Investigating similarities within city zones using Foursquare API data

Otoniel de Lima Filho

January 2020

1. Introduction

1.1. City of Osasco

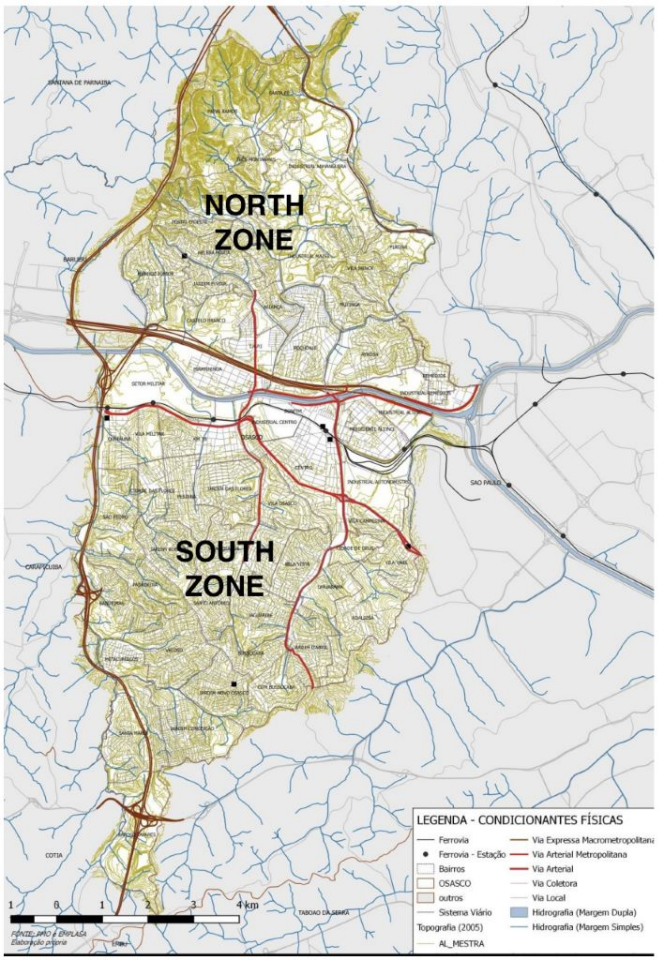
As decades go by, metropolitan areas have been increasing and merging cities whole around [1], which are common options for cheaper housing and services in comparison to the big capital cities.

The scope of our work is one of the cities around São Paulo, Brazil.
Osasco, founded in 1962, has 700000 inhabitants, and has already the 5th largest population in its state, and it has the 9th domestic gross product in the country, placing it among the richest cities in Brazil [2].

It is home for Brazil's largest private bank headquarters, two national television broadcasters and large service and e-commerce companies.

Although its economy is in good shape, the city has been struggling for decades with severe social problems, such as extreme poverty, high crime rate [3], as the city is currently ranked 4th most violent city in Brazil, unemployment around 15% and with disorderly growth.

Such problems have been reflecting in the current city layout, a wealthy south zone with shopping malls, hypermarkets, hospitals and plenty of transportation options, a poor north zone with large illegal land occupations, flood risk etc. While the south is home for new businesses every year, the north tends to be let behind.



1.2. Goals of this project

We want to use Data Sciene tools to display such different realities and also try to find similarities that would explain how the population manage to offer products and services to their neighbors even if the conditions are not ideal.

2. Data Acquisition

2.1. Foursquare API

This analysis will use the Foursquare API database to find many sorts of venues of retail, services and public utilities, with its 'Explore' endpoint.

As the search on Foursquare limits results to 50, we created functions to automate the queries and expanded the base by searching for venues of different categories, maximizing the results.

As Foursquare has a large variety of subcategories, we reduced the results to major categories, and thouse have been conviniently renamed to those groups.

2.2. Geopy

All queries have been performed within a radius of 10km from the center obtained by Geopy Geolocator. After collecting the data, we removed venues not registered exactly in Osasco and prepared the dataset of the valid venues

3. Methodology

3.1. Application of K-Means clustering

Clustering is na important part of the analysis, once the municipality of Osasco does not keep public data for neighborhood border coordinates or such, we will perform K-means clustering in order to find the closest delimitation of the North and South zones. For our reference, the zones are divided by the Tietê river that crosses the city from east to west. Hence we were able to find such division with 10 clusters (3 in North, 7 in South).

3.2. Libraries in use

We will use Python libraries such as Folium to display the data in maps and perform visual analysis. For K-Means clustering we use Scikit Learn library.

3.3. Venues classification

In a preliminary analysis we want to know our 10 clusters by listing the 10 most common venues and check for the distribution of some venues picked by our discretion.

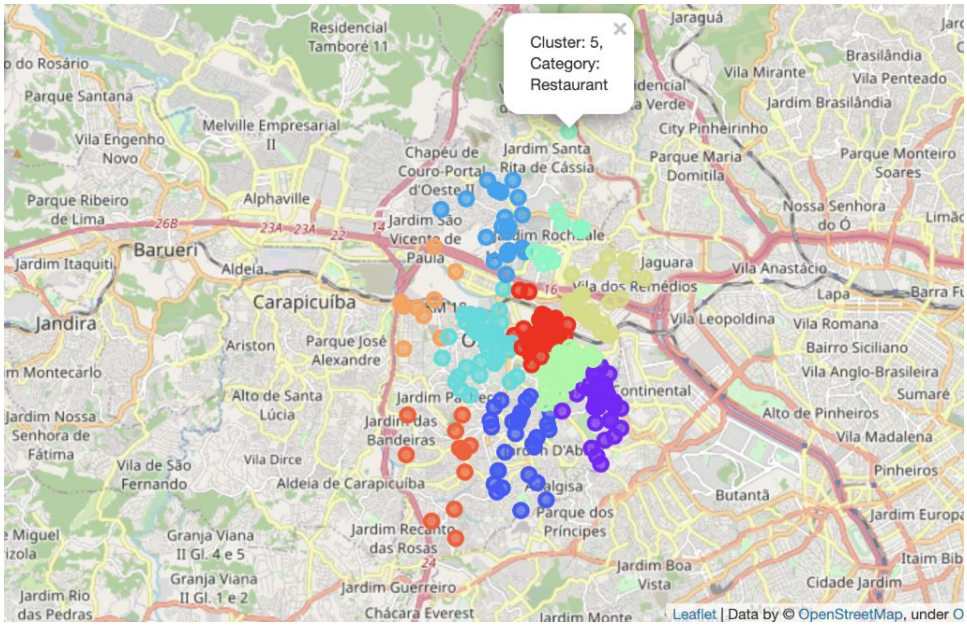
3.4. Similarity index

With the consolidated data from the 10 clusters we now group them into North and South zones and start the analysis of the results.

A similarity index is computed comparing the top 10 venues in both zones and counting the common spots.

4. Results

4.1. K-Means clustering in a map

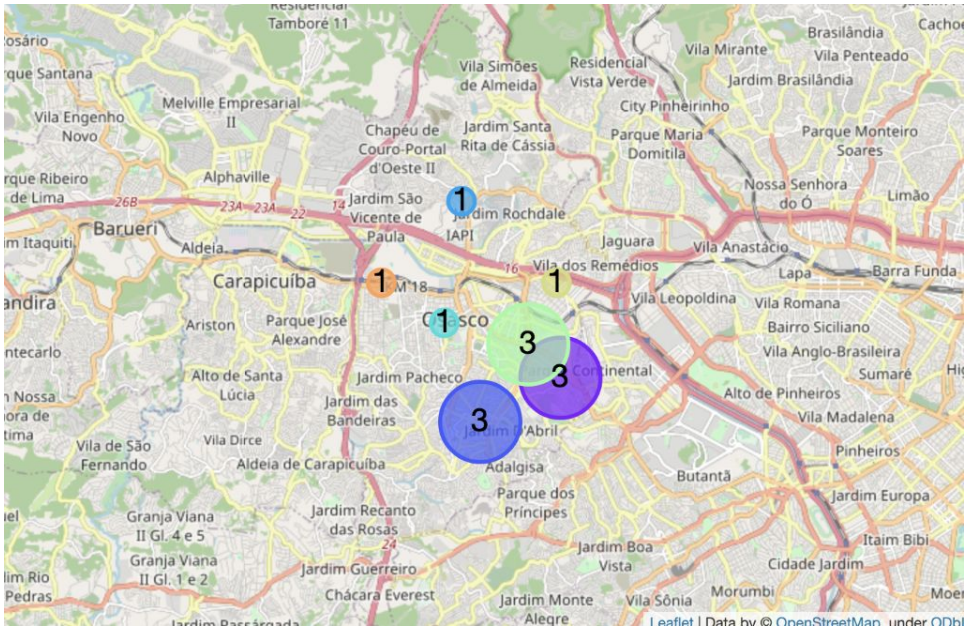


4.2. Top 10 venues for the 10 clusters

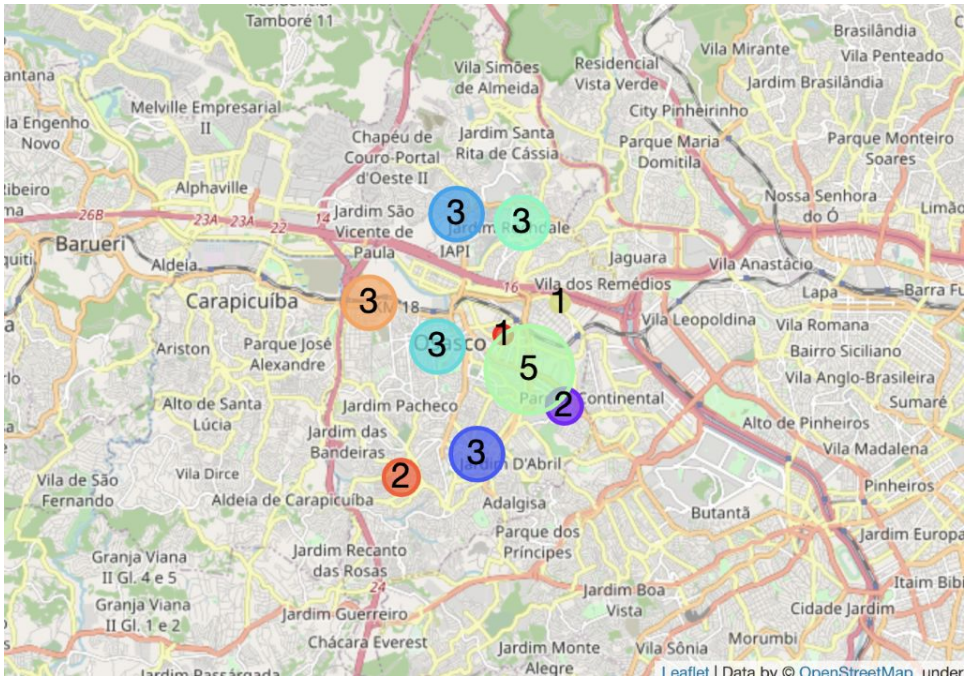
	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Shop	Mall	Hospital	Drugstore	Police	Restaurant	Gas Station	School	Church	Post Office
1	Restaurant	Nightlife	Shop	Recreation	College	Gym	School	Supermarket	Gas Station	Mall
2	Drugstore	Gas Station	Gym	Church	Restaurant	Supermarket	Recreation	Pet Shop	Shop	Police
3	School	Bus Station	Nightlife	Supermarket	Police	Post Office	Hospital	Gym	Drugstore	Stadium
4	Nightlife	Restaurant	Recreation	Gas Station	Supermarket	Post Office	School	College	Police	Parking
5	Supermarket	Restaurant	Pet Shop	Shop	Church	Recreation	Stadium	Mall	Taxi Stand	Police
6	Restaurant	Hospital	Shop	Nightlife	Drugstore	Supermarket	Hotel	Gas Station	Gym	Church
7	Restaurant	Hospital	Stadium	Drugstore	Gas Station	Nightlife	Parking	School	College	Post Office
8	Supermarket	Train Station	Bus Station	Gym	Post Office	Pet Shop	Gas Station	Taxi Stand	Hotel	Police
9	Church	Post Office	Supermarket	Restaurant	Drugstore	College	Pet Shop	Gas Station	Taxi Stand	Shop

4.3. Some venue distributions around the clusters

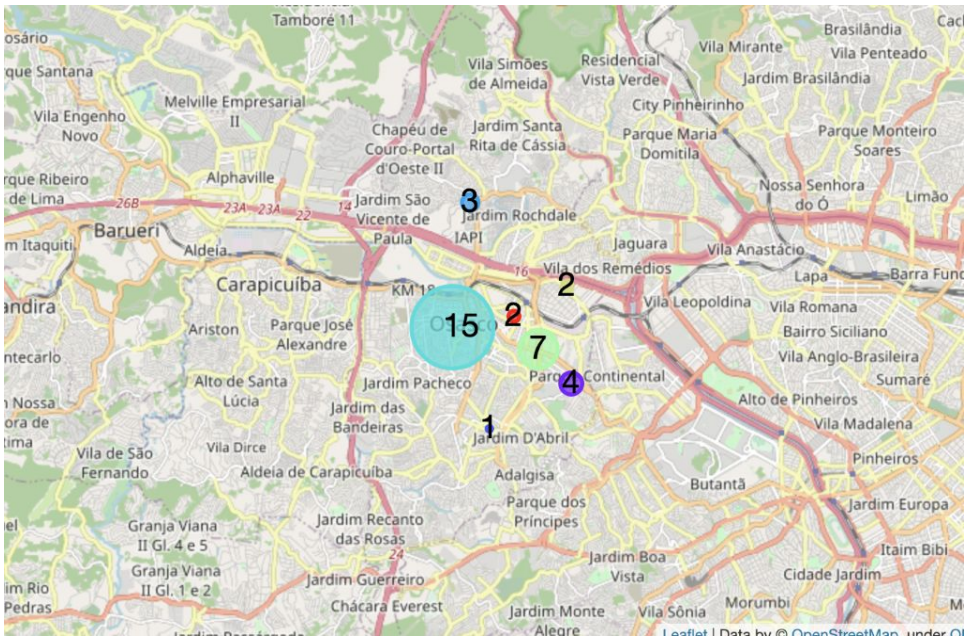
4.3.1. Gyms



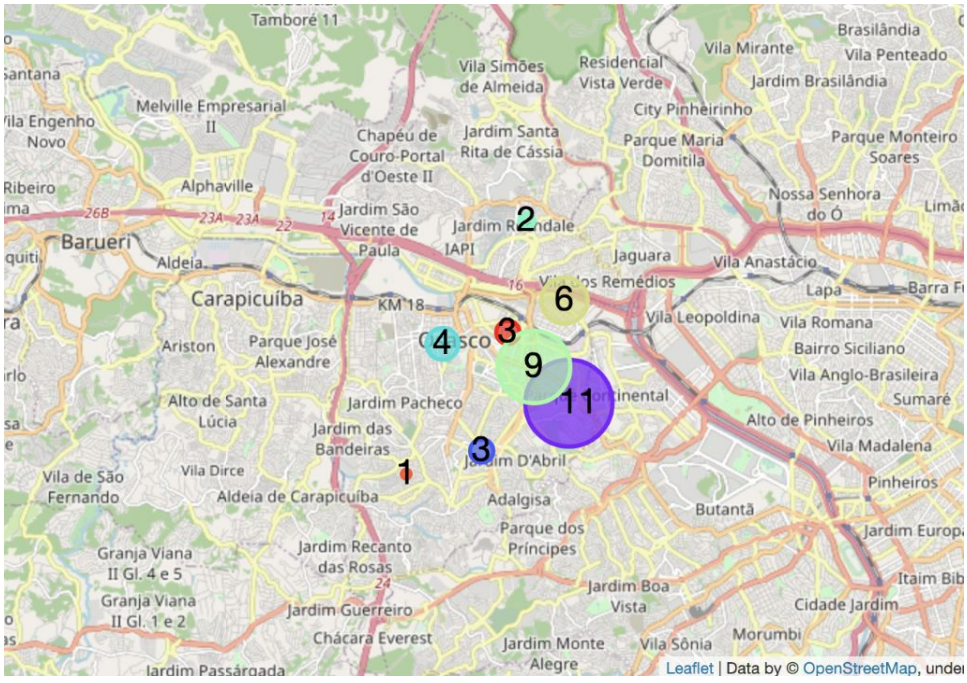
4.3.2. Supermarkets



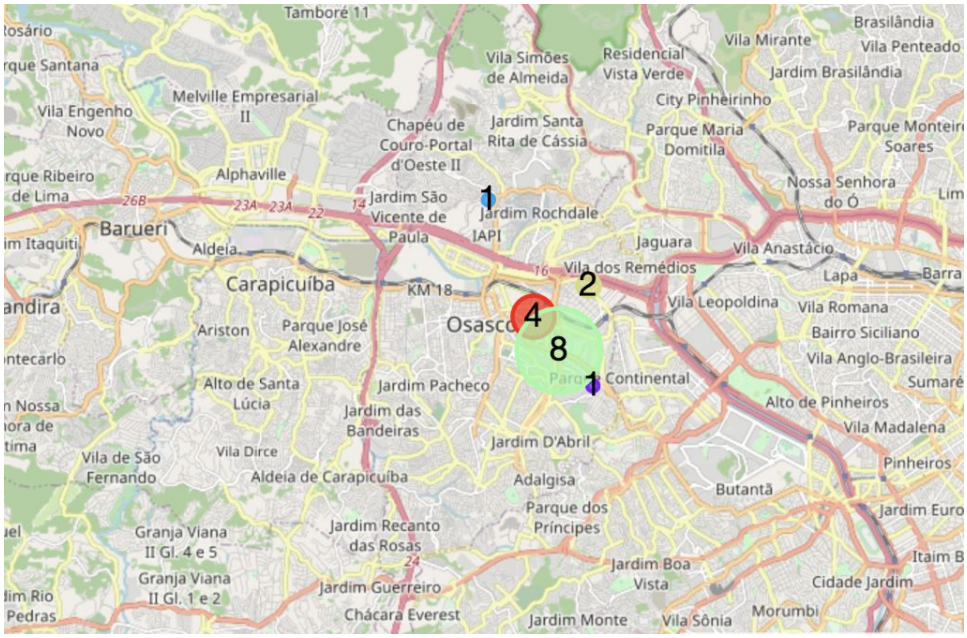
4.3.3. Nightlife



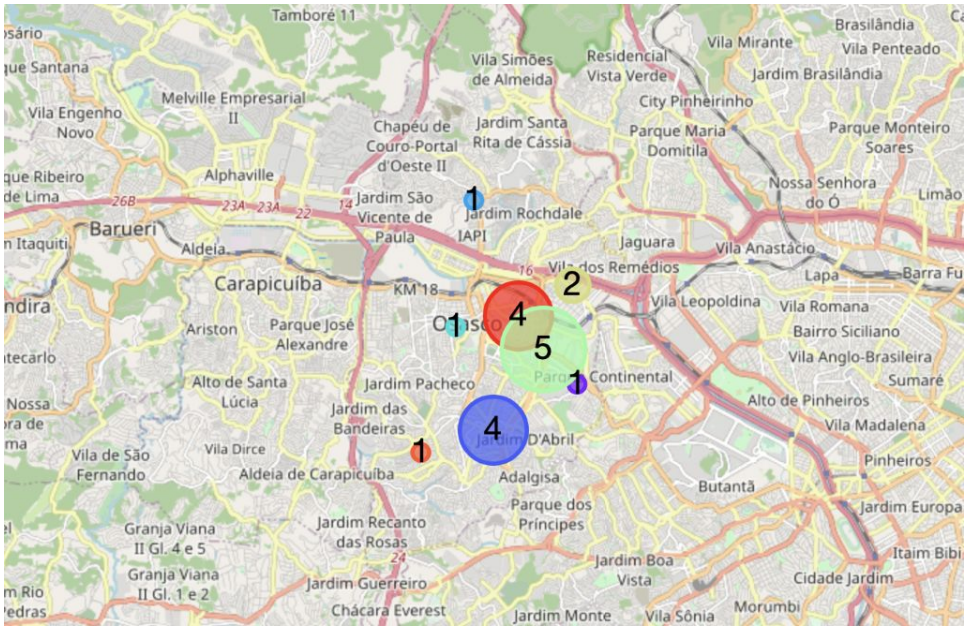
4.3.4. Restaurants



4.3.5. Hospitals



4.3.6. Drugstores



4.3.7. North vs South comparison of top 10 venues

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
North	Restaurant	Supermarket	Nightlife	School	Stadium	Post Office	Police	Gas Station	Bus Station	Drugstore
South	Restaurant	Nightlife	Shop	Supermarket	Gas Station	Drugstore	Hospital	Church	Recreation	Gym

Similarity index = 50%

5. Discussion

5.1. Similarity index

Since a similarity index was computed as 50% for north and south zones. In such way, we cannot conclude any significant difference, except for the number of venues, which is almost 4x higher for south zone, giving us na idea of the uneven distribution of venues. It is much higher when we compare the number of hospitals, 1 in north and 15 in south, leading us to a conclusion that the zone has shortages in such important public service, even though police stations are featured as a common spot in that area.

5.2. Influence of absent data

Although Foursquare API has a vast database of venues, including public utilities, commercial and service spots, it barely lists spots that are far from city center and might not reflect the reality in those communities. Also, relevant public data has not yet been available from government organizations of that city.

Such conditions impose a limitation to our analysis, requiring further research.

6. Conclusion

A data source such as Foursquare is a handful for introductory analysis of geographic sites which can help organizations to measure the potential of new businesses and, in comparison with other data sources, add significance to the problem analysis.
In this particular example, we managed to run a K-Means clustering, a powerful machine learning technique, to fill the data shortage related to the city's neighborhoods and helped us providing interesting visualizations in combination with Folium library.

Although not conclusive, this data analysis can be a good kickstart for projects applied to this specific city.

7. References

[1] https://news.un.org/pt/story/2019/02/1660701
[2] https://en.wikipedia.org/wiki/Osasco
[3] https://exame.com/brasil/as-cidades-mais-violentas-da-grande-sp-osasco-e-a-4a/