**Clustering Banks around Panamá City**

Leonel Castillo

**Introduction:**

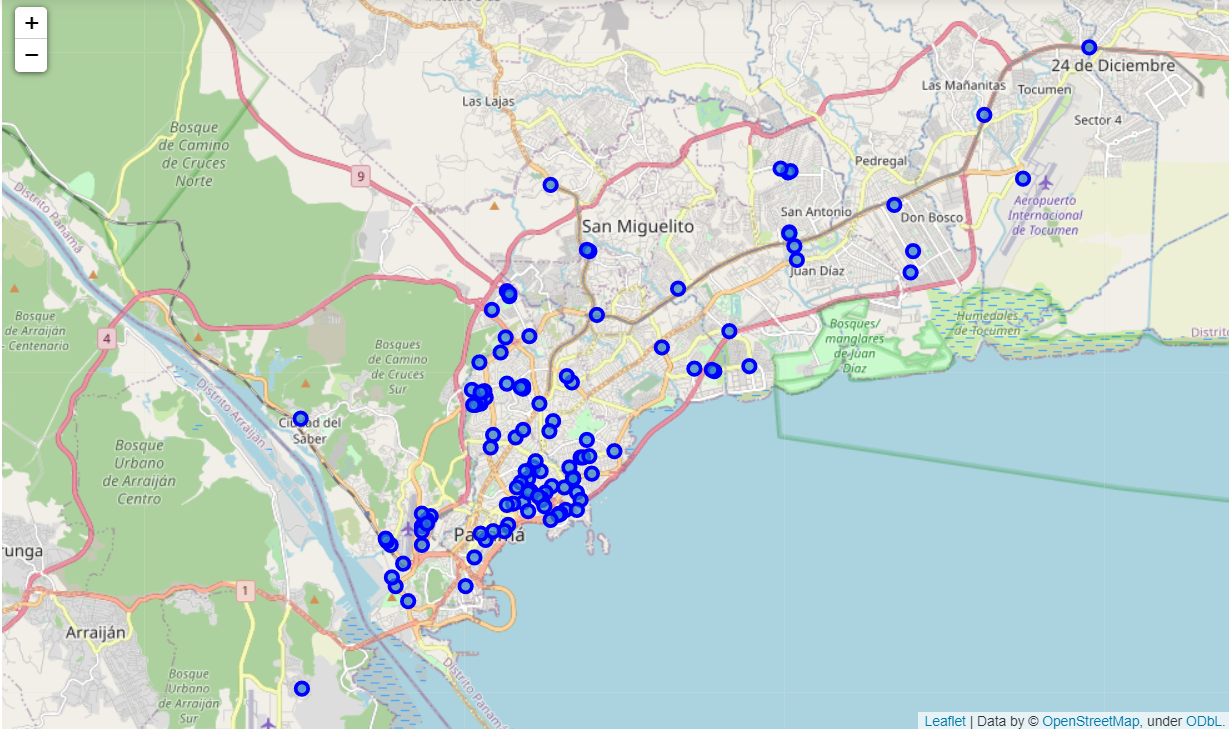
Panama City is well-known for its banking sector and many national and international banks have branches operating in the city competing for local business. Given this it is natural to consider the problem of how banks are distributed around the city and where should one open their first (or a new) branch given the distribution of business and population in the city but also factoring in what other banks are already doing.

This problem is formalized by considering the explicit problem of clustering banks based on the businesses around them. Beyond this we also have a goal of ensuring that the resulting clusters are explainable in simple terms so that the results of this analysis also include a deeper understanding of the local banking sector.

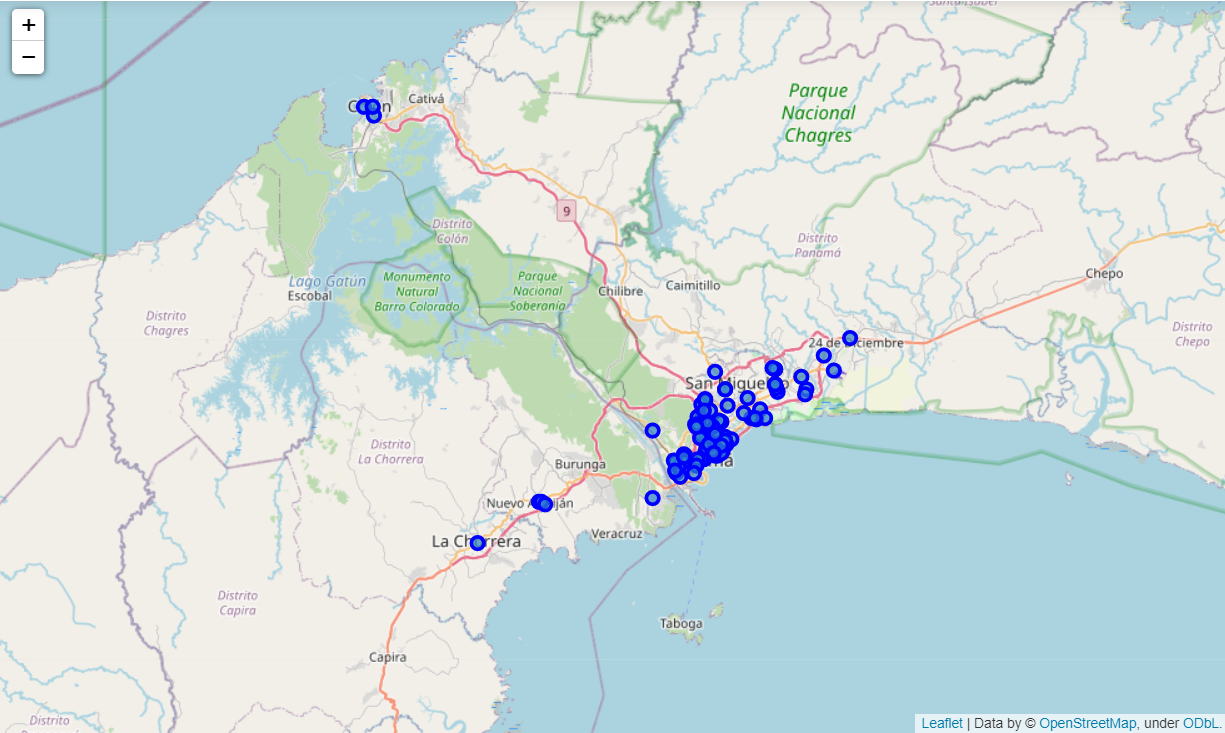
**Data:**

We will be using Foursquare’s API to find a list of all the local banks operating in and around the city and also to find a list of local businesses that are near these banks.

Zoomed in image of the banks plotted in folium:

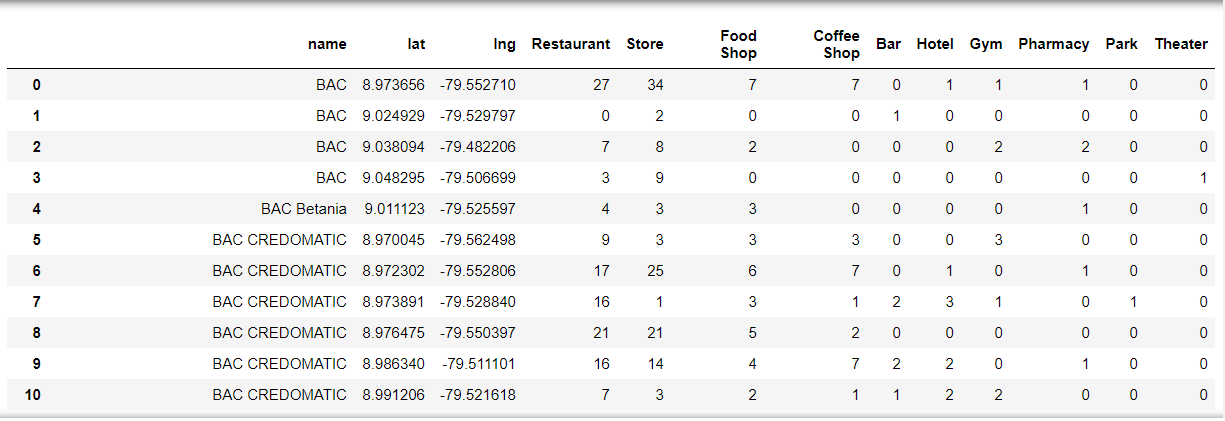


Zoomed out image of the map:



For this study we also decided to include farther away banks to observe in which clusters they would end up in.

The business data was compiled in a table like:



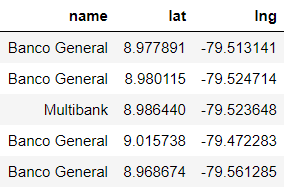
Where we can observe each banking branch and the number of certain types of businesses around them. Please notice that names repeat themselves, but the pairs of latitude and longitude are unique. This happens because many banks have many branches around the city, with some banks like Banco General having over 60 branches around the entire country. Also, the list of types of businesses was curated for the specific purpose of building an explainable model. Therefore, there are many categories that you would see on Foursquare but that were either merged inside a bigger category or removed from the study.

**Methodology:**

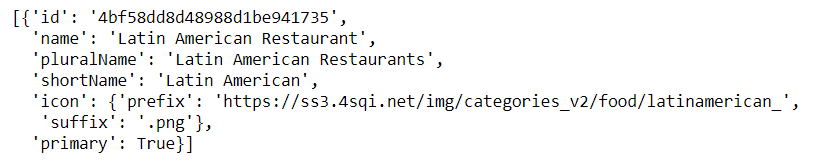
We used the Foursquare API to extract a list of the banks operating in the city. We used multiple scripts like:



To compile the list. We had to use multiple calls because we wanted our findings to be centered around different spots in the country so that we would also have banks that are not necessarily in the center of the city so that they could complement our results. After obtaining a list we were satisfied with we converted it into a dataframe with the name of the bank and the latitude and longitude pairs:



Again I stress that the unique pairs are the latitude and longitude as most banks have multiple branches around the city. After compiling this lists I used the Foursquare API to find a list of all businesses around each bank and specifically extracted the category. In this stage we did some exploratory data analysis because Foursquare categories include very specific names like “Latin American Restaurant” but for this analysis we did not want to have one variable for each type of restaurant.



Therefore I grouped many categories together until I arrived at the following list:



So a list of over 300 categories given by Foursquare was narrowed down to just 10 variables. It must be noted that this process was more art than science and was not guided by any mathematical principles. In other words the number 10 is arbitrary but it was guided by my goal of finding an explainable model. It would be possible to increase the accuracy of a model by adding more variables and breaking up some variables. One way I considered but did not pursue was breaking down the ‘Restaurant’ variable into ‘High-end’ and ‘Low-end’ restaurants so that the variables also had a wealth component. However the analysis required to properly do this is beyond the scope of this study.

After properly creating the 10 variables as the number of businesses around the bank of that category I decided to scale the data using the StandardScaler:



This is a tool that transforms all of my variables into variables with 0 mean and variance 1. We do this so that the scale of some variables do not give those variables a higher weight than they should have. This is important because in a clustering model the distance is a huge component and if a variable is bigger it can have more influence over the distance. But if we normalize it properly the risk of this happening is greatly reduced.

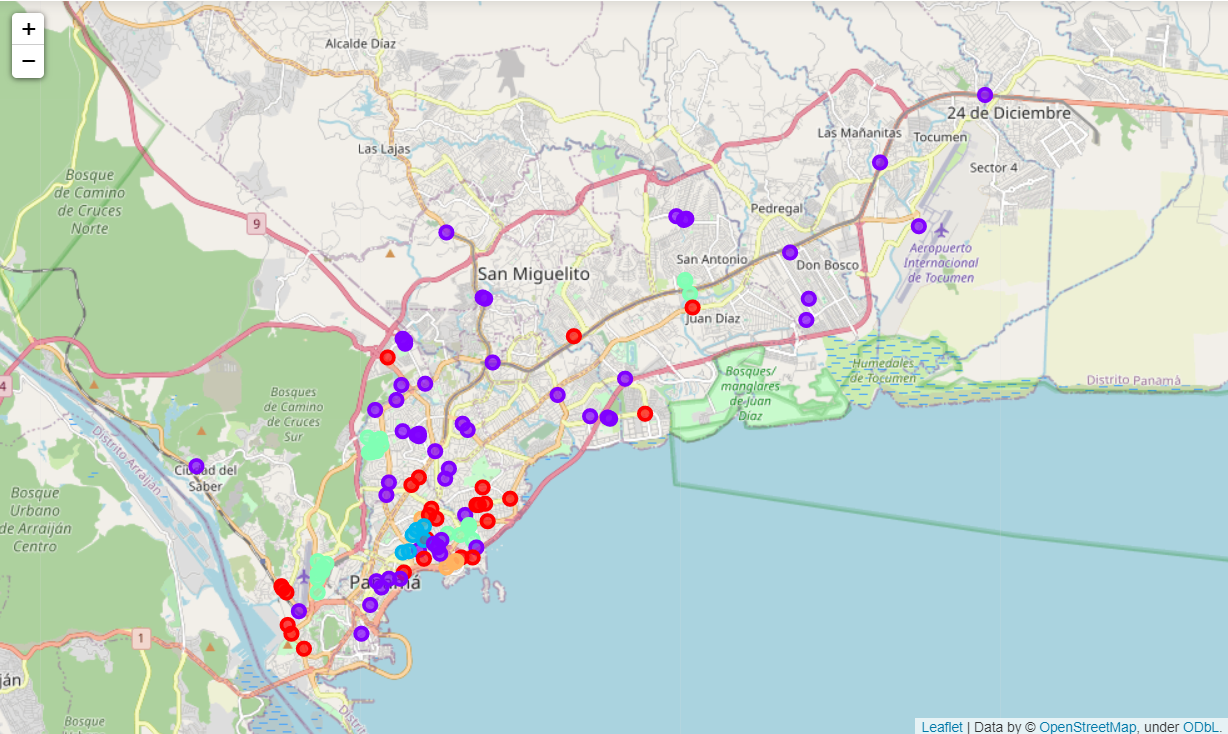
Finally, we decided to train a K-means model with K=5. We chose K-means because it is a very versatile model that is well-known to work perfectly for this kind of problem. I chose K=5 because from looking at the data I understood that in general these are the clusters I could expect:

1. ‘Central’ banks. I.e. banks deep within the financial areas of the city where every bank is expected to have a branch.
2. Remote banks i.e. banks that are opened not due to a great demand for banking but due to there being a significant enough presence of customers.
3. Special banks, i.e. banks that are placed around really important or popular spots around the city.

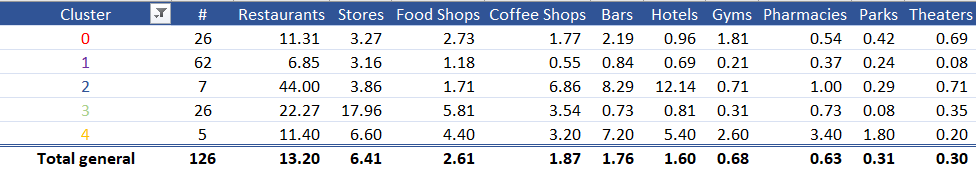
And so that the model had versatility I decided to add +2 to this list because there could be banking clusters that I could not predict on my own. But this initial analysis tells us that the number of clusters you want is at least 3, and not many more because the types of banking needs in a city are limited.

**Results:**

After processing the data, the clusters look like this:



And the mean of the variables looks like this:

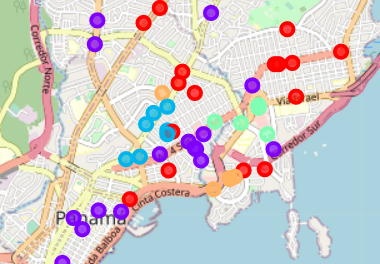


(the color on the table matches the color on the map).

**Discussion:**

After looking at the previous table and map I came up with the following profile for each cluster:

1. These are middle-ground banks. They are located in good strategic locations for bank placement but are not the best areas. A bank should only look to open a branch in one of these sectors if they already have a dominant presence in the area.
2. These are outreach banks. They are placed in below-average locations not necessarily for the benefit that that particular location brings in but to make their customer’s lives a little more convenient. This can be seen by how there are some banks in this cluster really far away from the city: 
3. The ultra-connected areas. These are the best placed banks in the whole region with the most access to local business. You can observe in the map that they are just one tiny cluster made of 7 banks in the middle of the city:



A new bank looking to open their first branch should definitely place their first branch around here.

1. Banks near stores. If you look at the table you’ll notice that this cluster has every other cluster beat, even the ultra-connected, in the stores category. This category includes many types of stores and this means that these banks are around dense shopping centers with plenty of businesses around them. Definitely a great second choice of placement.
2. General business banks. If you look at the table, you can notice that this cluster has a higher than average frequency of businesses in almost all categories. Therefore these are banks that are around a very varied selection of businesses. Likely around residential areas.

Using these results my advice would be that any bank looking to open their first branch must do so in the blue areas. This is the best region to start your local business. However, only one branch should be opened here and act as the main branch. As the bank expands it should start opening branches in the green and orange regions to capture a more varied set of customers but also in great spots. After the third of fourth branch the red areas should also be considered. And finally, no branches should opened on the purple areas unless the bank is already so dominant in the region that it can start thinking about opening more remote operations to better serve smaller sectors of their customers.