The Battle of Neighborhoods – Report

Applied Data Science Capstone by IBM/Coursera

Table of contents

- Introduction: Business Problem
- Data
- Methodology
- Results
- Discussion
- Conclusion

Introduction: Business Problem

As I am currently living in Madrid, I have decided to choose this city for my final project in order to make it more personal.

Madrid is the capital and most populous city in Spain. The city has almost 3.3 million inhabitants and a metropolitan area population of approximately 6.5 million covering an area of 4,609.7 square kilometers (1,780 sq mi). It is the second-largest city in the European Union, surpassed only by Berlin.

Madrid is also the political, economic, and cultural center of the country. The Madrid urban agglomeration has the third-largest GDP in the European Union and its influence in politics, education, entertainment, environment, media, fashion, science, culture, and the arts all contribute to its status as one of the world's major global cities. (wikipedia.org)

That is why the Spanish capital is so attractive for foreign and local investors which means that the market is highly competitive. And the cost of housing is also one of the highest - in terms of a real estate property Madrid is considered as one of the most expensive cities in Spain.

One of the reasons for choosing a real estate property could be the price as well as the social establishments. In this project, I will examine neighborhoods of Madrid by average price per square meter for second-hand housing in April 2020 and cluster them according to the venues density. The analysis will give an understanding not only about a cost but also the most popular venues in certain areas and will be useful to anyone who is interested in investing in property in Madrid.

Data

The purpose of this work is to demonstrate the skills I have been equipped with from the previous modules of the course. So my goal was to combine my knowledge in using location data to explore a geographical location with visualization of geospatial data, choropleth maps specifically.

To solve the problem I have used the following data about Madrid:

Data 1:
Madrid has a total of 21 districts and 131 neighborhoods. I have collected this data as well as the average prices of second-hand housing by a square meter in Madrid in April 2020 from the real estate portal idealista.com. I have used this data as an input for clustering analysis as well as for creating a choropleth map.

Data 2:
To explore the neighborhoods I have collected its latitude and longitude coordinates from GeoHack and merged them with the Data 1. This data was used as an input for the Foursquare API.

Data 3:
I have used the Foursquare API to get top venues of each neighborhood and explore them.

Data 4:
For creating a choropleth map I needed neighborhood divisions formatted in JSON. I have found them on GitHub.

So, the data acquired for this project is a combination of four data sources.

Methodology

For the aim of the project, I had to have sale prices of second-hand housing in each neighborhood in Madrid. I have collected this data from the real estate portal idealista.com and I have created a table.

To clean the data set, I have dropped all the columns with the information that was not necessary for the project, renamed columns I wanted to use and I have changed several districts' names to their composed names as they were outdated. And I have got the following input data that served me as a base for clustering analysis and choropleth map creation:

	District	Neighborhood	Price
0	Arganzuela	Acacias	3920
1	Arganzuela	Chopera	3578
2	Arganzuela	Delicias	3809
3	Arganzuela	Imperial	4099
4	Arganzuela	Legazpi	4436

To be sure about the right total of districts and neighborhoods, I have used the .shape method and got the following result:

```
The dataframe has 21 districts and 131 neighborhoods.
```

I have used the .describe method as well to see the statistical summary of the data set:

	Price
count	131.000000
mean	3436.488550
std	1351.452002
min	1400.000000
25%	2309.000000
50%	3252.000000
75%	4243.000000
max	8440.000000

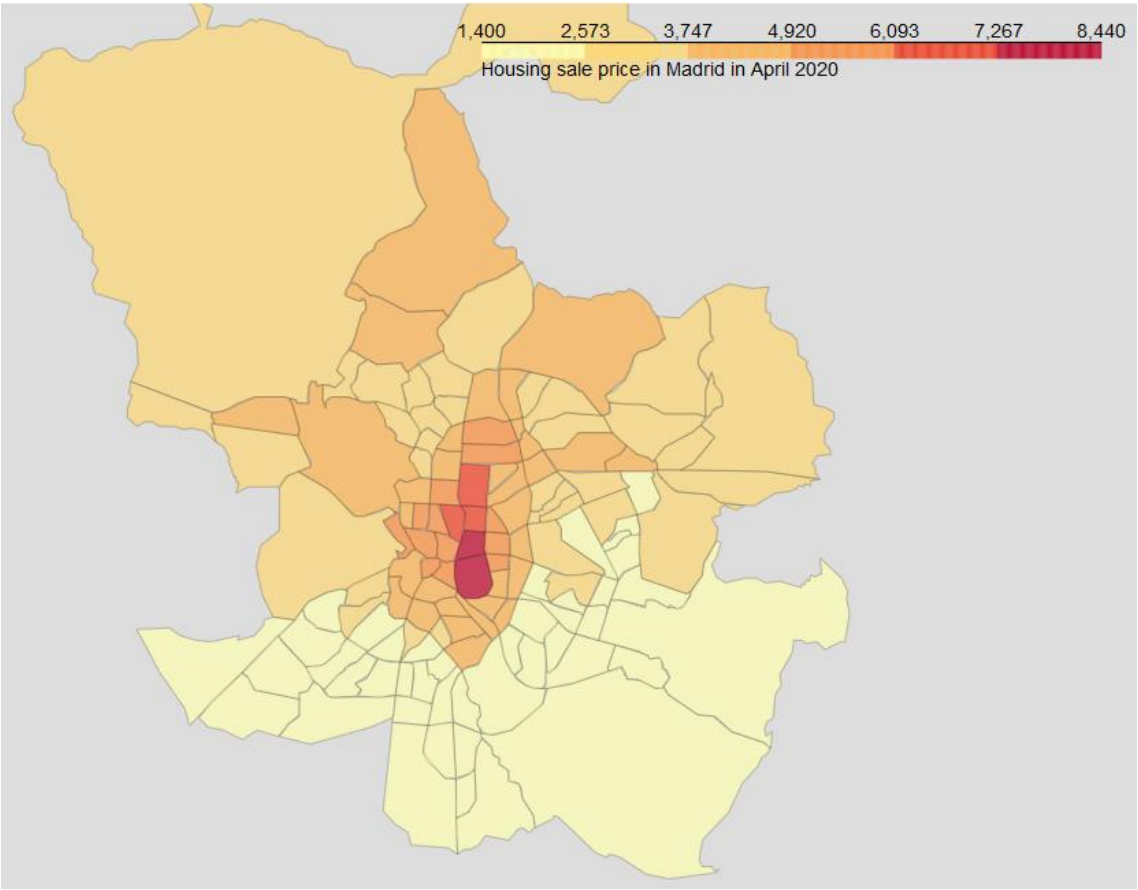
And the following conclusions are:

a) There are 131 unique values that represent the number of neighborhoods of Madrid;
b) The average price of square meter for second-hand housing in Madrid in April 2020 equals to 3.436 €/m2;
c) The minimum and maximum price of square meter for second-hand housing in Madrid in April 2020 equal to 1.400 €/m2 and 8.440 €/m2 respectively.

And the minimum and maximum prices belong to the following neighborhoods:

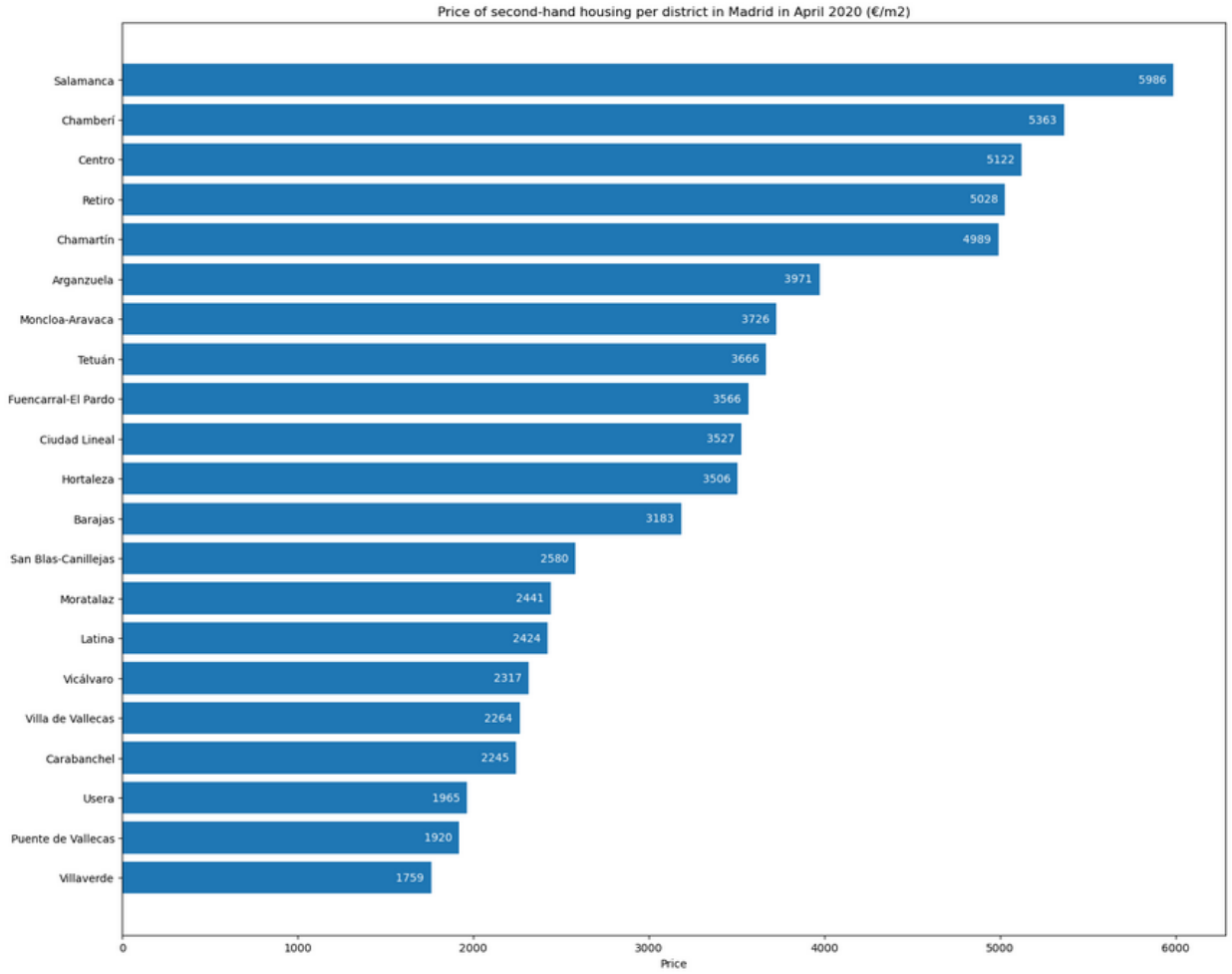
```
Max price per m2:
     District Neighborhood  Price
97  Salamanca    Recoletos   8440
Min price per m2:
       District   Neighborhood  Price
130  Villaverde  San Cristobal   1400
```

For a better understanding I have visualized Madrid´s neighborhoods using a choropleth map with shaded areas in proportion to sale prices:



Depending on the intensity of the color it is quite easy to see how sale prices vary across neighborhoods. We can see the most intense color in the center of the city and the less intense in its south.

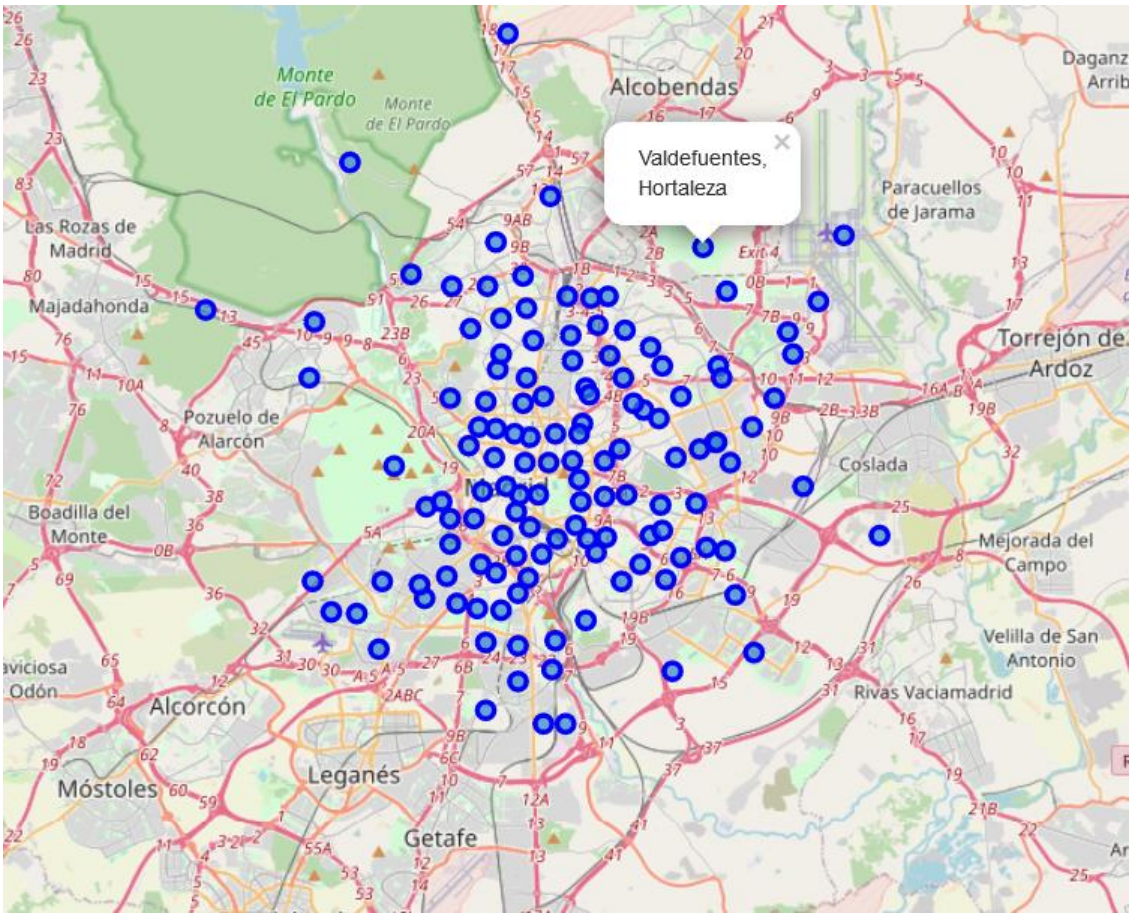
Based on the same data set I have also analyzed average prices by districts from where we can see Salamanca district with the average price of 5.986 €/m2 and Villaverde district with the average price of 1.759 €/m2:



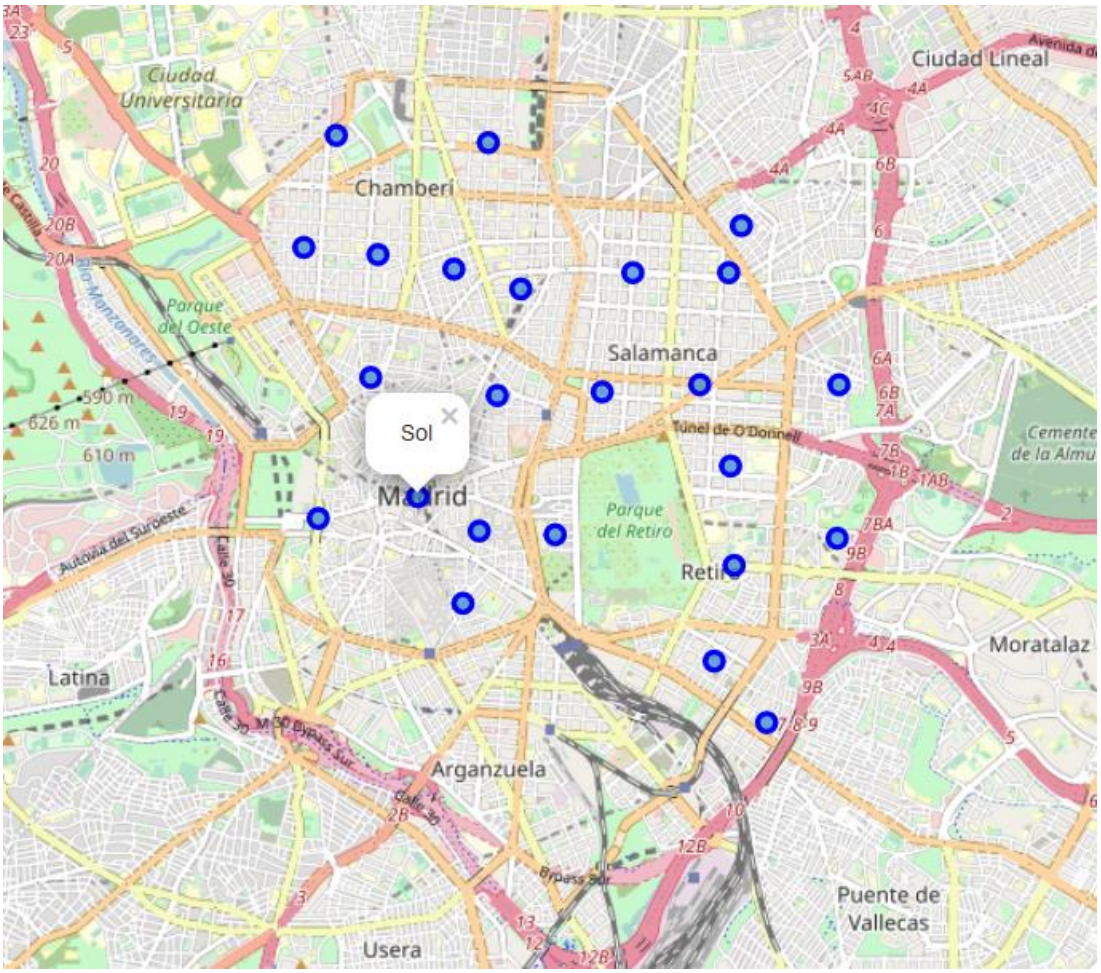
To explore the neighborhoods I have compiled their latitude and longitude coordinates from GeoHack and merged them with the first data set:

	District	Neighborhood	Price	Latitude	Longitude
0	Arganzuela	Acacias	3920	40.401422	-3.704936
1	Arganzuela	Chopera	3578	40.395000	-3.699444
2	Arganzuela	Delicias	3809	40.395833	-3.689444
3	Arganzuela	Imperial	4099	40.406667	-3.716944
4	Arganzuela	Legazpi	4436	40.388611	-3.695178

I have used GeoPy library to get the latitude and longitude values of Madrid and created a map of Madrid with neighborhoods superimposed on top:



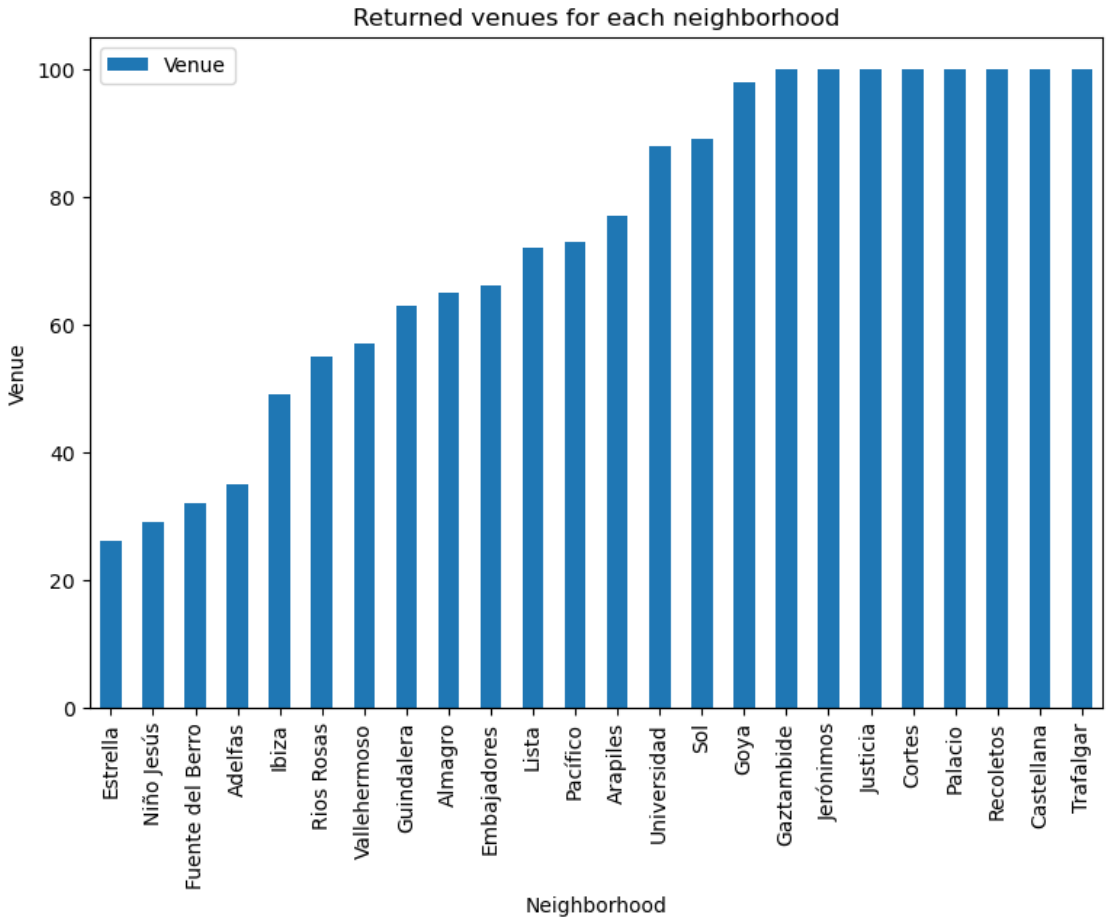
For illustration purposes, let's suppose our investor is interested in four city center districts such as Centro, Chamberí, Salamanca, and Retiro, and let's simplify the above map and segment and cluster only neighborhoods of the mentioned districts. So I have sliced the original data frame and created a new data frame of the city center data. Here is the visualization of the neighborhoods of four districts in the city center on the map:



Next, I have used the Foursquare API to explore the neighborhoods and segment them. I have set a limit of 100 venues that are in each neighborhood within a radius of 500 meters. And here is a head of a data frame generated:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Justicia	40.423889	-3.696389	Honest Greens	40.424880	-3.697894	Restaurant
1	Justicia	40.423889	-3.696389	DSTAgE	40.424729	-3.696305	Restaurant
2	Justicia	40.423889	-3.696389	Only YOU Hotel&Lounge	40.422227	-3.695762	Hotel
3	Justicia	40.423889	-3.696389	Bee Beer	40.421952	-3.696900	Beer Store
4	Justicia	40.423889	-3.696389	Plaza de Chueca	40.422724	-3.697588	Plaza

I have also counted the number of venues returned for each neighborhood and visualized them:



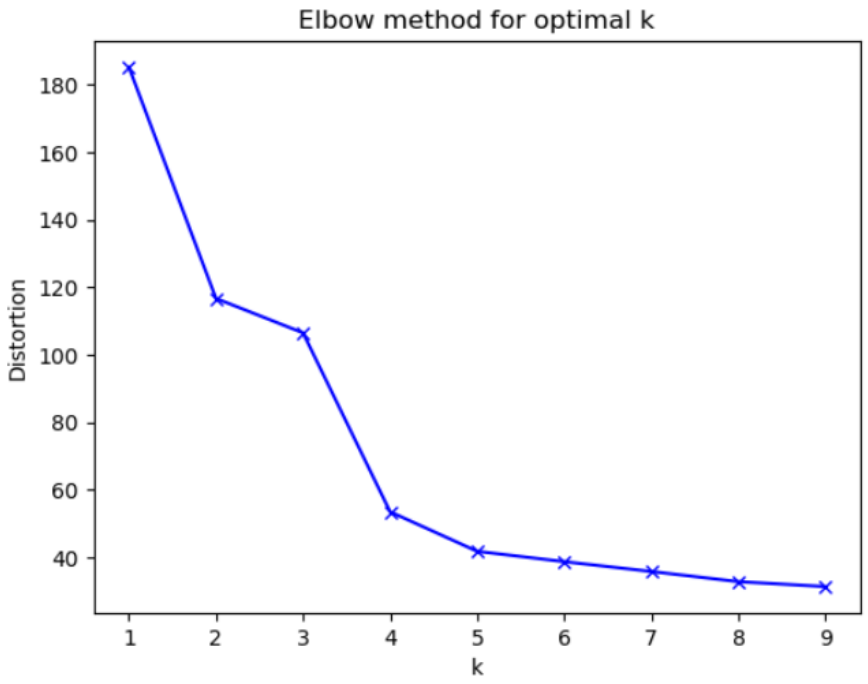
Out of 24 neighborhoods, 8 reached the limit of the number of venues, while the other 4 are below 40 venues per neighborhood.

The number of unique categories that can be curated from all the returned venues is 206.

For analyzing each neighborhood I have used one-hot encoding and reduced the number of returned venues to top 10:

	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
0	Adelfas	Fast Food Restaurant	Supermarket	Bar	Hotel	Grocery Store	Gym	Bookstore	Spanish Restaurant	Breakfast Spot	Brewery
1	Almagro	Restaurant	Spanish Restaurant	Italian Restaurant	Bar	Plaza	Nightclub	Pub	Breakfast Spot	Salad Place	Café
2	Arapiles	Spanish Restaurant	Bar	Café	Bakery	Tapas Restaurant	Burrito Place	Hotel	Restaurant	Sandwich Place	Multiplex
3	Castellana	Spanish Restaurant	Restaurant	Boutique	Coffee Shop	Tapas Restaurant	Mediterranean Restaurant	Japanese Restaurant	Gym / Fitness Center	Bar	Seafood Restaurant
4	Cortes	Bar	Café	Plaza	Hotel	Restaurant	Tapas Restaurant	Mediterranean Restaurant	Spanish Restaurant	Market	Pizza Place

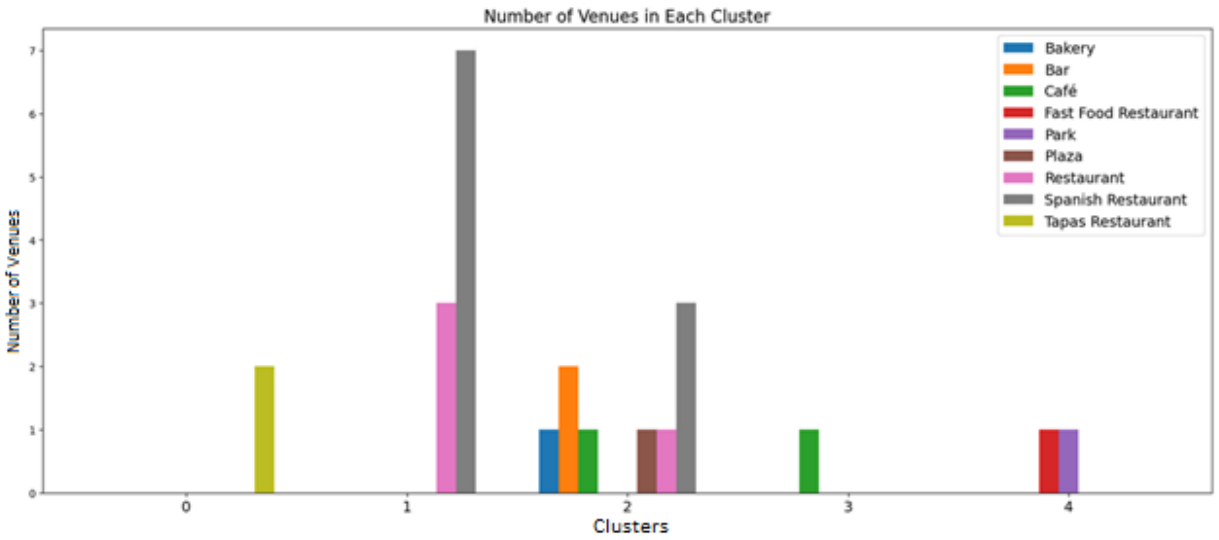
As there are common venues in the data set, I have used the K-means method as a type of unsupervised learning and one of the popular methods of clustering unlabeled data into k clusters. To find an optimal value of k, I have used the Elbow method that returned k = 5:



After that, I have created a new data frame that includes the clusters as well as the top 10 venues for each neighborhood:

	District	Neighborhood	Price	Latitude	Longitude	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	
0	Centro	Justicia	5983	40.423889	-3.696389	2	Bakery	Spanish Restaurant	Restaurant	Hotel	Bookstore	Deli / Bodega	Flower Shop	V
1	Centro	Cortes	5159	40.414167	-3.698056	2	Bar	Café	Plaza	Hotel	Restaurant	Tapas Restaurant	Mediterranean Restaurant	
2	Centro	Embajadores	4455	40.408889	-3.699722	2	Bar	Tapas Restaurant	Café	Spanish Restaurant	Market	Bookstore	Restaurant	
3	Centro	Universidad	5337	40.425278	-3.708333	2	Café	Bookstore	Hotel	Spanish Restaurant	Plaza	Tapas Restaurant	Cocktail Bar	
4	Centro	Palacio	4800	40.415000	-3.713333	0	Tapas Restaurant	Spanish Restaurant	Plaza	Bar	Restaurant	Church	Vegetarian / Vegan Restaurant	Med

I have visualized a number and a kind of venues in clusters so that it was easier to label each of them:



After examining the bar chart, I have named each label as follows:

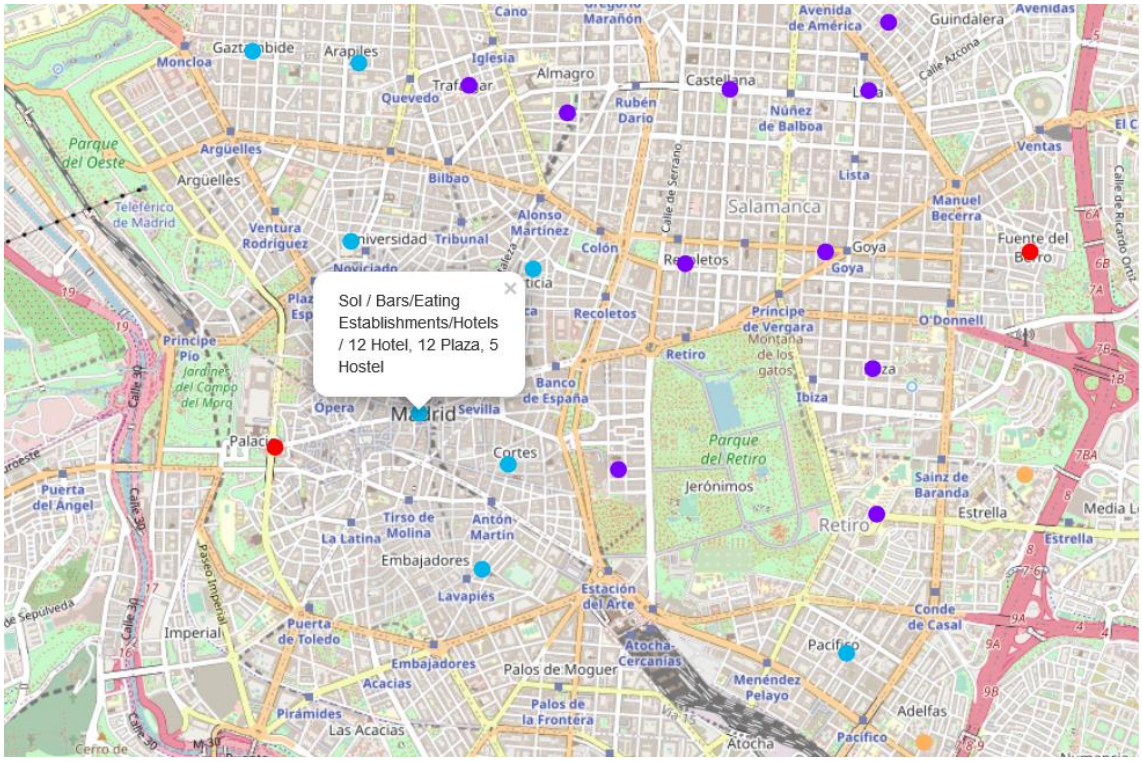
	Clusters	Labels
0	0	Tapas Restaurants
1	1	Restaurants
2	2	Bars/Eating Establishments/Hotels
3	3	Cafes
4	4	Fast Food Restaurants/Park

For last, I have grouped each neighborhood by the number of top 3 venues to show them on the map:

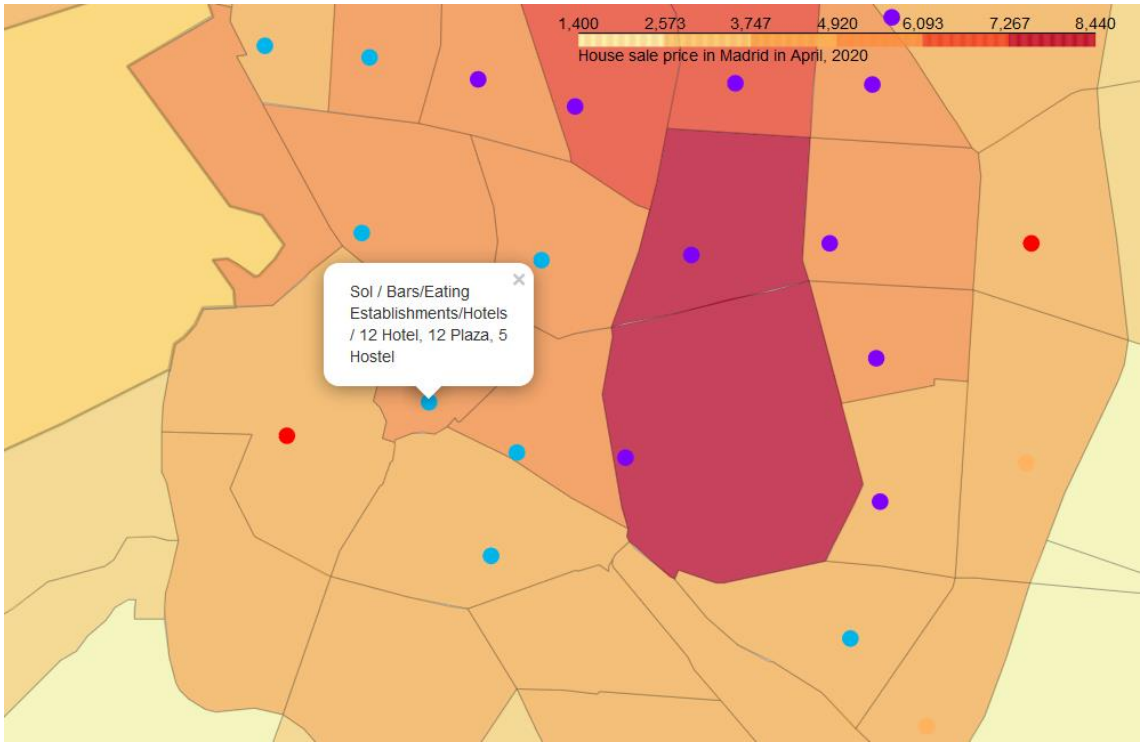
	Neighborhood	Top 3 venues
0	Adelfas	3 Fast Food Restaurant, 2 Bar, 2 Grocery Store
1	Almagro	8 Restaurant, 6 Spanish Restaurant, 3 Bar
2	Arapiles	8 Spanish Restaurant, 6 Bar, 5 Bakery
3	Castellana	16 Spanish Restaurant, 10 Restaurant, 8 Boutique
4	Cortes	7 Bar, 7 Café, 6 Hotel

Results

I have merged new variables (Top 3 venues and Labels) with related cluster information in the result table and visualized the clusters:



And finally, I have joined the cluster map with the choropleth map:



Discussion

The given analysis of Madrid neighborhoods was performed based on the housing sale prices of Madrid in April 2020 and venues density in the neighborhoods. For the project, I have utilized four data resources and two machine learning methods such as one-hot encoding and K-means. In addition, I have combined location data to explore a geographical location with visualization of geospatial data.

Housing sale prices and venues density data analysis could also be performed for the rest of the neighborhoods or based on more accurate parameters such as the number of venues and radius for segmenting them or even coordinates to explore the streets, for instance.

Also, venues density used for the project was one of the examples of choosing a real estate property and of course, the analysis could also be executed for another type of clusters. Comparison between housing sales pricing and, for instance, educational institutions, transport, or grocery shops accessibility could be another option to consider.

Certainly, this would depend on the purposes of the project and interested audience.

Conclusion

Madrid has a total of 21 districts and 131 neighborhoods. Housing sale prices vary across the neighborhoods. The center city neighborhoods have the highest cost per square meter and the neighborhoods in the south have the lowest cost. There is also a significant difference between the north and the south of the city, being the north more expensive than the south.

The average price of second-hand housing in the Spanish capital in April 2020 is 3.436 €/m2. Recoletos, in Salamanca district, with 8.440 €/m2 and San Cristobal, in Villaverde district, with 1.400 €/m2 have the maximum and minimum cost per square meter respectively.

For analysis and segmenting only 24 neighborhoods out of 131 were left and only 8 of them reached the venues limit of 100, 4 are below 40 venues per neighborhood. The number of unique categories that can be curated from all the returned venues is 206.

Due to the Elbow method, all the venues were clustered in 5 clusters with the following given names: Tapas Restaurants, Restaurants, Bars/Eating Establishments/Hotels, Cafes, Fast Food Restaurants/Park.

As a result, each neighborhood has information about its average sale price, cluster name, and top venues in it.

The obtained result can be useful for investors in making decisions in purchasing real estate property in Madrid.