Final Capstone Project: The Battle of the Neighborhoods

by Ting Ting Chen

Introduction: Business Problem

Opening a new business in Madrid

Madrid is the capital city of Spain and one of the main economic hubs for the country. It is also the destination of many tourists and many people seeking professional opportunities both from Spain and from the rest of the world. Therefore, there is no wonder why our client chose this place to open their new business.

Since our client made up their mind about the city where they would like to open their new business, a specialty cookie shop, a business idea imported from their home country. As they are not from Madrid and are not very familiar with the different boroughs within the city, they are seeking our help to make the best decision in regards to what district they should be targeting.

In order to solve their problem, we will analyze the different boroughs in Madrid: we will study the population density of each district, the venue categories in each of them (to make sure they aren't overcrowded with dessert shops). These aspects will helps us determine which district is best for our client.

Data

In order to execute our analysis, we will be using the following data sets/sources:

- Foursquare API will allow us to retrieve the venue categories and information
- Callejero_Vigente CSV from the official Madrid website, which contains information on each borough, district, postal code, type of road, coordinates, etc. This set will allow us to locate and list the different neighborhoods to be considered for the new business as well as their location.
- Population data set from Comunidad de Madrid, which includes population data for each district. This will be helpful for our analysis because they could be potential loyal and usual customers.

Methodology and Analysis

In this section, I proceed to explain the data wrangling and analysis part of this project.

I. District Geographical Information

For this section, I used the data from the Callejero_Vigente CSV available on the Madrid official website. This CSV contains extensive information for every single street, neighborhood, district, type of path, coordinates, and other attributes. Therefore, some wrangling had to be done as it also had to have matching district name formats with the other data set used for this project in order to easily merge them.

First, all the unnecessary columns were dropped.

Next, the coordinate format had to be changed so I had to use splits and operators to convert longitude and latitude to decimal formats looping through each row. As the data set was too large for my current server capabilities, I made an assumption by only considering the data of the first row of each district, as opposed to my original plan to calculate the mean of each of the streets in the district. This way, I was able to make the data set much smaller, allowing me to actually process the data.

Then, spaces had to be eliminated in order to make the district column comparable to the demographic data set.

The resulting data set, which allowed us to place each district on a map as well as to know the official list of districts of Madrid, would look like this:

	Distritos	Longitud	Latitud
0	ARGANZUELA	-3.703536	40.404592
1	BARAJAS	-3.596261	40.450250
2	CARABANCHEL	-3.719703	40.376583
3	CENTRO	-3.703944	40.418897
4	CHAMARTIN	-3.666369	40.451381

II. District Demographic Data

For this portion, I used the data set on the Community of Madrid website in regards to district and population density information. The CSV had information on each district along with population density per sq. KM.

This data set included columns that were not needed for this project so they had to be dropped.

Then, the district column had to be converted to upper case to make it the same as the district column on the previous data set, and spaces and special characters, such as accents, had to be eliminated, or else we wouldn't be able to easily merge both data sets.

The resulting data set looks like this:

	Distritos	Densidadkm2
0	CENTRO	25340.69
1	ARGANZUELA	23306.44
2	RETIRO	21867.53
3	SALAMANCA	26830.78
4	CHAMARTIN	15723.25

III. Merged District Data Set

The final data set to actually be used for our analysis and data visualization is the result of merging the previous two data sets. We prepared the row values to finally be able to merge both data sets into the following:

	Distritos	Longitud	Latitud	Densidadkm2
0	ARGANZUELA	-3.703536	40.404592	23306.44
1	BARAJAS	-3.596261	40.450250	1076.06
2	CARABANCHEL	-3.719703	40.376583	17316.88
3	CENTRO	-3.703944	40.418897	25340.69
4	CHAMARTIN	-3.666369	40.451381	15723.25

IV. Analyzing population

For this portion, I used a histogram to visualize the top 5 districts in terms of population density in order to narrow down our choices for the next parts of this analysis.

V. District locations on a map

Once I had the resulting five districts to further analyze, I used the folium library to plot them on the map of Madrid to visualize their actual location.

VI. Venue Categories per District and Clustering

Using the Foursquare API, I got nearby venues for each of the top 5 districts based on the coordinates provided on our merged data set. We limited the results to 100.

We were able to learn that there are 108 unique venue categories across the top 5 districts.

Then, using Panda's get dummies, grouping and mean functions, I was able to get the frequency of each of those venue categories per each of the top 5 districts in order to fetch the ten most popular venue categories for those districts.

This resulted in the following data set:

	Distritos	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	Bar	Tapas Restaurant	Restaurant	Spanish Restaurant	Café	Art Gallery	Plaza	Indie Theater	Coffee Shop	Bookstore
1	CENTRO	Plaza	Hotel	Clothing Store	Tapas Restaurant	Restaurant	Hostel	Café	Cocktail Bar	Bookstore	Gourmet Shop
2	CHAMBERI	Bar	Spanish Restaurant	Pub	Tapas Restaurant	Café	Restaurant	Sandwich Place	Beer Garden	Italian Restaurant	Coffee Shop
3	SALAMANCA	Restaurant	Spanish Restaurant	Bakery	Ice Cream Shop	Burger Joint	Café	Dessert Shop	Italian Restaurant	Tapas Restaurant	Clothing Store
4	TETUAN	Spanish Restaurant	Hobby Shop	Gym / Fitness Center	Ice Cream Shop	Mediterranean Restaurant	Mexican Restaurant	Park	Pizza Place	Restaurant	Breakfast Spot

Following this, using K means allowed us to group the districts, based on their similarity of frequency of types of venues. I chose a K of 4. This is a small data set as we narrowed it down earlier on, so clustering further would not be necessary.

The resulting data set of clustering would look like this data frame:

	Distritos	Longitud	Latitud	Densidadkm2	Cluster Label	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	10t Mos Commo Venu
0	ARGANZUELA	-3.703536	40.404592	23306.44	1	Bar	Tapas Restaurant	Restaurant	Spanish Restaurant	Café	Art Gallery	Plaza	Indie Theater	Coffee Shop	Bookstor
3	CENTRO	-3.703944	40.418897	25340.69	3	Plaza	Hotel	Clothing Store	Tapas Restaurant	Restaurant	Hostel	Café	Cocktail Bar	Bookstore	Gourm Sho
5	CHAMBERI	-3.711703	40.434694	29049.26	1	Bar	Spanish Restaurant	Pub	Tapas Restaurant	Café	Restaurant	Sandwich Place	Beer Garden	Italian Restaurant	Coffe Sho
14	SALAMANCA	-3.681403	40.421958	26830.78	2	Restaurant	Spanish Restaurant	Bakery	Ice Cream Shop	Burger Joint	Café	Dessert Shop	Italian Restaurant	Tapas Restaurant	Clothin Stor
16	TETUAN	-3.704361	40.464525	28664.25	0	Spanish Restaurant	Hobby Shop	Gym / Fitness Center	Ice Cream Shop	Mediterranean Restaurant	Mexican Restaurant	Park	Pizza Place	Restaurant	Breakfa Spo

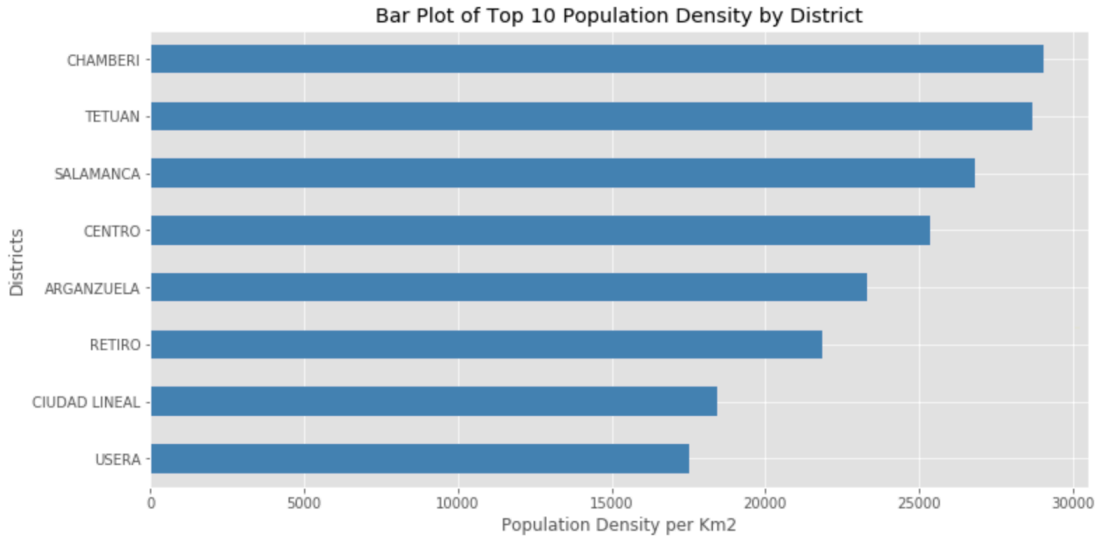
Our clusters were:

1. Residential
2. Dining
3. Dessert Area
4. Tourist Area

Then, using folium, the clusters were displayed on the map, and fetching the data of each of the data frames for each cluster I was able to further analyze results.

Results

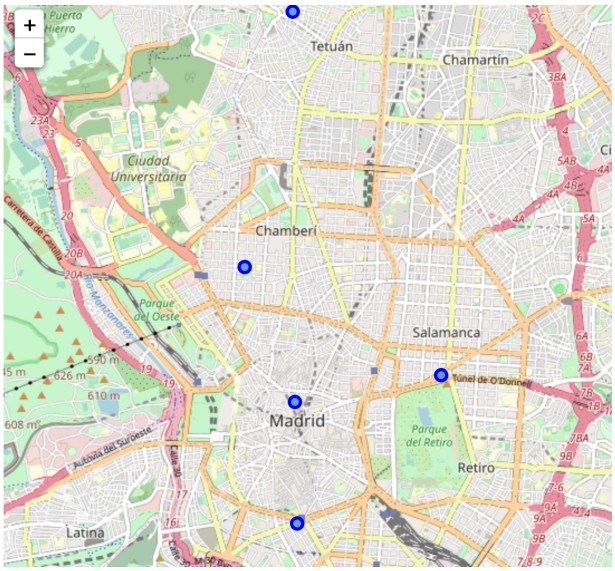
In terms of population density, the top 10 most dense districts are:



From these results, I decided to proceed with the top 5 most population dense districts to narrow down our analysis. Population density is important because it determines the amount of potential recurring customers due to proximity. The chosen districts are:

1. Chamberi
2. Tetuan
3. Salamanca
4. Centro
5. Arganzuela

The location of the five districts can be seen here:



The next step in the analysis is clustering to get the best venue category combination to choose the right district in terms of the nature of the businesses already located there. We had 4 resulting clusters:

1. Cluster 1: Residential

   I decided to call this the residential cluster as fitness centers are very common, there are parks and hobby shops are one of the most common categories. These indicate that this district isn't very touristy.

	Distritos	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
16	TETUAN	Spanish Restaurant	Hobby Shop	Gym / Fitness Center	Ice Cream Shop	Mediterranean Restaurant	Mexican Restaurant	Park	Pizza Place	Restaurant	Breakfast Spot

2. Cluster 2: Dining Area

   This cluster includes a variety of venues but mostly related to bars and restaurants. These districts are probably very busy as there are many dining places.

	Distritos	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	Bar	Tapas Restaurant	Restaurant	Spanish Restaurant	Café	Art Gallery	Plaza	Indie Theater	Coffee Shop	Bookstore
5	CHAMBERI	Bar	Spanish Restaurant	Pub	Tapas Restaurant	Café	Restaurant	Sandwich Place	Beer Garden	Italian Restaurant	Coffee Shop

3. Cluster 3: Dessert Area

   This cluster shows many different dessert/coffee places among their most common venues.

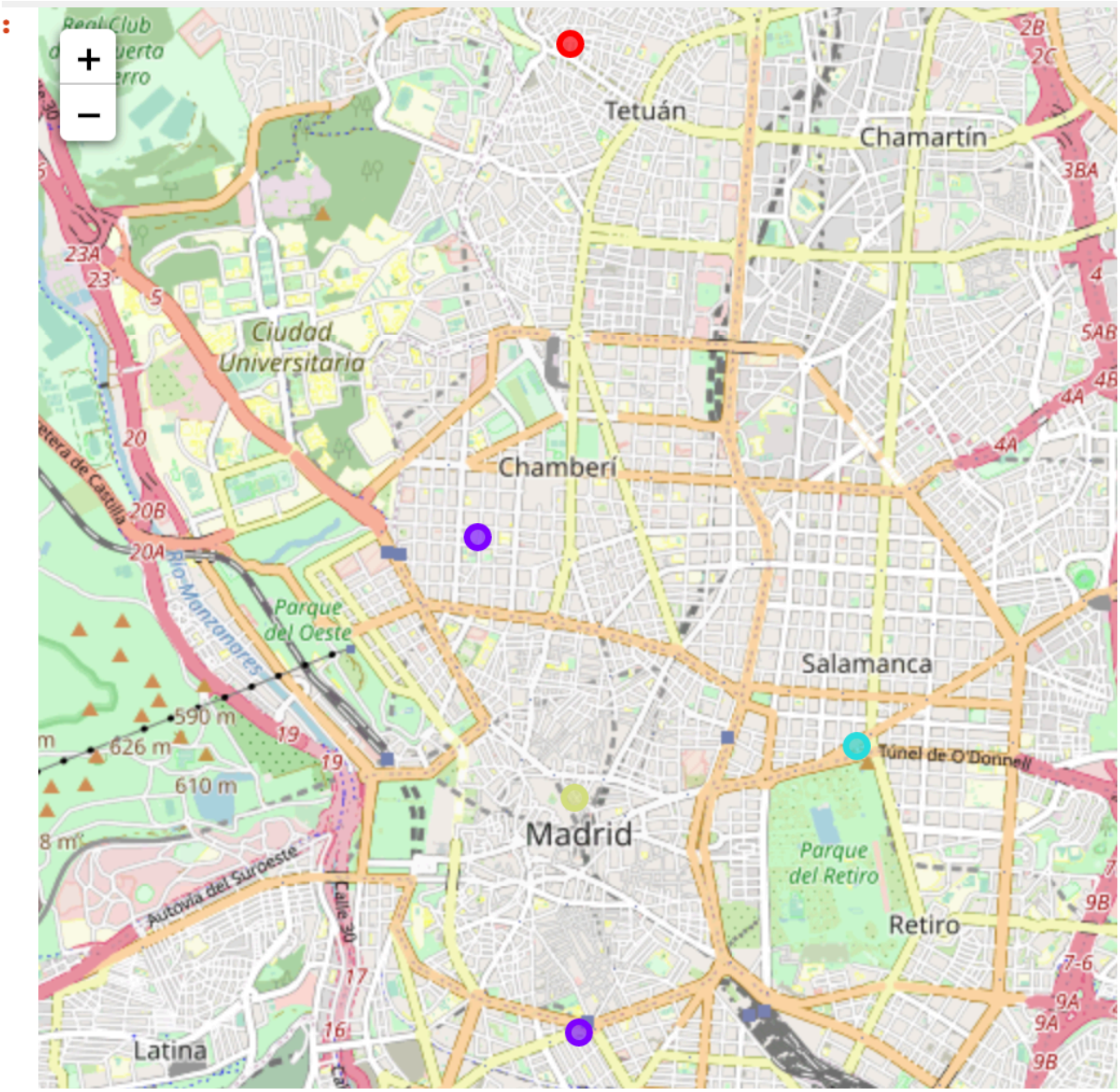
	Distritos	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
14	SALAMANCA	Restaurant	Spanish Restaurant	Bakery	Ice Cream Shop	Burger Joint	Café	Dessert Shop	Italian Restaurant	Tapas Restaurant	Clothing Store

4. Cluster 4: Tourist Area

   This cluster has been marked as tourist area as hotels and hostels are very common along with plazas and a good variety of shops and restaurants.

	Distritos	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
3	CENTRO	Plaza	Hotel	Clothing Store	Tapas Restaurant	Restaurant	Hostel	Café	Cocktail Bar	Bookstore	Gourmet Shop

The location of the clusters is shown here:



Discussion

Since we are looking to open a specialty cookie shop, we are looking for a district that does not have coffee shops or dessert shops as their most common venue (they are very similar in category to our own business idea) so we would not want to overcrowd the area. We are looking for an area with a good variety of venues as that guarantees that it is a busy area.

We are ruling Tetuan in cluster 1 out as two of the most common venues are a hobby shops and fitness centers. They don't seem to be very compatible with our product and target audience and they already have ice cream shops that could satisfy the dessert need in that area.

We are also ruling Salamanca in cluster 3 out as there are many bakeries, dessert shops and coffee shops, which all fall under the same category. There would be too much competition and not too big of a venue category variety.

Out of Arganzuela, Centro and Chamberi, Centro in cluster 4 would be our ideal candidate. Centro has a plaza as a most common venue, which means people might want something to snack on while there, hotels and stores are also present, so their guests and customers might also want something quick to eat and there are restaurants and food places, so our cookies could work for dessert. There are also coffee shops but they are only the 10th most common type of venue, so we will not overcrowd and this shows that this area is also not a no-go for our category. Our cluster 2 with districts Chamberi and Arganzuela, would be our second option as it is the dining cluster and dessert places could be fitting.

Based on our clustering and venue analysis, we would choose Centro as our target district. However, based on population, we would choose Chamberi, as it is the most populated district. Centro is our fourth most populated district.

The second cluster includes Chamberi and Arganzuela. Therefore, considering the results of population analysis and venue analysis, we would end up choosing Chamberi, being the most populated district and still in our second cluster of preference, still a very acceptable cluster. Choosing Centro would sacrifice our population density opportunity as it is the fourth most populated district, whereas Chamberi seems like the best compromise considering both variables.

Conclusion

Conclusively, I would advise our client to start their business in Chamberi. This seems like the best location considering both population density and the existing businesses in the area. Higher population density can result in more potential recurring customers and the categories of the existing businesses are an important aspect to consider as it helps us analyze our future competition.

References

1.Callejero: https://datos.madrid.es/egob/catalogo/200075-1-callejero.csv
2. Population density: http://datos.comunidad.madrid/catalogo/dataset/distritos_municipio_madrid#