

Recommending neighborhoods to open stores

Gabriel Santos

July 28, 2020

1. Introduction

The city of Sao Paulo is the biggest metropolitan area of the South Hemisphere and is the biggest, the most populated and the richest city in Brazil. This city has a lot of italians, spanish and japanese immigrants and there is a great diversification in terms of gastronomy with about 75 different ethnicities. We can go in a luxury restaurant or in a place to snack and drink, at each corner there is a lot of options to support the demand of people in the city.

With the growth of people in the city it's a nice spot to open food stores, but where should you open your store in this city ? The city is huge and has a lot of boroughs with different categories of venues. Search the best place manually it's not the best option to take, so we can use the data and the power of the machine learning to help us finding the best place in the city.

2. Data

This project will rely on public data from "CEP Lá" and Foursquare.

The *Dataset 1* is the treated data from the .csv file generated by the .txt file available at <http://cep.la/CEP-dados-2018-latin1.zip> that contains all the data from postal codes, cities, boroughs and neighborhoods. I uploaded the csv file inside the jupyter notebook and I inserted to code transforming into a dataframe.

The *Dataset 2* it's the final result applying the Foursquare API in the dataset 1 pushing the nearby venues with the categories from each neighborhood 100 meters around and inserting this venues in the dataframe.

3. Methodology

In the *Dataset 1* with the dataframe I limited to use only data from São Paulo/SP and to use only data from 9 boroughs (Consolação, Bela Vista, Sé, República, Bom Retiro, Brás, Liberdade, Santa Cecília, Cambuci) in the center of the city, the heart of the city. After that I rename the columns. So far this dataset is large so I select randomly 40 neighborhoods in those boroughs and we will analyze this neighborhoods.

Postcode	Borough	Neighborhood
01312010	Bela Vista	Praça Quatorze Bis - lado par
03043900	Brás	Rua Alegria, 300
03041010	Brás	Rua Paraná
01310919	Bela Vista	Avenida Paulista, 1728
01309030	Consolação	Rua Fernando de Albuquerque

I used the “Geocoder” package with “arcgis_geocoder” to obtain the latitude and longitude of the needed locations. Running a loop in the dataset 1 it’s possible to obtain the latitude and longitude for each neighborhood and insert in the dataframe. The dataframe now have these columns: Postcode, Borough, Neighborhood, Latitude, Longitude.

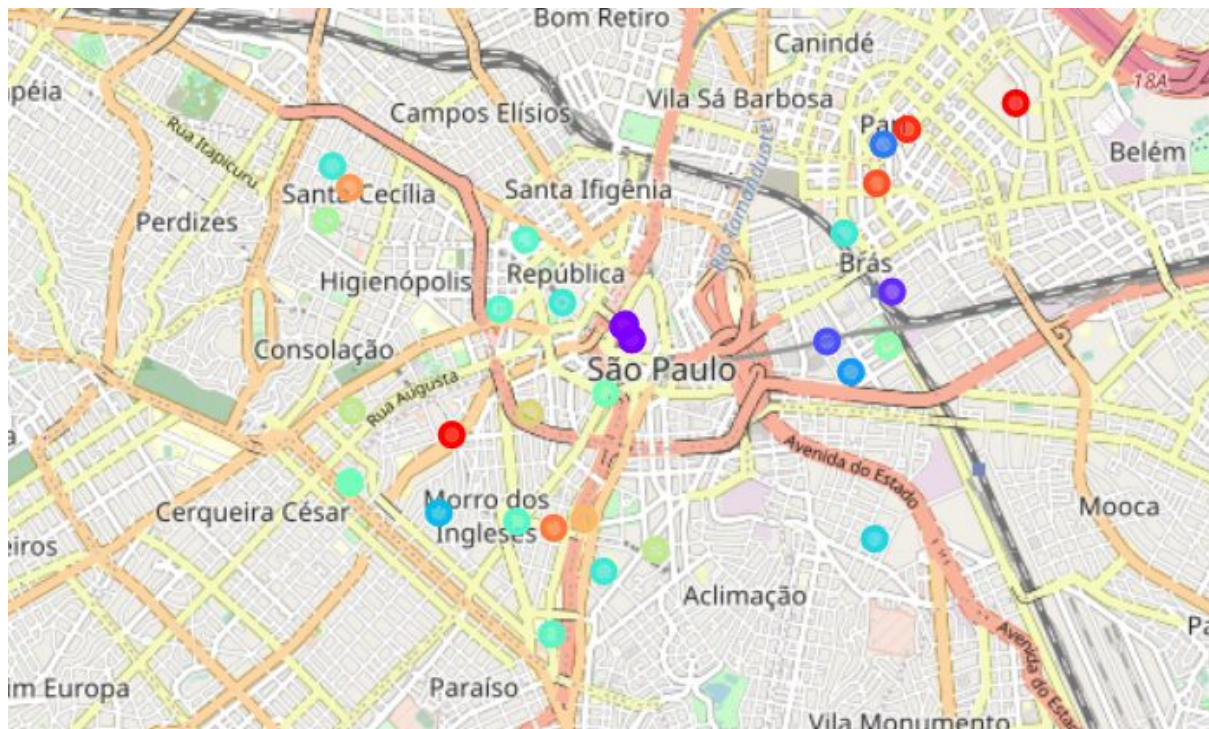
I applied the Foursquare API in the dataset 1 pushing the nearby venues with the categories from each neighborhood 100 meters around and inserting this venues in the dataframe. The dataframe now have these columns: Postcode, Postcode Latitude, Postcode Longitude, Venue, Venue Latitude, Venue Longitude, Venue Category. And now I have the *Dataset 2*.

Postcode	Postcode Latitude	Postcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
01312010	-23.555532	-46.650080	Frutaria E Cafeteria Bella Paim	-23.554799	-46.650001	Coffee Shop
01312010	-23.555532	-46.650080	Bar Brahma	-23.555664	-46.649738	Bar
01312010	-23.555532	-46.650080	Lanchonete Top Bis	-23.555549	-46.650022	Brazilian Restaurant
03043900	-23.549100	-46.615549	Restaurante Allegro	-23.549269	-46.615348	Brazilian Restaurant
03043900	-23.549100	-46.615549	Ponto Terminal Grajaú 5630-10	-23.548449	-46.616106	Bus Station

After that I applied `get_dummies` to convert the venues into numeric fields and applied “k-means++” with 20 clusters. With the result I used one function to return the most common venues for each postcode (neighborhood) and inserted in the *Dataset 1*. The dataframe now have these columns: Postcode, Borough, Neighborhood, Latitude, Longitude, Cluster Labels, 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.

Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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
01312010	Bela Vista	Prça Quatorze Bis - lado par	-23.555532	-46.650080	0.0	Coffee Shop	Bar	Brazilian Restaurant	Yoga Studio	French Restaurant	Dessert Shop	Empada House	Farmers Market	Fast Food Restaurant	Food & Drink Shop
03043900	Brás	Rua Alegria, 300	-23.549100	-46.615549	11.0	Bus Station	Brazilian Restaurant	Yoga Studio	French Restaurant	Dessert Shop	Empada House	Farmers Market	Fast Food Restaurant	Food & Drink Shop	Gastropub
03041010	Brás	Rua Paraná	-23.550919	-46.618377	5.0	Middle Eastern Restaurant	Yoga Studio	Food & Drink Shop	Department Store	Dessert Shop	Empada House	Farmers Market	Fast Food Restaurant	French Restaurant	Historic Site
01310919	Bela Vista	Avenida Paulista, 1728	-23.559007	-46.658125	10.0	Juice Bar	Dessert Shop	Gift Shop	Ice Cream Shop	Plaza	Brazilian Restaurant	Food & Drink Shop	Empada House	Farmers Market	Fast Food Restaurant
01309030	Consolação	Rua Fernando de Albuquerque	-23.553803	-46.657978	13.0	French Restaurant	Vegetarian / Vegan Restaurant	Bar	Café	Yoga Studio	Dessert Shop	Empada House	Farmers Market	Fast Food Restaurant	Food & Drink Shop

With the cluster labels I could applied one function to plot the result on a map.



After I plot the map I analyzed the results to recommend the best places to open a Pizza Place or a Gym. For each most common I searched if it was a Pizza Place or a Gym and I showed the borough and the neighborhood and the similar places inside the cluster of the recommended place.

4. Results and Discussion

With the cluster segmentation of the neighborhoods and the indicators of the most common venues by neighborhoods I could recommend the best places to help someone who wants to open a Pizza Place or a Gym.

PLACES WITH PIZZA AS 1st COMMON VENUE:

+++++

PLACES WITH PIZZA AS 2nd COMMON VENUE:

+++++

PLACES WITH PIZZA AS 3rd COMMON VENUE:

+++++

PLACES WITH PIZZA AS 4th COMMON VENUE:

+++++

PLACES WITH PIZZA AS 5th COMMON VENUE:

You can open your Pizza on REPÚBLICA around AVENIDA VIEIRA DE CARVALHO, 115
You can open your Pizza on BELA VISTA around AVENIDA BRIGADEIRO LUÍS ANTÔNIO, 733
Similar Places:
on REPÚBLICA around RUA REGO FREITAS, 354
on LIBERDADE around RUA MAESTRO CARDIM - DE 561/562 AO FIM

These are the recommendations for a Pizza Place, with the neighborhoods that I analyzed it's possible to find that two places where there is Pizza Place as 5th common venue in the neighborhood. With the cluster segmentation it's possible to show more two similar places that are in the same cluster as the neighborhood.

Recommendations for a Gym:

PLACES WITH GYM AS 1st COMMON VENUE:

You can open your Gym on SANTA CECÍLIA around RUA DOUTOR ALBUQUERQUE LINS, 574
Similar Places:

+++++

PLACES WITH GYM AS 2nd COMMON VENUE:

+++++

PLACES WITH GYM AS 3rd COMMON VENUE:

You can open your Gym on LIBERDADE around RUA MAESTRO CARDIM - DE 561/562 AO FIM
Similar Places:
on REPÚBLICA around RUA REGO FREITAS, 354
on REPÚBLICA around AVENIDA VIEIRA DE CARVALHO, 115
on BELA VISTA around AVENIDA BRIGADEIRO LUÍS ANTÔNIO, 733

+++++

PLACES WITH GYM AS 4th COMMON VENUE:

You can open your Gym on BRÁS around RUA DOUTOR ALMEIDA LIMA - ATÉ 599/600
Similar Places:

You can open your Gym on BRÁS around PRAÇA DOMINGOS FRANGIONE
Similar Places:

You can open your Gym on SANTA CECÍLIA around RUA DOUTOR VEIGA FILHO - LADO ÍMPAR
Similar Places:

You can open your Gym on BELA VISTA around RUA PEDROSO
Similar Places:

+++++

PLACES WITH GYM AS 5th COMMON VENUE:

You can open your Gym on BRÁS around RUA DOUTOR ALMEIDA LIMA - ATÉ 599/600
Similar Places:

You can open your Gym on BRÁS around PRAÇA DOMINGOS FRANGIONE
Similar Places:

You can open your Gym on SANTA CECÍLIA around RUA DOUTOR VEIGA FILHO - LADO ÍMPAR
Similar Places:

These are the recommendations for a Gym, with the neighborhoods that I analyzed it's possible to find some places where there is a Gym as 1st, 3rd, 4th and 5th common venue in the neighborhood. With the cluster segmentation it's possible to show some more similar places that are in the same cluster as the neighborhoods.

5. Conclusion

It's possible to create a machine learning program to recommend places to open a store and help someone. It's hard to do a manually research for a entire city so with this program we can indeed help these people. I'm learning the power of the machine learning and I am impressed with what I have found so far. This course helped me a lot in Data Science and I am grateful for what I can do know and for the opportunities that I will have with this knowledge.