**COUNTRIES WITH SIMILAR SOCIO-ECONOMIC CHARACTERISTICS TO KENYA.**

# Chapter 1:Background.

For many decades many organizations and governments of different countries in the world have found the need to collect and store data from various sources regarding their socio-economic characteristics. The data is of great importance for decision-making purposes for governments and organizations. Socio-economic data collected can be used by different countries, entrepreneurs, and donors for comparison purposes in order to foster: cooperation, investments and poverty eradication. This research project aims at analyzing 5 countries that have similar socio-economic characteristics to Kenya based on similarities of their socio-economic data in order to provide information to entrepreneurs, governments and donors.

# Chapter 2:Methodology.

**2.1 Data**

The dataused for this research was obtained from the UC Irvine data repository. It consists of 20 socio-economic variables collected from 221 different countries.

**2.2 Variables in the dataset.**

1. **Country**-The name of the country

2. **Region** -The region where the country is geographically located

3**. Area (sq. mi.)-**The area in square miles of the Country.

4.**Population**-The total number of people living within the boundaries of a particular country.

5.**Pop. Density (per sq. mi.)**-The number of people that live per square mile in a particular country.

6.**Coastline (coast/area ratio)**-The ratio of the area covered by the coastline to the total area covered by a country.

7.**Net migration**-The difference between the number of people who immigrate into a particular country to the people who emigrate from a particular country.

8.**Infant mortality (per 1000 births)**-The number of children who die during childbirth per every 1000 children born every year.

9.**GDP ($ per capita)**-The per capita income of a particular country.

10.**Literacy (%)-**The percentage of the number of people who know how to read and write in the whole population

11**.Phones (per 1000)**-The number of people who own and use phones per 1000 people.

12**.Arable (%)**-The percentage of the total area of land in a particular country that can be used for agricultural activities.

13. **Crops (%)**-The percentage of the total area of a country where growing crops can be practiced.

14. **Other (%)**-The percentage of area used for other activities other than agricultural activities.

15.**Climate**-The climatic conditions of a given country.

16.**Birthrate**-The rate at which people living in a particular country give birth.

17.**Deathrate**-The rate at which people living in a particular country lose their lives.

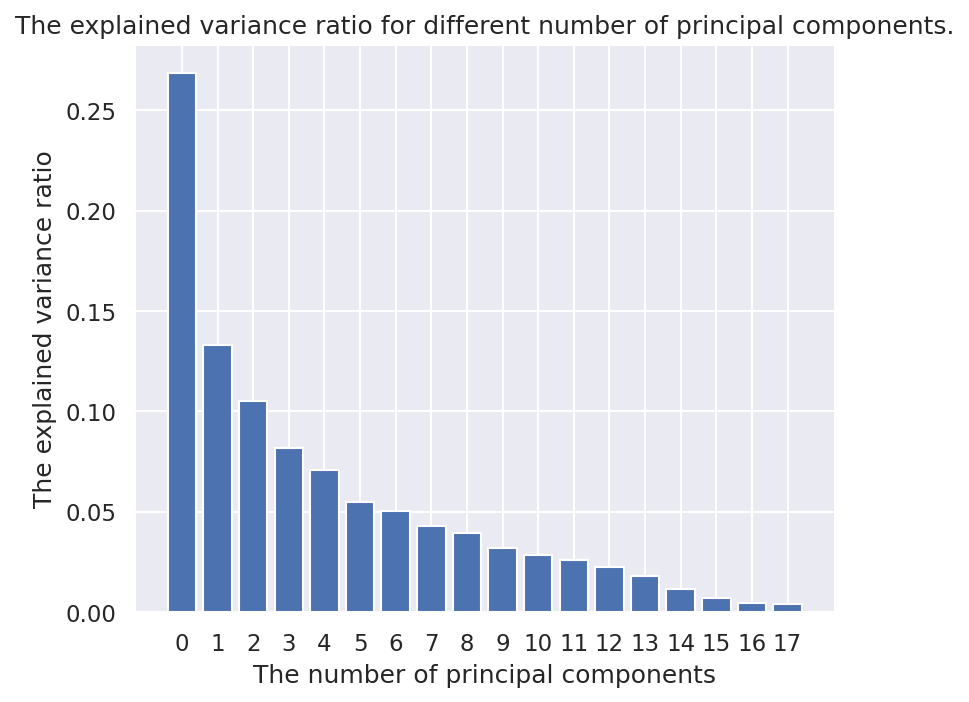
18.**Agriculture-**The percentage that agriculture contributes to GDP.

19.**Industry**-The percentage of GDP contributed by manufacturing industry.

20.**Service**-The percentage of GDP contributed by the Service industry.

**2.3 Cluster analysis to search for countries that have similarities to Kenya.**

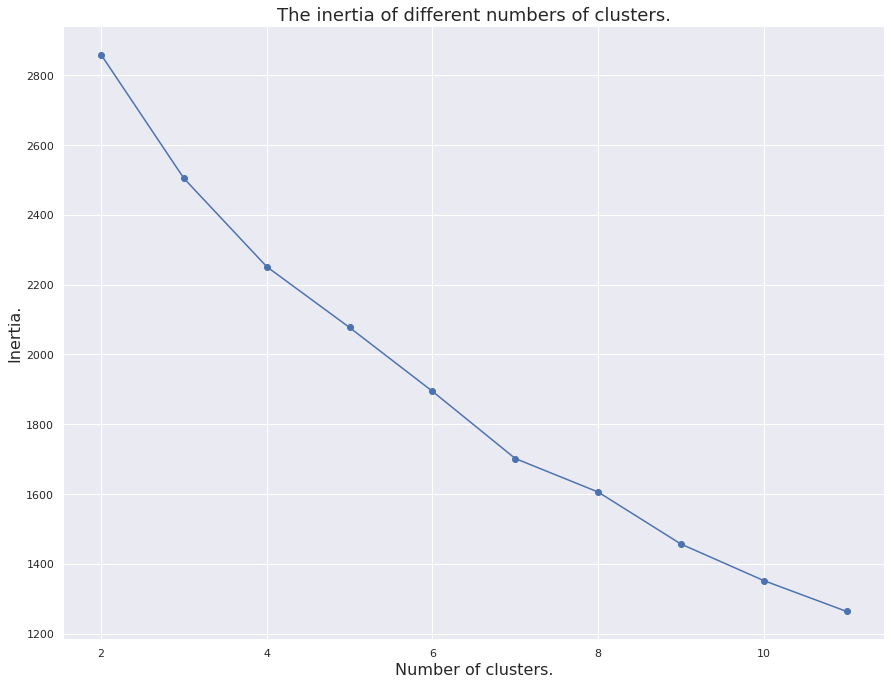
After standardizing the features in the data in order for them to have the same scale, cluster analysis was conducted to investigate which countries have the same characteristics as Kenya. Since the data set has a very high number of features, Principal component analysis was performed to reduce the number of features to prevent the curse of overfitting of the clustering model. An elbow plot was plotted to determine the most suitable number of features.



The graph indicates that 11 principal components would be a good choice since the change in the variance explained ratio from the 11th principal component to the 12th principal component is not significant. A choice of 11 principal components would also prevent the underfitting of the data.

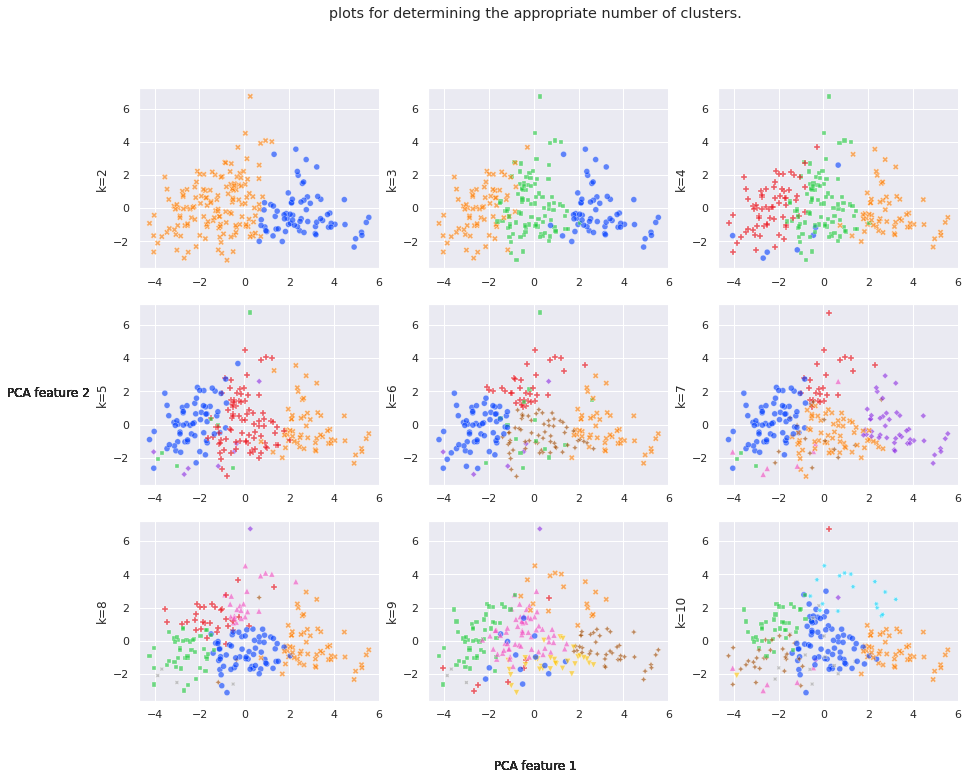
**2.4 Kmeans to cluster the countries.**

Since the data is unlabeled one of the most suitable algorithms to use for the labeling of the dataset into regions based on their similarities would be kmeans clustering due to its computational efficiency. In order to determine the most suitable number of clusters for this problem, inertia measures how close the data points are within a cluster vis-a-vis the centroid by setting the number of clusters to different integer values.



The elbow plot was not efficient in indicating a sharp decrease in inertia at any point in regards to the number of clusters hence cannot efficiently determine the suitable number of clusters.

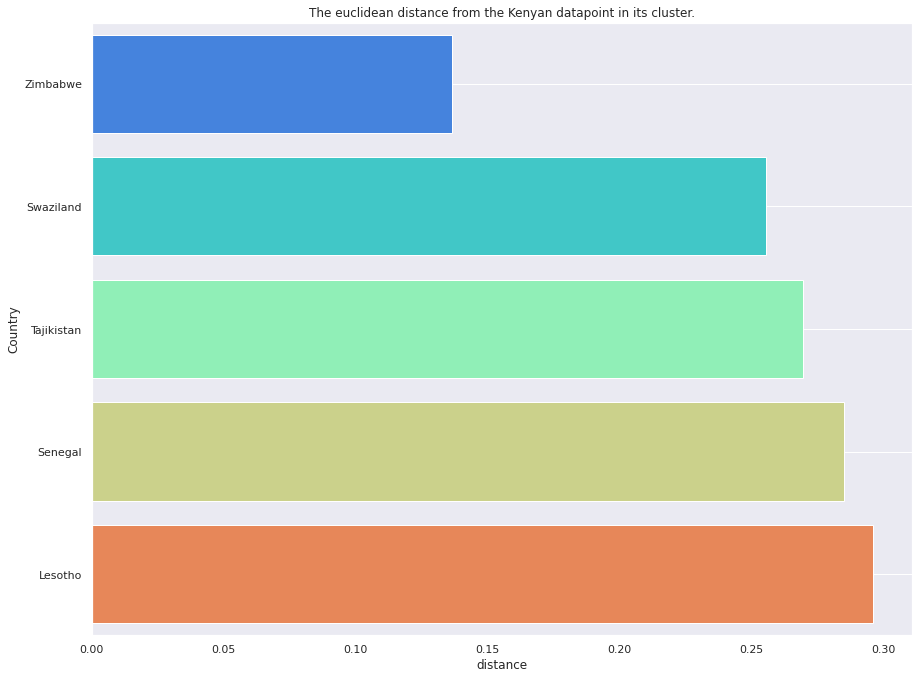
A scatter-plot was the go-to solution in order to determine the most suitable number of clusters. A scatter plot of the first two principal components with cluster labels grouped by a color palette to find out the most suitable number of clusters that separate the data into groups with very little overlapping.



**2.5 Finding the 5 similar characteristics to Kenya.**

Based on the subplot results, it seems that the most suitable number of clusters is 9 clusters that best splits the data into non-overlapping regions based on their socio-economic characteristics.

The 5 closest regions were then determined using the euclidean distance from the Kenyan data point in the cluster to which it was assigned. The Euclidean distance was computed using the first 2 principal components and the 5 countries whose data points were the closest to Kenya were selected.



The graph contains 5 countries that have the most similar characteristics to Kenya.

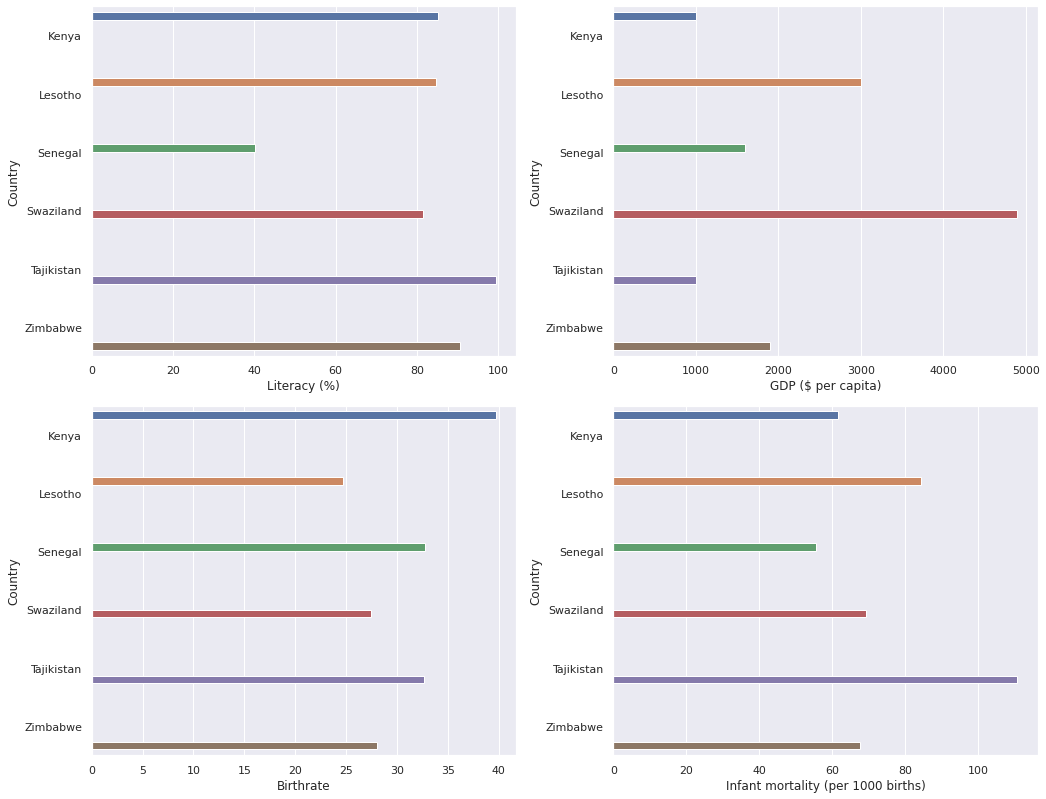
**2.6 Assessment of model performance.**

The model was assessed using silhouette distance as an accuracy metric.

The value of the silhouette distance was 0.1710861099862594 which is a fair performance of the K-means clustering model.

**Chapter 3: Results.**

After obtaining 4 features of the data of the 5 other countries' socio-economic features the similarities of the features were compared.



The plot containing the subplots of different features of the dataset shows that there is no significant difference in their socio-economic characteristics hence the model performed well in finding labeling the dataset according to socio-economic characteristics.