Here is my compendium of useful information about the course Project.

I am posting this now as some people are asking assignment questions in the first week (but I want to say, take the time to consider the content in the course in depth).

I am writing this as a CTA, a former student returning as a volunteer. And while I am not teaching staff, since when I the first test run of the course I have CTA'd the four public runs, so have seen the points where people get confused (I have also scored full marks on the assignment each time). So this is my advice. It includes every question I can recall being asked over the various sessions.

A favour to ask though, while I am quite happy to facilitate discussion about any and all these points, could people either join other existing threads on particular points or (if there isn't a thread) start a new one- if all the discussion takes place in this thread in will get unimaginably long and unweildy, so it would be really great if this particular thread could be for things I missed or links to other threads discussing items in more depth.

* **Where do I begin?**

For many people this is the first time you've encountered an open assignment, where parts of it are up to you. Think of it as getting an early chance to exercise some professional judgement under controlled conditions. Or, if you prefer, think of it more like writing an essay at University level- your choice of references is up to you in writing about a topic. In a similar way, your choice of steps is up to you to address the analysis. A bit like it being a good idea to plan out an essay, it is a good idea to plan out the analysis: break the requirements down to steps, work out what to do for each step, do that step (and check it worked). Just like doing an essay, a bit of background reading can be a good idea. The data file includes a readMe, it wouldn't be a bad plan to read it and have a think about how the data matches to it.

Take it step by step, get each step sorted before moving on.

* **The explanation is as important as the script, so make sure you make a ReadMe**

Some people have lost marks in previous courses for not making it easy for their reviewers to give them marks. Don't just make a tidy data set, make it clear to people reviewing it why it is tidy. When you given the variables descriptive names, explain why the names are descriptive. Don't give your reviewers the opportunity to be confused about your work, spell it out to them.

* **How do I put the data together**

If you are clipping data together, you often want a clipping data together function like rbind() or cbind(). To work out how various parts (x files, y files, subjects, train vs test) flow together try reading in the files and looking at their dimensions with a command like dim(), by matching sizes like lego bricks there really are a very limited number of ways the data can fit together even if you haven't read the ReadMe.

* **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.

* **can I have multiple scripts**

Though most people keep everything in one script, people have, in the past created separate scripts and used source() commands to bring it all together. If you are varying from the obvious spec it is really, really important to make it clear how what you are doing is within the brief for the assignment, so explaining how it is run\_analysis that is doing the work and what the various parts are contributing.

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

Based on 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 critieria 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 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. If you have 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 activity names means names based on the action the variable is recording, for example the Jerk of the body on the z axiz 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?**

I am going to say I have always uploaded the second (and have always gotten full marks). My reason for asserting this is the set created in step 5 is the only one explicitly described in the instructions as tidy. Once again, there is a request in for official rewording of this point. That said, while I have always put in the second set, I have always marked generously providing the data is actually tidy.

* **Is the wide or narrow form of the data tidy?**

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

Yes. The wide or narrow form is tidy. Go back and read Hadley Wickham's Tidy Data paper. I have my own elegant proof of this which the margins do not have room for. If you want to builtproof yourself on this point for the assignment, I suggest someone start a "Tidy data and the assignment" thread. We talk it out there. And that gives you something you can assert in your readMe citing the weblink to the discussion. That, people, is how you make sure there is no ambiguity for your markers.

* **what form should the text file be in?**

Of appropriate formats for this assignment and all the formats that Coursera will accept, a good intersection of the two is a txt file made with write.table. I sincerely hope by the end of week 3 anyone marking your work (and thus who has done the assignment) is capable of dealing with a text file. Particularly if you make it easy for your markers by explaining how to read it into R (for instance using read.table). Protip, if using write.table, set the option of not writing the line numbers as that can look a bit ugly when reading it back in. As you were very warned, **do not copy and paste data into the text box**, this would likely stop your submission working properly, resulting in no marks at all.

* **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)

* **And Submit**

This isn't a question, but there are two buttons at the bottom of the assignment "Save draft" and "Submit for grading". You have not submitted your assignment until you tick the honor code tickbox and "Submit for grading". If you do not actually do this step you will miss out on all the marks for the assignment.

* **Have I ruined my data?**

There are ways you can put the data together where it all goes wrong. All subjects should have done all activities. That is a pretty good rule of thumb.

* \*\*Why do I keep saying "make it easy for your markers"?

Because it is one of the basic principles of doing peer assignments for any course (not just this one). Most people want to give you marks, and if you actually explain about why your project meets the requirements they will know you have thought about it and be happy to give you marks (even if you did the project differently to them).