Capstone Project – Italian Restaurant in São Paulo, Brasil

Author: Matheus Cammarosano Hidalgo

Introduction

A client owns an Italian restaurant in São Paulo, SP, Brasil. His venue is placed in the district of Pinheiros. His restaurant is presenting good financial results and he wants to open a second unit, but he does not know where to place it.

São Paulo is the most populous city in Brasil and is the ninth largest city in the world in terms of inhabitants. Its population is of 12.11 million people. As it is normal in the biggest Brazilian cities, there are huge socioeconomics differences between districts, so it is important to include socioeconomics features in the analysis, because placing the restaurant in a poorer district might be a problem in terms of security and profitability and placing it in a wealthier neighborhood might be troublesome in terms of competition, which might affect profitability. Pinheiros is considered a wealthy district in São Paulo.

Thereby, this capstone project considers variables such as the Gross Domestic Product (GDP) per capita, Human Development Index (HDI) to evaluate the districts of São Paulo. Also, FourSquare Developer API is used to obtain venues characteristics of the neighborhoods in order to evaluate the competition that my client would face in each district.

In the next sections, the data used in this project is explained and shown. Also, methodology is explained later, followed by the results, discussion and conclusion.

Data

As mentioned in the Introduction, socioeconomic data of each district is gathered for the analysis. The initial source of data is the Wikipedia page: https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_S%C3%A3o_Paulo_por_popula%C3%A7%C3%A3o and from it the data of each district is obtained. The data of interest of each district is GDP per capita, which indicates how wealthy it is, the HDI, that shows how developed is the neighborhood. The HDI englobes a number of indexes and indicates how economically and socially developed is a determined place. Also, the population of each district is collected in order to quantify a potential amount of customers in a region.

Also, the geospatial coordinates of the districts with similar socioeconomics characteristics of Pinheiros are gathered. The data source is https://www.adistanciaentre.com/. With this information, the distance of the districts from Pinheiros (where the first restaurant is located) is calculated.

From FourSquare Developer API, the Italian restaurants and pizzas places from the districts and its respective price tiers and ratings are collected. In Brasil, it is relatively normal that Italian restaurants have pizzas in their menus, so the classification between Italians and pizzerias might have some intersections, so in this project both categories are analyzed together.

Methodology

The first step of the analysis is obtaining each district of São Paulo name, GDP per capita and HDI. With these informations, a K-means clustering algorithm is used to classify the districts according to these socioeconomics characteristics.

With the results given by K-means, the cluster that contains Pinheiros district (where the restaurant is located) can be identified and it presents every district with similar socioeconomics characteristics. Since Pinheiros showed to be an adequate choice of neighborhood for my client's restaurant, the other districts in this cluster might be good choices as well. Thus, a new dataframe is created containing only the districts that are in the same cluster as Pinheiros, which is named *potential* dataframe.

Thus, in the second part of the analysis, the geospatial coordinates of each district in *potential* dataframe is obtained by using Google search in my notebook and accessing a web site that returns the latitude and longitude of a given entry. Once obtaining the coordinates, the distance between each district of *potential* dataframe and Pinheiros (where the restaurant is located) is calculated. This feature is a measurement of how far the second unit of the restaurant would be of the first one, after all it is not desirable to have two units of the same brand of restaurants competing with each other.

Besides, the venues for each district in *potential* dataframe are obtained through FourSquare API. After that, the Italian restaurants and pizzas places are filtered, because they are the potential competition for my client's restaurants in the neighborhood. With the venue ID of each one of these places, it is possible to gather further information about them, such as price tier and rating, which are added to my dataframe.

My client's restaurant is of price tier 3 and have a rating of 8.9, so with the informations given above, it can be inferred the type of competition he might face in each district. It is defined an order of scenarios, from best to worse:

1. There are no Italian restaurants/pizza places in the district;
2. There are Italian restaurants or pizza places, but they are from a price tier far from my client's restaurant, that is, restaurants with price tiers different from 2, 3 and 4 (my client's price tier and its closer tiers);
3. There are Italian restaurants/pizzerias with price tier and/or 4, but not 3;
4. There is/are restaurant(s) with price tier 3 (direct competition).

Thereby, a ranking is made based on this criteria. However, if there is more than one districts that have no Italian/pizza restaurants, another criteria is necessary to indicate the best place to open the restaurant. Thus, the population of each district in my dataframe is calculated and the district best classified with the highest number of inhabitants is the best place to open the second unit of the restaurant.

Results

The result of the the first part of the analysis, in which the data with the socioeconomics indicators are clustered is show in Table 1. The GDP per capita is in Brazilian currency, that is, real (R$) and US$1 = R$3.80 (on January, 2019).

Table 1 – Dataframe with the clustered districts of São Paulo

	District	GDP per capita	HDI	Cluster ID
0	Grajaú (distrito de São Paulo)	2500.00	0.754	5
1	Sapopemba	3041.40	0.786	0
2	Jardim Ângela	1700.00	0.750	5
6	Cidade Ademar	5500.00	0.801	0
11	Campo Limpo (distrito de São Paulo)	1745.78	0.806	6
13	Cidade Dutra	4200.00	0.815	0
14	Itaquera	1058.87	0.795	6
15	Tremembé (distrito de São Paulo)	1120.66	0.826	6
16	Lajeado (distrito de São Paulo)	543.58	0.748	5
18	Pirituba	4200.00	0.904	1
19	Vila Curuçá	653.78	0.765	5
22	Cachoeirinha (distrito de São Paulo)	874.21	0.802	6
23	Jardim Helena	870.00	0.751	5
29	Cidade Líder	897.11	0.817	6
30	Vila Medeiros	894.53	0.836	6
33	Ermelino Matarazzo (distrito de São Paulo)	822.70	0.801	6
35	Vila Mariana (distrito de São Paulo)	7000.00	0.950	2
36	Santana (distrito de São Paulo)	2507.75	0.925	1
37	Guaianases (distrito de São Paulo)	1058.87	0.770	5
38	Saúde (distrito de São Paulo)	8000.00	0.942	2
39	José Bonifácio (distrito de São Paulo)	807.88	0.804	6
40	Vila Maria (distrito de São Paulo)	1014.48	0.824	6
44	Vila Matilde	1160.95	0.864	3
45	Perdizes (distrito de São Paulo)	6746.86	0.957	2
46	Raposo Tavares (distrito de São Paulo)	968.03	0.819	6
48	Vila Prudente (distrito de São Paulo)	1345.63	0.867	3
50	Ponte Rasa	919.10	0.834	6
52	Aricanduva (distrito de São Paulo)	3698.75	0.879	3
53	Cursino	1818.50	0.885	3
54	Jaçanã (distrito de São Paulo)	1776.56	0.816	6
55	São Domingos (distrito de São Paulo)	1279.40	0.854	3
57	Vila Formosa (distrito de São Paulo)	3755.40	0.884	3
58	Tucuruvi	2836.56	0.923	1
59	Perus	650.36	0.772	5
61	Itaim Bibi	5500.23	0.953	2
62	Água Rasa	2503.34	0.886	3
63	Jardim Paulista (distrito de São Paulo)	5144.03	0.957	2
64	Casa Verde (distrito de São Paulo)	1411.67	0.874	3
65	Tatuapé	3661.96	0.936	1
66	Moema (distrito de São Paulo)	12428.00	0.981	4
67	Carrão	2948.70	0.886	3
68	Parque do Carmo (distrito de São Paulo)	2200.18	0.859	3
69	Santa Cecília (distrito de São Paulo)	2505.76	0.930	1
70	Mooca	4098.75	0.909	1
71	Campo Belo (distrito de São Paulo)	6000.00	0.935	2
72	Pinheiros (distrito de São Paulo)	7000.00	0.960	2
73	Santo Amaro (distrito de São Paulo)	6000.00	0.943	2
74	Lapa (distrito de São Paulo)	5000.00	0.941	2
75	Liberdade (distrito de São Paulo)	2333.84	0.858	3
76	Bela Vista (distrito de São Paulo)	2435.70	0.940	1
78	Vila Guilherme	1393.41	0.868	3
79	Butantã	2584.46	0.928	1
80	Consolação (distrito de São Paulo)	4094.68	0.950	1
82	Jaguaré (distrito de São Paulo)	1487.11	0.849	3
83	Alto de Pinheiros	4809.46	0.955	2
84	Socorro (distrito de São Paulo)	3000.00	0.841	3
86	Morumbi (distrito de São Paulo)	13802.00	0.938	4
87	Vila Leopoldina	5737.79	0.907	2
88	Cambuci (distrito de São Paulo)	1604.97	0.903	1
89	Bom Retiro (distrito de São Paulo)	1358.39	0.847	3
90	Brás (distrito de São Paulo)	1240.11	0.868	3
91	Jaguara	1020.82	0.863	3
92	Sé (distrito de São Paulo)	978.31	0.858	3

And the clusters can be visualized in the scatter plot of Figure 1.

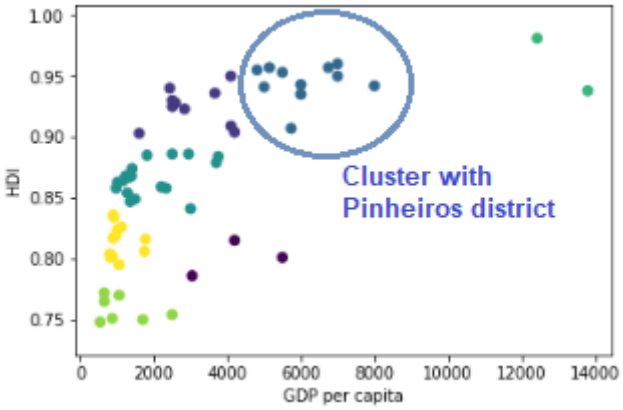

Figure 1 – Scatter plot of clustered data

The results in Figure 1 and the number of clusters assigned seem to be reasonable. The district of Pinheiros is placed in cluster 2, so now I can create my *potential* dataframe, with the first round of candidates, shown in Table 2.

Table 2 – Potential dataframe

	District	GDP per capita	HDI
0	Vila Mariana (distrito de São Paulo)	7000.00	0.950
1	Saúde (distrito de São Paulo)	8000.00	0.942
2	Perdizes (distrito de São Paulo)	6746.86	0.957
3	Itaim Bibi	5500.23	0.953
4	Jardim Paulista (distrito de São Paulo)	5144.03	0.957
5	Campo Belo (distrito de São Paulo)	6000.00	0.935
6	Pinheiros (distrito de São Paulo)	7000.00	0.960
7	Santo Amaro (distrito de São Paulo)	6000.00	0.943
8	Lapa (distrito de São Paulo)	5000.00	0.941
9	Alto de Pinheiros	4809.46	0.955
10	Vila Leopoldina	5737.79	0.907

As can be seen in Table 2, the districts in this dataframe have a very high HDI and a very good GDP per capita for Brazilian standards. As mentioned before, the district of Pinheiros is wealthy and K-means was able to find districts with similar characteristics. It can be noted that Moema and Morumbi are neighborhoods even wealthier, but they were placed in another cluster, because their GDP per capita is on another level. Placing these 2 districts in the same cluster might be troublesome, because in this neighborhoods, the client would have to compete with restaurants with high quality and price tiers. The price of the client's restaurant is relatively high and in these two places the competition would be harsh.

It is possible to note that the districts in potential dataframe have similar GDP per capita and HDI, which is expected. HDI, which is expected. Now the coordinates are obtained and the distance of these districts from Pinheiros are from Pinheiros are calculated. These results are synthetized in

Table 3.

Table 3 - Potential dataframe with coordinates and distance

	District	GDP per capita	HDI	Latitude	Longitude	Distance
0	Vila Mariana (distrito de São Paulo)	7000.00	0.950	-23.5895	-46.6346	7.423629
1	Saúde (distrito de São Paulo)	8000.00	0.942	-23.6184	-46.6394	8.673263
2	Perdizes (distrito de São Paulo)	6746.86	0.957	-23.5358	-46.6678	4.961725
3	Itaim Bibi	5500.23	0.953	-23.5846	-46.6878	2.531746
4	Jardim Paulista (distrito de São Paulo)	5144.03	0.957	-23.5693	-46.6565	4.750260
5	Campo Belo (distrito de São Paulo)	6000.00	0.935	-23.6264	-46.6855	6.886571
6	Pinheiros (distrito de São Paulo)	7000.00	0.960	-23.5666	-46.7030	0.000000
7	Santo Amaro (distrito de São Paulo)	6000.00	0.943	-23.6537	-46.7067	9.695455
8	Lapa (distrito de São Paulo)	5000.00	0.941	-23.5227	-46.7104	4.940949
9	Alto de Pinheiros	4809.46	0.955	-23.5544	-46.7084	1.464441
10	Vila Leopoldina	5737.79	0.907	-23.5381	-46.7311	4.272997

Now, FourSquare API can be used and the Italian/pizza restaurants information can be obtained. The results are shown in

Table 4.

Table 4 – Italian and pizza restaurants in the districts of interest

	District	District Latitude	District Longitude	Venue	Venue Category	Rating	Price Tier
7	Vila Mariana (distrito de São Paulo)	-23.5895	-46.6346	Sapore Cantina e Pizzaria	Pizza Place	8.5	2
11	Vila Mariana (distrito de São Paulo)	-23.5895	-46.6346	MiPi Pizzeria	Pizza Place	8.4	1
63	Perdizes (distrito de São Paulo)	-23.5358	-46.6678	Restaurante e Pizzaria Macedo's	Pizza Place	8.6	2
69	Perdizes (distrito de São Paulo)	-23.5358	-46.6678	Dona Veridiana Perdizes	Pizza Place	9.2	1
94	Itaim Bibi	-23.5846	-46.6878	Cristal Pizza Bar	Pizza Place	8.6	3
103	Itaim Bibi	-23.5846	-46.6878	Sallvattore	Italian Restaurant	9.2	2
120	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Veridiana	Pizza Place	9.3	3
121	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Tatini	Italian Restaurant	9.1	3
134	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Osteria Generale	Italian Restaurant	9.1	2
138	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Trattoria do Sargento	Italian Restaurant	8.2	2
141	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Taormina Ristorante	Italian Restaurant	8.7	2
145	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Cantina Nipolitana	Italian Restaurant	8.5	2
148	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Bráz Elettrica	Pizza Place	8.2	1
149	Jardim Paulista (distrito de São Paulo)	-23.5693	-46.6565	Prestíssimo Pizza Bar	Pizza Place	8.0	3
153	Campo Belo (distrito de São Paulo)	-23.6264	-46.6855	Vicolo Nostro	Italian Restaurant	9.3	4
207	Pinheiros (distrito de São Paulo)	-23.5666	-46.7030	Forneria Urbana	Pizza Place	8.1	1
244	Lapa (distrito de São Paulo)	-23.5227	-46.7104	Pizzaria Famiglia Lucco	Pizza Place	8.8	3
262	Lapa (distrito de São Paulo)	-23.5227	-46.7104	O Gordo e O Magro	Italian Restaurant	7.7	2
264	Lapa (distrito de São Paulo)	-23.5227	-46.7104	D'Lapanto	Pizza Place	8.4	1
283	Alto de Pinheiros	-23.5544	-46.7084	Vito Restaurante	Italian Restaurant	8.6	3
298	Alto de Pinheiros	-23.5544	-46.7084	Monte Verde Pizzaria	Pizza Place	7.2	1

This data can be reorganized in a way that is easier to comprehend and evaluate the best neighborhoods to place the restaurant. Table 5 synthesizes the districts data, englobing the quantity of Italian and pizza restaurants and its price tiers.

Table 5 – Districts and its Italians/piza restaurants

	District	GDP per capita	HDI	Latitude	Longitude	Distance	Qty Italians	Price Tier	Ranking
0	Vila Mariana (distrito de São Paulo)	7000.00	0.950	-23.5895	-46.6346	7.423629	2	2, 1	3
1	Saúde (distrito de São Paulo)	8000.00	0.942	-23.6184	-46.6394	8.673263	0	0	1
2	Perdizes (distrito de São Paulo)	6746.86	0.957	-23.5369	-46.6743	4.413271	4	2, 2, 2, 2	3
3	Itaim Bibi	5500.23	0.953	-23.5846	-46.6878	2.531746	2	3, 2	4
4	Jardim Paulista (distrito de São Paulo)	5144.03	0.957	-23.5693	-46.6565	4.750260	8	3, 3, 2, 2, 2, 2, 1, 3	4
5	Campo Belo (distrito de São Paulo)	6000.00	0.935	-23.6264	-46.6855	6.886571	1	4	3
6	Pinheiros (distrito de São Paulo)	7000.00	0.960	-23.5666	-46.7030	0.000000	1	1	2
7	Santo Amaro (distrito de São Paulo)	6000.00	0.943	-23.6537	-46.7067	9.695455	0	0	1
8	Lapa (distrito de São Paulo)	5000.00	0.941	-23.5227	-46.7104	4.940949	3	3, 2, 1	4
9	Alto de Pinheiros	4809.46	0.955	-23.5544	-46.7084	1.464441	2	3, 1	4
10	Vila Leopoldina	5737.79	0.907	-23.5381	-46.7311	4.272997	0	0	1

Table 5 shows that Saúde, Santo Amaro and Vila Leopoldina are tied in first place of the ranking, since neither of them have Italian and pizzas restaurants. The first two districts are far from Pinheiros and the third one is not close from the first restaurant but not that far away.

Now, the population of these districts is added to the dataframe in order to draw a definitive result, which is shown in Table 6.

Table 6 – Potential dataframe with population data

	District	GDP per capita	HDI	Latitude	Longitude	Distance	Qty Italians	Price Tier	Ranking	Population
0	Vila Mariana	7000.00	0.950	-23.5895	-46.6346	7.423629	2	2, 1	3	112952
1	Saúde	8000.00	0.942	-23.6184	-46.6394	8.673263	0	0	1	111308
2	Perdizes	6746.86	0.957	-23.5369	-46.6743	4.413271	4	2, 2, 2, 2	3	97706
3	Itaim Bibi	5500.23	0.953	-23.5846	-46.6878	2.531746	2	3, 2	4	80501
4	Jardim Paulista	5144.03	0.957	-23.5693	-46.6565	4.750260	8	3, 3, 2, 2, 2, 2, 1, 3	4	76033
5	Campo Belo	6000.00	0.935	-23.6264	-46.6855	6.886571	1	4	3	62530
6	Pinheiros	7000.00	0.960	-23.5666	-46.7030	0.000000	1	1	2	61711
7	Santo Amaro	6000.00	0.943	-23.6537	-46.7067	9.695455	0	0	1	60373
8	Lapa	5000.00	0.941	-23.5227	-46.7104	4.940949	3	3, 2, 1	4	58924
9	Alto de Pinheiros	4809.46	0.955	-23.5544	-46.7084	1.464441	2	3, 1	4	39477
10	Vila Leopoldina	5737.79	0.907	-23.5381	-46.7311	4.272997	0	0	1	30188

Of the 3 remaining candidates, Saúde is the district with the highest number of inhabitants by far. Therefore, Saúde is most the indicated place to open the second unit of the restaurant.

Finally, in Figure 2, it is displayed a map with the districts in *potential* dataframe. The blue marker represents Pinheiros, where the first restaurant is located. The green one represents Saúde and the red ones are the other districts of the cluster.

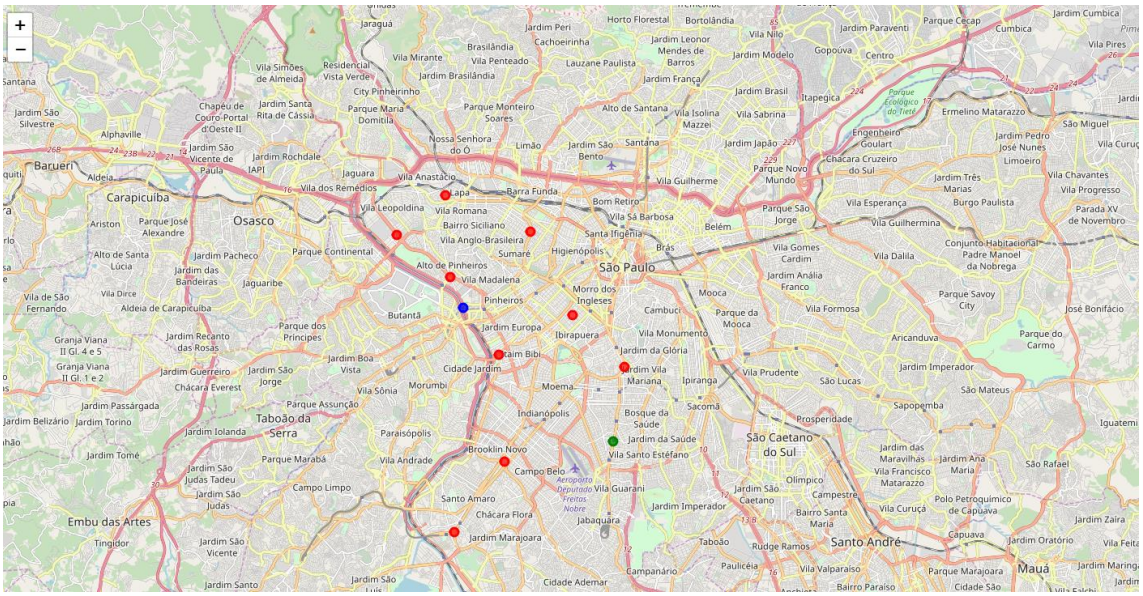

Figure 2 – Map of São Paulo with districts

In Figure 2, the two closest districts from Saúde in the cluster are Vila Mariana and Campo Belo, and neither of them have a price tier 3 restaurant. Thereby, the restaurant in Saúde might attract customers from these two districts.

Discussion

The data analysis showed that Saúde is the most indicated district to open the second unit of the restaurant. Firstly because it has similar socioeconomics characteristics than Pinheiros, thus the price rate from the restaurant will probably adequate for the neighborhood income.

Secondly because there is no competition in the district. FourSquare API did not find any other Italian restaurants and pizza places (hypothesis: the API is correct), so there is a lot of pent-up

demand. Thirdly because Saúde is far from Pinheiros, so there will not be a competition between the two restaurants and lastly, Saúde is a very populous district with a pent-up demand, so it seems to be an excellent place to open the restaurant there.

Besides, the closest districts from Saúde in my cluster of interest are Vila Mariana and Campo Belo, which do not have a tier 3 restaurant, so the client's place might also attract clients from these districts.

The second best option to open the restaurant is in Santo Amaro, it presents the same characteristics as Saúde, except for the population (54% of the number of inhabitants of Saúde) and the fact that is farther of the other districts of my cluster. Vila Leopoldina is the third best option, but its drawbacks are the lesser amount of inhabitants and its relative proximity to Pinheiros.

Conclusions

The data analysis in this report shows that the district of Saúde concentrates the best features to open the second unit of the customer's restaurant.

Santo Amaro would be the second best place to open the restaurant, followed by Vila Leopoldina.