Selecting the best investment in a food business

Álvaro César Leão Teixeira[1]

June 25, 2019

1. Introduction

Open a new business is always a hard challenge, but sometimes is even worst. Brazil is a country with so many problems, especially these days. When the economy crashes, we need to be creative to find ways to survive, and guarantee a minimum quality of life. A good way to assure this is to create a new business. But the competition is always enormous, especially as there are so many unemployed people in traditional jobs. My objective in this project is somehow facilitate to choose which kind of food business I should open in a particular neighborhood in one of the biggest cities in Brazil, Belo Horizonte. I will considering the top 10 most popular categories of this particular segment, and use Foursquare API, some data of the best neighborhood to open a restaurant and a clusterization technique to try to find one answer. Of course, not all the tools and techniques were applied in this analysis, and it is a theoretical work.

[1] Student from IBM Data Science Specialization .

2. Data

2.1 Data Sources

The main data that I will use is the list of postal codes/neighbourhood (available on URL) of one of the brazilian big cities, Belo Horizonte, specially one of the most festive neighbourhoods, Savassi. Then I will take the latitude and longitude data for every street in this neighbourhood, and cross this data with Foursquare, to retrieve the category and additional information of all the restaurant in the area, and choose the the category that least appears, using K-means to cluster the data .

LOGRADOURO	BAIRRO	CIDADE/ESTADO	CEP
Avenida Afonso Pena *- de 2022 a 3200 - lado par*	Savassi	Belo Horizonte, MG	30130-012
Rua Alagoas *- de 811/812 a 1099/1100*	Savassi	Belo Horizonte, MG	30130-167
Rua Alagoas *997*	Savassi	Belo Horizonte, MG	30130-912
Rua Alagoas *1314*	Savassi	Belo Horizonte, MG	30130-913
Rua Alagoas *- de 1101/1102 ao fim*	Savassi	Belo Horizonte, MG	30130-168
Rua Alagoas *- de 531/532 a 809/810*	Savassi	Belo Horizonte, MG	30130-165
Rua Antônio de Albuquerque *- até 539/540*	Savassi	Belo Horizonte, MG	30112-010

Figure 1 - a sample of the data source with neighborhood data. Retrieved from https://cep.guiamais.com.br/busca/savassi-belo+horizonte-mg

I used the BeautifulSoap tool to handle the data for this site. To complete the information with latitude and longitude data, I used the Geocode tool to.

2.2 Data Cleaning

The first step was to clean the data returned from the site where information about the neighborhood was found. This data was stored in a matrix at first, and all the garbage and unnecessary data (like HTML tags and irrelevant information), and the

duplicated record too, were cleaned. Then I created a dataframe and searched for latitude and longitude data using the geocode's geolocator

A second step was to clean the data from Foursquare. We used a little piece of all data retrieved from service. So, we need to clean all the unnecessary data unless the venue category, location, latitude and longitude.

	Street	Postcode	Neighborhood	Latitude	Longitude
0	Rua Alagoas	30130-167	Savassi	-19.9318	-43.9351
1	Rua Antônio de Albuquerque	30112-010	Savassi	-19.9395	-43.9302
2	Rua Bernardo Guimarães	30140-081	Savassi	-19.9309	-43.9333
3	Avenida Brasil	30140-008	Savassi	-19.9318	-43.9351
4	Rua Ceará	30150-314	Savassi	-19.937	-43.9297
5	Rua Cláudio Manoel	30140-105	Savassi	-19.9336	-43.9329
6	Avenida Cristóvão Colombo	30140-140	Savassi	-19.9384	-43.9353
7	Praça Diogo de Vasconcelos	30140-160	Savassi	-19.9379	-43.9356
8	Rua Fernandes Tourinho	30112-004	Savassi	-19.9398	-43.9342

Table 1 - a sample of the data after the cleaning process

3. Methodology

3.1. Exploratory Data Analysis

The first challenge was to collect data that would help find locations in the neighborhood with higher and lower incidence of food-related business establishments. I decided this would be done through the neighborhood street listing. To find them, I first searched for a site that contained all the streets of a neighborhood divided by postal address, and that were in some format that would facilitate the interpretation of the text through a program. The found site stored the data in a table. After reading this data, they were transformed and cleaned, as explained in the data cleaning session. Then Geopy library was used to bring the location data (latitude / longitude) of the neighborhood streets. Then, a chart was plotted with the map data to validate the search results

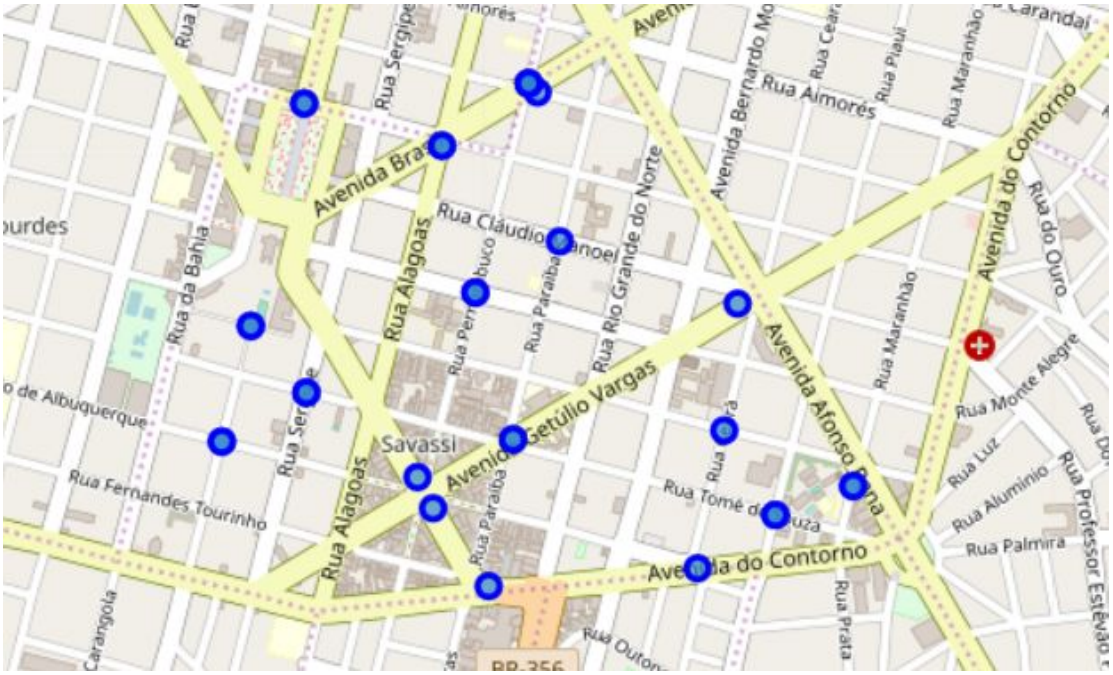

Figure 2 - The neighborhood streets plotted

Next, I searched the data of commercial establishments, cleaning and bringing only the venues categories related to food, and latitude and longitude data for each establishment. 41 unique categories found.

Venue Category	Street
Brazilian Restaurant	296
Restaurant	224
Café	200
Snack Place	159
Pizza Place	157
Burger Joint	118
Italian Restaurant	106
Vegetarian / Vegan Restaurant	97
Bakery	84
Buffet	54
Gastropub	53

Table 2 - the top 11 among the 41 categories found

So, first personal definition, I will select the 11th most common, because the top 10 is too usual. After this definition, the data were transformed to facilitate analysis and clustering. Then, the establishments were grouped by street, with the frequency of each one.

	Street	Asian Restaurant	BBQ Joint	Baiano Restaurant	Bakery	Bistro	Brazilian Restaurant	Breakfast Spot	Buffet	Burger Joint	Café	Chinese Restaurant	Creperie	Deli / Bodega	Diner	Empada House	Falafel Restaurant	Fast Food Restaurant	Food
0	Rua Pernambuco 1322	0.000000	0.020408	0.000000	0.051020	0.010204	0.163265	0.020408	0.020408	0.040816	0.081633	0.000000	0.010204	0.000000	0.010204	0.000000	0.000000	0.000000	0.040816
1	Avenida Brasil	0.000000	0.000000	0.000000	0.033333	0.033333	0.166667	0.016667	0.016667	0.066667	0.100000	0.000000	0.016667	0.000000	0.016667	0.000000	0.000000	0.000000	0.033333
2	Avenida Cristóvão Colombo	0.010000	0.010000	0.010000	0.030000	0.030000	0.090000	0.010000	0.050000	0.040000	0.080000	0.010000	0.000000	0.010000	0.000000	0.000000	0.000000	0.030000	0.000000
3	Avenida Getúlio Vargas	0.018868	0.018868	0.000000	0.037736	0.000000	0.226415	0.000000	0.000000	0.018868	0.075472	0.018868	0.000000	0.000000	0.000000	0.018868	0.000000	0.000000	0.018868
4	Praça Diogo de Vasconcelos	0.010000	0.010000	0.010000	0.020000	0.030000	0.100000	0.010000	0.030000	0.040000	0.080000	0.010000	0.000000	0.010000	0.000000	0.000000	0.000000	0.030000	0.000000
5	Rua Alagoas	0.000000	0.000000	0.000000	0.033333	0.033333	0.166667	0.016667	0.016667	0.066667	0.100000	0.000000	0.016667	0.000000	0.016667	0.000000	0.000000	0.000000	0.033333
6	Rua Antônio de Albuquerque	0.017241	0.034483	0.017241	0.068966	0.000000	0.086207	0.000000	0.000000	0.034483	0.103448	0.000000	0.000000	0.034483	0.000000	0.000000	0.017241	0.000000	0.000000
7	Rua Bernardo Guimarães	0.000000	0.021277	0.000000	0.042553	0.010638	0.170213	0.021277	0.021277	0.042553	0.063830	0.000000	0.010638	0.000000	0.010638	0.000000	0.000000	0.000000	0.042553
8	Rua Ceará	0.021277	0.000000	0.000000	0.042553	0.021277	0.191489	0.000000	0.021277	0.042553	0.085106	0.000000	0.000000	0.000000	0.000000	0.021277	0.000000	0.000000	0.000000

Table 3 - the frequencies of venues by streets

Then I analyzed the highest frequency of each establishment per street, returning the top 10 of each street (remembering, I decided to choose the 11 option of establishment, however I will analyze if this option is frequent in the specific street that will be chosen, to avoid a place very common). And then listed again, in tabular format, the top 10 choices per street:

	Street	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	Rua Pernambuco 1322	Brazilian Restaurant	Restaurant	Café	Snack Place	Bakery	Vegetarian / Vegan Restaurant	Burger Joint	Pizza Place	Italian Restaurant	Food
1	Avenida Brasil	Brazilian Restaurant	Restaurant	Café	Snack Place	Burger Joint	Vegetarian / Vegan Restaurant	Bakery	Pizza Place	Bistro	Mineiro Restaurant
2	Avenida Cristóvão Colombo	Pizza Place	Brazilian Restaurant	Café	Restaurant	Snack Place	Vegetarian / Vegan Restaurant	Buffet	Gastropub	Burger Joint	Sushi Restaurant
3	Avenida Getúlio Vargas	Brazilian Restaurant	Snack Place	Restaurant	Pizza Place	Café	Vegetarian / Vegan Restaurant	Steakhouse	Bakery	Sushi Restaurant	Gastropub
4	Praça Diogo de Vasconcelos	Brazilian Restaurant	Pizza Place	Restaurant	Café	Snack Place	Vegetarian / Vegan Restaurant	Gastropub	Burger Joint	Sushi Restaurant	Italian Restaurant

Table 4 - the top 10 establishments per street (5 streets example)

After that, the data were divided into 5 clusters, using the kmeans method:

> “The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:
>
> - The cluster cluster is the arithmetic mean of all points belonging to the cluster.
> - Each point is closer to its own cluster than to other cluster centers.
>
> Those two assumptions are the basis of the k-means model.”
>
> In Depth: K-Means Clustering, Python Data Science Handbook, 2008, ohttps: //jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html. Accessed June 23, 2019.

Then it was analyzed which cluster had the smallest number of streets. This was the cluster defined.

Cluster	Number of Streets
1	12
2	6
0	6
3	5
4	1

Table 5 - the top 10 establishments per street (5 streets example)

4. Results

The final step was to analyze and compile the clustering results, combined with the popularity data for each option. As mentioned earlier, the definition of which restaurant option would be chosen was made by a non-mathematical fact. But this would be reviewed later if the cluster with the least number of streets had among the most popular options the option chosen on all streets. But that's not what happened. Firstly, this cluster had only one street option, which reduced the analysis. And second, because this street did not have among the 10 most common options the type of commerce chosen. So the final option was an option that met the criteria.

To validate the results, I plotted the map with the streets identified by cluster, only to check if it was well distributed. As we can see, the cluster with the least number of streets (in this case, only one street), didn't have in the 10 most common options the Gastropub (the 11 most common general option, which had been previously selected). As it did not exist, it was defined that the restaurant will be a gastropub, opened in the Getúlio Vargas Avenue.

Figure 3 - The streets grouped by clusters

5. Discussion

About the result itself, many less explicit variables can be taken into account to determine whether a trade will succeed, like cost, revenue, salary range, and so many others. But relevant factors were used, including the area surveyed in person, and really is an area with a restaurant deficit compared to other areas of the neighborhood. But to make a deeper analysis, with more complex variables, more time and study will be necessary.

6. Conclusions

Although the analysis was not a thorough analysis, and it was a superficial work, it fulfilled the main purpose, which was to exercise not only the techniques learned in the course, but also the whole elaboration of the final work. This step was successfully completed. So, even being superficial, the work was successfully completed, and this possible business opportunity included more variable than many real cases.