



**MASINDE MULIRO UNIVERSITY OF SCIENCE AND  
TECHNOLOGY**

**SCHOOL OF COMPUTING AND INFORMATICS**

**DEPARTMENT OF COMPUTER SCIENCE**

**COURSE:PROJECT**

**COURSE CODE: BCS 417**

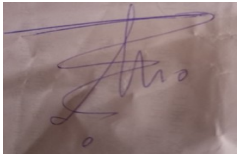
**SUBMITTED BY : BETTY KAMANTHE MUATHE**

**PROJECT TITLE : CUSTOMER SEGMENTATION**

## DECLARATION

I Betty Kamanthe Muathe of REG NO COM/B/01-00164/2017 hereby declare that this report is an original work and has not been published or submitted to this organization or any other institution of training for any academic award.

Signature:..... Date: 8<sup>th</sup> November 2021

A handwritten signature in purple ink, appearing to be 'Betty Kamanthe Muathe', is written on a light-colored surface.

## ACKNOWLEDGMENT

This project would have not been successful without the corporation and support of a number of people who guide us throughout the project coding and documenting this report. First and foremost, we want to thank the Almighty God for the charitable time, strength and aptitude that enable us to complete our project. We are very grateful to our supervisor Dr. Ujunju for providing direction throughout this project. I also thank my parents whose moral and financial support has remained unrelenting, dear friends and colleagues whose motivations kept us moving. I cannot end this list without paying tribute to the entire MMUST fraternity, Lecturers particularly those from the department of Information Technology for their constructive training and the knowledge they have imparted in us throughout the four years training. May God bless you abundantly.

## Table of Contents

DECLARATION.....	i
ACKNOWLEDGMENT.....	ii
ABSTRACT.....	V
1. CHAPTER ONE.....	1
1.1 BACKGROUND.....	1
1.2 STATEMENT OF THE PROBLEM.....	3
1.3 MAIN AIM OF THE PROJECT.....	4
1.4 SIGNIFICANCE.....	5
2. CHAPTER TWO.....	7
2.1 INTRODUCTION.....	7
2.2 Cluster analysis in market segmentation.....	10
2.3 Classification of Clustering.....	12
2.3.1 Hierarchical methods.....	12
2.3.2 Partitioning methods .....	13
2.4 CLUSTERING ALGORITHMS.....	15
2.4.1 K-Means.....	15
2.4.2 DIANA.....	15
2.4.3 K-Medoids.....	15
2.4.4 CLARA.....	16
2.4.5 FANNY.....	16
2.4.6 SOM.....	16
2.4.7 Model-Based Clustering.....	16
2.4.8 SOTA.....	17
2.5 CLUSTER VALIDATION.....	18
3. CHAPTER 3.....	19
3.1 RESEARCH DESIGN.....	19
3.2 RESEARCH DATA .....	19
3.3 UNSUPERVISED LEARNING.....	21
3.4 VALIDATION.....	30
4. CHAPTER 4.....	34

4.1.	INPUT DESIGN.....	34
4.2.	OUTPUT DESIGN.....	34
4.3.	DATABASE DESIGN.....	35
4.4.	NORMALIZATION.....	37
5.	CHAPTER 5.....	40
5.1.	Introduction .....	40
5.2.	SOFTWARE TESTING.....	41
5.3.	Architectural Goals and Constraints.....	44
5.4.	User Interactions with the system.....	44
5.5.	Results of Implementation.....	45
5.6.	Achievements of the project.....	45
5.7.	Limitations of the project.....	45
6.	CHAPTER FIVE: SUMMARY, RECOMMENDATION AND CONCLUSION.....	46
6.1.	Introduction.....	46
6.2.	Summary.....	47
6.3.	Conclusion.....	48
7.	REFERENCES.....	49

## **ABSTRACT**

To scale efficiently and effectively, expansion-stage companies need to focus their efforts on a specific subset of customers who are most similar to their current customers, not a broad universe of potential customers . Knowledge of customer behavior 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 becoming transformational, characterized 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 change with time is crucial. Customer segmentation and profiling has become an indispensable tool for organizations 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 data, 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, Spending score 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. Thereafter, classification into a specific homogeneous 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 marketing characterized 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 behavior 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. 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 characterized 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 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,  
1especially in emerging markets where most transactions still happen through cash.

## 1.2 Problem Statement

Businesses want to understand the customers who can easily converge(Target market), so that sense can be given to the marketing team and plan the strategy accordingly.

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 behavior, a combination of preference and ability. Additionally, segmentation has been mostly confined to the behavior 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 into homogeneous groups based on spending habits and then profile the groups by describing them based on their characteristics.

### **Specific objectives**

1. Track the difference between loyal customers vs visitors, perform heatmap analysis of their buying patterns.
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 empower our marketing department to make better strategic decisions.

## 1.4 Significance

This study delves into various clustering algorithms in consumer segmentation using the spending habits across various categories. 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 behavior. 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 behavior 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 . 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 are significant in bridging the gap above.

## CHAPTER 2: LITERATURE REVIEW

### 2.1 Introduction

Clustering can be defined as the procedure of splitting data with similar characteristics 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. That division is based on customers having similar:

1. Needs (i.e., so a single whole product can satisfy them)
2. Buying characteristics (i.e., responses to messaging, marketing channels, and sales channels, that a single go-to-market approach can be used to sell to them competitively and economically)

There are three main approaches to market segmentation:

1. **A priori segmentation**, the simplest approach, uses a classification scheme based on publicly available characteristics—such as industry and company size—to create distinct groups of customers within a market. However, a priori market segmentation may not always be valid since companies in the same industry and of the same size may have very different needs.
2. **Needs-based segmentation** is based on differentiated, validated drivers (needs) that customers express for a specific product or service being offered. The needs are

discovered and verified through primary market research, and segments are demarcated based on those different needs rather than characteristics such as industry or company size.

3. **Value-based segmentation** differentiates customers by their economic value, grouping customers with the same value level into individual segments that can be distinctly targeted.

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 categorized 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 behaviors by establishing homogeneous 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 an essential role to play in seeking improved comprehension of buyer behaviors 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 behavior. He utilizes several progressive clustering techniques that are adapted and activated for customer segment creation. An optimal 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 analyzed. 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 achieved for six distinct segments.

Ansari and Riasi (2016) used a combination of a 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.

The six standard segmentation schemes that can be applied to a customer segmentation research are:

1. Geographic base
2. Industry
3. Product class
4. Organization class
5. Product delivery
6. Needs

It is important to note that even if a market is divided into one of the schemes above, it is still not a valid segmentation of the market unless it results in meaningful differences in customers' values and needs, the company's value proposition, or the go-to-market strategy associated with each scheme. In such cases, it is merely a convenient organization of the market that has no strategic or operational value.

## 2.3 Classification of Clustering

Clustering techniques are commonly divided into the following broad categories:

1. Hierarchical clustering
2. Partitioning clustering
3. 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 or 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.

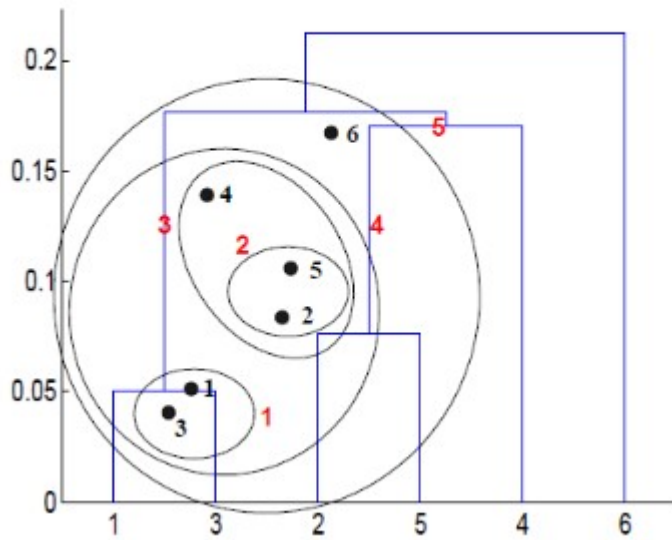


Figure 1: A Nested Cluster

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 approaches 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 Euclidean distance. Just like the K-means technique of partitioning, the number of clusters are 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:

1. reduction of data dimension with minimum loss of information. (These neural networks can single out sets of independent characteristics)
2. 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 hierarchy.



## 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 any market, replicable and reproducible in other similar markets. The data collected is based on a stratified design.

### **3.2 Research Data**

#### **3.2.1 Source of Data**

This research relies on primary data collected on kaggle. 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) is divided into  $12H$  strata. The strata construction is based on the three characteristics that are known – Age, gender and region. Each stratum  $h$  has  $N_h$  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.

#### **3.2.2 Volume of Data**

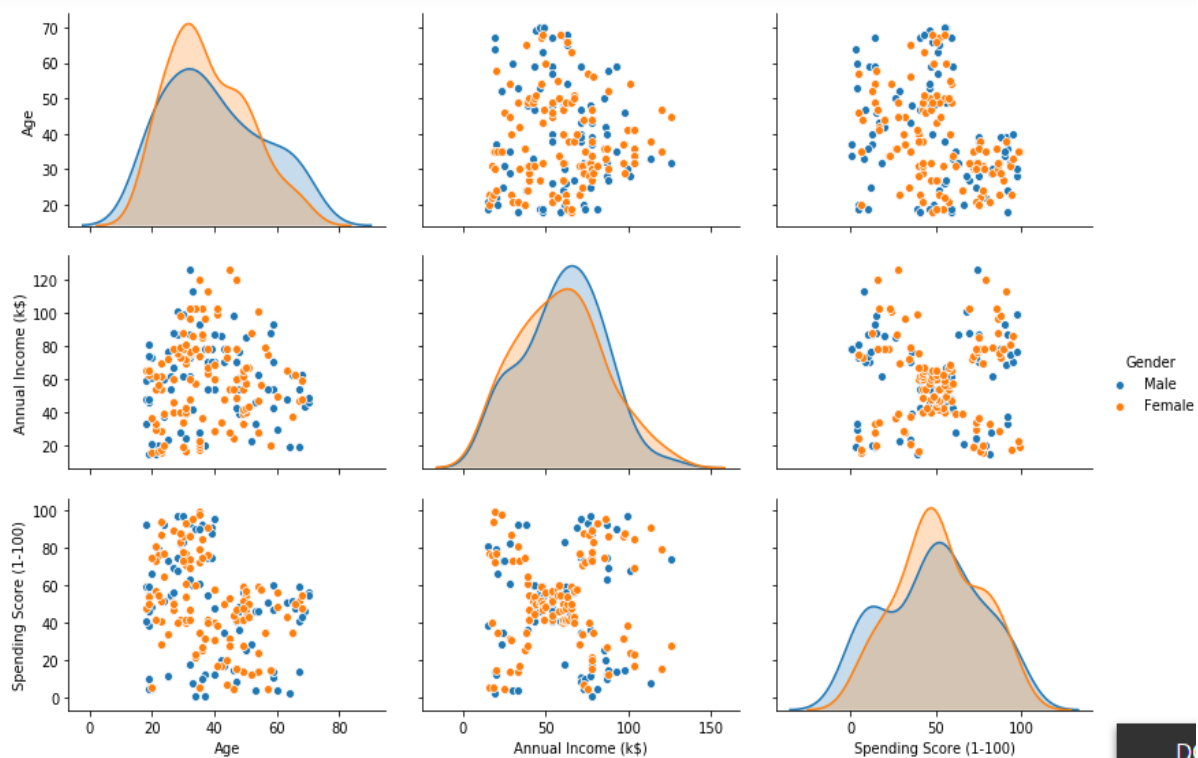
This paper will use data from surveys with consumers, which has been collected since April 2007. The panel has been engaged daily to report their expenditure details. The data is then aggregated into weekly and monthly measures to estimate the average expenditure (wallet size), the share of the wallet 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.

The clustering algorithms will be executed based on the aggregated average expenditure of each of the 200 consumers in the panel.

### 3.2.3 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:



### 3.3 Unsupervised Learning

Unsupervised Learning is a sub-field of Machine Learning, focusing on the study of mechanizing the process of learning without feedback or labels. Unsupervised learning models have no a-prior knowledge about the classes into which data can be placed. They use the features in the data set 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:

$x_{11}$

$x_{21}$

$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix}$$

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:

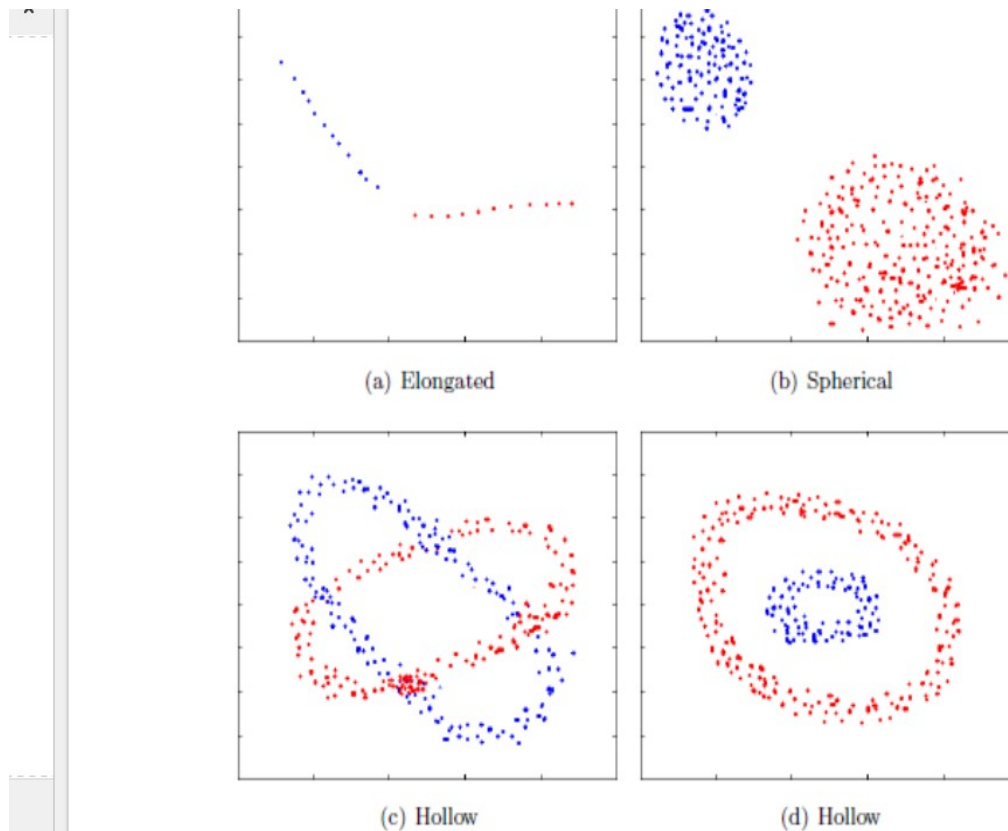


Figure 5: Possible Cluster Shapes

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:

$$\bigcup_{i=1}^c A_i = X, \quad (3.2)$$

$$A_i \cap A_j = \emptyset, \quad 1 \leq i \neq j \leq c, \quad (3.3)$$

$$\emptyset \subset A_i \subset X, \quad 1 \leq i \leq c \quad (3.4)$$

Expressed in the terms of membership functions:

$$\bigvee_{i=1}^c \mu_{A_i} = 1, \quad (3.5)$$

$$\mu_{A_i} \vee \mu_{A_j} = 0, \quad 1 \leq i \neq j \leq c, \quad (3.6)$$

$$0 \leq \mu_{A_i} < 1, \quad 1 \leq i \leq c \quad (3.7)$$

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

$$\mu_{ij} \in \{0,1\}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq c, \quad (3.8)$$

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

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

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

$$M_{hc} = \{U \in \mathbb{R}^{N \times c} | \mu_{ik} \in \{0,1\}, \forall i, k; \sum_{k=1}^c \mu_{ik} = 1, \forall i; 0 < \sum_{i=1}^N \mu_{ik} < N, \forall k\},$$

### b) Fuzzy partition

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

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

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

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

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

$$M_{fc} = \{U \in \mathbb{R}^{N \times c} \mid \mu_{ik} \in \{0,1\}, \forall i, k; \sum_{k=1}^c \mu_{ik} = 1, \forall i; 0 < \sum_{i=1}^N \mu_{ik} < N, \forall k\}$$

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:

$$\sum_{i=1}^c \sum_{k \in A_i} \|x_k - v_i\|^2$$

$A_i$  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



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

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 be 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 points in cluster  $i$  that minimizes

$$\sum_{i=1}^c \sum_{k \in A_i} \|x_k - x_{i^*}\|^2,$$

Stop when within-cluster variation does not change.

This algorithm shares the same properties as the 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:

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

With

$$V = [v_1, v_2, v_3, \dots, v_C], v_i \in \mathbb{R}^n,$$

$V$  represents the vector with the partition centres that must be established. The distance measure,  $\|x_k - v_i\|^2_A$  is known as a squared inner-product distance norm and is defined by:

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

Using Lagrange multipliers to establish the stationary points yields:

$$J(X; U, V, \lambda) = \sum_{k=1}^N \sum_{i=1}^C (\mu_{ik})^A D_{ikA} + \sum_{k=1}^N \lambda_k \left( \sum_{i=1}^C \mu_{ik} - 1 \right),$$

and by setting the gradients of  $\hat{J}$ , with respect to,  $U$ ,  $V$  and  $\lambda$  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  $U(k)$

$$V_i = \frac{\sum_{j=1}^N u_{ij}^m x_i}{\sum_{j=1}^N u_{ij}^m},$$

3. Update  $U(k), U(k+1)$
4.  $d_{ij} = \sqrt{\sum_{i=1}^m (x_i - v_i)^2}$

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

5. If  $\|U(k+1) - U(k)\| < \epsilon$ , 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:

$$D_{ik}A_i = (x_k - v_i)^T \cdot A_i (x_k - v_i), \text{ where } 1 \leq i \leq c \text{ and } 1 \leq k \leq N$$

For this algorithm, the objective function is computed by:

m

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

If we vary  $A_i$  to optimize the clusters while fixing the volume:

$$\|A_i\| = \rho_i, \rho_i > 0$$

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

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

With

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

### 3.4. Validation

Unlike in supervised learning where we have 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 analyze 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  $j$ th nearest neighbor 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 otherwise. For a particular clustering partition,  $C = \{C_1, C_2, \dots, C_k\}$  of the  $N$  observations into  $K$  disjoint clusters, the connectivity is given by:

$$conn(C) = \sum N$$

$$i=1 \sum_{j=1}^n x_{i,j}$$

## 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:

$$S(i) = (b_i - a_i) / \max(b_i, a_i)$$

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.

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

where  $C(i)$  is the cluster containing observation  $i$ ,  $\text{dist}(i, j)$  is the (e.g. Euclidean, Manhattan) distance between observations  $i$  and  $j$ , and  $n(C)$  is the cardinality of cluster  $C$ .

where  $C(i)$  is the cluster containing observation  $i$ ,  $\text{dist}(i, j)$  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 ratio of the tiniest separation of objects that are not in the same cluster to the biggest intra-cluster distance. It is computed as:

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

This index ranges from 0 to  $\infty$  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  $C^{i,l}$  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:

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

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:

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

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:

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

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  $\ell$ , FOM is calculated as follows:

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

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: SYSTEM DESIGN**

### **4.1 INPUT DESIGN**

Input design is the process of converting the user originated input to a computer based format. The design decision for handling input specific data are accepted for computer processing. Input design is a part of overall system design that needs careful attention. The input design of the system includes the design of:

1. Customer spending score
2. Customer income that has to rounded off

### **4.2 OUTPUT DESIGN**

One of the most important features of a system for user is the output it produces. Output design should improve the systems relationship with the user and help in decision making. Computer output is a process that involves designing necessary output that have to be given to various users according to their requirements. Efficient, intelligible output design should improve the system relationship with the user and help and in decision making. A major form of output is the use of simple sentences. The proposed system will utilize simple language that can be understood even by people who have not advanced in business knowledge.

The objective of output design is to define the controls and format of customer segment predictions that will be produced by the system. The output is the most important and direct source of information to the user.

For many end users output is the main reason for developing the system and the basis on which they will evaluate the usefulness of the application. Output generally refers to the system results. The output of the Intellibusiness is designed so as to include a number of reports. Reports reflect the output design. Output design is an ongoing activity, which start during study phase itself. Output generally refers to the results and information data are generated by the system. It can be in the form of operational documents and reports.

Objectives of Output Design

1. Design output to serve the intended purpose
2. Deliver appropriate type of output
3. Choose the right output method
4. Provide output on time

### 4.3 DATABASE DESIGN

A database is an organized mechanism that has the capability of storing information through which a user can retrieve stored information in an effective and efficient manner. The data in the proposed system's database include:

1. Customer id
2. Age
3. Spending score
4. Income
5. Gender

The data in the database is safe and easily accessed. The database design is a two level process. In the first step, user requirements are gathered together and a database is designed which will meet these requirements as clearly as possible. This step is called Information Level Design and it is taken independent of any individual DBMS.

In the second step, this Information level design is transferred into a design for the specific DBMS that will be used to implement the project repository system. This step is called Physical Level Design, concerned with the characteristics of the Project Repository System database. A database design runs parallel with the system design.

The organization of the data in the database is aimed to achieve the following two major objectives.

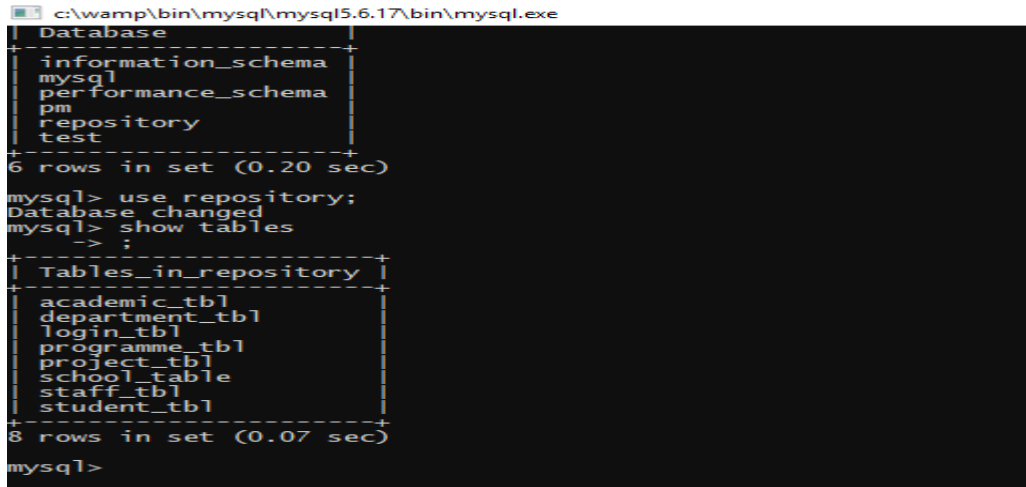
1. Data Integrity
2. Data independence

Normalization is the process of decomposing the attributes in an application, which results in a set of tables with very simple structure. The purpose of normalization is to make tables as simple as possible. Normalization is carried out in this system for the following reasons.

1. To structure the data so that there is no repetition of data, this helps in saving space.
2. To permit simple retrieval of data in response to query and report request.
3. To simplify the maintenance of the data through updates, insertions, deletions.
4. To reduce the need to restructure or reorganize data which new application requirements arise.

### 4.3.1 Database Architecture

As stated earlier in software requirements, the system uses MySQL server databases. Preview of the architecture of MySQL server is as shown below;



```
c:\wamp\bin\mysql\mysql5.6.17\bin\mysql.exe
Database
+-----+
| information_schema |
| mysql              |
| performance_schema |
| pm                 |
| repository          |
| test               |
+-----+
6 rows in set (0.20 sec)

mysql> use repository;
Database changed
mysql> show tables
+-----+
| Tables_in_repository |
+-----+
| academic_tbl          |
| department_tbl        |
| login_tbl             |
| programme_tbl         |
| project_tbl           |
| school_table          |
| staff_tbl             |
| student_tbl           |
+-----+
8 rows in set (0.07 sec)

mysql>
```

MySQL is a freely available open source Relational Database Management System (RDBMS) that uses Structured Query Language (SQL). SQL is the most popular language for adding, accessing and managing content in a database.

It is most noted for its quick processing, proven reliability, ease and flexibility of use. The server is responsible for storing data of the system such as vehicle information, driver information, users' information. etc.

### 4.3.2 Relational Database Management System (RDBMS)

A relational model represents the database as a collection of relations. Each relation resembles a table of values or file of records. In formal relational model terminology, a row is called a tuple, a column header is called an attribute and the table is called a relation. A relational database consists of a collection of tables, each of which is assigned a unique name. A row in a table represents a set of related values.

### 4.3.3 Relations Domains & Attributes

A table is a relation. The rows in a table are called tuples. A tuple is an ordered set of n elements. Columns are referred to as attributes. Relationships have been set between every table in the database. This ensures both Referential and Entity Relationship Integrity.

A domain D is a set of atomic values. A common method of specifying a domain is to specify a data type from which the data values forming the domain are drawn. It is also useful to specify a name for the domain to help in interpreting its values.

Every value in a relation is atomic, i.e. Not decomposable. Relationships Table relationships are established using Key. The two main keys of prime importance are Primary Key & Foreign Key. Entity Integrity and Referential Integrity Relationships can be established with these keys:

1. Entity Integrity enforces that no Primary Key can have null values.
2. Referential Integrity enforces that no Primary Key can have null values.
3. Referential Integrity for each distinct Foreign Key value, there must exist a matching Primary Key value in the same domain. Other keys are Super Key and Candidate Keys.
4. Relationships have been set between every table in the database. This ensures both Referential and Entity Relationship Integrity.

## 4.4 NORMALIZATION

As the name implies, it denoted putting things in the normal form. The application developer via normalization tries to achieve a sensible organization of data into proper tables and columns and where names can be easily correlated to the data by the user.

Normalization eliminates repeating groups at data and thereby avoids data redundancy, which proves to be a great burden on the computer resources. Normalization is the systematic technique of transforming data subject to a whole range of file maintenance problem into an organized data free from such problem Detecting tables through a number of levels of normalization.

It is carried out in four different steps:

1. Represent the unnormalized table or relation.
2. Transform the unnormalized table to the First Normal Form (1NF).
3. Transform of First Normal Form into Second Normal Form (2NF).
4. Transformation of Second Normal Form into the Third Normal Form (3NF).

These include:

1. Normalize the data.
2. Choose proper names for the tables and columns.

### 3. Choose the proper name for the data.

First Normal Form: The first step is to put the data into First Normal Form. This can be done by moving data into separate tables where the data is of similar type in each table. Each table is given a Primary Key or Foreign Key as per requirement of the project. This eliminated repeating groups of data.

Second Normal Form: This step helps in taking out data that is only dependent on a part of the key.

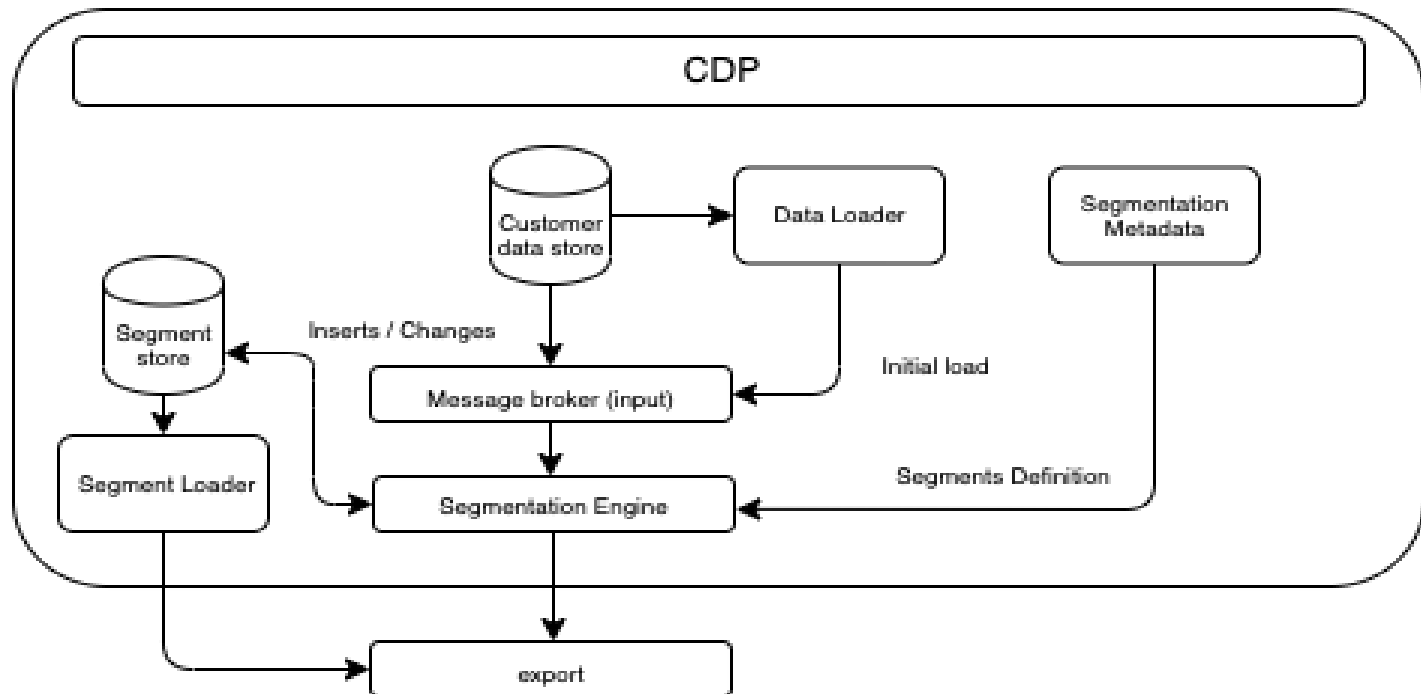
Third Normal Form: This step is taken to get rid of anything that does not depend entirely on the Primary Key.

#### Table design

| :

	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72

## SYSTEM DESIGN



The process is explained as below:

1. All data is stored in the data warehouse or data base
2. It is extracted.
3. The extracted data is cleaned and analyzed
4. The data is passed through the segmentation engine
5. After all the customers have been segmented the data is saved to a database.
6. Data can be exported in any form to be used for decision making

## **CHAPTER 5 : SYSTEM IMPLEMENTATION**

### **5.1 Introduction**

Project management system is based on these tools:

1. Jupyter Notebook - is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.
2. Flask - is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

The system is implemented in the form of a web application. Its hosted on a web server and the users can use it from their Internet access devices such as mobile phones and computers.

After the design was done, the actual development of the system took place to come up with the final product. I used python programming language as it's a powerful programming language that is able to do many operations. It is also an easy language to learn and implement.

I chose the approach of a web application as one can access more computational resources. We used SQL to access the database because it's easy to use, its secure, it has high performance, It has comprehensive Transaction support and it has flexibility of Open source.

## **5.2 SOFTWARE TESTING**

### **5.2.1 Unit testing**

Unit testing aims at testing the particular code blocks that together comprise the whole system. Unit testing is inspired by TDD test driven development whereby the unit requirements are orchestrated into a unit test, after the test is written the unit is developed and the test suite for the module is executed, if the test passes then the unit can be pushed to production.

Unit tests allow for further configuration which includes scope configuration. Defining the scope of the unit test will determine where the test is executed. For instance, if the scope of the test is development the test suite will execute only in the development environment whereas production scope will allow the test to run in production. The testing of the system was done using several computers to test its compatibility.

### **5.2.2 Module testing**

In this test scenario unit testing is applied to the whole module otherwise called functional testing. Since each module is a class in architecture functions independently the specialized function for the specific module are tested before the handshake between this module and other modules it integrates with are tested in integration testing.

### **5.2.3 Integration testing**

This test checks that when the modules are integrated they can combine to perform their respective functions. The essence of integration is to ascertain that when modules are combined they do not lose their efficiency and reliability. Though modules can still perform their intended functions solely, integrating them uncovers the errors that come with the whole handshaking process between the modules.



#### 5.2.4 User Acceptance Testing

After the requirements are integrated during development a test deployment is deployed to a limited number of users who test the new features. If the users, verify that the features work as required then the features are deployed to the full production environment. This restores confidence in the fault tolerance of the whole system in general. User acceptance testing is based on two main software engineering concepts white box testing and continuous integration CI/continuous deployment CD.

#### *The technologies used to develop the system*

- a) Encryption- this is a service that can authenticate users using only client-side code. It supports social login providers Facebook, GitHub, Twitter and Google. Additionally, it includes a user management system whereby developers can enable user authentication with email and password login stored with Firebase.
- b) Real time database- this is a service that provides application developers an API that allows application data to be synchronized across multiple clients and stored in firebase's cloud. The REST API uses the Server-Sent Events protocol, which is an API for creating HTTP connections for receiving push notification from a server.
- c) Machine learning – is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
- d) Flask - is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.

figure4 .0 below shows two flow chart that demonstrates how the system works.



INTELLIBUSINESS  
Growing Bigger

## Customer Segmentation

ANNUAL INCOME

SPENDING SCORE (1-100)

SUBMIT

The user enters the customer's annual income rounded off

Spending score - To calculate spending scores, you first need the values of three attributes for each customer: 1) most recent purchase date, 2) number of transactions within the period (often a year), and 3) total or average sales attributed to the customer (total or average margin works even better)

### 5.3 Architectural Goals and Constraints

This section describes the software requirements and objectives that have some significant impact on the architecture.

Technical Platform: The backend application will be deployed on a server.

Security: The system must be secured, so that unauthorized users cannot make changes. The basic security behaviors being:

- o Authentication: Login using at least a user name and a password
- o Authorization: according to their profile, user must be granted or not allowed to receive some specific services

For system access, the following requirements are mandatory

- o Confidentiality: sensitive data must be encrypted if any.
- o Safety: passwords must be hashed before storage
- o Data integrity: Data sent across the network cannot be modified by a tier
- o Auditing: Every sensitive action will be logged
- o Non-repudiation: gives evidence a specific action occurred

Persistence: Data persistence was addressed using a relational database specifically MYSQL.

Reliability/Availability: High availability is required since there are time and location issues related to the systems availability. The motorists should not be disappointed. The system's high availability will also ensure motorists satisfaction and loyalty. Targeted availability: 16 hours a day, 7 days a week (Maintenance at night)

Performance: Search queries should return %90 of the time below 5 sec. Status updates of the slot information was set to be after every 10 seconds. If a node fails to update, it is marked as unknown.

### 5.4 User Interactions with the system

This section provides a functional overview of the system by a use-case diagram.

a) Actors: These are the system users

i) Administrators: These are the people who create, edit and modify system settings.

ii) Managers: These are the people who are interested in observing the information generated by the system to make management decisions.

## **5.5 Results of Implementation**

The Researcher found out that it should have graphical user interface where User of the system (Business analysts) can perform all transaction intended to do i.e. can view client segments and , make marketing and strategic decisions based on the results.

## **5.6 Achievements of the project**

- Business analysts can make business decisions that are customer centered
- Marketing teams can strategically build marketing messages for their target audience
- Gives managers insight on their customer pool

## **5.7 Limitations of the project**

It can only be accessed online .

## **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 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

36agglomerative 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