Codebook

# Overview

This codebook describes how to use the data in the averages\_dataset.txt file (referred to as the *averages dataset* for the remainder of this document) and describes how it was created.

The *averages dataset* can be loaded into R using the following line of code:

avg\_ds <- read.table("averages\_dataset.txt", header = TRUE, stringsAsFactors = FALSE)

To recreate the *averages dataset*, run the run\_analysis.R script. The script will first download and unzip the raw data to the working directory (if not already present), clean and reshape the data, and write the resulting dataset to the file averages\_dataset.txt.

The script also saves the intermediate tidy set from which the averages are extracted to the file tidy\_dataset.txt. To load this dataset, use the following code:

tds <- read.table("tidy\_dataset.txt", header = TRUE, stringsAsFactors = FALSE)

# Study Design

This section describes how the data was obtained and modified.

## 2.1 Raw Data

The raw data made available for the course was the [*Human Activity Recognition Using Smartphones Data Set*](http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones). (The raw data files were obtained by downloading and unzipping this [zip file](https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip).)

The raw data is the result of 30 subjects performing six activities (walking, sitting, standing, etc.) while wearing a smartphone. The tri-axial acceleration data captured from the accelerometer and tri-axial angular velocity captured from the gyroscope were used to derive 561 measurements (referred to as “features”) with time and frequency domain variables.

The data was spread across multiple files in two directories, one directory with the datasets used to generate training data (train) and the other used to generate test data (test).

The raw data was made up of the following files:

| **File** | **Dimensions** | **Columns** | **Notes** |
| --- | --- | --- | --- |
| activity\_labels.txt | 6 rows,  2 columns | V1 (integer),  V2 (factor) | Associates descriptions (V2), with each activity code (V1) |
| features.txt | 561 rows,  2 columns | V1 (integer),  V2 (factor) | Each row represents a variable definition |
| train/  subject\_train.txt | 7352 rows,  1 column | V2 (integer) | Each row indicates the subject who performed the observation in the corresponding line of X\_train.txt |
| train/  X\_train.txt | 7352 rows,  561 columns | V1:V561 (numeric) | Each column represents the variables of an observation. Each variable corresponds to one of the variable definitions described in features.txt |
| train/  y\_train.txt | 7352 rows,  1 columns | V1 (integer) | Each row indicates the activity carried out for the observation in the corresponding line of X\_train.txt |
| test/  subject\_test.txt | 2947 rows,  1 columns | V1(integer) | Each row indicates the subject who performed the observation in the corresponding line of X\_test.txt |
| test/  X\_test.txt | 2947 rows,  561 columns | V1:V561 (numeric) | Each column represents the variables of an observation. Each variable corresponds to one of the variable definitions described in features.txt |
| test/  y\_test.txt | 2947 rows,  1 column | V1 (integer) | Each row indicates the activity carried out for the observation in the corresponding line of X\_test.txt |

Notes regarding the raw data:

* The raw data consisted of 3-axial measurements (i.e. having X, Y, and Z values) captured by an accelerometer (tAcc-XYZ) and a gyroscope (tGyro-XYZ) and consistent time intervals.
* The raw signals were filtered to remove noise. The accelerometer signal (tAcc-XYZ) was then separated into body (tBodyAcc-XYZ) and gravity (tGravityAcc-XYZ) acceleration signals.
* ‘Jerk’ values (changes in acceleration) were calculated by taking the derivative of the acceleration with respect to time.
* All measurement names beginning with ‘f’ represent frequency domain signals; those beginning with 't' represent time domain signals.
* See the README.txt file included with the raw data for more details.

## 2.2 Tidy Data

This section describes the steps that are carried out by the run\_analysis.R script to create the *averages dataset* from the raw data.

The tidy dataset is laid out following the principles described by Hadley Wickham in his paper [*Tidy Data*](http://vita.had.co.nz/papers/tidy-data.pdf), along with added insight provided by David Hood in his [*Getting and Cleaning Assignment*](https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-the-assignment/) blog post.

### Merging training and test sets

Per the course instructions, the script merges the training and test sets to create a single dataset.

Given that each of the two sets is spread across multiple files (subject\_\*, X\_\*, and y\_\*), I first followed the instructions in section 3.5 of Hadley Wickham’s [*Tidy Data*](http://vita.had.co.nz/papers/tidy-data.pdf) paper (*One type in multiple tables*) which calls for reading the files into a list of tables split by source file. A new column (named set) tracks the set an observation came from (and the source file name is used to determine its value: TEST or TRAIN). The two sets of data are then merged (using rbind).

I consider the 561 features to be *column headers used as values, instead of variable names* (as described in section 3.1 of the *Tidy Data* paper. I followed the instructions to melt the dataset and convert it from a wide format (10,299 rows by 564 columns) to a long format (5,777,739 rows by 5 columns).

### Summary choices

Only measurements on the mean and standard deviation for each measurement were extracted, per step #2 of the course project instructions.

* These are determined by searching for features that contain “mean()” or “std()” in their labels.
* Features derived from calls to “meanFreq()” are not included.

Per step #5 of the course project instructions, the *averages dataset* only includes the average for each of the mean/std measurements per subject and activity

### Variable Names

Variable names follow Hadley Wickham's [Style Guide](http://adv-r.had.co.nz/Style.html) for R programming: lowercase letters with underscores used to separate words.

### Factors

Per step #3 of the course project instructions, the activity ids are substituted with descriptive activity names (taken from the activity\_labels.txt file included with the raw data).

Per step #4 of the course project instructions, the V1:V561 measurement names are substituted with more descriptive variable names (taken from the features.txt file included with the raw data). To be consistent with the naming convention, the parentheses are removed from the variable names and hyphens are replaced with underscores.

## Variables not contained in the *averages dataset*

(This section has been included in the codebook following the guidance provided by JT Leek in [*How to share data with a statistician*](https://github.com/jtleek/datasharing).)

Per steps #2 and #5 of the [course project instructions](https://www.coursera.org/learn/data-cleaning/peer/FIZtT/getting-and-cleaning-data-course-project), only the features derived from the mean and standard deviation are included in the final dataset. For a description of the remaining variables, see the features.txt file included with the raw data.

# 3. The *averages dataset*

The *averages dataset* has 180 rows and 68 columns.

* The 180 rows are the result of 30 subjects performing 6 activities (30 x 6 = 180).
* The 68 columns are the result of 2 fixed variables plus the averages of 66 mean/std measurements (2 + 66 = 68)

Each row in the *averages dataset* contains the following fixed variables:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Description** | **Possible Values/Units** |
| subject\_id | Discrete | Identifies the subject who performed the activity | 1-30 |
| activity | Categorical | Identifies the activity that was performed | WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING |

The remaining variables contain the average value of a measurement per subject and activity. As in the original raw data files:

* All values are continuous and normalized/bounded within [-1,1]
* All measurement names beginning with ‘f’ represent frequency domain signals; those beginning with 't' represent time domain signals.

| **Variable name** | **Description** | **Units** |
| --- | --- | --- |
| fBodyAcc\_mean\_X  fBodyAcc\_mean\_Y  fBodyAcc\_mean\_Z | Average value of the body linear acceleration | standard gravity units (g)  (9.8 m/s2) |
| fBodyAcc\_std\_X  fBodyAcc\_std\_Y  fBodyAcc\_std\_Z | Standard deviation of the body linear acceleration | standard gravity units (g) |
| fBodyAccJerk\_mean\_X  fBodyAccJerk\_mean\_Y  fBodyAccJerk\_mean\_Z | Average value of the changes in acceleration (or Jerk) | gravity units per second  (9.8 m/s3) |
| fBodyAccJerk\_std\_X  fBodyAccJerk\_std\_Y  fBodyAccJerk\_std\_Z | Standard deviation of the changes in acceleration (or Jerk) | gravity units per second |
| fBodyAccMag\_mean | Average magnitude of the body linear acceleration | meters |
| fBodyAccMag\_std | Standard deviation of the magnitude of the body linear acceleration | meters |
| fBodyBodyAccJerkMag\_mean | Average magnitude of body jerk measurements | meters |
| fBodyBodyAccJerkMag\_std | Standard deviation of magnitude of body jerk | meters |
| fBodyBodyGyroJerkMag\_mean | Average magnitude of jerks in angular velocity | meters |
| fBodyBodyGyroJerkMag\_std | Standard deviation of magnitudes of jerks in angular velocity | meters |
| fBodyBodyGyroMag\_mean | Average magnitude of angular velocity | meters |
| fBodyBodyGyroMag\_std | Standard deviation of the magnitude of angular velocity | meters |
| fBodyGyro\_mean\_X  fBodyGyro\_mean\_Y  fBodyGyro\_mean\_Z | Average body angular velocity | radians/second |
| fBodyGyro\_std\_X  fBodyGyro\_std\_Y  fBodyGyro\_std\_Z | Standard deviation of the body angular velocity | radians/second |
| tBodyAcc\_mean\_X  tBodyAcc\_mean\_Y  tBodyAcc\_mean\_Z | Average body linear acceleration | standard gravity units (g) |