2- Data Acquisition and Cleaning

2.1 – Data Source

Data was all acquired using the free version of Foursquare API. As they describe themselves, "*Foursquare* is the most trusted, independent location data platform for understanding how people move through the real world." Foursquare has a reliable database that can provide useful insights on several kinds of venues, like location, services, tips, ratings, reviews, pictures and etc. Since it's a database that its collaborative and relies on its users to properly visit the places, evaluate and then provide feedback on the website/app, it can sometimes lack information when the location itself does not update their data, or when it's not visited enough by frequent foursquare active users.

2.2 – Data cleaning

The API allowed me to explore and analyze different venues on the neighborhood. Since the primary objective was to list venues that could be interesting to tourists, search parameters like 'Hotel', 'Museum', 'Bar' where chosen to pull data and start building the data frames, defining the search radius for 8km around the local airport, Santos Dumont, which is very well placed around important areas of the neighborhood.

Some of the problems encountered on the data: important fields like address were not properly filled. Thankfully, the API provided with accurate coordinate information for all venues, and that allowed us to correctly plot all desired venues. The search also returned several information that would not be useful for our problem. After checking and cleaning the data, I noticed that for some reason the API was pulling venues from outside the delimited 8km radius. The 'Distance' column was very helpful in filtering those outliers and keeping the results adequate to our problem. Before dropping those results, further confirmation was done by plotting those coordinates and confirming their real location.

During the cleaning process, I've also used the search parameters to help me define the venues categories. This information would be useful later, both during data and cluster analysis and for proper map plotting and visualization.

	id	name	categories	referralId	hasPerk	location.address	location.lat	location.lng	location.labeledLatLngs
0	57e59302498e6a859d6dd2df	Mr. Fit Fast Food Saudável	[{'id': '4bf58dd8d48988d16e941735', 'name': 'F...	v-1619049055	False	Av. Mal. Floriano, 127	-22.909987	-43.171241	[{'label': 'display', 'lat': -22.909987, 'lng'...
1	56427127498e4309fc73811e	Joe's Pub Food Truck	[{'id': '4bf58dd8d48988d11b941735', 'name': 'P...	v-1619049055	False	R. do Catete	-22.926608	-43.176610	[{'label': 'display', 'lat': -22.9266075299422...
2	5e5bf05498c49900087f5703	Vou Street Food	[{'id': '52939a643cf9994f4e043a33', 'name': 'C...	v-1619049055	False	Bossa Nova Mall (Praça De Alimentação)	-22.914657	-43.166351	[{'label': 'display', 'lat': -22.914657, 'lng'...
3	5f5f6c9624ae6f671d9fe64b	Kitchen Asian Food	[{'id': '4bf58dd8d48988d142941735', 'name': 'A...	v-1619049055	False	Marina da Glória	-22.920310	-43.170498	[{'label': 'display', 'lat': -22.92031, 'lng':...
4	57424806498e91d7fce0daa8	Cogu Food Truck	[{'id': '4bf58dd8d48988d1cb941735', 'name': 'F...	v-1619049055	False	NaN	-22.916962	-43.173560	[{'label': 'display', 'lat': -22.9169617799874...

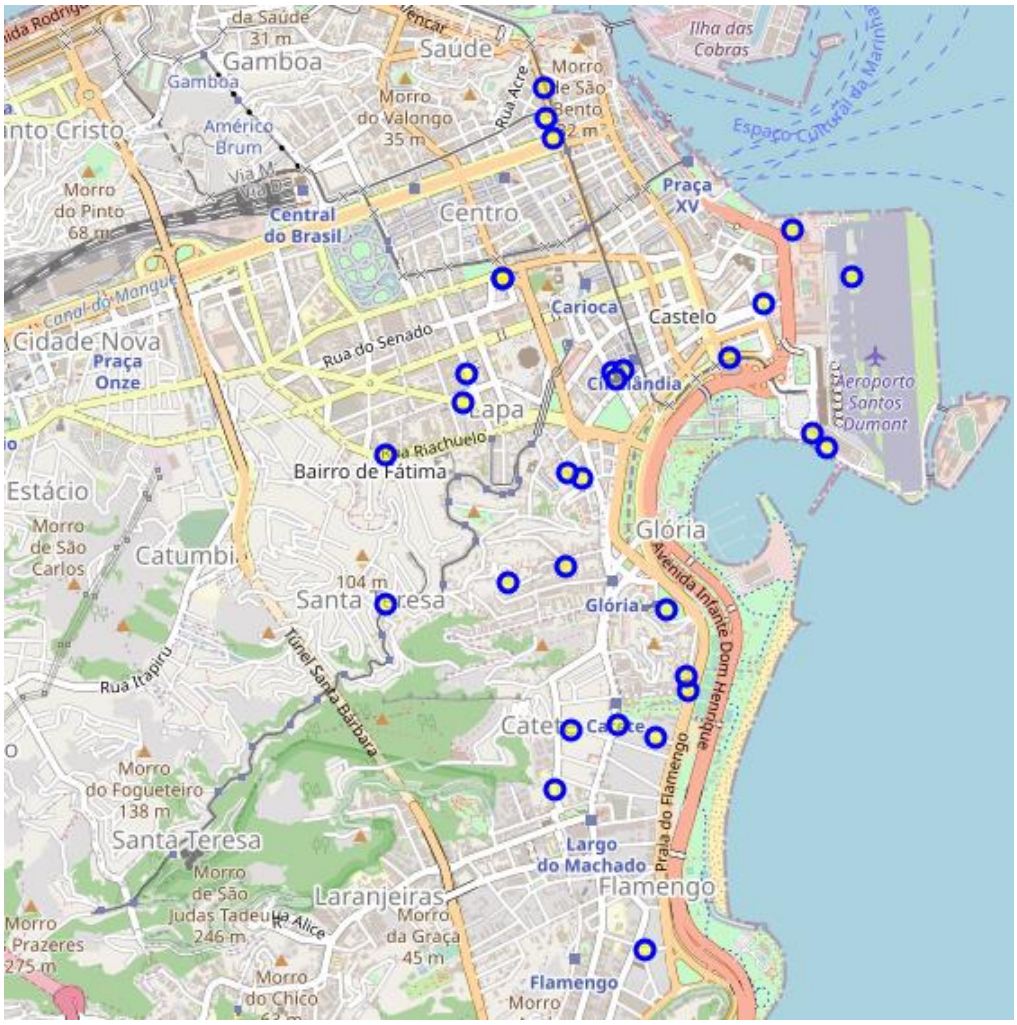
	name	latitude	longitude	distance	category
0	Mr. Fit Fast Food Saudável	-22.91	-43.1712	728	Food
1	Joe's Pub Food Truck	-22.9266	-43.1766	2180	Food
2	Vou Street Food	-22.9147	-43.1664	491	Food
3	Kitchen Asian Food	-22.9203	-43.1705	1248	Food
4	Rildy Carioca Food	-22.9058	-43.1748	1219	Food

Table of venues before and after cleaning, showing only the interesting variables for our analysis.

3- Exploring the data

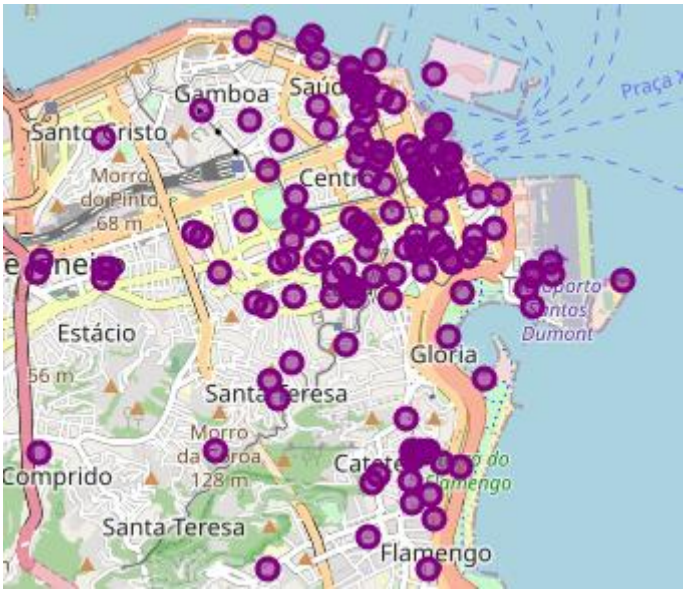
3.1 – Primary analysis

After cleaning and preparing the data frames, it was easy to plot and start thinking about the how I would proceed with the venues' clusters. I started by plotting the map with only the Hotels, to get a better idea of how they showed up in the map:



Initial plot of Hotel locations.

Hotel information was plotting correctly, and it was time to check the other venues data frame, with consisted of Restaurants/Food trucks and similar under the ‘Food’ category, Museums and art exhibitions under ‘Museu’, Coffee shops and similar under ‘Coffee’ and Drinks, nightlife and bars under ‘Drinks’. All information was checking and plotting correctly as well.



First plot of venues plotted on the map.

It was at this moment that I noticed the API was pulling results from a larger radio than intended. A quick correction by dropping those places based on the ‘distance’ column was enough to further clean the data and make the expected plot. Our final merged data frames had 159 venues, more than enough sample of the most popular and interesting venues in the neighborhood that would still allow me to look for the proper clusters.