**Code Book for Getting and Cleaning Data course project**

Human Activity Recognition Using Smartphones Data Set

R Sirmons 24 Aug 2014

NOTE: You may need to use read.table with header=TRUE to view XYtidy.txt

--Overview of received data:

The dataset is from experiments 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, 3-axial linear acceleration and 3-axial angular velocity measurements at a constant rate of 50Hz were captured by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). This data collected as multiple files with each having 2947 observations from the 30 subjects and their activities relating to the sensor and axis with values for each of the 128 windows sampled. This data was summarized into a dataset with 561 variables or ‘features’ that was then split into test and training datasets.

Relevant datasets provided include both a test and training version with 561 variables, or features (X\_test.txt and X\_train.txt), test and train sets specifying the subject for each observation (Y\_test.txt and Y\_train.txt), the feature names (features.txt), and the activity names (activity\_labels.txt). Supporting files of the 128 window inertial measurements for both test and training were provided but not used in this exercise.

--Variables

Since the thrust of this exercise is creating a tidy dataset of the averages of the mean and standard deviations estimates by subject and activity, the relevant variables in the tidy dataset (XYtidy.txt) are:

ActivityId (field [1]): V1 in Y\_test.txt and Y\_train.txt, link to V1 in activity\_labels.txt; id of activity

ActivityName (field [2]): V2 in activity\_labels.txt; description of activity

SubjectId (field [3]): V1 in subject\_test.txt; id for volunteer who conducted test

The following are the averages by subject and activity of the mean or standard deviation estimates. The original names, which were descriptive of axis and type of measurement, were prefixed by “Average of” to differentiate them. The units of the original data were not specified beyond being sensor signals. The units of the below variable are therefore either mean of sensor units per subject per activity or standard deviation of sensor units per subject per activity.

|  |  |
| --- | --- |
| Standard Deviation Averages (fields [4:36]): | Average of\_tBodyAcc-std()-X |
| Average of\_tBodyAcc-std()-Y | Average of\_tBodyAcc-std()-Z |
| Average of\_tGravityAcc-std()-X | Average of\_tGravityAcc-std()-Y |
| Average of\_tGravityAcc-std()-Z | Average of\_tBodyAccJerk-std()-X |
| Average of\_tBodyAccJerk-std()-Y | Average of\_tBodyAccJerk-std()-Z |
| Average of\_tBodyGyro-std()-X | Average of\_tBodyGyro-std()-Y |
| Average of\_tBodyGyro-std()-Z | Average of\_tBodyGyroJerk-std()-X |
| Average of\_tBodyGyroJerk-std()-Y | Average of\_tBodyGyroJerk-std()-Z |
| Average of\_tBodyAccMag-std() | Average of\_tGravityAccMag-std() |
| Average of\_tBodyAccJerkMag-std() | Average of\_tBodyGyroMag-std() |
| Average of\_tBodyGyroJerkMag-std() | Average of\_fBodyAcc-std()-X |
| Average of\_fBodyAcc-std()-Y | Average of\_fBodyAcc-std()-Z |
| Average of\_fBodyAccJerk-std()-X | Average of\_fBodyAccJerk-std()-Y |
| Average of\_fBodyAccJerk-std()-Z | Average of\_fBodyGyro-std()-X |
| Average of\_fBodyGyro-std()-Y | Average of\_fBodyGyro-std()-Z |
| Average of\_fBodyAccMag-std() | Average of\_fBodyBodyAccJerkMag-std() |
| Average of\_fBodyBodyGyroMag-std() | Average of\_fBodyBodyGyroJerkMag-std() |
|  |  |
| Mean Averages(fields [37:82]): |  |
| Average of\_tBodyAcc-mean()-X | Average of\_tBodyAcc-mean()-Y |
| Average of\_tBodyAcc-mean()-Z | Average of\_tGravityAcc-mean()-X |
| Average of\_tGravityAcc-mean()-Y | Average of\_tGravityAcc-mean()-Z |
| Average of\_tBodyAccJerk-mean()-X | Average of\_tBodyAccJerk-mean()-Y |
| Average of\_tBodyAccJerk-mean()-Z | Average of\_tBodyGyro-mean()-X |
| Average of\_tBodyGyro-mean()-Y | Average of\_tBodyGyro-mean()-Z |
| Average of\_tBodyGyroJerk-mean()-X | Average of\_tBodyGyroJerk-mean()-Y |
| Average of\_tBodyGyroJerk-mean()-Z | Average of\_tBodyAccMag-mean() |
| Average of\_tGravityAccMag-mean() | Average of\_tBodyAccJerkMag-mean() |
| Average of\_tBodyGyroMag-mean() | Average of\_tBodyGyroJerkMag-mean() |
| Average of\_fBodyAcc-mean()-X | Average of\_fBodyAcc-mean()-Y |
| Average of\_fBodyAcc-mean()-Z | Average of\_fBodyAcc-meanFreq()-X |
| Average of\_fBodyAcc-meanFreq()-Y | Average of\_fBodyAcc-meanFreq()-Z |
| Average of\_fBodyAccJerk-mean()-X | Average of\_fBodyAccJerk-mean()-Y |
| Average of\_fBodyAccJerk-mean()-Z | Average of\_fBodyAccJerk-meanFreq()-X |
| Average of\_fBodyAccJerk-meanFreq()-Y | Average of\_fBodyAccJerk-meanFreq()-Z |
| Average of\_fBodyGyro-mean()-X | Average of\_fBodyGyro-mean()-Y |
| Average of\_fBodyGyro-mean()-Z | Average of\_fBodyGyro-meanFreq()-X |
| Average of\_fBodyGyro-meanFreq()-Y | Average of\_fBodyGyro-meanFreq()-Z |
| Average of\_fBodyAccMag-mean() | Average of\_fBodyAccMag-meanFreq() |
| Average of\_fBodyBodyAccJerkMag-mean() | Average of\_fBodyBodyAccJerkMag-meanFreq() |
| Average of\_fBodyBodyGyroMag-mean() | Average of\_fBodyBodyGyroMag-meanFreq() |
| Average of\_fBodyBodyGyroJerkMag-mean() | Average of\_fBodyBodyGyroJerkMag-meanFreq() |

--Data transforms to create XYtidy.txt dataset

After reading all data the test data is processed first. First the activity labels are merged with the Y\_test data. Next the features data are applied to the X\_test data as the 561 variable names and friendly names (ActivityId, ActivityName, SubjectId) are applied to the subject data and the activity data. Finally the test data, the subject data and the activity data are combined into one data frame. This processing is then repeated for the training data.

Next the test and training data frames are combined into one data frame. That combined data frame is then filtered to only columns with “mean” or “std” in their names (plus ActivityId, ActivityName, and SubjectId). Then this data frame is aggregated by SubjectId and ActivityId and an average calculated for each data field for each aggregation. The columns are then renamed to prefix all ‘mean’ or ‘std’ columns with “Average of\_” to differentiate them from other variables in the original dataset. The activity labels are then merged with the data frame and the ActivityId, ActivityName, and SubjectId column names reapplied. Finally the ActivityName column is moved to place it next to the ActivityId column and the data frame written to the disk as XYtidy.txt using row.names =FALSE.