IBM Data Science Professional Certificate

IBM Capstone Project: Applied Data Science Capstone

Final report

1. Introduction/Business Problem section

This project aims to utilize data science concepts and machine learning tools to solve a problem for the entrepreneurs and lovers of tapas in Madrid, Spain: where is the best neighborhood to open a tapas restaurant? In this project, we will go through the process of defining the problem, preparing data, and using machine learning to improve business decisions.

Madrid is the capital and most populated city in Spain. The Madrid urban agglomeration has the third-largest GDP in the European Union and its influence in politics, education, entertainment, environment, media, fashion, science, culture, and the arts contribute to its status as one of the world's major global cities.

Madrid is reputed to have a "vibrant nightlife" and some of the main attractions in the city are tapas bars, cocktail bars, clubs, jazz lounges, live music venues, and flamenco theatres. The above makes Madrid a great place for entrepreneurs to start and grow their businesses.

1.2. Target Audience

Entrepreneurs who want to open a tapas restaurant. Our mission is to help them make the best decision by recommending the right borough and explaining why this place is better than others.

2. Description of the Data:

To attack the problem, we need different types of data. Which are:

- List of neighborhoods of Madrid
- Car Parking list and their geolocations of Madrid.
- Light Rail/Metro/Train/Tram/Bus Station list and their geolocations of Madrid

For the first point, we use this **database**. For the second point, we use **Google Geoencoding API** to find the approximate coordinates of each neighborhood. And for the third point, we get the datasets from the **Foursquare API**.

The main disadvantage of Foursquare API is the response limit of queries. In every query, the response of venues is limited to 100 results. In a posterior section, we explain how we tried to solve this limitation.

3. Methodology section

3.1. Business Understanding:

Our main goal is to find optimum borough for a new restaurant.

3.2. Analytic approach:

Choosing a neighborhood and location for an opening restaurant is an important and difficult step for entrepreneurs. Most of the time, it can be as crucial a factor as menu quality and service for the restaurant's success. The neighborhood is not only an important factor for customers but also this can affect the restaurant facade, including the interior designs, style of furniture, etc.

To choose the best location for the restaurant, we divide the project into two parts:

1. we take into account the suggestions of the article 4 Important Factors When Choosing a Location to Open a Restaurant which are
    - Parking. A restaurant should have its own parking lot. If that isn't an option (for example, in a major cities), consider partnering with a hotel in the area that has its own parking options. Also, it has an acceptable distance to public parking lots.
    - Accessibility. There's a reason that major restaurant chains are often located near highway exits: It makes them accessible for customers. Also, it is a good option to close distance to public transportations. Light Rail Station, Metro

Station, Train/Tram Station, Bus Stop around the restaurant is provide convenience for customers.

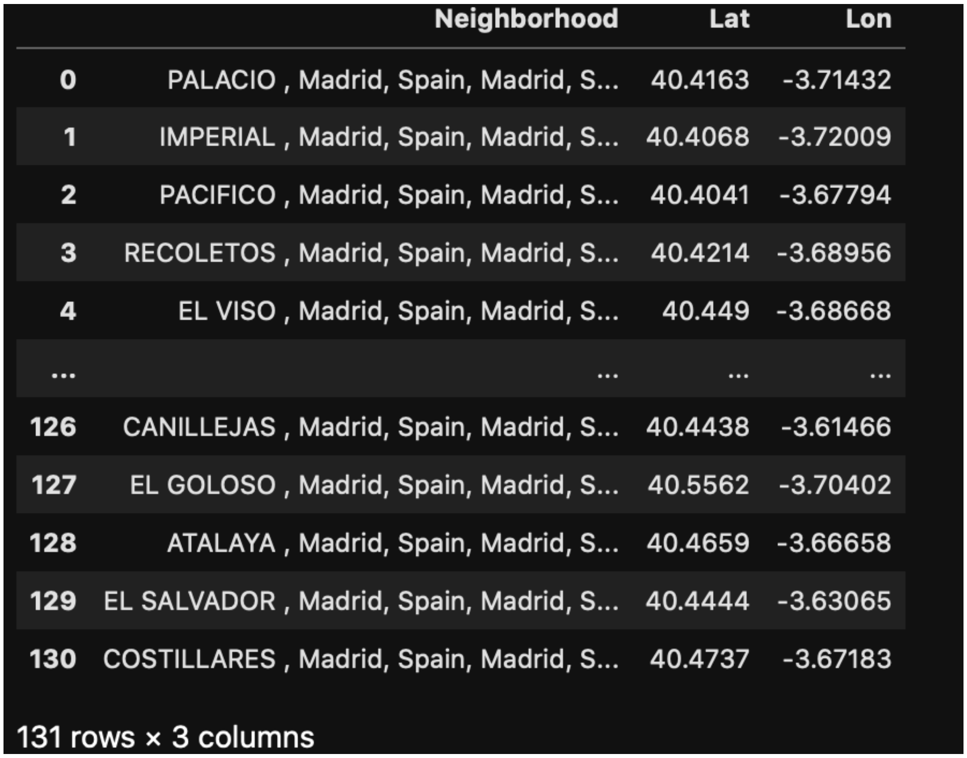
- Visibility. This goes along with accessibility and is very important for new restaurant locations. People have to know the restaurant is there, either in person or on their mobile devices.
- Population Base. There need to be enough people who live in or pass through the area regularly to keep the restaurant busy. So "the particular area's population base" is another important factor about suitable place for a new restaurant. Due to Madrid has a very crowded city with huge number of tourists, the Population Base is negligible.

The *Visibility* is out of the scope of this project. It is hard to keep in view, because it needs a different type of data like usage density of every street, frequently preferred routes and the preferred reason, etc. Maybe field research is also needed. Due to these difficulties, we focused on Parking and Accessibility and collected data about them.

2. From the first part, we take the best thirty neighborhoods with better parking and accessibility factor. Then, we will use Foursquare API to find the top 100 venues within a radius of 1000 meters of the center of those neighborhoods. Finally, we will run k-means clustering on the data to cluster neighborhoods.

3.3. Exploratory Data Analysis

This is the database of the boroughs that we used in this study. After cleaning the database, we added the latitude and longitude coordinates through the **Google Geoencoding API**. Then, we obtain the master data that we will use, which contains the components *Neighborhood, Latitude, and Longitude* of the city information.



	Neighborhood	Lat	Lon
0	PALACIO , Madrid, Spain, Madrid, S...	40.4163	-3.71432
1	IMPERIAL , Madrid, Spain, Madrid, S...	40.4068	-3.72009
2	PACIFICO , Madrid, Spain, Madrid, S...	40.4041	-3.67794
3	RECOLETOS , Madrid, Spain, Madrid, S...	40.4214	-3.68956
4	EL VISO , Madrid, Spain, Madrid, S...	40.449	-3.68668
...	...	...	...
126	CANILLEJAS , Madrid, Spain, Madrid, S...	40.4438	-3.61466
127	EL GOLOSO , Madrid, Spain, Madrid, S...	40.5562	-3.70402
128	ATALAYA , Madrid, Spain, Madrid, S...	40.4659	-3.66658
129	EL SALVADOR , Madrid, Spain, Madrid, S...	40.4444	-3.63065
130	COSTILLARES , Madrid, Spain, Madrid, S...	40.4737	-3.67183

131 rows × 3 columns

Fig. 1. A part of clear borough list from The London Datastore.

We used python **folium** library to visualize geographic details of Madrid neighborhoods

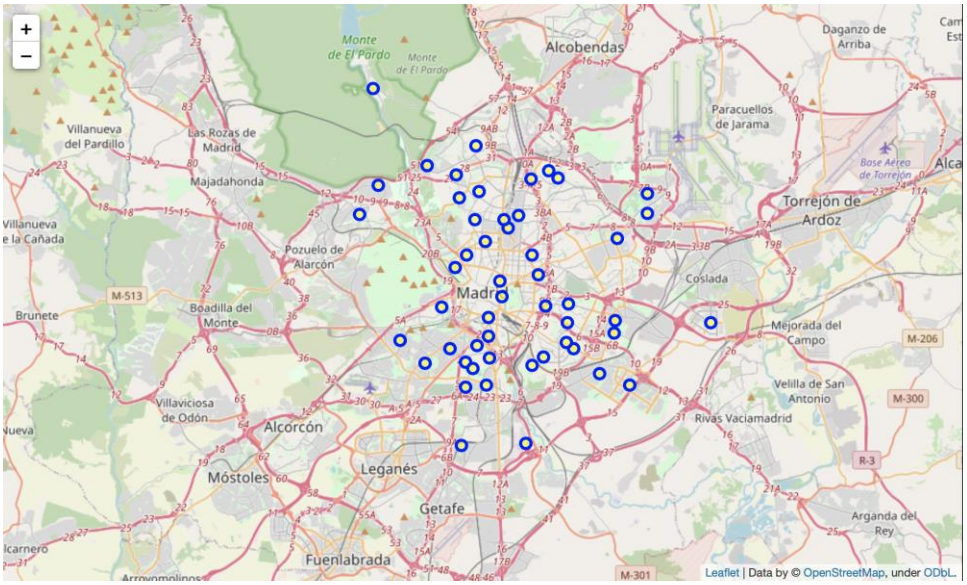
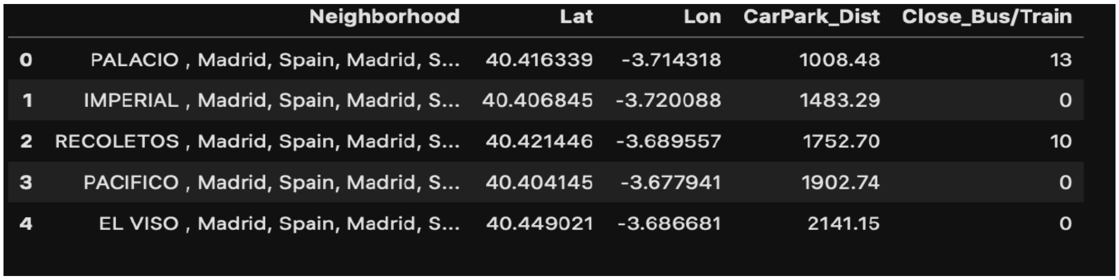


Fig. 2. Madrid Neighborhoods.

We utilized the **Foursquare API** to get the distances (meters) of the nearest parking lot to each neighborhood as well as the number of Light Rail/Metro/Train/Tram/Bus Station in a radius of 1 km. Next, we show the top 5 data of the extended data frame



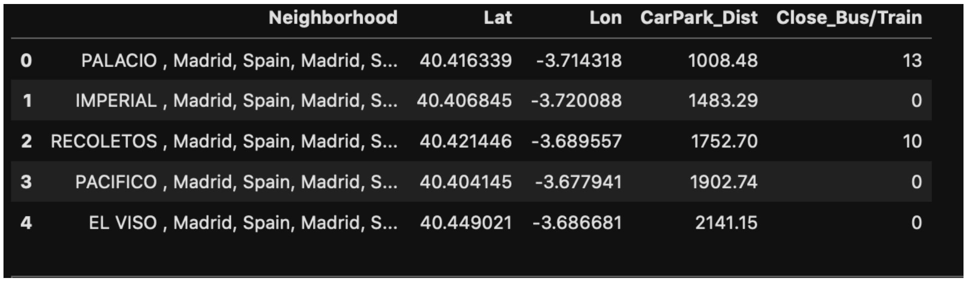
	Neighborhood	Lat	Lon	CarPark_Dist	Close_Bus/Train
0	PALACIO , Madrid, Spain, Madrid, S...	40.416339	-3.714318	1008.48	13
1	IMPERIAL , Madrid, Spain, Madrid, S...	40.406845	-3.720088	1483.29	0
2	RECOLETOS , Madrid, Spain, Madrid, S...	40.421446	-3.689557	1752.70	10
3	PACIFICO , Madrid, Spain, Madrid, S...	40.404145	-3.677941	1902.74	0
4	EL VISO , Madrid, Spain, Madrid, S...	40.449021	-3.686681	2141.15	0

3.4 Create DataSets For Comparison

We need 2 datasets for comparison between the neighborhoods

- Car Parking list and their geolocations of Madrid
- Light Rail/Metro/Train/Tram/Bus Station list and their geolocations of Madrid

We can get these data via Foursquare Api. Firstly, we establish the Foursquare API.



	Neighborhood	Lat	Lon	CarPark_Dist	Close_Bus/Train
0	PALACIO , Madrid, Spain, Madrid, S...	40.416339	-3.714318	1008.48	13
1	IMPERIAL , Madrid, Spain, Madrid, S...	40.406845	-3.720088	1483.29	0
2	RECOLETOS , Madrid, Spain, Madrid, S...	40.421446	-3.689557	1752.70	10
3	PACIFICO , Madrid, Spain, Madrid, S...	40.404145	-3.677941	1902.74	0
4	EL VISO , Madrid, Spain, Madrid, S...	40.449021	-3.686681	2141.15	0

Fig. 3. Data set with the parking lot distance and with the number of public transport stops.

3.5 Comparison of Neighborhoods

I In this step, we can start with the comparison of neighborhoods to determine the best suitable location for a new restaurant in Madrid. Next, we show the thirty best neighborhoods with better distances of parking lots and a greater number of public transport stations.

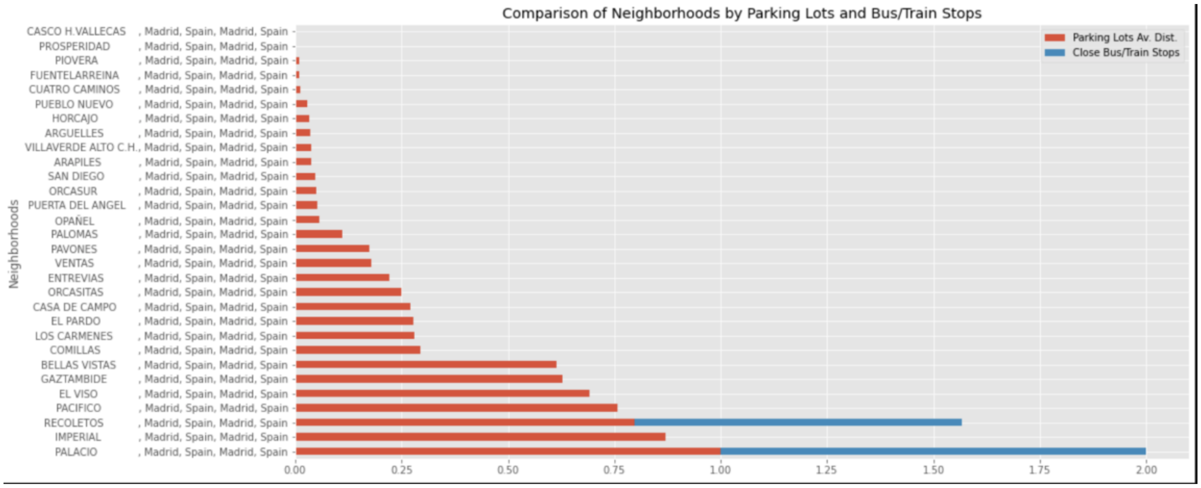


Fig. 4. Comparison among the first 30 neighborhoods.

We find that the top three neighborhoods are: 1 PALACIO, 2 RECOLETOS y 3 IMPERIAL. Next, we show a map with the best neighborhood, i.e., PALACIO with parking lots and public transport stations.

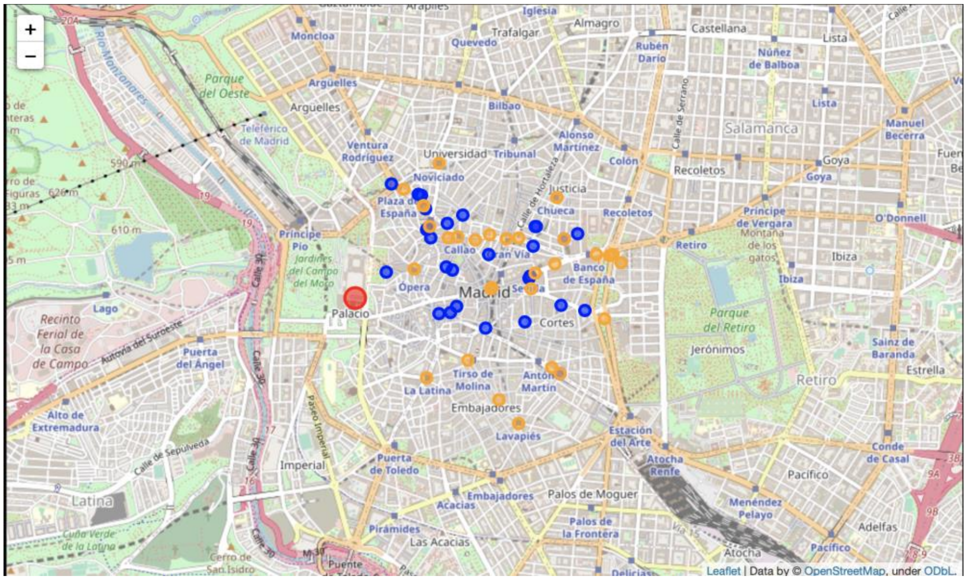


Fig. 5. Palacio, best neighborhood.

The **RED** circle marker is PALACIO, Parking Lots are **BLUE** circle markers, and Metro/Train/Tram/Bus Stations are **ORANGE** circle markers.

According to this preliminary analysis, the best neighborhood to open a restaurant is **Palacio**. Now, we will explore the top 100 venues that are in the neighborhoods within a radius of 1500 meters from the first thirty neighborhoods with the best parking lots and bus/train stations.

3.6 Explore Restaurants in Madrid

Before we go into clustering, we added new information to our data frame. This new information is about all tapas restaurants that were returned by the Foursquare API.

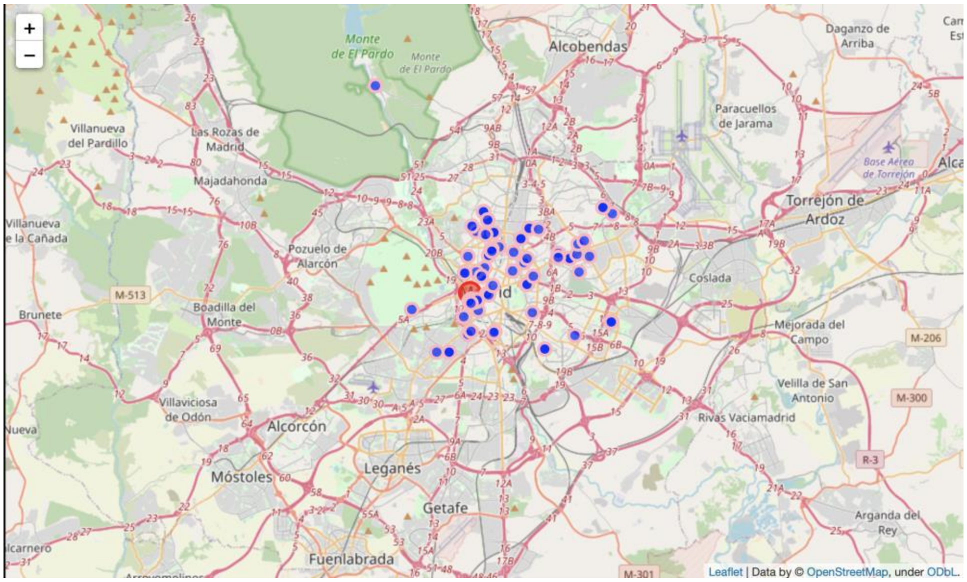


Fig. 6. T*apas restaurant for each borough in blue point and red the Palace neighborhood is shown.*

4.Results

4.1 Machine Learning - Cluster Neighborhoods

We are ready to go into machine learning! For this project, we using **k-means clustering**. To begin with, **a cluster is a collection of data points aggregated together based on their similarities**. Using machine learning algorithms, we can cluster the neighborhoods based on their similarities to each other. The K-means algorithm, in particular, first identifies k number of centroids and then assigns each data point to the cluster, such that the data point is closer to the centroid of that cluster than any other centroid. K-means algorithm repetitively runs this before the centroids are stabilized and the clusters are formed. We are using this method because it is an unsupervised learning method, which means that the algorithm will find the similarities between the data points for us since we do not know them in principle.

4.1.1 Find best K

One limitation of the k-means clustering is that the algorithm does not decide how many clusters to form on its own and we need to find the best K to make clustering more accurate. The Elbow Method is one of the most popular methods to determine this optimal value of k. We iterate the values of k from 1 to 10 and calculate the distortion and inertia values for each value of k in the given range. Distortion is the average of the squared distances from the cluster centers of the respective clusters while inertia is the sum of squared distances of samples to their closest cluster center.

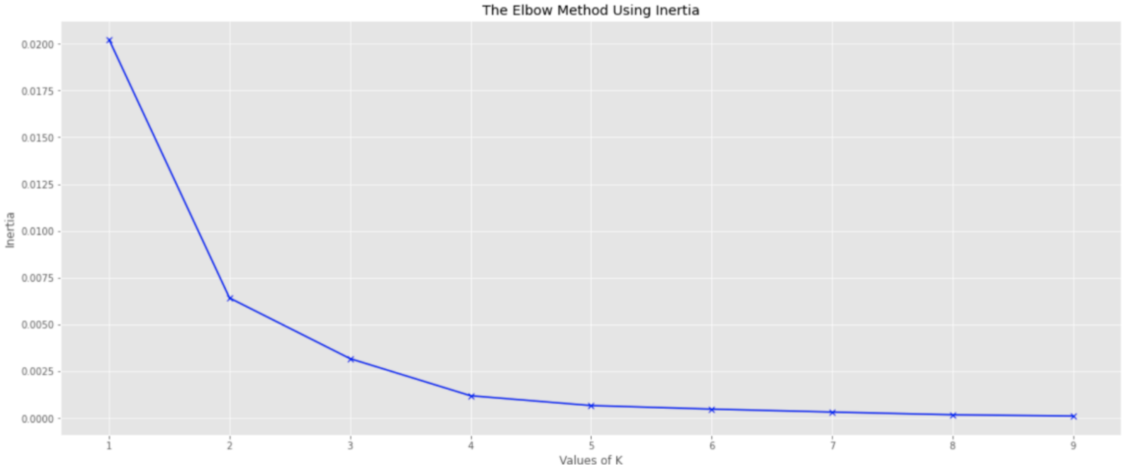


Fig. 7 The elbow method using inertia.

To determine the optimal number of K, we select the value of k at the “elbow” of the plots, the point after which the distortion/inertia starts decreasing linearly. Given these plots, we conclude that 4–6 clusters would work best for our data. Ultimately, we decided to go with 5.

Here, we show a map with the 5 cluster

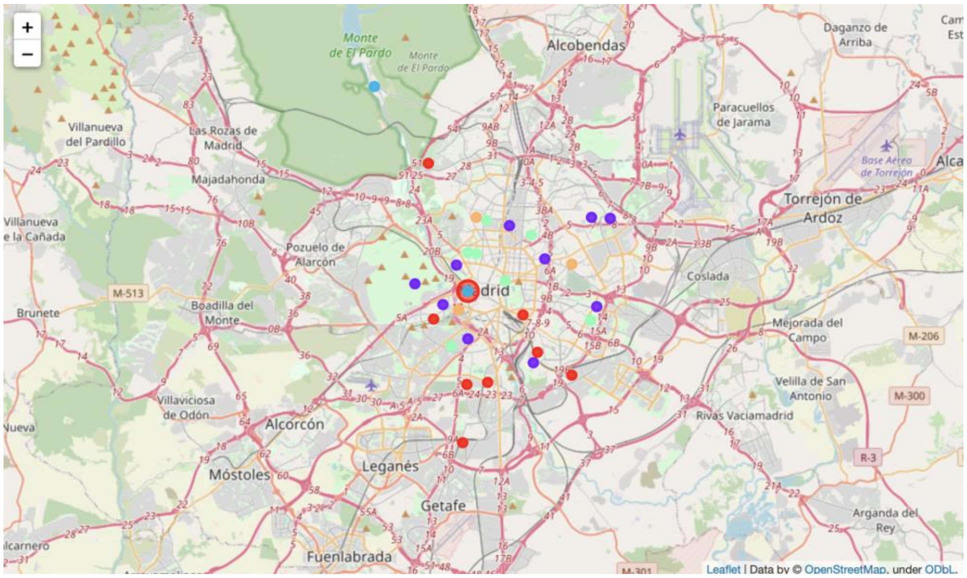


Fig. 8. Five clusters

The cluster are:

- Cluster 0 = Orange. It has an average frequency of Tapas Restaurants of 0.06666666666666667.
- Cluster 1 = Red. It has an average frequency of Tapas Restaurants of 0.0027380952380952383.
- Cluster 2 = Purple. It has average frequency of Tapas Restaurants of 0.027064030267977636.
- Cluster 3 = Blue. It has an average frequency of Tapas Restaurants of 0.09166666666666667.
- Cluster 4 = Green. It has an average frequency of Tapas Restaurants of 0.04817986611090059.
- Palacio Neighborhood = big red point

From the above map, we can see that the Palacio Neighborhood is in cluster 3 which has the higher average frequency of Tapas Restaurants. These two are a good indicator that the best place to open a tapas restaurant in the Palacio’s Neighborhood.

5. Discussion and Conclusions

In this project, we want to determine the best neighborhood in Madrid for the entrepreneurs to open a tapas restaurant. For this aim, we get the neighborhood list of Madrid, and then we use the Google Geoencoding API to find the approximate coordinates of the neighborhoods.

For comparison between the neighborhoods, we need the “Parking Lots” list and “Metro/Train/Tram/Bus Station list” with their geolocations. We get these data via Foursquare API. On the Foursquare website, we learned that the categoryId of our venues’ (Parking, Metro Station, Tram Station, and Bus Station) categoryIds and use these Ids to get and create the datasets. As we know, Foursquare API has a response limit (the number of venues returned by Foursquare API has a maximum limit is 100) of queries. Because there are more than 100 Bus/Train Stops in Madrid, we tried to send queries with sub-categoryIDs of venues and update the data frame with “for loop” with 1000 meters radius easily.

Then, we calculate the average distance of all “parking lots” for each neighborhood and calculate the total “Metro/Train/Tram/Bus Station number” below 1000 meters for each neighborhood.

After comparing the neighborhoods with these two factors, we reach that The best neighborhood for a new restaurant in Madrid is: **PALACIO**.

Palacio has:

- 13 public transport (Metro/Train/Tram/Bus) Stations in 1 km of distance.
- The average distance for 17 Parking Lots is around 1 km.

During our cluster analysis, we found that Cluster 1 has the lowest average frequency of tapas restaurants while Cluster 3 has the highest. However, Cluster 4 has the highest number of tapas restaurants, with a lower average frequency, which might be because Cluster 4 has a high number of neighborhoods compared to other clusters and that there are other common venues in the neighborhood which makes the frequency of tapas restaurants lower. On the other hand, cluster 3 has only two neighborhoods, Palacio and El Pardo.

Based on the previous result, we can conclude that the better place to open a new tapas restaurant is in the Palacio’s neighborhood. That is because it has the closest public transport station as well as the closest parking lots.

6. Final comments

It is only an exploratory project and it is not claimed that the result of the project is 100% proper for the trade and investing industry. The project framework needs more thinking and planning systematically. And also, the project needs more datasets, field and market research.

We should also note some limitations to this analysis. To start with, the Foursquare API limit to return only the top 100 venues within the radius of 1500 meters. But the neighborhoods are very different in shapes and sizes. Some neighborhoods are much larger but less populated while others are more densely populated with a smaller area. Thus, the Foursquare API might not have been able to capture all the tapas restaurants in each neighborhood. However, we calculated the frequency of these restaurants within the 1500-meter radius, which could still reflect the average frequency of tapas restaurants within that neighborhood. Thus, before opening a restaurant, it might be better to do some research on that specific neighborhood, for example, on its commercial pricing, consumers, competitors, and consider other factors. Here a list of some upgrades to do:

- We need different types of data like usage density of every street, frequently preferred routes, and the preferred reason.
- Field research is also important. For instance, a restaurant has 10 parking lots in the too-close distance but they haven't got enough capacity so most of the time it has no free parking lot. So here 10 close parking lots couldn't add positive effect to make a decision.
- Taxes are another important factor in determining the best location. So, we need extra datasets.
- More data about the neighborhood. For instance, bad smell in the area, construction in the street, closely a loud venue.

So, the project has insufficiencies.