<!DOCTYPE html>

—

title: “Exploring the BRFSS data” output: html_document: fig_height: 4 highlight: pygments theme: spacelab —

Setup

Load packages

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

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("brfss2013.RData")
```

Part 1: Data

This is an observational study and not an experimental study because other variables are not controlled for. Random assignment is missing. This means that we can study only correlation and not causation. The data seems to be fairly generalizable (as a random population is included) to the entire population as the sample size is large and diverse in terms of demographics. However since home phone numbers were reached out

to for the survey, this excludes people who do not have such a facility or might be busy at that time of the day to attend to such a call.

Part 2: Research questions

Research quesion 1: Is there a correlation between individual's weight and arthiritis diagnosis? More precisely, is there a trend in the ratio of people diagnosed positively for arthiritis to negative diagnosis as the person's weight category increases?

Usually people with higher weight are considered to be a higher risk of arthiritis as there is more weight on the knee joints and greater wear and tear.

Research quesion 2: Is there a correlation between bad mental health and number of hours per week for people with mental health issues? Is there an optimum level of weekly work hours at which bad mental health could be minimised? And does financial security also play a role?

Usually people with part time employment have a lot of time to them which could result in spending more time in anxious thoughts. Also people working extremely high hours burn out and might be in poor mental health as well. I'm interested in seeing if there is an optimum level and how mental health varies with people with and without financial security.

Research quesion 3: Is there a correlation between bad mental health and loneliness for people with mental health issues? Is this any different for one gender when compared to another?

Doctors usually ask people with poor mental health to not be alone even if that is what they want to do. Is this effective? How does the number of people around you affect it? Is it varying for different genders with the same number of people? Let's see!

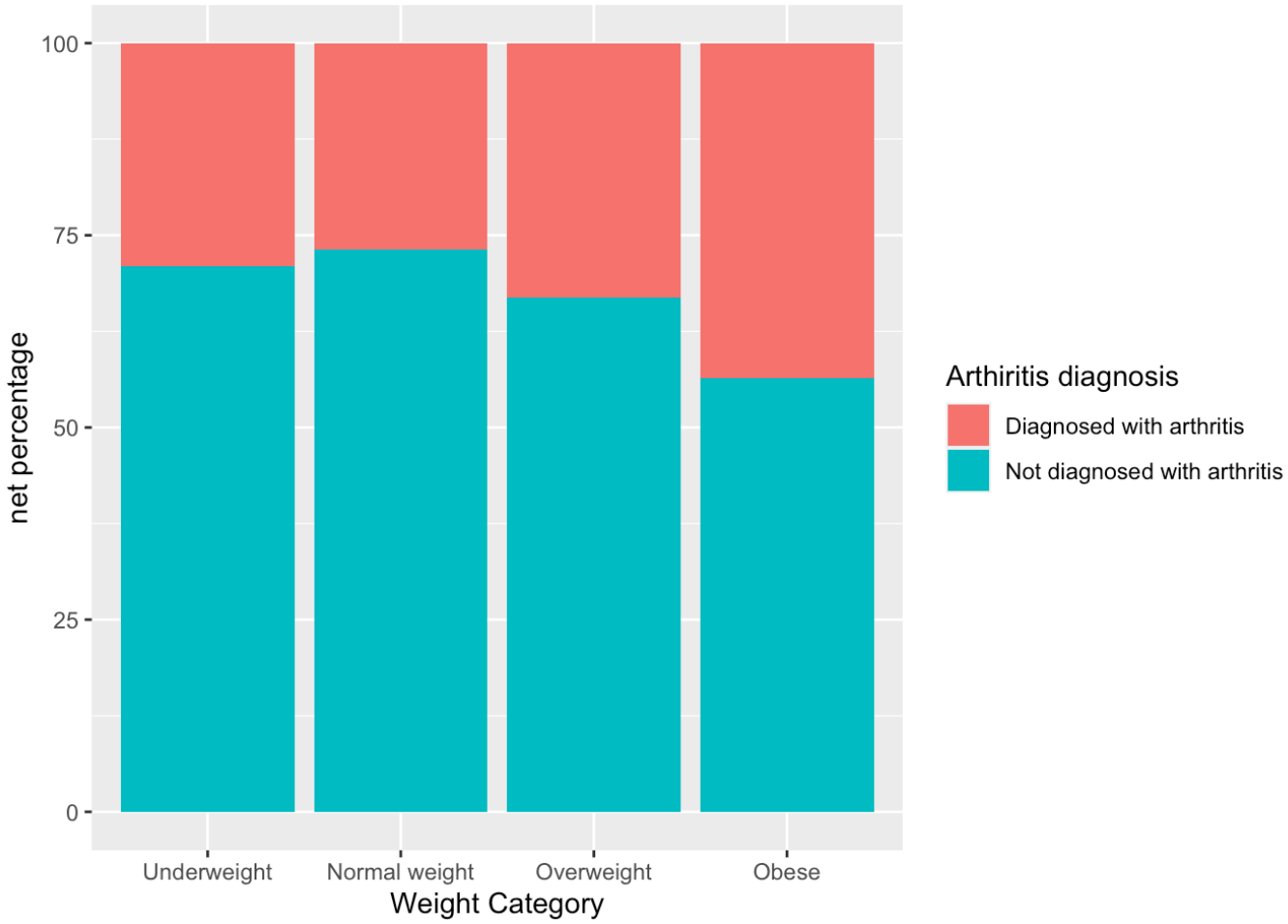
Part 3: Exploratory data analysis

Research quesion 1:

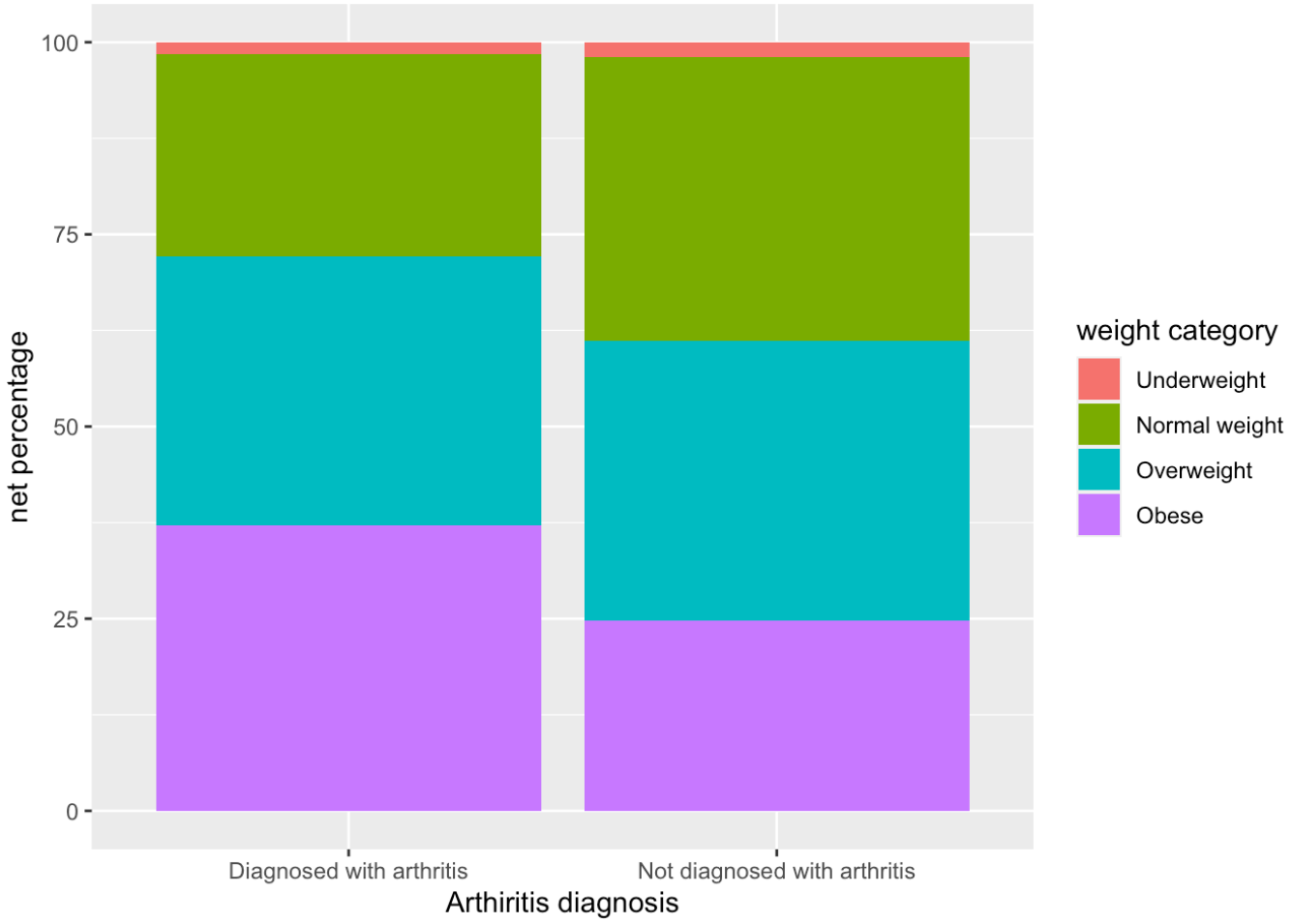
```
library(ggplot2)
library(dplyr)

##Summarising data by groups.2 cases exist for arthiritis diagnosis(X_drdxar1)
and 4 for weight categories(X_bmi5cat). A total of (2*4=8) groups exist.

q1dataset <- brfss2013 %>% filter(!is.na(X_bmi5cat),!is.na(X_drdxar1)) %>% grou
p_by(X_drdxar1,X_bmi5cat) %>% summarise(no_of_people=n())

##Obtaining the percentage of sub-category within the larger category. First ("
percentage") within the specific arthiritic condition adding the four different
weight categories. Second ("percentage2") within the weight category, we find t
he two ratios of arthiritic to non arthiritic diagnosis.

q1dataset$percentage = signif(100*with(q1dataset, ave(no_of_people, X_drdxar1,
FUN=function(x) x/sum(x))),4)

q1dataset$percentage2 = signif(100*with(q1dataset, ave(no_of_people, X_bmi5cat
, FUN=function(x) x/sum(x))),4)

## Now we have refined the summary statistics further to clearly show us percen
tage wise breakdowns in each category.

print(q1dataset)
```

```
## # A tibble: 8 x 5
## # Groups:   X_drdxar1 [2]
##   X_drdxar1                    X_bmi5cat     no_of_people percentage percent
age2
##   <fct>                        <fct>                <int>      <dbl>       <
dbl>
## 1 Diagnosed with arthritis     Underweight           2376       1.51
29.0
## 2 Diagnosed with arthritis     Normal weight        41356      26.3
26.8
## 3 Diagnosed with arthritis     Overweight           54919      35.0
33.0
## 4 Diagnosed with arthritis     Obese                58459      37.2
43.6
## 5 Not diagnosed with arthritis Underweight           5828       1.91
71.0
## 6 Not diagnosed with arthritis Normal weight       112689      36.9
73.2
## 7 Not diagnosed with arthritis Overweight          111239      36.4
67.0
## 8 Not diagnosed with arthritis Obese                75591      24.8
56.4
```

```
## Next, Plotting the graph with these ratios obtained above. Ratios would give us a better picture as compared to absolute numbers which differ significantly for each weight category and for different arthiritis diagnosis. So differences would be better viewed with ratios in this case.

df_base <- ggplot(data = q1dataset, aes(x=X_bmi5cat ,y=percentage2, fill=X_drdxar1))
df_base + geom_bar (stat = "identity") + labs(fill="Arthiritis diagnosis") + xlab("Weight Category") + ylab("net percentage")
```




```
df_base2 <- ggplot(data = q1dataset, aes(x=X_drdxar1 ,y=percentage, fill=X_bmi5cat))
df_base2 + geom_bar (stat = "identity") + xlab("Arthiritis diagnosis") + labs(fill="weight category") + ylab("net percentage")
```



As we can see in the first graph, a larger proportion of people are diagnose d with arthiritis as we move towards increasing weight categories. The second g raph shows that among people not diagnosed with arthiritis, a larger percentage are normal weight as compared to positively diagnosed ones. Positively diagnose d category has larger percentage of people in the overweight and obese categori es as compared to the undiagnosed category. Also the under weight category seem s to defy this, indicating that while risk of arthiritis increases above optimu m weight, it does not decrease if below optimum weight.Wow!

Research quesion 2:

```
library(ggplot2)
library(dplyr)

## selecting columns required for the analysis ("menthlth"-which indicates no o f days in a month with bad mental health. "scntwrk1","scntlwk1"- both indicate weekly hours of work and are combined to form another variable "hrs_worked" whi ch will be later converted to categorical data "hrs_work_ctgry" for analysis in box plot. Finally "scntmeal","scntmony" which refer to -Times Past 12 Months Wo rried/Stressed About Having Enough Money To Pay Your Nutrition and Rent- respec tively will be used to determine a new categorical variable "money_issue" indic ating whether is financially troubled. So two categorical variables that have b een derived will be used against the one numerical variable.
```

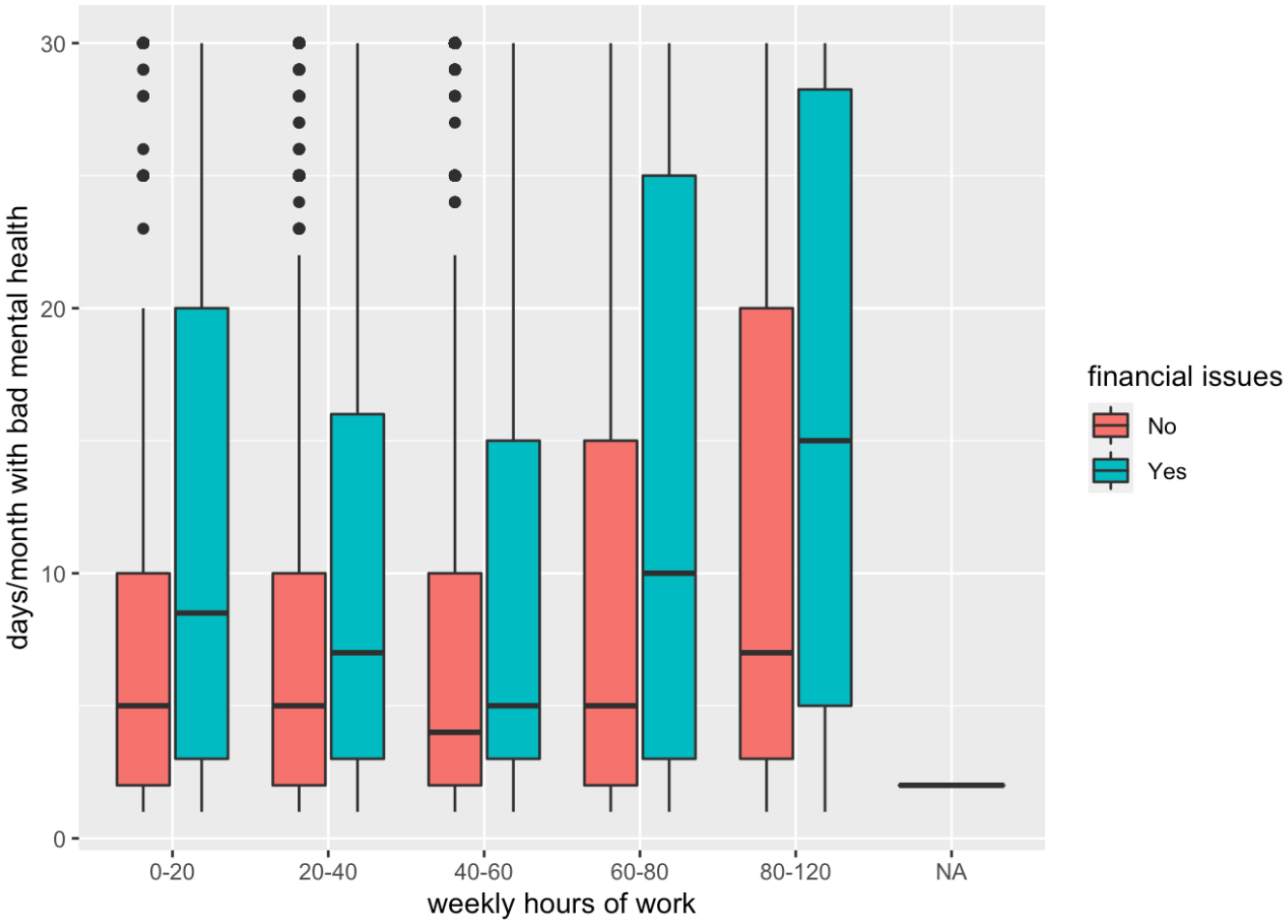
```
q2dataset <- brfss2013 %>% select (menthlth,scntwrk1,scntlwk1,scntmeal,scntmony
)

##Next, removing cases with no data on weekly hours worked,days/month with bad
mental health AND WITH 0 DAYS with bad mental health

q2dataset<- q2dataset %>% filter(!is.na(scntwrk1)|!is.na(scntlwk1),!is.na(menth
lth),menthlth>0)

##Next, creating a new column with weekly hours worked and removing NA values f
rom it.

q2dataset<- q2dataset %>% mutate(hrs_worked=ifelse(!is.na(scntwrk1),scntwrk1,sc
ntlwk1))

##removing cases with no data on financial health

q2dataset<- q2dataset %>% filter(!is.na(scntmeal)|!is.na(scntmony))

## Obtaining a new parameter "money_issue" indicating financial health by combi
ning two given parameters. The two parameters are refined to first remove "NA"
and  the data recorded in two new parameters. These two parameters are then com
bined to yield the new financial health parameter of "money_issue"

q2dataset<- q2dataset %>% mutate(scntmeal_new=ifelse(!is.na(scntmeal),scntmeal,
scntmony))
q2dataset<- q2dataset %>% mutate(scntmony_new=ifelse(!is.na(scntmony),scntmony,
scntmeal))
q2dataset<- q2dataset %>% mutate(money_issue=ifelse(scntmeal_new==5|scntmony_ne
w==5,"No","Yes"))

## Lets find the average and median number of days with bad mental health for o
ur two categories of people in "money_issue" variable. Also  let's find the ave
rage weekly hours worked for each category.

q2dataset %>% filter(money_issue=="No") %>% summarise(avg_days_rich=mean(menthl
th),median_days_rich=median(menthlth),avg_weekly_work_hours_rich=mean(hrs_worke
d))
```

```
##   avg_days_rich median_days_rich avg_weekly_work_hours_rich
## 1      8.463412                5                   42.62282
```

```
q2dataset %>% filter(money_issue=="Yes") %>% summarise(avg_days_poor=mean(menth
lth),median_days_poor=median(menthlth),avg_weekly_work_hours_poor=mean(hrs_work
ed))
```

```
##   avg_days_poor median_days_poor avg_weekly_work_hours_poor
## 1      11.29292                7                   40.97206
```

```
## It can be seen that while average weekly work hours are slightly less for pe
ople with money issues, their average number of days with poor mental health is
noticably higher.

## In the results above we can see that average number of days with bad mental
health is higher for people with money issues.

## Next, converting numeric data of "hrs_worked" to categorical data of "hrs_wo
rk_ctgry" for analysis in box plot

q2dataset<-q2dataset %>% mutate(hrs_work_ctgry=cut(hrs_worked,breaks = c(0,20,4
0,60,80,120),labels = c("0-20","20-40","40-60","60-80","80-120" )))

## analysis of data using box plot

plot1 <- ggplot(data=q2dataset,aes(x=hrs_work_ctgry,y=menthlth,fill=money_issue
))
plot1+geom_boxplot()+ xlab("weekly hours of work") + ylab("days/month with bad
mental health") +labs(fill="financial issues")
```



##Amazing!! We can see a U shaped graph of the median values of box plots. This indicates that number of bad mental health days reaches a minimum for people working 40-60 hrs per week. Also, the U shaped curve of median values of the box plots is more pronounced for people with financial issues as compared to financially secured groups. The higher ranges in box plots for people with financial issues can also be noted.

Research quesion 3:

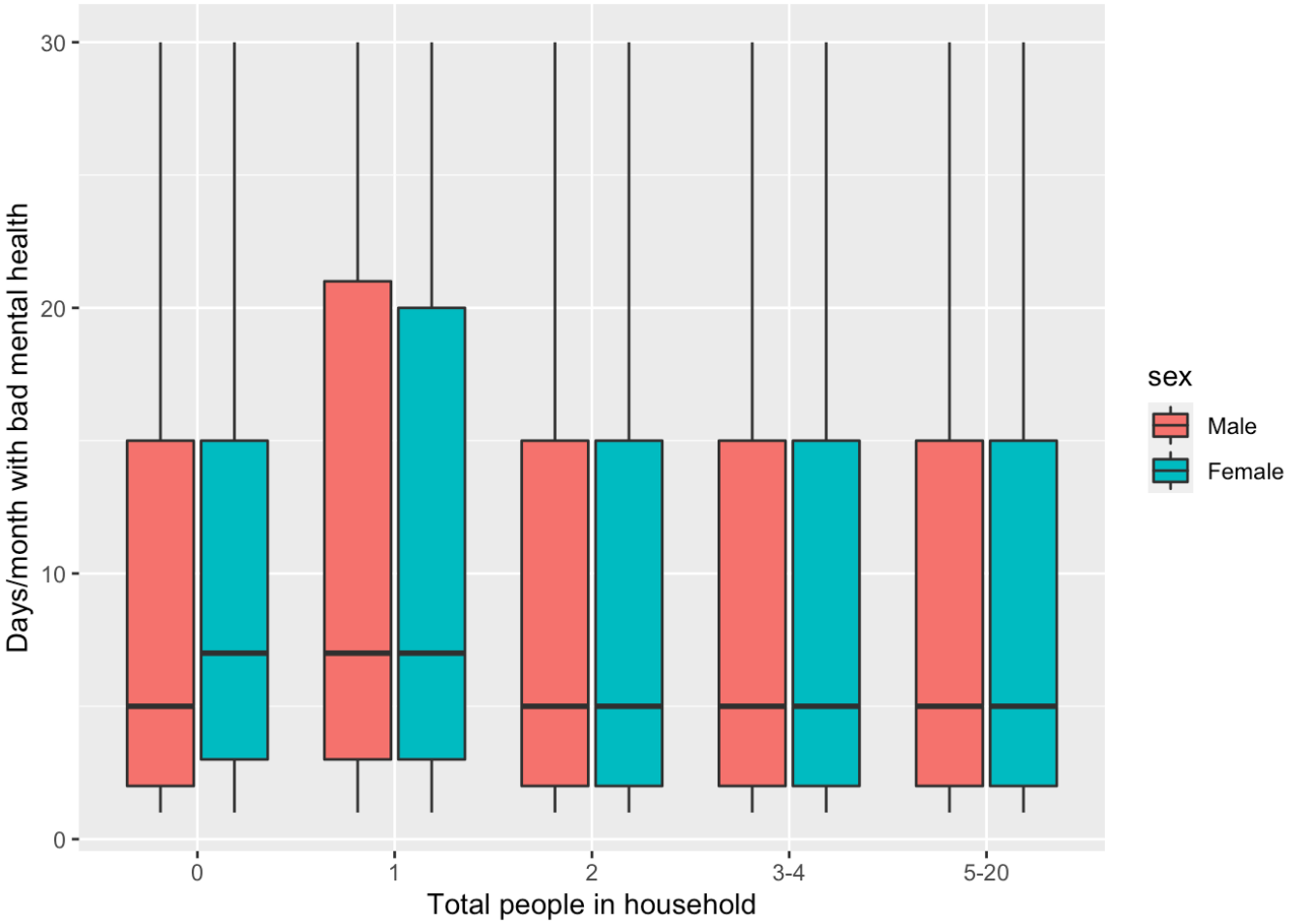
```
library(ggplot2)
library(dplyr)

## Selecting required variables ("sex"-represents gender, "menthlth" - indicates no of days in a month with bad mental health,"numadult"- number of adults in household,"children"- number of children in household) .And,filtering them for "NA". Also filtering for people WITH MENTAL HEALTH ISSUES.

q3dataset <- brfss2013 %>% select(sex,menthlth,numadult,children) %>% filter(!is.na(numadult)|!is.na(children),!is.na(menthlth),!is.na(sex),menthlth>0)

## Converting factor to integer for further calculations

q3dataset <- q3dataset %>% mutate(numadult=as.integer(q3dataset$numadult))

## Obtaining a new parameter ("total_people") indicating total people in household by combining two variables "numadult" and "child"

q3dataset <- q3dataset %>% mutate(total_people = ifelse(is.na(numadult),children,ifelse(is.na(children),numadult,ifelse(!is.na(numadult)|!is.na(children),numadult+children,"Unknown")) ))

## Converting integer data to categorical data for using in a box plot graph

q3dataset<-q3dataset %>% mutate(total_people_category=cut(total_people, breaks = c(0,1,2,3,5,21),labels = c("0","1","2","3-4","5-20"),include.lowest = TRUE,right = FALSE))

## Now let's find the average and median number of days with bad mental health for the two genders to see if there is any unexpected difference.

q3dataset %>% filter(sex=="Female") %>% summarise(avg_menthlth_fem=mean(menthlth),median_menthlth_fem=median(menthlth))
```

```
##   avg_menthlth_fem median_menthlth_fem
## 1         11.01115                   5
```

```
q3dataset %>% filter(sex=="Male") %>% summarise(avg_menthlth_male=mean(menthlth),median_menthlth_male=median(menthlth))
```

```
##   avg_menthlth_male median_menthlth_male
## 1          10.84885                     5
```

```
## The result verifies that there is no inherent difference in mental health by
virtue of gender.



## Next,Filtering to remove "NA" and to beautify graph

q3dataset<-q3dataset %>% filter(!is.na(total_people_category))

## Plotting the required graph

graph1<-ggplot(q3dataset,aes(x=total_people_category,y=menthlth,fill=sex))
graph1+geom_boxplot()+ xlab("Total people in household") + ylab("Days/month wit
h bad mental health")
```



Wow! Doctors seem to be kind of right. Having pople in household definately reduces bad mental health (and more so for women, rather surprisingly). Interes tingly it appears that having 2 people is as good as having 20, so increase in people above optimum is useless. There is a significant change from having one person to two indicating that having an elder or a child apart from having a si gnificant other could be helpful! Also note how mental health seems to be sligh tly adverse for women with zero or one person around (same median as compared t o previous category but a higher third quartile value) as compared to men for w hom mental health takes gets more bad with only one person around as compared t o noone around the household when compared to women(in case of men median and t hird quartile, both are up).