codebook.md

th3scr3am (but you can call me Blair)

15/04/2020

Course Project

Getting and Cleaning Data

Data Science Specialization

Johns Hopkins Univeristy

About the project The project takes data from 7 files and matches & merges them into one tidy dataset. These datasets have a combined 561 variables. We are interested in all variables that contain "mean" or "std" (standard deviation) in the descriptive variable name. We then create a second tidy data set which averages each of these variables, grouping these averages by six different activity types. Please refer to the README.txt file for more information on the contents of this data.

The variables activity - one of six activities used for observation: walking, walkingupstairs, walkingdownstairs, sitting, standing, laying. I used 86 additional variables in all from xtrain and xtest data sets. Each of the variables measured mean or standard deviation ("std"). For full technical explanation on these variables see: dataset/features.txt & dataset/feature_info.txt Below is a list of variables

```
## # A tibble: 87 x 1
##    x
##    <fct>
##  1 activity
##  2 tbodyacc-mean-x
##  3 tbodyacc-mean-y
##  4 tbodyacc-mean-z
##  5 tbodyacc-std-x
##  6 tbodyacc-std-y
##  7 tbodyacc-std-z
##  8 tgravityacc-mean-x
##  9 tgravityacc-mean-y
## 10 tgravityacc-mean-z
## # ... with 77 more rows
```

Step 1 The first step was to open the test files (call them ytest and xtest). y_test.txt contained 1 column with 2947 rows X_test.txt contained 561 column with 2947 rows activity_labels.txt contained 2 columns with 6 rows features.txt contained 1 column with 561 rows I checked table(ytest) to ensure each value in ytest did indeed correlate to one of the values (1-6) in the first column of activity_labels. The result can be seen here:

```
## ytest
##   1   2   3   4   5   6
## 496 471 420 491 532 537
```

I repeated these steps with y_train.txt and X_ train.txt, which had dimensions of 7352 x 1 and 7352 x 561, respectively.I checked table(ytrain) to ensure each value in ytrain did indeed correlate to one of the values (1-6) in the first column of activity_labels. The result can be seen here:

```
## ytrain
##    1    2    3    4    5    6
## 1226 1073  986 1286 1374 1407
```

Step 2: I complete the following in R script run_analysis.R. See comments in that R script for more information I first correlated activity_labels to ytest I observed that the activity labels used variables V1 and V2. I was interested in V2 (which were the names)since the V1 variables correlated to V2 by way of indices so V1 was redundant. I then converted the ytest numbers into these activity labels which are clearly descriptive. I took xtest and renamed the variables according to best practices. The variable names are based on the features_info.txt file which shows their descriptions (and thus the variables are descriptive). See the README.txt file where I've copied and pasted this info.

I added the converted ytest data as a new variable called activity and then assigned a new data set to xtest for variables including "mean" or "standard deviation (std)" as well as the activity labels.

Step 3: Also in the run_analysis.R script, I create a 2nd data set. See R script comments for more information I grouped the first dataset by activity to create a 2nd dataset. I then summarized each of the variables in this new data set in order to get the average of each variable.

Step 4: Auditing the new data sets I wanted to ensure that data was not lost or 'misplaced' when I combined them into one set. So I ran the following tests to ensure that. I created the R script "run_analysis_audit.R" to do achieve this. You can view the results by running "extra/run_analyis_audit.R"