HW 3: Building Simulation Models and Analyzing Customer Data

Purpose of This Assignment

In this assignment, we first build a simulation model. Building models is a useful skill to have in its own right (especially for making operational or financial decisions). It also gives you another chance to write loops or use an apply() function in context, and the concepts in the simulation problems will reinforce your understanding of probability. In the second problem, we explore data that Citi Bike, a bike-sharing company, has made public. Citi Bike made its data public so that it can effectively crowdsource its analytics efforts. For us, this presents an opportunity to build our data wrangling and exploratory data analysis (EDA) skills. In Problems #3 and #4, you build up your probability foundations through applied problems.

Simulation Model: Inventory Management (25 points total)

Your friend is running Brew and Blendz and wants to know how many muffins to stock tomorrow. The per unit cost of a muffin is \$0.20. The price for a muffin for customers (the per unit revenue) is \$2.50. Based on historical data from the last few months, the daily demand was found to be roughly Normally distributed with a mean of $\mu = 120$ and standard deviation $\sigma = 25$. Muffins have no salvage/resale value (unsold muffins are donated at the end of the day). How many muffins should be stocked tomorrow to maximize the expected profit? The first few questions will help you through the logic of the problem before trying to code anything.

1a. (1 point) If the demand tomorrow is 120 and only 100 muffins were stocked, then muffins are understocked. What would be the profit if only 100 muffins were stocked and it turns out that 100 hungry students and 20 hungry professors stopped by to buy muffins (with each person buying one muffin)?

1b. (1 point) If the demand is only 80 and 100 muffins were stocked, then muffins were overstocked. What would be the profit in this case?

1c. (1 point) What is the decision variable here (i.e., what lever is used to maximize expected profit)?

1d. (5 points) The expected profit changes for different stocking levels. What is the stocking level that maximizes the expected profit? (**Hint**: Writing a for-loop and using the which.max() function may help. But again, think through the logic of your approach before trying to code anything.)

1e. (3 points) At the optimal stocking level, what's the chance that all the demand for tomorrow will be met? **Note**: This is what operations managers call the *critical fractile*.

1f. (3 points) Conditional on running out, how much excess demand is there on average if stocking optimally?

1g. (3 points) Conditional on overstocking, how much excess supply is there on average if stocking optimally?

1h. (4 points) In this setting, how does higher demand variability impact expected profits? Create a plot of your expected profit at the optimal stocking level as a function of the standard deviation. In particular, your plot's x-axis should have σ between $\sigma = 0$ and $\sigma = 30$ in increments of 5. For each σ, figure out the optimal stocking level and the corresponding profit (essentially re-doing part (d) of the problem across different values of σ). Then plot (preferably using ggplot or plotly) the corresponding profit as a function of the demand's standard deviation, σ.

1i. (4 points) Suppose that the folks running the bakery Brew and Blendz buys the muffins from are celebrating their ten-year anniversary tomorrow by giving Brew and Blendz 100 muffins for free. How many additional muffins on top of the free 100, if any, should Brew and Blendz purchase tomorrow (at the regular price)? **Note**: Brew and Blendz is not giving muffins away for free, only their supplier is.

Exploratory Data Analysis: Citi Bike (35 points total)

In this problem, we will try to get a sense how much Citi Bike can make off of overage charges from its members and non-member casual customers in a single month.

```
url <- "https://s3.amazonaws.com/tripdata/202212-citibike-tripdata.csv.zip"
```

You can manually download the data to your hard drive, or you can do as we did in class.

```
temp <- tempfile()
download.file(url, temp)
citibike <- read.csv(unz(temp, "202212-citibike-tripdata.csv"),
                     stringsAsFactors = FALSE)
unlink(temp)
```

Some rows may have missing information. Remove all rows for which member_casual entry is missing.

```
  citibike.trips <- citibike %>%
  filter(member_casual == "member" | member_casual == "casual")
```

Citi Bike charges casual customers a fixed fee for renting, and members pay a monthly membership to forgo that fee. However, both casual customers and members are charged overage charges. In particular, casual customers who go over 30 minutes pay $0.26/min over the 30-minute mark. They start incurring this charge once their trip takes longer than 30 minutes. Members who go over 45 minutes pay $0.17/min over the 45-minute mark. For example a casual customer who spends 128.12 minutes would incur a charge of $\lceil(128.12 - 30)\rceil \times \$0.26 = 99 \times \$0.26 = \$25.74$, where $\lceil x \rceil$ denotes the smallest whole number greater than or equal to x (the ceiling() function applied to x). On the other hand, a member who spends 128.12 minutes would incur $\lceil(128.12 - 45)\rceil \times \$0.17 = 84 \times \$0.17 = \$14.28$ overage charge. Information can be found here: https://citibikenyc.com/pricing/single-ride and https://citibikenyc.com/pricing/annual.

Note: If the amount a person owes in overage charges doesn't exactly equal some cent value, then it would be rounded to the smallest cent value larger than what is owed. In other words, if a person owes $\$25.51\overline{0}$ exactly, then they would pay $2.51. However, if they owed $25.51000123, then they would be charged $25.52.

Note: **Ignore lost/stolen bikes** (bikes are considered lost/stolen if the trip duration is at least 24 hours). We just want to get an idea of the overage charges, and we don't want to include fines for lost/stolen bikes. **Also, some of the data is faulty. Remove observations for which the trip duration is negative.**

2a. (5 points) To simplify things, first consider the case when charges are continuously accrued. A charge of $0.26 every minute means that you're being charged $\frac{0.26}{60}$ dollars per second. Similarly, a charge of $0.17 every minute would amount to being charged $\frac{0.17}{60}$ dollars per second. For example, if a casual customer took the bike for 128.12 minutes, then they would owe $\lceil(128.12 \times 60 - 30 \times 60)\rceil \times \$\frac{0.26}{60} = 5888 \times \$\frac{0.26}{60} = \$25.51467$ in overage charges. They would actually pay $25.52, since the amount owed doesn't exactly equal some cent value and so it was rounded up to the smallest cent value larger than the amount owed.

Suppose Citi Bike's policy was to have overage charges accumulate continuously per second in this way. Focusing on trips that incurred an overage charge, what would be the average overage charge? Round to the nearest ten cents.

2b. (5 points) What would be the standard deviation of the overage charge? Round to the nearest ten cents.

2c. (5 points) Hypothetically, under the continuous charge policy, what would have been the total overage charge revenue from casual customers that month? What would have been the total overage charge revenue from Citi Bike members? Round to the nearest thousand dollars.

2d. (5 points) With Citi Bike's actual pricing policy, these overage charges are incurred in minute-by-minute increments instead of continuously accruing per second. With their per-minute charge policy, conditional on incurring an overage charge, what is the average overage charge for a trip? Round to the nearest ten cents.

2e. (5 points) What is the standard deviation of the overage charge? Round to the nearest ten cents.

2f. (5 points) Under Citi Bike's current policy, conditional on incurring an overage charge, what is the expected overage charge for a trip by a member? And similarly, conditional on incurring an overage charge, what is the expected overage charge for a trip by a casual customer? Round to the nearest ten cents.

2g. (5 points) Citi Bike asks you if they should modify their pricing policy. In particular, they are wondering which policy is better purely in terms of obtaining greater total overage charge revenue. Find the total overage charge revenue under their current pricing policy. Compare it to the sum of the overage charge revenue from members and casual customers (found earlier in Problem 2c). Would you recommend continuous charge or would you recommend that Citi Bike sticks with its current pricing policy?

Probability and Simulation Models: Inventory Management, Part II (15 points)

3. Going back to the Inventory Management problem above, let's re-do that problem with basic economics to see that the solution from the simulation model matches the theoretically optimal solution. While doing this problem, think about Problem 1i).

Suppose you were trying to solve for the theoretically optimal solution of the Inventory Management problem. You don't know what to stock because you cannot perfectly forecast the demand D. However, suppose you know the distribution of D based on historical data (specifically, the daily demand D has a Normal distribution with a mean of $\mu = 120$ and standard deviation $\sigma = 25$), and you want to figure out some stocking level (denoted as S^*) to maximize expected profit.

Thinking about it economically, if the cost of stocking one more muffin outweighed the expected benefit of having more stocked, then you would not want to stock more. On the other hand, if the expected benefit of stocking another muffin outweighed the cost of a muffin, then you would want to stock more. Therefore, at your optimal decision S^*, the expected marginal revenue of stocking an additional muffin (the marginal benefit) should equal the expected marginal cost.

3a. (2.5 points) What is the expected marginal cost of stocking one more muffin? **Hint**: The marginal cost is not a random variable. It is just the cost of stocking one more muffin, which is a constant rather than a random variable. Since marginal cost is not a random variable, the expected marginal cost is simply just the marginal cost.

3b. (2.5 points) What is the expected marginal revenue of stocking one more muffin? To help you, fill in the blank below:

$$\mathbb{E}[\text{Marginal Revenue}] = (\underline{\hspace{3cm}}) \times \$2.50$$

Hint: If you were to stock an additional muffin, then exactly one of two things will happen: 1) You sell that additional muffin you stock, or 2) You don't sell it. As a result, the marginal revenue is a discrete random variable which takes on one of two values. Think about what the two values that random variable takes on are, and then think about the definition of expected value for a discrete random variable.

Hint #2: If you currently stocked 1000 muffins, then the benefit of stocking one more is probably small (given the demand distribution for D). However, if you are currently stocking 20 muffins, then the benefit of stocking on more muffin is probably high (at least higher than the marginal benefit of stocking one more when you already had 1000 stocked). This suggests that the expected marginal revenue is a mathematical expression involving S^*.

3c. (10 points) Equating expected marginal cost with expected marginal revenue, what is the theoretically optimal stocking level S^*? Round to the nearest whole number. **Note**: I am asking for the theoretically optimal stocking level, which may or may not match your simulation result from 1d). Do not just copy/paste your work from 1d)'s simulation model into part c) of the problem. You need to find the theoretically optimal stocking level by equating part a) to part b) and solving for the optimal stocking level to get the points for this problem.

3d. (Optional, 0 points) **Reflecting on your work**: One measure of the service level of a process or service is the fraction of times demand is met. In the food/beverage service industry, how does the marginal cost of production impact the service level? Reflecting on COVID, if the marginal cost of production increases, would you expect the service level to go up or would you expect it to go down? How does the unit selling price (equivalently, the profit margin) impact the service level?

Instead of using the numbers provided in the original problem statement, try re-doing the questions above using a marginal cost of c and a per unit selling price of p. What do you find? Are your answers consistent with your intuitions?

Probability and Simulation Models: Lottery Game (25 points)

4. You are considering buying some raffle tickets in a global virtual lottery game. Imagine that each raffle ticket has a different number on it (no two people have the same raffle ticket number), and every three seconds, one raffle ticket is selected to be the winner of that raffle round. The lottery draws across several minutes of raffle draws look something like 250-174-13-23-37-228-139-168-148-39-115-111-58-235-55-25-56-156-190-118-105-53-54-159-134-138-8-144-184-126-93-69-87-56-232-252-181-11-177-77-58-246-131-183-16-168-54-218-205-85-116-192-148-24-151-172-97-187-182-88-187-220-73-215-190-250 79–0–194–163–0–255–232–122–28–108–255–174–68–207–27–82–228–78–23–118–232–201–44–43.

After a ticket is selected, it is virtually placed back into the pool of tickets (so that the same raffle ticket could be the winner in subsequent rounds after winning a raffle round). Each raffle draw is independent of the previous draw (so winning in a previous round does not decrease or increase your chance of winning in a future round). The raffle is expected to continue for the next ten years, and for simplicity, suppose that no further raffle tickets can be bought once the total supply of raffle tickets are bought. Suppose in total there are 20 billion raffle tickets, and assume throughout the rest of the problem that the system is at steady-state (so that 20 billion raffle tickets have been bought and no more will be bought). Throughout the rest of the problem, you will be asked questions about this lottery. You can either answer them with a simulation model or by finding the theoretical values based on a probability model. Clearly comment your code either way to justify your responses.

4a. (5 points) Suppose you have 10,000 raffle tickets. Then what is the chance of winning zero raffle draws within the next week (exactly seven days)? What is the chance of winning at least one raffle draw within the next week? Round to two decimal places.

4b. (5 points) How many raffle tickets would you need so that the chance of winning a raffle draw (at least one raffle draw) within a week (exactly seven days) is at least 80%? Round to the nearest ten thousand tickets.

4c. (5 points) If you have 10,000 raffle tickets and if the winner of a raffle round wins $250, then how much are you expecting to win over the next year (exactly 365 days)? Round to the nearest hundred dollars.

4d. (5 points) Imagine some of your friends are also buying raffle tickets, and each friend buys a different amount. They all want to answer the question, “How much do I expect to win over the next year?”. Make a plot of the average yearly reward amount (measured in $) as a function of how many raffle tickets you buy. The x-axis should contain values between 100 raffle tickets and 10,000,000 raffle tickets in increments of 100. It may be helpful to have the plot in log-scale (on the x-axis, so that x-axis goes from 1e+02 to 1e+07). One friend has 100,000 tickets, and another friend has 1,000,000 tickets. How much should they expect to win over the next year (exactly 365 days)? Round to the nearest thousand dollars.

4e. (5 points) Your friends come back thanking you for your calculations on their expected winnings. But they are also wondering how long on average they have to wait between winning raffle rounds. For the friend that has 100,000 raffle tickets, how long on average (measured in days) do you expect this friend to wait between winning raffle rounds? What about for your friend who has 1,000,000 tickets? Build a simulation model to find the expected waiting time between winning raffle rounds to answer this question. Round your answers to the nearest whole number. There are ways to get the answer to this question without a simulation model (a simple formula from stochastic processes theory), but try to come up with a simulation model. If you go with a simulation model, you should think about how to make your code efficient. Use system.time() to test different ways to write your code. When testing variations of code, you should start with a smaller number of simulations than you would for solving the problem.

Note: There are two reasons why we included these last few Probability and Simulation Models problems. First, we want you to notice that there are often different, independent ways to approach the same problem. If you solve a problem in different ways and they converge to the same answer, then you will become more confident that both of your approaches work. Second, we want you to understand that simulation models can be used to explore applied probability models. If the question is intractable in terms of explicit, analytical solutions (i.e., you cannot solve for the probability or expected value using pnorm() or related functions), then building a simulation to estimate the probability or expected value may be a good option.

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, and your last name. An example would be 'HW3_firstName_lastName.R' or 'HW3_firstName_lastName.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.** Also, make sure you clearly denote which problem is which problem. For example, use R code comments to write: #### Problem 2a ####. **If you find and modify a snippet of code from an external source (a YouTube video, StackExchange page, blog post, etc.), then write this in a comment before the line/chunk of code you are using from that source.**

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. Points will be taken off if the file name is not in the specified format or if you split your homework file into multiple R files.