

The Battle of neighborhoods: House hunters

Nelida Nkumu Mbomio Ada

May 2020

1. Introduction

1.1. Background

Looking for a place where to live has never been such an arduous task as it is nowadays. Not only we should find a nice apartment with all the facilities we could need according to our budget, it is equally important to live in an area that satisfies our needs.

Well, maybe we know a neighbourhood that fits our needs but it could be quite expensive, so we have to invest a lot of time and effort looking for another similar area where to live. Or maybe we are new in a city and we have no idea about where to start our search.

It may require a huge amount of time to analyze different neighborhoods in a city to determine if it fits or not.

1.2. Problem

Here, we are going to aim the stakeholders identifying neighborhoods that could be interesting for them in a city. How are we going to identify a neighborhood as *interesting*? Well, we have to provide an ideal neighborhood that will be used to look for similar ones into the desired city.

2. Data

2.1. Data extraction

For the purposes of our study, we will be focusing on:

1. Soto del Henares in Torrejón de Ardoz (Madrid, Spain) as our ideal neighborhood
2. Madrid city as the city where we want to live.

The data we will need to carry out this study will be the information about the neighborhoods in Madrid and the information about Torrejón de Ardoz. We will obtain the information about the neighborhoods in Madrid city parsing the information from https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid. There is a table with the districts and the neighborhoods in Madrid city.









District name (number) ⇅	District location ⇅	Number ⇅	Name ⇅	Image ⇅
Centro (1)		11	Palacio	
		12	Embajadores	
		13	Cortes	
		14	Justicia	
		15	Universidad	
		16	Sol	
		21	Imperial	

Figure 1. Source data from Wikipedia

After the use of BeautifulSoup to parser the table, we will obtain the following data.

	District	Neighborhood
0	Centro	Palacio
1	Centro	Embajadores
2	Centro	Cortes

128	Barajas	Casco Histórico de Barajas
129	Barajas	Timón
130	Barajas	Corralejos

Table 1. A view of districts and neighborhoods in Madrid city

The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. Thanks to this API we will be able to create the final dataset like follows.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	...	10th Most Common Venue
--------------	-----------------------	-----------------------	-----------------------	-----	------------------------

Table 2. Header of the dataset obtained after processing Foursquare output

The table above -excluding *Neighborhood* column- represents our final dataset. Our features will be the *nth Common venue*.

3. Methodology

3.1. Data preparation

Once we have the list of neighborhoods and districts, we will be using a geolocator to obtain the coordinates (latitude and longitude) for each neighborhood and include it in the dataframe. Latitude and longitude are needed in order to draw the neighborhoods in a map and get the venues in those areas.

	District	Neighborhood	Latitude	Longitude
0	Centro	Palacio	40.4151	-3.71562
1	Centro	Embajadores	40.4097	-3.70164
2	Centro	Cortes	40.4148	-3.69758
...
128	Barajas	Casco Histórico de Barajas		
129	Barajas	Timón	40.4736	-3.58217
130	Barajas	Corralejos	40.4682	-3.58707

Table 3. Neighborhoods dataset including latitude and longitude

In some cases, geolocator is not going to be able to retrieve the coordinates for a given neighborhood. Unfortunately, we will remove the data corresponding to neighborhoods without coordinates as we can see in record number 128 for Casco Histórico de Barajas, where no coordinates were retrieved.

Additionally, we have to obtain the longitude and latitude for Soto del Henares and include it as part of the neighborhoods to cluster.

	District	Neighborhood	Latitude	Longitude
0	Centro	Palacio	40.4151	-3.71562
1	Centro	Embajadores	40.4097	-3.70164
2	Centro	Cortes	40.4148	-3.69758
...
125	Barajas	Aeropuerto	40.4948	-3.57408
126	Barajas	Timón	40.4736	-3.58217
127	Barajas	Corralejos	40.4682	-3.58707
128	Torrejón de Ardoz	Soto del Henares	40.4605	-3.43996

Table 4. Filtered dataset including reference neighborhood

The number of neighborhoods we are going to analyze is 129, including Soto del Henares.

Now that we have the longitudes and latitudes, we are going to retrieve the list of venues in a radius of 500 metres for each neighborhood by using Foursquare API. We have obtained a list of 3480 venues in total with a total of 256 unique categories.

There were some neighborhoods where we could not retrieve any venue. Those neighborhoods will be removed from our study.

Neighborhood	N. Latitude	N. Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Palacio	40.415129	-3.715618	Santa Iglesia Catedral de Santa María la Real ...	40.415767	-3.714516	Church
Palacio	40.415129	-3.715618	Plaza de La Almudena	40.416320	-3.713777	Plaza
...

Table 5. List of venues retrieved

Finally, with all these data, we transform the data to get the list of top ten venues categories per neighborhood.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue		10th Most Common Venue
Abrantes	Restaurant	Soccer Field	Bakery	...	Fountain
Acacias	Bar	Park	Art Gallery	...	Hotel
Adelfas	Supermarket	Tapas Restaurant	Spanish Restaurant	...	Metro Station
Alameda de Osuna	Smoke Shop	Bakery	Hotel	...	Metro Station
Almagro	Spanish Restaurant	Restaurant	Hotel	...	Cocktail Bar
...

Table 6. List of neighborhoods most common venues

3.2. Clustering neighborhoods

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

We need to create different clusters in order to group those neighborhoods that may be similar.

k -means is a clustering algorithm that aims to partition n observations into k clusters in which each

observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

After running k-means with values for k in the range $[1, 50]$, we have to decide which value of k is most appropriate for this scenario.

```
In [35]: Ks = 50
final_cluster_size = np.zeros((Ks))
city_grouped_clustering = madrid_grouped.drop('Neighborhood', 1)
for n in range(1,Ks+1):

    #Train Model and Predict
    kmeans = KMeans(n_clusters=n, random_state=0).fit(city_grouped_clustering)
    soto_cluster = kmeans.labels_[110]
    final_cluster_size[n-1] = np.count_nonzero(kmeans.labels_==soto_cluster)
final_cluster_size

Out[35]: array([[126., 122., 109., 111., 99., 88., 64., 31., 72., 93., 40.,
 51., 47., 61., 20., 56., 31., 35., 51., 15., 23., 47.,
 27., 1., 52., 21., 59., 35., 59., 25., 52., 44., 9.,
 19., 6., 37., 20., 26., 1., 1., 1., 36., 23., 1.,
 1., 1., 1., 39., 1., 1.] )
```

Figure 2. Extract of code to evaluate k values in k -means algorithm

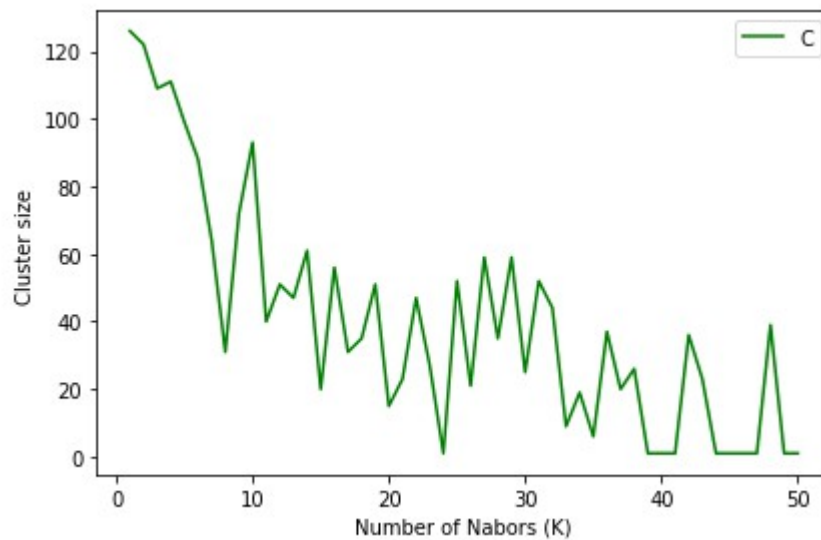


Figure 3. Plot of k values versus Cluster size

Due to the nature of the problematic and the final stakeholders, we should not consider those k that results in a huge cluster size for Soto del Henares; if the cluster is too big, the stakeholder will not be able to analyze all those neighborhoods and the problem will still be same. Analogously, we should not consider those clusters where the only neighborhood is Soto del Henares.

Finally, we decided to take $k = 35$, with a cluster size of 6 neighborhoods (including Soto del Henares).

4. Results

Figure 4 shows the view of the final map where all the neighborhoods have been clustered.

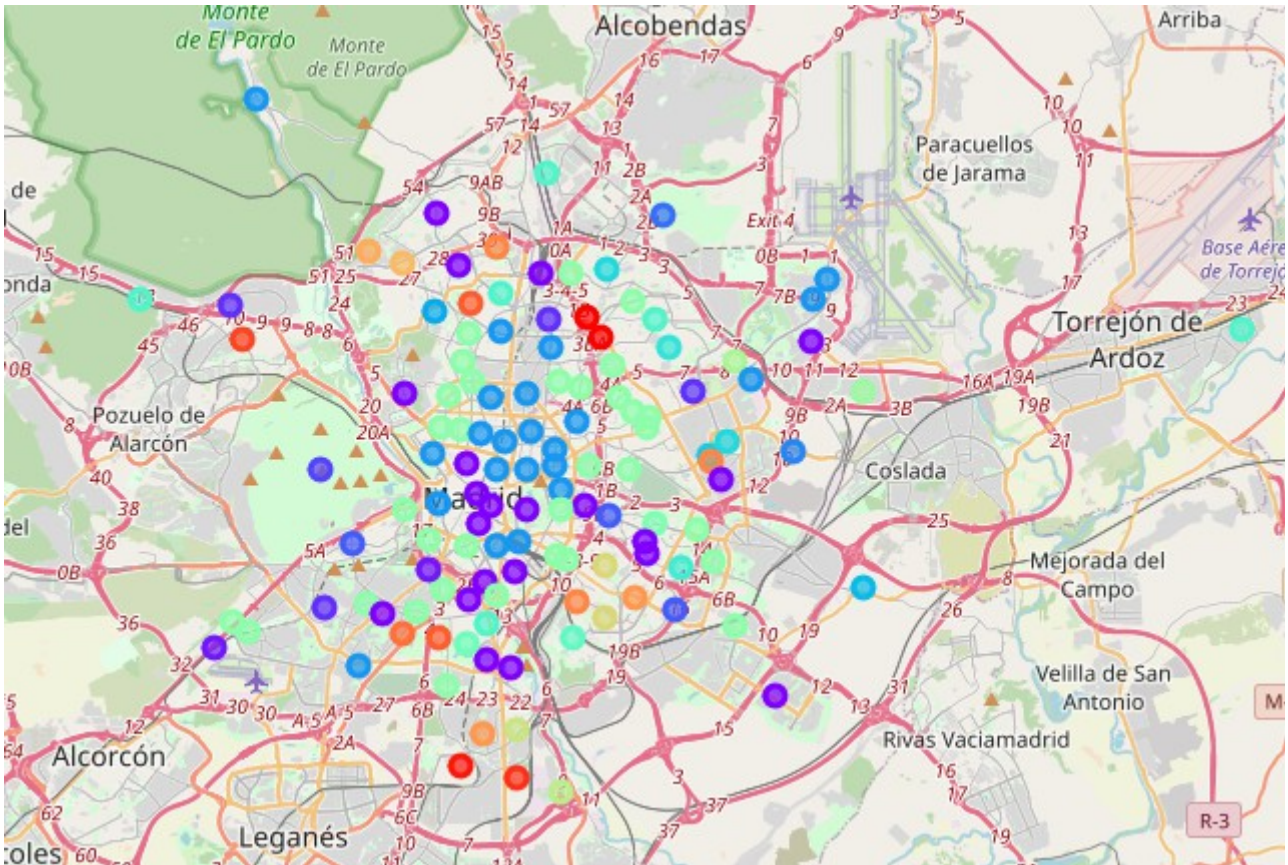


Figure 4. Clustered map including Soto del Henares in Torrejón de Ardoz (right side)

Drawing the cluster containing Soto del Henares, we can see the neighborhoods that our analysis consider similar to Soto del Henares.

We have Almenara, Valverde, El Plantío, Almendrales and Canillas.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	...	10th Most Common Venue
Almenara	Spanish Restaurant	BBQ Joint	Department Store	...	Gym / Fitness Center
Valverde	Italian Restaurant	Train Station	Breakfast Spot	...	Historic Site
El Plantío	Italian Restaurant	Spanish Restaurant	Asian Restaurant	...	Gym
Almendrales	Spanish Restaurant	Bar	Seafood Restaurant	...	Noodle House
Canillas	Spanish Restaurant	Italian Restaurant	Juice Bar	...	Sandwich Place
Soto del Henares	Gastropub	Train Station	Cafeteria	...	Furniture / Home Store

Table 7. Neighborhoods in Soto del Henares cluster

5. Discussion

Here we have finally analyzed a total of 126 neighborhoods in Madrid city. The size of our sample may be too big, and we could initially filter the neighborhoods by some criteria as close to an specific location – maybe the stakeholder workplace - or a crèche, school or university.

Even though this process allow us to find neighborhoods similar to a reference one, we have to keep in mind that this will not be a garanty of finding an apartment in this areas.

We should notice that even the area is supposed to be similar, prices may be not and the budget is always a very important issue when looking for an apartment. In this case, as a future work, we could cross our neighborhoods data with other databases containing some statistics about house prices as average price by squared meter.

6. Conclusions

In this study we have analyzed different neighborhoods in Madrid city and clustered them along with a reference neighborhood in order to get similar areas to the reference one and facilitate the process of search for a new apartment. We have used BeautifulSoup and Foursquare API to extract the data to be analyzed. Finally, running a k-means algorithm over the data, we were able to cluster all the neighborhoods (using the optimal value of k according to our problematic) and extracted. This procedure could help us not just in Madrid city center as desired city and Soto del Henares as reference neighborhood, looking for the appropriate data sources we could run this process with other cities and neighborhoods.