Exploring the BRFSS data

Setup

The Behavioral Risk Factor Surveillance System (BRFSS) is an ongoing surveillance system designed to measure behavioral risk factors for the non-institutionalized adult population (18 years of age and older) residing in the US.

The BRFSS 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. Factors assessed by the BRFSS in 2013 include 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.

BRFSS conducts both landline telephone- and cellular telephone-based surveys.Interviewers collect data from a randomly selected adult in a household.

There is a “Main Survey” and “Optional Modules”. The main survey consists of the following parts:

- Record Identification
- Health Status
- Healthy Days - Health-Related Quality of Life
- Health Care Access
- Inadequate Sleep
- Hypertension Awareness
- Cholesterol Awareness
- Chronic Health Conditions
- Demographics
- Tobacco Use
- Alcohol Consumption
- Fruits and Vegetables
- Exercise (Physical Activity)
- Arthritis Burden
- Seatbelt Use
- Immunization
- HIV/AIDS

Goal of the study: Analysis of the General Health of US citizens and its dependency on several factors such sex, BMI (Body Mass Index) and Educational level.

Load packages

```
library(ggplot2)
library(dplyr)
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 3.2.5
```

```
library("cowplot")
```

```
## Warning: package 'cowplot' was built under R version 3.2.5
```

Load data

```
load("/Users/anacaballero-herrera/Documents/Stat_Duke_Univ/project/brfss2013.RData")
```

Part 1: Data

The data is stored as a data frame with: 491775 obs. of 330 variables. The variables are Categorical and Numerical (Integers). Categorical variables are more common than numerical. The data presents many difficulties: It contains many non-answered questions (), there are too many non-answered questions thus difficulting the statistical analysis, there are variables that looks as numerical but are strings, there are "many" ??nrealistic answers and decisions have to be taken in order to accept or reject them in order to perform a correct statistical analysis.

The variables we are going to use are: General Health `genhlth` , Sex `sex` , Degree of Education `educa` , Weight `weight2` , Height `Height`

Several filtrations and changes have been performed:

1. **Height:** `height.m`
    - Change of units: from feet and inches –> cm. Heights lie between: 75cm< Height<240cm

```
my_data.fltr<- brfss2013 %>%
 filter(height3>106 & height3 < 803)   #803=251 cm
# Conversion feets & in to cm
my_data.fltr<-my_data.fltr %>%
  mutate(height.m=((height3%/%100)*30.48+(height3%%100)*2.54))
```

2. **Weight:** `weight.kg`
    - Change from categorical –> integer. Wheights in the interval 25 kg<Weight<200 Kg

```
# Weight is a categorical convert into integer
my_data.fltr<-my_data.fltr %>%
  mutate(weight.kg=as.integer(weight2))
# Filter unrealistic weights
my_data.fltr<-my_data.fltr %>%
  filter(weight.kg = (weight.kg <201 & weight.kg>24))
```

3. The **NA** observations are removed in the plots.
4. **Body Mass Index BMI** `BMI`
    - The Body Mass Index(BMI) is world wide extensively used for measuring the degree of "thikness/obesity" for men and women. The BMI is defined as:

$$\text{BMI} = \frac{\text{Weight(kg)}}{\text{Height}^2(\text{m}^2)}$$

```
my_data.fltr <-my_data.fltr %>%
  mutate(BMI=weight.kg/(height.m/100)^2)
```

5. **Classification by BMI:** `BMI.Class`

```
my_data.fltr <- my_data.fltr %>%
           mutate(BMI.Class = ifelse(BMI <18.5, "1.Underweight",
                      ifelse(BMI >= 18.50 & BMI< 25, "2.Normal",
                      ifelse(BMI >= 25 & BMI < 30, "3.Overweight",
                      ifelse(BMI >= 30 & BMI < 40,"4.Obese","5.Morbidity")))))
```

Part 2: Research questions

Q1: How is the general health among the USA population for females and males

Q2: How are the weigth, height and Body Mass Index (BMI) distributions among USA females and males:

Q3: Which factor has a greater impact on health BMI or Education ?

Part 3: Exploratory Data Analysis

Q1: General health

General Health - Percentage

```
x=table(my_data.fltr$genhlth)     #All interviewed -- genhlth is categorical (no summ
ary)
round(x/sum(x),2)    # Rate, two decimals
```

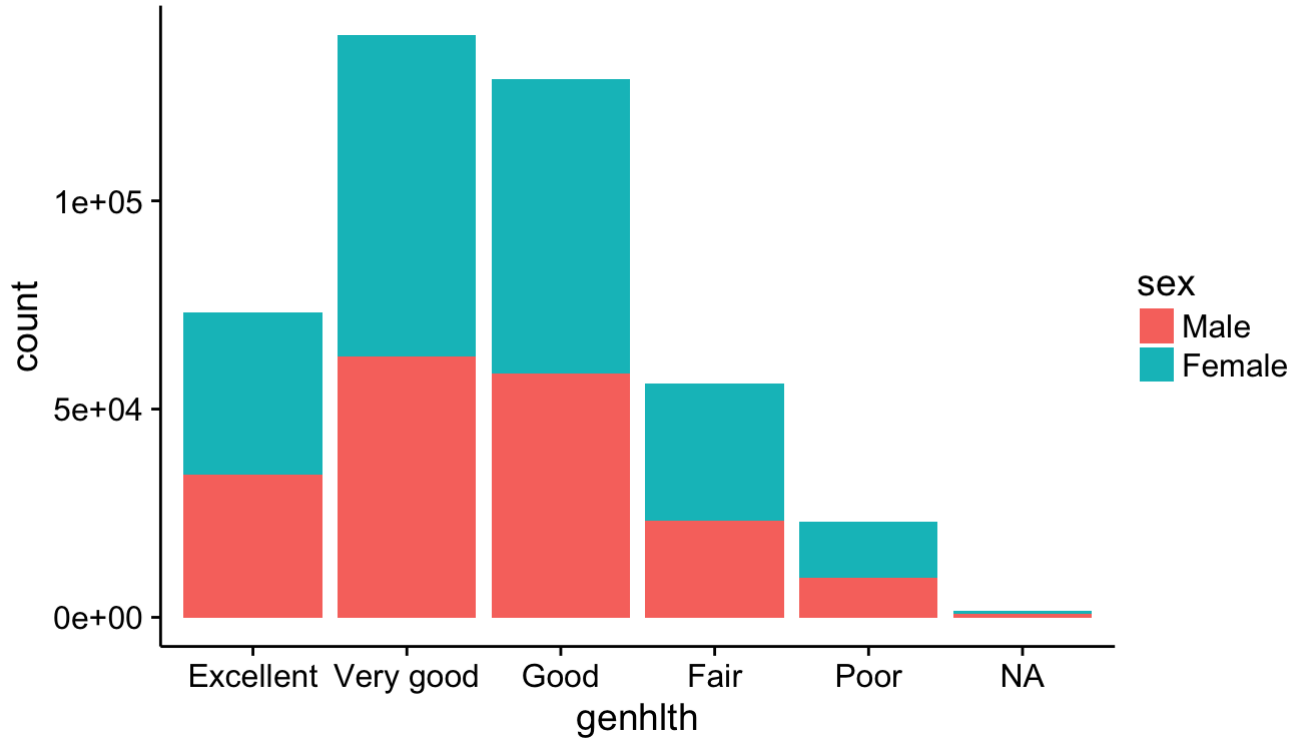
```
##
## Excellent Very good      Good      Fair      Poor
##      0.17      0.33      0.31      0.13      0.05
```

General Health vs. Sex Women seem to have a slightly worse general healht than men. The biggest difference is that there are around 2% more men who says to have a excellent general health than women. Also there are 2% more women than men who says to have a poor general health.

```
x=table(subset(my_data.fltr, select=c(genhlth,sex)))
x[,1]=round(x[,1]/sum(x[,1]),2)
x[,2]=round(x[,2]/sum(x[,2]),2); x;rm(x)
```

```
##            sex
## genhlth     Male Female
##   Excellent 0.18   0.17
##   Very good 0.33   0.33
##   Good      0.31   0.30
##   Fair      0.12   0.14
##   Poor      0.05   0.06
```

```
ggplot(my_data.fltr,aes(x=genhlth, fill=sex))+
    geom_bar()
```



The General health among the USA Population: Approximately 82 % of USA population is a good of better than good. There is a 5% of the populations that has a oop general health.

Differences in health between female and male are scarcely appreciables. Maybe the females have a slightly worse general health than males. (See Table genhlth/Sex)

Q2.1: Weight Distribution `weight.kg`

Weight-Percentage

```
round(summary(my_data.fltr$weight.kg),1)    # Weight is int
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    25.0    53.0    78.0    82.5   103.0   200.0
```

Weight vs. Sex: `genhlth VS sex`

```
stat.weight.sex<-my_data.fltr %>%
  group_by(sex) %>%
  summarise(grp.mean=round(mean(weight.kg),1), grp.sd=round(sd(weight.kg),1), gpr.med
ian=round(median(weight.kg),1), grp.iqr=round(IQR(weight.kg),1))
stat.weight.sex[1:2,]
```

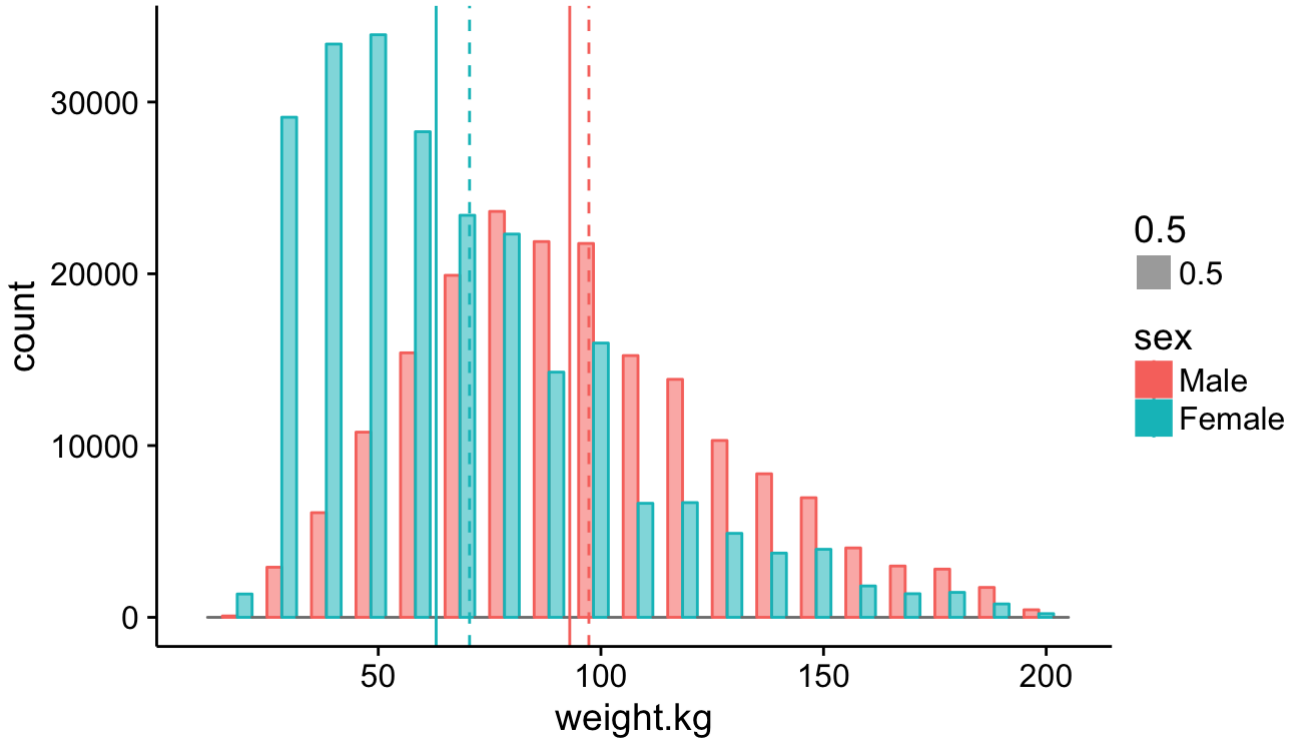
```
## Source: local data frame [2 x 5]
##
##      sex grp.mean grp.sd gpr.median grp.iqr
##   (fctr)    (dbl)  (dbl)      (dbl)   (dbl)
## 1   Male     97.3   35.0         93      45
## 2 Female     70.5   34.2         63      45
```

How is the weight distribution among the USA population in particular for females and males

The average weight for women and men in US is **71 Kg** and **97 Kg** respectively.The female and male weight distribution are right skeweed, althought the effect is hogher in the case of the females. The female weight median is 63 kg 8 kg under the mean. Whereas for the males the median is 93 kg 4 kg below the mean. The standard deviations in both cases are 35 kg.

```
p<-ggplot(my_data.fltr, aes(x=weight.kg,color=sex, fill=sex, alpha=0.5)) +
     geom_histogram(binwidth=10, position="dodge") +
     xlim(10,205)
p+geom_vline(data=stat.weight.sex[1:2,c(1,2,4)], aes(xintercept=grp.mean, color=sex),
               linetype="dashed") +    # Mean dashed
    geom_vline(data=stat.weight.sex[1:2,c(1,2,4)], aes(xintercept=gpr.median, color=s
ex),
               linetype="solid")  #Median Solid
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



Q2.2: Height `height.m`

Height

```
summary(my_data.fltr$height.m)    # Weight is int
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    76.2   162.6   170.2   170.1   177.8   236.2
```

Height vs. Sex: `height.m VS. sex`

```
stat.height.sex<-my_data.fltr %>%
  group_by(sex) %>%
  summarise(grp.mean=round(mean(height.m),1), grp.sd=round(sd(height.m),1), gpr.media
n=round(median(height.m),1), grp.iqr=round(IQR(height.m),1))
stat.height.sex[1:2,]
```

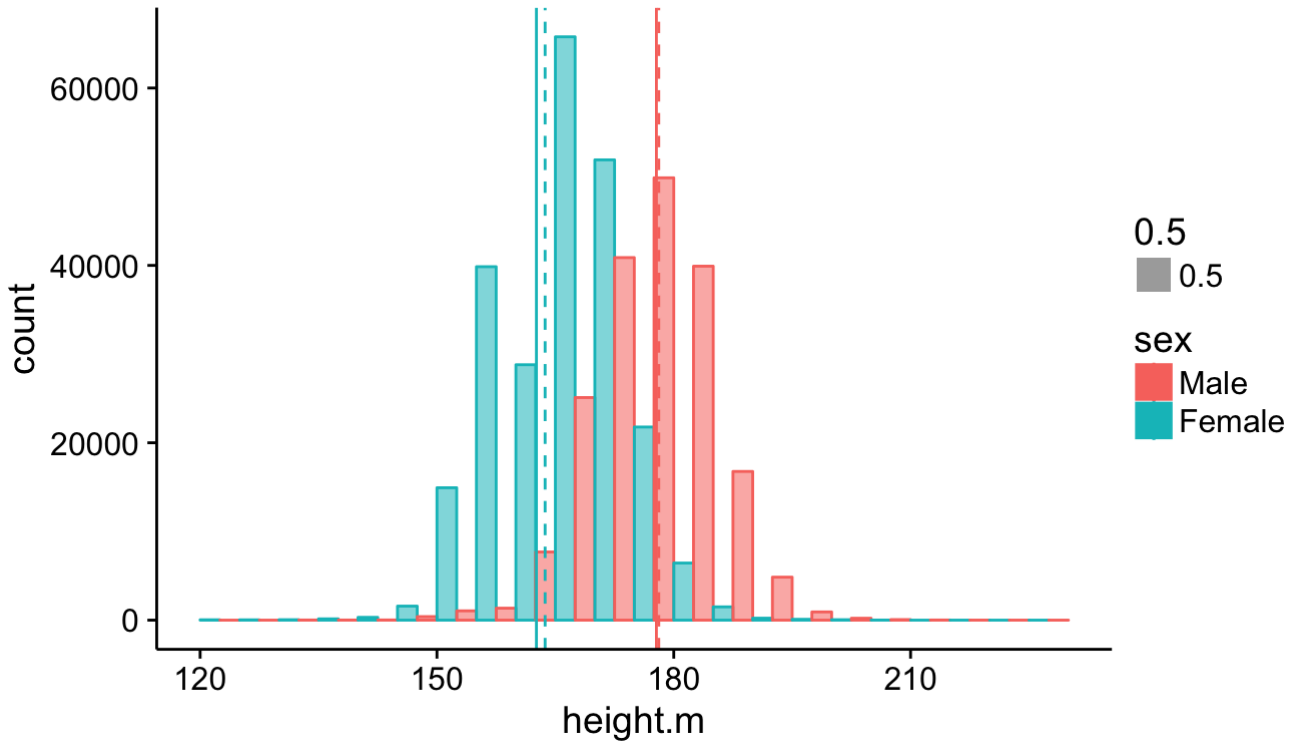
```
## Source: local data frame [2 x 5]
##
##      sex grp.mean grp.sd gpr.median grp.iqr
##   (fctr)    (dbl)  (dbl)      (dbl)   (dbl)
## 1   Male    178.1    7.6      177.8    10.2
## 2 Female    163.7    7.0      162.6     7.6
```

Q2.2: How is the Height distribution among the USA population in particular for females and males
The female and male height distribution are quite symmetrical as shown by the fact that mean and median lies approximately in the same place. The average height for women and men in US is **164 cm** and **178 cm** respectively.The female and male weight distribution are right skeweed, althought the effect is higher in the case of the females. The female height median 163 cm, 1 cm under the mean. Whereas for the males the median is 178 cm the same as the mean. The standard deviations in both cases are approximately 7 cm. The male distribution is a little bit wider indicated by a larger IQR.

```
p<-ggplot(my_data.fltr, aes(x=height.m,color=sex, fill=sex, alpha=0.5)) +
     geom_histogram(binwidth=5, position="dodge") +
     xlim(120,230)
p+geom_vline(data=stat.height.sex[1:2,c(1,2,4)], aes(xintercept=grp.mean, color=sex),
               linetype="dashed") +    # Mean Dashed
    geom_vline(data=stat.height.sex[1:2,c(1,2,4)], aes(xintercept=gpr.median, color=s
ex),
               linetype="solid")   #Median solid
```

```
## Warning: Removed 22 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Q2.3: BMI

BMI Histogram

```
round(summary(my_data.fltr$BMI),1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.5    19.4    26.3    28.1    34.8   180.8
```

```
stat.BMI.sex<-my_data.fltr %>%
  group_by(sex) %>%
  summarise(grp.mean=round(mean(BMI),1), grp.sd=round(sd(BMI),1), gpr.median=round(me
dian(BMI),1), grp.iqr=round(IQR(BMI),1))
stat.BMI.sex[1:2,]
```

```
## Source: local data frame [2 x 5]
##
##      sex grp.mean grp.sd gpr.median grp.iqr
##   (fctr)    (dbl)  (dbl)      (dbl)   (dbl)
## 1   Male     30.5   10.3       29.3    13.7
## 2 Female     26.2   12.5       23.5    15.9
```

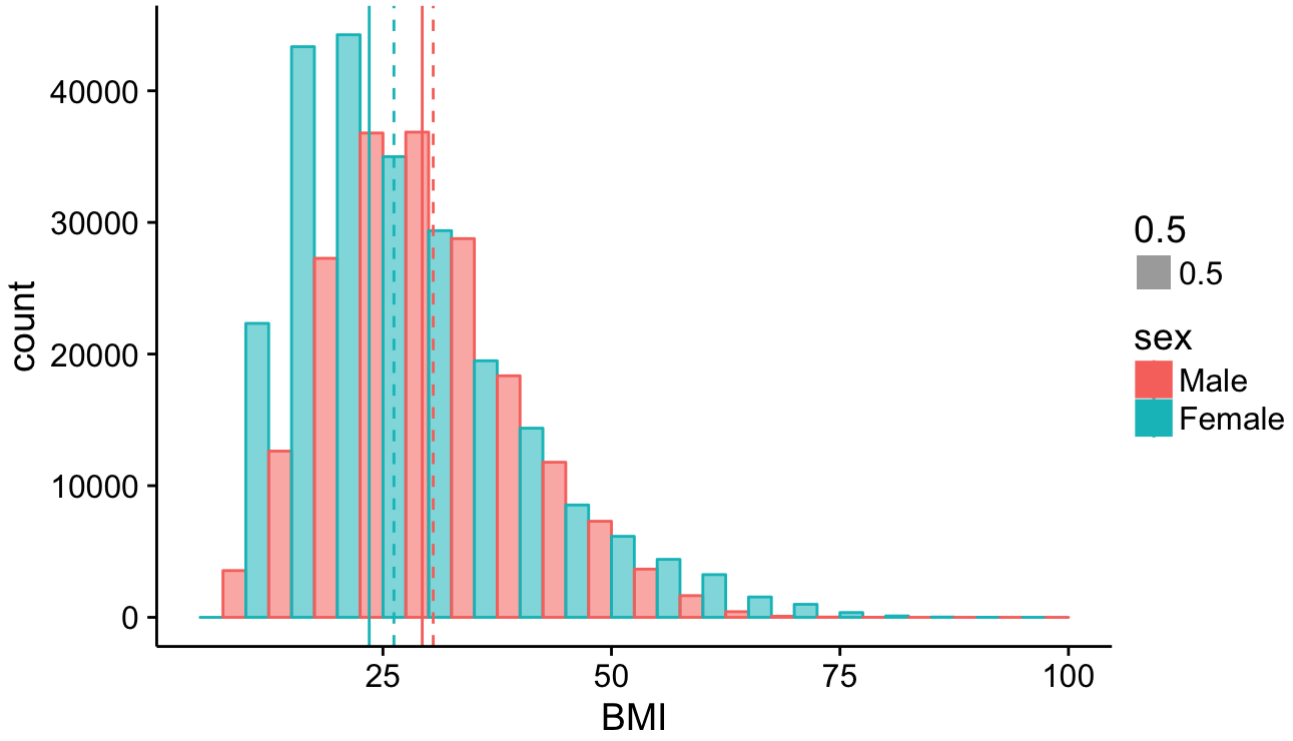
Q2.2: How is the BMI distribution in the USA population in particular for females and males

The female and male BMI distribution are right skeewed as shown by the fact that mean are larger than the median. Ths is more accusated in the case of the females. The average BMI for women and men in US is **26 units** and **31 units** respectively. This corresponds to the classification of **Overweith** and **Obese** for females and males. The female BMI median is 24, 2 units under the mean. Whereas for the males the median is 29 almost the same as the mean. The standard deviations in both cases are close to 10 units. **Females** mean and median are within the interval for **normal BM**I. However **male**'s mean and median are at the lower interval for **obesity**.
The female distribution is a little bit wider indicated by a larger IQR.

```
a1<-ggplot(my_data.fltr, aes(x=BMI,color=sex, fill=sex, alpha=0.5)) +
     geom_histogram(binwidth=5, position="dodge") +
     xlim(5,100)
a1+geom_vline(data=stat.BMI.sex[1:2,c(1,2,4)], aes(xintercept=grp.mean, color=sex),
              linetype="dashed") +    # print histogram
    geom_vline(data=stat.BMI.sex[1:2,c(1,2,4)], aes(xintercept=gpr.median, color=se
x),
              linetype="solid")
```

```
## Warning: Removed 7 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



BMI Classification vs. Sex

Underweight, Normal, Overweight, Obese, Morbidity definition:

- If $\text{BMI} < 18.5 \Rightarrow$ Underweigh
- If $28.5 \leq \text{BMI} < 25 \Rightarrow$ Normal
- If $25 \leq \text{BMI} < 30 \Rightarrow$ Overweight
- If $30 \leq \text{BMI} < 40 \Rightarrow$ Obese
- If $40 \leq \text{BMI} \Rightarrow$ Morbidity

```
x=table(my_data.fltr$BMI.Class)            #BMI Classification Table Percentage
round(x/sum(x),2)
```

```
##
## 1.Underweight      2.Normal  3.Overweight       4.Obese   5.Morbidity
##          0.22          0.23          0.17          0.23          0.16
```

```
x=table(subset(my_data.fltr, select=c(BMI.Class,sex))) #BMI Classification Table by s
ex
x[,1]=round(x[,1]/sum(x[,1]),2)
x[,2]=round(x[,2]/sum(x[,2]),2); x;rm(x)   #BMI Classification Sex Percentage
```

```
##                sex
## BMI.Class       Male Female
##   1.Underweight 0.11   0.32
##   2.Normal      0.22   0.23
##   3.Overweight  0.20   0.14
##   4.Obese       0.29   0.17
##   5.Morbidity   0.18   0.14
```

Q2.3: How is the USA population -female & male- Classified following the BMI Classification

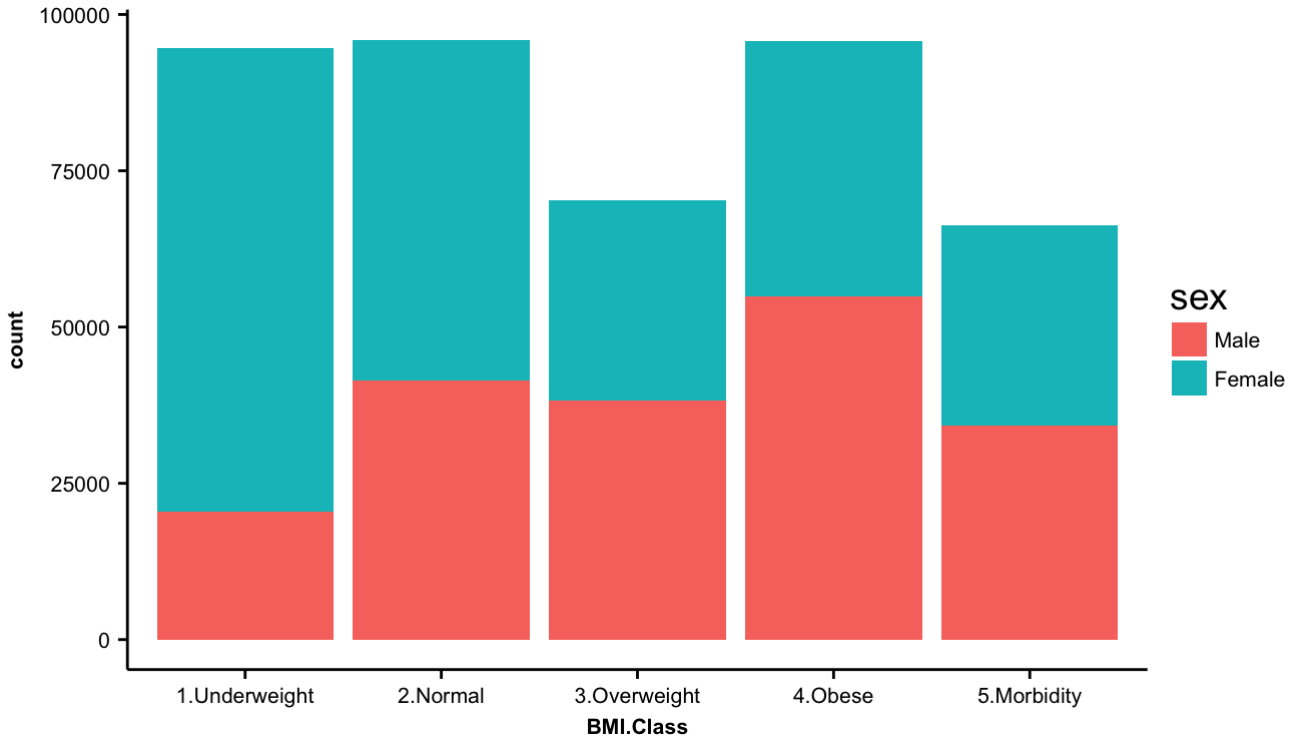
In general 55 % of the USA population has overweight, obesity or morbidity.

Female are “thiner” than males. It is clear to see that there are larger porcentage of female with normal and underweight status and lower percentages of women with overweight, obesity and Morbidity. The percentage of females women with overweight, obesity and Morbidity is 45% whereas for male is 67%.

Around 23% of Female and Male have a normal weight.

32% of Female are underweight (approximately 1 of 3 women) whereas onle 11% of males suffer underweight.

```
ggplot(subset(my_data.fltr,select=c(BMI.Class,sex)),aes(x=BMI.Class, fill=sex)) +
     geom_bar()+
theme(axis.text=element_text(size=8),
axis.title=element_text(size=8,face="bold", angle = 0))+
theme(legend.text = element_text(size = 8, colour = "black", angle = 0))+
theme(axis.text = element_text(colour = "black", angle=-0))
```



Q3: General Health vs. BMI/Education

Q3.1: General Health vs. BMI category & vs. Sex

Which rate of male/women with a particular BMI feels “Excellent”,“Very Good”, “Good”, “Fair” or “Poor”?

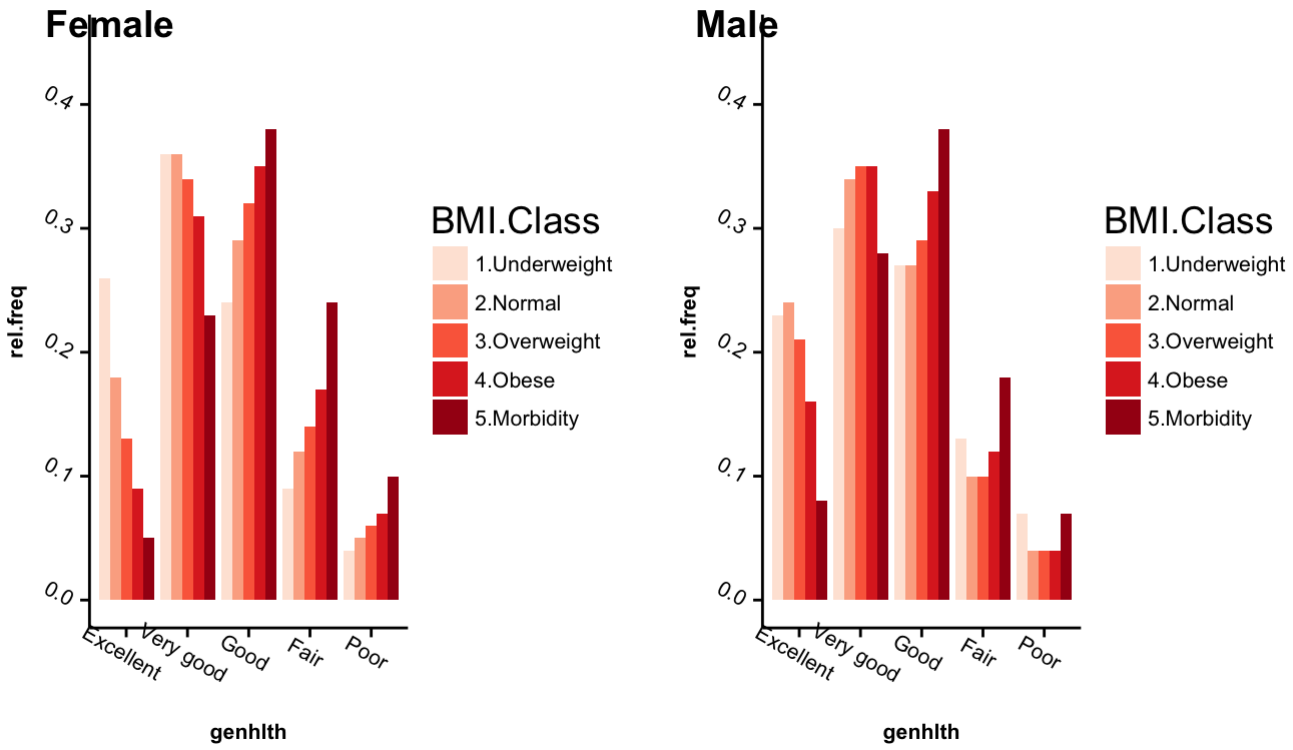
```
nr.sex<-my_data.fltr %>%                # nr. of each sex
     group_by(sex) %>%
     summarize(nr.sex=n())
nr.BMI.sex<-my_data.fltr %>%            # nr. of female/male vs. BMI category
     group_by(sex,BMI.Class) %>%
     summarise(nr.BMI.sex=n())
nr.BMI.sex<-nr.BMI.sex[complete.cases(nr.BMI.sex),]   # remove rows with NA
nr.BMI.sex.genhlth<-my_data.fltr %>%    #nr. of fem/mal with a particular BMI categor
y
      group_by(sex,BMI.Class,genhlth) %>%   # and particular level of genhlth
      summarise(nr.BMI=n())
nr.BMI.sex.genhlth<-nr.BMI.sex.genhlth[complete.cases(nr.BMI.sex.genhlth),]
x<-unlist(nr.BMI.sex$nr.BMI.sex)
vect.BMI<-data.frame(n=rep(x,each=5))
rm(x)
nr.BMI.sex.genhlth=data.frame(cbind(nr.BMI.sex.genhlth),vect.BMI)
nr.BMI.sex.genhlth<-nr.BMI.sex.genhlth %>%              # Calc the relative freq
                    mutate(rel.freq=round(nr.BMI/n,2))
# table((subset(my_data.fltr,select=c(BMI.Class,genhlth,sex))))
```

Plots

```
# Sex,BMI,Gen Health: Rel. Freq --> nr.sex.BMI.genhlth  --  Plot Female ---
q1<-ggplot(nr.BMI.sex.genhlth[26:50,],aes(genhlth,rel.freq, fill=BMI.Class))+
geom_bar(stat="identity", position="dodge") +
scale_fill_brewer(palette="Reds") +
theme(axis.text=element_text(size=8),
axis.title=element_text(size=8,face="bold", angle = 0))+
theme(legend.text = element_text(size = 8, colour = "black", angle = 0))+
theme(axis.text = element_text(colour = "Black", angle=-30)) +
ylim(0,0.45)
```

```
# Sex,BMI,Gen Health: Rel. Freq --> nr.sex.BMI.genhlth  --  Plot Male ---
q2<-ggplot(nr.BMI.sex.genhlth[1:25,],aes(genhlth,rel.freq, fill=BMI.Class))+
geom_bar(stat="identity", position="dodge") +
scale_fill_brewer(palette="Reds") +
theme(axis.text=element_text(size=8),
axis.title=element_text(size=8,face="bold", angle = 0))+
theme(legend.text = element_text(size = 8, colour = "black", angle = 0))+
theme(axis.text = element_text(colour = "Black", angle=-30)) +
  ylim(0,0.45)
```

```
plot_grid(q1, q2, labels=c("Female", "Male"), ncol = 2, nrow = 1)
```



We can see that there is not many differences between the results from females and males. The lager percentage of people with excellent and Very good health are Underweight nad Normal weight. Very few Morbidity people has excellent health. Reciprocally very few people with under and normal weighth have poor health. What is quite remarkable is that still 50% of people with morbidity says to have a Good Health. In general we can conclude that the tendency is that lighter people has better health.

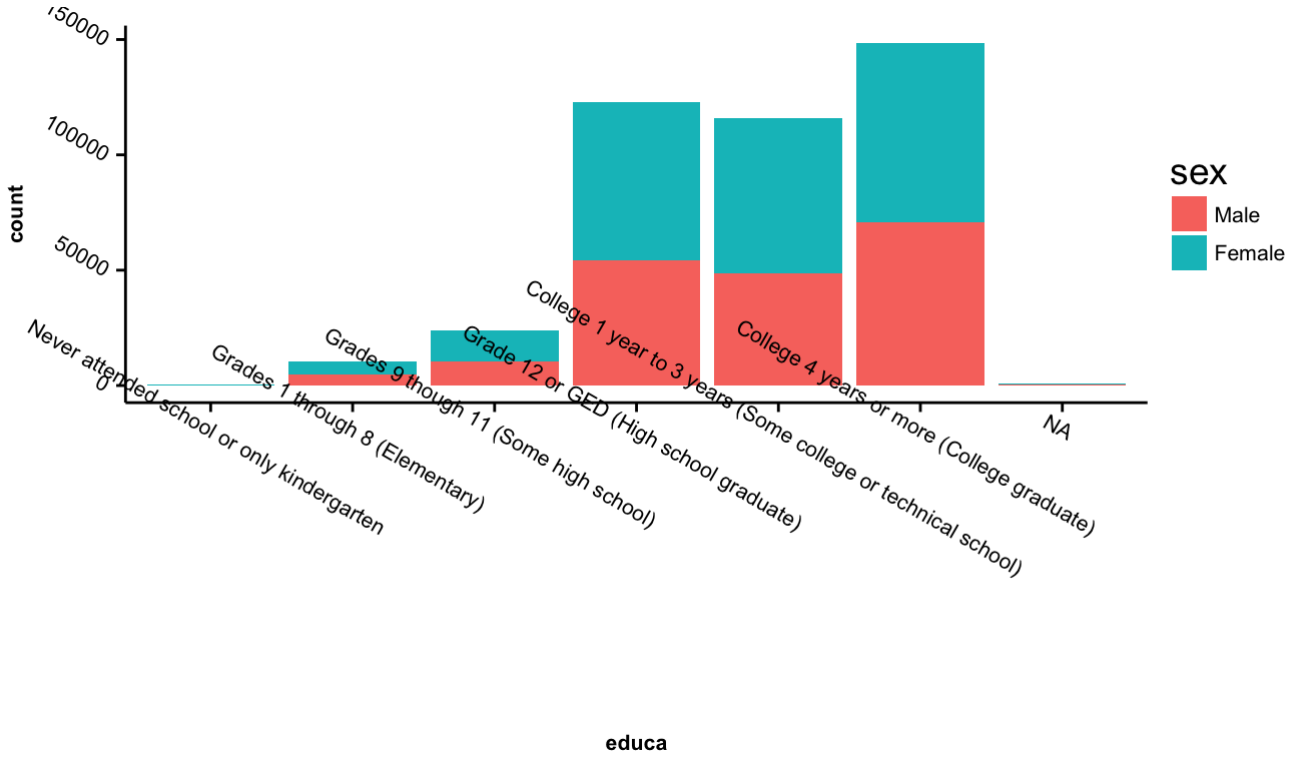
Q3.2: Educational level: `educa`

Education vs. Sex - Percentage

```
educa.rate<-table(subset(my_data.fltr,select=c(educa,sex)))
educa.rate[,1]=educa.rate[,1]/sum(educa.rate[,1])
educa.rate[,2]=educa.rate[,2]/sum(educa.rate[,2])
round(100*educa.rate,1)
```

```
##                                                          sex
## educa                                                     Male Female
##   Never attended school or only kindergarten               0.1    0.1
##   Grades 1 through 8 (Elementary)                          2.6    2.4
##   Grades 9 though 11 (Some high school)                    5.5    5.7
##   Grade 12 or GED (High school graduate)                  28.8   29.4
##   College 1 year to 3 years (Some college or technical school) 25.7   29.0
##   College 4 years or more (College graduate)              37.4   33.4
```

```
ggplot(subset(my_data.fltr,select=c(educa,sex)),aes(x=educa, fill=sex)) +
     geom_bar()+
theme(axis.text=element_text(size=8),
axis.title=element_text(size=8,face="bold", angle = 0))+
theme(legend.text = element_text(size = 8, colour = "black", angle = 0))+
theme(axis.text = element_text(colour = "black", angle=-30))
```



General Health vs. Edu & vs.Sex

Which rate of male/women with a particular level of education feels “Excellent”,“Very Good”, “Good”, “Fair” or “Poor”?

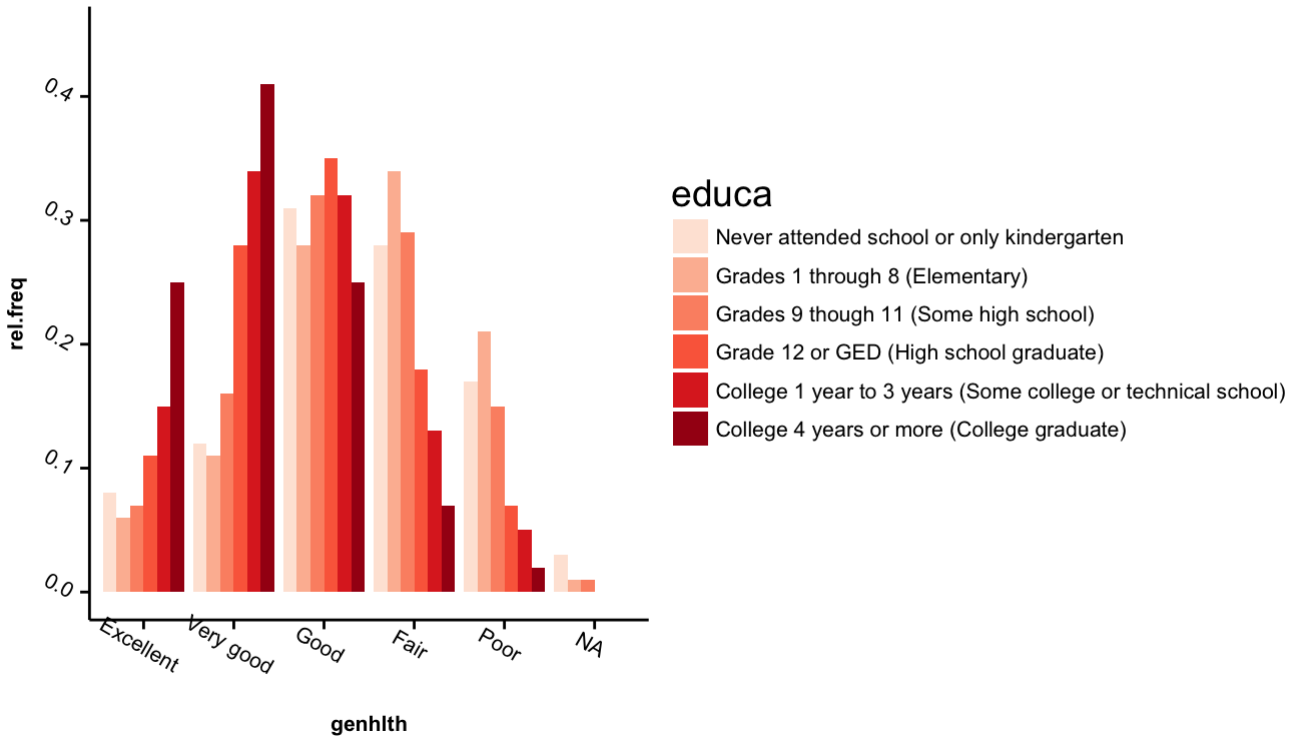
```
nr.sex<-my_data.fltr %>%                  # nr. of each sex
     group_by(sex) %>%
     summarize(nr.sex=n())
nr.sex.edu<-my_data.fltr %>%              # nr. of female/male vs. educational level
     group_by(sex,educa) %>%
     summarise(nr.sex.edu=n())
nr.sex.edu<-nr.sex.edu[1:14,]
nr.sex.edu.genhlth<-my_data.fltr %>%      #nr. of fem/mal with a particular edu level
      group_by(sex,educa,genhlth) %>%     # and particular level of genhlth
      summarise(nr.6edu.Hlth=n())
nr.sex.edu.genhlth<-nr.sex.edu.genhlth[1:84,]
x<-unlist(nr.sex.edu$nr.sex.edu)
vect.edu<-data.frame(n=rep(x,each=6))
rm(x)
nr.sex.edu.genhlth=data.frame(cbind(nr.sex.edu.genhlth),vect.edu)
nr.sex.edu.genhlth<-nr.sex.edu.genhlth %>%                # Calc the relative freq
                    mutate(rel.freq=round(nr.6edu.Hlth/n,2))
# table((subset(my_data.fltr,select=c(educa,genhlth,sex))))
```

```
# Sex, Edu,Gen Health: Rel. Freq --> nr.sex.edu.genhlth  --  Plot Female ---
p1<-ggplot(nr.sex.edu.genhlth[43:78,],aes(genhlth,rel.freq, fill=educa))+
geom_bar(stat="identity", position="dodge") +
scale_fill_brewer(palette="Reds") +
theme(axis.text=element_text(size=8),
axis.title=element_text(size=8,face="bold", angle = 0))+
theme(legend.text = element_text(size = 8, colour = "black", angle = 0))+
theme(axis.text = element_text(colour = "Black", angle=-30)) +
ylim(0,0.45)
```

```
# Sex, Edu,Gen Health: Rel. Freq --> nr.sex.edu.genhlth  --  Plot Male ---
p2<-ggplot(nr.sex.edu.genhlth[1:36,],aes(genhlth,rel.freq, fill=educa))+
geom_bar(stat="identity", position="dodge") +
scale_fill_brewer(palette="Reds") +
theme(axis.text=element_text(size=8),
axis.title=element_text(size=8,face="bold", angle = 0))+
theme(legend.text = element_text(size = 8, colour = "black", angle = 0))+
theme(axis.text = element_text(colour = "Black", angle=-30)) +
  ylim(0,0.45)
#plot_grid(p1, p2, labels=c("Female", "Male"), ncol = 2, nrow = 1)
```

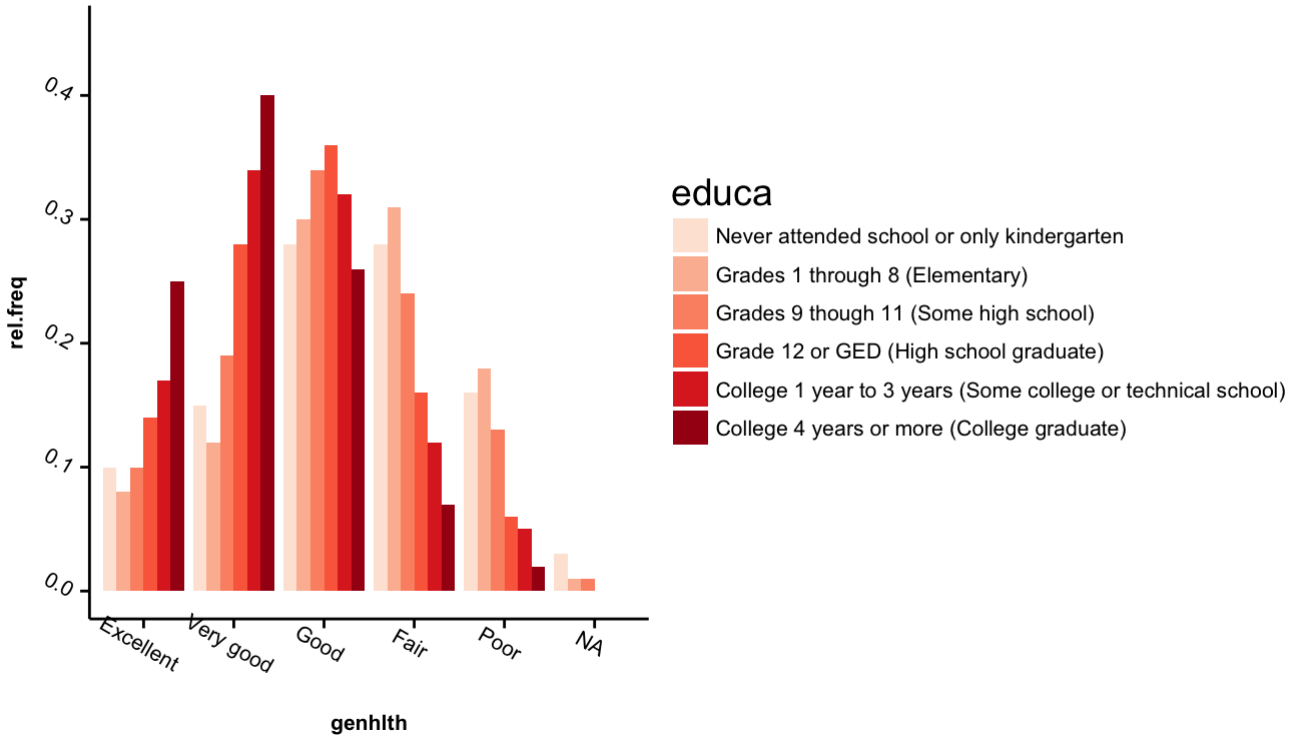
Female General Health-Education

```
p1 #### Female Health-Education
```



Male General Health-Education

```
p2 #### Male Health-Education
```



We can see that there is not many differences between the results from females and males. The lager percentage of people with excellent and Very good health are the people with largest education. Very few that just attended Some high school or less have excellent health. Reciprocally extremly few people with some college education or graduated have poor health. What is quite remarkable the correlation that appears here between health and educational level.

Q3: Conclusion

In general we can conclude that the tendency is that highly edeucated people and normal-thinner people has better health. However education seems to be a higher inpact than the BMI does.