UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

Clustering and Visualizing the Status of Child Health in Kenya: A Data Mining Approach.

NJIRU, NICHOLAS MURIUKI

(P52/72772/2014)

Supervisor

DR. ELISHA T. OPIYO OMULO

A project report submitted in partial fulfillment of the requirement for the award of Masters of Science in Computational Intelligence of the University of Nairobi

December 2015

Declaration

This project report is my original work and has not been presented for any award in any other University.

_____ _____

Signature Date

Nicholas Muriuki Njiru,

(P52/72772/2014)

This project report has been submitted for examination with my approval as the University supervisor.

_____ _____

Signature Date

Dr. Elisha T. Opiyo Omulo

Senior Lecturer

School of Computing and Informatics

University of Nairobi

Table of Contents

Table of Figures

List of Abbreviations

KDHS	-	Kenya Demographic and Health Survey
PCA	-	Principal Components Analysis
UNICEF	-	United Nations International Children's Emergency Fund
R	-	Revolution Analytics
CRISP-DM	-	Cross Industry Standard Process for Data Mining
KDD	-	Knowledge Discovery and Data Mining
SEMMA	-	Sample, Explore, Modify, Model and Assess
SAS	-	Statistical Analysis System
HCA	-	Hierarchical Cluster Analysis
WHO	-	World Health Organization
HIV/AIDS	-	Human Immune Deficiency /Acquired Immune Deficiency Syndrome
MCH	-	Maternal and Child Health
PAM	-	Partitioning Around Medoid
FANNY	-	Fuzzy Analysis
CLARA	-	Cluster Large Application
AGNES	-	Agglomerative Analysis
DIANA	-	Divisive Analysis

Acknowledgement

This research would not have been possible without the help provided by many people. First and foremost, I would like to thank the contributions of my supervisor Dr. Elisha T. Opiyo for his dedication and immense advice during my research work. I also want to thank the lecturers at the School of computing and Informatics for the knowledge they imparted me during the course work. I wish to commend the criticism, suggestions, and advice from the panelists Dr. Robert Oboko and Dr. Agnes Wausi for their contributions in this research. I would like to thank in a special way my wife Mary and daughter Bridget for being patient when I was not there for them whenever they needed me most. Finally, I would like to thank my family and twin Brother John for their constant inspiration and being there for me always. Above all, I would like to thank the mighty God for giving me intelligence and comprehension in my academic and professional accomplishments.

ABSTRACT

The inauguration of the new constitution in Kenya led to the devolution of health care in the counties. It is against this backdrop that necessitated a need to develop a model of grouping these counties into natural groups with similar characteristics that can influence the child health for the purpose of health care planning and regulation. Little research has explored a methodology that can be used to create such groupings in Kenya. The purpose of this research was to develop and explore a methodology of Clustering and Visualizing the status of the child health in Kenya. In this research we proposed a new model that clustered the counties based on the UNICEF indicators of child health. The cluster analysis methodology employed to achieve this was by use of K-Means clustering algorithm. Both hierarchical and non-hierarchical clustering algorithms were used to build a consensus with the results of clusters obtained by K-Means. The number of clusters selected was based on heuristic, integrating a statistical-based measure of cluster fit. Using data from literature, the clustering methodology developed grouped the 47 counties into three distinctive clusters. These three clusters were made up of 10, 8 and 29 counties respectively. The study classified the clusters as well-off, most marginalized and moderately marginalized counties respectively. The methodology developed was objective, replicable and sustainable to create the clusters. It was developed in a theoretically sound principle and can be generalized across applications requiring clustering. An examination of several clustering algorithms revealed similar results.

Keywords: Principal Component Analysis, K-Means, Clustering, Visualizing, Child Health Indicators, Data Mining, Dimensionality Reduction.

CHAPTER 1: INTRODUCTION

1.1 Background of the study

The inauguration of the new constitution has invoked the researchers in Kenya to do more research putting into considerations the devolved administrative regions called counties which have a wealth of information about them. The World Bank described the Kenya's devolution as one of the most ambitious globally. Under that consideration this research was meant to explore and develop a model that can be used by policy makers as a guide to be successful in achieving its mandate for provision of childcare by understanding the status quo of their regions. Health sector in Kenya has been centralized to the national government since independence. This led to spatial inequalities in different regions that have been inherited by the county governments. The research will support the stakeholders of child health in these counties such as the national government, non-governmental organizations and private individuals (consumers), researchers and planners in decision making and planning.

Children represent the future, and ensuring their healthy growth and development ought to be a prime concern of all societies (WHO).Child health refers to the state of physical, mental, intellectual, social and emotional well-being and does not imply just the absence of a disease or infirmity (WHO factsheet N220, 2014). The Child health is determined by the UNICEF indicators of child or other metrics. Article 1 of UNICEF convention on the child rights defines a child as a person below the age of 18 but allows laws of a particular country to set the legal age of a child (UNICEF factsheet). According to the Kenyan constitution children Act CAP 141, a child is any human being under the age of eighteen years. This research will concentrate on the cohort aged between 0 to 18years. In Kenya this age group account for 42.1% of which the populations male is 9,494,983 while that of female is 9,435,795(Kenya Demographics profile, 2014). To get healthy children, families, environments, and communities must provide them with the opportunity to help them grow into adulthood (Health Workgroup, 2007). To achieve optimal health, children are dependent upon adults in their family, government and community to provide them with an environment in which they can learn and grow (Health Workgroup, 2007).

The indicators identified by UNICEF have a great influence on child health. The direct and indirect expenditure related with child health are extremely huge. This has contributed to poor

economic performance of developing countries. In Kenya, previous research done on child health have mostly concentrated on diseases, family planning, HIV/AIDS and maternal health. This research focuses on taking a different approach by looking at the holistic view in creating a framework for visualizing the status of child health in the Kenyan counties based on the UNICEF indicators of health.

This framework was achieved through the data mining approach. Data mining is a multidisciplinary analytical technique made up of statistics, computer science, mathematics, and database technology (S. Fong, 2015). Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Over the past two decades there has been an explosion of big data stored in databases and other database applications in business and the scientific domain. This explosion of data stores electronically accelerated the relational model but little emphasis for the analysis of data was considered. Businesses discovered that these masses of data can be analyzed to uncover hidden patterns in these data and this gave birth to the concept of data mining. Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, and machine learning.

1.2 Statement of the Problem

Since independence, health in Kenya has been centralized to the national government with the power being centered at the capital Nairobi. This has resulted to a spatial inequality in the child health care. The large amount of data in the literature on child health, maternal health, water and sanitation, education and income have a lot of influence on child health and is available from the governments, non-governmental organizations and from medical health systems. The major challenge is the expertise of discovering the hidden pattern behind these data.

Cluster analysis is a rapidly growing technique used in variety of engineering and scientific disciplines that can be used in organizing these data by abstracting underlying structure that can be insightful. This research proposes cluster analysis in unearthing the counties most affected by these inequalities.

1.3 Research Objective

The purpose of this research is to use a data mining to develop a model of Clustering and Visualizing the status of child health in Kenya that can be used by players and policy makers in child health in decision making.

The objective of this study is:

i. Cluster the counties into "natural groups" according to their similarity or dissimilarity in child health status using data mining algorithms.
ii. To visualize the status of child health using visualization tools and techniques in data mining in 2-Dimension that can be conceptualized by easily by human beings.

1.4 Significance of the Study

The Government of Kenya has endorsed a range of these global initiatives, including the Global Strategy for Women's and Children's Health and the Millennium Development Goals (MDGs), specifically emphasizing MDGs 4, 5, and 6 (focusing on maternal health, children's health, and HIV/AIDS), and has made specific commitments to achieve them (T. Abuya, 2014).

By developing a framework for the most influential factors, it will assist the policy makers in child health to achieve the Millennium Development Goals. The new finding can also be used as a guide to further research. Kenya health expenditure accounts for 4.5% of GDP (2011).

The expected contribution to knowledge of this research:

i. The research will contribute to understanding of the status of child health in Kenya.
ii. It will combine theories and methodologies from various disciplines (Machine learning, Data Mining, Statistics, Computer science, Mathematics etc) to understand the requirements of the health for the Kenyan child.
iii. It will validate the applicability of existing theories of child health.
iv. It will contribute to the development of appropriate framework for child health policy analysis in Kenya.

1.5 Scope of the study

The scope of this research is limited to the UNICEF child health indicators and limited to 47 counties in Kenya.

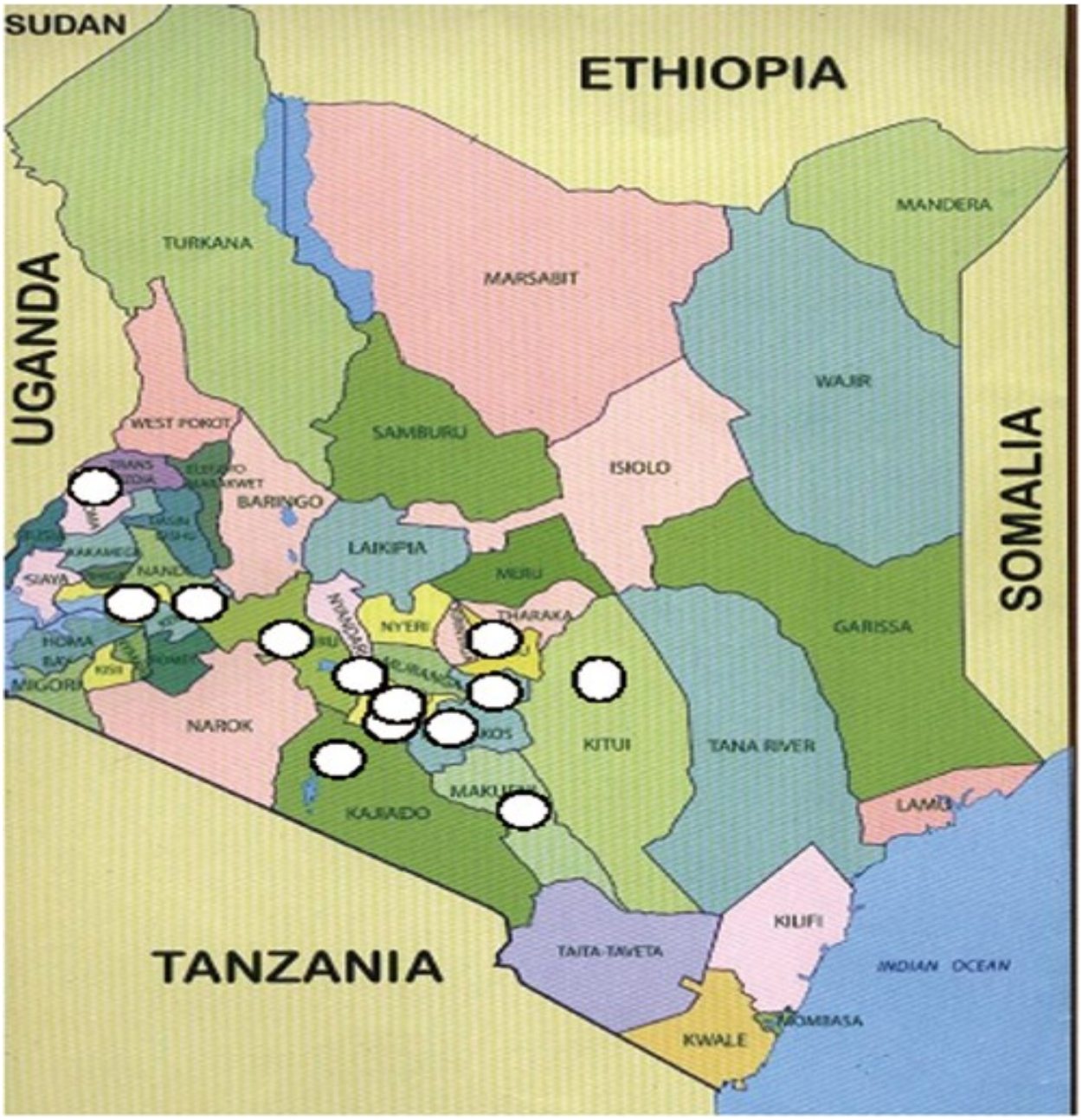


Figure 1-Kenya Counties Map

1.6 Limitations and Assumptions of the study

The major limitation in this study was the conversion of data in the literature to a form that could be analyzed since most of the research from literature was not supported by any data sets. The researcher attempted to find out if such data was available but was not in a format that could be used in our research. Due to time limitation, the focus of the observations and variables to be studied was limited. The assumption in the study was that the secondary data that was collected

from the previous research did not have errors. Time lag issues were assumed since the data could have not been a reflection of the present since it was historical data. Another assumption is that there was to be a linear relationship between variables and were suitable for data reduction and there was to be no significant outliers that could influence the results.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The data mining field proposes the development of methods and techniques for assigning useful meanings for data stored in databases. It gathers researches from many study fields like machine learning, pattern recognition, databases, statistics, and Artificial Intelligence, knowledge acquisition for expert systems, data visualization and grids. Data mining represents a set of specific algorithms of finding useful meanings in stored data (Jing He, 2009).

Data mining has become a well-established discipline within the domain of artificial intelligence (AI) and knowledge engineering (KE). It has its roots in machine learning and statistics, but encompasses other areas of computer science. It has received much interest over the last decade as advances in computer hardware have provided the processing power to enable large-scale data mining to be conducted. Data mining can be argued to be an application rather than a technology and thus can be expected to remain topical for the foreseeable future (Frans Coenen, 2011).

2.2 History of Data Mining

1970s: Computers have been used by statisticians for a long time to prove or disprove hypothesis on collected data. Statisticians used techniques such as Linear Regression, Nearest Neighbors in analysis. Fuzzy logic and other non-linear methods are new techniques for data analysis that have emerged. Statistics assumes that the researcher starts with hypothesis to establish the associations between the data attributes and aided by statistical tools approve and disprove the hypothesis. This is challenging while dealing with data with scores of variables since the model of hypothesize-and-test is time consuming during data analysis.

With the emergence of computers that could store huge volumes of data in sizes of Terabytes and Petabytes, data mining became an emerging trend in computer science to carry out a highly interactive data analysis. With the emergence of data warehouse, multidimensional data models have increased as users move data from the operational databases. This resulted to analysts to quest for development of more analytical way of viewing the data. Even with the aid of sophisticated visualization tools, human brain is extremely limited to analyze this huge data manually. During the 1970s, artificial intelligence branch of computer science was touted for its capability to analyze data.

1980s: In the 80s there was a continued development of AI algorithms that were designed to train machines and machine learning algorithms were embraced becoming realistic tools for dealing with large data sets. Unlike statistical techniques that required analysts to develop a hypothesis first, the Machine Learning algorithms had a capability of first analyzing data and identifying relationships between the variable and the entities in the data to develop models that would allow the domain experts who are not professional statisticians to visualize relationship between the attributes and the data sets. This relaxed the model of "hypothesize-and test" and gave birth to the "test-and-hypothesize" paradigm.

1990s: In 90s, data mining exploded with the development of the machine learning algorithms. Many financial and retail businesses applied the complex analytical capabilities of these algorithms to grow their customer base, fraud detection, study trends and pattern to predict fluctuations in interest rates, stock prices and economic demand. The popularity of data mining has been contributed by these successes. Unlike a human being who cannot deal with many attributes, it allows trends and data pattern recognition through automatic analysis of data.

Data mining is an amalgamation of several technologies such as data management, statistics, Machine Learning and visualization. Tools have been developed in AI that are now capable of classifying data sets, associating certain attributes or entities, segmenting the data into similar clusters, and identifying outliers in the data. Knowledge discovery from conventional databases consists of the process of collection, abstraction and cleansing of the data. Data mining tools have been developed to find patterns, validation and verification of patterns, visualization of the models and refinement of the collection process.

2.3 Curse of Dimensionality

Curse of dimensionality (Bellman, 1961) phenomena arise when analyzing and organizing data in high-dimensional spaces. When considering difficulties in optimization and dynamic programming, Richard E. Bellman (Haas, 1954) coined the phenomenon called the curse of dimensionality. The gravity of this phenomenon states that when the dimensionality expands for a set of data, the volume of the space increases disproportionally fast that inhabits most of the search algorithms. Central to this problem is the effect of combinatorial explosion.

2.4 Dimension reduction techniques in data mining

In data mining, situations are encountered where large number of variables is present. These variables could be highly correlated or uncorrelated with each other. Including such variables in data mining could lead to overfitting hence accuracy and reliability suffers. This large number of variables poses computational problems. When these variables are used in model development they can increase costs and time taken in processing these variables. Dimensionality of a model refers to the number of predictor variables used by the model. Reducing this dimensionality without sacrificing on accuracy is the key step in data mining. To achieve this practically an intelligent technique is required in data mining. Data reduction techniques exist which depends on the variables intended to be analyzed. PCA is intended for use in quantitative variable while categorical variables uses other methods such as correspondence analysis. This study is limited to quantitative variable analytics. In data mining there are two types of dimension reduction linear and non-linear techniques. This research focuses on linear technique because linear techniques perform dimensionality reduction by embedding the data into a subspace of lower dimensionality. Even though there are a variety of techniques to do so, PCA is by far the most popular (unsupervised) linear technique. Therefore, this research focuses only on PCA as a benchmark.

2.4.1 Principal Component Analysis (PCA)

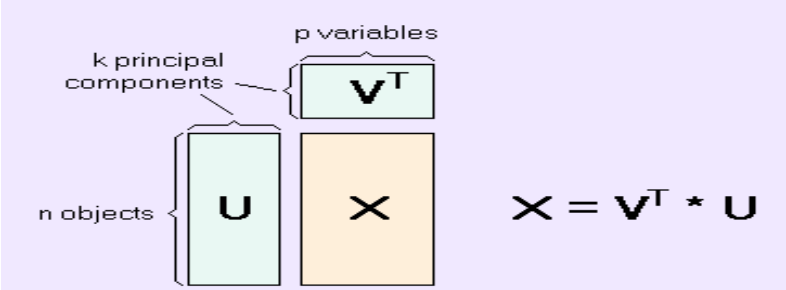
Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It is often used to make data easy to explore and visualize. The main goal of PCA is to reduce the dimensions of an M-dimensional dataset by providing it onto an N-dimensional subspace (where $N \leq M$) in order to increase the computational efficiency while retaining most of the information.

The steps involved in PCA.

1. The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.
2. PCA computes N orthonormal vectors which provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal components. The input data are a linear combination of the principal components.

3. The principal components are sorted in order of decreasing “significance" or strength. The principal components essentially serve as a new set of axes for the data, providing important information about variance.
4. Given that the components are sorted according to decreasing order of “significance", the size of the data can be reduced by eliminating the weaker components, i.e., those with low variance. Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

The concept behind the PCA is the decomposition of the data matrix X into two matrices V and U as shown in the figure below:



Matrices **V** and **U** are orthogonal (same as perpendicular). The matrix **V** is usually called the **loadings** matrix (based on variables), and the matrix **U** is called the **scores** matrix (based on observations). The loadings can be understood as the weights for each original variable when calculating the principal component. The matrix **U** contains the original data in a rotated coordinate system.

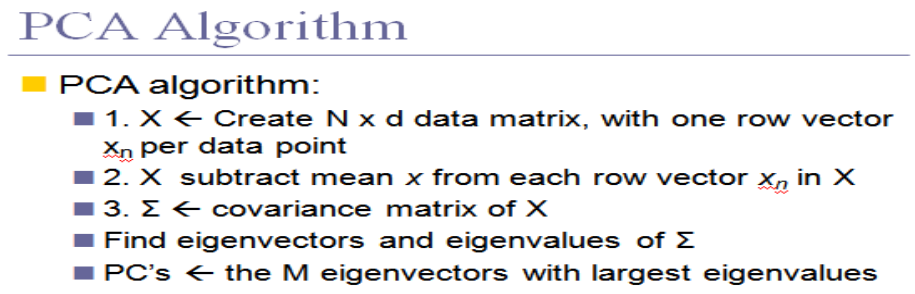


Figure 2-Source J. Leskovec (2006)

2.4.2 Linear Discriminant Analysis (LDA)

It is a dimensionality reduction technique used the pre-processing step for pattern-classification and machine learning applications. This technique is aimed at projecting a dataset onto a lower-dimensional space with good class separability in order avoid overfitting ("curse of

dimensionality") and also reduce computational costs. This technique was formulated by Ronald Fisher in 1936 and has practical uses as a classifier. The original linear discriminant was described for a 2-class problem, and it was then later generalized as "multi-class Linear Discriminant Analysis" or "Multiple Discriminant Analysis" by C. R. Rao in 1948. The general LDA approach is very similar to a Principal Component Analysis with an addition to finding the component axes that maximize the variance of our data (PCA); it additionally has interest in the axes that maximize the separation between multiple classes (LDA). The goal of an LDA is to project a feature space (a dataset n-dimensional samples) onto a smaller subspace k (where $k \leq n-1$) while maintaining the class-discriminatory information. In general, dimensionality reduction does not only help reducing computational costs for a given classification task, but it can also be helpful to avoid overfitting by minimizing the error in parameter estimation ("curse of dimensionality").

Both Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation techniques that are commonly used for dimensionality reduction. PCA can be described as an "unsupervised" algorithm, since it "ignores" class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset. In contrast to PCA, LDA is "supervised" and computes the directions ("linear discriminants") that will represent the axes that maximize the separation between multiple classes (Sebastian Raschka, 2014).

2.5 Clustering Methods

Clustering is defined as an unsupervised learning that occurs by observing only independent variables with no predefined classes. It is mostly used in studies of exploratory nature. The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data

2.5.1 Introduction to Cluster Analysis

Cluster analysis is a technique used for combining observations into groups or clusters such that:

i. Each group or cluster is homogeneous or compact with respect to certain characteristics. This means that the observations in each group are similar to each other.
ii. Each group should be different (heterogeneous) from other groups with respect to the same characteristics. That means that each observation of one group should be different from the observations of the other group

In a nutshell this technique wish to divide a set of n observations into k groups so that members who are more "similar" than the population members. The technique is normally used for discovering structures in data by clustering and hopes to find "natural" groups occurring in data. Clustering algorithms do not have predefined classes like classification and it's based on some similarity measure. The qualities of a good cluster are high intra-cluster similarity and low inter-cluster similarity.

Figure 3-Modeling clusters

Classification of Clustering Methods

Clustering methods is divided into hierarchical and non-hierarchical (Partitioning) methods.

2.5.2 Non-hierarchical or Partitioning Clustering Algorithms

This is a type of where data is divided into k partitions or groups each representing a cluster. In this method, the clusters must be known a priori unlike in non-hierarchical methods. Each group contains at least one observation and each observation should be in exactly one group. This means that k<=n. However, in many cases the researcher may not know how big k should be. To guide the researcher, various methods can be used for this.

i. The researcher can use some distance measure or ratio (like the "within-groups variance" and "between-groups variance" in general discriminant analysis).
ii. The researcher can also use a visual plot called Screeplot. This is subjective since the researcher can think otherwise.

This technique is the choice of cluster centroid and similarity measure. Partitioning methods can be classified into two categories; k-means and k-medoids. K-means is the most widely used than the k-medoids.

2.5.2.1 K-Means algorithm

K-Means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. It works by assigning the center of a cluster to the mean of that cluster. It begins by either some initial assignment of points to k groups or some choice of k group centers. After initialization iterate in the following ways.

Algorithm k-means (k , D)
1 Choose random k data points as initial Clusters Mean (cluster centers)
2 Repeat
3 For each data point x from D
4 Compute the distance between x and each cluster mean (centroid)
5 Assign x to the nearest cluster
6 End for
7 Re-compute the mean for current cluster collections
8 Until reaching stable clusters(current clusters means equals last clusters means)

Figure 4-Source J. Leskovec (2006)

2.5.2.2 K-medoids or Partitioning Around Medoids (PAM) Clustering algorithm

This is the process of partitioning (clustering) of the data into k clusters "around medoids", a more robust version of K-means. The k-medoids clustering is very similar to k-means, and the major difference between them is that: while a cluster is represented with its center in the k-means algorithm, it is represented with the object closest to the center of the cluster in the k-medoids clustering. The k-medoids clustering is more robust than k-means in presence of outliers. PAM (Partitioning Around Medoids) is a classic algorithm for k-medoids clustering. While the PAM algorithm is inefficient for clustering large data, the CLARA algorithm is an enhanced technique of PAM by drawing multiple samples of data, applying PAM on each sample and then returning the best clustering. It performs better than PAM on larger data.

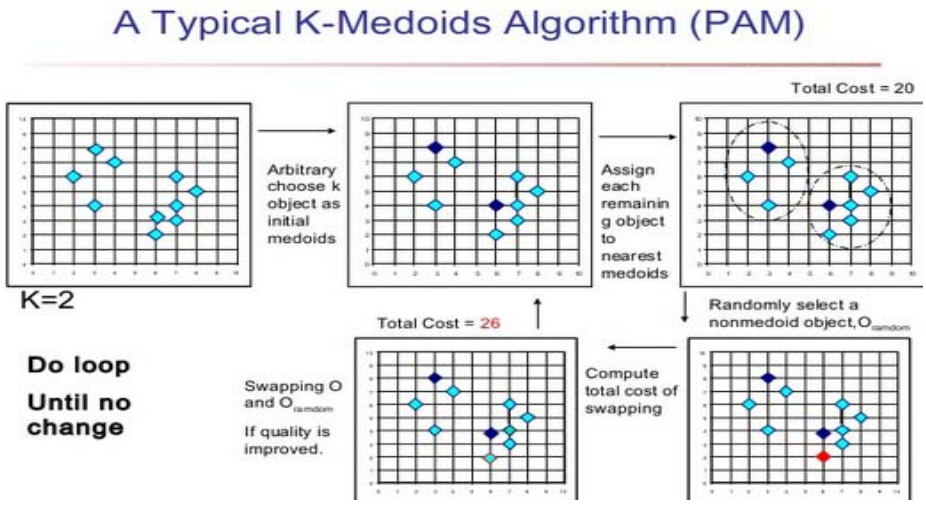


Figure 5-K-Medoids illustrated

2.5.2.3 Clustering Large Applications (CLARA) Clustering algorithm

Each sub-dataset is partitioned into k clusters using the same algorithm as in PAM. Once k representative objects have been selected from the sub-dataset, each observation of the entire dataset is assigned to the nearest medoid. The mean (equivalent to the sum) of the dissimilarities of the observations to their closest medoid is used as a measure of the quality of the clustering. The sub-dataset, for which the mean (or sum) is minimal, is retained. A further analysis is carried out on the final partition. Each sub-dataset is forced to contain the medoids obtained from the best sub-dataset until then. Randomly drawn observations are added to this set until sample size has been reached.

2.5.2.4 Fuzzy Analysis (Fanny) Clustering algorithm

According to Kaufman and Rousseeuw (1990), this algorithm computes a fuzzy clustering of the data into k clusters. In a fuzzy clustering, each observation is "spread out" over the various clusters. Denote by u(i, v) the membership of observation i to cluster v.

In this algorithm, instead of taking the whole set of data into consideration, the **CLARA** (**Clustering LARge Application**) algorithm randomly chooses a small portion of the actual data as a representative of the data. Medoids are then chosen from this sample using PAM. If the sample is selected in a fairly random manner, it should closely represent the original dataset.

2.5.3 Hierarchical Clustering Algorithms

These are algorithms where clusters falls into natural layers by producing a tree diagram called a dendrogram as shown below with each leaf being an observation.

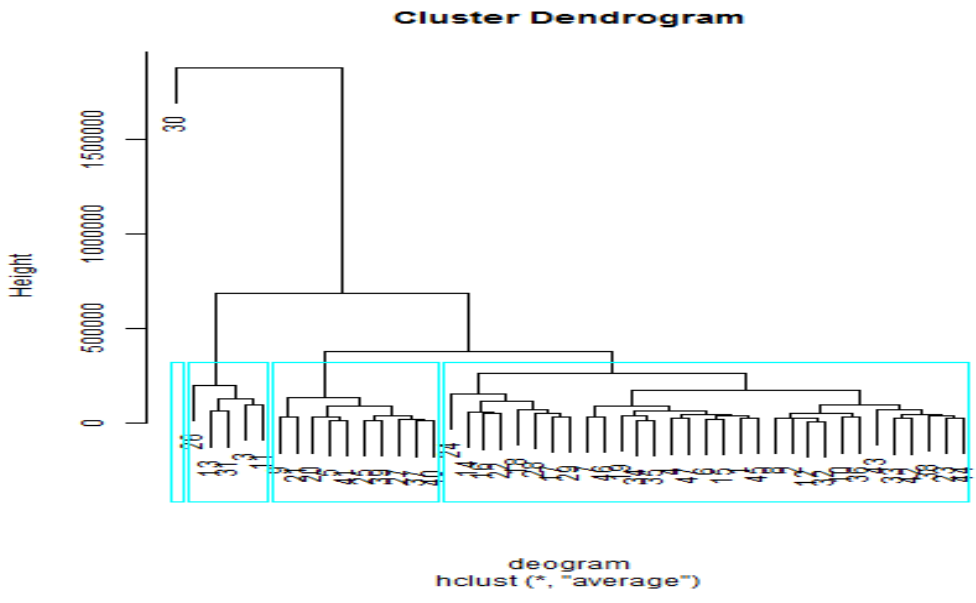


Figure 6-Example dendrogram

Hierarchical clustering algorithm is of two types of which both algorithms are exactly reverse of each other:

i) Agglomerative Hierarchical clustering algorithm or AGNES (agglomerative nesting) and

ii) Divisive Hierarchical clustering algorithm or DIANA (divisive analysis).

2.5.3.1 Agglomerative nesting (AGNES)

This algorithm computes agglomerative hierarchical clustering a dataset. At first, each observation is a small cluster by itself. Clusters are merged (bottom-up) until only one large cluster remains which contains all the observations. At each stage the two nearest clusters are combined to form one larger cluster.

Methods for AGNES clustering

Single link –This refers to the smallest dissimilarity between a point in the first cluster and a point in the second cluster (nearest neighbor method).

Complete link- This refers to the largest dissimilarity between a point in the first cluster and a point in the second cluster (furthest neighbor method).

Average link- In this method, the distance between two clusters is the average of the dissimilarities between the points in one cluster and the points in the other cluster.

2.5.3.2 Divisive Hierarchical clustering algorithm or DIANA (divisive analysis)

In the DIANA clustering algorithm, it constructs a hierarchy of clusters, starting with one large cluster containing all n observations and dividing them (top-down approach) until each cluster contains only a single observation. At each stage, the cluster with the largest diameter is selected. (The diameter of a cluster is the largest dissimilarity between any two of its observations.) To divide the selected cluster, the algorithm first looks for its most disparate observation (i.e., which has the largest average dissimilarity to the other observations of the selected cluster). This observation initiates the "splinter group". In subsequent steps, the algorithm reassigns observations that are closer to the "splinter group" than to the "old party". The result is a division of the selected cluster into two new clusters.

Methods	Advantages	Disadvantages
K-means Clustering	• It is a very simple clustering approach • Efficient • Less Complex	• Requires number of clusters a priori • Works only on numerical data • It is poor in discovering non-convex data. • Sensitive to outliers
Hierarchical methods	• They have an excellent visualization capability • They are easy to implement • Do not require specification of clusters in advance	• Slow due to time complexity • Affected by noise and outliers

2.5.4 Application of Clustering in Health

Geraci, N. S., Mukbel et al (2014), used k-means clustering in profiling of Human Acquired Immunity against the salivary proteins of Phlebotomus papatasi which revealed Clusters of Differential Immunoreactivity where the donors were segregated into four clusters distinguished by unique immunoreactivity profiles to varying combinations of the significantly immunogenic salivary proteins.

Shrivastava, K. (2014), used a modified k-means in medical image segmentation. He used a modified k-means clustering to prove that it gives better results for all performance measuring parameters such as structural similarity index measure, structural content, mean squares error and peak to signal ratio using when used in segmentation of Magnetic Resonance Images.

Messina, J. P. et al (2013), used clustering methodology by employing a 2-step K-means + Ward's clustering algorithm to group hospitals. The final number of clusters was selected using a heuristic that integrated both a statistical-based measure of cluster fit and characteristics of the

resulting Hospital Groups using hospital utilization data. The clustering methodology identified 33 Hospital Groups in Michigan.

2.5.5 Other General Application Areas of Clustering

Clustering has been applied in many areas such as marketing to help the marketers discover customer segments and use this knowledge for developing marketing programs. It has also been used in land use for identification of areas with similar land use in an earth observation data. The insurance firms use clustering in identifying groups of people with similar insurance policies with a high average claim cost so that they can detect fraud. The city planners have used it to identify groups of houses according to their house type, value and geographical locations.

2.6 Application of Data Mining in Health (Global and Kenya).

Several studies have previous researches have been done globally and locally in relation to child health. The following are similar researches that have been done to investigate factors influencing child health in various areas of study. However these studies cannot give an understanding of general health issues since the existing literature is focused on specific areas and no general determinants exists. These studies tended to focus on specific problems. The following are previous studies in child health in Kenya:

2.6.1 Maternal and Child Health Service Utilization

In their study (J.M Nzioki, R.O Onyango, J.H Omboka, 2015) carried out a research to explore the socio-demographic factors influencing Maternal and Child Health (MCH) service utilization in Mwingi district. In their research they used Binary logistic regression model to assess the influence of socio demographic characteristics on MCH service utilization.

The research motivation was not data mining approach and the technique used in this research differs greatly. The research scope was also limited to a particular district- Mwingi which is an administrative representation that is not representative of the current Kenya constitution. The presentation of their finding is also very different from those that will be done in this research.

2.6.2 Factors associated with stunting among children.

In their study (C. Shinsugi et al, 2015) did a research to determine factors associated with stunting among children according to the level of food insecurity in the household in Kwale District. In their research they used a cohort nested on the Health and Demographic Surveillance System (HDSS). The cohort recruited children under the age of six. The household

socioeconomic status (SES) was parameterized by the Principle Component Analysis (PCA) method.

This research was too specific and restricted to one area of finding children stunting and cannot generalized in making public policy in health. The research was also not motivated by Data Mining.

2.6.3 Factors Influencing children's blood pressure.

In their research (P. Wasiewicz et al , 2009) used data mining in the analysis of factors influencing children's blood pressure in a nation-wide health survey in Poland. Questionnaire on socio-economic factors: family structure, parents education, source of income, income category was filled by parents. The questionnaire included also information on child's overall physical activity and exposure to television and computer. In turn, school environment questionnaire was filled by school master or his/her delegate. The questionnaire included information on availability of energy-dense food(s) at school, food advertisements, school canteen and percentage of pupils who eat lunch at school, physical activity infrastructure and its usage. The motivation of this research was data mining. The researchers used Data Mining in the analysis of often large observational data sets to discover unsuspected relationships and to summarize the data in novel ways that was both understandable and useful to the data owner. They used decision tree to visualize their finding. They also used Multidimensional Scaling (MDS) for classification visualization.

2.7 Application of Data Mining in General Health.

As more and more data is being collected and becoming complex, the more there is an increased demand in data mining using the most efficient technologies and techniques in health. Health is spending account for 17% of the US GDP (Hersh W et al, 2011). Data mining has been used intensively and extensively by many organizations. In health, data mining is becoming increasingly popular, if not increasingly essential. Data mining applications can greatly benefit all parties involved in the health industry. For example, data mining can help health insurers detect fraud and abuse, health organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable health services (H.C Koh et al, 2003). Due to the complexity and voluminous nature of the large amount of data generated in health transactions, they cannot be analyzed using the

traditional methods. Methodologies and technologies have been devised in Data mining to transform them into useful information to aid decision makers.

2.7.1 Application of Analysis on gait kinematics

In his study (B. Fioser, 2011), Applied PCA to compare the kinematics of the knee joint during gait, in the frontal and sagittal planes, between a group of elderly women with and without diagnosis in the initial and moderate stages of Osteoarthritis (OA). Data from twenty seven acute and chronic hemiplegic patients were used and compared with data from five healthy subjects. The data was collected during walking along a 10-meter long path. The PCA was applied on a data set consisting of hip, knee, and ankle joint angles of the paretic and the non-paretic leg. The results point to significant differences in joint synergies between the acute and chronic hemiplegic patients that are not revealed when applying typical methods for gait assessment (clinical scores, gait speed, and gait symmetry). The results suggested that the PCA allows classification of the origin for the deficit in the gait when compared to healthy subjects; hence, the most appropriate treatment can be applied in the rehabilitation.

The PCA of joint angles recorded in healthy subjects shows that first two principal components account for about 88% of total variance (from 83% to 94%). First principal component (PC) described about 58%, while second PC described about 30% of total variance.

2.7.2 Prediction of Infertility Treatment Outcome

In the Prediction of infertility outcome study (A. J Milewska et al, 2014) used PCA and Logistic Regression in the prediction of infertility outcome. They used the PCA methods as the first step in analyzing data from IVF (in vitro fertilization). The next step and main purpose of the analysis was to create models that predict pregnancy. Therefore, 805 different types of IVF cycles were analyzed and pregnancy was correctly classified in 61% to 80% of cases for different analyzed groups in obtained models.

2.7.3 Use of data mining to prevalence of prostate cancer

In prevalence of prostate Cancer in Kenya (M.N Ngaruiya, 2014) used data mining to derive patterns which will be used in building a prognostic tool that helps in identification of the Gleason score once screened and advice on the treatment technique. The researcher used data mining tools (R Environment and WEKA) on a dataset containing 485 records and 7 variables. A 10-fold cross-validation was used in model building in comparing ANN and J48. The results showed that ANN is the most accurate predictor compared to J48 in all the instances.

2.7.4 Constructing an Area-based Socioeconomic Status Index

A research was carried (V. Krishnan, 2010) which was focused on development of a socioeconomic index that was used to differentiate disadvantaged areas from more privileged ones in a multivariate context. An index was derived from a Principal Components Analysis (PCA) of 2006 national census data from Alberta, at the Dissemination Area (DA) level where data on 26 variables measuring multiple aspects of socioeconomic status (e.g. income, education, occupation, housing, family and household, ethnicity) were utilized to extract their underlying constructs. Several statistical tests (e.g., KMO, Bartlett's Test of Sphericity) were used to assess the appropriateness of using PCA. Five factors were discovered which together explained 56 per cent of the total variation. Factor scores were utilized to derive standardized indices and quintiles. The PCA-based index suggested that a simple and robust measure, whose values and groupings could only be moderately affected by changes in the socioeconomic landscapes.

Even though most of the findings of these studies are commendable and useful, they were tailored for specific purposes. Assessing the impact of child demographic factors on child health will be incomplete without repeating such investigations in a more holistic view or more general approach. In the view of these gaps, there is a need for a holistic study that will contribute to the understanding of the UNICEF indicators of child health by use of data mining techniques.

2.7.5 Other Applications of Data Mining in Healthcare

i. Effective management of hospital resources- Data mining have been used in effective management of hospital resources by providing models for managing hospital resources. For example, J. Alapont et al proposed a tool for managing hospital resources effectively.
ii. Clustering is used in hospital ranking to analyze various hospital details in order to determine their ranks based on their ability to handle high risk patients.
iii. Recognition of high risk patients. Predictive models are developed to recognize the patients with high risks and improve their health quality and offer cost effective services to their customers.
iv. They are also used in health planning policies in order to improve the health quality as well as reducing the cost of the health services e.g. COREPLUS and SAFS models were developed using data mining

Indicators of Child Health Frameworks

This framework was proposed by Dr. Maheswari Jaiku in 2014.

Figure 7-Determinants of health

UNICEF CHILD HEALTH INDICATORS

DIMENSION	INDICATOR
Child Survival	Under-five child mortality and neonatal mortality
Child Health	Pneumonia, Diarrhea, Malaria and Immunization
Child Nutrition	Malnutrition, low birth weight and breast feeding data
Maternal health	Antenatal care, Delivery care and mortality rate
Water and sanitation	Water, sanitation and hygiene
Education	Literacy, primary and secondary education
Child protection	Birth registration, child labor, child marriage, FGM
Early Child Development	
Child disability	
HIV and AIDS	

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

Explanatory research design was used in this research. It began from the exploratory perspective where the researcher explored on the new idea identified to seek more information about the idea. This led to groundwork of more future research and investigated whether the findings coud be defined by the current existing theories. Descriptive statistics such as the correlation matrix, mean, standard deviations, principal component analytics and visualizations was used to explain the knowledge discovered in the research.

3.2 Research Design

In this research, CRISP-DM methodology was used. There are several Data mining methodologies such as CRISP_DM, SEMMA, and KDD that exist. The choice of this methodology was due to its acceptance in data mining and also because the model is designed as a general model and can be applied in a variety of fields industry and business problems. According to the 2014 KDD nuggets survey, the popularity rose from 42% in 2007 research to 43% in 2014 making it the most popular data mining methodology. SAS SEMMA popularity dropped from 13% in 2007 to 8.5% in 2014 while KDD Process model Popularity rose from

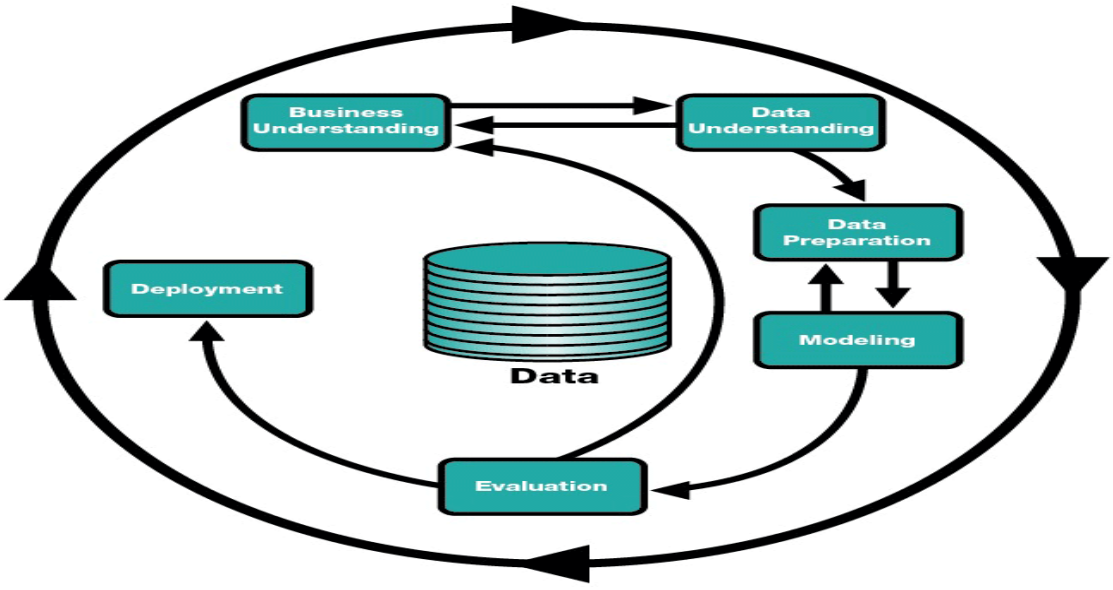


Figure 8 -CRISP-DM Process model

7.3% in 2007 to 7.5 % in 2014 (J.Taylor,2014).Source: CRISP-DM.

Available from: http://crisp-dm.eu/reference-model/

3.3 Overview of CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) that is extensively used process in data mining. The model is made up of steps intended as a cyclical process as shown in figure above.

i. **Business Understanding:** This step determines the business objectives, assessing the existing situation, establishing data mining goals, and developing a project plan.

ii. **Data Understanding:** After business objectives and the project plan have been established, data understanding then considers the data requirements. This includes initial data collection, data description, data exploration, and the verification of data quality. The data is explored and a summary statistics presented (This includes visual presentation of the categorical variables). Cluster analysis models are applied at some point in this stage, intention being the identification of patterns in the data.

iii. **Data Preparation:** On identifying the available resources, they are then selected, cleaned, built into desired form, and formatted. Data cleaning and data transformation in preparation of data modeling occurs at this stage. In depth data investigation at this stage and supplementary models are utilized. This provides an opportunity to observe patterns based on business understanding.

iv. **Modeling:** Data mining software tools such as visualization (Abstracting data to improve human recognition by plotting data and establishing their relationships) and cluster analysis (identification of variables that are related) are useful for primary analysis. Generalized rule induction tools can develop initial association rules. After greater data understanding is gained, more detailed models appropriate to the data type can be applied. Data needed for modeling is divided into training and test sets.

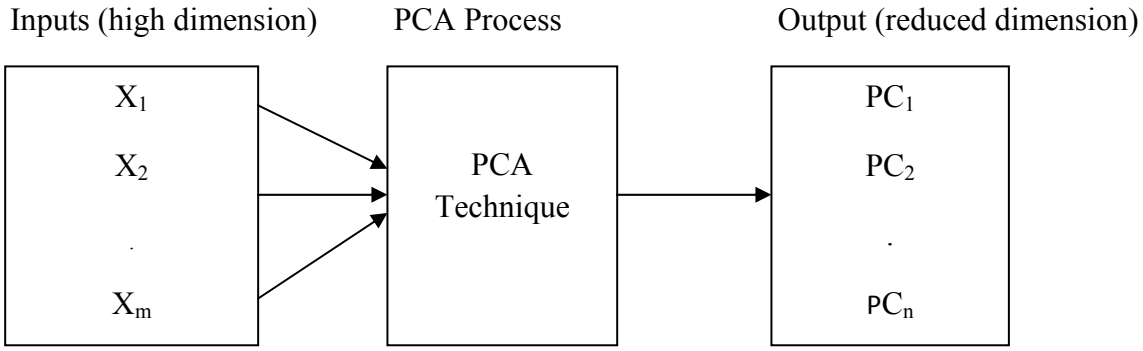
v. **Evaluation:** The model outcome is evaluated in the context of the business objectives established in the business understanding stage. This will leads to the identification of other needs through pattern recognition. The process then iterated to the first step of the CRISP-DM process to gain business understanding. New relationships that provide a deeper understanding of organizational operations are shown through visualization, statistical, and artificial intelligence tools.

vi. **Deployment:** Data mining can verify previously held hypotheses and for identification of useful knowledge. Sound models can be obtained from knowledge discovered in the

previous stages of the CRISP-DM process. The models are then monitored for modifications in the operating environment, because they vary with time. Any significant change occurring means that the model should be redone. The results of data mining projects should be documented for future reference.

CRISP-DM methodology is flexible and all phases need not to be applied by experienced analysts. The methodology was chosen due to the flexibility and great deal of backtracking.

3.4 PCA Model



Where, $n \leq m$

Figure 9-PCA model

PCA assumes that variables are linearly related and does not have any model for testing. PCA Analysis is like having a different viewpoint for the same data set. The viewpoint is changed by moving the origin of the coordinate system to the centroid of the data and then rotating the axes.

Consider a set of n variables (X_1, ..., X_m), PCA calculates a set of n linear combinations of the variables (PC_1, ..., PC_n) such that:

i. The total variation in the new set of variables or principal components is the same as in the original variables.
ii. The first PC contains the most variance possible, e.g. as much variance as can be captured in a single axis.
iii. The second PC is orthogonal to the first one (their correlation is 0), and contains as much of the remaining variance as possible.
iv. The third PC is orthogonal to all previous PC's and also contains the most variance possible.
v. Etc.

The above process is accomplished by calculating a matrix of coefficients where columns are referred to as eigenvectors of the variance-covariance or of the correlation matrix of the data set. The fundamental consequences of the process are that:

i. The entire original variables are involved in the computation of PC scores (i.e. the position of every observation in the new set of axis formed by the PC's).
ii. The sum of variances of the PC's equals the sum of the variances of the original variables when PCA is based on the variance-covariance matrix, or the sum of the variances of the standardized variables when PCA is based on the correlation matrix.
iii. There are n eigenvalues (n=number of variables in the data), each eigenvalues is associated with an eigenvector and a PC. Each eigenvalues is the variance of the data in each PC. Therefore, the sum of eigenvalues based on the variance-covariance matrix is equivalent to the summation of variances of the original variables.

PCA uses the correlation matrix which is similar to using PCA based on the variance-covariance of the standardized variables. Since standardized variables contain variance equal to 1, the totals of the eigenvalues is n, the number of variables.

3.5 Source of data and study Population

Secondary data was collected from Kenya National Bureau of Statistics, Commission of Revenue Allocation, Kenya HIV and AIDS profile per county, Statistical Abstract 2014, Kenya Economic report of 2014, and Kenya County Profile, Kenya Demographic and Health Survey of 2014 and e-health facilities.

The major demerit of secondary data collected by other researchers was that they controlled, decided what to collect and what to exclude and therefore the entire information desired for this research may not be available.

3.6 Data Analysis Tools and Presentation

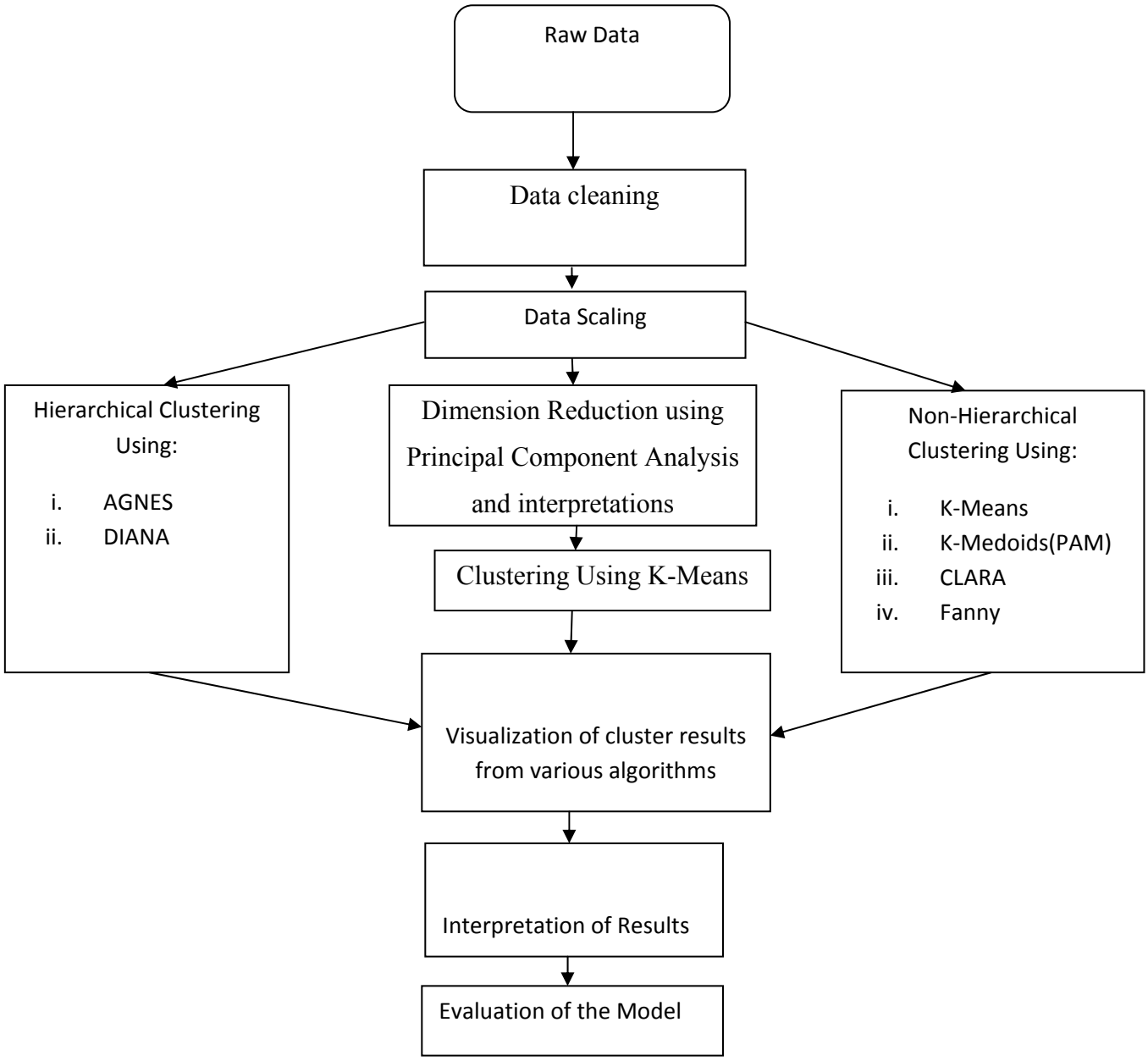
R studio was used for Analysis. The choice of the tool was because R is exceptional statistical software for analysis as it includes a broad range of analyses employed in data mining and machine learning analysis, as well as numerous routines for exploratory data analysis (EDA). The tool is open source and its implementation available for many systems for free.

R also has a powerful graphics capability. R supports a fairly broad range of graphic devices in addition to excellent on-screen plotting. Reflecting its origins on UNIX computers, it is quite good at Postscript output, but also includes other formats.

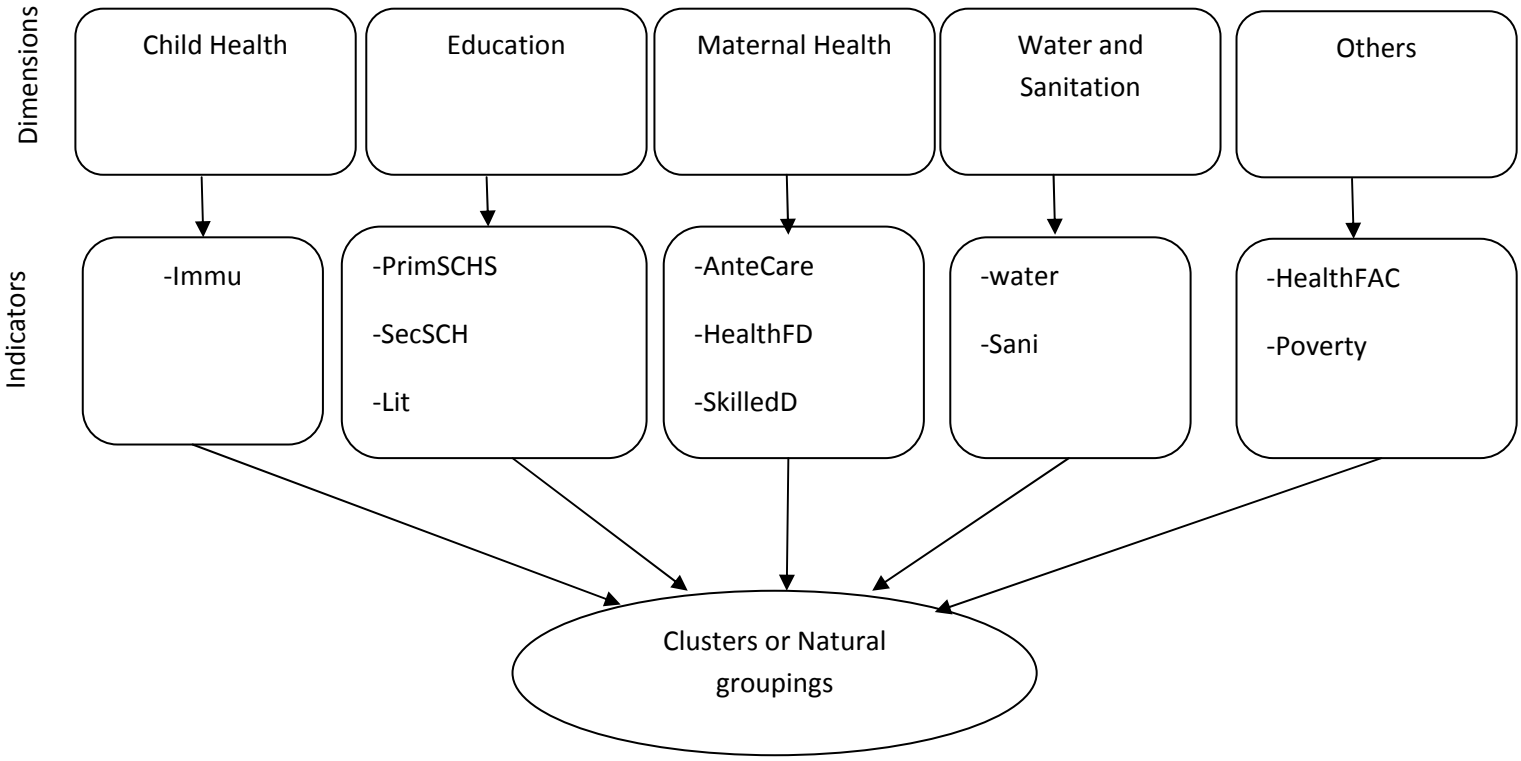
3.7 Data Analysis Methods, Justification and Limitation

This research proposed the use of Principal Component Analysis for data reduction and Clustering algorithms because it has a mathematical foundation using applied multivariate methods for analysis and produces results that can be proved mathematically. PCA and clustering being multivariate analysis techniques, enables the researchers is be in a position to study the effects of several variables acting concurrently instead of one by one. They can also be used in creating predictive models.

3.8 Proposed Framework



3.9 The Research Framework



Meaning of indicators abbreviations

-Immu	-	Immunization distribution
-PrimSCHS	-	Number of primary schools
-SecSCH	-	Number of secondary schools
-Lit	-	Literacy Level
-AnteCare	-	Occurrence seeking Antenatal care
-HealthFD	-	Occurrence delivered in health facilities
-SkilledD	-	Occurrence delivered by skilled attendants
-water	-	Water distribution
-Sani	-	Sanitation distribution
-HealthFAC	-	Number of health Facilities
-Poverty	-	Poverty inverse

CHAPTER 4: DATA NALYSIS AND DISCUSSION

4.1 Data Preprocessing

The data set was collected from literature of secondary data issued by the Kenya National Bureau of Statistics, Commission of Revenue Allocation, Kenya HIV and AIDS profile per county, Statistical Abstract 2014, Kenya Economic report of 2014, and Kenya County Profile, Kenya Demographic and Health Survey of 2014 and e-health facilities. The researcher eliminated any unusable data before transcribing.

4.1.1 Dataset Description

Variable name	Meaning	Data type
County	Name of the county	String
PrimSCHS	Number of primary schools	int
SecSCH	Number of secondary schools	int
HealthFAC	Number of health Facilities	int
AnteCare	Occurrence seeking Antenatal care	num
SkilledD	Occurrence delivered by skilled attendants	num
HealthFD	Occurrence delivered in health facilities	num
Poverty	Poverty inverse	num
Sani	Sanitation distribution	num
Immu	Immunization	num
Lit	Literacy Level	num
Water	Water distribution	num

Figure 10-Data description table

The dataset used in this research is made up of twelve variables with one qualitative and eleven quantitative attributes

4.1.2 Modeling Tools and Techniques

We used R with RStudio IDE which is a powerful tool used in machine learning, data mining and statistical analysis due to its productive user interface. RStudio is free and open source, and compatible with many operating system platforms such as Windows, Mac, and Linux. R studio comes with several packages such as cluster, animation, ggplots, initr, stats, and shinyapps. R is also very powerful in Visualization. We also used excel to store our data set in comma-separated value and also for saving our R results in CSV file format.

4.1.3 Data Exploration

We first checked the size, structure, dimension and the names of our dataset.

```
> dim(datascale)
[1] 47 11
```

Figure 11-Number of instances and attributes

Results

Our dataset was made up of 47 observations and 11 attributes. The attribute county was not included because it was to be used for labeling and not for analysis.

```
> str(datascale)
'data.frame': 47 obs. of  11 variables:
 $ PrimSCHS : num  0.0683 0.2025 1.0244 -0.7451 -0.8122 ...
 $ SecSCH   : num  -0.3491 0.00867 0.73415 -0.45842 -0.62737
 $ HealthFAC: num  0.0824 -0.5336 -0.2483 -0.6763 -0.579 ...
 $ AnteCare : num  -0.0476 0.0233 0.4387 0.4387 0.4894 ...
 $ SkilledD : num  -0.2068 -0.2888 -0.842 0.0339 0.3668 ...
 $ HealthFD : num  -0.1781 -0.4168 -0.8245 0.0506 0.3638 ...
 $ Poverty  : num  -0.211 0.556 0.107 -0.635 0.197 ...
 $ Sani     : num  -0.789 0.67 0.735 0.558 0.142 ...
 $ immu     : num  -1.885 -0.532 2.055 -0.209 0.63 ...
 $ Lit      : num  0.024 0.395 -0.317 -0.507 0.54 ...
 $ water    : num  -1.885 -1.512 1.643 1.25 -0.115 ...
>
```

Figure 12-Dataset datascale structure

4.1.4 Data Cleaning

The data provided to us from the various organizations was raw data with various impurities and was not in the form that we wanted. We therefore had to filter all that was of importance to our research. We then transcribed data from the literature to excel and saved it in Comma-Separated Values format required by our research for analysis.

4.1.5 Missing Values

We used the function “na.omit” to handle the not available (NAs) or the missing values by returning the object with incomplete cases removed. The importance of this was to remove all cases that had missing data anywhere in the data set as missing values could create problems for simple and complicated analyses.

4.2 Data Transformation

4.2.1 Scaling

We did normalization because our variables had different measurements using the scale() function. We normalized the dataset so that we could identify the structure regardless of the measurement scale it was taken. This was achieved by centering each variable. In other words,

subtract the mean from each variable divided by the standard deviation. The results produced a standardized value with the property that the sample variance matrix was the same as the correlation matrix.

$z = (X - \mu) / \sigma$

Where z is the z-score, X is the value of the element, μ is the population mean, and σ is the standard deviation.

```
> attributes(pcahealth)
$names
 [1] "PrimSCHS"  "SecSCH"    "HealthFAC" "AnteCare"  "SkilledD"
 [6] "HealthFD"  "Poverty"   "Sani"      "immu"      "Lit"
[11] "water"

$class
[1] "data.frame"
```

4.2.2 Principal Component Analysis

The first step was the dimensionality reduction method to make the information easier to visualize and analyze. We used the Principal Component Analysis (PCA) techniques for easier visualization of our high-dimensional set. On naively applying PCA to raw data and plotting, we saw that the first Principal component standard deviation was around 202% which accounted for 82% of the variance of the data. This was as a result of the differences in the variable measurements.

4.2.2.1 Verification of variance

To verify this, we plotted the variance of the columns and corrected this by using of the R scale() function and then verified the equality of variances for the different variables so that one variable could not dominate on applying the PCA. After applying the scale() function, there was constant variance across the variables.

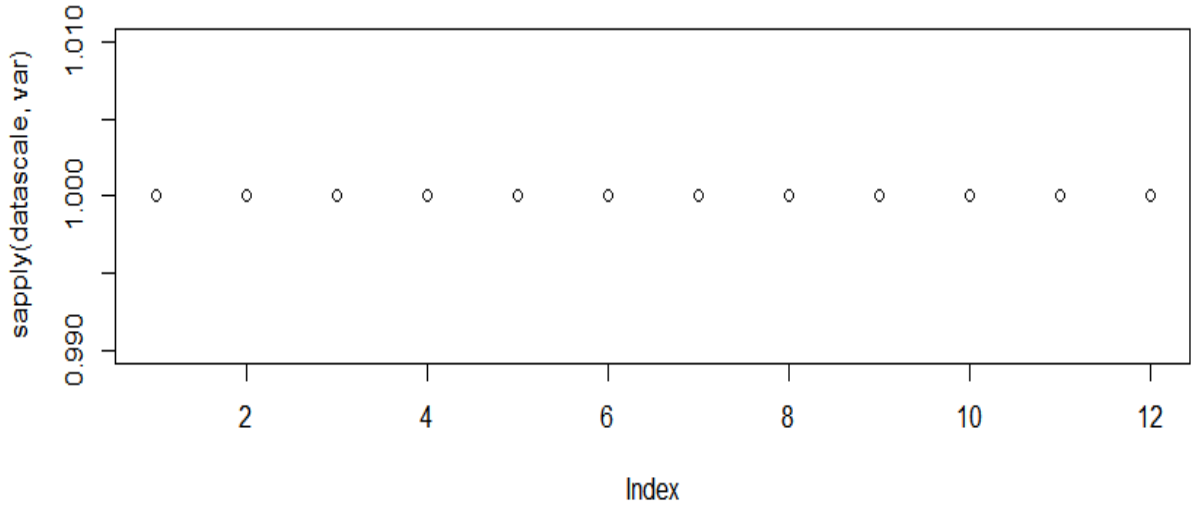


Figure 13-Verification of Variance plot

4.2.2.2 Bar and Line Screeplot

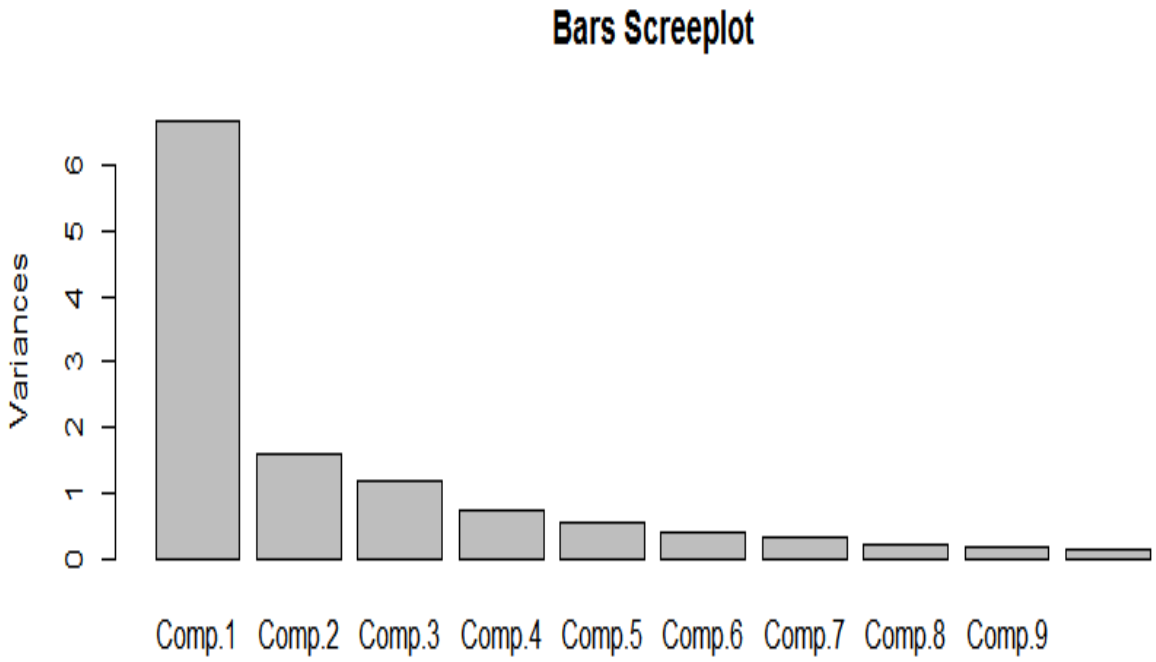


Figure 14-Bar Screeplot

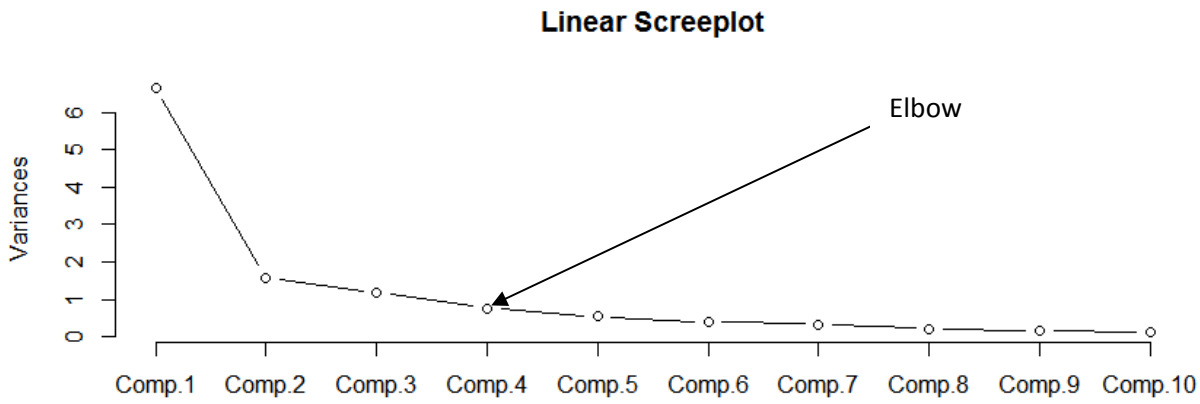


Figure 15-Line Screeplot

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4
Standard deviation     2.4300419 1.2455078 1.0795081 0.8526294
Proportion of Variance 0.5368276 0.1410263 0.1059398 0.0660888
Cumulative Proportion  0.5368276 0.6778539 0.7837937 0.8498825
                           Comp.5     Comp.6    Comp.7     Comp.8
Standard deviation     0.70710391 0.61381655 0.5745530 0.45712035
Proportion of Variance 0.04545418 0.03425189 0.0300101 0.01899627
Cumulative Proportion  0.89533671 0.92958860 0.9595987 0.97859497
                           Comp.9     Comp.10      Comp.11
Standard deviation     0.38420929 0.277970499 0.1028151422
Proportion of Variance 0.01341971 0.007024327 0.0009609958
Cumulative Proportion  0.99201468 0.999039004 1.0000000000
> |
```

Figure 16-Summary of importance of components

Results

We created the principal component for our dataset and plotted a Screeplot with a summary of our findings. The first four components in the Screeplot explained 85% of variance. We used the rule of thumb to select the number of principal components that were to be retained for our research. The rule of thumb can either be by picking the number of components that explains 85% of variance or greater or the Screeplot elbow. We retained the first four PC. We placed the

results into a new data frame and plotted by use of prcomp instead of princomp. The Screeplot plots the variances against the number of the principal component.

4.2.2.3 2-D and 3-D Plots

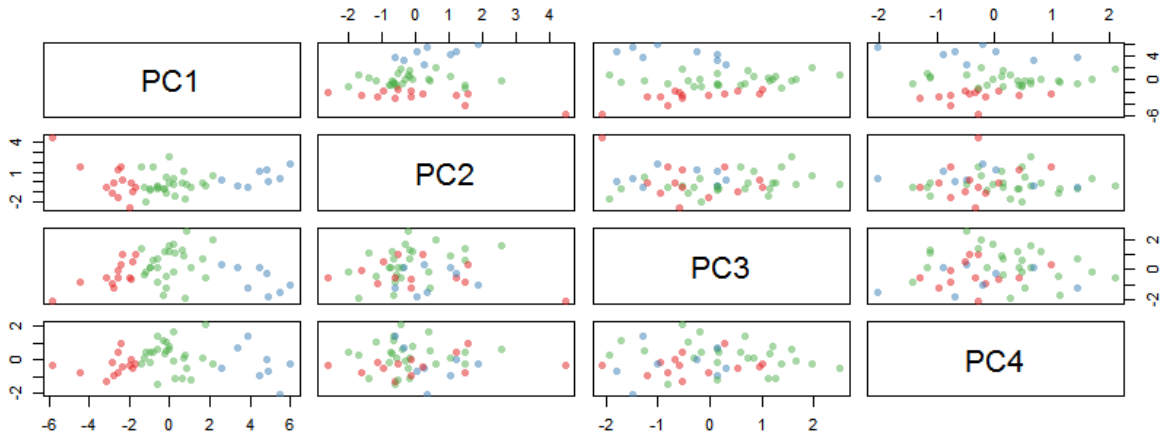


Figure 17-Correlation Matrix of the First Four PCs

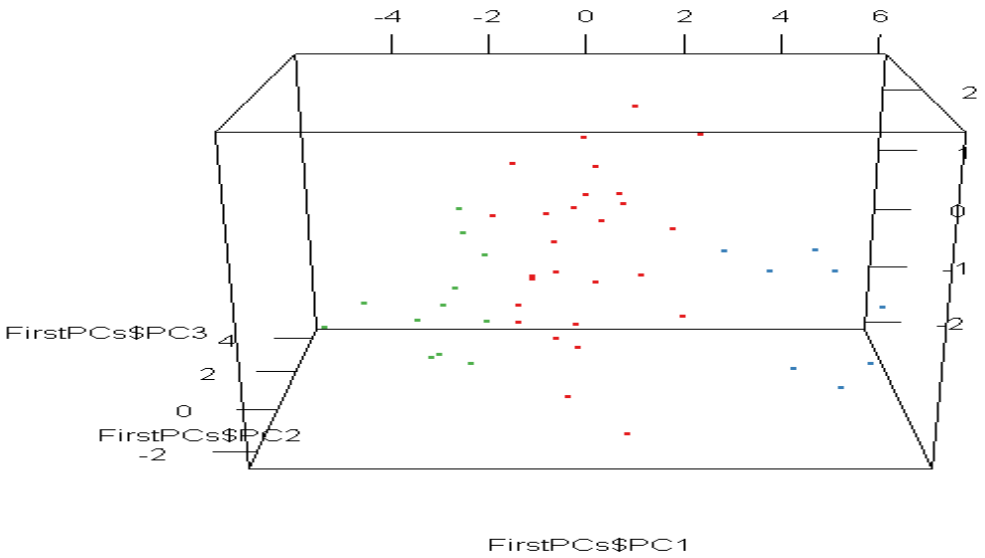


Figure 18-3-Dimension View of PC1, PC2 and PC3

Results

The figure 17 shows the 2-D projection of data which are on a 4-D space as it is easier to visualize than 3-D. We used 3-D (figure 18) to have an interactive visualization to allow us to

explore the space and avoided loosing meaning by collapsing the space into 2-D. By simplifying our complex dataset into a lower dimensional space, we were able to visualize, work and find patterns in the counties that were similar in child health status by use of the k-means unsupervised clustering algorithm.

The PCA enabled us to use the variations in our dataset which was described by 11 variables. By doing this we were able to reduce the 11 dimensions into 2 because more than three variables in the data set could have been very difficult in visualizing a multidimensional hyperspace. The initial variables were transformed into a new set of variables which was used to explain the variation in the data. These variables corresponded to linear combination of the originals and are called principal components. The PCA reduced the dimensionality of our data to two which could be visualized graphically with minimal loss of information.

4.2.2.4 Scatter plot

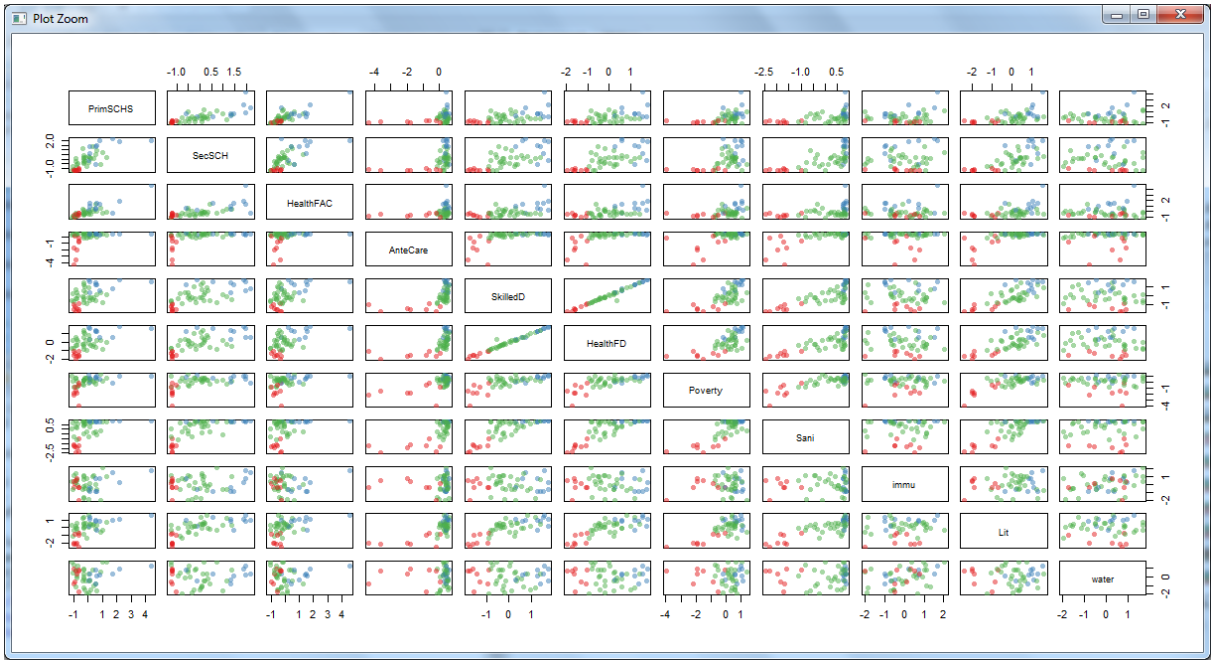


Figure 19-Scatter plot diagram

We did a scatter plot matrix to visualize all our variables. The scatter plot showed both positive and negative correlations. There was a remarkably almost linear positive correlation between skilled deliveries and health facilities deliveries variables. There was a negative correlation

between skilled delivery, health facility delivery & poverty with water. There was also a negative correlation between skilled delivery & health facility delivery with immunization.

4.2.2.5 Biplot

A biplot refers to an enhanced scatterplot that is used to display both points and vectors to represent structure of a dataset. It is used in Principal Component Analysis, where the axes of a biplot are a pair of principal components. These axes are labeled as Comp.1 (PC1) and Comp.2 (PC2) in our diagram. The biplot is used to represent the scores of the observations on the principal components. Vectors are used to represent the variables on the principal components. Points in these case are used to represent the counties and whereas the vectors represent the indicators of child health. The biplot shows vectors direction and length with pointers pointing away from the origin following some direction. The vector direction shows squared multiple correlations with the principal components. The length of the vector represents the proportional to the squared multiple correlation between the fitted values for the variable and the variable itself. Observations pointed furthest in the direction with most of what that variable measured, with those pointing in the middle having average amount and those pointing in opposite direction having the least. All vectors pointing in the same direction had similar influence by the child health indicators.

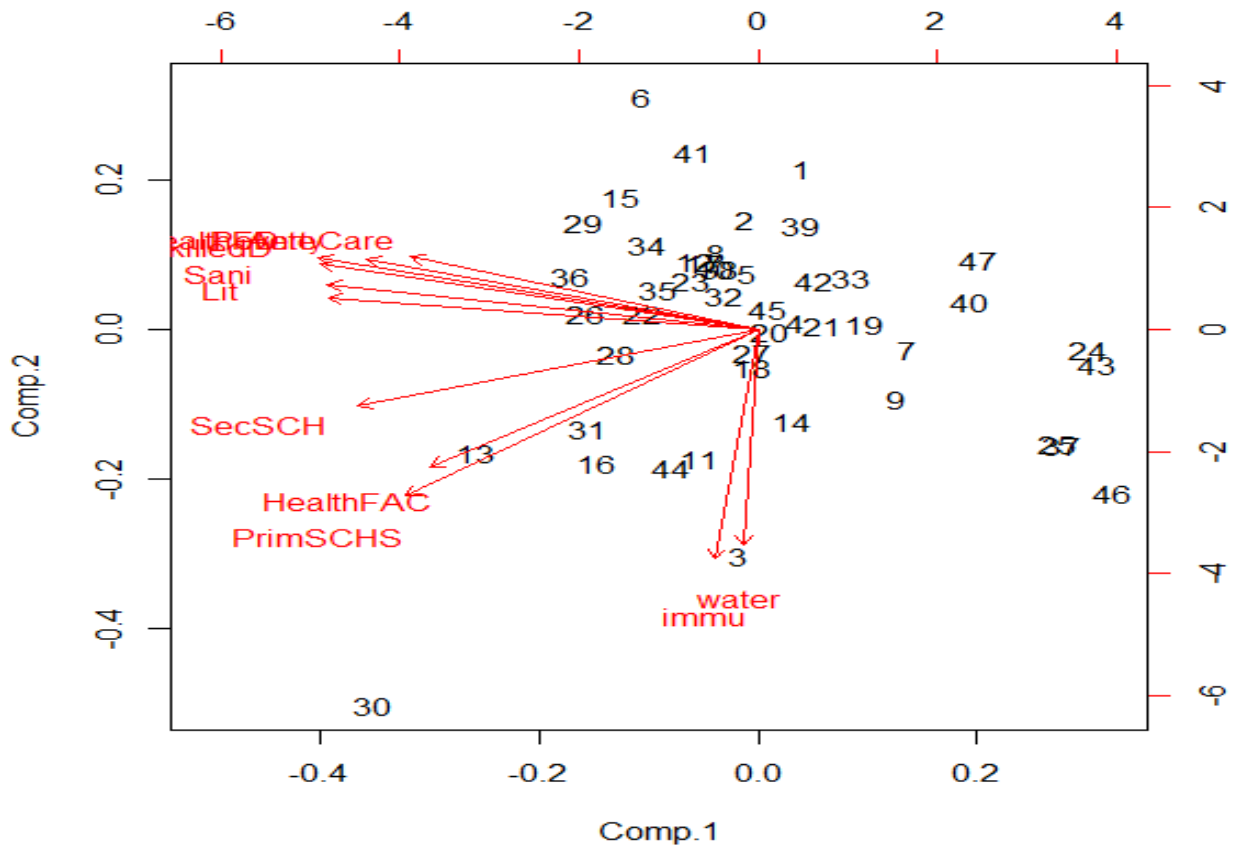


Figure 20-Biplot showing scores and loadings

Results

The relative locations of points that were close together were those counties that had similar scores on the components displayed in our plot. These components fitted well to our data and points corresponded to observations that had similar values on the variables. Counties that were close together had similar indicators of child health. The indicators HealthFAC and PrimSCHS rated Nairobi, Kiambu, Nakuru and Kisii counties highly. The counties of Kirinyaga, Nyamira, Murang'a and Embu were also rated highly although these points were far apart. The loading showed that the most influence in the highly rated counties was contributed by the variables SecSCH, HealthFAC and priSCHS. The position of the observation Bungoma County was mostly influenced by the variables water and immu with average influence on the county of Kitui. The counties of Nyeri, Meru and Murang'a were moderately influenced by the variables HealthD, HealthFAC, AnteCare, SkilledD, Sani, Lit and Poverty.

4.2.2.6 Correlation Matrix

	PrimSCHS	SecSCH	HealthFA(	AnteCare	SkilledD	HealthFD	Poverty	Sani	immu	Lit	water
PrimSCHS	1	0.747	0.823	0.319	0.515	0.513	0.437	0.449	0.228	0.516	0.147
SecSCH	0.747	1	0.683	0.448	0.621	0.624	0.477	0.675	0.137	0.671	-0.012
HealthFAC	0.823	0.683	1	0.264	0.523	0.524	0.363	0.359	0.063	0.446	0.095
AnteCare	0.319	0.448	0.264	1	0.556	0.581	0.608	0.693	0.017	0.675	0.033
SkilledD	0.515	0.621	0.523	0.556	1	0.989	0.688	0.769	-0.059	0.749	-0.038
HealthFD	0.513	0.624	0.524	0.581	0.989	1	0.689	0.772	-0.074	0.755	-0.056
Poverty	0.437	0.477	0.363	0.608	0.688	0.689	1	0.784	0.063	0.7	-0.068
Sani	0.449	0.675	0.359	0.693	0.769	0.772	0.784	1	0.061	0.783	0.11
immu	0.228	0.137	0.063	0.017	-0.059	-0.074	0.063	0.061	1	0.125	0.337
Lit	0.516	0.671	0.446	0.675	0.749	0.755	0.7	0.783	0.125	1	-0.026
water	0.147	-0.012	0.095	0.033	-0.038	-0.056	-0.068	0.11	0.337	-0.026	1

Figure 21-Correlation Matrix

2.2.7 Score and Loading plots

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
BARINGO	0.674057	1.846349	-1.85987	0.193323	-0.61633	0.260921	-0.2532	-0.42937	-0.32043	0.020022	0.020567
BOMET	-0.19341	1.261798	-0.32473	1.035381	-0.46205	0.299304	0.762041	-0.18619	-0.38256	0.519756	0.108366
BUNGOMA	-0.27907	-2.59091	1.560569	0.477063	-0.72153	0.014673	0.695571	0.634565	-0.40007	0.140616	-0.02584
ISIOLO	2.113295	-0.78637	1.908519	-0.24433	-0.07307	-0.15885	-0.78408	-0.19249	-0.05258	-0.18801	0.027076
KAJIADO	-0.66262	0.702363	-0.7718	-1.72724	-0.454	1.781788	0.172824	0.090272	-0.50825	-0.33717	0.023973
KAKAMEGA	-0.88566	-1.47847	0.588247	0.413616	-0.75576	-0.4593	0.880624	0.211412	0.433373	-0.06492	0.029858
KERICHO	-0.94046	0.780089	0.236773	0.362419	-0.0642	0.039683	0.18749	-0.15647	0.140363	0.039961	0.077074
KIAMBU	-4.29763	-1.40646	-0.89405	-0.70575	0.495904	-0.02713	-0.03161	0.142082	-0.35984	-0.44047	-0.03478
KILIFI	0.50931	-1.0433	0.853334	0.037808	-0.18625	-0.12456	-0.9768	-0.15292	0.38519	0.049783	-0.02075
KIRINYAGA	-2.09818	1.51688	-0.07049	-0.58068	1.223285	0.356484	-0.33683	0.501005	-0.10656	0.043267	0.000481
MAKUENI	-1.0188	0.558775	-0.32366	0.190954	-1.01696	-1.01379	0.61668	-0.6583	0.723903	0.166941	0.056976
MANDERA	5.038306	-0.23755	-2.05442	0.013171	2.371683	-0.58034	0.996108	-0.26261	-0.07453	0.42858	-0.03706
NAIROBI	-5.84923	-4.29295	-2.22832	-0.17139	0.117955	0.982442	-0.8992	-0.38036	-0.02174	0.523783	0.009997
NAKURU	-2.58506	-1.14278	-0.53077	0.451573	-0.07225	-0.46237	0.458627	0.115487	0.648862	-0.33345	-0.01825
NYERI	-2.8467	0.612058	-0.56699	-1.23268	0.570397	-0.18185	-0.30212	-0.60155	0.725097	0.038555	-0.023
SAMBURU	4.615722	-1.32757	-0.37686	0.251903	0.62402	-0.2016	-0.16215	0.181225	-0.08479	-0.02064	0.080384
SIAYA	-0.57459	0.690568	0.754096	1.488334	0.774034	-0.09707	-0.56996	0.72657	0.594253	-0.1848	0.032889
TAITA TAVETA	0.660931	1.184724	1.048849	-1.03473	0.16015	0.207268	0.037589	0.995084	0.107614	0.278527	0.019508
TURKANA	5.165136	-0.39468	-1.61172	-2.1365	-1.69959	-1.30834	-1.15213	0.511363	-0.35044	0.265899	0.00264
UASIN GISHU	-1.31117	-1.57422	1.308374	-0.29784	-0.08872	0.30765	0.447338	-0.37301	-0.81851	0.155234	0.010108
VIHIGA	0.139838	0.234224	1.721816	-0.20422	-0.27141	0.116128	0.692051	-0.02623	0.291808	0.058454	-0.02641
WAJIR	5.38541	-1.87662	-1.10515	-0.04596	1.39389	0.328175	0.71035	-0.44429	0.264434	-0.37088	-0.02008
WEST POKOT	3.341799	0.801711	-1.11961	1.220477	-0.57251	0.639237	-0.10552	-0.2337	-0.02688	-0.13125	0.02043

Figure 22-Scores plot

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
PrimSCHS	-0.2954	-0.39801	-0.33134	0.037128	-0.13392	0.219357	0.001143	-0.00011	-0.73825	0.179157	-0.00201
SecSCH	-0.3351	-0.18285	-0.23694	0.150741	-0.20185	-0.47861	0.442356	0.2277	0.121969	-0.49315	-0.00075
HealthFAC	-0.27464	-0.32851	-0.47648	-0.13167	-0.08947	0.248679	-0.24487	-0.07088	0.63158	0.196202	0.008172
AnteCare	-0.29145	0.175183	0.328702	-0.01012	-0.67895	-0.0068	-0.49258	0.256084	-0.02329	-0.07285	0.033753
SkilledD	-0.36725	0.15883	-0.02971	-0.16625	0.462259	-0.13082	-0.24308	0.136169	-0.07911	-0.08393	0.699749
HealthFD	-0.36907	0.173137	-0.03314	-0.16051	0.417599	-0.13425	-0.26924	0.149198	-0.07155	-0.07642	-0.71332
Poverty	-0.32814	0.169681	0.201454	0.145819	0.082982	0.749071	0.30441	-0.01333	0.075575	-0.36749	-0.0095
Sani	-0.36074	0.108214	0.269342	-0.0645	-0.02433	-0.09298	0.469528	0.213371	0.109951	0.70295	-0.00076
immu	-0.03633	-0.5494	0.410115	0.607658	0.26976	-0.04455	-0.21805	0.151697	0.107048	0.050846	-0.00618
Lit	-0.36015	0.076103	0.143952	0.166233	-0.06276	-0.23244	-0.03221	-0.86994	-0.00657	0.01724	-0.00326
water	-0.01238	-0.51596	0.441555	-0.6948	0.009397	-0.00556	0.09186	-0.10602	0.000826	-0.18959	-0.01307

Figure 23-Loading plot

Results

The score plot is a summary of the relationship among observations (samples) while is the loadings is a summary of the variables used as a means for interpreting the pattern seen in the score plot.

4.3 More Exploratory Data Analysis

4.3.1 Summary Statistics

PrimSCHS	SecSCH	HealthFAC	FertRate	AnteCare	SkilledD	HealthFD	Poverty	Sani	immu	Lit	water
Min. : 15.0	Min. : 11.0	Min. : 47.0	Min. :2.300	Min. :50.50	Min. :21.70	Min. :18.30	Min. :32.50	Min. :13.30	Min. :30.90	Min. :18.10	Min. :33.60
1st Qu.: 62.0	1st Qu.: 57.5	1st Qu.:132.0	1st Qu.:3.450	1st Qu.:93.55	1st Qu.:45.00	1st Qu.:43.00	1st Qu.:77.80	1st Qu.:63.45	1st Qu.:51.25	1st Qu.:56.05	1st Qu.:53.70
Median :124.0	Median :125.0	Median :178.0	Median :4.200	Median :96.70	Median :54.60	Median :57.40	Median :84.80	Median :90.60	Median :62.40	Median :70.40	Median :66.40
Mean :151.9	Mean :138.1	Mean :222.3	Mean :4.357	Mean :93.27	Mean :57.84	Mean :57.38	Mean :80.98	Mean :77.50	Mean :61.26	Mean :66.82	Mean :63.85
3rd Qu.:203.5	3rd Qu.:188.5	3rd Qu.:272.5	3rd Qu.:5.050	3rd Qu.:97.75	3rd Qu.:71.60	3rd Qu.:69.65	3rd Qu.:88.30	3rd Qu.:97.95	3rd Qu.:72.50	3rd Qu.:80.25	3rd Qu.:75.85
Max. :680.0	Max. :360.0	Max. :935.0	Max. :7.800	Max. :99.20	Max. :92.60	Max. :93.40	Max. :97.50	Max. :99.70	Max. :92.40	Max. :98.80	Max. :89.30

Figure 24-Summary statistics

Results

The 1st quantile represents 25% while the 3rd quantile represents 75%. We used summary which is a generic function used to produce result summaries of the results of various model fitting functions such as min, median, mean and maximum. For example the feature vector skilled delivery can be interpreted that the minimum percentage county women seeking skilled delivery is ~22% with the maximum being ~93%.Approximately 55% of women in all the counties seek skilled delivery. Out of the 25% of the first quantile, below 45% women seek skilled delivery

while 55% seeking for alternative methods and the 3rd quantile of 75%, women below ~72% seek for skilled delivery with the remaining 28% seeking for alternat.ive methods of delivery.

4.3.2 Histogram Plots

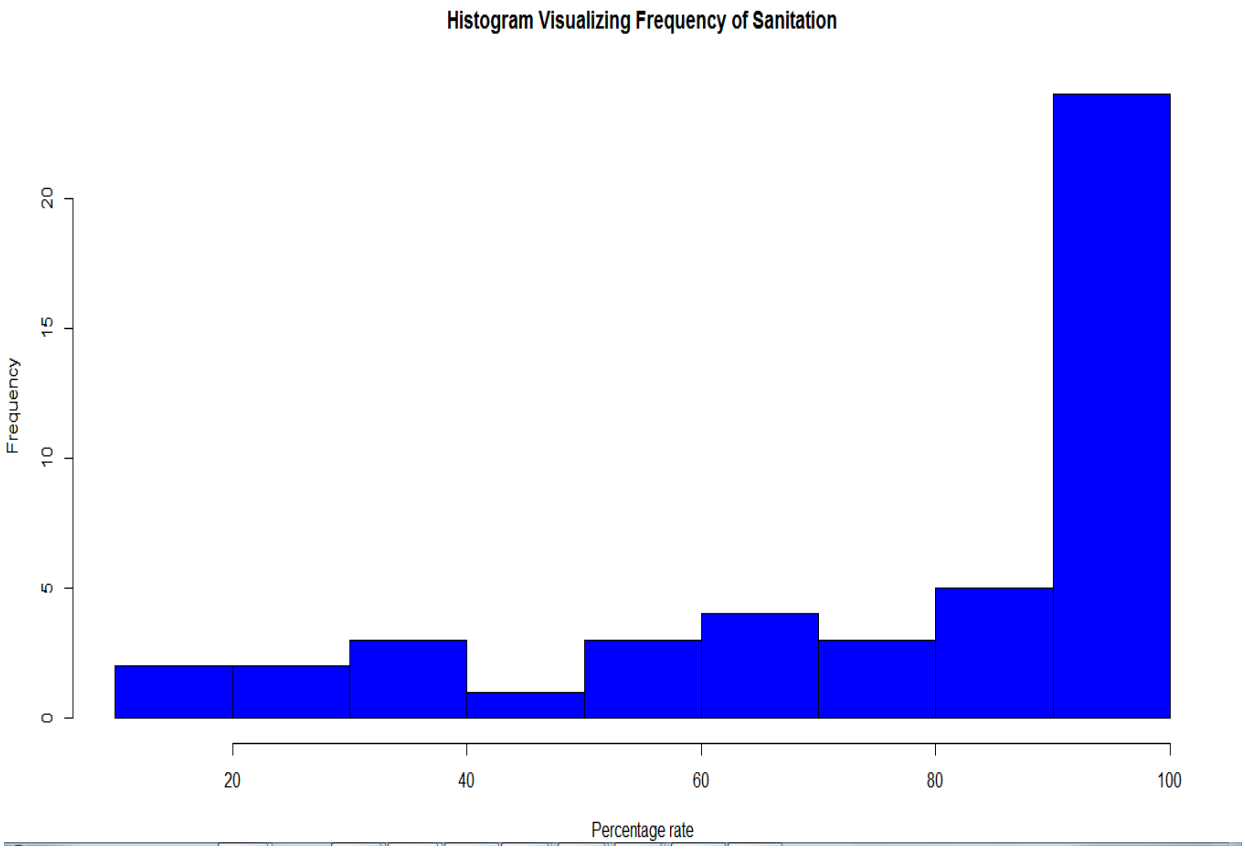


Figure 25--Histogram for Sanitation

We used histograms to give an idea of what different values are.

Results

The histogram is a plot of the frequency of sanitation against the percentage rate. It tells us that 20 counties have sanitation facilities of more that 90% whereas less than five counties have the sanitation facilities below 20%.

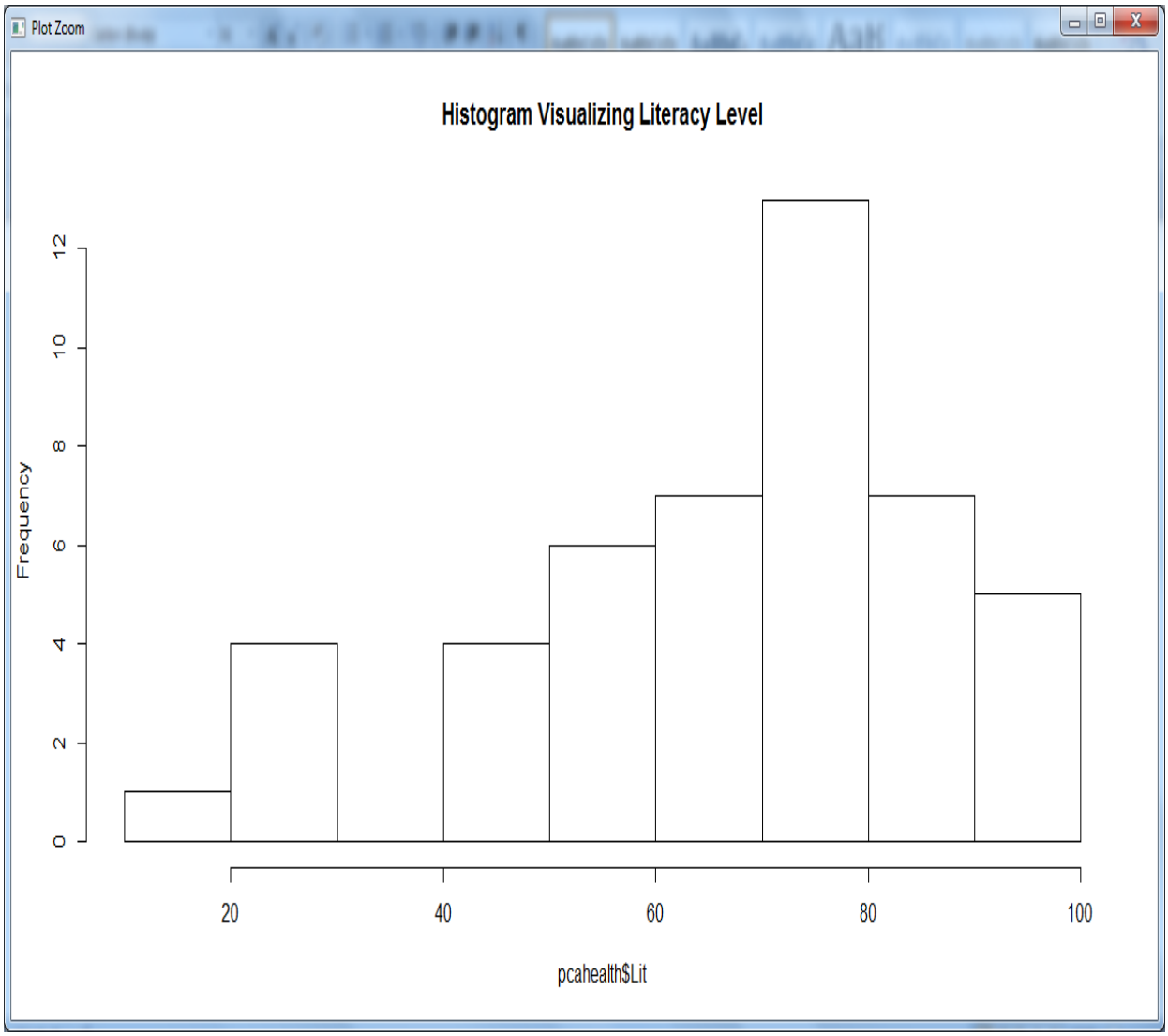


Figure 26-Histogram for Literacy

Results

The histogram depicts that five counties have a literacy level below 30% whereas 12 counties literacy level is above eighty percent. This means most counties are doing well academically.

4.3.3 Density Plots

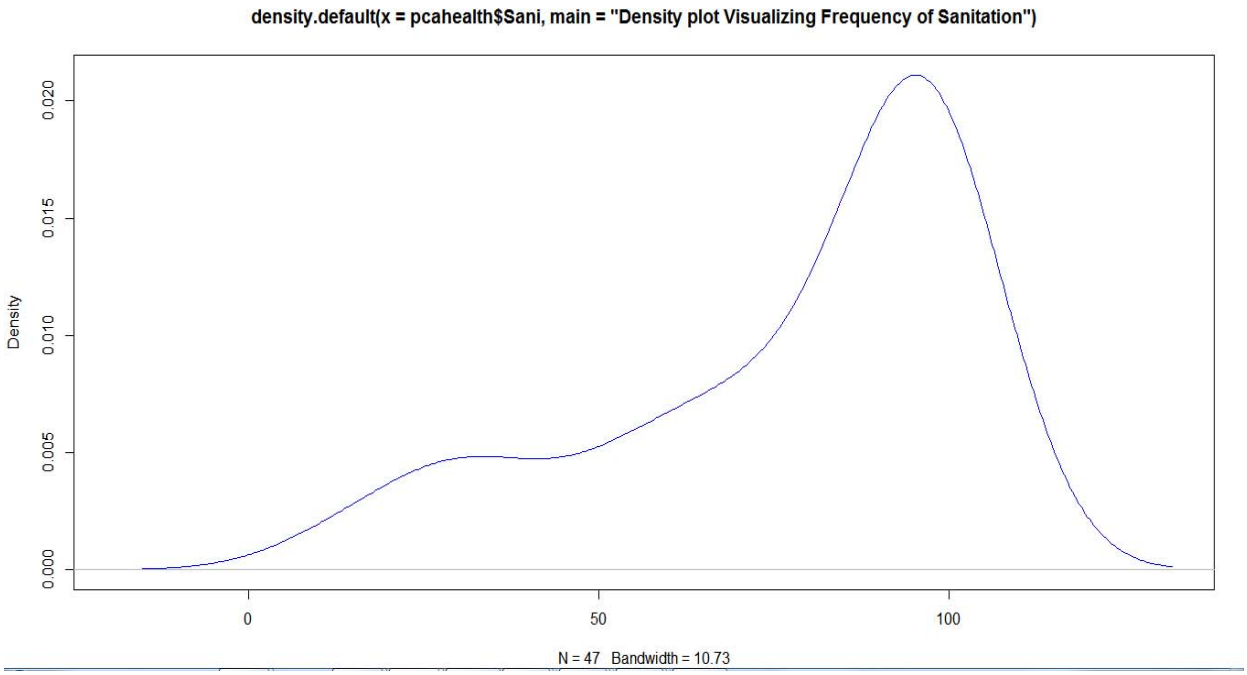


Figure 27-Density plot for Sanitation

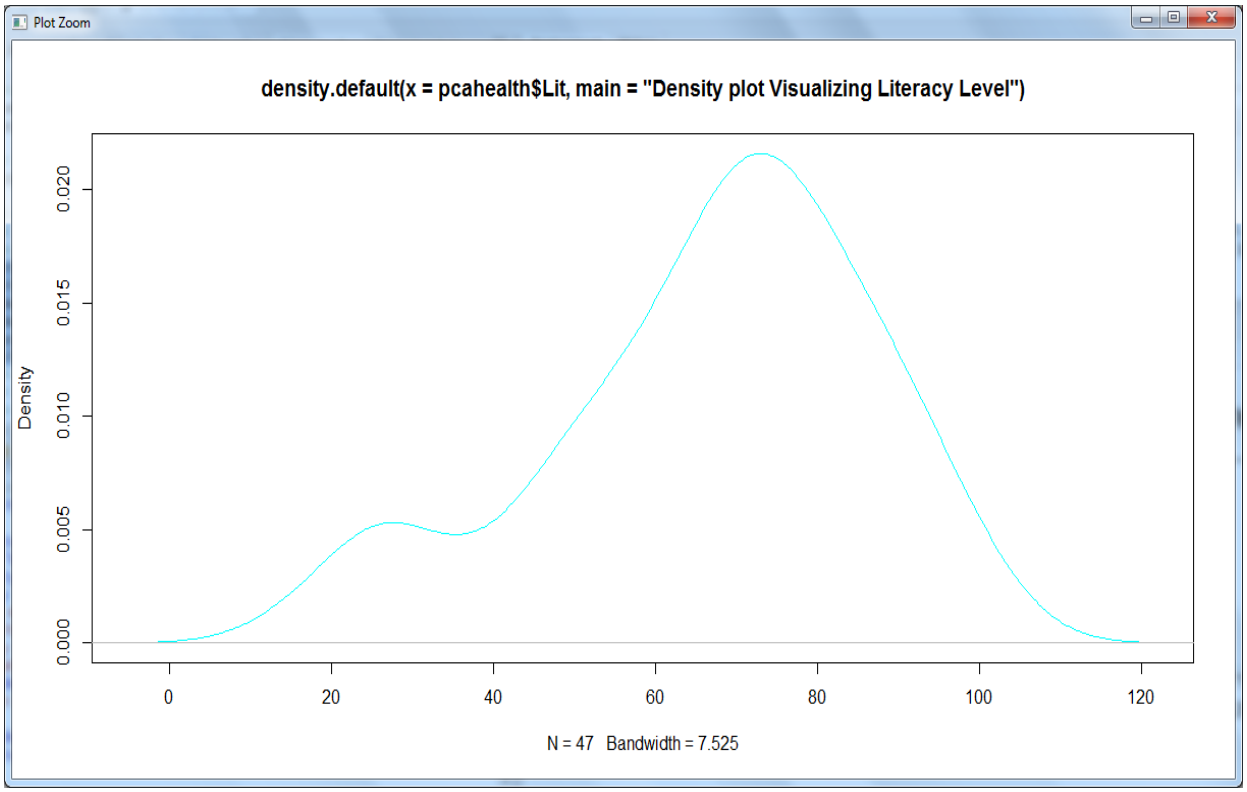


Figure 28-Density plot for Literacy

4.4 Modeling

4.4.1 Cluster Analysis

A cluster analysis is the process of summarizing a dataset by grouping similar observations together into clusters and observations are judged to be similar if they have similar values for a number of variables (i.e. a short *Euclidean distance* between them).

4.4.2 K-means Cluster Analysis

K-means algorithm cluster analysis was used to identify the naturally occurring groups present in the dataset. Using this non-linear clustering technique, each county was classified into one of the three groups according to the similarity of the counties based on the indicators of child health. Similarity using Euclidean distance measures between counties was calculated from the variables that went into these groups.

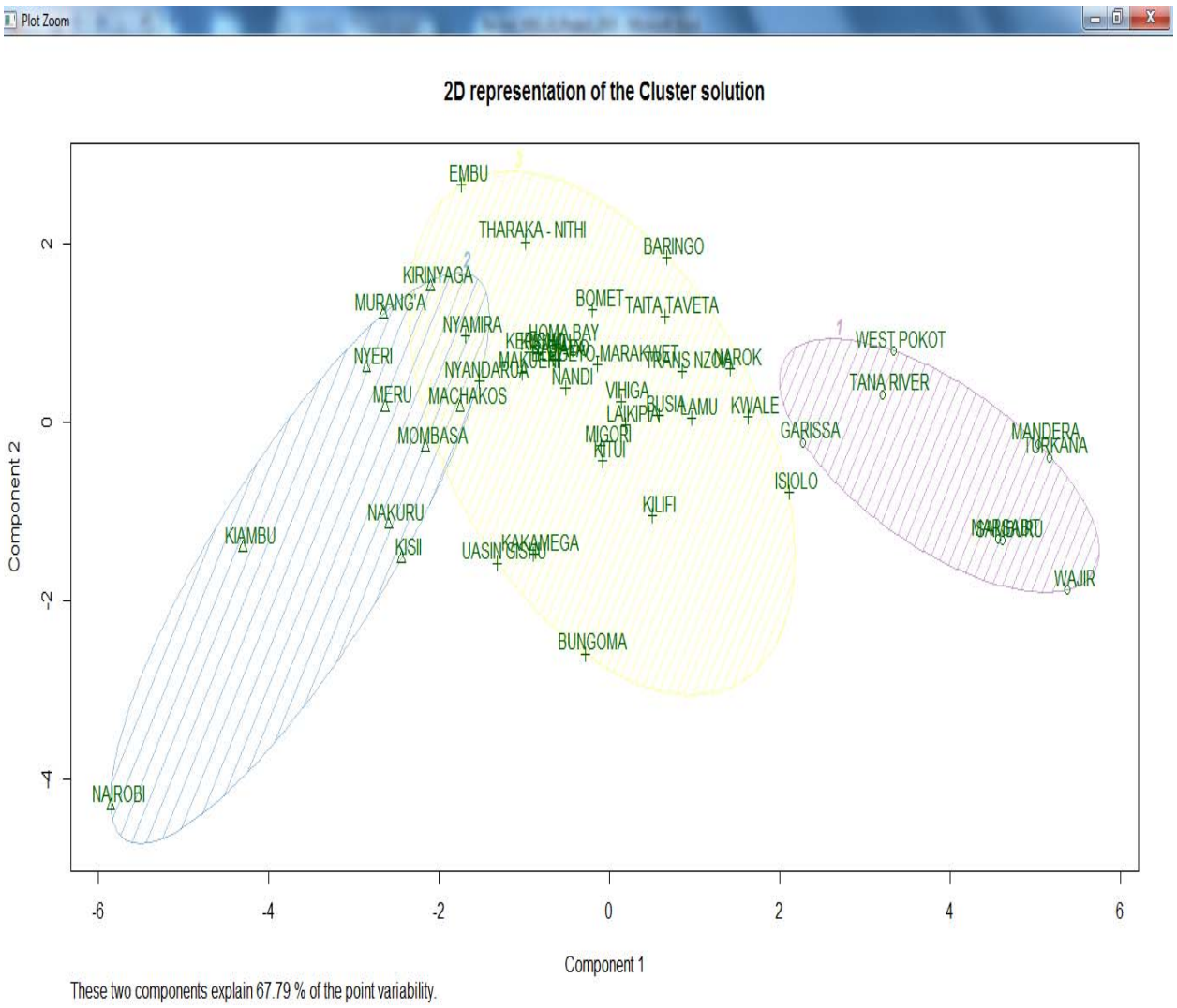


Figure 29-K-Means clustering results

KEY

Number	County
1	BARINGO
2	BOMET
3	BUNGOMA
4	BUSIA
5	ELEGEYO-MARAKWET
6	EMBU
7	GARISSA
8	HOMA BAY
9	ISIOLO
10	KAJIADO
11	KAKAMEGA
12	KERICHO
13	KIAMBU
14	KILIFI
15	KIRINYAGA
16	KISII
17	KISUMU
18	KITUI
19	KWALE
20	LAIKIPIA
21	LAMU
22	MACHAKOS
23	MAKUENI
24	MANDERA
25	MARSABIT
26	MERU
27	MIGORI
28	MOMBASA
29	MURANG'A
30	NAIROBI
31	NAKURU
32	NANDI
33	NAROK
34	NYAMIRA
35	NYANDARUA
36	NYERI
37	SAMBURU
38	SIAYA
39	TAITA TAVETA
40	TANA RIVER
41	THARAKA - NITHI
42	TRANS NZOIA
43	TURKANA
44	UASIN GISHU
45	VIHIGA
46	WAJIR
47	WEST POKOT

Figure 30-Counties' Key

Results

This was a creation of a bivariate plot visualizing a partition (clustering) of our dataset. All observations were represented by points in the plot, using principal components. An ellipse was drawn around each cluster.

4.4.3 Number of Clusters Determination

To determine the number of clusters to use, we used the within group sum of squares that guided us to group our dataset into three clusters as shown in the screeplot below.

Figure 31-Within the Sums of Squares plot

We used n-start parameter to avoid variable results for each run. By using n-start and iter-max parameters, we were able to get consistent results allowing us to have a proper interpretation of the Screeplot. The elbow was at k=4 but applied k-means clustering function with k=3 through heuristics and plotted the results.

We then looked at our clusters in order of increasing size. The first cluster contained 8 counties, second cluster contained 10 while the third cluster contained 29 counties. Cluster one was made up of the most marginalized counties, cluster two was made up of well-off counties while cluster three was made up of the moderately marginalized counties. Nairobi County is at its own rightly and is not an outlier. It is the county with the highest literacy level, health and educational facilities, and low poverty level.

4.4.4 Use of Box Plots

We used the box plots to compare literacy, healthcare delivery and health facilities in the clusters. In literacy, cluster two was the highest with an outlier, followed by the cluster three and cluster one had the lowest literacy level. Those seeking healthcare delivery was highest in cluster two followed by cluster three and lowest in cluster one. The sanitation was highest in cluster two followed by cluster three with the lowest being cluster one.

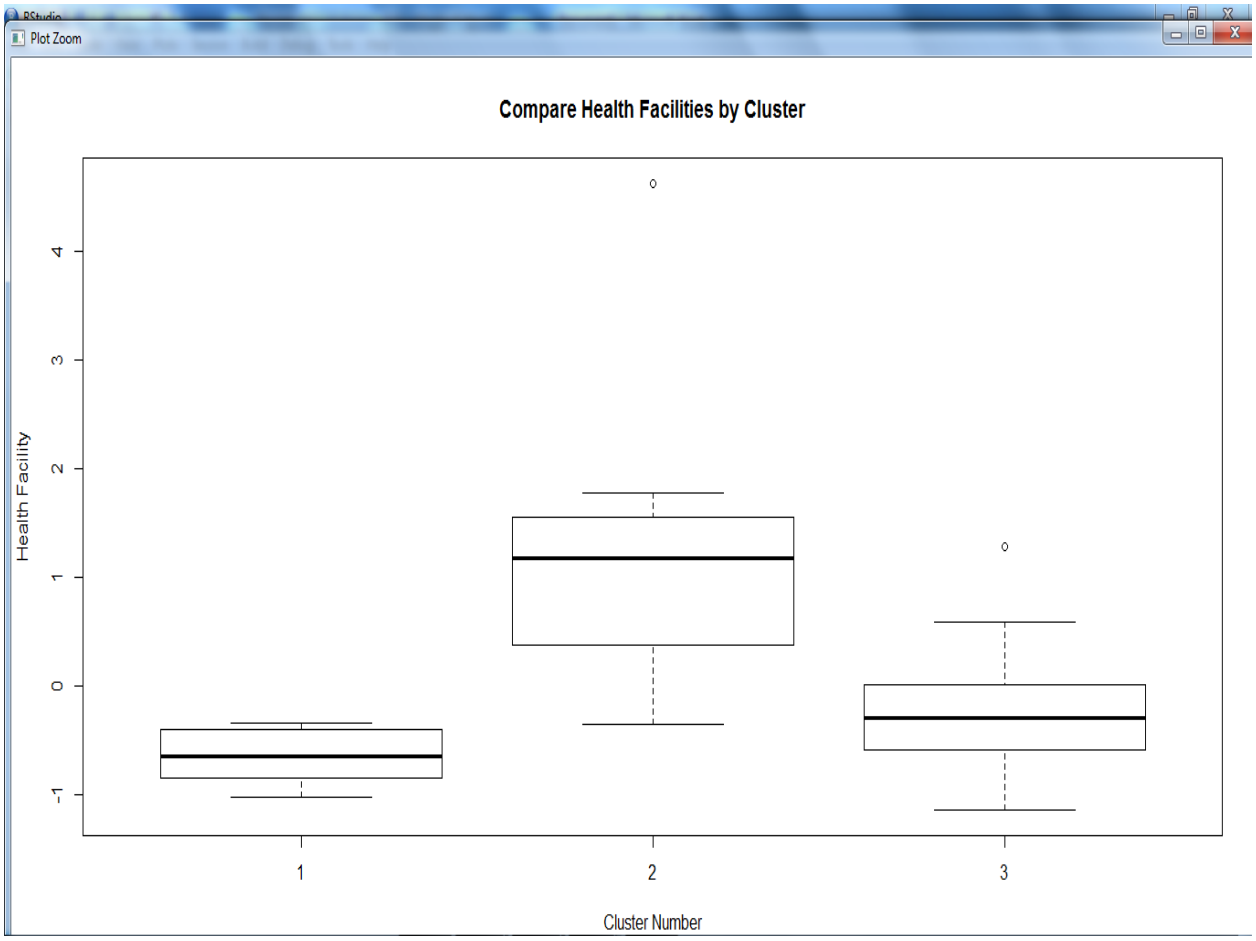


Figure 32-Comparing health facilities by cluster

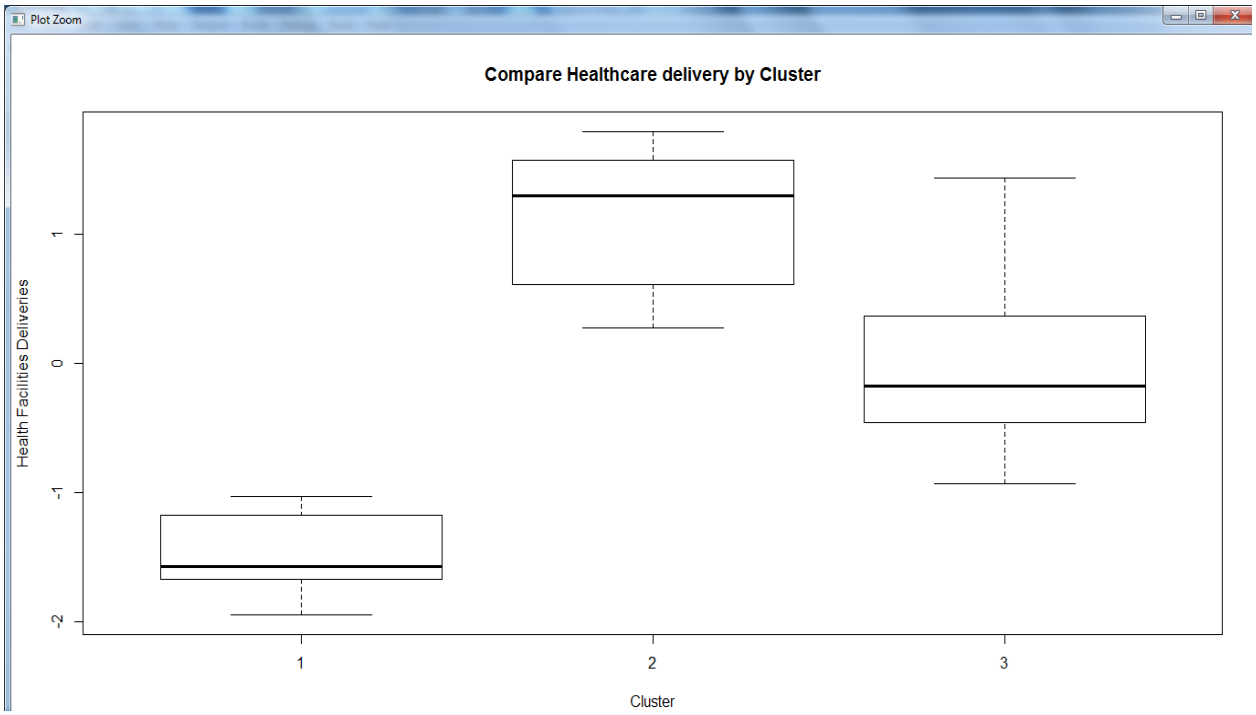


Figure 33-Comparing healthcare delivery by cluster

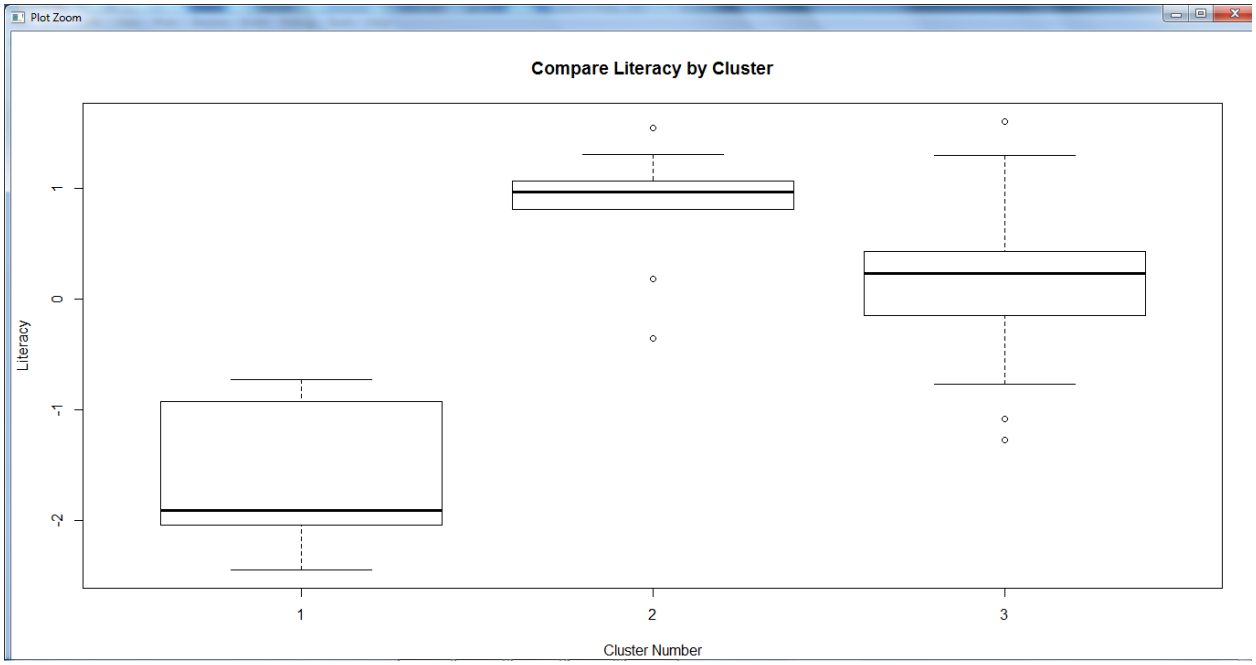


Figure 34-Compare Literacy by Cluster

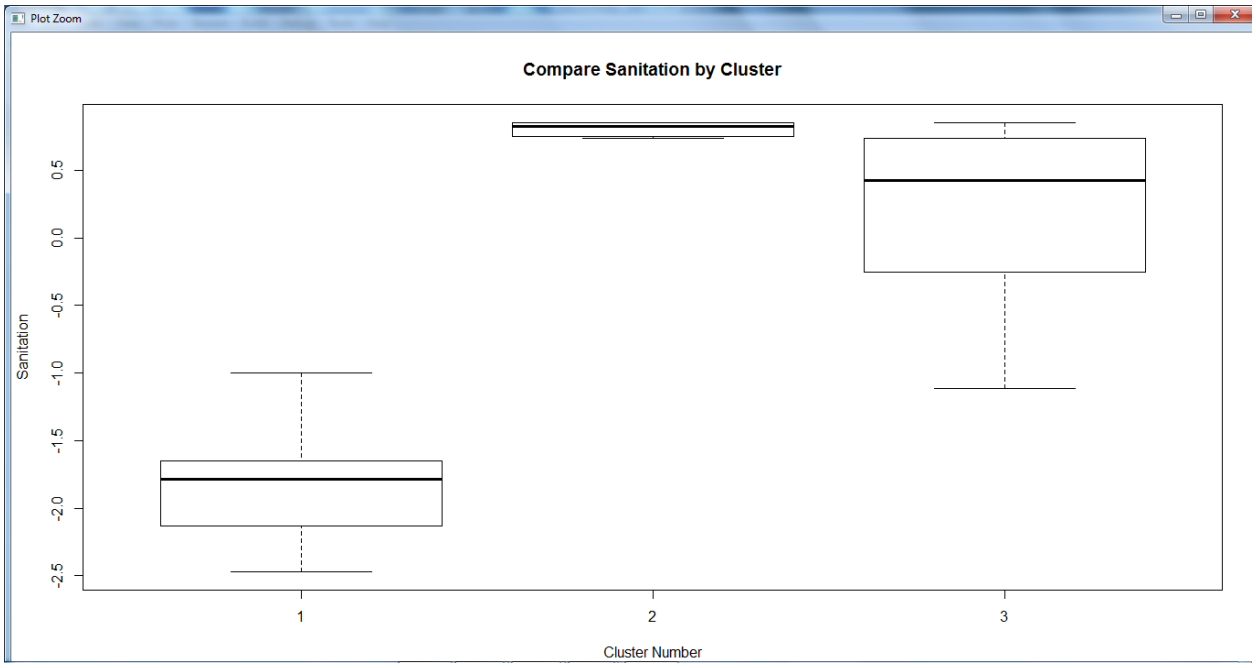
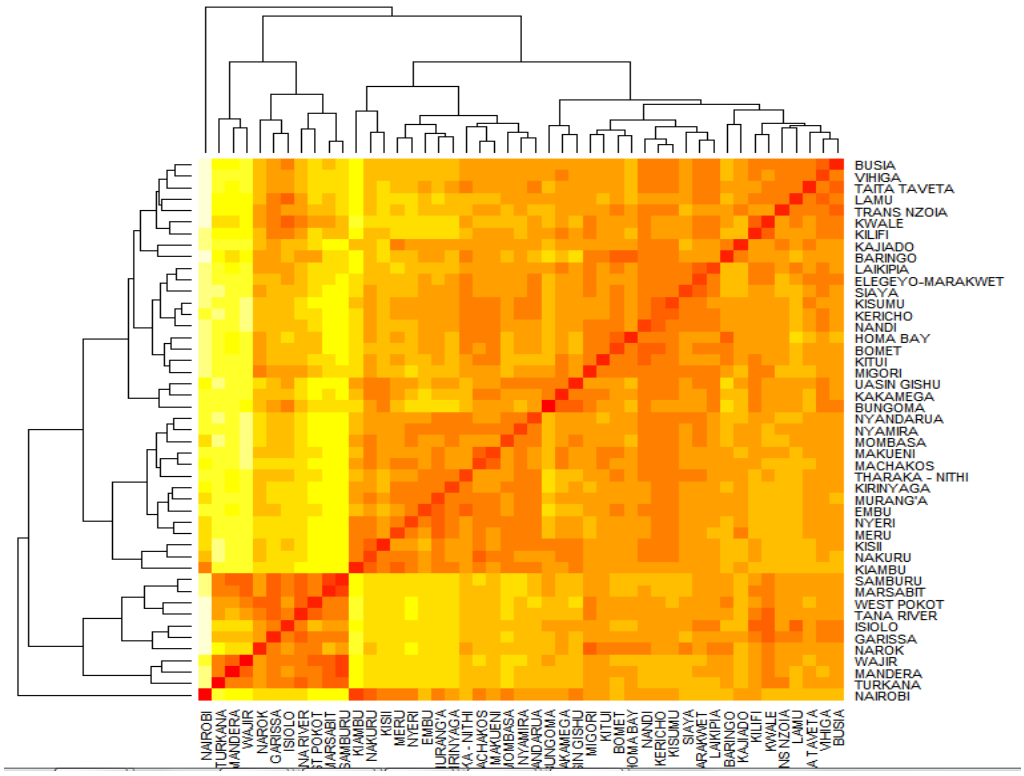


Figure 35-Compare Sanitation by Cluster

4.5 Dissimilarity Visualization

4.5.1 Heatmap

We used the heatmap to show the dissimilarities between the counties with color variations.



4.5.2 Dissimilarity Matrix

		BARINGO	BOMET	BUNGOMA	EMBU	GARISSA	KERICHO	KIAMBU	KILIFI	KIRINYAGA	MACHAKOS	MURANG'A	NAIROBI	SAMBURU	TANA RIVER	UASIN GIS	WEST POKO
2	BARINGO	0	0.128	0.297	0.21	0.217	0.186	0.429	0.166	0.261	0.233	0.272	0.543	0.363	0.245	0.275	0.231
3	BOMET	0.128	0	0.222	0.166	0.242	0.096	0.333	0.195	0.19	0.158	0.194	0.447	0.386	0.266	0.171	0.265
4	BUNGOMA	0.297	0.222	0	0.346	0.239	0.226	0.327	0.173	0.334	0.251	0.316	0.366	0.365	0.309	0.134	0.376
5	BUSIA	0.203	0.168	0.181	0.255	0.175	0.14	0.327	0.119	0.244	0.201	0.283	0.426	0.313	0.229	0.158	0.291
7	EMBU	0.21	0.166	0.346	0	0.373	0.143	0.255	0.3	0.134	0.164	0.109	0.359	0.523	0.402	0.234	0.412
8	GARISSA	0.217	0.242	0.239	0.373	0	0.263	0.458	0.164	0.358	0.319	0.4	0.572	0.173	0.12	0.297	0.146
13	KERICHO	0.186	0.096	0.226	0.143	0.263	0	0.255	0.178	0.143	0.095	0.147	0.368	0.403	0.284	0.124	0.342
14	KIAMBU	0.429	0.333	0.327	0.255	0.458	0.255	0	0.345	0.184	0.201	0.157	0.136	0.607	0.529	0.225	0.596
15	KILIFI	0.166	0.195	0.173	0.3	0.164	0.178	0.345	0	0.278	0.229	0.313	0.424	0.276	0.221	0.172	0.288
16	KIRINYAGA	0.261	0.19	0.334	0.134	0.358	0.143	0.184	0.278	0	0.181	0.108	0.315	0.507	0.391	0.231	0.434
28	MIGORI	0.131	0.119	0.191	0.237	0.225	0.132	0.352	0.095	0.25	0.193	0.245	0.43	0.328	0.235	0.167	0.279
29	MOMBASA	0.308	0.23	0.262	0.159	0.328	0.161	0.158	0.197	0.145	0.171	0.143	0.252	0.453	0.396	0.143	0.463
30	MURANG'A	0.272	0.194	0.316	0.109	0.4	0.147	0.157	0.313	0.108	0.114	0	0.273	0.549	0.429	0.221	0.459
31	NAIROBI	0.543	0.447	0.366	0.359	0.572	0.368	0.136	0.424	0.315	0.314	0.273	0	0.698	0.643	0.293	0.71
38	SAMBURU	0.363	0.386	0.365	0.523	0.173	0.403	0.607	0.276	0.507	0.468	0.549	0.698	0	0.131	0.423	0.164
39	SIAYA	0.207	0.143	0.228	0.17	0.289	0.103	0.309	0.171	0.192	0.193	0.203	0.378	0.387	0.291	0.176	0.339
40	TAITA TAVETA	0.235	0.179	0.256	0.209	0.206	0.142	0.323	0.194	0.198	0.174	0.237	0.438	0.33	0.25	0.198	0.302
44	TURKANA	0.351	0.435	0.443	0.513	0.22	0.463	0.663	0.355	0.545	0.515	0.575	0.776	0.175	0.22	0.501	0.214
45	UASIN GISHU	0.275	0.171	0.134	0.234	0.297	0.124	0.225	0.172	0.231	0.174	0.221	0.293	0.423	0.367	0	0.434
48	WEST POKOT	0.231	0.265	0.376	0.412	0.146	0.342	0.596	0.288	0.434	0.4	0.459	0.71	0.164	0.114	0.434	0
49																	
50																	

DissMatrix

Figure 36-Dissimilarity matrix

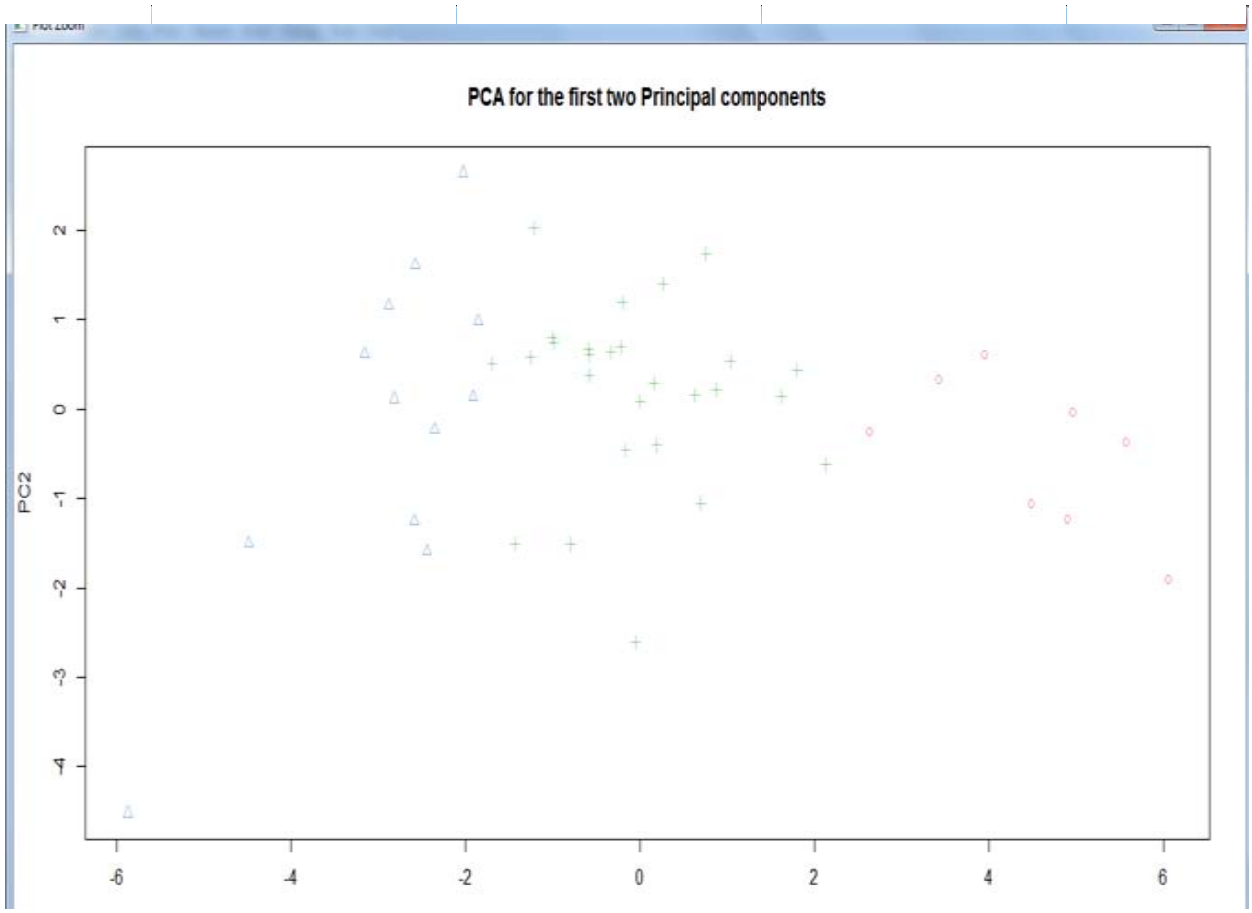


Figure 37-Scatterplot for the first two Principal Components.

4.6 Hierarchical Clustering and Bannerplot

Hierarchical Clustering draws a “banner”, i.e. basically a horizontal bar plot visualizing the (agglomerative or divisive) hierarchical clustering or any other binary dendrogram structure.

Agglomerative Coefficient (AC)

This refers to the measure of how much clustering structure exists in the data. A large AC (close to one) means that there is a strong clustering structure. A small AC means that the data is more evenly distributed hence a poor clustering structure.

4.6.1 Agglomerative Analysis (AGNES) and agglomerative coefficient

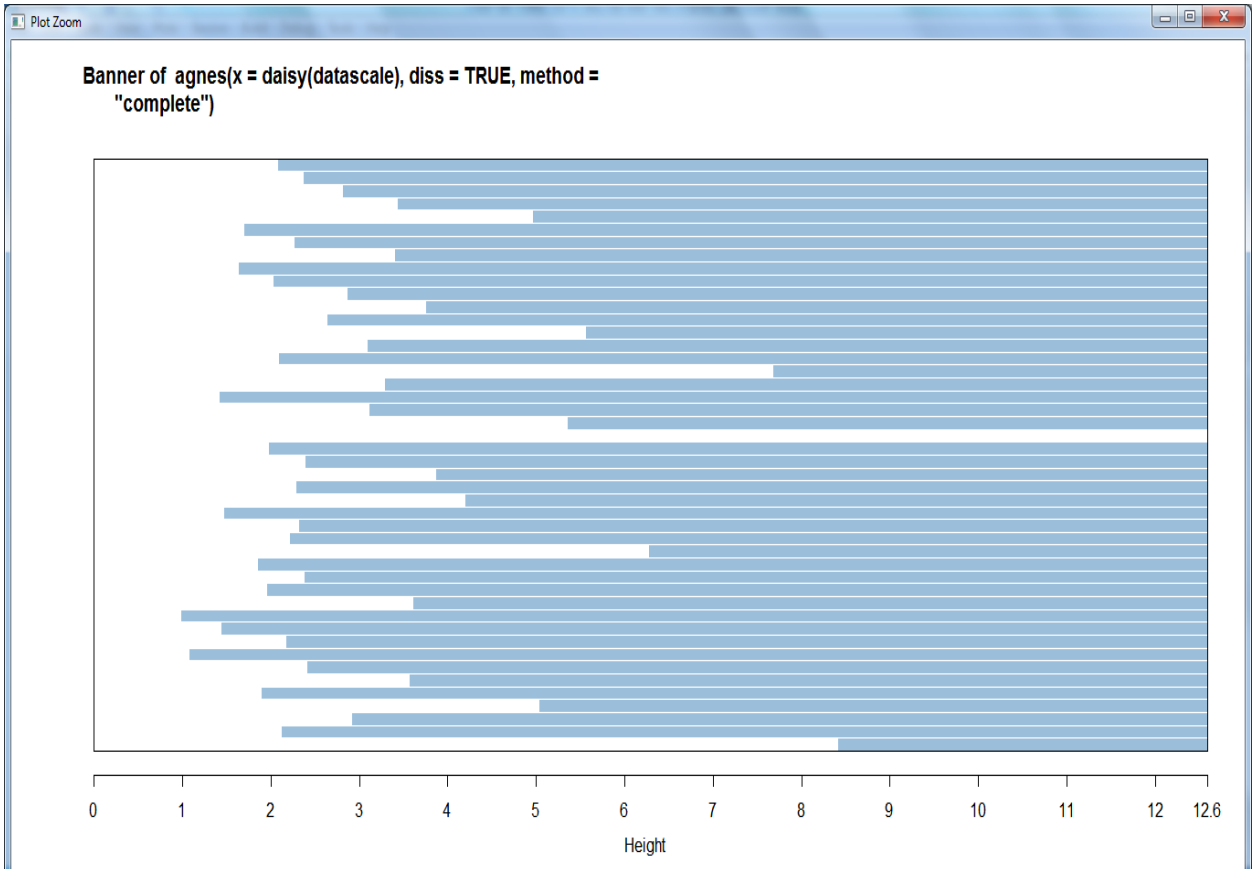


Figure 38-Banner plot of AGNES algorithm

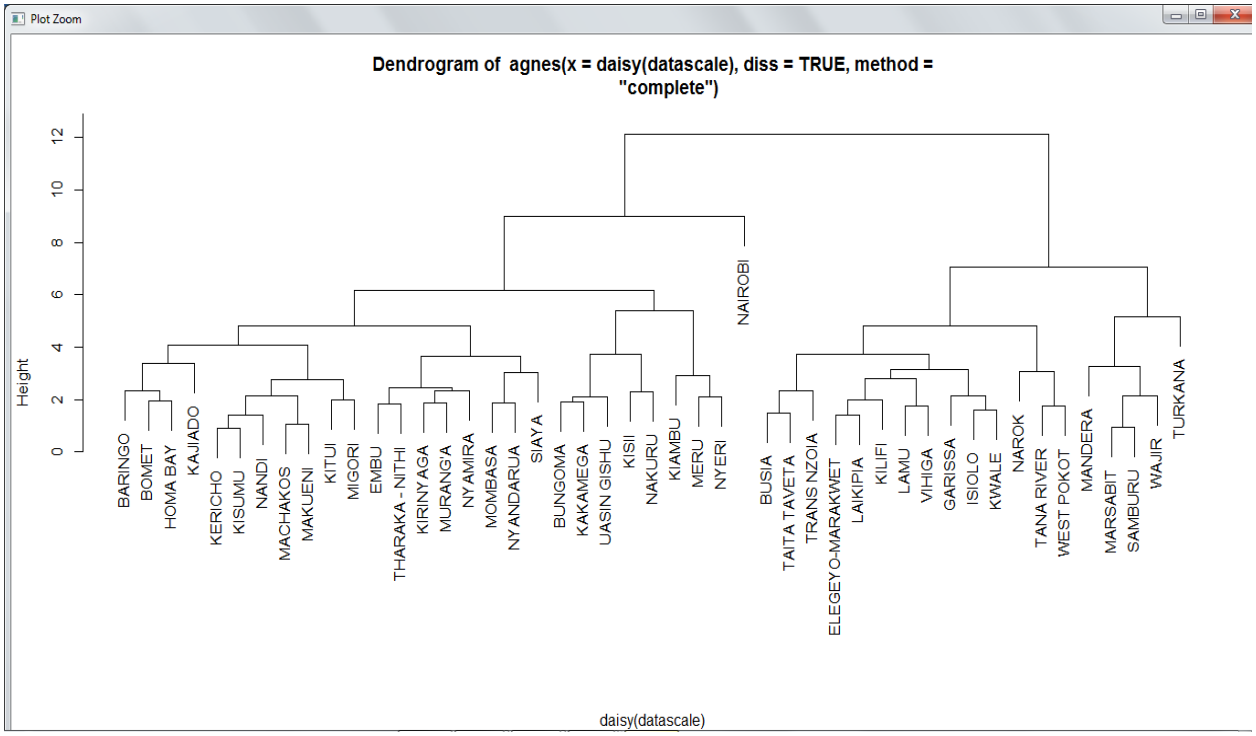


Figure 39-Dendrogram of AGNES algorithm

4.6.2 Divisive Analysis (DIANA) and divisive coefficient

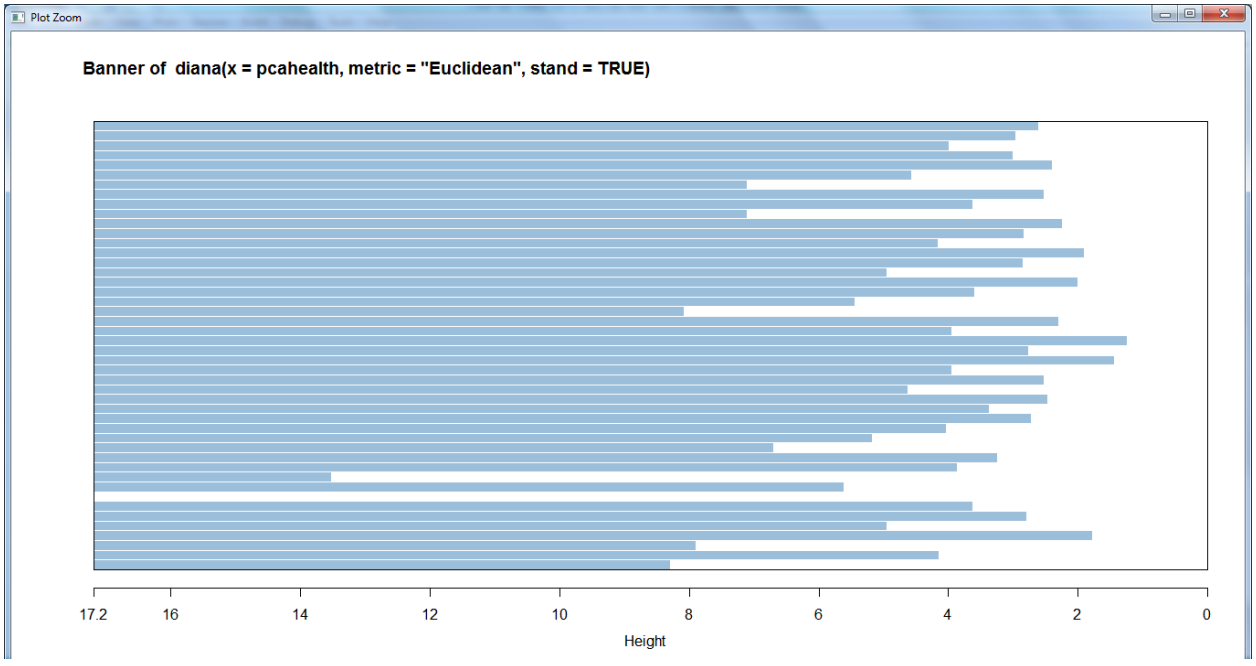


Figure 40-Banner plot for DIANA algorithm

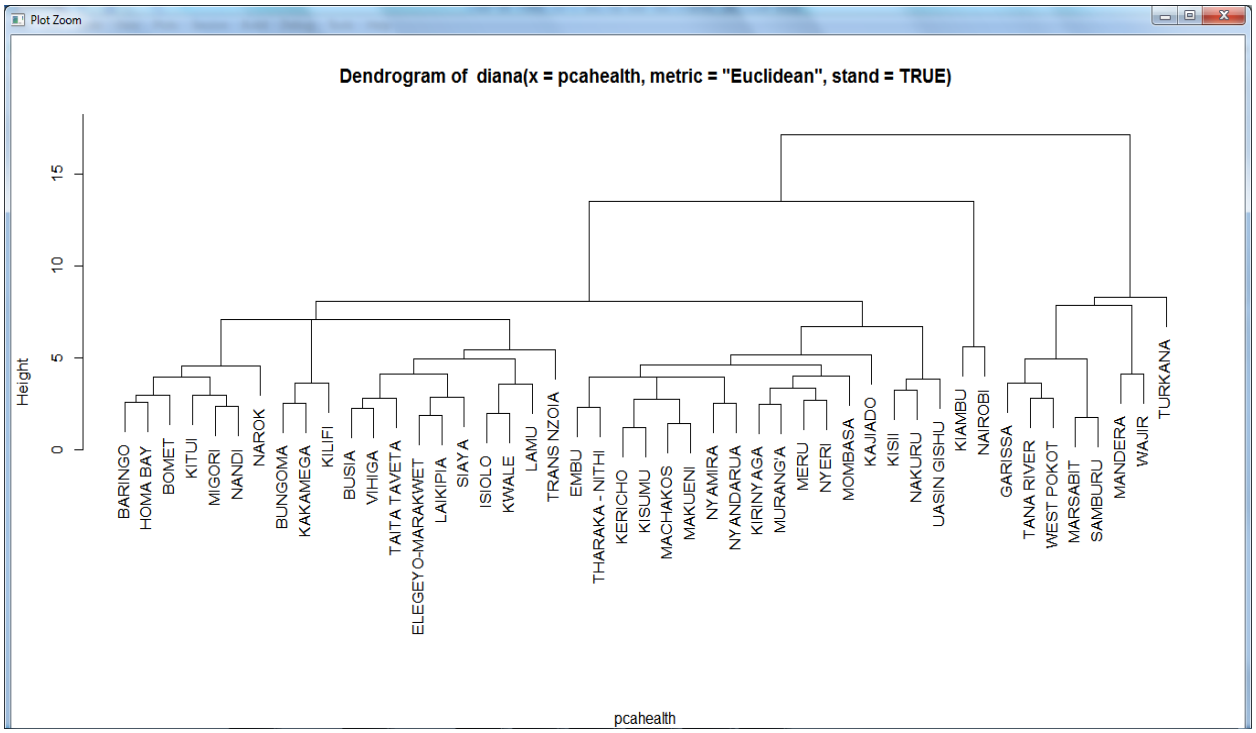


Figure 41-Dendrogram for DIANA algorithm

Silhouette Coefficient

Peter J. Rousseeuw (1986) described Silhouette as a method of interpretation and validation of consistency within clusters of data. This technique provides a succinct graphical representation of how well each object lies within its cluster.

Interpretation of Silhouette Coefficient

Silhouette Coefficient	**Explanations**
0.71-1.00	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial. Try additional methods of data analysis.
<=0.25	No substantial structure has been found

4.7 Other non-hierarchical Clustering Algorithms

4.7.1 Fuzzy Analysis (Fanny) and Silhouette Coefficient

Fuzzy clustering is a generalization of partitioning. In a partition, each object of the data set is assigned to one and only one cluster. It also allows for some ambiguity in the data, which often occurs in practice.

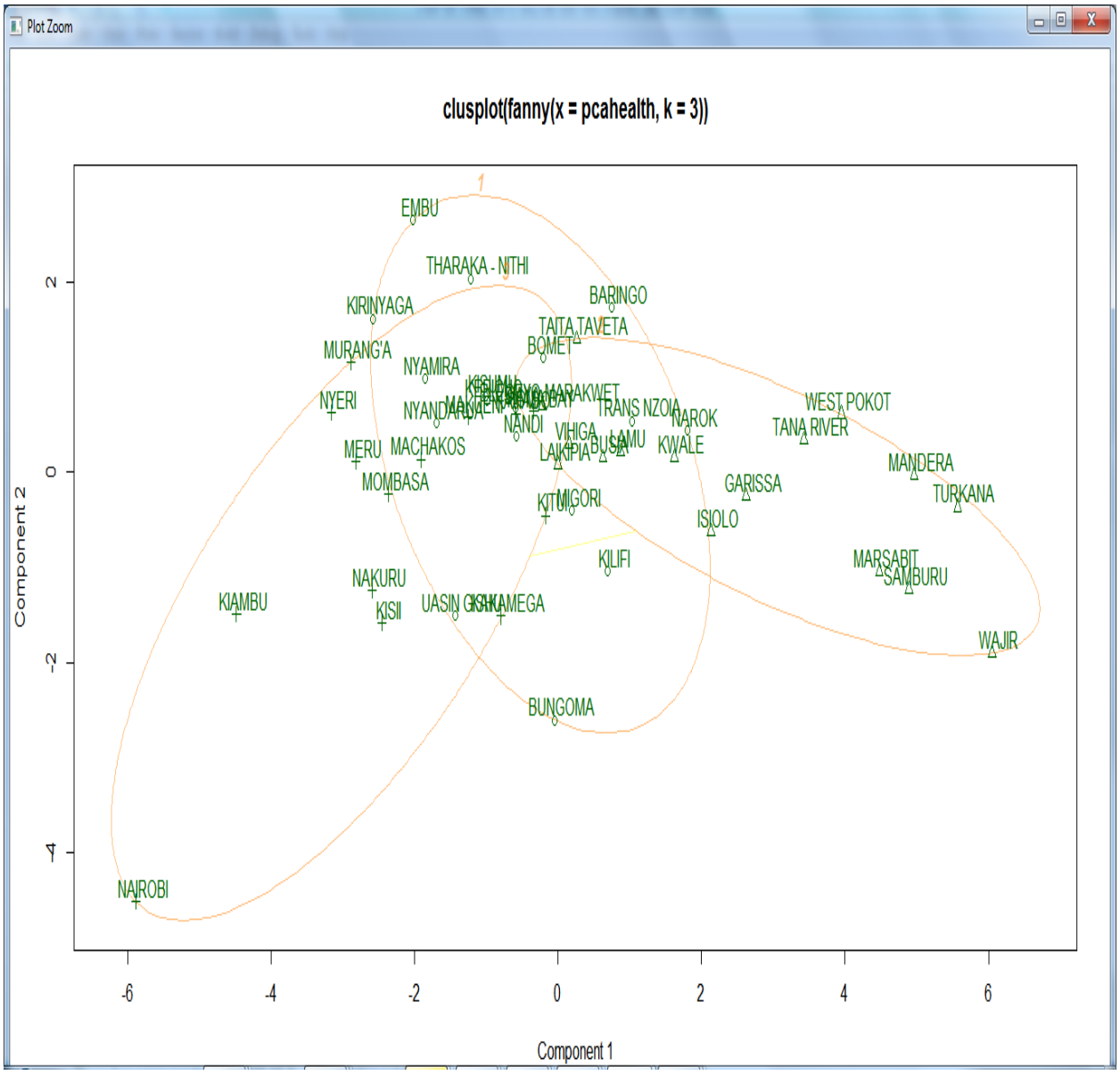


Figure 42-Clustering results using FANNY algorithm

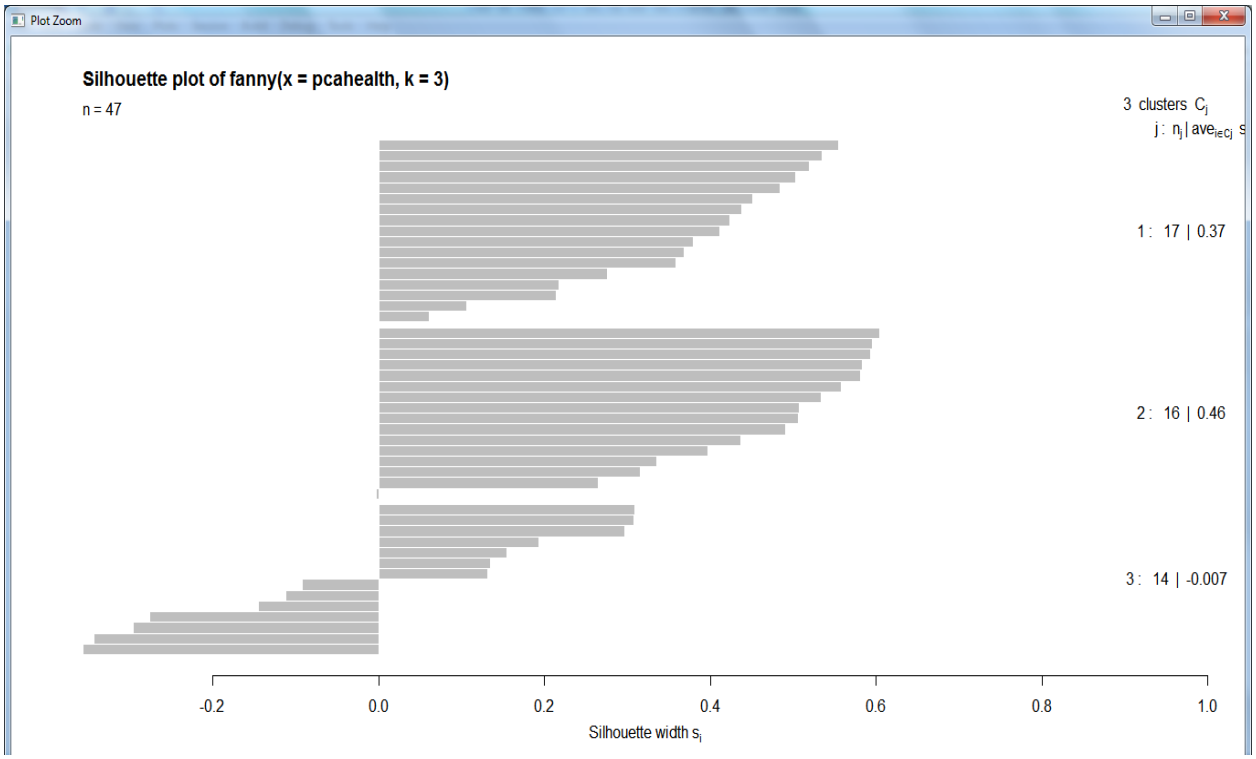


Figure 43-Silhouette from FANNY algorithm

Results

The fuzzy clustering algorithm classified our observation but into three clusters of with an average silhouette Coefficient of 0.29 which means that the structure was weak and artificial so another method was recommended. More analysis of the clusters is shown below.

```
Average silhouette width per cluster:
[1]  0.370624578  0.456362299 -0.006595817
Average silhouette width of total data set:
[1] 0.2874484

1081 dissimilarities, summarized :
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  25.266  136.070  206.990  256.570  316.640 1158.900
Metric :  euclidean
Number of objects : 47
```

Figure 44-More Fuzzy Analysis

4.7.2 Partitioning Around Medoids (PAM) and Silhouette Coefficient

We also tested our dataset using the Partitioning which is a more used for Partitioning (clustering) of the data into k clusters "around medoids", which is a more robust version of K-means. Compared to the k-means approach in k-means, the function PAM has the following

features: (a) it accepts a dissimilarity matrix; (b) it is more robust because it minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances;(c) it provides a novel graphical display, the silhouette plot.

Results

This algorithm generated a three cluster solution with the size of 24, 16 and 7. We however discarded its output because its silhouette coefficient was very low at 0.35 meaning that the structure was weak and could be artificial. More detailed results are shown below for silhouette width per cluster.

```
Average silhouette width per cluster:
[1] 0.2720443 0.5547723 0.1442289
Average silhouette width of total data set:
[1] 0.3492558

1081 dissimilarities, summarized :
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  25.266  136.070  206.990  256.570  316.640 1158.900
Metric :  euclidean
Number of objects : 47
```

Figure 45-Silhuoette width per cluster

4.7.3 Clustering Large Application (CLARA) and Silhouette Coefficient

This algorithm computes a "clara" object, that is, a list representing a clustering of the data into k clusters. This method can deal with large datasets as compared to PAM and FANNY.

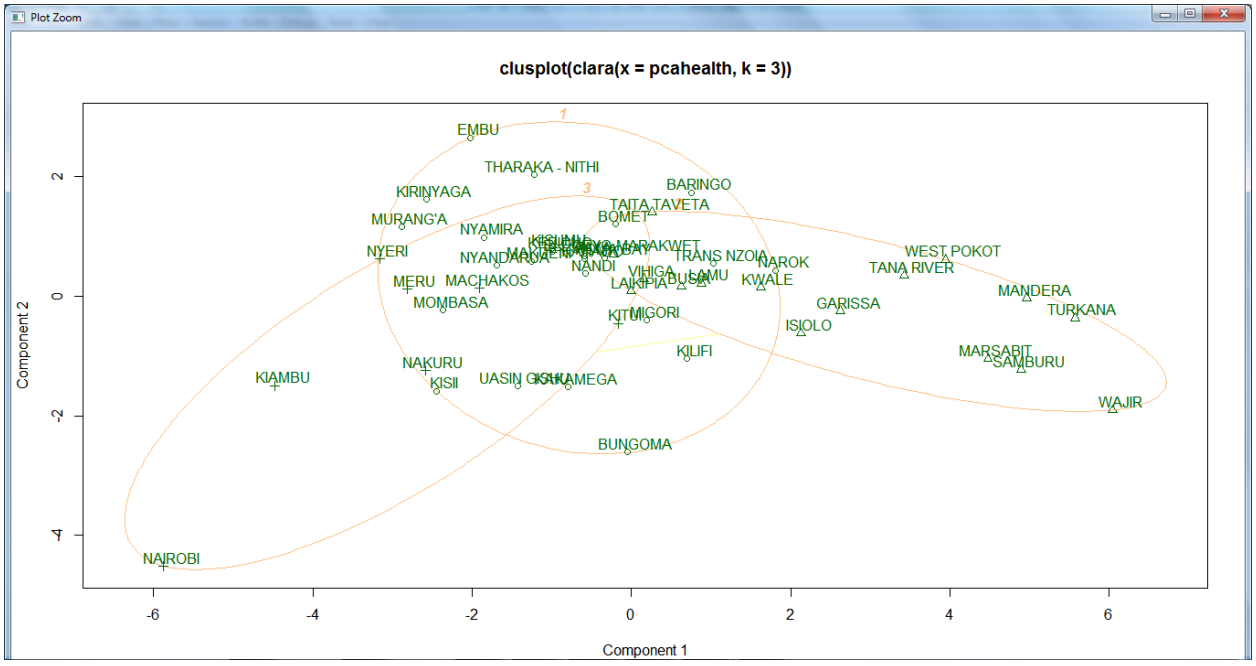


Figure 46-Results from CLARA algorithm

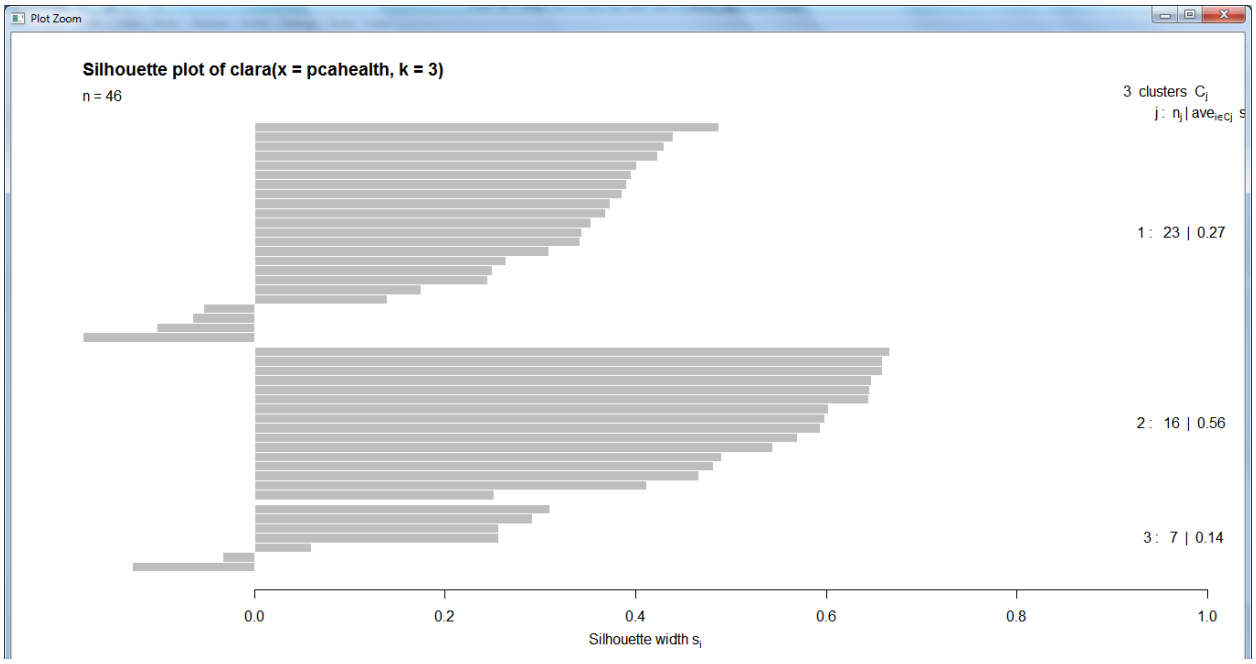


Figure 47-Silhouette results of CLARA algorithm

Results

The algorithm created three clusters of size 24, 16 and 7 with the two components explaining the variability of 68.68%. However we discarded the algorithm because the silhouette coefficient was very weak at 0.35 meaning the structure was weak. More detailed information on the clustering are as show below.

```
Numerical information per cluster:
     size max_diss   av_diss isolation
[1,]   24 283.5836 108.15323 1.5435630
[2,]   16 106.1514  73.30103 0.5777888
[3,]    7 659.6968 184.25007 2.2063206
Average silhouette width per cluster:
[1] 0.2658011 0.5580336 0.1448048
Average silhouette width of best sample: 0.3490347
```

Figure 48-CLARA algorithm Numerical Information

4.7.4 Results from various algorithms

Cluster/Algorithm	K-Means	PAM	CLARA	FANNY	AGNES	DIANA
Moderately Marginalized	29	24	24	17	27	37
Most Marginalized	08	16	16	16	19	08
Well-off	10	07	07	14	01	02

Figure 49-Results from various algorithms

CHAPTER FIVE-CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary of the Main Findings

This research concentrated on building a model for clustering and visualizing the status of child health in Kenya. A construct with five dimensions: Child health, Education, Maternal Health, Water and sanitation and others was used to develop the classification of three clusters of most marginalized, moderately marginalized and well-off counties. K-means clustering algorithm was used for modeling. We used other clustering algorithms such as Partitioning Around Medoids (PAM), CLARA, FANNY, AGNES and DIANA to compare the results from k-means which gave comparable results and also test the solutions' stability. We also used an expert child health to judge the validity our results who confirmed our findings were the reflection of reality. The k-means clustering algorithm generated the results shown in the table below.

Observations	%	Counties' Name	Cluster Class
10	21%	Kiambu, Kirinyaga, Kisii, Machakos, Meru, Mombasa, Murang'a, Nairobi, Nakuru, Meru.	Well-off
8	17%	Garissa, Mandera, Marsabit, Samburu, Tana-River, Turkana, Wajir, West-Pokot.	Most Marginalized
29	62%	Baringo, Bomet, Bungoma, Busia, Elgeyo-Marakwet, Homa-Bay, Isiolo, Kajiado, Kakamega, Kericho, Kilifi, Kisumu, Kitui, Kwale, Laikipia, Lamu, Makueni, Migori, Nandi, Nakuru, Nyandarua, Siaya, Taita Taveta, Tharaka Nithi, Trans Nzoia, Uasin Gishu, Vihiga, Embu, Nyamira.	Moderately Marginalized

Figure 50-Table showing K-Means results

This shows that 17% of the counties have the most disadvantaged children, 21% are well-off and 62% are moderately disadvantaged.

We used box plots to compare the three clusters in features; literacy and health care delivery, sanitation and health facilities. Cluster "well-off counties" was doing well in literacy, followed by cluster "moderately marginalized" with cluster "most marginalized" being highly disadvantaged. The literacy level in cluster "well-off counties" was above 80% but below 95%,

cluster “moderately marginalized” was between 60% and 70% whereas cluster “most marginalized” was below 45%.

There was much similarity in how observations were grouped, but also there were some differences. This was a reminder that different clustering methods often produce different groupings. In the application of different groupings, we were interested to observe how clustering patterns from different algorithms would vary.

By applying different cluster algorithms and data reduction methods, we were able to generate a consensus result describing the way the objects were grouped through the partitioning and hierarchical clustering algorithms. Partitioning method fanny allowed us to robustly assess objects to cluster and assess any ambiguities by looking at the fuzziness of objects. Plots that were generated by the algorithms enabled us to visualize the consensus grouping of objects.

5.2 Contribution of the Study

The study will contribute to the society by identifying the status of child health in Kenya. The study showed that the counties where the children are highly deprived of their rights of well being are Garissa, Mandera, Marsabit, Samburu, Tana River, Turkana, Wajir and West Pokot. The research was able to benchmark counties making the devolved government have a picture of the status of child health in their counties and help them in strategizing on the improvement of the indicators of the child health.

In academic, this study was a success as it utilized data mining tools and techniques that proved to have high contribution in deriving patterns that are useful in decision making. The significance of clustering status of child health patterns sheds light on potential application in healthcare and other research areas.

5.3 Recommendations

The devolved governments and the national government can create an opportunity by improving the child health by engaging them in the provision of the key services that promote child health such as the provision of improved sanitation, improved healthcare services, improving the household incomes, improve the delivery facilities, promote and improve education and infrastructure. There can also be a heighted advocacy by both the national and the county

government and other stakeholders in child wellbeing to oversee the implementation of these services in the counties.

5.4 Limitation of the study

The data collected from the literature was from various sources of past research and at different times and therefore quality and accuracy of research could have been compromised. The time lag was assumed since the data could have not been a reflection of the present since the data was historical. Some data on the indicators of child health could not be found from literature. We assumed the data we got was representative for our research. We used small samples to demonstrate clustering in our research even though in practice clustering is done on large data set

5.5 Recommendation for Future Work

In future we recommend a web and mobile based system using knitr and shinyapps packages provided by R studio to cluster and visualize the status in real-time. Further study with all UNICEF variables is required to prove this study.

5.6 Conclusion

Cluster analysis techniques can be constructive for exploring and describing data sets in child health. Through clustering, hidden relationships among variables that are not obvious to researchers were identified hence enhancing knowledge of data set which would serve as a preliminary point for future research. The technique used offers excellent results and can lead to an improvement in child health care. This research in cluster analysis has demonstrated how researchers can combine more than one clustering methods to explore data to reveal the underlying structure of objects.

References

1. Gupta, G. K. (2014). Introduction to Data Mining with Case studies, third edition. PHI Learning Private Limited, Delhi.
2. Amorim, R.C., & Hennig, C. (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors". Information Sciences 324: 126–145. doi:10.1016/j.ins.2015.06.039.
3. Koh, H. C., & Tan, G. (2005), "Data mining applications in healthcare," Journal of Healthcare Information Management, vol. 19, no. 2, pp. 64–72..
4. Nittel, S., Leung, K. T., & Braverman, A. (2003), "Scaling clustering algorithms for massive data sets using data stream," in Proceedings of the 19th International Conference on Data Engineering.
5. Rousseeuw, P. J. (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65. doi:10.1016/0377-0427(87)90125-7.
6. Shmueli, G., Patel, R., & Bruce, P. C. (2010). Data Mining for Business Intelligence. 2nd edition. New Jersey: Wiley.
7. Wasiewicz, P., Kulaga, Z., & M. Litwi (2009) .Data mining analysis of factors influencing children's blood pressure in a nation-wide health survey Author(s). Proc. SPIE 7502, Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2009, 75022R (6 August 2009); doi: 10.1117/12.838236
8. Rehnman, A. (2014). Socio –Economic and demographic factors affecting child health in Rural Areas of Tehsil Jehanian District Khanewal. Standard Scientific Research and Essays. Vol2 (12):652-656, December 2014 (ISBN: 2310-7502).
9. Nzioki, J.M., Onyango, R. O., & Ombaka, J. H. (2015). "Socio-Demographic Factors Influencing Maternal and Child Health Service Utilization in Mwingi; a Rural Semi-Arid District in Kenya." American Journal of Public Health Research 3.1 (2015): 21-30.
10. Shinsugi, C., Matsumura, M., Karama, M., Tanaka, J., Changoma, M., & Kaneko, S. (2015). Factors associated with stunting among children according to the level of food insecurity in the household: a cross-sectional study in a rural community of Southeastern Kenya. Shinsugi et al. BMC Public Health (2015) 15:441 DOI 10.1186/s12889-015-1802-6

11. Anand, S. S., & John, G. Data Mining: Looking Beyond the Tip of the Iceberg. Hughes Faculty of Informatics University of Ulster (Jordan town Campus) Northern Ireland.
12. Yim, H., Boo, Y., & Ebbeck, M. (2014). A Study of Children's Musical Preference: A Data Mining Approach. Australian Journal of Teacher Education, 39(2).
13. Jing, H. (2009).Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on (Volume: 1) Date of Conference: 21-22 Nov. 2009 Page(s): 634 - 636 Print ISBN: 978-0-7695-3859-4. DOI: 10.1109/IITA.2009.204 Publisher: IEEE
14. Coenen, F. (2011). Data mining: past, present and future, The Knowledge engineering Review, Vol. 26:1, 25–29.&Cambridge University Press. doi:10.1017/S0269888910000378
15. Gupta, G. K. (2006). Introduction to Data Mining with Case Studies. Easter Economy Edition, Prentice Hall of India.
16. Ott, R. L., & Longnecker, M. (2001). An introduction to Statistical Methods and Data Analysis. Sixth edition. Brooks/Cole Cenage Learning.
17. Joshi, M. C., & Moudgalya, K. N. (2004).Optimization Theory and Practice. Alpha Science International Ltd-Harrow UK.
18. Hsieh, W. W. (2009). Machine Learning in the Environmental Sciences: Neural Networks and Kernels. Cambridge University Press-NY,
19. Desikan, P., Hseu, K. W., Srivastava, J. (2011). Data Mining For Healthcare Management. 2011 SIAM International Conference on Data Mining, April 28-30, 2011.
20. Boaz, S. H. (2013). Principal Component Analysis on Higher Education Institution (HEI) Ranking Systems and Conceptual Framework for a New Age Hei Ranking System – Indian Context. Research Journal of Management Sciences. ISSN 2319–1171 Vol. 2(6), 27-32, Res. J. Management Sci.
21. Dean, J. (2014). Big Data, Data Mining, and Machine Learning; Value creation for business managers and practitioners. Wiley & SAS Business Series
22. Ngaruiya, M. N., & Moturi, C. (2014). Use of data mining to check the Prevalence of prostate cancer: case of Nairobi County.
23. Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. John Wiley & Sons, Inc. WIREs Comp Stat 2010 2 433–459.

Appendices

Sample Code

```
library(RColorBrewer)

library(cluster)

library(rgl)

library(scales)

#Setting the working directory and read CSV into R

setwd("C:/Users/nicken/Desktop/PROJREAL")

pcahealth=read.csv("clusteringHealth.csv",header=TRUE,row.names=1)

# Display Data structure

str(pcahealth)

#     Histograms and Density plots

histsani<-hist(pcahealth$Sani, main ="Histogram Visualizing Frequency of Sanitation")

densitywater<-density(pcahealth$Sani,main ="Density plot Visualizing Frequency of
Sanitation")

# verify by plotting variance of columns

mar <- par()$mar

par(mar=mar+c(0,5,0,0))

barplot(sapply(pcahealth, var), horiz=T, las=1, cex.names=0.8)

barplot(sapply(pcahealth, var), horiz=T, las=1, cex.names=0.8, log='x')

par(mar=mar)

pca.features2=pcahealth

#pca.features$County <- NULL

names(pca.features)

#View(pca.features)

#remove or estimate missing data
```

```
# Prepare Data
pca.features <- na.omit(pca.features2) # listwise deletion of missing
# Scale- rescale variables for comparability.
datascale<- data.frame(scale(pca.features))
write.csv(datascale, "scaleddata.csv")
# Verify variance is uniform
plot(sapply(datascale, var))
#PRINCIPAL COMPONENT ANALYSIS SECTION
#Performing Principal Component Analysis
pcaPC <- (princomp(datascale, cor=TRUE,scores=TRUE))
plot(pcaPC,main="Bars Screeplot")
plot(pcaPC,main="Linear Screeplot",type="l")#Plot the Screeplot diagram
print(summary(pcaPC))#view the proportion of the total
#variance explained by each component
summaryPC<-summary(pcaPC)
summaryPC
# Use of prcomp to get principal component vectors  instead of princomp
PCprcomp <- prcomp(datascale)
#      VISUALIZING IN TWO DIMENSION
# Shows the First four principal components
FirstPCs <- data.frame(PCprcomp$x[,1:4])
# Plotting the scatter plot
plot(FirstPCs, pch=16, col=rgb(0,0,0,0.5))
#      VISUALIZING IN THREE DIMENSION
library(rgl)
```

```
plot3d(FirstPCs$PC1, FirstPCs$PC2, FirstPCs$PC3)
plot3d(FirstPCs$PC1, FirstPCs$PC3, FirstPCs$PC4)
scores<- pcaPC$scores#Shows the scores of each observation
loadings<-pcaPC$loadings#show the loadings for each component
write.csv(loadings, "loadings.csv")
write.csv(scores, "scores.csv")
#+++++++COVARIANCE++++++++++++++++++
cell.cov<-round(cov(datascale),3)
cell.cov
write.csv(cell.cov,"Covariance.csv")
#+++++++CORRELATION MATRIX++++++++++++++++++
cell.cor<-round(cor(datascale),3)
cell.cor
write.csv(cell.cor,"Correlation.csv")
#Biplot************************************
plan0_sna <- na.omit(datascale)
matrix0 <- data.matrix (datascale)
matrix1<-matrix0[,1:ncol(matrix0)]
rownames(matrix1)<-plan0_sna$LOCAL
m2 <- princomp(matrix1, cor=TRUE)
biplot(m2)
#K-MEANS CLUSTERING SECTION-PARTITIONAL
# Determine number of clusters
wss <- (nrow(datascale)-1)*sum(apply(datascale,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(datascale,
```

```
                                centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
    ylab="Within groups sum of squares")
# Apply k-means Clustering with k=4
#no.group<-3
#no.iter<-500
k <- kmeans(FirstPCs, 3, nstart=25, iter.max=1000)
print(k)
clust <- names(sort(table(k$clust)))
palette(alpha(brewer.pal(9,'Set1'), 0.5))
plot(FirstPCs, col=k$clust, pch=16)
# Interactive version of3-D
plot3d(FirstPCs$PC1, FirstPCs$PC2, FirstPCs$PC3, col=k$cluster)
plot3d(FirstPCs$PC1, FirstPCs$PC3, FirstPCs$PC4, col=k$cluster)
#       Creating Centroids
library(cluster)
clusplot(pca.features,k$clust, main='2D representation of the Cluster solution',
        color=TRUE, shade=TRUE,
        labels=2, lines=0)
#Evaluate using PCA Scatter plot
pcaPC1<-pcaPC$scores[,1]
pcaPC2<-pcaPC$scores[,2]
plot(pcaPC1,pcaPC2,main ="PCA for the first two Principal components",
    xlab="PC1",ylab="PC2",col=c(k$cluster),pch=(k$cluster))
#------------Scatter plots----------------------------------------------------
```

```
plot(datascale, col=k$cluster, pch=16)
#Number of observation per cluster
table(k$cluster)
# Cluster sizes
sort(table(k$clust))
clust <- names(sort(table(k$cluster)))
# First cluster
row.names(datascale[k$clust==clust[1],])
# Second Cluster
row.names(datascale[k$clust==clust[2],])
# Third Cluster
row.names(datascale[k$clust==clust[3],])
# Fourth Cluster
#row.names(datascale[k$clust==clust[4],])
#------Within the sumsquares-----
withinSS<-k$withinss
withinSS
#------Between the sumsquares-----
btwnSS<-k$betweenss
btwnSS
#Confusion Matrix
table(pcahealth$County,k$cluster)
confMat<-table(pcahealth[,1],k$cluster)
write.csv(confMat,"ConfusionMatrix.csv")
# ======Average Hierarchical Clustering==============================
```

```
county.dist = dist(datascale)
county.hclust = hclust(county.dist)
plot(county.hclust,labels=pcahealth$County,main='Cluster Dendrogram Visualization')
#========Heat Map================================
distMatrix <- as.matrix(dist(datascale[,1:12]))
heatmap(distMatrix)
#----------Dissimilarity Matrix---------------------
DissMatrix<-round(as.matrix(daisy(pcahealth,metric = "gower")),3)
DissMatrix
write.csv(DissMatrix,"DissMatrix.csv")
#   Agglomerative Nesting(AGNES)
agn2 <- agnes(daisy(datascale), diss = TRUE, method = "complete")
print(agn2)
plot(agn2)
## Cut into 3 groups:
agn2 <- cutree(as.hclust(agn),k = 3)
table(agn2)
#   Divisive Analysis(DIANA)
diana <- diana(pcahealth, metric = "Euclidean", stand = TRUE)
print(diana)
plot(diana)
## Cut into 3 groups:
dv2 <- cutree(as.hclust(diana), k = 3)
table(dv2)
# calculate distances between objects and cluster centers
```

```
centers <- k$centers[k$cluster, ]
distances <- sqrt(rowSums((datascale - centers)^2))
# pick top 5 largest distances
outliers <- order(distances, decreasing=T)[1:5] # who are outliers
print(outliers)
# ======Other clustering alogrithms============================
#1) Partationing Around Medoids
thepamx<-pam(pcahealth,3)
thepamx
#1) Fuzzy
fannyx<-fanny(pcahealth,3)
fannyx
#1) Clustering Large Appliations
clarahealth<-clara(pcahealth,3)
clarahealth
summary(clarahealth)
plot(clarahealth,labels =2)
summary(thepamx)
plot(thepamx,labelS=2)
summary(fannyx)
plot(fannyx,labels =2)
```