CSE3019: Data Mining

Exploring the BRFSS data

Jacob John (16BCE2205)
Harish Narasimhan A (16BCE0637)

24 March 2018

Setup

Load packages

```r
library(ggplot2)        #Library to create plots
library(corrplot)       #Library to create correlation plot
library(dplyr)          #Data Manipulation
library(maps)           #For map data
```

Load data

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

Introduction

The goal of this project is to identify three research question similar to questions talked about in the CSE3019: Data Mining Course. All answers will be produced only using the Behavioral Risk Factor Surveillance System (BRFSS) dataset.

Part 1: Data

Background

The Behavioral Risk Factor Survelliance System (BRFSS) is a project adminstered by CDC's Population and Health And Survelliance Branches. It boasts more than 400,000 interviews annually which makes it the "largest continuously conducted health survey system". The BRFSS aims to collect data through a series of annual telephone surveys. With the exception of American Samoa, Federated States of Micronesia, and Palau that "collect data over a limited point-in-time (usually one to three months)". This project is intended to identify health associated, risk factors among non-instituationalized adults (by definition, personnels over the age of 18) residing in the US and recognize emerging trends.

Data collection

Participants are questioned about their health practices on the basis of their, "tobacco use, HIV/AIDS knowledge and prevention, exercise, immunization, health status, healthy days — health-related quality of life, health care access, inadequate sleep, hypertension awareness, cholesterol awareness, chronic health conditions, alcohol consumption, fruits and vegetables consumption, arthritis burden, and seatbelt use" as according to the BRFSS Codebook.

Note: The dataset contains 330 variables for a total of 491775 observations as of 2013. The missing values are primarily denoted by "NA".

Generalizability

Disproportionate stratified sampling (DSS) has been used for the telephone samples (with the expeption of Guam and Puerto Rico, who used random sampling), with the use of Random Digit Dialing (RDD) techniques for both landline and cellular telephone users. (Source) Furthermore, also due to the breadth of the survey (50 states and 500,000 observations), it is safe to state that the study's result captures enough of a random sample to make it generalizable at large.

Biases

There is a possibility of underreporting the data or causing a population bias. As according to the Survey, 2.5% of the population doesn't have access to landlines or celluar based telecommunication services for interviews. Hence, households without access to telephones lines aren't represented in the sample. To some extent, it can be hence stated that the study cannot be generalized to these households.

On average, the primary segment of the survey lasts for approximately 18 minutes as per BRFSS. Despite the fact that calls were made throughout the week, during mornings and evening hours, it introduces a non-response bias. It can be said that unavailablity or incomplete responses (e.g. by work professionals, due to lack of time) upon RDD could further establish this bias and add to the underreporting of data.

Since the responses aren't validated, answers for alcohol consumption, tobacco consumption, health care awarness, etc could be negatively notated due to reponse bias. Some participants might feel pressure to give answers that are socially acceptable, hence cause discrepancy in the data. Furthermore, overreporting of desirable traits such as height, exercise, etc and similarily underreporting of undesirable traits such as weight, alcohol consumption also alter responses. Another factor could be the difficulty in recalling information for responses that require details from up to 30 days ago.

The NAs present in the dataset represent missing values that hence state inconsistency accross variables (or questions sets) among respondants. One possibility could be nonresponse error or because candidates refused to give sensitive/personal information.

Causality

Since the given study is an observational exercise, we can only only establish an association or correlation. In addition to this, there is no explicitly stated random assignment to treatments nor was it confucted in an experimental setting. Thus, a causation cannot be inferred with the given dataset.

References

A list of the surveyed questions, codebooks and website for the aforementioned information are given as follows:

- BRFSS web site
- BRFSS Questionnaire (Mandatory and Optional Modules)
- BRFSS Codebook
- BRFSS Guide to Calculated Variables
- BRFSS Guide to Optional Modules Used, by State

Part 2: Research questions

Research quesion 1:

> Is there any association between health care coverage of an individual and the number of days "poor" physical or mental health affects their usual activities?

The analysis used the following variables:

- *poorhlth*: Poor Physical Or Mental Health
- *hlthpln1*: Have Any Health Care Coverage

Research quesion 2:

> What factor significantly contributes to an overall "Poor" general health in the "sickest" state?

The following variables will be used:

- *genhlth*: General Health
- *medcost*: Could Not See Dr. Because Of Cost
- *X_State*: State Fips Code
- *hlthpln1*: Have Any Health Care Coverage
- *drnk3ge5*: Binge drinking
- *checkup1*: Length Of Time Since Last Routine Checkup
- *employ1*: Employment Status
- *toldhi2*: Ever Told Blood Cholesterol High

Research quesion 3:

> Does a correlation exist between heart attacks to high cholestrol, strokes, bmi and drinking?

Variables to be used:

- *toldhi2*: Ever Told Blood Cholesterol High
- *cvdinfr4*: Ever Diagnosed With Heart Attack
- *cvdstrk3*: Ever Diagnosed With A Stroke
- *bmi5*: Computed Body Mass Index
- *maxdrnks*: Most Drinks On Single Occasion Past 30 Days

Part 3: Exploratory data analysis

Research quesion 1

Q: Is there any association between health care coverage of an individual and the number of days "poor" physical or mental health affects their usual activities?

The attribute *poorhlth* answers the question, "During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities?"

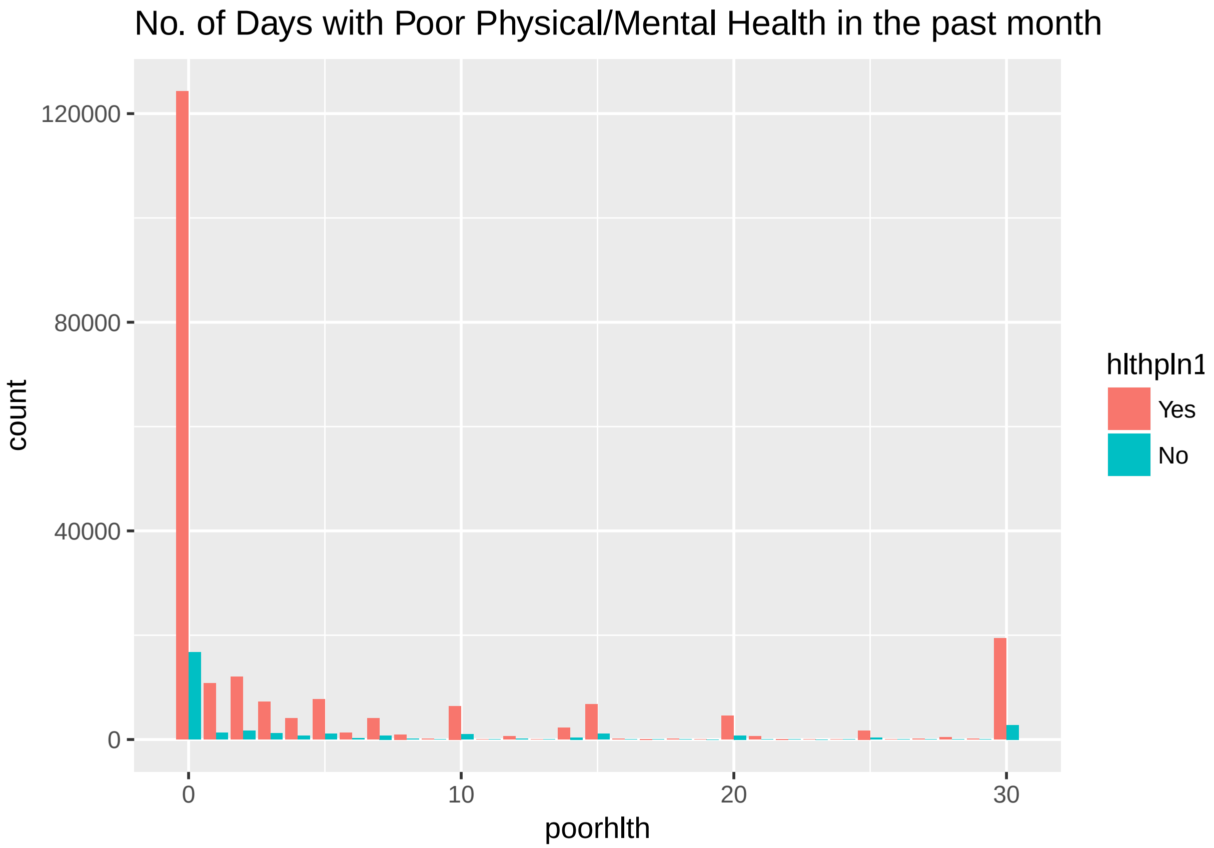
While the attribute *hlthpln1* is a response to, "Do you have any kind of health care coverage?"

The first step to perform data cleaning by filtering the data by omitting all the na values using the *na.omit()* function.

```
health <- select(brfss2013,poorhlth,hlthpln1) %>% #selecting only required variables
        na.omit()                                  #omitting all the Na values
```

Now, given whether someone has a Health Plan, how many respondants felt that poor physical or mental health kept them from doing their usual activities? This can be represented with a bar plot with the count on the y-axis and poorhlth or the x-axis.

```
ggplot(aes(x=poorhlth, fill=hlthpln1), data = health) +
        geom_bar(position=position_dodge()) +
        #Visualizing data using a bar plot
        ggtitle('No. of Days with Poor Physical/Mental Health in the past month')
```



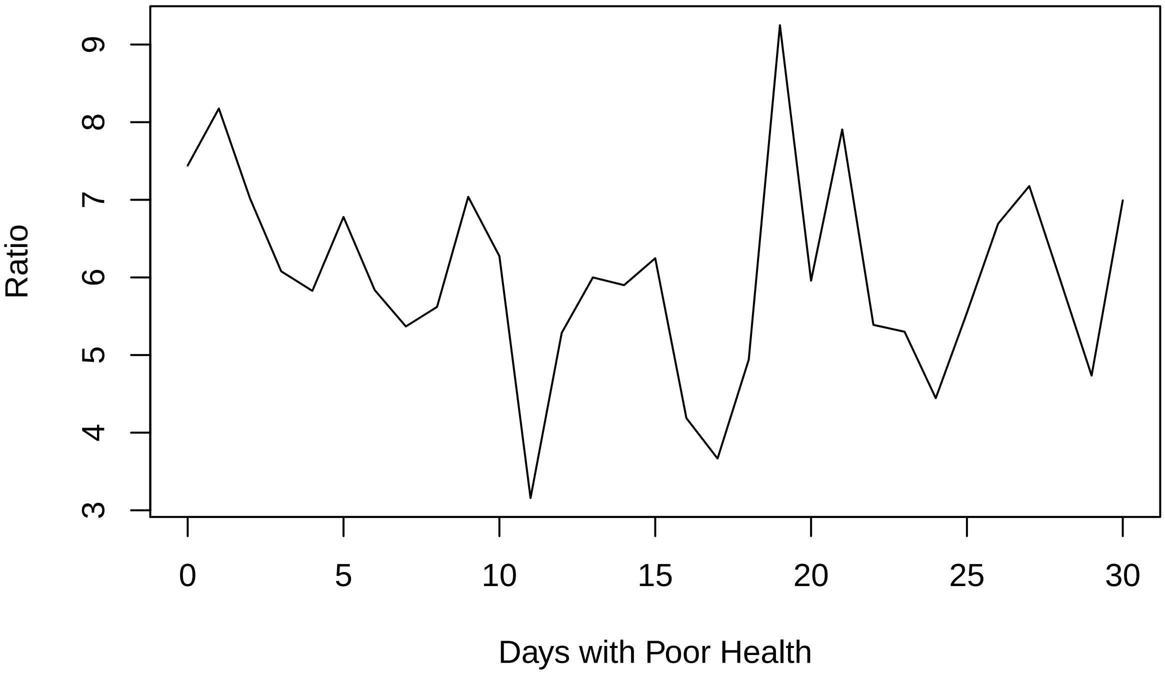
With an emergent peak at 0 for hlthpln1 = “Yes”, it is clearly evident that people who have a health care plan are in better physical and mental health as their usual activities are least affected (or not affected at all) by a poor mental or physical health.

Therefore, it implies that there may exist a correlation between having a health care plan and having 0 days when a poor mental or physical health impacts an individual’s acitivities. With the given conclusion, it could be further investigated whether a health care plan have a direct impact on physical or mental health.

Below, we will be taking the ratios between the people with poor health against those without. This is in order to verify the peak at “0” days.

```
#Filtering count by health plan = Yes
hlth_yes <- select(brfss2013,poorhlth,hlthpln1) %>%
        na.omit() %>%
        filter(hlthpln1=="Yes") %>%
        group_by(poorhlth) %>%
        summarize(count=n())

#Filtering count by health plan = No
hlth_no <- select(brfss2013,poorhlth,hlthpln1) %>%
        na.omit() %>%
        filter(hlthpln1=="No") %>%
        group_by(poorhlth) %>%
        summarize(count=n())
```

```r
#Ratio between those who have a health plan and those who don't by days
ratio_ds <- data.frame(ratio = (hlth_yes$count[hlth_yes$poorhlth==hlth_no$poorhlth]
                    /hlth_no$count[hlth_yes$poorhlth==hlth_no$poorhlth]),
           poor_hlth = (hlth_yes$poorhlth[hlth_yes$poorhlth==hlth_no$poorhlth]))

#Plotting ratios against days with poor health
plot(x=ratio_ds$poor_hlth,y=ratio_ds$ratio,type="l",
     xlab="Days with Poor Health",ylab="Ratio")
```



The values lie between the ratios “3” and “9”. The trend is noisy and its is neither exhibits a positive nor a negative correlation.

Therefore, with the given data it, there exists no apparent trend between having a health plan and the number of days their daily activity is affected by poor mental or physical health.

Conclusion

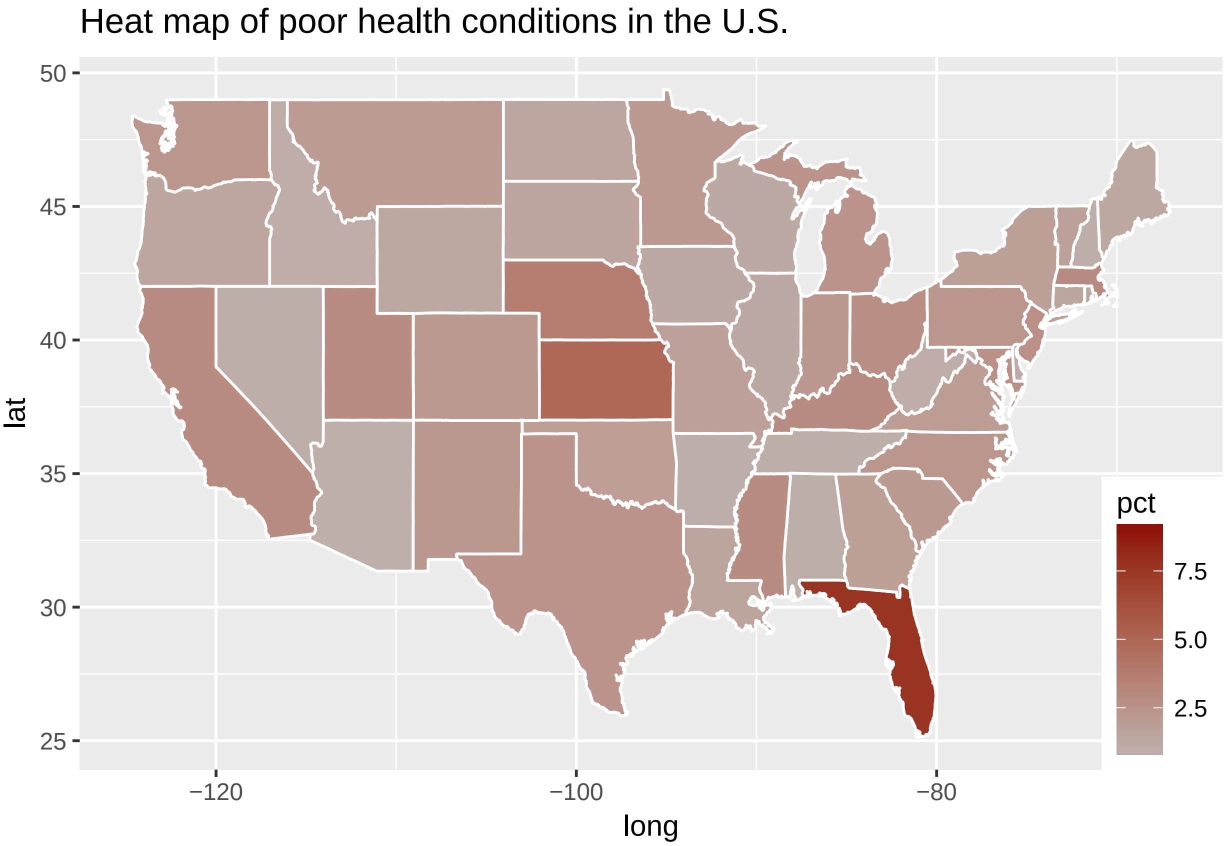
Research quesion 2

Q: What factor significantly contributes to an overall “Poor” general health in the “sickest” state?

What state is sickest and what is fittest? A Heatmap could be usefull!

```r
# ggplot2 function for heatmapping health status
all_states <- map_data("state")

health_states <- brfss2013 %>%
    filter(!is.na(genhlth)) %>% # omit on NA values!
        select(genhlth, X_state, poorhlth)
```

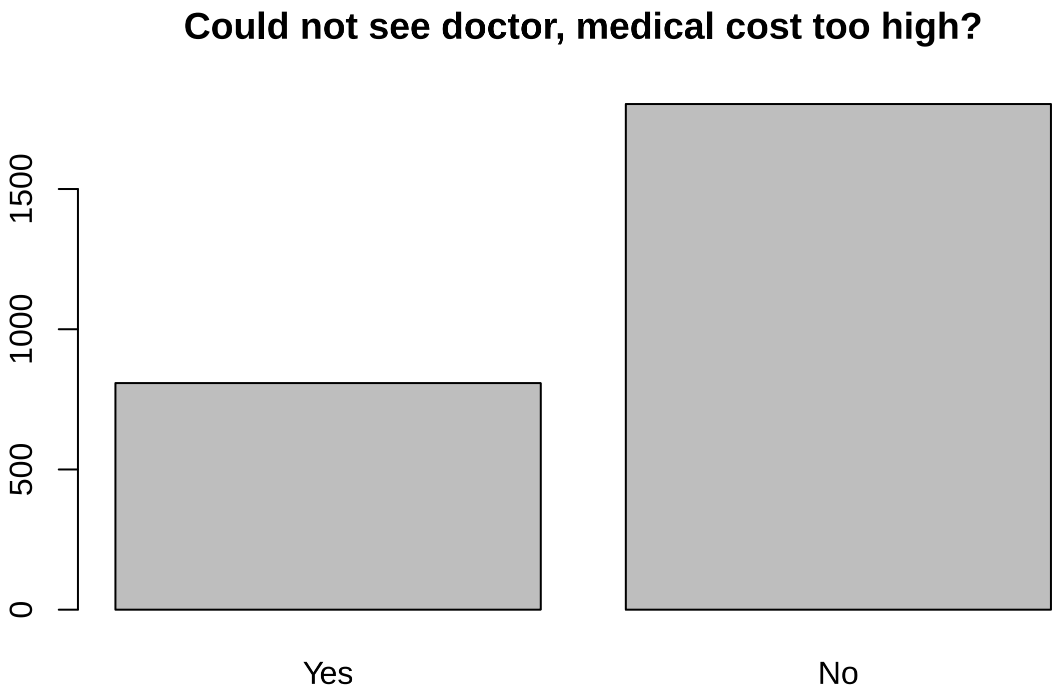
```r
health_states_poorhlth <- health_states %>%
    group_by(genhlth, X_state) %>%
    summarise(mean(poorhlth),n=n()) %>%
    mutate(pct = (n/sum(n))*100)

health_states_map <- health_states_poorhlth %>%
    mutate(region= tolower(X_state))

states_map <- merge(all_states, health_states_map, by="region")

# plotting map...
ggplot() +
  geom_polygon(data = states_map, aes(x= long, y = lat, group = group,
                                      fill = pct), color = "white") +
  ggtitle("Heat map of poor health conditions in the U.S.") +
  scale_fill_gradient2(low = "blue", mid = "grey", high = "darkred") +
  theme(legend.position = c(1, 0), legend.justification = c(1, 0))
```



I am using poorhlth variable to plot this heat map, which means that there is higher number of people with poor health in Florida. Why?

We can check if the respondants are unable to see a doctor due to high medical costs by plotting a bar plot for the variable *medcost* which is a answer to the question, “Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?”. This will be used to determine whether the respondants feel that the medical costs are too high.

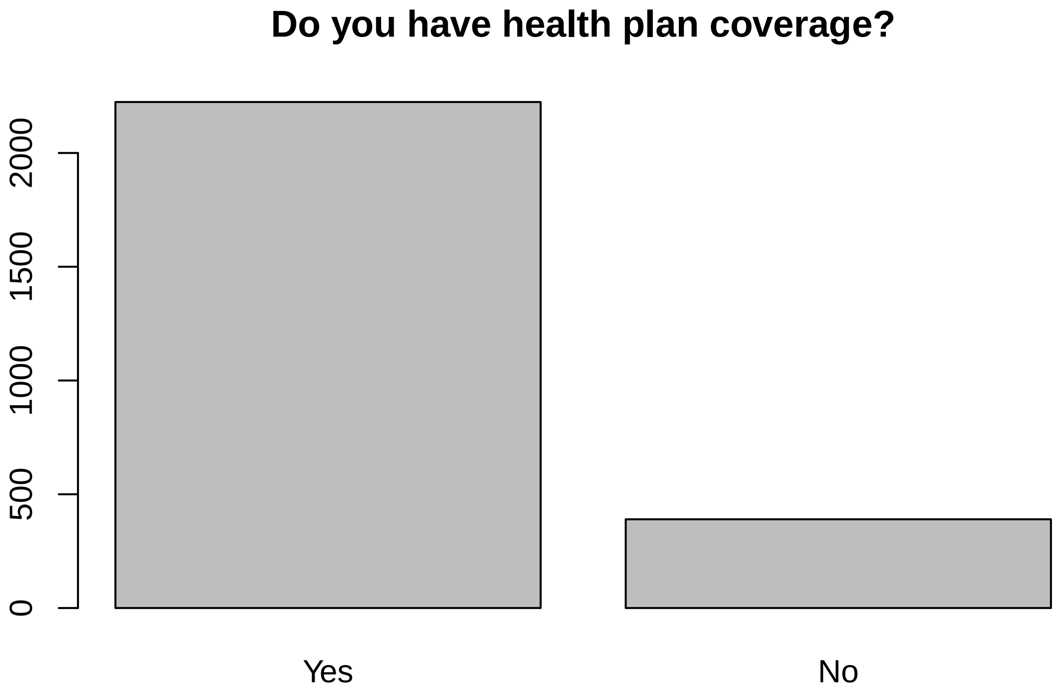
```
plot(brfss2013 %>%
       filter(!is.na(medcost), X_state=="Florida",genhlth=="Poor") %>%
       select(medcost), main="Could not see doctor, medical cost too high?"
     )
```



Majority of the responses feel that the medical costs do not prevent them from seeing the doctor. Therefore, medical costs does not play a significant role in poor health care.

Now we will check whether having a health plan coverage impacts poor health using the *hlthpln1* variable. This is similar to what we did in question one. However, here we're simply plotting the frequncy of "Yes" against "No".

```
plot(brfss2013 %>%
       filter(X_state=="Florida",genhlth=="Poor") %>%
       select(hlthpln1), main="Do you have health plan coverage?"
     )
```



As given the plot above, majority of the residents have a health care plan. Hence, we do not consider this variable

Now to test whether drinking causing an impact. In order to explore this, we investigate the variable *drink3ge5* which answers, "how many times during the past 30 days did you have 5 or more drinks (men) or 4 or more drinks (women) on an occasion?" to check whether binge drinking causes poor health in Florida.

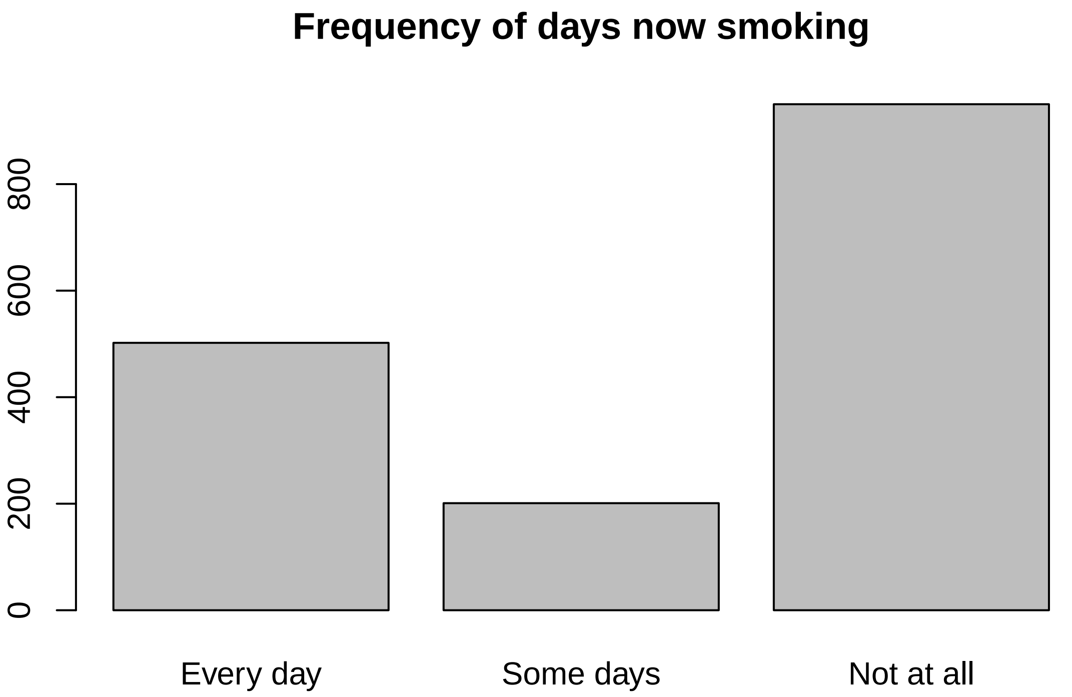
```
# drinking
summary(brfss2013 %>%
          filter(!is.na(drnk3ge5), X_state=="Florida",genhlth=="Poor") %>%
          select(drnk3ge5)
        )
```

```
##     drnk3ge5
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.000
##  Mean   : 1.675
##  3rd Qu.: 0.000
##  Max.   :30.000
```

Since the mean of the result is only 1.675 with a Q1 and Q2 of 0. The residents of Florida aren't heavy drinkers.

In order to test whether smoking impacts their "poor" health, the variable *smokday2* is used. This answers, "Do you now smoke cigarettes every day, some days, or not at all?". Thus enabling us to understand the prevalence of smoking by visualizing using a bar plot.

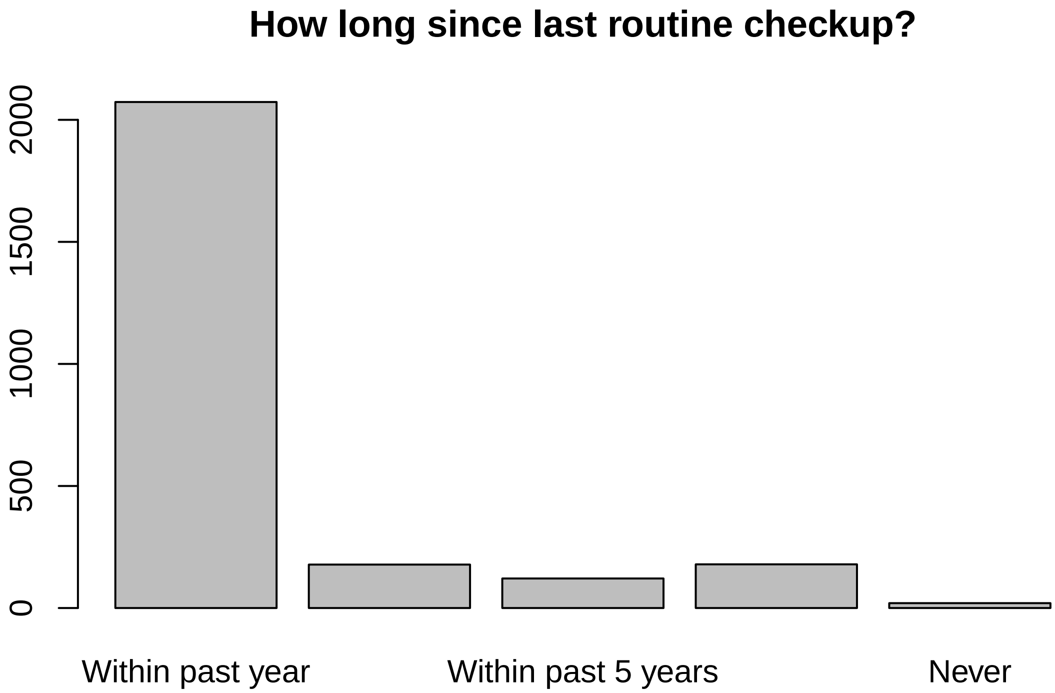
```
plot(brfss2013 %>%
       filter(X_state=="Florida",genhlth=="Poor") %>%
       select(smokday2), main="Frequency of days now smoking"
     )
```



Since “Not at all” comprised of majority of the responses, then they are not heavy smokers.

Using the *checkup1* attribute we can identify the “length of time since last routine checkup”. This helps us understand whether the candidates go for medical exams often and whether this impacts “poor” health.

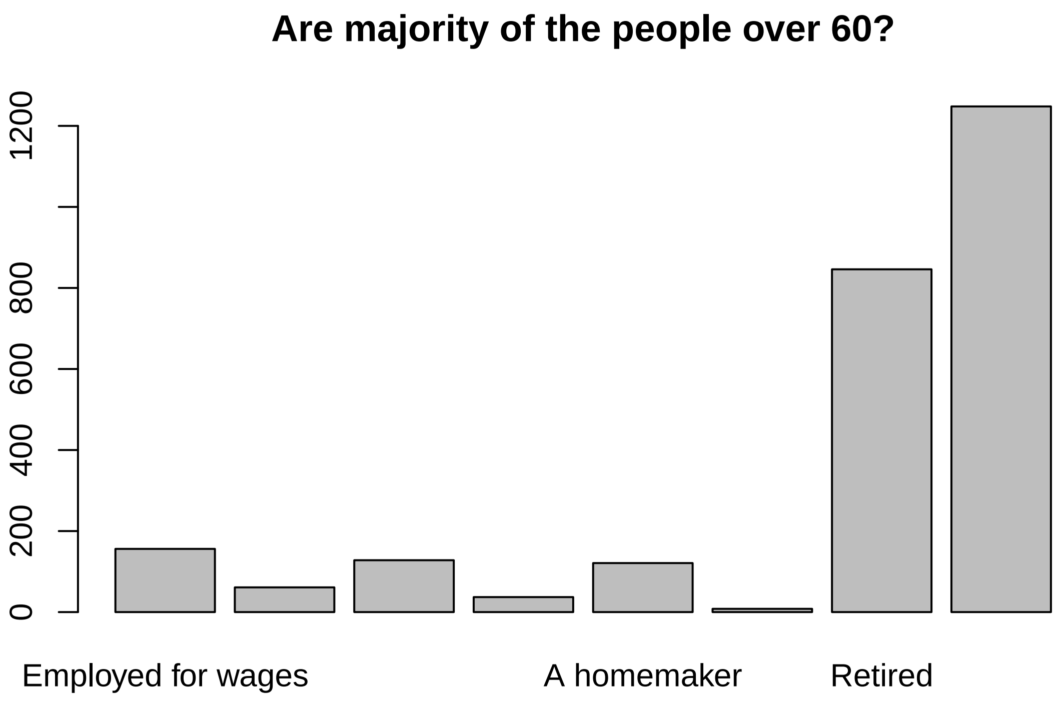
```
plot(brfss2013 %>%
       filter(X_state=="Florida",genhlth=="Poor") %>%
       select(checkup1), main="How long since last routine checkup?"
     )
```



Since a substaintial number of participants have gone for a medical check up within the past year, it doesn't signify "poor" health.

To test whether age plays a role in "poor" health, rather than testing the number of respondants per age, we use the categories presented by the employment status or *employ1* variable. It is obvious that anyone in the "Retired" category is going to be over the age of 60. Our goal here is to check whether this plays a role in "poor" health.

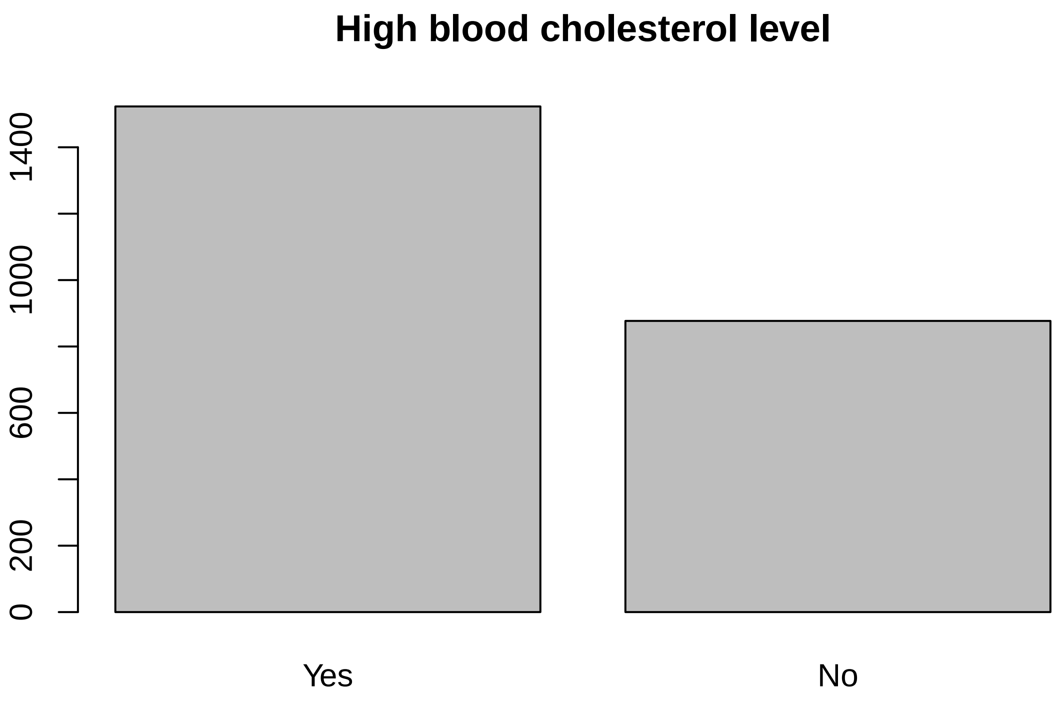
```
plot(brfss2013 %>%
       filter(X_state=="Florida",genhlth=="Poor") %>%
       select(employ1), main="Are majority of the people over 60?"
     )
```



Yes, several of the candidates are retired.

Now, to check if blood cholestrol level also impacts "poor" health, we use the *toldhi2* variable. This tells us whether the respondant has "ever been told by a doctor, nurse or other health professional that their blood cholesterol is high".

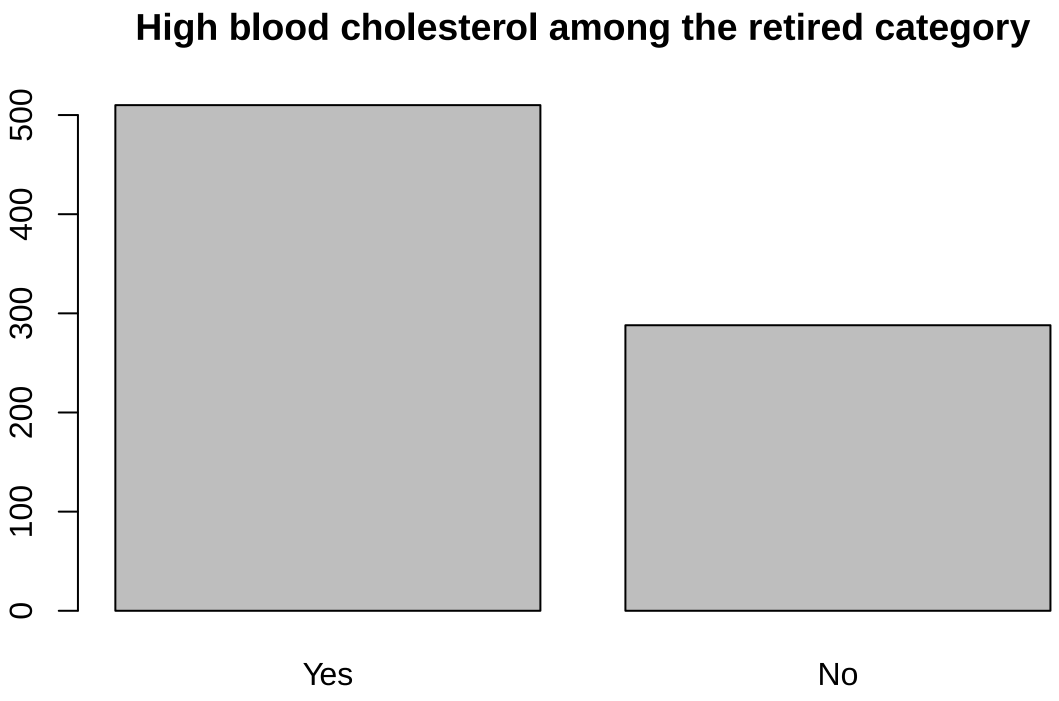
```
plot(brfss2013 %>%
       filter(X_state=="Florida",genhlth=="Poor") %>%
       select(toldhi2), main="High blood cholesterol level"
     )
```



Although not evident, there are several people with high blood cholestrol in Florida.

Now we can explore whether blood cholestrol issues are mainly been faced by retired personnel by plotting the filtered data.

```
plot(brfss2013 %>%
             # only selecting those with "Poor" general health and "Retired"
       filter(X_state=="Florida", employ1=="Retired",genhlth=="Poor") %>%
       select(toldhi2), main="High blood cholesterol among the retired category"
     )
```



```
# Summarizing the data
summary(brfss2013 %>%
       filter(X_state=="Florida", employ1=="Retired",genhlth=="Poor") %>%
       select(toldhi2)
     )
```

```
##  toldhi2
##  Yes :510
##  No  :288
##  NA's: 48
```

Hence, There are several retired people (adults over 60) in Florida with a high level of cholestrol. This plays a significant role in the overall "poor" general health of Florida.

Conclusion

From the derived heat map, it can be inferred that Florida is the sickest state in the US. Thereon, we understand that medical cost is not a reason behind the poor health of the population of Florida. The third graph tells us that most of the people have a health plan coverage. The fourth graph tells us that most of the people in Florida are non-smokers. Most of the people have done a routine check-up within the past year. Most of the people above age 60 are retired. The seventh graph states that several people in Florida suffer from high blood cholesterol. With 60 being set as the age of a retired citizen, we can see that many retired people have high blood cholesterol. Therefore, it is evident that due to the number of elderly in Florida, the number of "poor" general health responses deem it be a relatively sicker state.

Research quesion 3

Q: Does a correlation exist between heart attacks to high cholestrol, strokes, bmi and drinking?

To identify whether a participant has been diagnosed with a stroke, the *cvdinfr4* variable will be used that asks them whether they've "ever been diagnosed with a heart attack, also called a myocardial infarction".

Furthermore, the following variables are also used for correlation:

- *toldhi2*: Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high?
- *cvdstrk3*: (Ever told) you had a stroke.
- *X_bmi5*: Computed Body Mass Index
- *maxdrnks*: During the past 30 days, what is the largest number of drinks you had on any occasion?

First, we need to remove all the NA values using *filter()*.

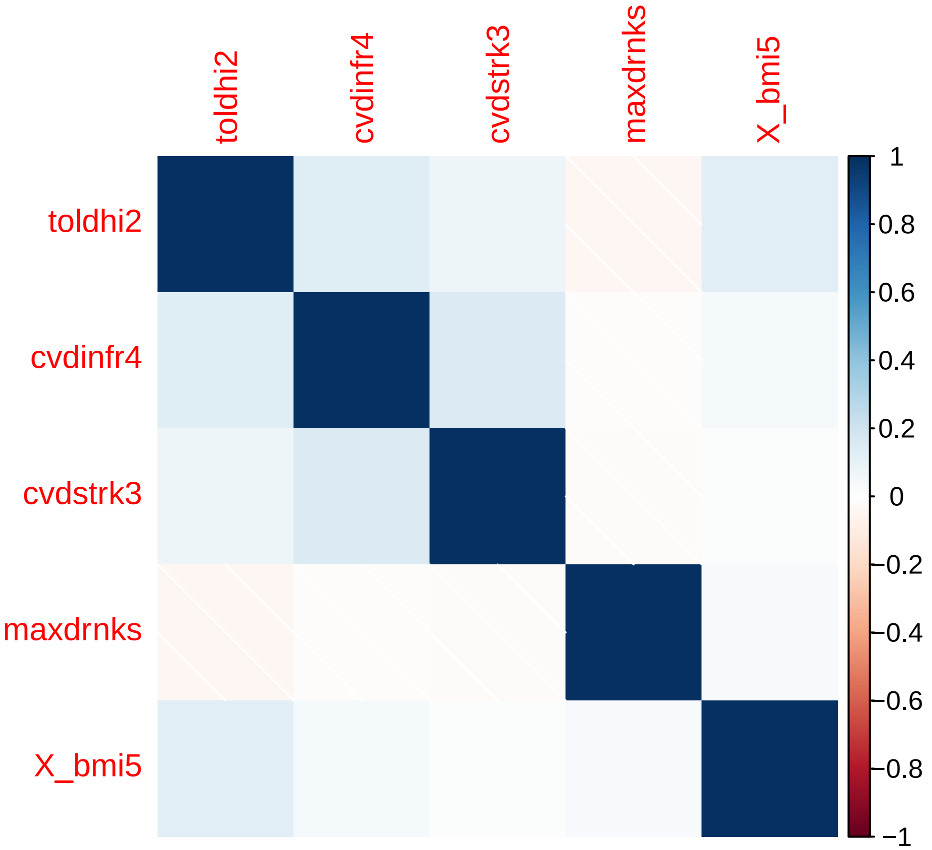
```
hr_strk <- brfss2013 %>%
        # omitting NAs
        filter(!is.na(toldhi2), !is.na(cvdinfr4), !is.na(cvdstrk3),
                !is.na(X_bmi5), !is.na(maxdrnks)) %>%
        select(toldhi2,cvdinfr4,cvdstrk3,maxdrnks,X_bmi5)
```

Since the values for *toldhi2*, *cvdstrk3* and *cvdinfr4* are categorical and comprise of “Yes” and “No” responses, we need to coerece them to TRUEs and FALSEs to allow for a correlation plot.

```
# Since the data set contains only Yes and No, we coerce them to TRUE and FALSE
hr_strk <- hr_strk %>%
  mutate(toldhi2 = ifelse(hr_strk$toldhi2=="Yes",TRUE,FALSE))
hr_strk <- hr_strk %>%
  mutate(cvdstrk3 = ifelse(hr_strk$cvdstrk3=="Yes",TRUE,FALSE))
hr_strk <- hr_strk %>%
  mutate(cvdinfr4 = ifelse(hr_strk$cvdinfr4=="Yes",TRUE,FALSE))
```

We then construct the correlation plot using *corrplot()*.

```
strk_cor<-cor(hr_strk)
corrplot(strk_cor, method="shade")
```



There is an extremely weak correlation between heart attacks and drinking in the last 30 days. This correlation coefficient lays around 0. Furthermore, another weak correlation exists between heart attacks and body mass index of an individual, however, this value is approximately 0.2. Lastly, there is a strong correlation between heart attacks and blood cholesterol and heart attacks and having a stroke in the past. Therefore, it can be said that having a stroke or a high blood cholesterol may increases chances of having an heart attack.

In summary,

```
table(data.frame(hr_strk$toldhi2,hr_strk$cvdstrk3,hr_strk$cvdinfr4))
```

```
## , , hr_strk.cvdinfr4 = FALSE
##
##                 hr_strk.cvdstrk3
## hr_strk.toldhi2  FALSE   TRUE
##           FALSE 107582   1470
##           TRUE   68425   2372
##
## , , hr_strk.cvdinfr4 = TRUE
##
##                 hr_strk.cvdstrk3
## hr_strk.toldhi2  FALSE   TRUE
##           FALSE   2136    347
##           TRUE    5071    865
```