Coursera – Getting and Cleaning Data – Assignment Course Project – Merrigan

Using the following data…

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

A script titled run\_analysis.r  was created to do the following:

1. Merges the training and the test sets to create one data set.
2. Extracts only the measurements on the mean and standard deviation for each measurement.
3. Uses descriptive activity names to name the activities in the data set
4. Appropriately labels the data set with descriptive variable names.
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.

Below is a description of each step beyond what is provided in the script notes.

1. **Get Data** 
   * Here we first must pull the necessary library (dplyr)
   * Download the data set highlighted above using download.file from the folder provided named UCI HAR Dataset
   * Lastly, name the data frames accordingly that are being pulled from the folder:
     + Name given = file : number rows, columns = *brief description*
     + features = features.txt : 561 rows, 2 columns = *accelerometer (tAcc-XYZ) and gyroscope (tGyro-XYZ) raw signals*
     + activities = activity\_labels.txt : 6 rows, 2 columns *Activities performed that were measured with these tools and its labels*
     + subject\_test = test/subject\_test.txt : 2947 rows, 1 column = *data of 9/30 participants*
     + x\_test = test/X\_test.txt : 2947 rows, 561 columns = *features test data*
     + y\_test = test/y\_test.txt : 2947 rows, 1 columns = *data of activities’ labels*
     + subject\_train = test/subject\_train.txt : 7352 rows, 1 column = *data of 21/30 participants*
     + x\_train = test/X\_train.txt : 7352 rows, 561 columns = *features train data*
     + y\_train = test/y\_train.txt : 7352 rows, 1 columns = *train activities’code labels*
2. **Question 1: Merges the training and the test sets to create one data set**
   * Due to the breakdown of files, first you must merge the rows of similar purposes, i.e. x rows, y rows, and subject rows. Then you are able to merge all columns and create the full data sheet.
   * Use rbind() to merge x\_train and x\_test and create X (10299 rows, 561 columns)
   * Use rbind() to merge y\_train and y\_test and create Y (10299 rows, 1 column)
   * Use rbind() to merge subject\_train and subject\_test and create Subject (10299 rows, 1 column)
   * Lastly, use cbind() to merge columns of Subject, Y, and X into the complete data set named Merged\_Data (10299 rows, 563 column)
3. **Question 2: Extracts only the measurements on the mean and standard deviation for each measurement**
   * Create a tidy data set, which includes only the variables of interest (i.e. mean and standard deviation), by creating a subset of Merged\_Data. Include only the subject, code, and measurements of mean and standard deviation and name TidyData, which will output 10299 rows and 88 columns)
4. **Question 3: Uses descriptive activity names to name the activities in the data set**
   * The activities file uploaded in part 1 contains a list of code for each activity, in files since the code column has included numbers. To make this easier to understand to future readers of the data set, you may replace the numbers of code in the code column of TidyData to the corresponding activity in activities variable (i.e. 1 will become walking).
5. **Question 4: Appropriately labels the data set with descriptive variable names**
   * Similar to question 3, and for the purposes of human readability, coding the variable names to read as full words is useful. Thus, the following is performed:
     + Code column renamed to activities
     + ACC renamed to Accelerometer
     + Gyro renamed to Gyroscope
     + BodyBody renamed to Body
     + Mag renamed to Magnitude
     + f renamed to Frequency
     + t renamed to Time
6. **Question 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**
   * To create this data set, the first step is grouping by activity and subject, then the means can be calculated accordingly, and the table written for mean of each activity for each subject to create a FinalData of 180 rows and 88 columns
   * Like we import tables, we can write this data set to create a txt file of the table
7. Lastly, just to double check the work you can use the str() function and display the names, degrees of freedom, variable type, and head the first few rows. This will allow a quick visual to whether the script was correctly run.