Exploring the BRFSS data

Setup

Load packages

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

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)
```

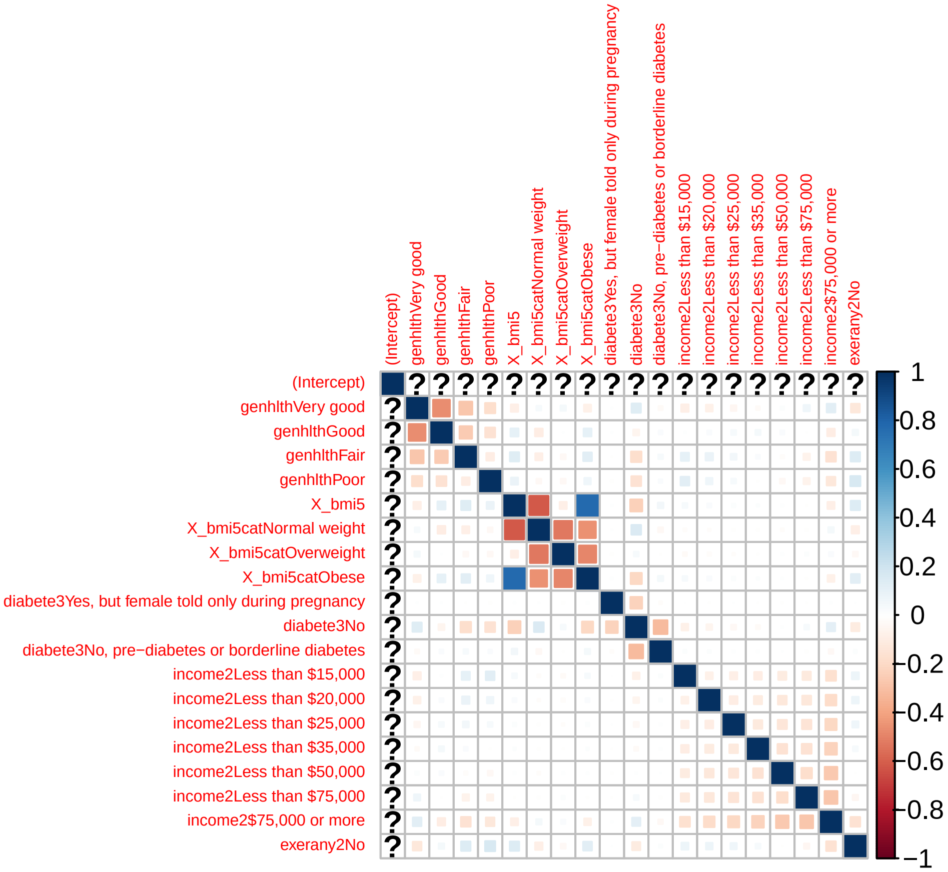
```
## [1] 378565      7
```

```
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
##
##
##
##
##
```

```r
# 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", "sleptim1", "cvdinfr4", "cvdcrhd4", "cvdstrk3", "asthma3", "asthnow"  )
brfss.q2 <- brfss2013[, col.q2]
brfss.q2 <- brfss.q2 %>%
              drop_na()
dim(brfss.q2)
```

```
## [1] 62645     7
```

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.

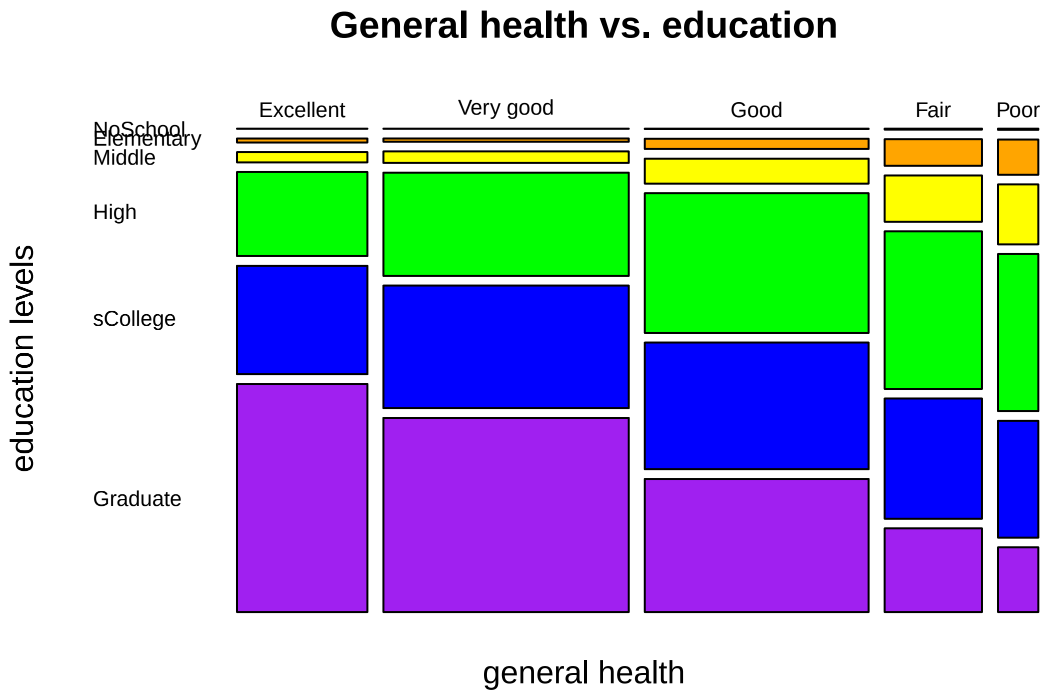
```r
# col.q1 <- c("genhlth", "educa", "income2", "hlthpln1", "exerany2")
# brfss.q1 <- brfss2013[, col.q1]
# plot the mosaic showing general health and education levels
summary( brfss.q1$educa)
```

```
##                   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
# the factor strings in educa column are too long, replace with short strings
# brfss.q1$education <- brfss.q1$educa
brfss.q1 <- brfss.q1 %>%
              mutate(education = case_when(
                educa == "Never attended school or only kindergarten" ~ "NoSchool",
                educa == "Grades 1 through 8 (Elementary)" ~ "Elementary",
                educa == "Grades 9 though 11 (Some high school)" ~ "Middle",
                educa == "Grade 12 or GED (High school graduate)" ~ "High",
                educa == "College 1 year to 3 years (Some college or technical school)" ~ "sCollege",
                educa == "College 4 years or more (College graduate)" ~ "Graduate"

              ))

# re-order the factor levels in $graduation column

brfss.q1$education = factor(brfss.q1$education, c("NoSchool", "Elementary", "Middle", "High", "sCollege

summary( brfss.q1$education)
```

```
##   NoSchool Elementary     Middle       High   sCollege   Graduate
##        396       9323      20430     109272     109061     144484
```

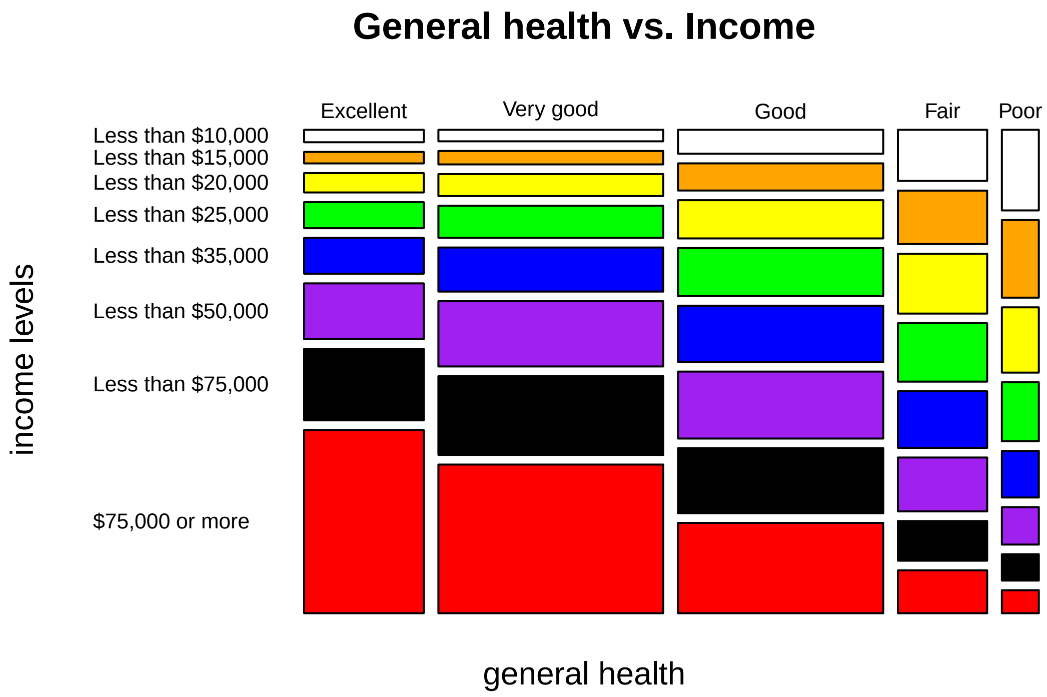
```r
#
# par(mar=c(5, 4, 4, 2) + 0.1)
mosaicplot( ~ genhlth + education,
            data = brfss.q1,
            xlab = "general health",
            ylab = "education levels",
            # direction = "v",
            color = c("red", "orange", "yellow", "green", "blue", "purple"),
            main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



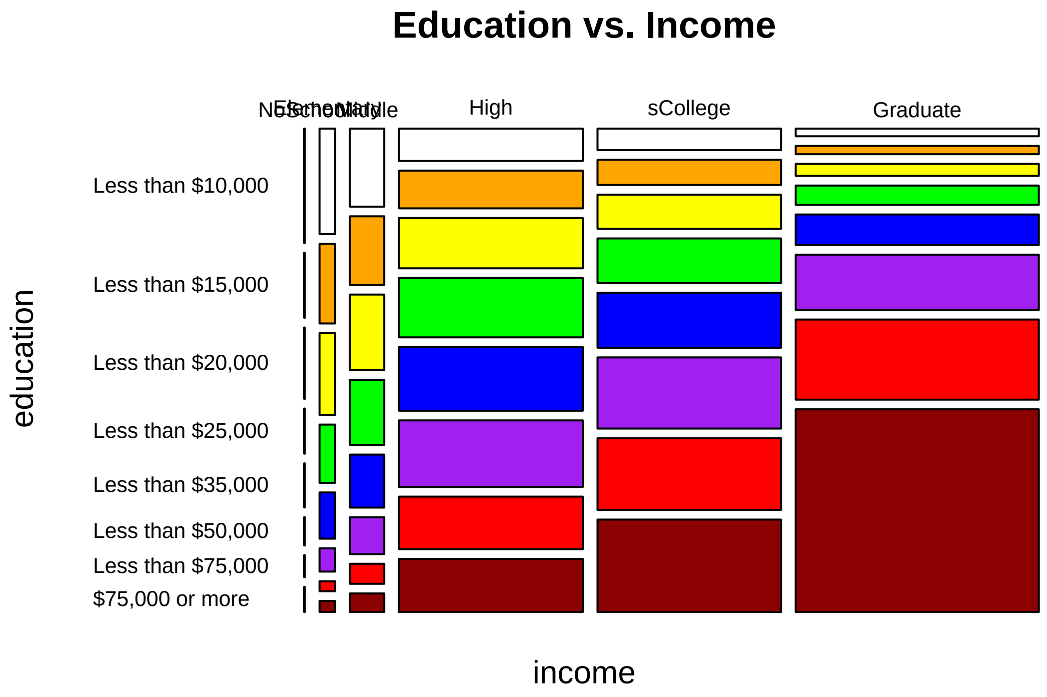
```
# mosaic( ~ genhlth + educa,
#          data = brfss.q1,
#          shade=TRUE,
#          legend = TRUE,
#          direction = "v",
#          rot_labels=c(0, 90, 0, 0),
#          #color = c("white", "orange", "yellow", "green", "blue", "purple", "black", "red"),
#          #xlab = "general health",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



```
# 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.

# plot income vs. education



mosaicplot( ~ education + income2,
            data = brfss.q1,
            xlab = "income",
            #xlab = list( "L15k", "L20k", "L25k", "L30k", "L35k", "L50k", "L75k", "More75k"),
            ylab = "education",
            color = c("white", "orange", "yellow", "green", "blue", "purple", "red", "darkred"),
            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

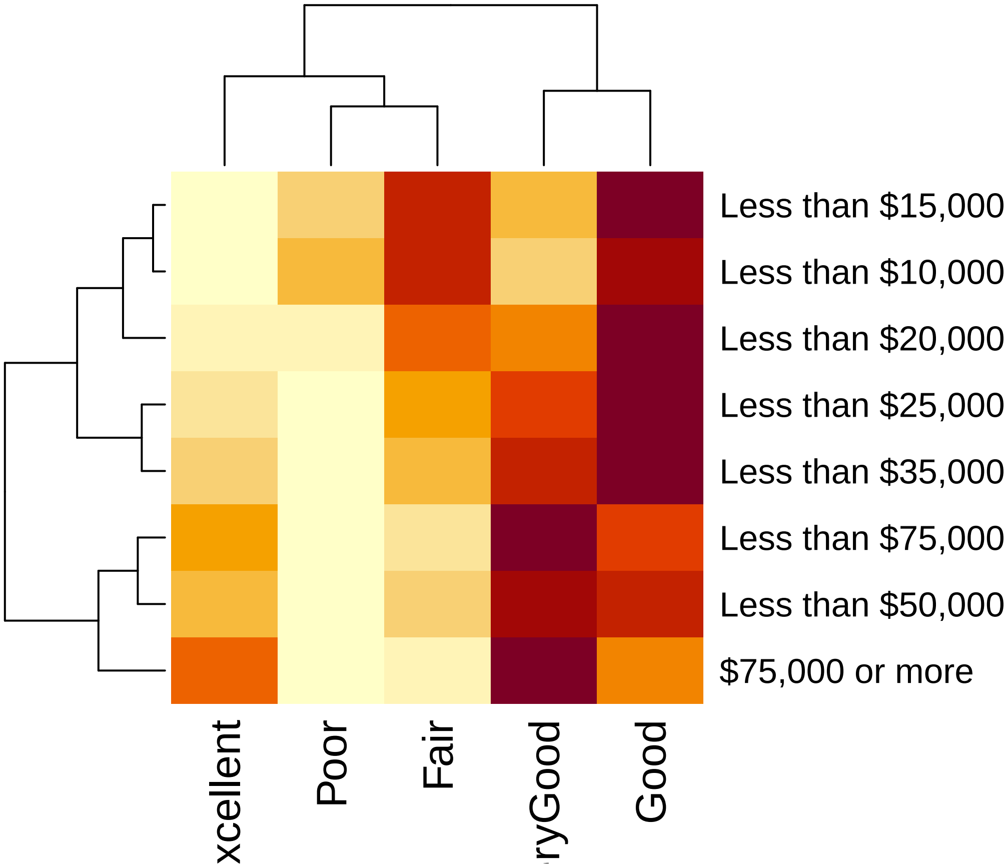
```
dim(brfss.q1.subgroup)
```

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

```
brfss.q1.subgroup
```

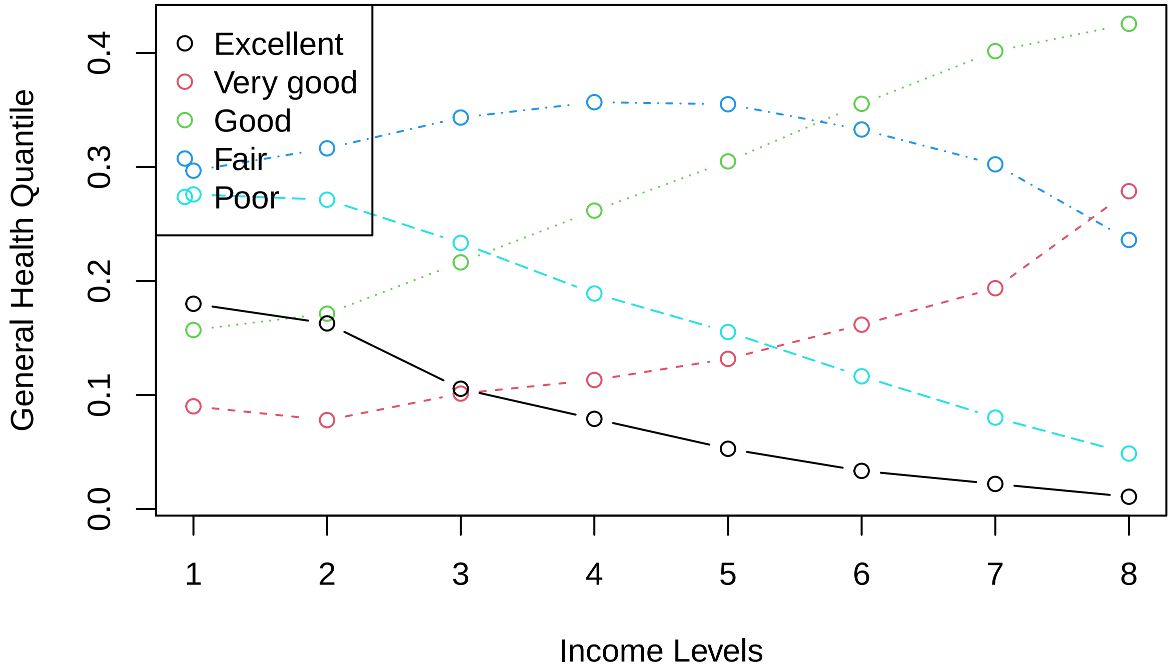
```
## # A tibble: 8 x 6
##   income2            Excellent VeryGood  Good   Fair   Poor
##   <fct>                  <dbl>    <dbl> <dbl>  <dbl>  <dbl>
## 1 Less than $10,000     0.0901    0.157 0.297 0.276  0.180
## 2 Less than $15,000     0.0780    0.171 0.316 0.271  0.163
## 3 Less than $20,000     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)
```

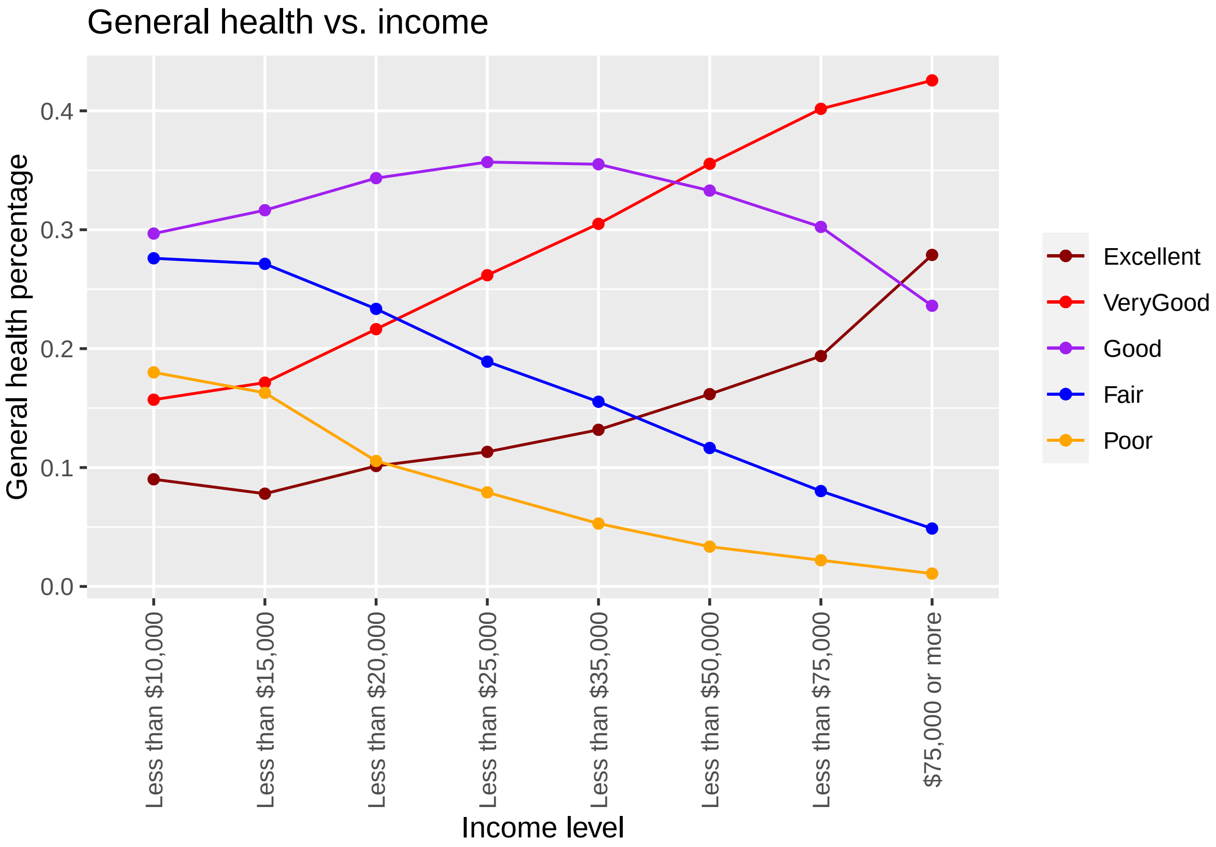


```
# plot multiple lines of health vs income
q1.df <- as.data.frame( brfss.q1.subgroup)
```

```
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")
                    ) +
  geom_point( aes( y = Excellent, group = 1, color = "Excellent")) +
  geom_point( aes( y = VeryGood, group = 1, color = "VeryGood")) +
  geom_point( aes( y = Good, group = 1, color = "Good")) +
  geom_point( aes( y = Fair, group = 1, color = "Fair")) +
  geom_point( aes( y = Poor, group = 1, color = "Poor")) +
  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

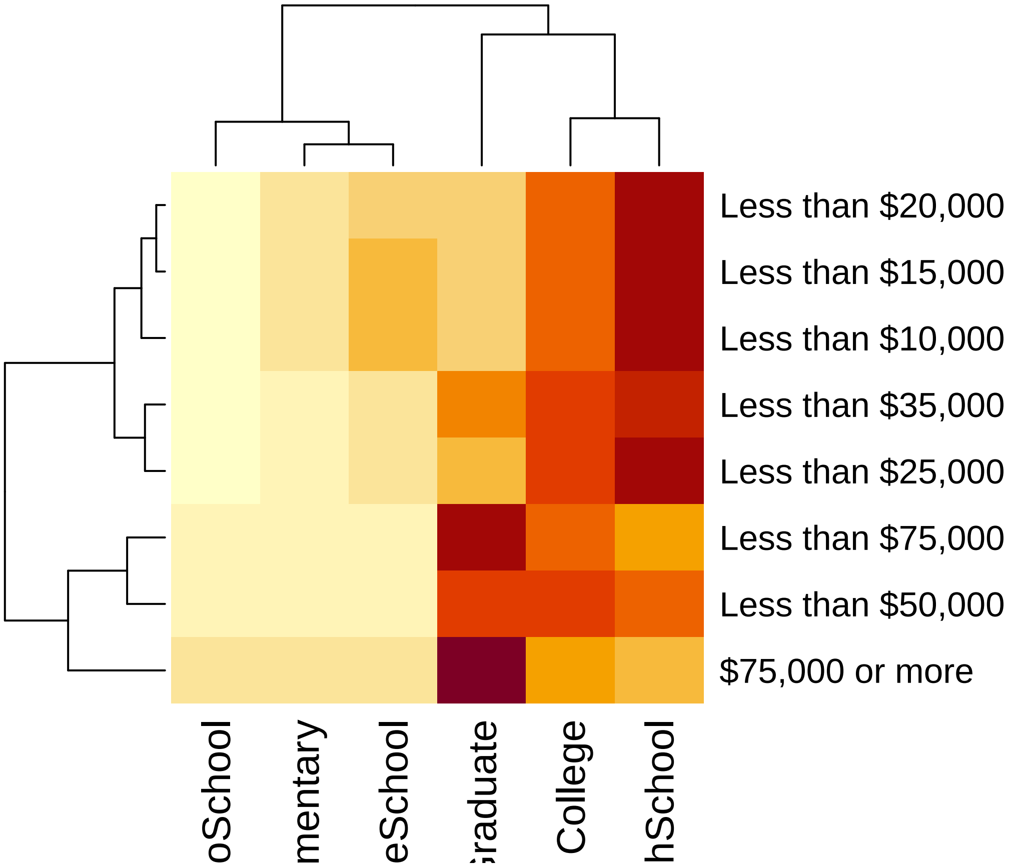
```
dim(brfss.q1.subgroup2)
```

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

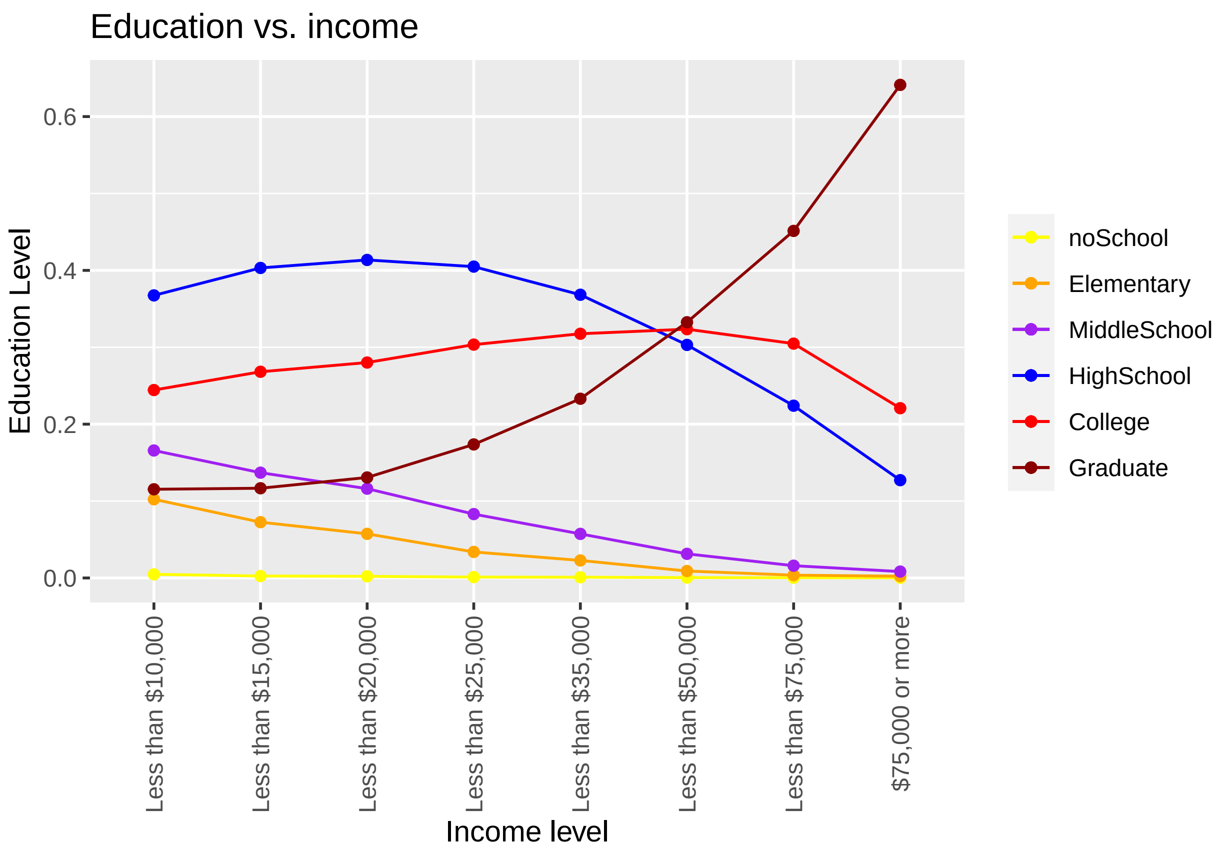
```
brfss.q1.subgroup2
```

```
## # 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")
                      ) +
  geom_point( aes( y = noSchool, group = 1, color = "noSchool")) +
  geom_point( aes( y = Elementary, group = 1, color = "Elementary")) +
  geom_point( aes( y = MiddleSchool, group = 1, color = "MiddleSchool")) +
  geom_point( aes( y = HighSchool, group = 1, color = "HighSchool")) +
  geom_point( aes( y = College, group = 1, color = "College")) +
  geom_point( aes( y = Graduate, group = 1, color = "Graduate")) +
  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: The second research question will focus on correlation between health stages and chronic health conditions. Note, because over 90% individuals checked NA for 'Still has asthma now', we could not drop all NAs in this dataframe.

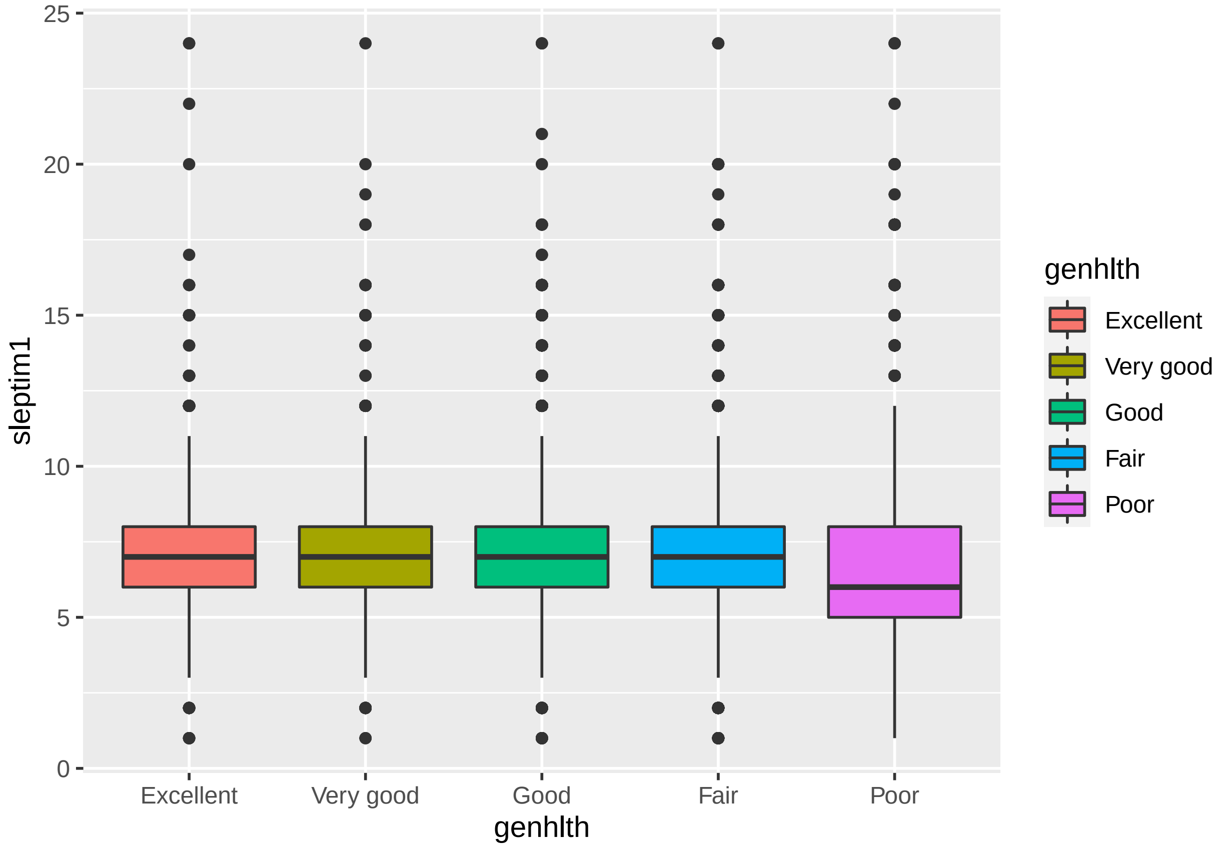
```
dim(brfss.q2)
```

```
## [1] 62645     7
```

```r
# s;eptim1: time sleep
# cvdinfr4: ever with heart attack
# cvdcrhd4: ever with angina or coronary heart disease
# cvdstrk3: ever with with a stroke
# asthma3: ever had asthma
# asthnow: still have asthma

summary( brfss.q2)
```

```
##       genhlth         sleptim1      cvdinfr4    cvdcrhd4    cvdstrk3
##  Excellent: 6618   Min.   : 1.000   Yes: 4853   Yes: 5437   Yes: 3839
##  Very good:16848   1st Qu.: 6.000   No :57792   No :57208   No :58806
##  Good     :19668   Median : 7.000
##  Fair     :12621   Mean   : 6.859
##  Poor     : 6890   3rd Qu.: 8.000
##                    Max.   :24.000
##  asthma3     asthnow
##  Yes:62645   Yes:43557
##  No :    0   No :19088
##
##
##
##
```

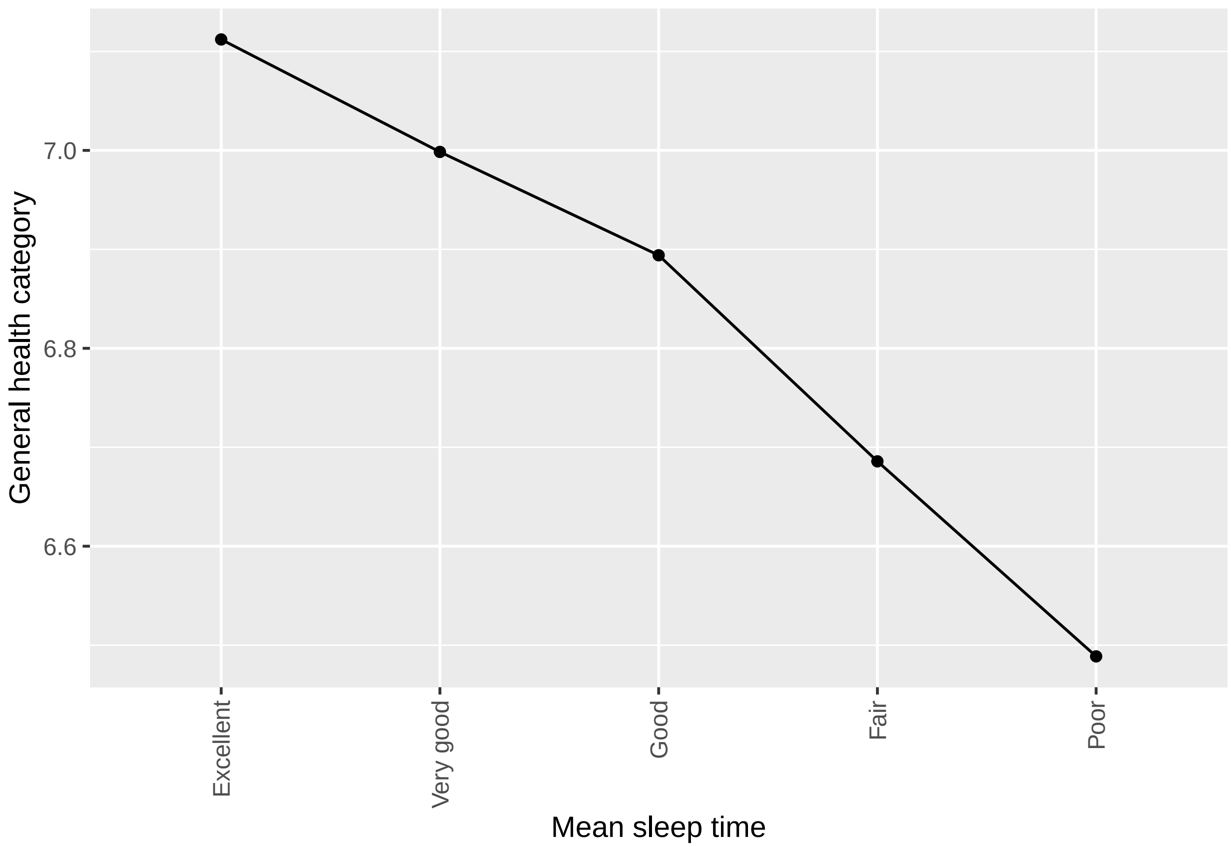
```r
# boxplot sleep time vs general health
ggplot( brfss.q2, aes( x = genhlth, y = sleptim1, fill=genhlth)) +
  geom_boxplot()
```



```
# line plot of each category of general health vs. mean sleep time.

brfss.q2.sub <- brfss.q2 %>%
                  group_by(genhlth) %>%
                    summarise(mSleep = mean( sleptim1))

summary(brfss.q2.sub)
```

```
##       genhlth      mSleep
##  Excellent:1   Min.   :6.489
##  Very good:1   1st Qu.:6.686
##  Good     :1   Median :6.894
##  Fair     :1   Mean   :6.836
##  Poor     :1   3rd Qu.:6.998
##                Max.   :7.112
```

```
ggplot( brfss.q2.sub, aes( x=mSleep, y = genhlth, group = 1))+
  geom_line( aes(genhlth, mSleep) ) +
  geom_point( aes(genhlth, mSleep)) +
  labs( x = "Mean sleep time",
        y = "General health category",
        main = "General health vs. sleep time") +
  theme( axis.text.x= element_text( angle = 90, vjust = 0.5, hjust = 1))
```



From the boxplot, we could see there's a significant drop of Q1 to Q3 sleeping range for individuals with poor health condition. From the sleep vs. general health category plot, we could see the mean sleeping time dropped as the health conditions went from excellent to poor. Those two plots show a positive correlation between sleeping time and general health.

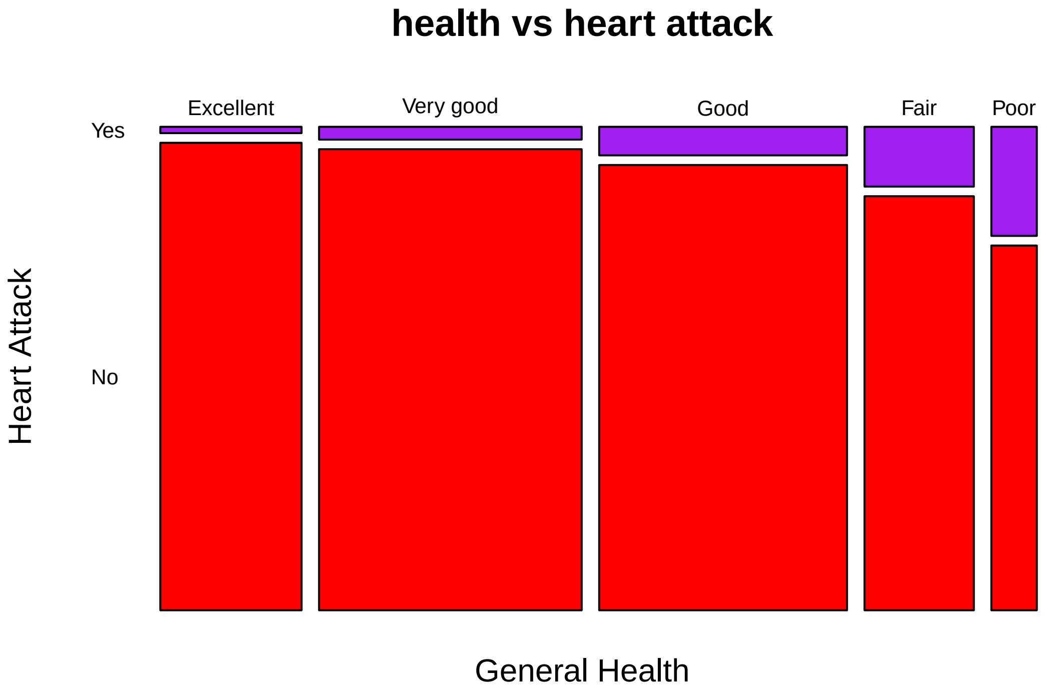
To further investigate, we would like to see the pattern between chronic conditions and sleeping conditions.

```
col.q2 <- c("genhlth", "sleptim1", "cvdinfr4", "cvdcrhd4", "cvdstrk3", "asthma3", "asthnow"  )
brfss.q2 <- brfss2013[, col.q2]

summary( brfss.q2)
```

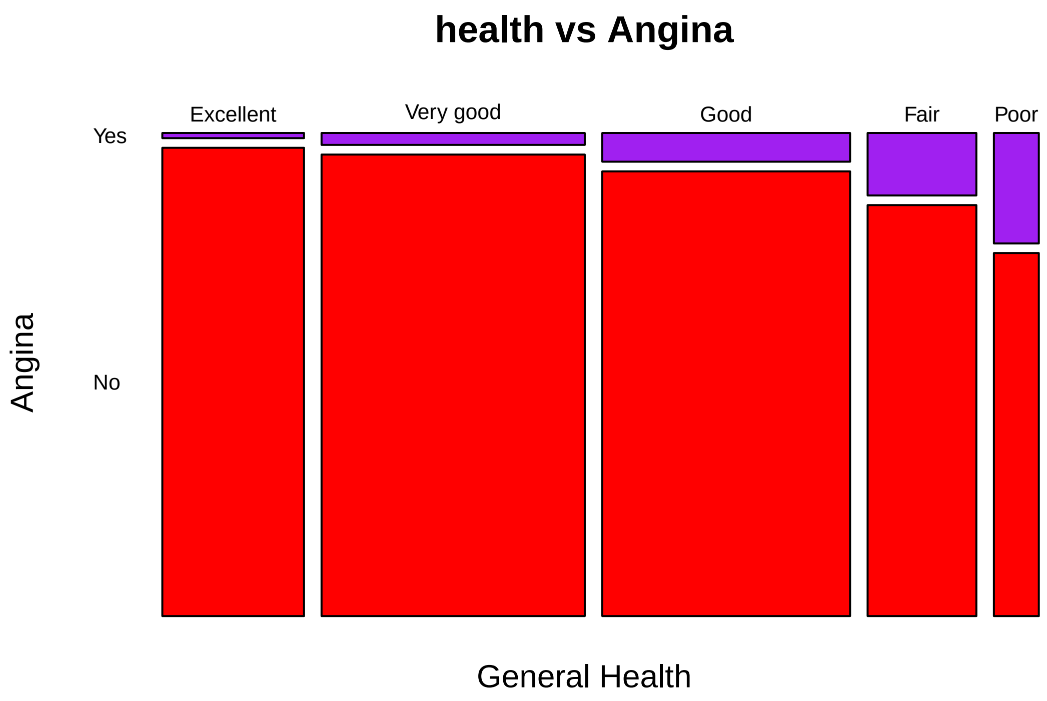
```
##       genhlth          sleptim1       cvdinfr4      cvdcrhd4      cvdstrk3
##  Excellent: 85482   Min.   :  0.000   Yes : 29284   Yes : 29064   Yes : 20391
##  Very good:159076   1st Qu.:  6.000   No  :459904   No  :458288   No  :469917
##  Good     :150555   Median :  7.000   NA's:  2587   NA's:  4423   NA's:  1467
##  Fair     : 66726   Mean   :  7.052
##  Poor     : 27951   3rd Qu.:  8.000
##  NA's     :  1985   Max.   :450.000
##                     NA's   :7387
##  asthma3       asthnow
##  Yes : 67204   Yes : 45644
##  No  :423012   No  : 19696
##  NA's:  1559   NA's:426435
##
##
```

```
##
##
```

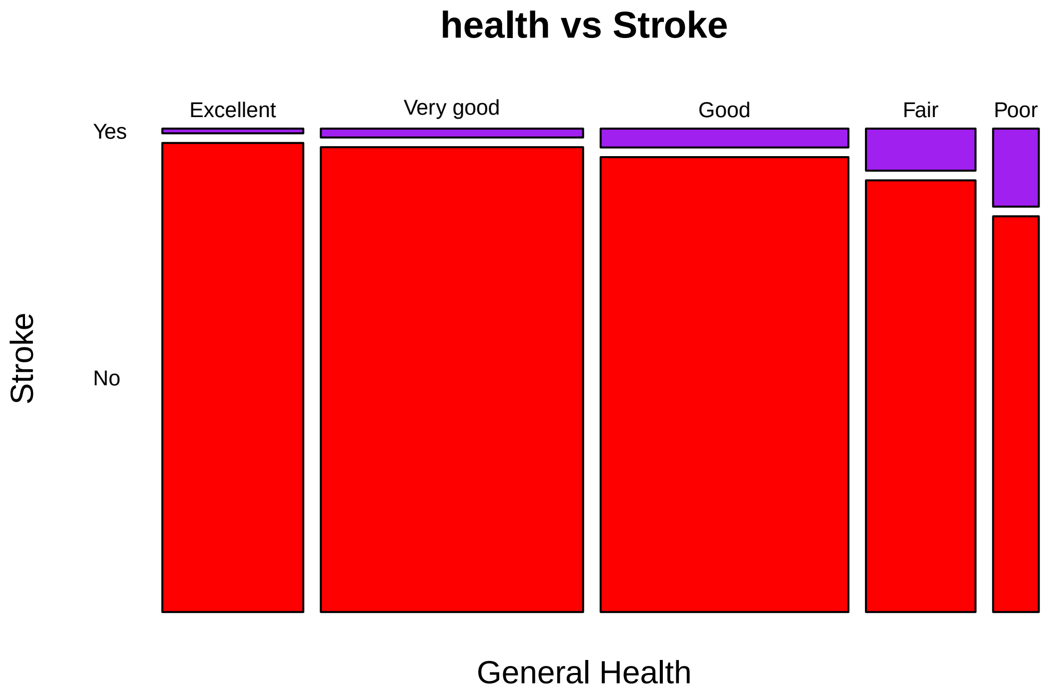
```
# par(mfrow = c(3, 2))
# cudinfr4: ever with heart attack

hAtt.plot <-mosaicplot( ~ genhlth + cvdinfr4,
                        data = brfss.q2,
                        xlab = "General Health",
                        ylab = "Heart Attack",
                        color = c( "purple", "red"),
                        main = "health vs heart attack",
                        las = 1
                        )
```



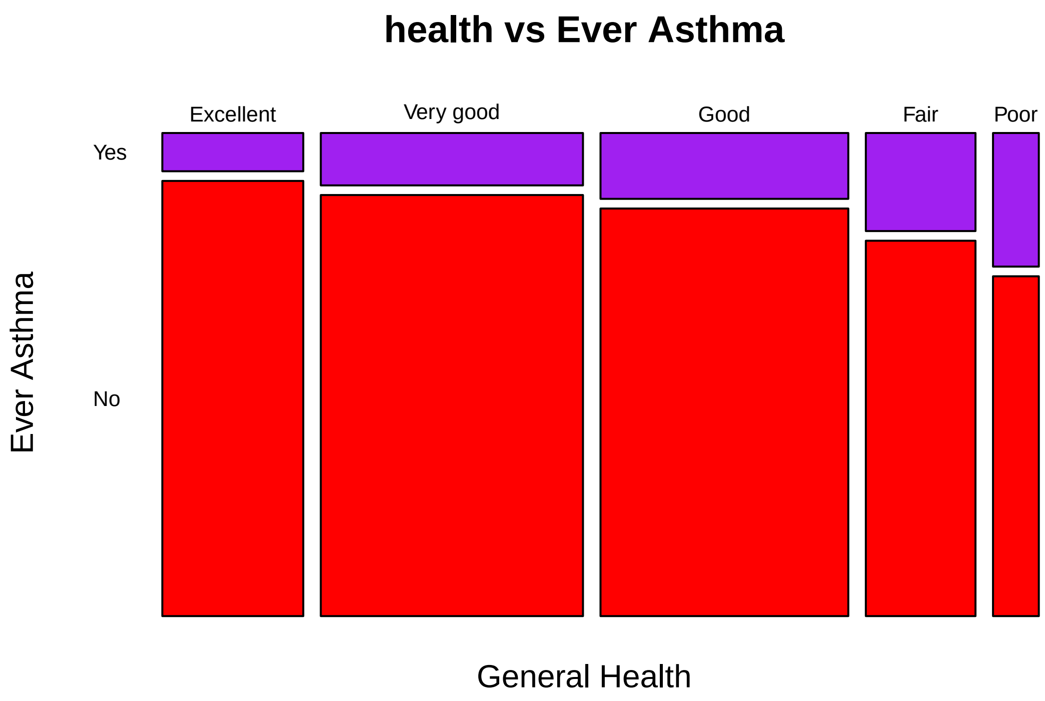
```
# cvdcrhd4: ever with angina or coronary heart disease

angina.plot <-mosaicplot( ~ genhlth + cvdcrhd4,
                        data = brfss.q2,
                        xlab = "General Health",
                        ylab = "Angina",
                        color = c( "purple", "red"),
                        main = "health vs Angina",
                        las = 1
                        )
```



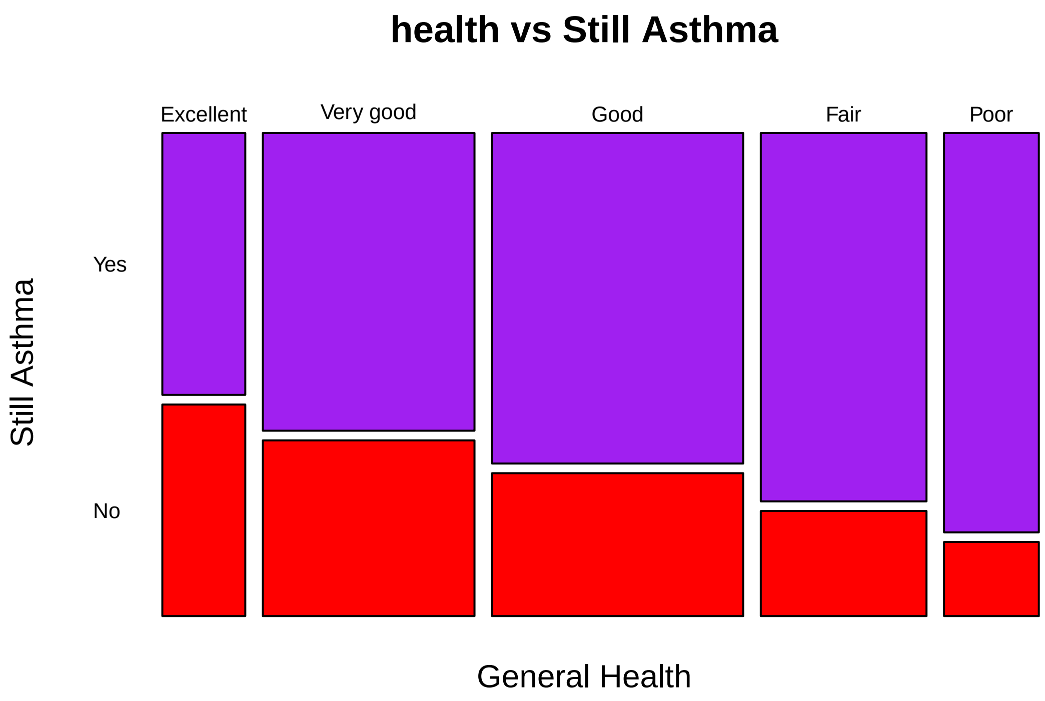
```r
# cvdstrk3: ever with with a stroke
strock.plot <-mosaicplot( ~ genhlth + cvdstrk3,
                      data = brfss.q2,
                      xlab = "General Health",
                      ylab = "Stroke",
                      color = c( "purple", "red"),
                      main = "health vs Stroke",
                      las = 1
                      )
```



```r
# asthma3: ever had asthma
eAsthma.plot <-mosaicplot( ~ genhlth + asthma3,
                      data = brfss.q2,
                      xlab = "General Health",
                      ylab = "Ever Asthma",
                      color = c( "purple", "red"),
                      main = "health vs Ever Asthma",
                      las = 1
                      )
```



```
# asthnow: still have asthma
sAsthma.plot <-mosaicplot( ~ genhlth + asthnow,
                      data = brfss.q2,
                      xlab = "General Health",
                      ylab = "Still Asthma",
                      color = c( "purple", "red"),
                      main = "health vs Still Asthma",
                      las = 1
                      )
```

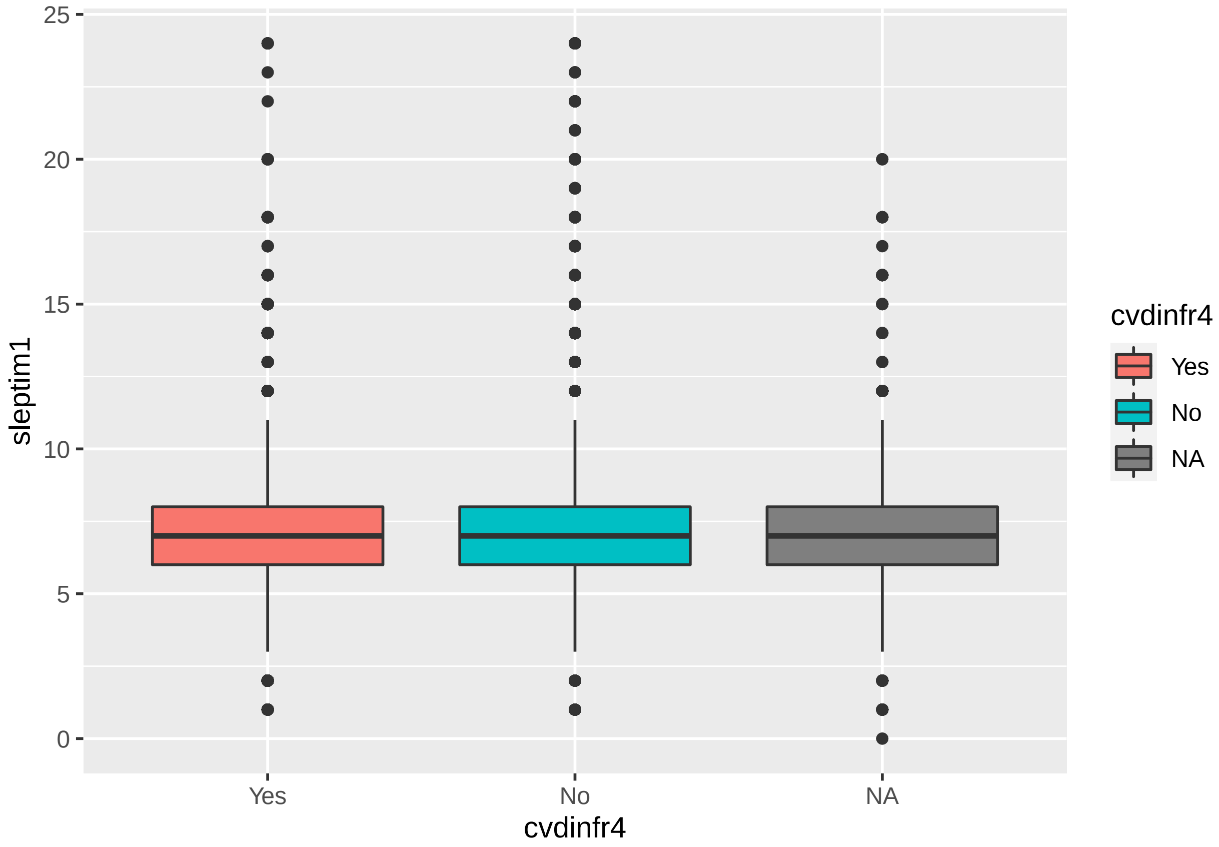


```r
# par(mfrow = c(1, 1))

# regroup by heart attack records
brfss.q2.hAttack <- brfss.q2 %>%
  select("sleptim1", "cvdinfr4")

brfss.q2.hAttack <- brfss.q2.hAttack %>%
                        drop_na()

brfss.q2.hAttack <- brfss.q2.hAttack %>%
                      group_by(cvdinfr4) %>%
                      summarize( mSleep = mean( sleptim1) )
brfss.q2.hAttack
```

```
## # A tibble: 2 x 2
##   cvdinfr4 mSleep
##   <fct>     <dbl>
## 1 Yes        7.07
## 2 No         7.05
```

```r
# boxplot sleepting time vs. heart attack
ggplot( brfss.q2, aes( x = cvdinfr4, y = sleptim1, fill = cvdinfr4)) +
  geom_boxplot() +
  scale_y_continuous( limits = c(0,24))
```

```
## Warning: Removed 7389 rows containing non-finite values (stat_boxplot).
```



```
# regroup by angina or coronary heart disease
brfss.q2.cvdcrhd4 <- brfss.q2  %>%
  select("sleptim1", "cvdcrhd4")

brfss.q2.cvdcrhd4 <- brfss.q2.cvdcrhd4 %>%
                        drop_na()

head( brfss.q2.cvdcrhd4)
```

```
##   sleptim1 cvdcrhd4
## 1        6       No
## 2        9       No
## 3        8       No
## 4        6       No
## 5        8       No
## 6        7       No
```

```
brfss.q2.cvdcrhd4 <- brfss.q2.cvdcrhd4 %>%
                      group_by(cvdcrhd4) %>%
                      summarize( mSleep = mean(sleptim1))

brfss.q2.cvdcrhd4
```

```
## # A tibble: 2 x 2
##   cvdcrhd4 mSleep
##   <fct>     <dbl>
## 1 Yes        7.06
## 2 No         7.05
```

```
# regroup by stroke records
# brfss.q2$cvdstrk3

brfss.q2.cvdstrk3 <- brfss.q2 %>%
                      select("sleptim1", "cvdstrk3")
brfss.q2.cvdstrk3 <- brfss.q2.cvdstrk3 %>%
                        drop_na()

head( brfss.q2.cvdstrk3)
```

```
##   sleptim1 cvdstrk3
## 1        6       No
## 2        9       No
## 3        8       No
## 4        6       No
## 5        8       No
## 6        7       No
```

```
brfss.q2.cvdstrk3 <- brfss.q2.cvdstrk3 %>%
                      group_by(cvdstrk3) %>%
                      summarize( mSleep = mean(sleptim1))

brfss.q2.cvdstrk3
```

```
## # A tibble: 2 x 2
##   cvdstrk3 mSleep
##   <fct>     <dbl>
## 1 Yes        7.13
## 2 No         7.05
```

After checking the mean sleeping time for individuals reporting chronic situations, the data did not show significant correlation between mean sleeping time and several chronic categories like heart attach, stroke and asthma records.

Alternatively, there is a significant relationship between general health and mean sleeping time.

-
-
-
-
-

Research quesion 3:

To explorer whether there's a correlation between Body Mass Index and diabetes, we first need to subset the dataset, only keep key columns like general health, bmi, and diet information like sugar drinks, and fruits/veggies consumsion.

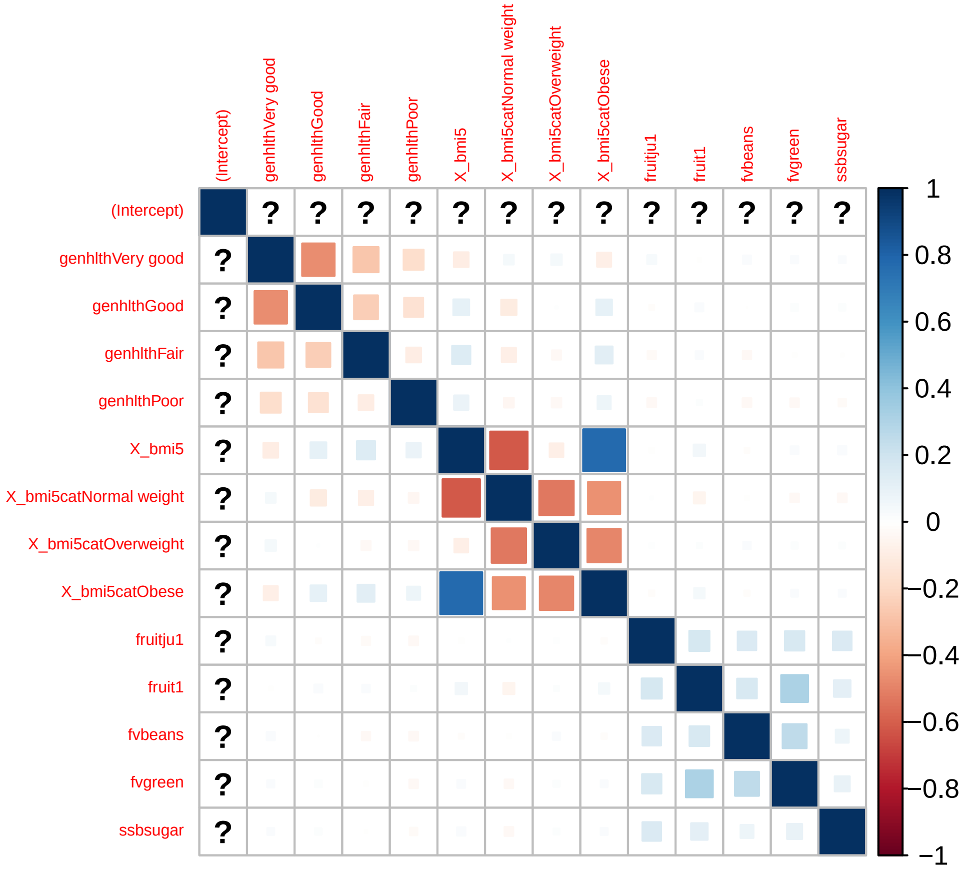
```r
col.q3 <- c("genhlth", "X_bmi5", "X_bmi5cat", "X_rfbmi5", "fruitju1", "fruit1", "fvbeans", "fvgreen", "s
brfss.q3 <- brfss2013[, col.q3]

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

dim(brfss.q3)
```

```
## [1] 94147     9
```

```r
summary( brfss.q3 )
```

```
##        genhlth          X_bmi5               X_bmi5cat     X_rfbmi5
##  Excellent:16911   Min.   :1213   Underweight   : 1505   No :32189
##  Very good:31898   1st Qu.:2373   Normal weight:30684   Yes:61958
##  Good     :27753   Median :2696   Overweight   :33813
##  Fair     :12243   Mean   :2799   Obese        :28145
##  Poor     : 5342   3rd Qu.:3100
##                    Max.   :8644
##     fruitju1          fruit1          fvbeans          fvgreen
##  Min.   :  0.0   Min.   :  0.0   Min.   :  0.0   Min.   :  0.0
##  1st Qu.:  0.0   1st Qu.:101.0   1st Qu.:201.0   1st Qu.:103.0
##  Median :101.0   Median :201.0   Median :204.0   Median :203.0
##  Mean   :134.6   Mean   :180.7   Mean   :211.6   Mean   :211.9
##  3rd Qu.:301.0   3rd Qu.:303.0   3rd Qu.:304.0   3rd Qu.:305.0
##  Max.   :399.0   Max.   :399.0   Max.   :399.0   Max.   :399.0
##     ssbsugar
##  Min.   :  0.0
##  1st Qu.:  0.0
##  Median :101.0
##  Mean   :124.7
##  3rd Qu.:301.0
##  Max.   :399.0
```

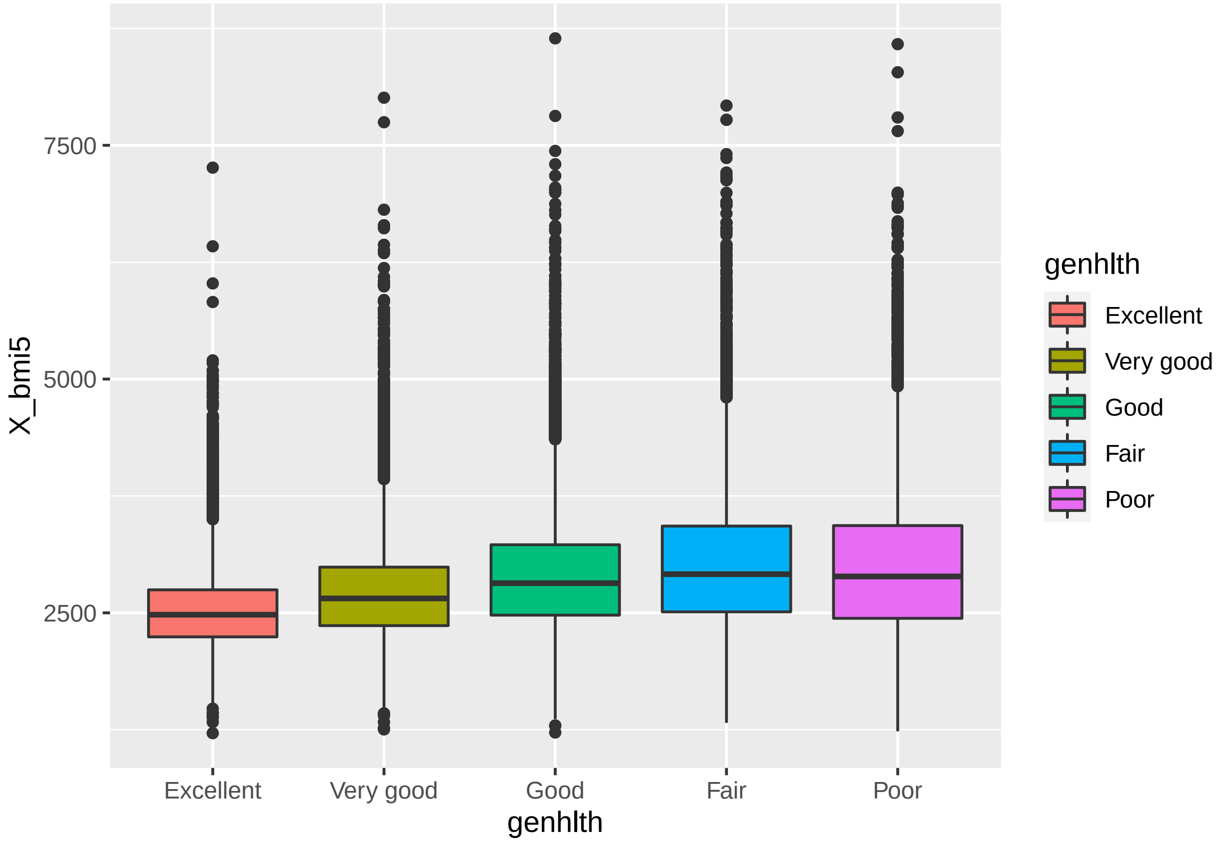
```r
# plot correlation matrix

brfss.q3.m <- model.matrix( X_rfbmi5 ~ ., data = brfss.q3)

corrplot( cor( brfss.q3.m), method = "square", tl.cex = 0.5)
```



From the correlation plot, we could see positive correlation between BMI and fruits/veggies consumsion.

We also want to explorer the mean Computed Body Mass Index in different general health categories.

```r
ggplot( brfss.q3, aes( x = genhlth, y = X_bmi5, fill = genhlth)) +
  geom_boxplot()
```

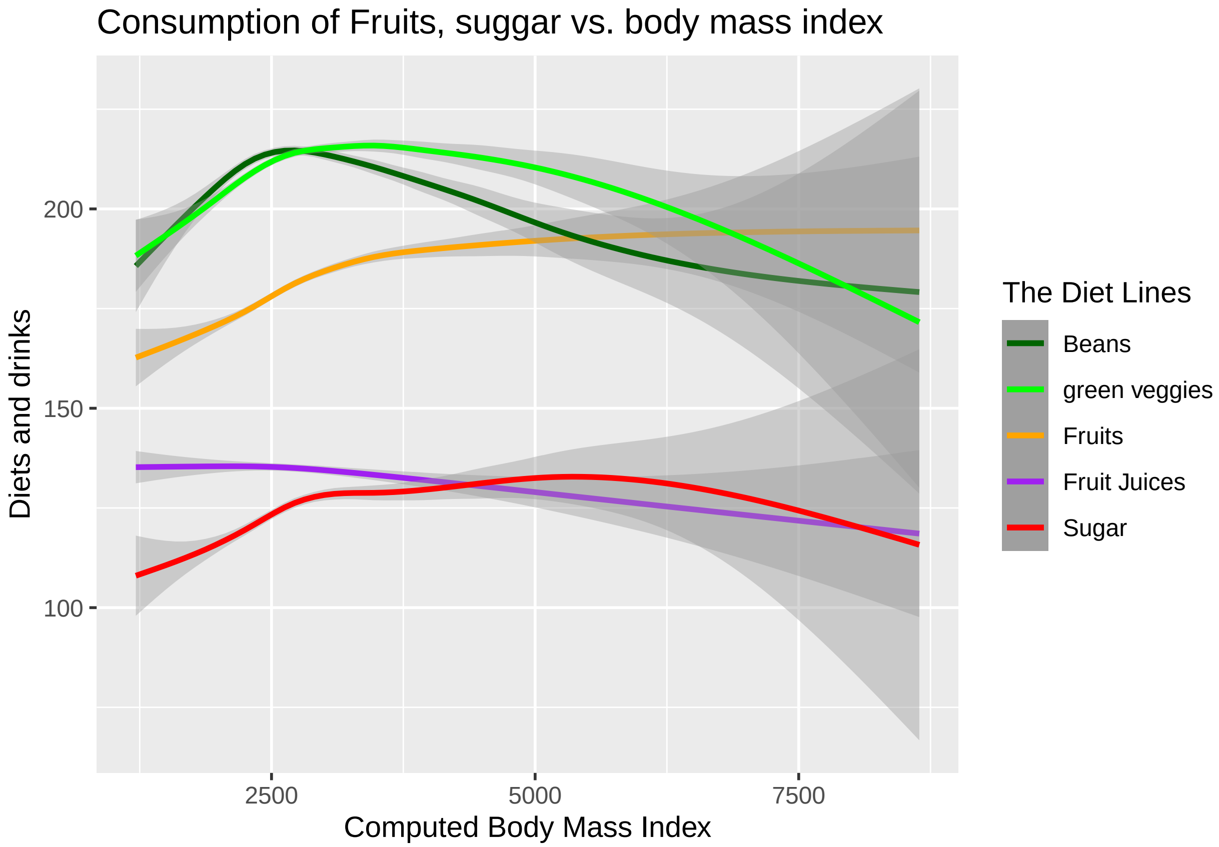


The above boxplot shows those individuals claimed to be in Excellent condition have BMI around 2500, as the general conditions go from excellent to very good, good, fair, and poor, the BMI also increase.

The next step, we would like to investigate whether there are correlations between BMI and diets and drinks.

```
ggplot() +
  geom_smooth( data = brfss.q3, aes( x = X_bmi5, y=fruitju1, color = "purple"))+
  geom_smooth( data = brfss.q3, aes( x = X_bmi5, y=fruit1, color = "orange"))+
  geom_smooth( data = brfss.q3, aes( x = X_bmi5, y=fvbeans, color = "darkgreen"))+
  geom_smooth( data = brfss.q3, aes( x = X_bmi5, y=fvgreen, color = "green"))+
  geom_smooth( data = brfss.q3, aes( x = X_bmi5, y=ssbsugar, color = "red"))+
  scale_color_manual( name = "The Diet Lines",
                      values = c("purple" = "purple", "orange"="orange", "darkgreen" = "darkgreen", "gr
                      labels = c("Beans", "green veggies", "Fruits", "Fruit Juices", "Sugar") )+
  labs( title = "Consumption of Fruits, suggar vs. body mass index",
        x = "Computed Body Mass Index",
        y = "Diets and drinks")
```

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



From the trend plot above, it is very interesting to see there's a positive correlation between consumption of fruits and computed body mass index. In our general opinion, eating more fruits would reduce the BMI, however that's not what we saw from the plot. For individuals eating beans and green veggies, we could see the positive correlation before BMI reachs 2500, after that the more of the consumpsion of beans and green veggies the less of the BMI. For sugar consumtion, we could see a rise of BMI as individuals drinking more sugar before the BMI reaches 6000, after that the BMI would drop even if individuals drink more suggar drinks. For individuals drinking fruit juice, there's a negative correlation with BMI, as individuals drink more fruit juice, their BMI would drop.

Discussion: In summary, all plots and tables generated in this document support most of our hypothesis in the three research questions we wanted to ask. 1. There are positive correlations between education levels and income levels, which would contribute positively to the general health categories.

2. The correlation between sleeping duration and chronic health conditions are not very significant.
3. There are significant relationships between diets and computed body mass index.