Exploring the BRFSS data

Setup

Load packages

```r
library(ggplot2)
library(dplyr)
library(tidyverse)
library(cowplot)
library(reshape2)
```

Load data

```r
load('brfss2013.Rdata')
```

Part 1: Data

This project aims to analyze the Behavioral Risk Factor Surveillance System (BRFSS) 2013 dataset through three concrete research questions (See research question part).
According to the BRFSS data user guide(https://www.cdc.gov/brfss/data_documentation/pdf/UserguideJune2013.pdf), BRFSS is "a cross-sectional telephone survey that state health departments conduct monthly over landline and cellular telephones with standardized questionnaires". **In other words, BRFSS 2013 is an observatory study to investigate the association among variables rather than the causal link**. The sample method is also detailed in the foregoing website. In brief, for landline telephone interviewers which accounted for ~80% of the questionnaires, a stratified strategy was applied based on the geographic division at multiple levels (termed "Disproportionate stratified sampling" in the document) At lowest geographic level (distinguished by phone area number), the interviewers were randomly selected from the population. For cellular interviewers which account for ~20% of the questionnaires, simple random sampling was conducted at the state's level. **In other words, the sampling of BRFSS 2013 is randomized and can be generalized to the population at large.**
In order to evaluate the non-response bias, the BRFSS protocol introduces a disposition code to score each interview attempt. The data quality report which includes response rates is available online (https://www.cdc.gov/brfss/annual_data/2013/pdf/2013_dqr.pdf).
As suggested by the data quality report, I quote the following claim:

> The response rate is the number of respondents who completed the survey as a proportion of all eligible and likely-eligible persons. The median survey response rate for all states, territories and Washington, DC, in 2013 was 46.4, and ranged from 29.0 to 60.3.For detailed information see the BRFSS Summary Data Quality Report.

Despite the low response rate, BRFSS applied a "raking" weight to scale the data based on demographic characteristics of respondents. The report claimed this ensured accurate estimates for most measures.
My project will be focused on the national-wide interpretation of BRFSS data in the U.S., so only only the data of 50 states will be included (i.e. data from the District of Columbia, Puerto Rico, and Guam are excluded). The analysis will be also focused on the data of main survey modules, since optional modules are not investigated by every state.

Part 2: Research questions

Research question 1: How is vegetarian diet correlated with diabetes condition? Diet is a well-recognized factor for diabetes. Vegetarian diet, which mainly consists of vegetable and fruit, is often recommended for diabetes patients. Hence, it is interesting to investigate if there is any association between vegetable/fruit consumption and diabetes condition within the BRFSS 2013 dataset.

Research question 2: How is inadequate sleep time correlated with heart attack?
Often times, there are reports claiming that sudden death of young people is associated with inadequate sleep time. Since most sudden death of young people are caused by heart attack, it is interesting to investigate whether the heart attack history is associated with inadequate sleep time in young people (age 18 ~ 34). To compare, the data from elder group (age 55 and above) will also be explored.

Research question 3: How is income level correlated with BMI categories (i.e. underweight, obese, etc.)?
During everyday talk, people often argue about the correlation between income level and BMI category. In the US, some people think that low income is associated with underweight, since people in poverty are struggling to get food; while others suggest that in developed country such as the US there is no real "poverty" characteristic of hunger, rather, low income often means obesity since "healthy food" like vegetables and fruit is generally much more expensive than "junk food" which contains much higher calories per unit of weight. It is thus interesting to look into the BRFSS 2013 data set to find out the association between these two indexes (BMI category and income level).

Part 3: Exploratory data analysis

Research question 1: How is vegetarian diet correlated with diabetes condition?

- Four variables are involved (the variable names used in table are indicated in parenthesis): States(*X_state*), Diabetes (*diabete3*), Total Vegetables Consumed Per Day (*X_vegesum*), Total Fruits Consumed Per Day (*X_frutsum*).

- **I will first select the variables of interest and filter out missing observations using *tidyverse* package; and then summarize the vegetable/fruit consumption indifferent groups using the *summary()* function; at last I will visualize the summary using box plot from the *ggplot2* package.**

Summary statistics

```
df1 <- brfss2013 %>%
  filter(X_state != 'Puerto Rico', X_state != 'Guam', X_state != 'District of Columbia', X_state != '0'
  select(diabete3, X_vegesum, X_frutsum) %>% # Select the variables of interest.
  drop_na() %>% # Remove missing observations.
  filter(diabete3 != 'Yes, but female told only during pregnancy', diabete3 != 'No, pre-diabetes or bor
# Below is for generating and printing a series of summary tables for different categories.
summary_1_1 <- summary(df1$X_vegesum[df1$diabete3 == 'Yes'])
summary_1_2 <- summary(df1$X_vegesum[df1$diabete3 == 'No'])
summary_1_3 <- summary(df1$X_frutsum[df1$diabete3 == 'Yes'])
summary_1_4 <- summary(df1$X_frutsum[df1$diabete3 == 'No'])
cat('Summary of vegetable consumption of those who in diabetes condition', sep = '\n')
print(summary_1_1)
cat('Summary of vegetable consumption of those who in non-diabetes condition', sep = '\n')
print(summary_1_2)
cat("Summary of fruit consumption of those who in diabetes condition", sep = '\n')
```

```r
print(summary_1_3)
cat("Summary of fruit consumption of those who in non-diabetes condition", sep = '\n')
print(summary_1_4)
```

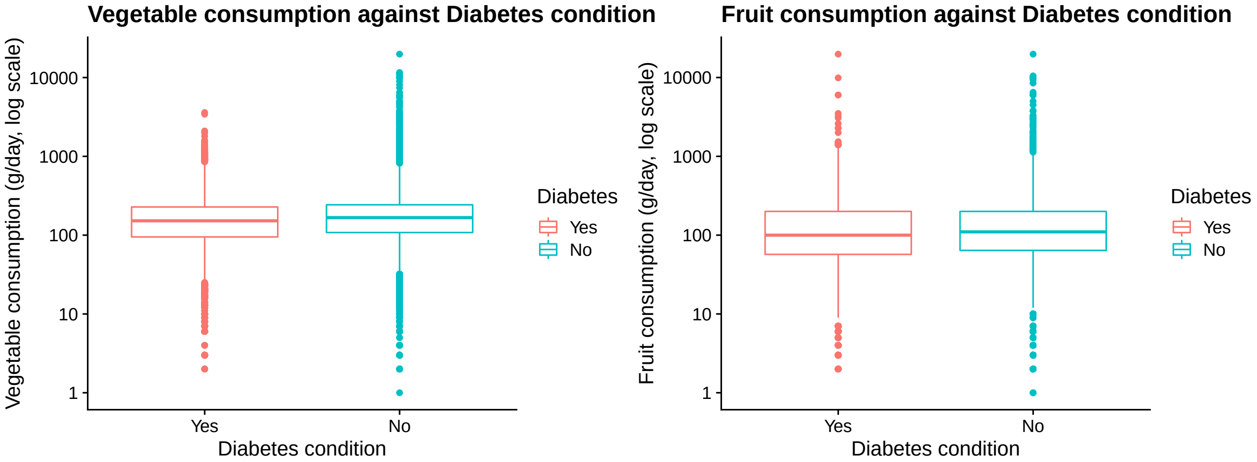
```
## Summary of vegetable consumption of those who in diabetes condition
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    93.0   150.0   175.1   228.0  3600.0
## Summary of vegetable consumption of those who in non-diabetes condition
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   107.0   167.0   192.6   243.0 19827.0
## Summary of fruit consumption of those who in diabetes condition
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    47.0   100.0   130.2   200.0 19800.0
## Summary of fruit consumption of those who in non-diabetes condition
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    57.0   104.0   141.4   200.0 19800.0
```

In the comparison of vegetable consumptions, diabetes patients tended to consume less than those who were in non-diabetes conditions (mean: 174.8 vs. 192.6; median: 150.0 vs. 167.0).
In the comparison of fruit consumptions, diabetes patients tended to consume slightly less than those who were in non-diabetes conditions (mean: 130.0 vs. 141.3; median: 100.0 vs. 104.0).
In other words, non-diabetes conditions were more likely to be associated with higher consumptions of vegetable and fruit.

It is worthy noting that there are some extreme observations which consume 100 times higher vegetables/fruit than the average; hence the median may be a better indicator. Box plots will be used to visualize the the vegetable/fruit consumptions against different diabetes conditions.

Data visualization (Box plot)

```r
gg1 <- ggplot(df1, aes(x = diabete3, y = X_vegesum, colour=diabete3)) +
  geom_boxplot() +
  scale_y_log10() +
  ggtitle("Vegetable consumption against Diabetes condition") +
  labs(x = "Diabetes condition", y = "Vegetable consumption (g/day, log scale)", colour = "Diabetes") +
  theme_cowplot()
gg2 <- ggplot(df1, aes(x = diabete3, y = X_frutsum, colour=diabete3)) +
  geom_boxplot() +
  scale_y_log10() +
  ggtitle("Fruit consumption against Diabetes condition") +
  labs(x = "Diabetes condition", y = "Fruit consumption (g/day, log scale)", colour = "Diabetes") +
  theme_cowplot()
plot_grid(gg1, gg2)
```



The box plot visualization is consistent with the statistic summary; in the US, people in non-diabetes condition consume more vegetables or fruit (On average, 10% more). Although this means only the correlation not the causality, we have a good evidence to support the recommendation of vegan diet for diabetes patients.

Research question 2: How is inadequate sleep time correlated with heart attack?

- Five variables are involved (the variable names used in table are indicated in parenthesis): States(*X_state*), Ages (*X_age_g*), Ever Diagnosed with Heart Attack (*cvdinfr4*), How Much Time Do You Sleep (*sleptim1*).

- **I will first select the variables of interest and filter out missing observations using *tidyverse* package; and then summarize the sleep time in different groups using the *summary()* function; at last I will visualize the summary using box plot from the *ggplot2* package.**

Summary statistics

```
young <- brfss2013 %>%
  filter(X_state != 'Puerto Rico', X_state != 'Guam', X_state != 'District of Columbia', X_state != '0'
  select(X_age_g, cvdinfr4, sleptim1) %>% # Select the variables of interest.
  filter(X_age_g == "Age 18 to 24" | X_age_g == "Age 25 to 34") %>% # Remove observations that are not
  drop_na() # Remove missing observations.
elder <- brfss2013 %>%
  filter(X_state != 'Puerto Rico', X_state != 'Guam', X_state != 'District of Columbia', X_state != '0'
  select(X_age_g, cvdinfr4, sleptim1) %>% # Select the variables of interest.
  filter(X_age_g == "Age 55 to 64" | X_age_g == "Age 65 or older") %>% # Remove observations that are n
  drop_na() # Remove missing observations.
# Below is for generating and printing a series of summary tables for different categories.
summary_2_1 <- summary(young$sleptim1[young$cvdinfr4 == 'Yes'])
summary_2_2 <- summary(young$sleptim1[young$cvdinfr4 == 'No'])
summary_2_3 <- summary(elder$sleptim1[elder$cvdinfr4 == 'Yes'])
summary_2_4 <- summary(elder$sleptim1[elder$cvdinfr4 == 'No'])
cat('Summary of the sleep time of young people who was diagnosed heart attack', sep = '\n')
print(summary_2_1)
cat('Summary of the sleep time of young people who was NOT diagnosed heart attack', sep = '\n')
print(summary_2_2)
cat("Summary of the sleep time of elder people who was diagnosed heart attack", sep = '\n')
print(summary_2_3)
cat("Summary of the sleep time of elder people who was NOT diagnosed heart attack", sep = '\n')
print(summary_2_4)
```

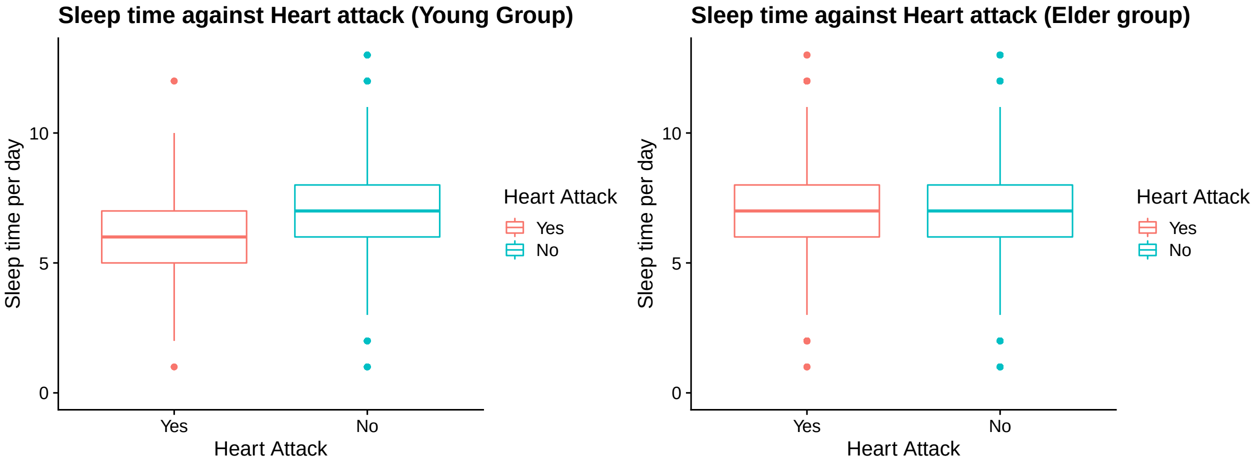
```
## Summary of the sleep time of young people who was diagnosed heart attack
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   6.000   6.195   7.000  18.000
## Summary of the sleep time of young people who was NOT diagnosed heart attack
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.000   7.000   6.969   8.000  24.000
## Summary of the sleep time of elder people who was diagnosed heart attack
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.000   7.000   7.164   8.000  24.000
## Summary of the sleep time of elder people who was NOT diagnosed heart attack
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.000   7.000   7.188   8.000  24.000
```

In young group, people who was diagnosed heart attack showed ~1 hour less sleep time than those who were not (mean: 6.195 vs. 6.969; median: 6.0 vs. 7.0).
In elder group, people who was diagnosed heart attack showed negligibly less sleep time (~1.5 minutes) than those who were not (mean: 7.164 vs. 7.188; median: 7.0 vs. 7.0).
In other words, inadequate sleep in young people is more likely to be associated with heart attack.

It is worthy noting that there are some extreme observations who sleep 24 everyday (likely fake reporting); hence the median may be a better indicator. Box plots will be used to visualize the the vegetable/fruit consumptions against different diabetes conditions.

Data visualization (box plot)

```r
young_plot <- ggplot(young, aes(x = cvdinfr4, y = sleptim1, color = cvdinfr4)) +
  geom_boxplot() +
  ylim(0, 13) +
  ggtitle("Sleep time against Heart attack (Young Group)") +
  labs(x = "Heart Attack", y = "Sleep time per day", colour = "Heart Attack") +
  theme_cowplot()
elder_plot <- ggplot(elder, aes(x = cvdinfr4, y = sleptim1, color = cvdinfr4)) +
  geom_boxplot() +
  ylim(0, 13) +
  ggtitle("Sleep time against Heart attack (Elder group)") +
  labs(x = "Heart Attack", y = "Sleep time per day", colour = "Heart Attack") +
  theme_cowplot()
plot_grid(young_plot, elder_plot)
```



The box plot visualization is consistent with the statistic summary; in the US, the inadequate sleep time of young people is more likely to be associated with heart attack than that of the elder group. Although this means only the correlation not the causality, we have a good evidence to ask young people to sleep well.

Research question 3: How is income level correlated with BMI categories (i.e. underweight, obese, etc.)?

- Three variables are involved (the variable names used in table are indicated in parenthesis): States(*X_state*), Computed Income Categories (*X_incomg*), Computed Body Mass Index Categories (*X_bmi5cat*).
- **I will first select the variables of interest and filter out missing observations using *tidyverse* package; and then summarize the contingency table (BMI category vs. Income level) using the *table()* function ; at last I will visualize the summary using stacked bar plot from the *ggplot2* package.**

Summary statistics

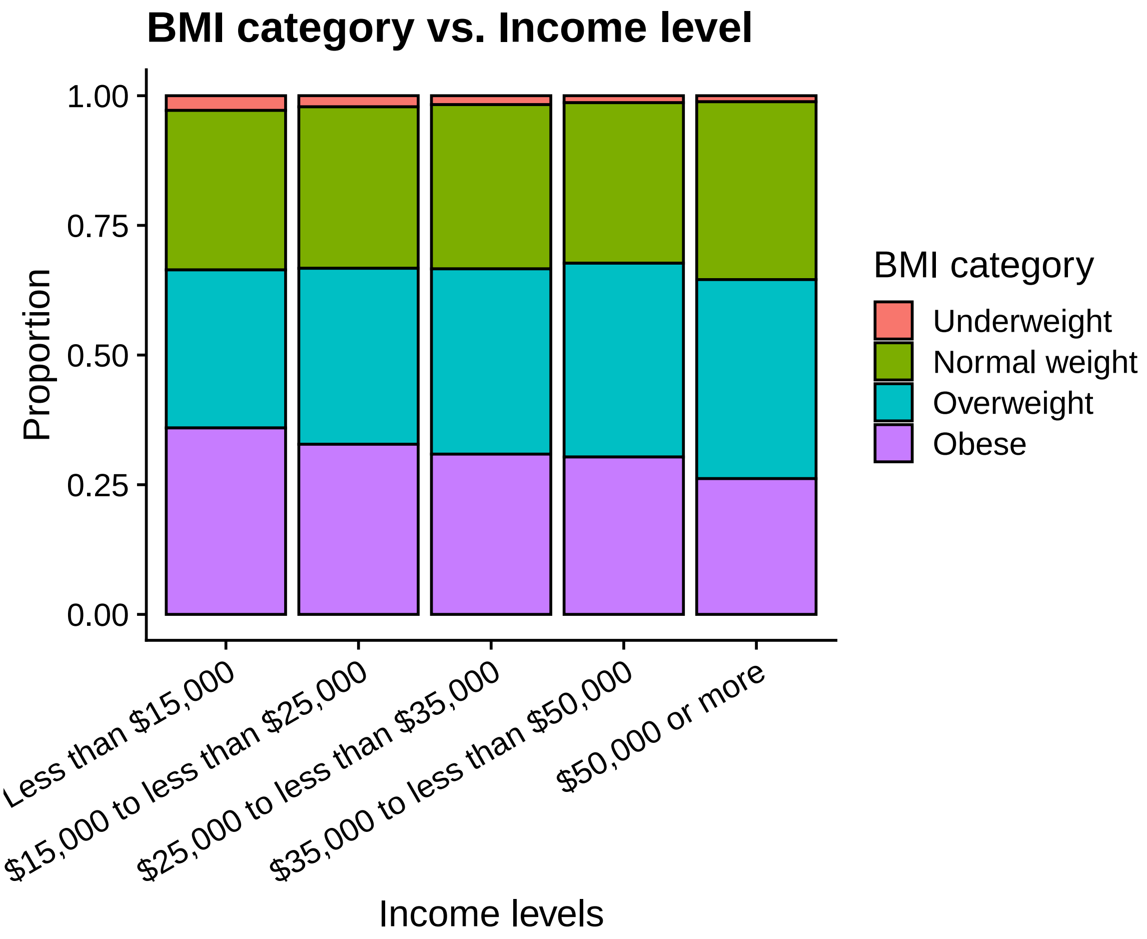
```r
moneyweight_df <- brfss2013 %>%
  filter(X_state != 'Puerto Rico', X_state != 'Guam', X_state != 'District of Columbia', X_state != '0'
  select(X_incomg, X_bmi5cat) %>% # Select the variables of interest.
  drop_na() # Remove missing observations.
moneyweight_table <- table(moneyweight_df$X_incomg, moneyweight_df$X_bmi5cat) # Make a contingency tabl
data <- as.data.frame(t(apply(moneyweight_table, 1, function(i) i*100/sum(i)))) # Transform numbers in
cat('Percentages of BMI categories in different income groups', sep = '\n')
print(data)
```

```
## Percentages of BMI categories in different income groups
##                               Underweight Normal weight Overweight    Obese
## Less than $15,000                2.836132      30.75021   30.46853 35.94513
## $15,000 to less than $25,000     2.148382      31.13178   33.92242 32.79741
## $25,000 to less than $35,000     1.728465      31.66016   35.71507 30.89631
## $35,000 to less than $50,000     1.348176      30.94801   37.35957 30.34425
## $50,000 or more                  1.167766      34.28551   38.36920 26.17752
```

From the distribution of BMI categories in each income group, we can see that as people get richer, the underweight rates decreases (2.8% > 2.1% > 1.7% > 1.3% > 1.2%); interestingly, the obesity rates also decrease (35.9% > 32.8% > 30.9% > 30.3% > 26.2%). This means the aforementioned claims (low income is associated with underweight vs. obesity) have certain grounds regardless of the causality.

Data visualization (stacked bar plot)

```r
data2<- melt(cbind(income_level= rownames(data),data), id = 'income_level') # Reshape the data frame to
data2$income_level <- factor(data2$income_level, levels = c('Less than $15,000','$15,000 to less than $2
ggplot(data2, aes(x=income_level, y=value, fill=variable)) +
  geom_bar(position="fill", stat="identity", colour = 'black') +
  ggtitle("BMI category vs. Income level") +
  labs(x = "Income levels", y = "Proportion", fill = "BMI category", colour = "BMI category") +
  theme_cowplot() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```



The visualization of stacked bar plot is consistent with the statistic summary; in the US, as people get richer, both the underweight rates and obesity rates decrease. In other words, high income is more likely to be associated with standard BMI.