**Data Cleaning Code Book**

**A. Folder and Files Setup:**

1. Create a working folder on your computer.
2. Copy the "data" folder and all its contents from this repo to your working folder. the "data" folder contains a sub folder "UCI\_HAR\_Dataset" which contains the "unclean" data. (Note the underscores and no spaces in the folder name)
3. Copy the R script file "run\_analysis.R" from this repo to your working folder.
4. Run the script.

**B. Running the script "run\_analysis.R" produces three data sets that satisfy the following requirements:**

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.

**C. Steps in The Data Cleaning and Analysis (Corresponds to section numbering in the script file "run\_analysis.R"):**

1. The descriptive activity labels (activity\_labels.txt) are joined to the training labels (y\_train.txt) and test labels (y\_test.txt) to create data-frames "labels\_ytrain" and "labels\_ytest" respectively. These data-frame contains descriptive activity names.
2. The descriptive feature names are extracted from "features.txt" to a vector "vfeatures" for use in the next step.
3. Descriptive column names are added to the training data. Then, the training subjects from "subject\_train.txt" are merged with the training labels and the training data and saved as a data-frame "subjecttrain". This data-frame contains descriptive activity names and descriptive column names.
4. Descriptive column names are added to the test data. Then, the test subjects from "subject\_test.txt" are merged with the test labels and the test data and saved as a data-frame "subjecttest". This data-frame contains descriptive activity names and descriptive column names.
5. The training data is merged with the test data into a new data-frame "alldata" containing the merged training and test data and written to the output file "/data/alldata.csv". This data-frame contains descriptive activity names and descriptive column names.
6. The mean and standard deviation for each measurement are extracted to a new data-frame "meanstd" and written to the output file "/data/meanstd.csv". This data-frame contains descriptive activity names and descriptive column names.
7. A second, independent tidy data set (named "average\_alldata") is created with the average of each variable for each activity and each subject and written to the output file "/data/summarized\_data.csv". This data-frame contains descriptive activity names and descriptive column names.

**D. Raw Data Files:**

For a listing and description of the raw data files please refer to the following files in the “data\ UCI\_HAR\_Dataset” folder:

1. README.txt
2. features\_info.txt

The raw data files may also be downloaded from this link:

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

Additional information about the data may be obtained from this link:

<http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

E. Clean Data Files

Three clean data file (listed below) are created when the "run\_analysis.R" script is successfully run. These files are saved in the “data/” folder:

1. **alldata.csv**:
   1. **description**: This dataset contains the merger training and test data and includes descriptive activity names and descriptive column names.
   2. **Size**: 10299 records x 563 columns
2. **meanstd.csv**:
   1. **description**: This dataset contains the extracted measurements on the mean and standard deviation for each measurement and includes descriptive activity names and descriptive column names.
   2. **Size**: 10299 records x 81 columns
3. **summarized\_data.csv:** 
   1. **description**: This data set contains the average of each variable for each activity and each subject and includes descriptive activity names and descriptive column names.
   2. **Size**: 180 records x 563 columns

For each of the three datasets mentioned above, the first column contains the subject id (column name = “subjected”) obtained from the “subject\_train.txt” and “subject\_test.txt” files included with the raw data.

The second column contains the descriptive activity name (column name = “activityname”) obtained from the “activity\_labels.txt” file included with the raw data.

The names for the remaining columns were obtained directly or extracted from the “features.txt” file as described in section C of this document. A description of the descriptive column names can be obtained from the “features\_info.txt” files included with the raw data.