**NAME: ANIZOBA EBUKA**

**IBM CAPSTONE PROJECT**

**PREDICTING THE BEST PLACES TO SIGHT AN OFFICE IN KANO**

1. **INTRODUCTION**

**1.1 Background**

Kano is one of the most populated places in Nigeria. It is also a developing area. It is filled by the northerners. These northerners speak Hausa language(which is among the three major languages in Nigeria. Apart from Lagos and Abuja, it is the next place most tourists would want to visit. Lagos and Abuja is already populated with offices

And even though they are good places, Kano is a rising place where one would love to open an office or headquarters

**1.2 Problem**

This analysis would be done by me and I don’t live in Kano. This means I would have to get various data including location data to be able to analyze the data as if I had been living there all my life. The project aims to predict the best place in Kano to open an office

**1.3 Interest**

Most people would love to open an office in a developing area for many reasons. One of the reasons could be that it would be cheaper to open an office in a developing neighborhood than in a developed one. Another reason could be that they want to get the best location for that neighborhood

1. **DATA ACQUISITION AND CLEANING**
   1. **Data Sources**

The data set used for this analysis was gotten from https://en.wikipedia.org/wiki/Kano\_State#:~:text=Population,populated%20by%20the%20Hausa%20people. The data in the table were Area, Census 2006, Administrative capital, Postal Code.

* 1. **Data Cleaning of data from Wikipedia**

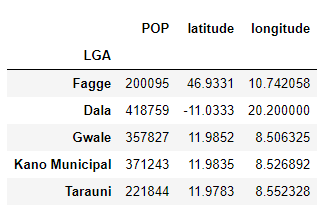
The table scraped from Wikipedia had to be cleaned. Some columns in the table were not needed. These were deleted and the needed data was scraped and put into a data frame. It should be noted that the only census data present was data from 2006.

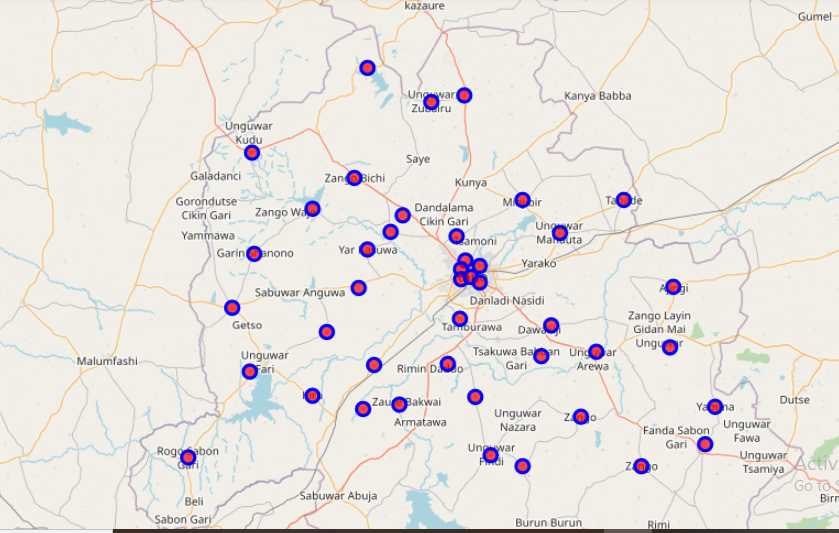
There was also a problem with the numbers. The numbers were written in the form of “200,142”. These numbers cannot be converted to an int so I had to clean it up and convert it to a form where it could be used “200142”

The data gotten had “\n” after the names of each local government which was used to represent next line when building the website but this is not needed in our data. The table was showing names as “Gwale\n” and we wanted it to show “Gwale”.

The data gotten from Wikipedia did not have longitude and latitude values and this was essential for our plotting using folium. Nominatim is used to get latitude and longitude of an address. We used Nominatim to get these values. It was done for each value in the index of the dataframe. The user agent used was “NG\_Explorer”

**Table 1 (data frame after longitude and latitude were added)**

****

**Map 1(Kano Municipal Map showing its LGAs) **

**2.3 Foursquare data**

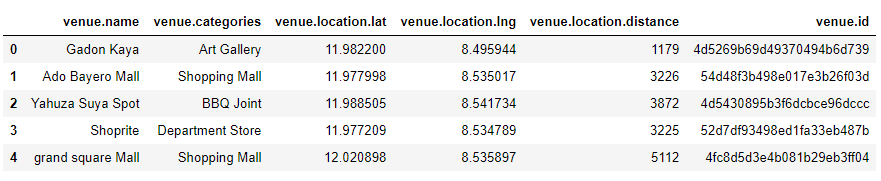
Foursquare is a free website that allows for users to search for nearby venues and see information and reviews about various places. Users can also give reviews on places in the website. The website leverages your location to give you the best insight on where to visit. Foursquare has over 75 million tips from over 105 million venues. It is one of the largest sources of location data. Others include Google except for the fact that Google’s Api is not free.

This project would leverage data for Kano Nigeria. The particular data needed is venues for various local Governments in Kano, their locations (latitude and longitude) . Foursquare data is accessible for developers through a free API (application programming interface). This API can be gotten through a free account. The website is <https://developer.foursquare.com/places.api> .

**2.4 Four Square data**

Four square data was used to get venues for a ten kilometer radius. Its limit was set to 200 venues at max. Four Square was used to get some specific columns which includes the venues names, categories, location(longitude and latitude), distance and venue id. Other columns were dropped.

**Table 2 (foursquare data for ‘gwale’ local government area)**



1. **METHODOLOGY**

This section describe data inferences, data exploration and machine learning models that were carried on the data to be able to give insights on places where

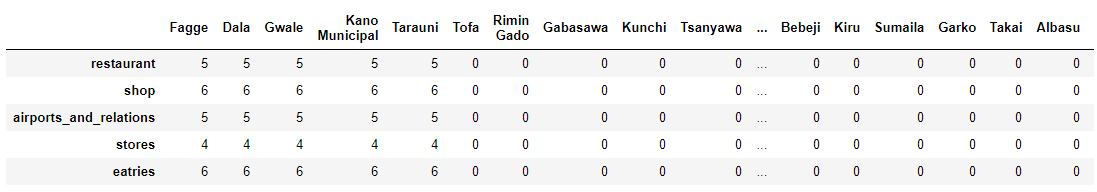
offices should be located.

* 1. **Data Exploration**

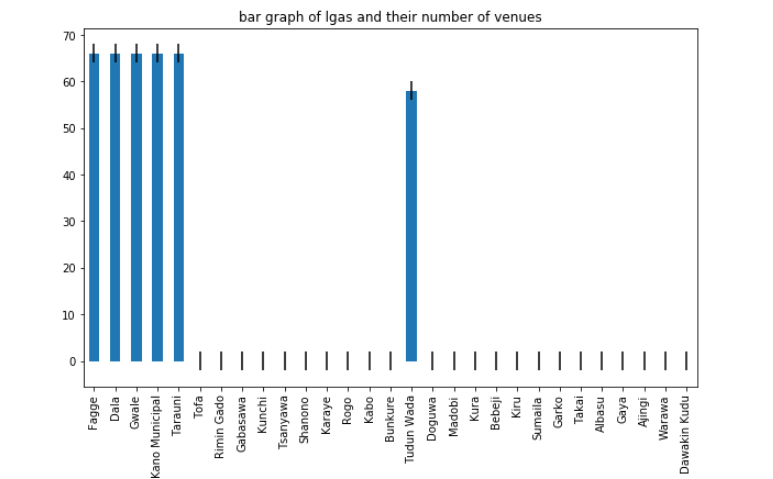
The 23 different kind of categories gotten from the various local governments in Kano were grouped into 8 categories in which they fall in which includes restaurant, shops, airports\_and\_relations, stores, eatries, gym\_and\_sports, funplaces, hotel

A data frame was created with these categories as index and the various local governments were used as the columns. The picture below shows the data frame.

**Table 3 (dataframe of categories and LGA’s)**



**Bargraph 1(Bargraph of summary\_df)**



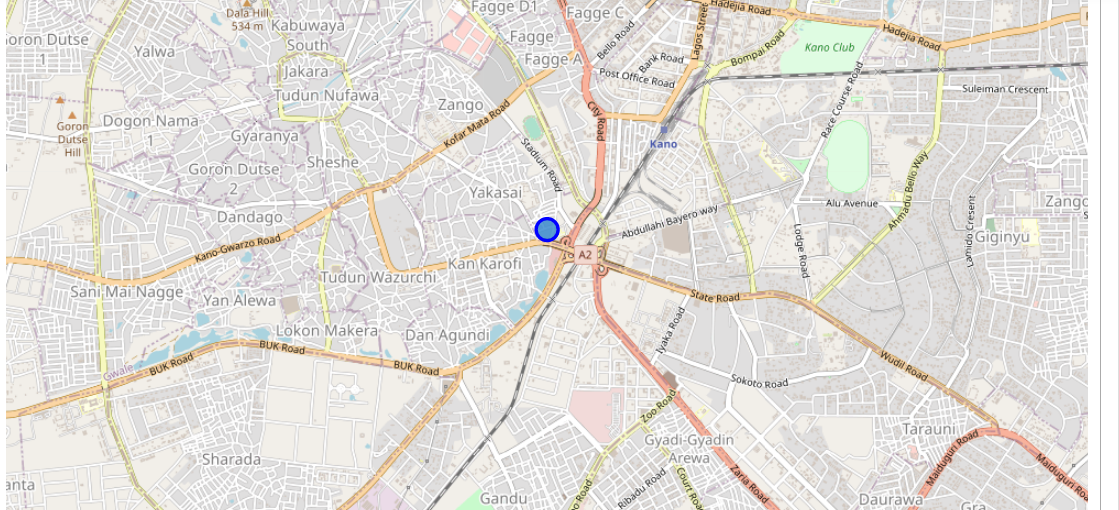
* 1. **Data Inferences**

I used the data to check for the local government area for the highest number of categories and the lowest number of missing categories i.e. the value of zero for a particular category. Fagge, Dala, Gwale, Kano Municipal, Taruni and Tundun Wada had more categories than the rest as seen above. We would be analyzing Kano Municipal

* 1. **Data Exploration of Kano Municipal**

The picture below is a folium map of Kano Municipal in Kano State Nigeria. The venues in Kano Municipal in a ten kilometer radius were gotten through four Square API.

**Map 2 (Kano Municipal Map)**



**Table 4 (Kano Municipal categories of places.)**



* 1. **Data Exploration of Kano**

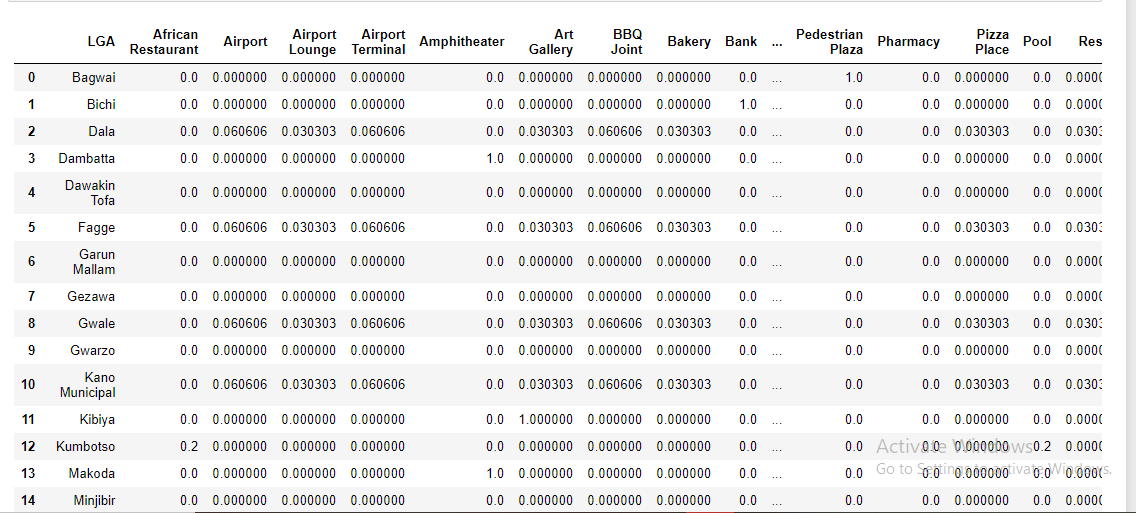
We have to explore the whole of Kano to determine what categories of places are most popular and what would be a good idea to open. This was done by getting venues of that are in 10 kilometer radius from the latitude and longitude of the local government areas. The picture below shows the first sixteen values from the 227 venues gotten

**Table 5**



The data was then grouped according to categories and their categories. The mean was also taken to give the table below.

**Table 6 (LGAs and the categories in them)**

****

The LGA’s were then grouped according to their most common venues. The

top ten most common venues were gotten and put in a table shown below.

**Table 7 ( LGAs and their top 10 categories)**

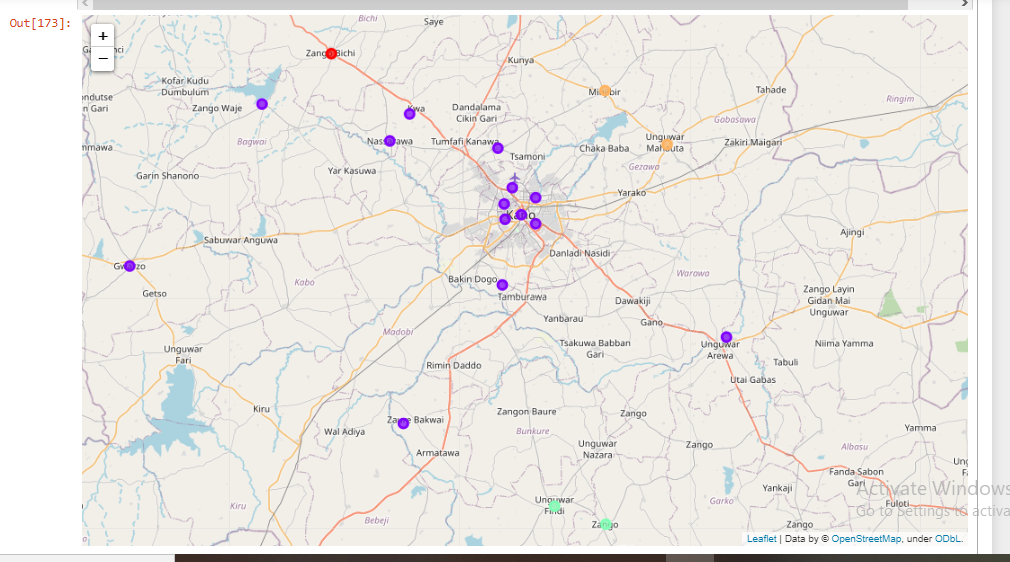


* 1. **Clustering**

The venue data will now be explored for similar distributions of venue types by selected neighbour-hoods. The clusters will be generated using the k-means algorithm, and displayed on the map of Kano

I selected five clusters. The cluster labels were 0,1,2,3,4. After the clusters were gotten, they were plotted in the map as shown below.

**Map 3**

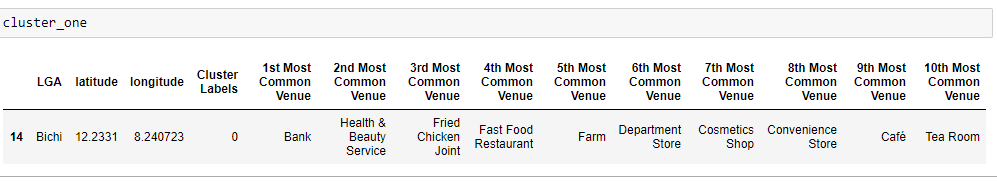


The map shows 5 different colors for different clusters

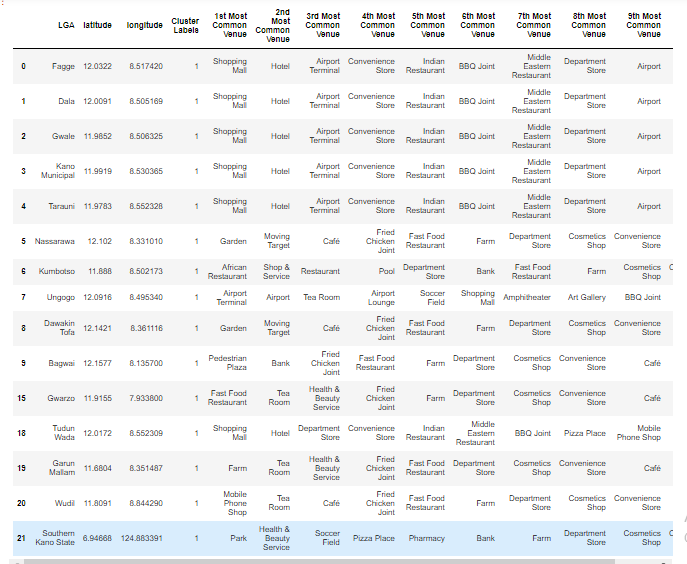
* 1. **Examining the clusters**

The clusters are clustered according to their similarities. This includes similarities in most popular venues.

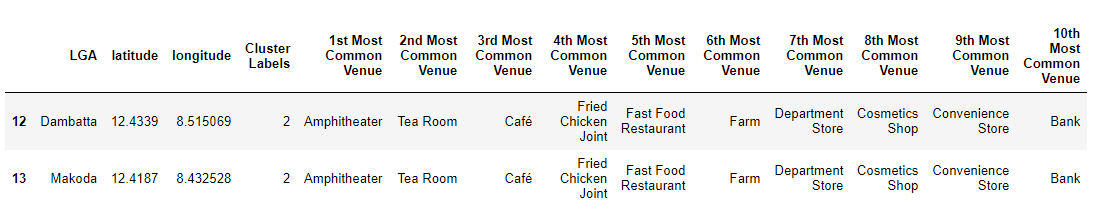
**Table 8 (cluster 0)**

****

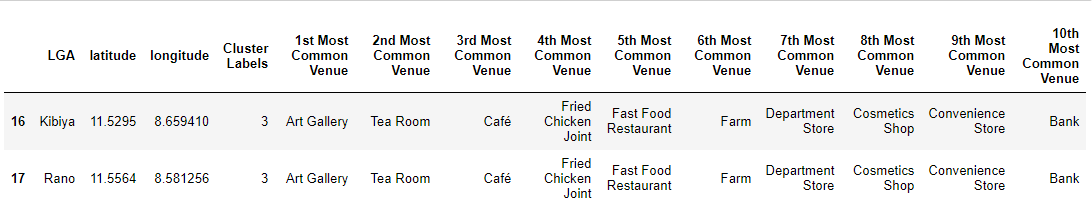
**Table 9 (cluster 1)**

****

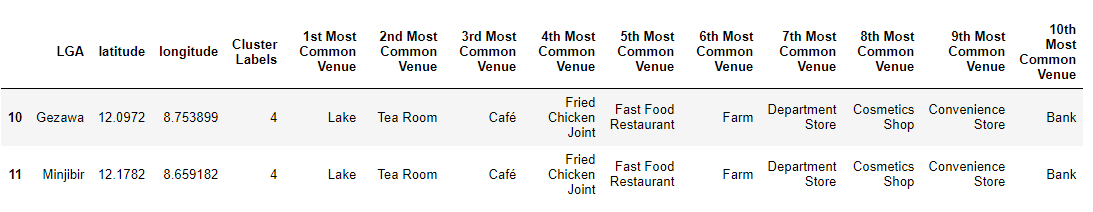
**Table 10 (cluster 2)**

****

**Table 11 (cluster 3)**

****

**Table 12 (cluster 4)**

****

The categories of places in a word cloud for Kano was gotten. A bigger word means it has more numbers. The word cloud is shown below



1. **RESULTS AND DISCUSSION**

This section shows results gotten from the outcomes of the methodology and their relevance to the problem of the research.

It was determined that Kano Municipal, Fagge, Dala, Gwale, and Taruni were great places to site an office due to availability of various venues in it. Kano Municipal was singled out among these good venues because it had the smallest amount of missing categories of venues.Kano Municipal’s categories and their numbers are shown in table 4.

It was seen that the highest number of categories was convenience store, then departmental store then cosmetic store as seen in the word cloud of Kano.

Cluster 0 has a lot of banks as its most popular category. The local government in it is Bichi. Cluster 1 has a lot of Shopping malls, parks and gardens as its most popular categories. Its local governments include Fagge, Dala, Gwale, Kano Municipal, Tarauni, Nassarawa, Kumbotso, Ungogo, Dawakin Tofa, Bagwai, Gwarzo, Tudun Wada, Garun Mallam, Wudil, Southern Kano State. Cluster 2 had Amphi theater as its most common first category. The local governments include Dambatta, Makoda. Cluster 3 had Art Gallery as its most popular category and its local governments includes Kibiya and Rano. Cluster 4 had lakes, tea rooms and cafes as its most popular categories. Its local governments includes Gezawa and Minjibir

1. **CONCLUSION**

The aim of this work is to help provide the necessary amenities to help people decide on where to set up. Using data gotten from Wikipedia, I was able to analyze Kano state data.foursquare data was used in getting the location data and Wikipedia was used in getting the local governments as well as their population . I used a machine learning algorithm known as KMeans Classification from the sklearn library to classify the data to 5 clusters

1. **DRAWBACKS.**

Foursquare data wasn’t enough to get all the locations in Kano. Universities were not shown. Small shops were not shown either. The conclusion gotten was just from the data available which needs a lot of room for improvement.

Some local governments had no data on them so where excluded from this analysis