Barcelona and it's neighbourhoods (2015-2019)

Clusterization and analysis of deriving trends

by: Elies Lahiguera Prats

Abstract

The purpose of this study is to show if systemic differences amongst Barcelona's neighbourhoods could be grouped using a machine learning clusterization algorithm (KMeans) for each of the five years being studied in pairs of years.

Our goal would be to repeatedly obtain clear and traceable distinct neighbourhood groups based on the socioeconomic factors present in the used data-set.

Ideally, revealed trends would have to resemble the existing reality in terms of neighbourhood changes, and should those changes be assessed by public-domain knowledge.

Introduction

We used public data from the City-Council Statistics dpt. to create the used data-set with the intention of representing yearly series of all the Barcelona city's neighbourhoods including variables historically related to displacement and neighbourhood change phenomena.
Data is conceptually divided in three categories:

1. Housing Factors: Here we include terms relative to the rent market as well as the real state market. From average prices to amount of registered purchases and a rent growth rate, calculated subtracting expired contracts from newly signed contracts. Also the area of each neighbourhood classified per registered activity, that provides a clearer sense of the studied neighbourhoods.

2. Resident Characteristics: Where we include percentages of population by certain conditions like higher education, foreigners and unemployed.

3. Economic Info: Here we include terms relative to wealth (i.e. disposable income) and it's inequality distribution (i.e. Gini index).

The chosen period spans from 2015 to 2019. This range was selected due to the next reasons:

1. Five years is a lapse long enough to reveal medium-term tendencies in a society.

2. This period's been chosen by its "inter-crisis" properties, since it begins seven years after the 2008 Sub-Prime crisis and lasts until a year before 2020 Pandemic crisis.

State of Art

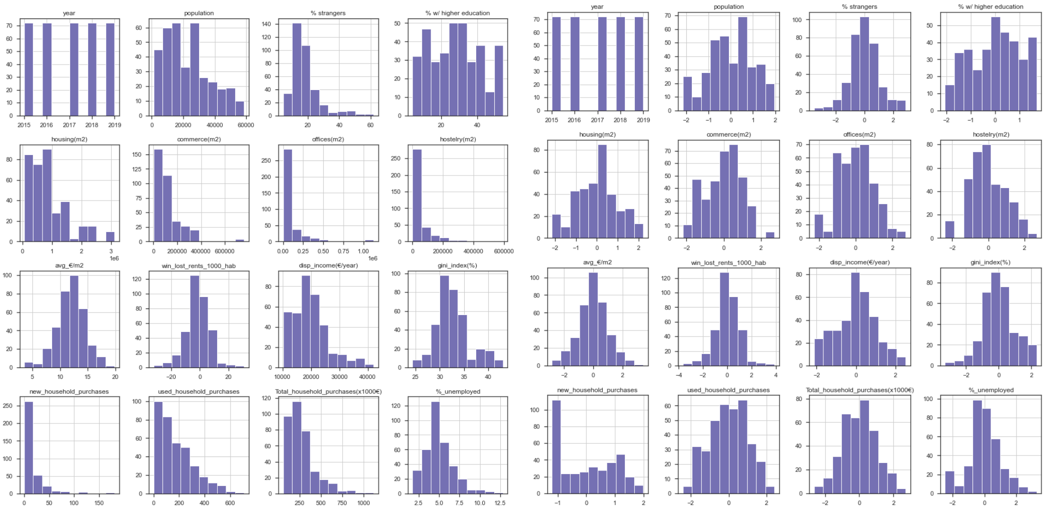
The problem with gentrification and displacement has been extensively studied in more thorough manners by the Academics, since the term was first coined by British urbanist Ruth Glass in 1964.

Resident characteristics, geolocation, fine-grain area splitting (i.e. censal areas), complex economical indicators, property type profiling (e.g. multi-housing units), proximity to roads, subsidized housing and many more features have been usually taken into account in the plethora of such studies.

In this sense, our aim was to deliver the most possibly compact and light data-set and see if the derived results provide enough meaningful insights with a minimum loss. Thus we could have a light-weight model yet capable of delivering enough information with lesser data.

Methodology

While Pre-Processing, outliers were conserved since were genuine. Missing values were imputed by the average percent growth of the more similar neighbourhoods within the same district. Afterwards, a Power Transformation (Yeo-Johnson) was applied to features to make data more Gaussian, as well as standardized.



(a) Features before transformation

(b) Features after transformation

Figure 1: Data Transforms

Then dimensionality reduction was applied trough Principal Component Analysis technique (PCA from now on). Cumulative explained variance was the key metric since we aimed to explain 95% of the data. Cluster distortion (namely WCSS, within cluster sum of squares) and Silhouette Visualizer where used to assess quality of the clustering based on the average silhouette coefficient.

Five was the number of clusters chosen, both because usually obtained results among the best scores and also fitted the logic behind this study, since there were a few groups strongly featured like; the wealthier, the touristic and the "less-favoured". So the remaining had to be the average-ones, but since we wanted to get an average group and a "pivotal-group" to show transition (like average but closer to more defined groups) we ultimately selected 5 clusters for the whole study.

Results

To depict evolution of time and follow redistribution of groups we used maps of Barcelona colored with the corresponding cluster color code and Sankey diagrams.

First two years only show minor changes, three in the neighbourhoods at the Northern hills surrounding el Carmel (el Coll, Can Baró and Vallcarca i els Penitents) and one closer to the East-side of Downtown (Hostafrancs).

Clusters are pretty steady, meaning if most of the clusters remain similar enough within that two year frame, consequently we could infer that, so do the differences amongst their features too.

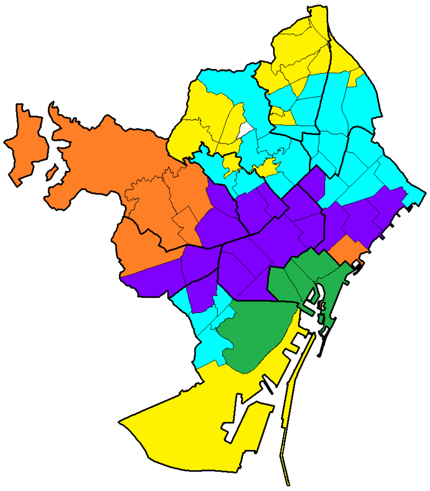


Figure 2: 2015 Cluster Map

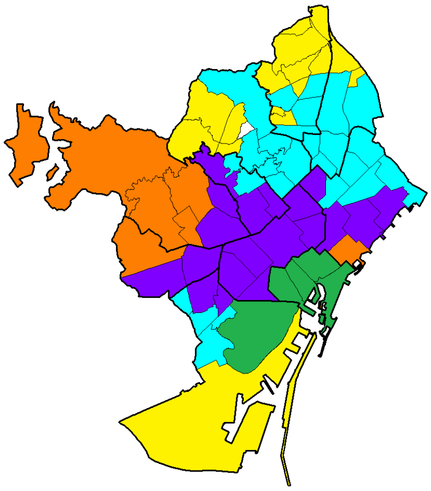


Figure 3: 2016 Cluster Map

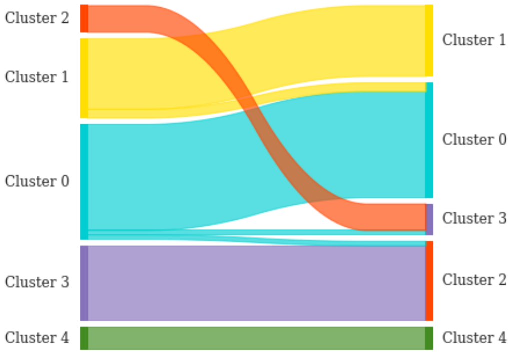


Figure 4: Sankey Diagram depicting neighbourhood redistribution between 2015-2016 clusters

Between 2016 and 2017, two big changes start to develop in the clustering involving two neighbourhood groups:

1. A group focused on the Northern Hills, mainly extending the trend initiated on 2016 to the surrounding areas.
2. A group that merges the two most central clusters (accounting for almost six complete Districts) into Cluster 1.

The fact that other clusters experienced minor changes, could imply some type of standardisation amongst feature values (i.e. lessen price differences) between neighbourhoods within Group 2.

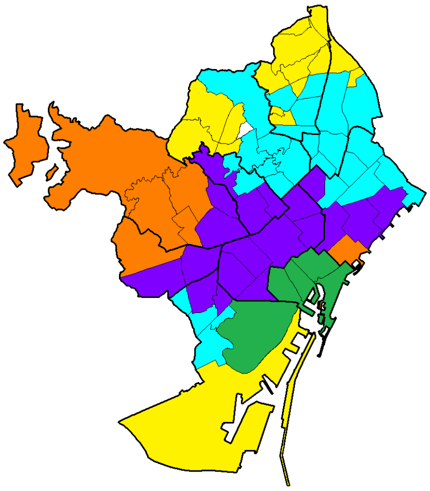


Figure 5: 2016 Cluster Map

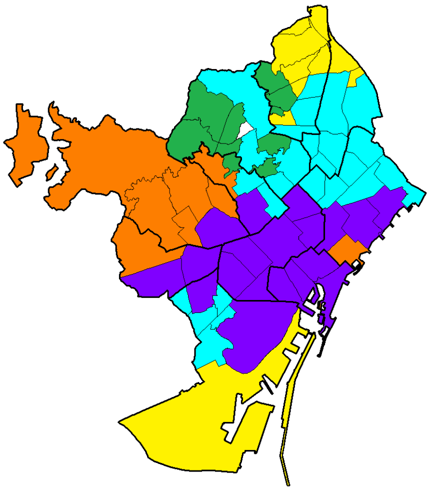


Figure 6: 2017 Cluster Map

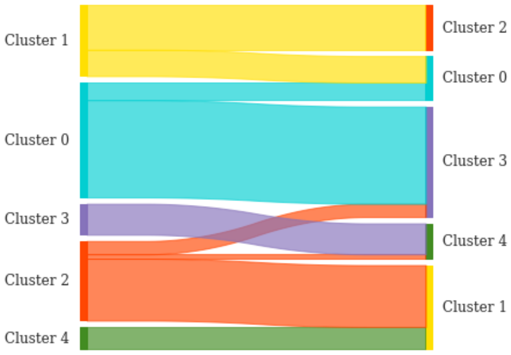


Figure 7: Sankey Diagram depicting neighbourhood redistribution between 2016-2017 clusters

Between 2017 and 2018 no major changes are seen although it's mostly the Group 2 from 2016 where minor changes take place. Interestingly two coastal areas change of cluster, but these two changes take no resemblance in each other since:

1. La Barceloneta is a neighbourhood with a quite strong touristic pressure due to it's proximity both to downtown and the Sea. Hence the first two years it's cluster contained the Downtown areas with most surface destined to tourism and consequently the highest percentages of foreign people.
   Nevertheless in 2018 it's unexpectedly clustered with some average neighbourhoods with no particular attractions, mostly from the outskirts and even some over-industrialized areas like el Bon Pastor o el Besós.

2. Diagonal Mar i el Front Marítim del Poblenou it's a revitalized zone with multiple offices and hotels being built. Sort of the newest growing coastal area, with tall luxury apartments by the sea. It has one of the wealthiest levels of disposable income per capita.

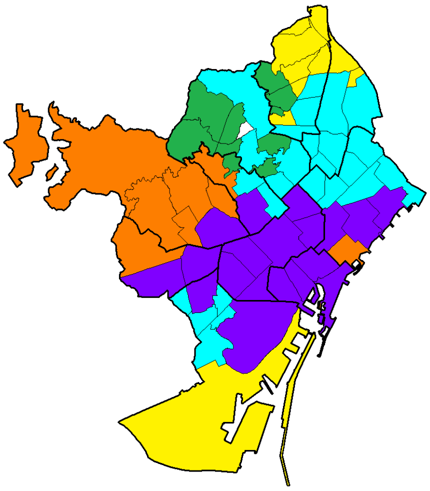


Figure 8: 2017 Cluster Map

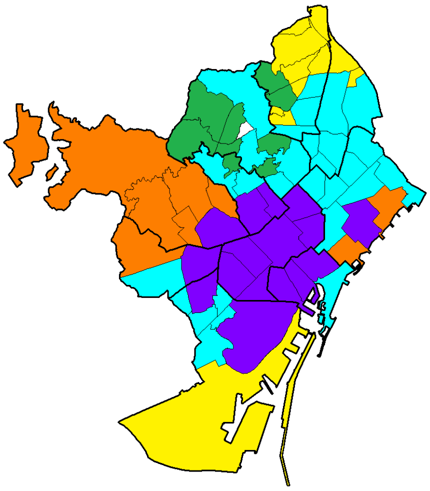


Figure 9: 2018 Cluster Map

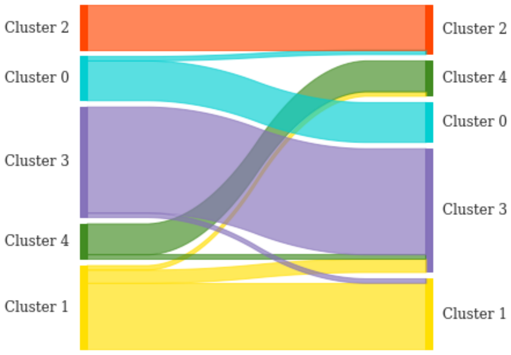


Figure 10: Sankey Diagram depicting neighbourhood redistribution between 2017-2018 clusters

Between 2018 and 2019 changes mimic the 2017 shape to a great extent. This standardization of central areas on the other hand comes along with bigger fragmentation amongst North-Eastern working class areas. This latter might be due to the fixed number of clusters, hence when two clusters approach their values enough to be considered as one cluster then the remaining clusters, even though if their values are steady, will most definitely be over-split.

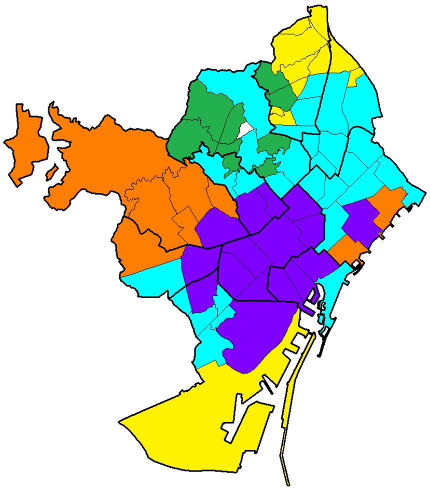


Figure 11: 2018 Cluster Map

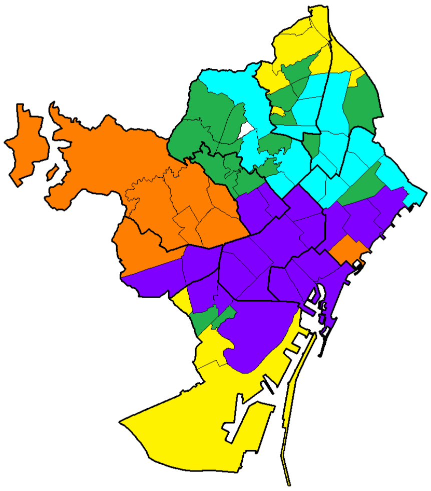


Figure 12: 2019 Cluster Map

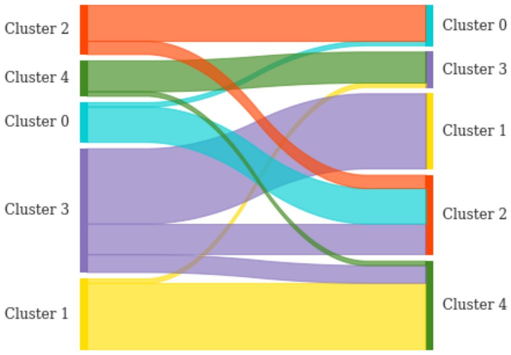
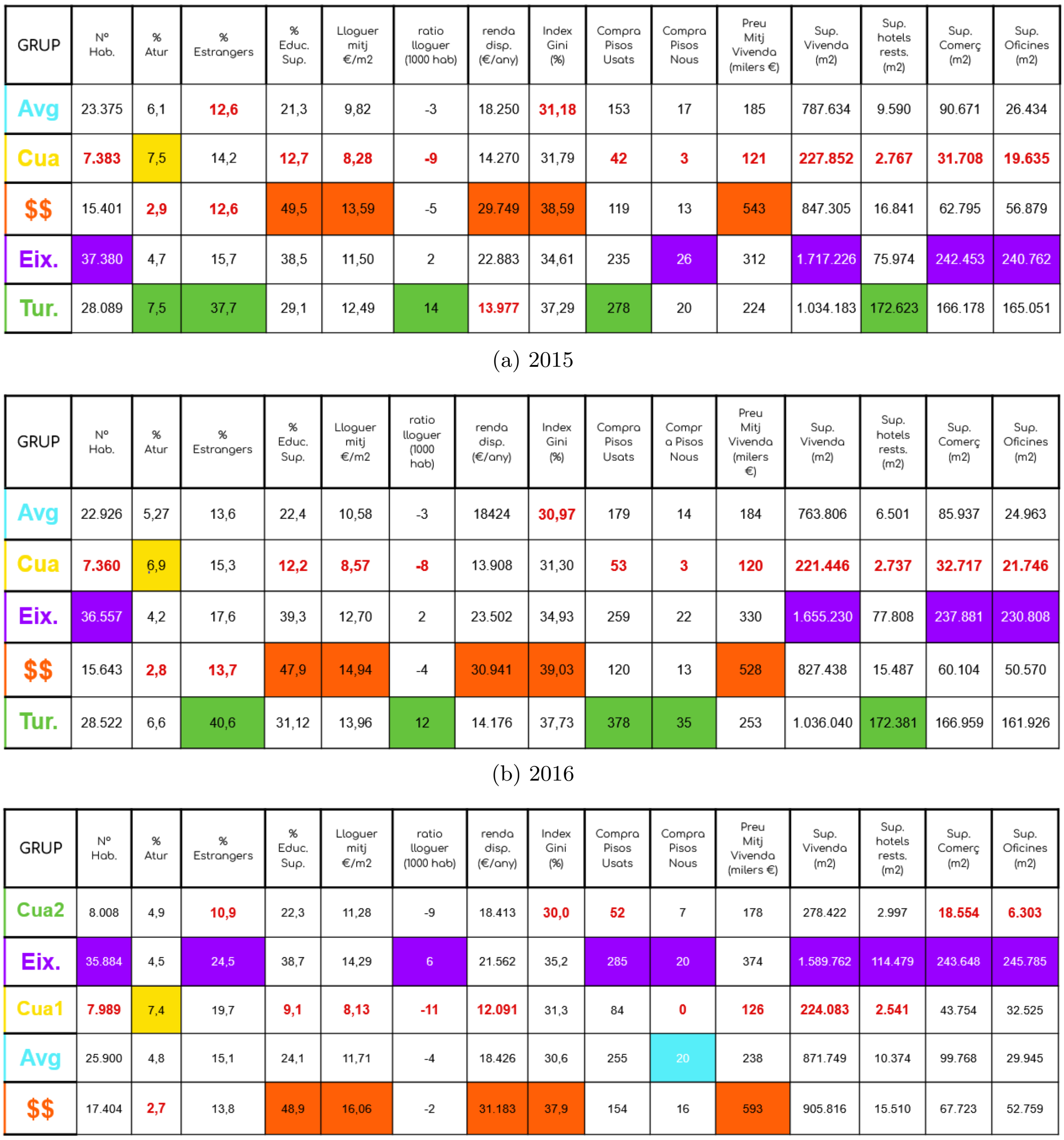


Figure 13: Sankey Diagram depicting neighbourhood redistribution between 2018-2019 clusters

Figure 14: Cluster Analysis Tables 2015-2017. ***Note:*** Max. values are colored with the color of it's cluster, Min. values are colored in red



GRUP	Nº Hab.	% Atur	% Estrangers	% Educ. Sup.	Lloguer mitj €/m2	ratio lloguer (1000 hab)	renda disp. (€/any)	Index Gini (%)	Compra Pisos Usats	Compra Pisos Nous	Preu Mitj Vivenda (milers €)	Sup. Vivenda (m2)	Sup. hotels rests. (m2)	Sup. Comerç (m2)	Sup. Oficines (m2)
Avg	23.375	6,1	**12,6**	21,3	9,82	-3	18.250	**31,18**	153	17	185	787.634	9.590	90.671	26.434
Cua	**7.383**	7,5	14,2	**12,7**	**8,28**	**-9**	14.270	31,79	**42**	**3**	**121**	**227.852**	**2.767**	**31.708**	**19.635**
$$	15.401	**2,9**	**12,6**	49,5	13,59	-5	29.749	38,59	119	13	543	847.305	16.841	62.795	56.879
Eix.	37.380	4,7	15,7	38,5	11,50	2	22.883	34,61	235	26	312	1.717.226	75.974	242.453	240.762
Tur.	28.089	7,5	37,7	29,1	12,49	14	**13.977**	37,29	278	20	224	1.034.183	172.623	166.178	165.051

(a) 2015

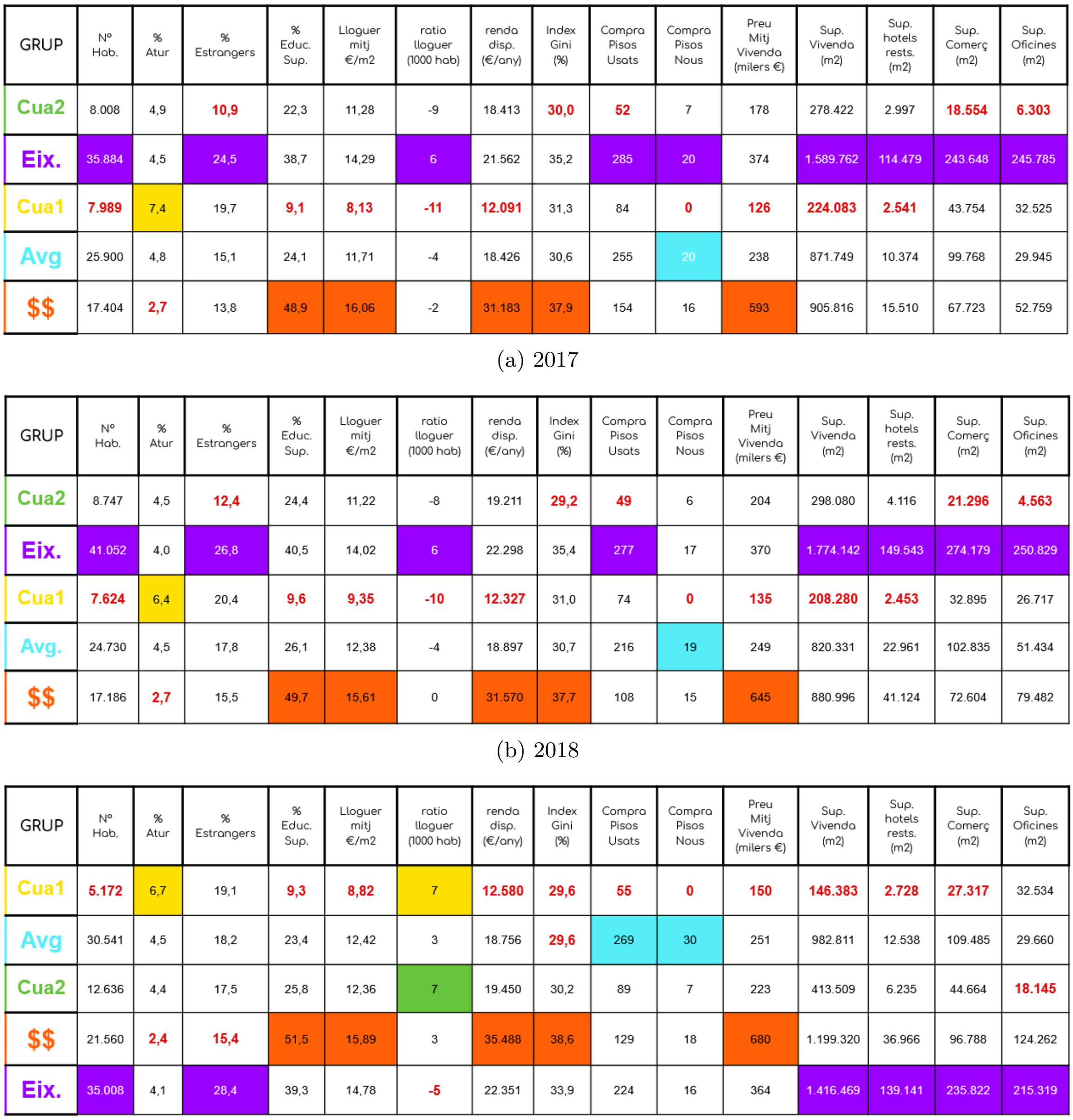
GRUP	Nº Hab.	% Atur	% Estrangers	% Educ. Sup.	Lloguer mitj €/m2	ratio lloguer (1000 hab)	renda disp. (€/any)	Index Gini (%)	Compra Pisos Usats	Compr a Pisos Nous	Preu Mitj Vivenda (milers €)	Sup. Vivenda (m2)	Sup. hotels rests. (m2)	Sup. Comerç (m2)	Sup. Oficines (m2)
Avg	22.926	5,27	13,6	22,4	10,58	-3	18424	**30,97**	179	14	184	763.806	6.501	85.937	24.963
Cua	**7.360**	6,9	15,3	**12,2**	**8,57**	**-8**	13.908	31,30	**53**	**3**	**120**	**221.446**	**2.737**	**32.717**	**21.746**
Eix.	36.557	4,2	17,6	39,3	12,70	2	23.502	34,93	259	22	330	1.655.230	77.808	237.881	230.808
$$	15.643	**2,8**	**13,7**	47,9	14,94	-4	30.941	39,03	120	13	528	827.438	15.487	60.104	50.570
Tur.	28.522	6,6	40,6	31,12	13,96	12	14.176	37,73	378	35	253	1.036.040	172.381	166.959	161.926

(b) 2016

GRUP	Nº Hab.	% Atur	% Estrangers	% Educ. Sup.	Lloguer mitj €/m2	ratio lloguer (1000 hab)	renda disp. (€/any)	Index Gini (%)	Compra Pisos Usats	Compra Pisos Nous	Preu Mitj Vivenda (milers €)	Sup. Vivenda (m2)	Sup. hotels rests. (m2)	Sup. Comerç (m2)	Sup. Oficines (m2)
Cua2	8.008	4,9	**10,9**	22,3	11,28	-9	18.413	**30,0**	**52**	7	178	278.422	2.997	**18.554**	**6.303**
Eix.	35.884	4,5	24,5	38,7	14,29	6	21.562	35,2	285	20	374	1.589.762	114.479	243.648	245.785
Cua1	**7.989**	7,4	19,7	**9,1**	**8,13**	**-11**	**12.091**	31,3	84	**0**	**126**	**224.083**	**2.541**	43.754	32.525
Avg	25.900	4,8	15,1	24,1	11,71	-4	18.426	30,6	255	20	238	871.749	10.374	99.768	29.945
$$	17.404	**2,7**	13,8	48,9	16,06	-2	31.183	37,9	154	16	593	905.816	15.510	67.723	52.759

(c) 2017

Figure 15: Cluster Analysis Tables 2017-2019. ***Note:*** Max. values are colored with the color of it's cluster and Min. values are colored in red



GRUP	Nº Hab.	% Atur	% Estrangers	% Educ. Sup.	Lloguer mitj €/m2	ratio lloguer (1000 hab)	renda disp. (€/any)	Index Gini (%)	Compra Pisos Usats	Compra Pisos Nous	Preu Mitj Vivenda (milers €)	Sup. Vivenda (m2)	Sup. hotels rests. (m2)	Sup. Comerç (m2)	Sup. Oficines (m2)
Cua2	8.008	4,9	**10,9**	22,3	11,28	-9	18.413	**30,0**	**52**	7	178	278.422	2.997	**18.554**	**6.303**
Eix.	35.884	4,5	24,5	38,7	14,29	6	21.562	35,2	285	20	374	1.589.762	114.479	243.648	245.785
Cua1	**7.989**	7,4	19,7	**9,1**	**8,13**	**-11**	**12.091**	31,3	84	**0**	**126**	**224.083**	**2.541**	43.754	32.525
Avg	25.900	4,8	15,1	24,1	11,71	-4	18.426	30,6	255	20	238	871.749	10.374	99.768	29.945
$$	17.404	**2,7**	13,8	48,9	16,06	-2	31.183	37,9	154	16	593	905.816	15.510	67.723	52.759

(a) 2017

GRUP	Nº Hab.	% Atur	% Estrangers	% Educ. Sup.	Lloguer mitj €/m2	ratio lloguer (1000 hab)	renda disp. (€/any)	Index Gini (%)	Compra Pisos Usats	Compra Pisos Nous	Preu Mitj Vivenda (milers €)	Sup. Vivenda (m2)	Sup. hotels rests. (m2)	Sup. Comerç (m2)	Sup. Oficines (m2)
Cua2	8.747	4,5	**12,4**	24,4	11,22	-8	19.211	**29,2**	**49**	6	204	298.080	4.116	**21.296**	**4.563**
Eix.	41.052	4,0	26,8	40,5	14,02	6	22.298	35,4	277	17	370	1.774.142	149.543	274.179	250.829
Cua1	**7.624**	6,4	20,4	**9,6**	**9,35**	**-10**	**12.327**	31,0	74	**0**	**135**	**208.280**	**2.453**	32.895	26.717
Avg.	24.730	4,5	17,8	26,1	12,38	-4	18.897	30,7	216	19	249	820.331	22.961	102.835	51.434
$$	17.186	**2,7**	15,5	49,7	15,61	0	31.570	37,7	108	15	645	880.996	41.124	72.604	79.482

(b) 2018

GRUP	Nº Hab.	% Atur	% Estrangers	% Educ. Sup.	Lloguer mitj €/m2	ratio lloguer (1000 hab)	renda disp. (€/any)	Index Gini (%)	Compra Pisos Usats	Compra Pisos Nous	Preu Mitj Vivenda (milers €)	Sup. Vivenda (m2)	Sup. hotels rests. (m2)	Sup. Comerç (m2)	Sup. Oficines (m2)
Cua1	**5.172**	6,7	19,1	**9,3**	**8,82**	7	**12.580**	**29,6**	**55**	**0**	**150**	**146.383**	**2.728**	**27.317**	32.534
Avg	30.541	4,5	18,2	23,4	12,42	3	18.756	**29,6**	269	30	251	982.811	12.538	109.485	29.660
Cua2	12.636	4,4	17,5	25,8	12,36	7	19.450	30,2	89	7	223	413.509	6.235	44.664	**18.145**
$$	21.560	**2,4**	**15,4**	51,5	15,89	3	35.488	38,6	129	18	680	1.199.320	36.966	96.788	124.262
Eix.	35.008	4,1	28,4	39,3	14,78	**-5**	22.351	33,9	224	16	364	1.416.469	139.141	235.822	215.319

(c) 2019

Discussion

As stated, we observed that clustering was repeatedly identifying groups based on certain conditions as follows:

1. Distribution amongst clusters followed a clear pattern with some clusters persistently getting the maximum values in some features and other group of clusters that usually got the lowest values.

2. Regarding their geographical situation, we could theorize that the conditions on clustering are inevitably defined by "purchasing-power" related features, i.e disposable income or prices. Thus the map depicts actual economic boundaries in a fairly accurate way.

   On the contrary, in a chicken and egg situation, it could be argued that it's precisely the location in relation to Downtown, and other assets i.e. Public Transport (that we didn't include in the given data-set), what explains the value of the properties and consequently sets the price barriers.

3. The cluster with the highest percentage of unemployed residents always showed the lowest percentage of residents with higher studies (amongst quite other features).

   This, on the other side could be explained due to the smallest housing area that this cluster accounts for, which in turn delivers the smallest population numbers. And that alone may be inflating the percentages.

According to the results the merge in 2018-2019 might show a standardization of conditions on virtually every neighbourhood under "La Diagonal" with a couple of exceptions expanding the central area to the north.

Conclusion

Although we managed to explain the model, to a certain extent, with the help of public-domain knowledge and the previous exploratory data analysis, we cannot assess complete reliability on the results of the model given it's lack of depth.

Arising questions can't be firmly explained with the current data available, besides steadiness is a common thing amongst quite a few variables and this latter alone is suggesting that those features were "short-framed" and their observable trends could be revealed in a bigger time-frame.
Therefore we can state that our objective working the model on a lighter database wasn't attained.

Improvements

Use of time series clustering algorithms should be considered when facing similar tasks since those are the algorithms specifically designed to tackle such duties. Ironically such algorithms were considered when we where running out of time...

Selecting the ideal number of clusters in unsupervised algorithms might still be a concerning subject, so considering KMeans Sensitivity regarding the selection of K or even switching to classification algorithms might render useful results.

A broader set of features and specially a broader time-frame may deliver more robust medium/long trends, since seemingly, cities don't grow overnight.

List of Figures

References

[Bro16] J. Brownlee. *Machine learning mastery with python.* Maching Learning. 2016.

[CP19] M. Cohen and K. Pettit. Guide to measuring neighborhood change to understand and prevent displacement. April 2019.

[GV18] Laia Gallego-Vila. El barraquismo en la ciudad de Barcelona durante el Franquismo. Primeras aproximaciones a una domesticidad desde los márgenes. 8, January 2018.

[Ins22] Inside Airbnb Project. Airbnb dataset listings. Web Project, 2022.

[Kno19] D. Knorr. Using machine learning to identify and predict gentrification in nashville, Tennessee. mathesis, Vanderbilt University, August 2019.

[Lah22] E. Lahiguera. Exploratory data analysis of Barcelona's neighbourhoods, 2022.

[ZC] Alice Zheng and Amanda Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly Media.

[LJMMMS20] Miquel Àngel Garcia-López, Jordi Jofre-Monseny, Rodrigo Martínez-Mazza, and Mariona Segú. Do short-term rental platforms affect housing markets? Evidence from Airbnb in Barcelona. *Journal of Urban Economics*, 119:103278, 2020.