Code Book

Project for Coursera: Getting and Cleaning Data, part of the Data Science specialization

Data Set

The purpose of this project is to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis. The data used for this project is collected from the accelerometers from the Samsung Galaxy S smartphones. A full description is available at the site where the data was obtained: Human Activity Recognition Data

The data for the project is accesible here: Project Data Set

Data Processing / Modifications

The data was processed according to the steps below:

1. Merging of the training and the test sets to create one data set

1.1. The data set was downloaded locally, extracted, read and assigned to variables as follows:

a) Features & Activity Labels

- features <- features.txt - *list of all the features (time and frequency domain variables) compiled from the accelerometers and gyroscopes. The units used for the accelerations (total and body) are 'g's (gravity of earth -> 9.80665 m/seg2). The gyroscope units are rad/seg.*
- activityLabels <- activity_labels.txt - *list of activity names and corresponding activity IDs*

b) Training Data

- trainSubjects <- subject_train.txt - *training data on study participants (21 out of 30 observations)*
- trainX <- X_train.txt - *training data features*
- trainY <- y_train.txt - *training data activities*

c) Test Data

- testSubjects <- subject_test.txt - *test data set on study participants (9 out of 30 observations)*
- testX <- X_test.txt - *test data features*
- testY <- y_test.txt - *test data activites*

1.2. Merging of the two data sets

- dataX - obtained by merging the training and test data features, *trainX* and *trainY*
- dataY - obtained by merging the training and test data activites, *trainY* and *testY*
- dataSubjects - obtained by merging the training and test data on study participants, *trainSubjects* and *testSubjects*
- mergedSet - obtained by merging *dataX*, *dataY* and *dataSubjects*

2. Extraction only of the measurements on the mean and standard deviation for each measurement

- selectedSet - obtained by subsetting the *mergedSet* and selecting the "subjects" and "id" columns (names previously set in the reading phase), together with the mean and standard deviation for each measurement

3. Use of descriptive activity names to name the activities in the data set

- The activity IDs were replaced with the activity names (*activityLabels*).

4. Appropriate labelling of the data set with descriptive variable names

- Renaming of the *id* column in *selectedSet* into *activity*
- Replacement of *^f* with *Frequency* in column names of *selectedSet*
- Replacement of *^t* with *Time* in column names of *selectedSet*
- Replacement of *Acc* with *Accelerometer* in column names of *selectedSet*
- Replacement of *BodyBody* with *Body* in column names of *selectedSet*
- Replacement of *Gyro* with *Gyroscope* in column names of *selectedSet*
- Replacement of *Mag* with *Magnitude* in column names of *selectedSet*
- Replacement of *tBody* with *TimeBody* in column names of *selectedSet*
- Replacement of *-mean()* with *Mean* in column names of *selectedSet*
- Replacement of *-std()* with *Standard Deviation* in column names of *selectedSet*
- Replacement of *-freq()* with *Frequency* in column names of *selectedSet*

5. Creation of a second, independent tidy data set with the average of each variable for each activity and each subject

- tidyData2 - obtained by grouping the *selectedSet* by "subjects"" and "activity" and taking the average (mean) of each variable for each activity and each subject
- TidyData.txt - saved output