Identification of real estate investment opportunities by crossed segmentation analysis of socio-economic development indicators and housing prices

Author: Francesco Fusaro

Date: April, 2021

Project prepared for submission for the IBM's Applied Data Science Capstone

Table of Contents

1 Introduction/Business problem

1.1 Background

Real estate investors are continuously looking for new areas in which to acquire properties. Obviously, the expectation is to ensure a good income from the rents generated by the property and to maximize the ratio between income derived from the rents and the initial property investment. From a real estate investor perspective, it is therefore interesting to identify at an early-stage, areas in which rents are expected to increase due to the area's positive economic-value development, while property prices do not yet (fully) reflect the development. As a side effect, due to its increase in market value over time, a property located in an area with such development prospects might also lead to higher benefits if the property is sold later on.

1.2 Business related key question

We are therefore trying to answer the following question of a real estate investor: "Which area of the city is expected to have higher ratios between the future income derived from renting housing space (rents) and the initial property investment?"

1.3 Approach/Value proposition

To answer this question, we will identify city areas for recommendation, whose current economic-value development indicates an increase of the rents level over time, but by today's comparison with area with similar economic-value, features lower property prices.

On one hand, we will try to assess the economic-value development of a city area by leveraging FourSquare location data to segment and compare the areas based on their most common venue categories according to popular places and venues recommended by FourSquare users. As further indication of area development we will consider the demographic evolution (population increase or decrease). On the other hand, current and historical property price and rents data will be used to segment the city areas in terms of property price level and rents. For more details on the data refer to Section 2.

By combining insights from economic-value and social development data with price and cost data we expect to be able to provide valuable data driven information to answer the question above.

We will prototype the approach on the neighborhoods of the city of Madrid, Spain

1.4 Business interest

Real estate investors and real estate agencies in Madrid would be equally interested in a product providing guidance for selecting areas in a city in order to have higher ratios between the future income derived from renting housing space (rents) and the initial property investment.

2 Data

2.1 Data sources

2.1.1 Geographical and demographic information

The city of Madrid is subdivided into 21 districts (distritos), which are subdivided into 131 neighborhoods (barrios administrativos) [1]. To characterize districts and neighborhood geographically and demographically the following information was retrieved from [1] and [2]:

- list of districts
- list of neighborhoods
- neighborhood surface area

Demographic evolution of each neighborhood was obtained from the statistical information database of the city of Madrid [3].

2.1.2 Prices of housing for sales and housing for rent

The City of Madrid's statistical database [3] contains actual and historical data about m2/€ housing selling prices and m2/€ housing rents for the last 5 years. However, whereas selling prices are available at the neighborhood level, rents are only reported at the district level.

The real estate web portal *idealista.com* has a report section which offers the possibility to consult Madrid area housing selling prices and rents at the neighborhood level [4],[5]. Quarterly variation, year on year variation, and all times maxima are also reported. The *idealista.com* data were used since we require data at the neighborhood level.

2.1.3 Madrid neighborhood location data (Places and venues in Madrid)

The `venues/explore` endpoint in the FOURSQUARE *Places API* was used to explore and retrieve recommended venues and their categories in each of Madrid's neighborhoods [6]. The search area was defined through the longitude/latitude of each neighborhood and a search radius of 700 m. (This corresponds to a surface area of 153 ha., which corresponds to the 3rd quartile of the distribution of the surface area of Madrid's neighborhoods).

The result of each was further processed to determine the frequency of each venue category in each of Madrid's neighborhoods (see section 2.2). The neighborhood's venue category frequency was used to try to segment neighborhoods in clusters of different economic-value development.

2.2 Data preparation

2.2.1 Geographical and demographic information

2.2.1.1 Madrid districts and neighborhoods subdivision

The name of Madrid's districts and their further subdivision in neighborhoods was obtained from [2]. A dataframe with the following columns was created (Figure 1):

- Neighborhood
- District
- Neighborhood surface area

	Neighborhood	District	Area (km2)
0	Palacio	Centro	1.471
1	Embajadores	Centro	1.032
2	Cortes	Centro	0.592
3	Justicia	Centro	0.742
4	Universidad	Centro	0.947
...	...	...	...
126	Alameda de Osuna	Barajas	1.961
127	Aeropuerto	Barajas	25.132
128	Casco Histórico de Barajas	Barajas	0.609
129	Timón	Barajas	9.595
130	Corralejos	Barajas	4.633

131 rows × 3 columns

Figure 1: mad_neighborhoods - Madrid's districts and neighborhoods subdivision. There are 21 districts which are subdivided in 131 neighborhoods.

2.2.1.2 Madrid's neighborhood demographic information

Demographic evolution data of each neighborhood in the years 2018, 2019, 2020 was downloaded from the statistical information database of the city of Madrid [3], namely:

- Neighborhood name
- Surface area
- Population density for the (years 2018-2020)
- Population (years 2018-2020)

The data were cleaned to ensure consistent data format. Moreover, two additional columns with the relative change of the population density with respect to the previous year were created:

- dDensity(2020)
- dDensity(2019)

The relative change for year *n* was calculated as:

$$dDensity(n) = \frac{(Density(n) - Density(n-1))}{Density(n-1)} \times 100 \qquad (2\text{-}1)$$

where *n* is the year 2020 or 2019.

The resulting dataset were merged with the one containing Madrid's districts and neighborhood subdivision obtained from Wikipedia [2], (see Section 2.2.1.1). (Note that prior to the merge some of the neighborhood names obtained from Wikipedia needed to be aligned to the ones from City of Madrid's database and the surface area from Wikipedia was dropped). A snapshot of the resulting dataframe is shown in Figure 2.

	Neighborhood	District	Area (Ha)	Density 2018 (Inh/Ha)	Population 2018 (Inh)	Density 2019 (Inh/Ha)	Population 2019 (Inh)	Density 2020 (Inh/Ha)	Population 2020 (Inh)	dDensity 2019 rel (%)	dDensity 2020 rel (%)
0	Palacio	Centro	146.99	153.17	22515	155.95	22923	160.51	23593	1.814977	2.924014
1	Embajadores	Centro	103.37	431.74	44630	437.82	45259	455.13	47048	1.408255	3.953680
2	Cortes	Centro	59.19	177.93	10531	177.13	10484	181.98	10771	-0.449615	2.738102
3	Justicia	Centro	73.94	224.20	16578	231.98	17153	243.72	18021	3.470116	5.060781
4	Universidad	Centro	94.80	325.91	30897	334.64	31725	352.50	33418	2.678654	5.337079
...	...	...	...	...	...	...	...	...	...	...	...
126	Alameda de Osuna	Barajas	197.03	98.69	19446	99.10	19526	100.59	19820	0.415442	1.503532
127	Aeropuerto	Barajas	2962.61	0.61	1794	0.62	1851	0.64	1900	1.639344	3.225806
128	Casco Histórico de Barajas	Barajas	54.94	133.53	7336	137.70	7565	139.84	7683	3.122894	1.554103
129	Timón	Barajas	509.45	23.06	11750	24.32	12388	25.23	12853	5.464007	3.741776
130	Corralejos	Barajas	468.25	16.04	7510	16.32	7642	16.56	7754	1.745636	1.470588

131 rows × 11 columns

Figure 2: mad_demographics – Evolution of demographic figures in the 131 neighborhoods of Madrid in the years 2018-2020.

2.2.2 Prices of housing for sales and housing for rent

Data for prices of housing for sales and housing for rent in the neighborhoods of Madrid were obtained from [4] and [5], respectively. Data were cleaned to ensure a consistent data format and data for housing sell price and housing rent prices were stored in two separate datasets.

2.2.2.1 Housing rent prices

Data for housing rent prices are shown in Figure 3 and are organized in the following columns:

- District
- Neigborhood
- Rent price (EUR/m2), February 2021
- Monthly rent variation (Monthly var rent (%))
- Quarterly price variation (Quarterly var rent (%))
- Yearly price variation (Yearly var rent (%))
- Historical maximum rent (Max rent EUR/m2)
- Variation with respect to maximum price (Max var rent (%))
- Year in which maximum price was achieved (Max rent year)

	District	Neighborhood	Rent price (EUR/m2)	Monthly var rent (%)	Quarterly var rent (%)	Yearly var rent (%)	Max rent (EUR/m2)	Max var rent (–)	Max rent year
0	Centro	Justicia	17.3	-1.0	-3.7	-14.4	20.9	-17.1	2020.0
1	Centro	Cortes	16.1	-0.7	-5.9	-13.7	19.6	-17.7	2018.0
2	Centro	Embajadores	15.9	-0.6	-3.6	-17.2	19.2	-17.2	2020.0
3	Centro	Universidad	16.7	0.0	-3.2	-16.7	20.0	-16.7	2020.0
4	Centro	Palacio	16.1	-0.9	-4.8	-11.6	18.2	-11.6	2020.0
...	...	...	...	...	...	...	...	...	...
125	Villaverde	Butarque	10.3	3.1	-1.9	-3.0	16.4	-37.0	2019.0
126	Villaverde	Ángeles	10.3	-4.4	-3.7	-4.0	11.4	-9.8	2020.0
127	Villaverde	Los Rosales	10.4	-0.4	-3.8	-5.3	11.8	-11.4	2020.0
128	Villaverde	Villaverde Alto, Casco Histórico de Villaverde	10.6	-1.4	-2.6	-2.6	11.8	-10.0	2019.0
129	Villaverde	San Cristóbal	NaN	NaN	NaN	NaN	NaN	NaN	NaN

130 rows × 9 columns

Figure 3: mad_rents_feb - Variation of prices of housing for rent in Madrid's neighborhoods.

2.2.2.2 Housing sale prices

Data for housing sell prices are shown in Figure 3 and are organized in the following columns:

- District
- Neigborhood
- Sale price (EUR/m2), February 2021
- Monthly sale price variation (Monthly var sale (%))
- Quarterly sale variation (Quarterly var sale (%))

- Yearly sale price variation (Yearly var sale (%))
- Historical maximum sale price (Max sale EUR/m2)
- Variation with respect to maximum price (Max var sale (%))
- Year in which maximum price was achieved (Max sale year)

	District	Neighborhood	Sale price (EUR/m2)	Monthly var sale (%)	Quarterly var sale (%)	Yearly var sale (%)	Max sale (EUR/m2)	Max var sale (-)	Max sale year
0	Centro	Justicia	5707.000	1.2	-1.7	-1.7	6120.0	-6.7	2019.0
1	Centro	Cortes	5229.000	1.5	3.0	-2.1	5481.0	-4.6	2018.0
2	Centro	Embajadores	4162.000	0.0	-1.7	-7.3	4489.0	-7.3	2020.0
3	Centro	Universidad	5051.000	0.3	-1.0	-4.4	5497.0	-8.1	2020.0
4	Centro	Palacio	4764.000	-0.1	0.8	1.8	5073.0	-6.1	2019.0
...	...	...	...	...	...	...	...	...	...
125	Villaverde	Butarque	2.092	0.4	-1.4	-2.0	2472.0	-15.4	2010.0
126	Villaverde	Ángeles	1.764	0.6	-0.7	-2.2	2487.0	-29.1	2010.0
127	Villaverde	Los Rosales	1.851	0.4	2.2	1.5	2299.0	-19.5	2010.0
128	Villaverde	Villaverde Alto, Casco Histórico de Villaverde	1.626	1.5	0.7	0.0	2691.0	-39.6	2009.0
129	Villaverde	San Cristóbal	1.308	-1.2	-3.4	-10.0	1727.0	-24.3	2010.0

130 rows × 9 columns

Figure 4: mad_sale_feb - Variation of prices of housing for sale in Madrid's neighborhoods.

2.2.2.3 Handling of missing housing sale and housing rent data

For few neighborhoods, some housing sale price data and housing rent data entries were not available in [4] and [5], respectively and therefore needed to be estimated from the available data. The corresponding estimation methods are summarized in Table 1 and Table 2.

Neighborhood	**District**	**Missing rent data**	**Missing rent data estimation assumption**
Atocha	Arganzuela	All housing rent data	District median
Fuentelarreina	Fuencarral-El Pardo	Quarterly and yearly variation	Same as monthly variation
Cuatro Vientos	Latina	All housing rent data	Same as Campamento (Latina)
El Plantío	Moncloa-Aravaca	All housing rent data	Same as Valdezarza (Moncloa-Aravaca)
Orcasitas	Usera	quarterly and yearly variation	Same as monthly variation
Orcasur	Usera	Yearly variation	Same as maximum variation
San Fermín	Usera	Yearly variation	Same as maximum variation
Pradolongo	Usera	Yearly variation	Same as maximum variation
Fontarrón	Moratalaz	Yearly variation	Same as maximum variation
Vinateros	Moratalaz	Yearly variation	Same as maximum variation
Pavones	Moratalaz	All housing rent data	Same as Fontarrón (Moratalaz)

Horcajo	Moratalaz	All housing rent data	Same as Marroquina (Moratalaz)
Apóstol Santiago	Hortaleza	Yearly variation	Same as maximum variation
San Cristóbal	Villaverde	All housing rent data	Same as Villaverde Alto
Santa Eugenia	Villa de Vallecas	All housing rent data	Same as maximum variation
El Cañaveral	Vicálvaro	All housing rent data	Same as Valdebernardo (Vicálvaro)
Hellín	San Blas-Canillejas	All housing rent data	Same as Amposta (San Blas-Canillejas)
Amposta	San Blas-Canillejas	Yearly variation	Same as maximum variation
Arcos	San Blas-Canillejas	Yearly variation	Same as maximum variation
Atalaya	Ciudad Lineal	Quarterly variation	Same as monthly one

Table 1: Summary of missing housing rent data and corresponding assumptions for their estimation.

Neighborhood	District	Missing sale data	Missing sale data estimation assumption
Atocha	Arganzuela	District median	District median
Cuatro Vientos	Latina	Same as Campamento (Latina)	Same as Campamento (Latina)
Pavones	Moratalaz	Same as Fontarrón (Moratalaz)	Same as Fontarrón (Moratalaz)
Horcajo	Moratalaz	Same as Marroquina (Moratalaz)	Same as Marroquina (Moratalaz)
Atalaya	Ciudad Lineal	All housing sale data	District median

Table 2: Summary of missing housing sale data and corresponding assumptions for their estimation.

2.2.3 Consolidation of neighborhood demographic data with neighborhood prices data

The housing price and rent datasets presented in Sections 2.2.2.1 - 2.2.2.3 were merged with the neighborhood demographic dataset from Section 2.2.1.2.

Note that the housing for sale and housing for rent price data [4],[5] does not fully follow the official neighborhood boundaries. On one hand, few neighborhoods are missing, as they are not relevant from a real estate point of view (e.g., the neighborhood “Aeropuerto” whose boundaries are drawn around Madrid’s international airport). These neighborhoods were dropped from the consolidated demographic and prices data. On the other hand, few areas interest have separate entries although officially they are part of one of the 131 neighborhoods. Those areas were added to the consolidated prices and demographic data.

A summary of the adjustments done to consolidate prices data and demographic data into one dataset are summarized in Table 3 and Table 4. The latter also reports how the surface area and population densities for the years 2018-2020 for the added areas was estimated. (The missing absolute population data was obtained by multiplication of the surface area and the population density).

A snapshot of the consolidated prices and demographic dataset is shown in Figure 5. The dataset contains 134 neighborhoods and areas.

Neighborhood dropped from consolidated demographic and prices data		
Name	***District***	***Reason***
El Pardo	Fuencarral-El Pardo	Park area, no urban area
El Goloso	Fuencarral-El Pardo	Military base, no urban area
Aeropuerto	Barajas	Airport, no urban area
Casco Histórico de Vicálvaro	Vicálvaro	Mainly non urban area; urban area covered by Ambroz

Table 3: Neighborhoods dropped from consolidated Madrid demographic and pricing data

Areas added to consolidated demographic and prices data			
Area Name	***Neighborhood***	***Surface Area***	***Population Densities***
Pau de Carabanchel	Cuatro Vientos (Latina) and Buenavista (Carabanchel)	See [7]	Same as Cuatro Vientos
Arroyo del Fresno	Mirasierra	See [8]	Same as Mirasierra
Las Tablas	Valverde	Median of all neighborhoods	Same as Valverde
Montecarmelo	El Goloso, Mirasierra	Median of all neighborhoods	Take Mirasierra, as El Goloso is a military base
Sanchinarro	Valdefuentes	Median of all neighborhoods	Take Valdefuentes
Virgen del Cortijo - Manoteras	Valdefuentes	Median of all neighborhoods	Take Valdefuentes
Ambroz	Casco Histórico de Vicálvaro	Median of all neighborhoods	Take Valderrivas

Table 4: Neighborhoods added to consolidated Madrid demographic and pricing data.

	Neighborhood	District	Area (Ha)	Density 2018 (Inh/Ha)	Population 2018 (Inh)	Density 2019 (Inh/Ha)	Population 2019 (Inh)	Density 2020 (Inh/Ha)	Population 2020 (Inh)	dDensity 2019 rel (%)	...	Max rent (EUR/m2)	Max var rent (-)	Max rent year	Sale price (EUR/m2)	Monthly var sale (%)	Quarterly var sale (%)	Yearly var sale (%)	Max sale (EUR/m2)	Max var sale (-)	Max sale year
0	Palacio	Centro	146.99	153.17	22515.0000	155.95	22923.0000	160.51	23593.0000	1.814977	...	18.2	-11.6	2020	4764.0	-0.1	0.8	1.8	5073.0	-6.1	2019
1	Embajadores	Centro	103.37	431.74	44630.0000	437.82	45259.0000	455.13	47048.0000	1.408255	...	19.2	-17.2	2020	4162.0	0.0	-1.7	-7.3	4489.0	-7.3	2020
2	Cortes	Centro	59.19	177.93	10531.0000	177.13	10484.0000	181.98	10771.0000	-0.449615	...	19.6	-17.7	2018	5229.0	1.5	3.0	-2.1	5481.0	-4.6	2018
3	Justicia	Centro	73.94	224.20	16578.0000	231.98	17153.0000	243.72	18021.0000	3.470116	...	20.9	-17.1	2020	5707.0	1.2	-1.7	-1.7	6120.0	-6.7	2019
4	Universidad	Centro	94.80	325.91	30897.0000	334.64	31725.0000	352.50	33418.0000	2.678654	...	20.0	-16.7	2020	5051.0	0.3	-1.0	-4.4	5497.0	-8.1	2020
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
129	Las Tablas	Fuencarral-El Pardo	135.64	68.97	9355.0908	70.37	9544.9868	72.00	9766.0800	2.029868	...	13.2	-9.0	2020	4130.0	0.0	0.0	-0.9	4324.0	-4.5	2019
130	Montecarmelo	Fuencarral-El Pardo	135.64	45.84	6217.7376	47.05	6381.8620	48.69	6604.3116	2.639616	...	13.9	-10.4	2018	4373.0	0.4	-0.7	-4.6	4630.0	-5.6	2019
131	Sanchinarro	Hortaleza	135.64	32.40	4394.7360	34.54	4685.0056	36.96	5013.2544	6.604938	...	13.5	-7.6	2019	4286.0	0.7	2.3	-3.5	4538.0	-5.6	2019
132	Virgen del Cortijo - Manoteras	Hortaleza	135.64	32.40	4394.7360	34.54	4685.0056	36.96	5013.2544	6.604938	...	14.1	-12.6	2019	3678.0	-0.4	-4.5	9.2	3914.0	-6.0	2018
133	Ambroz	Vicálvaro	135.64	282.66	38340.0024	282.07	38259.9748	282.20	38277.6080	-0.208731	...	13.0	-6.3	2019	2068.0	3.3	2.3	-1.5	2468.0	-16.2	2011

134 rows × 25 columns

Figure 5: mad_demo_prices - Consolidated prices and demographic dataset for Madrid's neighborhoods and areas.

2.2.4 Madrid neighborhood location data (Places and venues in Madrid)

In order to determine the frequency of the most recommended venue categories in each of Madrid's neighborhoods the following steps were applied:

1. Determine latitude and longitude data for each neighborhood/area in the consolidated demographic prices dataset using *Geopy (https://geopy.readthedocs.io/en/stable/)* package in conjuction with *Nominatim* geocoder[1]
2. Use *FourSquare Places* API `explore` endpoint to retrieve top recommended venues in each neighborhood/area by specifying the corresponding coordinates [6]. The venues were searched in a radius of 700 m and a limit of 100 venues per neighborhood was set.
3. Create a dataset containing the retrieved venues, their categories and their neighborhoods
4. Convert venue categories from categorical variables (e.g., 'Restaurant', 'Shop') in indicator variables using the `pandas.get_dummies()` method
5. Use the `pandas.groupby().sum()` method to group by neighborhood and sum the occurrences of each venue category in each of the neighborhoods. For normalization divide each entry by the total number of unique venue categories

A snapshot of the dataset with the frequency of the venue category in each of Madrid's neighborhood is shown in Figure 6. The dataset contains 280 venue categories and 133 neighborhoods/areas (one less than the demographic data and prices dataset, as for the 'El Cañaveral" neighborhood no venue recommendation was found in the specified radius).

	Neighborhood	Accessories Store	Airport	Airport Terminal	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	...	Used Bookstore	Vegetarian / Vegan Restaurant	Venezuelan Restaurant	Video Game Store	Vietnamese Restaurant	Whisky Bar	Wine Bar	Wine Shop	Women's Store	Yoga Studio
0	Abrantes	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	Acacias	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.014286	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.003571	0.0	0.0
2	Adelfas	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
3	Alameda de Osuna	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Almagro	0.0	0.0	0.0	0.003571	0.0	0.0	0.0	0.003571	0.003571	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
128	Virgen del Cortijo - Manoteras	0.0	0.0	0.0	0.003571	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
129	Vista Alegre	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
130	Zofío	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
131	Águilas	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
132	Ángeles	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

133 rows × 281 columns

Figure 6: mad_onehot_grouped – Frequency of venue categories in Madrid's neighborhoods according to venues recommended by FourSquare users

Finally, by sorting each row of the venue category frequency dataset, a dataset containing the top venue category in each neighborhood can obtained (see

[1] Plotting the locations on a map using the longitude and latitude retried from *Nominatim*, showed that in few cases the returned location was not within the City of Madrid. The coordinates of these locations were obtained from http://www.google.com/maps.

	Neighborhood	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	...	11th Most Common Venue	12th Most Common Venue
0	Abrantes	Pizza Place	Soccer Field	Bakery	Gym / Fitness Center	Fast Food Restaurant	Athletics & Sports	Restaurant	Yoga Studio	Fish & Chips Shop	...	Farm	Farmers Market
1	Acacias	Spanish Restaurant	Bar	Art Gallery	Tapas Restaurant	Coffee Shop	Pizza Place	Theater	Restaurant	Market	...	Pub	Café
2	Adelfas	Bar	Café	Grocery Store	Fast Food Restaurant	Spanish Restaurant	Bakery	Supermarket	Hotel	Gym	...	Sandwich Place	Farmers Market
3	Alameda de Osuna	Spanish Restaurant	Plaza	Bakery	Gym	Restaurant	Smoke Shop	Hotel	Bistro	Bar	...	Fried Chicken Joint	Scenic Lookout
4	Almagro	Restaurant	Spanish Restaurant	Plaza	Bar	Hotel	Japanese Restaurant	Coffee Shop	Mediterranean Restaurant	Italian Restaurant	...	Lounge	French Restaurant
...	...	...	...	...	...	...	...	...	...	...	...	...	...
128	Virgen del Cortijo – Manoteras	Spanish Restaurant	Burger Joint	Cosmetics Shop	Gastropub	Sushi Restaurant	Cafeteria	Restaurant	Bar	Mediterranean Restaurant	...	Train Station	Park
129	Vista Alegre	Grocery Store	Fast Food Restaurant	Coffee Shop	Plaza	Bakery	Pizza Place	Pub	Convenience Store	Comedy Club	...	Spanish Restaurant	Breakfast Spot
130	Zofío	Spanish Restaurant	Park	Theater	Beer Garden	Athletics & Sports	Asian Restaurant	Fish Market	Farm	Farmers Market	...	Fish & Chips Shop	Yoga Studio
131	Águilas	Bar	Spanish Restaurant	Shopping Mall	Restaurant	Gym Pool	Smoke Shop	Café	Tapas Restaurant	Convenience Store	...	Train Station	Seafood Restaurant
132	Ángeles	Bar	Restaurant	Spanish Restaurant	Grocery Store	Pet Store	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	...	Fish Market	Exhibit

133 rows × 21 columns

Figure 7: mad_venues_sorted – Most common venues categories by Madrid's neighborhood according to FourSquare users.

2.3 Feature Selection for segmentation into clusters

2.3.1 Madrid neighborhood location data (Places and venues in Madrid)

To segment Madrid's neighborhoods into clusters according to their venues and places, the dataset with the frequency of venue categories in Madrid's neighborhoods (see Section 2.2.4) was used as is.

2.3.2 Madrid neighborhood demographic and prices data

To segment Madrid's neighborhoods into clusters according to prices and demographic development the following features were selected from the demographic data and prices dataset (see Section 2.2.3):

- Population density for the year 2018
- dDensity(2020)
- dDensity(2019)
- Rent price (EUR/m2), February 2021
- Quarterly price variation (Quarterly var rent (%))
- Yearly price variation (Yearly var rent (%))
- Historical maximum rent (Max rent EUR/m2)
- Variation with respect to maximum price (Max var rent (%))
- Year in which maximum price was achieved (Max rent year)
- Sale price (EUR/m2), February 2021
- Quarterly sale price variation (Quarterly var sale (%))
- Yearly sale price variation (Yearly var sale (%))
- Historical maximum sale price (Max sale EUR/m2)
- Variation with respect to maximum price (Max var sale (%))
- Year in which maximum price was achieved (Max sale year)

3 Methodology

3.1 General

The comparison of Madrid neighborhoods is based on their segmentation (identification of groups with similar characteristics) according to socio-economic development indicators (represented by the mix and frequency of venue categories present in each neighborhood) on one hand and housing sale and rent prices on the other. The segmentation was carried out automatically by using clustering algorithms.

3.1.1 Clustering Algorithm

Both types (venues categories and housing prices) of automatic segmentations were based on the K-means algorithm in the ML library *sklearn* [9].

3.1.1.1 Selection of the number of clusters

The selection of the number of clusters was based on the cluster silhouette analysis as discussed in ref [10]. Reference [10] says that:

"Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster."

A formal definition of the silhouette coefficients can be found here [11].

The silhouette score defines the mean silhouette coefficient over all samples [11].

3.1.1.2 Remaining settings

Default *sklearn* settings were kept, except `random_state`, for which different values were tried.

3.2 Segmentation of Madrid's neighborhood by venues

For segmenting Madrid's neighborhoods by venues, the dataset *mad_onehot_grouped* was used as is. The dataset lists for each neighborhood the normalized frequency of occurrence for each of the 280 unique venue categories retrieved from the FourSquare Places API. For more details on how the dataset was obtained refer to section 2.2.4. A snapshot of the dataset is given in Figure 6.

K-means was run on the dataset with number of clusters varying in the range 2-9. The silhouette score (mean across all samples) was generally low (see Figure 8). This indicates that clusters overlap and cannot be well separated.

```
For n_clusters = 2 The average silhouette_score is : 0.4575156623177431
For n_clusters = 3 The average silhouette_score is : 0.3965370259017062
For n_clusters = 4 The average silhouette_score is : 0.3953164879059456
For n_clusters = 5 The average silhouette_score is : 0.2993940482929105
For n_clusters = 6 The average silhouette_score is : 0.3009017864106834
For n_clusters = 7 The average silhouette_score is : 0.2162211396208524
For n_clusters = 8 The average silhouette_score is : 0.23628576032091184
For n_clusters = 9 The average silhouette_score is : 0.1984208415248935
```

Figure 8: Silhouette score for increasing number of cluster for K-means runs on the mad_onehot_grouped dataset. Random_state was set to 1.

As a general trend, the silhouette score decreases with increasing number of clusters. However, local improvements can be observed when going from 5 to 6 clusters, or from 8 to 9.

Figure 9 shows the silhouette plots for different numbers of clusters. When the number of clusters is 2, the resulting clusters are very uneven in thickness. The thicker cluster is likely to contain neighborhoods characterized by a lower venue density and therefore difficult to characterize. Up to a certain point, the increase of the number of clusters leads to a split of the secondary cluster in a series of smaller cluster which are comparable in thickness and accuracy without jeopardizing to match the number of misclassified examples (silhouette score < 0).

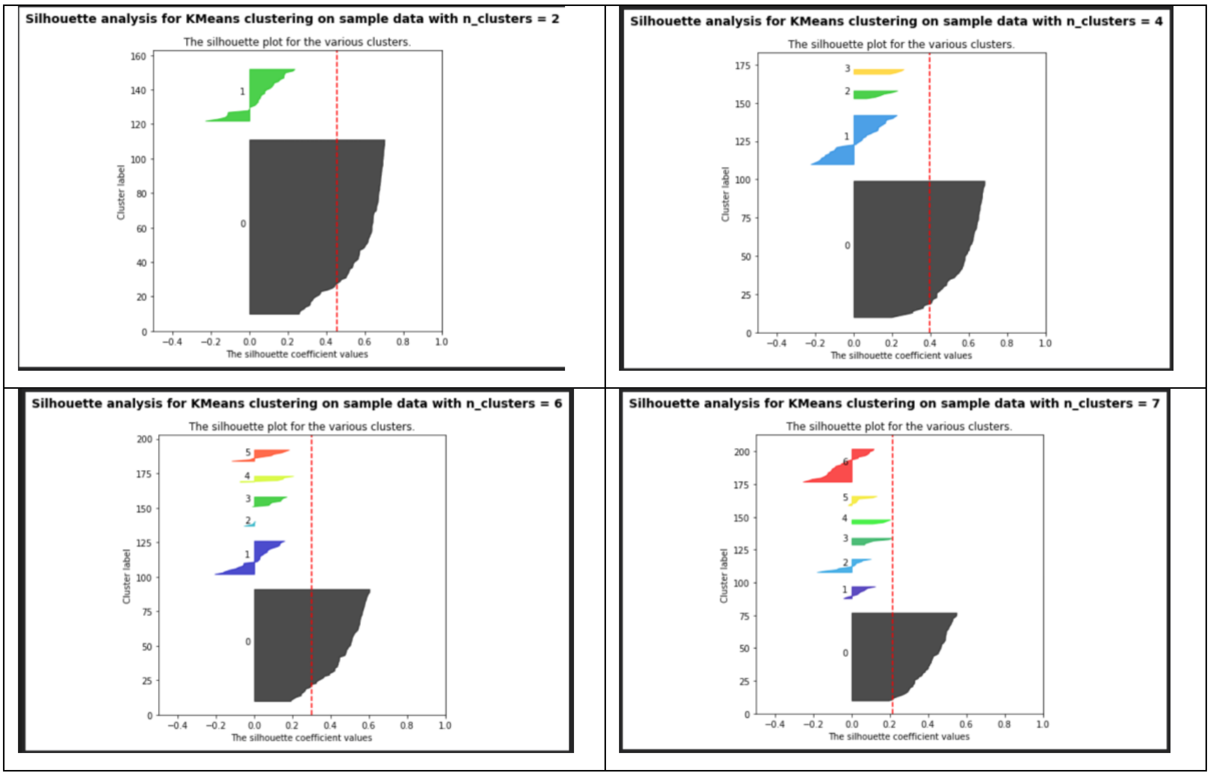


Figure 9: Silhouette plots for increasing number of cluster for K-means runs on the mad_onehot_grouped dataset. Random_state was set to 1.

Based on the arguments of homogenous silhouette thickness and silhouette accuracy, as well as local improvement of the silhouette score, the cluster number was set to 6. The highest score was achieved for `random_state` = 0. The corresponding silhouette plot and score are given in Figure 10 and the number of neighborhood assigned to each cluster, as well as the average number of venue per cluster is shown in Figure 11.

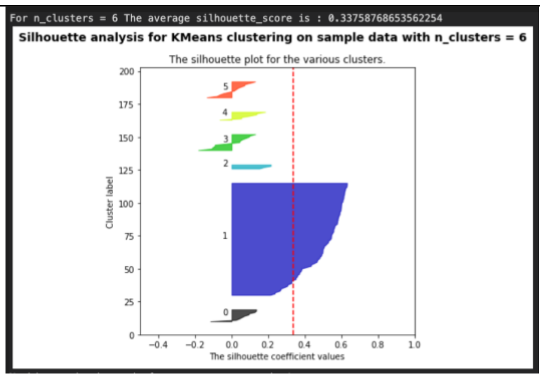


Figure 10: Silhoutte plot and silhouette score for selected neighborhood segmentation by venues (mad_onehot_grouped) dataset. Random_state was set to 0.

```
Number of neigborhoods per cluster:
Label
0     10
1     86
2      4
3     13
4      7
5     13
```

```
Average number of venues per cluster:
Label
0      90.800000
1      14.813953
2     100.000000
3      57.923077
4      97.428571
5      58.692308
```

Figure 11: Number of neighborhoods assigned to each cluster (left) and average number of venues per cluster (right).

The numbers in Figure 11 confirm that the biggest cluster groups the neighborhood with low venue density (compare left and right for cluster 1). Moreover, form the silhouette plot it can be inferred that the clusters with the highest misclassification rates are the ones with intermediate venue density, i.e., clusters 3 and 5.

3.3 Segmentation of Madrid neighborhood by housing prices and demographic development

In order to segment Madrid's neighborhoods by housing prices, the dataset mad_demo_prices was used. The dataset contains combined demographic information with price data for housing sale and rent for each neighborhood. For more details on how the dataset was obtained and details about its features refer to section 2.2.3 A snapshot of the dataset is given in Figure 5. Before running the K-means algorithm, the mad_demo_prices features were analyzed for correlations and trends to identify which feature to keep for the segmentation step.

3.3.1 Feature correlations and selection

First a boxplot of the price of housing for rent and sale in Madrid's neighborhood grouped by district is presented (see Figure 12):

- For both rent and sale prices, three distinct groups can be identified (high, medium, low)
- The price variation within a district can be significant
- Rent and sale price across districts are qualitatively very similar

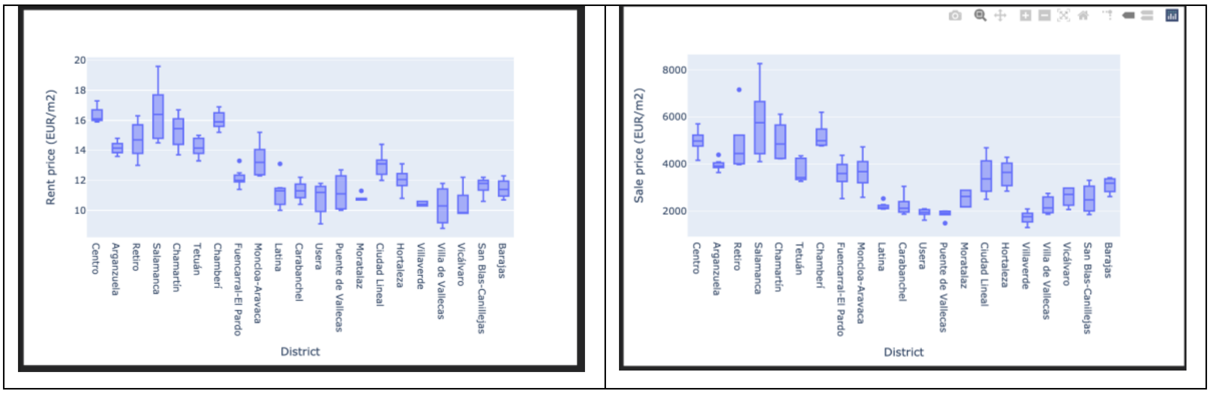


Figure 12: Boxplot of price of housing for rent (left) and sale (right) in Madrid's neighborhood grouped by district.

The Pearson correlation coefficients of the features *sale price* and *rent price* with remaining demographic and pricing features are shown in Figure 13. The correlation coefficients were grouped in 6 bins. For the current rent prices, pronounced linear correlations were found with the historical maximum rent prices (*max rent*), current sale price (*sale price*), and historical maximum sale price (*max sale*). In addition, current sale prices show strong correlation with the year in which the historical maximum was achieved (*max sale year*). The linear correlations were confirmed by corresponding scatter matrix plots.

```
Correlation coefficient for column: Rent price (EUR/m2)

Empty DataFrame
Columns: [Rent price (EUR/m2), bin]
Index: []

Empty DataFrame
Columns: [Rent price (EUR/m2), bin]
Index: []

                          Rent price (EUR/m2)          bin
Population 2018 (Inh)               -0.177366  neg[0.33-0]
Population 2019 (Inh)               -0.180840  neg[0.33-0]
Population 2020 (Inh)               -0.182688  neg[0.33-0]
dDensity 2019 rel (%)               -0.197200  neg[0.33-0]
dDensity 2020 rel (%)               -0.120766  neg[0.33-0]
Quarterly var rent (%)              -0.037248  neg[0.33-0]
Yearly var rent (%)                 -0.189957  neg[0.33-0]
Max var rent (%)                    -0.201412  neg[0.33-0]
Max rent year                       -0.019345  neg[0.33-0]
Monthly var sale (%)                -0.054663  neg[0.33-0]
Quarterly var sale (%)              -0.126154  neg[0.33-0]

                        Rent price (EUR/m2)       bin
Monthly var rent (%)               0.044844  [0-0.33]
Yearly var sale (%)                0.027204  [0-0.33]

                         Rent price (EUR/m2)          bin
Density 2018 (Inh/Ha)               0.390454  [0.33-0.67]
Density 2019 (Inh/Ha)               0.388187  [0.33-0.67]
Density 2020 (Inh/Ha)               0.386228  [0.33-0.67]
Max var sale (%)                    0.491045  [0.33-0.67]
Max sale year                       0.570527  [0.33-0.67]

                       Rent price (EUR/m2)       bin
Rent price (EUR/m2)               1.000000  [0.67-1]
Max rent (EUR/m2)                 0.963100  [0.67-1]
Sale price (EUR/m2)               0.889939  [0.67-1]
Max sale (EUR/m2)                 0.889687  [0.67-1]
```

```
Correlation coefficient for column: Sale price (EUR/m2)

Empty DataFrame
Columns: [Sale price (EUR/m2), bin]
Index: []

Empty DataFrame
Columns: [Sale price (EUR/m2), bin]
Index: []

                          Sale price (EUR/m2)          bin
Population 2018 (Inh)               -0.289015  neg[0.33-0]
Population 2019 (Inh)               -0.292519  neg[0.33-0]
Population 2020 (Inh)               -0.295790  neg[0.33-0]
dDensity 2019 rel (%)               -0.223340  neg[0.33-0]
dDensity 2020 rel (%)               -0.194331  neg[0.33-0]
Quarterly var rent (%)              -0.000902  neg[0.33-0]
Yearly var rent (%)                 -0.052091  neg[0.33-0]
Max var rent (%)                    -0.109653  neg[0.33-0]
Max rent year                       -0.145818  neg[0.33-0]
Quarterly var sale (%)              -0.036182  neg[0.33-0]

                         Sale price (EUR/m2)       bin
Density 2018 (Inh/Ha)               0.154535  [0-0.33]
Density 2019 (Inh/Ha)               0.149574  [0-0.33]
Density 2020 (Inh/Ha)               0.144174  [0-0.33]
Monthly var rent (%)                0.038241  [0-0.33]
Monthly var sale (%)                0.054299  [0-0.33]
Yearly var sale (%)                 0.154019  [0-0.33]

                   Sale price (EUR/m2)          bin
Max var sale (%)              0.637059  [0.33-0.67]

                      Sale price (EUR/m2)       bin
Rent price (EUR/m2)              0.889939  [0.67-1]
Max rent (EUR/m2)                0.838445  [0.67-1]
Sale price (EUR/m2)              1.000000  [0.67-1]
Max sale (EUR/m2)                0.982338  [0.67-1]
Max sale year                    0.681118  [0.67-1]
```

Figure 13: Pearson correlation coefficients of price of housing for rent (left) and price of housing for sale (right)

The scatter matrix plots indicated the following additional correlations for housing rent:

a) *Max var rent* vs *Yearly var rent*: (correlation coefficient = 0.75, p-value =7.15e-25)
(As a side note, values on the diagonal of the Yearly var rent vs Max var rent plot indicate that the variation from the maximum historical rent is the same as the yearly variation, i.e. the max historical rent occurred during the last year. Hence the closer we are to the diagonal, the higher the amount of price decline from historic maximum occurred during the last year)

The scatter matrix plots indicated the following additional correlations for housing sale:

b) *Max var sale (%)* vs *Max sale (EUR/m2)* (corr coeff = 0.48, p-value=4.44e-09)
(The higher the max historical sale price, the less the actual sale price decreased from the historical max)
c) *Sale price (EUR/m2)* vs *Max var sale (%)* (corr coeff = 0.64, p-value=2.18e-16)
(The higher the sale price, the less the actual sale price decreased from max)
d) Highest historical max sale prices peaked in 2020

The scatter matrix plots indicated that following variables are not correlated

e) *Sale price (EUR/m2)* vs *Max var rent (%)* (corr coeff = -0.11)
High sale price does not protect from decrease in rent (as it is the case for sale price, see b))
f) *Rent price (EUR/m2)* vs *Max var rent(%)* (corr coeff = -0.20)
High rent price, does not protect from decrease in rent price (as it is the case for sale price, see c))
g) In general rent prices had maximum in 2018, 2019, and 2020, irrespective of price
h) Yearly var sale (%) and Max var sale (%): (corr coeff = 0.39)

With respect to demographic data, strong correlation between population density changes in the years 2019 and 2020 exists:

i) *dDensity 2019* vs *dDensity 2020* (corr coeff = 0.81, p-value = 1.8e-32)
(Demographic trends in 2019 and 2020 were the same)

Feature selection:

In summary, as far as features related to prices of housing for rent are concerned the following features were kept:

- Max rent year
- Rent price (EUR/m2)
- Yearly var rent (%)
- Max var rent (EUR/m2)

By analogy the following features related to prices of housing for sale were considered:

- Max sale year
- Sale price (EUR/m2)
- Yearly var sale (%)
- Max var sale (EUR/m2)

The following sales related features were dropped:

- Monthly var sale

As far as demographic data are concerned the following features were kept:

- dDensity 2019
- dDensity 2020

The following demographic features were dropped:

- Population 2018
- Population 2019
- Population 2020
- Density 2018
- Density 2019
- Density 2020

In summary segmentation of the *mad_prices_dataset* was based on the following features:

- Max rent year
- Rent price (EUR/m2)
- Yearly var rent (%)
- Max var rent (EUR/m2)
- Max sale year
- Sale price (EUR/m2)
- Yearly var sale (%)
- Max var sale (EUR/m2)
- dDensity 2019
- dDensity 2020

3.3.2 Segmentation

Prior to segmentation, the features were normalized using *sklearn StandardScaler* [12].

The impact of varying cluster number the range 2-16 was assessed by means of silhouette analysis. The corresponding silhouette score (mean across all samples) are shown in Figure 14. The scores are generally low, indicating possible overlap. As a general trend, the silhouette score decreases with increasing number of clusters. However, local improvements can be observed when increasing the number of cluster from 2 to 3, from 6 to 7, from 13 to 14, and from 15 to 16. Silhouette plots for 2, 3, 7, and 14 clusters are shown in

```
For n_clusters = 2 The average silhouette_score is : 0.29147571864548405
For n_clusters = 3 The average silhouette_score is : 0.3025612848662443
For n_clusters = 4 The average silhouette_score is : 0.24723794928026013
For n_clusters = 5 The average silhouette_score is : 0.2560699235290197
For n_clusters = 6 The average silhouette_score is : 0.2246944979759733
For n_clusters = 7 The average silhouette_score is : 0.23244394013165767
For n_clusters = 8 The average silhouette_score is : 0.19459176279671875
For n_clusters = 9 The average silhouette_score is : 0.1718531250099079
For n_clusters = 10 The average silhouette_score is : 0.17075527429506449
For n_clusters = 11 The average silhouette_score is : 0.1703206805469778
For n_clusters = 12 The average silhouette_score is : 0.17401178852813645
For n_clusters = 13 The average silhouette_score is : 0.16322765211947526
For n_clusters = 14 The average silhouette_score is : 0.1831600808970761
For n_clusters = 15 The average silhouette_score is : 0.1743451180396892
For n_clusters = 16 The average silhouette_score is : 0.18049787409548732
```

Figure 14: Silhouette score for increasing number of cluster for K-means runs on the selected features of the mad_demo_prices dataset. Random_state was set to 1.

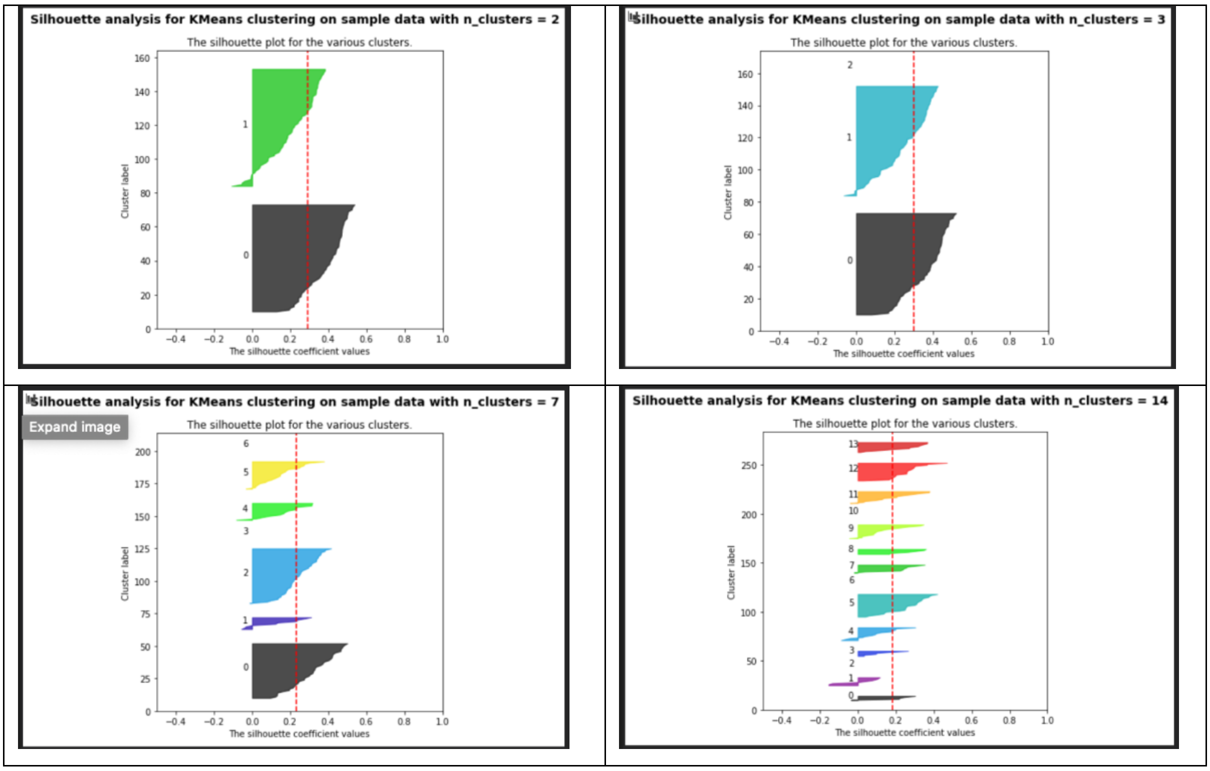


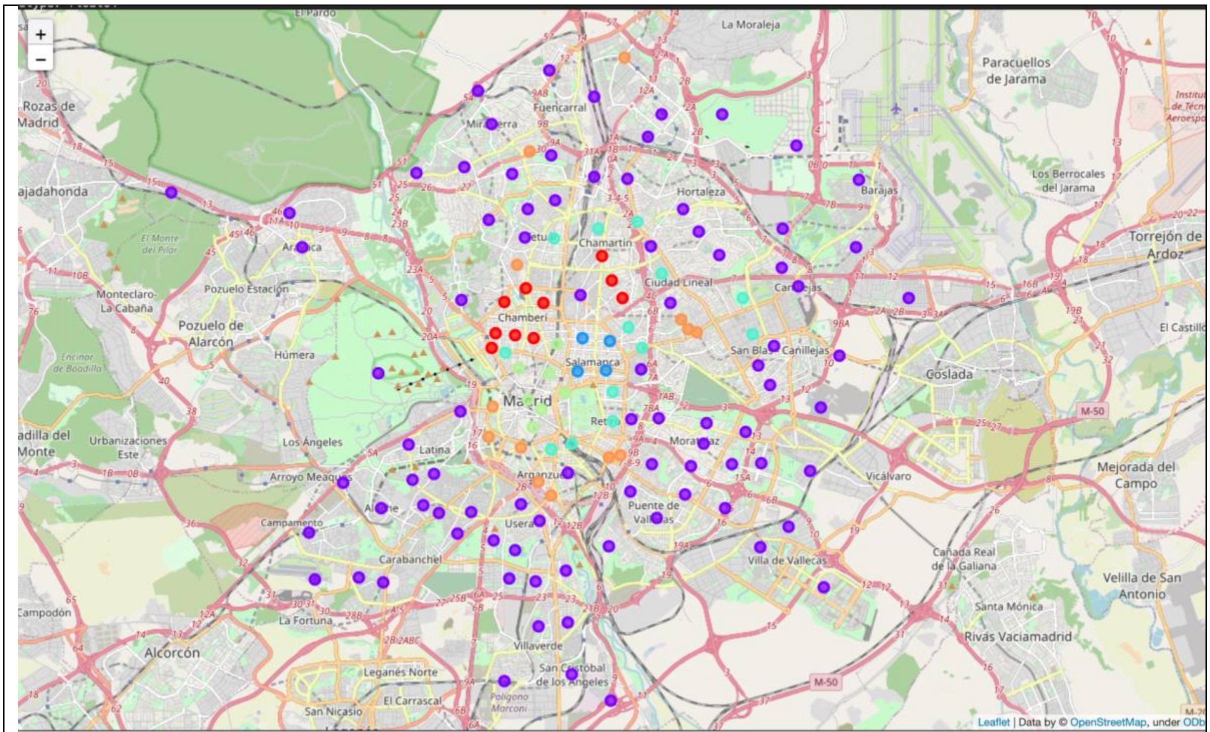
Figure 15: Silhouette plots for increasing number of cluster for K-means runs on the selected features of mad_demo_prices dataset. Random_state was set to 1.

The highest silhouette score is achieved for 3 clusters (thereby cluster 2 contains only one "outlier sample"). However, this result is in contrast with the box plots of rent and sale prices shown in Figure 12, which seem to indicate the existence of at least three groups with larger number of samples. Hence it was decided that a higher number of clusters would better represent the dataset (also considering that the additional uncorrelated features considered for segmentation next to price and rent are expected to increase the number of clusters). Based on the arguments of homogenous silhouette thickness as well as local improvement of the silhouette score, the cluster number was set to 14.

4 Results

4.1 Segmentation of Madrid’s neighborhoods by venue

The segmentation of Madrid’s neighborhood by venue is illustrated in Figure 16. As already noted, 60% of the neighborhoods are grouped in the large cluster 0 and surrounds the city. The remaining clusters are closer in size and concentrated inside the city. A characterization of the clusters is provided below.




```
Number of neigborhoods per cluster:
Label
0     10
1     86
2      4
3     13
4      7
5     13
```

```
Average number of venues per cluster:
Label
0      90.800000
1      14.813953
2     100.000000
3      57.923077
4      97.428571
5      58.692308
```

Figure 16: Segmentation of Madrid’s neighborhoods by venues. Cluster 0 (●), Cluster 1 (●), Cluster 2 (●), Cluster 3 (●), Cluster 4 (●), Cluster 5 (●).

Cluster 1 (Low density of interesting venues)

Largest cluster (86 neighborhoods) characterized by neighborhoods with low venue density (and/or small number of FourSquare recommendations). It is therefore difficult to characterize them in terms of venue category.

Cluster 2 (Classy neighborhoods)

Smallest cluster (4 neighborhoods) characterized by the presence of many (more expensive) restaurants of different fashion cuisines, luxury shopping, and hotels (see Figure 17). These neighborhoods correspond to the classiest in Madrid.

Cluster 3 (Pulse of the city)

Cluster with 13 neighborhoods characterized by restaurants, hotels, grocery shops/markets, shopping, art/museums/scenery, and nightlife (see Figure 18)

Cluster 4 (Historic, touristic, nightlife)

Cluster with 7 neighborhoods characterized by hotels, art/museum/scenery, restaurant, nightlife, and shopping (see Figure 19). These neighborhoods encompass the historic part of the city and people reward the squares and buildings found in them. They have a more “touristic” flavor. Note that in contrast to cluster 3, no general convenience good shops or grocery stores are listed.

Cluster 5 (Lively neighborhood / Outdoors)

Cluster with 13 neighborhoods characterized by restaurants, markets/supermarket/grocery health/fitness, garden/park/plaza and shopping (see Figure 20). People seem to value the presence of good restaurants, possibility to shop quality convenience goods, and good shopping opportunities. The neighborhoods also seem to be also highly appreciated for the recreational activities in parks and garden and for offering good opportunity for sport and fitness. These cluster of neighborhoods seems to be valued for its high quality of life.

Cluster 0 (Vibrant neighborhoods)

Cluster with 10 neighborhoods characterized by cafés, restaurants, grocery stores, nightlife, spas (see Figure 21). This cluster of neighborhoods appears to be valued for being a vibrant area by young adults.

```
Mediterranean Restaurant    4
Tapas Restaurant            4
Spanish Restaurant          4
Bakery                      4
Restaurant                  4
Hotel                       3
Coffee Shop                 3
Indian Restaurant           3
Japanese Restaurant         3
Burger Joint                3
Snack Place                 2
Gym                         2
Boutique                    2
Clothing Store              2
Shoe Store                  2
Café                        2
Plaza                       2
Breakfast Spot              2
Italian Restaurant          2
Seafood Restaurant          2
dtype: int64

mad_venues_cluster_dict['venue_cluster_2'].iloc[:,1:20].stack().value_counts()[20:40]

Mexican Restaurant        2
Ice Cream Shop            2
Sporting Goods Shop       1
Paella Restaurant         1
Spa                       1
Asian Restaurant          1
Cocktail Bar              1
Brewery                   1
Department Store          1
Bar                       1
Furniture / Home Store    1
Smoothie Shop             1
Supermarket               1
Dessert Shop              1
French Restaurant         1
Bookstore                 1
Grocery Store             1
Jewelry Store             1
Art Gallery               1
dtype: int64
```

Figure 17: Sorted total count of unique venues categories for cluster 2 (4 neighborhoods)

```
Mediterranean Restaurant    4
Tapas Restaurant            4
Spanish Restaurant          4
Bakery                      4
Restaurant                  4
Hotel                       3
Coffee Shop                 3
Indian Restaurant           3
Japanese Restaurant         3
Burger Joint                3
Snack Place                 2
Gym                         2
Boutique                    2
Clothing Store              2
Shoe Store                  2
Café                        2
Plaza                       2
Breakfast Spot              2
Italian Restaurant          2
Seafood Restaurant          2
dtype: int64
```

```
mad_venues_cluster_dict['venue_cluster_2'].iloc[:,1:20].stack().value_counts()[20:40]
```

```
Mexican Restaurant        2
Ice Cream Shop            2
Sporting Goods Shop       1
Paella Restaurant         1
Spa                       1
Asian Restaurant          1
Cocktail Bar              1
Brewery                   1
Department Store          1
Bar                       1
Furniture / Home Store    1
Smoothie Shop             1
Supermarket               1
Dessert Shop              1
French Restaurant         1
Bookstore                 1
Grocery Store             1
Jewelry Store             1
Art Gallery               1
dtype: int64
```

Figure 18: Sorted total count of unique venues categories for cluster 3 (13 neighborhoods)

```
mad_venues_cluster_dict['venue_cluster_4'].iloc[:,1:20].stack().value_counts()[0:40]
```

```
Plaza                       7
Restaurant                  7
Hotel                       7
Café                        7
Spanish Restaurant          6
Mediterranean Restaurant    6
Bookstore                   5
Tapas Restaurant            5
Theater                     5
Bar                         5
Coffee Shop                 5
Hostel                      4
Bistro                      3
Bakery                      3
Cocktail Bar                3
Italian Restaurant          3
Japanese Restaurant         3
Pizza Place                 3
Art Gallery                 2
Cosmetics Shop              2
Art Museum                  2
Deli / Bodega               2
Breakfast Spot              2
Argentinian Restaurant      2
Jazz Club                   2
Rental Service              1
Beer Bar                    1
Cheese Shop                 1
BBQ Joint                   1
Park                        1
Asian Restaurant            1
Pastry Shop                 1
Other Nightlife             1
Nightclub                   1
History Museum              1
Supermarket                 1
Seafood Restaurant          1
Miscellaneous Shop          1
Boutique                    1
Gourmet Shop                1
dtype: int64
```

Figure 19: Sorted total count of unique venues categories for cluster 4 (7 neighborhoods)

```
Spanish Restaurant          13
Tapas Restaurant            13
Bar                         12
Coffee Shop                  9
Supermarket                  9
Restaurant                   9
Burger Joint                 9
Brewery                      8
Bakery                       8
Café                         7
Grocery Store                6
Clothing Store               6
Pizza Place                  5
Park                         5
Fast Food Restaurant         5
Sporting Goods Shop          5
Italian Restaurant           4
Market                       4
Gym                          4
Soccer Field                 3
Art Gallery                  3
Taco Place                   3
Snack Place                  3
Garden                       3
Plaza                        3
Gastropub                    3
Mediterranean Restaurant     3
Food & Drink Shop            3
Breakfast Spot               3
Peruvian Restaurant          3
Sushi Restaurant             3
Indian Restaurant            3
Spa                          3
Beer Garden                  3
Sandwich Place               3
Hotel                        3
Farmers Market               3
Chinese Restaurant           3
Pool                         2
Mexican Restaurant           2
dtype: int64
```

Figure 20: Sorted total count of unique venues categories for cluster 5 (13 neighborhoods)

```
mad_venues_cluster_dict['venue_cluster_0'].iloc[:,1:20].stack().value_counts()[0:20]

Café                        10
Restaurant                  10
Bar                         10
Tapas Restaurant            10
Spanish Restaurant          10
Bakery                       9
Supermarket                  7
Pizza Place                  7
Italian Restaurant           6
Sandwich Place               6
Grocery Store                5
Coffee Shop                  5
Burger Joint                 4
Mediterranean Restaurant     4
Hotel                        4
Ice Cream Shop               4
Plaza                        4
Japanese Restaurant          4
Seafood Restaurant           4
Diner                        3
dtype: int64
```

```
mad_venues_cluster_dict['venue_cluster_0'].iloc[:,1:20].stack().value_counts()[21:40]

Theater                      3
Pub                          3
Cocktail Bar                 2
Sports Bar                   2
Middle Eastern Restaurant    2
Soccer Field                 2
Mexican Restaurant           2
Gastropub                    2
Food & Drink Shop            2
Multiplex                    2
Spa                          2
Deli / Bodega                2
Cheese Shop                  1
Beer Garden                  1
Concert Hall                 1
Fried Chicken Joint          1
South American Restaurant    1
Pastry Shop                  1
Furniture / Home Store       1
dtype: int64
```

Figure 21: Sorted total count of unique venues categories for cluster 0 (10 neighborhoods)

4.2 Segmentation of Madrid’s neighborhoods by housing prices and demographic data

The segmentation of Madrid’s neighborhood by housing prices and demographic development is illustrated on map in Figure 22; the cluster means for selected features sorted by sale price are given in Figure 23.

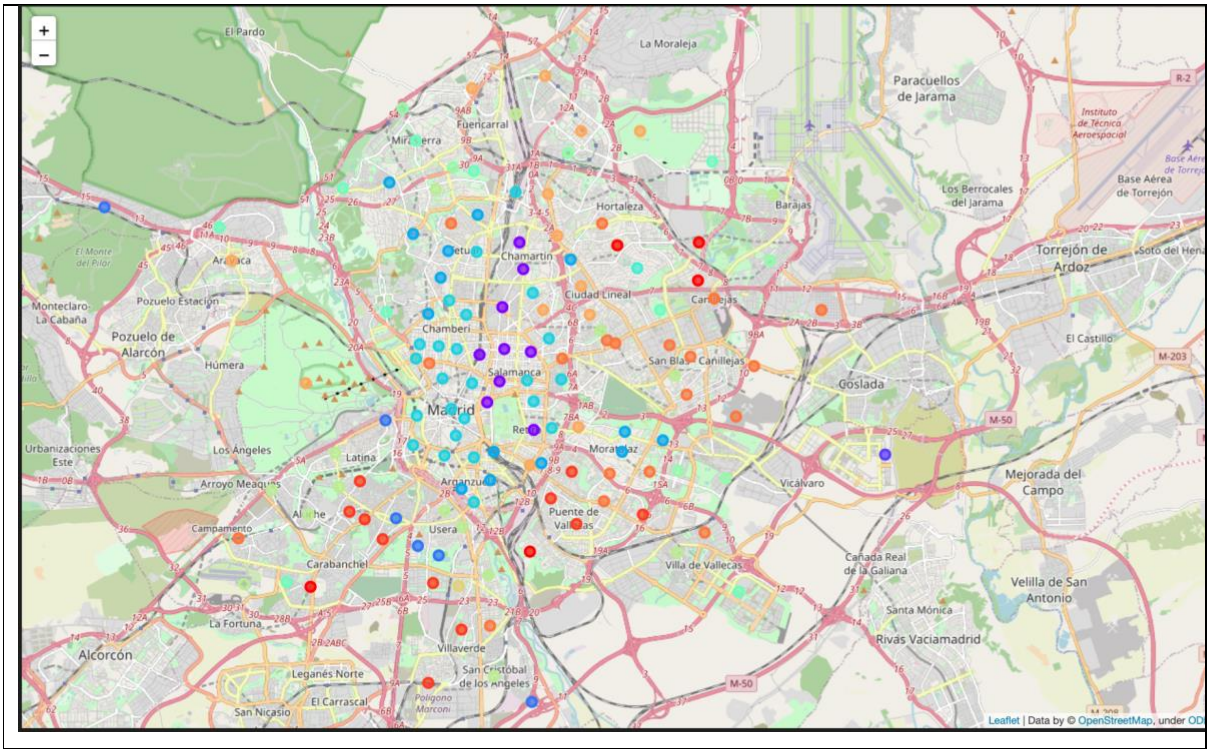


Figure 22: Segmentation of Madrid's neighborhoods by housing prices (sale and rent) and demographic data. Main clusters: Cluster 0 (●), Cluster 1 (●), Cluster 3 (●), Cluster 5 (●), Cluster 4 (●), Cluster 12 (●), Cluster 13 (●), Cluster 11 (●).

labels_both	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Sale price (EUR/m2)	Rent price (EUR/m2)	Yearly var sale (%)	Yearly var rent (%)	Max var sale (%)	Max var rent (%)
1	0.400546	0.791272	6219.777778	16.755556	-0.766667	-5.477778	-4.644444	-11.177778
5	0.982210	1.911839	4649.125000	15.283333	-1.700000	-12.183333	-6.404167	-15.020833
6	1.279591	0.189514	4257.000000	13.100000	0.900000	-7.500000	-18.200000	-10.000000
11	1.718085	2.138180	3890.423077	12.907692	-2.407692	-3.023077	-6.042308	-9.700000
4	0.772574	1.790371	3585.500000	13.075000	2.039286	-14.132143	-4.125000	-16.746429
7	2.221896	2.470013	3438.000000	12.400000	1.555556	2.933333	-12.077778	-4.588889
8	1.735670	2.341357	3328.500000	10.883333	4.166667	-7.733333	-1.350000	-12.050000
0	1.261142	1.442458	2730.600000	11.220000	-2.900000	0.960000	-17.220000	-13.600000
2	62.921348	56.551724	2444.000000	9.800000	2.200000	-8.100000	0.000000	-12.200000
12	1.135459	1.833705	2435.736842	11.484211	-0.121053	-7.273684	-13.300000	-10.421053
3	2.033762	2.455455	2224.666667	11.666667	-3.016667	-15.800000	-19.266667	-23.350000
9	1.800652	2.336828	2188.000000	11.446667	-8.780000	-7.126667	-18.486667	-11.273333
10	1.737708	2.501758	2101.000000	9.900000	17.500000	-12.600000	-14.800000	-12.600000
13	2.019389	3.042367	1894.909091	10.872727	-2.936364	-7.854545	-29.463636	-11.300000

Figure 23: Cluster means for segmentation of Madrid neighborhoods by housing prices and demographic development.

Table 5 summarizes the inter-cluster statistics for each cluster. The clusters are listed in order of decreasing mean sale price and a brief summary of the main differences compared to the preceding cluster is provided below each cluster's summary statistics.

<table>
<tr><th>Cluster</th><th>Cluster statistics</th></tr>
<tr><td>C1</td><td>
<table>
<tr><th></th><th>dDensity 2019 rel (%)</th><th>dDensity 2020 rel (%)</th><th>Rent price (EUR/m2)</th><th>Max rent year</th><th>Max rent (EUR/m2)</th><th>Yearly var rent (%)</th><th>Max var rent (%)</th><th>Sale price (EUR/m2)</th><th>Max sale year</th><th>Max sale (EUR/m2)</th><th>Yearly var sale (%)</th><th>Max var sale (%)</th></tr>
<tr><td>count</td><td>9.000000</td><td>9.000000</td><td>9.000000</td><td>9.000000</td><td>9.00000</td><td>9.000000</td><td>9.000000</td><td>9.000000</td><td>9.000000</td><td>9.000000</td><td>9.000000</td><td>9.000000</td></tr>
<tr><td>mean</td><td>0.400546</td><td>0.791272</td><td>16.755556</td><td>2019.222222</td><td>18.90000</td><td>-5.477778</td><td>-11.177778</td><td>6219.777778</td><td>2018.444444</td><td>6543.888889</td><td>-0.766667</td><td>-4.644444</td></tr>
<tr><td>std</td><td>0.581942</td><td>0.440997</td><td>1.266996</td><td>0.833333</td><td>1.96596</td><td>4.120005</td><td>3.648896</td><td>1041.322090</td><td>3.205897</td><td>1203.606398</td><td>4.484696</td><td>3.287899</td></tr>
<tr><td>min</td><td>-0.511991</td><td>0.108342</td><td>15.200000</td><td>2018.000000</td><td>17.00000</td><td>-10.500000</td><td>-15.800000</td><td>4815.000000</td><td>2010.000000</td><td>5036.000000</td><td>-7.100000</td><td>-8.900000</td></tr>
<tr><td>25%</td><td>0.024824</td><td>0.710322</td><td>16.100000</td><td>2019.000000</td><td>17.40000</td><td>-8.500000</td><td>-14.400000</td><td>5658.000000</td><td>2019.000000</td><td>5663.000000</td><td>-4.500000</td><td>-7.500000</td></tr>
<tr><td>50%</td><td>0.412579</td><td>0.771456</td><td>16.300000</td><td>2019.000000</td><td>18.90000</td><td>-4.800000</td><td>-11.000000</td><td>6118.000000</td><td>2019.000000</td><td>6614.000000</td><td>-0.600000</td><td>-4.500000</td></tr>
<tr><td>75%</td><td>0.917157</td><td>1.007136</td><td>16.900000</td><td>2020.000000</td><td>19.10000</td><td>-4.100000</td><td>-8.500000</td><td>6655.000000</td><td>2020.000000</td><td>7120.000000</td><td>3.100000</td><td>-1.400000</td></tr>
<tr><td>max</td><td>1.146288</td><td>1.452011</td><td>19.600000</td><td>2020.000000</td><td>23.30000</td><td>3.100000</td><td>-5.500000</td><td>8263.000000</td><td>2020.000000</td><td>8652.000000</td><td>6.100000</td><td>-0.100000</td></tr>
</table>

Highest prices for housing sale and rent
</td></tr>
</table>

C5

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	24.000000	24.000000	24.000000	24.000000	24.000000	24.000000	24.000000	24.00000	24.000000	24.000000	24.000000	24.000000
mean	0.982210	1.911839	15.283333	2019.416667	18.016667	-12.183333	-15.020833	4649.12500	2019.041667	4971.916667	-1.700000	-6.404167
std	0.914043	1.401460	1.088144	0.583592	1.509871	3.045833	2.197029	544.92866	0.750604	607.572570	2.235485	2.305094
min	-0.449615	0.206812	13.000000	2018.000000	15.400000	-19.700000	-19.800000	3838.00000	2018.000000	4128.000000	-7.300000	-11.600000
25%	0.270075	0.958625	14.725000	2019.000000	17.125000	-13.325000	-16.800000	4219.00000	2018.750000	4429.000000	-2.950000	-7.650000
50%	0.826146	1.586108	15.400000	2019.000000	18.050000	-12.100000	-15.050000	4615.00000	2019.000000	4968.500000	-1.650000	-6.500000
75%	1.447619	2.230620	16.025000	2020.000000	19.250000	-10.175000	-13.450000	5033.75000	2020.000000	5471.250000	-0.275000	-4.825000
max	3.470116	5.337079	17.300000	2020.000000	20.900000	-7.100000	-11.300000	5707.00000	2020.000000	6120.000000	3.100000	-1.200000

Diff to C1: Range for rent and sale prices lower; rent price decrease higher; population increase higher

C6

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean	1.279591	0.189514	13.1	2011.0	14.5	-7.5	-10.0	4257.0	2009.0	5203.0	0.9	-18.2
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	1.279591	0.189514	13.1	2011.0	14.5	-7.5	-10.0	4257.0	2009.0	5203.0	0.9	-18.2
25%	1.279591	0.189514	13.1	2011.0	14.5	-7.5	-10.0	4257.0	2009.0	5203.0	0.9	-18.2
50%	1.279591	0.189514	13.1	2011.0	14.5	-7.5	-10.0	4257.0	2009.0	5203.0	0.9	-18.2
75%	1.279591	0.189514	13.1	2011.0	14.5	-7.5	-10.0	4257.0	2009.0	5203.0	0.9	-18.2
max	1.279591	0.189514	13.1	2011.0	14.5	-7.5	-10.0	4257.0	2009.0	5203.0	0.9	-18.2

Diff to C5: Max rent year and max sale year unique. Rent and sale are within the range of C6

C11

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.000000	13.00000	13.000000	13.000000	13.000000
mean	1.718085	2.138180	12.907692	2019.538462	14.315385	-3.023077	-9.700000	3890.423077	2019.00000	4139.423077	-2.407692	-6.042308
std	2.383394	2.447059	1.017727	0.776250	1.292830	3.076566	2.981331	463.124616	0.57735	457.745191	1.892292	3.994572
min	-0.442478	-0.825397	11.300000	2018.000000	12.400000	-9.200000	-15.700000	2692.000000	2018.00000	2894.000000	-5.200000	-18.400000
25%	0.343763	0.584993	12.500000	2019.000000	13.300000	-5.500000	-10.500000	3678.000000	2019.00000	3944.000000	-3.500000	-5.800000
50%	0.641350	1.612903	12.600000	2020.000000	14.200000	-3.400000	-9.700000	4114.000000	2019.00000	4306.000000	-2.600000	-5.500000
75%	2.588130	2.522836	13.300000	2020.000000	15.500000	-0.100000	-7.600000	4151.000000	2019.00000	4407.000000	-1.300000	-4.500000
max	6.604938	7.006369	14.400000	2020.000000	16.000000	1.100000	-5.800000	4373.000000	2020.00000	4630.000000	1.900000	-1.300000

Diff to C5: Pop. density increase higher; rent price range below, sale price range below, rent price decrease weaker

C4

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000
mean	0.772574	1.790371	13.075000	2019.785714	15.703571	-14.132143	-16.746429	3585.500000	2018.857143	3742.321429	2.039286	-4.125000
std	0.575867	0.892963	1.470054	0.425815	1.684370	2.936744	2.950649	679.592836	2.381245	684.301992	3.147537	3.115887
min	-0.431514	-0.105063	10.800000	2019.000000	12.700000	-18.800000	-22.700000	2555.000000	2011.000000	2600.000000	-3.200000	-11.100000
25%	0.393114	1.480104	12.075000	2020.000000	14.875000	-15.275000	-18.775000	3135.250000	2019.000000	3371.500000	0.137500	-6.450000
50%	0.837066	1.641986	13.650000	2020.000000	16.300000	-13.600000	-16.250000	3425.500000	2019.000000	3598.000000	2.250000	-3.500000
75%	1.022952	2.240377	14.112500	2020.000000	16.400000	-12.100000	-14.537500	3920.250000	2019.750000	4215.875000	3.825000	-1.925000
max	1.620782	3.345229	15.200000	2020.000000	18.400000	-10.000000	-12.700000	4953.000000	2021.000000	4953.000000	9.300000	0.000000

Diff to C11: yearly var rent decrease stronger (similar C5); price and rent range similar to C11

C7

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000
mean	2.221896	2.470013	12.400000	2019.666667	12.988889	2.933333	-4.588889	3438.000000	2010.444444	3917.111111	1.555556	-12.077778
std	1.943199	2.094487	0.913783	0.500000	0.971396	4.069398	2.432306	457.813827	0.527046	510.894178	3.461254	6.102413
min	-0.051027	-0.357371	11.400000	2019.000000	12.200000	-1.700000	-7.600000	2755.000000	2010.000000	3251.000000	-5.300000	-22.300000
25%	0.940573	1.109823	11.800000	2019.000000	12.300000	1.800000	-6.700000	3056.000000	2010.000000	3603.000000	1.900000	-15.900000
50%	1.821936	2.027914	12.200000	2020.000000	12.500000	2.100000	-4.300000	3340.000000	2010.000000	3963.000000	2.200000	-13.800000
75%	2.639616	3.485654	12.300000	2020.000000	13.300000	4.000000	-3.300000	3800.000000	2011.000000	4207.000000	2.700000	-6.200000
max	5.464007	6.641760	14.200000	2020.000000	14.700000	12.400000	-0.300000	4194.000000	2011.000000	4864.000000	7.200000	-4.100000

Diff to C4: population density increase higher, rent increase compared to last year (instead of decrease), rent range similar, sale price range similar

C8

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.00000	6.000000	6.000000	6.000000	6.000000
mean	1.735670	2.341357	10.883333	2019.166667	12.416667	-7.733333	-12.050000	3328.50000	2020.166667	3381.500000	4.166667	-1.350000
std	2.594631	2.555316	1.041953	0.752773	1.188977	3.800351	0.543139	415.01747	1.169045	478.915337	2.468738	2.352658
min	-0.208731	0.046088	9.800000	2018.000000	11.200000	-12.500000	-12.600000	2978.00000	2018.000000	2978.000000	3.000000	-6.000000
25%	0.103861	0.947146	10.025000	2019.000000	11.425000	-10.125000	-12.425000	2979.25000	2020.000000	2979.250000	3.100000	-1.275000
50%	0.689097	1.309494	10.750000	2019.000000	12.250000	-8.100000	-12.200000	3202.00000	2020.500000	3212.500000	3.200000	-0.300000
75%	2.220400	2.990123	11.625000	2019.750000	13.225000	-5.550000	-11.750000	3613.75000	2021.000000	3796.000000	3.300000	0.000000
max	6.604938	7.006369	12.300000	2020.000000	14.100000	-2.200000	-11.200000	3933.00000	2021.000000	3994.000000	9.200000	0.000000

Diff to C7; similar density, similar rent price range, similar sale price, decrease in yearly rent price (instead of increase), sale price increase stronger

C0

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.000000	5.000000	5.000000
mean	1.261142	1.442458	11.220000	2018.600000	12.980000	0.960000	-13.600000	2730.600000	2010.20000	3289.200000	-2.900000	-17.220000
std	0.921545	0.767092	0.917606	0.894427	0.983362	3.194214	2.221486	916.461019	0.83666	1090.321833	2.329163	3.001166
min	-0.124177	0.640308	10.100000	2018.000000	11.800000	-2.600000	-15.800000	1490.000000	2009.00000	1871.000000	-5.100000	-20.300000
25%	0.771294	0.715474	10.600000	2018.000000	12.500000	-2.300000	-15.400000	2038.000000	2010.00000	2460.000000	-4.900000	-20.100000
50%	1.745636	1.470588	11.200000	2018.000000	12.600000	2.100000	-14.300000	3187.000000	2010.00000	3692.000000	-3.500000	-17.200000
75%	1.821936	2.027914	11.800000	2019.000000	13.900000	3.600000	-11.500000	3324.000000	2011.00000	3901.000000	-1.200000	-14.800000
max	2.091021	2.358003	12.400000	2020.000000	14.100000	4.000000	-11.000000	3614.000000	2011.00000	4522.000000	0.200000	-13.700000

Diff to C8: similar density, similar rent and sale range range, rent price decrease range weaker (or increase, as C7), however stagnant and decreasing sale prices (in contrast to C7)

C2

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean	62.921348	56.551724	9.8	2020.0	11.2	-8.1	-12.2	2444.0	2021.0	2444.0	2.2	0.0
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	62.921348	56.551724	9.8	2020.0	11.2	-8.1	-12.2	2444.0	2021.0	2444.0	2.2	0.0
25%	62.921348	56.551724	9.8	2020.0	11.2	-8.1	-12.2	2444.0	2021.0	2444.0	2.2	0.0
50%	62.921348	56.551724	9.8	2020.0	11.2	-8.1	-12.2	2444.0	2021.0	2444.0	2.2	0.0
75%	62.921348	56.551724	9.8	2020.0	11.2	-8.1	-12.2	2444.0	2021.0	2444.0	2.2	0.0
max	62.921348	56.551724	9.8	2020.0	11.2	-8.1	-12.2	2444.0	2021.0	2444.0	2.2	0.0

Diff to remaining clusters: extreme pop density increase

C12

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.000000	19.00000
mean	1.135459	1.833705	11.484211	2019.684211	12.831579	-7.273684	-10.421053	2435.736842	2009.894737	2800.789474	-0.121053	-13.30000
std	0.780222	0.850036	1.128550	0.477567	1.309815	2.776657	2.440305	440.933586	0.809303	424.637307	1.902215	4.60326
min	-0.208731	0.045992	8.800000	2019.000000	9.800000	-13.300000	-14.000000	1851.000000	2008.000000	2112.000000	-3.100000	-21.40000
25%	0.658614	1.423847	10.700000	2019.000000	11.950000	-9.050000	-11.600000	2124.500000	2009.500000	2484.000000	-1.700000	-17.00000
50%	1.023587	2.016454	11.800000	2020.000000	13.000000	-6.700000	-10.700000	2327.000000	2010.000000	2648.000000	-0.700000	-12.50000
75%	1.619174	2.243868	12.150000	2020.000000	13.600000	-5.150000	-9.700000	2759.000000	2010.000000	3166.000000	1.250000	-9.70000
max	2.481825	3.635817	14.000000	2020.000000	15.500000	-3.400000	-4.800000	3266.000000	2011.000000	3541.000000	3.600000	-6.00000

Diff to C0: yearly rent decrease range stronger, yearly sale decrease range weaker.

C3

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000	6.000000
mean	2.033762	2.455455	11.666667	2019.666667	15.316667	-15.800000	-23.350000	2224.666667	2009.666667	2768.666667	-3.016667	-19.266667
std	1.049998	2.150978	0.968848	0.516398	1.616684	7.112805	7.381531	284.753695	0.516398	432.713608	3.554950	4.907002
min	0.807309	-0.376884	10.300000	2019.000000	13.500000	-22.800000	-37.000000	1925.000000	2009.000000	2319.000000	-7.500000	-26.400000
25%	1.283760	1.611094	11.225000	2019.250000	14.275000	-19.700000	-24.975000	1996.000000	2009.250000	2429.250000	-5.725000	-22.675000
50%	1.849407	2.150641	11.550000	2020.000000	14.950000	-18.200000	-20.050000	2167.000000	2010.000000	2711.500000	-2.250000	-17.850000
75%	2.926369	2.984326	12.175000	2020.000000	16.150000	-13.850000	-19.100000	2461.750000	2010.000000	3021.500000	-1.400000	-15.800000
max	3.315881	6.114865	13.100000	2020.000000	17.900000	-3.000000	-17.600000	2590.000000	2010.000000	3410.000000	1.900000	-14.100000

Diff to C12: pop density increase higher, yearly (and max) rent decrease range stronger, yearly sale decrease range stronger

C9

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000
mean	1.800652	2.336828	11.446667	2019.400000	12.906667	-7.126667	-11.273333	2188.000000	2011.266667	2675.733333	-8.780000	-18.486667
std	0.929821	1.117646	0.952340	0.507093	1.176840	2.535594	2.479766	445.782618	3.172801	449.397131	2.775196	6.302592
min	0.255515	0.939767	10.000000	2019.000000	10.800000	-13.300000	-15.800000	1308.000000	2009.000000	1727.000000	-14.500000	-28.400000
25%	1.297898	1.578334	10.950000	2019.000000	12.250000	-7.950000	-12.700000	1911.500000	2010.000000	2496.000000	-9.850000	-22.900000
50%	1.698524	1.953212	11.500000	2019.000000	13.200000	-7.200000	-11.500000	2116.000000	2010.000000	2658.000000	-8.700000	-18.100000
75%	2.298662	2.924557	11.800000	2020.000000	13.550000	-6.050000	-9.950000	2406.000000	2010.500000	2837.500000	-6.800000	-14.250000
max	3.576168	5.266707	13.300000	2020.000000	15.100000	-2.600000	-6.000000	3032.000000	2019.000000	3474.000000	-5.200000	-9.000000

Diff to C3: same pop density increase, same rent and price level, less accentuated yearly rent price decrease, stronger yearly sale decrease

C10

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean	1.737708	2.501758	9.9	2018.0	11.4	-12.6	-12.6	2101.0	2012.0	2466.0	17.5	-14.8
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	1.737708	2.501758	9.9	2018.0	11.4	-12.6	-12.6	2101.0	2012.0	2466.0	17.5	-14.8
25%	1.737708	2.501758	9.9	2018.0	11.4	-12.6	-12.6	2101.0	2012.0	2466.0	17.5	-14.8
50%	1.737708	2.501758	9.9	2018.0	11.4	-12.6	-12.6	2101.0	2012.0	2466.0	17.5	-14.8
75%	1.737708	2.501758	9.9	2018.0	11.4	-12.6	-12.6	2101.0	2012.0	2466.0	17.5	-14.8
max	1.737708	2.501758	9.9	2018.0	11.4	-12.6	-12.6	2101.0	2012.0	2466.0	17.5	-14.8

Diff to remaining clusters: Exceptionally high yearly var sale increase

C13

	dDensity 2019 rel (%)	dDensity 2020 rel (%)	Rent price (EUR/m2)	Max rent year	Max rent (EUR/m2)	Yearly var rent (%)	Max var rent (%)	Sale price (EUR/m2)	Max sale year	Max sale (EUR/m2)	Yearly var sale (%)	Max var sale (%)
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	2.019389	3.042367	10.872727	2019.909091	12.272727	-7.854545	-11.300000	1894.909091	2009.272727	2693.363636	-2.936364	-29.463636
std	0.855000	1.254599	1.022830	0.539360	1.110937	2.650420	2.640455	176.079786	0.904534	216.193558	2.864358	6.223066
min	1.044386	1.335571	9.100000	2019.000000	9.800000	-11.900000	-15.700000	1617.000000	2008.000000	2376.000000	-10.700000	-41.100000
25%	1.251174	2.103785	10.350000	2020.000000	11.750000	-9.300000	-13.100000	1807.000000	2009.000000	2548.500000	-3.450000	-30.650000
50%	2.047211	3.073232	10.600000	2020.000000	12.500000	-8.600000	-11.400000	1938.000000	2009.000000	2691.000000	-2.600000	-28.300000
75%	2.494600	3.734527	11.300000	2020.000000	12.800000	-7.050000	-9.850000	1992.000000	2010.000000	2807.000000	-1.500000	-25.950000
max	3.469990	5.364250	12.700000	2021.000000	13.800000	-2.600000	-6.800000	2216.000000	2011.000000	3143.000000	0.000000	-21.300000

Diff to C9: max sale decrease higher, yearly max sale decrease higher

Table 5: Inter-cluster statistics for segmentation of Madrid neighborhoods by housing prices and demographic development. The clusters are listed in order of decreasing mean sale price and a brief summary of the main differences compared to the preceding cluster is provided below each cluster's summary statistics.

Finally, the box plot in Figure 24 provides a comparison of the inter-cluster sale price distributions of the resulting clusters.

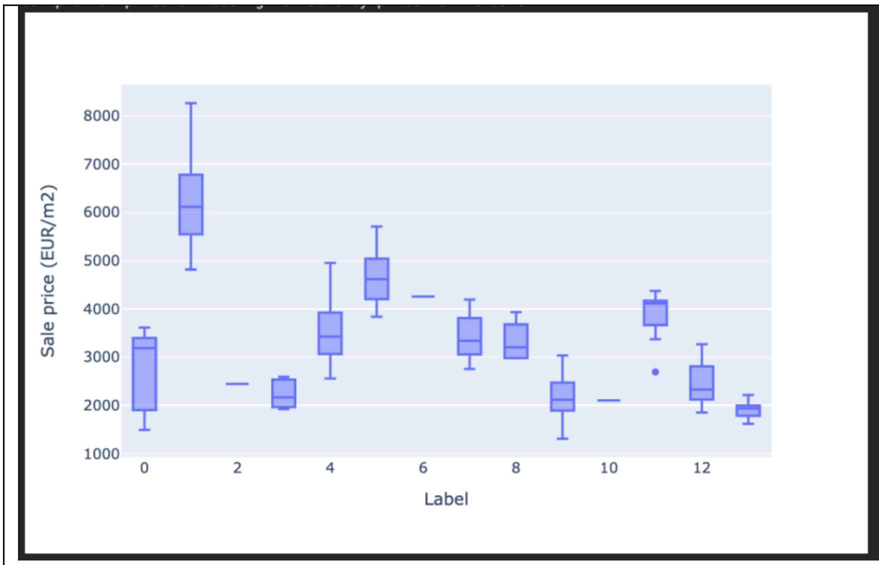


Figure 24: Box plot of price of housing for sale by cluster label

4.3 Combining the segmentation by venue with the one by housing prices and demographics.

The segmentation of the neighborhoods by venue was combined with the one for housing prices and demographic development by matching each neighborhood in a given venue cluster with the corresponding label from the housing price and demographic development segmentation. Note that given its low venue density vague characterization, venue-cluster 1 was excluded from the matching exercise. The matching between the remaining venue-clusters and housing prices clusters is given in Table 6. The labels for the venue-clusters and housing price-clusters are indicated by "Label_venue" and "Label_price", respectively.

It can be observed that neighborhoods within the same venue-cluster pertain to different housing price clusters. Within a given venue-cluster it is therefore possible to extract those neighborhoods that correspond to a housing price cluster with a lower price range. In Table 6, these neighborhoods are listed as opportunities and identified by their dataframe indices and a reference to their housing price clusters is also given.

<table>
<tr><td>Venue-cluster</td><td>Matching with housing prices and demographic development cluster</td></tr>
<tr><td>Cluster 0 (Vibrant neighborhoods)</td><td>

<code>mad_venues_prices_dict['venue_cluster_0'][['Neighborhood','District','Label_venue','Label_price']]</code>

<table>
<tr><th></th><th>Neighborhood</th><th>District</th><th>Label_venue</th><th>Label_price</th></tr>
<tr><td>11</td><td>Arapiles</td><td>Chamberí</td><td>0</td><td>5</td></tr>
<tr><td>14</td><td>Argüelles</td><td>Moncloa-Aravaca</td><td>0</td><td>5</td></tr>
<tr><td>32</td><td>Ciudad Jardín</td><td>Chamartín</td><td>0</td><td>5</td></tr>
<tr><td>40</td><td>Cuatro Caminos</td><td>Tetuán</td><td>0</td><td>5</td></tr>
<tr><td>53</td><td>Gaztambide</td><td>Chamberí</td><td>0</td><td>5</td></tr>
<tr><td>57</td><td>Hispanoamérica</td><td>Chamartín</td><td>0</td><td>1</td></tr>
<tr><td>95</td><td>Prosperidad</td><td>Chamartín</td><td>0</td><td>11</td></tr>
<tr><td>103</td><td>Ríos Rosas</td><td>Chamberí</td><td>0</td><td>5</td></tr>
<tr><td>115</td><td>Trafalgar</td><td>Chamberí</td><td>0</td><td>5</td></tr>
<tr><td>123</td><td>Vallehermoso</td><td>Chamberí</td><td>0</td><td>4</td></tr>
</table>

Price clusters with lower price range: 11, 4
Opportunities indices: 95, 123
</td></tr>
<tr><td>Cluster 2 (Classy)</td><td>

<code>mad_venues_prices_dict['venue_cluster_2'][['Neighborhood','District','Label_venue','Label_price']]</code>

<table>
<tr><th></th><th>Neighborhood</th><th>District</th><th>Label_venue</th><th>Label_price</th></tr>
<tr><td>28</td><td>Castellana</td><td>Salamanca</td><td>2</td><td>1</td></tr>
<tr><td>54</td><td>Goya</td><td>Salamanca</td><td>2</td><td>5</td></tr>
<tr><td>65</td><td>Lista</td><td>Salamanca</td><td>2</td><td>1</td></tr>
<tr><td>100</td><td>Recoletos</td><td>Salamanca</td><td>2</td><td>1</td></tr>
</table>

Price clusters with lower price range: 5
Opportunities indices: 54
</td></tr>
<tr><td>Cluster 3 (Pulse of the city)</td><td>

<code>mad_venues_prices_dict['venue_cluster_3'][['Neighborhood','District','Label_venue','Label_price']]</code>

<table>
<tr><th></th><th>Neighborhood</th><th>District</th><th>Label_venue</th><th>Label_price</th></tr>
<tr><td>16</td><td>Atalaya</td><td>Ciudad Lineal</td><td>3</td><td>11</td></tr>
<tr><td>17</td><td>Atocha</td><td>Arganzuela</td><td>3</td><td>4</td></tr>
<tr><td>30</td><td>Castillejos</td><td>Tetuán</td><td>3</td><td>5</td></tr>
<tr><td>35</td><td>Comillas</td><td>Carabanchel</td><td>3</td><td>12</td></tr>
<tr><td>44</td><td>El Salvador</td><td>San Blas-Canillejas</td><td>3</td><td>7</td></tr>
<tr><td>55</td><td>Guindalera</td><td>Salamanca</td><td>3</td><td>5</td></tr>
<tr><td>59</td><td>Ibiza</td><td>Retiro</td><td>3</td><td>5</td></tr>
<tr><td>75</td><td>Niño Jesús</td><td>Retiro</td><td>3</td><td>1</td></tr>
<tr><td>76</td><td>Nueva España</td><td>Chamartín</td><td>3</td><td>1</td></tr>
<tr><td>86</td><td>Palos de Moguer</td><td>Arganzuela</td><td>3</td><td>5</td></tr>
<tr><td>108</td><td>San Juan Bautista</td><td>Ciudad Lineal</td><td>3</td><td>11</td></tr>
<tr><td>112</td><td>Simancas</td><td>San Blas-Canillejas</td><td>3</td><td>12</td></tr>
<tr><td>125</td><td>Ventas</td><td>Ciudad Lineal</td><td>3</td><td>12</td></tr>
</table>

Price clusters with lower price range: 11,4,7,12
Opportunities indices: 16, 17, 35, 44, 108, 112, 125
</td></tr>
<tr><td>Cluster 4 (Historic, touristic, nightlife)</td><td>

<code>mad_venues_prices_dict['venue_cluster_4'][['Neighborhood','District','Label_venue','Label_price']]</code>

<table>
<tr><th></th><th>Neighborhood</th><th>District</th><th>Label_venue</th><th>Label_price</th></tr>
<tr><td>4</td><td>Almagro</td><td>Chamberí</td><td>4</td><td>1</td></tr>
<tr><td>38</td><td>Cortes</td><td>Centro</td><td>4</td><td>5</td></tr>
<tr><td>46</td><td>Embajadores</td><td>Centro</td><td>4</td><td>5</td></tr>
<tr><td>61</td><td>Justicia</td><td>Centro</td><td>4</td><td>5</td></tr>
<tr><td>67</td><td>Los Jerónimos</td><td>Retiro</td><td>4</td><td>1</td></tr>
<tr><td>113</td><td>Sol</td><td>Centro</td><td>4</td><td>5</td></tr>
<tr><td>116</td><td>Universidad</td><td>Centro</td><td>4</td><td>5</td></tr>
</table>

No opportunities identified
</td></tr>
</table>

Cluster 5 (Lively neighborhood / Outdoors)

```
mad_venues_prices_dict['venue_cluster_5'][['Neighborhood','District','Label_venue','Label_price']]
```

	Neighborhood	District	Label_venue	Label_price
1	Acacias	Arganzuela	5	5
2	Adelfas	Retiro	5	4
10	Apóstol Santiago	Hortaleza	5	8
18	Bellas Vistas	Tetuán	5	4
31	Chopera	Arganzuela	5	4
36	Concepción	Ciudad Lineal	5	9
60	Imperial	Arganzuela	5	5
63	Las Tablas	Fuencarral-El Pardo	5	11
64	Legazpi	Arganzuela	5	5
81	Pacífico	Retiro	5	11
82	Palacio	Centro	5	5
96	Pueblo Nuevo	Ciudad Lineal	5	12
99	Quintana	Ciudad Lineal	5	12

Price clusters with lower price range: 4, 8, 9, 11, 12

Opportunities indices: 2, 10, 18, 31, 36, 63, 81, 96, 99

Table 6: Matching of venue-clusters with housing prices and demographic development cluster. For the venue-clusters considered each neighborhood is matched with the corresponding housing prices and demographic development cluster label (Label price)

An overview of the neighborhoods identified as preferred opportunity for real estate investment is given in Figure 25. The list of neighborhoods is ordered by increasing cluster-venue and sale price. The location of the selected neighborhoods on the Madrid map is shown in

Neighborhoods with lower price range in given venue cluster

	Neighborhood	District	Label_venue	Label_price	Sale price (EUR/m2)	Rent price (EUR/m2)
95	Prosperidad	Chamartín	0	11	4228.0	14.40
123	Vallehermoso	Chamberí	0	4	4953.0	15.20
54	Goya	Salamanca	2	5	5599.0	16.50
35	Comillas	Carabanchel	3	12	2565.0	12.20
112	Simancas	San Blas-Canillejas	3	12	2633.0	11.80
125	Ventas	Ciudad Lineal	3	12	2663.0	12.00
44	El Salvador	San Blas-Canillejas	3	7	3311.0	12.20
17	Atocha	Arganzuela	3	4	3915.0	14.15
16	Atalaya	Ciudad Lineal	3	11	4132.5	14.40
108	San Juan Bautista	Ciudad Lineal	3	11	4151.0	13.30
96	Pueblo Nuevo	Ciudad Lineal	5	12	2506.0	12.10
99	Quintana	Ciudad Lineal	5	12	2909.0	12.50
36	Concepción	Ciudad Lineal	5	9	2942.0	13.10
10	Apóstol Santiago	Hortaleza	5	8	2983.0	10.80
18	Bellas Vistas	Tetuán	5	4	3440.0	14.30
31	Chopera	Arganzuela	5	4	3641.0	14.00
81	Pacífico	Retiro	5	11	3976.0	14.20
2	Adelfas	Retiro	5	4	4077.0	13.80
63	Las Tablas	Fuencarral-El Pardo	5	11	4130.0	12.00

Figure 25: List of neighborhoods identified as preferred opportunities for real estate investment through combination of venue and housing price and demographic data segmentation.

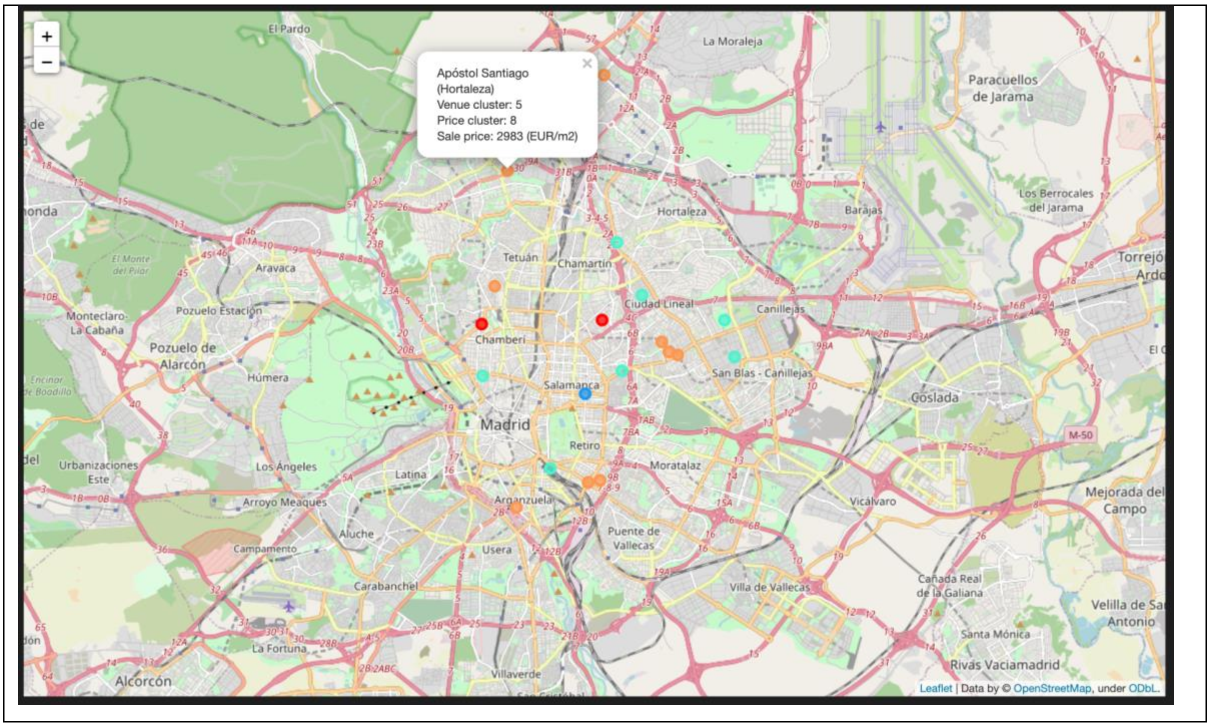


Figure 26: Location of neighborhoods identified as preferred opportunities for real estate investment through combination of venue and housing price and demographic data segmentation. The locations are colored by venue-cluster (see section 4.1): Cluster 0 (●), Cluster 2 (●), Cluster 3 (●), Cluster 4 (●), Cluster 5 (●).

5 Discussion

5.1 Segmentation of Madrid's neighborhoods by venue

The Segmentation by venue is characterized by two drawbacks:

a) Segmentation by venue was possible only on less than half the neighborhoods, as venue density in the remaining ones is too low

b) 280 different unique venue categories were retrieved from FourSquare. This means that the characterization of a given neighborhood might become very specific due the high granularity of the venue category, so that comparison with other neighborhoods might become more difficult. For characterization of neighborhoods, it would be more beneficial to rely on more general categories, i.e., compare three "fine restaurants", instead of one "sushi restaurant", one "Chinese restaurant", one "Argentinian restaurant" in one neighborhood and one "Italian restaurant", one "French restaurant" one "Spanish restaurant" in the other.

In spite of the drawbacks, the proposed automatic segmentation is in acceptable agreement with actual characteristics of Madrid neighborhoods, e.g., with respect to identification of the historic and more touristic neighborhood, or the classier areas famous for its shopping.

5.2 Segmentation of Madrid's neighborhoods by housing prices and demographic data

For the segmentation of neighborhoods based on housing prices and demographic data, several features such as such as historical maximum, yearly price variations, variation from historical maximum, year over year population density changes etc. were included. This approach favored the selection of a high number of clusters (14) with similar or overlapping sale price range, where the additional differentiating elements between the clusters become the additional features, such as historical maximum, yearly price variations, variation from historical maximum. In fact, based on consideration of sale price only 3-4 clusters would intuitively make more sense (see Figure 24). One advantage of the granularity on the price-cluster side allowed to identify neighborhoods which stands out within a venue cluster and can be subject of further scrutiny. Whether an analysis limited to less housing price features would improve the results, would need to be investigated.

5.3 Combining the segmentation by venue with the one by housing prices and demographics.

The selection of a neighborhood as potential investment opportunity was based on mean cluster price. Within a given venue-cluster, those neighborhoods were selected, whose price-cluster corresponds to a cluster with lower mean price if compared to the remaining neighborhoods in the same venue-cluster. This approach led to the identification of 19 neighborhoods distributed across 4 different venue-clusters. However, as discussed above certain price cluster feature a broad distribution of prices and prices might overlap in certain cases. Moreover, clusters are characterized by many more criteria than only actual price. Therefore, it remains to be assessed on a case-by-case basis, if the underlying reasons making a neighborhood stand out within a venue-cluster, justify an interest from real estate investment point of view. As an example, the Vallehermoso neighborhood (price-cluster 4, venue-cluster 0) was selected because its mean price-cluster 4 below the ones of price-cluster 5 and price-cluster 1, which represent the majority of price-clusters in the venue-clusters. However, Vallehermoso neighborhood's sale price corresponds to the maximum of its price-cluster and is comparable to third quartile of price cluster 5. In this case, it remains to be seen if other indicators justify the neighborhood a preferred location for investment.

6 Conclusion

The proposed combination of Madrid's neighborhood segmentation by venue categories with the one by prices for housing for rent and demographic development has led to the identification of 19 neighborhoods (out of more than 130) as potential investment opportunities for further screening.

The underlying segmentation by venue categories was able to correctly group neighborhoods agreeing with Madrid's situation. The segmentation quality might be improved by reducing the granularity of the original venue categories retrieved from FourSquare, thus easing the comparison between different neighborhoods.

The segmentation by housing prices and demographic development was also able to cluster neighborhoods that appear to be accordance with Madrid's actual real estate market situation. However, a relatively large number features has driven the selection of a large number of price-cluster, resulting in several overlapping clusters in terms of sale and price range. It remains to be investigated how the selection of lower number of features and/or clusters would impact the results.

7 References

[1] Anexo:Distritos de Madrid. In *Wikipedia*. Retrieved March 18, 2021, from https://es.wikipedia.org/wiki/Anexo:Distritos_de_Madrid

[2] Anexo:Barrios administrativos de Madrid. In *Wikipedia*. Retrieved March 18, 2021, from https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid

[3] http://www-2.munimadrid.es/CSE6/control/menuCSE

[4] Price evolution of housing for sale in Spain. In *idealista.com*. Retrieved March 18, 2021 from https://www.idealista.com/en/press-room/property-price-reports/

[5] Price evolution of housing for rent in Spain. In *idealista.com*. Retrieved March 18, 2021 from https://www.idealista.com/en/press-room/property-price-reports/rent/

[6] Venues Explore. In FourSquare/developers. Retrieved March 18, 2021, from https://developer.foursquare.com/docs/api-reference/venues/explore/

[7] La Peseta (Madrid). In *Wikipedia*. Retrieved March 18, 2021, from https://es.wikipedia.org/wiki/La_Peseta_(Madrid)

[8] Arroyo del Fresno. Retrieved March 18, 2021, from https://www.idealista.com/news/inmobiliario/vivienda/2011/11/23/365358-asi-es-arroyo-del-fresno-el-pau-mas-desconocido-y-privilegiado-de-madrid-video

[9] Sklearn.cluster.KMeans. Retrieved April 9, 2021, from https://sklearn.org/modules/generated/sklearn.cluster.KMeans.html - sklearn.cluster.KMeans

[10] Selecting the number of clusters with silhouette analysis on KMeans clustering. Retrieved April 9, 2021, from

- https://sklearn.org/auto_examples/cluster/plot_kmeans_silhouette_analysis.html - sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

[11] Silhouette Coefficient. Retrieved April 9, 2021, from https://scikit-learn.org/stable/modules/clustering.html - silhouette-coefficient

[12] sklearn.preprocessing.StandardScaler. Retrieved April 9, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html?highlight=standard scaler - sklearn.preprocessing.StandardScaler