DISCOVERING VENUES IN SAO PAULO FOR NEWCOMERS: A CLUSTERING APPROACH

Ivan E. M. Kühne

June 20, 2020

This report summarizes the Capstone of the IBM Data Science Professional Certificate offered by Coursera. The work summarized here was developed using the Brazilian city of Sao Paulo as a study case and was focused in discovering venues for newcomers using the Foursquere API as the main data source and Data Science techniques to analyze the available data. The goal of this analysis was to develope a guide to newcomers grouping venues that tipicaly are interesting to people arriving in a new city, either as tourists or new residents.

The Section 1 presents the introduction to the study case. The Section 2 describes the data sources and how they were used. The Section 3 describes the methodology that was used. The Section 4 described the results that were achieved. The Section 5 presents the discussion about these results. Finally, the Section 6 presents the conclusions about the work that was developed and the future directions.

1 Introduction

This section presents the introduction of this report. The Subsection 1.1 presents the background about the Brazilian city of Sao Paulo. The Subsection 1.2 describes the problem that was addressed in this work. Finally, the Subsection 1.3 describes the target audience.

1.1 Background

Sao Paulo is the biggest Brazilian city and our main finnancial center, being considered by the Globalization and World Cities Research Network (GaWC) as an “alpha city” and a "highly sufficient city" in the classification for 2018 [1]. Accordingly to the estimatives for 2019, it had more then 12 million inhabitants in the main city, reaching almost 22 million inhabitants when the whole metropolitan region was considered [2]. The city has a wide ethnical and cultural diversty, being home to people born in 196 different countries [3] and attracting workers from all the Brazilian regions.

Due to this wide diversity, it has a plethora of options related to entertainment, like different types of night clubs, restaurants, museums and sports gymnasiums. Also, there are categories of each of these venues that are more common in determinated regions. For example, there are regions where it is easier to find restaurants specialized in Italian or Japanese food, due to the ethnical roots of many people that live there. So, it is advantageous for tourists and newcomers to be informed about the regions that are more related to their personal interests.

1.2 Problem

The Foursquare API was used to discover the venues (like restaurants and parks) that exist around the main subway stations in the city of Sao Paulo. The subway stations were used instead of the neighborhoods because they are simplier to be used by a tourist or newcomer as reference points. They are well documented and the maps of the subway lines that are produced by the city's subway company are cleaner and easier to be understood when compared to more traditional urban maps, as showed in the Figure 1.

The resulting data were subject to a classification analysis, using a clustering algorithm, resulting in the identification of the regions that are more rich in certains categories of venues. The outcome of the process was documented and enriched using some unstructured data sources, generating a final product able to help the tourists and newcomers to discover the regions that are more compatible with their specific interests.

Figure 1: Sample of a Sao Paulo Subway Lines Map [4].

1.3 Target audience

The target audience of this project are people that are coming to Sao Paulo, both as tourists and new residents, and are interested in informations about the regions that are more related to their personal interests in the entertainment area. So, they can enjoy a smoother experience, having more fun in their trip or adaptation to the new city, without having to lose much time to discover interisting places.

2 Data Description

This section presents the data sources used in this work. The Subsection 2.1 describes the usage of the Foursquare API as a data source and how it was conducted, while the Subsection 2.2 describes the other data sources that were used.

2.1 Use of the Foursquare API

In this project, the Foursquare API was used to fetch the data about the venues (like restaurants, night clubs and parks) that exist around the main subway stations of the city of Sao Paulo. This phase was similar to the work done in the previous weeks of the Casptone Project. As explaned in the Section 1, the subway stations were used because they are simplier points of reference to tourists and newcomers when compared to the neighborhoods names. So, each one became the central point of 500 meters (about one third of a mile) radius circle.

The coordinates (latitude and logitude) of each station were used in a Foursquare request to discover the venues that exist in a 500 meters radius. This process is show in the figures 2 and 3, using the Station Japao Liberdade, one of the most famous subway stations in Sao Paulo, as an example. First, as show in the Figure 2 we use the Nominatim geocoder to retrieve the coordinates of the station. In sequence, as show in the Figure 3, we use these coordinates to retrieve the venues that are located in the desired radius.

Figure 2: Retrieving the Coordinates of Estacao Japao Liberdade.

```
address = 'Estacao Liberdade Sao Paulo'

geolocator = Nominatim(user_agent = "myExplorer")

location = geolocator.geocode(address)

latitude = location.latitude
longitude = location.longitude

print('The geograpical coordinates of Estacao Liberdade - Sao Paulo are {}° and {}°.'.format(latitude, longitude))
```

```
The geograpical coordinates of Estacao Liberdade - Sao Paulo are -23.5627171° and -46.6392845°.
```

Figure 3: Retrieving the Venues around Station Japao Liberdade.

```
In [15]: liberdadeVenues = getNearbyVenues(names = namesList, latitudes = latitudeList, longitudes = longitudeList, radius = 500)
         liberdadeVenues.head(10)
```

Out[15]:

	Location	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Estacao Liberdade	-23.562717	-46.639285	Sukiya	-23.562227	-46.638883	Japanese Restaurant
1	Estacao Liberdade	-23.562717	-46.639285	Vila Itororó - Canteiro Aberto	-23.561751	-46.641022	Cultural Center
2	Estacao Liberdade	-23.562717	-46.639285	Paraty- RJ	-23.561243	-46.641310	Rest Area
3	Estacao Liberdade	-23.562717	-46.639285	EVS - Espaço Vida Saudável Herbalife	-23.564457	-46.639275	Tea Room
4	Estacao Liberdade	-23.562717	-46.639285	Lola Espaço da Beleza	-23.562928	-46.641887	Spa
5	Estacao Liberdade	-23.562717	-46.639285	Academia Kyokushin Liberdade	-23.562084	-46.638917	Martial Arts Dojo
6	Estacao Liberdade	-23.562717	-46.639285	Combu - Produtos da Amazônia	-23.562827	-46.641920	Food & Drink Shop
7	Estacao Liberdade	-23.562717	-46.639285	Bar do bezerra	-23.564884	-46.638773	Bakery
8	Estacao Liberdade	-23.562717	-46.639285	Smart Fit	-23.561014	-46.638630	Gym / Fitness Center
9	Estacao Liberdade	-23.562717	-46.639285	A.C. Camargo Cancer Center - Café Médicos	-23.565202	-46.637603	Coffee Shop

When the research is done to all considered regions, the resulting data about the venues were subject to a clustering analysis, enabling the discovery of regions that are more similar to each one and what categories of venues are more common in each of them. The final result of this process was summarized in this report relatory and in a presentation, enabling it to be understood by the target audience.

2.2 Use of Other Sources of Data

When the clustering analyis was done, some geographical and social non-structured data were consulted. These data were used to provide a brief description of the regions of each cluster, enriching the final product that is delivered to the target audience. This process will help to provide a more in-depth contextualization about what the tourists and newcomers can expect in each cluster.

These data were retrieved using two guides that were published about the neighborhoods in Sao Paulo. The first one was published by the magazine Veja Sao Paulo and presents the best neighborhood in each of the twenty-three categories that were considered, including shopping, happy hour, restaurants and beauty [5]. The other one was published by the company Tegra Developer and presents some information about each of Sao Paulo neighborhoods [6].

3 Methodology

The available data, retrieved from the Fousquare API[1] was analyzed using a Jupyter Notebook created on the IBM Watson Studio[2]. The points used to retrieved the data about the venues were the subway stations, that are popular reference points in the city of Sao Paulo. So, first were listed fifteen subway stations around city downtown to be used as starting points of the retrieving process. The coordinates (latitude and longitude) of these stations were retrieved using the *Nominatim* method available on the *geopy* library.

Using the coordinates retrieved in the former step, a first version of the Sao Paulo map was created, centered around its downtown coordinates. A sample of this map is presented in the Figure 4, where the blue circles are the locations of the chosen subway station, accordingly to the retrieved coordinates. Comparing these results with the previous kowledge about the city geography, it was possible to verify that this first map was correct.

Since the subway stations were correctly located, the next step was using the Foursquare API to retrieve the venues that exist around each one using a previously develop method. This method generates a *DataFrame*, a high-level data structure available at the *Pandas* library to efficiently handle large amounts of data, containing the venues that exist in a 500 meters radius (about one third of a mile) around each of the subway stations. This distance was chosen as a trade-off between coverage and being reachable on foot. The numbers about the venues found around each subway station are presented in the Table 1.

[1]Platform available at https://developer.foursquare.com/.
[2]Platform available at https://cloud.ibm.com/login.

Figure 4: Sample of the First Generated Map.

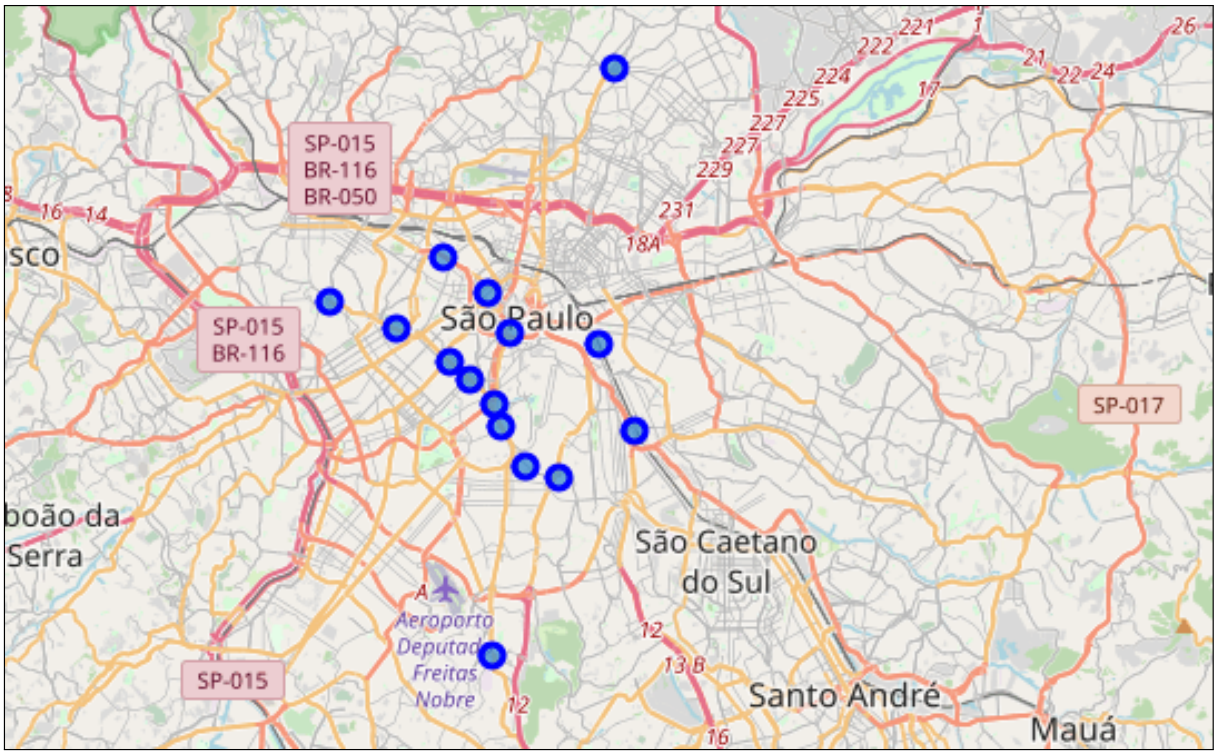


Table 1: Number of Venues by Subway Station

Subway Station	Number of Venues
Ana Rosa	59
Brigadeiro	60
Chacara Klabin	51
Clinicas	29
Ipiranga	56
Jabaquara	42
Japao Liberdade	55
Juventus Mooca	24
Marechal Deodoro	33
Paraiso	47
Republica	100
Santos Imigrantes	34
Trianon MASP	88
Tucuruvi	99
Vila Madalena	49

The results were grouped in a new *DataFrame* using the one-hot encoding apprach, so the number of venues in each venue category would become a new variable in this *DataFrame*. Due to this approach, it was possible to calculate the most common venue categories around the chosen subway stations and have a first impression about its surroudings. For example, the about the Station Japao Liberdade are presented in the Table 2, while the results about the Station Vila Madalena are presented in the Table 3.

Table 2: Top Five Most Common Venue Categories around Station Japao Liberdade

Order	Venue Category	Frequency
01	Japanese Restaurant	0.16
02	Sake Bar	0.09
03	Cosmetics Shop	0.05
04	Grocery Store	0.05
05	Bakery	0.04

Table 3: Top Five Most Common Venue Categories around Station Vila Madalena

Order	Venue Category	Frequency
01	Hostel	0.16
02	Burger Joint	0.08
03	Pharmacy	0.06
04	Gym	0.06
05	Thrift / Vintage Store	0.04

The results presented by the tables 2 and 3 are consistent with the previous knowledge about the city of Sao Paulo. The Japao Liberdade neighborhood has a large population of people with Asian ancestry, specially from Japan and China. So, we have a prevalence of venues related to Asian cultures, like Japanse restaurants and sake bars. On the other side, the Vila Madalena is a gathering point for the bohemian youth, so we have a prevalence of places like hostels, burger joints and vintage stores.

After this first inspection on the retrieved data, an object of the class *Kmeans* (available on the *Scikit-learn* library) was fitted to a copy of the grouped *Dataframe* to separate the subway stations in different clusters, based on their similarities (and dissimilarities) of available venues. After a few tests, the number of clusters was fixed in five, giving a good classification of the places without an excess of clusters. The results of this process are presented in the Section 4.

4 Results

First we have to describe the most common venues around each station, in decreasing order of frequency. These data are presented above.

a) Ana Rosa: Bakery, Japanese Restaurant, Brazilian Restaurant, Dessert Shop, Middle Eastern Restaurant, Café, Hotel, Candy Store, Restaurant and Burger Joint;

b) Brigadeiro: Japanese Restaurant, Middle Eastern Restaurant, Pizza Place, Coffee Shop, Ice Cream Shop, Hotel, Steakhouse, Cultural Center, Cosmetics Shop and Sushi Restaurant;

c) Chacara Klabin: Pizza Place, Dog Run, Burger Joint, Café, Gym/Fitness Center, Pet Store, Spa, Sushi Restaurant, Plaza and Pharmacy;

d) Clinicas: Brazilian Restaurant, Restaurant, Café, Pizza Place, Salon/Barbershop, Bus Stop, Snack Place, Motel, Sandwich Place and Coffee Shop;

e) Ipiranga: Jewelry Store, Dessert Shop, Chocolate Shop, Clothing Store, Food Court, Coffee Shop, Sporting Goods Shop, Steakhouse, Furniture/Home Store and Café;

f) Jabaquara: Bakery, Food Truck, Snack Place, Farmers Market, Gym/Fitness Center, Pharmacy, Martial Arts Dojo, Bookstore, Brazilian Restaurant and Pizza Place;

g) Japao Liberdade: Japanese Restaurant, Sake Bar, Grocery Store, Cosmetics Shop, Sushi Restaurant, Theater, Bakery, Bookstore, Nightclub and Asian Restaurant;

h) Juventus Mooca: Brazilian Restaurant, Café, Pizza Place, Diner, Restaurant, BBQ Joint, Gaming Cafe, Music Venue, Food Truck and Bakery;

i) Marechal Deodoro: Pizza Place, Restaurant, Coffee Shop, Brazilian Restaurant, Kosher Restaurant, Chocolate Shop, Street Art, Kids Store, Korean Restaurant and Bookstore;

j) Paraiso: Brazilian Restaurant, Café, Italian Restaurant, Sandwich Place, Middle Eastern Restaurant, Coffee Shop, Dessert Shop, Falafel Restaurant, Tennis Court and Gym;

k) Republica: Brazilian Restaurant, Italian Restaurant, Bar, Pizza Place, Coffee Shop, Theater, Café, Restaurant, Tea Room and Sandwich Place;

l) Santos Imigrante: Gym/Fitness Center, Dessert Shop, Burger Joint, Sushi Restaurant, Pharmacy, Movie Theater, Snack Place, Food Truck, Candy Store and Churrascaria;

m) Trianon MASP: Coffee Shop, Hotel, Italian Restaurant, Dessert Shop, Chocolate Shop, Cosmetics Shop, Restaurant, Health & Beauty Service, Sporting Goods Shop and Brazilian Restaurant;

n) Tucuruvi: Fast Food Restaurant, Pharmacy, Ice Cream Shop, Pizza Place, Dessert Shop, Clothing Store, Chocolate Shop, Market, Department Store and Snack Place;

o) Vila Madalena: Hostel, Burger Joint, Pharmacy, Gym, Farmers Market, Bakery, Thrift/Vintage Store, Ice Cream Shop, Health Food Store and Dessert Shop.

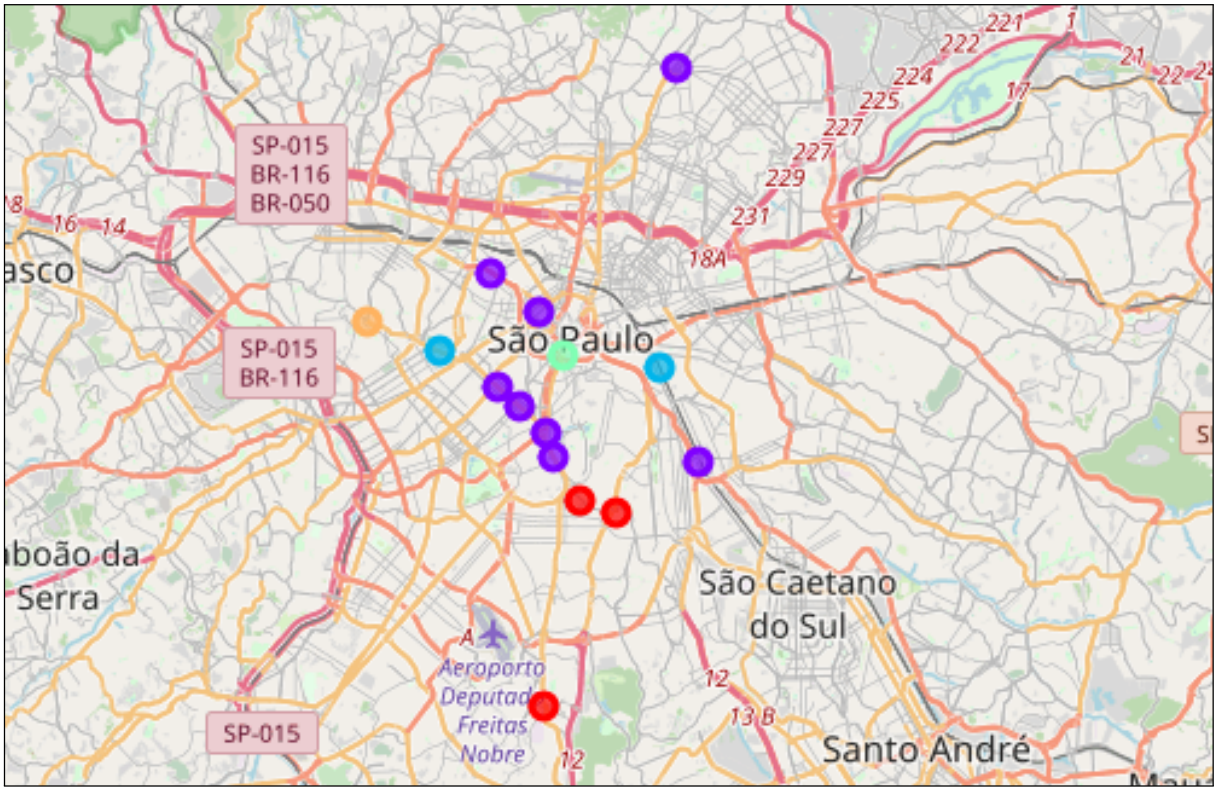
After the fitting of the *Kmeans* object, the subway station were groupped in five different clusters. The numbers of clusters was decided after a few texts, giving a good classification of the places without an excess of clusters. The results of the clustering process are show in the Table 4. This table shows that there eight subway stations in the Cluster 1, while the other ones have few stations.

The distribution of the clusters in the city of Sao Paulo is presented in the Figure 5. In this figure, the Cluster 0 stations are presented in red; the Cluster 1 stations are presented in purple; the Cluster 2 stations are presented in light blue; the Cluster 3 stations are presented in light green; and the Cluster 4 stations are presented in orange. Each cluster is described in further details in the oncoming subsections.

Table 4: Results of Clustering Process

Subway Station	Cluster
Ana Rosa	1
Brigadeiro	1
Chacara Klabin	0
Clinicas	2
Ipiranga	1
Jabaquara	0
Japao Liberdade	3
Juventus Mooca	2
Marechal Deodoro	1
Paraiso	1
Republica	1
Santos Imigrantes	0
Trianon MASP	1
Tucuruvi	1
Vila Madalena	4

Figure 5: Sample of the Clusters Map.



4.1 Cluster 0

The Cluster 0 includes the subway stations Chacara Klabin, Jabaquara and Santos Imigrantes. Looking for similarities between these surroundings, we see that, for people interested in food, there a lot of pizza places, food trucks, snack places. But, if you are feeling sick, there are a lot of pharmacies in this cluster. But, if you are interested in keeping yourself healthy before feeling sick, there are a lot of gyms and fitness centers.

4.2 Cluster 1

The Cluster 1 includes the subway stations Ana Rosa, Brigadeiro, Ipiranga, Marechal Deodoro, Paraiso, Republica, Trianon MASP and Tucuruvi. Looking for similarities between these surroundings, we see that many of them have Brazilian restaurants, restaurants, coffe shops and chocolate shops. At first glance, none of these are novelty for Brazilian people, being more atractive to people from outside the country that want to discover the local recipes.

But, looking a little deeper, we find some exotic places that may please some specific people, even when born in Brazil. Around the Station Marechal Deodoro, we can find restaurants specilized in kosher and Korean food that we do not see around the other stations. For people interested in these cuisines, it is the perfect place to explore. For people interested in Middle Eastern cuisine, the recomendation is to explore the surroundings of the stations Ana Rosa, Brigadeiro and Paraiso.

4.3 Cluster 2

The Cluster 2 includes the subway stations Clinicas and Juventus Mooca. Looking for similarities between these surroundings, we can find a prevalence of Brazilian restaurants, cafés, pizza places and restaurants in general. For people interested in another kinds of fun, we can see that around the station Juventus Mooca there are music venues and gaming cafes, venues that we don't find around the other subway stations.

4.4 Cluster 3

The Cluster 3 includes only the subway station Japao Liberdade. The top ten venue categories in this neighborhood are japanese restaurant, sake bar, grocery store, cosmetics shop, sushi restaurant, theater, bakery, bookstore, nightclub and asian restaurant. As presented before, this neighborhood congregates a lot of people with Asian ancestry, specially from Japan and China. For newcomers, this is a great place to try the Japanese and Asian food.

4.5 Cluster 4

The Cluster 4 includes only the subway station Vila Madalena. The top ten venue categories in this neighborhood are hostel, burger joint, pharmacy, gym, farmers market, bakery, thrift/vintage store, ice cream shop, health food store and dessert shop. As presented before, this neighboorhood is a gathering point for the bohemian youth. For newcomers that are adept of this lifestyle, it is a great place to meet people with similar interests in the hostels, burger joints, bakeries and dessert shops.

5 Discussion

Analysing the presented results, we can see differences in the nature of each generated clusters, as expected in the use of the clustering approach. Each cluster has a different atmosphere in the surroundings of the grouped subway stations, becoming a possible gathering point for people with different lifestyles and interests. This is specially true when talking about the clusters 3 and 4, that consist of only one station each and have venues that are distinct in relation to the ones that are common in the other clusters.

Looking deeper in the results, we could find that around the station Juventus Mooca there are some venues, like gaming cafes and music venues that are not found around the other subay stations, despite being grouped together with another station. However, with the station Clinicas, that is grouped in the same cluster, Juvetus Mooca share some categories of venues around it. On the other hand, speaking about the clusters 0 and 1, we can see more stations grouped an sharing some categories of venues.

6 Conclusion

In this work, I used the clustering approach to group fifteen well known subway stations of the Brazilian city of Sao Paulo based in the most common categories of venues that are located around them. The intent of this work was to be able to use these results as a guide to newcomers to the city, both as tourists and new residents, helping them to find places to have fun according to their main interests. The data about the venues was gathered using the Foursquare API and the Python programing language, that was used to apply the data cleaning and analysis techniques too. As a result of this approach, the fifteen stations were grouped in five differet clusters, each of one representing a different atmosphere and being capable of entertain people with specifics lifestyles and interests.

As future directions, the analysis techniques can be deepened to deal with some ambiguities in the naming of some categories of venues that lead us to think that some of them describe the some type of venues, despite having different names, like "Café" and "Coffe Shop". The same occurs with "Churrascaria", the Portuguese word for the restaurant that serves the Brazilian BBQ, and perhaps can be merged with the "BBQ Joint" category.

References

[1] GLOBALIZATION AND WORLD CITIES RESEARCH NETWORK (GaWC). GaWC City Link Classification 2018. https://www.lboro.ac.uk/gawc/world2018link.html, 2018. Online; acessed 07 February 2020.

[2] BRAZILIAN INSTITUTE OF GEOGRAPHY AND STATISTICS (IBGE). IBGE divulga as estimativas da população dos municípios para 2019. https://tinyurl.com/y3odhuwe, 2019. Online; acessed 07 February 2020.

[3] STATE OF SAO PAULO. As 10 menores comunidades estrangeiras de Sao Paulo. https://tinyurl.com/yd3rd62t, 2016. Online; acessed 07 February 2020.

[4] METROPOLITAN TRANSPORTS. Metropolitan Transport Network. http://www.metro.sp.gov.br/pdf/mapa-da-rede-metro.pdf, 2019. Online; acessed 07 February 2020.

[5] VEJA SAO PAULO. The Best Neghborhoods of the City in 23 Categories. https://vejasp.abril.com.br/cidades/bairros-campeoes-sao-paulo/, 2017. Online; acessed 08 February 2020.

[6] TEGRA DEVELOPER. Neghborhoods Guide: Best Neighborhoods in Sao Paulo. https://www.tegraincorporadora.com.br/sp/guia-de-bairro/, 2020. Online; acessed 08 February 2020.