Project: Getting and Cleaning Course Project

(Course #3 - Data Science Specialization by Johns Hopkins University)

# Purpose: demonstrate ability to collect, work with, & clean a data set

Goal: prepare tidy data that can be used for later analysis

Requirements: 1) tidy data set created by run\_analysis.R script such

that: A) each variable forms a column

B) each observation forms a row

C) each type of observational unit forms a table

(NOTE: steps done throughout run.analysis.R script

to tidy data in order to get final tidy data set)

2) link to Github repository with run\_analysis.R script

for performing analysis

3) code book (CodeBook.md)) describing variables,

data, & any transformations or work that was

performed to clean up the data

4) this README.md will only explain how run.analysis.R

works as it is the only one written for this project

(no explanation of scripts working together needed)

Background on data in project:

NOTE: data collected from the accelerometers

from the Samsung Galaxy S smartphone

1) Data Set Information

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. (more details in ‘features.info.txt’ file)

2) Data Attribute Information

Each record in the dataset provides:  
- 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.

3) Data Set Files

A) 'README.txt'

B) 'features\_info.txt': Shows information about variables used

on feature vector

C) 'features.txt': List of all features

D) 'activity\_labels.txt': Links class labels with their activity name

E) 'train/X\_train.txt': Training set

F) 'train/y\_train.txt': Training labels

G) 'test/X\_test.txt': Test set

H) 'test/y\_test.txt': Test labels

NOTE: following files for train & test data – descriptions equivalent

A) 'train/subject\_train.txt': row identifies subject who

performed activity for each sample - range 1 to 30.

B) 'train/Inertial Signals/total\_acc\_x\_train.txt': acceleration signal

from the smartphone accelerometer X axis in standard gravity

units 'g' with every row showing 128 element vector

(same description applies for 'total\_acc\_x\_train.txt' &

'total\_acc\_z\_train.txt' files for Y and Z axis)

C) 'train/Inertial Signals/body\_acc\_x\_train.txt': body acceleration

signal obtained by subtracting gravity from total acceleration.

D) 'train/Inertial Signals/body\_gyro\_x\_train.txt': angular velocity

vector measured by gyroscope for each sample

(units are radians/second)

NOTE: no single piece of documentation exists explaining how

different files relate to each other and how to combine them

into a single data set for subsequent analysis.

File content for each of 8 files shown below:

**File Description**

**------------------------------ ---------------------------------------------------------------------------------**

activity\_labels.txt File containing six rows of data, where each row

contains an activity identifier and an activity

description, a text label to associate with the

y\_train.txt and y\_test.txt files. The numeric

identifier and text labels are separated by a

blank space. The activity label words are

delimited by an underscore when an activity

contains more than a single word.

features.txt File listing the 561 different measurements

taken from the smartphone each time a person

was measured for one of the six activities

monitored during the experiment. Data is listed

in one row per measurement, where the line

number in the file is the assumed key to match

against the test and training data files.

**File Description**

**------------------------------ ---------------------------------------------------------------------------------**

subject\_test.txt File containing one column of data that

identifies the subject (i.e. person)

corresponding to each row of data in the test

measurement X\_test.txt file.

X\_test.txt File containing 561 measurements for each

observed experiment on one of the six activities

for a specific person

y\_test.txt File containing one column of data that

identifies activity corresponding to each

row of data in test measurement X\_test.txt file.

subject\_train.txt File containing one column of data that

identifies the subject (i.e. person) corresponding

to each row of data in test measurement

X\_train.txt file.

X\_train.txt File containing 561 measurements for each

observed experiment on one of the six activities

for a specific person.

y\_train.txt File containing one column of data that identifies

the activity corresponding to each row of data in

the test measurement X\_train.txt file.

File Analysis:

NOTE: column heading named for data sets for clarity

1) x data - combine x\_test & x\_train by rows

2) y data - combine y\_test & y\_train by rows

3) subject data - combine subject\_test & subject\_train by rows

4) merged data – combine x data, y data & subject data by columns

run.analysis.R script steps:

1) read data files for project

2) appropriately label data set with descriptive variable names

3) combine & merge training & test sets to create a data set

4) extracts only measurements of mean & standard deviation

for each measurement

5) uses descriptive activity names to name activities in data set

6) from data set created prior, create 2nd independent tidy

data set with average of each variable for each activity & subject