thoughtfulbloke aka David Hood

Getting and Cleaning Data, General Advice

Having been a long term (year and a half or so) CTA of the Getting and Cleaning Data course on Coursera, I've seen pretty much all the problems. So this, together with the week specific articles and the assignment ones, is my collected advice.

- Part 1 : General Advice (https://thoughtfulbloke.wordpress.com/2015/08/31/hello-world/)
- Part 2 : Getting and Cleaning Week 1 (https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-week-1/)
- Part 3 : Getting and Cleaning Week 2 (https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-week-2/)
- Part 4 : Getting and Cleaning Week 3 (https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-week-3/)
- Part 5 : Getting and Cleaning Week 4 (https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-week-4/)
- Part 6 : Getting and Cleaning the Assignment (https://thoughtfulbloke.wordpress.com/2015/09/09/getting-and-cleaning-the-assignment/)

There is nothing that breaches the honour code, and I am freezing the advice as of September 2015 (so there may be additional changes in R or the course) but here it is:

Troubleshooting Quiz answers

There are some general bit of advice for the quizzes that will take you a long way toward figuring out the correct answer if things go awry.

Downloading secure files

[**Update, May 6th 2016**: The release of R 3.3.0 solves most of the issues for download file and https connections (if you have the appropriate certificates on your machine) as it has cross platform default support for https. There are still binary problems, and I will keep the https advice in this document for legacy systems.]

The big tip here is upgrade to R 3.2.x where R introduced method="libcurl" that is a cross platform way of downloading. So in lectures where Jeff uses method="curl" to get secure files, everyone can use this option.

Otherwise, for older versions, because it goes through different operating systems you need different setting that work with that particular operating system. If you are using a Mac, or a Windows or Linux machine with curl installed and correctly configured so that the system can call it, then you can just follow along lecture code. But if you are using a Windows machine or a Linux machine without curl you will need to customise the code.

Windows: use the default method="internal" and precede the download.file command with

```
setInternet2(use = TRUE)
```

Linux: use method="wget"

As an alterntive for any OS, if the download is available as a non-secure download (as the cloudfront stored files are) you can just use download.file and turn the https into http.

Another general way of downloading secure files is to use the downloader package. In reality this just sets the above options to the kind of machine until it finds one that works, but it does mean identical code will run on different kinds of machines.

Downloading binary files

A lot of files you download for this course are text files (.txt, .csv, .for, .xml, .html) others are binary files (.xlsx, .pdf, .jpg). To successfully download a binary file you need to set **mode="wb"** in the download.file settings.

Checking your work

Sometimes you will do things that go wrong- you put a setting in the wrong function, you accidentally convert your data in a way you weren't expecting because you missed putting in an import setting. Here is a good plan for checking what is happening: After each step you do on the data run the commands

str(yourdata) to check both the dimensions of the data and the format of it.

View(yourdata) to see its current contents in a new RStudio tab (if it was a big data set I suggest just checking head() and tail().

By checking at each stage you can see if the data changes unexpected. For example, if you do as.numeric() on a factor variable and get different numbers to as.numeric(as.character()), chekcing with View() would show the unexpected change.

Pay attention to your folders

As people are not yet used to downloading files with r and saving them places, be sure to check what folder you are based in- what your working directory is, and make sure it is set to one that you can save things into. If you are trying to save things into a "data" subdirectory, you will need a data subdirectory to save into.

General studying advice

This is one of those classes of the kind where, if something interests you, you search for more information on the topic (this relates to the Hacker attitude discussed in the Toolbox prerequisite course).

In that spirit, all the course materials for all the data science courses are on GitHub, and could be found with a search like

```
    Coursera GitHub data science
```

for those who haven't had a lot of experience with Coursera, the discussion forums are a bit like a university study group or even a book club- a place where people are working through a text until they understand it. In this particular course, I have started some specific threads for the quiz questions to concentrate discussion but I encourage people to join in (or start new threads where one on a topic doesn't exist) as this is the place for discussing things until they are clear.

One of the jobs of the CTAs is to make sure the honour code is upheld, so we will be removing specific full solutions posted for quiz questions. But I want to make it clear, it is fine and helpful to point people in the direction they should be going. Explaining things is a great way to be sure you understand it.

If you are interested in a particular kind of data, now (in the sense of when it is covered in the lectures "now") is a golden opportunity to have a go at reading it into R, and having the forums available for help if you have problems trying it out.

One more small tip. If you are the sort of person who likes to calendar things, on the announcements page is a little calendar symbol toward the top right by the heading "Upcoming Deadlines"- that is a subscribable calendar with the course dates.

For getting help effectively in the discussion forums (and this applies to pretty much any coursera course) keep the following in mind:

It can be a big time saver to use the search box at the top of the main Discussion Forum area. If you are getting stuck installing rJava on Ubuntu, it might be an idea to do a search for rJava Ubuntu before posting your own question, to see if there is already collective wisdom to take advantage of.

If you are asking (or answering, being helpful is awesome) keep in mind the forum posting guidelines at the top of every forum text box. A couple of the key points are when asking questions about a problem you are having give lots of context information so it is easier for people to know how to help you, and when asking or answering questions don't post the answers (or full copy and paste code that gives the answers). A common way people slip up here is by posting all of their question code and asking what am I doing wrong, and once the answer (which is often something like a tiny typo was made) is known then that is effectively a full answer. A very good way of getting and giving help in questions like this is to talk in pseudocode- step by step English instructions of what you want the code to do. It is also just fine to chip in with discussion about particular functions (for example "Because this is a binary file download.file() needs the setting mode="wb"). Or providing test cases of data that people can use to work out where they went wrong. The goal is to help people get better at solving problems they hit when working with data, so that when they are no longer part of a course they can solve their own problems, this is why the best way to help is to guide people toward a solution rather than give it.

It helps to think of the quizzes as homework assignments, not in-class quizzes. The course wants you to experience doing things in practice and applying principles to new data sets, not just answering by repeating lecture slides. It is OK (and time efficient) to look at the quiz before watching the lectures to see what to pay a lot of attention to. You will be expected to use the lectures as starting points and research the answers.

About these ads (https://wordpress.com/about-these-ads/)

August 31, 2015 thoughtfulbloke Coursera, Getting and Cleaning, R