THE NEXT BAKERY

[CAPSTONE PROJECT]

## CARLOS GONZALEZ ROMERO

# INTRODUCTION / BUSINESS PROBLEM

### Background

The Levaduramadre brand is a gourmet bakery that is in full expansion in the city of Madrid, Spain. Although it was created in 2007, it is in 2017 that this expansion has been made. This expansion is happened mainly in the center and north-center of Madrid.

### Problem

To make the expansion in a homogeneus way, we will try to find an optimal location for a new bakery of the brand Levaduramadre in the south-center of Madrid. To do this, we will assume that the bakery owner gives us a choice of 3 specific locations, and we will have to justify our choice between those 3 options based on two criteria:

* How similar are the neighborhoods that currently have a Levaduramadre compared to the neighborhood of future Levaduramadre.
* How many bakeries there are around the future Levaduramadre and what its the distance between them.

### Interest

This report has been requested by the owner of other Levaduramadre interested in opening a new bakery in the south-center of Madrid.

# DATA

### Data sources

For this work we will use the data obtained from Foursquare API. Although it will be the same API, we will obtain three datasets, one for each aim:

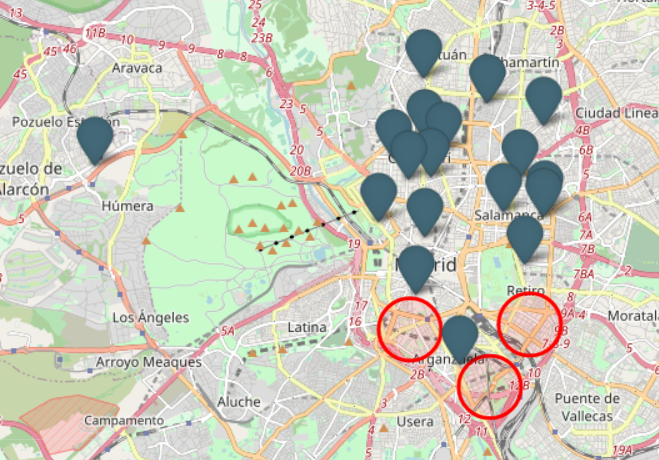
* **First dataset** will contain information about all Levaduramadre (existings and possibles future): This dataset will be used to obtain the second dataset.
* **Second dataset** will contain information about the venues around all Levaduramadre: This dataset will be used to create a cluster (using k-means clustering) of Levaduramadre’s locations to get the first criteria and help identify which should be the best location of next Levaduramadre.
* **Third dataset** will contain information about the bakeries around the possible locations: This dataset will be used to get the second criteria and help identify which should be the best location of next Levaduramadre.

### Data assumption

As said before, we assumed that Levaduramadre owner gives us 3 locations to choice the best option to open a new Levaduramadre.

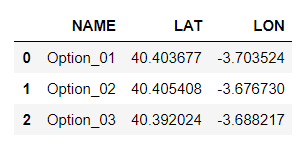
The locations of the possibles future Levaduramadre has been made by the following steps:

1. Search and paint on a map all existing Levaduramadre in Madrid. The search will be done with a call to Foursquare API, using the keyword “Levaduramadre” in the call, to obtain all venues named Levaduramadre in 20 km around Madrid center.
2. Select ([approximately](https://www.spanishdict.com/translate/approximately?langFrom=en)) three zones in the south-center of Madrid without Levaduramadre.



Picture 1.- Location of existings Levaduramadre and zones of possibles Levaduramadre.

1. Choose a random address in those zones, with the only criterion that it will be located on a street. Of the random address we get the latitude and longitude from google map.



Picture 2.- Dataframe with latitude and longitude of possibles future Levaduramadre

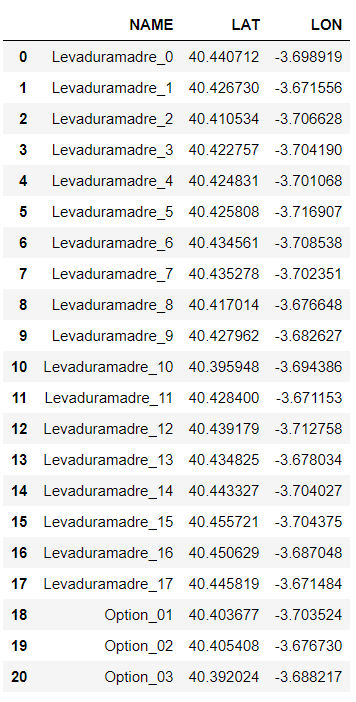
### Data cleaning

**First dataset**

To obtain the second dataset, I need clean the datas of the first dataset, because I obtain all venues named Levaduramadre, include the factory called “Obrador”, and other Levaduramadre that is not located in Madrid, so I deleted the venue with “Obrador” in its name, and filtered the rest by city equal to Madrid.

For easier identification of each venue, I will change the “name” each venue with the concatenated “Levaduramadre\_” plus a numerical value.

The last step for clean this first dataset will be get only the next information: name, latitude and longitude (also I will change the column name to match with the dataframe of possibles Levaduramadre). And then join with the dataframe of possibles future Levaduramadre.



Picture 3.- Example of first dataset

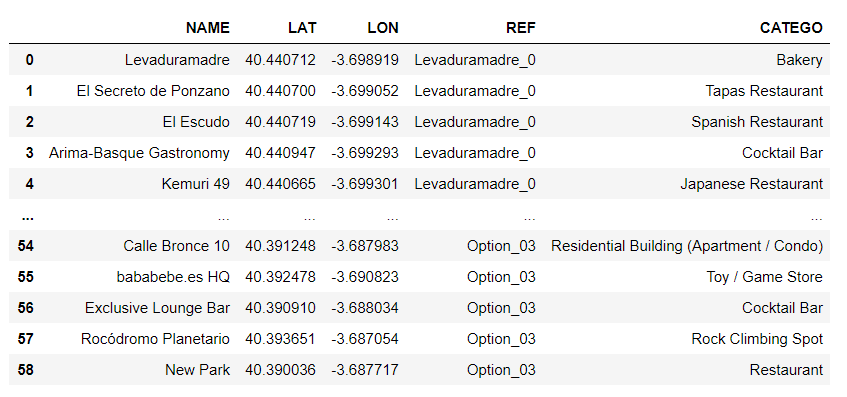
**Second dataset**

With the first dataset we can develop a search (with Foursquare API) and obtain the datas to make the second dataset. Obtaining those data will be made in the next 3 steps:

* Create an axiliar list.
* Make calls to Foursquare API and save the responses in auxiliar list generated: In this call we don´t use keyword, so we will obtein all venues in 200 meters around a point, in this case, each of venue in first dataset.
* Save the auxiliar list in a dataframe (named “df\_foursquare”).

**From the dataframe df\_foursquare** we get the useful information and create the second dataset. The useful information will be: name, latitude, longitude and the venue wich we use of reference in the search (in this case all Levaduramadre).

Then we add the category name to second dataset from the dataframe df\_foursquare.



Picture 4.- Example of second dataset

**Third dataset**

The process to create the third dataset is very similar to the second dataset, but we have to change the parameters of the call to Foursquare API, beacuse we will use a category ID (corresponding to Bakery) to make the calls. The response will be all bakeries around a point, in this case, the three possibles future Levaduramadre.

Once the call is made we proceed to obtain only the useful data: name, distance and the venue wich we use of reference in the search (in this case the possibles future Levaduramadre).



Picture 5.- Example of third dataset

# METHODOLOGY

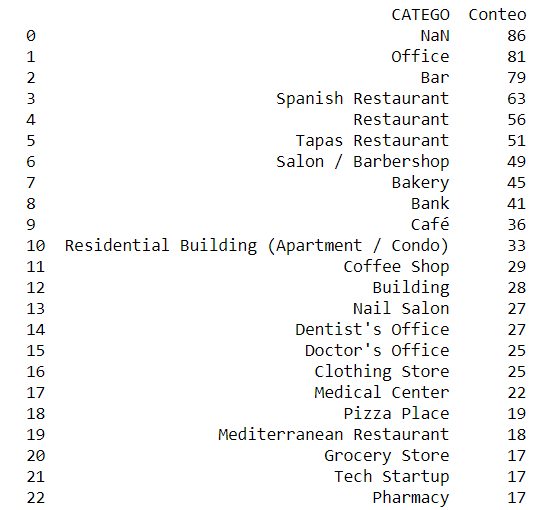
In this project we will calculate two criteria to choose where is the best locations to open a new bakery of the Levaduramadre´s branch.  
In **first step** we have collected and cleaned the required **data** and save it in two dataframes (called second and third dataset). To obtain the second dataset we had to created a first dataset with only Levaduramadre´s data.  
In the **second step** we use the datasets to develop the criteria:

* The second dataset will be used to develop the **first criterion**, which consist in make a clustering of the neighborhoods with a Levaduramadre and the neighborhoods with the possibility to have one. We will make a clustering with 4 cluster and other with 5 cluster (to have more data in the criterion).
* The third dataset will be used to develop the **second criterion**, which consist in make a serie of weighted average with the distances between the possibles future Levaduramadre and the bakeries around them in a 250 meters of ratio.

In a **third step** we develop an analysis of the criteria´s result and finllay we will make a **conclusion**.

### First criterion

To develop we will make a count of the categories in second dataset.



Picture 6.- Example of category count.

Then we will avoid the venues with the value 'NaN' in category, and only select the venues wich the categories have a count upper or equal than 15. With this process we can have relevant data to make the model of clustering.



Picture 7.- Data for clustering.

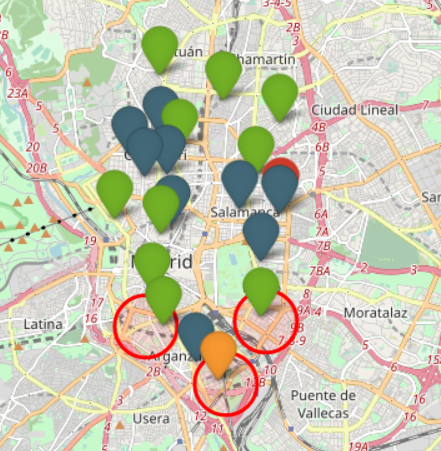
Now we will grouping of the datas by Levaduramadre (existing and possibles future), making the mean (normalized) about how many repetitions have a category for each Levaduramadre.



Picture 8.- Data clean for clustering.

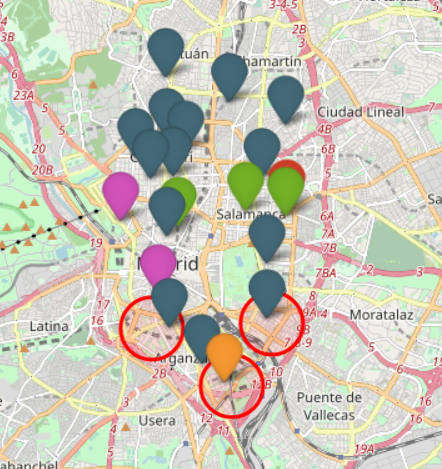
Then we develop two model (using k-means clustering), one with 4 clusters and other with 5 clusters.

The result of the model with 4 clusters is represent in the map below:



Picture 9.- Representation of result model with 4 clusters.

The result of the model with 5 clusters is represent in the map below:



Picture 10.- Representation of result model with 5 clusters.

### Second criterion

How the Option\_03 don´t have any bakery around (because there aren´t any data of Option\_03 in third dataset), we will do the calculations for Option\_01 and Option\_02.

First we make the average with the distances, obtain the following results:



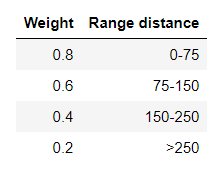
Picture 11.- Result of average calculation.

Then we are going to develop 3 cases, in each case we are going to calculate the weighted average with differents weights depending on the distance.

**CASE 01**

For this case we suposse the weights in the ranges of the shortest distances.

The weights considered are below:



Picture 12.- Weights of CASE 01.

And the results for CASE 01 are:

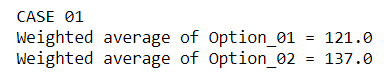
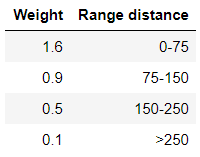


Ilustración 13.- Results of CASE 01.

**CASE 02**

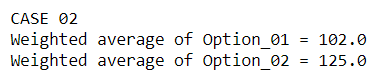
Now we will increase the weights in the ranges of the shortest distances. For this case we will double the first weight, we will sum the half of the own value to the second weight, and sum a quarter of the own value to the third weight. The last weight will be the half.

The weights considered are below:



Picture 14.- Weights of CASE 02.

And the results for CASE 02 are:

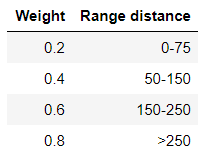


Picture 15.- Results of CASE 02.

**CASE 03**

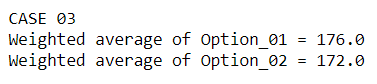
For this case we suposse the weights in the ranges of the longer distances.

The weights considered are below:



Picture 16.- Weights of CASE 03.

And the results for CASE 03 are:



Picture 17.- Results of CASE 03.

# DISCUSSION

### Discussion first criterion

As we can see from the results of the first criterion, Option\_01 and Option\_02 are in the most common cluster about Levaduramadre´s location (in both clustering models), so those locations can be considered good for opening a new Levaduramadre.

But Option\_03 has no similarity to the other Levaduramadre´s neighborhoods (in any of the clustering models), so it is easy to say that this location is not recommended to open a Levaduramadre, at least not without a high risk rate.

### Discussion second criterion

We only discuss about Option\_01 and Option\_02, because the Option\_03 don´t have any bakery around, which can tell us two things, that you can open a bakery without competition or is not the best place to open a bakery (I think more in the second).

So when we see the results of the average, the difference is 5 meters, so is very similar; but when we put a weight depending on the distance (when more shortest distance more weight) the Option\_01 decrease the average (in 29 meters) more than Option\_02 (in 18 meters) with a difference between them of 16 meters, and when we increase the weight to the shortest distance, the difference between them up to 23 meters, so we can say that the short distance between a future Levaduramadre and others bakeries affect more in Option\_01. But when we put weight depending on the longer distance, the difference between the two options is 4 meters, so the difference is simply maintained and we can conclude that bakeries with more distance affect less to the possibles future Levaduramadre.

# CONCLUSION

The aim of this project is select the best location to open a new Levaduramadre bakery. For that we have to chooce one between three options, and for that we considered two criteria.

With the first criterion we discard the Option\_03, due to the reason given in the Discussion part.

And due to the second criterion we consider Option\_02 as the best option, since the Option\_01 is most affected by the shortest distance, which is interpreted as having more competition.

Finally **our recommendation** is to open the next Levaduramadre in the **Option\_02 location**.