Exploring the BRFSS data

Setup

Load packages

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

Load data

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

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

Part 1: Data

In this case, BRFSS has conducted both landline telephone- and cellular telephone-based surveys since 2011. In operating the BRFSS landline telephone survey, interviews collect data from a randomly selected adult in the household. In accessing the cellular telephone version of the BFSS questionnaire, interviewers collect data from an adult who participates by using a cellular telephone and resides in a private residence or college housing.

This sampling is generalized because interviews collect data in a landline telephone survey from a randomly selected adult in a household. Since we are unsure whether data collecting of the cellular telephone version is random, we can't determine its generalizability. And from the dataset, we can tell the assignments are not given haphazardly. Therefore, we can conclude that this case is not causal but generalizable.

Part 2: Research questions

Research quesion 1: The relationship between female and their health quality.

Research quesion 2: The relationship between people's employment status and their sleeping hours.

Research quesion 3: The relationship between Internet and people's exercise engagement.

Part 3: Exploratory data analysis

With the increasing number of women engaging in the work field, we might be interested in exploring their health status(both physically and mentally). Here we'd like to know the relationship between female and their health quality.

Research question 1: The relationship between female and their health quality.

Step 1

We need to identify whether the variable "sex" is a factor entry. Here we use the dplyr chain to select the variable of interest and investigate the structure by adding str() at the chain's end.

```
brfss2013 %>%
  select(sex) %>%
  str()
```

```
## 'data.frame':    491775 obs. of  1 variable:
##  $ sex: Factor w/ 2 levels "Male","Female": 2 2 2 2 1 2 2 2 1 2 ...
```

According to the result, the variable "sex" is a factor entry.

Step 2

We use filter() to exclude NAs in the variable "sex" if there are any blanks in the data. And to find out the numbers of female respondents.

```
brfss2013 %>%
  filter(!is.na(sex), sex !="Male" )%>%
  group_by(sex) %>%
  summarise(count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 1 x 2
##   sex     count
##   <fct>   <int>
## 1 Female 290455
```

According to the result, 290,455 female respondents are attending this survey.

Step 3

Create a new variable called "phyhealthy" when people whose days of physical health not being good are 0.

```
brfss2013 <- brfss2013 %>%
  filter(!is.na(physhlth)) %>%
  mutate(phyhealthy = physhlth == 0)
```

Step 4

Create a new variable called "menhealthy" when people whose days of mental health not being good are 0.

```
brfss2013 <- brfss2013 %>%
  filter(!is.na(physhlth)) %>%
  mutate(menhealthy = menthlth == 0)
```

Step 5

We consider that people whose days of physical health not being fair and mental health not being good are both equal to 0 are healthy. Create a new variable called"healthy"; otherwise, it's another new variable called "unhealthy."

```
brfss2013 <- brfss2013 %>%
  mutate(health_quality = ifelse(phyhealthy == menhealthy, "healthy", "unhealthy") )
```

We group by their general health quality and use filter() to exclude "Male" in sex variable and count their numbers.

```
brfss2013 %>%
  filter(!is.na(health_quality), sex !="Male") %>%
  group_by(health_quality, ) %>%
  summarise(count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

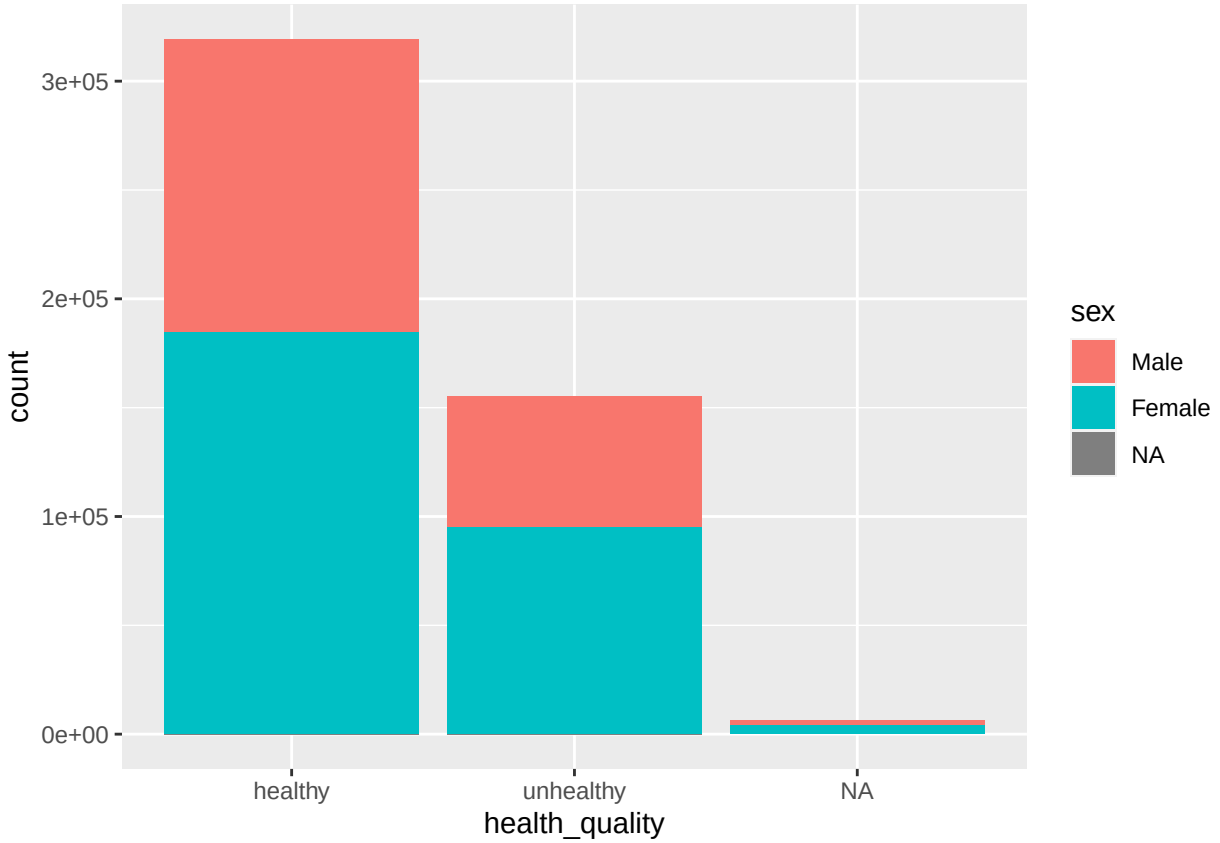
```
## # A tibble: 2 x 2
##   health_quality  count
##   <chr>           <int>
## 1 healthy        184554
## 2 unhealthy       94886
```

According to the result, we can conclude that more than 50% of female respondents' health quality is good (without either physical health or mental health problems), with 184,554 being healthy and 94,886 being unhealthy.

Step 6

Now we try to understand the distribution of the health quality between female respondents and male respondents in the barplot. Since the numbers of NA are small so they won't affect our results directly.

```
brfss2013 %>%
  filter(!is.na(sex)) %>%
ggplot(data = brfss2013, mapping = aes(x = health_quality, fill = sex)) +
  geom_bar()
```



All in all, we can reach a general conclusion that the health quality of more than 50% of female respondents is good (without suffering from either physical or mental health problems). Still, compared to male respondents, more of them are suffering from bad health quality problems.

With the development of technology, people are now facing unprecedented pressure from work. It might make us interested in exploring the relationship between employment status and people's sleeping hours under such a background.

Research question 2: The relationship between people's employment status and their sleeping hours.

Step 1

We can use summarize functions to have a general understanding of the respondents' sleeping status. Remember to exclude those NAs by using filter().

```
brfss2013 %>%
  filter(!is.na(sleptim1)) %>%
  summarise(slepmean = mean(sleptim1), slepmedian = median(sleptim1), slepsd = sd(sleptim1), minslep = m
```

```
##   slepmean slepmedian   slepsd minslep maxslep
## 1 7.050634          7 1.593205       0     450
```

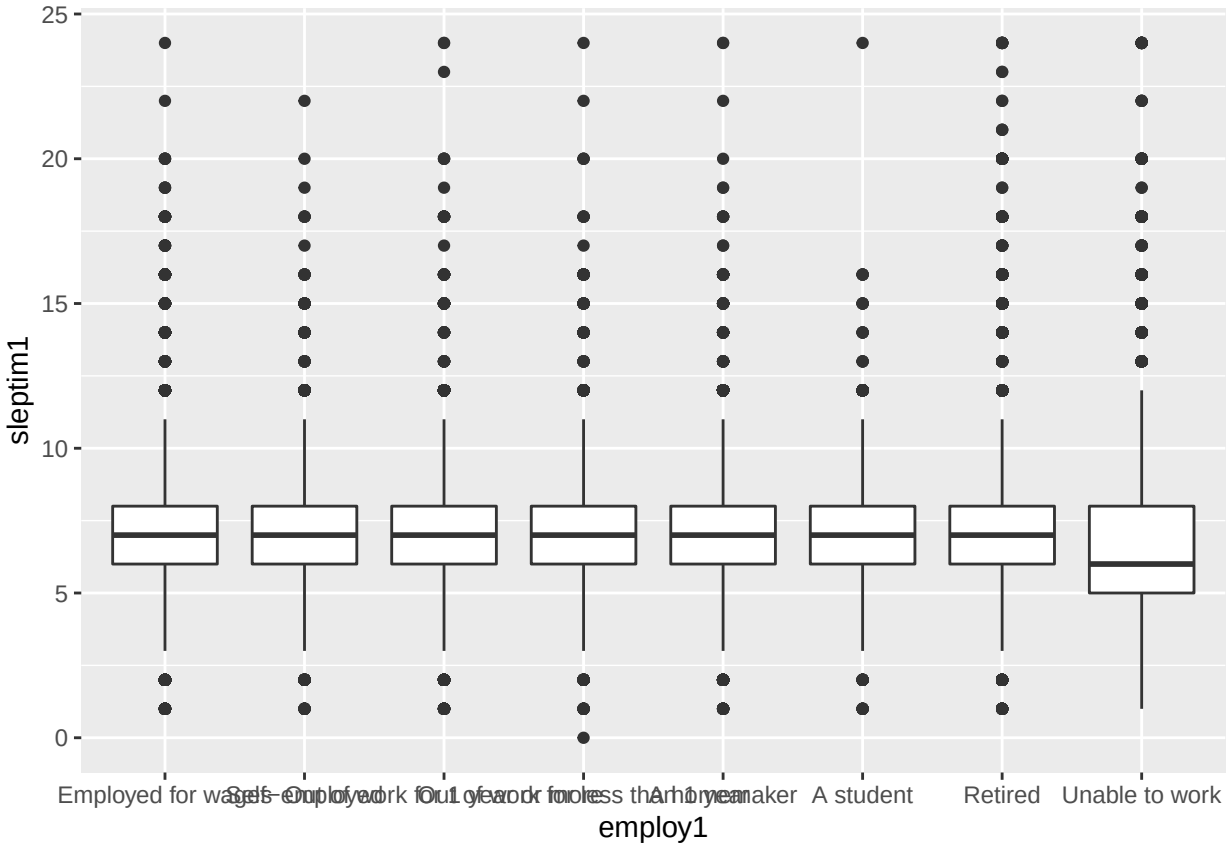
From the result, we can tell that the mean and median of sleeping hours are roughly equal, which means respondents' sleeping hours are symmetric.

Step 2

We use a boxplot to visualize the relationship between people's employment status and their sleeping hours. Also, we need to use filter() to exclude NAs in variable "employ1".

```
brfss2013 %>%
 filter(!is.na(employ1)) %>%
 ggplot(brfss2013, mapping = aes(x = employ1, y = sleptim1)) +
 geom_boxplot()
```

```
## Warning: Removed 6166 rows containing non-finite values (stat_boxplot).
```



From this boxplot box, we can see that the mean of nearly every emplotment staus is roughly the same except for people who are unable to work, which is less than the average level. Also, we know that the distributions are all left-skewed; among them, the sleeping hours of retired people distribute more normally. All in all, we can conclude that less than 50% of the data will be smaller than the mean, indicating less than 50% of respondents' sleeping hours are less than 7hrs.

It is well-known that more and more people indulge themselves on the Internet and less willing to do physical activities. Thus, we might be interested in exploring the relationship between the Internet and people's exercise engagement.

Research question 3: The relationship between Internet and people's exercise engagement.

Step 1

We group by the variable "internet" and variable "exerany2"(doing exercise in the past 30 days) to have a general understanding of the two variables. Also, we use filter() to exclude the NAs in the two variables.

```
brfss2013 %>%
  filter(!is.na(internet), !is.na(exerany2)) %>%
  group_by(internet, exerany2) %>%
  summarise(count = n())
```

```
## `summarise()` regrouping output by 'internet' (override with `.groups` argument)
```

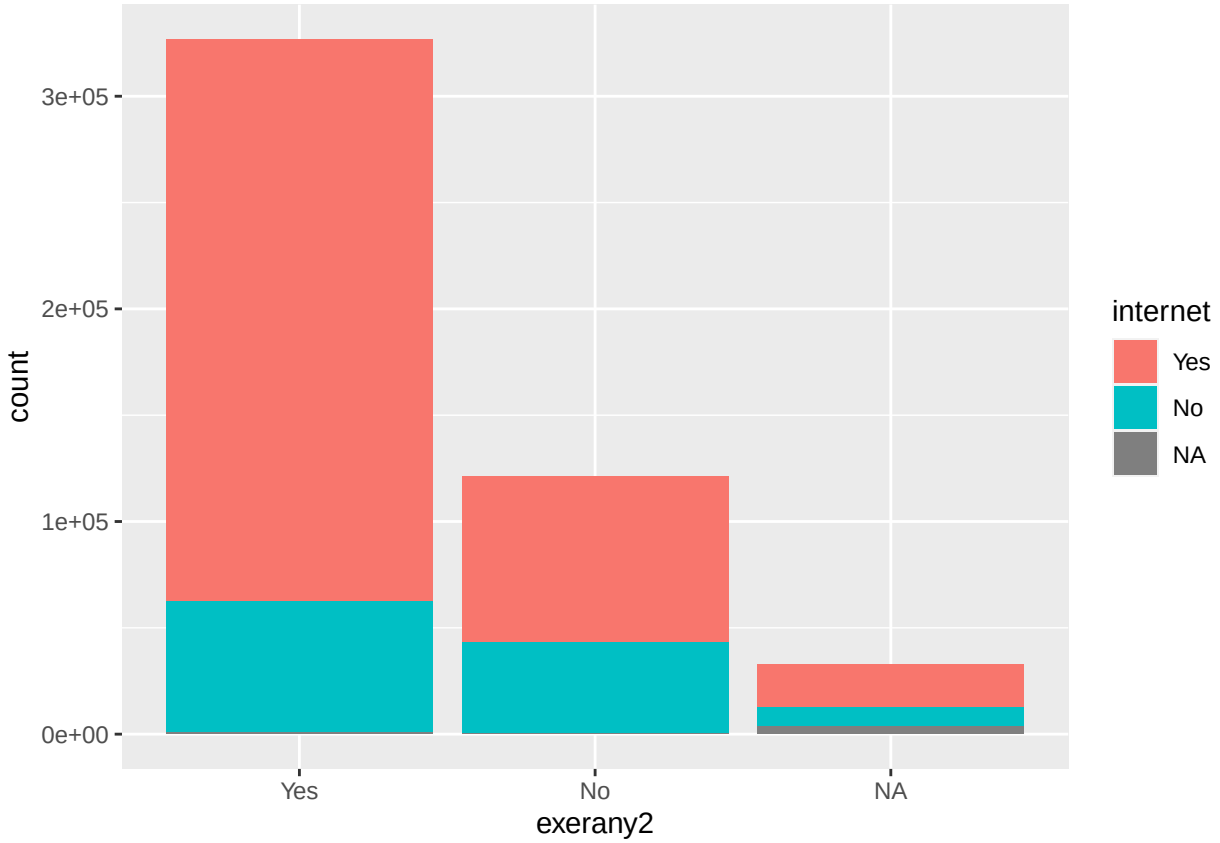
```
## # A tibble: 4 x 3
## # Groups:   internet [2]
##   internet exerany2  count
##   <fct>    <fct>     <int>
## 1 Yes      Yes      264240
## 2 Yes      No        78091
## 3 No       Yes       61811
## 4 No       No        42883
```

From the result, we can tell that people who used the Internet in the past days and did the exercise in the past 30 days are dominant.

Step 2

Here we use the bar plot again to help us visualize the distributions more clearly.

```
brfss2013%>%
  filter(!is.na(exerany2),!is.na(internet)) %>%
ggplot(data = brfss2013, mapping = aes(x = exerany2, fill = internet)) +
  geom_bar()
```



We can tell that most people can handle the balance between Internet using and physical exercise from this barplot. And people who haven't used the Internet or done any exercise account for the least distribution. Typically, based on the data, we can conclude that this result is generalizable that Internet using and exercise are not connected causally.