UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

CONSUMER SEGMENTATION AND PROFILING USING DEMOGRAPHIC DATA AND SPENDING HABITS OBTAINED THROUGH DAILY MOBILE CONVERSATIONS

By

Wambui Samuel Kamande

P52/85536/2016

Supervisor
Dr. Evans A. K. Miriti

A RESEARCH PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT FOR THE REQUIREMENTS OF THE AWARD OF DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE, SCHOOL OF COMPUTING AND INFORMATICS, UNIVERSITY OF NAIROBI

DECLARATION

Researcher's Declaration

This project report is my original work and has not been presented in any other institution for the purpose of an academic award. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the respective source.

SIGNATURE: _____ Date: _____

Wambui Samuel Kamande
Registration Number: P52/85886/2016

Supervisor's Approval
This project report has been submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University Supervisor.

SIGNATURE: _____ Date: _____

Dr. Evans A. K. Miriti
School of Computing and Informatics
University of Nairobi

DEDICATION

To my wife, Tracey Shiundu-Kamande

and

Mother, Sarah Wambui

ACKNOWLEDGEMENT

To God for the never-ending love, grace and strength over the two years. I cannot sufficiently express my gratitude.

I am immensely grateful to my supervisor, Dr. E.K. Miriti and panellists, Dr. C.K. Chepken and Dr. S.N. Mburu for their guidance, positive feedback and above all their valuable time and advice through my project work.

I am greatly obliged to mSurvey for not only the willingness to provide the data for this research, but also for availing resources to turn the results into a product and completely back its development.

Endless appreciation to my wife, Tracey Auma-Shiundu for her prayers as well as the serenity and stability she has provided for this journey to be near seamless. To Sarah Wambui, the most supportive mother a son could ask for, I am eternally grateful. I would not have the motivation and strength for this journey without her.

My classmates of two years at the School of Computing and Informatics made it all a fun journey, over and above the continuous learning from each other; and for that, I am much obliged.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Knowledge of customer behaviour helps organizations to continuously re-evaluate their strategies with the consumers and plan to improve and expand their application of the most effective strategies. The Kenyan consumer remains dynamic and the market is increasingly become transformational, characterised by high population growth, a youthful demographic, healthy urbanization, an emerging optimistic consumer class, albeit with unpredictable expenditure patterns. In addition to understanding demographic habits and product preferences, comprehensively factoring in consumer spending habits, their relationship to marketing reception and brand reception, and how they morph with time is crucial. Customer segmentation and profiling has become an indispensable tool for organisations to understand all these. The process is based on both internal data on expenditure, augmented by other research data. The consumer, however, does not spend in isolation. Every purchase they make affects another. Using expenditure data collected through daily mobile conversations with consumers in Kenya, this study sought to compare various clustering algorithms and establish one that best segments consumers, and subsequently providing profiles that provide a basis for marketing and brand strategy based on existing demographic data – age, gender, region and primary income source. K-Means, Hierarchical and Partitioning around Medoids (PAM) clustering algorithms were compared using internal and stability validation tests. Internal validation consists of three measures that compare the compactness, connectedness and separation of the cluster partitions through the Connectivity, Dunn index and Silhouette measures. Hierarchical clustering with four clusters had the best Connectivity (0.847) and Silhouette width (0.924) measures. Stability validation compares the results by removing a column, one at a time. Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance Between Means (AND) and Figure of Merit (FOM) were used to compare the algorithms. Again, Hierarchical clustering with four clusters was found to partition the data best. A rank aggregation of the two measures was not different. A four cluster Hierarchical fit performed best in four out of seven measures. The algorithm was fit into the data using an agglomerative approach and the four clusters profiled based on the available demographic characteristics. The study forms a basis for the use of additional profile descriptors once available to provide a firmer understanding of the customer segments built on expenditure data in Kenya. Thereafter, classification into a specific homogenous segment for marketing and brand targeting will be possible, given the consumers demographic characteristics.

CHAPTER 1: INTRODUCTION

1.1 Background

Consumer understanding is at the heart of product marketing and strategy in any industry. The deeper the understanding, the better. The current economy is a fast-moving and heavily dynamic world of a marketing characterised by both product and customer orientation. An imperative piece to achieving expansion in revenue and profitability is the management of customer treatment. Customer knowledge and comprehension of the behaviour can be very useful in the process of re-evaluating strategies with the goal of improving and expanding strategies that are effective for marketing teams (Hosseini & Shabani, 2015). The process of understanding consumers remains continuous and increasingly requires innovative ways to keep up with the dynamism of the consumers and their uptake of products overtime. The consumer has never been more dynamic and the market more transformational as characterised by the explosive population growth, the youthful demographic, healthy urbanization and an emerging optimistic consumer class. In addition to understanding their demographic habits and their product preferences, comprehensively factoring in their spending habits and how they morph is crucial. The spending habits of consumers shift in line with seasons, macro-economic environment as well as individual economic growth or lack thereof. The consumption of products is subsequently directly affected by these spending habits, thus making it more imperative now more than ever for consumer product manufactures and service providers to factor this into their tactics and strategies. With increased consumer data and computing power, all industries stand to benefit significantly.

Consumer segmentation and profiling has been an indispensable tool for organizations to understand the market, who to target with what product and how to optimize the marketing strategy. The two-step process is based on both internal data as well as survey data to establish the segments and profile to establish the parameters that best explain behaviour. Establishing accurate consumer spending habits and injecting this data into the available to demographic data for segmentation and profiling could significantly improve consumer understanding, thereby optimizing product design and marketing strategies. There are many ways of obtaining spending data. Daily conversations with consumers on what they spent money on the previous day through text messages provides a novel way of doing this,

especially in emerging markets where most transactions still happen through cash. In Kenya for instance, where this paper focuses on, livelihood transactions are mainly conducted via cash (about 77% on average), except for those who are employed, over half of whom receive their payments electronically (mobile financial services and bank transfers).

1.2 Problem Statement

Customer (and consumer) segmentation in emerging markets has been largely driven by market surveys and descriptive analysis of various characteristics to construct "personas" that advise product marketing. The surveys are limited by costs of gathering longitudinal data on variables such as spending habits. The segmentation and profiling thereby does not include key components that split consumers into homogeneous groups which best align with purchase behaviour, a combination of preference and ability. Additionally, segmentation has been mostly confined to the behaviour in relation to a specific product or category of products. This approach is beneficial to companies, but only to a certain extent in that it falls short of understanding the consumer wholly.

Whereas a key component across various organizations in product design and development is market segmentation (Pedro et al. 2015), the organization does not understand the consumer regarding other basic and secondary expenditure habits outside their own.

Existing approaches that do not generate from internal data alone do not cater for inclusion of other data sources to improve the knowledge and aid better, more accurate and effective sales and marketing strategies. There is a need to include more diverse data from non-traditional sources for enriched understanding. As the amount of the data collected increases, application of the proposed clustering and profiling algorithms will be automated using data mining techniques. This opens the ability to combine both structured and unstructured big data. The use of mobile phone surveys to collect self-reported spending information from previously unreachable consumers also makes it possible to leverage on technology to update the segments automatically in line with shifts in the market for an updated and consolidated understanding of the opportunities.

1.3 Objectives

The main objective of the research is to evaluate and compare the performance of clustering techniques and establish one that best segments the sample data collected via mobile into homogeneous groups based on spending habits in Kenya and then profile the groups by describing them based on their characteristics.

The specific objectives are as follows:

1. Compare clustering algorithms and identify the best performer for consumer segmentation based on spending data collected through mobile surveys and other behavioural data in Kenya
2. Identify the best profile descriptors of the established clusters based on available demographic characteristics
3. Provide practical recommendations on how the generated segments and profiles can provide more precision in the process of marketing design and strategy to (in)form improved propositions towards improving the consumer connections

1.4 Significance

This study delves into various clustering algorithms in consumer segmentation using the spending habits across various categories, information that has been gathered daily in a ten-month period. The algorithm that best divides the data into homogeneous segments will be selected and profiling of the clusters based on the available demographic variables done. Market segments were conceptualized and introduced by Wendell Smith (1956) and have since become an integral part of modern day marketing through multiple iterations and improvement. Smith proposed a market segmentation approach as an alternative evolution method to differentiate products in imperfectly competitive markets. A market segment is defined as a clearly identifiable group within the market based on a specific set of criteria. Consumers within a segment are assumed to be similar in their characteristics, needs and even behaviour. Additionally, other studies have demonstrated that formidable results are achievable through clustering methods for consumer segmentation to advance marketing strategies leading to growth in measurable revenue gains.

Traditional market segmentation has been dominantly based on the customer behaviour attached to an organization. Being responsive to customer demands in a timely manner is immensely beneficial in the establishment of strong and abiding relationships between an organization and its customers and intensify customer repeat purchase decisions (Ozer, 2001; Anderson et al., 2004; White & Yu, 2005; Chang & Ku, 2009). However, as organizations grapple with increased competition and disruption by new companies that are driven by technology innovate and move faster, understanding the consumer entirely is not a nice-to-have but a must-have. It is imperative that organizations understand the various groups of consumers in the market to drive new ways of engagement, anticipate and act on shifting consumer needs and take advantage of the opportunities. Shaw et al. (2001) articulate the goal of clustering as ensuring that all instances in each cluster have significance similarity to one another and are distinct from occurrences in the rest of the clusters.

Baines et al. (2010), in their consideration of the question whether the approach of segmenting the market has been ousted by other models of customer insights focus on three main research questions:

1. Have processes focusing on insights into the distinct consumer superseded market segmentation techniques?
2. How are segments defined in contemporary organizations?
3. How are various segmentation procedures implemented?

Following comprehensive literature survey and review, they draw the conclusion that segmentation process has largely focused on selection of bases whereas the anchor should be how a generated segmentation programme is used upon generation. However, some researchers such as (Dibb and Simkin, 1997; Dibb and Wensley, 2002; Dibb, 2005; Laiderman, 2005) go the extra step to illustrate how segments may be purposeful in the market understanding strategy procedure to (in)form robust propositions.

In this paper, the selection of a totally different group of variables and data and the provision of practical recommendations to help businesses make decisions based on the generated segments and profiles is significant in bridging the gap above.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Clustering can be defined as the procedure of splitting data into groups. The main objective is that instances/elements in each group have significantly more similarity between them than with those outside the group. The subsets/groups should be relevant based on a specific similitude quantification. Following the production of a specific number of significantly distinct and dissimilar groups in the feature set, clustering techniques are effectively used to obtain summaries and visualize data (Jain et al., 1999). There have been breakthrough applications of clustering methods in everyday life problems involving consumer segmentation, gene expression data, document grouping and many more examples (Shaw et al., 2001; Chan, 2008; Liu et al., 2008; Liang, 2010). Overall, clustering techniques are useful in the following main ways:

1. Summarisation – derivation of a miniaturized representation of the full data set
2. Discovery – finding and identifying contemporary insights into the structure of a data set

There are other numerous uses such as investigation of the validity of pre-existing group assignments and as a precursor to prediction by either regression or classification. Clustering is categorised as an unsupervised learning type of machine learning, where the machine receives inputs but no desired targets (outputs) or rewards from the surroundings. Usually, the objective is establishing patterns in the data above and beyond what would be considered noise.

2.2 Cluster analysis in market segmentation

Clustering techniques and analysis has become a commonly employed tool in market research for development of empirical arrangements of people, commodities or instances which might perform as a foundation for advanced analysis (Punj and Stewart, 1983). The primary use of cluster analysis in market research is market segmentation. In their work, Punj and Stewart note that clustering techniques have a paramount role to play in market research by seeking a superior grasp of buyer behaviours by establishing homogenous subsets of consumers. Researchers such as (Smith, 1956; Claycamp and Massy, 1968; Moorthy, 1984)

describe market segmentation as a long-established strategy that has been broken down and justified in every business devoted handbook over the years.

All segmentation research, regardless of the method used, is used designed to identify groups of entities that share certain common characteristics (attitudes, purchase propensities, media habits etc.). Without the specific data used to arrive at these and the detailed layout of the scope and objectives of the research, segmentation is equivalent to a grouping exercise. The two researchers add that clustering techniques also have had and essential role to play in seeking improved comprehension of buyer behaviours by establishing homogeneous classes of consumers.

Over the years, clustering techniques have been used across a wide array of industries to segment an organization's customers. Brito et al, (2015) delved into two separate techniques for customer segmentation: subgroup discovery and clustering. The models obtained produced six market segments and forty-nine rules that provided an improved comprehension of customer preferences in a tremendously customized organization dealing with fashion manufacturing.

Jansen (2007) performs segmentation and subsequent profiling of Vodafone customers based on usage call behaviour. He utilizes several progressive clustering techniques that are adapted and activated for customer segment creation. An optimality yardstick is defined to measure the performance of each and the best clustering technique is used to perform customer segmentation. A description of each segment is provided and followed by analysed. Finally, the Support Vector Machines (SVM) algorithm is employed to provide an estimate the group in which a customer will fall into by utilizing the provided profile. Based on the SVM approach, it is conceivable to categorize the group of a customer using its profile for the four-segment scenario in 80.3% of the cases. An accurate classification of 78.5% is achieves for six distinct segments.

Ansari and Riasi (2016) used a combination of and genetic algorithm and Fuzzy-C means techniques to segment the steel market customers. The customers were grouped into two segments by using the LRFM (length, recency, frequency, monetary value) variables model. From the results, customers in the first segment had a greater trade recency, higher relationship length, as well as trade frequency. However, their monetary value was lower in comparison to the mean values for these parameters across the customer base.

2.3 Classification of Clustering

Clustering techniques are commonly divided into the following broad categories:

- Hierarchical clustering
- Partitioning clustering
- Density-based clustering

However, this classification cannot be either forthright, or entirely canonical. The classes overlap in reality. (Rai, Singh, 2010).

2.3.1 Hierarchical methods

This method provides for construction of a hierarchy of clusters by allowing clusters to have their own sub-clusters, forming a systematic sequence of clusters. Each leaf in the sequence, also known as tree, represents a data instance. This is the tiniest possible group. The node at the root on the other hand represents the group that contains every data object. This is the biggest cluster possible. Every internal node within the sequence is a group whose components are all the objects in the nodes of the child (union of the sub-clusters). Designating an end of a given level provides the ability to extract a collection of non-overlapping objects.

Partition takes place sequentially. This process could in the end cluster all the instances into one group on n groups of one instance each. A two-dimensional diagram is used to illustrate hierarchical clustering by showing the divisions or fusions formed at each successive level of the clustering process. This diagram is referred to as a dendrogram.

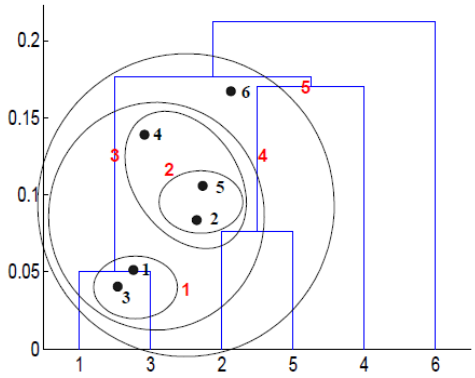

Figure 1: A Nested Cluster

Hierarchical methods are advantageous in that they provide embedded adaptability in as far as the extent of granularity and easily handle any typed of similarity or separation. They are also applicable to any attribute type, be it numeric or categorical. However, they tend to be vague when it comes to the termination criteria and most algorithms do not revisit preceding constructed clusters with the purpose of improvement.

2.3.2 Partitioning methods

These simply divide the objects/elements into a set of M groups, where each element has membership to one group. It is the most popular method. A unique centroid or cluster representative acts as the representative of each group. The centroid provides a near summary, if not a precise one, of the cluster objects. A precise characterization is dependent on the form of the object under consideration. In instances where the value of the data is available, the arithmetic mean of the variables for every object within a group gives a fitting representative. Whenever these values are unavailable, centroids in other forms may be needed.

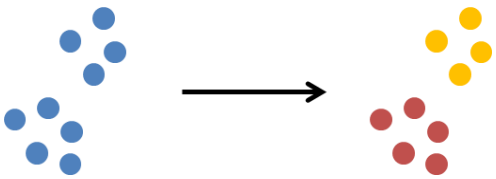

Figure 2: Partitioning Methods

2.4 Clustering Algorithms

2.4.1. K-means

This is an iterative technique whose objective is to minimize the sum of squares within a class for any number of clusters. The algorithm commences with the primary guess of centre for every cluster. Each instance is subsequently allocated in to a group to which it is most similar. This step is followed by updating the cluster centres, and the procedure is reiterated until the centres do not shift any more. An augmenting clustering technique such as hierarchical algorithm is normally applied at the onset to arrive at the cluster centre starting values.

2.4.2. DIANA

This is a divisive hierarchical approach in which every observation is placed in one cluster at the start. The algorithm subsequently divides the groups until each of them has a single observation. It is one of the few forms of the hierarchical type of clustering. Others in this approach include the Self Organizing Tree (SOTA).

2.4.3. K-Medoids

K-Medoids is not significantly different from K-means. It is, however, considered more potent owing to its ability to accommodate and utilize other measurements of dissimilarity besides the Euclidian distance. Just like the K-means technique of partitioning, the number of clusters is ordinarily defined initially. An accompanying batch of centres is necessary to initiate the algorithm.

The primary design of medoids clustering techniques is to establish K groups in n objects. This procedure begins with the arbitrary selection of a representative object per cluster. The rest of the objects are clustered with the medoid to which each is most similar. The K-medoids approach utilizes representative objects as the points of reference as opposed to using the mean values of each cluster. The approach takes the input parameter k, the number of groups to be partitioned among a collection of n objects.

2.4.4. CLARA

This algorithm is based on sampling and performs partitioning around medoids on several sub-groups of data (Kaufman and Rousseeuw, 1990). By using this sampling approach, run times are relatively brisk when there are many observations.

2.4.5. FANNY

This approach executes fuzzy clustering. Every object can be a partial member in every cluster. (Kaufman and Rousseeuw, 1990). Hence, there is a vector in each object that allows it to be partially a member of every one of the groups. A hard cluster is formed when every single object is assigned to the group in which it possesses the greatest membership.

2.4.6. SOM

The Self-organizing maps (SOM) technique has a firm foundation on neural networks. It is hugely considered for its capacity in mapping high-dimensional data and visualizing them to generate two dimensional depictions. According to early work by (Kohonen, 1995), SOM performs two types of data compression:

- reduction of data dimension with minimum loss of information. (These neural networks can single out sets of independent characteristics)
- reduction of data variety due to terminal composition prototypes separation. (Clustering and quantization of data sets)

2.4.7. Model-based clustering

Through this methodology, a finite combination of normal/gaussian distributions that form a statistical model is fit to the data. Every combination element serves as a cluster. The components for combinations and cluster memberships are estimated through maximum likelihood estimators (Fraley and Raftery 2001).

2.4.8. SOTA

The self-organizing tree algorithm (SOTA) is defined as an unsupervised technique whose binary tree structure has a divisive hierarchical.

2.5. Cluster Validation

A pertinent issue in clustering techniques is the assessment of outcomes from various algorithms to ratify the partitioning which optimally fits a certain data set (Halkidi et al., 2001). The assessment procedure must take on the following quantitatively expressed onerous questions:

i. The quality of clusters,
ii. The degree with which a clustering scheme fits a specific data set
iii. The optimal number of clusters in a partitioning.

Several methods have been put forward and tested for estimating the optimal number of clusters. The statistical elbow concept has been exploited by some. (Milligan and Cooper, 1985) summarized many of these approaches in the comprehensive survey. (Gordon, 1999) has also discussed in detail the best performers. More recent recommendations have come from (Tibshirani, Walther, and Hastie, 2001), (Sugar, 1998), and (Sugar, Lenert, and Olshen, 1999). Sufficient clarity is lacking, however, on if these approaches are extensively employed (Tibshirani, Walther).

(Guy et al. 2008), when developing an R package to perform cluster assessment, note that a wide range of criteria whose objective is to evaluate the results of a clustering procedure and establishing which technique provides the optimal categorization for a specific trial were recommended (Kerr and Churchill 2001; Yeung et al. 2001; Datta and Datta 2003). The substantiation is achieved through several ways. They proceed to provide three forms of validation – internal, stability and biological.

CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

3.1 Research Design

In this study, a quantitative methodology that uses applied research methods will be applied. Following a selection of the best clustering algorithm based on spending habits of a sample of consumers, profiling will be done for the segments and applicable descriptions. This will then be packaged as a product for use in the Kenyan market, replicable and reproducible in other similar markets. The data collected is based on a stratified design that is probabilistically proportional to the population of Kenya by age, gender, region and Living Standards Measures (LSM). The clustering algorithms will be run and tested using R-Gui.

3.2 Research Data

3.2.1 Mobile Data Collection in Kenya and other emerging markets

Over the years, adoption of cutting edge technologies to help research process has remained something that researchers are open to, without compromising on the integrity of the various aspects of the process.

In Kenya, mobile penetration currently stands at 88.7% (Communication Authority of Kenya, 2017). This provides for a wide-reaching means of data collection and engagement from people of all walks of life. mSurvey has been facilitating data collection through mobile phone conversations in Kenya since 2012. In 2016, the company, in collaboration with Safaricom, embarked on collecting spending data from a sample of 1,215 to map the cash economy through daily mobile conversations.

3.2.2 Source of Data

This research relies on primary data collected through daily mobile conversations. The panel answers daily questions about what products they spent on the previous day, how much they spent on each of the products and how they paid for it. The survey design and allocation is based on stratified and probability sampling. Through stratified sampling, partitioning of the population into non-overlapping groups is performed. The groups are known as strata. A random sample is then selected from each stratum by some design. The population of N sampling units (in this case people aged 18 and over, the legal age in Kenya) is divided into

H strata. The strata construction is based on the three characteristics that are known – Age, gender and region. Each stratum h has *Nh* sampling units. To guarantee that the sample distribution mirrors that of the population it is sampled from based on the three stratifying variables and the sample is a diminutive adaptation of the population, we make use of proportional allocation when designing the sample.

3.2.3 Volume of Data

This paper will use data from surveys with consumers, which has been collected since April 2007. The panel has been engaged daily to report their expenditure details. The data is then aggregated into weekly and monthly measures to estimate the average expenditure (wallet size), the share of the wallet based on the various categories, and the modes of payment. For this paper, data for nine months will be aggregated per respondent based on the 11 categories. This is sufficient to volumes for running the clustering algorithms and obtaining distinct segments for profiling.

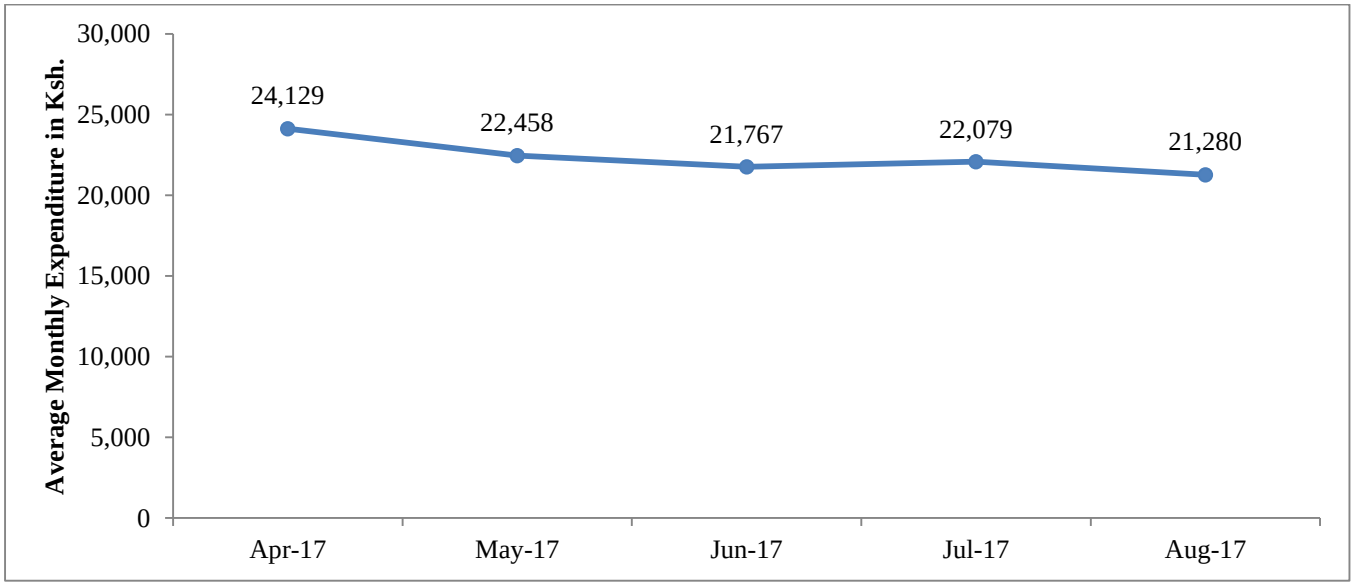

Figure 3: Average Monthly Expenditure in Kenya based on the first five month of Consumer Wallet

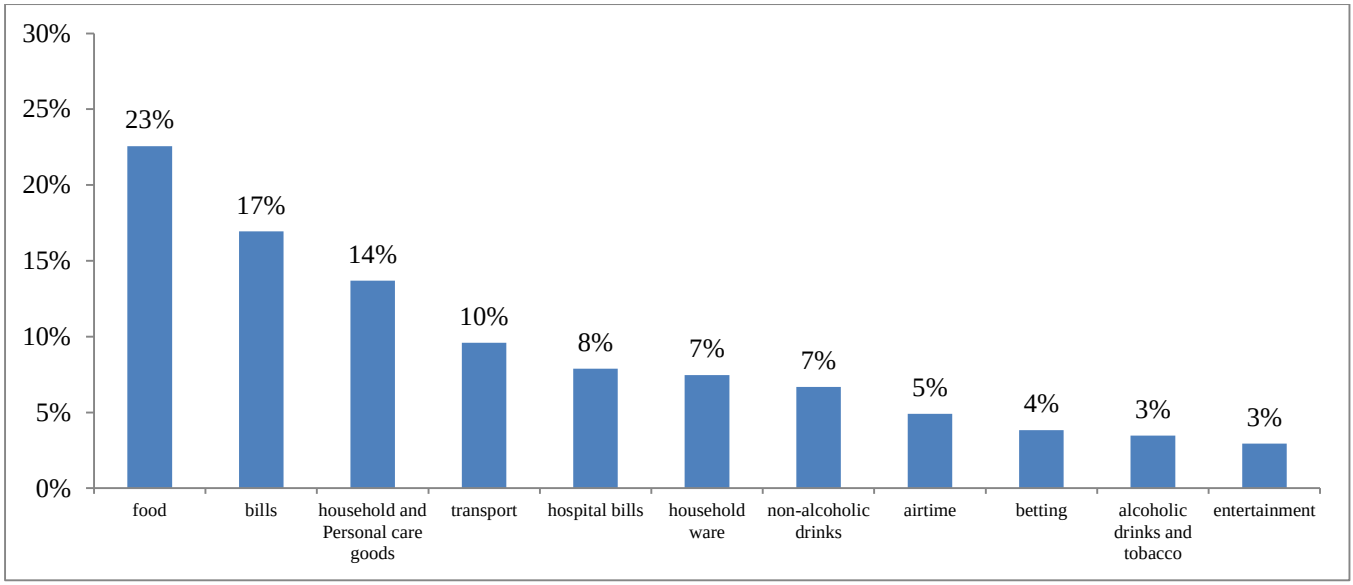

Figure 4: Share of consumer wallet in Kenya based on the first five month of Consumer Wallet

The clustering algorithms will be executed based on the aggregated average expenditure of each of the 1,215 consumers in the panel.

3.2.4 Consumer Demographics

To profile the consumers and construct profiles that form a solid anchor for product targeting, four demographic characteristics that are currently available will be used. With the following variables, the consumers' profiles can be broken down as:

Gender	Age Group	Region	LSM	Primary Expenditure Source
Male	18-29	Central	1-3	Borrowing
Female	30-39	Coast	4-6	Salary
	40-49	Eastern	7-9	Wages
	50 +	Nairobi	Above 9	Business profits
		North Eastern		
		Nyanza		
		Rift Valley		
		Western		

Table 1: Demographic characteristics for profiling

Using ungrouped age and LSM values or even relatively smaller class ranges would result into close relationships. The two variables thus must be grouped for the differences to be

significant enough. The table below shows the proportion of customers within the available demographic identifiers. Since three of these variables are also the stratification variables, the distribution across the groups in each variable is proportional to the population distribution in Kenya.

Gender	Female	Male						
	50.5%	49.5%						
Age Group	18-29	30-39	40-49	50 +				
	54.2%	24.6%	14.6%	6.7%				
Region	Central	Coast	Eastern	Nairobi	North Eastern	Nyanza	Rift Valley	Western
	13.4%	7.0%	12.2%	22.5%	5.3%	11.9%	18.8%	8.9%
LSM	1-3	4-6	7-9	10+				
	6.1%	37.2%	43.0%	13.7%				

Table 2: Proportional distribution of consumer wallet panel

With a profile built on these characteristics, a classification algorithm can be used to estimate the consumers' segment.

3.3 Unsupervised Learning

Unsupervised Learning is a subfield of Machine Learning, focusing on the study of mechanizing the process of learning without feedback or labels. Unsupervised learning models have no a-priori knowledge about the classes into which data can be placed. They use the features in the dataset to form groupings based on feature similarity.

3.3.1 Clustering

Clustering, as defined previously, is the process of partitioning a collection of observations into distinct groups so that the observations in each group as similar as possible to each other, relative to observations within other groups. These techniques are considered to be the most imperative of unsupervised learning methods. The approach does not utilize prior group identifiers of elemental patterns in a data set.

Simply, a cluster is thus described as a set of observations which have similarities between them and have dissimilarities with the observations in other partitions. The similarity measure

that was used in this case is distance. A different approach to partitioning a data set is conceptual clustering, where at least two observations are considered members of a group if one construes a concept typical to every other observation. Only distance-based techniques of clustering are used in this research.

Clustering techniques can be applied to numeric, categorical data, or a combination of the two. The clustering of numeric data (average monthly expenditure per category) is considered here. Each record of the consumer's expenditure by category is made up of of *n* collected values, organized into an *n*-dimensional row vector $x_k = [x_{k1},\ x_{k2},\ x_{k3}\ , \dots, x_{kn}]^T$, where $x_{kn} \in \mathbb{R}^n$. A set of N observations is denoted by $X = x_k | k = 1,2, \dots, N$ and is represented an $N \times n$ matrix:

$$X = \begin{matrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{matrix} \qquad (3.1)$$

The rows of the matrix represent individual consumers in the panel, whereas each column is the feature of their expenditure for each category under consideration. A given data set can reveal partitions of varying densities, geometrical shapes and sizes as shown in the figures below:

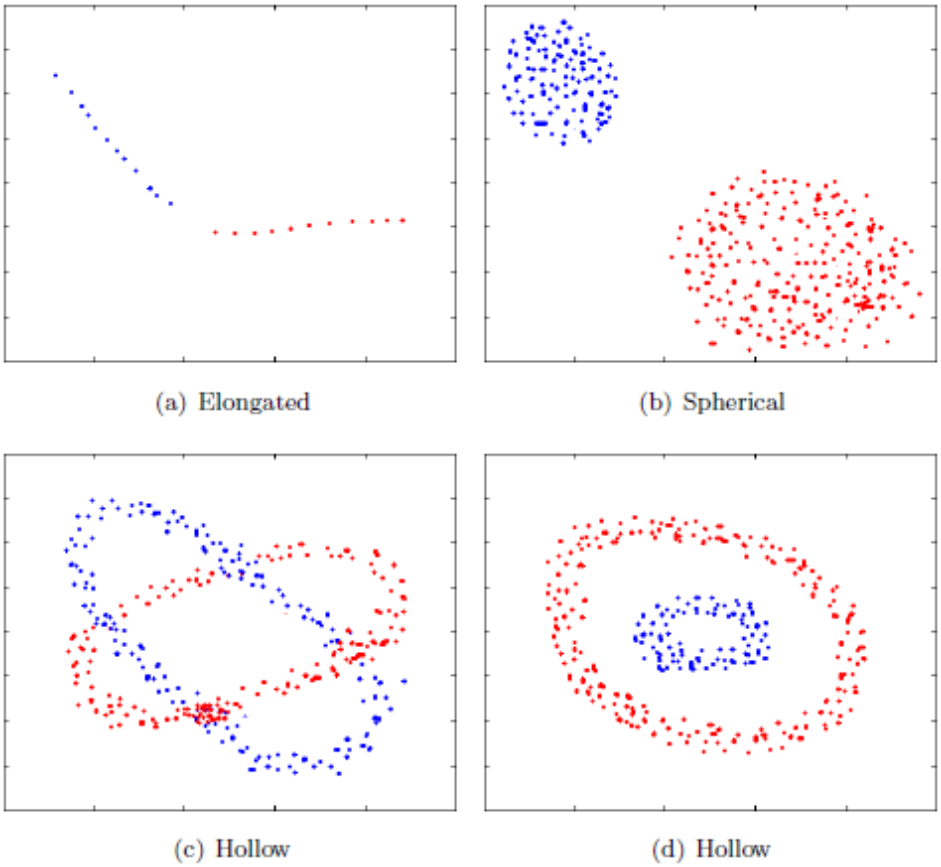

Figure 5: Possible Cluster Shapes

Clustering analysis techniques possess the unique ability to discover subspaces in a given data space. This renders them substantially dependable for identification. Groups arising from partitioning of the data space are categorized as either well-separated, continuously connected, or overlapping each other.

3.3.1.1 Cluster partitioning

Formally, clusters are primarily considered as subsets of a collection of observations. They can be classified and distinguished into two distinct categories:

- Fuzzy
- Crisp (Hard)

Crisp methods are founded on classical set theory. Every observation is required to either or not belong to a cluster.

a) Hard partition

The main purpose of partitioning in this case is grouping the data set X into c clusters. Using classical sets, a hard partition can be seen as a family of subsets$\{A_i|1 \leq i \leq c \subset P(X)\}$, its properties can be defined as follows:

$$\cup_{i=1}^{c} A_i = X_i \ , \tag{3.2}$$

$$A_i \cap A_j, 1 \leq i \neq j \leq c \ , \tag{3.3}$$

$$\phi \subset A_i, \subset X, \ 1 \leq 1 \leq c \tag{3.4}$$

Expressed in the terms of membership functions:

$$\vee_{i=1}^{c} \mu_{A_i} = 1, \tag{3.5}$$

$$\mu_{A_i} \vee \mu_{A_j}, \ 1 \leq i \neq j \leq c \ , \tag{3.6}$$

$$0 \leq \mu_{A_i} < 1, 1 \leq i \leq c \tag{3.7}$$

μ_{A_i} denotes the characteristic function of the subset A_i whose value is 0 or 1. Simplifying the notations, we use $\mu_i$instead of μ_{A_i} , and representing $\mu_i \ (x_k)$ by μ_{ik}, clusters can be denoted in a matrix notation. U = μ_{ik}, a NXc matrix, is a depiction of the hard partition if and only if its objects satisfy:

$$\mu_{ij} \in \{0,1\}, 1 \leq 1 \leq N, 1 \leq k \leq\leq c, \tag{3.8}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, 1 \leq i \leq N, \tag{3.9}$$

$$0 < \sum_{i=1}^{N} \mu_{ik} < N, 1 \leq k \leq c \tag{3.10}$$

In conclusion, let X be a finite data set and the number of clusters $2 \leq c < N \in N$. Then, the hard-partitioning space for X can be seen as the set:

$$M_{hc} = \{U \in R^{N_{xc}} | \mu_{ik} \in \{0,1\}, \forall i,k;\ \sum_{k=1}^{c} \mu_{ik} = 1, \forall i; 0 < \sum_{i=1}^{N} \mu_{ik} < N, \forall k\}, \quad (3.11)$$

b) Fuzzy partition

Consider the matrix $U = \mu_{ik}$, with the below conditions:

$$\mu_{ij} \in [0,1], 1 \leq i \leq N, 1 \leq k \leq c, \quad (3.12)$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, 1 \leq i \leq N, \quad (3.13)$$

$$0 < \sum_{i=1}^{N} \mu_{ik} < N, 1 \leq k \leq c \quad (3.14)$$

Let X be a data set which is finite and the cluster number $2 \leq c < N \in N$. It follows that, the fuzzy partitioning space for X is seen as the set:

$$M_{fc} = \{U \in R^{N_{xc}} | \mu_{ik} \in \{0,1\}, \forall i,k;\ \sum_{k=1}^{c} \mu_{ik} = 1, \forall i; 0 < \sum_{i=1}^{N} \mu_{ik} < N, \forall k\} \quad (3.15)$$

The $i-th$ column of U contains values of the membership functions of the $i-th$ fuzzy subset of X.

3.3.2. K-Means

The idea behind K-means clustering is to reduce the within-cluster variation to the smallest possible value. The technique distributes every object in the collection to a single set of the c groups to minimize the sum of squares within each cluster:

$$\sum_{i=1}^{c} sum_{k \in A_i} || x_k - v_i ||_2 \quad (3.16)$$

A_i stands for a collection of observations in the i-th group and v_i is the mean of the observations in group i. $|| x_k - v_i ||_2$ is a selected distance measure. v_i is the centre of cluster i:

$$v_i = \frac{\sum_{k=1}^{N_i} x_k}{N_i}, x_k \in A_i, \quad (3.17)$$

Where N_i is the number of points in A_i.

Some properties of K-means include:

- Within cluster variation decreases with each iteration of the algorithm
- The algorithm always converges, despite the initial cluster centres
- The ultimate clustering depends on the first cluster centres. Sometimes, various initial centres yield different final outputs. The algorithm is run multiple times with random initialization of cluster centres for each round, then choosing from the set of centres dependent on the one that provides the smallest within-cluster variation
- The algorithm is not guaranteed to deliver the clustering that globally minimizes within-cluster variation

3.3.3. K-Medoids

In some instances, we want each of the centres to the point itself. This algorithm is similar to K-Means in its calculations, only difference being that when fitting the centres, we restrict our attention to the points themselves.

The initial guess for the centres is $v_1, v_2, \dots, v_k$, then repeat.

1. Minimize over N; for each $i = 1,2, \dots, c$, find the cluster centre v_i closest to x_i
2. Minimize over $v_1, v_2, \dots, v_k$, ; for each $i = 1,2, \dots, c$, let $v_i = x_i^*$, the medoid of of points in cluster i that minimizes

   $\sum_{i=1}^{c} sum_{k \in A_i} || x_k - x_i ||_2$, (3.18)

   Stop when within-cluster variation does not change.

This algorithm shares the same properties as he K-means algorithm. It is computationally harder since computing the medoid is harder than computing the average, but it has the potentially important property that the centres are located among the points themselves. As a result, this algorithm performs more robustly when outliers and noise exist since the medoid is not as affected by other severe values compared to an average. The main drawback, however, is that this technique works adequately for relatively small sets of data but does not scale well for large ones.

3.3.4. Fuzzy C-means

Fuzzy set theory was initially submitted by Zadeh in 1965. It gave an idea of uncertainty of belonging which was described by a membership function. The Fuzzy C-means technique, which focuses on reducing an objective function called the C-means function is denoted as below:

$$J(X;U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m \, ||x_k - v_i||_A^2 \, , \tag{3.19}$$

With

$$V = [v_1, v_2, v_3, \dots, v_c], v_i \in R^n \, , \tag{3.20}$$

V represents the vector with the partition centres that must be established. The distance measure,

$||x_k - v_i||_A^2$ is known as a squared inner-product distance norm and is defined by:

$$D_{ikA}^2 = ||k_k - v_i||_A^2 = (x_k + v_i)^T A((x_k + v_i) \, , \tag{3.21}$$

Using Lagrange multipliers to establish the stationary points yields:

$$J(X;U,V,\lambda) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m \, D_{ikA}^2 + \sum_{k=1}^{N}\lambda_k \, (\sum_{i=1}^{c}\mu_{ik} - 1), \tag{3.22}$$

and by setting the gradients of $\hat{f}$, with respect to, U, V and λ to zero.

The algorithm works as follows when implemented:

1. Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$
2. At k-step, calculate the centre vectors $V^{(k)} = [v_j]$ with $U^{(k)}$

   $$V_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_i}{\sum_{j=1}^{N} u_{ij}^m}, \tag{3.23}$$
3. Update $U^{(k)}, U^{(k+1)}$

4. $$d_{ij} = \sqrt{\sum_{i=1}^{c}(x_i - v_i)} \tag{3.24}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{N}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}, \quad (3.25)$$

5. If $||U(k+1) - U(k)|| < \varepsilon$, then stop; otherwise, return to step 2.

This algorithm works by assigning membership to each data point corresponding to each cluster centre based on distance between the cluster centre and the individual point. The closer the distance between the point and the cluster centre, the more the point is assigned membership to the particular cluster. Adding up the membership of each element therefore results to one. After each iteration, cluster centres and membership are updated.

FCM is advantageous in that it converges, but has limitations owing to the long computation time, sensitivity to the initial guess (speed and local minima) and sensitivity to noise and outliers.

3.3.4. The Gustafson-Kessel algorithm

This technique is a variation on the Fuzzy c-means algorithm (Gustafson, Kessel, 1979). It applies a distinct and flexible measure of distance to identify geometrical shapes in the data. Every group in the data will possess a separate norm-inducing matrix. The matrix will satisfy the inner-product rule below:

$$D_{ikA}^2 = (x_k - v_i)^T . A_i (x_k - v_i), \text{ where } 1 \leq i \leq c \text{ and } 1 \leq k \leq N \quad (3.26)$$

For this algorithm, the objective function is computed by:

$$J(X; U, V, A) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D_{ikA_i}^2 \quad (3.27)$$

If we vary Ai to optimize the clusters while fixing the volume:

$$\left||A_i|\right| = \rho_i, \rho > 0 \quad (3.28)$$

ρ is the remaining constant for each cluster. Combining this with the Lagrange multiplier, A_i can be conveyed as follows:

$$A_i = [\rho_i \det(F_i)]^{1/n} F_i^{-1} , \quad (3.29)$$

With

$$F_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m (x_i - v_i)(x_i - v_i)^T}{\sum_{k=1}^{N} (\mu_{ik})^m} \quad (3.30)$$

3.4. Validation

Unlike in supervised learning where we a variety of measures to evaluate how good our model is, it is not as straightforward in unsupervised learning. It is necessary to evaluate clustering results to refrain from the possible peril of discovering patterns in noise, to analyse different algorithms, and to compare any two sets of clusters. The resulting groups are supposed to have robust statistical characteristics (compact, well-separated, connected, and stable) in an ideal situation, as well as provide results that are admissible to make better marketing strategy. Multiple methods for determining the optimal cluster number and validating the algorithms have been proposed. In their paper, (Guy et al. 2008) offer a package in the R statistical computing environment, clValid, which consists of an assortment of approaches for validation of cluster analysis outcomes. The presented measures are classified as below:

- Internal
- Stability
- Biological

We can choose and compare multiple algorithms through different validation measures, determine the optimal number of clusters for a given set of data. In this paper, we will use internal and stability measures for validation.

3.4.1. Internal Measures

These measures, the extent to which a group partitions are compact, connect and separate is validated.

a. Connectivity

Let $nn_{i(j)}$ be the jth nearest neighbour of observation i, and let $x_{i,nn_{i(j)}}$ be zero if i and $nn_{i(j)}$are in the same cluster and $1/j$ otherwise. For a particular clustering partition, $\mathbb{C} = \{C_1, C_2, \dots, C_k\}$ of the N observations into K disjoint clusters, the connectivity is given by:

$$conn(\mathbb{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}} \quad (3.31)$$

b. Silhouette width

Silhouette width is defined as the average of each observation's silhouette value, where the silhouette value measures the degree of confidence in the clustering assignment of an observation, with well-clustered observations having values near 1 and poorly clustered observations having values near - 1. For observation *i*, it is defined as:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (3.32)$$

Where a_i is the average distance between *i* and all the other observations in the same cluster and b_i is the average distance between *i* and the observations in the 'nearest neighbour cluster' i.e.

$$a_i = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist(i,j), \; b_i = \min_{C_k \in \mathbb{C} \setminus C(i)} \sum_{j \in C(i)} \frac{dist(i,j)}{n(C_k)} \quad (3.33)$$

where *C(i)* is the cluster containing observation *i*, *dist(i, j)* is the distance (e.g. Euclidean, Manhattan) between observations *i* and *j*, and *n(C)* is the cardinality of cluster *C*. The silhouette width thus lies in the interval [-1, 1], and should be maximized.

c. Dunn index

This index is defined as the ration of the tiniest separation of objects that are not in the same cluster to the biggest intra-cluster distance. It is computed as:

$$D(\mathbb{C}) = \frac{\min_{C_k, C_1 \in \mathbb{C}, C_k \neq C_1} \left(\min_{i \in C_k, j \in C_1} dist(i,j) \right)}{\max_{C_m \in \mathbb{C}} diam(C_m)} \quad (3.34)$$

This index ranges from 0 to ∞ and maximization is the target.

3.4.2. Stability Measures

a. Average proportion of non-overlap (APN)

Let $C^{i,0}$represent the cluster containing observation i using the original clustering (based on all available data), and $C^{i,\ell}$ represent the cluster containing observation *i* where the clustering is based on the dataset with column ℓ removed. Then, the APN measure is defined as:

$$APN(\mathbb{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{\ell=1}^{M}\left(1 - \frac{n(C^{i,\ell} \cap C^{i,0})}{n(C^{i,0})}\right) \quad (3.35)$$

b. Average distance (AD)

The AD measure computes the average distance between observations placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed. It is defined as:

$$AD(\mathbb{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{\ell=1}^{M}\frac{1}{n(C^{i,0})n(C^{i,\ell})}\left[\sum_{i\in C^{i,0}, j\in C^{i,\ell}} dist(i,j)\right] \quad (3.36)$$

c. Average distance between means (ADM)

This measure calculates the mean distance between centres for objects allocated to the same group by clustering for the complete data set and clustering for the same data set when a single column is omitted. It is defined as:

$$ADM(\mathbb{C}) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{\ell=1}^{M} dist(\bar{x}_{C^{i,\ell}}, \bar{x}_{C^{i,0}}) \quad (3.37)$$

where $\bar{x}_{C^{i,0}}$ is the mean of the observations in the cluster which contains observation *i*, when clustering is based on the full data, and $\bar{x}_{C^{i,\ell}}$ is similarly defined.

d. Figure of merit (FOM)

The FOM measures the average intra-cluster variance of the observations in the deleted column, where the clustering is based on the remaining (undeleted) samples. For a particular left-out column ℓ, FOM is calculated as follows:

$$FOM(\ell, \mathbb{C}) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(\ell)} dist(x_{i,\ell}, \bar{x}_{C_k(\ell)},} \quad (3.38)$$

An average of the final score is obtained across all the omitted columns. The average is a value between 0 and 1. A low value is an indication of superior performance.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction

This chapter presented the results of the study based on the research objectives as defined in the first chapter. At this stage, the tool was designed and tested iteratively with the variables evaluated at each stage. The results were then observed. The various algorithms are run and compared with an aim of increasing accuracy while reducing complexity as much as possible. Cluster analysis aims at grouping data points using only information that is found in the data set providing a description the objects and their relationships. The objective is for points in a group be as similar to each other and as dissimilar to points in other groups as possible (Velmurugan, 2012). Performance of the algorithms is therefore a measure of how well they can split the data into these dissimilar groups, and in this chapter, we compare the ability of various clustering algorithms to partition expenditure data and uncover some inherent structure.

4.2 Evaluation and comparison criteria

There currently exists numerous clustering algorithms for customer segmentation. With the increased need to understand customers, data mining has become an integral piece of any customer relationship management (CRM) system design. K-Means, PAM and hierarchical clustering algorithms were compared using internal and stability validation and then an aggregate measure. The best algorithm is then fit to the data and profiling data.

4.2.1 Internal validation

	CLUSTERS	4	5	6	7	8	9	10	11	12
hierarchical	Connectivity	0.847	2.338	5.100	5.100	7.844	11.541	11.541	12.087	15.762
	Dunn	0.095	0.136	0.063	0.092	0.107	0.072	0.082	0.130	0.136
	Silhouette	0.924	0.918	0.881	0.882	0.880	0.883	0.869	0.870	0.870
k-means	Connectivity	4.587	16.141	14.746	11.632	12.903	11.612	11.541	12.087	15.762
	Dunn	0.100	0.014	0.020	0.041	0.058	0.058	0.082	0.130	0.136
	Silhouette	0.919	0.917	0.882	0.885	0.885	0.884	0.869	0.870	0.870
Pam	Connectivity	8.516	11.649	9.286	16.612	16.612	16.612	13.498	14.044	14.044
	Dunn	0.024	0.003	0.008	0.012	0.011	0.011	0.012	0.013	0.013
	Silhouette	0.918	0.517	0.708	0.714	0.799	0.802	0.808	0.806	0.812

Table 3: Comparison of clustering algorithms by Internal Validation

	Score	Method	Clusters
Connectivity	0.8468	hierarchical	4
Dunn	0.1364	hierarchical	12
Silhouette	0.9235	hierarchical	4

Table 4: Optimal scores from Internal Validation

Hierarchical clustering with four clusters achieves the best performance in two of the cases, Connectivity and Silhouette width. There results of internal validation are also visualized below.

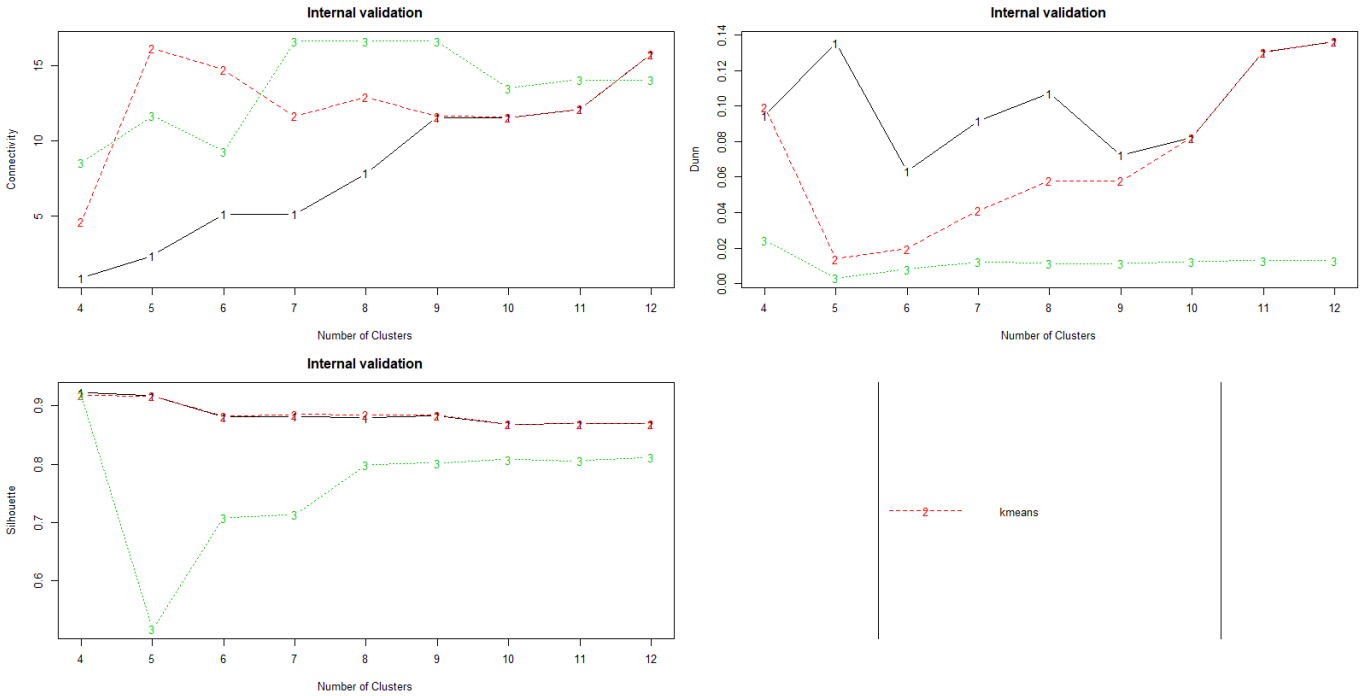

Figure 6: Plots of the connectivity measure, the Dunn Index, and the Silhouette Width

For relatively accurate partitioning, we aim at minimizing connectivity, while maximizing both the Dunn Index and the Silhouette Width. Hierarchical clustering (UPGMA) outperformed the other partitioning techniques under two validation measures. However, the optimal number of clusters was not as straightforward as the Dunn Index was maximized by twelve clusters and not four.

4.2.2 Stability validation

The measures of stability used here include the APN, AD, ADM, and FOM. The goal is to minimize all of them.

		4	5	6	7	8	9	10	11	12
hierarchical	APN	0	0	0	0	0	0	0	0	0
	AD	221,916	192,317	145,960	126,140	121,433	113,115	106,657	101,117	100,343
	ADM	0	0	0	0	0	0	0	0	0
	FOM	38,962	30,948	22,903	18,260	17,115	15,033	13,745	12,487	12,345
k-means	APN	0	0	0	0	0	0	0	0	0
	AD	222,173	187,602	143,381	124,251	118,341	113,184	106,657	101,117	100,343
	ADM	0	0	0	0	0	0	0	0	0
	FOM	37,902	29,991	22,065	17,734	16,111	15,018	13,745	12,487	12,345
pam	APN	0	0	0	0	0	0	0	0	0
	AD	221,884	183,042	139,941	107,438	88,923	69,743	57,682	51,661	45,202
	ADM	0	0	0	0	0	0	0	0	0
	FOM	37,935	36,571	30,845	20,838	20,232	15,505	13,686	12,049	10,391

Table 5: Comparison of clustering algorithms by Stability Validation

	Score	Method	Clusters
APN	0	hierarchical	4
AD	45,202	pam	12
ADM	0	hierarchical	4
FOM	10,391	pam	12

Table 6: Optimal scores from Stability Validation

A comparison of the algorithms by stability measures produced a tie between a hierarchical algorithm with 4 clusters and Partitioning around the Median clustering with 12 clusters.

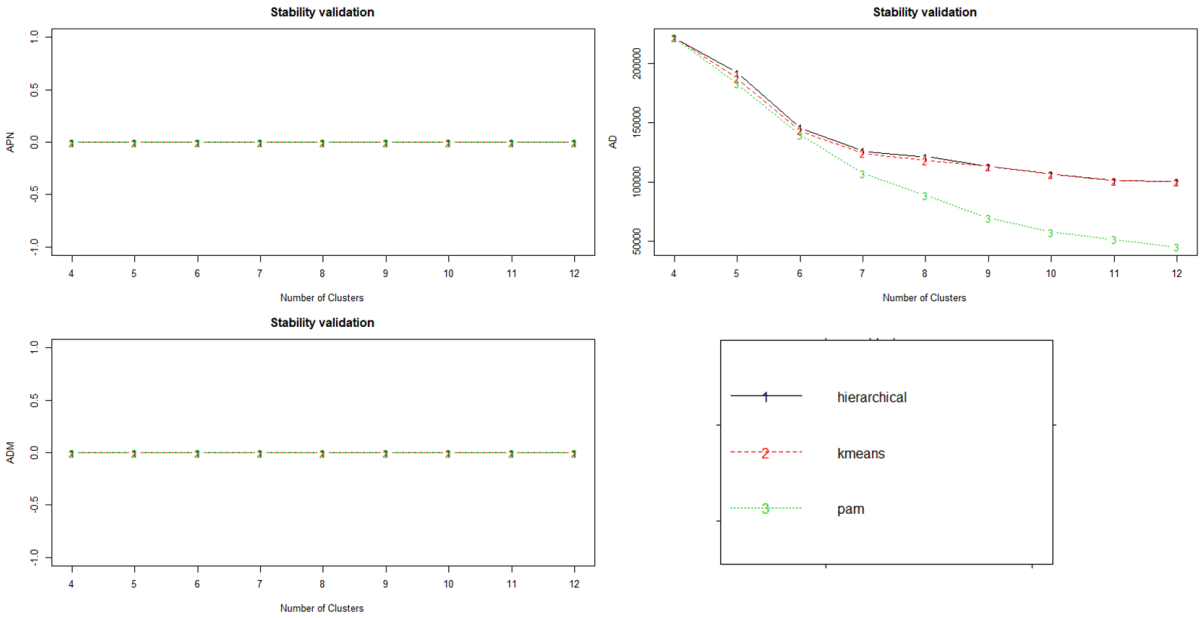

Figure 7: Plot of the APN, AD, and APN measures.

4.2.3 Rank aggregation

The order of the best clustering algorithms for the expenditure data was not same in the two validation measures. They however provided information for each measure to help in understanding what each is good at. The overall winner was then determined from a rank provided by an aggregate validation measure which uses the above measures simultaneously. The rank aggregation reconciles the ranks, producing a super-list (Brock et al., 2008). This stems from an idea that was suggested in the context of cluster analysis by (Pihur et al., 2007).

A combination of both internal and stability validation measures was used to rank the three algorithms with four to twelve clusters. The best three clustering algorithms for each of the validation measures are as follows:

	1	**2**	**3**
APN	hierarchical-4	hierarchical-5	hierarchical-6
AD	pam-12	pam-11	pam-10
ADM	hierarchical-4	hierarchical-5	hierarchical-6
FOM	pam-12	pam-11	hierarchical-12
Connectivity	hierarchical-4	hierarchical-5	kmeans-4
Dunn	hierarchical-12	kmeans-12	hierarchical-5
Silhouette	hierarchical-4	kmeans-4	pam-4

Table 7: Top three algorithms and cluster numbers

The results from the individual measures are confirmed here. Hierarchical clustering with four clusters performed best on four of the seven measures. Rank aggregation was used to establish the best five algorithms and the accompanying number of clusters.

1. Hierarchical - 4
2. Hierarchical - 5
3. PAM – 12
4. PAM – 11
5. K-Means - 12

Algorithm: CE

Distance: Spearman

Convergence was attained in 15 iterations, with a minimum objective function score of 5.766164.

Below are plots of the convergence properties and the ultimate performance measures:

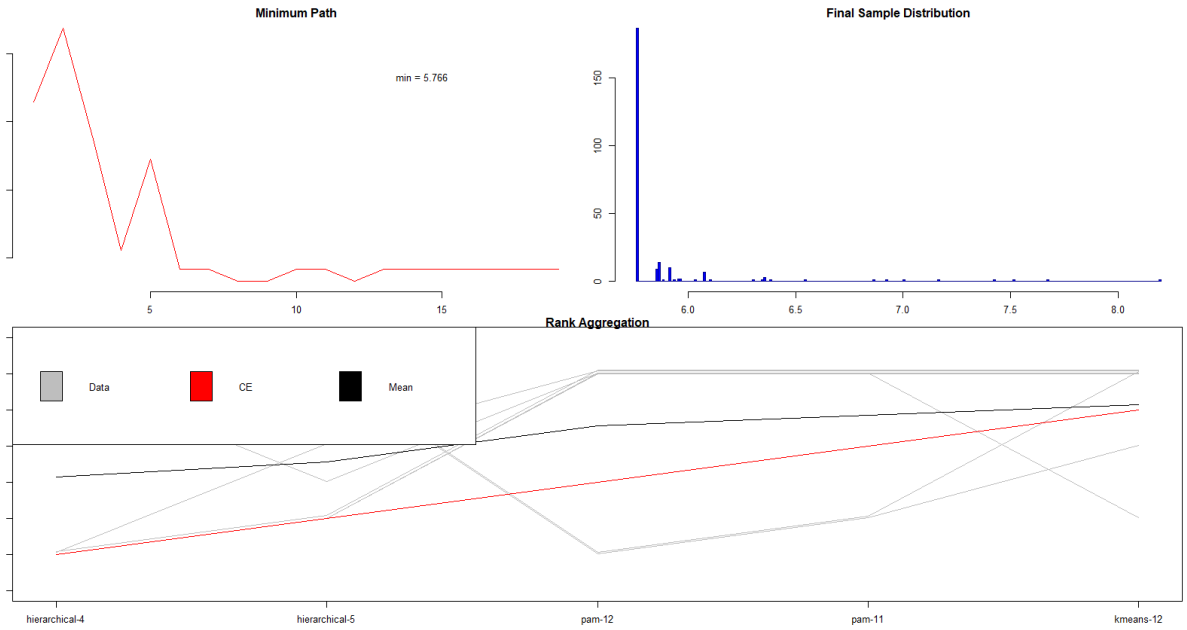

Figure 8: Optimal algorithms for clustering expenditure data

4.2 Fitting the Hierarchical Clustering algorithm

Based on the outcomes of the evaluation process, the best algorithm with the optimal number of clusters was fit into the data.

Hierarchical clustering can be classified into two principal forms: agglomerative Nesting (AGNES) and divisive Analysis (DIANA). The former partitions bottom-up, with the latter using a top-down approach.

Agglomerative clustering is great at establishing tiny groups of data, whereas divisive hierarchical clustering performs better at establishing bigger groups.

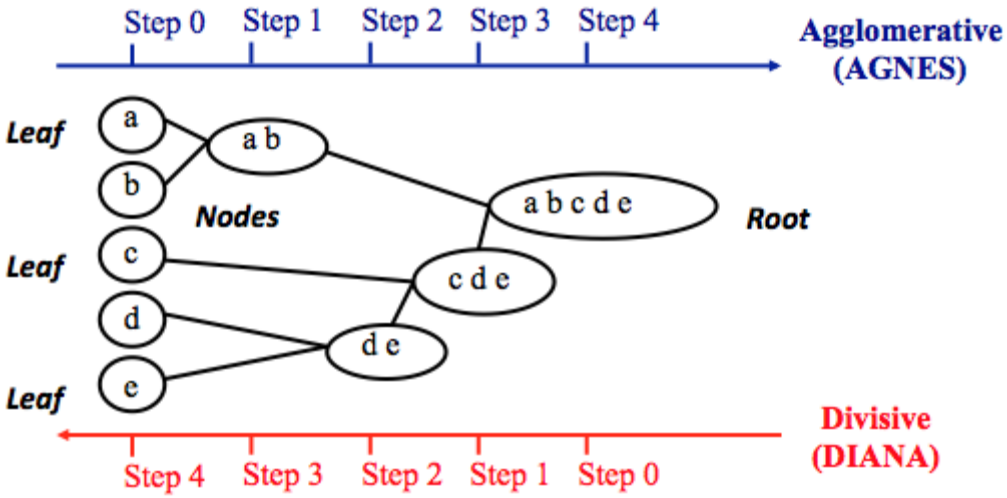

Figure 9: Agglomerative Vs. Divisive Hierarchical clustering approaches

In the two approaches used, the measure of dissimilarity between two clusters of observations was used to establish the clusters. Several different cluster agglomeration methods (i.e. linkage methods) have been developed and the most common types methods include minimum or single linkage clustering, Mean or average linkage clustering, Maximum or complete linkage clustering, Centroid linkage clustering, and ward's minimum variance method.

Using AGNES, however, the agglomerative coefficient was calculated. This coefficient measures the amount of clustering structure found. Ward's method identified the most robust structure in the data for the four techniques assessed. Below are the coefficients from the assessment of 4 methods:

Average	0.9998688
Single	0.9991312
Complete	0.9999476
Ward	0.9999943

Table 8: Agglomerative coefficients

The final clusters obtained from agglomerative nesting hierarchical clustering with four clusters for the expenditure data from 1,787 consumers are visualized below:

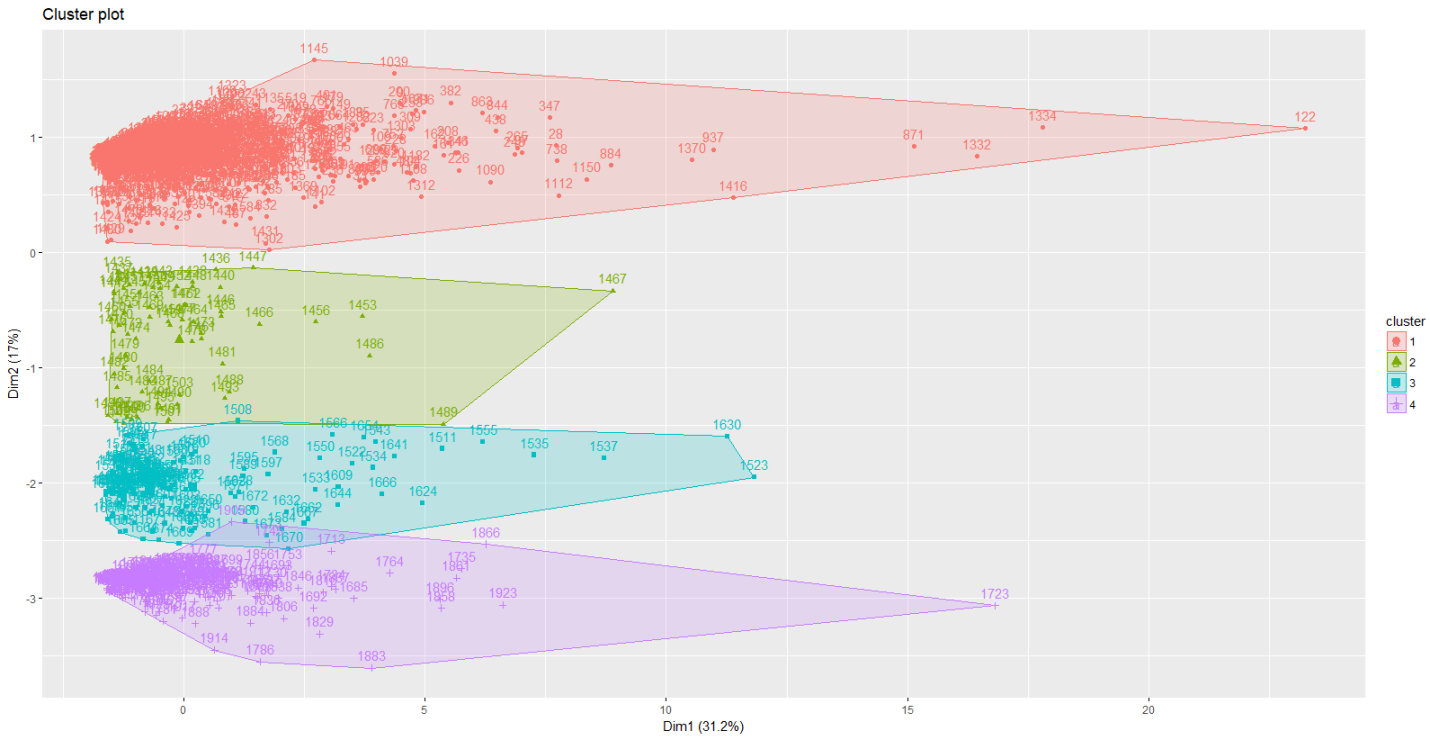

Figure 10: Clusters of consumers based on average expenditure across 11 categories

4.3 Profiling

The clusters obtained were described using the following available characteristics in addition to the spending habits:

- Age
- Gender
- Region
- Primary source

Segment 1

This segment is mainly composed of young consumers between eighteen and twenty-nine years of age. These consumers' spending habits revolve around short term needs such as food, airtime and alcoholic and non-alcoholic drinks. They have the lowest average expenditure, and 40% of it is either borrowed or provided by relatives.

Segment 2

This segment is relatively stable in terms of the source of finances. 42% of what these consumers spend comes from salary and wage savings, with another 25% coming from business profits and daily wages. The concentration of this segment is also in urban areas. There is a significant concentration for these consumers between twenty-five and fifty years. A significant proportion of their expenditures goes towards paying bills.

Segment 3

This segment consists of consumers below forty years. The source of their finances for daily expenditures is predominantly salaries and loans. On average, they spend more than any other consumer segment. There are slightly more male than female consumers in this segment, and they spend significantly more on entertainment and betting and have a proclivity towards alcoholic and non-alcoholic drinks, almost as much as the first segment. The proportion of their expenditure that goes towards food is significantly lower than all the other consumer segments.

Segment 4

More than half of the consumers in this segment are between thirty and fifty years of age. They are also relatively stable in that more than 75% of the expenditure comes from salary

savings and business profits. There are slightly more female than males in this segment, and their expenditure patterns seem focused on the household. They spend the least on entertainment, betting and alcoholic drinks, with the biggest piece (58%) going into food and household and personal care items.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

Segmenting and profiling consumers for better targeting, being an imperative focus for all consumer facing organizations, continues to evolve in approaches. The goal is to group the market into groups that are as homogeneous as possible, yet simple to understand and target. Traditional demographic traits no longer say enough to serve a as a basis for product and marketing strategy. Sound strategy depends on identifying segments that are potentially receptive to a product and brand category (Yankelovich and Meer, 2006). This paper used expenditure data in Kenya to identify the algorithm that best segments the market and then provided profiles for the segments based on available descriptors. The main challenge remains the availability of sufficient data to both segment as well as provide better segment descriptors to help organizations make better brand strategies.

Based on the findings of this study, it was concluded that expenditure data for eleven categories collected through daily mobile phone conversations with a sample of Kenya consumers provides a solid foundation for segmenting the market. The data consists of expenditure on eleven categories that are considered to constitute the significant proportion of the consumption in Kenya. Each of these expenditure data points is used as a variable in the comparison of various clustering algorithms to identify which best segments the consumers. Hierarchical, K-means and Partitioning around medoids (K-Medoid) clustering algorithms are compared based on internal and stability measures. Each of these is iterated across several pre-defined cluster sizes. Internal measures evaluate the compactness, connectedness and separation of the cluster partitions, while stability measures evaluate the results of clustering based on the full data and with one variable removed. Rank aggregation combined the two validation measures to determine the winning algorithm and corresponding optimal number of clusters. Hierarchical clustering with four clusters emerged best suited for this data. Using an

agglomerative approach to hierarchical clustering, the consumer data was segmented into four clusters with the minimum possible total within-cluster variance as measured by Ward's minimum variance method. These clusters were then described based on the available demographic data to provide profiles that can then be used by organizations to target brands and measure reception based on consumer expenditure.

5.2 Challenges and Limitation

The following are the challenges faced during the research project:

- Data quality – The research study is based on aggregate expenditure data obtained from daily surveys done on mobile. Being self-reported, there were instances of outlier and patterned records that needed detection and cleaning. Missing data also posed a challenge, and for these, incomplete records were omitted from the estimation of average individual expenditure.
- Data availability – despite the corporations that own the data making it available for the study, there was not sufficient profile characteristics for the consumers. The profile descriptors were therefore based on the few available variables, and there remains a huge opportunity to use other characteristics to not only provide rigorous and easily targetable profiles, but also for the classification purposes.

5.3 Recommendations

The researcher recommends that expenditure data be used for segmenting consumers for marketing and various brands. As opposed to looking at consumption patterns unilaterally based on purchases of one organization's products, leveraging on available data to construct segments based on cross-category expenditure provides a robust way of consumer

understanding. This data is available in Kenya and can be collected in numerous other ways, even with less frequency to start with. It is also recommended that as many demographic characteristics as possible be collected for each consumer to deepen the knowledge of each segment, thereby making marketing and brand strategy easier.

REFERENCES

1 Chin-Feng Lin, (2002) "Segmenting customer brand preference: demographic or psychographic", Journal of Product & Brand Management, Vol. 11 Issue: 4, pp.249-268, https://doi.org/10.1108/10610420210435443

2 OZER, M. (2001) User segmentation of online music services using fuzzy clustering, Omega, 29, 193–206.

3 ANDERSON, E.W., C. FORNELL and S.K. MAZVANCHERYL (2004) Customer satisfaction and shareholder value, Journal of Marketing, 68, 172–185.

4 WHITE, C. and Y.T. YU (2005) Satisfaction emotions and consumer behavioural intentions, Journal of Services Marketing, 19, 411–420.

5 CHANG, H.H. and P.W. KU (2009) Implementation of relationship quality for CRM performance: acquisition of BPR and organisational learning, Total Quality Management & Business Excellence, 20, 327–348.

6 SHAW, M.J., C. SUBRAMANIAM, G.W. TAN and M.E. WELGE (2001) Knowledge management and data mining for marketing, Decision Support Systems, 31, 127–137.

7 Bailey, C.; Baines, P.; Wilson, H. and Clarke, M. (2009), "Segmentation and customer insight in contemporary services marketing practice: why grouping customers is no longer enough", Journal of Marketing Management, Vol.25, No.3/4, pp.227-252.

8 Dibb, S. and Simkin, L. (1997), "A program for implementing market segmentation", Journal of Business and Industrial Marketing, Vol.12, No.1, pp.51-66.

9 Dibb, S. and Wensley, R. (2002), "Segmentation analysis for industrial markets: problems of integrating customer requirements into operations strategy", European Journal of Marketing, Vol.36, No.1/2, pp.231-251.

10 Laiderman, J. (2005), "A structured approach to B2B segmentation", Database Marketing and Customer Strategy Management, Vol.13, No.1, pp.64.75.

11 McDonald, M. and Dunbar, I. (2005) Market segmentation. Butterworth Heinemann, Oxford.

12 JAIN, A.K., M.N. MURTY and P.J. FLYNN (1999) Data clustering: a review, ACM Computing Surveys (CSUR), 31, 264–323.

13 LIANG, Y.H. (2010) Integration of data mining technologies to analyze customer value for the automotive maintenance industry, Expert Systems with Applications, 37, 7489–7496.

14 Central Bank of Kenya, Kenya National Bureau of Statistics & FSD Kenya. (2016). The 2016 FinAccess Household Survey on financial inclusion. Nairobi, Kenya: FSD Kenya.

15 Mattison, R., Data Warehousing and Data Mining for Telecommunications. Boston, London: Artech House, (1997).

16 Weiss, G.M., Data Mining in Telecommunications. The Data Mining and Knowledge Discovery Handbook (2005), pp. 1189-1201.

17 M.S.Yang," A Survey of fuzzy clustering" Mathl. Computer. Modelling Vol. 18, No. 11, pp. 1-16, 1993.

18 Handl J, Knowles J, Kell DB (2005). Computational Cluster Validation in Post-Genomic Data Analysis." Bioinformatics, 21(15), 3201-12.

19 Dunn JC (1974). \Well Separated Clusters and Fuzzy Partitions." Journal on Cybernetics, 4, 95-104.

20 Kaufman L, Rousseeuw PJ (1990). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

21 Datta S, Datta S (2006). \Methods for Evaluating Clustering Algorithms for Gene Expression Data using a Reference Set of Functional Classes." BMC Bioinformatics, 7, 397.

22 Yeung KY, Haynor DR, Ruzzo WL (2001). Validating Clustering for Gene Expression Data."Bioinformatics, 17(4), 309-18.

23 Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001) On Clustering Validation Techniques. Intelligent Information Systems Journal, 17(2-3): 107-145.

24 Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002) Cluster Validity Methods: Part II. SIGMOD Record, September 2002.

25 Punj, G., & Stewart, D. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. Journal of Marketing Research, 20(2), 134-148. oi:10.2307/3151680

26 Sekula, Michael N., "OptCluster : an R package for determining the optimal clustering algorithm and optimal number of clusters." (2015). Electronic Theses and Dissertations. Paper 2147.

27 Pedro QuelhasBrito, CarlosSoares, SérgioAlmeida, Ana Monte, Michel Byvoet (2015) Customer segmentation in a large database of an online customized fashion business 36, 93-100.

28 Kohonen, T. (1995) The Self-Organizing Maps. Springer, Berlin

29 W. Banzhafl, "Self-organizing systems", in Encyclopedia of Complexity and Systems Science, 2009, Springer, Heidelberg,

30 A. M. Turing, "The chemical basis of morphogenesis", Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, Vol. 237, No.641. pp.37-72, Aug. 14, 1952

31 W. R. Ashby, "Principles of the self-organizing system", E:CO Special Double Issue Vol. 6, No. 1-2, pp. 102-126, 2004

32 C. Fuchs, "Self-organizing system", in Encyclopaedia of Governance, Vol. 2, SAGE Publications, 2006, pp. 863-864

33 Yankelovich, Daniel and David Meer. "Rediscovering market segmentation." Harvard business Review 84.2 (2006): 122.