Starting a pharmacy business in Buenos Aires

José Ignacio Chajchir

April 26, 2021

Contents

[1. Introduction 1](#_Toc70357688)

[1.1 Background 1](#_Toc70357689)

[1.2 Problem 2](#_Toc70357690)

[1.3 Target Audience 2](#_Toc70357691)

[2. Data acquisition and cleaning 2](#_Toc70357692)

[2.1 Data targeting 2](#_Toc70357693)

[2.2 Data collection and cleaning 2](#_Toc70357694)

[2.2.1 How many pharmacies already exists in the neighborhood? 2](#_Toc70357695)

[2.2.2 How many persons live in the neighborhood? How old are them? 5](#_Toc70357696)

[2.2.3 Are these persons consumers of pharmacy products? 6](#_Toc70357697)

[2.3 Data sources 7](#_Toc70357698)

[2.4 Data preparation 7](#_Toc70357699)

# Introduction

## Background

Buenos Aires is the capital city of Argentina, the country where I was born and where I live. As the term “Buenos Aires” could represent different areas, the term “CABA” (stands for Ciudad Autonoma de Buenos Aires) is more accurate to differentiate the city from other places.

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

This report shall be of great interest for:

* 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

To build a useful and neat dataframe that could lead into meaningful results, I thought about which features would be useful by asking some questions for each neighborhood.

1. How many pharmacies already exist in the neighborhood?

The rate of persons per existing pharmacies should be a very important input parameter for the decision. If there are too many persons per pharmacies in the neighborhood, a new pharmacy will probably have clients.

1. How many persons live? How old are them?

As mentioned in the point before, the rate of persons per existing pharmacies is very useful information, so these questions should be answered in order to calculate the rate. The age of the persons could be also valuable as older people tend to buy more medication.

1. 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 collection and cleaning

### How many pharmacies already exists in the neighborhood?

For question “How many pharmacies already exists in the neighborhood?”, I used data from the search venue foursquare API. Although the API let you search only for pharmacies (by setting category ID) there´s a limit of 50 results per request. Therefore, data for the complete CABA area had to be collected using multiple requests, where each request belonged to a unique geographic point inside CABA area.

A grid of points was defined with certain parameters:

* The grid must represent CABA area. This area was obtained from a geojson file that contains Argentinian provinces.
* The distance within points must be defined based on:
  + Foursquare API radius parameter
  + 50 results per request limit

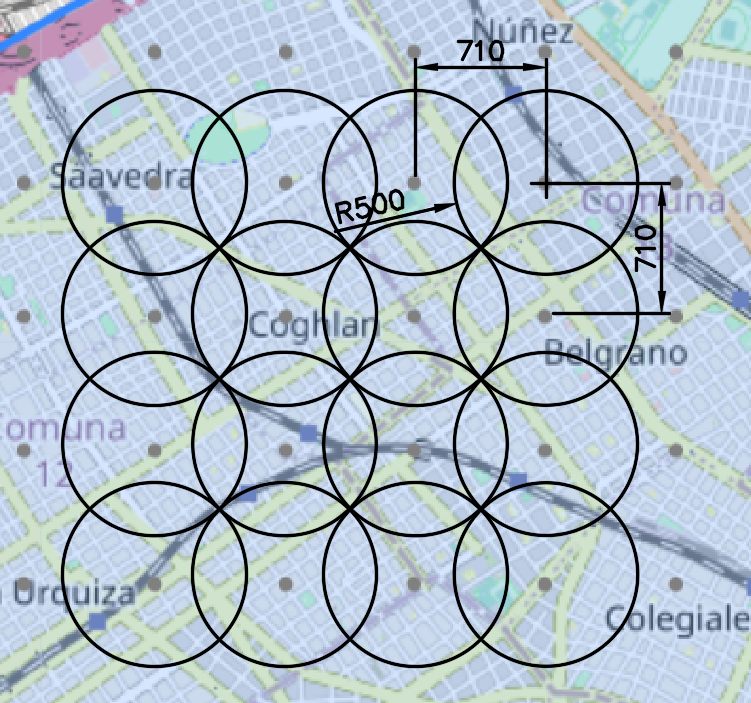


Figure 1: Shows chosen parameters for grid definition. With a radius of 500 meters and Longitude distance = Latitude distance = 710meters, the queries should find all the pharmacies in the target area. A radius of 500 meters doesn´t exceed the 50 results per query limit.

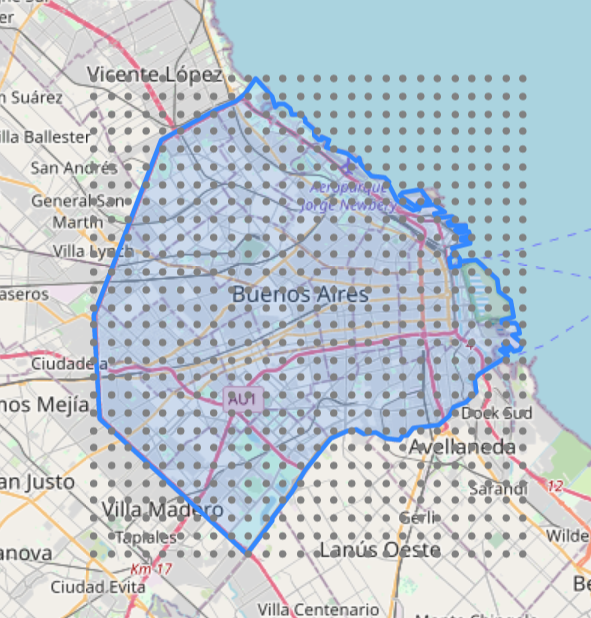


Figure 2: Grid of points and CABA polygon.

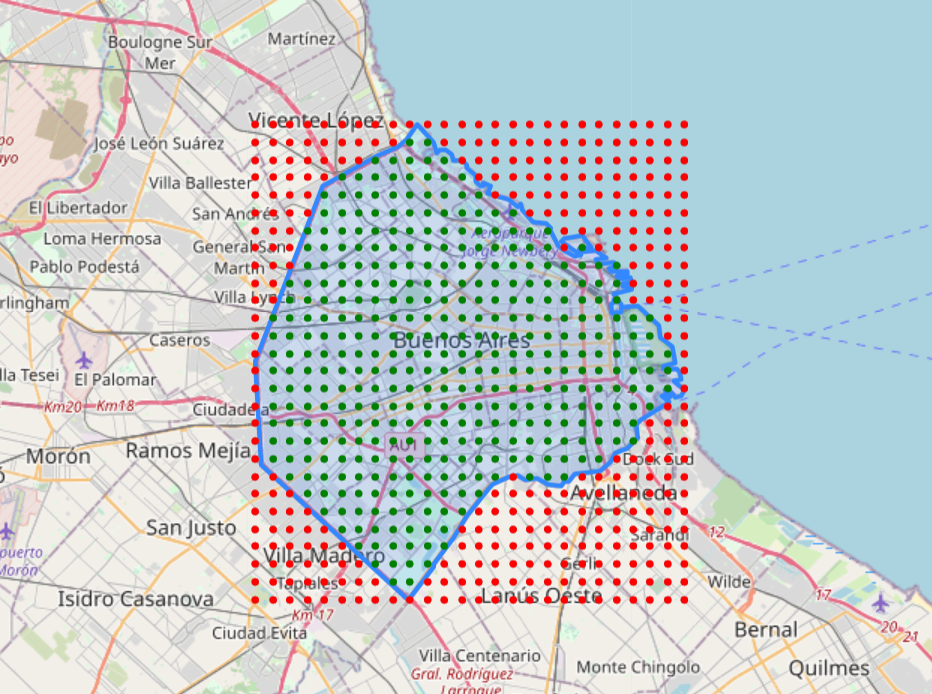


Figure 3: Grid of points classified (Green: inside CABA polygon – Red: outside CABA polygon)

Once the points inside CABA were defined, the foursquare API was tested with a few points to get familiar with the requested data.

I then created a function to obtain the data for every point inside CABA and compile the information into one single dataframe. After dropping repeated pharmacies (as showed in figure 1, intersection between circular areas were scanned twice), the dataframe looked like this:



Figure 4: Dataframe of unique pharmacies located in CABA area (total of 752 pharmacies).

Finally, I needed to find out in which neighborhood each pharmacy was located. Therefore, I decided to use a geojson file that contains CABA neighborhoods limits.

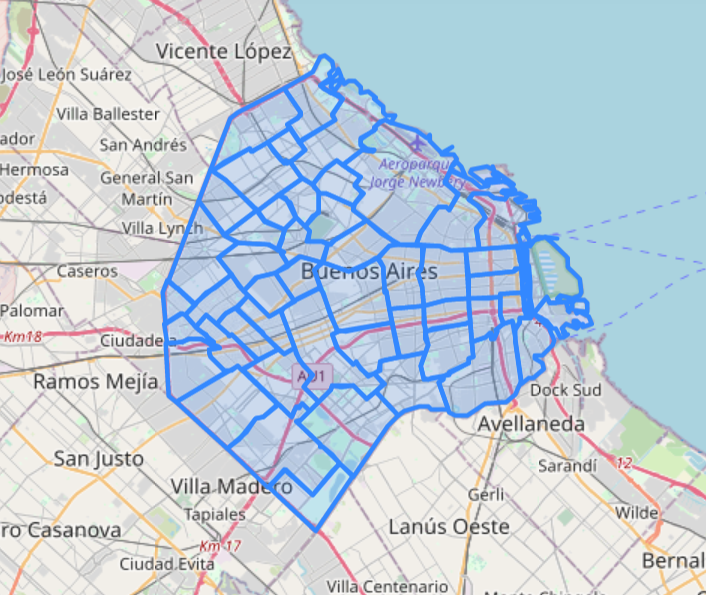


Figure 5: CABA neighborhoods geojson file.

After running a for loop, the mentioned dataframe was completed with the “Neighborhood” column, and using groupby function I finally got the answer of “How many pharmacies already exists in the neighborhood?”

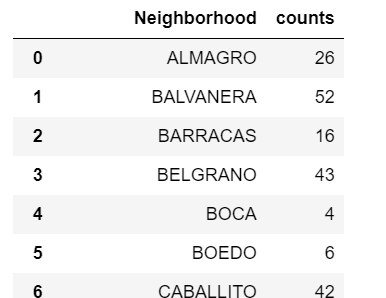


Figure 6: A total of 45 unique neighborhoods were listed in this dataframe

### How many persons live in the neighborhood? How old are them?

I explored several governmental websites to discover the answers. I finally found demographic statistical data of CABA with the desired information.

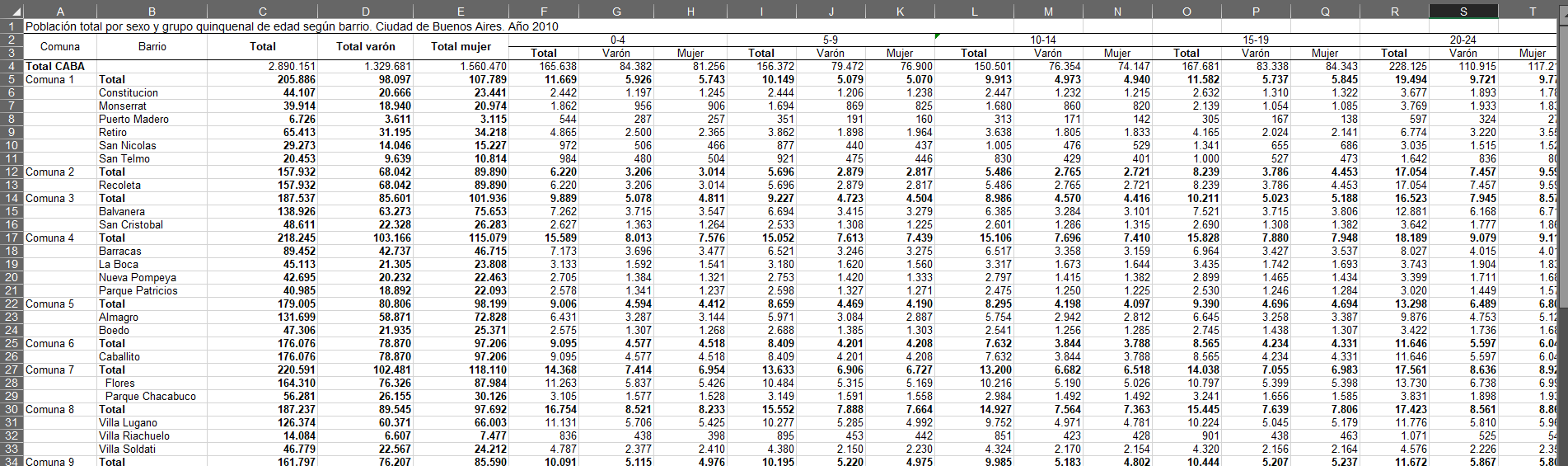


Figure 7: Excel file showing CABA population by sex, age group and neighborhood.

I then converted this excel file into a dataframe and performed cleaning operations. I also added a calculated column to get population above 65 years old. The desired dataframe was obtained:



Figure 8: “Barrio” means Neigborhood 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?

I searched online data of population income per neighborhood but unfortunately, I couldn´t find well-ordered and reliable information. What I did find instead, is real state information of price per square meter of each neighborhood. For the matter of this project, income and price per square meter were considered correlated.



Figure 9

## Data sources

Multiple datasets and geodata where needed to perform section 2.2:

* How many pharmacies already exists in the neighborhood? (2.2.1)
  + Zip file containing CABA shape file: <https://infra.datos.gob.ar/catalog/modernizacion/dataset/7/distribution/7.34/download/provincias.zip>
  + Existing Pharmacies in CABA: Foursquare API
  + CABA neighborhoods geojson file: <https://cdn.buenosaires.gob.ar/datosabiertos/datasets/barrios/barrios.geojson>
* How many persons live? How old are them? (2.2.2)
  + Statistics about neighborhood population:

<https://www.estadisticaciudad.gob.ar/eyc/?p=28008/PB_barrio_ARIP_CNP2010.xls>

* Are these persons consumers of pharmacy products? (2.2.3)
  + Neighborhoods Price per Square meter:

https://www.zonaprop.com.ar/noticias/zpindex/

## Data preparation

In this section I will describe how dataframes obtained in section 2.2 where merged and modified so as to get the appropriate data needed to feed a model intended to answer the main question of this project.

At first, output data of section 2.2.2 was merged with output data of section 2.2.1. Since they had 48 and 45 neighborhoods respectively, a left join merge was performed.

Secondly, the obtained dataframe was merged with output data from section 2.2.3. After performing cleaning operations and adding calculated columns, the following table was obtained:

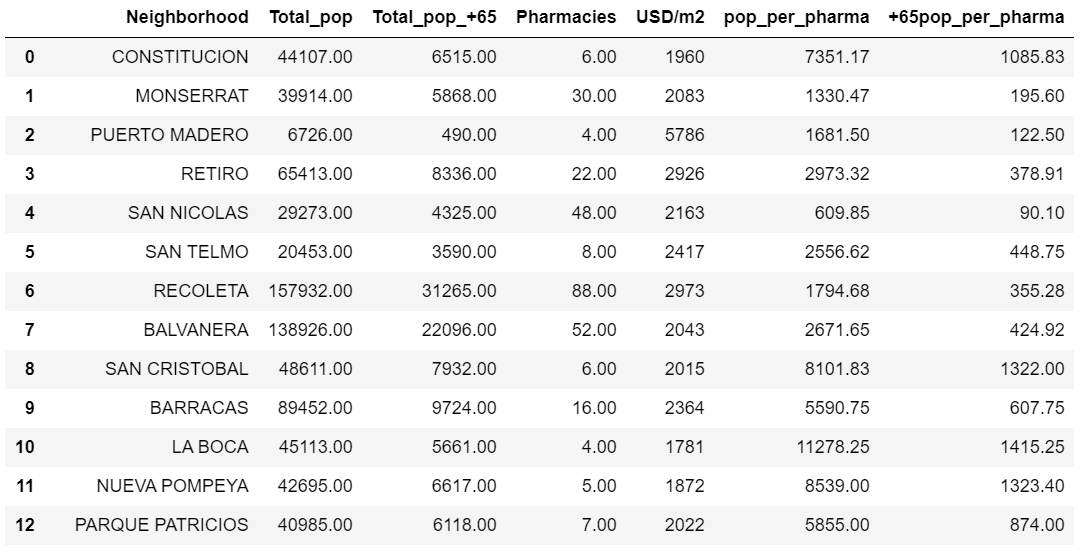


Figure 10

Columns “pop\_per\_pharma” and “+65pop\_per\_pharma” formulas:

* pop\_per\_pharma=Total\_pop/Pharmacies
* +65pop\_per\_pharma=Total\_pop\_+65/Pharmacies

As mentioned in section 2.1, the rate of persons per existing pharmacies is a very important feature. I decided to add +65pop\_per\_pharma as older people are more likely to make pharmacies purchases, especially medication. So, if neighborhood A has a similar “pop\_per\_phama” as neighborhood B but neighborhood A has a better “+65pop\_per\_pharma” than B Neighborhood A should be one step ahead than B to start a new pharmacy.