README.Rmd

Getting and Cleaning Data, Course Project, Tidy Data

PURPOSE The purpose of this project is to demonstrate the student's ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis.

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

According to the authors, "This dataset is distributed AS-IS and no responsibility implied or explicit can be addressed to the authors or their institutions for its use or misuse. Any commercial use is prohibited."

From the Author's README file:

"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."

LINKS FOR THE DATA 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.

FILES USED FOR THE ANALYSIS Seven files from the .zip file are 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". * Activity codes for each row of numerical data reside within "y_train.txt" and "y_test.txt". * The file "activity_labels.txt" contains names for each of the activity codes (a key, if you will).

FILES USED TO CREATE THE TIDY DATA

1) The R script called "run_analysis.R", located within my Git Respository (https://github.com/kmborchert/TidyDataProject.git) reads in data files described below which have been downloaded into the /data subfolder of the users working directory. The script does the following:

- Merges the training and the test sets to create one data set.
- Extracts only the measurements on the mean and standard deviation for each measurement.
- Uses descriptive activity names to name the activities in the data set
- Appropriately labels the data set with descriptive activity names.
- Creates a second, independent tidy data set with the average of each variable for each activity and each subject.

2) Also located within my Git Respository is a file called "CodeBook.md" which serves describes the variables, the data, and any transformations or work that you performed to clean up the data.
3) Finally, within the Git Respository there is a file called "SummarizedActivityDatabySubject.txt", which contains the cleaned-up and summarized data for each activity.

GENERAL 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 PREREQUISITE Working directory is set and files have been downleaded and unzipped into a subfolder entitled "data". Details are provided in the script

run_analysis.R: WHAT IT DOES

1) READS THE FILES
2) CONSTRUCTS THE DATAFRAME
3) SUBSETS ONLY MEAN AND STD MEASUREMENTS
4) MAKED TIDY COLUMN NAMES
5) REDUCES THE DATA: CALCULATES MEAN FOR EACH OF THE VARIABLES FOR EACH ACTIVITY AND SUBJECT
6) EXPORTS THE DATA TO A TXT FILE

Details for each of the steps can be found in the associated CodeBook.

Kristen Borchert May 20, 2014