BATTLE OF THE NEIGHBORHOODS

IBM Data Science Capstone Project

Juan Prieto

Business understanding

- Madrid is a major business hub in southwestern Europe and the Iberian Peninsula.
- Cost of living in Madrid is higher than in other Spanish cities, but the socioeconomic differences between different neighborhoods of the city can be used to find affordable places to live.
- A company wants to analyze socioeconomic data from Madrid to offer insights to prospective employees to help them find accommodation in the city.

Picture from https://www.smartcitylab.com/blog/es/gobierno-finanzas/que-pueden-aprender-otras-ciudades-de-los-errores-de-madrid-central/

Business understanding

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)

K Means clustering

The neighborhoods will be segmented and classified according to those features

Business understanding

Our audience:

New employees from the company wishing to move to Madrid and know a Little bit about the city before moving in.

Company's management expects to understand the rationale behind the recommendations made.

The general public could be also benefitted from this information.

Business understanding

Success criteria:

The project will be considered successful if a 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 – Geographical data

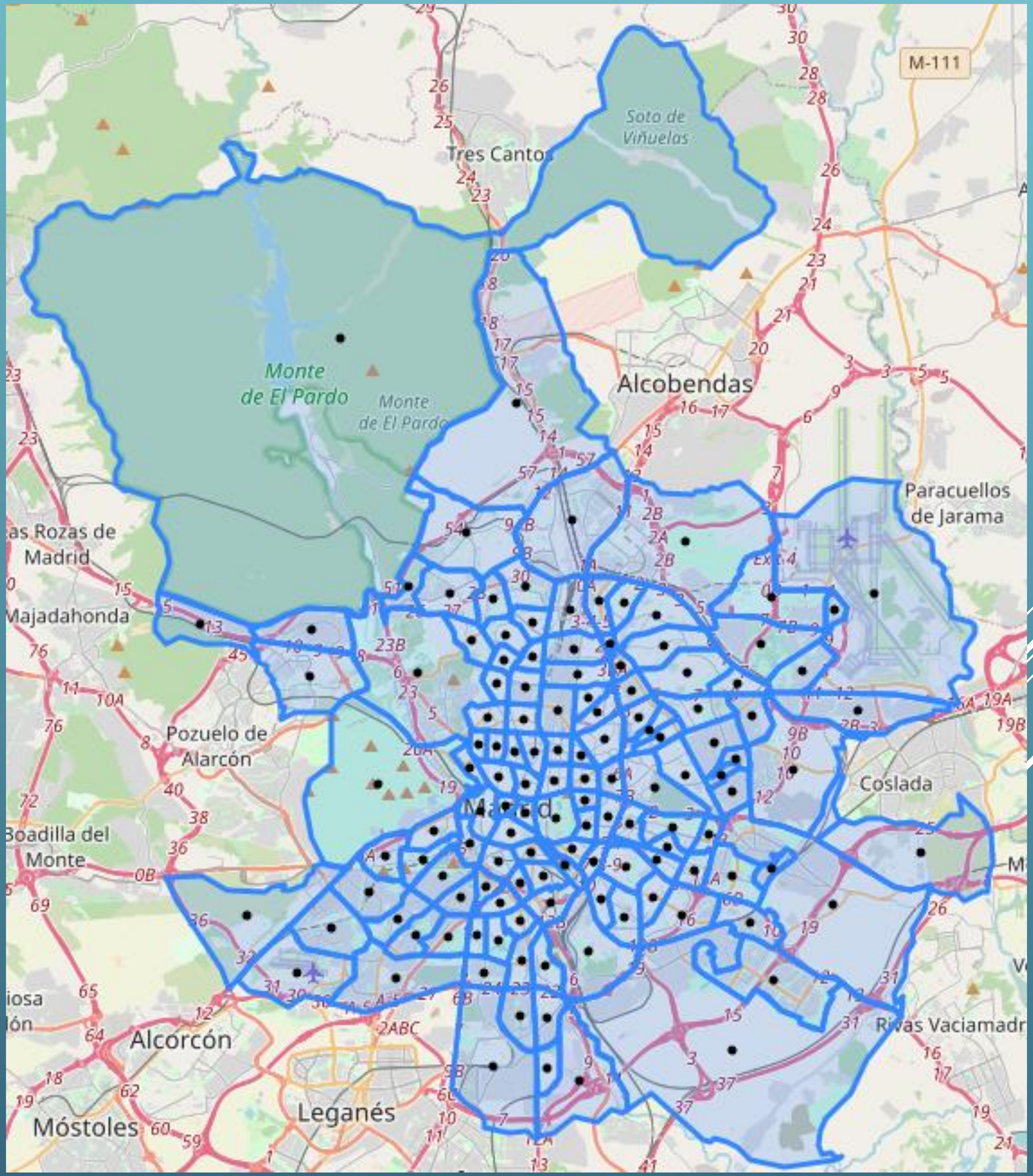
- Geopandas dataframe containing polygon shapes and total Surface of the 151 neighborhoods of the city of Madrid.
- Data taken from the City Council of Madrid

	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

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

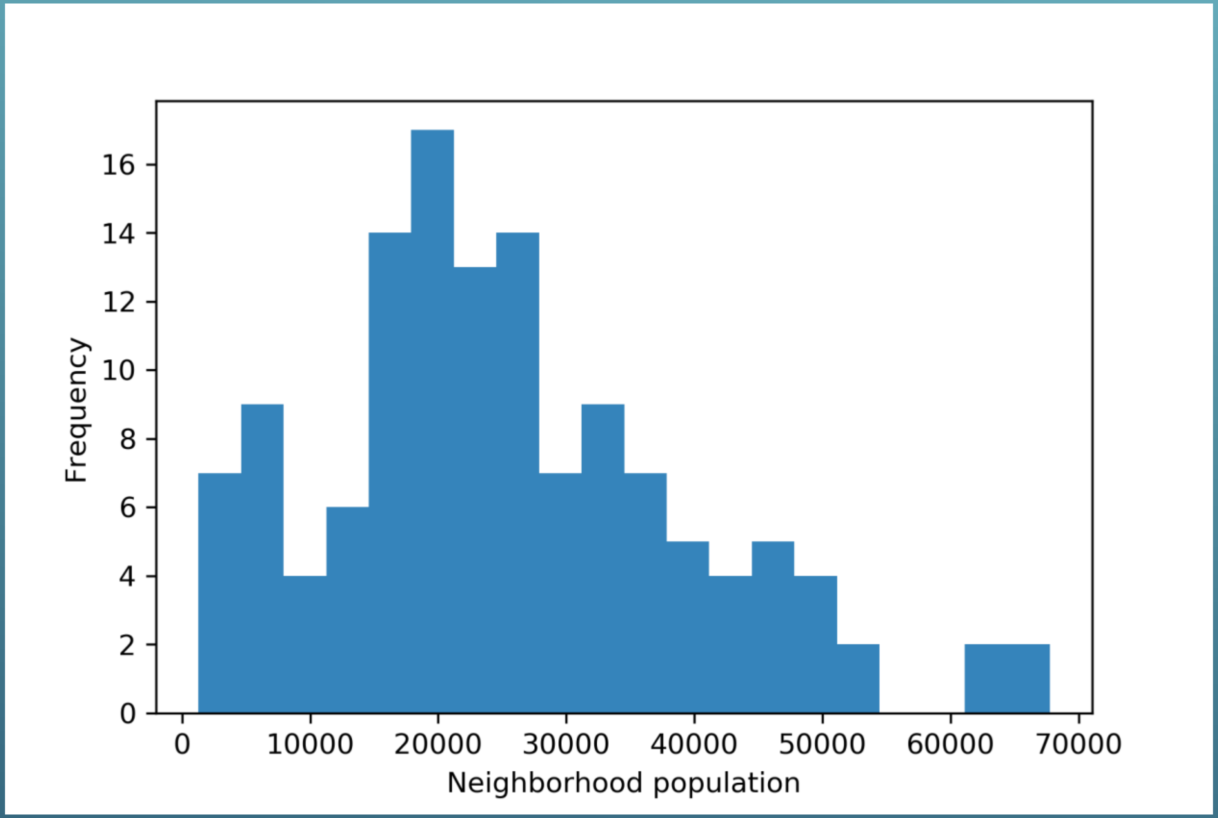
Data understanding – Geographical data

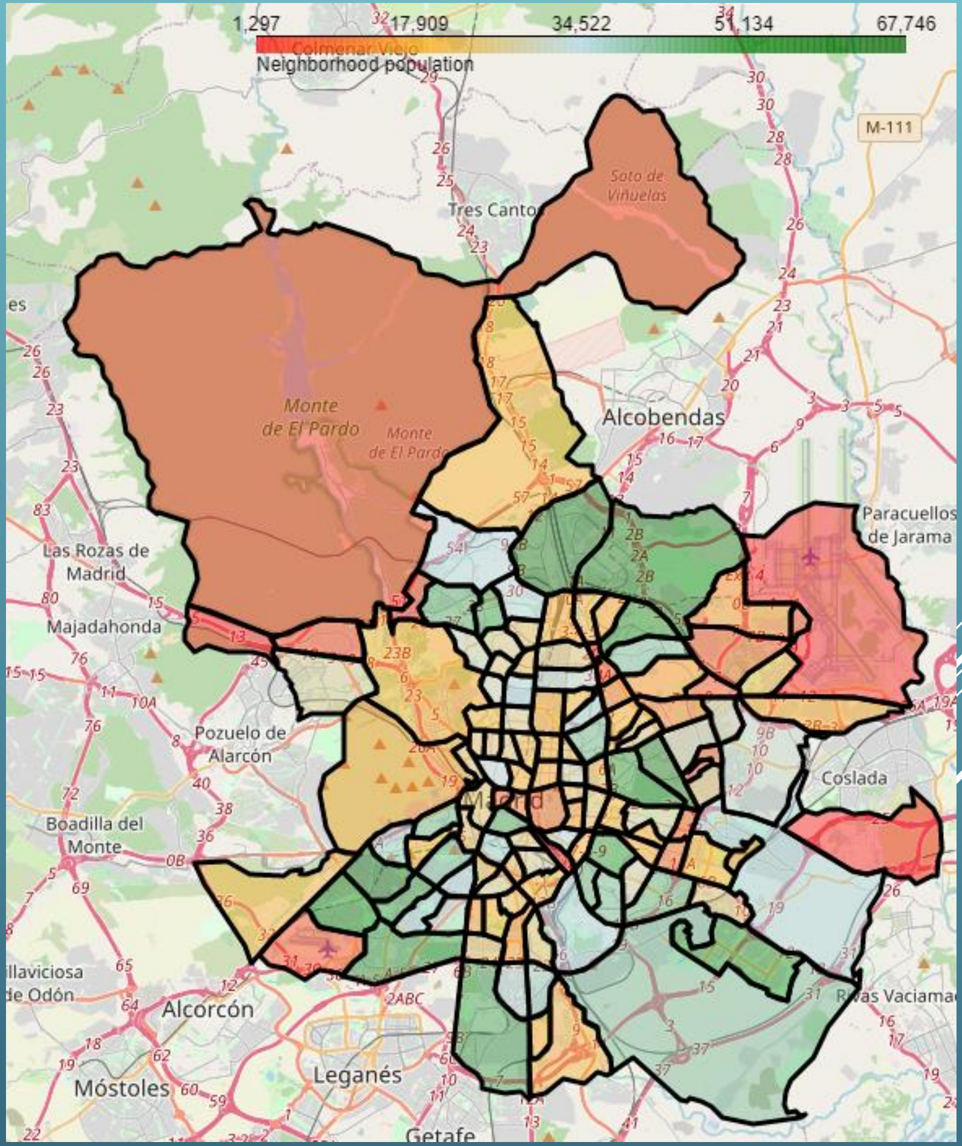
- Geopandas dataframe containing polygon shapes and total Surface of the 151 neighborhoods of the city of Madrid.
- Data taken from the City Council of Madrid



Madrid's neighborhoods (blue) and neighborhood centroids (black dots)

Data understanding – Population data

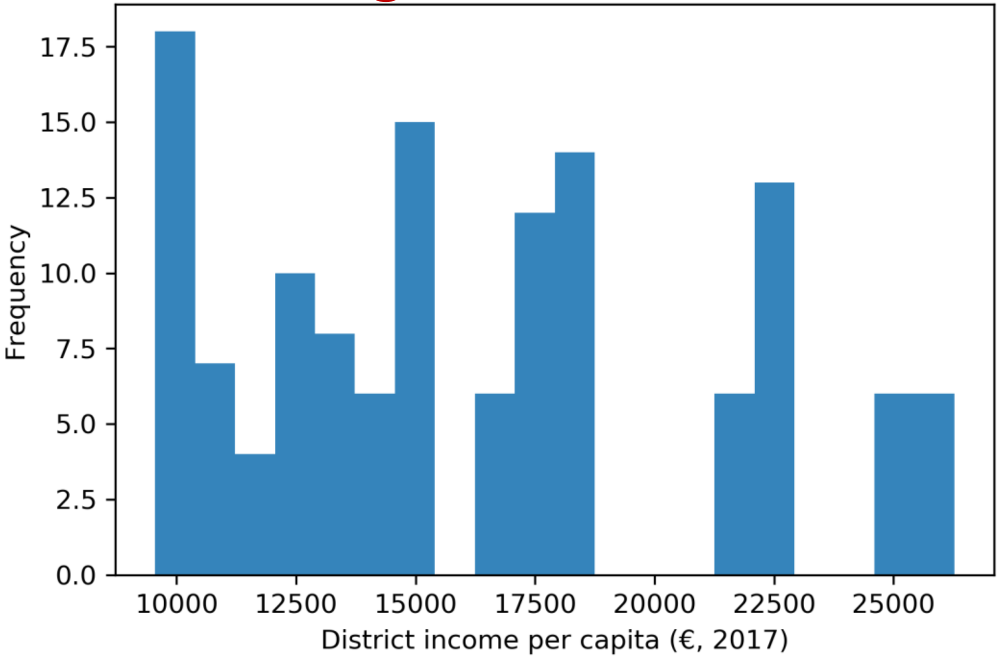


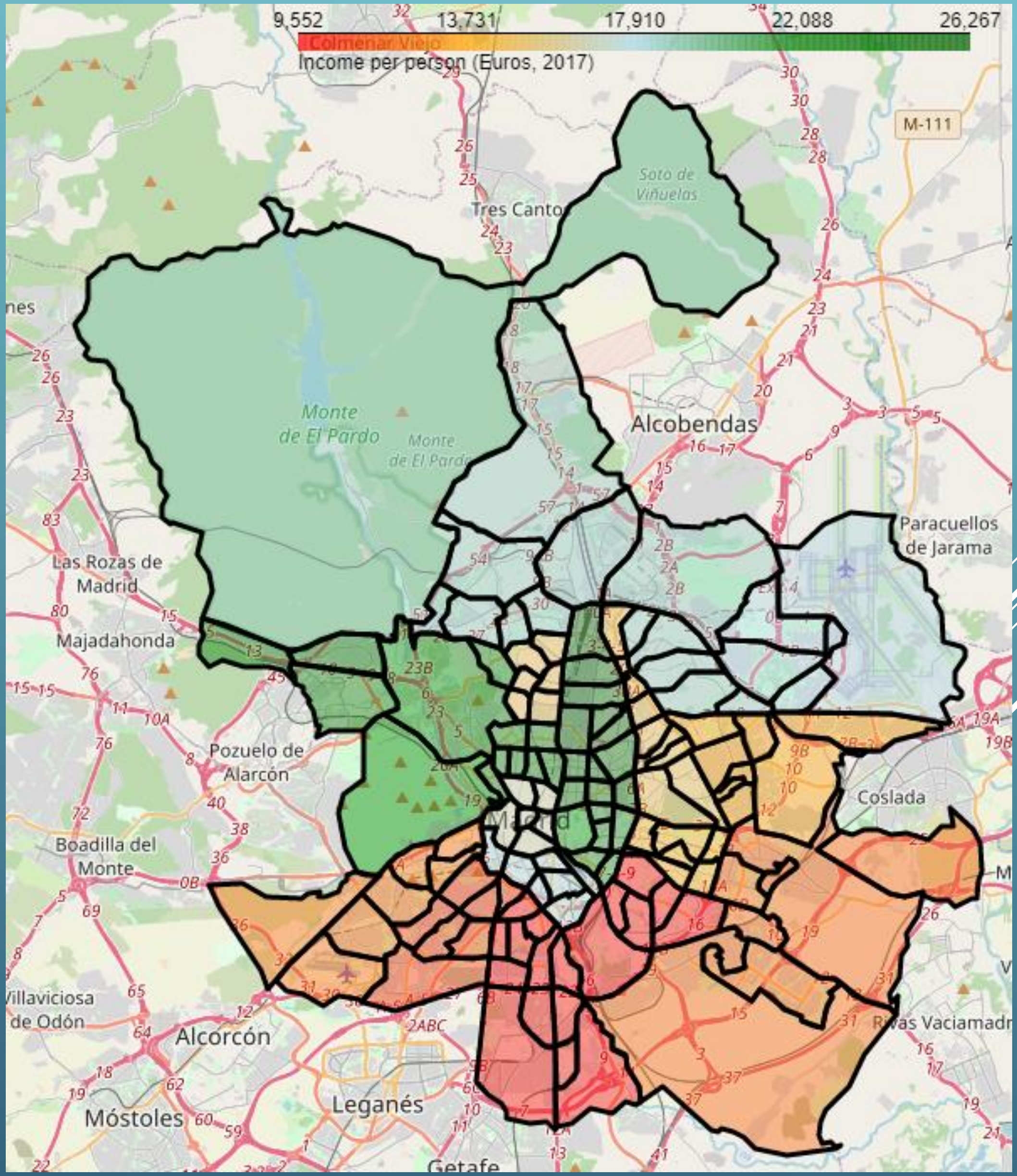


Data obtained from City Council of Madrid

Data understanding – Income data

Warning: Data at district level





Data taken from National Statistics Institute (INE)

Data understanding – Crime, Real Estate and amenities

Crime data taken from arrest reports per district of Madrid's Municipal Police

Real Estate data taken from Idealista, a Spanish realtor.

Amenities data taken from Foursquare using their API to obtain venues near the neighborhood centroids.

Data will be explained in the next section: feature engineering.

	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

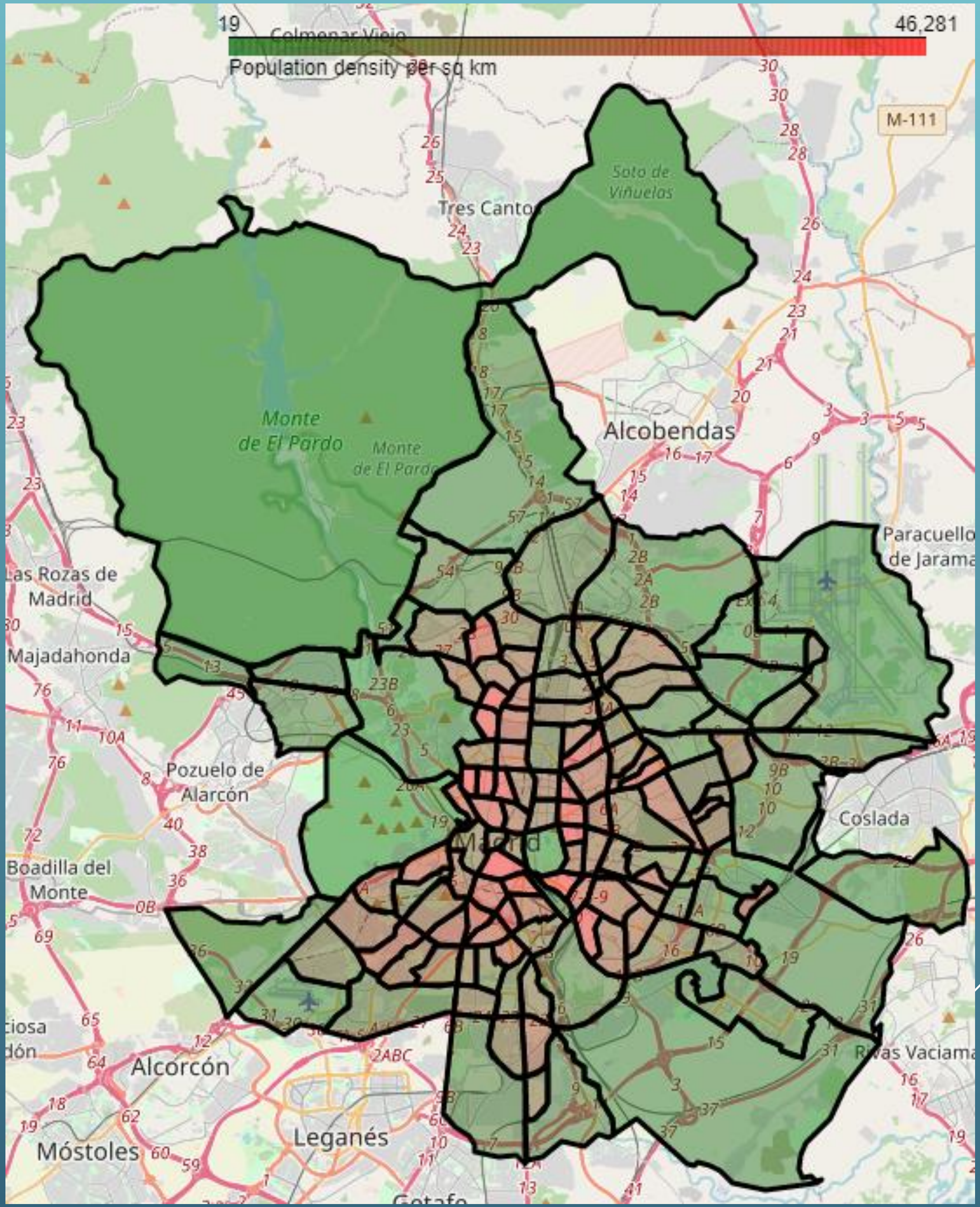
Data preparation – Feature Engineering

Population data

+

Surface data

Population density

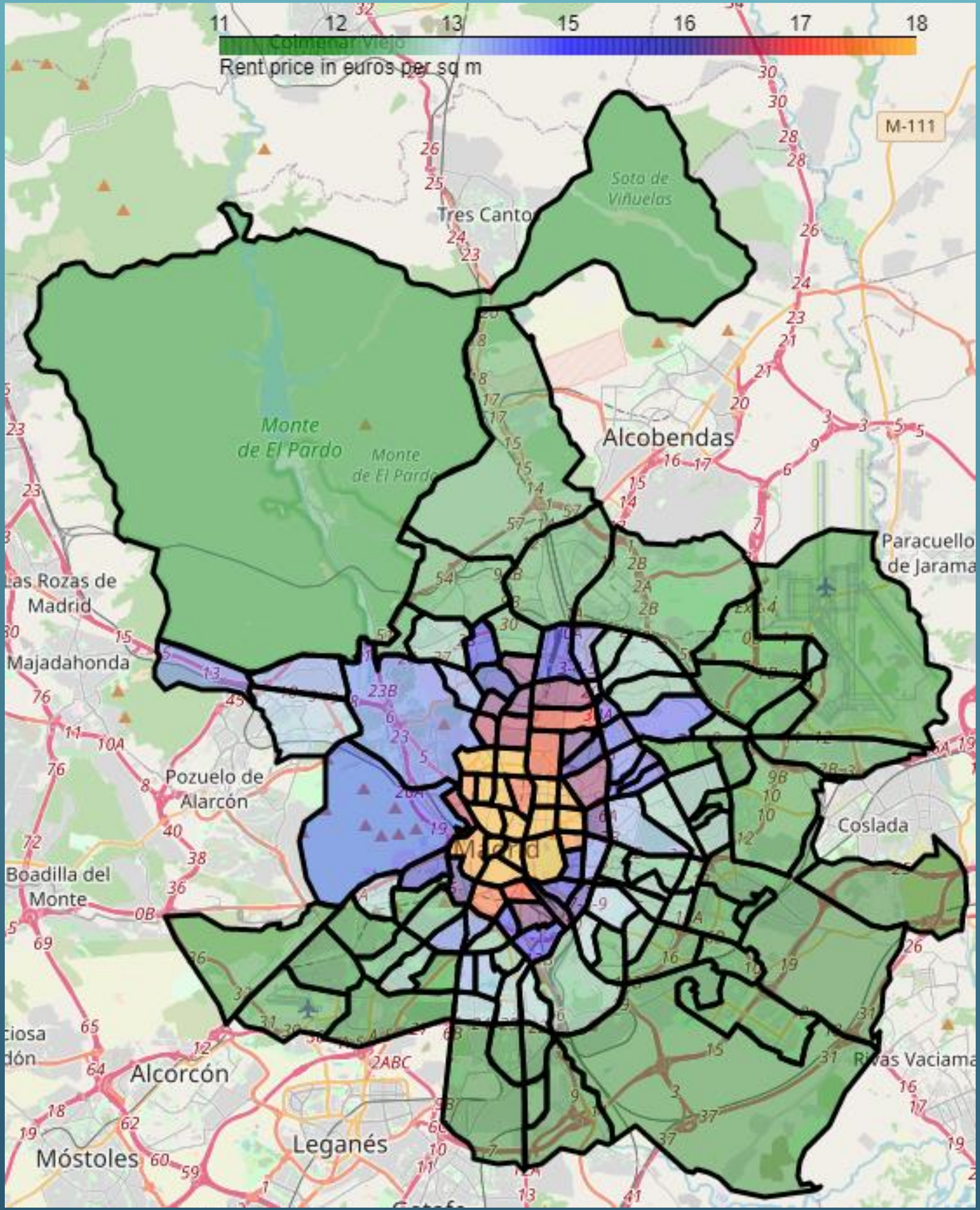


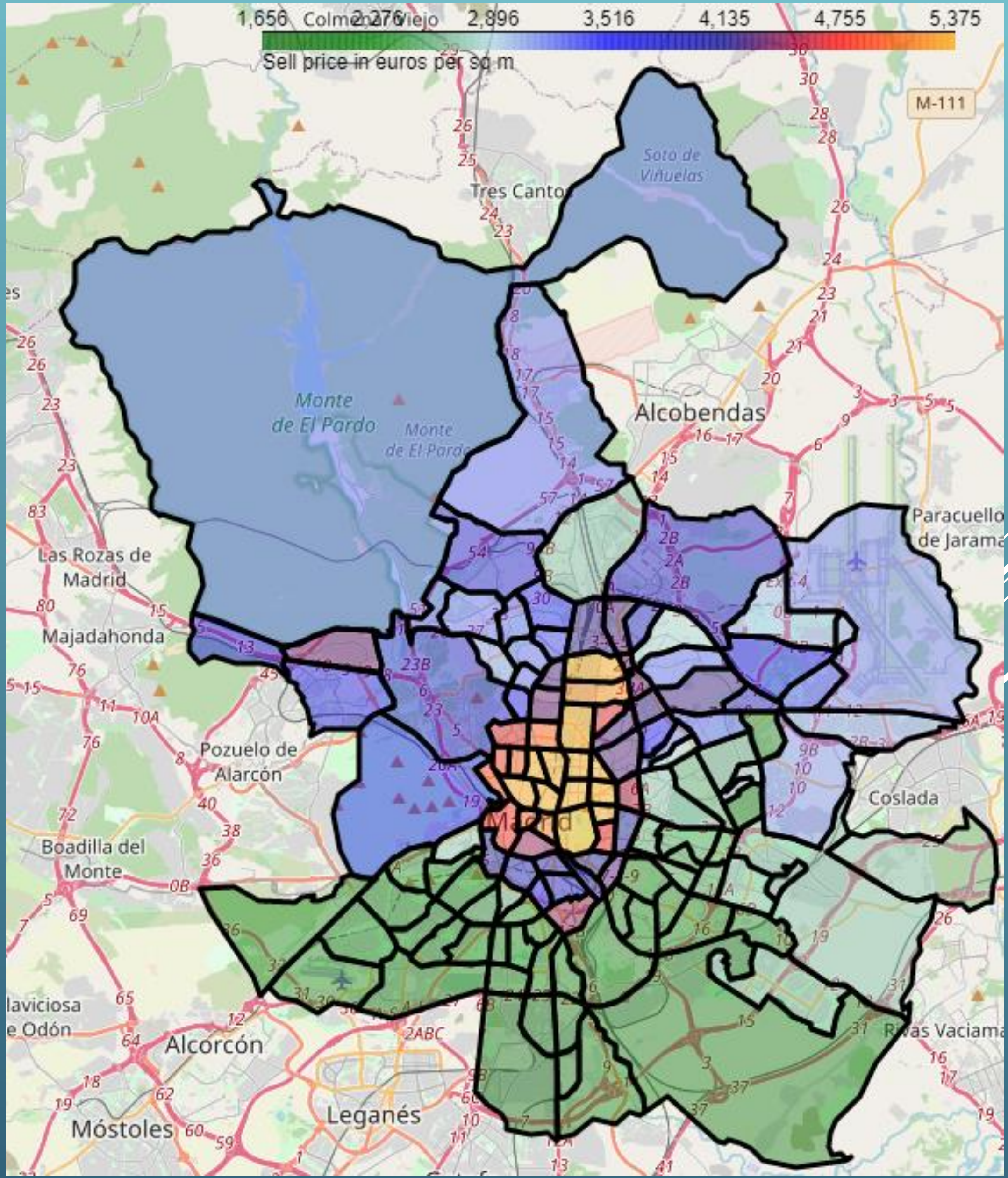
Data preparation – Feature Engineering

District	Sell price in euros per sq m	Rent price in euros per sq m
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

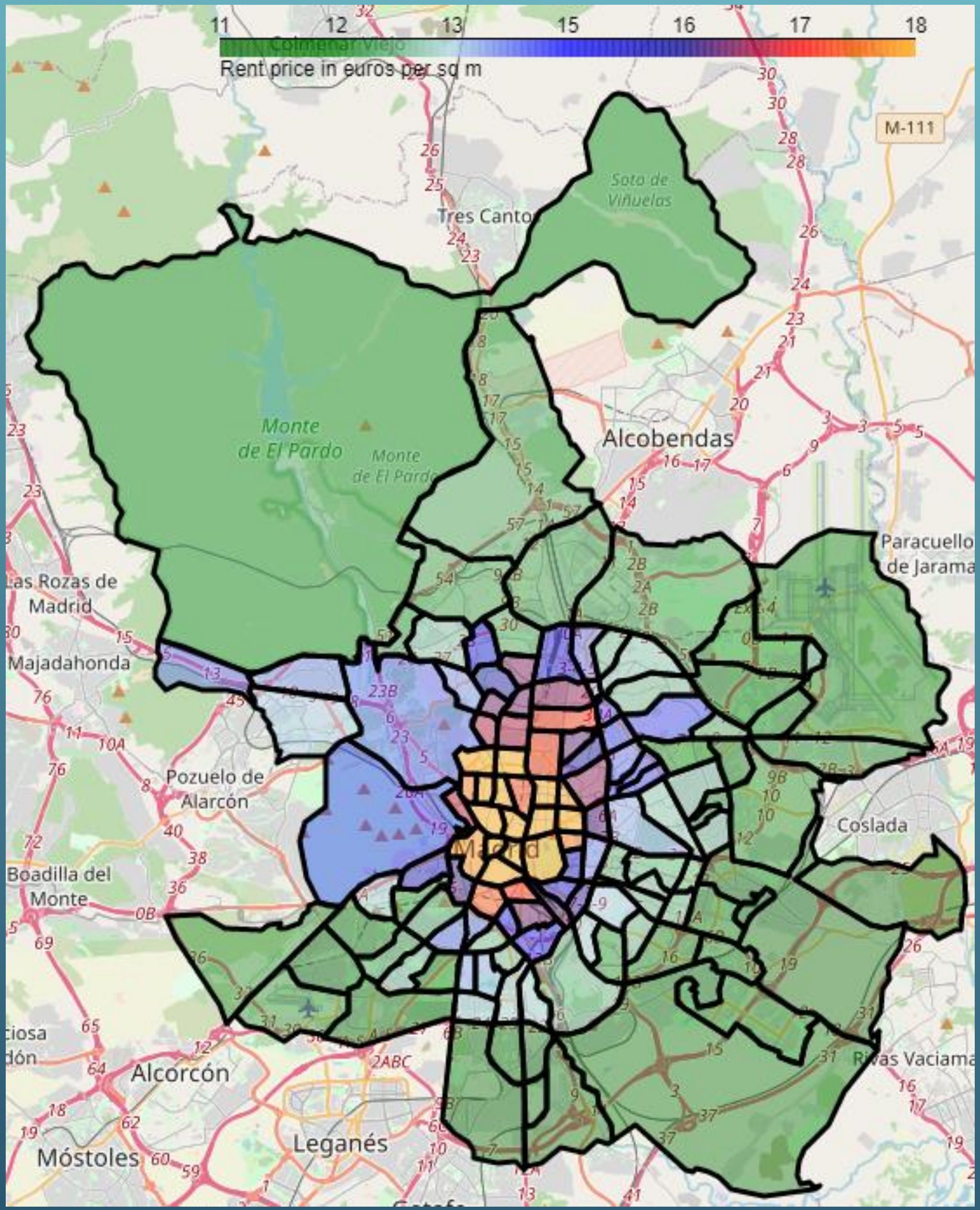
Real Estate data incomplete: Used median prices of the district for unknown values of the neighborhoods

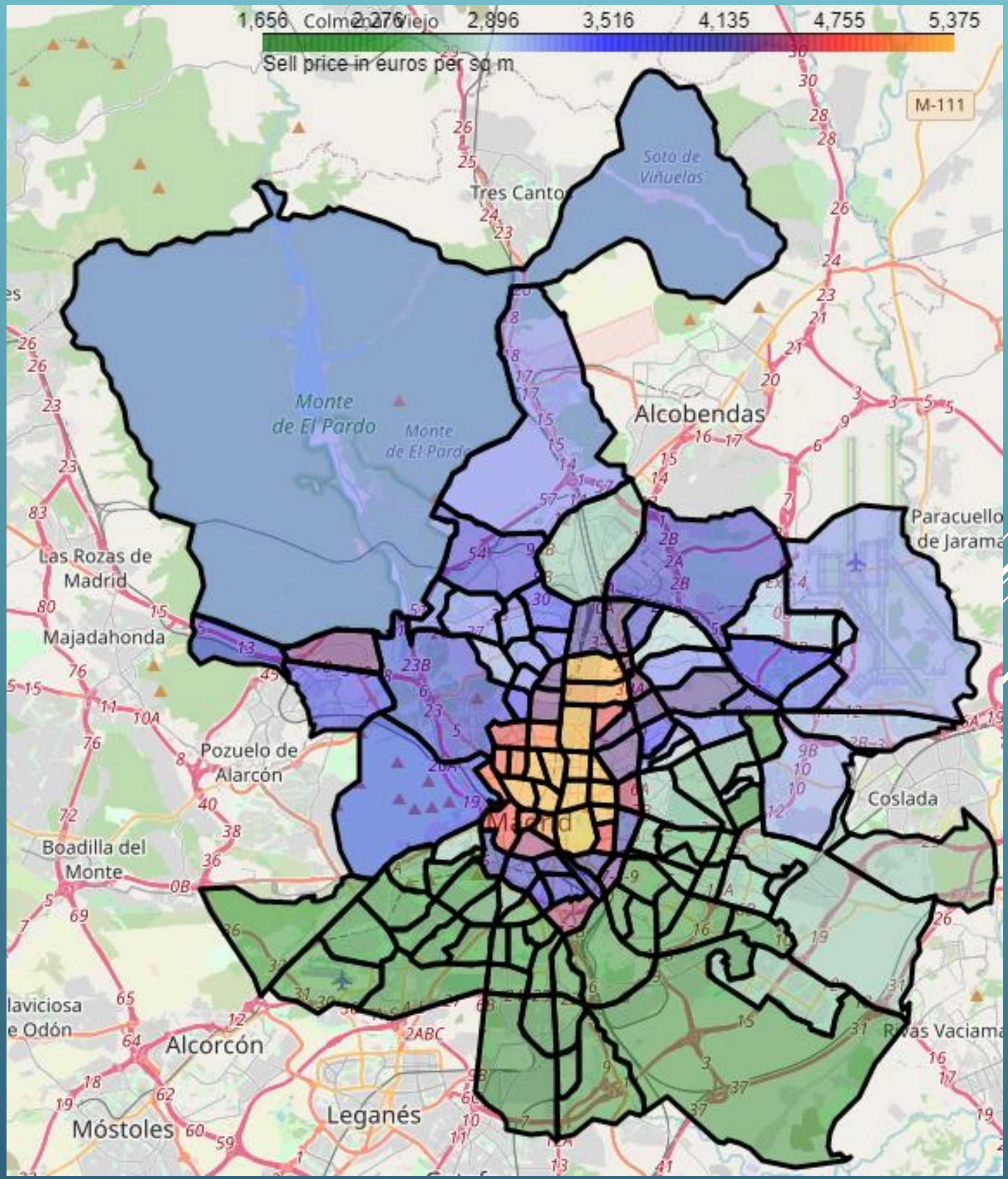
Data preparation – Feature Engineering





Data preparation – Feature Engineering





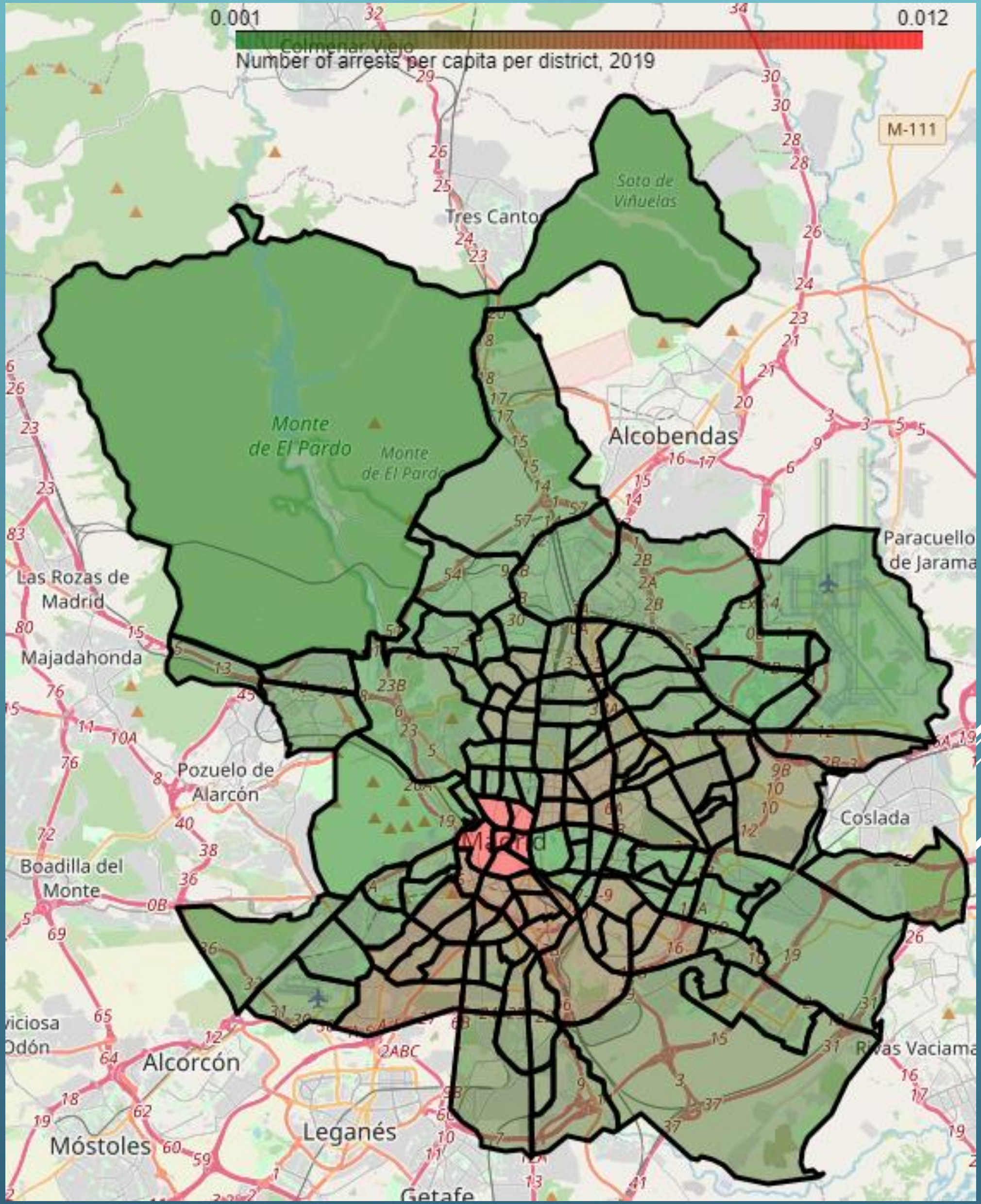
Data preparation – Feature Engineering

Crime data

+

Population data

Arrests per capita



Data preparation – Feature Engineering

	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

Amenities:

- Restaurants used as proxy of economic activity in the neighborhood.
- Counted number of restaurants per neighborhood.

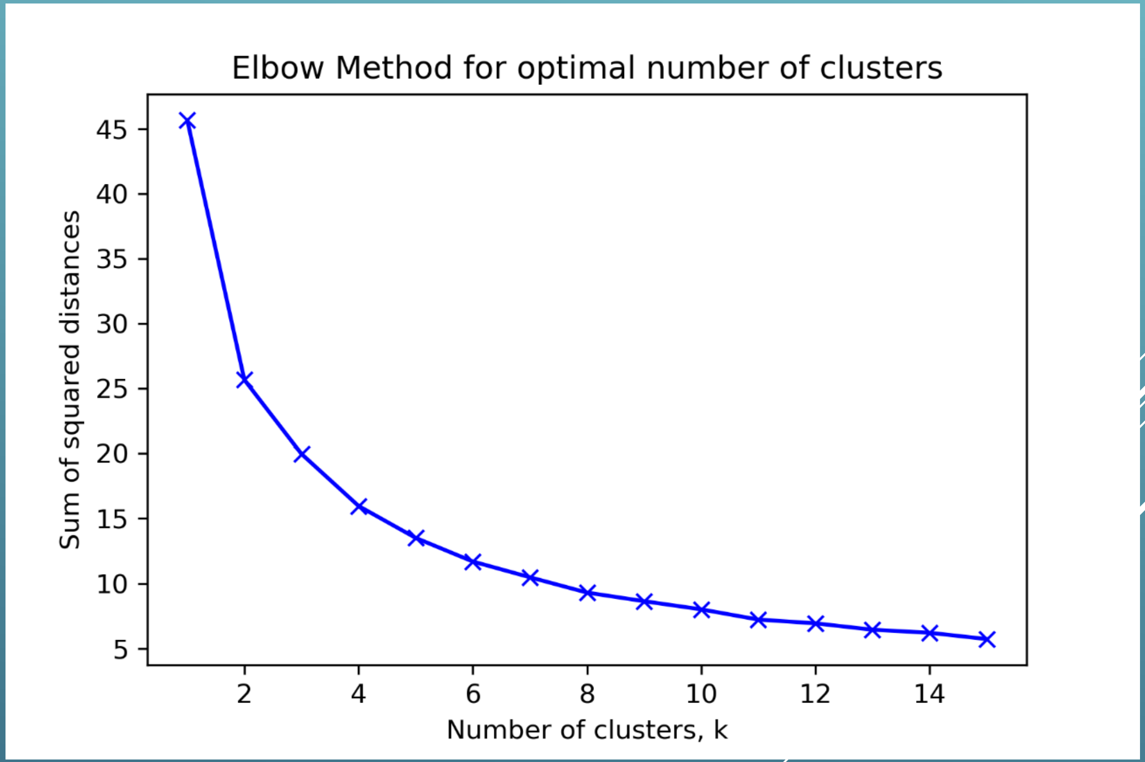
Modelling

Used a K-means clustering algorithm with the following features:

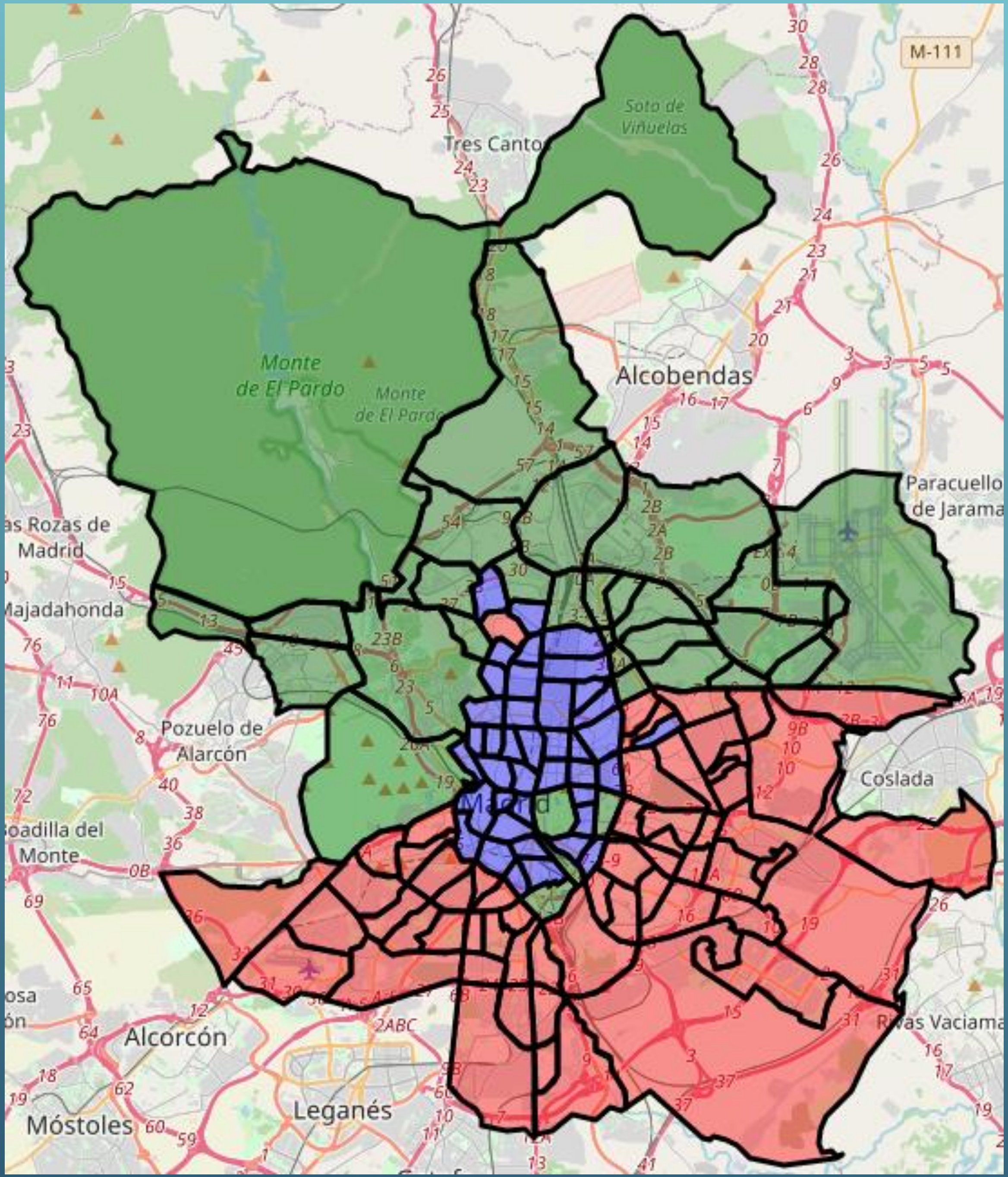
- Real Estate selling prices per square meter.
- Real Estate renting prices per square meter.
- Mean income per person.
- Population density in inhabitants per square kilometer.
- Number of arrests per capita.
- Number of restaurants in the neighborhood.

Data normalized using MinMaxScaler method.

Optimal number of clusters calculated using the elbow method.



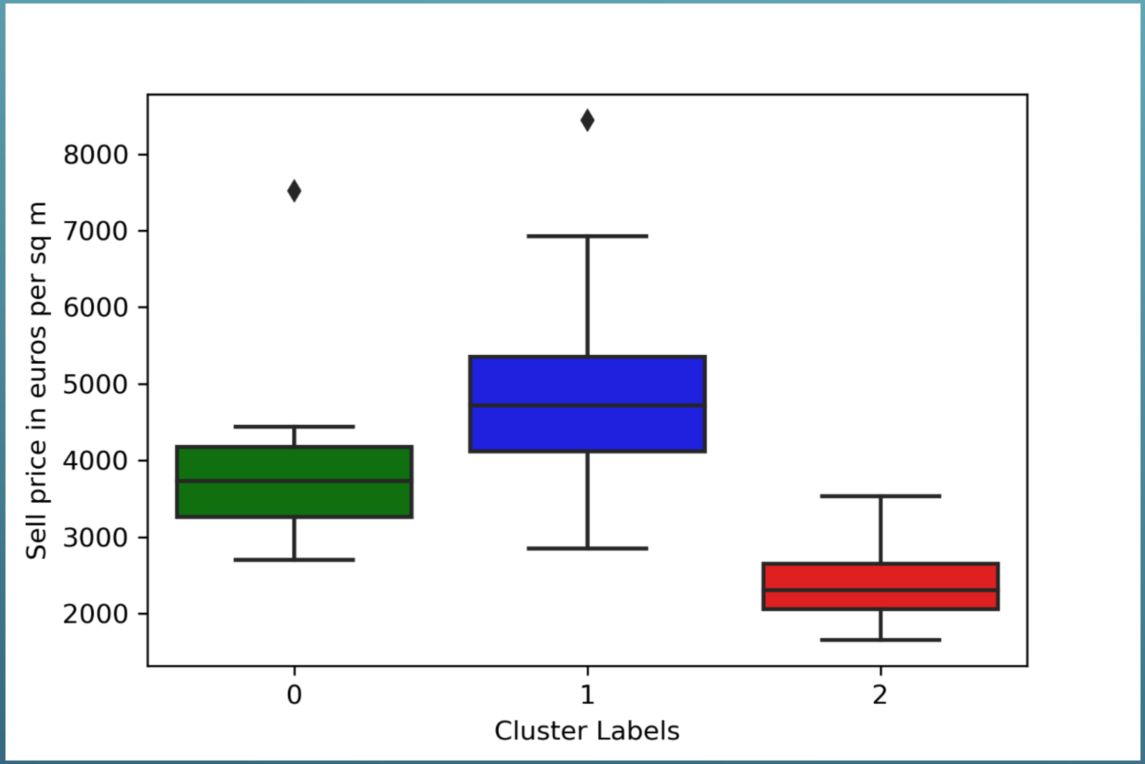
Modelling

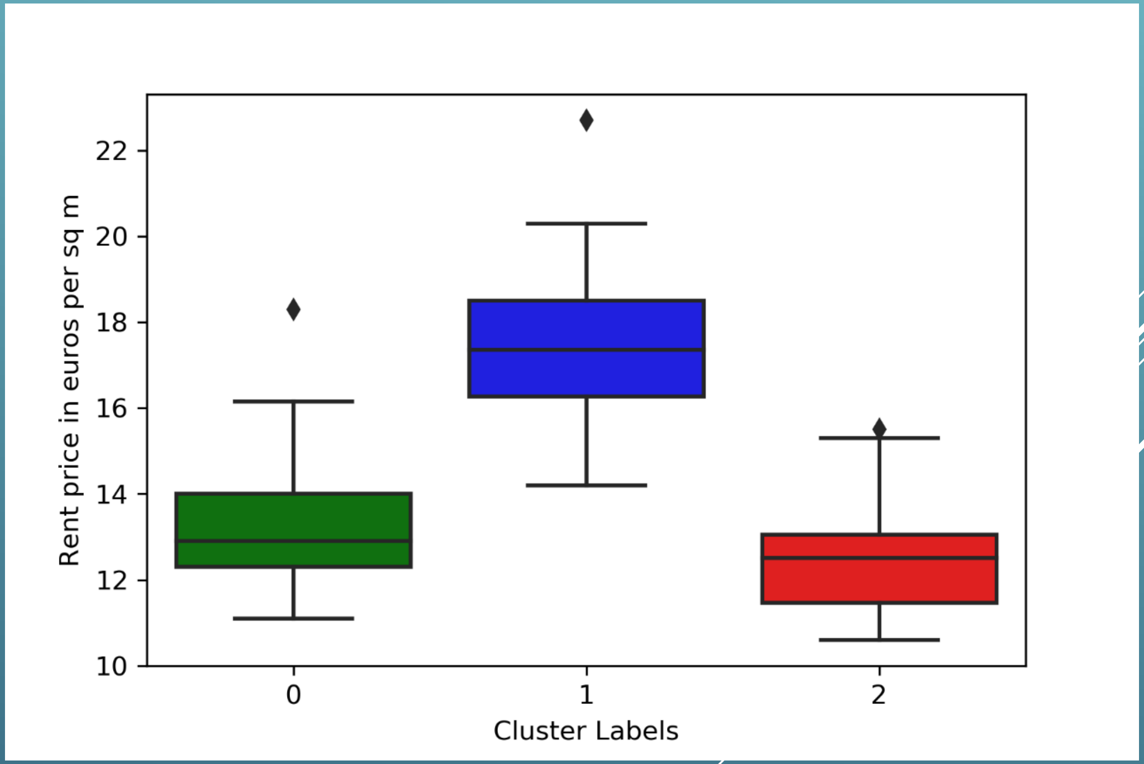


M-111
Soto de Viñuelas
Tres Cantos
Monte de El Pardo
Alcobendas
Paracuellos de Jarama
Las Rozas de Madrid
Majadahonda
Pozuelo de Alarcón
Boadilla del Monte
Madrid
Coslada
Alcorcón
Leganés
Móstoles
Rivas Vaciamadrid

Evaluation

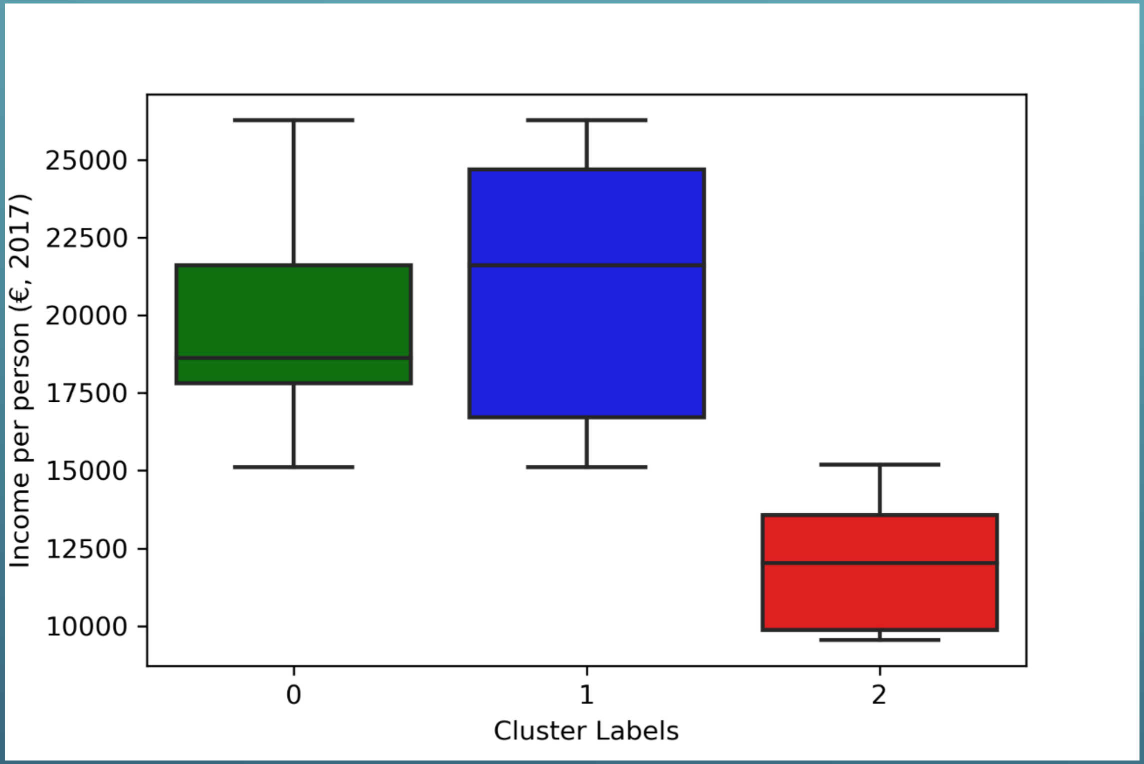
How are the features in each of the clusters?

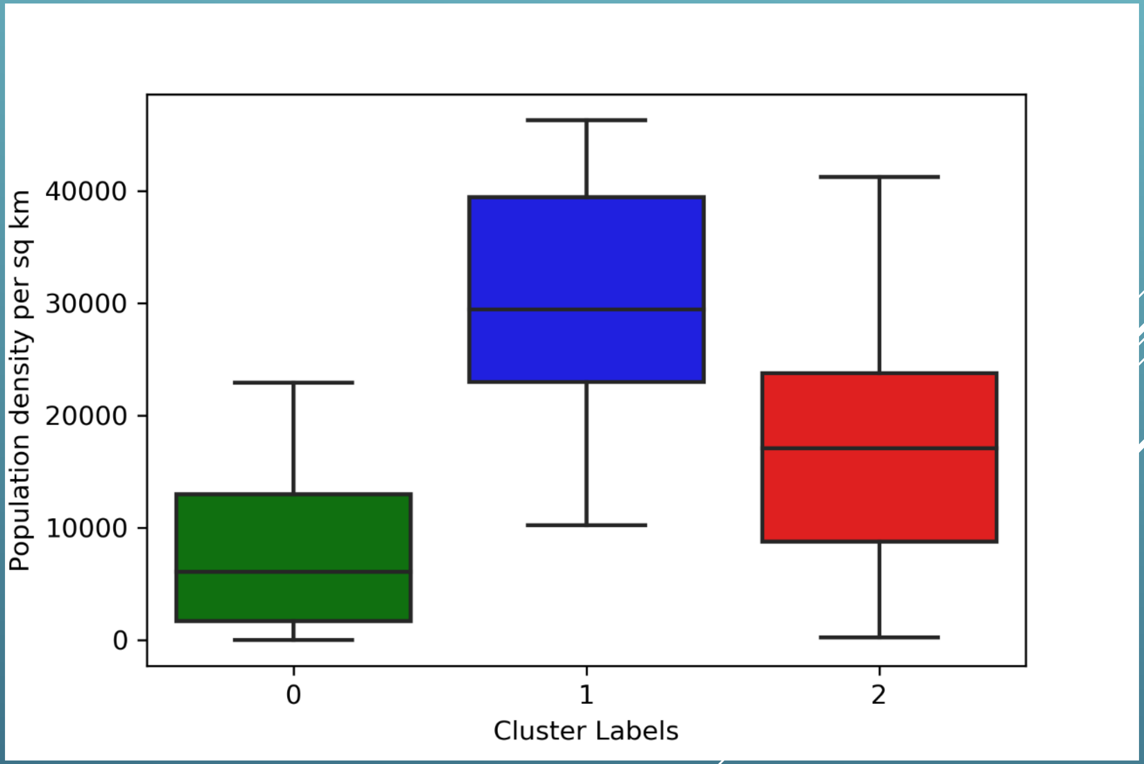




Evaluation

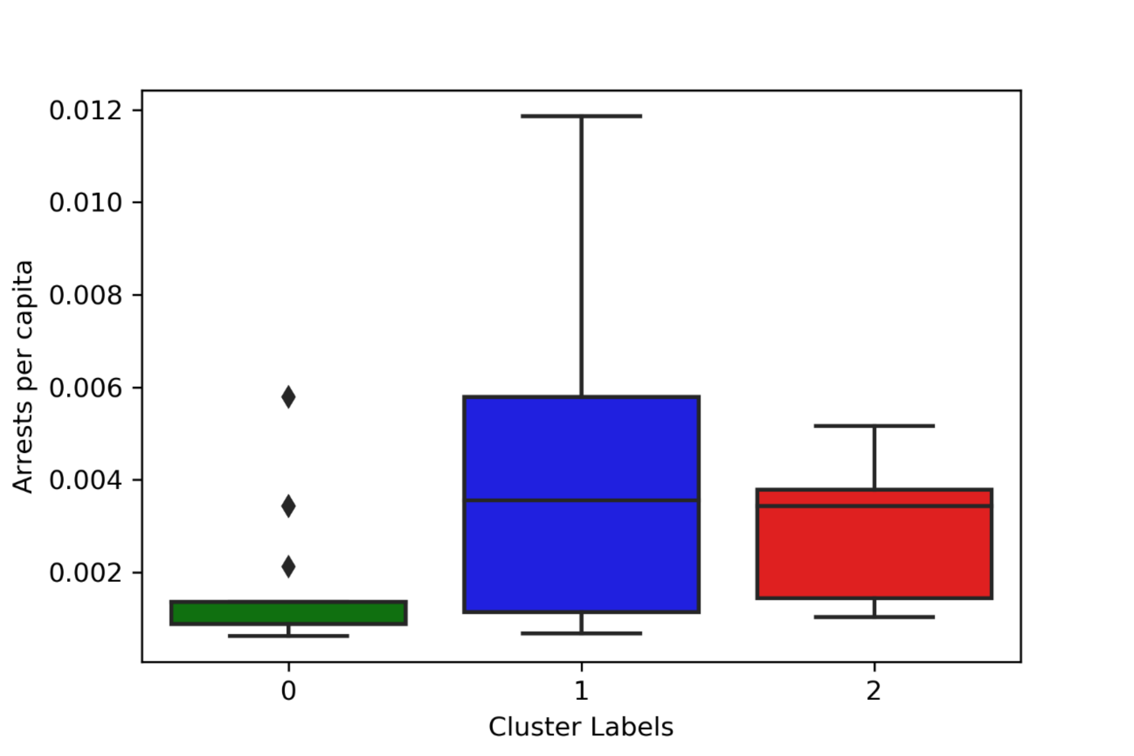
How are the features in each of the clusters?

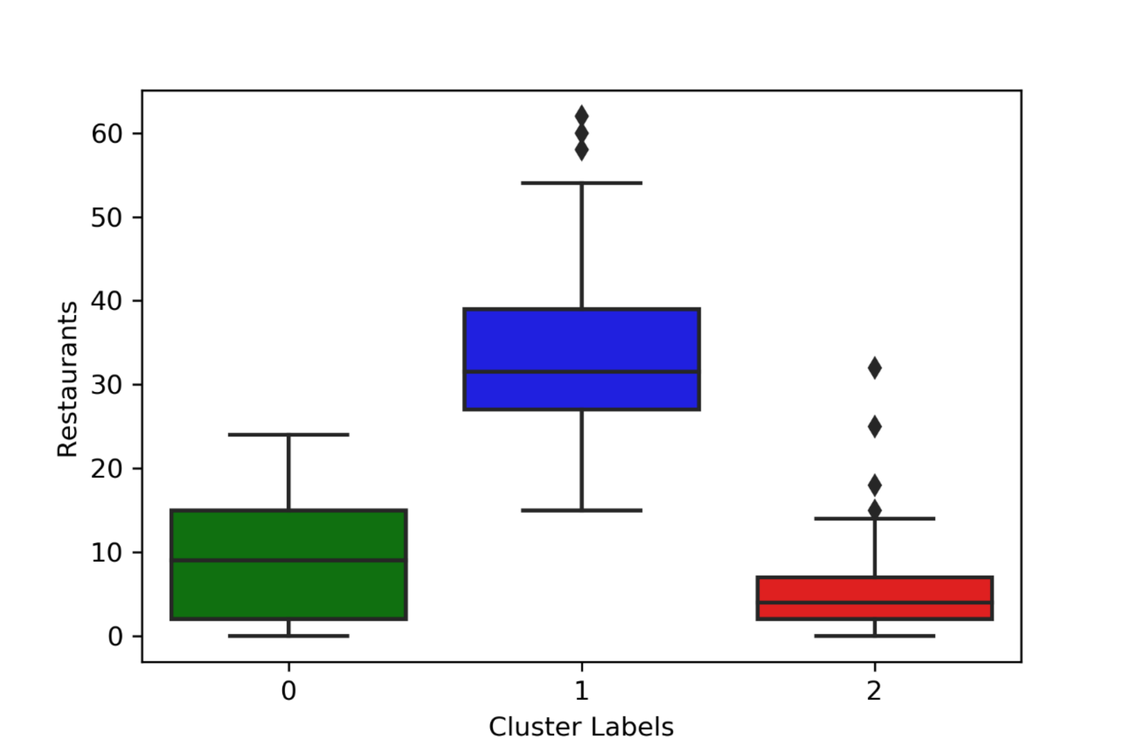




Evaluation

How are the features in each of the clusters?

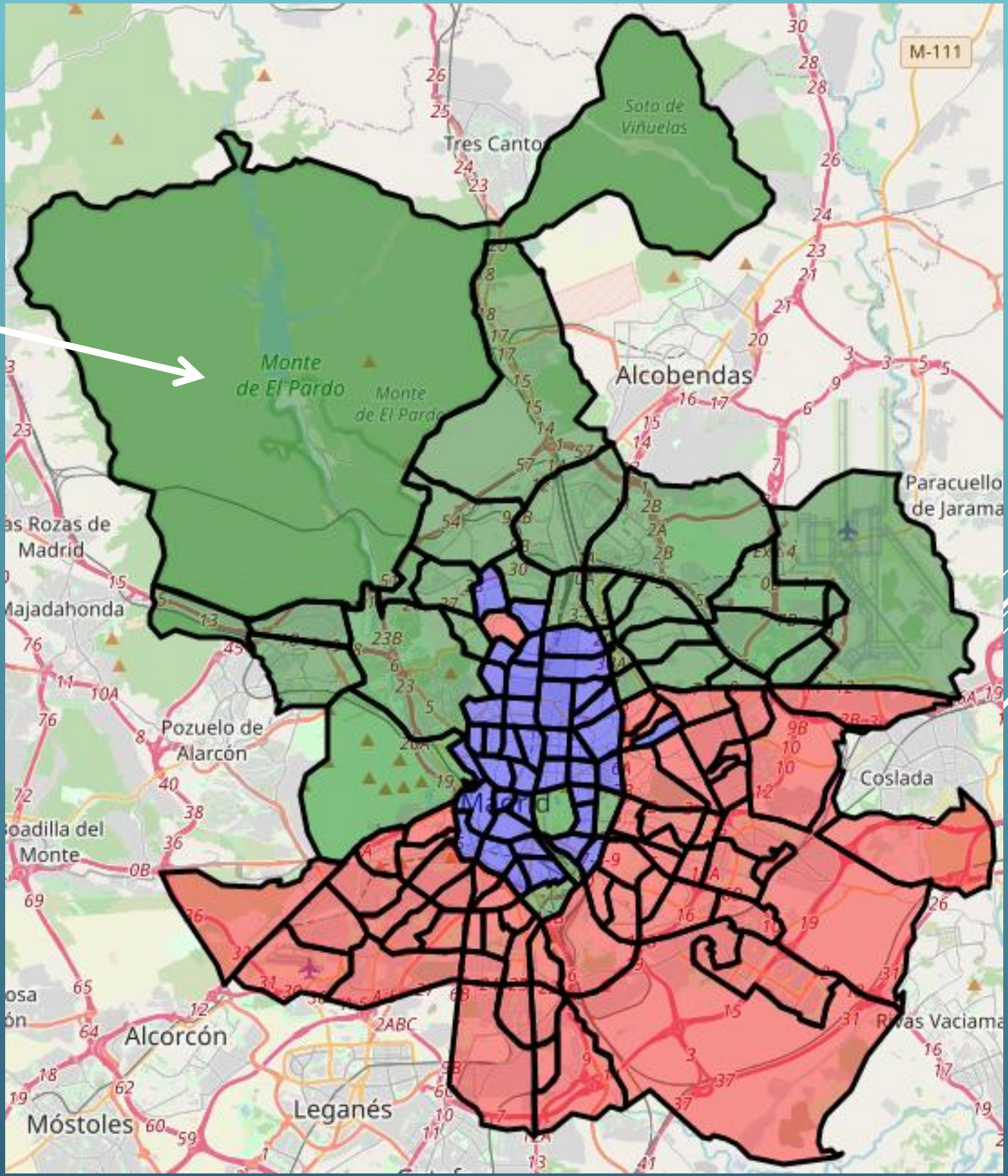




Evaluation

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

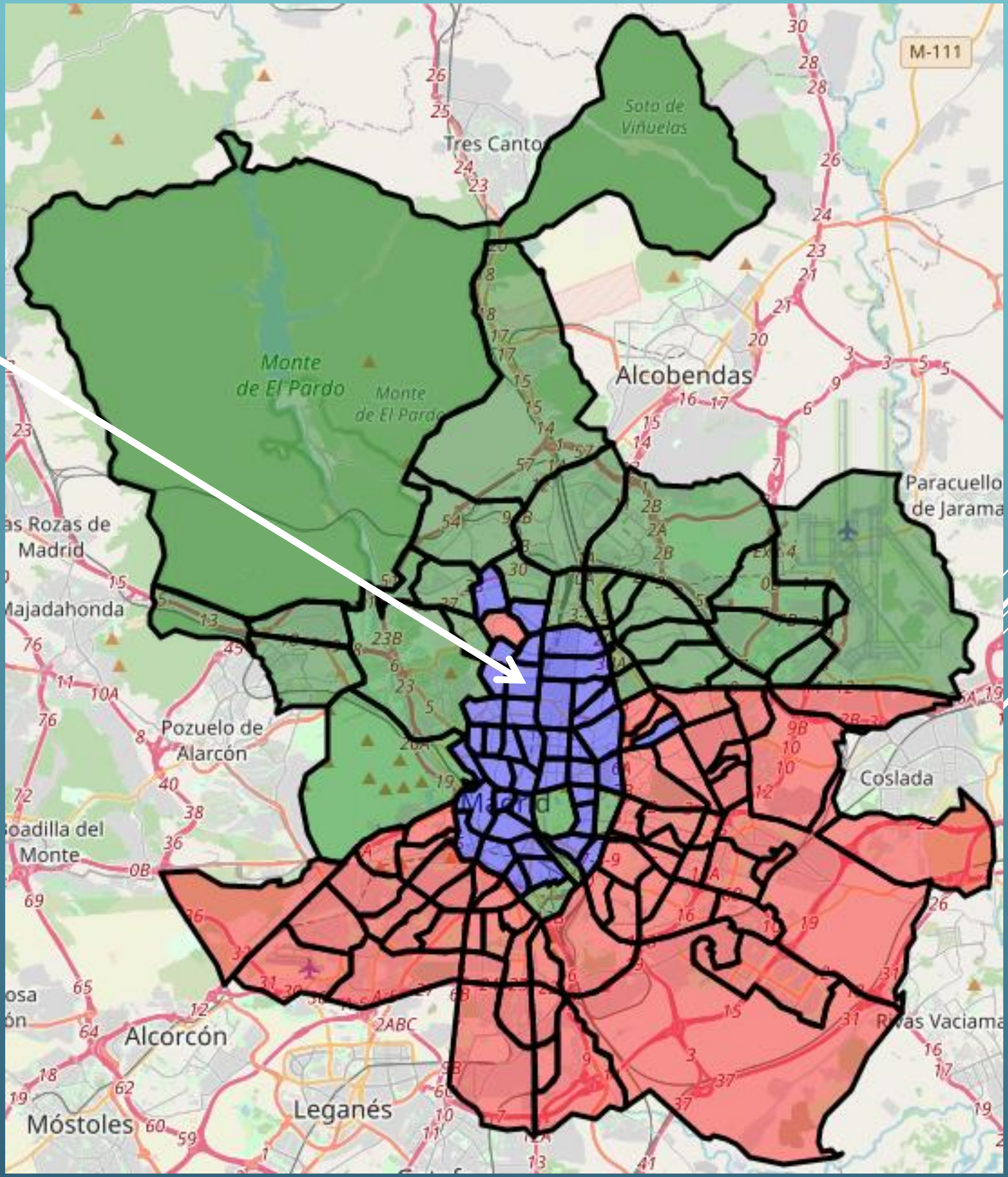
- Mainly the northern neighborhoods outside of the city center.
- Very low population density.
- High income.
- Moderate real estate prices for selling property.
- Low real estate prices for renting property.
- Very low crime levels.
- Low number of restaurants.



Evaluation

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

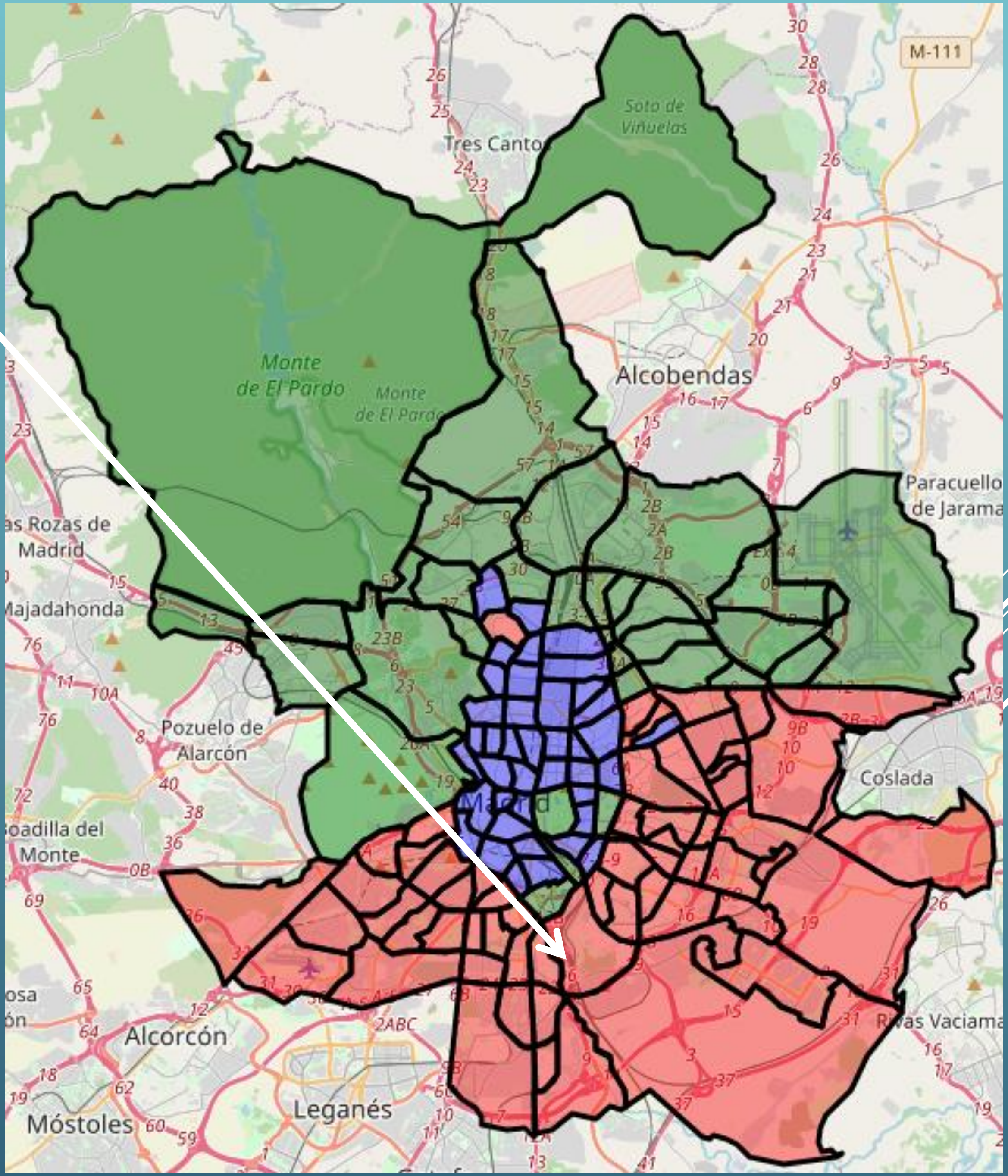
- Corresponding roughly to the city centre, inside the M-30 orbital motorway, the innermost ring road of the city.
- Highest population density of the three clusters.
- Highest income of the clusters.
- Highest real estate property value, especially for renting property.
- Highest numbers of arrest per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



Evaluation

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.



Conclusion

Main Goal:

Classify the neighborhoods of Madrid based on socioeconomic and business diversity in order to give information about living conditions in Madrid

Data taken from several official sources, realtors and Foursquare

The features were used to segment the neighborhoods into three clusters