IDENTIFYING THE OPTIMAL NEIGHBORHOODS TO OPEN AN ASIAN RESTAURANT IN THE CITY OF MADRID USING MACHINE LEARNING

A. Introduction

This project aims to explore the neighborhoods in the city of Madrid, Spain by using API calls from Squarespace and clustering these neighborhoods using an unsupervised machine learning algorithm – clustering. The goal of the project is to group the neighborhoods based on their similarity to one another in terms of the different kinds of venues located in each neighborhood.

B. Business Problem

Spain is one of the most exotic countries when it comes to food. International cuisines offered in all parts of the country contribute to the diversity of the country and is definitely one of the factors that invites a lot of tourists from all over the world. However, there seems to be a lack of authentic Asian food in some parts of the country. I managed to pinpoint Madrid as the starting point to identify the prevalence and popularity of Asian food in the country and from a high level point of view, there seems to be a significant lack of Asian restaurants in Madrid. So, the problem to be addressed here is to identify the areas in Madrid with no or the fewest number of Asian restaurants so these areas can be prioritized first when it comes to opening up an Asian restaurant.

This business case might be of particular interest to people willing to invest in the restaurant industry specializing in Asian cuisine. The business appeal of this project will be significant for the investors willing to either invest in Asian food services within the city of Madrid or open up an Asian restaurant chain in specific areas of Spain.

C. Data

The data to be used in this project will be a wikipedia entry about the neighborhoods in the Madrid city. The link to the entry is "https://en.wikipedia.org/wiki/List_of_neighborhoods_of_Madrid". The tabular data from this entry will be used in conjunction with the coordinates of the neighborhoods extracted using the Geocoder library as well as the Foursquare API.

The data from the Wikipedia page will be used to obtain the list of neighborhoods in Madrid. The data from the Geocoder library will be used to retrieve the latitude and longitude of each Madrid neighborhood and the data from Foursquare will be used to visualize and cluster the neighborhoods.

D. Methodology

D1. EXTRACTING THE TABULAR DATA FROM THE WIKIPEDIA PAGE

After obtaining all the necessary libraries and packages through the import process, the first step of the exploratory data analysis process is to extract the tabular data from the aforementioned Wikipedia entry

using the Pandas library. The result is a dataframe consisting of five columns – District name and number, District location, Number, Name (Neighborhood) and Image. There is a total of 131 total neighborhoods in the city of Madrid according to the tabular data extracted from the Wikipedia entry (11 of them are displayed below).

	District name (number)	District location	Number	Name	Image
0	Centro (1)	NaN	11	Palacio	NaN
1	Centro (1)	NaN	12	Embajadores	NaN
2	Centro (1)	NaN	13	Cortes	NaN
3	Centro (1)	NaN	14	Justicia	NaN
4	Centro (1)	NaN	15	Universidad	NaN
5	Centro (1)	NaN	16	Sol	NaN
6	Arganzuela (2)	NaN	21	Imperial	NaN
7	Arganzuela (2)	NaN	22	Acacias	NaN
8	Arganzuela (2)	NaN	23	Chopera	NaN
9	Arganzuela (2)	NaN	24	Legazpi	NaN
10	Arganzuela (2)	NaN	25	Delicias	NaN

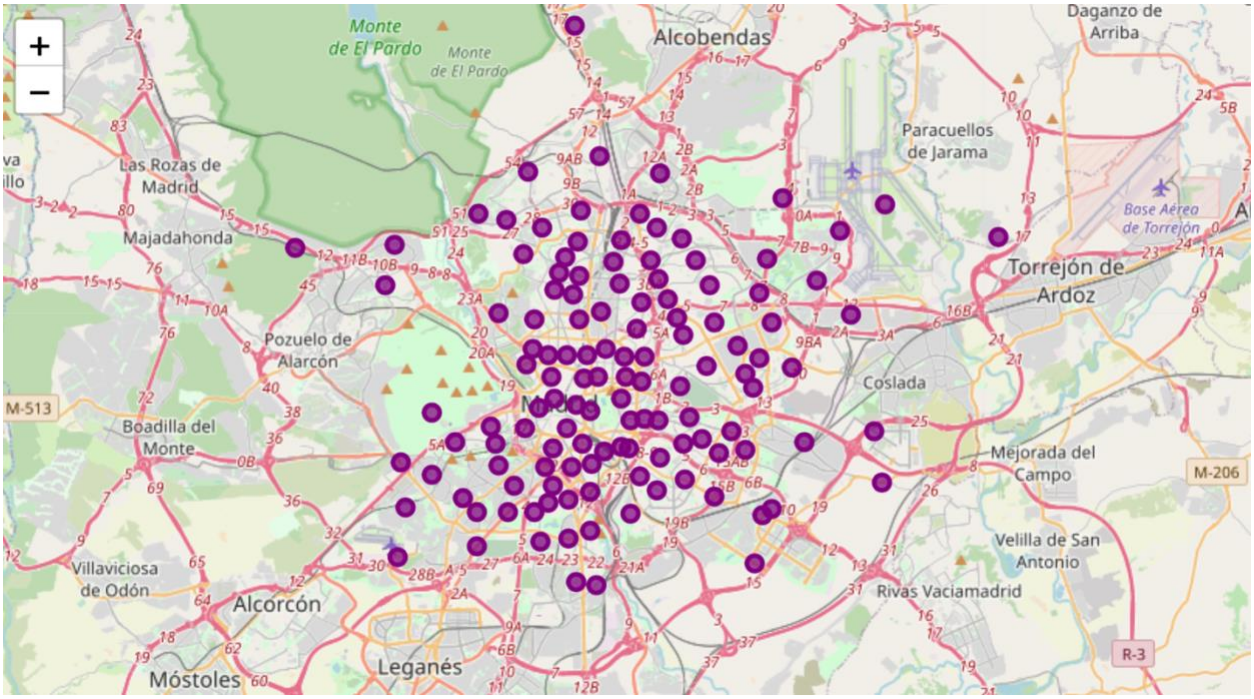
D2. RETRIEIVNG THE COORDINATES FOR THE NEIGHBORHOODS

Geocoder library is then imported to retrieve the latitude and longitude information for each neighbor hood in Madrid city and they are appended to the table above. The geographical coordinates of Mad rid, Spain, according to the Geocoder library, is 40.4167047, -3.7035825.

	District name (number)	Number	Name	Image	Latitude	Longitude
0	Centro (1)	11	Palacio	NaN	40.41517	-3.71273
1	Centro (1)	12	Embajadores	NaN	40.40803	-3.70067
2	Centro (1)	13	Cortes	NaN	40.41589	-3.69636
3	Centro (1)	14	Justicia	NaN	40.42479	-3.69308
4	Centro (1)	15	Universidad	NaN	40.42565	-3.70726
5	Centro (1)	16	Sol	NaN	40.41802	-3.70577
6	Arganzuela (2)	21	Imperial	NaN	40.40833	-3.71865
7	Arganzuela (2)	22	Acacias	NaN	40.40137	-3.70669
8	Arganzuela (2)	23	Chopera	NaN	40.39536	-3.69833
9	Arganzuela (2)	24	Legazpi	NaN	40.38702	-3.68990
10	Arganzuela (2)	25	Delicias	NaN	40.39613	-3.68946

D3. VISAULIZING THE NEIGHBORHOODS

The neighborhoods in the city are visualized using the Folium library.



D4. RETRIEVING THE VENUES IN EACH NEIGHBORHOOD

The venues located in a neighborhood can be retrieved using the Foursquare API. The following is a list of the first five venues in the neighborhood called “Palacio”. The latitude and longitude values of Palacio are 40.415170000000046 & -3.712729999999965 and there is a total of 77 venues around that area.

	name	categories	lat	lng
0	Zuccaru	Ice Cream Shop	40.417179	-3.711674
1	la gastroteca de santiago	Restaurant	40.416639	-3.710944
2	Charlie Champagne	Restaurant	40.413936	-3.712647
3	Santa Iglesia Catedral de Santa María la Real ...	Church	40.415767	-3.714516
4	Plaza de la Villa	Historic Site	40.415409	-3.710391

D5. IDENTIFYING THE NUMBER OF VENUES IN EACH CITY

The total number of venues for each neighborhood is identified. There are altogether 253 unique categories among all the neighborhoods.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Abrantes	8	8	8	8	8	8
Acacias	42	42	42	42	42	42
Adelfas	46	46	46	46	46	46
Alameda de Osuna	24	24	24	24	24	24
Almagro	100	100	100	100	100	100
Almenara	5	5	5	5	5	5
Almendrales	10	10	10	10	10	10
Aluche	13	13	13	13	13	13
Amposta	10	10	10	10	10	10

D6. FINDING THE MEAN OF EACH VENUE IN EACH NEIGHBORHOOD

The mean of each venue in each neighborhood is identified. Shown below is the first nine neighborhoods alongside some of the venues located in them.

	Neighborhood	Zoo Exhibit	Accessories Store	Adult Boutique	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	M
0	Abrantes	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0
1	Acacias	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0
2	Adelfas	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0
3	Alameda de Osuna	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0
4	Almagro	0.0	0.000000	0.00	0.010000	0.000000	0.000000	0.000000	0.010000	0
5	Almenara	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0
6	Almendrales	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0
7	Aluche	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0
8	Amposta	0.0	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0

D7. ASSIGNING CLUSTERS

After obtaining just the mean of Asian restaurants located in each neighborhood, a cluster label ranging from 0 to 4 is assigned to each neighborhood.

	Neighborhood	Asian Restaurant	Cluster Label
0	Abrantes	0.000000	0
1	Acacias	0.023810	4
2	Adelfas	0.021739	4
3	Alameda de Osuna	0.000000	0
4	Almagro	0.020000	4

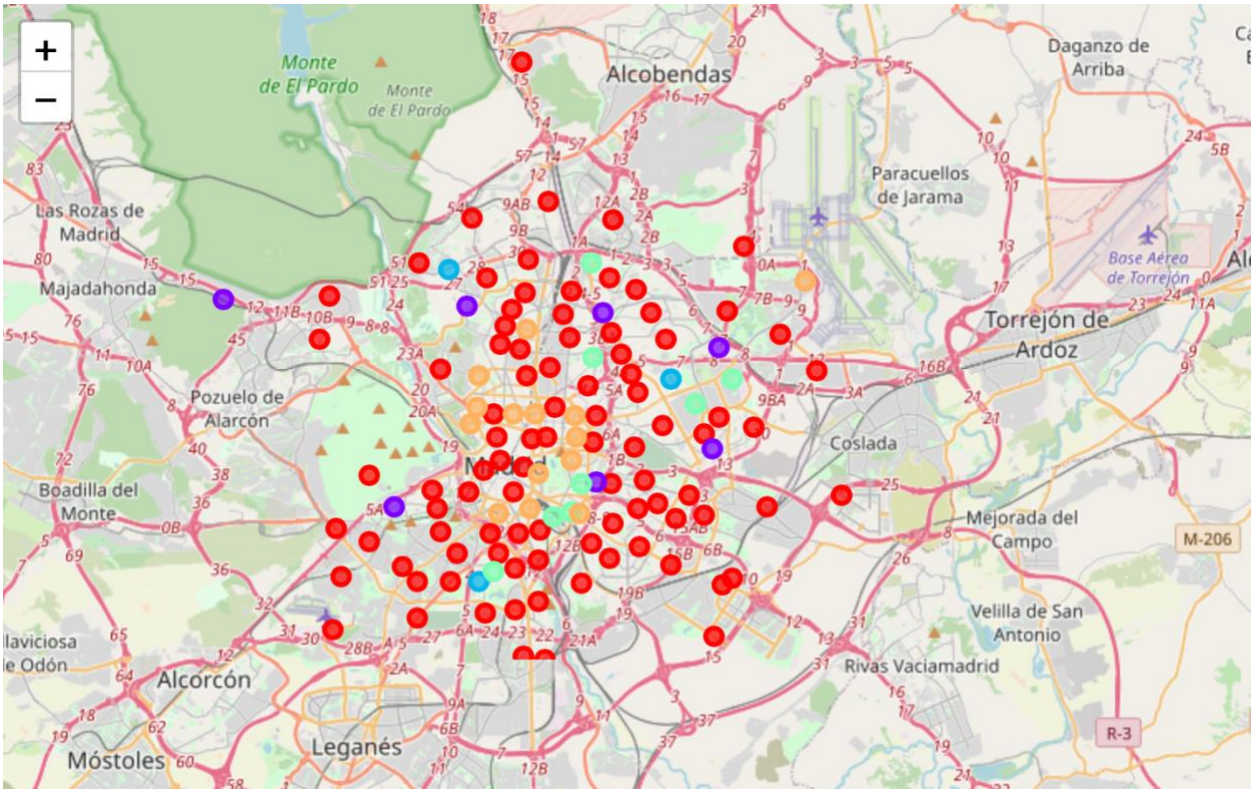
D8. PREPARING THE FINAL TABLE FOR VISUALIZATION

The coordinates table created in the beginning and the cluster table created from the previous step are joined so that the Folium library can be used to pinpoint each neighborhood on the map according to their cluster color.

	Neighborhood	Asian Restaurant	Cluster Label	District name (number)	Number	Image	Latitude	Longitude
0	Abrantes	0.000000	0	Carabanchel (11)	117	NaN	40.37980	-3.72636
1	Acacias	0.023810	4	Arganzuela (2)	22	NaN	40.40137	-3.70669
2	Adelfas	0.021739	4	Retiro (3)	32	NaN	40.40173	-3.67288
3	Alameda de Osuna	0.000000	0	Barajas (21)	211	NaN	40.45818	-3.58953
4	Almagro	0.020000	4	Chamberí (7)	74	NaN	40.43296	-3.69153

D9. VISUALIZING THE CLUSTERS

All the neighborhoods already assigned to a cluster are visualized on the map using the Folium library. Evidently, the red colored neighborhoods (assigned a cluster number of 0) are the most common on the map.



D10. ANALYZING EACH CLUSTER

All five clusters are analyzed to understand the distribution of Asian restaurants in each cluster. The mean of the "Asian Restaurant" column is the main determinant of which cluster has the most Asian restaurants.

	Neighborhood	Asian Restaurant	Cluster Label	District name (number)	Number	Image	Latitude	Longitude
37	Costillares	0.052632	3	Ciudad Lineal (15)	159	NaN	40.48041	-3.66827
93	Pradolongo	0.083333	3	Usera (12)	127	NaN	40.38297	-3.70865
15	Atocha	0.041667	3	Arganzuela (2)	27	NaN	40.40054	-3.68392
30	Ciudad Jardín	0.064516	3	Chamartín (5)	53	NaN	40.45046	-3.66771
109	Simancas	0.052632	3	San Blas-Canillejas (20)	201	NaN	40.43577	-3.62488
81	Pacífico	0.045455	3	Retiro (3)	31	NaN	40.40191	-3.67603
22	Canillejas	0.071429	3	San Blas-Canillejas (20)	207	NaN	40.44373	-3.60977
75	Niño Jesús	0.055556	3	Retiro (3)	36	NaN	40.41095	-3.67230

E. Results

According to the cluster analysis, the Cluster 1 appears to be the most suitable place to open an Asian restaurant due to the fact that the mean for Asian Restaurant is zero for almost all neighborhoods, indicating the lack of Asian restaurants in these neighborhoods. So, it should be reasonably concluded that the neighborhoods that belong to Cluster 1 should be considered for opening an Asian restaurant.

	Neighborhood	Asian Restaurant	Cluster Label	District name (number)	Number	Image	Latitude	Longitude
0	Abrantes	0.000000	0	Carabanchel (11)	117	NaN	40.37980	-3.72636
90	Pinar del Rey	0.000000	0	Hortaleza (16)	164	NaN	40.47225	-3.64943
89	Pilar	0.000000	0	Fuencarral-El Pardo (8)	84	NaN	40.47543	-3.71130
87	Pavones	0.000000	0	Moratalaz (14)	141	NaN	40.40004	-3.63300
85	Palomeras Sureste	0.000000	0	Puente de Vallecas (13)	134	NaN	40.38537	-3.63530
84	Palomeras Bajas	0.000000	0	Puente de Vallecas (13)	133	NaN	40.38756	-3.66029
82	Palacio	0.000000	0	Centro (1)	11	NaN	40.41517	-3.71273
80	Orcasur	0.000000	0	Usera (12)	122	NaN	40.37099	-3.69946
79	Orcasitas	0.000000	0	Usera (12)	121	NaN	40.36985	-3.71231
78	Opañel	0.000000	0	Carabanchel (11)	112	NaN	40.38915	-3.72375
77	Numancia	0.000000	0	Puente de Vallecas (13)	136	NaN	40.39851	-3.65901

F. Discussion

Although Cluster 1 is an ideal neighborhood cluster for opening an Asian restaurant due to the apparent lack of Asian restaurants in these areas, there is a possibility that these areas may not be conducive for such a business plan, which could be the reason there are apparently no Asian restaurants in the Cluster 1 neighborhoods in the first place. So, more research and analytical work should be conducted in order to make sure that the Cluster 1 neighborhoods are really ideal areas to open an Asian restaurant.

G. Conclusion

This project entails several aspects of a real life data science project. Identifying a business plan, structuring the approaches and the methodologies to tackle that problem, interpreting and exploring the data, designing a machine learning algorithm, analyzing the output and putting together recommendations to communicate to the relevant stakeholders are all different components of a data science project that are included in this capstone. The same ideas, principles and approaches employed in this project can be further extended to other real life problems and scenarios, which makes data science a truly worthwhile discipline to pursue for everyone who enjoys challenges and solving problems.