Lab 1: Intro to data

Key

Exercises:

Exercise 1: There are 20,000 cases and 9 variables in this data set. `genhlth` is ordinal categorical; `exerany`, `hlthplan`, and `smoke100` are categorical; `height`, `weight`, `wtdesire`, and `age` are continuous, though they've been measured discretely; `gender` is categorical.

```
cdc %>%
  group_by(exerany) %>%
  summarize(median_wt = median(weight))
```

Exercise 2:

```
## Source: local data frame [2 x 2]
##
##   exerany median_wt
##     (lgl)     (dbl)
## 1   FALSE       165
## 2    TRUE       165
```

The median weight of subjects who exercised in the last month and those who did not were the same: 165 lb.

```
cdc %>%
  select(gender) %>%
  table()
```

Exercise 3:

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

```
cdc %>%
  select(gender) %>%
  table()/20000
```

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

```r
cdc %>%
  select(genhlth) %>%
  table()/20000
```

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

There are 9569 males in the sample. 23.3% of the sample reports being in excellent health. This exercise was poorly worded, so you may have also looked at the joint distribution of these two variables, which is fine.

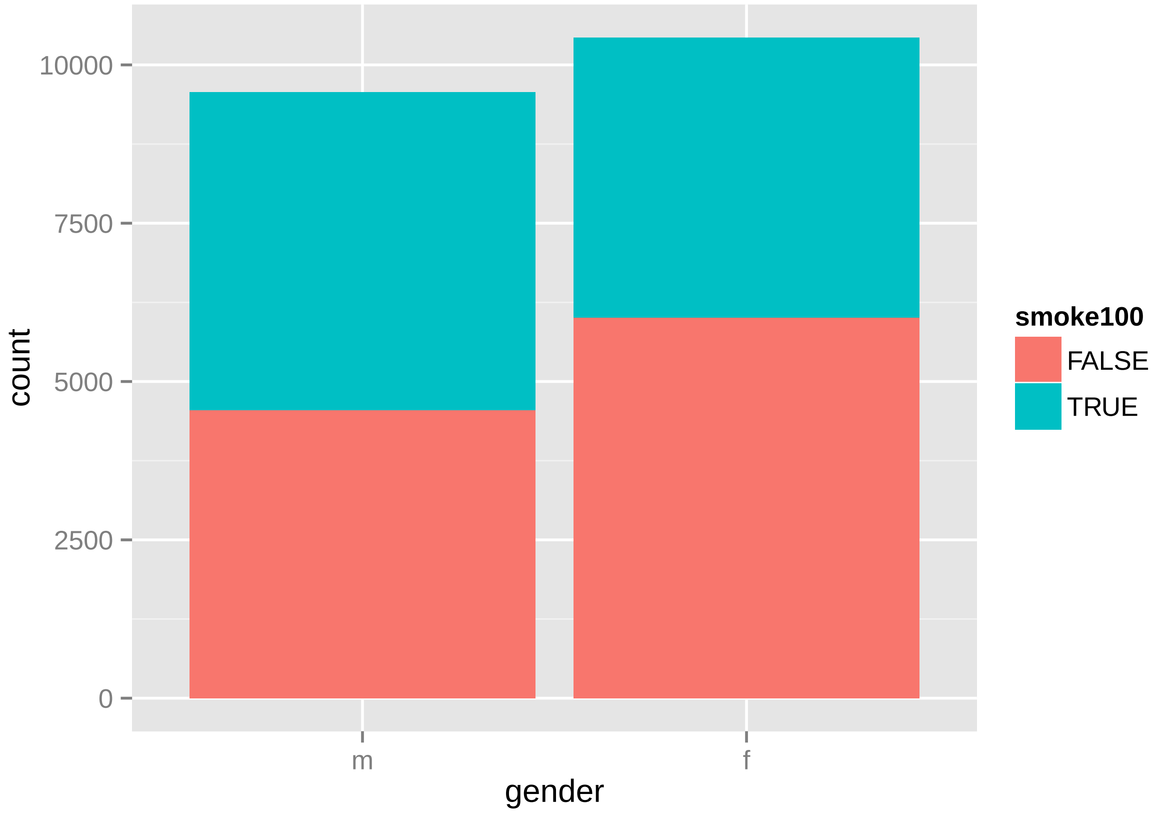
```r
cdc %>%
  select(gender, smoke100) %>%
  table()
```

Exercise 4:

```
##       smoke100
## gender FALSE TRUE
##      m  4547 5022
##      f  6012 4419
```

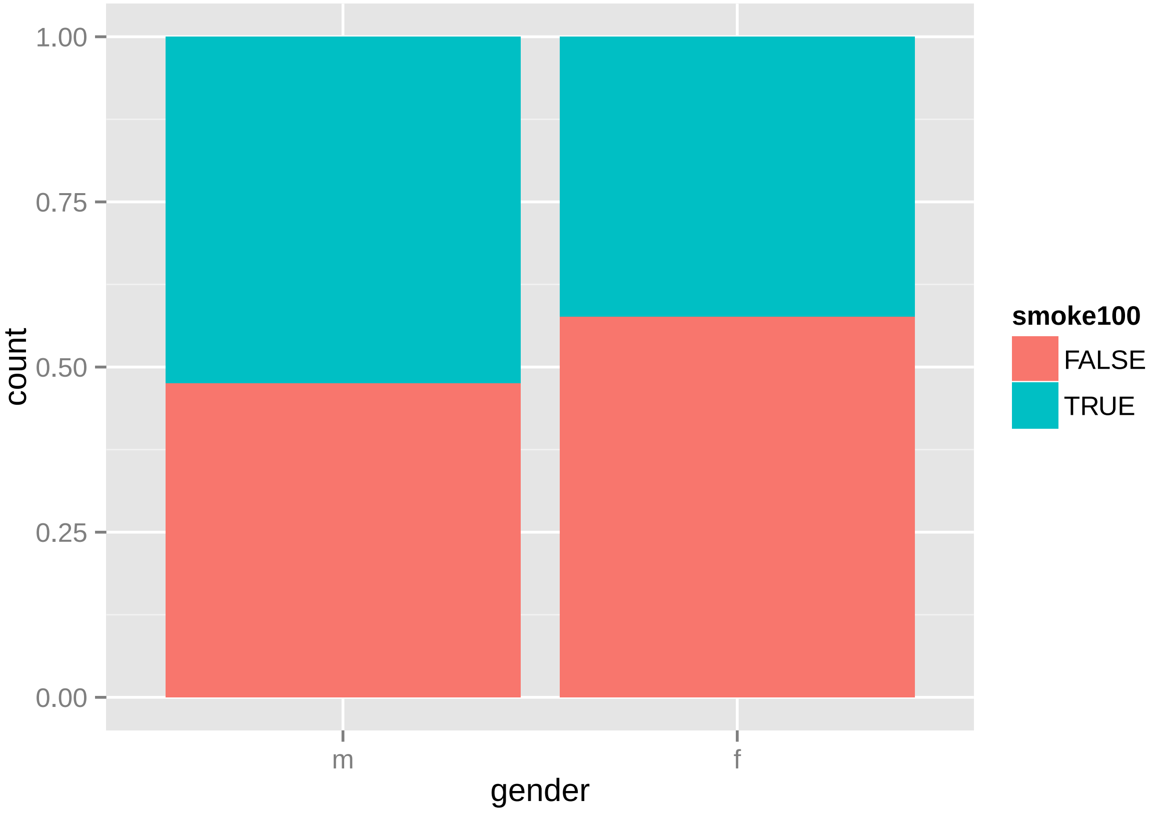
1 percent of men are smokers as are 6012 percent of women. 0 percent of the sample are male smokers while 0 percent of the sample are female smokers. The first pair of statistics is more useful in determining if men or women are more likely to be smokers since it controls for the possibility that there are different numbers of men and women in the sample.

```r
qplot(x = gender, fill = smoke100, data = cdc)
```



Exercise 5:

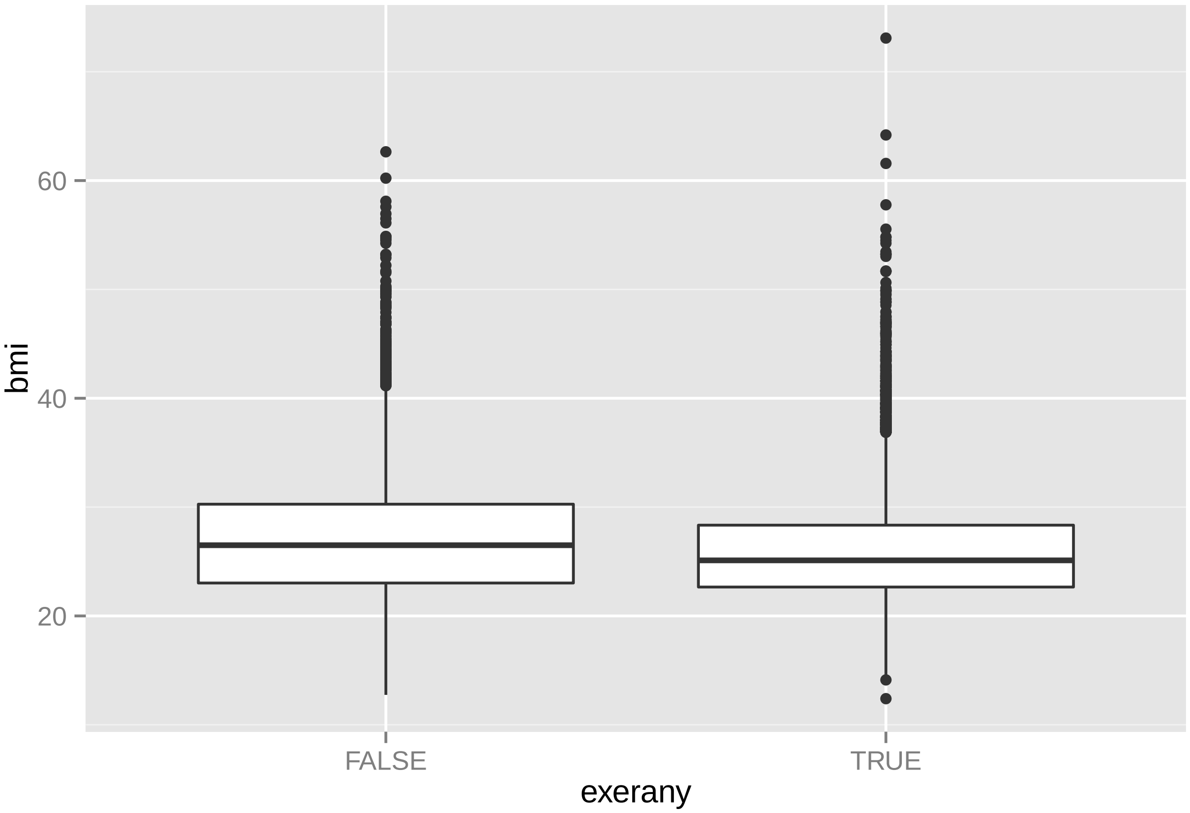
```
qplot(x = gender, fill = smoke100, data = cdc, position = "fill")
```



The first plot shows only the counts of smokers/non-smokers in each gender while the second plot controls for the number of males and females and shows the proportion of smokers/non-smokers in each group. Similar to the reasoning in exercise 4, this second plot gives better insight into the relative proportion of smokers in the two groups.

```
cdc <- cdc %>%
  mutate(bmi = (weight / height^2) * 703)

qplot(y = bmi, x = exerany, data = cdc, geom = "boxplot")
```



Exercise 6:

Body mass index is likely effected by the level of physical activity so I chose to compare `bmi` to the `exerany` variable. The distribution of BMI in both groups is right skewed with many outliers. The exercisers had a lower median and Q3 BMI as well as a slightly lower Q1 BMI. This agrees with the intuition that those that exercise more will have a lower BMI, on average. There is a notable outlier, though: the highest BMI belongs to an exerciser.

```
under23_and_smoke <- cdc %>%
  filter(smoke100 == TRUE & age < 23)
dim(under23_and_smoke)
```

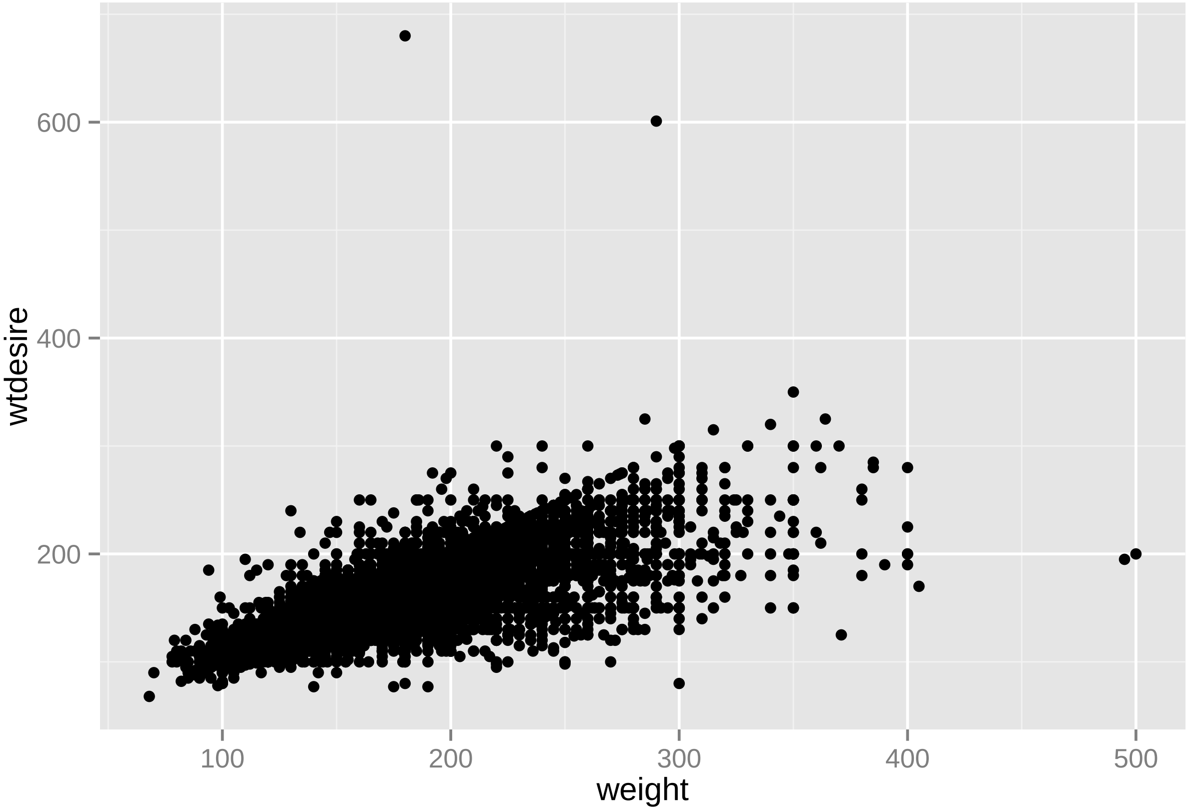
Exercise 7:

```
## [1] 620  10
```

There are 620 subjects that fit the criteria of being smokers and under the age of 23.

On your own:

```
qplot(x = weight, y = wtdesire, data = cdc, geom = "point")
```



1:

There is a positive and roughly linear association between subjects' weight and their desired weight. Notably, most of the points fall below the identity line, suggesting most people desire a weight that is lower than their recorded weight.

```
cdc <- cdc %>%
  mutate(wdiff = wtdesire - weight)
```

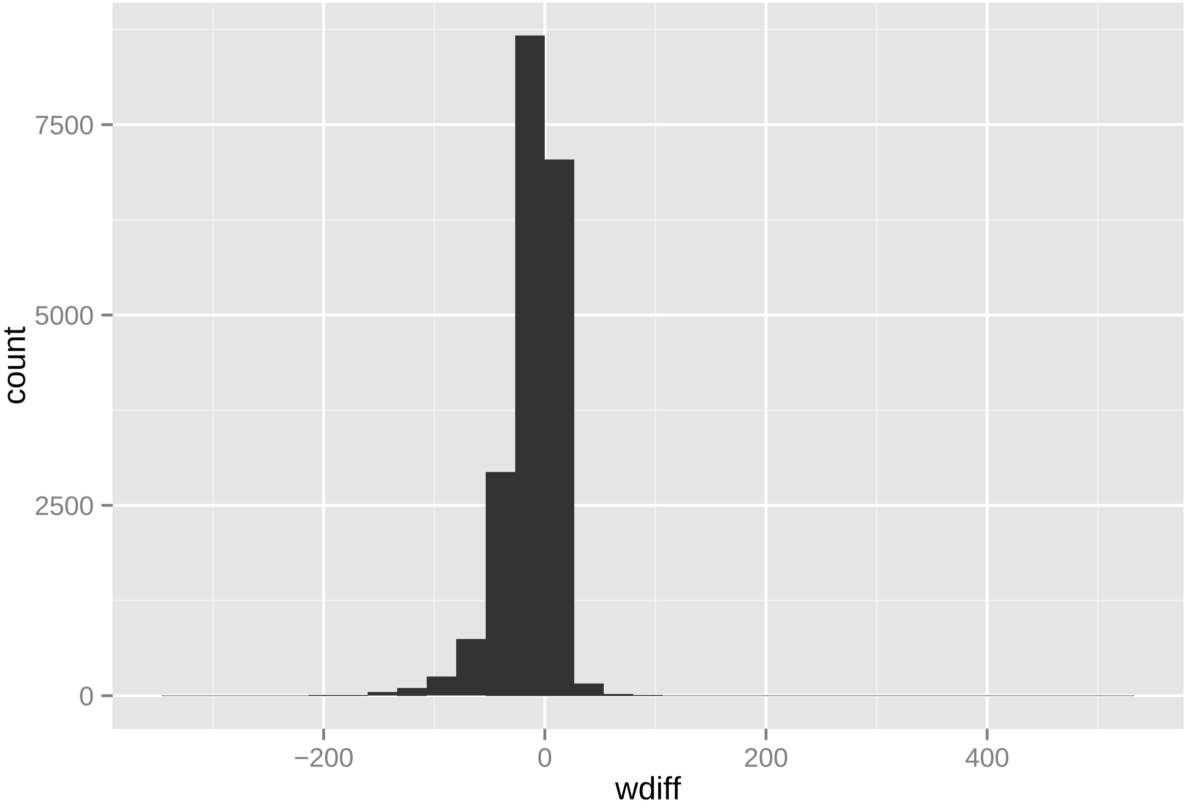
2:

3: `wdiff` is a continuous variable that has been measured discretely. A positive `wdiff` means that person wishes to gain weight; a negative `wdiff` means that person wishes to lose weight; and a `wdiff` of zero indicates that person is happy with their current weight.

```r
qplot(x = wdiff, data = cdc, geom = "histogram")
```

4:

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

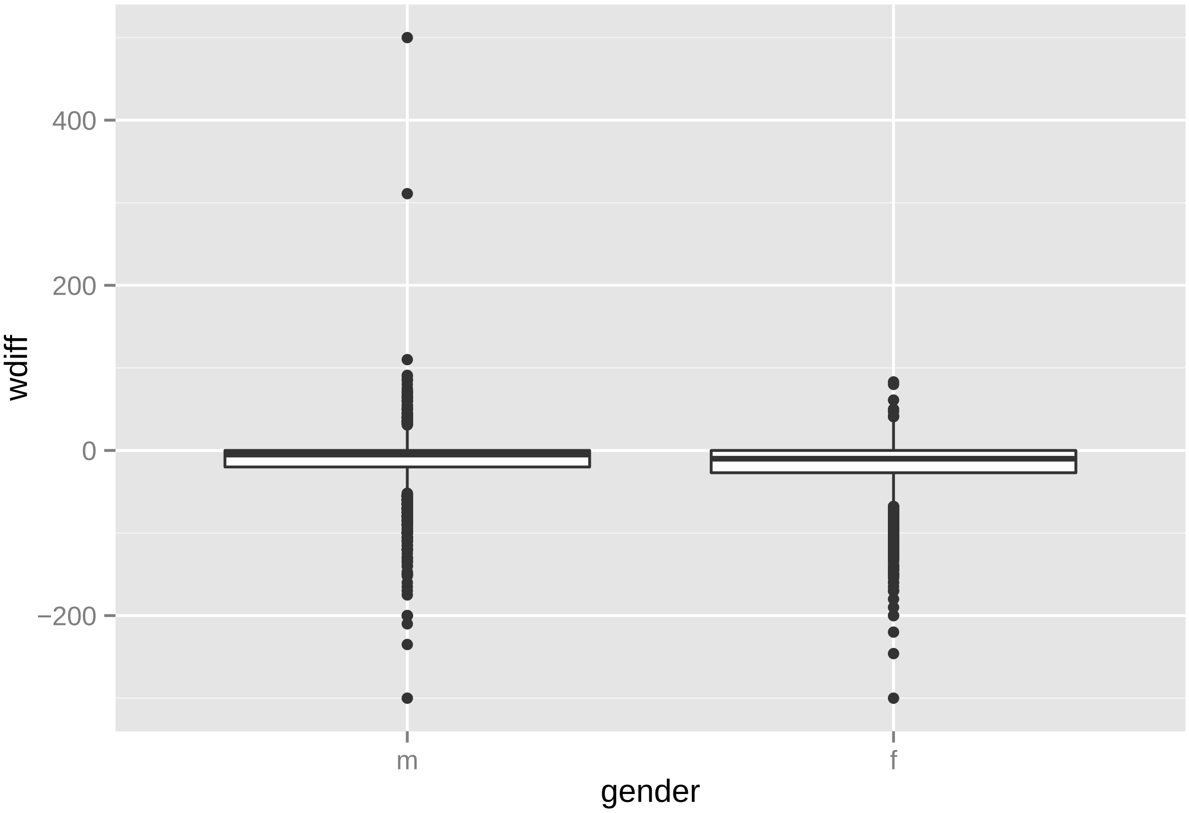


```r
summary(cdc$wdiff)
```

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

The `wdiff` data is centered at a median of -10 lb, meaning that the majority of people wish to lose weight. The IQR is 21 lb, though there is a low outlier at -300 lb and a high outlier at 500 lb. Excluding the outliers, the data appears to have a slight left skew.

```r
qplot(x = gender, y = wdiff, data = cdc, geom = "boxplot")
```



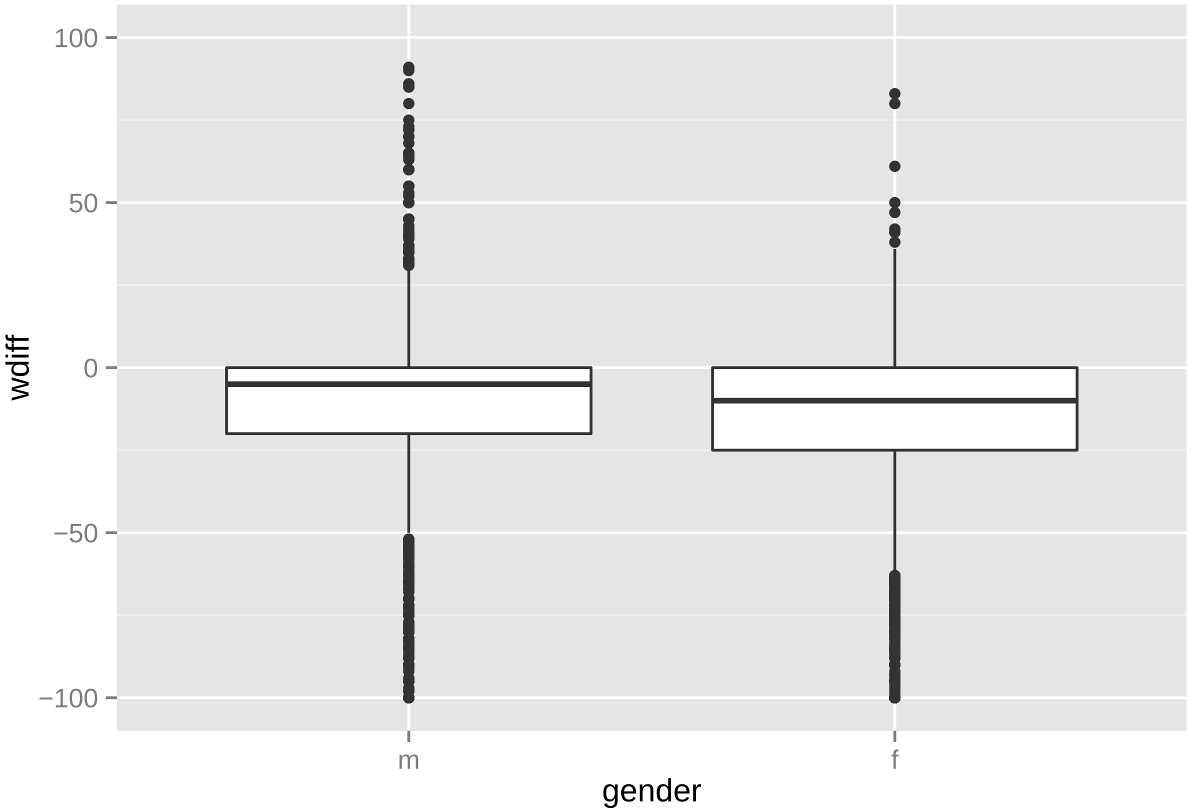
5:

The outliers dominate the range of the plot, so it's difficult to compare the two groups. Looking closely, women have a lower Q1 and median `wdiff` than do men, suggesting more women wish to lose more weight than men.

Note: these plots can be improve by restricting the limits of the y-axis using the following additional argument.

```
qplot(x = gender, y = wdiff, data = cdc, geom = "boxplot", ylim = c(-100, 100))
```

```
## Warning: Removed 184 rows containing non-finite values (stat_boxplot).
```



```
within_one_sd <- cdc %>%
  select(weight) %>%
  filter(weight < (mean(weight) + sd(weight)) &
           weight > (mean(weight) - sd(weight)))
dim(within_one_sd)
```

6:

```
## [1] 14152     1
```

0.7076 percent of the sample falls within one standard deviation of the mean weight.

FYI: you can create a single pipeline to create that proportion using something we haven't seen yet - the `ifelse` function.

```
cdc %>%
  mutate(within_one_sd = ifelse(weight < (mean(weight) + sd(weight)) &
                                  weight > (mean(weight) - sd(weight)),
                                "yes", "no")) %>%
  group_by(within_one_sd) %>%
  summarize(n = n()) %>%
  mutate(prop = n / sum(n))
```

```
## Source: local data frame [2 x 3]
```

```
##
##   within_one_sd     n   prop
##           (chr) (int)  (dbl)
## 1            no  5848 0.2924
## 2           yes 14152 0.7076
```