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.

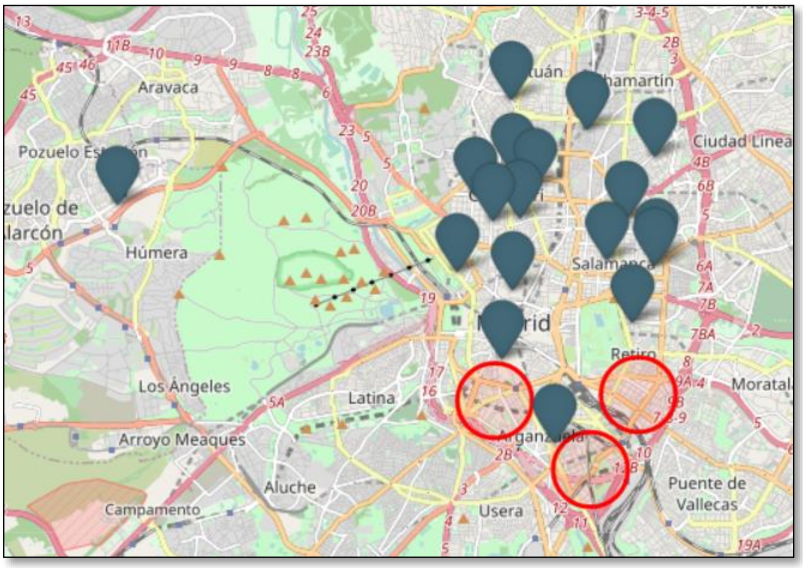
- **Third dataset** will contain information about the bakeries around the possible locations: This dataset will be used to get the second criteria and help identify which should be the best location of next Levaduramadre.

Data assumption

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

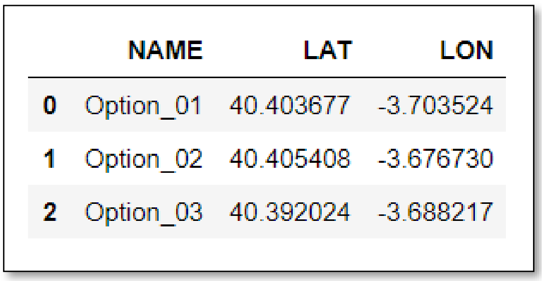
The locations of the possibles future Levaduramadre has been made by the following steps:

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) three zones in the south-center of Madrid without Levaduramadre.



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

3. 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.



	NAME	LAT	LON
0	Option_01	40.403677	-3.703524
1	Option_02	40.405408	-3.676730
2	Option_03	40.392024	-3.688217

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

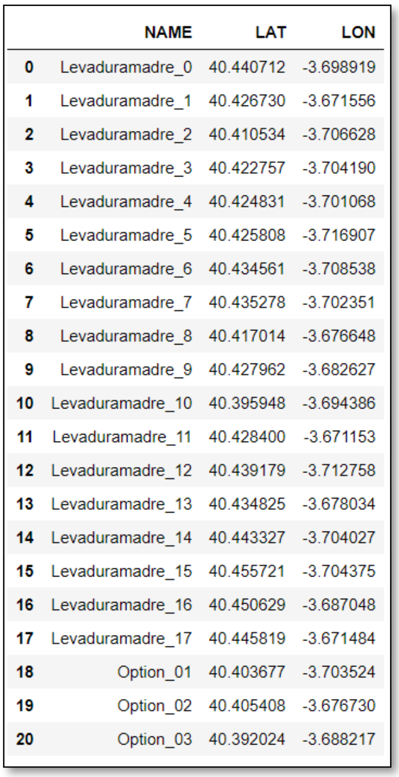
Data cleaning

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.



	NAME	LAT	LON
0	Levaduramadre_0	40.440712	-3.698919
1	Levaduramadre_1	40.426730	-3.671556
2	Levaduramadre_2	40.410534	-3.706628
3	Levaduramadre_3	40.422757	-3.704190
4	Levaduramadre_4	40.424831	-3.701068
5	Levaduramadre_5	40.425808	-3.716907
6	Levaduramadre_6	40.434561	-3.708538
7	Levaduramadre_7	40.435278	-3.702351
8	Levaduramadre_8	40.417014	-3.676648
9	Levaduramadre_9	40.427962	-3.682627
10	Levaduramadre_10	40.395948	-3.694386
11	Levaduramadre_11	40.428400	-3.671153
12	Levaduramadre_12	40.439179	-3.712758
13	Levaduramadre_13	40.434825	-3.678034
14	Levaduramadre_14	40.443327	-3.704027
15	Levaduramadre_15	40.455721	-3.704375
16	Levaduramadre_16	40.450629	-3.687048
17	Levaduramadre_17	40.445819	-3.671484
18	Option_01	40.403677	-3.703524
19	Option_02	40.405408	-3.676730
20	Option_03	40.392024	-3.688217

Picture 3.- Example of first dataset

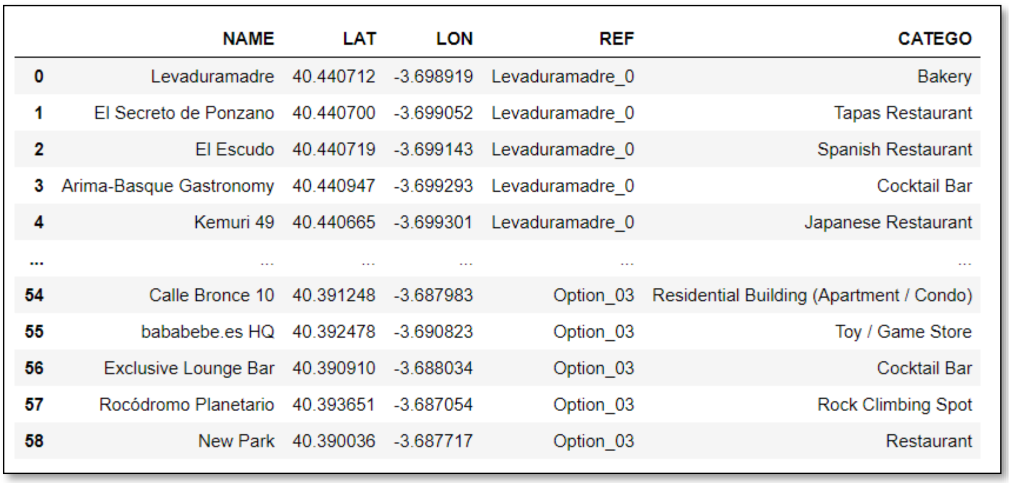
Second dataset

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.



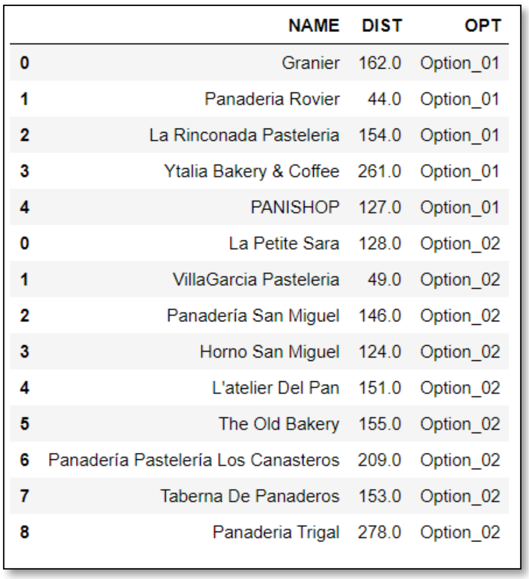
	NAME	LAT	LON	REF	CATEGO
0	Levaduramadre	40.440712	-3.698919	Levaduramadre_0	Bakery
1	El Secreto de Ponzano	40.440700	-3.699052	Levaduramadre_0	Tapas Restaurant
2	El Escudo	40.440719	-3.699143	Levaduramadre_0	Spanish Restaurant
3	Arima-Basque Gastronomy	40.440947	-3.699293	Levaduramadre_0	Cocktail Bar
4	Kemuri 49	40.440665	-3.699301	Levaduramadre_0	Japanese Restaurant
...	...	...	...	...	...
54	Calle Bronce 10	40.391248	-3.687983	Option_03	Residential Building (Apartment / Condo)
55	bababebe.es HQ	40.392478	-3.690823	Option_03	Toy / Game Store
56	Exclusive Lounge Bar	40.390910	-3.688034	Option_03	Cocktail Bar
57	Rocódromo Planetario	40.393651	-3.687054	Option_03	Rock Climbing Spot
58	New Park	40.390036	-3.687717	Option_03	Restaurant

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).



	NAME	DIST	OPT
0	Granier	162.0	Option_01
1	Panaderia Rovier	44.0	Option_01
2	La Rinconada Pasteleria	154.0	Option_01
3	Ytalia Bakery & Coffee	261.0	Option_01
4	PANISHOP	127.0	Option_01
0	La Petite Sara	128.0	Option_02
1	VillaGarcia Pasteleria	49.0	Option_02
2	Panadería San Miguel	146.0	Option_02
3	Horno San Miguel	124.0	Option_02
4	L'atelier Del Pan	151.0	Option_02
5	The Old Bakery	155.0	Option_02
6	Panadería Pastelería Los Canasteros	209.0	Option_02
7	Taberna De Panaderos	153.0	Option_02
8	Panaderia Trigal	278.0	Option_02

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.

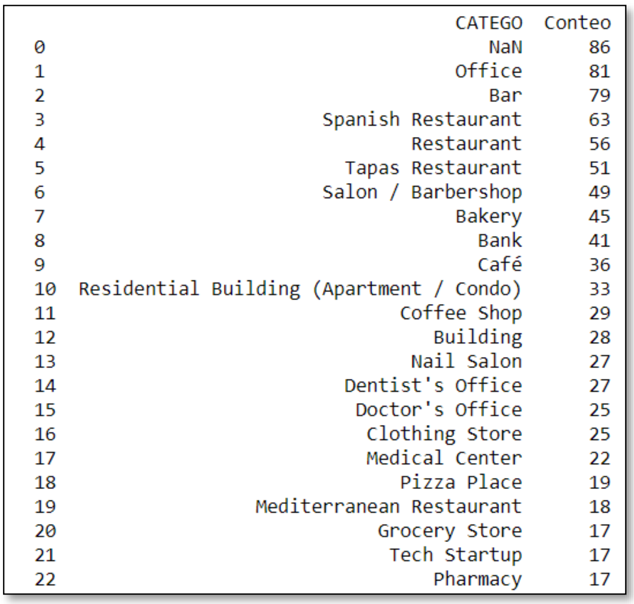
In the **second step** we use the datasets to develop the criteria:

- 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

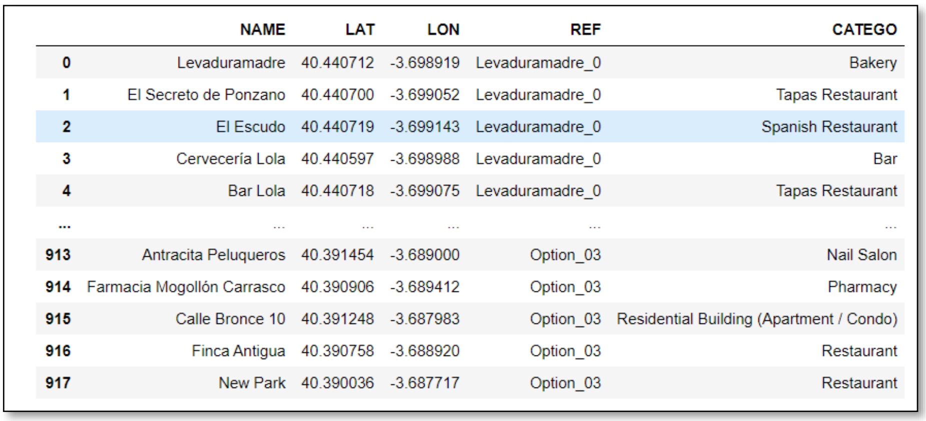
To develop we will make a count of the categories in second dataset.



```
                                     CATEGO  Conteo
0                                       NaN      86
1                                    Office      81
2                                       Bar      79
3                        Spanish Restaurant      63
4                                Restaurant      56
5                          Tapas Restaurant      51
6                        Salon / Barbershop      49
7                                    Bakery      45
8                                      Bank      41
9                                      Café      36
10  Residential Building (Apartment / Condo)     33
11                              Coffee Shop      29
12                                 Building      28
13                               Nail Salon      27
14                          Dentist's Office     27
15                           Doctor's Office     25
16                           Clothing Store      25
17                           Medical Center      22
18                              Pizza Place      19
19                 Mediterranean Restaurant      18
20                            Grocery Store      17
21                             Tech Startup      17
22                                 Pharmacy      17
```

Picture 6.- Example of category count.

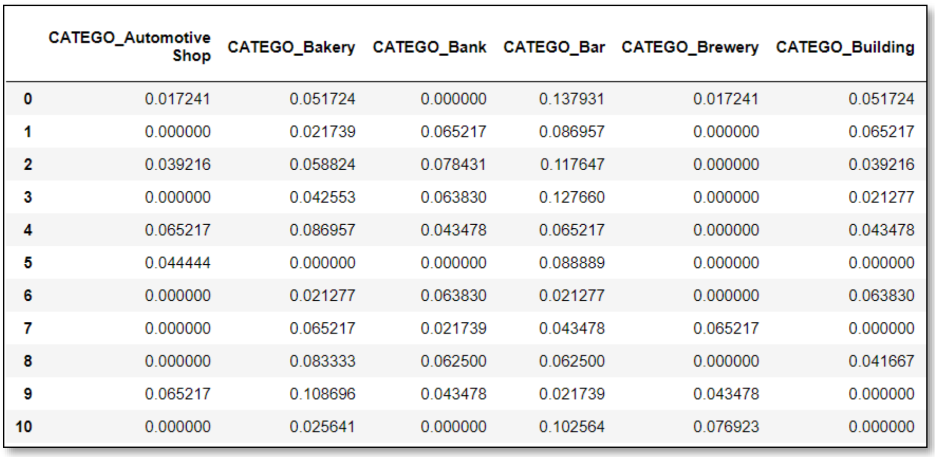
Then we will avoid the venues with the value 'NaN' in category, and only select the venues wich the categories have a count upper or equal than 15. With this process we can have relevant data to make the model of clustering.



	NAME	LAT	LON	REF	CATEGO
0	Levaduramadre	40.440712	-3.698919	Levaduramadre_0	Bakery
1	El Secreto de Ponzano	40.440700	-3.699052	Levaduramadre_0	Tapas Restaurant
2	El Escudo	40.440719	-3.699143	Levaduramadre_0	Spanish Restaurant
3	Cervecería Lola	40.440597	-3.698988	Levaduramadre_0	Bar
4	Bar Lola	40.440718	-3.699075	Levaduramadre_0	Tapas Restaurant
...	...	...	...	...	...
913	Antracita Peluqueros	40.391454	-3.689000	Option_03	Nail Salon
914	Farmacia Mogollón Carrasco	40.390906	-3.689412	Option_03	Pharmacy
915	Calle Bronce 10	40.391248	-3.687983	Option_03	Residential Building (Apartment / Condo)
916	Finca Antigua	40.390758	-3.688920	Option_03	Restaurant
917	New Park	40.390036	-3.687717	Option_03	Restaurant

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.

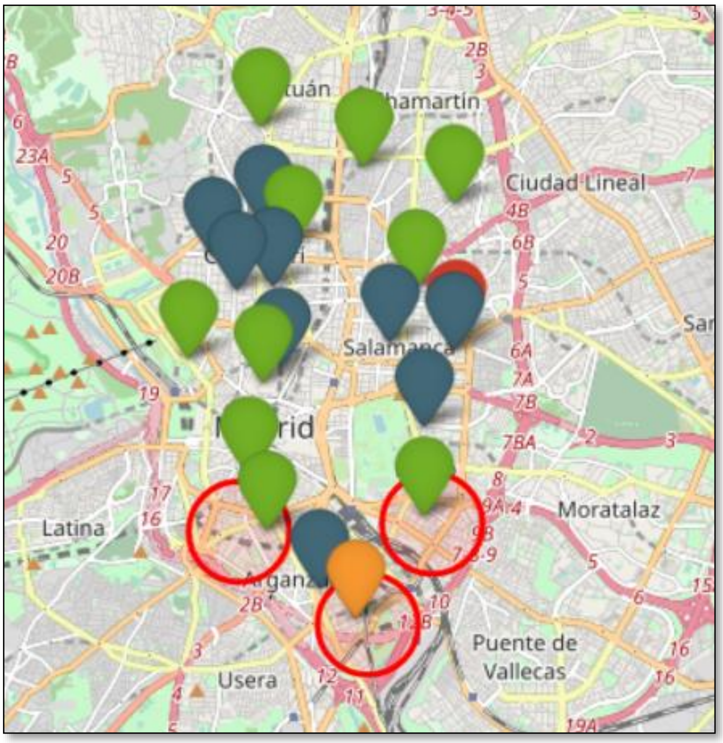


	CATEGO_Automotive Shop	CATEGO_Bakery	CATEGO_Bank	CATEGO_Bar	CATEGO_Brewery	CATEGO_Building
0	0.017241	0.051724	0.000000	0.137931	0.017241	0.051724
1	0.000000	0.021739	0.065217	0.086957	0.000000	0.065217
2	0.039216	0.058824	0.078431	0.117647	0.000000	0.039216
3	0.000000	0.042553	0.063830	0.127660	0.000000	0.021277
4	0.065217	0.086957	0.043478	0.065217	0.000000	0.043478
5	0.044444	0.000000	0.000000	0.088889	0.000000	0.000000
6	0.000000	0.021277	0.063830	0.021277	0.000000	0.063830
7	0.000000	0.065217	0.021739	0.043478	0.065217	0.000000
8	0.000000	0.083333	0.062500	0.062500	0.000000	0.041667
9	0.065217	0.108696	0.043478	0.021739	0.043478	0.000000
10	0.000000	0.025641	0.000000	0.102564	0.076923	0.000000

Picture 8.- Data clean for clustering.

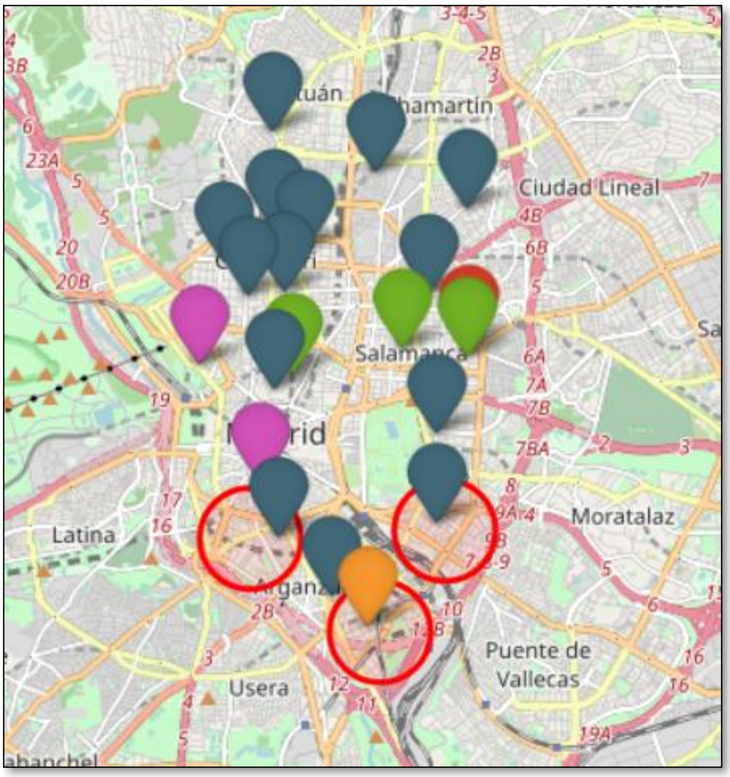
Then we develop two model (using k-means clustering), one with 4 clusters and other with 5 clusters.

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:

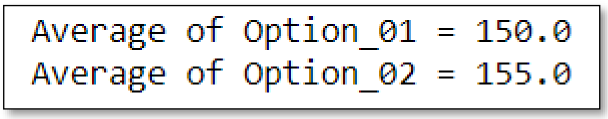


Picture 10.- Representation of result model with 5 clusters.

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:



```
Average of Option_01 = 150.0
Average of Option_02 = 155.0
```

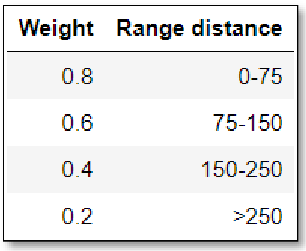
Picture 11.- Result of average calculation.

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.

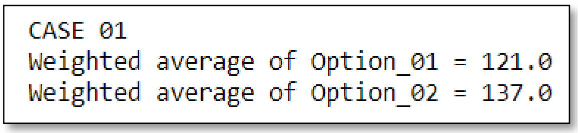
The weights considered are below:



Weight	Range distance
0.8	0-75
0.6	75-150
0.4	150-250
0.2	>250

Picture 12.- Weights of CASE 01.

And the results for CASE 01 are:



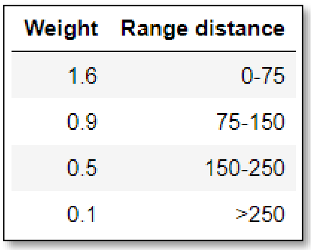
```
CASE 01
Weighted average of Option_01 = 121.0
Weighted average of Option_02 = 137.0
```

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.

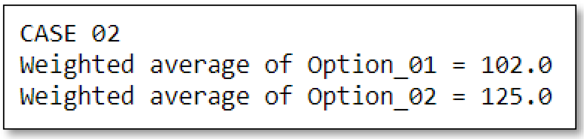
The weights considered are below:



Weight	Range distance
1.6	0-75
0.9	75-150
0.5	150-250
0.1	>250

Picture 14.- Weights of CASE 02.

And the results for CASE 02 are:



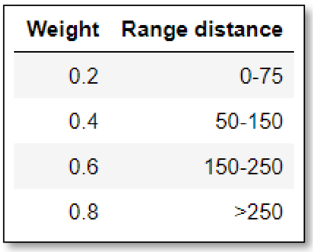
```
CASE 02
Weighted average of Option_01 = 102.0
Weighted average of Option_02 = 125.0
```

Picture 15.- Results of CASE 02.

CASE 03

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

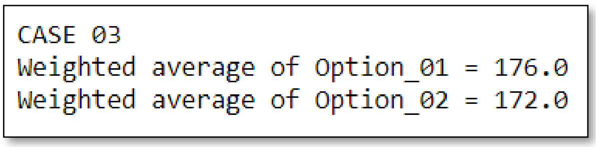
The weights considered are below:



Weight	Range distance
0.2	0-75
0.4	50-150
0.6	150-250
0.8	>250

Picture 16.- Weights of CASE 03.

And the results for CASE 03 are:



```
CASE 03
Weighted average of Option_01 = 176.0
Weighted average of Option_02 = 172.0
```

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**.