## Getting and Cleaning Data Course Project

## Anahita Saghafi 25/05/2014

##You should create one R script called run\_analysis.R that does the following.

##Merges the training and the test sets to create one data set.

##Extracts only the measurements on the mean and standard deviation for each measurement.

##Uses descriptive activity names to name the activities in the data set

##Appropriately labels the data set with descriptive activity names.

##Creates a second, independent tidy data set with the average of each variable for each activity and each subject.

##Good luck!

##Set working directory and check it

setwd("C:\\Users\\anahitas\\Documents\\coursera\\Getting and Cleaning Data\\getdata-projectfiles-UCI HAR Dataset\\UCI HAR Dataset")

getwd()

##Read "features.txt" and save as "Featutes" with column names "FeatureId and FeatureName" ## 561 obs. of 2 variables

Features <- read.table(".\\features.txt", sep="", header=FALSE,col.names=c("FeatureID", "FeatureName")) ## 2947 obs. of 561 variables

##Read test set data(X\_test.txt), and use FeatureName for column names

testData <- read.table(".\\test\\X\_test.txt", sep="", header=FALSE,col.names=Features$FeatureName) ## 2947 obs. of 561 variables

##Read test activity labels (y\_test.txt) and use "ActivityLabelID " for column names

testActivityLabels <- read.table(".\\test\\y\_test.txt", sep="", header=FALSE, col.names=c("ActivityLabelID")) ## 2947 obs. of 1 variables

##Read test subjects (subject\_test.txt) and use SubjectID for column names

testSubjects <- read.table(".\\test\\subject\_test.txt", sep="", header=FALSE, col.names=c("SubjectID"))

testActivityLabels <- read.table(".\\test\\y\_test.txt", sep="", header=FALSE, col.names=c("ActivityLabelID")) ## 2947 obs. of 1 variables

##Read train set data(X\_train.txt), and use FeatureName for column names

trainData <- read.table(".\\train\\X\_train.txt", sep="", header=FALSE, col.names=Features$FeatureName) ## 7352 obs. of 561 variables

##Read train activity labels (y\_train.txt) and use "ActivityLabelID " for column names

trainActivityLabels <- read.table(".\\train\\y\_train.txt", sep="", header=FALSE, col.names=c("ActivityLabelID")) ## 7352 obs. of 1 variables

##Read train subjects (subject\_train.txt) and use SubjectID for column names

trainSubjects <- read.table(".\\train\\subject\_train.txt", sep="", header=FALSE, col.names=c("SubjectID")) ## 7352 obs. of 1 variables

##Perform rbind to testData and trainData to create mergedSetData

mergedSetData <- rbind(testData, trainData) ## 10299 obs of 561 variables

## Perform rbind to testActivityLabels and trainActiviyLabels to create mergedSetActivityLabels ## 10299 obs. of 1 variables

mergedSetActivityLabels <- rbind(testActivityLabels,trainActivityLabels)

##Perform rbind to testSubjects and trainSubjects to create mergedSetSubjects

mergedSetSubjects <- rbind(testSubjects, trainSubjects) ## 10299 obs of 1 variables

##Read activity labels (activity\_labels.tct) and use "ActvityLabelID" and "ActivityLabelName" for column names ## 10299 obs. of 561 variables

ActivityLabelNames <- read.table(".\\activity\_labels.txt", sep="", header=FALSE,as.is=T, col.names=c("ActivityLabelID", "ActivityLabelName"))

##Replace activity label ids with activity label names and call it ActivityLabelName

ActivityLabelNames[, 2] = gsub("\_", "", tolower(as.character(ActivityLabelNames[, 2])))

mergedSetActivityLabels[,1] = ActivityLabelNames[mergedSetActivityLabels[,1], 2]

names(mergedSetActivityLabels) <- "ActivityLabelName" ## 10299 obs of 1 variables

## Get subset of features related to mean and std and apply to the mergedSetData to get mergedSubsetData

subsetFeatures <- grep(".\*mean\\(\\)|.\*std\\(\\)", Features$FeatureName) ## int[1:66] values

mergedSubsetData <- mergedSetData[,subsetFeatures] ## 10299 obs of 66 variables

##Add activity label names (ActivityLabelName) and subject id (SubjectID) to the merged sub data (mergetSubsetData)

mergedSubsetData$ActivityLabelName <- mergedSetActivityLabels$ActivityLabelName ## 10299 obs. of 67 variables

mergedSubsetData$SubjectID <- mergedSetSubjects$SubjectID ## 10299 obs of 68 variables

## Finally write clean data with descriptive activity names into mergedSubsetData.txt

write.table(mergedSubsetData, "mergedSubsetData.txt") ##10299 obs. of 68 variables

## Use library data.tabe to create a data table megedSubsetDataTable

library(data.table)

## Create mergedSubsetDataTable from mergedSubsetData

mergedSubsetDataTable <- data.table(mergedSubsetData)

## Use columns SubjectID and ActivilityLabelNames of mergedSubsetDataTable to create average data

AverageMergedSubSetData <- mergedSubsetDataTable[, lapply(.SD, mean), by=c("SubjectID", "ActivityLabelName")]

##OrderAverageMegedSubsetData by SubjectIDs

AverageMergedSubSetData <- AverageMergedSubSetData[order(AverageMergedSubSetData$SubjectID),] ##180 obs. of 68 variables

## Write the second data set "AverageMegedSubsetData" to complete the course project.

write.table(AverageMergedSubSetData, "averageCleanedData.txt") ## 10299 obs. of 67 variables