Day 01 - Exercise 02 - Data Manipulation

Victor Calderon

15 August, 2018

This are the responses/answers to the problems posed by the exercises in the `data-manipulation` file.

Preamble

> Suppose the data below represent body weight measurements for six participants in a weight loss intervention program, at three follow-up time points (t0, t1, and t2)

```
dat <- data.frame(id  = c( 1, 2, 3, 4, 5, 6),
                  age = c(47,52,35,28,62,44),
                  sex = factor(c("M","M","F","M","F","F")),
                  wt_t0 = c(278, 340, 239, 290, 244, 220),
                  wt_t1 = c(230, 302, 231, 277, 245, 201),
                  wt_t2 = c(211, 295, 231, 282, 243, 182))

head(dat)
```

```
##   id age sex wt_t0 wt_t1 wt_t2
## 1  1  47   M   278   230   211
## 2  2  52   M   340   302   295
## 3  3  35   F   239   231   231
## 4  4  28   M   290   277   282
## 5  5  62   F   244   245   243
## 6  6  44   F   220   201   182
```

Exercise 1

> 1. Convert these data to long format so that there is only one variable with weight measurements, and three rows per participant ('id')

Converting to *long format*:

```
data_melt <- melt(dat, id.vars=c('id','age','sex'))
head(data_melt)
```

```
##   id age sex variable value
## 1  1  47   M    wt_t0   278
## 2  2  52   M    wt_t0   340
## 3  3  35   F    wt_t0   239
## 4  4  28   M    wt_t0   290
## 5  5  62   F    wt_t0   244
## 6  6  44   F    wt_t0   220
```

We can examine the types of variables:

```
unique(data_melt$variable)
```

```
## [1] wt_t0 wt_t1 wt_t2
## Levels: wt_t0 wt_t1 wt_t2
```

and showing the lines for *Participant 1*

```
nrows_per_participant <- sum(ifelse(data_melt$id == 1, TRUE, FALSE))
nrows_per_participant
```

```
## [1] 3
```

This shows that there are **"3"** rows per participant!

Exercise 2

> Convert back to wide format, averaging across participant sex (resulting data frame should have one row for males and one for females)

Converting back to *wide* format:

```
dcast(data_melt, sex ~ variable, mean)
```

```
##   sex    wt_t0    wt_t1    wt_t2
## 1   F 234.3333 225.6667 218.6667
## 2   M 302.6667 269.6667 262.6667
```

Exercise 3

> Suppose there is a second data frame that contains additional information on participants and others, including marital status and other variables. Add marital status (only) to the above data (either the wide or long version) using the 'merge' command. What type of 'join' is this?

```
demo <- data.frame(id = 1:20,
                   married = sample(c(T,F), 20, replace=TRUE, prob=c(0.5,0.5)),
                   income  = rnorm(20, 75000, 10000),
                   state   = sample(state.abb, 20, replace=TRUE))
```

Let's examine the dataframe first

```
head(demo)
```

```
##   id married   income state
## 1  1    TRUE 82635.93    MA
## 2  2   FALSE 67009.91    VA
## 3  3   FALSE 63523.43    IA
## 4  4    TRUE 72105.38    MN
## 5  5    TRUE 72007.85    KY
## 6  6   FALSE 70884.89    NC
```

and its dimensions:

```
dim(demo)
```

```
## [1] 20  4
```

What about `data_melt`:

```r
dim(data_melt)
```

```
## [1] 18  5
```

Let's merge these two dataframes

```r
data_merged <-merge(data_melt, subset(demo, select=c('id','married')), by = "id", all.x = TRUE)
head(data_merged)
```

```
##   id age sex variable value married
## 1  1  47   M    wt_t0   278    TRUE
## 2  1  47   M    wt_t1   230    TRUE
## 3  1  47   M    wt_t2   211    TRUE
## 4  2  52   M    wt_t0   340   FALSE
## 5  2  52   M    wt_t1   302   FALSE
## 6  2  52   M    wt_t2   295   FALSE
```

Let's look at the *unique* elements for the column `married`:

```r
unique(data_merged$married)
```

```
## [1]  TRUE FALSE
```

Exercise 4

Suppose that a measure of 'irritability' (scale from 0 to 10) was also collected at each time point (see data below). Use melt and cast to convert this data to long format with three rows per participant, and one column for weights, and one column for irritability scores. (hint: after melting, you must add a 'measure' variable to indicate whether the corresponding value is a weight measurement or irritability measurement, and you must also add a time variable (i.e., 0, 1, or 2))

```r
dat <- data.frame(id  = c( 1, 2, 3, 4, 5, 6),
                  age = c(47,52,35,28,62,44),
                  sex = factor(c("M","M","F","M","F","F")),
                  wt_t0 = c(278, 340, 239, 290, 244, 220),
                  wt_t1 = c(230, 302, 231, 277, 245, 201),
                  wt_t2 = c(211, 295, 231, 282, 243, 182),
                  irr_t0 = factor(c(0, 1, 1, 0, 0, 0), levels=0:10, ordered=TRUE),
                  irr_t1 = factor(c(5, 3, 2, 1, 5, 7), levels=0:10, ordered=TRUE),
                  irr_t2 = factor(c(4, 3, 3, 1, 4, 6), levels=0:10, ordered=TRUE))
```

Exercise 5

Repeat the above task using the 'reshape' function.
