An Analysis on the BRFSS data

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

BRFSS is an ongoing surveillance system measuring behavioral risk factors for the non-institutionalized adult population, aged 18 years or older, who reside over the 50 U.S states plus the three U.S. territories. BRFSS's objective is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population.

Since 2011, BRFSS conducts both landline telephone- and cellular telephone-based survey to collect surveillance data on risk behaviors.In conducting the BRFSS landline telephone survey, interviewers collect data from a randomly selected adult in a household. In conducting the cellular telephone version of the BRFSS questionnaire, interviewers collect data from an adult who participates by using a cellular telephone and resides in a private residence or college housing.

According to BRFSS, in 2013, additional question sets were included as optional modules to provide a measure for several childhood health and wellness indicators, including asthma prevalence for people aged 17 years or younger.

The data from this project come from the "Behavioral Risk Factor Surveillance System". You can learn more about this survey on the following website: https://www.cdc.gov/brfss/.

For Generalizability, the sample is random and also a large representative,so the data sample is generalizable for the adult population of the US.

For Causality, BRFSS is kind of observational cross-sectional survey, so we cannot make causal inference through the data we have. We just make conclusions about association or correlation. Also, We will not assume that one outcome causes the others.

In this analysis, I would like to focus on the behavior of eating beans including cooked or canned beans and the health status of population.

Part 2: Research questions

In this analysis, I will be focus on answering three research questions:

Research question 1:

1. Is there any relation between the frequency of eating cooked or canned beans, such as refried, baked, black, garbanzo beans, beans in soup, soybeans, edamame, tofu or lentils and the participant's body mass index (BMI), overall, and between sexes?

- `beanday_` - Computed Bean Intake In Times Per Day
- `_bmi5cat` - Computed Body Mass Index Categories
- `sex` - Indicate sex of respondent

Background: Some studies have raised that eating beans could cause weight gain if eaten in large amounts too often. In this question, I want to explore this possible relation, under the hypothesis that people who eat less have a lower BMI.

Research question 2:

2. Are there any differences in the general health status between the frequency groups of the behavior of eating beans in terms of sex?

- `genhlth` - General Health
- `beanday_` - Computed Bean Intake In Times Per Day
- `sex` - Indicate sex of respondent

Background: It is widely believed that eating beans can contribute to the good health status. In this question, I want to check whether the health status of those eating more beans are better than those eating less beans or without eating-beans behaviors.

Research question 3:

3. Is there any tendency that people who take part in sporty activities including Walking, Running, Jogging, Or Swimming eat more beans than those who do not?

- `beanday_` - Computed Bean Intake In Times Per Day
- `exeroft1` - How many times per week or per month did you take part in this activity during the past month?

Background: Many people say that sporty activities like Walking, Running, Jogging, Or Swimming requires much energy and so people partake in these activities tend to consume more high-energy food, especially beans-based products. In this analysis, I am willing to figure out the behavior of eating beans of sporty people.

Part 3: Exploratory data analysis

Let us begin with the eating-beans behavior and the status of body mass index (BMI).

Research question 1:

Firstly, we should have a look at the chosen columns' data. About 5.17 percent of the data is NA, it is not a big problem for the large data sample with nearly 500K observations.

```
question1 <- brfss2013 %>% select(beanday_,X_bmi5cat,sex)
colSums(is.na(question1))
```

```
##  beanday_ X_bmi5cat       sex
##     37495     26727         7
```

```
percentage = mean(is.na(question1)) * 100
print ("percentage of missing values")
```

```
## [1] "percentage of missing values"
```

```
print (percentage)
```

```
## [1] 4.353549
```

Hence, we should remove the NAs and clean the data.

```
#Remove NAs
question1 <- brfss2013 %>%
   filter(!is.na(beanday_)&!is.na(X_bmi5cat)&!is.na(sex))%>%
   select(beanday_, X_bmi5cat,sex)
#Exploring the variables
str(question1)
```

```
## 'data.frame':    433134 obs. of  3 variables:
##  $ beanday_ : int  10 33 29 29 100 0 100 29 43 7 ...
##  $ X_bmi5cat: Factor w/ 4 levels "Underweight",..: 4 1 3 2 4 4 2 4 3 3 ...
##  $ sex      : Factor w/ 2 levels "Male","Female": 2 2 2 2 1 2 2 1 2 1 ...
```

Here, we have the summary statistics of the sample.

```
summary(question1)
```

```
##     beanday_              X_bmi5cat          sex
##  Min.   :   0.00   Underweight  :  7533   Male  :182429
##  1st Qu.:   7.00   Normal weight:143525   Female:250705
##  Median :  14.00   Overweight   :155479
##  Mean   :  27.85   Obese        :126597
##  3rd Qu.:  33.00
##  Max.   :9900.00
```

We also see the summary contribution by sex and BMI category in the below table. Obviously, there is little concern of the insufficient nutrition of the population in both sexes when seeing the population nutritional status. More people are in obese status (nearly 30%) rather than in underweight condition (appx.2% or less).

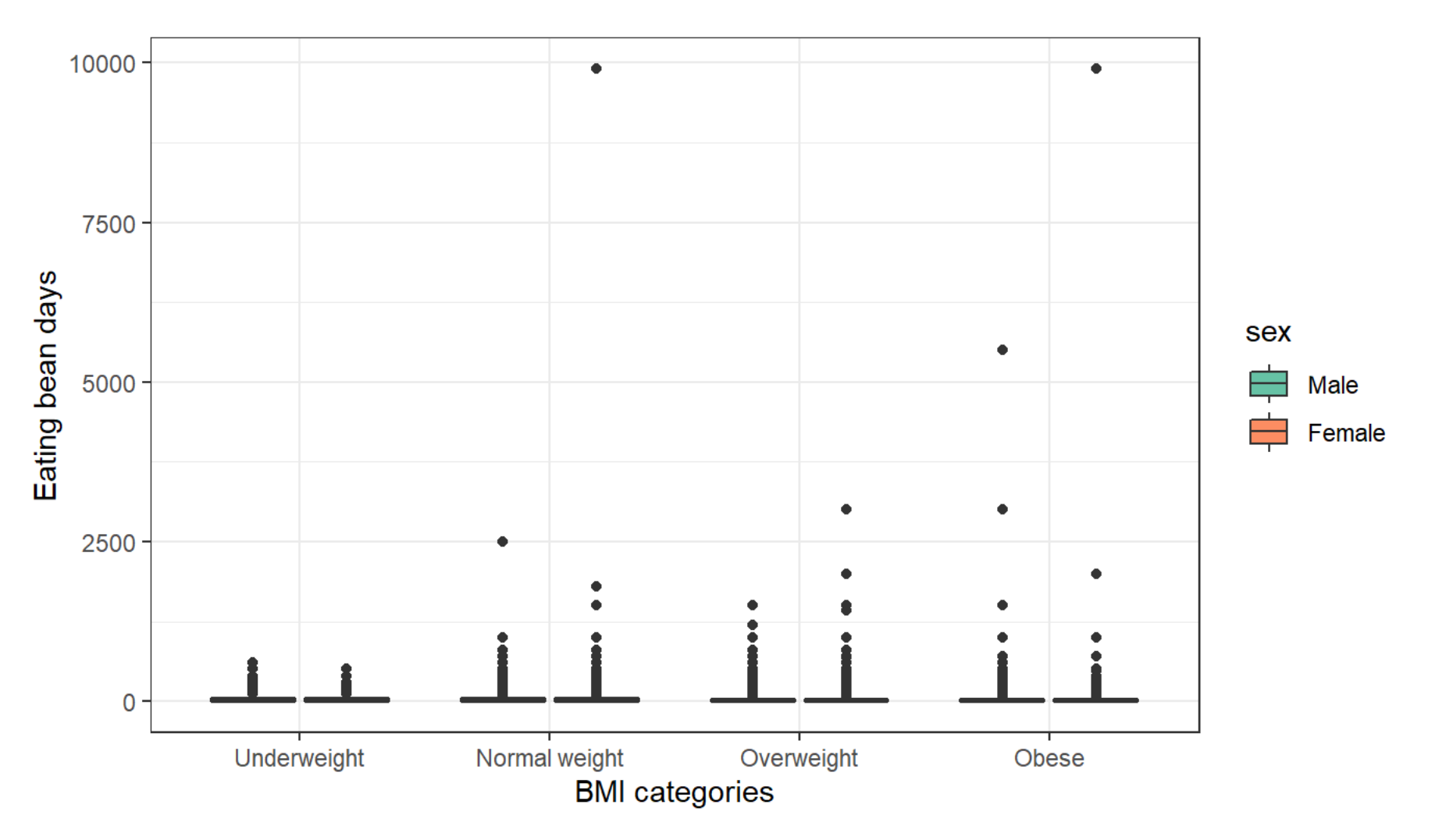
```
#Descriptive Statistics
table1 <- table(question1$sex, question1$X_bmi5cat)
```

```
prop.table(table1, 1)
```

```
##
##         Underweight Normal weight  Overweight       Obese
##   Male   0.009455733   0.267462958 0.429657565 0.293423743
##   Female 0.023166670   0.377862428 0.307520791 0.291450111
```

I want to check the outliers of the bean intake per day by using a boxplot. It is clear to pick out some big but I think impossible for eating-bean times per day.

```
#Plotting boxplot
ggplot(question1, aes(x=X_bmi5cat, y=beanday_, fill=sex)) +
  geom_boxplot() + labs(y = "Eating bean days", x = "BMI categories") +
  theme_bw() + scale_fill_brewer(palette="Set2")
```



We dig deep into the case by grouping the number of intake times.There are not many observations with the beandays_ value over 200 times per day - also the values that I believe is not realistic and usual in the sample. So, we will conduct a filtering process again to clear out these records (beandays_ > 200).

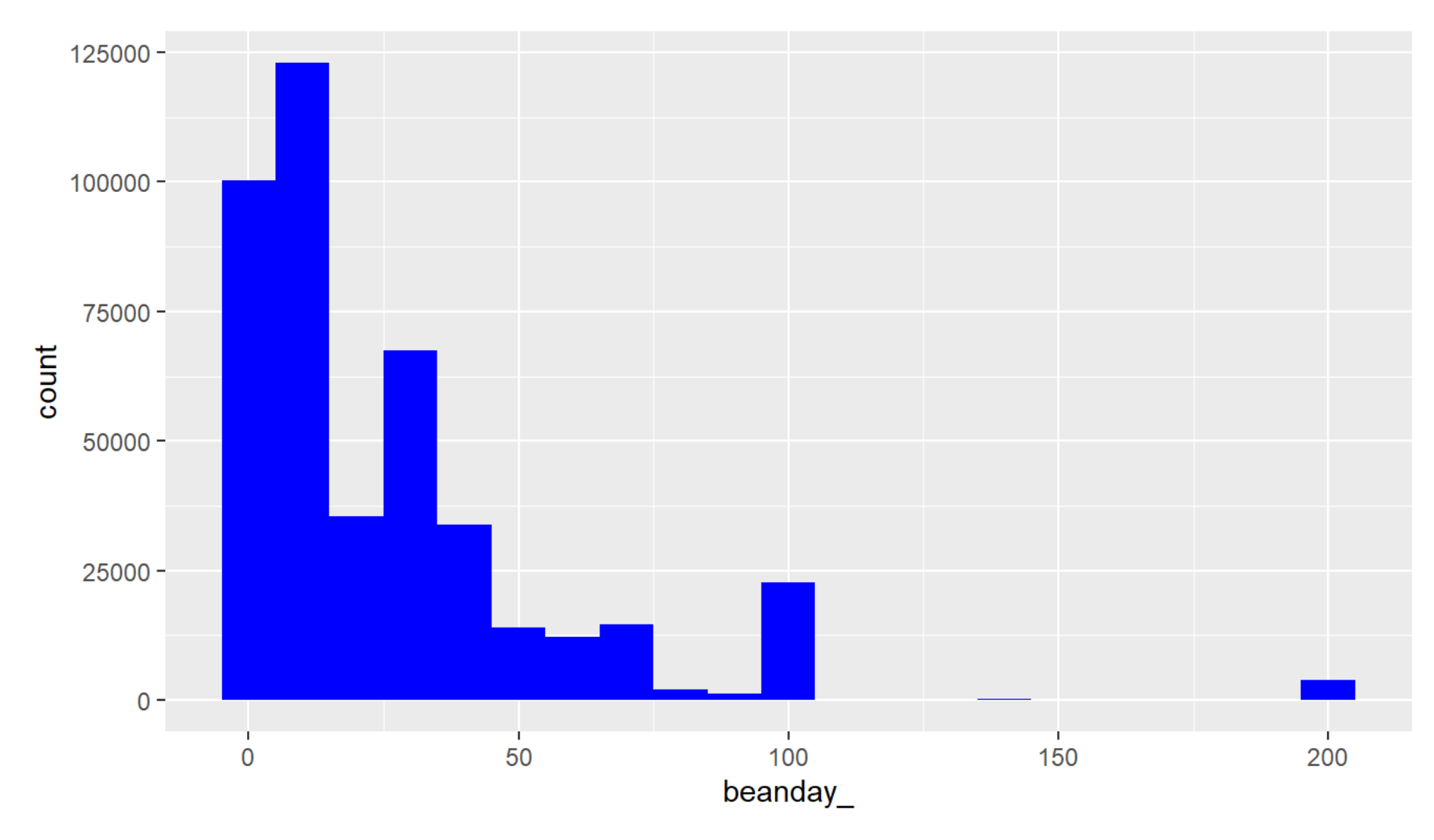
```
table(cut(question1$beanday_, breaks=c(0,200, 300, 500, 1000,Inf)))
```

```
##
##     (0,200]   (200,300]   (300,500] (500,1e+03] (1e+03,Inf]
##      357943        1196         542          54          16
```

```
question1 <- question1 %>%
  filter(beanday_<=200)
```

I plot to see the distribution of beanday_. It is nearly right-skewed distribution.

```
ggplot(question1,aes(x=beanday_))+
  geom_histogram(fill="blue", binwidth = 10)
```

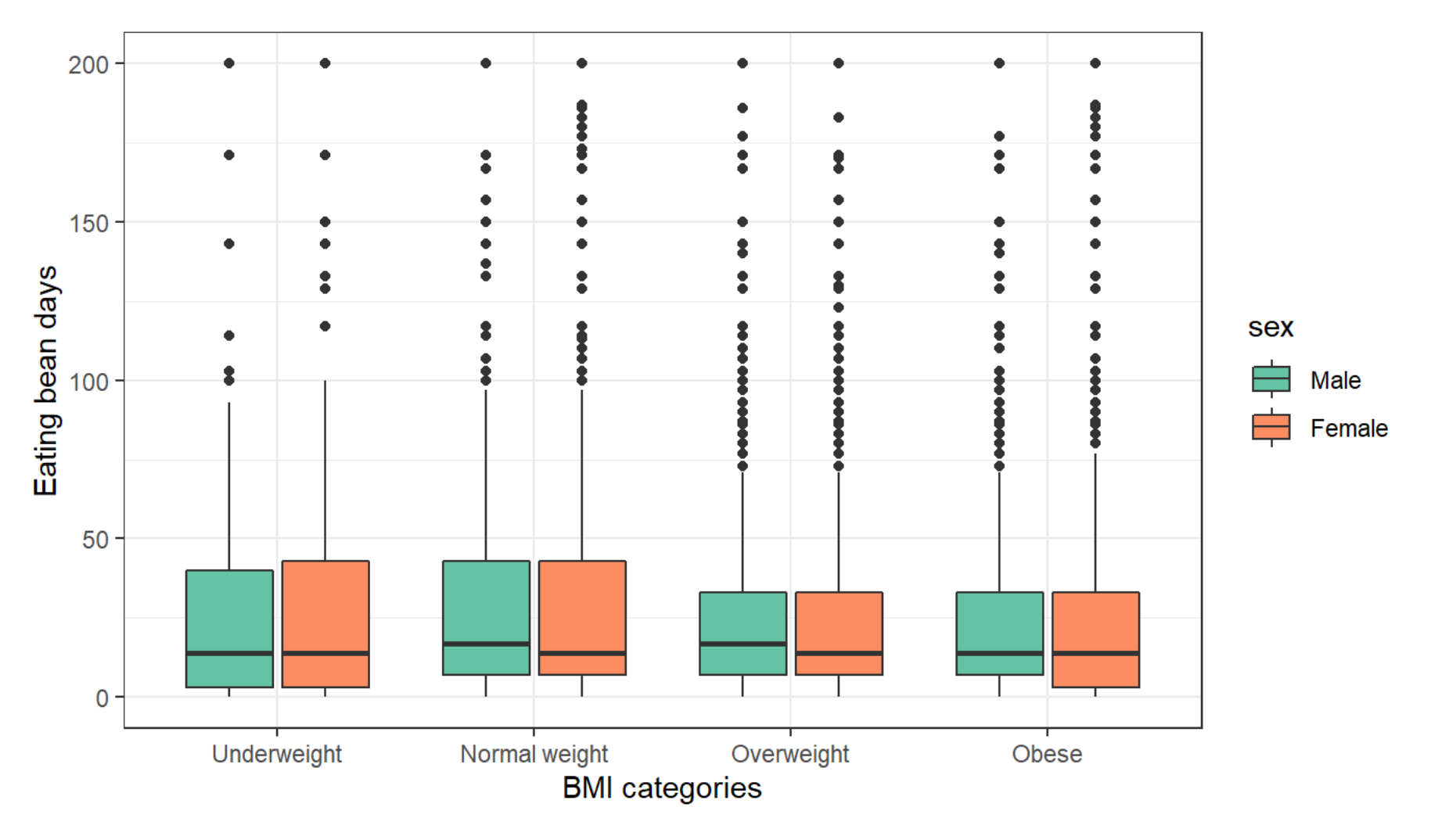


Here, it is more straightforward to recognize the relation between the bean intake and BMI of people. For both groups of BMI types and sexes, people tend to consume bean under 50 times per day.

It is noticeable that underweight and normal weight people have more frequent bean-eating behaviours while overweight and obese groups do not have a tendency of eating more beans daily or even less than normal people.

Here we can figure out that eating beans do not lead to overweight or obese status, but it is a good behavior for people health if kept in a moderate manner.

```
#Plotting boxplot
ggplot(question1, aes(x=X_bmi5cat, y=beanday_, fill=sex)) +
  geom_boxplot() + labs(y = "Eating bean days", x = "BMI categories") +
  theme_bw() + scale_fill_brewer(palette="Set2")
```



Research question 2:

Now let's move on to the next aspect of the relation between general health status and bean intake behaviour to clarify if eating beans can contribute to the good health status.

Firstly, we also prepare the data to answer the question.

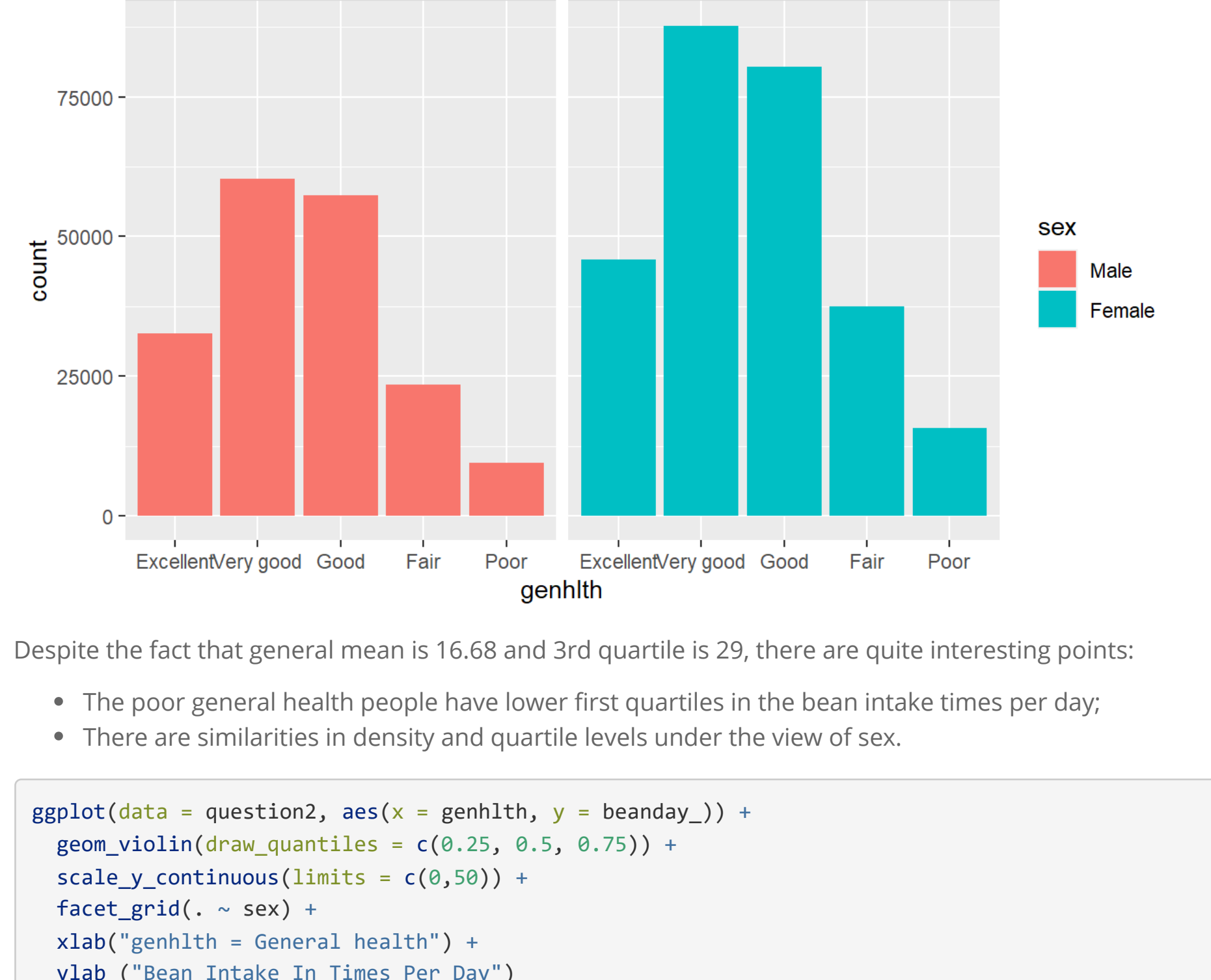
```
question2 <- brfss2013 %>%
  filter(!is.na(beanday_)&!is.na(genhlth)&!is.na(sex))%>%
  select(beanday_, genhlth,sex)
question2 <- question2 %>%
  filter(beanday_<=200)

summary(question2)
```

```
##     beanday_          genhlth          sex
##  Min.   :  0.00   Excellent: 78625   Male  :183512
##  1st Qu.:  7.00   Very good:148126   Female:267160
##  Median : 14.00   Good     :137764
##  Mean   : 26.41   Fair     : 60882
##  3rd Qu.: 33.00   Poor     : 25275
##  Max.   :200.00
```

I have curiosity on the general health status of the surveyed population by sex.

```
ggplot(question2, aes(x= genhlth,group=sex))+
  geom_bar(mapping = aes(fill=sex))+
  facet_grid(. ~ sex)
```

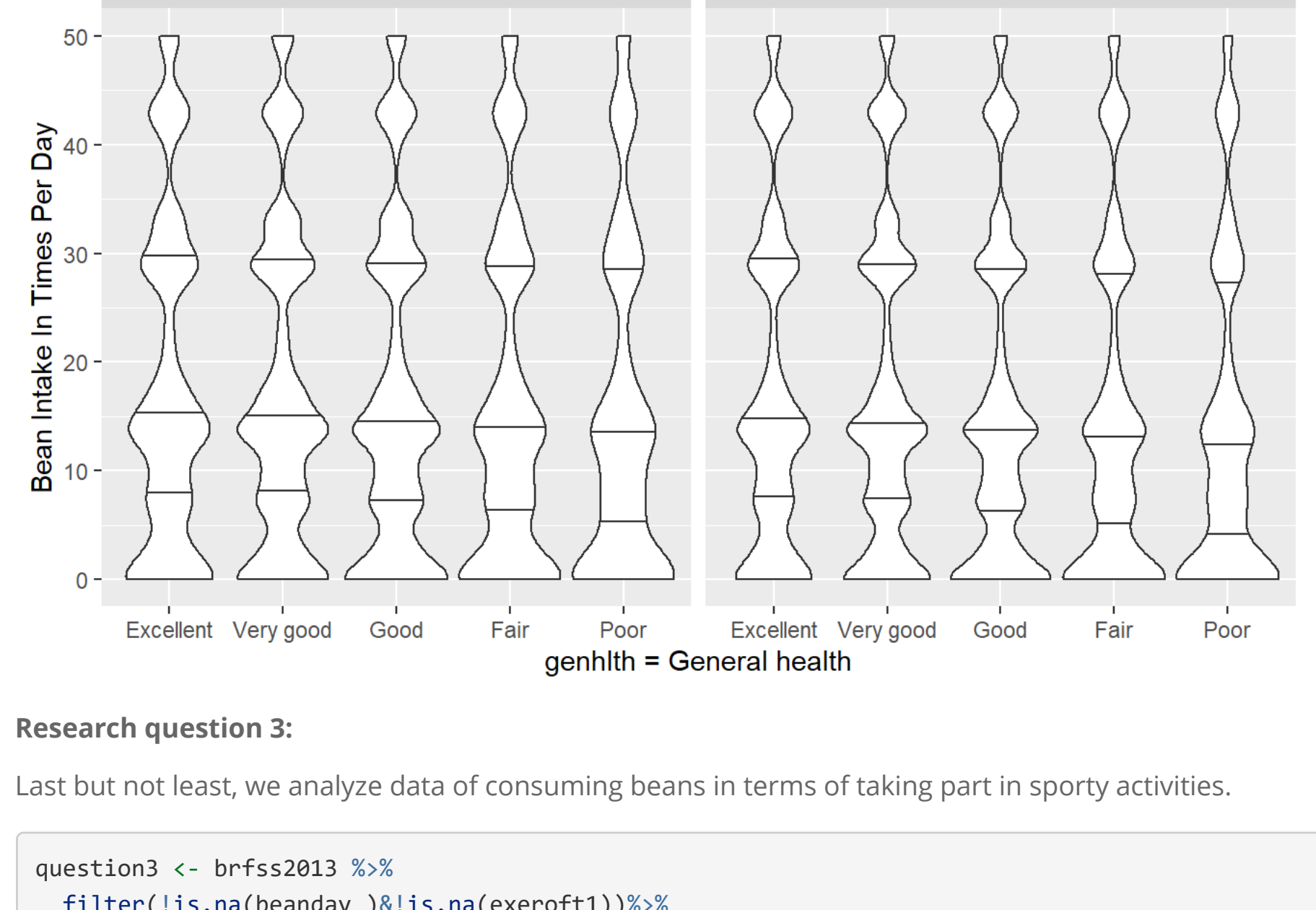


Despite the fact that general mean is 16.68 and 3rd quartile is 29, there are quite interesting points:

- The poor general health people have lower first quartiles in the bean intake times per day;
- There are similarities in density and quartile levels under the view of sex.

```
ggplot(data = question2, aes(x = genhlth, y = beanday_)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75)) +
  scale_y_continuous(limits = c(0,50)) +
  facet_grid(. ~ sex) +
  xlab("genhlth = General health") +
  ylab ("Bean Intake In Times Per Day")
```

```
## Warning: Removed 60315 rows containing non-finite values (stat_ydensity).
```



Research question 3:

Last but not least, we analyze data of consuming beans in terms of taking part in sporty activities.

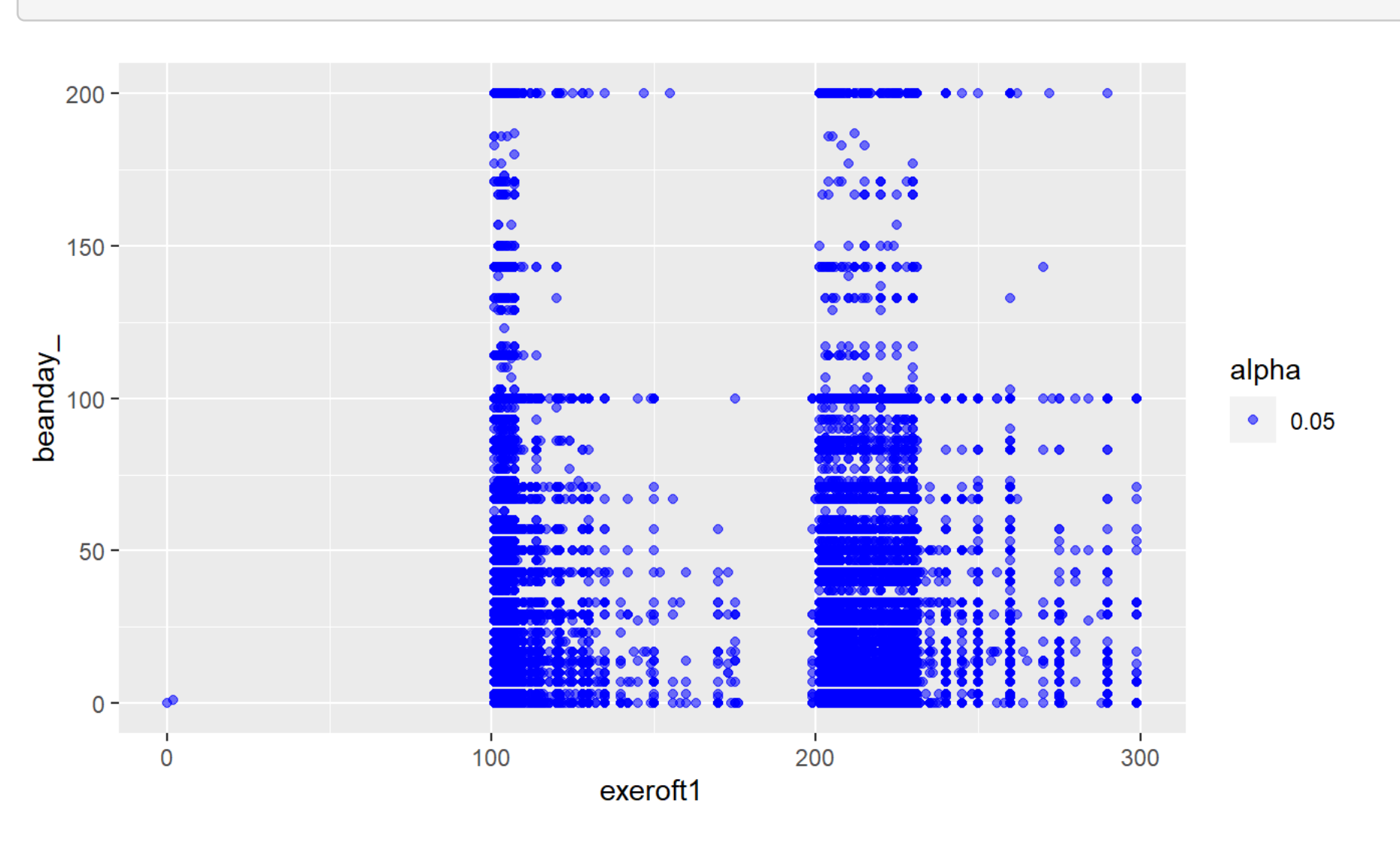
```
question3 <- brfss2013 %>%
  filter(!is.na(beanday_)&!is.na(exeroft1))%>%
  select(beanday_, exeroft1)
question3 <- question3 %>%
  filter(beanday_<=200)
```

We will not need to categorize the surveyed population into any groups. However, thanks to the scatter plot, we can see the two groups with values of [101 - 199]- Times per week and [201 - 299] - Times per month.

In my opinion, putting these two groups into one coordinate axis to compare is not meaningful and suitable, but in this situation, we just concentrate on the relation but not the absolute values.

It should be admitted that people who use beans more frequently do not have the tendency of taking part in sporty activities including Walking, Running, Jogging, Or Swimming. It even seems that sporty people consume beans in an appropriate frequency (higher density under 50).

```
ggplot(question3)+
  geom_point(mapping=aes(x=exeroft1,y=beanday_, alpha=0.05), colour="blue")
```



In summary, bean intake behavior can be considered the good behavor if we make use of beans in an appropriate usage level regarding sex, heath status, sporty people or not.

@ This analysis is conducted by KhuongDT (data introduction ref from BFRSS)