Applied Data Science Capstone

Tong Yu Bao

November 2020

1 Introduction - Business Problem

Due to the current pandemic (COVID-19), unemployment rates are growing exponentially and multiple small businesses have seen the end of its days. This is even more accentuated in Spain, where tourism and the small service industry represent a big portion of its national GDP. In order to survive these uncertain and difficult times, it is necessary to find a way to stay afloat by creating a safe business.

This idea is extremely difficult during this period. It is extremely risky and we need to find some idea that can generate income even with safety measures that the Government is taking such as lockdowns. After further consideration, we have decided that restaurants are a perfect fit given the situation we are in.

Even with restrictions, people will still need to eat outside, both takeaway and eat in. This means that restaurants are a pretty safe bet in this uncertain time.

Now comes the question of which type of restaurant and which location is the best combination to get the optimum result. The objective of this report is to explore possible types of restaurants and the optimal place to open one.

2 Data

In order to perform the analysis, data will be needed.

First of all, data related to the venues will be extracted from the Foursquare API. Data about income levels in different neighborhoods in Madrid will be taken from the official statistics website of the Madrid city council. Lastly, information about the neighborhood and its coordinates will be taken from Wikipedia

```
coordinates_df = pd.DataFrame( data = {


’Neighborhood’ : [’Centro’, ’Arganzuela’, ’Retiro’, ’Salamanca’,
    ’Chamartin’,
               ’ Tetun ’, ’ Chamber ’, ’Fuencarral-El Pardo’,
                  ’Moncloa-Aravaca’, ’Latina’, ’Carabanchel’,
               ’Usera’, ’Puente de Vallecas’, ’Moratalaz’, ’Ciudad
                  Lineal’, ’Hortaleza’, ’Villaverde’,
               ’Villa de Vallecas’, ’Viclvaro’, ’San
                  Blas-Canillejas’, ’Barajas’],
’Latitude’ : [’40.415347’, ’40.402733’, ’40.408072’, ’40.43’,
    ’40.453333’, ’40.460556’, ’40.432792’, ’40.478611’, ’40.435151’,
                ’40.402461’, ’40.383669’, ’40.381336’, ’40.398204’,
                    ’40.409869’, ’40.45’, ’40.469457’, ’40.345925’,
                    ’40.3796’,
                ’40.4042’, ’40.426001’, ’40.470196’],
’Longitude’ : [’-3.707371’, ’-3.695403’, ’-3.676729’, ’-3.677778’,
    ’-3.6775’, ’-3.7’, ’-3.697186’, ’-3.709722’, ’-3.718765’,
                 ’-3.741294’, ’-3.727989’, ’-3.706856’,
                     ’-3.669059’, ’-3.644436’, ’-3.65’, ’-3.640482’,
                     ’-3.709356’, ’-3.62135’,
                 ’-3.60806’, ’-3.612764’, ’-3.58489’]})
```

Madrid Neighborhood Coordinates

3 Methodology

Once all the data is loaded into the workspace, it is a good practice to plot the venues in a map using the Python package Folium

```
from folium import plugins
import folium
#madrid coordinates

madrid_geo=r’geojson/distritos-1.geojson’
latitude_mad = 40.4167754
```

```
longitude_mad = -3.7037902

madrid_map = folium.Map(location = [latitude_mad, longitude_mad],
   zoom_start = 12)

# instantiate a mark cluster object for the incidents in the
   dataframe
incidents = plugins.MarkerCluster().add_to(madrid_map)

# loop through the dataframe and add each data point to the mark
   cluster
for lat, lng, label, in zip(Madrid_venues[’Venue Latitude’],
   Madrid_venues[’Venue Longitude’], Madrid_venues[’Venue
   Category’]):
   folium.Marker(
       geo_data=madrid_geo,
       location=[lat, lng],
       icon=None,
       popup=label,
   ).add_to(incidents)
```

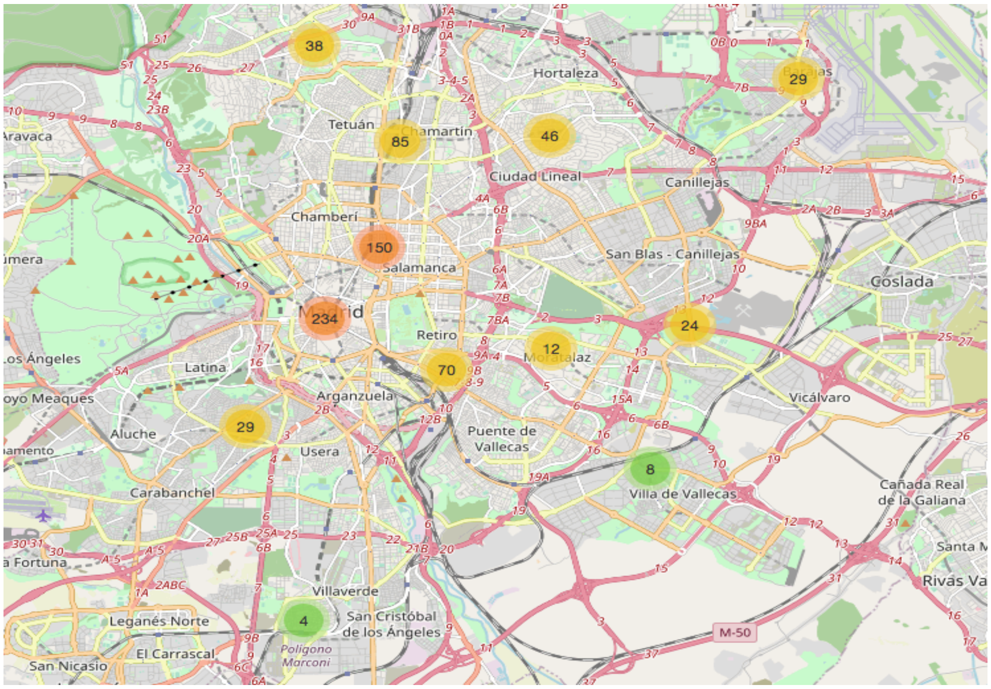


Figure 1: Map with venues

After the venues are loaded and displayed on the map, we are going to run a segmentation algorithm (K-Means) to group venues that are close to each other. Then an analysis on each cluster will be done to obtain the most frequent venues and analysis will be made with that information

4 Results

Once that we have loaded all the data into the workspace, we are going to segment the venues into different clusters attending to the distance between them. The objective of this segmentation is to group the venues that are close together, in walking distance from each other. The number of neighborhoods in Madrid is 21. This is why we have to choose several clusters higher than that. If we choose 21 or lower, we have the issue of segmenting into the existing neighborhoods.

However, when we choose a higher number, it can further divide the venues inside a neighborhood into the walking distance. We are going to choose 30 clusters

With the segmentation done using K-Means, a representation of the clusters is done using Folium.

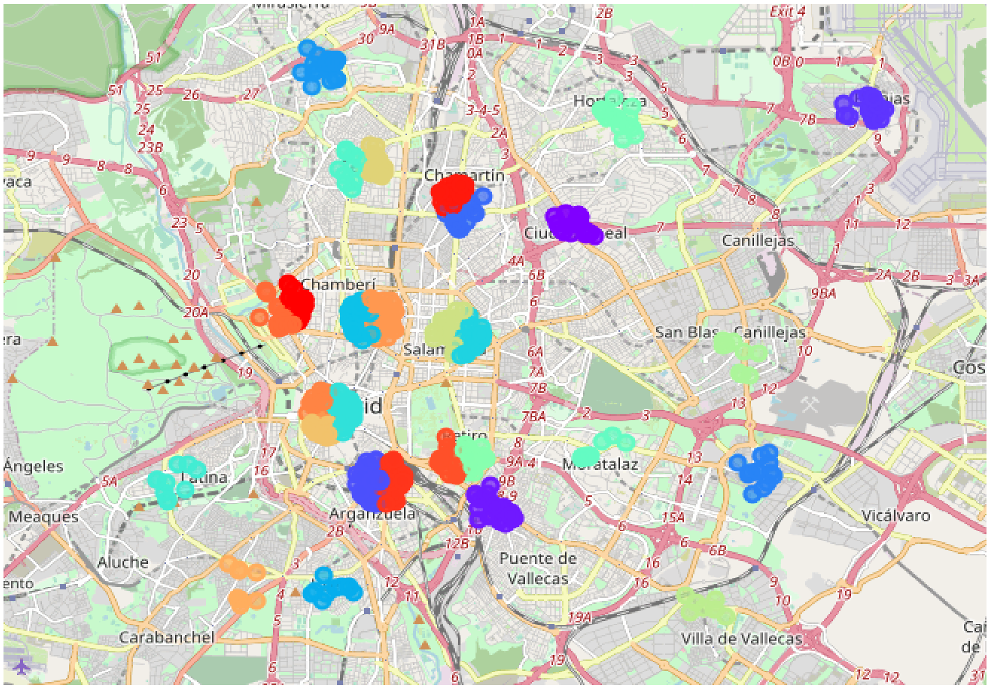
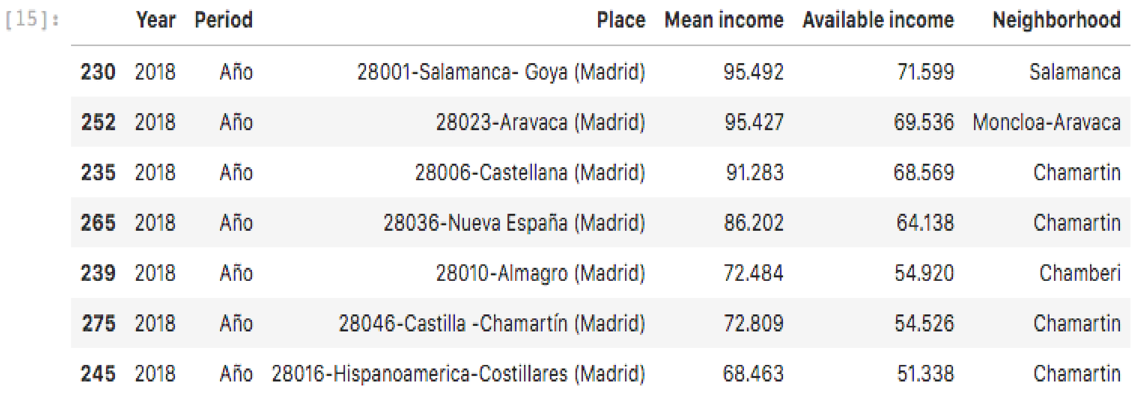


Figure 2: Clusters

Using the income data extracted from the city council website we can select the 5 places with the most income in Madrid and which neighborhood it belongs



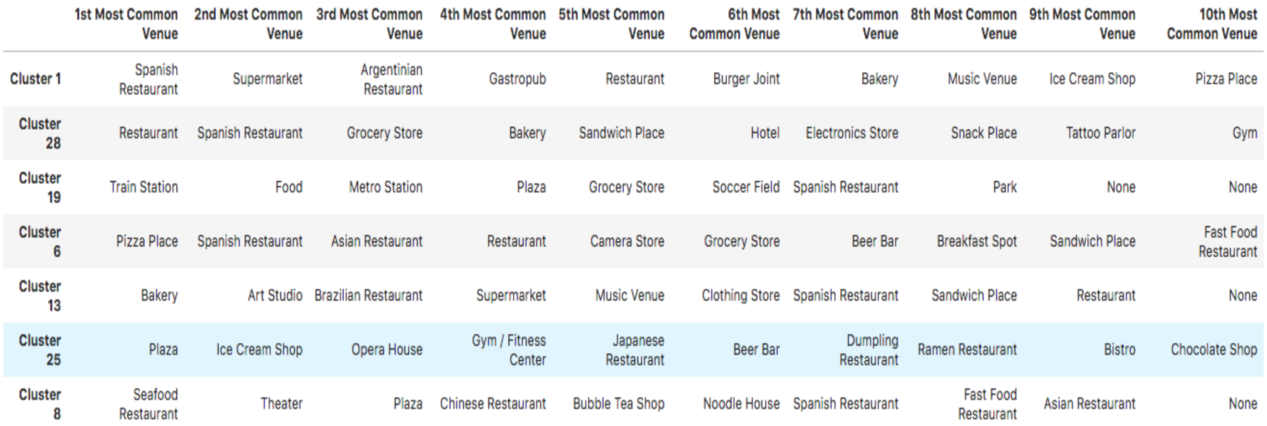
[15]:

	Year	Period	Place	Mean income	Available income	Neighborhood
230	2018	Año	28001-Salamanca- Goya (Madrid)	95.492	71.599	Salamanca
252	2018	Año	28023-Aravaca (Madrid)	95.427	69.536	Moncloa-Aravaca
235	2018	Año	28006-Castellana (Madrid)	91.283	68.569	Chamartin
265	2018	Año	28036-Nueva España (Madrid)	86.202	64.138	Chamartin
239	2018	Año	28010-Almagro (Madrid)	72.484	54.920	Chamberi
275	2018	Año	28046-Castilla -Chamartín (Madrid)	72.809	54.526	Chamartin
245	2018	Año	28016-Hispanoamerica-Costillares (Madrid)	68.463	51.338	Chamartin

Figure 3: Income levels

We can see that the 5 highest incomes are all situated in the northern part of Madrid. The neighborhoods with the most income are Salamanca, Moncloa-Aravaca, Chamartín, and Chamberí. These neighborhoods are perfect for opening a new restaurant. The clusters are 1-28-19 for Chamberí, 6 for Moncloa-Aravaca, 13-25 for Salamanca, and 8 for Chamartin.

Once we have this part, a data frame with the most frequent venues in each cluster is done with the following results.



	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
Cluster 1	Spanish Restaurant	Supermarket	Argentinian Restaurant	Gastropub	Restaurant	Burger Joint	Bakery	Music Venue	Ice Cream Shop	Pizza Place
Cluster 28	Restaurant	Spanish Restaurant	Grocery Store	Bakery	Sandwich Place	Hotel	Electronics Store	Snack Place	Tattoo Parlor	Gym
Cluster 19	Train Station	Food	Metro Station	Plaza	Grocery Store	Soccer Field	Spanish Restaurant	Park	None	None
Cluster 6	Pizza Place	Spanish Restaurant	Asian Restaurant	Restaurant	Camera Store	Grocery Store	Beer Bar	Breakfast Spot	Sandwich Place	Fast Food Restaurant
Cluster 13	Bakery	Art Studio	Brazilian Restaurant	Supermarket	Music Venue	Clothing Store	Spanish Restaurant	Sandwich Place	Restaurant	None
Cluster 25	Plaza	Ice Cream Shop	Opera House	Gym / Fitness Center	Japanese Restaurant	Beer Bar	Dumpling Restaurant	Ramen Restaurant	Bistro	Chocolate Shop
Cluster 8	Seafood Restaurant	Theater	Plaza	Chinese Restaurant	Bubble Tea Shop	Noodle House	Spanish Restaurant	Fast Food Restaurant	Asian Restaurant	None

Figure 4: 10 Most Common Venues

The first 3 clusters, which correspond to the neighborhood of Chamberi, have as most common venues some sort of restaurants (eg. Spanish restaurant, restaurant, Argentinian restaurant...). this means that this place is a very popular place for eating out. The available income of this part of Madrid is around 54920€ in 2018 which suggests that this is indeed a wealthy part so the positioning of the new restaurant can be luxurious.

Cluster 6, corresponds to Moncloa-Aravaca, which is where universities are located. And as expected, pizza places are the most popular due to the nature of the place and university students. This is followed by Spanish restaurants, which can be ”Bar de tapas” which confirms the location of universities.

Chamartin has the highest amount of seafood restaurants and also Asian styled venues such as Bubble Tea Shop, Chinese Restaurant, an Asian restaurant.

Lastly, clusters 13 and 25 are located in Salamanca, one of the highest-earning neighborhoods in Madrid. There is a variety of venues in this section and not dominated by restaurants as we can see in other sectors. This area is perfect to start a new business due to its high income and relative lack of restaurants.

The perfect restaurant for the area of Salamanca would be a seafood restaurant due to the high income of this neighborhood and the lack of options available.

Disclaimer: These data are extracted from Foursquare API which has a limit to the number of venues you can retrieve. Meaning that it does not offer a complete view of venues around the area.