Lab2-Idris Hayward

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

names(cdc)

## [1] "genhlth"  "exerany"  "hlthplan" "smoke100" "height"   "weight"
## [7] "wtdesire" "age"      "gender"
```

Exercise 1: How many cases are there in this data set? How many variables? For each variable, identify its data type (i.e., nominal, ordinal, discrete numeric, or continuous numeric).

Cases: 2000

Variables: 9. GenHealth (ordinal),Exerany (nominal), hlthplan (nominal), smoke100 (nominal), height (continuout numeric), weight (continuous numeric), wtdesire (continuous numeric), age (continuous numeric), gender (nominal)

Exercise 2: Generate 5-number summaries for both height and age (Hint: There should be 5 numbers in each summary, not 6), and compute the interquartile range for each.

Height

```
quantile(cdc$height)

##   0%  25%  50%  75% 100%
##   48   64   67   70   93
```

Height Interquartile Range

```
IQR(cdc$height)

## [1] 6
```

Age

```
quantile(cdc$age)
```

```
##   0%  25%  50%  75% 100%
##   18   31   43   57   99
```

Age Interquartile Range

```
IQR(cdc$age)
```

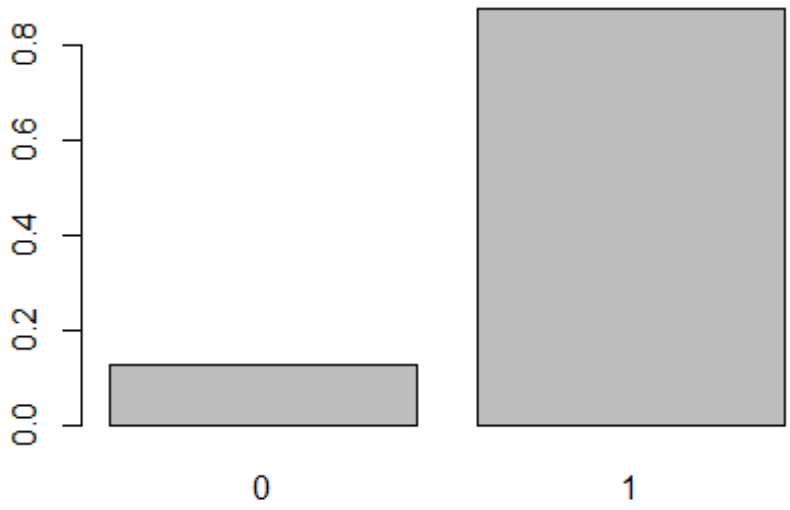
```
## [1] 26
```

Exercise 2 continued: Compute and plot the relative frequency distributions for hlthplan and exerany. How many females are in the sample? What percentage of the people in this sample report being in good health?

```
table(cdc$hlthplan)/nrow(cdc)
```

```
##
##      0      1
## 0.1262 0.8738
```

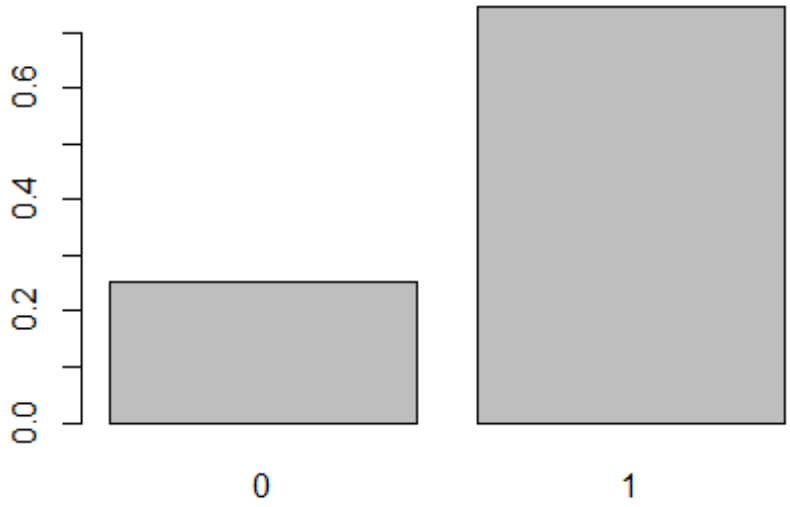
```
barplot(table(cdc$hlthplan)/nrow(cdc))
```




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

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

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




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

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

There are 104321 females in the sample

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

```
##
## excellent very good      good      fair      poor
##      4657      6972      5675      2019       677
```

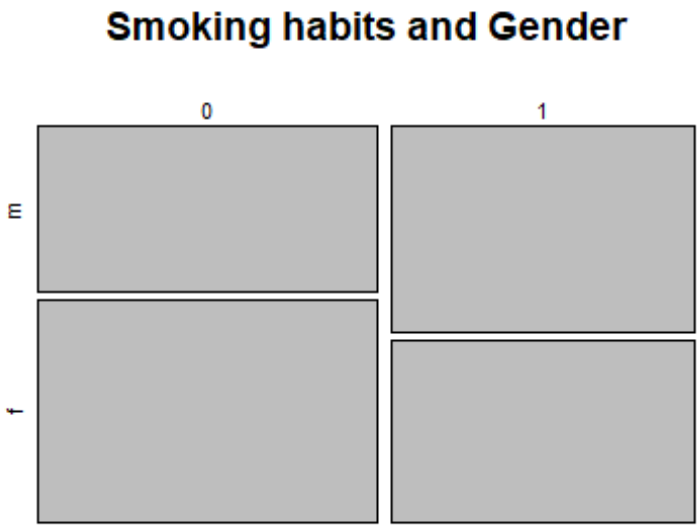
```
(5675/(4657+5675+2019+677+6972)) * 100
```

```
## [1] 28.375
```

28.37% report being in good health

Exercise 3 What does your mosaic plot reveal regarding the association between smoking habits and gender?

```
mosaicplot(table(cdc$smoke100, cdc$gender), main = "Smoking habits and
Gender")
```



#There is a greater number of Men who have smoked at least 100 cigarettes. However there are more females who have not smokes at least 100 cigarettes

Exercise 4: Create a new object called under23_and_smoke that contains all observations for respondents under the age of 23 who have smoked at least 100 cigarettes in their lifetimes. How many respondents are under age 23 and have smoked at least 100 cigarettes in their lifetimes? Include supporting R output as part of your answer to this exercise.

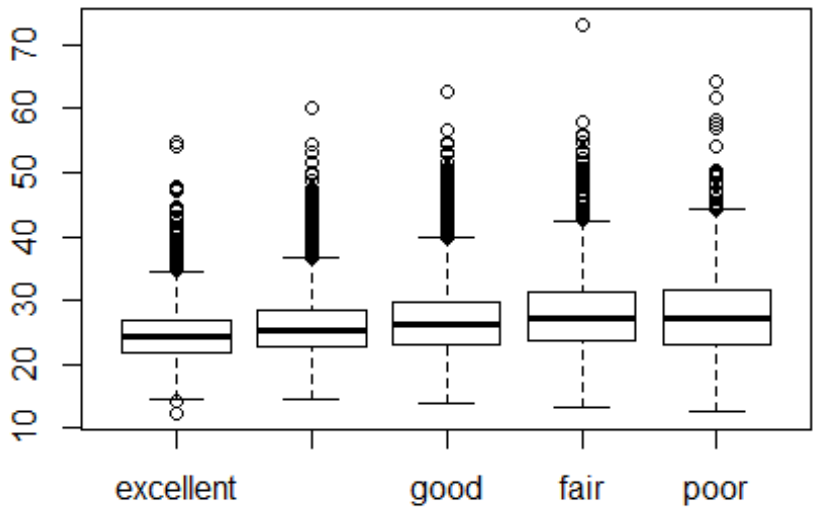
```
under23_and_smoke <- subset(cdc, cdc$smoke100 == TRUE & cdc$age < 23)
nrow(under23_and_smoke)

## [1] 620
```

620 people under the age of 23 have smoked at least 100 cigarettes

Exercise 5: What does this boxplot display suggest about the association between BMI and general health?

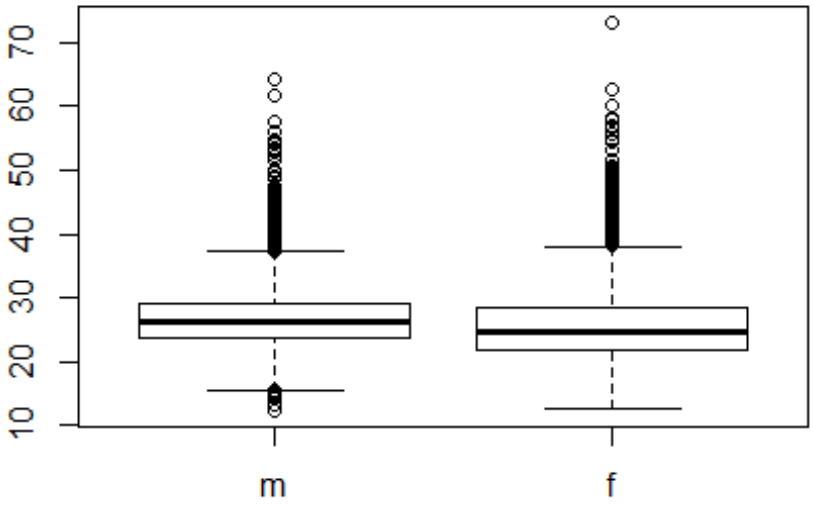
```
bmi <- (cdc$weight/cdc$height^2) * 703
boxplot(bmi ~ cdc$genhlth)
```



#The poorer a person is in health, the higher BMI they have.

Exercise 5 continued: Pick another categorical variable from the data set and generate a boxplot display to see how it relates to BMI. State the variable you chose, explain why you might think it would have a relationship to BMI, and indicate what the boxplot display seems to suggest.

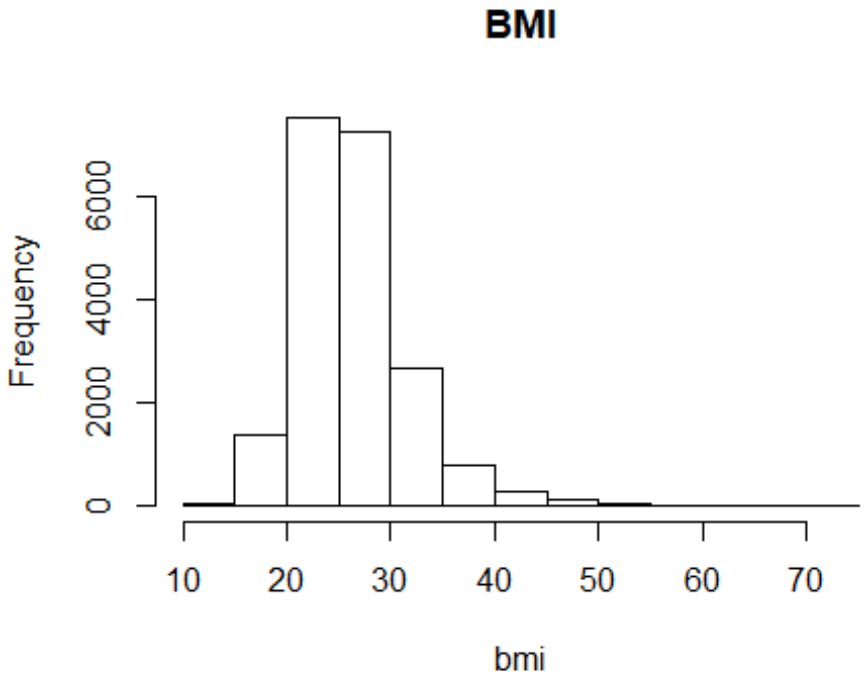
```
bmi <- (cdc$weight/cdc$height^2) * 703
boxplot(bmi ~ cdc$gender)
```



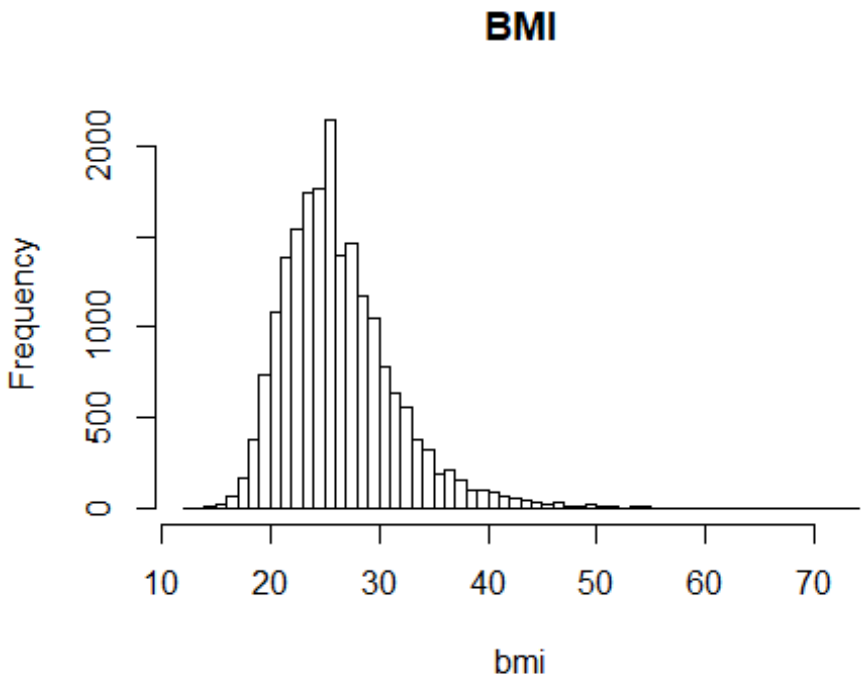
#Gender. The reason gender was chosen was because BMI is derrived from the height and weight of individuals, BMI by Gender would compare the two sexes and help determine which would be healthier. Females generally have a lower average BMI than men.

Exercise 6: Note that you can flip between plots that you've created by clicking the forward and backward arrows in the lower right region of RStudio, just above the plots. How do these two histograms compare? Describe the shape of the BMI distribution. Include both histograms as part of your answer to this question. NEED TO FINISH

```
hist(bmi, main ="BMI")
```




```r
hist(bmi, breaks = 50, main="BMI")
```

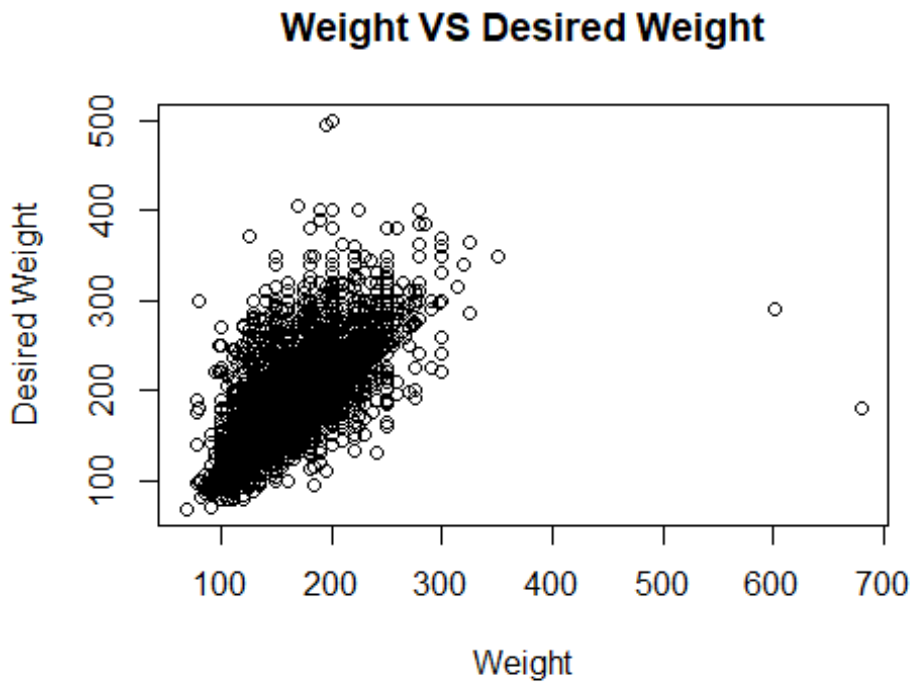


#The shape of both the histograms is skewed right. With breaks included with the histogram, it becomes more accurate because the breakdown of the data becomes more distinct.

Homework Assignment

1. . Make a scatterplot of weight (x) versus wtdesire (y). Describe the apparent relationship between these two variables.

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



#The relationship between desired weight and weight is that there are more people who desire to lose weight

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 in the order I specified above and assigning the result to a new object called wdiff. Answer this question with the R code you used to create this variable.

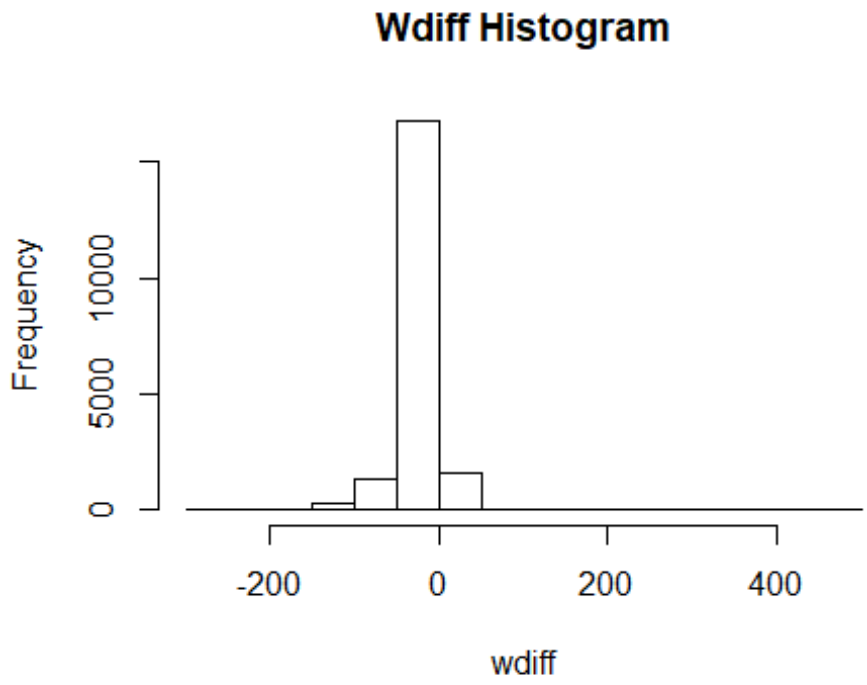
```
wdiff <- cdc$wtdesire - cdc$weight
```

3.If an observed wdiff is 0, how do a person's actual weight and desired weight compare? How do they compare if the wdiff is positive? Negative?

If wdiff is 0 then there a person weighs at their desired weight. If negative, a person wishes to lose weight. If postive, they wish to gain weight

4. Describe the distribution of wdiff in terms of its center, shape, and spread. What does this information tell us about how people feel about their current weight? Include a histogram and summary statistics as part of your answer.

```
hist(wdiff, main ="Wdiff Histogram")
```




```
summary(wdiff)
```

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

The center is the median which is 10. The shape of the histogram is skewed right The spread ranges from -200 to 500 This tells us that most people are interested in losing weight instead of gaining.