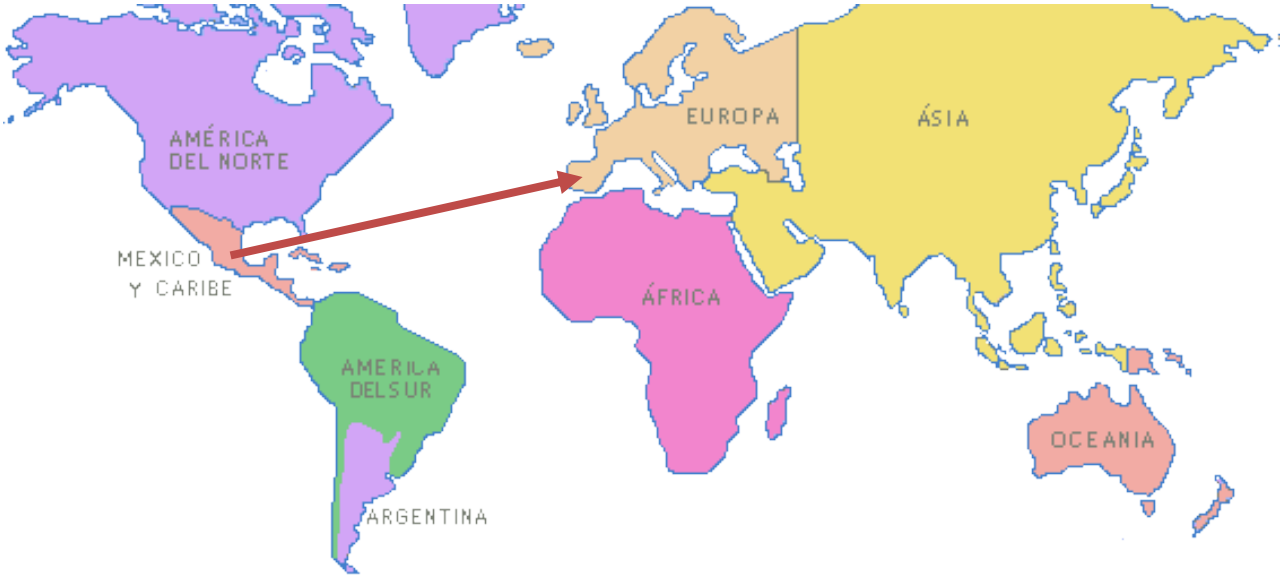
1 Moving to a similar neighborhood

1.1 Table of contents

1.2 Introduction

The approach of my project is based on the fact that I am going to move from Mexico City to Madrid. I want to use the techniques learned in the course to find neighborhoods in Madrid that have the same types of attractions and places of interest (venues) as those in the neighborhood where I currently live.



I will collect the places of interest of my current neighborhood using the Foursquare API, and then I will explore the venues for each of the neighborhoods of Madrid. Finally, I will use KMEANS to find neighborhoods in Madrid similar to my current neighborhood in terms of the places of interest found on Foursquare.

FOURSQUARE

1.3 Data

1.3.1 For my current location

I live in a neighborhood in Mexico City called "Coyoacan". It has the following

coordinates: 19.32804005 -99.1510634069359

I used the Foursquare API with the **explore** endpoint which returns a list of recommended venues near the current location. I limit the search to a radius of 1000 meters and 100 venues at most.

- GET https://api.foursquare.com/v2/venues/explore

The result were 84 venues returned with 41 unique categories.

1.3.2 For Madrid

To get data from Madrid neighborhoods I used data from Portal de datos abiertos del Ayun- tamiento de Madrid. Specifically I downloaded a CSV file titled Relación de barrios (superficie y perímetro).

This file is a list of 128 Districts and Neighborhoods in Madrid.

Distrito	Barrio
Arganzuela	Atocha
Arganzuela	Delicias
Arganzuela	Imperial
Arganzuela	La Chopera
Arganzuela	Las Acacias
…	…

To get geographical coordinates I used **Nominatim** from **geopy.geocoders**:

Distrito	Barrio	Latitud	Longitud
Arganzuela	Atocha	40.405477	-3.689800
Arganzuela	Delicias	40.397292	-3.689495
Arganzuela	Imperial	40.406915	-3.717329
Arganzuela	La Chopera	40.394893	-3.699705
Arganzuela	Las Acacias	40.400759	-3.706995
…	…	…	…

The resulting list includes 118 neighborhoods (for 10 neighborhoods *nominatim* returned "none").

Then, I used the Foursquare API with the **explore** endpoint to get the recommended venues for each neighborhood (using radius of 1000 meters and limit of 100 venues per neighbor- hood).

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Atocha	40.4054769	-3.68979999	Bodegas Rosell	40.403802520	-3.6906202941	Spanish Restaurant
Atocha	40.4054769	-3.68979999	Only You Hotel Atocha	40.407160659	-3.6884378646	Hotel
Atocha	40.4054769	-3.68979999	Running	40.40671358	-3.686904474	Sporting Goods

			Company Madrid			Shop
Atocha	40.4054769	-3.68979999	Estación de Madrid-Puerta de Atocha	40.406571297	-3.6906153807	Train Station
Atocha	40.4054769	-3.68979999	Pandora's Vox	40.40560029	-3.691992411	Music Venue
Atocha	40.4054769	-3.68979999	Jardín Tropical - Invernadero de Atocha	40.40686603	-3.6912039051	Garden
Atocha	40.4054769	-3.68979999	El Rosario	40.403587409	-3.690663146	Restaurant
Atocha	40.4054769	-3.68979999	Rodilla	40.40607987	-3.691218774	Sandwich Place
Atocha	40.4054769	-3.68979999	Plaza del Emperador Carlos V	40.40852749	-3.6926316754	Plaza
Atocha	40.4054769	-3.68979999	AS FONTES	40.404322737	-3.6906389337	Restaurant
Atocha	40.4054769	-3.68979999	Museo Nacional Centro de Arte Reina Sofía	40.40873653	-3.6938217695	Art Museum
…	…	…	…	…	…	…

The resulting dataframe has 3,767 rows with 256 unique categories.

1.4 Methodology

To find similar neighborhoods in Madrid to my neighborhood in Mexico City I used KMEANS to group similar neighborhoods in cluster. One of these cluster includes my neighborhood in Mexico City, so other neighborhoods in the same cluster will be similar to mine. First I transformed the data so all attributes are numeric. For this purpose I followed the following steps:

- Put together the location data of CDMX with those of the neighborhoods of Madrid (variable *geo_barrios*)
- Collect the data of places of interest of CDMX with those of Madrid (variable *madrid_venues*)
- Use "onehot encoding" to transpose the categories of the places of interest and convert them to numerical values
- Group the resulting matrix by neighborhood, using the average value of each category
- Apply KMEANS using different number of clusters (K)
- Measure clustering quality using Silhouette Coefficient and Calinski-Harabaz index
- Select the best K from these results

It turned out that best value for K was 4.

1.5 Analysis

For each neighborhood I looked for what are the 12 most frequent venue categories.

To further limit candidate neighborhoods, I used the distance for all points in cluster 2, which is the one that contains my neighborhood in Mexico City, to its centroid. Then I picked the closest neighborhoods in Madrid to mine.

- Find the distances of all the points to the cluster where the Mexico City neighborhood is located.
- Add cluster label to each point.
- Sort rows by cluster and distance.
- Keep only row for the cluster where the Mexico City neighborhood is located.
- Get the index of Mexico City in this last dataframe.
- Select the neighborhoods closest to the one in Mexico City, according to the distances to the centroid.

1.6 Results and Discussion

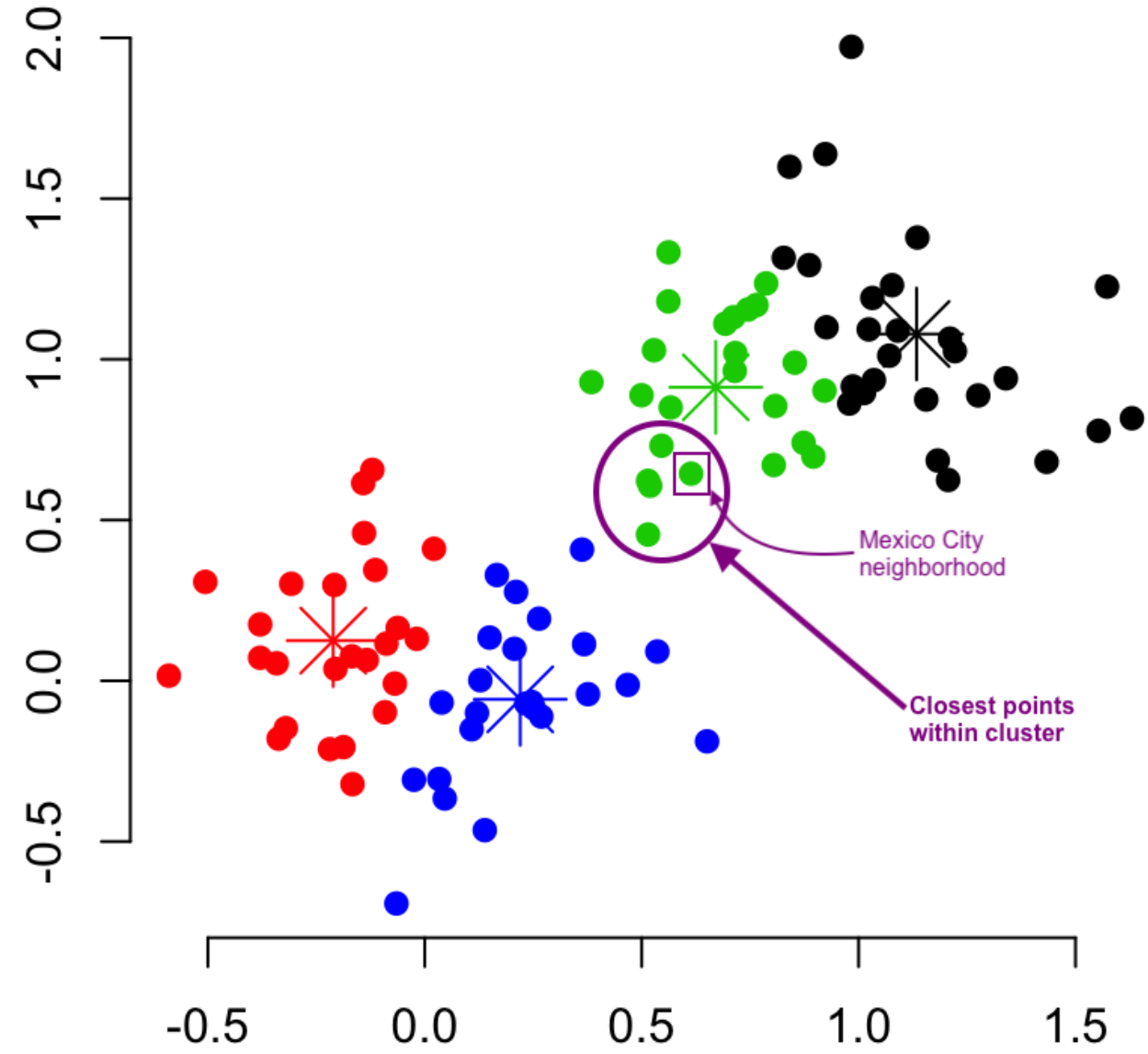
The result of this exercise shows that it is possible to help people who are in a situation similar to the one described in this case, that is, someone who has to move to another city and wants to find conditions similar to those in their current residence using public data available through the Foursquare API.

In KMEANS one of the difficulties is the choice of the value for K. To decide what value to use, I executed the algorithm with different K values and, for each case, I calculated the Silhouette Coefficient and the Calinski-Harabaz Index. These are 2 metric that allow us to decide if we obtain dense and well separated clusters. With both indicators the best value for K was 4.

The venues of my neighborhood in Mexico City were added to the venues of the 117 neighbor- hoods in Madrid and I generated 4 clusters. To further reduce the options of candidate neighbor- hoods, I searched within the cluster where my neighborhood in Mexico City was assigned (cluster 2), for those neighborhoods in Madrid that were closest considering the Euclidean distance of each one of them to the cluster centroid.

The characteristics that distinguish my neighborhood, according to the results of Foursquare, is the diversity of places to eat, shops and places to exercise. These same characteristics are present in almost all the selected neighborhoods.

The technique used to select the candidate neighborhoods is illustrated in the following figure:



1.7 Conclusion

Of course, the final decision cannot be based solely on the results of this analysis. Rather it should be considered as a tool to narrow the options that must be investigated in greater detail.

For example, one way to enrich the results of the analysis would be by adding demographic and socioeconomic attributes to each of the neighborhoods. This would result in more similar and homogeneous clusters.