**Code Book for Samsung Tidy Data Project**

**Introduction**

This project uses the data from the following site:

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

It builds on the code-book available with this dataset and lists only the additional steps taken to tidy the data.

All the files are stored in rjorapu/Samsung/tidy\_project at the following Github link:

https://github.com/rjorapu/Samsung\_tidy\_project.git

Libraries used: dplyr, tidyr, data.table

Datasets used with the definitions:

* File: ~/Features.txt

Documentation available in features\_info.txt

This is read into “features” datasets.

Dataset: features

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| No | Character | 1-561 |
| features | Character | tBodyAcc-mean()-X  tBodyAcc-mean()-Y  ..  angle(Z,gravityMean) |

* File: ~/activity\_labels.txt

This is read into activity\_labels

Dataset: activity\_labels

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Act.No | Character | 1-6 |
| Activity.Description | Character | Walking  Walking\_upstairs  Walking\_downstairs  Sitting  Standing  Laying |

* File: ~/train/X\_train.txt

Documentation available in ~/README.txt

This is read into X\_train

Dataset: X\_train: 7352 obs with 563 variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Act.No | Integer | 1-6 |
| Subject | Integer | 1-30 |
| Rest of 561 variables are those in features dataset | Character | tBodyAcc-mean()-X  tBodyAcc-mean()-Y  ..  angle(Z,gravityMean) |

* File: ~/train/subject\_train.txt

Documentation available in ~/README.txt

This is read into subject\_train

Dataset: subject\_train: 7352 obs with 1 variable

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Subject | Integer | 1-30 |

* File: ~/train/y\_train.txt

Documentation available in ~/README.txt

This is read into y\_train

Dataset: subject\_train: 7352 obs with 1 variable

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Act.No | Integer | 1-6 |

* File: ~/test/X\_test.txt

Documentation available in ~/README.txt

This is read into X\_test

Dataset: X\_test: 2947 obs with 563 variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Act.No | Integer | 1-6 |
| Subject | Integer | 1-30 |
| Rest of 561 variables are those in features dataset | Character | tBodyAcc-mean()-X  tBodyAcc-mean()-Y  ..  angle(Z,gravityMean) |

* File: ~/test/subject\_test.txt

Documentation available in ~/README.txt

This is read into subject\_test

Dataset: subject\_train: 2947 obs with 1 variable

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Subject | Integer | 1-30 |

* File: ~/test/y\_test.txt

Documentation available in ~/README.txt

This is read into y\_test

Dataset: subject\_train: 2947 obs with 1 variable

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Act.No | Integer | 1-6 |

* Dataset: X\_total:

This combines the two datasets: X\_test and X\_train

It has 10299 obs with 563 variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Act.No | Integer | 1-6 |
| Subject | Integer | 1-30 |
| Rest of 561 variables are those in features dataset | Character | tBodyAcc-mean()-X  tBodyAcc-mean()-Y  ..  angle(Z,gravityMean) |

* means: Integer vector

means contains the indices of all observations from features dataset where the string “mean” appears. This is used an column index to select only those variables corresponding to the mean.

This is derived using the grep as follows:

grep("mean", features$features) -> means

means is an integer vector with length 46 and values like means[1,2,3,41,42,43….]

* std: Integer vector

std contains the indices of all observations from features dataset where the string “std” appears. This is derived using the grep as follows:

grep("std", features$features) -> std

std is an integer vector with length 33 and values like

std[4,5,6, 44,45,46….]

* reqd\_features

this is used to select only means and std variables. It is an integer vector combining the above two vectors – means and std. It has a length of 79 (46 + 33).

However, the indices are incremented by 2 because 2 columns are inserted in X\_total (Act.No and Subject). Now, the new indices represent the correct column index in X\_total and so is used to select only required columns

* Dataset: X\_reqd:

This is a subset of X\_total with only those columns with indices in the vector reqd\_features

It has 10299 obs with 81 variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Value / Range |
| Act.No | Integer | 1-6 |
| Subject | Integer | 1-30 |
| Rest of 81 variables are those corresponding to mean and std columns from X\_total | Character | tBodyAcc-mean()-X  tBodyAcc-mean()-Y  ..  tBodyAcc-std()-X  tBodyAcc-std()-Y |

..

* Dataset: gr\_data:

This is a grouped dataset derived from X-reqd grouped by Act.No and Subject. It has 10299 obs with 81 variables

* Dataset: summary\_data:

This is a summarised dataset derived from gr\_data with means of all the variables of the grouped data in gr\_data. This is achieved by the function summarise\_each as follows:

summarise\_each(gr\_data, funs(mean)) -> summary\_data

It has 180 obs with 81 variables.

* Dataset: tidy\_data:

This is the tidy data set obtained by collapsing the 81 variables in summary\_data into 4 variables.

tidy\_data has 14220 obs and 4 variables – Act.No, Subject, parameter and value.

* Dataset: tidy\_final:

This is the final version of the tidy dataset with description of the activity instead of the numerical description. It has 14220 obs with 4 variables, namely Activity.Description, Subject, parameter and value.

* File: tidy\_final.txt:

This is the output of the tidy dataset written in the working directory. It has 4 columns, namely Activity.Description, Subject, parameter and value and has 14220 observations(rows).