# **House Prices in Madrid**

Abdurrahim AKBULUT

İstanbul, TR

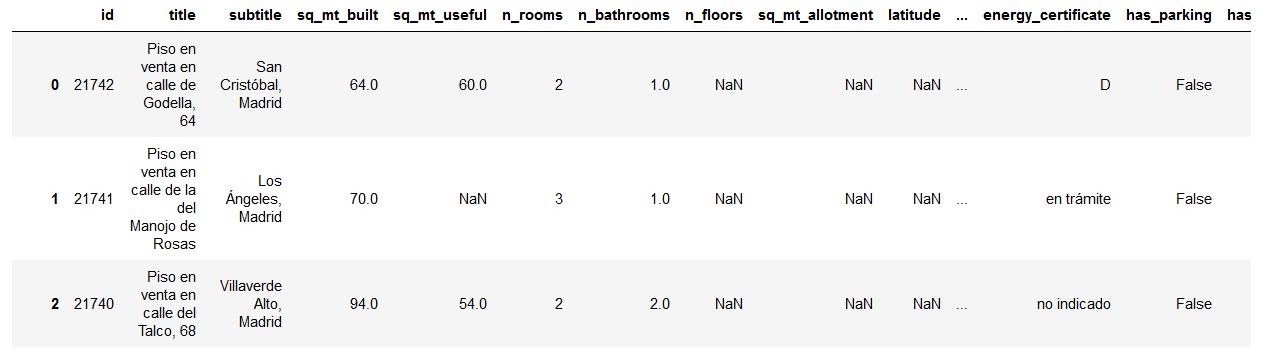
## **1.Introduction**

Madrid is the capital and most populous city of Spain. The city has almost 3.3 million inhabitants and a metropolitan area population of approximately 6.5 million. It is the second-largest city in the European Union (EU), surpassed only by Berlin. [1] There are 21 districts and 131 neighborhoods in Madrid. [2]

Imagine someone wants to buy a house in Madrid, but not sure about the location. It would be nice if they can compare each neighborhood based of some characteristics. So in this project, we'll analyze each neighborhood in Madrid based on the house prices and venues to be able to recommend a suitable neighborhood for customer needs.

## **2.Data**

In this project, ‘Madrid Real Estate Market’ dataset will be used. This dataset contains 21742 houses and can be found here: <https://www.kaggle.com/mirbektoktogaraev/madrid-real-estate-market>



**Image 1: ‘**Madrid Real Estate Market’ dataset overview

The venue data provided by Foursquare. The Places API offers real-time access to Foursquare’s global database of rich venue data. [3]

We’ll be creating a choropleth map to visualize house prices in Madrid. So, geojson files that required to create a choropleth map can be found here: <https://github.com/codeforamerica/click_that_hood/tree/master/public/data>

## **3.Methodology**

### **3.1 Data Preparation**

We’ll start by getting the list of neighborhoods. Wikipedia has a complete list of neighborhoods and districts. Pandas.read\_html method allows us to read tables from Wikipedia. I read that table and converted into a dataframe.

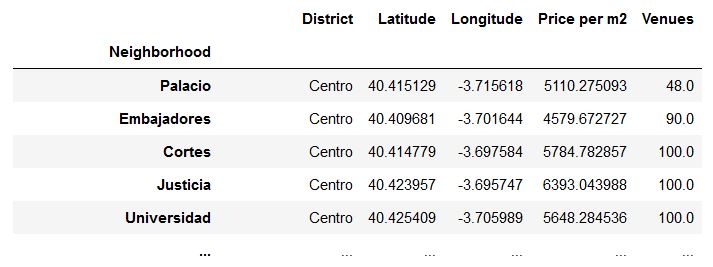
Next step is getting coordinates of neighborhoods. I tried to use Nominatim geocoder from GeoPy package but it returned some incorrect results. Thus, i had to update coordinates manually.

Original house prices dataset consists of 21742 rows and 58 columns. The ‘neighborhood\_id’ column contains neighborhood & district names. But some of the neighborhoods names is incorrect because in 2017 several changes made in administrative subdivions of Madrid. I updated these names by running some SQL queries. For example; the query given below replaces all ‘San Andrés’ (old name) instances with ‘Villaverde Alto’ (current name). I repeated this query for all outdated neighborhood names.



Next, i’ve read this dataset into a Pandas dataframe. Since most of the columns are unnecessary for this project, i select only the needed rows which are ‘neighborhood\_id’ and ‘buy\_price\_by\_area’. After calculating average house prices each neighborhood, the three dataframes (names, coordinates, house prices) merged.

Second part of data preparation is getting venue data for each neighborhood. I used Places API by Foursquare. For all neighborhoods, i send a get request that includes client\_id, client\_secret, latitude and longitude of neighborhood etc.. The API returned 3280 venues from 115 neighborhoods. After grouping by neighborhoods, the number of venues calculated for each neighborhood. Then i merged venue dataframe with main dataframe.

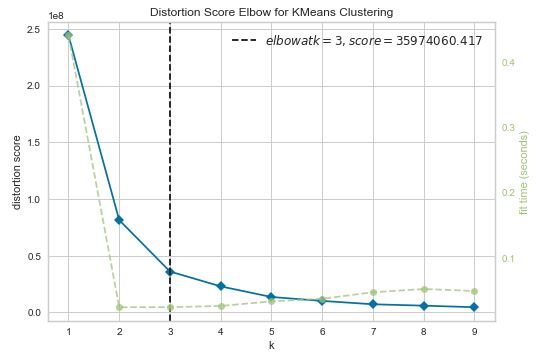


**Image 2:** Main dataframe

## **4. Analyzing Neighborhoods**

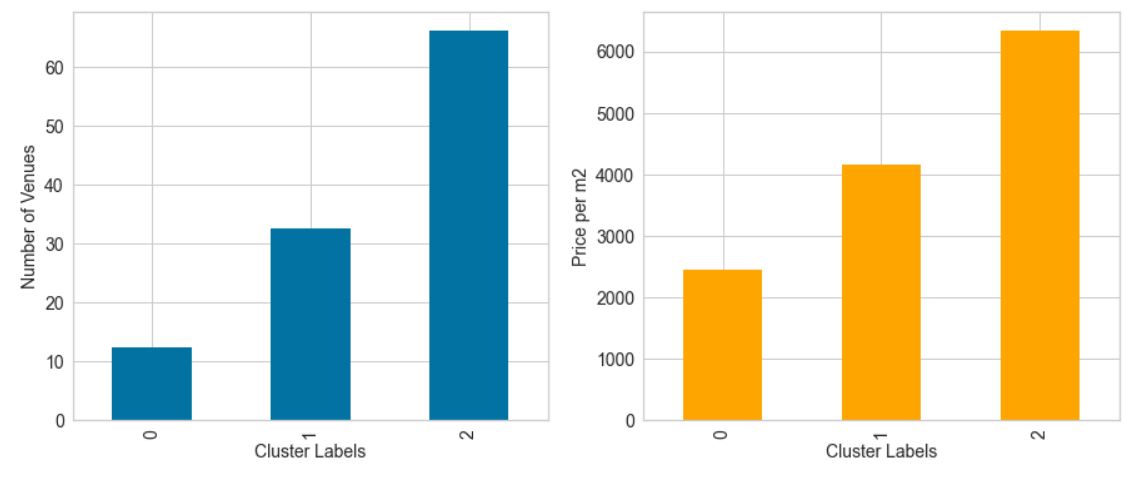
To be able to select a neighborhood among all others, we need to classify the neighborhoods based on their properties. Since goal of this project is selecting a suitable neighborhood for buying a house, we should classify the neighborhoods by house prices and venues.

I will use k-means clustering algorithm to categorize the neighborhoods. First we need to specify k (number of clusters) value. Elbow method provides a simple way to determine the number of clusters. So i used KElbowVisualizer from Yellowbrick library to select optimal *k* value.



**Image 3:** Elbow method indicates that k=3 should be selected

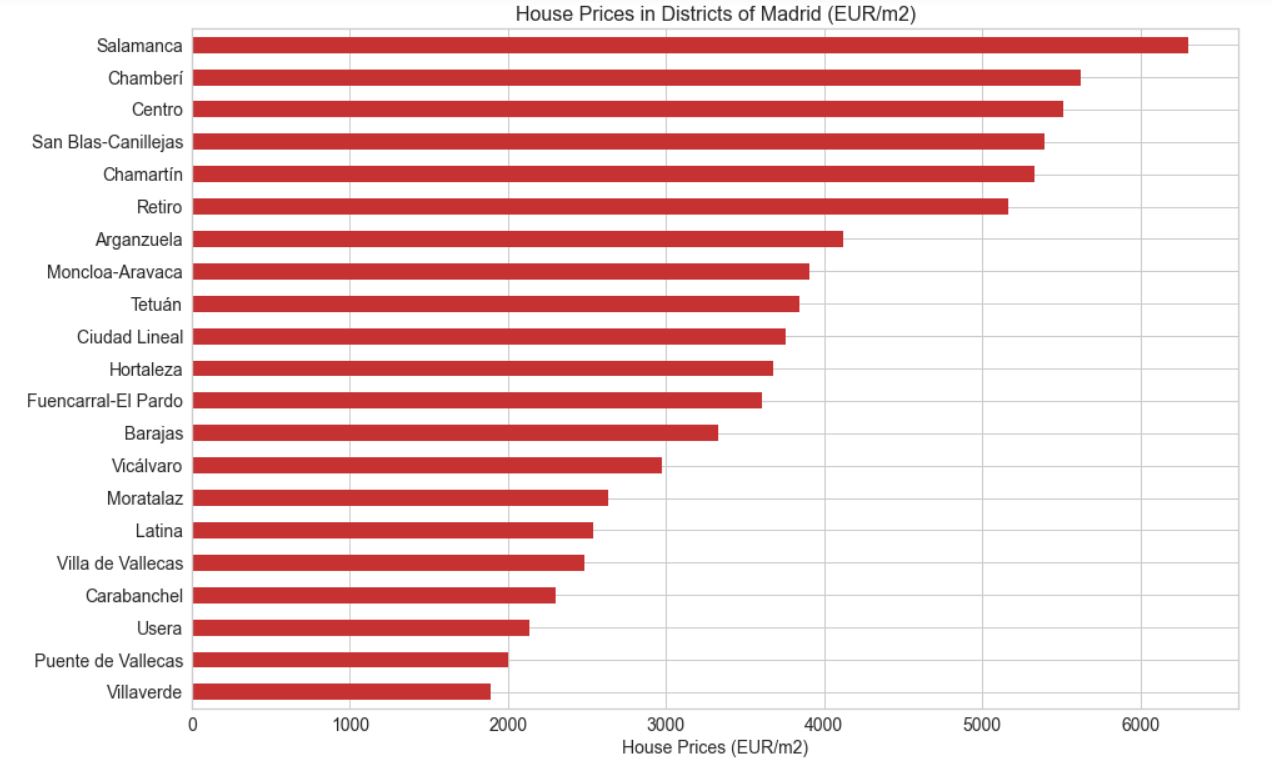
KMeans model from **scikit-learn** package is used for clustering. After the clustering, cluster labels of neighborhoods inserted to main dataframe. Then, i created two bar plots to examine the clusters.



**Image 4:** Analysis of clusters

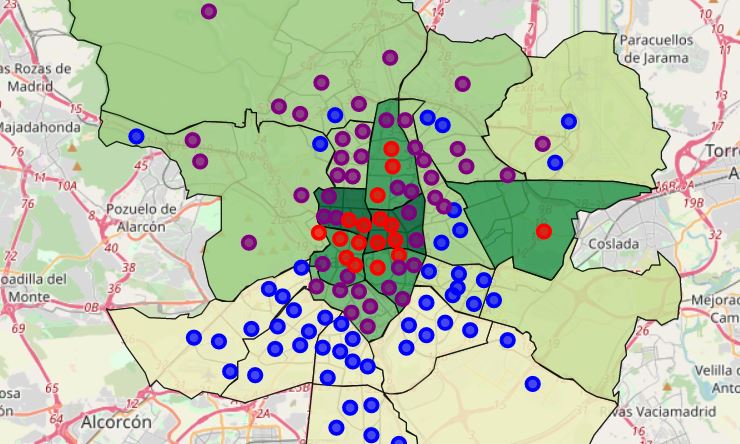
The figures above illustrates that neighborhoods in Cluster 0 has lowest house prices and around 10 venues while neighborhoods in Cluster 2 has the highest prices along with 65 venues average. Based on these plots it's possible to say that expensive neighborhoods also has more venues vice-versa.

Then house prices for each district calculated. The graph given below demonstrates that Salamanca district has the highest house prices while Villaverde district has the lowest.



**Image 5:** House prices in each district

Finally, i created a choropleth map of districts based on house prices. I encountered some problems while trying to create a choropleth map with folium. After some struggling, i noticed that map was not created correctly due to the special characters like é,í etc. in district names. So i replaced those names. I also updated the geojson file to be able to generate a correct map.



**Image 6:** Choropleth map of Madrid

## **5. Discussion**

There are 131 neighborhoods in Madrid while our main dataframe has 115. I had to drop these neighborhoods from dataframe because there was no pricing or venue data about them. To get better results, those neighborhoods should be added. Also some neighborhoods has very little information about either pricing or venues. Data for those neighborhoods also should be improved.

## **6. Conclusion**

Goal of this project was recommending a proper neighborhood to a customer who wants to buy a house in Madrid. Two properties has been used: house prices per m2 and number of venues in given neighborhood. It is possible to examine districts on the map based on house prices. Neighborhoods classified using k-means clustering by their properties :

Cluster 0: Low house prices & number of venues

Cluster 1: Middle house prices & number of venues

Cluster 2: High house prices & number of venues

As a result, customers can select suitable neighborhood for their needs by examining the choropleth map created during this project.

## **7. Resources**

[1] <https://en.wikipedia.org/wiki/Madrid> (accessed on May 29th, 2020)

[2] <https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid> (accessed on May 29th, 2020)

[3] <https://developer.foursquare.com/docs/places-api/> (accessed on May 30th, 2020)

[4] <https://github.com/codeforamerica/click_that_hood/tree/master/public/data>

[5] <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html> (accessed on May 31th, 2020)