**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 [data.txt](https://github.com/shuerga/GetAndCleanDataOnCoursera/blob/master/data.txt) and [data.csv](https://github.com/shuerga/GetAndCleanDataOnCoursera/blob/master/data.csv).

**Data Source Details**

* Original data: <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>
* Information provided by the authors: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

**Dataset Description**

Full details and explanations can be found in 'README.txt' in the ZIP file containing the original data source.

**Data Partitioning**

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.

**Measurements**

The following sensor signals were captured using the smartphone's embedded accelerometer and gyroscope:

* three-axial linear acceleration
* three-axial angular velocity at a constant rate of 50Hz

The captured signals 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). From each window, a vector of features was obtained by calculating variables from the time and frequency domain. See the section 'Feature Selection' below, also the file 'features\_info.txt' has complete details.

**Feature Selection**

The features selected for this database come from the accelerometer and gyroscope three-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 the following set of signals, suffix '-XYZ' denotes three-axial signals in the X, Y and Z directions:

1. tBodyAcc-XYZ
2. tGravityAcc-XYZ
3. tBodyAccJerk-XYZ
4. tBodyGyro-XYZ
5. tBodyGyroJerk-XYZ
6. tBodyAccMag
7. tGravityAccMag
8. tBodyAccJerkMag
9. tBodyGyroMag
10. tBodyGyroJerkMag
11. fBodyAcc-XYZ
12. fBodyAccJerk-XYZ
13. fBodyGyro-XYZ
14. fBodyAccMag
15. fBodyAccJerkMag
16. fBodyGyroMag
17. fBodyGyroJerkMag

The features were further combined with a variety of estimated variables, such as mean value, standard deviation, largest and smallest value in the set etc. This adds up to over 550 of different indicators in total. The file 'features.txt' lists all of the variables.

**Transformations**

As stipulated by the requirements the following transformations were made to keep the resulting output clean and tidy:

* The training and the test sets have been merged to form a single data set
* Out of the broad spectrum of features only the measurements on the mean and standard deviation were considered
* Descriptive activity names and labels were used appropriately to increase data readability
* A separate tidy dataset was created as the final step of data refinement efforts

In addition to the aforementioned mandatory transformations, the feature names have been amended to adhere to naming standards of the R language. That allowed for an elegant use of R languague features, such as column name filtering, and an increased code readability. I consider this additional transformation a minor change, which I believe doesn't have an adverse impact on the original information.

Feature name adjustment example:

* 'tBodyAcc-mean()-X' has been changed to 'tBodyAcc\_mean\_X'
* 'tBodyAcc-std()-Z' has been changed to 'tBodyAcc\_std\_Z'
* etc.

The transformations are achieved by the script called [run\_analysis.R](https://github.com/shuerga/GetAndCleanDataOnCoursera/blob/master/run_analysis.R), which:

1. Ensures that all non-standard R packages (dplyr, reshape2) are installed
2. Defines a number of helper functions to promote code reuse
3. Downloads the original dataset and verifies its content
4. Loads activity and label names datasets
5. Loads training and test datasets and enhances column names with appropriate labels
6. Merges the testing and the test datasets using dplyr's support for method chaining (pipe operator)
7. Creates an independent tidy dataset based on mean and standard deviations
8. Saves the tidy dataset as [data.txt](https://github.com/shuerga/GetAndCleanDataOnCoursera/blob/master/data.txt) and [data.csv](https://github.com/shuerga/GetAndCleanDataOnCoursera/blob/master/data.csv) so that it can be easily imported and viewed