Getting\_data\_course--run\_analysis.R

jennywang

Sun Oct 2 22:19:51 2016

## Code book for course "Getting and Cleaning Data" assignment:  
## Create one R script called run\_analysis.R that does the following:  
## using datasets from "https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip"  
## 1.Merges the training and the test sets to create one data set.  
## 2.Extracts only the measurements on the mean and standard deviation for each measurement.  
## 3.Uses descriptive activity names to name the activities in the data set  
## 4.Appropriately labels the data set with descriptive variable names.  
## 5.From the data set in step 4, creates a second, independent tidy data set with the average of   
## each variable for each activity and each subject.  
  
  
## load packages  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(data.table)

## -------------------------------------------------------------------------

## data.table + dplyr code now lives in dtplyr.  
## Please library(dtplyr)!

## -------------------------------------------------------------------------

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, last

library(tidyr)  
  
## Before started, the first is to download the datasets and unzip.  
setwd("~/Desktop/Coursera")  
fileurl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip"  
download.file(fileurl,destfile="Dataset.zip",method="curl")  
unzip(zipfile="Dataset.zip")  
  
## View the files in the Dataset folder  
list.files("./UCI HAR Dataset")

## [1] "activity\_labels.txt"   
## [2] "averagedata.rmd"   
## [3] "features\_info.txt"   
## [4] "features.txt"   
## [5] "fulldata.rmd"   
## [6] "Getting data course--run\_analysis.R"   
## [7] "Getting\_data\_course--run\_analysis.docx"   
## [8] "Getting\_data\_course--run\_analysis.R"   
## [9] "Getting\_data\_course--run\_analysis.spin.R"   
## [10] "Getting\_data\_course--run\_analysis.spin.Rmd"  
## [11] "MeanData.csv"   
## [12] "README.txt"   
## [13] "test"   
## [14] "TotalData.csv"   
## [15] "train"

list.files("./UCI HAR Dataset/test")

## [1] "Inertial Signals" "subject\_test.txt" "X\_test.txt"   
## [4] "y\_test.txt"

list.files("./UCI HAR Dataset/train")

## [1] "Inertial Signals" "subject\_train.txt" "X\_train.txt"   
## [4] "y\_train.txt"

## Extract the data in the file and store them into corresponding variable  
## If you are using R.studio, you can see the table str. in the right side upper window.  
## 1a. READ.TABLE the subject files and check col. names  
SubjectTest<- read.table("./UCI HAR Dataset/test/subject\_test.txt",   
 header=FALSE)  
SubjectTrain<- read.table("./UCI HAR Dataset/train/subject\_train.txt",   
 header=FALSE)  
names(SubjectTest)

## [1] "V1"

names(SubjectTrain)

## [1] "V1"

## 1b. READ.TABLE the dataset files and check col. names  
DataTest<- read.table ("./UCI HAR Dataset/test/X\_test.txt",   
 header=FALSE)  
any(duplicated(DataTest, MARGIN = 1)) ##FALSE, no duplicated rows

## [1] FALSE

DataTrain<- read.table("./UCI HAR Dataset/train/X\_train.txt",   
 header=FALSE)  
any(duplicated(DataTest, MARGIN = 1)) ##FALSE, no duplicated rows

## [1] FALSE

names(DataTest)

## [1] "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8" "V9" "V10"   
## [11] "V11" "V12" "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20"   
## [21] "V21" "V22" "V23" "V24" "V25" "V26" "V27" "V28" "V29" "V30"   
## [31] "V31" "V32" "V33" "V34" "V35" "V36" "V37" "V38" "V39" "V40"   
## [41] "V41" "V42" "V43" "V44" "V45" "V46" "V47" "V48" "V49" "V50"   
## [51] "V51" "V52" "V53" "V54" "V55" "V56" "V57" "V58" "V59" "V60"   
## [61] "V61" "V62" "V63" "V64" "V65" "V66" "V67" "V68" "V69" "V70"   
## [71] "V71" "V72" "V73" "V74" "V75" "V76" "V77" "V78" "V79" "V80"   
## [81] "V81" "V82" "V83" "V84" "V85" "V86" "V87" "V88" "V89" "V90"   
## [91] "V91" "V92" "V93" "V94" "V95" "V96" "V97" "V98" "V99" "V100"  
## [101] "V101" "V102" "V103" "V104" "V105" "V106" "V107" "V108" "V109" "V110"  
## [111] "V111" "V112" "V113" "V114" "V115" "V116" "V117" "V118" "V119" "V120"  
## [121] "V121" "V122" "V123" "V124" "V125" "V126" "V127" "V128" "V129" "V130"  
## [131] "V131" "V132" "V133" "V134" "V135" "V136" "V137" "V138" "V139" "V140"  
## [141] "V141" "V142" "V143" "V144" "V145" "V146" "V147" "V148" "V149" "V150"  
## [151] "V151" "V152" "V153" "V154" "V155" "V156" "V157" "V158" "V159" "V160"  
## [161] "V161" "V162" "V163" "V164" "V165" "V166" "V167" "V168" "V169" "V170"  
## [171] "V171" "V172" "V173" "V174" "V175" "V176" "V177" "V178" "V179" "V180"  
## [181] "V181" "V182" "V183" "V184" "V185" "V186" "V187" "V188" "V189" "V190"  
## [191] "V191" "V192" "V193" "V194" "V195" "V196" "V197" "V198" "V199" "V200"  
## [201] "V201" "V202" "V203" "V204" "V205" "V206" "V207" "V208" "V209" "V210"  
## [211] "V211" "V212" "V213" "V214" "V215" "V216" "V217" "V218" "V219" "V220"  
## [221] "V221" "V222" "V223" "V224" "V225" "V226" "V227" "V228" "V229" "V230"  
## [231] "V231" "V232" "V233" "V234" "V235" "V236" "V237" "V238" "V239" "V240"  
## [241] "V241" "V242" "V243" "V244" "V245" "V246" "V247" "V248" "V249" "V250"  
## [251] "V251" "V252" "V253" "V254" "V255" "V256" "V257" "V258" "V259" "V260"  
## [261] "V261" "V262" "V263" "V264" "V265" "V266" "V267" "V268" "V269" "V270"  
## [271] "V271" "V272" "V273" "V274" "V275" "V276" "V277" "V278" "V279" "V280"  
## [281] "V281" "V282" "V283" "V284" "V285" "V286" "V287" "V288" "V289" "V290"  
## [291] "V291" "V292" "V293" "V294" "V295" "V296" "V297" "V298" "V299" "V300"  
## [301] "V301" "V302" "V303" "V304" "V305" "V306" "V307" "V308" "V309" "V310"  
## [311] "V311" "V312" "V313" "V314" "V315" "V316" "V317" "V318" "V319" "V320"  
## [321] "V321" "V322" "V323" "V324" "V325" "V326" "V327" "V328" "V329" "V330"  
## [331] "V331" "V332" "V333" "V334" "V335" "V336" "V337" "V338" "V339" "V340"  
## [341] "V341" "V342" "V343" "V344" "V345" "V346" "V347" "V348" "V349" "V350"  
## [351] "V351" "V352" "V353" "V354" "V355" "V356" "V357" "V358" "V359" "V360"  
## [361] "V361" "V362" "V363" "V364" "V365" "V366" "V367" "V368" "V369" "V370"  
## [371] "V371" "V372" "V373" "V374" "V375" "V376" "V377" "V378" "V379" "V380"  
## [381] "V381" "V382" "V383" "V384" "V385" "V386" "V387" "V388" "V389" "V390"  
## [391] "V391" "V392" "V393" "V394" "V395" "V396" "V397" "V398" "V399" "V400"  
## [401] "V401" "V402" "V403" "V404" "V405" "V406" "V407" "V408" "V409" "V410"  
## [411] "V411" "V412" "V413" "V414" "V415" "V416" "V417" "V418" "V419" "V420"  
## [421] "V421" "V422" "V423" "V424" "V425" "V426" "V427" "V428" "V429" "V430"  
## [431] "V431" "V432" "V433" "V434" "V435" "V436" "V437" "V438" "V439" "V440"  
## [441] "V441" "V442" "V443" "V444" "V445" "V446" "V447" "V448" "V449" "V450"  
## [451] "V451" "V452" "V453" "V454" "V455" "V456" "V457" "V458" "V459" "V460"  
## [461] "V461" "V462" "V463" "V464" "V465" "V466" "V467" "V468" "V469" "V470"  
## [471] "V471" "V472" "V473" "V474" "V475" "V476" "V477" "V478" "V479" "V480"  
## [481] "V481" "V482" "V483" "V484" "V485" "V486" "V487" "V488" "V489" "V490"  
## [491] "V491" "V492" "V493" "V494" "V495" "V496" "V497" "V498" "V499" "V500"  
## [501] "V501" "V502" "V503" "V504" "V505" "V506" "V507" "V508" "V509" "V510"  
## [511] "V511" "V512" "V513" "V514" "V515" "V516" "V517" "V518" "V519" "V520"  
## [521] "V521" "V522" "V523" "V524" "V525" "V526" "V527" "V528" "V529" "V530"  
## [531] "V531" "V532" "V533" "V534" "V535" "V536" "V537" "V538" "V539" "V540"  
## [541] "V541" "V542" "V543" "V544" "V545" "V546" "V547" "V548" "V549" "V550"  
## [551] "V551" "V552" "V553" "V554" "V555" "V556" "V557" "V558" "V559" "V560"  
## [561] "V561"

names(DataTrain)

## [1] "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8" "V9" "V10"   
## [11] "V11" "V12" "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20"   
## [21] "V21" "V22" "V23" "V24" "V25" "V26" "V27" "V28" "V29" "V30"   
## [31] "V31" "V32" "V33" "V34" "V35" "V36" "V37" "V38" "V39" "V40"   
## [41] "V41" "V42" "V43" "V44" "V45" "V46" "V47" "V48" "V49" "V50"   
## [51] "V51" "V52" "V53" "V54" "V55" "V56" "V57" "V58" "V59" "V60"   
## [61] "V61" "V62" "V63" "V64" "V65" "V66" "V67" "V68" "V69" "V70"   
## [71] "V71" "V72" "V73" "V74" "V75" "V76" "V77" "V78" "V79" "V80"   
## [81] "V81" "V82" "V83" "V84" "V85" "V86" "V87" "V88" "V89" "V90"   
## [91] "V91" "V92" "V93" "V94" "V95" "V96" "V97" "V98" "V99" "V100"  
## [101] "V101" "V102" "V103" "V104" "V105" "V106" "V107" "V108" "V109" "V110"  
## [111] "V111" "V112" "V113" "V114" "V115" "V116" "V117" "V118" "V119" "V120"  
## [121] "V121" "V122" "V123" "V124" "V125" "V126" "V127" "V128" "V129" "V130"  
## [131] "V131" "V132" "V133" "V134" "V135" "V136" "V137" "V138" "V139" "V140"  
## [141] "V141" "V142" "V143" "V144" "V145" "V146" "V147" "V148" "V149" "V150"  
## [151] "V151" "V152" "V153" "V154" "V155" "V156" "V157" "V158" "V159" "V160"  
## [161] "V161" "V162" "V163" "V164" "V165" "V166" "V167" "V168" "V169" "V170"  
## [171] "V171" "V172" "V173" "V174" "V175" "V176" "V177" "V178" "V179" "V180"  
## [181] "V181" "V182" "V183" "V184" "V185" "V186" "V187" "V188" "V189" "V190"  
## [191] "V191" "V192" "V193" "V194" "V195" "V196" "V197" "V198" "V199" "V200"  
## [201] "V201" "V202" "V203" "V204" "V205" "V206" "V207" "V208" "V209" "V210"  
## [211] "V211" "V212" "V213" "V214" "V215" "V216" "V217" "V218" "V219" "V220"  
## [221] "V221" "V222" "V223" "V224" "V225" "V226" "V227" "V228" "V229" "V230"  
## [231] "V231" "V232" "V233" "V234" "V235" "V236" "V237" "V238" "V239" "V240"  
## [241] "V241" "V242" "V243" "V244" "V245" "V246" "V247" "V248" "V249" "V250"  
## [251] "V251" "V252" "V253" "V254" "V255" "V256" "V257" "V258" "V259" "V260"  
## [261] "V261" "V262" "V263" "V264" "V265" "V266" "V267" "V268" "V269" "V270"  
## [271] "V271" "V272" "V273" "V274" "V275" "V276" "V277" "V278" "V279" "V280"  
## [281] "V281" "V282" "V283" "V284" "V285" "V286" "V287" "V288" "V289" "V290"  
## [291] "V291" "V292" "V293" "V294" "V295" "V296" "V297" "V298" "V299" "V300"  
## [301] "V301" "V302" "V303" "V304" "V305" "V306" "V307" "V308" "V309" "V310"  
## [311] "V311" "V312" "V313" "V314" "V315" "V316" "V317" "V318" "V319" "V320"  
## [321] "V321" "V322" "V323" "V324" "V325" "V326" "V327" "V328" "V329" "V330"  
## [331] "V331" "V332" "V333" "V334" "V335" "V336" "V337" "V338" "V339" "V340"  
## [341] "V341" "V342" "V343" "V344" "V345" "V346" "V347" "V348" "V349" "V350"  
## [351] "V351" "V352" "V353" "V354" "V355" "V356" "V357" "V358" "V359" "V360"  
## [361] "V361" "V362" "V363" "V364" "V365" "V366" "V367" "V368" "V369" "V370"  
## [371] "V371" "V372" "V373" "V374" "V375" "V376" "V377" "V378" "V379" "V380"  
## [381] "V381" "V382" "V383" "V384" "V385" "V386" "V387" "V388" "V389" "V390"  
## [391] "V391" "V392" "V393" "V394" "V395" "V396" "V397" "V398" "V399" "V400"  
## [401] "V401" "V402" "V403" "V404" "V405" "V406" "V407" "V408" "V409" "V410"  
## [411] "V411" "V412" "V413" "V414" "V415" "V416" "V417" "V418" "V419" "V420"  
## [421] "V421" "V422" "V423" "V424" "V425" "V426" "V427" "V428" "V429" "V430"  
## [431] "V431" "V432" "V433" "V434" "V435" "V436" "V437" "V438" "V439" "V440"  
## [441] "V441" "V442" "V443" "V444" "V445" "V446" "V447" "V448" "V449" "V450"  
## [451] "V451" "V452" "V453" "V454" "V455" "V456" "V457" "V458" "V459" "V460"  
## [461] "V461" "V462" "V463" "V464" "V465" "V466" "V467" "V468" "V469" "V470"  
## [471] "V471" "V472" "V473" "V474" "V475" "V476" "V477" "V478" "V479" "V480"  
## [481] "V481" "V482" "V483" "V484" "V485" "V486" "V487" "V488" "V489" "V490"  
## [491] "V491" "V492" "V493" "V494" "V495" "V496" "V497" "V498" "V499" "V500"  
## [501] "V501" "V502" "V503" "V504" "V505" "V506" "V507" "V508" "V509" "V510"  
## [511] "V511" "V512" "V513" "V514" "V515" "V516" "V517" "V518" "V519" "V520"  
## [521] "V521" "V522" "V523" "V524" "V525" "V526" "V527" "V528" "V529" "V530"  
## [531] "V531" "V532" "V533" "V534" "V535" "V536" "V537" "V538" "V539" "V540"  
## [541] "V541" "V542" "V543" "V544" "V545" "V546" "V547" "V548" "V549" "V550"  
## [551] "V551" "V552" "V553" "V554" "V555" "V556" "V557" "V558" "V559" "V560"  
## [561] "V561"

##1c. READ.TABLE the activity files and check the col. names.  
ActTrain<- read.table ("./UCI HAR Dataset/train/y\_train.txt",   
 header=FALSE)  
ActTest<- read.table ("./UCI HAR Dataset/test/y\_test.txt",   
 header=FALSE)  
names(ActTest)

## [1] "V1"

names(ActTrain)

## [1] "V1"

## 1d. READ.TABLE the feature vector list and activity labels.  
Feature<-read.table("./UCI HAR Dataset/features.txt",   
 header=FALSE)  
ActLabel<-read.table("./UCI HAR Dataset/activity\_labels.txt",   
 header=FALSE)  
names(Feature)

## [1] "V1" "V2"

names(ActLabel)

## [1] "V1" "V2"

## Now it is ready to merge all the data into one big dataset. As we can see  
## that all of them have commom variable=V1, and it is time to change to a specific  
## name before column bind.  
## 1e. First of all, row bind Data, Act and Subject into one dataset.  
Data<- rbind(DataTest, DataTrain)  
any(is.na(Data)) ## Returns "FALSE", no invalid data present.

## [1] FALSE

Act<- rbind(ActTest, ActTrain)  
Subject<- rbind(SubjectTest, SubjectTrain)  
  
## 1f. Change the col.names V1 into descriptive names in Subject and Act.  
names(Act)<-sub("V1", "activity", names(Act))  
names(Subject)<-sub("V1", "subject", names(Subject))  
  
## 1g. Column bind three datasets into one big dataset. Check names to verify.  
Mixdf<- cbind(Subject, Act, Data)  
Mixdf<- arrange(Mixdf, subject, activity)  
  
## 1h. This step to form a dataset is to substitute the data variables with Feature   
## To make a Tidy Data, first it to make sure there is NO duplicated variables, in here, Feature.  
any(duplicated(Feature$V2)) ##this code returns "TRUE", duplication exist.

## [1] TRUE

names(Mixdf)[3:563]<-make.unique(as.character(Feature$V2))  
any(duplicated(names(Mixdf))) ## This returns "FALSE"

## [1] FALSE

## 1i. This is the final step (also required in question3) to replace the activity with corresponding value  
## using match function.  
Mixdf$activity<-ActLabel[match(Mixdf$activity,ActLabel$V1),2]  
head(Mixdf$activity,2) ## this step returns [1]STANDING STANDING. Now the dataset is formed.

## [1] WALKING WALKING  
## 6 Levels: LAYING SITTING STANDING WALKING ... WALKING\_UPSTAIRS

##2.Extracts only the measurements on the mean and standard deviation for each measurement.  
stdnames<-c("subject", "activity", grep("mean|std", names(Mixdf), value=TRUE))  
MeanStd<-Mixdf[, stdnames]  
head(MeanStd[1:3]) ## returns dataframe only contains the measurements with mean or std.

## subject activity tBodyAcc-mean()-X  
## 1 1 WALKING 0.2820216  
## 2 1 WALKING 0.2558408  
## 3 1 WALKING 0.2548672  
## 4 1 WALKING 0.3433705  
## 5 1 WALKING 0.2762397  
## 6 1 WALKING 0.2554682

##3. Uses descriptive activity names to name the activities in the data set  
## This is one is done in Step 1i  
  
##4. Uses descriptive names for each variable, such as replace "t" for "time,  
## "f" for "frequency", "acc" for "accelerator" and ect.  
names(Mixdf)<-tolower(names(Mixdf))  
names(Mixdf)<-gsub("^t", "time", names(Mixdf))  
names(Mixdf)<-gsub("^f", "frequency", names(Mixdf))  
names(Mixdf)<-gsub("acc", "accelerometer", names(Mixdf), fixed=TRUE)  
names(Mixdf)<-gsub("gyro", "gyroscope", names(Mixdf), fixed=TRUE)  
names(Mixdf)<-gsub("mag", "magnitude", names(Mixdf), fixed=TRUE)  
names(Mixdf)<-gsub("bodybody", "body", names(Mixdf), fixed=TRUE)  
  
##5a From the data set in step 4, creates a second,independent tidy data set with the average of   
## each variable for each activity and each subject.  
Data2<-aggregate(. ~subject + activity, Mixdf, mean)  
anyNA(Data2) ## returns [1] FALSE

## [1] FALSE

any(duplicated(Data2, margin=2)) ## returns [1] FALSE

## [1] FALSE

dim(Data2) ## returns [1] 180 563

## [1] 180 563

class(Data2) ## returns [1] "data.frame"

## [1] "data.frame"

## Save both data tables to desinated dir.  
write.csv(Mixdf, file="./UCI HAR Dataset/TotalData.csv") ## The total Data set  
write.csv(Data2, file="./UCI HAR Dataset/MeanData.csv") ## Dataset with average values  
  
##Final: to generate a codebook for this R script using R Markdown (Knit Document)  
##Alternatively, one can generate codebook for each dataset (Mixdf and Data2) using   
##codebook function using memisc package:  
## Write(codebook(Mixdf), file="./UCI HAR Dataset/fulldata.rmd")  
## Write(codebook(Data2), file="./UCI HAR Dataset/averagedata.rmd")