Coursera Capstone Project

Buenos Aires city Subway Station Analysis

Study for determining a new Store Shop location

Pablo Pedrazas

October 2019

Table of Contents

Introduction

Buenos Aires is the capital city of Argentina, and the most populated in the country, hosting around 3 millions of citizens in their populated neighborhoods. In this presentation we will define a problem to solve using Data Science techniques, based in the assessment of the Subway stations of Buenos Aires, and the exploration of Foursquare.com places found among them.

The problem selected for this example is to determine best places to open a Store Shop in the city of Buenos Aires in the proximity of subway stations.
We will review all the stores and referenced places in the nearby of Subway stations around the City, to determine the most common shop to find are Coffee Shops, so the goal will be to determine the location of future Coffee Shops locations.

There are a high number of variables to consider in such decision, but for the example of this practice, we will base our analysis in several data sets publicly available from Buenos Aires Subway network, Buenos Aires city government, and from data pulled using Foursquare,com exploration API, leaving other important data, measurable or not, outside of the scope of the study.

Problem Description

The problem selected is based in the city of Buenos Aires, in particular near the subway stations.

The Subway network concentrates a high number of persons traveling across the city daily, specially during workdays. This movement attracts Store owners to locate their business close to the Subway Stations, in an intention to make use of this concentration to their own revenue, and to provide services needed to this people while traveling.

In this case, the focus is set particularly in Coffee Shop, considering there are particularly popular in Argentina as meeting points for joining with friends, coworkers, and also for business

Analyzing the available data sets will conclude on best possible options to place a new store to host a Coffee Shop.

Target Audience

The target user for this case is any interested person in reviewing location information for planning the setup of a new Store Shop, in particular a Coffee Shop, in the city of Buenos Aires.

Data Section

The data sets used along this study are:

A. From the Government of the city of Buenos Aires, official information:

1. Neighborhood information, with geographical location
2. Population for each Neighborhood from same source, but based on 2010 official data
3. Subway stations information, including geographical location and line
4. Premetro stations information (additional stations complimenting the subway network)
5. Traffic information of the subway stations

B. From Foursquare.com

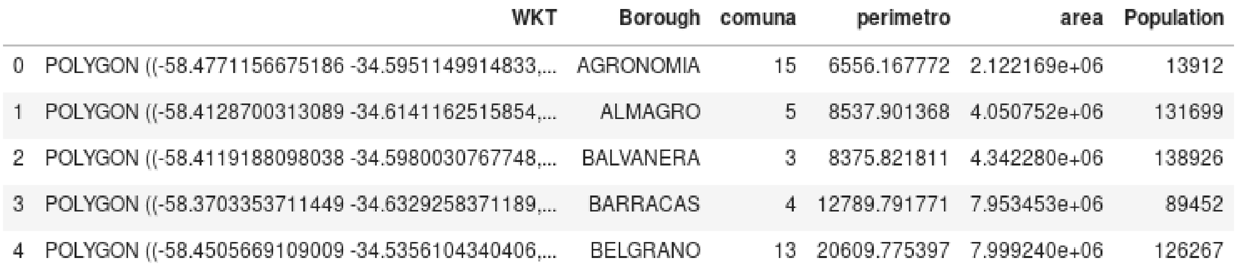
1. Exploration of all Venues near subways stations
2. Exploration of specific categories of Coffee Shops near subway stations

Data acquisition, cleaning and preparation

First two data sets were obtained form Government official information regarding the city itself.

Neighborhood information contains the distribution of the city of Buenos Aires in several neighborhoods, and includes the geographical location of each of them. The second data set was joined to this first collected data to include the population from 2010 official city census.

Acquiring and joining this information together provides the following data set

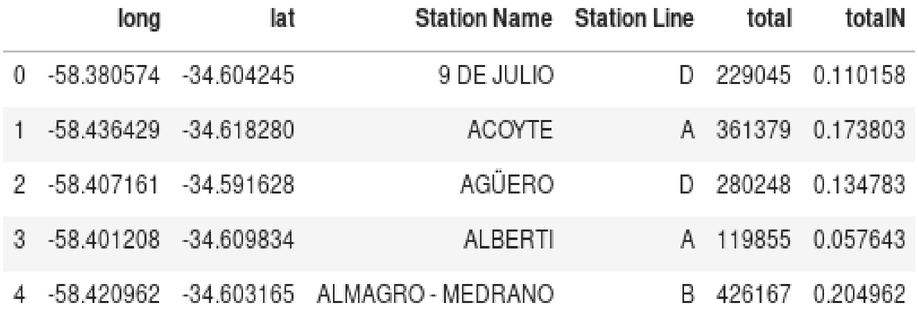


	WKT	Borough	comuna	perimetro	area	Population
0	POLYGON ((-58.4771156675186 -34.5951149914833,...	AGRONOMIA	15	6556.167772	2.122169e+06	13912
1	POLYGON ((-58.4128700313089 -34.6141162515854,...	ALMAGRO	5	8537.901368	4.050752e+06	131699
2	POLYGON ((-58.4119188098038 -34.5980030767748,...	BALVANERA	3	8375.821811	4.342280e+06	138926
3	POLYGON ((-58.3703353711449 -34.6329258371189,...	BARRACAS	4	12789.791771	7.953453e+06	89452
4	POLYGON ((-58.4505669109009 -34.5356104340406,...	BELGRANO	13	20609.775397	7.999240e+06	126267

The subway information contains the location information of each station, recognized by the station name, and part of a station line. Since there are a few stations with same name but in different lines, I kept both fields as referencing keys.

This information was joined with other data set containing the number of passengers passing though each station counter after purchasing a ticket, measured from initial date: 2019-01-01, final date 2019-08-31. It's a big set of data since contains 8300108 data records with valuable information to further review contain passengers by date and time, but out of the scope of the analysis of this problem. Since both data sets were not coincident in the index or Station names, some manual adjustments were performed to match both tables.

For the use of this work, this information was summarized to provide total number of people entering each station by month. An additional column normalized the data was added for further review. This is the final stage of the subway data set:



	long	lat	Station Name	Station Line	total	totalN
0	-58.380574	-34.604245	9 DE JULIO	D	229045	0.110158
1	-58.436429	-34.618280	ACOYTE	A	361379	0.173803
2	-58.407161	-34.591628	AGÜERO	D	280248	0.134783
3	-58.401208	-34.609834	ALBERTI	A	119855	0.057643
4	-58.420962	-34.603165	ALMAGRO - MEDRANO	B	426167	0.204962

The premetro data set contains the extension to the Subway network going at ground level. Since I could not find the traffic information. After checking foursqueare hits in this area, not many Venues where found, so left this data not used for the case of this study.

From foursquere.com, used the API to explore nearby subways stations for all possible hits as a first study. In a first exploratory research, a data set of a total of 7084 "venues" was returned from Foursquare near subway stations. Many of the 90 subway stations reached the limit provided by the Foursquare API reaching the 100 hits, so generated a second exploration running the API again, but this time to collect just the Category related to Coffee Shops.

In this second iteration, a total of 3684 Coffee or Tea shops were provided by Foursquare. Have discarded Tea shops for the scope of this analysis, since due Buenos Aires people behavior, are in a different category than Coffee Shops. A total of 3637 coffee shops were accounted against the subway stations.

More information is presented in the following section

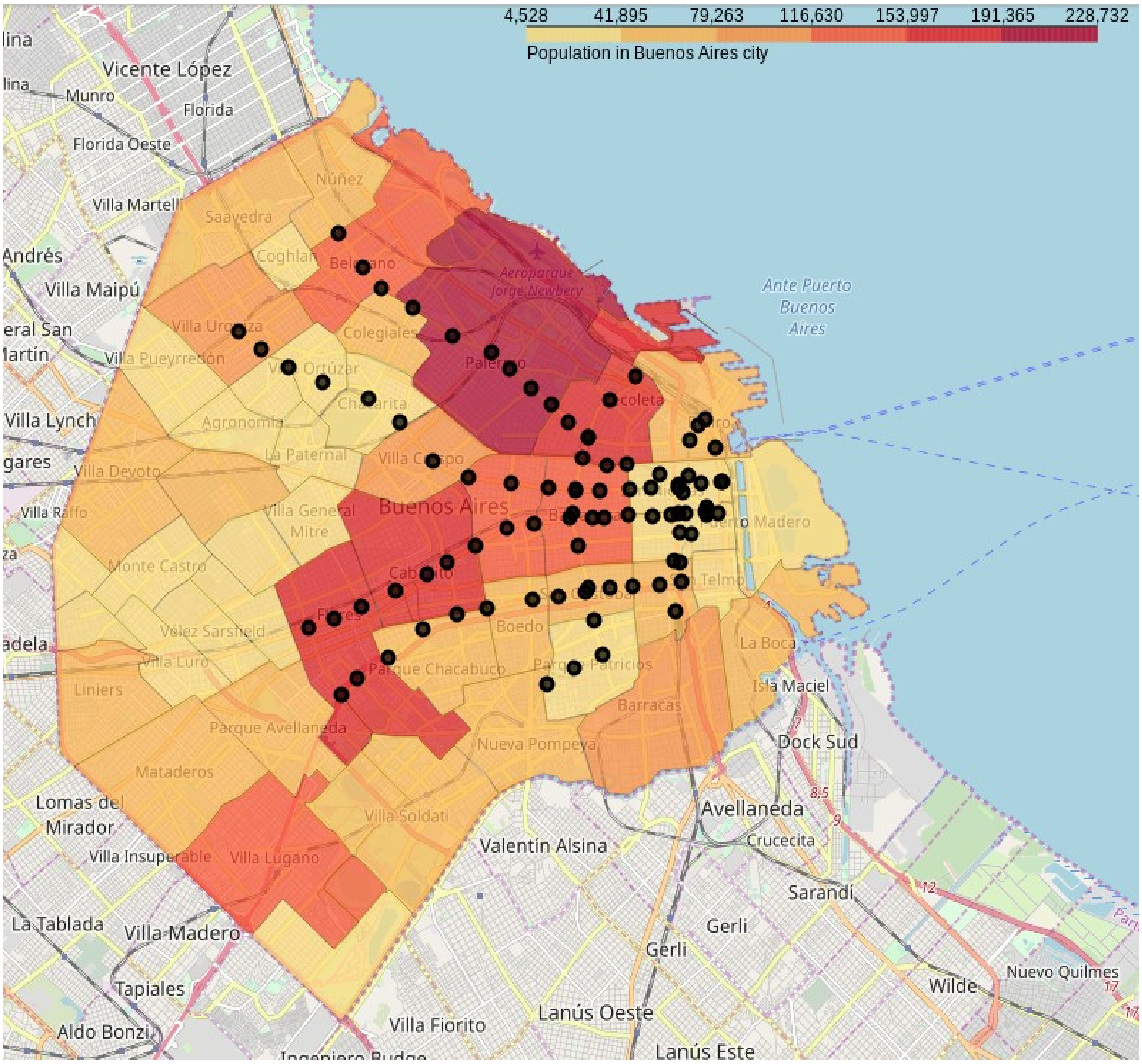
Analysis

In the following sections we presented the assessment made to different subjects

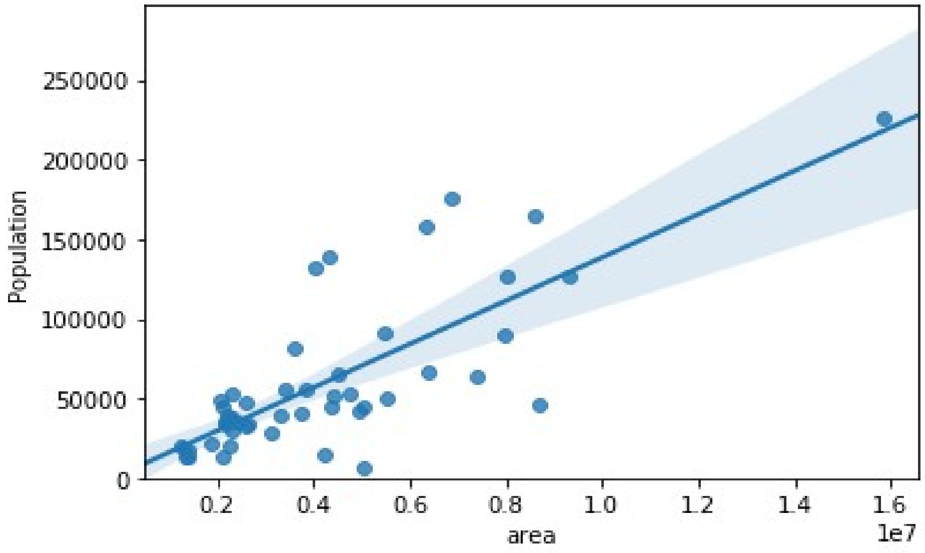
Buenos Aires Subway Stations

The data collected let us start representing the city of Buenos Aires in a Map together with it's population by neighborhood. To do that used Folium library and the Choropleth feature to show the population in a visual representation over the geographical area.

Over the Map I've added black circles to represent each of the subway stations. In a fist approach we see how the subway network is covering most populated neighborhoods.

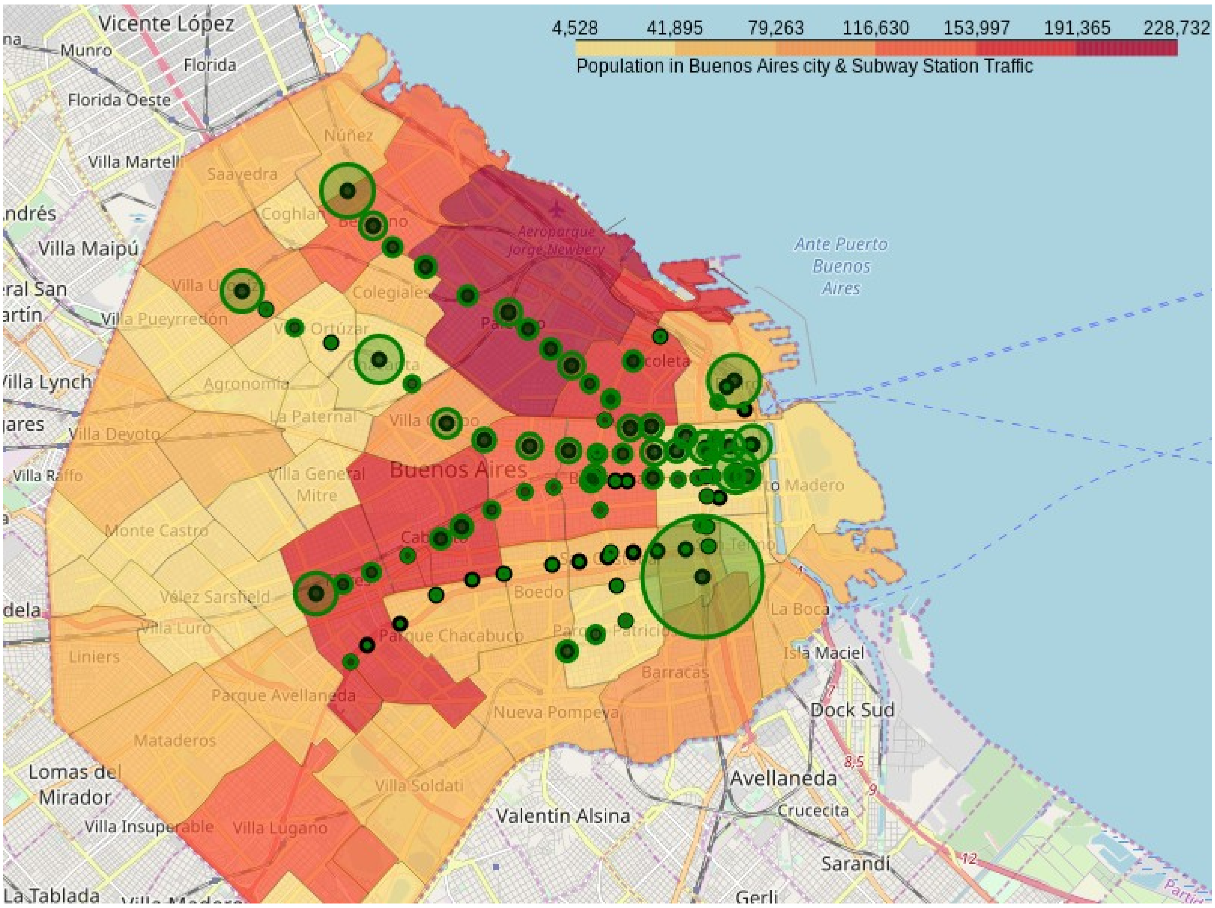


Other exploratory analysis was to confirm the obvious relation between the Population in each neighborhood and the geographical size of the area contained within its limits. This was made using a scatter plot and a regression line to confirm a positive linear relationship between both variables:



The subway information was transformed to show the traffic amount concentrated in each station, processing them grouping information after discarding irrelevant data for quicker data processing. The premetro data set was discarded since have missing information about the traffic, so concentrated in the core of the stations traffic data, using normalization technique to better study the traffic. Using the Simple Feature scaling method, found the higher station was Station Name "Constitución".

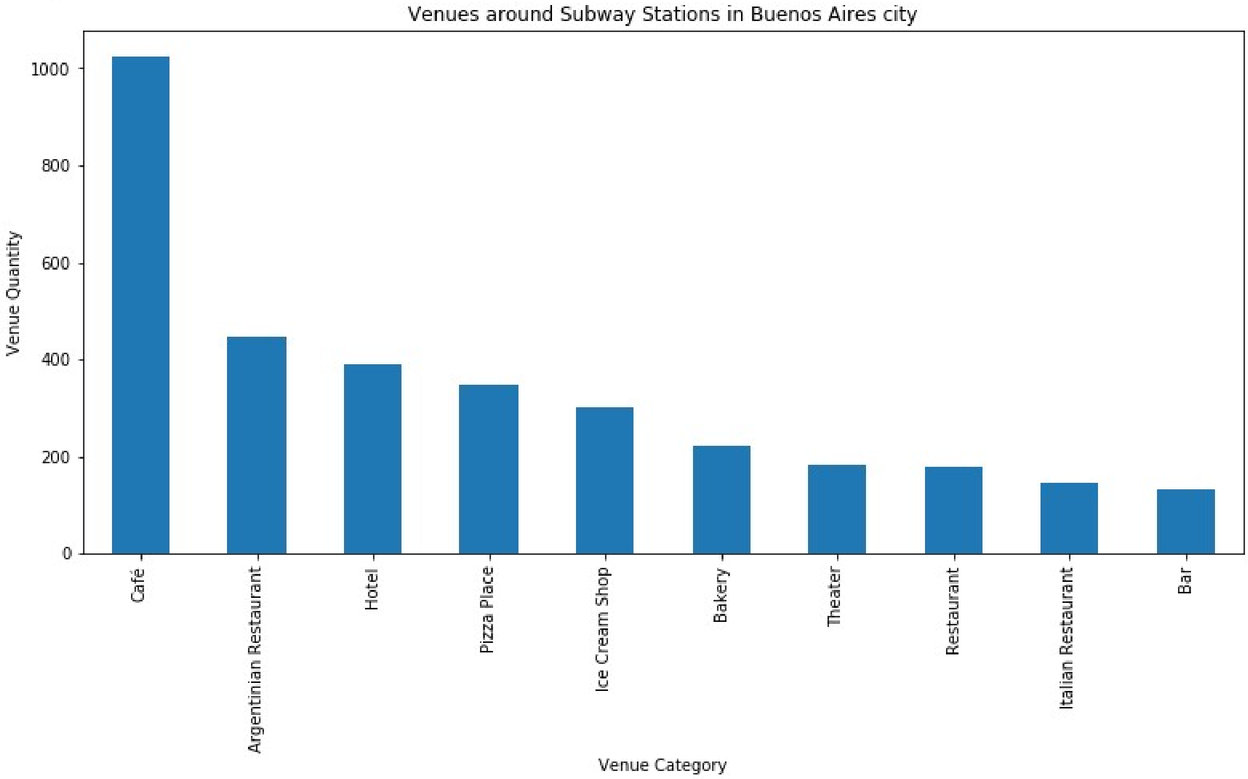
In the following figure it's represented with circles with their ratios proportional to the traffic accounted:



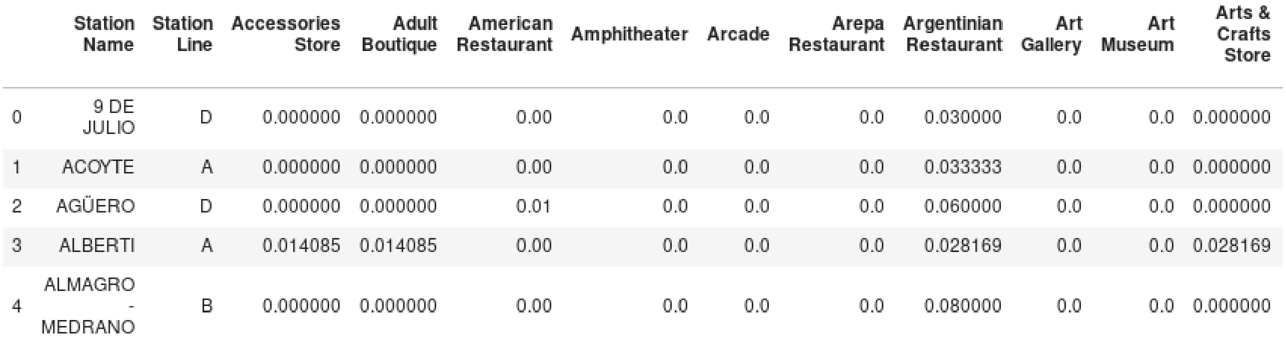
As any big city, with even more populated surroundings and subways network not covering all of them, the Stations with higher traffic are the ones located at the edges of each line, since people travel by bus or other way to reach the terminal and commute.

Foursquare Venues

The general exploration of places near the limit of 800 meters selected around each subway station, presented 7084 venues. After processed the data, and properly grouped, we found there were 249 unique categories. Among them, realized the most important place was for Coffee Shops. The most important categories are represented in the following bar chart:

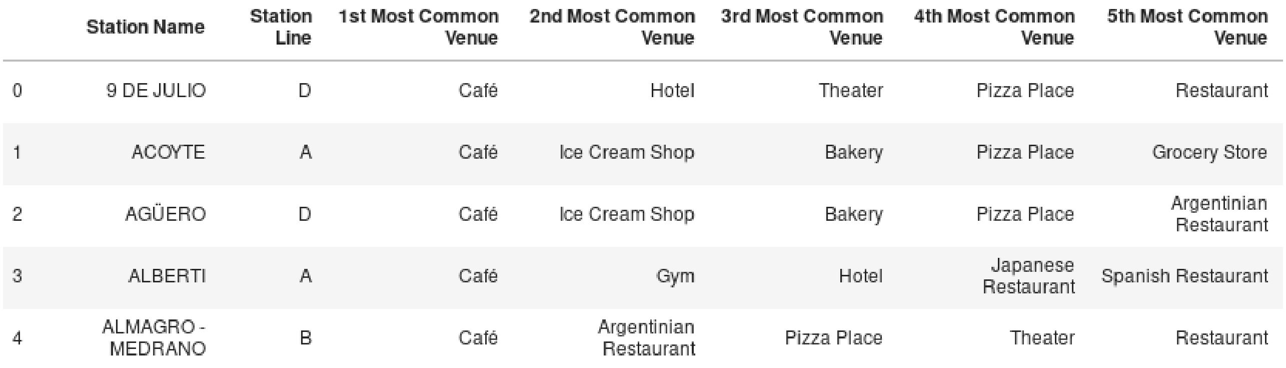


To analyze the hints by each station, processed the data grouping by station and normalizing the data by taking the mean of the frequency of occurrence of each venue category.. The mean method will return the average value of the feature in the data set:



	Station Name	Station Line	Accessories Store	Adult Boutique	American Restaurant	Amphitheater	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store
0	9 DE JULIO	D	0.000000	0.000000	0.00	0.0	0.0	0.0	0.030000	0.0	0.0	0.000000
1	ACOYTE	A	0.000000	0.000000	0.00	0.0	0.0	0.0	0.033333	0.0	0.0	0.000000
2	AGÜERO	D	0.000000	0.000000	0.01	0.0	0.0	0.0	0.060000	0.0	0.0	0.000000
3	ALBERTI	A	0.014085	0.014085	0.00	0.0	0.0	0.0	0.028169	0.0	0.0	0.028169
4	ALMAGRO - MEDRANO	B	0.000000	0.000000	0.00	0.0	0.0	0.0	0.080000	0.0	0.0	0.000000

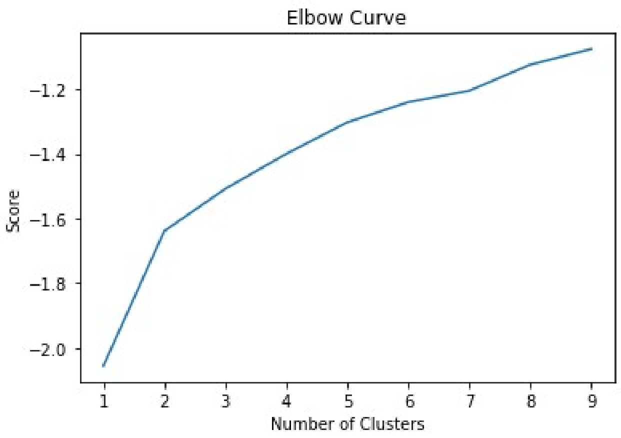
With the data normalized, it’s easier to order the importance of each feature to create a data set containing the most frequent venue category by each station:



	Station Name	Station Line	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	9 DE JULIO	D	Café	Hotel	Theater	Pizza Place	Restaurant
1	ACOYTE	A	Café	Ice Cream Shop	Bakery	Pizza Place	Grocery Store
2	AGÜERO	D	Café	Ice Cream Shop	Bakery	Pizza Place	Argentinian Restaurant
3	ALBERTI	A	Café	Gym	Hotel	Japanese Restaurant	Spanish Restaurant
4	ALMAGRO - MEDRANO	B	Café	Argentinian Restaurant	Pizza Place	Theater	Restaurant

Clustering techniques is a machine learning technique that let us group, classify or segment information even with no clear relation or even with unlabeled data. Using K-Means unsupervised clustering algorithm I processed the previous dataframe to find possibles segmentation. This model, if success, could better let me answer the question of where to place a new Store Shop.

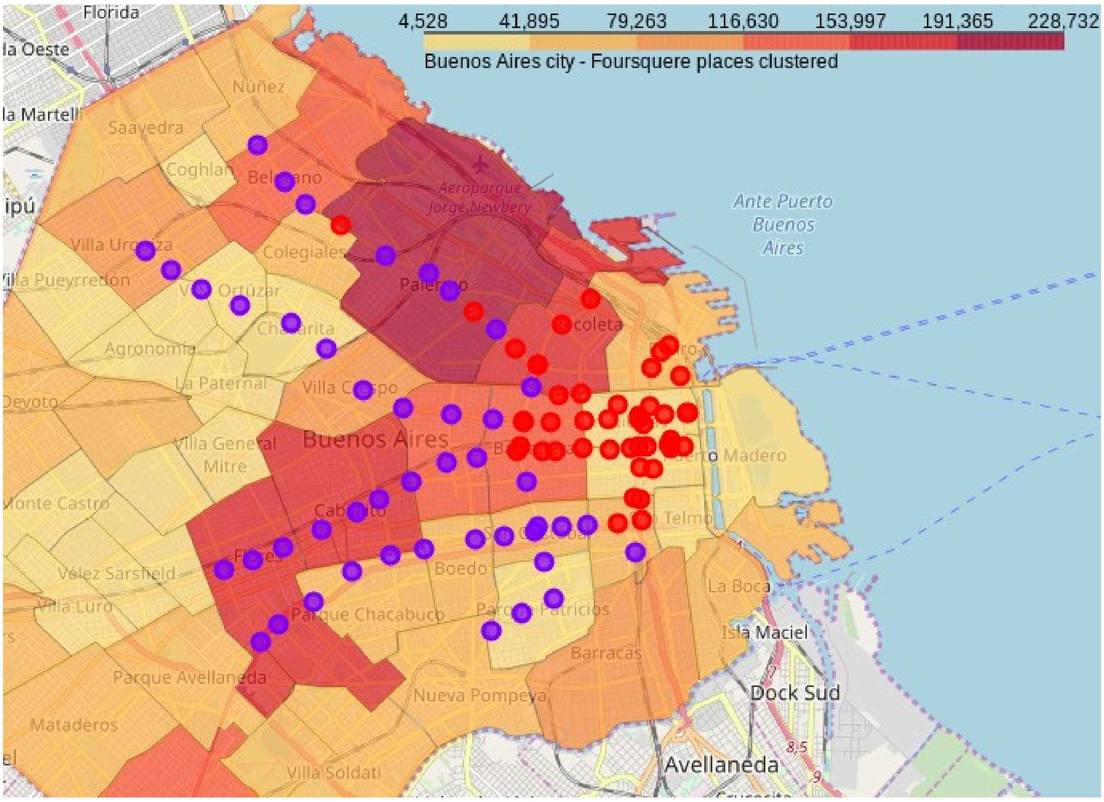
In order to better evaluate how many segments to find, we use the Elbow Curve method to better determine the proper clustering number:



The Elbow curve obtained suggests either 2 or 5 clusters needed. Running the cluster method with both options I was not able to conclude any relevant reason to describe segmenting in 5 groups.

The segmentation in two groups looks simpler describing a first group concentrated near downtown, higher traffic areas, characterized by the presence not only of Coffee Shops or “Cafes” as the most common venue, together with Hotels, Theaters and other attractions and food places.

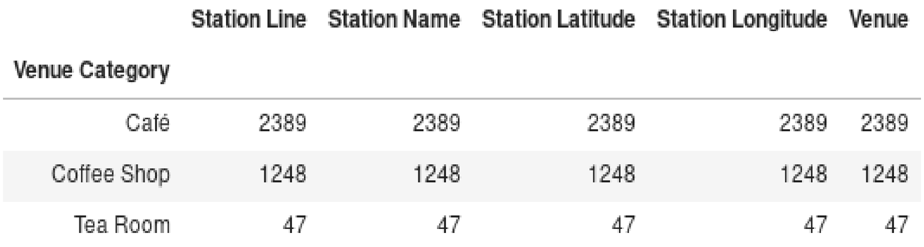
The remaining cluster was in the outer circle of the city, and even if also present food related places, the order is not the same and lacks of Hotels and attractions as theaters.



Foursquare Coffee Shops

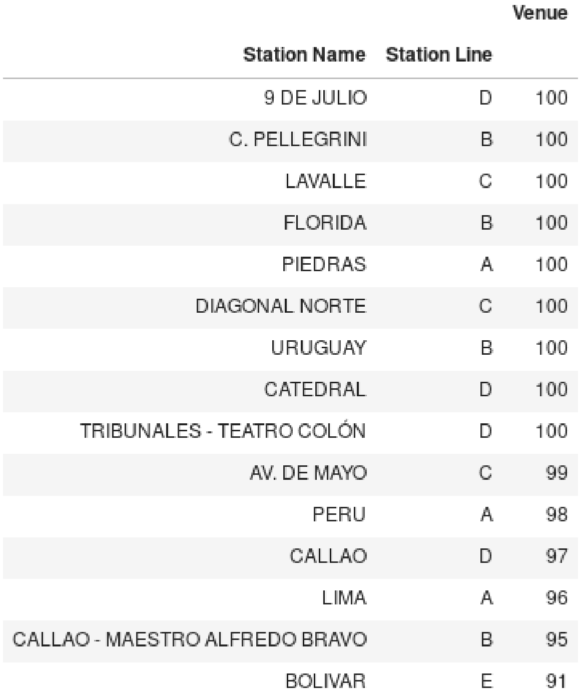
To better focus the study in the objective of Te business question proposed of where to place a new Coffee Shop, we run another exploration just targeting this category only. This way we explore the same nearby around each station, reducing the impact of the 100 limit of the standard free use of the API.

The category includes both Coffee or Tea shops, with a total of 3684 hints only in this category:



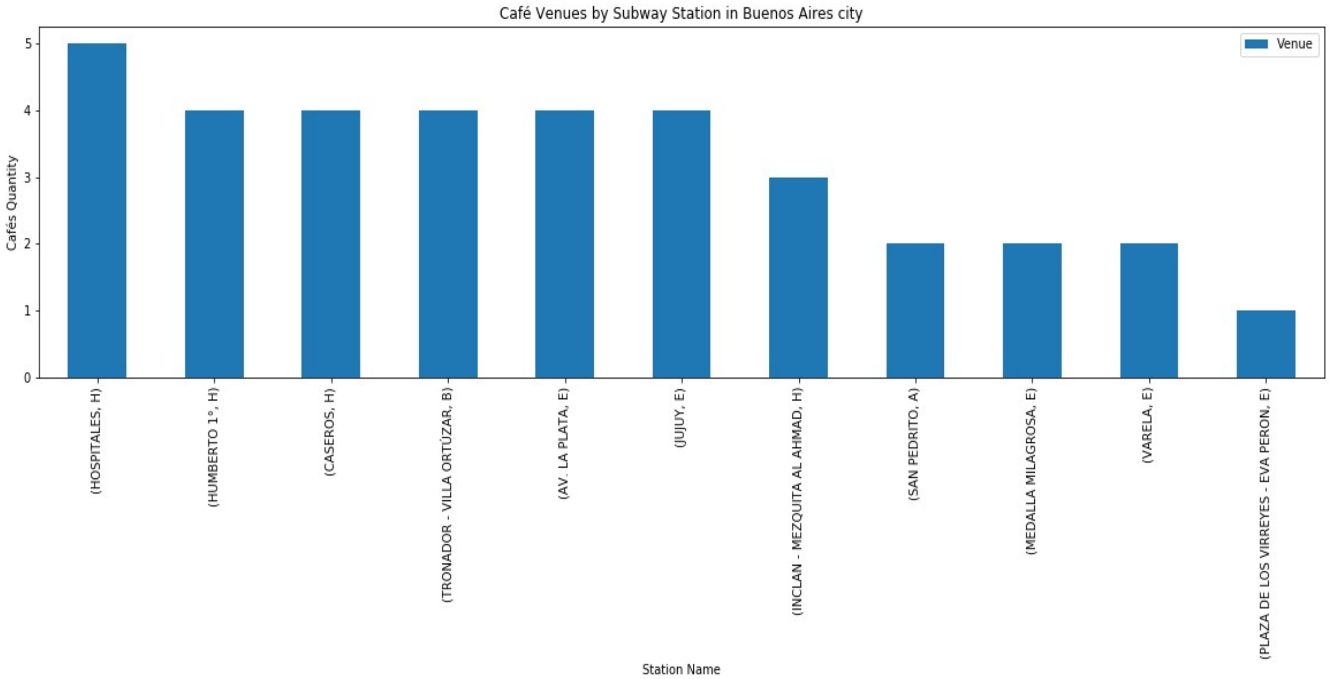
Venue Category	Station Line	Station Name	Station Latitude	Station Longitude	Venue
Café	2389	2389	2389	2389	2389
Coffee Shop	1248	1248	1248	1248	1248
Tea Room	47	47	47	47	47

After discarding Tea Rooms, and joining Cafe and Coffee Shops as a single category, proceed to account them and group by Station Name. Sorting descending, still found stations reaching the 100 limit of the API.

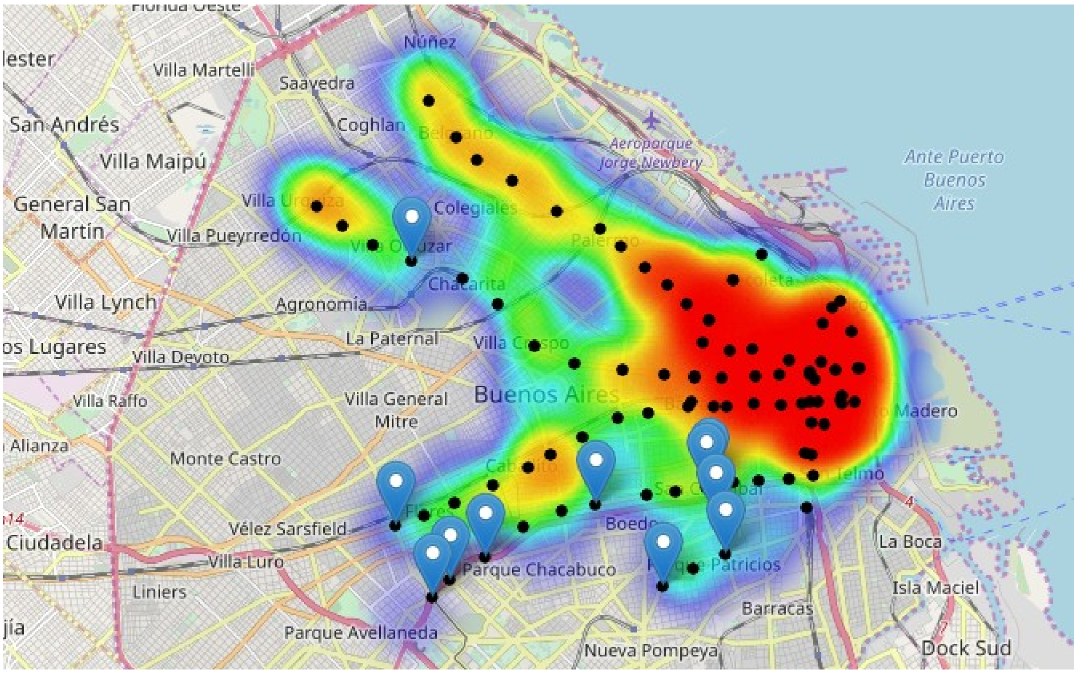


		Venue
Station Name	Station Line	
9 DE JULIO	D	100
C. PELLEGRINI	B	100
LAVALLE	C	100
FLORIDA	B	100
PIEDRAS	A	100
DIAGONAL NORTE	C	100
URUGUAY	B	100
CATEDRAL	D	100
TRIBUNALES - TEATRO COLÓN	D	100
AV. DE MAYO	C	99
PERU	A	98
CALLAO	D	97
LIMA	A	96
CALLAO - MAESTRO ALFREDO BRAVO	B	95
BOLIVAR	E	91

While this can hide some information if studying the top Stations, we will focus on the Stations with less offering of Coffee Shops places, as shown in this graphical representation:



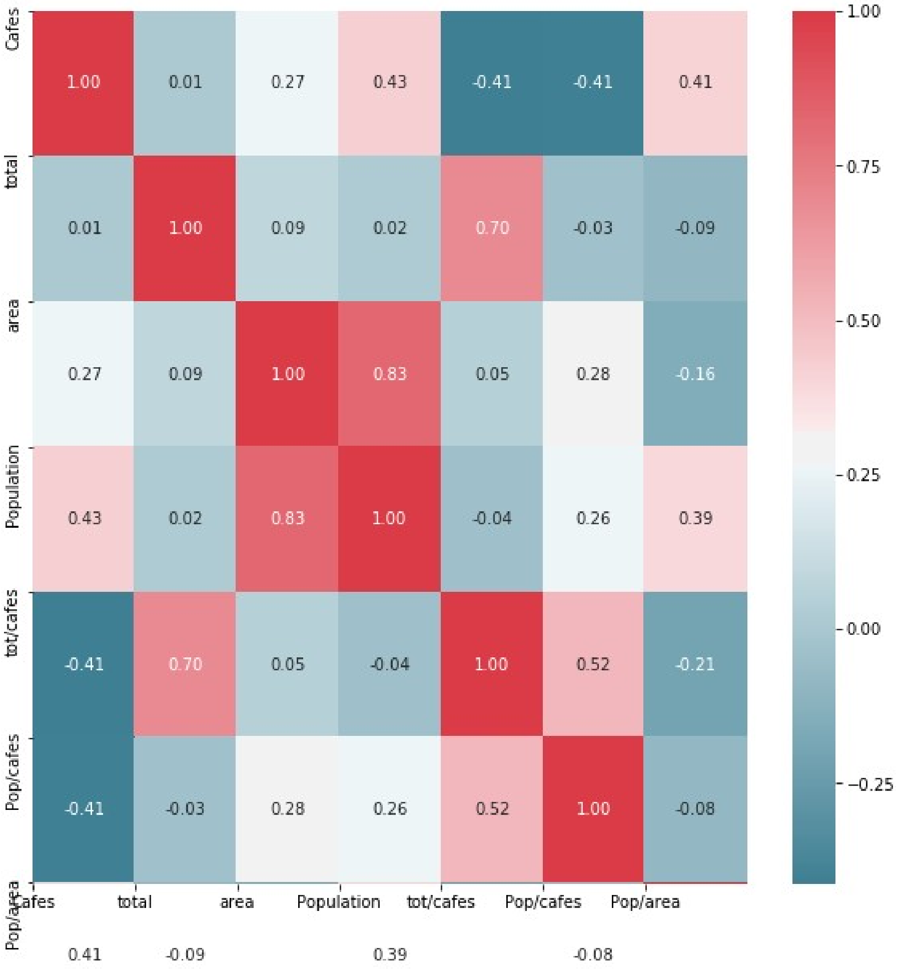
To better determine with geospatial information, created a Heatmap over the city representing the agglomeration of Coffee shops around the subway stations. The stations subject to further study, the less with Coffee places, are marked with the blue signal:



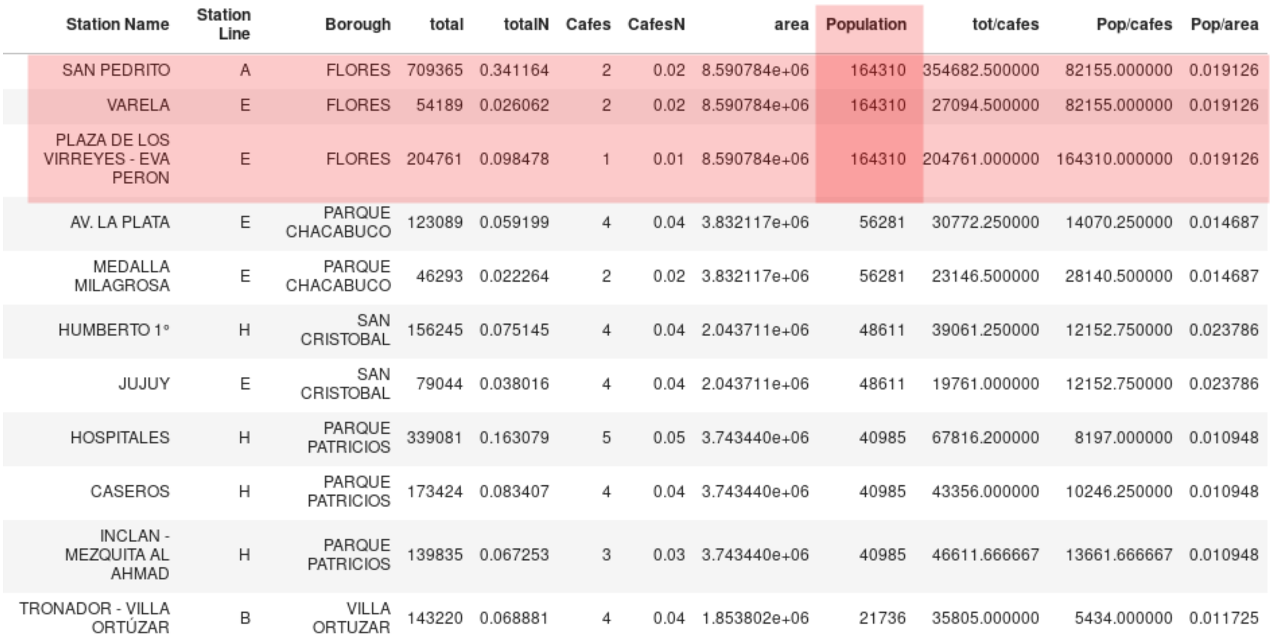
So having this graphical representation provides a better idea of where to look from. Of course can also use the map with the complete representation of each Coffee Shop in the map, but trying to find an analytic predictor, I added to the data set all other available features hoping some could work as predictor, or have a strong relationship with the number of Cafes found. To complete the study, I build a matrix of correlation working with the data sets available, consisting in:

- Neighborhood Area
- Neighborhood Population
- Monthly traffic per Station
- Area/Population

For a better visualization of the results, I build a heatmap color coding the better relations; high correlations are displayed in increasing red shades, while lower correlations trending into the blue hues. Unfortunately no direct correlation is found for the entire Stations data set, and although there are negative linear relationships those are still weak, and not representative. Reduced the set of data to the stations with less than 50 Coffee shops found and got this new representation:



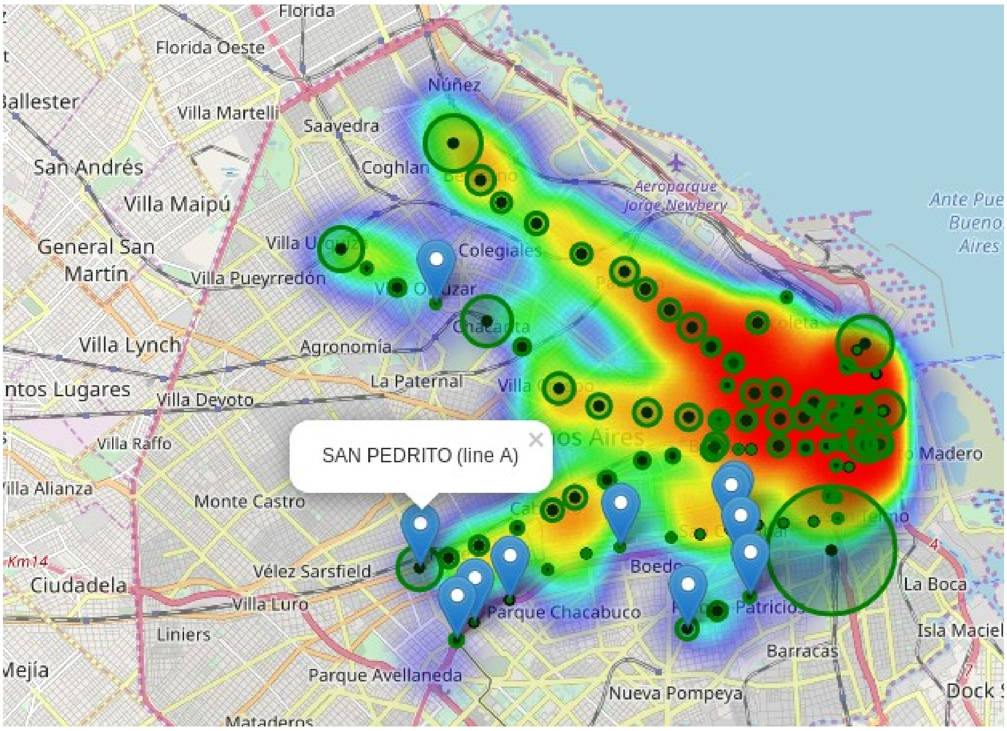
In this correlation matrix, the population data set has a positive linear correlation with the Cafes found. Using the sub data frame, with the less exploited Stations and sorted by this Population as predictor, determined the best 3 candidates for the target objective:



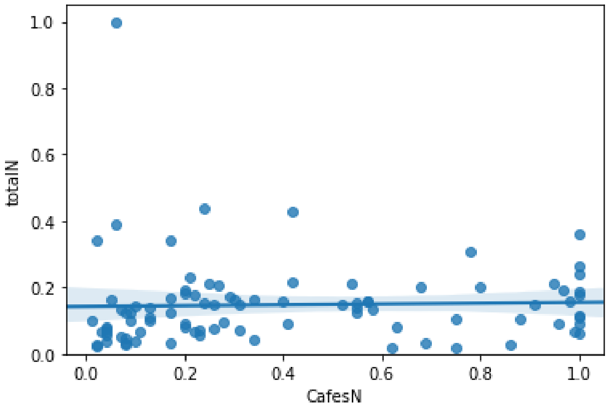
Station Name	Station Line	Borough	total	totalN	Cafes	CafesN	area	Population	tot/cafes	Pop/cafes	Pop/area
SAN PEDRITO	A	FLORES	709365	0.341164	2	0.02	8.590784e+06	164310	354682.500000	82155.000000	0.019126
VARELA	E	FLORES	54189	0.026062	2	0.02	8.590784e+06	164310	27094.500000	82155.000000	0.019126
PLAZA DE LOS VIRREYES - EVA PERON	E	FLORES	204761	0.098478	1	0.01	8.590784e+06	164310	204761.000000	164310.000000	0.019126
AV. LA PLATA	E	PARQUE CHACABUCO	123089	0.059199	4	0.04	3.832117e+06	56281	30772.250000	14070.250000	0.014687
MEDALLA MILAGROSA	E	PARQUE CHACABUCO	46293	0.022264	2	0.02	3.832117e+06	56281	23146.500000	28140.500000	0.014687
HUMBERTO 1°	H	SAN CRISTOBAL	156245	0.075145	4	0.04	2.043711e+06	48611	39061.250000	12152.750000	0.023786
JUJUY	E	SAN CRISTOBAL	79044	0.038016	4	0.04	2.043711e+06	48611	19761.000000	12152.750000	0.023786
HOSPITALES	H	PARQUE PATRICIOS	339081	0.163079	5	0.05	3.743440e+06	40985	67816.200000	8197.000000	0.010948
CASEROS	H	PARQUE PATRICIOS	173424	0.083407	4	0.04	3.743440e+06	40985	43356.000000	10246.250000	0.010948
INCLAN - MEZQUITA AL AHMAD	H	PARQUE PATRICIOS	139835	0.067253	3	0.03	3.743440e+06	40985	46611.666667	13661.666667	0.010948
TRONADOR - VILLA ORTÚZAR	B	VILLA ORTUZAR	143220	0.068881	4	0.04	1.853802e+06	21736	35805.000000	5434.000000	0.011725

Results

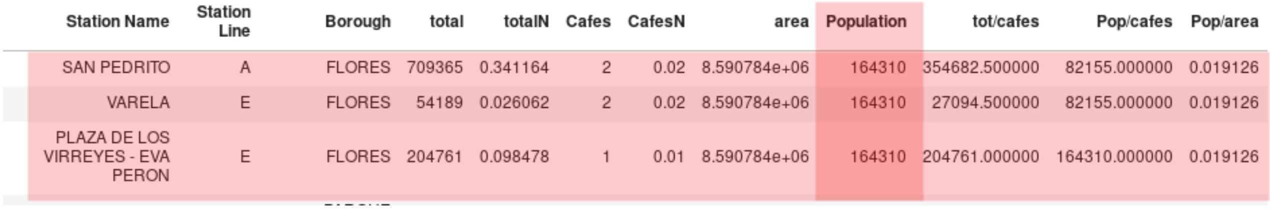
After combining the previous heatmap, with the traffic per station map represented before, we got this representation. The blue markers are pointing to the subway stations with less Coffee Shops found, and therefore candidates for new ones. Among them, “San Pedrito” station is the only one with higher traffic than the others, represented by the green circle.



Even I could not demonstrate using statistical correlation a strong relation between passengers traffic and the number of Coffee Shops found in the nearby, I estimated that Neighborhood population is a direct correlation that could influence this decision.



Not all neighborhoods had the same quantity of subway stations to test this relation, and what is most important, due the area of the Neighborhoods and the distribution of the subway lines, not all the population will make use of the subway. But with the help of the following data of the stations with less Coffee shops, and the ranking of the relations traffic/cafes and population/cafes, the top 3 Stations where to recommend the opening of a new Coffee Shop:



Station Name	Station Line	Borough	total	totalN	Cafes	CafesN	area	Population	tot/cafes	Pop/cafes	Pop/area
SAN PEDRITO	A	FLORES	709365	0.341164	2	0.02	8.590784e+06	164310	354682.500000	82155.000000	0.019126
VARELA	E	FLORES	54189	0.026062	2	0.02	8.590784e+06	164310	27094.500000	82155.000000	0.019126
PLAZA DE LOS VIRREYES - EVA PERON	E	FLORES	204761	0.098478	1	0.01	8.590784e+06	164310	204761.000000	164310.000000	0.019126

To arrive to this result, we support it using Geospatial representation over the city map. Folium maps library was used to map the stations over the city, and using chloroplets feature, I was able to add the neighborhood population where each station resides.

Recommended Station surrounding where to open new Coffee Shop:

1. San Pedrito – (Line A)
2. Varela (Line E)
3. Plaza de los Virreyes – Eva Peron (Line E)

Observations

The study is based in testing the data sets available, and using Data Science techniques to elaborate a response to the target problem of finding a good place to setup a new Coffee Shop in Buenos Aires.

In the data studied we included neighborhoods area and population. Considering the distribution of the subway lines is not evenly spread about all neighborhoods, and that some areas of them might not be suitable for the use of the subway, the analytic approach might not be exact, so included graphical representations to help in the decision to made.

There are many other variables not covered in this study that should follow this initial recommendation to compliment and therefore, confirm or reject the result provided. The market value for rent in the city of Buenos Aires should be included in addition to the available properties in the market, the state of each one, the investment depending on that and many others out of the scope of this study.

Conclusion

This study have the objective to study the Buenos Aires shops nearby the subways stations, in particular Coffee Shops distributions to find out possible places where to set up a new shop.

I used data collection techniques to incorporate relevant external data for the subject of this study. After cleaning and preparing it, grouped it properly started to study the different sets. I use statistics correlation to determine the incidence of the population and the passengers per station to became a predictive modeling. Tried using K-Means clustering to determine pattern recognition to segment places found, but no strong evidence was found to make a recommendation.

Collect additional information for the particular kind of bar commonly accepted in Buenos Aires, the Coffee Shops. Again, without finding a strong correlation to the passengers traffic, I made a correlation matrix with all possible variables. When worked the matrix with the Stations which less Coffee shops found, the neighborhood population feature come up as a predictor with a direct relation. Ordered the set of stations in study, in descendant order by this feature let me arrive to an analytic final response of 3 stations as possible candidates. The graphical representation was key to provide an approximation to the answer. Using geospatial data and representing it over a city map containing subways stations works as an aid to find the less serviced stations with this venue. Using an analytic approach, and grouping the data properly, the ranking for 3 stations with more population, more transit and less Coffee shops were suggested as result.