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

Santiago Maraggi

Grupo de Investigaciones en Ciencias Informáticas (REINVENT), Facultad de Ingeniería, Universidad de Buenos Aires
CITEDEF, Instituto de Investigaciones Científicas y Técnicas para la Defensa

{smaraggi}@gmail.com
https://www.linkedin.com/in/santiagomaraggi/

Abstract. The goal of the work is to extract an initial guideline to determine the best place to open a new Pizza Restaurant in Buenos Aires.

In this work the K-Means algorithm was applied to classify the neighborhoods of Buenos Aires, according with their most common venue types. For each neighborhood, top 50 most common venue types were established, and then, 10 neighborhood clusters were obtained with the mentioned algorithm in order to provide some clues about which neighborhoods could be best investment options.

Keywords: Business Analysis, City Venue Types, Neighborhood Clustering.

1 Introduction

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

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

This work proposes to combine some basic data science techniques applied to geographical information in order to cluster neighborhoods alike, in terms of their most common venue types, and select from the considered best possible group, neighborhoods less populated with Pizza Restaurants.

2 Data Section

First, the neighborhoods information was extracted from the official Buenos Aires Government Data service [1]. Neighborhood areas were obtained from this service and then, the geographical centroids were determined. A radius distance was established for each neighborhood, derived from its total area, in order to set a representative area for each neighborhood, based on the calculated geographical center and distance radius.

With the neighborhoods centers and radius determined, then the Foursquare API [2] was used to get the venues to characterize each neighborhood. For geographical visualization purposes, the library Folium was used, as well as other specific purpose Python libraries that were imported as required, such as pandas, numpy, json, geopy, requests, matplotlib, sklearn, urllib and math. The library os was also imported in order to cache locally the results of some queries, in order to enhance the testing cycles and not to overload the free tier license of Foursquare API used.

With the information provided by Foursquare, the most common venue types for each neighborhood were determined, and then, neighborhood clusters were built with the K-Means algorithm. The “best clusters” were determined selecting those with proportionally more Pizza Restaurants among other types of venues, considering these neighborhoods more likely to provide good conditions for these type of business, while the best neighborhoods within these clusters were selected with proportionally the less quantity of Pizza Restaurants among other venue types, making the assumption that, as these neighborhoods belong to “good clusters”, they still could be underexploited good candidates.

3 Methodology

To develop this work a Python environment was used within a Jupyter Notebook, running in the IBM Cognitive Labs [3] cloud service for machine learning and data science.

The neighborhoods were clustered in groups according with their 50 most common venue types, according to the Foursquare API results. The cluster with proportionally more Pizza Restaurants would be considered the best, and from it, the neighborhoods with proportionally less Pizza Restaurants would be the best candidates.

This model makes some assumptions and simplifications. The goal would be to set a new Pizza Restaurant in an interesting place, rather to a lonely area without competition. The neighborhoods clustering assumes that “venue types commonness” is, at the end, representative to characterize neighborhoods, clusters with higher “pizza restaurant commonness” would reflect good conditions for this particular type of business, while neighborhoods with low “pizza restaurant commonness” within these good clusters would be “underexploited”, but still interesting. The Foursquare API is also assumed to provide representative results, as these could be biased by the Foursquare business registration process, for example. Venues were obtained by each neighborhood representative area as described before, and again, this simplification is considering another simplification that could affect the results for some areas. Also, the centroid of neighborhoods were calculated averaging extreme coordinate values for each polygon, given the centroids were not provided by the official BA Data service, and that averaging for example all the points could lead to errors, because of irregular edges being more weighty than long regular segments, and the radius was determined from the overall neighborhood area. The "representative" area of neighborhoods finally is assumed to effectively being "representative".

Socio economic indicators from neighborhoods were left apart in this analysis, as the goal is to study interesting places in general for a Pizza Restaurant, from which a later economic consideration could be used to select from. Also, for neighborhood clustering, population considerations as used by Delmelle [4] were left apart, as the focus of this work was put purely in the commercial profile of neighborhoods.

4 Execution

BA Data service was used to determine the centroids for each neighborhood and then the area of each one was used to determine a radius distance. The distance formula used was finally the following (equation 1):

$$Radius = math.sqrt(neighborhood\_area / math.pi) * 3/4 \tag{1}$$

To explore the dataset, the Palermo neighborhood was used as a first testing ground. Then, the same methodology was applied to all neighborhoods. Some data visualization maps were issued to make sure that the results were reasonable. It was found that non-convex neighborhoods were more influenced by close neighborhoods, given sometimes the representative area showed an overlap with these border areas. A partial solution for a next iteration could be to use a few more centroids for each neighborhood and make a new call to the Foursquare API for each, to refine the neighborhood venue compilation.

With the neighborhoods data and the venues collected from their representative areas, 10 clusters were built with a K-Means algorithm (unsupervised learning), and considering the 50 most common venue types of each neighborhood.

5 Results

Table 1. Clustering Results (neighborhoods separated by comma)

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

6 Discussion

Lonely cluster neighborhoods were stable for the algorithm with different parameters (mostly varying the most common venue types amount to consider, from 10 to 50) and total cluster quantity. That is to say, in regards with the "venue types", the classification results seemed stable. The starting method to determine initial cluster amount was the "rule of thumb", proposed by Kodinariya and Makwana [5].

Single neighborhood clusters, obtained by common venue types, were found to contain in general the less developed neighborhoods. These were discarded because the interest was to install a "pizza place" (in the Foursquare API terminology) in a competitive and interesting place, rather than in a remote unexploited and atypical place.

Cluster 1 would have been the indisputable winner, according to the established criteria, however it was discarded because all four neighborhoods of this cluster had "pizza place" as the most common venue type, among the other 49, leaving no place for underexploited neighborhoods between them. From cluster 2, Boedo had "pizza place" as the 4th most common venue type, while for Vélez Sársfield it was 11th, so the later would be more advisable as less crowded with Pizza Restaurants, still in an interesting cluster. Cluster 3 was considered the "winner". From it, the most interesting neighborhoods were the following, considering "Pizza Place commonness": *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 was not concluded to be good for "pizza places", while cluster 9 yes, but being in 2nd and 7th position, it was discarded for being both neighborhoods plenty exploited in relation to the previously mentioned.

7 Conclusions

Neighborhoods could be classified with a K-Means algorithm, based on most common venue types. Single neighborhood clusters seemed stable, and neighborhood commercial affinities were reflected in the classification.

For an investor, according to the obtained results, it would be recommendable to open a new Pizza Restaurant in any of the mentioned neighborhoods from cluster 3, selecting one that goes along with development and socio-economic indicators preferred by the entrepreneur and depending on the targeted customer audience and the desired level of investment.

8 References

1. Buenos Aires Data: https://data.buenosaires.gob.ar/ (visited on 10/03/2020)

2. Foursquare API: https://developer.foursquare.com/ (visited on 10/03/2020)
3. IBM Cognitive Labs: https://labs.cognitiveclass.ai/ (visited on 10/03/2020)
4. Delmelle, E.C.: Five decades of neighborhood classifications and their transitions: A comparison of four US cities, 1970-2010. Applied Geography, 57, 1-11. 2015.
5. Kodinariya, T.M., Makwana, P.R.: Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies. 2013.