Code Book for Project

This file is the codebook for producing the tidy data set required in the project for the Coursera course "Getting and Cleaning Data".

Contents

Introduction:

This is the codebook for the tidy data set project. The goal is to get the file into the data format as described in Hadley Wickham's paper on tidy data found at http://www.jstatsoft.org/v59/i10/paper.

The final output data set meets the criteria:

1. Each column is a variable
2. Each row is an observation or measurement.
3. All of the non-factorial variables (Tester and Activity) are unit independent so there is no need to alter any of them of put some of them in a separate table.

The raw dataset comes from a study of Human Activity Recognition using Smartphones.[1] The README.txt file included with the downloaded dataset has the details of the experiment. It explains that there were 30 volunteers (I call them Testers) between the ages of 19-48 years. Each of them wore a smartphone (Samsung Galaxy S II) and performed 6 activities of WALKING, STANDING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, and LAYING. The tester and the activity are the fixed variables and I have placed them as the first two columns in the data set. For each of these activities, the testers were divided into two groups of trainers and testers and each is presented in a separate dataset. Our instructions are to merge the two datasets and calculate the mean of all variables which are themselves a mean or standard deviation.

Process:

1. Load R packages:
    a. dplyr
    b. stringr
2. Create a data directory if does not already exist in the current working directory.

1

[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

3. If the file does not already exist in ./data/Dataset.zip, then download it using download.file with the mode set to “wb” from https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip" and unzip it. Otherwise, do not download it.
4. The data will be in a directory in the current folder called “UCI HAR Dataset”. It contains the files and folders for the analysis.
    a. README.txt
    b. **activity_labels.txt**: These are the factors used to describe the activity the tester was engaging in at the time the measurement was made.
    c. **features.txt**: The variable names for the measurements in X_train.txt and X_test.txt.
    d. **features_info.txt**: A description of the variables in features.txt.
    e. folder – train: contains the training data sets.
        i. **subject_train**: A file listing which tester conducted the test in each row.
        ii. **y_train**: A file listing the activity the tester was engaged each row.
        iii. **X_train**: A file containing the measurements as described in the README.txt file downloaded with the data set.
    f. folder –test : contains the testing datasets:
        i. **subject_test**: A file listing which tester conducted the test in each row.
        ii. **y_test**: A file listing the activity the tester was engaged each row.
        iii. **X_test**: A file containing the measurements as described in the README.txt file
5. Read the **activity_labels.txt** file into the data frame **activity_labels** using **read.csv**.
    a. Assign it the column names **Record** and **ActivityType**.
    b. Print the file to view it contents.
6. Read the activity file **y_train.txt** and assign it to the data frame **y_train**.
    a. Assign it the column name **Activity**
    b. The number of levels in this file is 6 listed as the numbers 1:6.
    c. Rename the levels using the function factor with the levels and names coming from the **activity_labels** data set. As these are the column names that we will use for our data, we will condition these names in the data frame **features. Note:** this will not affect the raw data set file as we are not writing this file back out.
        i. Remove unwanted characters using the local function **f_remove_chars**. This function removes the following characters:
            1. (
            2.)
            3. ,
            4. 1 or more spaces in a row
            5. Hyphen –
            6. Underscore _
            7. Trims any leading and training spaces using **str_trim** from the **stringr** package.
        ii. Since CamelCode seems to the agreed on standard for variable names, then variable names in the df **y_train** will be changed. Let it be noted that some

people prefer all lower case. I changed to camel code because of the forums for the course. Also note that after we subset the columns to only the ones that we want, later there will only be three words which need to be fixed.

1. mean to Mean
2. std to Std and
3. gravity to Gravity

iii. I have chosen to leave the prefixed t and f on all variable names as it is a single character and represents the domain of the variable as to whether it is t for the time domain or f for the frequency domain.

7. The request is to keep only the columns which represent the mean and frequency of the measurements. In order to do this, the conditioned features data frame is subset using grep and assigned to a variable called **columns_to_keep**. In order to help create the code book **columns_to_ keep** is written to a file called **df_columns_to_keep.txt**.
8. Next we read in the subject data into **subject_train** and assign it the column name **tester.**
9. Now we read in the actual measurement variables in **X_table.txt** into **x_train**. Note that this data set is white space delimited. Sometime there is more than one space between variables. Therefore, we use read.table as it allows for any white space between variables if the separator is set to “”.
10. Subset the data into **sub_x_train** using the variable we created above **columns_to_keep**. Now we do not have any issues with duplicate column names. Note that if you do not remove the characters from the feature variables as we did above the R will treat some of the column names as duplicates.
11. Now we check **sub_x_train** for NA values. I found that there are none in the dataset.
12. Now, the data set **data_train** is created by column binding **subject_train**, **y_train**, and **sub_x_train**. This dataset has 7352 observations and 86 variables.
13. Next we read in the subject data into **subject_test** and assign it the column name **tester.**
14. Now we read in the actual measurement variabes in **X_test.txt** into **x_test**. Note that this data set is white space delimited. Sometime there is more than one space between variables. Therefor we use read.table as it allows for any white space between variables if the separator is set to “”.
15. Subset the data into **sub_x_test** using the variable we created above **columns_to_keep**. Now we do not have any issues with duplicate column names. Note that if you do not remove the characters from the feature variables as we did above the R will treat some of the column names as duplicates.
16. Now we check s**ub_x_test** for NA values. I found that there are none in the dataset.
17. Now, the data set **data_test** is created by column binding **subject_test**, **y_test**, and **sub_x_test**. This dataset has 2947 observations and 86 variables.
18. The next step is to append **rbind data_test** to **data_train. B**efore verified that:
    a. The dimensions of the data sets are the same.
    b. The tester is unique between the two files.
19. Now the two files are combined, so we want to get the mean of the columns. To do this use the **dplyr** package functions to create the tidy data set **out_data**:

a. Use **tbl_df** to convert data to a data from table which is required for the **dplyr** tools. The output dataset is **tbl_data.** The table has 10,299 observations and 88 variables.
b. Use **group_by** to group the table by **tester** and **activity**. The output dataset is **g_data**
c. Use **summarise_each** to apply the mean to the **non grouped** columns. The output dataset is the final dataset **out_data.** The final dataset has 180 observations of 88 variables.

20. Write out **out_data** to the tidy data set file **tidy_data_set.txt.**

The final data set has is of the form:

Testor	Activity	Mean of 1st variable	Mean of 2nd variable
1	WALKING	*value*	*value*
1	STANDING	*Value*	*value*

21.

Data Dictionary: Description of Variables:

Explanation of the variable and valid options: In the list below, a prefix of "**t"** indicates that the variable is in the time domain and a prefix of "**f"** indicates that it is in the frequency domain.

1. Tester – The person who conducted the test.
    a. Tester has values of 1 to 30 each one representing a different testor.
2. ActivityType -
    a. WALKING
    b. WALKING_UPSTAIRS
    c. WALKING_DOWNSTAIRS
    d. SITTING
    e. STANDING
    f. LAYING

Time Domain Variables (all units normalized to 1)

3. tBodyAccMeanX – the mean of the body acceleration in the X direction.
4. tBodyAccMeanY– the mean of the body acceleration in the Y direction.
5. tBodyAccMeanZ– the mean of the body acceleration in the Z direction.
6. tBodyAccStdX– the standard deviation of the body acceleration in the X direction.
7. tBodyAccStdY– the standard deviation of the body acceleration in the Y direction.
8. tBodyAccStdZ– the standard deviation of the body acceleration in the Z direction.
9. tGravityAccMeanX– the gravity acceleration mean in the X direction.
10. tGravityAccMeanY– the gravity acceleration mean in the Y direction.
11. tGravityAccMeanZ– the gravity acceleration mean in the Z direction.
12. tGravityAccStdX– the gravity acceleration standard deviation in the X direction.

13. tGravityAccStdY– the gravity acceleration standard deviation in the Y direction.
14. tGravityAccStdZ– the gravity acceleration standard deviation in the Z direction.
15. tBodyAccJerkMeanX– the mean of the body acceleration Jerk in the X direction.
16. tBodyAccJerkMeanY– the mean of the body acceleration Jerk in the Y direction.
17. tBodyAccJerkMeanZ– the mean of the body acceleration Jerk in the Z direction.
18. tBodyAccJerkStdX– the standard deviation of the body acceleration Jerk in the X direction.
19. tBodyAccJerkStdY– the standard deviation of the body acceleration Jerk in the Y direction.
20. tBodyAccJerkStdZ– the standard deviation of the body acceleration Jerk in the Z direction.
21. tBodyGyroMeanX– the mean of the body gyro motion in the X direction.
22. tBodyGyroMeanY– the mean of the body gyro motion in the Y direction.
23. tBodyGyroMeanZ– the mean of the body gyro motion in the Z direction.
24. tBodyGyroStdX– the Standard deviation of the body gyro motion in the X direction.
25. tBodyGyroStdY– the Standard deviation of the body gyro motion in the Y direction.
26. tBodyGyroStdZ– the Standard deviation of the body gyro motion in the Z direction.
27. tBodyGyroJerkMeanX– the mean of the body gyro jerk motion in the X direction.
28. tBodyGyroJerkMeanY– the mean of the body gyro jerk motion in the Y direction.
29. tBodyGyroJerkMeanZ– the mean of the body gyro jerk motion in the Z direction.
30. tBodyGyroJerkStdX– the Standard deviation of the body gyro motion in the Z direction.
31. tBodyGyroJerkStdY– the Standard deviation of the body gyro motion in the Y direction.
32. tBodyGyroJerkStdZ– the Standard deviation of the body gyro motion in the Z direction.
33. tBodyAccMagMean– the mean of the body acceleration mag.
34. tBodyAccMagStd– the standard deviation of the body acceleration mag.
35. tGravityAccMagMean -the mean of the gravity acceleration mag.
36. tGravityAccMagStd– the standard deviation of the gravity acceleration mag.
37. tBodyAccJerkMagMean- the mean of the body acceleration jerk motion.
38. tBodyAccJerkMagStd- the standard deviation of the body acceleration jerk motion.
39. tBodyGyroMagMean – the mean of the body gyro mag.
40. tBodyGyroMagStd -the standard deviation the body gyro mag.
41. tBodyGyroJerkMagMean - the mean of the body gyro jerk mag.
42. tBodyGyroJerkMagStd-the standard deviation the body gyro jerk mag.

Frequency Domain Variables (all units normalized to 1)

43. fBodyAccMeanX - the mean of the body acceleration in the X direction.
44. fBodyAccMeanY - the mean of the body acceleration in the Y direction.
45. fBodyAccMeanZ - the mean of the body acceleration in the Z direction.
46. fBodyAccStdX -the standard deviation of the body acceleration in the X direction.
47. fBodyAccStdY -the standard deviation of the body acceleration in the Y direction.
48. fBodyAccStdZ -the standard deviation of the body acceleration in the Z direction.
49. fBodyAccMeanFreqX – the mean frequency of the body acceleration in X direction.
50. fBodyAccMeanFreqY- the mean frequency of the body acceleration in Y direction.
51. fBodyAccMeanFreqZ- the mean frequency of the body acceleration in Z direction.
52. fBodyAccJerkMeanX -the mean of the body acceleration Jerk in the X direction.

53. fBodyAccJerkMeanY-the mean of the body acceleration Jerk in the X direction.
54. fBodyAccJerkMeanZ-the mean of the body acceleration Jerk in the Z direction.
55. fBodyAccJerkStdX– the mean of the body acceleration Jerk in the X direction.
56. fBodyAccJerkStdY– the mean of the body acceleration Jerk in the Y direction.
57. fBodyAccJerkStdZ– the mean of the body acceleration Jerk in the Z direction.
58. fBodyAccJerkMeanFreqX- the mean frequency of the body acceleration Jerk in the X direction.
59. fBodyAccJerkMeanFreqY -the mean frequency of the body acceleration Jerk in the X direction.
60. fBodyAccJerkMeanFreqZ -the mean frequency of the body acceleration Jerk in the X direction.
61. fBodyGyroMeanX -the mean of the body gyro in the X direction.
62. fBodyGyroMeanY-the mean of the body gyro in the Y direction.
63. fBodyGyroMeanZ-the mean of the body gyro in the Z direction.
64. fBodyGyroStdX-the standard deviation of the body gyro in the X direction.
65. fBodyGyroStdY- the standard deviation of the body gyro in the Y direction.
66. fBodyGyroStdZ -the standard deviation of the body gyro in the Z direction.
67. fBodyGyroMeanFreqX-the mean of the body gyro frequency in the X direction.
68. fBodyGyroMeanFreqY-the mean of the body gyro frequency in the Y direction.
69. fBodyGyroMeanFreqZ-the mean of the body gyro frequency in the Z direction.
70. fBodyAccMagMean– the mean of the body acceleration mag.
71. fBodyAccMagStd– the standard deviation of the body acceleration mag.
72. fBodyAccMagMeanFreq- body acceleration mag mean frequency.
73. fBodyBodyAccJerkMagMean- the body body acceleration jerk mag mean.
74. fBodyBodyAccJerkMagStd- the body body acceleration jerk mag standard deviation.
75. fBodyBodyAccJerkMagMeanFreq- the body body acceleration gyro man mean.
76. fBodyBodyGyroMagMean – the body body gyro mag standard deviation.
77. fBodyBodyGyroMagStd – the body body gyro mag standard deviation.
78. fBodyBodyGyroMagMeanFreq – the body body gyro mag mean frequency.
79. fBodyBodyGyroJerkMagMean – the body body gyro jerk mag mean.
80. fBodyBodyGyroJerkMagStd – the body body gyro jerk mag standard deviation.
81. fBodyBodyGyroJerkMagMeanFreq- the body body gyro jerk mag mean frequency.
82. AngletBodyAccMeanGravity- the angle body acceleration mean gravity.
83. AngletBodyAccJerkMeanGravityMean- the angle body acceleration jerk mean gravity.
84. AngletBodyGyroMeanGravityMean- the angle body gyro mean gravity.
85. AngletBodyGyroJerkMeanGravityMean- the angle body jerk mean gravity.
86. AngleXGravityMean the angle X gravity mean.
87. AngleYGravityMean – the angle Y gravity mean.
88. AngleZGravityMean the angle Z gravity mean.