README

James

5/10/2020

R Script Description

A note: this markdown file is more readable if you access the pdf file also uploaded under the same name.

There are three main steps within my code for this project:

1. Reading the data
2. Merging and filtering the data
3. Summarize the data

Reading the data

First, all of the relavent text files are read into R. This includes both the test and the training data as well as the corresponding data on activities and subjects for both test and training. The features list is also loaded into a dataframe called col_names while a mapping table between activity and activity code is also loaded in for use later in the code.

Merging and filtering the data

Next, the larger datasets are combined into one master dataframe. First, the subject and activities data is merged with the training and test data sets using cbind. Then, the full test and training datasets are combined into one large dataframe using rbind. THe column names are updated for this larger dataset using manually input names for the first two columns and using the provided feature list for the remaining columns.

Once the data has been combined, as above, we filter out the columns that are not a mean or standard deviation measurement. The assignment is ambiguous on whether the features with "meanFreq()" in their names should be included or excluded. I chose to exclude them, but it would be reasonable to do it either way.

There is a final merge step to add the activity definitions to the dataframe. Since we defined the column name on import to be "activityCode" and then later defined a column with the same name in the larger dataframe, the merge function will automatically match based on these columns.

Summarize the data

The first step is to tell R how to group the data - this was done by subject and activity description, per the directions of the assignment. Next, I found the summarise_at function to be helpful to calculate the means of each feature in the filtered dataset, without having to spell out every column name. The column numbering in this case is a bit unintuitive so, let me explain it:

The dimensions of the groupedData dataframe are 10,299 rows by 69 columns. The activityCode and activityDesc columns are are the second and last (69th) columns, respectively - these are the columns we have grouped the data by. The first column is the activityCode column which is the numerical code corresponding to the activity description, and thus is duplicate information that we will want to drop when we summarize the data. Thus, we want to calculate the mean for only columns 3:68 as these are the columns containing the feature data. However, when we use the summarise_at function, the first and last columns are invalid inputs (because they define the groups) and thus there are only 67 columns that we can specify. This is why 2:67 is specified, and not 3:68.

Lastly, to make the column names more descriptive, we have specified in the final table that all of the feature columns are calculated means by prefixing the column names with "mean of".

Code book

The following table shows a summary of the columns present in the final dataframe, summaryData:

Table 1: Column Descriptions for summaryData

Column Number	Column Name	Description	Possible Values
1	subjectID	The ID number related to the subject studied. There were thirty different subjects from which the data was captured.	Integers 1-30
2	activityDesc	The activity corresponding to the data in the remaining columns.	WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING
3-68	Various	The average of the means or standard deviations of the many measurements taken in this study.	[-1,1]

The dimensions of summaryData are 180 rows and 68 variables. It has 180 rows because there are 30 subjects, each of which performed 6 activities - 30*6 = 180.