Assignment: Getting and Cleaning Data Course Project

This read me file helps the reader walk through the “run_analysis.R” script.

The overall function of the “run_analysis.R” script is to

1. Collect and gather data from various sources (files).
2. Filter the data based on some given criteria.
3. Give meaningful names to the features. As well as, in one instance change values of an attribute to make it instantly understandable instead of requiring supplementary information.
4. Summarize the data based on given criteria.
5. Save the summarized tidy data.

The details of the steps are as follows:

1. The script first checks whether an unzipped directory “UCI HAR Dataset” exists in the working directory from where the R script is being run. If not, it downloads the data and unzip it. These steps result in the data being in downloaded in the current directory. The data used in this project comes from this zipped source: "https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip"

2. The script extracts all the relevant data from “UCI HAR Dataset” directory.

The directory contains the following files and directories:

i. "activity_labels.txt" - file containing the numeric codes and descriptive labels of 6 different activities.
ii. “features_info.txt" - file containing detailed description of where the feature comes from? How the measurements are done on those feature? Overall there are 561 features.
iii. “features.txt" - The complete list of 561 features.
iv. “README.txt" - The full details of Version 1.0 of the “Human Activity Recognition Using Smartphones Dataset”

v. "train" - directory containing data of approximately 70% of the subjects(users) in the study. 7353 observations.
    - The data consist of 3 files:
        - Subject_test.txt: IDs of the subjects corresponding to each observation in the test set.
        - X-test: the measurements of the 561 features corresponding to each observation in the test set.
        - Y-test: activity labels corresponding to each observation in the test set.

vi. "test" - directory containing data of approximately 30% of the subjects(users) in the study. 2947 observations.
    - The data consist of 3 files:
        - Subject_test.txt: IDs of the subjects corresponding to

each observation in the test set.
- X-test: the measurements of the 561 features corresponding to each observation in the test set.
- Y-test: activity labels corresponding to each observation in the test set.

vii. Note: the train and test directory contains one subdirectory each “Inertial Signals” that is not required for this project.

3. Following is the data about the subjects, features(measurements), and activities that is read in:

a. Reading all three files for train sets and clipping the train and test data together, such that, all the train data is consolidated in one data frame (7353 observations, each associated with a subject, activity and a record of 561 features)
b. Reading all three files for test sets and clipping the train and test data together, such that, all the train data is consolidated in one data frame (2947 observations, each associated with a subject, activity and a record of 561 features)
c. Reading the “activity_labels.txt” to get the description of activities that are recorded as 1 to 6 in Y-train/test.txt files.
d. Reading the “features.txt” file to get the feature labels.

4. Following are the 5 steps of the project:

Step1: Merges the training and the test sets to create one data set. (10299 observations for 30 users performing 6 different activities).

Step 2: Extracts only the measurements on the mean and standard deviation for each measurement. Since this step asked for the mean and standard deviation on each measurement, there are 8 * 3 (X,Y,Z-axis) + 9(magnitudes) = 33 signal measurements that are extracted from the instruments. A number of variables (including mean and standard deviation are estimated for each of these signals). Hence, a total of 66 features are extracted from the compete data. Note: there are five variables extracted from the original set by averaging the signals. In my view, these are not the raw measurements hence do not constitute the measurements of which mean and standard deviation should be extracted. Besides, these angle features do not have the standard deviation estimates. Thus, I extracted features mean and standard deviation where both were available for each measurement.

Step 3: Uses descriptive activity names to name the activities in the data set. Using the descriptive names in “activity_labels.txt” that was earlier read into R, update the activity identification information.

Step 4: Appropriately labels the data set with descriptive variable names. Each variable is renamed such that what information its values contain can easily be gauged by the name.

Step 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. The tidy data contains one piece of observation in each row. That is, one row for each activity performed by each user and 66 independent features encoding the mean and standard deviation of 33 different signals. The resulting data is lean with 180 observations on 66 distinct features.

5. Save the tidy data in a file named “UCI_HAR_Averages.txt”

6. The data can be read into a dataframe using this command:
    read.table("UCI_HAR_Averages.txt")