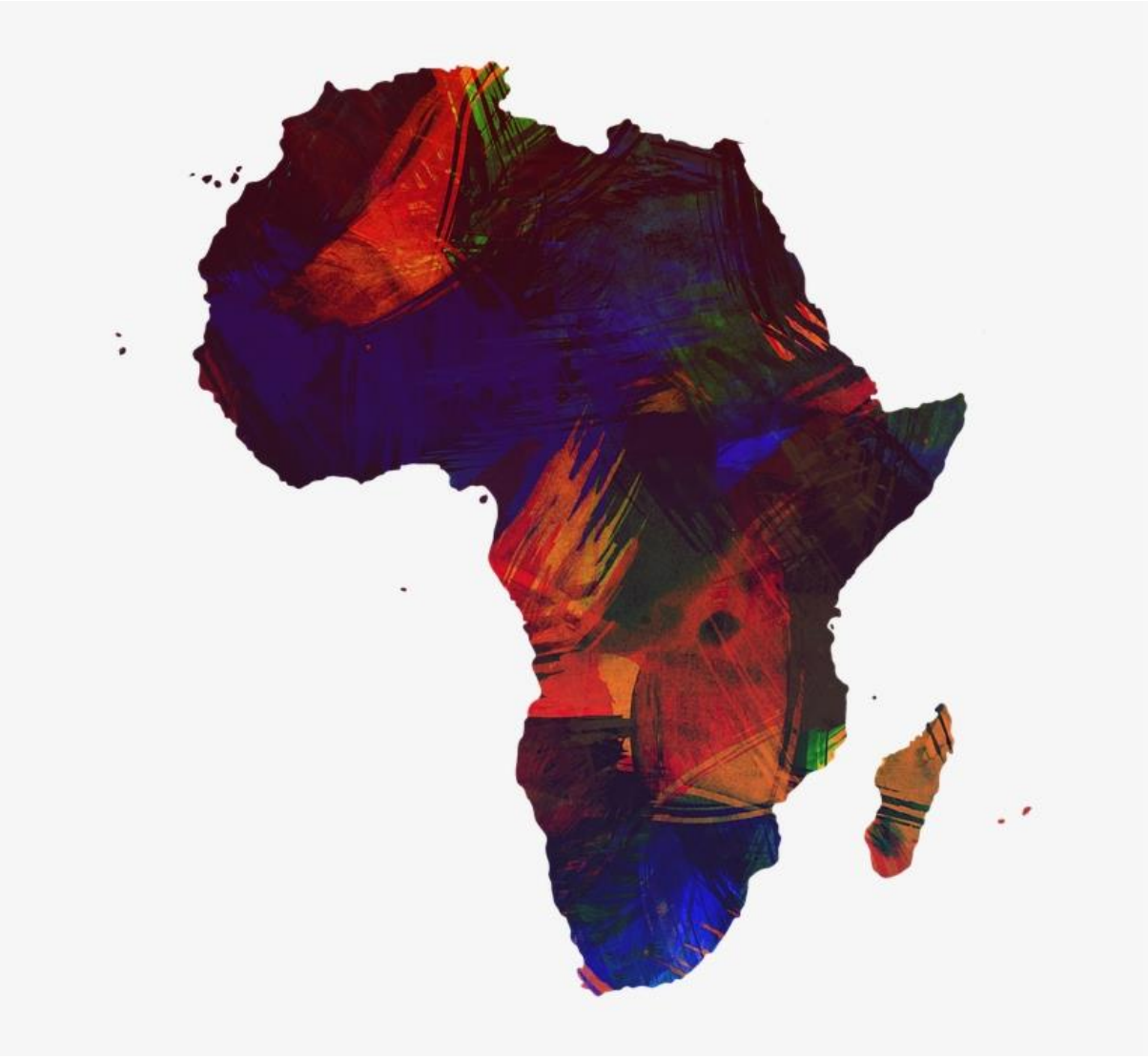
Clustering cities of Africa



Achraf OUGDAL

21/02/2021

1- Introduction :

• Background :

Africa is one of the most largest continents in the world. Despites of being one of the poorest continents of the world, it is still the most unique of them all. Whenever Africa is mentioned, most people think of roaming animals and jeep safaris. Whilst Africa has much more to offer than animals, the beauty that nature has bestowed on this continent cannot be overlooked. The animal inhabitants of Africa are amongst the most varied and beautiful in the world. Africa's beauty goes beyond this, its diversity cultural diversity is like pieces of art, where each piece is totally different than the others, and together they form a one big beautiful piece art.

• Problem :

The rich diversity of Africa never fails of amaze me. I also discovered that Africa is one of the most popular summer vacation destination.

I personally have a lot of colleagues that came from different countries Ivory Coast, Senegal, Burkina Faso and many more. The way they talk and act towards different situation and their style of clothing made me once wonder: **"having all this diversity, can we find any similar African cities ?** in other words, **Are there any similar cities in different countries and places all over Africa ?**

This project was developed to answer this question.

The problem is to cluster more than 900 cities in Africa on their location, population, density, surface, built-up area and their popular avenues. Then we will look at each cluster to find out which cluster offers most diverse characteristics so that we could choose a city of that cluster as our future summer vacation destination.

• Interest :

This project is develop to help people that looks for African cities that offer most diverse characteristics to find the best destination to travel to and enjoy every moment of their journey.

2- Data acquisition:

- ## Data sources:

To the address the problem described above, we need a dataset that will contain African cities with the following information:

- Name and country
- Coordinates: latitude, longitude and latitude
- Population and density
- Surface and built-up area
- Venues across each of these cities.

So, in order to prepare this information, we need data from different sources. mainly, these three:

- The name of the cities, their respective country code (ISO3), coordinates, population, density, surface and built-up area from the agglomeration dataset from https://africapolis.org/data. It offers an Excel Dataset that we will use for this project. The given dataset contains data about more than 9000 cities all over Africa, for this project, and for the sake of simplicity, we will only use a portion this dataset (almost 10%).
- Since the agglomeration dataset contains only the respective country code of each city in Africa (Example: MAR for Morocco), and in order to find the full name of each country, we will use the country data from https://africapolis.org/data and merge the two datasets in order to have the full country name of each city in Africa.
- The data about 100 venues within a 10 km radius of each city from Foursquare API

3- Methodology:

Data Cleaning and transformation:

For Data Cleaning, we followed the following process:

The agglomeration Data Downloaded from Africapolis contained multiple unnecessary columns. So, in order to clean it, I first drop ped all the unnecessary columns and only keep the ones that we will need for our task. I then renamed the other columns to have a much clearer name for each column (example: I renamed the “ Agglomeration_Name “ column to “ City ”). There were no missi ng values and all the columns had the right types. So finally,

I was left with 8 columns (City, ISO3, Longitude, Latitude, Altitude , Population, Density and Builtup).

For the Country Data, I followed the same process as the one above, and I kept only two columns (Country and ISO3). ”). There were no missing values and all the columns had the right types.

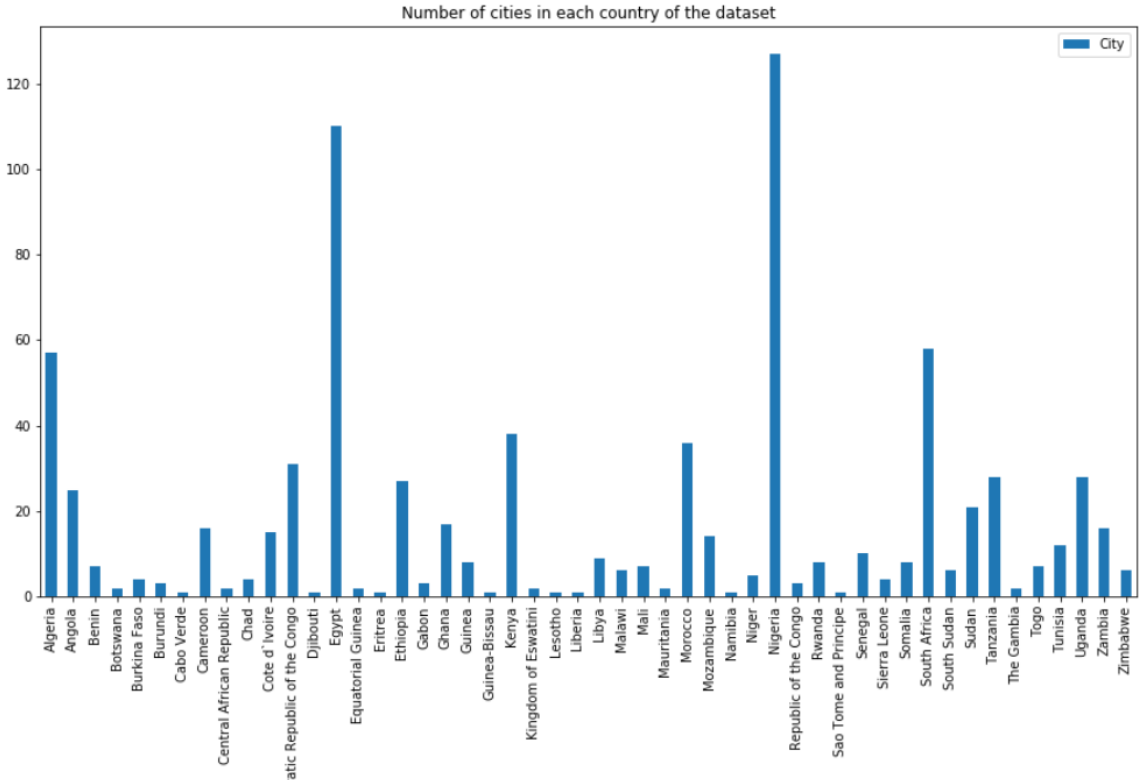
Now that the two parts were cleaned well, I merged the two dataframes to have one large dataframe that contains each city with its appropriate country. Then I dropped the ISO3 Column since I won’t need it anymore.

I then used Foursquare API to get, for each city, the venue’s name and their categories within a 10km radius. The resulted dataframe had 8289 venues each with its name and city. And finally, I dropped the “Venue_name” column since I won’t need it anymore. Finally, I created one-hot encoding for each cities using unique venue_category as feature set. I grouped data by city and by taking the number of occurrence of each category. Then I merged the dataframe with the one in the step before and finally got one final dataframe.

Exploratory Data Analysis :

Number of cities in each country :

We can see from the graph bellow, that most of the cities in our dataframe are Nigerian, Egyptian Algerian, and South African cities.

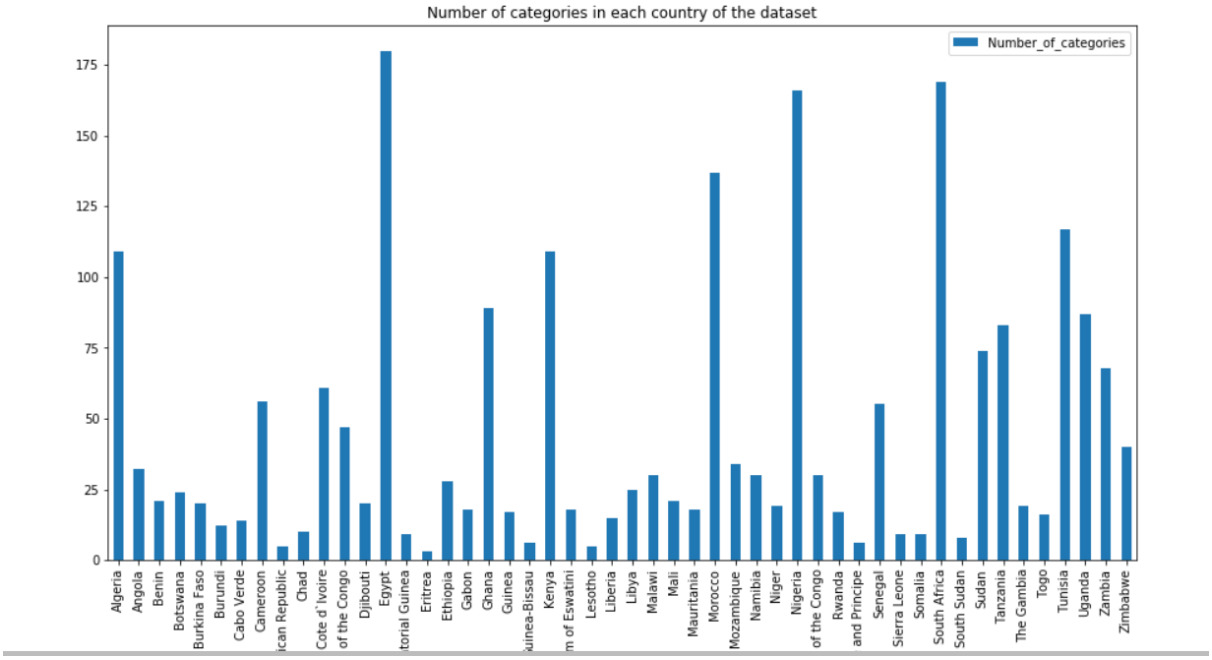


Location of each city in Africa

To visualize each city, a map of Africa was plotted using Folium with cities superimposed on top (as shown below).

Number of categories in each country:

Here, we can see that in this dataset, Egypt, Nigeria south Africa, Morocco, Tunisia, Kenya and Algeria have the more categories than all the other countries. One thing to notice is that in Morocco, even that he doesn't occur too many times in the dataset, yet it has more categories than Algeria.



After doing our exploratory data analysis, I finally created one-hot encoding of each country.

Clustering cities of Africa Using K-Means Algorithm:

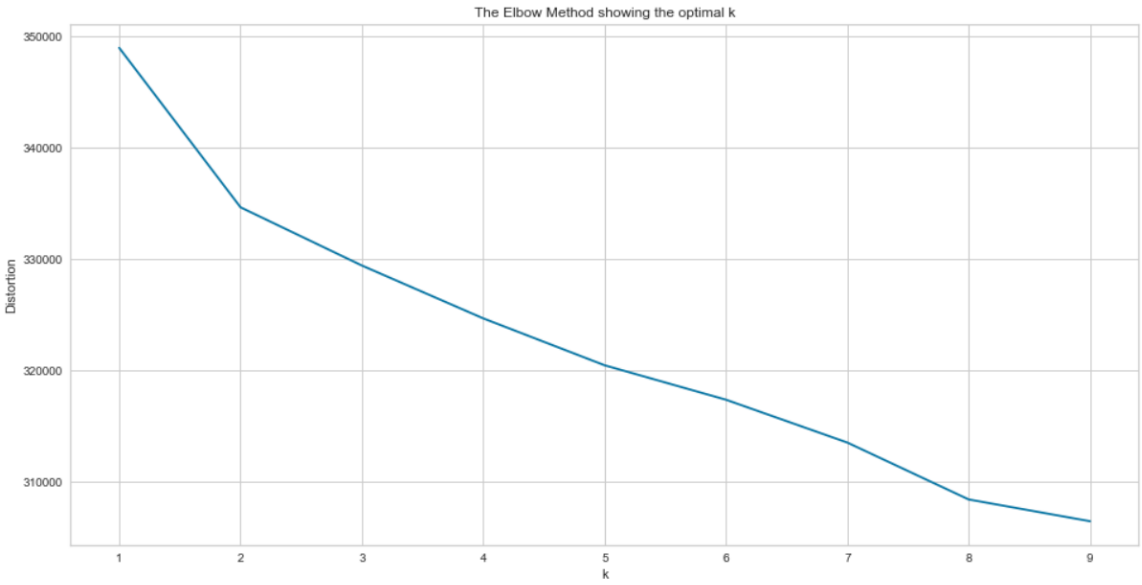
After the data has been prepared, now it’s time to cluster the cities into different clusters.

I used K-Means Algorithm for this purpose because it is much efficient on large datasets and large features.

In order to find the Best K for the K-Mean algorithm. I used the Elbow method. The Elbow method is a very popular technique and the idea is to run k-means clustering for a range of clusters k (let's say from 1 to 10) and for each value, we are calculating the sum of squared distances from each point to its assigned center (distortions).

When the distortions are plotted and the plot looks like an arm then the "elbow" (the point of inflection on the curve) is the best value of k.

Applying the elbow method on the dataset resulted the following graph.



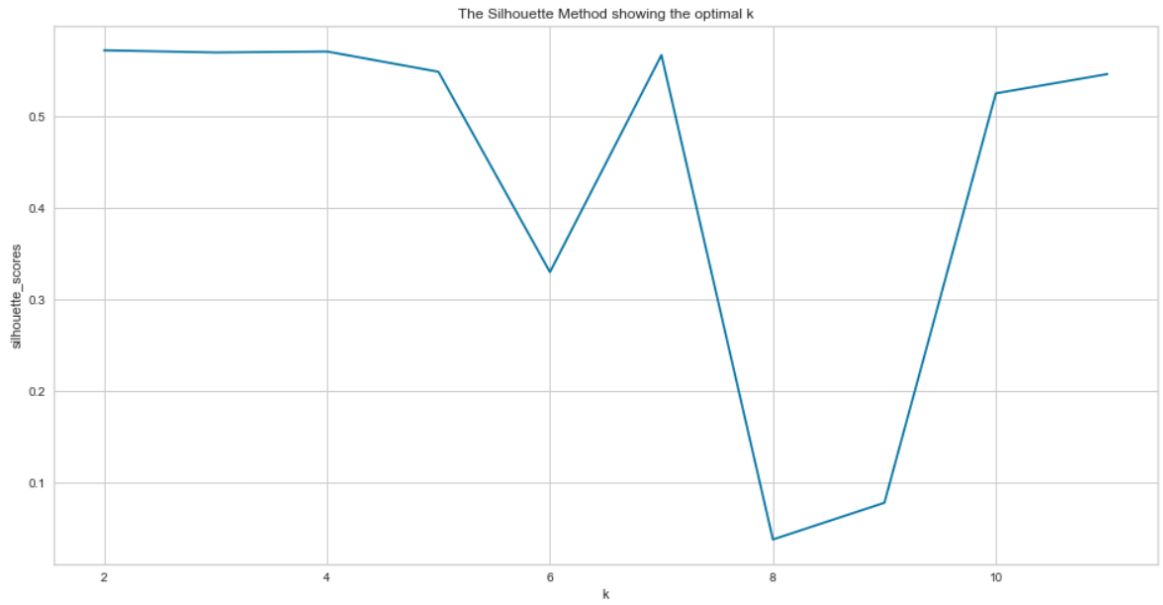
As we can see the plot is a bit confusing and there is no visible Elbow. so let's try another method to find the best K
This time, we will use the silhouette analysis to find the best K.

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters and negative

values indicate that those samples might have been assigned to the wrong cluster.

Applying the silhouette analysis on the dataset resulted the following graph.



As we can see here the plot is much clearer. The best value for K here is 7.

Building the Model with K = 7

Scikit Learn's KMeans has been used for this purpose. When clustering is done, we obtained dataset that contains the cluster labels along with the location of each city. The clusters are plotted on a folium map centered on Africa and the results are given in the results section of this report.

4- Results:

By plotting the results of the clustering, I found some interesting facts.
First, here is the resulted map.

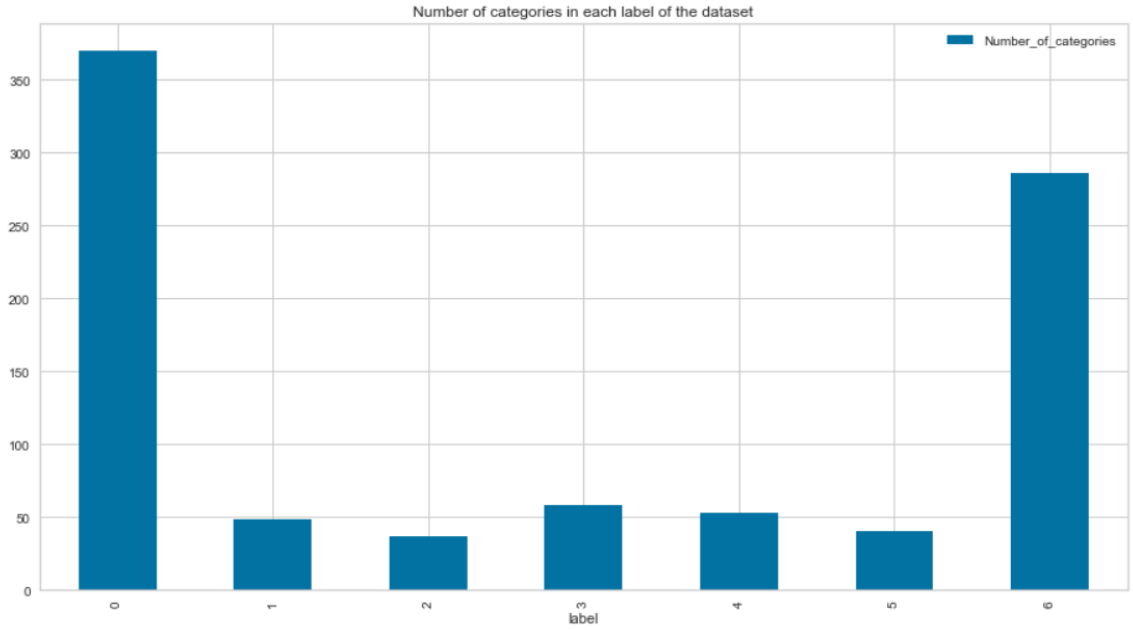
The first interesting this to notice about the results shown in the map above, is that the majority of the African cities are similar (cluster 0).

But noticing the pink clusters (cluster 6), we can see that cities that belongs to this cluster are located on the "corners" of Africa they are all coastal cities, so we can say that they share the same characteristics.

We can also notice that there are 4 cities that somehow don't belong to any cluster, but they formed a cluster of their. these cities are:

- Living stone : cluster 5 - orange
- Accra : cluster 1 - blue
- Sfax : cluster 4 - black
- Cairo : cluster 3 - purple
- Al-Mansoura : cluster 2 – green

We can see the number of different venues categories in each cluster by looking the following graph:



As we can see that the cluster number 0 and 6 are the most diverse clusters. they are respectively the cluster with the colors red and pink on the Folium Map. So, we can say that most of the cities of Africa are diverse and that each city has much more to offer in terms of characteristics.

5- Discussion:

From the obtained results we observed the following inferences. The red cluster is the most diverse cluster among all clusters, and since the majority of cities in Africa belong to that cluster, we can say that the majority of cities in Africa are similar and at the same time diverse and have rich characteristics. They have almost all the categories in the dataset.

For the pink cluster is also diverse and rich and has almost 60% of all the categories in the dataset, but what makes this cluster special, is the location of its cities, they are coastal cities and they are located on the “corners” of Africa. It is very interesting to know that such group exist and it is unique on its own.

Finally, 3 of the cities that formed different clusters on their own (Sfax, Cairo and Al-Mansoura) are located in Egypt and Tunisia. They are all Arabic countries but even though the 3 cities are similar to each other, and also similar their neighbours cities they somehow created clusters of their own.

6- Conclusion

To conclude, I would say that during this project, and even that I am a Moroccan student, I was surprised to know that Africa is really rich and diverse and has a lot more to offer than Safari and animals. It has its own beauty, every country in this beautiful continent has its own characteristics and at the same time almost all of the countries are similar. As I mentioned in the beginning of this report, this continent never fails to amaze me. I really enjoyed working on this project, and I am very proud of the results that I obtained. Also, we can say that for Adventure lovers, the majority of cities of Africa will be a great summer destination. The diversity of the cities of Africa is just perfect for adventure and exploring lovers.