BARCELONA CASE STUDY
Local coffee shop vs gourmet restaurant

Jaume Torres Bonet

January 9th, 2020

1. Introduction
1.1. Background

Barcelona is the second largest city in Spain with a population of around 1.6 million, situated in the coast of the autonomous community of Catalonia. This also makes it the fifth most populous urban area in Europe and a major economic center of southwestern Europe. It is also considered an extremely popular tourist destination, ranking high in number of international visitors year after year.

1.2. Problem

One of the first steps when opening a business is choose the best location to maximize profit. In this case study, we take on the job to find the most feasible location to open a local where food is served. Moreover, we will distinguish between two cases: opening a local coffee shop and opening a more exclusive gourmet restaurant. Barcelona is chosen in this example since it is the city I am currently residing in and the one I am more familiar with in the present.

To solve this problem, we will assume that the best location for a new business is usually where other similar businesses are, thus the location of most restaurants in Barcelona will be retrieved. Since we want to distinguish between a more local shop and a more exclusive one, we will also use the data regarding rent per capita, as well as the location of hotels to know where most of the tourists are staying, since they also contribute a lot to the economy of the city.

2. Data
2.1. Data sources

Different data sources will be used in this study. These include:

- Barcelona city hall data: It is the base of the study, includes a list of districts, neighborhoods and rent per capita.
- Wikipedia page for Barcelona districts: Used to get the names of the districts.
- Foursquare API: Used to get the data on the venues of each neighborhood.

3. Methodology
3.1. Data obtention

First of all, one should obtain a list of districts and neighborhoods of Barcelona, to then obtain further locations based on those. To do that, the website of Barcelona City Hall is scrapped and processed to obtain a dataframe with the base information. This was first chosen since this same website included population and rent per capita data as well, thus simplifying the job. The only thing left to do was to obtain the names of the districts, which

were scrapped from Wikipedia's page of Barcelona Districts. Once the data was obtained, numeric data was properly transformed from string to integer or float.

	District	District_name	Neigh_id	Neigh_name	Population	RFD
0	1	Ciutat Vella	1	el Raval	47986	71.2
1	1	Ciutat Vella	2	el Barri Gòtic	16240	106.1
2	1	Ciutat Vella	3	la Barceloneta	15101	79.6
3	1	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera	22923	99.4
4	2	Eixample	5	el Fort Pienc	32048	106.5

Fig. 3.1. Base dataframe with data on Barcelona's Neighborhoods.

The dataframe included the below columns:

- District id: Integer from 1 to 10.
- District name
- Neighborhood id: Integer from 1 to 73.
- Population: Number of inhabitants (2017).
- RFD: Index used to evaluate the rent per capita of the neighborhood, considering a RFD = 100 is the average in the city, thus anything above 100 is considered more affluent and vice versa. Data from 2017.

Then, we can use the *geopy* library to obtain the latitude and longitude of these neighborhoods. To do that we just create an "Address" column by concatenating "Barcelona" to the name of the neighborhoods so there are no confusions and look up that location. Once that data is obtained, the address column can be dropped.

	District	District_name	Neigh_id	Neigh_name	Population	RFD	Address	Latitude	Longitude
0	1	Ciutat Vella	1	el Raval	47986	71.2	el Raval, Barcelona	41.379518	2.168368
1	1	Ciutat Vella	2	el Barri Gòtic	16240	106.1	el Barri Gòtic, Barcelona	41.383395	2.176912
2	1	Ciutat Vella	3	la Barceloneta	15101	79.6	la Barceloneta, Barcelona	41.380653	2.189927
3	1	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera	22923	99.4	Sant Pere, Santa Caterina i la Ribera, Barcelona	41.388322	2.177411
4	2	Eixample	5	el Fort Pienc	32048	106.5	el Fort Pienc, Barcelona	41.395925	2.182325

Fig. 3.2. Base dataframe with data on Barcelona's Neighborhoods and obtained locations.

Once the latitude and longitude of each neighborhood is obtained, we can proceed to get the nearby venues using the Foursquare API, limiting the radius to 500m and a maximum of 100 venues to be retrieved.

Contrary to past assignments in the course, instead of looking for every type of venue, we will exclusively look for venues with the category "Restaurant" and with the category "Hotel". This is done to avoid differentiating too much between venues, since for instance, we are equally considering "Tapas Restaurants" and "Spanish Restaurants", and looking for every type of venue would differentiate between the two.

	Neigh_name	Restaurants
0	Baró de Viver	32
1	Can Baró	50
2	Can Peguera	47
3	Canyelles	15
4	Ciutat Meridiana	12
...	...	...
66	la Vila Olímpica del Poblenou	49
67	la Vila de Gràcia	50
68	les Corts	50
69	les Roquetes	45
70	les Tres Torres	50

	Neigh_name	Hotels
0	Can Baró	5
1	Can Peguera	2
2	Diagonal Mar i el Front Marítim del Poblenou	18
3	Horta	3
4	Hostafrancs	49
...	...	...
59	la Vila Olímpica del Poblenou	26
60	la Vila de Gràcia	43
61	les Corts	23
62	les Roquetes	2
63	les Tres Torres	8

Fig. 3.3 Obtained restaurant and hotels for each of the neighborhoods.

With these parameters in mind, 3.234 restaurants were retrieved, containing 122 unique subcategories. On the other hand, 1.176 hotels were retrieved, containing 17 unique subcategories. There were two neighborhoods that did not have any restaurant, whereas nine did not have any venue classified as hotel.

3.2. Data exploration

Before proceeding with the clustering part of the problem, it is wise to carry out some data analysis to see how our data is distributed and be sure there are no missing or erroneous values. Since there are a total of 73 neighborhoods, we will group our data in the 10 existing districts to make it easier to plot and arrive to conclusions. In this step, we will also express the population in thousands to make things clearer. The RFD has also been average weighted with the population of each neighborhood when the grouping was carried out.

	District	District_name	Population	RFD	Restaurants	Hotels
0	1	Ciutat Vella	102.250	84.305653	200	193
1	2	Eixample	267.184	122.445749	300	280
2	3	Sants-Montjuïc	182.354	84.592545	392	179
3	4	Les Corts	82.201	137.294018	146	41
4	5	Sarrià - Sant Gervasi	149.734	182.792187	252	100
5	6	Gràcia	121.566	105.342970	198	104
6	7	Horta-Guinardó	169.187	77.972005	515	69
7	8	Nou Barris	166.805	55.011607	469	24
8	9	Sant Andreu	147.693	74.620786	266	19
9	10	Sant Martí	236.163	88.139150	496	167

Fig. 3.4 Grouped data by district for the data exploration.

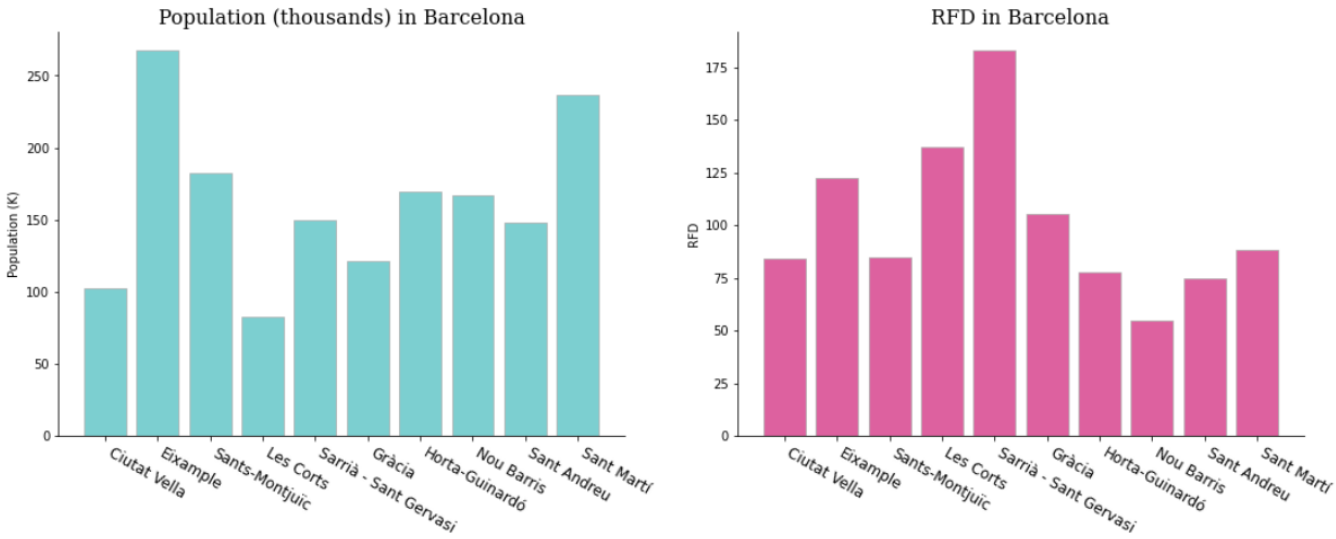

Fig. 3.5. Population by district (left) and RFD by district (right) in the city of Barcelona.

From these two barplots, we can see that the most populated district is Eixample, followed by Sant Martí, whereas the least populated district is Les Corts. On the other hand, the RFD in Barcelona is highest in Sarrià, followed by Les Corts, with the lowest being in Nou Barris. From this figure, we gather that usually the more affluent districts tend to be less populated (more exclusive), whereas most population will be situated in districts with less rent per capita (lower RFD).

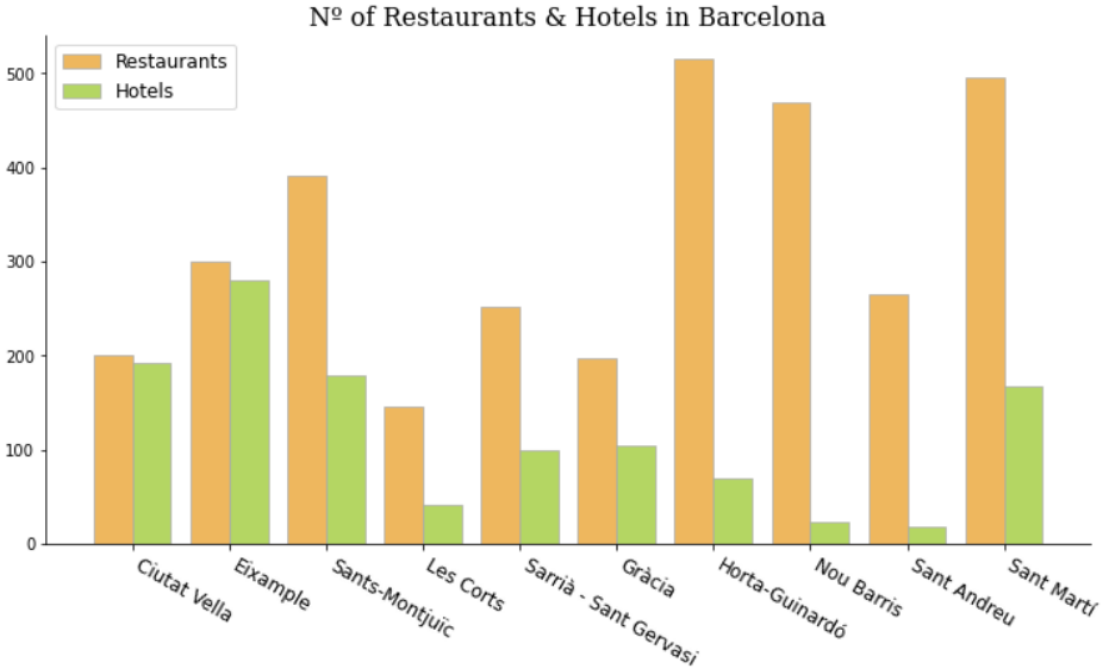

Fig. 3.6 Number of restaurants and hotels in Barcelona by district.

From Figure 3.6, we can see that most hotels concentrate in Eixample and Ciutat Vella, which are actually the most touristic parts of the city. In the restaurant category (this category includes any type of food location, from actual restaurants to coffee shops or fast food chains) we can see that, even though the touristic parts of the city do have a considerable amount of restaurants, most of them are actually situated in the last four districts in the plot (Horta-Guinardó, Nou Barris, Sant Andreu, Sant Martí). These districts

correspond to the districts with lowest RFD, which probably means that most of these locations are in fact small local bars or coffee shops.

Finally, before clustering, we can plot all this data onto a map to have an idea on what are we trying to group. The number of food locals by neighborhood can be qutie similar, with only low values in the neighborhoods outskirts of the city. Thus, in this case we will plot the number of hotels in the city to check where the touristic center is.

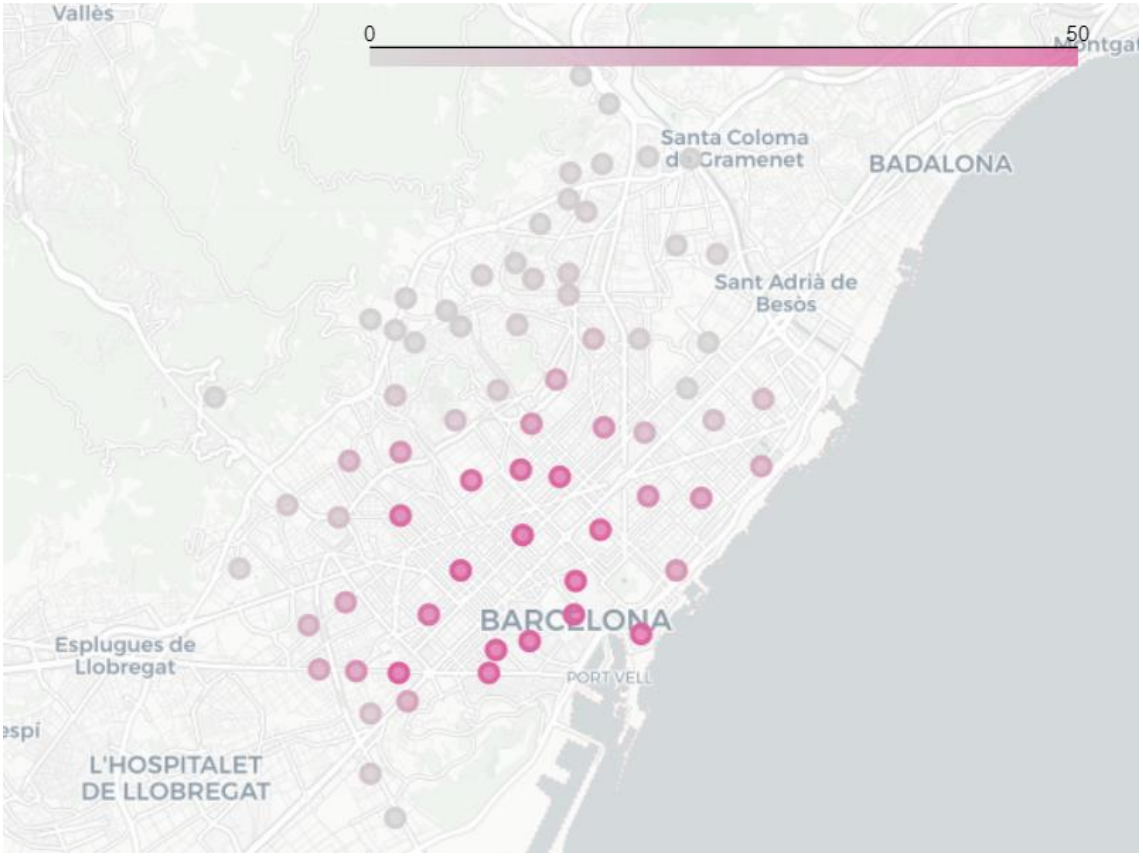

Fig. 3.7 Number of hotels in Barcelona by neighborhood.

3.3. Data transformation

Once the that has been checked for errors, we could proceed with the clustering, but first it has to be manipulated to correctly fit in a K-means model.

Rather than focus on absolute numbers, it might be more interesting to focus on the ratios. For example, we might consider more important the ratio restaurants/population, since there should be quite a difference between counting 100 restaurants per 1k population than 100k population. We can take the same approach with the number of hotels. By doing that, we will be able to drop the population column of our clustering dataframe for analysis. Finally, we will normalize the data to be ready for clustering.

3.4. Clustering

Once the data has been processed, we can apply the K-means model for clustering the neighborhoods, taking into account their RFD, restaurant/habitant ratio and hotel/habitant ratio (Fig. 3.8). To find out which is the optimal K number of clusters, we will carry out the calculation for sever values and plot the distortion and silhouette score, using the elbow method to pin point the optimal number (Fig. 3.9).

	RFD	Restaurants_ratio	Hotels_ratio
0	-0.529779	-0.341951	0.356583
1	0.293019	-0.138446	2.782026
2	-0.331742	-0.115244	3.058551
3	0.135061	-0.228127	1.767288
4	0.302449	-0.290178	0.896225

Fig. 3.8. Data used for the K-means clustering.

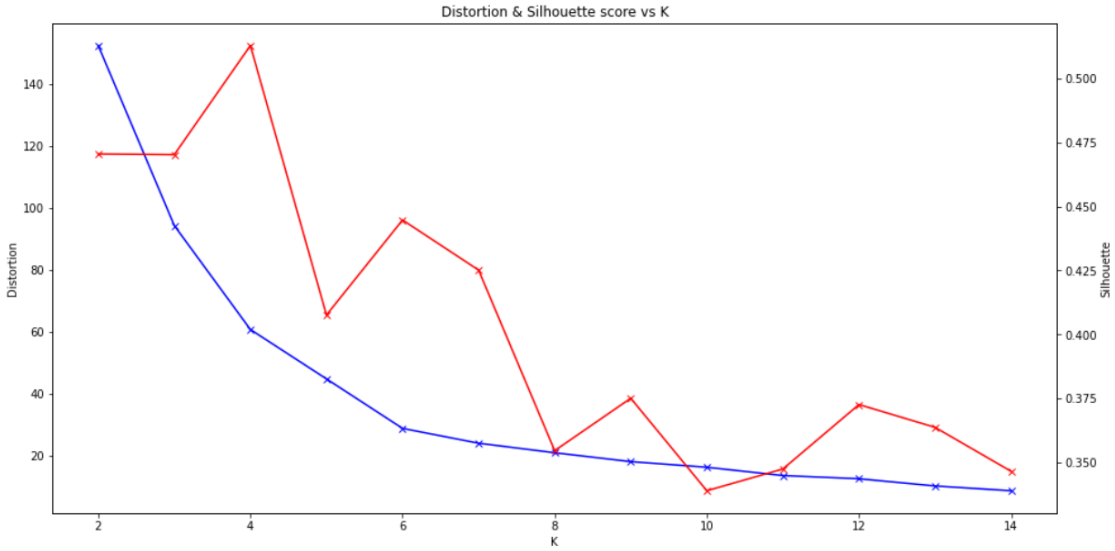

Fig. 3.9 Distortion score (blue) and silhouette score (red) for different values of K clusters.

From Figure 3.9, it is determined that the optimal number of clusters to take into account is k = 4.

4. Results

4.1. Clustering results

Once the clusters have been obtained, we can plot the results onto a map of Barcelona to visually check the results (Fig. 4.2).

We can also compute the features used for clustering for every of the four created clusters (Fig. 4.1), which will give us some hints on what type of areas are there and, more importantly, will provide us with the solution of the problem in hands. These results include the % of population, % of restaurants and % of hotels located in each of the clusters, as well as the weighted RFD index by the number of inhabitants in each cluster.

	Cluster	% Population	RFD	% Restaurants	% Hotels
0	0	8.45	102.92	12.31	26.87
1	1	75.30	83.46	70.90	54.25
2	2	0.11	58.55	2.88	0.26
3	3	16.14	175.58	13.91	18.62

Fig. 4.1 Results obtained for each of the clusters.

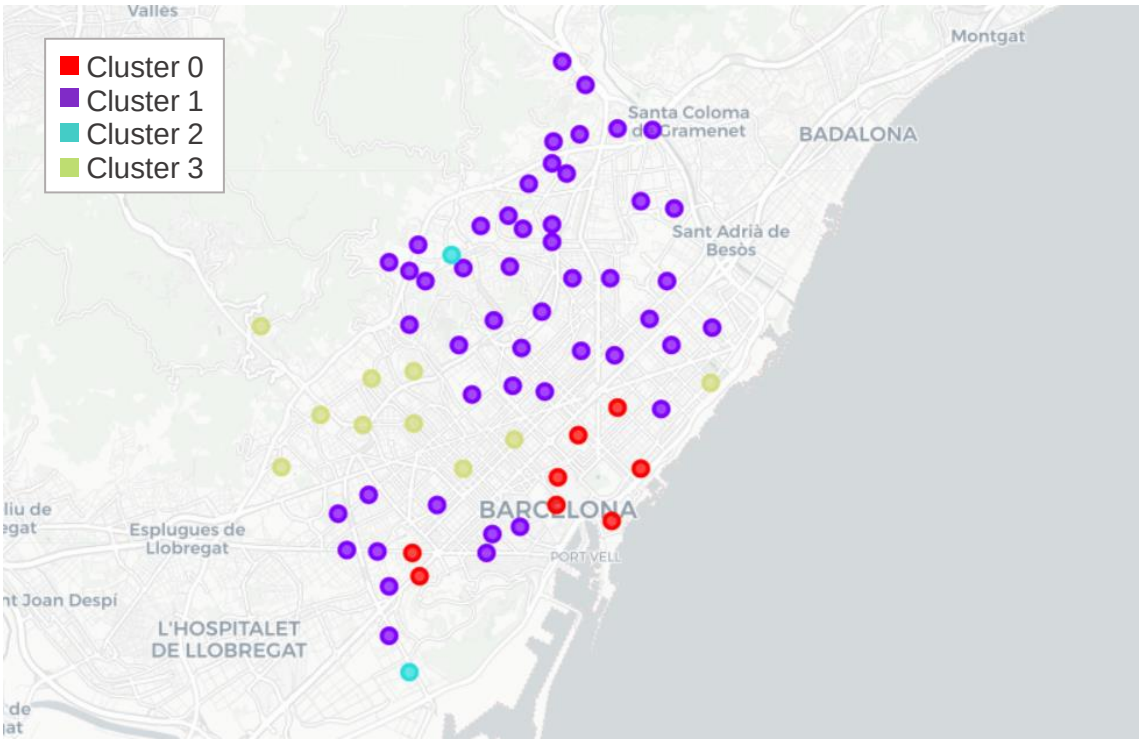

Fig. 4.2 The four obtained clusters in the city of Barcelona.

With this data in hands, we can proceed to identify each of the clusters:

1. **Cluster 0:** Not much population and slightly above average RFD. For the small population it has, there are quite a lot of restaurants and hotels. In the map is represented by the color red and it is mostly situated in the coast and center of Barcelona, most of the touristic places. This is a Touristic cluster.
2. **Cluster 1:** It includes most of the population in Barcelona (purple in the map), with most of its restaurants and hotels, even though the percentage of hotels lags behind compared to the percentage of population it represents (75% population vs 54% of the hotels). Adding to that, the RFD is 83.46 (below average), which implies that most of these neighborhoods are working class. This would be our Average cluster.
3. **Cluster 2:** Only two neighborhoods are represented in this cluster, which has way below average RFD and represent a small percentage of the population. This is our Underdeveloped cluster.
4. **Cluster 3:** It represents 16% of the population and has a higher percentage of hotels than restaurants. The RFD is way above the average (175) which implies these are quite affluent neighborhoods. In fact, we can see in the map that most of them are situated in the upper part of the city, which corresponds to the richer part. This would be our Affluent cluster.

Cluster	Label
0	Touristic
1	Average
2	Underdeveloped
3	Affluent

5. Discussion

Once the results have been obtained, we can now solve our problem: Where is the best location to open a local coffee shop? And a more exclusive gourmet restaurant?

5.1. Local coffee shop

The local coffee shop should be situated in the Average Cluster (1), purple in the map. It has the highest amount of population to serve and highest percentage of restaurants. The rent per capita indicates that it is working class, which would also be suitable for this kind of local.

5.2. Gourmet restaurant

The gourmet restaurant should be situated either in Touristic cluster (0) or Affluent cluster (3), depending on which kind of clientele it caters. If the touristic cluster is chosen, it should not be an extremely exclusive restaurant, serving typical spanish food, which would benefit from the higher number of hotels in the zone. If the affluent cluster is chosen, I think you could go all-in in the exclusiveness of the local.

6. Conclusion

In this project, data from different sources has been analyzed to determine where to open a local coffee shop and a gourmet restaurant in the city of Barcelona. The results have been obtained by analyzing rent per capita, population, number of restaurants and number of hotels in each of the neighborhoods of the city. Afterwards, this data has been processed and clustered in four clusters.

Analyzing these four clusters, it was determined that a local coffee shop should be open in a cluster determined as average. On the other hand, the gourmet restaurant could be situated in two different clusters, one determined as touristic and the other one as affluent.

Further data could be analyzed to obtain more detailed clusters, with, for example, other types of locals taken into account. Extra processing could be done to the data regarding the rent per capita which is from 2017, to project it to the current year. Other studies could focus on other types of locals or different cities.