# Final Project: Location analysis tool using data from Foursquare

Anthony Alarcón Aguirre
February, 2019

## 1. Introduction

Chances are, you've heard the term "location, location, location" more than a few times. But if you're in the throes of creating a spectacular menu for your new restaurant or finding wholesalers for your first retail store, it might not be the first thing on your mind.

It's time to put location at the top of your to-do list. If you're preparing to open a food or retail business with a storefront, putting your business in the proper location might be the single most important thing you do at startup. Of course you need a winning product, too, but how will anyone know about that product unless you get them through the door?

"In the brick-and-mortar retail world, it's said that the three most important decisions [you'll make] are location, location and location," affirms Irene Dickey, a lecturer in management and marketing at the University of Dayton's School of Business in Dayton, Ohio. "Careful determination of new sites is critical for most retail and consumer service businesses."

## 2. Business Problem

The success of a business depends a lot on its location. The street, the orientation and the views are fundamental. But so is the economic activity that is in the vicinity. Currently, the location of a company or mall is very important because it is a key factor for the generation of profits. That is why many entrepreneurs and entrepreneurs come to tools and market analysis that allow them to define the place.

A very important decision, after having chosen the product you will sell, is the geographical place where you will offer it. Your choice of this site can mean the success or failure of your business. Here's the business problem: Is there a tool or solution that helps entrepreneurs define the location of their business?

For this case, we will focus on a group of entrepreneurs who decide to open a pastry shop in Santiago de Surco - Lima, Peru.

Pastry shops or candy stores benefit from impulse buying and it is important to be in areas of high traffic:

- Near a schools. For the snack, the aperitif of the recess or a sweet to the exit.
- Residential areas. To satisfy the demand of children's whims or those weekend food cakes.
- Commercial areas. Impulse purchases need a lot of traffic.

### 3. Data acquisition and cleaning

In order to provide a tool alternative to define the location of a business, Foursquare location data will be used.

Foursquare is a technology company that uses smart location to build a meaningful consumer experience and business solutions.

For the present case, we will use the district of Santiago de Surco and we will analyze the most visited places by the people. As it was explained, for this purpose the Foursquare service will be used:

Example:

- District: Santiago de Surco, Lima - Peru
- Latitude: -12.1251049
- Length: -76.9819193843966

This information will be key to analyze through the incorporation of geospatial points.

### 4. Metodology

In the first place, the working methodology was the use of python as programming language and to achieve the connection with the necessary libraries to map the locations.

In the second place, the connection was made with the Foursquare service, in which the credentials were validated and the data could be accessed.

In the third place, geolocation services were used to map the places with the highest concentration of people and detect other nearby shopping centers that allow analyzing the location of the bakery.

Finally, with the obtained results, a clustering was performed using the Kmeans algorithm. This was necessary because we had the location, latitude and longitude properties of each business near the location under study. What we sought is to segment each business by proximity. With the aim of redirecting and analyzing where the pastry shop could be located.
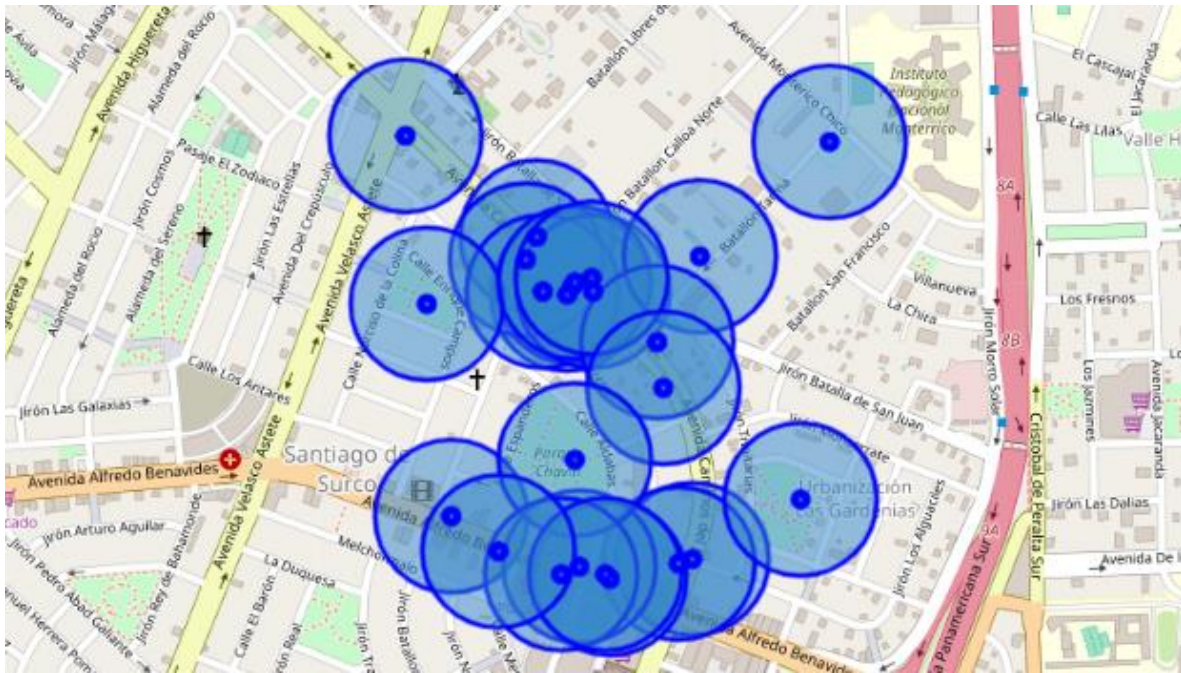
## 5. Results

The first objective of the project was to extract information from Foursquare locations. To do this, the connection was made and as the first procedure to follow was to define a latitude and longitude reference. As we stated in the discussion of the problem, we seek to find the best position in Santiago de Surco.
This is to seek the areas that have the highest concentration of people and that is as commercial as possible.

In which we can observe, it extracts information from businesses close to the chosen area. In addition, a very important variable to take into account is the business categories. As we know, what is proposed is to place a pastry shop, then we have to look for central areas and it is very important to analyze what kind of businesses are near or in which location we should open the business to maximize visits to the premises. On the other hand, we have the address of the premises, the distance to the point under study. And finally, the analysis variables that are latitude and longitude.

The results of the extraction of data by the Foursquare service shows the following structure:

| Name | Categories | Address | CC | City | Country | Distance | FormattedAddress | Lat | Lng |
|---|---|---|---|---|---|---|---|---|---|
| Dalcello | Ice Cream Shop | Caminos del Inca 1502 | PE | Surco | Perú | 152 | [Caminos del Inca 1502, Surco, Perú] | -12.12544 | -76.98328 |
| Il Piccolo Pomodoro | Italian Restaurant | Av. Caminos del Inca 1473 | PE | Surco | Perú | 202 | [Av. Caminos del Inca 1473, Santiago de Surco, Perú] | -12.12464 | -76.98372 |
| Tierra Santa | Middle Eastern Restaurant | Av. Caminos del Inca 1512 | PE | Surco | Perú | 142 | [Av. Caminos del Inca 1512, Santiago de Surco, 33, Perú] | -12.12533 | -76.98321 |
| La Verdad de la Milanesa | Restaurant | Av. Caminos del Inca 1478 | PE | Surco | Perú | 212 | [Av. Caminos del Inca 1478, Surco, LIMA 33, Perú] | -12.12495 | -76.98386 |

The following graph shows the point under study and the references that are found around it.

Mapping the points in analysis, an information was assembled with the following structure:

- Name
- Categories
- Latitud
- Longitud

This with the objective of running a segmentation of nearby shopping centers. For this the Kmeans algorithm was trained with 3 groups. Because the sample information obtained is small, the creation of only 3 groups was defined.

*K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:*

1. *The centroids of the K clusters, which can be used to label new data*

2. *Labels for the training data (each data point is assigned to a single cluster)*

*Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.*

*Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.*

With this we seek to give a categorization to each group.

| Cluster Label | Name | Categories | Lat | Lng |
|---|---|---|---|---|
| 2 | Dalcello | Ice Cream Shop | -12.125442 | -76.98328 |
| 2 | Il Piccolo Pomodoro | Italian Restaurant | -12.124636 | -76.98372 |
| 2 | Tierra Santa | Middle Eastern Restaurant | -12.125327 | -76.983209 |
| 2 | La Verdad de la Milanesa | Restaurant | -12.124947 | -76.983864 |
| 2 | Roll-Star Caminos Del Inca | Sushi Restaurant | -12.125382 | -76.983636 |
| 2 | Parque Juan Bielovuciv (Las Ardillas) | Park | -12.12556 | -76.985258 |
| 1 | Chifa San Joy Lao | Chinese Restaurant | -12.129051 | -76.981514 |
| 2 | Don Mamino | Bakery | -12.123241 | -76.985555 |
| 2 | Chiwake - comida peruana | Restaurant | -12.125389 | -76.982919 |

| | | | | |
|---|---|---|---|---|
| 2 | Chifa Fu Sen | Chinese Restaurant | -12.125264 | -76.983174 |
| 0 | Club Provincial Pacasmayo | Lounge | -12.124898 | -76.981413 |
| 1 | Chicharrones del Inca | Sandwich Place | -12.129128 | -76.981722 |
| 0 | Academia Alejo Aramburu | Tennis Court | -12.123343 | -76.979596 |
| 1 | Tai Loy | Paper / Office Supplies Store | -12.129343 | -76.982673 |
| 0 | Mis Costillitas | BBQ Joint | -12.12608 | -76.982019 |
| 0 | Parque Mochica | Park | -12.128243 | -76.979993 |
| 2 | Isushi | Asian Restaurant | -12.125194 | -76.982944 |
| 1 | Burger King | Fast Food Restaurant | -12.128464 | -76.984924 |
| 1 | Pollos y Parrilladas America | Fried Chicken Joint | -12.129173 | -76.983101 |
| 0 | La Choza de la Anaconda | Peruvian Restaurant | -12.126703 | -76.981923 |
| 1 | Parque Las Gardenias | Park | -12.127698 | -76.983175 |
| 1 | Chantilli | Bakery | -12.129279 | -76.983373 |
| 1 | Chifa Buena Fortuna | Chinese Restaurant | -12.129275 | -76.982742 |
| 1 | Roky's | BBQ Joint | -12.128966 | -76.984256 |

Finally, the following results were obtained, it was possible to segment the neighboring points and classify them by proximity. With the image we noticed 3 large defined groups that synthesize 3 large commercial balloons.

## 6. Discussion

As we can see in the following image, we managed to clusterize neighboring locations by proximity. Now with the results obtained, we can note that there are 3 large commercial balloons that make up a small part of Santiago de Surco. In this way, you can have a clearer idea of where to open the business. A very key option to open it at a point where it is adjacent to the groups or what is most acceptable by affinity to each group.
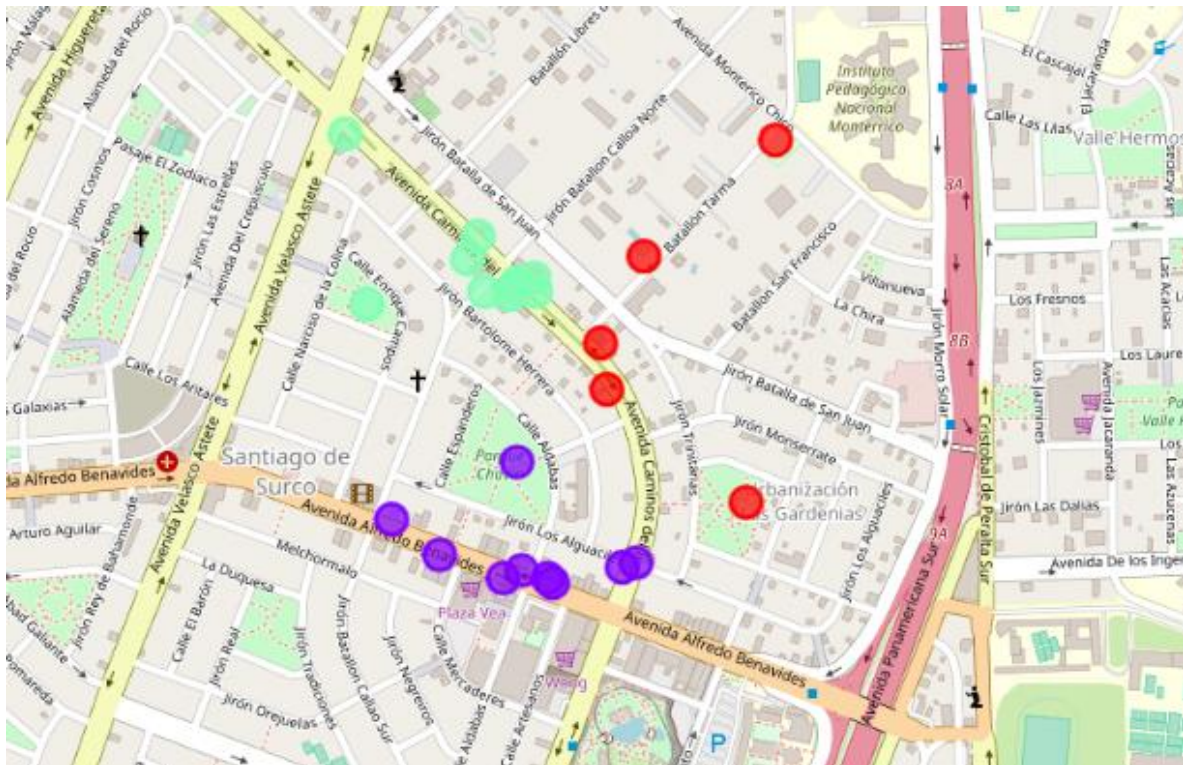


## 7. Conclusion

Finally, it can be concluded that advanced analytical tools are very enriching. Taking this case in specific, the location of a business is a key aspect that can define the success or failure of a business. That is why the constant alternatives that exist in the market to solve the problem. What has been exposed is one of the forms