Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(reshape2)
library(ggcorrplot)
library(tidyverse)
library(caret)
library(corrplot)
```

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

Briefly check the dim and summary of the whole dataset.

```
dim(brfss2013)
```

```
## [1] 491775    330
```

```
# str(brfss2013)
```

```
# summary(brfss2013)
# brfss2013 <- brfss2013 %>%
#              drop_na()
# after removing NAs there would be no row left,
# so we could not remove all NAs here, instead, we have to remove NAs in each subsection.
#
# dim(brfss2013)
# dim(brfss.complete)
```

The dataset is relatively large, with 491775 rows and 330 cols. So, the summary information for the whole dataset is commented out.

There are main survey sections and optional sections in the BRFSS-2013 dataset, and we would like to focus on main survey sections only, so the next step is to filter out optional sections, only keep the columns we would like to investigate.

```
# pre-test with a subset
brfss.sub <- brfss2013 %>%
                      select("genhlth",
                             "X_bmi5", "X_bmi5cat",
                             "diabete3",
                              "income2",
                             "hlthpln1","exerany2")


dim(brfss.sub)
```

```
## [1] 491775      7
```

```
# Have to drop NAs

brfss.sub <- brfss.sub %>%
              drop_na()
dim(brfss.sub)
```

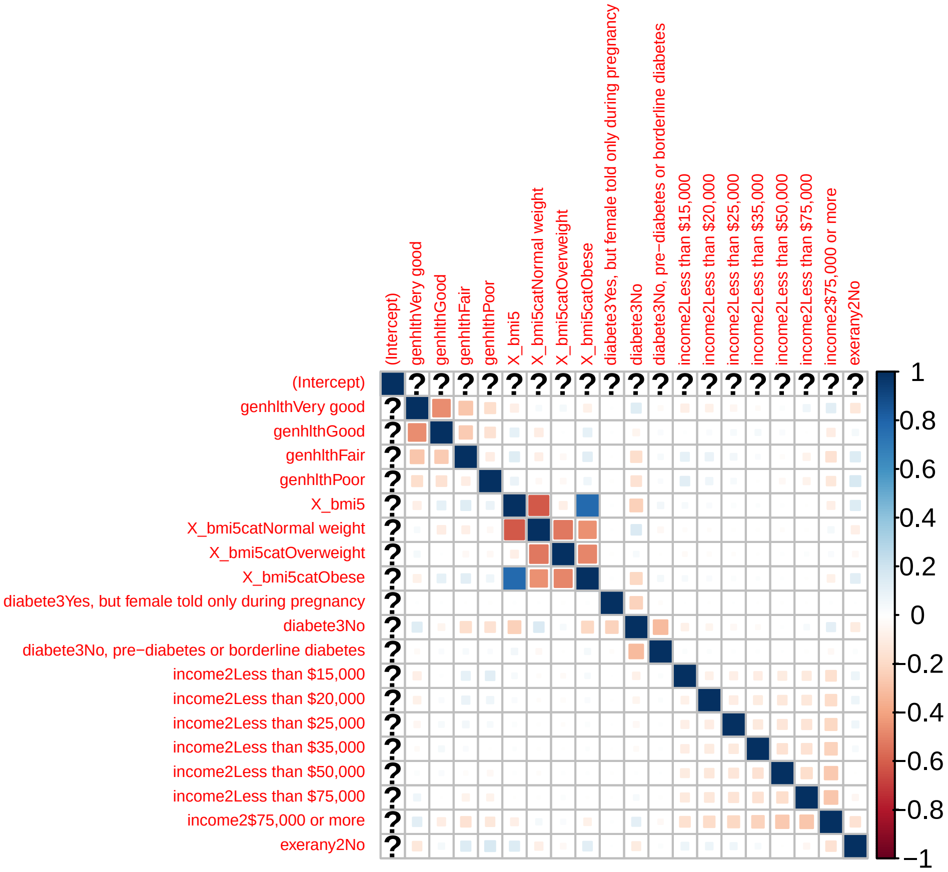
```
## [1] 378565      7
```

```
summary(brfss.sub)
```

```
##       genhlth          X_bmi5             X_bmi5cat
##  Excellent: 67390   Min.   :   1   Underweight   :  6133
##  Very good:126403   1st Qu.:2374   Normal weight:122888
##  Good     :114194   Median :2695   Overweight   :136914
##  Fair     : 49851   Mean   :2796   Obese        :112630
##  Poor     : 20727   3rd Qu.:3091
##                     Max.   :9769
##
##                                           diabete3                    income2
##  Yes                                          : 47773   $75,000 or more  :106568
##  Yes, but female told only during pregnancy:  3425   Less than $75,000: 59632
##  No                                           :320813   Less than $50,000: 55762
##  No, pre-diabetes or borderline diabetes   :  6554   Less than $35,000: 43712
##                                                         Less than $25,000: 36982
##                                                         Less than $20,000: 30516
##                                                         (Other)          : 45393
##  hlthpln1     exerany2
##  Yes:337427   Yes:278726
##  No : 41138   No : 99839
##
##
##
##
##
```

```
# brfss.data <- data.matrix( brfss.sub)
#
# check the numirical data matrix of brfss.sub
# summary(brfss.data)

# model.matrix( ~0+., data = brfss.sub) %>%
#   cor(use = "pairwise.complete.obs") %>%
#   ggcorrplot(show.diag = F, type = "lower", lab = TRUE, lab_size = 2)
#
#
# cor(brfss.sub, use = "pairwise.complete.obs")
#


brfss.m <- brfss.sub
brfss.m$X_state <- NULL
# summary( brfss.sub )

brfss.m <- model.matrix( hlthpln1 ~ ., data = brfss.sub)

# summary( brfss.m)
## brfss.dummy <- dummyVars( genhlth ~ ., data = brfss.sub)
# dim( brfss.m)
# brfss.cor <- cor( brfss.m, method = c("spearman"))
# dim(brfss.cor)

corrplot( cor( brfss.m), method = "square", tl.cex = 0.5)
```

```
## Warning in cor(brfss.m): the standard deviation is zero
```



Part 2: Research questions

Research quesion 1: The first research question we would like to focus on is to explorer whether there's a correlation between general health and education level. Our hypothesis is that there should be a direct correlation between education and income levels, which would contribute to the general health level of the individuals in this survery.

To address this research question, we need to subset the dataset with information about general health, education, and income.

```
# Subset Dataset for research question one
col.q1 <- c("genhlth", "educa", "income2", "hlthpln1", "exerany2")
brfss.q1 <- brfss2013[, col.q1]
brfss.q1 <- brfss.q1 %>%
              drop_na()
dim(brfss.q1)
```

```
## [1] 392966      5
```

Research quesion 2: The second research question we would like to focus on is to explorer whether there's a correlation between sleeping duration and chronic health conditions. Our hypothesis is that there should be a direct correlation between sleeping quality and chronic health states, which would contribute to the general health level of the individuals in this survery.

To address this research question, we need to subset the dataset with information about general health, ever diagnosed with heart attack, ever diagnosed with angina or coronary heart disease, ever diagnosed with a stroke, ever told had asthma, still have asthma.

```
# Subset Dataset for research question two
col.q2 <- c("genhlth", "cvdinfr4", "cvdcrhd4", "cvdstrk3", "asthma3", "asthnow"  )
brfss.q2 <- brfss2013[, col.q2]
brfss.q2 <- brfss.q2 %>%
              drop_na()
dim(brfss.q2)
```

```
## [1] 63698     6
```

Research quesion 3:

The third research question we would like to focus on is to explorer whether there's a correlation between Body Mass Index and diabetes. Our hypothesis is that there might be a correlation between obese and diabetes, which would contribute to the general health level of the individuals in this survery.

To address this research question, we need to subset the dataset with information about general health, overweight or obese calculated variable, computed body mass index, computed body mass index categories.

```
# Subset Dataset for research question three
col.q3 <- c("X_state","genhlth", "X_bmi5", "X_bmi5cat", "X_rfbmi5"  )
brfss.q3 <- brfss2013[, col.q3]

brfss.q3 <- brfss.q3 %>%
              drop_na()

dim(brfss.q3)
```

```
## [1] 463273      5
```

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 quesion 1: For research question one, we would like to explore the relationship between general health and education, then income level. We could first plot the mosaic correlation of income and general health.

```
# col.q1 <- c("genhlth", "educa", "income2", "hlthpln1", "exerany2")
# brfss.q1 <- brfss2013[, col.q1]
# plot the mosaic showing general health and education levels
summary( brfss.q1$educa)
```

```
##                    Never attended school or only kindergarten
##                                                           396
##                               Grades 1 through 8 (Elementary) 
##                                                          9323 
##                         Grades 9 though 11 (Some high school) 
##                                                         20430 
##                        Grade 12 or GED (High school graduate) 
##                                                        109272 
## College 1 year to 3 years (Some college or technical school) 
##                                                        109061 
##                    College 4 years or more (College graduate) 
##                                                        144484
```

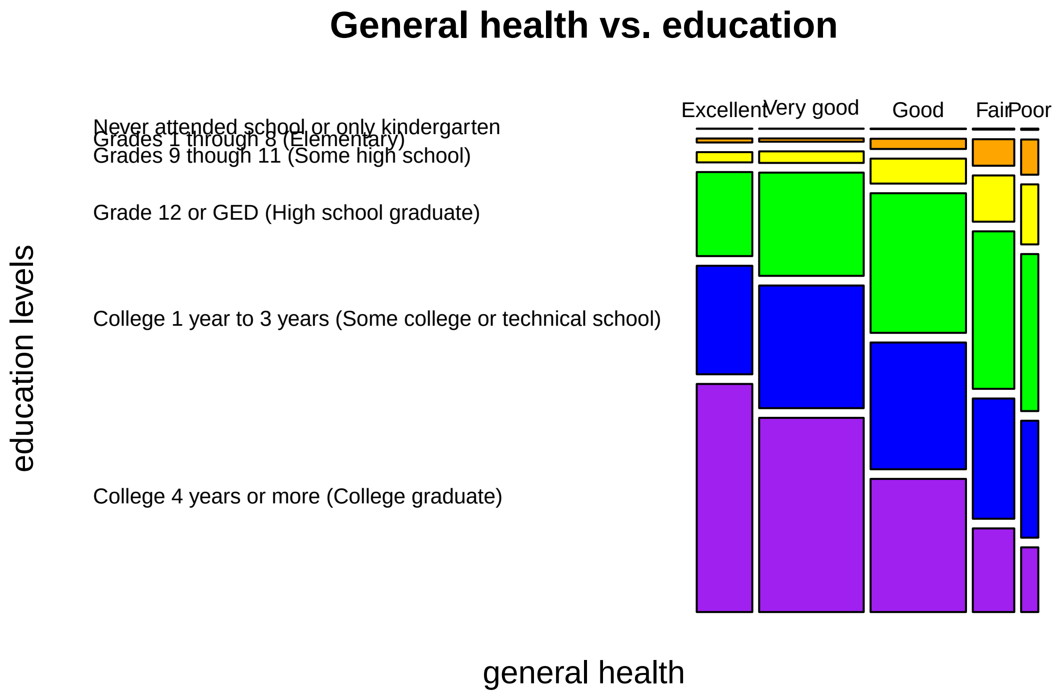
```r
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.0.5
```

```
## Loading required package: grid
```

```r
#
# par(mar=c(5, 4, 4, 2) + 0.1)
mosaicplot( ~ genhlth + educa,
            data = brfss.q1,
            xlab = "general health",
            ylab = "education levels",
            direction = "v",
            color = c("red", "orange", "yellow", "green", "blue", "purple"),
            main = "General health vs. education",
            las = 1)
```

```
## Warning: In mosaicplot.default(table(mf), main = main, ...) :
##  extra argument 'direction' will be disregarded
```

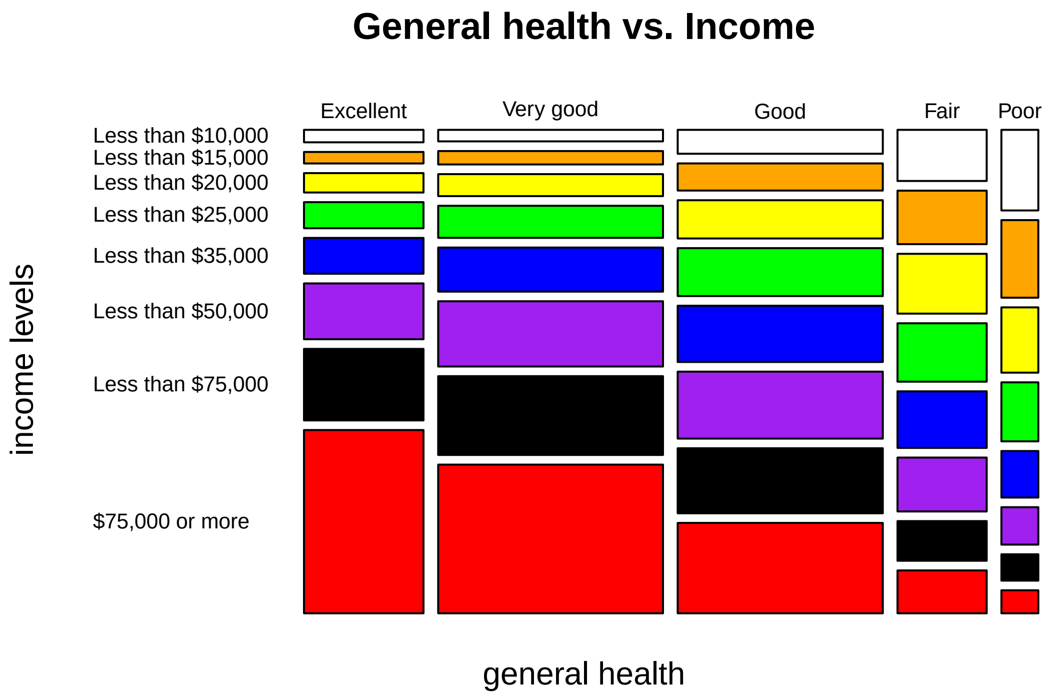


```
# mosaic( ~ genhlth + educa,
#         data = brfss.q1,
#         shade=TRUE,
#         legend = TRUE,
#         direction = "v",
#         rot_labels=c(0, 90, 0, 0),
#         #color = c("white", "orange", "yellow", "green", "blue", "purple", "black", "red"),
#         #xlab = "general health",
#         #ylab = "income levels",
#         main = "General health vs. education" )

# from the mosaic plot, we could see there's a trand that in general health "Excellent" and "Very good"
# there are more individuals with college 4 years or more education.

summary(brfss.q1$income2)
```

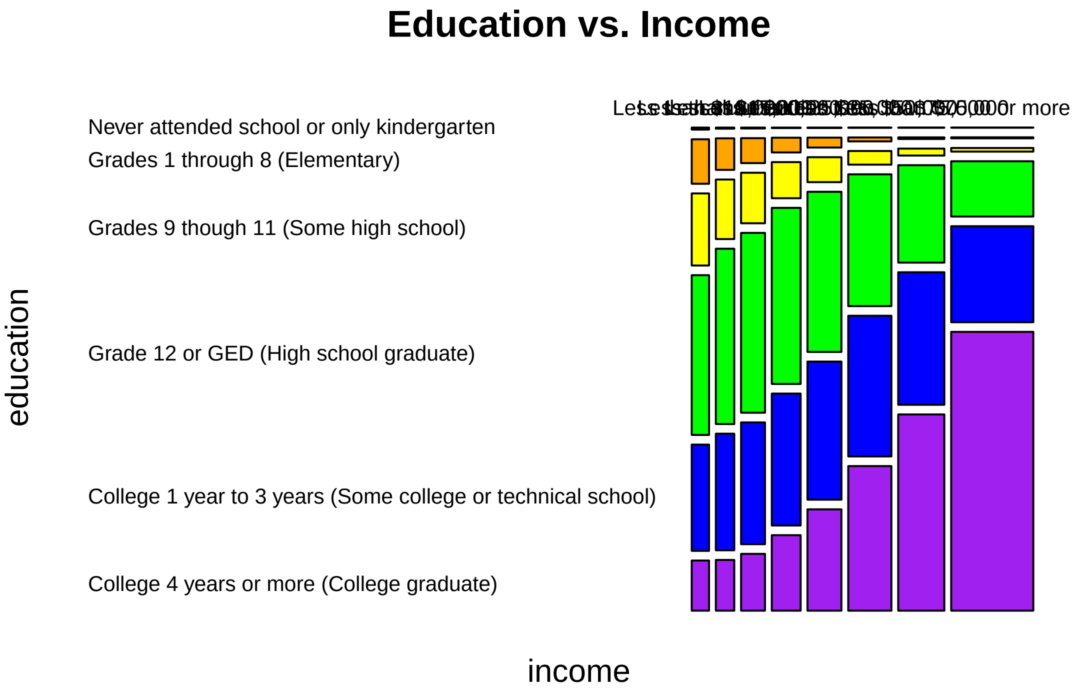
```
## Less than $10,000 Less than $15,000 Less than $20,000 Less than $25,000
##             23132             24613             31966             38540
## Less than $35,000 Less than $50,000 Less than $75,000   $75,000 or more
##             45449             57719             61679            109868
```

```
mosaicplot( ~ genhlth + income2,
            data = brfss.q1,
            xlab = "general health",
            ylab = "income levels",
            direction = "v",
            color = c("white", "orange", "yellow", "green", "blue", "purple", "black", "red"),
            main = "General health vs. Income",
            las = 1)
```

```
## Warning: In mosaicplot.default(table(mf), main = main, ...) :
##  extra argument 'direction' will be disregarded
```



```
# from the mosaic plot, we could see there's a trand that in general health "Excellent/Very good" categ
# there are more individuals with $75,000 or more income.

# mosaic( ~ genhlth + income2,
#         data = brfss.q1,
#         shade=TRUE,
#         #legend = TRUE,
#         direction = "v",
#         rot_labels=c(0, 90, 0, 0),
#         #margins = c(10, 10, 10, 10),
#         main = "General Health vs. Income",
#         )


# plot income vs. education
mosaicplot( ~ income2 + educa,
            data = brfss.q1,
            xlab = "income",
            ylab = "education",
            color = c("white", "orange", "yellow", "green", "blue", "purple"),
            main = "Education vs. Income",
            las = 1)
```



Those mosaic plots suggest that there are correlations between income and general health, as well as between education and income. So, the next step is to further investigate the correlations.

```
# sub group the individuals by their income levels
summary(brfss.q1$genhlth)
```

```
## Excellent Very good      Good      Fair      Poor
##     69507    130767    119324     51889     21479
```

```
brfss.q1.subgroup <- brfss.q1 %>%
                      group_by(income2) %>%
                      summarize( Excellent = sum( genhlth == "Excellent")/n(),
                                 VeryGood = sum(genhlth == "Very good")/n(),
                                 Good = sum(genhlth == "Good")/n(),
                                 Fair = sum(genhlth == "Fair")/n(),
                                 Poor = sum(genhlth == "Poor")/n())
```

Now we could plot the income vs. health status

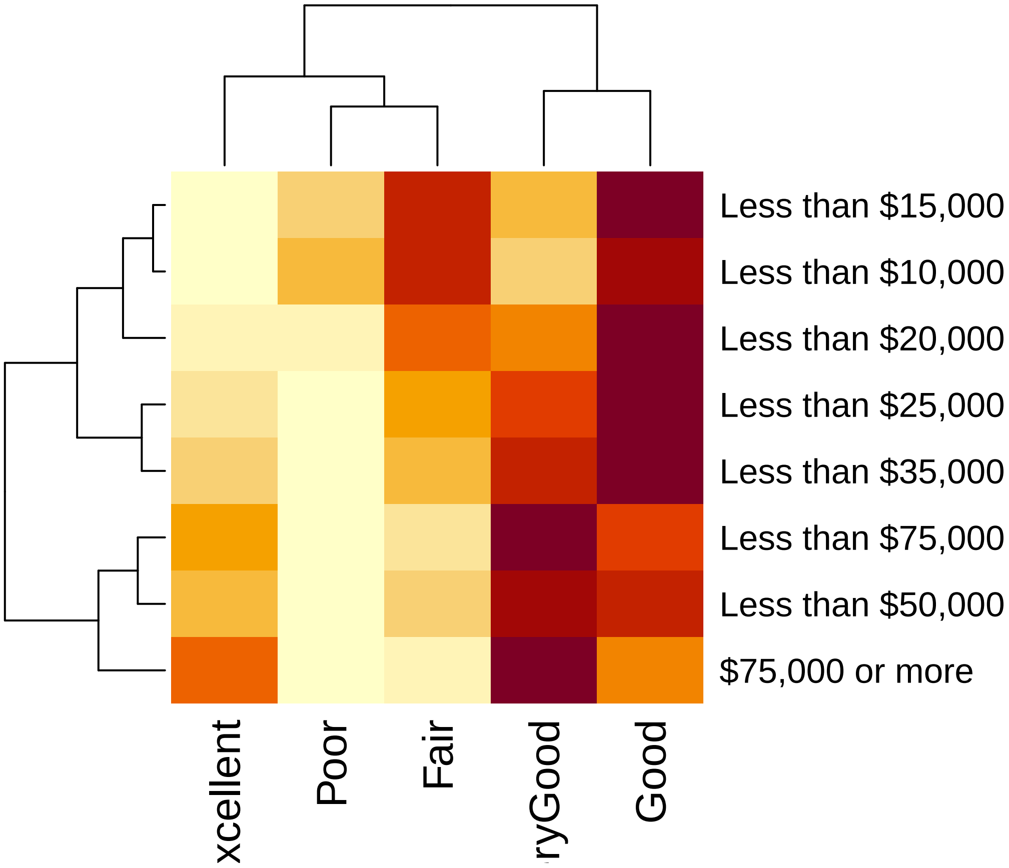
```
dim(brfss.q1.subgroup)
```

```
## [1] 8 6
```

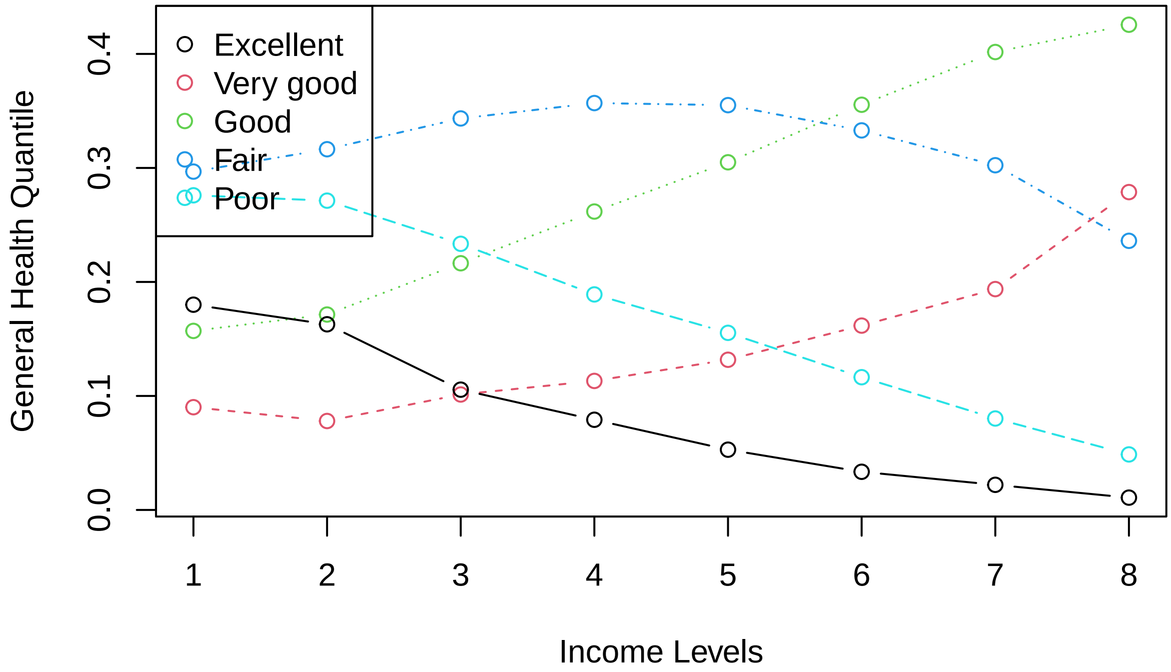
```
brfss.q1.subgroup
```

```
## # A tibble: 8 x 6
##   income2            Excellent VeryGood  Good   Fair   Poor
##   <fct>                  <dbl>    <dbl> <dbl>  <dbl>  <dbl>
## 1 Less than $10,000     0.0901    0.157 0.297 0.276  0.180
## 2 Less than $15,000     0.0780    0.171 0.316 0.271  0.163
## 3 Less than $20,000     0.101     0.216 0.343 0.233  0.106
## 4 Less than $25,000     0.113     0.262 0.357 0.189  0.0791
## 5 Less than $35,000     0.132     0.305 0.355 0.155  0.0529
## 6 Less than $50,000     0.162     0.355 0.333 0.116  0.0335
## 7 Less than $75,000     0.194     0.402 0.302 0.0802 0.0220
## 8 $75,000 or more       0.279     0.426 0.236 0.0487 0.0108
```

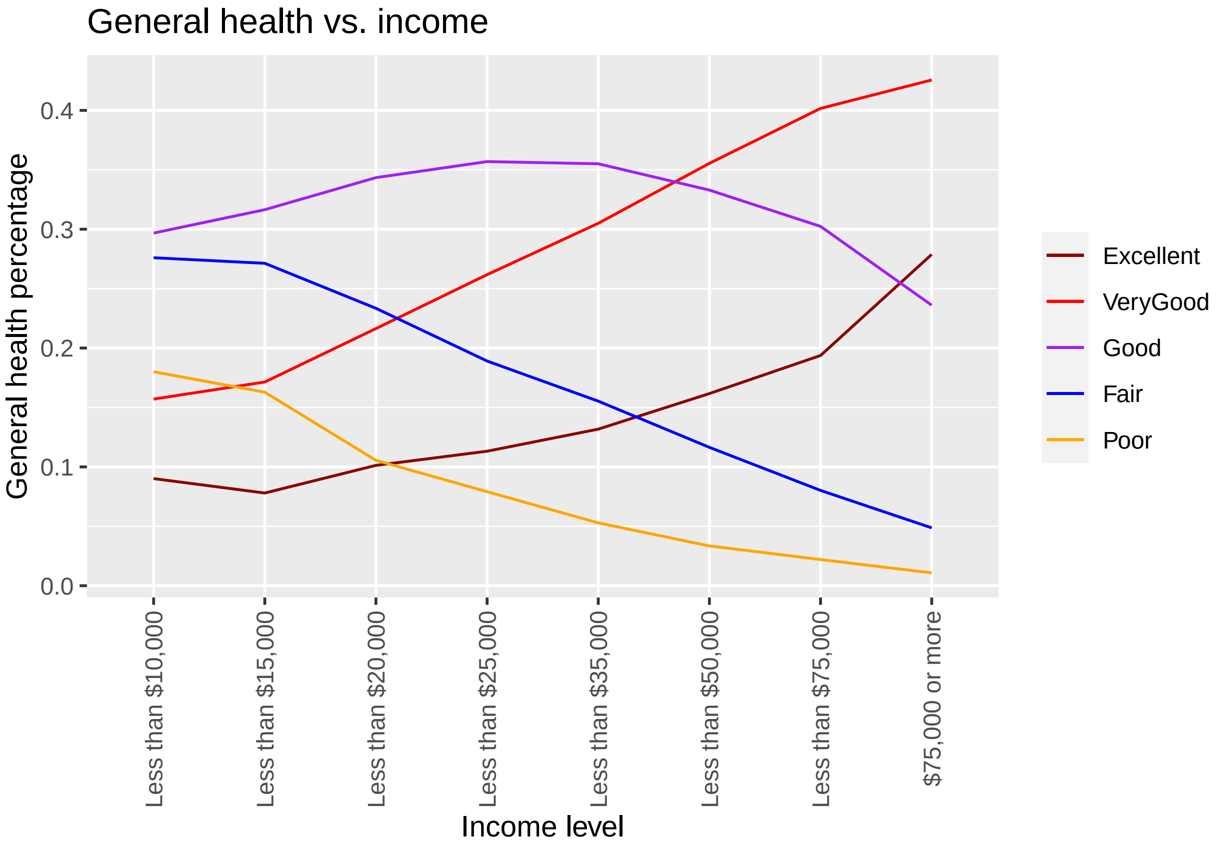
```
q1.df <- as.data.frame( brfss.q1.subgroup)
row.names(q1.df) <- q1.df$income2
q1.df$income2 <- NULL
q1.m <- as.matrix(q1.df)
heatmap(q1.m)
```



```
# plot multiple lines of health vs income
q1.df <- as.data.frame( brfss.q1.subgroup)


matplot( q1.df, type = c("b"),
         xlab = "Income Levels",
         ylab = "General Health Quantile",
         pch =1, col = 1:5)

legend("topleft",
       legend = c("Excellent", "Very good", "Good", "Fair", "Poor" ),
       col = 1:5,
       #xlab = "income2",
       pch = 1)
```



```
## try ggplot with multiple lines

ggplot( q1.df, aes(x = income2)) +
  geom_line( aes( y = Excellent, group = 1, color = "Excellent")) +
  geom_line( aes( y = VeryGood, group = 1, color = "VeryGood")) +
  geom_line( aes( y = Good, group = 1, color = "Good")) +
  geom_line( aes( y = Fair, group = 1, color = "Fair")) +
  geom_line( aes( y = Poor, group = 1, color = "Poor")) +
  scale_colour_manual("",
                      breaks = c("Excellent", "VeryGood", "Good", "Fair", "Poor"),
                     values = c("darkred", "red", "purple", "blue", "orange")
                     ) +
  labs( x = "Income level",
        y = "General health percentage",
        title = "General health vs. income") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



From the plots above, we could clearly see as the income level increase the proportions of individuals reporting “Excellent” and “Very Good” in general health category increased significantly, this positive correlation also confirms with our initial hypothesis.

For individuals reporting “Good” in general health category, there’s an increase when income increase from less than $10,000 to less than $35,000, then there’s a slighly drop when income levels are higher than $35,000 per year.

For individuals reporting “Fair” and “Poor” in general health category, we could see a clear negative correlation beteen the income level and the portion of general health.

Similarly, we shall plot the relationship between income level and education levels.

```
# sub group the individuals by their income levels
summary(brfss.q1$educa)
```

```
##                       Never attended school or only kindergarten
##                                                              396
##                                  Grades 1 through 8 (Elementary)
##                                                             9323
##                         Grades 9 though 11 (Some high school)
##                                                         20430
##                        Grade 12 or GED (High school graduate)
##                                                        109272
## College 1 year to 3 years (Some college or technical school)
##                                                        109061
##                    College 4 years or more (College graduate)
##                                                        144484
```

```
brfss.q1.subgroup2 <- brfss.q1 %>%
                      group_by(income2) %>%
                      summarize( noSchool = sum( educa == "Never attended school or only kindergarten")
                                 Elementary = sum(educa == "Grades 1 through 8 (Elementary)")/n(),
                                 MiddleSchool = sum(educa == "Grades 9 though 11 (Some high school)")/n
                                 HighSchool = sum(educa == "Grade 12 or GED (High school graduate)")/n(
                                 College = sum(educa == "College 1 year to 3 years (Some college or tec
                                 Graduate = sum(educa == "College 4 years or more (College graduate)")/
                                 )
```

Now we could plot the income vs. education status

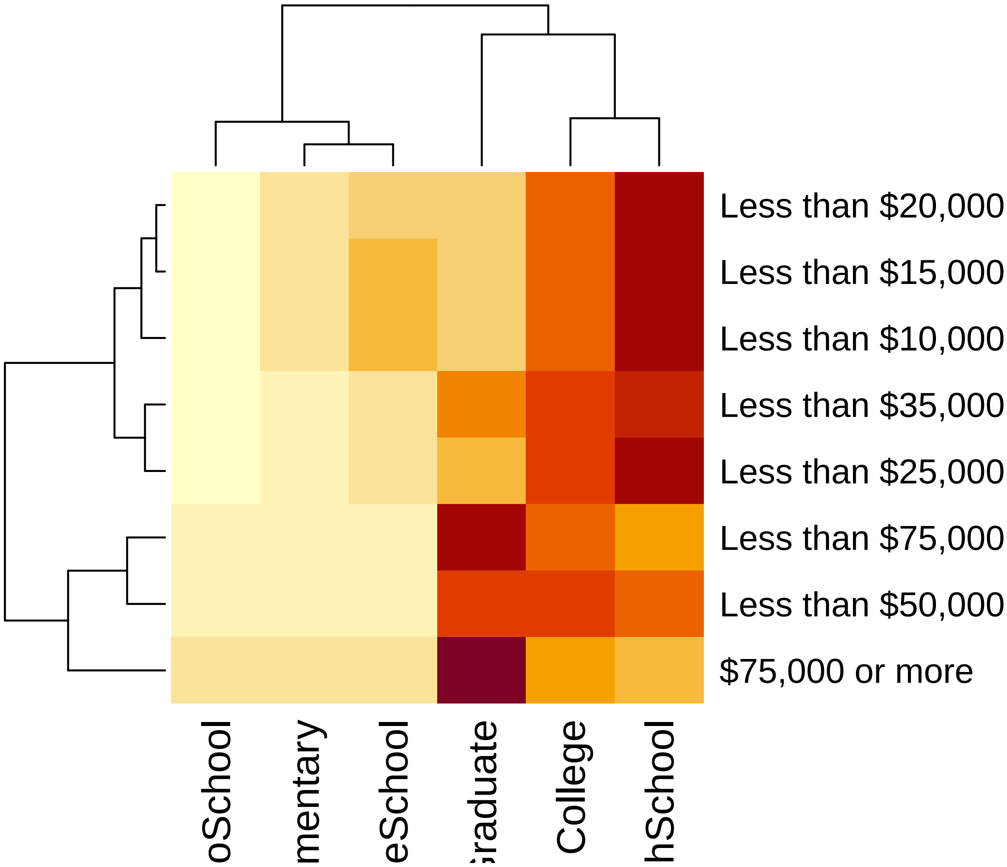
```
dim(brfss.q1.subgroup2)
```

```
## [1] 8 7
```

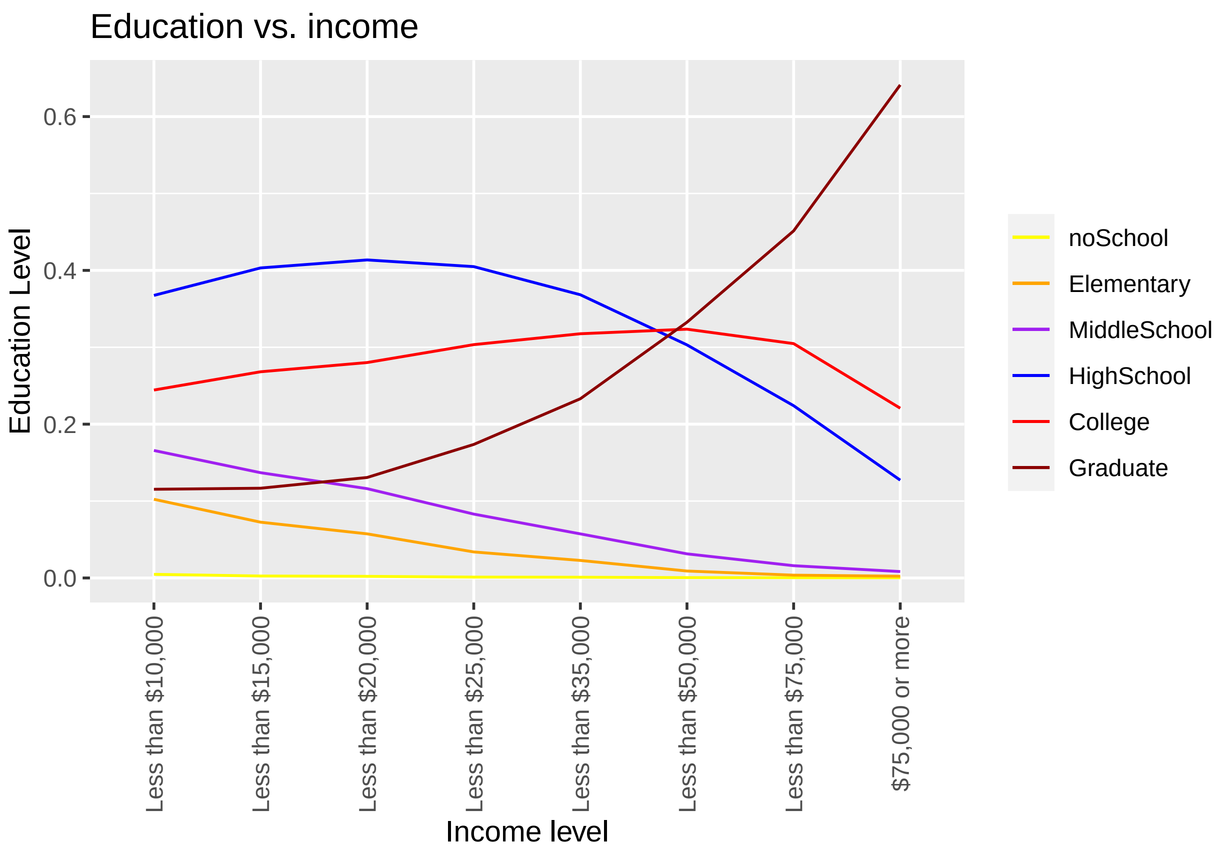
```
brfss.q1.subgroup2
```

```
## # A tibble: 8 x 7
##   income2           noSchool Elementary MiddleSchool HighSchool College Graduate
##   <fct>                <dbl>      <dbl>        <dbl>      <dbl>   <dbl>    <dbl>
## 1 Less than $10,000 0.00471     0.102        0.166        0.367   0.244    0.115
## 2 Less than $15,000 0.00252     0.0726       0.137        0.403   0.268    0.117
## 3 Less than $20,000 0.00213     0.0574       0.116        0.414   0.280    0.131
## 4 Less than $25,000 0.00112     0.0339       0.0831       0.405   0.303    0.174
## 5 Less than $35,000 0.000924    0.0228       0.0573       0.368   0.318    0.233
## 6 Less than $50,000 0.000468    0.00903      0.0314       0.303   0.324    0.333
## 7 Less than $75,000 0.000340    0.00363      0.0159       0.224   0.305    0.451
## 8 $75,000 or more   0.000218    0.00228      0.00832      0.127   0.221    0.641
```

```
q1.df2 <- as.data.frame( brfss.q1.subgroup2)

row.names(q1.df2) <- q1.df$income2

q1.m2 <- q1.df2
q1.m2$income2 <- NULL
q1.m2 <- as.matrix(q1.m2)
heatmap(q1.m2)
```



```
## try ggplot with multiple lines

ggplot( q1.df2, aes(x = income2)) +
  geom_line( aes( y = noSchool, group = 1, color = "noSchool")) +
  geom_line( aes( y = Elementary, group = 1, color = "Elementary")) +
  geom_line( aes( y = MiddleSchool, group = 1, color = "MiddleSchool")) +
  geom_line( aes( y = HighSchool, group = 1, color = "HighSchool")) +
  geom_line( aes( y = College, group = 1, color = "College")) +
  geom_line( aes( y = Graduate, group = 1, color = "Graduate")) +
  scale_colour_manual("",
                     breaks = c("noSchool", "Elementary", "MiddleSchool", "HighSchool", "College", "Gra
                     values = c("yellow", "orange", "purple", "blue", "red", "darkred")
                     ) +
  labs( x = "Income level",
        y = "Education Level",
        title = "Education vs. income") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



From the education vs. income plot above, we could clearly see the proportion of individuals who finished college increase as the income level increase. This trand is dropping in individuals without a college degree, and it is less significant.

Henceforce, we have confidence that from the dataset we are working on, individuals who have received better education are more likely to have better income, thus better income would contribute to better general health categories.

Research quesion 2:

Research quesion 3: