ANALYSIS OF VENUES CATEGORIES IN THE NEIGHBORHOODS OF MADRID

DATA

Title	Analysis of venues categories in the neighborhoods of Madrid	
Reference	N/A	
Revision	0	
Location	Madrid (Spain)	
Date	April 11, 2020	
Petitioner	N/A	
Author	Name	Daniel Sánchez Valenciano
	ID	N/A
	Qualification	BS in Telecommunications Engineering
	Official Member's N.	N/A

CONTENTS

1. INTRODUCTION/BUSINESS PROBLEM

1.1. Business problem

How are venues distributed among the neighborhoods in Madrid? Where are the neighborhoods in which a certain venue category is specially usual? Which is the most common venue category in each neighborhood? Is there any "outlier" neighborhood in Madrid? Could this information be leveraged in any way? **All these questions define the business problem** around which this data science project revolves.

1.2. Stakeholders

Madrid is a city brimming with culture, leisure, catering and hostelry, and so it attracts many entrepreneurs and investors from these fields. **These are the stakeholders in this business problem**. To boost the chances of success of their enterprises and investments, many of them rely on market researches which try to answer questions such as:

- If a business of a specific type is to be put into operation, where are the neighborhoods in Madrid which should be targeted? How are they distributed in the city? In the suburbs, in the down town, in the old city?

- If a the stakeholder plans to put a business of any type into operation in a specific neighborhood of Madrid, what business types should they consider? Is this neighborhood saturated with a particular type of business? Does this neighborhood lack any type of business? What is the market niche?

- Is there any "special" neighborhood in Madrid regarding its venues? Does it offer any special market opportunity to the entrepreneurs and investors?

- Finally, considering that venues open, close or simply change overnight, would it be possible to have this information dynamically updated?

This data science project aims to answer all of these questions for the stakeholders.

2. DATA

The following data is used:

- **Madrid boroughs and neighborhoods data**. They are freely available on Madrid City Council Open Data website [1]. They come in the form of geospatial vector data files of Madrid boroughs and neighborhoods. Once downloaded, these files are loaded in QGIS, a free and open-source cross-platform desktop geographic information system, in order to be processed. See Illustration 1.

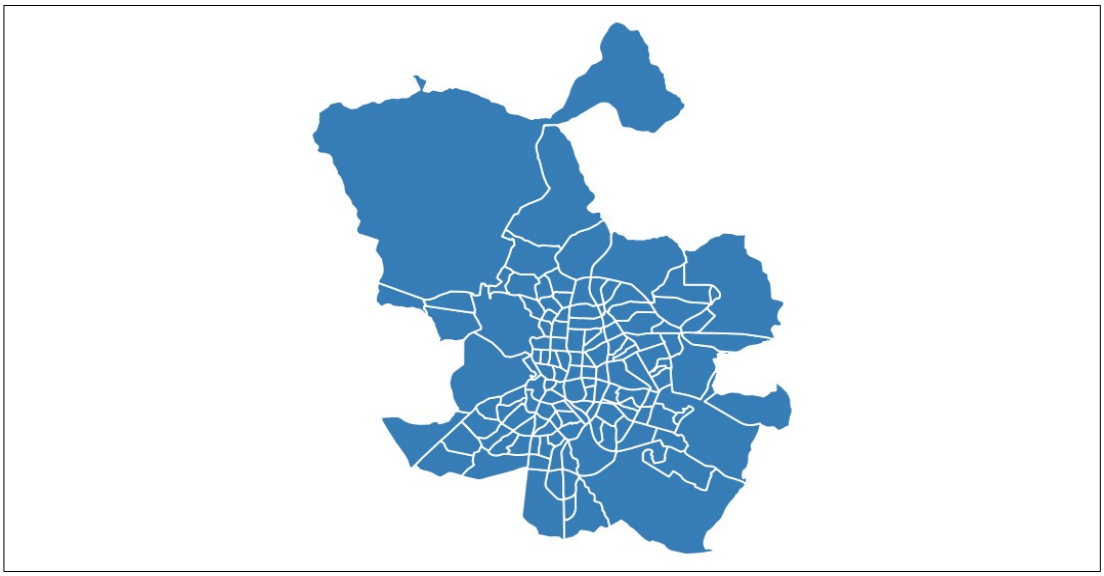

Illustration 1: Madrid boroughs and neighborhoods geospatial vector data files loaded in QGIS

The files are processed leveraging QGIS functionalities so that the corresponding attributes table contains the following information for each neighborhood of Madrid.

- Borough name.
- Neighborhood name.
- Area.
- Perimeter.
- X coordinate (ETRS89/UTM zone 30) of its centroid.
- Y coordinate (ETRS89/UTM zone 30) of its centroid.

See Illustration 2.

1https://datos.madrid.es/egob/catalogo/200078-10-distritos-barrios.zip

	NOMDIS	NOMBRE	area	perimeter	xcoord	ycoord
1	Arganzuela	Legazpi	1414470,6...	5141,6408...	441682,93...	4471151,4...
2	Arganzuela	Chopera	567787,01...	3203,4086...	440672,25...	4471808,8...
3	Arganzuela	Delicias	1054678,7...	4818,0027...	441453,46...	4472027,9...
4	Arganzuela	Acacias	1073437,8...	3950,3269...	439982,85...	4472513,4...
5	Arganzuela	Palos de M...	648442,55...	3647,7465...	441057,64...	4472774,2...
6	Arganzuela	Atocha	735680,59...	4957,6874...	442131,23...	4472353,0...
7	Arganzuela	Imperial	967678,64...	4557,9379...	439045,61...	4473089,6...
8	Barajas	Alameda d...	1970334,8...	6044,9403...	449852,67...	4478600,0...
9	Barajas	Corralejos	4682537,7...	9726,9510...	448524,15...	4479459,2...
10	Barajas	Casco Hist...	549393,99...	3388,6462...	450931,36...	4480530,1...
11	Barajas	Timón	5094465,9...	11840,044...	448908,39...	4480939,2...
12	Barajas	Aeropuerto	29626079,...	28744,094...	452241,98...	4481026,8...

Illustration 2: Madrid boroughs and neighborhoods geospatial vector data files final attributes table

The final attributes table is exported as a comma separated values file ("madrid_neighborhoods.csv"). Then, it is loaded into the Watson Studio project as an asset. See Illustration 3. Finally, this file is read from a Jupyter Notebook using methods from project_lib and pandas libraries.

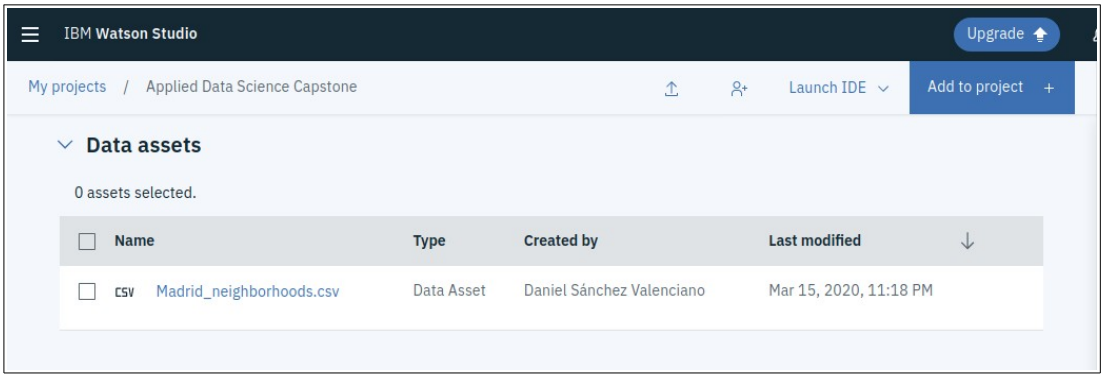

Illustration 3. "madrid_neighborhoods.csv" file loaded into the Watson Studio project as an asset

- **Foursquare data**.

Foursquare API is leveraged to explore venues in each neighborhood in Madrid. See Illustration 4.

	Name	Category	Latitude	Longitude
0	La Gelateria di Angelo	Ice Cream Shop	40.397951	-3.707739
1	Parque de la Arganzuela	Park	40.398330	-3.708686
2	Restaurante Peruano Mis Tradiciones	Peruvian Restaurant	40.399816	-3.711022
3	sushi raku	Sushi Restaurant	40.404623	-3.708216
4	Le Crust Pizza Bar	Pizza Place	40.400922	-3.709890

Illustration 4. Example of venues data from “Acacias” neighborhood retrieved from Foursquare

3. METHODOLOGY

The data of the neighborhoods in Madrid and the data of the venues are inspected, pre processed and merged into a final dataframe. It contains the one-hot encoded venues categories data grouped by neighborhood. I. e., **this final dataframe shows how many venues of each category are in each neighborhood in Madrid**.

In the analysis stage, **K-Means clustering algorithm is run against the aforementioned final dataframe in order to cluster the neighborhoods in Madrid according to the categories of their venues**. The scikit-learn library is leveraged to do so.

Then, **the resulting clusters are examined one by one to determine their distinctive features**, e. g., if they correspond to neighborhoods where the most common venues categories are clearly theaters and museums or whether, on the contrary, where the most common venues categories are supermarkets and groceries, and so on. The pandas library is used to add the clusters labels back to the dataframe and to process it to sum up the characteristics of each cluster.

Finally, **the neighborhoods in Madrid are displayed in a map and their icons are colored according to their clusters. This way, the distribution of the clusters in Madrid can be easily visualized**. At this point, it is possible to find out patterns in the distribution, "outlier" neighborhoods in Madrid, the areas of Madrid where it is common to find a specific business category, the most common business categories in a particular area of Madrid, and so on. The folium library is used to create the maps.

4. RESULTS

4.1. First approach

The first clustering produces very uneven clusters: 96.9% of the neighborhoods belong to only 1 cluster, while the remaining 3.1% of the neighborhoods are distributed among the rest of the clusters.

This main cluster is characterized by a clear predominance of restaurants (specially Spanish ones). See Illustration 5. As a result, the map is almost colored in only one color, what makes it impossible to find any pattern in the distribution or extract any further information. See Illustration 6. This is attributed to the fact that **restaurants seem to be ubiquitous in Madrid**.

```
************************************************** CLUSTER 1 **************************************************

There are 125 neighborhoods labelled as Cluster 1 (96.9 %).

There are 4826 venues in the neighborhoods labelled as Cluster 1.

                    Count  Percentage
Spanish Restaurant    478         9.9
Restaurant            332         6.9
Bar                   209         4.3
Tapas Restaurant      173         3.6
Supermarket           136         2.8
Hotel                 135         2.8
Coffee Shop           125         2.6
Italian Restaurant    125         2.6
Café                  124         2.6
Bakery                124         2.6
```

Illustration 5: main cluster characteristics (first approach)

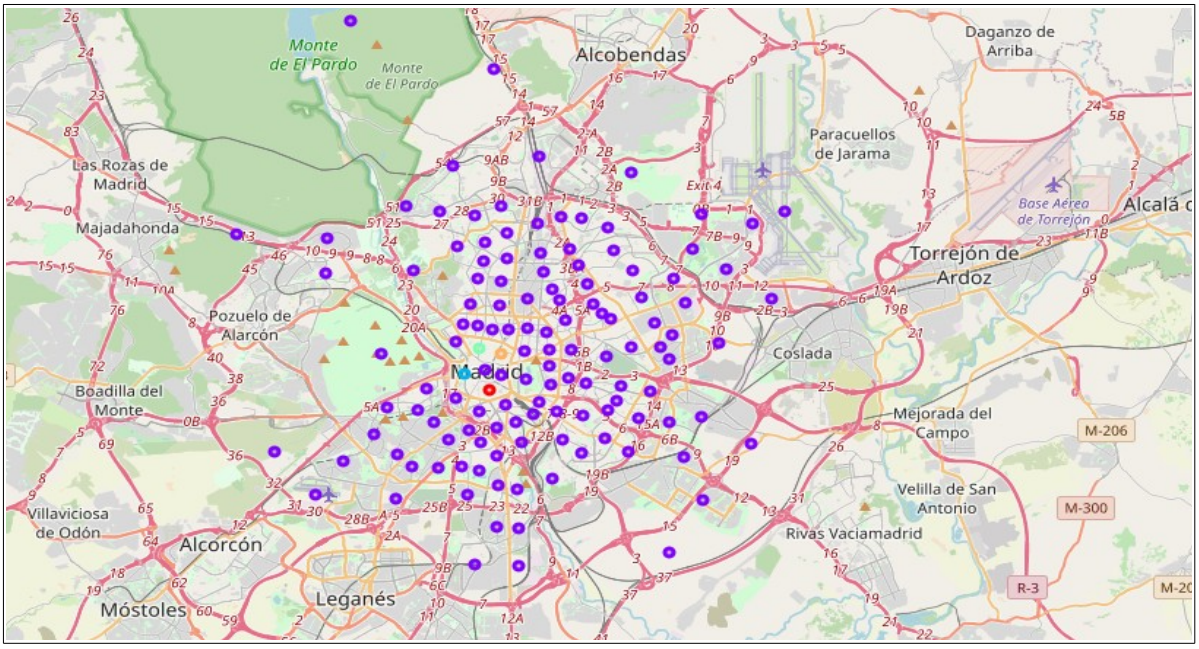

Illustration 6: final map (first approach). Neighborhoods belonging to the main cluster are colored in purple

4.2. Second approach

In the second approach, **the previous process is repeated after extricating the restaurants from the venues dataframe**, since they seem to be ubiquitous in Madrid and thus distort the analysis. The following results are obtained. See Illustration 7, Illustration 8, Illustration 9. Illustration 10. Illustration 11 and Illustration 12

```
************************************************** CLUSTER 0 **************************************************

There are 100 neighborhoods labelled as Cluster 0 (77.5 %).

There are 1601 venues in the neighborhoods labelled as Cluster 0.

                Count  Percentage
Park               97         6.1
Supermarket        92         5.7
Grocery Store      81         5.1
Hotel              64         4.0
Bakery             58         3.6
Coffee Shop        56         3.5
Gym                52         3.2
Café               49         3.1
Clothing Store     44         2.7
Plaza              41         2.6
```

Illustration 7: cluster 0 characteristics (second approach)

```
************************************************** CLUSTER 1 **************************************************

There are 5 neighborhoods labelled as Cluster 1 (3.9 %).

There are 274 venues in the neighborhoods labelled as Cluster 1.

                Count  Percentage
Hotel              28        10.2
Plaza              26         9.5
Café               16         5.8
Bookstore          10         3.6
Theater             9         3.3
Garden              9         3.3
Park                9         3.3
Coffee Shop         8         2.9
Hostel              7         2.6
Ice Cream Shop      6         2.2
```

Illustration 8: cluster 1 characteristics (second approach)

```
************************************************** CLUSTER 2 **************************************************

There are 22 neighborhoods labelled as Cluster 2 (17.1 %).

There are 976 venues in the neighborhoods labelled as Cluster 2.

                Count  Percentage
Bakery             69         7.1
Café               69         7.1
Coffee Shop        60         6.1
Hotel              54         5.5
Supermarket        42         4.3
Burger Joint       34         3.5
Plaza              31         3.2
Sandwich Place     30         3.1
Grocery Store      29         3.0
Brewery            21         2.2
```

Illustration 9: cluster 2 characteristics (second approach)

```
************************************************** CLUSTER 3 **************************************************

There are 1 neighborhoods labelled as Cluster 3 (0.8 %).

There are 53 venues in the neighborhoods labelled as Cluster 3.

                            Count  Percentage
Theme Park Ride / Attraction     13        24.5
Exhibit                           4         7.5
Grocery Store                     4         7.5
Park                              4         7.5
Pool                              3         5.7
Supermarket                       2         3.8
Student Center                    1         1.9
Theme Park                        1         1.9
Nightclub                         1         1.9
Metro Station                     1         1.9
```

Illustration 10: cluster 3 characteristics (second approach)

```
************************************************** CLUSTER 4 **************************************************

There are 1 neighborhoods labelled as Cluster 4 (0.8 %).

There are 66 venues in the neighborhoods labelled as Cluster 4.

                    Count  Percentage
Hotel                  10        15.2
Coffee Shop             9        13.6
Airport Lounge          8        12.1
Airport Service         5         7.6
Duty-free Shop          5         7.6
Rental Car Location     3         4.5
Accessories Store       2         3.0
Grocery Store           2         3.0
Breakfast Spot          2         3.0
Deli / Bodega           2         3.0
```

Illustration 11: cluster 4 characteristics (second approach)

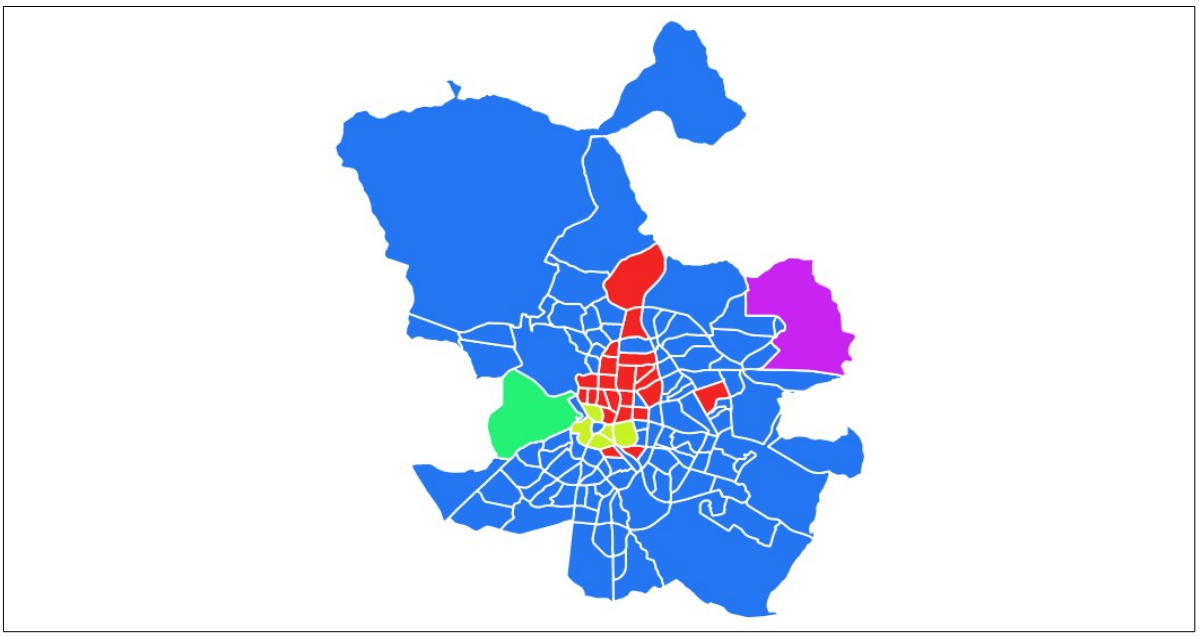

Illustration 12: final map (second approach). Blue: cluster 0; yellow: cluster 1; red: cluster 2; green: cluster 3; purple: cluster 4

5. DISCUSSION

Cluster 0 seems to correspond to neighborhoods where the main venues categories are parks, groceries stores and supermarkets. There are 100 neighborhoods (77.5 %) labeled as Cluster 0. The final map shows that these neighborhoods can be mostly found in boroughs which are outside the "Central Almond" of Madrid (i. e., outside M30 Motorway). They can be thought of as the **suburbs** (residential areas in the urban periphery). However, some of these neighborhoods are part of the "Central Almond" (though they are close to its limits): La Paz, Fuente del Berro, Legazpi, Ciudad Universitaria, etc.

Cluster 1 seems to correspond to neighborhoods where the main venues categories are hotels and plazas. There are 5 neighborhoods (3.9 %) labeled as Cluster 1. The final map shows that these neighborhoods can be mostly found in Centro and Retiro boroughs (i. e., **old town**).

Cluster 2 seems to correspond to neighborhoods where the main venues categories are cafés and coffee shops, bakeries and hotels. There are 22 neighborhoods (17.1 %) labeled as Cluster 2. The final map shows that these neighborhoods can be mostly found in boroughs such as Arganzüela, Chamartín, Chamberí, Retiro, Salamanca and Tetuán. These boroughs make up the "Central Almond" of Madrid (i. e., inside M30 Motorway), and can be thought of as the **downtown**. Only 2 neighborhoods which are not part of "Central Almond" are labeled as Cluster 2: Valverde (Fuencarral - El Pardo) and Simancas (San Blas - Canillejas).

Cluster 3 clearly corresponds to neighborhoods where the main venues categories are theme park rides and attractions. There is only 1 neighborhood (0.8 %) labeled as Cluster 3: **Casa de Campo**. It can be thought of as a special neighborhood, since it borders Casa de Campo, the largest public park in Madrid, which houses Madrid Amusement Park as well as Madrid Zoo.

Cluster 4 clearly corresponds to neighborhoods where the main venues categories are somehow linked to airports: hotels, coffee shops, airport lounges, airport services, duty free shops, etc.. There is only 1 neighborhood (0.8 %) labeled as Cluster 4: **Aeropuerto**. It can also be thought of as a special neighborhood, since it includes the territory around Aeropuerto Adolfo Suárez Madrid-Barajas, the main airport of Madrid.

6. CONCLUSION

The purpose of this project is t**o reveal the distribution of venues among the neighborhoods in Madrid**. I. e., to find out where are the neighborhoods in which a certain venue category is specially usual, which is the most common venue category in each neighborhood?, whether it is there any "outlier" neighborhood in Madrid regarding its venues, and so on. In this case, **the stakeholders are entrepreneurs and investors from the fields of culture, leisure, catering and hostelry**. They could potentially benefit from the results of this project in when it comes to make choices such as: where in Madrid they should put into operation a business of a specific type (the old city, the down town, the suburbs); what business types are more and less common in the area of Madrid which they have

targeted; are there any "outlier" neighborhoods in Madrid when it comes to venues and can they offer special market opportunities?

To do so, the following data have been used: **Madrid boroughs and neighborhoods data** (downloaded from Madrid Open Data website) and **venues data** (extracted from Foursquare). **These data have been pre processed to build a final dataframe** which shows how many venues of each category are in each neighborhood in Madrid.

Then, **K-Means clustering algorithm has been run against the aforementioned dataframe in order to cluster the neighborhoods in Madrid** in a several clusters. These clusters have been inspected to find out their distinctive features. **Finally, they have been represented in a map with different colors depending on their cluster labels, which allows to visually identify their distribution in Madrid**, find out patterns in the distribution, "outlier" neighborhoods in Madrid, the areas of Madrid where it is common to find a specific business category, the most common business categories in a particular area of Madrid, and so on.

The first conclusion to be drawn from the project is that r**estaurants (specially Spanish ones) seem to be ubiquitous in Madrid**. This led to a very uneven clustering that produced a very large cluster (it covered nearly 97% of the neighborhoods in Madrid) which characterized by a high preponderance of Spanish restaurants. So, the first idea for the stakeholders could be that, if they are planning to start a Spanish restaurant, it does not really matter where to do it, since they seem to work out everywhere in Madrid (or that Madrid is saturated with them).

In order to get more detailed results, **a second approach was adopted, but all restaurants were extricated from the dataframe this time**. Then, the previous process was repeated, and the following findings were made:

- **In the old town of Madrid (i. e., the neighborhoods in Centro borough) the top venues categories are clearly hotels and plazas**. This is the oldest area in the city and therefore it is full of historical buildings, monuments, cultural highlights and so on. Note that these neighborhoods account for only 3.9% of all neighborhoods count.

- **In the down town of Madrid (i. e., all boroughs inside the "Central Almond", such as Arganzüela, Chamartín, Chamberí, Retiro, Salamanca and Tetuán), the top venues categories are coffee shops and cafés, bakeries and hotels**. Though the population is slowly decreasing in this area, here is concentrated most of the economic activity. Note that these neighborhoods account for 17.1% of all neighborhoods count. Also note that there are 2 neighborhoods which are not part of "Central Almond" but have the same label: Valverde (Fuencarral - El Pardo) and Simancas (San Blas - Canillejas).

- **In the suburbs of Madrid (i. e., all boroughs outside the "Central Almond", such as Barajas, Carabanchel, Ciudad Lineal, Hortaleza and so on), the top venues categories are parks, supermarkets and groceries**. These are residential areas in the urban periphery. They are some times "bedroom communities", i. e., their inhabitants work in other areas of the city, and come back home only to rest. Note that these neighborhoods account for 77.5%

of all neighborhoods count. Also note that some of these neighborhoods are part of the "Central Almond" (though they are close to its limits): La Paz, Fuente del Berro, Legazpi, Ciudad Universitaria, etc.

- **Casa de Campo is one "outlier" neighborhood**, since it is the only neighborhood in its cluster. **It is very clearly characterized by the presence of theme park rides, attraction and exhibits**. This neighborhood borders Casa de Campo, the largest public park in Madrid, which houses Madrid Amusement Park as well as Madrid Zoo. Note that this neighborhood accounts for 0.8% of all neighborhoods count.

- **Aeropuerto is another "outlier" neighborhood**; it is also the only neighborhood in its cluster. **It is very clearly characterized by the presence of venues linked to airports: hotels, coffee shops, airport lounges, airport services, duty-free shop and so on**. This neighborhood includes the territory around Aeropuerto Adolfo Suárez Madrid-Barajas, the main airport of Madrid. Note that this neighborhood accounts for 0.8% of all neighborhoods count.