Data Acquisition

Erica Wong, Cindy Kang, and Abby Vogel

4/25/2017

Loading the Data:

```
data <- read.csv('../full_data_v3.csv', header = T)
data$X <- NULL
```

Number of Rows:

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

```
## [1] 737
```

First 5 Rows of Data:

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

```
##      id   int_wt  exam_wt psu stratum gender age        bmi opinion_wt
## 1 73579 63248.99 70708.03   1     110      F  12     normal     normal
## 2 73584 72700.38 72182.24   1     105      M  13 overweight overweight
## 3 73587 16220.74 15523.09   2     115      M  14      obese overweight
## 4 73599 25841.53 27288.18   2     107      F  13     normal     normal
## 5 73601 17430.86 18842.87   1     115      M  12     normal     normal
##   action_wt pe_yn freq_pe       enjoy_pe
## 1  maintain   yes       5          agree
## 2      lose   yes       3          agree
## 3      lose   yes       5 strongly agree
## 4  maintain   yes       3 strongly agree
## 5      lose   yes       5 strongly agree
```

Summary:

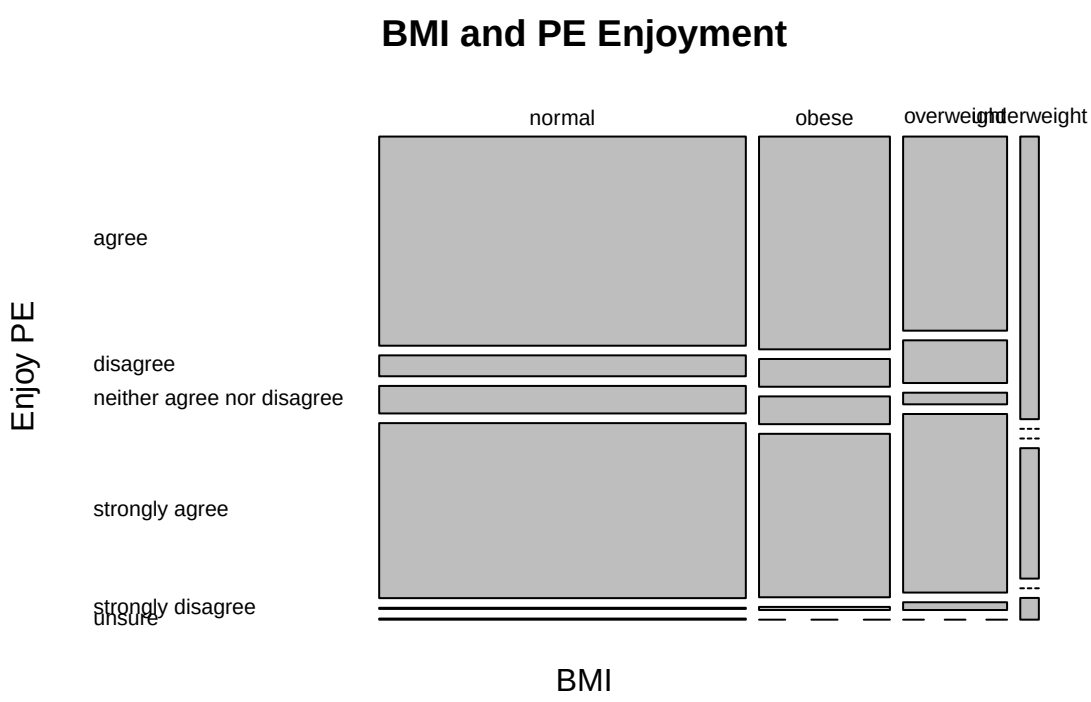
```
summary(data)
```

```
##        id            int_wt          exam_wt            psu
##  Min.   :73579   Min.   :  5874   Min.   :     0   Min.   :1.000
##  1st Qu.:76005   1st Qu.: 11723   1st Qu.: 11805   1st Qu.:1.000
##  Median :78623   Median : 15735   Median : 15438   Median :1.000
##  Mean   :78651   Mean   : 23418   Mean   : 23284   Mean   :1.446
##  3rd Qu.:81220   3rd Qu.: 21850   3rd Qu.: 22125   3rd Qu.:2.000
##  Max.   :83704   Max.   :102078   Max.   :104556   Max.   :2.000
##
##     stratum      gender       age                bmi
##  Min.   :104.0   F:362   Min.   :12.00   normal     :416
##  1st Qu.:107.0   M:375   1st Qu.:12.00   obese      :148
##  Median :111.0           Median :14.00   overweight :122
##  Mean   :111.1           Mean   :13.49   underweight: 21
##  3rd Qu.:115.0           3rd Qu.:14.00   NA's       : 30
##  Max.   :118.0           Max.   :15.00
##
##        opinion_wt     action_wt    pe_yn        freq_pe
##  normal     :484   gain    : 82   no  :123   Min.   :0.00
##  overweight :134   lose    :269   yes :556   1st Qu.:2.00
```

```
##  underweight: 53   maintain:184   NA's: 58   Median :3.00
##  NA's       : 66   nothing :135              Mean   :3.18
##                    unsure  :  1              3rd Qu.:5.00
##                    NA's    : 66              Max.   :5.00
##                                              NA's   :58
##                        enjoy_pe
##  agree                     :324
##  disagree                  : 40
##  neither agree nor disagree: 37
##  strongly agree            :265
##  strongly disagree         :  4
##  unsure                    :  2
##  NA's                      : 65
```

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 are rows that have NAs in them but currently we are in the process of tryng to fix that by imputation, probably going to use hot desk imputation and will fill it in by row.