Exploring the BRFSS data

Setup

Load packages

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

Load data

```
load("brfss2013.RData")
```

Part 1: Data

The observations in the sample were collected randomly through both landline and cellular survey by The Behavioral Risk Factor Surveillance System (BRFSS). The landline telephone survey collected data from randomly selected adult in a household and the cellular telephone survey from an adult who resides in a private residence or college housing.

The result of the study can be generalized to the non-institutionalized adult population, aged 18 years or older, who reside in the US and who have a landline telephone or cellular telephone. The data set was collected using random sampling.

This is an observational study because the study does not make use of random assignment. Thus, it is only allowed to make correlation statement. Since the study is not based on an experiment, there will be no causal connection between the variables. In other words, causation cannot be implied in this study.

Part 2: Research questions

Research question 1: Although health care insurance is important, not everyone has enough money to afford it. Moreover, women are more likely to earn less money than men, and they are also more likely to use health care insurance more often than men. Thus, my question is: Is it always the case that more women have health coverage than men regardless of their income level?

Research question 2: It seems to me that the majority of people from U.S sleep 6 hours. Therefore, I am curious about how much time American people sleep, on average, in a 24-hour period and how the data is distributed. The size of the sample is 484386.

Research question 3: Since I life in Minnesota, I would like to know what is the relationship between Minnesotans general health and the quantity of hours of sleep they get in a 24-hour period. For this analysis, I will assume that people who sleep less than 5 hours sleep bad, otherwise they sleep well.

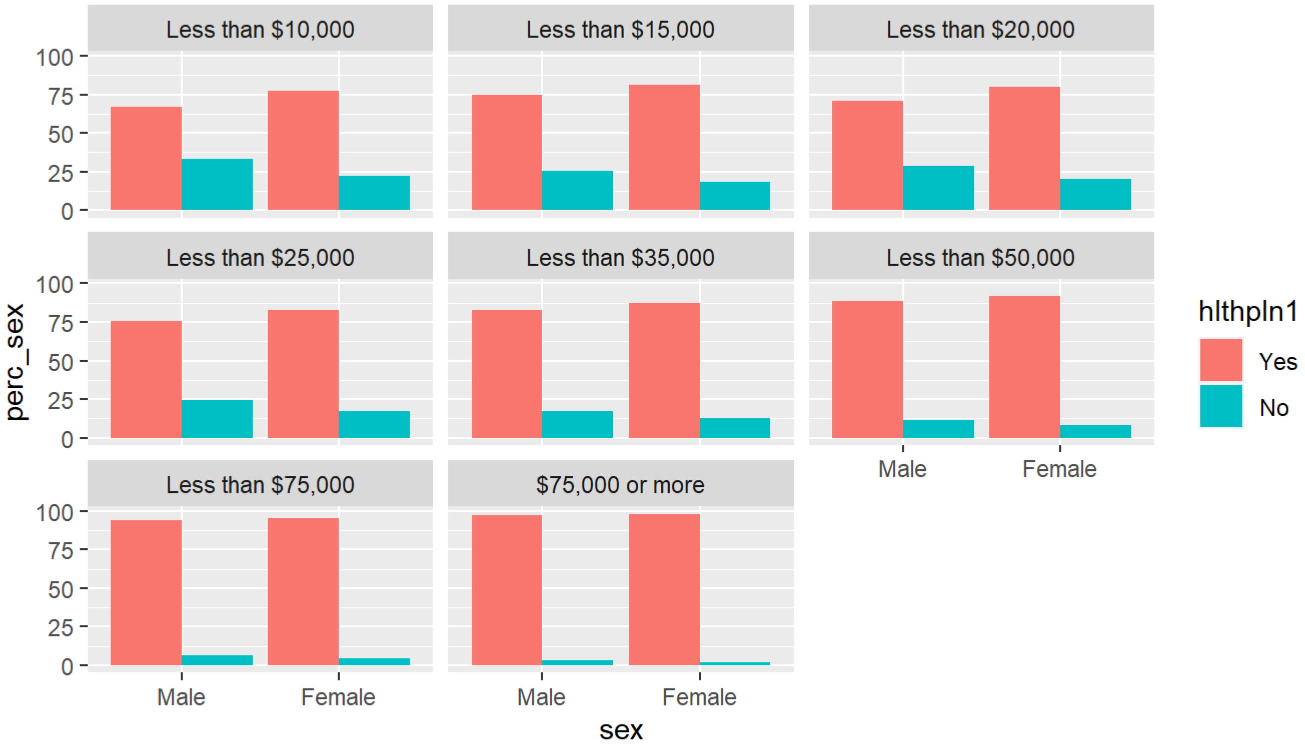
Part 3: Exploratory data analysis

Research quesion 1:

```
Income_level <- brfss2013 %>%
  filter(!is.na(sex), !is.na(income2), !is.na(hlthpln1)) %>%
  group_by(sex, income2, hlthpln1) %>%
  summarise(count = n())
```

```
## `summarise()` regrouping output by 'sex', 'income2' (override with `.groups` argument)
```

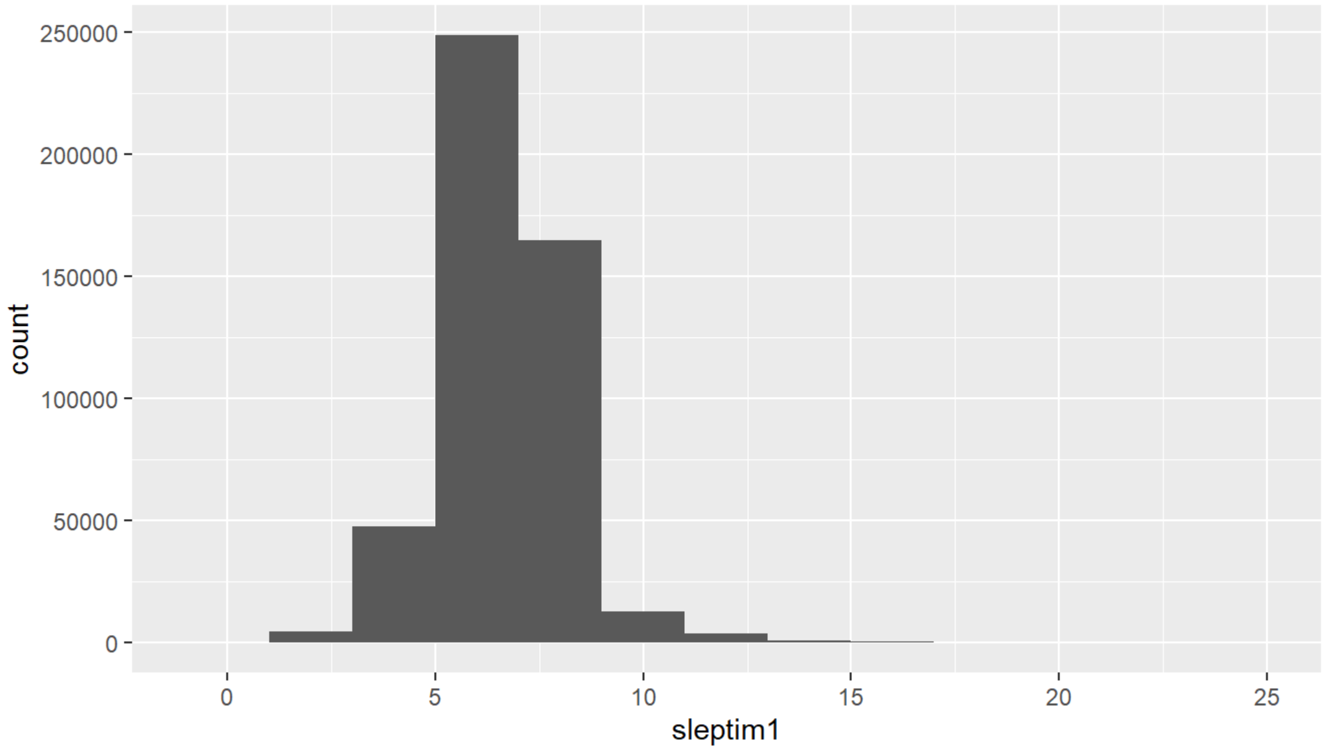
```
count_sex = aggregate(count ~ income2 + sex, Income_level, FUN  = sum)

Income_level2 = merge(Income_level, count_sex, by = c("income2", "sex"), suffixes = c("", "_tot"))
Income_level2$perc_sex= (Income_level2$count/Income_level2$count_tot)*100


ggplot(data = Income_level2, aes(y = perc_sex, x = sex, fill = hlthpln1))  + facet_wrap( ~ income2) +
  geom_bar(position = "dodge", stat = "identity")
```



Answer to research question 1: According to the analysis, it is always the case that more women have health care coverage than men regardless of their income level. However, the different between the percentage of women that have health insurance compared to men decreases as income level increases. This is because most people with higher income level tend to have health insurance.

Research question 2:

```
sleptime <- brfss2013 %>%
  filter(sleptim1<25)
ggplot(data = sleptime, aes(x = sleptim1)) +
         geom_histogram(binwidth = 2)
```



```
sleptime %>%
  summarise(mean_sl = mean(sleptim1), sd_sl = sd(sleptim1), n = n())
```

```
##    mean_sl    sd_sl      n
## 1 7.050986 1.465987 484386
```

Answer to research question 2: It seems like the variable called “Sleptim1” follow a normal distribution. On average, American people sleep 7 hours in a 24-hour period. Additionally, the data has aprox. a standard deviation of 1.46 in a sample of 484386 obs.

Research question 3:

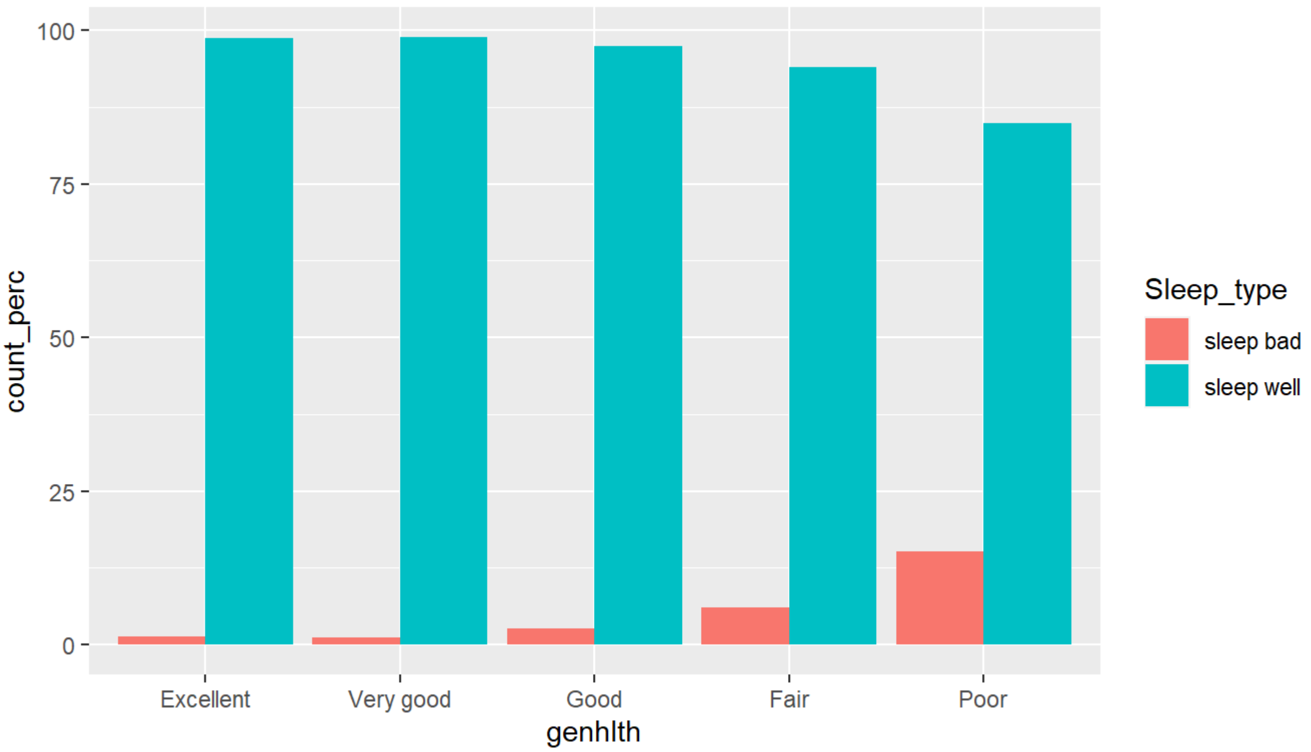
```
MN_general_health <- brfss2013 %>%
  filter(X_state == "Minnesota")
MN_general_health %>%
  filter(!is.na(genhlth)) %>%
  group_by(genhlth) %>%
  summarise(count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
## # A tibble: 5 x 2
##   genhlth   count
##   <fct>     <int>
## 1 Excellent  2907
## 2 Very good  5381
## 3 Good       4107
## 4 Fair       1403
## 5 Poor        501
```

```
MN_general_health <- MN_general_health %>%
  mutate(Sleep_type = ifelse(sleptim1 < 5, "sleep bad", "sleep well"))
df_count = MN_general_health %>%
  filter(!is.na(Sleep_type))%>%
  filter(!is.na(genhlth)) %>%
  group_by(genhlth, Sleep_type) %>%
  summarise(count = n())
```

```
## `summarise()` regrouping output by 'genhlth' (override with `.groups` argument)
```

```
df_count_totals = aggregate(count ~ genhlth, df_count, FUN = sum)
df_count2 = merge(df_count, df_count_totals, by = "genhlth", suffixes = c("", "_tot"))

df_count2$count_perc = (df_count2$count/df_count2$count_tot)*100
df_count3 = df_count2[order(df_count2$genhlth),]

ggplot(data = df_count3, aes(y = count_perc, x = genhlth, fill = Sleep_type)) +
  geom_bar(position = "dodge", stat = "identity")
```



Answer to research question 3: In general, people from Minnesota sleep well. However, people with poorer health tend to sleep less.