Exploring the BRFSS data

Setup

Load packages

```
# ggplot to plot figures out and dplyr to do the calculation
library(ggplot2)
library(dplyr)
```

Load data

This data is loaded from BRFSS website and we are going to use this dataset to finish this project.

```
# Load the data from local
load("brfss2013.RData")
```

Part 1: Data

Data Introduction:
The Behavioral Risk Factor Surveillance System (BRFSS) is a United States health survey that looks at behavioral risk factors. Begun in 1984, the BRFSS is run by Centers for Disease Control and Prevention and conducted by participating individual state health departments. The survey is administered by telephone and is the world's largest such survey. In 2009, the BRFSS began conducting surveys by cellular phone in addition to traditional "landline" telephones. Individual states can add their own questions to the survey instrument, which consists of a core set of questions on certain topics like car safety, obesity, or exercise. States get funding from the federal government to administer these questionnaires, and they pay for the additional questions themselves.

Generalizability:
Since this survey is conducted by phone call, so people that cannot be reached will not be covered as samples in this data set. But this dataset is still randomly sampled, no obvious bias shows up. So it could be generalized to peopled in all states living in the US.

Causality:
It cannot be causal because the data collector just collect those data for researchers, like us, to analyze the data. Researchers didn't do any random assignment around each observation they are using. It is apparently an observational experiment. So no casual result could this dataset bring to us in this project only. Correlated summaries are going to be more often to appear.

Part 2: Research questions

Research question 1: Check out the relationship between people's general health condition (genhlth) and the time they sleep (sleptim1). See if there's some positive or negative correlated factor between sleeping time and health condition.

There are so many people are sleeping less than the time they should have and still feeling fine. But actually their bodies are right at the red line. This result might have some value to show people the importance of adequate sleeping.

Research question 2: Do people's heart health (cvdinfr4) have some relationship with their sleeping time? How do heart attack affect their general health?

Inadequate sleeping time might have something to do with heart attack. This result might have some value to show people worried about their general health and people who sleep a little and feeling not so good about their heart health condition.

Research question 3: Despite the physical health condition, does inadequate sleeping have anything to do with mental disorders? What should people take attention to depressive disorder in this topic?

Mental and physical health condition are both importance to people. This result might have some value on researchers who care about the result of inadequate sleeping time and depressive.

Part 3: Exploratory data analysis

Research question 1: By using data "genhlth" and "sleptim1" to analyze the relationship between sleeping time and health general condition, the process is shown below.

```
# Select out only genlth and sleptim1 from brfss2013 data set and omit the NA values
health_sleep <- brfss2013 %>%
  select(genhlth, sleptim1) %>%
  na.omit()
```

We could group by the and count the quantity of each kind of health conditions and sleeping time in this data set. result shown in the chunk and table below.

```
# "sleep" represents each sleptim1 and the total quantity of each kind
sleep <- health_sleep %>%
  group_by(sleptim1)%>%
  summarise(slep_count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# "health" represents each genhelth and the total quantity of each kind
health <- health_sleep %>%
  group_by(genhlth)%>%
  summarise(hel_count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
# Mutate those two together
health_sleep_mutate <- health_sleep %>%
  group_by(genhlth, sleptim1) %>%
    summarise(hel_slep_count = n())
```

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

```
health_sleep_mutate
```

```
## # A tibble: 115 x 3
## # Groups:   genhlth [5]
##    genhlth   sleptim1 hel_slep_count
##    <fct>        <int>          <int>
##  1 Excellent        0              1
##  2 Excellent        1             26
##  3 Excellent        2             72
##  4 Excellent        3            254
##  5 Excellent        4           1212
##  6 Excellent        5           3661
##  7 Excellent        6          16013
##  8 Excellent        7          28824
##  9 Excellent        8          28478
## 10 Excellent        9           4260
## # … with 105 more rows
```

```
sleep
```

```
## # A tibble: 25 x 2
##    sleptim1 slep_count
##       <int>      <int>
##  1        0          1
##  2        1        228
##  3        2       1063
##  4        3       3466
##  5        4      14194
##  6        5      33290
##  7        6     105880
##  8        7     142090
##  9        8     140498
## 10        9      23688
## # … with 15 more rows
```

```
health
```

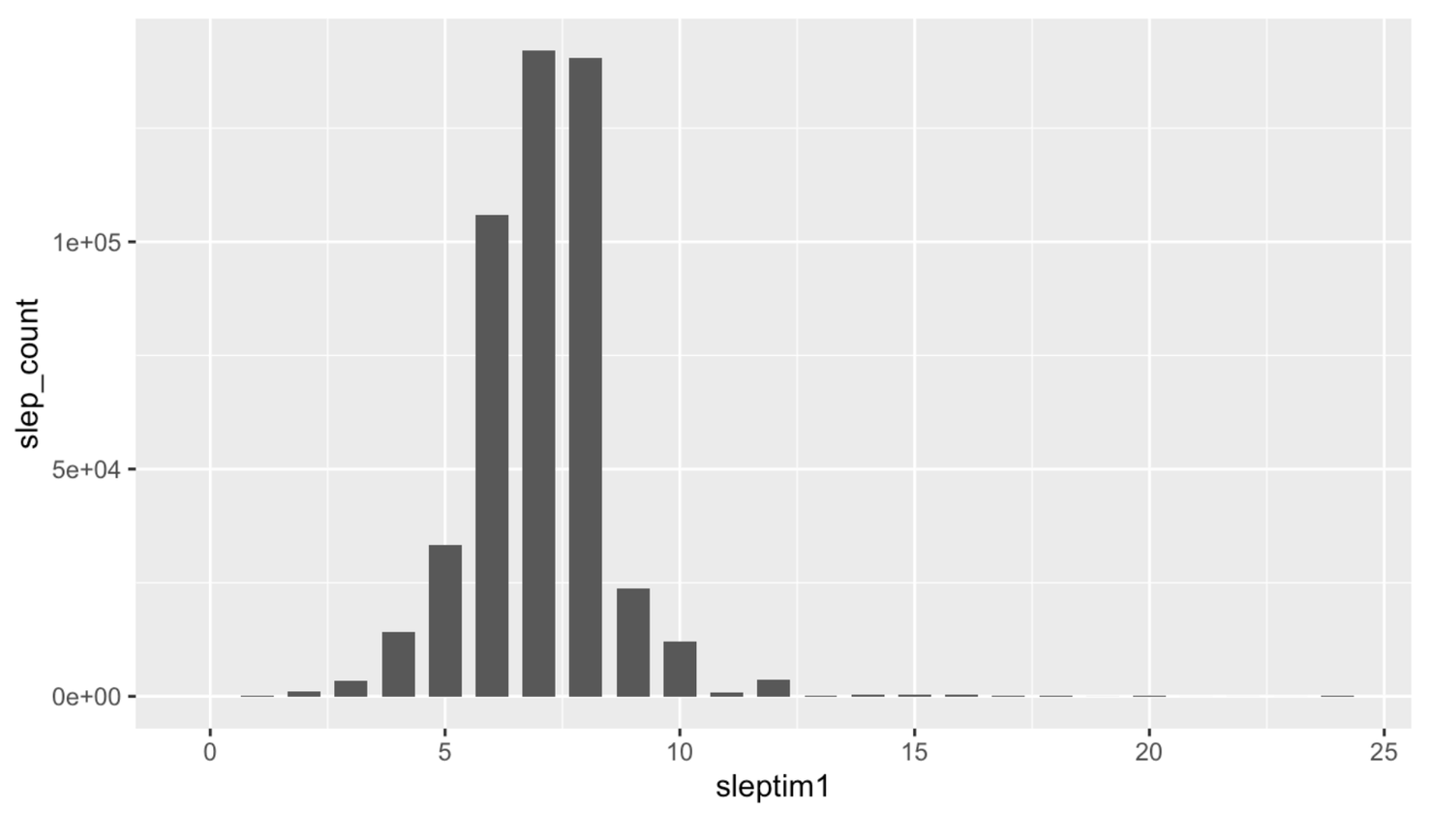
```
## # A tibble: 5 x 2
##   genhlth   hel_count
##   <fct>         <int>
## 1 Excellent     84822
## 2 Very good    157833
## 3 Good         148299
## 4 Fair          65012
## 5 Poor          26639
```

Here we use columns plots to show the sleeping time distribution and health condition numbers. From here we could tell, people are mostly at "Very Good" and "Good" condition and averaging 6-8 hours sleep per day.

```
cp_sleep <- ggplot(data = sleep, aes(x = sleptim1, y = slep_count)) + geom_col(stat = "count", width = 0.7)
```

```
## Warning: Ignoring unknown parameters: stat
```

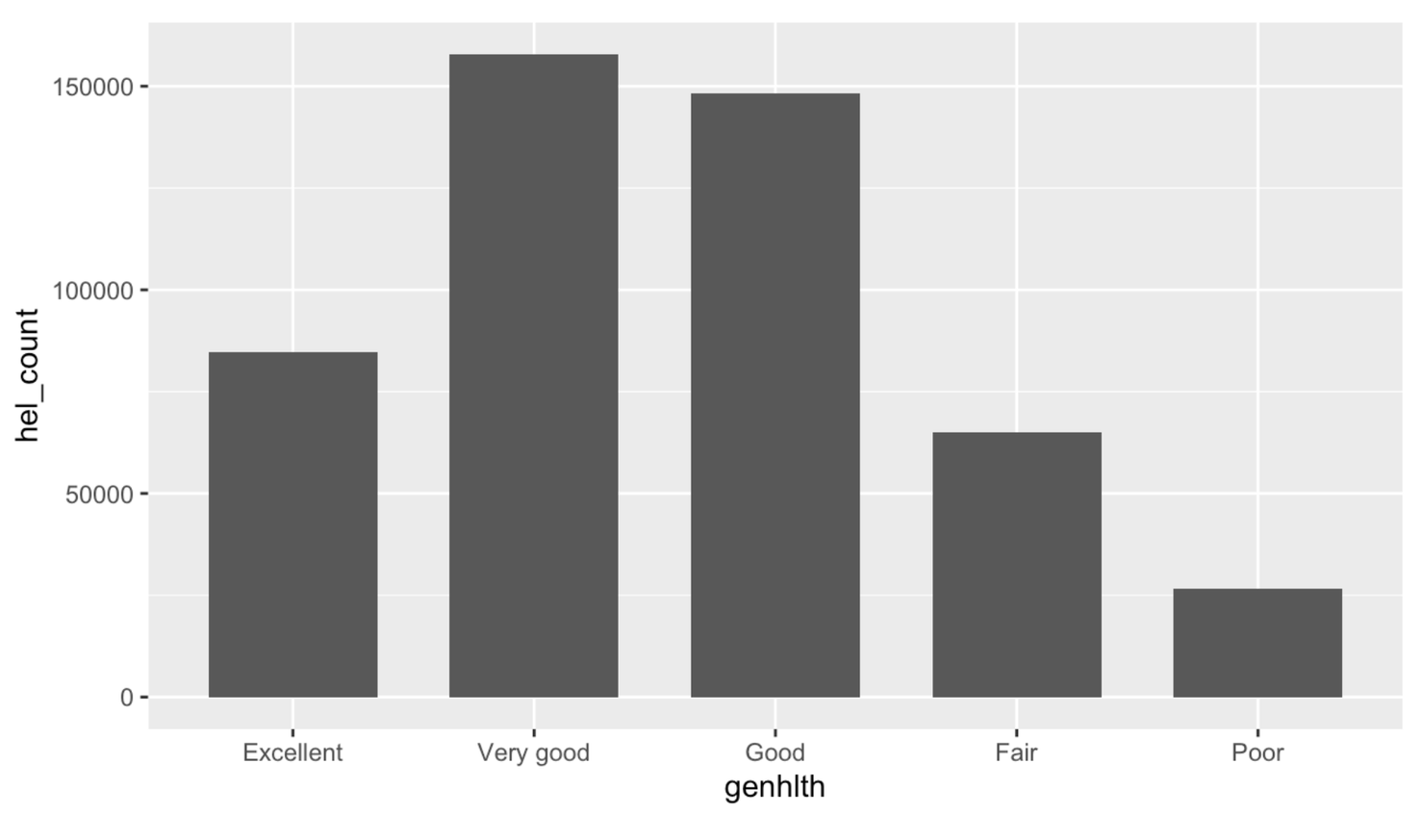
```
cp_sleep
```




```
cp_health <- ggplot(data = health, aes(x = genhlth, y = hel_count)) + geom_col(stat = "count", width = 0.7)
```

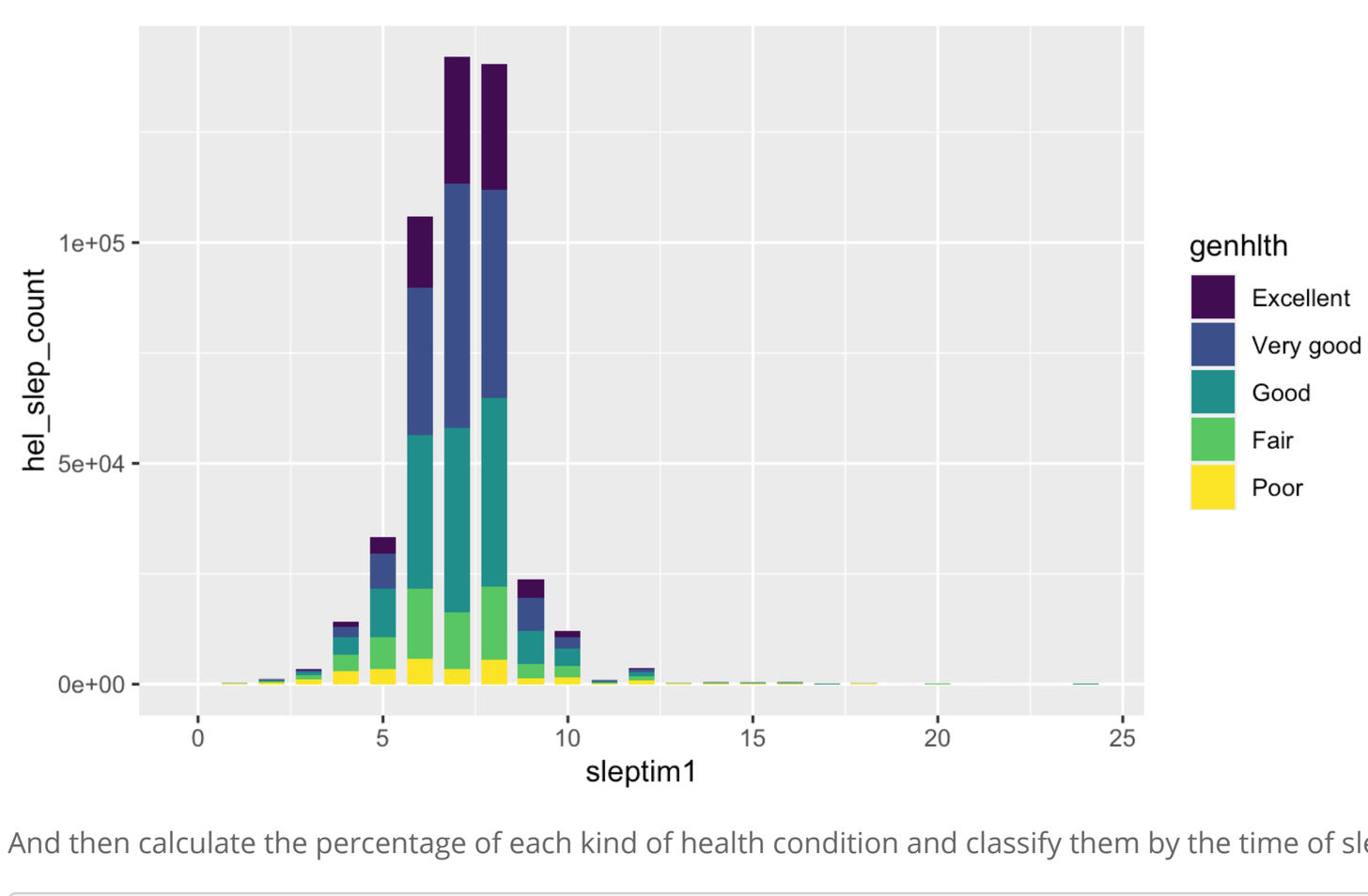
```
## Warning: Ignoring unknown parameters: stat
```

```
cp_health
```



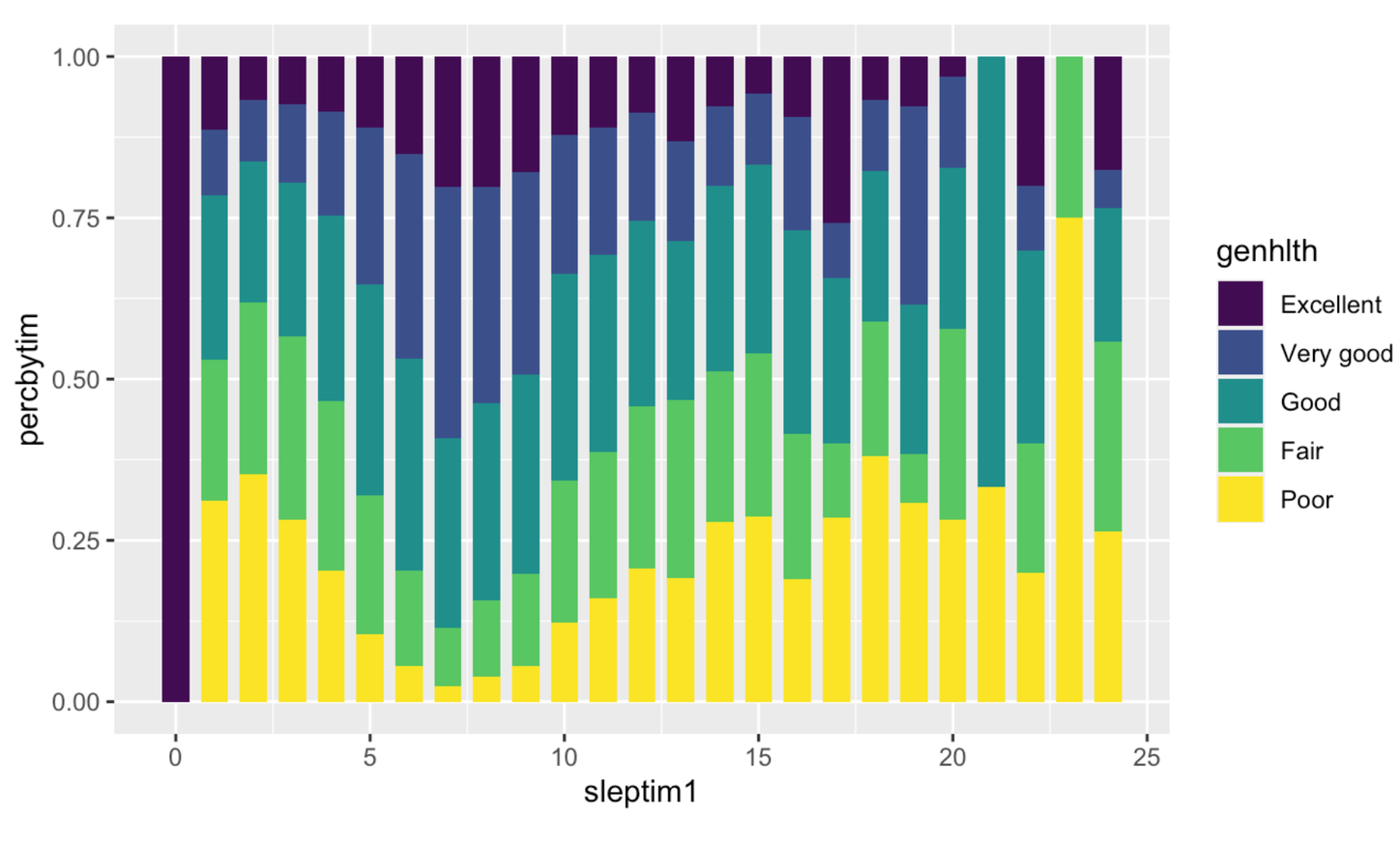
Now we want to see the percentage of each kind of health condition with how many hours they sleep every night. To do this, we have to combine these two plots together.

```
# Used for calculate the quantity of each "sleptim1" in order to be divied by "genhlth" numbers to calculate the pe
health_sleep_mutate$total_slep_count <- ifelse(health_sleep_mutate$sleptim1 == "0",1, ifelse(health_sleep_mutate$sle

# Calculate the percentage
health_sleep_mutate$percbytim <- health_sleep_mutate$hel_slep_count / health_sleep_mutate$total_slep_count

# Plot the figure
ggplot(health_sleep_mutate, aes(x = sleptim1, y = hel_slep_count))+geom_col(aes(fill = genhlth), width = 0.7)+scale_
```



And then calculate the percentage of each kind of health condition and classify them by the time of sleeping.

```
ggplot(health_sleep_mutate, aes(x = sleptim1, y = percbytim)) +
  geom_col(aes(fill = genhlth), width = 0.7) +
  scale_fill_viridis_d()
```



To make a brief summary here, people under this investigation are fairly healthy and most people are sleeping for 6-8 hours per day. However, there are still some observations that representing those who sleep less than 5 hours at night and also those cannot even wake up. The plots tell us that people who sleep around 6-8 hours every day holds the health conditions of "Excellence" and "Very Good" most, and barely be considered "Poor" in health. Less than 5-hour-sleep might have negative effects on your body. Too much sleep also might cause bad health situations as well.

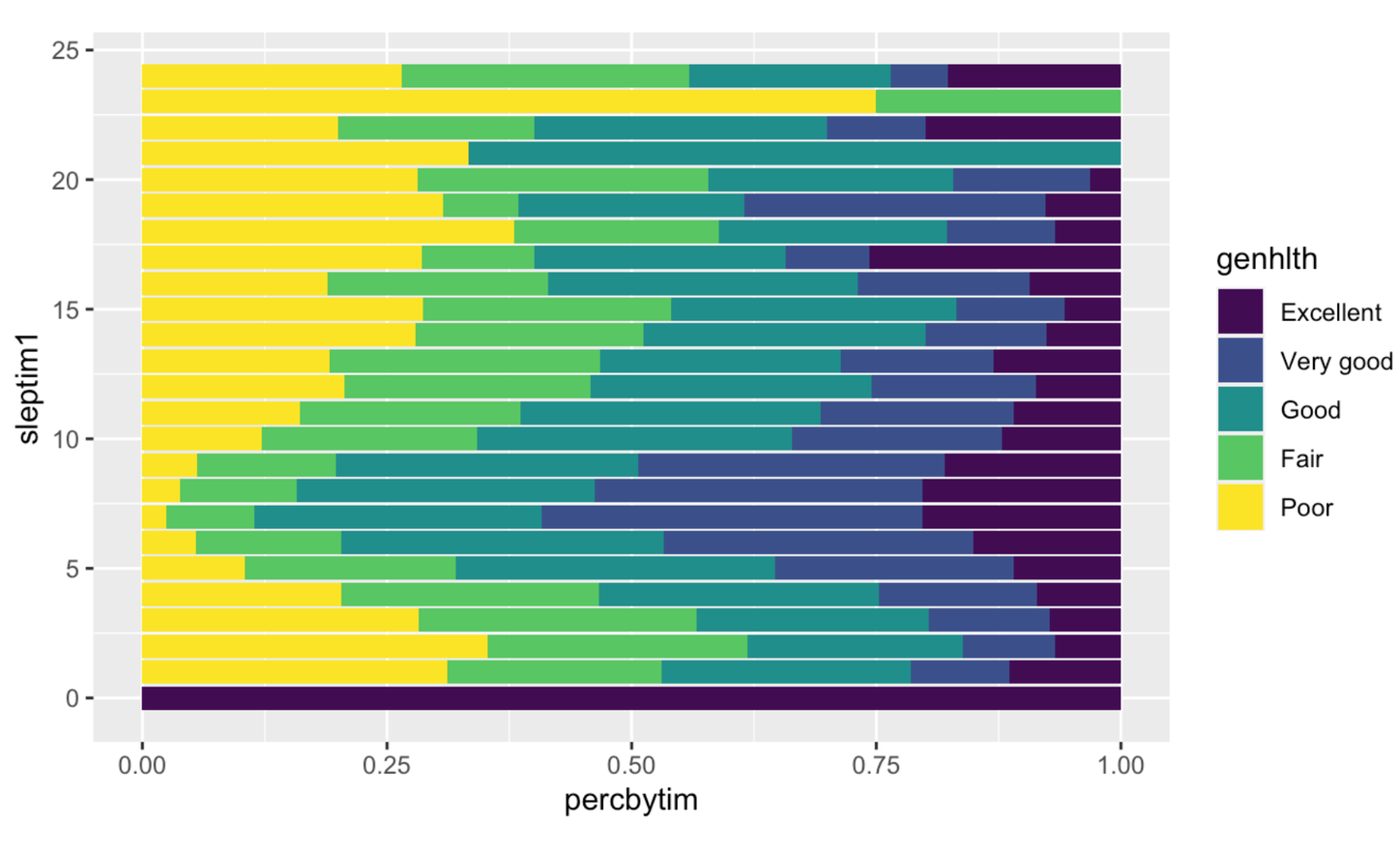
Research question 2:

Now add heart attack condition (cvdinfr4) into our discussion. "cvdinfr4" represents the data people ever diagnosed with heart attack. We make it visualized with classifying the data by "Yes" and "No" data.

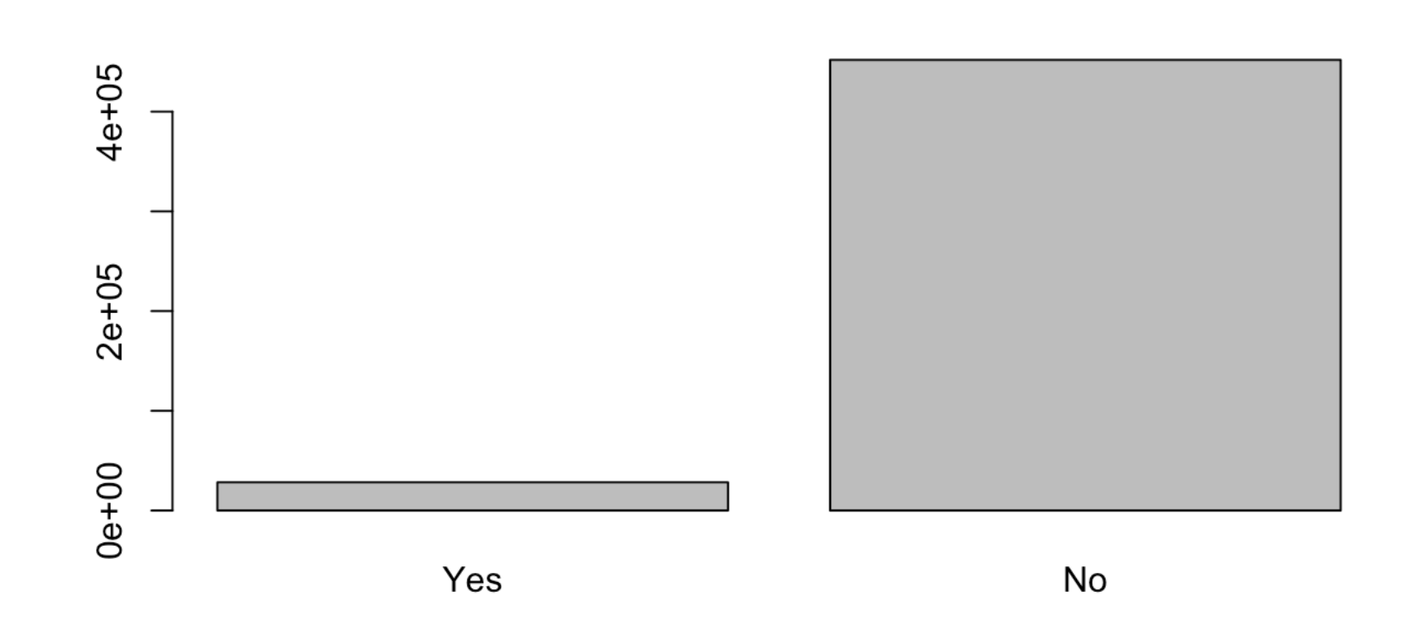
```
# Select out data that shows whether people have been diagnosed with heart attack problems
# Omit the NA values
health_sleep_heart_mutate <- brfss2013%>%
  select(genhlth,sleptim1,cvdinfr4)%>%
  na.omit()
```

Now to understand the figure more, I need to take down the bar chart made before, which shows the proportion of each kind of health condition with each sleeping time per day. And separate those data into tow columns: Did diagnosed heart attack & did not diagnosed heart attack.

```
# Turn over the chart of sleeping time and health condition
ggplot(health_sleep_mutate, aes(x = sleptim1, y = percbytim)) +
  geom_col(aes(fill = genhlth), width = 0.9) +
  scale_fill_viridis_d() +
  coord_flip()
```




```
# Plot and show the result
plot(health_sleep_heart_mutate$cvdinfr4)
```



Now we could plot the chart that shows the different result of whether have been diagnosed or not, with the reuglar health_sleep data set up before.

```
ggplot(health_sleep_heart_mutate, aes(x = sleptim1, fill = cvdinfr4)) +
  geom_bar(position = "fill") +
  facet_wrap( ~ genhlth, ncol = 5) +
  coord_flip()
```

This diagram tells us that people with poor health condition originally, may have a lot more opportunities to be diagnosed heart attack. And even people with excellent body, if they spend to much time on bed, heart attack might come up to them too. At last, no matter how good the body is, people stay up late and sleep less than 5 hours a day, need to check their body more often, maybe heart attack has already come up.

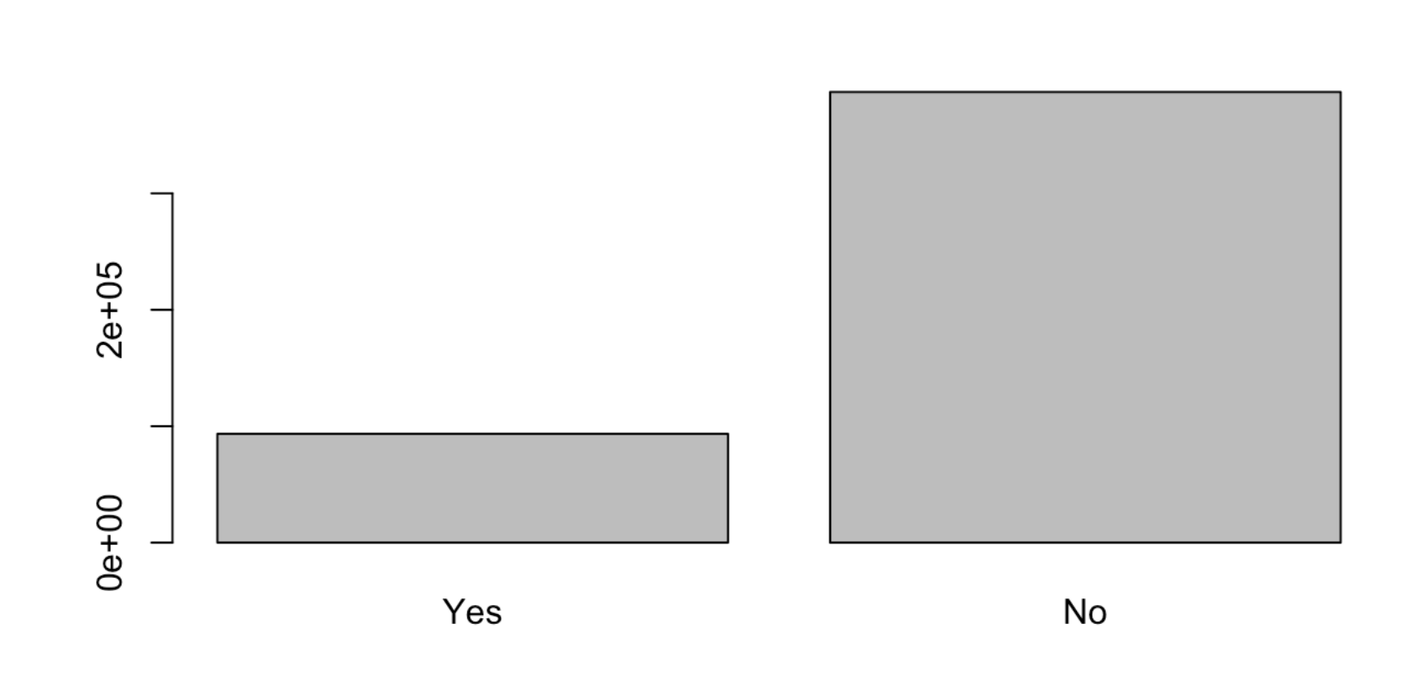
Research question 3:

Now add depressive disorder conditions (addepev2) into our discussion. "addepev2" represents the data people ever diagnosed with depressive disorder. We make it visualized with classifying the data by "Yes" and "No" data.

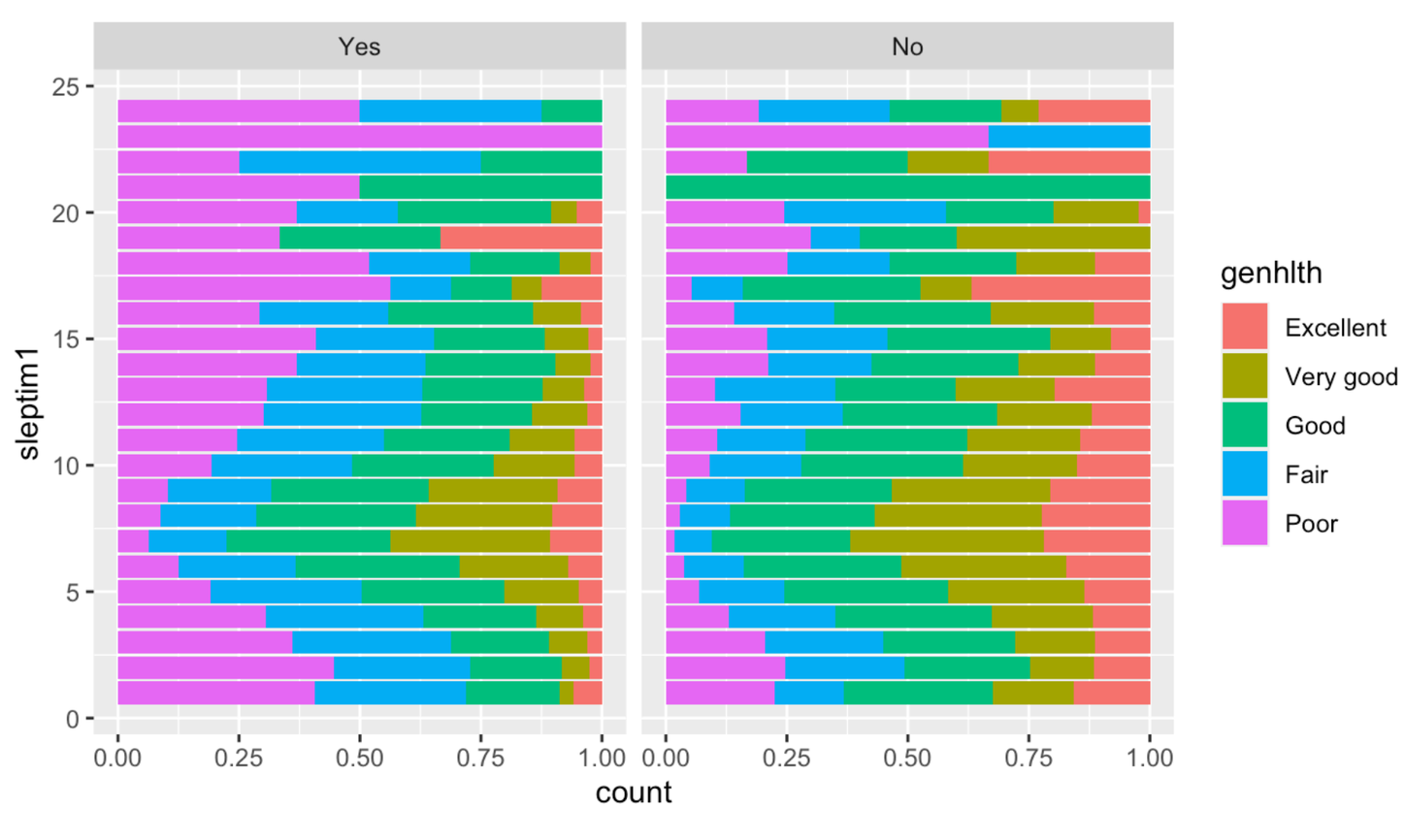
```
# Select out data that shows whether people have been diagnosed with depressive disorder problems
# Omit the NA values
health_sleep_depressive_mutate <- brfss2013%>%
  select(genhlth,sleptim1,addepev2)%>%
  na.omit()
```

And plot the figure of depressive disorder figure like the one above.

```
# Plot the figure of ever been told depressive disorder
plot(health_sleep_depressive_mutate$addepev2)
```




```
#Plot the figure of people's sleeping time and general health, under the condition of being depressive or not
ggplot(health_sleep_depressive_mutate, aes(x = sleptim1, fill = genhlth)) +
  geom_bar(position = "fill") +
  facet_wrap( ~ addepev2, ncol = 5) +
  coord_flip()
```



We can see here that people being told yes were much higher than heart attack data. Means that inadequate sleep might cause more negative effects on mental illnesses like depressive disorder than heart attack.

The second plot, it shows that mental disorder like depressive is indeed influenced by inadequate sleeping time. And at the same time, those people who diagnosed depressive disorder, are living with worse general health. Means that mental illnesses have some negative effects on a person's overall health, and might even make the person's physical health worse.

Conclusion

To make a conclusion here, people under this survey mostly sleep well every night at around 6-8 hours, which helps a lot to their both mental and physical health, like depressive and heart attack. Sleeping to long on bed every day might cause bad result on people's health. And those who sleep less than 5 hours, despite the general health, are holding much more chance to have heart attack. At last, depressive disorder maybe a factor of people's bad physical health.