Exploring the BRFSS data

Setup

Load packages

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

Load data

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

Part 1: Data

Data for this analysis is taken from the Behavioral Risk Factor Surveillance System or BRFSS. The BRFSS is a US state based random dialed telephone survey of individuals aged >= 18 years who live in the United States.

The dataset hasn't been collected using simple random sampling. Instead, it was collected using a more complex sampling method. Based on the qualifiers above, it should be understood that any results of this analysis cannot be generalized to the entire US population. Only the population of individuals who are US residents, >= age 18, and have ownership or access to a home landline or cellular phone could have results of analysis generalized to it.

In order to ensure that results are accurate and generalizable one needs to be sure that very specific subgroups such as specific geographic area have a large enough sample size, one greater than 50 records according to BRFSS guidelines.

Part 2: Research questions

Research quesion 1: Question: Which state has the highest mean sleep time, and does this state have better or worse on mental and physical health days on average than all states and territories surveyed? This question seeks to see if there is a relationship between sleep, as well as physical and mental health. There are many studies showing poorer sleep can affect your health. I am curious to see if this result will be shown in this dataset.

Research quesion 2: Question: Of those who gave a yes or no response to the question, is the probability that a college educated individual in the survey will have a cell phone for personal use higher or lower than one who has a GED? I am curious to see if there is a relationship between college education and education level.

Research quesion 3: What is the probability that a randomly selected survey respondent in Massachusetts exercised in the last 30 days, is this greater or lower than all states, and what the most likely exercise this individual would be doing? Massachusetts has reputation as a healthy state, and I would like to see if survey respondents are more likely to have exercised here than across the US as a whole as well as what type of exercise they're likely to have done recently.

Part 3: Exploratory data analysis

NOTE: Insert code chunks as needed by clicking on the "Insert a new code chunk" button (green button with orange arrow) above. Make sure that your code is visible in the project you submit. Delete this note when before you submit your work.

Research quesion 1:

```
top_sleep <- brfss2013 %>%
        group_by(X_state) %>%
        filter(!is.na(sleptim1)) %>%
        summarise(mean_dd = mean(sleptim1), sd_dd = sd(sleptim1), n = n())

top_sleep %>%
        arrange(desc(mean_dd))
```

```
## # A tibble: 54 x 4
##         X_state    mean_dd    sd_dd     n
##          <fctr>      <dbl>    <dbl> <int>
##  1            0 450.000000       NA     1
##  2      Wyoming   7.233648 1.424893  6360
##  3 South Dakota   7.200877 1.340321  6840
##  4     Colorado   7.184885 1.299639 13457
##  5       Kansas   7.176604 1.354266 22995
##  6       Oregon   7.163948 1.362241  5886
##  7     Nebraska   7.152486 1.363457 16972
##  8         Iowa   7.149845 1.331144  8075
##  9      Montana   7.148310 1.433593  9588
## 10  Mississippi   7.144986 1.721203  7180
## # ... with 44 more rows
```

```
#Wyoming is the state with the highest mean hours of sleep: 7.23

top_physical <- brfss2013 %>%
        group_by(X_state) %>%
        filter(!is.na(physhlth)) %>%
        summarise(mean_pp = mean(physhlth))

top_physical %>%
        filter(X_state != 0) %>%
        summarise(mean(mean_pp))
```

```
## # A tibble: 1 x 1
##   `mean(mean_pp)`
##             <dbl>
## 1        4.359257
```

```
#The mean bad physical days across all states is 4.35

top_physical %>%
  filter(X_state == "Wyoming")
```

```
## # A tibble: 1 x 2
##   X_state  mean_pp
##    <fctr>    <dbl>
## 1 Wyoming 4.208208
```

```
#Wyoming has a mean of 4.2, therefore it is slightly lower than all states

top_mental <- brfss2013 %>%
        group_by(X_state) %>%
        filter(!is.na(menthlth)) %>%
        summarise(mean_mm = mean(menthlth))

top_mental %>%
        filter(X_state != 0) %>%
        summarise(mean(mean_mm))
```

```
## # A tibble: 1 x 1
##   `mean(mean_mm)`
##             <dbl>
## 1        3.392186
```

```
#The mean bad physical days across all states is 3.39

top_mental %>%
        filter(X_state == "Wyoming")
```

```
## # A tibble: 1 x 2
##   X_state  mean_mm
##    <fctr>    <dbl>
## 1 Wyoming 2.799717
```

```
#Wyoming has a mean of 2.79, which is lower than all states
#It appears that there may be a relationship between sleep and sick days as well as b
ad physical days
```

Research quesion 2:

```
#This code find the probability that a respondant will have completed college

college_educated <- brfss2013 %>%
  select(educa, cpdemo1) %>%
  filter(educa == "College 4 years or more (College graduate)")

college_cell <- college_educated %>%
  filter(cpdemo1 == "Yes") %>%
  select(cpdemo1)

college_no_cell <- college_educated %>%
  filter(cpdemo1 == "No") %>%
  select(cpdemo1)

college_cell_cnt <- count(college_cell)
college_no_cell_cnt <- count(college_no_cell)

college_cell_prob <- college_cell_cnt / (college_cell_cnt + college_no_cell_cnt)

#The probability that a college graduate has a cell phone is:
college_cell_prob
```

```
##           n
## 1 0.8836861
```

```
#This code find the probability that a respondant will have completed their GED

ged_educated <- brfss2013 %>%
  select(educa, cpdemo1) %>%
  filter(educa == "Grade 12 or GED (High school graduate)")

ged_cell <- ged_educated %>%
  filter(cpdemo1 == "Yes") %>%
  select(cpdemo1)

ged_no_cell <- ged_educated %>%
  filter(cpdemo1 == "No") %>%
  select(cpdemo1)

ged_cell_cnt <- count(ged_cell)
ged_no_cell_cnt <- count(ged_no_cell)

ged_cell_prob <- ged_cell_cnt / (ged_cell_cnt + ged_no_cell_cnt)

#The probability that a GED grad has a cell phone is:
ged_cell_prob
```

```
##           n
## 1 0.7161043
```

```
#Thus we can see that it is more likely that a GED grad is less likely to have a cell
phone for personal use than a college graduate.
ged_cell_prob < college_cell_prob
```

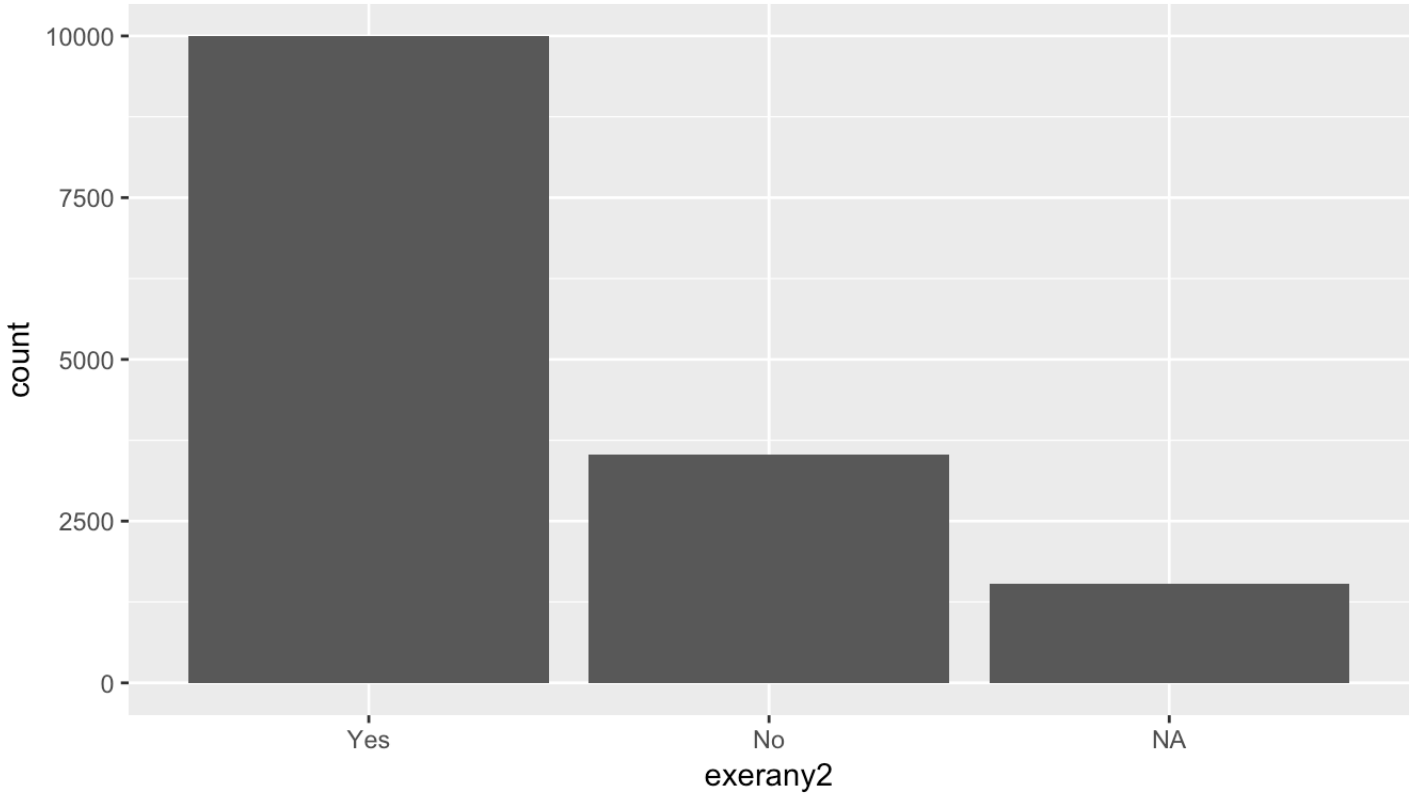
```
##         n
## [1,] TRUE
```

Research quesion 3:

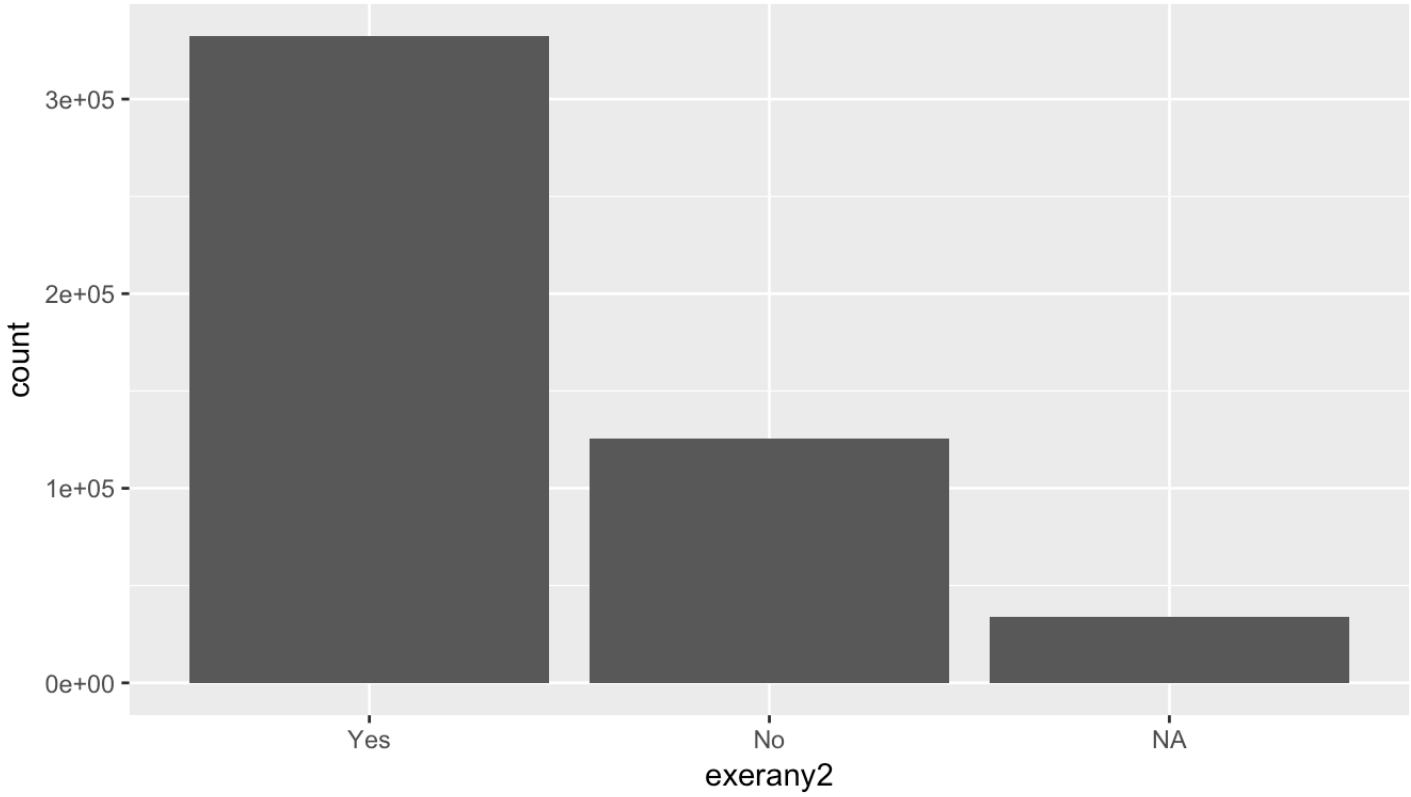
```
states_exer <- brfss2013 %>%
  group_by(X_state) %>%
  select(X_state, exerany2)

mass_exer <- brfss2013 %>%
    group_by(X_state) %>%
    select(X_state, exerany2) %>%
    filter(X_state == "Massachusetts")

yes_exer <- brfss2013 %>%
  group_by(X_state) %>%
  select(X_state, exerany2) %>%
  filter(X_state == "Massachusetts", exerany2 == "Yes")

a <- nrow(yes_exer)

no_exer <- brfss2013 %>%
  group_by(X_state) %>%
  select(X_state, exerany2) %>%
  filter(X_state == "Massachusetts", exerany2 == "No")

b <- nrow(no_exer)

na_exer <- brfss2013 %>%
  group_by(X_state) %>%
  select(X_state, exerany2) %>%
  filter(X_state == "Massachusetts", is.na(exerany2))

c <- nrow(na_exer)

total_value <- a + b + c

prob_ma_yes <- a / total_value

#the probability that a randomly selected survey respondent exercised in the last 30
days is 66%
prob_ma_yes
```

```
## [1] 0.663526
```

```
yes_exer_state <- brfss2013 %>%
  group_by(X_state) %>%
  select(X_state, exerany2) %>%
  filter(exerany2 == "Yes")

d <- nrow(yes_exer_state)

no_exer_state <- brfss2013 %>%
  group_by(X_state) %>%
  select(X_state, exerany2) %>%
  filter(exerany2 == "No")

e <- nrow(no_exer_state)

na_exer_state <- brfss2013 %>%
  group_by(X_state) %>%
  select(X_state, exerany2) %>%
  filter(is.na(exerany2))

f <- nrow(na_exer_state)

total_value_states <- d + e + f

prob_states_yes <- d / total_value_states


#we can see that a majority of survey respondents in MA exercised in the past 30 days
"
ggplot(data = mass_exer, aes(x= exerany2)) + geom_bar()
```




```
#across all states, it appears that the majority of individuals surveyed also exercis
ed in the last 30 days
ggplot(data = states_exer, aes(x= exerany2)) + geom_bar()
```




```
#The probability that a randomly selected survey respondant from Massachusetts exerci
sed in the last 30 days is:
prob_ma_yes
```

```
## [1] 0.663526
```

```
#The probability that a randomly selected survey respondant across all states exercis
ed in the last 30 days is:
prob_states_yes
```

```
## [1] 0.676049
```

```
exer_types <- brfss2013 %>%
  group_by(exract11) %>%
  filter(X_state == "Massachusetts", !is.na(exract11)) %>%
  select(exract11) %>%
  mutate(num = n())

exer_types_2 <- exer_types %>%
  group_by(exract11) %>%
  summarise(number = mean(num)) %>%
  arrange(desc(number))

exer_types_3 <- exer_types_2 %>%
  filter(number > 240)

#The most likely activity for an individual to have been doing is walking
exer_types_3
```

```
## # A tibble: 6 x 2
##                                          exract11 number
##                                            <fctr>  <dbl>
## 1                                          Walking   5543
## 2                                          Running    764
## 3                                            Other    448
## 4 Gardening (spading, weeding, digging, filling)    373
## 5                                     Weight lifting    353
## 6                       Bicycling machine exercise    246
```

```
#Overall, it does not appear that a Massachussets survey respondant is significantly
more likely to exercise than across all states
#The most likely activity for such a person to have done last is Walking
```