Exploring 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

According to the BRFSS document, the data had been collected by both landline telephone and cellular telephone-based surveys. In terms of the landline telephone survey, interviewers selected a random adult in a household. For the cellular telephone surveys, data was collected from a random adult who participated with a cellular telephone and resided in a private residence or college housing. Consequentely, a random stratified sampling was conducted when observations were collected, which means the data set is generalized to the populatoin at large.

However, because a random assignment was not conducted during the data collection, the association between variables can be observed though, the causality cannot be varified.

Part 2: Research questions

Research quesion 1:

People now have more career options compared to the past. However, self-employed may have more mental pressure and less sleep time than those employed by wages. So it maybe interesting to compare the sleep time and mental health status by people’s employment status.

Q1: What is the difference of people’s sleep time and mental health status by the employment status (employed by wages or self-employed)?

Used variables:

employ1: Employment Status

sleptim1: How Much Time Do You Sleep

menthlth: Number Of Days Mental Health Not Good

Research quesion 2:

Obesity is always related to high blood pressure and people have high income seems to be likely to have high blood pressure. So it seems interesting to prove this that whether people with high income are tend to be overweight and get high blood pressure.

Q2: What is the defference in income level and overweight status by whether peolpe are taking blood pressure medicatiion or not?

Used variables:

bpmeds: Currently Taking Blood Pressure Medication

income2: Income level

X_bmi5cat: Computed Body Mass Index Categories

Research quesion 3: We would like to know how common does a person drink alochol or smoke? People drinking alochol or smoking is so normal but we rarely care about whether people both drink alcohol and smoke. For example, what's the probability of a man smoke but don't drink alcohol?

Q3: Which one can be more universal, Smoking or drinking alcohol?

Used variables:

smokday2: Frequency Of Days Now Smoking

X_drnkdy4: Computed Number Of Drinks Of Alcohol Beverages Per Day

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.

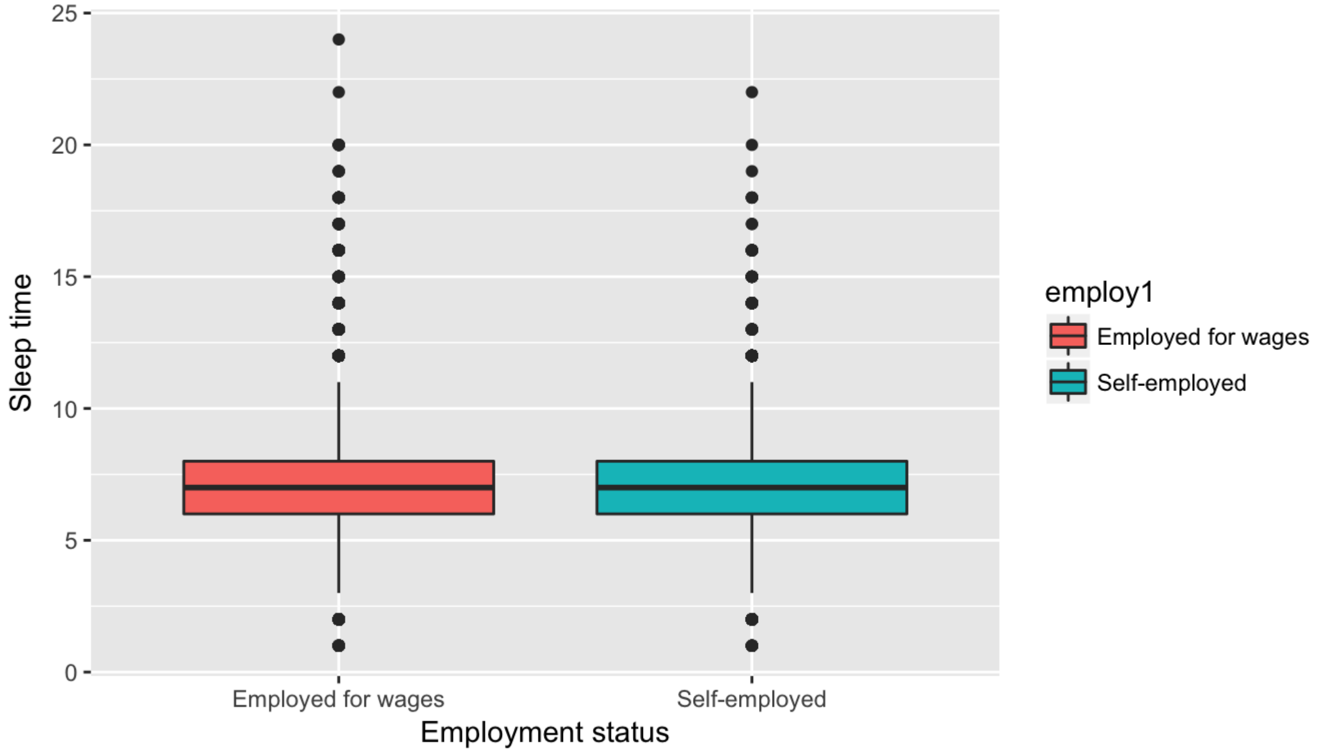
Research quesion 1:

To analyse the difference by people who employed by wages or self-employed, the first step is to clean up the dataset.

```
q1 <- brfss2013 %>%
  select(employ1, sleptim1, menthlth)%>%
  filter(employ1 %in% c("Employed for wages","Self-employed"))%>%
  filter(!is.na(sleptim1) & !is.na(menthlth))%>%
  group_by(employ1)
```

By plotting the distrubution of sleep time, it suggestes that there's no obvious differences between people employed for wages and self-employed.

```
ggplot(q1, aes(x=employ1, y=sleptim1, fill=employ1))+geom_boxplot()+xlab("Employment
 status")+ylab("Sleep time")
```

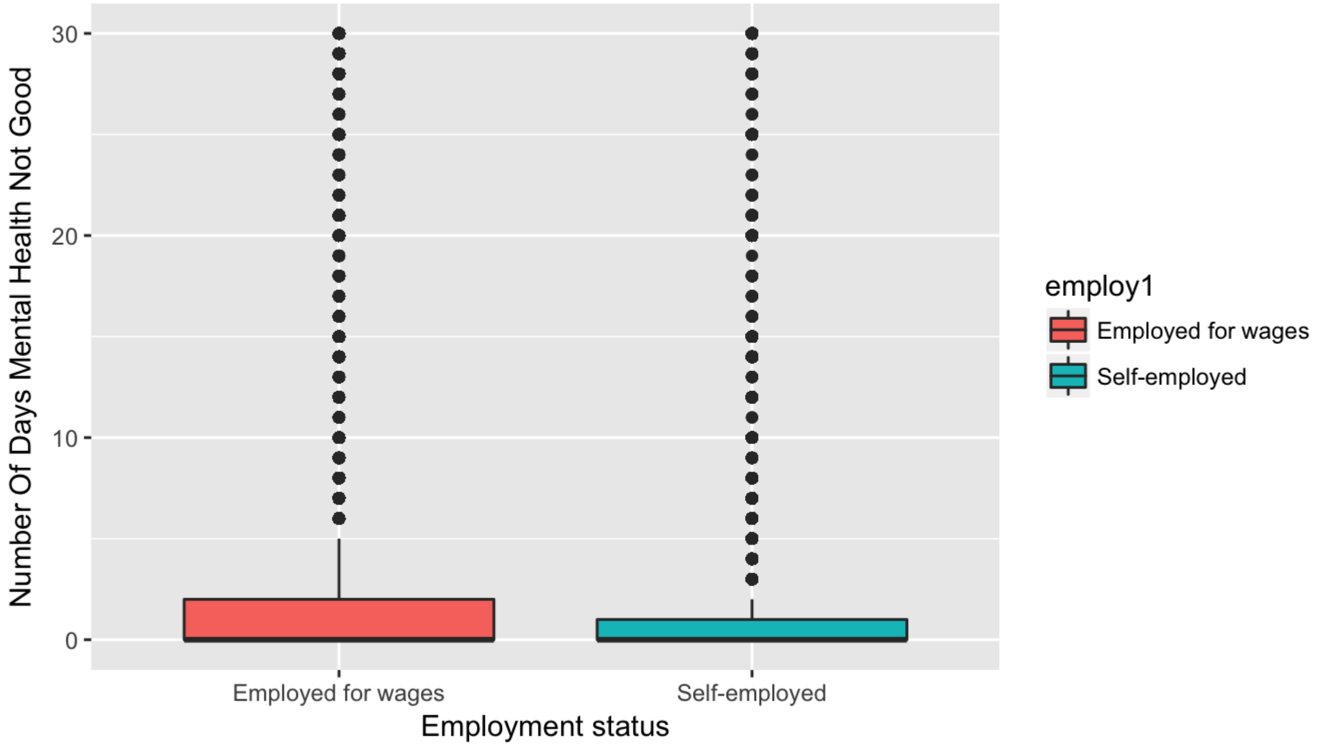


The summary data also proves it. The sleep time of both groups is concentrating on 6-8 hours and implies a right skewed distribution.

```
summarise(q1,sleptim_x_bar=mean(sleptim1),sleptim_sd=sd(sleptim1),sleptim_median=medi
an(sleptim1),sleptim_IQR=IQR(sleptim1))
```

```
## # A tibble: 2 x 5
##                employ1 sleptim_x_bar sleptim_sd sleptim_median sleptim_IQR
##                 <fctr>         <dbl>      <dbl>          <dbl>       <dbl>
## 1 Employed for wages      6.892329   1.209586              7           2
## 2      Self-employed      7.081981   1.255548              7           2
```

```
ggplot(q1, aes(x=employ1, y=menthlth, fill=employ1))+geom_boxplot()+xlab("Employment
 status")+ylab("Number Of Days Mental Health Not Good")
```



```
summarise(q1, menthlth_c_bar=mean(menthlth),menthlth_sd=sd(menthlth),menthlth_median=
median(menthlth),menthlth_IQR=IQR(menthlth))
```

```
## # A tibble: 2 x 5
##              employ1 menthlth_c_bar menthlth_sd menthlth_median
##               <fctr>          <dbl>       <dbl>           <dbl>
## 1 Employed for wages       2.685692    6.511640               0
## 2      Self-employed       2.394581    6.370896               0
## # ... with 1 more variables: menthlth_IQR <dbl>
```

The plot and summary data shows that the mental health status of self-employed people is slightly better than those employed by wages because the number of days mental health not good is less, but the difference is not that significant.

We need to know the situation of proportion of those who have bad mental health.

```
q1 %>%
  filter(employ1=="Employed for wages")%>%
  group_by(menthlth)%>%
  summarise(count=n())%>%
  mutate(prop=round(count*100/sum(count),1))
```

```
## # A tibble: 31 x 3
##    menthlth  count  prop
##       <int>  <int> <dbl>
##  1        0 137520  69.1
##  2        1   7939   4.0
##  3        2  11748   5.9
##  4        3   6283   3.2
##  5        4   3074   1.5
##  6        5   7886   4.0
##  7        6    745   0.4
##  8        7   2741   1.4
##  9        8    474   0.2
## 10        9     72   0.0
## # ... with 21 more rows
```

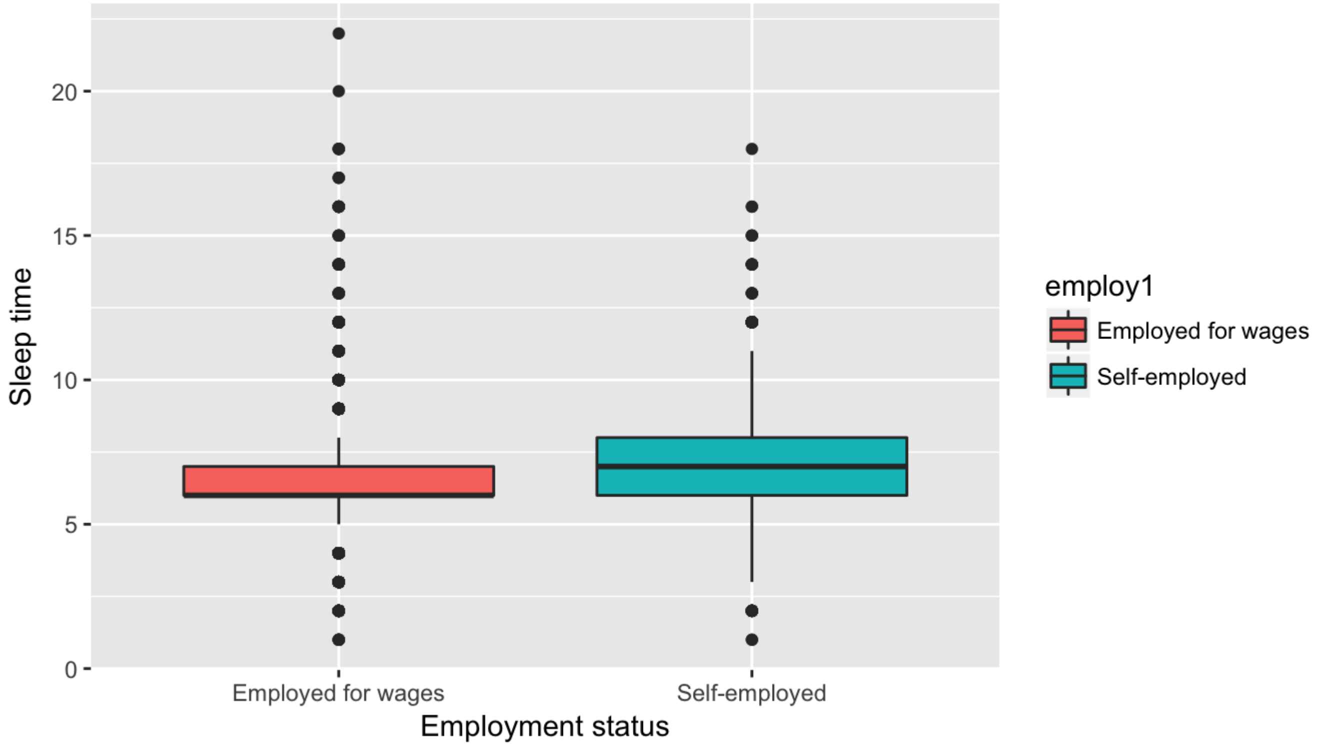
```
q1%>%
  filter(employ1=="Self-employed")%>%
  group_by(menthlth)%>%
  summarise(count=n())%>%
  mutate(prop=round(count*100/sum(count),1))
```

```
## # A tibble: 31 x 3
##    menthlth count  prop
##       <int> <int> <dbl>
##  1        0 28837  73.8
##  2        1  1351   3.5
##  3        2  1913   4.9
##  4        3  1025   2.6
##  5        4   511   1.3
##  6        5  1253   3.2
##  7        6   124   0.3
##  8        7   408   1.0
##  9        8    92   0.2
## 10        9    18   0.0
## # ... with 21 more rows
```

We found that the data of bad mental health days of both groups are gather between 0 and 5. So we define “good mental health” as the observation of menthlth is less or equal to 5. 88.3% of employed by wages and 89.3% of self-employed are have relatively good mental health status.

Consequently, it’s easy for us to analyse those who have bad mental health status.

```
q1_5 <- q1 %>%
  filter(menthlth>5)
ggplot(q1_5, aes(x=employ1, y=sleptim1, fill=employ1))+geom_boxplot()+xlab("Employmen
t status")+ylab("Sleep time")
```



So there’s a significant difference for those who have bad mental health status that, most people employed by wages have less sleep time and more extreme outliers are found.

Research quesion 2:

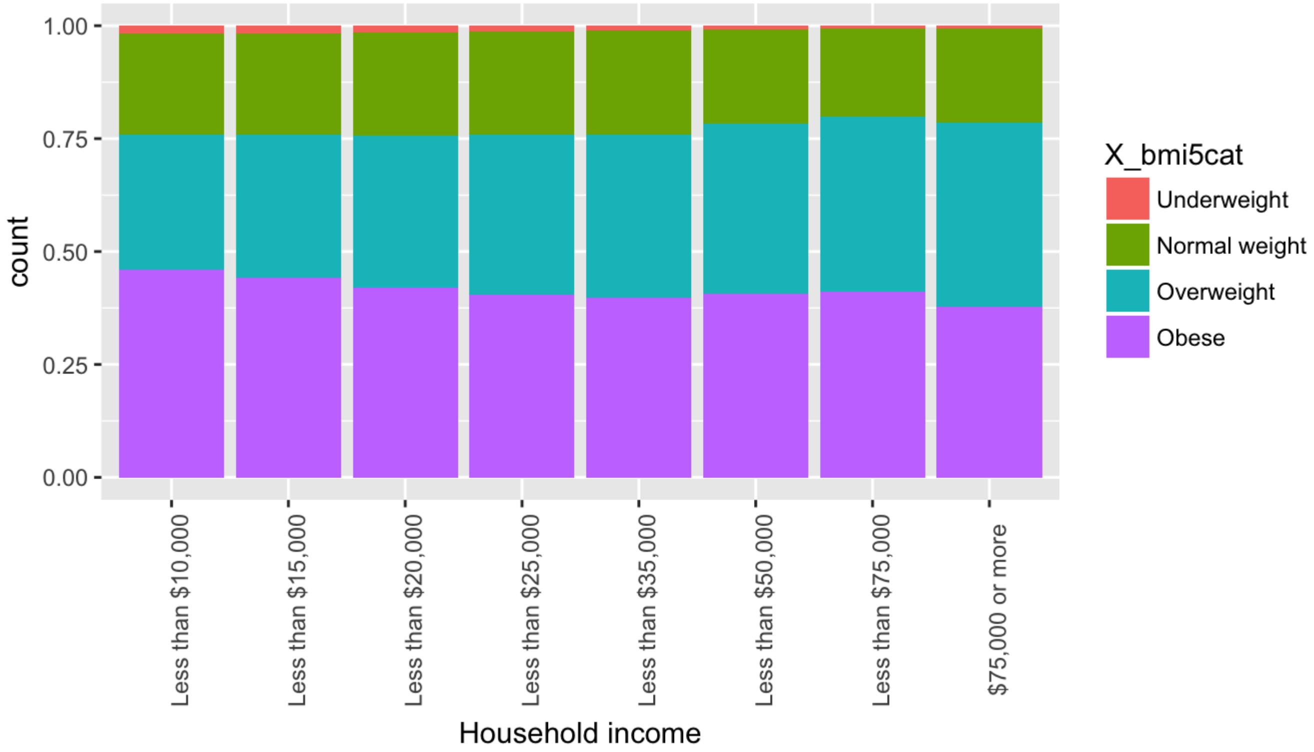
Firstly we would like to collect the data we need. The meanless data will be get rid of.

```
q2 <- brfss2013 %>%
  select(bpmeds, income2, X_bmi5cat)%>%
  filter(!is.na(bpmeds) & !is.na(income2) & !is.na(X_bmi5cat))
```

3 variables and 161833 observations are gathered to the new dataset named as “q2”.

Then we would like to see the association between income level and obesity status in general by bar plot.

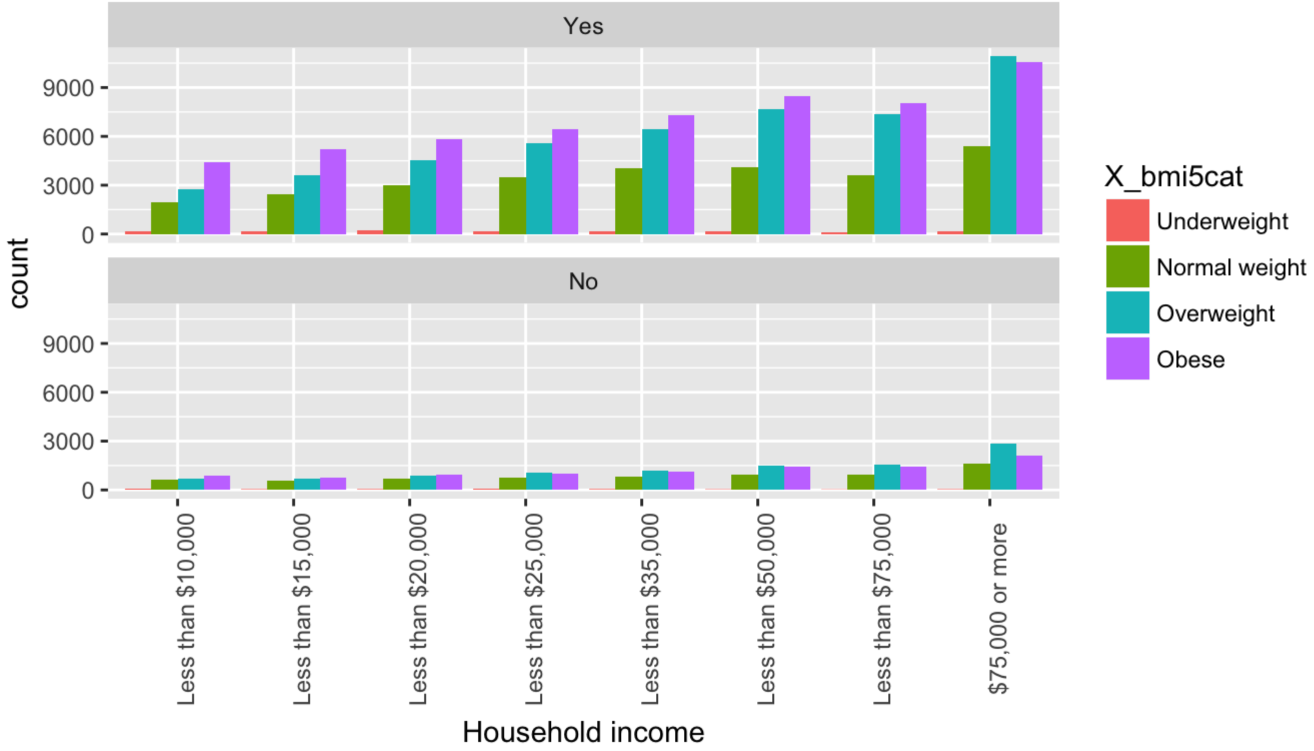
```
ggplot(q2, aes(x=income2, fill=X_bmi5cat))+geom_bar(position = "fill")+theme(axis.tex
t.x = element_text(angle = 90))+xlab("Household income")
```



The plot shows that as the income level improves, the proportion of Overweight" and “Obese” increase and the proportion of “Underweight” and “Normal weight” decrease.

To compare the status of whether people are taking blood pressure medication, we plot the comparison chart to examine it.

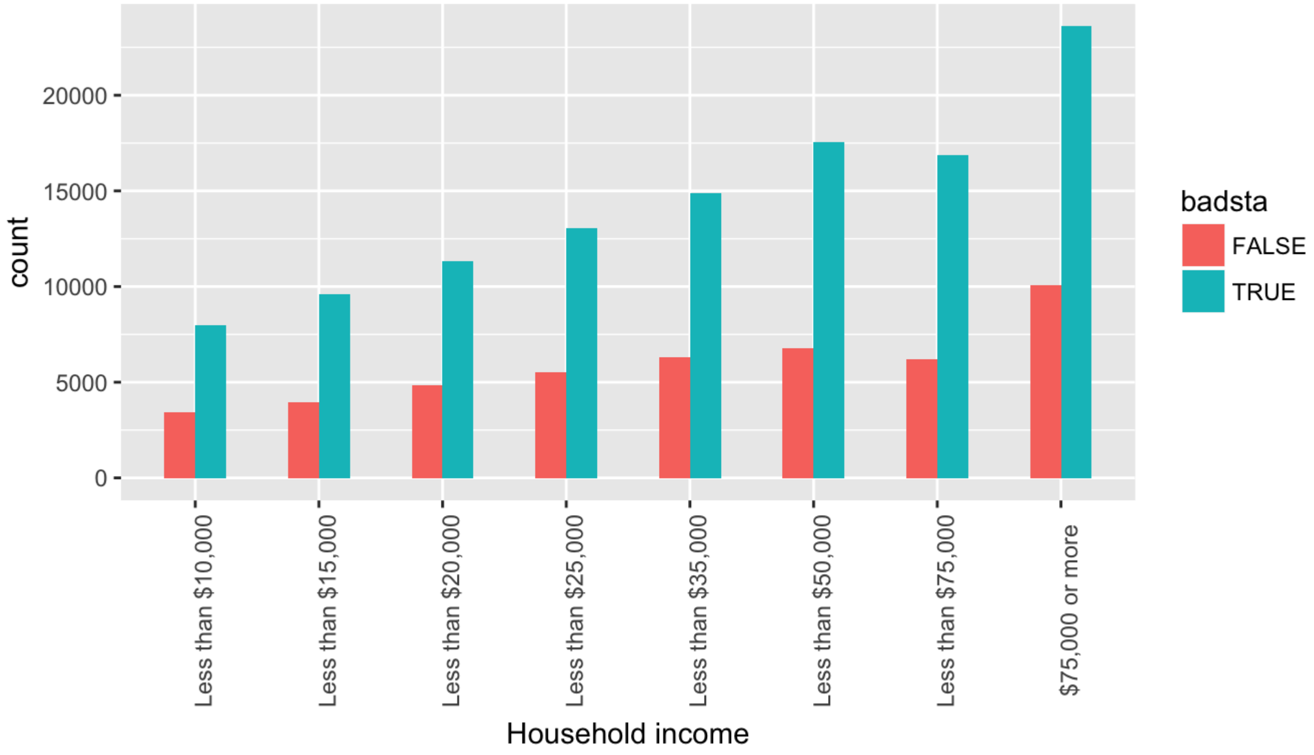
```
ggplot(q2, aes(x=income2, fill=X_bmi5cat))+geom_bar(position = "dodge")+facet_wrap(~b
pmeds, ncol = 1)+theme(axis.text.x = element_text(angle = 90))+xlab("Household incom
e")
```



We can find a significant increasing in both plots when the income level higher than $50,000.

Now we would like to focus on those who in “bad status”: be taking blood pressure medication and be overweight in the meanwhile.

```
q2 <- q2%>%
  mutate(badsta = case_when(bpmeds=="Yes" & X_bmi5cat=="Overweight" | X_bmi5cat=="Obe
se" ~ TRUE,TRUE ~ FALSE))
ggplot(data = q2, aes(x = income2, fill = badsta)) + geom_bar(width = 0.5, position =
 "dodge")+theme(axis.text.x = element_text(angle = 90))+xlab("Household income")
```



The dodge plot clearly shows a increasing trend of the number of people with bad status as the income level increasing, despite the observation of “Less than $75,000”.

```
q2 %>%
  filter(badsta=="TRUE")%>%
  group_by(income2)%>%
  summarise(count=n())%>%
  mutate(prop=round(count*100/sum(count),1))
```

```
## # A tibble: 8 x 3
##               income2 count  prop
##                <fctr> <int> <dbl>
## 1 Less than $10,000  7969   6.9
## 2 Less than $15,000  9592   8.4
## 3 Less than $20,000 11301   9.8
## 4 Less than $25,000 13042  11.4
## 5 Less than $35,000 14881  13.0
## 6 Less than $50,000 17571  15.3
## 7 Less than $75,000 16847  14.7
## 8   $75,000 or more 23608  20.6
```

There are about 50.6% of people with bad status are belong to the group of “high income level” (higher than $35,000), and as the income level increases, the proportion of number of people who have bad status increases too, despite the observation of “Less than 75,000” especially. Maybe something happend in this range but more detailed data are needed to interpret this phenomenon.

Research quesion 3:

Step1: Data preparation

Firstly, we would like to filter the data we need and remove the observations of NA values.

```
q3 <- brfss2013 %>%
  select(smokday2, X_drnkdy4)%>%
  filter(!is.na(smokday2) & !is.na(X_drnkdy4))
```

The new dataset “q3” includes 209034 observations and 3 variables.

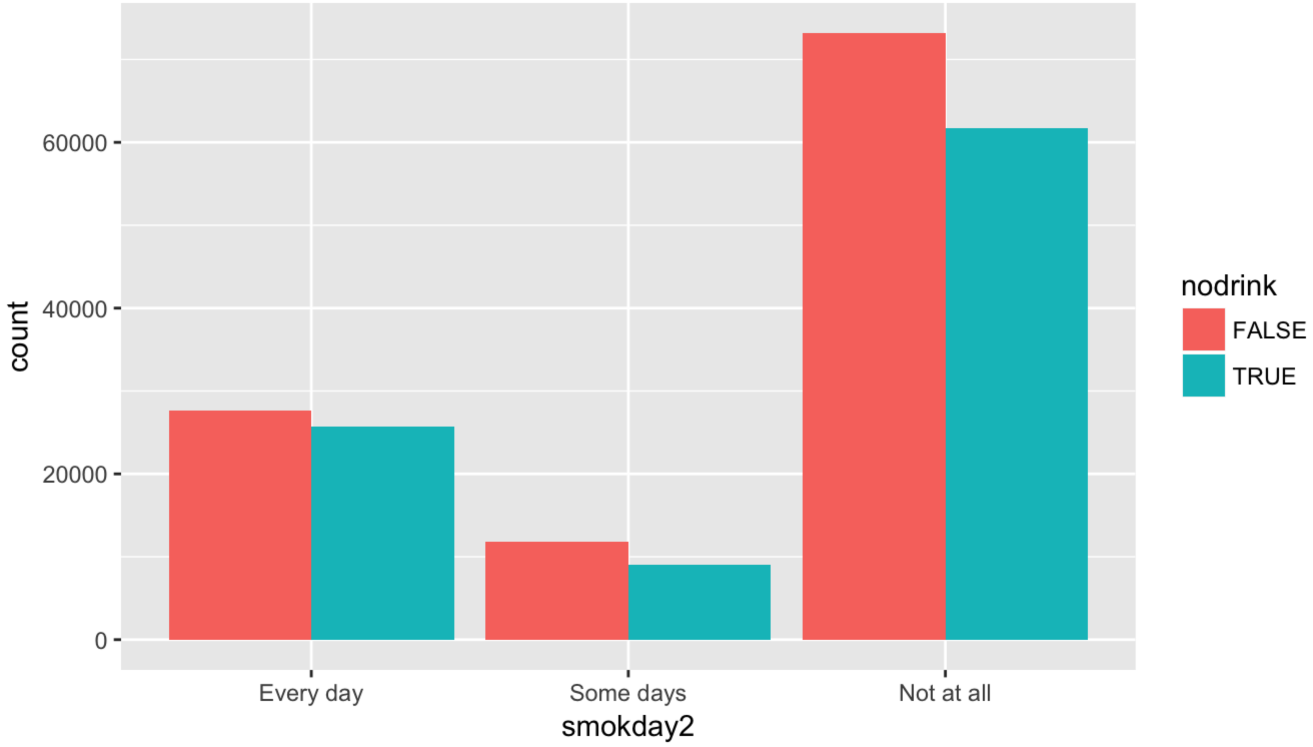
```
nrow(q3)/nrow(brfss2013)
```

```
## [1] 0.4250602
```

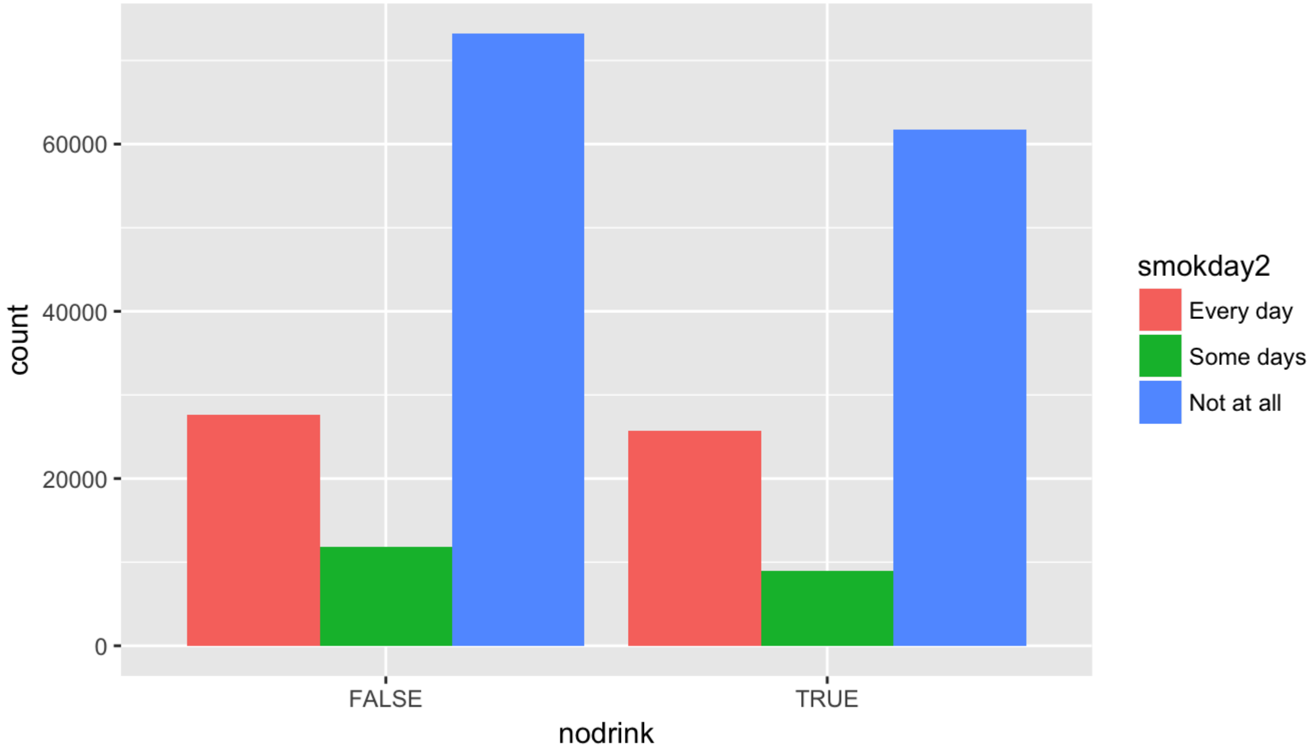
The data we chosen occupies 42.5% of the total data, which implies the analysis can be generalized to some extent.

Step2: Begin to EDA We would like to explore the association between smoking and drinking first.

```
q3 <- q3%>%
  mutate(nodrink=case_when(X_drnkdy4==0~TRUE,TRUE~FALSE))
ggplot(q3, aes(x=smokday2, fill=nodrink))+geom_bar(position = "dodge",)
```



```
ggplot(q3, aes(x=nodrink, fill=smokday2))+geom_bar(position = "dodge")
```



We can find that although there’s no significant difference of the frequency of people smoke in drinking alcohol and don’t drink alcohol, there are much more people who don’t smoke at all but drink alcohol than those have ever smoked.

```
q3%>%
  group_by(smokday2)%>%
  summarise(count=n())%>%
  mutate(prop=round(count/sum(count),3))
```

```
## # A tibble: 3 x 3
##     smokday2  count  prop
##       <fctr>  <int> <dbl>
## 1  Every day  53379 0.255
## 2  Some days  20767 0.099
## 3 Not at all 134888 0.645
```

We can see there are 64.5% of people don’t smoke at all which occpies most proportion of the total dataset.

```
q3%>%
  filter(smokday2=="Not at all")%>%
  group_by(nodrink)%>%
  summarise(count=n()) %>%
  mutate(share_nodrink=round(count/sum(count),3))
```

```
## # A tibble: 2 x 3
##   nodrink count share_nodrink
##     <lgl> <int>         <dbl>
## 1   FALSE 73200         0.543
## 2    TRUE 61688         0.457
```

P(nodrink|nosmoke)=45.7%. There are about 35% (54.3%x64.5%) of people who don’t smoke at all but drink alcohol and about 29.5%(45.7%x64.5%) of people who neither smoke at all nor drink alcohol.

```
q3%>%
  filter(nodrink=="TRUE")%>%
  group_by(smokday2)%>%
  summarise(count=n()) %>%
  mutate(share_nomoke=round(count/sum(count),3))
```

```
## # A tibble: 3 x 3
##     smokday2 count share_nomoke
##       <fctr> <int>        <dbl>
## 1  Every day 25764        0.267
## 2  Some days  8977        0.093
## 3 Not at all 61688        0.640
```

P(nosmoke|nodrink)=64%.

Summary

So it means that if a person don’t drink alcohol at all, he or she is likely to don’t smoke, which implies that drinking alcohol is more universal than smoking and people those don’t drink alcohol also tend to no smoke.