**Readme for JHU Data Cleaning Assignment**

**By Peter Johnson**

This document is submitted to fulfil the Assignment for the JHU Data Cleaning course on Coursera getting and cleaning Data . See [the Coursera assignment page](https://www.coursera.org/learn/data-cleaning/peer/FIZtT/getting-and-cleaning-data-course-project)

**Other related file**

* run\_analysis.R R script performing the actions required to get a tidy data set from original data project
* tidydataset.txt resulting tidy data set file from execution of above script
* Codebook.md describing the tidy data set structure and contents

**Information**

Unzipping this archive should create the directory UCI HAR Dataset where the script will look for data files.

**Some libraries are required** for the script to run:

* data.table for its fread function that behave well with white spaces
* dplyr and stringr for the mutate fonction (used to set/rename labels)

Please note that the script does not install required libraries if not present (for the same security reasons as for download mentioned at the begining of this document : any package installation should only be done by users).

Instead of removing the unwanted measures after loading data files, **the scripts does only read the necessary columns** from the X\_train.txt and X\_test.txt files.  
Those columns are the ones about either mean() or std() (standard deviation) measurements.

**To get some nice descriptive variable names**, the following is performed:

1. activities id are exchanged with their corresponding descriptive names (based on activity\_labels.txt)
2. variable names (columns) are renamed to more descriptive and understandable ones in both X\_train.txt and X\_test.txt files ; This could have been done in only one step after the merge but this way it is more easy to debug the script in case of some strange results.  
   It also could be arguable that names could be even more explicit or more grammaticaly correct but a balance between Shakespear and conciseness had to be found!

**The merge of X\_train and X\_test data** is done by concatenate the data tables by rows , set names to variables and merge columns to get the data frame. .

At last, **the final tidy data set is obtained** by *"averaging each variable for each activity and each subject"*...*)*

After executing run\_analysis.R script, **you can view the resulting tidy data set** with the following command:

Data2<-aggregate(. ~subject + activity, Data, mean)

Data2<-Data2[order(Data2$subject,Data2$activity),]

write.table(Data2, file = "tidydata.txt",row.name=FALSE)

Data2

**Do not forget to check the codebook (Codebook.md) for further information about the tidy data set content!**

**Credits for the data set**

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

Peter Johnson 09 June 2018.