# Code Book for Getting and Cleaning Data course on Coursera

This CodeBook that describes the variables, the data, and summaries that were used in run\_analysis.R script.

## Information about Source Data Experiment

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 were selected for generating the training data and 30% the test data.

## Original Data Source

Human Activity Recognition Using Smartphones Dataset Data for analysis is downloaded from the below URL: <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

## Structure of source Data

|  |  |
| --- | --- |
| **File Name** | **Details** |
| README.txt | Details of all the files in the downloaded folder |
| features\_info.txt | Details about variables used in the feature vector |
| features.txt | List of all features (i.e. measurement variables) |
| activity\_lables.txt | List of activity ID and corresponding activity name |
| train/X\_train.txt | Training data set |
| train/Y\_train.txt | Training activity ID labels |
| train/subject\_train.txt | Each row identifies the subject who performed the activity. Range 1 to 30 |
| test/X-test.txt | Test data set |
| test/Y\_test.txt | Test activity ID labels |
| test/subject\_train.txt | Each row identifies the subject who performed the activity. Range 1 to 30 |

Note: All the files in train/Inertial Signals and test/Inertial Signals folders will not be used for in this analysis

## Details about source data Files

|  |  |  |  |
| --- | --- | --- | --- |
| **File Name** | **# Rows** | **# Variables** | **Comments** |
| features.txt | 561 | 2 | Feature ID and feature Name |
| activity\_lables.txt | 6 | 2 | Activity ID and activity Name |
| X\_test.txt | 2947 | 561 | Values for measurements |
| y\_test.txt | 2947 | 1 | Activity ID per each observation |
| subject\_test.txt | 2497 | 1 | Subject ID per each observation |
| X\_train.txt | 7352 | 561 | Values for measurements |
| y\_train.txt | 7352 | 1 | Activity ID per each observation |
| subject\_train.txt | 7352 | 1 | Subject ID per each observation |

## Course assignment and requirements for run\_analysis.R script

1. Merge the training and the test sets to create one data set
2. Extract only the measurements on the mean and standard deviation for each measurement
3. Use descriptive activity names to name the activities in the data set
4. Appropriately label the data set with descriptive activity names
5. Create a second, independent tidy data set with the average of each variable for each activity and each subject

## Merge the training and the test sets to create one data set and

## 3 Use description activity names to name the activities in the data set

1. Make sure that **dplyr** and **reshape2** R packages are installed. Install them if they are not already installed.
2. Download the zip file from the URL mentioned above and unzip it to create UCI HAR Dataset folder.
3. Script reads and create factors common for both test and train datasets

|  |  |  |
| --- | --- | --- |
| **source file** | **factor** | **length** |
| activity\_labels.txt | activityLabels | 6 |
| features.txt | columnNames | 561 |

1. Script reads Train and Test datasets and creates data frames

|  |  |  |
| --- | --- | --- |
| **source file** | **data.frame** | **details** |
| X\_train.txt | x\_train | 7352 obs. of 561 variable |
| X\_test.txt | x\_test | 2947 obs. Of 561 variable |
| y\_train.txt | y\_train | 7352 obs. of 1 variable |
| y\_test.txt | y\_test | 2947 obs. Of 1 variable |
| subject\_train.txt | trainSubjects | 7352 obs. Of 1 variable |
| subject\_test.txt | testSubjects | 2947 obs. Of 1 variable |

1. Script creates additional variable called Set to designate if an observation came from Test or Training dataset
2. Script merges x\_train, y\_train , subject\_train and set data.frames using cbind function to form single train data.frame (7352 obs. of 564 variable)
3. Script merges x\_test, y\_test, subject\_test and set data.frames using cbind function to form single test data.frame (2947 obs. Of 561 variable)
4. Script merges train and test data.frames to form a single data.frame. The **rbind** function was used.

|  |  |
| --- | --- |
| **data.frame** | **Details** |
| data | 10299 obs. Of 564 variable |

1. Script removes all the temporary variables (set,train, test, tmp, x\_train, x\_test, y\_train, y\_test, trainSubjects, testSubjects, activityLabels, columnNames))

## Extract only the measurements on the mean and standard deviation for each measurement

1. According to the assignment script then extracts only the measurements on the mean and standard deviation for each measurement. **grep** function is used to search for occurrences of “mean”,” std”, “activity”, “set” in the names of data data.frame columns. Then the same function is used to exclude the occurrences of “meanFreq”. The resulting selecting gives us only 69 variables.

## Appropriately label the data set with descriptive activity names

1. gsub function and regular expressions were used to modify names of the variables. See the code below:

filtered\_names <- names(data)

filtered\_names <- gsub("\\(\\)", "", filtered\_names)

filtered\_names <- gsub("Acc", "-acceleration", filtered\_names)

filtered\_names <- gsub("^t(.\*)$", "\\1-time", filtered\_names)

filtered\_names <- gsub("^f(.\*)$", "\\1-frequency", filtered\_names)

filtered\_names <- gsub("(Jerk|Gyro)", "-\\1", filtered\_names)

filtered\_names <- gsub("BodyBody", "Body", filtered\_names)

filtered\_names <- gsub("mag", "-magnitude", filtered\_names)

filtered\_names <- tolower(filtered\_names)

names(data) <- filtered\_names

## 5 Create a second, independent tidy data set with the average of each variable for each activity and each subject

1. **reshape2** package was used to reshape the resulting data data.frame
2. **melt** function was used to reshape the data.frame utilizing these variables as ids: subject, activity, set. Resulting **melt\_data** data.frame has 679734 observations and 5 variable
3. dcast function was used to transform long-format data into wide format data. The resulting tidy\_data data.frame has 180 observations and 8 variables.
4. First two variables (subject and activity are id variables) and remaining 66 are measurement variables.
5. Each measurement variable [3 to 68] shows average value for each combination of subject and activity

## Saving results into tidy\_data.txt file

According to the assignment the **write.table** function was used with row.names = FALSE

## Description of variables in the Tidy Data set

|  |  |
| --- | --- |
| **Variable Name** | **Details** |
| subject | Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30 |
| activity | Activity label: WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING |
|  |  |
| body-acceleration-mean-x-time | Average of mean value time acceleration in x direction |
| body-acceleration-mean-y-time | Average of mean value time acceleration in y direction |
| body-acceleration-mean-z-time | Average of mean value time acceleration in z direction |
| body-acceleration-std-x-time | Average of standard deviation time acceleration in x direction |
| body-acceleration-std-y-time | Average of standard deviation time acceleration in y direction |
| body-acceleration-std-z-time | Average of standard deviation time acceleration in z direction |
| gravity-acceleration-mean-x-time | Average of mean value time domain gravity acceleration x direction |
| gravity-acceleration-mean-y-time | Average of mean value time domain gravity acceleration y direction |
| gravity-acceleration-mean-z-time | Average of mean value time domain gravity acceleration z direction |
| gravity-acceleration-std-x-time | Average of standard deviation time domain gravity acceleration x direction |
| gravity-acceleration-std-y-time | Average of standard deviation time domain gravity acceleration y direction |
| gravity-acceleration-std-z-time | Average of standard deviation time domain gravity acceleration z direction |
| body-acceleration-jerk-mean-x-time | Average of mean value time acceleration jerk in x direction |
| body-acceleration-jerk-mean-y-time | Average of mean value time acceleration jerk in y direction |
| body-acceleration-jerk-mean-z-time | Average of mean value time acceleration jerk in z direction |
| body-acceleration-jerk-std-x-time | Average of standard deviation time acceleration jerk in x direction |
| body-acceleration-jerk-std-y-time | Average of standard deviation time acceleration jerk in y direction |
| body-acceleration-jerk-std-z-time | Average of standard deviation time acceleration jerk in z direction |
| body-gyro-mean-x-time | Average of mean value time gyro in x direction |
| body-gyro-mean-y-time | Average of mean value time gyro in y direction |
| body-gyro-mean-z-time | Average of mean value time gyro in z direction |
| body-gyro-std-x-time | Average of standard deviation time gyro in x direction |
| body-gyro-std-y-time | Average of standard deviation time gyro in y direction |
| body-gyro-std-z-time | Average of standard deviation time gyro in z direction |
| body-gyro-jerk-mean-x-time | Average of mean value time gyro jerk signal in x direction |
| body-gyro-jerk-mean-y-time | Average of mean value time gyro jerk signal in y direction |
| body-gyro-jerk-mean-z-time | Average of mean value time gyro jerk signal in z direction |
| body-gyro-jerk-std-x-time | Average of standard deviation time gyro jerk signal in x direction |
| body-gyro-jerk-std-y-time | Average of standard deviation time gyro jerk signal in y direction |
| body-gyro-jerk-std-z-time | Average of standard deviation time gyro jerk signal in z direction |
| body-acceleration-magnitude-mean-time | Average of mean value time acceleration magnitude |
| body-acceleration-magnitude-std-time | Average of standard deviation time acceleration magnitude |
| gravity-acceleration-magnitude-mean-time | Average of mean value time domain gravity acceleration magnitude |
| gravity-acceleration-magnitude-std-time | Average of standard deviation time domain gravity acceleration magnitude |
| body-acceleration-jerk-magnitude-mean-time | Average of mean value time acceleration jerk magnitude |
| body-acceleration-jerk-magnitude-std-time | Average of standard deviation time acceleration jerk magnitude |
| body-gyro-magnitude-mean-time | Average of mean value time gyro magnitude |
| body-gyro-magnitude-std-time | Average of standard deviation time gyro magnitude |
| body-gyro-jerk-magnitude-mean-time | Average of mean value time gyro jerk magnitude |
| body-gyro-jerk-magnitude-std-time | Average of standard deviation time gyro jerk magnitude |
| body-acceleration-mean-x-frequency | Average of mean value frequency acceleration in x direction |
| body-acceleration-mean-y-frequency | Average of mean value frequency acceleration in y direction |
| body-acceleration-mean-z-frequency | Average of mean value frequency acceleration in z direction |
| body-acceleration-std-x-frequency | Average of standard deviation frequency acceleration in x direction |
| body-acceleration-std-y-frequency | Average of standard deviation frequency acceleration in y direction |
| body-acceleration-std-z-frequency | Average of standard deviation frequency acceleration in z direction |
| body-acceleration-jerk-mean-x-frequency | Average of mean value frequency acceleration jerk in x direction |
| body-acceleration-jerk-mean-y-frequency | Average of mean value frequency acceleration jerk in y direction |
| body-acceleration-jerk-mean-z-frequency | Average of mean value frequency acceleration jerk in z direction |
| body-acceleration-jerk-std-x-frequency | Average of standard deviation frequency acceleration jerk in x direction |
| body-acceleration-jerk-std-y-frequency | Average of standard deviation frequency acceleration jerk in y direction |
| body-acceleration-jerk-std-z-frequency | Average of standard deviation frequency acceleration jerk in z direction |
| body-gyro-mean-x-frequency | Average of mean value frequency gyro in x direction |
| body-gyro-mean-y-frequency | Average of mean value frequency gyro in y direction |
| body-gyro-mean-z-frequency | Average of mean value frequency gyro in z direction |
| body-gyro-std-x-frequency | Average of standard deviation frequency gyro in x direction |
| body-gyro-std-y-frequency | Average of standard deviation frequency gyro in y direction |
| body-gyro-std-z-frequency | Average of standard deviation frequency gyro in z direction |
| body-acceleration-magnitude-mean-frequency | Average of mean value frequency acceleration magnitude |
| body-acceleration-magnitude-std-frequency | Average of standard deviation frequency acceleration magnitude |
| body-acceleration-jerk-magnitude-mean-frequency | Average of mean value frequency acceleration jerk magnitude |
| body-acceleration-jerk-magnitude-std-frequency | Average of standard deviation frequency acceleration jerk magnitude |
| body-gyro-magnitude-mean-frequency | Average of mean value frequency body gyro magnitude |
| body-gyro-magnitude-std-frequency | Average of standard deviation frequency body gyro magnitude |
| body-gyro-jerk-magnitude-mean-frequency | Average of mean value frequency body gyro jerk magnitude |
| body-gyro-jerk-magnitude-std-frequency | Average of standard deviation frequency body gyro jerk magnitude |