CodeBook.Rmd

Getting and Cleaning Data Course Project: Creating Tidy Data

Kristen Borchert The intent of this document is to provide a summary of the data and variables that went into the Tidy Dataframe as well as the transformations that happened to data along the way.

The Data The data used in this study is from the Human Activity Recognition Using Smartphones Dataset Version 1.0. The reference for the data is listed below:

[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

The data used for this assignment represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description of the data collected may be found at: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. The data may be found in the zip file located here: https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip.

My Approach For this assignment I chose to create a tidy data set with all of the variables which had mean and standard deviation in the name. In my experience, it is better to be more inclusive than less; this appraoch could be revisted once exploratory plots are made. The same basic steps laid out in the script would be used.

run_analysis.R: WHAT IT DOES

1) READ THE FILES Read in all seven of the files needed for the analysis. Numerical data for training and test sets reside in the files "X_train.txt" and "X_test.txt". The column headings for the numerical data reside in the file "features.txt". Subject IDs for each of the rows of numerical data reside in the files "subject_train.txt" and "subject_test.txt", respectively, while the activity codes for each row of numerical data reside within "y_train.txt" and "y_test.txt". Finally, the file "activity_labels.txt" contains names for each of the activity codes (a key, if you will).

2) CONSTRUCT THE DATAFRAME A complete dataframe was constructed by taking the following steps: 1) Add column headers to the training and test numerical data using the features.txt file 2) Add Subject and ActivityId columns to the training and test data using cbind 3) Combine training and test data using rbind 4) Final step: add the Activity Labels to the dataframe by using the merge function. The final dataframe is called AllData

3) SUBSET ONLY MEAN AND STD MEASUREMENTS In this step only columns from the original dataframe (AllData) which contained "mean" and "std" in the Feature name were carried forward. Steps included: 1) Construct an index containing the names of all the columns containing mean and std.
2) Make a character vector containing the names of all the columns we want to keep (mean, std plus Subject, DataType and Activity). Note that I dropped ActivityId, as it wasn't needed anymore.

4) MAKE TIDY COLUMN NAMES In this step a series of gsub functions were used to remove "-" and "()" and to captialize "Mean" and "Std" The final variable names can be found in the next section of this document.

5) DATA REDUCTION: CALCULATE MEAN FOR EACH OF THE VARIABLES FOR EACH ACTIVITY AND SUBJECT The dataset was first melted down, leaving “Subject” and “Activity” intact; “DataType” was left off as it wasn’t needed for this particular analysis. Then dcast was used to calculate the mean for each variable then expand them back out so that each variable resided in its own column again.

6) EXPORT THE DATA TO A TXT FILE Finally, the data was exported to a .txt file called “SummarizedActivityDatabySubject.txt”

Variables in final, tidy data file: The final dataset contains 180 observations of 68 variables.

Subject: Subject ID - ranges from 1-30

Activity: Activity that the subject was engaged in. Entries include: LAYING, SITTING, STANDING, WALKING, WALKING_DOWNSTAIRS, WALKING_UPSTAIRS

66 Columns of Measurement Data. Each of the columns respresents a mean taken for each of the original parameters by Subject and Activity. All measurements range from 1 to -1. The column titles are as follows:

tBodyAccMeanX
tBodyAccMeanY
tBodyAccMeanZ
tBodyAccStdX tBodyAccStdY
tBodyAccStdZ
tGravityAccMeanX
tGravityAccMeanY
tGravityAccMeanZ
tGravityAccStdX
tGravityAccStdY
tGravityAccStdZ
tBodyAccJerkMeanX
tBodyAccJerkMeanY
tBodyAccJerkMeanZ
tBodyAccJerkStdX
tBodyAccJerkStdY
tBodyAccJerkStdZ
tBodyGyroMeanX
tBodyGyroMeanY
tBodyGyroMeanZ
tBodyGyroStdX
tBodyGyroStdY
tBodyGyroStdZ
tBodyGyroJerkMeanX
tBodyGyroJerkMeanY
tBodyGyroJerkMeanZ
tBodyGyroJerkStdX
tBodyGyroJerkStdY
tBodyGyroJerkStdZ
tBodyAccMagMean
tBodyAccMagStd
tGravityAccMagMean
tGravityAccMagStd
tBodyAccJerkMagMean
tBodyAccJerkMagStdtBodyGyroMagMean
tBodyGyroMagStd

tBodyGyroJerkMagMean
tBodyGyroJerkMagStd
fBodyAccMeanX
fBodyAccMeanY
fBodyAccMeanZ
fBodyAccStdX
fBodyAccStdY
fBodyAccStdZ
fBodyAccJerkMeanX
fBodyAccJerkMeanY
fBodyAccJerkMeanZ
fBodyAccJerkStdX
fBodyAccJerkStdY
fBodyAccJerkStdZ
fBodyGyroMeanX
fBodyGyroMeanY
fBodyGyroMeanZ
fBodyGyroStdX
fBodyGyroStdY
fBodyGyroStdZ
fBodyAccMagMean
fBodyAccMagStd
fBodyBodyAccJerkMagMean
fBodyBodyAccJerkMagStd
fBodyBodyGyroMagMean
fBodyBodyGyroMagStd
fBodyBodyGyroJerkMagMean
fBodyBodyGyroJerkMagStd