**Instructions**

The purpose of this project is to demonstrate your ability to collect, work with, and clean a data set.

Review criterialess

1. The submitted data set is tidy.
2. The Github repo contains the required scripts.
3. GitHub contains a code book that modifies and updates the available codebooks with the data to indicate all the variables and summaries calculated, along with units, and any other relevant information.
4. The README that explains the analysis files is clear and understandable.
5. The work submitted for this project is the work of the student who submitted it.

Getting and Cleaning Data Course Projectless

The purpose of this project is to demonstrate your ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis. You will be graded by your peers on a series of yes/no questions related to the project. You will be required to submit: 1) a tidy data set as described below, 2) a link to a Github repository with your script for performing the analysis, and 3) a code book that describes the variables, the data, and any transformations or work that you performed to clean up the data called CodeBook.md. You should also include a README.md in the repo with your scripts. This repo explains how all of the scripts work and how they are connected.

One of the most exciting areas in all of data science right now is wearable computing - see for example [this article](http://www.insideactivitytracking.com/data-science-activity-tracking-and-the-battle-for-the-worlds-top-sports-brand/) . Companies like Fitbit, Nike, and Jawbone Up are racing to develop the most advanced algorithms to attract new users. The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description is available at the site where the data was obtained:

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

Here are the data for the project:

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

You should create one R script called run\_analysis.R that does the following.

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. Subset columns so that mean and standard deviation are the only ones left
3. Uses descriptive activity names to name the activities in the data set - mutate
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.

Good luck!

* **Do we need the inertial folder**

Short answer- no. Long answer- If you work very hard you can attach this very raw data to the more processed train and test X data, but a subsequent steps calls on you to get rid of all the variables that are not to do with mean or standard deviation (worked out from the column names- the features) and you have no names for those columns so they go. Seems a lot easier just to not include them in the first place.

https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-the-assignment/

* **what columns are measurements on the mean and standard deviation**

Based on interpreting column names in the features is an open question as to is the the entries that include *mean()* and *std()* at the end, or does it include entries with *mean* in an earlier part of the name as well. There are no specific marking criteria on the number of columns. It is up to you to make a decision and explain what you did to the data. Make it easy for people to give you marks by explaining your reasoning.

* **Is descriptive activity names things like “Walking” and “Walking Up”**

Yes, you need to get the activity numbers in the data and replace them with descriptive terms which are words. Now, a lot of people watch the merge lecture in week three, and decide to use it in a “I am rushing through without checking each step kind of way” and fail to notice merge reorders the data (this is something you also might see in the week 3 quiz if you are paying attention). If you have not yet clipped all your data together (because you are doing the steps out of order) this will mean the things you are clipping together will be in a different order. If you are doing the steps in the set out order, you will never see this problem though. You can also apply the labels with subsetting. Another strategy is adjusting factor levels, or other even more exotic techniques.

* **Is step 4 the same as step 3?**

I am going to say **no**

Since it say *labels the data set* it is talking about the variable names (which at the moment are V1, V2, etc if you have be following the steps in there numbered order. Descriptive variable names means names based on the action the variable is recording, for example the Jerk of the body on the z axis of the phone. In general, the more descriptive is going to be the better. once again make it easy for your markers by noting why they are descriptive names (what they mean goes in the code book).

* **Is an average of a standard deviation even a thing?**

Short answer. Doesn’t matter, you are being asked to produce a average for each combination of subject, activity, and variable as a sign you can manipulate the data. Long answer, yes it is a thing.

* **Is it the first data set or the second we upload?**

Up load the set created in step 5. This is an independent set of data created by taking the results of step 4 and making a new set of averaged data. Step 4 is internal to the script only.

* **Are you sure either the wide or narrow form of the data is tidy?**

This question only makes sense when you have watched the reshaping lectures in week 3.

Yes. The wide or narrow form is tidy. The wording in the rubric has actually been clarified on this point to help people be clear in marking. Tidy data is one of the more important concepts in the is course. Go back and read Hadley Wickham’s Tidy Data paper (not that you have read this at the start of the course, but it is recommended reading in the lectures in both week one and week three before the assignment is due). It talks about how the specific form of “tidy” depends on the problem being solved, and this problem permits two forms. I have my own elegant proof of this which the margins do not have room for. If you want to bulletproof yourself on this point for the assignment, I am also going to start a similarly verbose thread about tidy data in the next few days. And that gives you something you can assert in your readMe citing the weblink to the discussion (hardly anyone cites things, but it is actually a brilliant plan). That, people, is how you make sure there is no ambiguity for your markers

* **should the saved text file look that weird?**

Tidy data is not made to be looked neatly at in programs like notepad (which is often the default for text files on windows), but if you saved the file with write.table according to the instructions, the command for reading it in and looking at it in R would be

data <- read.table(file\_path, header = TRUE) #if they used some other way of saving the file than a default write.table, this step will be different

View(data)

A person wanting to make life easy for their marker would give the code for reading the file back into R in the readMe. A person who varied the write.table settings should definitely help their marker by giving the variant instructions for reading the file in.

* **wasn’t there a Code Book?**

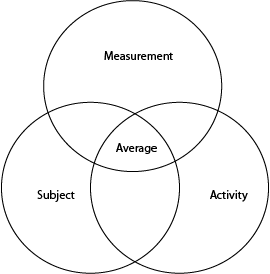
Yes, and it is really important you include it. Go back to quiz 1 and look at the codebook there for inspiration. Be sure it goes on github with the analysis script and the readme. People have lost major marks in previous sessions by having a brain-fade and forgetting about critical files so not getting the marks: you want a run\_analysis R script, a ReadMe markdown document, a Codebook markdown document, and a tidy data text file (this last goes on Coursera).

* **Is anyone else seeing the y or subject data file as gibberish**

Use Notepad++ as a high quality free Windows text editor.

Now, how I think about the data:

1 – I would say it is better to think of the problem as involving three sets of entities – subjects, activities, and readings and we are trying to get tidy data that represents a summary value (mean) for the intersection of each set. But this is because I think set theory (basically drawing venn diagrams about the entities involved) is a very useful lens for how to think about the tidiness of a set of data.

[](https://thoughtfulbloke.files.wordpress.com/2015/09/320.png)

In fundamental set theory speak, we are looking at the intersection of the three sets Subject ∩ Activity ∩ Measurement

2 – For this assignment we are only using the features involving the standard deviation and the mean as a subset of all the available features. imagine a small circle in the centre of the above three circles intersection.

3 – I would say they are discrete members of the set of observations, as it is possible for an action to change a y direction reading in the phone without changing a x or z direction reading (though practically that would be really hard).

**A brief example**

Let’s take the mtcars data

data(mtcars)

mtcars

Looking at it, I think we would all describe that as wide data.

Now let’s restrict the number of variables

narrow <- mtcars[, c("cyl", "gear", "vs", "mpg")]

narrow

Suddenly it looks a lot more like long (or narrow) data.

Now let’s say we are wanting to summarize the data, finding the maximum value of mpg for each combination of cyl, gear, and vs (there are, as you will know from the week 3 lectures, a lot of ways you could do this, I happen to like reshape2)

library(reshape2)

wide1 <- tidied <- dcast(narrow, cyl + gear ~ vs, max)

wide1

Putting aside that when there isn’t an entry max goes to -Inf and we should probably fix that, replacing with NA, this is wider than the original.

But if we go

wide2 <- tidied <- dcast(narrow, cyl + vs ~ gear, max)

wide2

That is the same data, arranged differently.

There is also

wide3 <- tidied <- dcast(narrow, gear + vs ~ cyl, max)

wide3

Same data, arranged differently.

For that matter, there is also

notverywide <- aggregate(mpg ~ gear + vs + cyl, data = narrow, max)

notverywide

A visual difference is that aggregate removed the -Inf entries, but for this the important issue is that it is the same narrow shape as the original.

Now, lets thrown in a clearly untidy result (though up until now we have only been talking long and wide rather than tidy)

untidy <- with(narrow, tapply(mpg, list(cyl, vs, gear), max))

untidy

Which I hope we can agree that dividing up the answers across multiple tables is not what we want (it may be useful in some cases, and the answers are correct, but the outcome is not one single tidy table for output).

So what’s going on with the long/wide thing? This all falls back ultimately to set theory. In this case we have a set of (cyl) a set of (vs) and a set of (gear). Imagine a cubic stack of boxes where the first dimension (width) is arranged so all the same cyls are at the same place going across, the second dimension (height) has all the same vs boxes at the same height, and the third dimension (depth) has all the same gears arranged at the same depth as we walk down the side of the cubic stack. Inside the boxes are the mpg readings for that combination of cyl, vs, and gear.

When we reshape this data, we can lay all the boxes out in a long line recording where they came from in the cubic static (so recording the cyl, vs, and gear) with the mpg still in the box. This is the long form.

In the wide forms above, we are laying out the data in a two dimensional arrangement, where one of the variables is the width and combinations of the other two are the height (With purely numeric data we could actually make a 3 dimensional matrix as well, but that is a little beyond the scope of this discussion).

There is a follow on to this- when we are reshaping our data we are changing the details of what our observations are (what the rows are). This is to do with where in Hadley Wickham’s paper he talked about some of the issues with tidy data being problem specific (to do with the question we are trying to answer). Some arrangements of observations are going to be better at address a particular problem than others. So within the broad set theory background it becomes what is the best tidy form of data to answer a specific question