APPLICATION OF QUEUING THEORY FOR OPTIMAL CUSTOMER CENTRICITY TO THE BANKING SECTOR IN KENYA

SAMWEL KISIANG'ANI JUMA

A Thesis Submitted to the Graduate School in Partial Fulfilment of the Requirements for the Award of the Degree of Master of Science in Applied Statistics of Chuka University

CHUKA UNIVERSITY

JUNE 2023

DECLARATION AND RECOMMENDATION

Declaration

This thesis is my original work and has not been submitted for award of diploma or conferment of degree in any University.

Signature..... Date 30/06/2023

Samwel Kisiang'ani Juma
SM18/39923/19

Recommendation

This thesis has been examined, passed and submitted with our approval as the University supervisors.

Signature..... Date 30/06/2023

Dr. Adolphus Wagala, PhD,
Bomet University College.
(A Constituent College of Moi University).

Signature..... Date 18/07/23

Dr. Gladys Njoroge, Ph.D.,
United States International University Africa.

DEDICATION

I dedicate this work to my parents, Mr. Juma Khisa, Mrs. Rosemary Anyango, wife, Mauline Nyarotso, and children Whiney Anyango, Joaness Juma, and Winnie Evelyn, for their prayers, support, and encouragement. May the Almighty God bless you abundantly in your endeavors.

ACKNOWLEDGEMENTS

First and foremost, I thank the almighty God for his grace and provision throughout my studies to the completion of this thesis.

Second, I am very grateful to my Chuka University internal supervisors, Dr. Adolphus Wagala, Ph.D., and Dr. Gladys Njoroge, Ph.D., for their diligent efforts, unwavering support, expertise, and guidance in completing my thesis.

Third, I appreciate all those who contributed to completing this thesis, including Chuka University departmental and faculty-level reviewers and external examiners.

Fourth, I would like to thank my wife Mauline Nyarotso, daughters Whitney Anyango and Winnie Evelyne, and son Joaness Juma not forgetting my parents for entrusting me and believing in my dreams. May God bless you for your concision, support, and encouragement through my academic journey.

Finally, to all my friends and colleagues at Chuka University and work, may God reward you abundantly for your encouragement and moral support throughout my academic journey.

ABSTRACT

Long queues and waiting times are common in banks, resulting in customer dissatisfaction and low customer retention. The study applied a descriptive research design to investigate queuing dynamics in a banking hall at a commercial bank in Kenya. A single server system (M/M/1) queuing model was used to estimate the average waiting time, system intensity, service time, and optimal number of staff during peak and off-peak periods (July). The study used secondary data on daily waiting times, service times, the number of customers, and servers for May and July 2019, 2020, and 2021 during working hours between 8.30 a.m. and 4 p.m. on Monday to Friday and 8:30 a.m. and 12 p.m. on Saturdays. Data analysis was done using R and Excel. The research findings indicated that the peak periods (May) recorded an average waiting time (AWT) of 13 minutes, 35 seconds in 2019, 10 minutes, 14 seconds in 2020, and 8 minutes, 36 seconds in May 2021. In the off-peak periods (July), an AWT of 3 minutes, 46 seconds, was registered in 2019, 5 minutes, 12 seconds in 2020, and 7 minutes, 42 seconds in 2021. An average service time (AST) of 1 minute 52 seconds in May 2019, 2 minutes 34 seconds in May 2020, and 2 minutes 27 seconds in May 2021. In the off-peak periods (July), an AST of 3 11 seconds was registered in 2019, 3 4 seconds in July 2020, and 2 43 seconds in July 2021. Overall, the system intensities are low to moderate, with the COVID-19 pandemic severely impacting the peak period more than the off-peak. In the peak periods, the service rates averaged 33, 24, and 25 persons per hour in May 2019, May 2020, and May 2021. The respective system intensities were 0.534, 0.360, and 0.492. In the off-peak periods, the average service rates were 19, 20, and 23 persons per hour in July 2029, July 2020, and July 2021. The respective associated system intensities of 0.535, 0.461, and 0.487. From the pooled data for 2019 and 2021, the study recommends that banks operate with an AWT of 6 minutes, 24 seconds, and an AST of 3 minutes. Further, the study established that a bank could work with an optimal four servers with an AST of 2 minutes, 35 seconds (a service rate of 20 people per hour), and achieve a moderate average service intensity of 0.552.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ART	Average Arrival Time
ASR	Average Service Time
ATM	Automated Teller Machine
BCP	Business Continuity Plan
CBK	Central Bank of Kenya
COVID-19	Coronavirus Disease-2019
CSAT	Customer Satisfaction
EQMS	Electronic Queue Management System
FCFS	First Come, First Served
FIFO	First in, First Out
MOH	Ministry of Health
PDF	Probability Distribution Function
SDGs	Sustainable Development Goals
SIRO	Service in Random Order
SLA	Service Level Agreement

CHAPTER ONE
INTRODUCTION

1.1 Background Information

Banks and financial institutions play a crucial role in the economy. Generally, financial institutions facilitate economic liquidity and allow higher economic movement. Banks and deposit-taking financial institutions use customer deposits to lend to other customers earning interest premiums. Banks provide financial transactions such as payment of goods and services and cash transfers, hence mediating economic activities. Banks give customers access to savings accounts, loans, insurance, remittances, wages, and payment services. Social transfer programs, such as in Kenya, have relied on banks to transfer payments to the old to improve their livelihoods. Besides, banks facilitate school fee payments in direct deposits, grants, or funds to the school accounts. As such, the banks' customer base has ballooned over time. While the positive growth reflects increased banks' profits, a social problem arises of queuing in the banking halls.

Despite the mobile-based banking services innovation, queuing in banks is inevitable due to cash-related and registering services. Mobile fund transfer or online banking mitigates the challenges of long queues since they can transact using their mobile devices (Ahmed *et al*., 2018; Billy *et al*., 2018). However, customer care desk services such as pin/cheque book collection, automated teller machine (ATM) card collection, statement collection/standing orders, or cash-related services such as cash deposits, cash withdrawals, bankers' cheque, fund transfer, and international services may constrain a customer to visit the bank. Despite the contemporary trends in telephone banking to reduce waiting times in queues, it has not yet rendered the expected outcome owing to the aforementioned specialized services. As such, queuing is inevitable in modern days. Generally, queues form when service demand exceeds supply (Odior *et al*., 2013). Due to natural variability hence unpredictability of the timing of demands and the duration of service, the dynamics of service systems are very complex (Green, 2006). Banks' success depends on how well service efficiency and queue management are handled (Mangkona & Murdifin, 2018). To remain competitive ahead of the competition, banks need to ensure that customers do not spend too much time in the banking hall (Odirichukwu *et al.*, 2014).

Queuing theory is a branch of mathematics that deals with analyzing and modeling waiting in queues (Berry 2006). Working for the Copenhagen Telephone Company, Erlang (1909) founded queuing theory by exploring how to minimize customer waiting time for telephone circuits. His primal approach to minimizing the waiting time gave birth to the modern queuing theory. The theory has been extended o examines every component of waiting in line, including the arrival times, service times, number of servers, and items which can be people, animals, or cars (Dhari & Rahman, 2013; Varma, 2016). Queuing theory, a part of operations research, can assist businesses in creating workflow systems that are more efficient and cost-effective. It has been applied in various fields, including agriculture, industrial engineering, education (Mwangi & Ombuni, 2015), banking (Asuming-Brempong & Antwi, 2013); Kabamba, 2019; Sheikh, Singh & Kashyap, 2013), health (Egbunu *et al*., 2020; Kiprono, 2017; Mayhew & Smith, 2006; Mehandiratta, 2011; Schlechter, 2015), technology (Menasce *et al.*, 2004), and transport sector (Odhiambo, Orwa, & Odhiambo, 2017). This thesis applied queuing model analysis to explore system intensity dynamics in a commercial bank during off-peak and peak periods.

Banks and other institutions utilize the Electronic Queue Management System (EQMS) to control queues to resolve the congestion problem in banking halls to boost customer satisfaction and increase the workers' overall productivity. A queueing system consists of one or more servers attending to customers arriving according to a well-defined stochastic process. Queuing systems have assumptions. First, it uses different distributions to describe the randomness and variability of the interarrival and service times. These include the exponential or Erlang distributions with different parameters and assumptions (Berry, 2006). Secondly, the service times depend on the arrival process and customer-related factors such as demographics and behaviours. Customarily, some of the service disciplines are first come, first served (FCFS), last come, first served (LCFS), Service in random order (SIRO), priority, processor sharing (PS), round-robin (RR) (Ghimire *et al*., 2017). Long waiting times in a banking hall are an opportunity cost since it can lead to a high level of dissatisfaction among the customers hence low customer retention (Ahmed *et al*., 2018). According to Hongna and Zhenwei (2010), queuing will increase waiting time and induce the customer to complain about dissatisfaction and the negative feeling transmitted to the staff, reducing

their service efficiency. How long a customer has to wait for a service is determined by how many customers are in line, how many servers are available, and how long it takes to serve each customer. Thus, queuing theory can help service managers in decision-making that increases customer and employee satisfaction.

An optimal queuing system reduces the waiting time and increases customer satisfaction (Xiao & Zhang, 2009) and customer retention. A queuing model isolates the components that relate to the system's ability to fulfil random service demands in their occurrence and duration (Ahmed *et al.,*2018). Fader (2020) suggests that companies can adopt and benefit from customer-centricity by using a systematic scoring method. Customer-centricity is a core part of the culture, a prerequisite for strategy, and a key factor for maintaining and growing customer loyalty and retention (Kohli *et al*., 2019). The extent of customer-centricity institutions is usually determined by customer satisfaction. Thus, institutions should identify good customer strategies and take necessary course corrections where necessary. This can include minimizing time wastage by understanding queues and determining how to manage them.

In a banking setting, it is challenging to forebode the congestion levels and the optimal length of service required to attain the desired performance level without a queueing model. The queuing theory attempts to minimize cost by minimizing inefficiencies and delays in a system (Odior, 2013). Analysts utilize queuing models to design and evaluate queuing systems' performance. The model allows for an evaluation of waiting in line with the prescribed corporate standard performance measures, for example, average waiting time in a queue and normal office usage, expected length of the queue, and the number of clients served at ago (Arghish *et al.*, 2012; Campbell & Frei, 2011; Eze & Odunukwe, 2012). As a probability modeling technique, it can evaluate the probability of giving up and the likelihood of a system being in a particular state, say, void or full (Olasore, 2013).

The queuing theory has been applied in several studies' performance analyses of waiting lines. Prasad *et al.* (2018) adopted a multi-server (M/M/S) queuing mode to establish the optimal number of servers, the expected number of customers, and the mean arrival and service rates. In another study, Quarm (2016) studied the multi-server queuing

model and computed the waiting time cost and service provided to hospital patients. In their study, Ni *et al*. (2010) applied the queuing theory to determine the optimal number of service stations and service rates. The main indicators of the system's effectiveness were the length, server intensity, and customer waiting time. Agyei, Asare-Darko, and Odilon (2015) applied queue system modeling techniques in a Ghana Commercial Bank Ltd case study. Kumasi Main Branch. The authors found that using five tellers was better than four or six based on time and total cost of operation. In another study, Bakari *et al.* (2014) applied queuing theory to examine customer service delivery at an ATM service point of Fidelity Bank Plc in Nigeria. They used the M/M/S queuing model and established a service intensity of 0.96. Similar to their work, in Nigeria, the study by Abiodun and Omosule (2015) used the M/M/S queuing model to compare the queuing patterns of customers of two Nigerian banks. They found that more servers lead to less waiting times.

One reoccurring issue in banks in Kenya is congestion in the banking hall. Like in any other banking institution, waiting for services is common in the banking sector in Kenya. In most cases, low service rates and high customer influx can lead to queues. This problem becomes worse during holidays, month-ends, and school openings. During this time, there is a higher influx of customers who need to make deposits for fees, borrow cash for their businesses or school fees, and withdraw for their transitionary purpose. In a case study of two Kenya Commercial Bank branch outlets (Kipande and Rongai), Genga (2018) revealed that customer challenges include long waiting times, unattended service delivery points, and inconsistent service rates. Besides, the banks' data showed that the two branches did not achieve the service level agreements of waiting and service times set by the bank. Several studies in Kenya have applied the queuing theory in the banking sector. Karoney, Kosgei, and Nyongesa (2019) applied queuing system theory to model the average waiting times in sampled Banks in Eldoret Town, Kenya. The deterministic and stochastic components assumed the D/D/1 and the M/M/c frameworks, respectively. Their study demonstrated that both delay components are compatible and depicted relatively equal waiting times for server utilization factors less than one and a substantially increasing delay at ρ above one.

In a recent study, Sewe (2019) applied single queue single server to single queue multiple server models in a post bank and Kenya commercial bank (KCB) in Kisumu, Kenya. The author employed M/M/1 and M/M/r models and established that the average waiting time in Post Bank was significantly higher than in the KCB at a 5% level. The findings can be attributed to bank-specific attributes such as customer base, queue management system, and efficiency of customer cares. Therefore, banks should strive to have bench-mark waiting times by other banks to evaluate their efficiency of service delivery. This will ensure that they maintain customer loyalty and satisfaction.

Despite the previous application of queuing system theory in Kenya's banking sector context, the recent Coronavirus disease-2019 (COVID-19) can create a tremendous shift in the banking sector. The long queue waiting for checkout is not an ideal banking environment for customers, especially when a country is battling COVID-19 because they must adhere to the ministry of health's measures, such as social distancing. Customer dissatisfaction may lead them to move to another bank that offers similar services at relatively equal costs or resort to mobile-based services offered by the host bank or other market competitors.

Whichever the case, customers registered at the banking halls amidst crisis and congestion may reduce the number of customers. Secondary consequences may include shrinking customer deposits and overall transactions, negatively influencing banks' profitably (Arshed and Kalim, 2021; Shah *et al*., 2023). However, the current look at the primal consequences of bank intensities. No empirical study has explored the system intensity dynamics in the context of the COVID-19 pandemic. Therefore, this study examined the applicability of queuing model for optimal customer-centricity in one of the commercial banks in Kenya, the Chuka Branch, within the COVID-19 context.

The study contextualizes the pandemic shock and usual monthly service variation in banks as a substantial determinant of peak and off-peak seasons. First, the peak season is considered the period before the outbreak of COVID-19 and during school opening days which usually occur in the months of May. Comparatively months of July are usually off-peak seasons when schools are closed. Thus, a comparison of waiting times,

service times, and customer size variation was collected, and the resultant service intensities were compared for May and July 2019, 2020, and 2021. May and July 2020 are purposively selected as stringent COVID was freshly in play. The data collected between the two periods run from 8.30 am to 4.00 pm. The analysis used a M/M/1 queuing model (treating the bank as a unit), with queuing discipline of the form First come - First served (FCFS) assumed in the queuing process. The queuing models were used to assess the queuing patterns of customers, that is, how they arrive and how they are served, respectively, with arrival assumed to have a Poisson distribution. On the other hand, the service rate has an Exponential distribution. The aim is to find the average service time, average waiting time, system intensities, and optimal servers at tolerable system intensities.

1.2 Statement of the Problem

Waiting in queues is a common phenomenon, especially when inquiring about essential services in public spaces like hospitals, banks, and finance departments in schools. Time spent on queues and in service is vital in determining customer-centricity. Lengthy queues often lead to high customer dissatisfaction and low customer retention. Customers of commercial banks in Kenya often face long queues and waiting times, especially during peak seasons. Amidst a free market system, banks with long waiting times can lose customers to market competitors offering similar banking services, such as other commercial banks, bank agents, microfinance institutions, and Saccos. This negative feeling transmits to the staff, can reduce their service efficiency, and breeds strife between clients and the servers at the bank, leading to customer churn and its long-term effects on the entire structure's closure.

Despite the previous application of queuing system theory in Kenya's banking sector context, the recent had a negative impact on the demand for banking services. Amidst the COVID-19 pandemic in 2020, the Ministry of Health Kenya laid down control measures such as social distancing, closure of schools, and quarantine that restricted movement in certain regions. Such measures reduce demand for critical services such as fee payments in cash and cheques. While Kenyan studies have examined the context of queuing models in the banking sector, no empirical research has explored the system intensity dynamics COVID-19 pandemic context (pre, during, and post-period). In this

regard, the research carried out an empirical analysis of the queuing theory to gain insights into the behaviour of queues, such as arrival times, waiting time, service time, system intensity, and departure times, in a case study of a commercial bank in Kenya.

1.3 Objectives of the Study

The objectives of this study are outlined as follows.

1.3.1 General Objectives

The study's main objective was to apply the queuing model method to analyse and understand the behaviour of queues in the banking industry, a case study of a selected commercial Bank in Kenya.

1.3.2 Specific Objectives

i. To estimate the average waiting time and the number of customers for cash-related services in a Kenya's commercial bank.
ii. To estimate Kenya's commercial bank system intensity using queuing models.
iii. To estimate the average service time for cash-related services at Kenya's commercial bank.
iv. To estimate an optimal number of staff (servers) during peak and off-peak periods in Kenya's commercial bank.

1.4 Research Questions

i. What is the average waiting time and the number of customers in Kenya's commercial banks for cash-related services?
ii. What is Kenya's commercial bank system intensity?
iii. What is the average service time for cash-related services at Kenya's commercial bank?
iv. What is the optimal number of staff (servers) during peak and off-peak periods in Kenya's commercial bank?

1.5 Significance of the Study

Financial institutions such as banks play an intermediation role in attaining SDGs by financing several activities towards poverty reduction, quality education, clean water, and sanitation. Banks give customers access to savings accounts, loans, remittances, wages, payment services, and social transfer payment programs. The banks' customer base has ballooned, creating a social problem of queuing in the banking halls. With competition in the market by other banks, optimizing the quality of the services provided while minimizing the time of the service is an ultimate goal in the banking sector. The queuing theory helps efficiently link queue lengths and times and service times.

In addition to the existing body of literature on modelling using the Queuing theorem, this research may assist the banking industry in providing excellent services with little waiting time to enhance customer-centricity. The study may also help the policymakers make evidence-based policies in the banking sector by management officials on staffing optimal service utility, especially during a health pandemic such as COVID-19, and provide a basis for evaluating the optimal level of service delivery. Understanding the nature of the queuing system in banks can help branch managers make business choices in constructing well-ordered and dynamic workflow systems. By obtaining the expected waiting time in the bank, clients get advanced information on the time they will probably take before being served in a commercial bank at the Chuka branch, Kenya. Moreover, the study also provides recommendations for further research based on the gaps that the study was not addressed in this study hence giving academicians a pathway to extend the Queuing models' applicability.

1.6 Assumptions of the Study

The customer interaction points in the bank are the service section and teller points, and the following assumptions were made for the queuing system:

i. The arrival of customers follows a Poisson distribution with an average rate of λ customers per minute.
ii. The queue discipline is First-Come, First-Served (FCFS); there is no precise preference upon arrival to any of the servers.
iii. The Service times are distributed exponentially, with an average of α Customers

per minute.

iv. There is no limit to the number of queues (That is, it assumes an infinite calling source capacity)

v. Per their employer's expectations, the service providers are working tirelessly to their maximum capacity.

vi. The average arrival rate is more than the mean service rate.

1.7 Operational Definition of Terms

Arrival process:	The probability distribution of customer arrival at any time.
Average arrival rate (λ)	is the number of customers arriving at the bank at a given period.
Average service rate (μ)	Is the average number of customers served at a given period.
Average service time	refers to the average service time received by customers seeking cash-related services.
Average waiting time	refers to customers waiting in the queue for cash-related services.
Customer-centricity	is a bank's culture that places the customer first and at the centre of its operations to enhance customer satisfaction and retention.
Off-peak periods	represent the months of July 2019, 2020, and 2021.
Optimal number of servers	refers to the number of servers required to serve at the bank without highly intensifying the bank's system.
Peak periods	represent the months of May 2019, 2020, and 2021.
Queuing system	consists of a number of servers at counters, interconnecting queues of customers waiting for bank services.

System intensity (**utilization) (ρ)** in a queuing system ratio of arrival and service rates.

CHAPTER TWO
LITERATURE REVIEW

2.1 Waiting Times and The Number of Customers in Commercial Banks.

Waiting times could be influenced by internal factors such as bank efficiencies in record keeping and external such as arrival rates of customers. Modern banks are more efficient in service delivery through innovations such as biometric verification systems that have reduced the verification process. Kasum *et al.* (2006) studied queue efficiency by comparing old and modern Nigerian banks. They found that the time spent in the queue for services in the old-generation bank is longer than in the new-generation bank. Waiting times in banks vary depending on the arrival rates. Cowdrey *et al*. (2018) established that waiting times for fast arrivals (interval rate of 3 minutes) under the FIFO queue is about 39.24 minutes for slow arrivals (interval rate of 10 minutes) is about 0.03 minutes. Banks' servers become overwhelmed at high arrival rates given their normal service rates; hence, congestion might arise at banks. This study postulates that peak periods such as school opening are often associated with longer waiting times due to high demand for cash-related services such as fee deposits.

Queuing Modelling was also applied by Zewude & Sodo (2016) in a comparative study of the two Commercial banks in the Wolaita zone of Ethiopia. The study established that the average waiting time (AWT) in the queue and system were 0.001 and 0.43 minutes in the Tona branch compared to 0.216 minutes and 0.828 minutes, respectively, in Dashen Bank. The authors suggested that the Tona branch needs to increase the number of servers (from 6) since it has a higher arrival rate of about 82 compared to 72 in Dashen Bank (four servers) to reduce AWR and consequently reduce congestion, improve efficiency, and overall performance.

In a Kenyan study, Mwangi & Ombuni (2015) empirically analysed Queuing Model. The study examined the queueing dynamics among students served at the finance office and how it relates to student satisfaction. The study applied the single-Queue Multiple Servers Model (M/M/c) to model students' flow as they form a single line waiting for service from three clerks. From a survey of 384 students, the authors established low satisfaction levels with waiting lines. This could be associated with the administration's remedy of turning away some students due to the long queues. Moreover, the study

revealed an average arrival rate of 22 persons per hour, and a service rate of 23.7 persons per hour, yielding a system intensity of 92.95%. The average number of students waiting was about 13, and the average waiting time was 33.42 minutes. The study concluded that the multi-server model (M/M/s) was better in the Finance department than the single-server (M/M/1) model. Thus, this study considers a single-server model in the banking sector to model the overall bank performance as a unit.

Azumah *et al.* (2021) adopted a multiple servers (M/M/s) Model to compare the waiting times of two selected banks on three consecutive days in July 2020. The authors established significant daily fluctuations in waiting times for customer service across banks. In Bank A, day two registered recorded the highest waiting time of three minutes and forty seconds (3 mins 40 secs), followed by day three (3 mins 20 secs), with day one recording the least waiting time of 2 mins and 30 secs. The variation in waiting times can be attributed to their arrival rates each day, respectively 28, 27, and 20 customers per hour. In Bank B, day 2 recorded the highest waiting time for customer service of 10 minutes, followed by day 3 with 4 minutes, and lastly, day 1 with three and half minutes (3:30). The variation in arrivals rates in each day, respectively, were 20, 9, 17 customers per hour. Notably, bank servers maintain their service delivery pace; hence the higher influx of customers within a given period will lead to congestion, thus, longer waiting times. This study demonstrates that peak seasons, usually Mays, are associated with longer waiting times than during off-peak seasons, usually Julys.

It is essential to note an overall decline in demand for on-site bank services due to the digitalization of financial services to mobile-based and the use of ATMs. Increased reliance on mobile baking makes it efficient for a single bank to service thousands of customers remotely. Deloitte Center for Financial Services report indicates that branch density is declining globally (Srinivas & Wadhwani, 2019). As customers increasingly use digital channels for regular transactions like paying bills or transferring money, many international banks are shutting down branches to reduce costs. For instance, more than 3000 commercial bank branches have been closed in the United States since 2010, declining to only 4,377 at the end of 2020 (Emmons, 2021). Banks with higher integration of services with mobile baking technologies are more likely to register few customers who physically visit the bank. Yet, essential services would force customers

to visit the bank, such as opening an account and high-value cash-related services that may not be authorized or exceed the capacity of bank mobile agents' outlets. Besides, there is an overall positive attitude and customer preference for on-site bank service to online or mobile channels. From a large cross-section study of 17,100 banking customers across 17 countries, Deloitte Center for Financial Services revealed that bank branches are more vital to overall customer satisfaction than online or mobile channels (Srinivas & Wadhwani, 2019). The study further established that customers prefer bank branches over digital channels when opening new accounts and cash-related services like loans. McKinsey & Company report also indicated that most individuals prefer using branches or ATMs due to low trust in banks and financial systems in Italy, Spain, the US, France, Germany, and the UK (Dallerup *et al.*, 2018).

There is no precise number of customers a bank expects, partly due to the aggregation of bank reports in their customer base and several confounders. The average monthly number of customers in a commercial bank significantly varies depending on various factors, such as country, the type of products and services, the size and location of the branch, the seasonality and frequency of customer visits, and the level of digitalization and automation of banking services. According to Statista Research Department (2022), the average number of customers per branch in the European countries in 2018 ranged from 1,500 in France to 14,500 in Estonia, equivalent to an average monthly number of customers per branch would be about 125 to 1208, respectively.

Banks' services are also affected by the size and location of the branch. Banks with a higher customer base situated in major cities in Kenya, such as Nairobi and Mombasa, are more likely to register few customers at the banking halls than other towns, such as Narok (Githae, Gatauwa, & Mwambia, 2018). Besides, there is a significant seasonal demand variation for bank services, especially in Kenya. Peak periods are often school opening months of May and July when banks are more likely to register more customers. Lastly, as discussed previously, banks that have intensified the digitalization and automation of banking services are more likely to get fewer customers since most of their services are done remotely (Srinivas & Wadhwani, 2019. Therefore, this study estimated the number of customers during peak and off-peak periods.

2.2 Commercial Bank System Intensity

Cowdrey *et al*. (2018) established that system intensity rates vary depending on the arrival rates. At first, arrivals, when a high influx of customers arrives within a given period, banks' servers [cashiers] are highly intensified. According to the authors, a typical fast arrival is where customers arrive at an interval of 3 minutes (Arrival rate [ART] of about) 20 customers/hour. In contrast, a slow arrival rate is at an interval of 10 minutes with an ART of 6 customers/hour. At fast and slow arrivals' service times are about 15 mins (4 customers/ hour) and 20 (3 customers/hour). Consequently, at first arrivals, banks tend to be highly intensified (utilization $[\rho] = 1$) compared to first arrivals ($\rho = 0.39$) (p 382).

Intensities vary from one server to another depending on the number of customers they receive in a given period. In Zewude and Sodo (2016), the six servers in Tona Bank had different intensities of 1.1898, 0.5949, 0.3964, 0.2978, .2379, and 0.1985 respectively, given their respective varied arrival rates of 163, 81.5, 54.3, 40.8,32.6, and 27.2 in consecutive of two days period (p 15). The same variation was also observed in Dashen Bank. The four servers in Dashen Bank had different intensities of 1.1633, 0.5816, 0.3878, and 0.2908, respectively, given their respective varied arrival rates of 114, 57, 38, and 28.5 in consecutive two days periods (p 15). Kenya's commercial banks, like other banks, have more than one server that might serve different customers daily, depending on their service rates and the type of service they offer. A teller with a higher service rate could dispense more customers in a day than a customer who is a little bit slower in their service delivery. Nonetheless, the study does not overlook the possibility that some services like funds transfer and cash deposit may take longer than NHIF/KRA and Bankers' cheque services. A server who coincidentally takes demanding services may end up serving fewer customers daily than their counterparts who take fewer demanding services. Thus, the current estimation of the service intensities of each teller was done alongside the composite intensity for the bank as a unitary entity.

Azumah *et al.* (2021) found that the second bank (Bank B) had a higher utilization factor than the first bank (Bank A), attributed to an increased number of customers and lower service rates in the banking hall of Bank B than that in Bank A. Banks' A utilization factors for the three consecutive days were 0.389, 0.354, and 0.370. The days

with higher system intensity registered a higher customer turnout than those with lower system intensity. The arrival rates each day were 28, 27, and 20 customers per hour, respectively. Banks' B utilization factors for the three consecutive days were 0.588, 0.750, and 0.568. The days with higher system intensity were also associated with a higher customer turnout than those with lower system intensity. The arrival rates each day were 20, 9, and 17 customers per hour. Service rates also contributed to the variation in the system intensities. The authors observed that Bank A had fewer service times, with customers spending four minutes completing their transactions much lower than 11 minutes in Bank B. The findings imply that waiting time for customer service affects the number of customers in the queue and, consequently, the system intensity utilization factor. Higher customer influx and low service rates lead to a higher system intensity.

2.3 Average Service Time at Commercial Bank.

Williams, Ogege, and Ideji (2014) did a case study using five big Nigerian banks within a queuing technique framework. The author examined how different aspects of customer services banks use affect their profitability in the banking sector. The authors established a mean service time of about 1.067 minutes per customer, translating to a service rate of 56.25 customers per hour. The study further highlighted the significance of waiting time on banks' profitability using Ordinary least square regression analysis. The results revealed that reduced waiting time improves the bank's profitability. Moreover, the findings also suggested that poor customer service management in the banks could cause their profitability to decline, creating financial difficulties for the banks. The study suggested that customer services, more so in the Nigerian banking sector, can be improved by giving more attention to the customer's waiting time in the queue, the average service time, and the probability that the bank cashier is idle. The study elucidates the relevance of estimating bank service rates which can serve as a proxy or banks financial performance. The current sought to establish service times in Kenya's commercial banks' context.

The average service time in banks largely depends on the type of services a customer receives. In a banking setting, two broad services are offered: non-cash and cash-related. Cash-related services include cash deposits, cash withdrawals, bankers' cheques, fund transfers, and international services such as MoneyGram and forex. Cash-related

services are primarily serviced at the customer care desk, including general inquiries, internet banking, pin/cheque book collection, ATM card collection, and statement collection/standing orders. Cowdrey *et al*. (2018) reported that bank services range between 2 and 30 minutes. Banking activities such as collecting banking statements would take 0 to 2 minutes, deposits and withdrawals between 2 to 5 minutes, and opening new accounts between 15 to 30 minutes. In contrast, loan processing could take more than 30 minutes (p 382). This study focused on waiting times for cash-related services only.

2.4 Optimal Number of Staff (Servers) In Commercial Banks

In a typical baking sector, bank intensity increases when service demands or arrival rates are high, and vice versa. Cowdrey *et al*. (2018) established that the waiting time for a fast arrival rate (interval rate of 3 minutes) of the customers under the FIFO technique is 87 minutes for one server (ρ = 1, λ = 0.10, μ = .05), 15 minutes for two servers (ρ = 0.91, λ = 0.10, μ = 0.05) 1.5 minutes for three servers (ρ = 0.65, λ = 0.33, 0.10), 0.2 minutes for four servers (ρ = 0.49, λ = 0.09, μ = 0.05), and 0.02 minutes for five servers (ρ = 0.49, λ = 0.09, μ = 0.05). At a slow arrival (interval rate of 10 minutes), waiting times are 156 minutes for one server (ρ = 1, λ = 0.33, μ = 0.32), 122 minutes for two servers (ρ = 1, λ = 0.33, 0.16) 91 minutes for three servers (ρ = 1, λ = 0.33, μ = 0.01), 63 minutes for four servers (ρ = 1, λ = 0.33, μ = 0.08), and 39 minutes for five servers (ρ = 1, λ = 0.33, μ = 0.07). Generally, the study demonstrates that banks become less intensified as more serves to increase and are likely to create idle time if arrivals are generally low. This study does not anticipate such a first interval rate of 3 minutes. From the fast and slow arrival rates, four servers would have yielded the same bank utilization as five without dissatisfying customers with long waiting times. Therefore, it is interesting to establish an optimal number of severs in Kenya's commercial bank context.

Burodo, Suleiman and Shaba (2019) examined queuing characteristics at the Kaura Namoda branch of First Bank Nigeria Ltd using one-server (M/M/1), two-server (M/M/2), and three-server (M/M/3) models. The authors established that 39.7%, 45.7%, and 84.4% of arriving customers per hour have to wait to be served on Monday, Tuesday, and Wednesday, respectively. Besides, the system's waiting and service times averaged

about 12 minutes, 3 minutes, and 2 minutes when the server was one, two, or three, respectively. The pattern of operations shows that more servers mean more service efficiency, and less servers mean less efficiency. The study concludes that more servers mean less average time in the system hence recommended that the bank should increase the number of servers to meet the dynamic expectations of swarming customers in banks. The study investigated the optimal number of servers for a commercial Bank in Chuka Town, Kenya.

In Ethiopia, Zewude and Sodo (2016) demonstrated that it might vary from one bank to another depending on the arrival rate. The authors recommended that Tona commercial bank improve its servers from six since it has a higher arrival rate estimated at about 82 compared to 72 in Dashen Bank with four servers. As a result, the current study anticipates that an optimal number of servers could be established. A simulation of possible intensities was shown by varying the number of servers.

Azumah *et al.* (2021) established that Bank B had a higher average utilization intensity (0.635) than Bank A (0.371) attributed, partly due to the lower number of servers in Bank B (2 servers) than Bank A (3 servers). Thus, the authors recommended that the management of Bank B adopt a three-server model. The findings imply that banks can vary their number of servers to meet their target system intensities. When the bank is highly intensified, it can increase its number of servers to meet the increasing service demand. Higher demand in this case study of Kenya's commercial bank occurs during peak seasons and declines during off-peak seasons of May. Hence, there is a need to determine the optimal number of serves that ensure a moderate and even system intensity during peak and off-peak seasons.

2.5 Overview of Customer Centricity

The overall impression of the company is the customer experience. It comprises all interactions customers have with the company's brand, with or without the company's knowledge (Meyer & Schwager, 2007). Customer-centricity is vital for customer growth and retention (Kohli, Jaworski, & Shabshab, 2019). Customer-centric goals and customer satisfaction measurement programs must align with company objectives and business performance measures related to financial performance.

Additionally, satisfied customers benefit an organization by spreading positive word-of-mouth, seemingly becoming a walky-talky advertisement (Nguli, 2016). Therefore, it reduces the cost of promotion to attract new customers (Anderson & Sullivan, 1993) and increases their competitive advantage (Nguli, 2016). Factors such as product innovation, management practices, banks' opening and closing times, response time to customers' complaints and workers' attitudes towards customers, and queue management systems influence the satisfaction of customers' banking sector (Mailu, 2014; Ngilu, 2016). Despite several banking innovations, including mobile banking, ATMs, and internet banking (Macesich, 2000), queuing up is the usual reason for customers' disgust with retail banking (Ngugi, 2016). Due to a higher turn-around time (time taken to complete a task), a long waiting time negatively affects customer satisfaction. Since customer satisfaction is a precursor for customer retention and loyalty, banks should seek to reduce waiting times.

Advanced companies will use predictive analytics and systems like Qmatics to design customer programs and manage the flow (Cederborg & Larsson 2015). With the Qmatic system, many companies know how many customers have been served, waiting, and service times. The customer-centricity is then measured using metrics like Customer Churn, CSAT, and Net Promoter Score for the respective branches (Diaz-Aviles *et al*.,2018). Several researchers have linked and compared waiting time to Customer satisfaction modelling using queuing theory and found that reduced waiting time increases customer satisfaction (Ahmed *et al*., 2018; Hongna & Zhenwei, 2010; Mwangi & Ombuni, 2015; Ngugi, 2016).

Chen *et al.* (2019) found that banks need to adopt a seasonal staffing policy due to the seasonality of customer demands. The authors observed that long waiting times had a negative effect on the behaviour of some customers who chose not to join (balked) or left the queue (reneged). Besides, the seasons of the year are divided into two peak and off-peak situations that are more complex in reality. Their study also looked at staff shifting between head offices and different branches. Branch management is faced with challenges regarding staffing due to economic and business environment changes and leaves management for the staff without interfering with customer expectations regarding waiting time. The main contributor to customer dissatisfaction is that waiting

time is managed differently, including adding staff, which means more operating costs and improper resource allocation weakens the resources' supporting role (Garvin, 1988). Thus, this study applies Queuing theory to model waiting times at different seasons' peaks and off-peak so that repeated customer dissatisfaction throughout the year could, if not eliminated but reduced due to informed data-driven decision-making despite managing operation costs for optimal customer-centricity.

2.6 Peak and Off-Peak Season

Consumer habits are constantly changing due to the fast adoption of new technologies, and banks are not sensible to open and close branches repeatedly to keep up (Kim, 2006). Banks should aim to provide a customer experience that meets customers' expectations and fully connects branch networks with all other digital channels. Customers view the bank and want to conveniently use both physical and online services. Some might prefer face-to-face services at the bank branch; for others, it may be a mobile-only experience. Traditional banks have adopted physical and digital service delivery channels to meet the customers' interests and preferences (Skorobogatov, 2012).

The bank usually faces significant seasonal variations yearly (Paxson, 1993). that is, peak and off-peak. Regarding customer arrival and service, customer arrivals are not constant throughout the year; the trend could be seen as a high influx of several customers during certain seasons, like school opening days, and then a decline in the arrival of customers after that. In Kenya, such occurrences are seen in January, May, and September, commonly called back-to-school seasons. However, other months register low service demand, predominately cheques for fee payments and school fees.

2.7 Queuing Systems

A queuing system comprises one or more servers serving arriving customers. Queuing systems derived from customers who find all the servers busy usually join one or more queues Odior, 2013). Queues depend on service times, service discipline, the number of servers, and inter-arrival times. Queuing systems have practical applications in many fields. In a typical bank queue system process, a web-based application assigns a queue number based on the service type to each customer on arrival. The customers then wait

for their number to be called to the correct terminal when their turn comes. The system is designed to improve queue management and service efficiency (Odirichukwu *et al*., 2014). From the time customers join a queue until they are served, there are specific steps they follow. These elements of queueing systems are briefly described as follows: (Güneş, 2012).

Calling population: Refers to the population of potential customers. It can be finite or infinite. Finite population model: The arrival rate is influenced by the number of customers served and waiting with a limit to the number of potential customers. For example, booking a flight operated by one plane. Infinite population model: The arrival rate depends on the number of customers served and waiting and a large or unlimited number of potential customers.

Customer: Refers to anything that needs service in a queuing system, which is mathematically analysed by queuing theory. It can be a customer at a bank, students in a finance office, or vehicles in a petrol station to emails waiting to be red or replied to.

The Arrival Process of Customers refers to how customers arrive at the queue and their theoretical distribution. It can be random, scheduled (one or in groups), finite (if it has a finite number of customers), or infinite (continuous). In a typical bank setting, customers' arrivals are not scheduled but arrive randomly based on their customer needs. Regardless, customers are assumed to follow a Poisson distribution or exponential inter-arrival times since they are counts and are positive.

System Capacity: Refers to a limit on the number of customers in the system where Limited capacity is when the queue can accommodate a limited number. For example, a bank's capacity can be limited by space or the number of seats. A full system limits further entry until more space is created and unlimited capacity is when queue can accommodate any number of customers, e.g., concert ticket sales.

Server: Refers to anything providing services to the customer. Examples include machines, airport runways, routers, and workers (Güneş, 2012). The table gives examples of components of a question system.

Table 1: Components of a queuing system

System	Customers	Server
Bank	People who need to deposit, withdraw, or transfer money	Teller or ATM
Supermarket	People who need to buy groceries or other items	Cashier or self-checkout machine
Call center	People who need to make or receive phone calls	Operator or agent
Airport	People who need to check in, go through security, or board a flight	Staff or scanner
Hospital	Patients who need to see a doctor, get a test, or receive treatment	Nurse or doctor
Network	Packets or users	Router

The Service Times: This is the time the server takes to serve the customers from when a customer joins a queue. The service time can be deterministic or exponentially distributed. It also depends on the queue length (Ghimir *et al.*, 2017). For instance, the processing rate of a machine can be increased if there is an increased number of jobs waiting to be processed. Successive times denoted by S_1, S_2, S_3 may be constant or random and often follow a distinct distribution such as Exponential, Weibull, and Gamma distributions.

Queue behaviour: How customers act when waiting for a service to start. As Ghimire *et al*. (2017) described, customers may patiently wait for their turn or leave. Thus, they described customers' behaviour as follows:

i. Balk: Customers choose to leave when they realize the queue is too long and have to wait longer than expected.
ii. Renege: leave when the queue moves too slowly or when they get tired.
iii. Jockey: move from one line to a shorter line. These customers can also re-join the queue they had left earlier either by balking or reneging and are considered jockeying.

Queue Discipline logical ordering of customers a queue determining who gets service next when a server is available. Some of the possibilities for the order of the service facilities are as follows:

i. First-in-first-out (FIFO)
ii. Last-in-first-out (LIFO)

iii. Service in random order (SIRO)

iv. Shortest processing time first (SPT)

v. Service according to priority (PR)

2.8 Probability Distributions

The Queuing theory is grounded in probability theory (Berry, 2006). The queuing model has several queuing characteristics, the distribution, input process, output process, queue discipline, and birth-death process; these characteristics are assumed to follow a given distribution (Bhat, 2015).

2.8.1 Exponential and Poisson Probability Distribution

In applying queueing models, the assumptions regarding the probabilistic nature of the arrival and service processes must be considered. Typically, the arrivals are assumed to follow the Poisson process because the number of arrivals in a specified period has a Poisson distribution. Thus, if $N(t)$ denotes the arrival counts at a period t and N(t) has a Poisson distribution, then;

$$Probability\,\{N(t)\} = \; n = \frac{e^{-\lambda t}(\lambda t)^n}{n!} \qquad \forall\, t \geq 0 \tag{1}$$

Where λ; is the rate and n; is the expected arrival counts per unit of time. The distribution determines the probability of arrivals in a given time(t). As an illustration, if $\lambda = 5$ customers per hour, then the expected number of arrivals in any 1-hour interval is 10, and the expected number in a 30-minute interval is 5. The Poisson process can also be characterized in terms of interarrival time. That is in terms of time between consecutive arrivals. Such a process follows an exponential distribution. Thus, if T is a random variable representing the interarrival time with exponential distribution with rate λ interarrival times, then;

$$P(IT \leq t) = 1 - e^{-\lambda t} \tag{2}$$

and

$$P(IT > t) = e^{-\lambda t} \tag{3}$$

Where; $1/\lambda$ is the average time between arrivals.

By setting n=0, the poison distribution in Equation 1 distribution reduces to $e^{-\lambda t}$ which is similar to $P(T > t)$ from the exponential distribution. The reduced function helps

to compare the probability of no arrivals occurring at a certain time with the probability that an interarrival time is of a certain length.

Berry (2006) sites two major reasons why exponential distribution is preferred. First, the exponential function is inversely proportional to t. After arrival, the subsequent waiting times before the next arrival tends to decline. Secondly, the time until the second arrival depends on the time elapsed or the last arrival, the so-called no memory property ("memoryless"). Thus, it is a suitable model for assessing customer arrivals since their decisions are independent. In this regard, the Poisson process is a 'random" arrival process (Green, 2006). Another defining property of the arrivals in a service system is that customers arrive one at a time.

2.8.2 Erlang Distribution

Erlang distribution is adopted if the exponential distribution is unsuitable when modelling interarrival times (Berry, 2006). Its density function depends on two parameters; the rate (R) and the shape (k). The distribution is expressed as in Equation 4.

$$f(t,R,k) = \frac{R(Rt)^{k-1}e^{-Rt}}{(k-1)!} \quad (4)$$

2.8.3 The Input Processes

According to Sevcik & Mitrani (1981), we define t_i as the time when i^{th} customer arrives $\forall\, i \geq 1$. Let $T_i = t_{i+1} - t_i$ to be i^{th} interarrival time. All $T_i's$ are independent and continuous random variables denoted by A with a PDF of $a(t)$. Generally,A has an exponential distribution with parameter λ is defined by Equation 5.

$$a(t) = \lambda e^{-\lambda t} \quad (5)$$

This implies that for all positive values of t and h;

$$P(A > t + h | A \geq t) = P(A > h) \quad (6)$$

The identity reflects the no-memory attribute of the exponential distribution useful in modelling inter-arrival times (Berry, 2006)

2.8.4 Birth and Death Process

Le, the number of people waiting in the queuing system, to be the state of the system at time t, and since future state t is affected by the initial state, the focus is reduced on the initial state at $t = o$ since the systems' state is starting. The system's initial state affects the future state at time t. Let $P_{ij}(t)$ be the probability that the state at time t given that the state at $t = o$ was i.

For a large t, $P_{ij}(t)$ becomes independent of i and approaches the limit π_j(Steady-state of j) (Hanin, 2001). A queueing system reaches a steady state when the probability of the number of customers in the system is independent of t. In such a case, the system is said to be stable since it does not change over time. It occurs when the arrival rates are less than or equal to the service rates; otherwise, the queue will keep increasing. A birth-death process is a process where the state of the system at any time t is a non-negative integer whose variables include;

λ_j, representing the birth rate at state j and symbolizes the probability of arrivals occurring over a period of time.

μ_j, representing the death rate at state j and symbolizes the probability that the completion of service occurs over a period of time.

A birth (death) process increases (decreases) the state by one. For the death process, $\mu_0 = 0$ since a negative state does not exist. Besides, Birth and death must be independent to be considered the birth and death process. The probability of birth occurring between t and $t + \nabla t$ is $\lambda_j h$, and increases the state from j to $j + 1$. While the probability of a death occurring between t and $t + h$ is $\mu_j \nabla t$, and decreases such a state from j to $j - 1$.

2.8.5 Steady-State Probabilities

According to Miller (1981), the steady-state probability π_j Is the probability that a state at time t will be j given that the state at $t = o$ was i for large t. The relationship between $P_{ij}(t + \nabla t)$ and $P_{ij}(t)$ for the system to end up reasonably sized t. The potential states at time t must be $j, j - 1, j + 1$.

To obtain π_j; the probabilities of the first three situations are added up since ∇t approaches zero; state j cannot be reached from the other states due to the independence of births and deaths and won't occur simultaneously. Based on the property of the multi-server queuing system model of no more than one event.

$P_{ij}(t+\nabla t) = \lambda_{j-1}\nabla t P_{i,j-1}(t) + \mu_{j+1}\nabla t P_{i,j+1}(t) + P_{ij}(t)\,(1-\lambda_j\nabla t - \mu_j\nabla t)$

We define steady-state probability $\pi_j = \lim_{\nabla t\to\infty} P_{ij}{}'(t)$,with further substitution, we have the flow balance equations as below.

$\lambda\pi_{j-1} + \mu\pi_{j+1} = (\lambda+\mu)\pi_j \;; j = 1,2,$

Second scenario

$P_0(t+\nabla t) = P_1(t)$ one service and no arrival $+P_0(t)$ no arrival and no service;

$P_0(t+\nabla t) = P_1(t)^*(1-\lambda\nabla t)\mu\nabla t + P_0(t)^*(1-\lambda\nabla t)$

$\frac{P_0(t+\nabla t)-P_0(t)}{\nabla t} = P_1(t)\mu - P_0(t)\lambda$, Given $\lim_{\nabla t\to\infty}\frac{P_0(t+\nabla t)-P_0(t)}{\nabla t} = 0$;

$$\mu\pi_1 = \lambda\pi_0; j = 0 \tag{7}$$

From Equation 7

$$\pi_1 = \frac{\lambda}{\mu}\pi_0 \tag{8}$$

From Equation 9, we have; $\mu\pi_2 = \lambda\pi_1$. Therefore,

$$\pi_2 = \left(\frac{\lambda}{\mu}\right)\pi_1 = \left(\frac{\lambda}{\mu}\right)^2\pi_0 \tag{9}$$

We define utilization factor; $\rho = \frac{\lambda}{\mu} \rightarrow \pi_1 = \rho\pi_0$. Generally,

$$\pi_n = \rho^j\pi_0 \tag{10}$$

Since $\pi_0 + \pi_1 + \pi_2 + \cdots = 1$, we have; $\pi_0 + \rho\pi_0 + \rho^2\pi_0 + \cdots = 1 ;\rightarrow \pi_0[\rho + \rho^2 + \rho^3 + \rho^4 + \cdots] = 1 \;;\rightarrow \pi_0\left[\frac{1}{1-\rho}\right] = 1$. This implies $\pi_0 = 1-\rho$; $\pi_1 = \rho(1-\rho)$Generally, the arrival rate for a steady state process is expressed as in Equation 11.

$$\pi_n = \rho^n(1-\rho) \tag{11}$$

2.9 Queuing Disciplines

Queuing discipline is a rule that determines the order of services of customers in a queue (Klausmeier *et al.*, 2020). In most cases, any arrival joins the end of the queue. First arrivals are served first in their arrival order. In most expansive organizations, the transaction that occurs with the help of electronic devices is based on dates. The

common service disciplines are outlined below (Baffour & Anokye, 2014).

FIFO (First In, First out): Customers arriving first are served first, and those arriving last are served last.
LIFO (Last in, First out): Customers arriving last are served first, and those arriving first are served last.
Random Service: Customers in the queue are served in random order.
Round Robin: The customers are served in a circular order. Customers at the front get a fixed amount of service time before switching to the end of the queue.
Priority Disciplines: The server chooses customers with the highest priority. s. An example of priority discipline in a bank is when customers with special needs, such as high-ranking profile citizens such as a Governor or pregnant women, are given priority over other customers.

Processor Sharing: Characterised by no queues. Customers are all served simultaneously, each receiving an equal fraction of the service capacity available such that the service rate is proportional to the number of customers in service. Often applicable in the telecommunication industry. For example, the channel shares the bandwidth among different users in a wireless network.

While selecting a given discipline will likely significantly affect waiting times for a particular customer, no one will want to arrive early in an LCFS discipline. In the LCFS discipline, all outcomes are constantly receiving service regardless hence no effect on waiting time.

2.10 Kendall-Lee Notation (A/B/C/D/E/F)

In a study by Kim (2012), the characters of the queuing system could be very wordy. Kendall (1918-2007) proposed a standard method known as Kendall's notation to categorise queueing systems. Thus, the study adopted the Kendall-Lee Notation system for parallel server queues denoted by: A/B/C/D/E/F. A and B denote the interarrival and service times distribution, respectively. In this case, both distributions are assumed to follow Poisson and Exponential distributions, respectively. Common symbols for the distribution of A and B are M: Markov, exponential distribution, E_k: Erlang of order k,

H: Hyper exponential, and G: General distribution. Thus, the symbols for A and B are M (Markov, exponential distribution). C denotes the number of parallel servers. D represents the service discipline for this study FCFS; E denotes the system capacity of the queuing system, while F denotes the calling population/ population size to draw from (Berry, 2006). This study assumes an infinite capacity and an infinite population, respectively. Thus, the infinity symbol represents E and F; otherwise, they can be dropped.

Examples:
M/M/1/FCFS/∞/∞ (similar to M/M/1): Single-server with unlimited capacity and calling population. Inter-arrival and service times follow an exponential distribution.
M/M/3/ FIFO/25/1200: Three parallel servers with a capacity 25, call-population 1200, and service discipline FIFO.

2.11 Little's Queuing Formula

Suppose we let L be the average number of customers in the queue at a given period and assume that a steady-state has been reached and W to be the average time a customer spends in the queuing system. Taking λ as the arrival rate of customers into the system per unit time. From the customer's perspective, four parameters of interest, whenever they enter a banking hall are; how many persons are in the queue, in service, the average amount of time spent in the queue, and in service.

We employ Little's queuing formula (Winston, 1991).in finding out the parameters of interest stated as

$$L = \lambda \, \mathrm{W} \tag{12}$$

Further breaking L into L_q and L_s to represent the average number of customers waiting in queue and in-service, respectively, that is

$$\mathrm{L} = L_q + L_s \tag{13}$$

While W into W_q and W_s represent the average time spent in the queue and in-service, respectively that is

$$\mathrm{W} = W_q + W_s \tag{14}$$

In summary, Little's formula to employ is presented in Equations 15 to 17.

$$L = \lambda \mathrm{W} \tag{15}$$

$$L_q = \lambda W_q \quad (16)$$

$$L_s = \lambda W_s \quad (17)$$

2.12 Queuing Model Formulation

This section covers different queuing model formulations based on the Kendall-Lee Notation framework explored in Section 2.7

2.12.1 Single Server Model: M/M/1/FCFS/∞/∞ Queuing System

Single server system with exponentials interarrival time and service time operating on FCFS queuing discipline, having an infinite system capacity and an infinite pool of customers to draw from the system can be modelled as a birth and death process where; (Singh *et al*., 2007)

$$\lambda_j = \lambda \; \forall \; j = 0,1,2, \ldots \quad (18)$$

$$\mu_i = \begin{cases} \mu_0 & for\ n = 0 \\ \mu\ for\ (n = 1,2,3 \ldots) ds \end{cases} \quad (19)$$

Substituting in the steady-state probability, we get;

$$\pi_j = \frac{\lambda^j}{\mu^j}\pi_0 \quad (20)$$

We define traffic intensity as before, given by $\rho = \frac{\lambda}{\mu}$. Since the sum of all steady-state probabilities is equal to 1, we've: $\pi_0[\rho + \rho^2 + \rho^3 + \rho^4 + \cdots + \rho^n] = 1$. Assuming $0 \leq \rho \leq 1$; we let $v = (\rho + \rho^2 + \rho^3 + \rho^4 + \cdots + \rho^n) = 1$, Then $v = \frac{1}{1-\rho}$ and $\pi_0 = 1 - \rho$. The steady-state probability of state j becomes; $\pi_j = \rho^j(1 - \rho)$. If $\rho \geq 1$, then v approaches infinity, and thus a steady-state cannot exist, but if $\rho \geq 1$, then $\lambda \geq \mu$ that is higher than the service rate, and the queuing system will grow without end.

Solving for the Length of the queue L using the system's steady-state probability calculated using Equation 21.

$$L = \sum_{j=0}^{\infty} j\pi_j = (1 - \rho)\sum_{j=0}^{\infty} j\rho^j \quad (21)$$

Let $v = \sum_{j=0}^{\infty} \rho^j = (\rho + 2\rho^2 + 3\rho^3 + 4\rho^4 + \cdots)$, then: $\rho v = (\rho^2 + \rho^3 + \rho^4 + \cdots)$. Subtracting $v - \rho v$;

$$\rightarrow v - \rho v = \rho + \rho^2 + \rho^3 + \rho^4 + \cdots$$

$$\rightarrow v(1 - \rho) = \rho\left(\frac{1}{1 - \rho}\right)$$

$$\rightarrow v = \sum_{j=0}^{\infty} \rho^j = = \frac{\rho}{(1-\rho)^2} \quad (22)$$

Substituting Equation 21 in Equation 22, the Length of the queue L is estimated using Equation 23.

$$L = (1-\rho)\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda} \quad (23)$$

To find L_s the customers in the system at a given period should be obtained. There will always be customer service for this system with one server unless there are no customers in the system or the bank.

$$L_s = P_0 + 1(\pi_1 + \pi_2 + \pi_3) = 1 - \pi_0$$

Thus, L_s We find how many customers are in the system at any given moment, defined in Equation 24.

$$L_s = 1 - (1-\rho) = \rho \quad (24)$$

Further,

$$L_q = L - L_s = \frac{\rho}{1-\rho} - \rho$$

The average number of customers waiting for que, L_q is represented in Equation 25.

$$L_q = \frac{\rho^2}{1-\rho} \quad (25)$$

Using Little's formula, $W, W_q$and W_s are found by dividing respective L by the value of μ and are represented in Equations 26 to 28.

$$W = \frac{L}{\lambda} \quad (26)$$

$$W_q = \frac{L_q}{\lambda} \quad (27)$$

$$W_s = \frac{L_s}{\lambda} \quad (28)$$

2.12.2 Multiple Server Model: M/M/S/FCFS/∞/∞ Queuing System

The system has Poisson arrivals and exponential service times with rates λ and μ (Koole, 1998). The system has S attendants willing and ready to serve incoming customers from a single line. If $j < s$, then all customers are being attended to, while if $j > s$, then the system has customers with A being served while remaining $j - s$ waiting in the line. The death rate dependents on how many attendants are serving. Thus, it is modelled as a death-birth system. The completion rate of each attendant is μ; to get

the actual death rate μ times the number of customers attended to at various service points in the bank.

Parameters of interest for the system are represented in Equations 29 to 30.

$$\lambda_j = \lambda \; for \; j = 0,1,2, \dots, \infty \tag{29}$$

$$\mu = \begin{cases} \mu & for \; j = 0,1, \dots, s \\ s\mu & \text{for } j = s+1, s+2, \dots, \propto \end{cases} \tag{30}$$

In this system, steady-state probabilities can be found using the below-defined utilization factor in the same manner above using the flow balance equations. The utilization factor or traffic intensity is estimated using Equation 31.

$$\rho = \frac{\lambda}{s\mu} \tag{31}$$

The probability of having a customer in the system is given by Equations 32 and 33.

$$\pi_0 = \left[\sum_{n=0}^{s-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^n \frac{s\mu}{s\mu - \lambda} \right]^{-1} \tag{32}$$

$$\pi_n = \begin{cases} \left(\frac{\rho^n}{n!}\right) \pi_0 \; if \; n \leq s \\ \frac{\rho^n}{(s! \, s^{n-s}) \pi_0} \; if \; n > s \end{cases} \tag{33}$$

When $n \geq s$, the number of customers in the system exceeds the number of servers, the next customer has to wait; that is

$$C(s, \rho) = \sum_{n=s}^{\propto} \pi_n = \frac{\rho^s}{s! \, (1 - \rho^s)} \pi_0 \tag{34}$$

The expected number of customers waiting in the queue is represented in Equation 2.36.

$$L_q = \left[\frac{1}{(s-1)!} \left(\frac{\lambda}{\mu}\right)^s \frac{\mu\lambda}{(s\mu - \lambda)^2} \right] \pi_0 \tag{35}$$

At any given time, the expected number of customers in the system is given by Equation 2.37.

$$L_s = L_q + \frac{\lambda}{\mu} \tag{36}$$

The expected waiting time for customers in the system and queue is obtained using Equations 37 and 38, respectively.

$$W_s = \frac{L_s}{\lambda} \tag{37}$$

$$W_q = \frac{L_q}{\lambda} \tag{38}$$

2.12.3 The M/G/∞/GD/∞/∞ and GI/G/∞/GD/∞/∞ Queuing Systems.

This kind of system is set to have infinite servers; thus, a customer never waits in a queue before receiving the service (Chaudhry, 1992). The best example is online self-service, like internet shopping. The average waiting time and queue length are represented in Equations 39 to 40.

$$W = {}^{1}/_{\mu} \tag{39}$$

$$L = {}^{\lambda}/_{\mu} \tag{40}$$

The steady-state probability at state j is estimated using Equation 41.

$$\pi_j = \frac{\left({}^{\lambda}/_{\mu}\right)^j}{j!} e^{-\left({}^{\lambda}/_{\mu}\right)} \tag{41}$$

2.12.4 The Machine Repair Model. M/M/R/GD/K/K Queue System

According to Krause & Musingwini (2007), the system has R servers, K as the customer population size, and the system's capacity. This model explains a scenario where there are K machines with a breakage rate of λ and R repair workers each can fix a machine at a rate μ. In this case, λ and μ depend on the number of machines left in the population or the number of repair workers. To model this as a birth-death process, it's noted that λ_j depends on the number of unused machines, as represented in Equation 42.

$$\lambda_j = (K - j)\lambda \tag{42}$$

μ_j is calculated by considering the number of employees in service. Similar to previous models, if a machine breaks down when all the employees are engaged, it waits in a queue to be attended to. The service rate is calculated using Equation 43.

$$\mu_j = \begin{cases} j\mu & for\ j = 0,1,\ldots,R \\ R\mu & for\ j = R+1, R+2,\ldots,K \end{cases} \tag{43}$$

The steady-stead probability for a system is represented in Equation 44.

$$\pi_j = \begin{cases} \binom{K}{j} p^j \pi_0 \ \ for\ j = 0,1,2,\ldots,R \\ \dfrac{\binom{K}{j}\, j\, p^j \pi_0}{R!\, R^{J-R}} \ for\ J = R+1, R+2, \ldots, K \end{cases} \tag{44}$$

2.12.5 The M/G/S/GD/S/∞ Queuing System.

This is another reasonable model for a customer to arrive, seeing all busy servers exits the system without being served. In such a case, no queues occur, and the system is said to be cleared (Berry, 2006). Moreover, $L_q = W_q = 0$ no queue ever formed. If λ is the arrival rate and $1/\mu$ is the mean service time, then,

$$W = L_q = 1/\mu \tag{45}$$

It works in such a way that arrivals are turned back when s customers exist, so π_s is equal to the proportion of arrivals turned back. This implies an average of $\lambda\pi_s$ arrivals per unit time cannot enter the system. Thus, $\lambda(1-\pi_s)$ arrivals per period enter the system, leading to Little's queuing formula deduction represented in Equation 46.

$$L = L_S = \frac{\lambda(1-\pi_s)}{\mu} \tag{46}$$

2.13 Conceptual Framework

Figure 1 represents the basic queueing model with three servers in the queueing system.

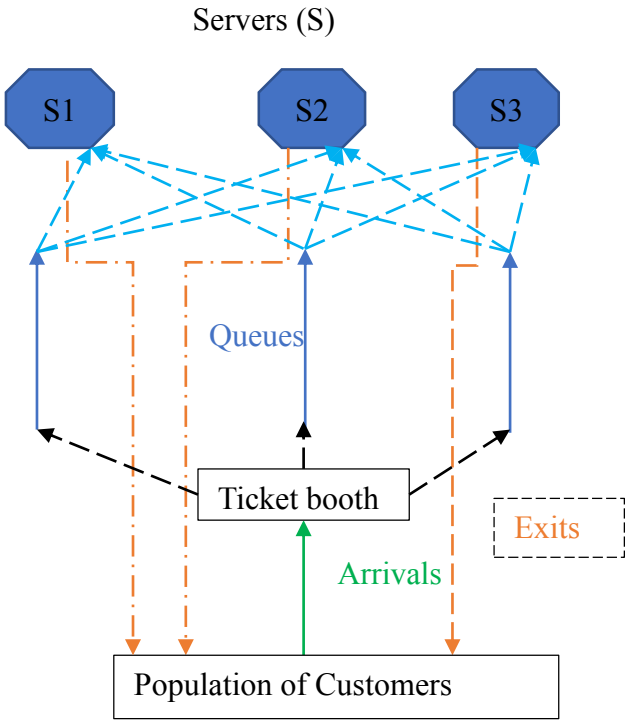

Figure 1: Queueing System

The customer arrival from a population of customers denotes the input process. The arrival of customers is assumed to follow a Poisson distribution with an average rate of λ customers per minute. The customers then pick a ticket and join the queue awaiting to receive the services from c servers. At this point, waiting time is used to proxy the bank's performance. Customers then proceed to the counter to receive a customer care desk (general inquiries, internet banking, Chequebook, and statement collection) or cash-related services (such as cash deposits, cash withdrawals, and fund transfers) on FCFS basis. That is, in ascending order of their ticket numbers) from any of the free servers at the bank. Each teller will have different service rates, depending on the type of customer's service, and consequently will serve a different number of customers. The presence of multiple but identical servers, c, providing services to customers denotes a multi-server queuing model (/M/c queuing model) since several servers are more than one customer can be processed simultaneously (Farayibi, 2016). Conjointly, the servers are assumed to operate together as a unit or a system; hence the entire banks' models converge to M/M/1 queuing model (Udoh, 2021). At last, the departure of a served customer from the bank ends the process.

CHAPTER THREE
METHODOLOGY

3.1 Location of the Study

The study was conducted in Chuka, Kenya, while applying queuing models to study, analyse, and understand the behaviour of queues in the banking industry. The selected bank requested anonymity due to the sensitivity of the data involved.

3.2 Research Design

The study applied a descriptive research design to investigate the dynamics of queuing in a banking hall in a commercial bank in Kenya. The design is appropriate since the study entails an in-depth analysis of queuing dynamics of a bank, including service times, waiting times, customer arrivals, and service intensity. A Single server system (M/M/1) queuing model (treating the bank as a unit), with queuing discipline of the form FCFS assumed in the queuing process. Performance metrics of banks based on waiting and service time for peak (the period before the outbreak of COVID and during school opening days) and off-peak seasons were compared.

3.3 Population of the Study

The study targeted all customers in the selected commercial bank banking hall during peak and off-peak months of service delivery. Peak periods occur mainly during the opening months of January, May, and September. The peaks can extend to a month before or after these peak months, while the rest are often off-peak periods. Besides, the outbreak of COVID-19 leads to the issuance of regulatory measures by the ministry of health of Kenya to ease congestion at work and in public places such as churches and markets. In 2020, the protocol to cap the spread of COVID-19 included closing churches and markets in most regions across the country. The banking halls recorded few customers, including fear instilled in the citizens of possible contagion. In this regard, the study considered the year before (2019) during stringent measures to cap the spread of COVID-19 (2020) and after easing the COVID-19 restrictions (2021). The study periods were explicitly chosen to compare the normal and peak periods.

3.4 Sampling Procedure and the Sampling size

The study did not encompass a probabilistic sampling technique since it is not a sample survey. Given that the sampling frame is months, May over 2019, 2020, and 2021 were purposively selected to represent peak periods, whereas July was purposively selected to represent off-peak periods. May is usually school opening month in Kenya; hence higher demand for services such as fee deposits and cheques. July is an off-peak period since it is two months after opening school; therefore, few or no customers still visit the bank to deposit school fees.

3.5 Data Collection

This study relied on secondary data recorded daily for May and July 2019, 2020, and 2021. The target variables are server ID, the number of customers, waiting time, type of service, and service times over purposive targets peak and off-peak periods of May and July across three years: 2019 to 20121. The data on the three target variables were collected on service days from Monday to Saturday within working hours. Banks' operating hours are 8.30 am to 4 pm (7 ½ hours) on weekdays and 8.30 am to 12 pm (3 ½ hours) on Saturdays. Servers were given anonymous codes denoted as Teller 1 (T1) to Teller n (Tn), where n is the number of tellers in a given period. The number of customers, waiting time, and service times were recorded aggregately by type of service and teller. The type of services were broadly classified as customer care desk service (General Enquiries, Shares, internet banking, pin/cheque book collection, ATM card collection, statement collection/standing orders) and cash-related services (cash deposits, cash withdrawals, bankers' cheque, fund transfer, and international services which are RTGS/SWIFT and Western Union/MoneyGram/forex). The data was compiled in a data collection schedule (Appendix I, page 84).

3.6 Data Analysis

Data analysis utilised descriptive statistics. Frequencies and percentages were used to summarize the distribution of customers registered across the study periods by type of service received at the Bank. The performance metrics comprised the length of the system or queue in times of time, waiting time, service time, system intensities, and service rates were estimated using Little's law by study period and server. The Bank's queuing performance metrics were computed assuming a M/M/1 queuing model on

FCFS queuing service discipline. The Kendall Notation of the model is M/M/1/C/P/D explained as follows. The first M: is the arrival process, assumed in this study to follow a Poisson distribution because the number of arrivals is counts. If λ_{itm} denotes the number of arrivals, i, during year t month m. λ_{tm} follows a Poisson distribution expressed in Equation 47.

$$P\{\lambda_{itm}\} = n = \frac{(\lambda_{tm})^n e^{-\lambda_{tm}}}{n!} \; \forall \; i \; \geq 0 \; ; 2019 \leq \; t \leq 2021, m = May, July \qquad (47)$$

From Little's queuing formula, where λ ; is the arrival rate, is the number of customers arriving in the system per unit of time; $1/\lambda$ is the average time between arrivals and n; is the expected number of arrivals per unit. Setting n = 0, we get a Poisson distribution giving us. $e^{-\lambda t} \; \forall \; t \geq 0$, which is an exponential distribution. The distribution determines the probability of arrivals in a given time(t). The reduced function is useful when relating the possibility that zero appearances occur in a given period.

The second M denotes service time distribution which is an exponential distribution. The number of customers served per period is called the service rate, μ. The bank Qmatic system, designated as a flow management system, is expected to bring order into the system and Equity. The fluctuation in the number of attendants from A to A + 1 will be considered to model the bank system. The bank's queue is a process with input values λ, μ, and A. The system has A attendants willing and ready to serve incoming customers from a single line. If $n < A$, then all customers are being attended to, while if $n > A$, then the system has customers with A being served while remaining $n - A$ waiting in line.

Define τ_i as the time when i^{th} customer arrive $\forall \; i \geq 1$; we define to be $\tau_i = t_{i+1} - t_i$ be interarrival time. All $\tau_i's$ are independent and continuous random variables denoted by A with a probability distribution function (pdf) of $a(t)$. A is has an exponential distribution with parameter λ defined by $a(t) = \lambda e^{-\lambda t}$ for the output process, assuming that service times are independent random variables denoted by S with probability density represented in Equation 48.

$$s(t) = \mu e^{-\mu t} \qquad (48)$$

Where; μ is the service rate, the number of customers per hour. The utilization factor or traffic intensity is estimated using Equation 49.

$$\rho = \frac{\lambda}{A\mu} \tag{49}$$

As defined above, the bank's queue is a process with input values λ, μ, and A. A set of output values is (L_q, L_s, W_q, W_s) and the intermediary values $(P_0, P_1, P_2, ...)$. The length of the queue, L, is estimated using Equation 49.

$$L = \frac{\rho}{1-\rho} \tag{50}$$

The steady-state probability of getting no customer in the system is represented using Equation 51.

$$P_0 = \left[\sum_{n=0}^{A-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{A!}\left(\frac{\lambda}{\mu}\right)^n \frac{A\mu}{A\mu-\lambda}\right]^{-1} \tag{51}$$

Thus, the probability that at least all the tellers at the bank were free at any point, P_0, is estimated using Equation 52.

$$P_0 = 1 - \frac{\lambda}{\mu} \tag{52}$$

Consequently, the probability that all the bank (tellers) is busy (P_b) is $1 - P_0$. The steady-state probabilities of having busy servers in the system / the customer in the system is represented in Equation 53.

$$P_n = \begin{cases} \left(\frac{\rho^n}{n!}\right) P_0 \; if \; n \leq A \\ \frac{\rho^n}{(A!\, A^{n-A}) P_0} \; if \; n > A \end{cases} \tag{53}$$

When $n \geq A$, the number of customers in the system exceeds the number of servers, the next customer has to wait, as represented using Equation 54.

$$C(\mathrm{A}, \rho) = \sum_{n=A}^{\infty} P_n = \frac{\rho^A}{A!\,(1-\rho^A)} P_0 \tag{54}$$

The length of queues (L) can be split into L_q and L_s to represent the average number of customers waiting in queue and in-service using Equation 55.

$$\mathrm{L} = L_q + L_s \tag{55}$$

The expected number of customers waiting in the queue is computed using Equation 56.

$$L_q = \left[\frac{1}{(A-1)!}\left(\frac{\lambda}{\mu}\right)^A \frac{\mu\lambda}{(A\mu-\lambda)^2}\right] P_0 \tag{56}$$

Application of Little's laws (Winston & Goldberg, 2004) gives the following relationships among (L_s, L_q, W_s) and W_q, at any given time. W can be integrated into W_q and W_s representing the average time spent in the queue and in-service, respectively, as represented in Equation 57.

$$\mathrm{W} = W_q W_s \tag{57}$$

The expected number of customers in the system is computed using Equation 58.

$$L_s = L_q + \frac{\lambda}{\mu} = \lambda W_s \tag{58}$$

The average number of customers in the queue (L_q) is computed using Equation 59.

$$L_q = \lambda W_q \tag{59}$$

The expected waiting time for customers in the system (W_s) and in the queue (W_q) are obtained as represented using Equations 60 and 61.

$$W_s = W_q + \frac{1}{\mu} \quad = \frac{L_s}{\lambda} \tag{60}$$

$$W_q = \frac{L_q}{\lambda} \tag{61}$$

Collectively, the system has Poisson arrivals with λ as the average arrival rate of customers and exponential service times with rates, μ, given "S" as the number of servers *k*.

From the Kendall Notation, S = 1 indicates a single server. While the bank has multi-server queues, each server is analysed individually, and the bank is treated aggregately as a unit. In accordance with this design, M/M/1 queuing model was adopted. C is the number of buffers (system capacity), the maximum number of customers that can be accommodated in the bank. It comprises those in the queue and those at the service point (Charity Ojochogwu Egbun *et al*., 2020). The number of tellers/servers in banks can determine the system's capacity. The more tellers available, the higher the system capacity and the shorter the customer waiting time (Kabamba, 2019). This study also

notes that the usual system capacity might have been driven by exogenous shocks such as the recent Covid-19 pandemic. In 2020, many organizations had to restructure their office layouts to ensure that seats are no closer than 1 meter away in accordance with the Ministry of Health's COVID-19 regulation. Thus, the system capacity could be approximately half the normal capacity.

P: is the population size. The size of the calling population is essential in queueing systems because customers enter a queue system since it influences arrival rates. Every arrival implies that the calling population reduces since the customers served are removed from the list of potential customers. For a finite population, the mean arrival rate will decrease with every arrival. If only a few possible customers or arrivals are present at a certain time, we say the calling population is infinite. This is the assumption for most queuing models (Oo, 2019). The assumption is that customers' arrival rate is unaffected by the number of customers (constant) already joining the queuing system. In the current study setting the commercial bank, the arrival rate is uneven and tends to be high in peak than off-peak periods.

D is the Service Discipline and refers to how customers are selected for service from the queue. This banking system in the selected commercial bank adopts a FCFS discipline. Under the FCFS service discipline in the banks' queueing systems, customers are served on a FCFS basis. The first customer in the queue is served first, followed by the second customer, *et cetera*.

Data analysis was carried out using two Software. The performance metrics were computed using R's (R Studio Team, 2020) 'queueing' package, version 0.2, 12 (Canadilla, 2019). R was preferred since it can be used to plot the resultant Poisson distribution for the arrival process in all the study periods. The resulting output was presented in Tables or line plots generated using Microsoft (MS) Excel (Microsoft Corporation, 2018) to indicate trends. MS Excel was also used to produce bar blots visualizing the distribution of customers by type of service in each of the study periods. The simulation tasks determining the optimal number of servers were also done using MS Excel.

3.7 Performance Metrics for the M/M/c Model Formulas

Performance metrics in queueing models include the average waiting time, the service times, and the probability of the system's states, such as empty, full, and having an available server or customers in the system. To measure the system's effectiveness, the study adopts performance metrics documented in the empirical literature, including Mwangi & Ombuni (2015) and Bakari et al. (2014). The parameters of the model are described below.

(i) Average arrival time (ART)

$$ART = \frac{\sum Inter\ arrival\ time}{Number\ of\ customers} \tag{62}$$

(ii) Average service rate (ASR)

$$ASR = \frac{\sum Service\ time}{Number\ of\ customers} \tag{63}$$

(iii) System intensity (utilization) (ρ)

$$\rho = \frac{ART}{ASR} \tag{64}$$

(iv) L= the average number of customers in the system and is computed as:

$$L = \frac{\rho}{1-\rho} \tag{65}$$

(v) L_q is the average length of the queue

$$L_q = \frac{\rho^2}{1-\rho} \tag{66}$$

(vi) W denotes the average time taken by the customer in the system

$$W = \frac{1}{\mu - \lambda} \tag{67}$$

(vii) W_q is the average time taken by the customer in the queue

$$W_q = \frac{\rho}{\mu - \lambda} \tag{68}$$

(viii) W(t) = the probability that a customer takes more than t units of time in the system.

$$W(t) = e^{-\frac{t}{w}} \quad (t \geq 0) \tag{69}$$

(ix) $(W_q(t)$ is the probability that a customer spends more than t unit of time in the queue is given as:

$$W_q(t) = \rho e^{-\frac{t}{w}} \quad (t \geq 0) \tag{70}$$

3.8 Ethical Considerations

The Chuka University Ethics Committee approved the research proposal (Appendix II, page 85). The research permit was obtained from NACOSTI before proceeding with the research (Appendix III, page 86). The selected bank requested anonymity due to the sensitivity of the data involved. Citations were used to avoid plagiarism.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Estimation of The Number of Customers Registered in The Bank.

This study applied queuing theory models to study, analyse and understand the behaviour of queues in the banking setting, a case study of a selected commercial Bank in Kenya. The study findings are presented and discussed objectively as follows. The study's first objective was to estimate the average time customers spend in the queue and the number of customers in Kenya's commercial bank for cash-related services. The results are presented in a period as follows.

4.1.1 Peak Periods

The total number of customers registered in May 2019 was 3,937, comprising 644 (16.36%) who received customer care desk services and 3293 (83.64%) who received cash-related services. Figure 2 depicts the distribution of customers by cash-related services offered during May 2019. The most common service offered is cash withdrawal registering 1, 738 customers (44.1%), followed by cash deposits (n = 1137, 28.9%), bankers' cheques (n = 314, 13.4%), statement election (n = 210, 5.3%). In contrast, the rest of the services constituted less than 5% of the total services, with the least being NHI/KRA services (n = 11, 0.3%). International services include RTGS/SWIFT and Western Union/MoneyGram/forex.

In May 2020, 2,383 customers were served, comprising 372 customers who received customer care desk services, excluding general inquiries (18.7%) and 1614 (81.3%) who received cash-related services. The total number of registered customers slightly dropped by about 39.5% compared to 3,937 registered in the same month in 2019. Figure 3 depicts the distribution of customers by cash-related services offered during May 2020. Like in May 2019, cash withdrawal registered the highest number of customers (703, 35.4%) but significantly less than customers receiving the same services in May 2019 by about 59.56%. Cash deposits ranked second in terms of services (n = 575, 29.0%). Other services ranked below 10% and are summarised in Figure 2.

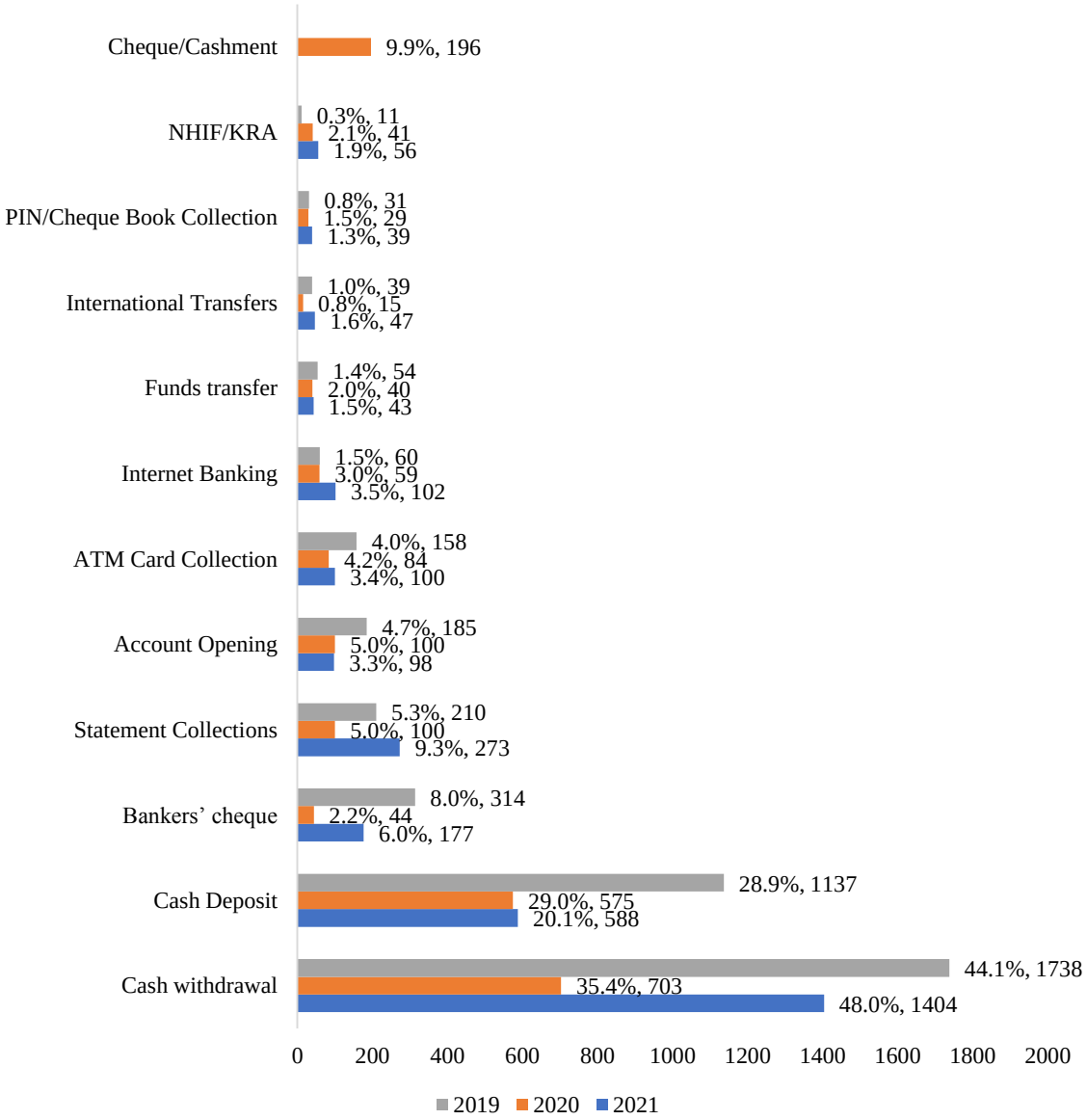

Figure 2: Number of customers registered in May 2019, 2020, and 2021 by service

The 2,927 customers served in May 2021 comprised 612 customers who received customer care desk service (20.9%) and 2315 (79.1%) who received cash-related services. The total number of registered customers increased by about 22.83% from 2020 (n = 2, 383) and decreased by approximately 34.51% compared to 2019 in the same month (n = 3, 937). Figure 3 depicts the distribution of customers by cash-related services offered during May 2021. Like in May 2019 and May 2020, cash withdrawals registered the highest number of customers. It denotes a 19.22% decrease from 2019 but a significant increase of 99.72% compared to May 2020 (Figure 2).

4.1.2 Off-peak Periods

Since schools are open, July is the off-peak period in banks, and fewer customer deposits are likely to be recorded. In July 2019, 2,201 customers were served with the common services offered in most banks, comprising 183 (8.3%) customers who received customer care desk services and those who received cash-related services in 2018 (91.7%). The total number of registered customers substantially dropped by about 21.39% compared to 2 800 registered in May 2019. The results are consistent with the expectations that July is off peaks compared to May, where banks register an increased demand for high cash-related services such as cheques and school fees deposits as high school resumes their second term studies. Figure 3 depicts the distribution of customers by cash-related services offered during July 2019, 2020, and 2021. Cash withdrawals registered the highest number of customers (1099, 49.9%) in July 2019, followed by cash deposits (n = 750, 34.1%). Other services are below 5% and are presented in Figure 3.

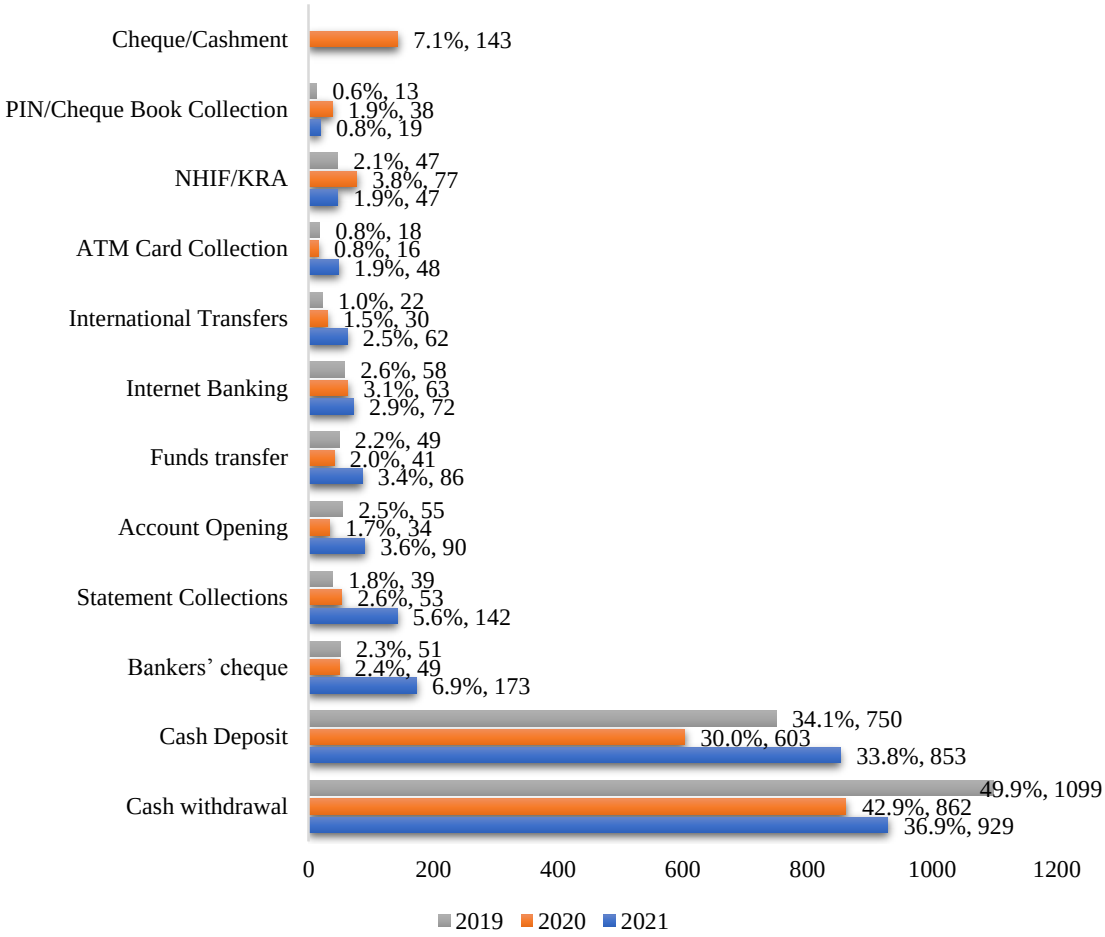

Figure 3: Number of customers registered in July 2019, 2020, and 2021 by service

In July 2020, 2,009 customers were served with the common services offered in most banks, including 204 (10.2%) customers who received customer care desk services and 1805 (89.8%) who received cash-related services. The total number of customers registered was slightly higher than those registered in July 2019 by about 20.31% due to increased customer care desk services but still a substantial drop by approximately 28.25% from May 2019. Cash withdrawals registered the highest number of customers (862, 42.9%) in July 2020, followed by cash deposits (n = 603, 30.0%). Other services are below 10% and are presented in Figure 3. The total number of customers registered in July 2021 was 2,521, comprising 371 (14.7%) who received customer care desk service and 2150 (85.3%) who received cash-related services. The most common service is cash withdrawal registering 929 customers (36.9%) in July 2021, followed by cash deposits (n = 853, 33.8%). Other services such as account opening, funds transfer, international services, and NHIF/KRA registered low proportions below 10%.

The distribution of customers by type of service provides important insights. The most dominant services at the banks are cash deposits, withdrawals, and cheques. Such services are unavoidable, especially for business individuals dealing with large transactions and needing privacy or security. Business individuals might be dealing with transactions worth millions that can be loaned to finance their business. ATMs and bank agents may not have the capacity or authorization to execute such services; hence the customers would have to visit the bank. Some schools have historically encouraged students to use banks to pay fees, partly promoted by financiers such as government and non-governmental organizations. And well-wishers who use cheques to dispatch their funds to the school's accounts. Srinivas and Wadhwani (2019).

The study further established that customers prefer bank branches over digital means when opening new accounts and cash-related services like loans due to low trust in online financial systems. The distribution also highlights how mobile banking services play a role in decongesting banks. Services like NHIF/KRA and international or customer to customers' transfers are now offered remotely using mobile-based applications without necessarily traveling to banks. Besides, some banks allow customers to create bank accounts remotely using their phones. Thus, such services are in less demand. It is also evident that May is the peak season since it coincides with

school opening days, usually January, May, and September in Kenya. However, other months register low service demand, predominately cheques for fee payments and school fees.

4.1.3 Comparison of Customers Registered for Cash-Related Services

A comparison of cash-related services (Cash withdrawal, Cash Deposit, banker's cheque, Funds transfer, International Transfers, NHIF/KRA) in the total cash-related services. Figure 4 compares cash-related services in July 2019, 2020, and 2021. The number of customers registered for cash-related services was 3293, 1614, and 2315 in May 2019, 2020, and 2021, respectively. The Cheque/Cashment records for 2019 (n = 196) were excluded in 2020 due to merging the paired records in 2019 and 2020. On average, all the cash-related services recorded in May 2020 were relatively lower than in 2019 and 2021, except NHI/KRA services, which seemed to have an increasing trend from 2019 to 2021. It was also observed that the margin of difference between services offered in May 2021 and May 2021 varied depending on the service offered. For instance, a relatively equal proportion of deposits was registered in 2020 and 2021, but bankers' cheques and cash withdrawals significantly increased in 2021 compared to 2020.

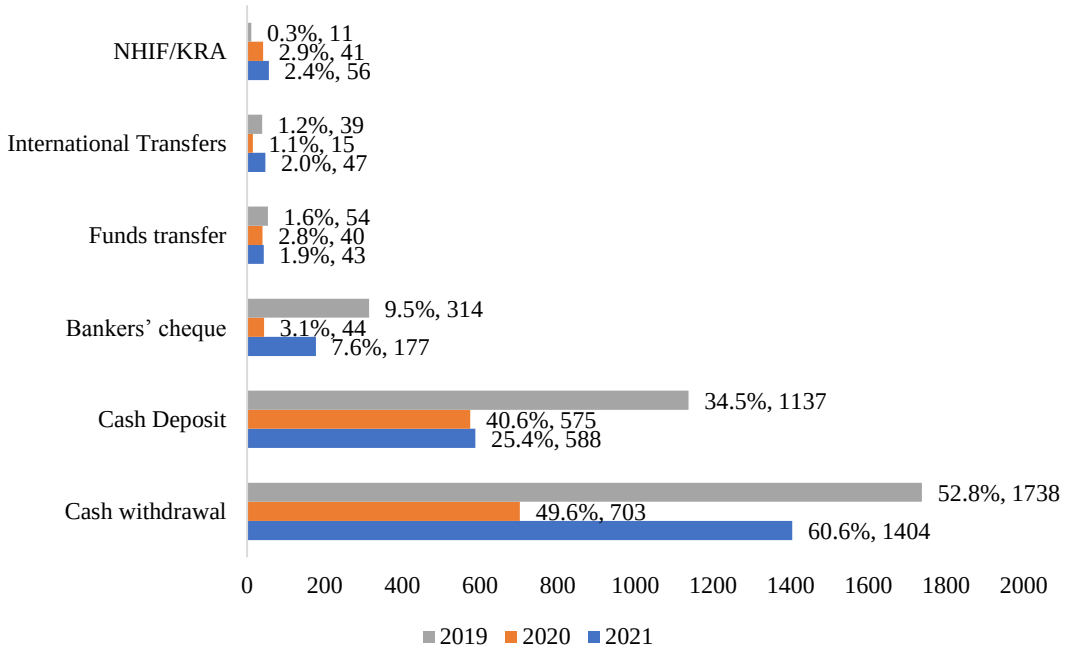

Figure 4: Comparison of cash-related services in May 2019, 2020, and 2021

Figure 5 compares cash-related services in July 2019, 2020, and 2021. In July, the number of customers registered for cash-related services was 2,018, 1,662, and 2,150 in 2019, 2020, and 2021, respectively. The Cheque/Cashment records for 2019 (n = 143) were excluded in 2020 to merge the paired records in 2019 and 2020. Like in May 2019 and 2021, the average cash-related services recorded in May 2020 were relatively lower than in 2019 and 2021, except for NHI/KRA services, which seemed to have increased in 2019 and decreased in 2021. The margin of difference between services offered in May 2021 and May 2021 varied by type of service. For instance, a relatively equal proportion of bankers' cheques were registered in 2020 and 2021. However, deposits and withdrawals declined in 2020 and rose in 2021.

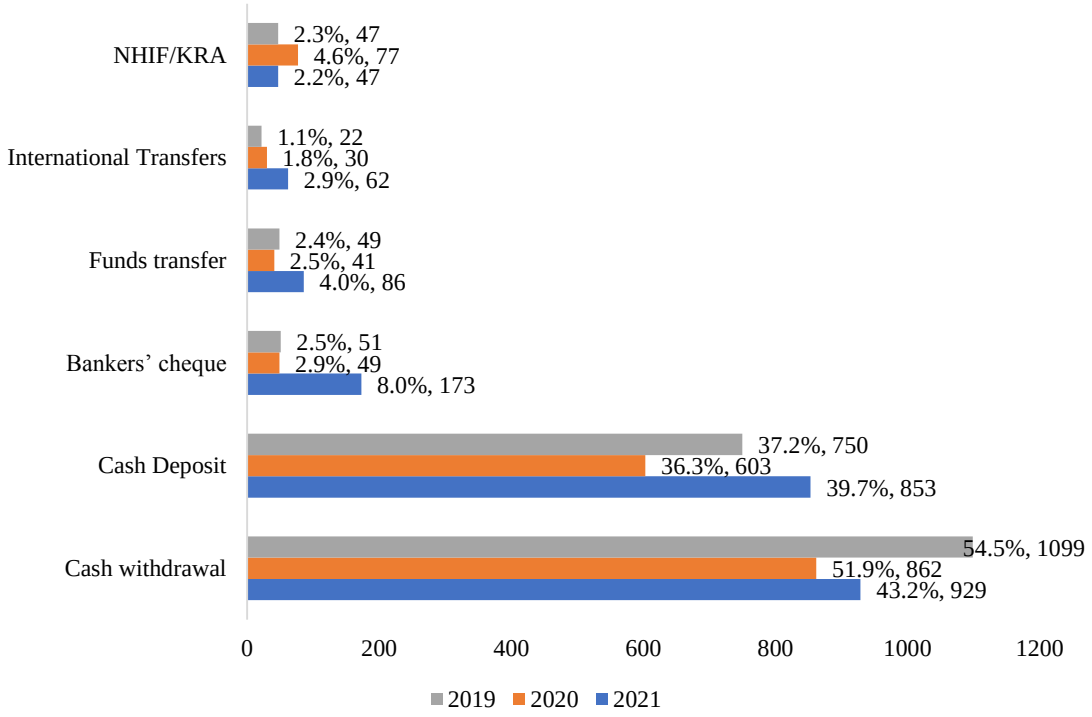

Figure 5: Comparison of cash-related services in July 2019, 2020, and 2021

4.2 Average Waiting Time and Service Times by Cash-Related Services Received

This section examines how average waiting and service times vary by type of cash-related service received. Average waiting time refers to customers waiting in the queue (Agyei *et al*., 2015). The study defines average service time as the length customers receive by type of cash-related service or teller. These performance metrics are analysed chronically, starting with the peak periods (Mays) followed by off-peak (July).

4.2.1 Peak Periods

In the banking sector, service times and mechanisms are dynamic and diverse depending on the time of the day, week, month, and services offered. Overall, the key metrics are usually the service times of consecutive arrivals into the baking halls denoted by $S_1, \ldots, S_n$ A sequence is assumed to be independent and identically distributed. Table 2 summarizes the average waiting and service time by type of cash-related services in May 2019.

Table 2: Performance metrics by service in May 2019

	Waiting Time				Service Time			
Service	Min	Max	Average	SD	Min	Max	Average	SD
NHIF/KRA	05:45	23:11	14:28	12:20	00:18	01:19	00:48	00:43
Bankers' cheque	08:29	23:15	14:00	06:26	00:24	01:47	01:02	00:37
International Transfers	12:02	26:05	15:32	05:56	00:22	01:43	01:10	00:32
Cash withdrawal	09:33	11:40	10:49	00:54	00:52	01:49	01:16	00:27
Funds transfer	03:59	52:48	24:13	20:57	00:13	04:03	02:17	01:41
Cash Deposit	00:58	11:15	04:39	04:46	03:17	04:31	03:49	00:30
Average	00:58	52:48	13:35	10:53	00:13	04:31	01:52	01:21

Across the tellers, an average service time of 1 minute 52 seconds ($\pm$ 1 min 21 sec) and an average waiting time of 13 minutes 35 seconds ($\pm$ 10 min 53 sec) was registered in May 2019. Most services requiring more service time include cash deposit opening (AWT = 3 minutes 49 seconds) and Funds Transfer (AWT = 02 minutes 17 seconds). Long waiting is an opportunity cost since less time is available for other activities, leading to high customer dissatisfaction (Ahmed *et al*., 2018; Hongna & Zhenwei, 2010; Mwangi & Ombuni, 2015; Ngugi, 2016). While the realized average waiting time is above 10 minutes, it is normal and sustainable given that other studies have identified waiting times higher than those obtained in this study. For instance, a Kenyan Mwangi and Ombuni (2015) survey revealed an average waiting time of 33.415 minutes. Regarding service times, Williams, Ogege, and Ideji (2014) established an average mean service time of 1.0667 minutes per customer in a case study of five big Nigerian banks. Thus, the average rates established in this study are realistic.

Table 3 displays the summary statistics of waiting time, service time, and types of cash-related services in May 2020. The waiting time was reduced to 10 mins 14 seconds

(Table 3) compared to May 2019, attributed to few registered customers. However, the service time increased to 2 mins 34 seconds which can be attributed to extra precautions that servers have to ensure that customers adhere to the laid down protocol, such as putting on the mask in the recommended position and ensuring they satanize their hands before they get served (Ministry of Health [MOH], 2020a; 2020b). Nonetheless, it is difficult to isolate whether the increase in the service times of services such as cash withdrawals, international transfers, and funds transfers relative to May 2019 is from service-related bulkiness or COVID-19 protocol.

Table 3: Performance metrics by service in May 2020

	Waiting Time				Service Time			
Service	Min	Max	Average	SD	Min	Max	Average	SD
Cheque/ Cashment	05:38	10:13	07:54	02:18	01:20	02:21	01:56	00:32
Bankers' cheque	04:12	20:32	11:51	08:13	01:39	02:12	01:56	00:17
NHIF/KRA	03:27	14:48	09:17	05:41	00:21	03:25	01:57	01:32
Cash Deposit	04:18	12:29	08:36	04:07	01:43	03:18	02:23	00:49
Cash withdrawal	05:10	09:35	07:41	02:16	01:44	03:28	02:31	00:53
International Transfers	06:48	18:56	12:36	06:05	01:32	05:03	02:44	02:00
Funds Transfer	06:19	23:59	13:43	09:11	03:32	05:11	04:29	00:51
Average	03:27	23:59	10:14	05:30	00:21	05:11	02:34	01:17

Table 4 summarizes the descriptive statistics of average waiting and service time by type of cash-related services in May 2021.

Table 4: Performance metrics by service in May 2021

	Waiting Time				Service Time			
Service	Min	Max	Average	SD	Min	Max	Average	SD
International Transfers	02:29	14:21	07:15	05:08	01:12	02:56	01:51	00:39
Cash withdrawal	07:21	12:04	09:04	02:04	01:46	02:39	02:11	00:26
Bankers' cheque	01:52	18:29	10:32	06:49	01:54	02:41	02:13	00:21
NHIF/KRA	03:31	05:16	04:24	01:14	01:45	03:19	02:32	01:06
Funds transfer	03:23	22:08	12:39	09:54	01:23	04:13	02:50	01:11
Cash Deposit	01:42	09:14	05:01	03:51	03:01	04:26	03:33	00:46
Average	01:42	22:08	08:36	05:58	01:12	04:26	02:27	00:52

The waiting time was slightly reduced to 8 mins 36 seconds in May 2021 compared to May 2020. The recorded waiting time is also substantially lower than in May 2019. The average service time was reduced to 2 mins 27 seconds, which is lower than the time recorded in May 2020. The finding could imply that most of the regulations initially laid out by the ministry to be observed in all public institutions, including the banks, such as hand sanitisation and temperature checks at the door, have been flexed. The reduced adherence checks reduce congestion at the banking hall. Besides, the reduced service and waiting time could be an internal regulation adjustment to ensure that the increased customer base relative to 2020 is serviced within the stipulated working hours. For instance, cash transfer and withdrawal times are substantially reduced relative to 2020. Nonetheless, service times like funds transfer and cash deposits increased from 2 mins 23 seconds in 2020 to 3 mins 33 seconds, indicating that the service times are independent of customer-based and time period.

4.2.2 Off-peak Periods

Table 5 summarizes the statistics of waiting time, service time, and types of cash-related services in July 2019.

Table 5: Performance metrics by service in July 2019

	Waiting Time				Service Time			
Service	Min	Max	Average	SD	Min	Max	Average	SD
Bankers' cheque	00:40	04:58	01:52	02:04	01:08	03:44	02:10	01:07
International Transfers	01:36	06:11	03:17	01:47	01:33	03:14	02:26	00:38
NHIF/KRA	01:04	13:01	07:03	08:27	03:11	03:39	03:25	00:20
Funds transfer	01:50	12:04	05:15	04:37	01:59	04:22	03:25	01:03
Cash withdrawal	01:41	08:13	03:44	03:05	02:47	04:43	04:00	00:55
Cash Deposit	01:02	05:11	02:56	02:06	04:01	04:21	04:13	00:10
Average	00:40	13:01	03:46	03:25	01:08	04:43	03:11	01:05

An average service time of 3 minutes 11 seconds and an average waiting time of 3 minutes 46 seconds was registered in July 2019. The service time slightly increases relative to May 2019, but the waiting time should significantly improve by about 14 minutes. Services that require more time are cash deposit (AWT = 4 minutes 13 seconds) and cash withdrawal (AWT = 04 minutes 00 seconds).

Table 5 summarizes the average waiting and service time by type of cash-related services in July 2020. The average waiting time slightly increased in July 2020 to 5 minutes and 12 seconds relative to 2019. The average service time was slightly reduced to 3 minutes and 4 seconds relative to July 2019. Like previous periods services that require more time are cash deposit (AWT = 3 minutes 41 seconds) and cash withdrawal (AWT = 3 minutes 41 seconds), which are slightly reduced relative to 2019. (Table 6). The reduction may not be directly linked to the COVID-19 pandemic but to the type or volume of transactions and customer's experience or knowledge.

Table 6: Performance metrics by service in July 2020

	Waiting Time				Service Time			
Service	Min	Max	Average	SD	Min	Max	Average	SD
Bankers' cheque	06:14	14:04	10:11	03:55	00:59	03:44	01:56	01:33
NHIF/KRA	02:33	03:59	03:13	00:43	01:44	03:01	02:31	00:41
Cheque/ Cashment	00:58	05:47	03:01	02:29	01:33	03:54	02:39	01:11
Cash Deposit	01:39	09:12	04:54	03:53	01:47	04:05	03:18	01:19
International Transfers	02:06	04:43	03:17	01:20	03:01	04:48	03:38	01:01
Cash withdrawal	04:11	14:51	07:58	05:58	02:05	04:51	03:41	01:26
Funds Transfer	01:11	05:39	03:48	02:20	02:55	04:50	03:45	00:59
Average	00:58	14:51	05:12	03:53	00:59	04:51	03:04	01:12

Table 7 presents the summary statistics of waiting time, service time, and types of cash-related services in July 2021.

Table 7: Performance metrics by service in July 2021

	Waiting Time				Service Time			
Service	Min	Max	Average	SD	Min	Max	Average	SD
Bankers' cheque	02:52	12:15	07:42	04:14	01:59	02:51	02:25	00:21
International Transfers	00:53	10:52	05:47	04:01	01:02	04:21	02:29	01:19
Funds Transfer	03:11	17:51	10:43	07:46	01:13	03:19	02:32	00:55
Cash Deposit	02:48	12:04	06:41	04:49	01:16	03:31	02:43	01:15
NHIF/KRA	07:47	08:19	08:03	00:23	01:11	05:10	03:11	02:49
Cash withdrawal	01:33	12:04	04:58	04:55	02:54	03:40	03:16	00:19
Average	00:53	17:51	07:12	04:53	01:02	05:10	02:43	01:04

In July 2021, the waiting time increased to 7 minutes 42 seconds compared to July 2019 and 2021 (Table 7). The average service time was also slightly reduced to 2 mins 43 seconds compared to previous off-peak periods. It is worth noting that there was a reduction in the number of customers in July 2021 relative to May 2021. The probability of services likely to take longer, such as cash withdrawals, dropped by a greater margin, lowering the overall waiting time. Overall, both services and average waiting times are relatively lower than those recorded in May, and July 2020, attributed to the tendency of flexed protocol laid down by the ministry to be observed in all public institutions, such as hand sanitisation and temperature checks at the door has been flexed. Such protocol contributes to the long-run billing of customers at the banking halls.

The empirical literature has no consensus on the baseline waiting times due to heterogeneity in services provided by banks and confounding factors. Many factors can affect the waiting times in banks during peak periods, including the system arrival rates, service rates, type of services, time of the day, and the servers' efficiency (Sarkar *et al*., 2011). The waiting times in Kenya's studied Commercial Bank are similar to those established by Azumah *et al.* (2021) in July 2020. The authors established that Bank A's waiting times on three consecutive days were two and a half minutes, three minutes forty-second-, and three-minutes twenty-second. In Bank B, the respective waiting times for customer service were three and a half minutes, 10 minutes, and 4 minutes. The authors attributed the variation to differences in arrivals rates in daily bank service rates. Bank A had higher arrival rates but high service rates than Bank B. In this study, the waiting times in the months of July ranged between 1 minute 52 seconds to ten minutes 43 seconds. The high deviation is attributed to intra-day variation in service rates and customer arrivals. Servers may be unable to dispense the same service rates for similar services based on the customer experience and volume of transactions involved. Thus, low service rates may increase customers' waiting time when the customer does not speed up transaction activities such as filling out deposit slips or when the currency denominations are high and would possibly take time to fill out the deposit slip forms. Alternatively, even with existing service rates, a higher influx of customers within a given period will lead to congestion and longer waiting times.

The current study findings also add to existing literature that season substantially determines waiting times. The results indicated that peak season, usually Mays, is associated with longer waiting times than during off-peak seasons, usually Julys. The results suggest that waiting time is seasonal and is consistent with vast empirical literature that has established that arrival rates of customers could influence waiting times and play a significant role in influencing waiting times. However, the reviewed studies indicate that waiting times depend on arrival rates. High arrival rates (peak periods in this study) are characterised by longer waiting times and vice versa, as established in the current study. Cowdrey *et al*. (2018) demonstrated that waiting times for fast arrivals (interval rate of 3 minutes) under the FIFO queue is about 39.24 minutes for slow arrivals (interval rate of 10 minutes) is about 0.03 minutes. Consistent with Cowdrey *et al*. (2018), this study established that peak periods, often during the school opening months of May, are often associated with longer waiting times due to high demand for cash-related services such as fee deposits, whereas off-peak periods, often Julys when demand for fee payment services are low register. The longer waiting times during peak season are associated with the high influx of customers, outweighing exiting service rates. Banks often have a fixed number of servers, as in this case, five to six servers. Thus, high arrival rates, given existing service rates, lead to congestion at the bank and longer waiting times.

Long waiting times are often addressed by increasing the capacity or efficiency of the servers during peak periods (Sarkar *et al*., 2011). For instance, this study found that the number of servers was reduced to five compared to six in 2019 and 2021. Such a remedy reduces staff's idle time since service demand is low. This study showed an evident reduction in service demand in the studied Kenyan commercial bank during 2020; hence, peak and peak periods registered shorter waiting times. In 2021, the bank resumed its server capacity of six.

4.3 Estimating the System Intensity

System intensities are determined based on the average waiting time and service times regardless of the service. Variation in system intensities by period gives banks' performance metrics useful in accessing a bank's service delivery capacity and resilience during peak seasons. Besides, variation of average waiting time and service

by the server is helpful when examining the optimal number of servers required in a bank without heavily intensifying the bank. The results are thus presented systematically as follows.

4.3.1 Estimating the System Intensity by Period

A queuing system is in a *steady-state* if the probability that the system is in a given state is time-independent: that is $P(L(t) = n) = P_n(t) = P_n$ is independent of time t. For the entire bank, the M/M/1 model is assumed since the bank is a single entity. For illustration, the bank's respective steady-state parameters in the month of May 2019 are computed as follows using Littles law. It is important to deduce the working hours in a month as a preliminary step. Banks' operating hours are 8.30 am to 4 pm (7 ½ hours) on weekdays and 8.30 am to 12 pm (3 ½ hours) on Saturdays. However, since banks are most likely to serve existing customers beyond 4 pm and 12 pm, the hours are rounded to 8 hours working weekdays and 4 hours on weekends. Banks are usually closed on Sundays and public holidays. Thus, in Mays, we have one public holiday, labour day, on 1st May. Thus, it may have 22 weekdays and 4 Saturdays, resulting in 192 working hours. The months of July had 23 working days and 4 Saturdays hence a total of 200 working hours. The associated arrival rate, λ, is about 18 persons per hour (3293 people/192 hours).

Since population arrival at the bank is infinite, the arrivals follow a Poisson distribution with a rate λ, arrivals per period. Besides, the inter-arrival times assume an exponential distribution with a mean λ. Another assumption also holds that service times are exponentially distributed with a mean $1/\mu$ where μ is the average service rate. The service rate, μ, is about 33 persons per hour 32.429 (60/AST [1.8667]). Therefore, the associated system intensity, ρ, is 0.5336 (λ/μ). The average number of customers in the system is, L is about 2 customers. The average length of the queue or number of customers in the queue, L_q, 0.6104. Multiplying by 10 minutes gives about 7 people every 10 minutes. Using Equation 67, a customer's average time in the system (sum of time in the queue and service time) is 4 mins 0 seconds. From Equation 68, the average time the customer takes in the queue, W_q is about 2 minutes and 6 seconds. From the above calculations, the probability that at least all the tellers at the bank were free at any point during working hours in May 2019, P_0, was 0.466, estimated using Equation

52. Consequently, the probability that all the bank (tellers) was busy (P_b) is $1 - P_0 = 0.534$. Henceforth, estimated ρ_s will mean P_b. Using the same procedures, the performance metrics across the months of May and July 2019, 2020, and 2021 are summarized in Table 8.

Table 8: Performance Metrics in the months of May 2019, 2020, and 2021

Year	λ	μ	ρ	L	w	w_q	L_q	P_0
May 2019	17.151	33	0.534	1.144	4 mins 0 sec	2 min 6 sec	0.6104	0.466
May 2020	8.406	24	0.360	0.562	4 min 48 sec	1 min 22 sec	0.2019	0.640
May 2021	12.057	25	0.492	0.970	4 min 50 sec	2 min 19 sec	0.4775	0.508
July 2029	10.090	19	0.535	1.152	6 min 51 sec	3 min 33 sec	0.6167	0.465
July 2020	9.025	20	0.461	0.856	5 min 42 sec	2 min 31 sec	0.3950	0.539
July 2021	10.750	23	0.487	0.948	5 min 18 sec	2 min 29 sec	0.4616	0.513

Overall, service rates, λ_s, were lower in 2020 relative to 2021, associated with the COVID-19 regulation protocol discussed previously. It is also evident that the COVID-19 pandemic hard hit the peak season (May 2020). The low average time the customer takes in the queue can be associated with the low customer base and, consequently, a decline in demand for services that require high service time, such as deposits due to the closure of schools (Ministry of Education, 2020). It can also be evident that the bank appeared to be busy with a probability of being busy (P_0) of 0.655 higher than other periods due to the tendency of stringent checks that customers follow the correct protocol, such as putting one on the mask correctly as prescribed by Kenya's MOH (MOH, 2020a; 2020b). The servers were also lower in 2020 (5) than in other years (6). While the servers were also five in July, July is an off-peak period hence P_0 was lower in July (0.539) than in May.

From Table 8, all the periods examined recorded an average service rate between 19 and 33 persons per hour. While there could be variability in service rates due to a varying number of servers and customers, it is assumed that banks can have a predisposition to higher servers based on their capacity. From this assumption, service rates can be compared with previous empirical studies with the caution that the comparison may not guarantee a higher degree of validity. Consequently, the study established average service rates comparable to those in previous studies. For instance, a Kenyan survey by Mwangia and Ombuni (2015) revealed an average service rate is 23.7 customers per hour, and the average waiting time was 33.415 minutes. In Nigeria,

Williams, Ogege, and Ideji (2014) did a case study using five big Nigerian banks and established a service rate of 56.250 customers per hour. Thus, the study findings demonstrate that the studied banks have service rates that are relatively low to guarantee high customer-centricity. While longer waiting times have been associated with low customer satisfaction (Ahmed *et al*., 2018; Hongna & Zhenwei, 2010; Mwangi & Ombuni, 2015). However, high service rates may be prone to errors, which lower customer satisfaction with the delivery of services.

The probability of having the customer (s) i; denoted P_i In the system are also important probabilities that can also be computed directly with reference to the above-reported parameters as $P_i = P_0 * \rho^i$ (Winston, 1991). An illustration is described with reference to the performance metrics recorded in May 2019, where $\rho = 0.534$ and $P_0 = 0.566$. Thus; $P_1 = 0.249$; $P_2 = 0.133$; $P_3 = 0.0127$; $P_4 = 0.071$; $P_{\geq 5} = 0.962$. Other are the probabilities that a customer takes more than t units of time in the system, W(t), and that a customer spends more than t units of time in the queue, $W_q(t)$. For hypothetical $t = 0,1, \ldots, 5$, W(t) and $W_q(t)$ are summarised in Table 9.

Table 9: Summary of performance metrics; $W(t)$ and $W_q(t)$

	$W_q(t) = e^{-t/w}$						$W_q(t) = \rho e^{-t/w}$					
Time	May			July			May			July		
(mins)	2019	2020	2021	2019	2020	2021	2019	2020	2021	2019	2020	2021
0	1.000	1.000	1.000	1.000	1.000	1.000	0.534	0.360	0.492	0.535	0.461	0.487
1	0.779	0.775	0.813	0.864	0.839	0.828	0.363	0.499	0.413	0.402	0.452	0.425
2	0.607	0.600	0.661	0.747	0.704	0.685	0.151	0.140	0.165	0.186	0.175	0.171
3	0.473	0.465	0.537	0.645	0.590	0.567	0.063	0.039	0.066	0.086	0.068	0.069
4	0.368	0.360	0.437	0.558	0.495	0.470	0.026	0.011	0.026	0.040	0.026	0.028
5	0.287	0.279	0.355	0.482	0.415	0.389	0.011	0.003	0.011	0.018	0.010	0.011

The functions $F\{W_q(t)\}$ and $F\{W(t)\}$ can be visualized to depict the cumulative probability distribution of wq and w, assuming FCFS queue discipline. By definition, $F\{W_q(t)\}$ is the probability of a customer waiting for $T \leq t$ before being served (Gross & Harris, 2008). As shown in Figure 6, the probability of a customer having to wait at least 1 minute at the banking hall is 1. Similarly, the probability of a customer spending 1 minute in the bank is 1.

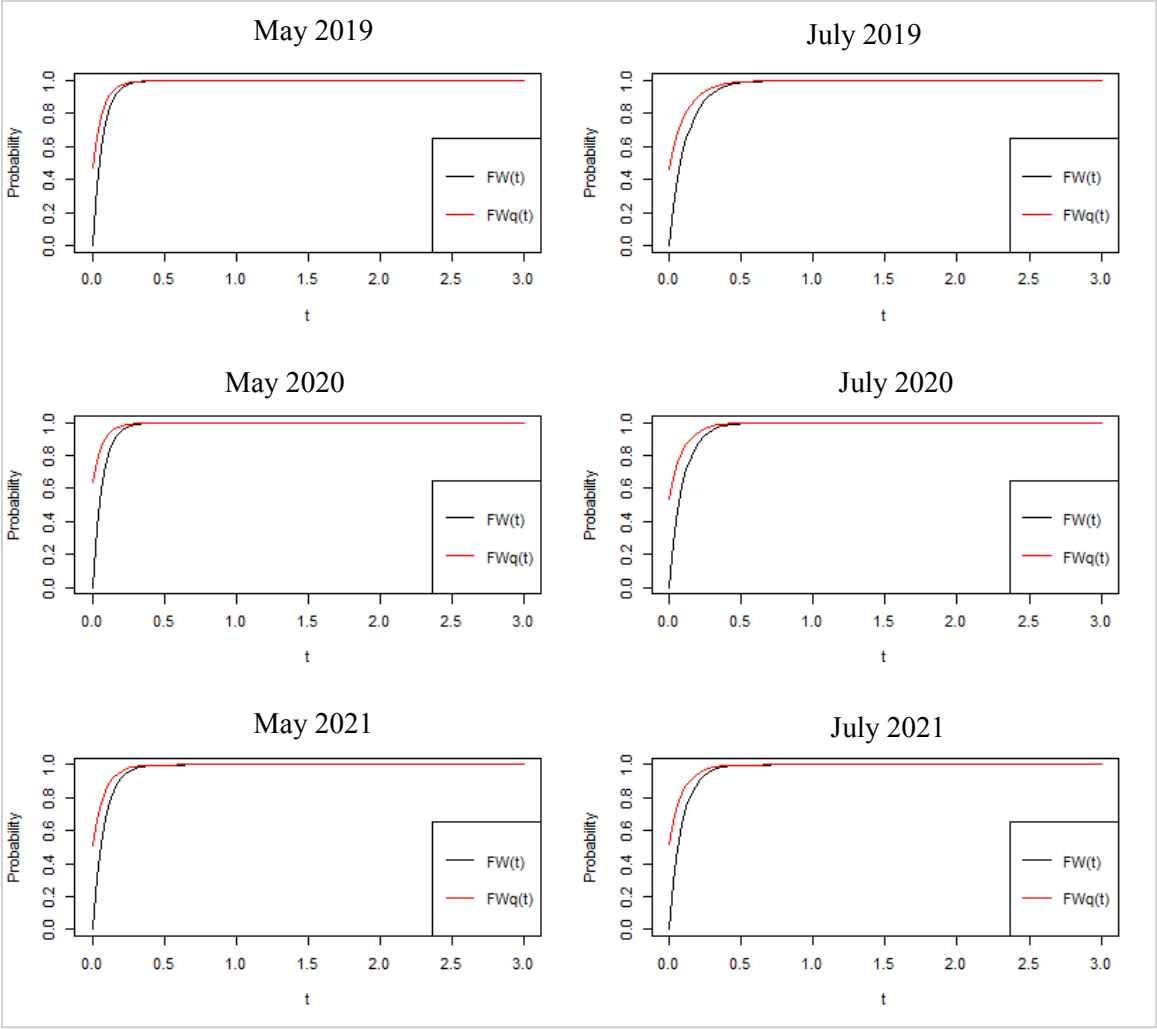

Figure 6: Distribution of random variables w and w_q across the study periods

4.3.2 Estimating the System Intensity by Server

This section provides the variation of average waiting time and service by the server and their associated intensities chronologically as follows.

4.3.2.1 Peak Periods

In a typical banking setting, a queuing system entails several service counters and interconnecting queues. Each service centre comprises several servers, c, working in parallel, serving on a FIFO basis. Parallel service mechanisms can either be a single server (c = 1) or multiple servers ($1 < c < n$). All systems are assumed to have a FIFO queue discipline where a bank teller serves the first available customer following ticket numbers provided, in most cases, by order of entry into the bank. The previous computations in 4.1 and 4.2 assumed a bank as a single entity server (c =1). Thus, looking at their service rates is vital to estimate the optimal number of servers. The

bank's setting considers a case where c = 6, whose respective performance metrics are summarized in Table 10.

Table 10: Performance Metrics in the month of May 2019 by server

Server	N	Average Waiting Time	Average Service Time	λ	μ	ρ	L	P_0
T6	353	01:37	02:23	1.839	25.175	0.073	0.079	0.927
T5	146	08:37	03:00	0.760	20.000	0.038	0.040	0.962
T4	592	06:23	03:03	3.083	19.672	0.157	0.186	0.843
T2	86	03:28	03:30	0.448	17.143	0.026	0.027	0.974
T3	652	01:47	03:32	3.396	16.981	0.200	0.250	0.800
T1	189	02:12	04:19	0.984	13.900	0.071	0.076	0.929
System	3293	13:35	01:52	17.151	32.143	0.534	1.144	0.466

The varying service rates and the number of customers served over the period reflect the variation in servers' intensities. While the service rate μ is inversely proportional to service time, the server intensity, ρ depends on the number of customers served. For instance, teller 6 (T6) had a higher service rate of 26 people per hour with a resultant service intensity of 0.073, given that the teller served 353 customers in May 2019. Comparatively, teller 1 recorded the least service rate of 14 persons per hour, with associated service intensity of 0.071, given that the teller served 189 customers over the period. Table 11 summarizes the performance metrics of each server based on the number of customers in May 2020.

Table 11: Performance Metrics in the month of May 2020 by server

Server	N	Average Waiting Time	Average Service Time	λ	μ	ρ	L	P_0
T2	244	04:33	01:21	1.271	44.444	0.0286	0.0294	0.971
T5	1142	09:44	02:00	5.948	30.000	0.1983	0.2473	0.802
T3	207	13:23	03:28	1.078	17.308	0.0623	0.0664	0.938
T4	13	23:59	03:32	0.068	16.981	0.0040	0.0040	0.996
T1	8	06:19	05:11	0.042	11.576	0.0036	0.0036	0.996
System	1614	10:14	02:34	8.406	23.377	0.3596	0.5615	0.640

Teller 2 had a higher service rate of 44 people per hour with a resultant service intensity of 0.029, having served 244 customers. Contrarily, teller 4 had the highest average service rate of 12 persons per hour with associated least service intensity of 0.0036, ranking third, having served the 8 customers. Table 12 summarizes the performance metrics of each server based on the number of customers in May 2021.

Table 12: Performance Metrics in the month of May 2021 by server

Server	N	Average Waiting Time	Average Service Time	λ	μ	ρ	L	P_0
T2	241	10:57	02:04	1.255	29.032	0.043	0.045	0.957
T6	1151	08:10	02:14	5.995	26.866	0.223	0.287	0.777
T4	272	03:57	02:15	1.417	26.667	0.053	0.056	0.947
T5	230	15:41	02:18	1.198	26.087	0.046	0.048	0.954
T3	224	07:15	02:29	1.167	24.161	0.048	0.051	0.952
T1	197	12:10	04:20	1.026	13.846	0.074	0.080	0.926
System	2315	08:36	02:27	12.057	24.490	0.492	0.970	0.508

Teller 2 had the highest service rate of 30 customers per hour with a resultant least service intensity of 0.043, having served 241 customers during May 2021. Teller 1 had the lowest service rate of 14 people per hour with associated service intensity of 0.074, ranking second, having served 197 customers over the study period.

4.3.2.2 Off-peak Periods

Table 13 summarizes the performance metrics of each server based on the number of customers in July 2019. Teller 6 had the highest service rate of 26 people per hour with an associated service intensity of 0.070, having served 353 customers during July 2021. Teller 1 had the lowest service rate of 14 people per hour with associated service intensity of 0.068, ranking fourth, having served 189 customers over the study period.

Table 13: Performance Metrics in the month of July 2019 by server

Server	N	Average Waiting Time	Average Service Time	λ	μ	ρ	L	P_0
T6	353	01:37	02:23	1.765	25.175	0.070	0.075	0.930
T5	146	08:37	03:00	0.730	20.000	0.037	0.038	0.964
T4	592	06:23	03:03	2.960	19.672	0.150	0.177	0.850
T2	86	03:28	03:30	0.430	17.143	0.025	0.026	0.975
T3	652	01:47	03:32	3.260	16.981	0.192	0.238	0.808
T1	189	02:12	04:19	0.945	13.900	0.068	0.073	0.932
System	2018	03:46	03:11	10.090	18.848	0.535	1.152	0.465

Table 14 summarizes the performance metrics of each server based on the number of customers in July 2020.

Table 14: Performance Metrics in the month of July 2020 by server

Server	N	Average Waiting Time	Average Service Time	λ	μ	ρ	L	P_0
T2	862	04:41	01:49	4.310	33.028	0.130	0.150	0.870
T3	355	06:09	03:15	1.775	18.462	0.096	0.106	0.904
T5	560	04:38	03:28	2.800	17.308	0.162	0.193	0.838
T1	16	05:39	03:31	0.080	17.062	0.005	0.005	0.995
T4	12	04:34	04:50	0.060	12.414	0.005	0.005	0.995
System	1805	05:12	03:04	9.025	19.565	0.461	0.856	0.539

Teller 2 had the highest service rate of 34 people per hour with an associated service intensity of 0.130, having served 862 customers during July 2020. Teller 4 had the lowest service rate of 13 people per hour with associated service intensity of 0.005, ranking fourth, having served 12 customers over the study period. Table 15 summarizes the performance metrics of each server based on the number of customers in July 2021.

Table 15: Performance Metrics in the month of July 2021 by server

Server	N	Average Waiting Time	Average Service Time	λ	μ	ρ	L	P_0
T5	275	11:05	01:15	1.146	48.000	0.024	0.024	0.976
T4	303	05:13	02:09	1.263	27.907	0.045	0.047	0.955
T2	145	03:51	02:23	0.604	25.175	0.024	0.025	0.976
T3	716	09:53	02:56	2.983	20.455	0.146	0.171	0.854
T1	330	10:20	03:25	1.375	17.561	0.078	0.085	0.922
T6	381	06:57	03:41	1.588	16.290	0.097	0.108	0.903
System	2150	07:12	02:43	8.958	22.086	0.406	0.682	0.594

Teller 6 had the highest service rate of 48 people per hour with an associated service intensity of 0.024, having served 275 customers during July 2021. Teller 6 had the lowest service rate of 17 people per hour with associated service intensity of 0.097, ranking second, having served 381 customers over the study period.

Generally, bank system intensities depend on the season and server. Due to high arrival rates, peak periods tend to have relatively higher system intensities. Given the same service rates, a high influx of customers at a given period leads to high system intensity since serves will have to dispense more customers. Contrarily, during off-peak periods, customers tend to have more idle time; hence the entire system may have a low utilization factor. Therefore, banks must check whether the number of servers is optimal by checking their individual utilization factors. More servers with low system intensities add to the unnecessary employment cost if few servers achieve tolerable

service intensities that ensure banks are not congested during off-peaks. The best criteria are to check individuals' service utilization factor and service rates as a proxy for retaining a few efficient employees. As demonstrated above, some servers have higher service rates with significant deviations in the number of customers they handle in a month. While several factors may affect the number of customers a sever handles in a month, the number of customers they individually serve ranged as low as 12 to 1 151 with service times of about 1 minute to about 5 minutes. For cash-related services, it is possible to complete the transaction in between 2 to three minutes (average). Thus, the variation could suggest that some servers might be more efficient in in-service times due to speed and experience, which needs further exploration by future duties. Nonetheless, the tendency that type of service delivered may influence server utilization and their service rates cannot be ignored. Cash deposits involving huge denominations may take longer than withdrawals or Cheques. Thus, Banks can use benchmark service times to check whether servers who spend more time than recommended are justified depending on the service they handle.

Generally, all system intensities are moderate and low (less than one), corroborating the existing empirical literature on commercial banks' system intensities (Cowdrey *et al*., 2018; Karoney, Kosgei & Nyongesa, 2019; Zewude & Sodo, 2016). Karoney, Kosgei, and Nyongesa (2019) findings established relatively equal waiting times for server utilization factors less than one in sampled Banks in Eldoret Town, Kenya. Cowdrey *et al*. (2018) established that system intensity rates vary depending on the arrival rates. At first arrivals, banks tend to be highly intensified ($\rho = 1$) compared to first arrivals ($\rho = 0.39$) (p 382). Their findings are consistent with this study's observation that peak periods tend to have relatively higher system intensities than off-peak periods due to high arrival rates. In another study, Azumah *et al.* (2021) did a cross-comparison of two banks and established that their arrival and service rates influence their system intensities. The authors revealed that Bank B had a higher utilization factor (0.588, 0.750, and 0.568) than Bank A (0.389, 0.354, and 0.370 in three consecutive days), attributed to a high number of customers and lower service rates in the banking hall of Bank B than that in Bank A.

The existing empirical literature and the findings established in this study demonstrate that banks do not have high system intensity. The results from this study implicitly imply that the servers are somewhat less intensified in terms of cash-related services. In an economic sense, the tellers are not optimally utilized; hence a bank, the probability of being idle is higher ($P_o > 1$) and consequently, probabilities of being busy (P_b) too low (all tellers have $\rho/P_b < 0.1$) expect T3 ($\rho = 0.146$). Thus, it is possible to reduce the number of servers as determined below.

4.3.3 Poisson distribution for the arrival process

The above analysis assumed that the arrivals follow a Poisson process. Thus, we compute the Poisson distribution for the arrivals in all the periods. For all months of May, the daily average number of customers was calculated by dividing the total number of customers in that particular period by the number of days. For example, in May 2019, 3293 customers were registered. Dividing by 26 days gives about 127 customers per day. Thus, arrival probabilities were plotted against 127 customers. Similarly, May 2020, 2021, July 2019, 2020, and 2021 averaged 63, 89,75, 67, and 80, respectively. For comparison, the minimum daily average number of customers of 63 was selected as the base line.

4.3.3.1 Peak Periods

The Poisson distribution for the arrivals in May 2019, 2020, and 2021 is shown in Figure 7.

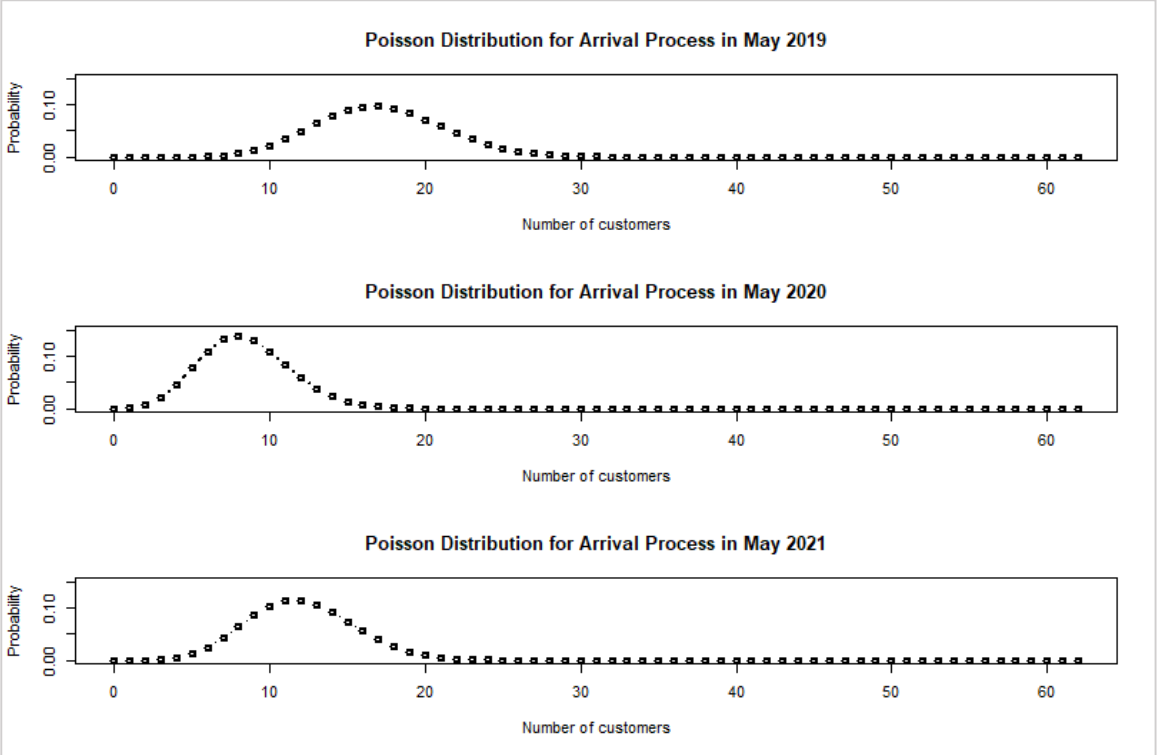

Figure 7: Poisson distribution for the arrival process during May 2019, 2020, and 2021

The distributions illustrate how COVID-19 negatively affects the utilization of Bank services. Notably, the Poisson distribution for the arrivals in May 2020 is more positively skewed than in May 2019 and May 2020. Notably, peak periods are associated with a higher influx of customers since students and parents or guardians need deposit services for fees or borrow cash for their businesses or school fees, and withdrawal for any transitionary or consumption motive. The onset of the COVID-19 pandemic led to the closure of schools. Besides, it negatively impacts business performance and prospects due to a reduction in overall consumer demand due to restricted supply chains of goods and services and job retrenchment/layoffs (Barasa *et al*., 2021; Demeke, Kariuki, & Wanjiru, 2020; Mukabana, 2021; Pahl *et al*., 2022; Xu *et al*., 2021).

4.3.3.2 Off-peak Periods

The Poisson distribution for the arrivals in July 2019, 2020, and 2021 is shown in Figure 8.

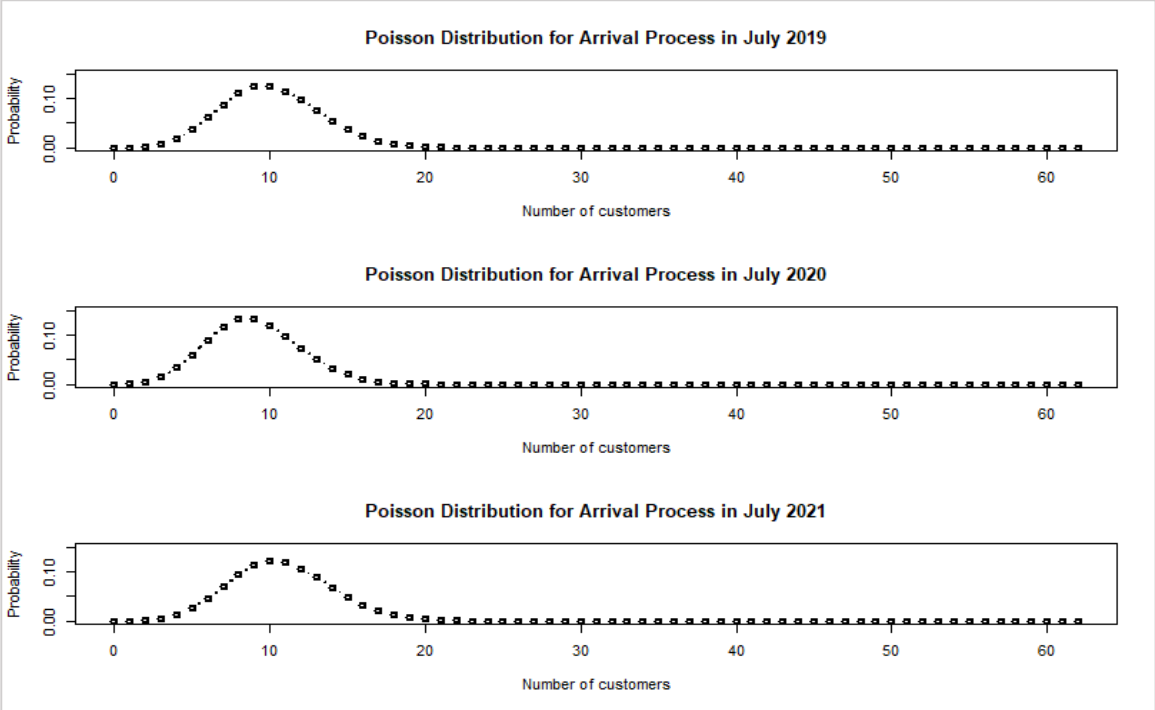

Figure 8: Poisson distribution for the arrivals during July 2019, 2020, and 2021

Compared to what was observed in the peak periods (month of Mays), the Poisson distribution for the arrivals in the off-peak periods illustrates the earlier observation that COVID-19 slightly affected the utilization of Bank services in off-peak periods. There is a slight deviation in the distribution for the arrival process in July 2020 from July 2019 and July 2020 (Figure 7) from the same periods in May (Figure 6). July is the off-peak period for banks. Since schools are open, fewer customer deposits are likely to be recorded. Thus, the impact of COVID-19 during this period tends to be lower than during peak periods, as in this case, the month of May. The perception is illustrated by the relatively low average number of customers receiving cash-related services during May than in July across the three years (see Figures 4 and 5, respectively).

4.4 Estimating Optimal Number of Staff (Servers)

Given the different service rates, the banks are constrained to minimize the service rates and increase customer satisfaction. High service rates would mean the system can clear out faster or avoid lengthy waiting times depending on the customer base recorded in a given period. Nonetheless, it is worth noting that high service rates do not always translate to low customer-centricity since services like inquiries barely take a minute or less. Besides, low service rates do not always imply high waiting time since the rate of

influx of customers within a given period is not fixed. The bank can record high arrival rates at a given hour than the other. Thus, average rates can be used as a performance metric at a given period, as in this case, monthly rates. Overall, the service intensity can be a good metric controlled by each server's average service rates. To simulate this optimization problem, a hypothetical situation is considered where a bank seeks to minimize the system intensity by employing servers with high service rates per unit of time.

As representative samples of peak and off-peak periods, the months of May and July 2019 are picked. The month of 2020 is excluded owing to the impact of COVID-19, which may not reflect the actual service and waiting times. Logistical challenges also constrained the study to consider data for 2019 and 2020 only. Severs were only 5 in 2020, May, and July, whereas 6 servers only served in May and July in 2019 and 202. Hypothetical tellers were created by averaging the number of customers, waiting times, and service rates by rank. For instance, T6 (n = 353, AST = 02:23) was the fastest in In May 2019, T2 in May 2020 (n = 241, 02:04), T6 in July 2019 (n = 353, AST = 02:23), and T5in Juk 2021 (n = 275, AST = 01:15). The averages give the first hypothetical server (HS1) who served 306 customers and an average service rate of 02 mins 01 seconds. The AWT is also computed using the same approach to all hypothetical servers HS1 to HS6. The simulated results are presented in Table 16.

Table 16: Simulated service rates and performance metrics

Server	N	Average Waiting Time	Average Service Time	λ	μ	ρ	L	P_0
HS1	306	06:19	02:01	1.561	29.752	0.052	0.055	0.948
HS2	437	07:40	02:36	2.230	23.077	0.097	0.107	0.903
HS3	400	05:08	02:41	2.041	22.360	0.091	0.100	0.909
HS4	280	08:08	03:03	1.429	19.672	0.073	0.078	0.927
HS5	465	05:17	03:14	4.574	18.557	0.246	0.327	0.754
HS6	239	05:53	04:10	1.219	14.400	0.085	0.093	0.915
System	2127	06:24	02:58	10.852	20.225	0.537	1.158	0.463

From the above results, the probability of an individual being busy (P_b) are low, with all below 0.1 expect HS5. Consequently, the probabilities of getting no client in the system are high ($P_0 > 0.7$). Assume the bank can higher one individual to serve all the 2127 cash-related services, and customers, during a given month for an average of 196 hours (Mays and July). Each individual will be highly intensified in such a case, as indicated in Table 17. For instance, HS1 had a resultant service rate of 30 persons per

hour. Given that the teller served 306 customers, HS1 initially had a service intensity of 0.052, which is low and is seen as an underutilization of the staff. However, by serving all the 2127 customers who visited the bank during a given month, the service intensity of HS1 increases to 0.365, an improvement in terms of server utilization. Table 17 presents the simulation results for all other hypothetical servers.

Table 17: Simulated service rates and performance metrics

Server	Average Waiting Time	Average Service Time	μ	ρ	L	P_0	P_b
HS1	06:19	02:01	29.752	0.365	0.574	0.635	0.365
HS2	07:40	02:36	23.077	0.470	0.888	0.530	0.470
HS3	05:08	02:41	22.360	0.485	0.943	0.515	0.485
HS4	08:08	03:03	19.672	0.552	1.230	0.448	0.552
HS5	05:17	03:14	18.557	0.585	1.409	0.415	0.585
HS6	05:53	04:10	14.400	0.754	3.059	0.246	0.754
System	06:24	02:58	20.225	0.537	1.158	0.463	0.537

Note. Each server is assumed to serve all N = 2127 customers; hence λ = 10.852 across

The main goal of the simulation is to identify how the bank can make sure that tellers are not "idle" or less intensified while at the same time ensuring that services are not delivered at high service rates at the expense of poor customer satisfaction. Thus, the simulation seeks to identify the optimal number that balances the two conflicting objectives. It is not feasible for a bank cannot operate with one server due to a possible influx of customers at peak hours or days. Empirically, Burodo, Suleiman & Shaba (2019) established that a three-server model was better than the one- or two-server model. Thus, a further hypothetical situation is considered where a bank can choose between retaining i servers ($\forall\, i = 1, 2., \ldots, 6$) based on their service rates. In this case, it is assumed that the bank rationally prefers servers with a higher service rate. To simulate this scenario, we assume that the bank can only recruit additional servers into the system with a higher marginal service rate than the potential servers ($6 - i$). In doing so, the simulated scenario also considers the distribution of the services from the ones with the least realistic service time, such as inquiries, to the most tasking services, such as opening an account. The resultant new performance metrics are computed using their cumulated service rates, starting from the server with the highest service rate to the least, as summarized in Table 18. Given that n is fixed at 2127, their system intensities are directly proportional to their intensities. As a result, λ is not considered.

Table 18: Hypothetical performance metrics by server

Server(s)	AWT	AWT	μ	ρ	L	P_0	P_b
HS1	06:19	02:01	29.752	0.365	0.574	0.635	0.365
HS1; HS2	06:59	02:19	25.899	0.419	0.721	0.581	0.419
HS1; HS2; HS3	06:22	02:26	24.658	0.440	0.786	0.560	0.440
HS1; HS2; HS3; HS4	06:49	02:35	19.672	0.552	1.230	0.448	0.552
HS1; HS2; HS3; HS4; HS5	06:30	02:43	22.086	0.491	0.966	0.509	0.491
HS1; HS2; HS3; HS4; HS5; HS6	06:24	02:58	20.225	0.537	1.158	0.463	0.537

Note. The computation of performance metrics is done considering N = 2127, λ = 10.852

A graphical depiction of the given hypothetical situation is shown in Figure 8. Hierarchically increasing the number of servers based on decreasing service rates reduces the system's service rates to a customer-centric level. From an eccentric customer point of view, the bank does not seek to achieve extremely low service intensity two-fold. First, low service intensity can imply that the service rates are too high to clear out customers as first as possible. Secondly, high or low system intensity can reflect a low customer base over a period. In this case, the service rates curve (μ) and the system intensity elbows after 3 servers. Thus, four servers give a moderate system intensity, hence can be an optimal number of servers.

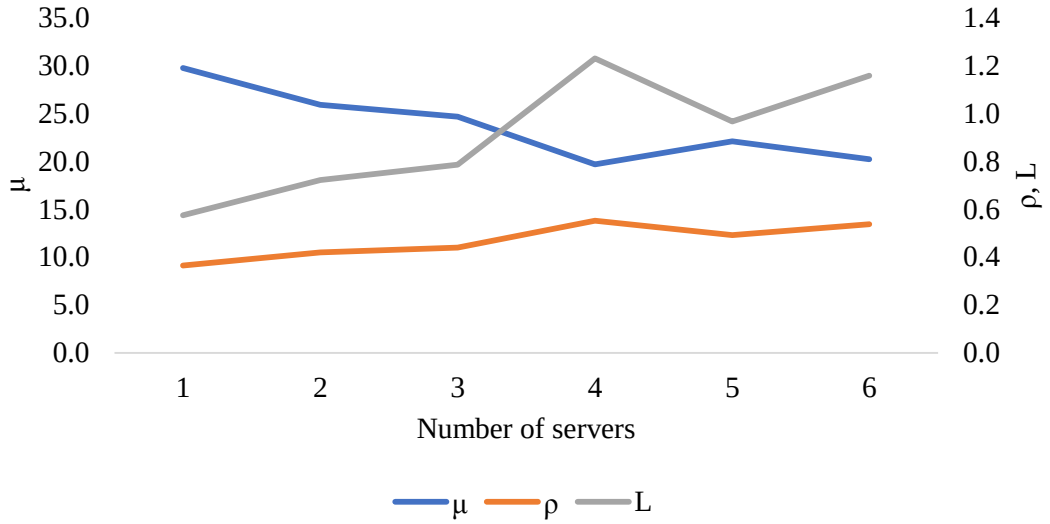

Figure 9: Graphical determination of the optimal number of servers in May 2019

The simulation analysis reveals that a bank can work with four servers with an average service rate of 02 minutes 35 seconds, with an associated service rate of 20 people per hour, and achieve an average service intensity or probability of being busy of 0.552. The results are comparable to those of Agyei *et al.* (2015), who studied a Ghanian Commercial Bank Ltd. Kumasi Main Branch found that using five-teller systems was

better than four or six-teller systems in terms of time and operational cost. During the COVID-19 pandemic in May and July 2020, as operationalised in this study, the studied bank relied on five tellers compared to other periods (2019 and 2021). However, this study reveals that the banks cannot be highly intensified with four customers. Five servers yield a system intensity of 0.491, whereas six yields an intensity of 0.537 compared to 0.522 with four. The reduced servers have the economic benefit of reduced banks' operational costs (Garvin, 1988).

Cowdrey *et al*. (2018) found that four servers would have yielded the exact bank utilization as would five servers under both first arrival rates of about 3 minutes interval (about 20 customers per hour) (ρ = 0.49) and during slow arrival rates of about 10 minutes interval (ρ = 0.33). The authors noted that one, two, and three servers lead to high intensities of 1.0, 0.91, and 0.65 under slow arrival rates and, respectively, 1.0 during first arrival rates. Therefore, their findings are consistent with the recommended four-server system in studied Kenya's commercial bank. In another study, Azumah *et al.* (2021) recommended that the management of Bank B adopt a three-server (M/M/3) model from the existing M/M/2 model since it had a higher average utilization factor (0.6349) than Bank A (0.3710) which relied on three servers. Since demand for banking services may vary across banks, system intensities provide a good proxy when recommending optimal number servers. In this study, Kenya's commercial bank should reduce its servers from six to four and achieve moderate system intensity as obtained under a five-server and a six-server model.

Nonetheless, Banks' customer influx and service duration are dynamic and unpredictable due to varying customer experiences (Green, 2006). For instance, customers with little experience filling out deposit forms may take more time to complete than those experienced and knowledgeable. This might explain the recorded longer service times for the same time point at different periods or by a teller. However, this recommendation tends to be average and is sustainable in the long run.

CHAPTER FIVE

SUMMARY, CONCLUSION, AND RECOMMENDATION

5.1 Summary of the Findings

It is common to find long queues in banks resulting in longer waiting times. Generally, queues occur when the demand for service exceeds the service rates. Consequently, longer waiting times can result in higher customer dissatisfaction and low customer retention. In turn, such negative feelings can transcend the staff leading to service inefficiency. The study adopted the M/M/1 queuing model and a FCFS queuing discipline to analyse and understand the behaviour of queues in the banking industry, a case study of a selected commercial Bank in Kenya. A purposive sample of two peaks (May) and off-peak (July) months were selected for 2019, 2020, and 2021. The study findings confirmed this assertion as the average number of customers receiving cash-related services was relatively lower in May than in July across the three years. July is the off-peak period for banks. Since schools are open, fewer customer deposits are likely to be recorded. On the contrary, banks can register a higher influx of customers during May as students and parents or guardians might need to make deposits for fees, borrow cash for their businesses or school fees, and withdraw for their transitionary purpose. Nonetheless, there was a higher margin in the difference between customers registered in May and July 2019 compared to other months' number of customers registered in 2019 and 2021.

The number of customers registered for cash-related services was 3293, 1614, and 2315 in May 2019, 2020, and 2021, respectively. The number of customers registered for cash-related services was 2,018, 1,805, and 2,150 in July 2019, 2020, and 2021, respectively. The findings also demonstrated the negative impact of COVID-19, which hard-hit the usual peak periods of May. However, unlike other cash-related services, which declined in 2020, NHI/KRA services seemed to have increased in 2020. The trend can be attributed to possible panic behaviour regarding the possibility of being hospitalised due to COVID-19 behaviours. The pandemic might have induced Kenyans to increase health insurance-seeking behaviour. However, studies such as Barasa *et al*. (2018) have indicated increased trends in Health Insurance Coverage in Kenya since 2008.

The study findings revealed that the average time customers take in the queue and the number of customers in the system is higher during peak periods than off-peak periods. In the peak periods, the average waiting time of 13 minutes 35 seconds was registered in May 2019, 10 minutes 14 seconds in May 2020, and 8 minutes 36 seconds in May 2021. In the off-peak periods, an average service time of 3 minutes 46 seconds was registered in July 2019, 5 minutes 12 seconds in July 2020, and 7 minutes 42 seconds in July 2021. The pooled data for 2019 and 2021 yielded an average waiting time of 6 minutes and 24 seconds. The high average time the customer takes in the queue during peak periods can be associated with the higher demand for cash deposits since months of Mays are usually opening days, when most students deposit their school fees. Comparatively, it is an expectation that the demand for cash deposits will decrease since schools tend to advocate for fee clearance at the beginning of school terms. School closure in 2020 could further explain why the time the customer spent in the queue was much lower than in 2019 and 2021.

The study findings further revealed that service time significantly increased during Covid-19. The average service time of 1 minute 52 seconds in May 2019, 2 mins 34 seconds in May 2020, and 2 mins 27 seconds in May 2021. In the off-peak periods, an average service time of 3 minutes 11 seconds was registered in July 2019, 3 minutes 4 seconds in July 2020, and 2 mins 43 seconds in July 2021. The increased service times during 2020 can be attributed to stringent checks that customers follow the correct protocol, such as putting one on the mask correctly as Kenya's MOH prescribes (MOH, 2020a; 2020b). For a reasonable generalization, 2020 data were excluded from the estimation of expected service time. Using the pooled data for 2019 and 2021, a customer should expect an average service time of 2 minutes and 58 seconds for cash-related service at a commercial bank.

The study findings further indicated that banks are less intensified during off-peak periods. In the peak periods, the service rates averaged 33, 24, and 25 persons per hour in May 2019, May 2020, and May 2021. The respectively associated system intensities are 0.534, 0.360, and 0.492. In the off-peak periods, the average service rates were 19, 20, and 23 persons per hour in July 2029, July 2020, and July 2021. The respectively associated system intensities of 0.535, 0.461, and 0.487. Overall, the system intensities are low to moderate. Notably, service rates were lower in 2020 than in 2021 associated

with the COVID-19 regulation protocol discussed previously. It is also evident that the COVID-19 pandemic hard hit the peak season (May 2020). COVID-19 did not substantially impact the number of customers registered in July 2020 since it is an off-peak period. The pooled data for 2019 and 2021 only yielded an average service rate of 20 persons per hour with a service intensity of 0.537.

Lastly, the study sought to estimate the optimal number of staff (servers) during a pandemic like COVID-19 in Kenya's banking sector. The studied bank utilized five servers during the COVID-19 pandemic (as registered in May and July 2020). However, this study established that a bank could work with an optimal four servers with an average service rate of 02 minutes 35 seconds, with an associated service rate of 20 people per hour, and achieve a moderate average service intensity or average probability of being busy of 0.552.

5.2 Conclusion

The study findings revealed that queuing dynamics in a typical Kenya commercial bank for cash-related services have seasonality. The monthly demand for cash-related services could range between 2300 to 3300 customers during peak periods and between 2000 and 2200 during off-peak periods. The average time customers take in the queue and the number of customers in the system is higher during peak periods, with an estimated waiting time between 7 minutes to 18 minutes than during off-peak periods, with an estimated waiting time between 2 to 13 minutes. On average, a customer would expect to stay in the queue for about 6 minutes and 24 seconds before receiving a service in a bank. Service times generally ranged between 1 minute to 4 minutes but substantially increased, on average, during the Covid-19 pandemic (2020). In any time period, a customer should expect an average service time for cash-related service of 2 minutes and 58 seconds at a commercial bank. Thus, a typical bank visit will last about 9 minutes and 22 minutes. The study findings further indicated that banks are highly intensified during peak periods, serving between 24 to 33 persons per hour and 19 to 23 persons per hour during off-peak. On average, a bank's service rate is about 20 persons per hour with a service intensity of 0.537. Generally, the commercial bank is not highly intensified and can operate with four servers with a moderate system intensity of 0.5.

5.3 Recommendation of the Study

The following recommendations are offered based on the findings of the case study.

i. Banks should achieve average waiting times of 6 minutes and service times of 3 minutes.
ii. Banks should seek to achieve an average service rate of 20 persons per hour with a service intensity of 0.537.
iii. Banks can work with four servers and achieve a moderate average service intensity.

5.4 Suggestions for Further Study

The following recommendations are provided for future studies applying queuing theory in the banking sector.

i. Since customers receiving different services might have different service-time distributions, future studies can consider computing service rates for each type of service.
ii. Future studies can also consider panel data to include more than one bank or cross-section data incorporating several banks.

REFERENCES

Abiodun, R., & Omosule, N. (2015). Queuing Model for Banking System: A Comparative Study of Selected Banks in Owo Local Government Area of Ondo State, Nigeria. *American Journal of Engineering Research (AJER)*, *4*(8), 191-195.

Agyei, W., Asare-Darko, C., & Odilon, F. (2015). Modeling and analysis of queuing systems in banks:(A case study of Ghana Commercial Bank Ltd. Kumasi Main Branch). *International Journal of Scientific & Technology Research*, *4*(07), 160-163.

Ahmed, S., Rahaman, S., Hamid, M., & Moral, I. (2018). Expected Actual Waiting Time and Service Delicery Evidence Using Queuing Theory in Selected Banking Institutions in Bangladesh. *Journal of International Business and Management*, *1*(2), 1-14.

Anderson, E., & Sullivan, M. (1993). The antecedents and consequences of customer satisfaction for firms. *Marketing science*, *12*(2), 125-143.

Arghish, O., Azadi, S., Anvari, A., & Honarvar, A. (2012). Cellular manufacturing system design with queueing approach", J. *Basic Appl. Sci. Res*, *2*(11), 11884-11890.

Arshed, N., & Kalim, R. (2021). Modelling demand and supply of Islamic banking deposits. *International Journal of Finance & Economics*, *26*(2), 2813-2831.

Asuming-Brempong, S., & Antwi, F. (2013). Improving customer service at the Agricultural Development Banks in Ghana: an application of the queuing theory. *African journal of management research*, *21*(1), 18-33.

Atefi, K., Yahya, S., Rezaei, A., & Erfanian, A. (2016, May). Traffic behavior of Local Area Network based on M/M/1 queuing model using poisson and exponential distribution.

Azumah, S., Addor, J., Twenefour, F., & Baah, E. (2021). Stochastic Model of Waiting Time: A Case of Two Selected Banks in the Sekondi-Takoradi Metropolis. *Open Journal of Statistics*, *11*(5), 906-924.

Baffour, K., & Anokye, Y. (2014). *Modelling queuing system in the banking industry: Case study Ghana Commercial Bank, Suame, Kumasi Ecobank, Ashtown, Kumasi and Barclays Bank of Ghana, Tanoso Branches* (Doctoral dissertation).

Bakari, H., Chamalwa, H.., & Baba, A. (2014). Queuing process and its application to customer service delivery (A case study of Fidelity Bank Plc, Maiduguri). *International Journal of Mathematics and Statistics Invention*, *2*(1), 14-21.

Barasa, E., Kazungu, J., Orangi, S., Kabia, E., Ogero, M., & Kasera, K. (2021). *Assessing the indirect health effects of the COVID-19 pandemic in Kenya* (Vol. 2021). Washington, DC, USA: Center for Global Development.

Barasa, E., Rogo, K., Mwaura, N., & Chuma, J. (2018). Kenya National Hospital Insurance Fund Reforms: implications and lessons for universal health coverage. *Health Systems & Reform*, *4*(4), 346-361.

Berry, R. Y. A. N. (2006). Queuing theory. *Senior Project Archive*, 1-14.

Bhat, U. (2015). *An introduction to queueing theory: modeling and analysis in applications*. Birkhäuser.

Billy, Brian, Joe Paru, Lui Masti, Herod Malo, and Mirzi Betasolo. (2018). "Application of Queuing Model in a Banking Service in PNG." *Available at SSRN 3108466*

Burodo, M., Suleiman, S., & Shaba, Y. (2019). Queuing Theory and ATM Service Optimization: Empirical Evidence from First Bank Plc, Kaura Namoda Branch, Zamfara State. *American Journal of Operations Management and Information Systems*, *4*(3), 80-86.

Campbell, D., & Frei, F. (2011). Market heterogeneity and local capacity decisions in services. *Manufacturing & Service Operations Management*, *13*(1), 2-19.

Canadilla, M. (2019). Package 'queueing'. *Analysis of queueing networks and models, version 0.2*, *12*. URL: https://CRAN.R-project.org/package=queueing>.

Cederborg, P., & Larsson-Hytte, A. (2015). What are the Entrepreneurial Management Practices and Their Impacts on Internationalization?: A study on Swedish SMEs, from a Dynamic Capabilities Perspective.

Chen, Y., Zhang, X., Bian, B., & Li, H. (2019). Optimal Staffing Policy in Commercial Banks Under Seasonal Demand Variation. *IEEE Access*, *7*, 121111-121126.

Cowdrey, K., de Lange, J., Malekian, R., Wanneburg, J., & Jose, A. C. (2018). Applying queueing theory for the optimization of a banking model. *Journal of Internet Technology*, *19*(2), 381-389.

Dallerup, K., Jayantilal, S., Konov, G., Legradi, A., & Stockmeier, H. (2018, July 18). *A bank branch for the Digital age*. McKinsey & Company. https://www.mckinsey.com/industries/financial-services/our-insights/a-bank-branch-for-the-digital-age

Demeke, M., Kariuki, J., & Wanjiru, M. (2020). Assessing the impact of COVID-19 on food and nutrition security and adequacy of responses in Kenya. *FAO Policy Briefing May*.

Dhari, K., & Rahman, T. (2013). Case Study for Bank ATM Queuing Models. *IOSR Journal of Mathematics*, 01-05.

Diaz-Aviles, E., Pinelli, F., Lynch, K., Nabi, Z., Gkoufas, Y., Bouillet, E., ... & Salzwedel, J. (2015, October). Towards real-time customer experience prediction for telecommunication operators. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1063-1072). IEEE.

Egbunu, C., Onyekwere, O., Rufai, M., Yange, T., & Atsanan, S. (2020). Queue Management in Non-Tertiary Hospitals for Improved Healthcare Service Delivery to Outpatients. *International Journal of Applied Information Systems (IJAIS)*, *12*(31).

Emmons, W. (2021, December 24). *Slow, steady decline in the number of U.S. Banks continues*. Saint Louis Fed Eagle. https://www.stlouisfed.org/on-the-economy/2021/december/steady-decline-number-us-banks

Eze, E., & Odunukwe, A. (2015). On Application of Queuing Models to Customers Management in Banking System. *American Research Journal of Bio Sciences*, *1*(2), 14-20.

Fader, P. (2020). *Customer centricity: Focus on the right customers for strategic advantage*. Wharton digital press.

Farayibi, A. (2016). Investigating the application of queue theory in the Nigerian banking system. *Available at SSRN 2836966*.

Garvin, D. (1988). *Managing quality: The strategic and competitive edge*. Simon and Schuster.

Genga, K. (2018). *Electronic Queueing Management System and Customer Service in Commercial Banks in Kenya: a Case Study of Kenya Commercial Bank* (Doctoral dissertation, University of Nairobi).

Ghimire, S., Thapa, G., Ghinire, R., & Silvestrov, S. (2017). A Survey on Queueing Systems with Mathematical Models and Applications. *American Journal of Operation Research*, *7*(1), 1-14.

Githae, L., Gatauwa, J., & Mwambia, F. (2018). Factors affecting uptake of agency banking services among customers in rural Kenya: A case of Narok County. *European Scientific Journal, ESJ*, *14*(16), 224-245.

Green, L. (2006). Queueing analysis in healthcare. In *Patient flow: reducing delay in healthcare delivery* (pp. 281-307). Springer, Boston, MA.

Gross, D., & Harris, C. (2008). Fundamentals of Queueing Theory (John Wiley & Sons). *Inc: New Jersey*.

Güneş, M. (2012). Modeling and Performance Analysis with Discrete-Event Simulation g y. *Computer Science, Informatik*, *4*.

Hanin, L. (2001). Iterated birth and death process as a model of radiation cell survival. *Mathematical biosciences*, *169*(1), 89-107.

Hongna, S., & Zhenwei, D. (2010, October). Simulation of banks queuing system based on WITNESS. In *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)* (Vol. 15, pp. V15-402). IEEE.

Kabamba, A. (2019). Modeling and Analysis of Queuing Systems in Banks:(A Case Study of Banque Commerciale du Congo-BCDC/Mbujimayi).

Karoney, K., Kosgei, M., & Nyongesa, K. (2019). Modelling the mean waiting times for Queues in Selected Banks in Eldoret Town-Kenya. *Asian Journal of Probability and Statistics*, 1-9.

Kasum, A., Abdulraheem, A., & Olaniyi, T. (2006). Queue Efficiency in Nigeria Banks: A Comparative Analysis of Old and New Generation Banks. *Ilorin Journal of Sociology*, *2*(1), 162-172.

Kim, A. (2006). *Community building on the web: Secret strategies for successful online communities*. Peachpit press.

Kim, S. (2012). Conceptual Design Based on Substance-Field Model in Theory of Inventive Problem Solving. *International Journal of Innovation, Management and Technology*, *3*(4), 306-309.

KIprono, D. (2017). Utility analysis of an intensive care unit model using queuing theory with improved touch loss function.

Klausmeier, D., Wong, K., Nguyen, Q., Sue, C., Hughes, D., Heitkamp, R., & Gomez, R. (2002). *U.S. Patent No. 6,430,191*. Washington, DC: U.S. Patent and Trademark Office.

Kohli, A., Jaworski, B., & Shabshab, N. (2019). Customer centricity: a multi-year journey. In *Handbook on Customer Centricity*. Edward Elgar Publishing.

Koole, G. (1998). Structural results for the control of queueing systems using event-based dynamic programming. *Queueing systems*, *30*(3-4), 323-339.

Krause, A., & Musingwini, C. (2007). Modelling open pit shovel-truck systems using the Machine Repair Model. *Journal of the Southern African Institute of Mining and Metallurgy*, *107*(8), 469-476.

Mangkona, S., & Murdifin, I. (2018). Implementation of Queue Model for Measuring the Effectiveness of Suzuki Car Maintenance.

Mayhew, L., & Smith, D. (2006). Using queuing theory to analyse completion times in accident and emergency departments in the light of the Government 4-hour target.

Mehandiratta, R. (2011). Applications of queuing theory in health care. *International Journal of Computing and Business Research*, *2*(2), 2229-6166.

Menasce, D., Almeida, V.., Dowdy, L., & Dowdy, L. (2004). *Performance by design: computer capacity planning by example*. Prentice Hall Professional.

Meyer, C., & Schwager, A. (2007). Understanding customer experience. *Harvard business review*, *85*(2), 116.

Microsoft Corporation. (2018). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Miller, D. (1981). Computation of steady-state probabilities for M/M/1 priority queues. *Operations Research*, *29*(5), 945-958.

Ministry Of Education, State Department of Early Learning and Basic Education (July 2020b). Guidelines On Health and Safety Protocols for Reopening of Basic Education Institutions Amid Covid-19 Pandemic. Available at https://www.education.go.ke/images/COVID-19_GUIDELINES.pdf

Ministry Of Health (10th April 2020a). Interim guidance for public use of face masks to reduce droplet transmission for COVID 19. Available at: https://www.health.go.ke/wp-content/uploads/2020/06/MoH-Guidance-on-use-of-face-masks-and-gloves-for-general-public_final2-1.pdf

Ministry Of Health (9th June 2020b) Interim Guidance for Health and Safety Measures in Workplaces in The Context of Covid-19. Available at: https://www.health.go.ke/wp-content/uploads/2020/06/INTERIM-GUIDANCE-FOR-HEALTH-AND-SAFETY-IN-WORKPLACE.pdf

Mukabana, S. (2021). *Effects of Covid-19 on Export Trade in Kenya: the Case of Leather Industry in Kariokor Market* (Doctoral dissertation, University of Nairobi).

Mwangi, S., & Ombuni, T. (2015). An empirical analysis of queuing model and queuing behaviour in relation to customer satisfaction at Jkuat students finance office. *American Journal of theoretical and applied statistics*, *4*(4), 233-246.

Nguli, R. (2016). *Internal factors affecting customer satisfaction of commercial banks in Kitui town* (Doctoral dissertation).

Ni, Q., Hu, L., Vinel, A., Xiao, Y., & Hadjinicolaou, M. (2010). Performance analysis of contention based bandwidth request mechanisms in WiMAX networks. *IEEE Systems Journal*, *4*(4), 477-486.

Odhiambo, F., Orwa, G. & Odhiambo, R. (2017). Application of queuing theory to vehicular traffic on Nakuru total road stretch. *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, *30*(1), 295-309.

Odior, A. (2013). Application of queuing theory to petrol stations in Benin-City area of Edo state, Nigeria. *Nigerian Journal of Technology*, *32*(2), 325-332.

Odirichukwu, J., Lekara, T., & Odii, J. (2014). Banking queue system in Nigeria. *Comput Inf Syst Develop Inform Allied Res J*, *5*(1), 95-106.

Oo, C. (2019). *Application Of Queuing Model for Banking System: A Case Study of Public and Private Banks in Yangon City* (Doctoral dissertation, MERAL Portal).

Pahl, S., Brandi, C., Schwab, J., & Stender, F. (2022). Cling together, swing together: The contagious effects of COVID□19 on developing countries through global value chains. *The World Economy*, *45*(2), 539-560.

Paxson, C. (1993). Consumption and income seasonality in Thailand. *Journal of political Economy*, *101*(1), 39-72.

Pei-Chun, L., & Ann, S. (2006): "Service efficiency evaluation of automatic teller machines- a study of Taiwan financial institutions with the application of queuing theory". *Journal of Statistics and Management Systems*, Vol.9, No.3, pp 555-570.

Prasad, S., Praveen, J., Tiwari, A., Prasad, K., Bindu, P., DON-THI, R., & Mahaboob, B. (2018). An application of LPP-graphical method for solving multi server queuing model. *International Journal of Mechanical Engineering and Technology*, *9*(4.10), 1066-1069.

Quarm, R. (2016). *Modelling Queuing system in Healthcare centres. A case study of the dental department of the Essikado Hospital, Sekondi* (Doctoral dissertation).

RStudio Team (2020). *RStudio: Integrated Development for R. RStudio*, PBC, Boston, MA URL http://www.rstudio.com/.

Sarkar, A., Mukhopadhyay, A., & Ghosh, S. (2011). Improvement of service quality by reducing waiting time for service. *Simulation Modelling Practice and Theory*, *19*(7), 1689-1698.

Schlechter, K. (2015). Hershey Medical Center to open redesigned emergency room. *The Patriot-News*.

Sevcik, K., & Mitrani, I. (1981). The distribution of queuing network states at input and output instants. *Journal of the ACM (JACM)*, *28*(2), 358-371.

Sewe, N. (2019). *Stochastic Analysis of Single Queue Single Server versus Single Queue Multiple Servers Models: A Case Study Of Post Bank and Kenya Commercial Bank* (Doctoral dissertation, Maseno University).

Shah, S., Gherghina, Ș., Dantas, R., Rafaqat, S., Correia, A., & Mata, M. (2023). The Impact of COVID-19 Pandemic on Islamic and Conventional Banks' Profitability. *Economies*, *11*(4), 104.

Sheikh, T., Singh, S., & Kashyap, A. (2013). A study of queuing model for banking system. *International Journal of Industrial Engineering and Technology*, *5*(1), 21-26.

Singh, S., Jaiswal, J., & Tiwari, S. (2007). Maximum entropy condition in queueing system M/M/1:∞/FCFS. *National Academy Science letters*, *30*(7-8), 237-242.

Skorobogatov, S. (2012). Physical attacks and tamper resistance. In *Introduction to Hardware Security and Trust* (pp. 143-173). Springer, New York, NY.

Srinivas, V., & Wadhwani, R. (2019). Recognizing the value of bank branches in a digital world. *Deloitte Insights. Deloitte Center for Financial Services. https://www2. deloitte. com/us/en/insights/industry/financial-services/bank-branch-transformation-digital-banking. html*.

Statista Research Department (2022, May 24). *Europe: Average number of customers per bank branch 2018*. Statista. https://www.statista.com/statistics/944084/average-number-of-customers-per-bank-branch-in-europe-by-country/

Tian, H., & Tong, Y., (2011): "Study on Queuing System Optimization of Bank Based on BPR". 3rd International Conference on Environmental Science and Information application Technology (ESIAT), pp 640-649.

Udoh, I. (2021). Resources Allocation Qu Multi-Server Petroleum Prod. IJTSRD. *International Journal of Trend in Scientific Research and Development (IJTSRD)*, 5(2), 2456 – 6470.

Varma, M. (2016). Minimization of Traffic congestion by using queueing theory. *IOSR Journal of Mathematics (IOSR-JM)*, *12*(1), 116-122.

Whitt, W. (1986). Deciding which queue to join: Some counterexamples. *Operations research*, *34*(1), 55-62.

Williams, H., Ogege, S., & Ideji, J. (2014). An Empirical Analysis of Effective Customers Service on Nigeria Banks Profitability. (A Queuing and Regression Approach). *Asian Economic and Financial Review*, *4*(7), 864.

Winston, W. (1991). *Operations research: applications and algorithms* (pp. 380-384). Boston: PWS-Kent publishing company.

Winston, W., & Goldberg, J. (2004). *Operations research: applications and algorithms* (Vol. 3). Belmont^ eCalif Calif: Thomson/Brooks/Cole.

Xiao, H., & Zhang, G. (2010, January). The queuing theory application in bank service optimization. In *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)* (Vol. 2, pp. 1097-1100). IEEE.

Xu, Z., Elomri, A., El Omri, A., Kerbache, L., & Liu, H. (2021). The compounded effects of COVID-19 pandemic and desert locust outbreak on food security and food supply chain. *Sustainability*, *13*(3), 1063.

Zewude, B., & Sodo, E. (2016). Queuing Modeling for Comparative Study of Banking System on Commercial Bank of Ethiopia Tona Branch and Dashen Bank: The Case of Wolaita Zone, Ethiopia. *Research Journal of Finance and Accounting*, *7*(21), 11-16.

APPENDICES

Appendix I: Data Collection Schedule

Period	Teller	Service	Waiting Time	Service Time
May 2019	T1	A	# mins # secs	# mins # secs
	T2	B	# mins # secs	# mins # secs
	TN			
May 2020	T1			
July 2021	TN	Z	# mins # secs	# mins # secs

Appendix II: Chuka University Ethics Committee Letter

CHUKA UNIVERSITY

knowledge is Wealth *(Sapientia divitia est)* Akili ni Mali

OFFICE OF THE DIRECTOR
BOARD OF POSTGRADUATE STUDIES

Telephones: 020-2310512/18
Chuka
Direct Line: 020-268 7625
www.chuka.ac.ke

postgraduate@chuka.ac.ke

P. O. Box 109-60400,
Website:

REF.: SM18/39923/19 **14th March, 2022**

Samwel Kisiang'ani Juma
Department of Physical Sciences

RE: REGISTRATION OF YOUR RESEARCH PROPOSAL

This is to acknowledge the receipt of your six loosely bound copies of proposal in the Graduate school.

I am pleased to inform you that your **Msc proposal** titled, **"Application of Queuing Theory for Optimal Customer Centricity to the Banking Sector in Kenya"** has been duly registered in the Graduate school of Chuka University. The following supervisor(s) have been appointed by BPGS on behalf of the Senate to work with you;

1. **Dr. Adolphus Wagala, PhD (Main Supervisor)**
2. **Dr. Gladys Njoroge**

You are now required to seek a Research Permit from **NACOSTI** and thereafter proceed with your Thesis. You will be required to keep in touch with your supervisor(s) and make sure you submit quarterly progress reports to the Graduate school, Chuka University. This is a **Mandatory** requirement and failure to do so shall lead to deregistration.

I wish you a successful and Fruitful Research.

Yours Sincerely,

for 2022

Prof. Moses Muraya
DIRECTOR
BOARD OF POSTGRADUATE STUDIES
Cc: DVC (ARSA)

Appendix III: National Commission for Science, Technology and Innovation (NACOSTI) Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **286066**

Date of Issue: **15/March/2022**

RESEARCH LICENSE

This is to Certify that Mr.. SAMWEL KISIANG'ANI JUMA of Chuka University, has been licensed to conduct research in Tharaka-Nithi on the topic: APPLICATION OF QUEUING THEORY FOR OPTIMAL CUSTOMER CENTRICITY TO THE BANKING SECTOR IN KENYA for the period ending : 15/March/2023.

License No: **NACOSTI/P/22/16250**

286066

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

Appendix IV: R-scripts

```
# queueing package
if(!require(queueing)){install.packages("queueing")}
library(queueing)
# Peak -----------------------------
#May 2019
N_M19<-3293
wh_M19<-192
lmd_M19<-N_M19/wh_M19

min_M19<-1
sec_M19<-52
AST_M19<- min_M19 +(sec_M19/60)
AST_M19
mu_M19<-60/AST_M19
mu_M19

input_mm1_M19 <- NewInput.MM1(lambda = lmd_M19, mu = mu_M19, n = 5)
# Create queue class object
output_mm1_M19 <- QueueingModel(input_mm1_M19)

# report
Report(output_mm1_M19)

# Get queue model summary
summary(output_mm1_M19)

par(mfrow=c(3,1))
3293/26
#Poisson Distribution Plot for Arrival Process
PDAP_1<-curve(dpois(x, input_mm1_M19$lambda),
     from = 0,
     to = 62,
     type = "b",
     lwd = 2,
     xlab = "Number of customers",
     ylab = "Probability",
     main = "Poisson Distribution for Arrival Process in May 2019",
     ylim = c(0, 0.15),
     n = 63)
#Exponential Distribution Plot for Interarrival Time

curve(dexp(x, rate = input_mm1_M19$mu),
     from = 0,
     to = 5,
     type = "l",
     lwd = 2,
     xlab = "Service Waiting Time",
     ylab = "Probaility",
     main = "Exponential Distribution for Service Process in May 2020",
     ylim = c(0, 1))
abline(h = 0)
#Exponential Distribution Plot for Service Process
curve(dexp(x, rate = 1/input_mm1_M19$lambda),
```

```
      from = 0,
      to = 10,
      type = "l",
      lwd = 2,
      xlab = "Interarrival Time",
      ylab = "Probaility",
      main = "Exponential Distribution for Interarrival Time",
      ylim = c(0, .2))
abline(h = 0)

#May 2020        ####
NM20<-1614
whM20<-192
lmdM20<-NM20/whM20

min_M20<-2
sec_M20<-34
AST_M20<- min_M20 +(sec_M20/60)
AST_M20
mu_M20<-60/AST_M20
mu_M20

input_mm1_M20 <- NewInput.MM1(lambda = lmdM20, mu = mu_M20, n = 5)
#  queue class object
output_mm1_M20 <- QueueingModel(input_mm1_M20)
# model report
Report(output_mm1_M20)

#  model summary
summary(output_mm1_M20)

1614/26
#Poisson Distribution Plot for Arrival Process
curve(dpois(x, input_mm1_M20$lambda),
      from = 0,
      to = 62,
      type = "b",
      lwd = 2,
      xlab = "Number of customers",
      ylab = "Probability",
      main = "Poisson Distribution for Arrival Process in May 2020",
      ylim = c(0, 0.15),
      n = 63)
#Exponential Distribution Plot for Interarrival Time

curve(dexp(x, rate = input_mm1_M20$mu),
      from = 0,
      to = 5,
      type = "l",
      lwd = 2,
      xlab = "Service Waiting Time",
      ylab = "Probaility",
      main = "Exponential Distribution for Service Process in May 2020",
      ylim = c(0, 1))
```

```
abline(h = 0)

#Exponential Distribution Plot for Service Process

curve(dexp(x, rate = 1/input_mm1_M20$lambda),
      from = 0,
      to = 10,
      type = "l",
      lwd = 2,
      xlab = "Interarrival Time",
      ylab = "Probaility",
      main = "Exponential Distribution for Interarrival Time in May 2020",
      ylim = c(0, .2))
abline(h = 0)

#May 2021        ####
N_m21<-2315
wh_m21<-192
lmd_m21<-N_m21/wh_m21

min_M21<-2
sec_M21<-27
AST_M21<- min_M21 +(sec_M21/60)
AST_M21
mu_M21<-60/AST_M21
mu_M21

input_mm1_M21 <- NewInput.MM1(lambda = lmd_m21, mu = mu_M21, n = 5)
# Create queue class object
output_mm1_M21 <- QueueingModel(input_mm1_M21)

# Get queue model report
Report(output_mm1_M21)

# Get queue model summary
summary(output_mm1_M21)

2315/26
#Poisson Distribution Plot for Arrival Process
curve(dpois(x, input_mm1_M21$lambda),
      from = 0,
      to = 62,
      type = "b",
      lwd = 2,
      xlab = "Number of customers",
      ylab = "Probability",
      main = "Poisson Distribution for Arrival Process in May 2021",
      ylim = c(0, 0.15),
      n = 63)
#Exponential Distribution Plot for Interarrival Time

curve(dexp(x, rate = input_mm1_M21$mu),
      from = 0,
      to = 5,
```

```
    type = "l",
    lwd = 2,
    xlab = "Service Waiting Time",
    ylab = "Probaility",
    main = "Exponential Distribution for Service Process in May 2021",
    ylim = c(0, 1))
abline(h = 0)

#Exponential Distribution Plot for Service Process

curve(dexp(x, rate = 1/input_mm1_M21$lambda),
    from = 0,
    to = 10,
    type = "l",
    lwd = 2,
    xlab = "Interarrival Time",
    ylab = "Probaility",
    main = "Exponential Distribution for Interarrival Time in May 2021",
    ylim = c(0, .2))
abline(h = 0)

#July 2019        ####
N_J19<-2018
wh_J19<-200
lmd_J19<-N_J19/wh_J19

min_J19<-3
sec_J19<-11
AST_J19<- min_J19 +(sec_J19/60)
AST_J19
mu_J19<-60/AST_J19
mu_J19


input_mm1_J19 <- NewInput.MM1(lambda = lmd_J19, mu = mu_J19, n = 5)
# Create queue class object
output_mm1_J19 <- QueueingModel(input_mm1_J19)

# Get queue model report
Report(output_mm1_J19)

# Get queue model summary
summary(output_mm1_J19)

2018/27
#Poisson Distribution Plot for Arrival Process
curve(dpois(x, input_mm1_J19$lambda),
    from = 0,
    to = 62,
    type = "b",
    lwd = 2,
    xlab = "Number of customers",
    ylab = "Probability",
    main = "Poisson Distribution for Arrival Process in July 2019",
```

```
    ylim = c(0, 0.15),
    n = 63)
#Exponential Distribution Plot for Interarrival Time

curve(dexp(x, rate = input_mm1_J19$mu),
    from = 0,
    to = 5,
    type = "l",
    lwd = 2,
    xlab = "Service Waiting Time",
    ylab = "Probaility",
    main = "Exponential Distribution for Service Process in July 2019",
    ylim = c(0, 1))
abline(h = 0)

#Exponential Distribution Plot for Service Process

curve(dexp(x, rate = 1/input_mm1_J19$lambda),
    from = 0,
    to = 10,
    type = "l",
    lwd = 2,
    xlab = "Interarrival Time",
    ylab = "Probaility",
    main = "Exponential Distribution for Interarrival Time in July 2019",
    ylim = c(0, .2))
abline(h = 0)

#July 2020        ####
N_J20<-1805
wh_J20<-200
lmd_J20<-N_J20/wh_J20

min_J20<-3
sec_J20<-04
AST_J20<- min_J20 +(sec_J20/60)
AST_J20

mu_J20<-60/AST_J20
mu_J20

input_mm1_J20 <- NewInput.MM1(lambda = lmd_J20, mu = mu_J20, n = 5)
# Create queue class object
output_mm1_J20 <- QueueingModel(input_mm1_J20)
# Get queue model report
Report(output_mm1_J20)

# Get queue model summary
summary(output_mm1_J20)

1805/27
#Poisson Distribution Plot for Arrival Process
curve(dpois(x, input_mm1_J20$lambda),
    from = 0,
```

```
    to = 62,
    type = "b",
    lwd = 2,
    xlab = "Number of customers",
    ylab = "Probability",
    main = "Poisson Distribution for Arrival Process in July 2020",
    ylim = c(0, 0.15),
    n = 63)
#Exponential Distribution Plot for Interarrival Time

curve(dexp(x, rate = input_mm1_J20$mu),
    from = 0,
    to = 5,
    type = "l",
    lwd = 2,
    xlab = "Service Waiting Time",
    ylab = "Probaility",
    main = "Exponential Distribution for Service Process in July 2020",
    ylim = c(0, 1))
abline(h = 0)

#Exponential Distribution Plot for Service Process
curve(dexp(x, rate = 1/input_mm1_J20$lambda),
    from = 0,
    to = 10,
    type = "l",
    lwd = 2,
    xlab = "Interarrival Time",
    ylab = "Probaility",
    main = "Exponential Distribution for Interarrival Time in July 2020",
    ylim = c(0, .2))
abline(h = 0)

#July 2021        ####
N_J21<-2150
wh_J21<-200
lmd_J21<-N_J21/wh_J21
min_J21<-2
sec_J21<-43
AST_J21<- min_J21 +(sec_J21/60)
AST_J21
mu_J21<-60/AST_J21
mu_J21

input_mm1_J21 <- NewInput.MM1(lambda = lmd_J21, mu = mu_J21, n = 5)
# Create queue class object
output_mm1_J21 <- QueueingModel(input_mm1_J21)
# Get queue model report
Report(output_mm1_J21)

# Get queue model summary
summary(output_mm1_J21)

2150/27
```

```
#Poisson Distribution Plot for Arrival Process
curve(dpois(x, input_mm1_J21$lambda),
      from = 0,
      to = 62,
      type = "b",
      lwd = 2,
      xlab = "Number of customers",
      ylab = "Probability",
      main = "Poisson Distribution for Arrival Process in July 2021",
      ylim = c(0, 0.15),
      n = 63)
#Exponential Distribution Plot for Interarrival Time
curve(dexp(x, rate = input_mm1_J21$mu),
      from = 0,
      to = 5,
      type = "l",
      lwd = 2,
      xlab = "Service Waiting Time",
      ylab = "Probaility",
      main = "Exponential Distribution for Service Process in July 2021",
      ylim = c(0, 1))
abline(h = 0)

#Exponential Distribution Plot for Service Process
curve(dexp(x, rate = 1/input_mm1_J21$lambda),
      from = 0,
      to = 10,
      type = "l",
      lwd = 2,
      xlab = "Interarrival Time",
      ylab = "Probaility",
      main = "Exponential Distribution for Interarrival Time in July 2021",
      ylim = c(0, .2))
abline(h = 0)
#Distribution function of w and wq in #####
par(mfrow= c(3,2))
gTitle <- "May 2019"
fw_M19 <- output_mm1_M19$FW
fwq_M19 <- output_mm1_M19$FWq
n <- 3
ty <- "l"
ylab <- "Probability"
xlab <- "t"
cols <- c("black", "red")
leg <- c("FW(t)", "FWq(t)")
curve(fw_M19, from=0, to=n, type=ty, ylab=ylab, xlab=xlab, col=cols[1], main=gTitle)
curve(fwq_M19, from=0, to=n, type=ty, col=cols[2], add=T)
legend("bottomright", leg, lty=c(1, 1), col=cols)

gTitle <- "July 2019"
fw_J19 <- output_mm1_J19$FW
fwq_J19 <- output_mm1_J19$FWq

curve(fw_J19, from=0, to=n, type=ty, ylab=ylab, xlab=xlab, col=cols[1], main=gTitle)
```

```
curve(fwq_J19, from=0, to=n, type=ty, col=cols[2], add=T)
legend("bottomright", leg, lty=c(1, 1), col=cols)

gTitle <- "May 2020"
fw_M20 <- output_mm1_M20$FW
fwq_M20 <- output_mm1_M20$FWq
curve(fw_M20, from=0, to=n, type=ty, ylab=ylab, xlab=xlab, col=cols[1], main=gTitle)
curve(fwq_M20, from=0, to=n, type=ty, col=cols[2], add=T)
legend("bottomright", leg, lty=c(1, 1), col=cols)

gTitle <- "July 2020"
fw_J20 <- output_mm1_J20$FW
fwq_J20 <- output_mm1_J20$FWq

curve(fw_J20, from=0, to=n, type=ty, ylab=ylab, xlab=xlab, col=cols[1], main=gTitle)
curve(fwq_J20, from=0, to=n, type=ty, col=cols[2], add=T)
legend("bottomright", leg, lty=c(1, 1), col=cols)

gTitle <- "May 2021"
fw_M21 <- output_mm1_M21$FW
fwq_M21 <- output_mm1_M21$FWq
curve(fw_M21, from=0, to=n, type=ty, ylab=ylab, xlab=xlab, col=cols[1], main=gTitle)
curve(fwq_M21, from=0, to=n, type=ty, col=cols[2], add=T)
legend("bottomright", leg, lty=c(1, 1), col=cols)

gTitle <- "July 2021"
fw_j21 <- output_mm1_J21$FW
fwq_J21 <- output_mm1_J21$FWq

curve(fw_j21, from=0, to=n, type=ty, ylab=ylab, xlab=xlab, col=cols[1], main=gTitle)
curve(fwq_J21, from=0, to=n, type=ty, col=cols[2], add=T)
legend("bottomright", leg, lty=c(1, 1), col=cols)
```