Getting and Cleaning Data Project CodeBook

Assignment By: Amr Elhennawy

7/14/2020

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.

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.
```

Problem Statement

In this assigment, it is required to:

```
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
```

How it works?

The run_analysis file starts by reading loading the necessary **dplyr** package which will be used in manipulating the data, downloading the file of the data and unzip it in the working directory.

Step 1:

Reads the training set and the test set along with the features file that contains the names of the 561 variables and list of associated activities. Then it binds the training set and testing set using the ubiquitous `rbind(TrainingData, TestingData)`

Step 2:

Extracts the indices of parameters that have either mean or std in its attribute. This is done through the use of the regular expression `"mean|std"` and the `grep()` function. Using these indices, the required values are extracted from the merged set in Step 1.

Step 3:

The list of activities used for the reading of step 1 is loaded and mapped between the indices of the activity and the name of the activity. A vector containing the names of the activities is generated and appended to the merged set of step 1 and the special set of step 2. A separate function is created called `ReplaceActivity(ActivityID)` which takes the activity ID and replaces it with a descriptive string indicating the activity name. `lapply()` is used to convert the integer vector of the activity IDs to vector of characters.

Step 4:

The data sets are labeled with the descriptive variable names generated in step 3.

Step 5:

The special set of step 2 which contains the means and standard deviations of readings is grouped by the person ID and activity name as required. Then the mean of every column for each group is calculated using the function `summarize(across(everything(), mean))` which generates a dataframe of 180 observations representing 6 activities and 30 persons involved in the experiment.