IBM Data Science Professional Certificate capstone project

Capstone Project - The Battle of Neighborhoods

Analysis of the communes / neighborhoods of the city of Buenos Aires

Julián Fernández

June 27, 2021

1. Introduction

1.1. Background

The city of Buenos Aires, Federal District and capital of Argentina, is the most populated city in the Argentine Republic with a population of approximately 3,000,000 (2010) and the Metropolitan Area of Buenos Aires (AMBA), made up of the city of Buenos Aires and 40 surrounding *municipios*, 15,000,000 (2010), according to the National Institute of Statistics and Censuses of the Argentine Republic.

The city of Buenos Aires is among the cities with the highest quality of life in Latin America, and its per capita income is among the three highest in the region. It is the most visited city in South America due to its cultural diversity, commerce, industry, politics, culture, and technology.

The 2010 national census estimated that approximately 2,000,000 immigrants born in another country resided in Argentina, equivalent to 4.5% of the population. At a global level, Argentina is ranked 29th for the number of immigrants in its territory, being the largest recipient of immigrants from Latin America and Buenos Aires being one of the main cities that houses this foreign population.

1.2. Business Problem

Given the large amount of immigrant population that the city of Buenos Aires has, it is presented as a great opportunity, to open a restaurant that satisfies a gastronomic need for typical food of some foreign country or countries, analyzing the amount of immigrant population by areas, as well the existing competition in them, in order to obtain as a result the best area to open the restaurant and the best type of food to sell, that is, if it will be for example, a Peruvian, Spanish or Indian restaurant.

2. Data

2.1 Data sources

The data used for this project are presented below:

1 - Foursquare Location Data: Venues information of each commune/neighborhood. We are only going to use the restaurants and their coordinates.

2 - Buenos Aires communes and neighbourhoods dataset: All the names of Buenos Aires communes and its neighborhoods.

https://cdn.buenosaires.gob.ar/datosabiertos/datasets/comunas/comunas.csv

3 - GeoJson dataset: Communes boundaries that will help to visualize communes on the map.

https://cdn.buenosaires.gob.ar/datosabiertos/datasets/comunas/CABA_comunas.geojson

4 - Coordinates of Buenos Aires communes: Latitude and longitude of each Buenos Aires commune.

https://www.municipalidad-argentina.com.ar/municipalidad-buenos-aires-ar.html

5 - Foreign population of communes dataset: Number of foreign population of each commune.

https://www.indec.gob.ar/indec/web/Nivel4-Tema-2-41-135

2.2 Data cleaning

We will start by downloading all the data we need and transform it into a Pandas dataframe and separating each neighborhood by comma. Unfortunately we do not have the coordinates of each commune in the dataset where the data was taken, therefore, we proceed to load it manually. The coordinates were taken from https://www.municipalidad-argentina.com.ar/municipalidad-buenos-aires-ar.html

Then we proceed to download the foreign population of each commune, getting the birthplace and the total foreign population by country.

Finally, wee look for the coordinates of the city of Buenos Aires, to later visualize it on a map and we use FourSquare to explore the area around the communes within a radius of 1 Km, to later filter this data where the venue category is restaurant or food.

3. Methodology

From now on our effort will focus on finding an area of the city of Buenos Aires where the density of restaurants (the competition) is low, as well as, finding the predominant foreign population that tells us what type of restaurant we should open, as long as this type of restaurant is not saturated in the area.

First, we will find the density of restaurants by communes.

Then taking into consideration the predominant foreign population, and the existing foreign food restaurants, we will look for the most promising area to open a new restaurant with typical food from the country of the predominant foreign population.

4. Analysis

Since we have the categorical variables such as the restaurants in each neighborhood, we proceed to apply the one-hot encoding technique to analyze each neighborhood. Afterwards we proceed to group rows by neighborhood and by taking the mean of the frequency of occurrence of each category.

Once the frequency of each type of restaurant has been found, we put the top 10 restaurant categories for each neighborhood in a new dataframe.

Now that we have the 10 most common restaurant categories by neighborhood, group them into clusters and observe the characteristics of each one.

To do this we will use the K-means algorithm that is one of the most common cluster method of unsupervised learning.

To select an optimal number of K clusters we will use two method, Elbow method and Silhouette method.

We run the Elbow cluster analysis using Cluster as 1 till 10, and the results are the following:

	Clusters	WCSS
0	1	3.218239
1	2	1.966600
2	3	1.060794
3	4	0.814606
4	5	0.650428
5	6	0.489234
6	7	0.382789
7	8	0.296772
8	9	0.223413

Figure 1: Elbow within Cluster Sum of Squared (WCSS).

Plotting these result, we obtain:

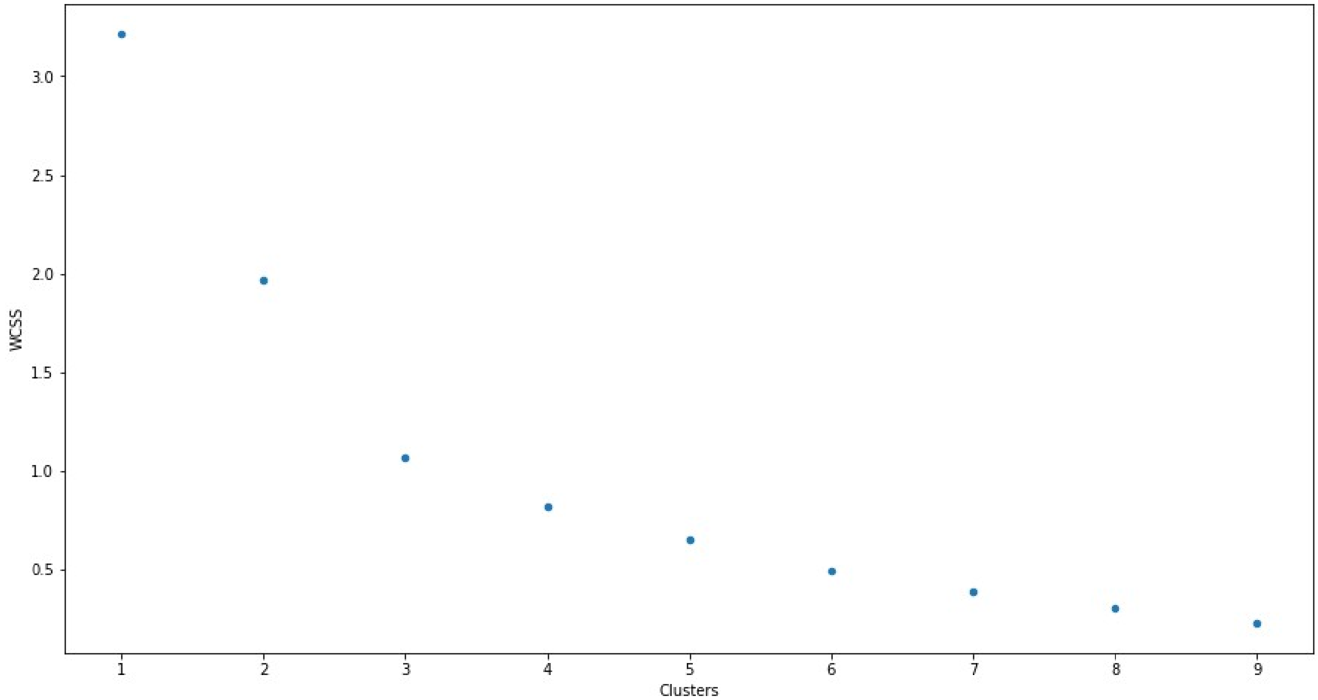

Figure 2: Elbow method graph.

We can see in the above graph the explained variation as a function of the number of clusters. The elbow of the curve in this case is k = 3 as the number of clusters to use.

To ensure that this amount of cluster is appropriate, we proceed to corroborate it with the Silhouette method.

```
Silhouette score for k = 2(clusters) is 0.41841303408853303
Silhouette score for k = 3(clusters) is 0.4459333520668169
Silhouette score for k = 4(clusters) is 0.3422741180343111
Silhouette score for k = 5(clusters) is 0.28756087544124714
Silhouette score for k = 6(clusters) is 0.2702451620281194
Silhouette score for k = 7(clusters) is 0.1881355835457263
Silhouette score for k = 8(clusters) is 0.178412665788756
Silhouette score for k = 9(clusters) is 0.18666878808806045
```

Figure 3: Silhouette method in range 2 till 10.

we can see above silhouette values as a measure of how similar an object is to its own group compared to other groups. The maximum value here is 3 K clusters, the same as the Elbow method throwed, so 3 K clusters is the correct option.

Now we can proceed to execute the K-mean algorithm to obtain the clusters, with a value of K = 3 and then add the cluster labels to each commune.

	Commune	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	
0	1	CONSTITUCION, MONSERRAT, PUERTO MADERO, RETIRO...	-34.6152	-58.3738	0	Argentinian Restaurant	Italian Restaurant	Japanese Restaurant	Fondue Restaurant	Vegetarian / Vegan Restaurant	Swiss Restaurant	Restaurant	M
1	2	RECOLETA	-34.5906	-58.3906	0	Argentinian Restaurant	Italian Restaurant	Restaurant	Empanada Restaurant	French Restaurant	Peruvian Restaurant	Spanish Restaurant	
2	3	BALVANERA, SAN CRISTOBAL	-34.6107	-58.4068	0	Fast Food Restaurant	Peruvian Restaurant	Argentinian Restaurant	American Restaurant	Vegetarian / Vegan Restaurant	Sushi Restaurant	Spanish Restaurant	
3	4	BARRACAS, BOCA, NUEVA POMPEYA, PARQUE PATRICIOS	-34.6464	-58.3843	2	Argentinian Restaurant	Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Paella Restaurant	
4	5	ALMAGRO, BOEDO	-34.6152	-58.4252	0	Argentinian Restaurant	Restaurant	Sushi Restaurant	Spanish Restaurant	Italian Restaurant	Paella Restaurant	Food Service	
5	6	CABALLITO	-34.6183	-58.4367	0	Argentinian Restaurant	Restaurant	Sushi Restaurant	Fast Food Restaurant	Empanada Restaurant	Food Service	Spanish Restaurant	
6	7	FLORES, PARQUE CHACABUCO	-34.6318	-58.4620	0	South American Restaurant	Fast Food Restaurant	Restaurant	Japanese Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	
7	8	VILLA LUGANO, VILLA RIACHUELO, VILLA SOLDATI	-34.6754	-58.4637	1	Fast Food Restaurant	American Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Paella Restaurant	Peruvian Restaurant	
8	9	LINIERS, MATADEROS, PARQUE AVELLANEDA	-34.6449	-58.5112	0	Argentinian Restaurant	Sushi Restaurant	Spanish Restaurant	Peruvian Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant	
9	10	FLORESTA, MONTE CASTRO, VELEZ SARSFIELD, VERSA...	-34.6243	-58.5043	2	Argentinian Restaurant	Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Paella Restaurant	
10	11	VILLA DEL PARQUE, VILLA DEVOTO, VILLA GRAL. MI...	-34.6116	-58.4801	0	Argentinian Restaurant	Restaurant	Spanish Restaurant	Latin American Restaurant	Mediterranean Restaurant	Mexican Restaurant	Middle Eastern Restaurant	
11	12	COGHLAN, SAAVEDRA, VILLA PUEYRREDON, VILLA URQ...	-34.5629	-58.4895	2	Argentinian Restaurant	Food & Drink Shop	American Restaurant	Thai Restaurant	Tapas Restaurant	Swiss Restaurant	Sushi Restaurant	
12	13	BELGRANO, COLEGIALES, NUÑEZ	-34.5524	-58.4572	0	Argentinian Restaurant	Peruvian Restaurant	Sushi Restaurant	Restaurant	Latin American Restaurant	Chinese Restaurant	Italian Restaurant	
13	14	PALERMO	-34.5889	-58.4306	0	Argentinian Restaurant	Italian Restaurant	Middle Eastern Restaurant	Thai Restaurant	Sushi Restaurant	Indian Restaurant	French Restaurant	
14	15	AGRONOMIA, CHACARITA, PARQUE CHAS, PATERNAL, V...	-34.5830	-58.4673	0	Italian Restaurant	Restaurant	Argentinian Restaurant	Vegetarian / Vegan Restaurant	Mexican Restaurant	South American Restaurant	Thai Restaurant	

Figure 4: caba most common type of restaurant per commune.

Extracting the 3 most common restaurants in each cluster in order to understand the trend that each cluster presents, we can say that:

* Cluster 0: Argentinian, Fast food, italian, japanese, middle eastern, peruvian, south american, spanish and sushi restaurant are the most common in this cluster.

* Cluster 1: American, Fast food and mediterranean are the most common in this cluster.

* Cluster 2: American, argentinian, food & drink and latin american are the most common in this cluster.

Now that we know the characteristics of each cluster, let's examine the top 10 foreign population of each cluster.

Cluster 0

0	Perú	46143
1	Bolivia	45350
2	Paraguay	42679
3	Resto de América	26770
4	Uruguay	23824
5	España	20084
6	Italia	15419
7	Resto de Europa	10882
8	Brasil	9282
9	Chile	8103

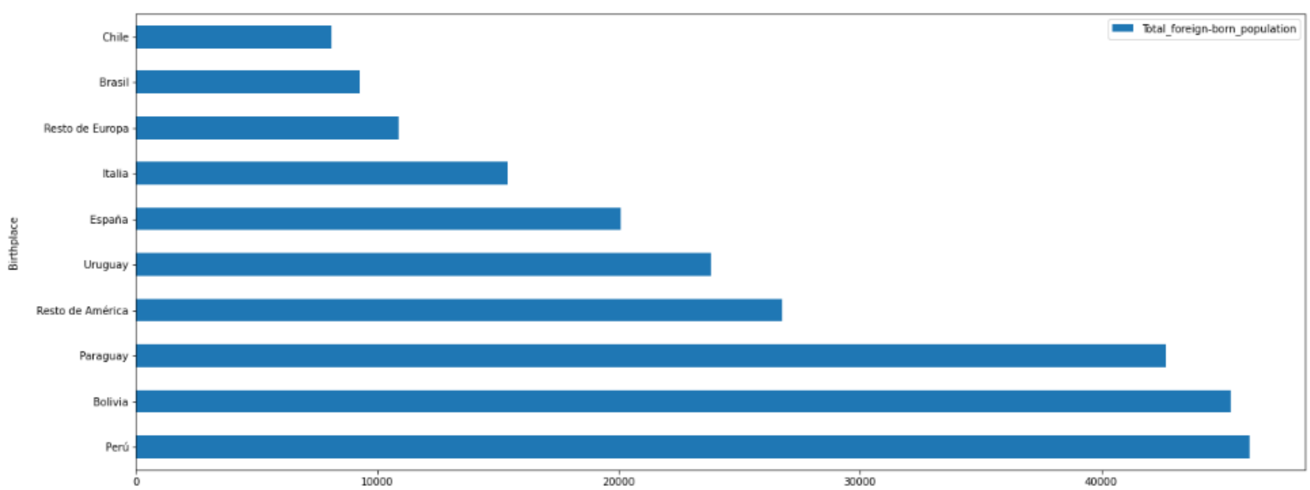

Figure 5: Foreign population in cluster 0.

Cluster 1

	Birthplace	Total_foreign-born_population
0	Bolivia	20365
1	Paraguay	16597
2	Perú	2265
3	Italia	1688
4	Uruguay	941
5	España	841
6	Chile	309
7	Resto de América	237
8	Resto de Europa	164
9	Brasil	149

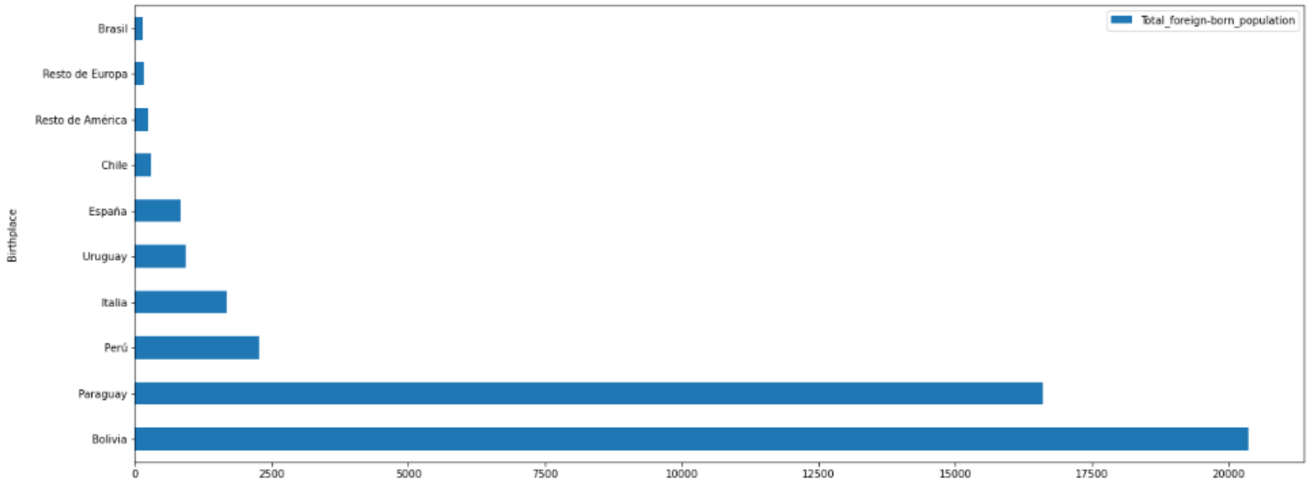

Figure 6: Foreign population in cluster 1.

Cluster 2

	Birthplace	Total_foreign-born_population
0	Paraguay	21874
1	Perú	12392
2	Bolivia	11021
3	Uruguay	6435
4	España	5356
5	Italia	4606
6	Resto de América	3988
7	Resto de Europa	2147
8	Chile	1778
9	Brasil	1423

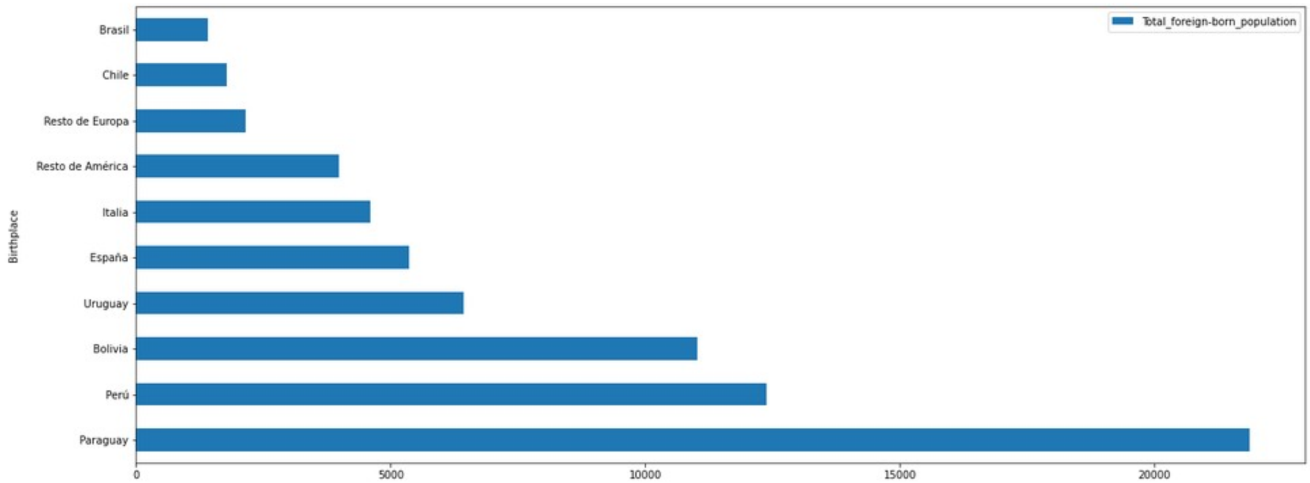

Figure 7: Foreign population in cluster 2.

We already know the predominant foreign population of each cluster.

Now we can proceed to identify the density of restaurants in each commune.

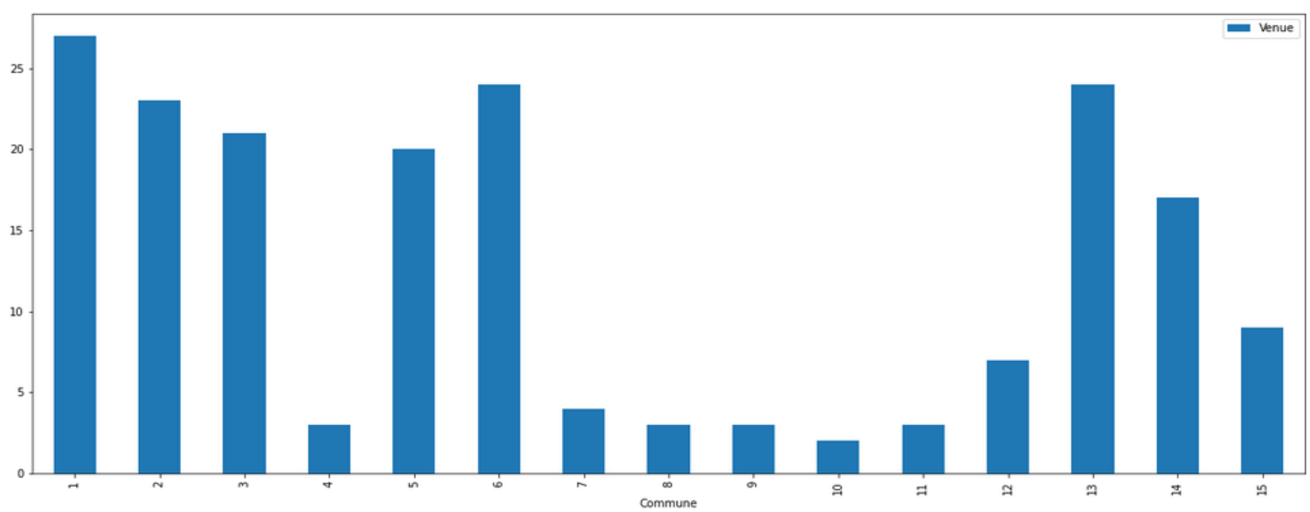

Figure 8: Number of restaurant per commune.

We can also view it in a choropleth map.

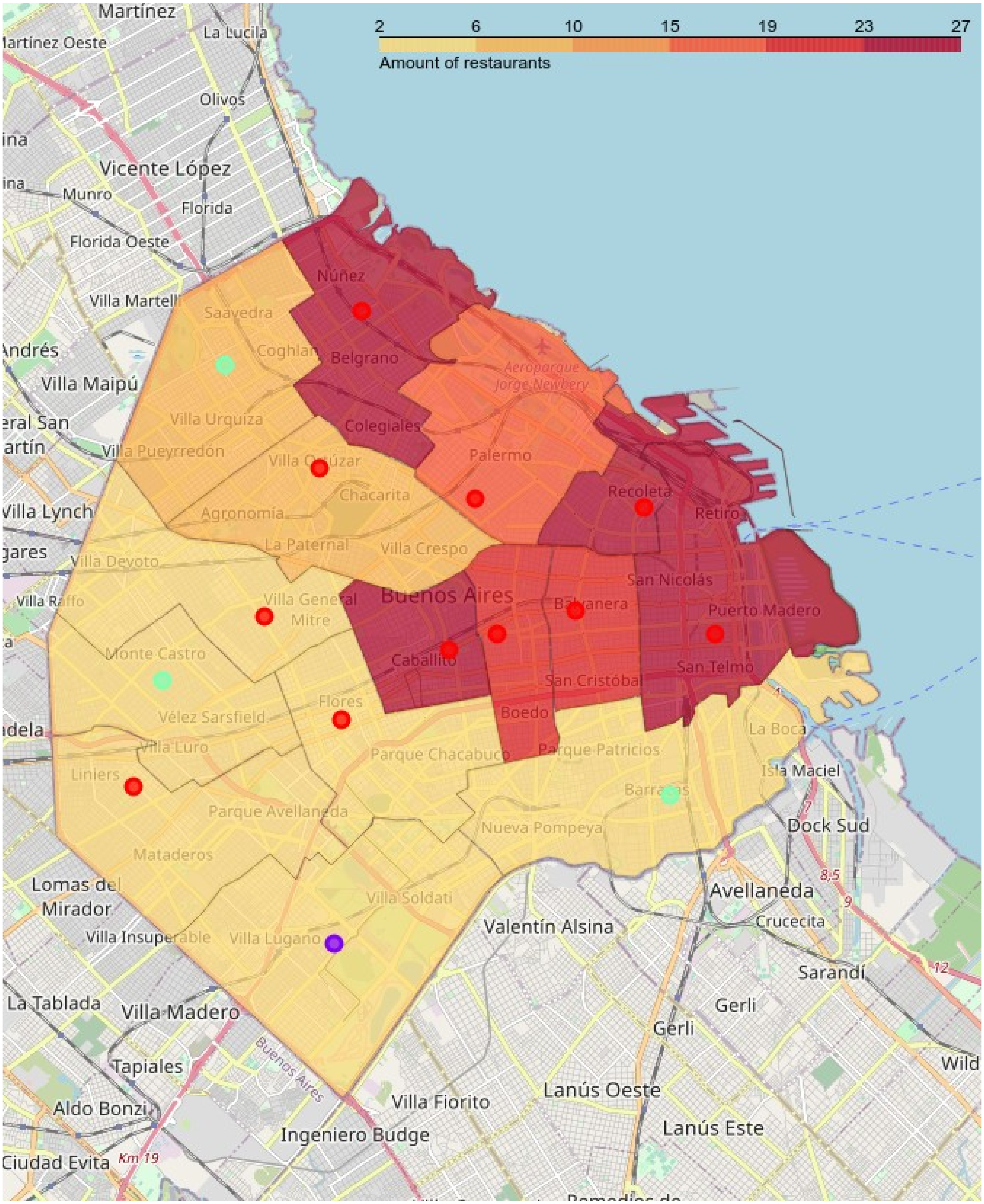

Figure 9: Map of Buenos Aires city identifying communes, clusters and restaurant density.

5. Results and Discussion

Our analysis shows that there is a great number of restaurants in Buenos Aires, but the highest concentration of restaurants was detected in commune 1, 2, 3, 5, 6, 13 and 14, so we focused our attention to all the others communes where the competention is not so high. These communes are 4, 7, 8, 9, 10, 11, 12 and 15.

With these last communes, we find in the cluster 0, commune 7, 9, 11 and 15. The foreign population that predominates in this cluster are from Peru, Bolivia and Paraguay. At the same time, the predominant restaurants in this cluster are peruvian and south american restaurants, among others, so we discard these communes.

In cluster 2, we have commune 4, 10 and 12. The foreign population that predominates in this cluster are from Paraguay, Peru and Bolivia. The predominant restaurants in this cluster are latin american restaurants, among others, so we also discard these communes.

In cluster 1, we have left commune 8. The foreign population that predominates in this cluster are from Bolivia, Paraguay and Peru. The predominant restaurants in this cluster are american, Fast food and mediterranean restaurants. Therefore, we can observe that restaurants of typical food from these countries do not predominate in this cluster. This is why the neighborhoods that make up commune 8 (VILLA LUGANO, VILLA RIACHUELO, VILLA SOLDATI) are presented as a potential area to open a bolivian, paraguayan, peruvian or Latin American restaurant.

This, of course, does not imply that this zone is actually an optimal location for a new restaurant! Purpose of this analysis was to only provide information on areas with poor competence and to look for an area where foreigners do not have restaurants with typical food from their countries in the areas close to their homes. Therefore, the recommended areas should be considered only as a starting point for a more detailed analysis that could eventually result in a location that takes into consideration many more factors than were evaluated in this project.

6. Conclusion

The purpose of this project was to identify Buenos Aires communes with low number of restaurants aiming to find low competition in an area where the predominant foreigners do not have nearby restaurants with typical food from their country. For this, the restaurants of each commune and their foreign population were analyzed. Finally, it was possible to find an optimal area that meets the requirements that were defined and the type of restaurant to open.