Final Capstone Project – Week 2

1. Introduction

A friend of mine (Our client) is the manager of a Gym Franchise in the city of Belém – PA, Brazil. He's looking for a good place at the city center to open a new Gym and expand his business, but has problems finding which neighborhood would be a good one for that.

1.1 Business Problem

Let's say you want to open a new business and need to find a place for that. Which neighborhood would be adequated to open your store? This project aims to explore the venues of the city center of Belém – PA, Brazil, and discover places that would be good to open the new Gym for our client, preferably places with as few gyms as possible.

1.2 Possible Stakeholders (Interest)

This project will focus on Gyms for the client, but we'll also explore the most common venues of the city. It's possible to find good places to open new restaurants, pubs, pharmacy, and many others kind of venues through a neighborhood. So any businessman looking for places to open new business, or expand their existing business with new stores would be interested.

2. Data

2.1 Data acquisition

This project will use the data from this wikipedia page to get the neighborhood of the center city to explore its venues. There's only 8 neighborhoods around the city center and no data about their coordinates, so we will be using the library geopy to get the right coordinates and continue our exploration.

2.2 Describing the Data

The BeautifulSoup library was used to extract the information of the necessary neighborhood from the wikipedia page. The data extracted was put into a DataFrame using the Pandas library.

With the data in hand the geopy library was used to find the coordinates (Latitude and Longitude) for each neighborhood. There was a problem here with the geopy returning

the wrong coordinates for 2 neighborhoods: "Nazaré" and "Campina"; so it was decided to use the coordinates of the main spots for these two neighborhoods: "Basílica" and "Praça da República" respectively.

The Foursquare API was used to search for the venues and the folium library was used to create the maps.

3. Methodology

I used the Jupyter Notebook to create the notebook used to explore this project. With the coordinates extracted, the map was created using the folium library (figure 1).
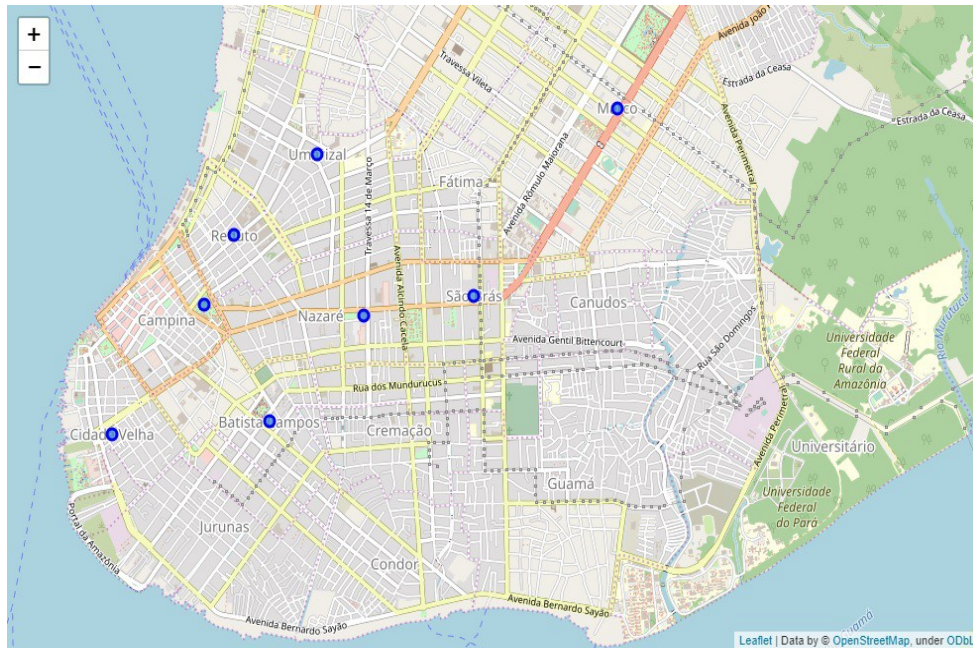

Figure 1. Map of Belém with the Center Neighborhoods

The Foursquare API was used to search for the venues in a 600m radius from the coordinates of the neighborhoods. There was a total of 596 venues with 156 unique categories distributed among the 8 neighborhoods.

With the data on hand, we decided to create a new DataFrame grouping the rows by Neighborhood and finding the mean of frequency for each venue category. This enables for further research letting us classify the top venues of each neighborhood and use the K-Mean algorithm to cluster them.

A sum of all columns related to Gyms was made to create a bar plot to analyze the situation of Gyms venues across these neighborhoods (Figure 2).
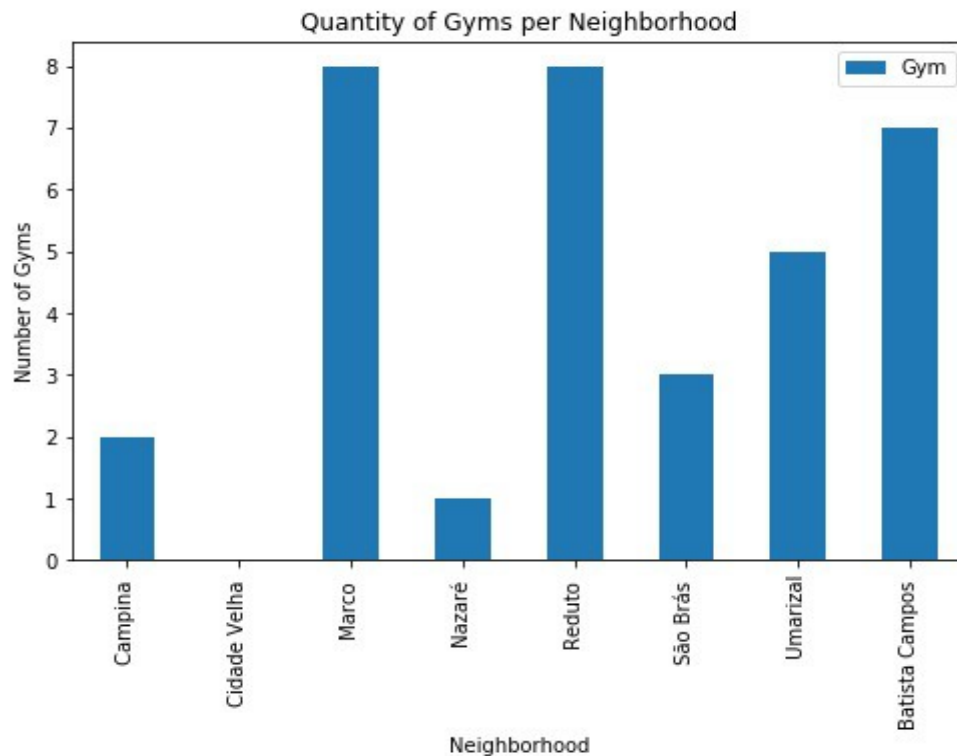
Figure 2. Bar Plot of Gyms per Neighborhood

The bar plot shows that "Cidade Velha" has no Gym, so that would be a good place to set up a new store; "Nazaré" and "Campina" also have veey few gyms around, so they are good places to set up the business. The other neighborhoods already have a good amount of gyms so they are not recommended to open new ones, since the possible customers would already be client of the various existing ones.

3.1 Clustering Neighborhoods

We used the K-Mean algorithm to segment the neighborhood and find similarities based on the most common venues. It was defined 5 clusters to partition the data into groups of similar categories and create a map based on it (Figure 3). Through these neighborhoods we can see that restaurants and pubs are predominant in almost all of them.

- "Campina" is a neighborhood with lots of tourists attraction;
- "Reduto", "São Brás" and "Umarizal" are more inclined to pubs and different kinds of restaurants;
- "Cidade Velha" is a good place for the local cuisine;
- "Nazaré" has lots of general stores and restaurants
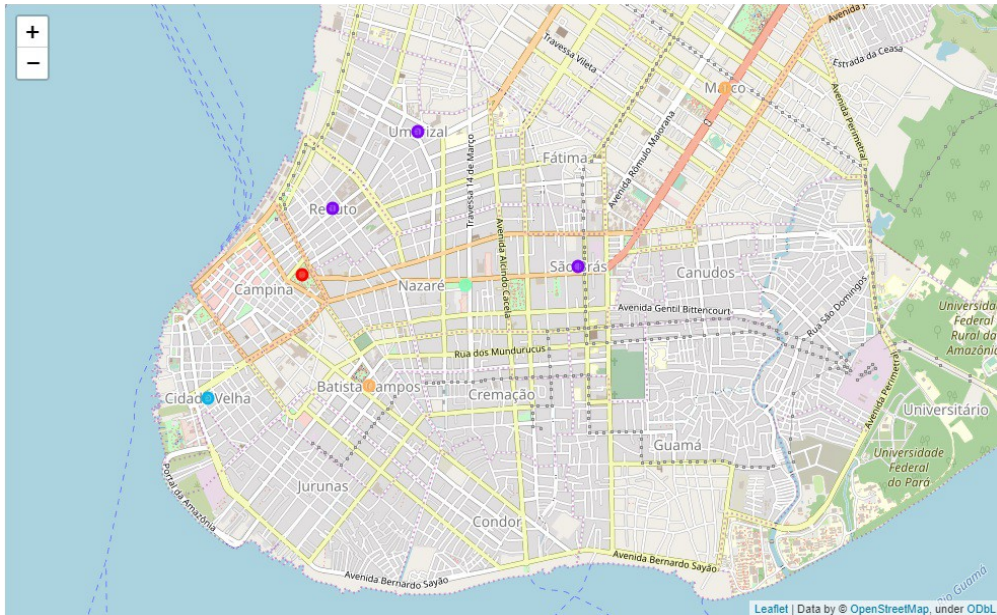- "Batista Campos" and "Marco" are aggregated by Gyms, Pharmacies and local restaurants.

Figure 3. Clustered Map

4. Discussion

Belém is not a big city compared to São Paulo or Rio de Janeiro, so the center of the city is not that big, but it has a good population density on these neighborhoods. Some people tend to go to squares on the city to jog/train, but the neighborhood "Cidade Velha" doesn't have a good square for people to do this. The possibility of a new Gym in the neighborhood is a good chance to attract customer and stabilish the only Gym in the vicinity.

The most common venues were also explored on these neighborhood and, on a further analysis, we can see that restaurants and pubs have a big presence on all of them, so opening these kind of venues are not recommended in the center of the city.

5. Conclusion

Python is a very powerful tool for data analysis. This project just showed a small amount of what it's capable of. We found the neighborhoods of center of Belém, its coordinates and explored the data providing the results. Businessmans could use these results in a good way to choose the best places to open their next store or office.