Getting and Cleaning Data Course Project.

Codebook

Max Serna

27/9/2020

Contents

CodeBook formats

If you´re reading the .rmd file, visit {https://maxserna.github.io/DataCourseProject/} to have a full view of this R Markdown file. Alternatively, you can read the PDF file `CodeBookPDF` located in this repo.

Overall description

This Codebook describes the data sets used for this assignment, as well as all transformations made to them and to their variables, as indicated by the Course instructions.

The process

In the following lines, there will be a brief explanation of the code performed in the `run_analysis.R` script.

1. Loading the libraries.
There are only two other R packages I used besides the base ones.

```
library(dplyr)
library(mgsub)
```

2. Downloading the data.
Instead of manually downloading the zip file containing the data, we can use the `tempfile()` function along with `download.file` and `zip` to internally get all the data from the next url. `unlink(temp)` closes the connections to the file after all necessary information has been extracted. For convenience, I won´t show all that code in here. Reader can reference to the `run_analysis.R` script to see the full example. Next lines show a brief description of the files needed.

The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ:
* `features` = `features.txt`. 2 cols and 561 rows

List of activities performed when the corresponding measurements were taken and its codes (labels):
* `activities` = `activity_labels.txt`. 2 cols and 6 rows

Contains test data of 9/30 volunteer test subjects being observed:
* `subject_test` = `test/subject_test.txt`. 1 cols and 2947 rows

Contains recorded features test data:
* `x_test` = `test/X_test.txt`. 561 cols and 2947 rows

Contains test data of activities'code labels:
* `y_test` = `test/y_test.txt`. 1 cols and 2947 rows

Contains train data of 21/30 volunteer subjects being observed:
* `subject_train` = `train/subject_train.txt`. 1 cols and 7352 rows

Contains recorded features train data:
* `x_train` = `train/X_train.txt`. 561 cols and 7352 rows

Contains train data of activities'code labels:
* `y_train` = `train/y_train.txt`. 1 cols and 7352 rows

3. Merging our files.

As pointed by the Course instructions: *1. Merges the training and the test sets to create one data set.* We end up with what I call a "raw" (`raw`) dataset, to which all modifications will be made.

```
test <- cbind(subject_test,
              y_test,
              x_test)
training <- cbind(subject_train,
                  y_train,
                  x_train)
raw <- rbind(training,
             test)
```

4. Keeping the mean and the standard deviation.

Just as instruction 2 conducts: *2. Extracts only the measurements on the mean and standard deviation for each measurement.* Hence we run the following:

```
step2 <- raw %>%
  select(subject,
         code,
         contains(c('mean',
                    'std')
         )
  )
```

5. Recoding activities column

Considering our `step2`data frame generated before, we can see that the following values are listed under the `code` column: 5, 4, 6, 1, 3, 2. Let´s also recall that this column contains the values from our `activities` data set but they´re coded rather than showing it´s more comprehensible values. Here's an example . Hence, instruction 3 from the Course asks us to: *3. Uses descriptive activity names to name the activities in the data set* We will assign `step2` to `step3` to continue with the instructions-structure. Note that all the script follows this fashion.

```
step3 <- step2
step3$code <- activities[step3$code, 2]
unique(step3$code)
```

```
## [1] "STANDING"           "SITTING"            "LAYING"
## [4] "WALKING"            "WALKING_DOWNSTAIRS" "WALKING_UPSTAIRS"
```

6. Renaming our variables

Point 4 in the Course instructions say: *4. Appropriately labels the data set with descriptive variable names.* As of right now, our column names look like these: tBodyAcc.mean...X, tBodyAcc.mean...Y, tBodyAcc.mean...Z. We can tell there are some incomprehensible words as tBodyAcc or Gyro (which in fact stand for Time, Body, Accelerometer or Gyroscope). We aim at correcting those by substituting them with clearer names. The following code will do the job.

```
tidy <- step3; rm(step3)
colnames(tidy)[1:2] <- c('Subject', 'Activities')
colnames(tidy) <- colnames(tidy) %>%
  mgsub(pattern = c('Acc', 'Gyro', 'BodyBody',
```

```
                    'Mag', '^t', '^f',
                    'tBody', 'mean', 'std',
                    'freq', 'angle', 'gravity'),
        replacement = c('Accelerometer', 'Gyroscope', 'Body',
                        'Magnitude', 'Time', 'Frequency',
                        'TimeBody', 'Mean', 'STD',
                        'Frequency', 'Angle', 'Gravity'),
        ignore.case = TRUE
  )
names(tidy)[3:15]
```

```
##  [1] "TimeBodyAccelerometer.Mean...X"     "TimeBodyAccelerometer.Mean...Y"
##  [3] "TimeBodyAccelerometer.Mean...Z"     "TimeGravityAccelerometer.Mean...X"
##  [5] "TimeGravityAccelerometer.Mean...Y"  "TimeGravityAccelerometer.Mean...Z"
##  [7] "TimeBodyAccelerometerJerk.Mean...X" "TimeBodyAccelerometerJerk.Mean...Y"
##  [9] "TimeBodyAccelerometerJerk.Mean...Z" "TimeBodyGyroscope.Mean...X"
## [11] "TimeBodyGyroscope.Mean...Y"         "TimeBodyGyroscope.Mean...Z"
## [13] "TimeBodyGyroscopeJerk.Mean...X"
```

```
dim(tidy)
```

```
## [1] 10299    88
```

7. Averaging averages and standard deviations (whatever that means)

For this final step, we will proceed to summarise our data by each subject in the experiment, as well as by activity, to obtain a nice and clean data frame that shows us some averages. This will be conducted by the next chunk of code:

```
tidy_means <- tidy %>%
  group_by(Subject,
           Activities) %>%
  summarise_all('mean')
write.table(tidy_means,
            file = 'C:/Users/max_s/Desktop/EconomíaUDG/Varios/Coursera/G.and.C.Data/Week4/tidy_means.tx
            row.names = FALSE)
tidy_means
```

```
## # A tibble: 180 x 88
## # Groups:   Subject [30]
##    Subject Activities TimeBodyAcceler~ TimeBodyAcceler~ TimeBodyAcceler~
##      <int> <chr>                <dbl>            <dbl>            <dbl>
##  1       1 LAYING               0.222         -0.0405          -0.113
##  2       1 SITTING              0.261         -0.00131         -0.105
##  3       1 STANDING             0.279         -0.0161          -0.111
##  4       1 WALKING              0.277         -0.0174          -0.111
##  5       1 WALKING_D~           0.289         -0.00992         -0.108
##  6       1 WALKING_U~           0.255         -0.0240          -0.0973
##  7       2 LAYING               0.281         -0.0182          -0.107
##  8       2 SITTING              0.277         -0.0157          -0.109
##  9       2 STANDING             0.278         -0.0184          -0.106
## 10       2 WALKING              0.276         -0.0186          -0.106
```

```
## # ... with 170 more rows, and 83 more variables:
## #   TimeGravityAccelerometer.Mean...X <dbl>,
## #   TimeGravityAccelerometer.Mean...Y <dbl>,
## #   TimeGravityAccelerometer.Mean...Z <dbl>,
## #   TimeBodyAccelerometerJerk.Mean...X <dbl>,
## #   TimeBodyAccelerometerJerk.Mean...Y <dbl>,
## #   TimeBodyAccelerometerJerk.Mean...Z <dbl>, TimeBodyGyroscope.Mean...X <dbl>,
## #   TimeBodyGyroscope.Mean...Y <dbl>, TimeBodyGyroscope.Mean...Z <dbl>,
## #   TimeBodyGyroscopeJerk.Mean...X <dbl>, TimeBodyGyroscopeJerk.Mean...Y <dbl>,
## #   TimeBodyGyroscopeJerk.Mean...Z <dbl>,
## #   TimeBodyAccelerometerMagnitude.Mean.. <dbl>,
## #   TimeGravityAccelerometerMagnitude.Mean.. <dbl>,
## #   TimeBodyAccelerometerJerkMagnitude.Mean.. <dbl>,
## #   TimeBodyGyroscopeMagnitude.Mean.. <dbl>,
## #   TimeBodyGyroscopeJerkMagnitude.Mean.. <dbl>,
## #   FrequencyBodyAccelerometer.Mean...X <dbl>,
## #   FrequencyBodyAccelerometer.Mean...Y <dbl>,
## #   FrequencyBodyAccelerometer.Mean...Z <dbl>,
## #   FrequencyBodyAccelerometer.MeanFrequency...X <dbl>,
## #   FrequencyBodyAccelerometer.MeanFrequency...Y <dbl>,
## #   FrequencyBodyAccelerometer.MeanFrequency...Z <dbl>,
## #   FrequencyBodyAccelerometerJerk.Mean...X <dbl>,
## #   FrequencyBodyAccelerometerJerk.Mean...Y <dbl>,
## #   FrequencyBodyAccelerometerJerk.Mean...Z <dbl>,
## #   FrequencyBodyAccelerometerJerk.MeanFrequency...X <dbl>,
## #   FrequencyBodyAccelerometerJerk.MeanFrequency...Y <dbl>,
## #   FrequencyBodyAccelerometerJerk.MeanFrequency...Z <dbl>,
## #   FrequencyBodyGyroscope.Mean...X <dbl>,
## #   FrequencyBodyGyroscope.Mean...Y <dbl>,
## #   FrequencyBodyGyroscope.Mean...Z <dbl>,
## #   FrequencyBodyGyroscope.MeanFrequency...X <dbl>,
## #   FrequencyBodyGyroscope.MeanFrequency...Y <dbl>,
## #   FrequencyBodyGyroscope.MeanFrequency...Z <dbl>,
## #   FrequencyBodyAccelerometerMagnitude.Mean.. <dbl>,
## #   FrequencyBodyAccelerometerMagnitude.MeanFrequency.. <dbl>,
## #   FrequencyBodyAccelerometerJerkMagnitude.Mean.. <dbl>,
## #   FrequencyBodyAccelerometerJerkMagnitude.MeanFrequency.. <dbl>,
## #   FrequencyBodyGyroscopeMagnitude.Mean.. <dbl>,
## #   FrequencyBodyGyroscopeMagnitude.MeanFrequency.. <dbl>,
## #   FrequencyBodyGyroscopeJerkMagnitude.Mean.. <dbl>,
## #   FrequencyBodyGyroscopeJerkMagnitude.MeanFrequency.. <dbl>,
## #   Angle.TimeBodyAccelerometerMean.Gravity. <dbl>,
## #   Angle.TimeBodyAccelerometerJerkMean..GravityMean. <dbl>,
## #   Angle.TimeBodyGyroscopeMean.GravityMean. <dbl>,
## #   Angle.TimeBodyGyroscopeJerkMean.GravityMean. <dbl>,
## #   Angle.X.GravityMean. <dbl>, Angle.Y.GravityMean. <dbl>,
## #   Angle.Z.GravityMean. <dbl>, TimeBodyAccelerometer.STD...X <dbl>,
## #   TimeBodyAccelerometer.STD...Y <dbl>, TimeBodyAccelerometer.STD...Z <dbl>,
## #   TimeGravityAccelerometer.STD...X <dbl>,
## #   TimeGravityAccelerometer.STD...Y <dbl>,
## #   TimeGravityAccelerometer.STD...Z <dbl>,
## #   TimeBodyAccelerometerJerk.STD...X <dbl>,
## #   TimeBodyAccelerometerJerk.STD...Y <dbl>,
## #   TimeBodyAccelerometerJerk.STD...Z <dbl>, TimeBodyGyroscope.STD...X <dbl>,
```

```
## #   TimeBodyGyroscope.STD...Y <dbl>, TimeBodyGyroscope.STD...Z <dbl>,
## #   TimeBodyGyroscopeJerk.STD...X <dbl>, TimeBodyGyroscopeJerk.STD...Y <dbl>,
## #   TimeBodyGyroscopeJerk.STD...Z <dbl>,
## #   TimeBodyAccelerometerMagnitude.STD.. <dbl>,
## #   TimeGravityAccelerometerMagnitude.STD.. <dbl>,
## #   TimeBodyAccelerometerJerkMagnitude.STD.. <dbl>,
## #   TimeBodyGyroscopeMagnitude.STD.. <dbl>,
## #   TimeBodyGyroscopeJerkMagnitude.STD.. <dbl>,
## #   FrequencyBodyAccelerometer.STD...X <dbl>,
## #   FrequencyBodyAccelerometer.STD...Y <dbl>,
## #   FrequencyBodyAccelerometer.STD...Z <dbl>,
## #   FrequencyBodyAccelerometerJerk.STD...X <dbl>,
## #   FrequencyBodyAccelerometerJerk.STD...Y <dbl>,
## #   FrequencyBodyAccelerometerJerk.STD...Z <dbl>,
## #   FrequencyBodyGyroscope.STD...X <dbl>, FrequencyBodyGyroscope.STD...Y <dbl>,
## #   FrequencyBodyGyroscope.STD...Z <dbl>,
## #   FrequencyBodyAccelerometerMagnitude.STD.. <dbl>,
## #   FrequencyBodyAccelerometerJerkMagnitude.STD.. <dbl>,
## #   FrequencyBodyGyroscopeMagnitude.STD.. <dbl>,
## #   FrequencyBodyGyroscopeJerkMagnitude.STD.. <dbl>
```

BOOM! This final data called `tidy_means` is to be exported as a text file named `tidy_means.txt`, that will be submitted to Coursera.