Exploring the BRFSS data: Therese Smit

Setup

Load packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.4.1
```

```
library(statsr)
```

Load data

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

Part 1: Data

The observations in the sample are collected via telephone surveys. This is an observaional study, since the data is collected in such a way as to not directly interfere with how the data arises. Stratified random sampling was used, since the population is divided by states (i.e. the states in United States of America), and then randomly sampled from within each state. Since the data is collected via telephone survey, there is a probability that there exists sampling bias of non-response. The study contains random sampling, however there is no random assignment. Therefore the study is generalisable, and not causal.

Part 2: Research questions

Research quesion 1: Which state has the highest rating in terms of general health status?
Finding out general health status of each state is of importance for the government e.g. where funding should go.

Research quesion 2: What is the relationship between the general health status of veterans and their ability to work?
Veterans unable to work cannot pay for health services, so benefits should be given to them.

Research quesion 3: What is the relationship between the education level and the income for each gender?
Discrimination in the work place based on gender, especially when the education level is the same, needs to be prevented.

Part 3: Exploratory data analysis

Research quesion 1:

Method: A new, empty dataframe was created. The raw data was reduced to the necessary variables, and cleaned so as to avoid complications by removing rows with NA data.
The frequency of the general health status of each state was determined and then mode of the frequencies was extracted to determine the highest rating general health status.

Interpretation of Results:

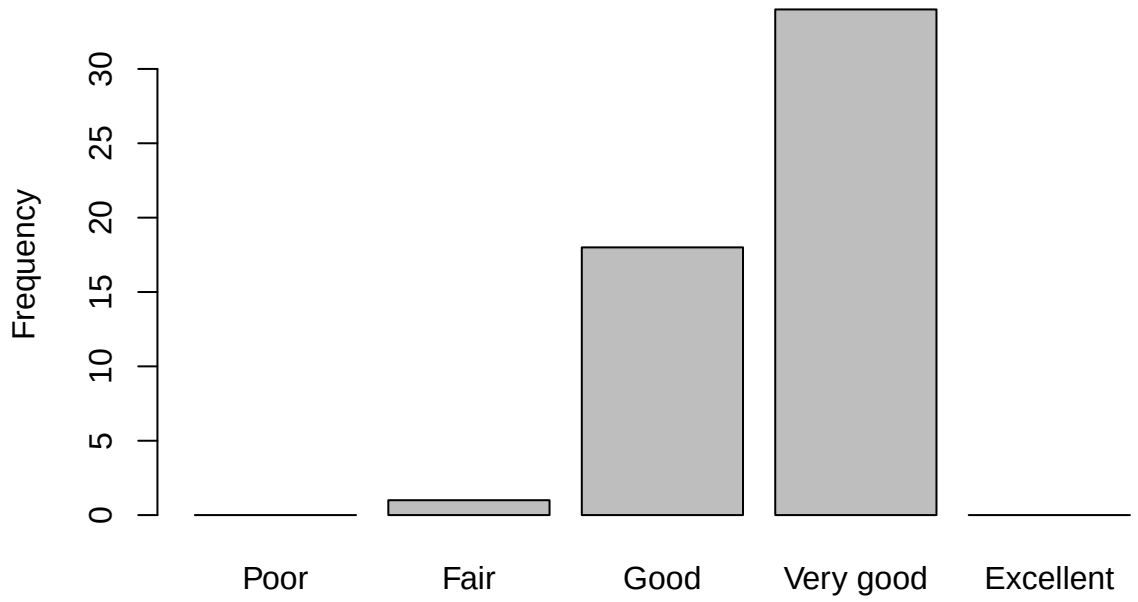
```
result <- data.frame('State'=character(), 'genhlth_mode'=character(), 'genhlth_perc'=numeric(), stringsA

x <- brfss2013[,c("X_state","genhlth")]
mydata <- na.omit(x)

state_names <- unique(mydata$X_state)
cnt <- 1
for(state in state_names) {

  y <- subset(mydata, X_state == state)
  genhlth_freq <- count(y,'genhlth')
  max_value <- genhlth_freq[which.max( genhlth_freq$freq),1]
  max_perc <- max(genhlth_freq$freq)/sum(genhlth_freq$freq)
  result[cnt,] <- c(state, as.character(max_value), as.numeric(max_perc))
  cnt <- cnt + 1
}
```

A bar graph is used to show the distribution of this frequency.

```
rdata = factor(result$genhlth_mode,levels=c('Poor','Fair','Good', 'Very good', 'Excellent'))
summary(rdata)
```

```
##      Poor      Fair      Good Very good Excellent
##         0         1        18        34         0
```

```
barplot(table(rdata), main = "The Distribution of the General Health Mode Across all States", ylab = "Fr
```



A table was then created to order the data in such a way that the state with the highest general health status could be determined.

```
ordered_result <- result[order(result[,2],result[,3]),]
ordered_result
```

```
##                   State genhlth_mode      genhlth_perc
## 53          Puerto Rico         Fair 0.345299716146268
## 43            Tennessee         Good 0.302852203975799
## 10              Florida         Good 0.307761597357754
## 41       South Carolina         Good 0.310209055967001
## 11              Georgia         Good 0.310704355885079
## 34       North Carolina         Good 0.310751104565538
## 19            Louisiana         Good 0.312980030721966
## 25          Mississippi         Good 0.315328663793103
## 18             Kentucky         Good 0.315590754213095
## 26             Missouri         Good 0.316063635083767
## 36                 Ohio         Good 0.317671691792295
## 49        West Virginia         Good 0.320645709430756
## 32           New Mexico         Good 0.321570736955352
## 4              Arkansas         Good 0.327865725729544
## 1               Alabama         Good 0.333642118264629
## 37             Oklahoma         Good 0.334349098000975
## 44                Texas         Good 0.339279043705641
## 12               Hawaii         Good 0.356687898089172
## 52                 Guam         Good  0.37710970464135
## 3               Arizona    Very good 0.314002828854314
```

```
## 5            California    Very good 0.314476885644769
## 33             New York    Very good 0.325118163403106
## 31           New Jersey    Very good 0.325445941026574
## 8              Delaware    Very good 0.326538461538462
## 15              Indiana    Very good 0.326572604350382
## 22        Massachusetts    Very good 0.326951203297434
## 47             Virginia    Very good 0.330209939508955
## 13                Idaho    Very good 0.331368345837799
## 27              Montana    Very good 0.331817242092206
## 14             Illinois    Very good 0.332738626226583
## 9  District of Columbia    Very good 0.334756097560976
## 39         Pennsylvania    Very good 0.335447498238196
## 29               Nevada    Very good 0.335561234519363
## 40         Rhode Island    Very good  0.34019668100799
## 21             Maryland    Very good 0.342352759790318
## 2                Alaska    Very good 0.342907257180443
## 28             Nebraska    Very good 0.345609727580966
## 51              Wyoming    Very good 0.348360018653816
## 17               Kansas    Very good 0.348458756672981
## 16                 Iowa    Very good 0.350767341927563
## 48           Washington    Very good 0.353490879683709
## 45                 Utah    Very good   0.3566773965612
## 23             Michigan    Very good 0.358224174445054
## 50            Wisconsin    Very good 0.358249772105743
## 7           Connecticut    Very good 0.359511116889871
## 42         South Dakota    Very good 0.360824742268041
## 35         North Dakota    Very good 0.367331189710611
## 38               Oregon    Very good 0.367880013481631
## 30        New Hampshire    Very good 0.371042216358839
## 20                Maine    Very good 0.371146582185939
## 6              Colorado    Very good 0.374825585664978
## 24             Minnesota    Very good 0.376320022379187
## 46               Vermont    Very good 0.381482061726461
```

From the bar graph, one can see that for majority of the states the general health status is classified as 'Very good'. This indicates that the general health status of all the states observed is rated highly.

The table shows us that Vermont has the highest rating in terms of general health status, compared to all the other states.

Research quesion 2:

Method: A new, empty dataframe was created. The raw data was reduced to the necessary variables, and cleaned so as to avoid complications by removing rows with NA data.
The dataframe was further reduced by selecting only veterans, and then ones that were unable to work.

Interpretation of Results:

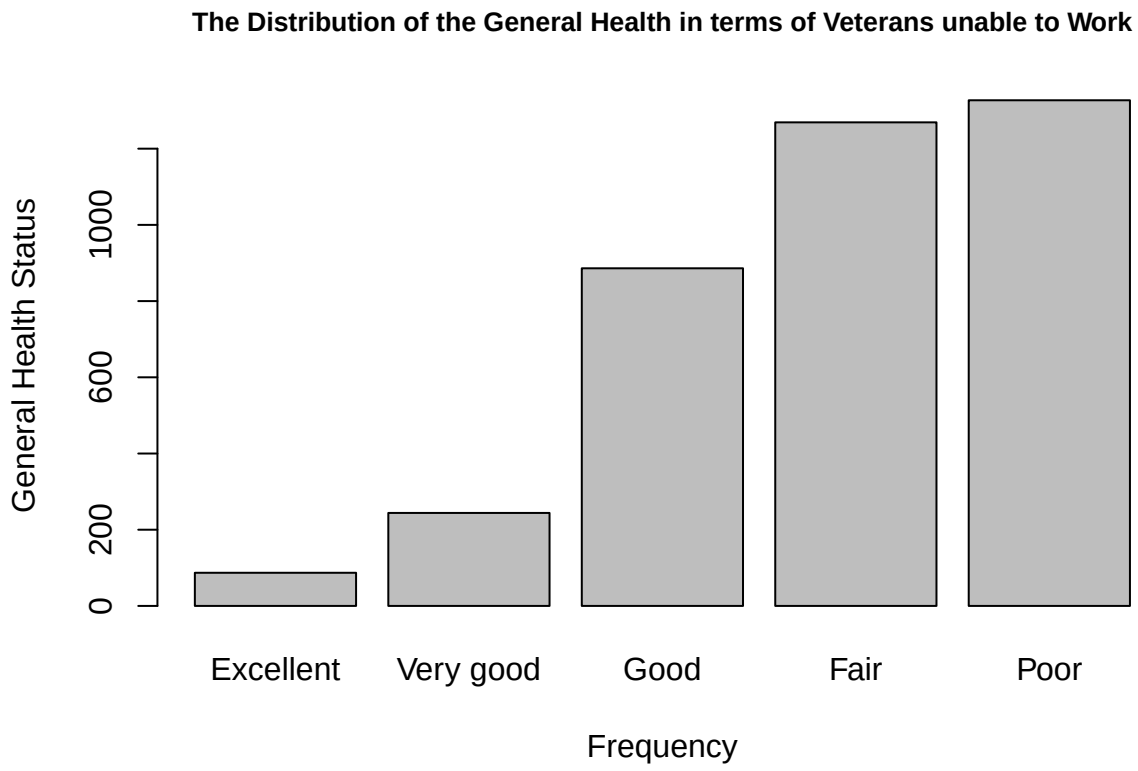
```
data <- brfss2013[,c("veteran3","genhlth","employ1")]
mydata <- na.omit(data)

veterans <- subset(mydata, veteran3 == "Yes")

unable <- subset(veterans, employ1 == "Unable to work")

freq<-unable$genhlth
```

A frequency table was then created, and then the data was plotted on a bar graph. This enabled us to view the relationship between general health status and employment in veterans.

```
summary(freq)
```

```
## Excellent Very good      Good      Fair      Poor
##        87       244       886      1269      1327
```

```
barplot(table(freq),main = "The Distribution of the General Health in terms of Veterans unable to Work"
```



The bar chart showing the frequency of the general health status for veterans that are unable to work indicates that majority of such veterans experience a poor general health status.

There is a strongly, positive relationship between the inability to work and a poor health status for veterans.

Research quesion 3:

Method: A new, empty dataframe was created. The raw data was reduced to the necessary variables, and cleaned so as to avoid complications by removing rows with NA data.
The dataframe was then separated into two different dataframes; one for males and one for females.
A subset of each dataframe was created, determing how many males and females were college graduates.
Then a frequency table was used to determine the relationship between having a college graduate education and income, for each gender.

Result:

```
x <- brfss2013[,c("sex","income2","educa")]
mydata <- na.omit(x)

newdata<- mydata[c('sex','income2','educa')]

male <- newdata[c(1:48),c("educa","income2")]
male_graduate <- subset(male, educa == "College 4 years or more (College graduate)")
male_freq <- count(male_graduate)
male_freq
```

```
##                                       educa           income2 freq
## 1 College 4 years or more (College graduate) Less than $10,000    1
## 2 College 4 years or more (College graduate) Less than $15,000    1
## 3 College 4 years or more (College graduate) Less than $35,000    1
## 4 College 4 years or more (College graduate) Less than $50,000    2
## 5 College 4 years or more (College graduate) Less than $75,000    6
## 6 College 4 years or more (College graduate)   $75,000 or more    8
```

```
female <- newdata[c(49:96),c("income2","educa")]
female_graduate <- subset(female, educa == "College 4 years or more (College graduate)")
female_freq <- count(female_graduate)
female_freq
```

```
##             income2                                       educa freq
## 1 Less than $20,000 College 4 years or more (College graduate)    1
## 2 Less than $50,000 College 4 years or more (College graduate)    2
## 3 Less than $75,000 College 4 years or more (College graduate)    4
## 4   $75,000 or more College 4 years or more (College graduate)    6
```

Interpretation:

As seen from the results above, there is a positive relationship between earning $75000 or more and being a college graduate i.e being a college graduate increases ones chance of earning a higher salary.

Also, within that scope, it is more likely for a college graduate male to earn $75000 or more than a college graduate female.