**Code Book – Getting and Cleaning Data – Course Project**

**Source Data**

Data for this project was sourced from:

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

***Extract from Source Read Me:***

==================================================================

Human Activity Recognition Using Smartphones Dataset

Version 1.0

==================================================================

Jorge L. Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto.

Smartlab - Non Linear Complex Systems Laboratory

DITEN - Università degli Studi di Genova.

Via Opera Pia 11A, I-16145, Genoa, Italy.

activityrecognition@smartlab.ws

www.smartlab.ws

==================================================================

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain. See 'features\_info.txt' for more details.

For each record it is provided:

======================================

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.

- Triaxial Angular velocity from the gyroscope.

- A 561-feature vector with time and frequency domain variables.

- Its activity label.

- An identifier of the subject who carried out the experiment.

The dataset includes the following files:

=========================================

- 'README.txt'

- 'features\_info.txt': Shows information about the variables used on the feature vector.

- 'features.txt': List of all features.

- 'activity\_labels.txt': Links the class labels with their activity name.

- 'train/X\_train.txt': Training set.

- 'train/y\_train.txt': Training labels.

- 'test/X\_test.txt': Test set.

- 'test/y\_test.txt': Test labels.

The following files are available for the train and test data. Their descriptions are equivalent.

- 'train/subject\_train.txt': Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.

- 'train/Inertial Signals/total\_acc\_x\_train.txt': The acceleration signal from the smartphone accelerometer X axis in standard gravity units 'g'. Every row shows a 128 element vector. The same description applies for the 'total\_acc\_x\_train.txt' and 'total\_acc\_z\_train.txt' files for the Y and Z axis.

- 'train/Inertial Signals/body\_acc\_x\_train.txt': The body acceleration signal obtained by subtracting the gravity from the total acceleration.

- 'train/Inertial Signals/body\_gyro\_x\_train.txt': The angular velocity vector measured by the gyroscope for each window sample. The units are radians/second.

**Data Background**

Data was downloaded from location advised previously and placed in a movement\_data dir below Home dir.

Source data files were not manipulated in any manner.

*Files used for the project were as follows:*

README - for background only

features\_info – for background only

X\_test and X\_train – these are the primary data files contain movement measurements per feature per activity.

These files contain a total of 10299 rows across 561 variables (measurements). An activity code is also recorded in these files. Project requires only mean and standard deviation measurements to be averaged per subject per activity per feature (variable).

Activity\_labels – 6 – contains the descripts for the codes in the X data files:

1 WALKING

2 WALKING\_UPSTAIRS

3 WALKING\_DOWNSTAIRS

4 SITTING

5 STANDING

6 LAYING

Subject\_test and Subject\_train – contain the subject id (1-30) for each X datafile – 10299 rows. Subjects may have multiple rows in each X data file.

**Data Processing**

All processing done via **run\_analysis.R** scrip supplied.

From the directory structure provided in the zip, the X test and train data files are imported from their respective sub dirs into data frames before being row bound into an interim df – fullx.

The list of features (source: features.txt) are also imported for use as column headings in the fullx df.

Subject data (ids 1-30) are imported are from their respective sub dirs into data frames before being row bound into an interim df – fullsubs.

Once all the X data and subject data from the test and training sets is combined it is then column bound together – fullsubsx df - to enable it to be subsetted for only the required mean and std measurements.

An empty df is created - subsetdf – to house the ‘long’ version of the data with only 4 columns – subject (1-30), features (one of the mean/std), measures (measurement recorded) and activity (one of the 6 described earlier).

A for loop over the column names then searches for mean/std features, if found this data is loaded into the subsetdf, with the data in the measurement column being split between the measurement value and the activity code. As each required column found is row bound into the df. Columns not required as written to xcolsremoved.txt for review.

Once the loop is finished the subsetdf contains only mean and std features with their measurements per activity and subject. Activity codes (1-6) are then tidied up to one of descriptions provided (e.g. 1= WALKING).

The final txt file required is then extracted via an SQL select on the subsetdf – Average measure value per subject per activity per feature (either mean or std).