Classifying Rio’s neighbourhoods according to restaurant types

Lucas das Dores

December 24, 2020

1. Introduction

1.1. Background

Rio de Janeiro is the second largest city in Brazil in terms of population with more than 6 million inhabitants. It officially has 163 neighbourhoods dividing its more than 1200 km² of area. It is natural to think that different neighbourhoods will have different types of restaurants due to different factors such as geographical position and socio-economic profile of its residents.

It is unquestionably valuable for new businesses to understand which type of food services are catered to each neighbourhood to make more informed decisions on their business strategies. Profiling these neighbourhoods can help gauge competition and market opportunities and further, to understand the consumer tendencies of each neighbourhood.

1.2. Business Problem

Explicitly, our aim is to classify the neighbourhoods of Rio according to the types of restaurants and bars we can find on it. More precisely we will try to answer the following questions.

- Can we determine the most common types of bars and restaurants of each neighbourhood?
- Can we find the tendency of a neighbourhood to have specific types of venues such as national or international cuisines, junk food or more slow-food types of restaurants?
- Can we find neighbourhoods with similar profiles of bars and restaurants?

1.3. Interest

The project is of interest of stakeholders aiming to open a new restaurant or business aiming to cater restaurants as it can gauge the profile of consumers of the area, the competition and possible market opportunities that might arise.

2. Data

2.1. Data Sources

I have collected the data of neighbourhoods of Rio using the Dataset of neighbourhoods of Rio from DataRio, that can be obtained here, to get the name of all neighbourhoods and the geolocation data of the limits of each neighbourhood.

Further, I have used the Foursquare data base to collect the name and geographical coordinates of food venues using the Foursquare API.

Finally, I have also used Google Maps to obtain coordinates to define centres for the searches performed through the API.

Before obtaining the venue data it was very important to try to visualize my search areas since given the varying size and shape of Rio's neighbourhoods, the search area could easily land on unpopulated natural parks and hills.

To do so I have defined an algorithm to determine a radius from the chosen centers which would still fall inside the neighbourhood limit so as to minimize the risk of getting venues from a neighbourhood and attributing it to another.

Afterwards, I drew these search areas on a folium map so as to check they covered a good area of each neighbourhood (with some possible overlaps). We can see the search area inside the red circles in the following map.

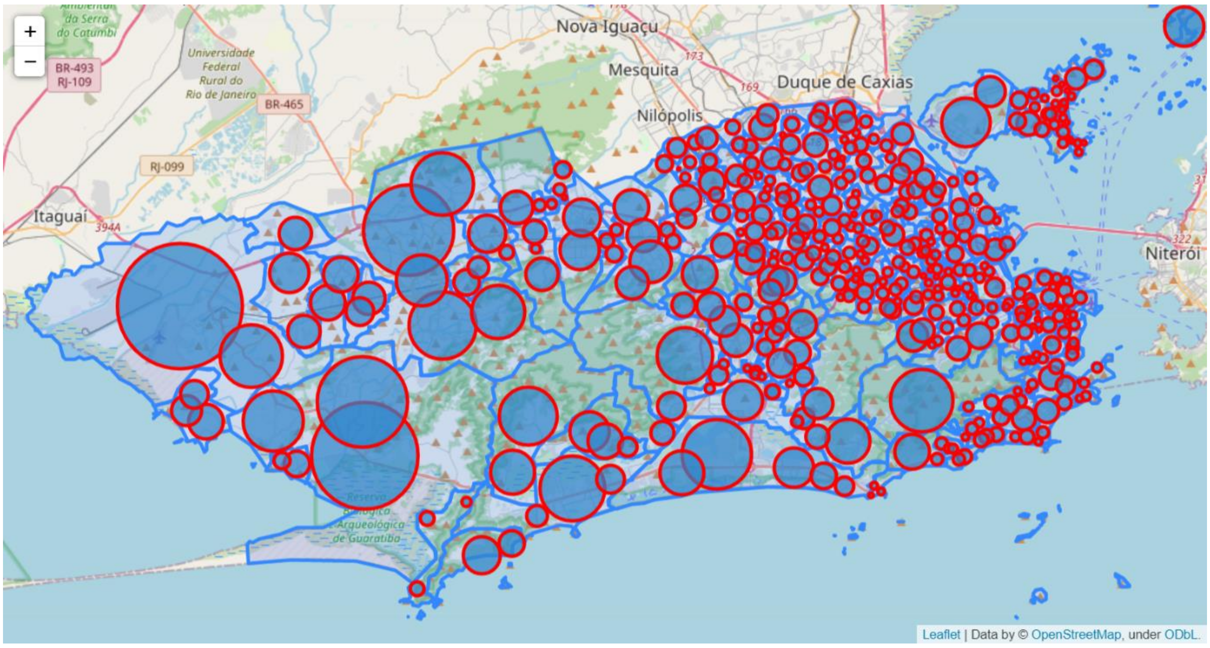

+
−
Ambiental da Serra do Catumbí
BR-493
RJ-109
Universidade Federal Rural do Rio de Janeiro
BR-465
RJ-099
Itaguaí
394A
Nova Iguaçu
Mesquita
173
169
Duque de Caxias
Nilópolis
Niterói
322
Reserva Biológica Arqueológica de Guaratiba
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 1. Search areas used on the API calls

I have then performed 10 searches with the search endpoint of the Foursquare API restricted to the food categories. The 10 search terms were the most common letters to start a word in Portuguese, that is D,A,E,C,P,S,O,M,N,Q. I have stored the results in a data frame together with their coordinates and venue type.

2.2. Data Cleaning and Preparation

Next, on the restaurant data frame we have identified a list of venues with undesired types such as Gas Stations and Convenience Stores. We promptly dropped them from the data frame. Moreover, since some of our search areas overlapped there were many repeated venues. The duplicates were removed by verifying entries with the same name *and* the same coordinates (since venues of a given chain all have the same name). The clean data frame returned 15085 unique venues as entries.

Once this was done we used one hot encoding to produce a data frame counting the number of venues and frequency of a given type of venue in the neighbourhood.

3. Exploratory Data Analysis

The data frame given by one hot encoding gave allowed to check the neighbourhoods with most restaurants and the ones with less as follows.

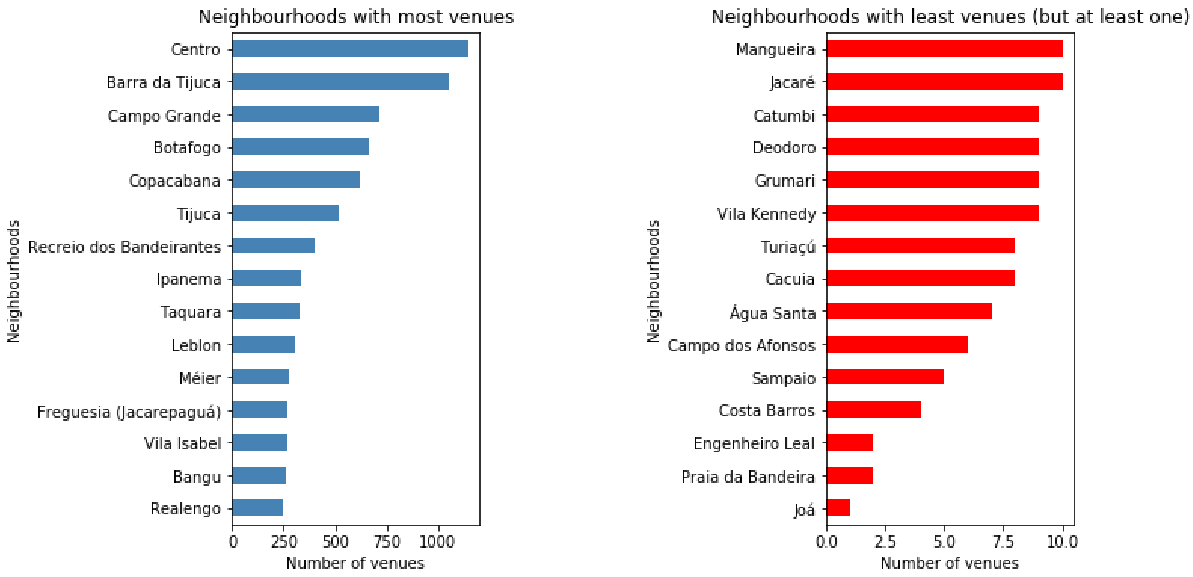
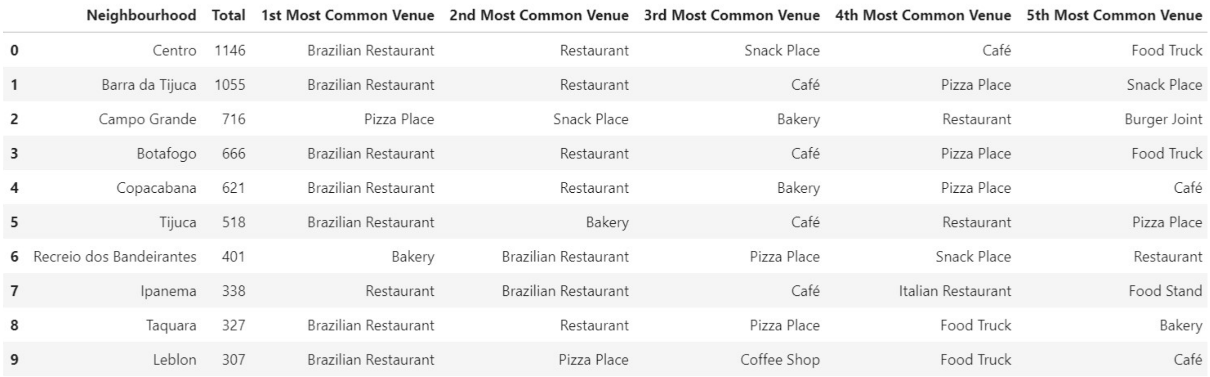


Figure 2. Neighbourhoods with most and least venues returned on the API calls

Once we computed the frequency of each type of restaurant we could also define a data frame containing the most common venues in each neighbourhood.



	Neighbourhood	Total	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Centro	1146	Brazilian Restaurant	Restaurant	Snack Place	Café	Food Truck
1	Barra da Tijuca	1055	Brazilian Restaurant	Restaurant	Café	Pizza Place	Snack Place
2	Campo Grande	716	Pizza Place	Snack Place	Bakery	Restaurant	Burger Joint
3	Botafogo	666	Brazilian Restaurant	Restaurant	Café	Pizza Place	Food Truck
4	Copacabana	621	Brazilian Restaurant	Restaurant	Bakery	Pizza Place	Café
5	Tijuca	518	Brazilian Restaurant	Bakery	Café	Restaurant	Pizza Place
6	Recreio dos Bandeirantes	401	Bakery	Brazilian Restaurant	Pizza Place	Snack Place	Restaurant
7	Ipanema	338	Restaurant	Brazilian Restaurant	Café	Italian Restaurant	Food Stand
8	Taquara	327	Brazilian Restaurant	Restaurant	Pizza Place	Food Truck	Bakery
9	Leblon	307	Brazilian Restaurant	Pizza Place	Coffee Shop	Food Truck	Café

The table above suggests that actually the most common type of venue in the neighbourhoods with most restaurants is the generically labelled Brazilian restaurant, with the notable exception of Campo Grande and Recreio dos Bandeirantes.

Other popular venues are generic restaurants, cafés and bakeries. Hence it wouldn't be a surprise that most clusters would have these as the most common types of venues and the clustering might have been poor with this was the only thing perceived. However this discovered inspired me to try to highlight the frequencies of certain collections types rather than individual types, which will be more thoroughly explained in the following section.

4. Methodology and Classification

One can use a K-means or DBSCAN clustering algorithms for clustering using the frequency data. I have chosen to use the K-means algorithm over the DBSCAN as the neighbourhoods seem to have similar frequencies when comparing the most frequent labels, such as Brazilian Restaurant, Pizza

Place or Snack Place, hence using the frequencies as coordinates results in that all neighbourhoods were “very close”. In this situation the DBSCAN algorithm would find only one cluster and fail to detect outliers.

Instead, the K-means algorithm had a chance to make a difference between less frequent labels and therefore would be more sensitive to different nuances in the profiles.

I have chosen the following subsets of venues to add to the analysis later on:

- *Brazilian cuisine venues*, e.g. Brazilian Restaurant, Pastelaria, Juice Bar, etc.
- *International cuisine venues*, e.g. Sushi Restaurant, Mexican Restaurant, Italian Restaurant, etc.
- *Niche cuisine venues*, e.g. Vegetarian / Vegan Restaurant, Deli / Bodega, Jewish Restaurant, Fondue Restaurant, etc.
- *Junk food venues*, e.g. Snack Place, Food Truck, Pizza Place, Burger Joint, etc.

For full list of the venues on each subset, please refer to the notebook.

On top of exploring the most common venues frequencies we will compare the frequencies of restaurants belonging to each of the subsets above to try to understand the clusters.

Before proceeding to the classification I have excluded every neighbourhood with less than 20 restaurants from the data frame and added them to a separate cluster. The reason is that with less than 20 restaurants it might be difficult to actually say what the profile of the neighbourhood means. Moreover, it is also an indicator that the search using the Foursquare API did not return sufficient data specially when the neighbourhood is densely populated such as the Complexo do Alemão neighbourhood. These neighbourhoods would only skew our classification to less accurate profiles. To solve this problem it might be needed to redo this classification using another API such as the one of Google Maps.

Next, we have decided on the number of clusters for the K-means algorithm on the following criteria:

- we want more than 2 clusters, so that we have minimally diversified profiles;
- we want clusters with more than eight neighbourhoods so the profiles are applicable to more neighbourhoods and;
- among these clusters we will take the one which will give smallest standard deviation between the number of neighbourhoods they possess, that is, the number of neighbourhoods on each cluster is reasonably spread.

Using this criteria we have decided on clustering the remaining neighbourhoods with in 7 clusters. Thus there would be a total of 8 clusters.

5. Results

We have named the 8 clusters returned by the algorithm with the name of the neighbourhoods with most restaurants within each cluster. The list of clusters is the following:

1. *Guaratiba cluster*: Guaratiba, Anchieta, Cosmos, Paciência, Pavuna, Parque Anchieta, Coelho Neto, Sepetiba, Cordovil, Santíssimo, Cidade de Deus, Ricardo de Albuquerque.
2. *São Cristóvão cluster*: São Cristóvão, Lapa, Santa Teresa, Jardim Botânico, Estácio, Saúde, S anto Cristo, Todos os Santos, Vasco da Gama, Manguinhos.
3. *Penha cluster*: Penha, Vila Valqueire, Praça Seca, Piedade, Jardim América, Pilares, Inhaúma, Jardim Sulacap, Osvaldo Cruz, Tauá, Parada de Lucas, Vista Alegre.

4. *Campo Grande cluster*: Campo Grande, Taquara, Bangu, Realengo, Santa Cruz, Anil, Engenho de Dentro, Vila da Penha, Pechincha, Curicica, Marechal Hermes, Guadalupe, Bento Ribeiro, Itanhangá, Vargem Pequena, São Conrado, Engenho Novo, Pedra de Guaratiba, Cascadura, Vidigal, Freguesia (Ilha), Vicente de Carvalho, Rocha, Magalhães Bastos, Engenho da Rainha.
5. *Barra da Tijuca cluster*: Barra da Tijuca, Botafogo, Copacabana, Tijuca, Recreio dos Bandeirantes, Ipanema, Leblon, Méier, Freguesia (Jacarepaguá), Vila Isabel, Jacarepaguá, Flamengo, Del Castilho, Lagoa, Jardim Guanabara, Laranjeiras, Grajaú, Gávea, Cachambi, Catete, Galeão, Humaitá, Benfica, Vargem Grande, Portuguesa, Padre Miguel, Tanque, Leme, Cosme Velho, Rocha Miranda, Abolição, Jardim Carioca, Alto da Boa Vista.
6. *Irajá cluster*: Irajá, Olaria, Andaraí, Quintino Bocaiúva, Ramos, Caju, Higienópolis, Cavalcanti, Brás de Pina, Senador Vasconcelos, Vigário Geral, Campinho.
7. *Centro cluster*: Centro, Madureira, Maracanã, Rio Comprido, Bonsucesso, Cidade Nova, Cidade Universitária, Maré, Gamboa, Penha Circular, Barra de Guaratiba, Glória, Praça da Bandeira, Inhoaíba, Gardênia Azul, Riachuelo, Senador Camará, Urca.
8. *Complexo do Alemão cluster*: Lins de Vasconcelos, Complexo do Alemão, Acari, Colégio, Encantado, Rocinha, Ribeira, Maria da Graça, Honório Gurgel, Jabour, Moneró, Cocotá, Parque Colúmbia, Paquetá, Jacarezinho, Camorim, Vaz Lobo, Vila Kosmos, Vila Militar, Bancários, São Francisco Xavier, Tomás Coelho, Barros Filho, Zumbi, Mangueira, Jacaré, Catumbi, Deodoro, Grumari, Vila Kennedy, Cacuia, Turiaçú, Água Santa, Campo dos Afonsos, Sampaio, Costa Barros, Engenheiro Leal, Praia da Bandeira, Joá, Pitangueiras, Gericinó.

We have created charts with the 10 most common venues within these clusters and their frequencies.

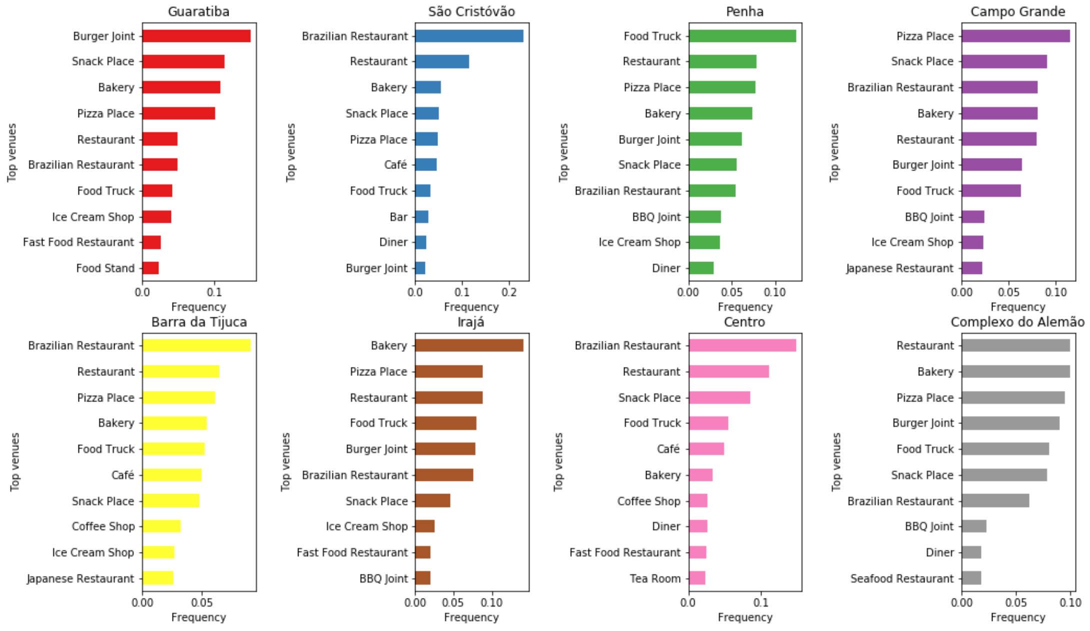


Figure 3. Frequency of most common venues in clusters

I have also produced a map with the geographical location of the clusters for comparison.

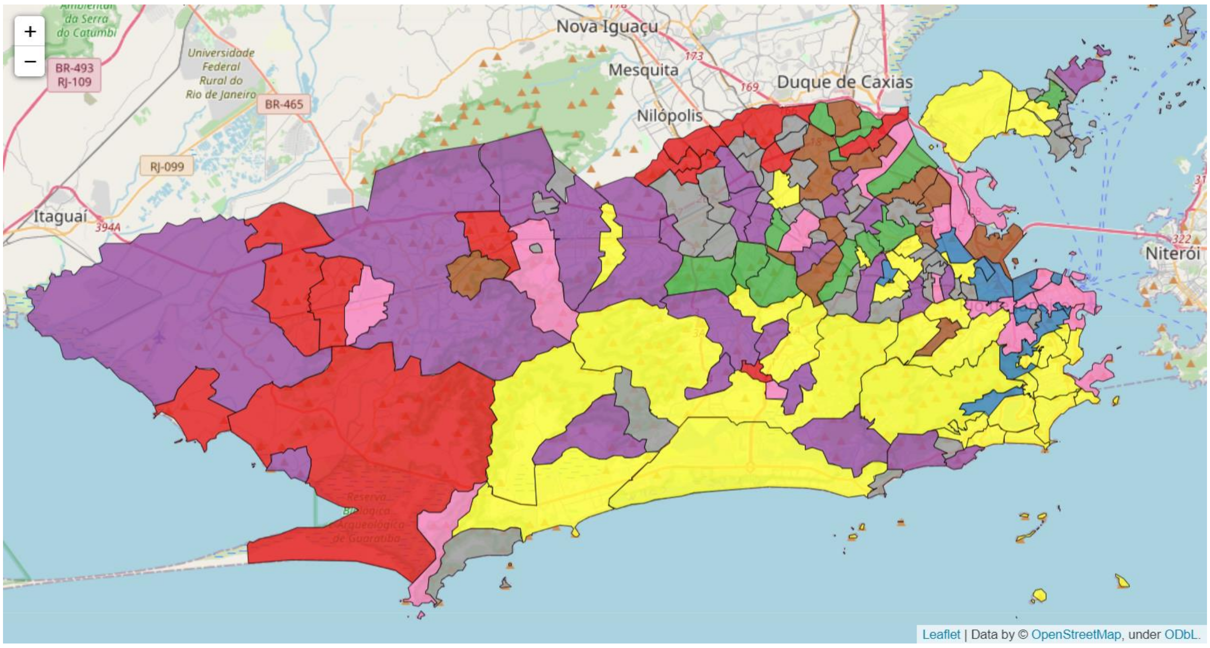
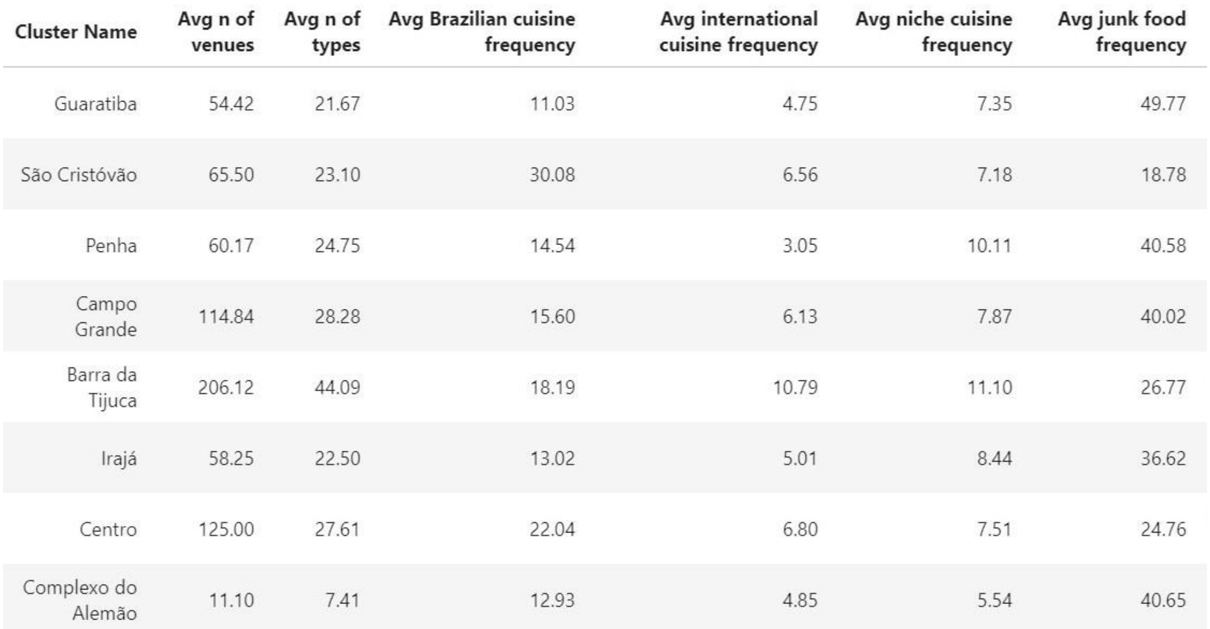

+
−
Ambiental da Serra do Catumbi
BR-493
RJ-109
Universidade Federal Rural do Rio de Janeiro
BR-465
RJ-099
Itaguaí
394A
Nova Iguaçu
Mesquita
173
169
Nilópolis
Duque de Caxias
Niterói
322
Reserva Biológica Arqueológica de Guaratiba
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 4. Map of clusters

Finally as mentioned before we have computed the average frequency of venues in the subsets of types selected in section 4 and added them to a final data frame of clusters.



Cluster Name	Avg n of venues	Avg n of types	Avg Brazilian cuisine frequency	Avg international cuisine frequency	Avg niche cuisine frequency	Avg junk food frequency
Guaratiba	54.42	21.67	11.03	4.75	7.35	49.77
São Cristóvão	65.50	23.10	30.08	6.56	7.18	18.78
Penha	60.17	24.75	14.54	3.05	10.11	40.58
Campo Grande	114.84	28.28	15.60	6.13	7.87	40.02
Barra da Tijuca	206.12	44.09	18.19	10.79	11.10	26.77
Irajá	58.25	22.50	13.02	5.01	8.44	36.62
Centro	125.00	27.61	22.04	6.80	7.51	24.76
Complexo do Alemão	11.10	7.41	12.93	4.85	5.54	40.65

Figure 5. Average frequency of cuisines.

According to the classification above the cluster with more average number of restaurants and variety is the Barra da Tijuca cluster, followed in average number of venues by the Centro and Campo Grande clusters.

It is worth noticing that the neighbourhoods of Barra da Tijuca, Centro and Campo Grande have many restaurants for different reasons:

- Barra da Tijuca neighbourhood is filled with shopping malls catered to the middle and upper class of Rio, many of the other neighbourhoods of its cluster also have this characteristic. Not incidentally there are many neighbourhoods of this cluster located on the so called 'south zone' of the city which also concentrates the main sights Rio is famous for. This cluster also seems to have more offer in international and niche cuisines, possibly for the reason outlined above.

- Centro neighbourhood is the historical city centre of Rio, its dynamics are different. Although there are sights in the city centre, a significant part of its buildings serve as offices and commercial structures. There is a very considerable contingent of people who commute to the city centre and need to eat out. It is very common to find 'self-service restaurants' catering the commuters. These are typical Brazilian types of restaurants on which a customer can freely choose the amount and type of food on a buffet and pay by the weight of their meal. It is possible that many of the venues labelled 'Brazilian Restaurant' fall in this category.

- Campo Grande is the most populous neighbourhood in Rio de Janeiro. Away from the usual sightseeing circuit of the city and geographically distant from the historical city centre, Campo Grande emerges as strong market on its own. One notices that the frequency of junk food venues on the cluster is radically higher than the previous two clusters. One possible reason is due to socio-economic factors, as the neighbourhood is far from the 'developed axis' of the city in the south zone and the residents might have lower income. An interesting remark is that even if that is the case the international and niche cuisine frequencies don't seem to fall much behind the ones of the historical city centre. In fact, niche cuisines are *more frequent* in the Campo Grande cluster rather than in the Centro cluster.

These are the clusters with the most neighbourhoods (excluding the Complexo do Alemão cluster, which we will address shortly).

Notice that the Penha, Guaratiba and Irajá clusters are similar in the average number of unique types per neighbourhood (between 20 and 25) and moreover they are similar to Campo Grande in that the frequency of junk food restaurants are all above 35% (reaching to staggering 49% in Guaratiba Cluster). The same observation made to Campo Grande might be valid to those clusters as they are all out of the main sightseeing circuit of Rio. Something that sets each of the clusters apart is the type of the most common venue (which for most of them is a junk food venue) which are Food Truck, Burger Joint and Bakery, respectively.

The São Cristóvão cluster has similarities with the Centro cluster with its offer of international and niche cuisines. Further, it is also geographically close to the Centro neighbourhood and its cluster. The differences are the slight lower average number of unique types (around 23), a higher concentration of typical Brazilian cuisine venues (around 30%) and lower concentration of junk food venues (below 20%).

Finally, the Complexo do Alemão cluster was created with neighbourhoods with less than 20 restaurants obtained on our search. Considering that we have made 10 API calls for each search radius and that most neighbourhoods had 3 search areas, obtaining less than 20 venues is underwhelming and reveals that there might be severe lack of data on these neighbourhoods on the Foursquare database. This is particularly evident when we are talking about very densely populated neighbourhoods such as Complexo do Alemão, Mangueira, Jacaré and Jacarezinho. We can't perform any conclusive analysis with this data and using a different database such as the one of Google as a replacement or complement might help us understand better the profiles of these neighbourhoods.

6. Further discussions

The classification above, although insightful, has a drawback, which is the exclusion of a significant part of the neighbourhoods of the city for the lack of data, those were all include in the Complexo do

Alemão Cluster. There are 41 neighbourhoods in this cluster, that is we have excluded from the classification algorithm something around 25% of the neighbourhoods of the city. To bypass this problem more data sources must be used. A possible solution is use the Google Maps API as a complementary source of the data. I believe this would bring significant improvement to the classification.

7. Conclusion

Despite the drawback outline above, with the available data we can see that the Barra da Tijuca and Penha clusters are the one with the most offer of niche cuisines, which includes, for instance, Vegetarian and Vegan restaurants. These might indicate two things:

- the consumers might look more for niche cuisines within these clusters which might help for a new restaurant in this profile to be established and;
- on the other hand this might also indicate there is a market opportunity for niche cuisines *outside* of these clusters.

Of course, complementary studies must be held in order to gauge the demand of such cuisines in the areas, as this one serves mostly to gauge their supply.

A similar conclusion could be drawn about the supply of international cuisines, whose largest frequency remains in the Barra da Tijuca cluster and less than 7% in every other cluster.

Finally, typically Brazilian cuisines are (rather unsurprisingly) frequent in all neighbourhoods (ranging from 10% to 30%). I believe it is safe to assume that there is always demand for this type of cuisine and, of course, that a lot competition will always be present. In that regard, other factors might be decisive on whether there is a market opportunity on opening a venue with this type of cuisine. For instance, the specific type of venue (e.g. would it be a self-service restaurant, a Pastelaria or a Juice Bar?) since there is more information on whether a cluster has a tendency to host more junk food or restaurant venues.