Housing Sales Prices versus Venues and Premises Data Analysis of Barcelona

Jordi Triquell

May 10, 2019

Table of Contents

Table of figures

1. Abstract

The aim of this data science research project is to segment neighbourhoods of Barcelona, using as main component the venue categories or activity sector of premises that are located on them. This is contrasted with information of the average housing prices in the area. Two main sources of data are used: Foursquare and the official census with the inventory of premises made by Barcelona City Hall. K-means non-supervised machine learning algorithm clustering algorithm is used to detect patterns in the data. This report may be useful for city residents that may to look for other neighbourhoods in the city with lower real estate prices but that have similar venues and shops to the one they are looking in, or investors that want to choose the better zona to start their business.

2. Introduction

2.1. Description and Discussion of the Background

This article is the main part of a Capstone Project in order to obtain the IBM Data Science Professional Certificate, where the main task is use data science techniques to compare neighbourhoods or cities of our choice or to come up with a problem where Foursquare location data can be used to solve it.

I have chosen to analyse the neighbourhoods of Barcelona which is the city where I currently live in. With a population of 1.6 million within city limits in an area of only 102.2 Km2, Barcelona is one of the most densely populated cities in Europe. The city is divided into 10 districts and there is also a new second-level administrative division of 73 neighbourhoods that was approved in 2006. This subdivision in neighbourhoods is what I am going to use in this research project.

When we think about **city residents,** they prefer to choose areas with lower real estate prices but at the same time they consider the venues, services and other commercial activity that are present in them. The aim of this study is to group similar neighbourhoods in terms of the density of venues and premises that are present on the area. If one specific city resident likes a specific type of neighbourhood because he likes the venues category that are on it, he could also look for other neighbourhoods that belong to the same cluster and have lower prices.

Another target audience of this research are **investors** that pretend to create a business, as they may want to select neighbourhoods where the business activity in which they want to invest is less intense and also detect unattended areas.

Finally, this study might give some insights about factors that influence in the real estate prices, which may be useful for **real estate agents**. We will study if clusters with a similar density of venues and commercial premises tend to have the same price in Barcelona, or on the contrary they are not a determinant factor for the price.

A significant factor that helps in this research is the fact that Barcelona has long been a leader in the movement of smart cities and it is part of a select group of pioneers in smart urban planning along with cities such as Singapore, Vienna, San Francisco and Copenhagen. As a result, it is very effective in collecting a huge amount of data and most of it is available for free to general public. About most than 400 datasets are available in Barcelona's City Hall Open Data Service, including information about venues located on each neighbourhood and purchase housing prices. In addition to this, cartographic data about administrative divisions is available for free in the cartographic division of Barcelona City Hall CartoBCN.

2.2. Data Description

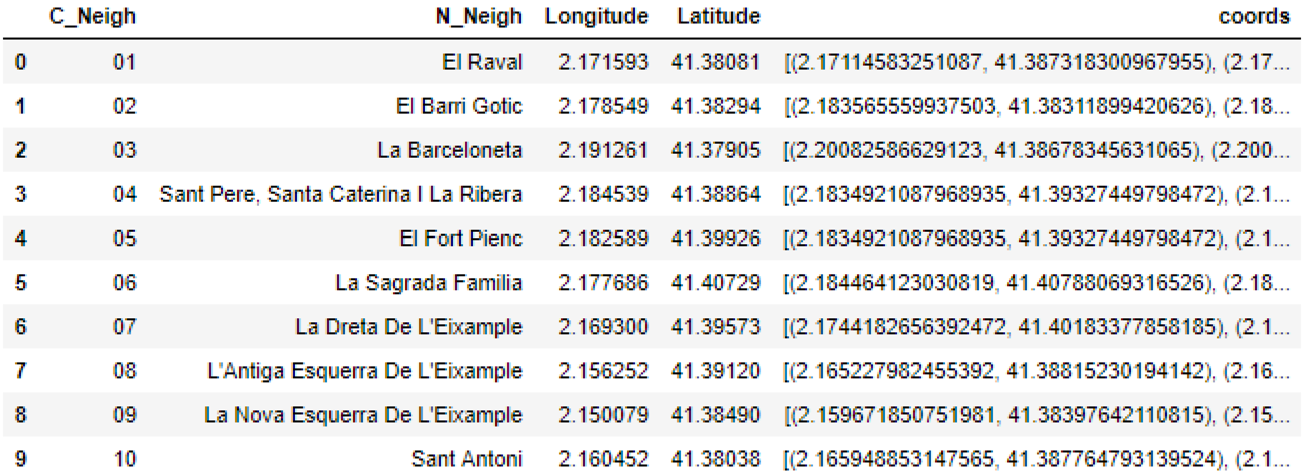
The following sources were used to carry out the project:

- ED50 Administrative Divisions in 2010 from Carto BCN. It contains the code, description and geographical position of the centre and the shape of each neighbourhood. It is possible to download the information by creating a free account. Conversion from UTM to WGS84 geographical coordinate system (used by GPS and Foursquare API) is needed in some cases.
- Purchase of registered properties of the city of Barcelona in 2018. Information about housing prices per square meter considering both new and second-hand properties.
- Venues around a certain location by using a **Foursquare API**
- Inventory of premises of the city of Barcelona in 2016. This was used to contrast results obtained by using Foursquare data.

3. Methodology

Source data was processed obtaining the following data-sets before applying machine learning algorithms.

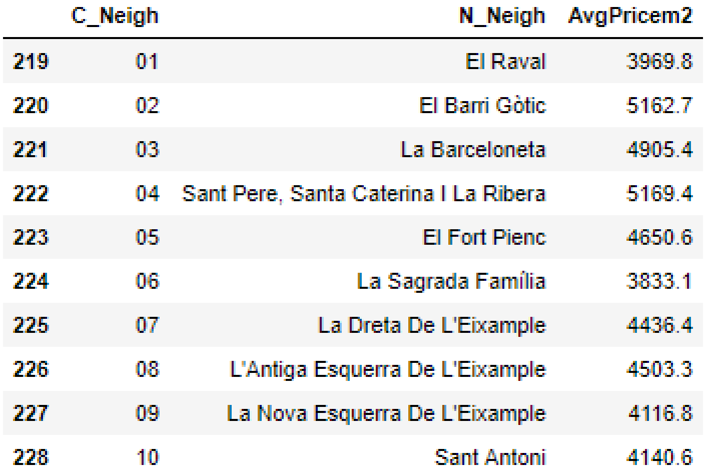
A sample of the processed data that contains geographical position of the centroid for each neighbourhood and a polygon that defines their boundaries is shown below:



	C_Neigh	N_Neigh	Longitude	Latitude	coords
0	01	El Raval	2.171593	41.38081	[(2.17114583251087, 41.387318300967955), (2.17...
1	02	El Barri Gotic	2.178549	41.38294	[(2.183565559937503, 41.38311899420626), (2.18...
2	03	La Barceloneta	2.191261	41.37905	[(2.20082586629123, 41.38678345631065), (2.200...
3	04	Sant Pere, Santa Caterina I La Ribera	2.184539	41.38864	[(2.1834921087968935, 41.39327449798472), (2.1...
4	05	El Fort Pienc	2.182589	41.39926	[(2.1834921087968935, 41.39327449798472), (2.1...
5	06	La Sagrada Familia	2.177686	41.40729	[(2.184464123030819, 41.40788069316526), (2.18...
6	07	La Dreta De L'Eixample	2.169300	41.39573	[(2.1744182656392472, 41.40183377858185), (2.1...
7	08	L'Antiga Esquerra De L'Eixample	2.156252	41.39120	[(2.165227982455392, 41.38815230194142), (2.16...
8	09	La Nova Esquerra De L'Eixample	2.150079	41.38490	[(2.159671850751981, 41.38397642110815), (2.15...
9	10	Sant Antoni	2.160452	41.38038	[(2.165948853147565, 41.387764793139524), (2.1...

Figure 1. Sample data with geographical data of neighbourhoods in Barcelona

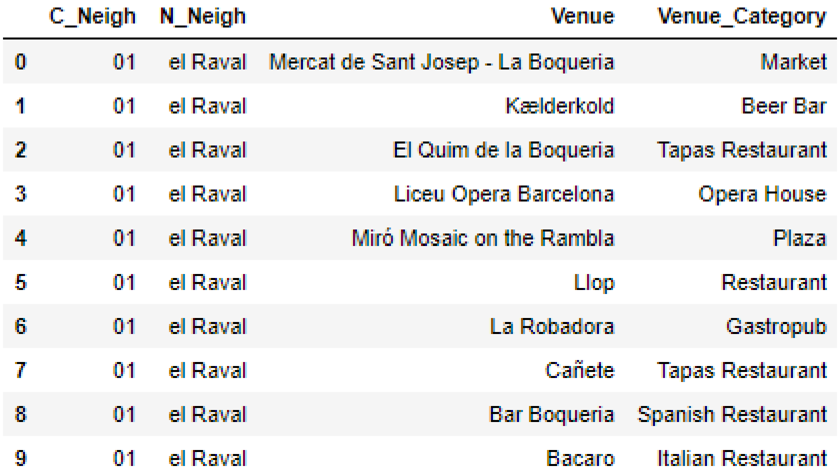
This is a sample of a processed file with average price (EUR per square meter) of new and second-hand purchases of registered properties on each neighbourhood:



	C_Neigh	N_Neigh	AvgPricem2
219	01	El Raval	3969.8
220	02	El Barri Gòtic	5162.7
221	03	La Barceloneta	4905.4
222	04	Sant Pere, Santa Caterina I La Ribera	5169.4
223	05	El Fort Pienc	4650.6
224	06	La Sagrada Família	3833.1
225	07	La Dreta De L'Eixample	4436.4
226	08	L'Antiga Esquerra De L'Eixample	4503.3
227	09	La Nova Esquerra De L'Eixample	4116.8
228	10	Sant Antoni	4140.6

Figure 2. Sample of pre-processed data with average housing prices per square meter

Foursquare API to explore the neighbourhoods and segment them. A radius of 500 meters around the centroid of the neighbourhood and a limit of 100 venues per neighbourhood were established. A total of **2878** venues in **272** distinct categories were obtained. Here, we show a sample of it:



	C_Neigh	N_Neigh	Venue	Venue_Category
0	01	el Raval	Mercat de Sant Josep - La Boqueria	Market
1	01	el Raval	Kælderkold	Beer Bar
2	01	el Raval	El Quim de la Boqueria	Tapas Restaurant
3	01	el Raval	Liceu Opera Barcelona	Opera House
4	01	el Raval	Miró Mosaic on the Rambla	Plaza
5	01	el Raval	Llop	Restaurant
6	01	el Raval	La Robadora	Gastropub
7	01	el Raval	Cañete	Tapas Restaurant
8	01	el Raval	Bar Boqueria	Spanish Restaurant
9	01	el Raval	Bacaro	Italian Restaurant

Figure 3. Sample of venues returned by Foursquare API

Results of the venues obtained for each neighbourhood is shown in a bar chart below. Barrio Gotic, Sant Pere, Santa Caterina i La Ribera, Sant Antoni, El Raval, Poblenou, Vila Gràcia, Barceloneta, Antiga Esquerra de l'Eixample and Nova Esquerra de l'Eixample reached the limit of 100 venues. On the other hand some neighbourhoods such Vallbona, La Clota, Torre Baró, Trinitat Vella, Trinitat Nova, Horta and Can Peguera reached less than 5 venues.

Add description

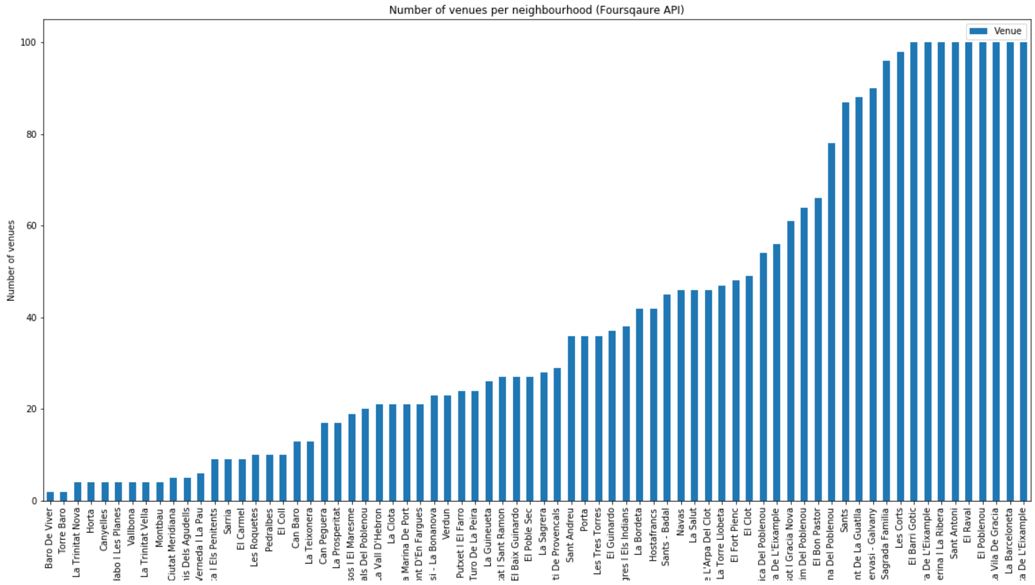
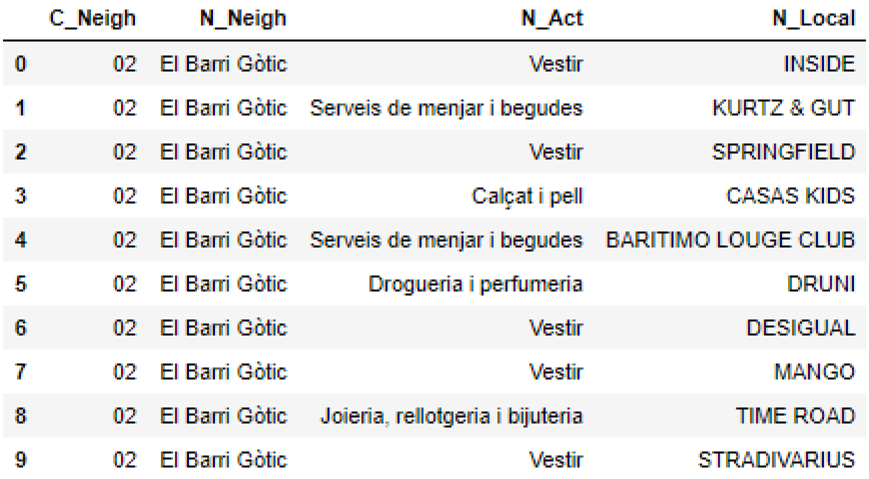


Figure 4. Number of venues per neighbourhood (Foursquare)

The inventory of premises in the city of Barcelona provided us much more data than Foursquare. Only active premises with an active business were considered. Instead of using the distance of the premise to the neighbourhood centroid we considered the neighbourhood where the premise was located as this information was easily available in the dataset. A total of **60.265** premises with an active economic activity and **63** categories were obtained. Here, we show a sample of it:



	C_Neigh	N_Neigh	N_Act	N_Local
0	02	El Barri Gòtic	Vestir	INSIDE
1	02	El Barri Gòtic	Serveis de menjar i begudes	KURTZ & GUT
2	02	El Barri Gòtic	Vestir	SPRINGFIELD
3	02	El Barri Gòtic	Calçat i pell	CASAS KIDS
4	02	El Barri Gòtic	Serveis de menjar i begudes	BARITIMO LOUGE CLUB
5	02	El Barri Gòtic	Drogueria i perfumeria	DRUNI
6	02	El Barri Gòtic	Vestir	DESIGUAL
7	02	El Barri Gòtic	Vestir	MANGO
8	02	El Barri Gòtic	Joieria, rellotgeria i bijuteria	TIME ROAD
9	02	El Barri Gòtic	Vestir	STRADIVARIUS

Figure 5. Sample data of premises (official census)

Similarly to the situation that happened with Foursquare venues, there were some neighbourhoods with less than 50 premises such as Vallbona, La Clota, Torre Baró and Can Peguera. On the other hand more 15 neighbourhoods had more than 1500 premises.

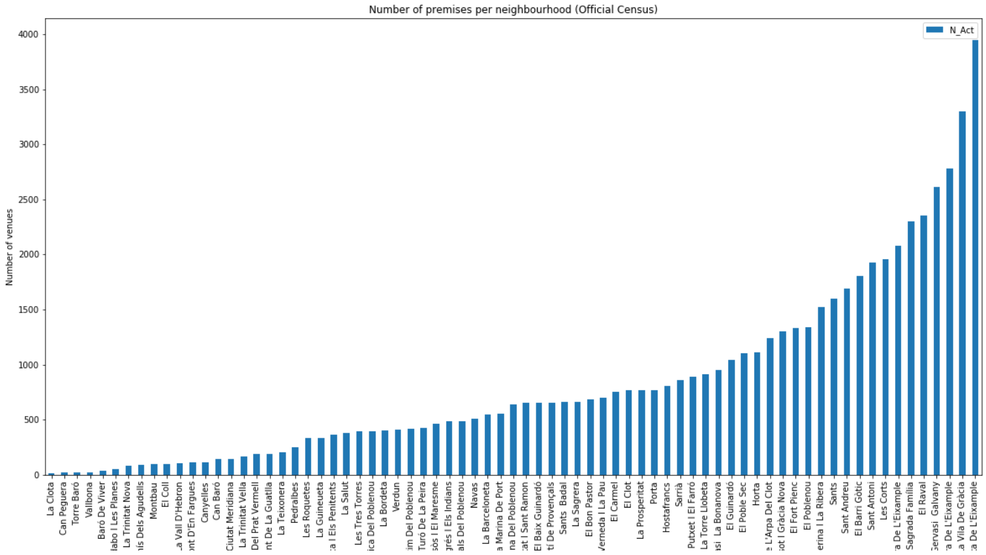


Figure 6. Number of premises per neighbourhood (official census)

Another difference between Foursquare and the inventory of premises is the fact that the first ones considers as venues some items that are not a premise such as theatres, parks, scenic lookouts or even metro stations.

Python Folium library has been used to visualise geographic details of Barcelona and its neighbourhoods. A map of Barcelona generated with the processed data of neighbourhoods' centroids and boundaries is shown below:

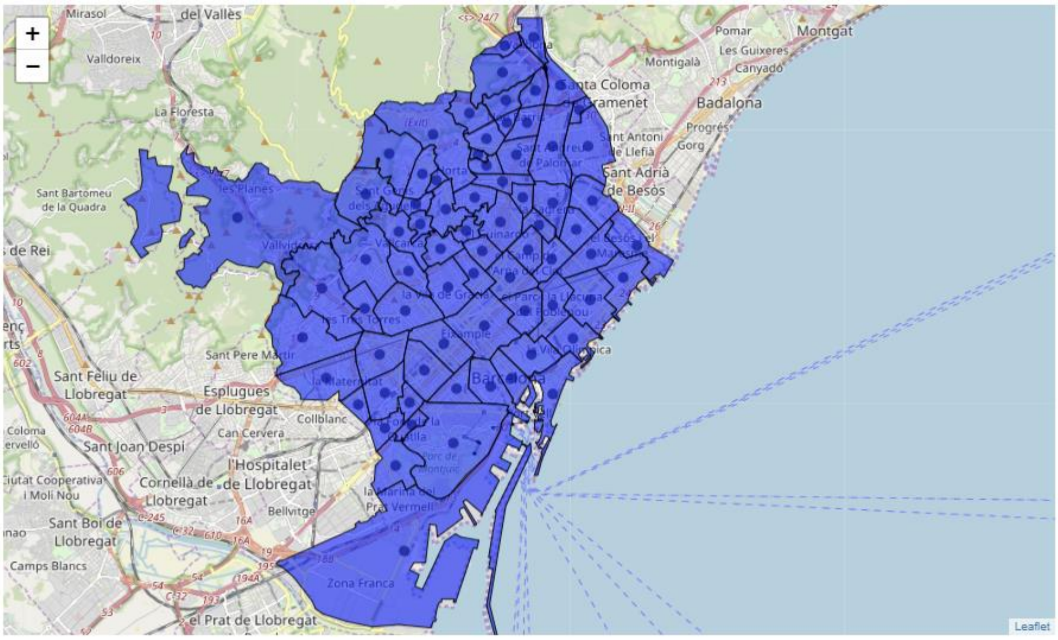


Figure 7. Neighbourhoods of Barcelona with their centroid and boundaries

On the other hand, k-means algorithm from Python Scikit-learn library has been used to cluster different neighbourhoods, as there are some common venue categories or premises with the same commercial activity in them. This algorithm is one of the most common cluster methods of unsupervised learning.

Before applying the algorithm, we have used of the criteria of calculating the frequency of each category (venue category or activity sector of the premise) versus the rest of categories for each neighbourhood. Therefore, the criteria for grouping neighbourhoods has been to find clusters that share similar frequency distributions in most of categories. However, other methods such as density of each category per square meter could be considered to apply a clustering algorithm.

In order to determine the optimal number of clusters we run **K-Means** in a range of 1 to 9 clusters and we have applied elbow method, which consist in measure the average distance from neighbourhoods to the cluster centre they belong. This is the most common and simple method to determine the optimal number of clusters when this algorithm is used.

In both cases elbow method did not show a clear number when adding more clusters doesn't give much better modelling of the data. This suggests that there is not a clear shape boundary between most of neighbourhoods, and that there are many neighbourhoods that act as a gradual transition zone. In our case, we decided to choose 5 cluster as it look a reasonable number to group the neighbourhoods.

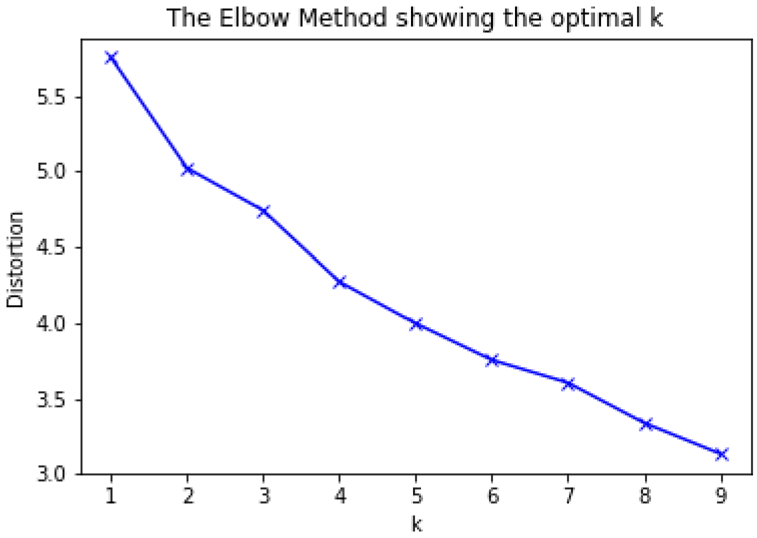


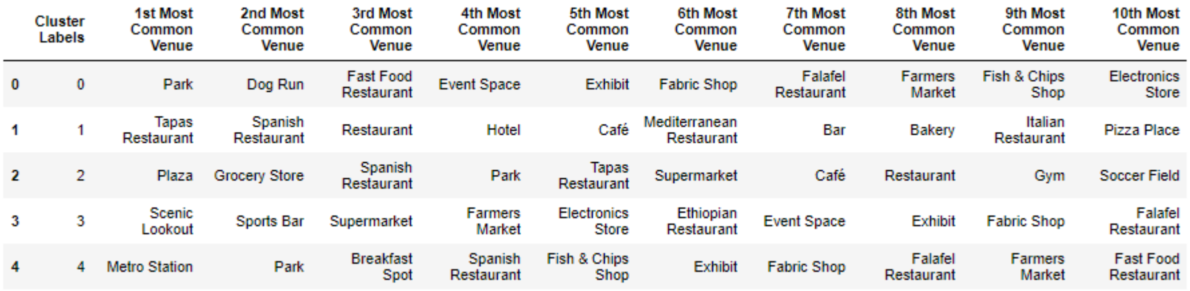
Figure 8. Average distance to the cluster centroid vs number of clusters (k)

A further documentation of the source code and libraries used to transform, visualise an algorithms is available on this GitHub Repository.

4. Results

a) Foursquare API

This table summarises the most common venues types for the 5 clusters that were found using a k-means algorithm using venue database of Foursquare:



	Cluster Labels	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	0	Park	Dog Run	Fast Food Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant	Farmers Market	Fish & Chips Shop	Electronics Store
1	1	Tapas Restaurant	Spanish Restaurant	Restaurant	Hotel	Café	Mediterranean Restaurant	Bar	Bakery	Italian Restaurant	Pizza Place
2	2	Plaza	Grocery Store	Spanish Restaurant	Park	Tapas Restaurant	Supermarket	Café	Restaurant	Gym	Soccer Field
3	3	Scenic Lookout	Sports Bar	Supermarket	Farmers Market	Electronics Store	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant
4	4	Metro Station	Park	Breakfast Spot	Spanish Restaurant	Fish & Chips Shop	Exhibit	Fabric Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant

Figure 9. Neighbourhood clusters obtained from Foursquare API

We can label cluster 1 as "High activity in catering and hotels", That is touristic neighbourhoods that are full of hotels, restaurants, bars and cafes. Cluster 2 is labelled as "Medium activity" and groups neighbourhoods with an average density of venues. Finally cluster 0, 3 and 4 are labelled as clusters with low activity as their main venues are metro stations, scenic lookouts or parks

Next figure shows the distribution of clusters in a map. Purple dots represent cluster 1 with high activity in catering and hotels. It is not a surprise that most of the map is covered in purple as Barcelona is a city full of bars, restaurants and hotels. Light blue dots show neighbourhoods with an average activity which include the north of Barcelona and some residential neighbours located on the tops of some hills. Finally, red, orange and green dots represent the neighbourhoods that are outliers and are located in industrial zones or isolated zones, and there is a low density of venues.

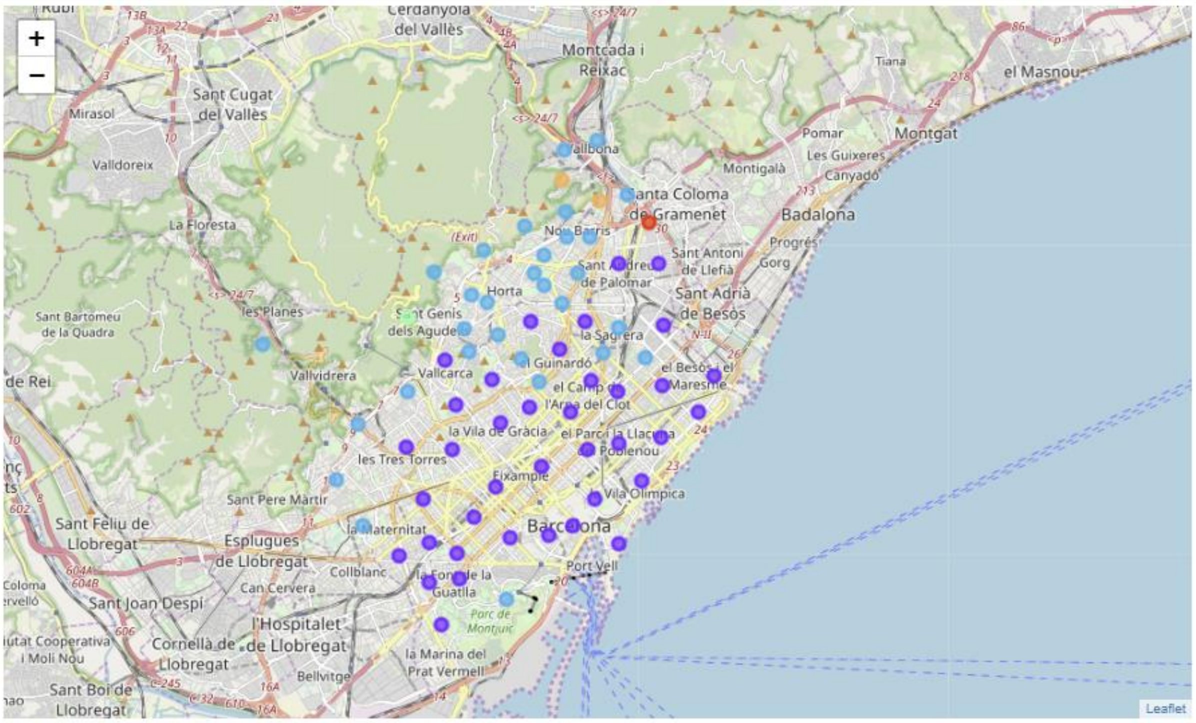


Figure 10. Neighbourhood clusters map (Foursquare source)

This information can be overlapped with a choropleth map that shows a colour gradient linked to the average housing prices per neighbourhood. The most obvious correlation that it is visually appreciated is that neighbourhoods labelled as outliers, exhibit lowest housing prices. Besides, on general terms neighbourhoods with an average density of venues exhibit lower prices (this can be seen in the north of the city). However, there is an exception as three of the most expensive in the western part of the city show also an average density of venues. This may be explained because this area is mainly residential with bigger houses and a minor availability of premises, if we compare with other neighbourhoods of the city.

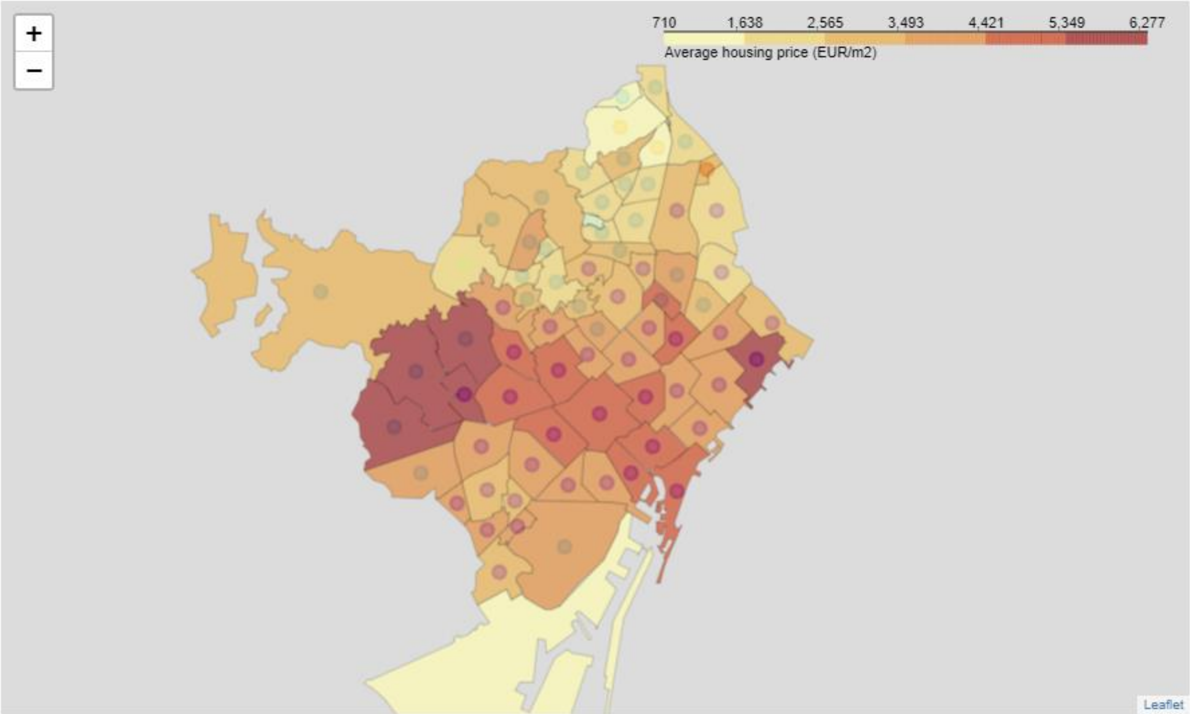
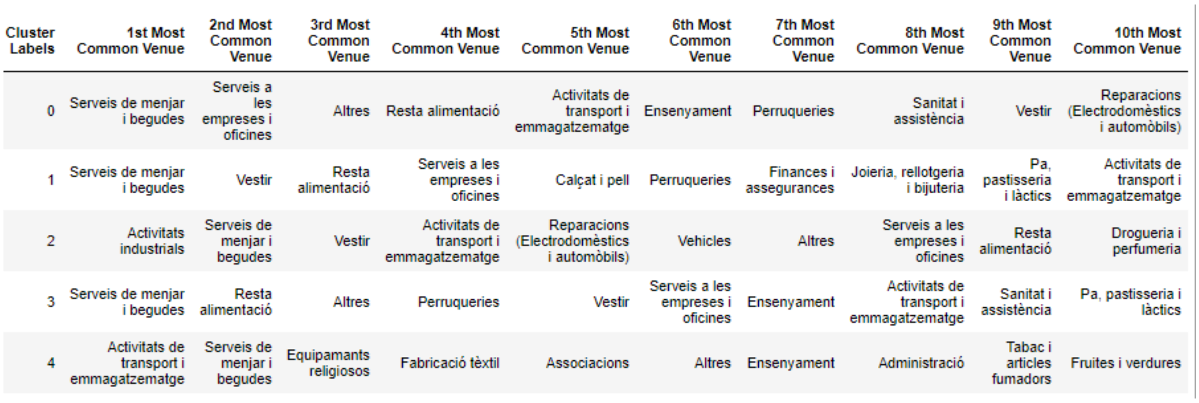


Figure 11. Neighbourhood clusters map and average housing prices (Foursquare source)

b) Inventory of premises

This table summarises the most common venues types for the 5 clusters that were found using a k-means algorithm using the activity of inventory of premises made by Barcelona City Hall:



Cluster Labels	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	Serveis de menjar i begudes	Serveis a les empreses i oficines	Altres	Resta alimentació	Activitats de transport i emmagatzematge	Ensenyament	Perruqueries	Sanitat i assistència	Vestir	Reparacions (Electrodomèstics i automòbils)
1	Serveis de menjar i begudes	Vestir	Resta alimentació	Serveis a les empreses i oficines	Calçat i pell	Perruqueries	Finances i assegurances	Joieria, rellotgeria i bijuteria	Pa, pastisseria i làctics	Activitats de transport i emmagatzematge
2	Activitats industrials	Serveis de menjar i begudes	Vestir	Activitats de transport i emmagatzematge	Reparacions (Electrodomèstics i automòbils)	Vehicles	Altres	Serveis a les empreses i oficines	Resta alimentació	Drogueria i perfumeria
3	Serveis de menjar i begudes	Resta alimentació	Altres	Perruqueries	Vestir	Serveis a les empreses i oficines	Ensenyament	Activitats de transport i emmagatzematge	Sanitat i assistència	Pa, pastisseria i làctics
4	Activitats de transport i emmagatzematge	Serveis de menjar i begudes	Equipamants religiosos	Fabricació tèxtil	Associacions	Altres	Ensenyament	Administració	Tabac i articles fumadors	Fruites i verdures

Figure 12. Neighbourhood clusters obtained from official census

We can label cluster 1 as "Clothes and footwear". This zone coincides with Barcelona downtown and areas frequented by tourists. This type of commercial activity was not detected by Foursquare as it has an inventory of venues and social places but not clothing stores. In addition to the city centre, this cluster contains the neighbourhood of Diagonal Mar (eastern part of the city) where there a big mall is located.

The distinction between clusters 0 and 3 is not easily to see, but in general cluster 3 respond to neighbourhoods where clothing and footwear is an important activity (such as Sants, Hostafrancs, Les Corts), and cluster 0 is more linked to neighbourhoods with a higher presence of teaching institutions (e.g. Sarria, Pedralbes, Tres Torres, Vallcarca and Penitents).

Finally clusters 2 and 4 are outliers with industrial activities or warehouses as main activities which coincide with the last industrial areas of the city in the south west and north east respectively (La Marina del Prat Vermell, Bon Pastor and La Clota).

All the clusters of course exhibit a high number of premises with catering activity and this is not a surprise as Barcelona, as many other Mediterranean city, is a city full of bars everywhere.

Next figure shows the clusters in a map. Purple dots are neighbourhoods that belong to cluster 1 where there is a high frequency of clothing and footwear stores. Green light dots represent cluster 0 with secondary shopping areas. Red dots are clusters are neighbourhoods to cluster 0, and finally blue and orange dots are outliers (cluster 2 and 4) with a lower presence of premises.

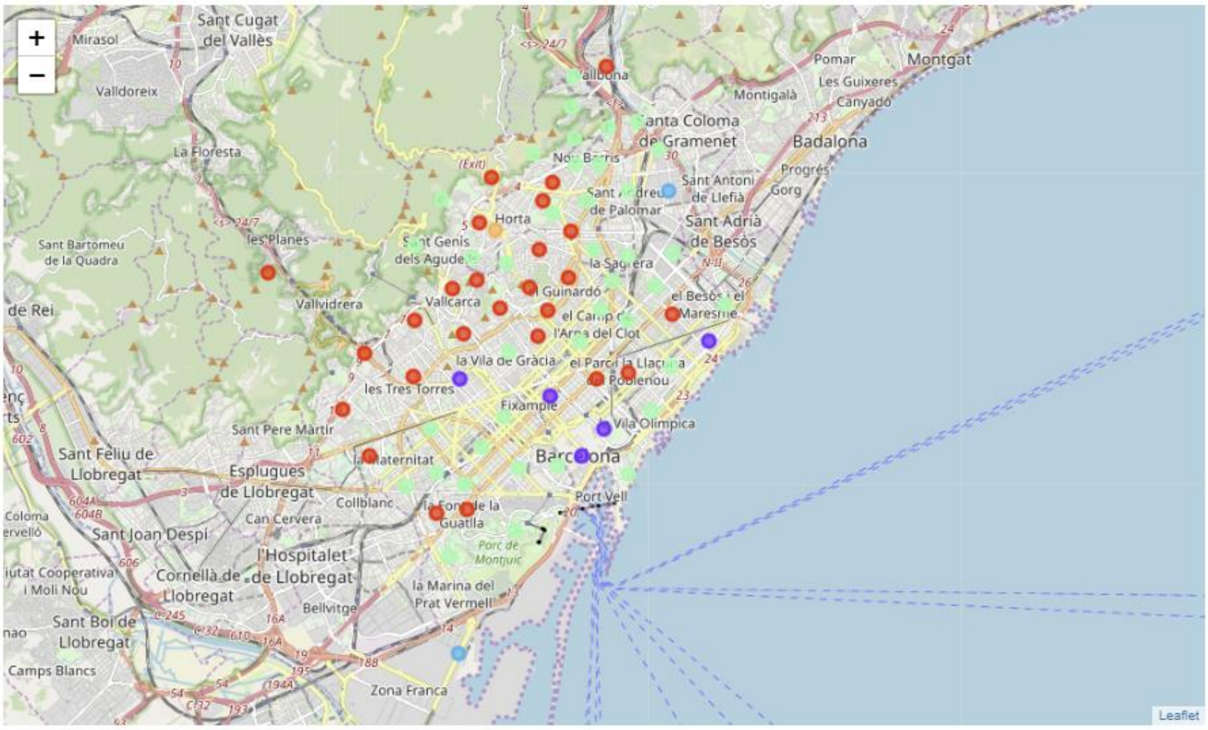


Figure 13. Neighbourhood clusters map (official census source)

As we have done with results obtained with Foursquare API we overlap the clusters with a choropleth map with gradient of average housing prices. Cluster 1 neighbourhoods with a high presence of "Clothing and footwear" are located in areas in expensive neighbourhoods (second quartile) but not in the most expensive area in the west which is mainly residential. The neighbourhoods located in industrial area exhibit lower prices than the average. However, no other clear correlations between average housing prices and premises can be appreciated visually. There is threshold that having more venues does not increase the housing prices and probably other factors are more determinant for the housing prices.

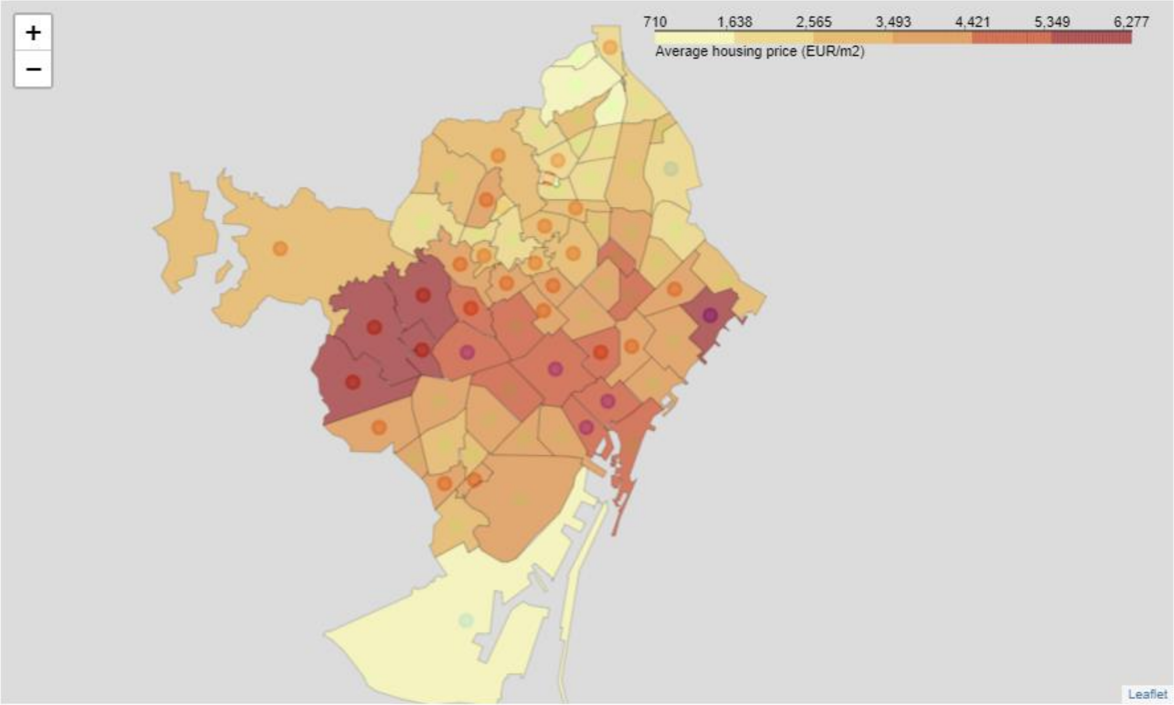


Figure 14. Neighbourhood clusters map and average housing prices (official census source)

5. Discussion

Barcelona is a city with a high population density as many other Mediterranean cities. The total number of measurements and population densities of the 73 neighbourhoods may be slightly different, but with the exemption of some outliers, such as some neighbourhoods placed on some of the last industrial zones of the city or in natural parks, the density is high and rather similar. We consider that 73 neighbourhoods are an accurate segmentation of geographical areas of the city, as smaller areas will be difficult to handle and bigger ones would not provide enough precision.

K-means algorithm has given different results using Foursquare API and the inventory of premises. Elbow method did not give us a well-defined optimal number of clusters and other methods such as Silhouette Value could be considered. On the other hand, a refinement of venues categories or premises could be considered.

With Foursquare API we have obtained 5 clusters but 3 of them contain only 1 or 2 neighbourhoods. These outlying neighbourhoods coincide with peripheral areas of the city that are not so populated or are not purely residential areas like the rest. The remaining two clusters consist of a cluster with a set of 15 of neighbourhoods placed in hilly areas on the north-east of Barcelona, while there is a super cluster of 50 neighbourhoods that covers most of the city. This may us think that neighbourhoods of Barcelona do not tend to concentrate one particular venue category and that most of venues categories are distributed among most of the neighbourhoods of the city or that there are many neighbourhoods that act as transition zones.

When we carry out the analysis by using premises census, some different results are obtained and are more similar to the ones that I expected. Of the 5 clusters obtained two clusters classify outlying neighbourhoods. Of the remaining three there is a first cluster that groups the most centric neighbourhoods (and the most visited by tourists) and Diagonal Mar, a neighbourhood located close to the beach that has become one of the most expensive areas of the city. There is second cluster that groups the districts with a high presence of premises and a last cluster with neighbourhoods with a minor density of premises.

In general terms the neighbourhoods of Barcelona are rather similar in terms of venues and commercial activity. There are cafes, bars, tapas bar, restaurant in almost every corner on the city and this is reflected in the fact that algorithm was not able to give a perfect separation of clusters. This does not mean that Barcelona is a homogeneous city but a cosmopolitan city that it is very heterogeneous in most of neighbourhoods.

Correlation in statistics does not necessarily explains causation, but there is a certain correlation in the neighbourhoods with a lack of venues and lower prices, but we have not seen that most expensive neighbourhoods have a higher density of venues. This can be explained in the fact that people may tend to prefer neighbourhoods with an active city life, but there is a threshold that does not bring more satisfaction (too noisy or crowded is not good).

We can also see that there is a certain correlation with cluster based on venues and premises and housing prices, but only when we are considering the neighbourhoods where the lowest housing prices, as they tend to have clusters with less venues or premises. However, no correlation is seen when we compare clusters between the neighbourhoods with medium and high prices.

Investors that want to acquire a property in Barcelona can consider most of neighbourhoods in terms of venues and facilities, as they are rather similar with the exemption of some outlying neighbourhoods. Therefore, they can save money by investing in neighbourhood with lower average prices, if their main criteria are the venues around it.

6. Conclusion

Information provided by public institutions such as datasets published by Barcelona City Hall and Barcelona cartographic institutions can be a reliable source to get insights about what's happening in big cities and take better decisions in consequence.

This may allow investors, city managers or public in general to take better decisions, by using this data, but we have also seen that this data needs to be contrasted with different sources of possible, as different outcomes may occur.

Further lines of research for this study should be tested and implemented. One could be using a venue category density on each neighbourhood instead of the venue category frequency that has been used in the study. Other line of research could be combining categories specially with results obtained from Foursquare as there are many categories that are very similar (e.g. many kind of restaurants) and k-means algorithm consider each category as orthogonal versus the others categories.

In general terms, we ca not conclude if it is better to use the official inventory of premises or Foursquare, as they provide complementary data in some cases. Combining information could be done in a future research.

7. References

- Barcelona - Wikipedia
- ED50 Administrative Divisions in 2010 from Carto BCN.
- Purchase of registered properties of the city of Barcelona in 2018.
- Inventory of premises of the city of Barcelona
- Venues around a certain location by using a Foursquare API