Getting and Cleaning Data - Course Project

A. Project Description

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

It is required to submit:

1.a tidy data set as described below

2.a link to a Github repository with your script for performing the analysis, and

3.a code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md. You should also include a README.md in the repo with your scripts. This file explains how all of the scripts work and how they are connected.

One of the most exciting areas in all of data science right now is wearable computing. Companies like Fitbit, Nike, and Jawbone Up are racing to develop the most advanced algorithms to attract new users. The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description is available at the site where the data was obtained: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

The data for the project is sourced from:
https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

==========

Human Activity Recognition Using Smartphones Dataset

Version 1.0

==========

B. Functional Background

==========

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.

For each record it is provided:

==========

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.

- Triaxial Angular velocity from the gyroscope.

- A 561-feature vector with time and frequency domain variables.

- Its activity label.

- An identifier of the subject who carried out the experiment.

==========

C. Input

==========

The dataset includes the following files:

==========

- 'README.txt'

- 'features_info.txt': Shows information about the variables used on the feature vector.

- 'features.txt': List of all features.

- 'activity_labels.txt': Links the class labels with their activity name.

- 'train/X_train.txt': Training set.

- 'train/y_train.txt': Training labels.

- 'test/X_test.txt': Test set.

- 'test/y_test.txt': Test labels.

The following files are available for the train and test data. Their descriptions are equivalent.

- 'train/subject_train.txt': Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.

- 'train/Inertial Signals/total_acc_x_train.txt': The acceleration signal from the smartphone accelerometer X axis in standard gravity units 'g'. Every row shows a 128 element vector. The same description applies for the 'total_acc_x_train.txt' and 'total_acc_z_train.txt' files for the Y and Z axis.

- 'train/Inertial Signals/body_acc_x_train.txt': The body acceleration signal obtained by subtracting the gravity from the total acceleration.

- 'train/Inertial Signals/body_gyro_x_train.txt': The angular velocity vector measured by the gyroscope for each window sample. The units are radians/second.

Notes:

======

-Getting and Cleaning Data Project

==========

D. The data processing thru run_analysis.R

==========

run_analysis.R

The cleanup script (run_analysis.R) does the following:

1.Merges the training and the test sets to create one data set.

2.Extracts only the measurements on the mean and standard deviation for each measurement.

3.Uses descriptive activity names to name the activities in the data set

4.Appropriately labels the data set with descriptive activity names.

5.Creates a second, independent tidy data set with the average of each variable for each activity and each subject.

Running the script performs 6 tasks:

1. Prepare the environment.
    a. # Set the Working directory
    b. # Load relevant Packages
        i. install.packages("readr")
        ii. install.packages("dplyr")
        iii. install.packages("data.table")
    c. # Load raw data
    d. # Define the tables in order to take advantage of the dplyr functionality
2. **Merge the training and the test sets and generate one data set.**
    a. Create the temporary *tbl*_training_sensor by combining the tbl_x_train with tbl_subject_train and tbl_y_train
    b. Create the temporary tbl_test_sensor by combining the tbl_x_test with tbl_subject_test and tbl_y_test
    c. Provide column names to senor data table

3. Do extract only the mean and standard deviation for each measurement of all observations
    a. Apply Pattern Matching Analysis by selecting the mean and standard deviation for each measurement of all observations by its, Subjectand Activity_Id and generate tbl_sensor_data_mean_std as result set
4. Apply descriptive activity names to name the activities in the data set
    a. Join the tbl_sensor_data_mean_std and tbl_activity_labels, by the common key "Activity_Id" and generate tbl_sensor_data_mean_std
    b. Documentation of the orginal_names of tbl_sensor_data_mean_std
5. Perform appropriate Labelling using descriptive names.

a. "Acc" to be replaced with "Accelerometer"
b. "Gyro" to be replaced with "Gyroscope"
c. "-mean()" to be replaced with "Mean"
d. "-std()" to be replaced with "StandardDeviation"
e. "BodyBody" to be replaced with " Body"
f. "angle" to be replaced with "Angle"
g. "gravity" to be replaced with " Gravity"
h. "tBody" to be replaced with " TimeBody"
i. "^f" to be replaced with " Frequency"
j. "^t" to be replaced with " Time"
k. "-freq()" to be replaced with " Frequency"
l. "FrequencyBody" to be replaced with "Frequency Body"
m. "Mag" to be replaced with "Magnitude"
n. Remove Special Characters
o. Documentation of new_names of tbl_sensor_data_mean_std

6. Generate a new tidy data set with the average of each variable for each activity and each subject.
    a. Remove redundant data not meeting Codd's 3rd normal form (Codd 1990) and generate tbl_sensor_data_mean_std_2
    b. Provide new tempory table tbl_sensor_data_mean_std_avg grouping by "subject" and "activity" the tbl_sensor_data_mean_std_2
    c. Calculate the average for each of the 79 measurements
    d. Apply the correct namining on all colums of tbl_sensor_data_mean_std_avg
    e. Provide tidy result file "sensor_data_avg_grpd.txt" to the Working directory: Write the clean dataset to disk

==========

E. Documentation Process

==========

1. The respective input and output data is listed in CodeBook.md
2. This Readme file documents the purpose, background and functional logic of the project
3. The Source Code of the program is run_analysis.R
4. The result set is provided with the file "sensor_data_avg_grpd.txt"

==========

F. Cleaned Data

==========

The resulting clean dataset is in this repository at: sensor_data_avg_grpd.txt. It contains one row for each subject/activity pair the respective mean of its observations of the respective measurements. standard deviation from the original dataset. Thus it matches the criteria of a tidy data set:

- Each variable is saved in its own column: here the all the single measurements
- Each observation is saved in its own row: here the mean on each subject/activity pair

Notes

A possible segregation into different observational units (multiple files) was not requested as the task explicitly stated to provide one file.

G. Content of this repository

•CodeBook.md: information about raw and tidy data set and elaboration made to transform them

•README.md: this file

•run_analysis.R: R script to transform raw data set in a tidy one