Analysis of COVID-19 in Madrid

Applying machine learning to a pandemic

Table of Contents

1. Introduction 3
   1.1. Background 3
   1.2. Stakeholders 3
2. Datasets 4
   2.1. Description 4
   2.2. Cleaning datasets 5
3. Methodology 8
4. Results 12
5. Discussion 13
6. Conclusion 13

Table of Figures

Figure 1 COVID-19 data by district initial version 5
Figure 2 COVID-19 data by district final version 5
Figure 3 Districts' coordinates 6
Figure 4 COVID-19 by district with coordinates 7
Figure 5 Data collected from *FourSquare* 7
Figure 6 Venues categories grouped by district 7
Figure 7 Venues categories mean frequency 8
Figure 8 Ten most popular venue categories by district 8
Figure 9 Districts labelled by their COVID-19 rate 9
Figure 10 Map of districts with their COVID-19 rate 9
Figure 11 Map of districts with their COVID-19 rate and markers 10
Figure 12 Ten most popular venue categories by district 10
Figure 13 *Districts clustered using K-means algorithm* 11
Figure 14 K-means graph 11
Figure 15 Districts labelled by number of infections and clustered by venue categories 12
Figure 16 Map of districts labelled by number of infections and clustered by venue categories 13

1. Introduction

1.1. Background

COVID-19 pandemic has marked a milestone in our lives: no matter your nationality, your job or your social status, this disease is being a major challenge for most of us. However, although this health and economic crisis has affected most of population world-wide, it is true that its impact has been much more serious in some countries or areas.

Factors such as ageing population, precarious jobs or economic models have been some of the reasons why some countries have suffered more infections and deaths and their economic recovery is likely to be more arduous.

Spain was one of the countries more affected during COVID-19 first wave, and factors mentioned above (ageing population, precarious jobs and its economic model) are thought to be the main reasons.

Despite the fact that Spain suffered a huge number of deaths, unemployment increase and serious hospital saturation due to COVID-19, the truth is this situation was unequal: though the situation was serious all around the country, it was critical in some areas.

Madrid, the capital city of Spain, was one of the areas where the crisis was uncontrollable: its large population (more than 3 million) and travellers (not only travellers from the rest of Spain, but also travellers from other countries) continuously coming and going are considered the principal causes.

Due to Madrid's great population, economic differences between its districts are notorious: whereas population of some of them is considered "working-class", the population of the rest is labelled as "upper class". It is confirmed that purchasing power is also a risk factor for COVID-19: people who live in areas with lower incomes are more likely to be infected. This is due to the fact that people who live in these areas often live with their whole family (grandparents, brothers and sisters, children, etc.), feel obliged to share a flat and expenses with flatmates and/or use public transport (which is usually overcrowded) instead of getting to their workplace in their own car.

Another risk factor for getting infected with COVID-19 is to frequent some venues with certain characteristics. Places which are indoors, usually overcrowded and/or where mask-faces are not obligatory, multiply the risk of infection.

Although everyone is exposed to COVID-19, some factors increase the risk of getting infected, and the ones mentioned in the two previous paragraphs are some of them. The study and the analysis this information can be useful to predict and avoid (or, at least, slow down) future outbreaks.

1.2. Stakeholders

As mentioned above, the place where you live and the venues you frequent can be risk factors for getting infected with COVID-19. Using the experience and the data acquired during the first wave of the pandemic, another similar situation can be avoided in the future.

Taking the districts of Madrid, their number of infections by COVID-19 and their most popular venues, these districts can be clustered in order to know which ones are more suitable to suffer another outbreak. Knowing it, governments will be able to take measures as soon as the number of infections starts to grow in any of these districts. Not only will this knowledge be useful for these specific districts, but it also can be applied in areas with similar characteristics.

Therefore, on the basis of a dataset which contains the number of COVID-19 infections in each district of Madrid, the first step of this project will consist of labelling them depending on the number of infections. The next step will be getting the most popular venues for each district: using the coordinates of each district and *FourSquare*, we will be able to add to each district its most popular venues. Finally, districts will be clustered by their most popular venues.

Having districts labelled by number of infections and clustered by their most popular venues, the type(s) of venue where an outbreak is more probable can be easily identified and restrictive measures can be sooner taken and customized for this/these kind(s) of venue.

2. Datasets

In order to get the aim of this project, three datasets will be used. Although each dataset will be shown in detail when loaded in the project, a brief description for each dataset is written below. Afterwards, the cleaning applied to each dataset will be explained.

2.1. Description

COVID-19 data by district

The first dataset, provided by the Community of Madrid[1], contains COVID-19 cumulative incidence on specific dates by town and, in the case of the city of Madrid, by district. Since this project will focus exclusively on the city of Madrid, only data relative to districts will be used.

This dataset can be downloaded from here: https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos

Districts' coordinates

In order to visualize districts over a map and relate them with *FourSquare* information, it is needed to get the latitude and the longitude of each district. Due to the number of districts (only 21) and the impossibility to find their coordinates in any open data repository, coordinates of each district will be got "by hand" using Google and saving the result in a custom data set, which will contain the name of every district and its coordinates.

Most popular venues by district

Using the coordinates of each district, most popular venues in each one will be got through calls to *FourSquare* API.

[1] Spain is divided into 17 communities and the capital city of Spain, Madrid, is located in the Community of Madrid. Therefore, "the Community of Madrid" is not the same as "the city of Madrid". In this project, we will study the city of Madrid.

2.2. Cleaning datasets

Since "*COVID-19 data by district*" is the only dataset which will be provided ("*Districts' coordinates*" will be made and "*Most popular venues by district*" is not a dataset as such.

Initially, "*COVID-19 data by district*" looks like that:

	municipio_distrito	fecha_informe	casos_confirmados_ultimos_14dias	tasa_incidencia_acumulada_ultimos_14dias	casos_confirmados_totales	tasa_incidencia_acumulada_total	codigo_geometria
0	Madrid-Retiro	2020/07/01 09:00:00	28.0	23,47	1691.0	1417,23	79603
1	Madrid-Salamanca	2020/07/01 09:00:00	23.0	15,74	1781.0	1218,99	79604
2	Madrid-Centro	2020/07/01 09:00:00	18.0	13,35	1282.0	950,7	79601
3	Madrid-Arganzuela	2020/07/01 09:00:00	16.0	10,4	1769.0	1150,17	79602
4	Madrid-Chamartín	2020/07/01 09:00:00	12.0	8,23	1800.0	1234,15	79605

Figure 1 COVID-19 data by district initial version

The steps to clean it will be the following:

1. Translate headers from Spanish to English
2. Remove data previous to July, 1st (we will only work with most recent data)
3. Remove data which does not belong to the city of Madrid: since our datasets contains data from all the towns which belongs to Madrid, only data from the districts of Madrid city will remain in the dataset
4. Convert to right data types: due to Spanish format (comma for decimals), some columns which represent numeric data are treated like string. These columns will be converted to number.
5. Remove columns which will not be used

After these steps, our dataset will look like this:

	town_district	total_positive_tests	total_cumulative_rate	district_code
0	Madrid-Retiro	1691.0	1417.23	79603
1	Madrid-Salamanca	1781.0	1218.99	79604
2	Madrid-Centro	1282.0	950.70	79601
3	Madrid-Arganzuela	1769.0	1150.17	79602
4	Madrid-Chamartín	1800.0	1234.15	79605
5	Madrid-Tetuán	2233.0	1413.09	79606
6	Madrid-Chamberí	1739.0	1247.68	79607
7	Madrid-Fuencarral-El Pardo	3290.0	1337.73	79608
8	Madrid-Moncloa-Aravaca	1568.0	1313.02	79609
9	Madrid-Latina	3080.0	1292.93	79610
10	Madrid-Carabanchel	2957.0	1168.32	79611
11	Madrid-Usera	1505.0	1076.99	79612
12	Madrid-Puente de Vallecas	3348.0	1425.55	79613
13	Madrid-San Blas - Canillejas	1584.0	1001.63	79620
14	Madrid-Barajas	464.0	948.00	79621
15	Madrid-Moratalaz	1517.0	1604.10	79614
16	Madrid-Ciudad Lineal	2431.0	1123.99	79615
17	Madrid-Hortaleza	1885.0	1001.66	79616
18	Madrid-Villaverde	1588.0	1066.16	79617
19	Madrid-Villa de Vallecas	1318.0	1194.22	79618
20	Madrid-Vicálvaro	929.0	1288.65	79619

Figure 2 COVID-19 data by district final version

The next dataset, which is "handmade" and contains the coordinates of each district, is shown below:

	district_code	latitude	longitude
0	79604	40.430000	-3.677778
1	79601	40.415278	-3.707500
2	79602	40.402778	-3.695278
3	79605	40.453333	-3.677500
4	79606	40.460556	-3.700000
5	79607	40.432778	-3.697222
6	79608	40.478611	-3.709722
7	79609	40.435470	-3.731700
8	79610	40.402500	-3.741389
9	79611	40.383611	-3.728056
10	79612	40.381389	-3.706944
11	79613	40.393540	-3.662000
12	79620	40.426111	-3.612778
13	79621	40.470278	-3.585000
14	79614	40.410000	-3.644444
15	79615	40.445050	-3.651320
16	79616	40.469444	-3.640556
17	79617	40.345833	-3.709444
18	79618	40.379722	-3.621389
19	79619	40.404167	-3.608056
20	79603	40.408056	-3.676667

Figure 3 Districts' coordinates

Once the two main datasets are loaded and cleaned, it is time to merge both using the column "*district_code*". The result of merge both looks like this:

	town_district	total_positive_tests	total_cumulative_rate	district_code	latitude	longitude
0	Madrid-Retiro	1691.0	1417.23	79603	40.408056	-3.676667
1	Madrid-Salamanca	1781.0	1218.99	79604	40.430000	-3.677778
2	Madrid-Centro	1282.0	950.70	79601	40.415278	-3.707500
3	Madrid-Arganzuela	1769.0	1150.17	79602	40.402778	-3.695278
4	Madrid-Chamartín	1800.0	1234.15	79605	40.453333	-3.677500
5	Madrid-Tetuán	2233.0	1413.09	79606	40.460556	-3.700000
6	Madrid-Chamberí	1739.0	1247.68	79607	40.432778	-3.697222
7	Madrid-Fuencarral-El Pardo	3290.0	1337.73	79608	40.478611	-3.709722
8	Madrid-Moncloa-Aravaca	1568.0	1313.02	79609	40.435470	-3.731700
9	Madrid-Latina	3080.0	1292.93	79610	40.402500	-3.741389
10	Madrid-Carabanchel	2957.0	1168.32	79611	40.383611	-3.728056
11	Madrid-Usera	1505.0	1076.99	79612	40.381389	-3.706944
12	Madrid-Puente de Vallecas	3348.0	1425.55	79613	40.393540	-3.662000
13	Madrid-San Blas - Canillejas	1584.0	1001.63	79620	40.426111	-3.612778
14	Madrid-Barajas	464.0	948.00	79621	40.470278	-3.585000
15	Madrid-Moratalaz	1517.0	1604.10	79614	40.410000	-3.644444
16	Madrid-Ciudad Lineal	2431.0	1123.99	79615	40.445050	-3.651320
17	Madrid-Hortaleza	1885.0	1001.66	79616	40.469444	-3.640556
18	Madrid-Villaverde	1588.0	1066.16	79617	40.345833	-3.709444
19	Madrid-Villa de Vallecas	1318.0	1194.22	79618	40.379722	-3.621389
20	Madrid-Vicálvaro	929.0	1288.65	79619	40.404167	-3.608056

Figure 4 COVID-19 by district with coordinates

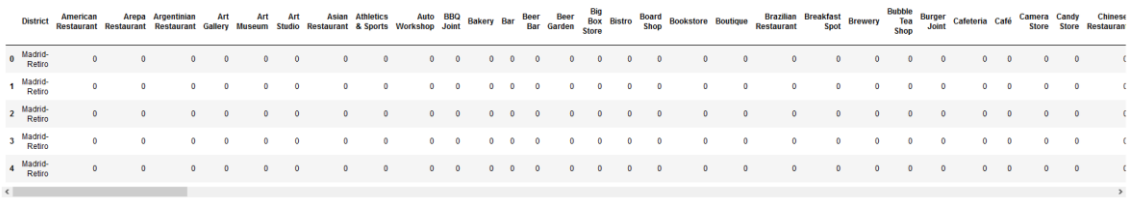
Our third and last dataset comes from *FourSquare*. To achieve the goal of this project, we have called *FourSquare* API to get the nearby venues in a radius of 500 meters for each district.

First of all, we will have a data frame which will have all the venues obtained and the information of each venue, like this:

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Madrid-Retiro	40.408056	-3.676667	Rincon De Extremadura	40.407017	-3.678272	Spanish Restaurant
1	Madrid-Retiro	40.408056	-3.676667	Don Giovanni	40.407648	-3.680188	Italian Restaurant
2	Madrid-Retiro	40.408056	-3.676667	Sidreria Astur II	40.405341	-3.677504	Spanish Restaurant
3	Madrid-Retiro	40.408056	-3.676667	Mezcal	40.406796	-3.677795	Mexican Restaurant
4	Madrid-Retiro	40.408056	-3.676667	Rosaleda del Retiro	40.410987	-3.680377	Garden

Figure 5 Data collected from *FourSquare*

Secondly, some operations will be executed in order to get the most common venue category for each district. The first step will be group venue categories by district:



	District	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art Studio	Asian Restaurant	Athletics & Sports	Auto Workshop	BBQ Joint	Bakery	Bar	Beer Bar	Beer Garden	Big Box Store	Bistro	Board Shop	Bookstore	Boutique	Brazilian Restaurant	Breakfast Spot	Brewery	Bubble Tea Shop	Burger Joint	Cafeteria	Café	Camera Store	Candy Store	Chinese Restauran
0	Madrid-Retiro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Madrid-Retiro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Madrid-Retiro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Madrid-Retiro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Madrid-Retiro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6 Venues categories grouped by district

Next, rows will be grouped again taking the mean of the frequency of occurrence of each venue category:

	District	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art Studio	Asian Restaurant	Athletics & Sports	Auto Workshop	BBQ Joint	Bakery	Bar	Beer Bar	Beer Garden	Big Box Store	Bistro	Board Shop	Bookstore	Boutique	Brazilian Restaurant	Breakfast Spot	Brewery	Bubble Tea Shop	Burger Joint	Ca
0	Madrid-Arganzuela	0.000000	0.023256	0.000000	0.000000	0.000000	0.000000	0.011628	0.000000	0.000000	0.000000	0.046512	0.000000	0.011628	0.023256	0.000	0.000000	0.000000	0.000000	0.000	0.000000	0.023256	0.011628	0.000000	0.023256	0.0
1	Madrid-Barajas	0.000000	0.000000	0.029412	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.029412	0.000000	0.000000	0.000	0.029412	0.000000	0.000000	0.000	0.000000	0.000000	0.029412	0.000000	0.000000	0.0
scroll output; double click to hide			0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.100000	0.0
3	Madrid-Centro	0.010204	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.030612	0.010204	0.000000	0.000	0.020408	0.000000	0.020408	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
4	Madrid-Chamartín	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.018868	0.056604	0.018868	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0
5	Madrid-Chamberí	0.013333	0.000000	0.000000	0.013333	0.013333	0.000000	0.013333	0.000000	0.000000	0.013333	0.026667	0.080000	0.000000	0.000000	0.000	0.000000	0.000000	0.013333	0.000	0.000000	0.013333	0.040000	0.000000	0.026667	0.0

Figure 7 Venues categories mean frequency

For each district, we will get its 10 most popular venue categories and the result will be saved in a data frame, getting the next result:

	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	Madrid-Arganzuela	Spanish Restaurant	Restaurant	Grocery Store	Tapas Restaurant	Bakery	Gym / Fitness Center	Burger Joint	Hotel	Sandwich Place	Falafel Restaurant
1	Madrid-Barajas	Hotel	Restaurant	Spanish Restaurant	Supermarket	Tapas Restaurant	Coffee Shop	Brewery	Fast Food Restaurant	Bistro	Japanese Restaurant
2	Madrid-Carabanchel	Metro Station	Soccer Field	Burger Joint	Fast Food Restaurant	Tapas Restaurant	Nightclub	Bakery	Plaza	Pizza Place	Pharmacy
3	Madrid-Centro	Tapas Restaurant	Spanish Restaurant	Plaza	Hostel	Bar	Ice Cream Shop	Restaurant	Cocktail Bar	Coffee Shop	Pastry Shop
4	Madrid-Chamartín	Spanish Restaurant	Restaurant	Grocery Store	Park	Tapas Restaurant	Bakery	Café	Japanese Restaurant	Gastropub	Coffee Shop
5	Madrid-Chamberí	Spanish Restaurant	Bar	Restaurant	Café	Tapas Restaurant	Brewery	Italian Restaurant	Plaza	Bakery	Coffee Shop
6	Madrid-Ciudad Lineal	Spanish Restaurant	Grocery Store	Hotel	Pizza Place	Italian Restaurant	Tapas Restaurant	Bar	Motorcycle Shop	Fast Food Restaurant	Park
7	Madrid-Fuencarral-El Pardo	Clothing Store	Italian Restaurant	Fast Food Restaurant	Tapas Restaurant	Burger Joint	Sandwich Place	Restaurant	Chocolate Shop	Pizza Place	Pharmacy
8	Madrid-Hortaleza	Pizza Place	Breakfast Spot	Supermarket	Spanish Restaurant	Dance Studio	Pharmacy	Plaza	Clothing Store	Donut Shop	Chinese Restaurant
9	Madrid-Latina	Pizza Place	Falafel Restaurant	Metro Station	Train Station	Food & Drink Shop	Park	Grocery Store	Asian Restaurant	Fast Food Restaurant	Bakery
scroll output; double click to hide		Spanish Restaurant	Gym	Pool	Pub	Stadium	Wine Shop	Falafel Restaurant	Food	Flea Market	Fish Market
11	Madrid-Moratalaz	Park	Bar	Bakery	Food Truck	Café	Nightclub	Soccer Field	Ice Cream Shop	Pizza Place	Plaza
12	Madrid-Puente de Vallecas	Gym	Pub	Pizza Place	Music Venue	Sandwich Place	Spanish Restaurant	Concert Hall	Performing Arts Venue	Camera Store	Café
13	Madrid-Retiro	Spanish Restaurant	Supermarket	Tapas Restaurant	Bar	Grocery Store	Mediterranean Restaurant	Diner	Brewery	Burger Joint	Food & Drink Shop
14	Madrid-Salamanca	Spanish Restaurant	Restaurant	Seafood Restaurant	Mediterranean Restaurant	Coffee Shop	Burger Joint	Supermarket	Tapas Restaurant	Mexican Restaurant	Bakery
15	Madrid-San Blas - Canillejas	Metro Station	Asian Restaurant	Shopping Mall	Snack Place	Supermarket	Gas Station	Pizza Place	Gym	Grocery Store	Dance Studio
16	Madrid-Tetuán	Spanish Restaurant	Bakery	Coffee Shop	Supermarket	Chinese Restaurant	Brazilian Restaurant	Grocery Store	Farmers Market	Pizza Place	Gym / Fitness Center
17	Madrid-Usera	Seafood Restaurant	Bubble Tea Shop	Spanish Restaurant	Asian Restaurant	Theater	Noodle House	Fast Food Restaurant	Chinese Restaurant	Wine Shop	Farmers Market
18	Madrid-Vicálvaro	Pizza Place	Spanish Restaurant	Beer Bar	Fast Food Restaurant	Sandwich Place	Café	Breakfast Spot	Metro Station	Grocery Store	Ice Cream Shop
19	Madrid-Villa de Vallecas	Metro Station	Grocery Store	Spanish Restaurant	Food	Soccer Field	Train Station	Department Store	Dessert Shop	Fast Food Restaurant	Food Court
20	Madrid-Villaverde	Pizza Place	Spanish Restaurant	Train	Diner	Falafel Restaurant	Food & Drink Shop	Food	Flea Market	Fish Market	Fast Food Restaurant

Figure 8 Ten most popular venue categories by district

3. Methodology

Once our datasets are ready to use, it's time to start to work seriously.

3.1. Classifying districts

We will start classifying each district by its COVID-19 total cumulative rate. We will define 3 categories, which will correspond to low incidence, medium incidence and high incidence. To achieve this, a range will be generated: it will have 3 categories starting by the lowest total cumulative rate and ending by the highest cumulative rate. Using the function "*linspace*" from "*numpy*" library, we will get this range. As soon as this range is created, the districts can be labelled depending on their total cumulative rate.

The districts with their coordinates and their labels by total cumulative rate are shown below:

	town_district	total_positive_tests	total_cumulative_rate	district_code	latitude	longitude	total_cumulative_rate_label
0	Madrid-Retiro	1691.0	1417.23	79603	40.408056	-3.676667	2
1	Madrid-Salamanca	1781.0	1218.99	79604	40.430000	-3.677778	1
2	Madrid-Centro	1282.0	950.70	79601	40.415278	-3.707500	0
3	Madrid-Arganzuela	1769.0	1150.17	79602	40.402778	-3.695278	0
4	Madrid-Chamartín	1800.0	1234.15	79605	40.453333	-3.677500	1
5	Madrid-Tetuán	2233.0	1413.09	79606	40.460556	-3.700000	2
6	Madrid-Chamberí	1739.0	1247.68	79607	40.432778	-3.697222	1
7	Madrid-Fuencarral-El Pardo	3290.0	1337.73	79608	40.478611	-3.709722	1
8	Madrid-Moncloa-Aravaca	1568.0	1313.02	79609	40.435470	-3.731700	1
9	Madrid-Latina	3080.0	1292.93	79610	40.402500	-3.741389	1
10	Madrid-Carabanchel	2957.0	1168.32	79611	40.383611	-3.728056	1
11	Madrid-Usera	1505.0	1076.99	79612	40.381389	-3.706944	0
12	Madrid-Puente de Vallecas	3348.0	1425.55	79613	40.393540	-3.662000	2
13	Madrid-San Blas - Canillejas	1584.0	1001.63	79620	40.426111	-3.612778	0
14	Madrid-Barajas	464.0	948.00	79621	40.470278	-3.585000	0
15	Madrid-Moratalaz	1517.0	1604.10	79614	40.410000	-3.644444	2
16	Madrid-Ciudad Lineal	2431.0	1123.99	79615	40.445050	-3.651320	0
17	Madrid-Hortaleza	1885.0	1001.66	79616	40.469444	-3.640556	0
18	Madrid-Villaverde	1588.0	1066.16	79617	40.345833	-3.709444	0
19	Madrid-Villa de Vallecas	1318.0	1194.22	79618	40.379722	-3.621389	1
20	Madrid-Vicálvaro	929.0	1288.65	79619	40.404167	-3.608056	1

Figure 9 Districts labelled by their COVID-19 rate

In order to visualize this information, we will use folium and choropleth. This will allow us, on the one hand, to see the districts over a map and, on the other hand, to visualize the distribution of COVID-19 total cumulative rate. The result is as follows:

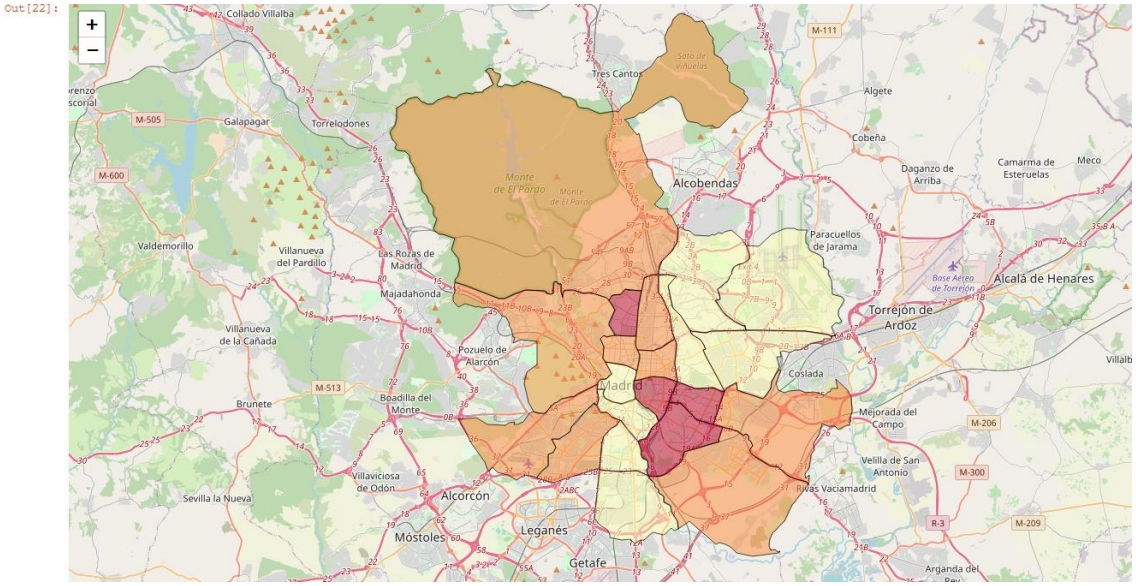


Figure 10 Map of districts with their COVID-19 rate

Now, and in order to double-check the coordinates, we will add markers to our map:

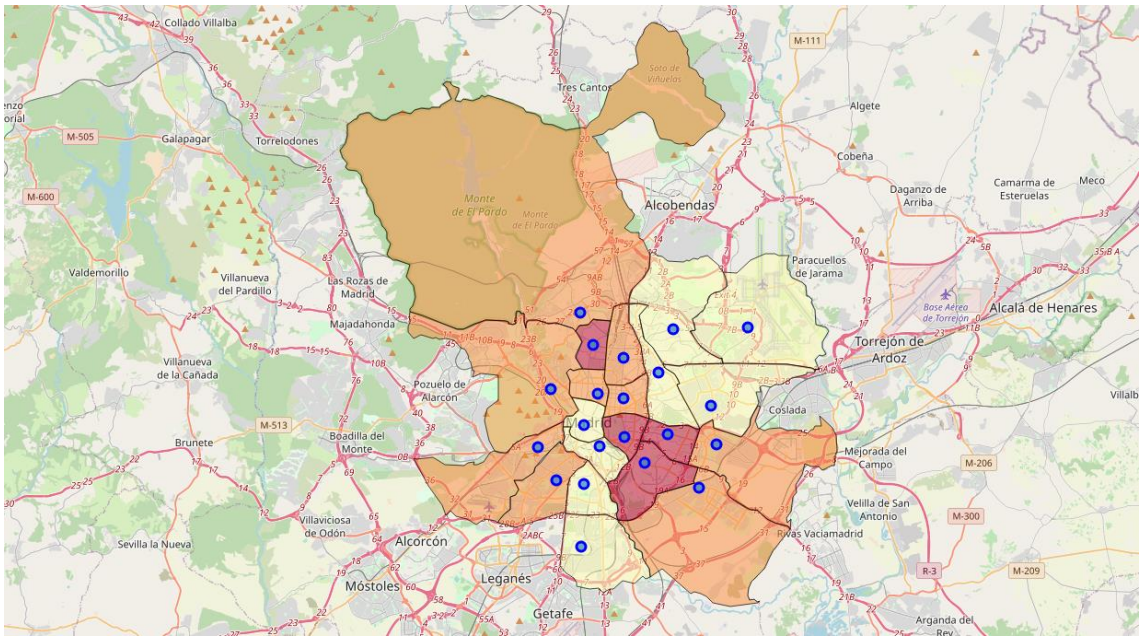


Figure 11 Map of districts with their COVID-19 rate and markers

3.2. Analysing data from FourSquare

Before applying machine learning to our datasets, we will analyse the data obtained from *FourSquare*.

As a reminder, the data frame which contains *FourSquare* information looks like this:

	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	Madrid-Arganzuela	Spanish Restaurant	Restaurant	Grocery Store	Tapas Restaurant	Bakery	Gym / Fitness Center	Burger Joint	Hotel	Sandwich Place	Falafel Restaurant
1	Madrid-Barajas	Hotel	Restaurant	Spanish Restaurant	Supermarket	Tapas Restaurant	Coffee Shop	Brewery	Fast Food Restaurant	Bistro	Japanese Restaurant
2	Madrid-Carabanchel	Metro Station	Soccer Field	Burger Joint	Fast Food Restaurant	Tapas Restaurant	Nightclub	Bakery	Plaza	Pizza Place	Pharmacy
3	Madrid-Centro	Tapas Restaurant	Spanish Restaurant	Plaza	Hostel	Bar	Ice Cream Shop	Restaurant	Cocktail Bar	Coffee Shop	Pastry Shop
4	Madrid-Chamartín	Spanish Restaurant	Restaurant	Grocery Store	Park	Tapas Restaurant	Bakery	Café	Japanese Restaurant	Gastropub	Coffee Shop
5	Madrid-Chamberí	Spanish Restaurant	Bar	Restaurant	Café	Tapas Restaurant	Brewery	Italian Restaurant	Plaza	Bakery	Coffee Shop
6	Madrid-Ciudad Lineal	Spanish Restaurant	Grocery Store	Hotel	Pizza Place	Italian Restaurant	Tapas Restaurant	Bar	Motorcycle Shop	Fast Food Restaurant	Park
7	Madrid-Fuencarral-El Pardo	Clothing Store	Italian Restaurant	Fast Food Restaurant	Tapas Restaurant	Burger Joint	Sandwich Place	Restaurant	Chocolate Shop	Pizza Place	Pharmacy
8	Madrid-Hortaleza	Pizza Place	Breakfast Spot	Supermarket	Spanish Restaurant	Dance Studio	Pharmacy	Plaza	Clothing Store	Donut Shop	Chinese Restaurant
9	Madrid-Latina	Pizza Place	Falafel Restaurant	Metro Station	Train Station	Food & Drink Shop	Park	Grocery Store	Asian Restaurant	Fast Food Restaurant	Bakery
scroll output; double click to hide		Spanish Restaurant	Gym	Pool	Pub	Stadium	Wine Shop	Falafel Restaurant	Food	Flea Market	Fish Market
11	Madrid-Moratalaz	Park	Bar	Bakery	Food Truck	Café	Nightclub	Soccer Field	Ice Cream Shop	Pizza Place	Plaza
12	Madrid-Puente de Vallecas	Gym	Pub	Pizza Place	Music Venue	Sandwich Place	Spanish Restaurant	Concert Hall	Performing Arts Venue	Camera Store	Café
13	Madrid-Retiro	Spanish Restaurant	Supermarket	Tapas Restaurant	Bar	Grocery Store	Mediterranean Restaurant	Diner	Brewery	Burger Joint	Food & Drink Shop
14	Madrid-Salamanca	Spanish Restaurant	Restaurant	Seafood Restaurant	Mediterranean Restaurant	Coffee Shop	Burger Joint	Supermarket	Tapas Restaurant	Mexican Restaurant	Bakery
15	Madrid-San Blas - Canillejas	Metro Station	Asian Restaurant	Shopping Mall	Snack Place	Supermarket	Gas Station	Pizza Place	Gym	Grocery Store	Dance Studio
16	Madrid-Tetuán	Spanish Restaurant	Bakery	Coffee Shop	Supermarket	Chinese Restaurant	Brazilian Restaurant	Grocery Store	Farmers Market	Pizza Place	Gym / Fitness Center
17	Madrid-Usera	Seafood Restaurant	Bubble Tea Shop	Spanish Restaurant	Asian Restaurant	Theater	Noodle House	Fast Food Restaurant	Chinese Restaurant	Wine Shop	Farmers Market
18	Madrid-Vicálvaro	Pizza Place	Spanish Restaurant	Beer Bar	Fast Food Restaurant	Sandwich Place	Café	Breakfast Spot	Metro Station	Grocery Store	Ice Cream Shop
19	Madrid-Villa de Vallecas	Metro Station	Grocery Store	Spanish Restaurant	Food	Soccer Field	Train Station	Department Store	Dessert Shop	Fast Food Restaurant	Food Court
20	Madrid-Villaverde	Pizza Place	Spanish Restaurant	Train	Diner	Falafel Restaurant	Food & Drink Shop	Food	Flea Market	Fish Market	Fast Food Restaurant

Figure 12 Ten most popular venue categories by district

As you can see, the most common categories in all districts are related with restaurant business (Spanish restaurants, Italian restaurants, tapas restaurants, nightclubs, cafés, etc.), which can mean a handicap for our goal. However, it's better not to hurry up.

3.3. Applying machine learning

Using their most common venues, we will classify the districts using K-means algorithm.

Having four clusters, the result after applying K-means looks as follows:

	Cluster Labels	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	3	Madrid-Arganzuela	Spanish Restaurant	Restaurant	Grocery Store	Tapas Restaurant	Bakery	Gym / Fitness Center	Burger Joint	Hotel	Sandwich Place	Falafel Restaurant
1	3	Madrid-Barajas	Hotel	Restaurant	Spanish Restaurant	Supermarket	Tapas Restaurant	Coffee Shop	Brewery	Fast Food Restaurant	Bistro	Japanese Restaurant
2	0	Madrid-Carabanchel	Metro Station	Soccer Field	Burger Joint	Fast Food Restaurant	Tapas Restaurant	Nightclub	Bakery	Plaza	Pizza Place	Pharmacy
3	3	Madrid-Centro	Tapas Restaurant	Spanish Restaurant	Plaza	Hostel	Bar	Ice Cream Shop	Restaurant	Cocktail Bar	Coffee Shop	Pastry Shop
4	3	Madrid-Chamartín	Spanish Restaurant	Restaurant	Grocery Store	Park	Tapas Restaurant	Bakery	Café	Japanese Restaurant	Gastropub	Coffee Shop
5	3	Madrid-Chamberí	Spanish Restaurant	Bar	Restaurant	Café	Tapas Restaurant	Brewery	Italian Restaurant	Plaza	Bakery	Coffee Shop
6	3	Madrid-Ciudad Lineal	Spanish Restaurant	Grocery Store	Hotel	Pizza Place	Italian Restaurant	Tapas Restaurant	Bar	Motorcycle Shop	Fast Food Restaurant	Park
7	0	Madrid-Fuencarral-El Pardo	Clothing Store	Italian Restaurant	Fast Food Restaurant	Tapas Restaurant	Burger Joint	Sandwich Place	Restaurant	Chocolate Shop	Pizza Place	Pharmacy
8	3	Madrid-Hortaleza	Pizza Place	Breakfast Spot	Supermarket	Spanish Restaurant	Dance Studio	Pharmacy	Plaza	Clothing Store	Donut Shop	Chinese Restaurant
9	0	Madrid-Latina	Pizza Place	Falafel Restaurant	Metro Station	Train Station	Food & Drink Shop	Park	Grocery Store	Asian Restaurant	Fast Food Restaurant	Bakery
10	1	Madrid-Moncloa-Aravaca	Spanish Restaurant	Gym	Pool	Pub	Stadium	Wine Shop	Falafel Restaurant	Food	Flea Market	Fish Market
11	0	Madrid-Moratalaz	Park	Bar	Bakery	Food Truck	Café	Nightclub	Soccer Field	Ice Cream Shop	Pizza Place	Plaza
12	0	Madrid-Puente de Vallecas	Gym	Pub	Pizza Place	Music Venue	Sandwich Place	Spanish Restaurant	Concert Hall	Performing Arts Venue	Camera Store	Café
13	3	Madrid-Retiro	Spanish Restaurant	Supermarket	Tapas Restaurant	Bar	Grocery Store	Mediterranean Restaurant	Diner	Brewery	Burger Joint	Food & Drink Shop
14	3	Madrid-Salamanca	Spanish Restaurant	Restaurant	Seafood Restaurant	Mediterranean Restaurant	Coffee Shop	Burger Joint	Supermarket	Tapas Restaurant	Mexican Restaurant	Bakery
15	0	Madrid-San Blas - Canillejas	Metro Station	Asian Restaurant	Shopping Mall	Snack Place	Supermarket	Gas Station	Pizza Place	Gym	Grocery Store	Dance Studio
16	3	Madrid-Tetuán	Spanish Restaurant	Bakery	Coffee Shop	Supermarket	Chinese Restaurant	Brazilian Restaurant	Grocery Store	Farmers Market	Pizza Place	Gym / Fitness Center
17	3	Madrid-Usera	Seafood Restaurant	Bubble Tea Shop	Spanish Restaurant	Asian Restaurant	Theater	Noodle House	Fast Food Restaurant	Chinese Restaurant	Wine Shop	Farmers Market
18	0	Madrid-Vicálvaro	Pizza Place	Spanish Restaurant	Beer Bar	Fast Food Restaurant	Sandwich Place	Café	Breakfast Spot	Metro Station	Grocery Store	Ice Cream Shop
19	0	Madrid-Villa de Vallecas	Metro Station	Grocery Store	Spanish Restaurant	Food	Soccer Field	Train Station	Department Store	Dessert Shop	Fast Food Restaurant	Food Court
20	2	Madrid-Villaverde	Pizza Place	Spanish Restaurant	Train	Diner	Falafel Restaurant	Food & Drink Shop	Food	Flea Market	Fish Market	Fast Food Restaurant

Figure 13 ***Districts clustered using K-means algorithm***

As we can see, "3" is the most common cluster, followed by cluster "0" and leaving "1" and "2" in last places.

In order to make out what have in common each cluster, we will try to visualize each cluster in a graph. Using *matplotlib*, the result is shown below:

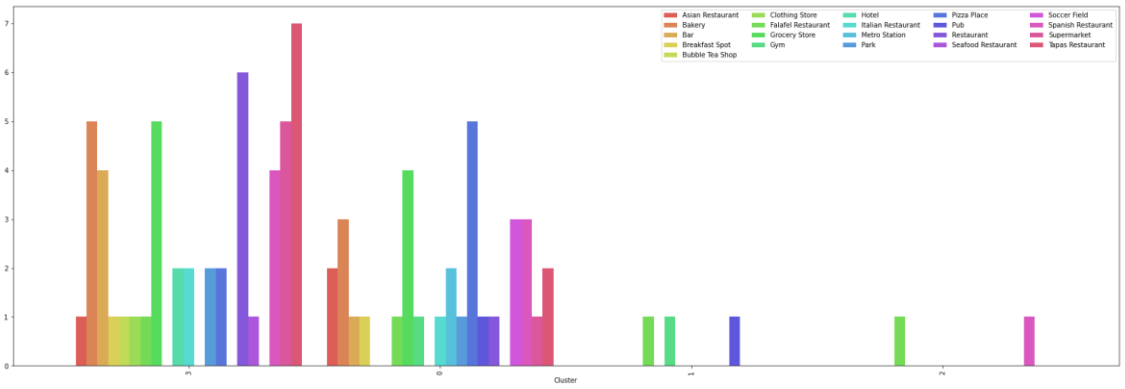


Figure 14 K-means graph

To create this graph, we have used the 1st, the 2nd and the 3rd most common venues for each district. As it was mentioned before, data collected from *FourSquare* reveals that most common venues in every category are relate to restaurants industry: only 6 put of 21 are not bars, restaurants cafés or similar.

Studying each cluster separately, we can come to some conclusions:

- That cluster "3" groups the most varied categories of venues (restaurants, bakery shops, hotels, metro stations, soccer fields and supermarkets). This cluster would probably correspond to districts very touristic: apart from places usually frequented by residents (bakery and grocery stores), places crowded by tourists are very common.
- That cluster "0" has a lot in common with cluster "3" but, in this case, parks are the most common venues, followed by falafel restaurants and bakeries. Besides, there are gyms. Keeping that in mind, the districts from these clusters would probably be less touristic than districts of cluster "3"

- That cluster “1” has a kind of restaurant, but also gyms and pubs. We can suppose that this cluster correspond to districts with more residents than tourists. We can also imagine that venues in these districts are less crowded than venues in previous clusters.
- That cluster “2” also has “Falafel restaurant” as one of their most common venues but, in this case, “Spanish restaurant” are the other more popular venue.

With this information, and knowing how COVID-19 spreads, we can expect that those districts which belong to cluster “3” and “0” have a high total cumulative rate. To confirm this, we will have to see together the districts labelled by COVID-19 incidence and the districts clustered by their most common venues.

4. Results

As we mentioned before, in order to know if COVID-19 infections are related with some kind of places, we need to see together the districts labelled by COVID-19 incidence and the districts clustered by their most common venues.

Merging the data frame which contains the districts labelled by the number of infections and the data frame which contains the districts clustered by their most common venues, the resulting one looks as follows:

	town_district	total_positive_tests	total_cumulative_rate	district_code	latitude	longitude	total_cumulative_rate_label	Cluster Labels	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	Madrid-Retiro	1691.0	1417.23	79603	40.408056	-3.676667	2	3	Spanish Restaurant	Supermarket	Tapas Restaurant	Bar	Grocery Store	Mediterranean Restaurant	Diner	Brewery	Burger Joint	Food & Drink Shop
1	Madrid-Salamanca	1781.0	1218.99	79604	40.430000	-3.677778	1	3	Spanish Restaurant	Restaurant	Seafood Restaurant	Mediterranean Restaurant	Coffee Shop	Burger Joint	Supermarket	Tapas Restaurant	Mexican Restaurant	Bakery
2	Madrid-Centro	1282.0	950.70	79601	40.415278	-3.707500	0	3	Tapas Restaurant	Spanish Restaurant	Plaza	Hostel	Bar	Ice Cream Shop	Restaurant	Cocktail Bar	Coffee Shop	Pastry Shop
3	Madrid-Arganzuela	1769.0	1150.17	79602	40.402778	-3.695278	0	3	Spanish Restaurant	Restaurant	Grocery Store	Tapas Restaurant	Bakery	Gym / Fitness Center	Burger Joint	Hotel	Sandwich Place	Falafel Restaurant
4	Madrid-Chamartín	1800.0	1234.15	79605	40.453333	-3.677500	1	3	Spanish Restaurant	Restaurant	Grocery Store	Park	Tapas Restaurant	Bakery	Café	Japanese Restaurant	Gastropub	Coffee Shop
5	Madrid-Tetuán	2233.0	1413.09	79606	40.460556	-3.700000	2	3	Spanish Restaurant	Bakery	Coffee Shop	Supermarket	Chinese Restaurant	Brazilian Restaurant	Grocery Store	Farmers Market	Pizza Place	Gym / Fitness Center
6	Madrid-Chamberí	1739.0	1247.68	79607	40.432778	-3.697222	1	3	Spanish Restaurant	Bar	Restaurant	Café	Tapas Restaurant	Brewery	Italian Restaurant	Plaza	Bakery	Coffee Shop
7	Fuencarral-El Pardo	3290.0	1337.73	79608	40.478611	-3.709722	1	0	Clothing Store	Italian Restaurant	Fast Food Restaurant	Tapas Restaurant	Burger Joint	Sandwich Place	Restaurant	Chocolate Shop	Pizza Place	Pharmacy
8	Madrid-Moncloa-Aravaca	1568.0	1313.02	79609	40.435470	-3.731700	1	1	Spanish Restaurant	Gym	Pool	Pub	Stadium	Wine Shop	Falafel Restaurant	Food	Flea Market	Fish Market
9	Madrid-Latina	3080.0	1292.93	79610	40.402500	-3.741389	1	0	Pizza Place	Falafel Restaurant	Metro Station	Train Station	Food & Drink Shop	Park	Grocery Store	Asian Restaurant	Fast Food Restaurant	Bakery
10	Madrid-Carabanchel	2957.0	1168.32	79611	40.383611	-3.728056	1	0	Metro Station	Soccer Field	Burger Joint	Fast Food Restaurant	Tapas Restaurant	Nightclub	Bakery	Plaza	Pizza Place	Pharmacy
11	Madrid-Usera	1505.0	1076.99	79612	40.381389	-3.706944	0	3	Seafood Restaurant	Bubble Tea Shop	Spanish Restaurant	Asian Restaurant	Theater	Noodle House	Fast Food Restaurant	Chinese Restaurant	Wine Shop	Farmers Market
12	Madrid-Puente de Vallecas	3348.0	1425.55	79613	40.393540	-3.662000	2	0	Gym	Pub	Pizza Place	Music Venue	Sandwich Place	Spanish Restaurant	Concert Hall	Performing Arts Venue	Camera Store	Café
13	Madrid-San Blas -	1584.0	1001.63	79620	40.426111	-3.612778	0	0	Metro Station	Asian Restaurant	Shopping Mall	Snack Place	Supermarket	Gas Station	Pizza Place	Gym	Grocery Store	Dance Studio

Figure 15 Districts labelled by number of infections and clustered by venue categories

If we visualize this data frame over a map, where “green” marks correspond to cluster “3”, “red” marks to cluster “0”, “purple” marks to cluster “1” and “blue” marks to cluster “2”:

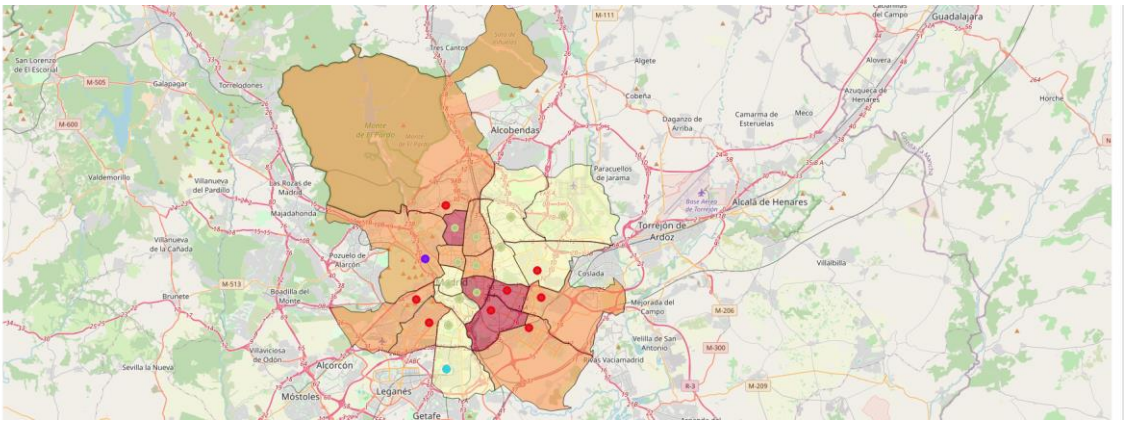


Figure 16 Map of districts labelled by number of infections and clustered by venue categories

5. Discussion

As we can see in the results, it is not possible to state by them which kinds of venues are the most dangerous when we talk about getting infected by COVID-19.

As we had supposed, districts which belong to clusters "3" and "0" are those which have more infections by COVID-19. This is probably due to the great variety of venues in those clusters and the traffic of people between them (not only tourists but also residents). It should be pointed out that the most common venues in these clusters were restaurants, but also there were hotels and metro stations.

On the contrary, districts which belong to clusters "1" and "2" (these were the clusters with less variety of venues) do not have high number of infections by COVID-19. Therefore, in this case we can conclude that results show what we expected: those areas with less crowded venues are safer than those which are overcrowded.

6. Conclusion

Although I was very enthusiastic when I started this project, the more I progressed, the less enthusiastic I was.

The first problems started when I tried to get data from public datasets. Although I found a lot of sources to get information about COVID-19 infections, geographical data was much more difficult to find.

Once I got all the datasets I needed, I noticed *FourSquare* might not provide the information I wanted. I would have preferred to get the most popular venues in a period of time in order to filter them by the previous months of the first breakout, but *FourSquare* API does not offer this possibility.

Apart from this problem, *FourSquare* might not be the best tool for this kind of analysis due to its use. In other words, people who frequently visit a venue (for example, the restaurant where you daily eat or the supermarket where you go every week) do not check-in in *FourSquare*. Tourists are probably the users who most use this application, and here we face another problem: tourists do not count as Madrid residents; therefore they do not count in Madrid statistics about COVID-19.

The very last problem is related to Spanish culture: as we all know (and *FourSquare* has also reminded us) restaurants and bars are one of the most common businesses in the country. And this supposes another problem: while *FourSquare* differentiates between, for example, Italian restaurants and Chinese Restaurants, for our purpose these categories should be the same: restaurants. In spite of this problem, this cultural characteristic might be one of the principal reasons for the great amount of COVID-19 infections during the first wave in Spain: this kind of venues is usually overcrowded and, during the first months of the year, people are indoors.

To draw the conclusion, I must admit that I expected much more about the results of this project. Perhaps using a dataset different from the provided by *FourSquare* such Google Maps, the information collected by Madrid Underground and/or using the population density, the mean income and the number of residents per home by district would have brought better results.