run_analysis() Code Book

## Summary

run_analysis() is an R program written for the final course project for the "Getting and Cleaning Data" course of the Data Science Specialization associated with Johns Hopkins University on Coursera. The program takes a series of .txt files, merges them, adds labels to each column, and then creates a clean, tidy data set.

## Files

There are 8 files that were used in creating a single tidy data set:

1. "activity_labels.txt"
   This file contains a list of the 6 activities tested in the data set ("X_train.txt" and "y_train.txt"): walking, walking_upstairs, walking_downstairs, sitting, standing, laying. Each activity has a numerical label from 1 to 6.
2. "features.txt"
   This file lists the variables that were tested in the data set. 30 participants were asked to perform the 6 activities listed in the "activity_labels.txt" file. Each participant wore a smart watch equipped with accelerometers and gyroscopes to collect acceleration. The "features.txt" is a list of the different measurements that were collected across the accelerometers and gyroscopes. These are essentially the column names for the data set.
3. "subject_train.txt"
   This file lists the participants used in the training data set. 21 of 30 participants were used for the training data set ("X_train.txt").
4. "subject_test.txt"
   This file lists the 9 participants used in the test data set ("X_test.txt").
5. "y_train.txt"
   This file is a single column data file that lists the activity for each record in the training data set ("X_train.txt").
6. "y_test.txt"
   This file is a single column data file that lists the activity for each record in the training data set ("X_test.txt").
7. "X_train.txt"
   The training data set
8. "X_test.txt"
   The test data set

## What the Code Does:

1. Sets the working directory to the file path of where the data is saved on my computer
2. Reads in the relevant files:
   a. Reads the file “activity_labels.txt” into R and transforms it into a data table called “activity_label” using the dplyr package

    b. Reads the file “features.txt” into an R data frame called “features_label”
    c. Reads the file “subject_train.txt” into R and transforms it into a data table called “train_subject” using the dplyr package
    d. Reads the file “y_train.txt” into R and transforms it into a data table called “train_activities” using the dplyr package
    e. Reads the file “X_train.txt” into R and transforms it into a data table called “train_data” using the dplyr package
    f. Reads the file “subject_test.txt” into R and transforms it into a data table called “test_subject” using the dplyr package
    g. Reads the file “y_test.txt” into R and transforms it into a data table called “test_activities” using the dplyr package
    h. Reads the file “X_test.txt” into R and transforms it into a data table called “test_data” using the dplyr package

3. Combines the test and training data into single tables using rbind:
    a. "train_subject" is combined with "test_subject" to create a data table called "subjects",
    b. "train_activities" is combined with "test_activities" to create a data table called "activities"
    c. "train_data" is combined with "test_data" to create a data table called "full_data"
4. Labels the column headers for “subjects”, “activities”, and “full_data”
    a. The one column in “subjects” is renamed “subject”
    b. The numeric labels (1-6) for “activities” are replaced with their descriptive names by performing an inner join of “activities” with the 6 x 2 data table called “activity_labels”. The first column of numeric labels are removed from “activities” so all that remains are descriptive labels for each activity. This single column is renamed “activity”
    c. The “full_data” data table column names are changed to the labels listed in “features_label”. The column names in “features_label” are actually in the second column, so the second column is pulled and transformed into a character class using as.character() and assigned to a variable called “data_names”. “data_names” now lists the column names for “full_data”. Using “data_names” the column names of “full_data” are replaced.
5. “subjects”, “activities” and “full_data” are combined into 1 data table called “complete_data” using cbind.
6. Next, only the columns in “complete_data” that are averages or standard deviations are pulled. “complete_data” currently has 563 columns, which will be reduced to 68 with the following steps:
    a. Using grepl, a true/false vector called “select_cols” is created by selecting the column names from “complete_data” that contain “mean//(“ or “std//(“. There are 66 of these columns
    b. Next, a new variable called “meanstddata” is created by selecting the columns 1 and 2 of “complete_data”, which contain the participant and activity information, followed by the 66 mean and std columns.

7. Next, a tidy data set is created from “meanstddata” by using the group_by() and summarize_each() functions from the dplyr package
    a. “meanstddata” is grouped by participant (of which there are 30) and activity (each participant performed 6 activities”; there should be 180 grouped records
    b. Next, the measurements (the 66 columns of mean and standard deviation data) are averaged for each combination of participant and activity using summarize_each(). This creates a data table called “tidy_data” which contains 180 records and 68 columns (partipant, activity, and the 66 measurements)
8. Finally “tidy_data” is written to a .txt file called “Course Project Tidy Data Set.txt”

**Variable Definitions:**

- “activity_label”: data table created by reading “activity_labels.txt”
- “features_label”: data frame created by reading “features.txt”
- “train_subject”: data table created by reading “subject_train.txt”
- “train_activities”: data table created by reading “y_train.txt”
- “train_data”: data table created by reading “X_train.txt”
- “test_subject”: data table created by reading “subject_test.txt”
- “test_activities”: data table created by reading “y_test.txt”
- “test_data”: data table created by reading “X_test.txt”
- “subjects” is a data table created by combining "train_subject" with "test_subject"
- “activities” is a data table created by combining "train_activities " with "test_activities”
- “full_data” is a data table created by combining “train_data" with "test_data"
- “data_names” is a vector created by pulling the column names listed in the second column of “features_label”; it is used to rename the column names of “full_data”
- “complete_data” is a data table created by combining “subjects”, “activities” and “full data” data tables
- “cnames” is a vector listing the column names from “full_data”
- “select_cols” is a true/false vector where true means that the column name contains either “mean(“ or “std(“
- “meanstddata” is a data table created by taking “complete_data” and removing the columns listed as false in “select_cols”
- “groupeddata” is created by grouping “meanstddata” by participant and activity
- “tidy_data” is created by taking the average of each measurement for each of the records for each grouping