Gentrification Map

Martín Ferrer Escobar

29 November 2019

1. Introduction

1.1. Background

During the past two decades the importance of cities has grown in a considerable way, being the center of the human activity. In a few years more than half of global population will live in cities for the first time in history, concentrating the principal economic activities, and for those reasons is important to rethink the way we understand the cities and approach new problems generated by these dynamics in an innovative way.

One of the problems that are having the main European and American cities is gentrification, a phenomenon that was first used by Ruth Glass in 1964 but specially in the past 15 years has become a popular concept. We can define gentrification as: “The process by which a place, especially part of a city, changes from being a poor area to a richer one, with the population being replaced by members of a higher social class”.

1.2. Problem

Gentrification is a complex issue and is difficult to face. The better way to act in front of the problem is to try to prevent it of developing. As we have many examples of cities and neighborhoods that have suffered gentrification, there are some characteristics that can help us to diagnose if certain area has certain symptoms, and act in consequence.

Here is where data plays a vital role solving the problem. Analyzing the actual situation of population, rent prices, typology of the venues, between others can help to prevent the situation to develop.

1.3. Interest

The city halls and public administrations are the principal interested in this methodology, to create documents that help the decision making in the urban planification and development of the measures to protect the citizenship.

2. Data acquisition and cleaning

2.1. Data Sources

As study case we will use the city of Barcelona, Catalonia. So all the information and data sets will be from that city.

There will be used three different data sets during this study. The first data set is one that will be used to have the list of neighborhoods and districts of Barcelona, the second data set have the information of rental prices of each neighborhood for the 2017 year, and the third data set is the information of the venues for each neighborhood and will be extracted from FourSquare.

2.2. Data Preparation

There is some processing that the data need to have in order to be useful for the final purpose. The first complete data set that is needed is one that contains the following information: “District Name”, “Neighborhood Name”, “Neighborhood Code”, “Neighborhood Latitude”, “Neighborhood Longitude”.

To obtain that data set we will use the Immigration registry for 2015 (any other data set containing neighborhood and district names and codes could be used). First, the columns that have information that is not useful for the final data set were dropped: “Year”, “Nationality”, “Name” and “District Code”. After changing the columns titles to English from Catalan we obtain the following data set:

```
(11396, 3)
```

	District	Neighborhood Code	Neighbourhood
0	Ciutat Vella	1	el Raval
1	Ciutat Vella	2	el Barri Gòtic
2	Ciutat Vella	3	la Barceloneta
3	Ciutat Vella	4	Sant Pere, Santa Caterina i la Ribera
4	Eixample	5	el Fort Pienc
5	Eixample	6	la Sagrada Família
6	Eixample	7	la Dreta de l'Eixample
7	Eixample	8	l'Antiga Esquerra de l'Eixample
8	Eixample	9	la Nova Esquerra de l'Eixample
9	Eixample	10	Sant Antoni
10	Sants-Montjuïc	11	el Poble Sec

Fig. 1. Data Set of neighborhoods and districts

As you can see the data set has 11396 rows, because at the start the information was the number of immigrants from all nationalities that each neighborhood had, so we had to delete all the rows that name of the name of the neighborhood repeated. After this process the final data set has 74 rows and 3 columns.

Finally, for this data set it was needed to obtain all the coordinates for each row, so the “geopy” library was selected. With a loop every name of the neighborhoods were introduced and a data set containing the X, Y coordinates was obtained. There was a small issue with this function and

is that it didn't recognized the name of three of the neighborhoods, so these were deleted from the data set as we couldn't work with them.

	District	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude
0	Ciutat Vella	el Raval	41.379518	2.168368
1	Ciutat Vella	el Barri Gòtic	41.383395	2.176912
2	Ciutat Vella	la Barceloneta	41.380653	2.189927
4	Eixample	el Fort Pienc	41.395925	2.182325
5	Eixample	la Sagrada Família	41.403479	2.174410

Fig. 2. Data set of the neighborhoods with their respective coordinates

The second data set that we needed to clean was the rental prices data set. The final data set that we need is one that only has two columns, "Neighborhood" and "Price (€/month)".

As in the first data set, the columns that are not needed are dropped: "Year"," Trimester"," District Code"," District Name". As this data is grouped by trimesters, there should be four rows for each neighborhood,292, but instead we find that there are 584. Making a quick exploration of the date we find that there is information of the average rent for flat and for m2. As we only want the information for average rent for flat, we should drop the rows containing the information per m2. For this action we use the "Lloguer Mitja" (Average Rent) column, and after dropping the unwanted values, the column is dropped too.

Next, the rent prices are grouped by neighborhoods applying the mean to have the average price for the whole year. And finally, all the prices are transformed from float numbers to integers, that is the format needed for the map visualization later.

	Neighbourhood	Price (€/month)
1	Can Baró	684
2	Can Peguera	407
3	Canyelles	681
4	Ciutat Meridiana	435
5	Diagonal Mar i el Front Marítim del Poblenou	1109

Fig. 3. Rent prices data set

For the third and last data set, FourSquare application is needed. FourSquare have a complete information from most of the venues in cities so it will be useful for determining the most popular type of venue for each neighborhood.

To obtain this data set, two functions must be defined, one that returns the category of each venue, and the other that returns a data frame of venues which contains coordinates, category, neighborhood and neighborhood coordinates.

(2898, 7)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	el Raval	41.379518	2.168368	Cera 23	41.378947	2.166180	Spanish Restaurant
1	el Raval	41.379518	2.168368	Arume	41.378953	2.166008	Spanish Restaurant
2	el Raval	41.379518	2.168368	Chulapio	41.379264	2.165905	Cocktail Bar
3	el Raval	41.379518	2.168368	A Tu Bola	41.380096	2.169054	Tapas Restaurant
4	el Raval	41.379518	2.168368	La Monroe	41.378795	2.170692	Spanish Restaurant

Fig. 4. Venues data set

3. Exploration of the Data

3.1. Rent Prices data set Exploration

Working with the first data set we can observe where are the neighborhoods with higher rent prices and lower rent prices, and the differences between them.

First, we need to know what magnitude the gap of rent prices between the extreme values of the data frame is. To accomplish this objective two data frames will be created, one with the top 10 values and other with bottom 10 values.

	Neighbourhood	Price (€/month)
10	Pedralbes	1785
72	les Tres Torres	1627
22	Sarrià	1353
16	Sant Gervasi - Galvany	1312
68	la Vila Olímpica del Poblenou	1248
17	Sant Gervasi - la Bonanova	1235
26	Vallvidrera, el Tibidabo i les Planes	1231
51	la Dreta de l'Eixample	1193
5	Diagonal Mar i el Front Marítim del Poblenou	1109
47	l'Antiga Esquerra de l'Eixample	1056

	Neighbourhood	Price (€/month)
24	Vallbona	302
2	Can Peguera	407
4	Ciutat Meridiana	435
23	Torre Baró	443
65	la Trinitat Vella	500
71	les Roquetes	534
64	la Trinitat Nova	540
27	Verdun	592
35	el Carmel	603
31	el Besòs i el Maresme	610

There are huge differences between extreme values of the rent prices, as the more expensive is nearly 6 times expensive than the one with the lower price. But as we approach the middle part of the tables the differences reduce drastically, as the cheaper of the top 10 is not even the double as the more expensive neighborhood of the bottom 10.

Studying the statistical values we can confirm that the majority of the values concentrate in the region between 600€ and 900€, meaning that the top 10 values represent less that 25%.

Fig. 5. Box plot of rent prices in Barcelona

Finally, it will be useful to determine if the more expensive and cheap neighborhoods are grouped in a certain geographical zone. To do that we will plot two maps, representing each of the two data frames.

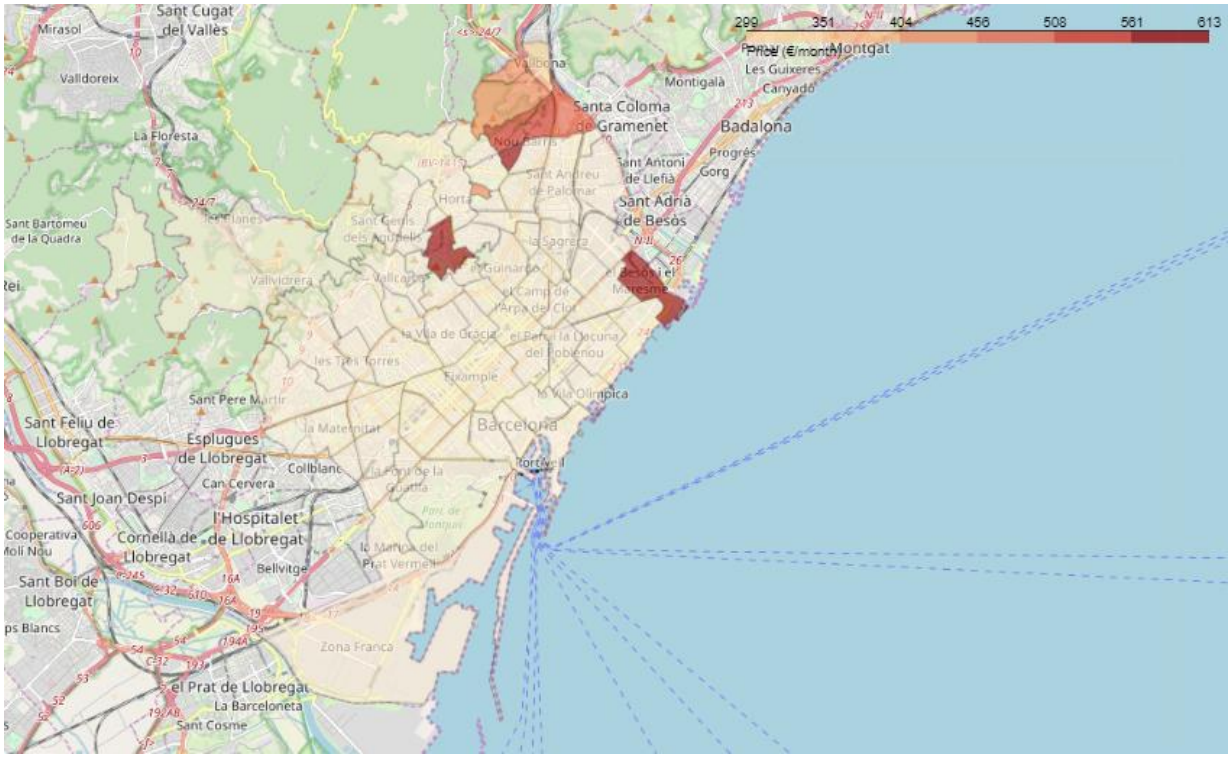

Fig. 6. Less expensive neighborhoods

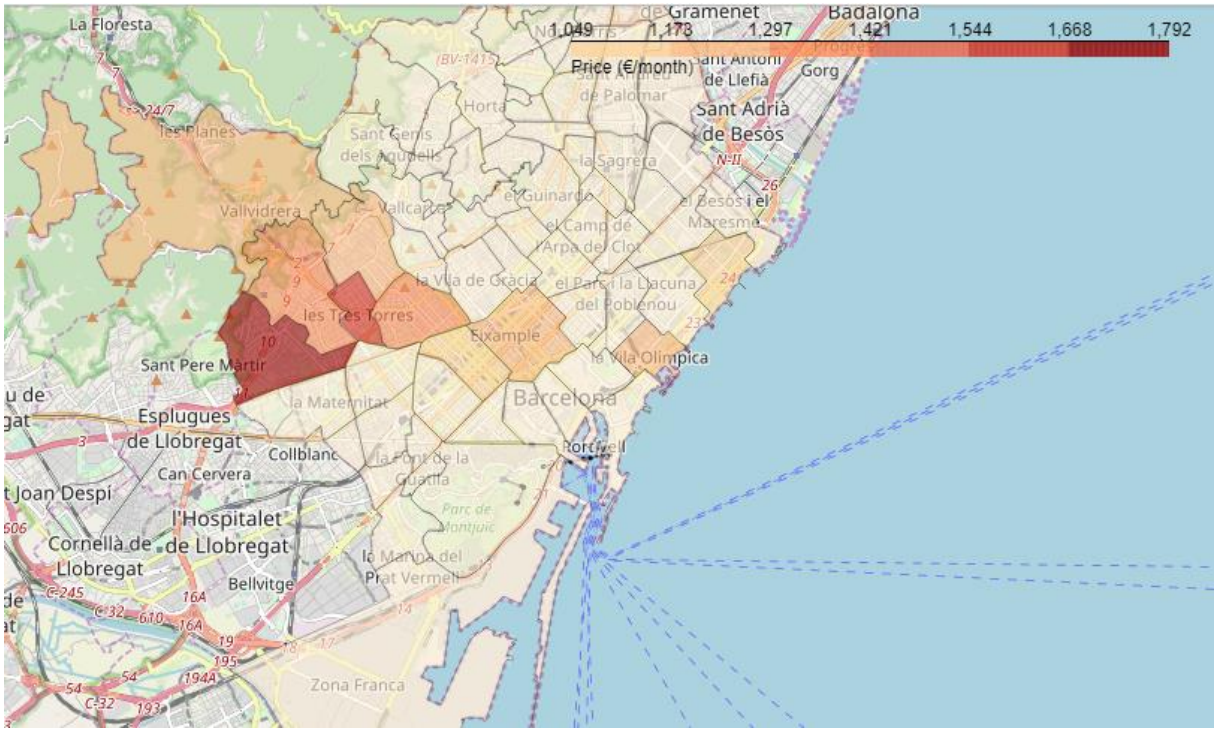

Fig. 7. Most expensive neighborhoods

There is a clear relationship between the geographic space and rent prices as the two groups can be identified in zone of the city. The neighborhoods with higher rents are placed in the west of the city and near the mountain, meanwhile the least expensive neighborhoods are in the north and near the peripheric towns.

3.2. Venues Data set Exploration

In the next chapter we will define which type of venues are at each neighborhood using the clustering method, but before it will be useful to know which are the most popular kind in order to understand the economic network of the city.

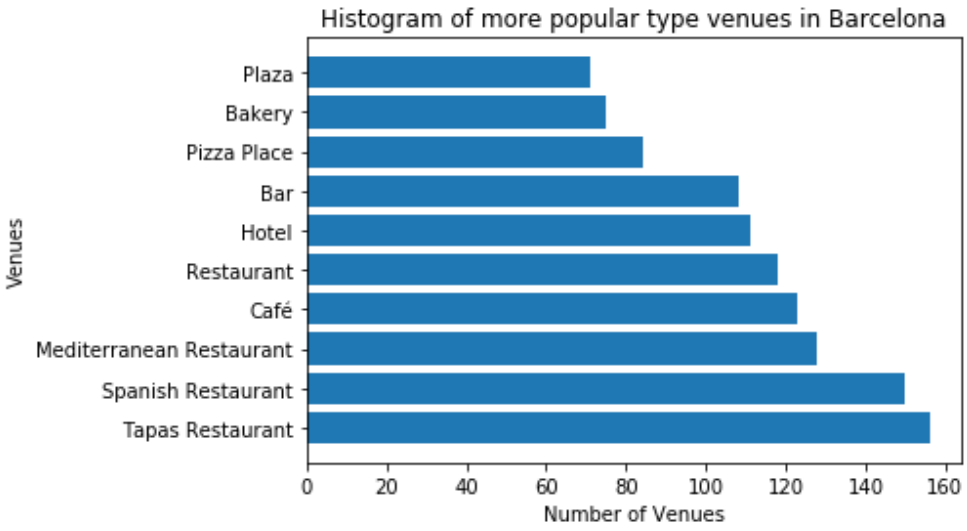

Fig. 8. Histogram of more popular venues in Barcelona

As we can see in the histogram, most venues are restauration and touristic-driven business, showing the principal economic activity in the city of Barcelona. The top three categories represent local food spots, followed by more generic third-sector venues such as hotels, restaurants or bars.

4. Results

4.1. Clusters definition

As the final objective of the study is to create a map that overlaps the rent prices information and the typology of the venues, we need to create clusters that groups venues by their categories, and then locate these venues in the map.

To define the cluster that we will be working with we will need to create a new data frame that contains the top 10 most popular venues for each neighborhood. A dummy matrix will be used to create a data frame containing one row for each venue and a binary code will be filled based n which category it belongs.

This dummy matrix will be grouped by neighborhoods to have the proportion that they have from each category and using a defined function we can extract the X (in our case, 10) more popular venues.

In the end the data frame is like this one:

5]:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Baró de Viver	Asian Restaurant	Dessert Shop	Toy / Game Store	Pet Store	Food	Restaurant	Park	Supermarket	Deli / Bodega	Furniture / Home Store
1	Can Baró	Spanish Restaurant	Grocery Store	Park	Historic Site	Chinese Restaurant	Tapas Restaurant	Big Box Store	Basketball Court	Trail	Café
2	Can Peguera	Park	Supermarket	Bar	German Restaurant	Tapas Restaurant	Restaurant	Grocery Store	Hostel	Plaza	Food & Drink Shop
3	Canyelles	Pizza Place	Mediterranean Restaurant	BBQ Joint	Soccer Stadium	Restaurant	Filipino Restaurant	Exhibit	Falafel Restaurant	Farmers Market	Fast Food Restaurant
4	Ciutat Meridiana	Metro Station	Grocery Store	Mediterranean Restaurant	Park	Supermarket	Plaza	Filipino Restaurant	Falafel Restaurant	Farmers Market	Fast Food Restaurant

Finally, to define the clusters the K-Means method is the chosen one, and the number of clusters is 5, this number was chosen to have a reduced number of clusters, and to avoid overfitted clusters.

4.2. Visualization

The result of the map visualization is the following:

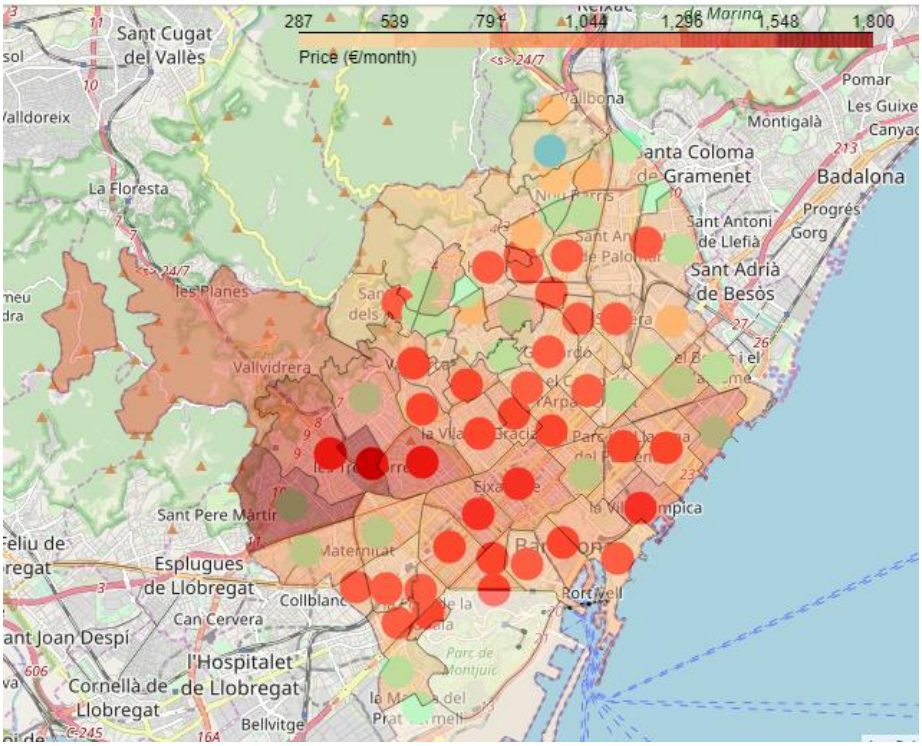

Fig. 9. Map overlaping rent prices and venue category clusters

There are three basic clusters in the map:

- **Red:** Touristic venues, are basically typical Spanish restaurants, cocktails bars, hotels and coffees.

- **Green:** mixed venues, this cluster is formed by neighborhoods that have presence of touristic type of venues mixed with more local driven locations like gyms, soccer fields, tennis courts or swimming pools.
- **Yellow:** local venues, in this cluster there can be found venues such as supermarkets, groceries stores, flea markets or farmers markets.

When analyzing the map, the first observation that, as we said before, most venues are service-oriented and are restaurants or bars. This characteristic can be seen as a symptom of a gentrification process, especially in the neighborhoods with lower rent prices, as it is easier for foreign and local investors to have benefits.

According to the visualization we can find a correlation between the category of the venues and the rent prices in each neighborhood. The ones with more touristic type of venues are nearer the center of the city and have higher rent prices, meanwhile the neighborhoods with less touristic venues have a lower rent price.

5. Conclusions

After analyzing the information that was obtained, we could extract some conclusions. To start, the general typology of the venues in Barcelona show a clear process of gentrification, as much of them are tourist oriented and, as said before, the typology of the venues is a symptom of higher rent prices.

The final map can also help us to understand how the process is evolving and which neighborhoods are more vulnerable. Actually the north-east neighborhoods are the ones that need a quick intervention, because they are the ones with lower rent prices and the adjacent neighborhoods are starting to change the typology of the venues, so is easy to assume that this "cheaper" neighborhoods could be of interest for real state agents and foreign investment funds, that could rise the rent prices and send off the local population.

The data driven investigation has been useful to create a map with the necessary information to make a diagnosis of the situation of the city in the gentrification issue and can be useful for the local government to try to protect the local population.

In the future this investigation can be pushed further and add new layers of information to the map, such as income and educational information of the population, historic situation of each neighborhood or even touristic rent in each area, in order to have a more complete diagnosis of the issue.