FINAL CAPSTONE

WHERE TO LIVE IN MADRID

MARCOS FLEIRE QUINTAS

Content

Introduction / Business Problem

The business problem is to find an appropriate new place to live for my girlfriend and me (We are the stakeholders). Now that we are working, we decided to stop sharing house and start living together.

We are considering three different neighborhoods, and we want to find the most appropriated one using data science. Our goal is to find a neighborhood with a certain preference of venues. We will give a certain value for each different venue, based on our interests, and we will calculate the winner neighborhood. The location chosen is Madrid, and we are considering four different neighborhoods identified by the Postal Codes (28001, 28008, 28050 and 28003)

Figure 1 Map of Madrid by Postal Code

DATA

The data used for this capstone was obtained from https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=b3c41f3cf6a6c410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default It is official data from a public website of the government it has multiple data that it is too detailed and useless for my study, but I can extract the coordinates form the different Postal Codes, which is the information that I actually will use.

Figure 1 Example of the original Data

After cleaning this is the data that I will be using:

	Barrio	CP	Lon	Lat
0	CASTELLANA	28001	-3.681086	40.429061
1	CIUDAD JARDIN	28002	-3.666369	40.451381
2	ALMAGRO	28003	-3.699867	40.438308
3	ALMAGRO	28004	-3.695542	40.428072
4	ACACIAS	28005	-3.703536	40.404592
5	CASTELLANA	28006	-3.680981	40.430419
6	ADELFAS	28007	-3.673142	40.402025
7	ARGUELLES	28008	-3.717186	40.431500
8	GOYA	28009	-3.680008	40.422697
9	ALMAGRO	28010	-3.692111	40.426914
10	ALUCHE	28011	-3.761086	40.401814
11	ACACIAS	28012	-3.702308	40.404411
12	CORTES	28013	-3.699481	40.419275
13	CORTES	28014	-3.693883	40.412444
14	ARAPILES	28015	-3.706394	40.429808
15	CASTILLA	28016	-3.671872	40.467547
16	PUEBLO NUEVO	28017	-3.637119	40.433400
17	PALOMERAS BAJAS	28018	-3.657567	40.378314
18	ABRANTES	28019	-3.720228	40.385092

Figure 2: Consolidated Data

Data Cleaning

All the data from the website mentioned earlier was excessive, the data that I was going to use was the CP (Postal Code in Spanish), a reference neighborhood name (This value is not especially important because the analysis will be held among four different postal code previously decided) and the latitude and longitude.

There were two challenges, the first one was to convert the coordinates in format º ' " to decimal. This was kind of easy simply with an excel formula. The greater challenge was to select the coordinate to represent each Postal Code, for doing this I calculated the average coordinate for each Postal Code and I selected the closer to the theoretical center.

Preview of the data

The first analysis performed to the data is getting the venues for the whole Madrid and see them on a list to decide which ones will be the venues that we will be focusing on.

```
array(['Lounge', 'Art Gallery', 'Plaza', 'Tapas Restaurant', 'Bakery',
       'Restaurant', 'Salon / Barbershop', 'Spanish Restaurant',
       'Japanese Restaurant', 'Liquor Store', 'Boutique',
       'Mexican Restaurant', 'Mediterranean Restaurant',
       'General Entertainment', 'Seafood Restaurant',
       'Sporting Goods Shop', 'Hotel', 'Juice Bar', 'Coffee Shop',
       'Jewelry Store', 'American Restaurant', 'Ramen Restaurant',
       'Market', 'Furniture / Home Store', 'Burger Joint',
       "Women's Store", 'Italian Restaurant', 'Shoe Store',
       'Thrift / Vintage Store', 'Ice Cream Shop', 'Snack Place',
       'Perfume Shop', 'Gym / Fitness Center', 'Clothing Store',
       'Cocktail Bar', 'Electronics Store', 'Spa', 'Wine Shop',
       'Middle Eastern Restaurant', 'Indian Restaurant', 'Supermarket',
       'Asian Restaurant', 'Café', 'Thai Restaurant', 'Rock Club', 'Park',
       'Gun Range', 'Pub', 'Music Venue', 'Breakfast Spot', 'Playground',
       'Paella Restaurant', 'Pizza Place', 'Bar', 'Beer Bar',
       'Steakhouse', 'Bistro', 'Food Stand', 'Multiplex',
       'Greek Restaurant', 'Theater', 'Science Museum', 'BBQ Joint',
       'Food & Drink Shop', 'Motorcycle Shop', 'Sandwich Place', 'Diner',
       'Metro Station', 'Brewery', 'Grocery Store', 'Burrito Place',
       'Dessert Shop', 'Deli / Bodega', 'Concert Hall', 'Creperie',
       'Pie Shop', 'Other Nightlife', 'Fried Chicken Joint',
       'Yoga Studio', 'Vegetarian / Vegan Restaurant', 'Movie Theater',
       'French Restaurant', 'College Classroom', 'Salad Place',
       'Event Space', 'Gymnastics Gym', 'Theme Restaurant', 'Beer Garden',
       'Bookstore', 'Gym', 'Sushi Restaurant', 'Indie Theater', 'Circus',
       'Indie Movie Theater', 'South American Restaurant', 'Gourmet Shop',
       'Flea Market', 'Photography Lab', 'Portuguese Restaurant',
       'Chinese Restaurant', 'Gastropub', 'Museum', 'Korean Restaurant',
       'Farmers Market', 'Gift Shop', 'Soccer Field', 'Football Stadium',
       'Athletics & Sports', 'Skate Park', 'Martial Arts School',
       'Department Store', 'Dumpling Restaurant', 'Fabric Shop',
       'Pastry Shop', 'Boxing Gym', 'Kebab Restaurant', 'Cheese Shop',
```

Figure 2 Example of the venues in Madrid

There were 214 different categories for venues in Madrid. For our study we will choose different categories of venues that will be our interests.

interests=['Gym','Grocery Store','Pharmacy','Yoga Studio','Athletics & Sports','Bar','Café','Kebab Restaurant','Movie Theater','Market','Beer Store','Athletics & Sports','Art Museum','American Restaurant','Bakery','Cafeteria','Cocktail Bar','Convenience Store']

Figure 3 List of interests

Value of the venues

To select the most appropriate neighborhood for my girlfriend and me we will give a certain value each venue to get a "fantasy league" like result and use data analysis to get the result.

VENUE	SCORE
Gym	5
Grocery Store	10
Pharmacy	8
Yoga Studio	4
Athletics & Sports	5
Bar	1
Café	2
Kebab Restaurant	3
Movie Theater	1
Market	10
Beer Store	1

Athletics & Sports	5
Art Museum	3
American Restaurant	5
Bakery	5
Cafeteria	2
Cocktail Bar	1
Convenience Store	10

Figure 4 Scores

Supermarkets, convenience stores and markets have the highest values because are venues used daily and basic on the day a day. The pharmacy has the second highest value because is always useful to have one close in case of an emergency.

Candidate Neighborhoods

With a preliminary analysis the Postal Code with the biggest amount of venues is the 28008 and the Postal Code with the least is the 28050.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	28003	40.438308	-3.699867	SimplyCity	40.435390	-3.704314	Grocery Store
1	28050	40.513994	-3.676703	Mercadona	40.512263	-3.677663	Grocery Store
2	28003	40.438308	-3.699867	Fide	40.437304	-3.699554	Bar
3	28003	40.438308	-3.699867	Artium	40.435162	-3.702124	Bar
4	28003	40.438308	-3.699867	Ferreras	40.435927	-3.704347	Bar
5	28008	40.431500	-3.717186	Bravas	40.433104	-3.716156	Bar
6	28008	40.431500	-3.717186	El Chapandaz	40.434099	-3.717356	Bar
7	28008	40.431500	-3.717186	Bar Morales El atómico	40.433170	-3.715881	Bar
8	28008	40.431500	-3.717186	Tuareg	40.433460	-3.714796	Bar
9	28003	40.438308	-3.699867	Cafetería Bretón	40.439621	-3.698994	Café
10	28003	40.438308	-3.699867	Santa kafeina	40.435885	-3.700121	Café
11	28008	40.431500	-3.717186	London	40.431835	-3.712844	Café
12	28008	40.431500	-3.717186	Los Arcos de Moncloa	40.434569	-3.718444	Café
13	28008	40.431500	-3.717186	Alpunto	40.434694	-3.719493	Café
14	28050	40.513994	-3.676703	Restaurante La Estación	40.511907	-3.673857	Café
15	28008	40.431500	-3.717186	Kebab House	40.433245	-3.717466	Kebab Restaurant
16	28001	40.429061	-3.681086	Mercado de la Paz	40.427132	-3.685696	Market
17	28003	40.438308	-3.699867	Mercado de Chamberí	40.436249	-3.697763	Market
18	28001	40.429061	-3.681086	Dingo	40.427247	-3.684072	American Restaurant
19	28001	40.429061	-3.681086	Mallorca	40.429077	-3.683929	Bakery
20	28001	40.429061	-3.681086	Le Pain Quotidien	40.432572	-3.683321	Bakery
21	28001	40.429061	-3.681086	Natas Belem	40.425075	-3.679807	Bakery
22	28003	40.438308	-3.699867	Orio Pan	40.435849	-3.704320	Bakery
23	28008	40.431500	-3.717186	Wicked Sweet	40.433951	-3.713285	Bakery
24	28001	40.429061	-3.681086	Milford	40.432896	-3.683972	Cocktail Bar
25	28001	40.429061	-3.681086	La Flaca	40.433114	-3.683466	Cocktail Bar

Figure 5 candidate neighborhoods.

Discussion

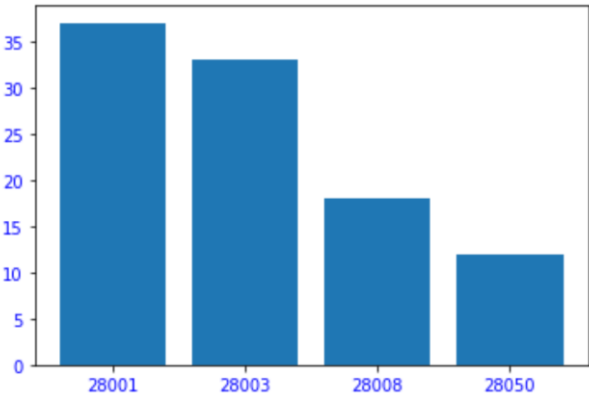After calculating the previously described fantasy value, the results are the next:

Figure 6 Final result

Even though the neighborhood with the postal code 28001 was not the one with the most venues, it had venues with high scores as it was a gym, a supermarket, and some bakeries. The

Postal Code 28008 had the highest number of venues however most of them were bars which had a low score.

Conclusions

The results are not as helpful as wanted because the free version of the Foursquare api has a limit of venues per request. However, I consider that the analysis and the methodology are well presented and make sense. With a premium account or another data provider the results would be much optimal, but the process would still be the same, which I consider is the important.