Clustering urban land candidates for a new restaurant in the city of Buenos Aires

Luis Martín Ghiglia

June 21, 2020

Introduction

Background

Buenos Aires is a city far more European than South American, an urban landscape steeped in tradition but offering modernity that never disappoints. From its numerous museums to its old-time cafés and milongas, or dance halls, for the seductive Tango, one will never find a dull moment. Watch a soccer match in the famous La Boca stadium and experience the bliss of *fútbol* (soccer) in Argentina. Wander the bustling streets and grand avenues of the city and then explore the numerous historic neighborhoods. Buenos Aires vibrates with the history and energy brought by numerous European immigrants at the turn of the 20th century. Visit La Boca, with its brightly painted homes, or San Telmo; both neighborhoods were founded by Italian immigrants. Then cross over to Recoleta or Barrio Norte, which boast a French aristocratic style.

One of the most recognized characteristics of the city is its cuisine. The cosmopolitan capital draws from Italian, German, British, Spanish, Arabic, and native cultures. Its beef is nonpareil, yet Buenos Aires also has superb fruit and vegetables, outstanding wine, wonderful dairy products, and sweets prepared in ways that borrow from the city's diverse cultural heritage. Less noted is the fact that Argentina has a long coastline teeming with fish, which is increasingly taking its place next to beef and lamb. Few cities benefit so completely from homegrown foods, making Buenos Aires both self-sufficient and supremely well-fed.

This situation turns commercial activities related to food as one of the main practices for new enterprises.

But, on the other hand, Buenos Aires, and Argentina in general, has been suffered the impact of economic and financial problems which generates issues related to pricing volatility, urban land offer shortage, and infrastructure problems, among others.

Problem

For all this, the risk for a new business implementation can be high, and some questions about the reliability of it appear:

- ***Is there a good offer of urban land?***
- ***Where do I have to install my new restaurant?***
- ***Is the price fair?***
- ***Are there structural problems, like waterlogging, which can compromise my business?***
- ***Will I have enough customers to make my business profitable?***

In this project, I tried to find insights that allow a stakeholder to reduce the risks while decide which is the best location for a restaurant in the city of Buenos Aires, Argentina.

Interest

The analysis is directed to investors who decide to start a business in the gastronomic field as an owner, getting information to minimize risks and get better conditions for their enterprise. Moreover, it can be useful to city administrators, interested in offering better conditions for investments.

Data

According to the definition of the business problem, factors that influence the results are:

- offer of properties in the city of Buenos Aires, Argentina
- structural problems like waterlogging
- number of existing venues in each zone

Analyzing these factors, this study aims to get valuable information about bests candidates among the urban land offer in the city of Buenos Aires, and its advantages for the stakeholders.

Data sources

The data sources used to extract information for the analysis were:

- List of neighborhoods of the city provided as a GEOJSON file, which includes the coordinates for graphical representation. The details of the dataset, according to the City of Buenos Aires Government are:

  ### Neighborhoods of Buenos Aires - main fields

NAME IN GEOJSON FILE	MEANING
BARRIO	Neighborhood Name
COMUNA	Commune Number
PERIMETRO	Neighborhood's Perimetert
AREA	Total Area
GEOMETRY.TYPE	Type of geometry
GEOMETRY.COORDINATES	Geospatial points for area definition

- List of offered urban land in the city of Buenos Aires as a CSV file, which includes features like price, total area, and coordinates for a better location. The main details of the dataset, according to the City of Buenos Aires Government are:

Urban Land Offer in Buenos Aires for 2018 - main fields

NAME IN TABLE	MEANING
LONG	Longitude - Degrees
LAT	Latitude - Degrees
M2TOTAL	Total Area of urban land
PRECIOUSD	Total price in USD

- Data from waterlogging zones as a CSV file, which includes coordinates for the location and grade of criticality. The main details of the dataset, according to the City of Buenos Aires Government are:

Waterlogging Zones in Buenos Aires, 2019

NAME IN TABLE	MEANING
WKT (WELL KNOWN TEXT)	Geometry type and coordinates
CLASIF	criticality level of the zone

- List of nearby venues to each urban land candidate in JSON format, which includes names, category, and location.

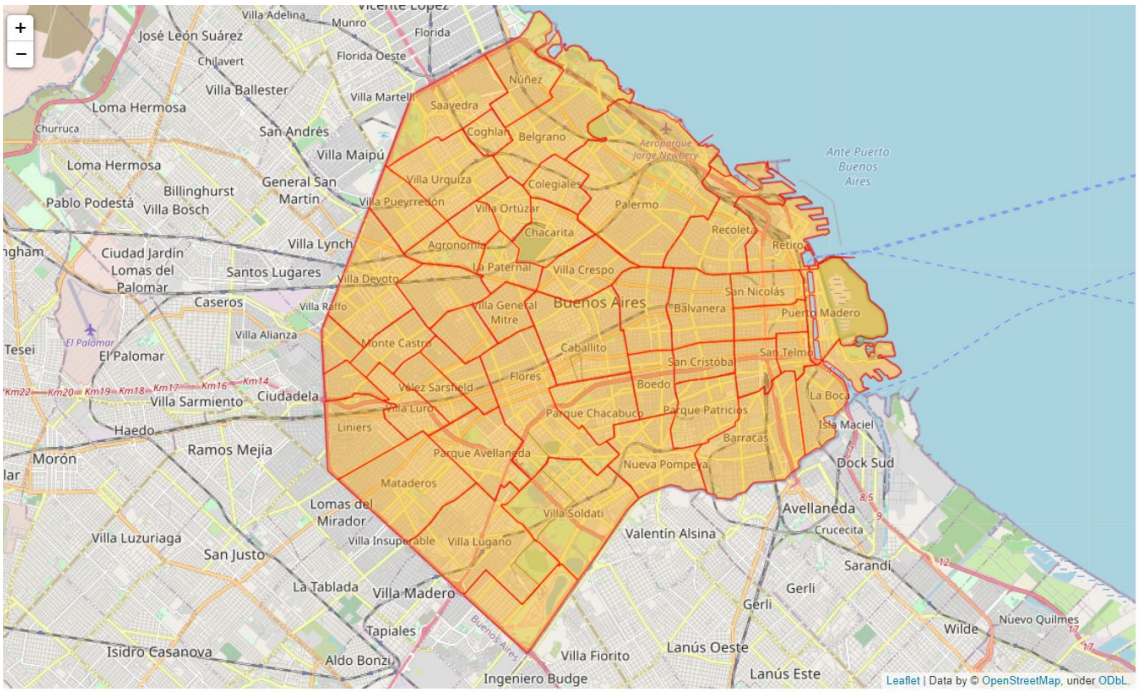
The Government of the city of Buenos Aires provides the datasets related to urban land, waterlogging, and neighborhoods in the city. (https://data.buenosaires.gob.ar/)

The Foursquare API provides the list of venues for the analysis. (https://developer.foursquare.com/docs/places-api/)

Data preparation

I started with a quick view of the city of Buenos Aires and its neighborhoods. I used Nominatim library to find the coordinates of the city and folium library to plot the map. For the data source, I used a GEOJSON file provided by the government of the city of Buenos Aires.

The city of Buenos Aires



The offer of urban land in the city

When an investor is commencing to think about a new business based on the gastronomic field, one of the principal resources to purchase is the property in which perform the desired activity. Since multiple variables affect the election of the appropriate place, like price, location, area, among others, it is necessary an extensive analysis to decide the better option. I continued my study by analyzing the price of urban land in the city. I utilized the CSV file provided by the government of Buenos Aires with information on urban land offer in 2018.

I did an initial exploration of the data and cleaned it, removing columns that weren't considered for the analysis. Another decision was to store the price in USD, for standardization.

After having described the dataframe, I found that there was, at least, one row with a value of zero for the total area ("m2total" column). I calculated that the total of zero values were 11 and replaced them with the relation between price and the average price by square meter.

I founded that the neighborhood NUÑEZ appeared misspelled, and the neighborhood BOCA should be LA BOCA. Both were corrected.

There was a null value in the column "barrio". Since it was the only one, I decided to remove it.

With the data corrected, I renamed the columns for a better understanding.

I assumed that storing the price of each square meter in USD gives a better understanding of the value of that land. So, I calculated this value for each row and added it to the dataframe in a new column named 'square_meter_price".

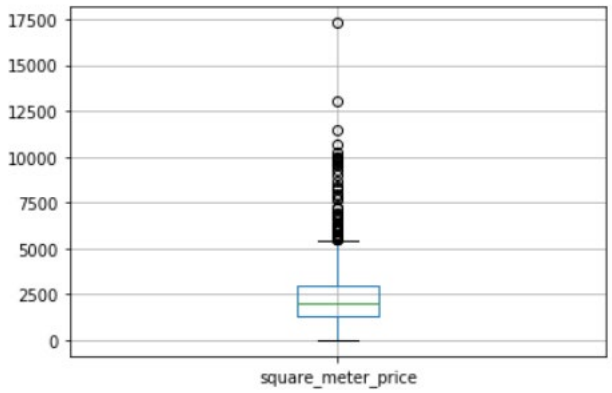
Urban land dataframe

	long	lat	area	price	neighborhood	square_meter_price
0	-58.414071	-34.618214	320.0	330000.0	ALMAGRO	1031.250000
1	-58.424566	-34.610097	320.0	690000.0	ALMAGRO	2156.250000
2	-58.424486	-34.613659	174.0	380000.0	ALMAGRO	2183.908046
3	-58.423651	-34.614752	650.0	1000000.0	ALMAGRO	1538.461538
4	-58.413725	-34.603124	283.0	850000.0	ALMAGRO	3003.533569

With this value in the dataframe, I explored the data to have a better idea of the prices.

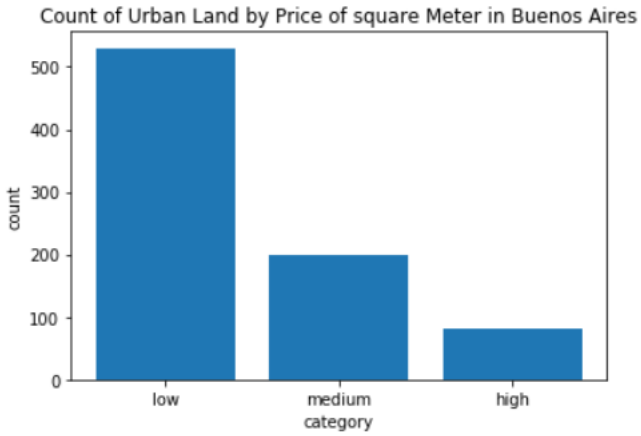
First, I ordered the dataframe by the square meter value, to see in which neighborhoods were the most expensive urban lands located and where were the cheapest ones. Since there was a big difference between the max and min value, I described the data to discover how it was distributed. This showed that 75% of the samples had a value of almost 3 thousand USD per square meter, which was a small value compared with the max. This was confirmed graphically.

Square meter price distribution



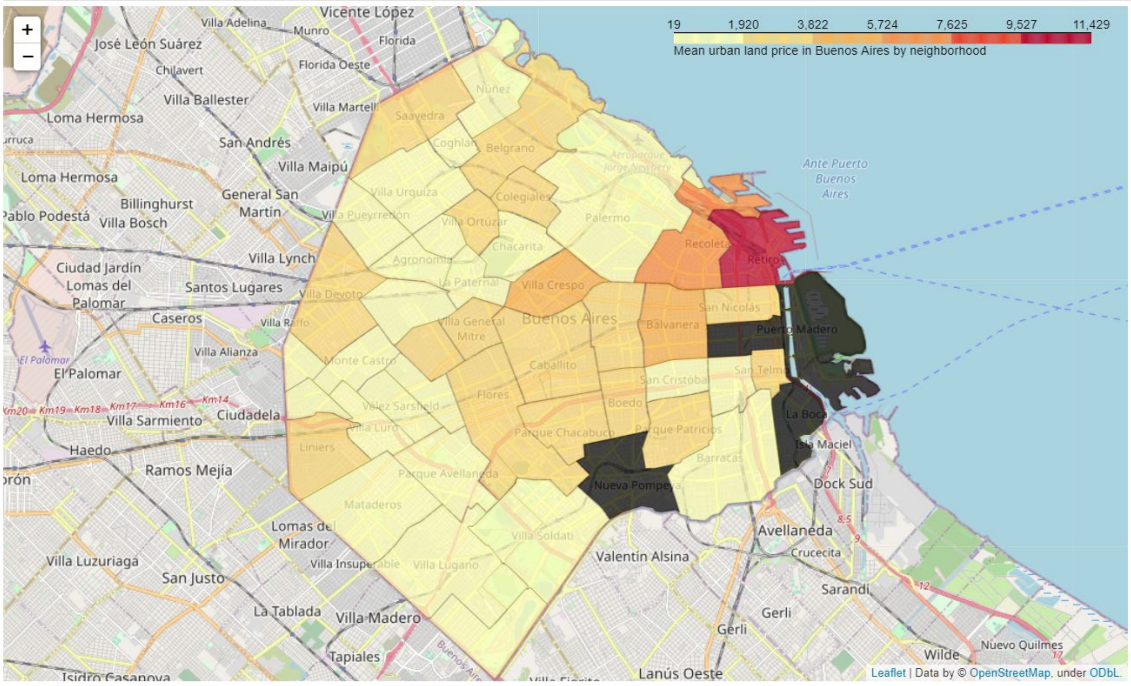
To get more information, I ignored some rows in the max and the min segment. I kept about 90% of the samples just to have a better view of the distribution. I created bins for categories of prices and counted how many lands fitted in each one. There were three bins for low, medium, and high values. Here is the result.

Urban Land by square meter price



I created another map using a choropleth style and marking the neighborhoods categorized by its square meter price of the urban lands.

Neighborhoods of Buenos Aires by square meter price



At this stage, I gained a notion about the value of urban land in the city of Buenos Aires and its distribution. I continued gathering information on other factors, including infrastructures like waterlogging zones and the existence of venues around the lands.

Waterlogging areas in Buenos Aires

To work with the waterlogging zones of the city, I collected the information from a CSV file from the government of the city of Buenos Aires.

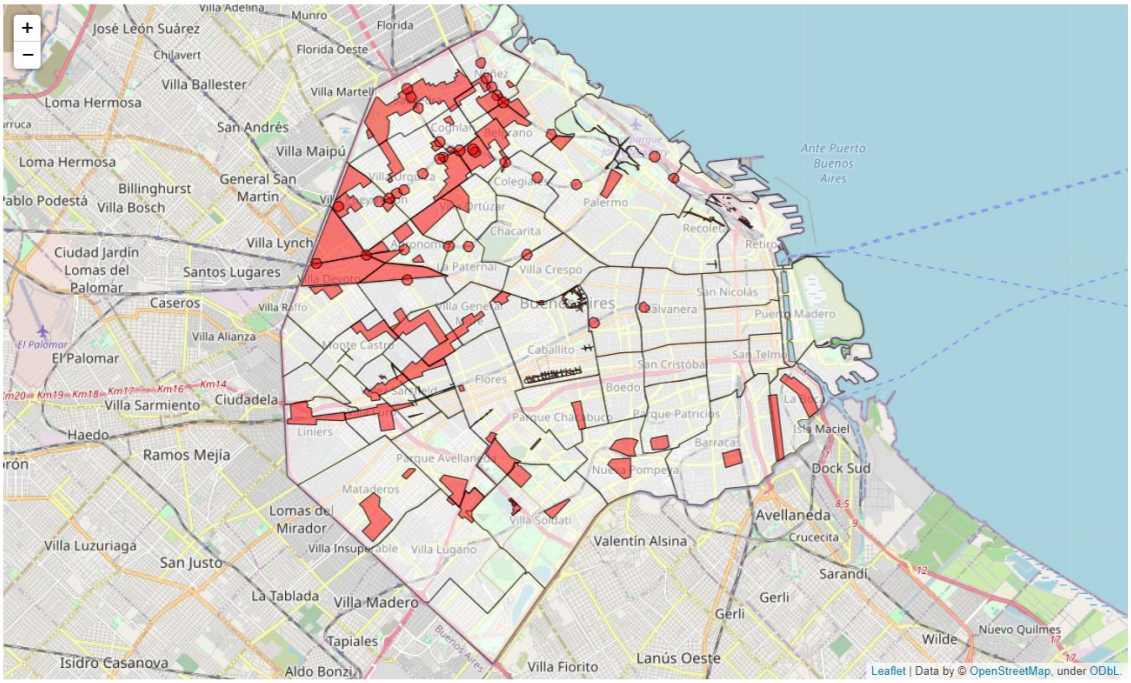
Then, I explored the data and cleaned it, dropping the column Comuna since it wasn't going to be necessary for the analysis

Next, I used the coordinates of each land to determine if it was included in the area of a waterlogging zone. To do this, I worked with Shapely, a library for manipulation and analysis of geometric objects in the Cartesian plane.

I stored in the dataframe the times that each land was included in a waterlogging zone, creating two columns: wl_critic and wl_warning.

Finally, I plotted a map with the waterlogging zones and the urban lands. I used red for the critic zones and yellow for the warning ones. I marked the lands using marker clusters.

Waterlogging zones in Buenos Aires



Foursquare

To obtain information about venues nearby the collected lands, I used the Foursquare API.

I decided to gather information on venues related to the food category, to have an idea of nearby competitors. Similarly, I also collected info related to categories that could be a customer source, as the travel category, like hotels, and to the nightlife category, like discotheques.

Methodology

The purpose of this project is to filter the best candidates among the urban land offer in the city of Buenos Aires, Argentina, considering a series of characteristics, like the value of the land, problems that the urban infrastructure, like waterlogging issues, may cause, and the density of venues around each land. The venues that are considered as important to the project are those categorized by Foursquare as Food, Travel & Transport, and Nightlife venues.

In the Data collection step, I collected the required data for the analysis. It includes information about neighborhoods in Buenos Aires, the offer of urban land, and the location of waterlogging zones identified by the city administration. I also collected the info related to the subcategories of venues that correspond to the main categories mentioned before.

For the analysis step, I worked with this data to filter the urban land candidates, removing those who were in zones identified as waterlogging zones. After that, I explored the nearby area of each one to get information about three different kinds of venues, attempting to recognize opportunities for the business, as the vicinity to hotels or travel venues, which is good, or to other food shops, which is not. I defined **vicinity** as a radius of 250 meters from a land location.

finally, I categorized the candidates in five different clusters, according to the characteristics mentioned before.

The result is valuable information for investors who want to install a restaurant in Buenos Aires, reducing risks related to waterlogging and with the addition of data about business opportunities.

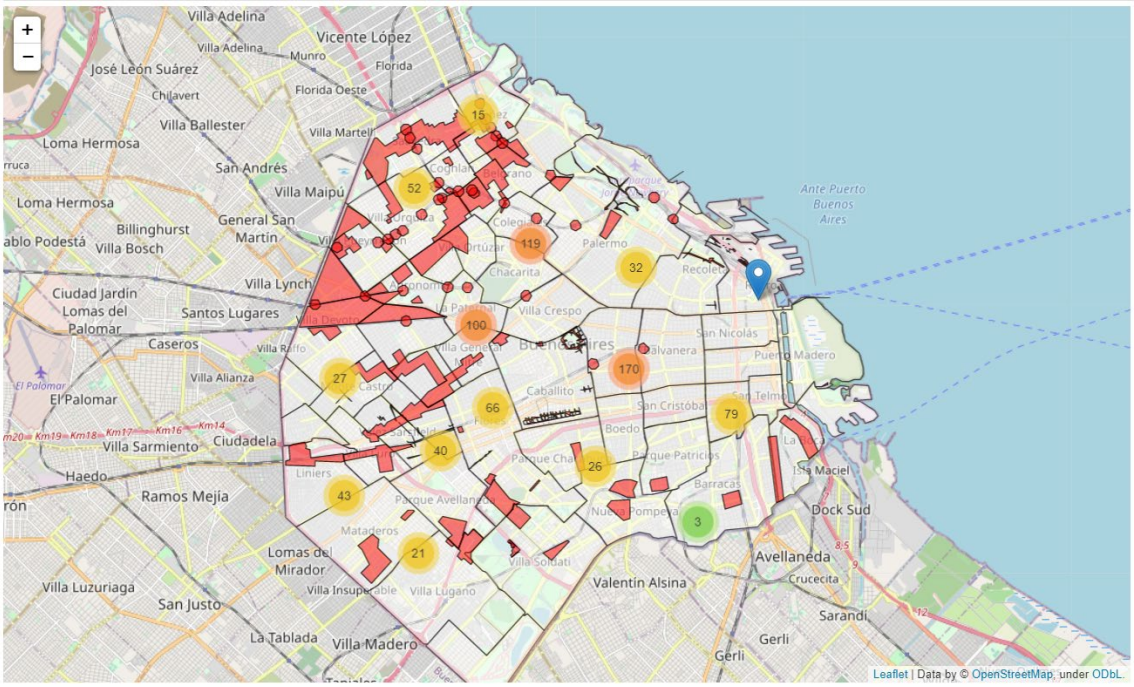
Analysis

The first step of the analysis stage was the filtering of candidates. Those included in a waterlogging zone were excluded. I utilized the coordinates of each land to determine it. To do this, I worked with Shapely, a library for manipulation and analysis of geometric objects in the Cartesian plane.

Later, for each land, I checked if it was in one of the waterlogging zones stored. If it was, then I counted how many times for each type. Finally, I stored those counts in the dataframe using two added columns: wl_critic and wl_warning.

At this point, it was possible to identify the lands located in waterlogging zones. I removed them from the list of candidates. The "good" candidates, those who weren't located in any waterlogging area, were stored in a new dataframe.

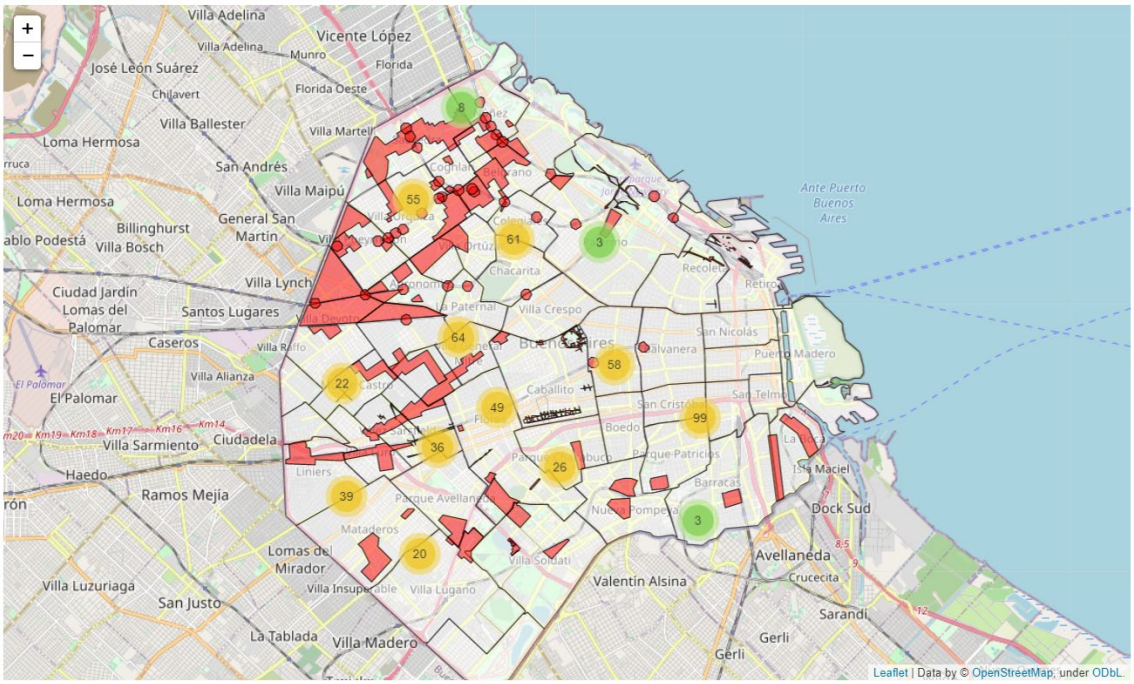
Urban land candidates after waterlogging zone filter



For the remaining candidates, I added another filter: I excluded those with more than four competitors in its vicinity.

To do it, I added a column for each type of venue, counted the venues of the defined categories in the vicinity of each land using the Foursquare API, and stored the result in the corresponding column.

Urban land candidates after >4 food venues filter



The last step was clustering. I used K-means. Before, I standardized the data for the columns of square meter price and the counters of venues.

The results were:

Cluster 0

According to the mean, these candidates have a high value, a considerable count of restaurants, and lower counts for the other categories.

	long	lat	area	price	square_meter_price	restaurant_counts	hotel_counts	other_counts	Clus_Db
count	29.000000	29.000000	29.000000	2.900000e+01	29.000000	29.000000	29.000000	29.000000	29.0
mean	-58.434736	-34.605149	621.379310	1.376828e+06	2163.341923	2.103448	2.517241	0.275862	0.0
std	0.041687	0.024772	586.457794	1.931002e+06	1327.923271	1.205488	0.870988	0.527565	0.0
min	-58.498081	-34.640429	127.000000	1.590000e+05	400.000000	0.000000	2.000000	0.000000	0.0
25%	-58.469616	-34.617798	230.000000	3.400000e+05	1222.493888	1.000000	2.000000	0.000000	0.0
50%	-58.435479	-34.610122	471.000000	8.500000e+05	1771.336554	2.000000	2.000000	0.000000	0.0
75%	-58.404490	-34.584251	755.000000	1.300000e+06	2874.664622	3.000000	3.000000	0.000000	0.0
max	-58.376014	-34.551452	2609.000000	7.800000e+06	4857.142857	4.000000	5.000000	2.000000	0.0

Cluster 1

According to the mean, these candidates have a lower value and lower counts for all the categories.

	long	lat	area	price	square_meter_price	restaurant_counts	hotel_counts	other_counts	Clus_Db
count	208.000000	208.000000	208.000000	2.080000e+02	208.000000	208.000000	208.000000	208.000000	208.0
mean	-58.463923	-34.620108	440.392286	5.474886e+05	1527.735525	0.538462	0.182692	0.105769	1.0
std	0.039303	0.031127	908.396093	7.150288e+05	661.450894	0.499721	0.387346	0.308284	0.0
min	-58.529674	-34.689147	62.000000	1.042200e+04	23.632653	0.000000	0.000000	0.000000	1.0
25%	-58.492480	-34.641427	190.000000	2.600000e+05	1006.792331	0.000000	0.000000	0.000000	1.0
50%	-58.473746	-34.627061	285.000000	3.800000e+05	1456.755162	1.000000	0.000000	0.000000	1.0
75%	-58.443889	-34.602829	390.000000	5.625000e+05	1934.991131	1.000000	0.000000	0.000000	1.0
max	-58.361857	-34.543445	9969.000000	7.500000e+06	3205.574913	1.000000	1.000000	1.000000	1.0

Cluster 2

According to the mean, these candidates have lower value but a considerable count of restaurants and hotels. The count of nightlife venues is lower, too.

	long	lat	area	price	square_meter_price	restaurant_counts	hotel_counts	other_counts	Clus_Db
count	18.000000	18.000000	18.000000	1.800000e+01	18.000000	18.000000	18.000000	18.000000	18.0
mean	-58.442552	-34.604749	369.607032	8.411111e+05	2876.908611	3.055556	0.333333	2.833333	2.0
std	0.030022	0.026045	316.838110	5.993490e+05	2126.311893	0.998365	0.594089	1.150447	0.0
min	-58.512162	-34.657823	83.000000	2.500000e+05	662.500000	1.000000	0.000000	2.000000	2.0
25%	-58.450260	-34.622985	174.250000	3.112500e+05	1540.555253	2.000000	0.000000	2.000000	2.0
50%	-58.436621	-34.598329	319.500000	6.550000e+05	2088.659208	3.000000	0.000000	2.000000	2.0
75%	-58.425241	-34.586078	391.500000	1.300000e+06	3448.296497	4.000000	0.750000	3.750000	2.0
max	-58.397531	-34.570557	1200.000000	2.400000e+06	9259.259259	4.000000	2.000000	6.000000	2.0

Cluster 3

According to the mean, these candidates have lower value but a considerable count of restaurants. Also, have zero *travel* venues and a lower count of nightlife venues.

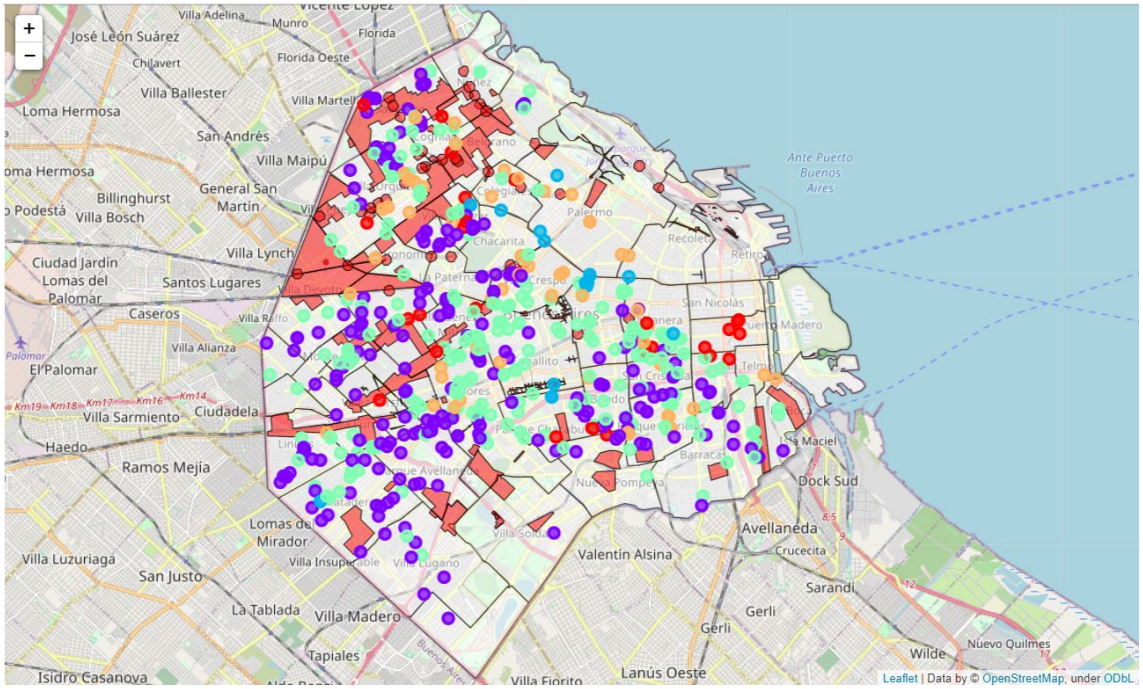
	long	lat	area	price	square_meter_price	restaurant_counts	hotel_counts	other_counts	Clus_Db
count	230.000000	230.000000	230.000000	2.300000e+02	230.000000	230.000000	230.000000	230.000000	230.0
mean	-58.450322	-34.614832	469.895495	6.843039e+05	1723.619470	2.778261	0.195652	0.108696	3.0
std	0.039067	0.026061	509.434973	5.543881e+05	670.268787	0.775417	0.397567	0.311936	0.0
min	-58.528688	-34.672245	86.000000	1.300000e+05	350.000000	2.000000	0.000000	0.000000	3.0
25%	-58.479987	-34.633030	224.000000	3.592500e+05	1223.722842	2.000000	0.000000	0.000000	3.0
50%	-58.455817	-34.617869	344.000000	5.300000e+05	1686.046512	3.000000	0.000000	0.000000	3.0
75%	-58.415021	-34.604626	484.000000	8.000000e+05	2242.692308	3.000000	0.000000	0.000000	3.0
max	-58.362995	-34.542595	4000.000000	5.500000e+06	3106.796117	4.000000	1.000000	1.000000	3.0

Cluster 4

According to the mean, these candidates have high value and a considerable count of restaurants and nightlife venues. The count of travel venues is low.

	long	lat	area	price	square_meter_price	restaurant_counts	hotel_counts	other_counts	Clus_Db
count	58.000000	58.000000	58.000000	5.800000e+01	58.000000	58.000000	58.000000	58.000000	58.0
mean	-58.445423	-34.599804	369.189655	1.713481e+06	4613.270205	2.500000	0.344828	0.086207	4.0
std	0.035074	0.026740	330.749519	1.707705e+06	1101.051535	1.013072	0.479463	0.283121	0.0
min	-58.509985	-34.639637	75.000000	3.100000e+05	3130.081301	1.000000	0.000000	0.000000	4.0
25%	-58.474872	-34.624172	179.750000	7.775000e+05	3718.054728	2.000000	0.000000	0.000000	4.0
50%	-58.447759	-34.597077	253.000000	1.125000e+06	4457.516340	2.500000	0.000000	0.000000	4.0
75%	-58.412631	-34.576762	420.000000	1.712500e+06	4940.497271	3.000000	1.000000	0.000000	4.0
max	-58.364223	-34.544165	1540.000000	7.500000e+06	7684.210526	4.000000	1.000000	1.000000	4.0

Candidates after clustering



Results

The analysis stage started with the examination of the offer of urban land in the city. I've found that the value of the land isn't evenly distributed and most of the offer can be categorized in a low-price class. This was represented in a graphical way using a box plot and a bar chart. Concerning the location, I've concluded that generally speaking, lands are well distributed over the city, except for a couple of neighborhoods and the west limit.

The next step was to identify lands located in high-risk zones, because of the existence of waterlogging problems in the city. Using geometry concepts applied to geospatial variables, it was possible to determine if a candidate was located in a waterlogging zone. Naturally, these candidates were excluded from the list of "good" ones to reduce those risks.

After that, I realized the exploration of the vicinity of each candidate, defined as a surrounding area in a radius of 250 meters, searching for venues related to food, travel, and nightlife. This search allowed me to count how many venues of each type surround the candidates. That information is used for filtering those with a considerable number of possible competitors. This info is also considered to detect good conditions for the business. For example, the closeness to a hotel could provide a substantial volume of customers.

To conclude, the remaining candidates were clustered if five categories according to the value of the land and the counts of different types of nearby venues. This categorization gives investors insights about business opportunities in the zone to take advantage of the proximity of hotels, dance clubs, pubs, or another possible source of customers, avoiding concentrations of competitors.

It's important to mention that this study only included three variables to construct lists to help stakeholders in their election of a candidate. However, many variables can be studied to identify the best option.

Conclusion

The objective of this analysis wasn't to define which candidates are the best but generate lists of similar ones that can fit basic requirements previously defined by stakeholders. These lists have fewer options than the original, its content has specific characteristics, which is useful to stakeholders, and doesn't include lands considered disadvantageous because of location inside waterlogging zones.

Considering that there is a universe of variables that can influence the election of urban land for restaurants, like the attractiveness of each location, security, social and economic dynamics, the volume of tourists, etc., there is plenty of room for improvement.