Name: Samuel Peoples

R Lab 1: Introduction to data

Please answer all the Exercises and the questions from the "On Your Own" section. If you use any graphs or charts to justify your answer, please include them.

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).

```
>view(cdc)
```

There are nine variables, of which there are 20,000 cases.

genhlth	Categorical, Ordinal
exerany	Categorical, Boolean Nominal
Hlthplan	Categorical, Boolean Nominal
smoke100	Categorical, Boolean Nominal
height	Numerical, Discrete
Weight	Numerical, Discrete
Wtdesire	Numerical, Discrete
Age	Numerical, Discrete
Gender	Categorical, Boolean 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?

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

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

IQR = 70-64=6

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

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

IQR = 57-31=26

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

```
    m     f
```

9569 10431
Relative frequency for males = 9569/(9569+10431)= .47845
Relative frequency for females = 10431/(9569+10431)= .52155

```
>table(cdc$exerany)
```

0 1
5086 14914
Relative frequency for "0" = 5086/(5086+14914)= .2543
Relative frequency for "1" = 14914/(5086+14914)= .7457

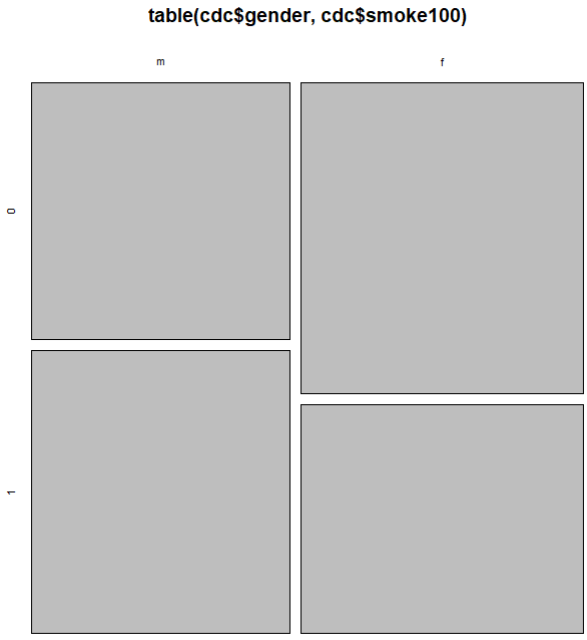
There are 9569 males.

```
>table(cdc$genhlth)
```

excellent very good good fair poor
4657 6972 5675 2019 677

There are 4657 people reporting "excellent" health.

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

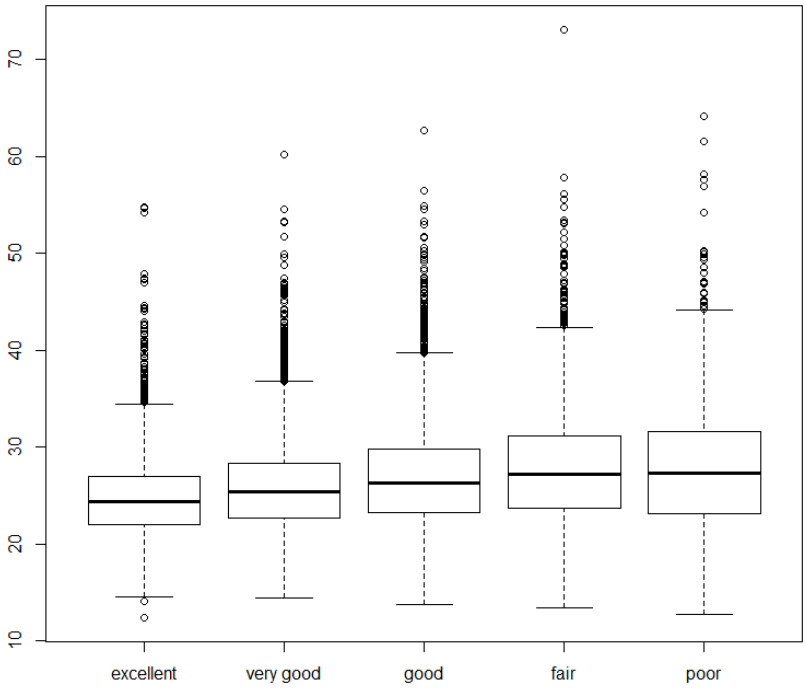


The mosaic plot reveals that more females than males report "0" for smoking, being more than half of the females, while more than half of males reported "1" for smoking.

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$smoke100 == "1" & age < 23)
```

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.



By entering the command:

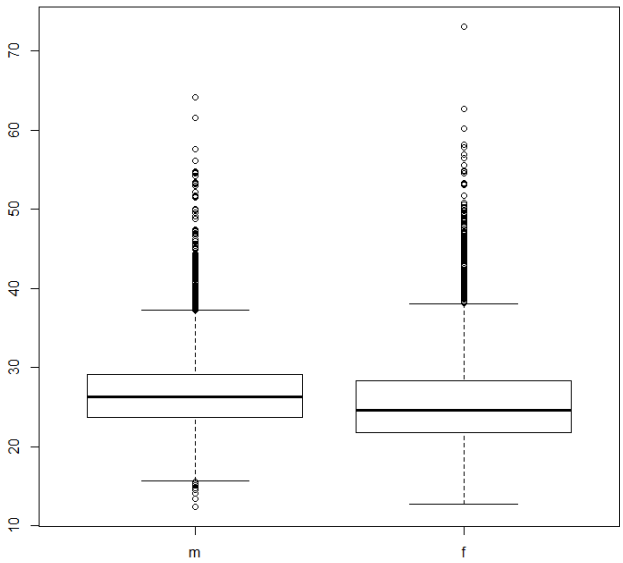
```
bmi <- (cdc$weight / cdc$height^2) * 703
boxplot(bmi ~ cdc$genhlth)
```

The new object bmi is defined by the BMI formula, $703 \cdot \frac{weight}{height^2}$, where weight is in pounds, and height is in inches. The boxplot then displays outliers as unfilled circles, horizontal lines noting the minimum and maximum BMI for each reported health condition. The boxes surround the 1st and 3rd quartiles, which are separated by a horizontal line denoting the median.

There seems to be a correlation between higher BMI and poor health, while having a lower BMI is less significant.

By entering the command:

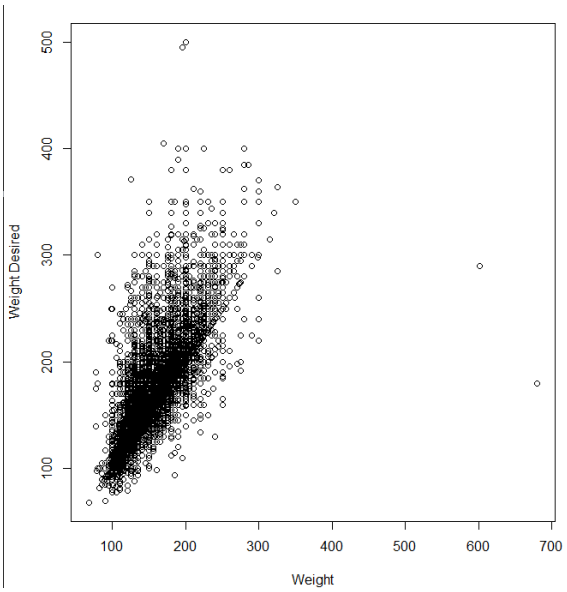
```
boxplot(bmi ~ cdc$gender)
```



Female BMI results are shown to have a slightly wider range than males, where males have a higher median BMI. So males have a higher likelihood of having a higher BMI than a female, while females' BMI have more variance.

On Your Own:

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




```
>plot(cdc$weight ~ cdc$wtdesire, xlab="Weight", ylab="Weight Desired")
```

There seems to be a positive correlation between a subject's weight and desired weight, where the scatterplot displays a trend with positive slope, with what may be more individuals desiring to be heavier than they are.

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`.

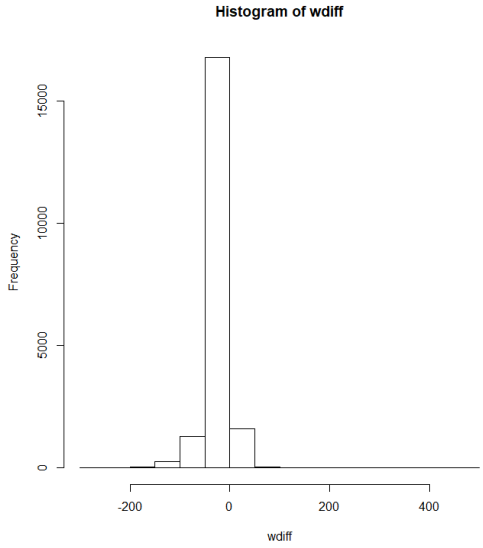
```
>wdiff <- (cdc$wtdesire - cdc$weight)
```

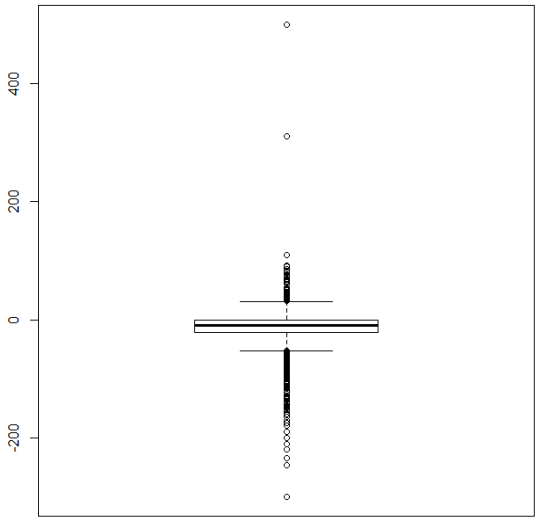
3. What type of data is `wdiff`? If an observation `wdiff` is 0, what does this mean about the person's weight and desired weight. What if `wdiff` is positive or negative?

The above command created a dataset of int values, where a result of zero means that individuals are currently their desired weight, negative values are indicative of a subject's desire to lose weight, and positive values indicate the desire to gain 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?

```
> hist(wdiff)
> boxplot(wdiff)
> summary(wdiff)
   Min. 1st Qu.  Median    Mean 3rd Qu. Max.
-300.00  -21.00  -10.00  -14.59    0.00 500.00
```

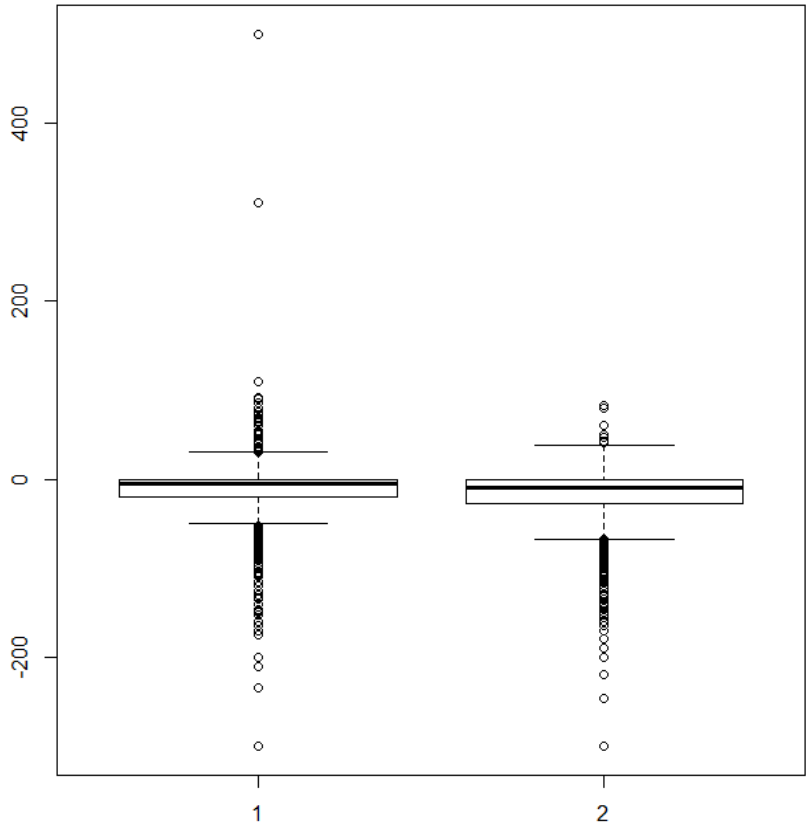




What the above commands display is that while there were some unorthodox responses, an IQR of 21, where the frequency of “0” drastically outweighs other results, one could infer that a majority of those sampled are satisfied with their current weight.

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

```
> males <- subset(cdc, cdc$gender == "m")
> females <- subset(cdc, cdc$gender == "f")
> m_wdiff <- (males$wtdesire - males$weight)
> f_wdiff <- (females$wtdesire - females$weight)
> boxplot(m_wdiff, f_wdiff)
> summary(m_wdiff)
   Min. 1st Qu.  Median    Mean 3rd Qu. Max.
-300.00  -20.00   -5.00  -10.71    0.00  500.00
> summary(f_wdiff)
   Min. 1st Qu.  Median    Mean 3rd Qu. Max.
-300.00  -27.00  -10.00  -18.15    0.00  83.00
```



The above commands display summaries and a boxplot of the difference in desired weights when compared between genders, where “1” denotes males, and “2” denotes females. Males have an IQR

of 20 versus the female's 27, with their Q3 both at zero. This means that a majority of both males and females desire to lose weight or are at their desired weight, where females tend to desire to lose more weight. Male results have more outliers in their responses, which is highlighted by the seemingly ridiculous responses of 300 and 500 pounds of weight that they would prefer to gain.

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(cdc$weight)

169.683

> sd(cdc$weight)

40.08097

> one_sd <- subset(cdc, cdc$weight > (mean(cdc$weight)-sd(cdc$weight)) &
cdc$weight < (mean(cdc$weight)+sd(cdc$weight)))

> nrow(one_sd)/nrow(cdc)

0.7076
```

70.76% of the reported weights are within one standard deviation of the mean of 169.683, where the reported weights are between 129.602, and 209.764, or the set [130,209].