Gentrification Map

Martín Ferrer Escobar

29 November 2019

1. Data acquisition and cleaning

1.1. 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.

1.2. 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