Codebook - Getting and Cleaning Data Course Project

Klismam Pereira

19/02/2020

Project and Data Overview

This project's objective was to "tidy" the original data tables using knowledge acquired during the Getting and Cleaning Data MOOC, by the Johns Hopkins University, available on the Coursera platform (https://www.coursera.org/learn/data-cleaning/home/welcome).

The data used in this project was generated in a study entitled "Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine" (IWAAL 2012). The study involved the measurement and process of 3-axial linear accelerations and 3-axial angular velocities obtained through the accelerometer and gyroscope of a smartphone. The device was attached to the volunteers' waist, which were asked to perform 6 activities while being recorded - walking, walking upstairs, walking downstairs, sitting, standing and laying. For more information on the study: http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.
For more information on how the script *run_analysis.R* works, please view the *README* file (.PDF or .md).

Variables

tidy_data_1

The resulting variables in the **tidy_data_1** archive are:

- **subject**: keys ranged from 1 to 30 that identify the the subject who carried out the experiment.
- **activity**: activity carried by the subject, can be one of six character values:
  - walking
  - walking_upstairs
  - walking_downstairs
  - sitting
  - standing
  - laying
- **domain**: indicate the domain of the signal:
  - time
  - frequency - a Fast Fourier Transform was applied to obtain the frequency domain of the signals
- **signal**: indicate the type of signal:
  - body_acceleration: filtered component of the acceleration signal

  - body_acceleration_magnitude: magnitude of the three-dimensional body acceleration signal calculated using Euclidean norm
  - body_acceleration_jerk: derivative of the body acceleration signal
  - body_acceleration_jerk_magnitude: magnitude of the three-dimensional body acceleration jerk signal calculated using Euclidean norm
  - gravity_acceleration: filtered component of the acceleration signal
  - gravity_acceleration_magnitude: magnitude of the three-dimensional gravity acceleration signal calculated using Euclidean norm
  - body_gyroscope: filtered gyroscope signal (angular velocity)
  - body_gyroscope_magnitude: magnitude of the three-dimensional body giroscope signal calculated using Euclidean norm
  - body_gyroscope_jerk: derivative of the gyroscope signal
  - body_gyroscope_jerk_magnitude: magnitude of the three-dimensional body giroscope jerk signal calculated using Euclidean norm

- **axis**: indicate the signal axis:

  - x
  - y
  - z
  - not_applicable: used for the magnitude signals

- **measure**: indicate which calculation the **values** column refer to:

  - mean: mean value
  - std: standard deviation

- **values**: The combination of the other variables characterize whether this is a mean or a standard deviation, if it's related to the time or frequency domain, etc.

```
str(read.table("./tidy_data_1.txt"))
```

```
## 'data.frame':    679735 obs. of  7 variables:
##  $ V1: Factor w/ 31 levels "1","10","11",..: 31 1 1 1 1 1 1 1 1 1 ...
##  $ V2: Factor w/ 7 levels "activity","laying",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ V3: Factor w/ 3 levels "domain","frequency",..: 1 3 3 3 3 3 3 3 3 3 ...
##  $ V4: Factor w/ 11 levels "body_acceleration",..: 11 1 1 1 1 1 1 1 1 1 ...
##  $ V5: Factor w/ 5 levels "axis","not_applicable",..: 1 3 3 3 3 3 3 3 3 3 ...
##  $ V6: Factor w/ 3 levels "mean","measure",..: 2 1 1 1 1 1 1 1 1 1 ...
##  $ V7: Factor w/ 649664 levels "-0.00010067201",..: 649664 632834 622030 620949 622637 620927 621877
```

---

tidy_data_2

The resulting variables in the **tidy_data_2** archive are the same as the tidy_data_1, except for the removal of **values** and the addition of:

- **average**: average of each variable for each activity and each subject.
  - Ex.: The combination "subject = 1, activity = walking_downstairs, domain = time, signal = body_acceleration, axis = x, measure = mean" has several observations in the column **values** of the tidy_data_1 set, the average of these observations is 0.272154196, and is shown in the appropriate row of the **average** column.

```
head(read.table("./tidy_data_2.txt"))
```

```
##        V1       V2     V3                 V4   V5      V6                 V7
## 1 subject activity domain             signal axis measure            average
## 2       1   laying   time body_acceleration    x    mean     0.22159824394
## 3       1   laying   time body_acceleration    x     std     -0.9280564692
## 4       1   laying   time body_acceleration    y    mean -0.0405139534294
## 5       1   laying   time body_acceleration    y     std   -0.83682740562
## 6       1   laying   time body_acceleration    z    mean   -0.11320355358
```

References

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012

[2] Getting and Cleaning Data, by Johns Hopkins University. https://www.coursera.org/learn/data-cleaning/home/welcome. Access on 17 of February, 2020.