Code Book
Course 3 Getting and Cleaning Data Final Project: run_analysis.R

DATASET LOCATION

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

DATASET LICENSE

Use of the dataset in publications must be acknowledged by referencing the following publication [1]

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Vitoria-Gasteiz, Spain. Dec 2012

ACTIVITIES (from activity_labels.txt, provided in the dataset)

1 = walking
2 = walking_upstairs
3 = walking_downstairs
4 = sitting
5 = standing
6 = laying

SUBJECTS

Thirty volunteers participated in the experiments and were provided an index from 1 to 30. subject_test.txt and subject_train.txt provided these codes for each data observation.

FEATURES

561 features are indexed in features_info.txt. These are the variables used in the analysis.

COLUMN HEADINGS
The first column is named “subject”, the second heading is named “activity”, and the remaining 561 heading names come from features.txt.

MERGED DATA
features_train and features_test are dataframes containing the data from the features files. activities_train and activities_test are character vectors containing the activity names corresponding to each observation. These files are merged with features_train and features_test using cbind to create features_activities_train and features_activities_test.

subjects_train and subjects_test are character vectors containing the subject codes corresponding to each observation. These files are merged with features_activities_train and features_activities_test using cbind to create subjects_features_activities_train and subjects_features_activities_test.

subjects_features_activities_train and subjects_features_activities_test are then combined using rbind to form data_train_test. This contains test and train data and includes features, activities, and subjects.

REMOVING DUPLICATE COLUMNS
The ‘duplicated’ function is used to remove duplicate columns from data_train_test. The resulting dataframe is uniquecols_data_train_test.

SELECTING MEAN AND STD VARIABLES
Only mean and standard deviation measurements are extracted from uniquecols_data_train_test using the ‘select’ function. The resulting dataframe is mean_std_data.

AVERAGING EACH VARIABLE FOR EACH ACTIVITY AND EACH SUBJECT
The data in mean_std_data are grouped by subject and activity to create the dataframe grouped_mean_std_data. The ‘aggregate’ function is then used on this dataframe to

calculate the mean of each variable indexed by subject and activity. The resulting dataframe is mean_grouped_mean_std_data. “activity” and “subject” column names are added back to this dataframe.

OUTPUT

The output dataframe is mean_grouped_mean_std_data.