**Code Book – Getting and Cleaning Data project**

**Project Overview:**

The purpose 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. The project requires having

1) a tidy data set

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 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 repo explains how all of the scripts work and how they are connected.

**Objective:**

An R script called run\_analysis.R need to be created 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.

**Source Data Overview:**

**Source system:**

Human Activity Recognition database built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors.

Raw data are obtained from UCI Machine Learning repository. In particular we used the *Human Activity Recognition Using Smartphones Data Set* , that was used by the original collectors to conduct experiments exploiting Support Vector Machine (SVM) [

Activity Recognition (AR) aims to recognize the actions and goals of one or more agents from a series of observations on the agents' actions and the environmental conditions. The collectors used a sensor based approach employing smartphones as sensing tools. Smartphones are an effective solution for AR, because they come with embedded built-in sensors such as microphones, dual cameras, accelerometers, gyroscopes, etc.

**How Data was generated?**

Generation of the dataset was based on the experiments 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, 3-axial linear acceleration and 3-axial angular velocity were captured at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually.

The obtained data set has been randomly partitioned into two sets, where 70% of the volunteers were selected for generating the training data and 30% the test data.

The Sensor signals -accelerometer and gyroscope 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 that has gravitational and body motion components were 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.

**Source File Location**

The source dataset is a zipped file from the URL: <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

**Processing File Location:**

The file is downloaded from the URL and placed it in the current working directory

**Dataset Information:**

**File Details:**

The Zipped file contain 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.

**Dataset Attribute Information:**

For each record in the dataset 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.

**Data Extraction:**

* Data is extracted into the Current processing directory .
* Once Extracted the files are zipped

**Data Loading and Transformation:**

**Scope:**

The scope of the project is extracting the data which has measurements on the mean and standard deviation. Once the data is fetched, the variables names and the column names need to corrected and replaced with proper labels applied.

**Data loading and Transformation Steps:**

* Load Raw Data Sets

1. The Training data set is loaded into table ‘ train’
2. Testing data set is loaded into table ‘test’
3. The features data is loaded into table – ‘features’

* Extract mean and standard deviation variables

1. Extract only the features that has the description as mean and standard deviation in their variables by searching for that description and load into table ‘ features\_mean\_std’
2. Extract only those subset of data from ’train’ and ‘test’ where the columns match the columns in ‘features\_mean\_Std’

* Merge training and test sets

1. Merge the extracted subset of data from table ‘train’ and table ‘test’ into ‘TrainTest’

* Label variables appropriately

1. Get the Feature Description from ‘features\_mean\_std’ and load into ‘features\_mean\_std\_desc’. This is required to label the columns.
2. Clean up the description by correcting the description of mean, std by removing special characters, so the description looks tidy.
3. Update the column names of the merged data ‘Traintest’ using the cleaned up ‘features\_mean\_std\_desc’

* Use descriptive activity names

1. Load activities data from Training data set into ‘TrainActivities’ and Testing dataset into ‘TestingActivities’
2. Combine the two Activities ‘TrainActivities’ and ‘TestingActivities’ data into ‘TrainTestActivities’
3. Load the activities label data into ‘activityLabels’
4. Rename the first column of ‘TrainTestActivities’ to ‘activity’
5. Update the column values of activity column in ‘TrainTestActivities’ data with the values from ‘activityLabels’ data to have descriptive values for the activity
6. Load the Subjects data from Training into ‘TrainSubjects’ and load the subjects data from Testing into ‘TestSubjects’
7. Combine the two subjects data ‘TrainSubjects ‘ and ‘TestSubjects’ into ‘TrainTestSubjects’
8. Rename the column in ‘TrainTestSubjects’ to ‘subject’
9. Combine ‘TrainTestSubjects’, TrainTestActivities’ with Traintest’ data into ‘AlltransformedData’

* Create a tidy data set

1. Calculate the mean of the measurements in ‘AlltransformedData’ into ‘Tidy\_AverageData’
2. Write the ‘Tidy\_AverageData’ into a ‘Tidy\_Averagedata.txt’

**Tidy Output Dataset:**

The tidy data set 81 variables divided in:

* activity : WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING
* Subject who carried out the experiment : 1, 3, 5, 6, 7, 8, 11, 14, 15, 16, 17, 19, 21, 22, 23, 25, 26, 27, 28, 29, 30
* 79 measurements columns that contain the measurement of each subject and activity.