# Getting Clean Data Project Code book

This code book describes the variables, the data, and any transformations or work performed to clean up the data.

## Study Design

+ The data come from the accelerometers on a Samsung Galaxy S II smartphone. It recorded various accelerometer measurements for 30 subjects during different prescribed activities. More information on the experiment can be found at: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

+ The data for the project is downloaded from: https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

+ This zip file has been saved in the working directory by the run\_Analysis.R script (analysis script) as ProjectData.zip

+ This file is unzipped in the working directory and creates the subdirectory 'UCI HAR Dataset' which contains all the raw data and a README.txt which describes the dataset.

+ The raw data is separated into test and training datasets. Within each of the test and training datasets the data are separated into files for measurement data, the activity labels, and the test subject labels.

## Code Book

The variables and processes will be outlined by the requirement steps for run\_Analysis.R

### 1. Merges the training and the test sets to create one data set.

The files for the test and training datasets raw data, activity labels and test subjects are imported into the following dataframes

+ test\_data

+ test\_labels

+ test\_subjects

+ train\_data

+ train\_labels

+ train\_subjects

Each corresponding datafile (data, labels, subjects) are combined by stacking the test and train data together in the same order for each to preserve matched rows across all combined datasets

+ combine\_data

+ combine\_labels

+ combine\_subjects

### 2. Extracts only the measurements on the mean and standard deviation for each measurement.

This step imports the variable names for measurement data. It finds the only the columns that contain variables that are the mean and standard deviation of measurements and subsequently subsets the combined data to only these variables.

+ features - This variables imports the dataset containing the measurement variable names

+ meanstdcols - Finds the positions in features which only contain variables that are for mean and standard deviation for measurements

+ dataMeanStd - subset of combine\_data with only the columns of variables that mean and standard deviations of measurements

### 3. Uses descriptive activity names to name the activities in the data set

The activity labels dataset is imported which contains the descriptive names of the activities linked with the numeric code for the activity. This is joined with the combine\_labels dataset and the output is a single variable dataset with descriptive activity labels for each row of the dataset.

+ activitynames - imported dataset with linking numeric coded activity with descriptive name

+ activities - output dataset that contains descriptively named activities for each row of the dataset

### 4. Appropriately labels the data set with descriptive variable names.

The measurement variable names for the mean and standard deviation variables are extracted from the features dataset and used to name the columns of the dataMeanStd data subset. The names are modified slightly for readability. The subject variable is also named and the subject, activites and dataMeanStd datasets are finally merged by binding all the columns together. The row order of these datasets have not changed so the rows in each of these datasets should still correspond to each other.

+ dataLabels - Variable labels for the dataMeanStd dataset. mean() and std() are changed to Mean and SD for readability. Hyphens are left in the variable names for readability.

+ combineall - Final dataset with descriptive variable names combining the columns for the combine\_subjects, activities, and dataMeanStd dataframes.

### 5. From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.

The combineall dataset is sorted by subject and activity for readability. Then a dataset is output that gives the Mean for each measurement variable for each Subject+Activity combination. Each measurement variable name is prefaced with "Mean\_" to indicate such.

+ meansummary - Outputs a dataset sorted by subject and activity with the mean for measurement variable for each combination of subject+activity.