Segmenting the potential market for the E-Systems®software development company

Alaa Mahjoub

June 2020

1. Introduction

1.1 Background

E-Systems® company is a software development company specialized in developing software systems for automating the business of hotels, coffeeshops and restaurants. The company develops three software products:

A- The Hotels Automation Software (HAS)
B- The Coffeeshops Automation Software (CAS)
C- The Restaurants Automation Software (RAS)

The company Marketing Department includes three Teams: A Hotel Automation Marketing Team, a Coffeeshop Automation Marketing Team and a Restaurant Automation Marketing Team.

The company is planning to expand its market by identifying potential overseas customers in relevant national capital cities across the whole world.

1.2 Problem

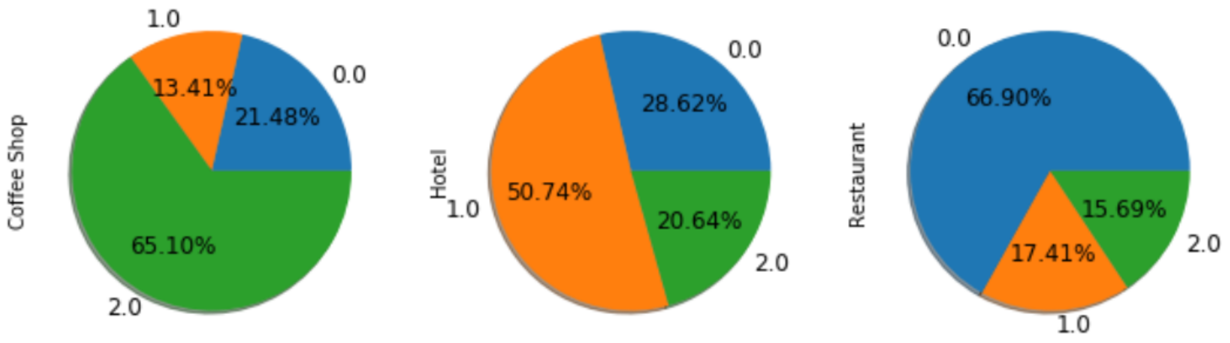
In order to expand its market, E-Systems® adopted a data driven approach and formulated a new market development strategy based on geo-demographic market segmentation. The data which will contribute to the market segmentation process includes:

- the national capital city name and its country name
- the national capital city geographical coordinates (i.e. the longitudes and latitude data of the city)
- the number and the category of potential customers in each national capital city (i.e. the number of hotels, the number of coffeeshops and the number of restaurants in thecity)

This project is a data clustering project and it is aimed to segment the national capital cities into three marketing segments (i.e. 3 clusters), Although the above mentioned data will contribute in the segmentation process, the segmentation itself will be done according to the number and the category of potential customers in each city(i.e. the number of hotels, the number of coffeeshops and the number of restaurants). In this approach, each Marketing Team will lead the new market development efforts in one of these three market segments internationally.

Figure 1 below depicts an example of the potential market segmentation, and Figure 2 depicts an example of a geo-demographic market segmentation based on this data clustering approach.

Figure 1 - An example of potential market segmentation

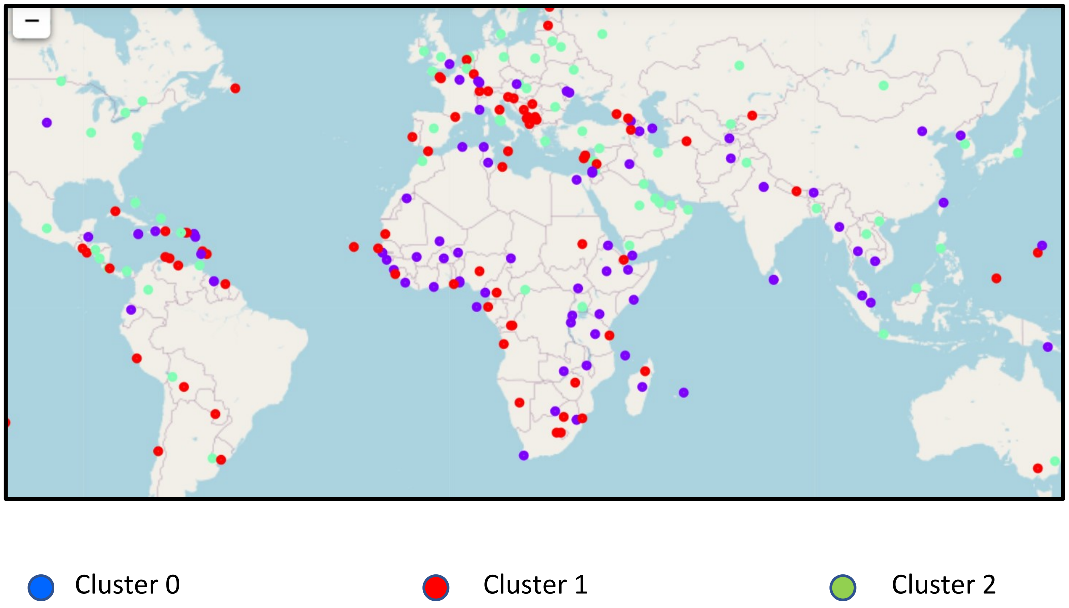


Market Segment 0
(National Capital City Cluster 0)

Market Segment 1
(National Capital City Cluster 1)

Market Segment 2
(National Capital City Cluster 2)

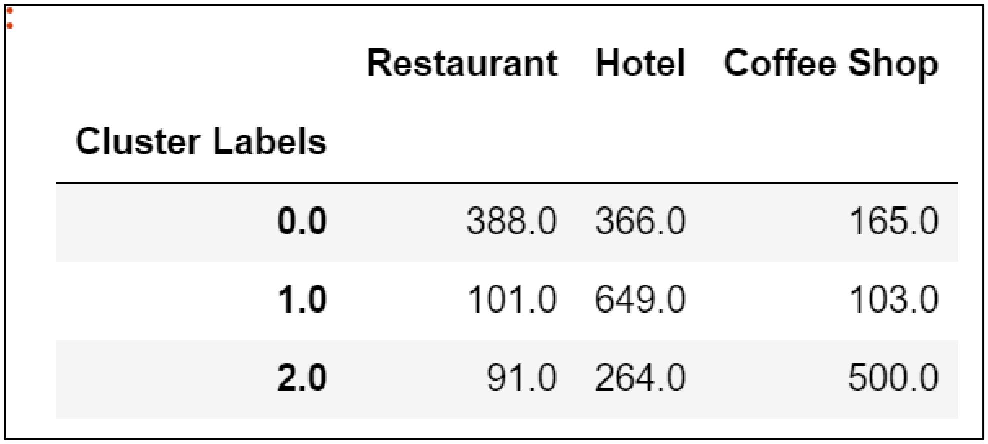
Figure 2 - An example of geo-demographic market segmentation



1.3 Interest

Obviously, it is in the interest of E-Systems® marketing department to know which capital cities will be managed by which marketing team. The size of the new potential market in each segment is also of interest to the department (see Table -1 below for an example). This information will help the Marketing Department develop or acquire the appropriate human capital competencies necessary to manage the new international market development activities. Other audiences who care about this problem include the E-Systems® company Senior Management and the company's shareholders.

Table 1- An example of potential market segment size (expressed in number of customers in each business line in each cluster)



	Restaurant	Hotel	Coffee Shop
Cluster Labels			
0.0	388.0	366.0	165.0
1.0	101.0	649.0	103.0
2.0	91.0	264.0	500.0

2. Data acquisition and cleansing

2.1 Data sources

Table 2 below describes the data sets used to build the clusters and their corresponding data sources.

Table 2- the data sets and their data sources

No	Data set	Description	Data Source
1	List of world-wide national capital cities	Data fields include City, Country and Notes. See Appendix I for an example of this data set.	I scraped the following Wikipedia site to obtain this data https://en.wikipedia.org/wiki/List_of_national_capitals
2	Geo-Location data of each national capital city	Data fields include the longitude and latitude coordinates of each national capital. See Appendix II for an example of this data set.	I obtained this data using the Python geocoding web services API.
3	Potential customers' data	Data fields include the venue name, category, longitude and latitude, See Appendix III for an example of this data set.	I obtained this data by exploring the national capitals venues using the Foursquare API
4	The world map GIS data	Data of world map with the national capitals across the world. See Appendix IV for an example if this data set.	I obtained this data using the Folium API

2.2 Data Cleansing

Data of national capitals are scraped from the Wikipedia page using Python. There were some missing data records which I discovered during searching the location data using the national capitals' names extracted from the Wikipedia. After investigation, I discovered that the missing data were due to some comments that were included in the Wikipedia page and put between round parentheses with some of the city names. So, I removed the parentheses and all data within them removed the parentheses and all data within them using the Pandas' based data cleansing module, and then I used the cleansed data to search the locations of the national capitals again. This time I got no missing data. However, I left this cleansing codes code which removes the parentheses and all data within them such that it can be used in future cases, should any update take place on the Wikipedia page.

Also, I have noticed that the column names in the Wikipedia page are not put in standard naming convention. Some column names use special characters, and this jeopardized the Python program code. So, I modified the column name to include only the standard alphabetic character set.

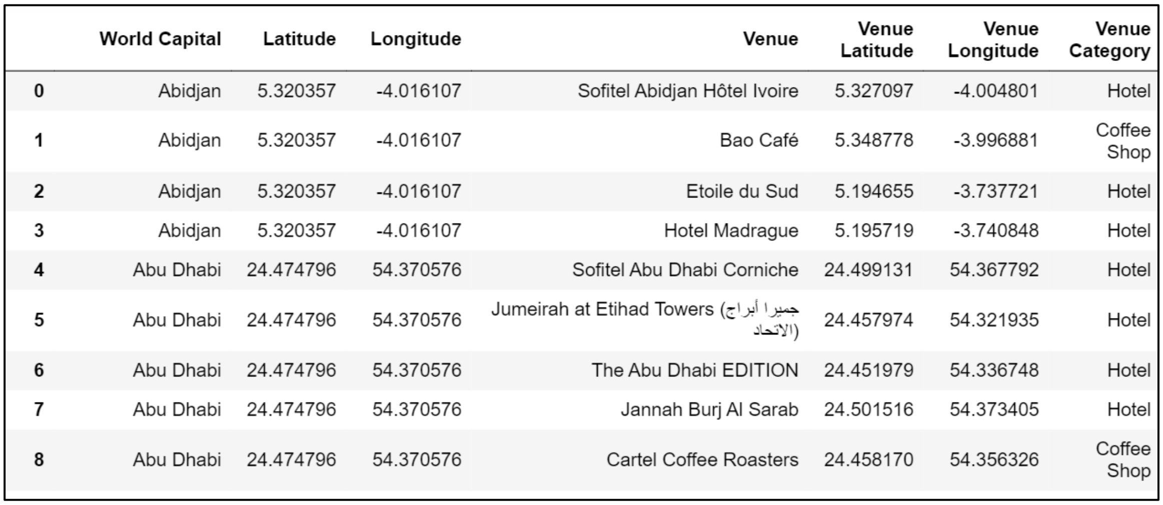
Then I inserted the latitude and longitude coordinate columns structure to the data frame structure of the table read from the Wikipedia page such that I can read the coordinates data from the geocoding web services and include it in the data frame.

I then obtained the national capitals' coordinates data using the geocoding web services. While doing that, I discovered that there are very few missing coordinates data that could not be retrieved by the API. So, I treated this data by displaying exception messages in the data acquisition software module, and then I dropped the rows with NaN values in latitude or longitude fields. This is quite acceptable since these missing data was associated with very few un famous towns. I then combined the venue data with the location data and the master data acquired from the Wikipedia (see Table 3 below).

I then used Folium to create a World Map with all national capitals superimposed on top and used this map to visually verify the correctness of acquired data on the map (see Appendix IV).

Having done all of that, the data quality became quite good and acceptable.

Table 3- Combined Wikipedia data, location data and venue data



	World Capital	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abidjan	5.320357	-4.016107	Sofitel Abidjan Hôtel Ivoire	5.327097	-4.004801	Hotel
1	Abidjan	5.320357	-4.016107	Bao Café	5.348778	-3.996881	Coffee Shop
2	Abidjan	5.320357	-4.016107	Etoile du Sud	5.194655	-3.737721	Hotel
3	Abidjan	5.320357	-4.016107	Hotel Madrague	5.195719	-3.740848	Hotel
4	Abu Dhabi	24.474796	54.370576	Sofitel Abu Dhabi Corniche	24.499131	54.367792	Hotel
5	Abu Dhabi	24.474796	54.370576	Jumeirah at Etihad Towers (جميرا أبراج الاتحاد)	24.457974	54.321935	Hotel
6	Abu Dhabi	24.474796	54.370576	The Abu Dhabi EDITION	24.451979	54.336748	Hotel
7	Abu Dhabi	24.474796	54.370576	Jannah Burj Al Sarab	24.501516	54.373405	Hotel
8	Abu Dhabi	24.474796	54.370576	Cartel Coffee Roasters	24.458170	54.356326	Coffee Shop

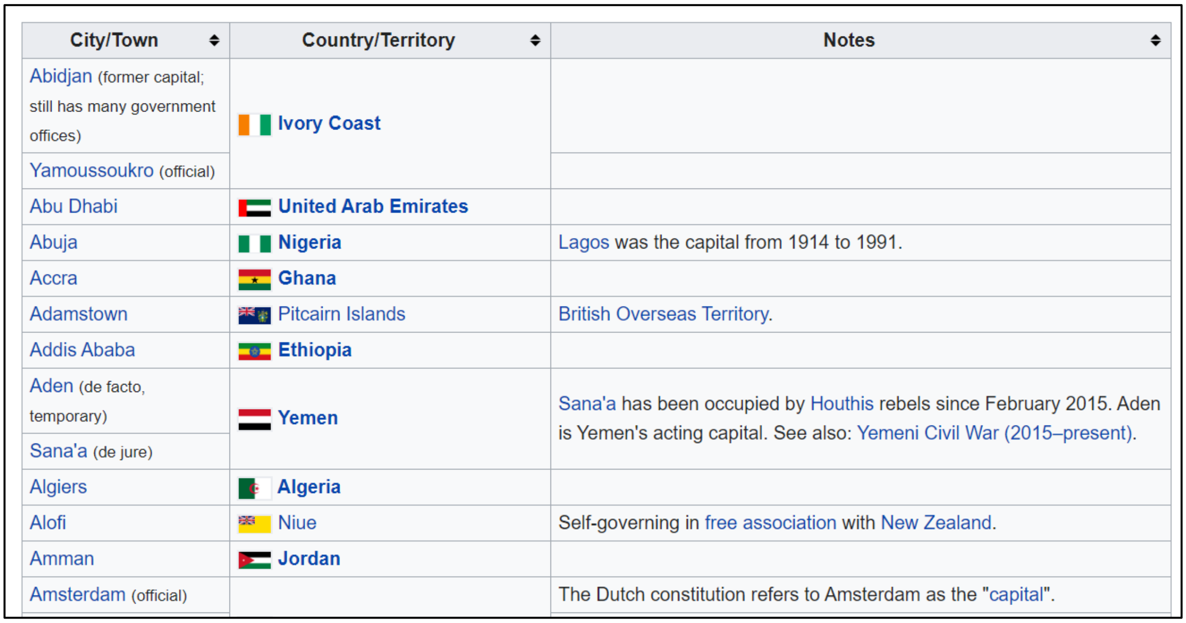
2.3 Feature Selection

After data cleansing, there were 16,702 samples and to know the total number of features (i.e. the number of venue categories of the national capitals), I calculated the number of unique categories curated from all the returned national capital venues. They were 522 unique venue categories, however, in this market segmentation project, we need only three of these features. These are the features marked as ‘Kept’ in the Feature selection Table -4 below:

Table 4. Feature selection during data cleaning

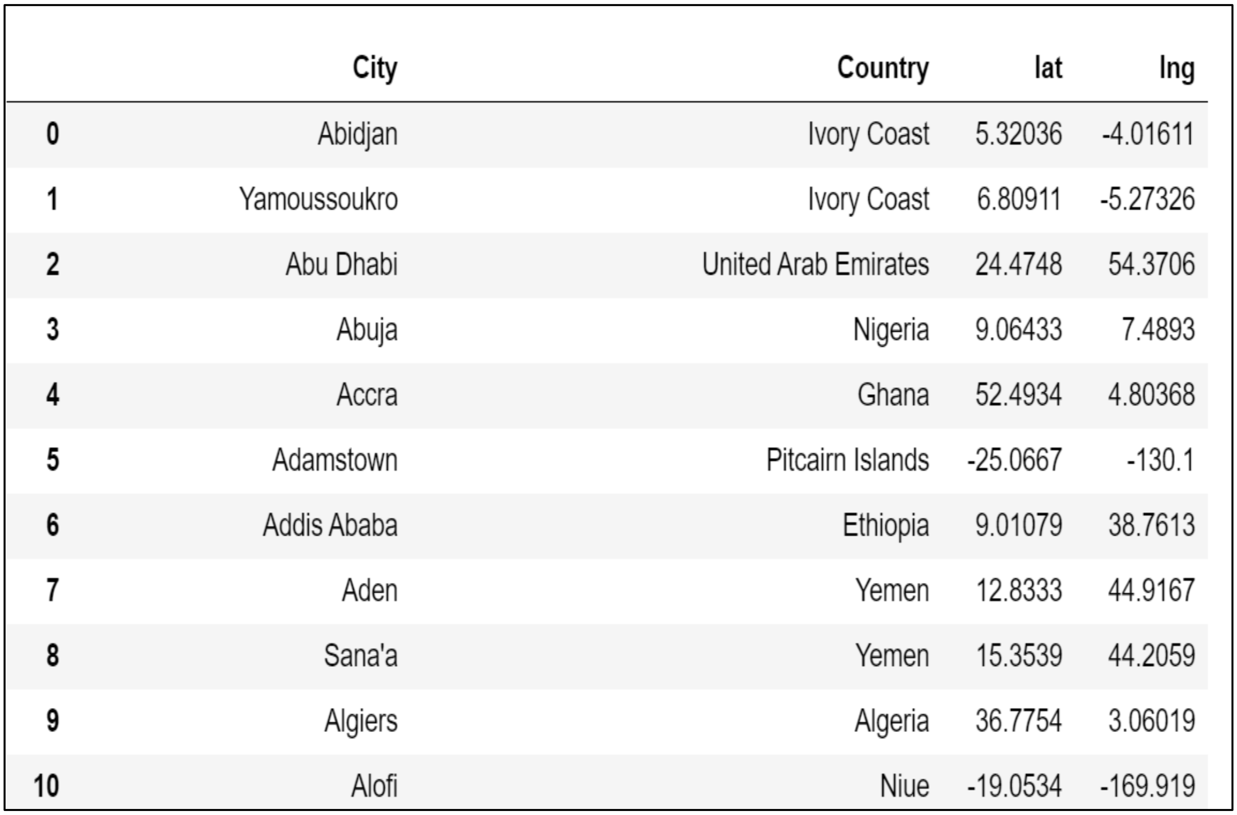
No	Feature	Type of variable	Kept/Dropped	Reason
1	Hotel Category Venue	Categorical	Kept	We need it to build our market segmentation cluster
2	Coffeeshop Category venue	Categorical	Kept	We need it to build our market segmentation cluster
3	Restaurant category venue	Categorical	Kept	We need it to build our market segmentation cluster
4	All other categorical variables such as Auto Workshop, Supplement Shop, Women's Store, etc.	Categorical	Dropped	We do NOT need them to build our market segmentation cluster

Appendix I – Example of data set 1
the Wikipedia List of world-wide national capitals



City/Town	Country/Territory	Notes
Abidjan (former capital; still has many government offices)	Ivory Coast	
Yamoussoukro (official)		
Abu Dhabi	United Arab Emirates	
Abuja	Nigeria	Lagos was the capital from 1914 to 1991.
Accra	Ghana	
Adamstown	Pitcairn Islands	British Overseas Territory.
Addis Ababa	Ethiopia	
Aden (de facto, temporary)	Yemen	Sana'a has been occupied by Houthis rebels since February 2015. Aden is Yemen's acting capital. See also: Yemeni Civil War (2015–present).
Sana'a (de jure)		
Algiers	Algeria	
Alofi	Niue	Self-governing in free association with New Zealand.
Amman	Jordan	
Amsterdam (official)		The Dutch constitution refers to Amsterdam as the "capital".

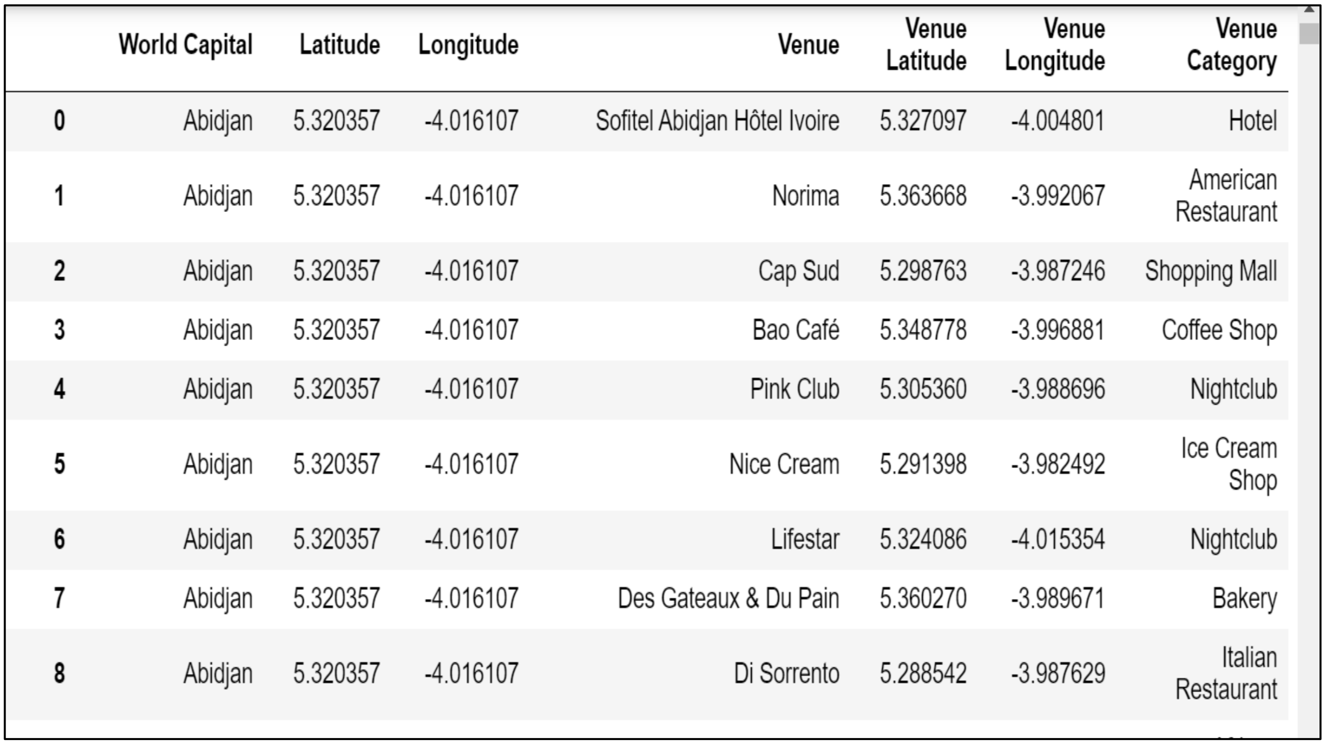
Appendix II – Example of data set 2
Geo-Location data of each national capital from the geocoding web services



	City	Country	lat	lng
0	Abidjan	Ivory Coast	5.32036	-4.01611
1	Yamoussoukro	Ivory Coast	6.80911	-5.27326
2	Abu Dhabi	United Arab Emirates	24.4748	54.3706
3	Abuja	Nigeria	9.06433	7.4893
4	Accra	Ghana	52.4934	4.80368
5	Adamstown	Pitcairn Islands	-25.0667	-130.1
6	Addis Ababa	Ethiopia	9.01079	38.7613
7	Aden	Yemen	12.8333	44.9167
8	Sana'a	Yemen	15.3539	44.2059
9	Algiers	Algeria	36.7754	3.06019
10	Alofi	Niue	-19.0534	-169.919

Appendix III – Example of data set 3

National capitals important venues from Foursquare API.



	World Capital	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abidjan	5.320357	-4.016107	Sofitel Abidjan Hôtel Ivoire	5.327097	-4.004801	Hotel
1	Abidjan	5.320357	-4.016107	Norima	5.363668	-3.992067	American Restaurant
2	Abidjan	5.320357	-4.016107	Cap Sud	5.298763	-3.987246	Shopping Mall
3	Abidjan	5.320357	-4.016107	Bao Café	5.348778	-3.996881	Coffee Shop
4	Abidjan	5.320357	-4.016107	Pink Club	5.305360	-3.988696	Nightclub
5	Abidjan	5.320357	-4.016107	Nice Cream	5.291398	-3.982492	Ice Cream Shop
6	Abidjan	5.320357	-4.016107	Lifestar	5.324086	-4.015354	Nightclub
7	Abidjan	5.320357	-4.016107	Des Gateaux & Du Pain	5.360270	-3.989671	Bakery
8	Abidjan	5.320357	-4.016107	Di Sorrento	5.288542	-3.987629	Italian Restaurant

Appendix IV – Example of data set 4

The world map GIS data from Folium