Coursera Capstone Project

Analysis of Rio de Janeiro’s Centro venues

Renan Rocha

April 22, 2021

1- Introduction

1.1 - Background

Rio de Janeiro is one of the most well-known cities in Brazil, and one of the most visited tourist locations in Latin America. It’s famous for its culture, music, artistic expression and natural settings. It’s one of Brazil’s former capitals, and the Centro neighborhood is home of several historical sites and museums, with a very active nightlife. With so many options, it would be helpful to map interesting locations around the neighborhood in a way that could provide a general overview to tourists coming to the city, and enable them to make well-informed choices regarding their stay, based on their personal interests.

1.2 - Problem

The collected data should be used to provide meaningful information that could help a tourist vising the city to choose the best hotel to stay, based on the locations they decide to visit within walking distance and his preferences in other activities. There is also potential of identifying locations where investors would find good spots for opening new businesses based on local demands.

2- Data Acquisition and Cleaning

2.1 – Data Source

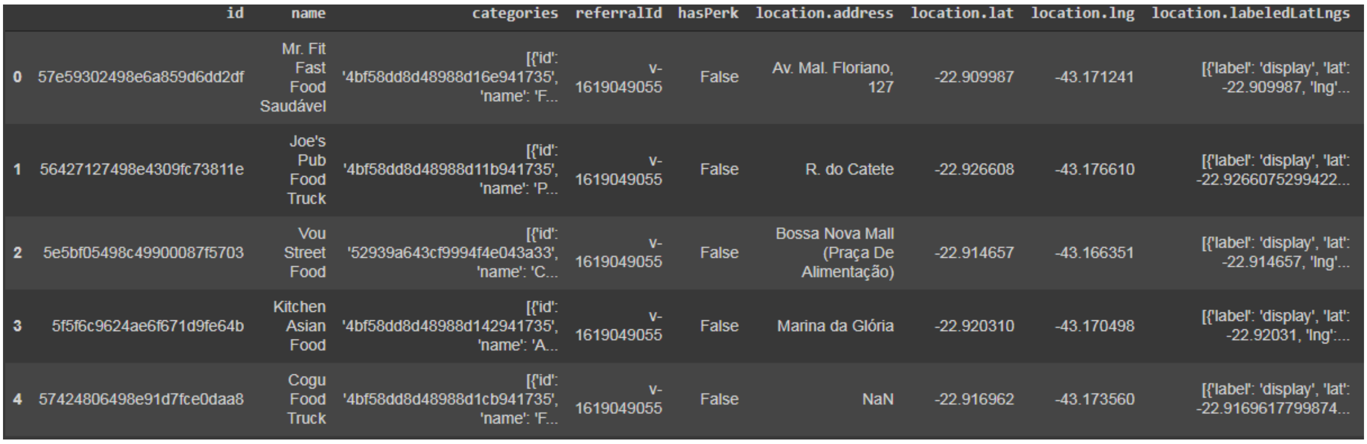
Data was all acquired using the free version of Foursquare API. As they describe themselves, “*Foursquare* is the most trusted, independent location data platform for understanding how people move through the real world.” Foursquare has a reliable database that can provide useful insights on several 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

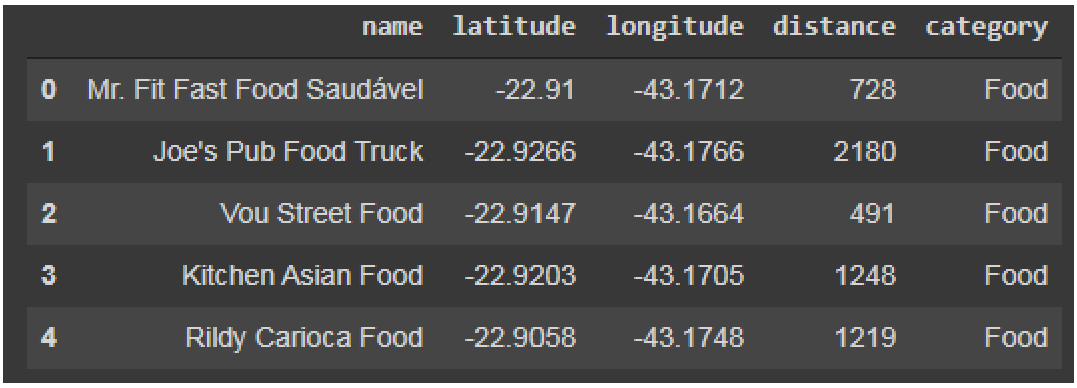
The API allowed me to explore and analyze different venues on the neighborhood. Since the primary objective was to list venues that could be interesting to tourists, search parameters like ‘Hotel’, ‘Museum’, ‘Bar’ where chosen to pull data and start building the data frames, defining the search radius 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...



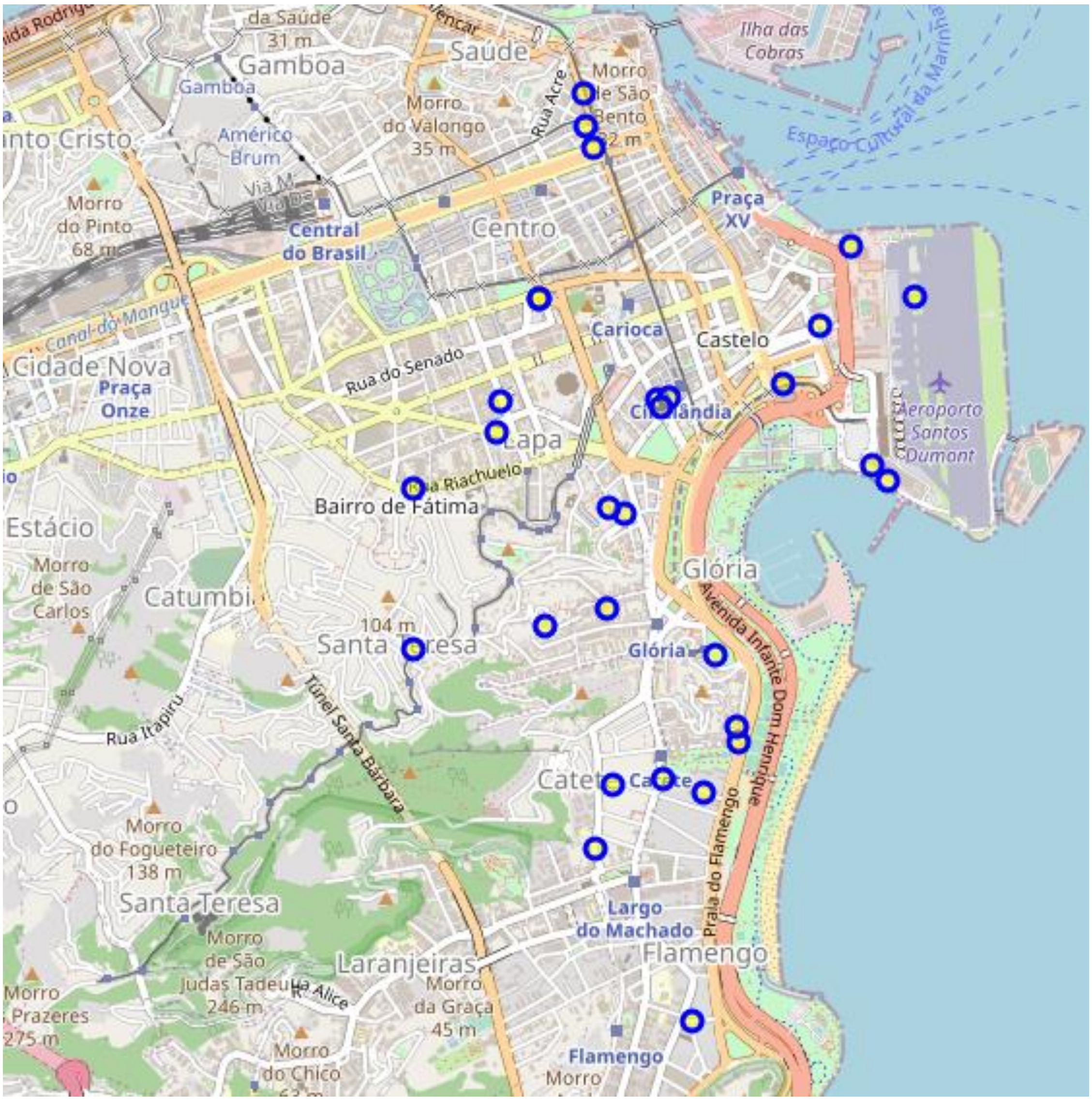
	name	latitude	longitude	distance	category
0	Mr. Fit Fast Food Saudável	-22.91	-43.1712	728	Food
1	Joe's Pub Food Truck	-22.9266	-43.1766	2180	Food
2	Vou Street Food	-22.9147	-43.1664	491	Food
3	Kitchen Asian Food	-22.9203	-43.1705	1248	Food
4	Rildy Carioca Food	-22.9058	-43.1748	1219	Food

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

3- Exploring the data

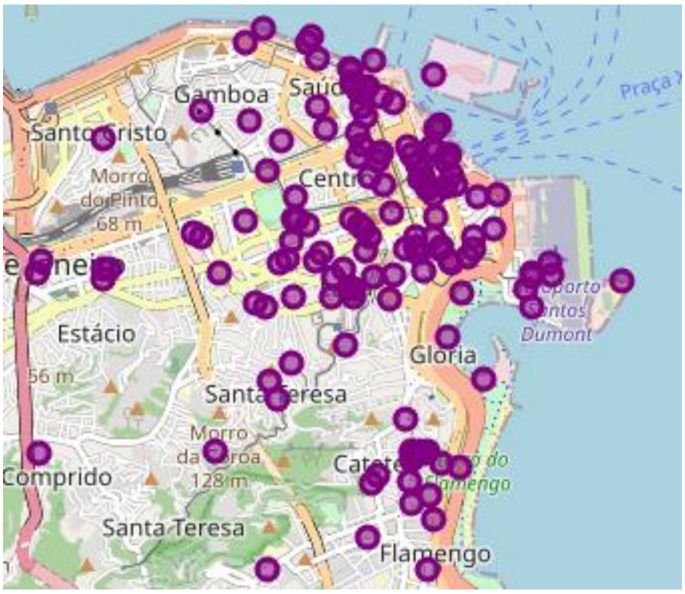
3.1 – Primary analysis

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.



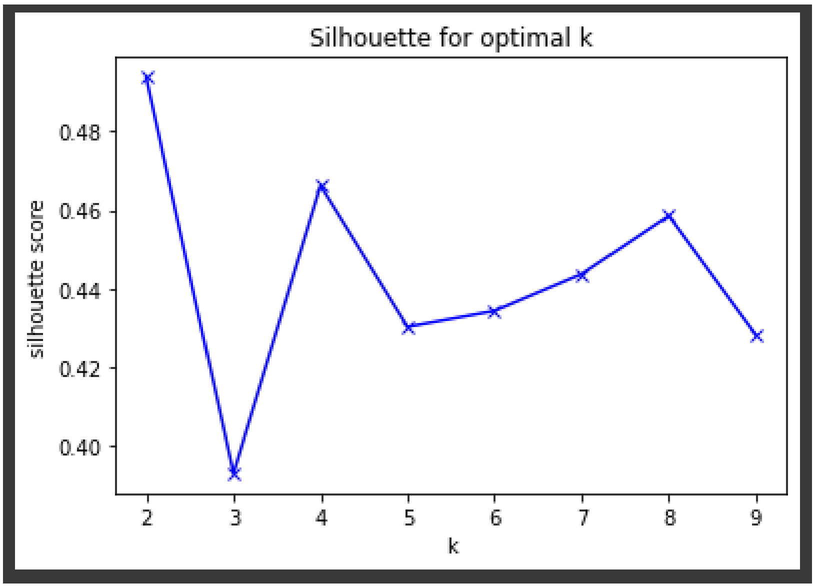
Initial detailed plot of venues on the map.

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.

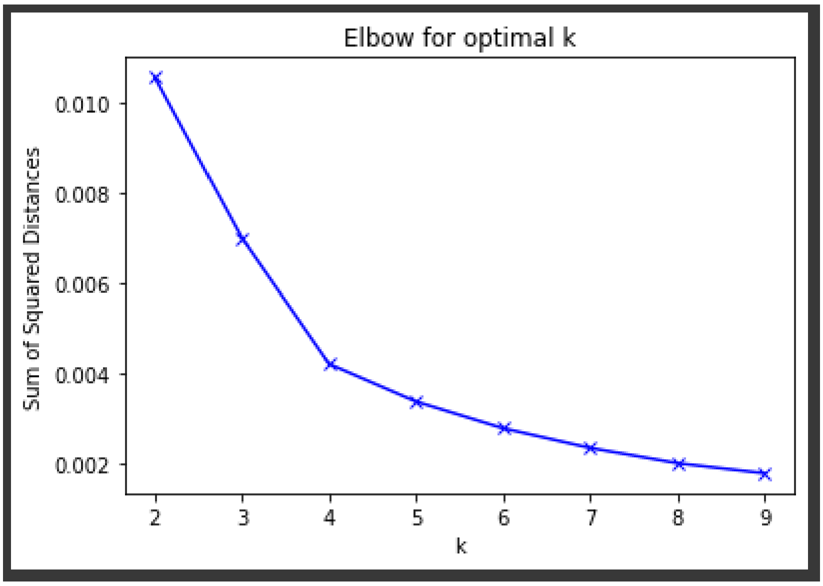
4- Cluster Analysis and further exploration

4.1 – Looking for the best number of clusters

To solve our clustering problem, I will use the cluster algorithm k-means. To help me define how many clusters the model should have, I've chosen to plot both the Silhouette score and Elbow method. They should provide me enough information to at least start visualizing the best option in this scenario. The Silhouette method will measure how similar a point is to its own cluster compared to other clusters.



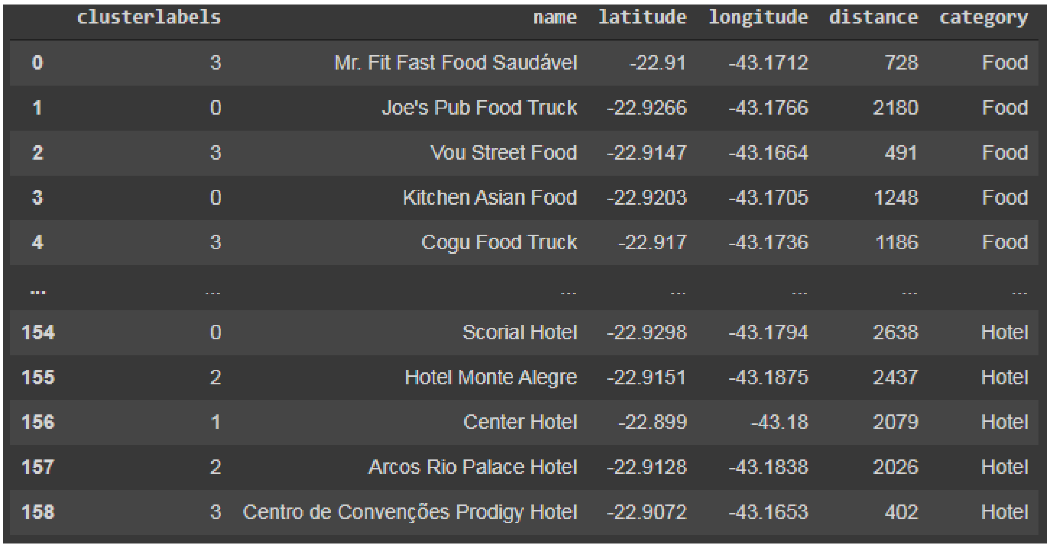
As we can see, the Silhouette score indicates 4 clusters as the best option for our problem. The score of 0.47 is a good indicative and it's relevant considering our data. An argument could be made for a k = 8, so I decided to plot the elbow method, which calculates the sum of squared errors for each value of k, to see if it would be able to provide more information to our model and validate our data:



Since our clusters are somewhat well defined, the elbow method is pretty clear and definitely helps me justifying my k as 4.

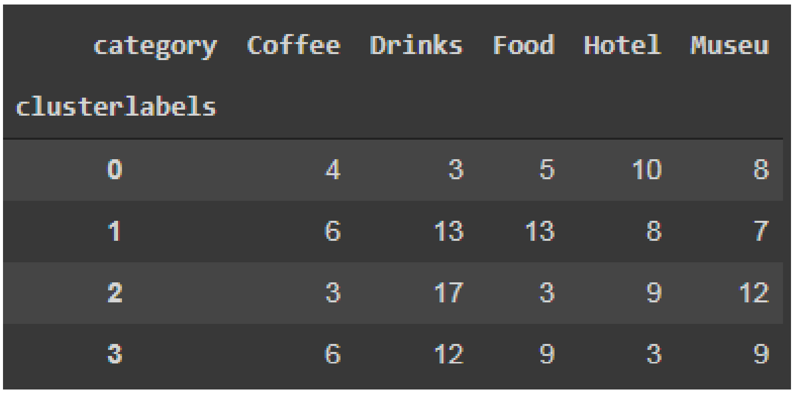
4.2 – Final data set adjustments

After fitting our model, I've added the clusters column to our dataset, to make it easier to plot and further analyze the data.



	clusterlabels	name	latitude	longitude	distance	category
0	3	Mr. Fit Fast Food Saudável	-22.91	-43.1712	728	Food
1	0	Joe's Pub Food Truck	-22.9266	-43.1766	2180	Food
2	3	Vou Street Food	-22.9147	-43.1664	491	Food
3	0	Kitchen Asian Food	-22.9203	-43.1705	1248	Food
4	3	Cogu Food Truck	-22.917	-43.1736	1186	Food
...	...	...	...	...	...	...
154	0	Scorial Hotel	-22.9298	-43.1794	2638	Hotel
155	2	Hotel Monte Alegre	-22.9151	-43.1875	2437	Hotel
156	1	Center Hotel	-22.899	-43.18	2079	Hotel
157	2	Arcos Rio Palace Hotel	-22.9128	-43.1838	2026	Hotel
158	3	Centro de Convenções Prodigy Hotel	-22.9072	-43.1653	402	Hotel

I’ve also decided to add a category count to our table, which would later prove useful when plotting and taking a deeper look on the clusters.

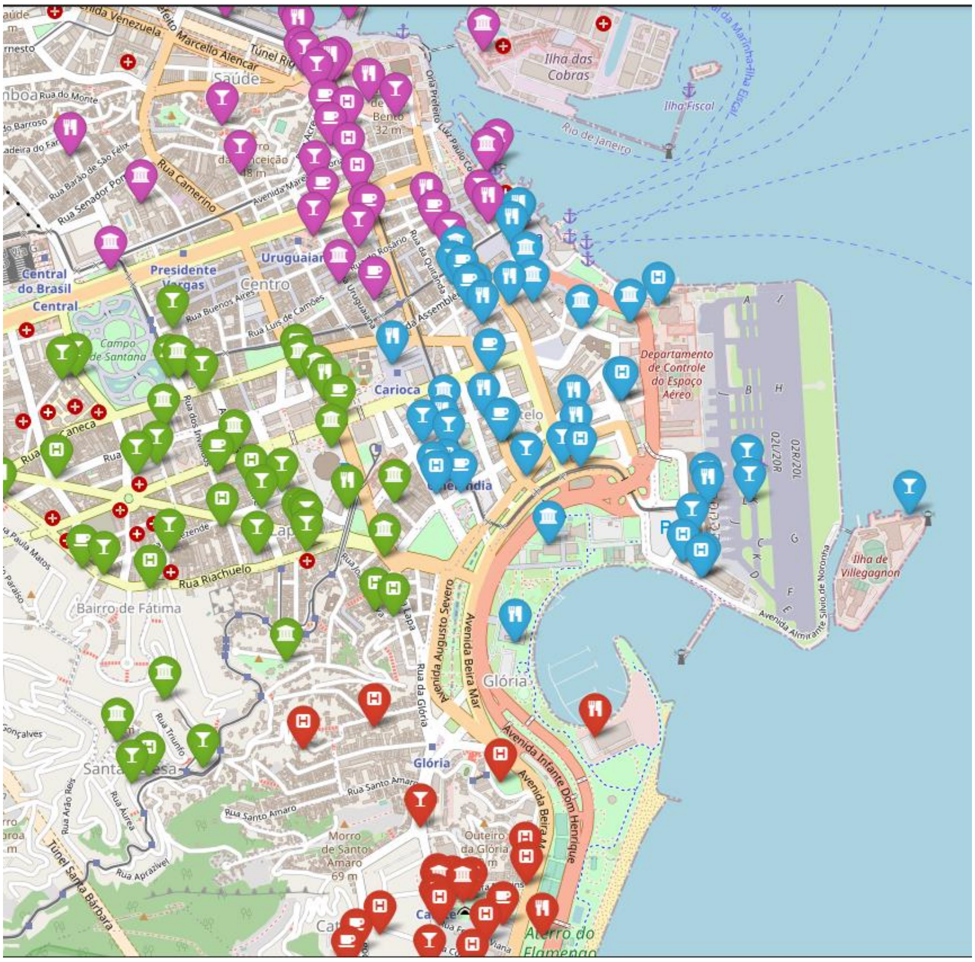


category	Coffee	Drinks	Food	Hotel	Museu
clusterlabels					
0	4	3	5	10	8
1	6	13	13	8	7
2	3	17	3	9	12
3	6	12	9	3	9

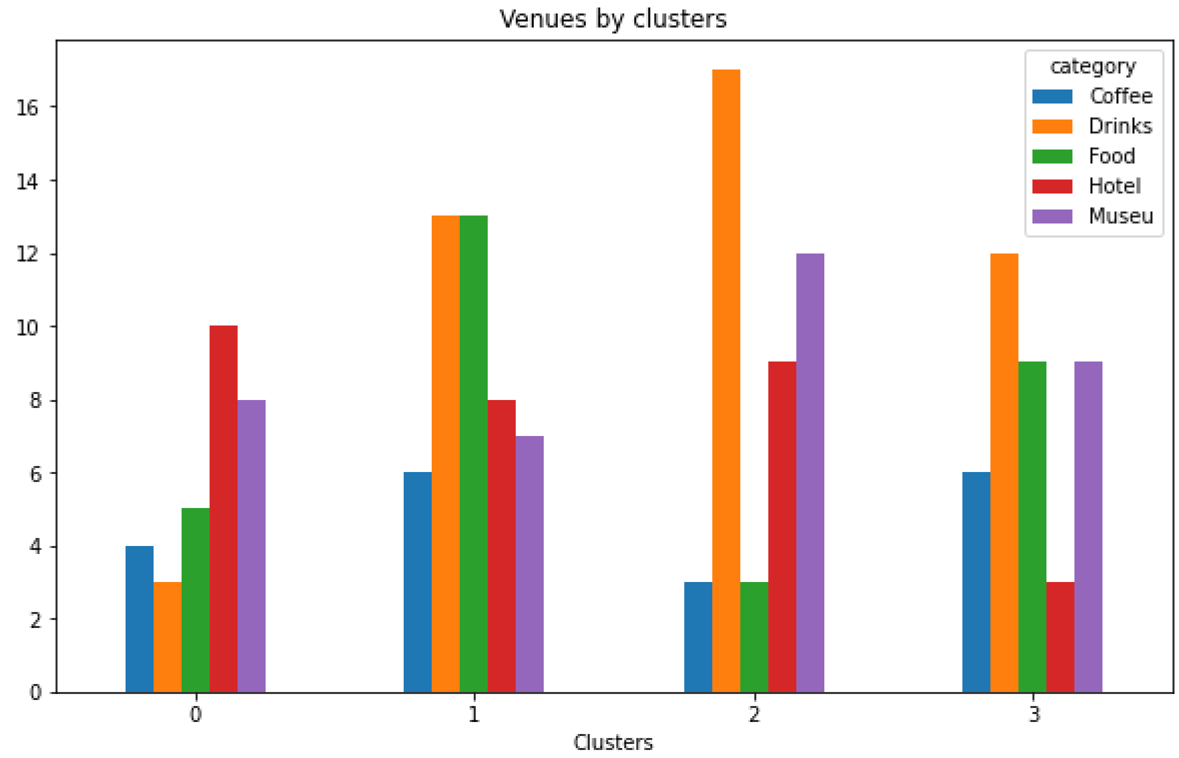
5- Visualization and final analysis

5.1 – Clean map plot and cluster study

Now that all the preparation work has been finalized, plotting was straightforward. To improve visuals and make our clusters easier to identify, I’ve used the Folium Marker map plot, and used icons from Font Awesome built-in Folium integration.



Each venue has its own icon to improve user readability. Using that information and the previous categorization of data, allow us to plot another graphic detailing each clusters content:



Our analysis shows that, while we have some options all around the neighborhood, some of the venues really concentrate in a few areas, while other presents good opportunities for new businesses.

- Zone 0 (red markers on the map) is the less crowded overall, but maybe that's because its located on the edge of Rio's Centro, and closer to the Zona Sul (South Zone), where some of the most famous beaches and neighborhoods like Copacabana are located. It could be a great option for those who are willing to travel a bit far and enjoy more of the City.

- Zone 1 (blue markers on the map), the nearest to the airport, is a well-balanced region, with good amount of different venues and activities, which could be an interesting option for a traveler who doesn't have a lot of time and wishes to visit a couple of interesting places on foot.

- Zone 2 (green markers on the map) has the most options of bars (17) and museums (12), while severely lacking on food (3) and coffee shop (3) options. This could mean a hotspot for tourists who would prefer a busy evening and lots of activities in the nigh life.

- Zone 3 (purple markers on the map) is a well-balanced area for activities, with a bit of everything, but one may find the hotel choices in the area a bit restrictive, as there are only 3 options to choose from.

Our clusters could also help potential investors to find good spots for opening new businesses. As an example, Zone 2 (green markers on the map), the one that has the most variety of museums (12) and night life (17), could really use more options of restaurants (3) or food trucks (3).

6- Conclusion & future directions

The purpose of this project was to map venues around Rio de Janeiro's Centro and provide a general overview to tourists coming to the city, so they would be able to make well-informed choices regarding their stay and based on their personal interests. The data gathered could also be used to indicate hotspots with bigger demand for potential new businesses around the neighborhood.

For further projects, we could implement a recommender system that would allow the user to choose, via a web interface their preferred spots and receive custom suggestions for hotels and interesting venues. Interactive filters based on current geolocation distance, rating filters and expansive information about chosen venues could improve user experience and usability.