Coursera - John Hopkins University

Getting and Cleaning Data - Week 4 Assignment

Project Description

Overview

The purpose of this assignment is to prepare tidy data from the provided data set which can then be used for further analysis. A summary set of data is also required. The stated purpose of the project is to demonstrate the student's ability to work with, and clean a data set.

Review Criteria

The following paragraphs address the stated Review Criteria, in terms of how they are intended to be addressed in this assignment submission

1. The submitted data set is tidy

It is perhaps a matter of interpretation as to what constitutes a 'tidy' data set in this instance. The rationale for my approach is detailed in the README file and later in this CodeBook.

2. The Github repo contains the required scripts.

All (three) scripts have been prepared in RStudio using RStudio Git and Github desktop functionality to then store to my Github account, in this repository

Scripts written are:

- 'run_analysis.R' - does all data loading, cleaning and output.
- 'README.Rmd' - queries data and prepares the README.md file (this document). PDF and HTML versions can also be run from the RStudio 'Knit' menu.
- 'Codebook.Rmd' - similar to above, produces the Codebook.md file (as well as PDF and HTML versions) by running the 'run_analysis.R' script and then querying and assembling appropriate data (augmented by manual text input)

3. GitHub contains a code book that modifies and updates the available codebooks with the data to indicate all the variables and summaries calculated, along with units, and any other relevant information.

As mentioned in previous section the github repository for this assignment contains Codebook in md, html and pdf versions as well as the 'Codebook.Rmd' script which produces and updates them.

4. The README that explains the analysis files is clear and understandable.

The assignment documentation further states the following in reltion to the README.md file:

"You should also include a README.md in the repo with your scripts. This repo explains how all of the scripts work and how they are connected."

This README document is intended to meet this requirement. As mentioned the 'README.Rmd' script which produces/updates this document is also available in the assignment repository.

The CodeBook also contains detailed information on the R code, and transformations undertaken.

5. The work submitted for this project is the work of the student who submitted it.

This work is solely my responsibility, supported by much reading of:

- Course notes
- Week 4 discussion forum (notably "Getting and Cleaning Data",David Hood 2015, link)
- Hadley Wickham's article[1] on 'Tidy Data'.
- Stack Overflow and other Google sources etc

The number of commits to my repository might support this.

Key output documents for this assignment comprise:

- 'run_analysis.R' - the R script which reads data and then formats and tidies data.
- 'all_data.txt' - output text file containing the 'tidied' data, not summarised.
- 'summary_data.txt' - output text file, summarised to show average data by subject, activity and signal.
- 'README.md - this document, which describes project approach and rationale.
- 'CodeBook.Rmd' - RMarkdown file which generates and updates the codebooks, as below:
  - 'CodeBook.md' - markdown version
  - 'CodeBook.pdf' - pdf version
  - 'CodeBook.html' - html version

The above are all provided on this Github repo. The original data sources are also saved on repo.

Running the Analysis

R Code Prerequisites

Two R packages are required to run the 'run_analysis.R' script (and I have not automated their installation if not already installed) :

- tidyverse - obviously actually quite a few separate packages will load.
- data.table - the 'fread' function is used for file reading preference and to allow selective column loads.

In addition, the CodeBook.Rmd and READMe.Rmd scripts use knitR package.

Processing

All script to run the analysis is in 'run_analysis.R'

Data input files should be in the folder 'UCI HAR Dataset' in the Working Directory. As shown in the 'Input Files' section below there are test and train subfolders required therein.

The two output files (all_data.txt and summary_data.txt) will be saved to the Working Directory.

Output - All_data dataframe

For information, first 6 rows of 'all_data' dataframe are as follows:

subject_id	activity	date_time_id	data_set	signal	mean	std_dev
1	LAYING	1	train	fBodyAcc-X	-0.9263140	-0.9103749
1	LAYING	1	train	fBodyAcc-Y	-0.8869485	-0.9056274

[1]Wickham H.(2014), "Tidy Data", *The Journal of Statistical Software*, vol.59.

subject_id	activity	date_time_id	data_set	signal	mean	std_dev
1	LAYING	1	train	fBodyAcc-Z	-0.9287921	-0.8810446
1	LAYING	1	train	fBodyAccJerk-X	-0.9491758	-0.9502229
1	LAYING	1	train	fBodyAccJerk-Y	-0.8980041	-0.9036754
1	LAYING	1	train	fBodyAccJerk-Z	-0.9729599	-0.9824606

Output - Summary_Data dataframe

Similarly, first 6 rows of summary data is as follows:

signal	activity	subject_id	average_mean	average_std_dev
fBodyAcc-X	LAYING	1	-0.9390991	-0.9244374
fBodyAcc-X	LAYING	2	-0.9767251	-0.9732465
fBodyAcc-X	LAYING	3	-0.9806656	-0.9836911
fBodyAcc-X	LAYING	4	-0.9588021	-0.9524649
fBodyAcc-X	LAYING	5	-0.9687417	-0.9649539
fBodyAcc-X	LAYING	6	-0.9391143	-0.9324629

Data Set Information

Source

This assignment uses data collected from the accelerometers from the Samsung Galaxy S smartphone as part of an experiment which is explained in detail on the following web site (webLink)

The data can be downloaded from the following site, although it is also stored in this repo. (webLink)

Overview of Original Data

The original data was produced from an experiment 'Human Activity Recognition Using Smartphones Dataset'[2]

An overview of the experiment as provided by the related README.txt (which is also contained in the repo for this assignment) is:

"The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data."

A detailed description of the data is contained in the CodeBook, filed in this repo.

Input Files

Key data input files sourced from the original data source are as follows:

[2]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

- ‘features.txt’ - List of all features.
- ‘activity_labels.txt’ - Links the class labels with their activity name.
- ‘train/X_train.txt’ - Training set.
- ‘train/y_train.txt’ - Training labels.
- ‘train/subject_train.txt’- subject_ids for each row of train set.
- ‘test/X_test.txt’ - Test set.
- ‘test/y_test.txt’ - Test labels.
- ‘test/subject_test.txt’ - subject_ids for each row if test set.

All these files are accessed by ‘run_analysis.R’ script.

The other key files used to document the original experiment data sets are:

- ‘README.md’ file - overview of experiment and related data files
- ‘features_info.txt’ - Shows information about the variables used on the feature vector.

Many other files are contained in the source data folders, but none were needed to complete this assignment, so are not referenced in this document or the CodeBook or any of the scripts.

Tidy Data - Rationale for Structure

The CodeBook for this assignment details the full ‘tidy data’ structure of the output data as contained in all_data.txt file.

The structure as illustated previously with the head() function can be characterised perhaps as long/narrow.

Reasons for this approach is as follows:

- The original experiment derived 33 signals for which they then calculated 17 measures across the 6 activities and 30 subjects. The 33 signals had the following components:
  - ‘f’ (frequency) v ‘t’ (time) signals
  - ‘X’ v ‘Y’ v ‘Z’ directional signals
  - ‘Body’ v ‘Gravity’
  - ‘Acc’ v ‘Gyro’ etc
- In the source data provided in the experiment data documentation these were all referred to as ‘Signals’.
- The documentation then referred to 17 variables which were calculated for each Signal (by Activity and Subject). Refer ‘features_info.txt’ file in source data files.
- Without additional infomration seems to be a reasonable approach to use the ‘Signals’ as observations and the 17 variables (but limited to mean and std, as per assignment requirements) as variables.
- I am not a subject matter expert so have no reason to break the signals’ names into their component parts,for example by extacting ‘f’ and ‘t’ components into separate columns for the mean and std_dev measures. Doing so, would also introduce NULL components, as ‘f’ and ‘t’ signals are not identical.
- Also, the Summary Analysis required can be easily accomplished with the data in this structure, so there is no reason to complicate the code further. I may have reconsidered if I knew what any further analysis might be, so could perhaps judge if another structure is more appropriate. I do not have that information however.