Exploring the BRFSS data

Setup

Load packages

The project has been completed using the grammar of `dplyr` and `ggplot2` packages from the `tidyverse` family.

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

Load data

The data used for this project has been stored as a `.RData` file.

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

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project between all of the states in the United States (US) and participating US territories and the Centers for Disease Control and Prevention (CDC). The BRFSS is administered and supported by CDC's Population Health Surveillance Branch, under the Division of Population Health at the National Center for Chronic Disease Prevention and Health Promotion.

The BRFSS questionnaire is comprised of an annual standard core, a biannual rotating core, optional modules, and state-added questions. Since only 5 states opted to include the optional module 17 as part of their survey in 2013, the amount of data available for the second research question reduced drastically, as is highlighted in the EDA for it.

States design samples within boundaries of sub-state geographic regions. States may determine that they would like to sample by county, public health district or other sub-state geography in order to make comparisons of geographic areas with their states. In order to conduct the BRFSS, states obtain samples of telephone numbers from CDC.

Thus, the results of any analysis on the data set can be generalized to the entire population of the US, provided a larger enough sample is obtained after filtering. Also, since there is no use for random assignment here, it being a survey, we cannot make any causality claims based on our analyses.

Part 2: Research questions

Research question 1

Do people who have higher minutes of total physical activity eat more fruits & vegetables than those with lesser minutes?

Research question 2

How do people belonging to different age groups and education levels perceive the usefulness of mental health treatment in leading a normal life?

Research question 3

Do people who smoke more than others, also drink more alcoholic beverages than others?

Part 3: Exploratory data analysis

Do people who have higher minutes of total physical activity eat more fruits & vegetables than those with lesser minutes?

We make use of two calculated factor variables available in the data set, `X_pa150r2` and `X_pa300r2`, to calculate a new factor variable- `phys_act_weekly`- which represents the total minutes(or equivalent vigorous minutes) of physical activity per week.

We calculate another factor variable- `frt_veg_intake_daily`- which represents whether or not the surveyee eats fruits and/or vegetables once or more every day, using two more calculated variables available in the data set, `X_frtlt1` and `X_veglt1`, representing whether the surveyee consumes fruits and vegetables, respectively, once or more in a day.

```
q1<-brfss2013  %>%
  mutate(
    phys_act_weekly=case_when(
      X_pa150r2=="0 minutes" & X_pa300r2=="0 minutes" ~ "0 minutes",
      X_pa150r2=="1-149 minutes" & X_pa300r2=="1-300 minutes" ~ "1-149 minutes",
      X_pa150r2=="150+ minutes" & X_pa300r2=="1-300 minutes" ~ "150-300 minutes",
      X_pa150r2=="150+ minutes" & X_pa300r2=="301+ minutes" ~ "301+ minutes"
                            ),

    frt_veg_intake_daily=case_when(
      X_frtlt1==levels(X_frtlt1)[1] & X_veglt1==levels(X_veglt1)[1]
             ~ "both, once or more",
      (X_frtlt1==levels(X_frtlt1)[2] & X_veglt1==levels(X_veglt1)[1]) |
      (X_frtlt1==levels(X_frtlt1)[1] & X_veglt1==levels(X_veglt1)[2])
             ~ "at least one, once or more",
      X_frtlt1==levels(X_frtlt1)[2] & X_veglt1==levels(X_veglt1)[2]
             ~ "both, less than once"
                                    )
        )  %>%
  filter(!is.na(phys_act_weekly)&!is.na(frt_veg_intake_daily))  %>%
  count(phys_act_weekly,frt_veg_intake_daily)  %>%  ungroup()
```

Taking a look at the data frame we just created, we see that the numbers vary quite a lot across different levels of the physical activity variable.

```
q1
```

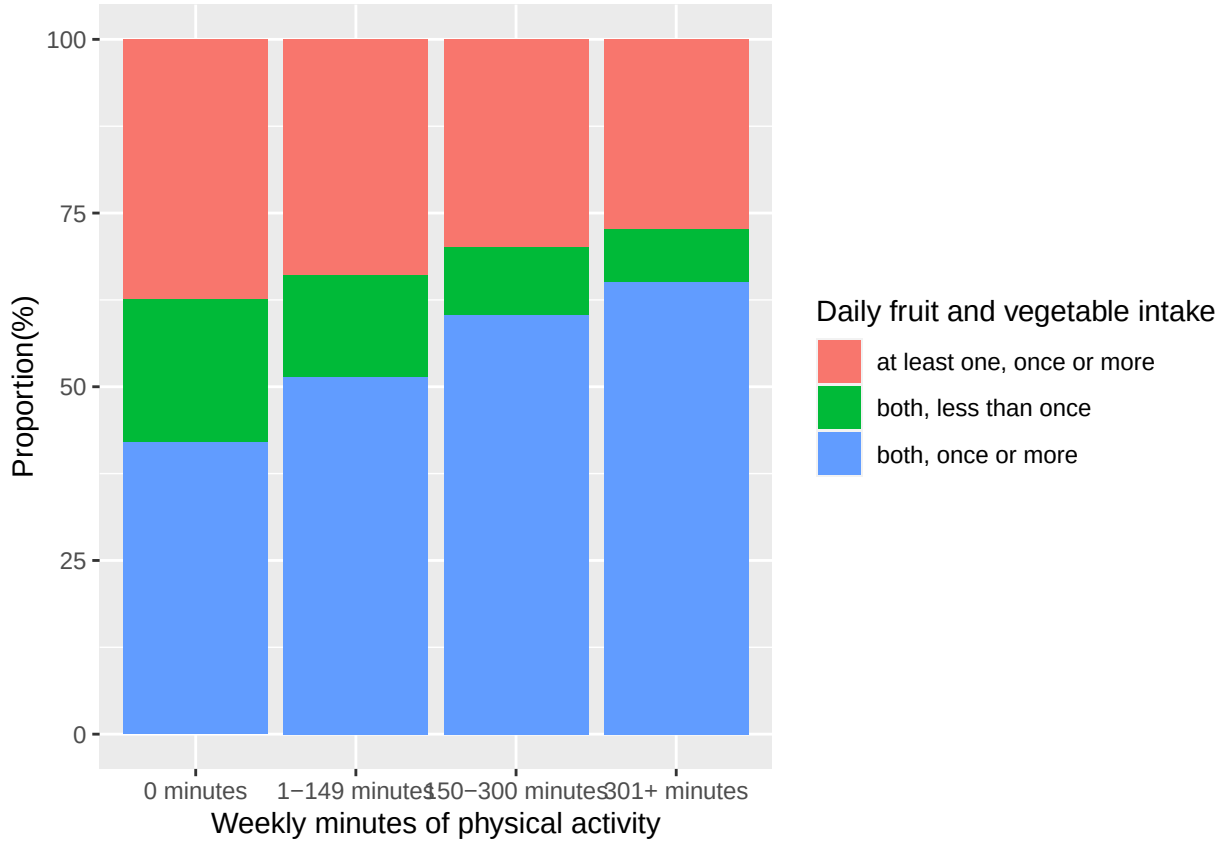
```
##    phys_act_weekly       frt_veg_intake_daily     n
## 1        0 minutes at least one, once or more 48536
## 2        0 minutes       both, less than once 26710
## 3        0 minutes         both, once or more 54369
## 4    1-149 minutes at least one, once or more 26347
## 5    1-149 minutes       both, less than once 11309
## 6    1-149 minutes         both, once or more 39831
## 7  150-300 minutes at least one, once or more 21320
## 8  150-300 minutes       both, less than once  6967
## 9  150-300 minutes         both, once or more 43011
## 10    301+ minutes at least one, once or more 36707
## 11    301+ minutes       both, less than once 10183
## 12    301+ minutes         both, once or more 87397
```

Thus, a look at the proportions might give a better picture of how many people belong in different categories. We add a variable to this data frame representing the what proportion of people in a certain level of `phys_act_weekly` belong to a certain level of `frt_veg_intake_daily`.

```
q1<-q1  %>%  group_by(phys_act_weekly)  %>%
  mutate(prop=100*n/sum(n))  %>%  ungroup()
```

Now, we can understand these values much better once we plot them on a bar plot.

```
ggplot(q1) + aes(phys_act_weekly, prop, fill=frt_veg_intake_daily) +
  geom_col() + labs(x="Weekly minutes of physical activity",
                    y="Proportion(%)", fill="Daily fruit and vegetable intake")
```



We can see from the plot that the proportion of people who eat both fruits and vegetables at least once a day increases consistently with an increase in physical activity minutes. This increase is accompanied by a correpsponding decrease in the proportion of people who eat less than one fruit and one vegetable daily.

This suggests that there is a positive relationship between being more physically active and having healthier eating habits.

How do people belonging to different age groups and education levels perceive the usefulness of mental health treatment in leading a normal life?

We calculate what proportion of people in each age group, given in the data set by the computed variable `X_ageg5yr`, have what views about the effectiveness of mental health treatment in leading a normal life, given by the variable `mistrhlp`.

```
q2a<-brfss2013  %>%  filter(!is.na(X_ageg5yr)&!is.na(mistrhlp))  %>%
  count(X_ageg5yr,mistrhlp)  %>%  ungroup()  %>%
  group_by(X_ageg5yr)  %>%  mutate(prop=100*n/sum(n))  %>%  ungroup()
```

We create a different data frame, calculating what proportion of people having different educational qualifications, given in the data set by the variable `educa`, have what views about the effectiveness of mental health treatment in leading a normal life, given by the variable `mistrhlp`.

```
q2b<-brfss2013  %>%  filter(!is.na(educa)&!is.na(mistrhlp))  %>%
  count(educa,mistrhlp)  %>%  ungroup()  %>%
  group_by(educa)  %>%  mutate(prop=100*n/sum(n))  %>%  ungroup()
```

We have to keep in mind, however, that the sample size of people belonging to each category of age group and education level has decreased drastically because of our subsetting of non-empty rows for the variable `mistrhlp`. This is so because in the year 2013, only 5 five states included the optional module 17, relating to mental illness and stigma, as part of their surveys for the Behavioral Risk Factor Surveillance System survey. This means that most of the people, as evidenced by the proportion of `NA`'s below, were not asked the questions relating to mental health.

```
brfss2013  %>%
  count(X_ageg5yr,mistrhlp)  %>%  ungroup()  %>%
  group_by(X_ageg5yr)  %>%  mutate(prop=100*n/sum(n))  %>%  ungroup()  %>%
  filter(is.na(mistrhlp))
```

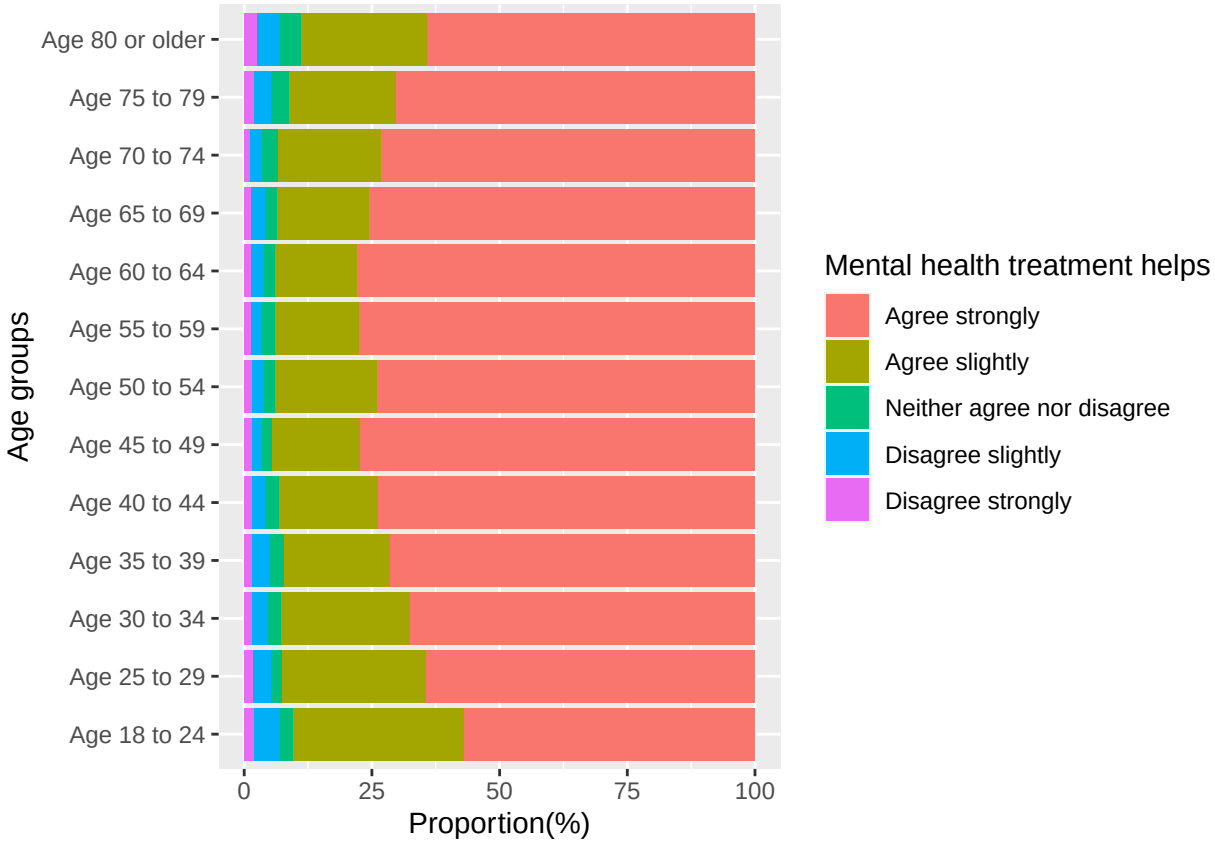
```
## # A tibble: 14 x 4
##    X_ageg5yr       mistrhlp     n  prop
##    <fct>           <fct>    <int> <dbl>
##  1 Age 18 to 24    <NA>     25642  94.3
##  2 Age 25 to 29    <NA>     21563  94.1
##  3 Age 30 to 34    <NA>     25468  93.6
##  4 Age 35 to 39    <NA>     26102  92.9
##  5 Age 40 to 44    <NA>     29158  92.5
##  6 Age 45 to 49    <NA>     33491  92.5
##  7 Age 50 to 54    <NA>     43665  92.6
##  8 Age 55 to 59    <NA>     48527  92.4
##  9 Age 60 to 64    <NA>     49488  92.2
## 10 Age 65 to 69    <NA>     46243  92.5
## 11 Age 70 to 74    <NA>     37147  92.9
## 12 Age 75 to 79    <NA>     28164  93.6
## 13 Age 80 or older <NA>     38120  94.1
## 14 <NA>            <NA>      4468  94.5
```

```
brfss2013  %>%
  count(educa,mistrhlp)  %>%  ungroup()  %>%
  group_by(educa)  %>%  mutate(prop=100*n/sum(n))  %>%  ungroup()  %>%
  filter(is.na(mistrhlp))
```

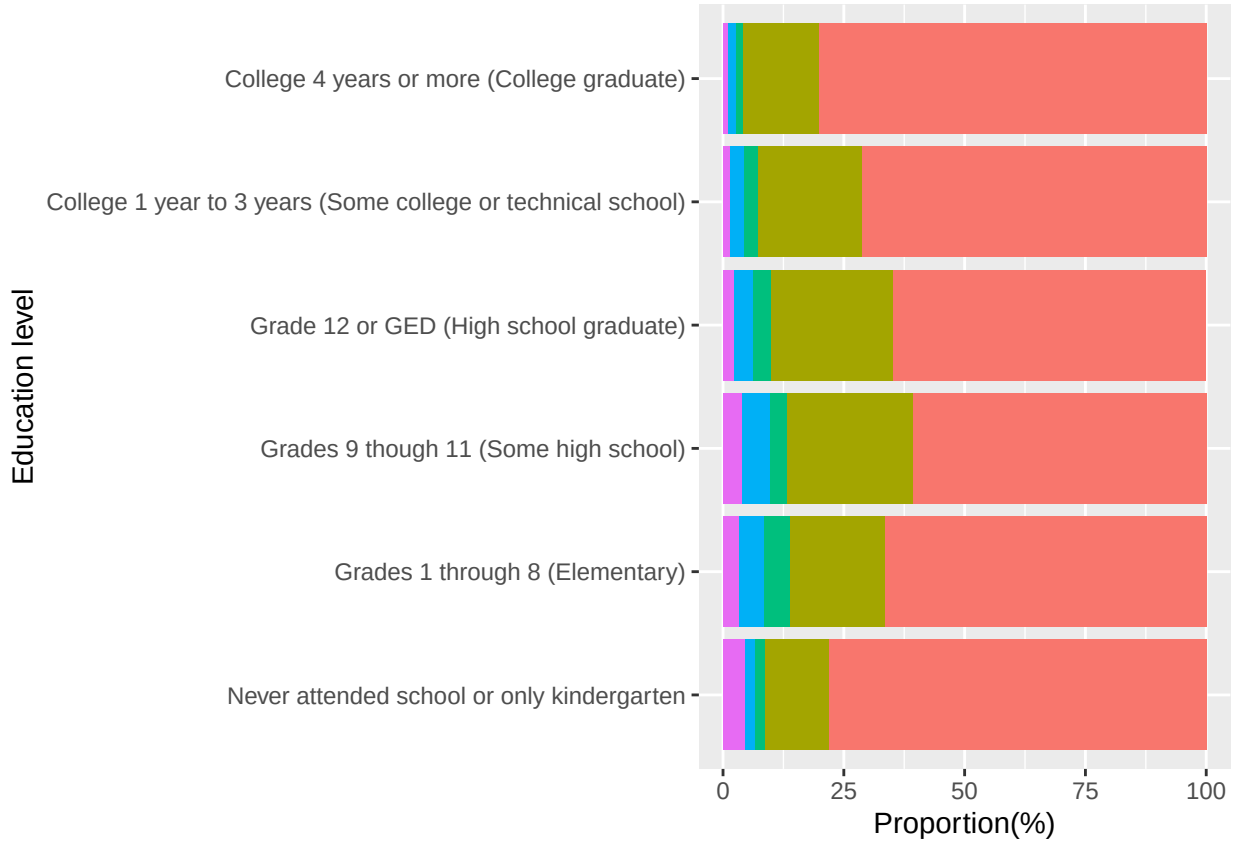
```
## # A tibble: 7 x 4
##   educa                                                mistrhlp      n  prop
##   <fct>                                                <fct>     <int> <dbl>
## 1 Never attended school or only kindergarten           <NA>        631  93.2
## 2 Grades 1 through 8 (Elementary)                      <NA>      12810  95.6
## 3 Grades 9 though 11 (Some high school)                <NA>      26841  95.4
## 4 Grade 12 or GED (High school graduate)               <NA>     134415  94.0
## 5 College 1 year to 3 years (Some college or technical sc~ <NA>  123882  92.3
## 6 College 4 years or more (College graduate)           <NA>     156465  92.0
## 7 <NA>                                                 <NA>       2202  96.8
```

Having taken note of this sample's eccentricity, we can nonetheless see whether there is a relationship between the variables in question by plotting them on a bar plot.

```
ggplot(q2a) + aes(prop,X_ageg5yr,fill=mistrhlp)+geom_col() +
  labs(x="Proportion(%)", y="Age groups", fill="Mental health treatment helps")
```




```
ggplot(q2b) + aes(prop,educa,fill=mistrhlp)+geom_col() +
  labs(x="Proportion(%)", y="Education level",
       fill="Mental health treatment helps") +
  theme(legend.position="none")
```



We can see from the plots that in each age group and education level, a majority strongly agrees that mental health treatment does help people in leading normal lives. Also, a very small proportion of people disagree with the statement.

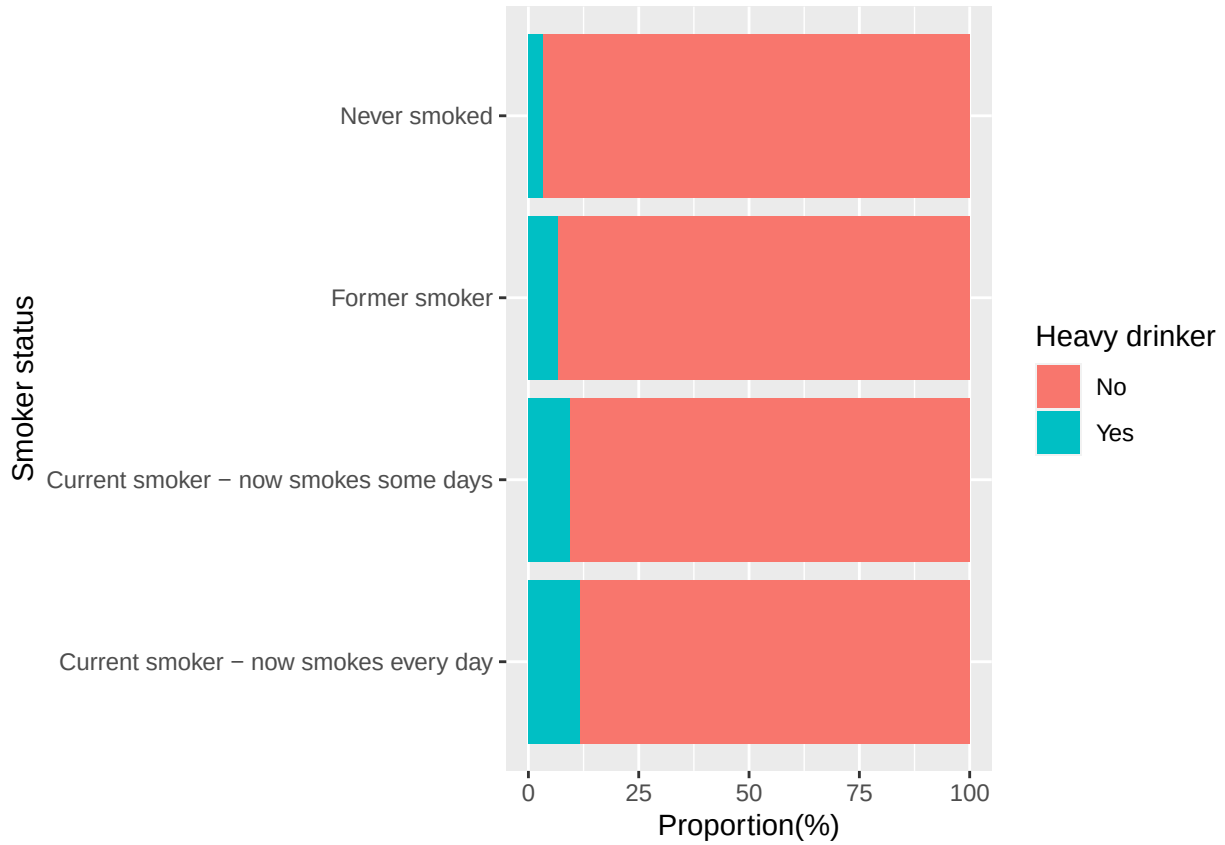
Do people who smoke more than others, also drink more alcoholic beverages than others?

We use two computed variables from the data set, `X_rfdrhv4` and `X_smoker3`, which represent whether the surveyee is a heavy drinker or not and the smoking status of the surveyee, respectively. We calculate the proportion of people in each smoker status who are heavy drinkers, and see if a relationship emerges.

```
q3<-brfss2013  %>%  filter(!is.na(X_rfdrhv4)&!is.na(X_smoker3))  %>%
  count(X_smoker3,X_rfdrhv4)  %>%  ungroup()  %>%
  group_by(X_smoker3)  %>%  mutate(prop=100*n/sum(n))  %>%  ungroup()
```

To see if a relationship actually exists or not, a bar plot would be much more useful.

```
ggplot(q3, aes(prop,X_smoker3,fill=X_rfdrhv4)) + geom_col() +
  labs(x="Proportion(%)", y="Smoker status", fill="Heavy drinker")
```



We can clearly see from the chart that people who smoke more have a higher proportion of people who are also avid drinkers. This suggests that a person who currently smokes every day is more likely to be a heavy drinker as well than those who do not smoke every day, or than those who never smoked.