CodeBook for Data\_Cleaning

Charles Bryan

8/18/2019

Purpose: The purpose of this codebook is to explain the variables, data, and transformations relative to this project:  
 Coursera-Data\_Cleaning\_Project  
  
---  
Variables  
---  
Paths The following are directory paths used to access the appropriate files  
xtrain\_path The path to the downloaded Xtrain data  
ytrain\_path The path to the downloaded ytrain data  
xtest\_path The path to the downloaded Xtest data  
ytest\_path The path to the downloaded ytest data  
features\_path The path to the downloaded features data  
<br>  
DataFrames The following are Dataframes loaded or made in the script  
xtrain The Xtrain data provided by the UCI HAR Dataset.   
 7352 obs of 561 variables. Variables are listed in the features dataframe.  
ytrain The ytrain data provided by the UCI HAR Dataset  
 7352 obs of 1 variables. These values correspond to the activity label of each xtrain observation.  
xtest The Xtrain data provided by the UCI HAR Dataset  
 2947 obs of 561 variables. Variables are listed in the features dataframe.  
ytest The Xtrain data provided by the UCI HAR Dataset  
 2947 obs of 1 variables. These values correspond to the activity label of each xtest observation.  
features A dataframe of the feature names provided by the UCI HAR Dataset.  
 Column 1 contains the feature's indexes starting at 1 and going to 561.  
 Column 2 contains the ordered feature names.  
train This variable is used to hold the merged xtrain and ytrain data.  
 7352 obs of 562 variables.  
test This variable is used to hold the merged xtest and ytest data.  
 2947 obs of 562 variables.  
train\_test This variable is used to hold the merged train and test data.  
 10299 obs of 562 variables.  
train\_test\_mean\_std This dataframe is a reduced version of train\_test.   
 From train\_test we took only the features that were mean() or std() using the   
 features\_mean\_std vector(66 features) and the final column, the activity label.   
 10299 obs. of 67 variables.  
variable\_means This tidy data set contains the average of each variable for each activity and each subject.  
 6 observations for 67 variables. The first variable being the activity labels and the  
 following 66 columns being the features extracted and used in train\_test\_mean\_std  
  
Values The following are Values loaded or made in the script  
  
features\_mean\_std This is a vector of integers. The integers correspond to the indexes of the features   
 dataframe that contain features either mean() or std() variables.  
features\_names This is a vector of characters. This vector contains the 66 feature names corresponding  
 to the features in our train\_test\_mean\_std dataframe.  
features\_names\_fixed This is a vector of characters. The features\_names vector using the names  
 copied from the features.txt file had an error in naming. In 6 cases the   
 text "BodyBody" appeared where I believe it should have been "Body". For this  
 reason, features\_names\_fixed contains the same 66 feature names of features\_names  
 but with the instances of "BodyBody" being replaced with just "Body"  
  
---  
The Data  
---   
Input Data  
  
The following is information on the original dataset used in this project. This data was retrieved in pieces  
and loaded into the following dataframes for cleaning purposes: xtrain, ytrain, xtest, ytest, features  
  
"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 was selected for generating the training data and 30% the test data.   
  
The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain. See 'features\_info.txt' for more details.  
  
Features are normalized and bounded within [-1,1]"  
[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012  
  
Output Data  
  
The output of this project are the following two data sets:  
  
1. train\_test\_mean\_std This is a restructuring and relabeling of the data provided and mentioned under "Input Data " above.  
 From the 561 variables captured in the provided Input Data, this data set contains the 66 variables  
 that captured mean(), std(), or the activity label. The numeric variables were labelled with the given  
 labels provided in the features.txt file. Of the 66 numeric variable names, 6 of which had a typo where   
 "Body" was replaced with "BodyBody". This was fixed in the labeling of this dataset. The final variable  
 of the 67 variables was labeled Activity Label. These values were originally numeric and were simply   
 replaced with the gievn activity as described in activity\_labels.txt. The numeric values of the 66 numeric  
 columns have been left unchanged.  
 10299 observations with 67 variables. The 67th variable is an Activity Label.  
   
2. variable\_means This tidy data set contains the average for each activity for each numeric variable within the dataset  
 train\_test\_mean\_std.   
 6 observations for 67 variables. The first variable being the Activity Labels and the following 66 columns   
 being the features extracted and used in train\_test\_mean\_std   
  
---  
Transformations  
---  
  
The numeric values were kept the same other than the activity labels column.  
The "Activity Label" column was given text for each of the 6 activities instead of a corresponding number.  
The feature names were labeled in the output dataset and mostly kept with the names given in the provided   
 data of features.txt except for 6 of the 66. These 6 had a typo where "Body" was replaced with   
 "BodyBody". This was fixed back to "Body" in the output datasets.  
The means of each variable based on the activity label were calculated when constructing the variable\_means  
 dataset.