Analysing Madrid districts for opening a cafeteria

COURSERA CAPSTONE PROJECT

Jairo Silvera | Applied Data Science Capstone | July 28, 2021

1. Introduction

1.1 Background

Madrid is the capital city of Spain, is a cosmopolitan city which houses the headquarters of the Government of Spain and its Ministries. It is an important economic center in Europe with a high nominal GDP which makes it a very dynamic city in several areas within the economic, educational, sports, tourism and cultural.

The commerce in Madrid is very active, the commercial premises type cafeteria or coffe shop are visited by locals and tourists who see an opportunity to chat and enjoy a good coffee and derivatives. Madrid is divided into districts which have their own characteristics, some are more touristic, others are educational and cultural, some are more residential, commercial and industrial.

1.2 Problem

The data can help any interested owner or holders interested in starting a coffee shop type business which leads to a better understanding of the possible location, district and neighborhoods. The goal of this project is to help locate the best district based on the use of Machine Learning techniques also verifying the level of COVID contagions.

1.3 Interest

Prospective and potential stakeholders interested in opening a coffeeshop business would be very receptive to this study which provides an insight into the business idea and where it could be located.

2. Data Acquisition and cleaning

2.1 Data sources

The data on location, districts, neighborhoods, COVID were extracted from official sources of the city of Madrid, the data source is free to consult example:

COVID data:

https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos/resource/877fa8f5-cd6c-4e44-9df5-0fb60944a841

Districts and neighborhoods:

https://datos.gob.es/en/catalogo/a13002908-covid-19-tia-por-municipios-y-distritos-de-madrid1

2.2 Data cleaning

The extracted data were grouped into datasets and loaded into the notebook for further processing.

The COVID dataset shows contagion information by district, the population column for each district was added, some data and columns were eliminated as they were irrelevant.

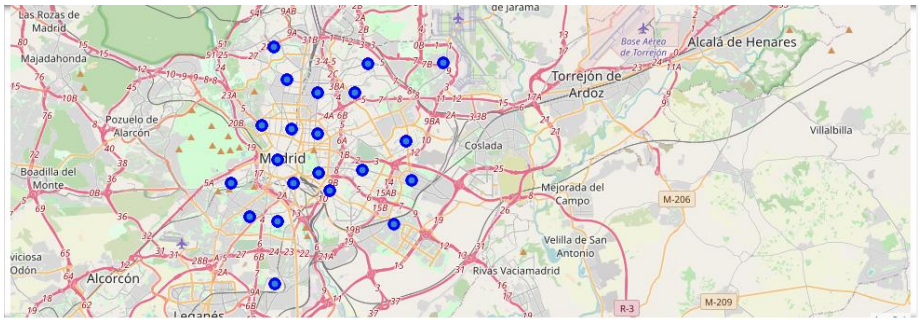
The location dataset by district and neighborhoods contains the basic information such as coordinates, names of neighborhoods and districts where also irrelevant data were removed.

3. Exploratory Data Analysis

3.1 Data wrangling

The location data were extracted directly from the dataset, the coordinates correspond to each district, which indicates the best location per district.

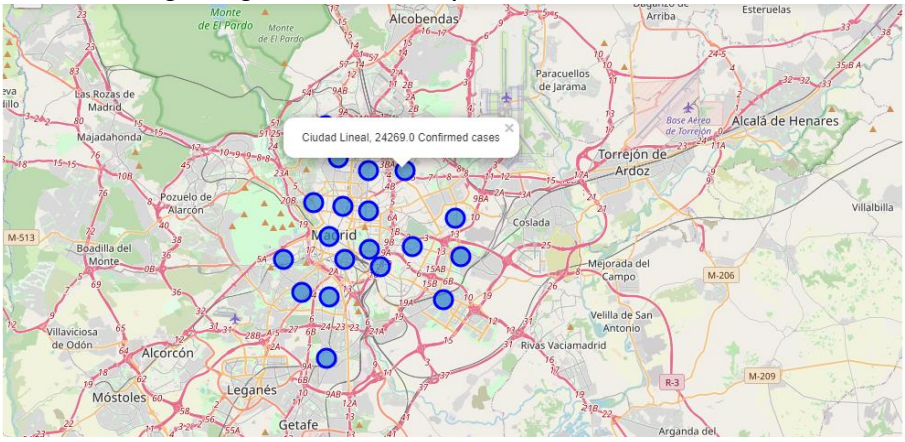
This folium maps image shows the 21 districts of Madrid.



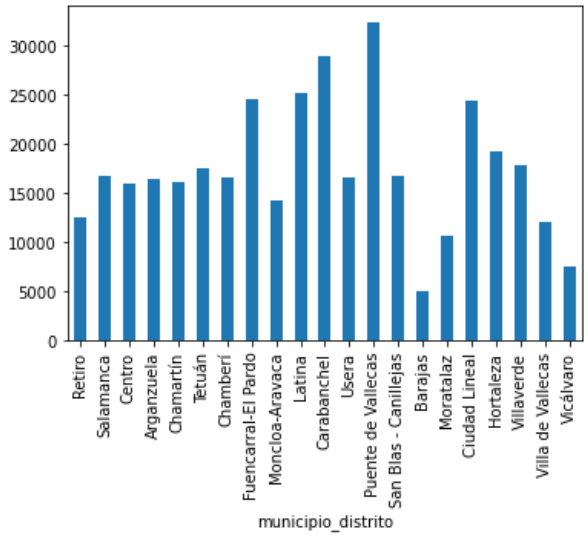
The most relevant data of the location dataset are the name of the district, neighborhood and the coordinates of each district.

The COVID dataset was also extracted from official sources where the most relevant data are district name, district coordinates and confirmed cases.

The following image of folium maps shows the districts and their confirmed cases.



The following image shows the confirmed cases per district



The average number of cases is 17382.23, the district with the most cases is Puente de Vallecas with a population of 241901 inhabitants, which is located between the center of the capital and the periphery.

4. Clustering model

The clustering model is used in conjunction with the Foursquare API which gives us an idea of the best sector to locate the potential new coffee shop business. The Clustering model is optimal for this case since the data is grouped by district and through the API we can see the most popular venues..

4.1 Applying cluster segmentation

According to the API data we extracted the venues by categories grouped by each district, there are very varied categories such as theaters, thai restaurant, wine bar, etc. The one needed for this exercise is cafeteria also coffe shop.

Finally, the 10 most popular venues in each district are grouped as follows:

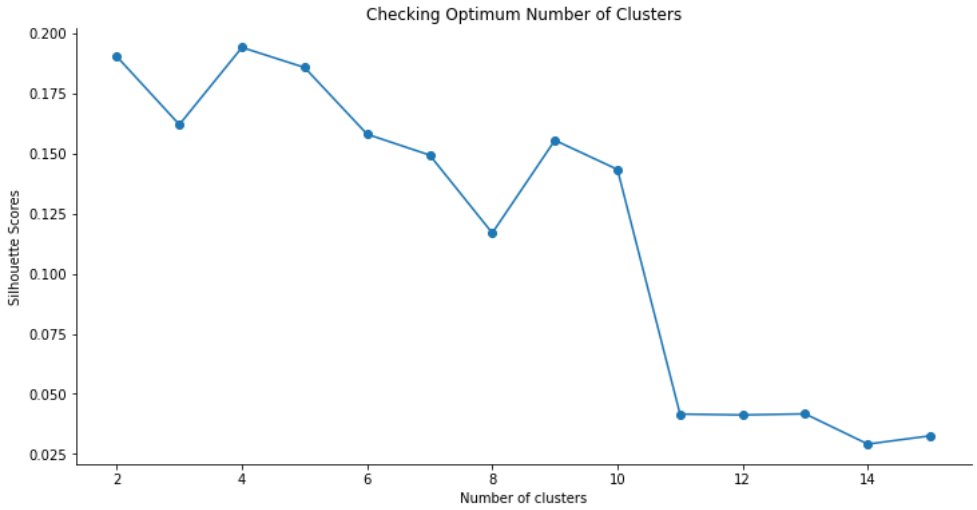
	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Arganzuela	Restaurant	Spanish Restaurant	Grocery Store	Bakery	Tapas Restaurant	Gym / Fitness Center	Beer Garden	Market	Falafel Restaurant	Burger Joint
1	Barajas	Hotel	Restaurant	Spanish Restaurant	Coffee Shop	Tapas Restaurant	Bar	Fast Food Restaurant	Brewery	Mexican Restaurant	Café
2	Carabanchel	Pizza Place	Nightclub	Fast Food Restaurant	Burger Joint	Tapas Restaurant	Bakery	Soccer Field	Metro Station	Plaza	Diner
3	Centro	Spanish Restaurant	Tapas Restaurant	Plaza	Hostel	Ice Cream Shop	Bookstore	Hotel	Pastry Shop	Cocktail Bar	Gym / Fitness Center
4	Chamartin	Spanish Restaurant	Restaurant	Bakery	Grocery Store	Tapas Restaurant	Café	Gastropub	Park	Coffee Shop	Japanese Restaurant
5	Chamberi	Spanish Restaurant	Restaurant	Bar	Brewery	Italian Restaurant	Japanese Restaurant	Café	Plaza	Mexican Restaurant	Tapas Restaurant
6	Ciudad Lineal	Restaurant	Spanish Restaurant	Park	Grocery Store	Argentinian Restaurant	Gastropub	Tapas Restaurant	Pharmacy	Café	Bus Line

In the image above the most popular venues are shown , they are diverse and there are different categories by district which makes Madrid a very active and varied city commercially. You can appreciate the culinary taste as the restaurant category is among the first, hotel and hostel are also important because it is touristy city.

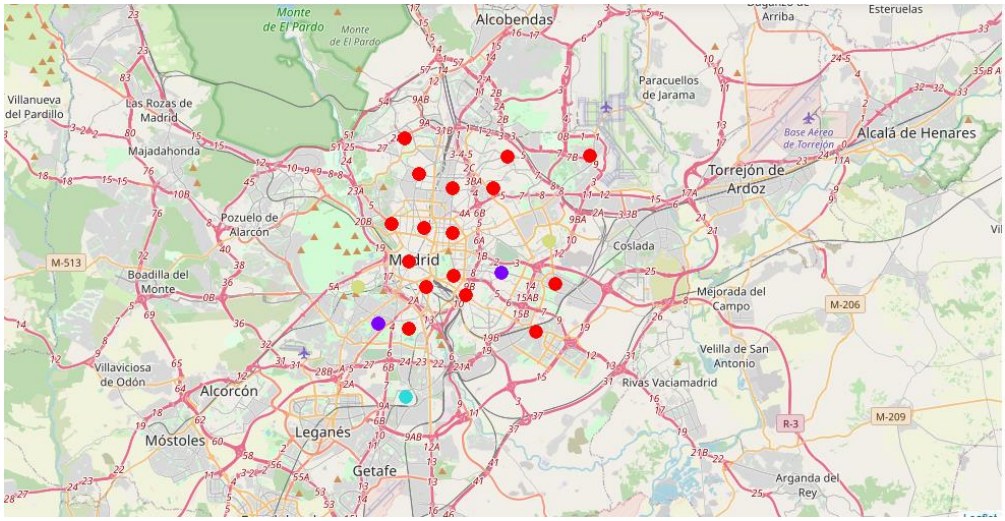
4.2 Clustering

Using the Silhouette Score technique we verify the optimal number of clusters.

In the following image it can be seen that the optimal number is 3 and 4 since they have the highest score, not necessarily the more clusters the better, with 14 clusters it has a very low score.



Once the number of clusters has been identified, the map looks like this:



Cluster 1 is the most suitable for the new business given the number of coffee shop venues and their popularity in these districts, which some are Chamartin, Barajas, Hortaleza.

5. Conclusions

In this study we analyzed the different districts of Madrid grouped into 4 clusters, it can be seen that the districts with more trend and more venues in the Foursquare API are those that are tourist, educational and commercial. Potential stakeholders may have a broader idea of where to start a new business type cafeteria in some of these sectors as they are widely reviewed in Foursquare, but not before taking into account the impact of COVID as it is currently fighting the disease for example Puente de Vallecas district with more than 30000 confirmed cases is a good district due to the venues in cafeteria, but is well above the average of contagions.

Moratalaz can be interesting as it does not have as many infections as other districts and Barajas also has the lowest number of infections and has a fourth place in cafeteria venues.

In addition, the venues are closely related there are venues type restaurants, breakfast, coffee shop, for a potential investor the culinary field is important in Madrid.

For example, the Latina district located in cluster 4 has high COVID cases (25000) and the most popular venues are restaurant type, pizza, fast food, Asian food, but not coffee shops could be a good indication as they are culinary type.

Moratalaz district for example has high venues in pizza place, food truck and also Café and at the same time it does not have as many cases as other districts with 10609 cases it is below the average 17382.23.