README

Dave Yachabach

8/14/2020

Files in this Project

This is the final project for Course 3 of Coursera's Data Science curriculum. Besides the normal Git Repository files, this repository contains the files:

- run_analysis.r - script defined by assignment
- tidy_dataset.csv - resulting dataset written by run_analysis
- codebook.pdf - documents variables in tidy_dataset
- codebook.rmd - constructs codebook.pdf from tidy_dataset

run_analysis Script

The script run_analysis.r does the following:

- accesses data from a wearable accelerometer device
- filters column variables to those containing "mean" or "std"
- averages all selected variables
- groups and orders the data by Subject then Activity
- writes a tidy dataset to tidy_dataset.csv

Data Source

The original data is in the file:

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

run_analysis Script Logic

run_analysis checks the working directory for a directory called:

```
./UCI HAR Dataset
```

If this directory is not in the working directory, run_analysis will download and unzip this file. The process can take up to 5 minutes.

Selecting Appropriate Variables

The assignment asks for an analysis of all MEAN and STD variables.

We'll use text manipulation on "features.txt" (provided in the .zip file) to find all the proper variables. run_analysis will use this file to create a data table called "filteredvarnames"

- filteredvarnames - data.table - only names with "mean" or "std"
  - V1 - variable index
  - V2 - variable names

Building Data Files of Only Required Variables

The provided data files include test and train data in two different files - "X_test.txt" and "X_train.txt". run_analysis will create data tables of only the required fields (using the data table filteredvarnames) and change column names appropriately.

- X_test - data.table of test data from "x_test.txt"

- X_train - data.table of training data from "x_train.txt"

Decoding Test and Train Activity Data

The activity files "y_test.txt" and "y_train.txt" are read into their own data tables:

- y_test - data table of "y_test.txt"

- y_train - data table of "y_train.txt"

These are left joined with a data table from "activity_labels.txt" to make the output more readable.

- activity_labels - data.table of activity labels and thier index
  - V1 - index
  - V2 - activity label (e.g. WALKING, STANDING, etc.)

- y_test_act - joined y_test with activity labels

- y_train_act - joined y_train with activity labels

Adding Subjects and Data to the Activities

The "y_" files are single-column files of data. They are the exact length as the "X_" files and the "subject_" files. I assumed, since there is no index available, that the resulting data.table(s) can be appended.

"subject_train.txt" and "subject_test.txt" are appended with the appropriate activity (y_) table and then with the respective data (X_) table.

- s_train - subject training data.table

- s_test - subject test data.table

- sy_test - appended activity and subject test files

- sy_train - appended activity and subject train files

- full_test_dt - X_test appended with sy_test

- full_train-dt - X_train appended with sy_train

Combining Test and Train Data

We will rbind these two tables for one large tidy data table. Then we can order, group, and take the mean of each col.

- rslt - Tidy data table:
  - of mean mean readings
  - of mean std readings
  - grouped by Subject and Activity
  - Ordered by Subject and Activity

Script Dependencies

- data.table

- dplyr