Data 606 - Lab 1

John Perez

1/31/2019

Exercises

Load Data

```
source("https://raw.githubusercontent.com/jreznyc/DATA606/master/Labs/Lab1/cdc.R")
```

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).

```
str(cdc)
```

```
## 'data.frame':    20000 obs. of  9 variables:
##  $ genhlth : Factor w/ 5 levels "excellent","very good",..: 3 3 3 3 2 2 2 2 3 3 ...
##  $ exerany : num  0 0 1 1 0 1 1 0 0 1 ...
##  $ hlthplan: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ smoke100: num  0 1 1 0 0 0 0 0 1 0 ...
##  $ height  : num  70 64 60 66 61 64 71 67 65 70 ...
##  $ weight  : int  175 125 105 132 150 114 194 170 150 180 ...
##  $ wtdesire: int  175 115 105 124 130 114 185 160 130 170 ...
##  $ age     : int  77 33 49 42 55 55 31 45 27 44 ...
##  $ gender  : Factor w/ 2 levels "m","f": 1 2 2 2 2 2 1 1 2 1 ...
```

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 of height variable & IQR
summary(cdc$height)
```

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

```
c("IQR height"=summary(cdc$height)[[5]]-summary(cdc$height)[[2]])
```

```
## IQR height
##          6
```

```
#summary of age variable & IQR
summary(cdc$age)
```

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

```
c("IQR age"=summary(cdc$age)[[5]]-summary(cdc$age)[[2]])
```

```
## IQR age
##      26
```

```
#relative freq distr gender & exerany
prop.table(table(cdc$gender, cdc$exerany))
```

```
##
```

```
##           0       1
##   m 0.10745 0.37100
##   f 0.14685 0.37470
```

```
#number of males in sample
summary(cdc$gender)['m']
```

```
##    m
## 9569
```

```
#proportion of the sample reports being in excellent health
table(cdc$genhlth)['excellent']/20000
```

```
## excellent
##   0.23285
```

Exercise 3. What does the mosaic plot reveal about smoking habits and gender?
It reveals that men are more likely to smoke than women.

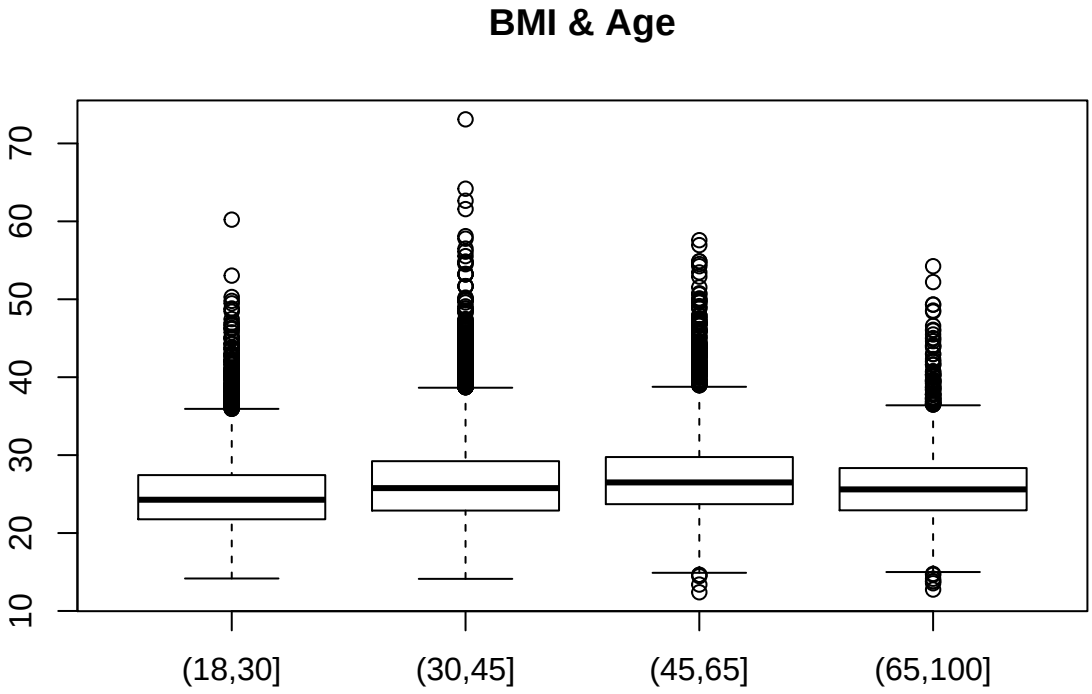
Exercise 4. Create a new object called `under23_and_smoke` that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime.

```
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.
The box plot shows that as general health goes from excellent to poor, bmi values become more spread out while at the same time having higher density of larger values for bmi.

For the second variable to compare with BMI, I'd choose age as I believe it should have a relationship to BMI because humans undergo significant physical and metabolical changes as they age. Below is a box plot of BMI and age group which shows that BMI is generally higher for age groups (30-45], and (45-65]

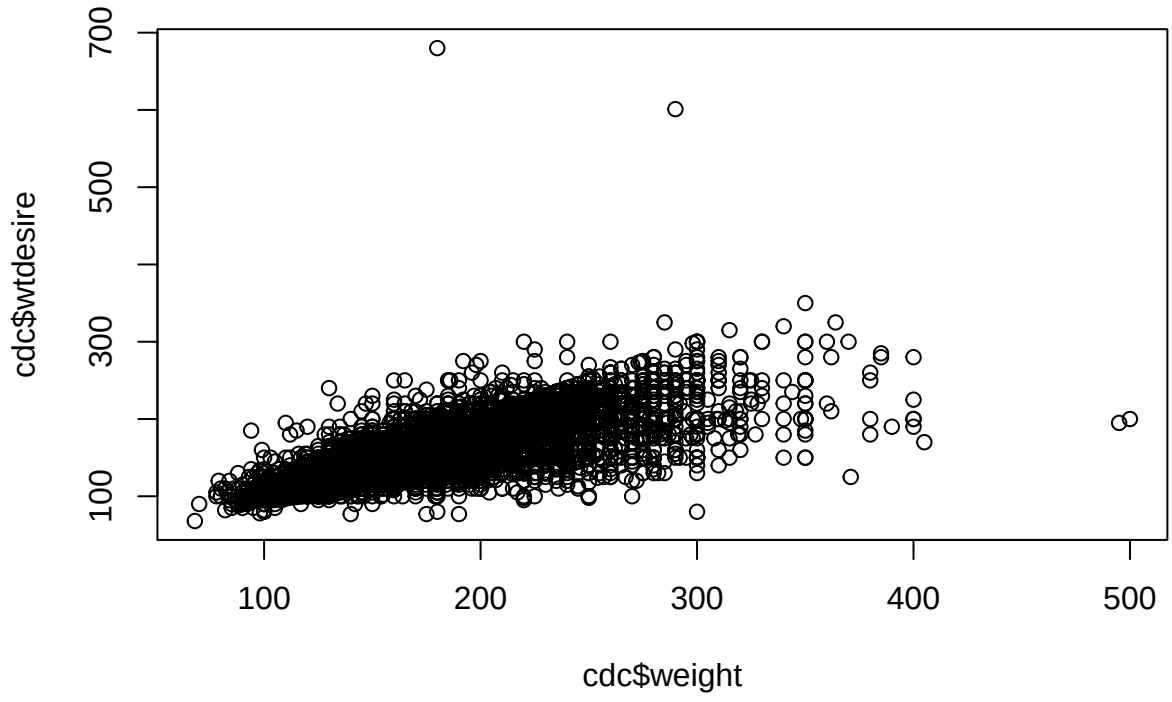
```
bmi <- (cdc$weight / cdc$height^2) * 703
age <-cut(cdc$'age', breaks=c(18,30,45,65,100), ordered_result = TRUE)
boxplot(bmi ~ age, main="BMI & Age")
```



On Your Own Questions

(1) Make a scatterplot of weight versus desired weight. Describe the relationship between these two variables.
Below is a scatterplot of weight versus desired weight. These two variables have a positive linear association.

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



(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?

wdiff is numerical data. If the observation is 0, it means that the person is at their desired weight. If wdiff is negative it means the person's desired weight is less than their current weight. If wdiff is positive it means the person's desired weight is greater than their current weight.

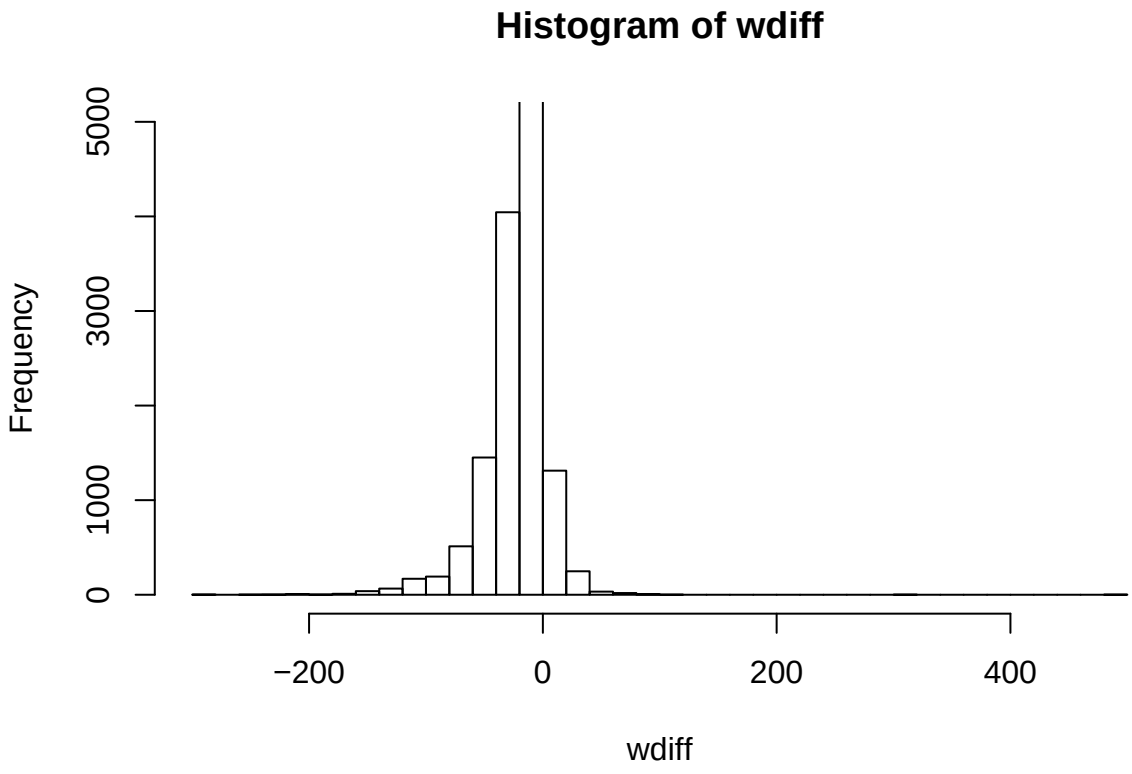
(4) Describe the distribution of wdiff in terms of its center, shape, and spread, including any plots you use. What does this tell us about how people feel about their current weight?

The distribution of wdiff is centered at -10, unimodal, and with a high spread and a significant number of outliers. This indicates that individuals have wildly varying differences between their current and desired weight.

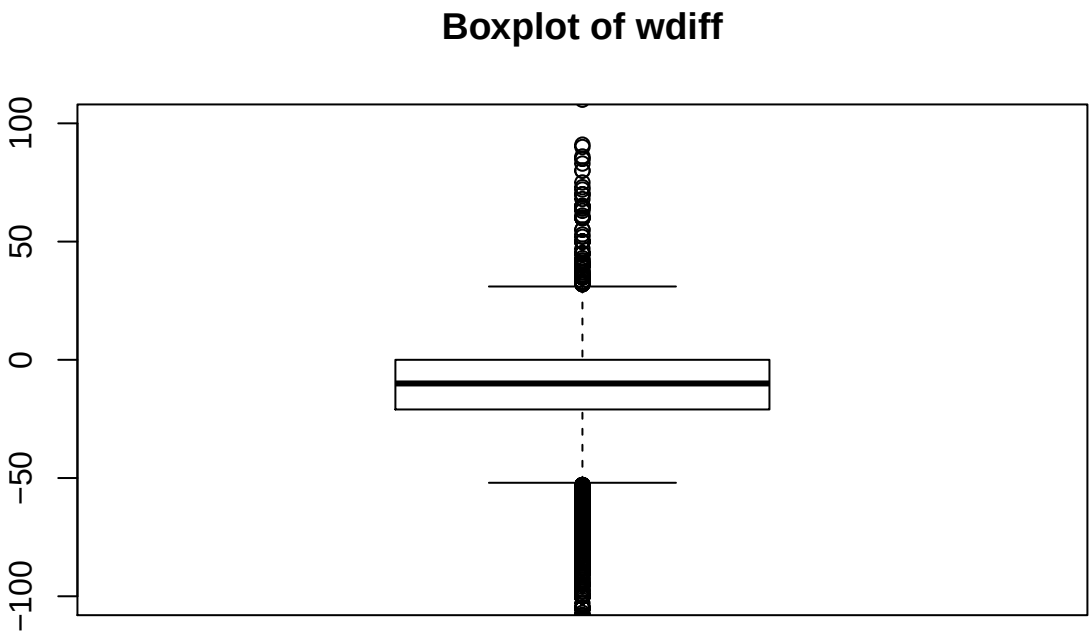
```
summary(wdiff)
```

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

```
hist(wdiff, breaks = 50, ylim=c(0,5000))
```




```
boxplot(wdiff, ylim=c(-100,100), main="Boxplot of wdiff")
```



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

Men are more likely to have a smaller difference between desired weight and actual weight. Women have a greater difference between their desired and actual weight. There is a large spread in values for both genders.

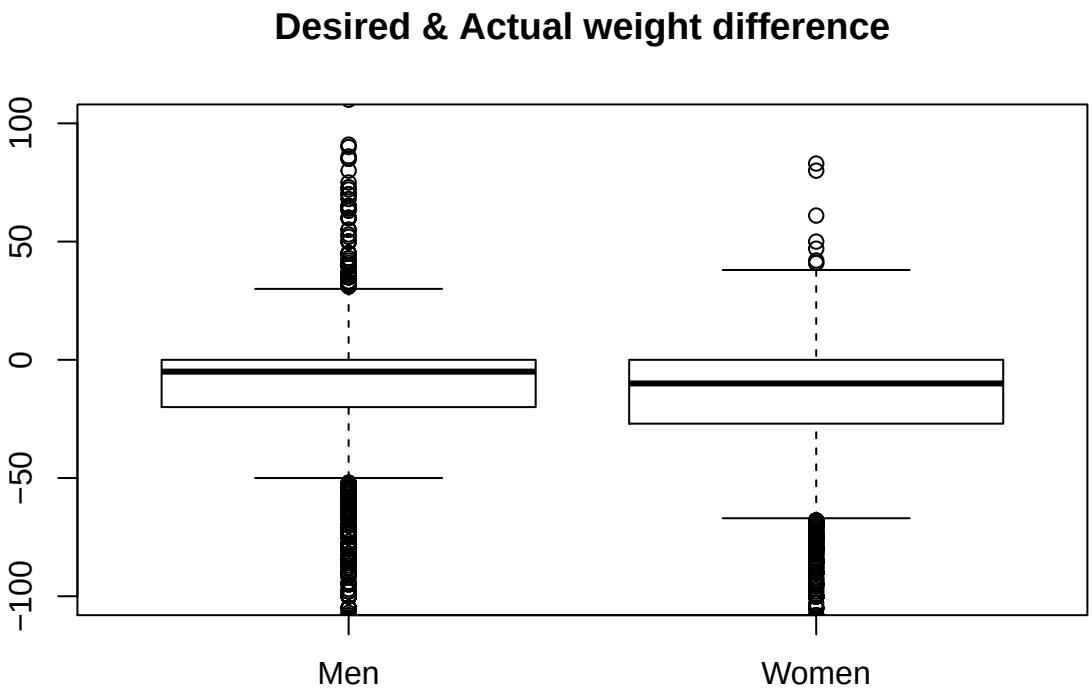
```
mdiff <-  cdc[cdc$gender=='m',]$wtdesire - cdc[cdc$gender=='m',]$weight
wdiff <-  cdc[cdc$gender=='f',]$wtdesire - cdc[cdc$gender=='f',]$weight
summary(mdiff)
```

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

```
summary(wdiff)
```

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

```
boxplot(mdiff,wdiff, ylim=c(-100,100), names=c('Men','Women'), main="Desired & Actual weight difference
```



(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.

```
sd <- sd(cdc$weight)
m <- mean(cdc$weight)
lo <- m - sd/2
hi <- m + sd/2
sum(cdc$weight>=lo & cdc$weight<=hi)/length(cdc$weight)
```

```
## [1] 0.39335
```