Capstone Project –
Predicting the best place for opening a restaurant

By Samuel Sadek

January 2021

1. Introduction

1.1. Business Problem

In this project we will attempt to answer this simple question: **where would you recommend opening a new restaurant?** As such, specifically this report will be targeted to stakeholders interested in opening a **Spanish restaurant** in **Madrid**, Spain.

Since there are lots of restaurants in Madrid, we will try to detect locations that are not already crowded with restaurants.

A brief summary about the city to set the scene: **Madrid** is the capital of Spain and is home to the Spanish Royal family as well as the Spanish Government. It is a modern metropolitan city and an economical and industrial centre of Spain, and, with its population of nearly 3,5 million people, is also the biggest city in Spain. It is located in the centre of the Iberian Peninsula and is surrounded by mountains and natural parks. Traditionally it the hub between different areas of Spain and is therefore connected to all major Spanish cities by train, road, or air.

In order to address the above question, we will leverage the data science techniques to come up with a few recommended suggestions on the best neighbourhoods and locations to enable the stakeholders to reach a final decision on the best location.

1.2. Target Audience

- A business entrepreneur looking into investing and opening up new restaurant in Madrid.
- Business Analyst or Data Scientists, looking to analyse the neighbourhoods of Madrid using python, jupyter notebook and some machine learning techniques.
- Someone curious about data that want to have an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.

2. Data section

Based on definition of our problem, there are several factors that will influence our decision are:

- Number of existing restaurants in the neighbourhood (and of any type of restaurant)
- Number of and distance to Spanish restaurants in the neighbourhood, if any

Following data sources will be needed to extract/generate the required information:

- Wikipedia webpage containing all of Madrid city´s neighbourhood data
- Number of restaurants and their type and location in every neighbourhood will be obtained using Foursquare API

First of all we needed to compile some information about the area of Madrid city such as districts, neighbourhoods, latitude, longitude etc. so the perfect avenue for gathering this information readily available within the public domain is Wikipedia which was the first place to take a look:

https://en.wikipedia.org/wiki/List_of_neighborhoods_of_Madrid

I began to scrape the content from this URL link and populated it into a Pandas data frame for further pre-processing purposes and use.

An example output of this is shown as follows:

[2]:

	District	Number	Neighborhood
0	Centro	11	Palacio
1	Centro	12	Embajadores
2	Centro	13	Cortes
3	Centro	14	Justicia
4	Centro	15	Universidad
...	...	...	...
125	San Blas-Canillejas	208	El Salvador
126	Barajas	211	Alameda de Osuna
127	Barajas	212	Aeropuerto
128	Barajas	213	Casco Histórico de Barajas
129	Barajas	214	Timón

130 rows × 3 columns

The second pre-processing step was to gather a shapefile and ancillary files for extracting all the Polygon coordinates data for each neighbourhood belonging to Spain´s capital city.

So, to this end, I had downloaded firstly the Madrid GeoJson file which contains this data via the following url:

https://raw.githubusercontent.com/codeforamerica/click_that_hood/master/public/data/madrid.geojson

Then after that, I had converted this GeoJson file into Shapefiles file format via following url:

https://mapshaper.org

Finally, I uploaded the generated Madrid.* files associated with Shapefiles formatting into my Python kernel runtime to be able to run this Python code.

Once I generated these files, I proceeded into the final step of pre-processing to generate a Pandas data frame object with the following headings needed for further processing purposes later on:

[6]:

	Neighborhood	Geometry	Latitude	Longitude
0	Palacio	POLYGON ((-3.70462 40.42147, -3.70503 40.42135...	40.416600	-3.712763
1	Embajadores	POLYGON ((-3.70262 40.41549, -3.70155 40.41510...	40.410419	-3.701153
2	Cortes	POLYGON ((-3.69178 40.42052, -3.69166 40.42006...	40.416025	-3.695522
3	Justicia	POLYGON ((-3.69836 40.42094, -3.69836 40.42094...	40.424839	-3.695351
4	Universidad	POLYGON ((-3.70028 40.42119, -3.70353 40.42144...	40.426850	-3.705760

3. Methodology

3.1. Business Understanding

The aim of this project is to find the best neighbourhood of Madrid to open a new restaurant.

3.2. Analytical Approach

There is a total of 130 neighbourhoods in Madrid, so we need to find a way to cluster them based on their similarities, that are the number and the kind of restaurant. Briefly, after some steps of Data Cleaning and Data Exploration, I will use a K-Means algorithm to extract the clusters, produce a map and make an argument on the final result.

3.3. Data Preparation

We have adopted “Folium” a python library for exploring data that can create interactive leaflet map using coordinate data.

Use geopy library to get the latitude and longitude values of Madrid City.

```
The geograpical coordinate of Madrid City are 40.4167047, -3.7035825.
```

I then proceeded using my Foursquare API credentials to extract all the venues associated with each neighbourhood in Madrid within the 'food' category, and here was the output:

```
(6278, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Palacio	40.4166	-3.712763	la gastroteca de santiago	40.416639	-3.710944	Restaurant
1	Palacio	40.4166	-3.712763	La Esquina del Real	40.417356	-3.710364	French Restaurant
2	Palacio	40.4166	-3.712763	Café de Oriente	40.418081	-3.711867	Spanish Restaurant
3	Palacio	40.4166	-3.712763	Charlie Champagne	40.413936	-3.712647	Restaurant
4	Palacio	40.4166	-3.712763	Gyoza Go!	40.416179	-3.708612	Dumpling Restaurant

Since there was a big concern of receiving duplicate venues from the retrieved dataset from foursquare API, I initially pruned down the list to unique venue names, latitudes, longitudes, and category types which halved the original rows amount, which was great.

However, I was then with another dilemma of using two centroids that were too close to one another, could in itself be also responsible for extracting some more duplicate venues.

To solve this problem, I linked the unique venue with the right neighbourhood by using polygons ("geometry"), retrieved originally from

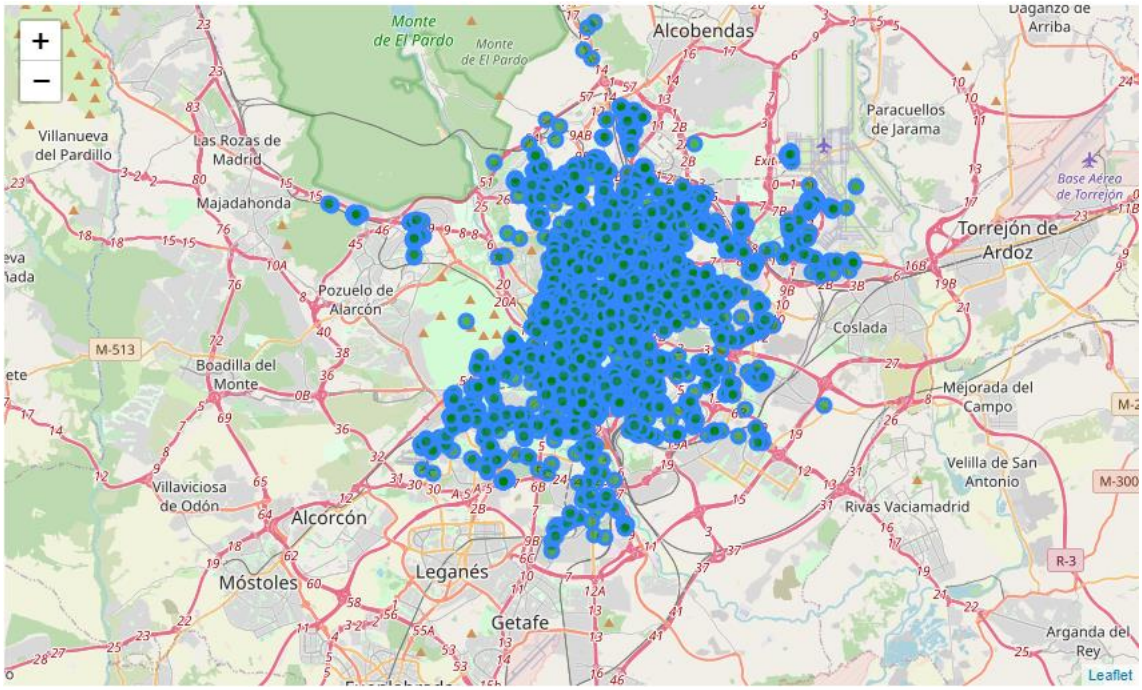
the Madrid shapefile and populated into the above data pre-processing steps.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rios Rosas	40.443574	-3.696422	(H)arina	40.446962	-3.694539	Bakery
1	Butarque	40.338369	-3.674994	100 Montaditos	40.347108	-3.677199	Sandwich Place
2	Buenavista	40.368278	-3.744770	100 Montaditos	40.363294	-3.739031	Sandwich Place
3	Palomeras Bajas	40.385940	-3.657573	100 Montaditos	40.381375	-3.663684	Sandwich Place
4	Aguilas	40.382983	-3.769777	100 Montaditos	40.386003	-3.763597	Sandwich Place

As you can see from the below data frame record counts, I had successfully managed to remove a lot of duplicates, by half:

```
(6296, 7)
(3182, 7)
```

I then populated the retrieved data points for the recovered Madrid venues into a Folium map which can be seen as follows:



After plotted all the available data points onto a map, we continued further pre-processing and incorporated one-hot encoding to convert the categorical predictor namely Venue Category, into a numerical value by calculating the mean of the frequency of occurrence of each category for each neighbourhood showing the following output:

	Neighborhood	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bistro	Brazilian Restaurant	...	Taco Place	Tapas Restaurant	Taverna
0	Abrantes	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.125000	0.0	0.125	...	0.000	0.000000	0.0
1	Acacias	0.000000	0.0	0.0	0.050000	0.000000	0.0	0.025000	0.0	0.000	...	0.025	0.100000	0.0
2	Adelfas	0.000000	0.0	0.0	0.052632	0.000000	0.0	0.052632	0.0	0.000	...	0.000	0.105263	0.0
3	Aeropuerto	0.000000	0.0	0.0	0.000000	0.076923	0.0	0.153846	0.0	0.000	...	0.000	0.076923	0.0
4	Aguilas	0.027027	0.0	0.0	0.000000	0.000000	0.0	0.027027	0.0	0.000	...	0.000	0.135135	0.0

5 rows × 91 columns

Also, since we did not want to distract or divert the attention of potential investor into investing needless establishments such as a cafeteria or a bakery shop which were brought in as standard in the original foursquare API dataset, we filtered on such avenues:

	Neighborhood	Italian Restaurant	Pizza Place	Restaurant	Japanese Restaurant	Seafood Restaurant	Sandwich Place	Sushi Restaurant	Chinese Restaurant	Bistro	...	Ramen Restaurant	Americ Restauran
0	Abrantes	0.000000	0.125000	0.000000	0.0	0.000000	0.125000	0.000	0.000000	0.0	...	0.0	0.00000
1	Acacias	0.000000	0.075000	0.075000	0.0	0.000000	0.000000	0.025	0.050000	0.0	...	0.0	0.00000
2	Adelfas	0.000000	0.000000	0.052632	0.0	0.000000	0.052632	0.000	0.052632	0.0	...	0.0	0.00000
3	Aeropuerto	0.076923	0.000000	0.153846	0.0	0.000000	0.076923	0.000	0.000000	0.0	...	0.0	0.00000
4	Aguilas	0.054054	0.027027	0.135135	0.0	0.027027	0.054054	0.000	0.054054	0.0	...	0.0	0.02702

5 rows × 35 columns

Calculated the top-ten most visited types of establishments or venues for each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each category. This was crucial last data preparatory step to generate the final dataset which would be subjected into a clustering technique explained in the next section of this report:

	Italian Restaurant	Pizza Place	Restaurant	Japanese Restaurant	Seafood Restaurant	Sandwich Place	Sushi Restaurant	Chinese Restaurant	Bistro	Burger Joint	...	Ramen Restaurant	American Restaurant	Re
0	0.000000	0.125000	0.000000	0.0	0.000000	0.125000	0.000	0.000000	0.0	0.125000	...	0.0	0.000000	
1	0.000000	0.075000	0.075000	0.0	0.000000	0.000000	0.025	0.050000	0.0	0.025000	...	0.0	0.000000	
2	0.000000	0.000000	0.052632	0.0	0.000000	0.052632	0.000	0.052632	0.0	0.000000	...	0.0	0.000000	
3	0.076923	0.000000	0.153846	0.0	0.000000	0.076923	0.000	0.000000	0.0	0.000000	...	0.0	0.000000	
4	0.054054	0.027027	0.135135	0.0	0.027027	0.054054	0.000	0.054054	0.0	0.054054	...	0.0	0.027027	

5 rows × 34 columns

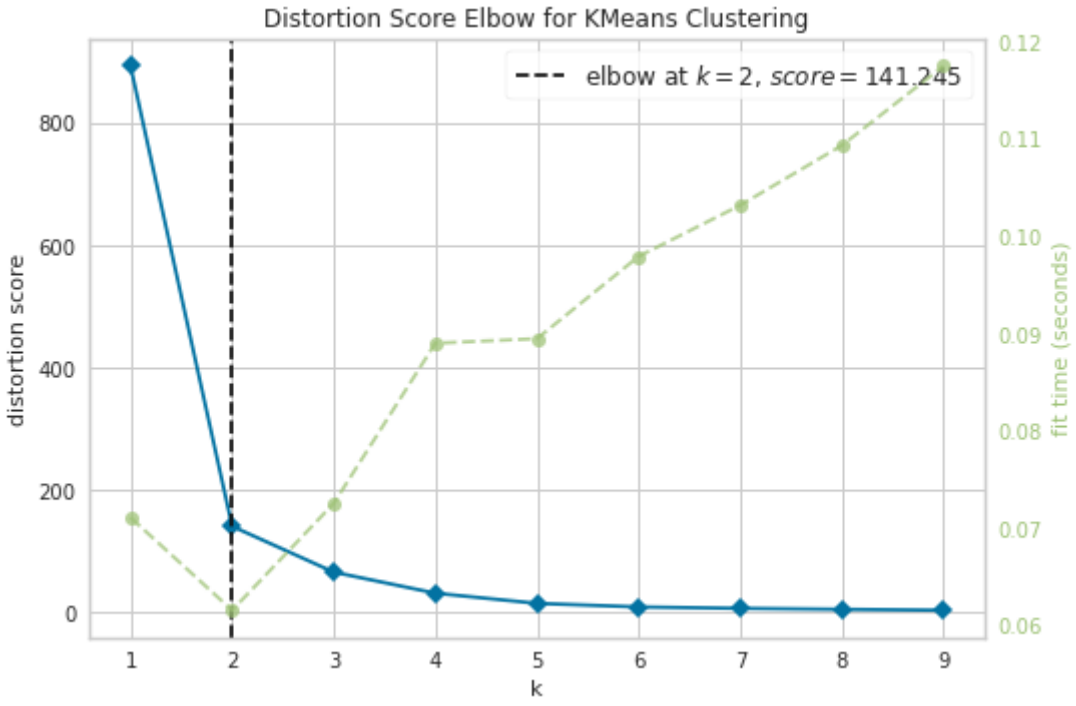
3.4. Clustering

To address the problem of opening up a new restaurant and to perform analysis on the neighbourhood data such as in the Madrid city area in this project, I have adopted K-means clustering algorithm: a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups).

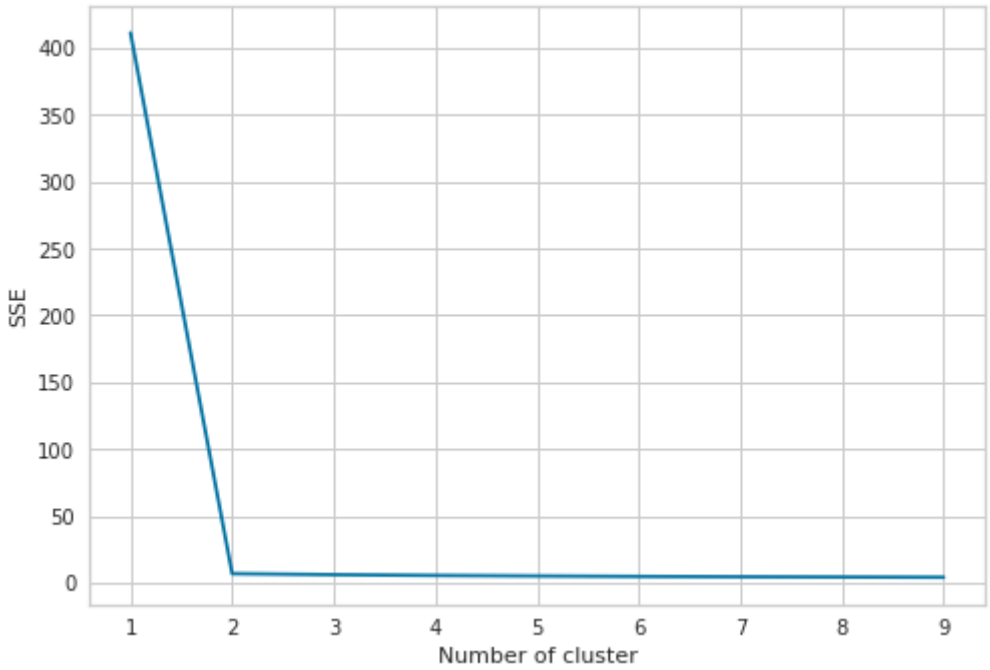
The goal of this algorithm is to **find groups in the data**, with the number of groups or clusters represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

So, the initial step taken was to identify the best optimal value **K** using a famous analytical approach: the **elbow** method.

So, let's proceed.



As you can see from the above plot, we can deduce that the best K is **2**. But before running the kmeans algorithm, let us try and validate that indeed this is the best K value through some other evaluation methods or approaches. For instance, let us test this with normal inertia to calculate elbow as follows:



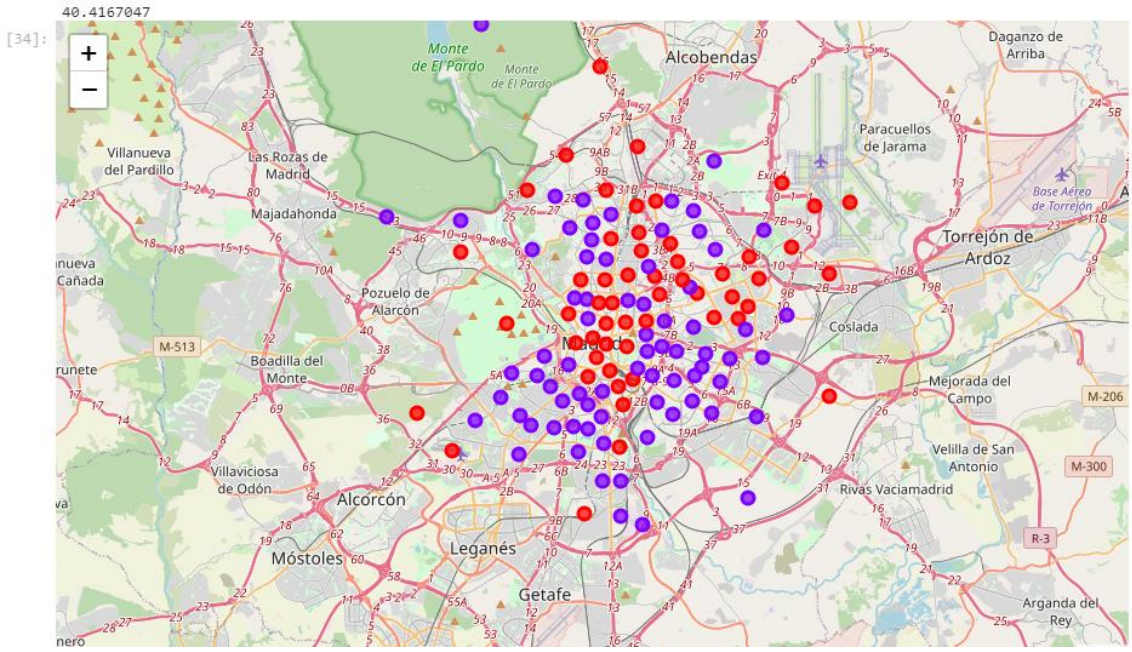
So, from the above, we can conclude it has brought back with a similar result of k = 2. So, it is now time to run the kmeans algorithm using this k value:

	Neighborhood	Latitude	Longitude	Cluster Labels	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 M Comr Ve
0	Palacio	40.416600	-3.712763	0	Spanish Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Gastropub	Argentinian Restaurant	Japanese Restaurant	Bistro	Bu J
1	Embajadores	40.410419	-3.701153	0	Spanish Restaurant	Restaurant	Pizza Place	Italian Restaurant	Breakfast Spot	Vegetarian / Vegan Restaurant	Seafood Restaurant	S Restau
2	Cortes	40.416025	-3.695522	0	Restaurant	Spanish Restaurant	Seafood Restaurant	Mediterranean Restaurant	Breakfast Spot	Argentinian Restaurant	Japanese Restaurant	S Restau
3	Justicia	40.424839	-3.695351	0	Restaurant	Spanish Restaurant	Bistro	Mediterranean Restaurant	Italian Restaurant	American Restaurant	Pizza Place	Sandv P
4	Universidad	40.426850	-3.705760	1	Spanish Restaurant	Restaurant	Gastropub	Pizza Place	Vegetarian / Vegan Restaurant	Italian Restaurant	Mediterranean Restaurant	Argentii Restau

Please note the incorporation of the Cluster Labels column added into the above dataset along with the top-ten most visited or frequent venue. This dataset will be used to draw our final conclusions of this report.

4. Results and Discussion

Prior to analysing all the clusters, let's take a look on a folium map by creating a map of Madrid with neighbourhoods superimposed on top:



As we can see, each cluster belong to a colour with different characteristics. **Red** plots belong to Cluster 1 and **purple** ones to Cluster 2.

Let us now examine the clusters in each turn.

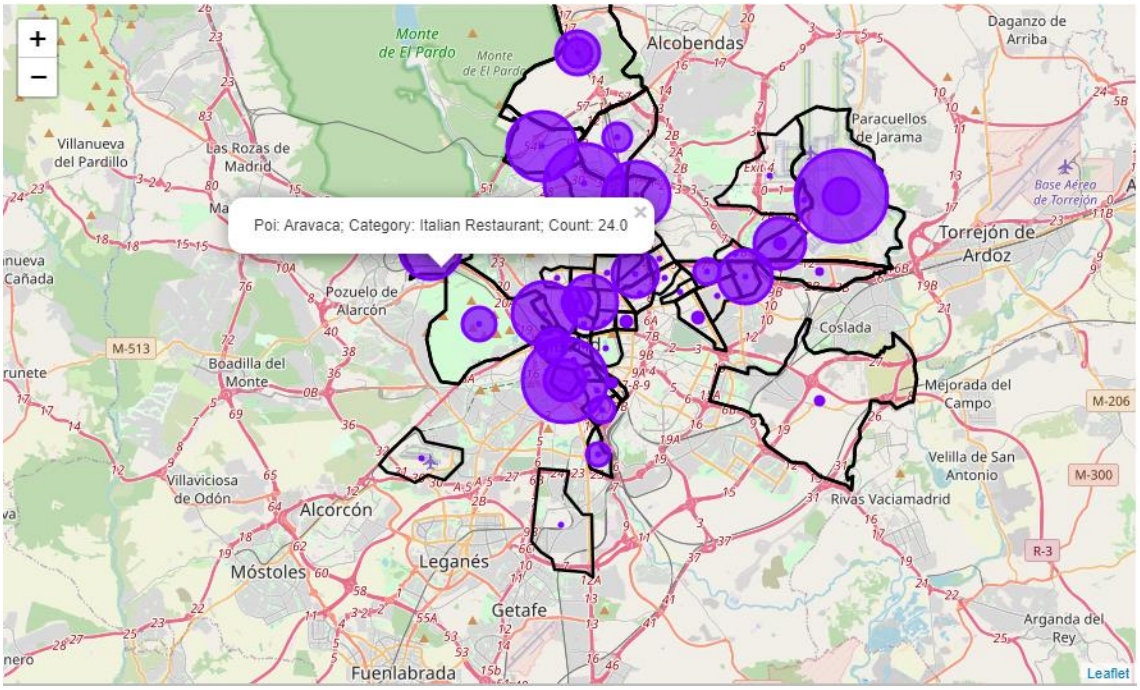
Below extract is data belonging to **Cluster 1**:

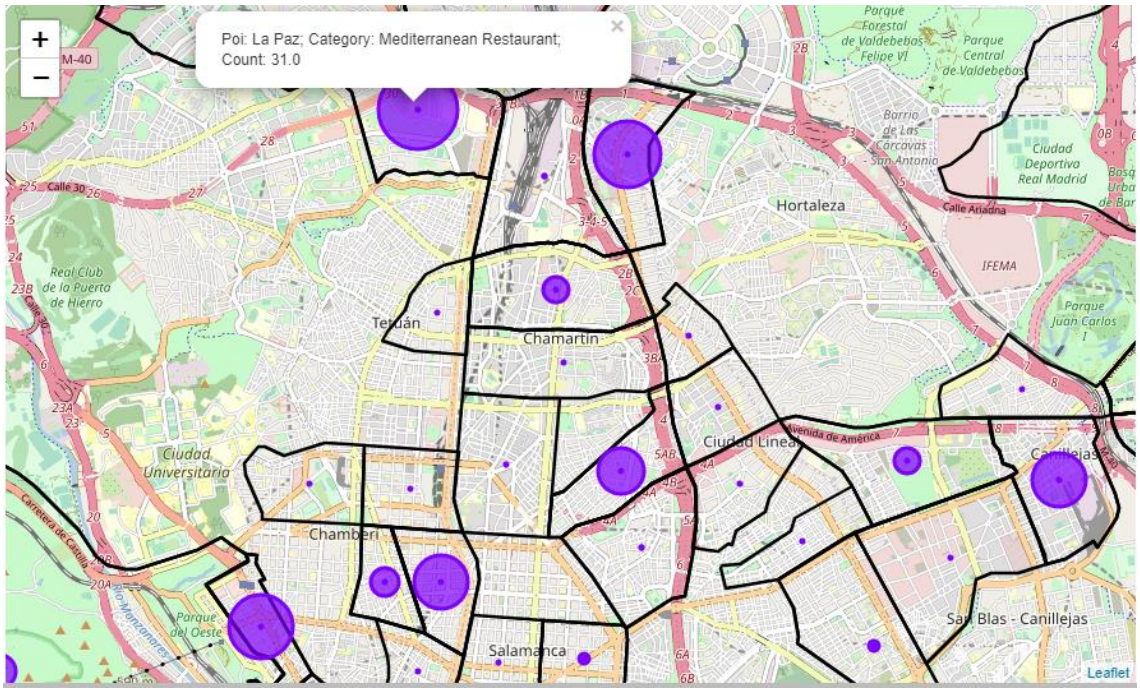
	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mc Comm Ven
0	Palacio	40.416600	-3.712763	0	Spanish Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Gastropub	Argentinian Restaurant	Japanese Restaurant	Bistro	Burg Jo
1	Embajadores	40.410419	-3.701153	0	Spanish Restaurant	Restaurant	Pizza Place	Italian Restaurant	Breakfast Spot	Vegetarian / Vegan Restaurant	Seafood Restaurant	Su Restaura
2	Cortes	40.416025	-3.695522	0	Restaurant	Spanish Restaurant	Seafood Restaurant	Mediterranean Restaurant	Breakfast Spot	Argentinian Restaurant	Japanese Restaurant	Su Restaura
3	Justicia	40.424839	-3.695351	0	Restaurant	Spanish Restaurant	Bistro	Mediterranean Restaurant	Italian Restaurant	American Restaurant	Pizza Place	Sandwi Pla
5	Sol	40.418492	-3.703244	0	Spanish Restaurant	Restaurant	Argentinian Restaurant	Italian Restaurant	Mediterranean Restaurant	Pizza Place	Seafood Restaurant	Mexic Restaura

Let us plot each venue category from Cluster 1 into a Folium map based on the frequency size of establishment type driven from the Count column which will help to guide the potential investors into opening up a new Spanish restaurant in this cluster.

First, we will create a new folium map using the below for data visualization analysis purposes. These maps are interactive in nature and can be consulted and queried upon by investors to learn about the frequency of occurrence of the venue or establishment type. This could useful statistic for determining whether or not to open up a new location or restaurant nearby:

	Neighborhood	Geometry	Latitude	Longitude	Venue Category	Count
0	Acacias	POLYGON ((-3.70124 40.40625, -3.70040 40.40513...	40.402248	-3.705948	Pizza Place	33.0
1	Acacias	POLYGON ((-3.70124 40.40625, -3.70040 40.40513...	40.402248	-3.705948	Spanish Restaurant	16.0
2	Aeropuerto	POLYGON ((-3.53038 40.44818, -3.53179 40.44783...	40.477296	-3.557254	Diner	14.0
3	Aeropuerto	POLYGON ((-3.53038 40.44818, -3.53179 40.44783...	40.477296	-3.557254	Restaurant	36.0
4	Alameda De Osuna	POLYGON ((-3.59194 40.45123, -3.59194 40.45123...	40.457732	-3.590131	Chinese Restaurant	20.0





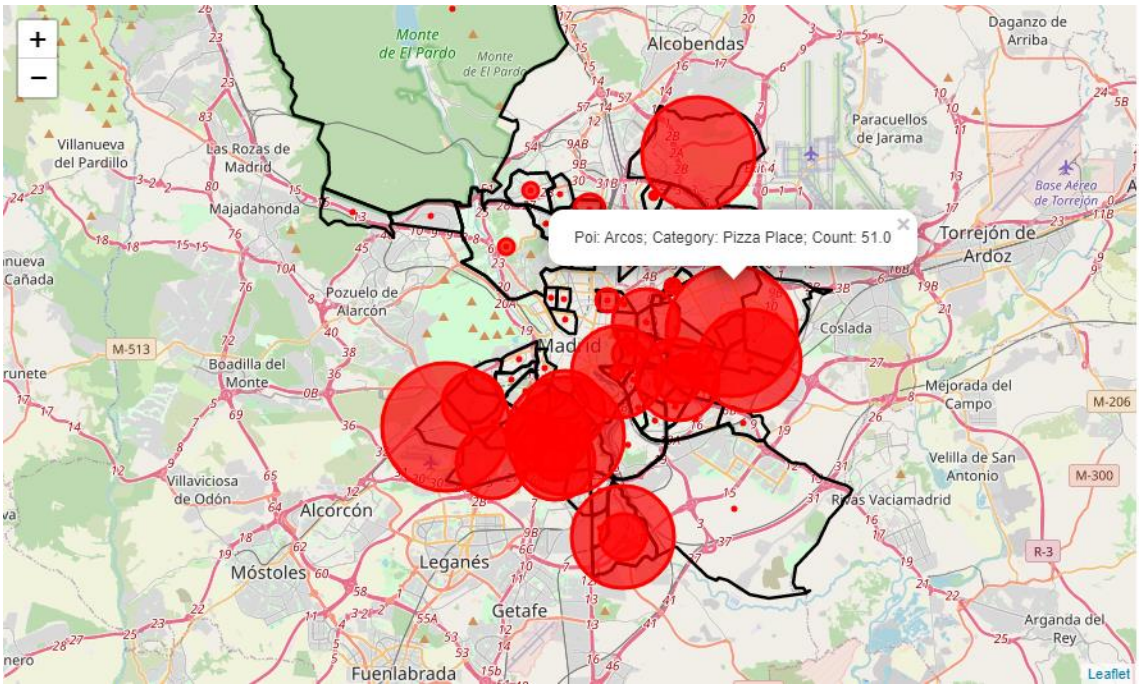
We will repeat the above extracts for Cluster 2 and draw up the conclusions in this report.

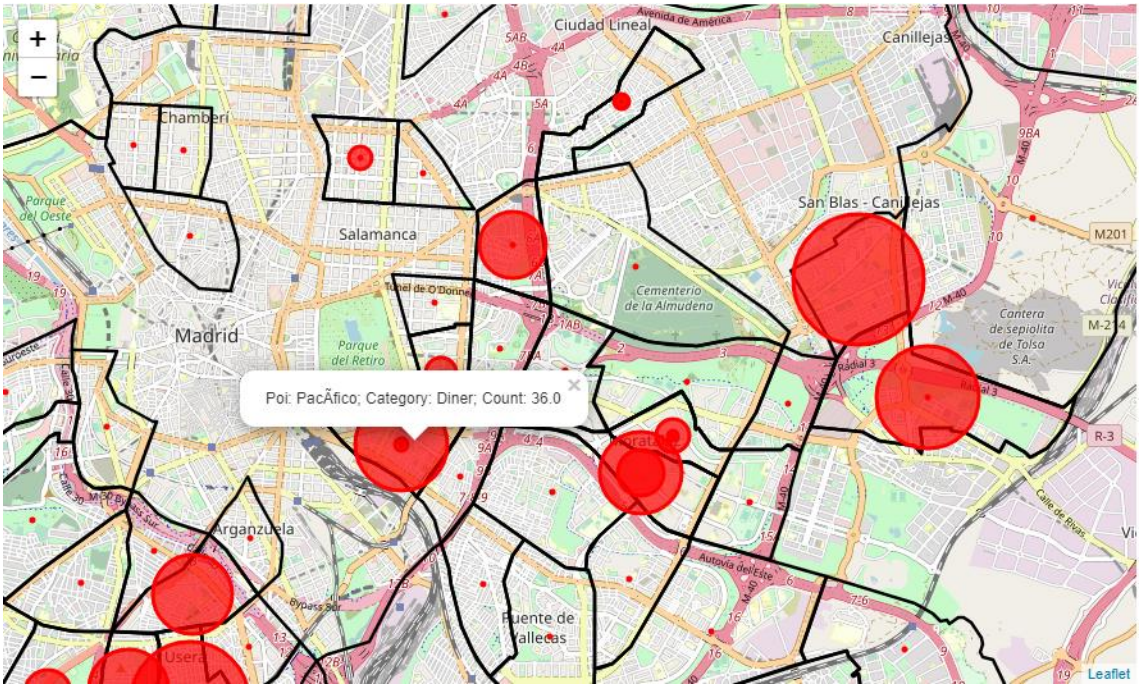
Cluster 2 dataset:

	Neighborhood	Latitude	Longitude	Cluster Labels	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
4	Universidad	40.426850	-3.705760	1	Spanish Restaurant	Restaurant	Gastropub	Pizza Place	Vegetarian / Vegan Restaurant	Italian Restaurant	Mediterranean Restaurant	Argentinian Restaurant
6	Imperial	40.407370	-3.717051	1	Spanish Restaurant	Pizza Place	Restaurant	Sandwich Place	Chinese Restaurant	Fast Food Restaurant	Argentinian Restaurant	Japanese Restaurant
8	Chopera	40.395942	-3.697762	1	Spanish Restaurant	Restaurant	Burger Joint	Pizza Place	Seafood Restaurant	Chinese Restaurant	Mediterranean Restaurant	Italian Restaurant
13	Pacífico	40.405536	-3.677421	1	Spanish Restaurant	Diner	Indian Restaurant	Pizza Place	Burger Joint	Italian Restaurant	Mediterranean Restaurant	Falafel Restaurant
14	Adelfas	40.402297	-3.669665	1	Spanish Restaurant	Breakfast Spot	Restaurant	Korean Restaurant	Sandwich Place	Asian Restaurant	Fast Food Restaurant	Chinese Restaurant

Table extract for data visualization analysis purposes:

	Neighborhood	Geometry	Latitude	Longitude	Venue Category	Count
0	Abrantes	POLYGON ((-3.71625 40.38646, -3.71691 40.38450...	40.380159	-3.724857	Spanish Restaurant	18.0
1	Adelfas	POLYGON ((-3.67221 40.39550, -3.67221 40.39550...	40.402297	-3.669665	Spanish Restaurant	1.0
2	Aguilas	POLYGON ((-3.75800 40.38481, -3.75754 40.38441...	40.382983	-3.769777	Restaurant	50.0
3	Almenara	POLYGON ((-3.68581 40.47533, -3.68624 40.47384...	40.471839	-3.693000	Salad Place	12.0
4	Almenara	POLYGON ((-3.68581 40.47533, -3.68624 40.47384...	40.471839	-3.693000	Spanish Restaurant	1.0


Poi: Arcos; Category: Pizza Place; Count: 51.0
Monte de El Pardo
Alcobendas
Daganzo de Arriba
Paracuellos de Jarama
Villanueva del Pardillo
Las Rozas de Madrid
Majadahonda
Torrejón de Ardoz
Pozuelo de Alarcón
Coslada
Boadilla del Monte
Mejorada del Campo
Villaviciosa de Odón
Velilla de San Antonio
Alcorcón
Rivas Vaciamadrid
Móstoles
Leganés
Getafe
Arganda del Rey
Fuenlabrada
M-513
M-206
M-300
R-3
Leaflet


Poi: PacÃfico; Category: Diner; Count: 36.0
Ciudad Lineal
Canillejas
Chamberí
Salamanca
San Blas - Canillejas
Madrid
Cementerio de la Almudena
Parque del Retiro
Arganzuela
Fuente de Vallecas
M201
M-214
R-3
Leaflet

5. Conclusion

Our analysis shows that although there is a great number of restaurants in Madrid (~6300).

Created candidates and were then clustered to create zones of interest which contain greatest number of location candidates.

Result of all this is 130 neighbourhoods shown and split into two separate clusters in the above maps showing the size of the plots are directly proportionate to the frequency of the same type of restaurant establishment found in each neighbourhood. The idea behind this depiction is to help guide the potential investor about where to open up a new Spanish restaurant in this cluster taken into account of existing competition. These recommended areas or zones should therefore be considered only as a starting point for more detailed analysis.

This capstone project has attempted to show Madrid as an international city with many different types of new restaurant business to offer and I think we have gone through the process of identifying the business problem, specifying the data required, clean the datasets, performing a machine learning algorithm using k-means clustering and providing some useful tips to our stakeholder.

Perhaps, as a recommended next step, to try out other clustering algorithms particularly **DBSCAN** which could help isolate the less densely populated venue areas from the higher ones particularly for cities with huge built-up population areas.

6. References

Deliverable	Description	URL
Jupyter Notebook	Capstone project Python codebase	https://github.com/ssadek1976/Coursera_Capstone/blob/master/Capstone_Battle_of_Neighborhoods_project%20(final).ipynb
Project Report	Detailed Capstone project report documentation	https://github.com/ssadek1976/Coursera_Capstone/blob/master/Predicting_best_location_for_a_new_Spanish_restaurant_Report.pdf
Project Presentation	Executive presentation containing project finds and conclusions	https://github.com/ssadek1976/Coursera_Capstone/blob/master/Predicting_best_location_for_a_new_Spanish_restaurant_Presentation.pdf