TAKE-HOME MIDTERM

Instructions:

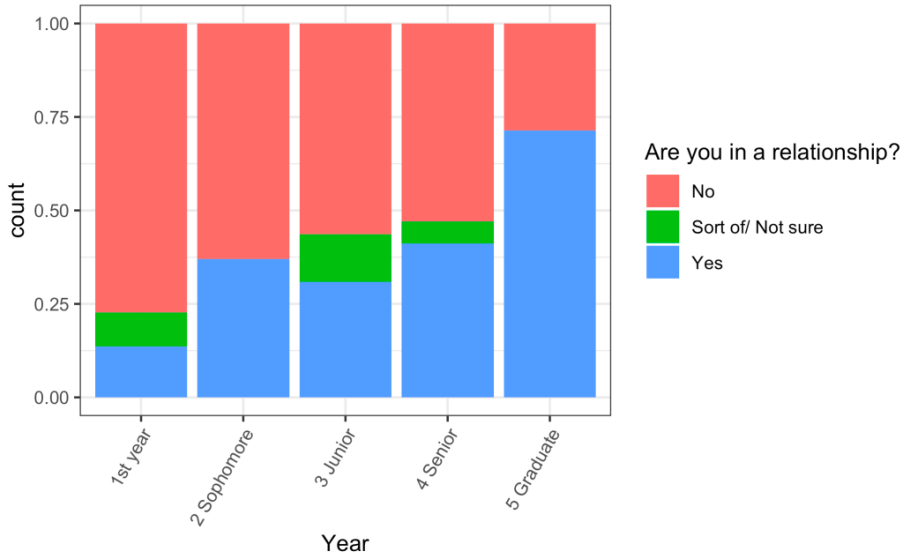
- Edit the Rmd script with code for all of the questions, and answering the written-answer questions in the 'white part' of the Rmd script.
- When reproducing graphs, **be sure to match** labels, backgrounds, treatment of missing values, etc.
- Make your code as neat and readable as possible; use comments. (Among other reasons, it will help us give you partial credit if we are trying to interpret your code!)
- Use methods that we learned in class. Small deviations are ok, but you must show that you have learned the basic techniques we've covered in class. For example, use dplyr and filter for removing values, and **do not use theme()**, except theme_ for backgrounds.
- Knit your Rmd document as html. Open the html document in your browser, then Save as (or Print to) pdf. Save your pdf with the name: MT*yourlastname.pdf*, and upload the file using the blue link.

The rules: You may not consult any other person while doing this test, either in person, email, texting, etc. You may not look at other persons' work, or show your work to others. You may use your text, your notes, class programs, homework assignments, class handouts, and class review materials. This test is focused on doing programming tasks using the methods that we've used in class. Minor deviations from class methods are ok, but major differences are not. You may Google problems, but know that this can lead you to methods not used in this course, and you may not get full credit. Finally, neither the TA nor I can help you do the problems. If you have a 'procedural' question (e.g., where is the data file?) you can certainly contact either of us.

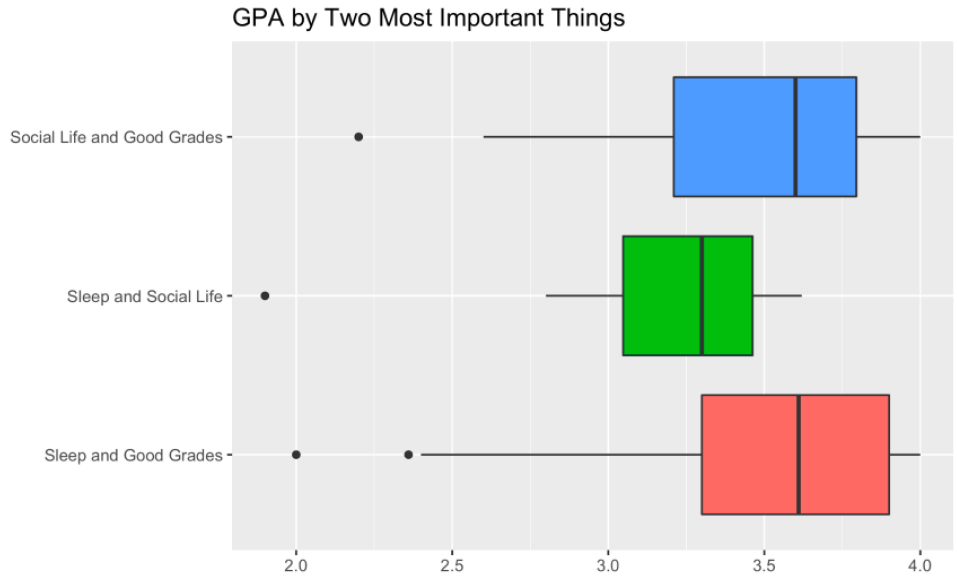
Introduction to Data Science

1. **Include** these in the set-up code chunk:
   - Load the packages that you need.
   - Read in the data, using `read.csv("SurveyforMTF21.csv", na.strings = '', stringsAsFactors = TRUE)`
   - Run a `summary()` of the data frame.

2. Write the code to create the plot below. **Use ggplot** for the plot, and include any other code if you use that, too (e.g., dplyr code for removing missing values). Note the theme used, the legend label, and the appearance of '1st year', '2 Sophomore', etc.. ***Write*:** Describe the trend that you observe, in terms of these variables, and why you believe it exists.

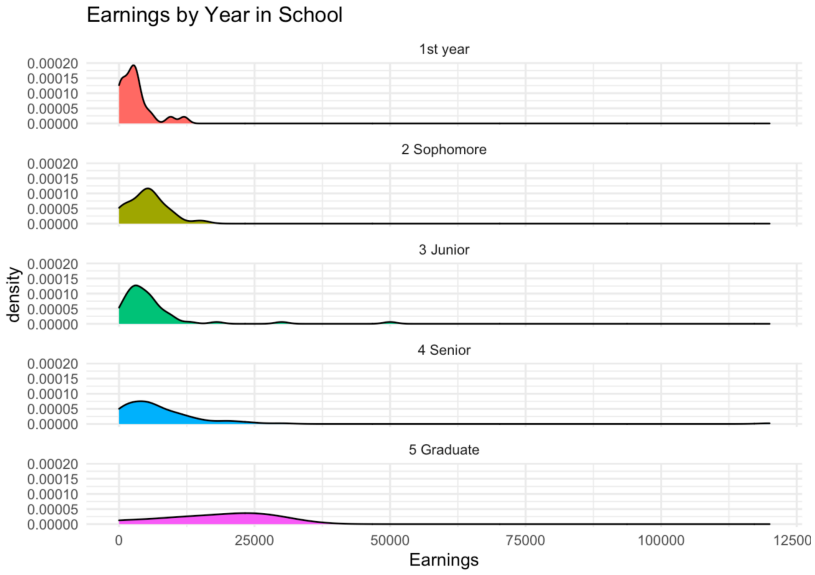


3. Write the code to create the plot below. **Use ggplot** for the plot, and include any other code needed (e.g., dplyr code). Note the theme, labels, and that there is no legend. ***Write*:** Describe the trend that you observe, in terms of these variables, and why you believe it exists.



4. Using **dplyr**, have R print the mean and median earnings by year in school, and the number of students in each year in school. Use the method we used in class, i.e., don't create separate data frames. Be sure there are no NAs in your output.

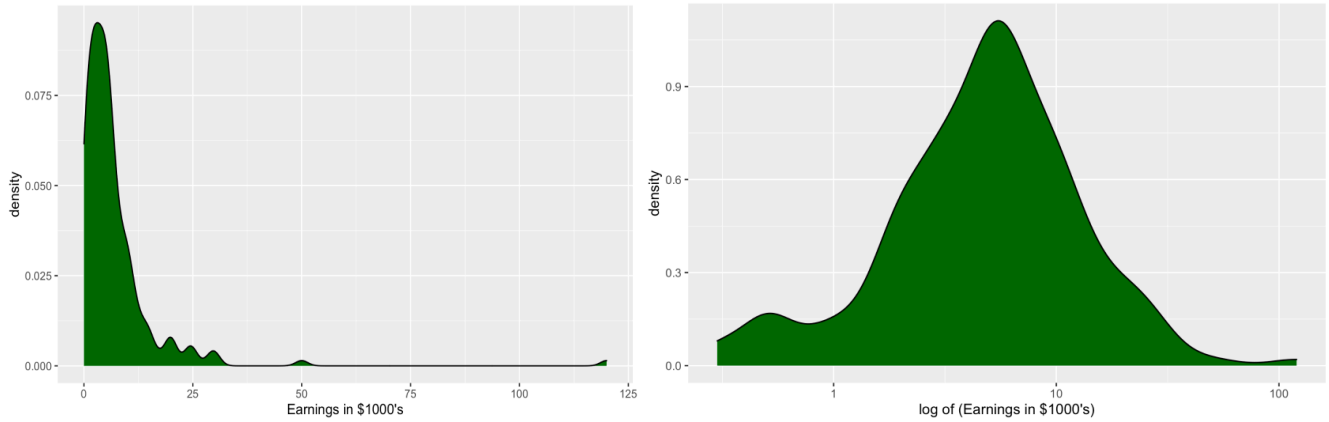
5. Write the code to create the plot below. **Use ggplot** for the plot, and include any other code needed (e.g., dplyr code). Note the title, and that there is no legend. ***Write*:** Based on your graph, and on the stats you calculated in part 4, does it appear that year in school is related to earnings? If so, why?



6. Using **dplyr**, print data for the **8 students with the highest GPAs**. Print only their year in school, college GPA, time spent on their phone, hours of sleep, days eating breakfast per week, and the number of books they read for pleasure. Filter out any students that have missing values for any of these variables, before printing the list of 8 students. Next, print the same type of list for the **8 students with the lowest GPAs**.

   Write: Do you notice any trends? That is, are the two groups of students different from each other in terms of phone time, sleep, breakfast, and books ?

7. Finally, use **dplyr** to create a <u>new earnings variable</u> that gives earnings in **thousands of $,** rather than $. Then use ggplot to make these two plots. Note color and x axis labels. (They don't have to be side-by-side in your pdf – they can be two separate plots).



Introduction to Data Science