Selecting the best regions for a Pizza Place operation in the city of São Paulo

Guilherme Macahyba

March 10, 2020

1. Introduction

Starting a new venture in the food business is complicated, from selecting the type of food you want to cook to the location of the venue, the enormity of the task can overwhelm the stakeholders responsible for making the critical decisions needed. This project aims to tackle a potential solution to the problem concerning the location of the venue. The business case presented below is fictitious, but it is within the realm of possible scenarios a business owner may face.

A small-sized company wants to enter the restaurant business, specifically they want to start a Pizza Place. They plan to buy and remodel available a real estate property in the center of the city of São Paulo, the financial capital of Brazil and the country's largest consumer of pizza, but do not know which location would best leverage their efforts to stand out in the food industry. So, in order to achieve their goal, they want to use the power of machine learning and data science to find their answer and thrive in their newfound business.

2. Data acquisition and cleaning

2.1 Data sources

The data used in this project was mainly geographical and its source is a .csv file found on this delivery website: hastedesign.com.br/lab/planilha-areas-de-entrega-por-ceps-de-sao-paulo-woocommerce/. The file contained the CEP (the Brazilian equivalent of the USA's Zip Code), the boroughs, the neighborhoods nested in each borough and, finally, the delivery fee for each. The file can be found on the same repository as this report.

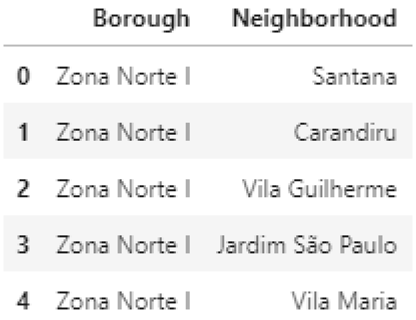
2.2 Data cleaning

Firstly, the delivery fee rate was dropped as it was not relevant to the scope of the project. Next, the CEP's column was dropped as well for it did not contain an accurate representation of the Zip Codes per neighborhood. NaN values were also dropped from the table.

All that was left then is the boroughs and its respective neighborhoods. Since the neighborhoods were grouped in a single cell the procedure performed next involved splitting them and expanding the information contained in each cell. After this, the DataFrame was ready to be used.

2.3 Feature selection

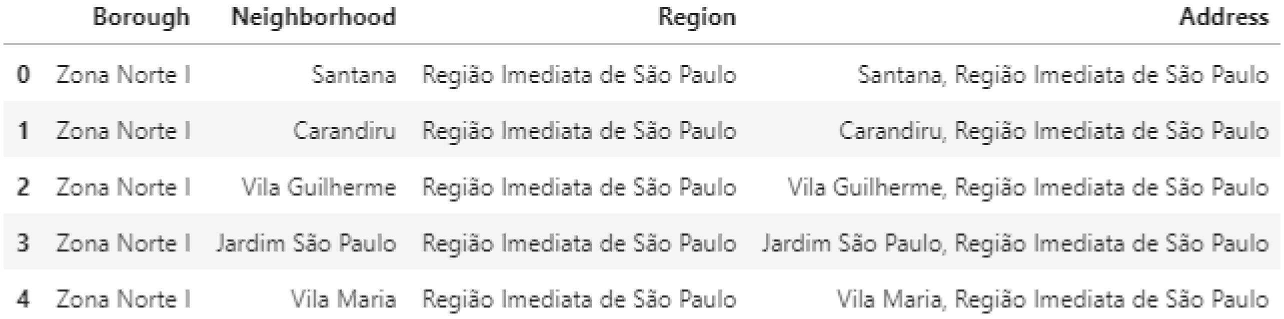
The features selected from the previous step were the Boroughs and Neighborhoods. They were chosen as they, especially the Neighborhoods, housed crucial information which would be fed to the Geocode library, with the purpose of extracting the geographical information of each location. The result of the cleaning and manipulation of the data can be seen below.



	Borough	Neighborhood
0	Zona Norte I	Santana
1	Zona Norte I	Carandiru
2	Zona Norte I	Vila Guilherme
3	Zona Norte I	Jardim São Paulo
4	Zona Norte I	Vila Maria

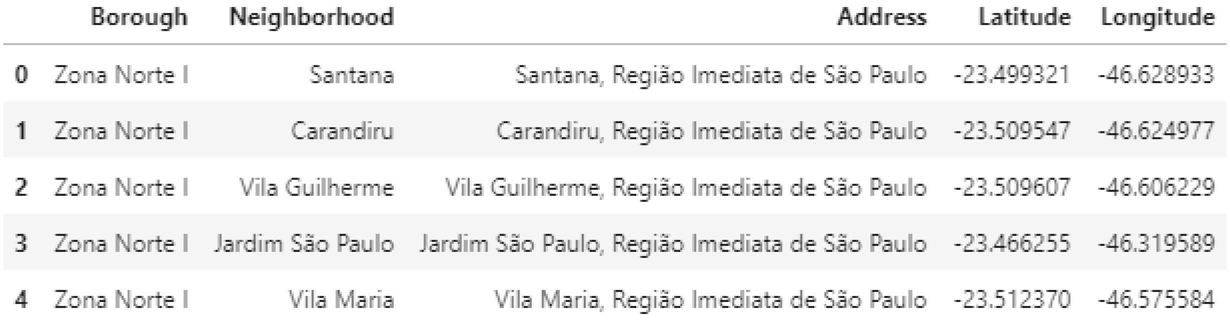
2.4 Obtaining the Geographical Data

Next, to get the geographical coordinates of the cities a new structure was attached to the Dataframe and a new column was added, one which will be fed to Geocode. The resulting Dataframe lies below:



	Borough	Neighborhood	Region	Address
0	Zona Norte I	Santana	Região Imediata de São Paulo	Santana, Região Imediata de São Paulo
1	Zona Norte I	Carandiru	Região Imediata de São Paulo	Carandiru, Região Imediata de São Paulo
2	Zona Norte I	Vila Guilherme	Região Imediata de São Paulo	Vila Guilherme, Região Imediata de São Paulo
3	Zona Norte I	Jardim São Paulo	Região Imediata de São Paulo	Jardim São Paulo, Região Imediata de São Paulo
4	Zona Norte I	Vila Maria	Região Imediata de São Paulo	Vila Maria, Região Imediata de São Paulo

Then the new column ['Address'] was, as written in the last paragraph, fed into Geocode, which after a few minutes yielded the resulting Dataframe:

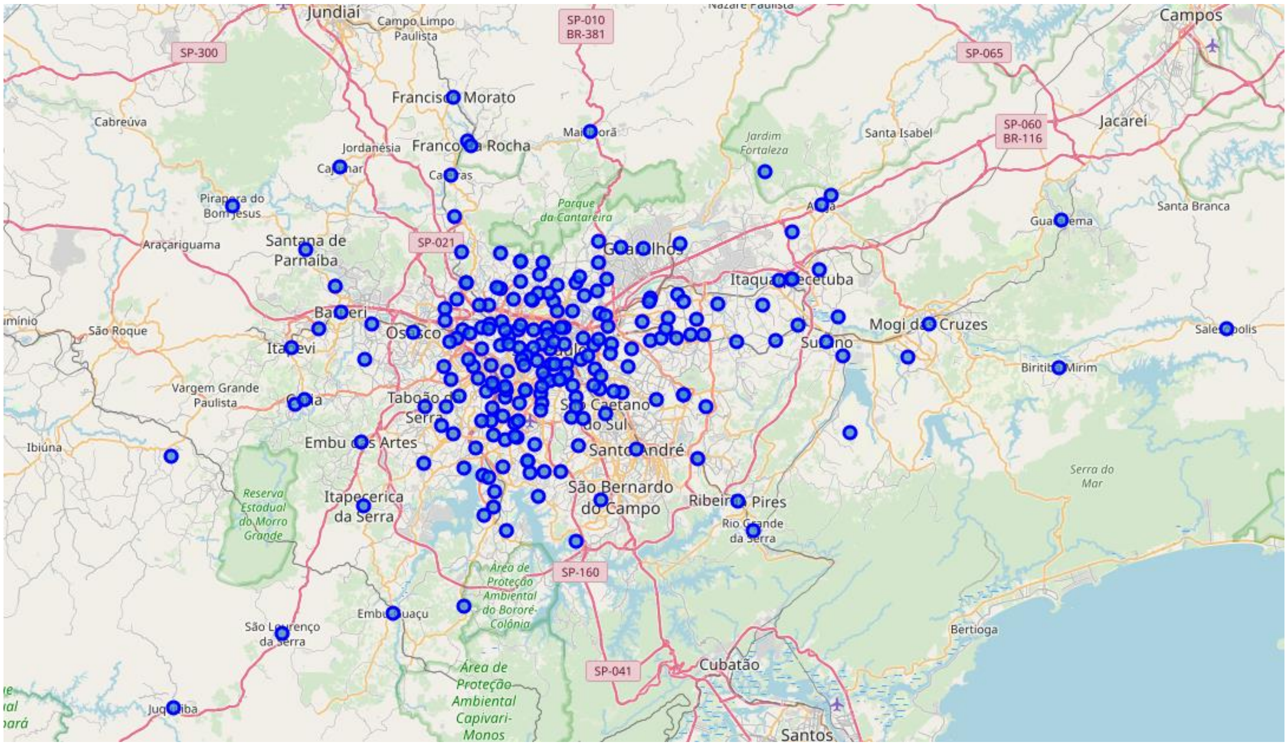


	Borough	Neighborhood	Address	Latitude	Longitude
0	Zona Norte I	Santana	Santana, Região Imediata de São Paulo	-23.499321	-46.628933
1	Zona Norte I	Carandiru	Carandiru, Região Imediata de São Paulo	-23.509547	-46.624977
2	Zona Norte I	Vila Guilherme	Vila Guilherme, Região Imediata de São Paulo	-23.509607	-46.606229
3	Zona Norte I	Jardim São Paulo	Jardim São Paulo, Região Imediata de São Paulo	-23.466255	-46.319589
4	Zona Norte I	Vila Maria	Vila Maria, Região Imediata de São Paulo	-23.512370	-46.575584

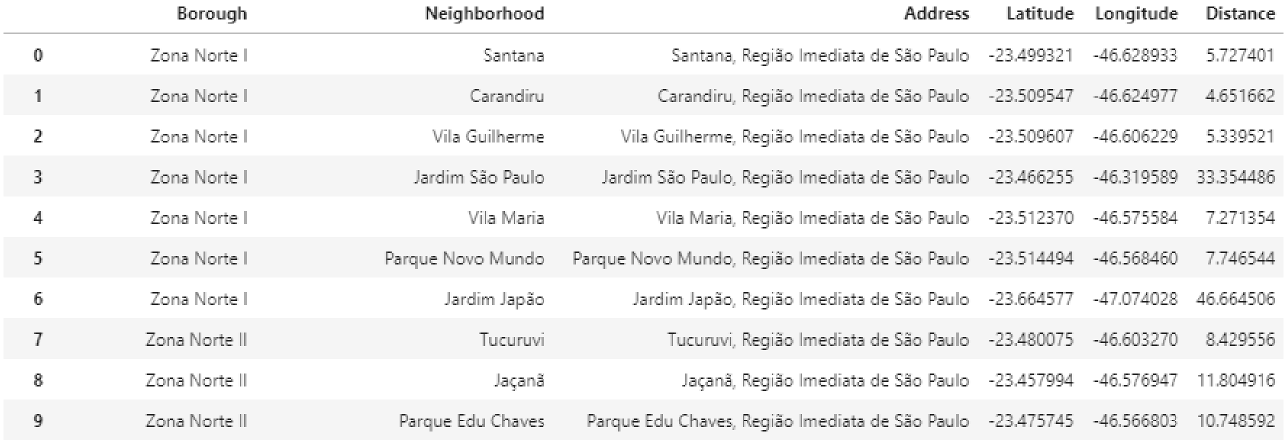
Next, the results obtained will be used to generate maps to further the understanding of the geographical layout.

3. Visualizations and Distances

After the geographical location of each neighborhood was retrieved, a programming library named **folium** was used to generate a map view of these points.

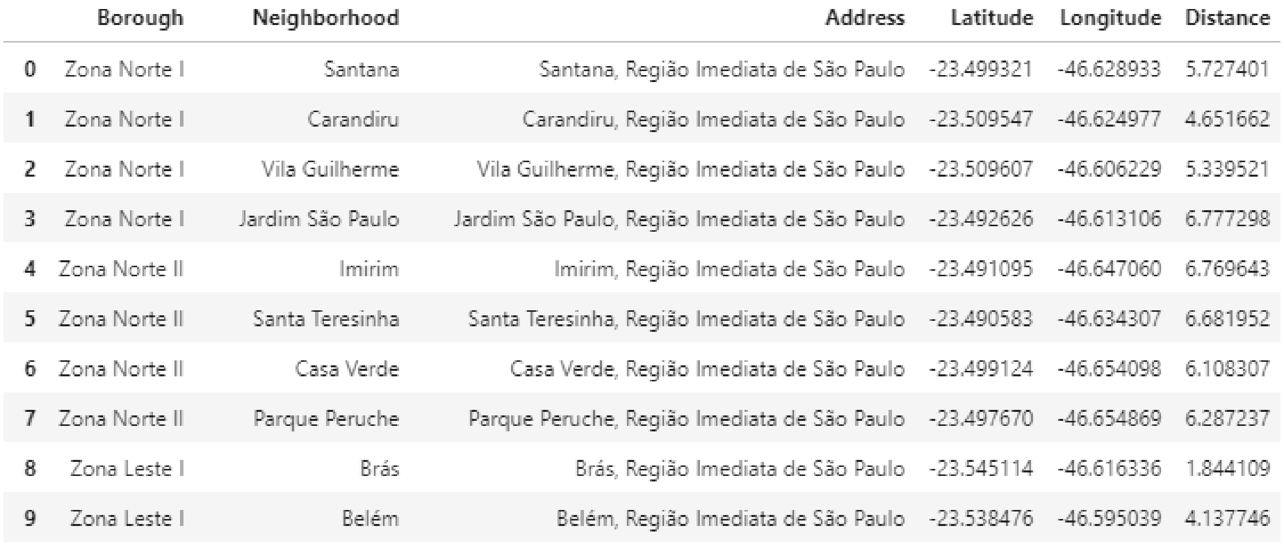


Since we only want the centermost regions of São Paulo to be considered for further analysis, the Haversine equation was plugged in a function to gauge the distance between the center of the city with each neighborhood. The latitude and longitude of the center are: **(-23.5506507, -46.6333824)**. The result can be seen in the new column ['Distance'].



	Borough	Neighborhood	Address	Latitude	Longitude	Distance
0	Zona Norte I	Santana	Santana, Região Imediata de São Paulo	-23.499321	-46.628933	5.727401
1	Zona Norte I	Carandiru	Carandiru, Região Imediata de São Paulo	-23.509547	-46.624977	4.651662
2	Zona Norte I	Vila Guilherme	Vila Guilherme, Região Imediata de São Paulo	-23.509607	-46.606229	5.339521
3	Zona Norte I	Jardim São Paulo	Jardim São Paulo, Região Imediata de São Paulo	-23.466255	-46.319589	33.354486
4	Zona Norte I	Vila Maria	Vila Maria, Região Imediata de São Paulo	-23.512370	-46.575584	7.271354
5	Zona Norte I	Parque Novo Mundo	Parque Novo Mundo, Região Imediata de São Paulo	-23.514494	-46.568460	7.746544
6	Zona Norte I	Jardim Japão	Jardim Japão, Região Imediata de São Paulo	-23.664577	-47.074028	46.664506
7	Zona Norte II	Tucuruvi	Tucuruvi, Região Imediata de São Paulo	-23.480075	-46.603270	8.429556
8	Zona Norte II	Jaçanã	Jaçanã, Região Imediata de São Paulo	-23.457994	-46.576947	11.804916
9	Zona Norte II	Parque Edu Chaves	Parque Edu Chaves, Região Imediata de São Paulo	-23.475745	-46.566803	10.748592

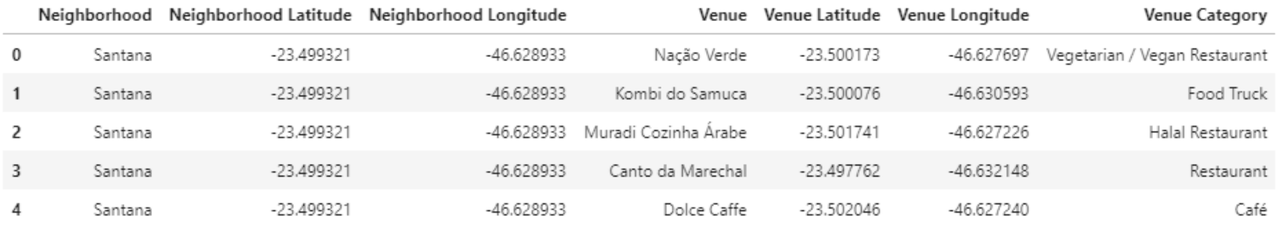
A maximum distance of 7km was considered as a suited distance in which boundaries to work, so only neighborhoods within that distance will be analyzed. The result of the selection is the following after the criteria is applied:



	Borough	Neighborhood	Address	Latitude	Longitude	Distance
0	Zona Norte I	Santana	Santana, Região Imediata de São Paulo	-23.499321	-46.628933	5.727401
1	Zona Norte I	Carandiru	Carandiru, Região Imediata de São Paulo	-23.509547	-46.624977	4.651662
2	Zona Norte I	Vila Guilherme	Vila Guilherme, Região Imediata de São Paulo	-23.509607	-46.606229	5.339521
3	Zona Norte I	Jardim São Paulo	Jardim São Paulo, Região Imediata de São Paulo	-23.492626	-46.613106	6.777298
4	Zona Norte II	Imirim	Imirim, Região Imediata de São Paulo	-23.491095	-46.647060	6.769643
5	Zona Norte II	Santa Teresinha	Santa Teresinha, Região Imediata de São Paulo	-23.490583	-46.634307	6.681952
6	Zona Norte II	Casa Verde	Casa Verde, Região Imediata de São Paulo	-23.499124	-46.654098	6.108307
7	Zona Norte II	Parque Peruche	Parque Peruche, Região Imediata de São Paulo	-23.497670	-46.654869	6.287237
8	Zona Leste I	Brás	Brás, Região Imediata de São Paulo	-23.545114	-46.616336	1.844109
9	Zona Leste I	Belém	Belém, Região Imediata de São Paulo	-23.538476	-46.595039	4.137746

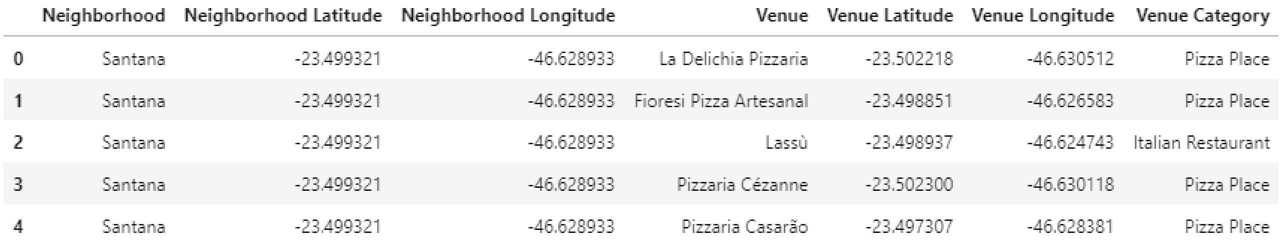
4. Cluster analysis

In this section the **FourSquare API** was used to retrieve the kind of venues we want to analyze. To achieve this result, a user-defined function was used which, when we plug in the values of Latitude and Longitude of each neighborhood of the previous table produced the following Dataframe:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Santana	-23.499321	-46.628933	Nação Verde	-23.500173	-46.627697	Vegetarian / Vegan Restaurant
1	Santana	-23.499321	-46.628933	Kombi do Samuca	-23.500076	-46.630593	Food Truck
2	Santana	-23.499321	-46.628933	Muradi Cozinha Árabe	-23.501741	-46.627226	Halal Restaurant
3	Santana	-23.499321	-46.628933	Canto da Marechal	-23.497762	-46.632148	Restaurant
4	Santana	-23.499321	-46.628933	Dolce Caffe	-23.502046	-46.627240	Café

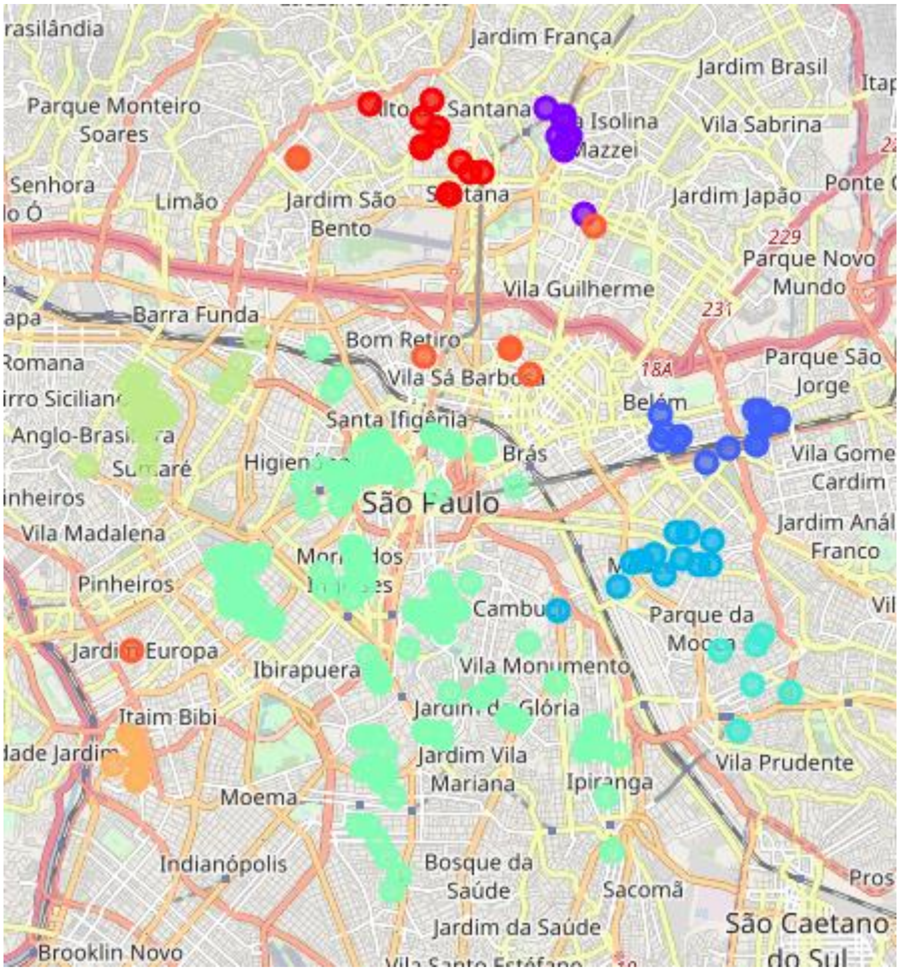
But since we only want Pizza Places and Italian Restaurants (which are not excluded from serving pizza), a new criterion was applied, and it resulted in 247 venues in the 7km region previously defined.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Santana	-23.499321	-46.628933	La Delichia Pizzaria	-23.502218	-46.630512	Pizza Place
1	Santana	-23.499321	-46.628933	Fioresi Pizza Artesanal	-23.498851	-46.626583	Pizza Place
2	Santana	-23.499321	-46.628933	Lassù	-23.498937	-46.624743	Italian Restaurant
3	Santana	-23.499321	-46.628933	Pizzaria Cézanne	-23.502300	-46.630118	Pizza Place
4	Santana	-23.499321	-46.628933	Pizzaria Casarão	-23.497307	-46.628381	Pizza Place

Now we will use a Machine Learning clustering algorithm. Its name is DBSCAN, which stands for Density-

Based Spatial Clustering Applications with Noise. It works by finding the core of regions with high density of user-defined components and from these cores it expands the clusters. It will help us better see the patterns in the location of the venues. The result of the application of such tool yielded the following map:



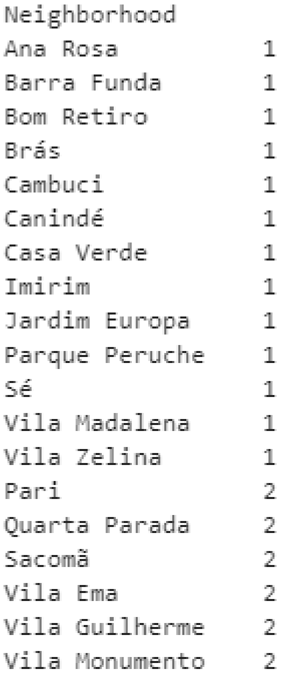
This result was achieved with the following function parameters:

- Epsilon = 1.25/km_per_radians, where km_per_radian = 6371.0088
- Min_samples = 5

Numerous combinations were tested and that one produced the best results, with the least noise and greater visual representation.

As can be seen the algorithm divided all those 247 venues in approximately 10 clusters. From the observation of the map a trend can be ascertained: there are high-density regions in the South and West directions, leaving a measurable chunk of the eastern portion of the map unattended.

Since the visualization alone may be insufficient to determine the most suited locations, a pandas operation will be performed, one that will assess the neighborhoods that do not have many restaurants/pizza places. The result of this lies in the image below:



```
Neighborhood
Ana Rosa          1
Barra Funda       1
Bom Retiro        1
Brás              1
Cambuci           1
Canindé           1
Casa Verde        1
Imirim            1
Jardim Europa     1
Parque Peruche    1
Sé                1
Vila Madalena     1
Vila Zelina       1
Pari              2
Quarta Parada     2
Sacomã            2
Vila Ema          2
Vila Guilherme    2
Vila Monumento    2
```

5. Results and Discussion

As per the analysis conducted it could seen that there are 247 Pizza Places/Italian Restaurants in the relatively small area that studied. But as can be seen in the map above those 247 venues are not evenly spaced. There are empty pockets in the area which do not cointain a single venue, and on the other hand, neighborhoods which house many venues closely packed together.

Combining the visualization with the data analysis it can be said that the best locations, regarding relative distance to competitors and distance to the center of the city, lie in the following pockets:

- The northeastern region between **Vila Guilherme** and **Pari**;
- The eastern region between **Brás** and **Belém**; and
- The southeastern region between **Moóca** and **Aclimação**.

That, naturally, does not imply that these are actually the most suited locations to house a new Pizza Place. These areas were chosen due to the fact that they are not crowded with the same type of restaurant and their distance to the center of the city is optimal if we think about accessibility to all the other regions in the city (in the event the venue also doing deliveries). But, we cannot discard this places as potential locations for the new business and they will serve as a starting point for further analysis as the business project unfolds.

6. **Conclusion**

The objective of this project was to determine the suitability of neighborhoods in São Paulo to house a new Pizza Place. Far from being a definitive answer, the results yielded by the analysis are merely a starting point. Many other factors should be considered, besides the relative distance to competitors, when opening a venue: socioeconomic profile of the population you want to service, the region's foot traffic, accessibility to suppliers and costumers, for example, are only a few of these factors.

The decision for the most suited location for a new business is one taken by the stakeholders of the project. They will ponder the factors mentioned above, and many others, to reach a final verdict, but they do not have to do so without the help of data science and machine learning, as was done in the past. This project aimed to, and hopefully succeeded, in giving a peek of the potential of these fields relating to any type of business.

7. Future directions

As discussed in the previous section, the analysis that can be done to select the best location for the new business venture can be greatly expanded. Socioeconomic demography, foot traffic, price of square meter for commercial properties and accessibility to customers and suppliers can also be used to narrow the decision-making process even further. Depending of the needs of the stakeholders the outcome may greatly change from its expected path. One might find out that, for instance, the pizza place business is a saturated market and it would be a sound idea to study other types of food businesses. This is one example, of many that could be given. Of course, reliable data sources for those factors must be secured first, be it through webscraping or access to research papers or any kind of document that contains this kind of data.

These are just a few suggestions to enhance and customize the analysis to the specific needs of the stakeholders. The goal is of course, to supply them with the best analysis possible, as to ensure the success of the business' operation