CodeBook - UCI HAR Tidy Datasets

This is a code book that describes the variables, the data, and any transformations or work performed to clean up the data.

Getting the data

The data was downloaded via Safari browser from https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip and unzipped automatically to a folder named "UCI HAR Dataset". The folder was renamed to "UCI_HAR_Dataset" for convenience.

Merging the Data

1. The training and text data for X (data values) are merged in the following order:
    - train/X_train.txt
    - test/X_test.txt
2. The training and text data for y (activities) are merged in the following order:
    - train/y_train.txt
    - test/y_test.txt
3. The training and text data for subjects are merged in the following order:
    - train/subject_train.txt
    - test/subject_test.txt

Meaningful Names

The data in the y and test files is named 'Activity' and the subject data is named 'Subject'. The x column labels are read in from the file **features.txt** and are cleaned up to make them easier to work with.

The substitutions are:

Find	Replace	Reason
open bracket	empty	Cleaner for working with column names
close bracket	empty	Cleaner for working with column names
comma	underscore	Cleaner for working with column names
dash	underscore	Cleaner for working with column names
meanFreq	MeanFreq	So as not to pick up when using grep 'mean' for column names
mean_X	X_mean	For easier sorting
mean_Y	Y_mean	For easier sorting
mean_Z	Z_mean	For easier sorting
std_X	X_std	For easier sorting
std_Y	Y_std	For easier sorting
std_Z	Z_std	For easier sorting

Descriptive activity names are used to name the activities in the data set. These are hard-coded from the information contained in the file activity_labels.txt, as follows:

Value	Substitution
1	WALKING
2	WALKING_UPSTAIRS
3	WALKING_DOWNSTAIRS
4	SITTING
5	STANDING
6	LAYING

Tidy dataset 1 - dataExtract

Measurements on the mean and standard deviation for each measurement are extracted from the data frame named **x** to create a data frame named **dataExtract**. The output file is **UCI_HAR_DataExtract.txt**. All the variables are numerical apart from Activity and Subject:

```
Activity
Factor w/ 6 levels "WALKING","WALKING_UPSTAIRS",..: 5 5 5 5 5 5 5 5 5 5 ...

Subject
int [1:10299] 1 1 1 1 1 1 1 1 1 1 ...
```

The numerical variable names are:

```
[1]  "fBodyAcc_X_mean"           "fBodyAcc_X_std"            "fBodyAcc_Y_mean"
[4]  "fBodyAcc_Y_std"            "fBodyAcc_Z_mean"           "fBodyAcc_Z_std"
[7]  "fBodyAccJerk_X_mean"       "fBodyAccJerk_X_std"        "fBodyAccJerk_Y_mean"
[10] "fBodyAccJerk_Y_std"        "fBodyAccJerk_Z_mean"       "fBodyAccJerk_Z_std"
[13] "fBodyAccMag_mean"          "fBodyAccMag_std"           "fBodyBodyAccJerkMag_mean"
[16] "fBodyBodyAccJerkMag_std"   "fBodyBodyGyroJerkMag_mean" "fBodyBodyGyroJerkMag_std"
[19] "fBodyBodyGyroMag_mean"     "fBodyBodyGyroMag_std"      "fBodyGyro_X_mean"
[22] "fBodyGyro_X_std"           "fBodyGyro_Y_mean"          "fBodyGyro_Y_std"
[25] "fBodyGyro_Z_mean"          "fBodyGyro_Z_std"           "tBodyAcc_X_mean"
[28] "tBodyAcc_X_std"            "tBodyAcc_Y_mean"           "tBodyAcc_Y_std"
[31] "tBodyAcc_Z_mean"           "tBodyAcc_Z_std"            "tBodyAccJerk_X_mean"
[34] "tBodyAccJerk_X_std"        "tBodyAccJerk_Y_mean"       "tBodyAccJerk_Y_std"
[37] "tBodyAccJerk_Z_mean"       "tBodyAccJerk_Z_std"        "tBodyAccJerkMag_mean"
[40] "tBodyAccJerkMag_std"       "tBodyAccMag_mean"          "tBodyAccMag_std"
[43] "tBodyGyro_X_mean"          "tBodyGyro_X_std"           "tBodyGyro_Y_mean"
[46] "tBodyGyro_Y_std"           "tBodyGyro_Z_mean"          "tBodyGyro_Z_std"
[49] "tBodyGyroJerk_X_mean"      "tBodyGyroJerk_X_std"       "tBodyGyroJerk_Y_mean"
[52] "tBodyGyroJerk_Y_std"       "tBodyGyroJerk_Z_mean"      "tBodyGyroJerk_Z_std"
[55] "tBodyGyroJerkMag_mean"     "tBodyGyroJerkMag_std"      "tBodyGyroMag_mean"
[58] "tBodyGyroMag_std"          "tGravityAcc_X_mean"        "tGravityAcc_X_std"
[61] "tGravityAcc_Y_mean"        "tGravityAcc_Y_std"         "tGravityAcc_Z_mean"
[64] "tGravityAcc_Z_std"         "tGravityAccMag_mean"       "tGravityAccMag_std"
[67] "Activity"                  "Subject"
```

Tidy dataset 2 - dataSummary

A second, independent tidy data set is created with the average of each variable for each activity and each subject. This data frame is named dataSummary and output file is **UCI_HAR_DataSummary.txt**.

The reshape2 library melt function was used to transform the data into a long narrow format. This resulted in 4 variables:

```
'data.frame':    679734 obs. of  4 variables:
 $ Subject : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Activity: Factor w/ 6 levels "WALKING","WALKING_UPSTAIRS",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ variable: Factor w/ 66 levels "fBodyAcc_X_mean",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ value   : num  -0.995 -0.997 -0.994 -0.995 -0.997 ...
```

Next, the with and tapply functions were used to create a multi-dimensional array for Activity, Subject and variable.

```
 num [1:6, 1:30, 1:66] -0.2028 -0.4043 0.0382 -0.9796 -0.9952 ...
 - attr(*, "dimnames")=List of 3
  ..$ : chr [1:6] "WALKING" "WALKING_UPSTAIRS" "WALKING_DOWNSTAIRS" "SITTING" ...
  ..$ : chr [1:30] "1" "2" "3" "4" ...
  ..$ : chr [1:66] "fBodyAcc_X_mean" "fBodyAcc_X_std" "fBodyAcc_Y_mean" "fBodyAcc_Y_std" ...
```

The final dataSummary data frame was created from the array to result in the following structure:

```
'data.frame':    11880 obs. of  4 variables:
 $ Activity: Factor w/ 6 levels "WALKING","WALKING_UPSTAIRS",..: 1 2 3 4 5 6 1 2 3 4 ...
 $ Subject : num  1 1 1 1 1 1 2 2 2 2 ...
 $ Feature : Factor w/ 66 levels "fBodyAcc_X_mean",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Mean    : num  -0.2028 -0.4043 0.0382 -0.9796 -0.9952 ...
```

Finally, the two datasets are written to files with write.table and row.name as FALSE:

- UCI_HAR_DataExtract.txt
- UCI_HAR_DataSummary.txt