Christian Sebastian Russo

Analysis of neighborhoods where it is convenient to build houses or buildings.

BUENOS AIRES, 2020

Introduction

- In Argentina, many construction companies have problems finding the best neighborhood to build houses or buildings.
- This is because the places to build have different prices, but not only that, but it also depends on the commercial movement of the neighborhood, companies, industries, proximity to transport, shops, businesses, among others.
- On the other hand, the construction of a house is not the same as that of a building, the commercial and movement analysis of an area is different.
- It is for this reason that different companies have difficulty and ask themselves: What is the best place to start a new venture?

Description of the problem

- In the province of Buenos Aires there are neighborhoods where a property costs around USD 70,000, in turn near this property we can find a property with similar characteristics at USD 200,000, that is, a price of 285% higher.
- The price of a property is given, among other things, by the proximity to places of interest; for example, a property near a supermarket / shopping that is close to a cemetery will not have the same price.
- Among many of the factors that are taken into account when choosing where the venture is considered most feasible, is the proximity to different points of interest, such as: shops, supermarkets, bars, train stations, bus stations, avenues, etc.

Solution approach

- In the present work, the first thing we will do is obtain the necessary geographic information of the neighborhoods of Buenos Aires (correctly called departments) to be able to work with the Foursquare API to obtain the points of interest of each neighborhood.
- Once this information is obtained, we will calculate, for each neighborhood, what proportion of each category of places of interest counts, finally calculating a total value (the sum of the percentages of each category) and thus be able to measure how feasible it is to build in that neighborhood.
- Then, we will use the Machine Learning algorithm (called k-means) to be able to determine similarities between neighborhoods, and with these similarities find a neighborhood similar to another with the desired characteristics.

Points of interest

- Train Stations
- Shops
- Restaurantes
- Gyms
- Plazas
- Cinemas
- Bar
- Shoppings
- Supermarket
- Etc

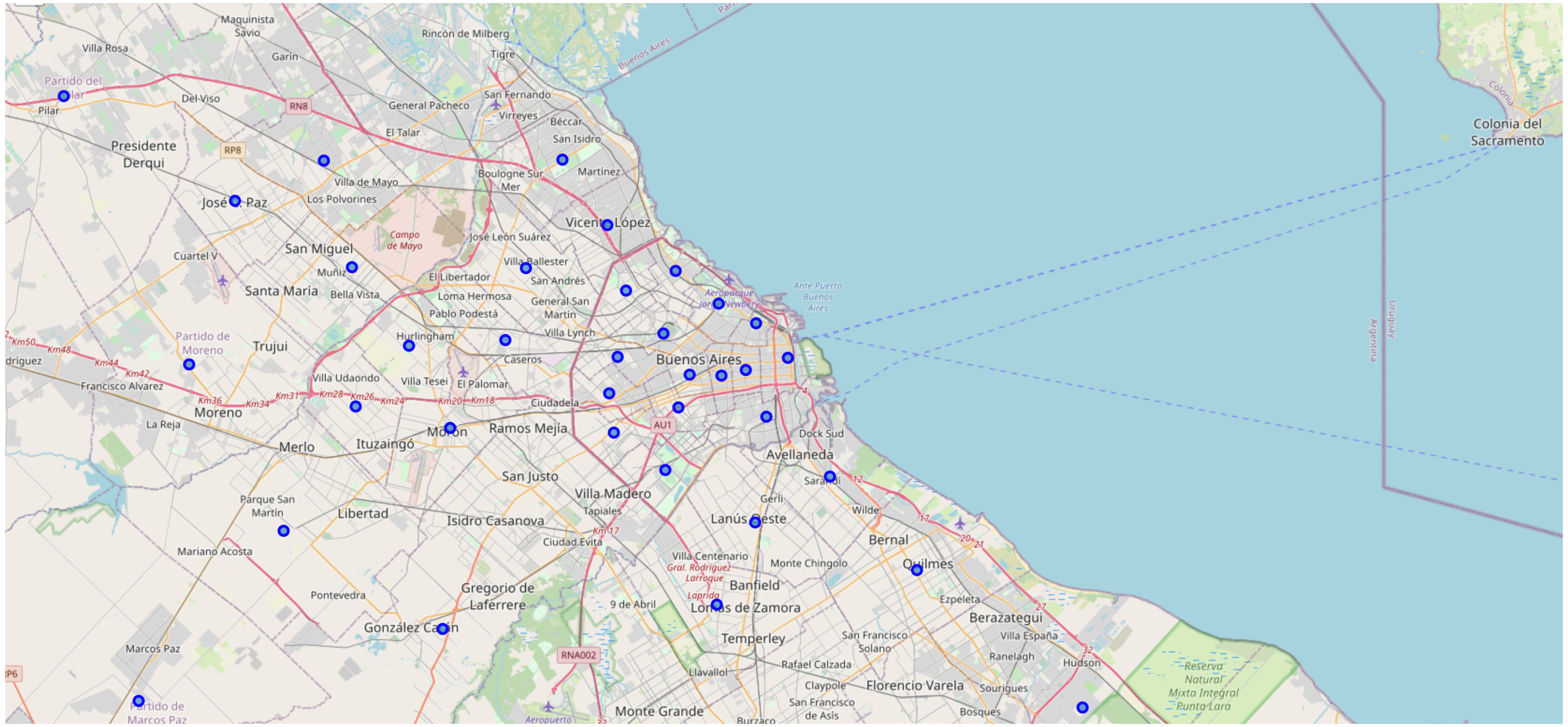
Data analysis

- The first thing we will do is analyze the data obtained from https://www.datos.gob.ar/. Once the data in Buenos Aires and Capital Federal has been filtered, the result of those data from the neighborhoods of Argentina will be shown.

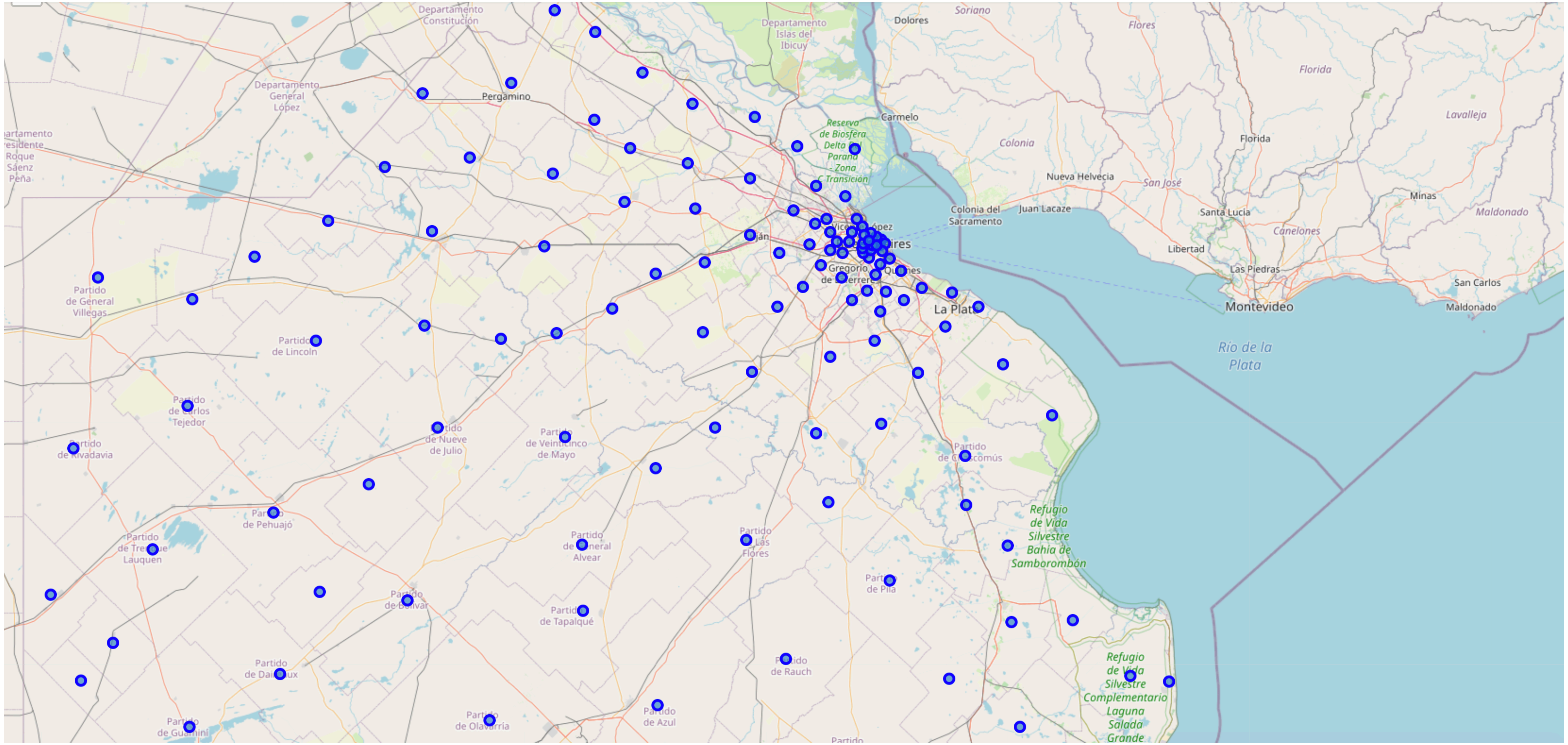
	categoria	centroide_lat	centroide_lon	fuente	id	nombre	nombre_completo
0	Partido	-34.511876	-58.777671	ARBA - Gerencia de Servicios Catastrales	6412	José C. Paz	Partido de José C. Paz
3	Partido	-37.964225	-60.249078	ARBA - Gerencia de Servicios Catastrales	6014	Adolfo Gonzales Chaves	Partido de Adolfo Gonzales Chaves
4	Partido	-37.153214	-57.230787	ARBA - Gerencia de Servicios Catastrales	6315	General Juan Madariaga	Partido de General Juan Madariaga
5	Partido	-37.335655	-59.181806	ARBA - Gerencia de Servicios Catastrales	6791	Tandil	Partido de Tandil
13	Partido	-38.147744	-61.264659	ARBA - Gerencia de Servicios Catastrales	6196	Coronel Pringles	Partido de Coronel Pringles

Data analysis

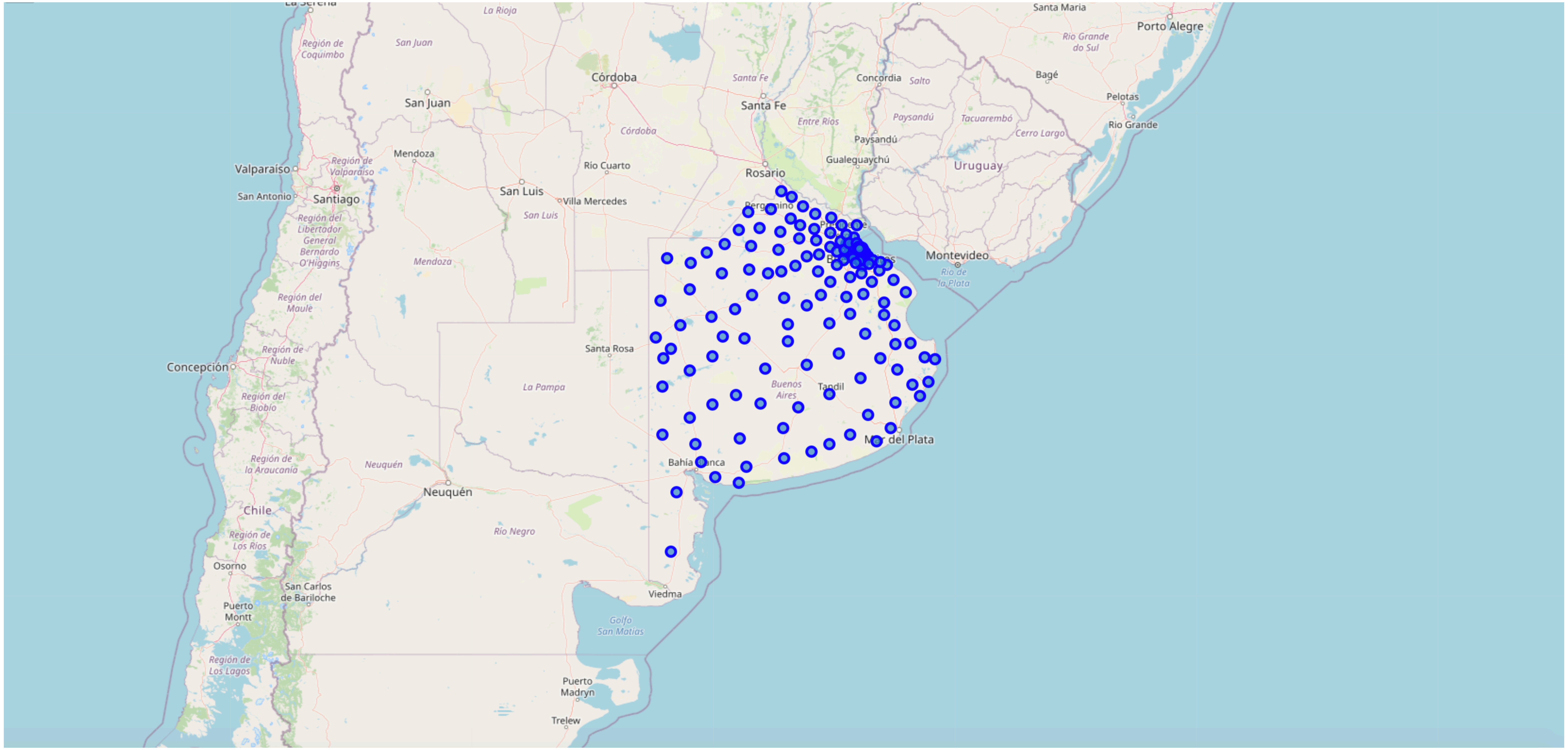
- Using the geographic coordinates of each neighborhood (latitude and longitude) we will draw a map to verify the information obtained:



Data analysis



Data analysis



Data analysis

- Then, using the Foursquare API (https://api.foursquare.com/v2/venues/explore), we will calculate for each neighborhood its points of interest.
- Note, that for the use of the API we need to configure a limit and radius of results for each neighborhood. For this instance we will configure the radius at 8000 meters and the limit of results at 500 units.
- Once the API was used, we found 3904 points of interest, and for each of these its category, latitude, longitude, name, type, etc.
- Finally, what we will do, using the One Hot Encoding technique and grouping data will be to generate a table that for each neighborhood we can see how many points of interest it has separated by category.

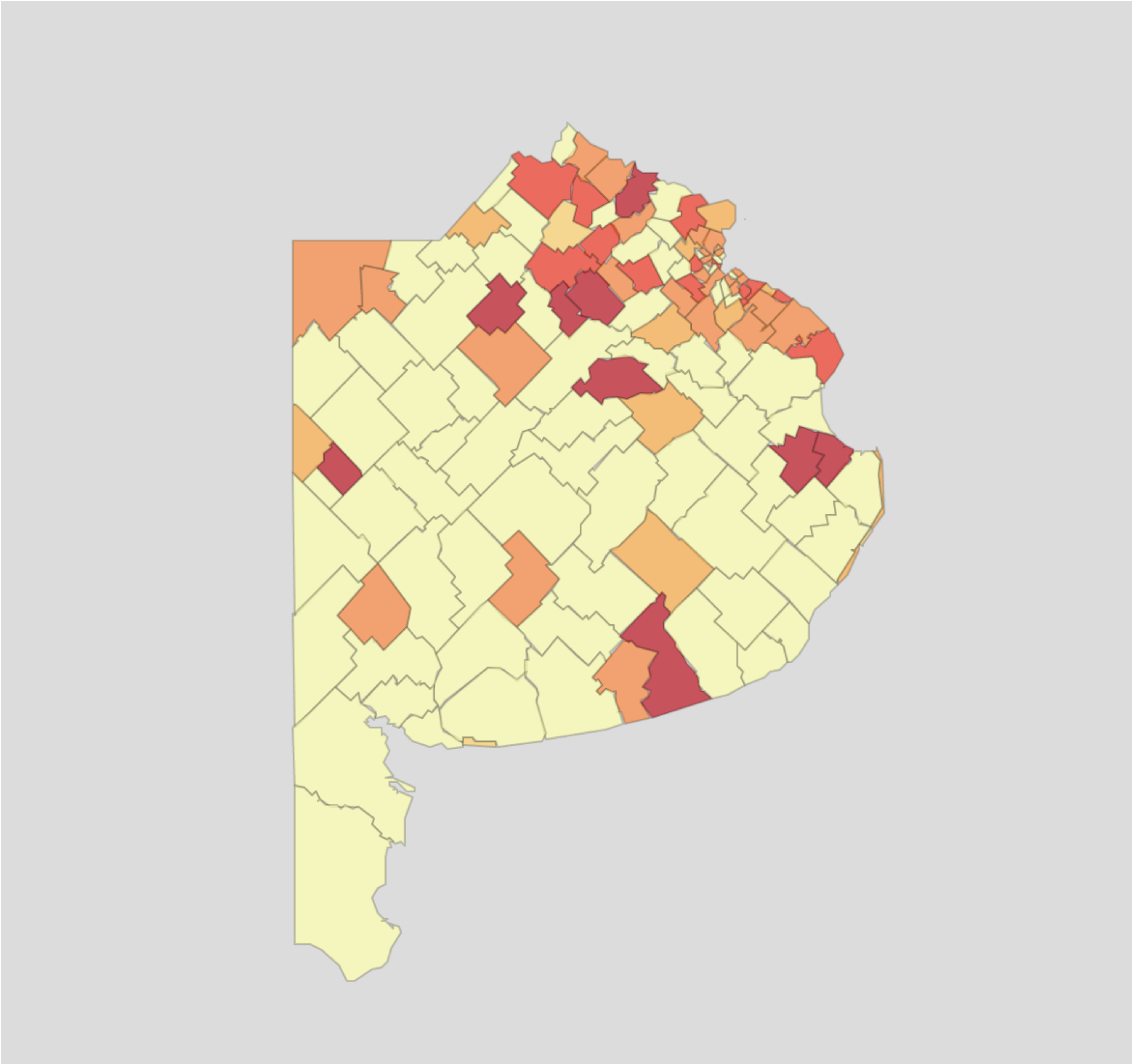
Data analysis

	Neighborhood	Zoo	Accessories Store	Airport	Airport Lounge	Airport Terminal	American Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop
0	9 de Julio	0.0	0.00	0.0	0.0	0.5	0.000	0.500000	0.00	0.00	0.000000	0.000000	0.0	0.0
1	Alberti	0.0	0.00	0.0	0.0	0.0	0.000	0.000000	0.00	0.00	0.000000	0.000000	0.0	0.0
2	Almirante Brown	0.0	0.00	0.0	0.0	0.0	0.000	0.032258	0.00	0.00	0.016129	0.032258	0.0	0.0
3	Arrecifes	0.0	0.00	0.0	0.0	0.0	0.125	0.250000	0.00	0.00	0.000000	0.000000	0.0	0.0
4	Avellaneda	0.0	0.00	0.0	0.0	0.0	0.000	0.130000	0.02	0.01	0.000000	0.000000	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
106	Tres Lomas	0.0	0.00	0.0	0.0	0.0	0.000	0.000000	0.00	0.00	0.000000	0.000000	0.0	0.0
107	Tres de Febrero	0.0	0.00	0.0	0.0	0.0	0.020	0.050000	0.00	0.00	0.000000	0.020000	0.0	0.0
108	Vicente López	0.0	0.01	0.0	0.0	0.0	0.020	0.040000	0.00	0.00	0.000000	0.020000	0.0	0.0
109	Villa Gesell	0.0	0.00	0.0	0.0	0.0	0.000	0.075000	0.00	0.00	0.000000	0.000000	0.0	0.0
110	Zárate	0.0	0.00	0.0	0.0	0.0	0.000	0.666667	0.00	0.00	0.000000	0.000000	0.0	0.0

Data analysis

- We will add the information in the table for all the neighborhoods, to obtain a value per neighborhood and be able to draw it obtaining the following graph. It is important to note that not all the categories were added, but only the desired category:

Data analysis

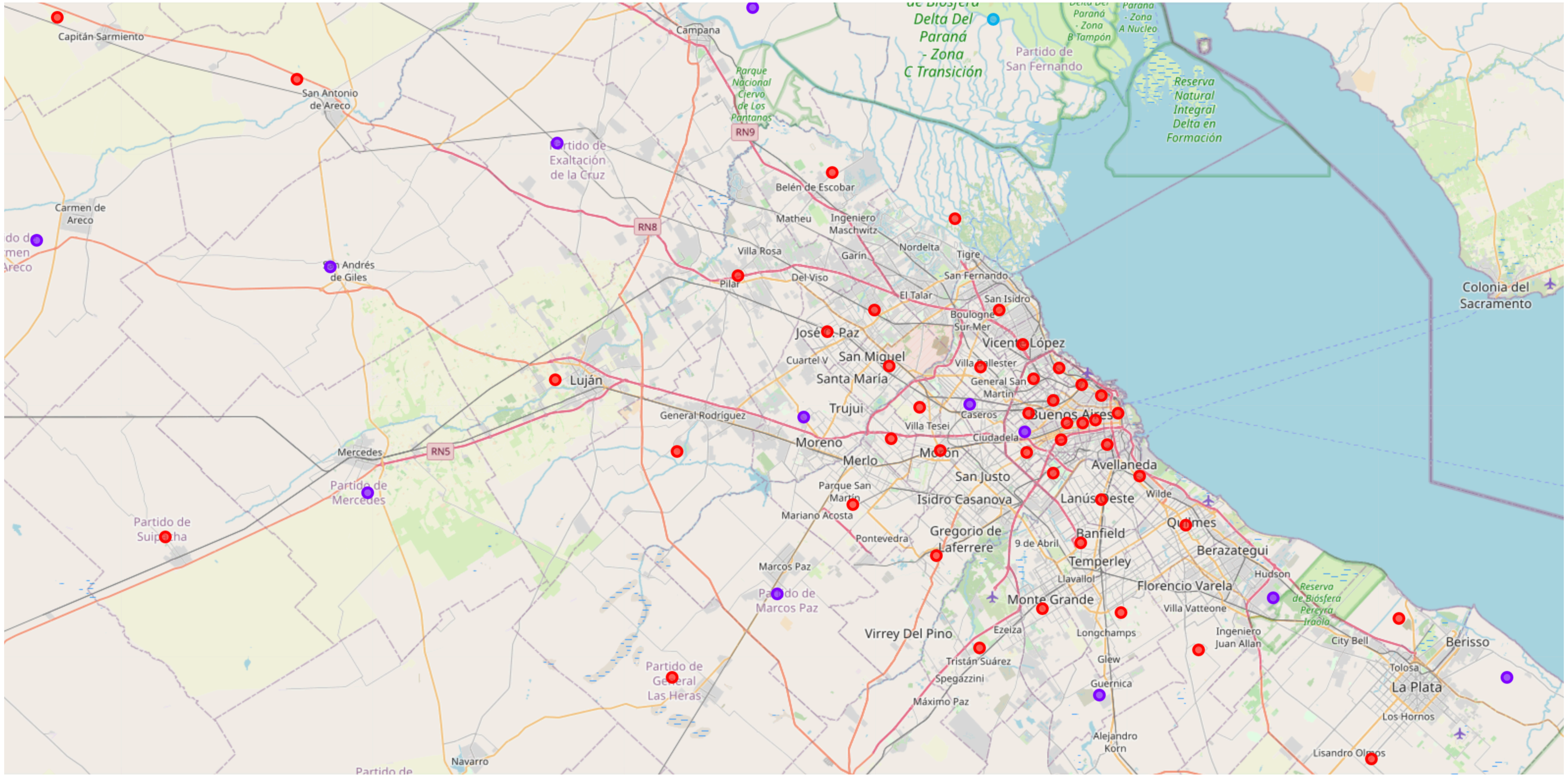


Cluster Neighborhood

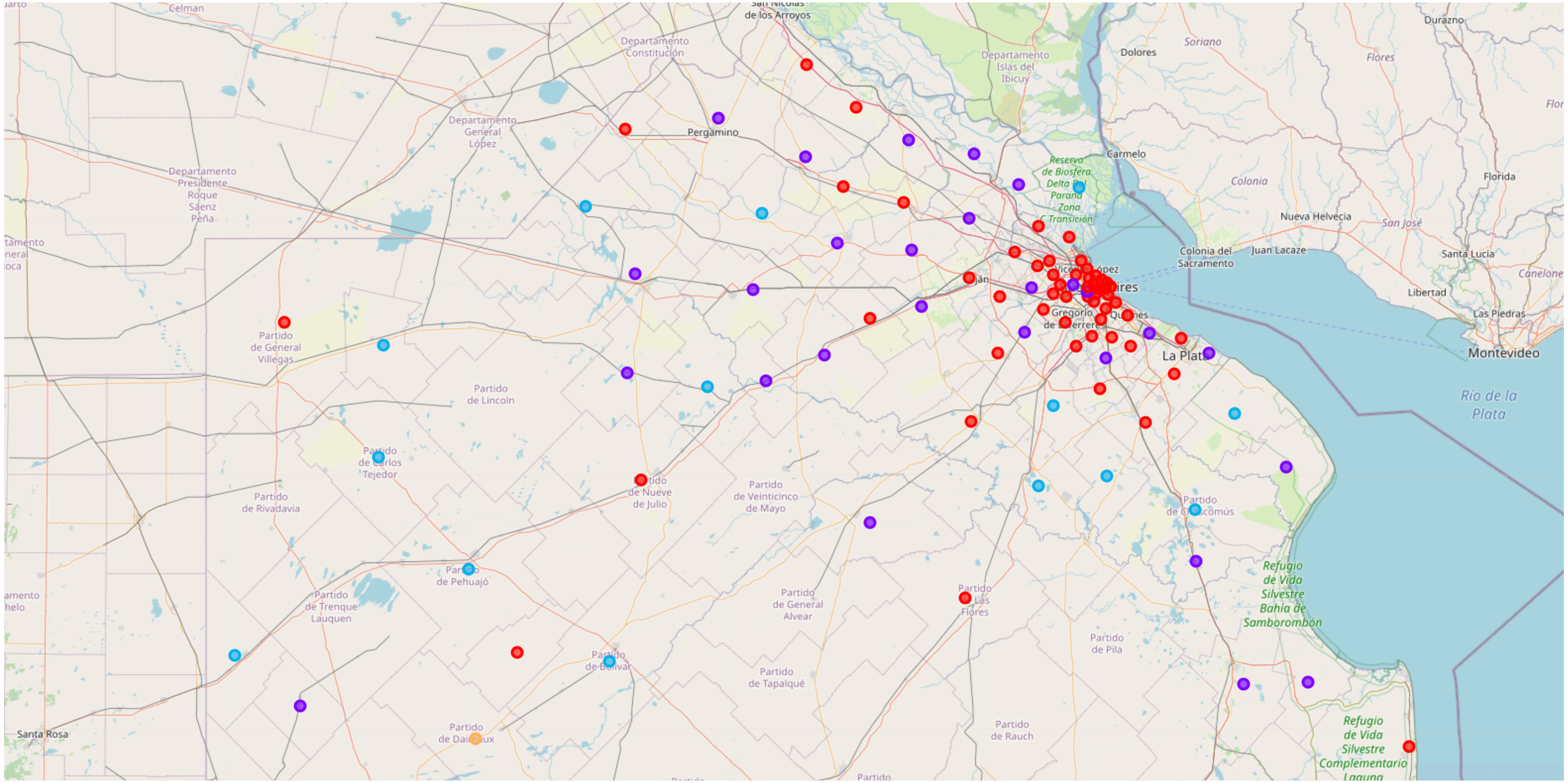
- With the Machine Learning algorithm called KMeans, we will generate 5 clusters to be able to find similarity between different neighborhoods, this could help construction companies in their analysis and thus determine where it is convenient for them to invest.

- To better see the results and see the clusters of our algorithm we draw it on the map

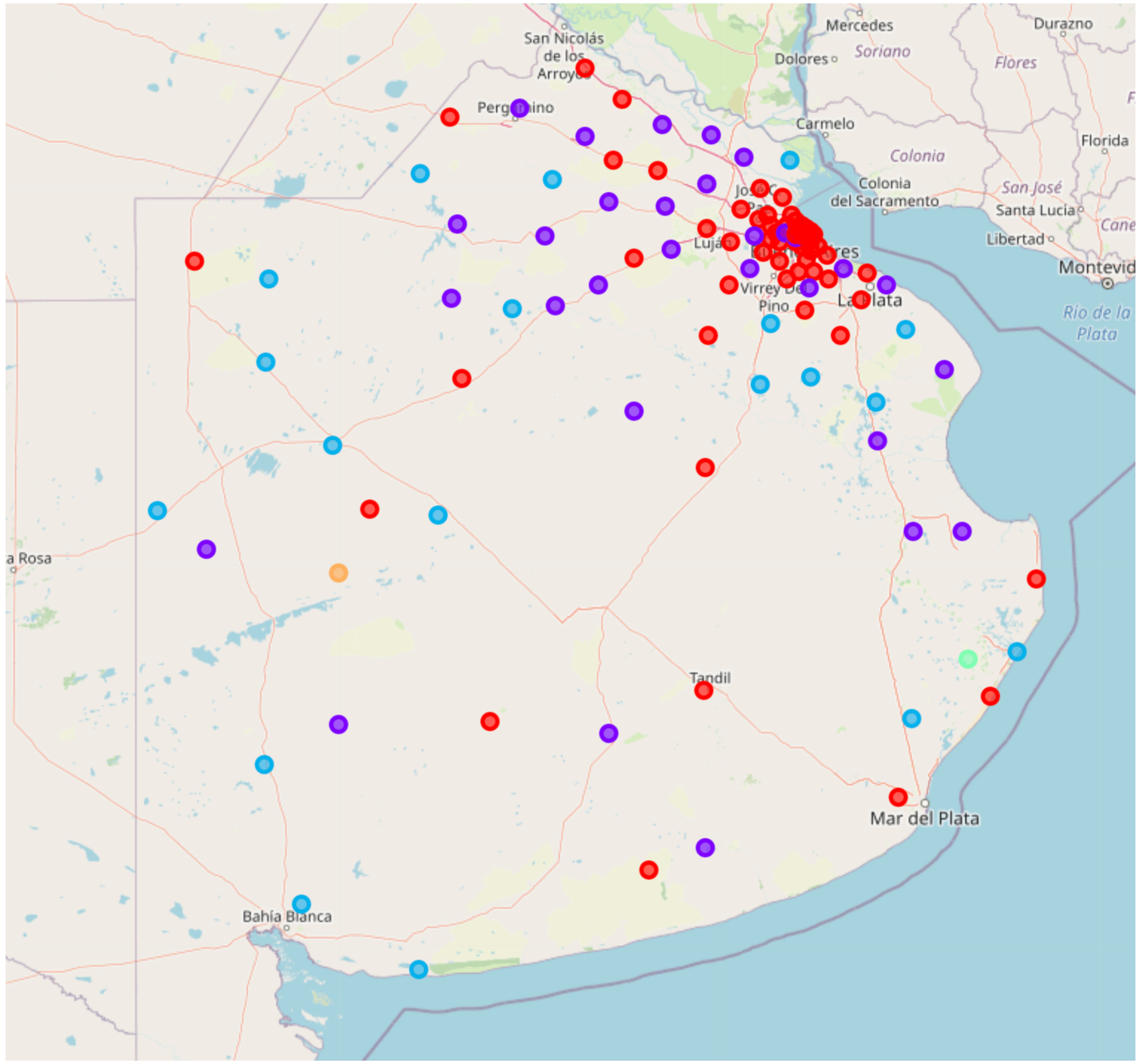
Cluster Neighborhood



Cluster Neighborhood

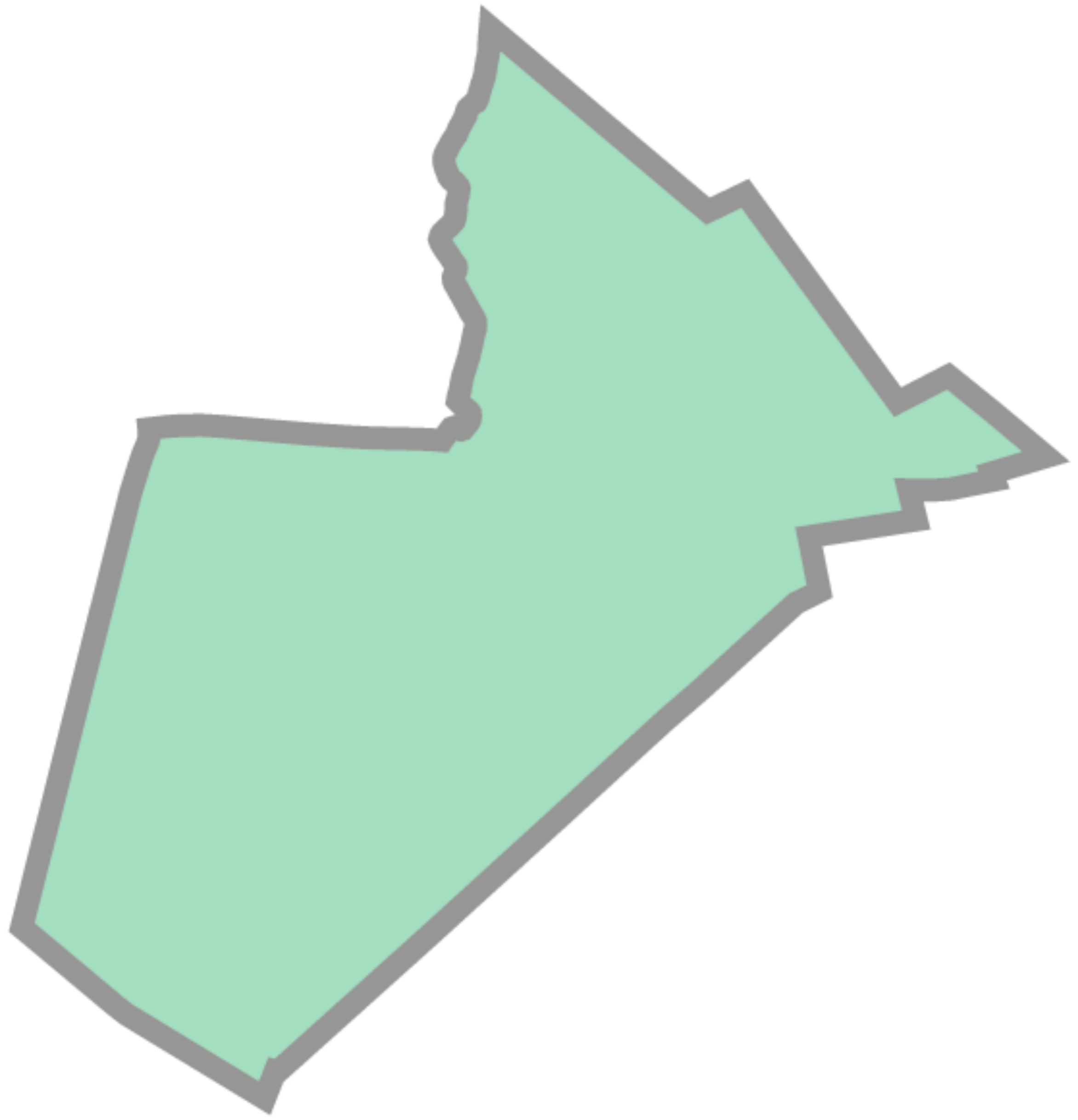


Cluster Neighborhood



Analysis of a particular neighborhood

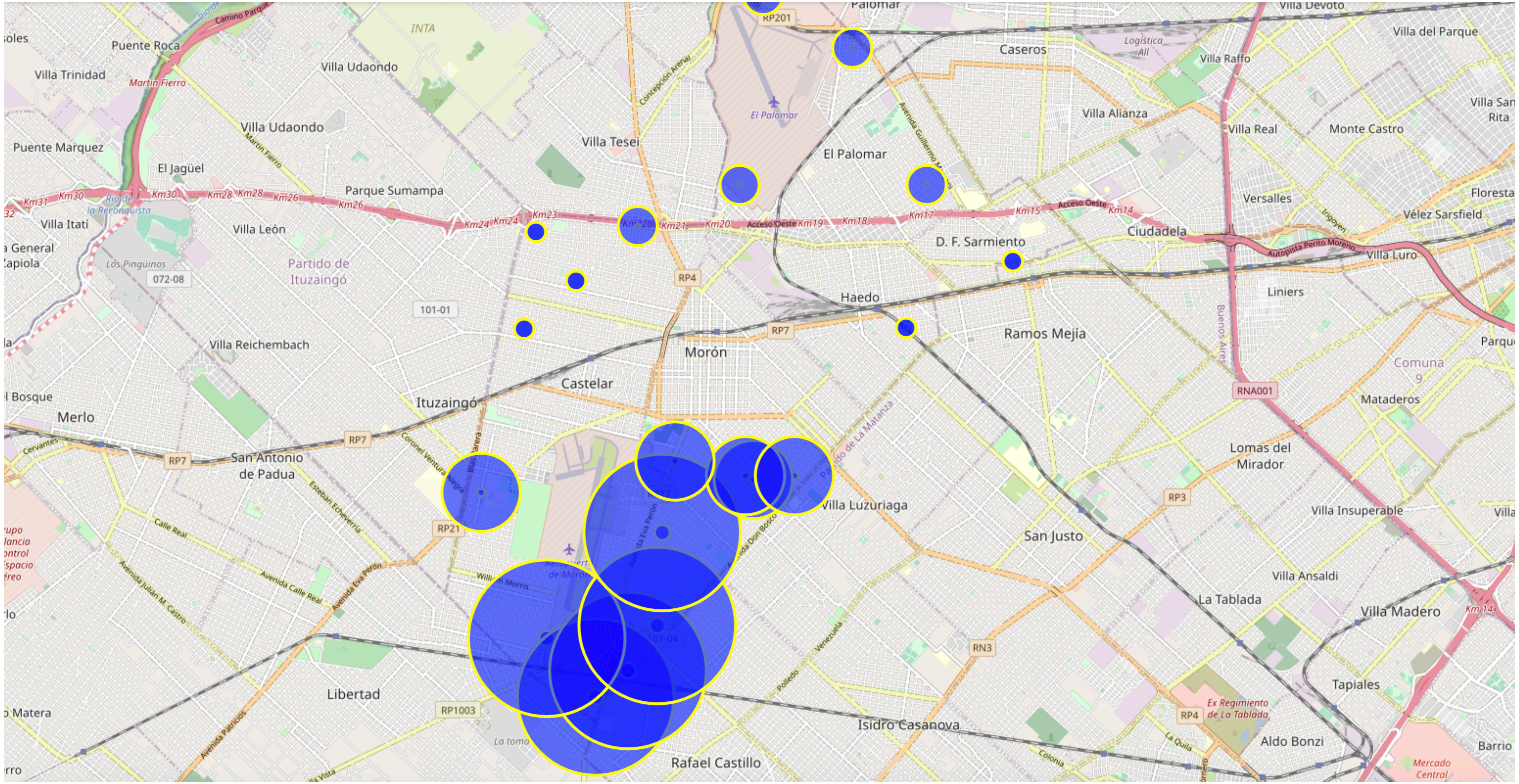
- We will evaluate the same logic used in the previous chapter, but instead of the entire province of Buenos Aires, we will only take the neighborhood called Morón.
- The first thing we will do is determine the geographical limits of the neighborhood and obtain its polygon.



Analysis of a particular neighborhood

- We will generate random points within the limits of the neighborhood. For this example we will execute only 20 points.
- We repeat the previous process, that is, we generate a table with One Hot Encoder and group by a unique identifier (in this case we use the tuple made up of latitude and longitude as the identifier). We calculate the total sum of the percentages of movement at each point of interest and make it on a map.
- You can see the marked points and the radius is equivalent to the percentage of movement of the point. That is, the more radius the circle has centered on the point, it is because that point has more movement than the desired points of interest.

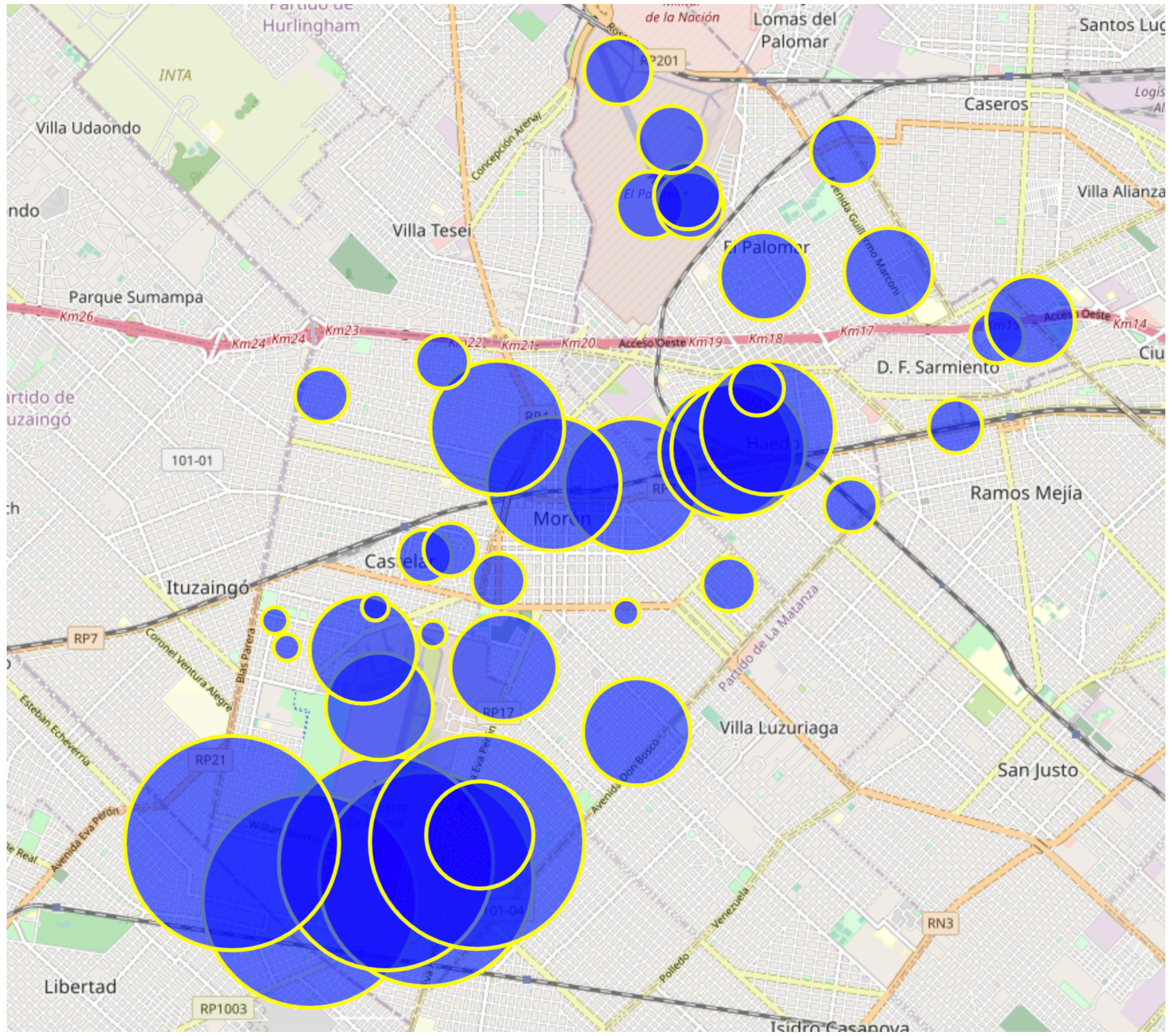
Analysis of a particular neighborhood



CHRISTIAN SEBASTIAN RUSSO, BUENOS AIRES 2020

Analysis of a particular neighborhood

- Now let's try 40 new random points



Thank you