Intro to Data & Probability - Exploring the BRFSS data

The purpose of this project is to use an exploratory data analysis in an attempt to answer a few research questions from a large size statistical study.

Setup

In this project we shall user a number o R packages `ggplot2` and `gridExtra` for the visualisation of results and `dplyr` to perform transformative operations on datasets

Load packages

```r
library(ggplot2)
library(gridExtra)
library(dplyr)
```

Load data

The BRFSS dataset is loaded from a file `brfss2013.rdata` which can be read by R directly with the load() function creating a `brfss2013` data.frame object.

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

Part 1: Data

The dataset used in this project comes from the ("The Behavioral Risk Factor Surveillance System," n.d.) (BRFSS) 2013 edition. Every year the US Center for Disease Control and Prevention (CDS) asks to conduct a detailed questionnaire of health related questions within a large sample of the US population residing in all US states and territories. The BRFSS conducts land-line telephone and cellular telephone-based surveys. The data is collected uniformly across all states and all days of the edition's year through a random sampling of US adults. Groups of studied factors include general health status, number of unhealthy days in a past month, health care access, practice of a physical activity, alcohol consumption, fruits and vegetables consumption, hypertension awareness, cholesterol awareness and chronic health conditions. The BRFSS 2013 edition contains a dataset of over 491 thousands of individuals spanning over several hundreds of variables. This is an example of a very large continuous observational statistical study with a generalizability of the measured data.

Part 2: Research questions

Research question 1: Having a health insurance increases the quality of life. We are interested to study how people make decisions whether to get a health insurance coverage if they have limited means to afford it. Does the level of education plays a role when deciding about taking or forgoing the health insurance plan? More precisely we shall try to answer a question if college educated people with low income but in a good health are more likely to make an economic decision to forgo a health insurance when compared to people with a precollege education from the same income category and the same physical condition.

Research question 2: In this section we shall study a variable which represents a number of days int last 30 days people felt to be in a poor physical health. The dataset has a variable on general health. But arguably the latter is more qualitative and the former is more quantitative. Studying numbers of sick days that people take has a direct economic interest reflecting a net wealth expense rather than generation by the employed population. We are interested to study factors which can potentially reduce the number of sick days taken. Among various factors we concentrate on a physical activity levels as well as on a consumption of healthy food. Thus does a physical activity, a consumption of a healthy food negatively correlate with the number of sick days taken? if yes, which factor is more negatively correlated with the number of sick days?

Research question 3: In this section we are interested in checking the normality of a distribution of human height. Thus are the studied population heights normally distributed? Our interest is driven to verify similar claims (see Cetinkaya-Rundel, Barr, and Diez 2015, 137–39). Taller people have higher risks of health (Kabat et al. 2013), (Ho 2011, 400) and if a human height is normally distributed the frequency of very tall humans is rare and under the 'thin tails' of the normal distribution.

Part 3: Exploratory data analysis

Research question 1:

We start with extracting variables from a general dataset relevant for the subject 1. `genhlth,X` is a general health status, `X_educag` contains levels of education, `X_incomg` represents various income categories and `hlthpln1` indicates if a subject has a health insurance plan.

```
sel_brfss2013 <- brfss2013 %>% select (genhlth,X_rfhlth,X_educag,X_incomg,hlthpln1)
str(sel_brfss2013)
```

```
## 'data.frame':    491775 obs. of  5 variables:
##  $ genhlth : Factor w/ 5 levels "Excellent","Very good",..: 4 3 3 2 3 2 4 3 1 3 ...
##  $ X_rfhlth: Factor w/ 2 levels "Good or Better Health",..: 2 1 1 1 1 1 2 1 1 1 ...
##  $ X_educag: Factor w/ 4 levels "Did not graduate high school",..: 4 3 4 2 4 4 2 3 4 2 ...
##  $ X_incomg: Factor w/ 5 levels "Less than $15,000",..: 5 5 5 5 4 5 NA 4 5 2 ...
##  $ hlthpln1: Factor w/ 2 levels "Yes","No": 1 1 1 1 1 1 1 1 1 1 ...
```

To distinguish individuals with college vs pre-college education a new variable `edu_cat` is created.

```
#2. create a variable on education level
sel_brfss2013 <- sel_brfss2013  %>%
mutate(edu_cat = factor( ifelse( is.na(X_educag),NA
                                ,ifelse(X_educag %in% c("Attended college or technical school"
                                                       ,"Graduated from college or technical school")
                                                        ,"college education","pre-college education")))
```

To have a first idea about the research problem it is useful to plot relationships between studied variables.

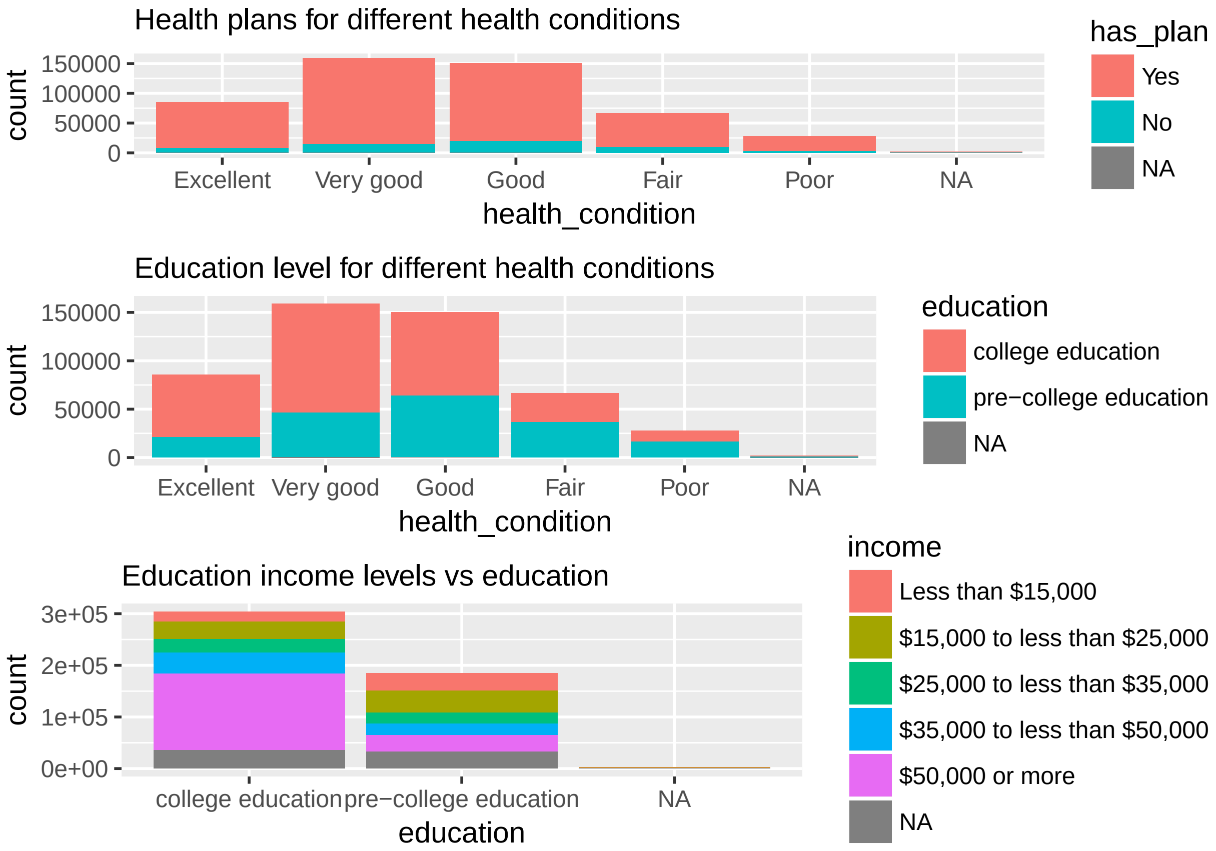
```
sel_brfss2013$health_condition<-sel_brfss2013$genhlth
sel_brfss2013$has_plan<-sel_brfss2013$hlthpln1
sel_brfss2013$education <-sel_brfss2013$edu_cat
sel_brfss2013$income <-sel_brfss2013$X_incomg

p1 <- ggplot(sel_brfss2013, aes(x=health_condition, fill =has_plan )) +
  geom_bar() +
  ggtitle("Health plans for different health conditions")  +
  theme(plot.title = element_text(size = 11))

p2 <- ggplot(sel_brfss2013, aes(x=health_condition, fill =education )) +
  geom_bar() +
  ggtitle("Education level for different health conditions")  +
  theme(plot.title = element_text(size = 11))

p3 <- ggplot(sel_brfss2013, aes(x=education, fill = income )) +
  geom_bar() +
  ggtitle("Education income levels vs education") +
  theme(plot.title = element_text(size = 11))

grid.arrange(p1,p2,p3,nrow=3,heights=c(0.80, 0.99, 0.99) )
```



We observe that only a minority of people do not have a health coverage plan and that income level seems to be positively correlated with a college education. It can also be observed 'Fair' and 'Poor' health condition is seemingly equally distributed among people with college and pre-college education

For studied factors we create new variables with less categories, 2 for health and 3 for income. We shall concentrate on individuals in the "low" income category with people making up to 25000 USD.

```
# create a health condition variable
sel_brfss2013 <- sel_brfss2013  %>%
mutate(g_health =factor(ifelse(is.na(genhlth), NA
                       ,ifelse(genhlth %in% c("Excellent", "Very good", "Good"),
                        "excellent or good", "poor or fair"))) )

# create a variable on income category  low middle upper
sel_brfss2013 <- sel_brfss2013  %>%
mutate(incom_cat = factor(ifelse(is.na(X_incomg),NA
                        ,ifelse(X_incomg %in%
                        c("Less than $15,000","$15,000 to less than $25,000"),"low"
                        ,ifelse(X_incomg %in% c("$25,000 to less than $35,000","$35,000 to less than $5
                        ,"middle","upper")))) )
```

Now lets see the summary statistics for subjects of interest to our research question. Those are the individuals from a 'low' income category who do not have a health insurance plan and who are in good health. Because they are in good health theoretically they can afford to avoid taking a health coverage plan.

```
## show summary statistics
sel_brfss2013  %>%
filter( !is.na(edu_cat)
        & g_health =="excellent or good"
        & incom_cat %in%c("low")
        & has_plan == "No" ) %>%
group_by(has_plan,g_health,edu_cat,incom_cat) %>%
summarize(count=n()) %>%
arrange(edu_cat)
```

```
## Source: local data frame [2 x 5]
## Groups: has_plan, g_health, edu_cat [2]
##
##   has_plan          g_health               edu_cat incom_cat count
##     <fctr>            <fctr>                <fctr>    <fctr> <int>
## 1       No excellent or good     college education       low  8599
## 2       No excellent or good pre-college education       low 11136
```

The same summary statistics using xtabs function:

```
# using xtabs
xt<-xtabs(~edu_cat+incom_cat
      ,data=sel_brfss2013 %>% filter(!is.na(edu_cat)
                                   & g_health =="excellent or good"
                                   & !is.na(incom_cat)
                                   & has_plan == "No") )
print(xt)
```

```
##                         incom_cat
## edu_cat                    low middle upper
##   college education       8599   5633  3710
##   pre-college education  11136   5040  1879
```

A Relative frequency of such individuals with 'college education'

```
xt[1,1] / sum(xt[,1])  # 8599 / (8599+11136) = 0.43572
```

```
## [1] 0.4357233
```

A Relative frequency of individuals with 'pre-college education'

```
xt[2,1] / sum(xt[,1])  # 11136 / (8599+11136) = 0.5642
```

```
## [1] 0.5642767
```

Thus we conclude that healthy people with low income and college degree are still more likely to take the health coverage plan compared to those with a pre-college degree.

Research question 2:

As with the previous question lets extract variables from a general dataset relevant to our current research question. The variable `X_frutsum` for a total of consumed fruits in a past month, `X_vegesum` a total of consumed vegetables in a past month, `X_pacat` a variable for a physical activity with 4 factor levels and `physhlth` is a number of days with bad physical health in a past 30 days.

```
# selection of variables for question 2
sel1_brfss2013 <- brfss2013 %>%
select (X_frutsum  # sum of consumed fruits in 30 days
       ,X_vegesum  # sum of consumed vegatables in 30 days
       ,X_pacat1  # sport activity
       ,physhlth) # number of days sick in past 30 days
print ( str(sel1_brfss2013) )
```

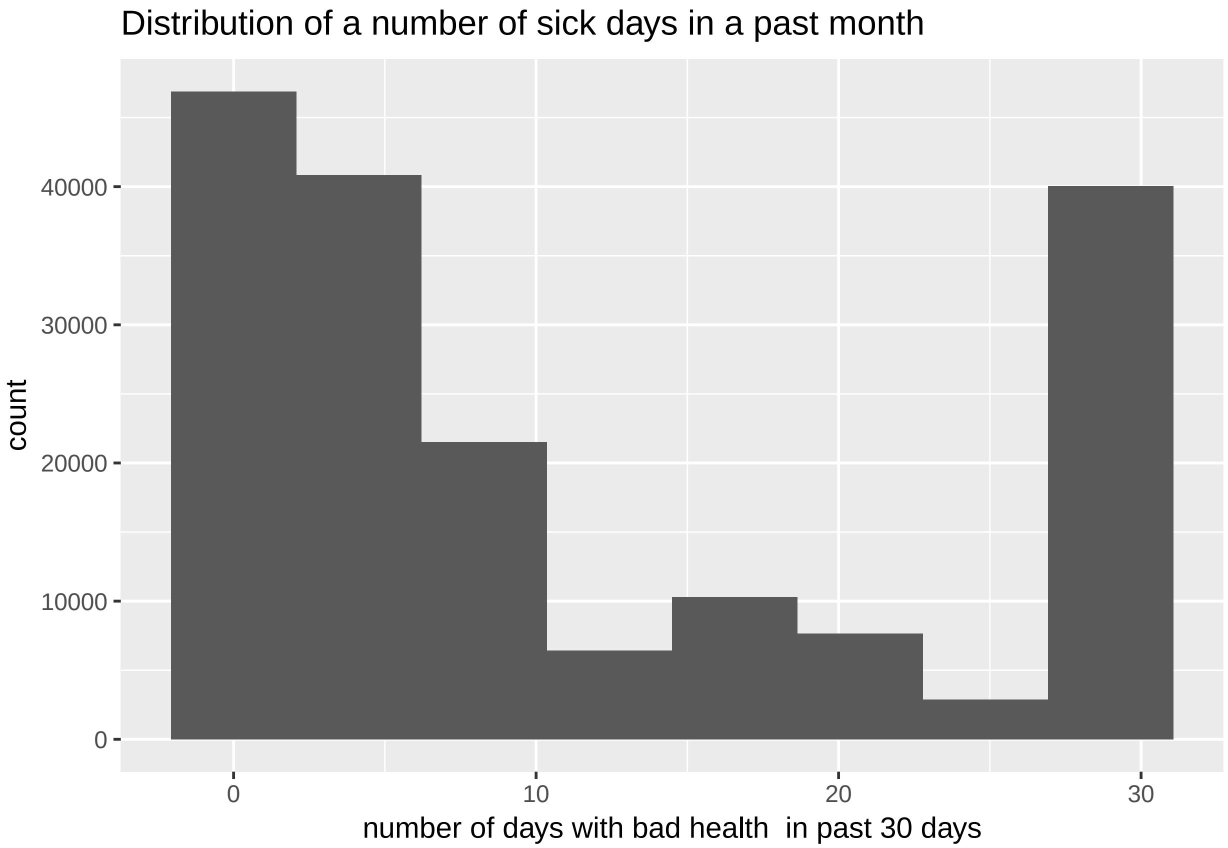
```
## 'data.frame':    491775 obs. of  4 variables:
##  $ X_frutsum: num  413 20 46 49 7 157 150 67 100 58 ...
##  $ X_vegesum: num  53 148 191 136 243 143 216 360 172 114 ...
##  $ X_pacat1 : Factor w/ 4 levels "Highly active",..: 4 3 4 3 4 3 2 3 2 2 ...
##  $ physhlth : int  30 0 3 2 10 0 1 5 0 0 ...
## NULL
```

For an initial view of the data we plot a histogram of a number of sick days for individuals who had reported at least one sick day within a past month.

```
print (summary(sel1_brfss2013$physhlth) )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   0.000   4.353   3.000  60.000   10957
```

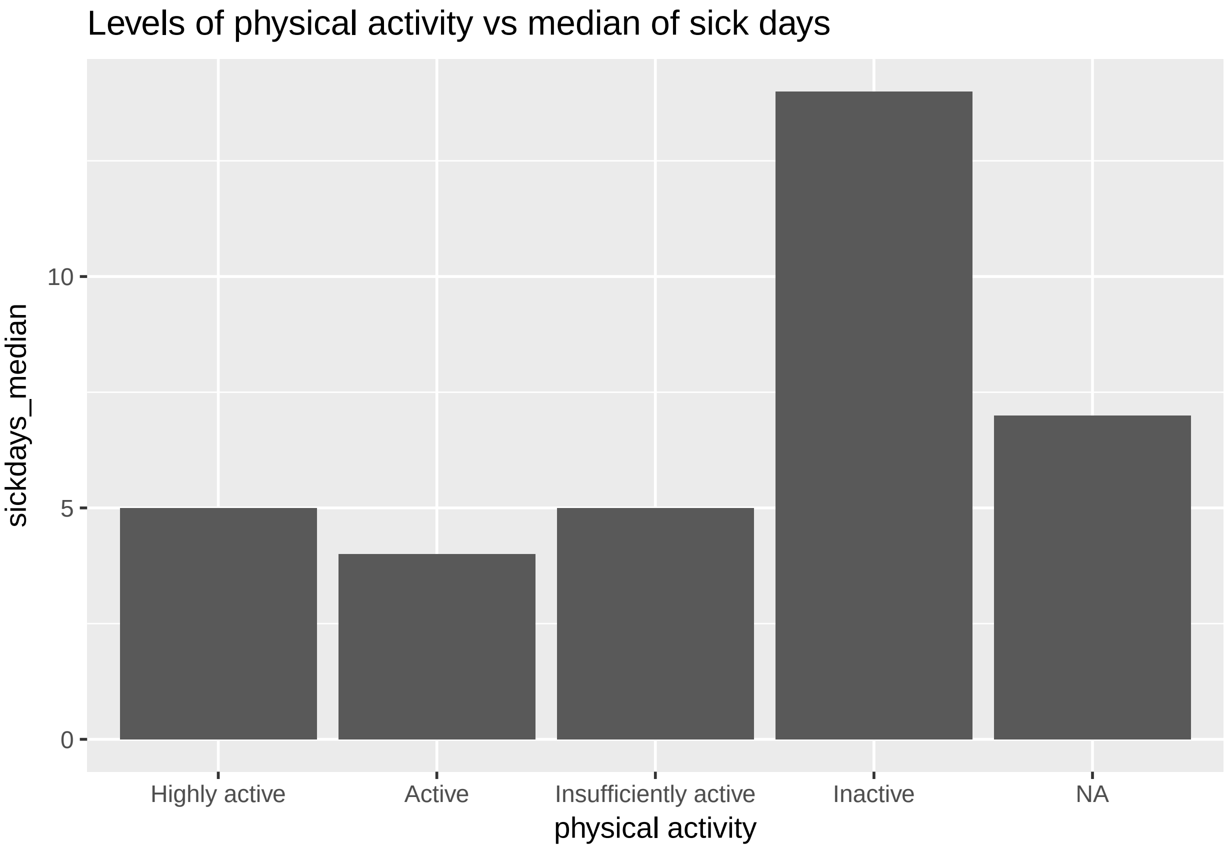
```
p1 <- ggplot(data = sel1_brfss2013 %>% filter(physhlth > 0.99 & physhlth <= 30) ,  aes(x = physhlth)) +
  geom_histogram(bins=8) +
  labs(x="number of days with bad health  in past 30 days") +
  ggtitle("Distribution of a number of sick days in a past month")
print(p1)
```



A filter of less or equal of thirty sick days in past 30 days is used to cut off cases with bad data. We observe a bimodal distribution with peaks for 1, 2 sick days and 30 days. It is interesting to observe a high peak for 30 sick days.

We can also plot a median of sick days for different levels of physical activity.

```
sel1_brfss2013 %>% filter(physhlth>0.99 & physhlth < 32) %>%
group_by(X_pacat1) %>% summarize(sickdays_median=median(physhlth)) %>%
ggplot(aes(x=X_pacat1, y=sickdays_median)) +
    geom_bar(stat="identity") +
    ggtitle("Levels of physical activity vs median of sick days") +
    labs(x="physical activity")
```



The chart suggests a dependency between the number of sick days and especially an ‘Inactive’ level of physical activity. It is interesting to observe a relatively high value of the median of sick days for people who did not report their level of physical activity (NA category).

To study a possible relation between a type of food consumed and a number of sick days we wish to compute a variable reflecting a consumption of healthy food based on two underlying variables: consumption of fruits `X_frutsum` and consumption of vegetables `X_vegesum`.

```
sel1_brfss2013 <- sel1_brfss2013  %>%

#divide by 100 as the original variable implies two decimal pionts
mutate(fruit_consum = X_frutsum/100)

sel1_brfss2013 <- sel1_brfss2013  %>%
mutate(vegetable_consum = X_vegesum/100)

summary(sel1_brfss2013$fruit_consum)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.570   1.030   1.395   2.000 198.000   28702
```

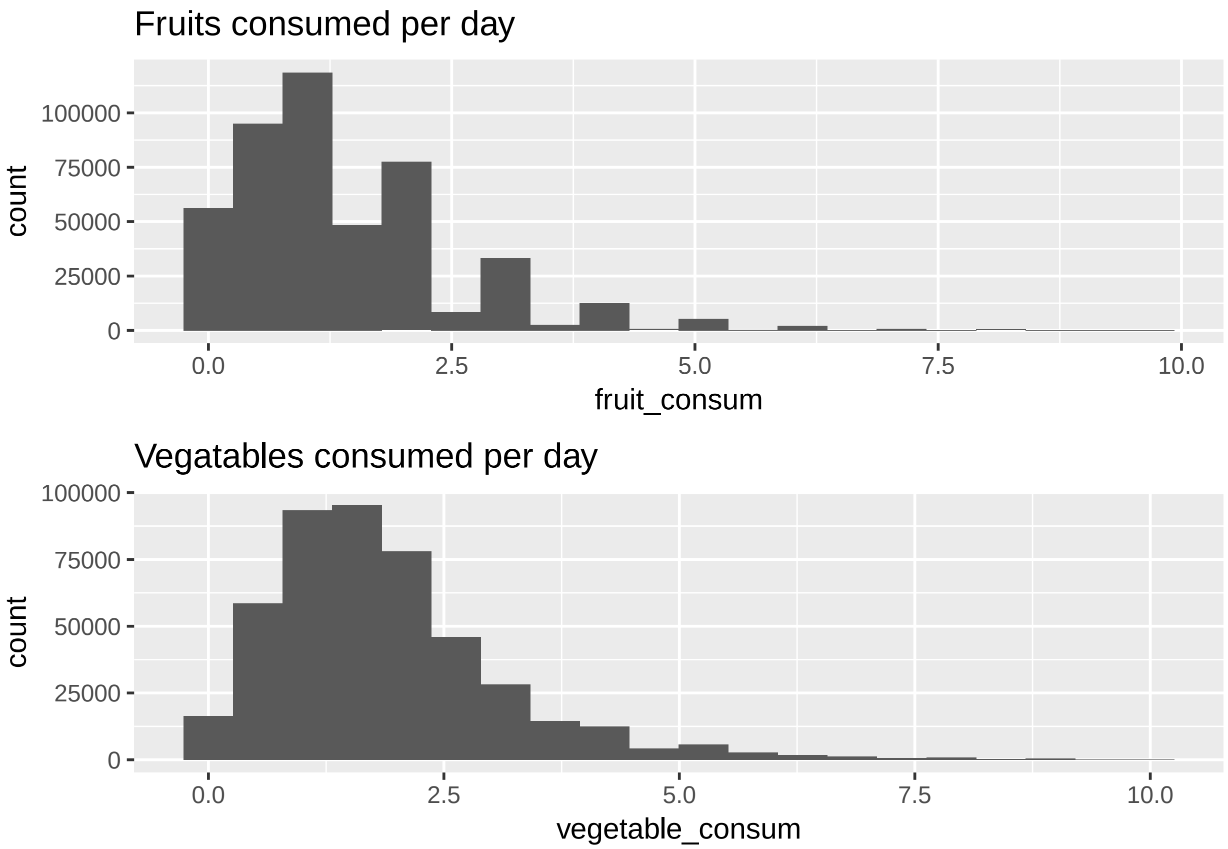
```
summary(sel1_brfss2013$vegetable_consum)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    1.04    1.65    1.90    2.43  198.30   30165
```

```
#motivating charts
p1<-ggplot(data=sel1_brfss2013 %>% filter(fruit_consum < 10.0)
    ,aes(x=fruit_consum)) +
      geom_histogram(bins=20) +
      ggtitle("Fruits consumed per day")

p2<-ggplot(data=sel1_brfss2013 %>% filter(vegetable_consum < 10.0)
    ,aes(x=vegetable_consum)) +
     geom_histogram(bins=20) +
      ggtitle("Vegatables consumed per day")

grid.arrange(p1,p2,nrow=2)
```



The chart allow us to determine a meaning cutoff of a variable (10 fruits or vegetables per day) to exclude bad data. For both food factors we observe an unimodal right skewed distribution.

We create a `healthyfood` variable with 2 categories. A 'very good' value reflects a simultaneous consumption of vegetables and fruits above their medians.

```
#create healthy food variable
fruit_med <- median( (sel1_brfss2013 %>% filter(fruit_consum < 10.0))$fruit_consum )
vegetable_med <- median( (sel1_brfss2013 %>% filter(vegetable_consum < 10.0))$vegetable_consum )

sel1_brfss2013 <- sel1_brfss2013  %>%
mutate(healthyfood =  factor(ifelse( is.na(fruit_consum) | is.na(vegetable_consum), NA
                       ,ifelse( fruit_consum > fruit_med & vegetable_consum > vegetable_med
                        ,"very good", "poor and average")))  )

summary(sel1_brfss2013$healthyfood)
```

```
## poor and average        very good             NA's
##          316571          143812           31392
```

Next we create binary variables for having at least 1 sick day and having an active and above physical activity.

```
#  create a health condition variable  is_sick  (had any sick days in past 30 days)
sel1_brfss2013 <- sel1_brfss2013  %>%
mutate(is_sick =factor(ifelse(is.na(physhlth), NA
                       ,ifelse(physhlth >0.99 & physhlth <32, "yes", "no"))))

#create physical condition active variable
sel1_brfss2013 <- sel1_brfss2013  %>%
mutate(is_sport_active =factor(ifelse(is.na(X_pacat1), NA
                       ,ifelse(X_pacat1 %in%c("Highly active", "Active"), "yes", "no"))))

summary(sel1_brfss2013$is_sick)
```

```
##     no    yes   NA's
## 304266 176552  10957
```

```
summary(sel1_brfss2013$is_sport_active)
```

```
##     no    yes   NA's
## 208619 206278  76878
```

A contingency table for `is_sick` versus `is_sport_active` variables will help us to compute conditional probabilities such as a probability to have at least one sick day given an active physical activity.

```
ms_sick_vs_sport<-xtabs(~is_sick+is_sport_active, data=sel1_brfss2013)
print(ms_sick_vs_sport)
```

```
##        is_sport_active
## is_sick     no    yes
##     no  115447 140787
##     yes  87871  62590
```

A joint probability to have at least 1 sick day and to have an active physical activitiy is:

```
p_has_sickdays_and_sport <- ms_sick_vs_sport[2,2]/sum(ms_sick_vs_sport)
```

We also need to know general probabilities to have at least 1 sick days in the last 30 days and a probability to have an active physical activity.

```
#probability to have at least 1 sick day
p_has_sickdays <- nrow(sel1_brfss2013 %>% filter(is_sick=="yes")) /
                  nrow(sel1_brfss2013 %>% filter(!is.na(is_sick)))

#probability to be have a high level physical activity
p_is_sport_active <- nrow(sel1_brfss2013 %>% filter(is_sport_active=="yes")) /
                     nrow(sel1_brfss2013 %>% filter(!is.na(is_sport_active)))
```

Using a conditional probability formula $P(A|B) = \frac{P(A \cup B)}{P(B)}$ the conditional probability have have at least 1 sick day given a good physical activity level:

```
p_has_sickdays_given_sport_active <- p_has_sickdays_and_sport / p_is_sport_active
```

We perform similar calculations for the variable of healthy food. A contingency table for `is_sick` and `healthfood` variables is

```
ms_sick_vs_healthfood<-xtabs(~is_sick+healthyfood, data=sel1_brfss2013)
print(ms_sick_vs_healthfood)
```

```
##        healthyfood
## is_sick poor and average very good
##     no            192186     92286
##     yes           117191     49036
```

A joint probability to have sick days and consume healthy food and a conditional probability to have sick days given a consumption of healthy food :

```
#joint probability
p_has_sickdays_and_healthyfood <- ms_sick_vs_healthfood[2,2] / sum(ms_sick_vs_healthfood)

#general probability to eat healthy food
p_has_healthyfood <- nrow(sel1_brfss2013 %>% filter(healthyfood %in%c("very good"))) /
                     nrow(sel1_brfss2013 %>% filter(!is.na(healthyfood)))
#conditional probability
p_has_sickdays_given_healthyfood <- p_has_sickdays_and_healthyfood  / p_has_healthyfood
```

Printing the calculated probabilities yields:

```
fmt<-"%2.4f"
#p to have at least 1 sick days
sprintf(fmt,p_has_sickdays)
```

```
## [1] "0.3672"
```

```
#p to have  sick days given active physical activity
sprintf(fmt, p_has_sickdays_given_sport_active)
```

```
## [1] "0.3095"
```

```
#p to have sick days given consuming healthy food
sprintf(fmt, p_has_sickdays_given_healthyfood)
```

```
## [1] "0.3483"
```

Thus we first observe that both factors of healthy food and of active physical activity seems to be negatively correlated with the number of sick days. It is also clear that the factor of active physical activity has a stronger correlation with the number sick days.

Research question 3:

This question concentrates on a single variable `htm4` representing a height of humans in [cm]. To check the 'normalness' of the distribution of this variable we will check its unimodality, symmetry, percentiles for multiples of standard deviations and its normal probability plot. First the summary of this variable provides hints to eliminate 'NA' and extreme values.

```
summary(brfss2013$htm4)
```

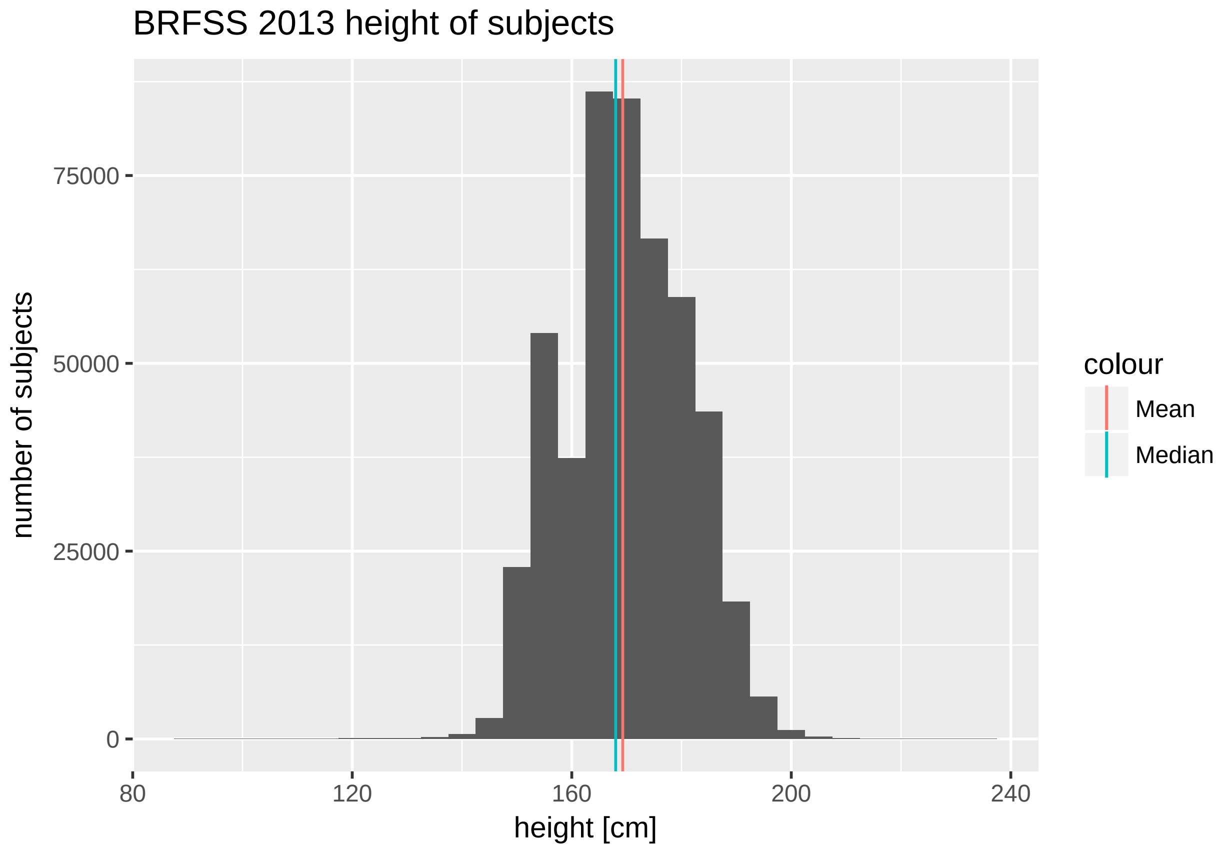
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##     1.0   163.0   168.0   169.3   178.0  2469.0    7643
```

The actual lower and upper cutoff values come from Wikipedia ("List of Shortest People," n.d.) and ("List of Tallest People," n.d.).

```
cbrfss2013<-brfss2013 %>% filter(!is.na(htm4) & htm4 > 50 & htm4 < 280)
```

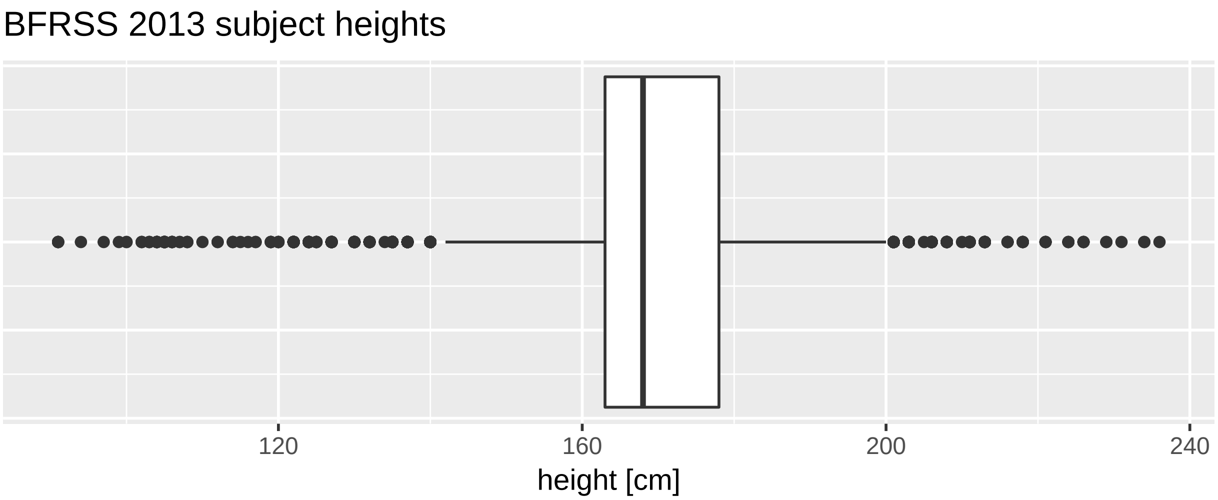
For an initial check of the distribution of heights let us plot its histogram adding the distribution's mean and median on top of it.

```
print(
ggplot(data = cbrfss2013,  aes(x = htm4)) +
  geom_histogram(bins=30)
  + ggtitle("BRFSS 2013 height of subjects")
  + labs(x="height [cm]",y="number of subjects")
  + geom_vline(aes(xintercept=mean(cbrfss2013$htm4),color="Mean"))
  + geom_vline(aes(xintercept=median(cbrfss2013$htm4),color="Median"))
)
```



We observe that the distribution is in fact unimodal and seemingly symmetric with the mean and median lines relatively very close to each other. We can use a boxplot chat for a further check of symmetry.

```
print(
ggplot(data=cbrfss2013, aes(x=0, y=htm4))
  + geom_boxplot() + coord_flip()
  + ggtitle("BFRSS 2013 subject heights")
 + labs(y="height [cm]", x="" )
 + theme(axis.title.y=element_blank(), axis.ticks.y=element_blank(), axis.text.y=element_blank(), aspec
)
```



On the box-and-whisker plot we observe that the median is slightly skewed towards the first quartile. Our next check is to verify the share of probability falling within one, two and 3 standard deviations of the mean of the distribution. For a normal distribution those probabilities are 68, 95 and 99.7 % known as “68-95-99.7%” rule.

```
#mean and a standard deviation of htm4 variable
ssd<-sd(cbrfss2013$htm4)
smu <-mean(cbrfss2013$htm4)

#create 3 binary variables to check being within 1,2,3 standard deviations
cbrfss2013 <- cbrfss2013 %>%
mutate(in_1sd = ifelse(abs(htm4 -smu) < 1*ssd, 1, 0)
      ,in_2sd = ifelse(abs(htm4 -smu) < 2*ssd, 1, 0)
      ,in_3sd = ifelse(abs(htm4 -smu) < 3*ssd, 1, 0))

#number and percentage of subjects having a height within 1 standard deviation from the mean
cbrfss2013 %>% filter(in_1sd ==1) %>% summarize(count=n(), pct=n()/nrow(cbrfss2013))
```

```
##    count      pct
## 1 308366 0.636954
```

```
#number and percentage of subjects having a height  within 2 standard deviation from the mean
cbrfss2013 %>% filter(in_2sd ==1) %>% summarize(count=n(), pct=n()/nrow(cbrfss2013))
```

```
##    count      pct
## 1 466406 0.963398
```

```
#number and percentage of subjects having a height  within 3 standard deviation from the mean
cbrfss2013 %>% filter(in_3sd ==1) %>% summarize(count=n(), pct=n()/nrow(cbrfss2013))
```

```
##    count      pct
## 1 482738 0.997133
```

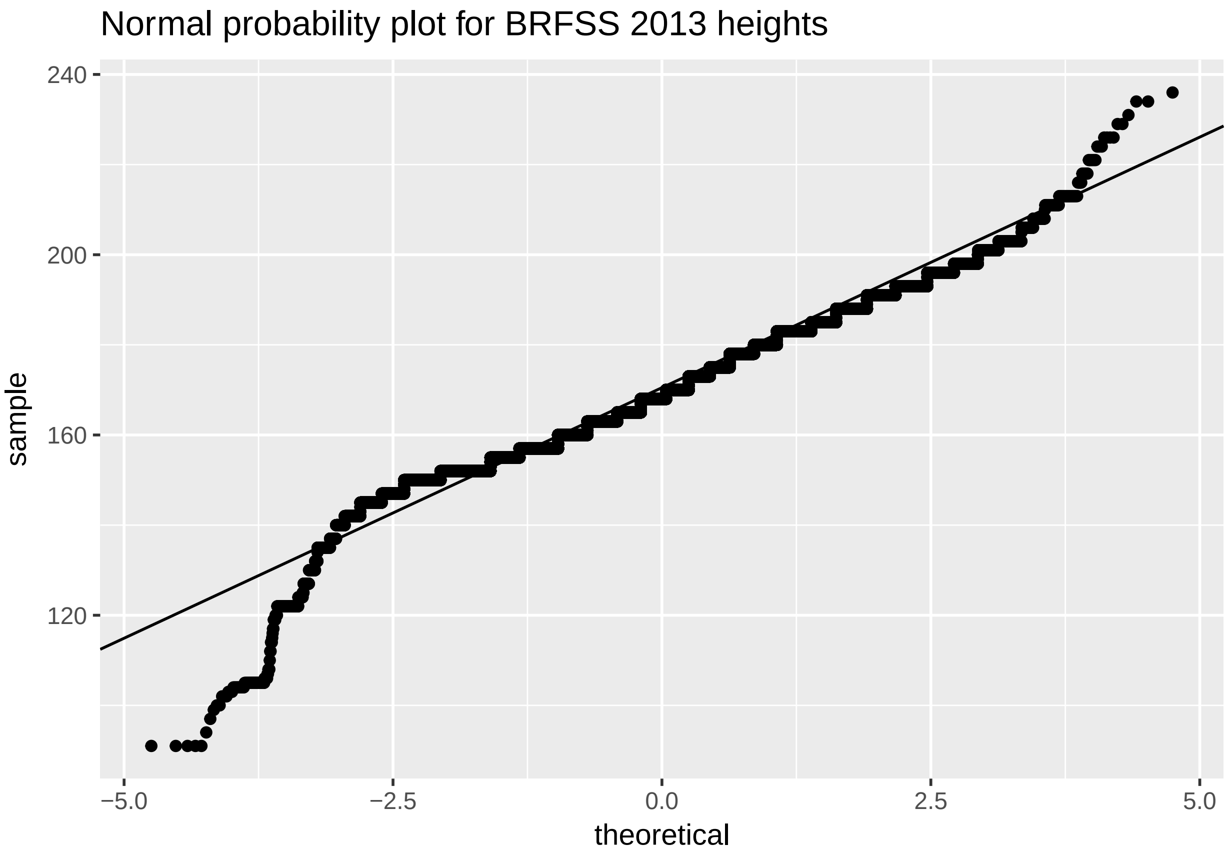
We observe a good match of the probability percentile within 3 standard deviations from the mean and an approximate match for probabilities within 1 and 2 standard deviations from the mean.

The most sensitive measure of ‘normality’ is a normal probability plot.

```
# normal probaility plot
qqplot.qq <- function (vec, titlestr) # argument: vector of numbers
{
  y <- quantile(vec[!is.na(vec)], c(0.25, 0.75))
  x <- qnorm(c(0.25, 0.75))
  slope <- diff(y)/diff(x)
  int <- y[1L] - slope * x[1L]
  d <- data.frame(resids = vec)

  ggplot(d, aes(sample = resids)) + stat_qq() + geom_abline(slope = slope, intercept = int) + ggtitle(t
}

qqplot.qq(cbrfss2013$htm4, 'Normal probability plot for BRFSS 2013 heights')
```



The plotted chart suggests a good linear approximation within 3 standard deviations from the mean. Further than 3 standard deviations the chart shows a large presence of data points in the distribution's tails compared to the theoretical normal distribution.

Thus we can conclude that human height from the BRFSS dataset is distributed only nearly normally.

Appendix: List of used fields in the BFRSS dataset

The following list is extracted from ("BFRSS Codebook," n.d.).

Variable	Question	Datatype
`htm4`	2184-2186: Height in meters [2 implied decimal places]	int
`genhlth`	80: Would you say that in general your health is	factor
`X_rfhlth`	2101: Adults with good or better health	factor
`X_educag`	2199: Level of education completed	factor
`X_incomg`	2200: Income categories	factor

Variable	Question	Datatype
`hlthpln1`	87: Health Care Access	factor
`X_frutsum`	2247-2252: Number of Fruits consumed per day (two implied decimal places)	int
`X_vegesum`	2253-2258: Number of Vegetables consumed per day (two implied decimal places)	int
`X_pacat1`	2348: Physical Activity Categories	factor
`physhlth`	81-82: For how many days during the past 30 days was your physical health not good?	int

References

"BFRSS Codebook." n.d. http://www.cdc.gov/brfss/annual_data/2013/pdf/CODEBOOK13_LLCP.pdf.

Cetinkaya-Rundel, M, D Barr, and D Diez. 2015. *OpenIntro Statistics. Third Edition*. Book. https://www.openintro.org/stat/textbook.php.

Ho, K. 2011. *Growth Hormone Related Diseases and Therapy: A Molecular and Physiological Perspective for the Clinician*. Book. Springer Science & Business Media.

Kabat, G, M Anderson, M Heo, Hosgood D, V Kamenasky, and T Rohan. 2013. "Adult Stature and Risk of Cancer at Different Anatomic Sites in a Cohort of Postmenopausal Women." Journal Article. *Cancer Epidemiol Biomarkers Prev* 22(8): 1–11. http://cebp.aacrjournals.org/content/early/2013/07/25/1055-9965.EPI-13-0305.abstract.

"List of Shortest People." n.d. Article. https://en.wikipedia.org/wiki/List_of_shortest_people.

"List of Tallest People." n.d. Article. https://en.wikipedia.org/wiki/List_of_tallest_people.

"The Behavioral Risk Factor Surveillance System." n.d. https://www.cdc.gov/brfss/.