Moving to Madrid - Which neighborhood to choose

Rebeca Diaz Rubio

May 2020

1. Introduction

1.1. Background and Problem Description

When deciding which neighborhood moving to, there are several key aspects. One of course is budget or price per square foot. Another one is criminal rate. Another one is communications and last one is finding in that neighborhood what matters to the person that is moving. We are going to focus on the latest.

In this project, we are going use data science and Foursquare to recommend families with children interested in moving to Madrid which neighborhood(s) to choose depending on the venues that are more important to them. That would be schools, daycare centers, parks and groceries.

1.2. Interest

Deciding where to move is a big decision. Especially if you don't know the city and when several people are involved, including their belongings, luggage, furniture, etc. This decision can be better made with some data science to back it up and make sure it is the right one from the beginning.

2. Data description and cleaning

2.1. Data sources

To do that, we get the data from several data bases:

- Borughs and neighborhoods in Madrid from Wikipedia:
https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid

- Latitude and longitude of each neighborhood using Geopy library

- High criminality and noisy boroughs from City Council of Madrid:
https://datos.madrid.es/egob/catalogo/212616-74-policia-estadisticas.xlsx, so that we can discard them as they are not recommendable for families.

- Number of each desired venue and Top 10 venues by neighborhood from Foursquare data

Let's describe them. From the mentioned Wikipedia page, we can extract the boroughs and neighborhoods of this city and put them into a pandas dataframe so that we can operate with it and save it. As seen when preparing this data, Madrid has 21 boroughs and 131 neighborhoods.

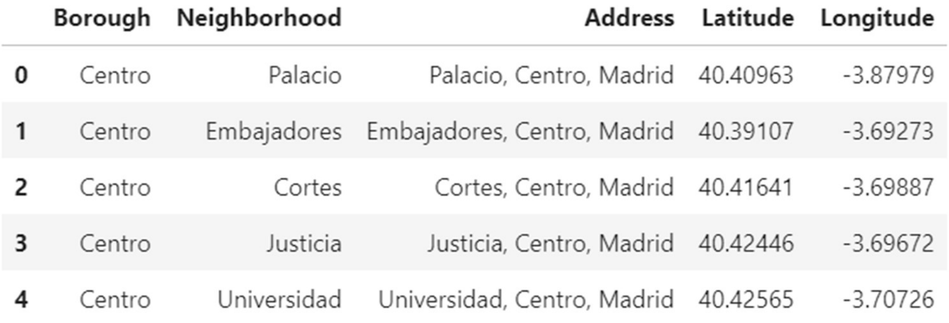
Next, we use Geopy which is a geocoding service useful to locate the coordinates of addresses, cities, countries, and landmarks across the globe. We use it to get longitude and latitude of Madrid city and its neighborhoods.

On City Council of Madrid’s web there is a lot of interesting information about criminality in this city by month. We use the latest data among all the available. On this downloaded excel we find 17 sheets about robberies, arrested people and reports for drinking alcohol in public places, carrying weapons, street trading and reported businesses. We focus on the information by boroughs: arrested people (for any reason), reported people for drinking in the streets and reported businesses, as boroughs with high numbers of these crimes must be noisy, dangerous and non-recommendable for families.

Foursquare data is a location data provider that can be used to explore a zone. For each interesting neighborhood, we can determine what types of venues exist within a defined radius from the center of the neighborhood. That way we can find out if a neighborhood is full of parks and schools, and therefore appropriate for children, or on the contrary, if it is full of bars and discos and therefore non-appropriate for children and families.

2.2. Data cleaning and selection

After downloading the information about boroughs and neighborhoods, we get the coordinates of each neighborhood using Geopy library. To do that, we just need the address formed by the name of the neighborhood and borough, creating a string. For example: “Palacio, Centro, Madrid”. This is the head of the dataframe we get:



	Borough	Neighborhood	Address	Latitude	Longitude
0	Centro	Palacio	Palacio, Centro, Madrid	40.40963	-3.87979
1	Centro	Embajadores	Embajadores, Centro, Madrid	40.39107	-3.69273
2	Centro	Cortes	Cortes, Centro, Madrid	40.41641	-3.69887
3	Centro	Justicia	Justicia, Centro, Madrid	40.42446	-3.69672
4	Centro	Universidad	Universidad, Centro, Madrid	40.42565	-3.70726

As indicated previously, Madrid has 21 boroughs and 131 neighborhoods. In order to reduce the number of neighborhoods to be compared, we are also using data from Madrid’s City Council, to narrow the search and decide which boroughs to compare. Boroughs with high criminality will be excluded. This way the number of boroughs to study is reduced to 4: Arganzuela, Retiro, Chamartín and Villa de Vallecas; and 22 neighborhoods.

3. Methodology

3.1. Madrid Map

First, it is interesting to situate Madrid (Spain) and its neighborhoods on the map. To visualize it and the results, we use visualization library Folium and the coordinates we calculated previously. We also need Madrid’s coordinates, which we get using Geopy:

```
The geograpical coordinate of Madrid are 40.4167047, -3.7035825.
```

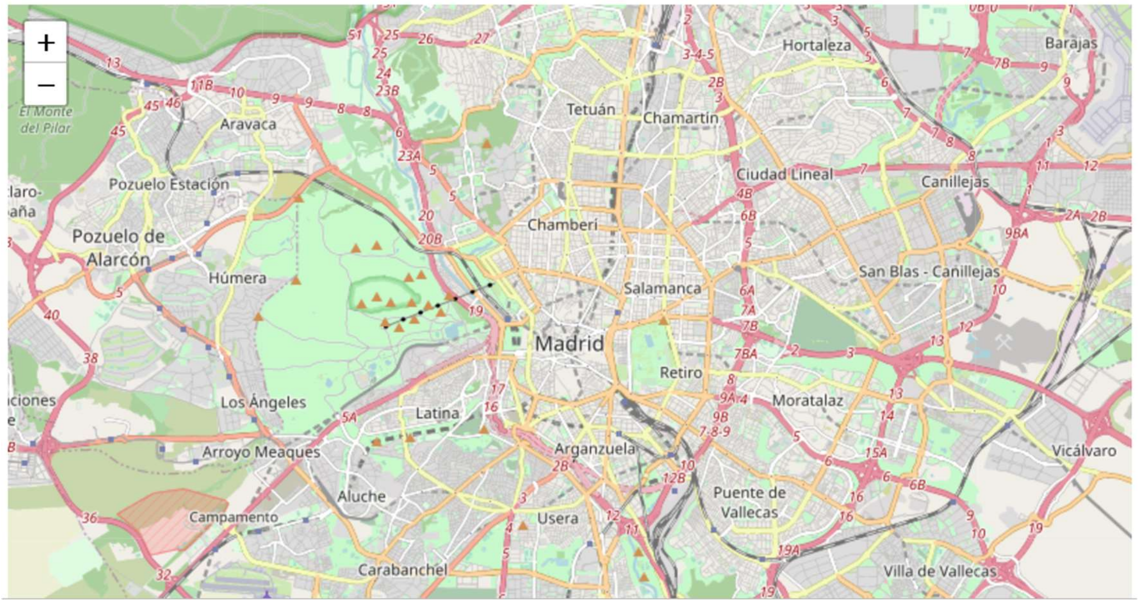
These are the maps we get. Spain:

+
−
Nantes
Centre-Val de Loire
France
München
Österreich
Schweiz/Suisse/Svizzera/Svizra
Graz
Slovenija
Milano
Auvergne-Rhône-Alpes
Nouvelle-Aquitaine
Venezia
Torino
Genova
Bologna
Hrvatska
Oviedo / Uviéu
Monaco
Città di San Marino
Occitanie
Vitoria-Gasteiz
Marseille
Andorra la Vella
Castilla y León
Aragón
Roma
Italia
Napoli
Mar Balear
Palma
España
València
Portugal
Lisboa
Palermo
Murcia
Sevilla
Andalucía
Málaga
Gibraltar
Oran
Algiers
Constantine
تونس
سوسة
Malta

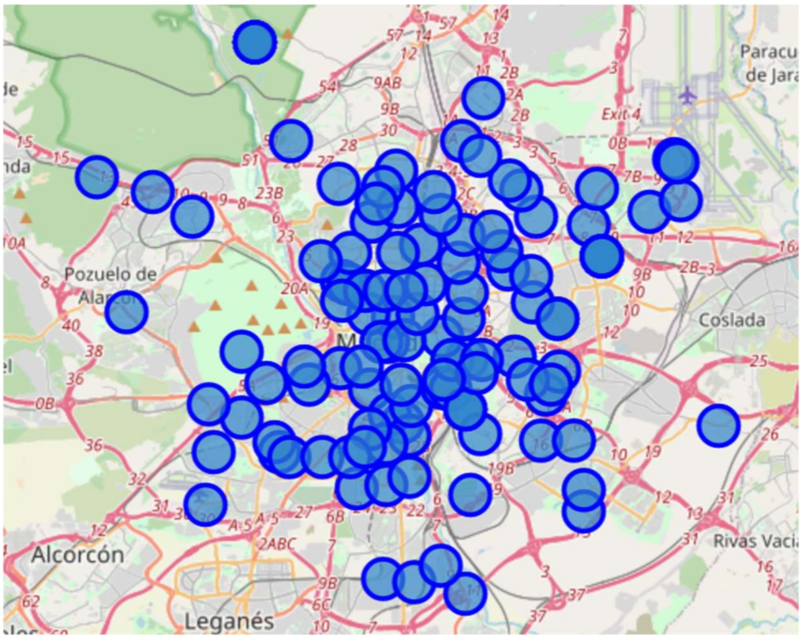
Stamen Terrain map of Madrid:

Galapagar
de Manzanares
Daganzo de Arriba
orial
Las Matas
Meco
arejo
Las Rozas de Madrid
San Sebastián de los Reyes
Villanueva del Pardillo
Majadahonda
Madrid
Brunete
Pozuelo de Alarcón
Mejorada del Campo
Torres de la Alame
Villaviciosa de Odón
evilla
ueva
Navalcarnero
Arganda
Humanes

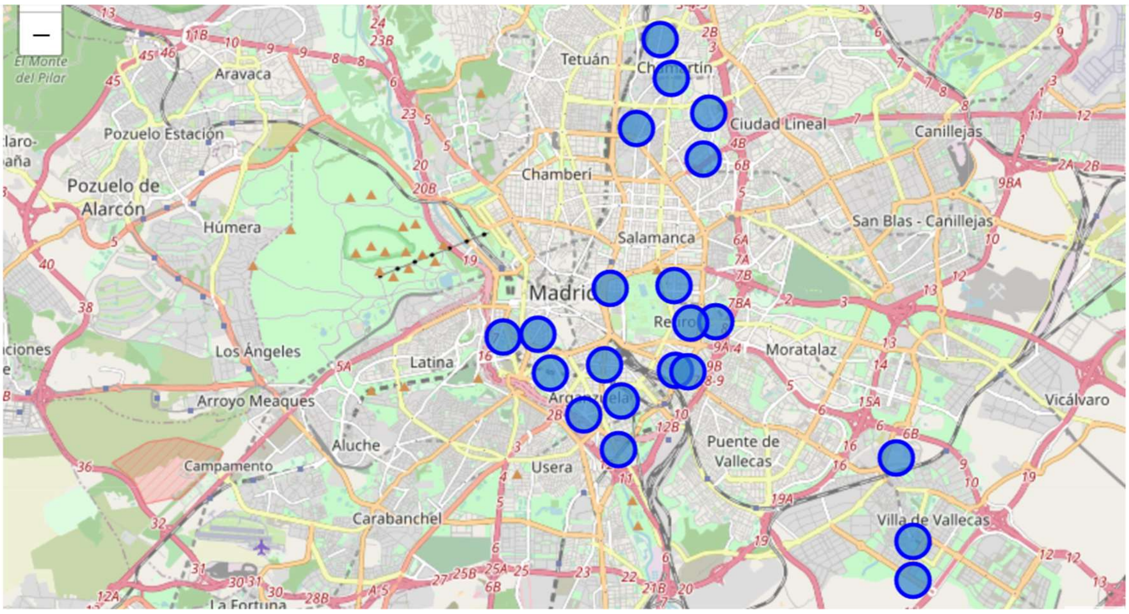
Map of Madrid using zoom 12:



This is how all the neighborhoods look on the map:



After reducing the neighborhoods we are interested in, we can draw them on the map:



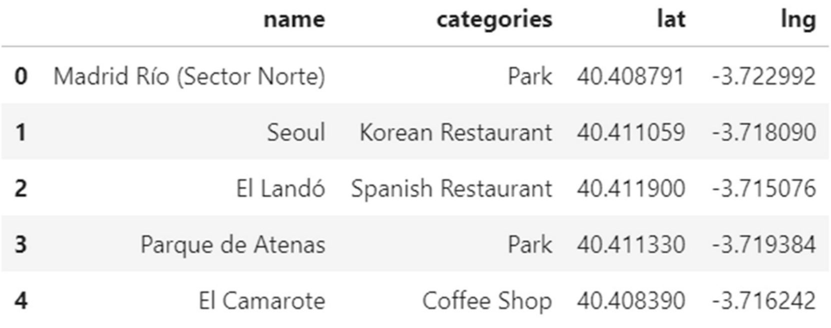
3.2. Foursquare

We utilize Foursquare API to explore the neighborhoods and segment them. First, we need to define Foursquare Credentials and Version, we get them when signing up on Foursquare page.

Then we explore the first neighborhood in our dataframe to see if it works properly. We do that by getting the top 100 venues that are in the first neighborhood within a radius of 500 meters. First neighborhood in the dataframe is Imperial:

```
Latitude and longitude values of Imperial, Arganzuela, Madrid are 40.408330000000035, -3.71864
9999999968.
```

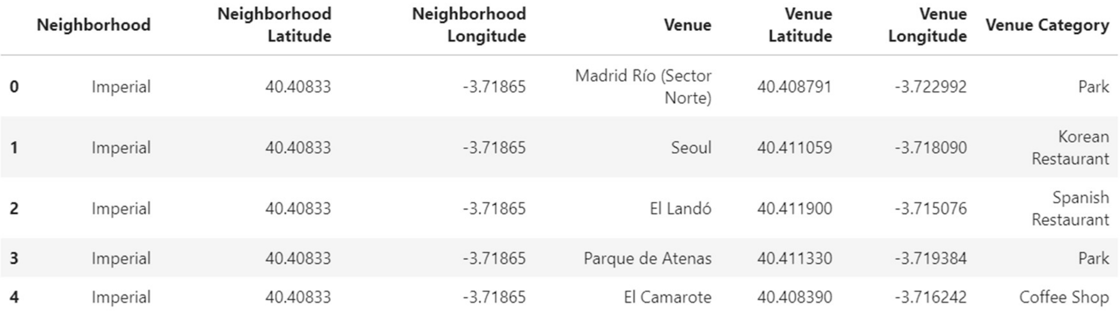
We get 25 venues returned by Foursquare and put them in a datafrma as follows:



	name	categories	lat	lng
0	Madrid Río (Sector Norte)	Park	40.408791	-3.722992
1	Seoul	Korean Restaurant	40.411059	-3.718090
2	El Landó	Spanish Restaurant	40.411900	-3.715076
3	Parque de Atenas	Park	40.411330	-3.719384
4	El Camarote	Coffee Shop	40.408390	-3.716242

3.3. Venues in the city

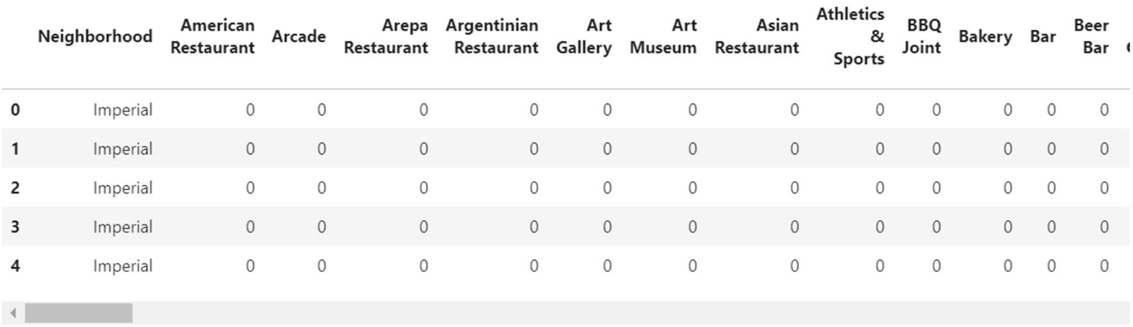
Next step is getting all the venues in the boroughs we are interested in. We get 757 venues in those 22 neighborhoods and put them into a dataframe:



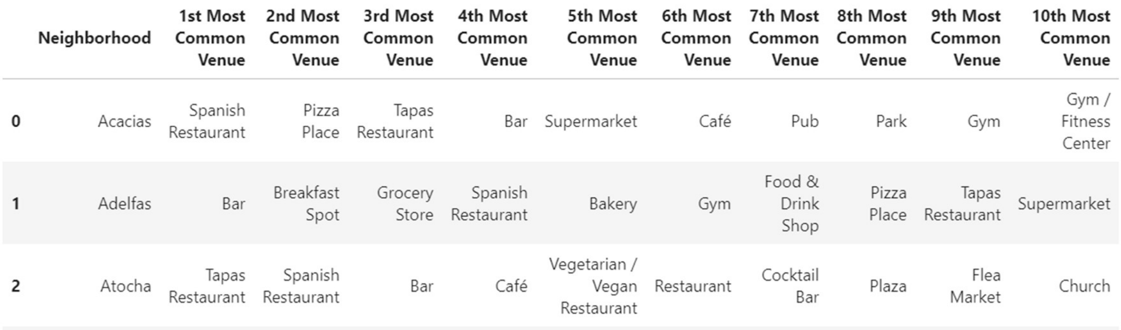
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Imperial	40.40833	-3.71865	Madrid Río (Sector Norte)	40.408791	-3.722992	Park
1	Imperial	40.40833	-3.71865	Seoul	40.411059	-3.718090	Korean Restaurant
2	Imperial	40.40833	-3.71865	El Landó	40.411900	-3.715076	Spanish Restaurant
3	Imperial	40.40833	-3.71865	Parque de Atenas	40.411330	-3.719384	Park
4	Imperial	40.40833	-3.71865	El Camarote	40.408390	-3.716242	Coffee Shop

By counting the number of venues for each neighborhood we can Casco Histórico de Vallecas, Santa Eugenia and Castilla have very few venues and 143 unique categories.

We just need the neighborhoods and categories, so we create a venue matrix indicating with number 1 the venue type of each one we found.



	Neighborhood	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bar	Beer Bar
0	Imperial	0	0	0	0	0	0	0	0	0	0	0	0
1	Imperial	0	0	0	0	0	0	0	0	0	0	0	0
2	Imperial	0	0	0	0	0	0	0	0	0	0	0	0
3	Imperial	0	0	0	0	0	0	0	0	0	0	0	0
4	Imperial	0	0	0	0	0	0	0	0	0	0	0	0



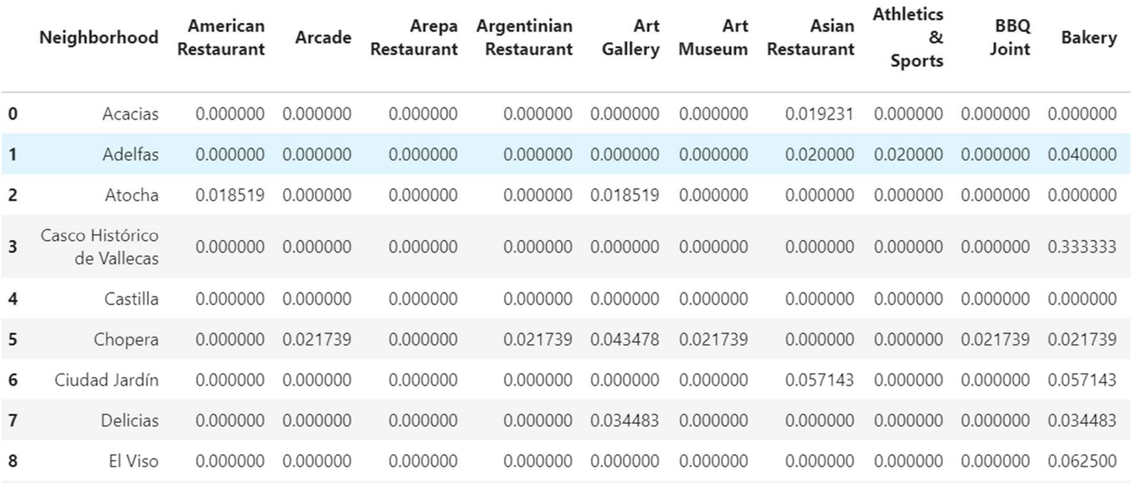
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Acacias	Spanish Restaurant	Pizza Place	Tapas Restaurant	Bar	Supermarket	Café	Pub	Park	Gym	Gym / Fitness Center
1	Adelfas	Bar	Breakfast Spot	Grocery Store	Spanish Restaurant	Bakery	Gym	Food & Drink Shop	Pizza Place	Tapas Restaurant	Supermarket
2	Atocha	Tapas Restaurant	Spanish Restaurant	Bar	Café	Vegetarian / Vegan Restaurant	Restaurant	Cocktail Bar	Plaza	Flea Market	Church

3.4. Cluster Neighborhoods

We run k-means algorithm to cluster the neighborhood into 4 clusters. We use the Folium library to visualize the neighborhoods in Madrid and their emerging clusters.

4. **Results**

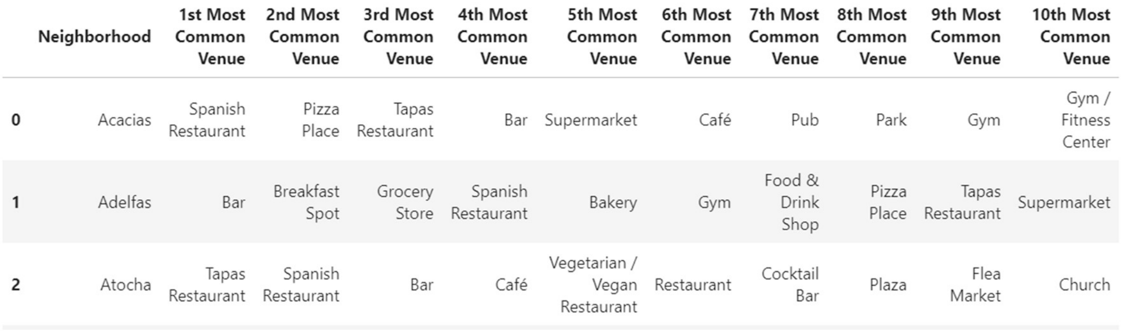
We have found and grouped the venues of each neighborhood and displayed them taking the mean of the frequency of occurrence of each category.



	Neighborhood	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery
0	Acacias	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.019231	0.000000	0.000000	0.000000
1	Adelfas	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.020000	0.020000	0.000000	0.040000
2	Atocha	0.018519	0.000000	0.000000	0.000000	0.018519	0.000000	0.000000	0.000000	0.000000	0.000000
3	Casco Histórico de Vallecas	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333
4	Castilla	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Chopera	0.000000	0.021739	0.000000	0.021739	0.043478	0.021739	0.000000	0.000000	0.021739	0.021739
6	Ciudad Jardín	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.057143	0.000000	0.000000	0.057143
7	Delicias	0.000000	0.000000	0.000000	0.000000	0.034483	0.000000	0.000000	0.000000	0.000000	0.034483
8	El Viso	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.062500

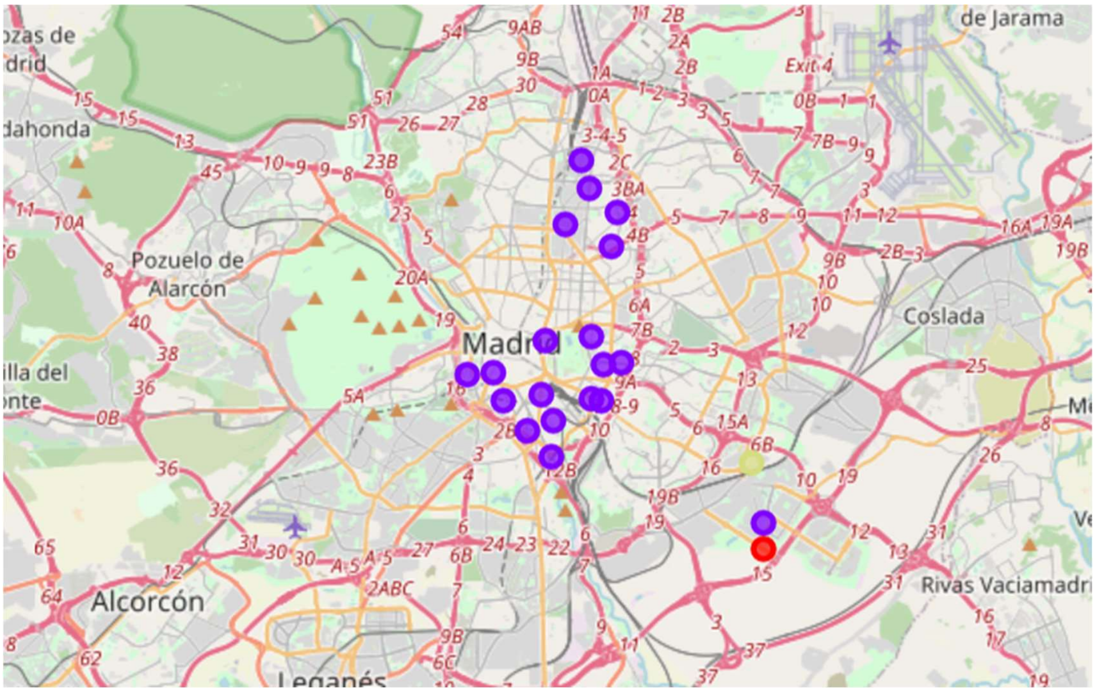
This way we get a dataframe with 22 lines or neighborhoods and 144 types of venues.

Then we create a new dataframe and display the top 10 venues for each neighborhood. For example:



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Acacias	Spanish Restaurant	Pizza Place	Tapas Restaurant	Bar	Supermarket	Café	Pub	Park	Gym	Gym / Fitness Center
1	Adelfas	Bar	Breakfast Spot	Grocery Store	Spanish Restaurant	Bakery	Gym	Food & Drink Shop	Pizza Place	Tapas Restaurant	Supermarket
2	Atocha	Tapas Restaurant	Spanish Restaurant	Bar	Café	Vegetarian / Vegan Restaurant	Restaurant	Cocktail Bar	Plaza	Flea Market	Church

We have ended the study by visualizing the data and clustering the information on the map.



We have divided the neighborhoods in 4 clusters:

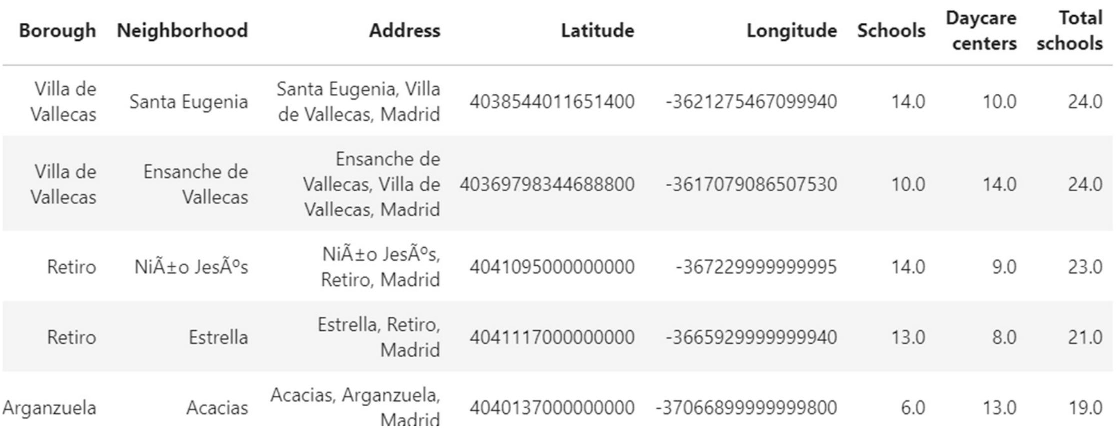
- Cluster 1. Only one neighborhood Casco Histórico de Vallecas fits into this cluster. This neighborhood is basically based on food shops with a slight touch of outdoors activities.
- Cluster 2. In this cluster we find most neighborhoods we have previously selected. As we can see, they are very alike.
- Cluster 3. Only one neighborhood Castilla fits into this cluster. In this neighborhood most venues are markets and restaurant businesses. Therefore, it is not very recommendable for our purpose.
- Cluster 4. Only one neighborhood Santa Eugenia fits into this cluster. In this neighborhood there are gyms and other type of businesses non-related to our purpose.

5. Discussion

As seen in the results section, our perfect neighborhood should be in cluster 1 or cluster 2.

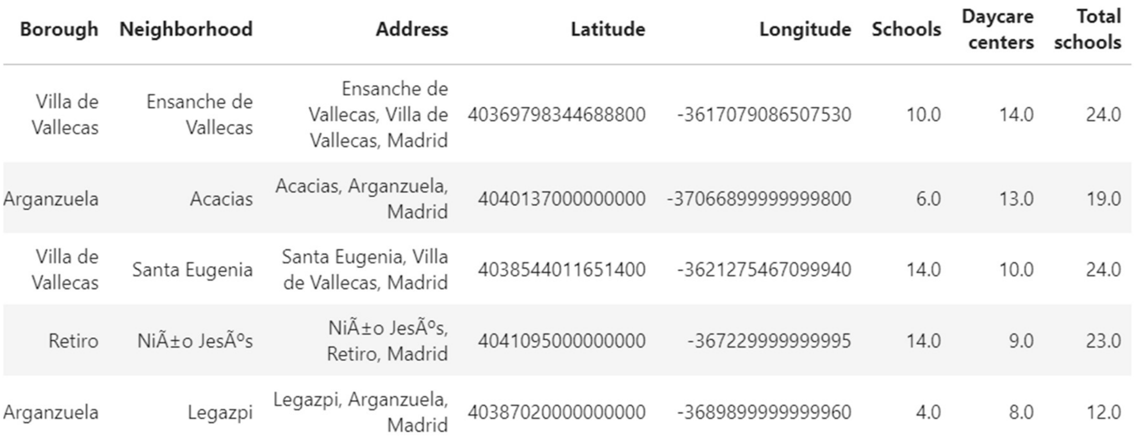
Unlikely, on Foursquare no information about schools and daycare centers is available, which would be a must-have in this perfect neighborhood we are looking for. We have extracted this information from other tool, we used Googleapis by proximity, put it into a dataframe and sorted it by higher values:

By total number of schools and daycare centers:



Borough	Neighborhood	Address	Latitude	Longitude	Schools	Daycare centers	Total schools
Villa de Vallecas	Santa Eugenia	Santa Eugenia, Villa de Vallecas, Madrid	4038544011651400	-3621275467099940	14.0	10.0	24.0
Villa de Vallecas	Ensanche de Vallecas	Ensanche de Vallecas, Villa de Vallecas, Madrid	40369798344688800	-3617079086507530	10.0	14.0	24.0
Retiro	NiÃ±o JesÃºs	NiÃ±o JesÃºs, Retiro, Madrid	4041095000000000	-367229999999995	14.0	9.0	23.0
Retiro	Estrella	Estrella, Retiro, Madrid	4041117000000000	-3665929999999940	13.0	8.0	21.0
Arganzuela	Acacias	Acacias, Arganzuela, Madrid	4040137000000000	-37066899999999800	6.0	13.0	19.0

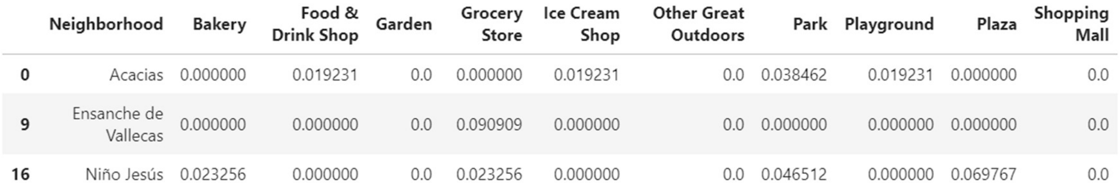
By number of daycare centers:



Borough	Neighborhood	Address	Latitude	Longitude	Schools	Daycare centers	Total schools
Villa de Vallecas	Ensanche de Vallecas	Ensanche de Vallecas, Villa de Vallecas, Madrid	40369798344688800	-3617079086507530	10.0	14.0	24.0
Arganzuela	Acacias	Acacias, Arganzuela, Madrid	4040137000000000	-37066899999999800	6.0	13.0	19.0
Villa de Vallecas	Santa Eugenia	Santa Eugenia, Villa de Vallecas, Madrid	4038544011651400	-3621275467099940	14.0	10.0	24.0
Retiro	NiÃ±o JesÃºs	NiÃ±o JesÃºs, Retiro, Madrid	4041095000000000	-367229999999995	14.0	9.0	23.0
Arganzuela	Legazpi	Legazpi, Arganzuela, Madrid	40387020000000000	-3689899999999960	4.0	8.0	12.0

The neighborhoods with most daycare centers and schools are: Ensanche, Acacias, Niño Jesus and Santa Eugenia. As discussed before, Santa Eugenia is Cluster 4 (Others) non-recommendable for our purpose.

In addition to selecting the neighborhood regarding the type of neighborhood, we would also need to select it depending on the important venues. We have listed the important or interesting venues for the narrowed list of neighborhoods, getting:



	Neighborhood	Bakery	Food & Drink Shop	Garden	Grocery Store	Ice Cream Shop	Other Great Outdoors	Park	Playground	Plaza	Shopping Mall
0	Acacias	0.000000	0.019231	0.0	0.000000	0.019231	0.0	0.038462	0.019231	0.000000	0.0
9	Ensanche de Vallecas	0.000000	0.000000	0.0	0.090909	0.000000	0.0	0.000000	0.000000	0.000000	0.0
16	Niño Jesús	0.023256	0.000000	0.0	0.023256	0.000000	0.0	0.046512	0.000000	0.069767	0.0

As it can be seen, Ensanche has no parks of playgrounds, so it is not recommendable for families with children. Niño Jesus is the neighborhood families would be most interested to live in and Acacias the second one for having everything they will need close by.

6. Conclusions and Future directions

Madrid, as most capital cities, is a wide city with a wide range of neighborhoods and deciding which neighborhood moving to can be tricky specially if you are a foreigner. By using data science, we have reduced the number of districts of interest and analyzed them.

This study has focused on one type of interested people: families with children. However, it is important to highlight that this study can be flexible and customized depending on the type of interested profile. For example, young couple or single youngsters may be interested in moving into other types of neighborhoods: livelier and with more leisure areas.

Also, it can be used the other way around. Analyzing the venues of each neighborhood, we can tell the profile of the people living there. This can be useful for new businesses and for advertising campaign.

Another interesting variable to be added in future lines and analyzed would be the cost of cost of the house by square foot.

7. References

[1] Wikipedia: https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid

[2] Geopy library

[3] City Council of Madrid: https://datos.madrid.es/egob/catalogo/212616-74-policia-estadisticas.xlsx

[4] Forsquare API

[5] Google Map