FINDING THE BEST PLACE TO INSTALL A JAPANESE RESTAURANT IN NATAL - BRAZIL

GEORGE THÓ

August 10, 2020

1 INTRODUCTION

1.1 *Background*

Located in the northeast of Brazil, Natal is the capital of the state of Rio Grande do Norte and its main economic sector is tourism, where its natural beauty, especially its beaches, attract thousands of visitors every year [1].

With an estimated population of 900 thousand inhabitants, distributed in 37 neighborhoods, divided into two large regions (North and South) by the Potengi River, it attracts companies from the most varied segments, many of them linked to the hotel and tourism sector. The sector that interests us, given the introduction above is the branch of Japanese restaurants in the municipality of Natal.

1.2 *Problem*

A large national chain of Japanese restaurants wants to open a branch in the city, and for that, it wants to know which location, which neighborhood has the best return possibilities for the company.

1.3 *Interest*

This type of problem is very recurrent not only for large companies that can use consulting to find the best place to install, but for medium and small companies around the world. In addition, the solution we found is a portable model for various types of business and for countless cities in Brazil, it is worth considering that we will have a problem of lack of information for smaller cities, where the primary information source we use will not have as much data, leading to a search for additional information in order to fill the possible gaps in these locations.

[1] Environment reported before the Covid, as tourism was one of the sectors most affected by the pandemic.

	BAIRRO	RM até 3 Sal	RM até 10 Sal	RM acima 10 Sal	Alfabetizado
0	Alecrim	90.95	8.20	0.68	93.70
1	Areia Preta	62.36	23.87	13.78	96.99
2	Barro Vermelho	46.99	43.56	9.40	97.80
3	Bom Pastor	98.72	1.23	0.04	82.44
4	Candelária	50.75	39.89	9.33	97.61
5	Capim Macio	43.67	45.94	10.37	97.98
6	Cidade Alta	77.39	19.13	3.23	90.49
7	Cidade da Esperança	89.71	8.62	1.66	93.27
8	Cidade Nova	97.89	2.02	0.08	84.93
9	Dix-sept Rosado	95.67	4.19	0.15	88.33
10	Felipe Camarão	98.57	1.29	0.11	83.09

Table 1: A extract from Natal neighborhoods.

2 DATA ACQUISITION AND CLEANING

2.1 *Data Sources*

In order to answer which is the best neighborhood for a Japanese restaurant to install, we used web scraping on the website containing information about the hometown on wikipedia in order to list which neighborhoods exist in the target city. Through an analysis of the city's master plan (urbanistic) on the city hall website, we were able to compile, in addition to the list of existing neighborhoods, data on the remuneration and literacy level of the residents of each neighborhood, which is why we stopped using the return from wikipedia webscrapping to use the csv file (Table 1) from the city hall website.

The first column of the file contains the name of each neighborhood, the second column ('RM up to 3 Sal') contains the percentage of people with low income in the locality, as well as the third ('RM up to 10 Sal') and fourth ('RM above 10 Sal') columns represent the percentage of people in the middle and upper classes respectively. The fifth column represents the number of people who have at least primary education and are residents of the neighborhood. Adding this additional information that Api from foursquare does not allow us was very important, since in our country the consumer profile of this restaurant is people with high education and average income.

The list of neighborhoods was obtained by reading the csv file of the city hall, but to start doing searches for establishments via foursquare it is necessary to have a reference gps position for each neighborhood. We got this information through the library **geocoder**, where having as input argument a string with desired address, such as *Ponta Negra, Natal, RN, Brazil*, it returns as output a gps position referring to that address , as in the figure 1.

From this, we survey which venues are contained in each neighborhood, using foursquare services, noting that there is a larger proportion of establishments that are not registered in this tool in relation to American cities, as the application is not so widely used in Brazil. One way to improve this

```
0 Alecrim -5.7988797 -35.2188577
1 Areia Preta -5.7869014 -35.1902633
2 Barro Vermelho -5.7978488 -35.2092207
3 Bom Pastor -5.8094982 -35.2462143
4 Candelária -5.8437235 -35.2214483
5 Capim Macio -5.8576398 -35.2014489
6 Cidade Alta -5.7856388 -35.20926
7 Cidade da Esperança -5.8250648 -35.235023
8 Cidade Nova -5.8372291 -35.2389607
9 Dix-sept Rosado -5.8089827 -35.2272512
10 Felipe Camarão -5.8257002 -35.2547225
11 Guarapes -5.8366106 -35.271913
12 Igapó -5.7690001 -35.2618893
13 Lagoa Azul -5.7233255 -35.2614748
14 Lagoa Nova -5.8226937 -35.2126403
15 Lagoa Seca -5.8073333 -35.2106714
16 Mãe Luiza -5.7950652 -35.1862162
17 Neopólis -5.8666079 -35.2088062
18 Nordeste -5.7930033 -35.2444009
19 Nossa Senhora Apresentação -5.7465241 -35.27878
20 Nossa Senhora de Nazaré -5.8174362 -35.2313961
21 Nova Descoberta -5.8269204 -35.1985475
22 Pajuçara -5.7372347 -35.2360433
23 Petropólis -5.7845495 -35.1984438
24 Pitimbu -5.8633093 -35.2324324
25 Planalto -5.8521764 -35.2568875
26 Ponta Negra -5.8846468 -35.1775119
27 Potengi -5.7505819 -35.253447
28 Praia do Meio -5.7789823 -35.194817
29 Quintas -5.7972181 -35.2316702
30 Redinha -5.7540506 -35.2152587
31 Ribeira -5.7767142 -35.2062156
32 Rocas -5.7737244 -35.1997909
33 Salinas -5.7756019 -35.2293544
34 Santos Reis -5.9152014 -35.269882
35 Tirol -5.8032096 -35.2002054
```

Figure 1: A GPS information retrieved by Geocoder library

query would be to use the google api, unfortunately a paid service that does not meet our educational purpose at the moment.

2.2 *Data Cleaning*

With the acquisition of these different databases, we cleaned up the data before integrating the data set in order to correct the existing problems of capturing this information. In retrieving the location information of the neighborhoods, the Santos Reis neighborhood had the GPS position erroneously assigned (figure 2. Due to previous knowledge, as a resident of the city, we know that it is a military neighborhood with little trade, we decided to remove it from the neighborhood assessment.

2.3 *Feature Selecting*

With the gps information returned from each neighborhood in the city of Natal, we made inquiries on the foursquare service to return the existing establishments within a radius of one thousand meters from the central position of the neighborhood.
There is a plethora of information returned by the foursquare service as shown in the table 2.

We extract only the name, category, and gps position of the establishments, referring to the neighborhood that was consulted at the time. So the dataset at this moment contains Neighborhood, Name, GPS and Category of the Establishment.
Also in this dataset we include the socioeconomic data of the inhabitants of each neighborhood, such as minimum wage, divided into three ranges, low, medium and high income, in addition to the literacy percentage of each neighborhood.

Figure 2: A Wrong gps information retrieved by Santos Reis neighborhood

Information	Description
Summary	Venues Summary
Venue Name	Store's name
Venue Location Address	Venue's Address
Venue Location lat	Latitude
Venue Location lng	Longitude
Venue Distance	Distance from store to gps point passed by query
Venue Postal Code	Postal Code
Venue City	City
Venue State	State
Venue Country	Country
Venue Formatted Address	Full venue's address
Venue Categories Name	Kind of venue
Venue Photos	Internet address containing store's pictures

Table 2: Information gathered by Foursquare.

In this way, the original dataset containing the information we want was just like in the 3 table.

3 EXPLORATORY DATA ANALYSIS

We can get an idea of the socio-economic distribution referring to Table 1, viewing the boxplot that represents the dataset, as shown in Figure 3.

We can observe that the first two columns are far more spread than the others columns, which can says that basic education don't differ too

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alecrim	-5.7988797	-35.2188577	Artkasa Design	-5.799280	-35.219787	Furniture / Home Store
1	Alecrim	-5.7988797	-35.2188577	Atacado da Pesca	-5.798980	-35.216604	Fishing Store
2	Alecrim	-5.7988797	-35.2188577	Cricricell Apple - Assistência Técnica Apple	-5.797596	-35.216645	Mobile Phone Shop
3	Alecrim	-5.7988797	-35.2188577	Docelândia	-5.796091	-35.216857	Candy Store
4	Alecrim	-5.7988797	-35.2188577	Padaria Santa Cecilia	-5.797099	-35.220302	Bakery

Table 3: A extract from dataset after foursquare queries.

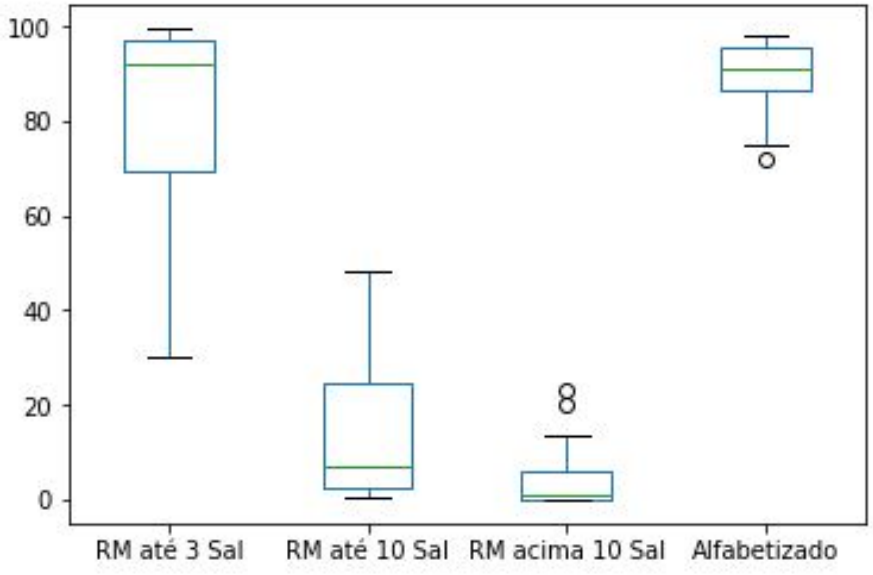

Figure 3: Boxplot representing the population from Natal

much from all neighborhood. It would be more interesting if we use more columns about education, like high school percentage, and graduated and pos graduated percentage of population.

With the data obtained from the city hall file, verified in Table 1, it was necessary to normalize this data, in order to later use distance-based clustering algorithms. Thus, we use the function **StandardScaler** in the columns referring to the average monthly income, as well as the indicative column of education, obtaining a new dataset according to Table 5.

We also visualized through the histogram, the distribution of establishments by neighborhood, according to the figure 4 below. Note that there is a considerable number of neighborhoods that the number of establishments returned was 100, indicating that the maximum number of establishments for each consultation in the foursquare api is one hundred, even though the request argument has changed to two hundred. In the next implementations we have to work around this limitation of the api.

In the figure 5, we see the distribution of restaurants in the municipality, as well as in the figure 6 we see the distribution of Japanese restaurants in the city. It is interesting to note that the distribution of restaurants in general is much more diffuse than the distribution of Japanese restaurants, which are mostly located in the economic and wealthiest center of the city, starting in the Petrópolis District in a straight line to the Candelária neighborhood ,

	RM até 3 Sal	RM até 10 Sal	RM acima 10 Sal	Alfabetizado
count	36.000000	36.00000	36.000000	36.000000
mean	80.881944	15.31750	3.784167	90.353889
std	21.039943	15.78388	5.818864	6.332244
min	30.120000	0.32000	0.000000	72.210000
25%	69.462500	2.53500	0.107500	86.750000
50%	92.010000	7.13500	0.845000	91.040000
75%	97.377500	24.66500	6.030000	95.447500
max	99.560000	48.24000	23.200000	98.000000

Table 4: Quartiles representing income and education from neighborhoords in NAtal

	Neighborhood	RM até 3 Sal	RM até 10 Sal	RM acima 10 Sal	Alfabetizado
0	Alecrim	0.485309	-0.457331	-0.541033	0.535920
1	Areia Preta	-0.892810	0.549536	1.742200	1.062853
2	Barro Vermelho	-1.633688	1.814707	0.978798	1.192584
3	Bom Pastor	0.859845	-0.905185	-0.652581	-1.267505
4	Candelária	-1.452445	1.578893	0.966598	1.162154
5	Capim Macio	-1.793722	1.967633	1.147862	1.221414
6	Cidade Alta	-0.168322	0.244970	-0.096587	0.021800
7	Cidade da Esperança	0.425537	-0.430344	-0.370227	0.467050
8	Cidade Nova	0.819837	-0.854424	-0.645609	-0.868701
9	Dix-sept Rosado	0.712826	-0.714992	-0.633409	-0.324150

Table 5: Normalizing income and education from neighborhoords in NAtal

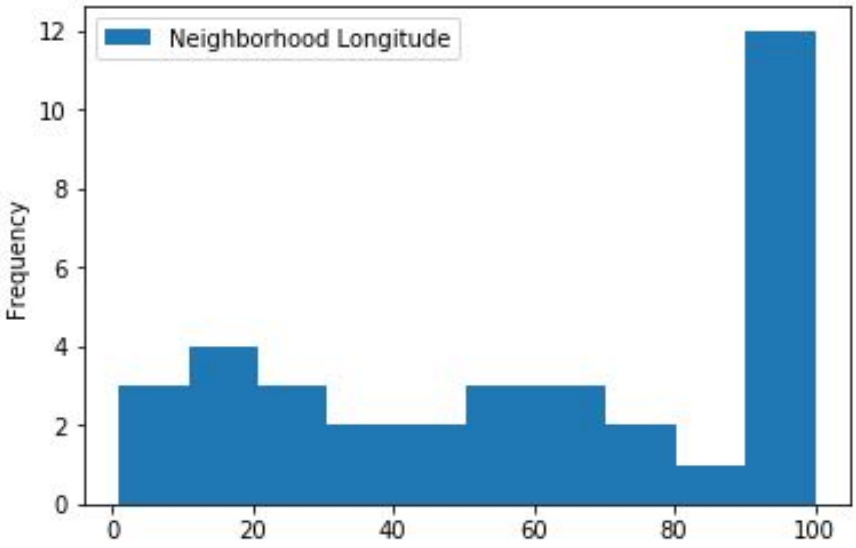

Figure 4: Histograma com as quantidades de estabelecimentos por bairro

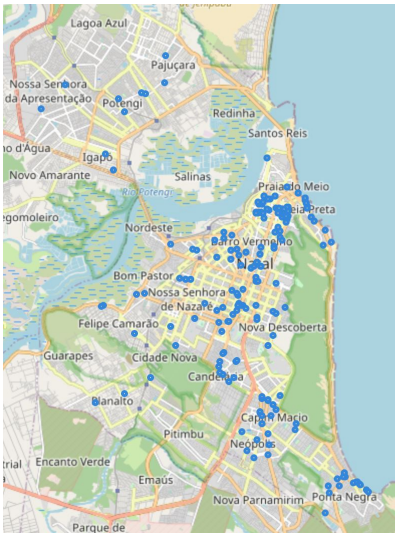

Figure 5: GPS position of all restaurants in Natal

	Neighborhood	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	Alecrim	Bakery	Bar	Snack Place	Supermarket	Acai House	Pharmacy	Department Store	Restaurant	Dessert Shop	Brazilian Restaurant
1	Areia Preta	Restaurant	Hotel	Bakery	Snack Place	Ice Cream Shop	Brazilian Restaurant	Shopping Mall	Italian Restaurant	Plaza	Seafood Restaurant
2	Barro Vermelho	Bakery	Dessert Shop	Pharmacy	Acai House	Furniture / Home Store	Bar	Electronics Store	Deli / Bodega	Clothing Store	Chinese Restaurant
3	Bom Pastor	Convenience Store	Bakery	Brazilian Restaurant	BBQ Joint	Gym	Gym / Fitness Center	Fast Food Restaurant	Warehouse Store	Burger Joint	Plaza
4	Candelária	Bar	Restaurant	Bakery	Pizza Place	Snack Place	Gym	Seafood Restaurant	Burger Joint	Sushi Restaurant	Café

Table 6: Top 10 Categories Venues Neighborhoods.

making a new straight to the Ponta Negra neighborhood, in an area similar to the letter L.

We created a function that generated a table that contained the 10 main categories of each neighborhood, in order to have a better idea of the profile of each Christmas area, as in the table 6.

4 MODELING - CLUSTERING

In order to cluster the neighborhoods to find the possible neighborhoods that would fit into a desirable aspect for Japanese restaurant installation, we used the k-means clustering algorithm.

In order to have more effectiveness in the solution, for the distance based algorithm it is necessary to normalize the algorithm's input data. The columns referring to the average salary and literacy level of the neighborhoods are not normalized, it was necessary to normalize them before running kmeans.

After cleaning data, selecting important characteristics and normalizing socioeconomic data, the dataset we modeled to use neighborhood clustering was just like in the table 7.

Where each row represents a neighborhood, and each column are the existing category types in the city of Natal. Each cell between this intersection represents the proportion among the establishments in the neighborhood. If the neighborhood of Alecrim had only bakery and restaurants, all other

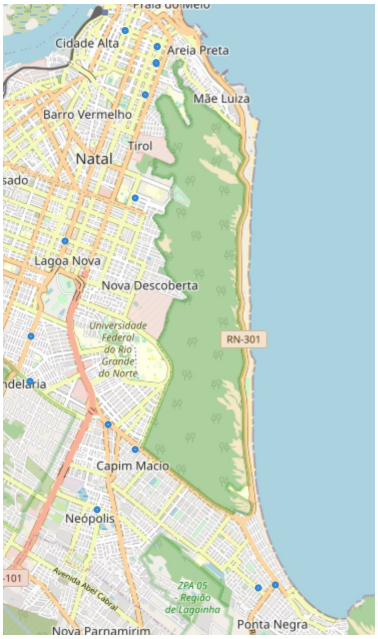

Figure 6: GPS position of all japanese restaurants in Natal

columns would have zero as information and the two columns would have the proportion between them, for example, 0.6 for bakery and 0.4 for restaurant if there were 12 bakeries and 8 restaurants in the locality.

In the first case, we use **3 centers for the clustering function**, and as we can see, the neighborhoods were well defined in terms of low, medium and high wellness, where the green center corresponds a poor locations, the red center corresponds a not so rich and not so poor neighborhoods, where the last color, purple, refers to a better and richer locations from natal.

We can see how the clustering was divided into 3 groups in the figure 8.

In the second case, we use **5 centers for the clustering function**, and as we can see, the neighborhoods were defined like this:

- Red Center (center 0): Contains **the richest area from city**, Tirol and Petropólis are in red in this clustering;
- Blue Center (center 2): **Upper Middle Class Neighborhoods**, contains Ribeira, Areia Preta, Barro Vermelho, Lagoa Nova, Candelária, Capim Macio;
- Green Center (center 3): **Lower Middle Class Neighborhoods**, contains Ponta Negra, Lagoa Seca, Neópolis, Pitimbu, Cidade Alta, Nova Descoberta and Praia do Meio. These neighborhoods areas in this category are clearly divided where some part of them are very rich, and other part is very poor. In cluster Blue, these differences are not so big as we can see here;
- Purple Center(center 1): Peripherical with **low income and low number of venues in city**, contains almost all west and north locations except Rocas and Mae Luiza;
- Orange Center (center 4): Contains Guararapes and Salinas neighborhood, **the poorest locations in town**;

	Neighborhood	Acai House	Accessories Store	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Astrologer	...	Vegetarian / Vegan Restaurant	Veterinarian	Volleyball Court	Warehouse Store	Waterfront	Whisky Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	Alecrim	0.050000	0.000000	0.00000	0.00	0.000000	0.000000	0.010000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.01	0.00
1	Areia Preta	0.000000	0.000000	0.01087	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.010870	0.000000	0.00	0.00
2	Barro Vermelho	0.030000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.01	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.010000	0.000000	0.00	0.00
3	Bom Pastor	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.076923	0.00	0.000000	0.000000	0.000000	0.00	0.00
4	Candelária	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
5	Capim Macio	0.030000	0.010000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.010000	0.000000	0.00	0.01
6	Cidade Alta	0.010000	0.000000	0.00000	0.00	0.050000	0.010000	0.000000	0.01	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
7	Cidade Nova	0.052632	0.000000	0.00000	0.00	0.000000	0.000000	0.052632	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
8	Cidade da Esperança	0.024691	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.012346	0.00	0.000000	0.000000	0.000000	0.00	0.00
9	Dix-sept Rosado	0.024691	0.012346	0.00000	0.00	0.000000	0.000000	0.012346	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
10	Felipe Camarão	0.024390	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
11	Guarapes	0.250000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
12	Igapó	0.000000	0.020408	0.00000	0.00	0.000000	0.000000	0.020408	0.00	0.00	...	0.000000	0.00	0.000000	0.061224	0.00	0.000000	0.000000	0.000000	0.00	0.00
13	Lagoa Azul	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
14	Lagoa Nova	0.010000	0.000000	0.00000	0.01	0.000000	0.000000	0.000000	0.00	0.00	...	0.010000	0.01	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.03	0.00
15	Lagoa Seca	0.020000	0.010000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.01	0.000000	0.000000	0.000000	0.01	0.00
16	Mãe Luiza	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
17	Neopólis	0.011111	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
18	Nordeste	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
19	Nossa Senhora Apresentação	0.000000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
20	Nossa Senhora de Nazaré	0.010000	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.01	...	0.000000	0.00	0.000000	0.010000	0.00	0.000000	0.000000	0.000000	0.01	0.00
21	Nova Descoberta	0.018868	0.000000	0.00000	0.00	0.000000	0.000000	0.018868	0.00	0.00	...	0.018868	0.00	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.00	0.00
22	Pajuçara	0.051282	0.000000	0.00000	0.00	0.000000	0.000000	0.000000	0.00	0.00	...	0.000000	0.00	0.000000	0.051282	0.00	0.000000	0.000000	0.000000	0.00	0.00

Figure 7: Final Dataset input for Kmeans Clustering.

In the third case, we use **7 centers for the clustering function**, and as we can see, the neighborhoods were defined like this:

- Dark Blue Center (center 2): Contains **the richest area from city**, Tirol and Petropólis are in red in this clustering;
- White Cream Center (center 5): **Upper Middle Class Neighborhoods**, contains Barro Vermelho, Lagoa Nova, Candelária and Capim Macio;
- Light Blue Center (center 3): **Lower Middle Class Neighborhoods**, contains Pitimbu, Neopólis, Ponta Negra, Nova Descoberta, Lagoa Seca, Cidade Alta and Praia do Meio;
- Purple Center(center 1): Contains Ribeira and Areia Preta neighborhood;
- Red Center (center 0): Contains Potengi, Cidade Nova, Nossa Senhora da Nazaré, Alecrim and Rocas neighborhood, **comercial low class areas**;
- Orange Center (center 6): Contains Lagoa Azul, Pajuçara, Nossa Senhora da Apresentação, Igapó, Redinha, Mãe Luiza, Planalto, Felipe Camarão, Bom Pastor, Nordeste, Quintas, Dix-Sept Rosado and Cidade da Esperança neighborhood, **locations very poor in our city**;
- Green Center (center 4): Contains Guararapes and Salinas neighborhood, **the poorest locations in town**;

5 CONCLUSIONS

Resuming our study for this kind of problem, we have:

- In Brazil, there is a specific kind of public that usually eats Japanese food;

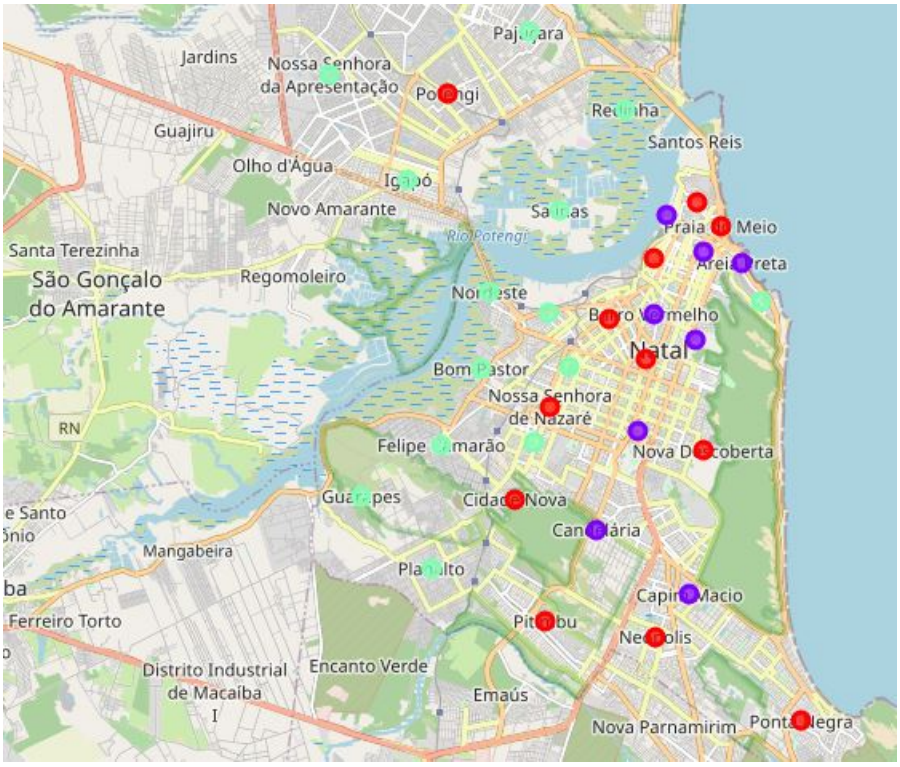

Figure 8: Clusterizing neighborhoods with 3 centers

- We have found data about income and education per capita, in each neighborhood, by collecting information from Town City Hall;
- The foursquare APIs was used in order to get local piece of information and make a neighborhood profile. Therefore, we have obtained the best location area for a new Japanese restaurant;
- It was realized that the main area that has Japanese restaurants is located in 'economic central spine in town', a L-shape area situated between Areia Preta and Ponta Negra districts, including Candelaria;
- Many clustering attempts were successfully used to group neighborhoods and find the best spot to install a restaurant. We have tried clustering with three, five and seven centers, and we noticed the following standard:
  - Petropolis and Tirol were always in the same center cluster and it means that we could choose any of them to install the restaurant;
  - Neighborhoods located in the West and North Zones are not recommended to install a Japanese restaurant, since they basically have the same profile: low income, low education and few venues in this area;

Finally, checking the distribution of Japanese restaurants in the city, according to the figure 6 and having discovered the path of the richest economic neighborhoods in the city, in an area that comprises the existing neighborhoods between Petrópolis - Candelária - Ponta Negra, we recommend installing a new Japanese restaurant between the neighborhoods of

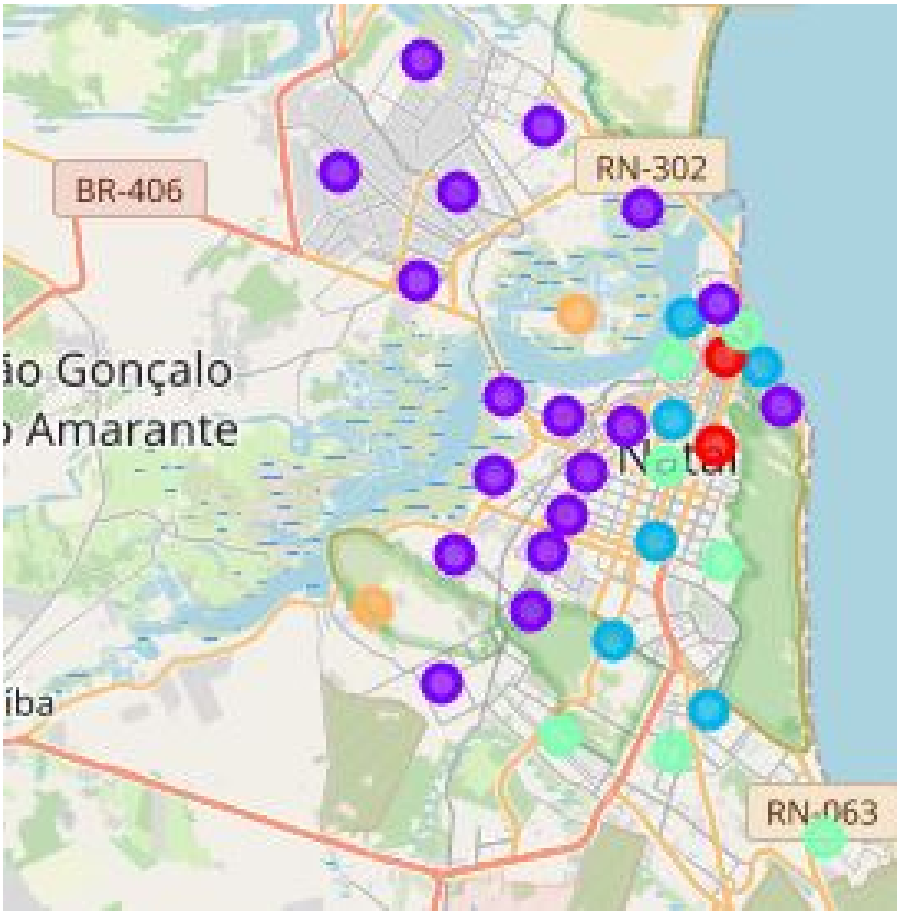

Figure 9: Clusterizing neighborhoods with 5 centers

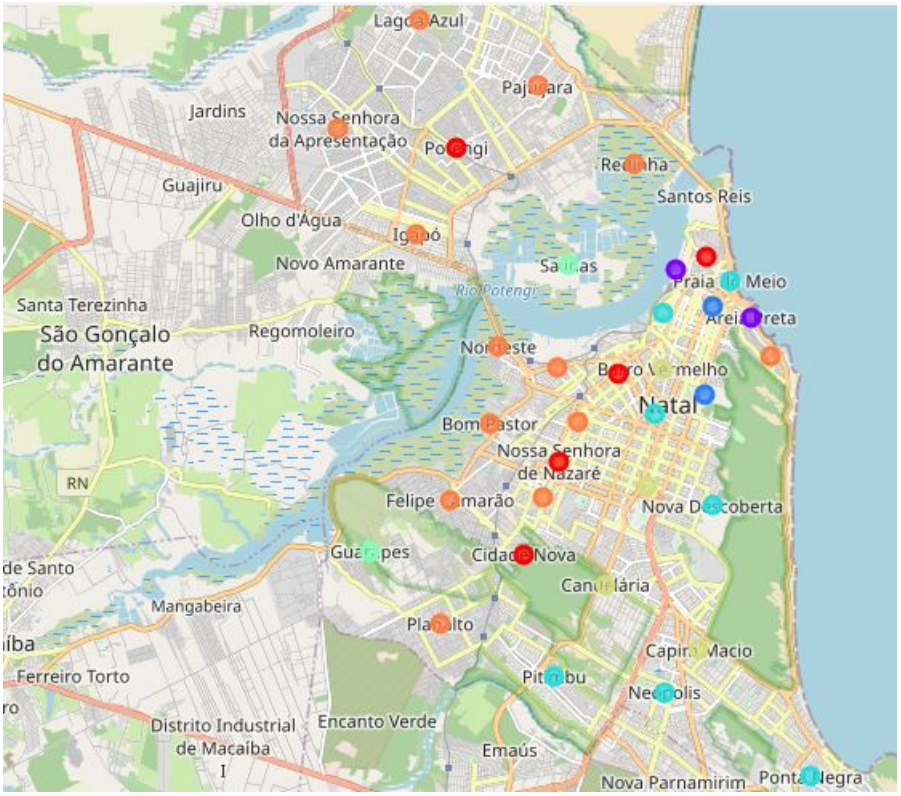

Figure 10: Clusterizing neighborhoods with 7 centers

Ponta Negra and Capim Macio, as there is a large region between these neighborhoods with no restaurant in this culinary style.

6 FUTURE DIRECTIONS

In order to give more assertiveness in solving this problem, we can retrieve the positions through the postal code, since this shape is more accurate than just one position per neighborhood, as a neighborhood can have several postal codes, depending on its size.
In addition, we would have to add more information than foursquare gives us, performing webscrapping on the return of google maps to find the largest number of available establishments, because as reported, the number of Japanese restaurants returned by google is higher than that foursquare returns.

In order to generalize this study to other types of restaurants, we must also raise the specific consumer profile of each gastronomy, as well as acquire more social and economic information about the residents of the neighborhoods.