"Getting and Cleaning Data" Course Project

Initial data for research

The script is invented to analyze the data from UCI HAR Dataset. It's supposed that archive is extracted to the working directory.

The following files from the initial dataset is used:

1. ***features.txt*** - includes the descriptions for features measured
2. ***train/X_train.txt*** - includes the measurements of the features in train set (one row - 1 measurement of 561 features)
3. ***test/X_test.txt*** - includes the measurements of the features in test set
4. ***train/subject_train.txt*** - subject for each measurement from the train set
5. ***test/subject_test.txt*** - subject for each measurement from the test set
6. ***train/y_train.txt*** - activity (from 1 to 6) for each measurement from the train set
7. ***test/y_test.txt*** - activity (from 1 to 6) for each measurement from the test set

How script works

Script involves the following stages:

1. Downloads to R ids and descriptions for features being measured in experiment from file ***features.txt***.

2. Independently loads complete data for train and test sets. Let's revoke these loading process considering train set:
   a. Firstly loads the measurements from ***X_train.txt*** as a data frame
   b. For these data frame column names are updated to be more user friendly using features description loaded on the previous stage. (**STEP 4**: *Appropriately label the data set with descriptive variable names* of Course Project
   c. activity labels and subjects for measurements are also loaded from files ***train/y_train.txt*** and ***train/subject_train.txt*** and added to data frame as a separated columns.

Similar steps are made for test dataset and finally 2 rows of 2 data frames are merged together to form are data frame with complete data (**STEP 1**: *Merge the training and the test sets to create one data set* of assignment)

3. To extract measurements that involves only mean and standard deviation values script uses grep, that finds column names that includes "mean()" or "std()" (also

columns activity and subject are added to filtered data frame, since they are important dimensions). After that all new data frame with only necessary columns is created. (**STEP 2**: *Extract only the measurements on the mean and standard deviation for each measurement* of assignment)

4. To provide descriptive values for activity labels a new variable "*activitylabel*" is added to dataset, that is a factor variable with levels mentioned in file activity_labels.txt (**STEP 3**: *Use descriptive activity names to name the activities in the data set* of assignment)

5. Creates a melted data frame using activity label and subject as ids, after that mean values for all variables are calculated grouped by activity and subject using dcast() function and tidy data frame is created. (**STEP 5**: *Create a second, independent tidy data set with the average of each variable for each activity and each subject*)