Coursera Data Science Project:

The Battle of Neighborhoods

Juan Prieto

27-5-2020

Index

Introduction

This is the report for the IBM Data Science Professional certificate capstone project. Here, all the steps taken in the code notebook will be explained.

This project will allow the certificate candidates to demonstrate the knowledge learnt during the eight previous courses. The topic is free, and the only constraint is to use Foursquare location data (by using their API) as one of the data sets employed for the project.

Business Understanding

Background

Madrid is not only Spain's capital, but also its principal business hub. It is in almost the geographical center of the Iberian Peninsula and Spain's infrastructure stems from the city in a radial network. This makes Madrid very attractive for business due to its beneficial fiscal aids from its local and regional government, the ease of travel from and to the city and its culture of entrepreneurship.

The city is a major center for banking, finance, retailing, trade, media, services and tourism not only in the country, but in the European southwest. Madrid is an alpha city according to the 2018 Globalization and World Cities Research Network[1] (GaWC). According to GaWC, "Alpha level cities are linked to major economic states and regions and into the world economy, and are classified into four sections, Alpha ++, Alpha +, Alpha, and Alpha − cities".

Due to this, the cost of living in Madrid is higher than in other Spanish cities, but the socioeconomic differences between the different neighborhoods of the city can be used to find affordable places to live.

Problem description

Recently, a business in the city has been expanding its operations and needs to recruit talent from outside the city (both nationally and internationally). To help their prospective employees, they want to analyze the level of living throughout the city to better recommend their new workforce where to settle. The insights derived from this analysis will give a good understanding of the socioeconomic conditions of the different neighborhoods of the city and will allow to tailor real estate recommendations to all their prospective employees independently of their salaries.

The aim of this is to increase worker satisfaction with the company and reduce the employee churn rate and retain talent. A good worker satisfaction would also raise the customer's opinion on the company, with a possible positive effect on their sales.

[1] https://www.lboro.ac.uk/gawc/world2018t.html

The key indicators employed to analyze Madrid's neighborhoods will be:

- Population
- Average income
- Crime level
- Amenities in the neighborhood
- Real estate and rent prices (per square meter)

Target Audience

The objective is to study the socioeconomic levels of the city in order to provide housing and living expenses recommendations to new employees of the client company. The company's management also expects to understand the rationale behind the recommendations made.

The insights extracted from this analysis would also interest anyone interested in living in the city.

Success criteria

The project will be considered successful if a tiered list of Madrid's neighborhoods based on socioeconomic and business diversity in the neighborhood can be presented to the client to inform its prospective employees of their living choices in the city.

Data Understanding

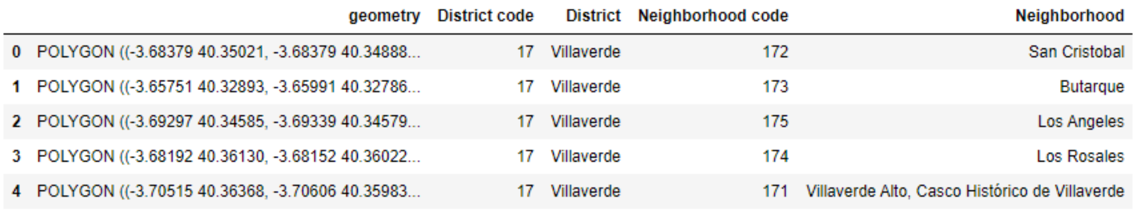
Several datasets pertaining the city of Madrid (Spain) will be used. All of them will be described in the following section. In this section an exploratory analysis of the data will also be performed.

Geographical data

All geographical data was taken from one source: The list of administrative divisions (current districts and neighborhoods) of Madrid, taken from the open data page of the city of Madrid. The files in geographic format for both neighborhoods and districts (*Barrios en formato geográfico* , and *Distritos en formato geográfico* , respectively) contain not only the neighborhood and district names and codes for each of them, but also the polygon shapes (in .shp format) so they can be placed in a map. This files were converted to geojson format by means of an external converter. It is important to specify the output geographical coordinate reference system of the data to WGS84, which is the standard for geojson format. Once this is done, the geojson files in '/data/geodata' are obtained, for both the neighborhoods and the districts.

The areas of the neighborhoods were taken from the csv files from "Relación de barrios (superficie y perímetro)" (List of neighborhoods, surface and perimeter) from the same page as the geographic files.

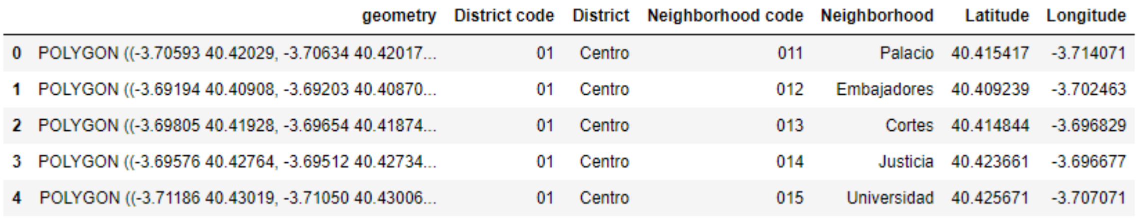
The first thing that will be done is extracting the useful data (neighborhood name and number, district name and number, and geographical coordinates of the neighborhoods) from the geojson file.



	geometry	District code	District	Neighborhood code	Neighborhood
0	POLYGON ((-3.68379 40.35021, -3.68379 40.34888...	17	Villaverde	172	San Cristobal
1	POLYGON ((-3.65751 40.32893, -3.65991 40.32786...	17	Villaverde	173	Butarque
2	POLYGON ((-3.69297 40.34585, -3.69339 40.34579...	17	Villaverde	175	Los Angeles
3	POLYGON ((-3.68192 40.36130, -3.68152 40.36022...	17	Villaverde	174	Los Rosales
4	POLYGON ((-3.70515 40.36368, -3.70606 40.35983...	17	Villaverde	171	Villaverde Alto, Casco Histórico de Villaverde

Figure 1 Geopandas dataframe with the geometry of the neighborhoods, their name and code and their district name and code.

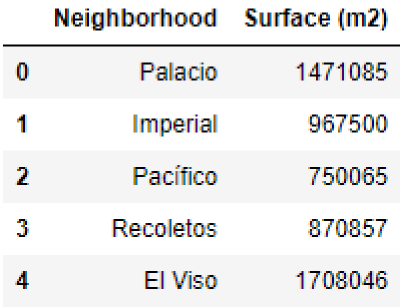
The centroid for each neighborhood was calculated and their latitudes and longitudes added to the dataframe.



	geometry	District code	District	Neighborhood code	Neighborhood	Latitude	Longitude
0	POLYGON ((-3.70593 40.42029, -3.70634 40.42017...	01	Centro	011	Palacio	40.415417	-3.714071
1	POLYGON ((-3.69194 40.40908, -3.69203 40.40870...	01	Centro	012	Embajadores	40.409239	-3.702463
2	POLYGON ((-3.69805 40.41928, -3.69654 40.41874...	01	Centro	013	Cortes	40.414844	-3.696829
3	POLYGON ((-3.69576 40.42764, -3.69512 40.42734...	01	Centro	014	Justicia	40.423661	-3.696677
4	POLYGON ((-3.71186 40.43019, -3.71050 40.43006...	01	Centro	015	Universidad	40.425671	-3.707071

Figure 2 Geopandas dataframe with the centroids' latitude and longitude from each neighborhood polygon added.

The data from the list of neighborhoods, surface and perimeter was also extracted:



	Neighborhood	Surface (m2)
0	Palacio	1471085
1	Imperial	967500
2	Pacífico	750065
3	Recoletos	870857
4	El Viso	1708046

Figure 3 Dataframe with Surface information per neighborhood

In this dataframe, there were three neighborhoods without surface information. The neighborhoods of "Ensanche de Vallecas", "Valderrivas", and "El Cañaveral" had a surface, according to the dataframe, of 0 m^2. The surface was added manually with appropriate data taken from Google Maps.

There were also some names written differently in both dataframes, so the names were scanned to search for differences in spelling and then amended. Both dataframes were merged, and a new one was created with the following information:

1. Neighborhood code.
2. Neighborhood name.
3. District name.
4. District code.
5. Latitude of the neighborhood centroid.
6. Longitude of the neighborhood centroid.

7. Surface of the neighborhood.

In Figure 4 the histogram of the neighborhood surface will be shown.

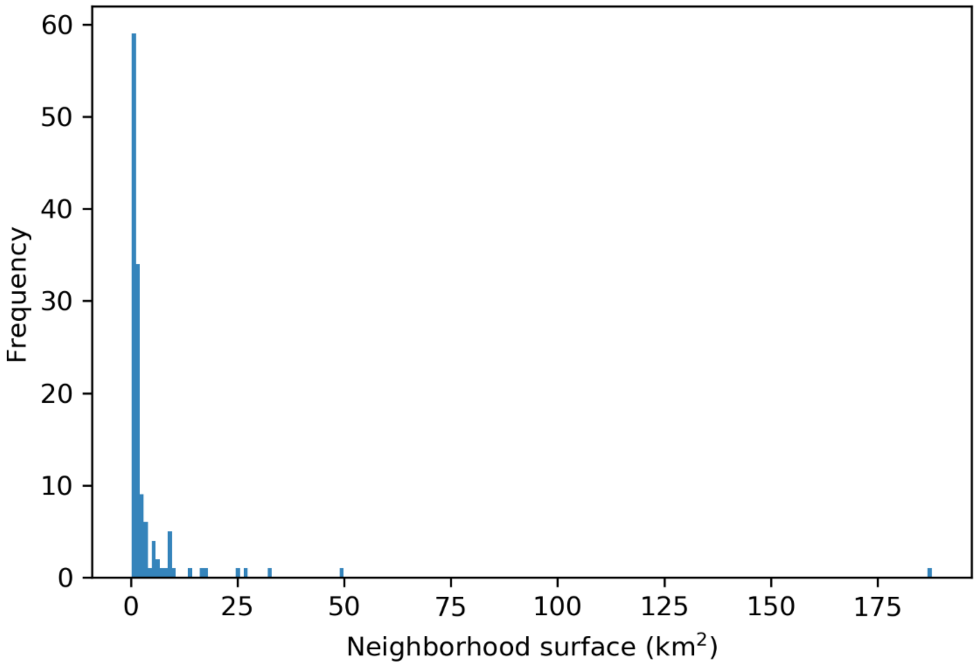


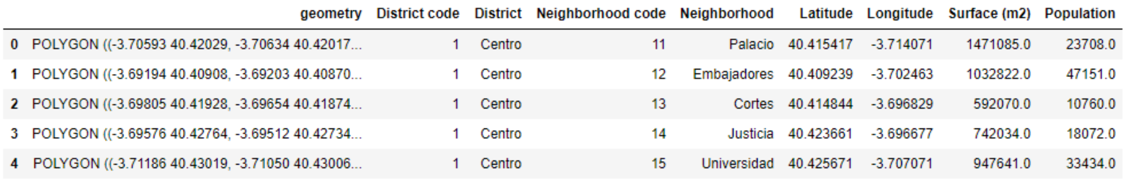
Figure 4 Histogram of the Surface of Madrid's Neighborhoods. Most of the neighborhoods have a surface of less than 15 km^2, and there is a neighborhood with a surface value of more than 175 km^2.

The mean surface is of 4.76 km^2, with a standard deviation of 173.21 km^2, due to the extreme value of one neighborhood. The 75-percentile has a value of 2.43 km^2, showing that most of the neighborhoods are in the first bins.

Next, in Figure 6, a map of the neighborhoods with their centroids will be presented.

Population data

Population data for the city of Madrid was taken from the data bank of the City Council of Madrid, here. The information extracted is the number of inhabitants per neighborhood. After cleaning the data, the population information was added to the main dataframe, as shown in .



	geometry	District code	District	Neighborhood code	Neighborhood	Latitude	Longitude	Surface (m2)	Population
0	POLYGON ((-3.70593 40.42029, -3.70634 40.42017...	1	Centro	11	Palacio	40.415417	-3.714071	1471085.0	23708.0
1	POLYGON ((-3.69194 40.40908, -3.69203 40.40870...	1	Centro	12	Embajadores	40.409239	-3.702463	1032822.0	47151.0
2	POLYGON ((-3.69805 40.41928, -3.69654 40.41874...	1	Centro	13	Cortes	40.414844	-3.696829	592070.0	10760.0
3	POLYGON ((-3.69576 40.42764, -3.69512 40.42734...	1	Centro	14	Justicia	40.423661	-3.696677	742034.0	18072.0
4	POLYGON ((-3.71186 40.43019, -3.71050 40.43006...	1	Centro	15	Universidad	40.425671	-3.707071	947641.0	33434.0

Figure 5 Dataframe containing population data for all neighborhoods of Madrid.

A histogram of the population can be seen in Figure 7.

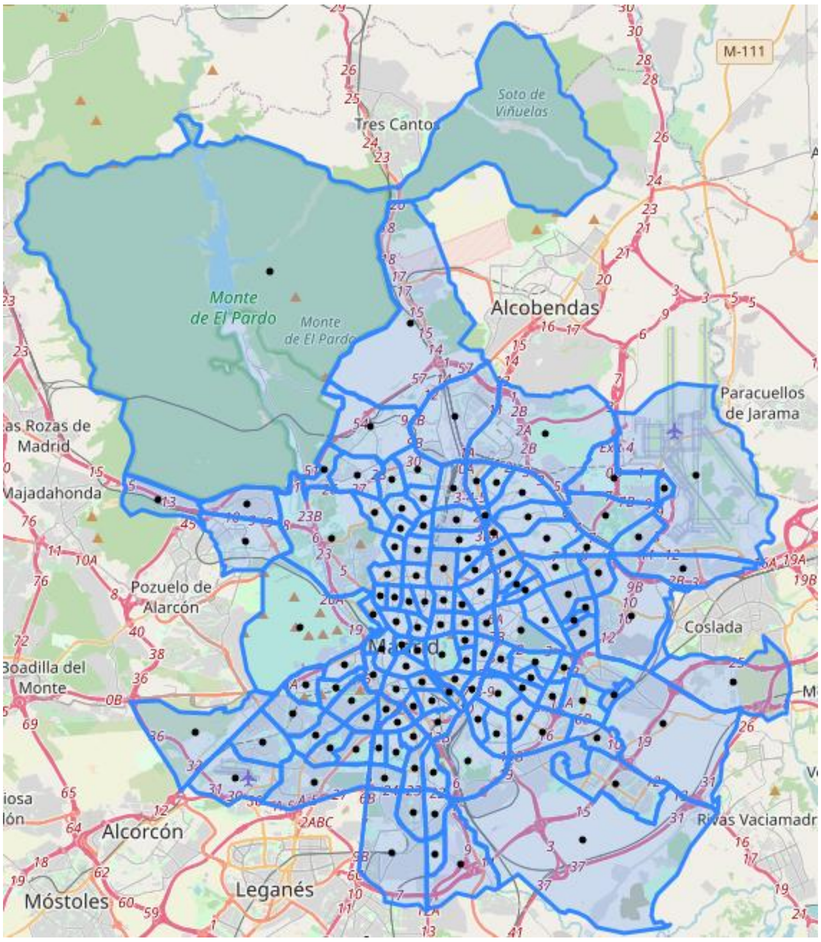


Figure 6 Map of the different neighborhoods of Madrid (blue surfaces), along with its centroids (black dots). An interactive map is available in the GitHub repository.

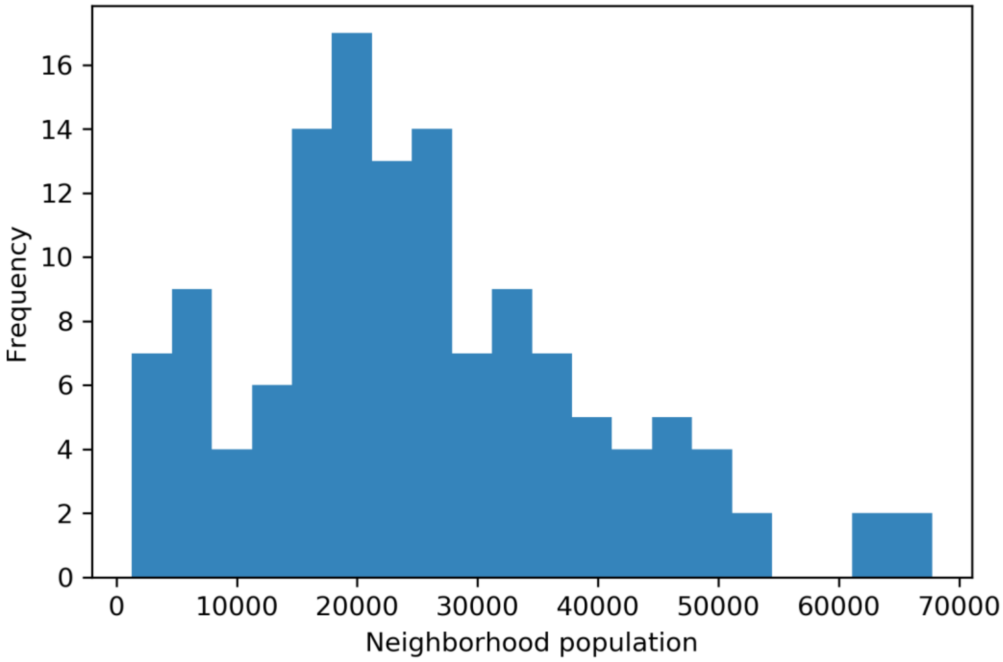


Figure 7 Histogram of the different population of the neighborhoods of Madrid.

In the case of the population values per neighborhood the data is more grouped than the surface of the neighborhoods. 75% of the neighborhoods have a population less than approximately 34000 people, with the most populous neighborhood housing around 68000 people. In this case the mean and the median are approximately the same, with a difference of about 8% between those two numbers. Next, in Figure 8, the population will be visualized in a map.

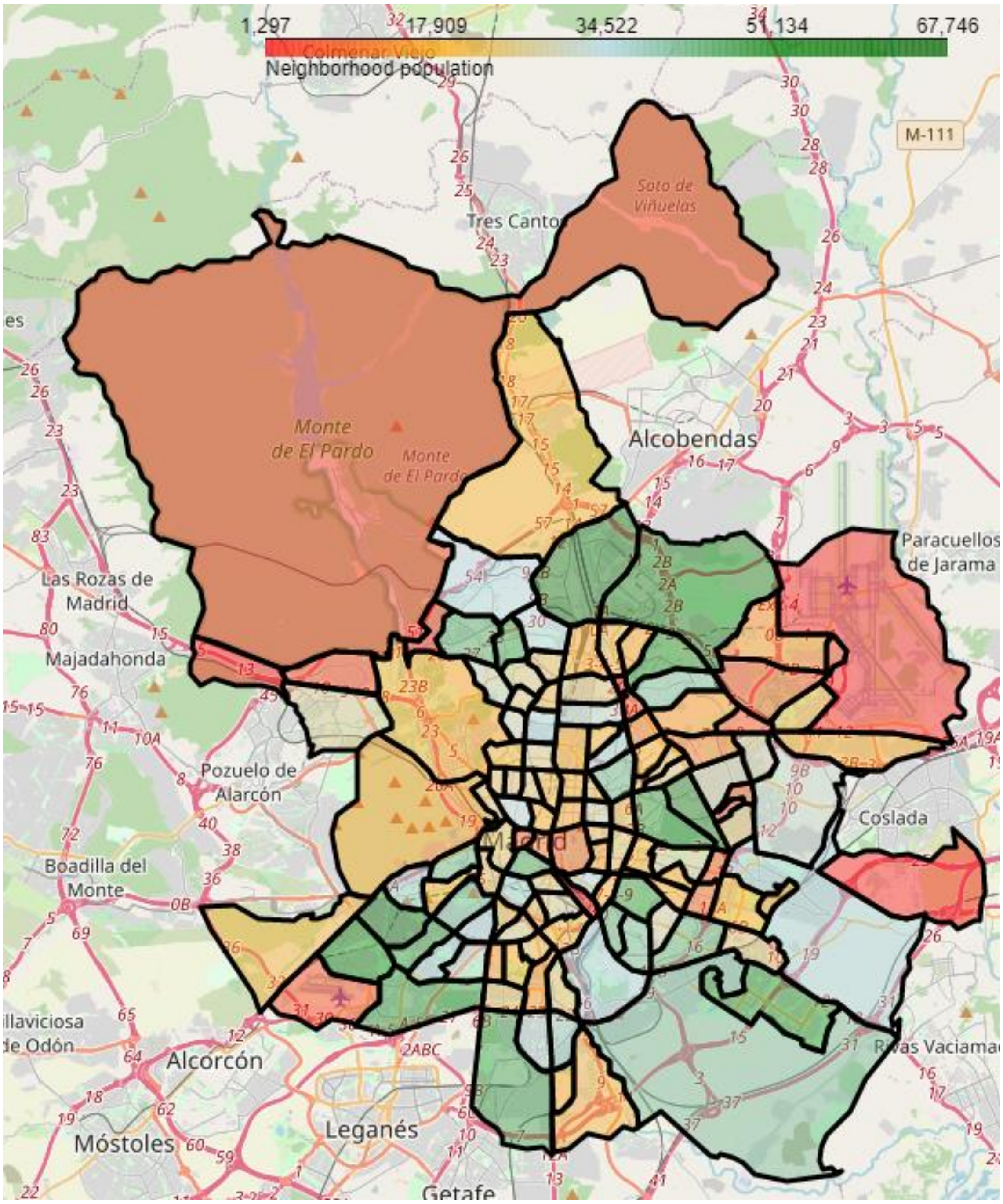


Figure 8 Color map of the different neighborhoods of Madrid. Red indicates less population, whereas green indicates more. An interactive version of the map exist in the GitHub repository.

Income data

Data corresponding to the average income per person for the year 2017 was taken from Spain's Instituto Nacional de Estadística, INE, (National Statistics Institute). This data was taken at a district level, as there is no data at neighborhood level (there is data on a sub-neighborhood level, but sadly we do not have any geojson files regarding those divisions). The downloaded file is a csv with all the district data for the city of Madrid. The data was cleaned and added to the main dataframe. A histogram of the income data per district per capita can be seen in Figure 9.

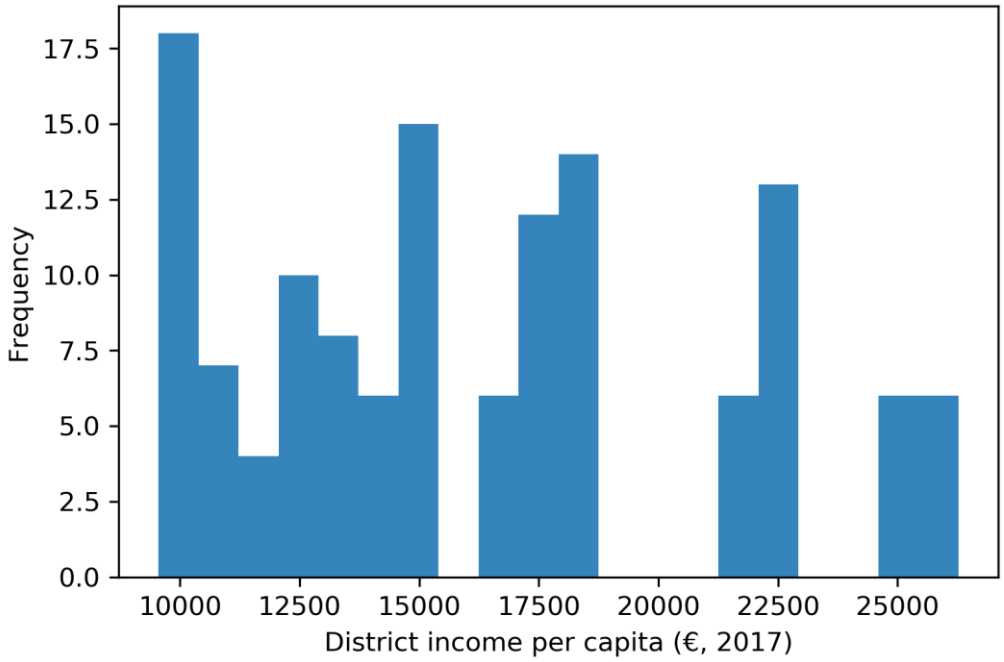


Figure 9 Histogram of the district income per capita. There was no information available at a neighborhood level, so the district level information was employed.

The mean value is around 16400 €, and the median value is 15180 €. There is a clear trend of lower income frequency as the average income per person goes up. This information can also be visualized in a map, as shown in Figure 10. Here we can see for the first time the geographical divide in the city between high rent neighborhoods in the north and city center, and the blue-collar neighborhoods in the south. There are some exceptions in the north (the Tetuán district) but this north-south trend will also be seen in other socioeconomic indicators.

It is important to notice that the data is at a district level. In the case of Tetuán, there is a remarkable difference in income levels between some of its neighborhoods, with Azca and Castillejos having income levels way higher than the average income level of the district.

Real Estate data

The information for the average price of the square meter at a neighborhood level will be scraped from Idealista data website. Idealista is one of the biggest real estate agencies in Spain, and they have a very interesting data analysis and visualization available to the public.

The prices for both selling and renting real estate will be scraped for all neighborhoods in Madrid. Idealista has in place a system to avoid scraping its webpage, so all the different pages for each district for both rent and sell prices was manually stored in the computer and the web scraping was done from there.

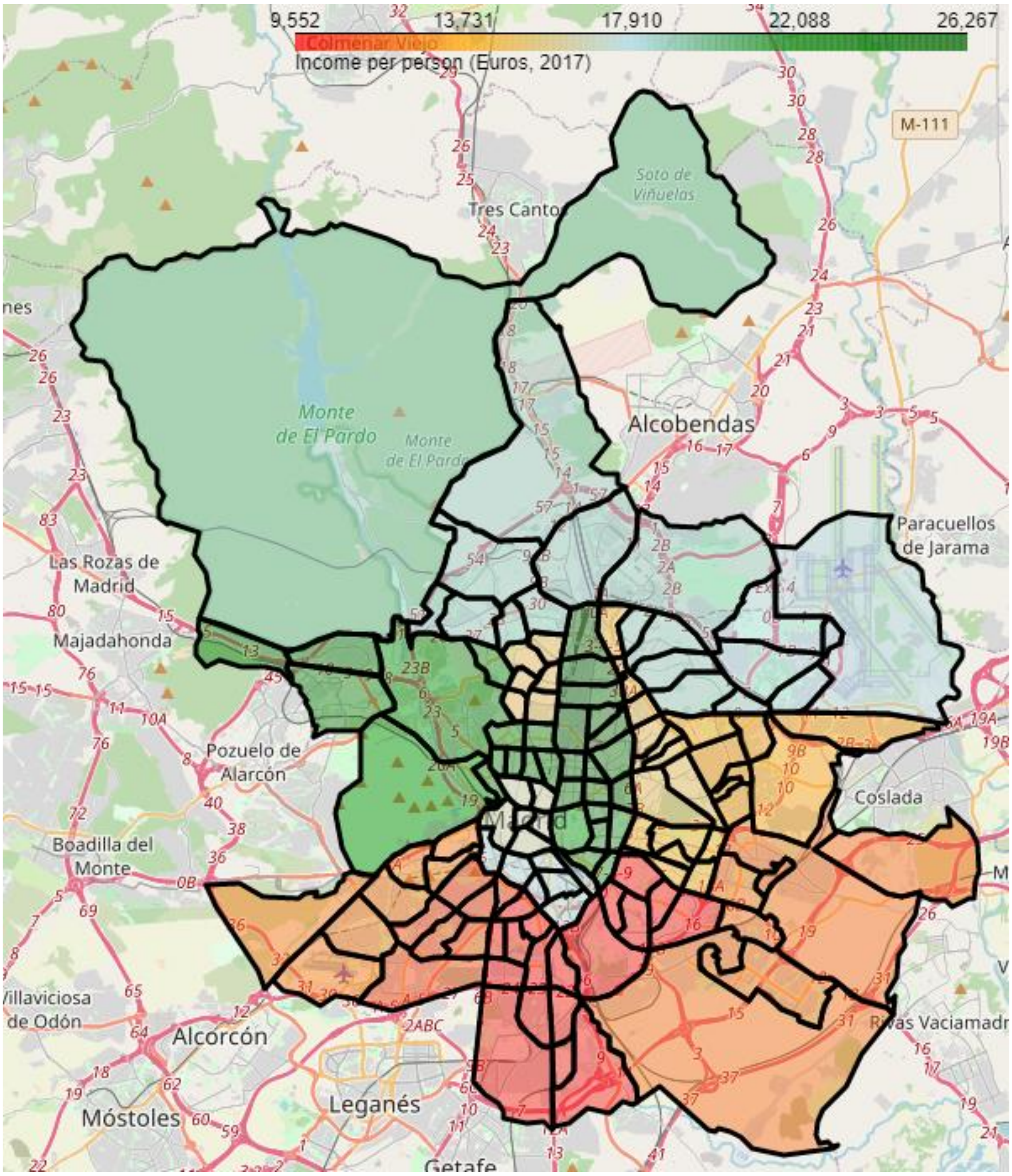


Figure 10 Color map of income distribution in Madrid. Red indicates lower incomes than green. An interactive version of the map can be found in the GitHub repository.

After loading the data into a dataframe, we have noticed that Idealista does not use the official neighborhood names and sometimes combines them. Some data needed their neighborhood names to be changed. But there were still some neighborhoods without data. In the Data Engineering section an explanation of the process to interpolate data for those neighborhoods will be explained. There, the exploratory analysis of the data will be described.

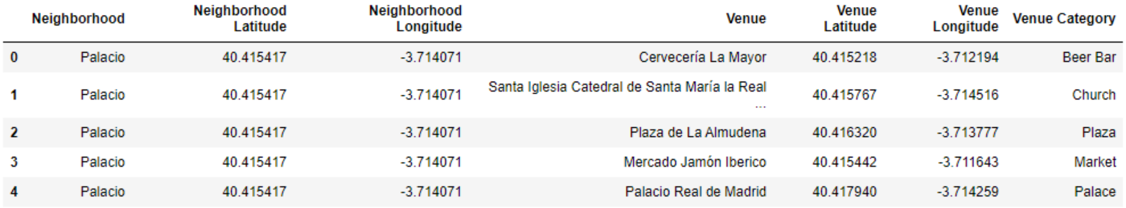
Crime data

Crime data is not readily available from single sources. In order to have a general idea of the crime levels in the city, only data from Madrid's Municipal Police will be taken. Excel files for each month of the year and district are available. The data corresponding to arrests will be taken from all files from the year 2019 (Jan 2019 to Dec 2019). Some operations will be performed in that dataset and the exploratory analysis for crime will be also performed in the next section.

Amenities data

The number of venues in each neighborhood and district will be taken from Foursquare using their API. This data will be used as a proxy of economic activity in each neighborhood.

The data obtained will be combined for all neighborhoods (using information on a district level where neighborhood level is not available, such as crime and income) and the data set will be fed to a K-means algorithm to segment the neighborhoods. An example of the information taken can be seen in Figure 11.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Palacio	40.415417	-3.714071	Cervecería La Mayor	40.415218	-3.712194	Beer Bar
1	Palacio	40.415417	-3.714071	Santa Iglesia Catedral de Santa María la Real ...	40.415767	-3.714516	Church
2	Palacio	40.415417	-3.714071	Plaza de La Almudena	40.416320	-3.713777	Plaza
3	Palacio	40.415417	-3.714071	Mercado Jamón Iberico	40.415442	-3.711643	Market
4	Palacio	40.415417	-3.714071	Palacio Real de Madrid	40.417940	-3.714259	Palace

Figure 11 Dataframe containing venues classified by type and neighborhood in which they are situated. This data will be used later to assess the economic activity of each neighborhood.

Data preparation

Feature engineering

Population data

We will combine the population and surface of the neighborhood to get the population density. The population density will be used in the clustering algorithm instead of those two features. The population density can be visualized in the map shown in Figure 12.

In general, the population density in the city center and some of the blue-collar neighborhoods in the south is higher than in the rest of the city, especially the big neighborhoods in the outskirts, where there is some sub-urban development but not the high density development one should expect in downtown Madrid.

Real Estate data

In section 4.4 we have seen that there are some neighborhoods without Real Estate data. This is because some data in the real estate database is shared between different neighborhoods due to proximity, or because the name of the real estate area does not correspond with a real neighborhood.

In this section, we will assign the median price (for both rent and sell) of the district to which they belong for all those neighborhoods without data. We will use the median instead of the mean because it is more robust to the influence of distribution skewness and outliers. A table with the median prices per square meter in each of Madrid's districts can be seen in Figure 13. Additionally, these same values in form of a map can be seen in Figure 14 and Figure 15.

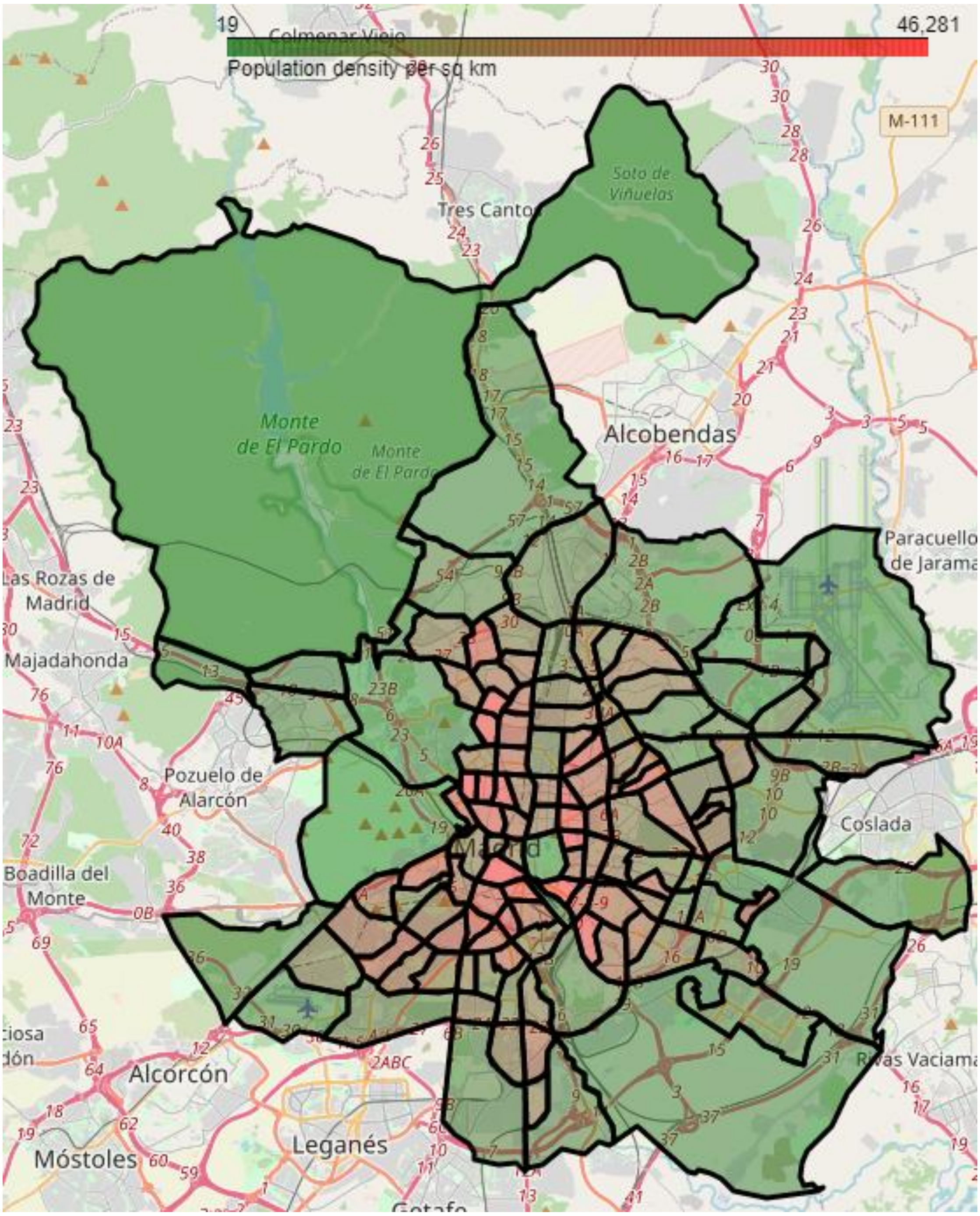
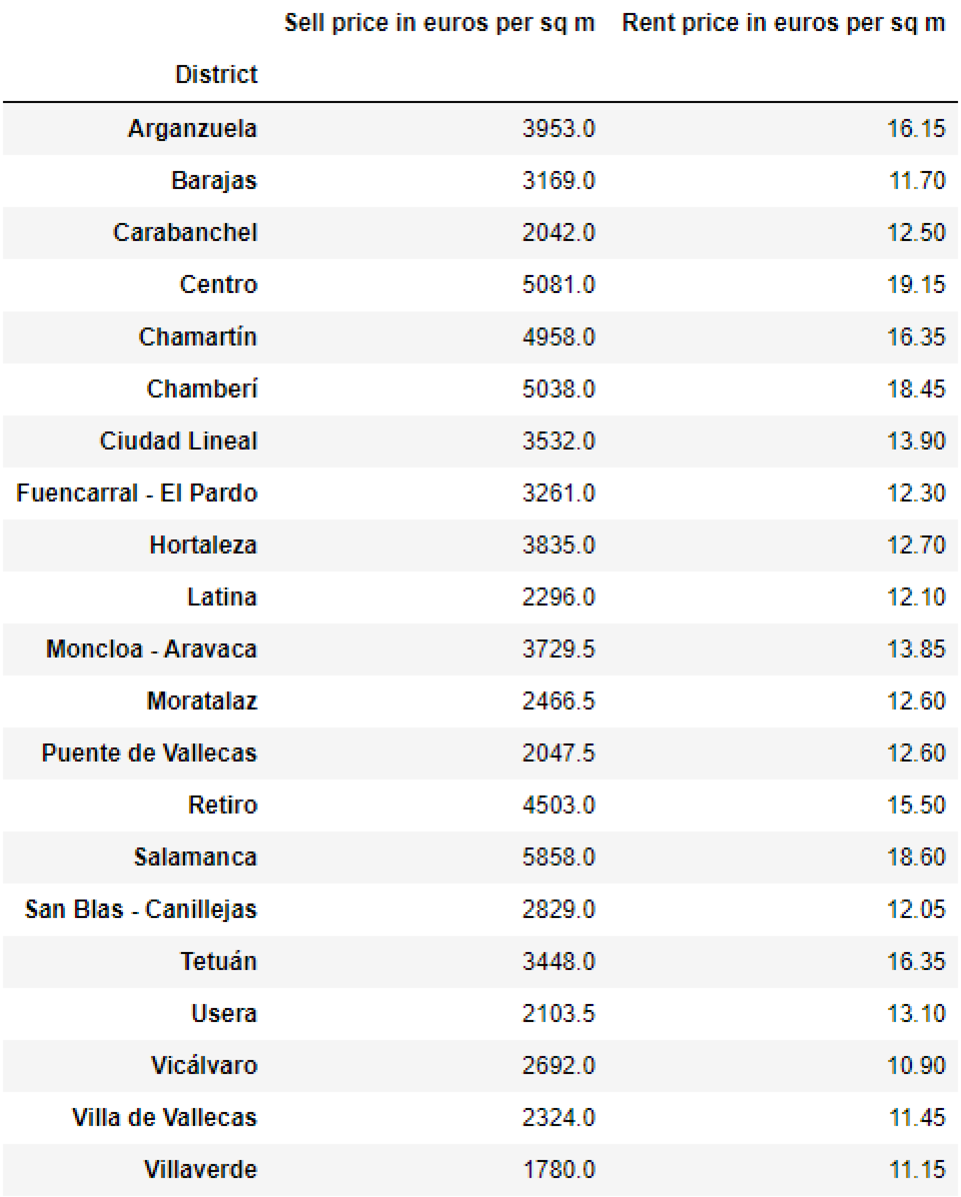


Figure 12 Population density in inhabitants per square kilometer in the different neighborhoods of Madrid. Green indicates low population density, whereas red indicates high population density. An interactive version of the map can be found in the GitHub repository.



	Sell price in euros per sq m	Rent price in euros per sq m
District		
Arganzuela	3953.0	16.15
Barajas	3169.0	11.70
Carabanchel	2042.0	12.50
Centro	5081.0	19.15
Chamartín	4958.0	16.35
Chamberí	5038.0	18.45
Ciudad Lineal	3532.0	13.90
Fuencarral - El Pardo	3261.0	12.30
Hortaleza	3835.0	12.70
Latina	2296.0	12.10
Moncloa - Aravaca	3729.5	13.85
Moratalaz	2466.5	12.60
Puente de Vallecas	2047.5	12.60
Retiro	4503.0	15.50
Salamanca	5858.0	18.60
San Blas - Canillejas	2829.0	12.05
Tetuán	3448.0	16.35
Usera	2103.5	13.10
Vicálvaro	2692.0	10.90
Villa de Vallecas	2324.0	11.45
Villaverde	1780.0	11.15

Figure 13 Median prices of the square meter of real estate property in Madrid for each of the 21 districts of the city. This data will be used to fill the gaps in neighborhood's real estate prices.

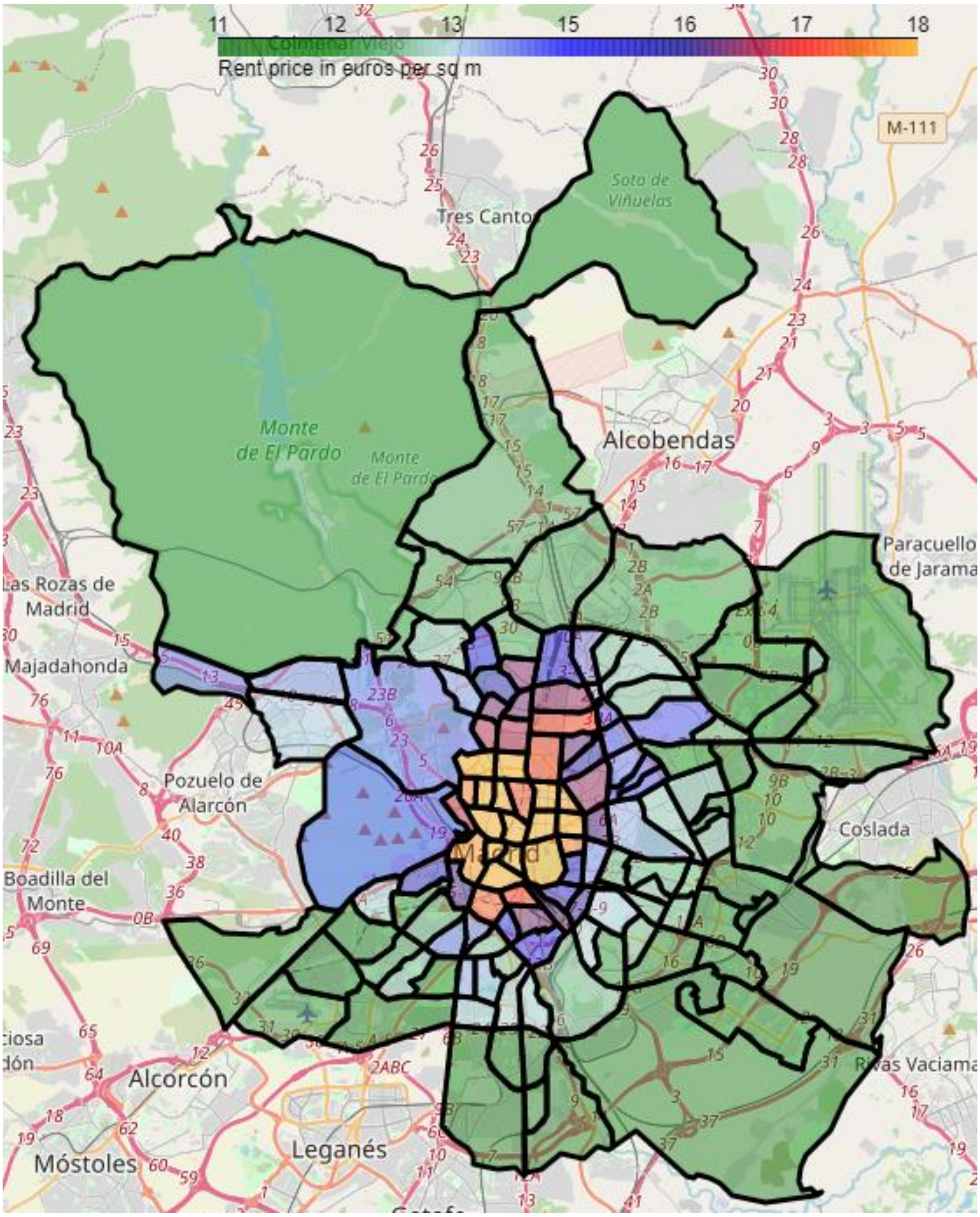


Figure 14 Color map of rent prices in euros per square meter for all the neighborhoods of Madrid. Green indicates a low number, transitioning to blue for medium numbers, red for high numbers and orange for very high prices. An interactive map can be found in the GitHub repository.

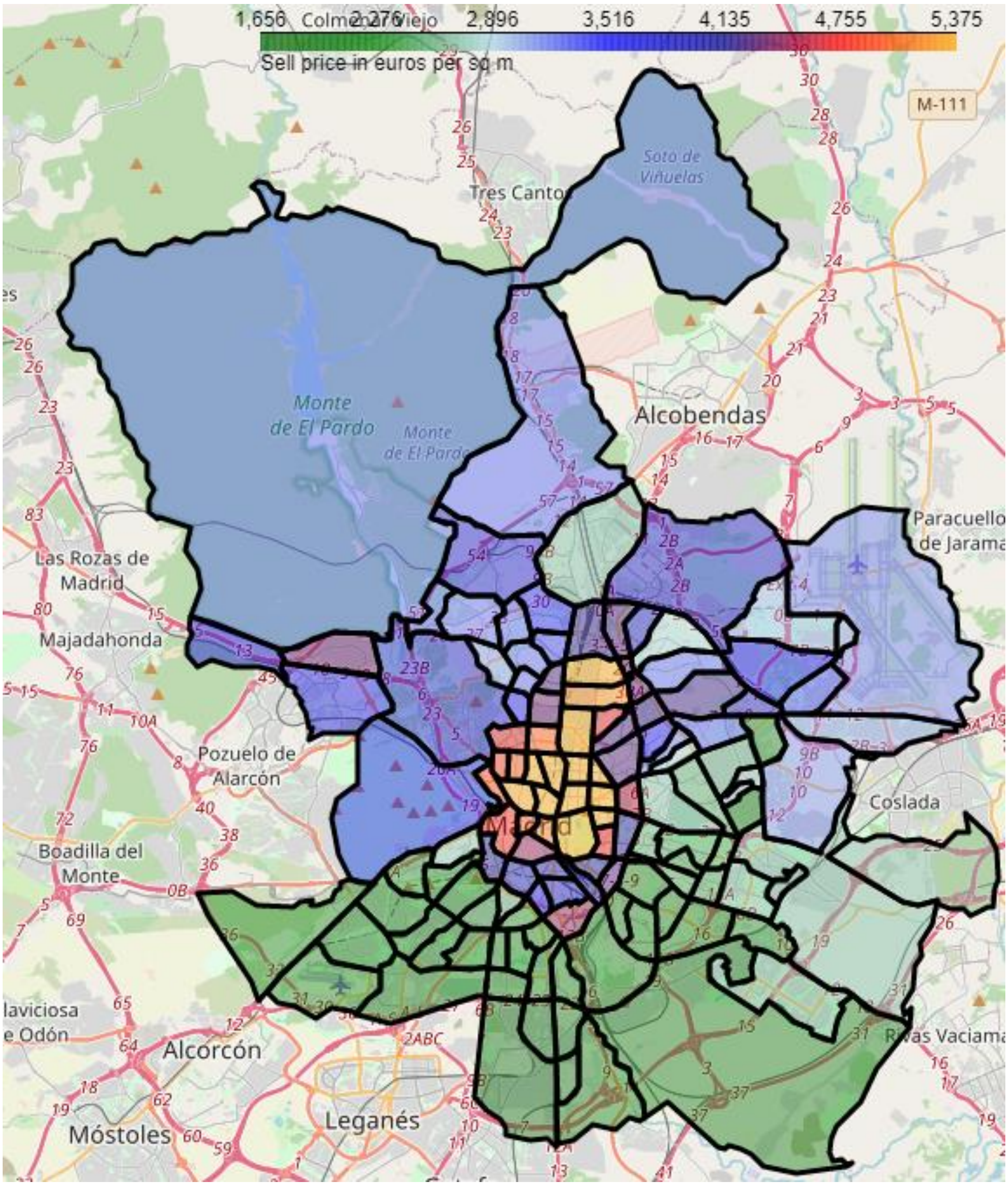


Figure 15 Color map of sell prices in euros per square meter for all the neighborhoods of Madrid. Green indicates a low number, transitioning to blue for medium numbers, red for high numbers and orange for very high prices. An interactive map can be found in the GitHub repository.

In both maps a north-south divide and a center-periphery divide can be seen. In general, real estate prices are higher in the northern part of the city than in the south; and they are also higher in the central area inside of the city ring M-30 than outside that road. Some notable high prices can be seen in the Salamanca district, which houses luxury shops and the so-called golden mile of the city.

Crime data

Data from the arrests per district in the year 2019 will be summed and then divided by the population of the district to obtain a ratio of arrests per capita. The results will be visualized in the map shown in Figure 16.

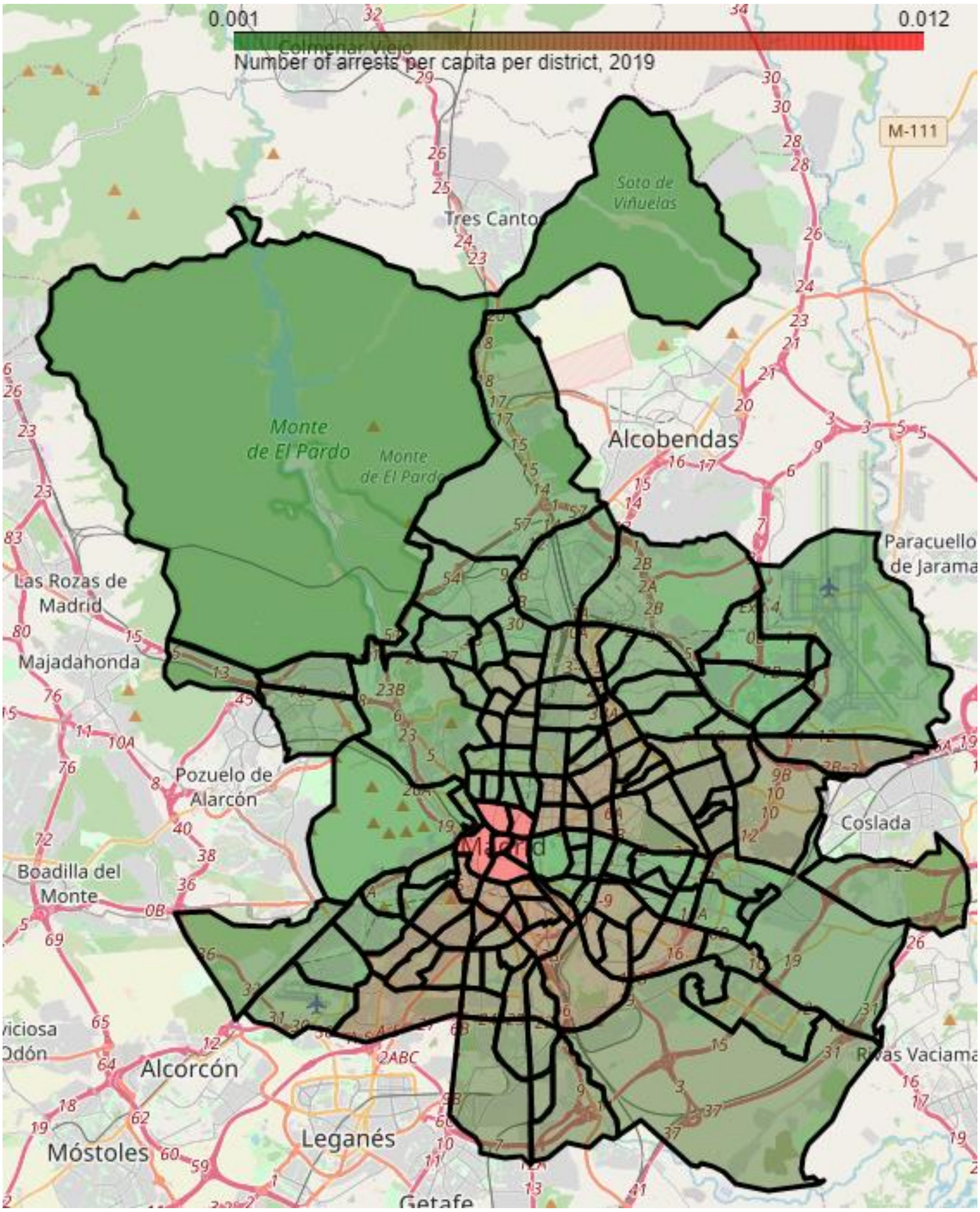
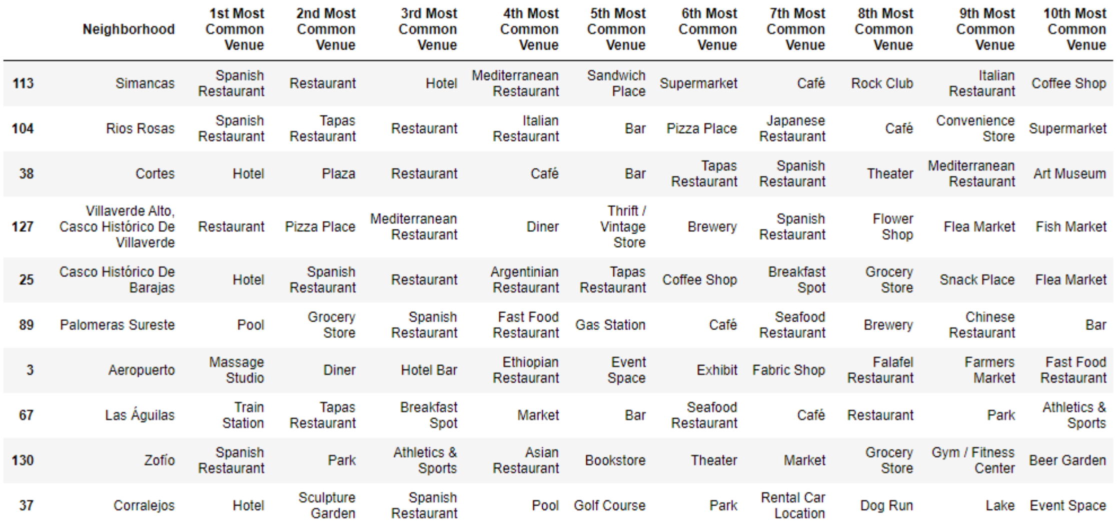


Figure 16 Number of arrests per capita for each district of Madrid. Green indicates a low number, whereas red indicates a high number. An interactive version of the map can be found in the GitHub repository.

The districts with higher crime levels in the city are the central district, where there is a very large influx of people (inhabitants of the city and also tourists), the Salamanca district (same as the Central district) and some districts in the south part of the city. Spain is a very safe country compared with even other European countries, scoring as the 34th safest country in the world according to Numbeo Crime Index by Country for the year 2020.

Amenities data

We will extract the data from the Foursquare data frame in order to collect the 10 most common venues in each neighborhood. Then that data will be examined to search for indicators of economic activity to add to the main dataframe. The venues will be grouped by neighborhood, so we have the percentage of venues of a certain type for each neighborhood. Then, the ten most common venues will be shown in the dataframe. Figure 17 shows a sample of some neighborhoods and their most common venues.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
113	Simancas	Spanish Restaurant	Restaurant	Hotel	Mediterranean Restaurant	Sandwich Place	Supermarket	Café	Rock Club	Italian Restaurant	Coffee Shop
104	Rios Rosas	Spanish Restaurant	Tapas Restaurant	Restaurant	Italian Restaurant	Bar	Pizza Place	Japanese Restaurant	Café	Convenience Store	Supermarket
38	Cortes	Hotel	Plaza	Restaurant	Café	Bar	Tapas Restaurant	Spanish Restaurant	Theater	Mediterranean Restaurant	Art Museum
127	Villaverde Alto, Casco Histórico De Villaverde	Restaurant	Pizza Place	Mediterranean Restaurant	Diner	Thrift / Vintage Store	Brewery	Spanish Restaurant	Flower Shop	Flea Market	Fish Market
25	Casco Histórico De Barajas	Hotel	Spanish Restaurant	Restaurant	Argentinian Restaurant	Tapas Restaurant	Coffee Shop	Breakfast Spot	Grocery Store	Snack Place	Flea Market
89	Palomeras Sureste	Pool	Grocery Store	Spanish Restaurant	Fast Food Restaurant	Gas Station	Café	Seafood Restaurant	Brewery	Chinese Restaurant	Bar
3	Aeropuerto	Massage Studio	Diner	Hotel Bar	Ethiopian Restaurant	Event Space	Exhibit	Fabric Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant
67	Las Águilas	Train Station	Tapas Restaurant	Breakfast Spot	Market	Bar	Seafood Restaurant	Café	Restaurant	Park	Athletics & Sports
130	Zofío	Spanish Restaurant	Park	Athletics & Sports	Asian Restaurant	Bookstore	Theater	Market	Grocery Store	Gym / Fitness Center	Beer Garden
37	Corralejos	Hotel	Sculpture Garden	Spanish Restaurant	Pool	Golf Course	Park	Rental Car Location	Dog Run	Lake	Event Space

Figure 17 Sample of ten Madrid's neighborhoods and their ten most common venues.

It seems that a lot of venues are different kinds of restaurants. There is evidence that restaurants can be used as a proxy of socioeconomic activities in a neighborhood in absence of other data (the study can be seen here "Predicting neighborhoods' socioeconomic attributes using restaurant data" by Lei Dong, Carlo Ratti, and Siqi Zheng. PNAS July 30, 2019 116 (31) 15447-15452. All the venue names corresponding to restaurants will be added together and a new feature counting the number of restaurants in each neighborhood will be created. This feature will be added to the main dataframe.

Modelling

With all the information we have we will perform a segmentation of the neighborhoods using a K-means clustering algorithm. The dataframe fed to the segmentation algorithm will contain the following features for each neighborhood:

- Real Estate selling prices per square meter.
- Real Estate renting prices per square meter.
- Mean income per person.
- Population density in inhabitants per square kilometer.
- Number of arrests per capita.
- Number of restaurants in the neighborhood.

This data will be normalized employing the MinMaxScaler preprocessing tool. The clustering was also tested by using the StandardScaler tool, but the results were equivalent.

We need to determine the optimal number of clusters for the algorithm. To do this, the elbow method will be used. For each k value, k-means will be initialized, and the inertia attribute will be employed to identify the sum of squared distances of samples to the nearest cluster center. The results of the test are shown in Figure 18.

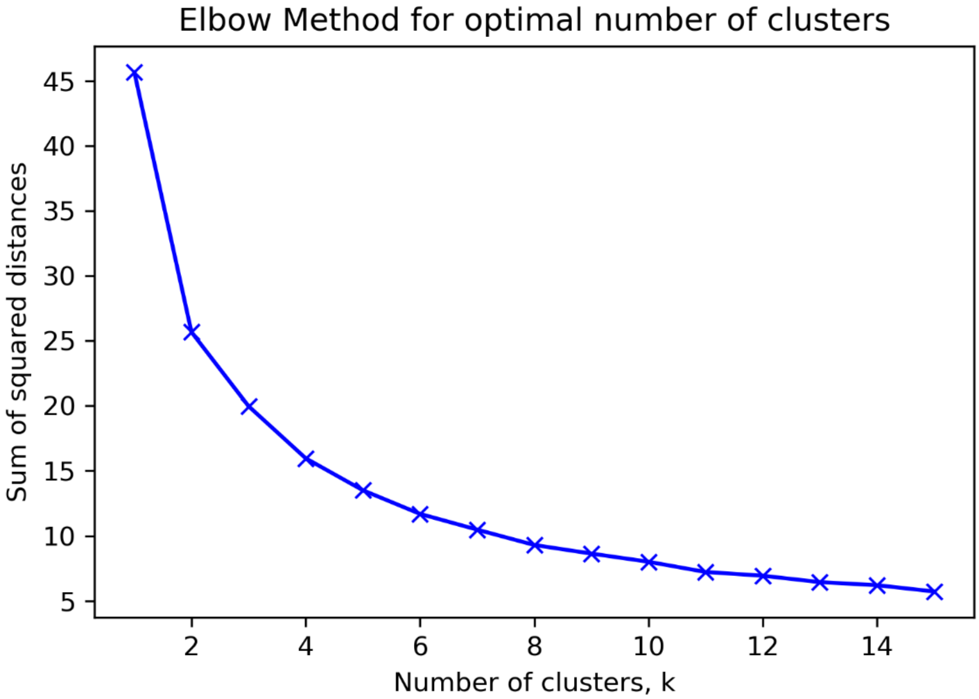


Figure 18 Results of Elbow method to find the optimal number of clusters in the segmentation algorithm.

From Figure 18 the optimal number of clusters for this dataset is 3. The segmentation algorithm was run with that number of clusters and the k-means++ method. The rest of options were set to the default of the method. The results of the clustering can be seen color coded in the map in Figure 19. These colors do not have any particular meaning, and they only are used to separate the different labels that resulted from the segmentation algorithm.

Evaluation

Let us look at the different features employed in the algorithm in form of box plots to see how they are segmented. Figure 19 shows the neighborhoods of Madrid segmented using a three-color code. Figures 20-25 show how the different features are distributed throughout the labels.

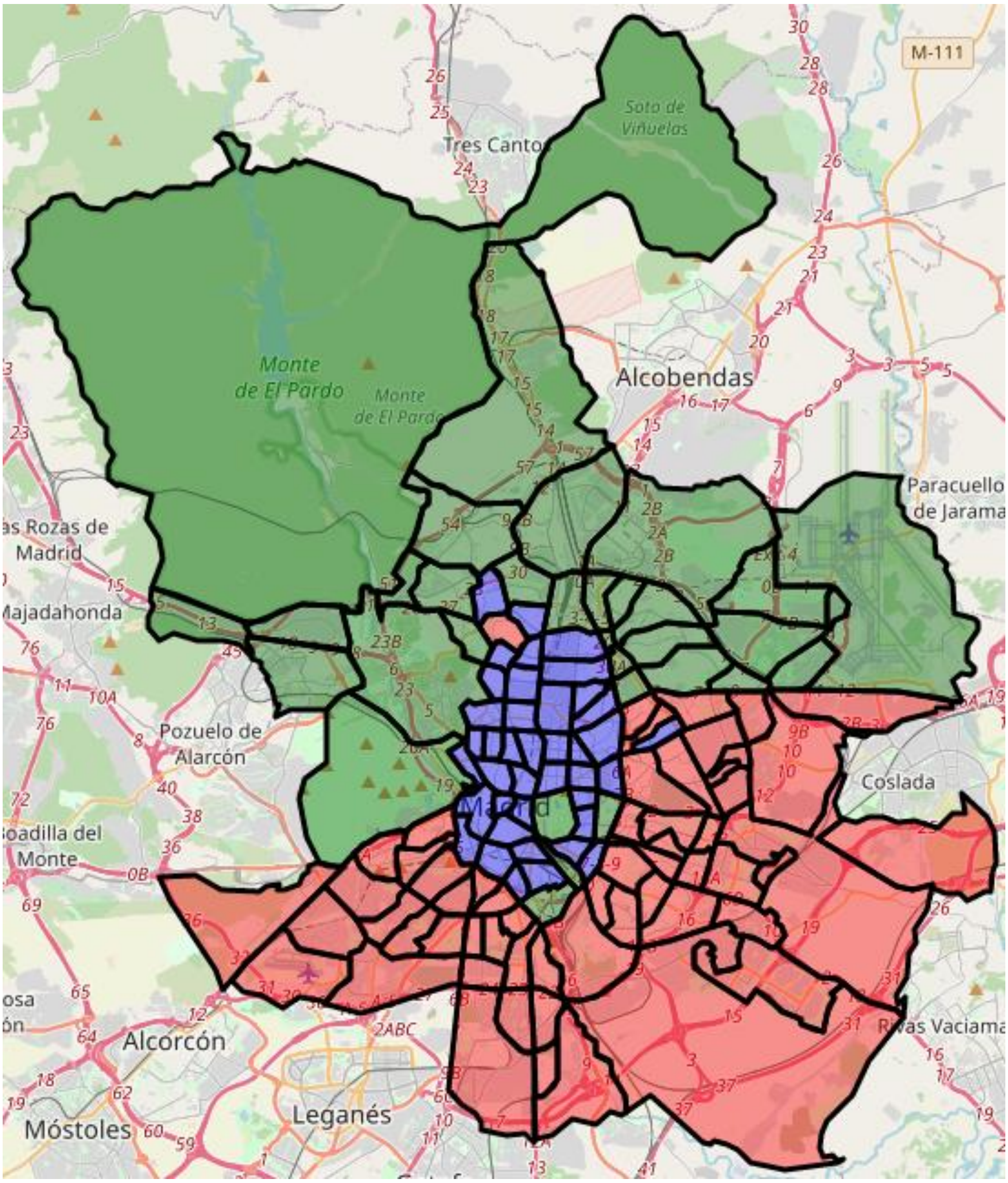


Figure 19 Neighborhoods of Madrid segmented into three different groups using a K Means segmentation algorithm. Colors do not have any meaning except to indicate the group to which the neighborhoods belong. Blue: first cluster; Red: second cluster; Green: third cluster. An interactive map version can be found in the GitHub repository.

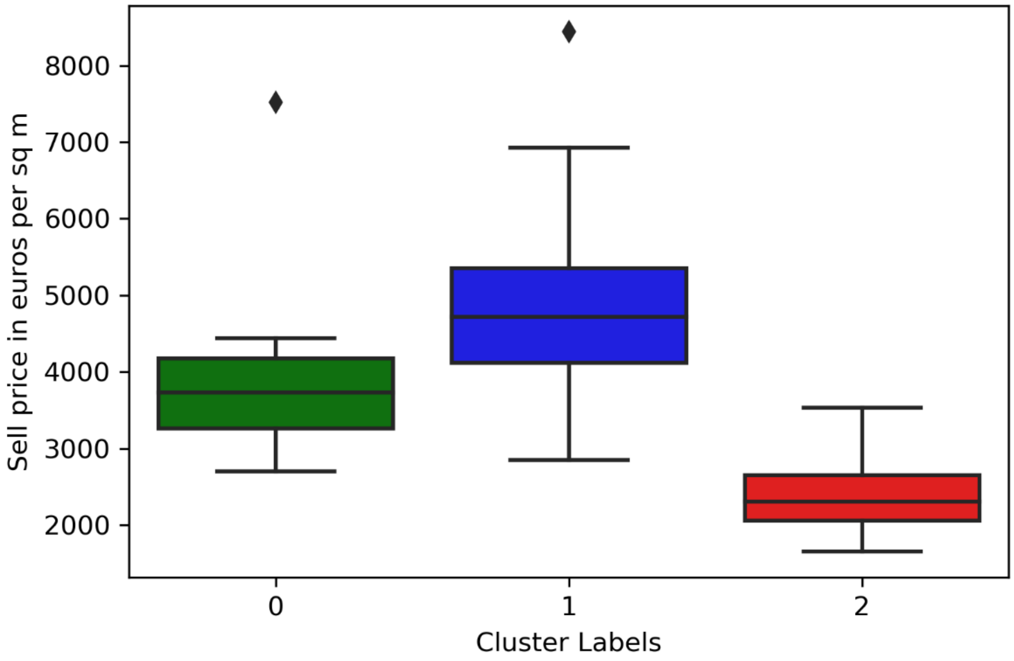


Figure 20 Box plot of the real estate selling prices per square meter as a function of the cluster label. The colors of the boxes are the same as the clusters in the map.

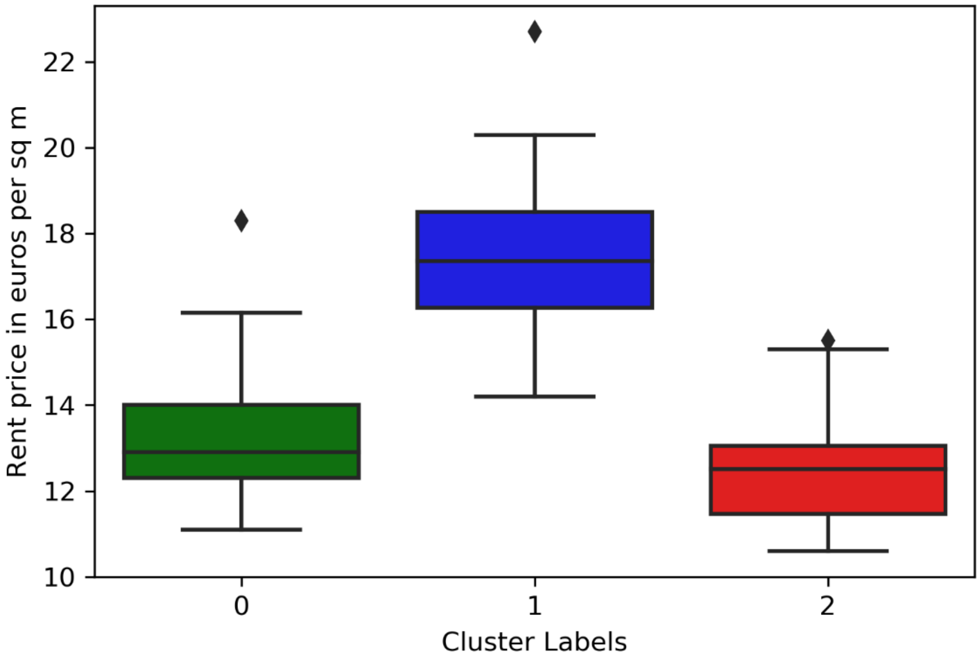


Figure 21 Box plot of the real estate renting prices per square meter as a function of the cluster label. The colors of the boxes are the same as the clusters in the map.

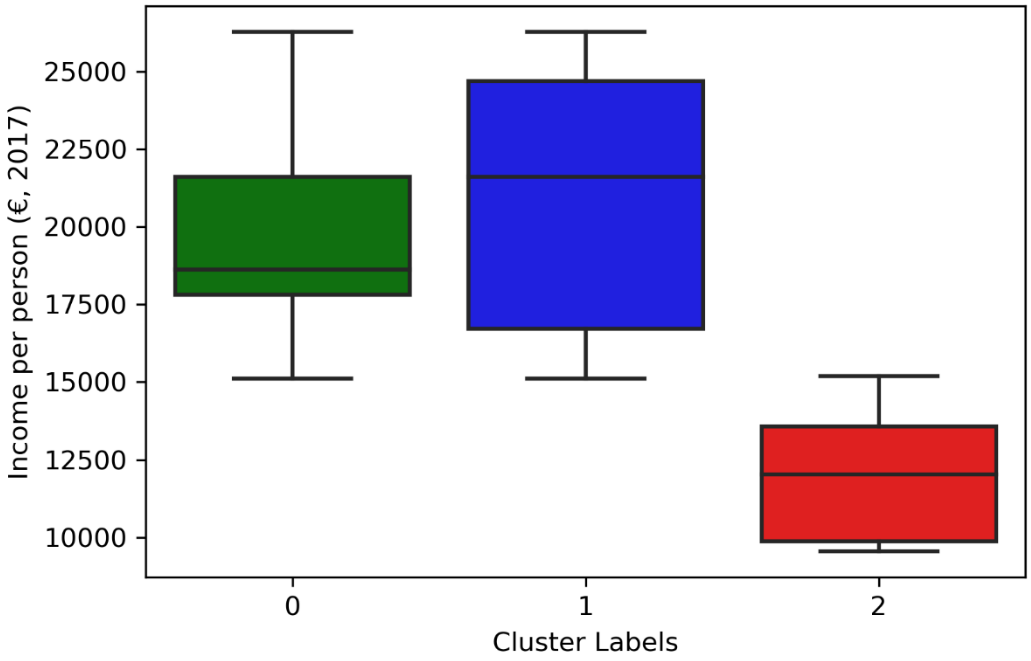


Figure 22 Box plot of the mean income per person as a function of the cluster label. The colors of the boxes are the same as the clusters in the map.

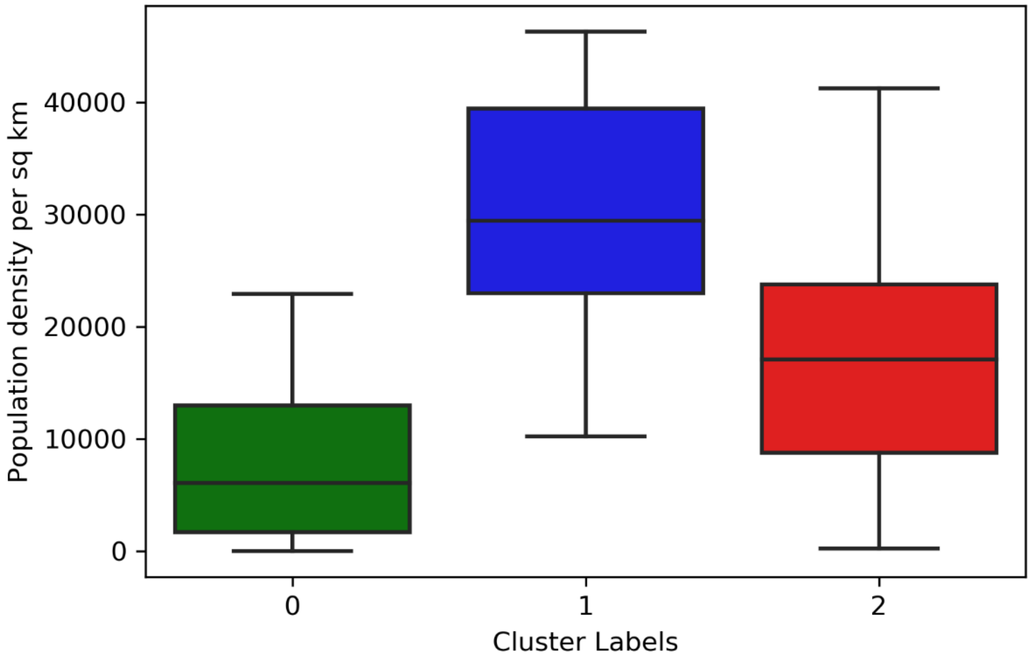


Figure 23 Box plot of the population density per square km as a function of the cluster label. The colors of the boxes are the same as the clusters in the map.

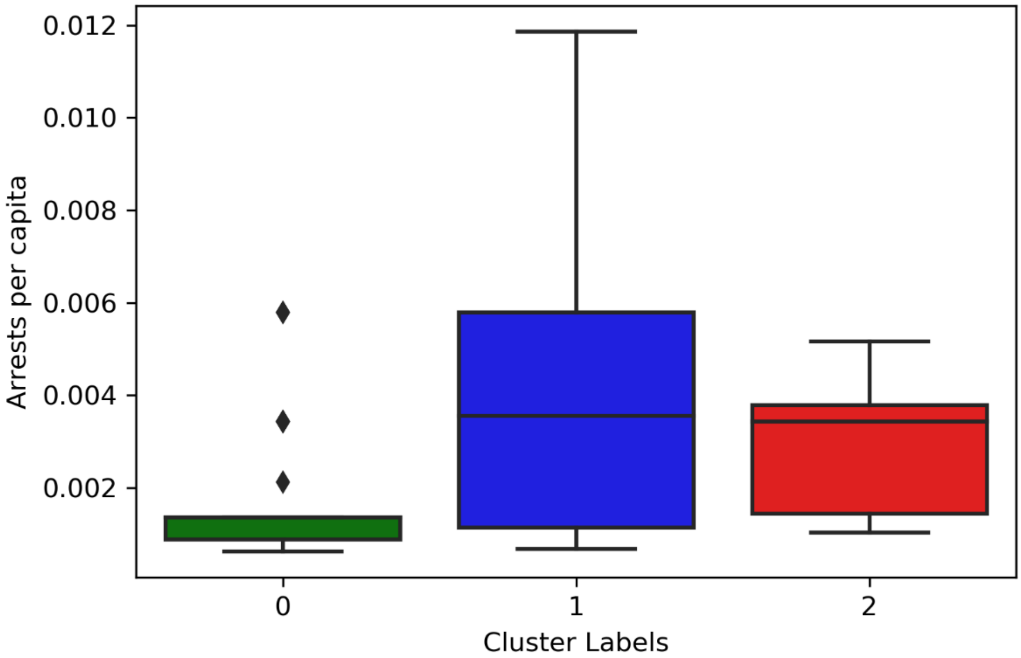


Figure 24 Box plot of the arrests per capita as a function of the cluster label. The colors of the boxes are the same as the clusters in the map.

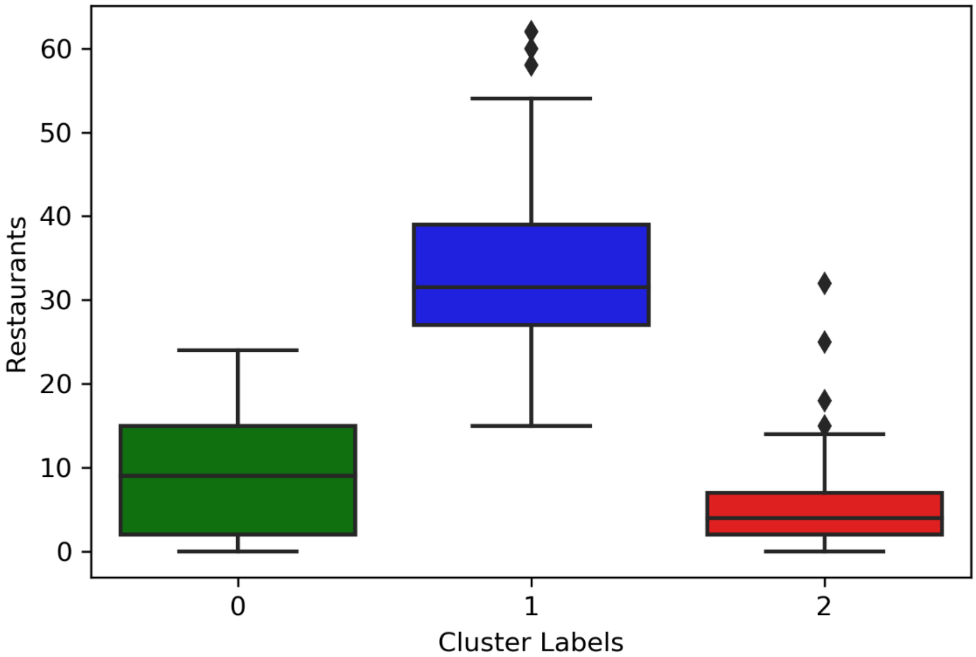


Figure 25 Box plot of number of restaurants per neighborhood as a function of the cluster label. The colors of the boxes are the same as the clusters in the map.

We can obtain some insights from this analysis. The classification of the neighborhoods of Madrid would be as follows:

Neighborhoods in the first cluster (green in the map):

- Mainly the northern neighborhoods outside of the city center.
- Very low population density.
- High income.
- Moderate real estate prices for selling property.
- Low real estate prices for renting property.
- Very low crime levels.
- Low number of restaurants.

These features indicate neighborhoods located usually in the northern outskirts of Madrid, with high prices for selling real estate (but not for renting real estate) and high rent levels. It seems that this segment refers to the more upscale neighborhoods in the north of Madrid, with suburban areas and neighborhoods in Madrid with ample green areas (as the Retiro park).

Neighborhoods in the second cluster (blue in the map):

- Corresponding roughly to the city centre, inside the M-30 orbital motorway, the innermost ring road of the city.
- Highest population density of the three clusters.
- Highest income of the clusters.
- Highest real estate property value, especially for renting property.
- Highest numbers of arrest per capita.
- Plenty of restaurants.

This cluster tends to be the city centre of Madrid, with tourist activity (hotels, exhibits...) and more rent than the southern part. Crime is higher because there is also more population and more tourist activity.

Neighborhoods in the third cluster (red in the map):

- Southern part of the city, even though there is a neighborhood in the north.
- Moderate population density.
- Lowest prices for buying/selling and renting real estate of the three clusters.
- Moderate number of arrests per capita.
- Very low number of restaurants.

This cluster shows mostly the more blue-collar part of the city since the income is lower. Mainly residential neighborhoods situated in the south, not so touristic and with higher crime levels than the outskirts in the north.

Conclusion

The main goal of this project was to classify the neighborhoods of Madrid based on socioeconomic and business diversity in order to give information about living conditions in Madrid to new employees of our client to narrow down their housing search after they are hired to work there.

To do that, information about population density, income levels, crime levels, real estate information (renting and buying prices), and amenities in the neighborhood were collected from several official government bodies, such as the Spanish national statistics institute and the city council of Madrid and from business data from Foursquare.

The data was converted to a data frame which was normalized and fed to a K Means clustering algorithm that segmented the neighborhoods in the optimal number of clusters using the elbow method, which was three clusters.

Finally, the clusters were examined to search common features for the neighborhoods, as shown in the previous section. Final selection of living arrangements will be performed by our client's new employees based on specific features of the clusters.