Madrid District Classification Analysis

Introduction

- Madrid, Spain is one of the most important and biggest cities in Europe. It is full of diverse features and options for many tastes and interests.
- Madrid has 21 districts with differences in aspects such as real estate values, security and places of interest or venues.
- Depending on people's preferences and possibilities, some districts might be more convenient for them than others.
- The two main objectives of this study are:
  - Characterize how the different characteristics of each district have impact on the average unit area real estate price.
  - Define a number of district classes and group similar districts into a same class.
- Once we get enough insights we will describe what profiles are better suited for each classified district.
- All our processes, tables, calculations and plots will be executed in a Python notebook with the aid of different useful libraries such as Pandas, Numpy, BeautifulSoup, Folium, Scikit-Learn, Nominatim and Seaborn.

Data input types

- ***Geographical***: Extract district information a from Wikipedia source and coordinate information from *Nominatim* functions.
- ***Real Estate***: Extract property price values from different Real Estate information sources such as Idealista and El País-Cinco Días.
- ***Security***: Extract security incident statistics from Madrid police public data published in datos.madrid.es.
- ***Venues (places of interest)***: Extract Madrid venue information from Foursquare.com database.

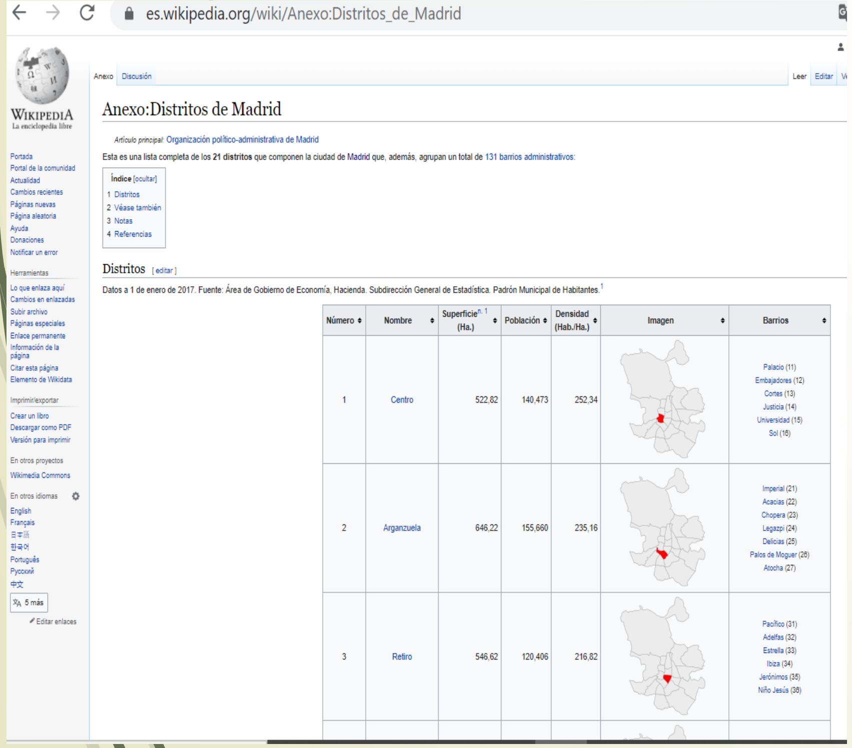
Analytic methods

- ***Plots***: Graphically represent the district features as bar plots, crossplots, choropleth maps and pairplots to better understand them and discover our first district insights. We rely on *Matplotlib* functions, as well as *Folium* and *Seaborn* libraries to generate our graphics.
- ***Multilinear Regression***: We will try to understand how the different characteristic impact each district unit area price. We use *Scikit-Learn's Linear Regression* module to execute this method.
- ***K-Means clustering***: Based on processed and normalized district attributes, we group similar disctricts into different clases. We run some Elbow method tests to determine an optimum quantity of clases. We rely on *Cluster Scikit-Learn's KMeans* module to perform the clustering step and *cdistance* module from *Geopy's Spatial Distance* to select the optimum k number of classes.

Geographical District Data Mining

- Extract table information from [Wikipedia page]() to a *Pandas* data frame (table) using *BeautifulSoup* library.
- Get District coordinates from *Nominatim-Geocod* commands

Original Wikipedia page table



Nominatim – Geocod: Get district coordinates

Built Pandas dataframe

DistNumber	District	Neighborhood	LAT	LONG
1	Centro	[Palacio, Embajadores, Cortes, Justicia, Unive...	40.417653	-3.707914
2	Arganzuela	[Imperial, Acacias, Chopera, Legazpi, Delicias...	40.396954	-3.697289
3	Retiro	[Pacífico, Adelfas, Estrella, Ibiza, Jerónimo...	40.411150	-3.676057
4	Salamanca	[Recoletos, Goya, Fuente del Berro, Guindalera...	40.427045	-3.680602
5	Chamartín	[El Viso, Prosperidad, Ciudad Jardín, Hispanoa...	40.458987	-3.676129
6	Tetuán	[Bellas Vistas, Cuatro Caminos, Castillejos, A...	40.460578	-3.698281
7	Chamberí	[Gaztambide, Arapiles, Trafalgar, Almagro, Río...	40.436247	-3.703830
8	Fuencarral-El Pardo	[El Pardo, Fuentelarreina, Peñagrande, Pilar, ...	40.556346	-3.778591
9	Moncloa-Aravaca	[Casa de Campo, Argüelles, Ciudad Universitari...	40.439495	-3.744204
10	Latina	[Los Cármenes, Puerta del Ángel, Lucero, Aluch...	40.403532	-3.736152
11	Carabanchel	[Comillas, Opañel, San Isidro, Vista Alegre, P...	40.374211	-3.744676
12	Usera	[Orcasitas, Orcasur, San Fermín, Almendrales, ...	40.383894	-3.706446
13	Puente de Vallecas	[Entrevías, San Diego, Palomeras Bajas, Palome...	40.383553	-3.654535
14	Moratalaz	[Pavones, Horcajo, Marroquina, Media Legua, F...	40.405933	-3.644874
15	Ciudad Lineal	[Ventas, Pueblo Nuevo, Quintana, Concepción, S...	40.448431	-3.650495
16	Hortaleza	[Palomas, Piovera, Canillas, Pinar del Rey, Ap...	40.472549	-3.642552
17	Villaverde	[Villaverde Alto, San Cristóbal, Butarque, Los...	40.345610	-3.695956
18	Villa de Vallecas	[Casco Histórico de Vallecas, Santa Eugenia, E...	40.373958	-3.612163
19	Vicálvaro	[Casco Histórico de Vicálvaro, Valdebernardo, ...	40.396584	-3.576622
20	San Blas-Canillejas	[Simancas, Hellín, Amposta, Arcos, Rosas, Reja...	40.428919	-3.604002
21	Barajas	[Alameda de Osuna, Aeropuerto, Casco Histórico...	40.473318	-3.579845

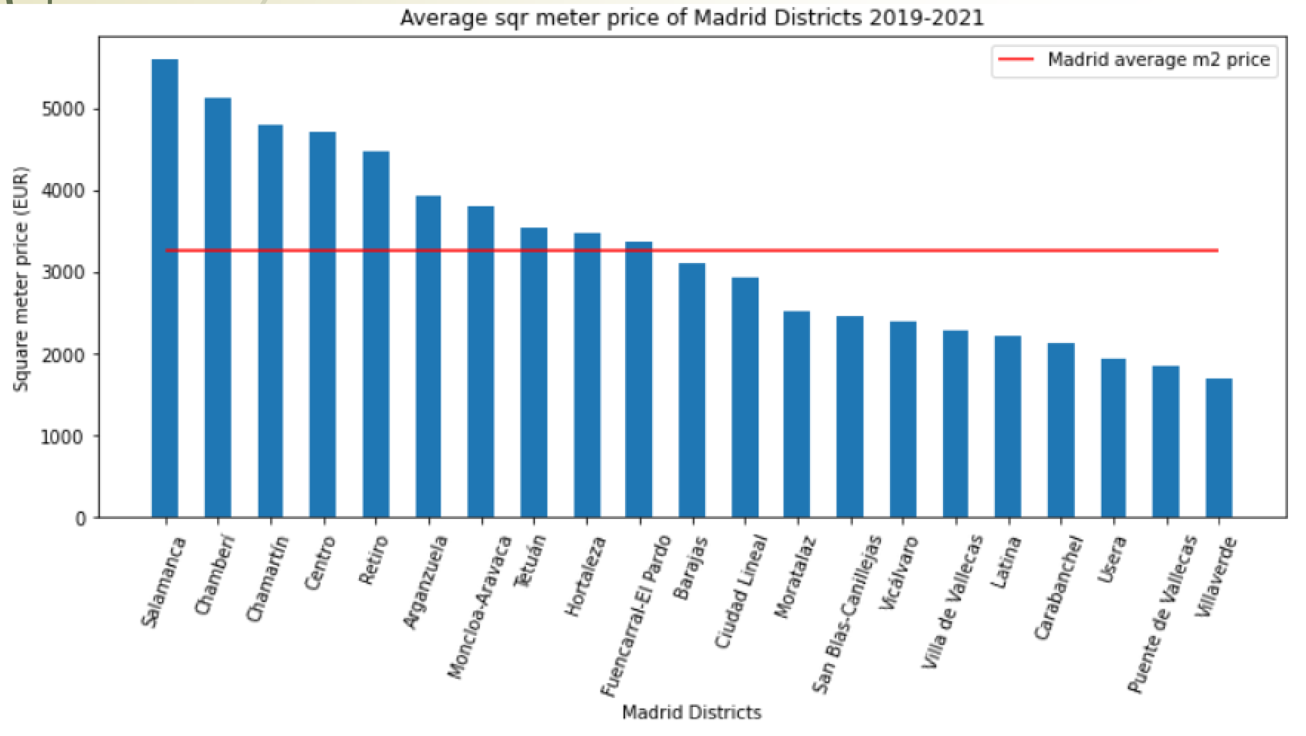
Madrid unit area prices per district data

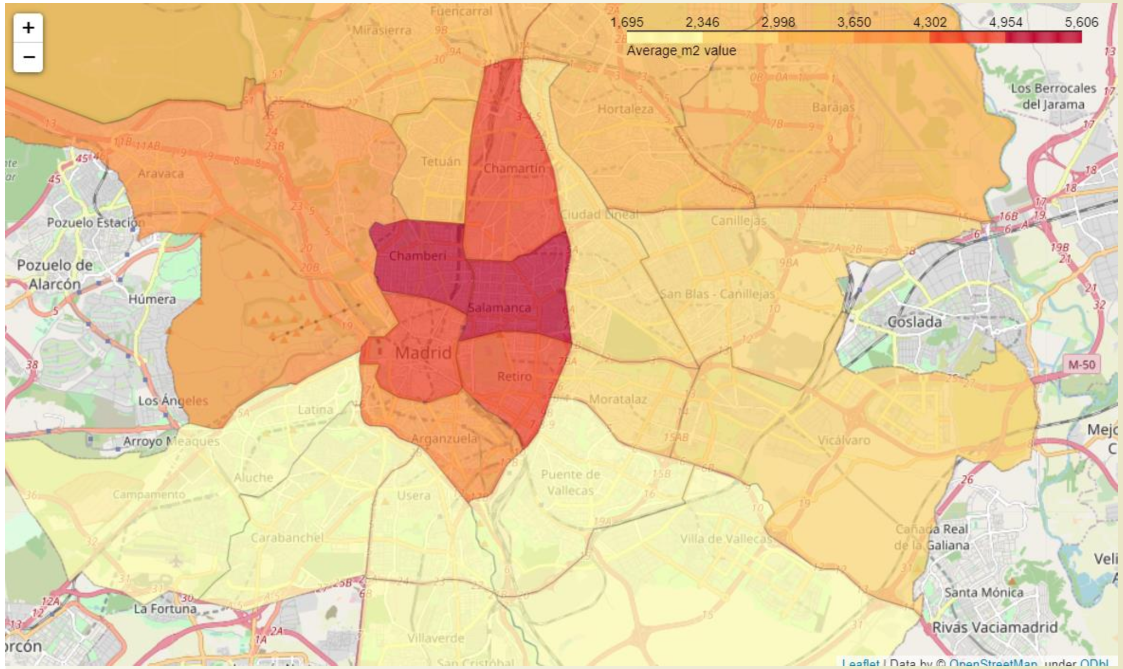
- *2019 data*: Average district square-meter (SQM) prices from an October 2019 article of the media site El País-Cinco Días.
- *2021 data*: Mid-year average SQM district prices from Idealista real-estate page.
- Both datasets were extracted as spreadsheets and loaded as *Pandas* data frames.
- Average prices and percent variation calculated from both datasets.
- Normalized to mean: Each district average price was divided by the mean of all average prices in order to have as reference a relative district value from Madrid mean SQM price.
- Linear normalization: District price minus the minimum price divided by the price range.. Output values range from 0 (minimum price) to 1 (maximum price).

District	SQMprice2021	SQMprice2019	Var	meanSQM	SQMnormMean	SQMlinNorm
Arganzuela	4000	3878	0.031460	3939.0	1.208997	0.573894
Barajas	3225	3001	0.074642	3113.0	0.955473	0.362695
Carabanchel	2123	2156	-0.015306	2139.5	0.656677	0.113782
Centro	4793	4621	0.037221	4707.0	1.444720	0.770263
Chamartín	5137	4468	0.149731	4802.5	1.474032	0.794682
Chamberí	5347	4905	0.090112	5126.0	1.573323	0.877397
Ciudad Lineal	3004	2883	0.041970	2943.5	0.903449	0.319356
Fuencarral-El Pardo	3504	3248	0.078818	3376.0	1.036196	0.429941
Hortaleza	3812	3154	0.208624	3483.0	1.069037	0.457300
Latina	2254	2183	0.032524	2218.5	0.680924	0.133981
Moncloa-Aravaca	3999	3613	0.106836	3806.0	1.168176	0.539887
Moratalaz	2585	2451	0.054672	2518.0	0.772850	0.210560
Puente de Vallecas	1902	1807	0.052573	1854.5	0.569202	0.040910
Retiro	4705	4244	0.108624	4474.5	1.373358	0.710816
Salamanca	6063	5148	0.177739	5605.5	1.720496	1.000000
San Blas-Canillejas	2595	2329	0.114212	2462.0	0.755662	0.196241
Tetuán	3664	3433	0.067288	3548.5	1.089141	0.474048
Usera	1997	1863	0.071927	1930.0	0.592375	0.060215
Vicálvaro	2448	2321	0.054718	2384.5	0.731875	0.176425
Villa de Vallecas	2394	2193	0.091655	2293.5	0.703944	0.153158
Villaverde	1717	1672	0.026914	1694.5	0.520093	0.000000

Average square meter price per District graphics

- 10 districts have square-meter (SQM) average prices above the city mean, around 3258 €/m2.
- 11 districts have SQM prices bellow the city mean.
- Central districts have the highest square-meter average price near or above 4000 €/m2.
- Northern districts are more expensive than southern districts which are around 2000 €/m2 or below.





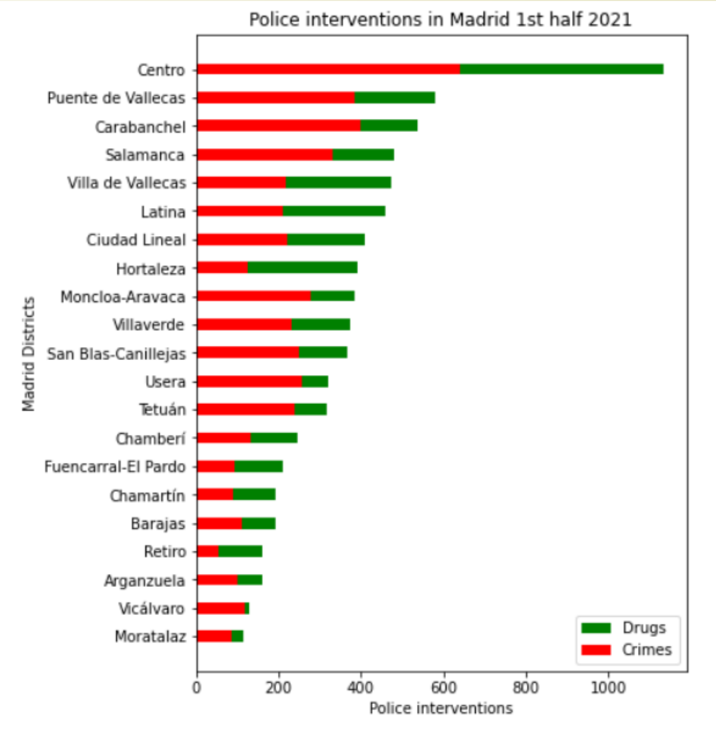
Madrid security indicators

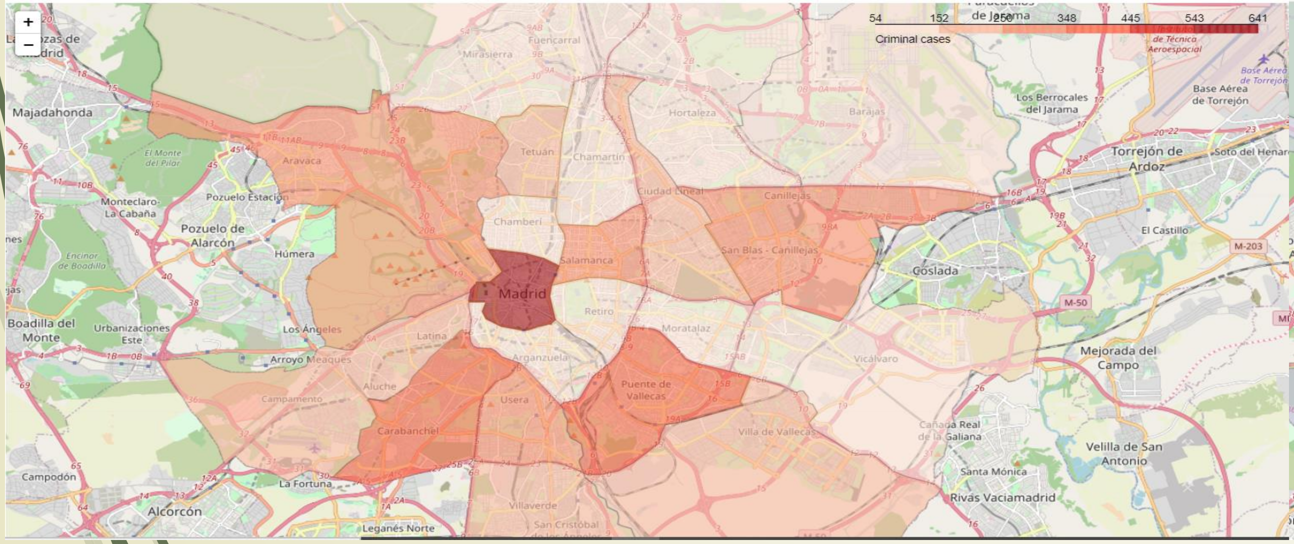
- Monthly reports of police interventions per districts downloaded form the website datos.madrid.es
- Statistics integrated in a single first-half-2021 table
- Categories related with people, with assets and weapon possession were grouped in "Crimes" category. Drug possession and drug consumption were grouped in "Drugs" category.
- Proportional versions of "Crimes" and "Drugs" categories compared to the total Madrid cases were also generated for further analytic calculations.

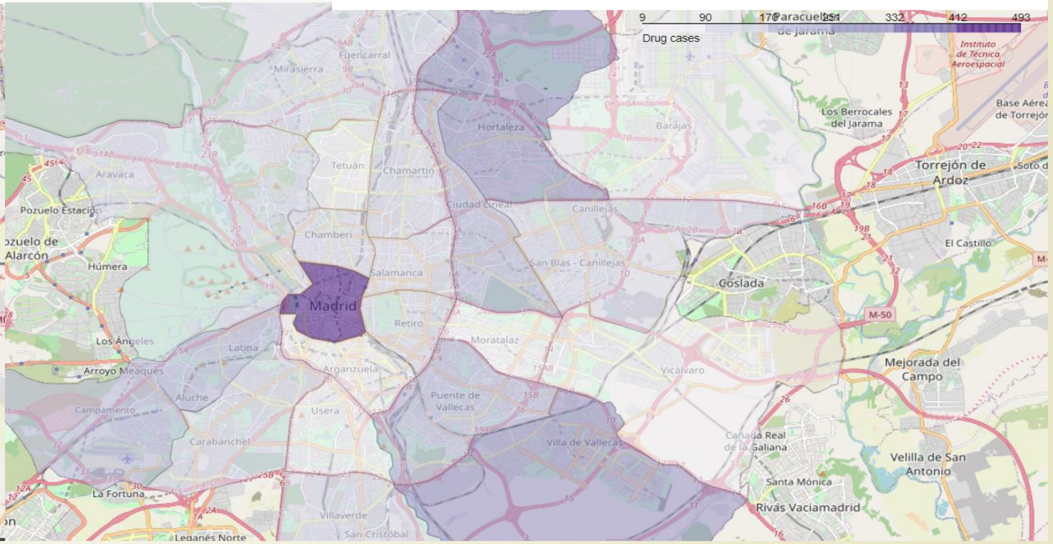
District	People	Assets	Weapons	DrugPos	DrugCons	Crimes	Drugs	Total	CrimesProp	DrugsProp
Moratalaz	44	41	2	24	3	87	27	114	0.019096	0.008789
Vicálvaro	67	52	0	7	2	119	9	128	0.026119	0.002930
Arganzuela	34	50	14	37	25	98	62	160	0.021510	0.020182
Retiro	21	28	5	88	18	54	106	160	0.011853	0.034505
Barajas	43	66	1	60	23	110	83	193	0.024144	0.027018
Chamartín	31	49	9	88	17	89	105	194	0.019535	0.034180
Fuencarral-El Pardo	40	44	7	112	6	91	118	209	0.019974	0.038411
Chamberí	58	63	10	98	15	131	113	244	0.028753	0.036784
Tetuán	109	114	15	62	18	238	80	318	0.052239	0.026042
Usera	120	134	3	48	14	257	62	319	0.056409	0.020182
San Blas-Canillejas	111	126	13	108	8	250	116	366	0.054873	0.037760
Villaverde	116	110	5	119	24	231	143	374	0.050702	0.046549
Moncloa-Aravaca	147	121	9	92	17	277	109	386	0.060799	0.035482
Hortaleza	36	57	32	214	52	125	266	391	0.027436	0.086589
Ciudad Lineal	108	101	12	149	38	221	187	408	0.048507	0.060872
Latina	106	83	20	204	46	209	250	459	0.045874	0.081380
Villa de Vallecas	108	91	18	109	147	217	256	473	0.047629	0.083333
Salamanca	44	274	12	139	11	330	150	480	0.072432	0.048828
Carabanchel	208	171	18	113	27	397	140	537	0.087138	0.045573
Puente de Vallecas	194	162	28	192	5	384	197	581	0.084284	0.064128
Centro	203	366	72	416	77	641	493	1134	0.140694	0.160482

Security indicator per District graphics

- Madrid Centro highlights for its considerably high quantity of police interventions compared to the rest of districts, likely related to the high tourist traffic that could be target of burglars and/or drug dealers.
- Northern districts are relatively safer than the southern ones.
- It seems that the central districts, excluding Centro, have less drug issues than the border districts.
- Eastern and southern districts have more drug cases than the western and northwestern districts.







Venues per district mining

- URL search with *Request* library from the Foursquare API of items and json information related to places of interest or venues from each district central coordinates with a search radius of 1 kilometer. The search allowed us to populate a table with venue names, coordinates and categories.
- Then we counted the number of venues and the number of different venue categories associated to each district.

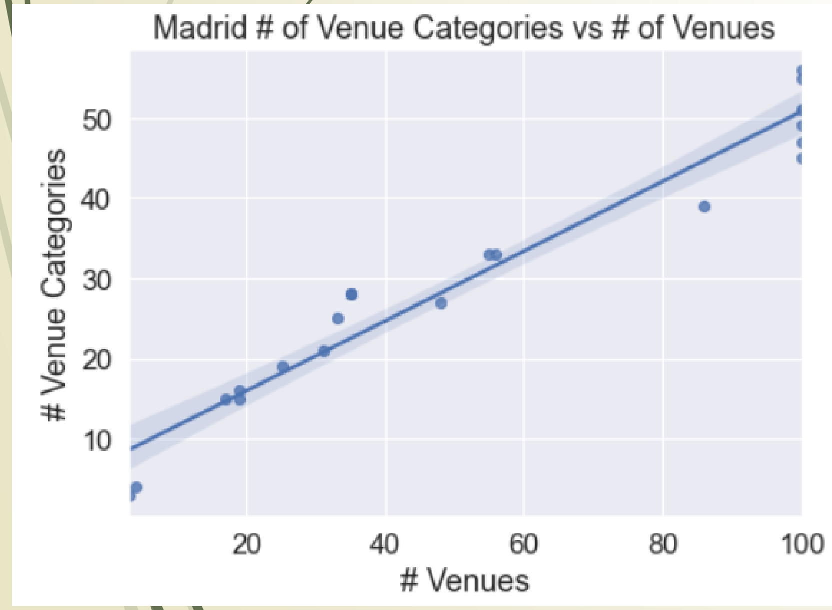
	District	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Centro	40.417653	-3.707914	Plaza de Isabel II	40.418114	-3.709397	Plaza
1	Centro	40.417653	-3.707914	Plaza Mayor	40.415527	-3.707506	Plaza
2	Centro	40.417653	-3.707914	La Esquina del Real	40.417356	-3.710364	French Restaurant
3	Centro	40.417653	-3.707914	Zen Zoo	40.416263	-3.707174	Smoothie Shop
4	Centro	40.417653	-3.707914	Torrons Vicens: Artesa D' Agramunt	40.416095	-3.708119	Pastry Shop
5	Centro	40.417653	-3.707914	Mercado de San Miguel	40.415443	-3.708943	Market
6	Centro	40.417653	-3.707914	TOC Hostel	40.417264	-3.705928	Hostel
7	Centro	40.417653	-3.707914	Cerveceria Erte	40.419241	-3.707470	Bar
8	Centro	40.417653	-3.707914	Gran Meliá Palacio de los Duques *****	40.419835	-3.709494	Hotel
9	Centro	40.417653	-3.707914	Trattoria Malatesta	40.416788	-3.707182	Italian Restaurant

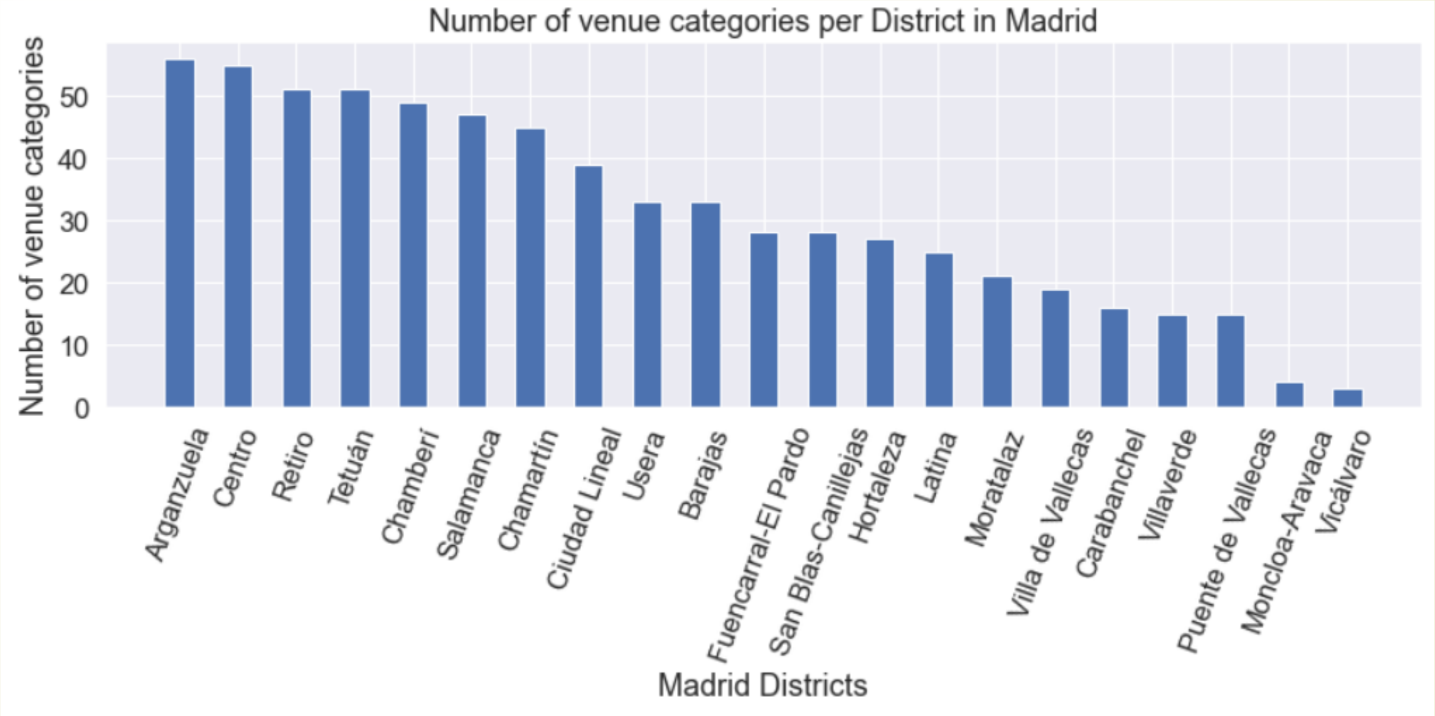
Count # of venues and # of different venue categories per district

	District	Venue	Venue Category
0	Arganzuela	100	56
1	Barajas	55	33
2	Carabanchel	19	16
3	Centro	100	55
4	Chamartín	100	45
5	Chamberí	100	49
6	Ciudad Lineal	86	39
7	Fuencarral-El Pardo	35	28
8	Hortaleza	48	27
9	Latina	33	25
10	Moncloa-Aravaca	4	4
11	Moratalaz	31	21
12	Puente de Vallecas	19	15
13	Retiro	100	51
14	Salamanca	100	47
15	San Blas-Canillejas	35	28
16	Tetuán	100	51
17	Usera	56	33
18	Vicálvaro	3	3
19	Villa de Vallecas	25	19
20	Villaverde	17	15

Venues per District graphics

- Number of retrieved venues were limited to 100. Clipped results are visible on crossplot.
- Number of different venue categories are highly correlated to number of venues: the more venues there are the more diverse they trend to be.
- Central districts of Madrid are the ones with more variety of venues. As we get farther from downtown the diversity becomes low to very low.

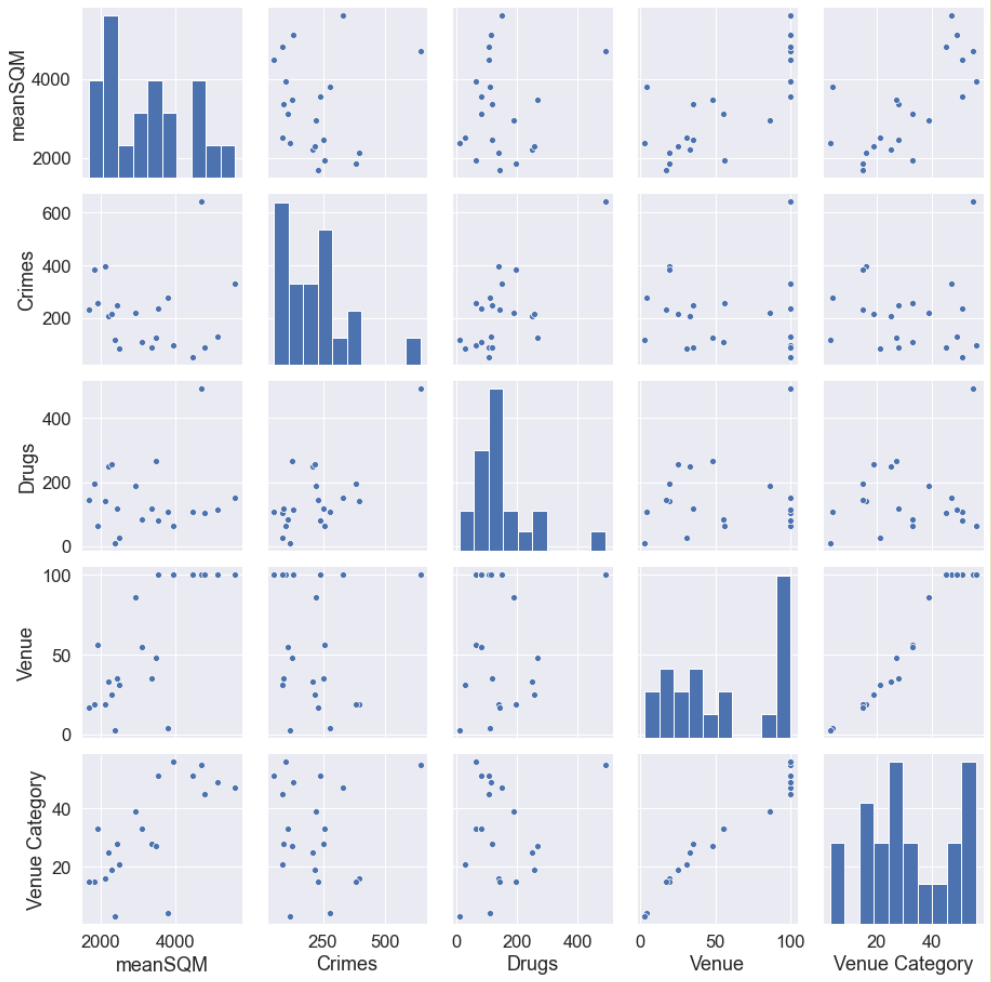




Selected features for regression analysis and relationship observations

- Features selected for the multilinear regression analysis: average square-meter prices, crime and drug cases, number of venues and venue categories.
- The pair plots that compares all the variables indicate that the area prices appear to be more correlated to the quantity of venues and the variety of venues.
- Also, the area prices seem to be inversely correlated to the number of reported crimes. Crime and drug cases seem to have some degree of correlation. Number of venues and venue categories show strong correlation between them.

	District	meanSQM	Crimes	Drugs	Venue	Venue Category
0	Centro	4707.0	641	493	100	55
1	Arganzuela	3939.0	98	62	100	56
2	Retiro	4474.5	54	106	100	51
3	Salamanca	5605.5	330	150	100	47
4	Chamartín	4802.5	89	105	100	45
5	Tetuán	3548.5	238	80	100	51
6	Chamberí	5126.0	131	113	100	49
7	Fuencarral-El Pardo	3376.0	91	118	35	28
8	Moncloa-Aravaca	3806.0	277	109	4	4
9	Latina	2218.5	209	250	33	25
10	Carabanchel	2139.5	397	140	19	16



Multilinear regression modeling of district features

```python
regr=linear_model.LinearRegression()
x=np.asanyarray(madrid_features[['Crimes','Drugs','Venue','Venue Category']])
y=np.asanyarray(madrid_features[['meanSQM']])
regr.fit(x,y)
# The coefficients
print('Coefficients: ',regr.coef_)
print('Intercept: ',regr.intercept_)
print('Variance score: %.2f' %regr.score(x,y))
```

```
Coefficients:  [[ -0.64387148   1.5255613   55.3718756  -73.14628878]]
Intercept:  [2399.01894308]
Variance score: 0.64
```

- Model 1 considered all the variables. It suggests that the prices are positively influenced by the drug cases and number of venues and negatively influenced by the crimes and variety of venues. But we saw that the number of venue categories showed a positive trend with the area prices. This could be explained by the fact that the independent variables "number of venues" and "venue categories" have collinearity, that is, they are correlated. For this reason, we decided to stay with only one of those variables, venue category in our case.

Multilinear regression modeling of district features

```
regr2=linear_model.LinearRegression()
x2=np.asanyarray(madrid_features[['Crimes','Drugs','Venue Category']])
y2=np.asanyarray(madrid_features[['meanSQM']])
regr2.fit(x2,y2)
# The coefficients
print('Coefficients: ',regr2.coef_)
print('Intercept: ',regr2.intercept_)
print('Variance score: %.2f' %regr2.score(x2,y2))
```

```
Coefficients:  [[-0.59706497  0.6596417  49.11058566]]
Intercept:  [1747.63438778]
Variance score: 0.49
```

- Model 2 considered the crime and drug cases and variety of venues. The variance score dropped to nearly 50% where the price is explained to be considerably benefited from a greater variety of venue categories, negatively affected by the crime occurrences and positively from drug cases. But we have observed some degree of correlation between crimes and drugs. Thereby we proceeded to our next simplified model by selecting one of the security categories. We chose the number of crimes, since it visually better correlates with the unit area prices and we consider that crimes negatively affect more the willingness to invest or rent.

Multilinear regression modeling of district features

```
regr3=linear_model.LinearRegression()
x3=np.asanyarray(madrid_features[['Crimes', 'Venue Category']])
y3=np.asanyarray(madrid_features[['meanSQM']])
regr3.fit(x3,y3)
# The coefficients
print('Coefficients: ',regr3.coef_)
print('Intercept: ',regr3.intercept_)
print('Variance score: %.2f' %regr3.score(x3,y3))
```

```
Coefficients:  [[-0.25789678 49.84272906]]
Intercept:  [1747.5369784]
Variance score: 0.49
```

- Model 3 was simpler but as accurate as the model 2. It suggests that for every new crime the average square meter price drops by 0.25 euros and that for every new venue category present in a district, the square meter price increases 49.8 euros. This model does not allow us to accurately predict property prices for a district because of the low model accuracy, but it gives us qualitative insights indicating that Madrid districts get better real estate values by the presence of more variety of places of interests for the residents and visitors and they are less affected by the occurrence of security incidents that require police intervention.

K-means clustering analysis

- The purpose of this analytic method is to group the districts into different classes according their different features such as normalized property prices, proportional criminal and drug indicators, proportional variety of venues and the proportional quantity of each venue type
- K-means method was chosen because it allowed us to select in how many classes we wanted to group the different districts. A quantity fair enough to guarantee that the districts within a class had to be as similar as possible.
- The input table with selected features only (without district column) had the following configuration:

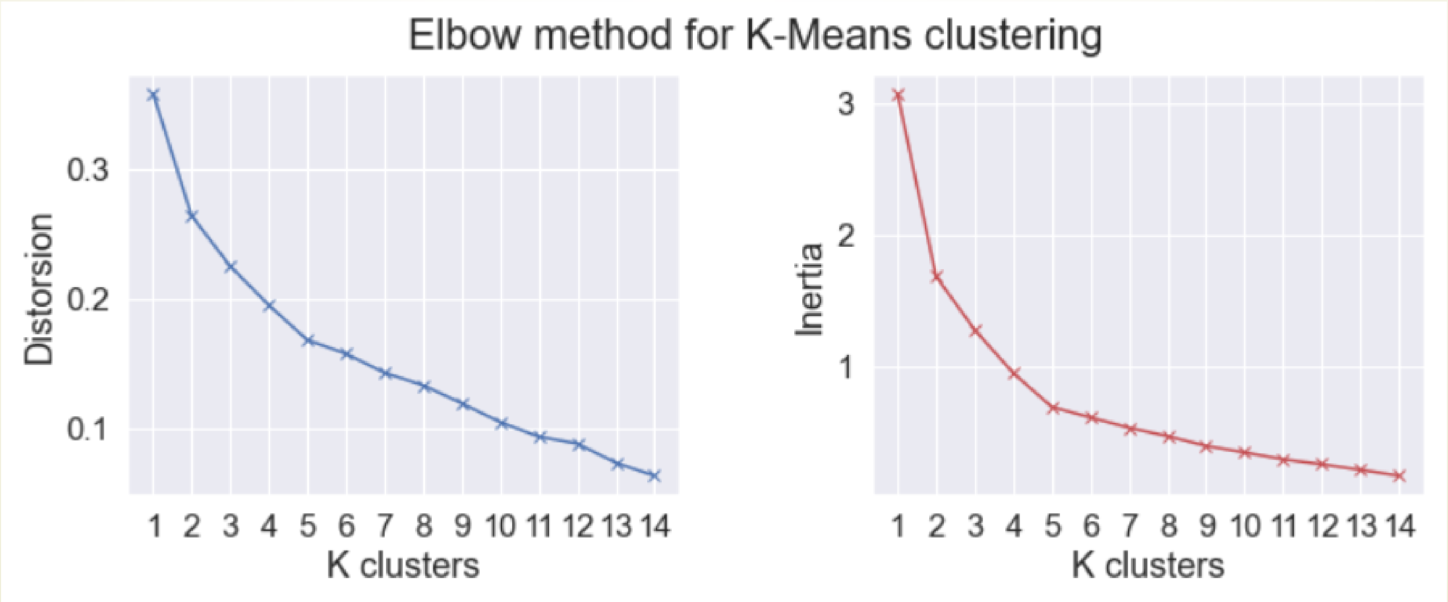
	Accessories Store	Airport	Airport Lounge	Airport Service	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant
0	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.01	0.020000
1	0.00	0.018182	0.018182	0.054545	0.000000	0.00	0.00	0.036364
2	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000
3	0.01	0.000000	0.000000	0.000000	0.010000	0.00	0.00	0.000000
4	0.00	0.000000	0.000000	0.000000	0.010000	0.01	0.00	0.000000
5	0.00	0.000000	0.000000	0.000000	0.010000	0.00	0.00	0.000000
6	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.023256
7	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000
8	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.020833
9	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000

…

	Trade School	Train Station	Vegetarian / Vegan Restaurant	Venezuelan Restaurant	Vietnamese Restaurant	Wine Bar	Wine Shop	SQMlinNorm	CrimesProp	DrugsProp	VenueCatProp
0	0.01	0.000000	0.00	0.000000	0.00	0.010000	0.000000	0.573894	0.140694	0.160482	0.277778
1	0.00	0.000000	0.00	0.000000	0.00	0.018182	0.000000	0.362695	0.021510	0.020182	0.282828
2	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.113782	0.011853	0.034505	0.257576
3	0.00	0.000000	0.01	0.000000	0.00	0.020000	0.000000	0.770263	0.072432	0.048828	0.237374
4	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.794682	0.019535	0.034180	0.227273
5	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.010000	0.877397	0.052239	0.026042	0.257576
6	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000	0.319356	0.028753	0.036784	0.247475
7	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.028571	0.429941	0.019974	0.038411	0.141414
8	0.00	0.000000	0.00	0.020833	0.00	0.000000	0.000000	0.457300	0.060799	0.035482	0.020202
9	0.00	0.030303	0.00	0.000000	0.00	0.000000	0.000000	0.133981	0.045874	0.081380	0.126263

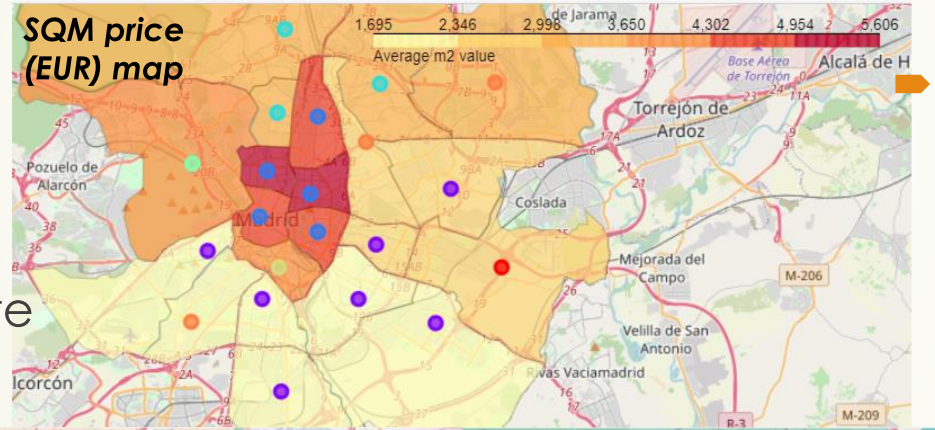
K-means clustering, k selection

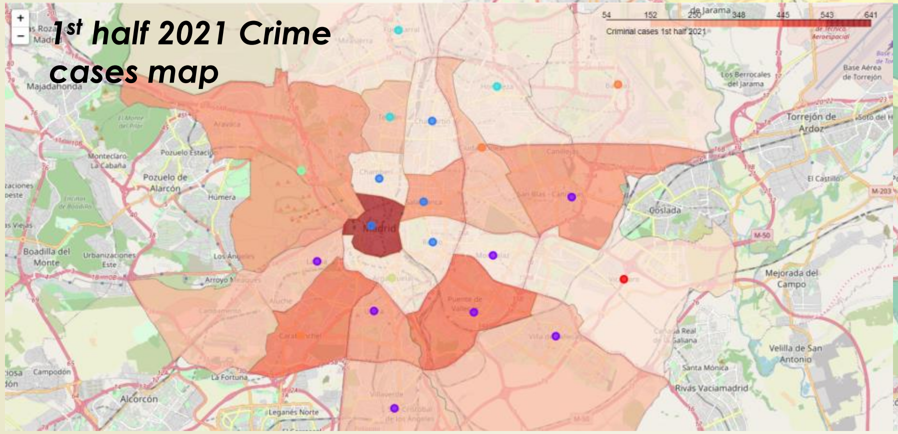
- Elbow method: To determine an optimal number of "k" classes, not too large or too small, we measured feature-Euclidean-distance metrics such as distortion and inertia for different k iterations and plotted them to identify the "elbow" point from where Euclidean distance reduction is smoother and negligible.
- We observe distortion and inertia elbows at k=5, more remarkable with the inertia measure. From these two plots we decide that 7 would be a fair number of clusters that will assure us relatively low distortion and inertia values, lower than elbow values. Also 7 is the third of the total number of Madrid districts, 21. We would then expect about 3 districts per class.

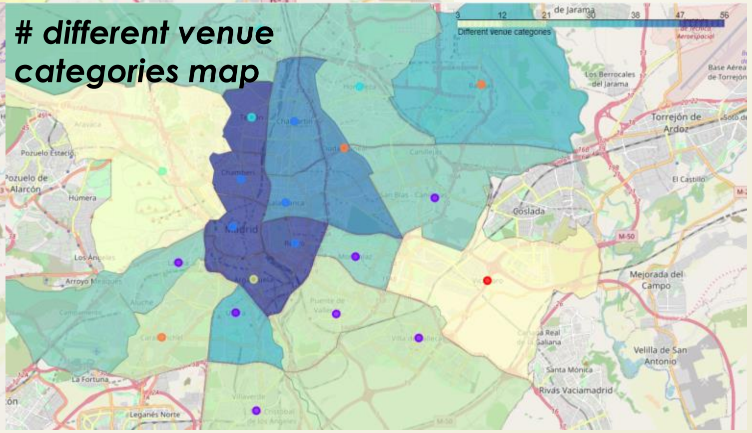


Clustering graphics

- Most of the central districts, the ones with the most expensive square-meter (SQM) areas, share the same class.
- One class gathers most of the southern districts and some eastern ones. They show average to below-average SQM prices, intermediate amount of different venue categories and intermediate to high crime indicators.







- The northern districts are gathered into two classes that have intermediate unit area values, relatively low crime indicators and intermediate to high numbers of different venue categories.
- There are three single-district classes and two with five or more districts gathered.

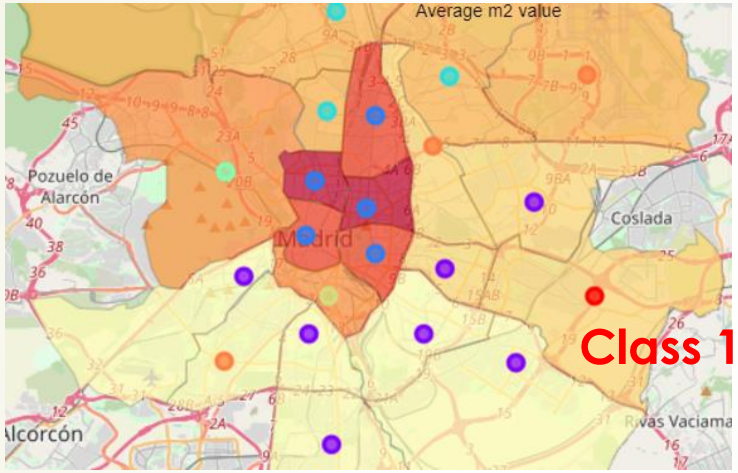
Most common venues per district

- In order to better qualitatively describe the clustered districts, we organized in a data frame the 10 most common venue types per district

	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Arganzuela	Spanish Restaurant	Restaurant	Tapas Restaurant	Grocery Store	Bakery	Park	Gym	Bar	Market	Indie Theater
1	Barajas	Hotel	Spanish Restaurant	Restaurant	Airport Service	Tapas Restaurant	Coffee Shop	Snack Place	Duty-free Shop	Breakfast Spot	Argentinian Restaurant
2	Carabanchel	Tapas Restaurant	Restaurant	Spanish Restaurant	Candy Store	Supermarket	Bakery	Café	BBQ Joint	Cafeteria	Athletics & Sports
3	Centro	Plaza	Tapas Restaurant	Spanish Restaurant	Hotel	Café	Hostel	Bookstore	Bar	Pastry Shop	Gourmet Shop
4	Chamartín	Spanish Restaurant	Restaurant	Mediterranean Restaurant	Plaza	Bar	Grocery Store	Pizza Place	Tapas Restaurant	Gastropub	Japanese Restaurant
5	Chamberí	Tapas Restaurant	Café	Bar	Spanish Restaurant	Restaurant	Theater	Ice Cream Shop	Plaza	Japanese Restaurant	Italian Restaurant
6	Ciudad Lineal	Spanish Restaurant	Grocery Store	Restaurant	Park	Chinese Restaurant	Hotel	Italian Restaurant	Café	Bar	Pharmacy
7	Fuencarral-El Pardo	Restaurant	Spanish Restaurant	Soccer Field	Bar	Wine Shop	Tapas Restaurant	Salad Place	Fast Food Restaurant	Metro Station	Bookstore
8	Hortaleza	Spanish Restaurant	Supermarket	Restaurant	Tapas Restaurant	Sandwich Place	Pizza Place	Plaza	Soup Place	Irish Pub	Coffee Shop
9	Latina	Grocery Store	Park	Pizza Place	Bar	Supermarket	Fast Food Restaurant	Bowling Alley	Sandwich Place	Bakery	Food
10	Moncloa-Aravaca	Hookah Bar	College Cafeteria	Park	Tennis Court	Wine Shop	Dog Run	Flea Market	Fish Market	Fast Food Restaurant	Farmers Market

District class 1 (label 0) description

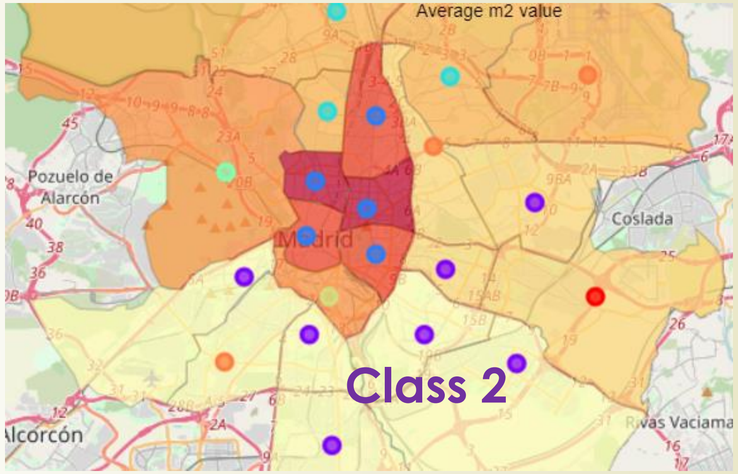
District	SQMnormMean	CrimesProp	DrugsProp	Venue	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Vicálvaro	0.731875	0.026119	0.00293	3	3	0	Construction & Landscaping	Mediterranean Restaurant	Toll Booth	Wine Shop	Duty-free Shop	Food & Drink Shop	Food	Flea Market	Fish Market	Fast Food Restaurant



- Vicalvaro is an easterly border district with the lowest variety of venues, property values below the average and one of the lowest crime rates. This district is in growth. It has a relatively new university and urbanistic development in progress. We might see in the near future a different picture of this district, with new venue opportunities and increase of property value.

District class 2 (label 1) description

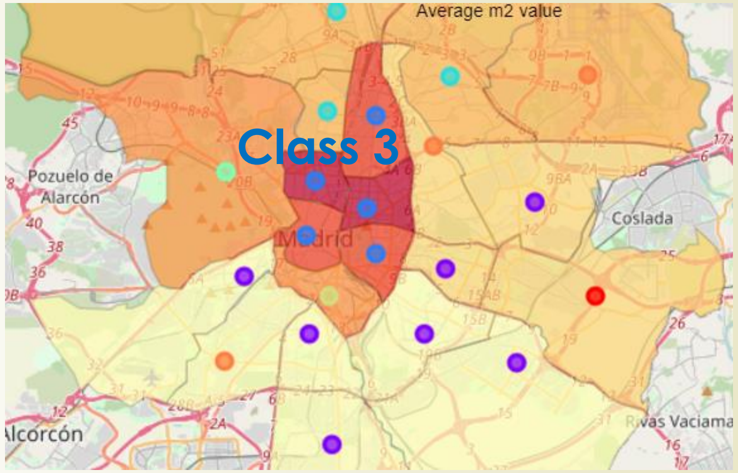
District	SQMnormMean	CrimesProp	DrugsProp	Venue	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Latina	0.680924	0.045874	0.081380	33	25	1	Grocery Store	Park	Pizza Place	Bar	Supermarket	Fast Food Restaurant	Bowling Alley	Sandwich Place	Bakery	Food
Usera	0.592375	0.056409	0.020182	56	33	1	Spanish Restaurant	Beer Garden	Grocery Store	Seafood Restaurant	Coffee Shop	Bar	Bakery	Fast Food Restaurant	Clothing Store	Gastropub
Puente de Vallecas	0.569202	0.084284	0.064128	19	15	1	Clothing Store	Park	Bar	Supermarket	Grocery Store	Concert Hall	Sandwich Place	Fast Food Restaurant	Shopping Mall	Spanish Restaurant
Moratalaz	0.772850	0.019096	0.008789	31	21	1	Park	Bar	Bakery	Café	Pizza Place	Playground	Plaza	Skating Rink	Soccer Field	Metro Station
Villaverde	0.520093	0.050702	0.046549	17	15	1	Metro Station	Spanish Restaurant	Pizza Place	Grocery Store	Brewery	Mediterranean Restaurant	Bus Station	Electronics Store	Furniture / Home Store	Gastropub
Villa de Vallecas	0.703944	0.047629	0.083333	25	19	1	Tapas Restaurant	Pizza Place	Gym	Restaurant	Soccer Field	Spanish Restaurant	Pharmacy	Church	Pet Store	Plaza
San Blas-Canillejas	0.755662	0.054873	0.037760	35	28	1	Grocery Store	Pizza Place	Tapas Restaurant	Gym	Seafood Restaurant	Playground	Soccer Stadium	Sports Club	Sporting Goods Shop	Spanish Restaurant



- This class gathers most of the southern and eastern districts of Madrid, which are characterized by property prices below the city average, intermediate to high criminal case reports and an intermediate variety of venues that include parks, local restaurants, pizza places, markets and sport businesses. This is more a “working-class” cluster. This is a popular group for people who are willing to live and enjoy of areas with good variety of options for eating, purchasing and recreation. However, they should to take some security precautions, but not as if they were in Centro (downtown).

District class 3 (label 2) description

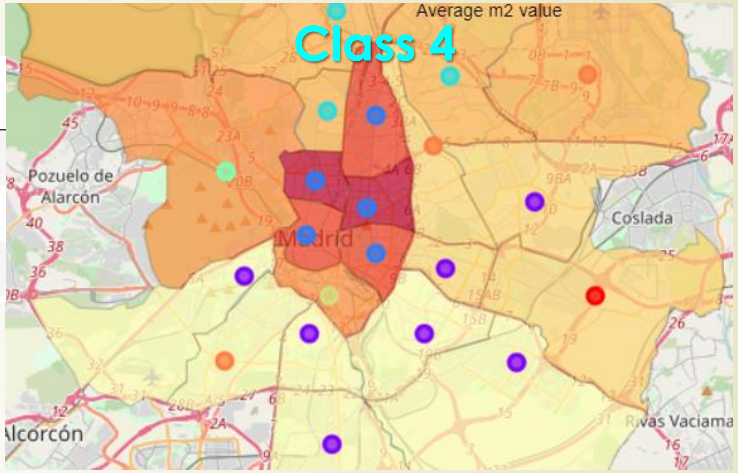
District	SQMnormMean	CrimesProp	DrugsProp	Venue	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Centro	1.444720	0.140694	0.160482	100	55	2	Plaza	Tapas Restaurant	Spanish Restaurant	Hotel	Café	Hostel	Bookstore	Bar	Pastry Shop	Gourmet Shop
Retiro	1.373358	0.011853	0.034505	100	51	2	Spanish Restaurant	Bar	Italian Restaurant	Brewery	Hotel	Tapas Restaurant	Supermarket	Gym	Garden	Bakery
Salamanca	1.720496	0.072432	0.048828	100	47	2	Spanish Restaurant	Restaurant	Italian Restaurant	Clothing Store	Boutique	Burger Joint	Hotel	Tapas Restaurant	Furniture / Home Store	Mexican Restaurant
Chamartín	1.474032	0.019535	0.034180	100	45	2	Spanish Restaurant	Restaurant	Mediterranean Restaurant	Plaza	Bar	Grocery Store	Pizza Place	Tapas Restaurant	Gastropub	Japanese Restaurant
Chamberí	1.573323	0.028753	0.036784	100	49	2	Tapas Restaurant	Café	Bar	Spanish Restaurant	Restaurant	Theater	Ice Cream Shop	Plaza	Japanese Restaurant	Italian Restaurant



- This cluster includes the most centric and expensive districts of Madrid. They also have a high to very high variety of venues including typical restaurants, bars, cafés, hotels, gastropubs, shops, etc. Most of Madrid's tourist activity and traffic lies in this cluster. Unsafety and drug occurrences are mixed. It includes Madrid Centro, which is the district with more criminal and drug cases. It is logical due to the intense tourist transit and associated amusement options.

District class 4 (label 3) description

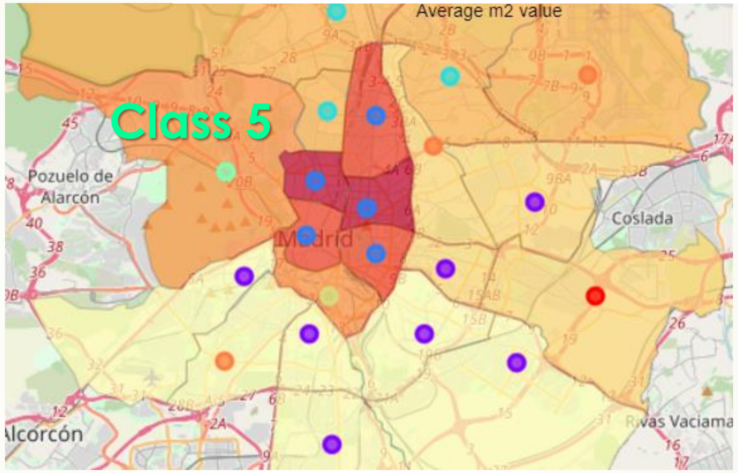
District	SQMnormMean	CrimesProp	DrugsProp	Venue	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Tetuán	1.089141	0.052239	0.026042	100	51	3	Spanish Restaurant	Restaurant	Chinese Restaurant	Hotel	Supermarket	Japanese Restaurant	Pub	Burger Joint	Paella Restaurant	Brazilian Restaurant
Fuencarral-El Pardo	1.036196	0.019974	0.038411	35	28	3	Restaurant	Spanish Restaurant	Soccer Field	Bar	Wine Shop	Tapas Restaurant	Salad Place	Fast Food Restaurant	Metro Station	Bookstore
Hortaleza	1.069037	0.027436	0.086589	48	27	3	Spanish Restaurant	Supermarket	Restaurant	Tapas Restaurant	Sandwich Place	Pizza Place	Plaza	Soup Place	Irish Pub	Coffee Shop



- This class gathers most of the northern districts which are characterized by average unit area prices, intermediate to high venue varieties that include traditional and international restaurants, spirits businesses and supermarkets. and relatively low criminal and drug cases. We could consider it a “middle class” cluster. People willing to taste different local and exotic flavors, have fun and invest with intermediate housing budget seem to fit better in this class of districts.

District class 5 (label 4) description

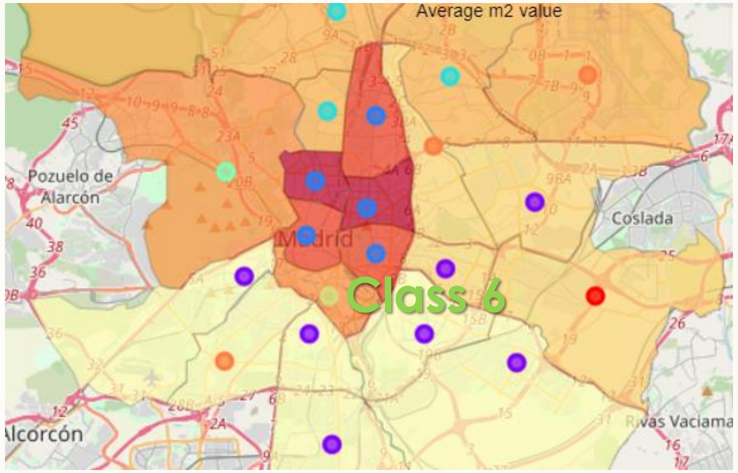
District	SQMnormMean	CrimesProp	DrugsProp	Venue	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Moncloa-Aravaca	1.168176	0.060799	0.035482	4	4	4	Hookah Bar	College Cafeteria	Park	Tennis Court	Wine Shop	Dog Run	Flea Market	Fish Market	Fast Food Restaurant	Farmers Market



- This could be tagged as the green cluster of Madrid, which only includes Moncloa-Aravaca district. There are not so many variety of places of interest compared to its neighbor districts, but they include many green areas and parks, including the famous Casa de Campo park. The property square meter prices are above Madrid's average. Despite the district's average unit area price is not as high as the central district prices (class 3), the neighborhood of Aravaca holds some of the most expensive and largest Madrid houses.

District class 6 (label 5) description

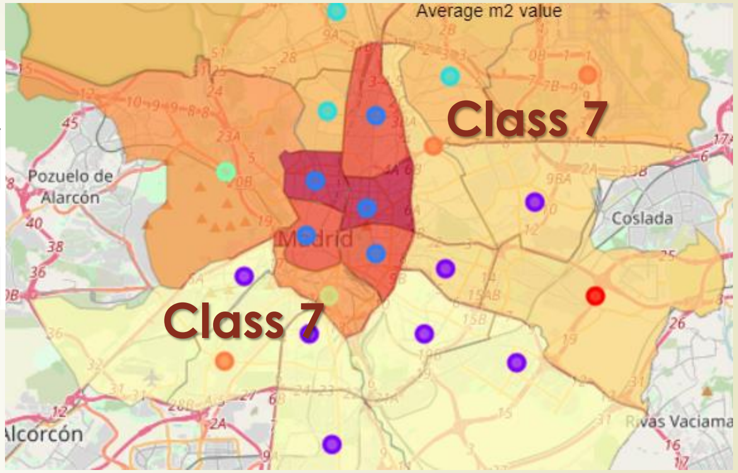
District	SQMnormMean	CrimesProp	DrugsProp	Venue	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Arganzuela	1.208997	0.02151	0.020182	100	56	5	Spanish Restaurant	Restaurant	Tapas Restaurant	Grocery Store	Bakery	Park	Gym	Bar	Market	Indie Theater



- This class only includes Arganzuela district, which is located just south of the central district. With property values above the average, it has a wide variety of venues, mainly typical restaurants and a mix of market types and recreation venues such as gyms, parks and bars.

District class 7 (label 6) description

District	SQMnormMean	CrimesProp	DrugsProp	Venue	Venue Category	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Carabanchel	0.656677	0.087138	0.045573	19	16	6	Tapas Restaurant	Restaurant	Spanish Restaurant	Candy Store	Supermarket	Bakery	Café	BBQ Joint	Cafeteria	Athletics & Sports
Ciudad Lineal	0.903449	0.048507	0.060872	86	39	6	Spanish Restaurant	Grocery Store	Restaurant	Park	Chinese Restaurant	Hotel	Italian Restaurant	Café	Bar	Pharmacy
Barajas	0.955473	0.024144	0.027018	55	33	6	Hotel	Spanish Restaurant	Restaurant	Airport Service	Tapas Restaurant	Coffee Shop	Snack Place	Duty-free Shop	Breakfast Spot	Argentinian Restaurant



- the majority of the districts included in this class, Ciudad Lineal and Barajas, are also northern districts similar to those of cluster 4 (label 3), except that these are cheaper and have more traditional restaurants, hotels, grocery stores and cafes than the other northern districts that have more specialized businesses. We also have in this cluster the southern district of Carabanchel, which has a more popular profile, probably better associated with the southern districts of class 2. However, what Carabanchel has more in common with this group's northern districts are a greater proportion of Spanish, tapas and other types of restaurants, as well as coffee businesses. Thereby, in terms of more common venues, Carabanchel is more related to its fellow cluster districts, but in terms of real estate and security indicators, it could be related to the other southern districts of class 2.

Conclusions

- The property values of Madrid seem to be mainly influenced by the variety of venues present at each district. In Madrid, the more venues there exist, the more diverse they are. It is not a monotonous city in terms of places of interest.
- Security problems apparently represent a slightly negative factor affecting the property values. Drug cases don’t seem to represent a factor that could affect the property values. But given its slight correlation with criminal cases, it is an issue not to be ignored.
- Despite the negative correlation between the unit area values and the local security incidents, we cannot affirm that the poorer areas are less safe and the wealthiest areas are safer. In fact, Madrid Centro is one of the most expensive real estate districts and it is by far the district with more criminal and drug incidents. On the other hand, Moratalaz is an economic zone and still is the district with less security incidents.

Conclusions

- Linear regression analytic method helped us understand the greater and lesser impact of the different district attributes that could be affecting the property value performance of each district. More robust predicting models should include more variables and more focus on specific neighborhoods and districts.
- The K-means clustering methods did a good work identifying similar districts mainly by their property values, variety of venues and more frequent venues. The classes followed a geographically logical pattern.
- The central districts can be described as expensive districts full with wide variety and quantity of venues. But they require the most attention by the police.
- The southern districts are economic districts that have a fair variety of venues and intermediate to relatively high intervention by the police.
- The northern districts can be described as average property price zones with an interesting variety of international gastronomic options and relatively low needs of police interventions.
- The western part of Madrid is the greenest one, with some of the most expensive neighborhoods. It is a place for people who afford to pay more for spacious homes in green environments and enjoy the relatively low variety, but open space recreational venues.
- The eastern part of Madrid is still economic and low in venue variety and quantity. But it is an expanding zone that could later become more competitive to the other districts.