**Codebook** – “run\_analysis.R” assignment

|  |
| --- |
| title: run\_analysis.R |
| author: Jeffrey Gonlin |
| date: 3.April.2016 |
| output: |
| word\_documents:  “Codebook run\_analysis.R assignment” and  “Instruction List for run\_analysis.R assignment” |
| keep\_md: yes |
| --- |
|  |
| # Project Description |
| This work was done as an assignment in the Coursera Data Cleaning course, week 4. It massages data sourced from the link below, where a project description can be found: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>  Citation: Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013. |
|  |
| #Study design and data processing  Data was collected from cell phone sensors while 30 subjects each engaged in 6 different activities. Noise filters were applied to the raw signals, and a Butterworth low-pass filter was used to separate gravitational effects on acceleration. Subjects and data were separated into two groups, with 9 assigned to the “test” group and 21 to the “train” group. See: “UCI HAR Dataset/README.txt.” and “UCI HAR Dataset/features\_info.txt” in the source zip file for further details of study design, pre-processing and filtering of original raw signal samples. |
|  |
| #Collection of the raw data |
| Data was collected from cell phone sensors while 30 subjects each engaged in 6 different activities using the same model Samsung phone. See: “UCI HAR Dataset/README.txt.” and “UCI HAR Dataset/features\_info.txt” in the source sip file for fuller description. |
|  |
| #Notes on the original (raw) data |
| Data derived from cell phone sensors and found in ‘inertial’ data files was transformed (e.g., to adjust for influence of gravity) and used to estimate variables for training and testing. These estimates are found in “UCI HAR Dataset/test/X\_test.txt” and “UCI HAR Dataset/test/X\_train.txt”. I do not use the original raw data; I do use the resulting ‘test’ and train’ data..  **Units**: note that datapoints in these files have already been normalized with a range between -1 to +1. For information on original inertial readings, see “README” and “features\_info” .txt files 3 and 4 mentioned below |
|  |
| #Creating the tidy datafile  The reader of this codebook and reviewer of my coding is forewarned:  - rather than sketch out a strategy in advance, this journey through the assignment has been done with more of a ‘straight out of the head wandering’ approach.  - the coding is neither elegant nor efficient (at ~60 seconds execution and 169 lines of code and comment). Think of it as a rambling first draft.  - Plenty of ‘thinking out loud’ comments have been included throughout the code, so my hope is that one can follow my train of thought while reading along.  The outcomes prescribed for this exercise, and the approximate code line one should seek for review, are:  1) Merges the training and the test sets to create one data set. See: lines 67 – 69 for ‘subjects’, ‘activities’ and ‘feature\_vectors’ test/train combined files. See: line 99 for ‘subjects & activities’ combined file See: line 150 for “one\_data\_set” containing both test & train subjects, activities and feature\_vectors measurements.   1. 2) Extracts only the measurements on the mean and standard deviation for each measurement. See: lines 114, 115 for colnames of mean and standard dev measurements respectively. See: line 139 for combined list of colnames to be extracted See: line 144 for dataset containing only mean and std variables and valules. 2. 3) Uses descriptive activity names to name the activities in the data set See: line 88 for naming of activities 3. 4) Appropriately labels the data set with descriptive variable names. See: lines 96, 97 for ‘subjects’ and ‘activites’ colnames See: line 109 for ‘feature\_vectors’ colnames 4. 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. See: lines 156 for establishing and 162 for then grouping and summarizing the dataset.   Note: **Detailed code** can be found in the word.doc “**Instruction List** for run\_analysis.R assignment”. |
|  |
| #Guide to creating and cleaning the tidy data file  fileUrl = "https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip"  download.file(fileUrl, destfile = "./Dataset.zip")  #**NOTE**: **After downloading but before running the entire code, please set up the WD file as follows**:  ## **TO PRESERVE THE ORIGINAL DATA**: Go to the desktop / WD 'Assignment' folder and open "Dataset.zip"  ## **COPY the "UCI HAR Dataset"** folder but take no other action with it.  ## **GO UP ONE LEVELAND PASTE IT. THIS IS THE FILE I WILL WORK WITH**.  ## Both "Dataset.zip" and "UCI HAR Dataset" should be found in: "C:/Users/Jeff/Desktop/Week4 Assignment"  ## Sys.Date @ time of test run was 2016-04-02 |
| **Dataset.zip** contains 32 entries. The file “UCI HAR DSataset” is of interest.  “UCI HAR Dataset” contains 4 .txt files and folders “test” and “train”:  Files 3 and 4, “features\_info” and “README” are best read in a text program.  Within both “test” and “train” folders one finds:   * 1. folders relating to “Inertial Signals” files. These are original raw data machine readings, which samplings have been processed further to produce the ‘test’ and ‘train’ measurements of interest.   2. Files 16:18 and 30:32, relating to “subject\_train” “subject\_test”, “X\_train”, “X\_test”, and “y-train”, “y\_test” are files we are interested in.  | **Name** | | | **Length** | **Date** | | --- | --- | --- | --- | --- | |  |  |  | |  | | **1** | UCI HAR Dataset/activity\_labels.txt | 80 | | 2012-10-10 15:55:00 | | **2** | UCI HAR Dataset/features.txt | 15785 | | 2012-10-11 13:41:00 | | **3** | UCI HAR Dataset/features\_info.txt | 2809 | | 2012-10-15 15:44:00 | | **4** | UCI HAR Dataset/README.txt | 4453 | | 2012-12-10 10:38:00 | | **5** | UCI HAR Dataset/test/ | 0 | | 2012-11-29 17:01:00 | | **6** | UCI HAR Dataset/test/Inertial Signals/ | 0 | | 2012-11-29 17:01:00 | | **7** | UCI HAR Dataset/test/Inertial Signals/body\_acc\_x\_test.txt | 6041350 | | 2012-11-29 15:08:00 | | **8** | UCI HAR Dataset/test/Inertial Signals/body\_acc\_y\_test.txt | 6041350 | | 2012-11-29 15:08:00 | | **9** | UCI HAR Dataset/test/Inertial Signals/body\_acc\_z\_test.txt | 6041350 | | 2012-11-29 15:08:00 | | **10** | UCI HAR Dataset/test/Inertial Signals/body\_gyro\_x\_test.txt | 6041350 | | 2012-11-29 15:09:00 | | **11** | UCI HAR Dataset/test/Inertial Signals/body\_gyro\_y\_test.txt | 6041350 | | 2012-11-29 15:09:00 | | **12** | UCI HAR Dataset/test/Inertial Signals/body\_gyro\_z\_test.txt | 6041350 | | 2012-11-29 15:09:00 | | **13** | UCI HAR Dataset/test/Inertial Signals/total\_acc\_x\_test.txt | 6041350 | | 2012-11-29 15:08:00 | | **14** | UCI HAR Dataset/test/Inertial Signals/total\_acc\_y\_test.txt | 6041350 | | 2012-11-29 15:09:00 | | **15** | UCI HAR Dataset/test/Inertial Signals/total\_acc\_z\_test.txt | 6041350 | | 2012-11-29 15:09:00 | | **16** | UCI HAR Dataset/test/subject\_test.txt | 7934 | | 2012-11-29 15:09:00 | | **17** | UCI HAR Dataset/test/X\_test.txt | 26458166 | | 2012-11-29 15:25:00 | | **18** | UCI HAR Dataset/test/y\_test.txt | 5894 | | 2012-11-29 15:09:00 | | **19** | UCI HAR Dataset/train/ | 0 | | 2012-11-29 17:01:00 | | **20** | UCI HAR Dataset/train/Inertial Signals/ | 0 | | 2012-11-29 17:01:00 | | **21** | UCI HAR Dataset/train/Inertial Signals/body\_acc\_x\_train.txt | 15071600 | | 2012-11-29 15:08:00 | | **22** | UCI HAR Dataset/train/Inertial Signals/body\_acc\_y\_train.txt | 15071600 | | 2012-11-29 15:08:00 | | **23** | UCI HAR Dataset/train/Inertial Signals/body\_acc\_z\_train.txt | 15071600 | | 2012-11-29 15:08:00 | | **24** | UCI HAR Dataset/train/Inertial Signals/body\_gyro\_x\_train.txt | 15071600 | | 2012-11-29 15:09:00 | | **25** | UCI HAR Dataset/train/Inertial Signals/body\_gyro\_y\_train.txt | 15071600 | | 2012-11-29 15:09:00 | | **26** | UCI HAR Dataset/train/Inertial Signals/body\_gyro\_z\_train.txt | 15071600 | | 2012-11-29 15:09:00 | | **27** | UCI HAR Dataset/train/Inertial Signals/total\_acc\_x\_train.txt | 15071600 | | 2012-11-29 15:08:00 | | **28** | UCI HAR Dataset/train/Inertial Signals/total\_acc\_y\_train.txt | 15071600 | | 2012-11-29 15:08:00 | | **29** | UCI HAR Dataset/train/Inertial Signals/total\_acc\_z\_train.txt | 15071600 | | 2012-11-29 15:08:00 | | **30** | UCI HAR Dataset/train/subject\_train.txt | 20152 | | 2012-11-29 15:09:00 | | **31** | UCI HAR Dataset/train/X\_train.txt | 66006256 | | 2012-11-29 15:25:00 | | **32** | UCI HAR Dataset/train/y\_train.txt | 14704 | | 2012-11-29 15:09:00 |   See: line 17 of the code or the “datasetunzip” data in the Global Environment to view.  Steps taken after perusal of the various datasets:   1. Combine ‘test’ and ‘train’ data - established 3 combined datasets “subjects”, “activities” and “feature\_vectors”  for the dataset pairs  ‘subject\_test’ & ‘subject\_train’, ‘y\_test’ & ‘y\_train’ and ‘X\_test’ & ‘X\_train’ respectively. 2. Substituted activity names (e.g., walking) for numerical codes representing the activities 3. Renamed columns (e.g., ‘subject’ instead of V1) 4. Merged ‘subjects’ and ‘activities’ data 5. Extracted 79 “mean” and “std” (standard deviation) related measurement variables from the total 561 variables,  - first by assembling a vector with column names and  - then applying it to the larger “feature\_vectors” dataset  - to arrive at the “feature\_vectors\_mean\_std\_cols” dataset.  (Note: the “X”, “Y” and “Z” mean dimensional variables have been included in this set. I found the instructions a touch vague on this point, and reckon they can more easily be removed than reinstated, if ultimately not wanted.) 6. Obtained “one\_data\_set” by combining the “feature\_vectors\_mean\_std\_cols” dataset with “”subjects\_activites” dataset. 7. Using “one\_data\_set” I grouped and averaged the readings per subject, per activity to arrive at “tidy\_data\_grouped\_averaged”. |
|  |
|  |
| One can argue whether separate tables for mean and std should be established. If data had simply been grouped, I would have used two tables. At this point we have and intend to have only one set of averages for these selected measurement variables. Therefore I am keeping them on one table - the table relating to grouped averaged data.  Fortunately, examination of the structure of the datasets revealed they were in good order in terms of missing data or other problems (perhaps due to the mechanical recording of data and previous processing), so this became more an exercise in otherwise wrangling the data.  See: “readme” for the code and comments. |
|  |
| ##Description of the variables:  (You are encouranged to read the original documentation on variables and measurements along with this description.)  **“subject”** - the persons, by number, who took part in this project. 30 persons total took part, with 9 test subjects and 21 train subjects.  str(subject\_test)  'data.frame': 2947 obs. of 1 variable:  $ V1: int 2 2 2 2 2 2 2 2 2 2 ...  str(subject\_train)  'data.frame': 7352 obs. of 1 variable:  $ V1: int 1 1 1 1 1 1 1 1 1 1 ...  “**activity**” - ‘activity\_labels.txt’ contains a list of 6 activities engaged in as listed below. “activities” is the name given to the y\_test and y\_train combined file. These data files use an integer code to represent each type of activity, e.g., 5 = Standing. “activity” is the colname used for this variable.  str(activities)  Classes ‘data.table’ and 'data.frame': 10299 obs. of 2 variables:  $ activity: chr "standing" "standing" "standing" "standing" ...  $ index : int 1 2 3 4 5 6 7 8 9 10 ...  code activity  # 1 WALKING  # 2 WALKING\_UPSTAIRS  # 3 WALKING\_DOWNSTAIRS  # 4 SITTING  # 5 STANDING  # 6 LAYING  str(y\_test)  'data.frame': 2947 obs. of 1 variable:  $ V1: int 5 5 5 5 5 5 5 5 5 5 ...  str(y\_train)  'data.frame': 7352 obs. of 1 variable:  $ V1: int 5 5 5 5 5 5 5 5 5 5 ...  Features / mean & std /  str(features)  'data.frame': 561 obs. of 2 variables:  $ V1: int 1 2 3 4 5 6 7 8 9 10 ...  $ V2: Factor w/ 477 levels "angle(tBodyAccJerkMean),gravityMean)",..: 243 244 245 250 251 252 237 238 239 240 ...  “**features**” are the measurement variables derived from the inertial measurements. The nomenclature used below, although perhaps not ‘tidy’, was considered better than improvements considered. These features were combined with various statistical measurement labels as detailed in “features\_info.txt”. We are only interested in variables relating to mean and standard deviation (mean, std).  tBodyAcc-XYZ  tGravityAcc-XYZ  tBodyAccJerk-XYZ  tBodyGyro-XYZ  tBodyGyroJerk-XYZ  tBodyAccMag  tGravityAccMag  tBodyAccJerkMag  tBodyGyroMag  tBodyGyroJerkMag  fBodyAcc-XYZ  fBodyAccJerk-XYZ  fBodyGyro-XYZ  fBodyAccMag  fBodyAccJerkMag  fBodyGyroMag  fBodyGyroJerkMag  str(X\_train)  'data.frame': 7352 obs. of 561 variables:  $ V1 : num 0.289 0.278 0.28 0.279 0.277 ...  $ V2 : num -0.0203 -0.0164 -0.0195 -0.0262 -0.0166 ...  $ V3 : num -0.133 -0.124 -0.113 -0.123 -0.115 ...  str(X\_test)  'data.frame': 2947 obs. of 561 variables:  $ V1 : num 0.257 0.286 0.275 0.27 0.275 ...  $ V2 : num -0.0233 -0.0132 -0.0261 -0.0326 -0.0278 ...  $ V3 : num -0.0147 -0.1191 -0.1182 -0.1175 -0.1295 ...  A combined dataset will have dimensions 10299 obs. Of 561 variables. That dataset has been called “featrure\_vectors”.  A full listing of the 561 variables can be found in the dataset “fearures” (see: line 103).  The mean and std variables is to be found in the dataset “feature\_vectors\_mean\_std\_labels” (see: line 139). |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
| ##Sources |
| Data for this assignment was sourced from: <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip> |
|  |