Midterm Clarifications Q&A

This is a list of all the questions Dr. Li and the teaching fellows have answered as clarifications for other students, so all students can have access to the same information. The question numbers below refer to the question on the midterm. There is also a "general advice/tips" section at the bottom of this document.

Question 1

- Q: Q1 says to use "mutate" function. I was wondering if it is a specific requirement for this question, or just a code suggestion. Because "mutate" results in a new date column, but I have used another way to just transform the current column, then I don´t need to add a new one.
- A: It's just a suggestion, you don't have to use the mutate function

- Q: By 'Only include the dates "1/22/20" - "10/31/20".', do you mean to include all dates from 1/22/20 to 10/31/20, or only those two dates (1/22/20 and 10/31/20)?
- A: We mean all dates.

Question 2

- Q: I am having a lot of trouble grouping by country given the back-slash in "Country/Region."
  A: Try using the rename function and renaming Country/Region to just country

Question 3

- Q: For question 3, we have to make a plot showing a few different countries one being the United States. However, the United States is not a country included in the dataset.
  A: The United States is listed as US in that data frame.

Question 4

Question 5

- Q: When I try to group by country and arrange the dates, it does not work (it seems to arrange by date first and then have each country listed)
- A: The group_by() to arrange() pipeline doesn't quite work out of the box as you might expect. If you need to arrange a grouped data frame within groups, take a look at the .by_group (notice the period) argument in the arrange function. (https://www.statology.org/dplyr-arrange-by-group/)

  However, note that if the whole data frame is sorted by some variable and then grouped, even if the data frame doesn't print out as sorted by that variable within group, any function applied at the group level (eg. mutate) applies to only one group at a time. Hence, the overall data frame being sorted is equivalent to it being sorted within each group for the purposes of calculating something at the group level, and thus you don't need to worry about the .by_group argument if you don't want to (that only matters if you like to see the sorting w/in each group when the data frame is printed or viewed).

In other words, if you arrange and then group_by, the order you want (by date) will still be in place. Make sure to arrange and group_by before using mutate.

- Q: For question 5 in the midterm where it asks for new cases each day, I see some negative numbers for new cases.
- A: There are data quality problems in the JHU data we are using – sometimes the case counts were revised. This isn't inherently an issue, and shouldn't affect any specific details of this problem.

  However, if you are getting negative values for the **first day** in each country you probably are making a small mistake. Take a look at Lab 03 Q1.

Question 6

- Q: For question 6 on the midterm, when we were asked to plot again. Do you want us to still filter to ("China", "Colombia", "Germany", "Nigeria", "US") as well, or plot all countries?
- A: Still just those countries

- Q: Do we keep the transformation or just plot it as raw?
- A: No transformation needed.

- Q: Can we combine the code for Q5 and Q6 into 1 step?
- A. Yes, that's fine. Just be sure to mention this in writing.

Question 7

- Q: For question 6 on the midterm, when we were asked to plot again. Do you want us to still filter to ("China", "Colombia", "Germany", "Nigeria", "US") as well, or plot all countries?
- A: Still just those countries

Question 8 (Bonus question)

General Advice / Tips

- Check that your code is doing what you want it to do along the way
  - Example: see what output you get every time you use the pipe
- Make sure you answer **every** part of every question.
- Do not type text answers in code chunks - type them in the main body of the file, outside of code chunks.