# Clustering African Capitals

Mohamed Abu Bakr

July-2019

## 1. Introduction :

### 1.1 Background:

**Africa** is the world's second largest and second most-populous continent, being behind Asia in both categories. At about 30.3 million km2 (11.7 million square miles) including adjacent islands, it covers 6% of Earth's total surface area and 20% of its land area. With 1.2 billion people as of 2016, it accounts for about 16% of the world's human population.

The continent is surrounded by the Mediterranean Sea to the north, the Isthmus of Suez and the Red Sea to the northeast, the Indian Ocean to the southeast and the Atlantic Ocean to the west. The continent includes Madagascar and various archipelagos. It contains 54 fully recognized sovereign states (countries), nine territories and two de facto independent states with limited or no recognition. The majority of the continent and its countries are in the Northern Hemisphere, with a substantial portion and number of countries in the Southern Hemisphere.

**The African Union (AU)** is a 55-member federation consisting of all of Africa's states. The union was formed, with Addis Ababa, Ethiopia, as its headquarters, on 26 June 2001. The union was officially established on 9 July 2002 as a successor to the Organization of African Unity (OAU).

### 1.2 Problem:

Considering diversity of African countries, this study aims to **cluster African capitals into three groups**. Each group will combine capitals with similar economic, demographical, and venues features.

### 1.3 Interest:

The African Union as a continental foundation with a lot of funding and construction plans on hand , would be highly interested in such study which may help greatly in resources allocations and fair funds distribution between countries based on current condition for each group of countries.

Also any investors aiming to start business in Africa, would make use of domestic, and demographic measures provided by study.

## 2. Data Acquisition and cleaning:

### 2.1 Data Requirements:

A lot of features could be used to cluster African capitals. This would include social, economic, geographic, and demographical features. Unfortunately not all of these data available online. Therefore, our clustering factors will include GDP, Population density, and average life in years of individuals. In addition to, different venues (hotels, stations, stores, etc.) located in the capital area.

Each cluster would combine African capitals with obvious similarity of the above mentioned features.

### 2.2 Data Sources:

Demographic measures including population density, average birth rate …etc., were scrapped from following webpage (http://doheth.co.uk/info/countries-of-the-world.php).

Location data (Latitude, and longitude) was scrapped from following web-page (http://techslides.com/list-of-countries-and-capitals).

Then different venues (including hotels, stations, shops, etc.) in each capital as a measure of how facilities exist and distributed will be obtained from foursquare platform.

Clustering will be limited to capitals where all above data is available (some capitals not covered by above online sources).

### 2.3 Data Cleaning and Preprocessing:

Records were available in the first two data sources for the whole world capitals (245 records from 1st source, and 199 records from 2nd source) .Therefore, we filtered data to only African capitals (keeping 58 records from 1st source, and 55 records from 2nd source).

As we scrapped data from two different sources, we checked length of data frame form both sources, found some capitals were covered by one of them, while ignored by the other source. We will limit our analysis to capitals covered by both sources (45 capital).

A check for missing values was performed, observed only two records missing only the value of country code [Windhoek, and Hargeisa]. As country code will not be used for our analysis, we kept records to make use of other columns data. More check revealed missing values in other columns in Hargeisa record, so we dropped it.

A check for data types was performed, found most columns not in numerical format, which will obstruct mathematical, and statistical analysis. Therefore, we converted columns to numerical to be easily processed.

## 2.4 Feature Selection :

Data retrieved from first source includes following features: ['Country Name', 'Capital Name', 'Capital Latitude', 'Capital Longitude', 'Country Code', ‘Continent Name']

Data retrieved from second source includes following features: ['Name (English)', ‘Location (Continent)', 'Capital city (official)', 'Currency (poss. multiple)', 'Population (inhabitants)', 'Area (km²)', 'Pop. density (inhabitants/km²)', 'GDP (nominal) (millions of USD)', 'Life exp. (Years)', 'Birth rate (births/1,000)', 'Death rate (deaths/1,000)']

Data retrieved from Foursquare include different venues such as hotels, shops, etc. will be encoded.

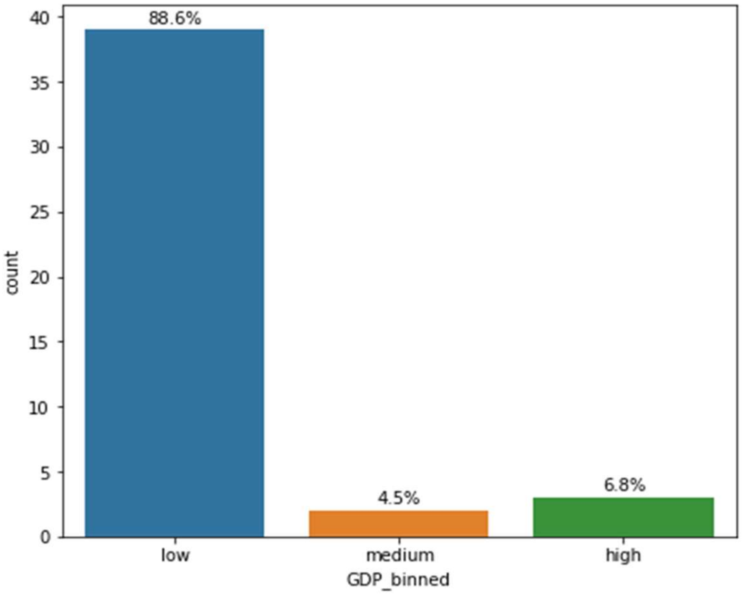
From first source will use location data (latitude, and longitude) to feed Foursquare with required dimensions to get available venues. Meanwhile, from second source we will use population density (it is more comprehensive and covers output of area, and population columns). Also, we will use GDP as the main economic measure in our analysis. Finally, we will use Life exp. (years) (again, it is more representative than birth rate or death rate)

# 3. Exploratory Data Analysis:

## 3.1 **Ranking Capitals:**

Binning: Values binning into groups rather than wide range of continuous values is accomplished for three main columns of GDP, Population density, and Life exp.it would help to get count of each bin. It would help in categorization of values into low, medium, and high categories.

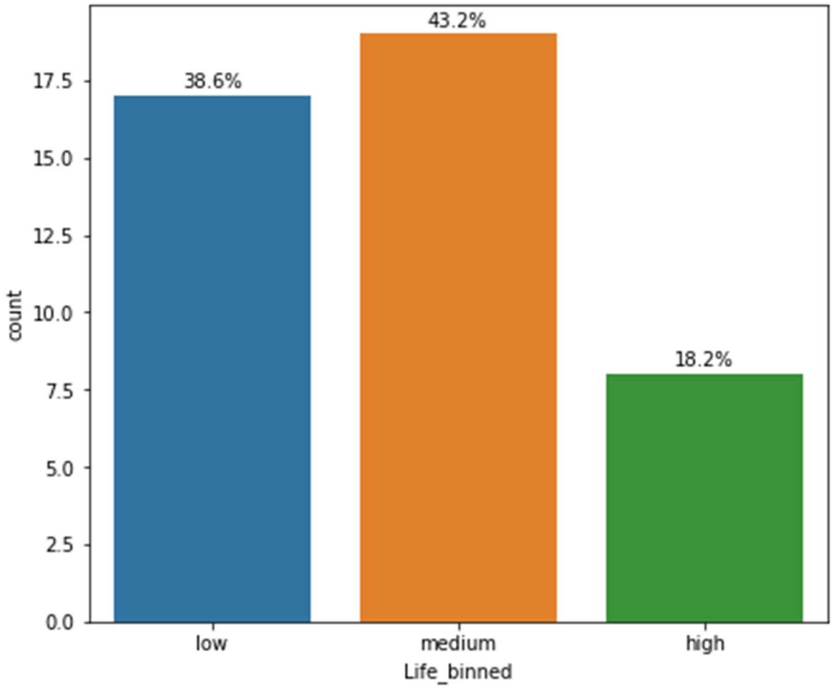
Regarding GDP (GDP_binned Figure), it is observed that 88.6 % of African capitals covered by study have low GDP (it is in good match with the fact that most African countries are developing countries with poor economic indices).

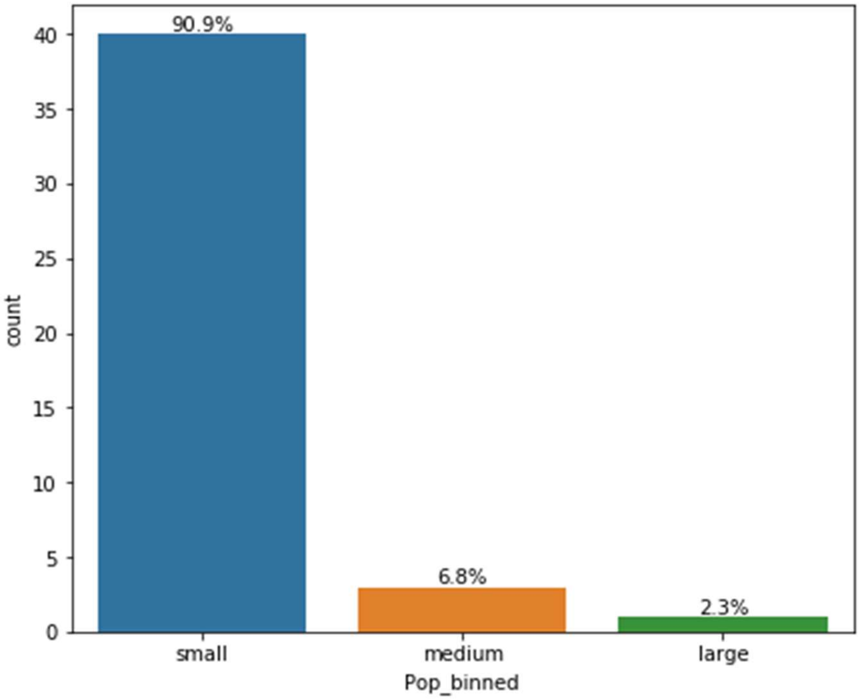


Regarding population density of African capitals, it is observed that only 2.3 of African capitals are highly populated.

Regarding Life exp. It did not show such sever difference, with about 43 % of capitals with medium values.

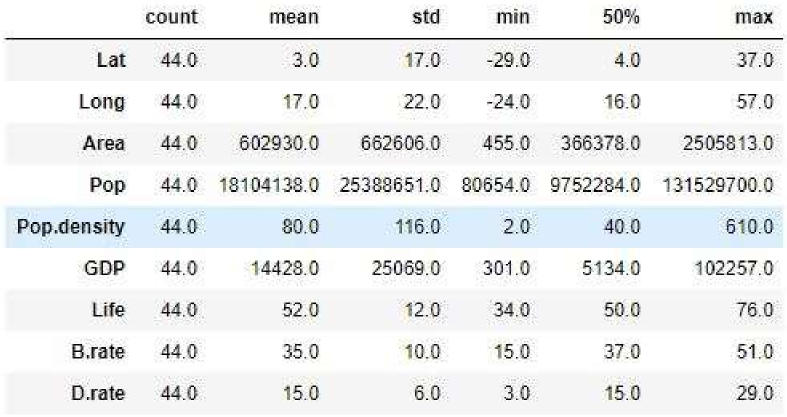
It is clearly charted in below graphs.


43.2%
38.6%
18.2%
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
count
low
medium
high
Life_binned


90.9%
6.8%
2.3%
40
35
30
25
20
15
10
5
0
count
small
medium
large
Pop_binned

## 3.2 Descriptive Statistics:

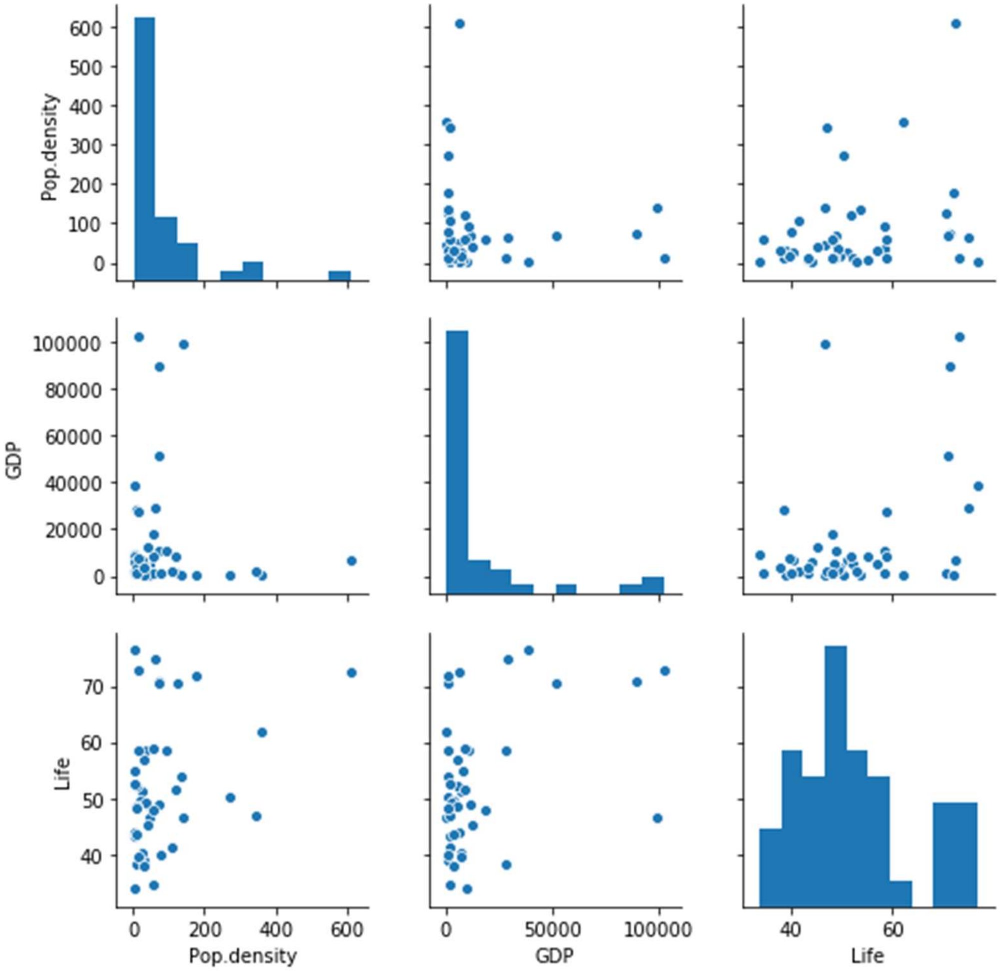
General descriptive analysis was performed, below table shows some statistical parameters including mean, standard deviation, minimum and maximum values.



| | count | mean | std | min | 50% | max |
|---|---|---|---|---|---|---|
| Lat | 44.0 | 3.0 | 17.0 | -29.0 | 4.0 | 37.0 |
| Long | 44.0 | 17.0 | 22.0 | -24.0 | 16.0 | 57.0 |
| Area | 44.0 | 602930.0 | 662606.0 | 455.0 | 366378.0 | 2505813.0 |
| Pop | 44.0 | 18104138.0 | 25388651.0 | 80654.0 | 9752284.0 | 131529700.0 |
| Pop.density | 44.0 | 80.0 | 116.0 | 2.0 | 40.0 | 610.0 |
| GDP | 44.0 | 14428.0 | 25069.0 | 301.0 | 5134.0 | 102257.0 |
| Life | 44.0 | 52.0 | 12.0 | 34.0 | 50.0 | 76.0 |
| B.rate | 44.0 | 35.0 | 10.0 | 15.0 | 37.0 | 51.0 |
| D.rate | 44.0 | 15.0 | 6.0 | 3.0 | 15.0 | 29.0 |

## 3.3 Relation between parameters:

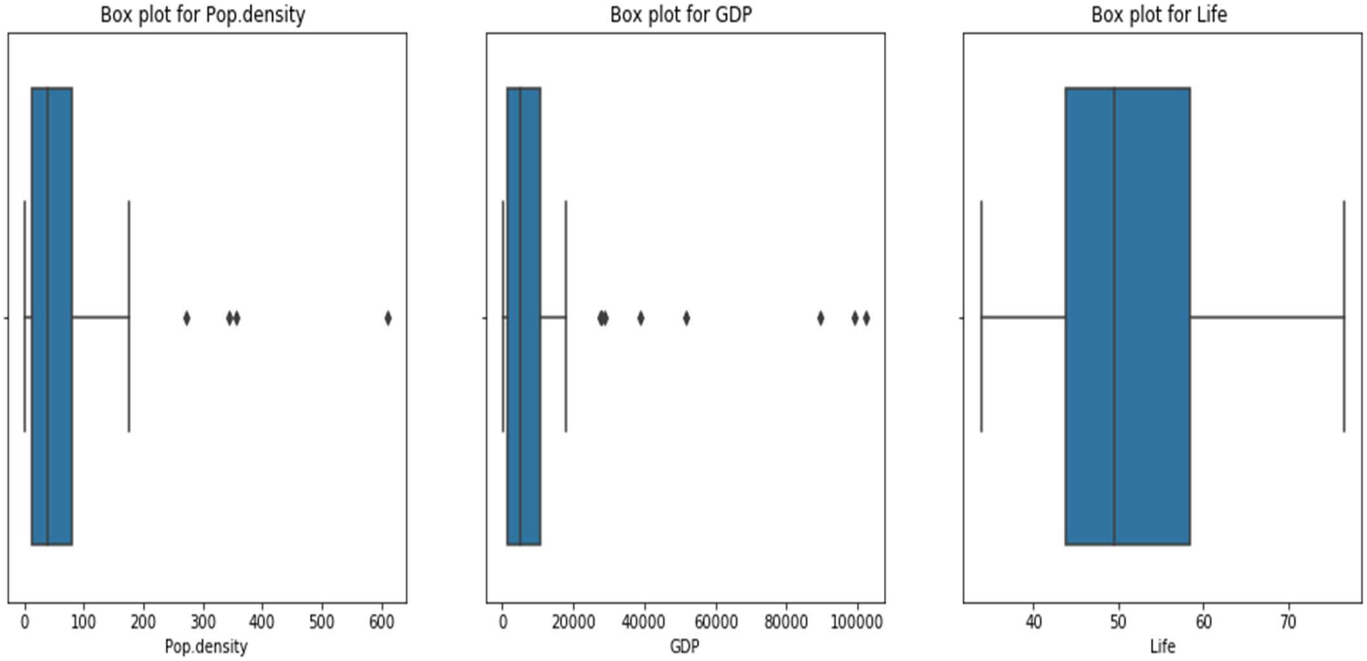
Plotting scatter plots between main features (GDP, Population density, and Life Exp.), does not show strong correlated between variable as per below figure.



### 3.4 Distribution of values:

Box plots were used to show how data centered about mean value. As per below figure, capitals with values of GDP more than 50,000 million of USD were encountered.

Algiers recorded the maximum GDP (102257.0 Millions of USD). Meanwhile, Port-Louis capital of Mauritius recorded maximum population density (610 inhabitants/km²)



## 4. Modeling:

### 4.1 Model Selection:

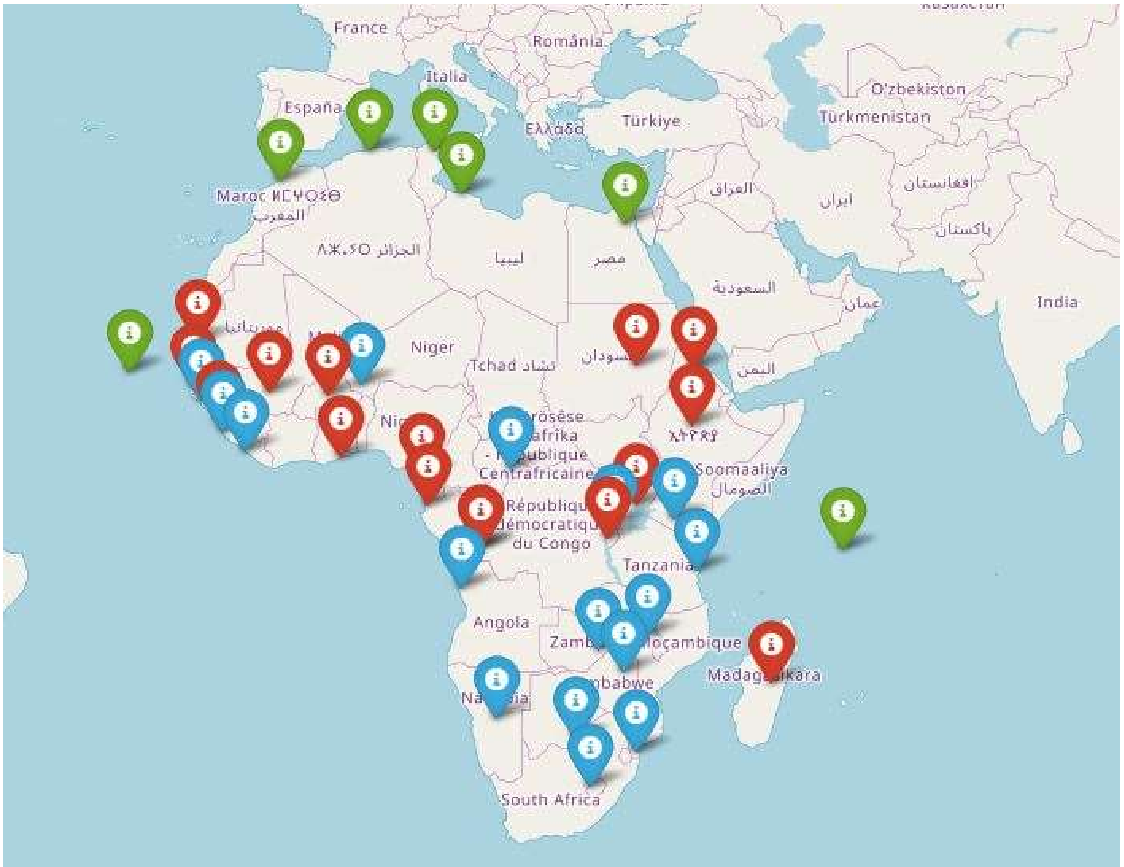
As problem requires grouping of capitals into clusters, unsupervised K-mean clustering algorithm is selected. It is direct and easy algorithm that should suit our small data set.

Model evaluation not applied here, only we should use domain knowledge to judge if clusters have real common features.

One hot encoding applied to all features (Venues, GDP, Life Exp., and Population density) to feed model with numerical values. Data frame with 39 records, and 140 columns was fed to K-means clustering algorithm.

## 4.2 Results:

Capitals are clustered into three main groups as per below map.



### 1st Red Cluster:

Most Common features here are low GDP & medium Life Exp. with medium to low population density, with hotel and bars are the most common venues (in good match with continent central countries actual conditions).

***2nd Blue Cluster:***

Most Common features here are low Life Exp. and low GDP, with diversity in the most common venues (in good match with continent southern countries actual conditions).

***3rd Green Cluster:***

Most Common features here are high Life Exp. & with medium to high GDP & low population density (in good match with continent central countries actual conditions).

## 5. Discussion :

From results, it is clear that neighbor countries share similar features. We may say that clustered capitals divided into southern, central, and northern groups.

## 6. Conclusion:

African capitals are clearly separated into three groups, with obvious economic, demographic differences.

Northern countries seem to have better economic measures, and low population density (considering larger area). Better business success chances would be achievable. Also, its capitals contain a lot of hotels that may accommodate tourists or visitor for any continental events.

Southern and central countries need more fund to help economy gets better, with more effort to provide health care.

## 7. Future Directions:

More updated data would help to make model more realistic. Also, we have to drop some capitals as no data available on hand. Finally, adding more social, and income measure would help to provide better modeling.