Week 12 Challenge Problem

CSCI E-5a: Programming in R (Spring 2022)

Overview

Starting in Week 12, we don't have any more official problem sets. However, I wanted to provide you with something to work on and give you some more practice. This challenge problem is completely voluntary, and you don't have to do it if you don't want to. In fact, there won't even be an entry in Gradescope! This is just for your benefit, and if you're still finishing up some of the problem sets then you should be focus on that and leave this challenge problem for later.

Motivation

I also wanted to give you a taste of a more-or-less "real world" application, so you could see how to use R to solve an actual problem instead of the artificial problems that we work on in class. In general, this is not possible, because genuine real-world challenges tend to demand detailed knowledge of a specific substantive topic. For instance, if we study a problem in genetics, then it's very helpful to know a lot about genetics, because then you can just focus on the R programming concepts. But students who don't have a background in genetics will find it difficult to think about these programming concepts, because they have to spend a lot of energy trying to understand the details of the genetic analysis. And the converse is true as well – if we study an example from finance, then the genetics students might find this difficult to follow. So it's hard to find an example that is both genuinely challenging while being accessible to the whole class.

But there is ONE very real-world problem that everyone in the class will understand: scoring your problem set submissions! Even though the method for doing this is a little complicated, you should have a strong familiarity with the details of the process by now. Determining your final grade on a problem set involves a variety of techniques that we've studied in this course, and it serves as a nice challenge problem to help you to integrate many different R concepts.

Review

Let's review how I calculate your final score for a problem set based on your scores across the three grading sessions for that problem set. First, for each individual problem, your final score is the maximum score across all three grading sessions. That means that if you receive a score of 1 on any of the grading sessions, you'll get a final score of 1 for that problem. Then, once I've calculated the final score for each individual problem, then I add them all together to

obtain your final combined score for the problem set, and this is what is posted on Canvas Grades.

For instance, suppose a student has these scores for Problem 1:

- First Grading Session: 0
- Second Grading Session: 1
- Third Grading Session: 0

The maximum score across all three grading sessions is 1, so the final score for Problem 1 is 1. You could also think of this as: a student receives a final score of 1 on a problem if and only if there was at least one grading session where the student received a score of 1. On the other hand, if a student received a score of 0 on that problem for each of the three grading sessions, then the student receives a 0 for the final score for that problem.

Notice that we do NOT simply add up the points across the three grading sessions. For instance, suppose a student has these scores for Problem 1:

- First Grading Session: 1
- Second Grading Session: 1
- Third Grading Session: 1

In this case, the student scored a 1 for each grading session, and so the student should receive 1 point for that problem. However, if we simply added the scores together, the student would receive a score of 3 for this problem. Don't do that! I think you'll find that the simplest approach is to just take the maximum of the three grading sessions, for each problem.

About Gradescope Downloads

In Gradescope we treat each grading session as a completely separate assignment, and there's no way to coordinate the scores across these separate assignments. Thus, there's no way in Gradescope for us to determine your combined score for a problem set, that is, the overall score that you receive across all three grading sessions. Instead, I download the scores for each grading session as separate .CSV files, and then run an R program that will take these three .CSV files and calculate the final combined score. These scores are then exported to a .CSV file that can then be uploaded into the Canvas Gradebook.

For this challenge problem, I've created three files representing the output from Gradescope for a hypothetical problem set. These files consist of simulated data, and contain no actual student data. However, the format of the files is completely the same as what I receive from Gradescope, and it has exactly the same columns and headers as a "real" Gradescope download. Also, just like a "real" Gradescope download, there isn't much in the way of documentation! For the most part, the column headers are sufficiently informative that you can figure out the contents without a lot of explanation.

For our purposes, the columns that really matter are the SID column and the columns for each individual problem. The SID column is very important, because this is the unique identifier for each student; also, I need this information in order to upload the scores into Canvas. Of course the columns for each individual problem are very important as well! You can discard all the other columns.

When Gradescope outputs the scores for an assignment, the students are not listed consistently in the same order. That is, the order of the students will not be the same across the three grading session files. You'll have to think about how to deal with this so that when you combine the data for all three grading sessions the scores for each student are properly aligned.

The Challenge

The challenge here is to take the three Gradescope files in CSV form and construct a file of combined scores. The procedure here is conceptually straightforward:

- Read in each of the three Gradescope files.
- Combine them together so that all the data for each student is on one row.
- For each individual problem, take the maximum score across the three grading sessions, and this becomes the final score for that individual problem.
- Add the final scores for the individual problems together to obtain the final total score for the assignment.
- Write the final total score and the SID identifier out to a file.

You're on your own for this problem! You'll have to figure out how to organize your files, how to read in the data, what columns to use, how to properly align the data, etc. There are many different ways to approach this, and there's no single correct method. However, remember that we are trying to determine student grades here, and I can tell you that even the slightest inaccuracy will not be tolerated. This is a situation where you MUST get the correct answer, for each person!

I've included one more .CSV file called "Answer Check.csv". This file consists of the SID value and the final combined score for each student, so you can check your work.

Have fun with this, and I'll release my solution next week.