# HW 2: Building Decision Support Tools

## Purpose of This Assignment

In this assignment, you will build your own functions in R. Being able to write your own functions is a critical programming skill. This is not only because you learn the coding language and practice thinking logically but also because it gets you in the habit of tackling complex problems at a high level first before diving into the details. Also, it forces you to think about how to break a complicated process down step by step. Writing your own functions in R will be relevant once you start working on your projects later this semester. In the last problem, we introduce you to the essence of simulation modelling. Throughout the entire assignment, do not think of yourself as a student working on a homework assignment. Imagine that you were asked to build something for a client, and that your reputation is on the line if this product you were asked to build breaks down or does not satisfy the client's requirements.

## Midterm Retake Policy (50 points)

1. Due to COVID-19, many instructors are experimenting with novel online examination procedures. A professor has asked me for your help to design the online midterm retake policy for their class.

For the midterm, each student will have one initial attempt. The score of that attempt is that student's pre-retake score. After the midterm, students can retry the exam an unlimited number of times on D2L (before a specified deadline), learning from their mistakes along the way.

The midterm retake policy takes the following form. If your pre-retake score is less than $P$, then if your highest retake attempt score $R$ is at least $P+B$, then your updated midterm grade will be $P+B$. Otherwise, it will be $R$. If your pre-retake score $S$ is at least $P$, then if your highest midterm retake attempt score $R$ is at least $S+B$, then your updated midterm grade will be $S+B$. Otherwise, it will be $R$. In other words, there is a cap on how high a student's post-retake score can be which depends on the student's initial score.

For example, suppose the midterm policy had $P = 30$ and $B = 30$. The policy would be: If your pre-retake score is less than 30, then if your highest retake attempt score $R$ is at least 60, then your updated midterm grade will be 60. Otherwise, it will be $R$. If your pre-retake score $S$ is at least 30, then if your highest midterm retake attempt score $R$ is at least $S+30$, then your updated midterm grade will be $S+30$. Otherwise, it will be $R$. For example, if you got a pre-retake score of 45, then your highest possible post-retake midterm grade would be 75. If you got a pre-retake score of 60, then your highest possible post-retake midterm grade would be 90. If you do not get at least 90 on a retake attempt (for example, suppose your highest retake attempt score was $R = 85$), then your post-retake midterm grade would be your highest retake attempt score (85).

**To get credit for the questions below, show your work as R comments in your homework file.**

1) Suppose you were in that class. If you initially got a score of 21 on the midterm. Also suppose that the midterm retake policy was set so that $P = 30$ and $B = 30$. Suppose that you attempted the midterm over and over again until you scored $R = 50$. Then what would be your post-retake midterm score?

2) Suppose you initially got a score of 21 on the midterm. Further suppose that the midterm retake policy was set so that $P = 30$ and $B = 30$. Suppose that you attempted the midterm over and over again until you scored $R = 60$. Then what would be your post-retake midterm score?

3) Suppose you initially got a score of 21 on the midterm. Further suppose that the midterm retake policy was set so that $P = 30$ and $B = 30$. Suppose that you attempted the midterm over and over again until you scored $R = 70$. Then what would be your post-retake midterm score?

4) Suppose you initially got a score of 21 on the midterm. Further suppose that the midterm retake policy was set so that $P = 30$ and $B = 30$. What would be your highest possible post-retake score?

5) Suppose you initially got a score of 54. Further suppose that the midterm retake policy was set so that $P = 22$ and $B = 30$. Suppose that you attempted the midterm over and over again until you scored $R = 67$. Then what would be your post-retake midterm score?

6) Suppose you initially got a score of 54. Further suppose that the midterm retake policy was set so that $P = 22$ and $B = 30$. Suppose that you attempted the midterm over and over again until you scored $R = 78$. Then what would be your post-retake midterm score?

7) Suppose you initially got a score of 54. Further suppose that the midterm retake policy was set so that $P = 22$ and $B = 30$. Suppose that you attempted the midterm over and over again until you scored $R = 89$. Then what would be your post-retake midterm score?

8) Suppose you initially got a score of 54. Further suppose that the midterm retake policy was set so that $P = 22$ and $B = 30$. Then what would be your highest possible post-retake midterm score?

9) Throughout the rest of the problem, assume that all students achieve their highest possible post-retake scores. In other words, assume that each student keeps attempting the midterm retake until he or she gets a retake score $R$ that achieves his or her maximum possible post-retake score. For example, suppose you initially got a score of 22 on the midterm. Further suppose that the midterm retake policy was set so that $P = 30$ and $B = 30$. What would be your highest possible post-retake score?

10) Suppose you initially got a score of 86 on the midterm. Further suppose that the midterm retake policy was set so that $P = 30$ and $B = 30$. What would be your highest possible post-retake score? Assume that there are enough bonus problems on the midterm that your score can be over 100 (for example, imagine that there are 50 points worth of bonus problems on the midterm).

**Building the Midterm Retake Policy Function**

Given a data frame that consists of a column of pre-retake midterm scores, **build a function** that tells the professor what pre-retake threshold $P$ and boost amount $B$ to use so that the maximum post-retake average is between 70 and 75. What is meant by 'maximum post-retake average' is that **you should assume each student achieves their highest possible post-retake score**. Your function should satisfy the following:

1. $P$ and $B$ are positive multiples of 2.

2. If possible, choose $P$ and $B$ so the maximum post-retake average is between 70 and 75 (strict inequality) and that $P + B \geq 60$. If there are multiple $P$ and $B$ that satisfy this criterion, then choose the one whose standard deviation of post-retake scores is closest to the standard deviation of the pre-retake scores.

3. (Optional) If no $P$ and $B$ with $P + B \geq 60$ can make the maximum post-retake average lie between 70 and 75, then if possible, choose $P$ and $B$ so that the post-retake average is between 70 and 75 and that $P + B$ is as close to 60 as possible. If multiple $P$ and $B$ satisfy this, use the standard deviation condition above.

4. (Optional) If there is no $P$ and $B$ that would make the maximum post-retake average lie between 70 and 75, then choose $P$ and $B$ to get a post-retake average above 75 but as close to 75 as possible.

**D2L Questions**: What $P$ and $B$ do your function output when applying your function to the five RDA files in the HW 2 D2L folder containing pre-retake midterm scores?

**Tip #1**: Think about the structure of the problem. Only once you **understand the problem** should you begin to think about how you could **design a step-by-step process** to find $P$ and $B$ to satisfy the problem.

**Tip #2**: After you have a potential solution approach, then begin to think about the code to implement that approach. One way to see if your code works is to apply your functions on the datasets provided in the HW folder for this assignment and use the D2L quiz as a way to check if you got it right/wrong. Another way is to construct a simple dataset yourself, so that you can more easily track how pieces of your code are working on that simple dataset. For example, you can construct a dataset with eight students with scores: 10, 20, 30, 40, 50, 60, 70, 80. If you take this second approach, then repeat this process again for other datasets with simple numbers you construct to 'test out your function(s).' Once you are confident that your code works on any dataset, then you can become more confident that your code should work for any number of students and any set of pre-retake scores whose average is less than 70.

**Note**: **Make sure you have built a user-defined function for this problem.** I should be able to run a user-defined function you built to answer the D2L questions for this problem. **You will lose 25 points for this problem if your answers to the D2L questions about what P and B to choose are not outputs of a function you built.**

WARNING: From a student's perspective, HW #2 has historically been the hardest homework assignment of the semester for this course. If you feel that this homework assignment is very challenging, almost all of your classmates will also agree with you. From an instructor's perspective, HW #2 has historically been the assignment that has tempted many students to commit academic integrity violations. If you are tempted to share your solution or copy and modify a solution from a friend, please reconsider. Working together is great but only to the extent that you discuss solution approaches at a high level. Do not share your code, screenshot your code or someone else's code, let someone take a quick peek at your code for some inspiration, etc. On top of the TAs carefully marking homework submissions, we have an automated tool that can examine code across everyone's submissions. If your submission is flagged, the investigation may take weeks, since we do not want to accuse you of anything you did not do. If we decide there is enough evidence to submit to the appropriate office, then we will submit the evidence from our investigation. Your score for the assignment would be updated to a zero, and you can feel free to appeal the charge through the formal appeals process conducted by the office in charge of the investigation. The process will be handled entirely by the office from that point forward, and they would ultimately decide on the penalties.

WARNING #2: The above warning was written and displayed visibly on the assignment in a prior semester, and several students still decided to take the risk anyways. If you are struggling on the homework, then schedule an office hours appointment to get help. Don't ask your friends to give you their solutions. We set high standards to make sure that everyone comes out of the program with a solid set of foundational skills. There is a zero-tolerance policy for academic misconduct in this class.

## Random Student Selector (10 points)

2. One of the challenges with moving courses online is that instructors are worried that students may tune out more easily. An instructor at Haskayne wants to keep students on their toes by calling on students throughout the class. However, the instructor does not want to always pick the same students (students that come to office hours, students whose names are at the top of the class roster, etc.). Instead, the instructor wants to be more fair about how the names are chosen by calling students at random.

In this problem, you will create the Random Student Selector function. In particular, given a number $N$ and a csv file containing the first and last names of students (as provided on D2L in the HW 2 folder), **build a function** that prints or returns $N$ randomly selected students' full names. Your function should have a number $N$ and a class roster (as a csv file) as an input. The requirement of the function is to print or return $N$ randomly selected students' full names. **Comment your code so that the TA and I can know how to use your function.**

To test your function, you can apply your function on our class roster. If you have an error with the read.csv() function, make sure that the csv file is in the current working directory. To check your current working directory, use the getwd() function. You can change your working directory with the setwd() function or using RStudio's tools.

**Tip #1**: When reading the csv, using the read.csv() function, set the stringsAsFactors parameter to FALSE. That way, you don't have your columns be factors (levels of a categorical variable in regression) as default.

**Tip #2**: Make sure you have return() somewhere in the body of your function so that it's clear to you (and to the TA and I) what the output of your function was supposed to be. Using the return() function in your user-defined function is not necessary (by default, R will assume the object that was last defined in your function is your output), but it is helpful to be explicit about what the output of that function was supposed to be. This tip applies for any user-defined function you make (especially when working on team projects).

**D2L Questions**: How many arguments does your Random Student Selector function have? Could an instructor of BTMA 601 use your function to randomly select three BTMA 601 students? Could an instructor of MGST 217 use your function to randomly select four MGST 217 students? If I asked you to randomly select five students from this class, is there a chance that you would be selected? Assuming no two students have the same name, could the output of your function ever print the same student twice when running your function once?

**Warning**: On the D2L quiz for this homework, there are hypothetical Yes/No or simple numeric questions for this problem. Do not just guess the right Yes/No answer without making sure your code matches your response. If you have done the problem above, then you will have created a user-defined function. You must have defined a function in your solution above (I should see function(...) somewhere in your code). If you did not build a user-defined function but somehow got the questions on D2L correct (i.e., you said that you created a function on the D2L quiz, but you did not make one), then 25 points will be deducted from your assignment score for misrepresenting your solution in the assignment. If you (whether intentionally or accidentally) misrepresented your solution on D2L (for example, if I asked you how many inputs your function has, and your D2L answer is correct but does not match what your function looks like), then you will lose 20 points. In either case, note that the penalty is greater than the number of points for the problem.

**Make sure your D2L answers are consistent with your function.**

**Note**: In prior years, we had relied on an honor code system, in the sense that we did not have the Warning message above explicitly written this way but expected students to answer truthfully on D2L. Unfortunately, a significant fraction of students misrepresented their solutions on D2L. The ability to learn from your mistakes by having unlimited attempts on the homework is a core feature of the class, but when used improperly, this feature turns the assignment into a game of clicking buttons until you get a perfect score (which is not the point of giving students the chance to learn from their mistakes). The above message should not be interpreted as accusing you of potential misconduct. Instead, it is meant to set very clear expectations in terms of academic standards.

## Peer Review Assignment Function (20 points)

3. In a class at Haskayne, students will be evaluating other students' projects. The professor of the class is asking for your help to develop a tool to quickly assign students to mark other students' projects. Each student will evaluate three other students' projects (students may not evaluate their own projects).

Given a data frame that contains two columns (assume that the first column is "Last Name" and the second column is called "First Name," as that is how the names are displayed when exporting the roster from D2L), **build a function** that randomly assigns each student to be a judge of three other students' projects. Make sure your function is such that each student has three classmates evaluating his or her project (so that it's not the case that every student evaluates the same student and no other students get feedback). You can design your output however you want, but make sure to **comment your code** to make it clear how to use your function and interpret the output. **Make your function output either a list or a data frame.** Figure out how to structure that list or data frame so that the instructor can run your function see in a glance what they want to see. In other words, if your function just prints a few students' names, then that's not enough. Your function **should work on any class roster of any size with at least four students.**

**D2L Questions:** Would your function work on any class roster with at least four students? How many inputs does your function have? If I used your function, would it be possible that some student has two or fewer peer evaluators? If I used your function, would it be possible that some student evaluates four or more of their peers' projects? If I used your function, would it be possible for a student to evaluate his or her own project?

**Tip #1**: As always, think about the problem conceptually before jumping into the code (and perhaps even draw a diagram, if you are a visual thinker). Write down your approach before jumping into writing the code itself so that you have a high-level sense of what your function does. It is possible to write such a function with only a few lines of code.

**Tip #2**: Be careful of calling objects within your function that were defined outside of your function. If you are calling an object that was defined outside of your function (i.e., an object that is already in the Global Environment), then this is very risky to do unless you are sure this is what you want to do. If someone tries to run your function without that object already defined, then your function won't work.

**Tip #3**: Students often lose points on this question because they did not guarantee that their function would satisfy the requirements. For example, the output of their function might have that a student evaluates his or her own work or that two of the reviewers might be the same person. As a Quality Assurance step, it is recommended that you create a function to ensure quality (in addition to the function you are asked to build). This function should guarantee that each student has three distinct peer evaluators and that that nobody is evaluating his or her own projects. In essence, if you build this quality assurance function, then you can can "prove" that your function works and satisfies the specified requirements that it was supposed to satisfy.

**Warning**: The D2L questions associated to this problem are Yes/No questions. Do not misrepresent your solution in the D2L quiz; answer the questions truthfully. The same penalties from Problem #2 apply if you misrepresent your solution for this problem on the D2L Quiz.

## Let's Play a Game (20 points)

4. This game illustrates the basic idea of simulation models. One of the homework problems in HW 3 builds on top of the skills developed here to tackle a common problem in operations management. **For each sub-question below, comment your code so that the TA and I know where to look for work for 4a, 4b, 4c, 4d, 4e, and 4f.**

The game goes as follows. First, you choose a whole number between 0 and 1000 (inclusive). Then I use R's sample() function to randomly select a whole number between 0 and 1000 (inclusive). Then you pay me the square of the difference between the numbers. For example, if you choose 2 and the random number generator says 5, then you pay me $\$(3)^2 = \$9$. On the other hand, if you choose 80 and the random number generator says 20, then you would pay me $\$(60)^2 = \$3600$. You always pay me unless you perfectly guess my number.

4a) (2 point) To the nearest five thousand dollars (do not include the decimal or the cents, just give the whole dollar value), how much would you expect to pay me if you chose 30?

4b) (2 point) To the nearest five thousand dollars (again, leave out the decimal and cents, only writing the dollar value), how much would you expect to pay me if you chose 950?

4c) (2 point) To the nearest five thousand dollars, how much would you expect to pay me if you chose 450?

4d) (5 points) If you had to, what number would you choose to minimize your expected loss? Round to the nearest multiple of 5.

**Note**: Guessing the answer is not sufficient. **You need to show your work to justify that your choice indeed incurs the lowest expected loss out of all your possible choices (for example, by using which.min() or by creating a plot of how the expected loss changes in the decision variable).**

4e) (2 point) What is your expected loss at this chosen number? Round to the nearest five thousand dollars, with the same format as in the previous questions.

**Hint/Note**: If you want more context, you can think of this as a model of politics. You are about to make a speech to the press and the general public. How do you position yourself, knowing that people on both sides of the aisle may toss eggs at you or write scathing articles about you if you say the wrong thing?

**Super Hint**: Feel free to use the following code as a template.

```
N <- 10000 # Number of simulations. Take it to N <- 1000000 when ready.

random.draws <- sample(0:1000, size = N, replace = TRUE) # Random simulation draws

# If it helps, think of the entries of that vector
# as simulation draws from different universes.

# Across the multiverse, different numbers were drawn.
# You don't know which particular universe you happen to be in.

# You want to make a decision to minimize your average loss across all potential outcomes.

# For each possible choice, you want to figure out your expected loss for that choice.

choices <- 0:1000
expected.loss.vec <- numeric(0) # Defines empty numeric vector to store values in

for(i in 1:length(choices)){
# Fill in the for-loop below to complete the problem.

}
```

**Note**: Do not worry about copy/pasting the chunk of code above into your homework submission. Feel free to use the template if you wish. The automated detection tool can handle common code chunks as inputs, and you will not be flagged for investigation if your code is very similar to another student's code because both of you used the template above.

4f) (7 points) I'm going to change the rules of the game. Instead of choosing a whole number between 0 and 1000, I will choose a number from a random set of numbers (specifically, a set of *real numbers*, which may include non-whole numbers, negative numbers, etc.). I will give you that set of numbers a few minutes before we play the game, and you have to decide what number to choose to minimize your expected loss (using the same squared-loss penalty as before). Build a function that takes in a vector of numbers as the input and returns the choice that minimizes your expected loss as the output. To get the points for this question, you must build a function. **Make sure your function can work on any numeric vector as an input (i.e., that it won't break down if I put in a random vector of arbitrary numbers).** Does your function work on any input vector without the user having to modify the function? What number do you get when applying your user-defined function to the dataset provided in the HW 2 folder called 'HW2Q4.rda'? **Round your answer to the nearest multiple of 5.**

**Note**: This is getting into the essence of simulation modelling. Think carefully about the structure and nature of the problem to guide you. Your intuition will help here, but you will need to formalize your thinking to let it generalize to more complex problems.

## Grading Scheme

### R Code

Submit your R code as **a single R file** to the D2L dropbox folder. In the file name, include the homework number, your first name, your last name, and your section number. An example would be 'HW2_firstName_lastName_L02.R' or 'HW2_firstName_lastName_L02.Rmd' (depending on whether or not you used RMarkdown). Submit your work as a single R file with the stated naming convention. **Don't create a zip file with multiple R files.** Make sure you clearly denote which problem is which problem. For example, use R code comments to write: #### Problem 3a ####.

If you have no R code showing your work, I will give no credit. If outputs from your R file do not match your numeric D2L responses on the HW 2 D2L Quiz, I will give no credit for your submission. Furthermore, if your answers to the conceptual questions do not match what you have in your R code, you will be penalized for misrepresenting your solution. For multiple choice questions, do not guess until you correctly get the answer if your code does not match your D2L quiz response. For simple numeric questions (for example, the first ten conceptual questions of the Midterm Retake Policy problem), show your work as R code comments.

**Aggregate Penalty Policy**: Your functions must work, and for the Midterm Retake Policy, the output of your function for each data set must match your D2L responses. You will incur penalties (worth more than the sub-problem itself) if you answer a question correctly on D2L without having made a function for the problem when requested to do so. The maximum aggregate penalty across Problems #1 – #4 will be capped at 50 points. For example, if you didn't know what was meant by 'build a function' (which is the purpose of the assignment) and did not build any functions when asked to build them (or if your functions did not satisfy the requirements specified in the problem), then you will have points deducted throughout the assignment, but the total penalty will be capped at 50 points.

**Caution**: Early on in the semester, students do not follow instructions carefully. If you follow the steps below, you can almost guarantee that you will not lose points due to your R code not running properly on my computer. Carefully making sure your code works fine and does not contain irrelevant chunks of code is good practice, and you want to have formed these habits prior to working together in teams on projects.

1) Save your current homework R file, and close RStudio.
2) Open your homework R file again.
3) Start at Line 1 (the first line in your code). Press Ctrl + Enter (Cmd + Enter for Macs).
4) Line by line, keep pressing Ctrl + Enter (Cmd + Enter) and check the outputs. Make sure that your outputs match your D2L quiz responses for the assignment.
5) Remove all extraneous bits or chunks of code that were not needed to produce your output. For example, if you had bits of code that were dead ends and the objects you created were not used in your solution, then delete those lines of code. Students have things like 'Attempt 4' or 'Attempt 10' in their code. Before doing this step, you may want to save a separate copy that contains all your dead ends (in case you accidentally delete something you needed). That's great to have a separate copy of your work (saved as a separate R file) containing the dead ends, but in your final submission, only submit code that is part of your working solution. Otherwise, if you include all prior attempts before your working solution, then the TA might accidentally mark the wrong approach and flag your work for academic misconduct (in which case, I would have to investigate).
6) **Repeat Steps 1 - 5 again (to make sure you didn't remove a needed line of code).**
7) **If you use setwd() in your code, then either use the load() function to load the datasets or write as R comments what files you are using from your working directory. I don't have access to your computer, so your code will not work on my computer if you just use setwd() and load the data in the Console. I need those datasets in my working directory to run your code.**
8) Save your file with your first name and last name as stated in the instructions, and submit your work to D2L. Your work for the assignment should all be in a single R file, not multiple R files.