Exploring the BRFSS data

Setup

Load packages

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

Load data

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

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) (https://www.cdc.gov/brfss/) is a premier system of health-related telephone surveys. This surveillance system aims at measuring behaviral risk factors for the non-institutionalized adult population residing in the United States. Through monthly telephone surveys, state-specific data about U.S. residents regarding their health-related risk behaviors and preventive health practices are collected. For example, respondents need to answer questions related to tobacco use, HIV/AIDS knowledge and prevention, exercise, and etc.

Data Collection:

The BRFSS data were collected from all 50 states in the U.S., the District of Columbia, Puerto Rico, Guam, U.S. Virgin Islands, American Samoa, Federated States of Micronesia, and Palau. Interviewers utilize both landline telephone- and cellular telephone-based surveys to collect data from randomly selected individuals. According to the BRFSS Data User Guide (https://www.cdc.gov/brfss/data_documentation/pdf/UserguideJune2013.pdf), dispropotionate stratified sampling (DSS) has been used for the landline sample and simple random sampling (SRS) has been used for the cellular telephone sample. The brfss2013 dataset that we are using contains 330 variables for a total of 491,775 obervations aged 18 years or older in the 2013 BRFSS surveys. “NA” are used to denote any missing responds.

Generalizability:

Based on the information collected from the BRFSS Data User Guide (https://www.cdc.gov/brfss/data_documentation/pdf/UserguideJune2013.pdf), this sample data for 2013 survey could be used to generalize to the population of interest. The sample of this specific dataset is representative:

- Firstly, the sample size is large which is a total of 491,775 observations.
- Also, the observations in the sample come from different territories of the U.S. instead of a single geographic location.

- Finally, the dispropotionate stratified sampling and simple random sampling used have ensured that all the data are collected in a random framwork from the population.

However, potential biases may also exist due to several reasons:

- Since it is a telephone survey, certain populations were excluded from the survey (e.g. individuals do not have access to any types of phones, individuals who did not respond to the phone surveys.)
- It is possible that interviewees may overreport or underreport their responses on different variables due to misremembering information and other related factors.

Casuality:

We could not make casual conclusions based on this particular dataset. The BRFSS is an observational study which merely observe rather than interfering with how the data arise. Thus, random assignment was not used in the BRFSS. We can only use the data to assess the association between variables.

Part 2: Research questions

Research quesion 1:

Is the respondent's education level related to the access to health care coverage? Is there any difference between genders?

This question looks at how individuals with different eduaction levels perceive the importance of obtaining health care coverage. Since individuals with different education levels have different knowledge of health and health-related benefits, it would be interesting to see if there is any trends and relationship between these two variables. Also, it would be interesting to analyze the gender differences so we can figure out different perceptions between males and females.

Variables used in this analysis:

- hlthpln1: Have Any Health Care Coverage
- educa: Education Level
- sex: Respondents Sex

Research quesion 2:

Is there a relationship between general health status and sleep time? Can we observe any gender differences?

This question aims to find an association between health and sleep time. We all know that sleep is a very important component of health. It would be interesting to analyze if there is any specific connection between these two variables. Also, we are interested in assessing the gender differences. We want to see if males and females have different patterns of sleep time on different levels of general health.

Variables used in this analysis:

- genhlth: General Health
- sleptim1: How Much Time Do You Sleep
- sex: Respondents Sex

Research quesion 3:

Is there a relationship between how often individuals experience depression, their opinions on treatment, and whether they receive treatment or not?

This is a very interesting question because it compares individuals' opinions, actions, and actual mental health status. Does individuals agree with the usefulness of treatment are more willing to receive treatment whatever the mental health status they have? This analysis could be useful to see an individual's strategy when making treatment-related decisions.

Variables used in this analysis:

- misdeprd: How Often Feel Depressed Past 30 Days*
- mistrhlp: Mental Health Treatment Can Help People Lead Normal Life*
- mistmnt: Receiving Medicine Or Treatment From Health Pro For Emotional Problem*

Part 3: Exploratory data analysis

Research quesion 1:

Is the respondent's education level related to the access to health care coverage? Is there any difference between genders?

```
# Select relevant variables from the dataset and omit NAs.
q1 <- brfss2013 %>%
 select(hlthpln1, educa, sex) %>%
 filter(!is.na(hlthpln1), !is.na(educa), !is.na(sex))
```

Three variables were selected in this analysis which are hlthpln1 (Access to health care coverage), educa (Education levels), and sex (Respondents sex). In order to use valid data, we omitted NA responses.

```
# Present the totals on different levels of hlthpln1 variable.
q1 %>%
  group_by(hlthpln1) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   hlthpln1  count
##   <fct>     <int>
## 1 Yes      432677
## 2 No        55004
```

The hlthpln1 variable is a categorical variable which contains two levels "Yes" and "No" with 432,677 and 55,004 responses respectively.

```
# Present the totals on different levels of educa variable.
q1 %>%
  group_by(educa) %>%
  summarise(count=n())
```

```
## # A tibble: 6 x 2
##   educa                                                  count
##   <fct>                                                  <int>
## 1 Never attended school or only kindergarten               670
## 2 Grades 1 through 8 (Elementary)                        13322
## 3 Grades 9 though 11 (Some high school)                  27923
## 4 Grade 12 or GED (High school graduate)                142171
## 5 College 1 year to 3 years (Some college or technical school) 133750
## 6 College 4 years or more (College graduate)            169845
```

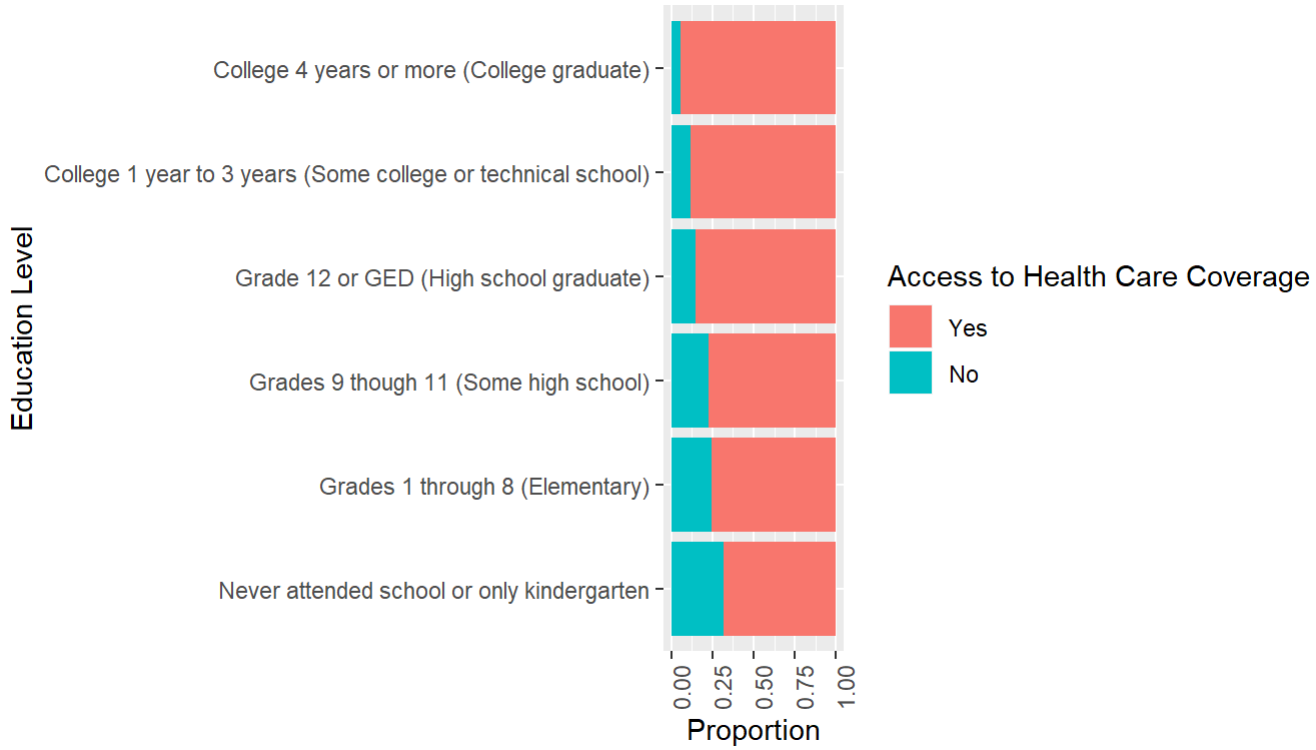
Respondents in this data come from six education levels, ranging from never attended school or only kindergarten to college graduate.

```
# Present the totals on different levels of sex variable.
q1 %>%
  group_by(sex) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   sex     count
##   <fct>   <int>
## 1 Male   199436
## 2 Female 288245
```

The data used for this analysis contains 199,436 males and 288,245 females.

```
# Use bar plot to visualize the relationship between hlthpln1 and educa variables.
ggplot(data=q1, aes(x=educa, fill= hlthpln1)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 90))+
  xlab("Education Level") +
  ylab("Proportion") +
  scale_fill_discrete(name="Access to Health Care Coverage")+
  coord_flip()
```

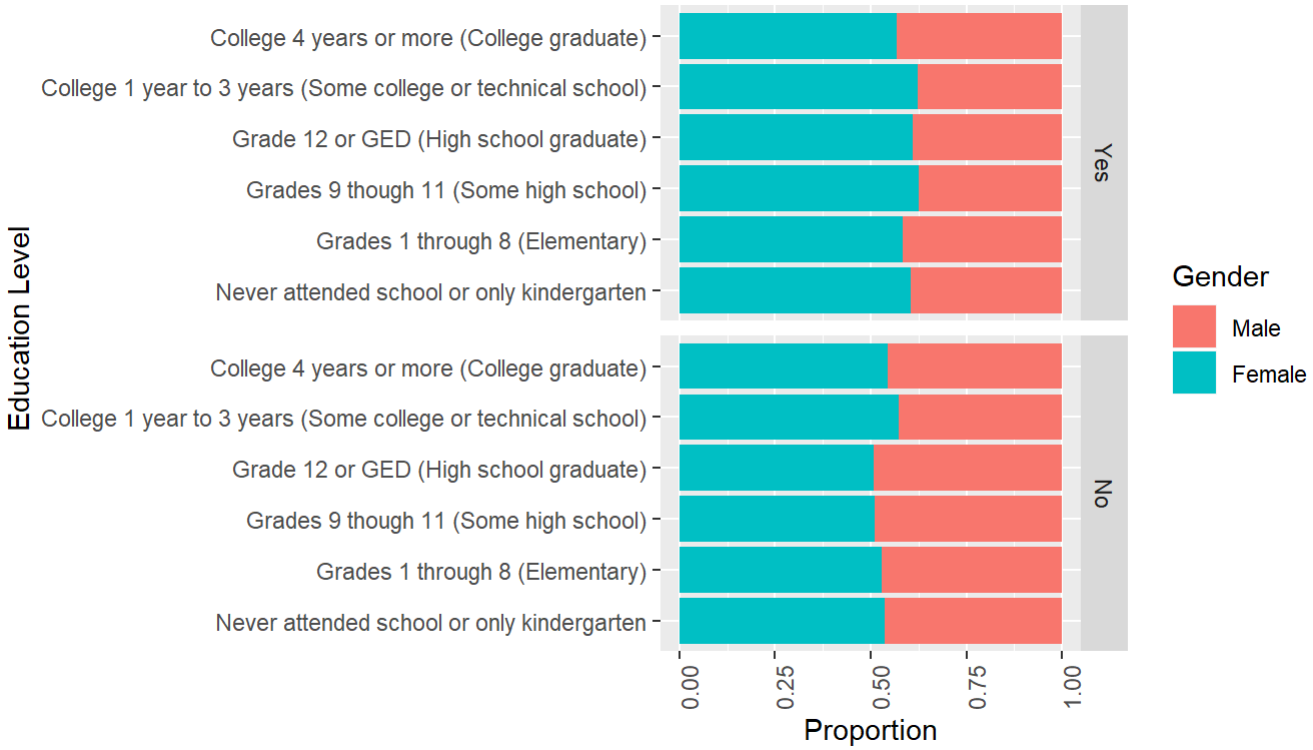


According to the bar plot, several trends can be observed:

- The proportion of responding "Yes" regarding access to health care coverage increases when individuals with higher education levels. This indicates that individuals with higher education levels value health care coverage more than those with lower education levels.
- The access to health care coverage seems to increase greatly after individuals graduate from certain education levels (College graduate vs. Some college or techinical school, High school graduate vs. Some high school). This may indicate that individuals' perceptions of the importance of health care coverage change greatly after receving certain degrees.

Let's look at the effect of the gender variable?

```
# Use bar plot to visualize the relationship between hlthpln1, educa, and sex variables.
ggplot(data=q1, aes(x=educa, fill= sex)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 90))+
  facet_grid(hlthpln1 ~ .) +
  xlab("Education Level") +
  ylab("Proportion") +
  scale_fill_discrete(name="Gender")+
  coord_flip()
```



We can observe the following:

- In all different education levels, females respond more “Yes” regarding access to health care coverage than males.This indicates that females pay more attention to the benefits they can obtain from health care coverage than males.
- The “Yes” responses of females decrease after they graduate from certain education levels. On the other hand, males who have graduated response more “Yes” than those who are still in school. This could indicate that how males and females perceive the importance of health care coverage after graduation.

In summary, this analysis indicates that there is a relationship between education level and the access to health care coverage, as well as gender differences.

Research quesion 2:

Is there a relationship between general health status and sleep time? Can we observe any gender differences?

```
# Select relevant variables from the dataset and omit NAs.
q2 <- brfss2013 %>%
   select(genhlth, sleptim1, sex) %>%
   filter(!is.na(genhlth), !is.na(sleptim1), !is.na(sex))
```

Three variables were selected in this analysis which are genhlth (General Health), sleptim1 (How much time do you sleep), and sex (Respondents sex). In order to use valid data, we omitted NA responses.

```
# Present the totals on different levels of genhlth variable.
q2 %>%
  group_by(genhlth) %>%
  summarise(count = n())
```

```
## # A tibble: 5 x 2
##   genhlth    count
##   <fct>      <int>
## 1 Excellent  84821
## 2 Very good 157832
## 3 Good      148299
## 4 Fair       65012
## 5 Poor       26639
```

The genhlth variable is a categorical and contains 5 levels: “Excellent”, “Very Good”, “Good”, “Fair”, and “Poor.”

```
# Present the descriptive statistics of sleptim1 variable.
q2 %>%
  summarise(mean_slep = mean(sleptim1), median_slep =  median(sleptim1),
            sd_slep = sd(sleptim1), iqr_slep = IQR(sleptim1),
            max_slep = max(sleptim1), min_slep = min(sleptim1))
```

```
##   mean_slep median_slep  sd_slep iqr_slep max_slep min_slep
## 1  7.050559           7 1.464554        2       24        1
```

The sleptim1 variable is numerical ranging from 1 to 24. The mean is 7.050559, the median is 7, the standard deviation is 1.464554, and the interquartile range is 2.

```
# Present the totals on different levels of sex variable.
q2 %>%
  group_by(sex) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   sex     count
##   <fct>   <int>
## 1 Male   198148
## 2 Female 284455
```

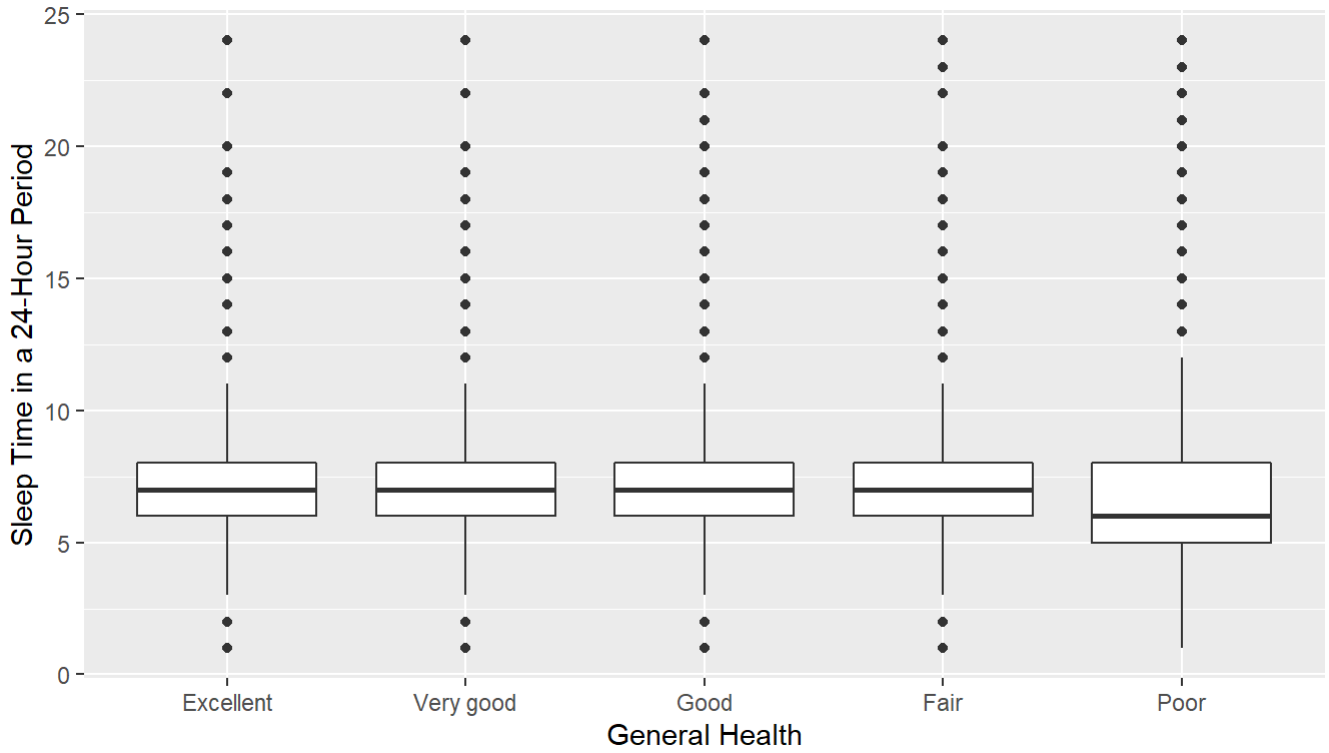
The data used for this analysis contains 198,148 males and 284,455 females.

```
# Present the descriptive statisctis of sleptim1 variable on different levels of genhlth variabl
  e.
q2 %>%
  group_by(genhlth) %>%
  summarise(mean_slep = mean(sleptim1), median_slep = median(sleptim1),
            sd_slep = sd(sleptim1), iqr_slep = IQR(sleptim1),
            max_slep = max(sleptim1), min_slep = min(sleptim1))
```

```
## # A tibble: 5 x 7
##   genhlth   mean_slep median_slep sd_slep iqr_slep max_slep min_slep
##   <fct>         <dbl>       <dbl>   <dbl>    <dbl>    <int>    <int>
## 1 Excellent      7.19           7    1.21        2       24        1
## 2 Very good      7.10           7    1.21        2       24        1
## 3 Good           7.04           7    1.44        2       24        1
## 4 Fair           6.90           7    1.81        2       24        1
## 5 Poor           6.74           6    2.39        3       24        1
```

This is a table of numerical summary of sleptim1 on different levels of genhlth. It would be more effective to visualize these data in a boxplot.

```
# Use bar plot to visualize the relationship between genhlth and sleptim1 variables.
ggplot(data = q2, aes(x = genhlth, y = sleptim1)) +
  geom_boxplot() +
  xlab ("General Health") +
  ylab ("Sleep Time in a 24-Hour Period")
```

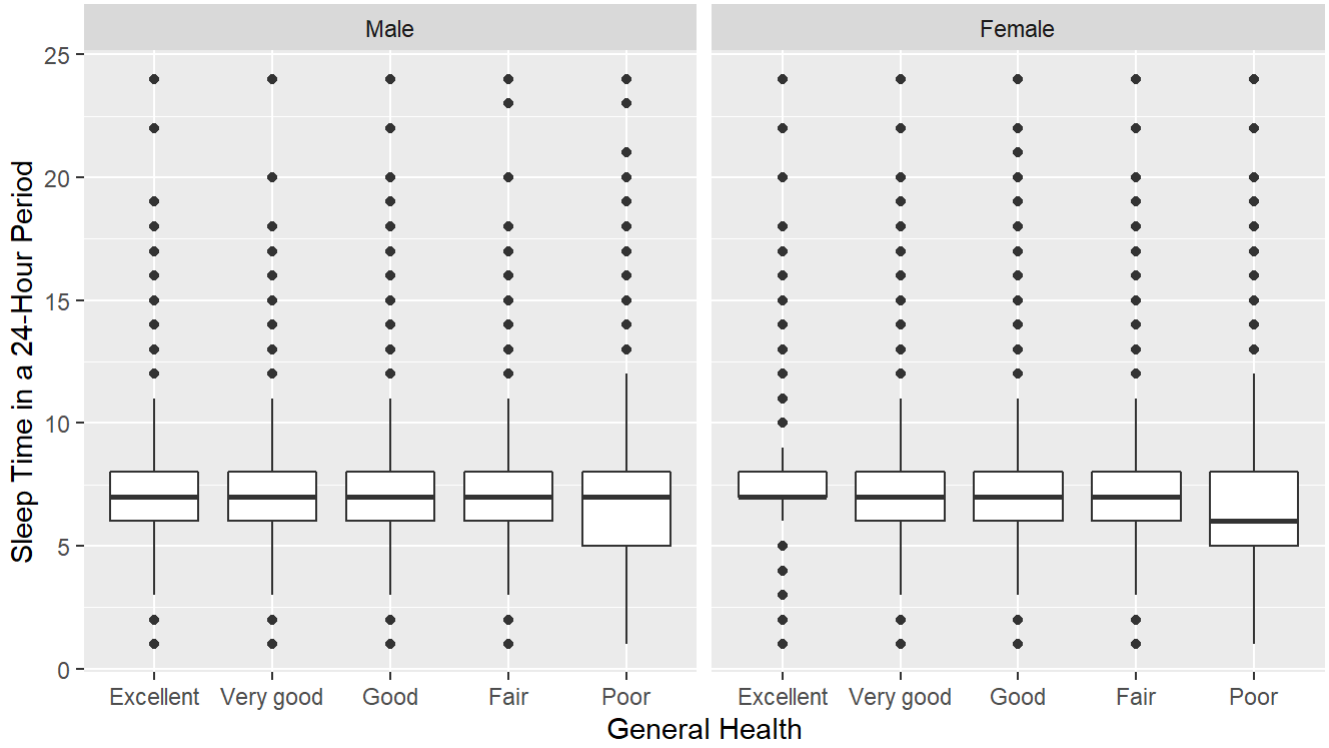


We can observe the following from the boxplot:

- The median of sleep time for individuals with poor health status is much lower than others. This indicates that individuals with poor health sleep fewer hours than other health status.
- The distribution patterns of sleep time for those who reported “Excellent”, “Very Good”, “Good”, and “Fair” health status are quite similar. This could indicate that when individuals with at least fair health status, their sleep hours are consistent.
- The distribution varies more when individuals reported worse health status. This may indicate that the sleep time for individuals with worse health status is more inconsistent than those who are healthier.

Let’s add the gender variable.

```
# Use bar plot to visualize the relationship between genhlth, sleptim1, and sex variables.
ggplot(data = q2, aes(x = genhlth, y = sleptim1)) +
  geom_boxplot() +
  facet_grid(. ~ sex) +
  xlab ("General Health") +
  ylab ("Sleep Time in a 24-Hour Period")
```



We can observe the following:

- The sleep hours of females with poor health are much fewer than females with other health status. This may indicate the relationship between less sleep time and poor health status.
- The distributions seem to be similar for males and females with poor health. However, the median of females is much smaller than males. This could indicate that females tend to sleep fewer hours than males when they have poor health status.
- The distrution of sleep hours for females with excellent health varys much less than that for males. This indicates that females with excellent health have more consistent sleep hours than males with the same health status.

In summary, this analysis indicates that there is a relationship between general health status and sleep hours. Also, we can observe gender differences.

Research quesion 3:

Is there a relationship between how often individuals experience depression, their opinions on treatment, and whether they receive treatment or not?

```
# Select relevant variables from the dataset and omit NAs.
q3 <- brfss2013 %>%
  select(misdeprd, mistrhlp, mistmnt) %>%
  filter(!is.na(misdeprd), !is.na(mistmnt), !is.na(mistrhlp))
```

Three variables were selected in this analysis which are misdeprd (How Often Feel Depressed Past 30 Days), mistrhlp (Mental Health Treatment Can Help People Lead Normal Life), and mistmnt (Receiving Medicine Or Treatment From Health Pro For Emotional Problem). In order to use valid data, we omitted NA responses.

```
# Present the totals on different levels of misdeprd variable.
q3 %>%
  group_by(misdeprd) %>%
  summarise(count = n())
```

```
## # A tibble: 5 x 2
##   misdeprd count
##   <fct>    <int>
## 1 All        236
## 2 Most       502
## 3 Some      1482
## 4 A little  2410
## 5 None     29710
```

The misdeprd variable is a categorical variable which contains five levels: “All”, “Most”, “Some”, “A little”, and “None.”

```
# Present the totals on different levels of mistrhlp variable.
q3 %>%
  group_by(mistrhlp) %>%
  summarise(count = n())
```

```
## # A tibble: 5 x 2
##   mistrhlp                   count
##   <fct>                      <int>
## 1 Agree strongly             24984
## 2 Agree slightly              6965
## 3 Neither agree nor disagree   906
## 4 Disagree slightly            976
## 5 Disagree strongly            509
```

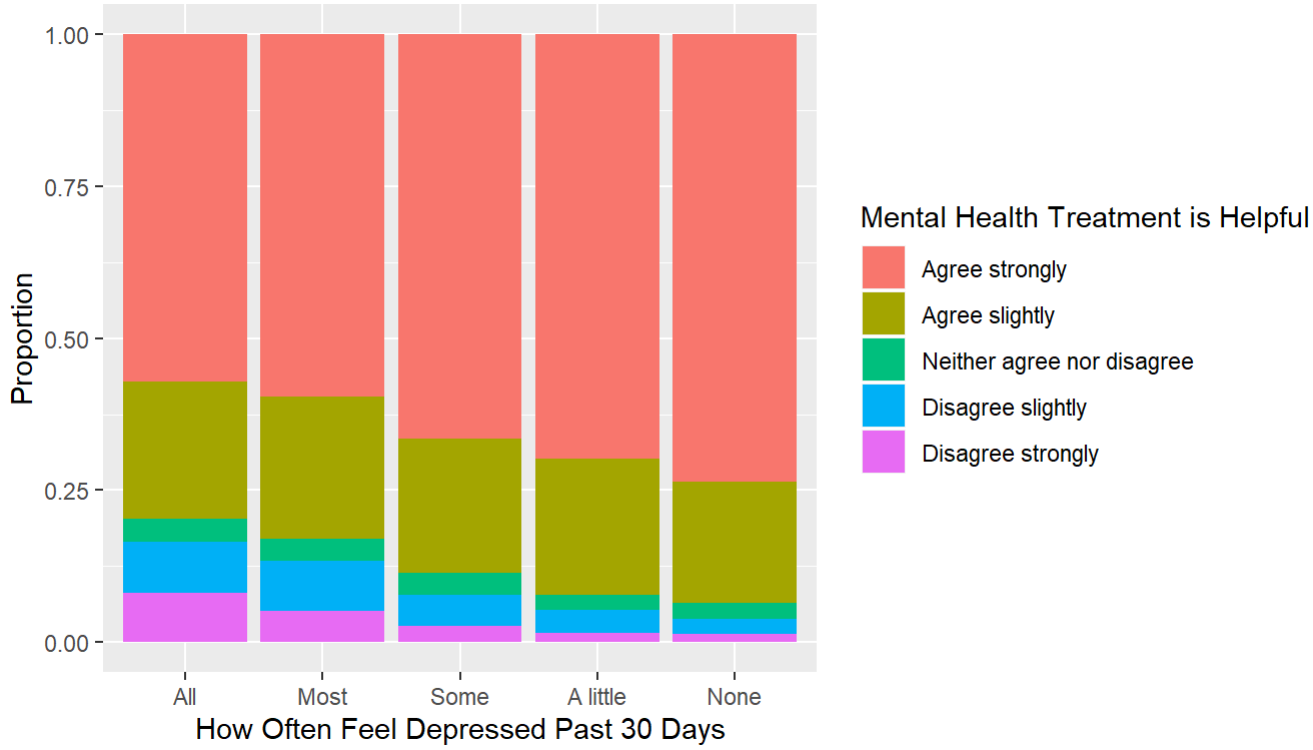
The mistrhlp variable represents how individuals perceive the usefulness of treatment. The levels range from “Agree strongly” to “Disagree Strongly.”

```
# Present the totals on different levels of mistmnt variable.
q3 %>%
  group_by(mistmnt) %>%
  summarise(count = n())
```

```
## # A tibble: 2 x 2
##   mistmnt count
##   <fct>   <int>
## 1 Yes      5214
## 2 No      29126
```

5,214 respondents receive treatment for their emotional issues, and 29,126 respondents do not.

```
# Use bar plot to visualize the relationship between misdeprd and mistrhlp variables.
ggplot(data = q3, aes(x = misdeprd, fill = mistrhlp)) +
  geom_bar(position = "fill") +
  xlab("How Often Feel Depressed Past 30 Days") +
  ylab("Proportion") +
  scale_fill_discrete(name="Mental Health Treatment is Helpful")
```



We can observe the following from the bar plot:

- The proportion of “Agree strongly” response increases when individuals have fewer days of feeling depressed. On the other hand, individuals, who feel depressed more often, are more possible to provide “Disagree strongly” responses. This may indicate that individuals who feel less depressed trust the usefulness of mental health treatment than those who are more depressed.
- The proportion of Neither agree nor disagree" seems to be very stable in different bars. This may indicate that the frequency of depression is more related to extreme opinions like “Agree strongly” or “Disagree Strongly.”

Let’s see what the relationship will look like when we add mistmnt variable.

```
# Use bar plot to visualize the relationship between misdeprd, mistrhlp, and mistmnt variables.
ggplot(data = q3, aes(x = misdeprd, fill = mistmnt)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 90)) +
  facet_grid(. ~ mistrhlp) +
  xlab("How Often Feel Depressed Past 30 Days") +
  ylab("Proportion") +
  scale_fill_discrete(name = "Receiving Treatment or Not")
```

It is interesting to observe that:

- Except the “Disgree strongly”, it is more possible for individuals who experienced more days of depression to receive treatment. In other words, individuals who have fewer days of depression is more associated with not receiving treatments. However, according to the analysis above, individuals who experienced fewer days of depression tends to strongly agree with the usefulness of treatments. This may indicate that these individuals tends to use other ways to treat their mental health.
- On all levels of perceptions of treatment, the “Yes” repsonses increase greatly when individuals experienced depression (“None” to “A little”). This could indicate that individuals are sensitive to their mental health status and tend to use treatment as a way to deal with issues related to depression.

In summary, this analysis indicates that there is a relationship between how often people experience depression, perceptions of treatment, and receiving treatment or not.