Coursera

Data Cleaning Course

Course Project Code Book

Describing each variable and its values in tidy data set

**Table of Contents**

Introduction 1

List of relevant files 2

Summary of Choices Made: 2

Study Design 3

Code Book 4

Instruction List (R-Script) 7

The Environment 7

Description of the R program Script: run\_analysis.R 7

Performance: 7

Instructions for running the script 8

The R program script run\_analysis.R 8

License: 10

References 10

# Introduction

The original data files were provided by the course faculty in the form of a zip file, namely, dataset.zip. Extracting this file creates a top level folder “UCI HAR Dataset” under current directory. This folder will contain other sub-folders and files.

This codebook describes the final tidy version of dataset created, detailing the variables and units of measure etc. This codebook also contains the instruction set, namely, R-program scrip that extracts and transforms the data in the original data set into the new tidy dataset.

The Summary of Choices Made section details the choices made during the process of cleaning up the data to produce the desired tidy data set.

The codebook section below details the variables in the tidy data set.

The Instruction List section contains the R-Program scrip run\_analysis.R that creates the tidy dataset.

# List of relevant files

* Original UCI HAR Dataset zip file provided for the Data Cleaning project, which is assumed to be extracted under a “current directory” that creates UCI HAR Dataset folder and contains several subfolders and files.
* CodeBook.docx (This current file)
* CodeBook.md (same as CodeBook.docx but in MarkDown format)
* Readme.md (A file that is to be read first before using any other files as a part of the course project submission).
* run\_analysis.R (R Script file that reads the raw data files and produces two tidy data set csv files, as required for the course submission). The script takes no parameters.
* tidy\_full\_data.csv (output file created by run\_analysis.R script in the “current directory”) containing the tidied data with Subject, Activity and 86 variables that contain “mean” or “std” as a part of the variable name.
* tidy\_avgs\_data.csv (output file created by run\_analysis.R script in the “current directory”) containing tidied data with averages for each variable for each Subject-Activity combination.

# Summary of Choices Made:

1. It is assumed that the raw data zip file is extracted in the current directory, so that the “UCI HAR Dataset” folder is created there.
2. The run\_analysis.R script is also assumed to be located under the current directory.
3. run\_analysis.R is the R-program script that creates the tidy data set from the original raw data files under “UCI HAR Dataset” folder
4. There are 561 features (variables) in the original data. However, only a subset of these variables, namely, those variables that contain “mean” or “std” in their names were included the tidy dataset. For this purpose, case-insensitive search is programmatically made to identify such variables. Only 86 variables from the original set of 561 variables satisfy this criterion.
5. For deriving meaningful names, the parentheses from the original variable names have been removed in the tidy dataset
6. Subject and Activity variables are added to the data set, by combining data from appropriate raw data files .
7. Activity column in the tidy data set shows the description (and not a numerical code) in the tidy data set as per the course project instructions.
8. Certain libraries are used in the R Script and it is assumed that they will be available.
9. The tidy data set files are created by the run\_analysis.R script in the current directory.

# Study Design

Scope of this course project does not include Study design or experiment. The data from a study “Human Activity Recognition Using Smartphones Dataset” is provided as a part of the course project instructions. However, the dataset’s readme.txt file provides the description of the study.

The following excerpt is taken from readme.txt file of the original raw data zip file, that can be located under UCI HAR Dataset folder

**BEGIN EXCERPT**

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain. See 'features\_info.txt' for more details.

For each record it is provided:

- Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.

- Triaxial Angular velocity from the gyroscope.

- A 561-feature vector with time and frequency domain variables.

- Its activity label.

- An identifier of the subject who carried out the experiment.

**END EXCERPT**

# Code Book

**Description of the variables including Units of measure**

There are totally 88 variables in both the tidy data sets. The first two are Subject and Activity. The next 86 variables constitute a subset variable names of the original data set. This subset is derived by including in it only the variables with the names that include “mean” or “std” key words. Also, the parentheses from the original names removed.. The last 3 variables are added to include details of subjects and activities. They are described below.

**The First Two variables**

The first two variables are Subject and Activity.

**Subject**: Subject is the rowed of person who performed the activity for each window sample. Its range is from 1 to 30

**Activity**: The activity is not a numeric code but the descriptive text. The values of Activity are: one of the following six:

* WALKING,
* WALKING\_UPSTAIRS
* WALKING\_DOWNSTAIRS
* SITTING
* STANDING
* LAYING

**The next 86 variables**

* Features are normalized and bounded within [-1,1].
* Each feature vector is a row on the text file.

The next 86 variables are normalized and bounded within [-1,1]. The description of various variables are included in the original raw data set, (see the README.TXT in the original raw data, under UCI HAR Dataset, so they are not being repeated here. A relevant excerpt is included in previous sections. Please refer to features\_info.txt for more info. Keep in mind that the variables names in the tidy data set have slightly different from the original variable names in the raw data, as described in summary of choices made section above – namely, the parenthesis in the names have been removed.

Hence, no additional description about the units of measure are included.

The variable names indicate whether a variable is a “mean” or a “std”. The original variable names are in the features.txt. These original names are transformed in two ways: (a) any parenthesis () from the names are removed. The original data has 561 variables. However, only those variables that have “mean” or “std” in the names have been included in this tidy data set, as per the course project instructions.

[1] "Subject"

[2] "Activity"

[3] "tBodyAcc-mean-X"

[4] "tBodyAcc-mean-Y"

[5] "tBodyAcc-mean-Z"

[6] "tBodyAcc-std-X"

[7] "tBodyAcc-std-Y"

[8] "tBodyAcc-std-Z"

[9] "tGravityAcc-mean-X"

[10] "tGravityAcc-mean-Y"

[11] "tGravityAcc-mean-Z"

[12] "tGravityAcc-std-X"

[13] "tGravityAcc-std-Y"

[14] "tGravityAcc-std-Z"

[15] "tBodyAccJerk-mean-X"

[16] "tBodyAccJerk-mean-Y"

[17] "tBodyAccJerk-mean-Z"

[18] "tBodyAccJerk-std-X"

[19] "tBodyAccJerk-std-Y"

[20] "tBodyAccJerk-std-Z"

[21] "tBodyGyro-mean-X"

[22] "tBodyGyro-mean-Y"

[23] "tBodyGyro-mean-Z"

[24] "tBodyGyro-std-X"

[25] "tBodyGyro-std-Y"

[26] "tBodyGyro-std-Z"

[27] "tBodyGyroJerk-mean-X"

[28] "tBodyGyroJerk-mean-Y"

[29] "tBodyGyroJerk-mean-Z"

[30] "tBodyGyroJerk-std-X"

[31] "tBodyGyroJerk-std-Y"

[32] "tBodyGyroJerk-std-Z"

[33] "tBodyAccMag-mean"

[34] "tBodyAccMag-std"

[35] "tGravityAccMag-mean"

[36] "tGravityAccMag-std"

[37] "tBodyAccJerkMag-mean"

[38] "tBodyAccJerkMag-std"

[39] "tBodyGyroMag-mean"

[40] "tBodyGyroMag-std"

[41] "tBodyGyroJerkMag-mean"

[42] "tBodyGyroJerkMag-std"

[43] "fBodyAcc-mean-X"

[44] "fBodyAcc-mean-Y"

[45] "fBodyAcc-mean-Z"

[46] "fBodyAcc-std-X"

[47] "fBodyAcc-std-Y"

[48] "fBodyAcc-std-Z"

[49] "fBodyAcc-meanFreq-X"

[50] "fBodyAcc-meanFreq-Y"

[51] "fBodyAcc-meanFreq-Z"

[52] "fBodyAccJerk-mean-X"

[53] "fBodyAccJerk-mean-Y"

[54] "fBodyAccJerk-mean-Z"

[55] "fBodyAccJerk-std-X"

[56] "fBodyAccJerk-std-Y"

[57] "fBodyAccJerk-std-Z"

[58] "fBodyAccJerk-meanFreq-X"

[59] "fBodyAccJerk-meanFreq-Y"

[60] "fBodyAccJerk-meanFreq-Z"

[61] "fBodyGyro-mean-X"

[62] "fBodyGyro-mean-Y"

[63] "fBodyGyro-mean-Z"

[64] "fBodyGyro-std-X"

[65] "fBodyGyro-std-Y"

[66] "fBodyGyro-std-Z"

[67] "fBodyGyro-meanFreq-X"

[68] "fBodyGyro-meanFreq-Y"

[69] "fBodyGyro-meanFreq-Z"

[70] "fBodyAccMag-mean"

[71] "fBodyAccMag-std"

[72] "fBodyAccMag-meanFreq"

[73] "fBodyBodyAccJerkMag-mean"

[74] "fBodyBodyAccJerkMag-std"

[75] "fBodyBodyAccJerkMag-meanFreq"

[76] "fBodyBodyGyroMag-mean"

[77] "fBodyBodyGyroMag-std"

[78] "fBodyBodyGyroMag-meanFreq"

[79] "fBodyBodyGyroJerkMag-mean"

[80] "fBodyBodyGyroJerkMag-std"

[81] "fBodyBodyGyroJerkMag-meanFreq"

[82] "angletBodyAccMean,gravity"

[83] "angletBodyAccJerkMean,gravityMean"

[84] "angletBodyGyroMean,gravityMean"

[85] "angletBodyGyroJerkMean,gravityMean"

[86] "angleX,gravityMean"

[87] "angleY,gravityMean"

[88] "angleZ,gravityMean"

# Instruction List (R-Script)

## The Environment

The following R program is used to read the data files and create the tidy data set files in csv format. The programs are developing in RStudio. The software version details are below:

**RStudio**

Version 0.99.483 – © 2009-2015 RStudio, Inc.

Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_11\_5) AppleWebKit/601.6.17 (KHTML, like Gecko

**R**

R version 3.2.1 (2015-06-18)

## Description of the R program Script: run\_analysis.R

R program script is run\_analysis.R. This script is interspersed with comment lines to make it readable. The code is divided into sections with section headings in comment lines. The sections or STEPs correspond to the functionality bullet points for run\_analysis.R contained in the Course Project Instructions These are

1. Merging training and test data sets,
2. Extracting data only for variables that have “mean” or “std” in the variable names in the original data sets,
3. Using Descriptive Activity names, rather than codes,
4. Using descriptive variable names (we removed parentheses “()”]
5. Creating a new tidy data sets that shows the averages for each variable for each Subject-Activity combination.
6. Writing the tidied full data set and tidied averages-for-subject-activity combinations data set.

There are also several write statements included in the code to see log of activities being performed while the code is running.

## Performance:

It took about 2-3 minutes for the entire code to run and generate the csv formatted tidy data set files on Macbook Pro with 20 GB ram.

## Instructions for running the script

* Download the original raw data zip file provided by the course instructors from [https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip](https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip" \t "_blank)
* Extract the zip file in the current folder
* Start RStudio (preferably, though you can as well start R) in the current folder
* From RStudio run the script using the following at the RStudio Console prompt:

**source(“run\_analysis.R”)**

* The script creates two tidy data files in the current folder

## The R program script run\_analysis.R

# Gopal G

# Data Cleaning Course Coursera

# Course Project: Data Cleaning

# Script: run\_analysis.R

# Step 1: Read two following two files and merge the data sets

# ./test/x\_test.txt

# ./train/x\_train.txt

library(data.table)

library (dplyr)

# STEP 1 MERGING TRAINING AND TEST DATA SETS

print("STEP 1: Merging the test and training datasets")

print(" Reading data files into R")

print(" Reading features.txt into features")

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

print(" Reading activity\_lables.txt into activities")

activities <- read.table("./UCI HAR Dataset/activity\_labels.txt",stringsAsFactors = FALSE)

print(" Reading X\_test.txt into Xtest")

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

print(" Reading Y\_test.txt into Ytest")

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

print(" Reading subject\_test.txt into SubjectTest")

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

print(" Reading X\_train.txt into Xtrain")

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

print(" Reading y\_train.txt into Ytrain")

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

print(" Reading subject\_train.txt into SubjectTrain")

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

print(" Reading data files Complete" )

print(" Merging test and training datasets")

print(" Merging (rbind) Xtest and Xtrain into X")

X <- rbind(Xtest,Xtrain)

print(" Merge(rbind) Ytest and Ytrain into Y")

Y <- rbind(Ytest,Ytrain)

print(" Merge(rbind) SubTest and SubTrain into Subject")

Subject <- rbind(SubTest,SubTrain)

print(" Merging test and training datasets complete")

# STEP 2 EXTRACTING ONLY MEAN AND STD COLUMNS

print("STEP 2: Extracting only mean and std columns")

print(" Naming the columns of dataset from features")

colnames(X) <- gsub("[()]","", features$V2)

print(" Giving name Subject to the Subject DF col")

names(Subject) <- "Subject"

print(" Giving name Activity to the Y DF col")

names(Y) <- "Activity"

#select(data\_set, Subject, everything())

print(" Finding columns with std/mean in the name")

print(" into mean\_std\_cols")

mean\_std\_cols <- grep("mean|std",colnames(X), ignore.case = TRUE, value=TRUE)

print(" Extracting only columns with std/mean")

X <- X[,mean\_std\_cols]

print(" Merging(cbind) Activity and X into data\_set")

data\_set <- cbind(Y, X)

print(" Merging(cbind) Subject and data\_set into data\_set")

data\_set <- cbind(Subject,data\_set )

# STEP # USING DESCRIPTIVE ACTIVITY NAMES

print("STEP 3: Using Descriptive Activity Names ")

data\_set$Activity <- as.factor(data\_set$Activity)

levels(data\_set$Activity) <- activities[,2]

# STEP 4 LABEL WITH DESCRIPTIVE VARIABLE NAMES

print("STEP 4: label with descriptive variable names")

print(" Accomplished in STEP 2 above")

# STEP 5 CREATING TIDY DATA SET OF AVERAGES

# FOR ALL SUBJECT ACTIVITY COMBINATIONS

print("STEP 5: create an independent data set")

data\_set\_avgs <- group\_by(data\_set,Subject,Activity) %>% summarise\_each(funs(mean))

print("WRITING TIDY DATA FILES FOR SUBMISSION")

print(" Writing full tidy data set")

write.csv(data\_set, "./tidy\_full\_data.csv", row.names = FALSE)

print(" Writing tidy data set of avgs")

write.csv(data\_set\_avgs, "./tidy\_avgs\_data.csv", row.names = FALSE)

# License:

The following references are reproduced as per the licensing requirements of the original data set from which the tidy data set is constructed.

## References

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012

This dataset is distributed AS-IS and no responsibility implied or explicit can be addressed to the authors or their institutions for its use or misuse. Any commercial use is prohibited.

© Jorge L. Reyes-Ortiz, Alessandro Ghio, Luca Oneto, Davide Anguita. November 2012.