**Codebook**

**Project**: Coursera Getting & Cleaning Data Class Project

**Study Design**

**Original data**: As directed for the assignment, the original project for this data consisted of the datasets downloaded from <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

A full description of the data can be found at <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

The original data came from a study of wearable computing. It consisted of accelerometer and gyroscope measurements from the Samsung Galaxy S smartphone, for 30 participants, performing six different activities. For a detailed description of the variables in the original dataset, please refer to the codebook which comes with it when downloaded.

**This study**: For this study, the data was downloaded directly from the internet, and then extracted from the zip file.

To tidy and analyze the data, a computer program was written in the R programming language. The program consists of a single script, which is named “run\_analysis.R”.

**Inputs**: The program uses the following original data files as inputs:

1. X\_test.txt – measurements for the members of the test group.
2. y\_test.txt – identify, by row, which activity was performed for each measurement of the test group, using a numerical code.
3. X\_train.txt – measurements for the members of the training group.
4. y\_train.txt – identify, by row, which activity was performed for each measurement of the training group, using a numerical code.
5. features.txt – names of all measured variables.
6. activity\_labels.txt – one label for each of the six activities, in two columns: the first column contains the numerical code, and the second column contains the corresponding word or phrase.
7. subject\_train.txt – identifies, by row, which measurements belong to which members of the training group.
8. subject\_test.txt – identifies, by row, which measurements belong to which members of the test group.

**Output:** The output of the study is a file named “tidy\_averages.txt”. The file consist of the mean values of all measurement variables from the original data set which contained the strings “mean” or “std” (standard deviation), regardless of case or position within the original variable name. (Recall the original variable names are found in “features.txt”.) These mean values were taken over all measurements of each participant performing each activity.

**Example**: for the subject corresponding to “subjectID” = 1, all “tBodyAcc-mean()-X” measurements while “activity” = “WALKING” were averaged to produce the mean. If the table is read into R using

df <- read.table("tidy\_averages.txt", header = TRUE)

this value will be found in df[1, 3].

**Test and training groups:** Originally the data was organized into two groups, the test group and the training group. The subsetting method used by the program rearranged the rows into numerical order, by “subjectID”. From the original data, the participants of each group were as follows, by “subjectID”:

Test group: 2, 4, 9, 10, 12, 13, 18, 20, 24.

Training group: 1, 3, 5, 6, 7, 8, 11, 14, 15, 16, 17, 19, 21, 22, 23, 25, 26, 27, 28, 29, 30.

**Variable names**: all variable names of all measurements remain the same as in the original data set, due to their technical significance. The word “mean” was not added, as it should be understood that this file consists entirely of the means described above (along with subject and activity identifiers).

**Codebook**

Variable names and units:

1. subjectID – unique numerical identifier for each of the 30 participants in the study. The values range from 1 – 30.
2. activity – the name of each of the six activities (WALKING, WALKING\_UPSTAIRS, SITTING, STANDING, and LAYING).

The names of the remaining 86 variables are the same as in the original data set (see subsection **Variable names** above). Recall that each of these is now the *mean* value of the collection of measurement given by the variable name, for each participant performing each activity. Since calculating a mean does not change the units, all (mean) values have the same units as the original measurements.

For the original, technical meaning of each of these variables, please refer to the Appendix, which is the “features\_info” file from the original data.

1. tBodyAcc-mean()-X
2. tBodyAcc-mean()-Y
3. tBodyAcc-mean()-Z
4. tBodyAcc-std()-X
5. tBodyAcc-std()-Y
6. tBodyAcc-std()-Z
7. tGravityAcc-mean()-X
8. tGravityAcc-mean()-Y
9. tGravityAcc-mean()-Z
10. tGravityAcc-std()-X
11. tGravityAcc-std()-Y
12. tGravityAcc-std()-Z
13. tBodyAccJerk-mean()-X
14. tBodyAccJerk-mean()-Y
15. tBodyAccJerk-mean()-Z
16. tBodyAccJerk-std()-X
17. tBodyAccJerk-std()-Y
18. tBodyAccJerk-std()-Z
19. tBodyGyro-mean()-X
20. tBodyGyro-mean()-Y
21. tBodyGyro-mean()-Z
22. tBodyGyro-std()-X
23. tBodyGyro-std()-Y
24. tBodyGyro-std()-Z
25. tBodyGyroJerk-mean()-X
26. tBodyGyroJerk-mean()-Y
27. tBodyGyroJerk-mean()-Z
28. tBodyGyroJerk-std()-X
29. tBodyGyroJerk-std()-Y
30. tBodyGyroJerk-std()-Z
31. tBodyAccMag-mean()
32. tBodyAccMag-std()
33. tGravityAccMag-mean()
34. tGravityAccMag-std()
35. tBodyAccJerkMag-mean()
36. tBodyAccJerkMag-std()
37. tBodyGyroMag-mean()
38. tBodyGyroMag-std()
39. tBodyGyroJerkMag-mean()
40. tBodyGyroJerkMag-std()
41. fBodyAcc-mean()-X
42. fBodyAcc-mean()-Y
43. fBodyAcc-mean()-Z
44. fBodyAcc-std()-X
45. fBodyAcc-std()-Y
46. fBodyAcc-std()-Z
47. fBodyAcc-meanFreq()-X
48. fBodyAcc-meanFreq()-Y
49. fBodyAcc-meanFreq()-Z
50. fBodyAccJerk-mean()-X
51. fBodyAccJerk-mean()-Y
52. fBodyAccJerk-mean()-Z
53. fBodyAccJerk-std()-X
54. fBodyAccJerk-std()-Y
55. fBodyAccJerk-std()-Z
56. fBodyAccJerk-meanFreq()-X
57. fBodyAccJerk-meanFreq()-Y
58. fBodyAccJerk-meanFreq()-Z
59. fBodyGyro-mean()-X
60. fBodyGyro-mean()-Y
61. fBodyGyro-mean()-Z
62. fBodyGyro-std()-X
63. fBodyGyro-std()-Y
64. fBodyGyro-std()-Z
65. fBodyGyro-meanFreq()-X
66. fBodyGyro-meanFreq()-Y
67. fBodyGyro-meanFreq()-Z
68. fBodyAccMag-mean()
69. fBodyAccMag-std()
70. fBodyAccMag-meanFreq()
71. fBodyBodyAccJerkMag-mean()
72. fBodyBodyAccJerkMag-std()
73. fBodyBodyAccJerkMag-meanFreq()
74. fBodyBodyGyroMag-mean()
75. fBodyBodyGyroMag-std()
76. fBodyBodyGyroMag-meanFreq()
77. fBodyBodyGyroJerkMag-mean()
78. fBodyBodyGyroJerkMag-std()
79. fBodyBodyGyroJerkMag-meanFreq()
80. angle(tBodyAccMean,gravity)
81. angle(tBodyAccJerkMean),gravityMean)
82. angle(tBodyGyroMean,gravityMean)
83. angle(tBodyGyroJerkMean,gravityMean)
84. angle(X,gravityMean)
85. angle(Y,gravityMean)
86. angle(Z,gravityMean)

**Appendix**

Feature Selection

=================

The features selected for this database come 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.

Subsequently, the body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyAccJerk-XYZ and tBodyGyroJerk-XYZ). Also the magnitude of these three-dimensional signals were calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag).

Finally a Fast Fourier Transform (FFT) was applied to some of these signals producing fBodyAcc-XYZ, fBodyAccJerk-XYZ, fBodyGyro-XYZ, fBodyAccJerkMag, fBodyGyroMag, fBodyGyroJerkMag. (Note the 'f' to indicate frequency domain signals).

These signals were used to estimate variables of the feature vector for each pattern:

'-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.

tBodyAcc-XYZ

tGravityAcc-XYZ

tBodyAccJerk-XYZ

tBodyGyro-XYZ

tBodyGyroJerk-XYZ

tBodyAccMag

tGravityAccMag

tBodyAccJerkMag

tBodyGyroMag

tBodyGyroJerkMag

fBodyAcc-XYZ

fBodyAccJerk-XYZ

fBodyGyro-XYZ

fBodyAccMag

fBodyAccJerkMag

fBodyGyroMag

fBodyGyroJerkMag

The set of variables that were estimated from these signals are:

mean(): Mean value

std(): Standard deviation

mad(): Median absolute deviation

max(): Largest value in array

min(): Smallest value in array

sma(): Signal magnitude area

energy(): Energy measure. Sum of the squares divided by the number of values.

iqr(): Interquartile range

entropy(): Signal entropy

arCoeff(): Autorregresion coefficients with Burg order equal to 4

correlation(): correlation coefficient between two signals

maxInds(): index of the frequency component with largest magnitude

meanFreq(): Weighted average of the frequency components to obtain a mean frequency

skewness(): skewness of the frequency domain signal

kurtosis(): kurtosis of the frequency domain signal

bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.

angle(): Angle between to vectors.

Additional vectors obtained by averaging the signals in a signal window sample. These are used on the angle() variable:

gravityMean

tBodyAccMean

tBodyAccJerkMean

tBodyGyroMean

tBodyGyroJerkMean

The complete list of variables of each feature vector is available in 'features.txt'