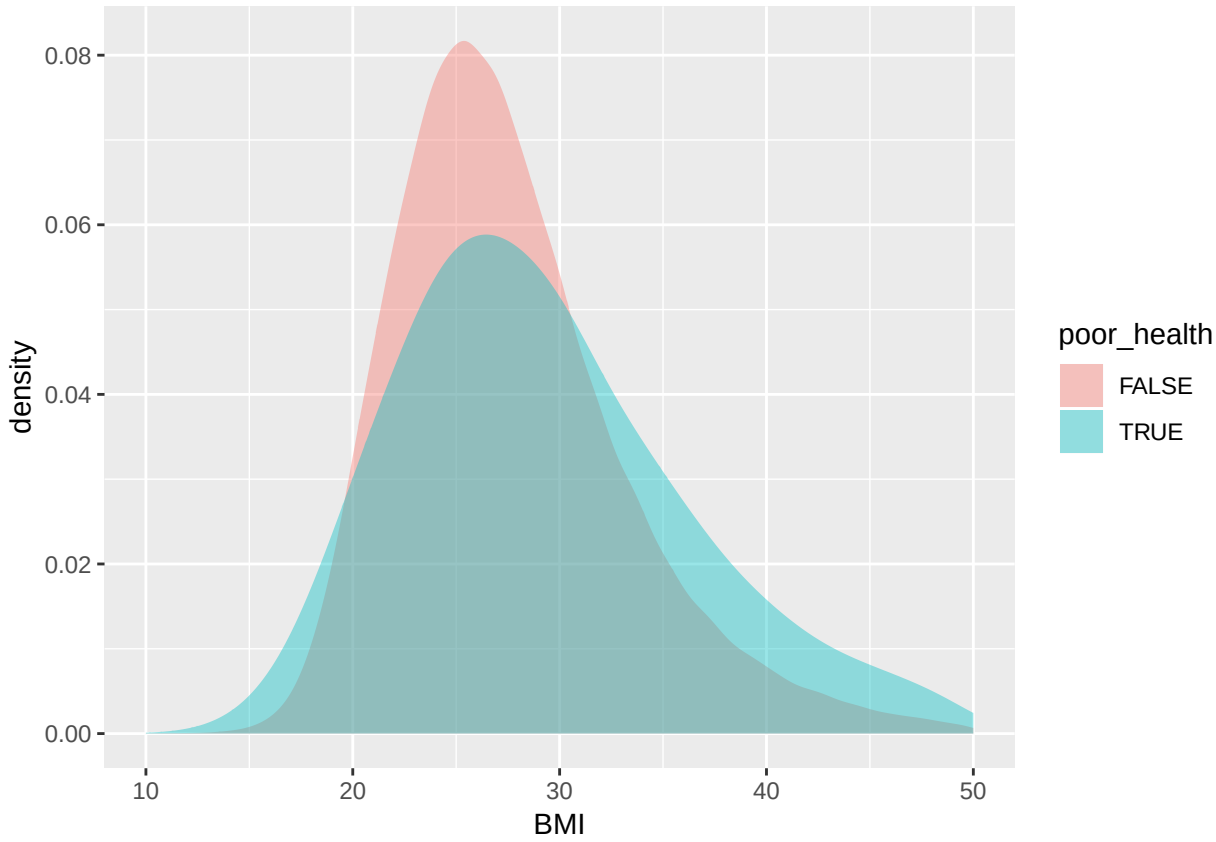
```
## Warning: Removed 24913 rows containing non-finite values (stat_density2d).
```




We see weight and height are correlated from the above graph. However it is not quite clear how weight and height would affect health together. We know that in general tall people are heavier than short people. So we want to take account of height when discussing whether people's weight are in a healthy range. Body mass index (BMI, https://en.wikipedia.org/wiki/Body_mass_index), defined as weight/height^2, categorize people as underweight, normal weight, overweight, or obese. We are now going to see the distribution of BMI within healthy and unhealthy groups.

```
health_data$BMI = health_data$weight / (health_data$height * health_data$height) * 10000

health_data %>%
  filter(!is.na(poor_health)) %>%
    ggplot(aes(x=BMI, fill=poor_health), na.rm = TRUE) +
    geom_density(col=NA, alpha=0.4, adjust=2) +
    xlim(10,50)
```

```
## Warning: Removed 26730 rows containing non-finite values (stat_density).
```




Clearly, BMI too large (overweight) or too small (underweight) may indicate poor health.

**Research quesion 3:**

There is a correlation between mental health and physical health, let's check the correlation coefficient between the number of days physical health is not good and the number of days mental health is not good.

```
#ggplot(data=health_data, aes(x=physhlth, y=menthlth), na.rm = TRUE) + geom_count() +
#   xlim(0,30) + ylim(0,30)
health_data$physhlth[health_data$physhlth > 30] = NA
health_data$menthlth[health_data$menthlth > 30] = NA
health_data$sleptim1[health_data$sleptim1 > 24] = NA
cor(health_data$physhlth, health_data$menthlth, use = "complete.obs")
```

```
## [1] 0.349241
```

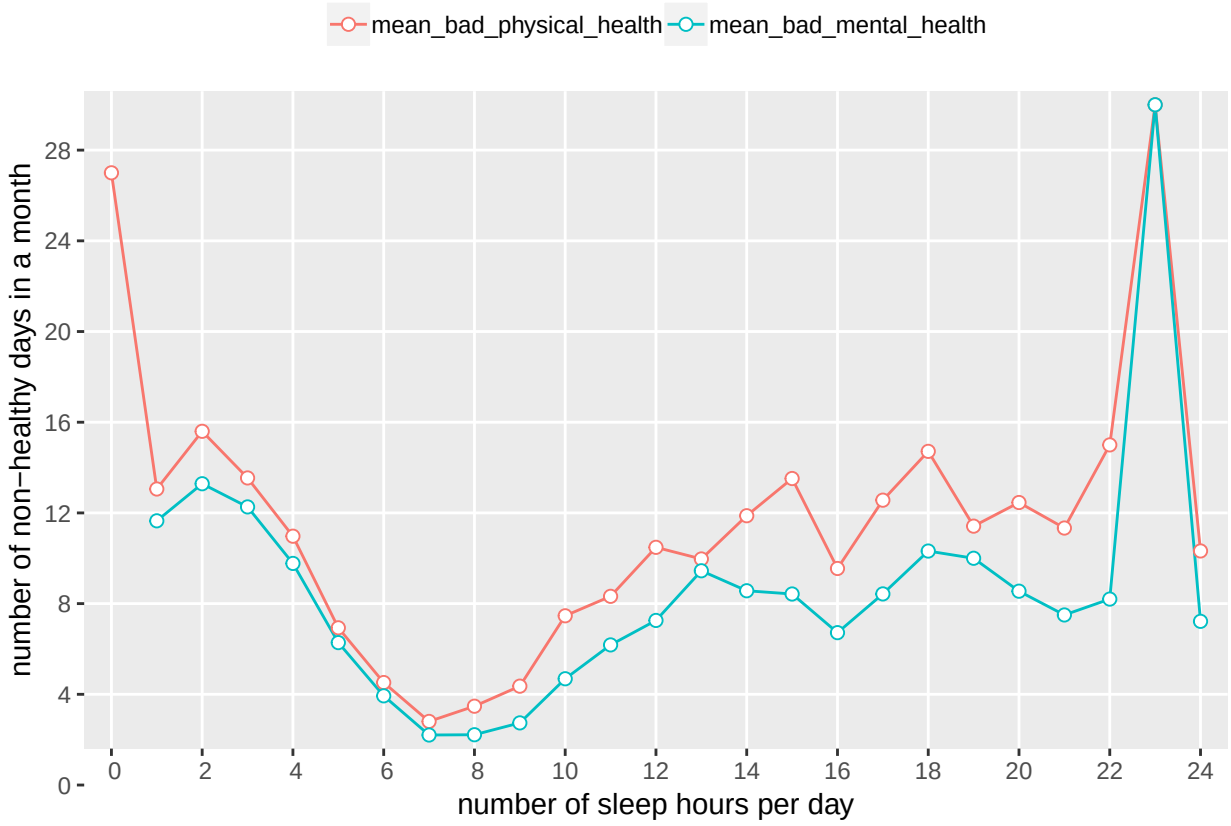
There is a positive correlation. Next, we want to show how sleep affect physical and mental health. We plot the mean of unhealthy days agains number of sleep hours.

```
library("reshape2")

sleep_health_table <- health_data %>% group_by(sleptim1) %>%
  summarise(mean_bad_physical_health = mean(physhlth, na.rm=TRUE),
            mean_bad_mental_health = mean(menthlth, na.rm=TRUE))

mdf <- melt(sleep_health_table, id.vars="sleptim1", value.name="n_days", variable.name="health")

ggplot(data=mdf, aes(x=sleptim1, y=n_days, group = health, colour = health)) +
  geom_line() +
  geom_point(size=2, shape=21, fill="white") +
```

```
  scale_x_discrete(limits=seq(0,24,2), name ="number of sleep hours per day") +
  scale_y_discrete(limits=seq(0,30,4), name ="number of non-healthy days in a month") +
  theme(legend.title = element_blank()) +
  theme(legend.position="top")
```

```
## Warning: Removed 3 rows containing missing values (geom_path).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```




We can see that people sleep for too long or too short are more likely to feel unhealthy, both physically and mentally.