Project on Cleaning Data Course

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.

I. Human Activity Recognition Using Smartphones Data Set

The data linked to this project is the data collected from the accelerometers from the Samsung Galaxy S smartphone.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 were 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.

- Each record consists of the following measurement data:
- 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.

A full description along with the data is available at the following site:
http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

II. Project Data Set

The dataset includes the following files:

- '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.

III. R script "run_analysis.R" written by Ramakrishna Neti

The R script "run_analysis.R" performs the following functions:

1. Download the zip file containing the data set from the website https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip
2. Load data from test and training files into R data frames
3. Use descriptive activity names to name the activities in the data set
4. Label the data sets appropriately with descriptive activity names
5. Merge test and training sets into one data set called MergedData, including the activities
6. Extract only the measurements on the mean and standard deviation for each measurement
7. Create a second, independent tidy data set with the average of each variable for each activity and each subject
8. Upload the tidy data as a txt file created with write.table() using row.name=FALSE

IV. R Packages Used in the project

- "downloader"
- "data.table"
- "reshape2" that includes "plyr"

V. Variables (data frames, tables and files used in the R program)

- **testDataSet** – data frame containing test data (tBodyAcc, fBodyAcc, tBodyAcc-energy, fBodyAcc-energy, tBodyAcc-iqr, fBodyAcc-iqr tBodyAcc-entropy, fBodyAcc-entropy, tBodyAcc-arCoeff, tBodyAcc-correlation()-X,Z, tBodyAcc-correlation()-Y,Z, tGravityAcc-min()-X, tGravityAcc-max()-X, tBodyAccJerk, tBodyGyro, angle etc.)
- **testDataLabels** - data frame containing test data activity labels (STANDING, SITTING, LAYING, WALKING_DOWNSTAIRS, WALKING_UPSTAIRS etc.)
- **testDataSubj** - data frame containing test data subjects (subject who performed the activity for each window sample. Its range is from 1 to 30)
- **trainDataSet** – data frame containing train data (tBodyAcc, fBodyAcc, tBodyAcc-energy, fBodyAcc-energy, tBodyAcc-iqr, fBodyAcc-iqr tBodyAcc-entropy, fBodyAcc-entropy, tBodyAcc-arCoeff, tBodyAcc-correlation()-X,Z, tBodyAcc-correlation()-Y,Z, tGravityAcc-min()-X, tGravityAcc-max()-X, tBodyAccJerk, tBodyGyro, angle etc.)
- **trainDataLabels** - data frame containing training data activity labels (STANDING, SITTING, LAYING, WALKING_DOWNSTAIRS, WALKING_UPSTAIRS etc.)
- **trainDataSubj** - data frame containing test data subjects (subject who performed the activity for each window sample. Its range is from 1 to 30)
- **testDataLabels$V1** - name the activities in the test data set

- **trainDataLabels$V1** - name the activities in the test data set
- **features** – data frame containing descriptive activity name
- **MergedData** – data frame that contains data from testDataSet and trainDataSet merged along with activity labels and subjects
- **DT** – Data table that contains means and standard deviations for each variable in the test and training data sets.
- **Tidy.txt** – this file contains the merged data sets along with the mean and standard deviation values.