Lab - Intro to data

Tabito

01/23

Lab report

```
source("http://www.openintro.org/stat/data/cdc.R")
```

Load data:

```
tail(cdc)
```

Exercise 1:

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

There are 20,000 cases and 9 variables. We can see that general health,exerany, hlthplan, smoke100, and gender are categorical data, while height, weight, wtdesire, and age are numerical data.

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

Exercise 2:

```
##    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$weight)
```

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

```
190-140
```

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

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

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

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

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

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

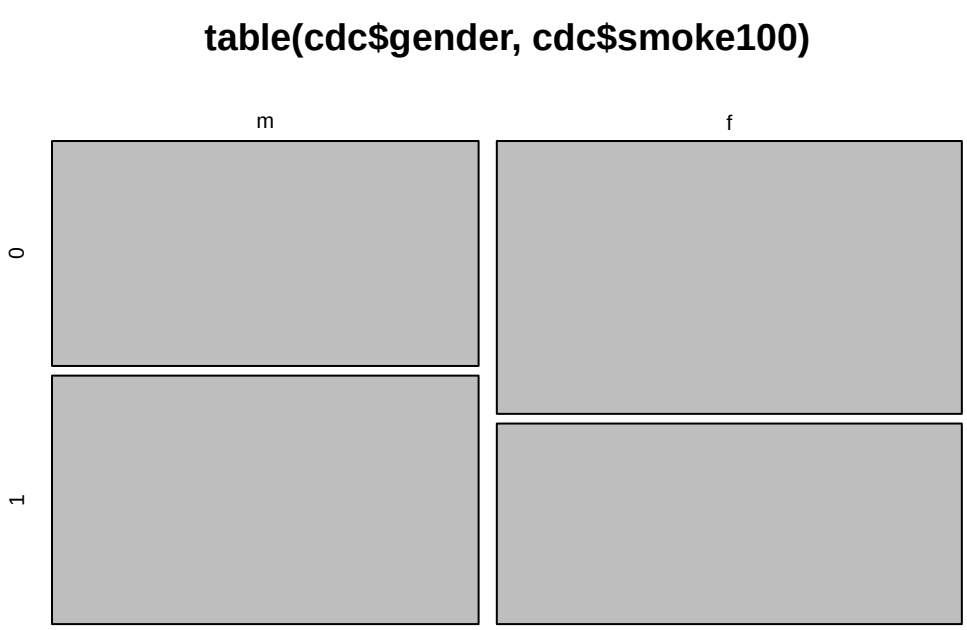
```
##
##      0      1
## 0.2543 0.7457
```

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

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

IQR for height = 6 IQR for weight = 50 9569 Males 23.285% of sample report excellent health

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

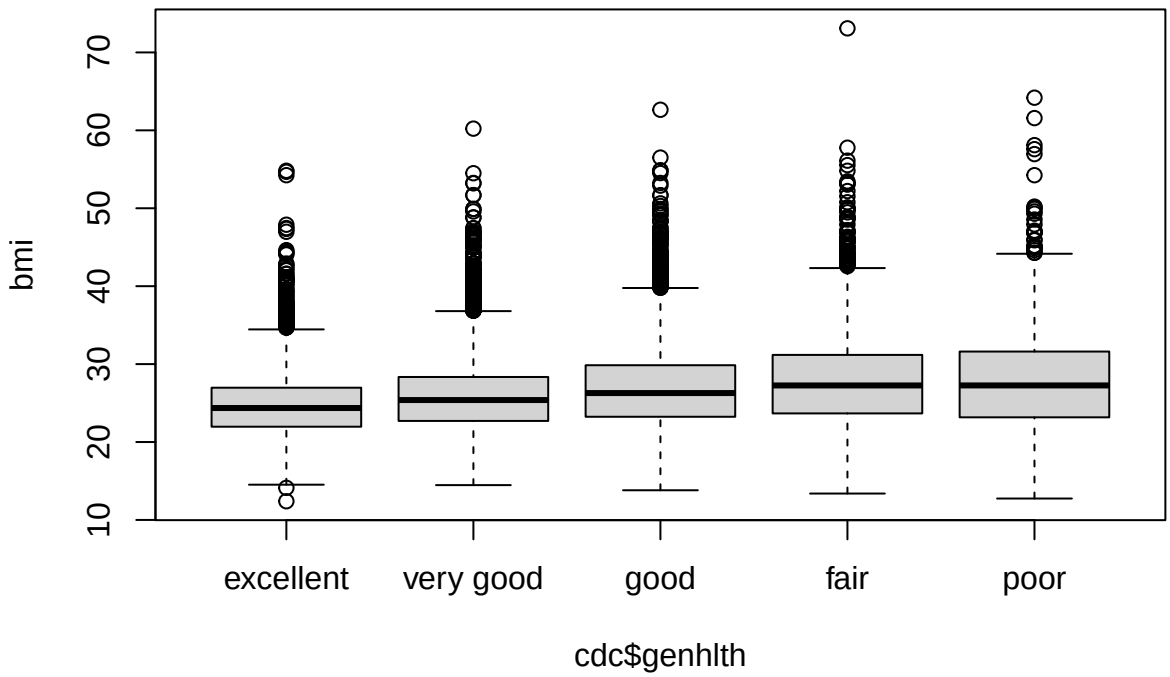


Exercise 3:
Males are slightly more likely to have smoked at least 100 cigarettes in their lifetimes. #### Exercise 4:

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

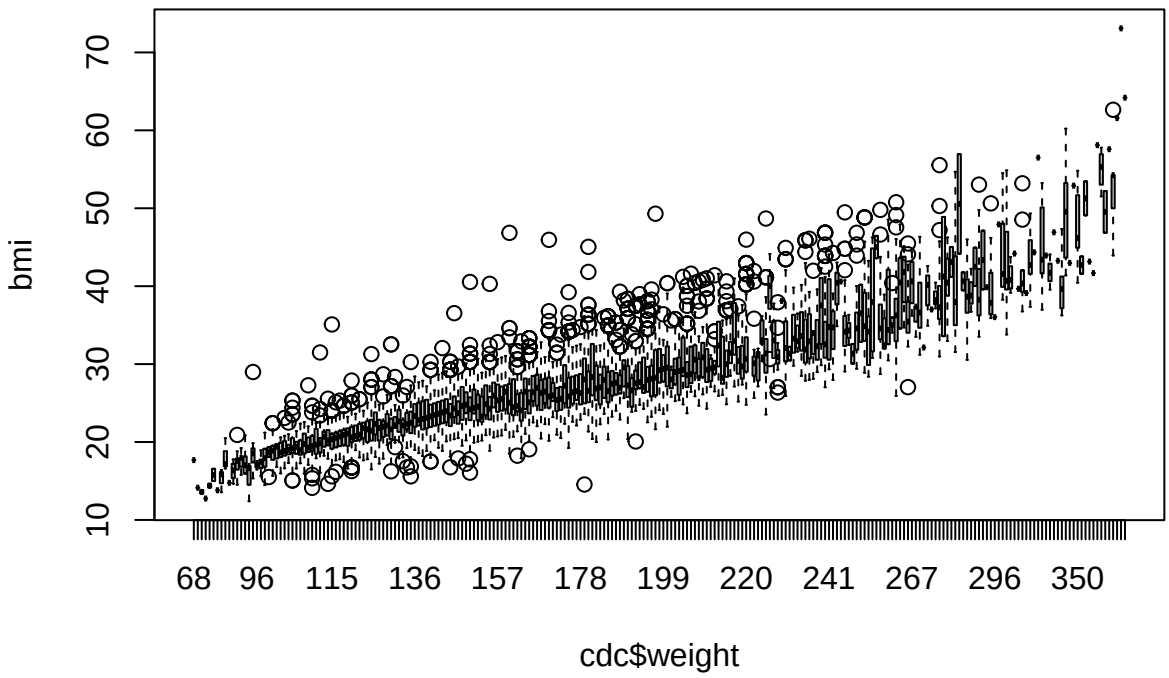
```
##       genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 13  excellent       1        0        1     66    185      220  21      m
## 37  very good       1        0        1     70    160      140  18      f
## 96  excellent       1        1        1     74    175      200  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
```

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



Exercise 5:

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



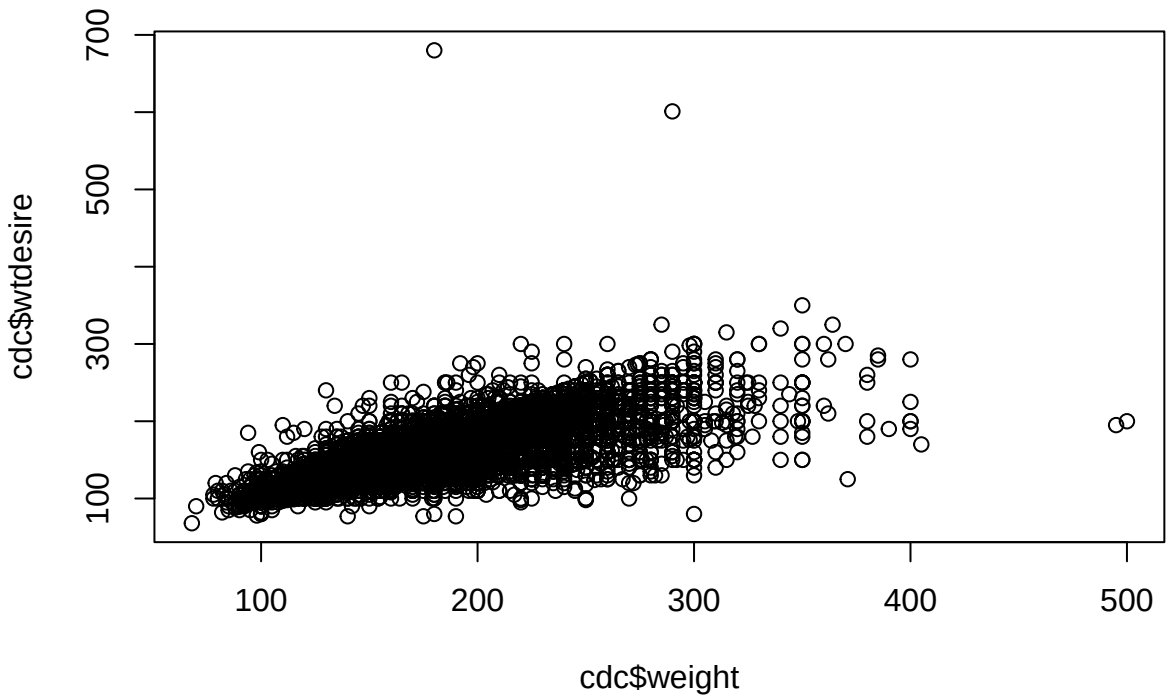
Shows that the median of bmi gets lower as general health becomes better.

For the second plot I chose weight, as I believe it would have a relationship with BMI as weight is a part of the formula when solving for it. Median of bmi gradually declines as weight declines as well.

On your own:

1: People generally want to stay around the same weight, if not lower.

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



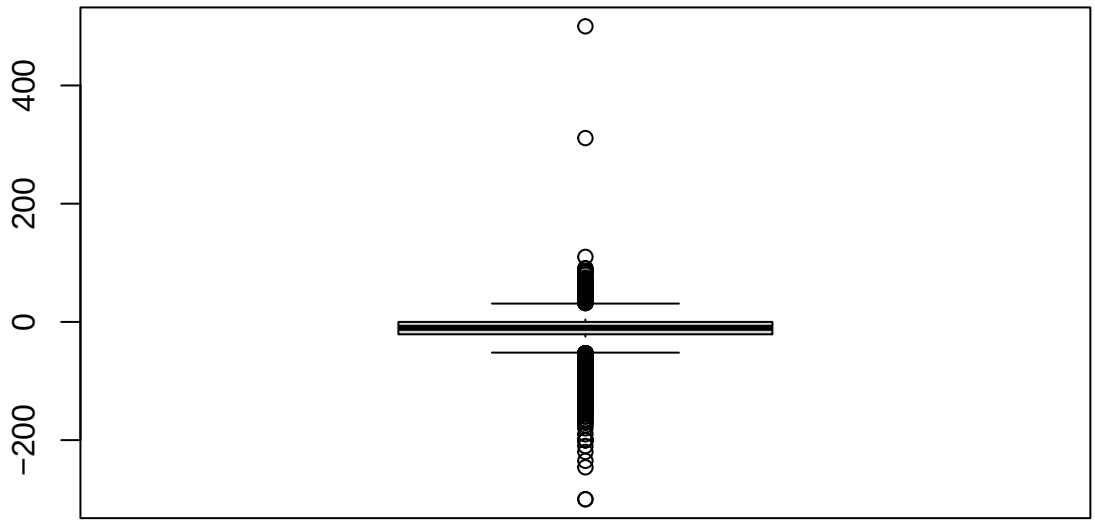
2:

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

3: wdiff is a numerical data. If an observation is 0, the person desires zero change in weight. If positive, that person wants to gain weight, if negative, they want to lose weight.

4: Since mean and median are negative, we can see that people want to lose weight on average.

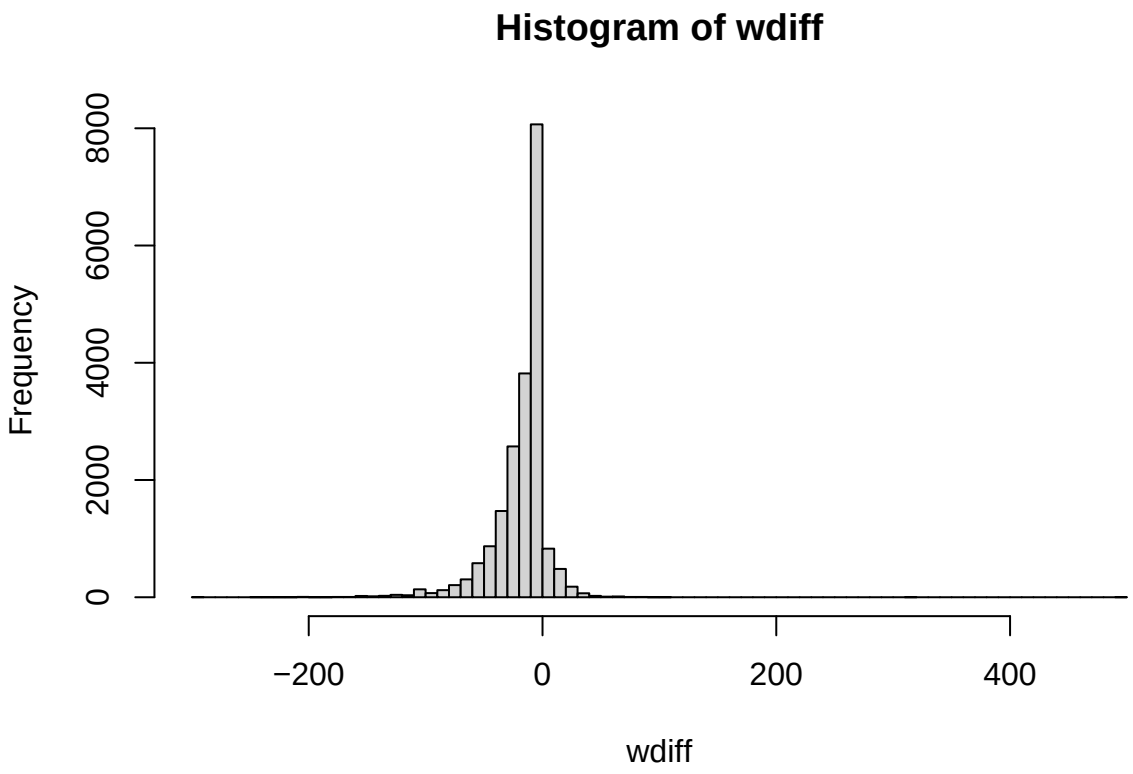
```
boxplot(wdiff)
```




```
summary(wdiff)
```

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

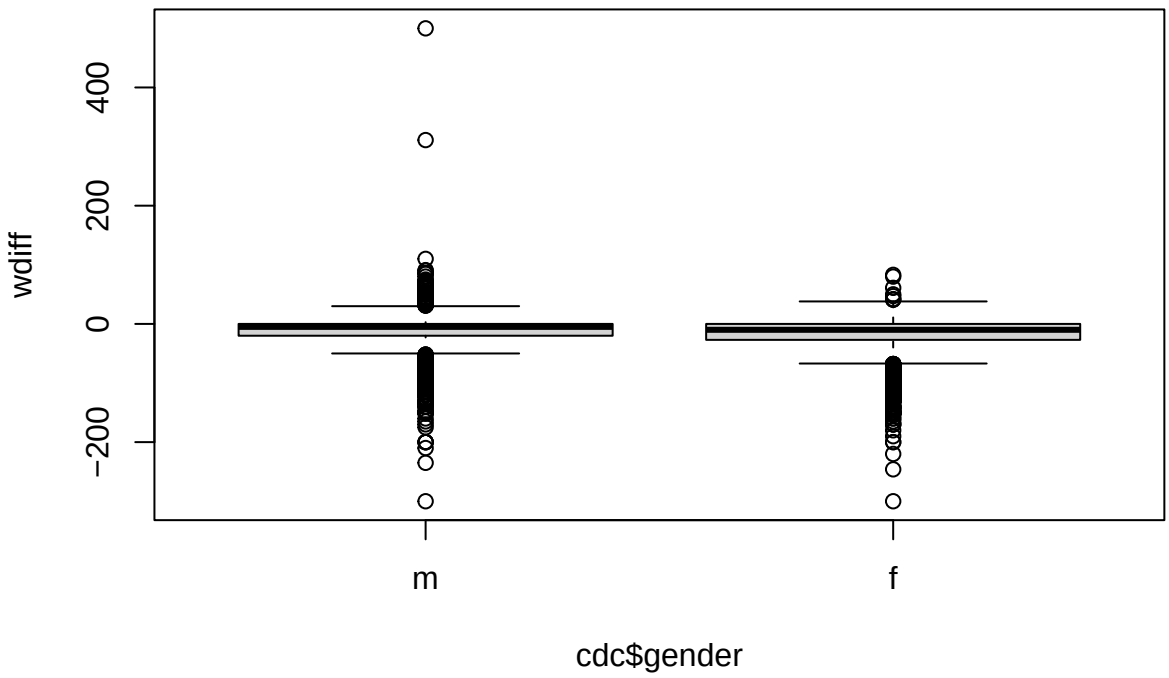
```
hist(wdiff, breaks = 100)
```



5:

Overall women desire a lager weightloss then men.

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




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

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

```
fdata <- subset(wdiff, cdc$gender == "f")
summary(fdata)
```

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

6: 70.76% of the weights are within one standard deviation.

```
mw <- mean(cdc$weight)
sdw <- sd(cdc$weight)
lower <- mw - sdw
upper <- mw + sdw
total_length <- length(cdc$weight)
within_sd <- length(subset(cdc$weight, cdc$weight >= lower & cdc$weight <= upper))

prop_of_weight_within_sd <- within_sd/total_length
```

Teamwork report

Team member	Attendance	Author	Contribution %
Name of member 1	Yes / No	Yes / No	25%
Name of member 2	Yes / No	Yes / No	25%
Name of member 3	Yes / No	Yes / No	25%
Name of member 4	Yes / No	Yes / No	25%
Total			100%