Exploring the BRFSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(pander)
library(tidyr)
```

Load data

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

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRSS2013) objective is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population. In conducting the BRFSS landline telephone survey, interviewers collect data from a randomly selected adult in a household. 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. Some questions ask the respondent information about actions and events happened in the past 30 days of even before, and some people might not remember accurately what they have done such long time before. Another issue is that the interview are contacted over the phone and participation is voluntary, therefore there is no random assignment and also people who do not own a telephone (landline and/or mobile) and do not live in a private residence are excluded from the study. In fact, according the Centers for Disease Control and Prevention (CDC) website, “No direct method of accounting for non-telephone coverage is employed by the BRFSS”. From people who have these requisite, living in a private residence/college and own a landline or mobile line, of non-institutionalized adult population, aged 18 years or older, a random sample from each state has been selected, therefore this is a stratified random sample (according to the CDC website : Home telephone numbers are obtained through random-digit dialing). We can assume independence of the random sampling even if some of the interviews were contacted over mobile phones and therefore there might be the chances of having interviewed two people from the same household on their personal mobile phones. However we can consider this chance very small. This is an observational study, because it is asking respondent about their past actions, it is not an experimental study as there has not been any treatment randomly assigned to respondents. For this reason there is no causality but we can only observe association between the variables.

Part 2: Research questions

Research quesion 1: Does sleeping time affect physical and mental health? Are people who play sport more affected?

Research quesion 2: Do people with higher income have also better health conditions? Are the results affected by whether one smokes or not?

Research quesion 3: Do people who earn more have fewer chances to get depressive disorder?

Part 3: Exploratory data analysis

Research quesion 1:

Does sleeping time affect physical and mental health? Are people who play sport more affected?

I want to explore the effect of sleeping time on physical and mental health. I will check then if the results are affected by playing any sport. In order to answer this question I am using the following variables from the BFRSS2013 dataset. 1. sleptim1: How Much Time Do You Sleep (On average, how many hours of sleep do you get in a 24-hour period?) 2. poorhlth: Poor Physical Or Mental Health (During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?) 3. exerany2: Exercise In Past 30 Days (During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?)

I clean these variables from NA values and I store them in a new dataset called q1.

```
q1<-brfss2013 %>%filter(!(is.na(sleptim1)),!(is.na(poorhlth)),!(is.na(exerany2)))
```

I then explore the variables of time of sleeping (sleptim1) and days that the respondent reported poor physical and menthal health in the prevuso 30 days. We explore the variable with a summary statistics of each:

```
q1 %>%
  summarise(slepmean = mean(sleptim1), slepmedian = median(sleptim1), slepsd = sd(sle
ptim1), slepmin = min(sleptim1), slepmax = max(sleptim1))
```

```
##   slepmean slepmedian   slepsd slepmin slepmax
## 1 6.922381          7 1.611705       1      24
```

```
q1 %>%
  summarise(hlthmean = mean(poorhlth), hlthmedian = median(poorhlth), hlthsd = sd(poo
rhlth), hlthmin = min(poorhlth), hlthmax = max(poorhlth))
```

```
##   hlthmean hlthmedian   hlthsd hlthmin hlthmax
## 1 5.180428          0 9.299675       0      30
```

We can see that the mean of hours of sleeping is 6.9 hours with a standard deviation of 1.6 and the mean of the days reporting poor health in the past 30 days is 5.1 days with a standard deviation of 9.3. The variable reporting whether the respondent has done any physical exercises in the past 30 days, exerany2 , is a discrete variable with answers yes and no, we tabulate the data to see the responses.

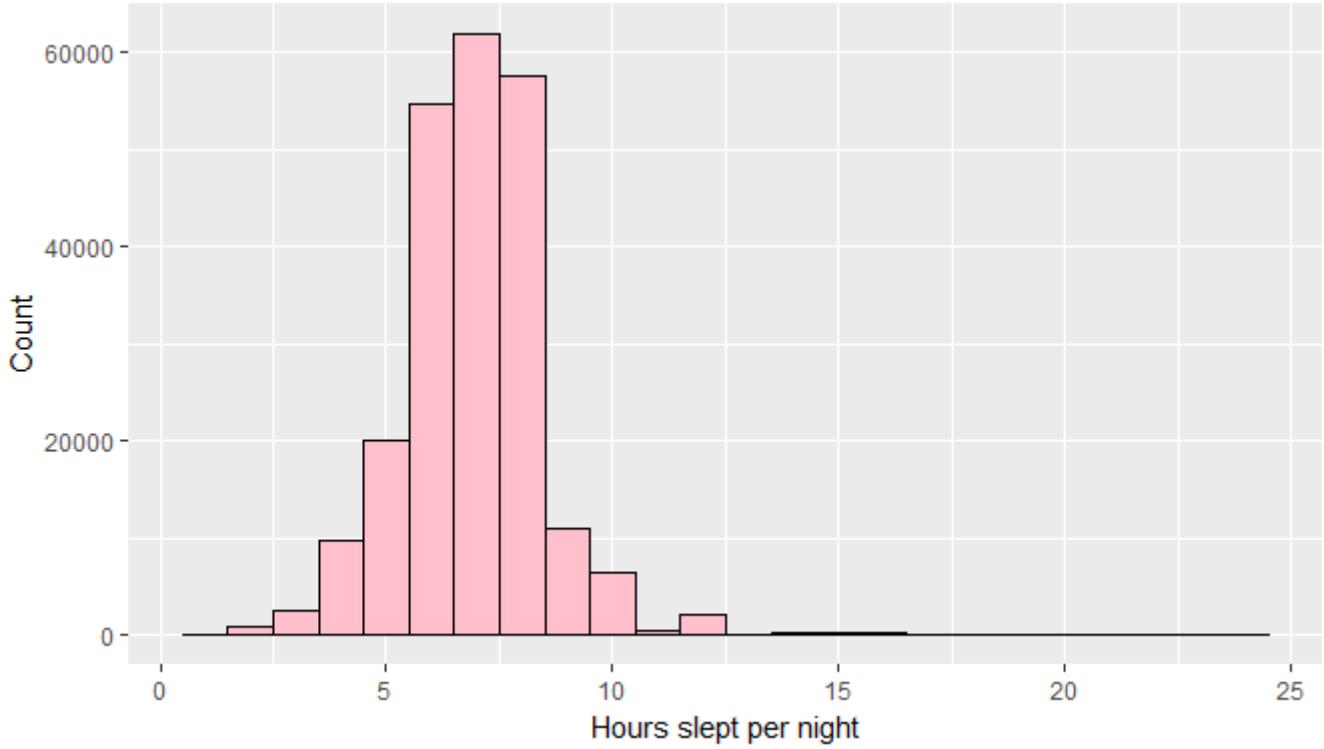
```
q1 %>%
group_by(exerany2) %>%
summarise(count = n())
```

```
## # A tibble: 2 x 2
##   exerany2  count
##   <fct>     <int>
## 1 Yes      157688
## 2 No        71367
```

From the results, there are 157688 respondent in the dataset who exercises in the last 30 days and 71367 who did not.

I now create a graphical representation of the respondents on the hour of sleeping variable.

```
PlotSleep <- ggplot(q1) + aes(x=sleptim1)
PlotSleep <- PlotSleep + geom_histogram(binwidth = 1, color="black", fill="pink")
PlotSleep <- PlotSleep + xlab("Hours slept per night") + ylab("Count")
PlotSleep
```



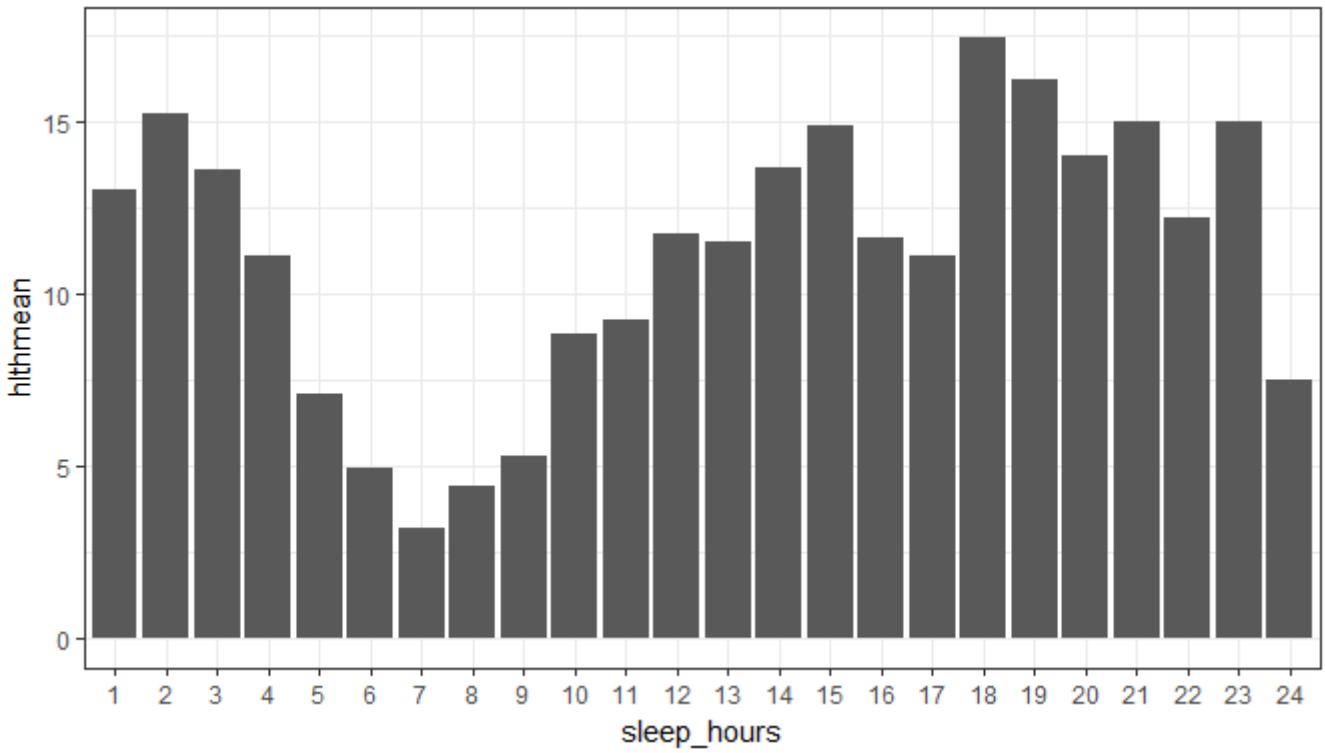
I create a new variable slep_hours encoding sleptim1 as a factor and hlthmean containing the mean of the variable poorhlth. I store these variables in the dataset sleephealth. I tabulate the average of reported days of poor mental and physical health against the hours of sleep.

```
sleephealth<-q1 %>%
group_by(sleep_hours = as.factor(sleptim1))%>%
summarise(hlthmean=mean(poorhlth),count=n())
sleephealth
```

```
## # A tibble: 24 x 3
##    sleep_hours hlthmean count
##    <fct>          <dbl> <int>
##  1 1               13.0    157
##  2 2               15.2    784
##  3 3               13.6   2545
##  4 4               11.1   9748
##  5 5                7.10 20104
##  6 6                4.96 54808
##  7 7                3.22 61958
##  8 8                4.41 57683
##  9 9                5.29 10920
## 10 10               8.83  6552
## # ... with 14 more rows
```

Then I visualize the results in a graph

```
ggplot(sleephealth,aes(x=sleep_hours,y=hlthmean))+geom_bar(stat = 'identity') + theme_bw()
```



As we can see, those who reported the least days of poor mental and physical health, are those who sleep on average 7 hours. As the number on hours of sleep move away from the average f 7 hours, the number of days of poor health increases. This means that sleeping a lot more than 7 hour or a lot less, are reporting the highest number of poor health. I subset the data to include only those who exercise, because I want to see if the above findings are changing if people do some physical exercise.

```
q1ex<-subset(q1, exerany2=="Yes")
```

I now recalculate a summary statistics of the variable sleep and poor health and I create a new graph which now will include only those who exercise.

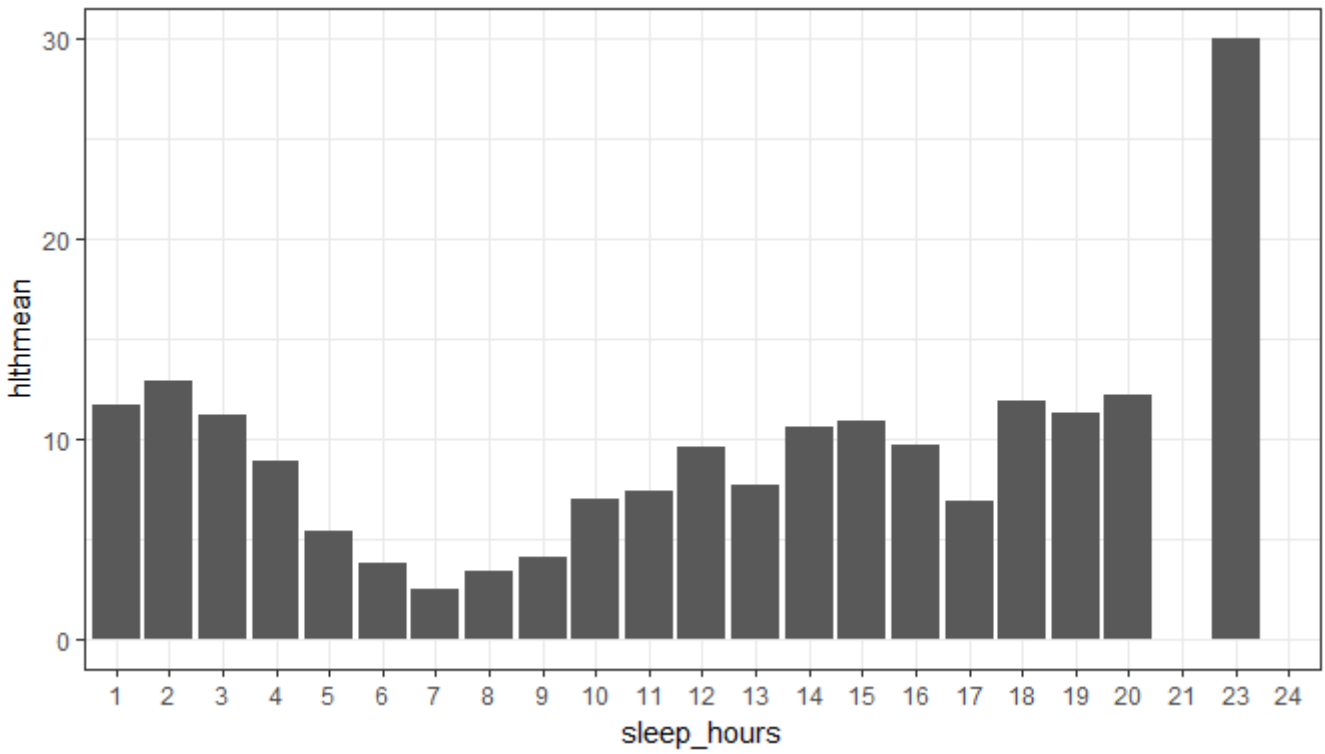
```
q1ex %>%
  summarise(slepmean = mean(sleptim1), slepmedian = median(sleptim1), slepsd = sd(sle
ptim1), slepmin = min(sleptim1), slepmax = max(sleptim1))
```

```
##   slepmean slepmedian   slepsd slepmin slepmax
## 1 6.941207          7 1.473396       1      24
```

```
q1ex %>%
  summarise(hlthmean = mean(poorhlth), hlthmedian = median(poorhlth), hlthsd = sd(poo
rhlth), hlthmin = min(poorhlth), hlthmax = max(poorhlth))
```

```
##   hlthmean hlthmedian   hlthsd hlthmin hlthmax
## 1 3.864746          0 7.824401       0      30
```

```
sleephealthex<-q1ex %>%
group_by(sleep_hours = as.factor(sleptim1))%>%
summarise(hlthmean=mean(poorhlth),count=n())
ggplot(sleephealthex,aes(x=sleep_hours,y=hlthmean))+geom_bar(stat = 'identity') + the
me_bw()
```



From the results we see that the means, median and standard deviations have not changed much and also the graph looks very similar to the previous one and has the same trend specifically, the average of reported days of poor health is the lowest at 7 hours of sleep and increases as the hours of sleep increase or decrease. Therefore we can infer that the there is a correlation between the hours a person sleep per day and they general health, and this is irrespective of whether that person practice exercise regularly or not.

Research quesion 2:

Do people with higher income have also better health conditions? Are the results affected by whether one smoke or not? The variables that I use to answer this question are: 1. educa: education level - What is the highest grade or year of school you completed? 2. genhlth: Would you say that in general your health is?

3. X_rfsmok3: Adults who are current smokers

```
brfss2013%>%group_by(educa)%>%summarise(count=n())
```

```
## # A tibble: 7 x 2
##   educa                                                    count
##   <fct>                                                    <int>
## 1 Never attended school or only kindergarten                 677
## 2 Grades 1 through 8 (Elementary)                          13395
## 3 Grades 9 though 11 (Some high school)                    28141
## 4 Grade 12 or GED (High school graduate)                  142971
## 5 College 1 year to 3 years (Some college or technical school) 134197
## 6 College 4 years or more (College graduate)              170120
## 7 <NA>                                                      2274
```

```
brfss2013%>%group_by(genhlth)%>%summarise(count=n())
```

```
## # A tibble: 6 x 2
##   genhlth    count
##   <fct>      <int>
## 1 Excellent  85482
## 2 Very good 159076
## 3 Good      150555
## 4 Fair       66726
## 5 Poor       27951
## 6 <NA>        1985
```

```
brfss2013%>%group_by(X_rfsmok3)%>%summarise(count=n())
```

```
## # A tibble: 3 x 2
##   X_rfsmok3  count
##   <fct>      <int>
## 1 No        399786
## 2 Yes        76654
## 3 <NA>       15335
```

We explore the variables and see that all of them have some NAs values. I create a new dataset with the variables cleaned of the NAs values and I summarize them again.

```
q2<-brfss2013 %>%filter(!(is.na(genhlth)),!(is.na(educa)),!(is.na(X_rfsmok3)))
q2%>%group_by(educa)%>%summarise(count=n())
```

```
## # A tibble: 6 x 2
##   educa                                                    count
##   <fct>                                                    <int>
## 1 Never attended school or only kindergarten                 616
## 2 Grades 1 through 8 (Elementary)                          12647
## 3 Grades 9 though 11 (Some high school)                    26889
## 4 Grade 12 or GED (High school graduate)                  137559
## 5 College 1 year to 3 years (Some college or technical school) 130158
## 6 College 4 years or more (College graduate)              165436
```

```
q2%>%group_by(genhlth)%>%summarise(count=n())
```

```
## # A tibble: 5 x 2
##   genhlth    count
##   <fct>      <int>
## 1 Excellent  82438
## 2 Very good 154370
## 3 Good      145207
## 4 Fair       64355
## 5 Poor       26935
```

```
q2%>%group_by(X_rfsmok3)%>%summarise(count=n())
```

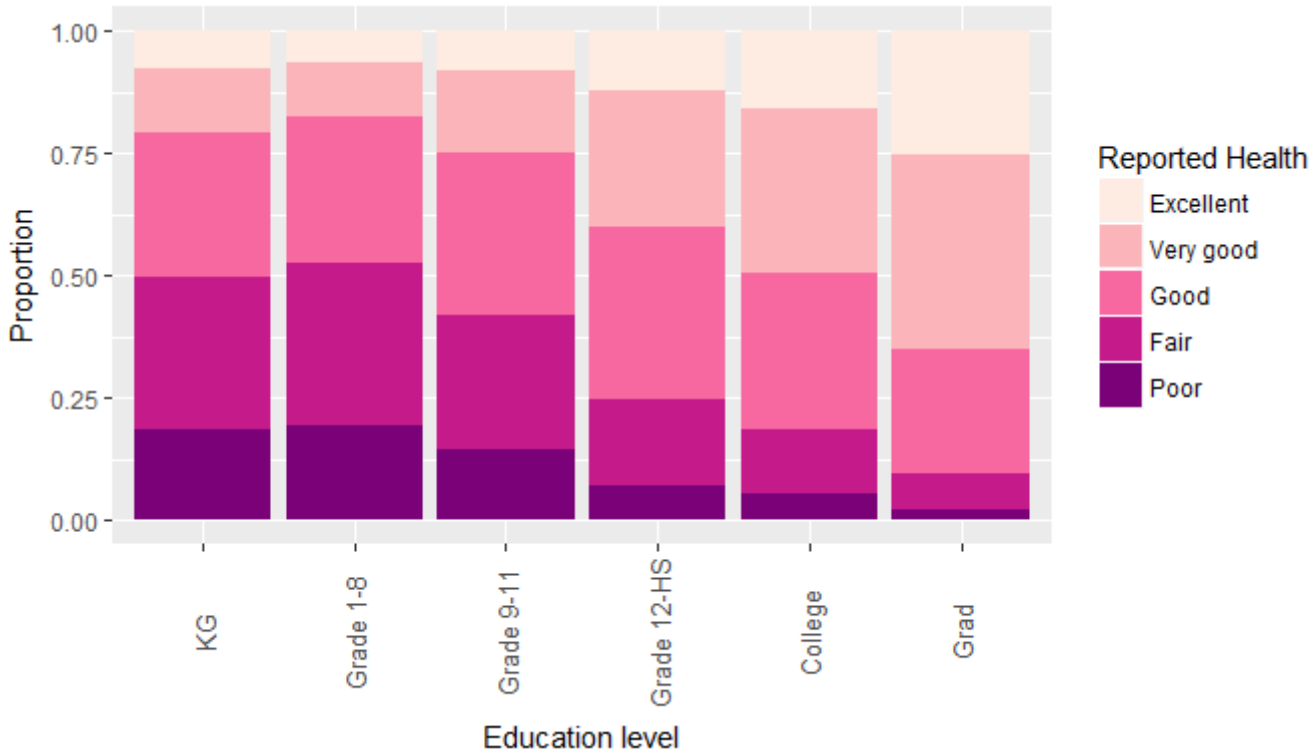
```
## # A tibble: 2 x 2
##   X_rfsmok3  count
##   <fct>      <int>
## 1 No        397105
## 2 Yes        76200
```

I change the values of the education variable (educa) variable coding education levels with shorter labels that are easier to read and to work with.

```
levels(q2$educa) <- c("KG", "Grade 1-8", "Grade 9-11", "Grade 12-HS", "College", "Gra
d")
```

I plot the two variables education level and level of general health in a barplot.

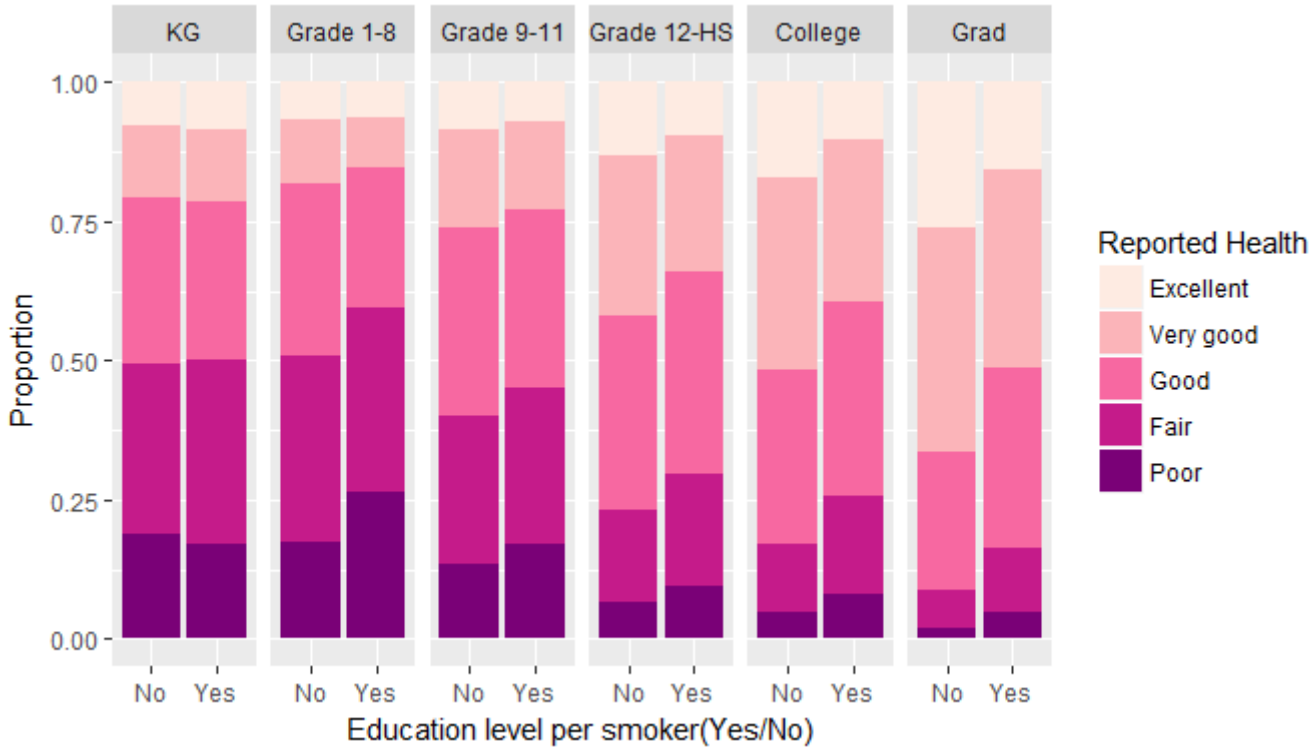
```
g <- ggplot(q2) + aes(x= educa,fill=genhlth) + geom_bar(position = "fill")
g <- g + xlab("Education level") + ylab("Proportion") + scale_fill_brewer(name="Repor
ted Health", palette = "RdPu") + theme(axis.text.x = element_text(angle =90, size = 1
0, vjust = 0.5))
g
```



This graph shows that as level of education completed increases, the proportion of people reporting a better general health increases. In fact the proportion of respondent reporting excellent health for example, is highest in the group of people holding a college degree.

I now add to the graph the variable for smokers and therefore each group in the education level will be separated according if they smoke or not, to see if there is any difference between the two groups.

```
gsmoke <- ggplot(q2) + aes(x= X_rfsmok3,fill=genhlth) + geom_bar(position = "fill") +
facet_grid(.~educa)
gsmoke <- gsmoke + xlab("Education level per smoker(Yes/No)") + ylab("Proportion") +
scale_fill_brewer(name="Reported Health", palette = "RdPu")
gsmoke
```



From this graphic representation we see that indeed there is a difference in the health reported by respondent according to whether they are smokers or not. There are some interesting findings from this graph. The difference of reported health amongst respondent with lowest education, does not vary much according to whether they are smokers of not. Also the gap of non-smokers reporting better health condition than smokers increases as education level increases. We can assume that amongst all the respondent people with higher education have the highest proportion of good/excellent health, and among them the highest proportion reporting excellent health is of those who do not smoke. Therefore being a smoker seems to make a difference on reported health especially on those who have a higher education. This is an interesting finding, as we might generally think that the effect of smoking on the body should be irrelevant of the education of the individual.

Research quesion 3:

Do people who earn more have fewer chances to get depressive disorder? This question wants to explore whether from the BRFSS we can find some correlation between the level of income of the respondent and the having a depressive disorder. The reason for exploring this possible correlation is to see if the level of income has also an impact on psychological life and not only on material life. The variable I am using for this analysis are: 1. income2: income2: Income Level - Is your annual household income from all sources 2. addepev2: ever told you had a Depressive Disorder - (Ever told) you that you have a depressive disorder, including depression, major depression, dysthymia, or minor depression?

I know from the codebook that these variables have NAs value, so I remove them from the variables and store them in a new dataset called q3.

```
q3 <- brfss2013 %>%
filter(!is.na(income2),!is.na(addepev2)) %>%
select(income2, addepev2)
```

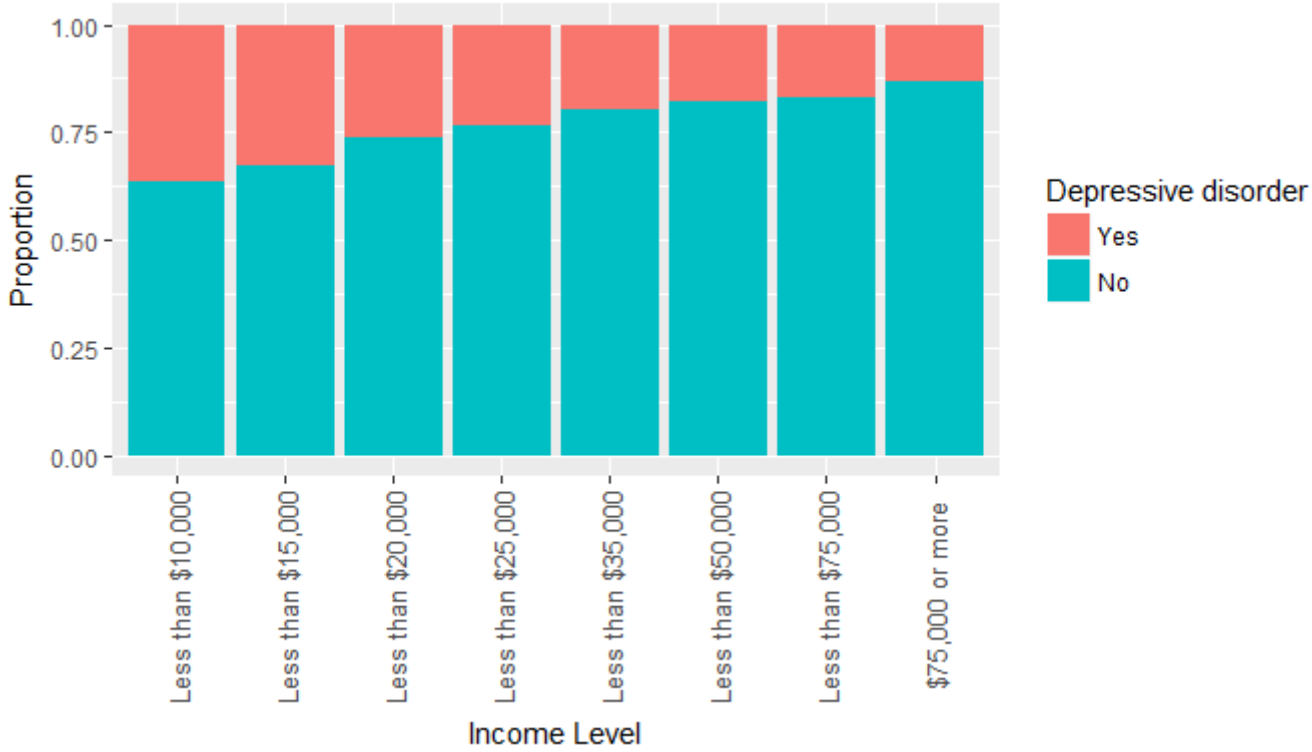
I create a table to summarize the number of respondent per income level according to whether they have been ever diagnosed a depressive disorder or not. And showing the percentage of those who have been diagnosed some type of depressive disorder, per each income level.

```
Table <- q3 %>%
group_by(income2, addepev2) %>%
summarize(Sum = n()) %>%
spread(addepev2, Sum) %>%
mutate(Sum = Yes+No, `% Yes` = round(Yes/Sum*100, digits =1))
colnames(Table)[1] <- "Income Level"
pandoc.table(Table, caption = "Fig. 1 - Summary Statistics for if Respondent has ever
been diagnosed a depressive disorder by Income Level", justify = "center")
```

```
##
## ------------------------------------------------------
##    Income Level        Yes       No       Sum      % Yes
## ------------------- ------- -------- -------- -------
##  Less than $10,000    9240    15983    25223     36.6
##
##  Less than $15,000    8731    17892    26623     32.8
##
##  Less than $20,000    9152    25564    34716     26.4
##
##  Less than $25,000    9717    31811    41528     23.4
##
##  Less than $35,000    9609    39061    48670     19.7
##
##  Less than $50,000    11056   50259    61315      18
##
##  Less than $75,000    10894   54187    65081     16.7
##
##   $75,000 or more     15448   100238   115686    13.4
## ------------------------------------------------------
##
## Table: Fig. 1 - Summary Statistics for if Respondent has ever been diagnosed a dep
ressive disorder by Income Level
```

From the table we can see that generally the proportion of people diagnosed with a depressive disorder increases as the income level decrease. In fact the highest proportion is to be found amongst those with the lowest income level. I also visualize the results in a graph.

```
g3 <- ggplot(q3) + aes(x=income2,fill=addepev2) + geom_bar(position = "fill")
g3 <- g3 + xlab("Income Level") + ylab("Proportion") + scale_fill_discrete(name="Depr
essive disorder")+ theme(axis.text.x = element_text(angle =90, size = 10, vjust = 0.5
))
g3
```



In this graph we see the same results as the table, and namely when we consider the respondent according to their income level, the number of those who have ever been diagnosed a depressive disorder is higher amongst those with a lower income. Further study can be conducted to analyze this association and see if this is due to change or if there is an actual correlation between income level and the having a depressive disorder, and in case the reasons behind this.