Lab-1

Bin Lin

2016-9-4

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

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).

20000 observations, 9 variables. The data types of each variables are shown in the following R code.

```
str(cdc)
```

```
## 'data.frame':    20000 obs. of  9 variables:
##  $ genhlth : Factor w/ 5 levels "excellent","very good",..: 3 3 3 3 2 2 2 2 3 3 ...
##  $ exerany : num  0 0 1 1 0 1 1 0 0 1 ...
##  $ hlthplan: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ smoke100: num  0 1 1 0 0 0 0 0 1 0 ...
##  $ height  : num  70 64 60 66 61 64 71 67 65 70 ...
##  $ weight  : int  175 125 105 132 150 114 194 170 150 180 ...
##  $ wtdesire: int  175 115 105 124 130 114 185 160 130 170 ...
##  $ age     : int  77 33 49 42 55 55 31 45 27 44 ...
##  $ gender  : Factor w/ 2 levels "m","f": 1 2 2 2 2 2 1 1 2 1 ...
```

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?

```
#Summary for height and age and their interquatile range
summary(cdc$height)
```

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

```
70-64
```

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

```
summary(cdc$age)
```

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

```
57-31
```

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

```
#Relative frequency tables of gender and exerany
table(cdc$gender)/20000
```

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

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

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

```
#Number of males
0.47845*20000
```

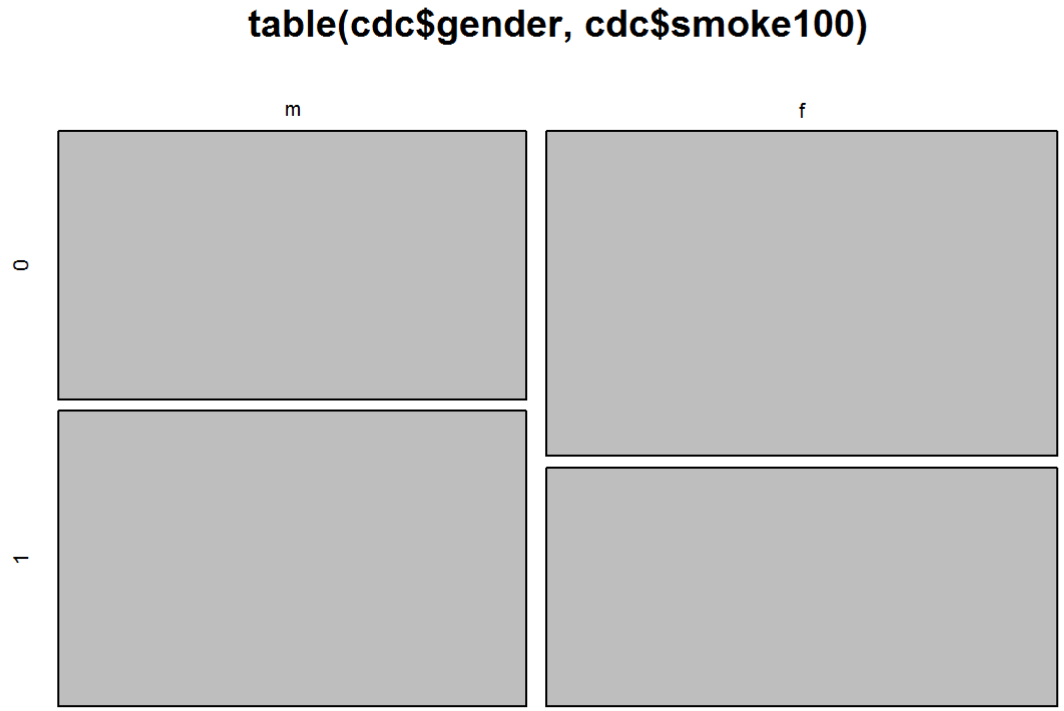
```
## [1] 9569
```

```
#Proportion of the sample being in excellent health = 23.285%
table(cdc$genhlth)/20000
```

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

Exercise 3. What does the mosaic plot reveal about smoking habits and gender?

```
#It tells us that male has relatively higher proportion of smoking population compare to female
mosaicplot(table(cdc$gender,cdc$smoke100))
```

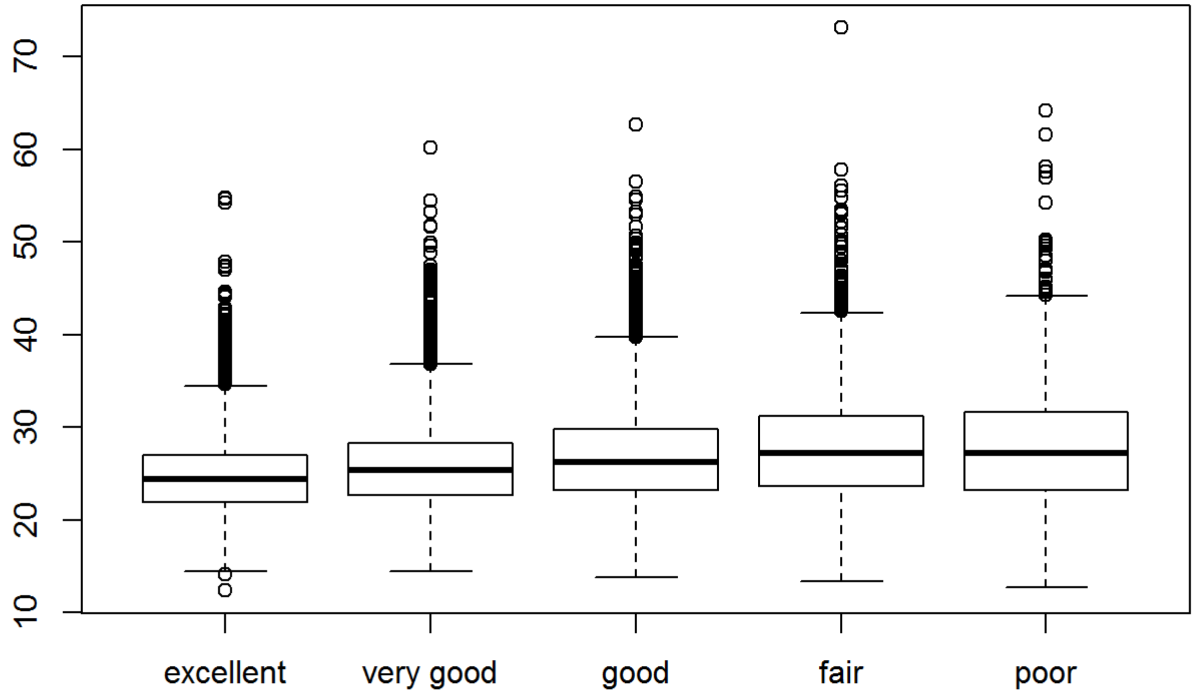


Exercise 4. Create a new object called under23_and_smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

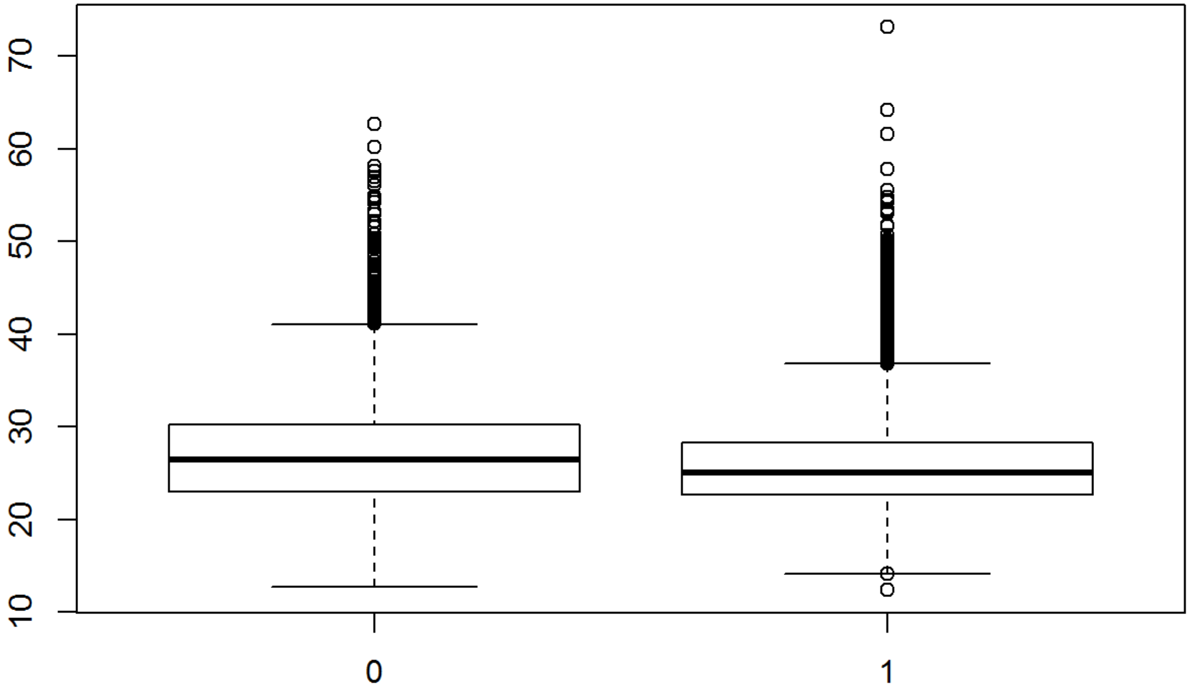
```
under23_and_smoke <- subset(cdc, cdc$age < 23 & cdc$smoke100 == 1)
```

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.

```
#AS the health status of  people gets poorer and poorer, their BMI tend to go up.

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

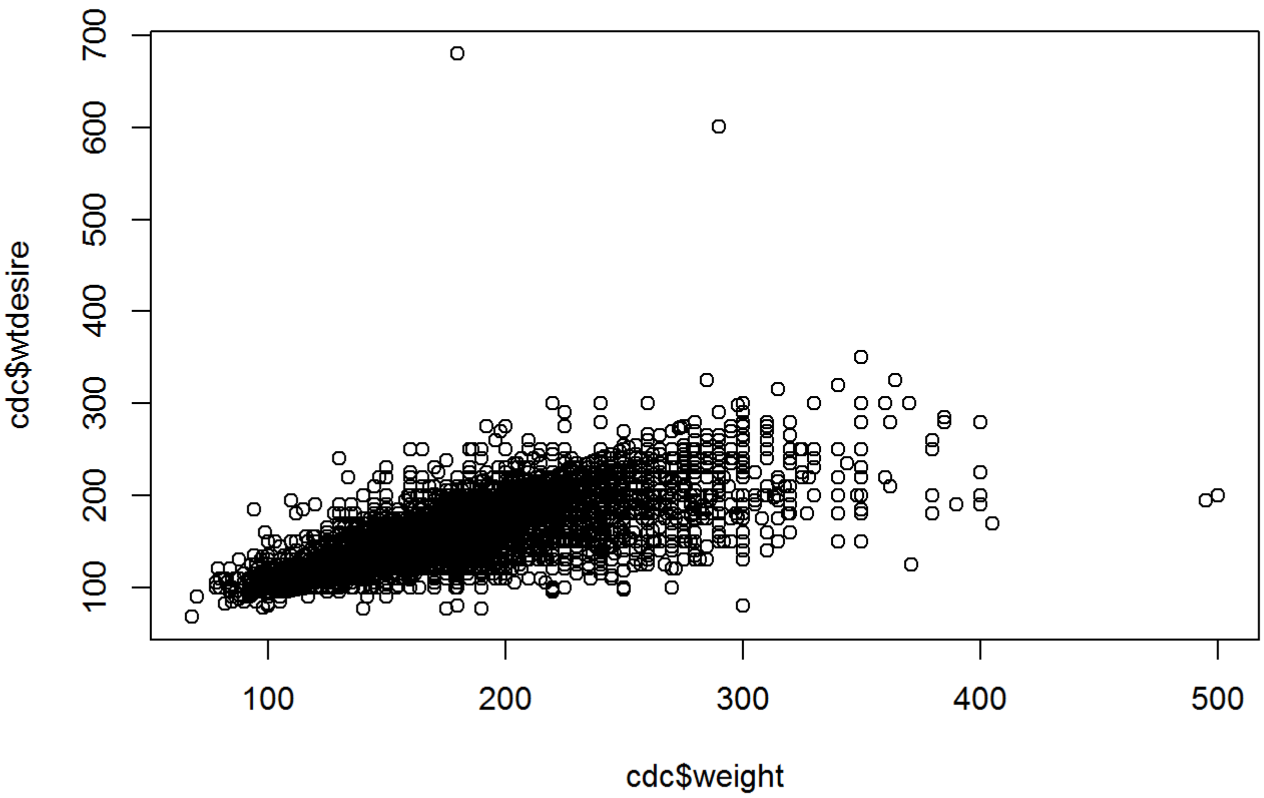


```
#This boxplot shows if people have been exercising, their BMI tend to be smaller. Because I think exercise will burn more ca
lories, therefore poeple lose weight. Their BMI will decrease.
boxplot(bmi ~ cdc$exerany)
```



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

```
#There is a positive linear relationship betweem people's weight and their desired weight. As the weight increases, the corr
esponding desired weight also increases.

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



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$weight - cdc$wtdesire
```

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?

```
#The data type of wdiff is integer. If wdiff is 0, that means person's weight equals to his or her desired weight. If wdiff is positive, that means person's weight is greater the desired weight. He or she needs to lose weight to achieve desired body weight. If wdiff is negative, that means that person's wight is less than the desired weight. He or she needs to gain weight to achieve desired body weight.

str(wdiff)
```

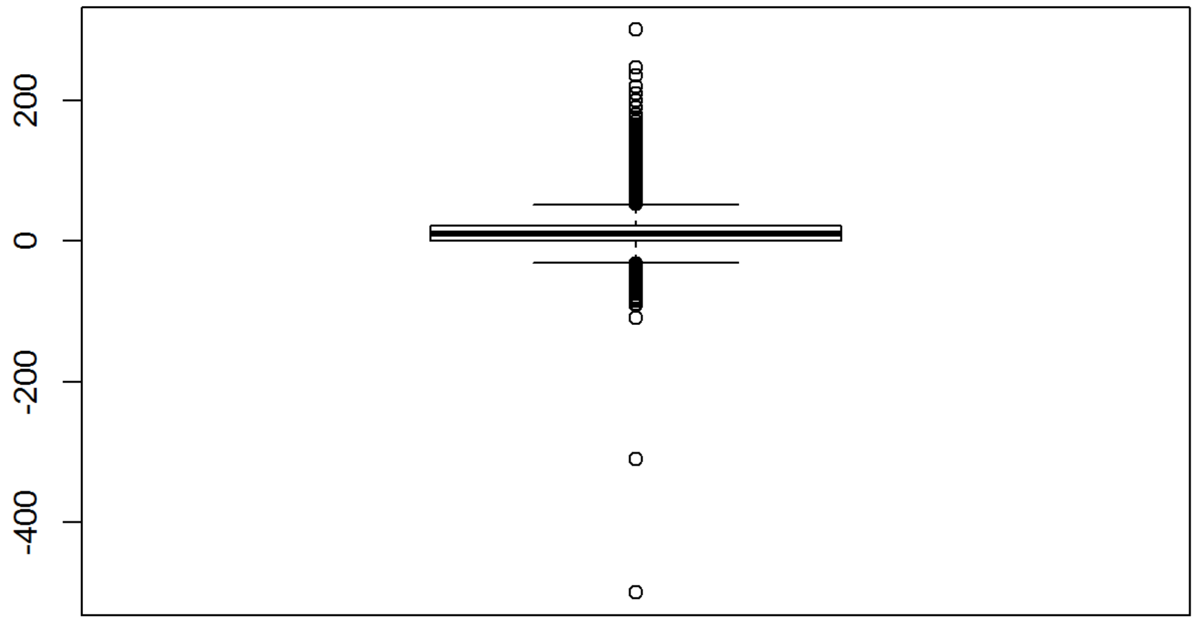
```
##  int [1:20000] 0 10 0 8 20 0 9 10 20 10 ...
```

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?

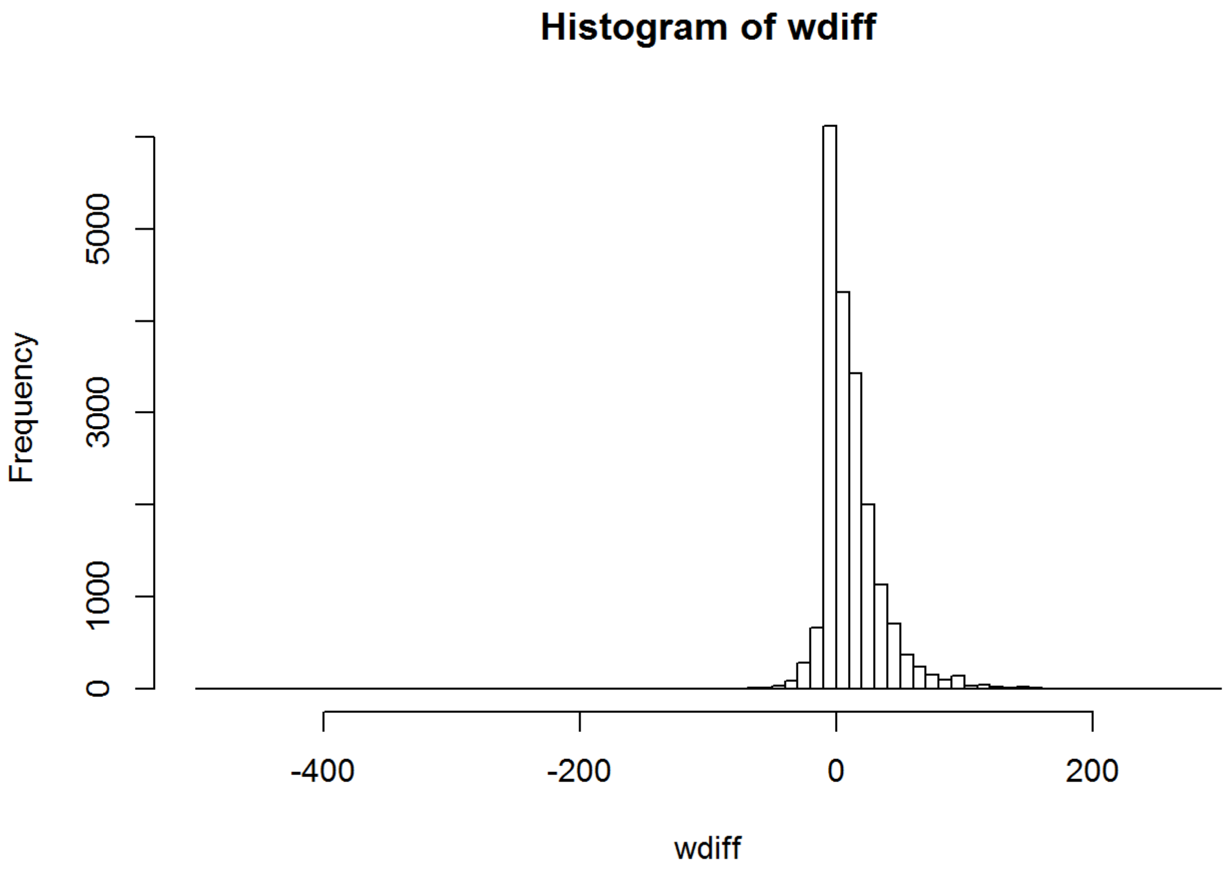
```
#From the summary statement, we know the median of this dataset is 10, the range is from -500 to 300lb. So based on the raw data, there are a lot of outliers. For example, someone whose weight is 500lb less than the desired body weight, and someone whose weight are 300 lb more than the desired body weight. The graphs I created based on wdiff is very difficult the tell the spread, shape and center of the data. That is why I subset the dataset to include all the data that fall into (-100, 200) range. I got much better graph. The center of the data set is about 10. It only has one peak which is between (-5, 0), so it is a unimodel shape. The distribution is right skewed, according to the histogram. The figure tell us most people feel they are overweight by 10lb or so. Very few people feel they are overweight by more than 100lbs.

summary(wdiff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -500.00    0.00   10.00   14.59   21.00  300.00
```

```
boxplot(wdiff)
```



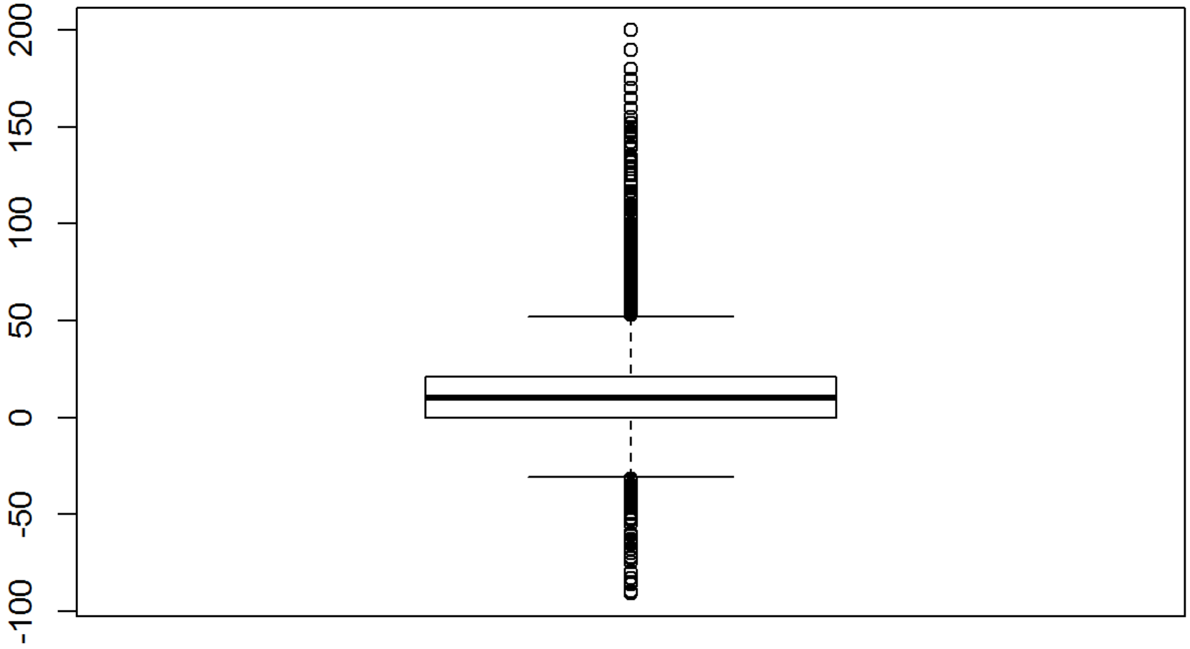
```
hist(wdiff, breaks = 100)
```



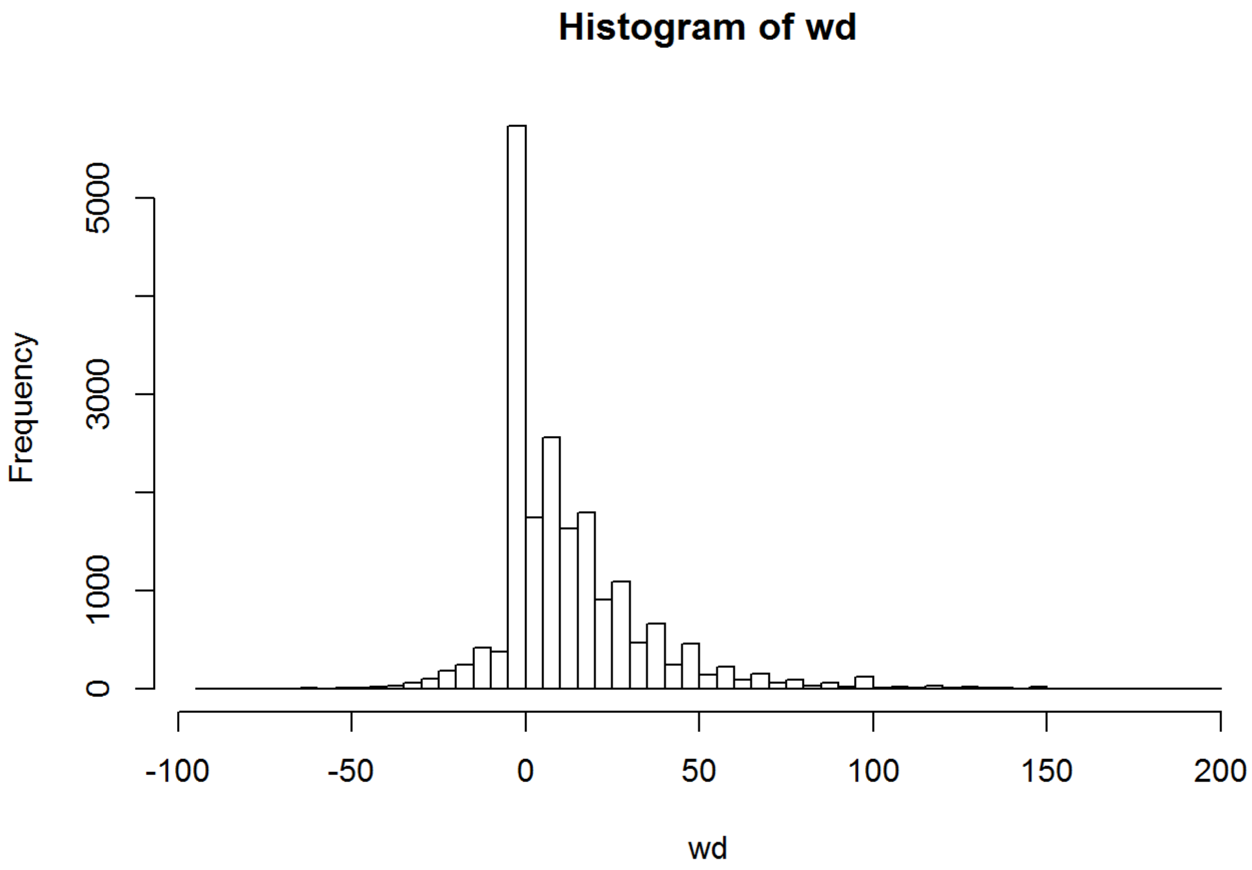
```
wd <- subset(wdiff, wdiff>= -100 & wdiff <=200)

summary(wd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -91.00    0.00   10.00   14.57   21.00  200.00
```

```
boxplot(wd)
```

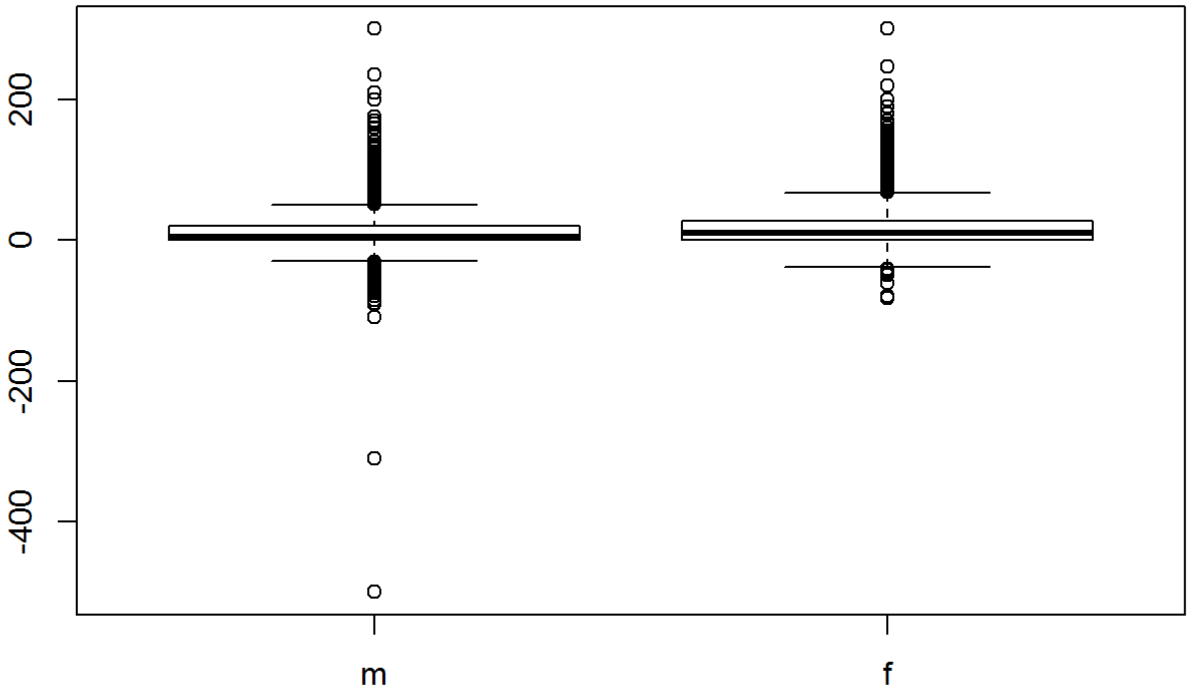


```
hist(wd, breaks = 100)
```



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

```
#based on this boxplots, we can visually see (even though it is difficult), the female population will perceive themselves t
o be more overweight compare to the male population. As we can see most of the inner box of the female boxplot lie above tha
t of the male boxplot.In addition to that, male population have more observations that have desired and actual weight differ
ence lower than the lower whisker line. As shown in the figure, the dots are much darker in the male boplot under below the
 lower whisker than those in the female boxplot.

boxplot(wdiff ~ cdc$gender)
```



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.

```
m <- mean(cdc$weight)
s <- sd(cdc$weight)

#the proportion is about 70.76%
sub <- subset(cdc, cdc$weight > (m-s) & cdc$weight <= (m+s))
dim(sub)/20000
```

```
## [1] 0.70760 0.00045
```