Capstone Project

Open a hotel in Madrid

Introduction

Most of the tourists that Spain receives go to sun and beach destinations, however, in recent years, Madrid (capital of Spain) has tried to grow this number of tourists using other skills such as museums, parks and leisure.

In 2019 more than 10 million tourists arrived in Madrid, 2.1% more than last year and spent an average of about 270 euros a day, 12.6% more than the previous year[1].

With these data of possible future growth it could be interesting to open a hotel in the city of Madrid, however it is difficult to choose where.

Madrid could be an interesting place within Spain to open a hotel since compared to other communities it has a much lower number of hotels.

For example, the case of Galicia, with 2.7 million inhabitants, has an average of 1,409 hotels open, while Madrid, with 6.6 million inhabitants, has an average of 1,169 open hotels[2].

Madrid also has the incentive that it has the Barajas airport, through which 57.8 million passengers passed in 2018.

Principal problem

Madrid has 21 neighborhoods and depending on the hotel we plan to build, it will be better to choose one or the other. For example, building a hotel in Barajas where Madrid Airport is located may not attract young people who prefer to spend some night at a party in the center of Madrid, but it does attract some of the millions of travelers that we have previously commented that They spend the year at the airport to spend a night at the hotel. On the other hand, a family that plans to spend a few days visiting the parks and museums and eating in the restaurants may want to get away from the center a little to avoid night noise but stay relatively close to visit these places that are mostly in the center.

At the same time, doubts arise about which of these 21 neighborhoods have the greatest number of hotels since, if there is an excess, it is very likely that our hotel that has just opened will have difficulty making a place for itself among the others.

Data

Necessary information

1. To solve this problem we will need the neighborhoods of Madrid. It is essential since we want to know in which of these it may make more sense to build our hotel.
2. Coordinates of the neighborhoods of Madrid, we need these data to join it with the previous information and to be able to obtain the next step.
3. Information of all venues in Madrid, we will segment this information with the neighborhoods to carry out the analysis. It is key to our project with information about hotels, plazas, parks, museums, restaurants ... since in this information we will base our subsequent analysis

Data sources

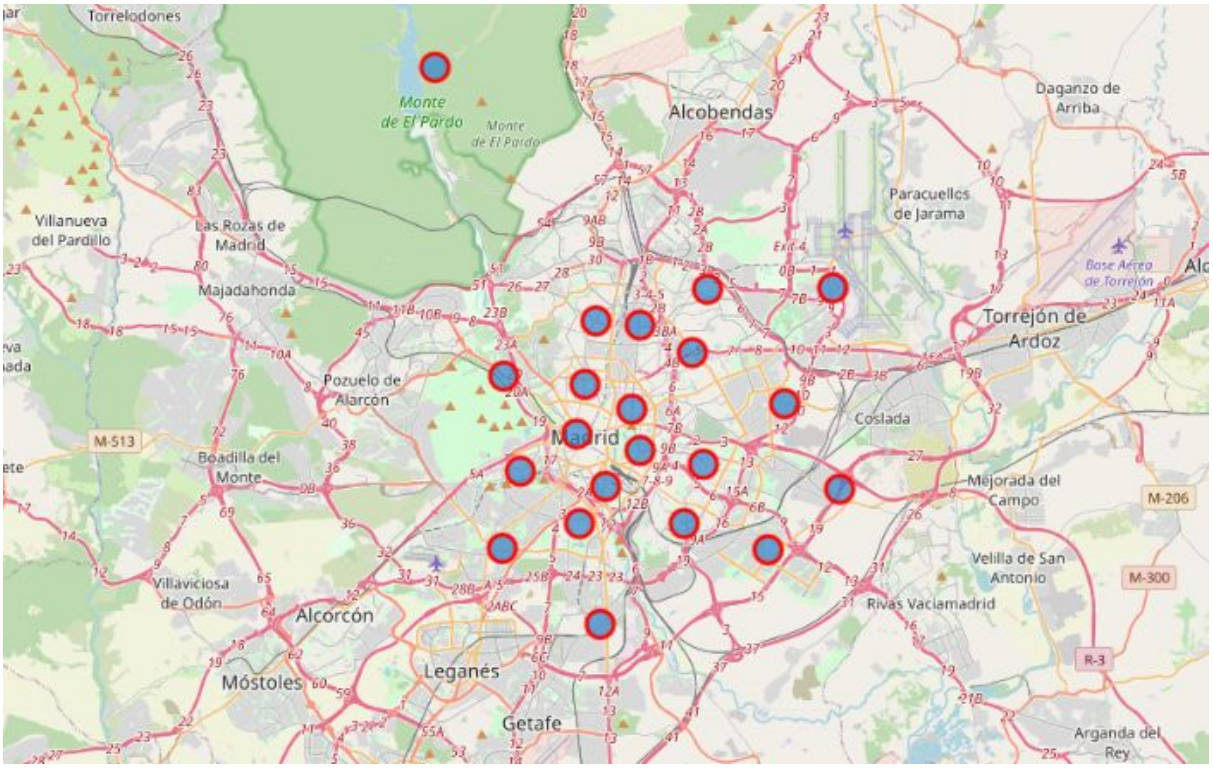
1. We will obtain the data of the neighborhoods of Madrid from Wikipedia, specifically from https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid will have to do scraping of the page to be able to obtain them. Specifically, we will use Python together with the BeautifuSoup package to obtain the data.
2. Once we have obtained the neighborhoods we will use the Python Geocoder tool to obtain the longitude and latitude of each neighborhood.
3. Finally, once we have linked the information of each neighborhood with its coordinates, we will use the Foursquare API to obtain the information of all venues of each neighborhood. Foursquare will give us the category of each venue, which will allow us to analyze relevant information for our project, such as how many hotels there are per neighborhood, parks, museums ...

Methodology

Exploratory data analysis

The first thing we needed was to obtain the neighborhoods of Madrid, we found them in Wikipedia and as we have commented in the previous section, using scraping with the BeautifulSoup tool we were able to obtain them.

The next step was to obtain the coordinates of each neighborhood, in this way we could later join each neighborhood with its venues. Once we got each neighborhood in Madrid with its coordinates, using the folium tool we could see how each neighborhood was geographically located:



This step is interesting before carrying out the analysis since we can see how there are some neighborhoods such as: Centro, Salamanca or Retiro very central. Others like Fuencarral-El pardo or Vicálvaro peripherals and others like Barajas or Hortaleza near the airport.

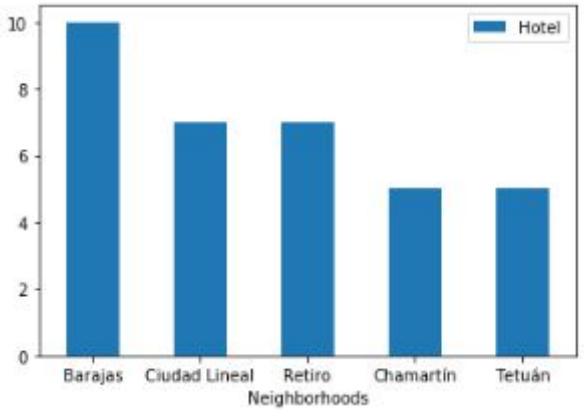
Once we have the neighborhoods of Madrid with their coordinates, we can use the foursquare API to mix the data of each neighborhood and its venues, in this way we

can analyze how many shops and places of interest there are in each neighborhood, as well as how many hotels there are. for further analysis later.

It is important to note that since the account we have on Foursquare is limited, the sample that has been analyzed in this project has been the same size as that used in the laboratories (100 venues) and a radius of 2 km. If a more precise analysis is wanted in the future, this size could be increased.

Once the foursquare API was used, I obtained a table with one row for each venue, in order to better manage this information, I made a grouping by neighborhood and thus being able to visualize each neighborhood with all the venues categories and the number of each one.

If we transform the table in such a way that we filter only by hotels and we use a simple bar graph to be able to visually analyze which neighborhoods have the most hotels, we obtain the following results:



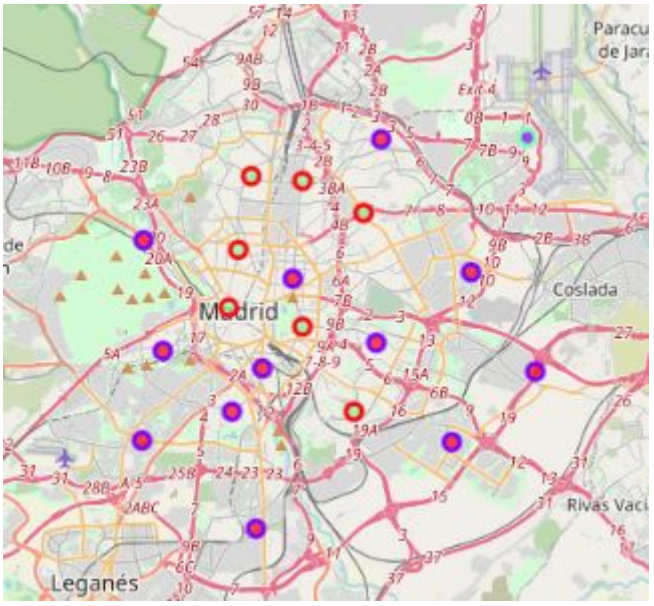
(The five neighborhoods with the largest number of hotels)

We can see how Barajas has a higher concentration of hotels than the rest, as we mentioned at the beginning, since more than 50 million travelers pass through the Barajas airport a year, it is very possible that this area has enough hotels to accommodate some of these travelers who, for example, want to stop in Madrid and spend the night near the airport to catch their next flight.

Clustering

The next step was to group the neighborhoods of Madrid, using k-means. This was done in order to see which groups of neighborhoods have the highest concentration of hotels.

Grouping in 3 clusters I obtained the following results seen on the Madrid map:



(Result of the 3 clusters located on the Madrid Map)

The result of grouping by clusters obtained the following results:

- Group 0: Neighborhoods with a high concentration of hotels were grouped in this cluster. As we can see on the Madrid map (in red color), most of these points are located in the center of Madrid and in tourist places. In this group are:

	Neighborhoods	Hotel	Latitude	Longitude	Cluster Labels
3	Centro	0.040000	40.417653	-3.707914	0
4	Chamartín	0.050000	40.458987	-3.676129	0
5	Chamberí	0.040000	40.436247	-3.703830	0
6	Ciudad Lineal	0.070000	40.448431	-3.650495	0
11	Puente de Vallecas	0.036145	40.383553	-3.654535	0
12	Retiro	0.070000	40.411150	-3.676057	0
15	Tetuán	0.050000	40.460578	-3.698281	0

(Table with the neighborhoods of Madrid classified in group 0)

- Group 1: Group 1 classified the majority of neighborhoods furthest from the center (purple on the map) and therefore more residential. With the exception of the Salamanca neighborhood, the rest of the neighborhoods are of this type. These are the neighborhoods with the fewest hotels.

	Neighborhoods	Hotel	Latitude	Longitude	Cluster Labels
0	Arganzuela	0.010000	40.398068	-3.693734	1
2	Carabanchel	0.010000	40.374211	-3.744676	1
7	Hortaleza	0.022727	40.472549	-3.642551	1
8	Latina	0.010000	40.403532	-3.736152	1
9	Moncloa-Aravaca	0.000000	40.439495	-3.744204	1
10	Moratalaz	0.000000	40.405933	-3.644874	1
13	Salamanca	0.020000	40.427045	-3.680602	1
14	San Blas-Canillejas	0.014493	40.428919	-3.604002	1
16	Usera	0.000000	40.383894	-3.706446	1
17	Vicálvaro	0.000000	40.396584	-3.576622	1
18	Villa de Vallecas	0.011628	40.373958	-3.612163	1
19	Villaverde	0.029412	40.345610	-3.695956	1

(Table with the neighborhoods of Madrid classified in group 1)

- Group 2: Finally, the k-means algorithm has classified the Barajas neighborhood as the only neighborhood in this group (blue color at the top right), this is the neighborhood with the largest number of hotels.

	Neighborhoods	Hotel	Latitude	Longitude	Cluster Labels
1	Barajas	0.121951	40.473318	-3.579845	2

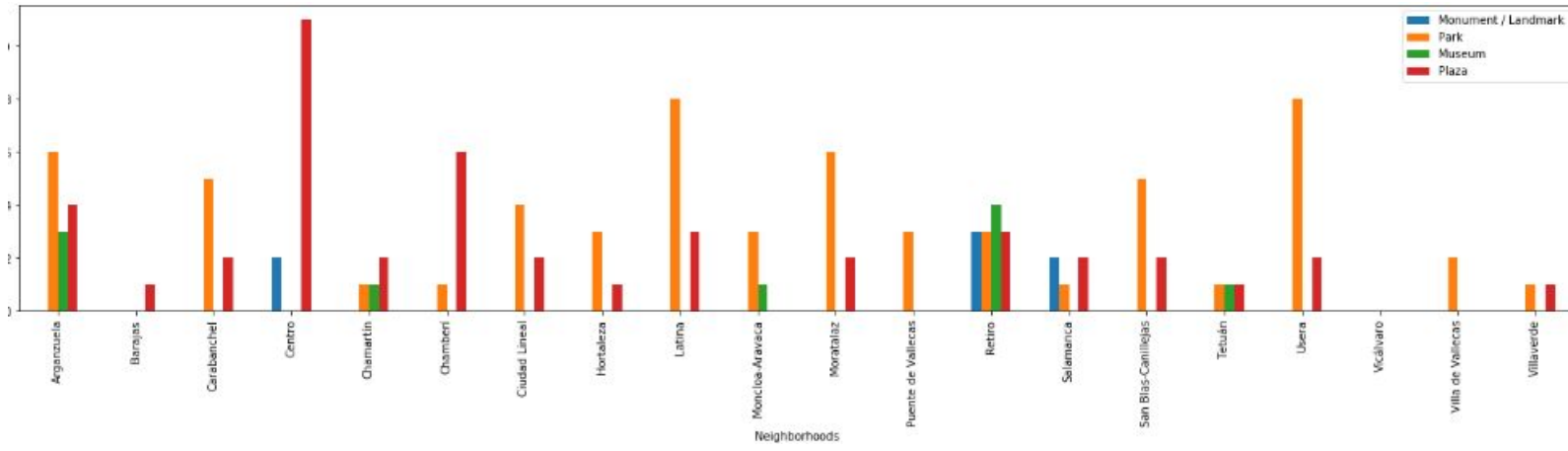
(Table with the neighborhoods of Madrid classified in group 0)

Discussion

We have started talking about the importance that it could have when opening an airport in Madrid to locate this one near the airport since, we could receive part of those 50 million passengers that pass through the airport each year. However, given the results obtained, we can verify that the Barajas neighborhood that is located next to the airport has been classified as the cluster with the largest number of hotels and therefore, if we open a hotel here, it could have great competition. Therefore this neighborhood would not be one of the best options to open our hotel.

The rest of the remaining neighborhoods (20) are classified in groups 0 and 1, as we have seen previously, the neighborhoods in group 1 are the ones with the fewest number of hotels, so it could be interesting to choose one of the neighborhoods in this cluster.

When traveling, most of the time we usually choose the location of the hotel depending on what we are going to do in the destination. That is why using the data provided by the foursquare API we can analyze what these neighborhoods offer us. Since Madrid does not have a beach, one of the main activities carried out by tourists who travel to this city is visiting the parks, museums, squares and monuments. Using a bar graph we can see what each neighborhood in Madrid offers us in this sense

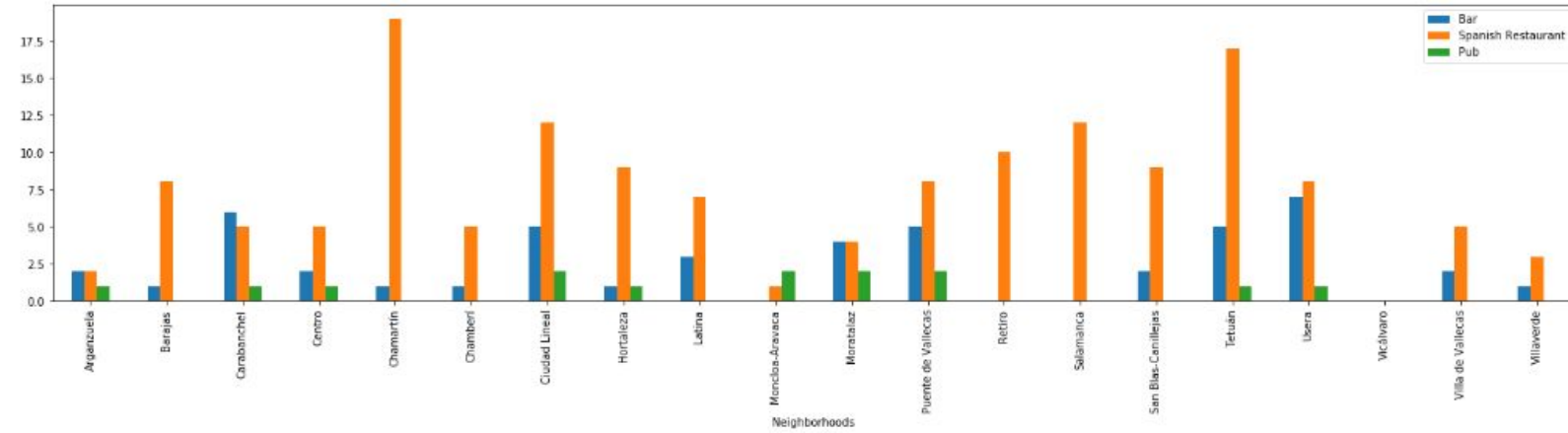


Thanks to this graph we can see how, as we had previously commented, Barajas is the neighborhood with the largest number of hotels thanks to the airport, since it is one of the neighborhoods with the least number of parks, museums or squares.

Very interesting neighborhoods like Retiro appear, in the previous map we can see that it is a neighborhood near the center and although it is located in the intermediate group of neighborhoods with hotels, it has a very varied offer of parks, museums ...

Two places that appear well positioned for a possible opening of a hotel could be Salamanca and Arganzuela, although they have a smaller number of museums and parks than Retiro, they have the advantage that k-means has classified these two neighborhoods within the neighborhoods with the lowest number of hotels. They are

also two fairly central neighborhoods and therefore are closer to neighborhoods like Centro, which are the ones with the greatest number of museums and parks.



Many tourists travel to Madrid to go to a pub, bar or restaurant. That is why it is also interesting to see that neighborhoods offer us the most of these venues.

We can see thanks to the graph that there are quite a few neighborhoods in which the restaurant clearly predominates but there are hardly any bars or pubs, it is the case of neighborhoods such as Retiro, Salamanca or Chamartín. Probably because they are residential neighborhoods.

To find bars and pubs you have to move to central neighborhoods or to the periphery.

To finish, the decision to choose will depend on the type of hotel that we are thinking of opening, if for example we want it to be a family hotel in which tranquility is sought for guests and eat well and also have museums and parks nearby, we could choose the neighborhood Salamanca, which is also among the neighborhoods classified in the group with the fewest hotels.

If we prefer to open a more central hotel with a greater variety of pubs and bars, we can choose neighborhoods such as Latina, which is also classified within the neighborhoods with the least amount of hotels.

Another alternative if the hotel that we are going to open is for example a large hotel chain that we are sure can steal guests from nearby hotels and therefore we do not care that it is a grouped neighborhood among those that already have a normal number of hotels. It could be Retiro for example, since it has been the neighborhood

with the largest number of parks, museums, monuments and squares, it is close to the center and has a large number of restaurants.

Conclusion and future directions

Based on the data collected, we have been able to separate the neighborhoods of Madrid into 3 groups, thus classifying the hotel occupancy of each group.
Furthermore, using the foursquare API, we have been able to analyze and obtain results about where to open our hotel.
For a possible future work it would be advisable to increase the number of venues to obtain a better result, obtain information about the score of each hotel in order to be able to locate ours in an area in which our hotel can stand out and obtain information on the price that would have to create the hotel in every neighborhood