



Starting a pharmacy business in Buenos Aires

Jose Ignacio Chajchir

April, 2021

Introduction

Background

Buenos Aires (CABA) is the capital city of Argentina.

Large cities have a lot of diverse Neighborhoods and CABA is not an exception. If someone is interested in starting a business, location will be one of the most important factors for that business to be profitable.

Problem

The goal of this project is to, based on data, identify which are the most suitable neighborhoods in CABA to start a pharmacy business.

Target Audience

- Individual investors, especially those with background in the pharmacy business.
- Pharmacy chains interested in opening a new store

It should be useful for any organization or Specialists who perform demographic analysis.

Data acquisition and cleaning

Data targeting

- a. How many pharmacies already exist in the neighborhood?
 - *If there are too many persons per pharmacies in the neighborhood, a new pharmacy will probably have clients.*
- b. How many persons live? How old are them?
 - *If there are too many persons per pharmacies in the neighborhood, a new pharmacy will probably have clients.*
- c. Are these persons consumers of pharmacy products?
 - *If people living in the neighborhood have a low income, they will probably avoid spending money in esthetic or cosmetic products.*

Data acquisition and cleaning

How many pharmacies already exist in the neighborhood?

Foursquare API has a limit of 50 results per request. Therefore, data for the complete CABA area had to be collected using multiple requests. A mesh grid was created:

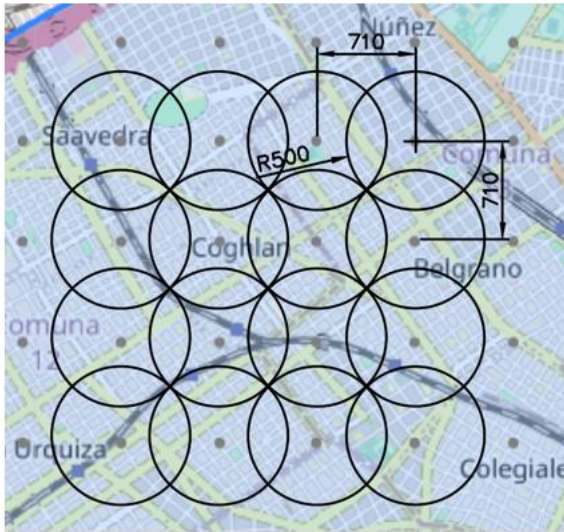


Figure 1: Shows chosen parameters for grid definition. With a radius of 500 meters and Longitude distance = Latitude distance = 710 meters, the queries should find all the pharmacies in the target area. A radius of 500 meters doesn't exceed the 50 results per query limit.

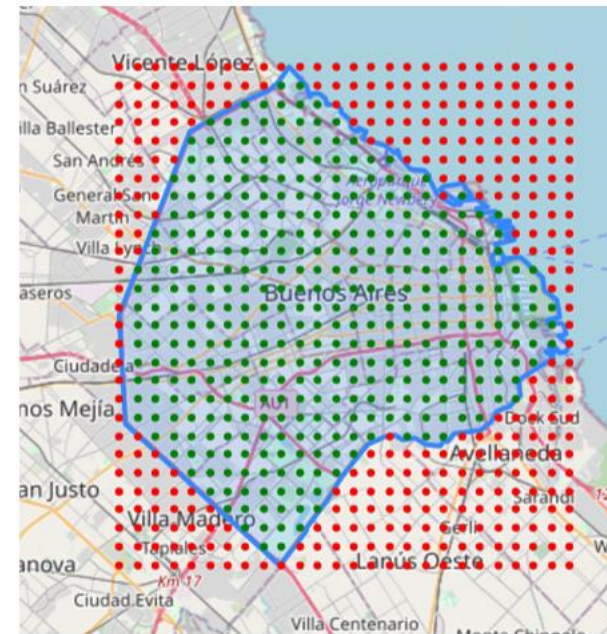


Figure 3: Grid of points classified (Green: inside CABA polygon – Red: outside CABA polygon)

Data acquisition and cleaning

How many pharmacies already exist in the neighborhood?

Neighborhood's data (geojson file) was collected, and the neighborhood of each point was obtained. Finally, the counts of pharmacy per Neighborhood was acquired



Figure 5: CABA neighborhoods geojson file.

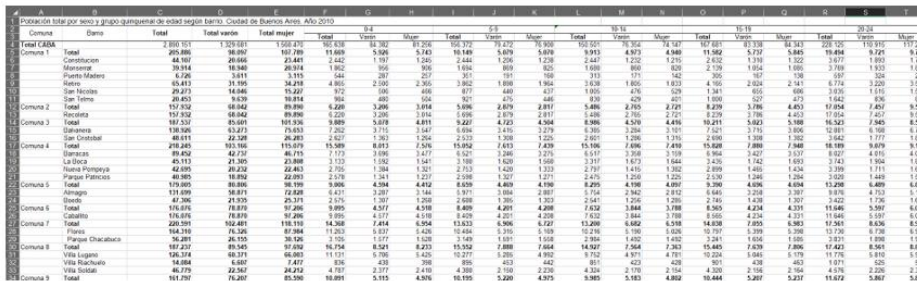
	Neighborhood	counts
0	ALMAGRO	26
1	BALVANERA	52
2	BARRACAS	16
3	BELGRANO	43
4	BOCA	4
5	BOEDO	6
6	CABALLITO	42

Figure 6: A total of 45 unique neighborhoods were listed in this dataframe

Data acquisition and cleaning

How many persons live? How old are them?

Demographic data of CABA neighborhoods was found in governmental websites



The image shows a screenshot of an Excel spreadsheet with multiple columns and rows. The columns are labeled with letters A through T, and the rows are labeled with numbers 1 through 14. The data represents population statistics for various neighborhoods in CABA, categorized by sex and age group.

Figure 7: Excel file showing CABA population by sex, age group and neighborhood.

	Neighborhood	Total	Total_more_65
0	CONSTITUCION	44107.0	6515.0
1	MONSERRAT	39914.0	5868.0
2	PUERTO MADERO	6726.0	490.0
3	RETIRO	65413.0	8336.0
4	SAN NICOLAS	25273.0	4325.0
5	SAN TELMO	20453.0	3590.0
6	RECOLETA	157932.0	31265.0
7	BALVANERA	138926.0	22096.0
8	SAN CRISTOBAL	48611.0	7932.0
9	BARRACAS	89452.0	9724.0
10	LA BOCA	45113.0	5661.0
11	NUEVA POMPEYA	42695.0	6617.0
12	PARQUE PATRICIOS	40885.0	6118.0
13	ALMAGRO	131699.0	23199.0
14	BOEDO	47306.0	7601.0

Figure 8: "Barrio" means Neighborhood in Spanish. "Total" represents the total population and "Total_more_65" represents the population aged more than 65.

Are these persons consumers of pharmacy products?

Real state information of price per square meter of each neighborhood was obtained. For the matter of this project, income and price per square meter were considered correlated.

zonapropnoticias

Home Mercado Inmobiliario Zonaprop Data Date un respiro

	BARRIO	ESTRENAR	POZO	INDEX	USADO
1.	Puerto Madero	6,202	7,059	5,734	5,537
2.	Palermo	3,398	3,085	3,242	3,165
3.	Belgrano	3,466	3,203	3,099	2,956
4.	Nuñez	3,093	2,846	2,989	2,934
5.	Recoleta	3,139	2,923	2,912	2,858
6.	Retiro	3,201	4,098	2,882	2,834
7.	Colegiales	3,131	2,974	2,825	2,719

	Neighborhood	USD/m2
0	Puerto Madero	5786
1	Palermo	3313
2	Belgrano	3164
3	Nuñez	3039
4	Recoleta	2973
5	Retiro	2926
6	Colegiales	2888
7	Villa Urquiza	2801
8	Coghlan	2690
9	Chacarita	2643

Figure 9

Data preparation

Dataframes obtained in the previous section were merged and modified so as to get the appropriate data needed to feed a model intended to answer the main question of this project.

After performing cleaning operations and adding calculated columns, the following table was obtained:

	Neighborhood	Total_pop	Total_pop_+65	Pharmacies	USD/m2	pop_per_pharma	+65pop_per_pharma
0	CONSTITUCION	44107.00	6515.00	6.00	1960	7351.17	1085.83
1	MONSERRAT	39914.00	5868.00	30.00	2083	1330.47	195.60
2	PUERTO MADERO	6726.00	490.00	4.00	5786	1681.50	122.50
3	RETIRO	65413.00	8336.00	22.00	2926	2973.32	378.91
4	SAN NICOLAS	29273.00	4325.00	48.00	2163	609.85	90.10
5	SAN TELMO	20453.00	3590.00	8.00	2417	2556.62	448.75
6	RECOLETA	157932.00	31265.00	88.00	2973	1794.68	355.28
7	BALVANERA	138926.00	22096.00	52.00	2043	2671.65	424.92
8	SAN CRISTOBAL	48611.00	7932.00	6.00	2015	8101.83	1322.00
9	BARRACAS	89452.00	9724.00	16.00	2364	5590.75	607.75
10	LA BOCA	45113.00	5661.00	4.00	1781	11278.25	1415.25
11	NUEVA POMPEYA	42695.00	6617.00	5.00	1872	8539.00	1323.40
12	PARQUE PATRICIOS	40985.00	6118.00	7.00	2022	5855.00	874.00

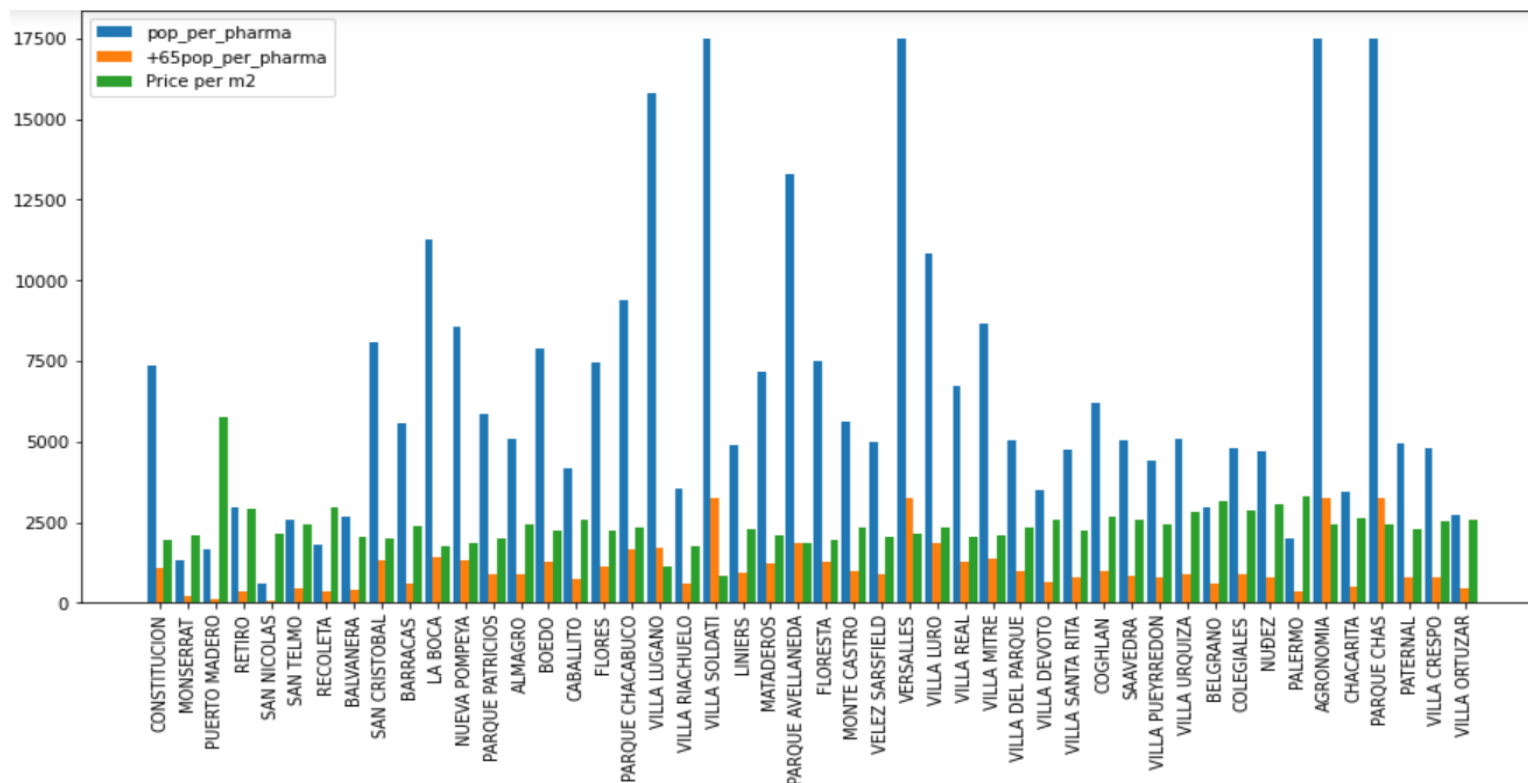
Figure 10

Methodology

Exploratory analysis

The 3 main features were plotted in a Bar chart. Observations:

- plenty of variation over the 46 neighborhoods
- grouping the neighborhoods would be the best approach to build the outcome of this project



Methodology

Model

As my goal was to identify groups of neighborhoods based on their similarity, I decided to use **K-means**. Data was sliced it into 2 different datasets and run separate K-means clustering algorithms. The purpose of this division was to evaluate the results considering two groups of residents: total residents and residents age 65-plus.

After running both algorithms, the results were procured:

	Neighborhood	Total_pop	Total_pop_+65	Pharmacies	USD/m2	pop_per_pharma	+65pop_per_pharma	cluster_A	cluster_B
0	CONSTITUCION	44107.0	6515.0	7.0	1960	6301.000000	930.714286	0	3
1	MONSERRAT	39914.0	5868.0	30.0	2083	1330.466667	195.600000	2	3
2	PUERTO MADERO	6726.0	490.0	4.0	5786	1681.500000	122.500000	3	2
3	RETIRO	65413.0	8336.0	22.0	2926	2973.318182	378.909091	1	0
4	SAN NICOLAS	29273.0	4325.0	48.0	2163	609.854167	90.104167	2	3

Figure 12

Results

Two pivot charts were created to understand the properties of the clusters generated from dataset A and dataset B.

Dataset A Pivot Chart:

	pop_per_pharma					USD/m2				
	my25	median	my75	mean	std	my25	median	my75	mean	std
cluster_A										
0	7391.366162	7993.083333	8853.729167	8320.793771	1537.371980	1952.25	2075.0	2229.75	2079.583333	178.458275
1	2949.558140	3501.105263	4777.363636	3802.563851	1318.784161	2591.00	2801.0	2973.00	2829.615385	243.788343
2	3308.663462	4862.336601	5034.950175	4129.154186	1564.791434	2074.75	2282.0	2376.50	2235.375000	202.111809
3	1681.500000	1681.500000	1681.500000	1681.500000	NaN	5786.00	5786.0	5786.00	5786.000000	NaN
4	16219.812500	17489.000000	17489.000000	16510.000000	1708.806433	1304.25	2000.5	2354.75	1802.166667	677.952629

Figure 13: pivot chart results dataset A clustering

Dataset B Pivot Chart:

	+65pop_per_pharma					USD/m2				
	my25	median	my75	mean	std	my25	median	my75	mean	std
cluster_B										
0	454.281250	713.090226	836.709091	661.099453	222.061422	2588.75	2745.5	2961.25	2808.000000	247.794580
1	1267.333333	1339.325000	1663.000000	1455.698611	233.443425	1864.50	2040.5	2146.50	1977.500000	324.094964
2	122.500000	122.500000	122.500000	122.500000	NaN	5786.00	5786.0	5786.00	5786.000000	NaN
3	582.500000	799.000000	916.000000	718.061710	288.909237	2050.00	2261.0	2357.00	2201.823529	191.114898
4	3277.000000	3277.000000	3277.000000	3277.000000	0.000000	1828.25	2289.5	2422.75	1961.500000	760.791036

Figure 13: pivot chart results dataset B clustering

Results

Dataset A Clusters:

Cluster A 0:

- Medium rate of persons per pharmacy
- Low income
- Cluster_A_0 neighborhoods: ['CONSTITUCION', 'SAN CRISTOBAL', 'LA BOCA', 'NUEVA POMPEYA', 'BOEDO', 'FLORES', 'PARQUE CHACABUCO', 'MATADEROS', 'FLORESTA', 'VILLA LURO', 'VILLA REAL', 'VILLA MITRE']

Cluster A 1:

- Low rate of persons per pharmacy
- Medium income
- Cluster_A_1 neighborhoods: ['RETIRO', 'RECOLETA', 'CABALLITO', 'VILLA DEVOTO', 'COGHLAN', 'SAAVEDRA', 'VILLA URQUIZA', 'BELGRANO', 'COLEGIALES', 'NUÑEZ', 'PALERMO', 'CHACARITA', 'VILLA ORTUZAR']

Cluster A 2:

- Low rate of persons per pharmacy
- Medium income
- Cluster_A_2 neighborhoods: ['MONSERRAT', 'SAN NICOLAS', 'SAN TELMO', 'BALVANERA', 'BARRACAS', 'PARQUE PATRICIOS', 'ALMAGRO', 'VILLA RIACHUELO', 'LINIERS', 'MONTE CASTRO', 'VELEZ SANSFIELD', 'VILLA DEL PARQUE', 'VILLA SANTA RITA', 'VILLA PUEYRREDON', 'PATERNAL', 'VILLA CRESPO']

Cluster A 3:

- Very low rate of persons per pharmacy
- Very high income
- Cluster_A_3 neighborhoods: ['PUERTO MADERO']

Cluster A 4:

- Very high rate of persons per pharmacy
- Low income
- Cluster_A_4 neighborhoods: ['VILLA LUGANO', 'VILLA SOLDATI', 'PARQUE AVELLANEDA', 'VERSALLES', 'AGRONOMIA', 'PARQUE CHAS']

Results

Dataset B Clusters:

Cluster B 0:

- Very low rate of persons per pharmacy
- Medium income
- Cluster_B_0 neighborhoods: ['RETIRO', 'RECOLETA', 'CABALLITO', 'VILLA DEVOTO', 'COGHLAN', 'SAAVEDRA', 'VILLA URQUIZA', 'BELGRANO', 'COLEGIALES', 'NUÑEZ', 'PALERMO', 'CHACARITA', 'VILLA CRESPO', 'VILLA ORTUZAR']

Cluster B 1:

- Medium rate of persons per pharmacy
- Low income
- Cluster_B_1 neighborhoods: ['SAN CRISTOBAL', 'LA BOCA', 'NUEVA POMPEYA', 'BOEDO', 'PARQUE CHACABUCO', 'VILLA LUGANO', 'MATADEROS', 'PARQUE AVELLANEDA', 'FLORESTA', 'VILLA LURO', 'VILLA REAL', 'VILLA MITRE']

Cluster B 2:

- Low rate of persons per pharmacy
- Very high income
- Cluster_B_2 neighborhoods: ['PUERTO MADERO']

Cluster B 3:

- Low rate of persons per pharmacy
- Medium income
- Cluster_B_3 neighborhoods: ['CONSTITUCION', 'MONSERRAT', 'SAN NICOLAS', 'SAN TELMO', 'BALVANERA', 'BARRACAS', 'PARQUE PATRICIOS', 'ALMAGRO', 'FLORES', 'VILLA RIACHUELO', 'LINIERS', 'MONTE CASTRO', 'VELEZ SARSFIELD', 'VILLA DEL PARQUE', 'VILLA SANTA RITA', 'VILLA PUEYRREDON', 'PATERNAL']

Cluster B 4:

- Very high rate of persons per pharmacy
- Low income
- Cluster_B_4 neighborhoods: ['VILLA SOLDATI', 'VERSALLES', 'AGRONOMIA', 'PARQUE CHAS']

Discussion

Clusters_A and Clusters_B show a clear inverse correlation between persons per pharmacy rate and income. But which of them is more important to select the right cluster?

With our current results, **I would recommend clusters Cluster_A_4 and Cluster_B_4 for starting a pharmacy.** Although they have the lowest value for the income feature (represented as USD/m²) they are not far away from the average of the cluster values.

Conclusion

Final recommendation

Final recommendation of this study:

- Highly recommended neighborhoods: ['VILLA SOLDATI', 'VERSALLES', 'AGRONOMIA', 'PARQUE CHAS']
- Alternative recommended neighborhoods: ['VILLA LUGANO', 'PARQUE AVELLANEDA'].



Figure 14: Recommended neighborhoods

Conclusion

Future enhancements

This model could be improved by boosting accuracy of existing data and getting additional information to generate new features.

Boosting accuracy:

- Using real Income data instead of price per square meters should be considered. Probably this kind of data could be provided by the Government of CABA city.

Getting additional data:

- Data of nearby hospitals and medical centers.
- Data of nearby commercial areas/shopping malls.

Conclusion

Q & A section

