Code Book

Get Clean Data Assignment

The purpose of this project was to apply Tidy Data Principles to a source of data (see below), namely:

1. One variable measured per column
2. Each different observation of a variable to be in a different row
3. Variable Names to be Human readable but without spaces, hyphens, underscores and dots.
4. Removal of redundant variables

The data methods used to create a tidy dataset collected from Samsung wearable devices that systematically recorded linear and angular movement over time using R functions. The filename of the Tidy Dataset created is a text file called: “samsungDatasetSummarised.txt”. This created tidy dataset contains the average and standard deviation for a range of linear and angular distances, velocities and accelerations for 30 different subjects each having performed each of the six activities. The dataset consists of 180 rows by 88 columns. The variables names are listed in a Table below.

The R script file name used to create this dataset file :

run\_analysis.R.

Note: R version version used : 3.1.2 (2014-10-31) using Windows 7 Operating System 32 bit

The script requires the dplyr package. The script requires to be in the same directory as the source data from the unzipped data documented below.

The dataset used to create the datset was obtained “samsungDatasetSummarised.txt” was created from the following data source:data:

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

The files used from this zip file are listed n the table below:

|  |  |  |
| --- | --- | --- |
| File Name | Contents | Dimension |
| activity\_labels.txt | Code file to enable descriptive names for activity | 6:2 |
| features.txt | Vector containing measured variable names | c |
| ./test/X\_test.txt | Matrix of Measured variables of subjects doing an activity during the test component | r by c |
| ./test/Subject\_test.txt | Subject identification from each measurement and activity being done by the subject in Test Component of the study | r |
| ./test/Y\_test.txt | Activity of subject at the time when measurements were being made during the test component. |  |
| ./train/X\_train.txt | Matrix of Measured variables of subjects doing an activity during the training component. |  |
| ./train/Subject\_train.txt | Subject identification from each measurement and activity being done by the subject in Training Component of the study. |  |
| ./train/Y\_train.txt | Activity of subject at the time when measurements were being made during the training component. |  |

Note: Inertial Signal datafiles were not used.

Details about Source Data: There were two sources of subjects whose activities were measured: those from the Test Study Component (9 subjects) and subjects from the Training Component (21 subjects). The linear and angular distances, velocities and accelerations were measure for each of the 6 different activities each subject did over time.

The data from each of these Components were aggregated together resulting .

The steps used to generate “samsungDatasetSummarised.txt” dataset are as follows:

1. Preliminary Steps
   1. Read in activity labels and assign understandable variable names
   2. Read in vector containing original Variable Names
   3. Adjust Variable names by removing parentheses and commas
   4. Then create vector with variable names representing means
   5. Then create vector with variable names representing stds
   6. Combine vectors of variable names
2. Create Dataset from Training Component data
   1. # Read in X\_train.txt dataset
   2. # Select only those columns which have Mean or Std in their name
   3. # Read in subject training Data
   4. # Read in y-train data and allocate variable name to activityCode
   5. # Combine columns for subjectTrain, yTrain and XTrainDataset dataframes for Training Data
3. Repeat steps of Training Dataset for the Test Componet Data
   1. Read in X\_test.txt dataset
   2. Read in subject test Data
   3. Read in y-test data and allocate variable name activityCode
   4. Combine columns for subjectTest, yTest and xTestDataset dataframes for test Dataset
   5. Combine trainDataset and testDataset together using rbind function
4. Created New Variable List
   1. Replace old variable list with new Variable List
5. Then Merge activity Labels into samsungDataset
   1. Examine distribution of subjects by activityCode and by activity
6. Create summarised samsung dataset where the mean of each variable is calculated for each subject-activity level by first using group\_by function **f**ollowed by using the summarise\_each function from dplyr to obtain the average of each of the mean and standard deviation variables
   1. Drop activityCode variable as it is redundant
   2. Make subject variable a factor
   3. Remove dots from variable names as per Tidy Data specifications
7. Write samsungDatasetSummarised dataset to text file in working directory

The code for each of these steps is listed in run\_analysis.R file.

The resulting dataset created was a single table containing 86 variables of measured average linear and angular velocities (along with measures of their average standard deviation) variables from each of the six activities for each of the 30 subjects. Each variable was normalised between -1 and 1 and are therefore dimensionless.

The Variable List of “samsungDatasetSummarised.txt”

[1] "subject" "activity"

[3] "tBodyAccXmean" "tBodyAccYmean"

[5] "tBodyAccZmean" "tGravityAccXmean"

[7] "tGravityAccYmean" "tGravityAccZmean"

[9] "tBodyAccJerkXmean" "tBodyAccJerkYmean"

[11] "tBodyAccJerkZmean" "tBodyGyroXmean"

[13] "tBodyGyroYmean" "tBodyGyroZmean"

[15] "tBodyGyroJerkXmean" "tBodyGyroJerkYmean"

[17] "tBodyGyroJerkZmean" "tBodyAccMagmean"

[19] "tGravityAccMagmean" "tBodyAccJerkMagmean"

[21] "tBodyGyroMagmean" "tBodyGyroJerkMagmean"

[23] "fBodyAccXmean" "fBodyAccYmean"

[25] "fBodyAccZmean" "fBodyAccFreqXmean"

[27] "fBodyAccFreqYmean" "fBodyAccFreqZmean"

[29] "fBodyAccJerkXmean" "fBodyAccJerkYmean"

[31] "fBodyAccJerkZmean" "fBodyAccJerkFreqXmean"

[33] "fBodyAccJerkFreqYmean" "fBodyAccJerkFreqZmean"

[35] "fBodyGyroXmean" "fBodyGyroYmean"

[37] "fBodyGyroZmean" "fBodyGyroFreqXmean"

[39] "fBodyGyroFreqYmean" "fBodyGyroFreqZmean"

[41] "fBodyAccMagmean" "fBodyAccMagFreqmean"

[43] "fBodyBodyAccJerkMagmean" "fBodyBodyAccJerkMagFreqmean"

[45] "fBodyBodyGyroMagmean" "fBodyBodyGyroMagFreqmean"

[47] "fBodyBodyGyroJerkMagmean" "fBodyBodyGyroJerkMagFreqmean"

[49] "angletBodyAccMeangravity" "angletBodyAccJerkMeangravitymean"

[51] "angletBodyGyroMeangravitymean" "angletBodyGyroJerkMeangravitymean"

[53] "angleXgravitymean" "angleYgravitymean"

[55] "angleZgravitymean" "tBodyAccXstd"

[57] "tBodyAccYstd" "tBodyAccZstd"

[59] "tGravityAccXstd" "tGravityAccYstd"

[61] "tGravityAccZstd" "tBodyAccJerkXstd"

[63] "tBodyAccJerkYstd" "tBodyAccJerkZstd"

[65] "tBodyGyroXstd" "tBodyGyroYstd"

[67] "tBodyGyroZstd" "tBodyGyroJerkXstd"

[69] "tBodyGyroJerkYstd" "tBodyGyroJerkZstd"

[71] "tBodyAccMagstd" "tGravityAccMagstd"

[73] "tBodyAccJerkMagstd" "tBodyGyroMagstd"

[75] "tBodyGyroJerkMagstd" "fBodyAccXstd"

[77] "fBodyAccYstd" "fBodyAccZstd"

[79] "fBodyAccJerkXstd" "fBodyAccJerkYstd"

[81] "fBodyAccJerkZstd" "fBodyGyroXstd"

[83] "fBodyGyroYstd" "fBodyGyroZstd"

[85] "fBodyAccMagstd" "fBodyBodyAccJerkMagstd"

[87] "fBodyBodyGyroMagstd" "fBodyBodyGyroJerkMagstd"

Go to: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

to obtain more information about variables if required.