Instruction List for run\_analysis.R - ##Annotated

Jeff Gonlin April 2016

This code has been written in as part of the Coursera Data Cleaning course, week 4 assignment. The code below is taken straight out of the R file “run\_analysis.R” and contains a considerable volume of ##commetary which should shed some light on my mental ramblings as I sketched out the code.

## Coursera Data Week4 Assignment

## WD should be set to "Desktop", where I've established a folder "Week4 Assignment", i.e.,

setwd("C:/Users/Jeff/Desktop/Week4 Assignment")

library(plyr)

library(dplyr)

library(tidyr)

library(R.utils)

library(data.table)

fileUrl = "https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip"

download.file(fileUrl, destfile = "./Dataset.zip")

Sys.Date() #2016-04-02 as of test run

## PLEASE READ BEFORE PROCEDING:

## TO PRESERVE THE ORIGINAL DATA: Go to the desktop / WD 'Assignment' folder and open "Dataset.zip"

## I WOULD DO THIS MANUALLY:

## COPY the "UCI HAR Dataset" folder but take no other action with it.

## GO UP ONE LEVEL AND PASTE IT. THIS IS THE FILE I WILL WORK WITH.

## Both "Dataset.zip" and "UCI HAR Dataset" should be found in: "C:/Users/Jeff/Desktop/Week4 Assignment"

## The original "Dataset.zip" download will remain untouched as original source data.

#######################################################################################################

datasetunzip <- unzip("./Dataset.zip",list=TRUE)

str(datasetunzip)

View(datasetunzip)

## subject, test and train files of interest. The "features" and "features\_info" files will also find use.

subject\_test <- read.table("./UCI HAR Dataset/test/subject\_test.txt")

subject\_train <- read.table("./UCI HAR Dataset/train/subject\_train.txt")

X\_test <- read.table("./UCI HAR Dataset/test/X\_test.txt", sep = "")

X\_train <- read.table("./UCI HAR Dataset/train/X\_train.txt", sep = "")

y\_test <- read.table("./UCI HAR Dataset/test/y\_test.txt", sep = "")

y\_train <- read.table("./UCI HAR Dataset/train/y\_train.txt", sep = "")

## Let's combine 'subject', 'X' and 'y' files and call them "subjects", "activities" and "feature\_vectors" (in keeping with ReadMe).

subjects <- bind\_rows(subject\_test, subject\_train)

activities <- bind\_rows(y\_test, y\_train)

feature\_vectors <- bind\_rows(X\_test, X\_train)

## set up file to decode progressively the activities and substitute descriptive text.

activity\_labels <- read.table("./UCI HAR Dataset/activity\_labels.txt")

activities1 <- data.table(gsub(1, "walking", activities$V1))

activities2 <- data.table(gsub(2, "walkingupstairs", activities1$V1))

activities3 <- data.table(gsub(3, "wakingdownstairs", activities2$V1))

activities4 <- data.table(gsub(4, "sitting", activities3$V1))

activities5 <- data.table(gsub(5, "standing", activities4$V1))

activities6 <- data.table(gsub(6, "laying", activities5$V1))

rm(activities, activities1, activities2, activities3, activities4, activities5, activity\_labels) # Not needed

## Now add an index, give proper colnames and merge 'subjects' and 'activities6' files

subjects <- mutate(subjects, index = seq\_along(V1))

activities6 <- mutate(activities6, index = seq\_along(V1))

colnames(subjects) <- c("subject", "index")

colnames(activities6) <- c("activity", "index")

subjects\_activities <- merge(subjects, activities6)

features <- read.table("./UCI HAR Dataset/features.txt", sep = "")

str(features) ## dim = 561 obs x 2 variables.

## These entries correspond to the 561 colname variables (V1, V2, etc.) in 'feature\_vectors'.

## Make tidy and use them as is to add the col names to 'feature\_vectors'.

featuresV2 <- c(as.character(features$V2))

featuresV2 <- tolower(featuresV2)

featuresV2 <- gsub("[()]", "", featuresV2)

feature\_vectors <- setNames(feature\_vectors, featuresV2)

## Could remove "-" symbols, however I find the names much more readable with, so will retain them.

## Extract variables relating to: mean(): Mean value and std(): Standard deviation

feature\_vectors\_meanlabels <- data.frame(select\_vars(colnames(feature\_vectors), matches(".mean.")))

feature\_vectors\_stdlabels <- data.frame(select\_vars(colnames(feature\_vectors), matches(".std.")))

## I don't believe the "Angle" vectors (e.g., "Angle(X, gravityMean")) should be included. Select out.

feature\_vectors\_meanlabels\_unwanted <- data.frame(select\_vars(rownames(feature\_vectors\_meanlabels), matches("^Angle.")))

feature\_vectors\_meanlabels <- slice(feature\_vectors\_meanlabels, -c(47:53))

rm(feature\_vectors\_meanlabels\_unwanted)

## Retain the rownames as they will be needed in column selection

rownames(feature\_vectors\_meanlabels) <- feature\_vectors\_meanlabels[ ,1]

## Enforce a common colname on both 'label' files so they can be combined; remove uneeded files.

colnames(feature\_vectors\_meanlabels) <- c("mean\_std\_labels")

colnames(feature\_vectors\_stdlabels) <- c("mean\_std\_labels")

feature\_vectors\_mean\_std\_labels <- rbind(feature\_vectors\_meanlabels, feature\_vectors\_stdlabels)

rm(feature\_vectors\_meanlabels, feature\_vectors\_stdlabels)

feature\_vectorsDT <- as.data.table(feature\_vectors)

feature\_vectors\_mean\_std\_cols <- feature\_vectorsDT[,c(rownames(feature\_vectors\_mean\_std\_labels)), with=FALSE]

## Now combine 'subjects\_activities' with 'feature\_vectors\_mean\_std\_cols' to get one single complete dataset.

feature\_vectors\_mean\_std\_cols <- mutate(feature\_vectors\_mean\_std\_cols, index=seq\_len(10299))

one\_data\_set <- merge(subjects\_activities, feature\_vectors\_mean\_std\_cols)

one\_data\_set <- as.data.table(one\_data\_set)

View(one\_data\_set)

## 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.

tidy\_data\_set <- one\_data\_set

tidy\_data\_set <- tidy\_data\_set[,index:= NULL]

## Group and summarize:

tidy\_data\_grouped\_averaged <- ddply(tidy\_data\_set, .(subject, activity), colwise(mean))

## Clean up!

rm(activities6, subjects, subjects\_activities, subject\_train, subject\_test, X\_test, X\_train, y\_test, y\_train, features)

View(tidy\_data\_grouped\_averaged)

write.table(tidy\_data\_grouped\_averaged, file = "C:/Users/Jeff/Desktop/Week4 Assignment/tidy\_data\_grouped\_averaged\_saved", sep="\t", row.name=FALSE)