FAQ-Course-Project

David Hood

February 4, 2015

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 year 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 all the main question I can recall being asked over the various sessions.

I want to make it clear though, this is not a guide to what you need to do to do the assignment, this is a guide to to doing the assignment in a way to get as reliable a mark as possible when graded by peers- you could meet the rubrics with less, but you are also leaving more to the judgement of your peers (and thus gambling on that judgement).

But, as well as being a guide to this assignment, it should also be seen as a guide to how to approach assignments on Coursera for any paper, and the golden rule is "Make it Easy for Your Markers". Your markers will give your points if you make it clear how the assignment meets the criteria, but this signposting and spelling out how an assignment fits the rubric is a learned skill. It takes people years of high school to learn to write University quality essays, so there is some reflective learning to be had to get good at this. What we are discussing in here is the gold-plated plan to make sure your markers know you have done the right thing, when it comes to doing marking, you may well encounter more ambiguous assignments.

## 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 (p.s. read their 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 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 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 suggest someone start a "Tidy data and the assignment" thread. We talk it out there (at least once the tidy data question in quiz 1 reaches hard deadline, we can't actually give away the answer to that). 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. A person careful about ethically noting their sources might cite this thread (though that is a pretty minor bit of code). There are fancier ways of reading the file in, like from a web address on the clipboard, but that is probably a topic for a stand lone thread if people were really curious, rather than extending this thread.

## 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).

## Feel free to start a thread about what a good codebook is.

Is anyone else seeing the y data file as gibberish You must be on Windows using notepad. Notepad is not a very clever text editor, and can only cope with basic Windows text encodings. Use Notepad++ as a high quality free Windows text editor (if you want a good free Macintosh one I would suggest TextWrangler)

## Should I decompose the variable names

No. For two reasons. One is that no-one ever does so correctly. The other is that you need to write a really excellent ReadMe and Codebook that makes it clear to your markers how what you've done is tidy, and for reasons of the first part this is a problem. This is one of those ideas that is better in theory than in practice. People (possibly inspired by the tidyr swirl tutorial) go "I can split the x,y, and z, and all the others into different columns". The trouble in practice is that you don't actually get clear one variable per column because the the entries in each column are not independent, mutually exclusive, members of the same set. It is like seeing red, dark green, light green, pink, and blue as categories and thinking it is a good idea to make it tidier by putting the light and dark in a separate column. You introduce a bunch of NA values for all the other entries, and introducing a bunch of NA values where there were not previous ones (or a functional equivalent term like "other") is a pretty clear sign the data is a best no tidier (and is probably worse).

## 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. If you want to share an aggregate mean value for a particular subject/activity/measurement combination, go right ahead. But the all subjects did all activities normally covers potential problems.

## 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).

## Can I vigorously defend what I believe to be the one truly tidy arrangement of the data

Go right ahead. But please keep it to the discussion forums rather than carrying it into how you are marking (and please start another thread for that). This can be one of those times where people overestimate their understanding of the structure of the data if you are certain of your opinion. I've been helping out in this course for nearly a year and I think it is arguable. What we are marking for is the data tidy- nicely categorized data so we can do an analysis easily- has it got headings that make it clear which column is which, is it one column per variable (though what the variables are will be different in the long and wide form), is it one row per observation (though what the observations are will be different in the long and wide form). Tidy data is not a certain number of rows and columns, if the question was looking for that, the rubric would specific it (that said, for example if their data left out entries because they didn't include the test data, the data set might be tidy but the script itself would have a serious flaw).

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.