Discrete Event Simulation - Expanding on the Coffee Shop

Group 3

Julia O’Keeffe, Paul Yaginuma, Graham Lovell, Su Hyun Byeon, & Jacob Zott

MSDS 460 - Decision Analytics

Professor Thomas W. Miller, PhD

March 2025

Problem definition:

For this project we model the process of customers visiting a coffee shop by conducting discrete event simulation (DES). We expand on the original coffee shop model by incorporating agent-based simulations where the customers can make the decision to leave at various steps in the process. By using DES, we are better able to understand the flow of events and study the impact that different key variables—such as the number of baristas or the percentage of rushed customers—have on the overall system.

Abstract:

We conducted numerous simulations that modeled the coffee shop process varying the value of key variables such as the number of baristas working and the percent of rushed customers. Overall we found that the number of baristas had the biggest impact on operations. An increase in baristas led to more customers served, fewer reneging, and shorter wait times (Figures 1 & 2). We also found that increasing the percentage of rushed customers increased the number that balked, but counterintuitively decreased reneging–likely due to balking causing shorter lines.

Introduction:

Queue-based processes are ubiquitous in daily life for everything from ordering coffee to waiting at the DMV. While seemingly trivial, there are many decision points that can impact efficiency and customer experience. Optimizing for the right balance of staff, service time, and cost can be informed by modeling tools like DES. DES is useful for analyzing these complex systems because it models each action as a discrete step that alters the state of the system and dictates what follow-on actions can occur.

We ran simulations to evaluate the sequence of events as customers enter a coffee shop, decide on whether to get in—and stay in—line, then place and receive their order. We model agent-based actions, where customers can decide to not join the line if it is too long—which we call balking—or choose to leave a line if it's taking too long—which we call reneging. This results in six distinct possible events: a customer arriving, balking or joining the line, reneging or starting service, and ending service (Figure 1).

Literature Review:

We conducted a literature review of studies on queue-based systems to learn about prior experimental results and to inform our modeling. In Kuaban 2020, the authors study healthcare waiting times and unsurprisingly conclude that increased balking and reneging reduces congestion, but increases customer losses. One interesting result they found was that reneging seemed to have a social contagion effect, where patients who observed another patient leave would be more likely to leave themselves. This created patterns of what the authors called bursty reneging where several patients would leave one after another.

Another set of interesting findings is presented in Zhang 2000, where the authors conclude that having a single queue served by multiple staff is more efficient than having separate lines. However, they also find that a single queue makes people perceive a line as being longer, which can lead to more balking as customers arrive. Finally, they conclude that customers switching between lines creates inefficiencies compared to a single line, but gives people more sense of control and a higher overall satisfaction, despite the average wait times being longer.

There is also the story of an airport that reduced complaints about luggage wait times, not by improving the system, but by switching gates to make the walk to baggage claim longer. Despite taking the same amount of time, customers spent less idle time at the baggage claim and complaints were reduced (Burkeman 2018). These findings demonstrate that customer behavior and satisfaction can be counterintuitive at times. They demonstrate the strengths of DES where agent actions may be difficult to predict, but can be observed and modeled effectively. While one strategy may be most efficient or cost optimized, there may be hidden costs in terms of customer satisfaction, which in turn can affect throughput and retention.

Research Design, Algorithms, and Modeling Methods:

To conduct the coffee shop simulations we used SimPy to create a model that tracks customers as they arrive, make decisions, and are served. We experimented with variables such as the number of baristas, service time, and the proportion of customers that were rushed. Our model generates a stream of customers, tracks them as they step through each event, and produces an event log that captures the state of the system at each step.

The model first generates a new customer based on a random wait time drawn from an exponential distribution. Each customer is then randomly assigned as rushed or relaxed based on a pre-determined distribution. A rushed customer balks if the line is longer than five while a relaxed one tolerates 10. Similarly a rushed customer reneges if the wait is more than five times the mean service time, while a relaxed one will wait up to 10x. Service time is also drawn randomly from an exponential distribution, but with minimum and maximum caps to prevent unrealistic times from appearing in the simulation.

Once all of these parameters are established the SimPy DES model generates a continuous stream of customers and processes them through the system for a simulated 10 hour period. For example, as each customer arrives, they are assigned characteristics and ‘request’ a barista from the environment. The model identifies how long the line is, tracks how long they have waited, and yields—puts their request on hold—until a barista becomes available. How many customers get served during each simulation period is determined by the various factors at play, including the frequency of customer arrivals and the randomly drawn service times.

Results, Interpretation, and Management Recommendations

After running our simulations, we identified that increasing the number of baristas had the largest impact on key business metrics. Increasing from one to two baristas increased customers served by about 200, while adding a third increased customers served by about 100. Average wait times also dropped by about 10-15 minutes per barista added. Unsurprisingly we found that this decreased wait time reduced reneging significantly from over 400 with one barista to under 50 with three baristas (Figures 1 & 2).

We also identified that an increase in the percentage of rushed customers increases balking, reduced reneging, and slightly reduces the number of customers served (Figures 4 & 5). Increasing the percentage of customers that were rushed by 25% led to about 30 additional customers balking, with the largest jump coming between 50% and 75%. Interestingly, this led to a decrease in reneging as the long lines drove customers away instead of long waits.

Based on our findings we would make three recommendations. First, increase the number of baristas from one to two, which gives the biggest return in terms of customers served and likely profit. Second, add an additional barista during peak periods when the percentage of people rushed is over 50%, which could be measured using surveys. Finally, we would recommend collecting feedback to identify two things 1) whether balking or reneging more negatively impacts long-term customer retention and 2) the impact of using a single feeder queue vs. multiple queues on customer satisfaction. If balking is less damaging to a business than reneging–since leaving immediately might be less frustrating than waiting then leaving –optimizing for less balking may be preferable. Based on the findings in Zhang 2000, it's likely that a single feeder queue would achieve a better balance of balking to reneging and lead to greater customer satisfaction overall, despite slightly longer wait times.

Appendix:

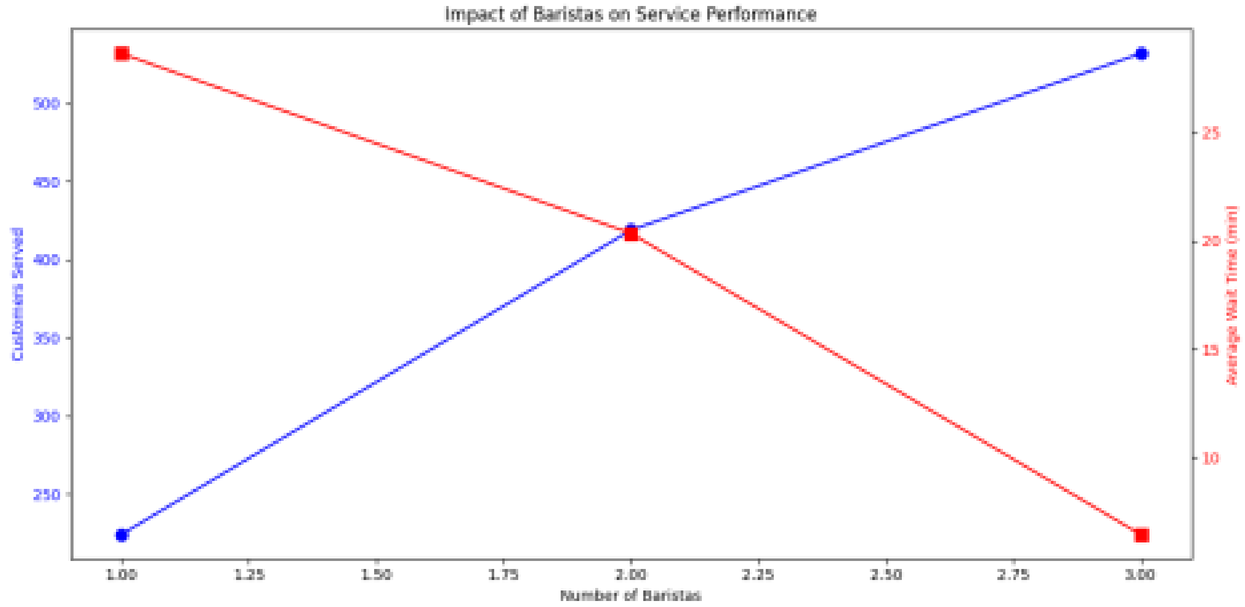

Figure 1: Customers served and average wait times vs. number of baristas

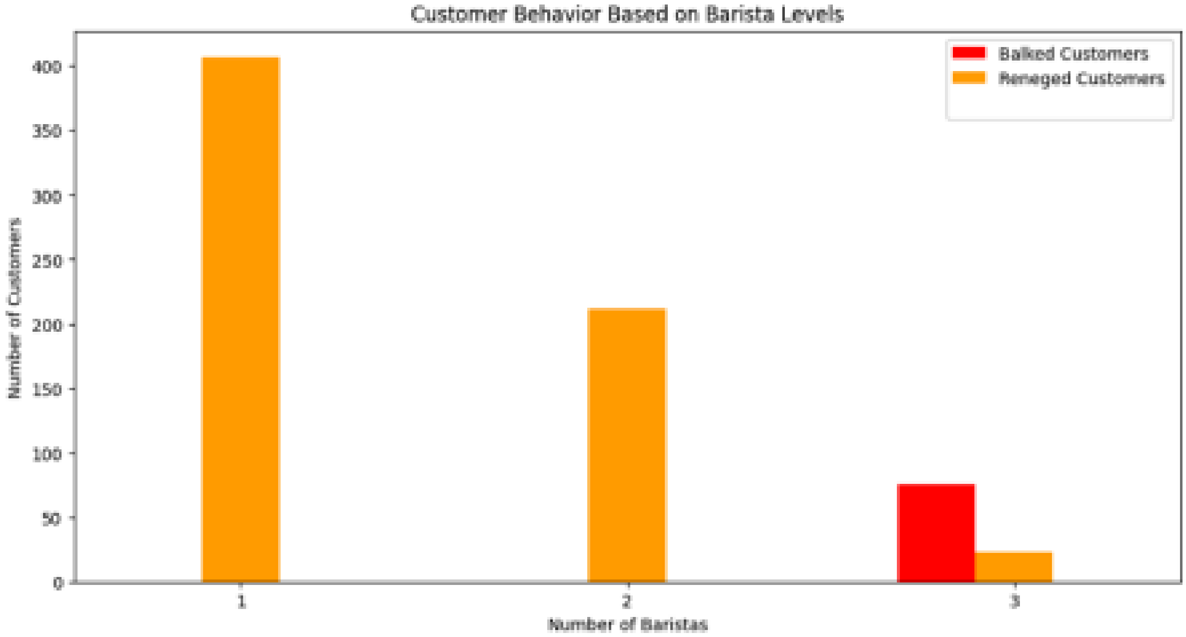

Figure 2: Number of customers who reneged or balked vs. number of baristas

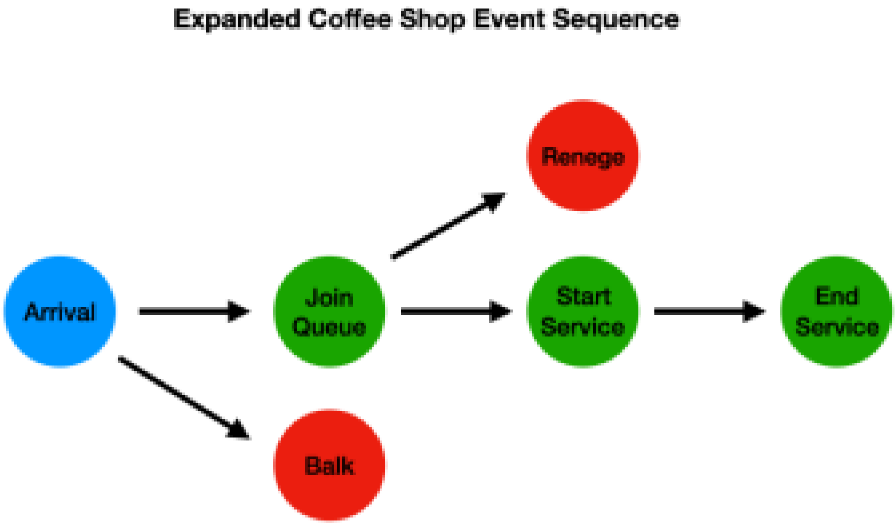

Figure 3: Expanded Coffee Shop Event Sequence

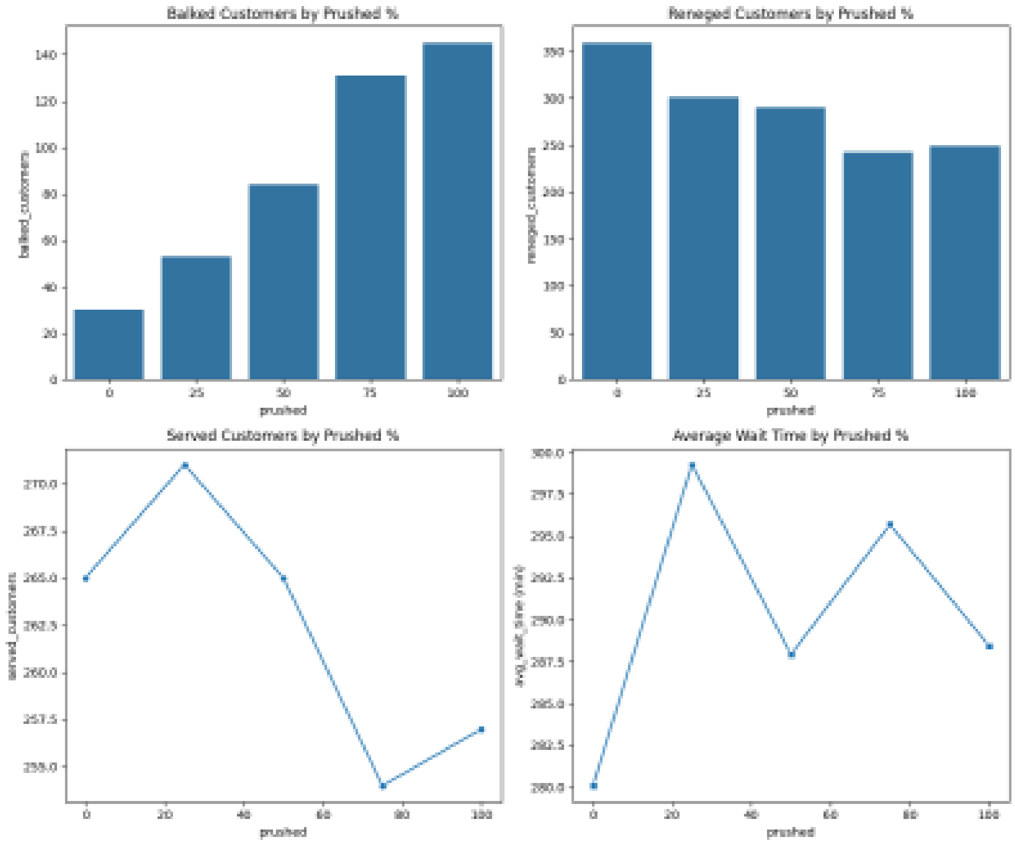

Figure 4: Number of customers that balked, reneged, were served, and average wait times vs. percent of rushed customers

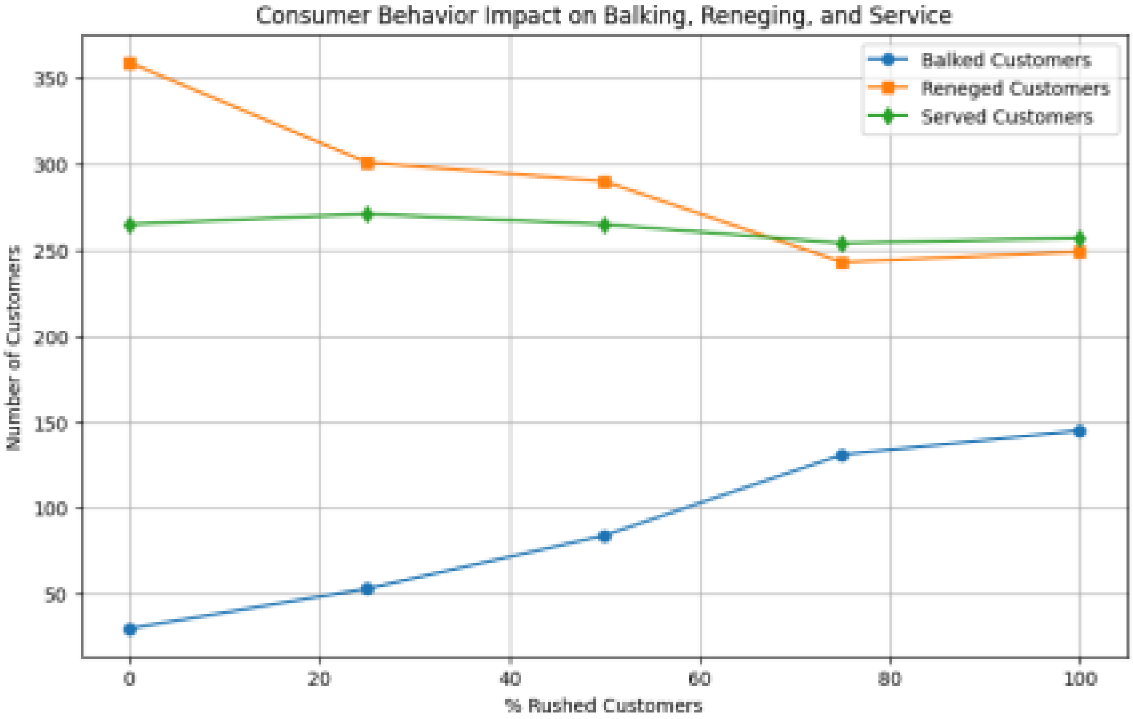

Figure 5: Number of customers served, balked, and reneged vs. percent rushed

References:

Kuaban 2020

Kuaban, G. S., Kumar, R., Soodan, B. S., & Czekalski, P. (2020). A multi-server queuing model with balking and correlated reneging with application in health care management. IEEE Access, 8, 169122–169136. https://doi.org/10.1109/ACCESS.2020.3024259

Zhang 2000

Zhang, L. J., Ng, W. W. J. L., & Tay, S. C. (2000). Discrete-event simulation of queuing systems. Proceedings of the Sixth Youth Science Conference, Ministry of Education, Singapore. Retrieved from https://phyweb.physics.nus.edu.sg/~phytaysc/articles/queue.pdf

Burkeman 2018

Burkeman, O. (2018, September 7). How a longer walk to baggage reclaim cut complaints. The Guardian.
https://www.theguardian.com/lifeandstyle/2018/sep/07/how-to-beat-bottlenecks-oliver-burkeman