thoughtfulbloke aka David Hood

Getting and Cleaning week 3

So we are mostly past machine configuration problems, and people should be getting used to actually checking the help, so thing should start getting easier this week (of course there is still the assignment)

- 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/)

Quiz 3 advice

Question 1

A few people are not clear on what a logical vector is. If I had a list of numbers

```
ml <- c(1, 2, 3, 4, 5, 6)
```

and I write some subsetting criteria

```
    lv <- ml > 4
```

lv is now a logical vector, a list of true and falses telling you does the entry match the rule, so if you printed it out you get

```
    FALSE, FALSE, FALSE, FALSE, TRUE, TRUE
```

Logical Vectors are the basis of subsetting (identifying which entries you want to use through the use of rules) so this question ties to the Subsetting and Sorting lecture.

Basically, if you actually subsetted the data you went a step to far

Question 2

Because people in previous courses have had trouble with finding information between lots of different threads on particular topics, I am creating some initial threaids for particular quiz questions to give people an obvious place to look for help.

One way people go wrong with this question is in downloading the file. Images, like Excel files, are binary files, so mode needs to be set accordingly, the "it doesn't fully work in image viewing programs" is a very similar symptom to "the Excel file won't open. **mode="wb"**.

As to why some linux systems produce an answer 638 different for the 30th? I don't know. I could speculate, but that is all it would be

Question 3

My interpretation of Quiz 3 Question 3 is the absolutely critical learning objective is to get used to extremely common issues of working with real data. This is one of those "solve the problem by any means necessary" kind of questions that hark back to the Hacking Skills discussion in the perquisite Toolbox course, so might involve some looking things up and checking what is going on with your data when you do something to it (rather than finding an answer on a lecture slide).

This question involves putting together pretty much everything you have learned to date. Because there is so much going on, you might want to try a practice question where it is just fine to discuss code to solve it:

https://github.com/thoughtfulbloke/faoexample (https://github.com/thoughtfulbloke/faoexample)

And while there are many different pathways to a solution, some include elements of:

- using read.csv but changing some of the settings when reading the data in (many of the previous weeks lectures in this course involve this kind of step)
- using subsetting to pick out the areas of interest (the prerequisite R programming course, but recapped in the Week 3 subsetting lectures in this course)
- Identifying the kind of column, and changing it if needed (the prerequisite R programming course, but recapped in the Week 3 summarising lectures in this course)
- joining the two tables together (Week 3 merging data lecture).
- checking what effect all of the above are having on the ordering of the data (checking you data is the prerequisite R programming course, but recapped in the Week 3 summarising lectures in this course)

I do want to stress the checking what you are doing as you go, as mentioned in the general advice for doing this course, because (for example as a rhetorical question) if you have information about 190 countries and merge it with more information about those countries then what should be the logical upper bound of the number of countries in the results file? Getting something outside that range is a pretty big clue you should revisit your data.

Equally if the question says to read in the 190 ranked countries, and straight after you do a read you check what you have got, and it is not the 190 ranked countries, you might want to investigate that.

In general, problems are easier to fix earlier than repair later- if you are writing you code in a script you can just change the earlier part. If you are trying to fix things later, you will almost certainly have to do some conversions from factor, so consider this a warning:

```
numbersAsText <- c("10", "100", "11", "9","1000")
nAsFactors <- as.factor(numbersAsText)
convert2number <- as.numeric(nAsFactors)
convertViacharacter <- as.numeric(as.character(nAsFactors))
sum(convert2number)
sum(convertViacharacter)
convert2number
convertViacharacter
```

But you will not see this problem if you get the right early enough in the steps.

Question 4

Everything I posted about in question 3 applies here, so check above, including checking your work as you go.

A few people rush reading the question a little- we are using GDP rank not GDP.

Question 5

Everything I posted about in question 3 applies here, so check above.

My advice here is that this question is for practicing skills from the Making New Variables lectures.

If you wind up with a group of 37 rather than 38, look to how you are grouping the data into bins (ranges) with regard to the options outlined lecture. The subtle variation in how you divide up you data boils down to the difference for x between y <= x < z compared to y < x <= z so if you have data sitting right on the threshold that can be very important to be clear about. You might also think about where you have threshold values and why. In particular what happens around putting thresholds at the same place as the highest or lowest value.

You can force the various ways to divide up the data to use the other one than the default (which end is "open") but it is why the question expressly tells you that the correct answer should involve a set of size 38.

September 9, 2015 thoughtfulbloke Coursera, Getting and Cleaning, R