Discrete Event Simulation Challenge

Cafe Operations: Bottleneck Identification and Service Improvements

Author: Gabriel Carvalho Domingos da Conceição

LinkedIn: https://www.linkedin.com/in/gabriel-carvalho-conceicao/
Repository: https://github.com/gcarvalho-DataAI/discrete-event-simulation

Date: February 6, 2026

Contents

List of Tables

List of Figures

1 Abstract

This report presents a discrete-event simulation model for a small cafe with separate queues for ordering, preparation, and pickup. Customers are segmented into fast, medium, and slow types with distinct arrival rates, service times, and patience thresholds. The simulation captures peak behavior, identifies bottlenecks, and compares scenarios across multiple datasets. The primary objective is to locate the dominant bottleneck under peak conditions and propose operational improvements.

2 Introduction

Service systems such as cafes experience short periods of high demand where small capacity mismatches generate large delays. Discrete-event simulation (DES) is well suited for this environment because state changes occur at discrete events: arrivals, service start, service end, and abandonment.

3 Theoretical Background

The cafe can be modeled as a multi-stage queueing system with heterogeneous customers and reneging. Key foundations:

- **Queueing theory:** arrival processes are modeled as Poisson processes, leading to exponential inter-arrival times. Service times are modeled using distributions such as triangular or lognormal.
- **Reneging (impatient customers):** customers may abandon the queue after exceeding a patience threshold, a common extension to M/G/c systems.
- **State-dependent service:** preparation time can increase with queue length due to congestion and human workload effects.
- **DES methodology:** systems evolve via events, and performance is measured via throughput, utilization, and waiting times.

Representative references are listed in the References section.

4 Model Specification

4.1 System Components

- Queues: ordering, preparation, pickup.
- Resources: order attendants, baristas (preparation), pickup attendant.
- Customer types: fast, medium, slow.

Note on queues: the model keeps two customer-visible queues (order and pickup) as stated in the prompt. The “preparation queue” represents internal WIP/backlog that affects preparation time, matching the requirement that service time varies with preparation load.

4.2 Event Flow

- Arrival → order queue → preparation queue → pickup queue → exit.
- Reneging occurs if waiting time in the order queue exceeds customer patience.
- Preparation time increases with queue length by factor $(1 + \alpha \cdot q)$.

4.3 Metrics

The simulation reports:

- Abandonment rate.
- Average waiting times per stage.
- Total time in system.
- Resource utilization.
- Average queue lengths.

---

5 Simulation Setup

- Horizon: 4 hours of operation with 0.5 hour warm-up (default).
- Inter-arrival times: exponential by customer type.
- Service times: triangular around the type mean.
- Patience: exponential by customer type.
- Congestion: prep time multiplied by $(1 + \alpha \cdot q)$.
- Random seed: 42 (default).

---

6 Data and Instance Generation

Three public datasets were used to estimate arrival profiles and product mix:

- Maven Analytics (Coffee Shop Sales): `https://mavenanalytics.io/data-playground/coffee-shop-sales`
- Hugging Face (CoffeeSales): `https://huggingface.co/datasets/tablegpt/CoffeeSales`
- Kaggle (Coffee Sales Dataset): `https://www.kaggle.com/datasets/saadaliyaseen/coffee-sales-dataset`

ussy
ormalsize
Story of instance collection and generation:

- The Maven dataset was downloaded as an XLSX file and parsed to extract transaction timestamps and product names.
- The Hugging Face dataset was obtained as a CSV (vending-machine transactions) and used to validate arrival patterns.
- The Kaggle dataset was downloaded manually as a ZIP, then extracted to CSV with date and time columns.

- Product names were mapped into fast/medium/slow categories using keyword rules.
- Peak-hour rates and product mix were computed per dataset and then used to generate synthetic peak scenarios.

From each dataset, peak-hour rates and product mix were derived to generate instances. For each dataset, 180 instances were created by varying:

- Demand scale: $0.7, 0.85, 1.0, 1.15, 1.3$
- Staffing combinations (order/barista/pickup)
- Customer patience levels
- Preparation congestion factor α

Additional synthetic instances were generated by combining multiple staffing and patience levels with congestion sensitivity, in order to stress-test the system under different peak conditions.

---

7 Environment and Reproducibility

OS	Linux 6.8.0-94-generic
Machine	x86$_6$4
CPU	Intel(R) Core(TM) i9-14900K
RAM	62Gi
Python	3.12.3
UV	uv 0.9.28 (0e1351e40 2026-01-29)

This table records the execution environment to ensure reproducibility across machines and future runs.

Usage:

```
uv venv
uv pip install -r requirements.txt
uv run python src/cafe_sim.py
uv run python scripts/build_report.py
pdflatex reports/cafe_sim_report.tex
```

---

8 Solver and Simulation Choices

This problem is a discrete-event simulation, not a mathematical optimization solved by a MILP/CP solver. The core engine is SimPy (process-based DES), which is appropriate for queueing systems with reneging and state-dependent service times.

- Solver choice: not applicable (SimPy event scheduling).
- Event calendar: SimPy event queue.
- Randomness: exponential inter-arrival, triangular service times, exponential patience.
- SimPy: `https://simpy.readthedocs.io/en/latest/`

---

9 Aggregate Results

This table reports the average performance metrics across datasets. Maven clearly shows higher delays and abandonment, indicating peak stress, while Hugging Face and Kaggle remain low-load baselines.

Table 1: Aggregate metrics by dataset (mean across instances).

10 Bottleneck Analysis

The preparation stage consistently dominates queueing time in high-demand scenarios. In the Maven dataset, preparation utilization approaches saturation and abandonment becomes significant. Kaggle and Hugging Face datasets remain below capacity and serve as baseline (low-stress) scenarios. This indicates that improvements should focus on preparation capacity and variability rather than on ordering or pickup.

sectionPlots by Dataset

Maven

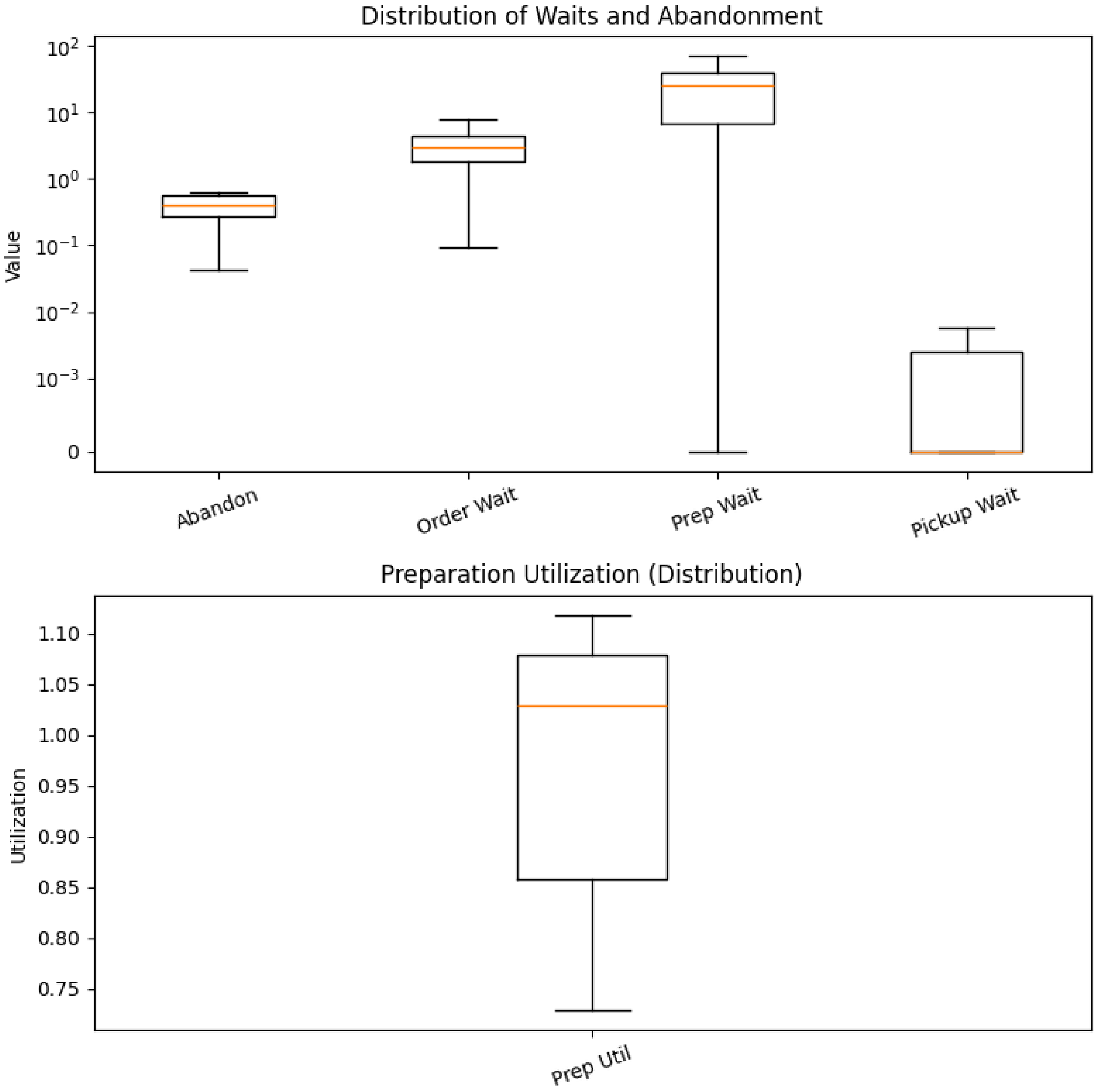


Figure 1: Maven: distributions of waits and abandonment

- Variability is high for preparation waits and abandonment under peak-like demand.
- P95 values highlight extreme congestion not visible in averages.
- Preparation utilization is consistently near saturation in worst cases.

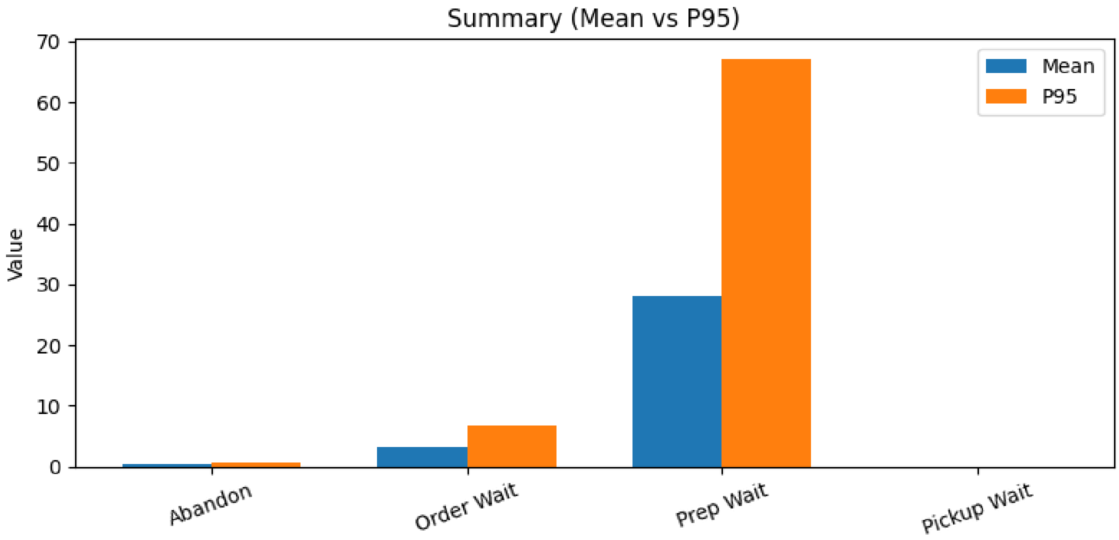


Figure 2: Maven: summary (mean vs P95)

Hugging Face

- Waits and abandonment remain near zero; distributions are compressed.
- P95 values confirm low stress across scenarios.
- Utilization stays far below saturation.

Kaggle

- Metrics indicate low congestion across scenarios.
- P95 values remain close to means, indicating stable service.
- Utilization is low, consistent with non-peak demand.

11 Top-10 Scenarios (Maven)

Top-10 is computed only from Maven instances to reflect peak-like demand. The score is:

$$Score = 100 \cdot Abandon + 2 \cdot T_{system} + W_{order} + W_{prep} + 0.5 \cdot W_{pickup}$$

Lower is better. This table summarizes the best-performing Maven scenarios under peak-like demand.

Table 2: Top-10 scenarios (Maven only).
resizebox!

Rank	Instance	Dataset	Abandon	Total Time	Prep Wait	Prep Util
1	maven_coffee_shop_sales_025.json	Maven (Coffee Shop Sales)	0.313	6.690	0.000	0.729
2	maven_coffee_shop_sales_026.json	Maven (Coffee Shop Sales)	0.313	6.690	0.000	0.729
3	maven_coffee_shop_sales_027.json	Maven (Coffee Shop Sales)	0.313	6.690	0.000	0.729
4	maven_coffee_shop_sales_019.json	Maven (Coffee Shop Sales)	0.366	6.444	0.000	0.764
5	maven_coffee_shop_sales_020.json	Maven (Coffee Shop Sales)	0.366	6.444	0.000	0.764
6	maven_coffee_shop_sales_021.json	Maven (Coffee Shop Sales)	0.366	6.444	0.000	0.764
7	maven_coffee_shop_sales_055.json	Maven (Coffee Shop Sales)	0.384	6.232	0.000	0.759
8	maven_coffee_shop_sales_056.json	Maven (Coffee Shop Sales)	0.384	6.232	0.000	0.759
9	maven_coffee_shop_sales_057.json	Maven (Coffee Shop Sales)	0.384	6.232	0.000	0.759
10	maven_coffee_shop_sales_022.json	Maven (Coffee Shop Sales)	0.379	6.910	0.000	0.762

The results indicate which staffing and congestion combinations minimize abandonment and system time.

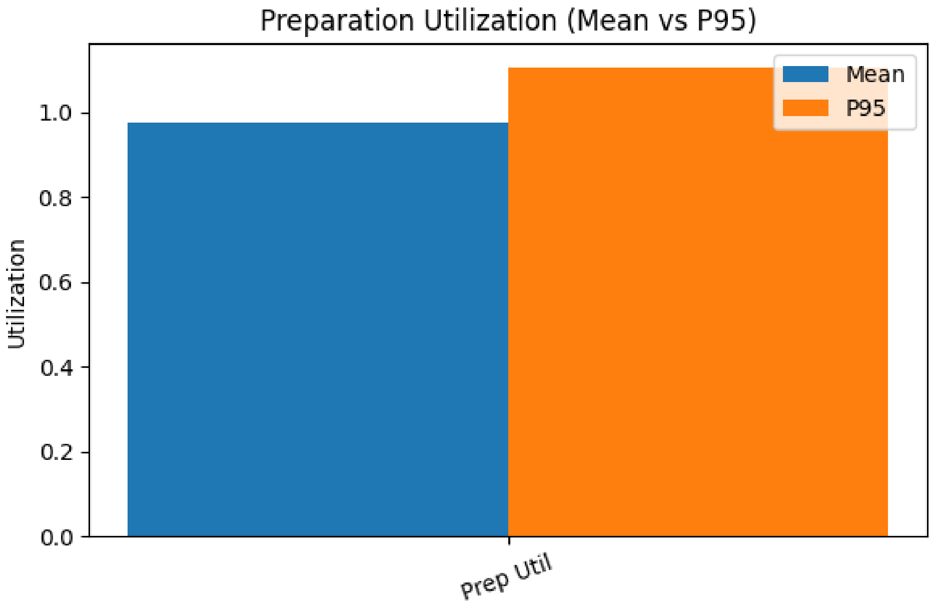


Figure 3: Maven: preparation utilization summary

12 Conclusions

The simulation confirms that preparation is the dominant bottleneck under peak-like demand, while ordering and pickup contribute marginally to total delay. Maven-derived scenarios show high utilization and abandonment, indicating that staffing or process improvements should prioritize the preparation stage. Kaggle and Hugging Face scenarios remain low-load baselines and should be scaled if used to emulate peak conditions.

13 Recommendations

- Increase preparation capacity (additional barista) or reduce prep variability.
- For Maven-like peak conditions, focus on prep queue control rather than order or pickup.
- For Kaggle/HF datasets, scale arrival rates to simulate true peak demand.

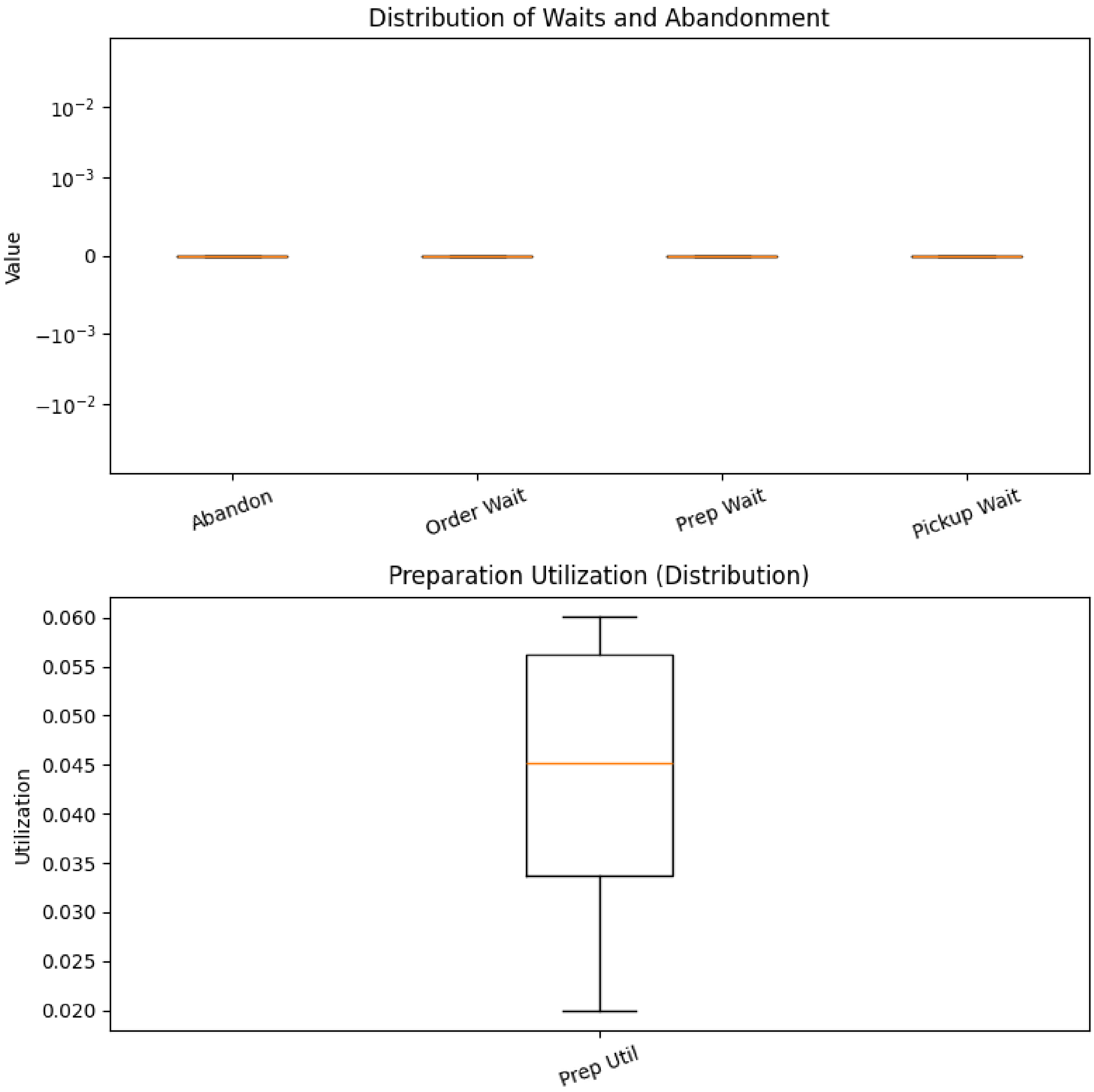


Figure 4: Hugging Face: distributions of waits and abandonment

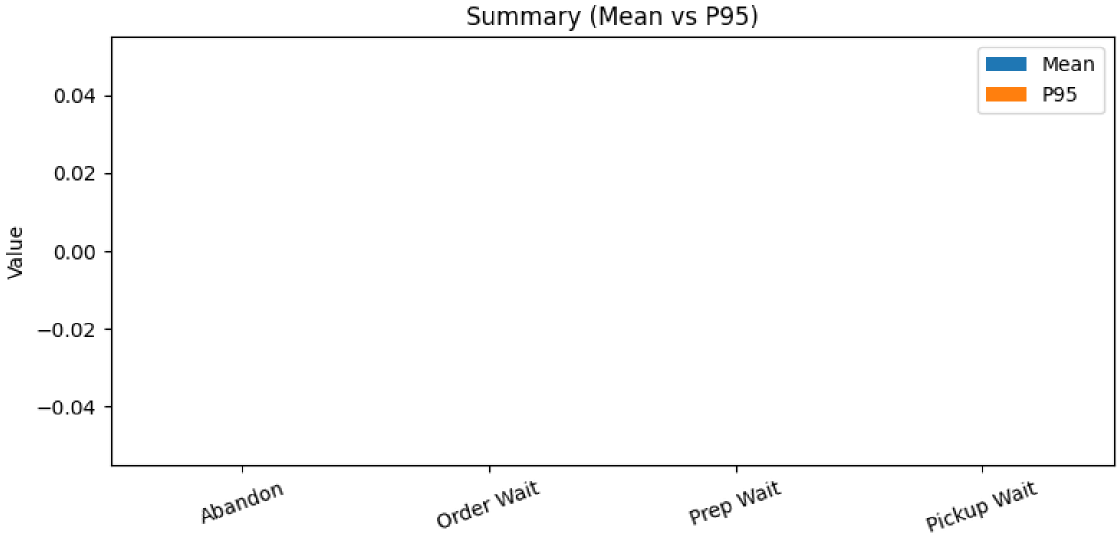


Figure 5: Hugging Face: summary (mean vs P95)

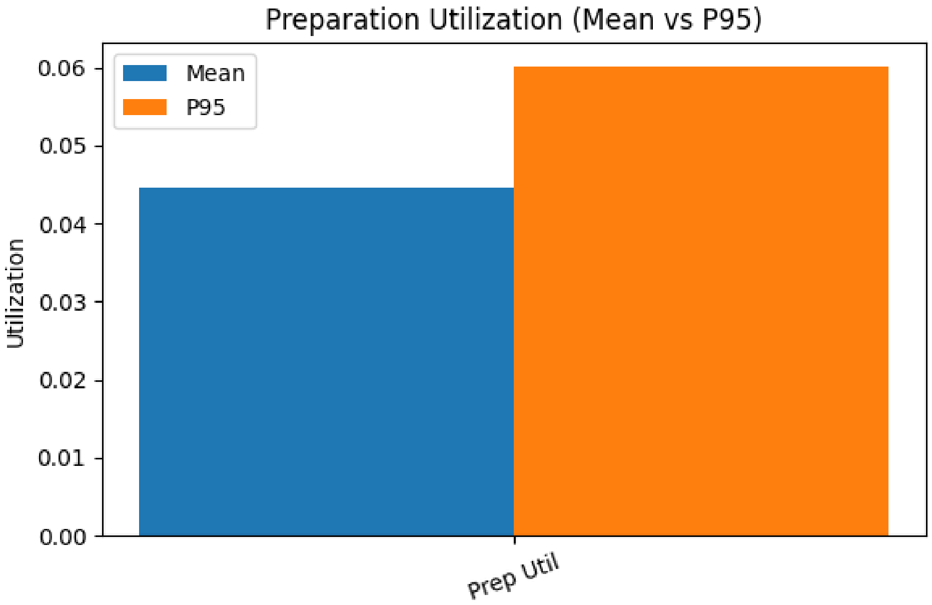
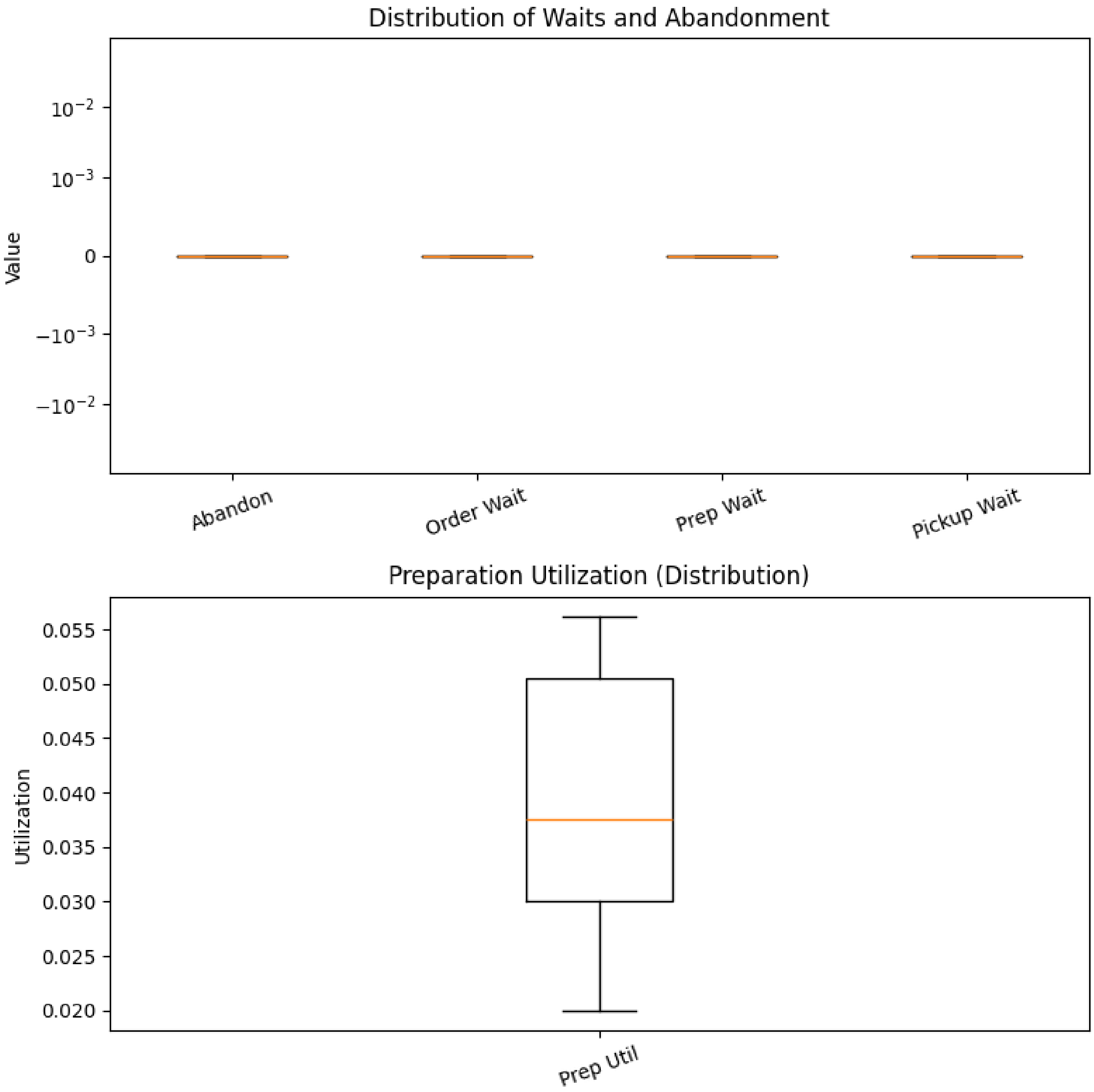


Figure 6: Hugging Face: preparation utilization summary



Distribution of Waits and Abandonment

Preparation Utilization (Distribution)

Figure 7: Kaggle: distributions of waits and abandonment

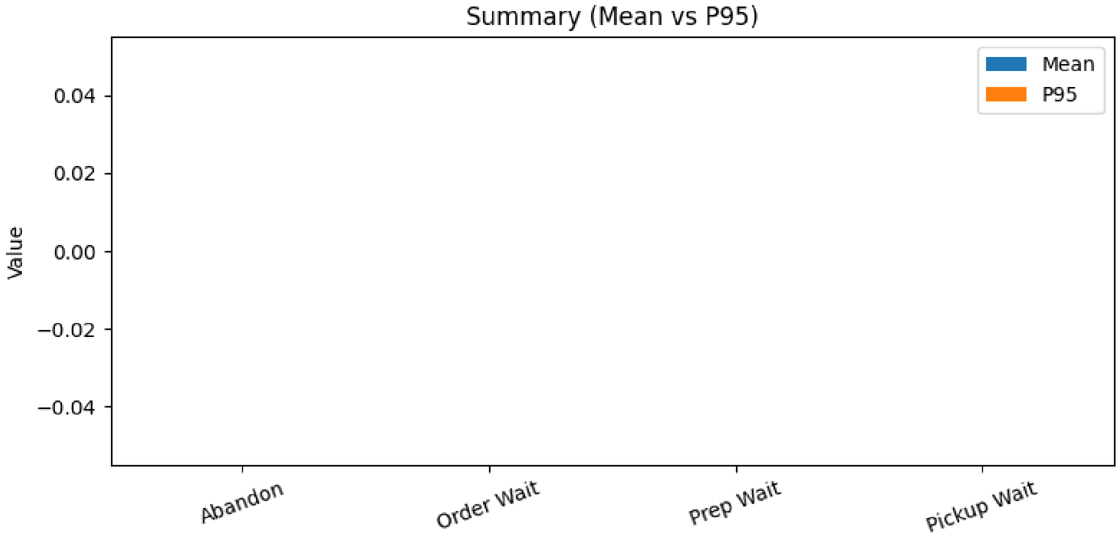


Figure 8: Kaggle: summary (mean vs P95)

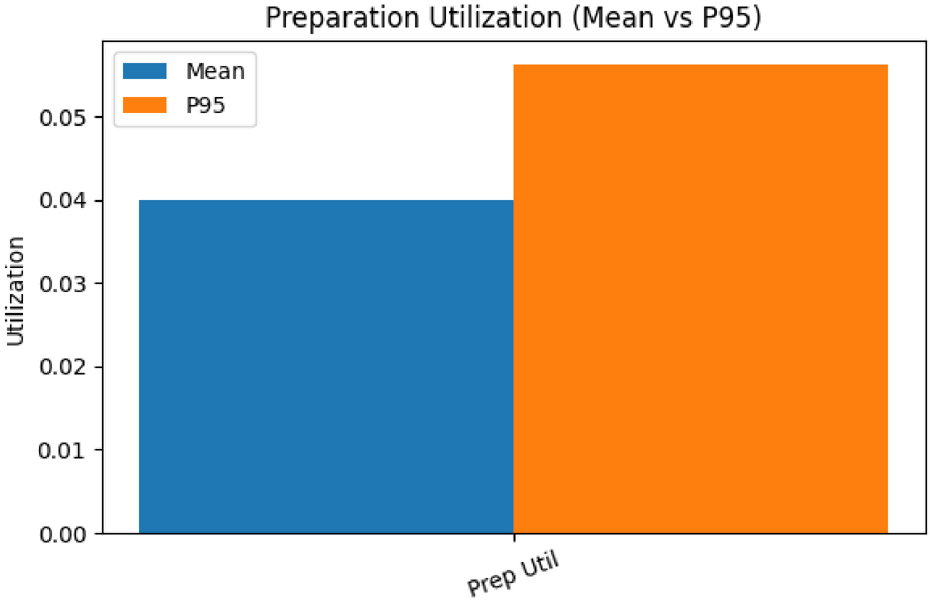


Figure 9: Kaggle: preparation utilization summary

14 References

- Banks, J., Carson, J. S., Nelson, B. L., & Nicol, D. M. (2010). *Discrete-Event System Simulation.* Pearson.
- Law, A. M. (2015). *Simulation Modeling and Analysis.* McGraw-Hill.
- Kleinrock, L. (1975). *Queueing Systems, Volume 1: Theory.* Wiley.
- Baccelli, F., & Hebuterne, G. (1981). On Queues with Impatient Customers.
- Zohar, E., Mandelbaum, A., & Shimkin, N. (2002). Adaptive Behavior of Impatient Customers in Queues.
- George, J. M., & Harrison, J. M. (2001). Dynamic Control of a Queue with Variable Service Rate.
- KC, D. S., & Terwiesch, C. (2009). Impact of Workload on Service Time and Quality: An Analysis of Hospital Operations.