Codebook

Hiro

2021/10/29

Codebook

The codebook explains the data source (original files), variables, function, and names of the measurements(values) in the final output "selectedGroupedSummary"

Original files were obtained from the web-site ("WEBSITE") below.
https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

Original data files
X_test.txt :Test set
y_test.txt :Test labels
subject_test.txt :Links the subject with Test set
X_train.txt :Training set.
y_train.txt :Training labels
subject_train.txt :Links the subject with Train set
activity_labels.txt :Links the class labels with their activity name
features.txt :List of all features

data
Stored original data
data_X_test: X_test.txt
data_y_test: y_test.txt
data_subject_test: subject_test.txt
data_X_train: X_train.txt
data_y_train: y_train.txt
data_subject_train: subject_train.txt
data_activity_labels: activity_labels.txt
data_features: features.txt

Transformed files
test: temporary table to store the "test" table
train: temporary table to store the "train" table
dataAll: combined table consisting of "test" and "train" table
selected: subset table consisting of "Subject", "Activity", and all of the extracted variables containing mean() or std()
selectedGrouped: transformed table of dataAll, grouped by "Subject" and "Activity"
selectedGroupedSummary: final tidy data set containing averages of each measurement of selectedGrouped

Variables
g: extracted characters (values) ending as "mean()" or "std()" in the features data.
activityNames: descriptive names of activities
valueNames: descriptive variable names of measurement values

function
function f() was created to to replace activity values with descriptive activity names
1: WALKING
2: WALKING_UP
3: WALKING_DW
4: SITTING
5: STANDING
6: LAYING

fixed values
Subject
number of 1 to 30

Activity
WALKING
WALKING_UP
WALKING_DW
SITTING
STANDING
LAYING

**names of the variables(average) in the final output*
mean(tBodyAccMagMean)
mean(tBodyAccMagStd)
mean(tGravityAccMagMean)
mean(tGravityAccMagStd)
mean(tBodyAccJerkMagMean)
mean(tBodyAccJerkMagStd)
mean(tBodyGyroMagMean)
mean(tBodyGyroMagStd)
mean(tBodyGyroJerkMagMean)
mean(tBodyGyroJerkMagStd)
mean(fBodyAccMagMean)
mean(fBodyAccMagStd)
mean(fBodyBodyAccJerkMagMean)
mean(fBodyBodyAccJerkMagStd)
mean(fBodyBodyGyroMagMean)
mean(fBodyBodyGyroMagStd)
mean(fBodyBodyGyroJerkMagMean)
mean(fBodyBodyGyroJerkMagStd)

Transformations
Original data files obtained from the WEBSITE were first stored in data, e.g. "data_X_test".

"test" and "train" are temporary table, that merge "Subject" and "Activity", horizontally. "dataAll" is one data set, created by merging train and test, vertically.

“selected” table is a subset table from “dataAll”, selecting “Subject”, “Activity”, and all of the column name containing “mean()” or “std()” at the end. In this process, “grep” function was applied to extract them from the data_features, and stored in variable “g”.

At this timing, “Activity” values were transformed as descriptive names, e.g. “WALKING”.In this process, custom made function “f” was used as well as “lappy” and “unlist” functions.

Next, lengthy variable names containing “-mean()” or“-std()” were appropriately adjusted by using “sub” function and “regular expression”.

To get final output, “selected” table was grouped by “Subject” and “Activity”, by applying “grouped_by” function. Created data was named as “selectedGrouped”

Finally, “selectedGrouped” was summarized calculating average or mean value of each variables. The Number of the variables applied is eighteen(18). The resulted name of the tidy data variables are automatically assigned, e.g.“mean(tBodyAccMagMean)”.

The final output was named as “selectedGroupedSummary”, consisting of 180 rows and 20 columns including two fixed variables “Subject”and “Activity”, and 18 variables, e.g. “mean(tBodyAccMagMean)”.