lab_2_Introduction to data

Salma Elshahawy

9/4/2019

Introduction to data

```r
source("cdc.R")
names(cdc)
```

```
## [1] "genhlth"  "exerany"  "hlthplan" "smoke100" "height"   "weight"
## [7] "wtdesire" "age"      "gender"
```

Exercise_1: How many cases are there in this data set? How many variables? For each variable, identify its data type (e.g. categorical, discrete).

```r
dim(cdc)
```

```
## [1] 20000     9
```

There are 20,000 cases. There are nine(9) variables.

variable name	type of the variable	type 2
index	numerical	continuous
genhlth	categorical	ordinal
exerany	categorical	
hlthplan	categorical	
smoke100	categorical	
height	numerical	continuous
weight	numerical	continuous
wtdesire	numerical	continuous
age	numerical	continuous
gender	categorical	

```r
head(cdc)
```

```
##     genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 1      good       0        1        0     70    175      175  77      m
## 2      good       0        1        1     64    125      115  33      f
## 3      good       1        1        1     60    105      105  49      f
## 4      good       1        1        0     66    132      124  42      f
## 5 very good       0        1        0     61    150      130  55      f
## 6 very good       1        1        0     64    114      114  55      f
```

```r
tail(cdc)
```

```
##         genhlth exerany hlthplan smoke100 height weight wtdesire age
## 19995      good       0        1        1     69    224      224  73
## 19996      good       1        1        0     66    215      140  23
## 19997 excellent       0        1        0     73    200      185  35
## 19998      poor       0        1        0     65    216      150  57
## 19999      good       1        1        0     67    165      165  81
## 20000      good       1        1        1     69    170      165  83
##       gender
## 19995      m
## 19996      f
## 19997      m
## 19998      f
## 19999      f
## 20000      m
```

```r
summary(cdc$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    68.0   140.0   165.0   169.7   190.0   500.0
```

```r
190 - 140
```

```
## [1] 50
```

```r
mean(cdc$weight)
```

```
## [1] 169.683
```

```r
var(cdc$weight)
```

```
## [1] 1606.484
```

```r
median(cdc$weight)
```

```
## [1] 165
```

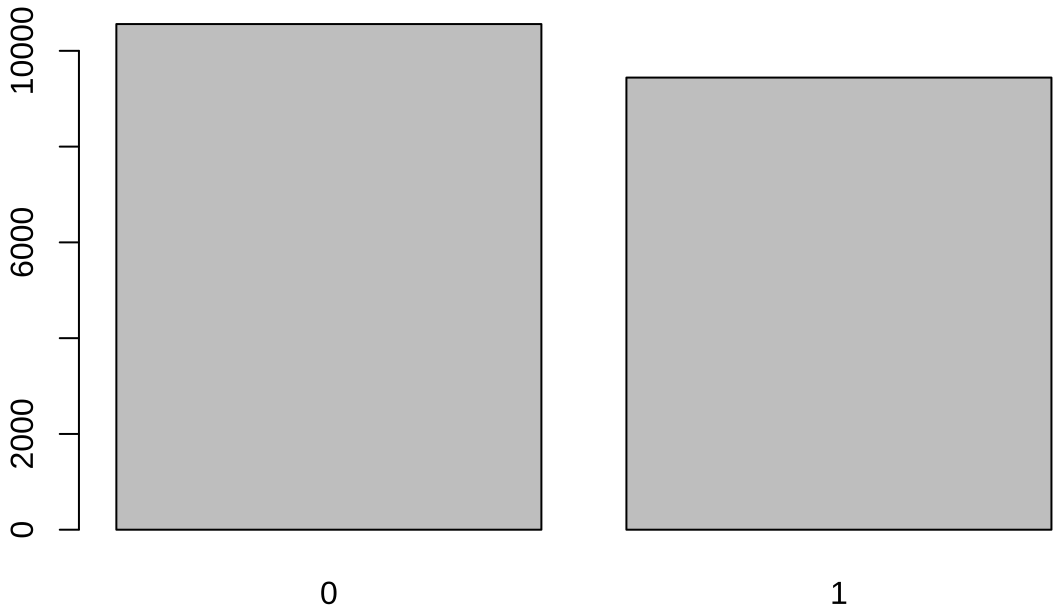
```r
table(cdc$smoke100)
```

```
##
##     0     1
## 10559  9441
```

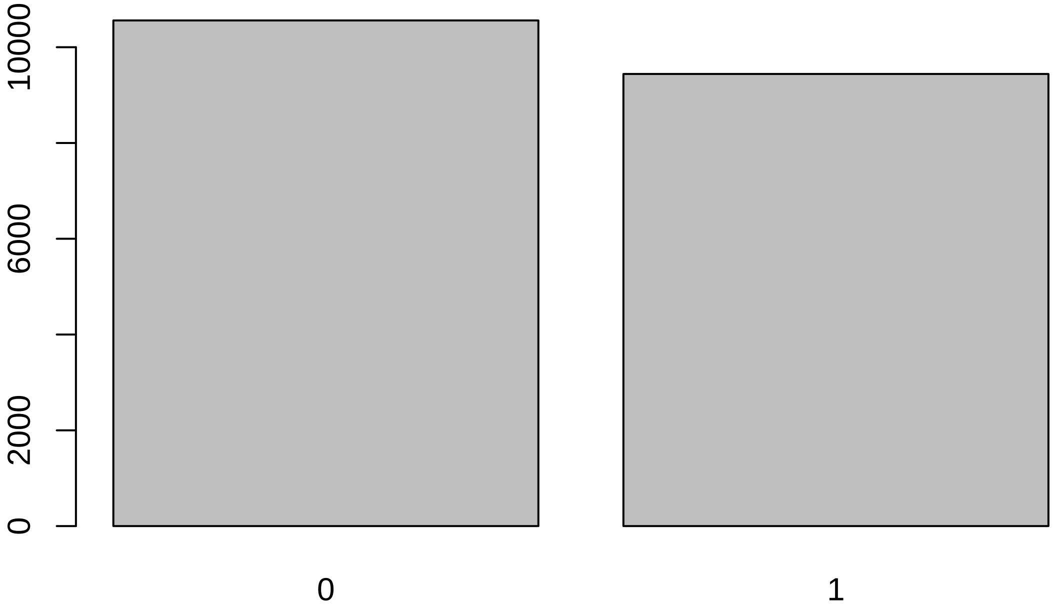
```r
table(cdc$smoke100)/20000
```

```
##
##       0       1
## 0.52795 0.47205
```

```r
barplot(table(cdc$smoke100))
```



```r
smoke <- table(cdc$smoke100)
barplot(smoke)
```



### Exercise_2: Create a numerical summary for height and age, and compute the interquartile range for each. Compute the relative frequency distribution for gender and exerany. How many males are in the sample? What proportion of the sample reports being in excellent health?

```
# getting summary for height
summary(cdc$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   48.00   64.00   67.00   67.18   70.00   93.00
```

```
# getting the interquartile range
70 - 64
```

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

```
#summary for age
summary(cdc$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.00   31.00   43.00   45.07   57.00   99.00
```

```
# interquartile for age
57 - 31
```

```
## [1] 26
```

the relative frequency distribution for gender

```
table(cdc$gender)/20000
```

```
##
##       m       f
## 0.47845 0.52155
```

How many males are in the sample?

```
table(cdc$gender)
```

```
##
##     m     f
##  9569 10431
```

there are 9,569 males in the sample

the relative frequency distribution for exerany

```
table(cdc$exerany)/20000
```

```
##
##      0      1
## 0.2543 0.7457
```

What proportion of the sample reports being in excellent health

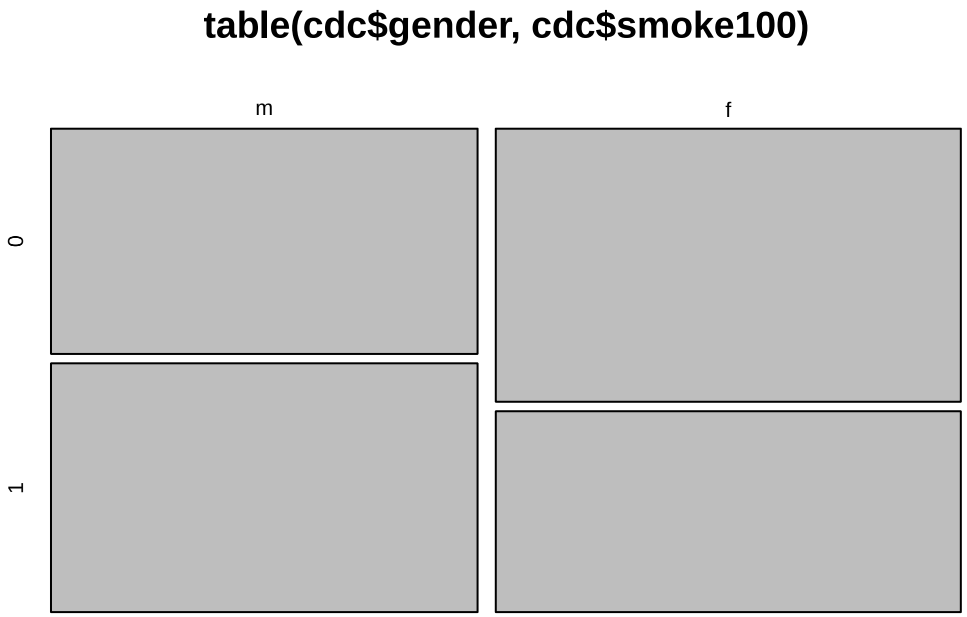
```
table(cdc$genhlth)['excellent']/20000
```

```
## excellent
##   0.23285
```

```
table(cdc$gender,cdc$smoke100)
```

```
##
##        0    1
##   m 4547 5022
##   f 6012 4419
```

```
mosaicplot(table(cdc$gender,cdc$smoke100))
```



Exercise_3: What does the mosaic plot reveal about smoking habits and gender?

Males smoking more then 100 cigerattes than females.

```
dim(cdc)
```

```
## [1] 20000     9
```

```
cdc[567, 6]
```

```
## [1] 160
```

```
cdc[1:10, 6]
```

```
##  [1] 175 125 105 132 150 114 194 170 150 180
```

```
cdc[1:10, ]
```

```
##      genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 1       good       0        1        0     70    175      175  77      m
## 2       good       0        1        1     64    125      115  33      f
## 3       good       1        1        1     60    105      105  49      f
## 4       good       0        1        0     66    132      124  42      f
## 5  very good       0        1        0     61    150      130  55      f
## 6  very good       1        1        0     64    114      114  55      f
## 7  very good       1        1        0     71    194      185  31      m
## 8  very good       0        1        0     67    170      160  45      m
## 9       good       0        1        1     65    150      130  27      f
## 10      good       1        1        0     70    180      170  44      m
```

```
mdata <- subset(cdc, cdc$gender == "m")
head(mdata)
```

```
##      genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 1       good       0        1        0     70    175      175  77      m
## 7  very good       1        1        0     71    194      185  31      m
## 8  very good       0        1        0     67    170      160  45      m
## 10      good       1        1        0     70    180      170  44      m
## 11 excellent       1        1        1     69    186      175  46      m
## 12      fair       1        1        1     69    168      148  62      m
```

Exercise_4: Create a new object called under23_and_smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

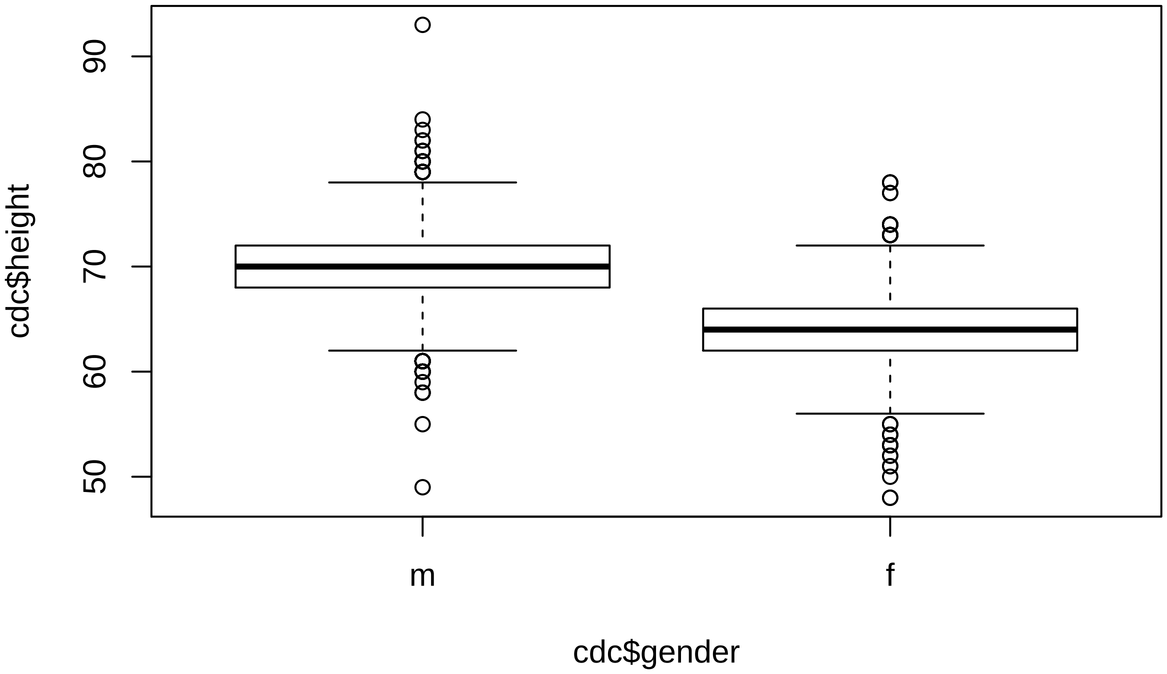
```
under23_and_smoke <- subset(cdc, smoke100 == 1 & age < 23)
head(under23_and_smoke)
```

```
##       genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 13  excellent       1        0        1     66    185      220  21      m
## 37  very good       1        0        1     70    160      140  18      f
## 96  excellent       1        1        1     74    175      200  22      m
## 180      good       1        1        1     64    190      140  20      f
## 182 very good       1        1        1     62     92       92  21      f
## 240 very good       1        0        1     64    125      115  22      f
```

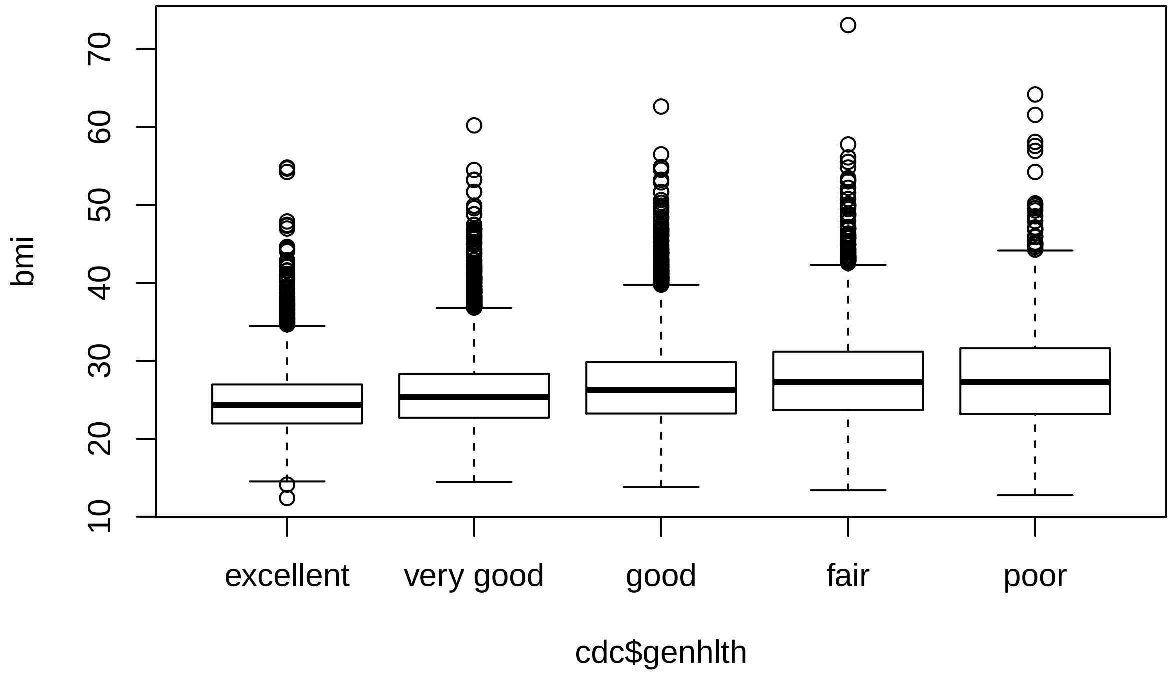
```
summary(cdc$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   64.00   67.00   67.18   70.00   93.00
```

```
boxplot(cdc$height ~ cdc$gender)
```



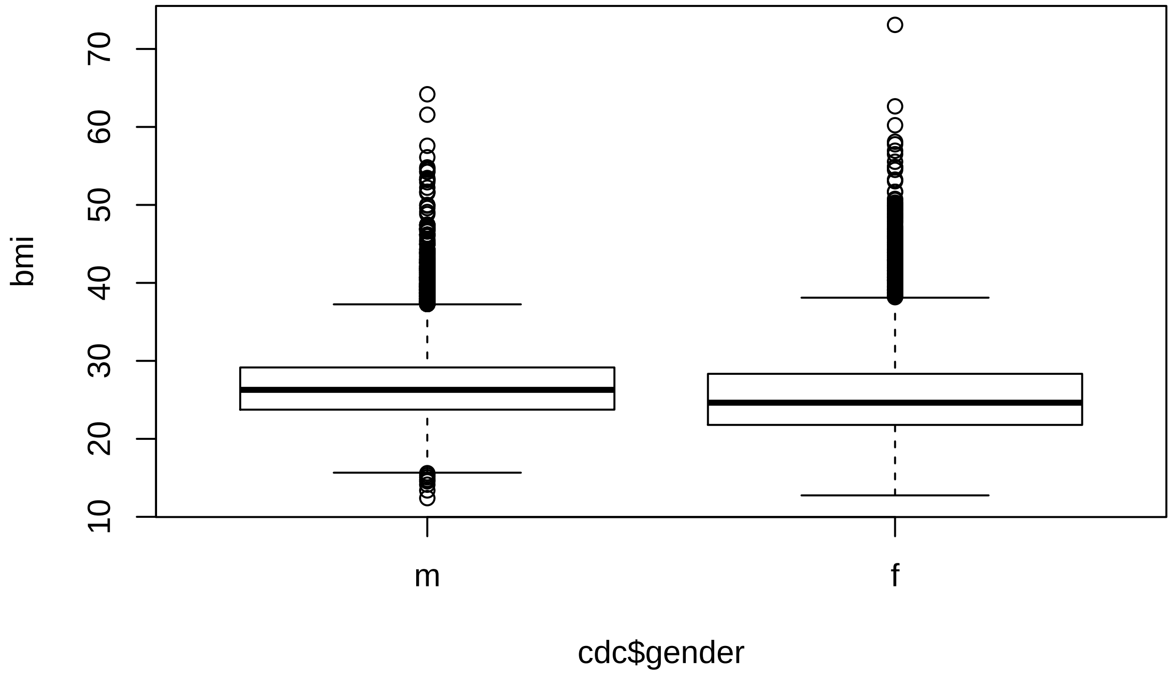
```
bmi <- (cdc$weight / cdc$height^2) * 703
boxplot(bmi ~ cdc$genhlth)
```



Exercise_5: What does this box plot show? Pick another categorical variable from the data set and see how it relates to BMI. List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.

It shows the calculated BMI for all participants corresponds to genhlth variable. As illustrated, it shows an increasing in the BMI.

```
boxplot(bmi ~ cdc$gender)
```

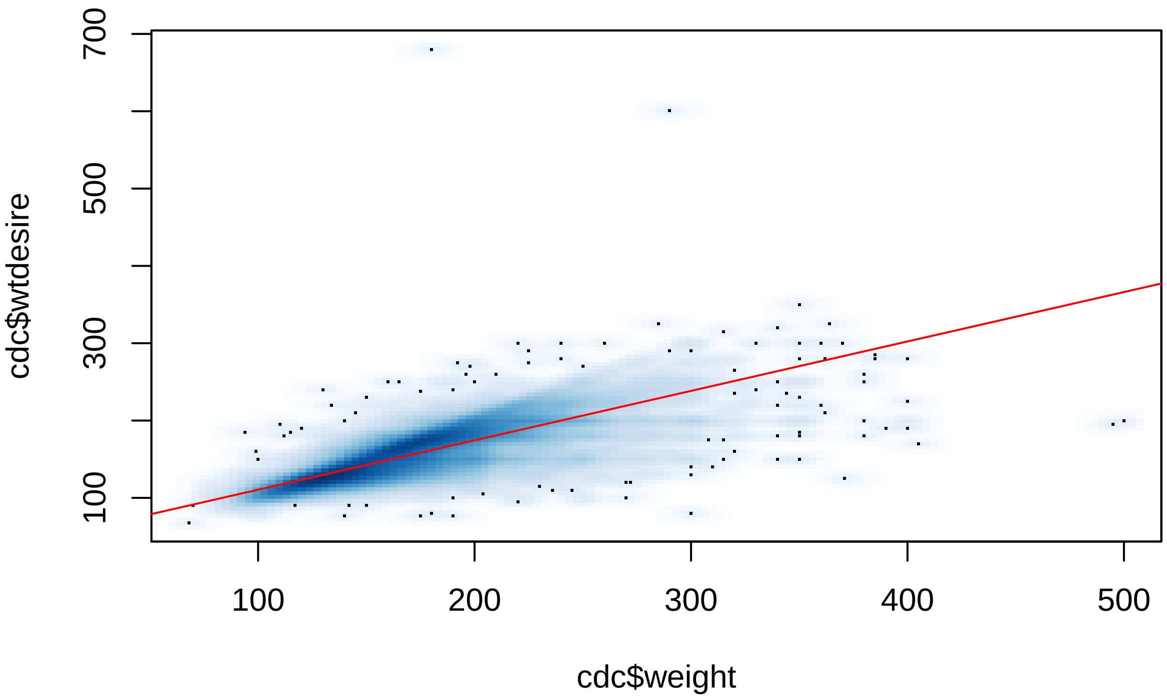


The boxplot shows a likely similar BMI for both genders. However, males seem to have BMI below 30.

On Your Own

1. Make a scatterplot of weight versus desired weight. Describe the relationship between these two variables.

```
smoothScatter(cdc$wtdesire ~ cdc$weight)
abline(lm(cdc$wtdesire~cdc$weight), col="red")
```



The relationship is positive.

2. Let's consider a new variable: the difference between desired weight (wtdesire) and current weight (weight). Create this new variable by subtracting the two columns in the data frame and assigning them to a new object called wdiff.

```
wdiff <- (cdc$wtdesire - cdc$weight)
```

3. What type of data is wdiff? If an observation wdiff is 0, what does this mean about the person's weight and desired weight. What if wdiff is positive or negative?

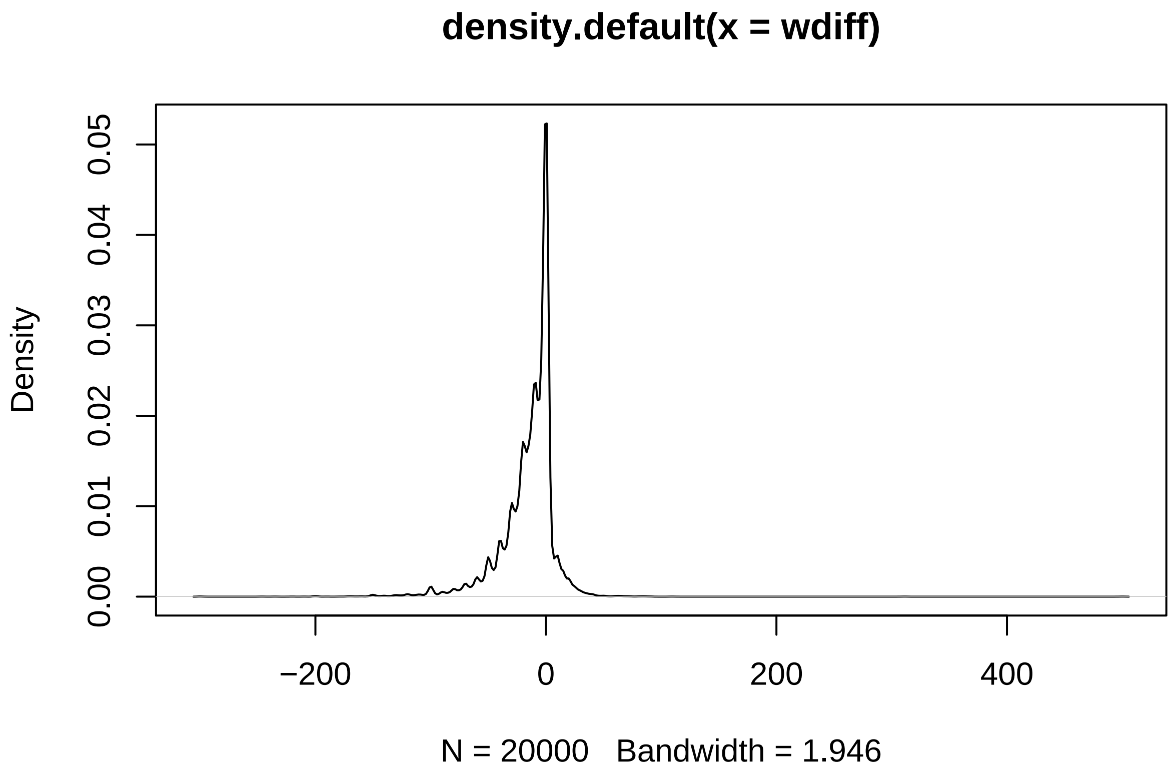
```
typeof(wdiff)
```

```
## [1] "integer"
```

If the observation of wdiff is 0 means that the person has an ideal weight (his weight is same as desired). If the wdiff is positive means that the person needs to gain weight to reach ideal. However, if the wdiff is negative means that the person needs to lose weight.

4. Describe the distribution of wdiff in terms of its center, shape, and spread, including any plots you use. What does this tell us about how people feel about their current weight?

```
differ <- density(wdiff)
plot(differ)
```



This density plot reflects that most of the responders are happy with their weight(mode is 0)

```
mean(wdiff)
```

```
## [1] -14.5891
```

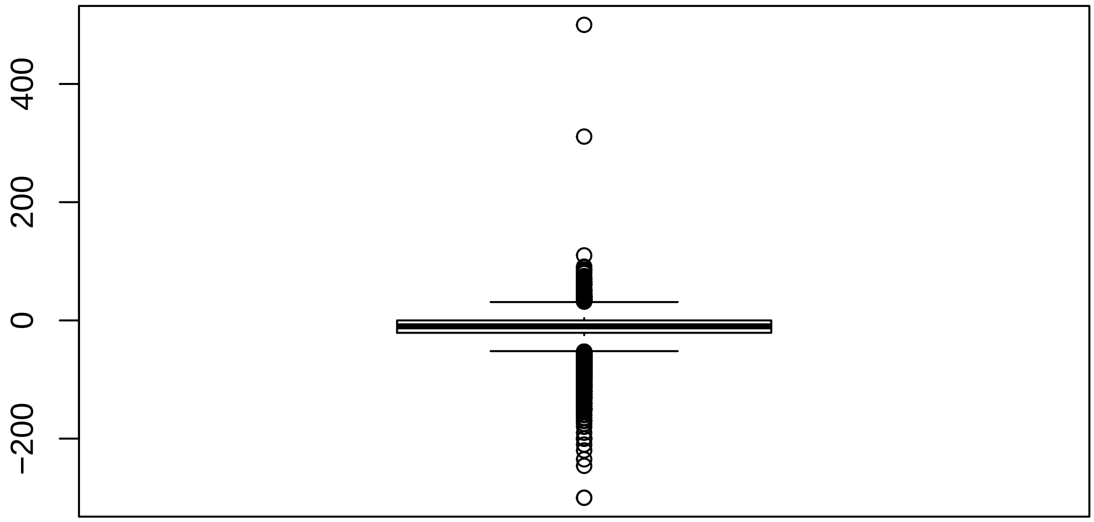
```
median(wdiff)
```

```
## [1] -10
```

```
quantile(wdiff)
```

```
##   0%  25%  50%  75% 100% 
## -300  -21  -10    0  500
```

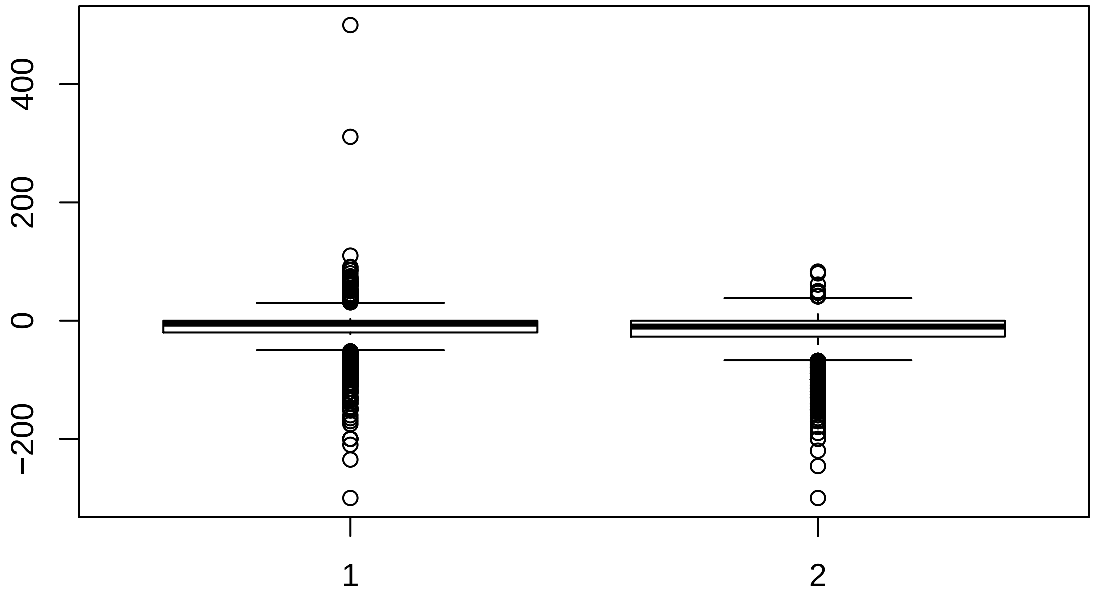
```
boxplot(wdiff)
```



From the boxplot, we can see the outlier points of people who think they should be 250 IB heavier.

4. Using numerical summaries and a side-by-side box plot, determine if men tend to view their weight differently than women.

```
m_desire <- subset(cdc, cdc$gender == 'm')$wtdesire
f_desire <- subset(cdc, cdc$gender == 'f')$wtdesire
m_weight <-subset(cdc, cdc$gender == 'm')$weight
f_weight <- subset(cdc, cdc$gender == 'f')$weight
boxplot(m_desire - m_weight, f_desire - f_weight)
```



```
summary(m_desire - m_weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -300.00  -20.00   -5.00  -10.71    0.00  500.00
```

```
summary(f_desire - f_weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -300.00  -27.00  -10.00  -18.15    0.00   83.00
```

The summary and boxplot showing that males are likely think that they like to lose weight.

6. Now it's time to get creative. Find the mean and standard deviation of weight and determine what proportion of the weights are within one standard deviation of the mean.

```
mean(cdc$weight)
```

```
## [1] 169.683
```

```
sd(cdc$weight)
```

```
## [1] 40.08097
```

```
below_mean <-subset(cdc, cdc$weight > mean(cdc$weight)-sd(cdc$weight))
above_mean <-subset(cdc, cdc$weight < mean(cdc$weight)+sd(cdc$weight))
within_sd <-subset(below_mean, below_mean$weight < max(above_mean$weight))
nrow(within_sd)/nrow(cdc)
```

```
## [1] 0.7071
```