Clustering postal code areas in Haiti

Grégory PINCHINAT

July 17, 2020

1. Introduction

1.1. Background

The Republic of Haiti is located on the Island of Ayiti, with the Dominican Republic on its right side. As an island, our main and directly exploitable natural resource is our landscapes. So it constitutes a competitive advantage to the Haitian economy. Therefore, providing foreigners as well as locals with tools such as a map with places clustered based on venues that are around may encourage them to visit new places in Haiti, and this will be of a big importance for the whole country.

1.2. Problem

According to the World Bank's data, Haiti's GDP was 8.499 billion dollars[1] in 2019. The overall contribution of tourism to the GDP in 2019 was of about 11.41%, the foreigners have spent a total 970 million dollars for leisure and business combined and the locals have spent around 25% of this amount (Haiti Open Data)[2]. Also, according to Haiti Open Data, since 2015 (the last presidential elections' year), political unrests have made the country less stable, so we have observed a decrease in the number of capital investment in the tourism sector. Entrepreneurs are looking towards our neighbor, the Dominican Republic for their next investments in the Tourism Sector, while locals are still leaving on tourism abroad. While we can't solve the political problems and can't stop everyone from leaving, we can encourage the Haitians who can't afford going abroad to visit local places they believe might be safe, instead.

[1] https://data.worldbank.org/country/haiti

[2] http://opendata.investhaiti.ht/wqldnsf/tourism-in-haiti?action=export#

1.3. Interest

For entrepreneurs of all kind (touristic ones included), the map may reveal places that are already crowded with a given type of business and inspire for new ideas of investment. For consumers of all kinds (especially local or foreign travelers), the map may serve as a guide to know which venues they may find in a radius of ten 10 kilometers around the area where they are headed to or where they already are.

2. Data

2.1. Sources

The Data needed for clustering neighborhoods in Haiti Postal Codes, Postal Codes' areas along with their administrative boroughs, which were scraped from Wikipedia[3]. Geographical coordinates were requested and obtained from various open sources such as OpenStreet Maps[4] and Google Maps[5]. And finally, venues data were obtained from Foursquare[6].

2.2. Cleaning

The data scraped from an html-format was put in a dataframe, based on the boroughs (Department) and Postal Codes. Then, the geographical coordinates (Nei_lat, Nei_lng) were added as two new columns in the dataframe. There were mistakes in the names of some neighborhoods which kept from getting the right coordinates. Said mistakes were corrected and the missing coordinates were searched for and appended manually.

2.3. Feature Selection

The Postal Codes only served as a means to sample neighborhoods from the Haitian Territory. Many different neighborhoods may share a common postal code, so we had to split cells containing more than one neighborhood for just one postal code

[3] https://fr.wikipedia.org/wiki/Code_postal_en_Ha%C3%AFti

[4] www.openstreetmap.org

[5] https://www.google.com/maps

[6] https://foursquare.com/

into more rows, in which the same data but the neighborhood cell name. Nonetheless, the feature of interest was the Neighborhood column, which contains a list of all neighborhoods based on their respective postal codes and borough (Department).

3. Methodology

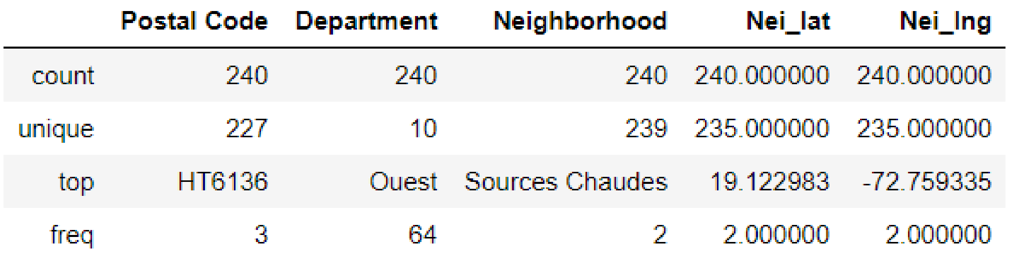
3.1. Technology and Infrastructure

Python 3 was used to carry all the analyses in a jupyter notebook. The geographical data were mainly obtained thru the "Geopy" package. A loop was used to get them automatically with exception handling, but 25 neighborhoods still had missing geographical data after this process. Said coordinates were searched and appended manually. For further information about changes made on the dataset for purely technical purpose, the reader may refer to the jupyter notebook given as an annex.

3.2. Exploratory Data Analysis

3.2.1. Neighborhoods

After the wrangling, the table was left with 240 entries, of which no null nor missing values were to be accounted for. 227 unique Postal Codes and 10 Unique Departments exist according to the administrative division of the State of Haiti.



	Postal Code	Department	Neighborhood	Nei_lat	Nei_lng
count	240	240	240	240.000000	240.000000
unique	227	10	239	235.000000	235.000000
top	HT6136	Ouest	Sources Chaudes	19.122983	-72.759335
freq	3	64	2	2.000000	2.000000

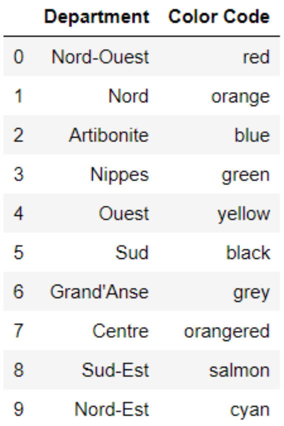
One of the most frequent Postal code is that of HT6136 which corresponds to *Lamentin, Mariani and Merger* (three neighborhoods in the commune of "**Carrefour**", Department "**Ouest**"). The Department with the most neighborhoods recognized by the Haitian Postal Services is that of the Capital: the Department "**Ouest**". There should also be 240 unique neighborhoods, but only 239 are taken into account because two different neighborhoods with the same name and different geographical data have been found in the data.

3.2.2. Venues

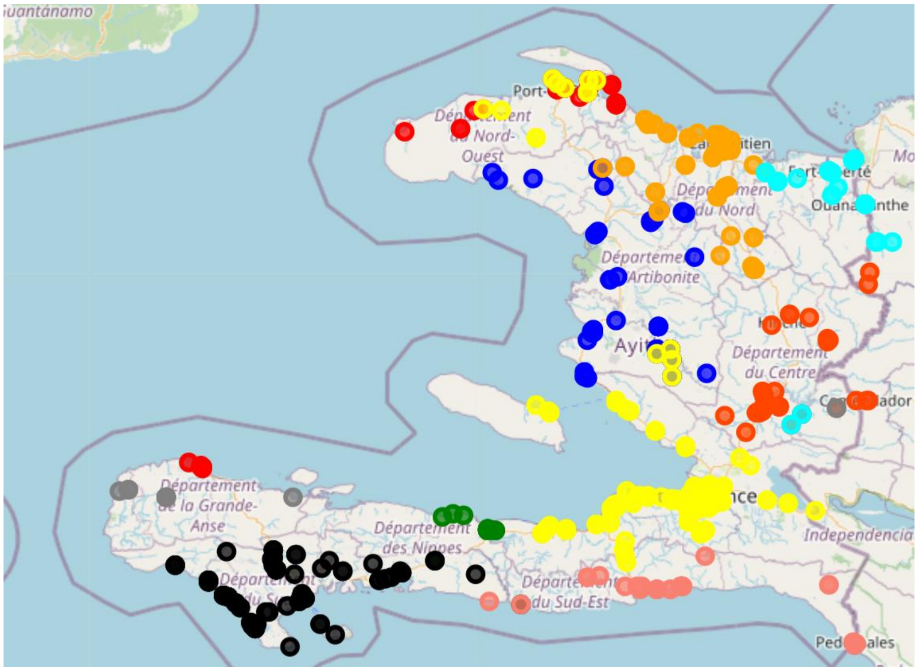
In Haiti, Foreign Direct Investments from 2008 to 2018 have accounted for less than 4%, on a yearly basis[7]. As of 2014, only 5% of all businesses[8] are formal. Internet penetration rate stands about 33% of the population[9] in 2020.

All of the above might explain why not many Haitian businesses can be found on Foursquare. The segmentation of the territory by postal codes might be biased as well, because of some areas being way larger than others, while having very few businesses listed.

At first, a radius of 1 kilometers was chosen, then of 1.5, then of 5, but less than 500 venues were returned. Finally, a radius of 10 kilometers around the neighborhood's central coordinates were explored in order to get a total of 4260 venues. Here is a graphical overview of all the venues grouped by department:



	Department	Color Code
0	Nord-Ouest	red
1	Nord	orange
2	Artibonite	blue
3	Nippes	green
4	Ouest	yellow
5	Sud	black
6	Grand'Anse	grey
7	Centre	orangered
8	Sud-Est	salmon
9	Nord-Est	cyan



Graph 1.- Map of venues in Haiti (10 km around neighborhoods)

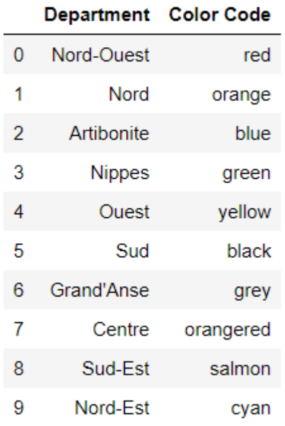
[7] https://tradingeconomics.com/haiti/foreign-direct-investment-net-inflows-percent-of-gdp-wb-data.html

[8] Private Sector Assessment Report of Haiti, Inter-American Development Bank, 2014

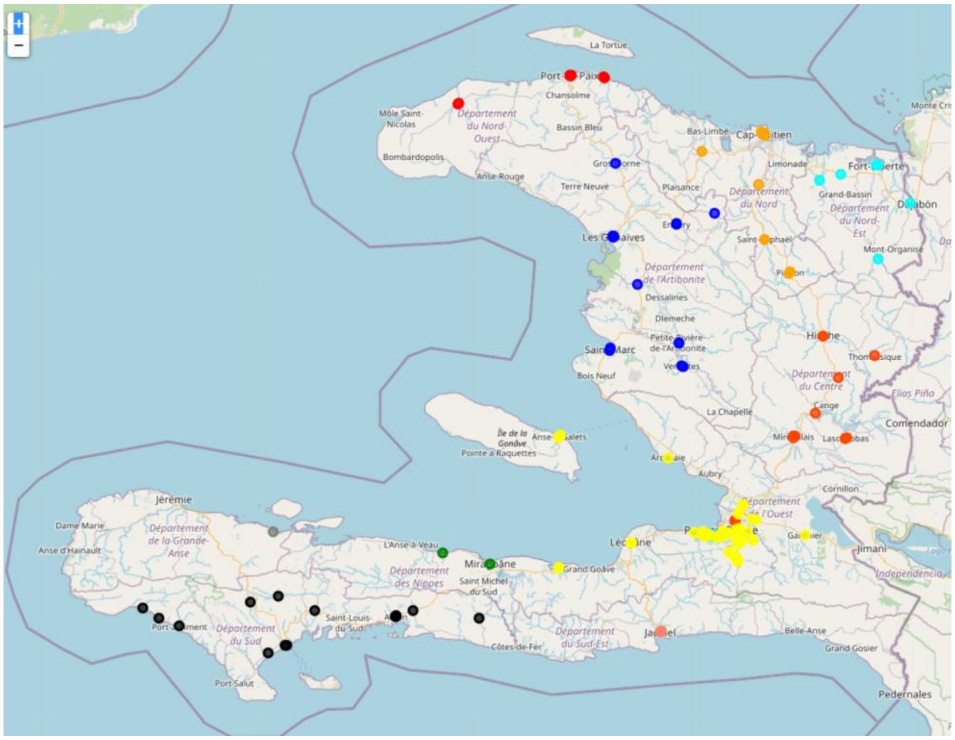
[9] https://datareportal.com/reports/digital-2020-haiti

Such a large radius may have returned more venues, but most venues were not in the vicinity of their respective neighborhood at all. For example the yellow markers are supposed to be at the proximity of neighborhoods in the Department Ouest only, but some of them fell all the way up on the North region. And while red dots are supposed to remain in the North-West of the country, some were found on the South-Western region.

For remaining realistic, a radius of 500 meters was then imposed around each neighborhoods, which is close enough the neighborhood as well with a means of transportation as on foot. This time we obtained only 189 venues. The new overview then rendered this way:



	Department	Color Code
0	Nord-Ouest	red
1	Nord	orange
2	Artibonite	blue
3	Nippes	green
4	Ouest	yellow
5	Sud	black
6	Grand'Anse	grey
7	Centre	orangered
8	Sud-Est	salmon
9	Nord-Est	cyan



Graph 2.- Map of venues in Haiti (500 meters around neighborhoods)

As can be seen on the map above, all neighborhoods remained inside their respective departments[10]. This is the data that were finally used to cluster the neighborhoods.

3.3. K-Means Clustering of Neighborhoods

78 different categories were accounted for in the venues dataset, which were clustered in 5 superior categories using K-Means Clustering Methodology. For doing so, some quantitative variables were needed for computing the dissimilarities. The frequencies of venues per respective categories around the neighborhoods were then used for this computation.

The more a neighborhood is likely to have the same categories of venue around it as another neighborhood, the more both neighborhoods would be similar, and the more likely the algorithm would be to place them in the same cluster. The centroids have been initialized at random[11] and the algorithm has been run 20 times, after which the best result would be kept.

4. Results

After the clustering completed, a new table was created with all the previous data (apart from the postal codes and the categories), including the cluster labels (0 thru 5) and the 10 most frequent venue categories around each neighborhood. This table can be subset for each cluster label, and the neighborhoods can be represented by icon markers on a map.

4.1. Cluster 1 (label 0) : NIGHT CLUBS

This cluster is represented by the icon NIGHT CLUB and groups only night clubs.

4.2. Cluster 2 (label 1) : BARS & BOUTIQUES

[10] The dots on the map are not unique, most of them are actually a ‘set’ of points representing venues around several close neighborhoods. A zoom allows to view the map with more precision. An html file is attached so the user may either print the map on a greater format in order to view the neighborhoods more distinctly or to zoom in from a browser and view more details.

[11] The random state has been fixed to the value of 0 (zero) for reproducibility.

This cluster is represented by the icon and groups mostly bar, bar restaurants, shops and other convenience stores.

4.3. Cluster 3 (label 2) : CREOLE RESTAURANTS

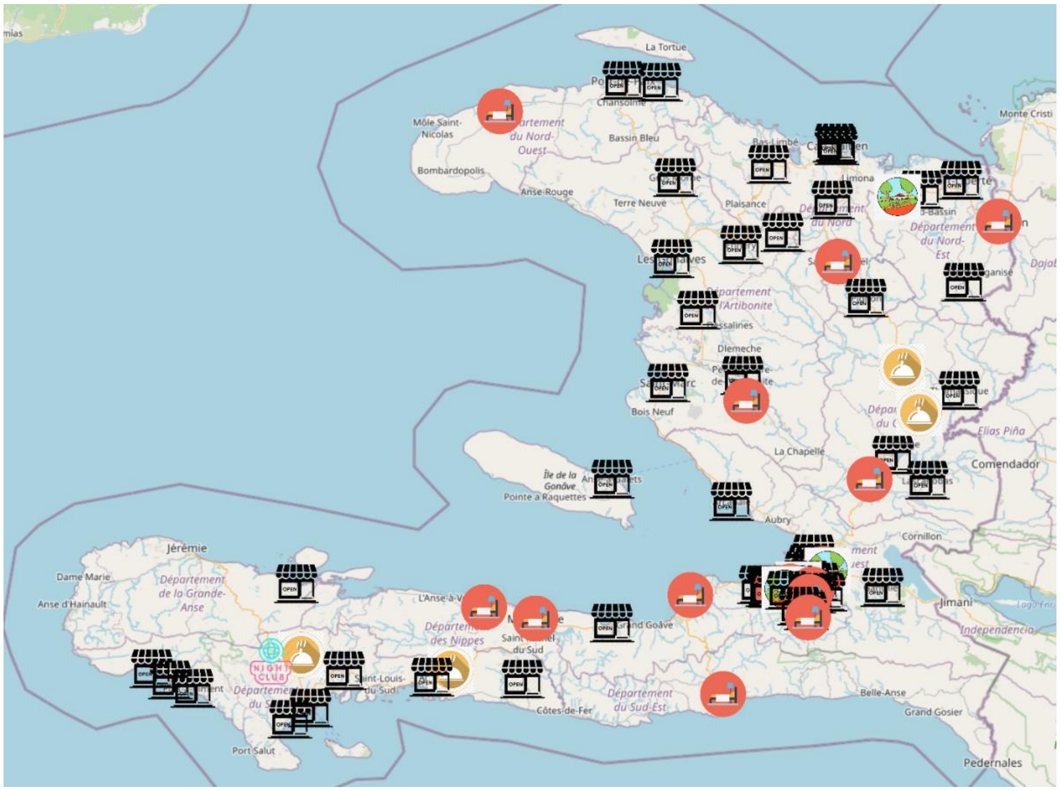
This cluster is represented by the icon and groups mostly cajun cuisine or creole cuisine restaurants.

4.4. Cluster 4 (label 3) : HOSTING PLACES

This cluster is represented by the icon and groups mostly restaurants, diners, hotels, hotels that have restaurant.

4.5. Cluster 5 (label 4) : PARKS

This cluster is represented by the icon and groups mostly parks and other public open-air meeting spots.



Graph 3.- Map of clustered postal code areas in Haiti (based on most frequent venues in a radius of 500 km)

5. Discussion

The first two things that cannot go unnoticed on the map, is the only night club cluster on the map and the variety of clusters that characterize the neighborhoods around the capital. The clusters, although not completely dissimilar, remain distinguishable a priori.

So bars and boutiques are fairly distributed around the territory, showing a preference for small businesses, while hosting places are found above all in coastal regions (most touristic places on an island, obviously).

Creole restaurants are remarkable in two Departments: Centre and Sud. And finally, most parks are found in the metropolitan area of Port-au-Prince.

6. Conclusion

The objective of this study was to cluster neighborhoods in Haiti in order to provide a tool to all kinds of users based on their interests (local tourism and investment on leisure or business). The neighborhoods were chosen by their attached postal codes, and neighborhoods with the same postal codes were taken individually. Using a Foursquare Developer API access, venues in a radius of 500 meters around each of these neighborhoods were explored, then grouped by the similarity of the venues' categories with K-Means Clustering Methodology, yielding 5 individual clusters. The algorithm did not completely converge, a lot of null values were returned. Those null values were dropped and only neighborhoods assigned to a given cluster were kept for the creation of the tool. This study has limits: it is made in a country where technology is not really used, nor advanced, where 95% of businesses are not formal or don't have any digital imprint, and where there very few assets invested. This map may be quite useful for all planning purposes, but may not be representative of larger (greater than one-kilometer-wide) postal areas' neighborhoods in Haiti.