Madrid District Classification Analysis

Applied Data Science Capstone: Peer graded Project

Luis Andrade, August 2021

INTRODUCTION

The objective of this study is to analyze the different patterns that characterize the different districts that conform the city of Madrid, Spain. Each of them have pros and cons depending on the people's profile. There are people to prefer to live in a district full of fun and gastronomic venues, while others prefer a quieter environment with more open-space recreational spaces. Some might pursuit one of these options assuming a high profile budget whereas others are only willing to afford economic places. Security could also be an issue for some if there were considerable flaws. We will figure out what aspects could impact more the average property value of each district and also, we will try to understand the similarities among them. To achieve these goals, we will mine some public data of Madrid, apply regression modeling to the mined data and perform some clustering techniques to group similar districts.

Before running into the scope of this study it is important to review the main aspects of the Spanish capital and understand its importance. Madrid is the capital of Spain and also Europe's second largest city, after Berlin. With approximately 6,8 million inhabitants according to 2019 projections, it is European Union's second most populated city after Paris. It also holds the third largest GDP in all EU. Madrid is the house of two of the most important European Clubs such as Real Madrid and Atlético de Madrid. Madrid residents are around 75% are Spain-born people and the rest are immigrants mainly from the former Spanish colonies in the Americas.

Madrid is also a very important touristic hotspot. It receives around 10 million visitors from all over the world who come to enjoy it's fascinating Renaissance architecture, culture, art, gastronomy and nightlife options, as well as to enjoy of the sport events, mainly the local football league (La Liga) and European championships.

Madrid is also the headquarter of important Spanish multinational enterprises such as Repsol, Telefonica, Iberia, BBVA, among other companies as well as branches of other international companies. Despite the central districts are the ones with more economical weight for the city, it's continuous expansion towards suburban areas has opened opportunities for more industries to choose Madrid as its facility headquarters.

Madrid is connected through the air by the Adolfo Suarez International Airport, in the vicinity of Barajas, which is commonly known as Barajas airport. It is Europe's sixth busiest hub with nearly 62 million travelers using this airport in 2019. Madrid is also very well connected by land with a very efficient road and railroad network to the neighboring communities and rest of the country and European Union territory. This include high-speed train services, known as AVE. Madrid also counts with a suburban rail network known as Cercanias, an urban metro network with 12 lines and supplementary bus services that keep residents and visitors very well connected from origin to destination at reasonable travel times.

The presence of many universities such as Universidad Complutense and business schools such as Instituto Empresa has attracted many students from all over the world, specially from Latin America, to Madrid. This means that more people come with location and service needs, thereby, more opportunities.

In terms of safety, according to Numbeo.com, Madrid in 2021 mid-year holds a crime index of 29.76, occupying the place 107 of 155 European cities, where Bradford, UK holds Europe's highest index of 70.08. In terms of safety index, Madrid has a score of 70.24, which means that it is the 49th safest of 155 European cities, being Zurich the safest city with an 83.67 score. What we see is that Madrid has an intermediately high performance in terms of safety, according to European references which are low if we compare to worldwide cases. Thereby, safety is not a very big concern for Madrid residents and visitors. According to several sources and reviews, most of criminal cases are related to pickpockets. Not often we hear about serious or harmful incidents in Madrid. Despite this, we don't know at this point if the security aspect has an impact in property value of the affected communities or if there are some communities more affected than others.

As we can see, Madrid is a city with multiple features distributed among its 21 metropolitan districts that fit for very diverse people profiles. We are going to extract geographical, real estate, safety and venue information for each of Madrid's district. We will mine each database and integrate the different types of information to better understand each district, find correlation of the different aspects and try to group similar districts and determine in what they are similar.

DATA DESCRIPTION AND ANALYTIC METHODOLOGIES

Data origins and descriptions

Geographical data: First, we scrapped the Wikipedia page that described each Madrid district and neighborhood by using Beautiful Soup library in order to extract the corresponding table and arrange it as data frame. Then we extracted the coordinates of each district geographical center with the aid of the *Nominatim* module of *Geopy* library. These coordinates were subsequently added to the district data frame, from where we plotted in a map through the *Folium* library each district marker according to its coordinates. These markers were later used to map district clustering results.

Real Estate data: In this work, we only focus on district property values represented as average price per square meter. After reviewing several web sources, the chose to extract values from two sources. The first one comes from the Real Estate site Idealista.com, where we extracted a table with 2021 values that were copied and pasted to an Excel spreadsheet. The second source was an article of El País-Cinco Días reporting October 2019 square meter prices, where a table image of Madrid districts was transcribed to an Excel spreadsheet. We chose to work with 2019 and 2021 data and not 2020's because the later was an odd year to the whole World. With these values we calculate for this study price variations, average prices and normalized prices to improve our observations and include them as new features in our main dataset for further analysis and modeling.

Safety data: From Madrid municipality data repository datos.madrid.es we downloaded 2021 monthly police intervention tables where they described the number of police interventions per district and

intervention type. These tables were then loaded and mined in our Jupyter notebook. Then we totalized the number of cases per category and district with the help of *Numpy* library. Afterwards, we grouped the cases by similar categories. For example, the cases where people and assets were affected and guns were seized, were grouped as an "crimes" category, because these are more serious threats to security. Drug possession and consumption were grouped as "drugs" category. These cases don't necessarily affect directly the community's security but they might affect its quality of life. From the "crimes" and "drugs" attributes we generated a proportional version of each of them, with respect to the corresponding total of cases. All these generated attributes were integrated to our main data frame and later used in our analytic and modeling steps.

Venues per district: From the Foursquare API, we extracted the main diverse venues existing in each district of Madrid. Once extracted, we then proceeded to count them per districts, to count the different venue types also per district and to group these new attributes with the real estate and security ones that served as input for our next analysis and modeling.

Analytic methodologies

Plots: From our mined data we plotted the pre-processed attributes to understand their behaviors across the different districts and to identify any existing trend or correlation among them. Through the plot observations we expect to get some initial insights about each district and understand the data inputs for the next analytic methods. To perform the plotting steps we mainly relied on *Matplotlib Pyplot, Folium* and *Seaborn* libraries. Also we plotted these attributes as Chropleth maps to understand their geographical behavior.

Linear regression: One of the questions we wanted to answer is how the quantity and variety of venues, the unsafety and drug cases affect the property values per district. We evaluate all and some of them to conclude what are the most important factors that might be related or affect the average district property values and how confident we are affirming this. For this step, we relied on the *Scikit-Learn's Linear Regression* module.

K-Means clustering: Our next analytic step is the clustering of the districts according to their different features such as property prices, safety and drug indicators, total and variety of venues and the quantity of each venue type. We selected this method because it allowed us to choose the quantity of classes that we wanted: A fair quantity to sufficiently differentiate the districts and keep the same class districts as similar as possible. We used as analytic tools the *KMeans* module of the Cluster *Scikit-Learn* library to perform the clustering step and *cdistance* module from *Geopy's Spatial Distance* library to help us quantify the internal differences of the resulting clustered classes. Once again, we will then use the *Folium* library to plot the district markers colored and labeled with their class output over background Choropleth maps representing different district attributes. Afterwards we are going to check the districts included in each class along with their attributes to discuss and understand what they have in common and how they can work for the different people profiles and interests.

District Geographical Data Mining

Thanks to the scrapping of the Wikipedia page dedicated to the districts of Madrid and the use of *Nominatim* library to extract each district central coordinate, we could build a data frame that described for each district, its number of District, Neighborhoods, latitude and longitude, which is visible at figure 1:

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

Figure 1: List of Madrid's districts and neighborhoods with district coordinates

Thanks to this mined information we could then generate our first geographical plot that was generated with the *Folium* library which allowed us to visually locate each district and better explore the locations of the neighborhoods of Madrid, as we can see on figure 2.

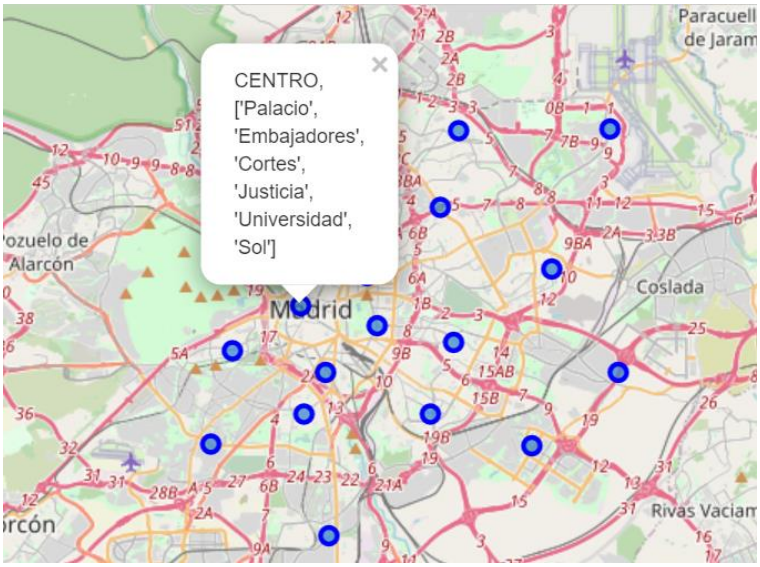

Figure 2: Folium map with markers containing district and neighborhoods information, centered on each district central coordinates

District Unit Area Value Exploration

As mentioned in the introductory section, we proceeded to extract average square meter price per district from two web sources:

- October 2019 average district prices from an article of the media site El País-Cinco Días, part of the prestigious El País Spanish journal. The necessary information was extracted from an image chart and put into an Excel spreadsheet that was later imported in our project notebook.
- Mid-year 2021 average district prices from Idealista real estate platform. There was a table that was later accommodated in an Excel spreadsheet that was then imported to our project notebook.

After some preprocessing and data selection we then proceeded to perform some calculations:

- Average: We extracted the average of the square meter prices per districts from both sources and times to assure more data consistency and stability along time.
- Variation: We estimated the price variations from 2019 to 2021 in order to check their variations
- Normalized to mean value prices: We divided each district average price by the mean of all average prices in order to have as reference a relative district value from Madrid mean value.
- Linear Normalization: District price minus the minimum price divided by the price range. This parameter works as input for the future classification exercise that require variable scaling processes. 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

Figure 3: Average square-meter prices per districts, 2-year-variation and normalized means

The resulting table, illustrated on figure 3 showed that from 2019 to 2021 the prices changed in a range from -1.53% to 20.86%. Carabanchel (southwest Madrid) was the only district to have the negative variation. Hortaleza (northeast Madrid) had the highest variation. In average, Madrid districts increased their values in 7.89% in this 2-year term.

Then we plotted each average district price as bars and included as reference a line representing the average Madrid price of 3857 euros/m2, as we can see on figure 4. From there we can see what districts are above and below this global reference, practically split 48% above and 52% below the average. The prices range from 1694 euros/m2 in Villaverde (south Madrid) to 5606 euros/m2 in Salamanca (downtown Madrid). The price variation per district goes gradually, with some minor jumps and plateaus. At least in Madrid we don't see important gaps of price ranges which in turn mean that there are options for almost all budgets between 1700 and 5600 euros/m2.

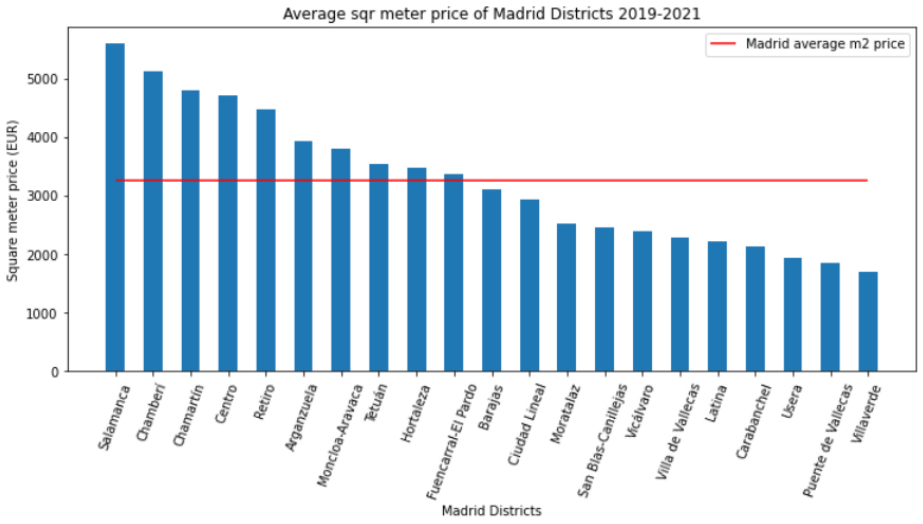

Figure 4: Average square meter prices per district. Red line represents the city's average

Plotting the area prices in a Choropleth map give us a better idea of how the prices change geographically. Downtown districts turn to be the most expensive areas. On the other hand, the southern and eastern districts have the lowest prices. The northern and western districts offer intermediate prices. Of course, if we go into more detail we will find neighborhoods moderately or significantly more expensive than others in a same district. But that could be the scope of a more detailed analysis.

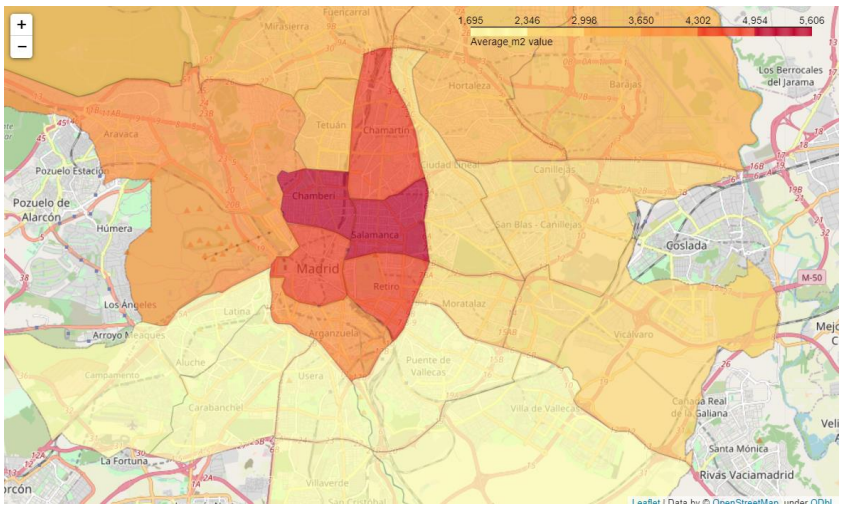

Figure 5: Average square-meter prices per districts in euros

District Security Indicators

From the web site datos.madrid.es we downloaded monthly reports of police interventions per districts that are separated per categories. We downloaded six spreadsheet files with reports from January to June 2021. Then we integrated these tables to estimate the total values per category and district for this first half of 2021 period. The reported categories were related with people, with assets, weapon possession, drug possession and drug consumption. In order to simplify the categories, we grouped the first three categories as security incidents or "crimes", a short alias to better handle coding names, and the drug reports as "drugs" categories. In other words, we summed person, asset and weapon cases as "crimes" and drug possession and consumption as "drugs". We made this gathering step based on the premise that the people trend to be more concerned about assaults and/or weapon incidents that represent personal threats rather than drug problems that often involve more health problems. This doesn't mean that we discard any connection between these two generated categories.

Also, we generated proportional calculations of criminal and drug cases compared to their corresponding category totals. These proportions will serve as inputs for the classification steps.

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

Figure 6: Totalized security incidents, grouped in Crimes (People, Assets and Weapons) and Drugs (consumption and possession) categories, absolute and proportional versions

The table illustrated in figure 6 could be represented as bar plots ordered by total cases reported by the police with the details of criminals and drug cases per districts. From the bar plot showed in figure 7 we realize that Madrid Centro stands out for its considerably high quantity of cases compared to the rest of districts. We believe this is related to the high presence of tourists that are target of burglars and/or drug dealers.

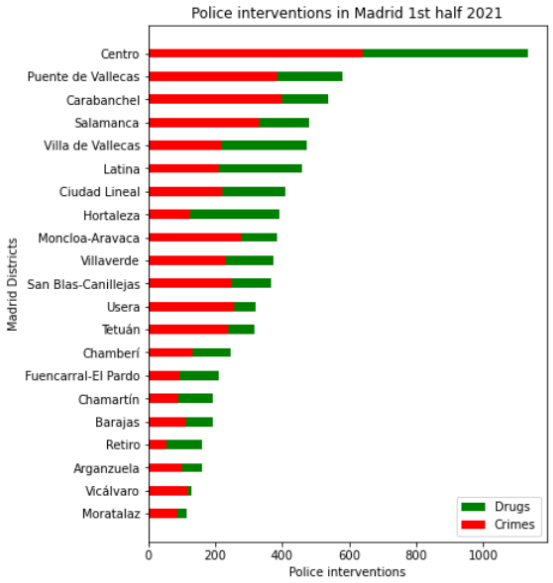

Figure 7: Total Madrid police interventions per category per district

Now we can observe how the criminal and drug cases are geographically distributed through the following choropleth maps illustrated in figure 8:

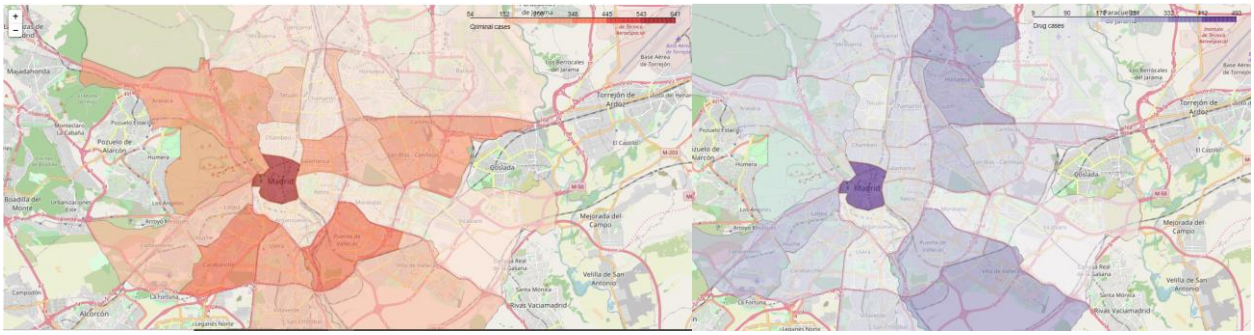

Figure 8: Left: Criminal cases per districts. Right Drug cases per district

From the Crime choropleth map, figure 8 left, it looks obvious that Madrid Centro is a relatively hot spot in terms of security. But this concern doesn't seem to chase away the tourists or the investors. We also can observe as general overview that the northern districts are relatively safer than the southern ones. By observing the drug cases, figure 8 right, it seems that the central districts have less drug issues than the border districts. Of course, this excludes Centro (Downtown) which exhibits the highest number of cases related to drugs. Also, in general, eastern and southern districts have more drug cases than the western and northwestern districts.

Before proceeding to the next step, we checked if there was any correlation between criminal cases and drug reports. We generated a crossplot with the “regplot” option of the Seaborn library, which can be seen on figure 9. It turns out that there seems to be a positive correlation between both parameters. However, the degree of dispersion might be considered as significant.

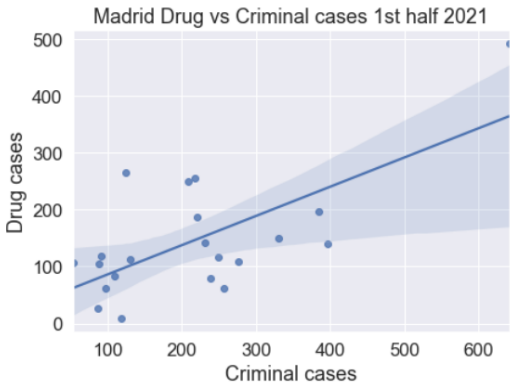

Figure 9: Crossplot of drug cases compared to criminal cases, showing a slight positive correlation

Venue Extraction per District with FourSquare

Part of our Madrid district analysis includes the determination of the quantity and variety of venues that each district has, which may help to differentiate the districts or associate them. By using the FourSquare API, and using our personal credentials we could perform a global venue search centered on each Madrid district coordinates with a search radius of 1 kilometer.

The venue search was accomplished by retrieving from a URL request the 'json' 'item' responses that allowed us to populate a table with venue names, coordinates and categories. The resulting table had the layout illustrated on figure 10:

	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

Figure 10: Venue table obtained from Foursquare API venue search

From there, it was possible to count the number of venues and the number of different venue categories associated to each district. We observed the relationship between these two variables, which turned to be high, as we can see on figure 11.

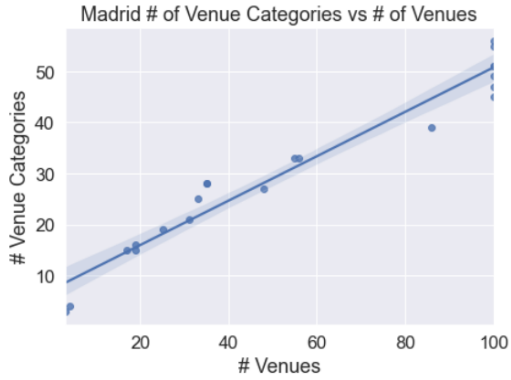

Figure 11: Crossplot of the relationship between the number of venues and their variety per district

The venue quantities seem to be clipped due to Foursquare search limitations. But the venue categories don't seem to be clipped and it highly correlates with the venue quantities. Thereby, we proceed to generate a bar plot with the quantity of different venue categories per district. The resulting bar plot is showed on figure 12.

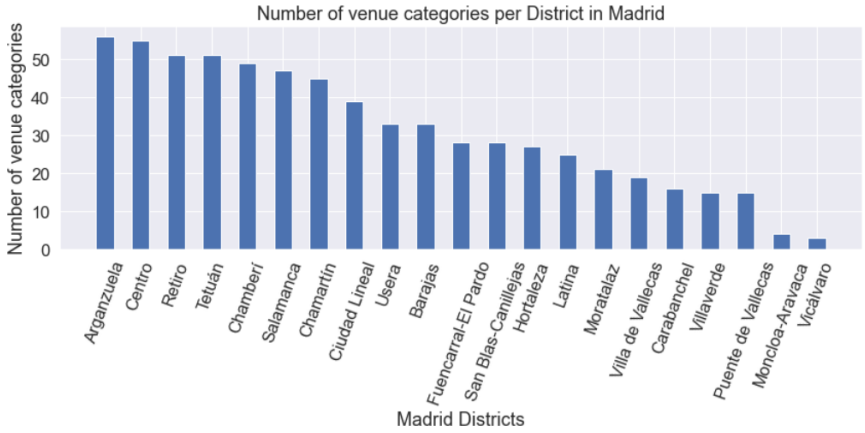

Figure 12: Variety of venues per Madrid district in 1-Km-radius from each district central coordinates

The districts inside or near downtown Madrid are the ones with more variety of venues. Naturally, as we get farther from downtown the diversity becomes low to very low. These new venue attributes were gathered with the previous real estate and security indicators to generate the necessary features for our next multilinear regression and clustering steps.

ANALYTIC METHODS

Linear Regression Modeling

At this point we have already processed for each district parameters like criminal cases, drug cases, average square meter price, number of venues and number of venue categories. With the help of *Seaborn* library, we generated pair plots for all the variables to visually observe trends among all the possible pairs.

	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

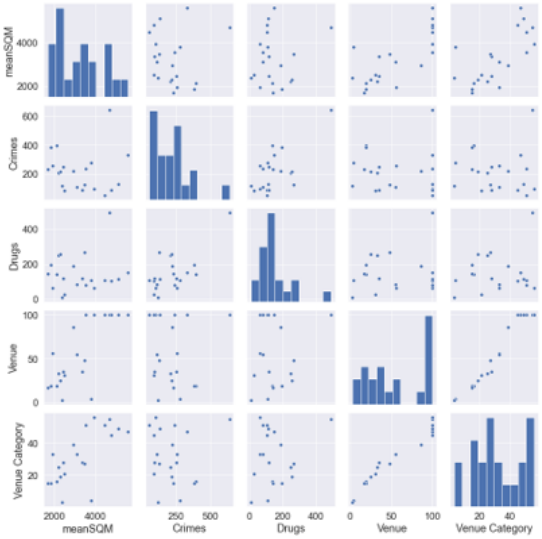

Figure 13: Crossplots of all the variables considered in our analysis (pairplots)

In most cases, shown on figure 13, the correlations were not very clear. 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 crime quantities. Crime and drug cases seem to have some degree of correlation. Number of venues and venue categories show strong correlation between them.

Since we have several parameters, we are going to test some multilinear models where we try to predict the square meter price per district based on security flaws, drug cases and venues.

```
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
```

Figure 14: Multilinear regression model considering all the variables

Our first model considered all the available variables. Its results are shown on figure 14. This model with a modest variance score 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. This doesn't seem very logic since we previously 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. We chose to stay with venue category variable because the number of venue variable showed evidence of clipping that affects the linear correlation.

```
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
```

Figure 15: Multilinear regression model considering all the variables except number of venues

Our second model took into account the crime cases, drug cases and variety of venues. The results are displayed on figure 15. 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. Since we previously saw some degree of correlation between crimes and drugs, we proceeded to our next model even simpler were we stayed with one of the security categories. We favored in this case the number of crimes, since it visually better correlates with the unit area prices and we consider that crimes negatively affect more the people's willingness to invest or rent in specific locations than the drug cases.

```
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
```

Figure 16: Multilinear regression model considering total crimes and number of different venue categories

Our third model was simpler but as accurate as the second model and it is displayed on figure 16. It suggests that for every new crime the average square meter price drops by 0.25 euros. On the other hand, for every new venue category present in a district, the square meter price increases 49.8 euros. Of course, we should not rely on this model to try to accurately predict property prices for a district because of the low model accuracy, the broad scale of the analysis and the limited variables taken into account. There are many more variables that impact the property values that require more localized studies and feature selections, which are out of the scope of this study. However, this model 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 slightly affected by the occurrence of security incidents that require police intervention.

K-Means classification

After having got some insights with the multilinear models about the possible factors that more likely affect the district's property values, we moved forwards and proceeded to the clustering analysis. The purpose of this analytic method was to group the districts into different classes according their different features such as property prices, criminal and drug indicators, total and variety of venues and the quantity

of each venue type. As mentioned in the introductory section, we chose the K-means method because it allowed us to determine the quantity of classes we wanted the different districts to be grouped at. But this number of classes had to be fair enough to guarantee that the distortion and inertia (measures of how big are the Euclidean distances between points and cluster centroids inside each cluster) were low enough. That is, the districts within a class had to be as similar as possible among them and as different as possible to the other district classes. We performed some tests with the Elbow method to select a number of classes high enough to reduce the Euclidean distances of the clusters and low enough to assure a number of classes significantly lower than the total of samples (districts in our case). We used as tools the *KMeans* module of the *Cluster Scikit-Learn* library to perform the clustering step and *cdistance* module from *Geopy's Spatial Distance* library to help us quantify the internal differences of the resulting clustered classes.

Prior to our clustering steps we still needed to prepare part of our input data. First, we needed to determine the proportions of each venue category per district, ranging from 0 to 1. Then, it was necessary to transform each venue output to one hot encoding where the corresponding venue category as column was assigned 1 value whereas the rest of non-corresponding venue categories, also as columns, were assigned 0 values. Afterwards, the results were grouped by district and mean values were estimated, also ranging from 0 to 1. As result, we obtained a table with proportion of venue categories per district, as it can be seen on figure 17.

	District	Accessories Store	Airport	Airport Lounge	Airport Service	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art Studio	Asian Restaurant	Athletics & Sports	Auto
0	Arganzuela	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.01	0.020000	0.02	0.01	0.00	0.010000	0.000000	0.(
1	Barajas	0.00	0.018182	0.018182	0.054545	0.000000	0.00	0.00	0.036364	0.00	0.00	0.00	0.000000	0.000000	0.(
2	Carabanchel	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.00	0.000000	0.052632	0.(
3	Centro	0.01	0.000000	0.000000	0.000000	0.010000	0.00	0.00	0.000000	0.00	0.01	0.00	0.010000	0.000000	0.(
4	Chamartín	0.00	0.000000	0.000000	0.000000	0.010000	0.01	0.00	0.000000	0.00	0.00	0.00	0.000000	0.010000	0.(
5	Chamberí	0.00	0.000000	0.000000	0.000000	0.010000	0.00	0.00	0.000000	0.00	0.01	0.00	0.010000	0.000000	0.(
6	Ciudad Lineal	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.023256	0.00	0.00	0.00	0.000000	0.023256	0.(
7	Fuencarral-El Pardo	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.00	0.000000	0.028571	0.(
8	Hortaleza	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.020833	0.00	0.00	0.00	0.000000	0.000000	0.(
9	Latina	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.00	0.00	0.00	0.030303	0.030303	0.(

Figure 17: Different venue categories transformed to one hot encoding and grouped by mean criterion, with value ranges from 0 (no existence) to 1 (total existence in 100% of cases)

We also generated a table with the ten most common venues per district, which we listed and showed on figure 18. This information would not be part of the K-means clustering process but would be used for qualitative description of the grouped districts by cluster. The steps required to build this table included column creation according to the defined number of top venues (ten in our case), sorting in descending order (from greater to lower) the row values representing the frequency of venue categories and retrieving the name of each venue category.

	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

Figure 18: The most common features per district

Going back to our K-means clustering process, we had our input data per district included the mean real estate normalized values, crime proportions, drug case proportions, proportional quantity of different venues (compared to the total venue categories), and proportions of each venue category.

We then performed the Elbow method tests with different number of classes “k” and observed the resulting distortion and inertia values.

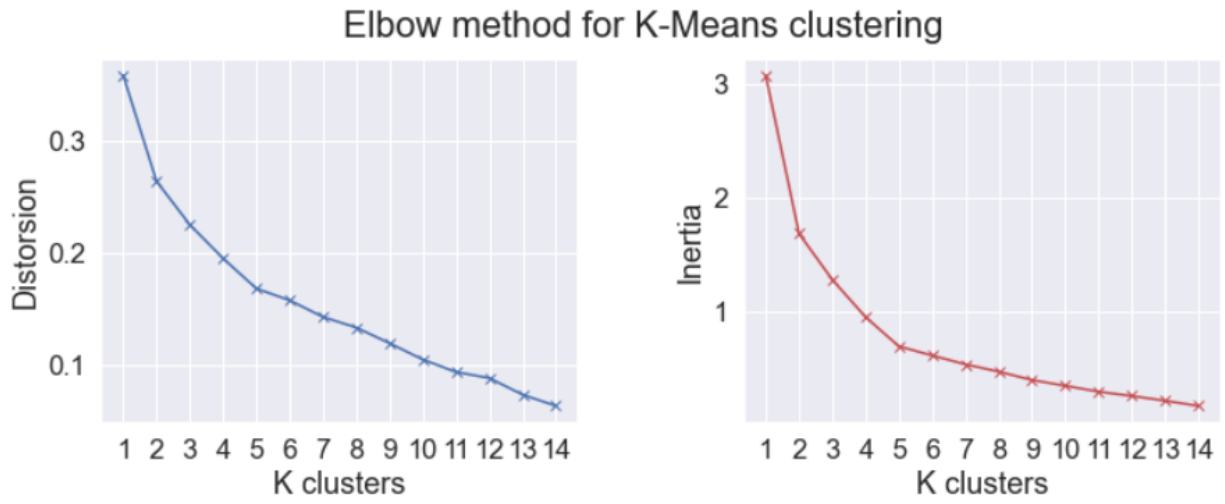

Figure 19: Euclidean distances represented as distortion and inertia for each number of k classes

Observing the results on figure 19, we realize that from the distortion perspective, k beyond 10 does not drop it considerably. We observe a distortion elbow at k=5, but in my opinion its value is still slightly high. After observing the inertia plot, we find a more remarkable elbow at k=5. Beyond that value, inertia starts to drop more slowly. 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.

We then grouped the districts into 7 classes with the K-means method. It generated a class label that was then appended to our merged data frame. Then we proceeded to map the district classes as markers in

Folium maps over choropleth maps of square meter prices, crimes and number of different venue categories per district. We can observe them on figure 20.

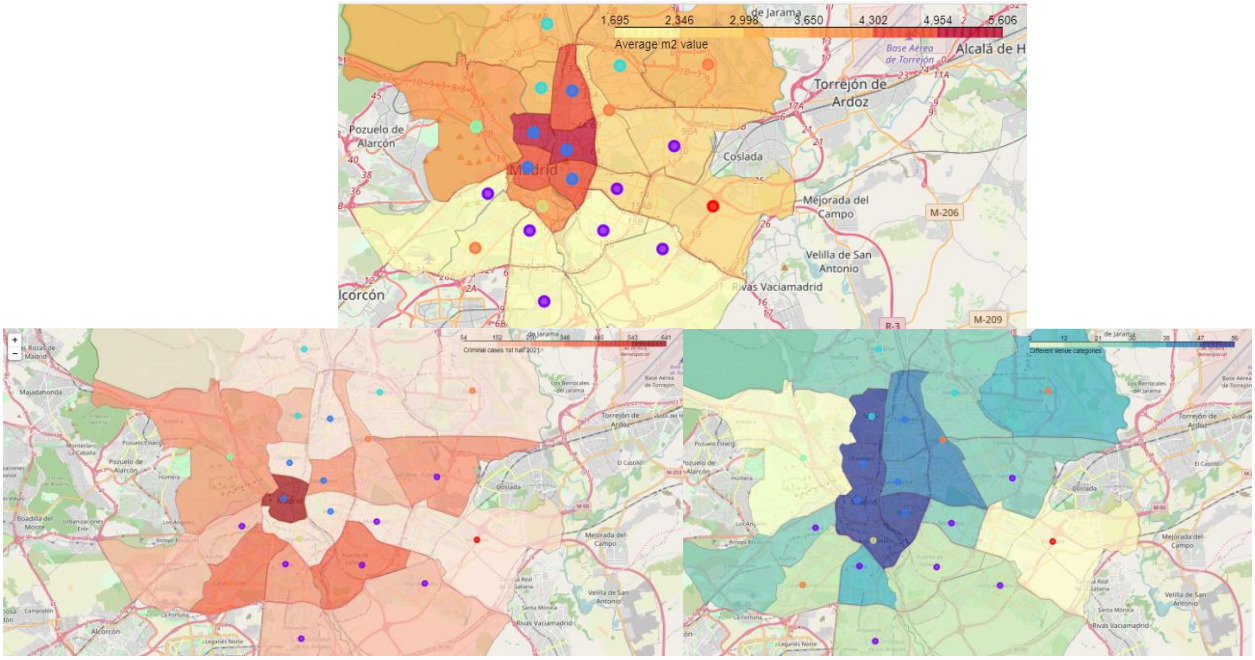

Figure 20: Districts grouped in 7 classes with different marker colors compared to the unit are prices, crimes and number of different venue types per district

Visually we realize that most of the central districts, the ones with the most expensive unit areas, share the same class. There is a class that gathers most of the southern districts and part of the eastern ones, which show average to below-average unit area prices, intermediate amount of different venue categories and intermediate to high crime indicators. The northern districts are gathered in 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.

Classification result analysis

After completing the visual observation of the geographical trend of the cluster classification, we then analyzed the result of each class based on the input clustering features and the 10 most common venue categories per district that we previously ordered and tagged. Please note that the output labeled classes start from 0 and reach to 6, giving 7 labels in total. But in the descriptive analysis we do in this section we name each class starting from 1 to 7. In other words, class 1 is label 0, class 2 is label 1 and so on until our last class number 7 that corresponds to label number 6.

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

Figure 21: Different features of class # 1 (label 0)

Class 1 (label 0): Figure 21 summarizes our first class. 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	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

Figure 22: Different features of class # 2 (label 1)

Class 2 (label 1): This class, as we can see on figure 22, 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	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

Figure 23: Different features of class # 3 (label 2)

Class 3 (label 2): This cluster described in figure 23 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	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

Figure 24: Different features of class # 4 (label 3)

Class 4 (label 3): This class listed in figure 24 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	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

Figure 25: Different features of class # 5 (label 4)

Class 5 (label 4): This could be tagged as the green cluster of Madrid, which only includes Moncloa-Aravaca district. As we can see on figure 25, 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 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	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

Figure 26: Different features of class # 6 (label 5)

Class 6 (label 5): It only includes Arganzuela district as it can be seen on figure 26. This district is located immediately 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	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

Figure 27: Different features of class # 7 (label 6)

Class 7 (label 7): From figure 27 we observe that 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, which have more specialized businesses. We also see included 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 that could be found on 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, from recreational to gastronomic.

Security problems apparently represent a negative factor affecting the property values, but slightly, since Madrid is not particularly a very unsafe city in Europe. Drug cases don't represent a factor that could positively or negatively affect the property values. However, there is a slight correlation between drug cases and other criminal cases. It is an issue not to be ignored.

Despite that there exists some 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 shows square meter prices below Madrid average and still is the district with less security incidents.

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 models for predicting purposes should include more internal and external variables and probably 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, more frequent venues and to a lesser degree, the criminal and drug-type incidents. The classes followed a geographically logical pattern, even though geographic coordinates were not included in the clustering process.

The central districts can be grouped and described as expensive districts full with wide variety and quantity of venues. But they require the most attention by the police.

The southern districts can be grouped and described as economic districts that have a fair variety of venues and intermediate to relatively high intervention by the police.

The northern districts grouped 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 can afford to pay more for spacious homes surrounded by green environments and enjoy the relatively low variety, but open space recreational venues.

The eastern part of Madrid is still relatively economic and low in venue variety and quantity. But it is an expanding zone that could later become more competitive to the other districts.