Data Acquisition

Erica Wong, Cindy Kang, and Abby Vogel

4/25/2017

Loading the Data:

```r
data <- read.csv('full_data_v2.csv', header = T)
data$X <- NULL
```

Number of Rows:

```r
(nrow(data))
```

```
## [1] 671
```

First 5 Rows of Data:

```r
head(data, 5)
```

```
##      id gender age        bmi opinion_wt action_wt pe_yn freq_pe
## 1 73579      F  12     normal     normal  maintain   yes       5
## 2 73584      M  13 overweight overweight      lose   yes       3
## 3 73587      M  14      obese overweight      lose   yes       5
## 4 73599      F  13     normal     normal  maintain   yes       3
## 5 73601      M  12     normal     normal      lose   yes       5
##         enjoy_pe
## 1          agree
## 2          agree
## 3 strongly agree
## 4 strongly agree
## 5 strongly agree
```

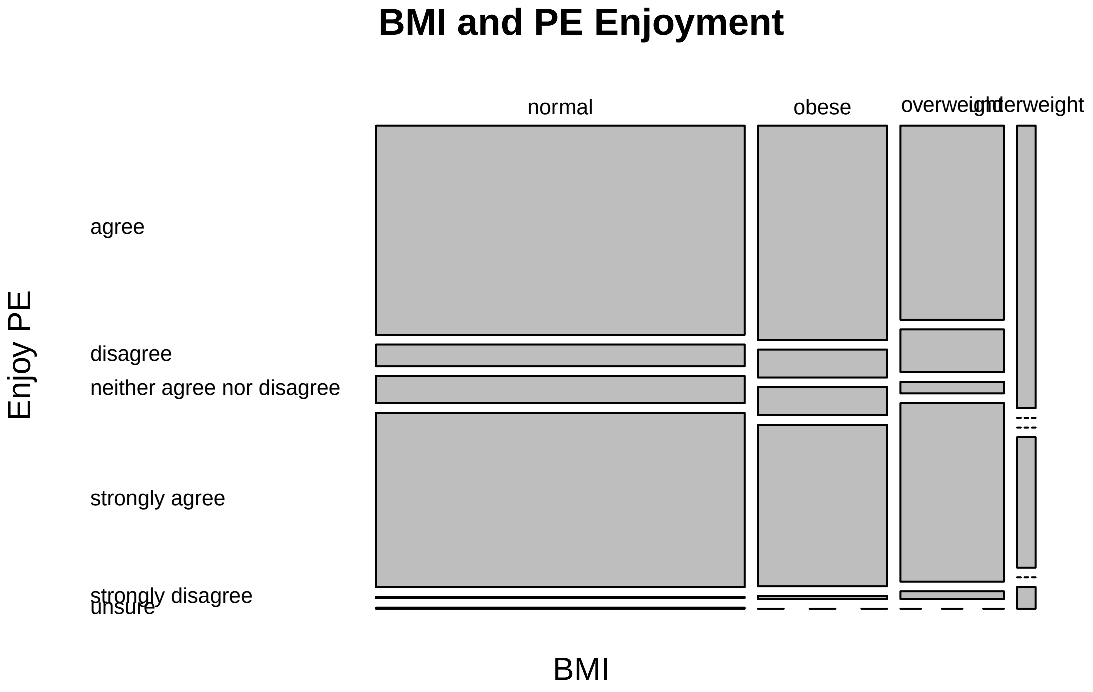
Summary:

```r
summary(data)
```

```
##        id          gender       age                  bmi
##  Min.   :73579   F:330   Min.   :12.00   normal     :399
##  1st Qu.:75879   M:341   1st Qu.:12.00   obese      :140
##  Median :78576           Median :14.00   overweight :112
##  Mean   :78608           Mean   :13.48   underweight: 20
##  3rd Qu.:81184           3rd Qu.:14.00
##  Max.   :83704           Max.   :15.00
##        opinion_wt     action_wt   pe_yn         freq_pe
##  normal     :484   gain    : 82   no :122   Min.   :0.00
##  overweight :134   lose    :269   yes:549   1st Qu.:2.00
##  underweight: 53   maintain:184             Median :3.00
##                    nothing :135             Mean   :3.17
##                    unsure  :  1             3rd Qu.:5.00
##                                             Max.   :5.00
##                         enjoy_pe
##  agree                     :324
##  disagree                  : 40
##  neither agree nor disagree: 37
##  strongly agree            :264
##   strongly disagree          :  4
##   unsure                     :  2
```

Plot:

```
library(ggplot2)
er <- table(data$bmi, data$enjoy_pe)
mosaicplot(er, las=1, xlab="BMI", ylab="Enjoy PE", main="BMI and PE Enjoyment")
```



What We Did:

In order to get the data to this point, we first looked at the codebook of each of the datasets that we were interesed in using. Then using the plyr and dplyr packages, we were able to join all of the different SAS files that we were interested in using based upon the given ID. Because the column names are coded by something that is impossible to understand without having the codebook open, we started off by renaming our columns to something that can be understood by a person who is reading our code. Additionally, since all the variables are coded by numbers, we replaced the numbers with informative factors that allow us to know what the data is telling us about the subjects without needing to look up what each number for each column means. Finally, we looked at our data and realized that there were some missing values. One of columns that was giving us a lot of NAs was freq_pe, which is the frequency of PE class. This is because this column was marked as NA when a student didn't have PE at school. So, we recoded the missing value to 0. Additionally, there were columns that were missing certain pieces of information, if there were less than 3 blanks, then we would look at other rows with similar characteristics and replace that value with the mode because we assumed that they would act like the majority of people. For rows with more than 3 blanks, there was just too much missing to make a reasonable assumption, so we decided to throw away those rows. We understand this may skew our data a little bit, but we could not find another more reasonable thing to do because we did not want to create bias either.