Code Book

The goal of this combined data set is to make a very clean and tidy data set.

The entire data set includes a READ ME file, a Code Book and two data frames, testingDF and testing2. The testingDF and testing2 includes data from test and train, all linked by the same 563 column names, 561 variables, and two others. SubjectID ranges from 1 to 30, one for each of the 30 subjects, that were 19 to 48 years old, and Activity identifies six separate activities, walking, walking upstairs, walking downstairs, standing, sitting, and laying down performed by each subject. Each row measurement includes the subject id and activity name. The subjects performed these activities wearing a Samsung Galaxy S II smartphone attached to their waist. The data frames represent three different types of observations from the accelerometer and gyroscope in the smartphone, body acceleration measured in “g”s, or meters per second squared, m/seg2, total acceleration measured also measured in m/seg2and the gyroscope measurements are in rad/seg.

The mean and standard deviation were extracted from the testingDF data frame. A second data frame was created to compute the mean of each variable, by subject and activity.

The original data set can be found here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

A list of all the files in the original data set are included in the READ ME file.

Several modifications have been made to the original data in order to make a tidier data set.

1. A column (SubjectID) has been added to identify which subject performed each measurement. Previously this information was stored in a separate file.
2. A column (Activity) was added to identify by descriptive name which of the six activities were being performed for the measurement. Previously this information was stored in two files – one to identify each row activity by number and another file to match the number to the specific named activity.
3. Several column names were the same although the data was different. For example, rows 303, 317, and 331 were all named “fBodyAcc-bandsEnergy()-1,8”. In order to have unique column names, necessary for subsetting, the column number was added to the beginning of the column name. The data frame now includes, 303\_fBodyAcc-bandsEnergy()-1,8, 317\_fBodyAcc-bandsEnergy()-1,8, and 331\_fBodyAcc-bandsEnergy()-1,8.
4. For statistical purposes, the subject data was randomly divided into two separate categories, train and test. This created two separate data frames. These have now been combined into one data frame including all data.