ANALYSIS OF RESIDENTIAL NEIGHBORHOODS IN MADRID, SPAIN

Capstone Project: The battle of Neighborhoods

Madrid Downtown – Getty images

Author: Raúl Castellanos

1. INTRODUCTION

As a part of Coursera IBM Data Science Professional Certificate Capstone Project, we were asked to elaborate an analysis of a city and its neighborhoods.

A common problem nowdays is the uprising prices of houses and cost of living. This is more problematic in the big cities as are the most crowded places so I find pretty interesenting for every family that wants to move to a new city to be able to know which are the best neighborhoods in their new destiny. As I am from Madrid, Spain I decided to elaborate this kind of research in this city and retrieve a list of its neighborhoods, find their geographical coordinates and use the coordinates as input of the Foursquare API, which we have used previously in the course, to obtain the top venue categories in each neighborhood. Using the frequency of venue categories we can use the k-means algorithm to cluster neighborhoods of similar venue categories and identify which ones are residential neighborhood, so a family can decide where to move to live in this city.

In addition to venue categories, I introduced data from the public schools of the city in each neighborhood in order to check if the information obtained from the cluster analysis is correct, and there is a great number of schools in those neighborhoods as it would be expected from a residential neighborhood.

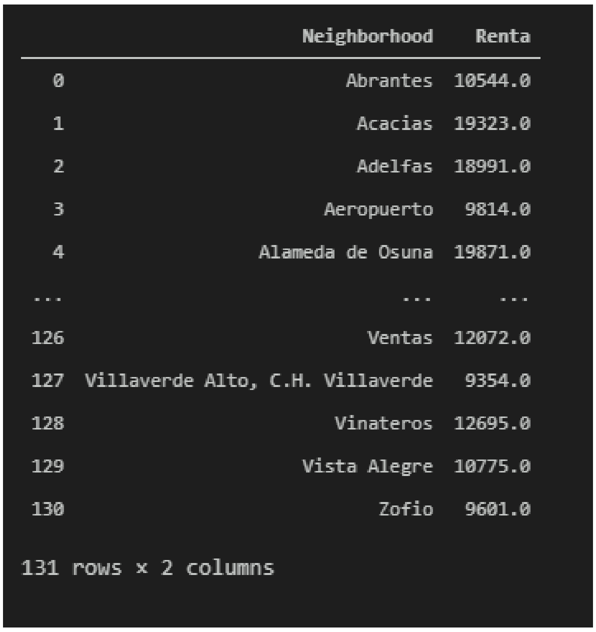
As a result we will be able to identify which are the most suitable parts of the city to live avoiding those neighborhoods with less facilities for a new family.

2. DATA

In this section I introduce the datasets that will be used and their sources.

Neighborhoods information

My initial dataset is a csv downloaded from the Madrid’s public wepage where we can find the name of every neighborhood and its Per Capita Income.



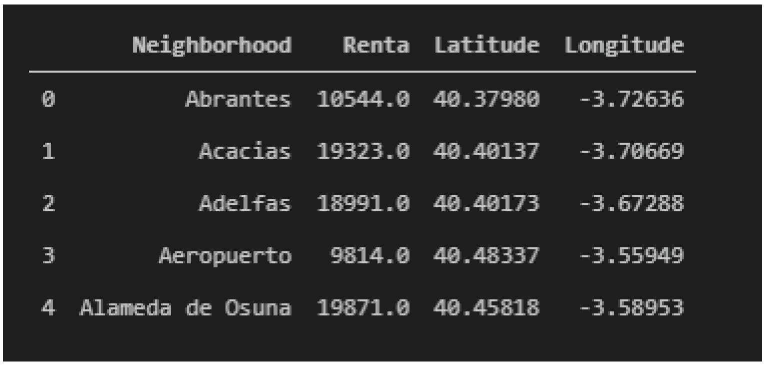
	Neighborhood	Renta
0	Abrantes	10544.0
1	Acacias	19323.0
2	Adelfas	18991.0
3	Aeropuerto	9814.0
4	Alameda de Osuna	19871.0
...	...	...
126	Ventas	12072.0
127	Villaverde Alto, C.H. Villaverde	9354.0
128	Vinateros	12695.0
129	Vista Alegre	10775.0
130	Zofio	9601.0

131 rows × 2 columns

Figure 1: Per Capita Income and neighborhoods name in Madrid.

Geographical coordinates

The information of every neighborhood will be enriched with geographical coordinates using the Geopy library from which I will add its latitude and longitude.

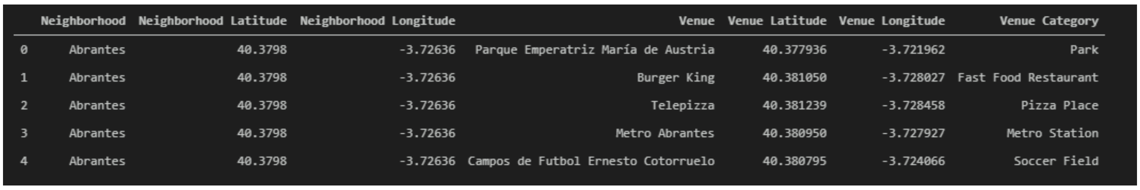


	Neighborhood	Renta	Latitude	Longitude
0	Abrantes	10544.0	40.37980	-3.72636
1	Acacias	19323.0	40.40137	-3.70669
2	Adelfas	18991.0	40.40173	-3.67288
3	Aeropuerto	9814.0	40.48337	-3.55949
4	Alameda de Osuna	19871.0	40.45818	-3.58953

Figure 2: Neighborhoods in Madrid with latitude and longitude information.

Venue categories

Next, I will use the Foursquare API, using the latitude and logitude included from the Geopy for every neighborhood to retrieve venues in a given radius around each location.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abrantes	40.3798	-3.72636	Parque Emperatriz María de Austria	40.377936	-3.721962	Park
1	Abrantes	40.3798	-3.72636	Burger King	40.381050	-3.728027	Fast Food Restaurant
2	Abrantes	40.3798	-3.72636	Telepizza	40.381239	-3.728458	Pizza Place
3	Abrantes	40.3798	-3.72636	Metro Abrantes	40.380950	-3.727927	Metro Station
4	Abrantes	40.3798	-3.72636	Campos de Futbol Ernesto Cotorruelo	40.380795	-3.724066	Soccer Field

Figure 3: Sample Venue Categories returne by Foursquare API per neighborhood.

Public schools

Finally, information of amount of public schools in every neighborhood is added from the public dataset from https://datos.madrid.es/portal/site/egob/

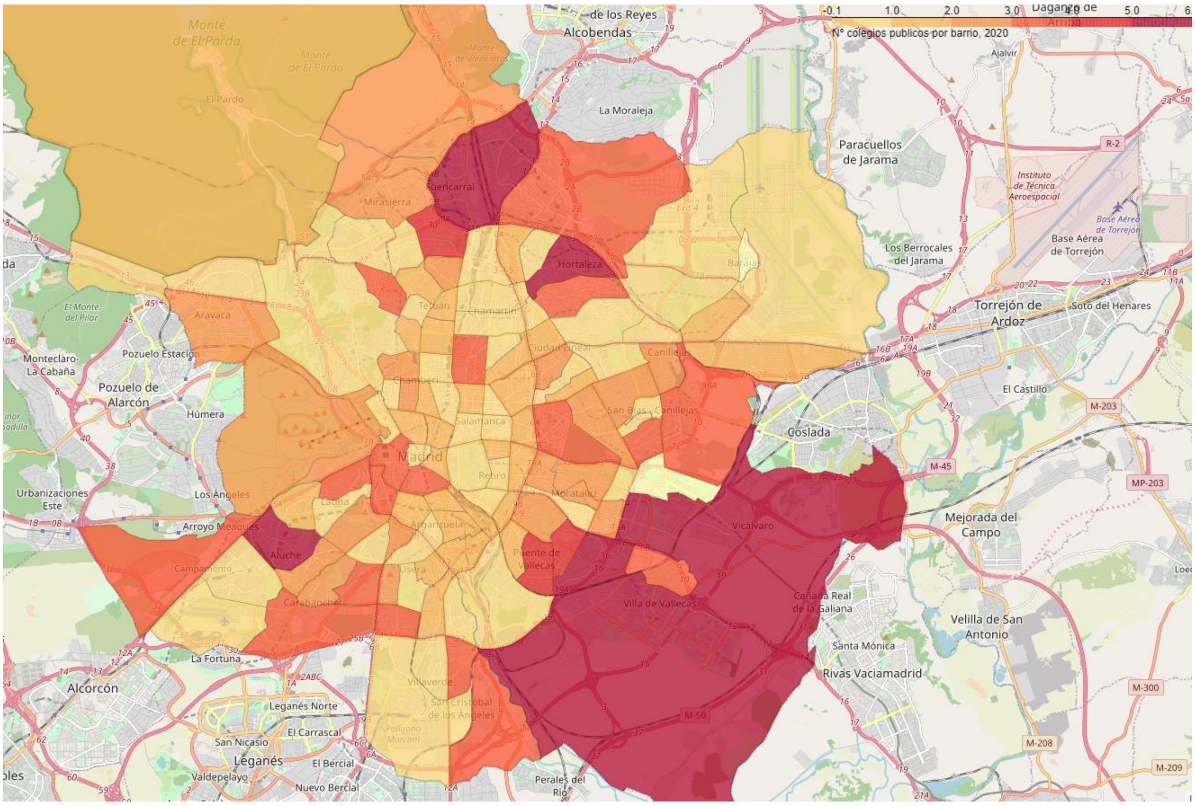
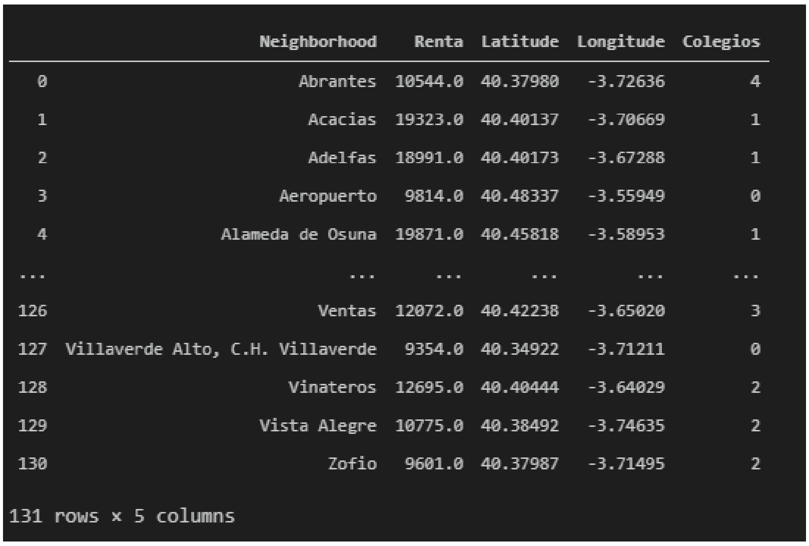


Figure 4: Number of schools per neighborhoods in Madrid.

3. METHODOLOGY

First I collected all my data and stored it into dataframes. I started creating a data frame with the names and Per Income data and i used that dataframe to get the coordinates from Geopy. Once I got the geographical coordinates I added them to the previous dataframe like you can see in figure 2.

Next, I introduced the number of schools in each neighborhood in a different dataframe and I joined it into the previous one to have all the mentioned information in a single dataset.

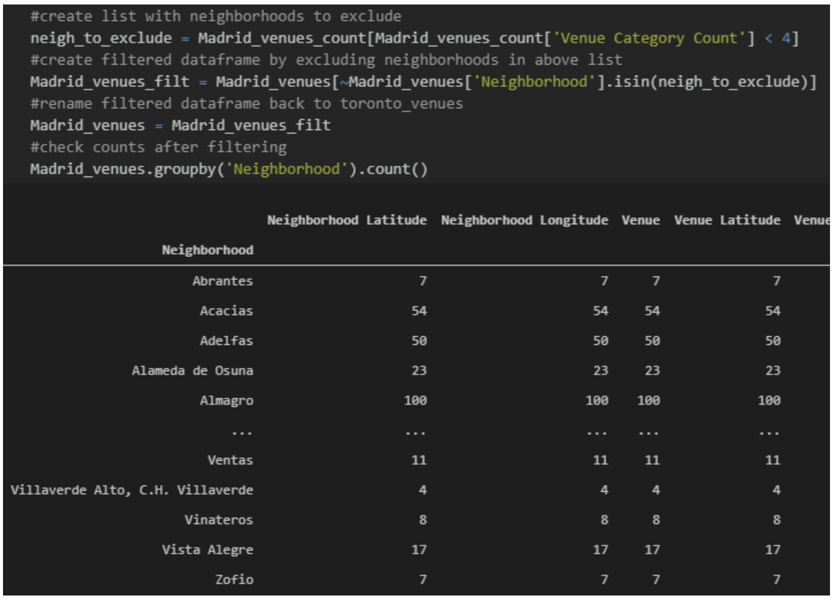


```
                              Neighborhood    Renta  Latitude  Longitude  Colegios
  0                               Abrantes  10544.0  40.37980   -3.72636         4
  1                                Acacias  19323.0  40.40137   -3.70669         1
  2                                Adelfas  18991.0  40.40173   -3.67288         1
  3                             Aeropuerto   9814.0  40.48337   -3.55949         0
  4                        Alameda de Osuna  19871.0  40.45818   -3.58953         1
...                                    ...      ...       ...        ...       ...
126                                 Ventas  12072.0  40.42238   -3.65020         3
127   Villaverde Alto, C.H. Villaverde       9354.0  40.34922   -3.71211         0
128                              Vinateros  12695.0  40.40444   -3.64029         2
129                            Vista Alegre  10775.0  40.38492   -3.74635         2
130                                  Zofio   9601.0  40.37987   -3.71495         2

131 rows × 5 columns
```

Figure 5: Final dataframe with all information used.

Once I had it, I used the Foursquare API to explore the neighborhoods. I pased the geographical coordinates of every neighborhood to the API and it returned a list of venues in a given radius of 500m and a limit of máximum 200 venues per neighborhood. The resulting dataset becase a list of all neighborhood by name with added venues and venue categories. To have a better analysis of the neighborhoods I decided to exclude those neighborhoods that after researching in the Foursquare API only had less than 4 venues in its data.



```
#create list with neighborhoods to exclude
neigh_to_exclude = Madrid_venues_count[Madrid_venues_count['Venue Category Count'] < 4]
#create filtered dataframe by excluding neighborhoods in above list
Madrid_venues_filt = Madrid_venues[~Madrid_venues['Neighborhood'].isin(neigh_to_exclude)]
#rename filtered dataframe back to toronto_venues
Madrid_venues = Madrid_venues_filt
#check counts after filtering
Madrid_venues.groupby('Neighborhood').count()
```

```
                                  Neighborhood Latitude  Neighborhood Longitude  Venue  Venue Latitude  Venue
                    Neighborhood
                        Abrantes                      7                       7      7               7
                         Acacias                     54                      54     54              54
                         Adelfas                     50                      50     50              50
                Alameda de Osuna                     23                      23     23              23
                         Almagro                    100                     100    100             100
                             ...                    ...                     ...    ...             ...
                          Ventas                     11                      11     11              11
Villaverde Alto, C.H. Villaverde                      4                       4      4               4
                       Vinateros                      8                       8      8               8
                    Vista Alegre                     17                      17     17              17
                           Zofio                      7                       7      7               7
```

Figure 6: Number of values per neighborhood in Madrid.

Finally I ran the k-means clustering algorithm on the above dataframe to derive clusters of neighborhoods by postal code using 5 as the number of clusters.

4. RESULTS

First, I visualized the resulting clusters on a map as below:

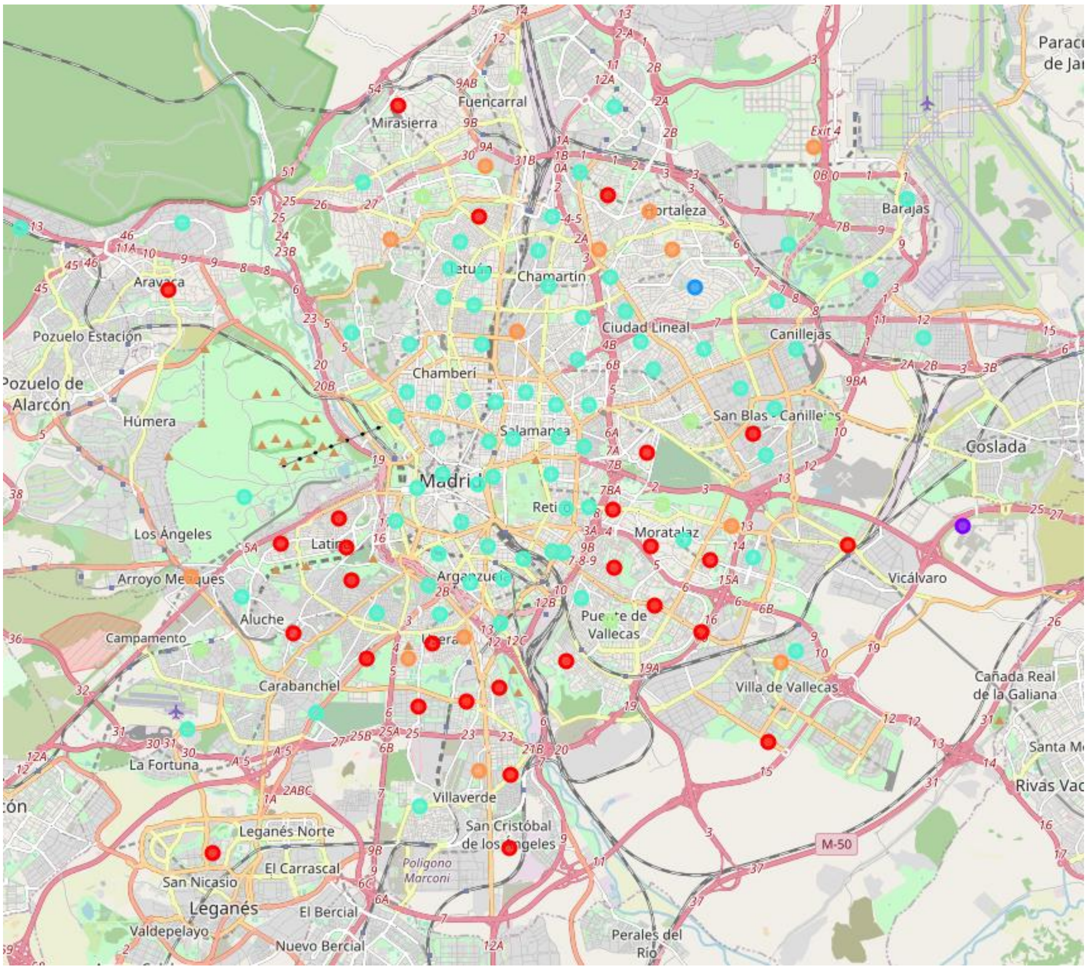


Figure 7: Resulting clusters displayed on a map of Madrid.

The different colored dots represent clusters:

- 🔴 Cluster 0
- 🟣 Cluster 1
- 🔵 Cluster 2
- 🟠 Cluster 3
- 🟢 Cluster 4

Then I explored each cluster to determine the common venue categories that define each cluster and I named the clusters accordingly.

Cluster 0: Residential

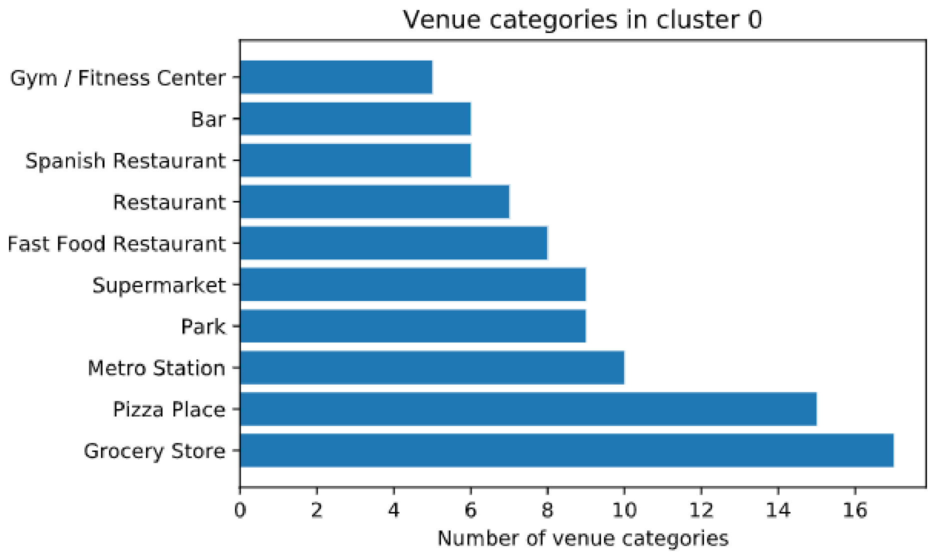


Figure 8: Venues categories of cluster 0.

The amount of grocery stores, Metro stations, parks and supermarkets indicates that it are probably mainly residential areas.

Cluster 1: Shopping centre

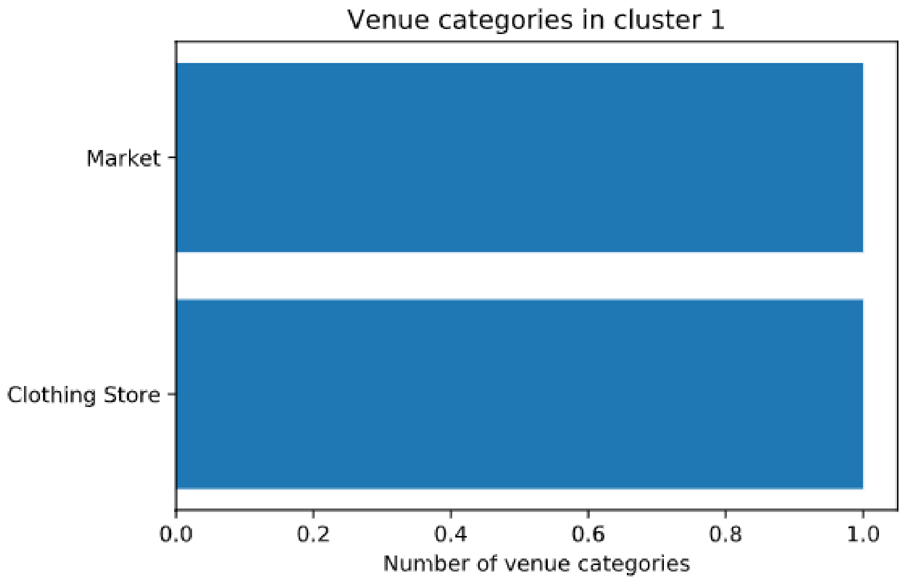


Figure 9: Venues categories of cluster 1.

This is just a neighborhood where there must be a shopping centre.

Cluster 2: Downtown

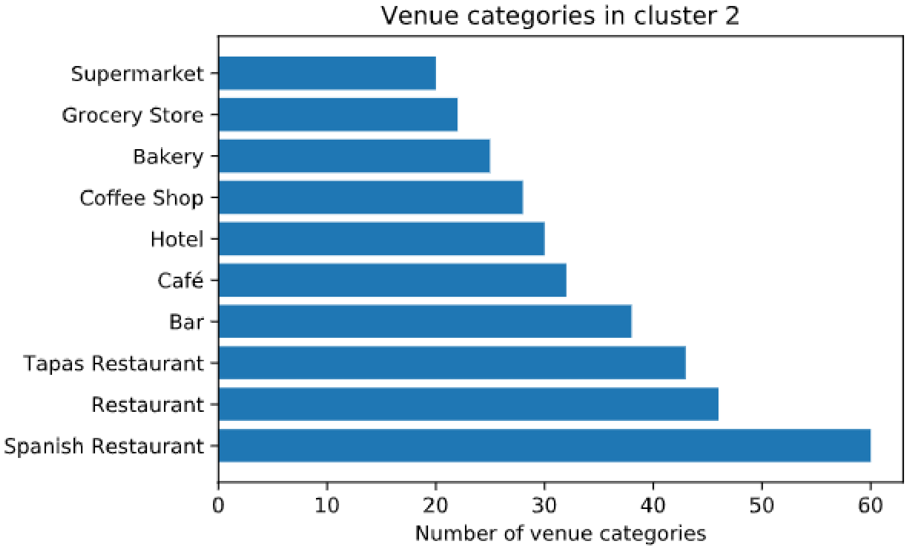


Figure 10: Venues categories of cluster 2.

This cluster has mostly high end restaurants and hotels with a couple of coffe shop which suggests it is made up of downtown neighborhoods.

Cluster 3: Sport lovers residential area

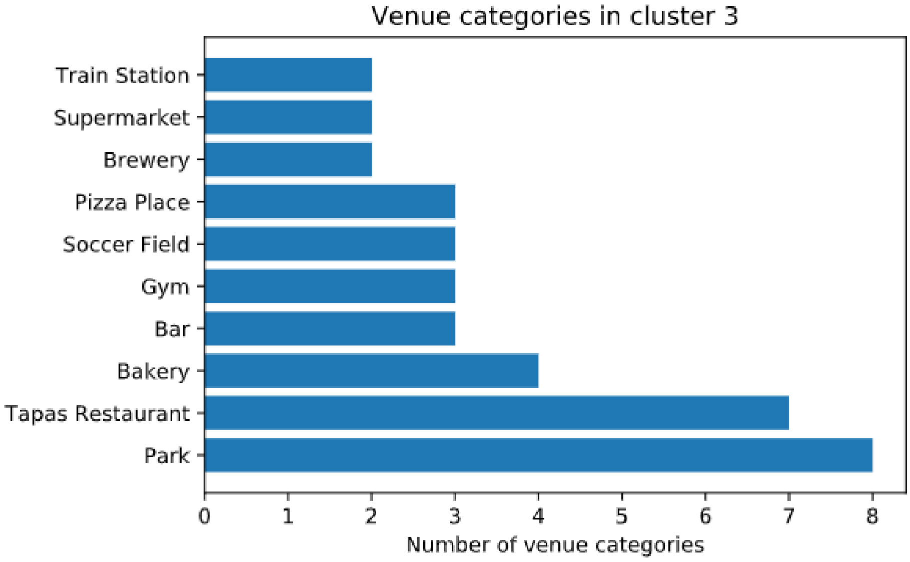


Figure 11: Venues categories of cluster 3.

Similar to the first residential neighborhoods where there are a lot of parks and supermarket but also gym and even soccer fields suggesting a newer residential neighborhood.

Cluster 4: Food lovers residential area

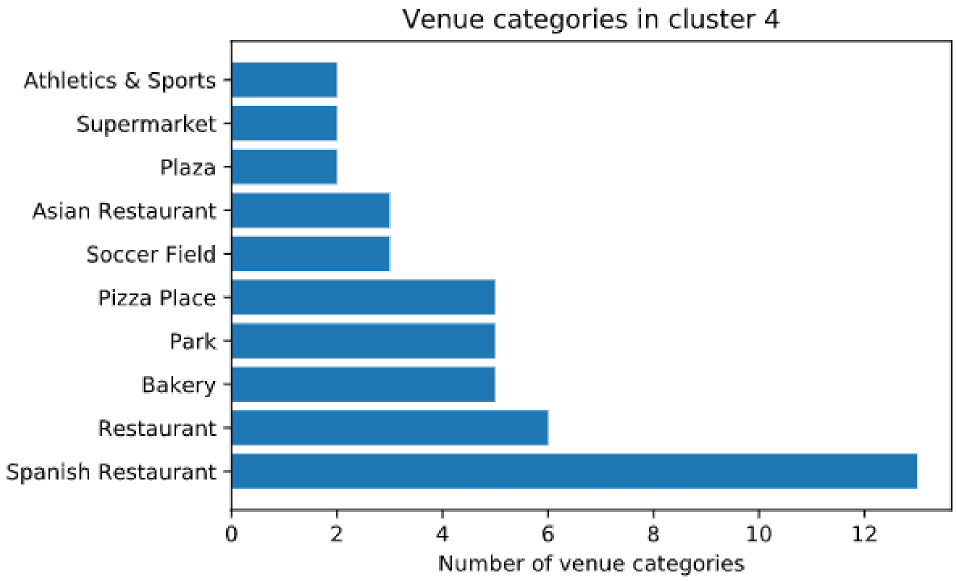


Figure 12: Venues categories of cluster 4.

Pretty similar to the cluster 3 with tipical residential services like supermarkets, parks but with a higher amount of restaurants.

5. DISCUSSION

Taking into account these results we can guess which are the most suitable clusters to be residential:

▨ ▨ ▨ Clusters 0, 3 and 4: These clusters have those kind of venue categories that are useful for a familiar life like supermarkets, parks, sports centers and some restaurants.

▨ Cluster 2: This cluster seems to reperesent all the downtown of the city which is the most crowded zone and its venues are useful for tourist like hotels and plenty different high end restaurants

If we plot these clusters over a map with the information of public schools we can see that as we expected clusters 0, 3 and 4 included most of the neighborhoods with more public schools as would be expected from residential neighborhoods.

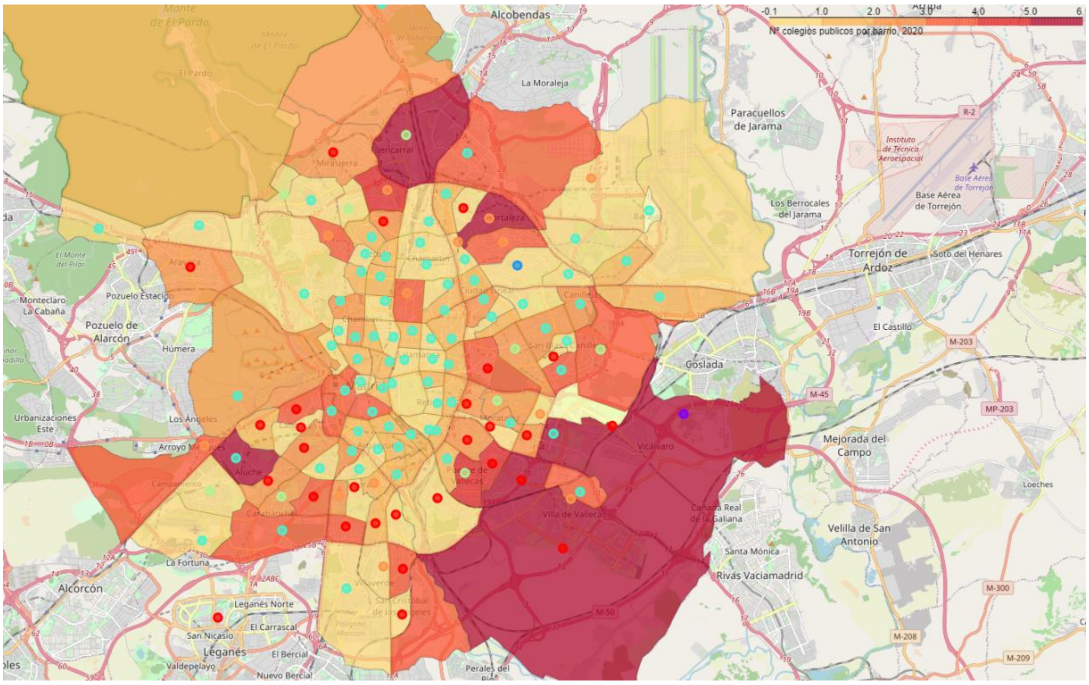


Figure 13: Resulting clusters and number of schools displayed on a map of Madrid

6. CONCLUSIONS

Our research have proved to be pretty useful at identifying the type of neighborhoods of this city and any family could choose a neighborhood from the clusters 0,3 and 4 as a place to live. Moreover, from the individual analysis from every cluster the family could choose among those 3 clusters depending if they prefer a more classical residential neighborhood (cluster 0 🔴) a more sport-like residential neighborhood (cluster 3 🟠) or a food-lover residential neighborhood (cluster 4 🟢)