Code Book - Project Assignment Getting & Cleaning Data Course

Christof

2023-03-02

Raw Data

Human Activity Recognition Using Smartphones Dataset, Version 1.0

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

Data set: https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip

The script "run_analysis.R" processes the raw data

There 4 data sets created from the source.

- activities_data
  - Factor data (key-value pairs) for activities, for example activity 5 is the identifier for the activity STANDING. There are 6 different activities
  - Created using "activity_labels.txt" file
- features_list
  - measurment names
  - Created using "features.txt" file
- tidy_data_full
  - Contains per activity - subject - measurement combination the value for each measuremt that contains "mean()" or "std() in the name.
  - Created using test and training data from the study
- tidy_data_mean_bymeasurement
  - Contains per activity - subject - measurement combination the mean value of all occurrences for the combination.
  - Created using the tidy_data_full data set (see previous bullet)

For each data set is described how these are created the next section. The data set variables are described in the last section

Processing of the raw data

Creating data set "activities_data"

- File: activity_labels.txt
- Transformation to the data frame:
  - Adding descriptive names to the columns

Creating data set "features_list"

- File: features.txt
  - Tranformation to data frame
    * Adding descriptive names to the columns
    * Create an additional column to create unique names. Reason: there are duplicate names

Creating data set "tidy_data_full"

Steps

1. For both test and training data:
   - retrieve activity identifiers into own data frame - including creation of descriptive name
   - retrieve subject identifiers into own data frame - including creation of descriptive name
   - retrieve measurement data
   - Combine the 3 data frames into 1 - resulting in 2 data frames (1 for test and 1 for training data)
2. Merge both test and training data frames - by row
3. Joined the merged data frame with the activities_data frame to add the activity names
4. Calculate the mean per row of all measurements and add this to the merged data frame excluding any NA values if present
5. Calculate the standard deviation per row of all measurements and add this to the merged data frame excluding any NA values if present
6. Melt the data frame to create a tidy data frame with for each row:
   1. Activity - Subject combination
   2. Variable: calculated mean (step #4) or calculated stand deviation (step #5)
   3. Value of the variable

Files were used

- Measurement: X_test.txt and X_train.txt
  - This contains the actual measurements (= observations)
  - Each row is a set of measurement values (561) of 1 type of activity performed by 1 subject (person)
- Activity id (per measurement): y_test.txt, y_train.txt
  - This contains the type of activity for each row of the measurements

  - Each row has 1 integer value which is the identifier of the activity

- Subject id (per measurement): subject_test.txt, subject_train.txt
  - This contains the subject for each row of measurement
  - Each row has 1 inter value which is the identifier of the subject

Creating data set "tidy_data_mean_bymeasurement"

The starting point for this data set is tidy_data_full created as described in the previous section.

Steps

1. Create grouping: activity_id, subject_id, measurement
2. Summarize using the grouping with the mean of the value

Data dictionary

activities_data (data frame - 6 obs of 2 variables)

Column name	Class	Size	Example
act ivity_id	integer	1	5
activ ity_name	Factor, c haracter	6 levels	WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING

features_list (data frame - 561 obs. of 3 variables)

Column name	Class	Size	Example
f eature_id	integer	1	2
fe a ture_name	ch aracter		tBodyAcc-mean()-Y
u n iquenames	ch aracter		2tBodyAcc-mean()-Y

tidy_data_full (tibble - 679,734 × 4)

Name	Class	Size	Example
activity_id	Factor, character	6 levels	WALKING, WAL K I NG_UPSTAIRS, WA LKI N G _DOWNSTAIRS, SITTING, STANDING, LAYING
subject_id	Factor, integer	30 levels	26

Name	Class	Size	Example
measurement	Factor, character		1tBod y Acc-mean()-X
value	numeric	double	0.277 -0.0174

tidy_data_mean_bymeasurement (tibble - 11,880 × 4)

Name	Class	Size	Example
activity_id	Factor, character	6 levels	WALKING, WAL K I NG_UPSTAIRS, WA LKI N G _DOWNSTAIRS, SITTING, STANDING, LAYING
subject_id	Factor, integer	30 levels	26
measurement	Factor, character		1tBod y Acc-mean()-X
mean	numeric	double	0.277 -0.0174