**Final Project. Getting and Cleanning Data.**

**Sergio Ardila**

1. **Description of Project.**

The purpose of this project is to demonstrate ability to collect, work with, and clean a data set.

**Review criteria**

1. The submitted data set is tidy.
2. The Github repo contains the required scripts.
3. GitHub contains a code book that modifies and updates the available codebooks with the data to indicate all the variables and summaries calculated, along with units, and any other relevant information.
4. The README that explains the analysis files is clear and understandable.
5. The work submitted for this project is the work of the student who submitted it.

**Getting and Cleaning Data Course Project**

The goal is to prepare a tidy data that can be used for later analysis. We are required to submit: 1) a tidy data set as described below, 2) a link to a Github repository with the script for performing the analysis, and 3) a code book that describes the variables, the data, and any transformations or work performed to clean up the data called CodeBook.md. One should also include a README.md in the repo with your scripts. This repo 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 - see for example [this article](http://www.insideactivitytracking.com/data-science-activity-tracking-and-the-battle-for-the-worlds-top-sports-brand/) . 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 following site and wthe data for the project in the second one:

<http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

<https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

One should create one R script called run\_analysis.R that 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 variable names.
5. From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.
6. **Description of experiments and data (done by the researchers)**

===================================================================================================

Human Activity Recognition Using Smartphones Dataset

Version 1.0

===================================================================================================

Jorge L. Reyes-Ortiz(1,2), Davide Anguita(1), Alessandro Ghio(1), Luca Oneto(1) and Xavier Parra(2)

1 - Smartlab - Non-Linear Complex Systems Laboratory

DITEN - Universit� degli Studi di Genova, Genoa (I-16145), Italy.

2 - CETpD - Technical Research Centre for Dependency Care and Autonomous Living

Universitat Polit�cnica de Catalunya (BarcelonaTech). Vilanova i la Geltr� (08800), Spain

activityrecognition '@' smartlab.ws

===================================================================================================

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.

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:

======

- Features are normalized and bounded within [-1,1].

- Each feature vector is a row on the text file.

- The units used for the accelerations (total and body) are 'g's (gravity of earth -> 9.80665 m/seg2).

- The gyroscope units are rad/seg.

- A video of the experiment including an example of the 6 recorded activities with one of the participants can be seen in the following link: http://www.youtube.com/watch?v=XOEN9W05\_4A

For more information about this dataset please contact: activityrecognition '@' smartlab.ws

License:

========

Use of this dataset in publications must be acknowledged by referencing the following publication [1]

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.

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.

Jorge L. Reyes-Ortiz, Alessandro Ghio, Luca Oneto, Davide Anguita and Xavier Parra. November 2013.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

1. Description of work done.

## Step 1: Download the dataset and unzip to a folder in the project directory.

## Important: I used a directory associated with my coursera class !!!! and

## files dowloaded previously to that directory

## The files needed according to instructions are:

# test/subject\_test.txt

# test/X\_test.txt

# test/y\_test.txt

# train/subject\_train.txt

# train/X\_train.txt

# train/y\_train.txt

## Stpe 2: Read data from the files into clearly named variables

## 2.1.Activity files (remember that they are under two subdirectories: text and train)

## 2.2. Subject files

## 2.3. Features files

## Test: Look at data sets properties using str()

## Step 3: Merge the training and the test sets to create one data set (point #1 of Project)

## 3.1.Build one data table each for Subjects, Activity and Features,

## adding rows for Train and Test

## 3.2 Give names to datasets columns using "suject", Activity",

## and the names in file "fetaures.txt" for dataFeatures

## 3.3. Bind data sets by columns to get one data frame for all data

## Step 4: Extracts only the measurements on the mean and standard deviation

## for each measurement. (Point #2 of the project)

## 4.1. Subset dataFeaturesNames that contain "mean()" and "std()" as part of it

## Extract using first grep to identify the indices, and then apply it

## 4.2. Create list of selected column names, plus "subject" and "activity"

## to use select an allData frame

## Step 5: Use descriptive activity names to name the activities in the data set (Point 3 of project)

## 5.1. Read descriptive activity names from "activity\_labels.txt" file

## 5.2. Assign labels to substitute numeric values of Variable "activity" using

## descriptive activity names imported just above

## Step 6: Appropriately labels the data set with descriptive variable names (Point 4 of project). Use function gsub()

## Step 7: From the data set in point 4 of the project, create a second, ordered

## independent tidy data set, with the average of each variable for each subject and activity. Finally, print the data to a file so you can report it.

The script contains plenty of comments so it is very easy to follow. It is possible to print intermediate data.frames so the transformations are being observed.