Exploring the BRFSS data

Setup

Load packages

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

Load data

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

Part 1: Data

In the brfss, stratified sampling was used. If we miss the fact that participants in such long questionnaires are category of people who can't be absolutely random, this study is observational and generalizable for the entire population. Since participants were not randomly divided to groups and assigned to perform different tasks, etc., causation can not be infered from data.

There should also be inaccuracies and distortions in the data, since it is unlikely that people can accurately and honestly answer such questions as: "How many of the past 30 days have been active or healthy?"

Part 2: Research questions

Research quesion 1: For the general population in the US, is there a correlation between the amount of sleep, the person's cognitive functioning and his income level? For my own experience I believe that people who sleep regularly are more productive and sucsess in study. I wonder if there associative relationship between sleep hours, cognitive health and signs of productivity, such as high wages?

Research quesion 2: For the general population in the US, is there a pattern in relation between month and reported number of days with poor mental health? Many people associate their depression or poor mental health with season. There is also a Seasonal affective disorder which is occurs in same season every year, more often in winter. I wonder if that facts can affect a data.

Research quesion 3: For the general population in the US, is there any relation between smoking, drinking alcohol, weight and having depression disorder? Several studies that I

read earlier found a correlation between depression, obesity, alcohol, and smoking. I will try to find a correlation between these variables in the proposed data.

Part 3: Exploratory data analysis

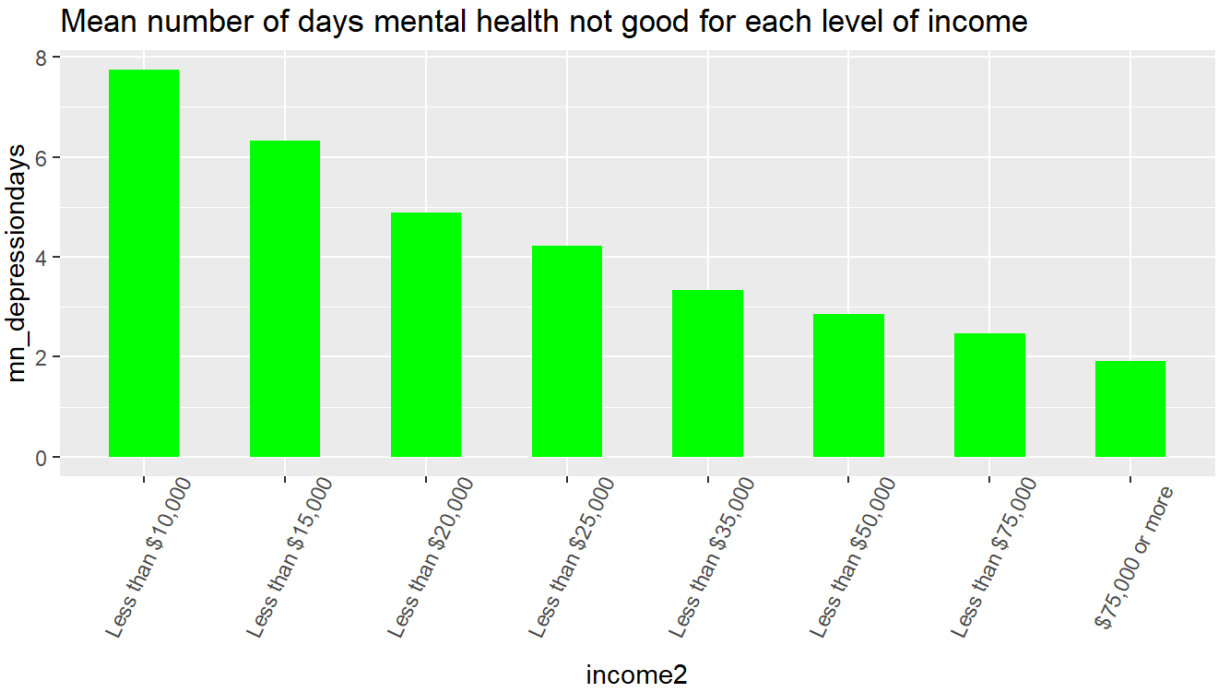
Research quesion 1:

All meaningless emissions and observations with missing data were filtered out. To determine whether there is an association between income level and depression level, all data were grouped by income level, and then the average number of days with poor mental health was calculated. All data was visualized using bars.

```
#Filter observation with no data
mydata <- filter(brfss2013, !is.na(income2), !is.na(menthlth),menthlth < 3
1, !is.na(sleptim1), sleptim1 < 25)

myvars <- c("income2", "menthlth", "sleptim1")
mydata <- mydata[myvars]

# Create df income level and mean of sleep time for each income level cata
gory.
income_M_Hlth <- mydata %>% group_by(income2) %>% summarise(mn_depressiond
ays = mean(menthlth))

# display
income_M_Hlth
```

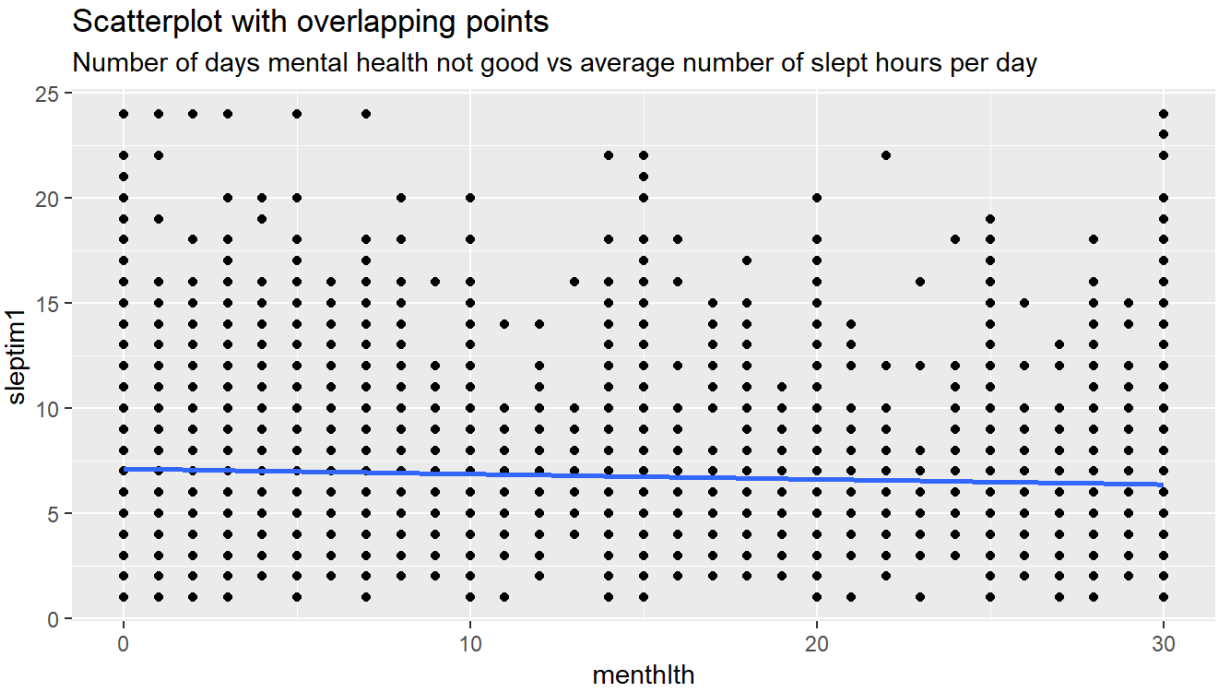
```
## # A tibble: 8 x 2
##   income2           mn_depressiondays
##   <fct>                         <dbl>
## 1 Less than $10,000              7.75
## 2 Less than $15,000              6.32
## 3 Less than $20,000              4.88
## 4 Less than $25,000              4.21
## 5 Less than $35,000              3.33
## 6 Less than $50,000              2.86
## 7 Less than $75,000              2.47
## 8 $75,000 or more                1.93
```

```
ggplot(income_M_Hlth, aes(x=income2, y=mn_depressiondays)) +
  geom_bar(stat="identity", width=.5, fill="green") +
  labs(title="Mean number of days mental health not good for each level of
income") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



There is a correlation between lower income and more depressive, stressful days, which is not at all surprising. The main question is: is there a correlation between the average hours of sleep per day and the number of days with poor mental health? Average hours of sleep per day stored in “sleptim1” variable.

```
g <- ggplot(mydata, aes(menthlth, sleptim1))
g + geom_point() +
  geom_smooth(method="lm", se=F) +
  labs(subtitle="Number of days mental health not good vs average number o
f slept hours per day",
       y="sleptim1",
       x="menthlth",
       title="Scatterplot with overlapping points")
```



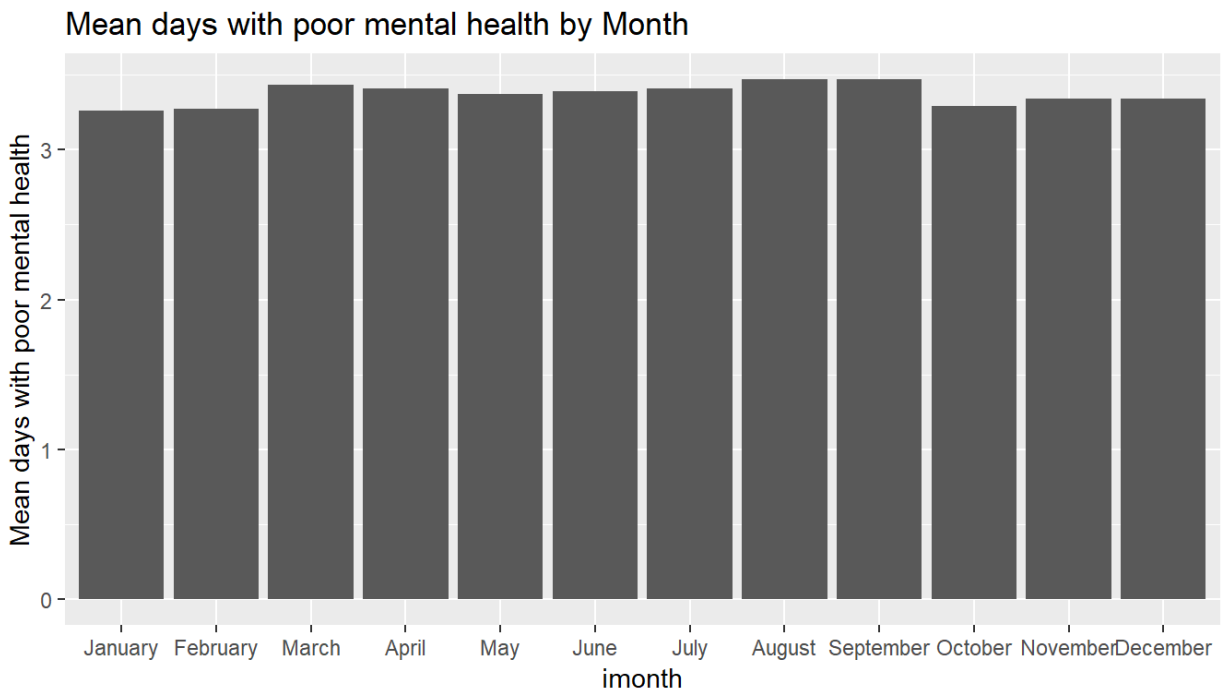
Looking at the plot above, there does not appear to be a relation between time sleeping and the number of days with poor mental health.

Research quesion 2:

```
#All observations with missed data have been filtered out. Then all respon
ds was grouped by month, added, and divided on number of responses.
Sum_poorhlthday <- filter(brfss2013, !is.na(menthlth), !is.na(imonth) )
Sum_poorhlthday <- aggregate(Sum_poorhlthday$menthlth, by=list(sum=Sum_poo
rhlthday$imonth), FUN=sum)

reports <- brfss2013 %>% filter(!is.na(menthlth), !is.na(imonth)) %>% grou
p_by(imonth) %>% summarise(n=n() )
reports <- mutate(reports, mean = round(Sum_poorhlthday$x/reports$n, 2))

#print results
reports
```

```
## # A tibble: 12 x 3
##    imonth        n  mean
##    <fct>     <int> <dbl>
##  1 January   33645  3.26
##  2 February  42128  3.27
##  3 March     43723  3.43
##  4 April     42184  3.41
##  5 May       39771  3.37
##  6 June      37204  3.39
##  7 July      42864  3.41
##  8 August    41559  3.47
##  9 September 37840  3.47
## 10 October   41507  3.29
## 11 November  41175  3.34
## 12 December  39547  3.34
```

```
#Plot data using bars

ggplot(aes(x=imonth, y=mean), data=reports) + geom_bar(stat = 'identity')
+ ggtitle('Mean days with poor mental health by Month')+ ylab('Mean days w
ith poor mental health')
```



I was trying to find out whether people respond their mental health condition differently in the different month. For example, are people more likely to say they have poor mental health in winter or autumn? It appears that there was no obvious pattern.

Research quesion 3:

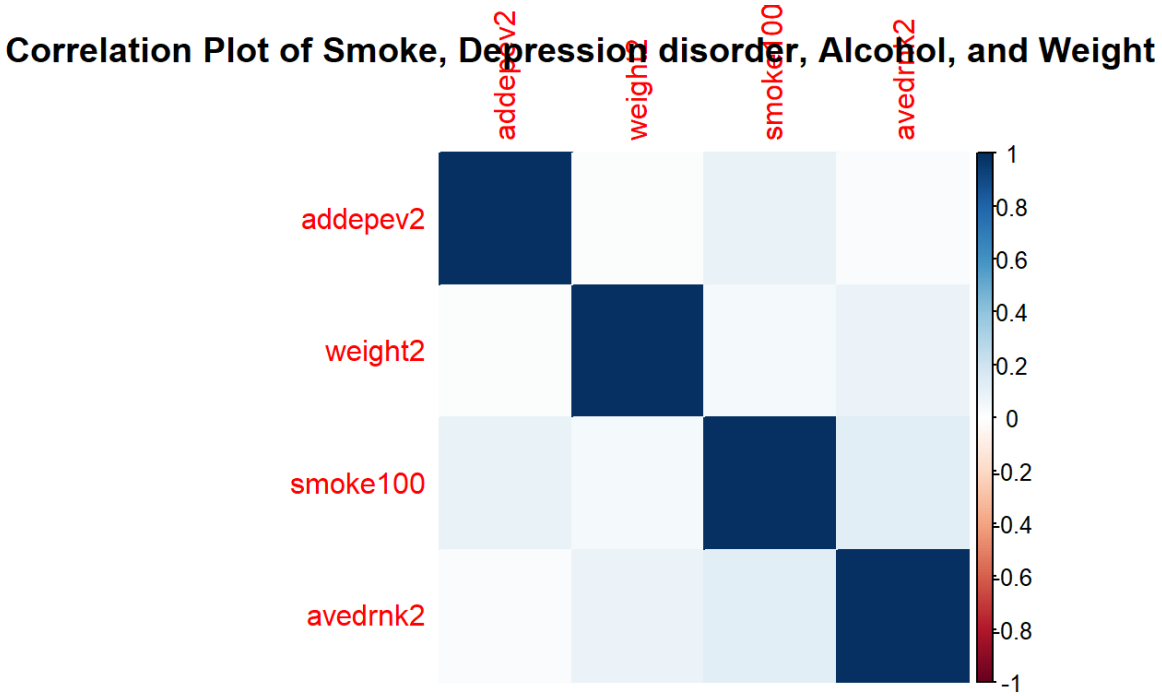
For research proposal several variables were chosen: "smoke100" - Smoked At Least 100 Cigarettes "avedrnk2" - Avg Alcoholic Drinks Per Day In Past 30 "weight2" - Reported

Weight In Pounds "addepev2"- Ever Told You Had A Depressive Disorder

```
#Select variables which are we interested in
vars <- names(brfss2013) %in% c('smoke100','addepev2', 'avedrnk2',  'weigh
t2')
#filter out observations with missed data

temp <- brfss2013 %>% filter(!is.na(smoke100), !is.na(addepev2), !is.na(av
edrnk2), !is.na(weight2))
selected_var <- temp[vars]

#transformate data to numeric
selected_var$addepev2 <- ifelse(selected_var$addepev2=="Yes", 1, 0)
selected_var$weight2 <- as.numeric(selected_var$weight2)
selected_var$smoke100 <- ifelse(selected_var$smoke100=="Yes", 1, 0)
selected_var$avedrnk2 <- as.numeric(selected_var$avedrnk2)

library(Hmisc)
```

```
library(corrplot)
```

```
#plot data
corr.matrix <- cor(selected_var)
corrplot(corr.matrix, main="\n\nCorrelation Plot of Smoke, Depression diso
rder, Alcohol, and Weight", method="color")
```



It is appears that there is no strong correlation between these variables.