Codebook.md

Iris

9/21/2020

Run_analysis.R Codebook

Dataset used is from: *Human Activity Recognition Using Smartphones.*

Dataset Zip File: *UCI HAR Dataset*

The run_analysis code is organized into 6 parts.

1. Reading the files in
2. Combining files into one dataset
3. Extracting mean and standard deviation entries
4. Naming data set activities
5. Renaming data set variables
6. Creating a new tidy dataset, AvgData

1. Reading the files into R

Each file in the UCI HAR Dataset was assigned a variable and column names *features.txt*: lists each of the measurements taken. The measurements are recorded in the order of this list

- features <- features.txt

- columns:
  - n : the column that each measurement corresponds to in the data

  - features : the name of the measurements

activity_labels.txt: lists each of the activities the participants do

- activities <- activity_labels.txt

- columns:
  - activity_label : the numeric code for each of the activities in the data table

  - activity : the name each code corresponds to

Data in the Test file

subject_test.txt: lists which participant corresponds to each row in the data files

- subject_test <- subject_test.txt.
- columns:
  - participant : lists which participant corresponds to the row

y_test.txt: lists which activity the participant did in each row

- y_test <- y_test.txt

- columns:
  - ActivityLabel: The numeric codes for each activity observed in the corresponding column

x_test.txt: The table of observations. Each row corresponds to a participant in x_subject and an activity in y_test. Each column corresponds to an observation in features

- y_test <- y_test.txt

- columns:
  - features$functions is the complete list of observations corresponding to each column of y_test

Data in the train file

subject_train.txt: lists which participant corresponds to each row in the data files

- subject_train <- subject_train.txt.
- columns:
  - participant : lists which participant corresponds to the row

y_train.txt: lists which activity the participant did in each row

- y_train <- y_train.txt

- columns:
  - ActivityLabel: The numeric codes for each activity observed in the corresponding column

x_train.txt: The table of observations. Each row corresponds to a participant in x_subject and an activity in y_train. Each column corresponds to an observation in features

- y_train <- y_train.txt

- columns:
  - features$functions is the complete list of observations corresponding to each column of y_train

2. Combining all the files into a single dataset

The files in this dataset are related as such:

- The "x_[test|train]" files are the data files that contain the observations collected for each participant
- The "y_[test|train]" files contain the activities corresponding to each row of the "x_[test|train]" files
- The "subject_[test|train]" files contain the participant that corresponds to each row of the "x_[test|train]" and "y_[test|train]" files
- The "features" files contains the list of observations that corresponds to the columns of the "x_[test|train]" files.

These files are the files that will be combined into one dataset. The "activity_label" files translate the numeric code of the activity as recorded in the "y_[test|train]" files to the activity's descriptive name.

To combine these files, each of the [test|train] pairs (x_[test] with x_[train], etc) were bound by row using rbind() to form 3 dfs: X (with the x_data), Y(with the y_ data) and participants(with the subject_data).

Then thee participants, Y, and X files, were bound in this order, using cbind() to make a single table called merge.

3. Extracting only the measurements on mean and standard deviation.

For this step, I used grep() on the names of the columns to identify the columns that contain "mean" or "std"' in it – i.e. the measurements on mean and standard deviation. I set value = TRUE to get the character names of the columns instead of the indices to check my work, but this is not necessary. This was saved as selected_names.

Then, I indexed participants, activities and selected_names from merge, in that order, to get a data table with the participant and activity columns as well as all the observations on mean or standard deviation, named selected.

4.Naming data set activities

Using the data file "activity_label.txt" saved as activities, I translated the numeric codes for the activities from selected to the descriptive names by indexing

5. Renaming the dataset with descriptive variable names

gsub() was used on the names of the data set to find and replace undescriptive parts of the name with more descriptive counterparts, as well as removing unnecesary symbols and periods.

6. Create a tidy dataset

aggregate() was used to apply the mean() function to each observation column of the data table, grouped by participant and activity.

This was saved into a file called AvgData.txt