



UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

CUSTOMER BEHAVIOUR SEGMENTATION AMONG MOBILE SERVICE PROVIDERS  
USING ALGORITHMS (COMPARISON OF K-MEANS AND SELF-ORGANIZING MAPS)

BY

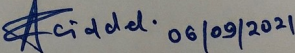
GIDDEL AGOI EMBALO

A PROJECT REPORT SUBMITTED TO UNIVERSITY OF NAIROBI IN PARTIAL  
FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF MASTER OF SCIENCE IN  
INFORMATION TECHNOLOGY MANAGEMENT DEGREE OF THE UNIVERSITY OF  
NAIROBI.

AUGUST 2021

## DECLARATION

This project report is my original work and to the best of my knowledge this research work has not been submitted for any other award in any University.


Sign:  06/09/2021

Date: 6<sup>th</sup> September 2021

Name: Giddel Agoi Embalo

Admin. No: P54/79498/2015

This project report has been submitted in partial fulfilment of the requirement of the Master of Science Degree in Information Technology Management of the University of Nairobi with my approval as the University supervisor.

Sign: 

Date: 9th September, 2021

FOR Name: Ms. Pauline Wambui

Position: Lecturer

School of Computing and Informatics

## **ACKNOWLEDGEMENTS**

My appreciation goes to the Almighty God for his grace throughout this academic journey and for granting me courage, good health, wisdom and inspiration that was essential to carry out this demanding study.

## **DEDICATION**

This paper is dedicated to my institution mentors under whose constant guidance I have completed this dissertation. They not only enlightened me with academic knowledge but also gave me valuable advice whenever I needed it the most.

## ACRONYMS AND ABBREVIATIONS

**ANN:** Artificial neural networks

**BI:** Business Intelligence

**BIC:** Bayesian Information Criterion.

**CDR:** Call Detail Records

**CRM:** Customer Relationship Management

**DT:** Decision Trees

**ETL:** Extraction, Transformation and Loading

**KDD:** Knowledge Discovery in Databases

**KVQ:** Kohonen vector quantization

**LR:** Logistic Regression

**OLAP:** On-Line Analytical Processing

**PCA:** Principal Component Analysis

**RFM:** Recency, frequency and monetary

**SOM:** Self-Organizing Map

**SPM:** Sequential pattern mining

**SVC:** Support Vector Clustering

**SVM:** Support Vector Machines

**UCAM:** Unique Clustering through Affinity Measure

**CPA:** Customer portfolio analysis.

## DEFINITION OF OPERATIONAL TERMS

**Data mining:** is the most common way of separating valuable information from a dataset and introducing it in an easy to use design for dynamic.

**Segmentation:** Is the division of an organization's client base into bunches known as client fragments, every one of which incorporates customers who meet comparative market models.

**Clustering:** can be characterized as the grouping of information. In bunching, class marks are obscure, and it is up to the grouping calculation to discover fitting classes.

**Association:** It is the study of the frequency of things happening together in transactional databases, and it determines frequent item sets based on a support threshold.

**Business intelligence:** is the most common way of breaking down an enormous volume of information and conveying a significant level arrangement of reports that concentrate the significance of that information into the establishment of business tasks, helping supervisors in settling on every day choices.

## LIST O TABLES

Table 2.5.1 shows the parameters used to compare the algorithms.....	34
----------------------------------------------------------------------	----

## LIST OF FIGURES

Figure 1: A self-organizing map. The input data $X$ is distributed to a set of models ( $M_i$ ). The best matching unit ( $M_c$ ) and the units in its neighborhood (large circle) are modified to better match input unit $X$ . (Kohonen 2013) .....	32
Figure 1. Conceptual Framework.....	37
Figure 2: Design Process.....	39



## ABSTRACT

Client division is a significant space of Business Insight where clients are totaled into bunches with comparative qualities like segment, geographic, or conduct attributes. Thus, every individual from the section has comparable requirements, wants, and attributes. Division can give a multidimensional perspective on the client, which can be utilized to educate a treatment technique. Not many examinations in the Kenyan business climate have utilized this way to deal with managing dimensionality; accordingly, it is suitable to utilize it. The hole that this examination expects to fill is trying the exhibition of these two calculations in taking care of enormous datasets in a Kenyan setting. This examination investigated client conduct division among versatile specialist co-ops utilizing K-intends to deal with multidimensionality and SOM to distinguish anomalies. The examination's fundamental objective was to give client conduct division from subliminally gathered versatile telecom utilization information utilizing K-implies and the SOM Calculation, and to think about the viability of the two strategies. The particular objectives are to think about the multidimensionality information taking care of capacities of the K-means and SOM calculations and give indisputable outcomes. Dealing with anomalies in huge informational collections utilizing the SOM calculation, just as leading a writing survey to think about the SOM and K-implies calculations' application on enormous informational collections in the Kenyan portable assistance industry. The aftereffects of the tests led in this investigation show that SOM beats the K-implies calculation. Much of the time, any business might want to play out a division that puts each client into some effectively portrayed bunch, which SOM in this investigation doesn't accomplish. K-implies, then again, gives an approach to see how bunches from huge informational indexes identify with the business. Comparable to client conduct, the presentation of the two calculations was estimated utilizing boundaries, for example, number of groups, map geography, mistake rate, precision, calculation time, intricacy, and execution time. It is presumed that SOM delivers better outcomes when managing anomalies in huge datasets, though K-implies is more qualified to multidimensionality in enormous datasets in light of the fact that it permits the utilization of different calculations inside it.

## CONTENTS

DECLARATION.....	ii
------------------	----

ACKNOWLEDGEMENTS.....	iii
DEDICATION.....	iv
ACRONYMS AND ABBREVIATIONS.....	v
DEFINITION OF OPERATIONAL TERMS.....	vi
LIST O TABLES.....	vii
LIST OF FIGURES.....	viii
ABSTRACT.....	ix
CHAPTER ONE.....	1
INTRODUCTION.....	1
1.0 Background of the study.....	1
1.1 Kenya Telecommunication Industry.....	4
1.2 Statement of the problem.....	5
1.3 General objectives.....	6
1.3.1 Specific objectives.....	6
1.4 Research questions.....	7
1.5 Significance of the study.....	7
1.6 Scope of the study.....	7
CHAPTER TWO.....	8
LITERATURE REVIEW.....	8
2.0 Introduction.....	8
2.1 Business intelligence and data mining.....	8
2.2 Customer segmentation and data mining techniques.....	10
2.3 Concept of Data Mining.....	13
2.3.2 Classification techniques for customer segmentation.....	14
2.3.3 Clustering techniques for customer segmentation.....	15
2.4 Empirical analysis of SOM and K-means algorithms.....	22
2.4.1 K-Means Algorithm.....	22
2.4.2 Self-Organization Maps Algorithm.....	22
2.5 How will SOM and K-means algorithms be compared?.....	23
2.6 Research gap in existing literature.....	23
2.7 Conceptual Framework.....	26
CHAPTER THREE.....	28
METHODOLOGY.....	28

3.0	Introduction.....	28
3.1	Research Philosophy.....	28
3.2	Research Approach.....	28
3.3	Research Design.....	28
3.2	Population.....	30
3.3	Data Collection.....	30
3.4	Creating a target dataset.....	30
3.5	Data cleaning and preprocessing.....	31
3.6	Data Presentation.....	32
3.7	Research Limitations.....	32
3.8	Ethical Review.....	32
3.9	Chapter Summary.....	32
	CHAPTER FOUR.....	33
	DATA ANALYSIS, EXPERIMENT RESULTS AND DISCUSSION.....	33
4.0	Introduction.....	33
4.1	Data Size and processing.....	33
4.1	Handle outliers in large data using SOM and K-means algorithm.....	35
4.1.1	Outliers Detection using the box plot.....	35
4.1.2	Outliers detection using distance between Neurons in SOM.....	39
4.2	Handle multidimensional data using Multiple Correspondence Analysis (MCA).....	41
4.3	Comparison of k-Means and SOM algorithms on the large data set.....	43
	CHAPTER FIVE.....	44
	SUMMARY AND CONCLUSION.....	44
5.0	Introduction.....	44
5.1	Summary.....	44
5.2	Conclusion.....	44
5.3	Recommendation and Future Work.....	45
	REFERENCE.....	46
	Appendix I:.....	50
	Handle multidimensional data using Multiple Correspondence Analysis (MCA).....	51



## CHAPTER ONE

### INTRODUCTION

#### 1.0 Background of the study

The broadcast communications industry was among quick to utilize information mining procedures. This is probably on the grounds that broadcast communications organizations as often as possible gather and store a lot of top notch information, have a huge client base, and work in a speedy, exceptionally aggressive climate. Broadcast communications organizations use information mining to further develop their promoting endeavors, recognize extortion, and deal with their organizations all the more adequately. Notwithstanding, because of the huge extent of their informational indexes, the consecutive and worldly parts of their information, and the necessity to anticipate extremely surprising occasions continuously—like client extortion and organization disappointments—these organizations face various information mining difficulties (Thakur and Workman2016). The reception of information mining in the telecoms business can be seen as a characteristic augmentation of the business' utilization of master frameworks.

Market segmentation has been a basic idea in marketing theory and practice over the years. The process of dividing a huge market into smaller groups or clusters of clients is known as segmentation. The commonalities within each segment can be used to determine the similarity of purchase behavior. Due to the restricted number of high-profit customers and strong market competition, there is an unprecedented demand for customer comprehension. Enterprises should be able to segment clients in order to suit their needs and improve their satisfaction. Varied clients have different preferences, values, profit margins, and so on. Market segmentation has been a basic idea in marketing theory and practice over the years. The process of dividing a huge market into smaller groups or clusters of clients is known as segmentation. The commonalities within each segment can be used to determine the similarity of purchase behavior. Due to the restricted number of high-profit customers and strong market competition, there is an unprecedented demand for customer comprehension. Enterprises should be able to segment clients in order to suit their needs and improve their satisfaction. Varied clients have different preferences, values, profit margins, and so on. According to customer segmentation theory, groups of customers with similar demands and purchasing habits are more likely to respond to marketing programs in a consistent manner (Buttle & Maklan 2015). Enterprises

may organize the right products, services, and resources for each target customer cluster and create a tight relationship with them via efficient market segmentation. As a result, market segmentation has come to be recognized as one of the most important aspects of modern marketing and customer relationship management.

The ascent in business rivalry, just as the accessibility of gigantic verifiable information vaults, has driven the boundless utilization of information mining methods to discover significant and vital data covered in organization data sets (Wei et al., 2013). Telecom organizations would now be able to monitor their customers because of mechanical progressions. Dissecting recorded information empowers organizations to discover existing buyers' standards of conduct, which could hugely affect gauging future client conduct. CPA (client portfolio examination) is a valuable technique for investigating client conduct. "CPA will likely gap customers into classifications (Thakur and Workman 2016)". Client division is the division of clients into comparable gatherings dependent on various qualities utilizing verifiable information (Hsu et al. 2012). The association would have the option to find the buyers who are deliberately significant and productive through the client division system. Clients can be isolated into two gatherings: "those with a high future lifetime esteem and those with a high volume (Buttle and Maklan 2015). Contenders squarely prepared to offer similar administrations and items at a lower cost and of unrivaled quality. Clients will essentially leave at lesser costs or better quality (Keramati et al. 2014). On account of the misfortune in deals, losing customers likewise brings about a promising circumstance cost (Verbeke et al. 2011). Past research has exhibited that holding ebb and flow significant buyers is generously more affordable than enlisting new ones. The way toward separating valuable data from a dataset and introducing it in an intelligible manner for choice help is known as information mining. Measurements, computerized reasoning, AI, and information base frameworks all converge with information mining draws near. Bioinformatics, climate determining, extortion discovery, monetary examination, and client division are only a couple of the uses of information mining.

Client division is the parted of an organization's client base into groupings called client portions, every one of which incorporate customers with practically identical market rules. This division depends on components that can impact the market or business straightforwardly or in a roundabout way, like item inclinations or assumptions, area, and conduct (Gupta, 2014). Client division is the parted of an organization's client base into groupings called client fragments, every one of which incorporate customers with equivalent market rules. This division depends on components that can

impact the market or business straightforwardly or in a roundabout way, like item inclinations or assumptions, area, and conduct (Gupta, 2014). The significance of client division incorporates, *inter alia*, the capacity of a business to tweak market programs that will be appropriate for every one of its client portions. Business choice help in hazardous circumstances like acknowledge connections for clients; recognizable proof of items related with each fragment and how to oversee request and supply powers; uncovering some inert conditions and relationship among clients, items, or clients and items that the business may not know about; capacity to foresee custom (Arora and Malik, 2015).

Client division, as indicated by Hoegele, Schmidt, and Torgler (2016), is the most common way of collection an enormous gathering of changed clients into bunches with equivalent provisions, conduct, or requests. Organizations figure out how to more readily contact their clients by distinguishing and profiling the main classes. Organizations that have a superior comprehension of client conduct can offer more customized types of assistance and items (Hamka et al., 2014). Accordingly, division has demonstrated to be an important information hotspot for promoting, deals, and vital arranging. Organizations have consistently sectioned their items and administrations dependent on statistical surveying and client input. Notwithstanding, with the coming of information mining devices, it is presently conceivable to fragment clients dependent on their real conduct.

Bunching has been exhibited to be powerful in distinguishing minor however strategic examples or connections stowed away inside a store of unlabeled datasets. This kind of learning is sorted as unaided learning. Bunching strategies like K-Means, k-Nearest Neighbor, Self-Organizing Map (SOM), and others are accessible (Floh et al., 2014). These calculations can perceive groups in a dataset without earlier information on the dataset by over and over again contrasting info designs until stable bunches in the preparation models are accomplished utilizing the bunching basis or measures. Each bunch contains information focuses that are very like each other while being limitlessly not the same as information focuses in different groups. Bunching is very helpful in design acknowledgment, picture examination, bioinformatics, and different fields (Arora and Malik, 2015). Data segmentation is the partitioning of data into groups that are comparable in some way. This strategy is used in the market to target groups of clients that have similar requests or preferences. This allows the manufacture of a product to be studied, as well as the like and disliking of the product by buyers. It is a standard internet strategy that has been validated and covered in every business guidebook. Operators understand that they can completely please every customer, both intellectually and practically. As a result, the idea will be to divide clients into teams, then focus marketing and

advertising efforts on the most appealing sector. In this example, charm denotes profitability as well as long-term viability (Hamka et al., 2014). The primary goal of segmentation was to distinguish between homogeneous and heterogeneous objects in the external market (the consumers). On the other hand, regardless of whether process is frequently utilized, the choice is almost never computerized or data driven. The outcome of segmentation is mostly determined by knowledge variables, which can be gleaned from market, psychographic, regional, and lifestyle data, among other sources. Clustering and subgroup finding are two kinds of customer segmentation. The goal was to break up the external market. It is quite rare for the final decision to be automatic or entirely based on data (Arora & Malik, 2015). Many essential decisions must be made, such as segment selection (which segment to choose), segment identification, and segment relative size. The input variable influences segmentation. Demographic, psychographic, regional, and lifestyle characteristics are subdivided further in the input variables.

### **1.1 Kenya Telecommunication Industry**

Expanded contest in the versatile market can spread to different spaces of the telecom area, driving administrators to depend on packaged and merged administrations to advance their market positions. Kenya's broadcast communications market is an energetic illustration of how expanded contest in the portable market can gush out over into all spaces of the telecom area, compelling administrators to go to packaged and merged administrations to advance their market positions. The Communication Authority of Kenya (CAK) (2016) published a 2016 statistics report on the mobile phone sector in Kenya for the period January to March 2016. Mobile money figures, internet penetration, and market share data for all telecommunication service providers are all included in the paper. It also provides information on courier and postal services. Kenya's mobile penetration increased by 3.5 million to 38.3 million subscriptions during the quarter, up from 37.7 million the previous quarter. As a result, during the quarter under review, mobile penetration increased by 1.5 percentage points to 89.2 percent, up from 87.7 percent the previous quarter (Keter, 2015).

Safaricom increased its subscriber market share by 0.9 percent to 65.6 percent, up from 64.7 percent in the previous quarter. Its overall subscriptions increased by 3.4 percent to 25.1 million from 24.4 million subscribers. Airtel, on the other hand, saw its market share drop by 1.7 percent to 17.5 percent. This was as a result of the regulator's new SIM card requirements. Over 500, 000 Airtel



subscribers were disconnected as a result of this. Its entire subscriber base was 6.7 million, down from 7.2 million in the previous quarter. Orange increased their market share by 0.1 percent to 12.5 percent, with 4.8 million users, up from 4.6 million. Equites, a subsidiary of Equity Bank, increased its market share to 4.4 percent, with 1.6 million members (Arasa & Githinji, 2014). During that time, Sema Mobile, a new operator, obtained an MVNO license and managed to acquire 158 users.

Current trends indicate that the Kenyan Communications Commission's decision to issue unified licenses in 2008 will result in the fixed market experiencing the rapid growth and development that was anticipated. Kenya's mobile market became significantly more competitive in 2008. The rate at which new technology is implemented, which expands market potential by bringing new services and providing important players with new capabilities while simultaneously lowering costs, is the primary driver of growth in the telecommunications and information technology sectors (Premkumar & Rajan, 2017). Other factors influencing competition and growth in the sector include global deregulation and privatization, as well as government efforts to change the monopoly position of national communication carriers. Cell phone usage in Kenya is increasing, as evidenced not only by an increase in the number of subscribers and providers, but also by the variety of services available. It's a growing industry with plenty of room for innovation and expansion, as well as a diverse range of services to meet the needs of every consumer (Ndambuki, Bowen & Karau, 2017).

## **1.2 Statement of the problem**

Customer segmentation is a crucial area of Business Intelligence that groups consumers together based on comparable demographic, geographic, or behavioral criteria. As a result, each segment member has comparable requirements, wants, and characteristics. The purpose of segmentation is to get to know your customers better and use that knowledge to boost profits, save operating costs, and improve customer service. For a better treatment plan, segmentation can provide a multi-dimensional view of the consumer (Frenkel et al., 2013).

Currently, operators create segmentation models by combining data from several attributes. This is supported by research conducted by (Kohonen, 2013), who examined Self-Organization Maps and discovered that they lack functionality for dealing with high dimensionality in complex data. It is also difficult to implement better segments using SOM as expected because information may be lost if a complex set of data is mapped onto an array with a resolution too small to display the data) SOM

however, shows more accuracy in classifying most to the objects to their clusters than other algorithms, it shows good results when using small datasets, therefore, it will make sense to compare it with k-means and use it as a yardstick in clustering of the data. On the other hand, according to Larose (2014), the SVC algorithm does not take into account dimensionality reduction when dealing with high dimensionality and fast accessibility of data during segmentation. It is clear that these models are weak and insufficient to deal with high dimensionality and fast accessibility of data during segmentation. There is a definite need for Business Intelligence (BI) to assist in acquiring, providing access to, and analyzing data or information about business organizations in order to enable better-informed decision-making (Wei et al., 2013). For Circular Invariant Clustering, “Ramirez-Ortiz et al., (2015) utilized a Modified K-Means Algorithm”. The examination found that when the quantity of bunches, k, develops bigger, the exhibition of the SOM calculation falls behind that of the K-mean methodology. The utilization of K-intends to recognize nonlinearly divisible groups was examined by Peker et al. (2017). It was shown that with regards to portioning informational indexes that are particular or very much isolated from each other, K-implies outflanks SOM. K-Means was analyzed by Hickman et al., (2013). Dimensionality decrease was accomplished utilizing head segment investigation and direct change, and the underlying centroid was determined prior to utilizing K-Means. “Just K-means could permit the utilization of different calculations like PCA and direct change something which is absent in other bunching strategies such as Self-arranging maps (SOM) (Rodriguez and Laio, 2014)”. K-implies in dimensionality decrease. It was tracked down that K-implies calculation “handles the high-dimensionality of information during the iterative advances”. Because of the set number of studies that have used this way to deal with manage dimensionality, it is more adequate to utilize it in the Kenyan setting, and the adaptability of K-implies is the hole that this investigation needs to fill. Monetary components (mean month to month costs) for each successive help will be used as contributions to K-implies for Segmentation to test its exhibition. Finally, the outcomes of clustering will be assessed, and the most profitable portions will be identified using a completely randomized design (CRD).

### **1.3 General objectives**

To segment customer behavior among mobile service providers and compare the capability of K-means and SOM algorithms in handling large data.

### **1.3.1 Specific objectives**

The following are specific objectives used: -

- 1) To handle outliers in large data using SOM and K-means algorithm
- 2) To compare multidimensionality data handling capability of K-means and SOM algorithms and provide conclusive findings

### **1.4 Research questions**

- 1) How can we handle outliers in large data using SOM algorithm and K-means?
- 2) What is the difference between K-means and SOM algorithms in handling data dimensionality?

### **1.5 Significance of the study**

This study is very crucial to all stakeholders especially the players in telecommunication industry where the competition is very intense. This study would provide players in the telecommunication industry with tested algorithms as tools they can use to comfortably handle their multidimensional data. They would be able to classify their customers based on their behavior hence would be able to concentrate on the most reliable and untapped markets. This study would further add to the body of literature concerning customer segmentation using clustering methodologies and techniques especially in the Kenyan telecommunication industry.

### **1.6 Scope of the study**

Geographically, this study will be carried out in Kenya and it looked at the telecommunication industry in Kenya. Furthermore, this project will purely focus Safaricom, Airtel and telecom current and past subscribers. The study will concentrate on the Network Quality, Customer Care, Cost or affordability as the information was sought on these issues to help in establishing customer segments.

## **CHAPTER TWO**

### **LITERATURE REVIEW**

#### **2.0 Introduction**

This section begins with a brief overview of data mining methods, then moves on to discuss various models, data mining functionalities, the relationship between data mining and business intelligence, before narrowing it down to various data clustering methods and providing pros and cons, and finally, a summary.

#### **2.1 Business intelligence and data mining**

Business Intelligence (BI) is the way toward assessing a tremendous measure of information and delivering an undeniable level arrangement of reports that center the information's significance into the center of business tasks, helping directors in settling on every day choices (Maheshwari, 2014). ETL, information warehousing, data set question and revealing, multidimensional/on-line scientific preparing (OLAP) information investigation, and information mining are all possible with BI software.

Executives, for example, cannot make choices purely on the basis of financial records comparing last month's outcomes to a budget established a year ago; they must also take into account current facts. Do they need knowledge that allows them to answer swiftly to simple questions like what was the sale last year? What is still available this year? What is some money-saving ideas? What expenses can be eliminated without inflicting long-term harm? Business intelligence solutions assist you in answering critical questions by transforming huge amounts of data from functional frameworks into a more noticeable and open organization (Braha, 2013). BI programming permits a firm, or even certain offices inside it, to follow current and long-haul patterns while likewise conveying consistent criticism on the adequacy of their choices.

In any industry, the accessibility of business issues that can be effectively tended to and taken care of with assistance of Business Intelligence complemented by Data Mining advancements, just as the accessibility of information for such innovations' sending, are two basic variables. These two main prerequisites for the Telecommunications business are completely met in the majority of the evaluated literature sources. The telecommunications industry has access to large amounts of “high-

quality data”, which is a critical success factor for business intelligence and data hungry applications, and it faces numerous business problems that require immediate attention through the application of innovative, powerful methods and tools.

Telecommunications data is divided into three categories and is produced by various operating systems. The first kind is customer contractual data, which includes personal information such as a customer's name and address, administration plan, contact data, FICO assessment, family pay, and installment history. Second, call detail information - itemized consider records that contain the starting and ending telephone numbers, the call date and time, and the length of customer exchanges, just as charging information from which information about customers' calling behavior may be retrieved at the customer level; Third, there's network data, which is derived by the operation of extremely complex systems. The three categories of telecoms data that have been identified are as follows: (Witten and coworkers, 2016) (Floh et al., 2014) investigate data multidimensionality, which is one of the most essential characteristics in many BI and Data Mining applications. The above-mentioned abundance of telecommunications data is a major contributor to the industry's recent spike in interest in business intelligence and data mining. The table below summarizes BI and Data Mining applications: -

**Table 2.0: BI and Data Mining applications**

BI and Data Mining Application Areas	Business Problems Addressed
Marketing, Sales, and CRM	<ul style="list-style-type: none"> <li>• Customer profiles are created using call detail information, and these profiles are mined for marketing purposes.</li> <li>• Identifying and maintaining profitable clients via measuring customer value</li> <li>• Getting the most profit out of each consumer</li> <li>• To help media transmission administrations, analysts are finding affiliation and grouping designs to acquiring new customers</li> <li>• Analysis of Segmentation</li> <li>• Predicting a client's conduct when it is probably going to happen is known as division expectation.</li> <li>• Understanding why certain customers are portioned and making endeavors to keep them is division the executives.</li> </ul>

## Fraud Detection

- Detection of potentially fraudulent users and their out-of-the-ordinary usage patterns (subscription fraud)
- Detecting attempts to gain unauthorized access to customer accounts (supposed fraud)
- Identifying unusual patterns that may require special attention, such as busy hours, frustrated call attempts, switch and route congestion patterns, and so on.

## Network Management

- Identification of flaws in the network; Relationship between alarms (for relating multiple alarms to a single fault)
- Fault prediction in a network
- Identifying and contrasting various types of data traffic
- The system's workload management
- Resource utilization management
- User groups' activities

---

## 2.2 Customer segmentation and data mining methods

Client division is the way toward arranging a customer base into various groups. Client division, to put it another way, is the way toward isolating clients into homogeneous gatherings dependent on shared or normal provisions. The objective of division is to study your purchasers and use that data to build income, save expenses, and further develop client assistance. Division can give a multi-layered picture of the buyer for a superior treatment approach (Frenkel et al., 2013).

To segment customers, data must be acquired, sorted, and analyzed. In order to increase an organization's revenue, it is possible to assess a customer's reliability/loyalty through appropriate data segmentation. Segmentation is a technique for forming appropriate customer groups based on individual explanation qualities and behavior. There are two primary strategies for segmentation, according to Kwach, Flora, and Rajagopal (2014): The main kind of division is requirements based division, which is the act of separating clients into bunches dependent on their necessities. The

method involved with fragmenting customers dependent on their qualities, perspectives, or practices is alluded to as attributes based division.

Segmentation models based on volume of sales, according to Aggarwal and Reddy (2013), should be used, implying that marketing efforts should be focused on clients who make a large number of transactions. As per the "profound half theory," as this technique is known, one-half of customers can represent up to 80% of complete pay. The amplexness of multivariate systems for anticipating deal tendency was raised doubt about during the 1970s, inciting the advancement of better speculative models of purchaser conduct. "Lee, Verma, and Roth (2015) made a generalizable psychographic division model that fragmented the market dependent on friendly class, way of life, and character characteristics".

However, the functional execution difficulties of this convoluted division approach have been widely acknowledged. The division cycle diagrams the qualities of the buyer gatherings (called portions or groups) inside the information, and it is additionally seen as an approach to have more designated commitment with clients. Because of the variety of customer wants and purchasing conduct, which is impacted by way of life, pay levels, and age, past division procedures have demonstrated ineffective. Thus, numerous cutting edges promoting division techniques depend on derived customer conduct from exchange records or reviews. The information is explored utilizing information mining procedures, for example, bunch investigation. They researched how information mining might be utilized to portion data. Client division is routinely referred to in research as the best methodology for expanding customer productivity by exactly focusing on clients. Lee, Verma, and Roth (2015) found market fragments for another PC framework utilizing bunch examination and information from a client overview. To make division simpler, item ascribes were given higher importance.

In their book "Data Mining: A hands-on approach for business professionals Wang and Fan (2014) summarize and explain the current state of data mining, as well as discuss some prominent tools on the market that could be useful to anyone considering data mining". He looked into Knowledge Seeker, a data mining software that employs the decision tree technique, and discovered that data could be structured in the most efficient manner possible, which could be extremely useful in market segmentation research.

Wang and Fan (2014) summarize and explain the current state of data mining in their book "Data Mining: A hands-on approach for business professionals, as well as discuss some prominent tools on the market that could be useful to anyone considering data mining". He looked into Knowledge

Seeker, a data mining software that employs the decision tree technique, and discovered that data could be structured in the most efficient manner possible, which could be extremely useful in market segmentation research. Another industry where data mining is commonly used, according to the author, is direct mail and mailing. Direct marketing is used by a considerable number of merchants, and its primary goal is to gain insight into client segmentation, which is a clustering problem in data mining. Traditional survey-based market research has several disadvantages, one of which is that it generates a large amount of data on a small number of clients. However, understanding the characteristics of all clients is usually required in order to properly implement market research findings. To put it another way, market research can lead to the identification of interesting customer segments (Guha & Mishra, 2016). They must then “be projected onto the existing client base” using available data. Behavioral data, which is frequently compiled from transaction and billing histories, can be especially useful in this regard. Customers must be identified as part of the market research in order to understand the behavior of the participants. In today's competitive marketplaces, this method is insufficient and inefficient. In order to gain competitive advantage, organizations, on the other hand, require a comprehensive view of their clients. They must then be projected onto the existing client base using available data. In this regard, behavioral data, which is frequently compiled from transaction and billing histories, can be especially useful. Customers must be identified as part of the market research in order to understand the participants' behavior. This method is insufficient and inefficient in today's competitive marketplaces. Organizations, on the other hand, require a comprehensive view of their clients in order to gain a competitive advantage. They must then “be projected onto the existing client base” using data that is readily available. In this regard, behavioral data, which is frequently compiled from transaction and billing histories, can be especially useful. Customers must be identified as part of the market research in order to understand the participants' behavior. This method is insufficient and inefficient in today's competitive marketplaces. Organizations, on the other hand, require a comprehensive view of their clients in order to gain a competitive advantage. In contrast, in segmentation, a group of people with similar needs, characteristics, or behaviors is typically grouped together.

Then, using available data, they must be “projected onto the existing client base. In this regard, behavioral data”, which is frequently compiled from transaction and billing histories, can be especially useful. Customers must be identified as part of the market research in order to understand the participants' behavior. This method is insufficient and inefficient in today's competitive



marketplaces. Organizations, on the other hand, require a comprehensive view of their clients in order to gain a competitive advantage. There are several types of segmentation based on the segmentation criteria used. Customers can be classified based on their value, socio-demographic and life-stage information, as well as behavioral, need/attitude, and loyalty factors. Your business goal and target audience will determine the type of segmentation you use. Different criteria and segmentation approaches are appropriate for different contexts and commercial goals (Witten et al., 2016).

Customers are divided into groups based on their behavior and consumption habits in behavioral segmentation. Despite the fact that business rules can be used to create behavior fragments, there are some drawbacks to this methodology. It can only handle a few division fields adequately, and its objectivity is called into question because it is based on the genuine beliefs of a business master. Information mining, on the other hand, can be used to create information-driven conduct portions. Bunching calculations can investigate client behavior, recognize regular groupings, and provide an answer based on observed data designs. Information mining methods can recognize groups with unmistakable profiles and characteristics, resulting in rich division plans with business value.

As a rule, when joining an enormous number of division factors, utilizing a group model to uncover the sections is vital. Rather than business manages, a bunch model can deal with an enormous number of traits and uncover information driven components that aren't known early (Guha and Mishra, 2016). Utilizing a bunch model to uncover the portions is basic when incorporating countless division boundaries. A group model, rather than business rules, can deal with a high number of qualities and uncover information driven segments that aren't known early (Guha and Mishra, 2016). Information mining can likewise be utilized to build division plans dependent on the current or anticipated worth of purchasers. These classifications are vital for focusing on client support and showcasing activities dependent on the significance of every client. We will likewise give techniques to profiling the bunches and featuring their separating attributes, as appreciating the importance of the resultant groups is a vital piece of the division interaction. One methodology to discover social sections is to utilize the bunching calculations depicted in this procedure produces gatherings of comparative customers, however it tends to be hard to sort out how these gatherings identify with the business. Normally, "an organization would like to do a division that arranges the entirety of its clients into one of a few classifications (Guha and Mishra, 2016)". These sections are habitually made considering a showcasing evenhanded, like membership recharging or high spending levels. For this kind of division, choice tree methods, for example, those depicted in are great.

### **2.3 Concept of Data Mining**

Information mining is the way toward removing substantial, significant, already obscure, lastly coherent information from enormous data sets. Information mining is viewed as a stage in the general interaction of information disclosure. Information mining strategies can help any business application that utilizes information, like expanding specialty unit and generally productivity, “understanding client wants and needs, finding beneficial customers, and acquiring new ones (Braha, 2013)”.

As per “Zheng (2015), information mining is the way toward removing information from monstrous measures of unstructured information”. Because of the broad accessibility of huge measures of information and the need to change such information into important data and information, it impressively affects the data business, remarkably in the media communications field, as of late. Market examination is only one illustration of how the information accumulated can be put to utilize. Information mining is a strategy for gaining from information that comprises of a bunch of decides and conditions that might be utilized to find intriguing information designs, investigate conduct, and foresee it. It tends to be arranged into two classifications dependent on their motivations: regulated/prescient learning and unaided learning. Administered and solo learning are two very unique learning frameworks, as their names propose. Solo learning doesn't need outer oversight, however managed learning does. Arrangement calculations are utilized in managed figuring out how to figure out which of a bunch of determined classes another article has a place with (Braha, 2013). In unaided picking up, bunching calculations are utilized to attempt to bunch a bunch of items and check whether there is a connection between them. Information mining is to uncover substantial, inventive, possibly helpful, and reasonable examples and associations in existing information. Discovering helpful examples in information is alluded to by an assortment of words, including information extraction, data disclosure, data gathering, information paleo history, and information design handling.

The expression "information mining" is utilized by analysts, data set specialists, and money managers. The contraction KDD (Knowledge Discovery in Databases) alludes to the way toward removing valuable data from information as a general rule, with information mining as a subset. The KDD cycle incorporates steps like information readiness, information determination, information cleaning, and fitting understanding of information mining results to guarantee that significant data is removed from the information. Information mining is a subset of customary information examination

and factual methodologies that consolidates insightful devices from an assortment of fields, including computerized reasoning, AI, OLAP, and information perception (Shmueli and Lichtendahl, 2017).

### **2.3.2 Classification techniques for customer segmentation**

Grouping examination is otherwise called managed arrangement. Characterization examination is the way toward recognizing a model (or capacity) that clarifies and isolates information classes or thoughts to utilize the model to anticipate the class of items whose class name is obscure. Order can be utilized to characterize information into a particular class. Class names are utilized to coordinate the items in the information assortment in a sensible way. The order model is quite possibly the most broadly utilized regulated demonstrating procedures. To characterize information, a client should break it into portions and afterward develop one of a kind non-covering gathering. To isolate material into classes, a client should know certain data about the thing to be fragmented. The reason for characterization issues is to distinguish the attributes that recognize which bunch every model has a place in. This example can be utilized to both comprehend existing information and foresee how new cases will act. By inductively discovering an expectation design in recently grouped information (cases), information mining creates order models (Frenkel et al., 2013).

In most grouping frameworks, all things in a preparation set are as of now connected with set up class names. Following that, the order calculation makes a model dependent on the preparation information. The model is utilized to order new things. To put it another way, order is a two-venture measure in which preparing information is utilized to develop an arrangement model, which is hence used to classify new information. Characterization errands have been embraced in the CRM area for an assortment of reasons. Ocumpaugh et al., (2014) utilized a choice tree to distinguish purchasers and foster procedures dependent on customer lifetime esteem. To decide the slant of the client lifecycle, Shmueli and Lichtendahl (2017) utilized a Bayesian organization classifier. The creator demonstrated Bayesian organization classifiers as a “valuable apparatus in the toolset of CRM investigators in the use of ascertaining the incline of the client lifecycle of long-haul customers. Sheu et al. (2009)” explored the likely connection between key impacting components and customer dependability utilizing a choice tree. The results of this examination motivate us to dissect the relationship between customer buy amounts and segment and conduct factors utilizing a choice tree, with an accentuation on the qualities of high-and low-spending customers.

### **2.3.3 Clustering techniques for customer segmentation**

Clustering is a technique for grouping similar objects in a collection of physical or abstract objects. Clustering is also known as unsupervised classification because the categorization is not dictated by the sequence of input class labels. There are a variety of clustering approaches, “all of which are based on maximizing intra-class similarity while minimizing similarity between items of different classes (inter-class similarity)”. “Grouping is similar to arrangement in that classes are not predefined, and the bunching technique is in charge of locating suitable characterizations (Tsipitsis and Chorianopoulos, 2011)”. Changing the gathering by removing factors that were used to organize events is frequently required because the client recognizes them as superfluous or irrational upon closer examination. Bunches that adequately section the data set are discovered, and these groups are then used to characterize new information. Kohonen include guides and K-implies are two commonly used bunching techniques. Bunching and division are not synonymous terms. Bunching is a procedure for segmenting information into previously unidentified groups, whereas division is the more general issue of recognizing bunches with similar elements. Bunching is a technique for identifying common information clusters known as groups. A bunch is a collection of information that is linked in some way. Bunching, also known as unaided learning, can be used to group customers who have similar practices and make business decisions (Berkhin, 2006). Solo learning is an arrangement method with an ambiguous goal, which implies that the class assigned to each case is ambiguous. The goal is to categorize the situations into distinct classes that are input homogeneous. In grouping research, there are no reliant factors. Bunching is perhaps the most helpful exercises in the information digging measure for finding gatherings and recognizing intriguing appropriations and examples with regards to the hidden information. The bunching issue requires gathering information into groups with the goal that information focuses in a similar bunch are more comparative than information focuses in different groups. Berkhin (2006) utilizes a grouping method to fragment customers and commercial centers. The K-implies grouping method and the Kohonen self-putting together guide are the two most famous bunching calculations.

Samira et al. (2007) utilized a proposed distance capacity to section customers of Iran's Exchange Advancement Association, which utilizes affiliation rules standards to decide dissimilarities between send out containers of various nations. The RFM model is then applied to each bunch to decide the best procedure for advancing each section. "The worth of the gathering items," "the kind of gathering

wares," and "the connection between fare bunch wares" are a portion of the factors utilized as division standards.

Pramod et al. carefully describe how to section shopper profiles in a retail firm utilizing bunching (2011). As per the discoveries, K-Means grouping empowers retailers to acquire a superior comprehension of their clients and settle on information driven choices to offer customized and productive assistance. Huang et al. (2009) explored client an incentive for a chasing store in Taiwan utilizing the K-implies strategy, Fluffy C-implies grouping technique, and sacked bunching calculation, and found that the packed away bunching calculation outflanked the other two methodologies.

Hickman and partners examined K-Means (2013). This investigation utilizes "head part examination and straight change" to lessen dimensionality and characterize the underlying centroid, which is then applied to the K-Means bunching method. Just K-implies permits you to utilize methods like PCA and straight change that other grouping draws near, for example, Self-coordinating guides, don't permit (SOM). Since this technique has been utilized in less investigations, it is more adequate in Kenya, and the adaptability of K-implies is a need that this examination will address. Manzano et al. (2015) employed developing constraints in the K-means approach and their elimination in data mining. According to Manzano et al., K-means is widely used because it is "conceptually simple, computationally fast, and memory efficient; however, it has a number of limitations that make extraction difficult when compared to other clustering and data mining approaches (2015)".

Ahn et al. (2011) give a customer grouping approach that might be utilized to work with strategically pitching in the versatile telecom organization using different techniques, in view of this blend of strategies. To begin, grouping methods including fake neural organizations (ANN), strategic relapse (LR), and choice trees (DT) are used individually to gauge future item arranges. This is more confounded and tedious, and it likewise requires broad information mining aptitude. Second, each model creates an outcome, which is then joined with the consequences, all things considered, and K-implies grouping of clients to survey whether a client would purchase another item. A versatile telecom administrator in Korea is scrutinizing the system. Subsequently, the model accomplished remarkable strategically pitching results. With regards to customer division, K-implies is probably the best datum mining approaches in view of its straightforwardness and speed.

Ramirez-Ortiz et al. (2015) developed a Modified K-Means Algorithm for Circular Invariant Clustering. They discovered that vector extraction and vector bunching are used in a variety of

fundamental example recognition applications. The Enhanced Moving K-Means (EMKM) Algorithm is used to segment images. In a Modified Version of the K-Means Algorithm, a Cluster Symmetry-Based Distance is calculated. According to the scientists, the separating method is a simple and effective implementation of Lloyd's k means bunching calculation. Because it is based on a kd-tree as its primary information structure, this strategy is simple to implement. The investigation also discovered that as the number of bunches,  $k$ , increases, the display of the SOM calculation lags behind that of the K-mean method.

In this study, experts propose a modified version of the K-infers strategy to pack data because it reduces the adequacy of the data mining measure.

Bose and Chen (2010) employ a between group examination strategy to look for financial potential in client data from various administrations. The creators employ a client grouping procedure to determine which clients are most likely to accept versatile administrations. Bundling is frequently done in terms of administration quality, revenue, utilization, and client groupings. In this work, information mining methods such as K-means and Kohonen vector quantization (KVQ) were used to sort customers based on their characteristics. The delivered bunches are then subjected to between bunch examination, which determines how clients are distributed across various groups of characteristics. In Hong Kong, a cell phone administrator provided information on customer exchanges. The data set contains 50,000 records spanning a one-year period beginning September 1, 2004 and ending September 1, 2005. The information was classified into four categories. The results show that the K-means and EM calculations outperformed the progressive grouping calculation. A few studies, particularly in Kenya, have used K-implies, so based on the writing survey alone, this will be an excellent opportunity to employ the K-implies calculations. As a result of this research, it was determined that K-implies is far superior to KVQ and other leveled bunching strategies.

According to “Mehta et al., (2017), K-means and SOM have direct unpredictability, with  $O(ndk)$  and  $O(nd)$ , separately, where  $n$  is the number of data tests,  $d$  is the number of estimations, and  $k$  is the number of gatherings”. As a result, the two systems are adaptable and work well with large illuminating files. The third strategy, SVC, has a computational complexity of where  $n$  is the number of help vectors, not data tests. Finding the Lagrange multiplier, which is used to find the bundle line regards, is a critical stepping stone for SVC with large informational collections. For example, the Sequential Minimal Optimization (SMO) computation has been proposed as a solution to this

problem. As a result, each of the three approaches is adaptable for large data sets, and execution should be unaffected.

“Rodriguez and Laio (2014) made a viable K-Means Clustering calculation for sorting protein groupings and breaking down quality articulation information, among other bioinformatics assignments”. In spite of their far-reaching use, a few of these calculations have the accompanying defects: (i) depend on foreordained boundary tweaking, for example, deduced information on the quantity of groups; (ii) utilizing nondeterministic techniques that produce conflicting outcomes. Subsequently, a structure that tended to these imperfections was attractive. An information driven system was given, which had two stages that were interconnected. The originally was SVD-based measurement decrease, while the second was a computerized framework-based boundary change (s). The measurement decrease measure is all around tuned for truly huge datasets. The Bayesian Information Criterion, an interior assessment measure, was utilized to track down the best boundary setting (BIC). The presentation of most grouping approaches can be improved by joining them into this system. When utilizing numerous properties for division, the information can be hard to see and appreciate. A few bunching models were set up for conduct division, including normal income per client, visit check per each event, and download volume per distinctive event. (Rodriguez and Laio, 2014) (Rodriguez. “This high-dimensionality influences the exhibition of the calculations and nature of the groups, so the less the measurements are, the quicker the execution and more minimized appropriation of information tests are, K-implies calculation handles the high-dimensionality of information during the iterative advances”.

As per “Peker et al., bit implies is an augmentation of the standard means bunching strategy that recognizes nonlinearly distinct groups (2017)”. The worldwide piece implies technique proposes a deterministic and reformist answer for bit based grouping. At each stage, a worldwide inquiry methodology comprising of numerous part - implies runs from proper instatements added one bunch. Because of its gradual nature and search strategy, this methodology didn't need group instatement, identified nonlinearly detachable bunches, and situated close ideal arrangements while staying away from helpless nearby minima. Besides, two changes were made to decrease the computational expense without compromising the nature of the result. The procedures given have been improved to deal with weighted information focuses, making them reasonable for chart parting. The proposed strategy beat K-implies with irregular restarts in this preliminary with a scope of informational

indexes. It was shown that K-implies outflanks SOM with regards to dividing informational collections that are particular or all around confined from each other.

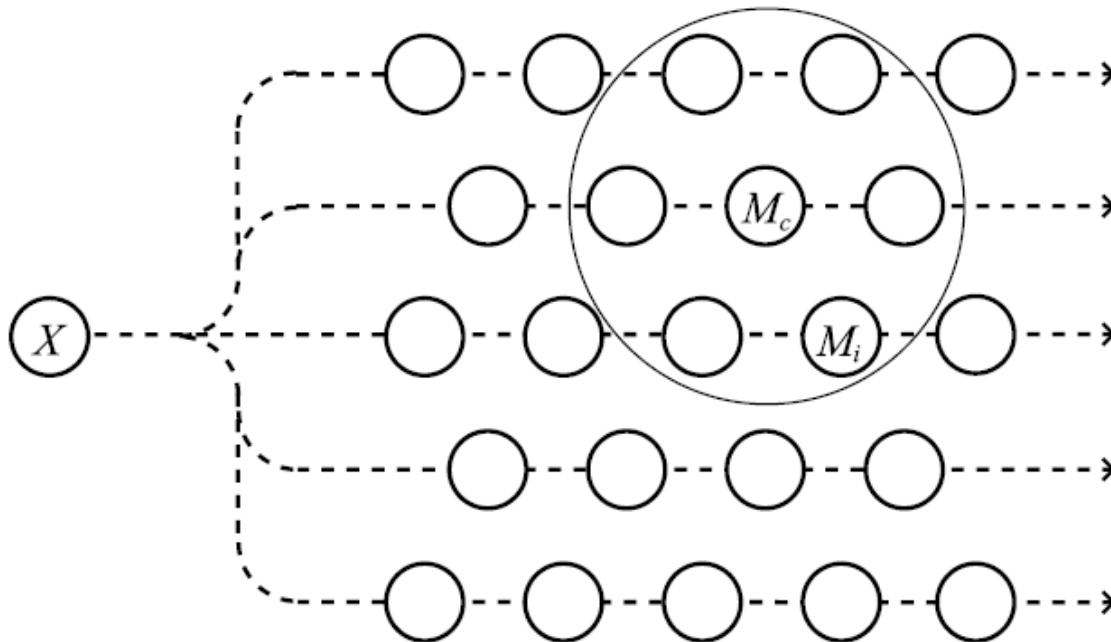
“Crespo and Weber (2015) investigated the K-Means bunching procedure and found that the underlying seeds”, which were picked consecutively or arbitrarily, affected the nature of the last group. It was hard to anticipate the quantity of bunches and first seeds for a huge continuous information base. To get around this issue, a few arrangements have been proposed. They were known as “Unique Clustering through Affinity Measure (UCAM)” and worked without characterizing introductory seeds, the quantity of resultant groups to frame, or remarkable bunching. Not at all like SOM, K-implies has no exception treatment highlights; all things being equal, anomalies influence bunching in light of the fact that they are dealt with similarly with any remaining information. This will naturally bring about lower bunching results relying upon the quantity of exceptions in the informational collection. Since neither K-implies nor the grouping approach represent missing information, the solitary choice is to erase them physically before bunching. SVC, then again, utilizes a delicate edge consistent (C) to deal with exceptions and missing qualities consequently, permitting these information focuses to stay outside the encasing circle. Accordingly, the calculation's capacity to manage exceptions and missing information will be tried.

Hosseini et al. (2010) utilized the K-implies calculation to characterize client dependability dependent on RFM esteems. “Cheng and Chen (2009) utilized K-means and harsh set hypothesis to isolate client esteem dependent on RFM esteems”. Chen et al. (2009) recognized buying patterns utilizing successive examples. Migueis et al. (2012) proposed a strategy for fragmenting clients relying upon the kinds of items they buy. This strategy depends on bunching calculations, what partition customers into bunches dependent on their ways of life. To reinforce client devotion and increment deals, the creator ordered customers of an European retailing organization into way of life divisions and planned special plans customized to each segment. The creator utilized the VARCLUS procedure, which is remembered for SAS programming, to bunch the items. The methodology involves construing the way of life related with each bunch of items by analyzing the sort of items included in each group. K-implies end up being quick, strong, and genuinely productive in all circumstances:  $O(knd)$ , where  $n$  is the quantity of items,  $k$  is the quantity of bunches,  $d$  is the quantity of article measurements, and  $t$  is the quantity of emphases Normally,  $k, t, d \ll n$ .



The self-organizing map (SOM) is an unaided neural organization-based methodology for information investigation, perception, and grouping. The SOM is a two-dimensional yield map that converts high-dimensional data to two-dimensional yield data. The yield map saves the first information's topological connections (Kiang, Hu et al. 2006). Because SOM, unlike many other bunching calculations, can group and extend information at the same time, its group geographies are reasonable without the need for post-processing (Yao, Sarlin et al. 2014).

The SOM is a two-layer bogus neural organization that consists of information and yield layers. The SOM models are linked to the yield layer's hubs, which are typically a two-dimensional cluster of hubs. As shown in Figure 1, each info information perception is communicated to all SOM models and coordinated with the model that best matches the info thing. The exhibit's triumphant model (the best coordinating with model) and its spatial neighbors are then changed for worked on coordinating. As a result of this cycle, models that are similar to one another on the SOM exhibit will be associated with hubs that are close to one another, whereas models that are less comparable will be associated with hubs that are further apart on the cluster (Kohonen 2013).



**Figure 4: An example of a self-organizing map. The input data  $X$  is distributed to a group of models ( $M_i$ ). The best matching unit ( $M_c$ ) and its neighbors (large circle) are modified to better match input unit  $X$ . Kohonen (2013)**

The yield layer of the SOM is typically addressed by a two-dimensional exhibit of hubs. The number of hubs can be predetermined so that all data is planned to a variety of the SOM investigator's choice. The size of the cluster should not be determined solely by the size and complexity of the information provided. Because predicting the exact size of the cluster early on is difficult, an experimentation technique should be used to determine the final exhibit size (Kohonen 2013). Every exhibit hub can be thought of as its own group (Yao, Sarlin et al. 2014). If the informational index contains a couple of groups, it is simple to depict these groups with a couple of hubs. However, data may be lost if a complex arrangement of information is planned onto an exhibit with a goal that is insufficient to address the information information's small constructions (Kohonen 2013). However, if the SOM is used with a large number of hubs, the SOM's preparation begins to deviate from bunch examination, making it difficult to distinguish the groups in the data (Yao, Sarlin et al. 2014). A second degree of grouping is frequently used to group the hubs of a larger SOM cluster (Sarlin, Yao 2013). When compared to results obtained by physically bunching the information, using a second level of bunching has a negative impact on grouping exactness. The SOM has a number of advantages that make it an ideal tool for data mining and segmentation research. When conducting customer

segmentation research, Bloom (2004) identified three significant advantages of SOM over cluster analysis. The self-organizing map outperforms standard clustering techniques in terms of robustness; missing values in the sample data have no effect on the SOM's performance. Missing data elements, on the other hand, would have a greater impact on cluster analysis, limiting its effectiveness.

No prior knowledge of the underlying distribution of the sample data is required when performing a SOM analysis. The analyst must make several assumptions about the data before performing cluster analysis. Cluster analysis requires the analyst to know ahead of time how many clusters exist. There are no such requirements for the self-organizing map.

Apart from the Bloom benefits, one of the most notable benefits of using the SOM as a data mining approach is its ability to display complex multidimensional data in a simple and easy-to-understand manner. The usability of the SOM as a quick data exploration method is also important. Because it is trained using unsupervised learning and organizes itself based on observed patterns, the SOM is an excellent tool for data analysis. It can be used to locate and analyze key data points in a dataset quickly.

## **2.4 Empirical analysis of SOM and K-means algorithms**

### **2.4.1 K-Means Algorithm**

K-means partitioning is a well-known partitioning method. Objects are categorized into one of three K groups. The centroid for each group (the multidimensional version of the mean) is calculated and each object is assigned to the group with the closest centroid. By relocating cluster members iteratively, this method reduces total within-cluster dispersion.

### **2.4.2 Self-Organization Maps Algorithm**

SOM, which is based on neural networks in the brain, achieves unsupervised learning through a competition and cooperation mechanism. In the basic SOM, a set of nodes is arranged in a geometric design, typically a 2-dimensional lattice. Each node has a weight factor the same size as the input space. The goal of SOM is to find a good mapping from the high-dimensional input space to the two-dimensional representation of the nodes. When using SOM for clustering, consider the objects in the input space that are represented in the same node as clustered into a cluster. When using SOM for clustering, consider the objects in the input space that are represented in the same node as clustered

into a cluster. Each object in the input is represented to the map during training, and the best matching node is identified.

## 2.5 How will SOM and K-means algorithms be compared?

Specific parameters, as indicated in table 2.5.1 below, will be used to compare the two algorithms. For each algorithm, the outcomes will be analyzed and compared after each run. The outcomes were written down. This procedure is done for each parameter.

Table 2.5.1 shows the parameters used to compare the algorithms:

S.No.	Parameters	Algorithm
1	No. of Clusters	Value
2	Map Topology	Value
3	Iterations	Value
4	Error Rate	Value
5	Computational Time	Time in milliseconds
6	Accuracy	High / Low
7	Time Complexity	High $O(nkl)$ / Low $O(S)^2$
8	Space Complexity	High $O(N)^2$ / Low $O(k+n)$
9	Execution Time	Fast / Slow

## 2.6 Research gap in existing literature

As the literature study shows, several researchers have employed various data mining techniques to establish consumer segmentation in the telecommunications business. According to Sharma, Panigrahi, and Kumar (2013), logistic regressions, classification, clustering, and decision trees are all effective data mining approaches for identifying consumer segmentation, and they use survival analysis and the hazard function in their suggestion. Rather than pointing out the flaws in logistic regression, classification, and clustering, they argued that K-means outperforms Fuzzy and other prediction models in consumer segmentation.

In real-world consumer transaction data, for example, outliers and missing values can be found. These are frequently problematic for clustering because they distort the results, lowering the cluster quality. SOM transforms the data into a two-dimensional map, which aids in the detection of outliers. Outliers can be identified on the map because they are the furthest away from the other groups. When a data miner discovers such a group, he or she can either delete the outliers immediately or investigate further to see if the group has any intriguing characteristics. In the case of missing values, SOM can simply be tuned to exclude missing values from training if the number of missing values is not excessive in comparison to the size of the data set.

According to the review of literature, K-means and SOM are comparable in that they calculate distances and attempt to reduce a certain error rate in order to enforce good cluster quality. SOM, on the other hand, creates clusters based on the similarities of the input data samples, clustering more similar clusters together. They perform better in terms of segmentation but not prediction. Basic K-means is restricted to numeric data and can only be applied to data that has a computed mean or median. As a result, categorical data is incompatible with basic K-means. Furthermore, K-means can only find spherical clusters of the same size, which can be restricted in some cases. It is a simple and efficient implementation of Lloyd's k means clustering technique, also known by researchers as the filtering algorithm. This algorithm is simple to implement because it only employs a kd-tree as its primary data structure. The result provides new insights into the observed effectiveness of PCA-based data reductions, beyond the conventional noise-reduction explanation that PCA, via singular value decomposition, provides the best low-dimensional linear approximal. In some cases, the performance of the SOM method on dimension reduction falls behind that of the K-mean approach as the number of clusters,  $k$ , rises. In terms of learning, the findings provide great solutions for K-means data grouping. The revised boundaries are within 0.5-1.5 percent of the ideal values, according to experiments. In comparison to other common clustering algorithms, K-means clustering enhances segmentation performance. Because of the nature of this study and the type of data, K-means is the best option. As a result, this study will make use of this strength, which is a gap in the literature review.

In the telecommunications industry, a lot of study has been done on segmentation all over the world. This is demonstrated by Bravo et al., (2017), Keramati et al., (2014) in Iran, Gerpott, Ahmadi & Weimar (2015) in Pakistan, Farquad, Ravi & Raju (2014) in South India, and others in Europe. "Ramirez-Ortiz et al., (2015) implemented a Modified K-Means Algorithm for Circular Invariant

Clustering.” The study discovered that as the number of clusters,  $k$ , increases, the performance of the SOM algorithm lags behind that of the K-mean technique. Peker et al. discussed using K-means to find nonlinearly separable clusters (2017). When it comes to segmenting data sets that are distinct or well isolated from one another, K-means outperforms SOM. Hickman et al. investigated K-Means (2013). “Principal component analysis and linear transformation were used to reduce dimensionality, and the initial centroid was calculated before using K-Means”. Only K-means lets you to apply techniques like PCA and linear transformation that aren't available in other clustering approaches like Self-organizing maps (SOM). K-means were utilized in dimensionality reduction by (Rodriguez and Laio, 2014). During the iterative steps of the K-means method, it was discovered that the high-dimensionality of data is handled well. In the Kenyan business environment few studies have used this approach in handling dimensionality, therefore, it is appropriate to use it. Testing the performance of these two algorithms in handling large datasets in a Kenyan context is the gap this study intends to utilize.

It is, however, difficult to anticipate the performance of segmentation data mining methodologies in relation to one another. When reviewing the literature, previous studies reveal a variety of outcomes about the performance of the tactics. Each technique appears to have studies to back up its superiority over others. Using k-means data methods, this study evaluated the useful patterns of consumer behavior with SOM. Because there is a paucity of literature on the use of K-means in consumer segmentation in Kenya, our study took pride in researching and contributing to it.

## **2.7 Conceptual Framework**

The Big Data Application Process in Strategic Communication is presented in this section, which is based on (Wiencierz and Roettger, 2019). The practice of strategic communication planning has been supplanted by algorithmic segmentation of big data. The term "goal definition" has been renamed "customer segmentation." The secondary data collected by the Kenyan Communication Commission is the independent variable.

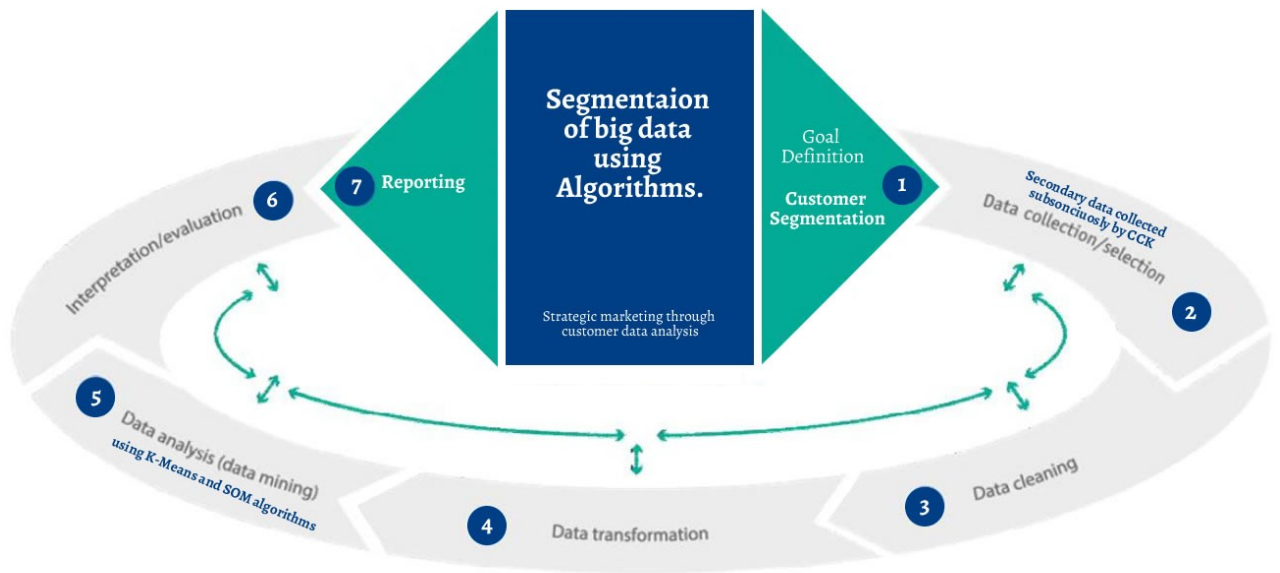
The complexity of data production and processing was captured conceptually using Fayyad et al.(1996a, 's 1996bKnowledge) Discovery in Databases (KDD) technique, which originated in (commercial) information systems. The KDD process is defined as a "nontrivial process of identifying authentic, novel, possibly beneficial, and ultimately comprehensible patterns in data" (Fayyad et al., 1996b, p. 29). The overall procedure in our study comprises a comparison of two

algorithms to massive data sets in the telecommunications industry that run smoothly, as well as the interpretation and dissemination of results for strategic marketing goals.

A comprehensive grasp of the region in which big data will be employed, as well as a clear formulation of the goal, are required in the first step of KDD (Sharafi, 2013, pp. 61–62). Customers are segmented to specify the goals to be achieved with big data while using algorithms to cluster it. In the second phase, data selection, the needed data to accomplish the stated communication goals is developed or acquired; in our study, it was gathered from the CCK database. This step is important to the project's success since "the entire study may fail if any key qualities are missing" (Maimon & Rokach, 2010, p. 3). The third phase of the KDD process is data cleaning. Outliers, inconsistencies, and faults in the data, such as random measurement and transfer mistakes, are removed at this stage (Maimon & Rokach, 2010). The data transformation phase follows, which comprises turning raw data into various data formats that can be further processed if necessary. Data reduction is the primary goal of this phase, which can be accomplished by merging variables (Fayyad et al., 1996b; Sharafi, 2013, p. 64). Data mining is divided into three steps in KDD, according to Fayyad et al. (1996a), which we have integrated into one step in this model. In this fifth step, data mining, (1) the data mining goal must be established before (2) the K-means and SOM algorithms are applied to multidimensional data to (3) look for patterns. This data mining technique will focus on summarizations, classifications, regressions, and clustering.

K-Means and SOM will be used as a meter to measure data correctness throughout data analysis and clustering. To predict future consumer behavior, classification rules are discovered using demographic variables (age, gender, location, etc.) and K-mean values of client groupings.

Managers can use the proposed model to develop more effective marketing strategies that fully utilize data mining and K-Mean-SOM analysis. Because not all clients have transacted the same amount, some have transacted more frequently, and some have transacted more recently, it is useful for anticipating client behaviors based on demographic variables.



**Figure 3. Conceptual Design**



## **CHAPTER THREE**

### **METHODOLOGY**

#### **3.0 Introduction**

Research methodology is a way to systematically solve a research problem by logically adopting various steps (Scridhar, 2008). This chapter looks at the various methods and procedures adopted in conducting the study to try and answer the research questions presented in chapter one.

#### **3.1 Research Philosophy**

A research philosophy is a set of beliefs about how to collect, process, and apply data about a phenomenon. Rather than doxology (what is thought to be true), the term epistemology (what is known to be true) encompasses a wide range of research methodologies. The chosen research philosophy of the researcher incorporates critical assumptions that serve as the foundation for the research plan (saunders, Lewis & Thornhill 2009). Data mining generates knowledge and comprehension of our environment, with or without establishing knowledge operational bounds, and serves as a philosophical study as a forerunner to technology and application. From the obtained data, the researcher will gain knowledge on how to use K-means and SOM algorithms as qualitative methods to understand people's culture and behavior in telecommunication.

#### **3.2 Research Approach**

From the given historical data, this study will take a deductive method. Existing theories and models are investigated using a deductive research approach. In this scenario, we'll look at and test existing algorithms, as well as see how K-means and SOM handle multidimensionality.

#### **3.3 Research Design**

The structure through which data is collected, measured, and analyzed, according to Kothari (2010), is referred to as study design. The role of the research design is to assist or steer the researcher in gathering essential information while taking into account time and money or expense. The archival research strategy will be used in this study's model, which will be executed utilizing R studio software. To construct consumer segments from achieved data, K-means and SOM clustering

methods will be used. The behavioral factors of customers will be used as input to the k-means and SOM algorithms. The outcomes of clustering will be assessed, and the most profitable sector will be identified using a “completely randomized design (CRD)”. The proposed strategy is a two-phased method. Data will be retrieved from the Communication Authority database from the target demographic in the first phase, followed by data purification. It will entail removing the noise and extraneous data, “followed by formatting to the appropriate standard. In second phase, clustering phase, the research will generate the clusters using K-means and SOM algorithms, then clustering results evaluated and the most profitable segment will be determined”. Consequently, results will be analyzed and presented. This process can be presented in the figure below: -

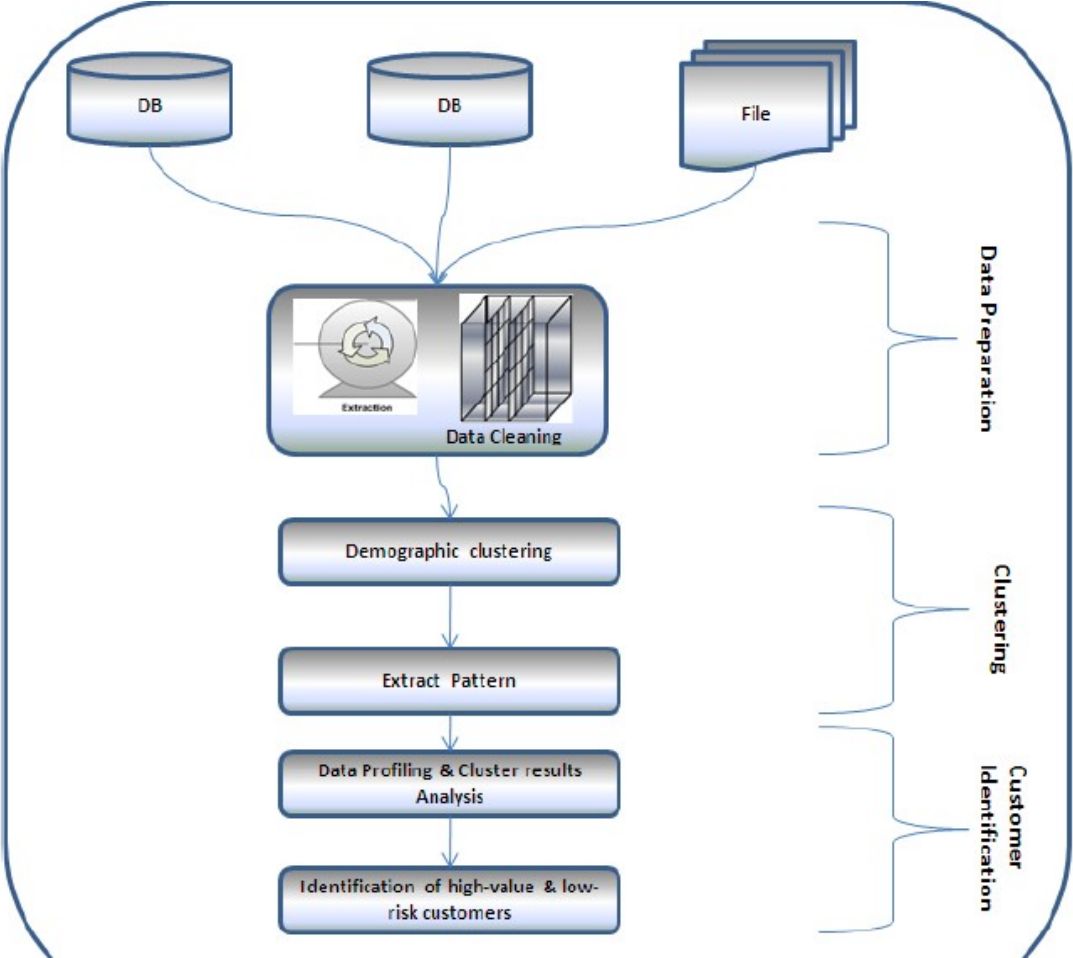


Figure 4: Design Process

### **3.4 Population**

A population is any group that is the focus of a research project (Goddard et al, 2007). The term population refers to the entire set of objects or events under investigation from which inferences can be drawn. The population is the larger set of observations, while the sample is the smaller set. When the goal of the research is to better understand the larger population, the sample should be as representative of the larger population as possible. This study included the entire group of people with characteristics that were of interest to the study, which included all of the different mobile subscribers available in the database. Customers of Safaricom, Airtel, and Telecom will be of particular interest, as these are Kenya's three major service providers.

### **3.3 Data Collection**

This is how the sample's data was gathered. The Communication Commission provided secondary data for this investigation. Traditional methods of extracting knowledge and useful information from data rely primarily on manual analysis and interpretation, which is time-consuming, costly, and prone to error. Traditional methods of managing large amounts of data have become problematic as the amount of data collected and stored in operational transaction systems and databases has grown at an unprecedented rate, necessitating the development of more sophisticated and cost-effective data mining technologies. Data mining has gotten a lot of attention because of its ability to access, model, and visualize essential links in data, and it's been used in a variety of fields, including science, marketing, investing, fraud detection, manufacturing, and communications. CRM data mining aims to help businesses improve their marketing and sales performance by gaining a better and deeper understanding of their customers and forecasting their behavior with greater accuracy.

### **3.4 Creating a target dataset**

Data relevant to the analysis will be chosen here. First, data from various sources, such as data warehouses, web servers, enterprise resource planning (ERP) systems, marketing databases, and external data providers, must be combined into a single dataset. Efforts should sometimes be concentrated on a subset of variables and data samples. The following information will be gathered:

- Demographic data: gender, age, marital status, and level. Data on purchasing habits such as the number of lines, contract type, frequency of use, churn rate, and amount spent.

### **3.5 Data cleaning and preprocessing**

R-software was used to analyze the data. Before the data can be used for the actual data mining process, it must be cleaned and formatted. In the real world, erroneous data is common. It has been demonstrated that up to 40% of data collected is unclean in some way (Maimon & Rokach, 2005). There are six of these tasks. They include discovering and repairing inconsistencies in data formats and encoding, as well as spelling errors, abbreviations, and punctuation.

Unwanted data fields are removed. From an analysis standpoint, data may contain many meaningless fields, such as production keys and version numbers.

The interpretation of codes into text or the replacement of text with meaningful numbers.

Data may include cryptic codes. These codes must be supplemented and replaced with text that is recognizable and equivalent.

Combining data from multiple tables, such as customer data, into a single common variable.

Identifying multiple used fields. One method is to count or list all of the distinct variables in a field.

Before mining, the following data preparations will be carried out:

Checking for out-of-bounds, abnormal, or ambiguous values. Some of these outliers may be correct, but they are extremely rare and thus difficult to explain.

Examining data fields that are missing or have been replaced by a default value.

Making computed fields available as inputs or targets.

Using decision trees, for example, to map continuous values into ranges.

Variables are normalized. Normalization can be classified into two types. The first type involves normalizing the values between [0, 1]. The second method is to reduce the variance to one.

Converting nominal data to metric scales (for example, yes/no answers).

Converting text to numeric data.

To avoid the problem of "garbage in – garbage out," the data should be preprocessed before mining. To improve the speed and quality of the process, the data validity is improved by cleansing, such as imputing missing values, smoothing out noisy data, and removing outliers. This will be accomplished through the use of Excel functions.

### **3.6 Data Presentation**

This is how the analyzed data will be presented for easy understanding and help in understanding trends or relationships between variables. The data in this study will be presented in tables, and some aspect of narrative form.

### **3.7 Research Limitations**

This study as any other is expected to encounter some limitations and shortcomings, the study will confine itself to data from the Communications Authority database.

### **3.8 Ethical Review**

The ethical standards of any research must be maintained. Since the researcher was handling data of a personal nature, the standards were upheld in this study. To begin with, the researcher requested approval from the commission of science and technology (NACOSTI) to be allowed to conduct the study. In addition, request for research data was officially done through the institution. Furthermore, data provided was not recent, as the study focused on its objectives which were not pegged on the dates the data was captured rather the size of the data. In conclusion, in the study, data received will be used for its intended purpose which is for academic use only.

### **3.9 Chapter Summary**

This chapter looked at the methodology that will be used to carry out the study. The subsequent chapters will present the findings of the study, recommendations and conclusions.

## CHAPTER FOUR

### DATA ANALYSIS, EXPERIMENT RESULTS AND DISCUSSION.

#### 4.0 Introduction

This chapter covers the data size, data processing, and results presentation. Multiple Correspondence Analysis (MCA) with K-Means was used to find and represent underlying patterns in a data set by modeling large amounts of data in a low-dimensional space, which was preceded by outlier detection with SOM and confirmed with the box plot approach. As a result, visualizing takes on entirely new significance. This method is the inverse of Principal Component Analysis. After presenting the large data in a low-dimensional space (multidimensionality data reduction), clusters were estimated using a stability plot, and the clustering results were validated using silhouette width. Financial and socio-demographic data for each frequently used service were then input into k-means for segmentation, and the clustering results were validated using silhouette width.

#### 4.1 Data Size and processing

The R-studio software was used to implement the research model. To create consumer groupings, the K-means clustering technique is used. The k-means and SOM algorithms were used to input customer behavioral information. The most profitable sector was found using a completely randomized design after the clustering findings were analyzed (CRD). The proposed strategy is a two-phased method. Data was gathered from the Communication Commission database and the target population in the first step, followed by data purification. It entailed eliminating the noise and extraneous data before formatting it to the appropriate quality. A total data of over 5000 entries was retrieved and edited.

Table 4.1 shows a sample of the extracted data and coded data imported into R. studio.

**Table 4. 1: A sample of the extracted data imported into R studio.**

Gender	Marital Status	Dependent	Age	Multiple Lines	Contract-Type	Monthly Charges	Service used Frequently
1	1	1	10	0	1	2	2
1	0	1	4	0	1	2	2
1	0	0	6	1	1	2	4
1	1	1	3	1	1	2	1
1	1	1	2	0	1	2	1
0	1	0	5	0	1	2	2
1	1	1	0	1	1	2	3

0	1	1	1	0	1	2	2
1	0	0	8	1	1	1	2
1	0	1	0	0	1	2	3
1	0	0	0	1	1	2	4
1	0	0	0	1	1	1	3
0	1	1	0	0	1	2	2
0	1	1	7	0	1	2	2
0	1	0	9	1	1	1	2
0	1	1	0	1	1	1	3
1	1	1	0	0	1	2	4

---

**KEY TO THE CODES**

**KEY**

**Gender**

1=Male; 0=Female

**Marital Status**

1= Married; 0=unmarried

**Dependent**

1= Dependent; 0 independent

**Age**

0 <18  
1 18-19  
2.00 20-24  
3.00 25-29  
4.00 30-34  
5.00 35-39  
6.00 40-44  
7.00 45-49  
8.00 50-54  
9.00 55-59  
10.00 60-64  
11.00 65-69  
12.00 70-74  
13.00 75-79  
14.00 80+

**Number of Lines**

1= Multiple lines; 0 = one line

**Contract-Type**

1= Postpaid; 2= Prepaid

**Monthly Charges**

1= 0-1500, 2, > 1500

**Service used frequently**

1= Voice call; 2 SMS; 3 Internet bundles; 4 = Mobile money

#### 4.1 Handle outliers in large data using SOM and K-means algorithm

The initial goal is to deal with outliers using SOM and K-means methods. SOM is trained using unsupervised learning. The main goal of SOM when translating multidimensional data into a lower dimensional space is to preserve the topology. One of the most difficult aspects of data mining is dealing with outliers. There are five broad categories of techniques for detecting outliers in the literature. Distribution, grouping, distance, density, and depth-based methods. The techniques based on statistical distributions employ standard methods for estimating statistical distributions. The depth-based methods seek to detect outliers by computing a distance measure from a specific data item to the data's centroid. The distance-based methods seek to calculate the distance between a data item and its nearest neighbor. Density-based approaches compare the density surrounding a data item to the density surrounding its local neighbors. Outlier detection methods can also be classified as supervised, semi-supervised, or unsupervised.

Outlier detection is a task that searches for items that differ or are inconsistent with the rest of the data. It has many applications, such as fraud detection, network intrusion detection, and disease clinical diagnosis. Clustering methods are commonly used to detect outliers. Clustering algorithms only consider outlier detection if it does not interfere with the clustering process.

##### 4.1.1 Outliers Detection using the box plot

This is the process of spotting and eliminating outliers from a given dataset. An outlier is an extreme value or data that diverse drastically from a given average or mean of the dataset. The standard definition for an outlier is a value which is less than Q1 or greater than Q3 by more than 1.5 times the interquartile range ( $IQR=Q3-Q1$ ). That is, an outlier is any value less than  $Q1 - (1.5 \times IQR)$  or greater than  $Q3 + (1.5 \times IQR)$ . For outlier's detection in the data, the median, quartiles were computed. A summary of the median, lower, and upper quartiles (Defined as the 25<sup>th</sup> and 75<sup>th</sup> percentiles) of the data obtained is as shown in Table 4.2

**Table 4. 2: Outliers Detection using the box plot**

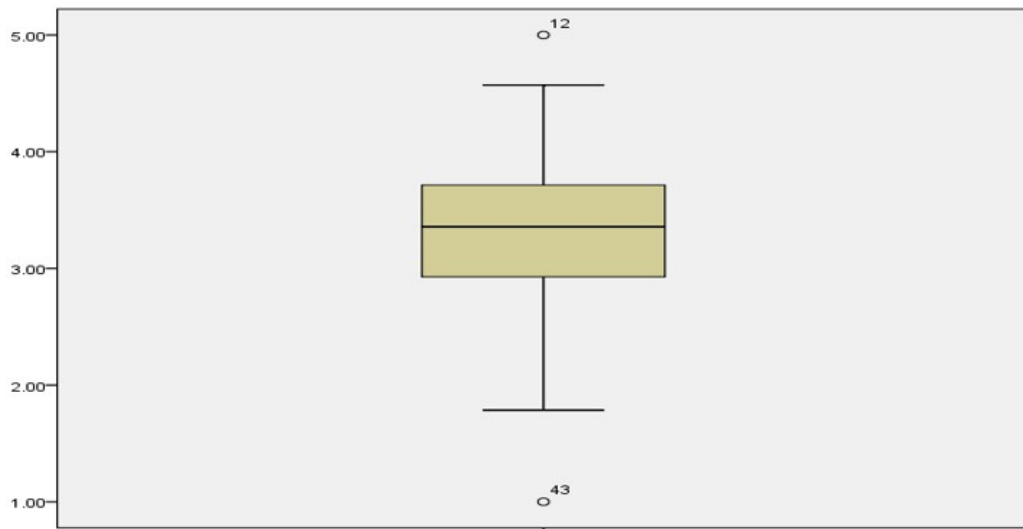
SMS	Voice Call	Data Plan	Mobile Money
Min.:18.8	Min: 18.0	Min. 18.85	Min. 19:05
Mean: 19.0	45.0	25.0	40.0
1 <sup>st</sup> Qu.: 369.8	1 <sup>st</sup> Qu: 429.2	1 <sup>st</sup> Qu: 373.39	1 <sup>st</sup> Qu: 473.00
Median: 1266.5	Median: 1450.0	Median: 1253.53	Median: 1518.83



Median: 1518.83	Mean: 2325.4	Mean: 2161.81	Mean: 2406.99
3rd Qu.:3542.1	3rd Qu.:3932.9	3rd Qu.:3479.01	3rd Qu.:4022.34
Max.:8529.5	Max.: 8476.5	Max.: 8468.20	Max.: 8564.75

---

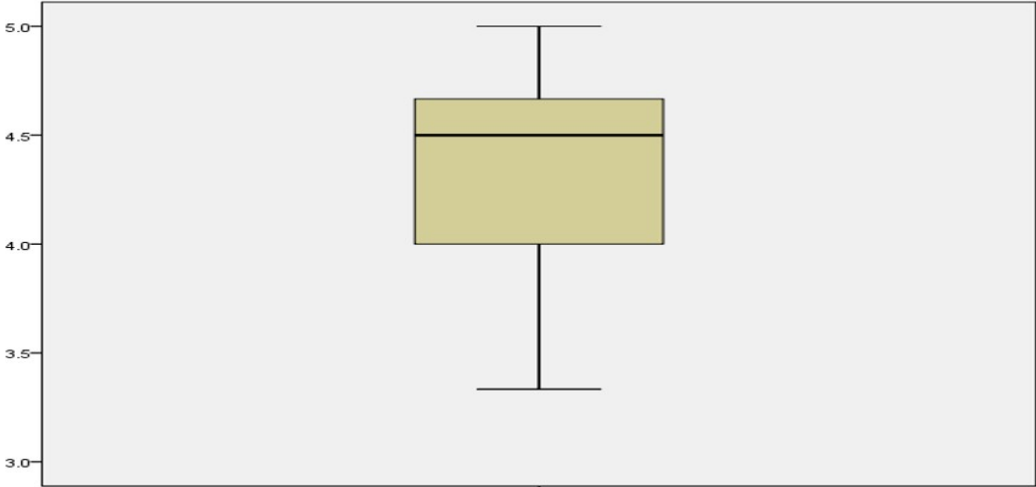
One of the problems that arose in this study was the need to transform the raw data. An effort was made to collect, clean, categorize, and gain insight from a large dataset containing 26,717 instances and 4 attributes, with an additional 4 columns comprised of formula derived values or classes to categorize the data. To address this issue, the derived attributes were performed in accordance with the needs of the learning machine scheme. Descriptive statistics were created. Table 4.2's summary of the median and quartiles was used to plot the box and whiskers in order to detect outliers and remove them from the data. A box and whisker plot reveal a "box" with the bottom edge at Q1. A box plot is a useful graphical representation for describing the behaviour of data in the middle and at the ends of a distribution. It is critical to understand that derived attributes are new variables based on original variables. Those derived variables that represent something in the real world, such as customer behaviour, are the most effective. Total values, average values, and ratios are all examples of derived variables. The derived variable of the average value, mean, mode, and median was used in this study.



**Figure 4. 1:SMS Outlier Detection a**

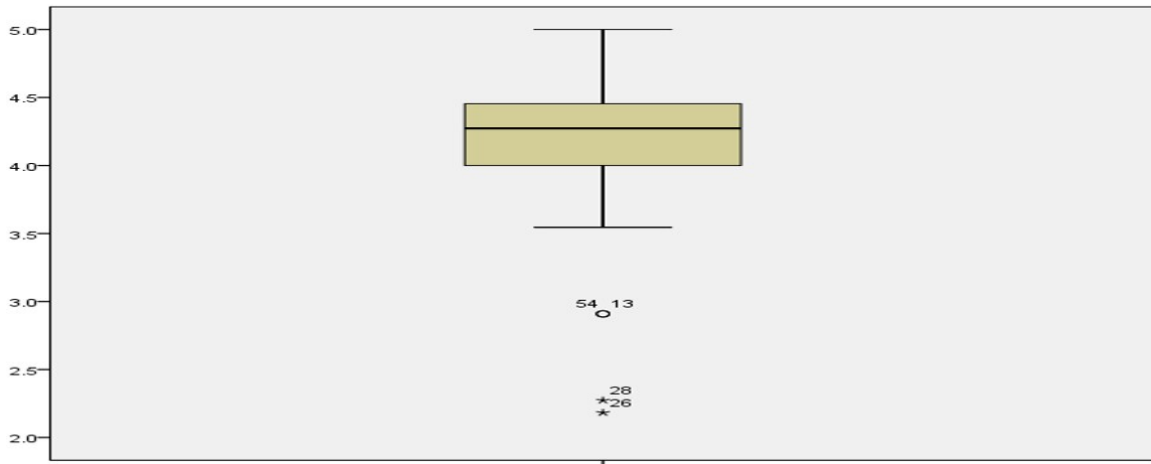
From Figure 4.1 in the outlier detection, only two points shows point of outlier detection which has no effect on the overall data.

Figure 4.1 Indicates the customers using SMS over a period. The minimum age of those who are using SMS are 18 years of age hence they are the segments which should be targeted when it comes to using short message services. Mean age is given as 19 years for SMS hence young people are the ones using short message service mostly.



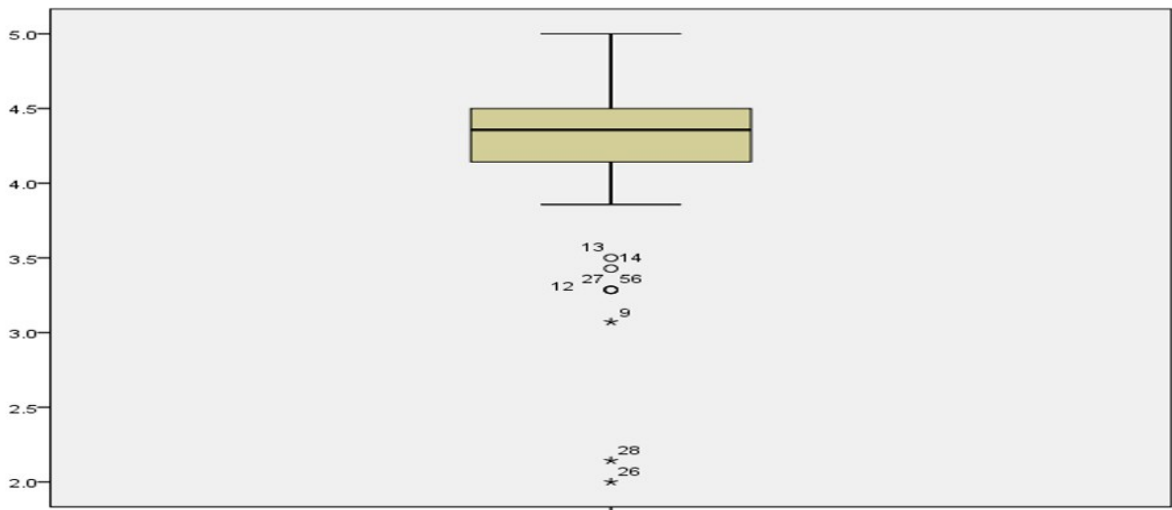
**Figure 4. 2: Voice Call Outlier Detection.**

Figure 4.2, the data does not contain any outlier since all the data are within the box plot. It means the data does not require normalization. When it comes to segmentation, age is a very important attribute, and in this study, the extracted attribute of customer age groups categorizes the customer's age into five groups, 0-14, 15-24, 25-44, 45-64, and 65+. These groups were selected based on Kenyan data groupings to allow for future comparisons. It first performs an error check to ensure that the age column is a number, then it determines which group that number belongs to and returns the value for that group. The decision tree classifier simplifies the classification process in terms of the customer's age. From the table, mean age of people using voice call frequently are 45 years and above. This indicates that the marketing on voice calls should specifically target this group ensuring that time for advertisement and promotion specifically meets their needs.



**Figure 4. 3: Data Plan Outlier Detection**

From figure 4.3, there are data points which lies outside Box plot hence in the data plan, normalization will be done to reduce the presence of outlier in the data. The minimum age of those using data is 18 years, that is per the communication authority data. The average age of customers using data is 25 years of age. This gives an impression of people who always using social media, mostly college students and those who are just in



**Figure 4. 4: Mobile Money Transfer Outlier Detection**

In Figure 4.4, several points lie outside the box plot indicating the presence of outliers in the data.

The top edge is at Q3, the "middle" of the box is at Q2 (the median), and the maximum and minimum are labeled as "whiskers." The box plot is depicted in Figures 4.1-4.4. A boxplot is a graphical representation of data distribution. The whiskers represent the data's highest and lowest non-outlier points. Outliers are represented as distinct points beyond the whiskers of a box plot. Any number less than Q1 (1.5IQR) or greater than Q3+ (1.5IQR) for outlier detection. Because no value in the data fell outside of this range, the presence of an outlier is indicated on the box plot. That is, the data contained outliers. Mobile money appeals to people of all ages, with people aged 40 and up being the most prevalent in the segment. This is the working group with money to send and receive money from the business, so it is critical for the sales team.

#### 4.1.2 Outliers detection using distance between Neurons in SOM

Anomaly detection is the detection of patterns in each data set that do not conform to a predefined, normal behavior. The detected patterns are known as anomalies, and they frequently translate to critical and actionable information in a variety of application domains. Anomalies are also known as outliers, surprises, aberrants, deviations, and peculiarities. The shortest path between all neurons in SOM of Classes 1 (red), 2 (light blue), and 3 is depicted in Figure 4.5. (yellow).



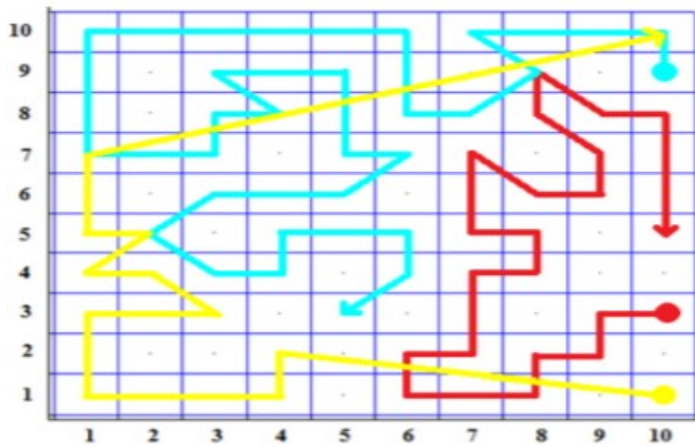
**Figure 4. 5: SOM Paths between Neurons**

The majority of Class 3 members are located in the bottom left corner of SOM. However, one member is in the upper right corner and two are in the lower right corner.

These three individuals are considered outliers.

In the SOM maps gender was used predominantly to establish the customer segments, the first and second segment is female dominated with high level of texting and data bundles usage while male counterpart dominates the second class with voice calls and data bundles. This aspect is crucial for marketing as it helps in tailor making of advertisement and product development. The main

contributions of this work can be summarized by the creation of a self-organized map from data. This allows for an examination of the major factors that contribute to the emergence of clustering. After that, customer segmentation can be created by introducing a loyalty rate. Such an analysis is useful for strategic planners, who should be aware of the main factors causing cluster formation and segmentation. This enables them to create better marketing plans that are tailored to the needs of their customers. Using this method with telecommunications customers is critical because this type of usage is heavily reliant on nonlinear factors with complex relationships. The segments are depicted in Figure 4.6 below. The shortest paths between all the same class members in SOM are found in the next step of the algorithm, as illustrated in figure 4.6.



**Figure 4. 6: Shortest path between classes**

The starting points are denoted by circles. Because the members of Class 1 and Class 2 are close to one another in Fig. 4.6, all distances between these members in the shortest path are less than the parameter value. Based on the results of the two experiments, it is possible to conclude that outlier rejection is affected by two factors: 1) a bound that indicates how many members of the same class can be so far away from their cluster that they are considered outliers; 2) a distance that indicates how far away some data should be from other data of the same class that they are also considered outliers. It is for this reason that this study has both theoretical and practical significance in developing the self-organized map algorithm and in using the results to adapt marketing future strategies.

#### **4.2 Handle multidimensional data using Multiple Correspondence Analysis (MCA)**

The first goal of this research is to determine the multidimensionality of the data handling capabilities of the K-means and SOM algorithms. K-means clustering is a cluster analysis method that aims to divide  $n$  observations into  $k$  clusters, with each observation belonging to the cluster with the closest

mean. The number of clusters is presumptively fixed. The k-means method can quickly and efficiently cluster a large data set. Self-Organizing maps (SOMs) have also been used successfully to reduce dimensionality and select features for face space representations. “Only K-means allows the use of multiple algorithms such as PCA and linear transformation, which other clustering methods such as Self-organizing maps (SOM) do not.” (Rodriguez & Laio, 2014) used K-means to reduce dimensionality. A data analysis technique called multiple correspondence analysis is used to detect and represent underlying structures in a data set. It accomplishes this by displaying high-dimensional data in a low-dimensional Euclidean space. As a result, the procedure appears to be the categorical data counterpart of principal component analysis. The results of multiple correspondence Analysis on the large data set are shown in Table 4.4.

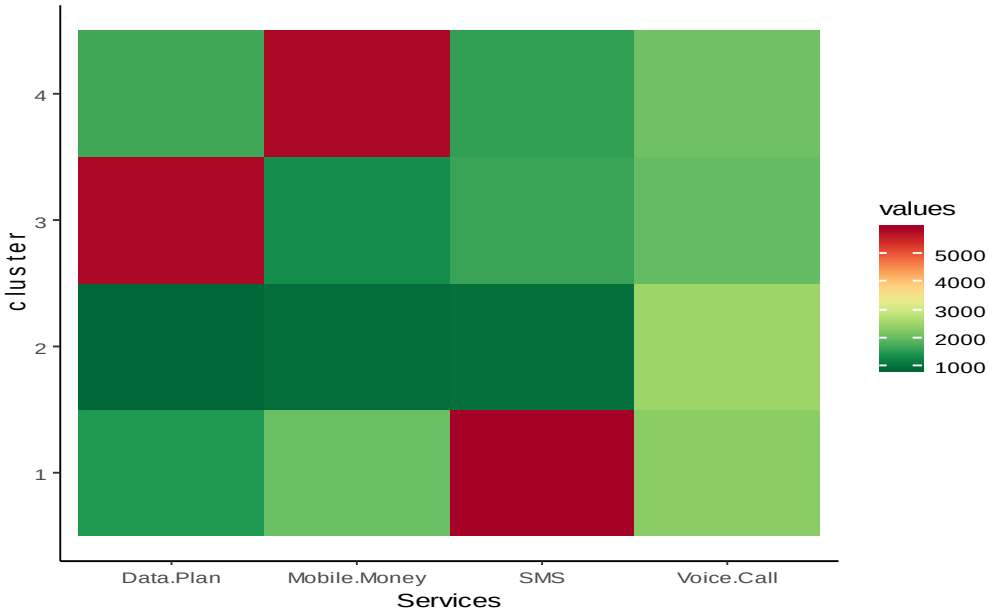
**Table 4. 3: Experiment Results for MCA on data.**

Component	Eigenvalue	Percentage of Variance	Cumulative Percentage of variance
MCA1(Services used Frequently)	0.262569	19.69265	19.69265
MCA2(Multiple Lines)	0.208968	15.67262	35.36527
MCA3(Monthly Charges)	0.170294	12.77208	48.13735
MCA4(Age)	0.168205	12.61539	60.75274
MCA5(Dependents)	0.164686	12.35148	73.10421
MCA6(Contract Type)	0.145304	10.89778	84.00199
MCA7(Gender)	0.128301	9.622567	93.62456
MCA8(Marital Status)	0.085006	6.375442	100

From Table 4.4 above MCA represented the large data into 8 components: services and frequency, multiple lines, monthly charges, age, dependents, contract type, gender and marital status.

**Using financial and socio-demographic variables as inputs in k-means for segmentation.**

In this objective, customers were clustered using the financial and socio-demographic variables. The customers are clustered according to the Mean Monthly Charge for each of the four services they frequently used, i.e., Voice Call, SMS, Data plan & Money transfer. The results are represented using a heat map to visualize the contrasts existent within and between the different clusters and services.



**Figure 4. 7: Heat Map**

Heat Map showing the four Clusters against Services used frequently

**Table 4. 4. Cluster means used to generate heat map.**

Cluster	Data Plan	M. Money	SMS	Voice Call
4	1748.08	5765.98	1563.13	2119.26
3	5803.15	1401.83	1693.67	2042.91
2	957.42	1023.55	1093.24	2578.91
1	1535.37	2198.27	5938.52	2327.55

### 4.3 Comparison of k-Means and SOM algorithms on the large data set

The two algorithms were compared based on specific parameters as shown in table 4.5 below. After each run the results were studied and compared for each algorithm. The results were written down. This step was repeated for all the parameters.

**Table 4. 5: Comparative results of both algorithms**

S.No	Parameters	Kohonen-SOM	K-Means
1	No. of Clusters	8	8
2	Map Topology	8	8
3	Iterations	5	5
4	Error Rate	0.5544	0.5726
5	Computational Time	953 ms	5578ms
6	Accuracy	High	Low
7	Time Complexity	Low $O(S)^2$	High $O(nkl)$
8	Space Complexity	High $O(N)^2$	Low $O(k+n)$
9	Execution Time	Fast	Slow

- The following conclusions can be drawn from the above table.
- In comparison, Kohonen SOM has a lower error rate (55%) than K-means (57 percent). The error rate ranges between 0 and 1.
- The computation time of the Kohonen SOM is significantly less than that of K-Means on the same data set.
- SOM produces more accurate results than k-means.
- SOM has a low time complexity but a high space complexity. Both of these parameters point to the superior algorithm.
- When compared to the K-means algorithm, the Kohonen SOM algorithm takes less time to execute.



#### 4.4 Discussion

The findings from this study are in line with (Floh et al., 2014), K-Means algorithm which is classified under unsupervised form of learning, was capable of identifying clusters without any knowledge of the dataset beforehand. “The cluster model was able to manage the large number of attributes and reveal previously unknown data-driven segments” (Guha & Mishra, 2016).

According to “Mehta et al. (2017), both K-Means and SOM have linear complexity,  $O(ndk)$  and  $O(nd)$ , respectively, where  $n$  is the number of data samples,  $d$  is the number of dimensions, and  $k$  is the number of clusters.” As a result, both methods scale to large datasets and perform well; however, in this study's comparative analysis of SOM and K-Means on the same dataset, SOM proved to be more accurate by 2% and required less computational time.

Since K-Means is the only algorithm which allows the use of multiple algorithms, MCA within PCA was used in this study to achieve multidimensionality handling. MCA represented the large data into 8 components: services and frequency, multiple lines, monthly charges, age, dependents, contract type, gender and marital status. “(Rodriguez & Laio, 2014) used K-Means in dimensionality reduction. It was found that, K-Means algorithm handles the high-dimensionality of data during the iterative steps”, in this study customers were further clustered according to the Mean Monthly Charge. Clusters were further reduced to 4 (Voice Call, SMS, Data plan & Money transfer) from the 8 components achieved by MCA. However, it is not evident how K-Means through other algorithms within it will handle multidimensionality on different datasets besides telecommunication data, especially in a Kenyan context. How can K-Means handle multidimensionality of medical or mobile taxi application use data for instance among other sources of large data with multidimensionality components?

## CHAPTER FIVE

### SUMMARY AND CONCLUSION

#### 5.0 Introduction

This chapter gives the discussion of the results and the conclusion on the research topic and findings.

#### 5.1 Summary

This examination investigated customer conduct division among portable specialist co-ops utilizing K-means calculations. The significant reason for the investigation was to provide customer behavior segmentation from subconsciously collected mobile telecommunication usage data using K-means while comparing the effectiveness of the SOM Algorithm. The researcher wanted to see how multidimensional data was handled by algorithms within K-means. To handle outliers in large datasets using the SOM algorithm in the Kenyan mobile service industry, and to conduct a literature review to compare and contrast SOM and K-means algorithms in handling large data. The findings of this study are consistent with those of Guha & Mishra, (2016), who showed that SOM just produces clusters of similar customers, but that the K-means algorithm helps them to understand how these clusters relate to the business. Normally, a business would rather undertake a segmentation that divides all of its clients into one of several groups that can easily be interpreted to inform a marketing strategy.

#### 5.2 Conclusion

The mobile telecommunications market is incredibly cutthroat. Administrators commonly need to plan different promoting strategies relying upon the assorted propensities for their mobile clients to further develop their advertising results and benefit. Client utilization patterns are depicted in Call Detail Records. They have more information than billing system data to portray customer behavior. Clustering analysis based on call detail records can provide marketing managers with more information than traditional clustering analyses. We developed a client life cycle model that included historical contributions, predicted value, and churn risk all at the same time. The client segmentation model that was used. Marketing managers can identify client segmentation with more balanced viewpoints by using three perspectives on customer value (current value, potential value, and customer loyalty). Understanding subscriber behavior and loyal groups enables the establishment of mobile customer clusters and the development of a strategic plan for each group to achieve customer happiness. According to this study, Kohonen SOM outperforms K-means. The number of clusters,

map topology, error rate, accuracy, computation time, complexity, and execution time are all used to evaluate the performance of both algorithms. Finally, when tested under identical working conditions, Kohonen SOM outperforms K-Means as a clustering technique. Future research could concentrate on methods to maintain cluster quality and optimality while reducing temporal complexity. Experiments with natural datasets with diverse features will be carried out in the future. We used the K-means approach to create segments with clearly defined borders based on the number of segments created by the SOM to aid in the examination of common properties underlying these segments.

### **5.3 Recommendation and Future Work**

Kohonen SOM surpasses K-means, as evidenced by the cluster's validation findings, while K-means gives better reporting that may be used to make key business choices. As a result, these strategies can be used to mine large data sets and clustering in mobile phone client segmentation. On the other hand, the clustering approaches must be reviewed further in connection to the large data sets available in Kenya's many sectors. This therefore, is an area for future studies where other researchers can use other tested clustering algorithms in comparison to K-means and SOM, to be able to come up with clustering from large data-sets, while measuring multi-dimensionality handling and outlier detection.

## REFERENCE

- Aggarwal, C. C., & Reddy, C. K. (2013). *Data clustering: algorithms and applications*. CRC press.
- Ahn, H., Ahn, J. J., Oh, K. J., & Kim, D. H. (2011). Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques. *Expert Systems with Applications*, 38(5), 5005-5012.
- Arasa, R., & Githinji, L. (2014). The relationship between competitive strategies and firm performance: A case of mobile telecommunication companies in Kenya. *International Journal of Economics, Commerce and Management*, 2(9), 1-15.
- Arora, D., & Malik, P. (2015). Analytics: Key to go from generating big data to deriving business value. In *Big Data Computing Service and Applications (BigDataService)*, 2015 IEEE First International Conference on (pp. 446-452). IEEE.
- Bahari, T. F., & Elayidom, M. S. (2015). An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Computer Science*, 46, 725-731.
- Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping multidimensional data*, 25, 71.
- Borg, W. & Gall, M. (2006). *Educational Research: An introduction*. New York: Longman Inc.
- Bose, I., & Chen, X. (2010). Exploring business opportunities from mobile services data of customers: An inter-cluster analysis approach. *Electronic Commerce Research and Applications*, 9(3), 197-208.
- Braha, D. (2013). *Data mining for design and manufacturing: methods and applications* (Vol. 3). Springer Science & Business Media.
- Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for segmentation prediction in telco. *Expert Systems with Applications: An International Journal*, 85(C), 204-220.
- Chen, K., Hu, Y. H., & Hsieh, Y. C. (2015). Predicting customer segmentation from valuable B2B customers in the logistics industry: a case study. *Information Systems and e-Business Management*, 13(3), 475-494.
- Chen, Y. L., Kuo, M. H., Wu, S. Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electronic Commerce Research and Applications*, 8(5), 241-251.

- Crespo, F., & Weber, R. (2015). A methodology for dynamic data mining based on fuzzy clustering. *Fuzzy Sets and Systems*, 150(2), 267-284.
- Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Segmentation prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing*, 19, 31-40.
- Gerpott, T. J., Ahmadi, N., & Weimar, D. (2015). Who is (not) convinced to withdraw a contract termination announcement? –A discriminant analysis of mobile communications customers in Germany. *Telecommunications Policy*, 39(1), 38-52.
- Floh, A., Zauner, A., Koller, M., & Rusch, T. (2014). Customer segmentation using unobserved heterogeneity in the perceived-value–loyalty–intentions link. *Journal of Business Research*, 67(5), 974-982.
- Frenkel, D., Wilkinson, K., Zhao, L., Hickman, S. E., Means, T. K., Puckett, L., & El Khoury, J. (2013). Scara1 deficiency impairs clearance of soluble amyloid- $\beta$  by mononuclear phagocytes and accelerates Alzheimer's-like disease progression. *Nature communications*, 4.
- Guha, S., & Mishra, N. (2016). Clustering data streams. In *Data Stream Management* (pp. 169-187). Springer Berlin Heidelberg.
- Gupta, G. K. (2014). *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd.
- Hamka, F., Bouwman, H., De Reuver, M., & Kroesen, M. (2014). Mobile customer segmentation based on smartphone measurement. *Telematics and Informatics*, 31(2), 220-227.
- Hickman, S. E., Kingery, N. D., Ohsumi, T. K., Borowsky, M. L., Wang, L. C., Means, T. K., & El Khoury, J. (2013). The microglial sensome revealed by direct RNA sequencing. *Nature neuroscience*, 16(12), 1896-1905.
- Hoegel, D., Schmidt, S. L., & Torgler, B. (2016). The importance of key celebrity characteristics for customer segmentation by age and gender: Does beauty matter in professional football? *Review of Managerial Science*, 10(3), 601-627.
- Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved segmentation prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, 24, 994-1012.
- Keter, A. (2015). *Challenges of Strategy Implementation in the Telecommunication Industry in Kenya: A Case of Safaricom Limited* (Doctoral dissertation, United States International University-Africa).
- Kothari, C. R. (2009). *Research methodology: Methods and techniques*. New Age International

- Kwach, J., Flora, J., & Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, 5(1), 420-430.
- Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.
- Lee, M. K., Verma, R., & Roth, A. (2015). Understanding customer value in technology-enabled services: A numerical taxonomy based on usage and utility. *Service Science*, 7(3), 227-248.
- Mahajan, V., Misra, R., & Mahajan, R. (2017). Review on factors affecting customer segmentation in telecom sector. *International Journal of Data Analysis Techniques and Strategies*, 9(2), 122-144.
- Maheshwari, A. (2014). *Business Intelligence and Data Mining*. Business Expert Press.
- Malhotra, A., & Kubowicz Malhotra, C. (2013). Exploring switching behavior of US mobile service customers. *Journal of Services Marketing*, 27(1), 13-24.
- Manzano, B. L., Means, B. K., Begley, C. T., & Zechini, M. (2015). Using Digital 3D Scanning to Create “Artifictions” of the Passenger Pigeon and Harelip Sucker, Two Extinct Species in Eastern North America: The Future Examines the Past. *Ethnobiology Letters*, 6(2), 232-241.
- Mehta, S., Cronkite, D. A., Basavappa, M., Saunders, T. L., Adiliaghdam, F., Amatullah, H., ... & Lauer, G. M. (2017). Maintenance of macrophage transcriptional programs and intestinal homeostasis by epigenetic reader SP140. *Science immunology*, 2(9).
- Mugenda & Mugenda, G.A. (2003). *Research Methods, Qualitative and Quantitative Approaches*, Kenya: ACTS press
- Murtagh, F., & Legendre, P. (2014). Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion? *Journal of Classification*, 31(3), 274-295.
- Ndambuki, A., Bowen, M., & Karau, J. (2017). The effects of business strategies on growth of market share in the telecommunications industry in Kenya: a case study of Telkom Kenya. *European Journal of Business and Strategic Management*, 2(4), 16-32.
- Ocuppaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487-501.
- Peker, S., Peker, S., Kocyigit, A., Kocyigit, A., Eren, P. E., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence & Planning*, 35(4), 544-559.

- Premkumar, G., & Rajan, J. (2017). Customer Retention in Mobile Telecom Service Market in India: Opportunities and Challenges. *Ushus-Journal of Business Management*, 12(2), 17-29.
- Ramirez-Ortiz, Z. G., Prasad, A., Griffith, J. W., Pendergraft III, W. F., Cowley, G. S., Root, D. E., ... & Means, T. K. (2015). The receptor TREML4 amplifies TLR7-mediated signaling during antiviral responses and autoimmunity. *Nature immunology*, 16(5), 495-504.
- Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, 344(6191), 1492-1496.
- Sharma, A., Panigrahi, D., & Kumar, P. (2013). A neural network-based approach for predicting customer segmentation in cellular network services. *arXiv preprint arXiv:1309.3945*.
- Shmueli, G., & Lichtendahl Jr, K. C. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons.
- Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.
- Wang, W., & Fan, S. (2014). Application of Data Mining Technique in Customer Segmentation of Shipping Enterprises. In *Database*
- Wei, J. T., Lee, M. C., Chen, H. K., & Wu, H. H. (2013). Customer relationship management in the hairdressing industry: An application of data mining techniques. *Expert Systems with Applications*, 40(18), 7513-7518.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3), 29.

## Appendix I:

### Loading data

```
# Telcom.data <- read.csv (file. choose() , header = T)

#head(telcom.data)

#install.packages("ClustOfVar", dependencies = TRUE)

#install.packages("cluster", dependencies = TRUE)

#library (ClustOfVar)
```

## Appendix 2

### Outliers detection

```
> df<-read.csv("R PROJECT-New.csv")
> df1<-na.omit(df)
> summary(df1)
> boxplot(df1[,2], col = rgb(0,0,1,0.5), main = "Boxplot of df1[,2]")
```



## Appendix 3

### Handle multidimensional data using Multiple Correspondence Analysis (MCA)

```
#telcom.data <- read.csv("DATA.csv")
#telcom.data $customerID = NULL
res.mca = MCA (telcom.data, quanti.sup=c (7:8))
dimdesc(res.mca)
res.mca$eig
# column coordinates
head(res.mca$var$coord)
#Screeplot
fviz_screepLOT(res.mca, addlabels = TRUE, ylim = c(0, 45))
```

## Appendix 4

### Determine the value of parameter K

```
#stab <- stability (tree, B=50) # "B=50"
#plot(stab)
```

## Appendix 4

### Segmentation using K-means algorithm

#### Import data

```
<- read.csv ("R PROJECT-New.csv")
>View(df)
>
```

```

>
>kmean_withinss<- function(k) {
+   cluster<- kmeans(na.omit(df), k)
+   return (cluster$tot.withinss)
+ }

```

```

>

```

### Testing the function

```

>kmean_withinss(4)

```

```

[1] 9062904805

```

```

>

```

```

> # Set maximum cluster

```

```

>max_k<-50

```

```

> # Run algorithm over a range of k

```

```

>wss<- sapply(4:max_k, kmean_withinss)

```

```

>km_cluster<- kmeans(na.omit(df), 4)

```

```

>km_cluster

```

K-means clustering with 4 clusters of sizes 143, 264, 184, 139

### Cluster means:

Cluster	Data Plan	M. Money	SMS	Voice Call
4	1748.08	5765.98	1563.13	2119.26
3	5803.15	1401.83	1693.67	2042.91
2	957.42	1023.55	1093.24	2578.91
1	1535.37	2198.27	5938.52	2327.55

Clustering vector:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  
1 1 4 3 2 1 4 1 1 4 1 2 4 3 4 3 1 1 3 1  
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40  
1 3 4 1 3 1 1 2 3 3 1 4 4 1 1 2 1 2 3 3  
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60  
2 1 1 4 3 3 1 1 1 4 1 3 1 1 1 1 2 2 2 1  
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80  
3 2 1 2 1 2 1 2 3 4 2 1 1 4 3 3 4 4 1 1  
81 82 83 84 85 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101  
1 2 4 1 2 1 4 1 2 2 1 3 2 1 1 2 2 3 2 3  
102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121  
1 1 1 3 4 2 4 4 1 3 1 2 1 4 3 3 1 4 2 4  
122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141  
4 2 1 2 4 1 1 2 4 1 4 3 1 2 1 1 1 1 1 4  
142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161  
2 3 2 3 2 1 1 2 4 4 3 2 3 3 3 2 4 2 1 2  
162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181  
1 1 3 3 4 3 4 4 4 3 4 1 3 4 4 3 1 3 3 3  
182 183 184 185 186 187 189 190 191 192 193 194 195 196 197 198 199 200 201 202  
2 1 1 3 2 4 3 1 2 1 1 1 1 1 4 1 3 1 1 1  
203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222  
1 1 1 1 1 1 1 3 1 1 1 4 4 4 2 2 4 3 1 1  
223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242  
4 1 4 4 4 3 1 4 3 1 4 1 3 2 4 1 2 2 1 1  
243 244 245 246 247 248 249 250 251 252 253 254 255 257 258 259 260 261 262 263  
1 1 2 1 2 1 3 1 1 1 2 4 1 2 2 2 1 2 1 1  
264 265 266 267 268 269 270 271 272 274 275 276 277 278 279 280 281 282 283 284

1 4 2 3 1 1 3 2 4 3 1 3 4 4 1 2 3 3 2 1  
285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304  
2 2 4 4 1 2 2 1 1 4 1 1 4 3 1 4 1 4 1 3  
305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324  
2 1 3 4 1 3 2 4 4 1 4 1 3 4 1 2 3 1 4 1  
325 326 327 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345  
1 1 1 2 1 4 2 4 2 3 3 4 2 4 2 3 3 1 4 4  
346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365  
4 3 1 4 1 1 3 3 4 3 2 1 2 1 3 1 1 3 1 1  
366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385  
4 1 1 1 3 2 4 2 4 3 1 3 1 4 1 1 3 1 4 3  
386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405  
1 2 4 4 1 1 4 1 4 1 1 4 4 2 4 4 1 4 1 1  
406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425  
1 1 3 2 4 4 4 1 2 4 4 2 1 1 2 4 1 2 3 2  
426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445  
4 2 4 1 2 1 1 2 4 4 1 4 3 4 4 2 1 2 3 4  
446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465  
1 4 2 4 4 3 1 3 3 4 1 3 2 2 1 1 3 4 4 3  
466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485  
3 1 1 1 3 2 1 1 1 1 3 1 1 3 2 4 3 2 3 4  
486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505  
1 1 1 2 4 1 1 2 4 4 3 2 2 2 1 3 4 1 3 1  
506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525  
1 1 1 4 4 1 4 2 2 1 1 1 1 4 4 1 3 3 4 1  
526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545  
3 3 2 3 2 1 2 3 3 1 1 1 4 3 1 4 1 4 4 2  
546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565

```

3 4 4 4 1 1 1 1 2 4 1 1 4 4 1 4 1 2 4 4
566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585
1 4 2 3 4 3 4 3 3 2 4 2 1 1 3 1 1 4 1 4
586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605
3 3 3 1 2 1 3 2 1 4 3 3 4 3 1 3 1 2 1 1
606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625
1 1 1 2 4 4 4 1 1 1 2 3 1 1 3 2 1 3 2 1
626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645
2 1 4 1 4 2 1 1 1 1 4 3 2 1 1 4 2 4 4 1
646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665
1 4 1 3 1 3 4 1 2 3 1 3 1 1 3 4 3 2 1 4
666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685
2 1 1 4 2 4 3 3 3 3 4 1 4 1 1 2 1 1 4 4
686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705
1 3 4 2 1 1 4 2 3 3 1 1 4 3 2 4 2 2 4 2
706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725
2 3 2 1 1 1 1 1 4 2 1 1 1 1 1 1 4 1 1 3
726 727 728 729 730 731 732 733 734 735
1 3 1 1 1 1 1 1 2 1

```

Available components:

```

[1] "cluster" "centers" "totss" "withinss" "tot.withinss"
[6] "betweenss" "size" "iter" "ifault"
>km_cluster$withinss
[1] 2403798571 1724343679 1498314243 1983538493
>km_cluster$cluster

```

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  
 1 1 4 3 2 1 4 1 1 4 1 2 4 3 4 3 1 1 3 1  
 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40  
 1 3 4 1 3 1 1 2 3 3 1 4 4 1 1 2 1 2 3 3  
 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60  
 2 1 1 4 3 3 1 1 1 4 1 3 1 1 1 1 2 2 2 1  
 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80  
 3 2 1 2 1 2 1 2 3 4 2 1 1 4 3 3 4 4 1 1  
 81 82 83 84 85 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101  
 1 2 4 1 2 1 4 1 2 2 1 3 2 1 1 2 2 3 2 3  
 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121  
 1 1 1 3 4 2 4 4 1 3 1 2 1 4 3 3 1 4 2 4  
 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141  
 4 2 1 2 4 1 1 2 4 1 4 3 1 2 1 1 1 1 1 4  
 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161  
 2 3 2 3 2 1 1 2 4 4 3 2 3 3 3 2 4 2 1 2  
 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181  
 1 1 3 3 4 3 4 4 4 3 4 1 3 4 4 3 1 3 3 3  
 182 183 184 185 186 187 189 190 191 192 193 194 195 196 197 198 199 200 201 202  
 2 1 1 3 2 4 3 1 2 1 1 1 1 1 4 1 3 1 1 1  
 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222  
 1 1 1 1 1 1 1 3 1 1 1 4 4 4 2 2 4 3 1 1  
 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242  
 4 1 4 4 4 3 1 4 3 1 4 1 3 2 4 1 2 2 1 1  
 243 244 245 246 247 248 249 250 251 252 253 254 255 257 258 259 260 261 262 263  
 1 1 2 1 2 1 3 1 1 1 2 4 1 2 2 2 1 2 1 1  
 264 265 266 267 268 269 270 271 272 274 275 276 277 278 279 280 281 282 283 284  
 1 4 2 3 1 1 3 2 4 3 1 3 4 4 1 2 3 3 2 1

285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304  
2 2 4 4 1 2 2 1 1 4 1 1 4 3 1 4 1 4 1 3  
305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324  
2 1 3 4 1 3 2 4 4 1 4 1 3 4 1 2 3 1 4 1  
325 326 327 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345  
1 1 1 2 1 4 2 4 2 3 3 4 2 4 2 3 3 1 4 4  
346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365  
4 3 1 4 1 1 3 3 4 3 2 1 2 1 3 1 1 3 1 1  
366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385  
4 1 1 1 3 2 4 2 4 3 1 3 1 4 1 1 3 1 4 3  
386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405  
1 2 4 4 1 1 4 1 4 1 1 4 4 2 4 4 1 4 1 1  
406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425  
1 1 3 2 4 4 4 1 2 4 4 2 1 1 2 4 1 2 3 2  
426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445  
4 2 4 1 2 1 1 2 4 4 1 4 3 4 4 2 1 2 3 4  
446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465  
1 4 2 4 4 3 1 3 3 4 1 3 2 2 1 1 3 4 4 3  
466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485  
3 1 1 1 3 2 1 1 1 1 3 1 1 3 2 4 3 2 3 4  
486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505  
1 1 1 2 4 1 1 2 4 4 3 2 2 2 1 3 4 1 3 1  
506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525  
1 1 1 4 4 1 4 2 2 1 1 1 1 4 4 1 3 3 4 1  
526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545  
3 3 2 3 2 1 2 3 3 1 1 1 4 3 1 4 1 4 4 2  
546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565  
3 4 4 4 1 1 1 1 2 4 1 1 4 4 1 4 1 2 4 4

```

566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585
  1  4  2  3  4  3  4  3  3  2  4  2  1  1  3  1  1  4  1  4
586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605
  3  3  3  1  2  1  3  2  1  4  3  3  4  3  1  3  1  2  1  1
606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625
  1  1  1  2  4  4  4  1  1  1  2  3  1  1  3  2  1  3  2  1
626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645
  2  1  4  1  4  2  1  1  1  1  4  3  2  1  1  4  2  4  4  1
646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665
  1  4  1  3  1  3  4  1  2  3  1  3  1  1  3  4  3  2  1  4
666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685
  2  1  1  4  2  4  3  3  3  3  4  1  4  1  1  2  1  1  4  4
686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705
  1  3  4  2  1  1  4  2  3  3  1  1  4  3  2  4  2  2  4  2
706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725
  2  3  2  1  1  1  1  1  4  2  1  1  1  1  1  1  4  1  1  3
726 727 728 729 730 731 732 733 734 735
  1  3  1  1  1  1  1  1  2  1
>km_cluster$centers
>km_cluster$iter
[1] 3
>km_cluster$ifault
[1] 0
>km_cluster$betweenss/km_cluster$tot.withinss
[1] 0.9584079
>
>library(tidyr)
>

```



```

> # create dataset with the cluster number
>
>cluster<- c (1: 4)
> center_df1 <- data.frame(cluster,center)
>
>
> # Reshape the data
>center_reshape<- gather (center_df1, Services, values, SMS: Mobile.Money)
>head(center_reshape)
>library (RColorBrewer)
> # Create the palette
>hm.palette<-colorRampPalette(rev(brewer.pal(10, 'RdYlGn'))),space='Lab')

```

### **Plot the heat map**

```

>ggplot (data = center_reshape, aes (x = Services, y =cluster, fill = values)) +
+   scale_y_continuous (breaks = seq (1, 4, by = 1)) +
+   geom_tile () +
+   coord_equal () +
+   scale_fill_gradientn (colours = hm.palette(90)) +
+   theme_classic ()

```

#Within clustering sum of the squares by cluster:

```
# [1] 1. 1.0823726 0.5850840.7593950 2.6586792 2.7054773
```

```
# (between_SS / total_SS = 94.7 %)
```

## Appendix 5

### Cluster Validation Analysis

```
> km_stats
```

```
$`n`
```

```
[1] 730
```

```
$cluster.number
```

```
[1] 4
```

```
$cluster.size
```

```
[1] 162 281 144 143
```

```
$min.cluster.size
```

```
[1] 143
```

```
$clus.avg.silwidths
```

```
1 2 3 4
```

```
0.2650204 0.3570662 0.1945318 0.2523378
```

```
$avg.silwidth
```

```
[1] 0.2840627
```

```
cluster size ave.sil.width
```

```
R-Codes Validation
```

```
> set_wd <-("C:\\Users\\user\\Desktop")
```

```
> read.csv <- ("C:\\Users\\user\\Desktop\\R PROJECT-New.csv")
```

```
> df <-read.csv ("C:\\Users\\user\\Desktop\\R PROJECT-New.csv")
```

```
> library(factoextra)
```

```
> library(fpc)
```

```

> library (NbClust)
> new_df <- na.omit(df)
> rescale_df <- scale(new_df)
> km.df <- eclust(rescale_df, "kmeans", k = 4, nstart = 50, graph = FALSE)
> fviz_cluster (km.df, geom = "point", ellipse.type = "norm", palette = "jco", ggtheme =
theme_minimal())
> km_stats <- cluster.stats(dist(rescale_df), km.df$cluster)
> km_stats$dunn
[1] 0.04333782
> km_stats
$`n`
[1] 730
$cluster.number
[1] 4
$cluster.size
[1] 162 281 144 143
$min.cluster.size
[1] 143
$noisen
[1] 0
$diameter
[1] 4.445769 3.978445 5.278754 5.189887

$average.distance
[1] 1.848909 1.581565 2.170621 1.998190
$median.distance
[1] 1.791695 1.448536 2.141611 1.912333

```

\$separation

[1] 0.2349072 0.2287697 0.3161671 0.2287697

\$average.toother

[1] 2.948614 2.795530 3.124452 3.111752

\$separation.matrix

    [,1]  [,2]  [,3]  [,4]

[1,] 0.0000000 0.2349072 0.3161671 0.4454469

[2,] 0.2349072 0.0000000 0.3704251 0.2287697

[3,] 0.3161671 0.3704251 0.0000000 0.3330158

[4,] 0.4454469 0.2287697 0.3330158 0.0000000

\$ave.between.matrix

    [,1]  [,2]  [,3]  [,4]

[1,] 0.000000 2.639024 3.231568 3.272038

[2,] 2.639024 0.000000 2.909249 2.858315

[3,] 3.231568 2.909249 0.000000 3.425985

[4,] 3.272038 2.858315 3.425985 0.000000

\$average.between

[1] 2.972462

\$average.within

[1] 1.770791

\$n.between

[1] 193255

\$n.within

```
[1] 72830
$max.diameter
[1] 5.278754
$min.separation
[1] 0.2287697
$within.cluster.ss
[1] 1481.411
$clus.avg.silwidths
      1      2      3      4
0.2650204 0.3570662 0.1945318 0.2523378
$avg.silwidth
[1] 0.2840627
$pearsongamma
[1] 0.532691
$dunn
[1] 0.04333782
$dunn2
[1] 1.215793
$entropy
[1] 1.341116
$wb.ratio
[1] 0.5957321
$ch
[1] 234.3512
$cwidegap
[1] 1.146254 0.645371 2.228716 1.289683
```

```
$widestgap
```

```
[1] 2.228716
```

```
$index
```

```
[1] 0.4286963
```

```
$corrected.rand
```

```
NULL
```

```
$vi
```

```
NULL
```

```
> fviz_cluster(res.fanny, geom = "point", ellipse.type = "norm", repel = FALSE,palette = "jco",  
ggtheme = theme_minimal(),legend = "right")
```

```
> library(fpc)
```

```
> library (NbClust)
```

```
> new_df <- na.omit(df)
```

```
> km.df <- eclust(new_df, "kmeans", k = 4, nstart = 50, graph = FALSE)
```

```
> fviz_silhouette(km.df, palette = "jco",ggtheme = theme_minimal())
```

```
FCM-Algorithm Experiment
```

```
> set_wd <-("C:\\Users\\user\\Desktop")
```

```
> read.csv <- ("C:\\Users\\user\\Desktop\\R PROJECT-New.csv")
```

```
> df <-read.csv ("C:\\Users\\user\\Desktop\\R PROJECT-New.csv")
```

```
> library(factoextra)
```

```
> library(cluster)
```

```
> new_df <- na.omit(df)
```

```
> fanny (new_df, 4, metric = "euclidean", stand = FALSE)
```

```
Fuzzyness coefficients:
```

```
  dunn_coeff  normalized
```

```
2.500000e-01 -1.854072e-14
```

```
Closest hard clustering:
```

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20  
1 1 2 2 2 1 2 2 1 2 2 2 1 2 2 2 1 1 2 1  
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40  
1 1 2 2 2 1 1 2 2 2 1 2 1 1 1 2 2 2 1 2  
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60  
2 1 1 2 2 1 1 1 1 1 1 1 2 1 1 2 2 1 2 1  
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80  
1 2 1 2 1 2 1 1 1 2 2 1 1 2 2 2 1 2 1 1  
81 82 83 84 85 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101  
1 1 2 2 2 1 2 2 2 2 2 1 2 1 1 2 1 2 2 2  
102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121  
1 1 1 1 2 2 2 2 1 2 1 2 2 1 2 2 2 2 2 1  
122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141  
2 2 1 2 1 1 1 2 2 2 2 2 1 2 1 1 2 1 1 2