# Code Book

This code book describes variables, data, and any transformations or work performed to clean up the data. These efforts resulted into a clean and tidy dataset - please see SummarisedData.txt

## Data Source

* Original data is taken from the following location: <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>
* Further information on the data set provided by the authors: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

## Dataset Description

* The experiments involved a group of 30 volunteers (19 - 48 years of age)
* Each person performed six activities while wearing a smartphone (Samsung Galaxy S II) on their waist
* The activities consisted of walking flat, walking upstairs, walking downstairs, sitting, standing and laying
* The experiments have been video recorded.
* The obtained dataset was randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

## Measurements Captured

* Using the smartphone’s embedded accelerometer and gyroscope, the following signals and measurements were captured
  + 3-axial linear acceleration
  + 3-axial angular velocity
* 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 cut-off frequency was used.
* From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

For each record in the dataset 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 i.e. walking flat, walking upstairs, walking downstairs, sitting, standing or laying
* An identifier of the subject who carried out the experiment.

# Features

The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ.

* The time domain signals were captured at a constant rate of 50 Hz, prefix 't' denotes time
* The signals were then filtered to remove noise
* The accelaration signal was separated into body and gravity acceleration signals:
  + tBodyAcc-XYZ
  + tGravityAcc-XYZ
* Subsequently, the body linear acceleration and angular velocity were derived in time to obtain Jerk signals:
  + tBodyAccJerk-XYZ
  + tBodyGyroJerk-XYZ
* Next, the magnitude of these three-dimensional signals were calculated using the Euclidean form:
  + tBodyAccMag
  + tGravityAccMag
  + tBodyAccJerkMag
  + tBodyGyroMag
  + tBodyGyroJerkMag
* Finally a Fast Fourier Transform (FFT) was applied to some of these signals, prefix 'f' indicates frequency domain signals:
  + fBodyAcc-XYZ
  + fBodyAccJerk-XYZ
  + fBodyGyro-XYZ
  + fBodyAccJerkMag
  + fBodyGyroMag
  + fBodyGyroJerkMag

That leaves us with all of the signals above collected, '-XYZ' denotes three-axial signals in the X, Y and Z directions, hence 3 captures for each of the signals above

The features were further combined with a variety of estimated variables, such as mean, standard deviation, largest and smallest value in the set etc. This adds up to 561 of indicators in total. The file 'features.txt' lists all of the variables.

## Transformations from Raw to Tidy Data

The following transformations were made in order to tidy the data and make it as useful as possible:

* First the test and training data sets for activity codes, subject number and all the various readings from the sensor signals (accelerometer and gyroscope) were combined. Using rbind() to combine train and test, and then cbind() to combine the 3 separate datasets together. This total combined data set was 10,299 observations of 563 variables.
* Next the dataset was reduced so that only measurements of mean and standard deviation of each of the features was considered. This was done using the grep() function to select out only variable names with mean or std in them. This reduced the data set to 10,299 observations of 69 variables.
* Descriptive activity names were attached to the data in order to make it more readable and sensible to interpret.
* All of the variables created from the signals captured were renamed to make them more intuitive and so the names were more descriptive of the contents. So for all the variables listed above the following replacements were made
  + t = Time - at the start of a variable name
  + f = Frequency - at the start of a variable name
  + Acc = Accelerometer
  + Gyro = Gyroscope
  + Mag = Magnitude
  + mean() = MEAN
  + std() = SD
  + activity = ActivityName
  + code = ActivityNum

so for example we have the following changes to variable names:

* 'tBodyAcc-mean()-X' has been changed to 'TimeBodyAccelerometer -MEAN-X'
* 'tBodyAcc-std()-Z' has been changed to 'TimeBodyAccelerometer -SD-Z'
* tBodyGyroJerk-std()-Y has been changed to ‘TimeBodyGyroscopeJerk-SD-Y’
* ‘tBodyAccJerkMag-std()’ has been changed to ‘TimeBodyAccelerometerJerkMagnitude-SD’
* Etc…
* Final step was to create final tidy dataset with the average of each variable for each activity and each subject. 10299 instances were split into 180 groups (30 subjects and 6 activities) and 66 mean and standard deviation features are averaged for each group.
* The resulting Summarised data table has 180 rows and 69 columns – 33 Mean variables + 33 Standard deviation variables + 1 Subject(1 of of the 30 test subjects) + ActivityName + ActivityNum

## Code to Achieve Tidy Data

All of the above data tidying and transformations are performed by running the script called run\_analysis.R, which:

* Ensures that all non-standard R packages (dplyr etc..) are installed if they are used throughout the code
* Defines a number of helper functions to promote code reuse
* Downloads the original dataset and verifies its content
* Imports activity and label names datasets
* Imports and loads into data tables the training and test datasets and enhances column names with appropriate labels
* Merges the testing and the test datasets using rbind() and cbind()
* Creates an independent tidy dataset based on mean and standard deviations
* Saves the final tidy dataset as SummarisedData.txt so that it can be easily imported and viewed

Further details on how this code works and what is taking place at different stages of the code can be found in README.md