This is a Code Book for the tidy data set produced in the “Getting and Cleaning Data Course Project”.

# Code Book

This section describes each variable and its unit.

Column 3 – 81 gives the **average** of the “mean”, “meanFreq” or “standard deviation” of each of the measurements.

|  |  |  |  |
| --- | --- | --- | --- |
| Column | Name of Variable | Unit | Description |
| 1 | subjectId | Integer  (range 1-30) | This is the id of the subject who performed the test. Values is from 1-30, as there are 30 subjects |
| 2 | activity | Character  (values are “walking”, “walking\_upstairs”, “walking\_downstairs”, ”sitting”, “standing”, “laying”) | This describes the activities undertaken by each subject. |
| 3 | tBodyAcc-mean()-X | Numeric | This data comes from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise.  Similarly, the acceleration signal was then separated into body and gravity acceleration signals (tBodyAcc-XYZ and tGravityAcc-XYZ) using another low pass Butterworth filter with a corner frequency of 0.3 Hz.  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean value of tBodyAcc-XYZ and tGravityAcc-XYZ. |
| 4 | tBodyAcc-mean()-Y | Numeric |
| 5 | tBodyAcc-mean()-Z | Numeric |
| 6 | tGravityAcc-mean()-X | Numeric |
| 7 | tGravityAcc-mean()-Y | Numeric |
| 8 | tGravityAcc-mean()-Z | Numeric |
| 9 | tBodyAccJerk-mean()-X | Numeric | The body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyAccJerk-XYZ).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean value of tBodyAccJerk-XYZ. |
| 10 | tBodyAccJerk-mean()-Y | Numeric |
| 11 | tBodyAccJerk-mean()-Z | Numeric |
| 12 | tBodyGyro-mean()-X | Numeric | This data comes from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise.  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean value of tBodyGyro-XYZ. |
| 13 | tBodyGyro-mean()-Y | Numeric |
| 14 | tBodyGyro-mean()-Z | Numeric |
| 15 | tBodyGyroJerk-mean()-X | Numeric | The body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyGyroJerk-XYZ).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean value of tBodyGyroJerk-XYZ. |
| 16 | tBodyGyroJerk-mean()-Y | Numeric |
| 17 | tBodyGyroJerk-mean()-Z | Numeric |
| 18 | tBodyAccMag-mean() | Numeric | The magnitude of these three-dimensional signals were calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag)  The values here are the **average** of the mean value of tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag. |
| 19 | tGravityAccMag-mean() | Numeric |
| 20 | tBodyAccJerkMag-mean() | Numeric |
| 21 | tBodyGyroMag-mean() | Numeric |
| 22 | tBodyGyroJerkMag-mean() | Numeric |
| 23 | fBodyAcc-mean()-X | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals producing fBodyAcc-XYZ. (Note the 'f' to indicate frequency domain signals).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean value of fBodyAcc-XYZ. |
| 24 | fBodyAcc-mean()-Y | Numeric |
| 25 | fBodyAcc-mean()-Z | Numeric |
| 26 | fBodyAcc-meanFreq()-X | Numeric | meanFreq(): Weighted average of the frequency components to obtain a mean frequency.  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean frequency value of fBodyAcc-XYZ. |
| 27 | fBodyAcc-meanFreq()-Y | Numeric |
| 28 | fBodyAcc-meanFreq()-Z | Numeric |
| 29 | fBodyAccJerk-mean()-X | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals fBodyAccJerk-XYZ (Note the 'f' to indicate frequency domain signals).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean value of fBodyAccJerk-XYZ. |
| 30 | fBodyAccJerk-mean()-Y | Numeric |
| 31 | fBodyAccJerk-mean()-Z | Numeric |
| 32 | fBodyAccJerk-meanFreq()-X | Numeric | meanFreq(): Weighted average of the frequency components to obtain a mean frequency.  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean frequency value of fBodyAccJerk-XYZ. |
| 33 | fBodyAccJerk-meanFreq()-Y | Numeric |
| 34 | fBodyAccJerk-meanFreq()-Z | Numeric |
| 35 | fBodyGyro-mean()-X | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals fBodyGyro-XYZ. (Note the 'f' to indicate frequency domain signals).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean value of fBodyGyro-XYZ. |
| 36 | fBodyGyro-mean()-Y | Numeric |
| 37 | fBodyGyro-mean()-Z | Numeric |
| 38 | fBodyGyro-meanFreq()-X | Numeric | meanFreq(): Weighted average of the frequency components to obtain a mean frequency.  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the mean frequency value of fBodyGyro-XYZ. |
| 39 | fBodyGyro-meanFreq()-Y | Numeric |
| 40 | fBodyGyro-meanFreq()-Z | Numeric |
| 41 | fBodyAccMag-mean() | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals producing fBodyAccMag (Note the 'f' to indicate frequency domain signals).  The values here are the **average** of the mean value of fBodyAccMag. |
| 42 | fBodyAccMag-meanFreq() | Numeric | meanFreq(): Weighted average of the frequency components to obtain a mean frequency.  The values here are the **average** of the mean frequency value of fBodyAccMag. |
| 43 | fBodyBodyAccJerkMag-mean() | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals producing fBodyAccJerkMag. (Note the 'f' to indicate frequency domain signals).  The values here are the **average** of the mean value of fBodyAccJerkMag. |
| 44 | fBodyBodyAccJerkMag-meanFreq() | Numeric | meanFreq(): Weighted average of the frequency components to obtain a mean frequency.  The values here are the **average** of the mean frequency value of fBodyAccJerkMag. |
| 45 | fBodyBodyGyroMag-mean() | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals producing fBodyGyroMag. (Note the 'f' to indicate frequency domain signals).  The values here are the **average** of the mean value of fBodyGyroMag. |
| 46 | fBodyBodyGyroMag-meanFreq() | Numeric | meanFreq(): Weighted average of the frequency components to obtain a mean frequency.  The values here are the **average** of the mean frequency value of fBodyGyroMag. |
| 47 | fBodyBodyGyroJerkMag-mean() | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals producing, fBodyGyroJerkMag. (Note the 'f' to indicate frequency domain signals).  The values here are the **average** of the mean value of fBodyGyroJerkMag. |
| 48 | fBodyBodyGyroJerkMag-meanFreq() | Numeric | meanFreq(): Weighted average of the frequency components to obtain a mean frequency.  The values here are the **average** of the mean frequency value of fBodyGyroJerkMag. |
| 49 | tBodyAcc-std()-X | Numeric | This data comes from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise.  Similarly, the acceleration signal was then separated into body and gravity acceleration signals (tBodyAcc-XYZ and tGravityAcc-XYZ) using another low pass Butterworth filter with a corner frequency of 0.3 Hz.  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the standard deviation value of tBodyAcc-XYZ and tGravityAcc-XYZ. |
| 50 | tBodyAcc-std()-Y | Numeric |
| 51 | tBodyAcc-std()-Z | Numeric |
| 52 | tGravityAcc-std()-X | Numeric |
| 53 | tGravityAcc-std()-Y | Numeric |
| 54 | tGravityAcc-std()-Z | Numeric |
| 55 | tBodyAccJerk-std()-X | Numeric | The body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyAccJerk-XYZ).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the standard deviation value of tBodyAccJerk-XYZ. |
| 56 | tBodyAccJerk-std()-Y | Numeric |
| 57 | tBodyAccJerk-std()-Z | Numeric |
| 58 | tBodyGyro-std()-X | Numeric | This data comes from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise.  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the standard deviation value of tBodyGyro-XYZ. |
| 59 | tBodyGyro-std()-Y | Numeric |
| 60 | tBodyGyro-std()-Z | Numeric |
| 61 | tBodyGyroJerk-std()-X | Numeric | The body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyGyroJerk-XYZ).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the standard deviation value of tBodyGyroJerk-XYZ. |
| 62 | tBodyGyroJerk-std()-Y | Numeric |
| 63 | tBodyGyroJerk-std()-Z | Numeric |
| 64 | tBodyAccMag-std() | Numeric | The magnitude of these three-dimensional signals were calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag)  The values here are the **average** of the standard deviation value of tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag. |
| 65 | tGravityAccMag-std() | Numeric |
| 66 | tBodyAccJerkMag-std() | Numeric |
| 67 | tBodyGyroMag-std() | Numeric |
| 68 | tBodyGyroJerkMag-std() | Numeric |
| 69 | fBodyAcc-std()-X | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals producing fBodyAcc-XYZ. (Note the 'f' to indicate frequency domain signals).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the standard deviation value of fBodyAcc-XYZ. |
| 70 | fBodyAcc-std()-Y | Numeric |
| 71 | fBodyAcc-std()-Z | Numeric |
| 72 | fBodyAccJerk-std()-X | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals fBodyAccJerk-XYZ (Note the 'f' to indicate frequency domain signals).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the standard deviation value of fBodyAccJerk-XYZ. |
| 73 | fBodyAccJerk-std()-Y | Numeric |
| 74 | fBodyAccJerk-std()-Z | Numeric |
| 75 | fBodyGyro-std()-X | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals fBodyGyro-XYZ. (Note the 'f' to indicate frequency domain signals).  '-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.  The values here are the **average** of the standard deviation value of fBodyGyro-XYZ. |
| 76 | fBodyGyro-std()-Y | Numeric |
| 77 | fBodyGyro-std()-Z | Numeric |
| 78 | fBodyAccMag-std() | Numeric | A Fast Fourier Transform (FFT) was applied to some of the signals producing fBodyAccMag, fBodyAccJertMag, fBodyGyroMag, fBodyGyroJerkMag. (Note the 'f' to indicate frequency domain signals).  The values here are the **average** of the standard deviation value of fBodyAccMag, fBodyActJertMag, fBodyGyroMag, fBodyGyroJerkMag. |
| 79 | fBodyBodyAccJerkMag-std() | Numeric |
| 80 | fBodyBodyGyroMag-std() | Numeric |
| 81 | fBodyBodyGyroJerkMag-std() | Numeric |

# Study Design

This section has information of how I collected the data. It also describes the summary choices I made.

## Raw data provided

The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description is available at the site where the data was obtained:

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

Here are the data for the project:

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

## What are the requirements for data cleaning

The following are the data and analysis we are interested in - requirement is to create one R script called run\_analysis.R that does 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.

## Raw data used

The data is from the zip file given in section 2.1.

The files are extracted from the zip file into the default working directory : getdata-projectfiles-UCI HAR Dataset\UCI HAR Dataset.

It is assumed that the R script - run\_analysis.R, is in directory “UCI HAR Dataset”.

There are many files in the directory “UCI HAR Dataset”. However, after analysis of the data (by reading “readme.txt” and “feature\_info.txt”) and the requirements, the following data files contain the raw data needed:

1. Test Data
   1. UCI HAR Dataset\test\X\_test.txt - Test set
   2. UCI HAR Dataset\test\subject\_test.txt - Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.
   3. UCI HAR Dataset\test\y\_test.txt – Test labels
2. Train Data
   1. UCI HAR Dataset\test\X\_train.txt – Train set
   2. UCI HAR Dataset\test\subject\_train.txt - Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.
   3. UCI HAR Dataset\test\y\_train.txt - Training labels

The following data were also used in the R script:

1. 'features.txt': List of all features
   1. The actual name of the feature was extracted and used from this file in the R script
2. 'activity\_labels.txt': Links the class labels with their activity name.
   1. The actual name of the activity was extracted from this file. As there are only 6 names, a decision was taken to not read this file, and to “hardcode” the names in the R script, when I converted the given Activity Id (numeric) to their description name.

## Summary Choices Made

The analysis asks to extract only the measurements on the mean and standard deviation for each measurement. To do this, I made the assumption that

* all mean will have the word “mean” in the names found in ‘'features.txt'.
* all standard deviation will have the word “std” in the names found in ‘'features.txt'.

As a result of this, the mean values also included “meanFreq()”.

I have chosen to keep the name from “features.txt” exactly, so that it is easy to find this in the original raw data. (I could have simplify the names by removing “-“, “\_” and “()”, but have chosen not to do so).