Getting and Cleaning Data Course Project - Tidy Data Set

Maurice Shaffer

10/02/2020

Introduction

This codebook describes a tidy dataset created from data collected from the accelerometers from the Samsung Galaxy S smartphone. The data set created through this project is a tidy data set because it has the following three characteristics as defined by Hadley Wickham in the Tidy Data Paper:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Source Data

A full description of the source data is available at the site where the data was obtained:

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

Here is a link to a zip file containing that original source data:

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

Data Processing

The data processing for this project was done by a single R script called run_analysis.R that does the following.

1. Merges the training and the test sets to create one data set.
2. Extracts only the measurements on the mean and standard deviation for each measurement.
3. Uses descriptive activity names to name the activities in the data set
4. Appropriately labels the data set with descriptive variable names. The feature (measurement) labels were used as the variable names. The feature labels in the original data set contained special characters that needed to be replaced to make the labels useful as column names in a data frame. As a result, the following transformations were done on the feature labels:
    - Replace “()” with "_val"
    - Drop any remaining “)” at the end of a label
    - Replace remaining special characters with "_" these include: “,”, “-”, “(”, and “)”
5. Using the data set created above, the R script creates a second, independent tidy data set with the average of each variable for each activity and each subject. The new data set contains 30 test subjects with the measurements averaged for each of 6 activities for each subject. As a result there are 180 rows of data:

- 30 (subjects) x 6 (activities) = 180 (data rows).

6. And finally, writes the contents of the new tidy data set out to the text file: Getting-Cleaning-Data-Course-Tidy-Data.txt

This new tidy data set can be loaded and viewed using the following R code:

```
data <- read.table("Getting-Cleaning-Data-Course-Tidy-Data.txt", header = TRUE)

View(data)
```

The remainder of this codebook describes the content of this new tidy data set.

Variables

- **[,1] "subject_id"**
  - *Type:* numeric
  - *Description:* ID for the test subject who performed the activity measured. There were 30 test subjects so the IDs are in the range 1:30
- **[,2] "activity_label"**
  - *Type:* character
  - *Description:* a label describing the activity that the measurements are associated with. The valid activity labels are:
    - WALKING
    - WALKING_UPSTAIRS
    - WALKING_DOWNSTAIRS
    - SITTING
    - STANDING
    - LAYING
- **[,3] "tBodyAcc_mean_val_X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAcc-X measurement. This variable is now the mean of that variable for each activity and each subject.
- **[,4] "tBodyAcc_mean_val_Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAcc-Y measurement. This variable is now the mean of that variable for each activity and each subject.
- **[,5] "tBodyAcc_mean_val_Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAcc-Z measurement. This variable is now the mean of that variable for each activity and each subject.
- **[,6] "tGravityAcc_mean_val_X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tGravityAcc-X measurement. This variable is now the mean of that variable for each activity and each subject.
- **[,7] "tGravityAcc_mean_val_Y"**

  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tGravityAcc-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,8] "tGravityAcc__mean__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tGravityAcc-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,9] "tBodyAccJerk__mean__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAccJerk-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,10] "tBodyAccJerk__mean__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAccJerk-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,11] "tBodyAccJerk__mean__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAccJerk-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,12] "tBodyGyro__mean__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyro-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,13] "tBodyGyro__mean__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyro-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,14] "tBodyGyro__mean__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyro-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,15] "tBodyGyroJerk__mean__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyroJerk-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,16] "tBodyGyroJerk__mean__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyroJerk-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,17] "tBodyGyroJerk__mean__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyroJerk-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,18] "tBodyAccMag__mean__val"**

  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAccMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,19] "tGravityAccMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tGravityAccMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,20] "tBodyAccJerkMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyAccJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,21] "tBodyGyroMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyroMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,22] "tBodyGyroJerkMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the tBodyGyroJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,23] "fBodyAcc__mean__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyAcc-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,24] "fBodyAcc__mean__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyAcc-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,25] "fBodyAcc__mean__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyAcc-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,26] "fBodyAccJerk__mean__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyAccJerk-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,27] "fBodyAccJerk__mean__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyAccJerk-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,28] "fBodyAccJerk__mean__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyAccJerk-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,29] "fBodyGyro__mean__val__X"**

  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyGyro-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,30] "fBodyGyro__mean__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyGyro-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,31] "fBodyGyro__mean__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyGyro-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,32] "fBodyAccMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyAccMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,33] "fBodyBodyAccJerkMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyBodyAccJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,34] "fBodyBodyGyroMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyBodyGyroMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,35] "fBodyBodyGyroJerkMag__mean__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the mean of the fBodyBodyGyroJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,36] "tBodyAcc__std__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAcc-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,37] "tBodyAcc__std__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAcc-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,38] "tBodyAcc__std__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAcc-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,39] "tGravityAcc__std__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tGravityAcc-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,40] "tGravityAcc__std__val__Y"**

  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tGravityAcc-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,41] "tGravityAcc__std__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tGravityAcc-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,42] "tBodyAccJerk__std__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAccJerk-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,43] "tBodyAccJerk__std__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAccJerk-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,44] "tBodyAccJerk__std__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAccJerk-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,45] "tBodyGyro__std__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyro-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,46] "tBodyGyro__std__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyro-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,47] "tBodyGyro__std__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyro-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,48] "tBodyGyroJerk__std__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyroJerk-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,49] "tBodyGyroJerk__std__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyroJerk-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,50] "tBodyGyroJerk__std__val__Z"**
  - *Type:* numeric

  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyroJerk-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,51] "tBodyAccMag_std_val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAccMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,52] "tGravityAccMag_std_val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tGravityAccMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,53] "tBodyAccJerkMag_std_val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyAccJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,54] "tBodyGyroMag_std_val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyroMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,55] "tBodyGyroJerkMag_std_val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the tBodyGyroJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,56] "fBodyAcc_std_val_X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyAcc-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,57] "fBodyAcc_std_val_Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyAcc-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,58] "fBodyAcc_std_val_Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyAcc-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,59] "fBodyAccJerk_std_val_X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyAccJerk-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,60] "fBodyAccJerk_std_val_Y"**
  - *Type:* numeric

  - *Description:* the variable in the original data set was the standard deviation of the fBodyAccJerk-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,61] "fBodyAccJerk__std__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyAccJerk-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,62] "fBodyGyro__std__val__X"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyGyro-X measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,63] "fBodyGyro__std__val__Y"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyGyro-Y measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,64] "fBodyGyro__std__val__Z"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyGyro-Z measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,65] "fBodyAccMag__std__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyAccMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,66] "fBodyBodyAccJerkMag__std__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyBody-AccJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,67] "fBodyBodyGyroMag__std__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyBody-GyroMag measurement. This variable is now the mean of that variable for each activity and each subject.

- **[,68] "fBodyBodyGyroJerkMag__std__val"**
  - *Type:* numeric
  - *Description:* the variable in the original data set was the standard deviation of the fBodyBody-GyroJerkMag measurement. This variable is now the mean of that variable for each activity and each subject.