2015 COURSERA DATA SCIENCE

GETTING & CLEANING DATA COURSE PROJECT

By: Dheeraj Agarwal

Email: dheeeraj.agarwal@gmail.com

Date: August 19, 2015

Contents

FAQs

What is this document?

This codebook acts as a basic reference document for the course project output generated as part of the 2015 Coursera Getting & Cleaning data project. Simply put it is a data dictionary put together to explain what the contents of the final output mean. It will details the possible values, meaning of the columns and also provide additional notes where ever necessary.

How to use this codebook?

Each section serves a different purpose and it may be helpful to look at the R code as well as the final output simultaneously to understand the definitions listed here.

The column definition section will help define what each column is named, what it means, field length, the possible values and any factor translation if required.

Use the following commands in R console or RStudio or any other IDE that you may use to work with R while referring to this document. The list of command(s) is not exhaustive and is by no means meant to indicate as the only way to understand this document better.

List of useful commands

- View(result)
- names(result)
- dim(result)
- head(result)
- unique(result$Activity)
- summary(result[n])   #where 'n' is between 3 and 68.

Where can I find the R code?

R code is posted publicly on GitHub at the user repository available here. This location will also have a README.md file as well as this codebook.

What is the source of this tidy data?

This tidy data is a result of cleaning activity on the source data which can be found at

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

A full description is available at the site where the data was obtained.

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

The rationale for coming up with this tidy data

Course objective:

To create one 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.
5. From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.

Rationale:

Given the above objectives, the R code, populates the column names from the features.txt file and then filters out any records that do not have the phrase 'mean()' or 'std()' in it. There are certain columns in the original data set that had phrases like 'meanFreq' etc. but have not been taken into consideration.

The original data set had 561 columns. 2 new columns were added to include subject and activity making a total of 563 columns in the data set.

This resultant was then subset using the criteria for mean and standard deviation to arrive at the table that had 68 columns in total.

The number of rows at this point in time was greater than 10K. Per the objective, the mean of each column was calculated and result grouped by Subject and Activity.

Total Subjects = 30

Total Activities = 6

Hence after aggregation and calculation of mean, total observations will be 180.

The final output hence will have **180 rows & 68 columns.**

Column Definitions

Descriptive columns [1:2]

Subject

Identifies the volunteer who was subjected to the test

2

1 – 30

[n] – Volunteer [n]   #example: n=7 will represent volunteer 7

Activity

The test is based on 6 different activities named 1-6 in the original data source, but with descriptions in the Tidy data set.

6

1 – 6

WALKING – 1

WALKING_UPSTAIRS – 2

WALKING_DOWNSTAIRS – 3

SITTING – 4

STANDING – 5

LAYING – 6

Measurement Columns [3:68]

Each table name is a composite of several parameters.

- The first alphabet in the table name denotes either a time or a frequency signal. If the signal is time then the first letter is 't' else it is 'f'.
- The second component is a breakdown of the third signal (see third component) where the signal is either 'Body' or 'Gravity'
- The third component is a measurement that captures the source of the signal. If the source is an accelerometer then it is represented by 'Acc'. If the source is gyroscope then it is represented by 'Gyro'.
- The fourth component is either 'mean()' or 'std()' which represents the mean of the measurement and the standard deviation respectively.
- Each of the above measurements or components together apply to a 3-axial measurement system, where each axis is represented by either 'X' or 'Y' or 'Z'.
- Certain columns with have a fifth element which could be either 'Jerk' or 'Mag' or a combination represented as 'JerkMag'. 'Jerk' represents a body's linear acceleration and angular velocity. 'Mag' is the magnitude of the three-dimensional signals.

Examples:

'tBodyAcc-mean()-X' is read as: *Mean of the time signal of Body measurement from an accelerometer in X axis.*

'fBodyGyroJerkMag-std()' is read as: *Standard Deviation of the magnitude of the frequency signal of the linear acceleration & angular velocity (jerk) of Body measurement from a gyroscope.*

tBodyAcc-mean()-X

17

-1..1

Mean of time domain for body acceleration in X direction

tBodyAcc-mean()-Y

19

-1..1

Mean of time domain for body acceleration in Y direction

tBodyAcc-mean()-Z

18

-1..1

Mean of time domain for body acceleration in Z direction

tBodyAcc-std()-X

18

-1..1

Standard deviation of time domain for body acceleration in X direction

tBodyAcc-std()-Y

17

-1..1

Standard deviation of time domain for body acceleration in Y direction

tBodyAcc-std()-Z

18

-1..1

Standard deviation of time domain for body acceleration in Z direction

tGravityAcc-mean()-X

17

Mean of time domain for gravitational acceleration in X direction

tGravityAcc-mean()-Y

18

-1..1

Mean of time domain for gravitational acceleration in Y direction

tGravityAcc-mean()-Z

12

-1..1

Mean of time domain for gravitational acceleration in Z direction

tGravityAcc-std()-X

    18

    -1..1

    Standard deviation of time domain for gravitational acceleration in X direction

tGravityAcc-std()-Y

    18

    -1..1

    Standard deviation of time domain for gravitational acceleration in X direction

tGravityAcc-std()-Z

    18

    -1..1

    Standard deviation of time domain for gravitational acceleration in X direction

tBodyAccJerk-mean()-X

    18

    -1..1

    Mean of time domain for body linear acceleration and angular velocity in X direction

tBodyAccJerk-mean()-Y

    18

    -1..1

    Mean of time domain for body linear acceleration and angular velocity in Y direction

tBodyAccJerk-mean()-Z

    20

    -1..1

    Mean of time domain for body linear acceleration and angular velocity in Z direction

tBodyAccJerk-std()-X

    18

    -1..1

    Standard deviation of time domain for body linear acceleration and angular velocity in X direction

tBodyAccJerk-std()-Y

    18

    -1..1

    Standard deviation of time domain for body linear acceleration and angular velocity in Y direction

tBodyAccJerk-std()-Z

    18

    -1..1

    Standard deviation of time domain for body linear acceleration and angular velocity in Z direction

tBodyGyro-mean()-X

15

-1..1

Mean of time domain for body using gyroscope in X direction

tBodyGyro-mean()-Y

19

-1..1

Mean of time domain for body using gyroscope in Y direction

tBodyGyro-mean()-Z

18

-1..1

Mean of time domain for body using gyroscope in Z direction

tBodyGyro-std()-X

18

-1..1

Standard deviation of time domain for body using gyroscope in X direction

tBodyGyro-std()-Y

19

-1..1

Standard deviation of time domain for body using gyroscope in Y direction

tBodyGyro-std()-Z

18

-1..1

Standard deviation of time domain for body using gyroscope in Z direction

tBodyGyroJerk-mean()-X

19

-1..1

Mean of time domain for body linear acceleration and angular velocity using a gyroscope in X direction

tBodyGyroJerk-mean()-Y

19

-1..1

Mean of time domain for body linear acceleration and angular velocity using a gyroscope in Y direction

tBodyGyroJerk-mean()-Z

19

-1..1

Mean of time domain for body linear acceleration and angular velocity using a gyroscope in Z direction

tBodyGyroJerk-std()-X

    18

    -1..1

    Standard deviation of time domain for body linear acceleration and angular velocity using a gyroscope in X direction

tBodyGyroJerk-std()-Y

    18

    -1..1

    Standard deviation of time domain for body linear acceleration and angular velocity using a gyroscope in Y direction

tBodyGyroJerk-std()-Z

    18

    -1..1

    Standard deviation of time domain for body linear acceleration and angular velocity using a gyroscope in Z direction

tBodyAccMag-mean()

    18

    -1..1

    Mean of the magnitude of the time domain for body acceleration

tBodyAccMag-std()

    18

    -1..1

    Standard deviation of the magnitude of the time domain for body acceleration

tGravityAccMag-mean()

    18

    -1..1

    Mean of the magnitude of the time domain for gravitational acceleration

tGravityAccMag-std()

    18

    -1..1

    Mean of the magnitude of the time domain for gravitational acceleration

tBodyAccJerkMag-mean()

    18

    -1..1

    Mean of the magnitude of time domain for body linear acceleration and angular velocity

tBodyAccJerkMag-std()

    19

    -1..1

    Standard deviation of the magnitude of time domain for body linear acceleration and angular velocity

tBodyGyroMag-mean()

18

-1..1

Mean of the magnitude of time domain for body using a gyroscope

tBodyGyroMag-std()

18

-1..1

Standard deviation of the magnitude of time domain for body using a gyroscope

tBodyGyroJerkMag-mean()

18

-1..1

Mean of the magnitude of time domain for body linear acceleration and angular velocity using a gyroscope

tBodyGyroJerkMag-std()

18

-1..1

Standard deviation of the magnitude of time domain for body linear acceleration and angular velocity

fBodyAcc-mean()-X

18

-1..1

Mean of frequency domain for body acceleration in X direction

fBodyAcc-mean()-Y

18

-1..1

Mean of frequency domain for body acceleration in Y direction

fBodyAcc-mean()-Z

18

-1..1

Mean of frequency domain for body acceleration in Z direction

fBodyAcc-std()-X

18

-1..1

Standard deviation of frequency domain for body acceleration in X direction

fBodyAcc-std()-Y

14

-1..1

Standard deviation of frequency domain for body acceleration in Y direction

fBodyAcc-std()-Z

  18

  -1..1

  Standard deviation of frequency domain for body acceleration in Z direction

fBodyAccJerk-mean()-X

  18

  -1..1

  Mean of frequency domain for body linear acceleration and angular velocity in X direction

fBodyAccJerk-mean()-Y

  19

  -1..1

  Mean of frequency domain for body linear acceleration and angular velocity in Y direction

fBodyAccJerk-mean()-Z

  18

  -1..1

  Mean of frequency domain for body linear acceleration and angular velocity in Z direction

fBodyAccJerk-std()-X

  18

  -1..1

  Standard deviation of frequency domain for body linear acceleration and angular velocity in X direction

fBodyAccJerk-std()-Y

  14

  -1..1

  Standard deviation of frequency domain for body linear acceleration and angular velocity in Y direction

fBodyAccJerk-std()-Z

  18

  -1..1

  Standard deviation of frequency domain for body linear acceleration and angular velocity in Z direction

fBodyGyro-mean()-X

  18

  -1..1

  Mean of frequency domain for body using a gyroscope in X direction

fBodyGyro-mean()-Y

  18

  -1..1

  Mean of frequency domain for body using a gyroscope in Y direction

fBodyGyro-mean()-Z

18

-1..1

Mean of frequency domain for body using a gyroscope in Z direction

fBodyGyro-std()-X

18

-1..1

Standard deviation of frequency domain for body using a gyroscope in X direction

fBodyGyro-std()-Y

19

-1..1

Standard deviation of frequency domain for body using a gyroscope in Y direction

fBodyGyro-std()-Z

18

-1..1

Standard deviation of frequency domain for body using a gyroscope in Z direction

fBodyAccMag-mean()

18

-1..1

Mean of the magnitude of the frequency domain for body

fBodyAccMag-std()

18

-1..1

Standard deviation of the magnitude of the frequency domain for body

fBodyBodyAccJerkMag-mean()

19

-1..1

Mean of the magnitude of frequency domain for body linear acceleration and angular velocity

fBodyBodyAccJerkMag-std()

18

-1..1

Standard deviation of the magnitude of frequency domain for body linear acceleration and angular velocity

fBodyBodyGyroMag-mean()

18

-1..1

Mean of the magnitude of frequency domain for body using a gyroscope

fBodyBodyGyroMag-std()

18

-1..1

Mean of the magnitude of frequency domain for body using a gyroscope

fBodyBodyGyroJerkMag-mean()

18

-1..1

Mean of the magnitude of frequency domain for body linear acceleration and angular velocity using a gyroscope

fBodyBodyGyroJerkMag-std()

18

-1..1

Standard deviation of the magnitude of frequency domain for body linear acceleration and angular velocity using a gyroscope

NOTE:

This document is only for reference and created as an original document based on my understanding of the course objective and its intention. If you are planning to take this course or currently taking it, please use this only as a reference and a guide to understand my viewpoint and approach. Do not copy stuff from here. It may help you complete the course but will not help you learn write a codebook.