ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 6 No. 1 (2026) 185-191
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Analysing Customer Waiting Time Using Queuing Models

Hafizuddin Saari[1], Norzuria Ibrahim[1]*,

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA.*

*Corresponding Author: norzuria@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2026.06.01.023

Article Info

Received: 4 January 2026
Accepted: 23 January 2026
Available online: 9 July 2026

Keywords

Restaurant, Queuing Models, Single Server, Waiting Time, Server

Abstract

Customer experience plays a crucial role in determining the long-term success of service-oriented businesses and waiting time is one of the most influential factors affecting customer satisfaction. Long queues, delays in seating and slow order processing often lead to frustration, negative reviews and reduced customer retention, which ultimately harm a business's reputation and revenue. As organizations strive to balance efficient operations with high-quality service, reducing waiting time becomes a key priority especially during high-volume periods or when staffing levels are limited. This study applies queuing theory to analyse customer waiting times in two different service environments: a coffee shop operating under a single-server structure and a convenience store operating under a multi-server structure. Data were collected over one week in both locations and analysed using MATLAB to compute performance measures such as arrival rate, service rate, server utilization and average waiting time. The findings highlight how queuing models can be adapted to different service configurations, offering data-driven insights into managing queues more effectively and supporting operational improvements tailored to each environment.

1. Introduction

This study focuses on analysing customer waiting times in restaurants through the application of queuing models specifically the single-server model at a coffee shop and multi-server model at a convenience store Waiting is a common challenge in daily life and delays frequently occur in service environments such as restaurants, airports and banks [1]. These delays affect both customers and staff, making it essential for businesses to adopt systematic approaches that balance efficiency, service quality and profitability [2]. As competition increases and customer patience decreases the ability to manage queues effectively becomes a critical factor for sustaining service performance.

Queuing theory is a branch of mathematics that studies systems in which customers, tasks or data wait in line to receive service [3]. It provides a structured framework for analysing arrival patterns, service mechanisms and queue configurations, enabling researchers and practitioners to understand system behaviour and identify opportunities for improvement. The theory was first introduced by Erlang in 1917 to address congestion in telephone networks and has since become a widely used analytical tool in telecommunications, manufacturing, healthcare and hospitality [4].

1.1 Data Collection

The data was collected within 1 week at Coffee Shop and Convenience Store. The observations were conducted from Monday to Sunday covering weekdays and weekends.

2. Methodology

The methodology for this study is based on the application of queuing models to analyse customer waiting times at two selected service locations. Since the observed coffee shop operated with one primary service point during the data collection period the M/M/1 model was identified as the most appropriate representation of its real operating conditions. The convenience store operated with two service counters therefore the M/M/2 model was applied to represent its multi-server. The aim of this methodological approach is to evaluate how customer arrivals and service processes influence waiting times and overall system performance, allowing the study to determine whether the existing service structures are efficient or require improvement.

The M/M/1 model assumes that customer arrivals follow a Poisson distribution and that service times follow an exponential distribution which aligns with typical behaviour in restaurant and café environments. The M/M/2 model extends this framework by considering two parallel servers each serving customers independently under the same assumptions of Poisson arrivals and exponential service times. Customer arrival rates and service rates were collected through direct observation recorded at consistent intervals and processed using MATLAB.

The analysis focuses on interpreting these performance measures to identify bottlenecks inefficiencies and opportunities for improvement. This methodological approach ensures that the findings are grounded in established queuing theory principles while remaining directly applicable to the real operational conditions of both the coffee shop and the convenience store.

2.1 Equations of Single Server Model M/M/1

The arrival and service time is exponentially distributed in the M/M/1 queuing model. The following factors will be explored for the counter M/M/1 queuing model analysis:

Customer arrival rate, λ

Service rate, μ

Server utilization, ρ

$$\rho = \frac{\lambda}{\mu} \tag{1}$$

Expected number of customers, L_s

$$L_s = \frac{\rho}{1 - \rho} \tag{2}$$

Average number of customers waiting in queue, L_q

$$L_q = \frac{\rho^2}{1 - \rho} \tag{3}$$

Average waiting time in system, W_s

$$W_s = \frac{1}{\mu - \lambda} \tag{4}$$

Average waiting time in queue, W_q

$$W_q = \frac{\rho}{\mu(1 - \rho)} \tag{5}$$

2.2 Equations of Multi Server Model M/M/2

The arrival and service time is exponentially distributed in the M/M/2 queuing model The following factors will be explored for the counter M/M/2 queuing model analysis:

Customer arrival rate, λ

Service rate, μ

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) p. 185-191

Server utilization, ρ

$$\rho = \frac{\lambda}{\mu} \tag{6}$$

Probability of no customers, P_0

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!}\left(\frac{1}{1-\frac{\rho}{c}}\right)\right]^{-1}, \frac{\rho}{c} < 1 \tag{7}$$

Expected number of customers, L_s

$$L_s = L_q + \rho \tag{8}$$

Average number of customers waiting in queue, L_q

$$L_q = \frac{\rho^{c+1}}{(c-1)!\,(c-\rho)^2} P_0 \tag{9}$$

Average waiting time in system, W_s

$$W_s = \frac{L_s}{\lambda} \tag{10}$$

Average waiting time in queue, W_q

$$W_q = \frac{L_q}{\lambda} \tag{11}$$

2.3 Data

The data is collected within 1 week at Coffee Shop. The observations are conducted from Monday to Sunday covering weekdays and weekends.

Table 1 Number of customers arrived at Coffee Shop

Date	Day	Number of Customer		Total Customer
		10am-12pm	6pm-8pm	
1-Aug	Monday	29	38	47
2-Aug	Tuesday	14	25	39
3-Aug	Wednesday	17	16	33
4-Aug	Thursday	18	18	36
5-Aug	Friday	16	28	44
6-Aug	Saturday	39	34	73
7-Aug	Sunday	37	32	69

Table 1 shows the total customers recorded from 1 August 2025 to 7 August 2025 at the Coffee Shop in Kuala Lumpur. The collected data is analysed using MATLAB software.

Table 2 Number of customers arrived at Convenience Store

Date	Day	Number of Customer		Total Customer
		10am-12pm	6pm-8pm	
3-Nov	Monday	20	14	34
4-Nov	Tuesday	23	20	43
5-Nov	Wednesday	25	21	46

6-Nov	Thursday	33	18	51
7-Nov	Friday	22	18	40
8-Nov	Saturday	24	24	48
9-Nov	Sunday	20	16	34

Table 2 shows the total customers recorded from 3 November 2025 to 9 November 2025 at the convenience store. The collected data is analysed using MATLAB software.

3. Results and Discussion

Table 3 MATLAB Software Result for Coffee Shop

Date	Arrival Rate, (λ)	Service Rate, (μ)	Average Waiting Time In System (Ws)	Server Utilization, (ρ)
1-Aug	0.2792	0.5630	3.52 minutes	49.58%
2-Aug	0.1625	0.3824	4.55 minutes	42.50%
3-Aug	0.1375	0.4583	3.12 minutes	30.00%
4-Aug	0.1500	0.4091	3.86 minutes	36.67%
5-Aug	0.1833	0.5176	2.99 minutes	35.42%
6-Aug	0.3042	0.4740	5.89 minutes	64.17%
7-Aug	0.2875	0.4313	6.96 minutes	66.67%
Average	0.2149	0.4622	4.41 minutes	46.43%

Table 3 shows the result of a single-server queuing system. The performance of the single-server queuing system from 1 to 7 August 2025 shows that the system operated under stable and efficient conditions. Throughout the week, the service rate consistently exceeded the arrival rate, with averages of 0.4622 and 0.2149 customers per minute respectively, ensuring that the server was never at risk of overload. Server utilization ranged from 30.00% to 66.67%, reaching its lowest level on 3 August and peaking on 7 August. These fluctuations reflected natural variations in customer arrival patterns but remained within a moderate range, resulting in a weekly average utilization of 46.43%. This level of utilization indicates that the server had sufficient idle capacity to handle demand smoothly without congestion.

These utilization patterns directly influenced customer waiting times in the system. The shortest waiting time, 2.40 minutes, occurred on 5 August when utilization was relatively low, while the longest waiting time of 6.96 minutes was recorded on 7 August, coinciding with the highest utilization. On average, customers spent 4.41 minutes in the system, showing that delays were generally minimal and manageable. Overall, the results demonstrate that the queuing system performed reliably throughout the week, maintaining smooth service flow and ensuring that customers experienced short and predictable waiting times.

The line graph illustrates the pattern of server utilization over the seven days from 1 to 7 August, showing noticeable fluctuations in how heavily the server was used throughout the week. The utilization values begin at a moderate level of 49.58% on 1 August before dipping to 42.50% and then reaching the lowest point of 30.00% on 3 August. This early-week decline suggests that the system experienced relatively light demand during these days. Afterward, the utilization gradually rises again, moving through 36.67% and 35.42% on 4 and 5 August, indicating a slight recovery in system load as the week progresses.

A more significant shift occurs toward the end of the week, where server utilization increases sharply to 64.17% on 6 August and peaks at 66.67% on 7 August. This upward trend, also reflected in the dotted trend line, suggests growing demand or higher customer arrival rates during the final days. The overall pattern highlights a transition from low to high utilization, demonstrating how system load can vary substantially within a single week.

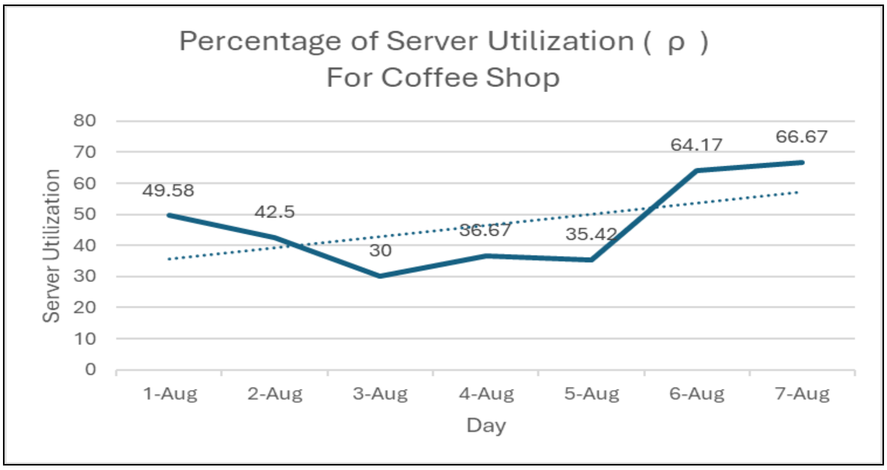

Fig. 1 Graph Percentage of Server Utilization (ρ) For Coffee Shop

Table 4 MATLAB Software Result for Convenience Store

Date	Arrival Rate, (λ)	Service Rate, (μ)	Average Waiting Time In System (Ws)	Server Utilization, (ρ)
3-Nov	0.1417	0.5667	1.79 minutes	12.50%
4-Nov	0.1792	0.5584	1.84 minutes	16.04%
5-Nov	0.1917	0.5349	1.93 minutes	17.92%
6-Nov	0.2125	0.5152	2.03 minutes	20.62%
7-Nov	0.1667	0.5714	1.79 minutes	14.58%
8-Nov	0.2000	0.6000	1.71 minutes	16.67%
9-Nov	0.1833	0.5366	1.91minutes	17.08%
Average	0.1822	0.5547	1.86 minutes	16.48%

Table 4 shows the result of a multi-server queuing system at the convenience store. The performance of the multi-server system from 3 to 9 November indicates that the service structure operated under highly stable and efficient conditions throughout the week. The average service rate of 0.5547 customers per minute consistently exceeded the average arrival rate of 0.1822 customers per minute ensuring that the servers were never at risk of overload. Server utilization remained low ranging from 12.50% to 20.62% with a weekly average of 16.48%. These values reflect light demand relative to service capacity resulting in substantial idle time across servers but guaranteeing smooth handling of customer arrivals without congestion.

These utilization levels directly influenced customer waiting times in the system. The shortest waiting time 1.71 minutes occurred on 8 November when utilization was relatively low while the longest waiting time of 2.03 minutes was recorded on 6 November coinciding with the highest utilization. On average customers spent 1.86 minutes in the system showing that delays were minimal and highly manageable. Overall, the results demonstrate that the multi-server queuing system performed reliably throughout the week maintaining efficient service flow and ensuring that customers experienced consistently short and predictable waiting times.

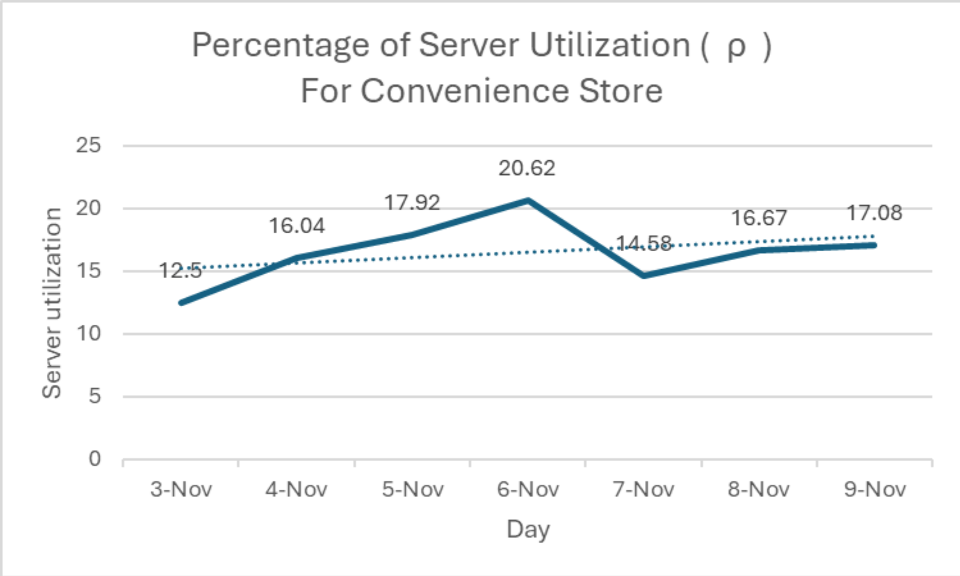

Fig. 2 Graph Percentage of Server Utilization (ρ) For Convenience Store

Fig. 2 illustrates the pattern of server utilization over the seven days from 3 to 9 November showing noticeable fluctuations in how heavily the servers were used throughout the week. The utilization values begin at a relatively low level of 12.50% on 3 November before rising steadily to 16.04% on 4 November and 17.92% on 5 November. This gradual increase suggests that customer demand was built during the early part of the week.

A more significant change occurs on 6 November when utilization reaches its highest point of 20.62% indicating the busiest day of the observation period. Afterward the values decline slightly to 14.58% on 7 November before recovering again to 16.67% on 8 November and 17.08% on 9 November. The dotted trend line reflects this overall upward movement suggesting a general increase in demand as the week progressed.

4. Conclusion

This study investigated customer waiting times in two service environments using queuing theory models. For the coffee shop, the single-server M/M/1 model was applied to evaluate performance. The results show that the system remained stable throughout the week with the service rate consistently exceeding the arrival rate. Server utilization ranged from 30.00% to 66.67% which indicates that the server had sufficient idle capacity to prevent congestion. Waiting time followed a similar pattern with shorter delays on low utilization days and longer delays when utilization increased. On average customers spent 4.41 minutes in the system demonstrating that the single-server setup was generally able to manage demand efficiently. These findings suggest that the coffee shop's current service structure performs reliably although higher demand periods may benefit from operational adjustments to maintain consistently low waiting times

For the convenience store, the multi-server M/M/2 model was used to capture the dynamics of its two service counters. The results show that utilization levels were consistently low ranging between 12.50% and 20.62% which indicates that the system had ample capacity to handle customer demand. Waiting times remained short throughout the week with values between 1.71 minutes and 2.03 minutes. The overall average waiting time was 1.86 minutes confirming that customers were served quickly and efficiently. The dotted trend line in the utilization graph highlights a gradual upward movement suggesting a general increase in demand as the week progressed. Overall, the dual-server structure ensured smooth service flow with minimal delays and consistently manageable system load.

The findings demonstrate that both single-server and multi-server structures were effective in their respective contexts. The coffee shop's single-server system managed demand efficiently under moderate utilization while the convenience store's dual-server system provided ample capacity and consistently short waiting times. These results highlight the importance of aligning service structures with customer arrival patterns to maintain efficiency and customer satisfaction.

Acknowledgement

The author would like to thank to Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia for the support.

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) p. 185-191

Conflict of Interest

Authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Hafizuddin Saari, Norzuria Ibrahim;* ***data collection****: Hafizuddin Saari;* ***analysis and interpretation of results****: Hafizuddin Saari, Norzuria Ibrahim;* ***draft manuscript preparation****: Hafizuddin Saari, Norzuria Ibrahim. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] A. Afolalu, K. O. Babaremu, S. O. Ongbali, A. A. Abioye, A. Abdulkareem, and S. B. Adejuyigbe, "Overview impact of application of queuing theory model on productivity performance in a banking sector," Journal of Physics: Conference Series, vol. 1378, no. 3, 2019. doi: 10.1088/1742-6596/1378/3/032033.

[2] J. J. Patel, "Minimize the waiting time of customer and gain more profit in restaurant using queuing model," International Journal of Engineering Research and Applications, vol. 2, no. 3, pp. 1422–1425, 2012.

[3] H. A. Taha, Operations Research, 10th ed. Pearson Education, 2017.

[4] P. O. Peter and R. Sivasamy, "Queueing theory techniques and its real applications to health care systems – outpatient visits," International Journal of Healthcare Management, 2019. doi: 10.1080/20479700.2019.1616890.

[5] Z. Motamedi, A. Ghodratnama, and S. H. R. Pasandideh, "Queueing theory application in food supply chain management: A classified review," Journal of Quality Engineering and Production Optimization, vol. 5, no. 2, 2020. doi: 10.22070/jqepo.2021.5627.1162.