Getting and Cleaning Data Project by George Drakulic

download data from

http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+

unzip it at your working directory

it will create directoy called “UCI HAR Dataset”

in that directory there are two other directories: “test” and “train”,

as well as labels of variour columns of the data in txt files

create a data frame “subject” by reading subject_test.txt, which has rows of a subject's number,

and name the column “subject”

```
subject <- read.table("UCI HAR Dataset/test/subject_test.txt")
colnames(subject) <- c("subject")
```

create a data frame “activity” by reading y_test.txt, which has rows of an activity's number,

and name the column “activity”

```
activity <- read.table("UCI HAR Dataset/test/y_test.txt")
colnames(activity) <- c("activity")
```

change the numbers in activity data frame into textual activities listed in “activity_labels.txt”

```
activity <- factor(activity$activity, levels = c(1, 2, 3, 4, 5, 6), labels = c("walking",
    "walking_upstairs", "walking_downstairs", "sitting", "standing", "laying"))
```

merge two data frames into one (subject and activity into dat1)

```
dat1 <- cbind(subject, activity)
```

read X_test data with 561 columns

```
X_test <- read.table("UCI HAR Dataset/test/X_test.txt")
```

get the names of the features (which will be column names of X_test data)

```
features <- read.table("UCI HAR Dataset/features.txt")
```

rows should be columns, so we will transpose rows and columns, and get just the names of the features

```
feature_name <- t(features)[2, ]
```

rename the columns of X_test data with the column names of the features from features.txt

```
colnames(X_test) <- feature_name
```

get the index of columns that have “mean()” in them, and also “meanFreq()

```
g_mean <- grep("mean()", feature_name)
```

get the index of columns that have "std()” in them

```
g_std <- grep("std()", feature_name)
```

combine those indices into one

```
g_all <- append(g_mean, g_std)
```

sort the indices

these are the indices we need in our data

```
sorted_indices <- sort(g_all)
```

the data set with columns of needed data of mean, and std computations

```
X_test_mean_and_std <- X_test[, sorted_indices]
```

merge subject + activity (dat1), and X_test_mean_std

```
X_test_merged <- cbind(dat1, X_test_mean_and_std)
```

just to check the final result: nrow(X_test_merged) gives me 2947,

ncol(X_test_merged) gives me 81, so the test merged data frame is 2947 x 81

Do the same for the train data

```
subjectTrain <- read.table("UCI HAR Dataset/train/subject_train.txt")
colnames(subjectTrain) <- c("subject")

activityTrain <- read.table("UCI HAR Dataset/train/y_train.txt")
colnames(activityTrain) <- c("activity")

activityTrain <- factor(activityTrain$activity, levels = c(1, 2, 3, 4, 5, 6),
    labels = c("walking", "walking_upstairs", "walking_downstairs", "sitting",
        "standing", "laying"))
dat2 <- cbind(subjectTrain, activityTrain)
colnames(dat2) <- c("subject", "activity")

X_train <- read.table("UCI HAR Dataset/train/X_train.txt")
colnames(X_train) <- feature_name
X_train_mean_and_std <- X_train[, sorted_indices]

X_train_merged <- cbind(dat2, X_train_mean_and_std)
```

just to check the final result: nrow(X_train_merged) gives me 7352,

ncol(X_train_merged) gives me 81, so the train merged data frame is 7352 x 81

merging the test and the training sets should give me (2947 + 7352) x 81 data frame

i.e. 10299 x 81

```
X_final <- rbind(X_test_merged, X_train_merged)
```

nrow(X_final) gives me 10299, which is as expected, ncol(X_final) is 81 as well

empty data frame for tidy data

```
X_tidy <- data.frame()
```

the data is split by subject, activity, and other variables, so it computes

the mean of each variable for each subject and his/her activity,

e.g. subject 1, activity 1, all means of variables; subject 2, activity 1, means,

up to last subject, activity 1

then, it starts with subject 1, activity 2 …, last s, act 2 and so on

```
data_split <- split(X_final[, 3:81], list(X_final$subject, X_final$activity))
X_tidy <- sapply(data_split, colMeans)
```

nrow(X_tidy) = 79, ncol(X_tidy) = 180

output with row.names=F so the data is properly aligned in MS Excel, but the row names

are, of course, missing

```
write.table(X_tidy, file = "./X_tidy.txt", sep = "\t", row.names = FALSE)
```

output with row.names=T so the data is NOT properly aligned in MS Excel

```
write.table(X_tidy, file = "./X_tidy_with_row_names.txt", sep = "\t")
```

```
head(X_tidy, 2)
```

```
##                   1.walking 2.walking 3.walking 4.walking 5.walking
## tBodyAcc-mean()-X   0.27733   0.27643   0.27557   0.27858   0.27784
## tBodyAcc-mean()-Y  -0.01738  -0.01859  -0.01718  -0.01484  -0.01729
```

```
##                   6.walking 7.walking 8.walking 9.walking 10.walking
## tBodyAcc-mean()-X    0.2837   0.27559   0.27469   0.27850    0.27857
## tBodyAcc-mean()-Y   -0.0169  -0.01865  -0.01866  -0.01809   -0.01702
##                   11.walking 12.walking 13.walking 14.walking 15.walking
## tBodyAcc-mean()-X    0.27182    0.27713    0.27588    0.27196    0.27390
## tBodyAcc-mean()-Y   -0.01665   -0.01556   -0.01857   -0.02178   -0.01708
##                   16.walking 17.walking 18.walking 19.walking 20.walking
## tBodyAcc-mean()-X    0.27602    0.27234     0.2739    0.27393     0.2726
## tBodyAcc-mean()-Y   -0.02043   -0.01849    -0.0178   -0.01918    -0.0212
##                   21.walking 22.walking 23.walking 24.walking 25.walking
## tBodyAcc-mean()-X    0.27918    0.27886    0.27321    0.27698    0.27899
## tBodyAcc-mean()-Y   -0.01816   -0.01672   -0.01836   -0.02254   -0.01865
##                   26.walking 27.walking 28.walking 29.walking 30.walking
## tBodyAcc-mean()-X    0.27926    0.27685    0.28123    0.27200    0.27641
## tBodyAcc-mean()-Y   -0.01543   -0.01665   -0.01568   -0.01629   -0.01759
##                   1.walking_upstairs 2.walking_upstairs 3.walking_upstairs
## tBodyAcc-mean()-X            0.25546            0.24716             0.26082
## tBodyAcc-mean()-Y           -0.02395           -0.02141            -0.03241
##                   4.walking_upstairs 5.walking_upstairs 6.walking_upstairs
## tBodyAcc-mean()-X            0.27088            0.26846             0.26823
## tBodyAcc-mean()-Y           -0.03198           -0.03253            -0.02724
##                   7.walking_upstairs 8.walking_upstairs 9.walking_upstairs
## tBodyAcc-mean()-X            0.24871            0.25888             0.26244
## tBodyAcc-mean()-Y           -0.02756           -0.02824            -0.01951
##                   10.walking_upstairs 11.walking_upstairs
## tBodyAcc-mean()-X             0.26712             0.26378
## tBodyAcc-mean()-Y            -0.01439            -0.03032
##                   12.walking_upstairs 13.walking_upstairs
## tBodyAcc-mean()-X             0.27297             0.25820
## tBodyAcc-mean()-Y            -0.02636            -0.02774
##                   14.walking_upstairs 15.walking_upstairs
## tBodyAcc-mean()-X             0.26242             0.27019
## tBodyAcc-mean()-Y            -0.02044            -0.02875
##                   16.walking_upstairs 17.walking_upstairs
## tBodyAcc-mean()-X             0.25599             0.25260
## tBodyAcc-mean()-Y            -0.01437            -0.02286
##                   18.walking_upstairs 19.walking_upstairs
## tBodyAcc-mean()-X             0.26540              0.2421
## tBodyAcc-mean()-Y            -0.02221             -0.0304
##                   20.walking_upstairs 21.walking_upstairs
## tBodyAcc-mean()-X             0.25210             0.26519
## tBodyAcc-mean()-Y            -0.02823            -0.02372
##                   22.walking_upstairs 23.walking_upstairs
## tBodyAcc-mean()-X             0.24839             0.25000
## tBodyAcc-mean()-Y            -0.02686            -0.03238
##                   24.walking_upstairs 25.walking_upstairs
## tBodyAcc-mean()-X              0.2699             0.27800
## tBodyAcc-mean()-Y             -0.0252            -0.02699
##                   26.walking_upstairs 27.walking_upstairs
## tBodyAcc-mean()-X             0.27269              0.2658
## tBodyAcc-mean()-Y            -0.02816             -0.0201
##                   28.walking_upstairs 29.walking_upstairs
## tBodyAcc-mean()-X             0.26201             0.26542
## tBodyAcc-mean()-Y            -0.02794            -0.02995
##                   30.walking_upstairs 1.walking_downstairs
## tBodyAcc-mean()-X             0.27142             0.289188
## tBodyAcc-mean()-Y            -0.02533            -0.009919
##                   2.walking_downstairs 3.walking_downstairs
```

```
## tBodyAcc-mean()-X              0.27762              0.29242
## tBodyAcc-mean()-Y             -0.02266             -0.01936
##                   4.walking_downstairs 5.walking_downstairs
## tBodyAcc-mean()-X             0.279965             0.293544
## tBodyAcc-mean()-Y            -0.009802            -0.008501
##                   6.walking_downstairs 7.walking_downstairs
## tBodyAcc-mean()-X              0.27705              0.28031
## tBodyAcc-mean()-Y             -0.01954             -0.01663
##                   8.walking_downstairs 9.walking_downstairs
## tBodyAcc-mean()-X              0.28348               0.2959
## tBodyAcc-mean()-Y             -0.02111              -0.0204
##                   10.walking_downstairs 11.walking_downstairs
## tBodyAcc-mean()-X               0.29040               0.29161
## tBodyAcc-mean()-Y              -0.02001              -0.01781
##                   12.walking_downstairs 13.walking_downstairs
## tBodyAcc-mean()-X               0.28152               0.29491
## tBodyAcc-mean()-Y              -0.01808              -0.01437
##                   14.walking_downstairs 15.walking_downstairs
## tBodyAcc-mean()-X               0.29342               0.28020
## tBodyAcc-mean()-Y              -0.02001              -0.00563
##                   16.walking_downstairs 17.walking_downstairs
## tBodyAcc-mean()-X               0.29559               0.29392
## tBodyAcc-mean()-Y              -0.01839              -0.01674
##                   18.walking_downstairs 19.walking_downstairs
## tBodyAcc-mean()-X               0.28844               0.26269
## tBodyAcc-mean()-Y              -0.01687              -0.01459
##                   20.walking_downstairs 21.walking_downstairs
## tBodyAcc-mean()-X              0.296144               0.30146
## tBodyAcc-mean()-Y             -0.009641              -0.01732
##                   22.walking_downstairs 23.walking_downstairs
## tBodyAcc-mean()-X                0.2845               0.28990
## tBodyAcc-mean()-Y               -0.0198              -0.01621
##                   24.walking_downstairs 25.walking_downstairs
## tBodyAcc-mean()-X               0.28863               0.29133
## tBodyAcc-mean()-Y              -0.01457              -0.02102
##                   26.walking_downstairs 27.walking_downstairs
## tBodyAcc-mean()-X               0.27928               0.29754
## tBodyAcc-mean()-Y              -0.01263              -0.01356
##                   28.walking_downstairs 29.walking_downstairs
## tBodyAcc-mean()-X               0.29364               0.29314
## tBodyAcc-mean()-Y              -0.02202              -0.01494
##                   30.walking_downstairs 1.sitting 2.sitting 3.sitting
## tBodyAcc-mean()-X               0.28319  0.261238   0.27709  0.257198
## tBodyAcc-mean()-Y              -0.01744 -0.001308  -0.01569 -0.003503
##                   4.sitting 5.sitting 6.sitting 7.sitting 8.sitting
## tBodyAcc-mean()-X  0.271538  0.273694   0.27678   0.28467  0.267491
## tBodyAcc-mean()-Y -0.007163 -0.009901  -0.01459  -0.01461 -0.006726
##                   9.sitting 10.sitting 11.sitting 12.sitting 13.sitting
## tBodyAcc-mean()-X   0.24833    0.27061    0.27659    0.27501   0.274328
## tBodyAcc-mean()-Y  -0.02702   -0.01504   -0.01492   -0.01579  -0.005877
##                   14.sitting 15.sitting 16.sitting 17.sitting 18.sitting
## tBodyAcc-mean()-X   0.279991    0.27290    0.28077    0.27736    0.27727
## tBodyAcc-mean()-Y  -0.008706   -0.01172   -0.01025   -0.01416   -0.01287
##                   19.sitting 20.sitting 21.sitting 22.sitting 23.sitting
## tBodyAcc-mean()-X    0.27383    0.27805     0.2775    0.27358    0.27335
## tBodyAcc-mean()-Y   -0.01674   -0.01472    -0.0144   -0.01235   -0.01339
##                   24.sitting 25.sitting 26.sitting 27.sitting 28.sitting
## tBodyAcc-mean()-X    0.27348    0.27854   0.258244    0.27394    0.27698
```

```
## tBodyAcc-mean()-Y   -0.01313   -0.01477  -0.007134   -0.01553   -0.01854
##                   29.sitting 30.sitting 1.standing 2.standing 3.standing
## tBodyAcc-mean()-X    0.27718   0.268336    0.27892    0.27791    0.28005
## tBodyAcc-mean()-Y   -0.01663  -0.008047   -0.01614   -0.01842   -0.01434
##                   4.standing 5.standing 6.standing 7.standing 8.standing
## tBodyAcc-mean()-X   0.280500   0.282544    0.28035    0.28272    0.27962
## tBodyAcc-mean()-Y  -0.009489  -0.007004   -0.01812   -0.01457   -0.01481
##                   9.standing 10.standing 11.standing 12.standing
## tBodyAcc-mean()-X    0.28231     0.27665      0.2777      0.2774
## tBodyAcc-mean()-Y   -0.02005    -0.01554     -0.0172     -0.0169
##                   13.standing 14.standing 15.standing 16.standing
## tBodyAcc-mean()-X     0.27776     0.28055     0.27892      0.2835
## tBodyAcc-mean()-Y    -0.01679    -0.01521    -0.01835     -0.0166
##                   17.standing 18.standing 19.standing 20.standing
## tBodyAcc-mean()-X     0.27794     0.27846     0.27817     0.27808
## tBodyAcc-mean()-Y    -0.01741    -0.01664    -0.01542    -0.01807
##                   21.standing 22.standing 23.standing 24.standing
## tBodyAcc-mean()-X     0.27695     0.27905     0.27790     0.28035
## tBodyAcc-mean()-Y    -0.01671    -0.01586    -0.01775    -0.01448
##                   25.standing 26.standing 27.standing 28.standing
## tBodyAcc-mean()-X     0.27801     0.28113     0.27957     0.27780
## tBodyAcc-mean()-Y    -0.01636    -0.01666    -0.01659    -0.01726
##                   29.standing 30.standing 1.laying 2.laying 3.laying
## tBodyAcc-mean()-X     0.27797     0.27711  0.22160  0.28137  0.27552
## tBodyAcc-mean()-Y    -0.01726    -0.01702 -0.04051 -0.01816 -0.01896
##                   4.laying 5.laying 6.laying 7.laying 8.laying 9.laying
## tBodyAcc-mean()-X   0.2636   0.2783  0.24866  0.25018  0.26125  0.25920
## tBodyAcc-mean()-Y  -0.0150  -0.0183 -0.01025 -0.02044 -0.02123 -0.02053
##                   10.laying 11.laying 12.laying 13.laying 14.laying
## tBodyAcc-mean()-X   0.28023   0.28059   0.26011   0.27672   0.23328
## tBodyAcc-mean()-Y  -0.02429  -0.01766  -0.01752  -0.02044  -0.01134
##                   15.laying 16.laying 17.laying 18.laying 19.laying
## tBodyAcc-mean()-X   0.28948   0.27423   0.26978   0.27469   0.27265
## tBodyAcc-mean()-Y  -0.01663  -0.01661  -0.01685  -0.01739  -0.01714
##                   20.laying 21.laying 22.laying 23.laying 24.laying
## tBodyAcc-mean()-X   0.23951   0.27133   0.27996   0.27404   0.27285
## tBodyAcc-mean()-Y  -0.01444  -0.01842  -0.01426  -0.02166  -0.01736
##                   25.laying 26.laying 27.laying 28.laying 29.laying
## tBodyAcc-mean()-X   0.25079   0.27165   0.27410   0.27591    0.2873
## tBodyAcc-mean()-Y  -0.01889  -0.01919  -0.01799  -0.01675   -0.0172
##                   30.laying
## tBodyAcc-mean()-X   0.28103
## tBodyAcc-mean()-Y  -0.01945
```