Grando-1

John Grando

January 31, 2017

Lab 1 Assignment

First, let's set the working directory and source the data.

```
setwd("~/Documents/Masters/DATA606/Week1/Lab/Lab1")
source("more/cdc.R")
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(plyr)
```

```
## Loading required package: plyr
```

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:

The number of cases is the number of rows in the data frame:

```
nrow(cdc)
```

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

The number of variables is the number of columns:

```
ncol(cdc)
```

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

Data Type:

genhlth - categorical, ordinal

exerany - categorical, nominal. Could be ordinal if it were viewed as "less than once in the last month" and "more than once in the last month"

hlthplan - categorical, nominal

smoke100 - categorigal, nominal. Could be ordinal if it were viewed as "less than 100 cigarettes" and "more than 100 cigarettes" options

height - numeric, discrete

weight - numeric, discrete

wtdesire - numeric, discrete

age - numeric, discrete

gender - categorical, nominal

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?

Answer:

Numerical 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
```

Numerical 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
```

The interquartile range for height is 70 - 64 = 6

```
h <- summary(cdc$height)
h["3rd Qu."] - h["1st Qu."]
```

```
## 3rd Qu.
##       6
```

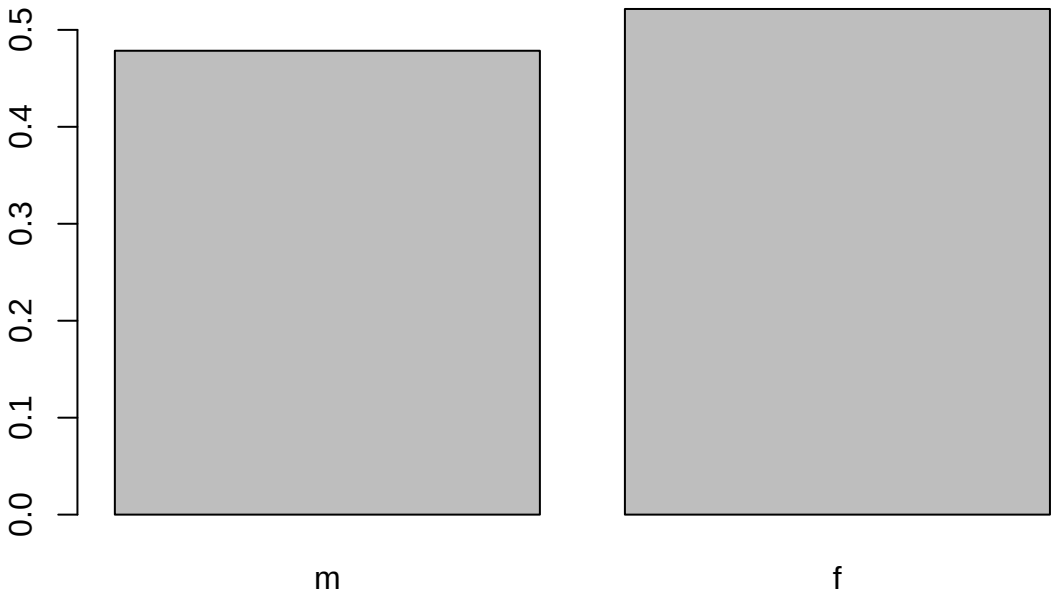
The interquartile range for age is 57 - 31 = 26

```
a <- summary(cdc$age)
a["3rd Qu."] - a["1st Qu."]
```

```
## 3rd Qu.
##      26
```

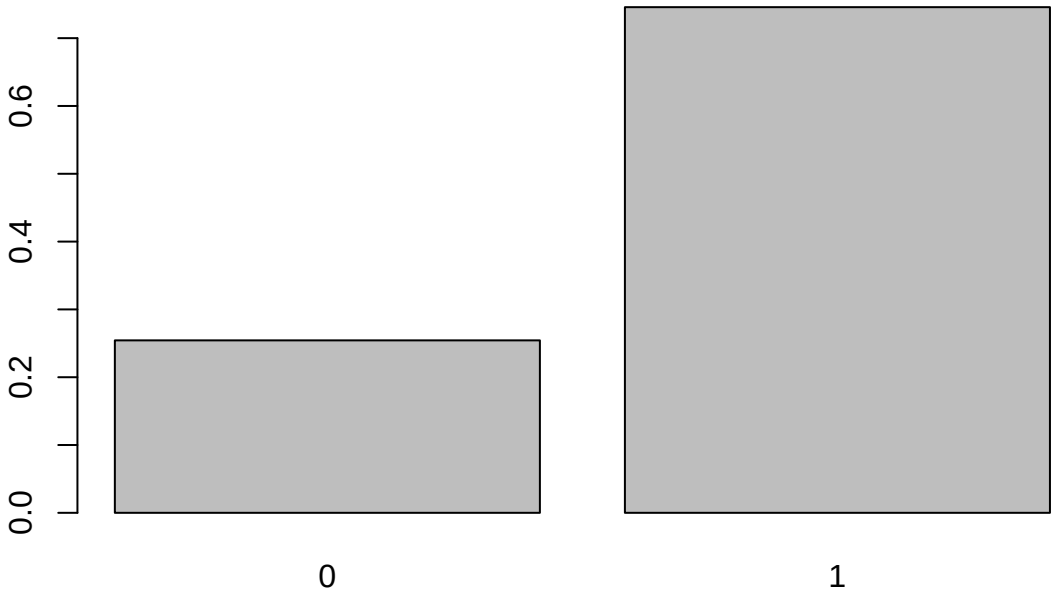
Relative frequency distribution for gender

```
barplot(table(cdc$gender)/nrow(cdc))
```



Relative frequency distribution for exerany

```
barplot(table(cdc$exerany)/nrow(cdc))
```



Males in the sample

```
nrow(subset(cdc, gender == "m"))
```

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

Proportion of sample in excellent health

```
nrow(subset(cdc, genhlth == "excellent"))/nrow(cdc)
```

```
## [1] 0.23285
```

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

Answer:

It appears that there may be a higher proportion of men who smoke more than 100 cigarettes in their lifetime than women.

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:

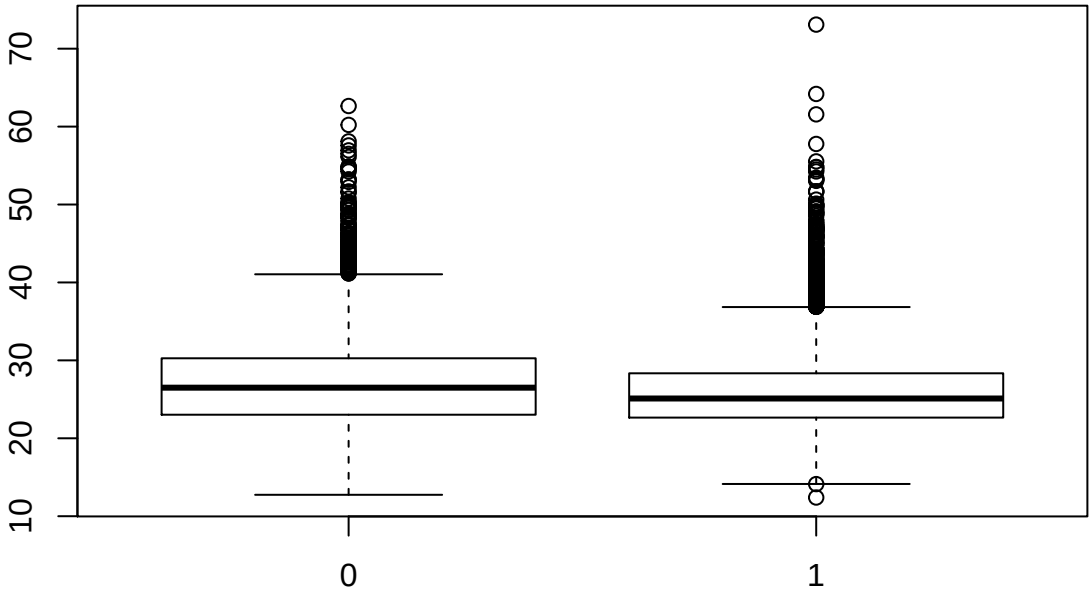
```
under23_and_smoke <- subset(cdc, cdc$age < 23 & cdc$smoke100 ==
    1)
```

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.

Answer:

It shows that the median of bmi is slightly lower and the variability slightly smaller for those who report excellent health. Although these values slightly increase (median and variability) as reported health gets poorer, it does not appear to be by much. This indicates that self-assessment of general health may not have a strong association to bmi.

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

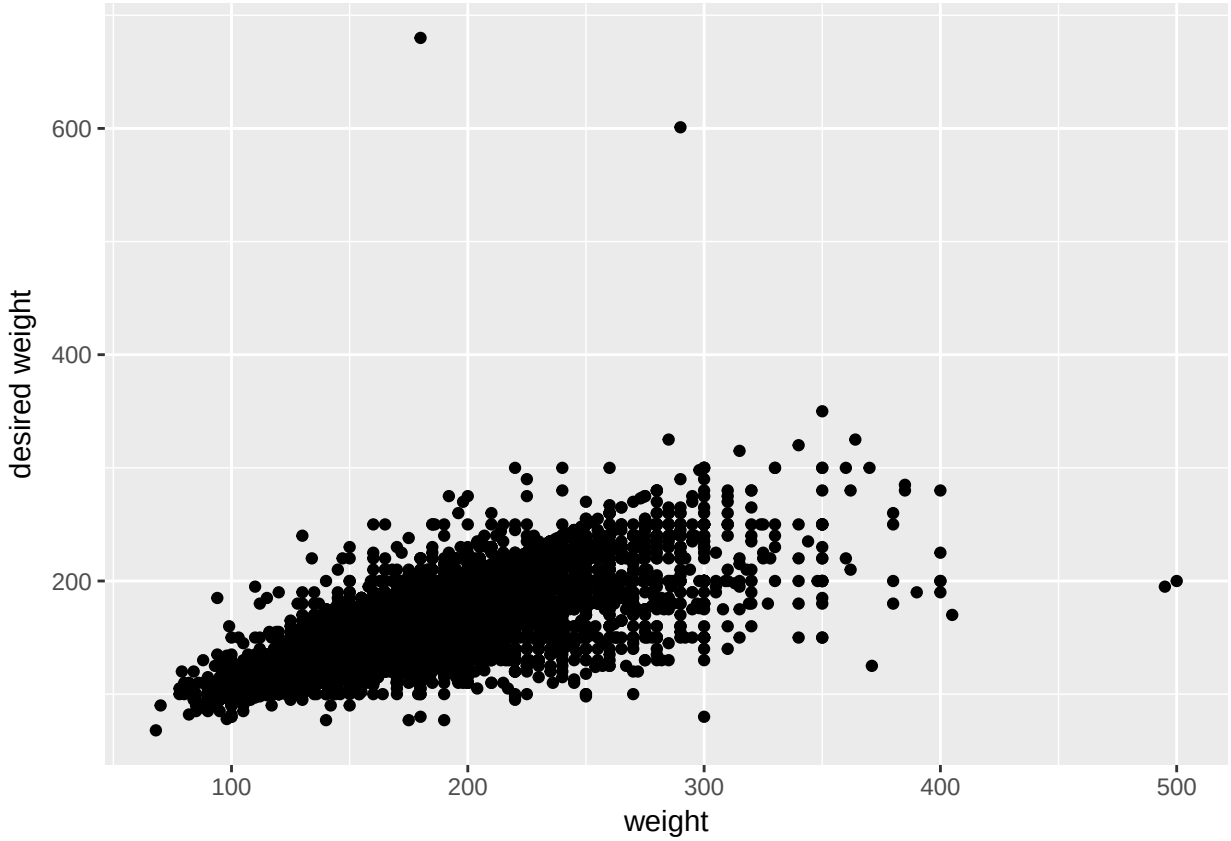


I chose exerany as my categorical variable because it seemed that a person who hasn't exercised within the last month, doesn't do it often, and would have a higher weight. My assumption is that height is independent with exerany. It appears that the median is slightly lower for those individuals who have worked out and the variability is lower. However, note that the outliers are just as extreme (actually even more) than the non exercising group.

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

Now I will make the scatter plot.

```
ggplot(cdc, aes(y = wtdesire, x = weight)) + geom_point() + labs(y = "desired weight")
```



It appears that there is a positive linear association between weight and desired weight.

Question 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.

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

Question 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?

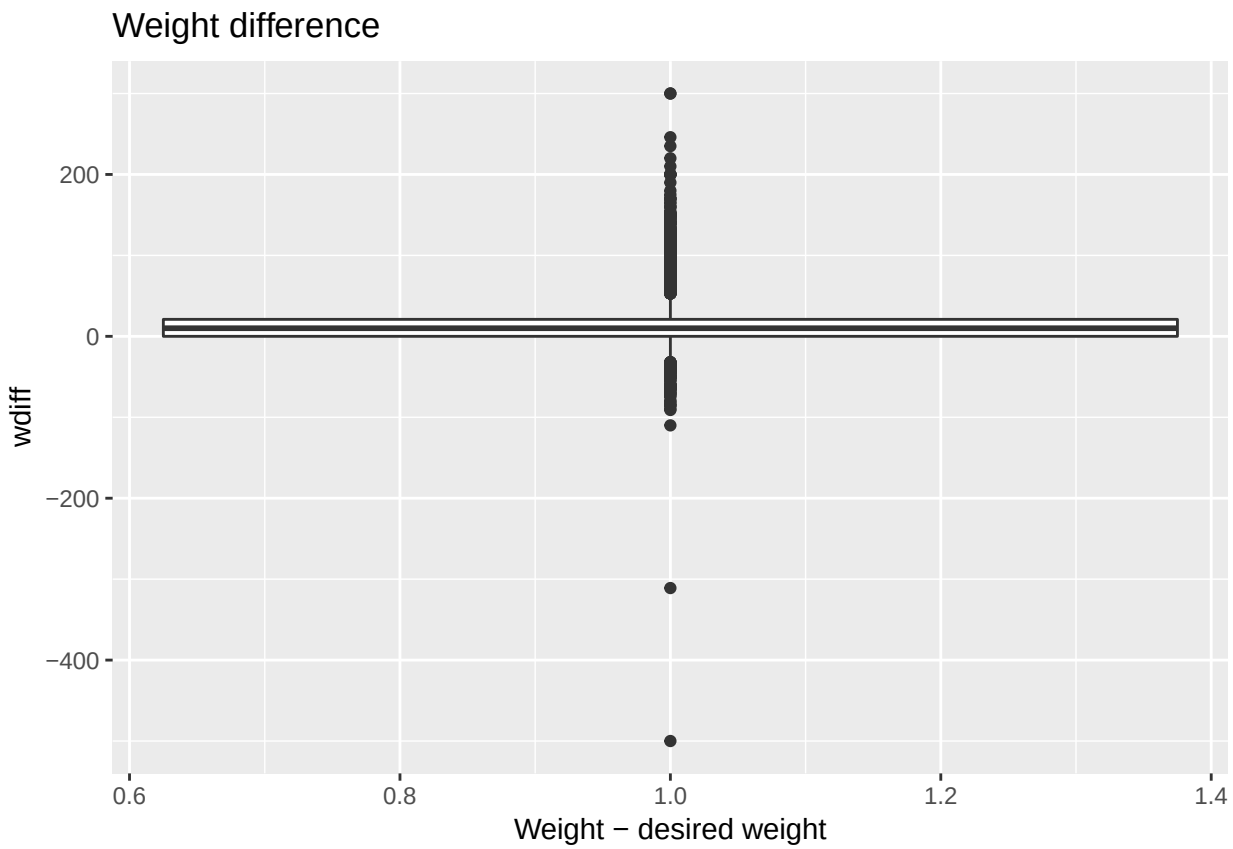
wdiff is numerical data because performing arithmetic operations on this data would still have meaning. If an observation of wdiff is 0, then that indicates a person is at their desired weight. Since I have subtracted desired weight from weight, a positive value would indicate that a person is heavier than they wish to be. A negative value would indicate a person is lighter than they wish to be.

Question 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?

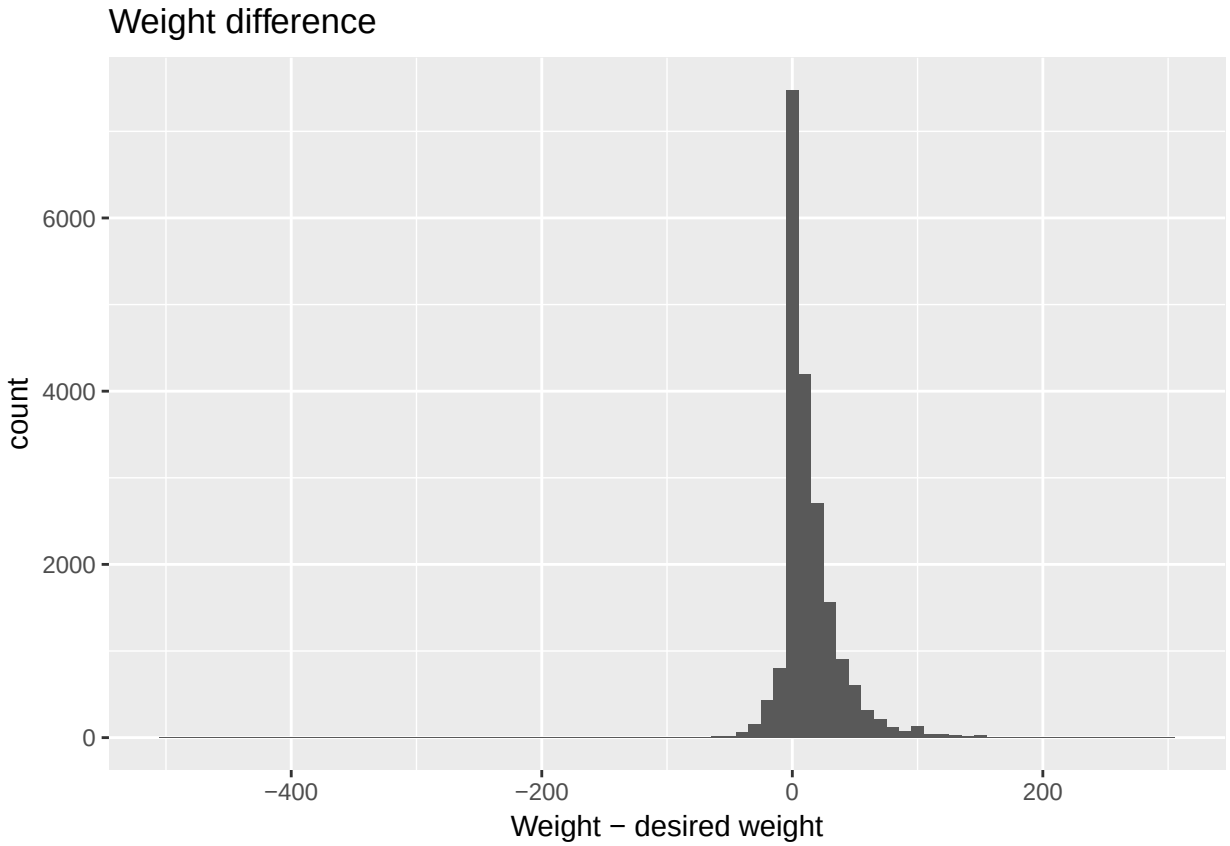
```
summary(cdc$wdiff)
```

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

```
ggplot(cdc, aes(y = wdiff, x = 1)) + geom_boxplot() + ggtitle("Weight difference") +
    labs(x = "Weight - desired weight")
```




```
ggplot(cdc, aes(x = wdiff)) + geom_histogram(binwidth = 10) +
    ggtitle("Weight difference") + labs(x = "Weight - desired weight")
```



there are some pretty extreme outlying negative values (person who is lighter than desired) which appear to be errors. Let's check the head and tail of the data.

```
head(arrange(cdc, wdiff))
```

```
##     genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 1      good       0        1        0     69    180      680  24      m
## 2 very good       1        1        1     73    290      601  56      m
## 3      good       0        1        0     69    130      240  26      m
## 4 excellent       1        1        1     73     94      185  52      m
## 5      good       1        0        1     74    160      250  20      m
## 6      poor       0        1        1     75    134      220  71      m
##   wdiff
## 1  -500
## 2  -311
## 3  -110
## 4   -91
## 5   -90
## 6   -86
```

```
tail(arrange(cdc, wdiff))
```

```
##       genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 19995    poor       1        1        1     72    400      190  50      m
## 19996    fair       1        0        0     65    300       80  45      f
## 19997    poor       1        1        0     68    405      170  32      m
## 19998    poor       0        1        1     67    371      125  35      f
## 19999    poor       1        1        0     74    500      200  45      m
```

```
## 20000    fair       1        1        1     69    495      195  32      f
##       wdiff
## 19995   210
## 19996   220
## 19997   235
## 19998   246
## 19999   300
## 20000   300
```

It looks like the tail of wdiff (positive vlaues) is pretty valid but the head is extreme and nonsensical. I doesn't stand to reason that a person would want to gain 110, 311, or 500. Therefore, I would suggest using median and IQR to determine the average expected weight difference and variability.

```
summary(cdc$wdiff)
```

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

While there are some extreme outliers, the overall data seems to have a right skew. Additionally, givent the median and IQR, it appears the data is showing us that people generally indicate their ideal weight to be less than their current weight.

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

First, i'll maket two new data frames, one for men and one for women.

```
cdcm <- subset(cdc, gender == "m")
cdcf <- subset(cdc, gender == "f")
```

next, i'll print the summary statistics for these

```
summary(cdcm$wdiff)
```

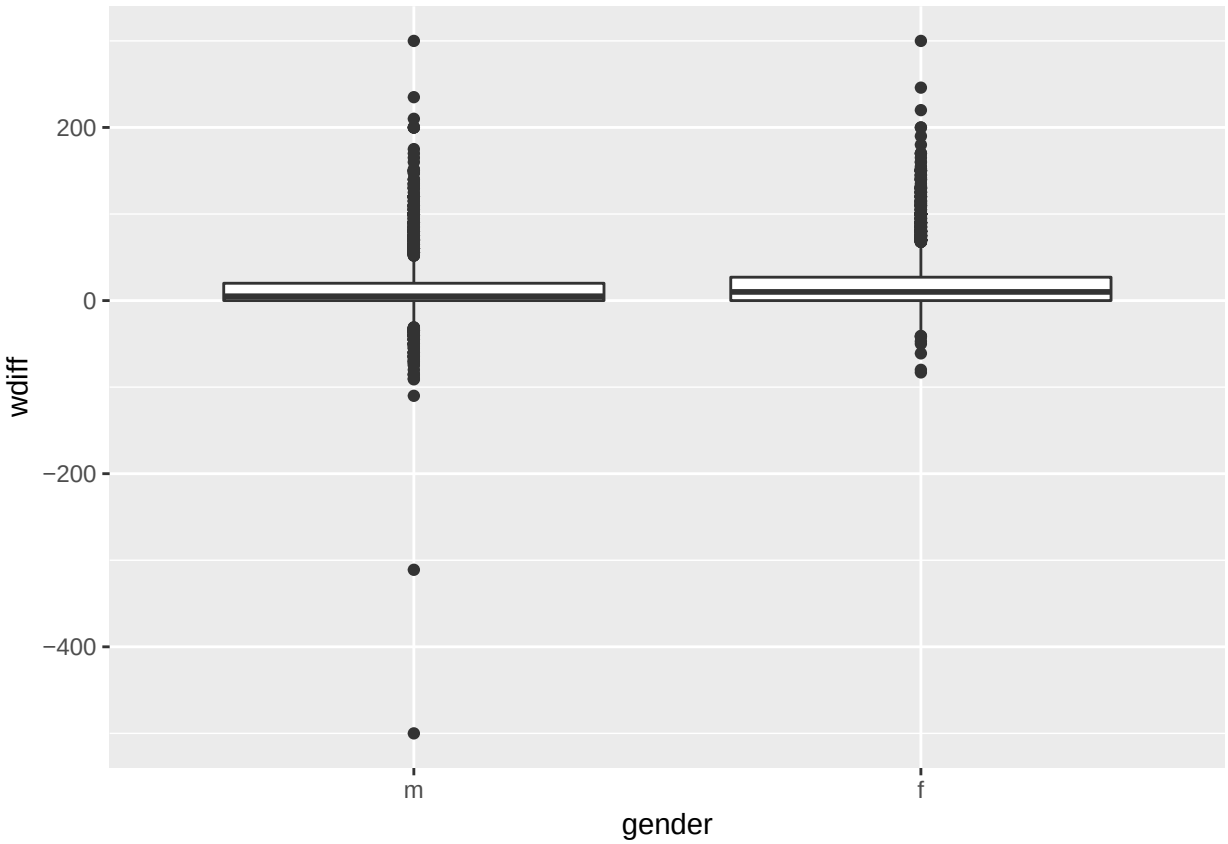
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -500.00    0.00    5.00   10.71   20.00  300.00
```

```
summary(cdcf$wdiff)
```

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

Then I will print out the boxplot

```
ggplot(cdc, aes(y = wdiff, x = gender)) + geom_boxplot()
```



It's hard to tell anything from the boxplots themselves due to the large difference for some samples; however, the summary statistics show that women tend to have a higher median wdiff which means that men tend to view their weight as closer to their desireable weight than women. Additionally, the IQR for men is smaller so that inidcates there is less variability in their responses.

Question 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 of weight

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

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

Standard deviation of weight

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

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

Proportion of weigths with one standard deviation of the mean

```
sum(mean(cdc$weight) + sd(cdc$weight) > cdc$weight & mean(cdc$weight) -
    sd(cdc$weight) < cdc$weight)/nrow(cdc)
```

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

70.76% of the weights in the data set are within one standard deviation of the mean.