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

Nicholas M Njiru, Elisha T O Opiyo

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

Nicholas M. Njiru
Multimedia University of Kenya
Email: nnjiru@mmu.ac.ke
Elisha T.O. Opiyo
University of Nairobi
Email: opiyoAuonbi.ac.ke

Elisha T.O. Opiyo
University of Nairobi
Email: opiyoAuonbi.ac.ke

ABSTRACT
The inauguration of the new constitution in Kenya has led to the devolution of health care in the counties. It is against this backdrop that has necessitated the need to develop a model of grouping these regions 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 the 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 propose a new model that clusters 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 12, 8 and 27 observations respectively. The study classified the clusters as well-off, most marginalized and moderately marginalized counties. The methodology developed was objective, replicable and sustainable to create the clusters. It was developed in a theoretically sound principle and can generalize 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.

I. INTRODUCTION

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 has 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. Thedirect and indirect expenditure related with child health are extremely huge. This has contributed to poor economic performance of developing countries. In Kenya previous research has been 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.

II. METHODOLOGY

Introduction

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

Research Design

In this research, CRISP-DM methodology will be used. There several Data mining methodologies such as CRISP_DM, SEMMA, KDD that exist. The choice of this methodology is due to its acceptance in data mining and also because the model is designed for 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 (J.Taylor, 2014). Available from:

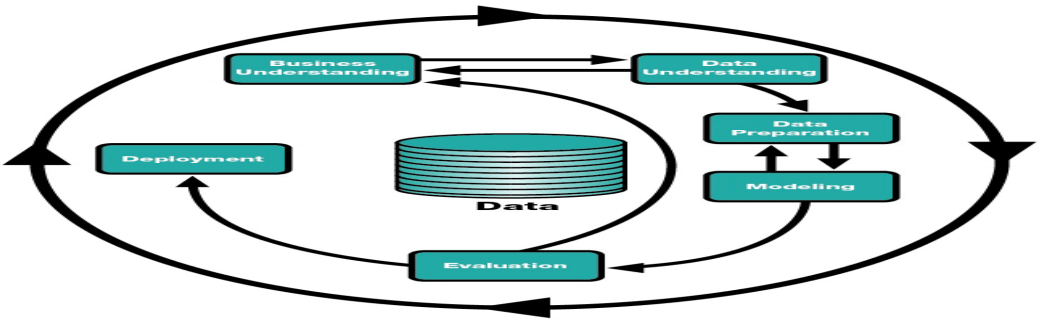


Figure 1: CRISP-DM Process model

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

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.

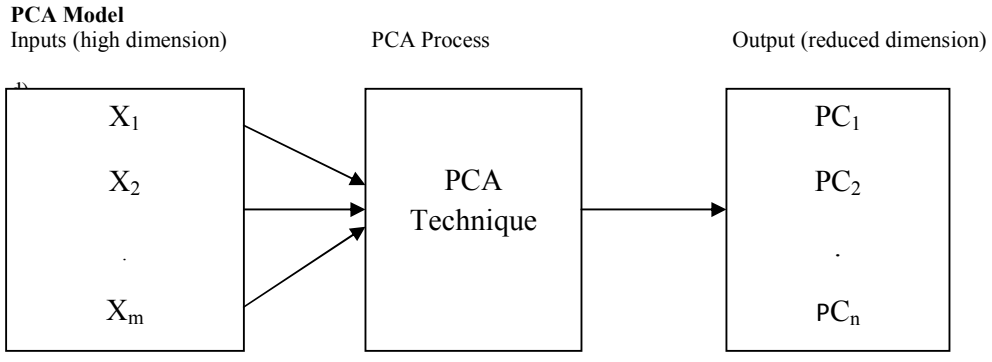


Figure 2: PCA model

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.

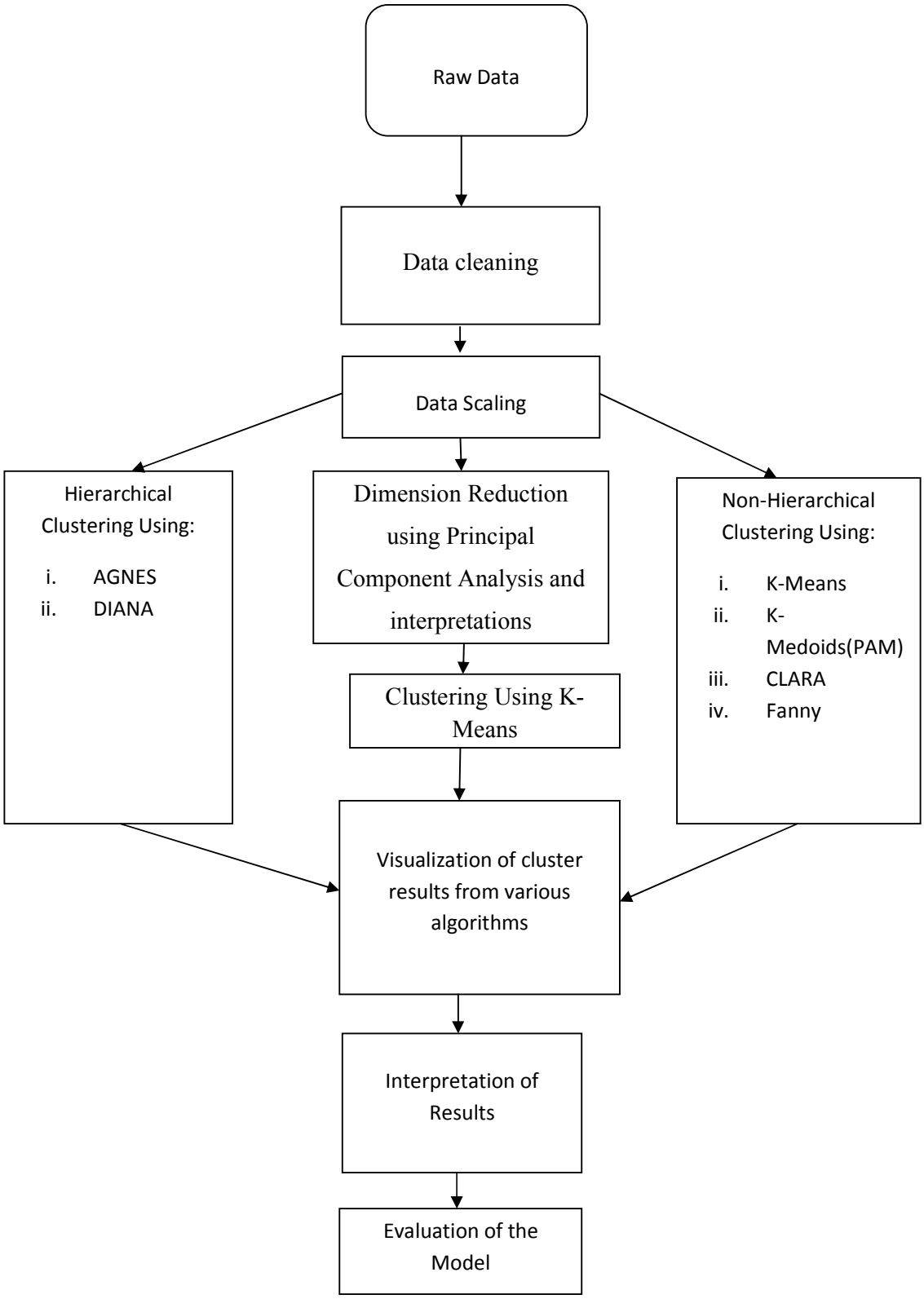
PCA uses the correlation matrix which is similar to using PCA based on the variance-covariance of the standardized variables. Since standardized variables contain variance equal to 1, the totals of the eigenvalues is n, the number of variables.

Source of data and study Population

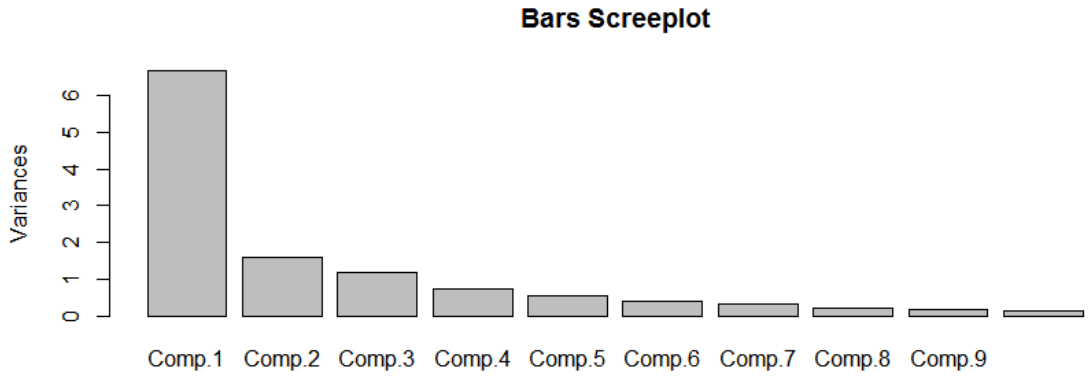
Secondary data collected from Kenya National Bureau of Statistics, Commission of Revenue Allocation, Kenya HIV and AIDS profile per county, Statistical Abstract 2014, Kenya Economic report of 2014, and Kenya County Profile, Kenya Demographic and Health Survey of 2014 and e-health facilities.

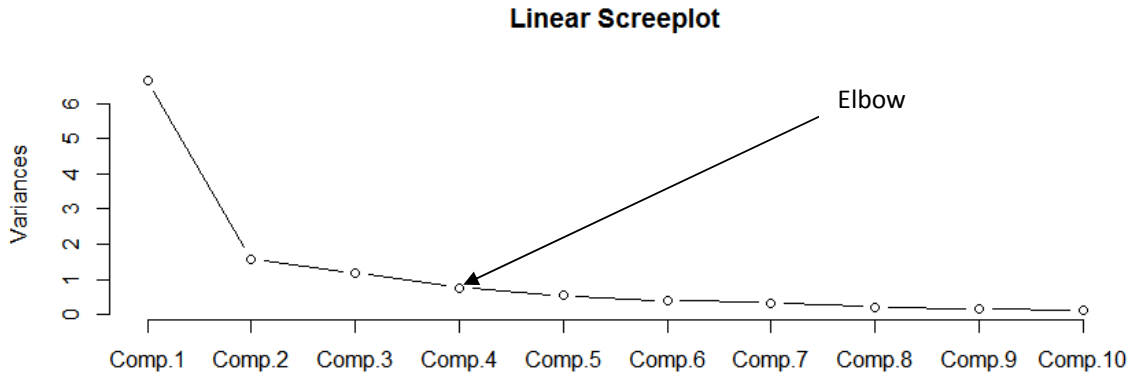
The major demerit of secondary data collected by other researchers is that they controlled, decided what to collect and what to exclude and therefore the entire information desired for this research may not be available.

Proposed Framework



III. RESULTS






```
Importance of components:
                          Comp.1     Comp.2     Comp.3
Standard deviation     2.5811064 1.2566677 1.08499441
Proportion of Variance 0.5551759 0.1316011 0.09810107
Cumulative Proportion  0.5551759 0.6867770 0.78487808
                           Comp.4     Comp.5     Comp.6
Standard deviation     0.86948560 0.73269429 0.6265889
Proportion of Variance 0.06300043 0.04473674 0.0327178
Cumulative Proportion  0.84787852 0.89261526 0.9253331
                           Comp.7     Comp.8     Comp.9
Standard deviation     0.57456224 0.45722116 0.40475197
Proportion of Variance 0.02751015 0.01742093 0.01365201
Cumulative Proportion  0.95284321 0.97026414 0.98391616
                          Comp.10     Comp.11
Standard deviation     0.35086932 0.243701746
Proportion of Variance 0.01025911 0.004949212
Cumulative Proportion  0.99417526 0.999124476
                           Comp.12
Standard deviation     0.1025001641
Proportion of Variance 0.0008755236
Cumulative Proportion  1.0000000000
```

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.

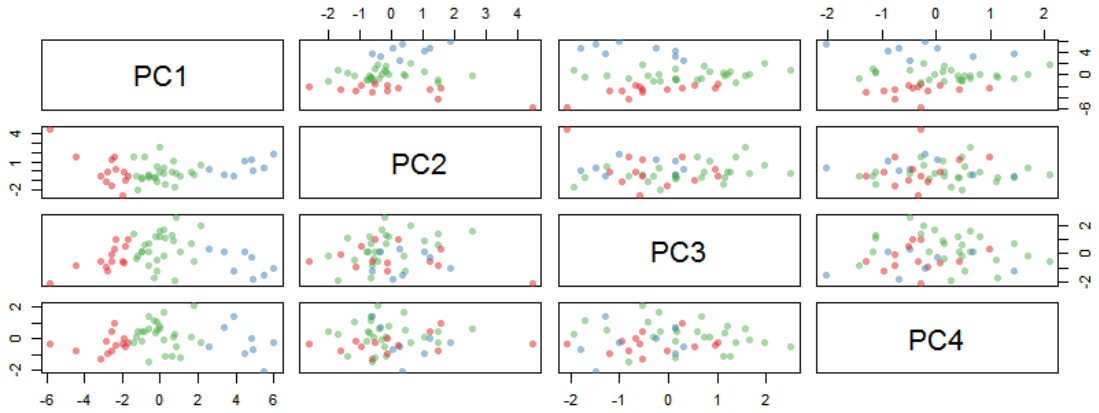


Figure 3 - Correlation Matrix of the First Four PCs

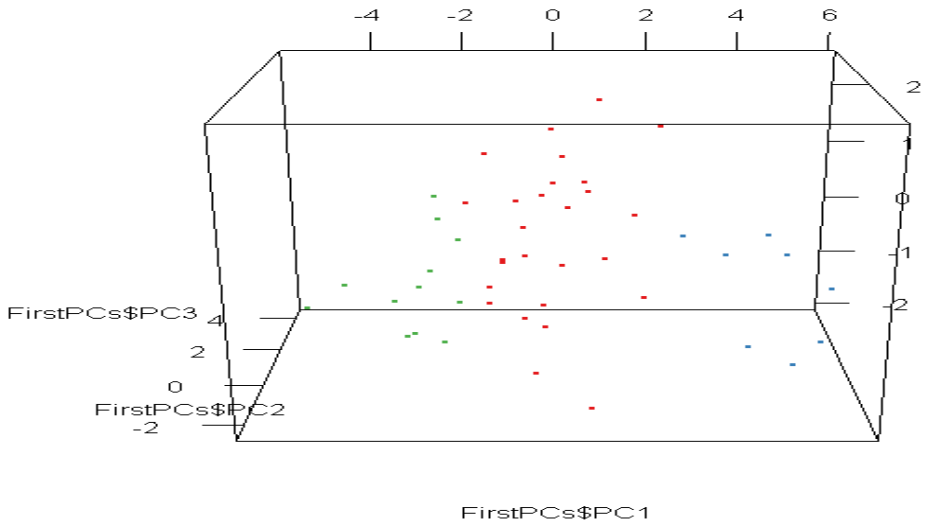


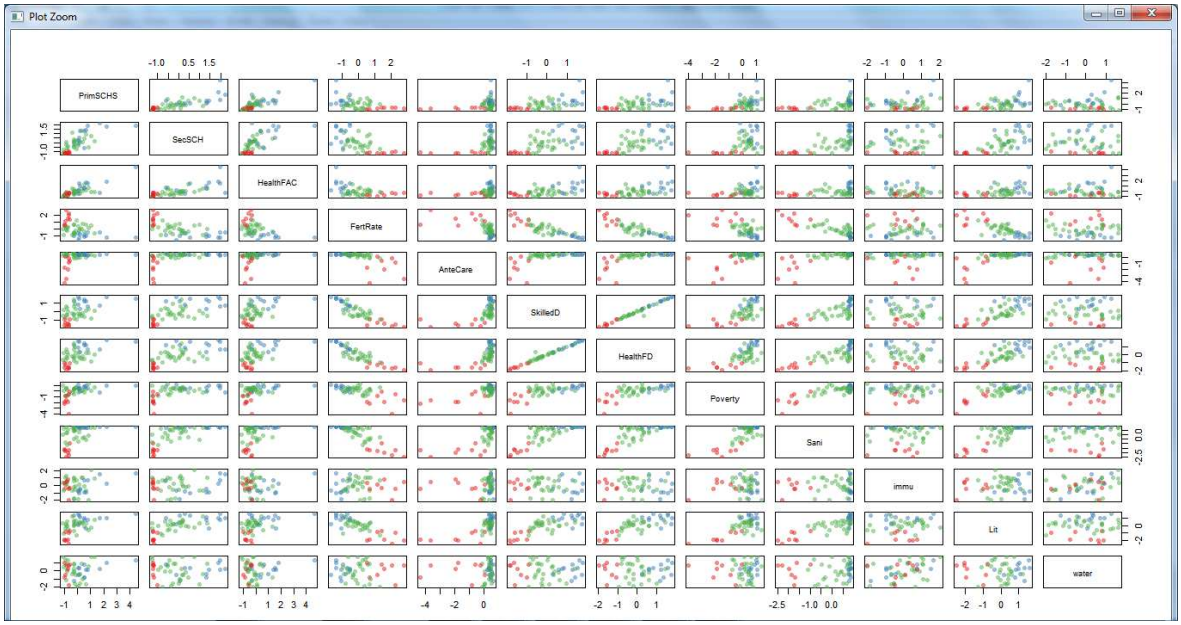
Figure 4-3-Dimension View of PC1, PC2 and PC3

Results

The figure 12 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 13) 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 12 variables. By doing this we were able to reduce the 12 dimension 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.

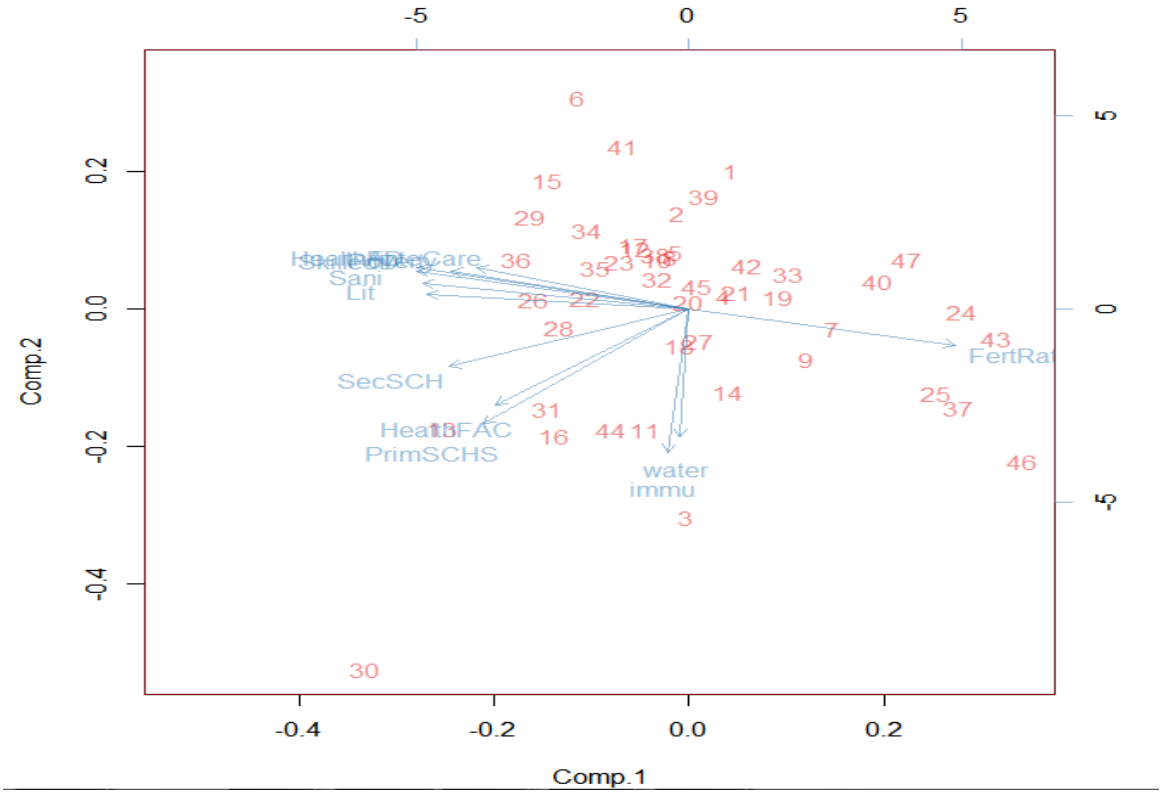
4.2.2.4 Scatter plot



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 variables. There was a strong negative correlation between fertility rate and skilled deliveries, health facilities, poverty, sanitation, literacy and secondary schools.

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 in 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.



Results

Fertility rate was the variable that had the most influence of component one. 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 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 county of Bungoma was relatively high and variables water and immu were the most influential variable. The position of the observation Turkana County was

mostly influenced by the variable FertRate with average influence of the county of Garissa. The counties of Kirinyaga, Nyamira, Murang'a and Embu were highly influenced by the variables HealthD, HealthFAC, AnteCare, SkilledD, Sani, Lit and Poverty.

4.2.2.6 Correlation Matrix

	PrimSCHS	SecSCH	HealthFAC	FertRate	AnteCare	SkilledD	HealthFD	Poverty	Sani	immu	Lit	water
PrimSCHS	1	0.747	0.823	-0.438	0.319	0.515	0.513	0.437	0.449	0.228	0.516	0.147
SecSCH	0.747	1	0.683	-0.606	0.448	0.621	0.624	0.477	0.675	0.137	0.671	-0.012
HealthFAC	0.823	0.683	1	-0.456	0.264	0.523	0.524	0.363	0.359	0.063	0.446	0.095
FertRate	-0.438	-0.606	-0.456	1	-0.59	-0.873	-0.878	-0.655	-0.841	0.025	-0.764	-0.056
AnteCare	0.319	0.448	0.264	-0.59	1	0.556	0.581	0.608	0.693	0.017	0.675	0.033
SkilledD	0.515	0.621	0.523	-0.873	0.556	1	0.989	0.688	0.769	-0.059	0.749	-0.038
HealthFD	0.513	0.624	0.524	-0.878	0.581	0.989	1	0.689	0.772	-0.074	0.755	-0.056
Poverty	0.437	0.477	0.363	-0.655	0.608	0.688	0.689	1	0.784	0.063	0.7	-0.068
Sani	0.449	0.675	0.359	-0.841	0.693	0.769	0.772	0.784	1	0.061	0.783	0.11
immu	0.228	0.137	0.063	0.025	0.017	-0.059	-0.074	0.063	0.061	1	0.125	0.337
Lit	0.516	0.671	0.446	-0.764	0.675	0.749	0.755	0.7	0.783	0.125	1	-0.026
water	0.147	-0.012	0.095	-0.056	0.033	-0.038	-0.056	-0.068	0.11	0.337	-0.026	1

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	Comp.12
BARINGO	0.752371	1.731245	-1.97481	0.275183	0.612171	-0.13334	-0.25703	0.425179	-0.1396	0.295099	0.067386	0.023438
BOMET	-0.20058	1.196818	-0.43821	1.114372	0.335993	-0.14677	0.756775	0.173749	-0.01472	0.64168	-0.29723	0.108553
BUNGOMA	-0.04854	-2.61082	1.59714	0.626459	0.693524	0.096829	0.694051	-0.63117	-0.3184	0.255218	-0.12624	-0.02953
ISIOLO	2.131484	-0.62344	1.992128	-0.24756	0.046433	0.186312	-0.78336	0.191833	-0.01919	0.020858	0.228924	0.033105
KAJIADO	-0.58677	0.615991	-0.82438	-1.45004	1.09817	-1.80767	0.160798	-0.08841	-0.43288	0.143576	0.307252	0.026733
KAKAMEGA	-0.79747	-1.51553	0.601359	0.492573	0.606797	0.599459	0.881755	-0.21694	0.326832	-0.29495	-0.0085	0.027929
KERICHO	-0.99238	0.749499	0.167102	0.431744	0.068458	-0.03886	0.187288	0.155866	0.092987	-0.10212	-0.08254	0.074307
KIAMBU	-4.48116	-1.4972	-0.83376	-0.77601	-0.35574	-0.12841	-0.02832	-0.12522	-0.49545	-0.10141	0.356488	-0.03086
KILIFI	0.69476	-1.04843	0.863174	0.12533	0.233661	0.110275	-0.9749	0.157666	0.153743	-0.39808	-0.2347	-0.03106
KIRINYAGA	-2.57833	1.614888	-0.03979	-0.77271	-1.13399	-0.52998	-0.33768	-0.50479	0.022528	0.233492	0.119167	0.009332
MAKUENI	-1.25518	0.581481	-0.32207	0.152435	0.643716	1.293807	0.61861	0.633905	0.862322	-0.06297	-0.01012	0.064149
MANDERA	4.964497	-0.03582	-1.817	-0.6919	-2.73417	0.320177	1.002265	0.256043	0.17909	0.466004	-0.14448	-0.02887
NAIROBI	-5.87976	-4.52046	-2.09435	-0.2959	-0.05642	-0.91118	-0.90843	0.361086	0.298245	0.448442	-0.33336	0.009018
NAKURU	-2.59004	-1.2489	-0.53274	0.431928	-0.04753	0.463297	0.462193	-0.11408	0.34141	-0.64738	0.135934	-0.02066
NYERI	-3.15664	0.619976	-0.53418	-1.30752	-0.40288	0.022015	-0.29816	0.601969	0.468785	-0.52454	-0.17481	-0.02847
SAMBURU	4.895493	-1.23494	-0.24897	0.014247	-0.79323	0.15331	-0.1604	-0.18066	-0.04884	0.079026	0.060749	0.082903
SIAYA	-0.59383	0.674298	0.669796	1.446314	-0.91572	-0.01583	-0.56747	-0.72339	0.282019	-0.58713	-0.00844	0.027883
TAITA TAVETA	0.263387	1.403212	1.132875	-1.13939	-0.14596	-0.14262	0.03339	-1.01669	0.423792	0.416162	0.029453	0.031694
TURKANA	5.568702	-0.36753	-1.51246	-2.05653	1.708604	1.502979	-1.1441	-0.49733	-0.39649	0.150517	-0.31309	-0.00605
UASIN GISHU	-1.43036	-1.50653	1.391545	-0.2834	0.14213	-0.27328	0.444223	0.371855	-0.45065	0.735546	0.088305	0.018171
VIHIGA	0.156044	0.2891	1.688994	-0.03553	0.414421	-0.12995	0.691766	0.027045	0.152307	-0.25846	-0.17152	-0.03314
WAJIR	6.046304	-1.90716	-1.01355	-0.20612	-1.19134	-0.68475	0.714651	0.469225	-0.26449	-0.79006	-0.05946	-0.03573
WEST POKOT	3.94757	0.611616	-1.29644	1.452209	0.660931	-0.5898	-0.10984	0.241502	-0.2118	-0.28778	-0.11373	0.008024

Figure 5-Scores plot

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12
PrimSCHS	-0.26658	-0.43079	-0.32682	0.0484	0.14955	-0.18589	-0.00095	0.00315	-0.4805	0.58795	-0.02656	0.00064
SecSCH	-0.30913	-0.21449	-0.2445	0.15397	0.11639	0.4892	0.44608	-0.22028	-0.07542	-0.36388	0.36941	0.00281
HealthFAC	-0.25045	-0.35877	-0.46438	-0.1375	0.11167	-0.20299	-0.24786	0.05743	0.58618	-0.24763	-0.22504	0.00395
FertRate	0.34381	-0.13507	-0.10207	0.21405	0.26853	-0.18581	0.00506	0.029	-0.47402	-0.5507	-0.41392	-0.02275
AnteCare	-0.27361	0.15592	0.2904	0.11688	0.6859	0.12171	-0.49408	-0.25816	-0.0056	0.00533	0.0743	0.03494
SkilledD	-0.35154	0.14347	-0.04013	-0.17477	-0.36706	-0.0218	-0.23895	-0.12058	-0.28846	-0.21557	-0.07997	0.69337
HealthFD	-0.35331	0.15696	-0.04499	-0.16528	-0.32887	-0.00632	-0.26532	-0.13482	-0.26225	-0.19042	-0.04909	-0.71902
Poverty	-0.30737	0.14435	0.16402	0.21945	0.05864	-0.76247	0.29899	0.00962	0.04381	-0.16551	0.32557	-0.00479
Sani	-0.34435	0.09813	0.25003	-0.01968	0.0615	0.06845	0.47065	-0.21395	0.12317	0.13301	-0.70715	-0.01465
immu	-0.02734	-0.53683	0.42384	0.54982	-0.37786	0.03635	-0.21789	-0.15425	0.10839	-0.0205	-0.03917	-0.00626
Lit	-0.33942	0.05479	0.11846	0.20076	0.04625	0.21296	-0.02973	0.87506	-0.05985	-0.06582	-0.07338	-0.00668
water	-0.01267	-0.47986	0.48609	-0.66824	0.11843	-0.04847	0.09315	0.11211	-0.10334	-0.1572	0.11643	-0.0131

Figure 6-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.

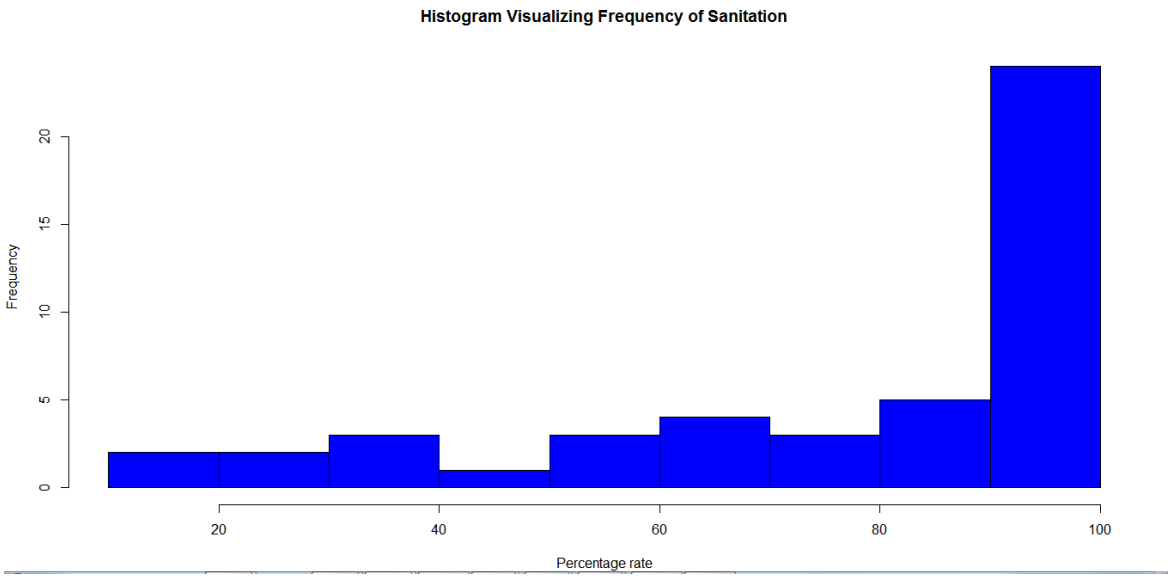
Summary Statistics

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		PrimSCHS	SecSCH	HealthFAC	FertRate	AnteCare	SkilledD	HealthFD	Poverty	Sani	immu	Lit	water	
2		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	
3		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	
4		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	
5		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	
6		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	
7		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	

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%.Aprroximately 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.

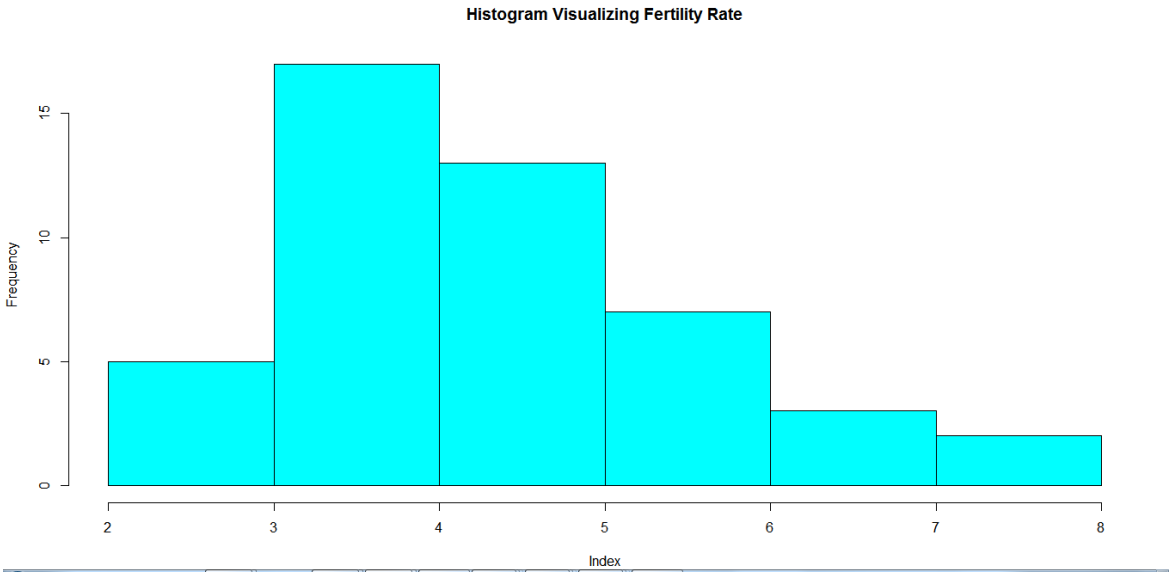
Histogram Plots



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%.



Results

The histogram depicts approximately 16 counties fertility rate is in the range of index 3 to 4 with majority counties are concentrated between the index of 3 to 6.

Modeling

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).

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.

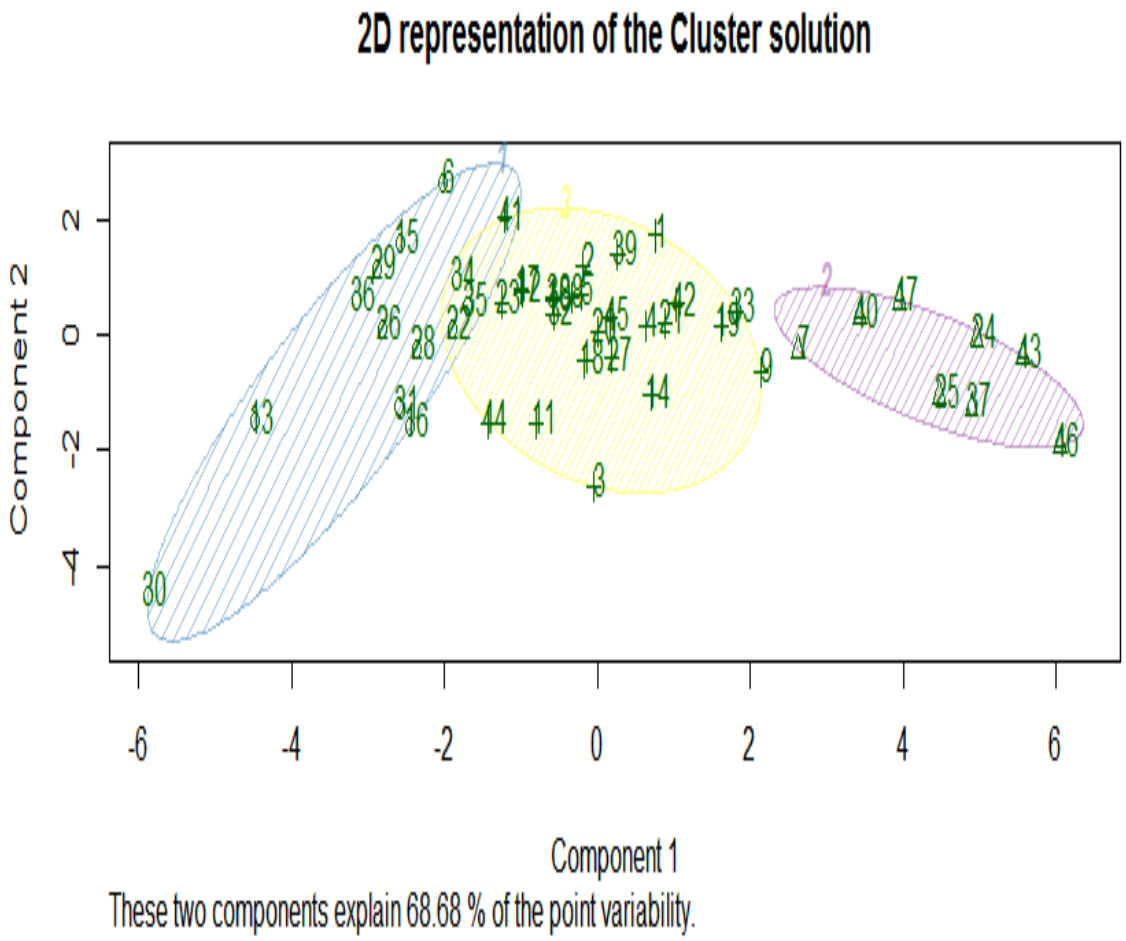


Figure 7: k-means clustering results

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

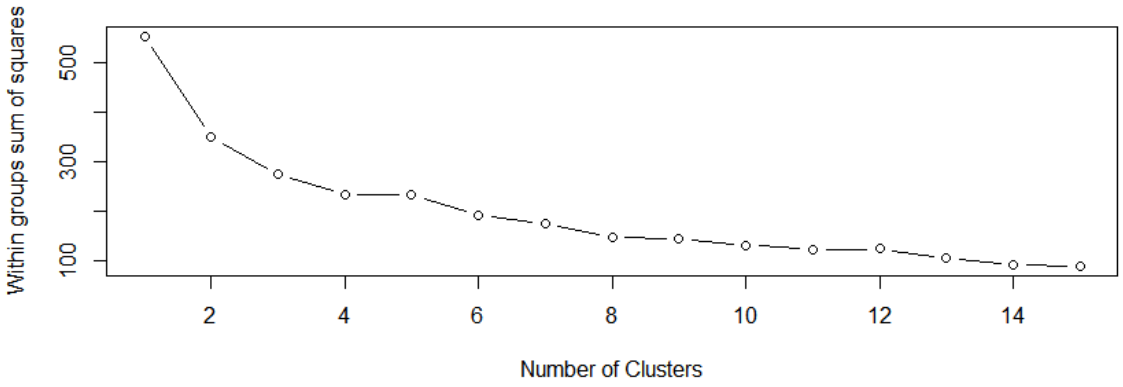
Figure 8: Counties' Key

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 representing the clusters.

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.



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 and therefore applied k-means clustering function with k-4 and plotted the results.

We then looked at our clusters in order of increasing size. The first cluster contained 12 counties, second cluster contained 8 while the third cluster contained 27 counties. Cluster one was made up of the well-off counties, cluster two was made up of the most marginalized 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.

Use of Box Plots

We used the box plots to compare, literacy, healthcare delivery and fertility rates in the clusters. In literacy, cluster one was the highest with an outlier, followed by the cluster three and then cluster two had the lowest literacy level. The fertility rate is very low in cluster one followed by cluster three but highest in cluster two. Those seeking healthcare delivery was highest in cluster one followed by cluster three and lowest in cluster two. The sanitation was highest in cluster one followed by cluster three with the lowest being cluster two.

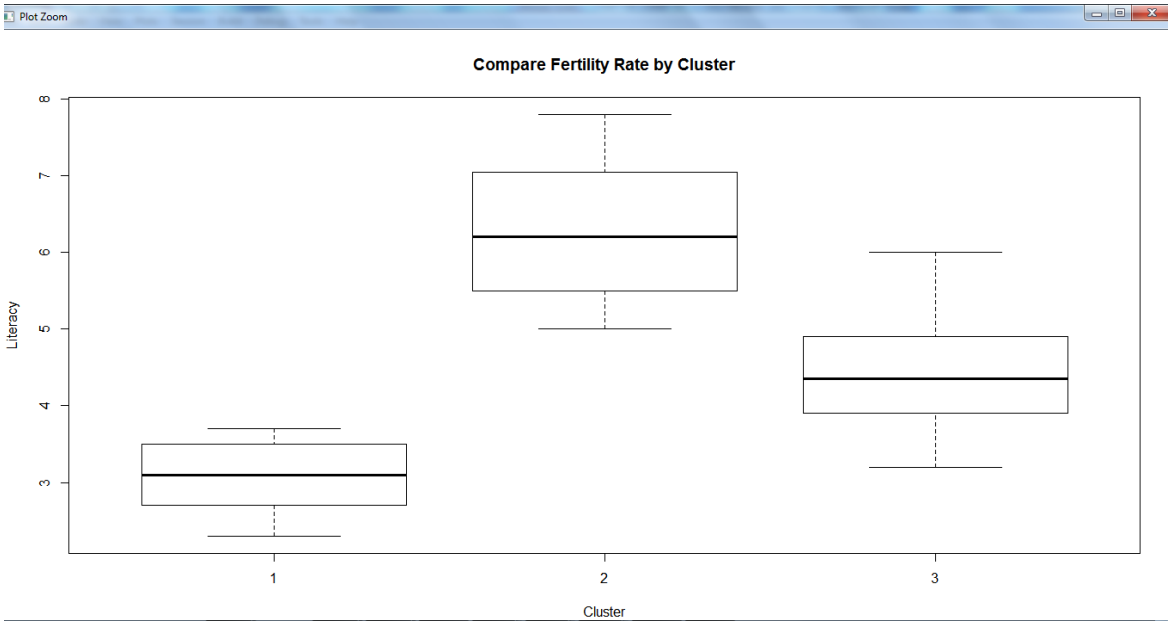


Figure 9: Comparing Fertility Rate by Cluster

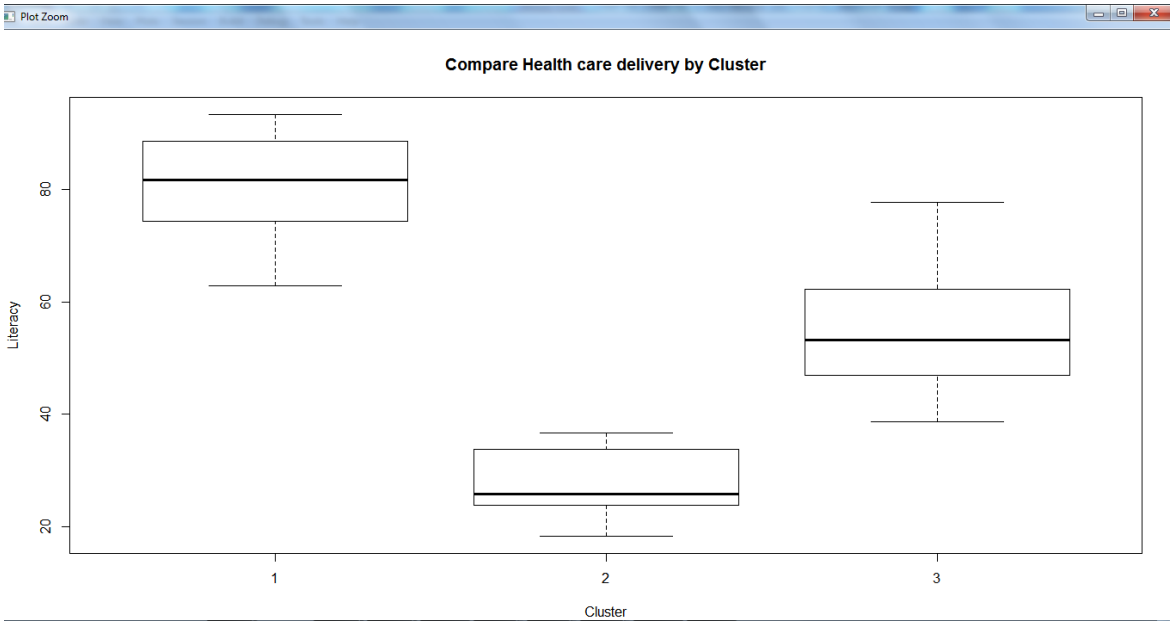


Figure 10: Comparing Healthcare by cluster

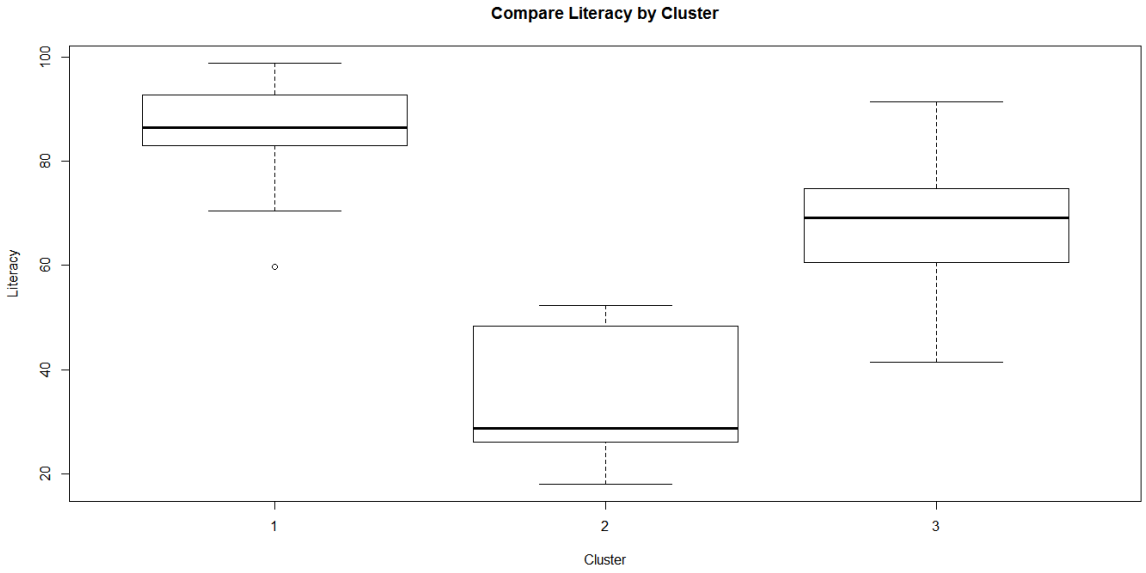


Figure 11-Compare Literacy by Cluster

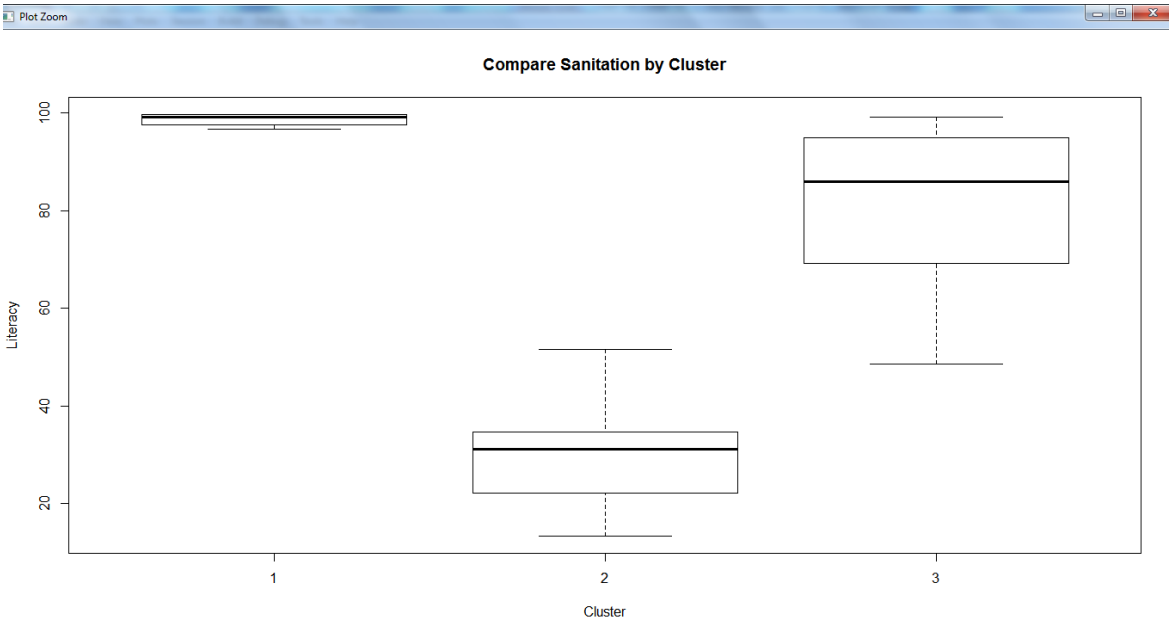
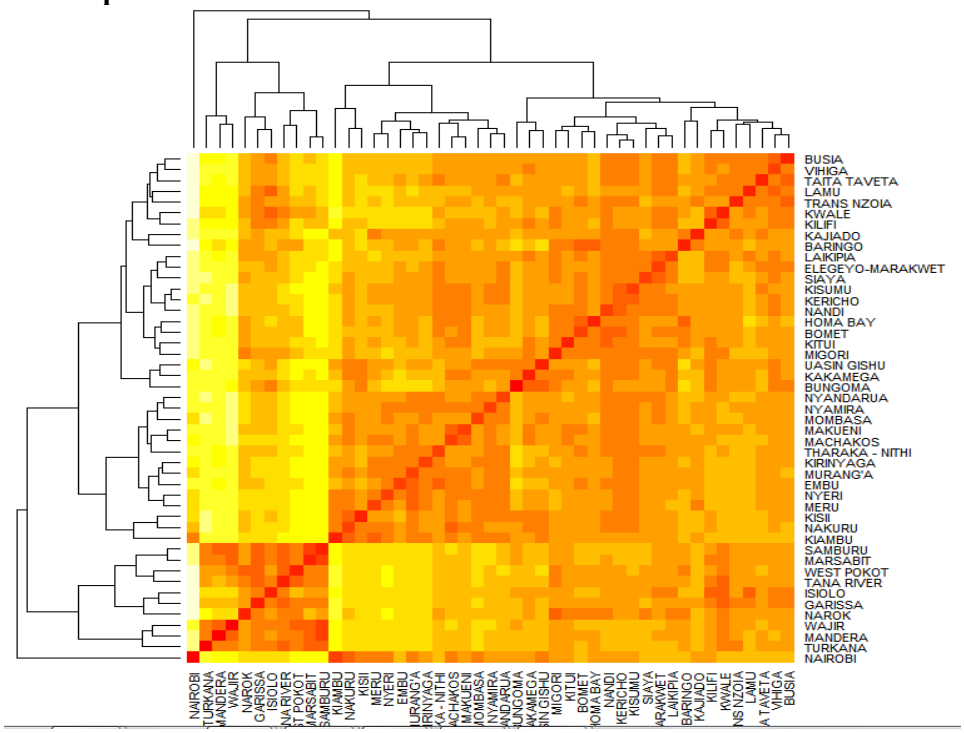


Figure 12-Compare Sanitation by Cluster

Dissimilarity Visualization

Heatmap

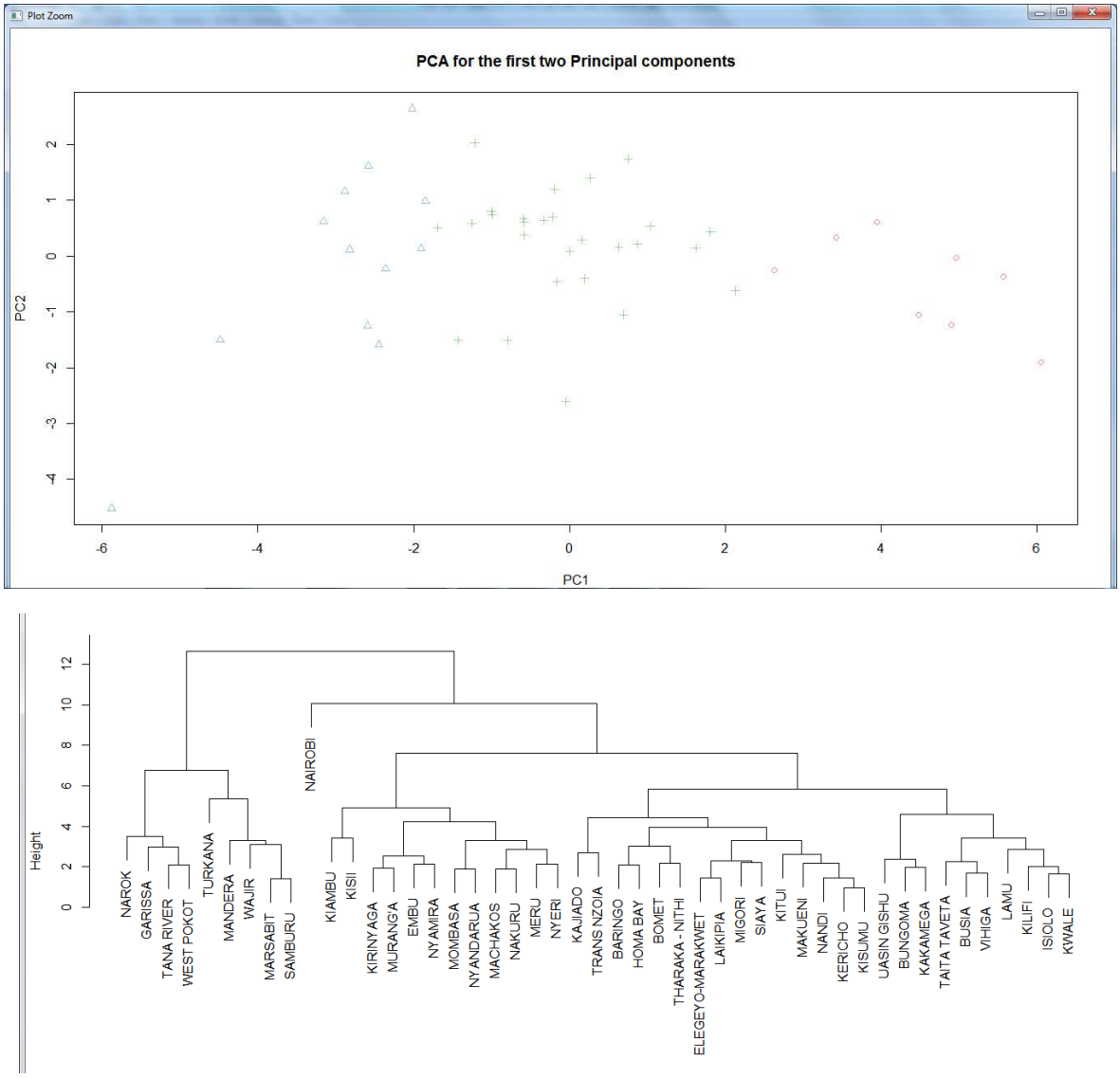


Dissimilarity Matrix

		BARINGO	BOMET	BUNGOMA	EMBU	GARISSA	KERICHO	KIAMBU	KILIFI	KIRINYAGA	MACHAKOS	MURANG'A	NAIROBI	SAMBURU	TANA RIVER	UASIN GISI	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

Ready



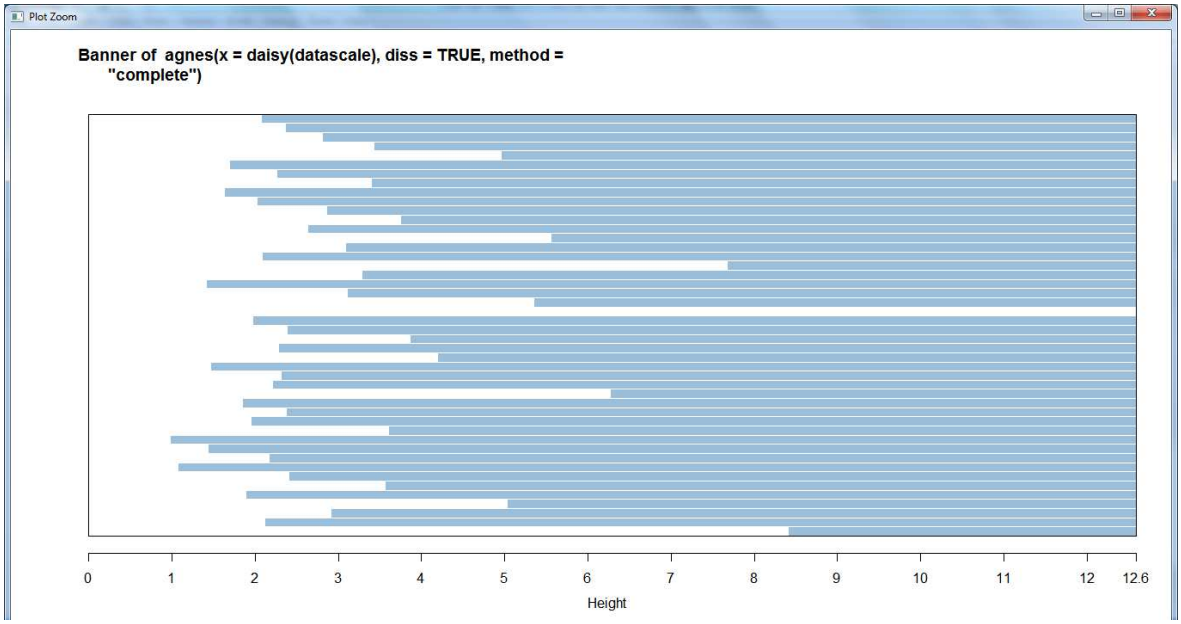
Hierarchical Clustering and Bannerplot

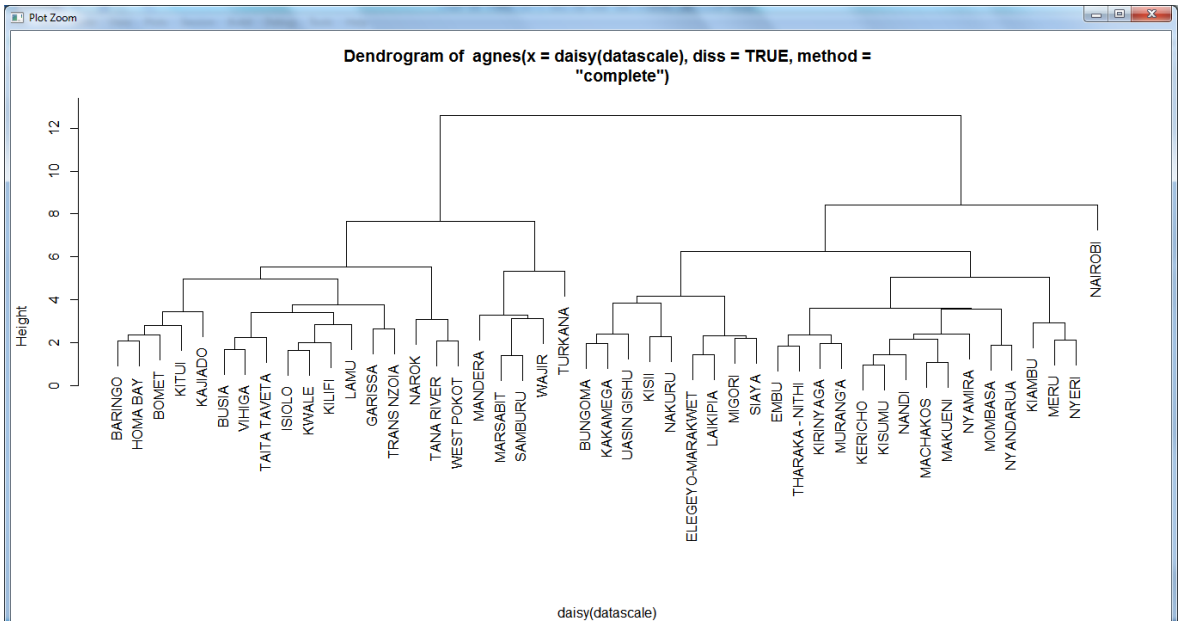
Hierarchical Clustering draws a "banner", i.e. basically a horizontal bar plot visualizing the (agglomerative or divisive) hierarchical clustering or any other binary dendrogram structure.

Agglomerative Coefficient (AC)

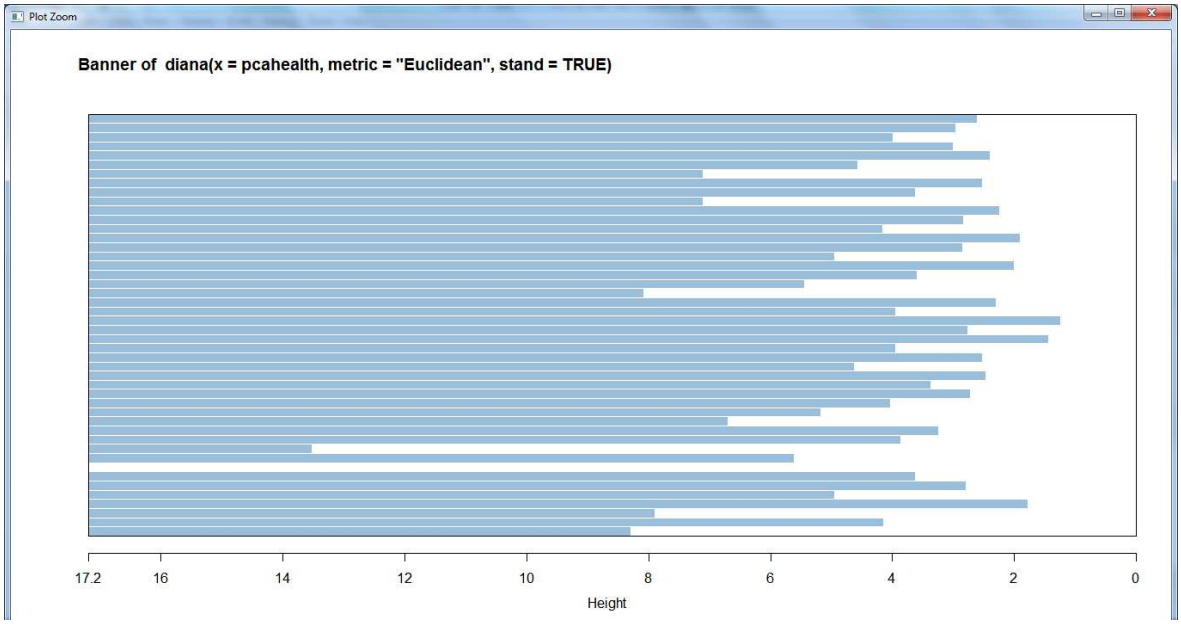
This refers to the measure of how much clustering structure exists in the data. A large AC (close to one) means that there is a strong clustering structure. A small AC means that the data is more evenly distributed hence a poor clustering structure.

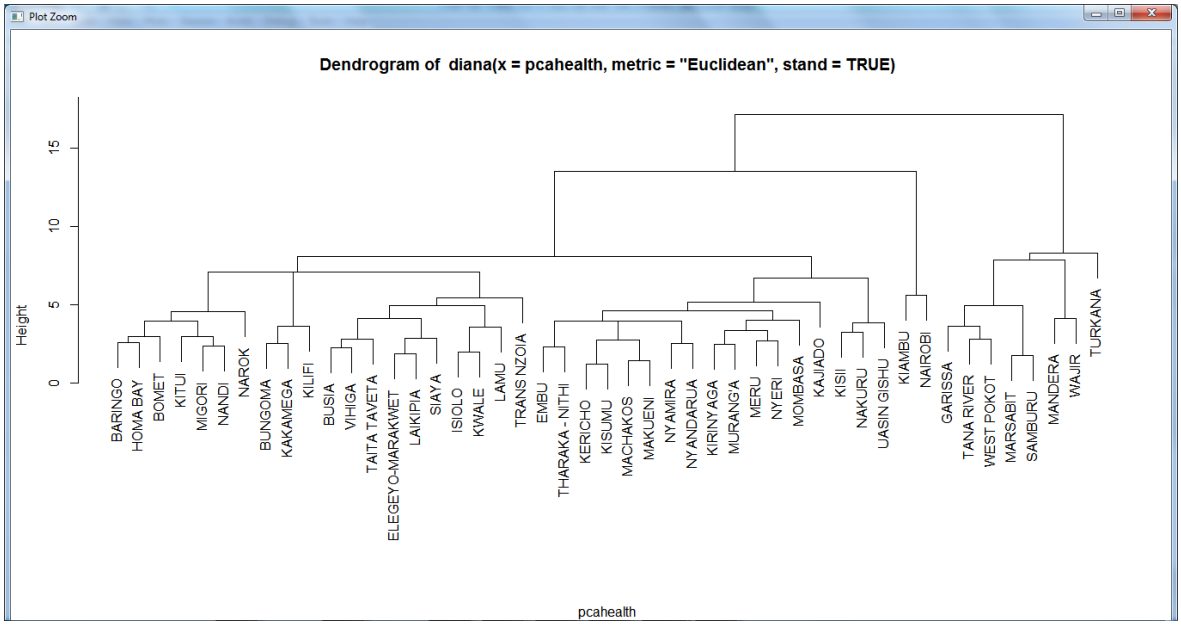
Agglomerative Analysis (AGNES) and agglomerative coefficient





Divisive Analysis (DIANA) and divisive coefficient





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.

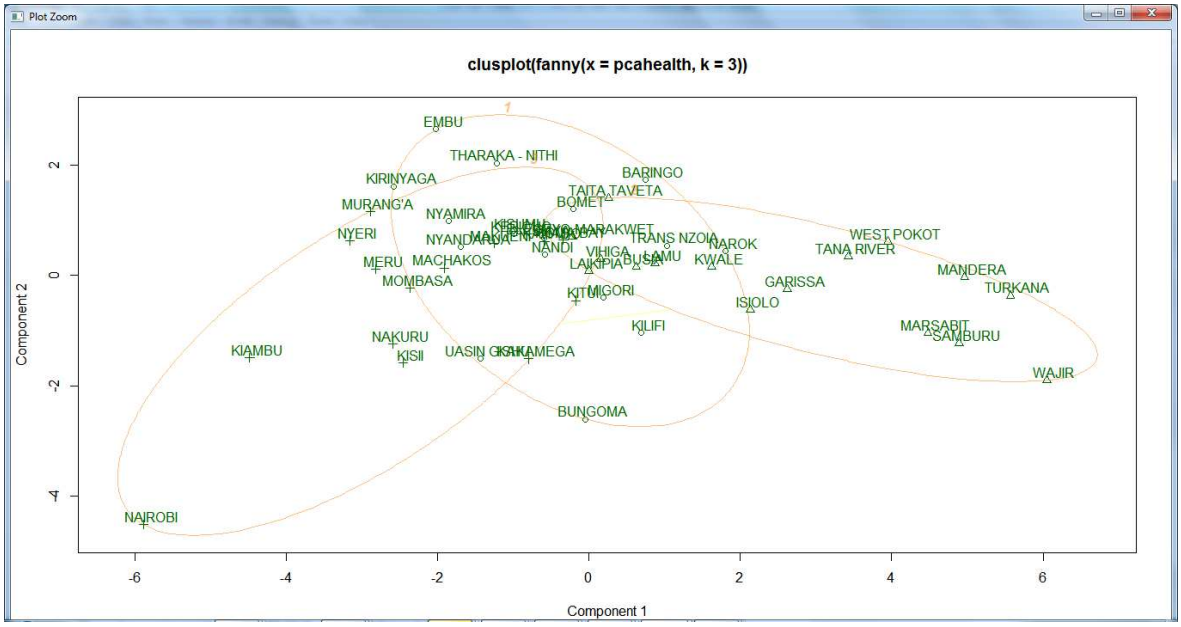
Interpretation of Silhouette Coefficient

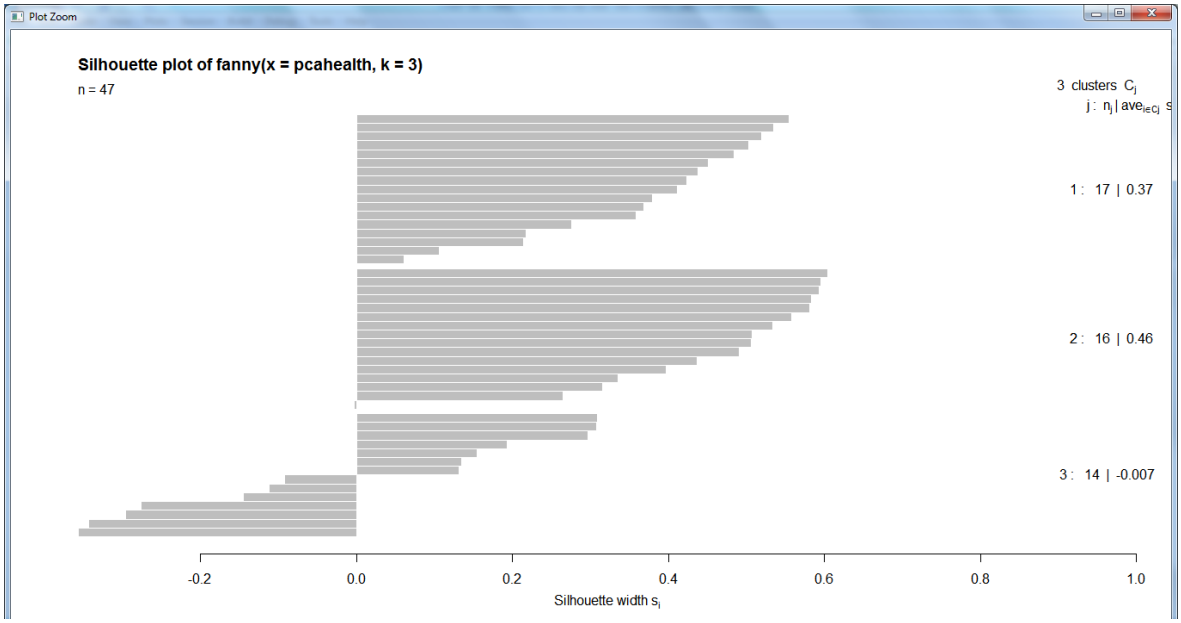
Silhouette Coefficient	Explanations
0.71-1.00	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial. Try additional methods of data analysis.
<=0.25	No substantial structure has been found

Other non-hierarchical Clustering Algorithms

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 fuzzy allows for some ambiguity in the data, which often occurs in practice.





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
```

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

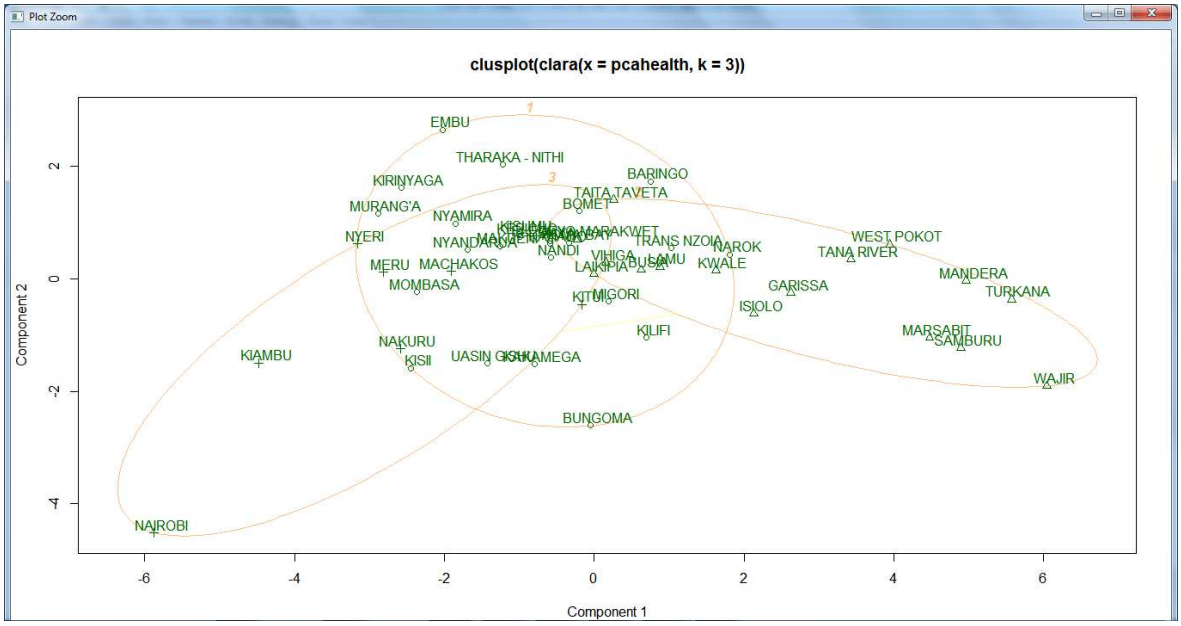
This algorithm generated a three cluster solution with the size of 24, 16 and 7. We however discarded its output because its silhouette coefficient was very low at 0.35 meaning that the structure was weak and could be artificial. More detailed results are shown below for silhouette width per cluster.

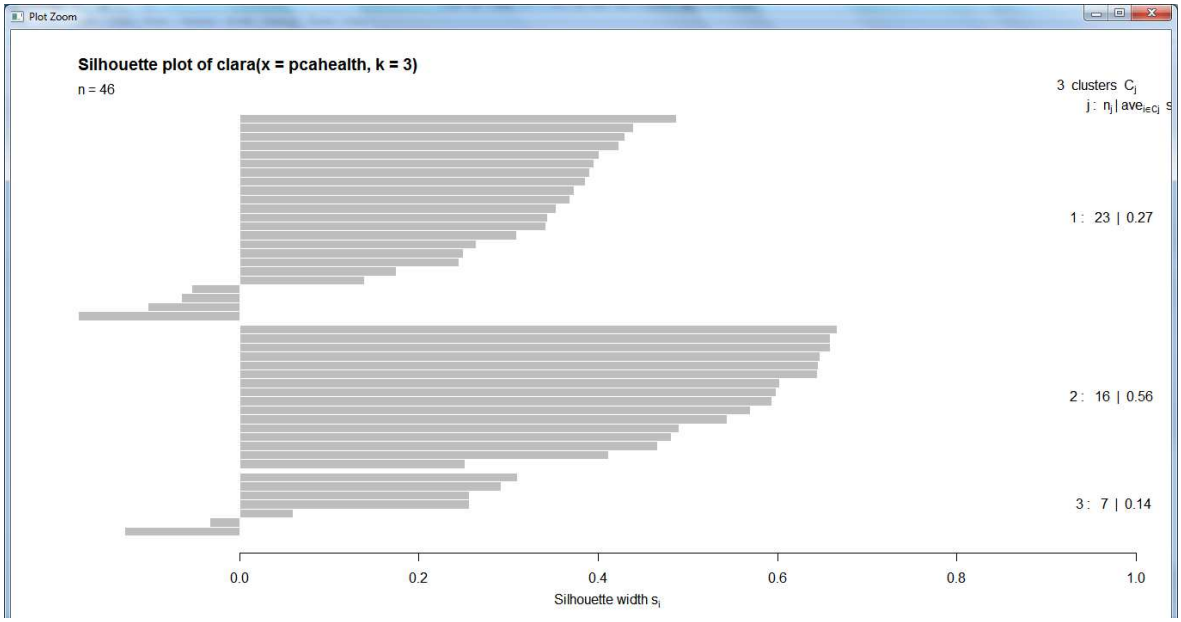
```
Average silhouette width per cluster:
[1] 0.2720443 0.5547723 0.1442289
Average silhouette width of total data set:
[1] 0.3492558

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
```

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.





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
```

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.

Cluster	Observations	%	Counties Name	Class
1	12	26%	Embu, Kiambu, Kirinyaga, Kisii, Machakos, Meru, Mombasa, Murang'a, Nairobi, Nakuru, Nyamira, Meru.	Well-off
2	8	17%	Garissa, Mandera, Marsabit, Samburu, Tana-River, Turkana, Wajir, West-Pokot	Most Marginalized
3	27	57%	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.	Moderately Marginalized

This shows that 17% of the counties have the most disadvantaged children, 26% are well-off and 57% are moderately disadvantaged.

We used box plots to compare the three clusters of literacy, health care delivery, sanitation and fertility rates. Cluster one was doing well in literacy, followed by cluster three and cluster two was highly disadvantaged. The literacy level in cluster one was above 80% but below 95%, cluster two was below 45% whereas cluster three was between 60% and 70%. Cluster two health care deliveries and sanitation was below 30%. In contrast the fertility rate for cluster two was very high with an index of between 5.5 and 7.

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.

DISCUSSIONS AND CONCLUSION

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.

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. Since the fertility rate of the most marginalized counties is very high, creating awareness towards sustainable Family Planning practices among marginalized counties is necessary. This can be done by helping women and couples realize the reproduction intentions so as to get healthy families. To achieve this there should be increased knowledge of the family planning methods and services through the assistance of the community health workers and non-governmental organizations to provide accessible family planning services.

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.

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.

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. 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 from the panelists Dr. Oboko and Dr. Wausi for it has enhanced my view of research.

References

1. G. K. Gupta (2014). Introduction to Data Mining with Case studies, third edition. PHI Learning Private Limited, Delhi.
2. R.C. de Amorim, C. Hennig (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. H. C. Koh and G. Tan (2005), "Data mining applications in healthcare," Journal of Healthcare Information Management, vol. 19, no. 2, pp. 64–72.
4. S. Nittel, K. T. Leung, and A. Braverman (2003), "Scaling clustering algorithms for massive data sets using data stream," in Proceedings of the 19th International Conference on Data Engineering, U. Dayal, K. Ramamritham, and T. M. Vijayaraman, Eds., IEEE Computer Society, Bangalore, India.
5. Peter J. Rousseeuw (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, Galit, R. Patel, and Peter C. Bruce (2010). Data Mining for Business Intelligence. 2nd edition. New Jersey: Wiley.
7. P. Wasiewicz, Z. Kulaga, 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. A. Rehnman (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. J.M. Nzioki, R.O. Onyango, J.H. Ombaka (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. C. Shinsugi, M. Matsumura, M. Karama, J. Tanaka, M. Changoma, S.Kaneko (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. S. S. Anand, 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 He (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