Lab1

Priya Shaji

February 4, 2019

Load the cdc dataset

```
source("more/cdc.R")
```

view the names of the variables

```
dim(cdc)
```

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

```
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).

Answer 1.

Cases: 20,000
variables: 9 Data Type of each variable genhlth - Categorical exerany - Categorical hlthplan - Categorical smoke100 - Categorical height - Numeric, continuous weight - Numeric, continuous wtdesire - Numeric, continuous age - Numeric, discrete(since age can take only integer values, therefore it's discrete) gender - Categorical

First few entries (rows) of our data

```
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
```

Last few entries (rows) of our data

```
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
```

summary of column weight

```
summary(cdc$weight)

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

calculate the mean, median, and variance of weight

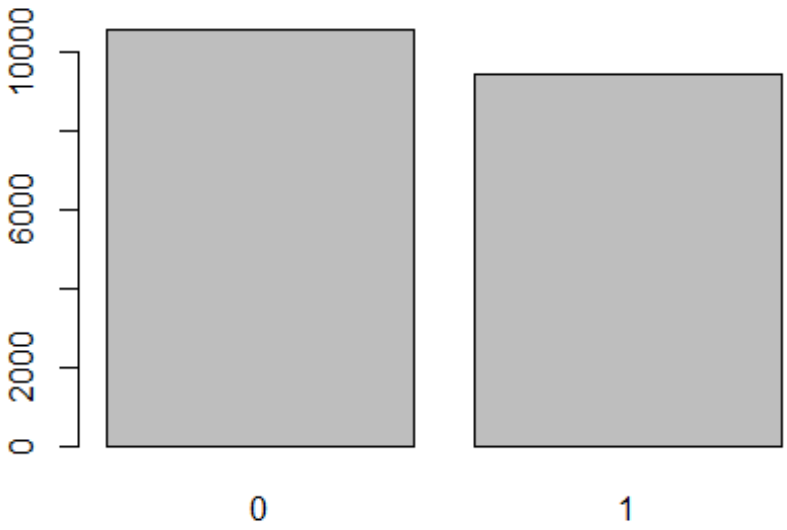
```
mean(cdc$weight)

## [1] 169.683

var(cdc$weight)

## [1] 1606.484

median(cdc$weight)

## [1] 165
```

For categorical data, we consider their sample frequency or relative frequency distribution For example, to see the number of people who have smoked 100 cigarettes in their lifetime

```
table(cdc$smoke100)

##
##     0     1
## 10559  9441

table(cdc$smoke100)/20000

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

bar plot of the entries in the table

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



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?

Answer2.

summary for height and age

```r
##height
summary(cdc$height)

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

##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 range for each

```r
##height
70.00-64.00
```

```
## [1] 6

##age
57.00-31.00

## [1] 26
```

relative frequency distribution for gender and exerany

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

##
##       m       f
## 0.47845 0.52155

##exerany
table(cdc$exerany)/20000

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

How many males are in the sample

```
table(cdc$gender)

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

No. of males are: 9569 47.8% of the sample are males

What proportion of the sample reports being in excellent health?

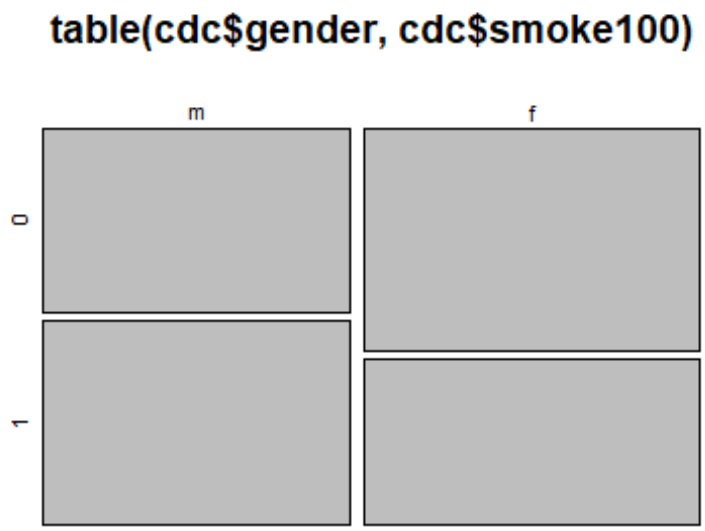
```
table(cdc$genhlth)/20000

##
## excellent very good      good      fair      poor
##   0.23285   0.34860   0.28375   0.10095   0.03385
```

The table command can be used to tabulate any number of variables

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

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

create a mosaic plot of this table

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



Exercise 3

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

Answer 3.

Mosaic plot of above table shows us that the smoking habits of males are higher as compared to females

sixth variable of the 567th respondent

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

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

the weights for the first 10 respondents

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

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

all of the data for the first 10 respondents

```
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       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
## 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
```

to extract just the data for the men in the sample,create a subset

```
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
```

to extract just the data for the men and also who are over 30 in the sample

```
m_and_over30 <- subset(cdc, gender == "m" & age > 30)
```

to extract just the data for the men orwho are over 30 in the sample

```
m_or_over30 <- subset(cdc, gender == "m" | age > 30)
```

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.

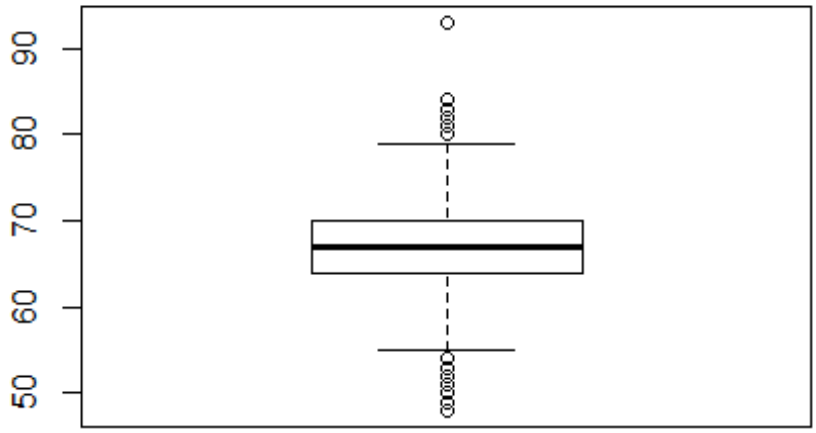
Answer 4.

```
under23_and_smoke<-subset(cdc, smoke100 = 1 & age < 30)
head(under23_and_smoke)
```

```
##     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
```

construct a box plot for a single variable

```
boxplot(cdc$height)
```



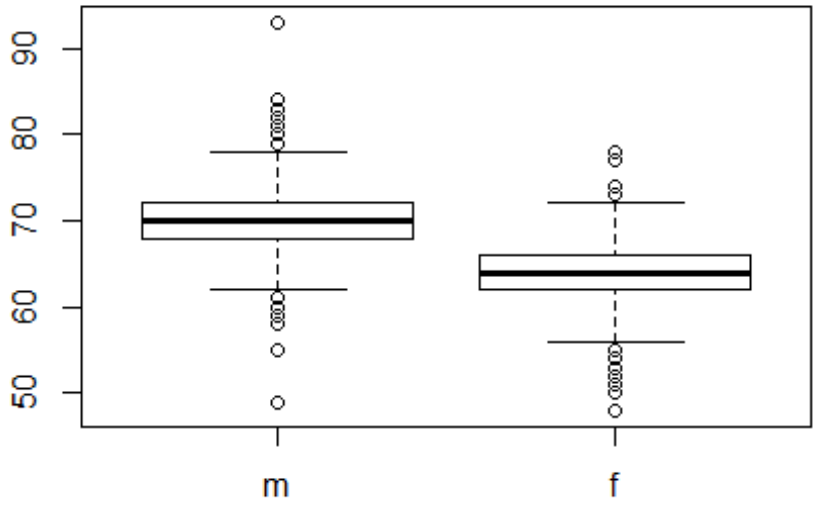
compare the locations of the components of the box by examining the summary statistics.

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

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

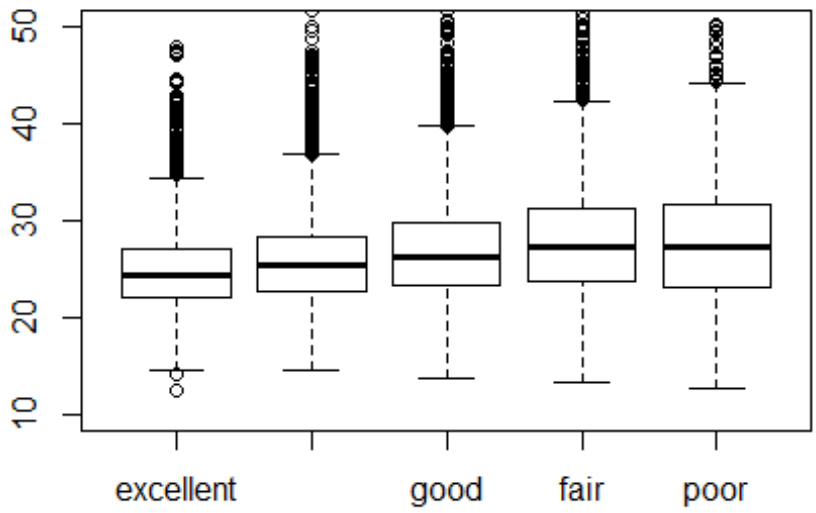
compare the heights of men and women using boxplot

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



make a new object called bmi and then creates box plots of these values, defining groups by the variable cdc$genhlth.

```
bmi <- (cdc$weight / cdc$height^2) * 703
boxplot(bmi ~ cdc$genhlth,ylim=c(10,50))
```



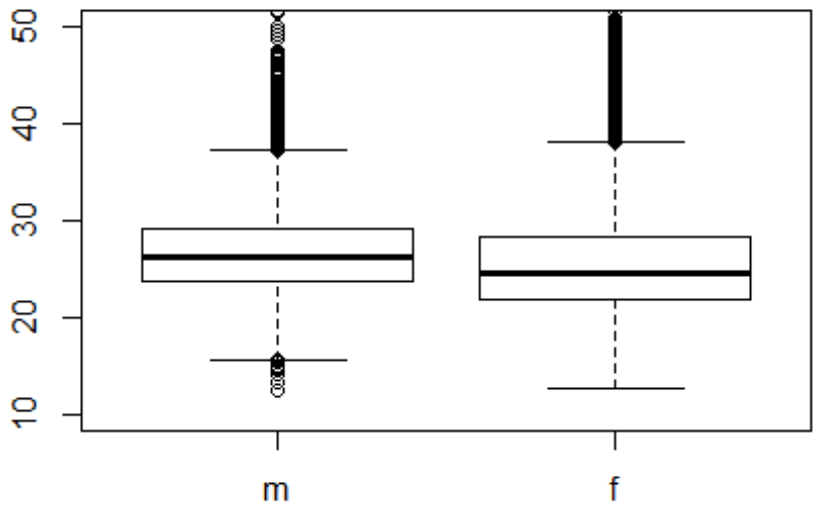
Exercise 5

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.

Answer 5.

bmi does not seem to depend much on the general health facor. Respondants with excellent to very good health seems to have lower bmi's when combared to bmi's of respondants of good, fair,poor heath

```
## using 'gender' as a variable
bmi <- (cdc$weight / cdc$height^2) * 703
boxplot(bmi ~ cdc$gender,ylim=c(10,50))
```



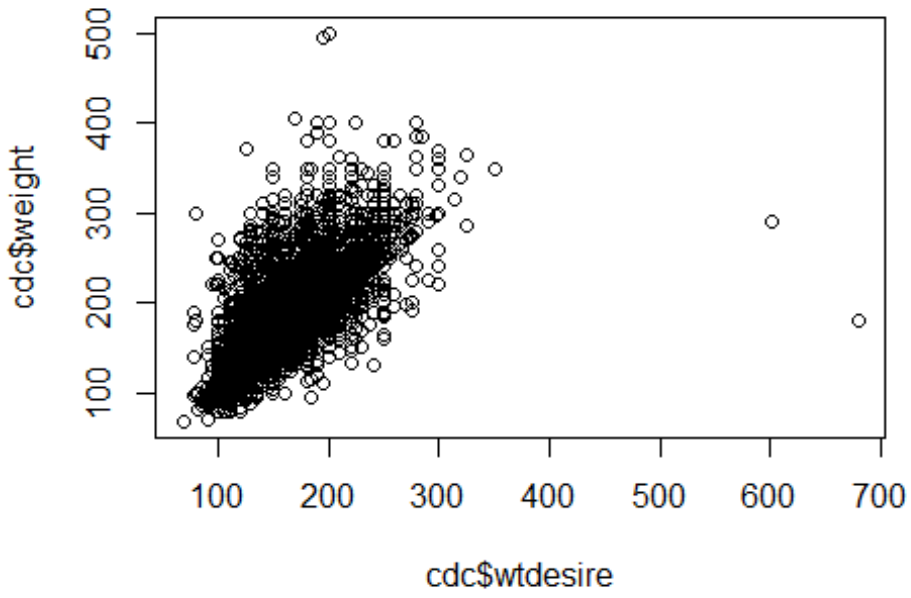
males seems to have lower bmi's as compared to females

On Your Own

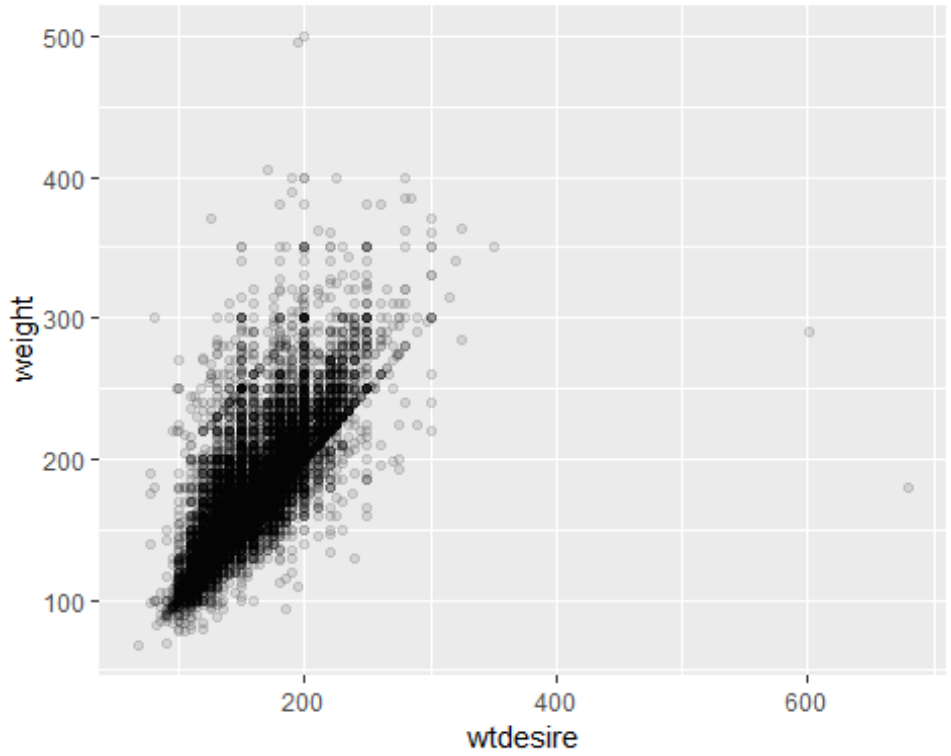
1. Make a scatterplot of weight versus desired weight. Describe the relationship between these two variables.

Answer 1.

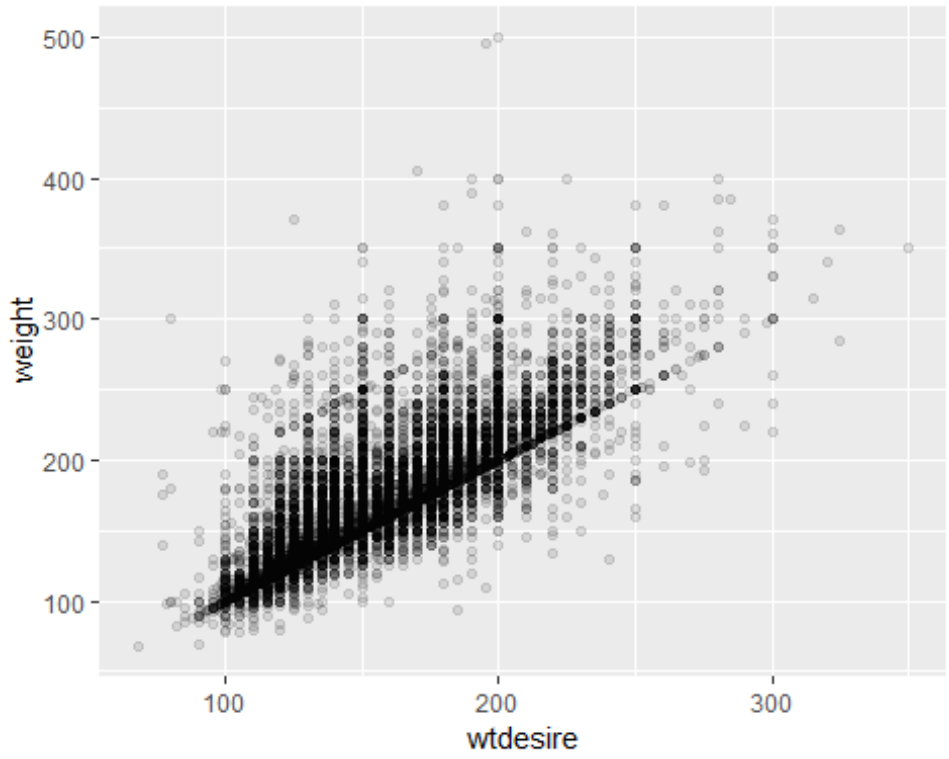
```
plot(cdc$weight~cdc$wtdesire)
```




```
library(ggplot2)
ggplot(cdc,aes(wtdesire,weight)) + geom_point(alpha=1/10)
```



```
ggplot(cdc[which(cdc$wtdesire < 400),],aes(wtdesire,weight)) +
geom_point(alpha=1/10)
```



Current weight and desired weight factors are correlated.Seems like more number of respondants have their desired weight equal to their current weight. For some people current weight tends to be higher than their desired weight.

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.

Answer 2.

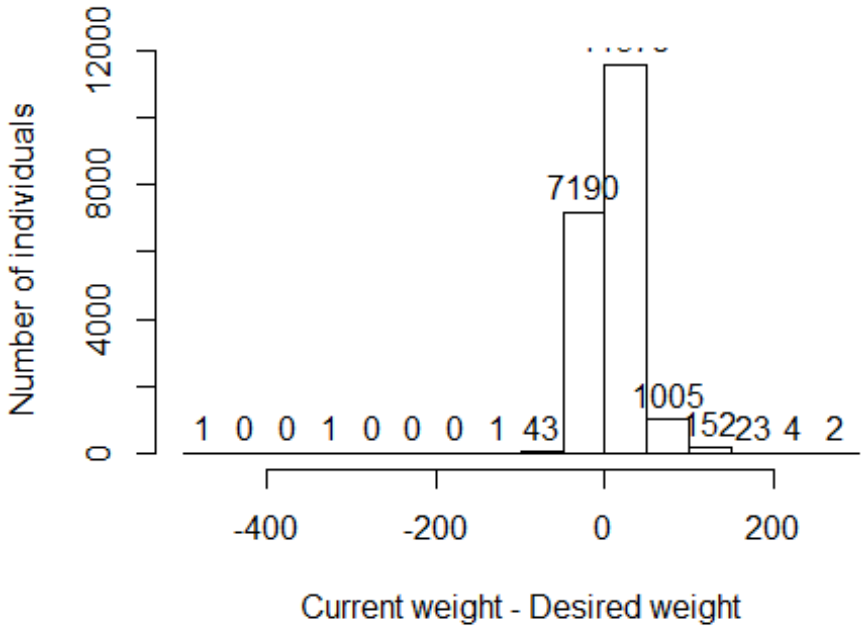
```
wdiff<-cdc$weight-cdc$wtdesire
head(wdiff)

## [1]  0 10  0  8 20  0
```

the data type of 'wdiff' is: numeric and discrete. If an observation wdiff is 0: The respondant have the same current weight and desired weight if wdiff is positive: respondants current weight > desired weight(They want to loose weight) if wdiff is negetive: respondants current weight < desired weight(They may or may not want to loose 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?

Answer 4.

```
hist(wdiff,xlab="Current weight - Desired weight",ylab="Number of
individuals",main="",labels=TRUE)
```




```
median(wdiff)

## [1] 10
```

By the above obervation , there are more respondants who wish to loose weight or their current weight is higher that the desired weight.

Center of the data is 10 , means half of the respondants current weight is higher than the desired weight and wish to loose 10 pounds.

The plot is left-skewed, means there are only a few respondants who weigh less than their desired weight. So they wish to loose less weight or they do not wish to loose weight at all.

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

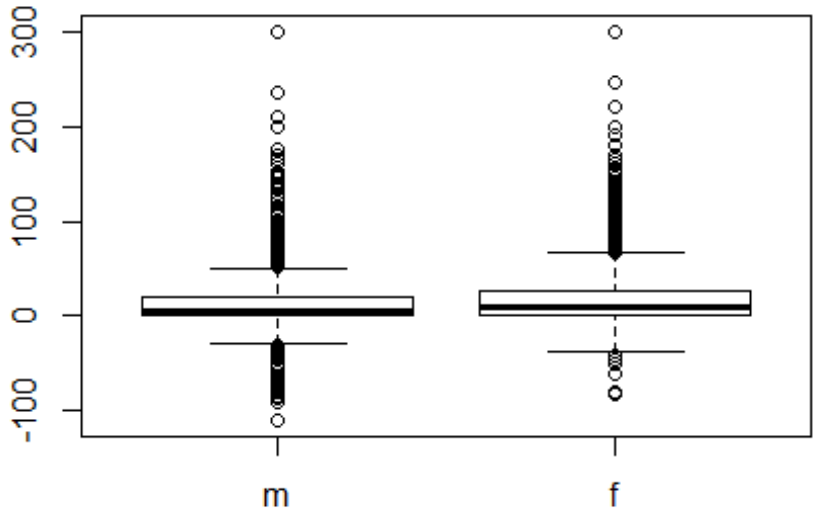
Answer 5.

```
##Analyzing for males
summary(cdc$weight,cdc$gender=='m')

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

```
##Analyzing for females
summary(cdc$weight,cdc$gender=='f')

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

create a dataframe

```
weight_gender<-data.frame(wdiff = wdiff,gender = cdc$gender)

## remove the two outliers who's data points show 600+ pounds weight
weight_gender <- weight_gender[which(cdc$wtdesire < 600),]
## generate the side by side boxplot
boxplot(wdiff ~ gender,data=weight_gender)
```



By observing the above plot, we infer that females wants to loose more weight as compared to men. And also female's current weight is greater that the desired weight . With men's weight data points are more towards the negative scale, it shows that most of the men's current weight is less than the desired 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.

Answer 6.

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

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

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

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

what proportion of the weights are within one standard deviation of the mean.

```
below_sd <- mean(cdc$weight) - sd(cdc$weight)
above_sd <- mean(cdc$weight) + sd(cdc$weight)
length(which(cdc$weight >= below_sd & cdc$weight <= above_sd))/20000
```

```
## [1] 0.7076
```