Post by mentor, **Guy Maskall**

I've seen a lot of posts about students being a bit stumped with this assignment, so I thought I'd mention an important couple of soft skills that you rarely see specifically highlighted:
* don't get intimitated, and
* break a problem down into steps.

I think the Samsung data is a great little assignment that really will give you a sense of achievement that is very relevant to the real world when you start to want to pull sources of data together into a more directly useful form. I thought I'd put down my thoughts on hints that might help make this assignment a little more digestible.

There seem to be a lot of files involved, but take stock and you'll make sense of it. I'll ignore the distinction between train and test for now. There are actual (predictor) measurements in a file. You'll want to take a peek at this file to see its structure. Would you use read.csv? Why not? The column (variable/predictor) names are given in a separate file. If you read this in, you'd get a vector of the same length as the number of columns in the measurements file. What function do you know that gets/assigns column names? Ah yes, but you only want certain columns. Well, either a good educated guess or reading some of the other information files will point you towards which columns contain mean and standard deviation values. What R functions do you know that can find text patterns in character vectors? Can you think how to use the output of such a function to give you a way to select only certain columns? (If you are still working with the separate data frame and vector of column names, make sure you index both consistently).

Okay, so that's the columns, what about the rows? Can you find the file that codes which observation is associated with particular subjects? And the file that codes which observation is associated with a particular activity? Obviously in both cases, the files you're looking for will have the same length as the number of observations. How do you now combine them with your data (measurement) data frame? And how do you interpret the activity, which is just a number? We want something more meaningful. Which file relates activity number to its label/description? Can you now use this to add a new column that has the appropriate activity label?

Are you here now? Great, you've built up a data frame that contains data about a load of measurements, the ID of the subject doing it, and what they were doing at the time. How about adding in a column that specifies whether the data belongs to the train or test set? (This label doesn't seem specifically required in the assignment, but you're merging two data sets and you should ALWAYS consider retaining a distinction between them - this might be the source filename or day etc., depending on your data sources). If you've now got two data frames, one for each of train and test, can you join them together?

Now you've done the data "heavy lifting" and you're ready to think about the last part, which is an aggregation task.