# UCI HAR Datasets cleaning & processing

*by Ameen AboDabash 29-Feb 2016* as part of Data Cleaning course [assignment](https://www.coursera.org/learn/data-cleaning/peer/FIZtT/getting-and-cleaning-data-course-project), for more information about raw data and information about experiments carried on to load the raw data [available here](http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones).

we encasulate getting the raw data and clean it in four different functions located in ( <.R>/run\_analysis.R ), we're using two packages and by sourcing the run\_analysis.R file it will automatically load the required packages: just call the follwing line of code:

packages<-c("data.table", "dplyr")  
sapply(packages, require, character.only = TRUE)

you've to call run\_analysis() function and it will do the following: 1- download raw data if not donwloaded before, and extract it to <./data-raw> dir 2- read only required datasets and merg it into different sets (we will descipe this in more details later) 3- tidying data. 4- export pretidying data "mergedData.txt" & tidy dataset "tidyData.txt" to <./data> dir

tidyDataDF <- run\_analysis()   
## there's optional parameter refering to raw-data donwload url

lets digg into run\_analysis() function code:

run\_analysis <- function(extURL = "..."){  
   
 dfList <- readRequiredFiles(extURL)  
   
   
 ## tidying activities data  
 activitiesDF <- dfList[[1]] %>%   
 select(V2) %>% ## select only Activity Name column- no need for IDs  
 setNames("activity") ## set new DF col name  
   
   
 ## tidying subjects data  
 subjectsDF <- dfList[[2]] %>%   
 setNames("subject") ## set new DF col name  
   
   
 ## tidying featuers data  
 featuersDF <- dfList[[3]] %>%   
 ## Select only measurments with mean and sd using grep  
 ## resulting arround 66 cols from 500+ cols  
 select(num\_range("V",grep(x = dfList[[4]]$V2, "mean\\(\\)|std\\(\\)"))) %>%  
 ## set new DF col names using the same techniqe  
 setNames(nm = getDescriptiveFeaturesLabels( grep(x = dfList[[4]]$V2, "mean\\(\\)|std\\(\\)",value = T)))  
   
   
 ## creating a datatable (easy to maiting package data.table)  
 ## combine all tidied DF (subjects, activities & featuers) using cbind  
 tempTidyDF <- data.table( cbind(subjectsDF,activitiesDF,featuersDF) )   
   
   
 ## cooking the second tidying dataset  
 ## aggregate by ( subject & activity ), and apply mean for remaining cols (.)   
 secondTidyDF <- aggregate(. ~ subject - activity, data = tempTidyDF, mean) %>%  
 ## reorder data  
 arrange(subject,activity) %>%  
 ## add the perfix "avg." to all measurments vairables  
 setNames( nm= c("subject","activity",paste("avg.", getDescriptiveFeaturesLabels(grep(x = dfList[[4]]$V2, "mean\\(\\)|std\\(\\)",value = T )) ,sep = "") ) )   
   
  
 ## write both files to output directory  
 write.table(tempTidyDF,file = "data\\MergedData.txt", row.name=FALSE ) ## not required  
 write.table(secondTidyDF,"data\\TidyData.txt", row.name=FALSE) ## required as per assignment  
   
   
 secondTidyDF  
  
}