Codebook for Tidying up of UCI HAR dataset containing records of experiments conducted under “Human Activity Recognition Using Smartphones - Version 1.0 by Smartlab, Italy.

BACKGROUND

“Getting and Cleaning Data” Course Project (Coursera), requires demonstration of ability to collect, work with, and clean a dataset. Following dataset was prescribed for the project.

Data Source

The source data pertains to observations that were recorded during an experiment on wearable computing algorithms using Samsung smartphones.

- For this project data is sourced from the following URL: https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip
- To know about the origin and related work, one may check at:

  http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

Source files

UCI HAR Dataset containing the following files

activity_labels.txt
features.txt
features_info.txt
Filelist.txt
README.txt
test/subject_test.txt
test/X_test.txt
test/y_test.txt
test/Inertial Signals/body_acc_x_test.txt
test/Inertial Signals/body_acc_y_test.txt
test/Inertial Signals/body_acc_z_test.txt
test/Inertial Signals/body_gyro_x_test.txt
test/Inertial Signals/body_gyro_y_test.txt
test/Inertial Signals/body_gyro_z_test.txt
test/Inertial Signals/total_acc_x_test.txt
test/Inertial Signals/total_acc_y_test.txt
test/Inertial Signals/total_acc_z_test.txt
train/subject_train.txt
train/X_train.txt
train/y_train.txt
train/Inertial Signals/body_acc_x_train.txt
train/Inertial Signals/body_acc_y_train.txt
train/Inertial Signals/body_acc_z_train.txt
train/Inertial Signals/body_gyro_x_train.txt
train/Inertial Signals/body_gyro_y_train.txt
train/Inertial Signals/body_gyro_z_train.txt
train/Inertial Signals/total_acc_x_train.txt
train/Inertial Signals/total_acc_y_train.txt
train/Inertial Signals/total_acc_z_train.txt

The challenge – convert messy data to tidy data

The source data does not confirm to Tidy-data structure for following reasons.

- Every observation is broken into 3 files instead of placing them in one file.
- Descriptive headings are absent for all the values in their respective files.

Further the Items of interest are measurements on mean and standard deviation, which form a subset of 561 variables in the source data.

Task is to extract a subset of these variables from source files, suitably transform and store them in a tidy-data structure that facilitates subsequent use.

VARIABLES

Features:

"Features.txt" file lists 561 variables. Their values are held in source files "train_x.txt" and "test_x.txt". More details are available in "Features_info.txt" of the source files. Focus of this document is restricted to the Mean and Standard Deviation values that are relevant.

- All the variables were derived from the 3-axial raw time domain (denoted by prefix "t" of variable names) signals that were captured from accelerometer (tAcc-XYZ) and gyroscope (tGyro-XYZ) as part of the experiments.
- Noise was removed from these raw signals and separated into body (tBodyAcc) and gravity acceleration (tGravityAcc) signals along X,Y and Z dimensions.
- Jerk signals (tBodyAccJerk and tBodyGyroJerk) were derived from body linear acceleration and angular velocity computations.
- Magnitude (Mag) was calculated for 3 dimensional signals using Euclidean norm for corresponding measures.
- Fast Fourier Transform (TFT) was applied to some of the above measures and resultant variables are indicated with a prefix of "f".
- The above signals were used to estimate variables of the feature vector for each pattern: '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.
    - tBodyAcc-XYZ
    - tGravityAcc-XYZ
    - tBodyAccJerk-XYZ
    - tBodyGyro-XYZ
    - tBodyGyroJerk-XYZ
    - tBodyAccMag
    - tGravityAccMag
    - tBodyAccJerkMag
    - tBodyGyroMag
    - tBodyGyroJerkMag
    - fBodyAcc-XYZ
    - fBodyAccJerk-XYZ
    - fBodyGyro-XYZ
    - fBodyAccMag
    - fBodyAccJerkMag
    - fBodyGyroMag
    - fBodyGyroJerkMag
- Following variables were also estimated from these signals as follows:
    - mean(): Mean value
    - std(): Standard deviation
    - mad(): Median absolute deviation
    - max(): Largest value in array
    - min(): Smallest value in array
    - sma(): Signal magnitude area
    - energy(): Energy measure. Sum of the squares divided by the number of values.
    - iqr(): Interquartile range
    - entropy(): Signal entropy
    - arCoeff(): Autorregresion coefficients with Burg order equal to 4
    - correlation(): correlation coefficient between two signals
    - maxInds(): index of the frequency component with largest magnitude
    - meanFreq(): Weighted average of the frequency components to obtain a mean frequency
    - skewness(): skewness of the frequency domain signal

  - kurtosis(): kurtosis of the frequency domain signal
  - bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.
  - angle(): Angle between to vectors.
- Additional vectors were obtained by averaging the signals in a signal window sample. These are used on the angle() variable:
  - gravityMean
  - tBodyAccMean
  - tBodyAccJerkMean
  - tBodyGyroMean
  - tBodyGyroJerkMean

Subject identifiers of participants

Each row in “train/subject_train.txt” and “test/subject_test.txt” identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.

- The file “train/subject_train.txt” contains 7352 rows pertaining to training group.
- The file “test/subject_test.txt” contains 2947 rows pertaining to test group.

There were 30 participants in the experiment. Of which 21 were part of training group and 9 were part of test group.

- Id’s of training group: 1 , 3, 5, 6 , 7, 8 , 11 , 14, 15, 16, 17, 19, 21, 22, 23, 25, 26, 27, 28, 29, 30
- Id’s of test group: 2 ,4, 9, 10, 12, 13, 18, 20, 24

Activities

Each participant performed 6 activities each (walking, walking_upstairs, walking_downstairs, sitting, standing, laying). These activities are represented by values and labels in “activity_lables.txt” as follows:

1. LAYING
2. SITTING
3. STANDING
4. WALKING
5. WALKING_DOWNSTAIRS
6. WALKING_UPSTAIRS

Each row in files "train/y_train.txt" and “test/y_test.txt” identifies the activity performed by the subject for each window sample

- The file “train/y_train.txt” contains 7352 rows pertaining to training group.
- The file “test/y_test.txt” contains 2947 rows pertaining to test group.

DATA

Each row of “train/x_train.txt” and “test/x_test.txt” holds 561-feature vector with time and frequency domain variables for each window sample. These hold observations of training and test groups respectively. For this project we use only the features that represent mean and standard deviation measures.

TRANSFORMATIONS

1. **Source file “dataset.zip” is downloaded to current working directory and unzipped**

Obtain UCI HAR Dataset containing the following files. We do not deal with data held under Inertial Signals for the purposes of the project.

2. **Merge the training and the test sets to create one data set.**

Source holds data in two sets, one pertaining to training group and the other to test group. Each observation is held in 3 files for each group as indicated below.

Group	**SUBJECT**	**ACTIVITY**	**561-FEATURE VECTOR**
Training	train/subject_train.txt	train/y_train.txt	train/x_train.txt
Test	test/subject_test.txt	test/y_test.txt	test/x_test.txt

- The 3 files of each group are read into R as data-frames and then their columns are glued together using cbind() to obtain a complete record set for that group.
- Append the record set of Test group to the record set of Training group to get complete data set.

3. Extract the measurements on the mean and standard deviation for each measurement.

- Create a new data-frame using a sub-set of 68 relevant columns from the complete data set.
- This subset is comprised of first two columns (subject and activity) and 66 other variables that have either "mean()' or "std()" in their feature names.

Arithmetic Mean variables *	Standard deviation variables *	Description
a. tBodyAcc-mean()-X b. tBodyAcc-mean()-Y c. tBodyAcc-mean()-Z d. tGravityAcc-mean()-X e. tGravityAcc-mean()-Y f. tGravityAcc-mean()-Z g. tBodyGyro-mean()-X h. tBodyGyro-mean()-Y i. tBodyGyro-mean()-Z	a. tBodyAcc-std()-X b. tBodyAcc-std()-Y c. tBodyAcc-std()-Z d. tGravityAcc-std()-X e. tGravityAcc-std()-Y f. tGravityAcc-std()-Z g. tBodyGyro-std()-X h. tBodyGyro-std()-Y i. tBodyGyro-std()-Z	tBodyAcc : body acceleration signals tGravityAcc : gravity acceleration signals tBodyGyro : body angular velocity of the signals
a. tBodyAccJerk-mean()-X b. tBodyAccJerk-mean()-Y c. tBodyAccJerk-mean()-Z d. tBodyGyroJerk-mean()-X e. tBodyGyroJerk-mean()-Y f. tBodyGyroJerk-mean()-Z	g. tBodyAccJerk-std()-X h. tBodyAccJerk-std()-Y i. tBodyAccJerk-std()-Z j. tBodyGyroJerk-std()-X k. tBodyGyroJerk-std()-Y l. tBodyGyroJerk-std()-Z	tBodyAccJerk : Jerk measure derived from body linear acceleration of the signal tBodyGyroJerk : Jerk measure derived from angular velocity of the signal
a. tBodyAccMag-mean() b. tGravityAccMag-mean() c. tBodyAccJerkMag-mean() d. tBodyGyroMag-mean() e. tBodyGyroJerkMag-mean()	a. tBodyAccMag-std() b. tGravityAccMag-std() c. tBodyAccJerkMag-std() d. tBodyGyroMag-std() e. tBodyGyroJerkMag-std()	Magnitude (Mag) was calculated for 3 dimensional signals using Euclidean norm for corresponding measures
a. fBodyAcc-mean()-X b. fBodyAcc-mean()-Y c. fBodyAcc-mean()-Z d. fBodyAccJerk-mean()-X e. fBodyAccJerk-mean()-Y f. fBodyAccJerk-mean()-Z g. fBodyGyro-mean()-X h. fBodyGyro-mean()-Y i. fBodyGyro-mean()-Z j. fBodyAccMag-mean() k. fBodyAccMag-std() l. fBodyBodyAccJerkMag-mean() m. fBodyBodyGyroMag-mean() n. fBodyBodyGyroJerkMag-mean()	a. fBodyAcc-std()-X b. fBodyAcc-std()-Y c. fBodyAcc-std()-Z d. fBodyAccJerk-std()-X e. fBodyAccJerk-std()-Y f. fBodyAccJerk-std()-Z g. fBodyGyro-std()-X h. fBodyGyro-std()-Y i. fBodyGyro-std()-Z j. fBodyBodyAccJerkMag-std() k. fBodyBodyGyroMag-std() l. fBodyBodyGyroJerkMag-std()	Fast Fourier Transform (TFT) was applied to some of the above measures and resultant variables are indicated with a prefix of "f".

*Note : Suffix X,Y,Z denote measure along the corresponding dimension of 3 axial signals

4. **Uses descriptive activity names to name the activities in the data set**
    - Source file has activity codes (1 to 6) to represent various activities performed by the subject during the each window sample.
    - Factor the values in column that has activity values in the data set that was created in step 3 above
    - Use the following levels and labels while factoring activity

        1 - LAYING
        2 - SITTING
        3 - STANDING
        4 - WALKING
        5 - WALKING_DOWNSTAIRS
        6 - WALKING_UPSTAIRS

5. **Appropriately label the data set with descriptive variable names.**
    - Source data does not have descriptive variable names.
    - First column is labelled as "Subject" and second column is "Activity"
    - Hence we select variable names for from "Features.txt" and remove special characters
        - Filter feature names to those that were selected during step 3 above (features having phrase "mean()" or "std()" in their names)
        - Remove all special characters like ",-()" i.e., comma, hyphen and parenthesis
        - Name the columns with right feature-name
6. **From the data set in the previous step, create a second, independent tidy data set with the average of each variable for each activity and each subject.**
    - Average of each variable for each activity and each subject is computed using gather(), group_by(), summarize() and spread() functions. The code is…

```
averageValues<- selectData %>%
        gather(Features,value,-Subject,-Activity) %>%   # gather() works similiar to melt()
        group_by(Subject,Activity,Features) %>%         # group_by(), summarize()
        summarize(MeanValue = mean(value)) %>%          #   & spread() work similar to dcast()
        spread(Features,MeanValue)
```

    - All variable names are suffixed with "Avg".
    - The resultant tidy data set is stored as a txt file created with write.table() using row.name=FALSE.

STRUCTURE OF RESULTANT TIDY DATASET

The R program "run_analysis.R" provided with this codebook generates "tidy_mean_std_averages.txt", a txt file using write.table() with row.name = FALSE. While attempting to read it, please note that it has a header. The resultant data set has following structure.

```
'data.frame':  180 obs. of  68 variables:
 Subject                  : int  1 1 1 1 1 1 2 2 2 2 ...
 Activity                 : Factor w/ 6 levels "LAYING","SITTING",..: 4 6 5 2 3 1 4 6 5 2 ...
 AvgtBodyAccmeanX         : num  0.277 0.255 0.289 0.261 0.279 ...
 AvgtBodyAccmeanY         : num  -0.01738 -0.02395 -0.00992 -0.00131 -0.01614 ...
 AvgtBodyAccmeanZ         : num  -0.1111 -0.0973 -0.1076 -0.1045 -0.1106 ...
 AvgtBodyAccstdX          : num  -0.284 -0.355 0.03 -0.977 -0.996 ...
 AvgtBodyAccstdY          : num  0.11446 -0.00232 -0.03194 -0.92262 -0.97319 ...
 AvgtBodyAccstdZ          : num  -0.26 -0.0195 -0.2304 -0.9396 -0.9798 ...
 AvgtGravityAccmeanX      : num  0.935 0.893 0.932 0.832 0.943 ...
 AvgtGravityAccmeanY      : num  -0.282 -0.362 -0.267 0.204 -0.273 ...
 AvgtGravityAccmeanZ      : num  -0.0681 -0.0754 -0.0621 0.332 0.0135 ...
 AvgtGravityAccstdX       : num  -0.977 -0.956 -0.951 -0.968 -0.994 ...
 AvgtGravityAccstdY       : num  -0.971 -0.953 -0.937 -0.936 -0.981 ...
 AvgtGravityAccstdZ       : num  -0.948 -0.912 -0.896 -0.949 -0.976 ...
 AvgtBodyAccJerkmeanX     : num  0.074 0.1014 0.0542 0.0775 0.0754 ...
 AvgtBodyAccJerkmeanY     : num  0.028272 0.019486 0.02965 -0.000619 0.007976 ...
 AvgtBodyAccJerkmeanZ     : num  -0.00417 -0.04556 -0.01097 -0.00337 -0.00369 ...
```

```
AvgtBodyAccJerkstdX        : num  -0.1136 -0.4468 -0.0123 -0.9864 -0.9946 ...
AvgtBodyAccJerkstdY        : num  0.067 -0.378 -0.102 -0.981 -0.986 ...
AvgtBodyAccJerkstdZ        : num  -0.503 -0.707 -0.346 -0.988 -0.992 ...
AvgtBodyGyromeanX          : num  -0.0418 0.0505 -0.0351 -0.0454 -0.024 ...
AvgtBodyGyromeanY          : num  -0.0695 -0.1662 -0.0909 -0.0919 -0.0594 ...
AvgtBodyGyromeanZ          : num  0.0849 0.0584 0.0901 0.0629 0.0748 ...
AvgtBodyGyrostdX           : num  -0.474 -0.545 -0.458 -0.977 -0.987 ...
AvgtBodyGyrostdY           : num  -0.05461 0.00411 -0.12635 -0.96647 -0.98773 ...
AvgtBodyGyrostdZ           : num  -0.344 -0.507 -0.125 -0.941 -0.981 ...
AvgtBodyGyroJerkmeanX      : num  -0.09 -0.1222 -0.074 -0.0937 -0.0996 ...
AvgtBodyGyroJerkmeanY      : num  -0.0398 -0.0421 -0.044 -0.0402 -0.0441 ...
AvgtBodyGyroJerkmeanZ      : num  -0.0461 -0.0407 -0.027 -0.0467 -0.049 ...
AvgtBodyGyroJerkstdX       : num  -0.207 -0.615 -0.487 -0.992 -0.993 ...
AvgtBodyGyroJerkstdY       : num  -0.304 -0.602 -0.239 -0.99 -0.995 ...
AvgtBodyGyroJerkstdZ       : num  -0.404 -0.606 -0.269 -0.988 -0.992 ...
AvgtBodyAccMagmean         : num  -0.137 -0.1299 0.0272 -0.9485 -0.9843 ...
AvgtBodyAccMagstd          : num  -0.2197 -0.325 0.0199 -0.9271 -0.9819 ...
AvgtGravityAccMagmean      : num  -0.137 -0.1299 0.0272 -0.9485 -0.9843 ...
AvgtGravityAccMagstd       : num  -0.2197 -0.325 0.0199 -0.9271 -0.9819 ...
AvgtBodyAccJerkMagmean     : num  -0.1414 -0.4665 -0.0894 -0.9874 -0.9924 ...
AvgtBodyAccJerkMagstd      : num  -0.0745 -0.479 -0.0258 -0.9841 -0.9931 ...
AvgtBodyGyroMagmean        : num  -0.161 -0.1267 -0.0757 -0.9309 -0.9765 ...
AvgtBodyGyroMagstd         : num  -0.187 -0.149 -0.226 -0.935 -0.979 ...
AvgtBodyGyroJerkMagmean    : num  -0.299 -0.595 -0.295 -0.992 -0.995 ...
AvgtBodyGyroJerkMagstd     : num  -0.325 -0.649 -0.307 -0.988 -0.995 ...
AvgfBodyAccmeanX           : num  -0.2028 -0.4043 0.0382 -0.9796 -0.9952 ...
AvgfBodyAccmeanY           : num  0.08971 -0.19098 0.00155 -0.94408 -0.97707 ...
AvgfBodyAccmeanZ           : num  -0.332 -0.433 -0.226 -0.959 -0.985 ...
AvgfBodyAccstdX            : num  -0.3191 -0.3374 0.0243 -0.9764 -0.996 ...
AvgfBodyAccstdY            : num  0.056 0.0218 -0.113 -0.9173 -0.9723 ...
AvgfBodyAccstdZ            : num  -0.28 0.086 -0.298 -0.934 -0.978 ...
AvgfBodyAccJerkmeanX       : num  -0.1705 -0.4799 -0.0277 -0.9866 -0.9946 ...
AvgfBodyAccJerkmeanY       : num  -0.0352 -0.4134 -0.1287 -0.9816 -0.9854 ...
AvgfBodyAccJerkmeanZ       : num  -0.469 -0.685 -0.288 -0.986 -0.991 ...
AvgfBodyAccJerkstdX        : num  -0.1336 -0.4619 -0.0863 -0.9875 -0.9951 ...
AvgfBodyAccJerkstdY        : num  0.107 -0.382 -0.135 -0.983 -0.987 ...
AvgfBodyAccJerkstdZ        : num  -0.535 -0.726 -0.402 -0.988 -0.992 ...
AvgfBodyGyromeanX          : num  -0.339 -0.493 -0.352 -0.976 -0.986 ...
AvgfBodyGyromeanY          : num  -0.1031 -0.3195 -0.0557 -0.9758 -0.989 ...
AvgfBodyGyromeanZ          : num  -0.2559 -0.4536 -0.0319 -0.9513 -0.9808 ...
AvgfBodyGyrostdX           : num  -0.517 -0.566 -0.495 -0.978 -0.987 ...
AvgfBodyGyrostdY           : num  -0.0335 0.1515 -0.1814 -0.9623 -0.9871 ...
AvgfBodyGyrostdZ           : num  -0.437 -0.572 -0.238 -0.944 -0.982 ...
AvgfBodyAccMagmean         : num  -0.1286 -0.3524 0.0966 -0.9478 -0.9854 ...
AvgfBodyAccMagstd          : num  -0.398 -0.416 -0.187 -0.928 -0.982 ...
AvgfBodyBodyAccJerkMagmean : num  -0.0571 -0.4427 0.0262 -0.9853 -0.9925 ...
AvgfBodyBodyAccJerkMagstd  : num  -0.103 -0.533 -0.104 -0.982 -0.993 ...
AvgfBodyBodyGyroMagmean    : num  -0.199 -0.326 -0.186 -0.958 -0.985 ...
AvgfBodyBodyGyroMagstd     : num  -0.321 -0.183 -0.398 -0.932 -0.978 ...
AvgfBodyBodyGyroJerkMagmean: num  -0.319 -0.635 -0.282 -0.99 -0.995 ...
AvgfBodyBodyGyroJerkMagstd : num  -0.382 -0.694 -0.392 -0.987 -0.995 ...
```

CONCLUSION:

The messy data is turned into a tidy data and saved for future use. The final outcome meets Hadley Wickham's definition of tidy data:

- each variable is a column,
- each observation is a row, and
- each type of observational unit is a table.