22/02/2020

(A simple) Data Science Experiment:

Where would you open...
a new Pizza Restaurant in Buenos Aires?

By **Santiago Maraggi**

Final work for the IBM Data Science specialization (Coursera)

Email: smaraggi@gmail.com

LinkedIn: https://www.linkedin.com/in/santiagomaraggi/

Business Problem / Introduction

The goal of this work is to recommend a neighborhood to open a new Pizza Restaurant in Buenos Aires City.

Buenos Aires is a high profile touristic city and the Capital City of Argentina. Founded two times, in 1536 and in 1580, it has a rich colonial history, it had defended two big scale invasions during the beginning of 19th Century and it became a very important hub for commerce and politics, since the 18th. Currently, the city's Italian influences in food and general culture are widely known and it is also famous for being the origin of Tango music.

Pizza, on the other hand, is the most well known Italian dish all over the world. Businessmen and investors always struggle to decide good places to set new businesses.

Finding an appropriate environment for an entrepreneurship is a challenging task, to ensure best chances for the business to flourish. Best places for common business, however, tend to be overpopulated with well-established actors.

This work proposes to apply some basic data science techniques combined with geographical information in order to group common neighborhoods and select from the best possible group, the neighborhood less populated with Pizza Restaurants.

Data Section

First, the neighborhood information will be extracted from the official Buenos Aires Government Data service (BA Data).

Official website is:
https://data.buenosaires.gob.ar/

From this data service, the neighborhood areas will be obtained, and then, from these areas, the geographical centroid will be determined. A radius distance will be calculated from each neighborhood area, so that that area will be considered as if it were obtained from a perfect circle, and from that radius a representative fraction will be considered.

Buenos Aires neighborhoods geojson data:
http://cdn.buenosaires.gob.ar/datosabiertos/datasets/barrios/barrios.geojson

Once the neighborhoods centers and radio distances are determined, the FourSquare API will be used to get the venues representative for each neighborhood. For geographical visualization purposes, the library folium will be used, as well as other specific purpose Python libraries were imported as required, such as: pandas, numpy, json, geopy, requests, matplotlib, sklearn, urllib and math. The library os was also imported to cache locally the results of some queries, in order to enhance the testing cycles and not to overload the free licence calls quota for the Foursquare API.

The most common venue types for each neighborhood will be determined, and from this common data, neighborhood clusters will be determined with a K-Means cluster algorithm. As it was detailed, the "best cluster" will be the one with more Pizza Restaurants, while the "best neighborhood" will be the one with less pizza restaurants from the "best cluster".

Methodology

In order to fulfill the purpose of this work, neighborhoods are going to be clustered in groups by similarities in the most common venues in each one of them.

The cluster with proportionally more Pizza Restaurants on it will be considered the most interesting to open a new business of this type, and then, the neighborhood from this cluster with less Pizza Restaurants will be considered a good candidate to open the new pizza shop.

This model makes a few assumptions and simplifications:

- The business is assumed to be of a well established company, so the aim is to compete in interesting places, rather than set a restaurant in areas completely without pizza restaurants.
- The neighborhood clusters, classified by “most common venues”, are assumed to be representative at some point.
- Neighborhood clusters with most pizza restaurants are considered to have common properties that makes them interesting for this kind of business.
- From the "most interesting cluster of neighborhoods", extracted with the mentioned criteria, and given that clusters are considered to be representative of common features that make them interesting for the target business type, the neighborhood with less pizza restaurants will be considered "underexploited", but still interesting.
- Venues for each neighborhood will be obtained from the FourSquare public API (free licence).
- Venues will be obtained from the geographic centroid of each neighborhood, within a radius proportionally related to the overall neighborhood area. This could leave some venues near the border of each neighborhood outside consideration, so we are assuming that statistically this will not be very significant.
- Centroid of neighborhoods, as they are not provided by BA Data service, will be calculated from the neighborhood polygon averaging extreme latitude and longitude, a fast method that is still considered to be representative.
- Of course, the first assumption is still that you want to open a Pizza Restaurant in Buenos Aires and make a profit out of it.

Execution

The mentioned original dataset obtained from the official Buenos Aires Government Data Service did not provide the geographical center of each neighborhood, but only their border coordinates in a geojson file.

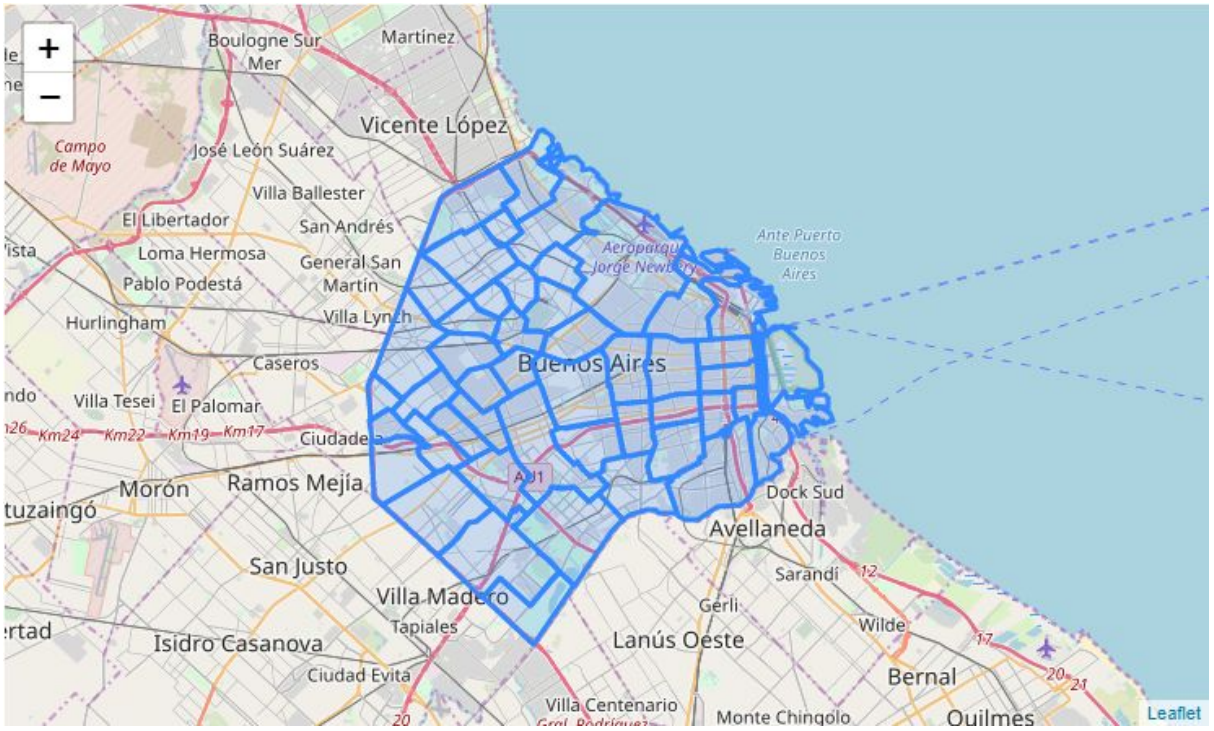

Figure 1. Neighborhood boundaries provided by the Buenos Aires Government official Data Service.

In order to determine the geographical centers, averaging the boundaries coordinates was discarded, because the most irregular and detailed segments of the boundaries would have more relative weight than straight long segments. An infinitesimal integration, in order to get the exact central value was considered out of the scope, given the goals and scope of the work. After a small analysis, it was decided that averaging extreme latitude and longitude values could result in a better approach. It was of interest to find one point that could be considered representative for the neighborhood. This could be not the best for non-convex neighborhoods, because the center could be very close to a border area and far from some other internal areas. However, still in these cases the next neighborhood was considered influential for the neighborhood center, so it was acceptable to capture some areas of a close neighborhood too in a few cases.

Exploring the dataset, the test for this technique was applied to the neighborhood of Palermo first.

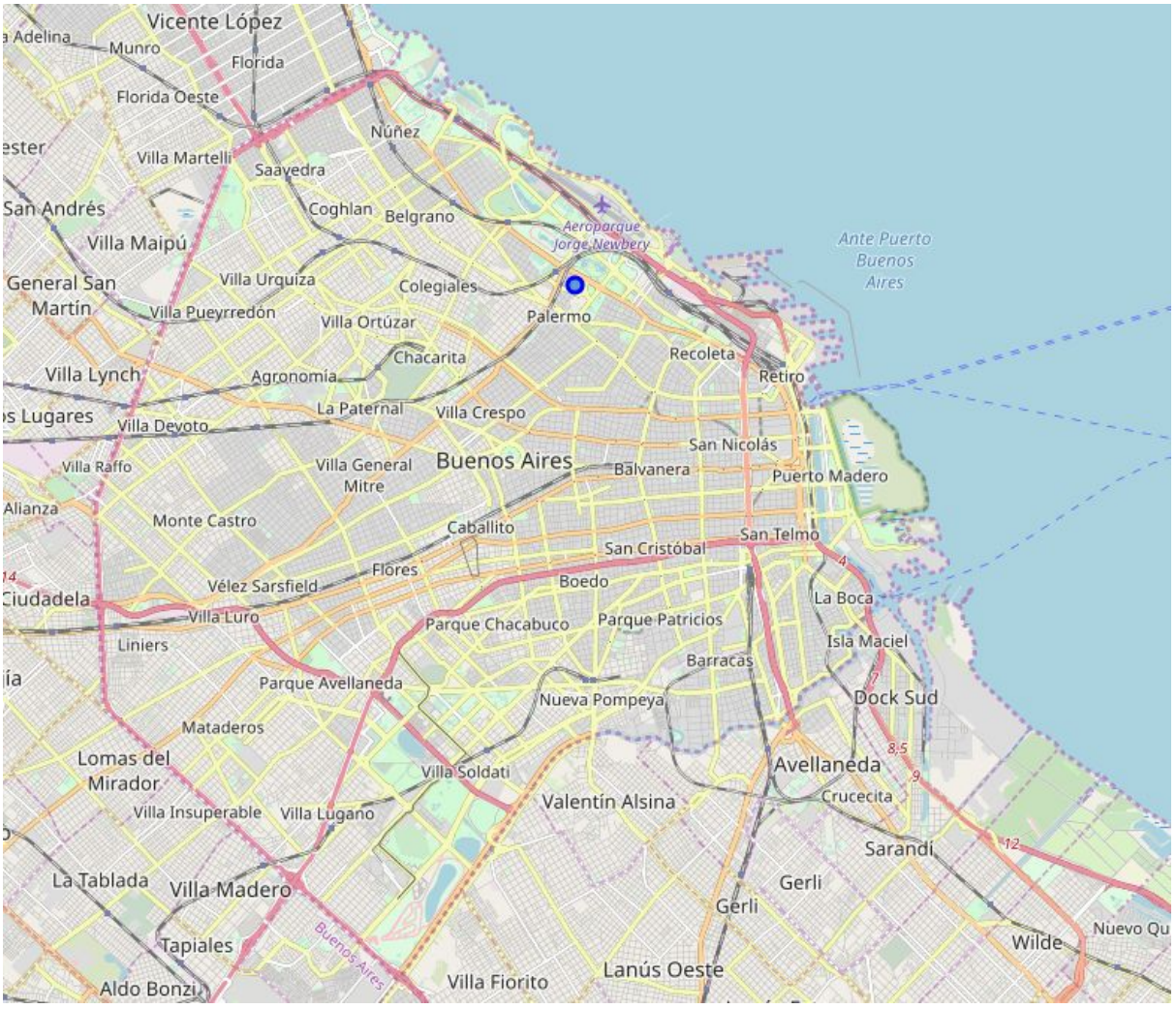

Figure 2. Estimated geographical centroid for the neighborhood of Palermo.

The result obtained for Palermo indicated that the technique could provide good approximations, so it was further tested with all the neighborhoods.

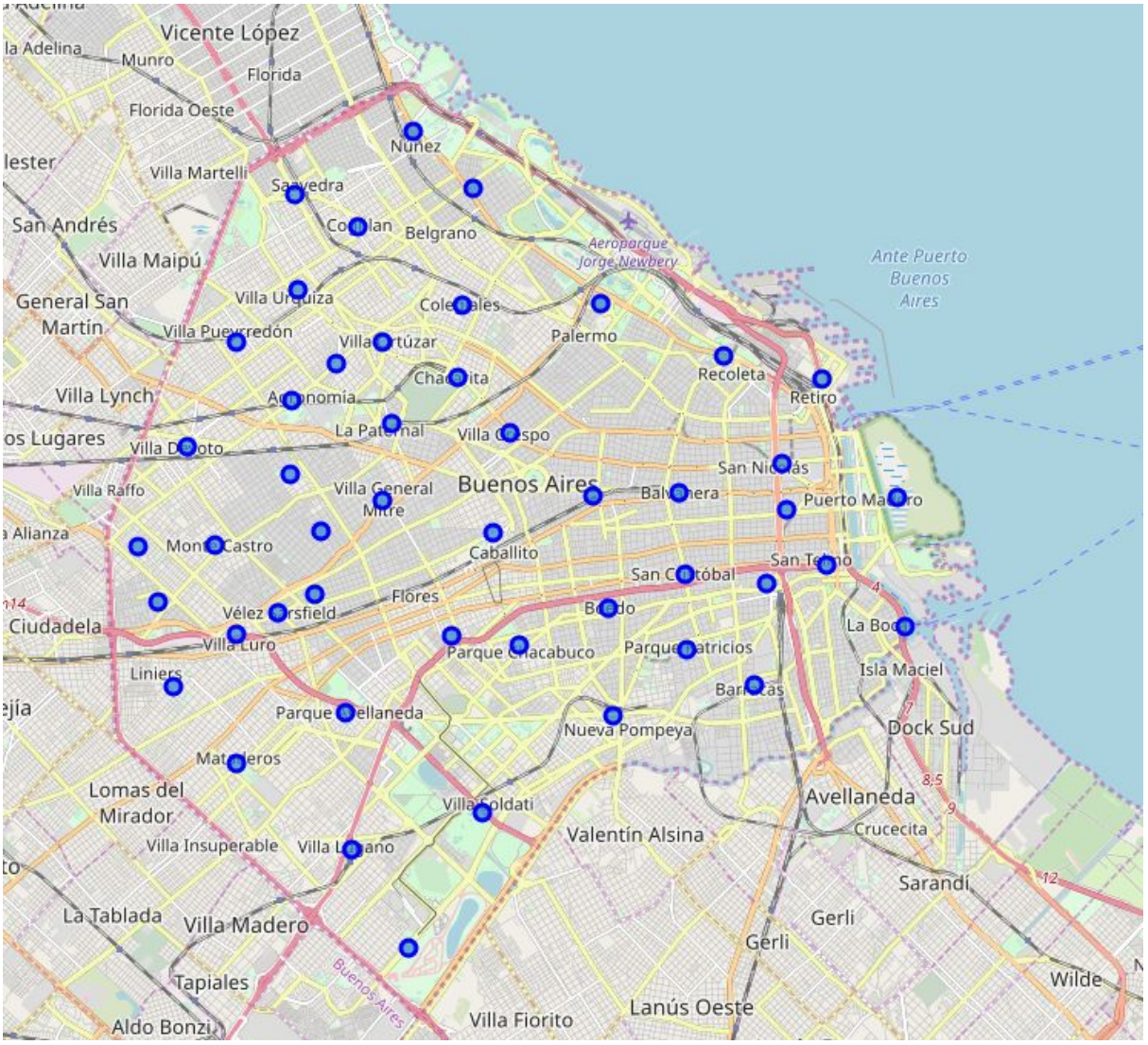

Figure 3. Calculated centers for all the neighborhoods of Buenos Aires.

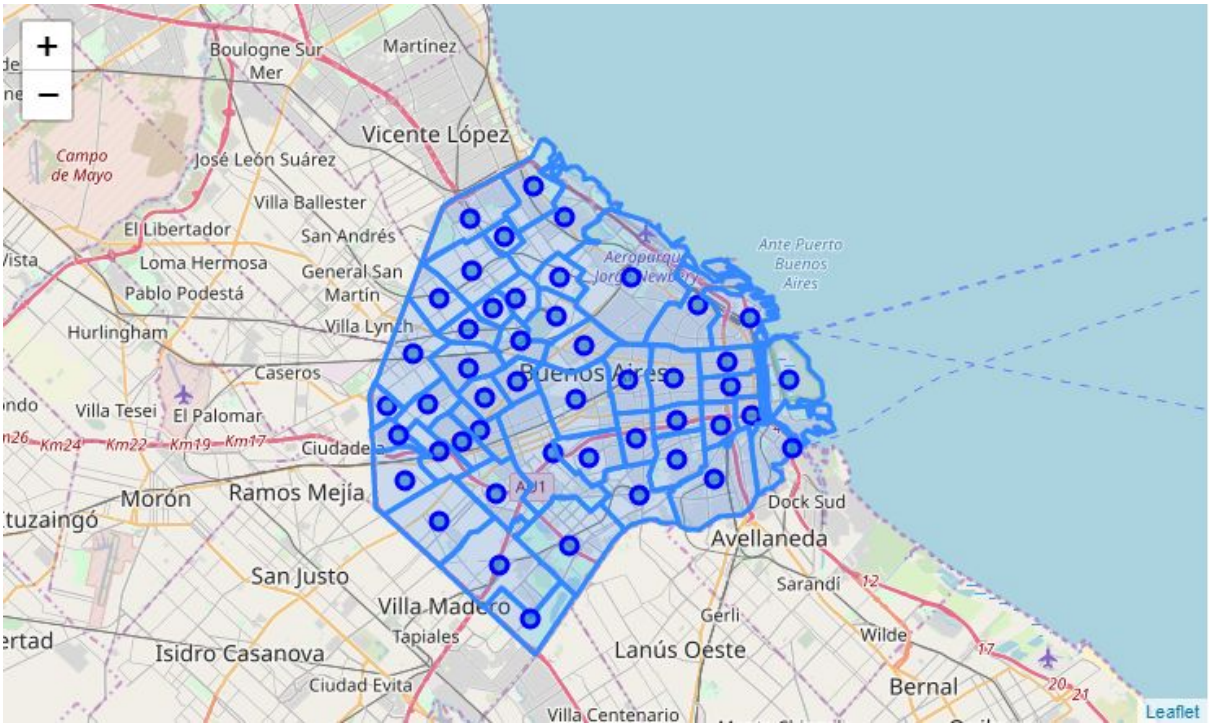

Figure 4. Comparison between the estimated neighborhood centers and the neighborhood boundaries provided by the official BA Data service. Non-convex neighborhoods can be seen with centers close to some borders, however still aligned with the logic of the implemented methodology.

For each neighborhood, then a radius formula was determined, in order to find a representative area for each one around its geographical center. The formula finally used was the following, given a “neighborhood_area”:

```
radius = math.sqrt(neighborhood_area / math.pi) * 3 / 4)
```

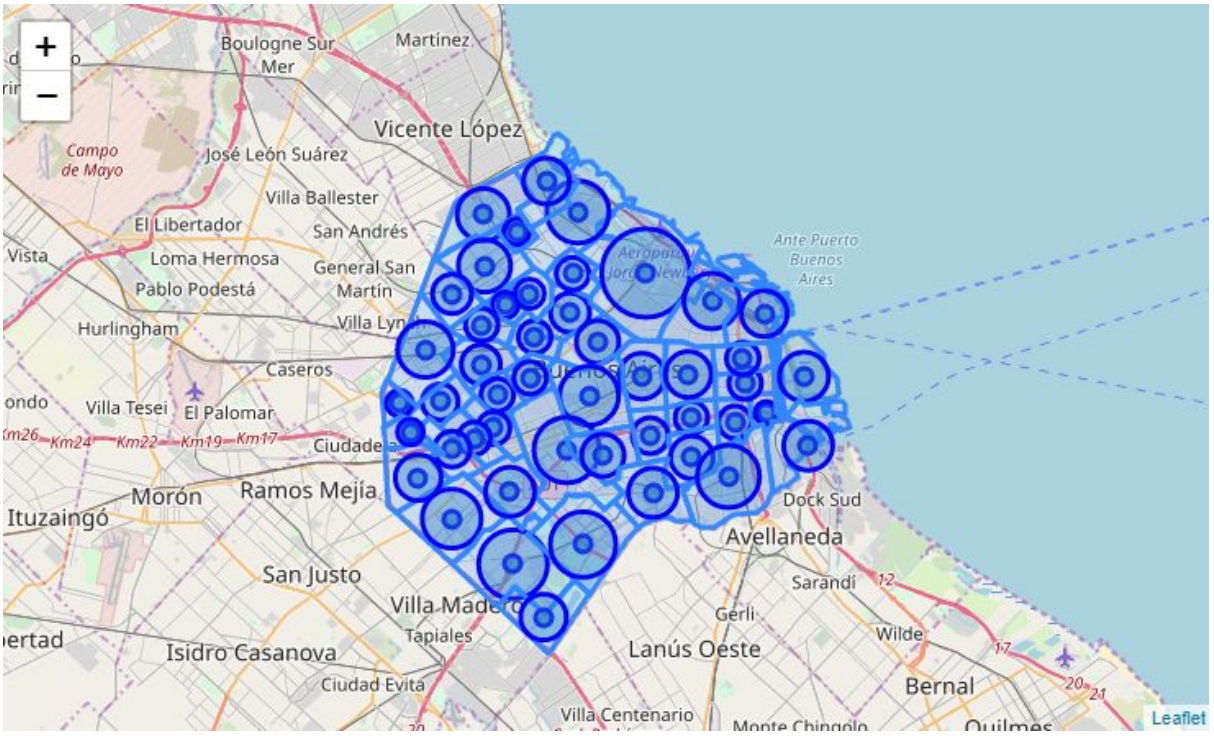

Figure 5. Neighborhoods centers with their radius of influence, compared against each neighborhood delimited area. Non-convex neighborhoods are more influenced by close neighbors, however for the purpose of this work this solution was found to be a reasonable approximation.

Results

Neighborhoods were grouped in clusters, using the K-Means algorithm for non-supervised classification, using the Python library for machine learning scikit-learn (sklearn).

The criteria for classification was “most common 50 venue types”. For each neighborhood, all venues within the radius of interest from the neighborhood center were obtained from the Foursquare API. Within each neighborhood, then, venues were grouped by type, and then, the most common venue types were established. The venues total amount was cut to a maximum of 100 for all the neighborhoods.

The finally obtained clusters were the following:

CLUSTER 1
MATADEROS VILLA LUGANO NUEVA POMPEYA LINIERS

CLUSTER 2
BOEDO VÉLEZ SARSFIELD

CLUSTER 3
CHACARITA VILLA CRESPO VILLA DEL PARQUE ALMAGRO CABALLITO VILLA SANTA RITA FLORES FLORESTA VILLA LURO PARQUE PATRICIOS SAN TELMO SAAVEDRA COGHLAN VILLA URQUIZA COLEGIALES BALVANERA AGRONOMÍA VILLA ORTÚZAR BARRACAS PARQUE CHACABUCO PALERMO VILLA DEVOTO VERSALLES PUERTO MADERO MONSERRAT SAN NICOLÁS BELGRANO RECOLETA RETIRO NÚÑEZ BOCA

CLUSTER 4
VILLA REAL SAN CRISTÓBAL VILLA GENERAL MITRE VILLA PUEYRREDÓN

CLUSTER 5
PATERNAL

CLUSTER 6
PARQUE AVELLANEDA

CLUSTER 7
VILLA RIACHUELO

CLUSTER 8
VILLA SOLDATI

CLUSTER 9
MONTE CASTRO CONSTITUCIÓN

CLUSTER 10
PARQUE CHAS

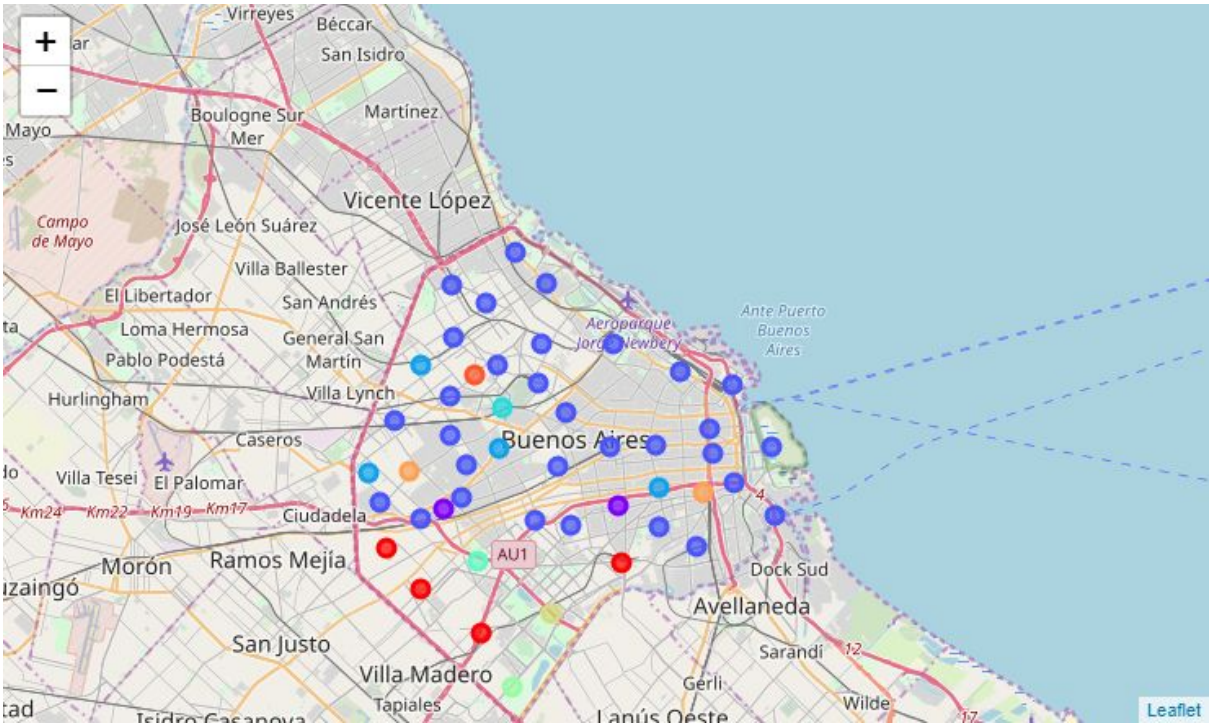

Figure 6. Neighborhood centers signaled with different colors according to their cluster number. Characterizing neighborhoods by their most common venues, it was found that most neighborhoods belong to a few clusters, and some particular neighborhoods remain alone differentiated from the rest.

DISCUSSION

Goal of the Work

As stated initially, the goal is to find the most interesting cluster of neighborhoods to install a Pizza Restaurant ("Pizza Place", in terminology of the Foursquare API).

We are considering for this work that the interest is to install a high-end touristic oriented Pizza Shop, and not a simple family business, so the idea is to compete in an interesting neighborhood, that shares characteristics with neighborhoods with successful "Pizza Places", but still is the less populated with business of this type from its group ("cluster").

A Few Important Model Limitations

It is to note that, among neighborhoods, many different economic and development indicators could be found. For this regards, venue types were expected in some way to be representative of economic development, however a further analysis could consider more specific and precise indicators, as this study only focuses on “venue types” within neighborhood locations.

Venue types, on the other hand, are only obtained by the Foursquare API. This leaves outside consideration a lot of businesses and venues that are not registered in this geographical information service. Crossing different geographical services could help to enhance the results of this study. Also, venue types are considered as presented by this API. Grouping all types of restaurants, for example, could also improve neighborhood characterization.

Single Neighborhood Clusters

Neighborhood classification for this work is achieved by most common venue types. The amount of "most common venue types" and the total of "clusters" for neighborhood classification were the 2 main variables to play with.

The occurrence of single neighborhood clusters appeared to be a big deal when running the K-means cluster algorithm with 5 clusters and only the 10 most common venues, leaving 4 clusters with only one neighborhood and one with all the rest. Increasing the numbers to 10 clusters and 50 most common venue types, still 5 clusters only contain more than one neighborhood, so that is to say that differences of single neighborhood clusters, with this classification criteria, are quite stable.

Single neighborhood clusters, classified by this method, were found to contain in general not the most developed neighborhoods of the city in general. These neighborhoods will be discarded, because it is of interest to install a competitive business, and not to find a hot-spot in areas that are not common in terms of commercial venues, given their apparent unique characteristics in relation to the most traditional neighborhoods of the city.

Most Interesting Clusters Analysis

After the previously mentioned considerations, clusters 1, 2, 3, 4, and 9 remained for further analysis to be selected as most interesting to determine the best candidates.

Cluster 1

This cluster would be the undisputable winner, according to the criteria previously established. All four neighborhoods of this cluster have “Pizza Place” as the most common venue type, however this leaves no room to select the “best candidate” of this cluster to recommend. Given this finding, this cluster is finally not recommended for a new pizza restaurant, given it seems to be saturated for this type of business.

As a side comment, these 4 neighborhoods are not the most developed of the city. This opens the suggestion that a new further research with some more specific socio economic indicators could be of some use for this research.

Cluster 2

Boedo and Vélez Sarsfield were found to share the same cluster alone. Boedo has “Pizza Place” as the 4th more common venue, and Vélez Sarsfield has it as its 11th. Among the two, Vélez Sarsfield would be the most recommendable neighborhood for a new Pizza Restaurant.

Cluster 3

This is the most important cluster, containing 31 out of the total 48 neighborhoods. This big cluster could suggest that further refinement classification methods could be of good use, to refine the results of this work. On this cluster, 21 out of 31 neighborhoods have “Pizza Place” on their top ten most common venues, so this could be considered a cluster that groups neighborhoods with common characteristics that make pizza restaurants prosper.
Good candidates of this cluster could be considered the neighborhoods that don’t have “Pizza Places” among their top ten most common venues. Those neighborhoods would be:

CLUSTER 3 RECOMMENDED NEIGHBORHOODS	
NEIGHBORHOOD	“PIZZA PLACES” venue type commonness position
Palermo	>50th
Puerto Madero	>50th
Floresta	>50th
Retiro	>50th
Versalles	>50th
Saavedra	44th
Monserrat	20th
Recoleta	18th
Villa Ortúzar	16th
Villa Santa Rita	12th

Cluster 4

This cluster has only 2 neighborhoods with “Pizza Places” in their top ten venue types (3rd for San Cristóbal and 10th for Villa General Mitre) and 2 neighborhoods without “Pizza Places” in their top 50 venue types, so the cluster could not be determined to be significant to group convenient neighborhoods for Pizza Restaurants.

Cluster 9

This cluster only contains two neighborhoods, for which “Pizza Places” is 2nd and 7th respectively. The cluster, similar to cluster 1, seems to group adequate neighborhoods for Pizza Restaurants, however all the members seem to be highly saturated with this type of venue. No recommendation can be extended from this cluster, then, however the neighborhood of Constitución could be established as more convenient than Monte Castro.

CONCLUSIONS

- Neighborhoods could have been grouped successfully according to their most common venue types, however some other elements, such as socio economic indicators could result of use to enhance the classification results.
- The winner cluster, cluster 1, could not be of use to recommend a place of convenience to open a new Pizza Restaurant, given that all the neighborhoods within it had “Pizza Places” as the most common venue type.
- Socioeconomic indicators for neighborhoods were left appart from this study. It would result in a broader analysis to determine the development of neighborhoods, as “Pizza Places” could comprehend small businesses and top international restaurants as well. This study, however, could be considered a first guide for further refinement, given it provides a full scope of venue types and their locations all through the city of Buenos Aires.
- Many neighborhoods remained alone in exclusive clusters, even changing the classification algorithm variables and increasing the scope of analysis. This indicates that, in terms of most common venue types, there are neighborhoods that are significantly different from the rest. From the results obtained, it can be seen that these are more commonly peripheral neighborhoods or neighborhoods being developed, mostly and in general, rather than traditional and well established commercial and cultural hubs of the city.
- This work could provide a general insight and very basic guide to start an analysis to open a new business type, in this case applied to Pizza Restaurants. It is important, however, to understand the limitations of the model and the simplifications made along the way, to use these results as an initial guide and an introductory and exploratory analysis.
- **For an investor, it would be recommended to open a new Pizza Restaurant in any of the neighborhoods recommendable from the cluster 3, selecting one that goes along with development and economic indicators preferred by the entrepreneur, depending on the targeted customer audience.**