Brooks-Lab1

Dan Brooks

February 6, 2016

```
#Show the top few rows of the matrix
head(cdc)
```

```
##     genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 1      good       0        1        0     70    175      175  77      m
## 2      good       0        1        1     64    125      115  33      f
## 3      good       1        1        1     60    105      105  49      f
## 4      good       1        1        0     66    132      124  42      f
## 5 very good       0        1        0     61    150      130  55      f
## 6 very good       1        1        0     64    114      114  55      f
```

```
#Show the names of each column in the matrix
names(cdc)
```

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

```
#genhlth (Categorical, Ordinal)
#exerany (Catergorical, Variable)
#hlthplan (Catergorical, Variable)
#smoke100 (Categorical, Variable)
#height (Numerical, Continuous)
#Wieght (Numnerical, Continous)
#wtdesire (Numerical, Continuous)
#age (Numerical, discrete)
#gender (Cantegorical, variable)

#Show the length (Num of cases) of an individual column
length(cdc$genhlth)
```

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

```
#Show the width (Num of Variables) of the overall matrix
length(cdc)
```

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

```
#There are 20,000 cases within this data set.
#There are 9 different variables in this data set
```

```
#five number summary of height
summary(cdc$height)
```

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

```r
#IQR for height
IQRH <- 70.00 - 64.00
IQRH
```

```
## [1] 6
```

```r
#five number summary of age
summary(cdc$age)
```

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

```r
#IQR for age
IQRA <- 57.00 - 31.00
IQRA
```

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

```r
#Relative Distrisution of gender and health rating
table(cdc$gender, cdc$exerany)
```

```
##
##        0    1
##   m 2149 7420
##   f 2937 7494
```

```r
#total number of males in the study
TotMales <- 2149 + 7420
TotMales
```

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

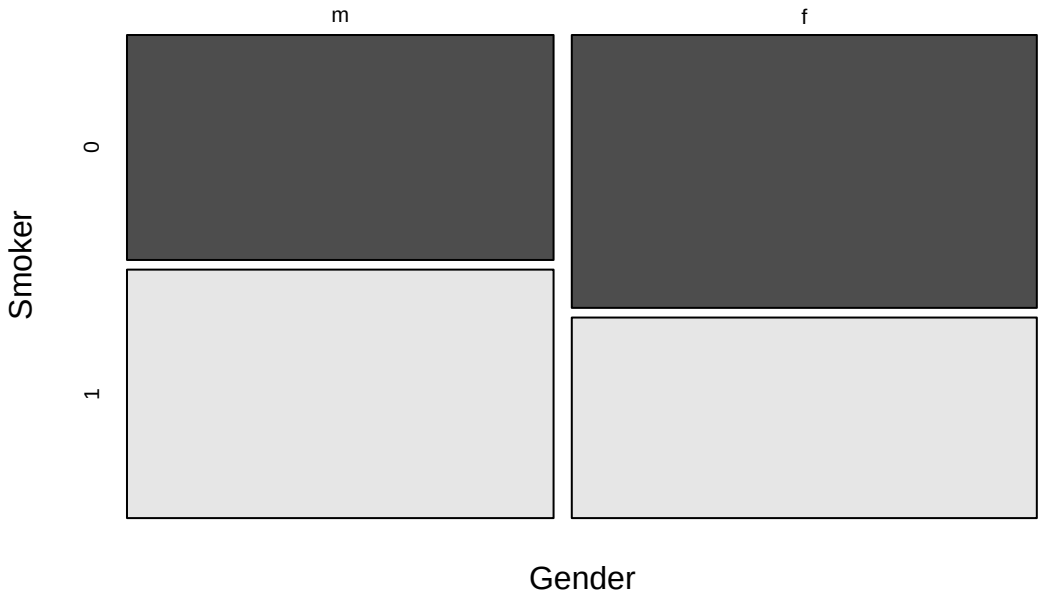
```r
#proportion of sample in excellent health
TotalPeople <- 2149 + 7420 + 2937 + 7494
TotalPeople
```

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

```r
#Excellent Health
ExHealth <- 7420 + 7494
#Proportion of participants in excellent health
ExProp <- ExHealth / TotalPeople
ExProp
```

```
## [1] 0.7457
```

```r
#Create a mosaic plot of smoking habits Vs gender
mosaicplot(table(cdc$gender,cdc$smoke100), xlab = 'Gender', ylab = 'Smoker', color = TRUE)
```




```
#It appears that males have more of a smoking habit than females.
```

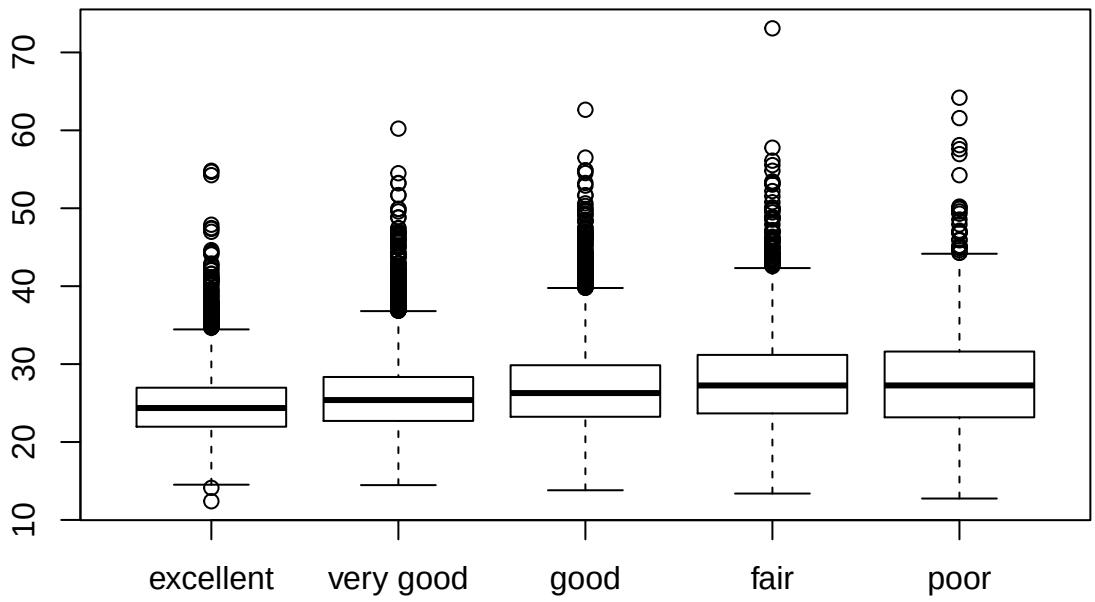
```
#Subset the CDC matrix to only people uder the age of 23 and smoke more than 100 cigarettes
under23_and_smoke <- subset(cdc, age < 23 & smoke100 == 1)
head(under23_and_smoke)
```

```
##       genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 13  excellent       1        0        1     66    185      220  21      m
## 37  very good       1        0        1     70    160      140  18      f
## 96  excellent       1        1        1     74    175      200  22      m
## 180      good       1        1        1     64    190      140  20      f
## 182 very good       1        1        1     62     92       92  21      f
## 240 very good       1        0        1     64    125      115  22      f
```

```
dim(under23_and_smoke)
```

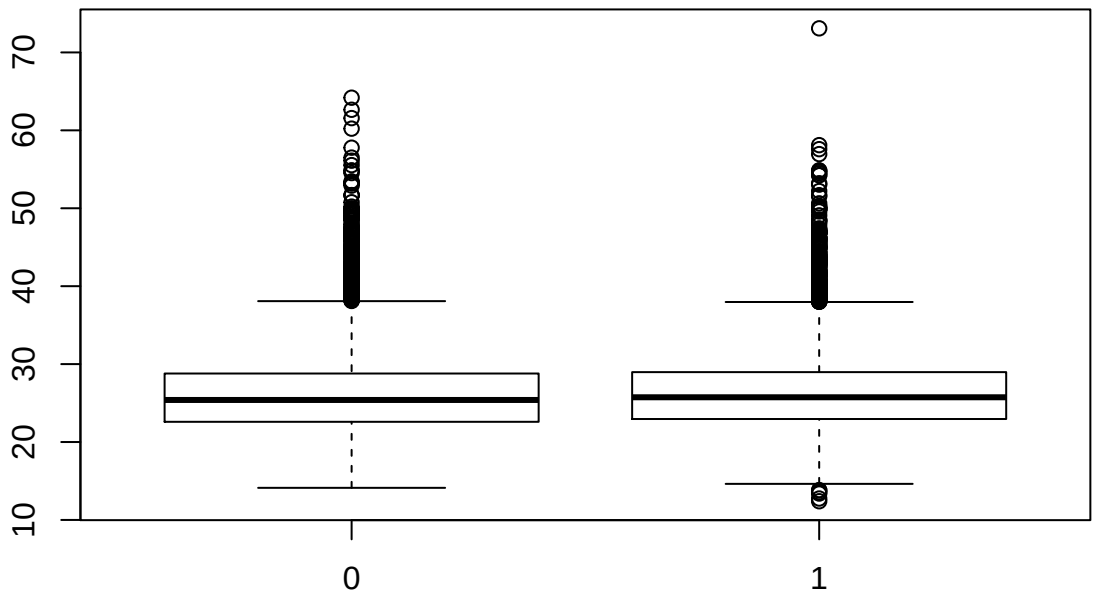
```
## [1] 620   9
```

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



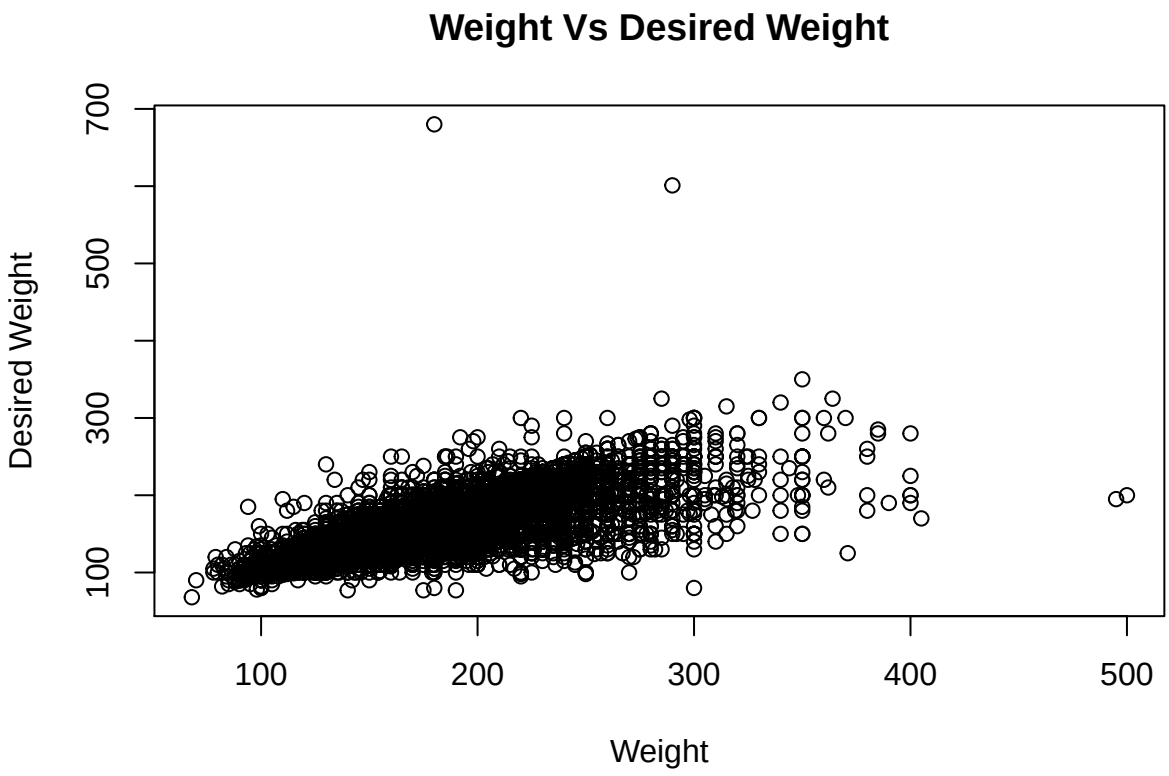

```
#This boxplot is showing the BMI (Body Mass Index) of the participants in the case study
#compared to their general overall health. Each box plot shows the BMI of the participants
#but it is serparated out by their general health.

boxplot(bmi ~ cdc$smoke100)
```




```
#I chose to compare the BMI and amount of cigraettes in a liftime
#I chose that because I have heard that smoking can cause mass amounts of weigth
#loss. Poeple loo sickly and malnurished. I thought I was going to see the mean
#very close to zero. Instead I am seeing that both of the graphs look almost
#indentical. That was not what I was expecting.
```

```
plot(cdc$weight, cdc$wtdesire, title(main = 'Weight Vs Desired Weight'), xlab = 'Weight',
     ylab = 'Desired Weight')
```




```
#The two variables seem to have a linear relationship. Most of the people look to be about 100 pounds
#less than their current weight. It appears that most of the people would like to lose the same
#amount.

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

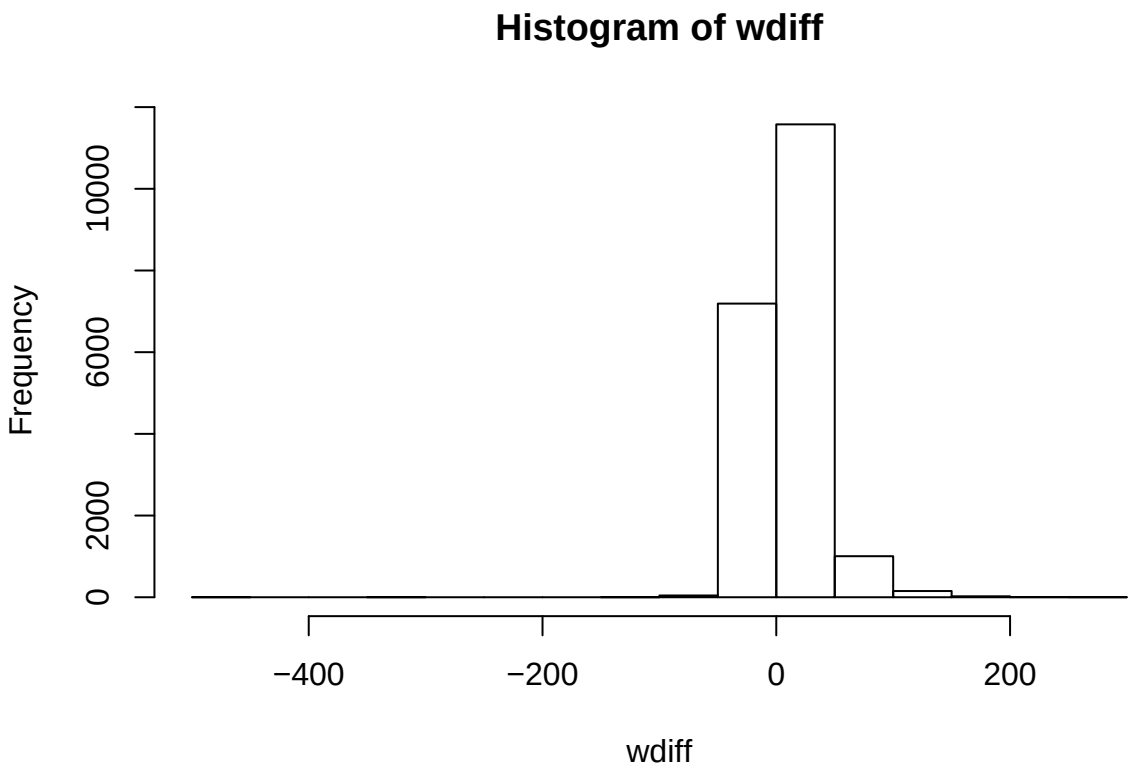
```
##  int [1:20000] 0 10 0 8 20 0 9 10 20 10 ...
```

```
#The new object wdiff is and integer data type
#If widff is 0 then the person is already at their desired weight
#If widff is negative then their desired weight is more than their current weight
#If widff is positive than their current wieght is larger than their desired weight

summary(wdiff)
```

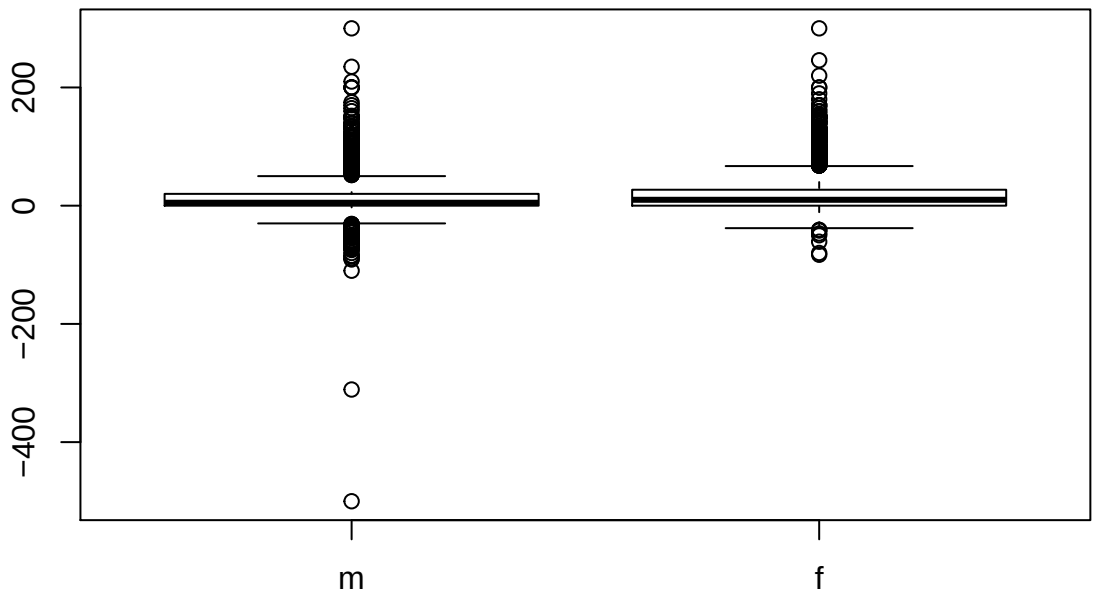
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -500.00    0.00   10.00   14.59   21.00  300.00
```

```
hist(wdiff)
```




```
#It actually appears to be a pretty normal distribution. The graph is centered around 0
#which means that a lot of the people are happy with their current weight. There are a
#few outliers, but overall people are happy with their weight or want to make very
#minute changes.

boxplot(wdiff ~ cdc$gender)
```




```
#It appears than me actually want to gain some weight, while women are more about losing it.
#I am guessing women want to lose weight to keep a sloim figure, while men probably want
#to gain more muscle.

mean(cdc$weight)
```

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

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

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

```
PosSd <- sd(cdc$weight) + mean(cdc$weight)
NegSd <- -sd(cdc$weight) + mean(cdc$weight)
OneSDTotal <- subset(cdc, cdc$weight > NegSd & cdc$weight < PosSd)
dim(OneSDTotal)
```

```
## [1] 14152     9
```

```
PropOneSD <- 14152/20000
PropOneSD
```

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