Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `brfss2013` . Delete this note when before you submit your work.

```
load("~/R/stats-r/brfss2013.RData")
```

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world, according to the CDC’s website on the BRFSS 2013 data.

Individual respondents are randomly selected from all adults, aged 18 years and older, living in a household and are interviewed in accordance with BRFSS protocol, according to the user guide for the BRFSS 2013 data.

It can be generalized to the U.S. population because the participants are selected randomly and it is across all 50 states, 3 territories, and the District of Columbia.

Part 2: Research questions

Research quesion 1:

```
#Are the states, territories, or districts (X_state) with the highest amounts of tetanus shots
          (tetanus) rhe same states with the highest amount oof pneuomnia (pneuvac3) and flu (fl
         ushot6) shots?
```

It is important to consider the relationship between the state one is living in and the rates for tetanus, pneumonia, and flu shots because we can analyze the policies in the states with the highest rates for each of these vaccines to apply to other states to increase vaccination rates.

Research quesion 2:

```
#Are rates of depression (addepev2), stroke (cvdstrk3), and/or coronary heart disease (cvdcrhd4)
         correlated with veteran status (veteran3)?
```

It is important to consider the relationship between veteran status and rates of depressino, stroke, and coronoary heart disease because we can understand how to assist returning veterans with their health issues.

Research quesion 3:

```
#How is the amount of exercise one has had in the past 30 days (exerany2) correlated with the in
         take of fruits (fruits) and vegetables (vegetables)?

#*fruits was created by adding fruitju1, fruit1 rowwise for each observation
#**vegetables was created by adding fvbeans, fvgreen, fvorang, vegetab1 rowwise for each observa
         tion
```

It is important to consider the relationship between the amount one consumes fruits and vegetables and the amount one exercises in the past 30 days because we can learn more about the relationship between healthy eating and exercise.

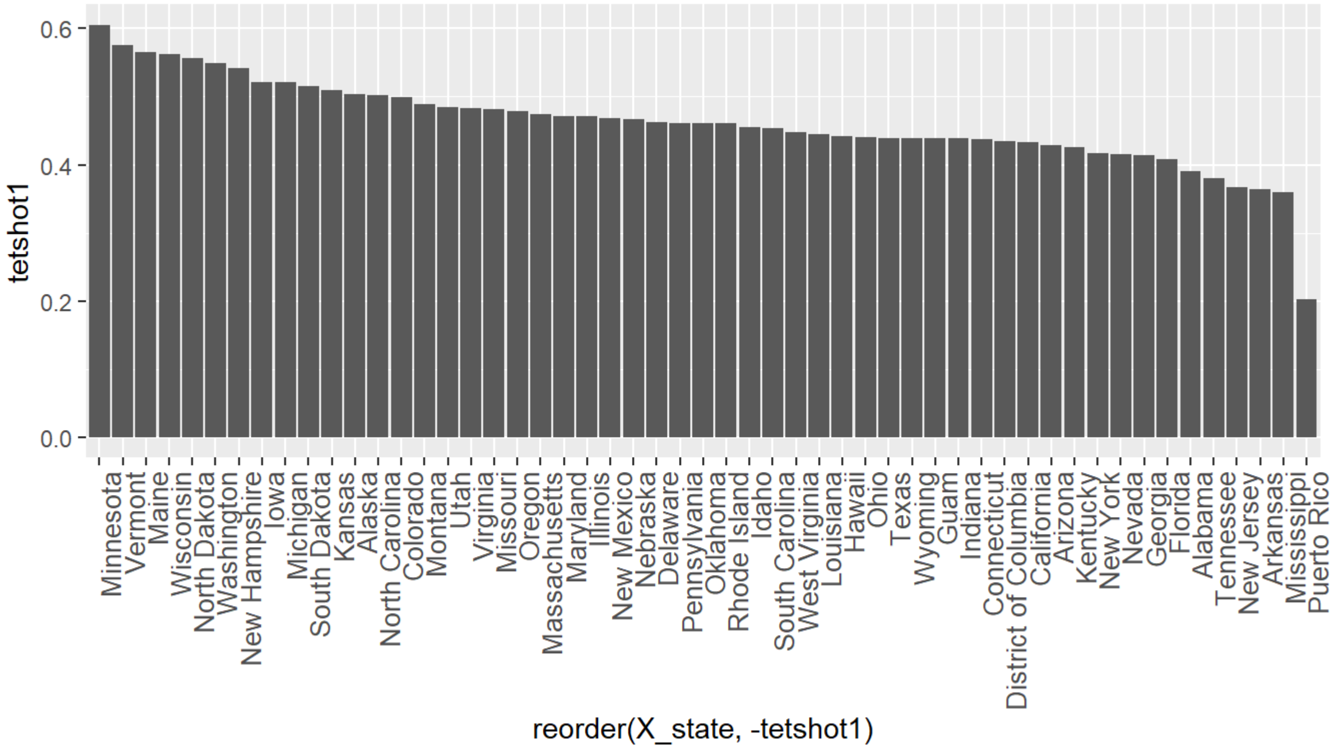
Part 3: Exploratory data analysis

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:

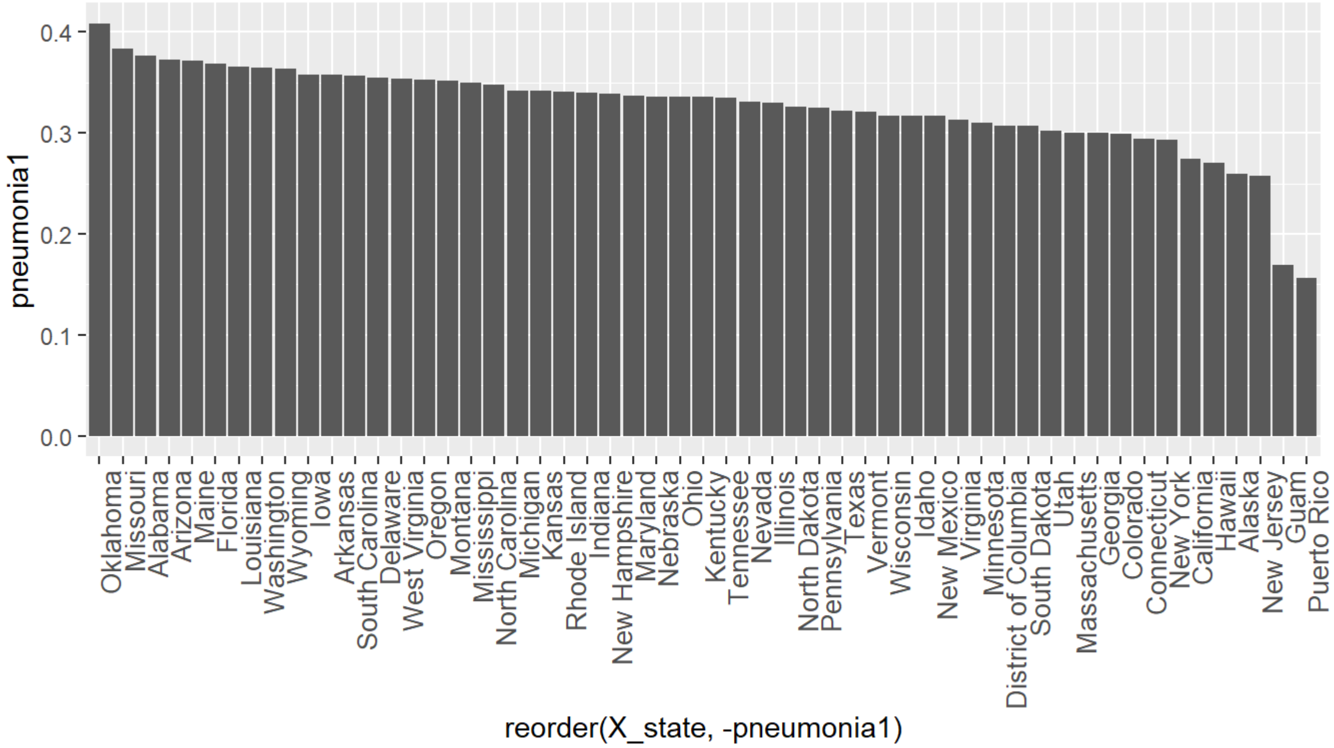
```
brfss2013 <- brfss2013 %>% mutate(tetshot = grepl("Yes", tetanus))
tetanus_by_state <- brfss2013 %>% group_by(X_state) %>% summarize(tetshot1 = mean(tetshot)) %>%
         filter(X_state != 0, X_state != 80) %>% droplevels()

#Minnesota has the highest amount of tetanus shots in America
ggplot(tetanus_by_state, aes(x=reorder(X_state, -tetshot1), y=tetshot1)) + geom_bar(stat="identi
         ty") + theme(axis.text.x = element_text(size = 10, angle = 90, hjust=1))
```



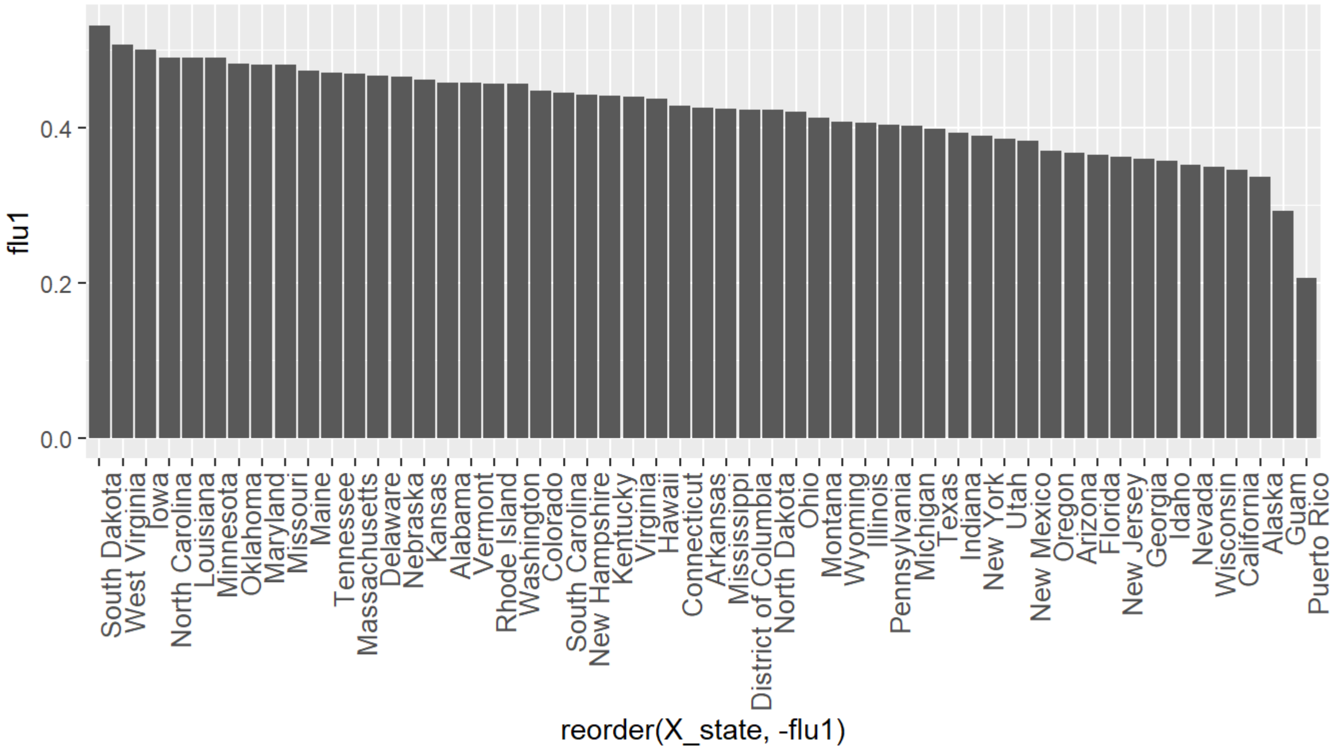
```
brfss2013 <- brfss2013 %>% mutate(pneumonia = grepl("Yes", pneuvac3))
pneumonia_by_state <- brfss2013 %>% group_by(X_state) %>% summarize(pneumonia1 = mean(pneumoni
        a)) %>% filter(X_state != 0, X_state != 80)

#Oklahoma has the highest amount of pneumonia shots in America
ggplot(pneumonia_by_state, aes(x=reorder(X_state, -pneumonia1), y=pneumonia1)) + geom_bar(stat=
        "identity") + theme(axis.text.x = element_text(size = 10, angle = 90, hjust=1))
```



```
brfss2013 <- brfss2013 %>% mutate(flu = grepl("Yes", flushot6))
flu_by_state <- brfss2013 %>% group_by(X_state) %>% summarize(flu1 = mean(flu)) %>% filter(X_state != 0, X_state != 80)

#South Dakota has the highest amount of flu shots in America
ggplot(flu_by_state, aes(x=reorder(X_state, -flu1), y=flu1)) + geom_bar(stat="identity") + theme(axis.text.x = element_text(size = 10, angle = 90, hjust=1))
```



I am answering this research question by collecting the states with the top ten rates for tetanus, pneumonia, and flu shots, which are my summary statistics.

These were the ten states with the highest rates of tetanus shots: Minnesota, Vermont, Maine, Wisconsin, North Dakota, Washington, New Hampshire, Iowa, Michigan, South Dakota. Of these states, three appeared in the top ten for pneumonia shots (Maine, Washington, Iowa), and four appeared in the top ten for flu shots (South Dakota, Iowa, Minnesota, Maine). Therefore, only Maine and Iowa are the states that have the highest amount of tetanus, pneumonia, and flu shots, which indicates that the majority of the states with the highest amount of tetanus shots are not the same states with the highest amount of pneumonia and flu shots.

Research quesion 2:

```
vet_brfss <- brfss2013 %>% filter(veteran3 == "Yes") %>% mutate(depression = grepl("Yes", addepev2))
nonvet_brfss <- brfss2013 %>% filter(veteran3 == "No") %>% mutate(depression = grepl("Yes", addepev2))
vet_dep <- mean(vet_brfss$depression)
nonvet_dep <- mean(nonvet_brfss$depression)
t.test(vet_brfss$depression, nonvet_brfss$depression)
```

```
##
##  Welch Two Sample t-test
##
## data:  vet_brfss$depression and nonvet_brfss$depression
## t = -23.423, df = 83538, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04077387 -0.03447712
## sample estimates:
## mean of x mean of y
## 0.1619145 0.1995400
```

The mean of depression among non-veterans is ~0.199 and the mean of depression among veterans is ~0.162, so depression is higher among non-veterans. The p-value is 2.2e-16.

```
vet_brfss <- brfss2013 %>% filter(veteran3 == "Yes") %>% mutate(stroke = grepl("Yes", cvdstrk3))
nonvet_brfss <- brfss2013 %>% filter(veteran3 == "No") %>% mutate(stroke = grepl("Yes", cvdstrk
         3))
Vet_stroke <- mean(vet_brfss$stroke)
nonvet_stroke <- mean(nonvet_brfss$stroke)
t.test(vet_brfss$stroke, nonvet_brfss$stroke)
```

```
##
##  Welch Two Sample t-test
##
## data:  vet_brfss$stroke and nonvet_brfss$stroke
## t = 24.844, df = 72669, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02354021 0.02757260
## sample estimates:
## mean of x mean of y
## 0.0638284 0.0382720
```

The mean of stroke among non-veterans is ~0.04 and the mean of depression among veterans is ~0.06, so stroke is higher among veterans. The p-value is 2.2e-16.

```
vet_brfss <- brfss2013 %>% filter(veteran3 == "Yes") %>% mutate(coronary = grepl("Yes", cvdcrhd
         4))
nonvet_brfss <- brfss2013 %>% filter(veteran3 == "No") %>% mutate(coronary = grepl("Yes", cvdcrh
         d4))
vet_coronary <- mean(vet_brfss$coronary)
nonvet_coronary <- mean(nonvet_brfss$coronary)
t.test(vet_brfss$coronary, nonvet_brfss$coronary)
```

```
##
##  Welch Two Sample t-test
##
## data:  vet_brfss$coronary and nonvet_brfss$coronary
## t = 51.803, df = 69585, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06754465 0.07285679
## sample estimates:
##  mean of x  mean of y
## 0.12052859 0.05032788
```

The mean of coronary heart disease among non-veterans is ~0.121 and the mean of depression among veterans is ~0.050, so coronary heart disease is higher among non-veterans. The p-value is 2.2e-16.

I am answering this research question by finding the mean rates of depression, pneumonia, and coronary heart disease among veterans and non-veterans and the p-values for these computations, which are my summary statistics.

By the above findings, the rates for depression and coronary heart disease were significantly higher among non-veterans, and the rates for stroke were signficantly higher among veterans. Therefore, we can conclude that that veteran status is associated with different diseases. Long-term diseases are more common among the non-veteran population while short-term medical conditions are more common among veterans. Further research is required to either prove or discredit this finding.

Research quesion 3:

```
brfss2013 <- brfss2013 %>% rowwise() %>% mutate(vegetables = sum(fvbeans, fvgreen, fvorang, vege
         tab1, na.rm = TRUE))
brfss2013 <- brfss2013 %>% rowwise() %>% mutate(fruit = sum(fruitju1, fruit1, na.rm = TRUE))

df <- select(brfss2013, fruit, vegetables, exerany2)
df$exerany2 <- as.numeric(df$exerany2)

#The mean of fruits eaten across everyone is 288.8803 and the mean of vegetables eaten across ev
         eryone is 769.2092. The difference is 480.3289. The p-value is 2.2e-16.
t.test(df$fruit, df$vegetables)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$fruit and df$vegetables
## t = -881.69, df = 753970, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -481.3967 -479.2612
## sample estimates:
## mean of x mean of y
##  288.8803  769.2092
```

```
#The mean of fruits eaten across people who exercise is 309.3814 and the mean of vegetables eate
         n across people is 829.6003. The difference is 520.2189 The p-value is 2.2e-16. Exercis
         e slightly increases the difference between eating fruits and vegetables.
t.test(df$fruit[df$exerany2 == 1], df$vegetables[df$exerany2 == 1])
```

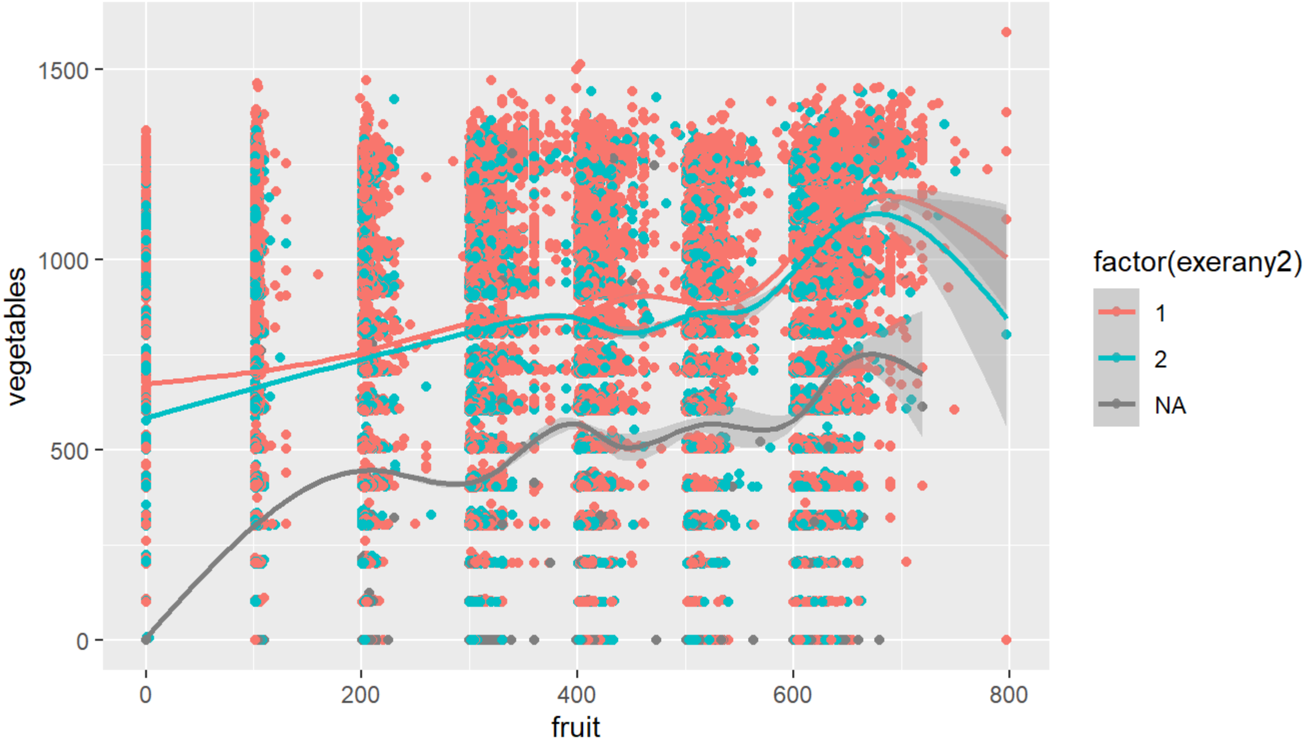
```
## 
##  Welch Two Sample t-test
## 
## data:  df$fruit[df$exerany2 == 1] and df$vegetables[df$exerany2 == 1]
## t = -922.26, df = 552632, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -521.3245 -519.1134
## sample estimates:
## mean of x mean of y 
##  309.3814  829.6003
```

```
#The mean of fruits eaten across people who exercise is 296.9171 and the mean of vegetables eate
         n across people is 791.9892. The difference is 495.0721. The p-value is 2.2e-16.
#Limiting the data to those who do not exercise increases the difference between eating fruits a
         nd vegetables, but not as much as with those who do exercise.
t.test(df$fruit[df$exerany2==2], df$vegetables[df$exerany2==2])
```

```
## 
##  Welch Two Sample t-test
## 
## data:  df$fruit[df$exerany2 == 2] and df$vegetables[df$exerany2 == 2]
## t = -505.66, df = 202011, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -496.9910 -493.1531
## sample estimates:
## mean of x mean of y 
##  296.9171  791.9892
```

```
#The trends of eating fruits and vegetables for those who exercise and don't exercise take on ap
         proximately the same shape, but the trend for those who don't exercise is slightly lowe
         r.
#Those with NA values ate much fewer vegetables for every fruit.
#My speculation is that they do not exercise very frequently because people might be embarrassed
         to admit they don't exercise and therefore left the question blank, an example of socia
         l desirability bias
ggplot(df, aes(fruit, vegetables, color = factor(exerany2))) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



I am answering this research question by performing t-tests on the fruits and vegetables eaten among the general population, people who exercise, and people who don’t exercise, which provide my summary statistics.

Exercise is found to slightly increase the difference between people eating fruits and vegetables compared to the general population. Not exercising yields the same result though to a lesser extent than exercising. Therefore, we can conclude that exercising does correlate with the amount of fruits and vegetables eaten than not exercising at all. Further research is needed on this topic to either prove or disprove this conclusion.