# Study Design

## The experiment

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

## The data set

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

## The process for the resulting 2 data tables

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.

**Steps taken by run\_analasys.r**

1. Read column name data fie ( features.txt )
2. Read Activity table file ( activity\_labels.txt )
3. Read both the test and Train Data
   1. Read subject column data and clean up its names
      1. For example remove ‘(‘ and ‘)’ characters
   2. Read activity column data
   3. Read data for all the 561 columns
      1. This will also assign each of the columns it name from the previously read column name data file
   4. Remove all NON mean and STD columns
      1. Leaves you 79 columns
   5. Merge Subject Column & Activity Column & Data Columns
      1. Giving you a total of 81 columns
   6. Replace Activity numbers with descriptive Activity names
      1. For example, a 1 in the active column becomes WALKING
   7. Write the resulting table to file

# Code Book

## Data set # 1 returned in data.txt from the run\_analysis. r script.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column / Variable name** | **Column #** | **Data Type** | **Description** |
| Activity | 1 | Factor | Expresses on of the 6 activities  WALKING  WALKING\_UPSTAIRS  WALKING\_DOWNSTAIRS  SITTING  STANDING  LAYING |
| Subject | 2 | Numeric | A number identifying one of the 30 subjects 1 threw 30 |
| tGravity\*  tBody\* | 3 to 81 |  | 79 variables from the original data set that represent the mean and SDT .  The variable names have been cleaned up to remove ‘-‘ , ‘()” from them.  For example, the variable from the original data set  tBodyAcc-mean()-X  was renamed  tBodyAcc.mean.X |

## Data set # 2 returned in data.txt from the run\_analysis. r script.

This data set has average of each variable for each activity and each subject. The ask “average of each variable for each activity and each subject “is a bit ambiguous. As its not asking explicitly for a mean for each combination of activity and subject, I concluded that is must be asking means for activities and subjects independent. A total of 36 means across 78 variables

To represent this in a single table the 70 variables from the first data set are now the observations named rows. And each column Activity Laying, Subject 30 , .. is now the variable. That means we will have 6 columns to represent the activity mean and 30 Colum to represent the Subject means. For a total of 36 columns.

The table snippet below

|  |
| --- |
| A.LAYING.mean A.SITTING.mean …  tBodyAcc.mean.X 0.26864864 0.27305961 …  tBodyAcc.mean.Y -0.01831773 -0.01268957 …  ……………………………………………………………………………………………………………………………. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Column #** | **Data type** | **Description** |
| A.<Activity>.mean | 1 threw 6 | Numeric | Where <Activity> is one of the 6 activities  WALKING  WALKING\_UPSTAIRS  WALKING\_DOWNSTAIRS  SITTING  STANDING  LAYING |
| S.<Subject>.mean | 7 threw 37 | Numeric | Where <Subject> is one of the 30 numbers representing a subject |