Exploring the BRFSS data

Setup

Load packages

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

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `brfss2013`. Delete this note when before you submit your work.

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

Part 1: Data

##About our Data Our Data is a complete survey of the health status and health issues of adult U.S. citizens which is conducted by BRFSS.The dataset "brfss2013" contains the data of *491775* randomly selected citizens and have a total of *330* variables. The data is Non - Response bias since the data only comes from people willing to answer the call and going through the long survey

##Process of Data Collection Data from 50 U.S. states, the District of Columbia, and three U.S. territories is collected and more than 400,000 adult BRFSS surveys are conducted every year. The Survey is taken using landline or celephone telephones

1) In conducting the BRFSS landline telephone survey, interviewers collect data from a randomly selected adult in a household
2) In conducting the cellular telephone version of the BRFSS questionnaire, interviewers collect data from an adult who participates by using a cellular telephone and resides in a private residence or college housing.

####Scope of inference

##Data is Generalizable Since the data is collected by sampling random adults from each household. Also none of the researchers interfered with the data hence its an observational study.The data is collected through stratified sampling and the sampling is random hence data is generalizable

##Data is Non Causal Since the participants were not randomly assigned to groups for specific tasks this cannot be identified as an experimental study and hence no causality can be inferred

---

Part 2: Research questions

Research quesion 1: Is there any correlation between people who smoked at least 100 cigarettes and use smokeless tobacco products Every day, in having other type of cancer(not skin cancer)? Variables: smoke100,usenow3,chcocncr

Reason for the question: It is believed that people who smoke and use smokeless tobacco are prone to getting cancer, this research question is used to show the statistics in the answers of yes and no with a fair count

Research quesion 2: Is there a correlation between sleeptime, mental health and physical health? Variables: sleptim1,menthlth,physhlth

Reason for the question: It is believed that people who sleep very less or very much are more prone to having bad physical health and bad mental health, this research question is used to show the statistics of whether there is a correlation or not between the entities

Research quesion 3: Can gender(respondent's sex) and income level be correlated with poor physical or mental health ? variables: poorhlth,income2,sex

Reason for the question: The question is to brief the possibility that different genders at different income levels experience different levels of healthiness. It is believed that more males than females experience poor health and it is also believed that higher the salary of a person the less is the chance of the person having poor health. These two things can be easily answered with a detailed plot through data analysis. * * *

Part 3: Exploratory data analysis

NOTE: Insert code chunks as needed by clicking on the "Insert a new code chunk" button (green button with orange arrow) above. Make sure that your code is visible in the project you submit. Delete this note when before you submit your work.

Research question 1:

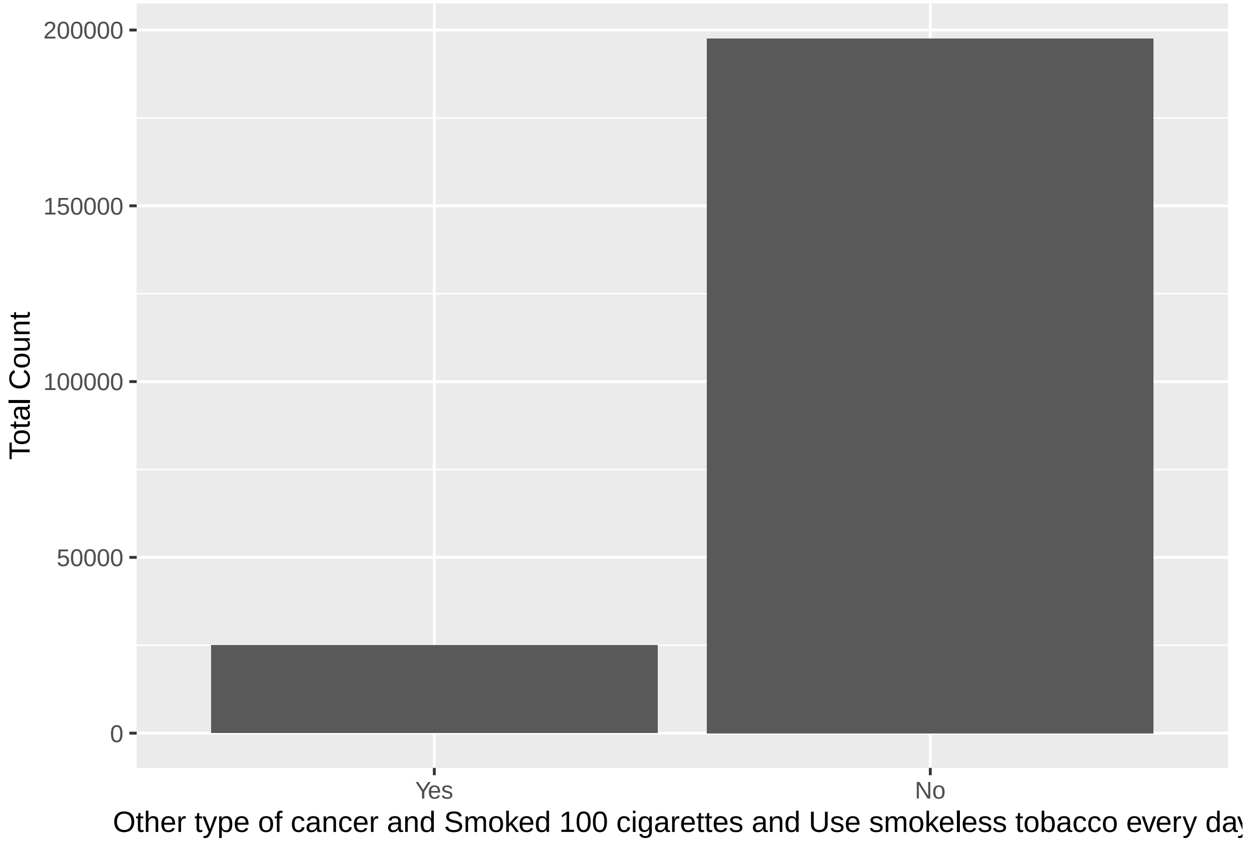
```
sm_100 <- brfss2013 %>%
  filter(!is.na(smoke100),!is.na(usenow3),!is.na(chcocncr))

##summary stats
cancer100 <- sm_100 %>%
  group_by(chcocncr) %>%
  summarise(smoke_and_tob = sum(smoke100 == 'Yes',usenow3 == 'Every day'))

cancer100
```

```
## # A tibble: 2 x 2
##   chcocncr smoke_and_tob
##   <fct>            <int>
## 1 Yes              25011
## 2 No              197638
```

```
##plot
ggplot(data = cancer100,aes(x = chcocncr , y = smoke_and_tob ))+geom_bar(stat = 'identity')+ylab('Total
```



##Narative
As we can see from the given plot and summary stats that citizens of U.S. who smoked 100 cigarettes and use smokeless tobacco everyday got a *11.23%* of chance of getting different types of cancer other than skin cancer which is still a huge amount when generalized to the whole country We can see in the visuals and stats that *25011* out of *222,649* calculated count had other type of cancer Hence we can say that the research question proved to be useful and when generalized shows that if you *smoke at least 100 cigarettes* and *use smokeless tobacco products every day* you have a *11.23%* of chance of getting *cancer other than skin cancer*

Research question 2:

```r
##Creating a small data sample for easy calculations
sampledata = brfss2013[70000:90000,]
sampledata = tbl_df(sampledata)
```

```
## Warning: ‘tbl_df()‘ was deprecated in dplyr 1.0.0.
## i Please use ‘tibble::as_tibble()‘ instead.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```r
##Summary Statistics
stats = summarize(sampledata, "sleptim1", mean(sleptim1, na.rm=T))
colnames(stats) = c("Variable", "mean")
physhlthmean = summarize(sampledata, "physhlth", mean(physhlth, na.rm=T))
colnames(physhlthmean) = c("Variable", "mean")
menthlthmean = summarize(sampledata, "menthlth", mean(menthlth, na.rm=T))
colnames(menthlthmean) = c("Variable", "mean")
```

```r
stats = rbind(stats, physhlthmean,menthlthmean)

stats
```

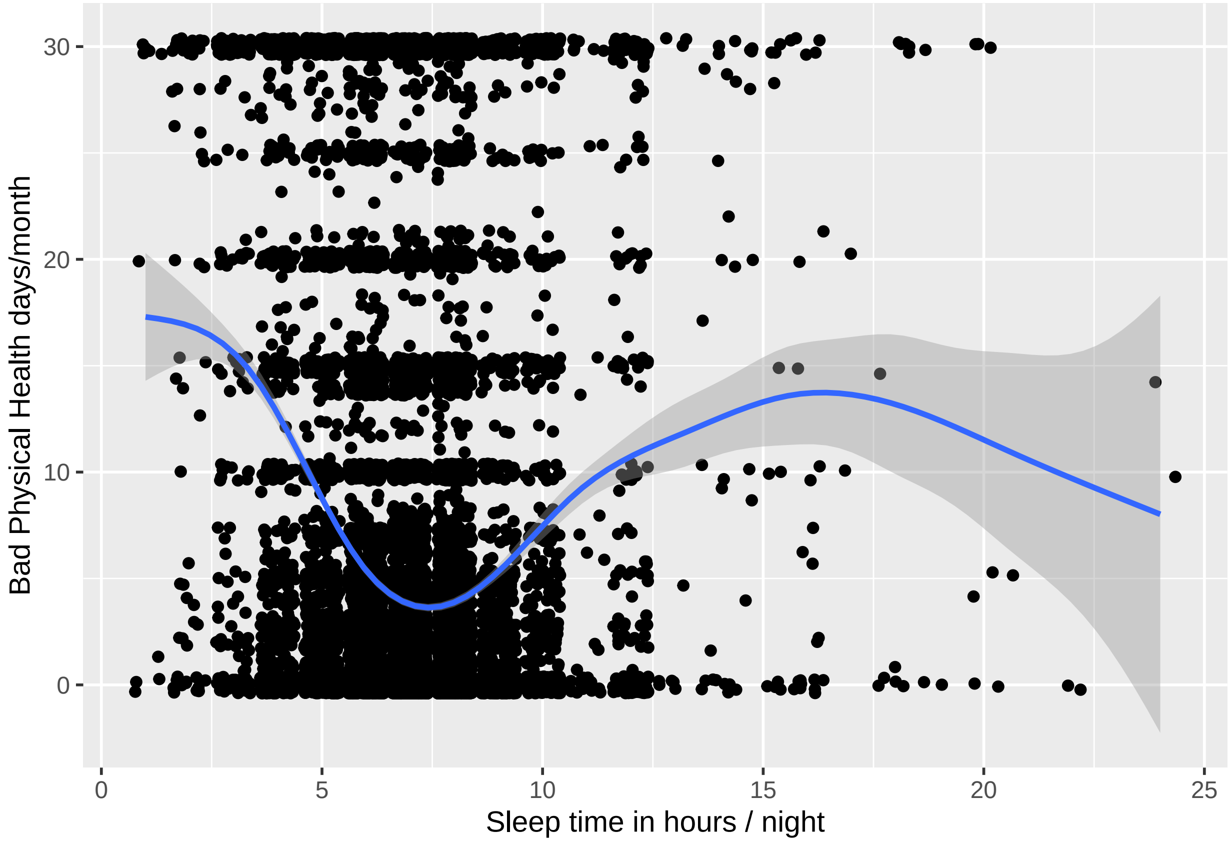
```
## # A tibble: 3 x 2
##   Variable  mean
##   <chr>    <dbl>
## 1 sleptim1  7.09
## 2 physhlth  5.35
## 3 menthlth  3.72
```

```r
##Plot for correlation between sleeptime and physical health
ggplot(sampledata, aes(y= physhlth, x = sleptim1)) + geom_jitter() +geom_smooth()+labs(x="Sleep time in
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1026 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1026 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
##Correlation Value for sleeptime and physical health
cor(sampledata$sleptim1,sampledata$physhlth,use ="complete.obs")
```

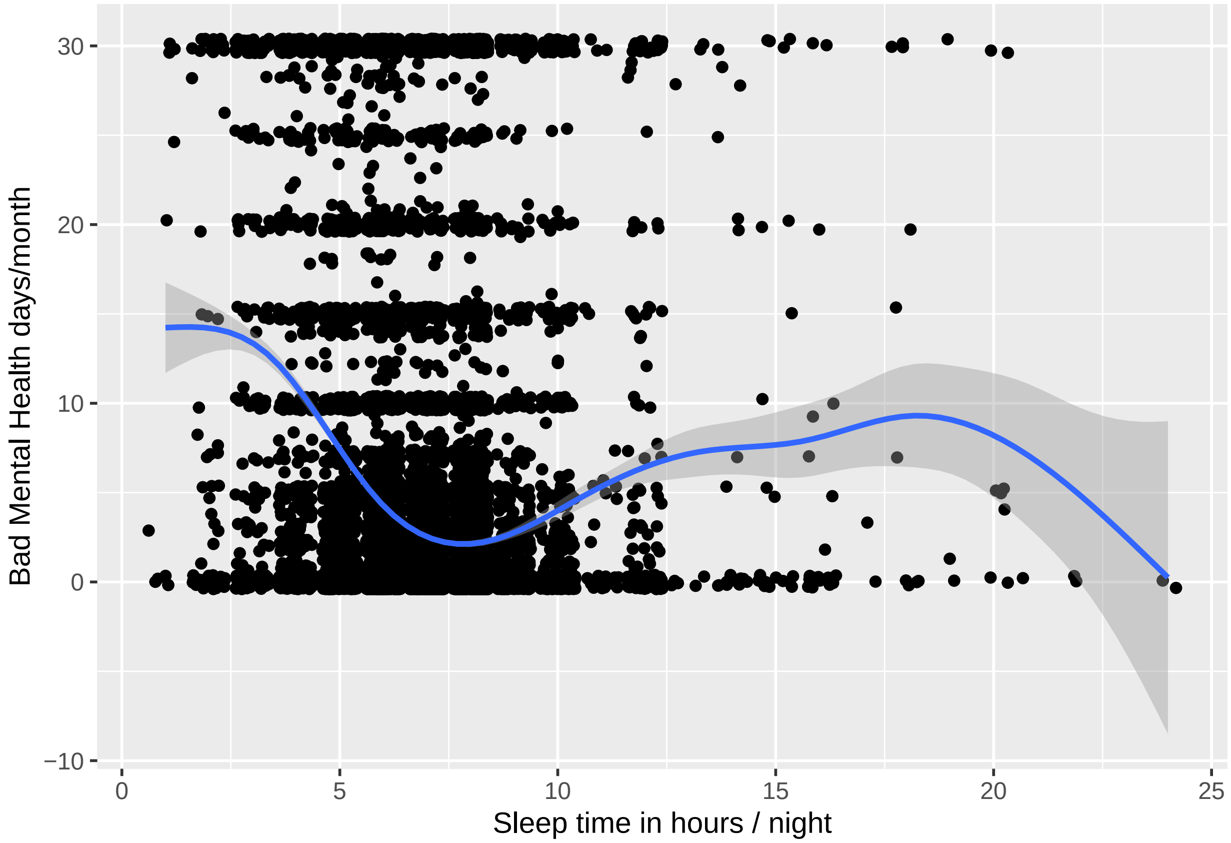
```
## [1] -0.09275033
```

```
##Plot for correlation between sleeptime and mental health
ggplot(sampledata, aes(y= menthlth, x = sleptim1)) + geom_jitter() +geom_smooth()+labs(x="Sleep time in
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 864 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 864 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
##Correlation value for sleeptime and mental health
cor(sampledata$sleptim1,sampledata$menthlth,use ="complete.obs")
```

```
## [1] -0.163537
```

##Narative
In visual representation of physical health and sleeptime and mental health and sleeptime we find that there

is no clear correlation or linear relationship and even the correlation value is negative which is not significant, But in the visual representations we can barely see some dips in plotted regression line which shows that people on both extremes can have higher chances of having bad physical or mental health days per month

Research question 3:

```
healthtable <- brfss2013 %>% filter(!(is.na(sex)), !(is.na(poorhlth)), !(is.na(income2)))
healthtable <- healthtable %>% mutate(poorhlth = ifelse(poorhlth > 8, "9+", poorhlth))
healthtable$poorhlth <- factor(healthtable$poorhlth)

##Summary Statistics
healthtable %>% group_by(income2, sex) %>% summarize(count=n())
```

```
## ‘summarise()‘ has grouped output by ’income2’. You can override using the
## ‘.groups‘ argument.
```

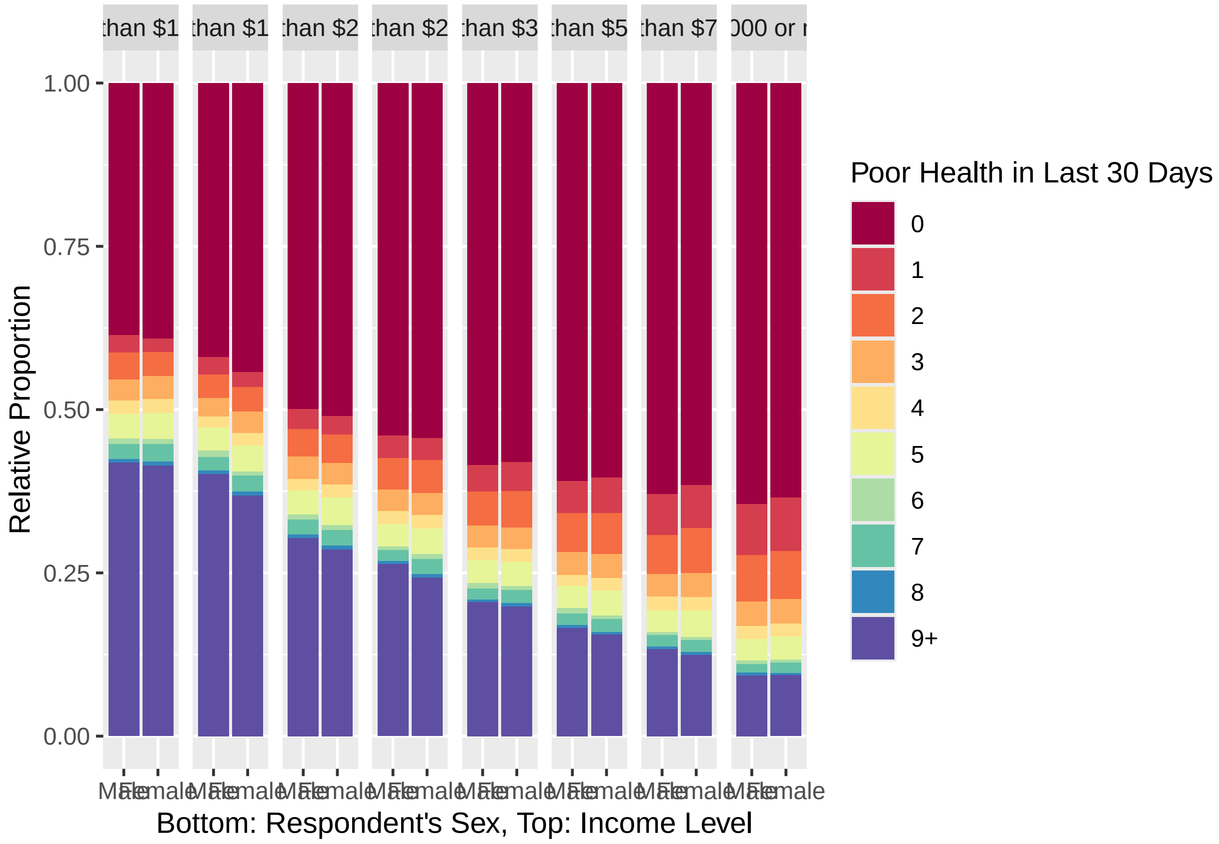
```
## # A tibble: 16 x 3
## # Groups:   income2 [8]
##    income2           sex    count
##    <fct>             <fct>  <int>
##  1 Less than $10,000 Male    5402
##  2 Less than $10,000 Female 11819
##  3 Less than $15,000 Male    5708
##  4 Less than $15,000 Female 11683
##  5 Less than $20,000 Male    7176
##  6 Less than $20,000 Female 13550
##  7 Less than $25,000 Male    8251
##  8 Less than $25,000 Female 15054
##  9 Less than $35,000 Male    9397
## 10 Less than $35,000 Female 15778
## 11 Less than $50,000 Male   11888
## 12 Less than $50,000 Female 18124
## 13 Less than $75,000 Male   12354
## 14 Less than $75,000 Female 18558
## 15 $75,000 or more   Male   21116
## 16 $75,000 or more   Female 28900
```

```
healthtable %>% group_by(poorhlth, sex) %>% summarize(count=n())
```

```
## ‘summarise()‘ has grouped output by ’poorhlth’. You can override using the
## ‘.groups‘ argument.
```

```
## # A tibble: 20 x 3
## # Groups:   poorhlth [10]
##    poorhlth sex    count
##    <fct>    <fct>  <int>
##  1 0        Male   46658
##  2 0        Female 74760
##  3 1        Male    4197
##  4 1        Female  6690
##  5 2        Male    4558
##  6 2        Female  7644
##  7 3        Male    2799
##   8 3         Female  4692
##   9 4         Male    1570
##  10 4         Female  2697
##  11 5         Male    2796
##  12 5         Female  5106
##  13 6         Male     526
##  14 6         Female   828
##  15 7         Male    1421
##  16 7         Female  2752
##  17 8         Male     354
##  18 8         Female   648
##  19 9+        Male   16413
##  20 9+        Female 27649
```

```
##Plot
plot <- ggplot(healthtable) + aes(x = sex, fill = poorhlth) + geom_bar(position = "fill") + facet_grid(
plot <- plot + scale_fill_brewer(name="Poor Health in Last 30 Days", palette = "Spectral") + xlab("Botto
plot
```



##Narative
The research question has been answered,As we can see from the given plot that people with higher salaries have less days of poor health. In the middle segment of the graph of the income level we can clearly see that when it comes to genders, more males experience poor health than women when given the same amount of salary just by a little bit though. We can clearly notice the difference between people who are given high income and people who are given low income. We can also see that females are offered lower salaries in the summary statistics.