Purpose

This project consists of collecting, working and cleaning "*Human Activity Recoginition Using Smartphones"* data set.

Data source

A full description of the source data set is available at the site where the data was obtained:
http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

Here are the data for the project:
https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

Files used are:

- '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.
- 'train/subject_train.txt': Each row identifies the training set subjects who performed the activity for each window sample. Its range is from 1 to 30.
- 'test/X_test.txt': Test set.
- 'test/y_test.txt': Test labels.
- 'train/subject_test.txt': Each row identifies the test set subjects who performed the activity for each window sample. Its range is from 1 to 30.

R libraries used

tidyr
plyr
dplyr

Study design

Steps

1. Data download if not present in working directory.
2. Build training data. Training set is joined to training activity and subject identifiers.
3. Build test data. Test set is joined to test activity and subject identifiers.
4. Join training and test data into the same data frame.
5. Name all features measured/calculated using provided feature names.
6. Extract only features on the mean and the standard deviation.
7. Clean feature names.
8. Set descriptive activity names using activity labels provided.
9. Aggregate data, calculating average of each feature for each activity and each subject.
10. Tidy data:
    - 10.1. separate features into identified variables
    - 10.2. order columns and rows
    - 10.3. rename values and variable names where required
    - 10.4. create factors

Relevant data produced

- all_data - data frame with all raw data (step 4).
- mean_std_data - data frame with selected raw data, clean feature names and activity labels (step 8).
- messy_data - data frame with aggregated untidy data (step 9).
- tidy_data - tidy data set produced containing variables described in this code book (step 10).

Variables in tidy data

subject

Identifier for each subject who carried out the experiment
Integer number from 1 to 30

activity

Activity performed by subjects during the experiment
Factor, levels:

WALKING
WALKING_UPSTAIRS
WALKING_DOWNSTAIRS

SITTING
STANDING
LAYING

domain

Domain of the signal measured
Factor, levels:
frequency
time

signal

Signals measured
Factor, levels:
BodyAcc
BodyAccJerk
BodyAccJerkMag
BodyAccMag
BodyGyro
BodyGyroJerk
BodyGyroJerkMag
BodyGyroMag
GravityAcc
GravityAccMag

axis

Spatial axis of the signal measured
Factor, levels:
X
Y
Z
NA - axis not defined / not applicable

mean

Average of all measured means for each subject, activity, domain, signal and axis
Float number from -1.0 to 1.0

standardDeviation

Average all measured standard deviations for each subject, activity, domain, signal and axis
Float number from -1.0 to 1.0