Comparative clustering and visualization of socioeconomic and health indicators: A case of 47 counties in Kenya.

Evans Kiptoo Korir*
University of Szeged, Hungary.
evanskorir6@gmail.com
April 11, 2024

Abstract

In this study, we used principal component analysis (PCA) to reduce the dimensionality of the data and used a hierarchical and K-means clustering technique to stratify counties in Kenya into five clusters. The grouped counties were then projected onto a geographic map to understand the relationship between their location and socioeconomic and health indicators. The results obtained may be useful to the county and state governments in future plans to promote inclusive and sustainable economic development.

Keywords— Cluster analysis, socioeconomic and health indicators, counties, Dimensionality reduction, Principal component analysis, Hierarchical and K-means clustering.

1 Introduction

Prior to the enactment of the Kenyan constitution in 2010, the country was structured into 8 provinces [24]. The provincial setup was replaced by a decentralized governance system overseen by the Ministry of Decentralization, which designated 47 counties as administrative units. The delineation, size, and geographical location of these counties are based on the 47 legal regions outlined in the 2010 Constitution [20, 22]. The complete list of the 47 counties is displayed along the horizontal axis in Table 4a. As a result of the establishment of this county system, various crucial responsibilities of the national government, including healthcare, education, agriculture, and transportation, were delegated to the counties [20, 22, 24]. Each county is led by an elected governor chosen by the county's residents and is supported by an executive committee and members of the county assemblies.

*Your affiliation and email address here

In the past ten years, significant economic reforms have been implemented by counties, leading to continuous economic growth, enhanced healthcare, and increased school enrollment rates. Nevertheless, development progress has been hindered by various challenges like corruption, poverty, inequality, lack of transparency and accountability, interference from the national government, climate change, and low investment appeal [1, 21]. These obstacles have resulted in socioeconomic and health gaps among counties, prompting migration between them. One notable distinction lies in the population of each county. The national population stood at 47.56 million in 2019, exhibiting a growth rate of 2.3 percent [20]. By comparison, in 2009, the population was 38.6 million, marking a fivefold increase since the country's independence in 1963. The demographic composition is predominantly youthful, with individuals aged 15 and below constituting 39 percent of the population, a decrease from 43 percent in 2009 [36]. Employment is observed in 57 percent of the population (aged 15 to 64), with 29 percent representing the youth demographic. The elderly population (aged 60 and above) accounts for 6 percent of the total population [20]. Nairobi County boasts the highest population percentage at 9.24 percent, while Isiolo has the lowest at a mere 0.56 percent [6, 12, 13, 27, 36].

As per the findings of the Kenya Economic Report 2020, a mere 15.0 percent of the 47 counties exhibit notable manufacturing activities, with the majority of counties relying heavily on agriculture as their primary economic activity. Agriculture emerges as the predominant economic sector in most areas, with the service sector following closely behind [28]. Counties with more established manufacturing and agricultural sectors tend to have larger populations as well [22]. Disparities in poverty rates among counties are evident, with Nairobi County reporting the lowest rate at 16.7 percent, while Turkana County records the highest rate at 79 percent. The report highlights that 22 counties still lag below the national poverty level of 36.1 percent, whereas the poverty rate in 16 counties falls below the national core poverty rate of 8.6 percent [17, 20, 27].

The government has intensified its efforts to allocate more resources to address poverty and inequality in counties through the budget [8, 29]. Turkana and Mandera, which are among the least developed counties, are receiving a relatively larger portion of the revenue allocation. This allocation is determined by the poverty factor in the Commission's Revenue Allocation Formula (CRA), which constitutes 18 percent of the revenue allocation. Nevertheless, there is a lack of clarity regarding the additional indicators that should be incorporated into the CRA formula to gain a better understanding of the counties in need of special attention [8, 29].

Recent research has highlighted a strong connection between socioeconomic elements and health indicators. Cluster analysis, a method that involves grouping multivariate data points into clusters, has been employed to reveal hidden patterns related to regional economic advancement [4, 18, 31, 37]. By comparing and describing different counties, it becomes feasible to investigate the theory that disparities in socioeconomic status and health play a role in the spatial variations seen across regions. The incorporation of cluster techniques into broader regional development strategies offers an opportunity to boost the competitiveness of counties.

In this study, we utilized principal component analysis to reduce dimensions and an agglomerative hierarchical method to depict clusters through a dendrogram at different levels of granularity. We employed 24 socioeconomic and health metrics to represent the variations across 47 counties. To improve the precision of the clustering results, we standardized the data and validated the clusters derived from the hierarchical approach with K-means clustering.

2 Literature Review

Before this investigation, to our knowledge, no previous research has been carried out on the task of categorizing all regions in Kenya into specific groups based on socioeconomic and health indicators. A methodology was developed and explored to group and visualize the child's health status in Kenya in a study conducted by [25]. The research introduced a novel model to cluster counties in Kenya using UNICEF indicators of child health. The clustering technique utilized in this study was the k-means clustering algorithm. To validate the results obtained from k-means, both hierarchical and nonhierarchical clustering algorithms were also employed. The determination of the optimal number of clusters was based on a heuristic approach that integrated a statistical measure of the fit of the cluster. By applying this clustering methodology to data from the literature, the 47 counties of Kenya were grouped into three distinct clusters, comprising 12, 8, and 27 observations, respectively. These clusters were identified as well-off, most marginalized, and moderately marginalized counties based on their child health indicators.

A different investigation, conducted in Kenya, aimed to explore the concept and circumstances of marginalization [7]. The study included a national survey carried out in June 2012 in the 47 counties to identify counties experiencing marginalization. A sample of 3,707 participants representative of the population from various sectors in all counties was selected, who completed individual and group questionnaires. Marginalization was assessed based on the degree of deprivation experienced by the population. The research suggested using multiple indicators of deprivation to identify the locations, causes, and mechanisms of marginalization, and presented strategies to address it. The results revealed that Turkana, Marsabit, Mandera, Lamu, Wajir, Isiolo, Samburu, Tana River, West Pokot, and Garissa are among the most marginalized counties in Kenya. Furthermore, each of the 47 counties contains at least one marginalized area, community, or group.

In the study conducted by [23], a total of 666 instances of confirmed anthrax outbreaks in livestock from 1957 to 2017 were examined. The main objective was to explore how the disease spread across different time periods and geographical locations, aiming to identify and describe specific areas with a high frequency of anthrax cases. The results revealed that particular administrative subcounties displayed clustering in terms of spatial distribution, with roughly 12% of these subcounties contributing to over 30% of all anthrax occurrences. Notably, about 36% of the subcounties did not report any anthrax incidents during the entire 60-year period. Except for Turkana County, situated in the far northwest of the country, which did not document any anthrax incidents. Eight counties (constituting 17.0%) reported

more than 20 events each, with most of them situated in the western and southern parts of the country, including Kiambu, Meru, Narok, Nyeri, Nakuru County, and Bomet County, respectively.

However, some studies have grouped a limited number of regions in Kenya according to different criteria. For example, [35] used the Kulldorff spatial scan Poisson model to detect clusters with a high number of individuals aged 15 to 64 years infected with HIV. Researchers classified people living with HIV (PLHIV) into higher prevalence or lower prevalence (HP / LP) groups and then examined the distributions of sociodemographic and biobehavioral factors related to HIV risk and their connections with the clustering. They identified clusters in various regions, the largest in the Nyanza region and several in Nairobi, as well as in the Central-Rift Valley, Central, and Coast regions.

Numerous research projects have aimed to categorize areas according to socioeconomic and health criteria in different countries. In their work, Merzouki et al. (2021) used combined survey and demographic information to categorize 29 countries in Sub-Saharan Africa into three clusters using 48 indicators related to socioeconomic status and HIV. Principal Component Analysis (PCA) was used to determine the key variables that account for the most significant variance. The findings suggest that the implementation of strategies that integrate social and behavioral aspects could be beneficial in reducing HIV transmission rates in Sub-Saharan African nations.

In a different investigation, [15] used a standardized contact matrix to analyze 32 African nations. They merged data from the World Bank country website, which encompassed various indicators such as social, economic, environmental, institutional, governance, health, well-being, education, gender inequality, and other factors related to development to categorize countries. Through principal component analysis, they visualized the socioeconomic similarities between countries and pinpointed the indicators with the highest variability. To mitigate the issue of high dimensionality, they employed the $(2D)^2PCA$ method to condense the synthetic contact matrices for each country. Subsequently, hierarchical agglomerative clustering was applied to identify clusters of countries that exhibit similar social patterns, considering their socioeconomic performance. Analysis revealed the formation of four distinct clusters, each characterized by unique characteristics. The researchers concluded that social contacts differed between clusters but showed similarities within each group, suggesting that a country's socioeconomic status played a significant role in shaping these groupings.

Socioeconomic and demographic variables were utilized to categorize 146 WHO member nations in order to pinpoint disease-related fatalities and propose strategies for lowering death rates in different regions [4]. Through PCA, the researchers pinpointed fundamental data trends and then employed the hierarchical clustering technique with Ward linkage to cluster countries with comparable mortality causes. The research revealed a connection between a country's income, expenditure, education, and causes of death. The disparities between clusters were shown to be statistically meaningful.

A different research [32] assessed 180 countries regarding their COVID-19 cases and fatalities to assess their readiness and management of the pandemic. By analyzing the daily COVID-19 cases of these countries, the study employed hierarchical clustering to categorize them

into five distinct groups. The clustering revealed that highly developed nations tended to be grouped together, whereas the least developed countries, notably those in Africa, were clustered separately. These findings indicate a strong correlation between a country's economic status, healthcare system, and the number of deaths caused by the pandemic.

To achieve efficient clustering and a structured dendrogram, different linkage methods such as average, complete, Ward, and single linkages are utilized. [19] examined these methods in the context of identifying the factors influencing poverty in North Sulawesi. Through the utilization of agglomerative hierarchical clustering with various linkage approaches and the evaluation of the root mean square standard deviation (RMSSTD) value, it was found that the Ward linkage method yielded the smallest RMSSTD values compared to other methods, indicating its suitability for a thorough analysis.

3 Methodology

3.1 Data

To group counties in Kenya with similar characteristics into a reduced number of clusters, socioeconomic and health data for all counties were obtained from various sources, including reports of the National Council for Population and Development, Kenya Economic Growth 2020, Ministry of Decentralization, Individual County Website, Kenya, the census report collected in 2019, Knoema data website, and the Kenya Property Developers Association (KPDA) website. The data collected encompass residential structure, demographic, health, social, and economic details, providing a comprehensive view of different aspects of the districts. A total of 24 original and derived metrics were chosen for the clustering analysis, as described in Table 1.

The original data underwent processing to standardize specific characteristics, as factors such as county GDP, population and density, household size, infant mortality rates, and crime index display considerable disparities and are essential for optimal model learning. The standard scalar technique from scikit-learn, which normalizes attributes by eliminating the mean and adjusting to a unit variance [31], was utilized to standardize the variables.

3.2 Research design

Various research procedures were carried out in this investigation. The research framework of the study is described in Figure 1.

Statistical analysis, PCA, hierarchical, and K-means clustering were conducted within the Python object-oriented programming framework, utilizing both built-in libraries and open-source packages. The visualization of cluster maps was achieved through the Folium python library. Scikit Learn was employed for the development and execution of PCA and clustering algorithms. The code is openly accessible on GitHub.

Socioeconomic and Health Indicators		
Fertility measures	**Mortality measures**	**Social measures**
contraceptive prevalence	Infant Mortality rates	Employment rate
Fertility rates	Under-five mortality rates	Crime index
Birth rate	Death rates	Poverty rates
Household size	Healthcare facility delivery	Unemployment rate
Education level measures	**Population and Development**	**Access to social services**
HIV prevalence rates	Population size	Urbanization rates
Education level	Population density	Electricity access
Literacy rates	Gross Domestic Product	land size
Child marriage rates	Growth rates	Healthcare facility Density

Table 1: The 24 socioeconomic and health indicators categorized into 6 measures that forms the basis for clustering counties.

3.3 Principal Component Analysis

In contemporary statistical applications, dealing with large data sets that contain a larger number of features than data points is common. To mitigate the challenge of dimensionality that can affect the performance of learning models and training when handling high-dimensional data, techniques such as feature selection and reduction in dimensionality, such as principal component analysis (PCA), are frequently employed [9, 15, 16, 34]. PCA is used for various purposes, including identifying relationships between variables, detecting outliers, recognizing patterns, and reducing the dimensionality of the data. In this study, we use PCA to reduce dimensionality, since we want to map our raw data ($X = x_1, x_2, ..., x_n$) from a high-dimensional space $\mathbb{R}^m$ to a lower-dimensional space $\mathbb{R}^d$ to capture maximum variation, where n denotes the total number of our observations [5, 18, 19]. The PCA space comprises principal components d that are orthogonal, uncorrelated, and indicate the directions of maximum variance. These principal components can be easily calculated using either the covariance method or the singular value decomposition (SVD) [16, 26, 34]

PCA was conducted on the 24 socioeconomic and health datasets using the Singular Value Decomposition (SVD) method. The data was then projected and converted from a 24-dimensional space ($\mathbb{R}^{24}$) to a four-dimensional feature space ($\mathbb{R}^4$), as illustrated in Fig. 4a. The principal components obtained and the proportion of variance accounted for by each component are presented in Fig. 2b.

The optimal number of principal components, which is 4, was determined by analyzing the graph in Figure 2b. These four components account for 74% of the variance in the data set shown in Figure 2b. The first principal component explains 37% of the variance, followed by the second component with 15%, the third component with 14%, and the fourth component with only 8%. In particular, the latter components contribute minimally to the overall variability.

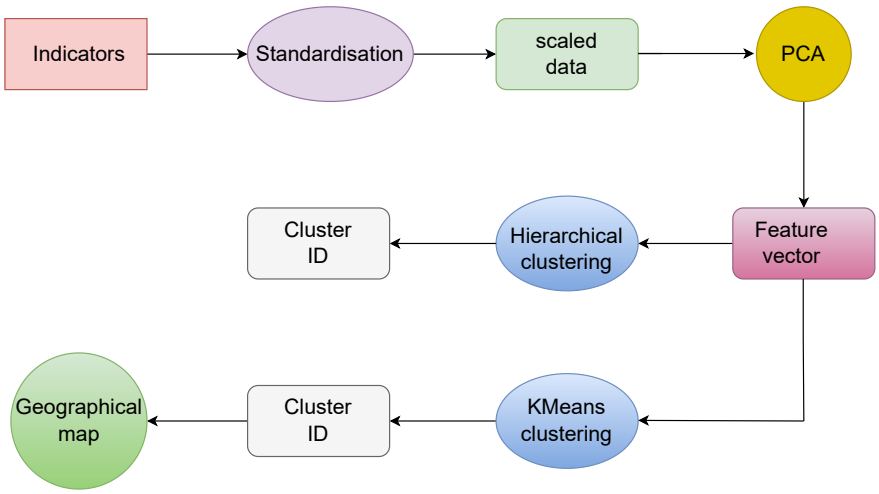
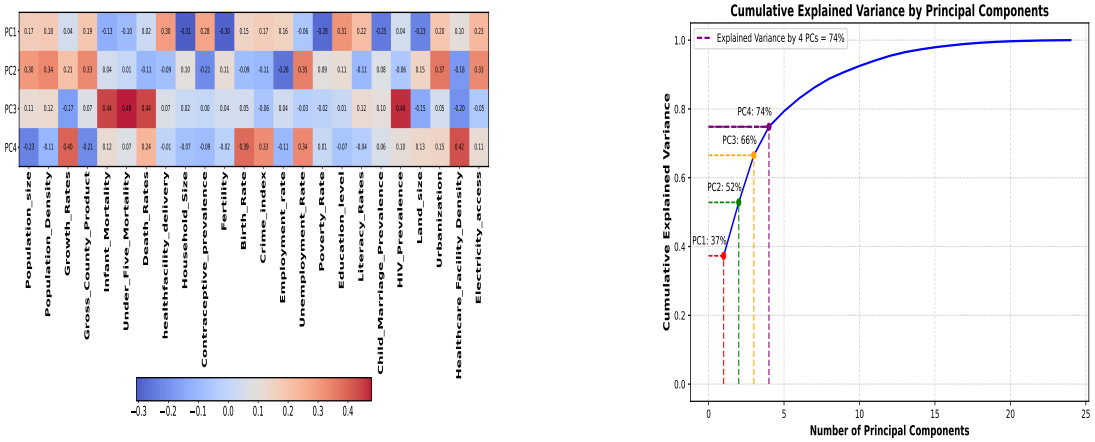


Figure 1: Flowchart summarizing the research design. Rectangles show the output of a module, represented by ellipsoids that contain function calls that execute a specific part of the method. As can be seen, the socioeconomic and health indicators are the input of the study. We applied standardization to the indicator data to obtain the scaled data. In the next step, we apply PCA to the scaled data. The PCA output feature vectors are used in the hierarchical and K-means clustering algorithm to obtain the clusters. The geographical map visualizes the clusters in a geographical context.

3.4 Clustering

The categorization of extensive data collections has become increasingly crucial in a variety of fields, including health, medicine, social sciences, and spatial analysis [19]. This process of grouping data necessitates the use of cluster analysis, like clustering, which is an unsupervised algorithm designed to identify clusters within a dataset by maximizing the similarity within clusters while minimizing the dissimilarity between them. Various approaches exist for clustering, such as feature selection, distance-based methods, partition-based techniques, density-based algorithms, probabilistic methods, grid-based approaches, and others [10, 16, 31, 32].

In this research, distance-based algorithms are utilized due to their wide range of applications, simplicity, and ease of implementation compared to other clustering methods [16]. Initially, the study explored agglomerative clustering, where each observation was initially treated as an individual cluster. The process involves merging the closest cluster pairs at each stage and updating the dissimilarity matrix accordingly [19, 31, 32]. This agglomerative merging process continues until the final maximum cluster is achieved. Clusters are combined based on decreasing similarity levels until a single cluster is reached that encompasses all data points. This signifies the completion of the dendrogram and the conclusion of the merging process [10]. In this investigation, Ward's criterion is employed to consolidate the individual clusters using a distance measure that aims to minimize the sum of squared errors (SSE). For any pair of clusters, C_1 and C_2, the linkage is calculated by assessing the increase in the



(a) (b)

Figure 2: Figure (a) corresponds to the heat map of factor loading for the first four principal components. The colors represent the correlation coefficient between the original feature and the principal component. Their values are provided by the horizontal bar on the lower side. The cumulative variance plot (b) shows the total percentage of variance explained versus the number of principal components.

clustering SSE resulting from their merger, $C_1 \cup C_2$. Ward's criterion can be expressed as:

$$W(C_{1\cup 2}, c_{1\cup 2}) - W(C, c) = \frac{N_1 N_2}{N_1 + N_2} \sum_{v=1}^{M} (C_{1v} - C_{2v})^2$$

$$= \frac{N_1 N_2}{N_1 + N_2} d(C_1 - C_2)$$

where d represents the selected distance metric. For this particular investigation, the Euclidean distance metric is utilized. Subsequently, determining the number of clusters becomes straightforward after constructing the dendrogram. As a general guideline, the vertical distance that shows the greatest separation between two groups is typically considered for the determination of the group [10, 11].

In this study, k-Means is employed to validate the outcomes achieved through hierarchical clustering. In this method, the mean of each cluster serves as the representative for partitioning. The calculation of distances is based on the Euclidean distance metric. Alternatively, other metrics such as the Manhattan distance and the cosine similarity can also be utilized. Let $D = \{x_1, x_2, ..., x_N\}$ represent a data set comprising N points, and let $C = \{C_1, C_2, ..., C_k..., C_K\}$ denote the resulting clustering after applying k-Means. The goal is to identify a clustering configuration that minimizes the Sum of Squared Errors score (SSE) [10].

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|2$$

$$c_k = \frac{\sum_{x_i \in C_k} x_i}{\mid c_k \mid}$$

where c_k is the centroid of the cluster C_k.

4 Results

We analyzed the unprocessed data and performed PCA to capture the variability in our dataset, which consisted of 24 socioeconomic and health metrics. The original factors were converted to a new set of variables to account for the variance in the data. By reducing the complexity of our data set to a lower dimension, we could visually explore, analyze, and identify patterns in regions with comparable socioeconomic and health characteristics using the Hierarchical Unsupervised Clustering technique. We used a K-means clustering algorithm to compare the hierarchical results which gave comparable results.

4.1 Visualization of socioeconomic and health characteristics

The socioeconomic and health variables shown in Figure 3 were classified into six larger measures, as described in Table 1. The values of these variables were then visualized using a heat map, as shown in Fig. 4a, which facilitates easy interpretation. The variables are displayed in the same order as in Table 1 and as shown in Figure 4b, with the values for each county for each variable shown as small triangles. In Fig.4a, when examining access to social services, urbanization is indicated on the top, access to electricity on the right, land size at the bottom, and healthcare facility density on the left. The magnitude of the values is reflected in the intensity of the color. For example, Nairobi and Mombasa exhibit higher levels of urbanization and access to electricity (pictured in green and red, respectively). On the contrary, Narok and Nyandarua show low urbanization rates, while Mandera and the Tana River have limited access to electricity (lightest shade). Similarly, Lamu demonstrates the highest density of health facilities (purple shade), while Marsabit and Turkana boast the largest land area (blue shade in the lower triangle).

The data presented in Figure 4a illustrate various indicators related to population and development, which are categorized into population size (top), population density (right), GDP (bottom), and growth rates (left), as shown in Figure 4b. Among the major cities, Nairobi and Mombasa exhibit the highest population density, while Nairobi and Kiambu have the largest population size, with Nairobi also showing a substantial GDP. On the other hand, marginalized counties such as Marsabit and Isiolo experience high population growth rates.

HIV prevalence rates, educational attainment, literacy rates, and child marriage rates presented in Table 1 were consolidated into indicators of educational achievement, as shown in Figures 4a and 4b. According to the illustration, Siaya, Kisumu, Homabay and Migori exhibit the highest HIV prevalence rates. In particular, Wajir, Isiolo, and Samburu counties demonstrate elevated marriage rates.

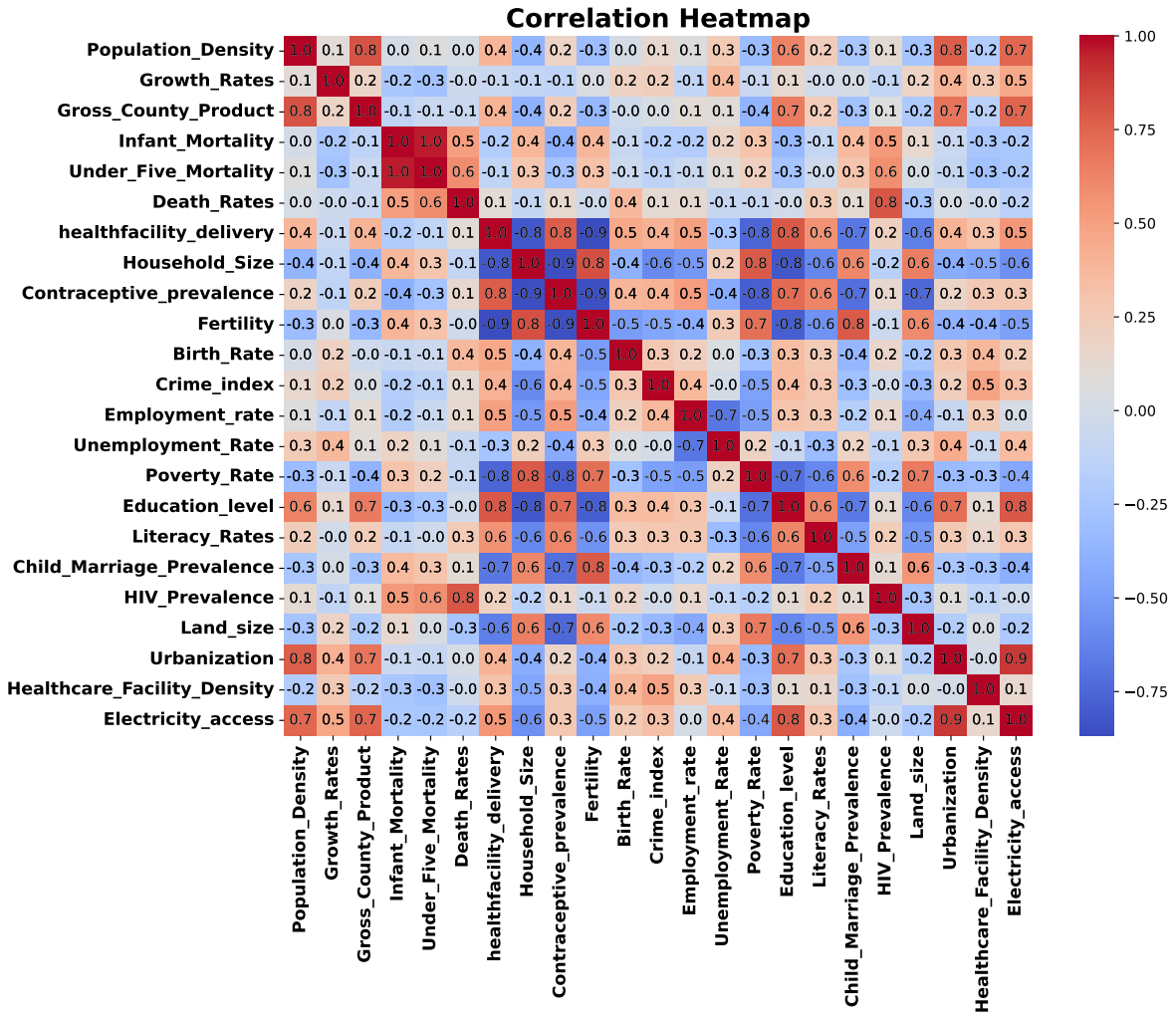
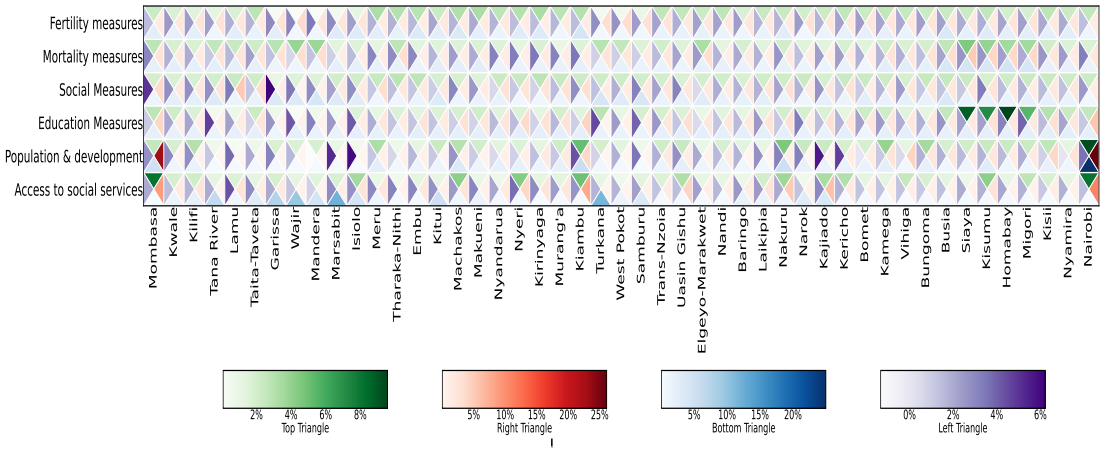


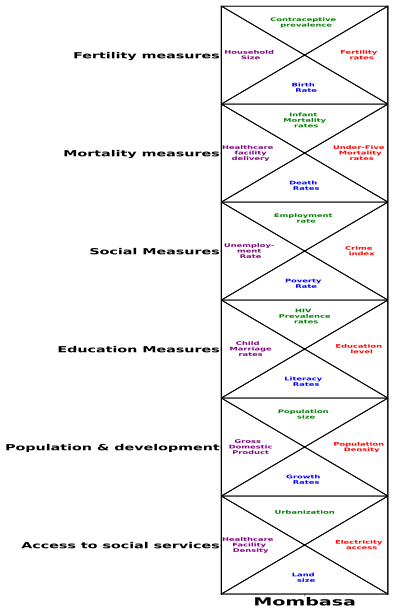
Figure 3: Heatmap illustrating the pairwise correlations among socioeconomic and health indicators. The color intensity indicates the strength and direction of correlation, ranging from deep blue for a perfect negative correlation to deep red for a perfect positive correlation. Annotations display correlation coefficients rounded to one decimal place, providing insights into the relationships between variables.

In this manuscript, social metrics include the employment rate, the crime index, the poverty rate and the unemployment rate as specified in Table 1. Kilifi, Tana River, Isiolo, Machakos, Makueni, Kajiado and Nairobi counties exhibit elevated levels of unemployment. On the contrary, the Lamu, Mombasa and Taita Taveta counties demonstrate the highest incidences of crime.

The article focuses on mortality indicators such as infant mortality rates, under-five mortality rates, death rates, and the availability of health facilities. Siaya, Kisumu, Homabay, and Migori have documented the highest mortality rates, partly attributed to an elevated prevalence of HIV. Similarly, other less developed counties, such as Garissa, Wajir, Man-



(a)



(b)

Figure 4: Socioeconomic and health factors are grouped into six measures and displayed vertically for all 47 counties (figure (a)). These variables are listed in the same sequence as shown in Table 1 and Figure 4b. For example, in access to social services, urbanization is at the top, access to electricity is on the right, land size is at the bottom, and health facility density is on the left. Values represent the percentage contribution of each county to the respective variable. The intensity of the color in the horizontal bars reflects the magnitude, with lighter shades indicating lower percentages and darker shades representing higher values in the counties. Figure (b) has been annotated with the color maps of figure (a), and Mombasa county has been utilized as an illustration.

dera, and Marsabit, exhibit similar patterns. In contrast, Mombasa, Meru, Tharaka-Nithi, Embu, Siaya, Kisumu and Kisii counties demonstrate better provision of health facilities, as indicated by the purple color in Figure 4a.

Contraception usage, fertility rate, birth rate, and household size are the factors that influence fertility rates (see figure 4b). Regions bordering Lake Victoria, such as Siaya, Kisumu, Homabay, and Migori, exhibit the highest birth rates. Meanwhile, Kwale, Garissa, Wajir, Marsabit, Turkana, and West Pokot counties are characterized by larger household sizes, as indicated by the purple color in the data.

The six-group measures considered exhibit strong positive correlations within the same group and weak correlations between different groups. For example, within mortality measures, infant mortality, under-five mortality, and death rates show high correlations (small squares). The relationships between these indicators and variables in other groups are generally weak, with the exception of the HIV prevalence rate, which typically reflects mortality but is treated as an indicator of education in this study. Similar correlation patterns are observed across the various measures in Figure 3.

4.2 County clustering and examination of related socioeconomic and health factors

The hierarchical arrangement of the 47 countries produced a dendrogram as depicted in Figure 5. An advantage of this clustering method is the ability to manually select the hierarchy level to cut and extract the clusters without the need to rerun the algorithm. Using a general guideline, clusters can be distinguished by setting a threshold. The selection of this threshold is entirely at the discretion of the user. By segmenting the dendrogram at a cluster distance of 10, five distinct clusters were established, as illustrated in Figure 5 and Table 2.

The first group of counties (shown in blue) is characterized by less productive economic sectors, leading to a minimal contribution to the national GDP. This group consists of counties located in the northern region of Kenya (see Figure 7), known for its arid conditions with temperatures reaching up to 35 degrees Celsius in certain areas, which are not conducive to substantial economic endeavors such as agriculture. The inhabitants of these counties are primarily engaged in nomadic pastoralism. Cluster 2, shown in green, includes three important developed regions in Kenya; Nairobi, Mombasa, and Kiambu. Nairobi, as the capital of the nation, has the largest population, makes the largest contribution to GDP, and is more industrialized (Figure 4a). Mombasa, the second most populous city after Nairobi, is particularly suited for international trade due to its primary port. This port not only serves the Kenyan inland regions, but also serves markets in neighboring countries such as Uganda. It ranks as the second largest port in sub-Saharan Africa in terms of tonnage and containers handled, after Durban. Kiambu, located adjacent to Nairobi, is undergoing rapid expansion as a result of its significant population and flourishing commercial operations. This region shows elevated levels of progress in contrast to others, and Nairobi serves as a notable example.

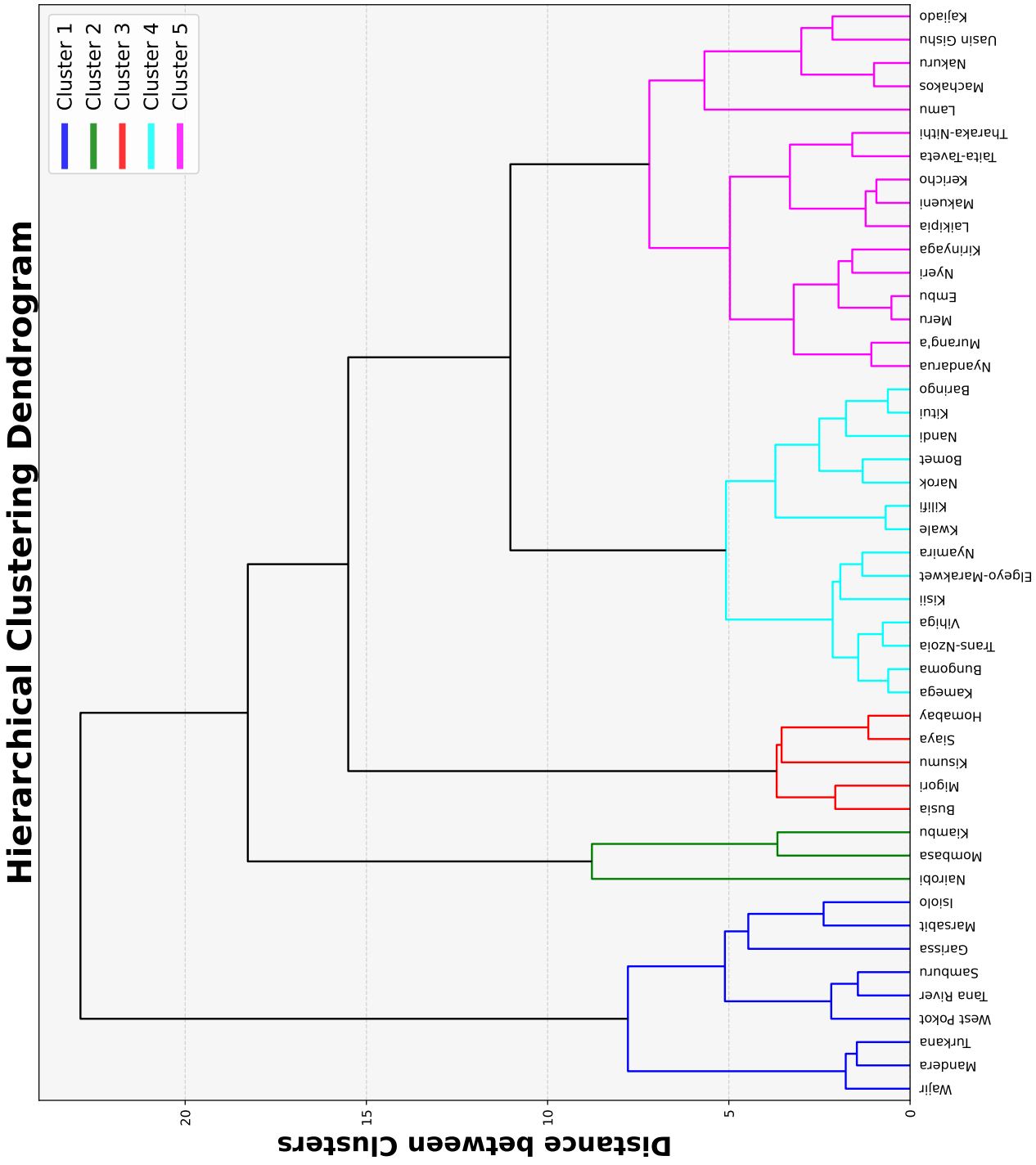


Figure 5: Dendrogram related to the grouping of counties based on the socioeconomic and health indicators. The vertical axis indicates the distance between the two groups that are joined. The agglomerative nature of the algorithm can be seen by following the tree from the bottom up. Counties with the same hue belong to the same cluster.

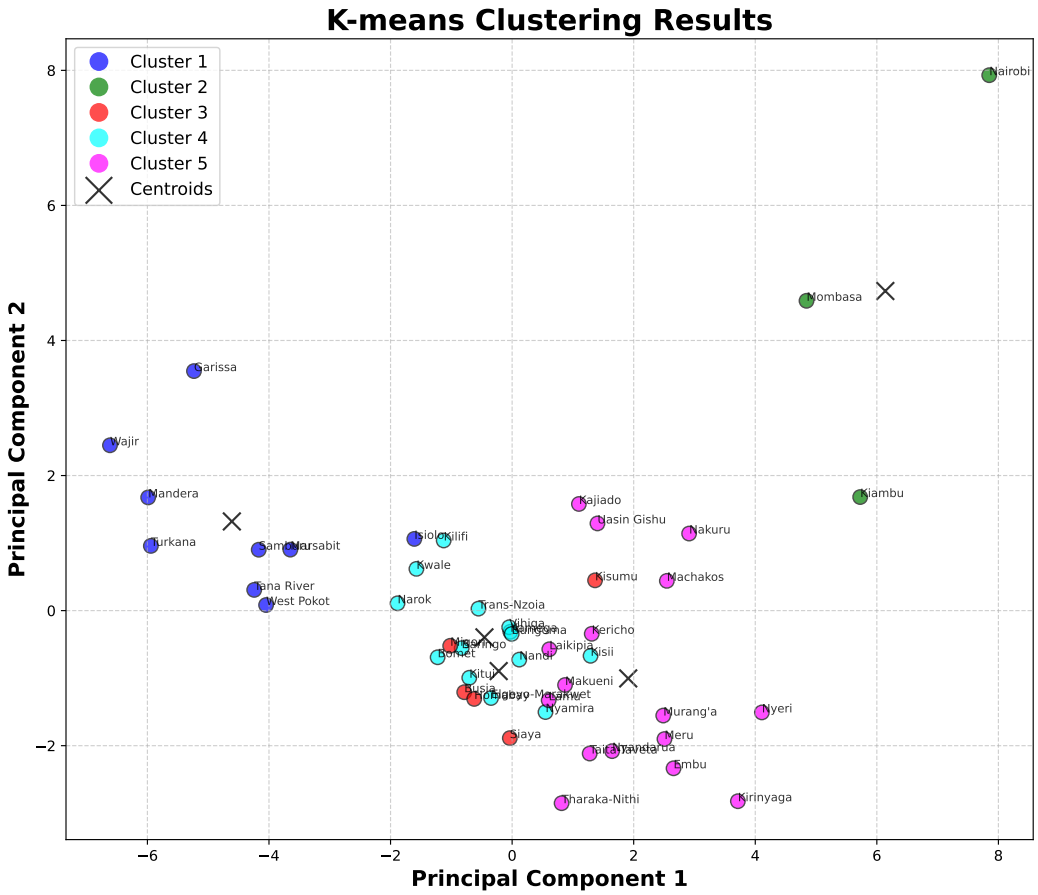


Figure 6: Grouping of counties based on the socioeconomic and health indicators using K-means algorithm. Centroids refer to the center points of each cluster. The output graph clearly shows the five different clusters with different colors.

The third group, colored red, represents counties in Nyanza province, where fishing is the main economic activity. These counties are characterized by high rates of HIV prevalence and mortality. Both clustering algorithms generated identical clusters for this specific group. The fourth group of cyans consisted of counties of various regions, excluding the northern part, as shown in Figure 7, with socioeconomic and health indicators comparable to those of the second and fifth groups (Figure 4a). Most counties in the western region are represented in this group. Cluster 5, represented in purple, encompasses counties that are geographically spread throughout the country. Although these counties are not highly developed, they exhibit stronger economic performance compared to the counties in groups 1, 3, and 4. They are characterized by having larger urban centers, moderate levels of development and average socioeconomic indicators, as agriculture, fishing, and commerce are the main

economic activities in these areas. The conclusive cluster lists generated by hierarchical and K-means clustering methods can be found in Table 2.

CLUSTER	COUNTY NAMES	SIZE	%	CLASS
1	Wajir, Mandera, Turkana, West Pokot, Tana River, Samburu, Garissa, Marsabit, Isiolo.	9	19%	Marginalized.
2	Nairobi, Mombasa, Kiambu.	3	6%	Prosperous.
3	Kisumu, Siaya, Homabay, Busia, Migori.	5	11%	Average.
4	Vihiga, Kakamega, Bungoma, Trans-Nzoia, Elgeiyo-marakwet, Kwale, Kilifi, Narok, Bomet, Nandi, Kitui, Baringo.	12	26%	Stable.
5	Meru, Embu, Nyeri, Kirinyaga, Laikipia, Kericho, Taita-Taveta, Makueni, Nyandarua, Muranga, Tharaka-Nithi, Kisii, Nyamira, Lamu, Machakos, Nakuru, Uasin Gishu, Kajiado.	18	38%	Well-off.

Table 2: Clusters were generated for the 47 counties in Kenya by utilizing socioeconomic and health indicators through the application of hierarchical clustering method and K-means algorithm.

5 Discussion.

This paper discusses the socioeconomic indicators that characterize a group of counties, along with the key factors that contribute to the economic success of a county. Child mortality, infant mortality, and HIV prevalence rates are closely related, as they serve as markers for mortality rates and life expectancy as shown in Figure 3. Counties with high mortality rates tend to form one group. It is clear that these counties have limited contraceptive use, a moderate incidence of early marriages, and a relatively low poverty rate in the figure 4a. The GDP of a county is a crucial metric for its development, influenced by factors such as population size, population density, access to electricity for industrial advancement, education, and urbanization, which all contribute to providing residents with essential and affordable services. GDP growth is affected by high birth and mortality rates, as evidenced in this study [8, 29]. Counties with low GDP levels typically exhibit high mortality rates, fertility rates, household sizes, and the prevalence of early marriages (see Figures 3 and 4a). In this investigation, the algorithm grouped the counties into five clusters - marginalized, prosperous, average, stable, and well-off - according to their socioeconomic indicators as in Table 2.

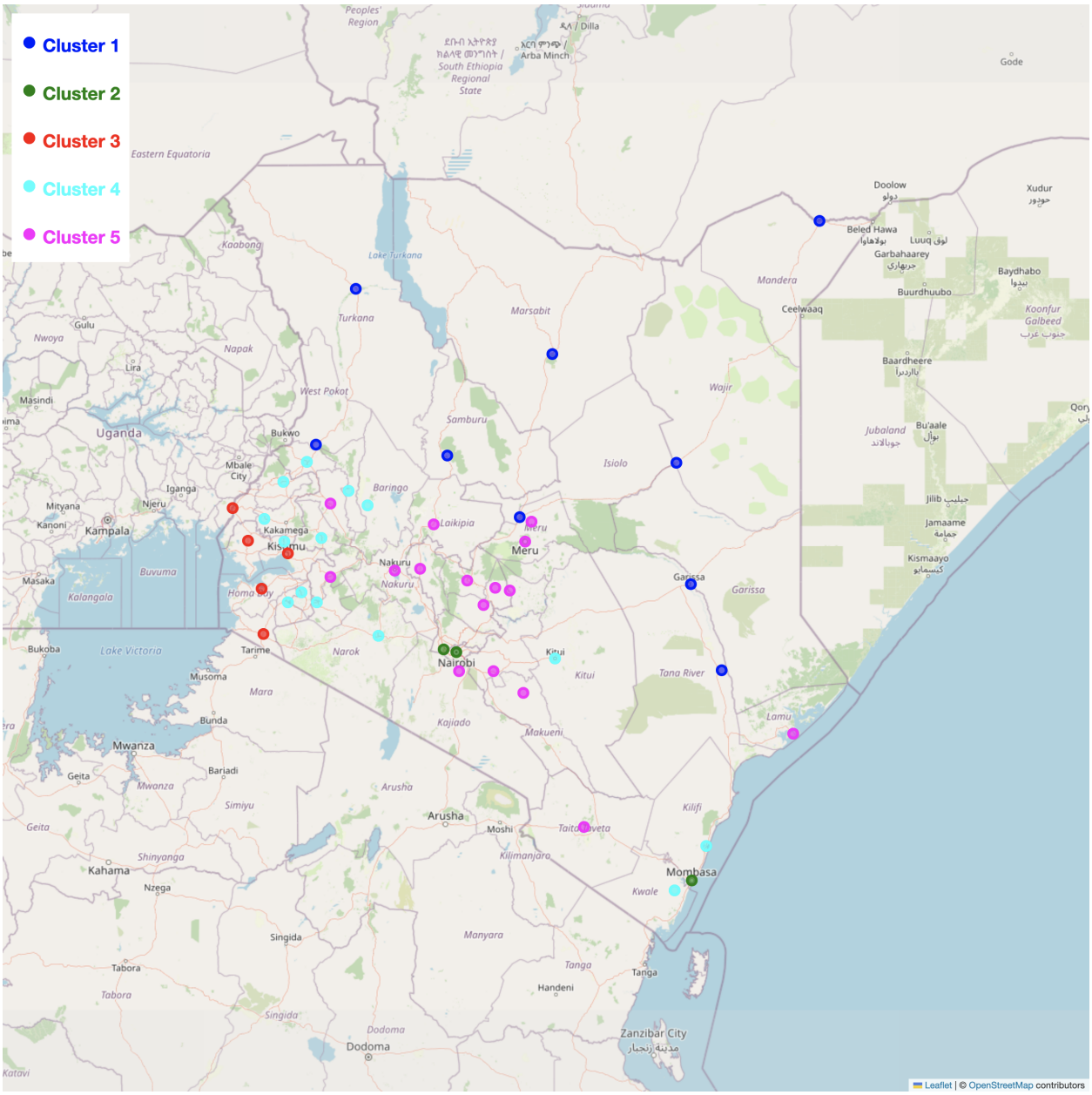


Figure 7: Visualizing the county's socioeconomic and health indicators and their geographic context using Folium library in python. The 47 Kenyan counties were divided into five clusters based on the hierarchical and K-Means algorithm.

Cluster 1 comprises 19% of the counties, as illustrated in Table 2. These counties are located in the northern region of Kenya (see Figure 7). In particular, they are located in arid and semi-arid zones where livestock farming is the main economic activity. It is significant to mention that these counties exhibit a low level of GDP of approximately 1%. Given that agriculture is the main industry in Kenya, these regions cannot sustain most agricultural practices, leading to a low population density. Consequently, these counties exhibit reduced

economic productivity, with widespread food poverty prevailing, averaging over 60 percent in these areas. Consequently, these counties also experience elevated poverty rates compared to others, with high poverty levels observed in Mandera, Wajir, and Garissa. Due to a strong positive correlation between household size and poverty rate, the counties in this group have larger households, as shown in Figure 4a. In the study by [25], these counties were clustered as a group according to child health, education, maternal health, water and sanitation and classified as the most marginalized. Similarly, the findings of [7] indicate that these counties are the most marginalized in Kenya.

The second cluster comprises counties known as prosperous because of their high level of development. This cluster is made up of only three counties: Nairobi, Mombasa, and Kiambu (see Figure 5). These counties have significant cities and towns and can be characterized as developed areas. Mombasa, the second largest city in Kenya after Nairobi (the capital), has a unique advantage due to its strategic location on the eastern coast of the Indian Ocean, making it a popular beach and tourist destination (see Figure 7). On the other hand, Kiambu, which shares a border with Nairobi County, has experienced substantial benefits from urbanization and expansion of the capital city. These counties are among the most urbanized, with Nairobi and Mombasa having an urbanization rate of 100%, followed by Kiambu in the third place. They serve as hubs for numerous Kenyan companies and are well integrated into the national electricity grid. Furthermore, these counties exhibit well-established industries and a thriving manufacturing sector, leading to the creation of numerous job opportunities at both the national and county levels, consequently boosting the GDP of the respective regions. Nairobi alone contributes a quarter of total GDP, with Kiambu and Mombasa following as the second and third largest contributors, respectively [8, 29]. These counties were clustered as "well-off" among the 26% counties in [25].

Cluster 3 includes the counties located in Nyanza province in southwestern Kenya near Lake Victoria, as shown in Figure 7. This area is mainly inhabited by the ethnic groups Luo and Kisii, who are Bantu speakers. Due to its proximity to the lake, the main economic activity in these counties is fishing. However, the potential of lake water remains underutilized, leading to food shortages due to limited commercial irrigation and agricultural practices in the region, and thus the cluster has been labeled "average." The communities within these counties have rich social and cultural customs. The prevalence of HIV / AIDS is significant in this area, reflecting the larger HIV burden in Kenya, where an estimated 1.6 million people lived with HIV in 2013 [12]. Women are disproportionately affected by the disease. The HIV epidemic is particularly concentrated in Homabay, Siaya, Kisumu, and Migori counties, where prevalence rates are notably high. These high rates of HIV are closely related to increased mortality, especially among children and infants. Homabay County has the highest mortality rates, followed by Siaya. Infant mortality rates exceed 100 per 1,000 live births in all counties within this cluster, with Siaya County recording the highest rate at 142. Furthermore, the mortality rate among children under five years of age is notably high, and Siaya County reported the highest rate, as shown in Figure 4a. This region was identified as having the highest HIV prevalence in Kenya and was classified by the authors of [35], while [25] classified it as moderately marginalized among 57% of the counties.

Cluster 4 comprises moderately developed and stable counties located in the Western, Coast, and Rift Valley provinces, as shown in Figure 7. These counties do not have major urban centers and primarily focus on agriculture and trade. They have a limited number of major towns and are looking for markets for their agricultural goods in neighboring counties within clusters 2 and 5. The lack of significant manufacturing industries in the area still constrains their GDP growth. However, these counties outperform those of cluster 1 in terms of socioeconomic and health indicators. In terms of health indicators, they perform better than cluster 3, as depicted in Figure 4a. In the study by [25], this group was classified as moderately marginalized.

Cluster 5 represents economically stable counties and is classified as a well-off group. It is the largest group, accounting for 38%, as indicated in Table 2 and Figure 5. These counties are located in various parts of Kenya, except the northern region, which is predominantly represented by the counties in cluster 1 (see Figure 7). They are currently undergoing urbanization, but continue to rely heavily on agriculture as their primary economic sector, in addition to trade and commerce. Although they have notable cities and towns, they are not sufficient to establish a large and sustainable market for agricultural products. However, major cities have attracted a larger population, resulting in inter-county migration that has marginalized at least one area within the county [7]. These regions are clearly still in the development phase towards achieving the standards of cluster 2. They are still lacking substantial manufacturing sectors that could enhance the region's GDP. This also clarifies the limited regional transformation at the regional level. A strong manufacturing, agricultural, and service industry is expected to generate valuable employment opportunities at the regional level for the local population. The group outperforms others in terms of socioeconomic factors compared to Cluster 1 and health indicators compared to Cluster 2. They show high rates of school enrollment, improved healthcare care, reliable access to electricity, increased urbanization rates, and controlled levels of prevalence rates of fertility, mortality, household size, and child marriage based on Figure 4a. However, there are still peripheral areas within this cluster as well as other clusters. This group contributes to the highest percentage of counties grouped as "moderately marginalized" in [25].

Our research has certain limitations. First, there is a gap in the data due to variations in the years of collecting socioeconomic and health indicators. Demographic data, such as the county population, were sourced from 2019, while education levels were from 2018, poverty rates from 2016, and other metrics such as infant mortality, crime index, and gender index from 2009. Consequently, there is a risk that these indicators may not accurately reflect the current circumstances. Second, the use of a limited number of principal components to elucidate the raw features can present challenges when interpreting their significance and assigning labels. Although the first principal component signifies the direction of maximum variance, it may not capture the most relevant component for the study, as it only offers a broad overview of the data. This limitation also extends to the other selected components. Lastly, the choice of socioeconomic indicators was based solely on the accessibility of the data. The four components utilized in the analysis were derived from the existing literature and may be modified in the future. However, the attributes of the study led to robust and reliable results. The socioeconomic and health status presented may be influenced by the implementation of new policies and development initiatives by the new national government.

Consequently, it is advisable to interpret the results cautiously, and further research will be necessary to review the clusters and reflect the evolving socioeconomic landscape of the county.

6 Conclusion

Cluster analysis methods can be advantageous for investigating and explaining socioeconomic and health discrepancies. By uncovering hidden connections between variables that may not be immediately apparent to researchers, clustering can improve understanding of the data set, providing a foundation for future studies. Using easily accessible variables, Principal Component Analysis (PCA) was employed to decrease the complexity of the data, followed by the application of hierarchical clustering methods and K-means to categorize Kenyan counties into five distinct clusters. Then, these grouped counties were visualized on a geographical map to explore the correlation between their geographical location and socioeconomic and health metrics. Cluster analysis revealed notable variations according to indicators. Utilizing a stacked heat map provided valuable information on each county's performance and its associated variables. The findings obtained could be beneficial for county and state authorities in developing strategies to advance inclusive and sustainable economic growth.

7 Ethics approval and consent to participate

Not applicable.

8 Consent for publication

Not applicable.

9 Data Availability

Code and county-level data used for this analysis are available on Github:

`https://github.com/Evanskorir/kenya-counties`

10 Competing interests

The authors declare that they have no competing interests.

Funding

No financial support was received for this study.

11 Acknowledgement

The Stipendium Hungaricum Scholarship Program with Application No. 118250 supported the author.

References

[1] Achoki, Tom, et al. "Health disparities across the counties of Kenya and implications for policy makers, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016." The Lancet Global Health 7.1 (2019): e81-e95.

[2] Anderberg, M. R. (2014). Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks (Vol. 19). Academic press.

[3] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer google schola, 2, 645-678.

[4] Çağlar, M., Gürler, C. (2022). Sustainable Development Goals: A cluster analysis of worldwide countries. Environment, development and sustainability, 24(6), 8593-8624.

[5] Carrillo-Larco, Rodrigo M., and Manuel Castillo-Cara. "Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach." Wellcome open research 5 (2020).

[6] Chae, Sophia, and Thoai Ngo. "The global state of evidence on interventions to prevent child marriage." (2017).

[7] Commission on Revenue Allocation. "Survey report on marginalized areas/counties in Kenya." CRA Work Paper No. 2012/03 (2012).

[8] Danis, Ouma, and Jennifer M. Kilonzo. "Resource allocation planning: Impact on public sector procurement performance in Kenya." International Journal of Business and Social Science 5.7 (2014): 1.

[9] Dunteman, George H. Principal components analysis . Vol. 69. Sage, 1989.

[10] Gan, Guojun, Chaoqun Ma, and Jianhong Wu. Data clustering: theory, algorithms, and applications . Society for Industrial and Applied Mathematics, 2020.

[11] Friedman, Jerome. "The elements of statistical learning: Data mining, inference, and prediction." (No Title) (2009).

[12] Kwena, Zachary A., et al. "HIV prevalence, spatial distribution and risk factors for HIV infection in the Kenyan fishing communities of Lake Victoria." PloS one 14.3 (2019): e0214360.

[13] Lowe, Hattie, et al. "'If she gets married when she is young, she will give birth to many kids': a qualitative study of child marriage practices amongst nomadic pastoralist communities in Kenya." Culture, health sexuality 24.7 (2022): 886-901.

[14] Łuczak, Aleksandra, and Sławomir Kalinowski. "Fuzzy clustering methods to identify the epidemiological situation and its changes in European countries during COVID-19." Entropy 24.1 (2021): 14.

[15] Korir, Evans Kiptoo, and Zsolt Vizi. "Clusters of African countries based on the social contacts and associated socioeconomic indicators relevant to the spread of the epidemic." arXiv preprint arXiv:2303.17332 (2023).

[16] Korir, Evans Kiptoo, and Zsolt Vizi. "Clustering of countries based on the associated social contact patterns in epidemiological modelling." International Symposium on Mathematical and Computational Biology. Cham: Springer Nature Switzerland, 2022.

[17] Macharia, Peter M., Eda Mumo, and Emelda A. Okiro. "Modelling geographical accessibility to urban centres in Kenya in 2019." PLoS One 16.5 (2021): e0251624.

[18] Merzouki, Aziza, et al. "Clusters of sub-Saharan African countries based on sociobehavioural characteristics and associated HIV incidence." PeerJ 9 (2021): e10660.

[19] Mongi, C. E., et al. "Comparison of hierarchical clustering methods (case study: Data on poverty influence in North Sulawesi)." IOP Conference Series: Materials Science and Engineering. Vol. 567. No. 1. IOP Publishing, 2019.

[20] Mose, Naphtali. "Determinants of regional economic growth in Kenya." African Journal of Business Management 15.1 (2021): 1-12.

[21] Muthoka, James M., et al. "Mapping Opuntia stricta in the arid and semi-arid environment of kenya using sentinel-2 imagery and ensemble machine learning classifiers." Remote Sensing 13.8 (2021): 1494.

[22] Mwenda, Albert K. "Economic and administrative implications of the devolution framework established by the constitution of Kenya." (2010).

[23] Nderitu, Leonard M., et al. "Spatial clustering of livestock Anthrax events associated with agro-ecological zones in Kenya, 1957–2017." BMC Infectious Diseases 21 (2021): 1-10.

[24] Ngenoh, Peter K. Challenges of implementing devolution and planning objectives by the ministry of devolution and planning in Kenya . Diss. University of Nairobi, 2014.

[25] Njiru, Nicholas M. Clustering and visualizing the status of child health in Kenya: A data mining approach. Diss. University of Nairobi, 2015.

[26] Nicholson, Charles, et al. "A machine learning and clustering-based approach for county-level COVID-19 analysis." Plos one 17.4 (2022): e0267558.

[27] Nyariki, Dickson M., and Dorothy A. Amwata. "The value of pastoralism in Kenya: Application of total economic value approach." Pastoralism 9.1 (2019): 1-13.

[28] Nyoro, James K. Agriculture and rural growth in Kenya. Tegemeo Institute, 2019.

[29] ONYANGO, Jared Abongo, et al. "Adequacy of the Commission on Revenue Allocation Parameters for Equitable Revenue Sharing with Counties in Kenya." International Journal of Innovative Finance and Economic Research 3.4 (2015): 1628.

[30] Rios, Ricardo A., et al. "Country transition index based on hierarchical clustering to predict next COVID-19 waves." Scientific reports 11.1 (2021): 15271.

[31] Rizvi, Syeda Amna, Muhammad Umair, and Muhammad Aamir Cheema. "Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators." Chaos, Solitons Fractals 151 (2021): 111240.

[32] Sadeghi, Banafsheh, Rex CY Cheung, and Meagan Hanbury. "Using hierarchical clustering analysis to evaluate COVID-19 pandemic preparedness and performance in 180 countries in 2020." BMJ open 11.11 (2021): e049844.

[33] Semaan, Gustavo Silva, et al. "A hybrid heuristic with Hopkins statistic for the automatic clustering problem." IEEE Latin America Transactions 17.01 (2019): 7-17.

[34] Tibshirani, Robert, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63.2 (2001): 411-423.

[35] Waruru, Anthony, et al. "Finding hidden HIV clusters to support geographic-oriented HIV interventions in Kenya." JAIDS Journal of Acquired Immune Deficiency Syndromes 78.2 (2018): 144-154.

[36] Wiesmann, Urs Martin, Boniface Kiteme, and Zachary Mwangi. Socio-economic atlas of Kenya: Depicting the national population census by county and sub-location. Kenya National Bureau of Statistics, Centre for Training and Integrated Research in ASAL Development, Centre for Development and Environment, 2014.

[37] Yakovenko, NV, IV Komov, and RV Ten. "Cluster approach in assessing the level of socio-economic development of the municipal districts (Voronezh region)." International Science and Technology Conference" FarEaston"(ISCFEC 2019) . Atlantis Press, 2019.