# Getting and cleaning data - Assingment #1

## by Zzhuk

## Introduction

The purpose of this project is to demonstrate your ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis. You will be graded by your peers on a series of yes/no questions related to the project. You will be required to submit: 1) a tidy data set as described below, 2) a link to a Github repository with your script for performing the analysis, and 3) a code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md. You should also include a README.md in the repo with your scripts. This repo explains how all of the scripts work and how they are connected.

One of the most exciting areas in all of data science right now is wearable computing - see for example this article . Companies like Fitbit, Nike, and Jawbone Up are racing to develop the most advanced algorithms to attract new users. The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description is available at the site where the data was obtained:

<http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

Here are the data for the project:

<https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

You should create one R script called run\_analysis.R that does the following 5 steps:

## Part 1

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

## considering zip file is downloaded and saved under working directory  
## unzip("getdata-projectfiles-UCI HAR Dataset.zip")  
  
## test data:  
XTest<- read.table("UCI HAR Dataset/test/X\_test.txt")  
YTest<- read.table("UCI HAR Dataset/test/Y\_test.txt")  
SubjectTest <-read.table("UCI HAR Dataset/test/subject\_test.txt")  
  
## train data:  
XTrain<- read.table("UCI HAR Dataset/train/X\_train.txt")  
YTrain<- read.table("UCI HAR Dataset/train/Y\_train.txt")  
SubjectTrain <-read.table("UCI HAR Dataset/train/subject\_train.txt")  
  
## features and activity  
features<-read.table("UCI HAR Dataset/features.txt")  
activity<-read.table("UCI HAR Dataset/activity\_labels.txt")  
  
##Part1 - merges train and test data in one dataset (full dataset at the end)  
X<-rbind(XTest, XTrain)  
Y<-rbind(YTest, YTrain)  
Subject<-rbind(SubjectTest, SubjectTrain)

Dimension of new datasets:

dim(X)

## [1] 10299 561

dim(Y)

## [1] 10299 1

dim(Subject)

## [1] 10299 1

## Part 2

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

index<-grep("mean\\(\\)|std\\(\\)", features[,2])  
length(index) ## count of variables

## [1] 66

X<-X[,index] ## getting only variables iwth mean/stdev  
dim(X) ## checking dim of subset

## [1] 10299 66

## Part 3

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

Y[,1]<-activity[Y[,1],2] ## won't reorder Y set  
head(Y)

## V1  
## 1 STANDING  
## 2 STANDING  
## 3 STANDING  
## 4 STANDING  
## 5 STANDING  
## 6 STANDING

## Part 4

Appropriately labels the data set with descriptive variable names.

names<-features[index,2] ## getting names for variables  
   
 names(X)<-names ## updating colNames for new dataset  
 names(Subject)<-"SubjectID"  
 names(Y)<-"Activity"  
   
 CleanedData<-cbind(Subject, Y, X)  
 head(CleanedData[,c(1:4)]) ## first 5 columns

## SubjectID Activity tBodyAcc-mean()-X tBodyAcc-mean()-Y  
## 1 2 STANDING 0.2571778 -0.02328523  
## 2 2 STANDING 0.2860267 -0.01316336  
## 3 2 STANDING 0.2754848 -0.02605042  
## 4 2 STANDING 0.2702982 -0.03261387  
## 5 2 STANDING 0.2748330 -0.02784779  
## 6 2 STANDING 0.2792199 -0.01862040

## Part 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.

CleanedData<-data.table(CleanedData)  
TidyData <- CleanedData[, lapply(.SD, mean), by = 'SubjectID,Activity']  
dim(TidyData)

## [1] 180 68

write.table(TidyData, file = "Tidy.txt", row.names = FALSE)

First 12 rows and 5 columns in Tidy dataset:

head(TidyData[order(SubjectID)][,c(1:4), with = FALSE],12)

## SubjectID Activity tBodyAcc-mean()-X tBodyAcc-mean()-Y  
## 1: 1 STANDING 0.2789176 -0.016137590  
## 2: 1 SITTING 0.2612376 -0.001308288  
## 3: 1 LAYING 0.2215982 -0.040513953  
## 4: 1 WALKING 0.2773308 -0.017383819  
## 5: 1 WALKING\_DOWNSTAIRS 0.2891883 -0.009918505  
## 6: 1 WALKING\_UPSTAIRS 0.2554617 -0.023953149  
## 7: 2 STANDING 0.2779115 -0.018420827  
## 8: 2 SITTING 0.2770874 -0.015687994  
## 9: 2 LAYING 0.2813734 -0.018158740  
## 10: 2 WALKING 0.2764266 -0.018594920  
## 11: 2 WALKING\_DOWNSTAIRS 0.2776153 -0.022661416  
## 12: 2 WALKING\_UPSTAIRS 0.2471648 -0.021412113