IBM Capstone Project – Battle of the Neighborhoods

1. Introduction

Where would I find a certain cuisine of my choice in the city of Madrid, Spain? This project will look at the districts in Madrid and the types of restaurants in those districts. I was planning to travel to Madrid pre-covid and so decided to look at the different cuisines available in the city and where these restaurants are located. I am hoping my fellow tourists may also be interested to know what type of food is available in which area of the city.

Madrid is the capital and most-populous city of Spain. In 2018, Madrid received 10.21 million tourists of which 53.3% were international tourists. There are many art museums, cultural centers, and bullfighting events. The main annual international events held in Madrid are cycling and tennis Madrid Open. Since 2019, it also hosts Davis Cup which is the finals of the major tournament between men's national teams.

2. Data

To get the information about the districts in Madrid, I will use Wikipedia. The geographical data we need should include the name of the districts, district number and wards of Madrid.

Source for Madrid:

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

We will use Foursquare API to get the venue categories. Foursquare is a database of more than 105 million places worldwide and an API that provides location data for many companies such as Apple, Samsung, Microsoft, and Uber among others.

3. Data Exploring and Cleaning

For the analysis purposes, all null values will be excluded. I will also remove unnecessary data which is not required for the analysis of data in this project.

The district column was renamed to Neighborhood when district data from Wikipedia was merged with the venue data from Foursquare.

4. Methodology

The district data of the city of Madrid was acquired from Wikipedia. The coordinates of Madrid and all the districts in Madrid was obtained using geopy geocoders. Madrid has 20 main districts with many wards in each district. Foursquare API was used to find the venues in these districts and then search was narrowed to restaurants only in these districts.

Folium, which is a Python library for visualizing the geospatial data, was used to map and display restaurants clusters in the districts. Visualization of the Top 10 types of restaurants found in Madid was done using Seaborn visualization library.

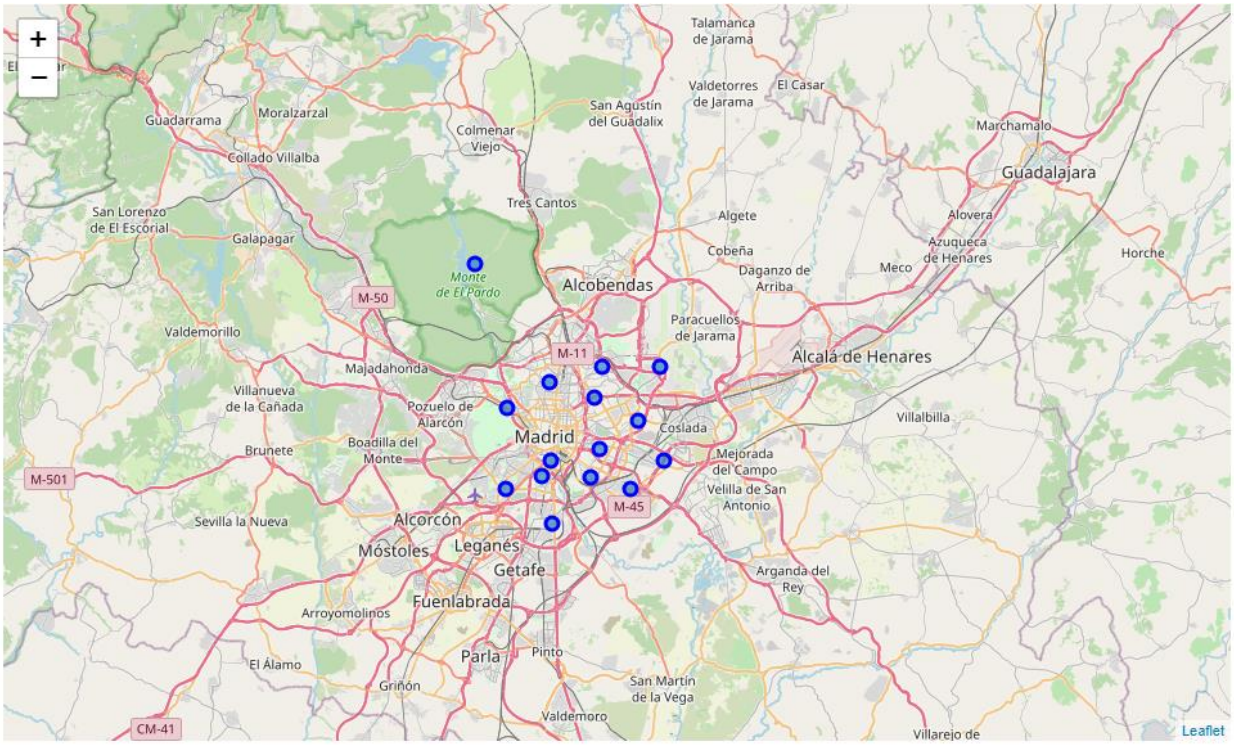
The unsupervised machine learning technique K-Means was used to cluster similar neighborhoods. Elbow method plot was used to find the correct K or the optimal number of clusters to be used in K-Means technique.

5. Results

After cleaning the data, we find there are 20 main districts in Madrid with many wards in each district. Using the geopy geocoder we get the longitude and latitude of these districts.

	District Number	Name	Administrative wards	dist_coord	Latitude	Longitude
0	1.0	Centro	Palacio (11)Embajadores (12)Cortes (13)Justici...	(47.5490251, 1.7324062)	47.549025	1.732406
1	2.0	Arganzuela	Imperial (21)Acacias (22)Chopera (23)Legazpi (...	(40.3969535, -3.6972891)	40.396954	-3.697289
2	3.0	Retiro	Pacífico (31)Adelfas (32)Estrella (33)Ibiza (3...	(6.06171475, -75.51064152773847)	6.061715	-75.510642
3	4.0	Salamanca	Recoletos (41)Goya (42)Fuente del Berro (43)Gu...	(40.9651572, -5.6640182)	40.965157	-5.664018
4	5.0	Chamartín	El Viso (51)Prosperidad (52)Ciudad Jardín (53)...	(40.7018688, -4.9570085)	40.701869	-4.957008
5	6.0	Tetuán	Bellas Vistas (61)Cuatro Caminos (62)Castillej...	(40.4605781, -3.6982806)	40.460578	-3.698281
6	7.0	Chamberí	Gaztambide (71)Arapiles (72)Trafalgar (73)Alma...	(45.5662672, 5.9203636)	45.566267	5.920364
7	8.0	Fuencarral-El Pardo	El Pardo (81)Fuentelarreina (82)Peñagrande (83...	(40.55634555, -3.7785905137518054)	40.556346	-3.778591
8	9.0	Moncloa-Aravaca	Casa de Campo (91)Argüelles (92)Ciudad Univers...	(40.43949485, -3.7442035396547055)	40.439495	-3.744204
9	10.0	Latina	Los Cármenes (101)Puerta del Ángel (102)Lucero...	(41.45952605, 13.012591212188894)	41.459526	13.012591
10	11.0	Carabanchel	Comillas (111)Opañel (112)San Isidro (113)Vist...	(40.3742112, -3.744676)	40.374211	-3.744676
11	12.0	Usera	Orcasitas (121)Orcasur (122)San Fermín (123)Al...	(40.383894, -3.7064459)	40.383894	-3.706446
12	13.0	Puente de Vallecas	Entrevías (131)San Diego (132)Palomeras Bajas ...	(40.3835532, -3.65453548036571)	40.383553	-3.654535
13	14.0	Moratalaz	Pavones (141)Horcajo (142)Marroquina (143)Medi...	(40.4059332, -3.6448737)	40.405933	-3.644874
14	15.0	Ciudad Lineal	Ventas (151)Pueblo Nuevo (152)Quintana (153)Co...	(40.4484305, -3.650495)	40.448431	-3.650495
15	16.0	Hortaleza	Palomas (161)Piovera (162)Canillas (163)Pinar ...	(40.4725491, -3.6425515)	40.472549	-3.642552
16	17.0	Villaverde	Villaverde Alto (171)San Cristóbal (172)Butarq...	(40.3456104, -3.6959556)	40.345610	-3.695956
17	18.0	Villa de Vallecas	Casco Histórico de Vallecas (181)Santa Eugenia...	(40.3739576, -3.6121632)	40.373958	-3.612163
18	19.0	Vicálvaro	Casco Histórico de Vicálvaro (191)Valdebernard...	(40.3965841, -3.5766216)	40.396584	-3.576622
19	20.0	San Blas-Canillejas	Simancas (201)Hellín (202)Amposta (203)Arcos (...	(40.428919050000005, -3.604002428077398)	40.428919	-3.604002
20	21.0	Barajas	Alameda de Osuna (211)Aeropuerto (212)Casco Hi...	(40.4733176, -3.5798446)	40.473318	-3.579845

Folium map shows the districts in the city of Madrid.



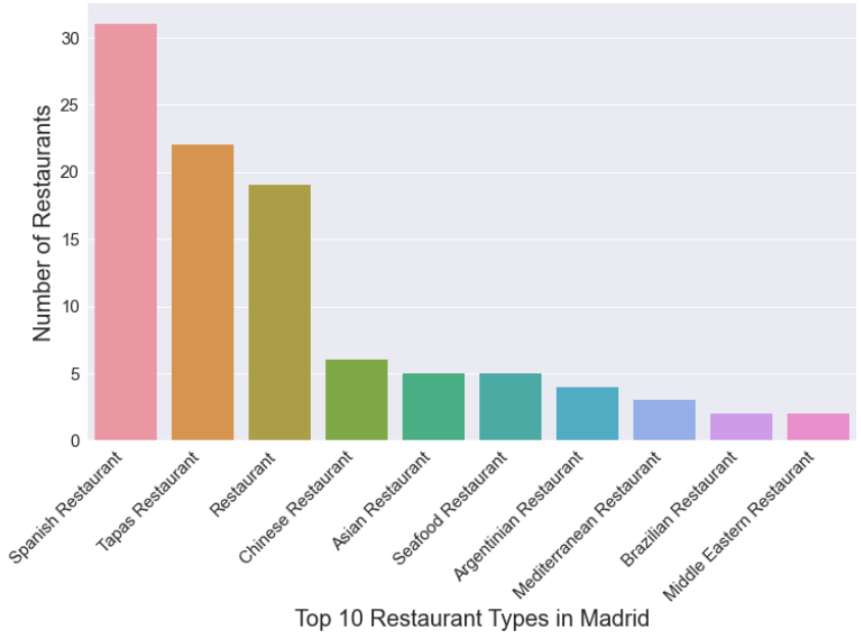
Foursquare API gets the venue name and categories with coordinates in each district.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arganzuela	40.396954	-3.697289	Tres Cerditos	40.397316	-3.694184	Chinese Restaurant
1	Arganzuela	40.396954	-3.697289	Go Hyang Mat	40.396512	-3.699201	Korean Restaurant
2	Arganzuela	40.396954	-3.697289	Trattoria In Crescendo	40.394582	-3.698388	Italian Restaurant
3	Arganzuela	40.396954	-3.697289	Las tinajas	40.396993	-3.697779	Tapas Restaurant
4	Arganzuela	40.396954	-3.697289	Salón de Té Al Yabal	40.399015	-3.700249	Cocktail Bar

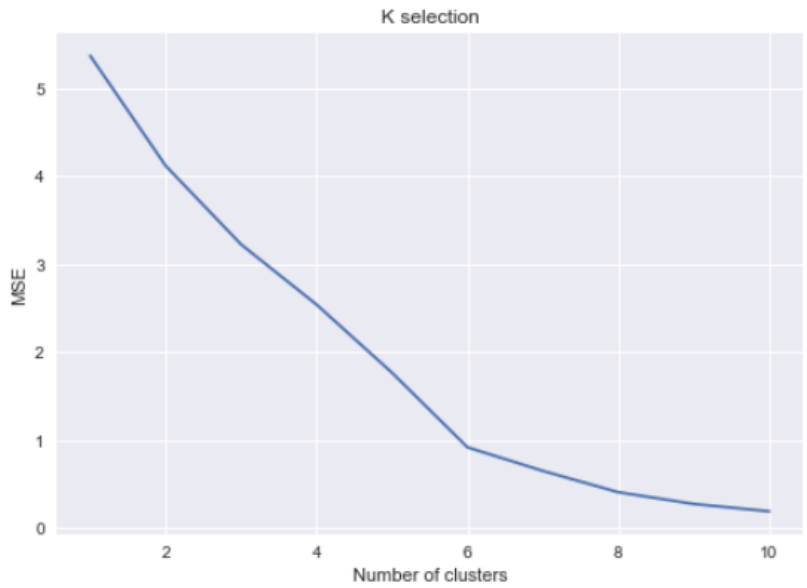
Next, we look at the 10 most common restaurant in each district.

	Neighborhood	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th Most Common Restaurant
0	Arganzuela	Tapas Restaurant	Spanish Restaurant	Restaurant	Chinese Restaurant	Mediterranean Restaurant	Korean Restaurant	Italian Restaurant	Falafel Restaurant	Turkish Restaurant	Fast Food Restaurant
1	Barajas	Spanish Restaurant	Restaurant	Argentinian Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Fast Food Restaurant	Himalayan Restaurant	Turkish Restaurant	Mexican Restaurant
2	Carabanchel	Spanish Restaurant	Restaurant	Tapas Restaurant	Turkish Restaurant	Himalayan Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant
3	Chamberí	French Restaurant	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fast Food Restaurant	Himalayan Restaurant
4	Ciudad Lineal	Spanish Restaurant	Argentinian Restaurant	Restaurant	Chinese Restaurant	Comfort Food Restaurant	Tapas Restaurant	Himalayan Restaurant	Asian Restaurant	Brazilian Restaurant	Falafel Restaurant

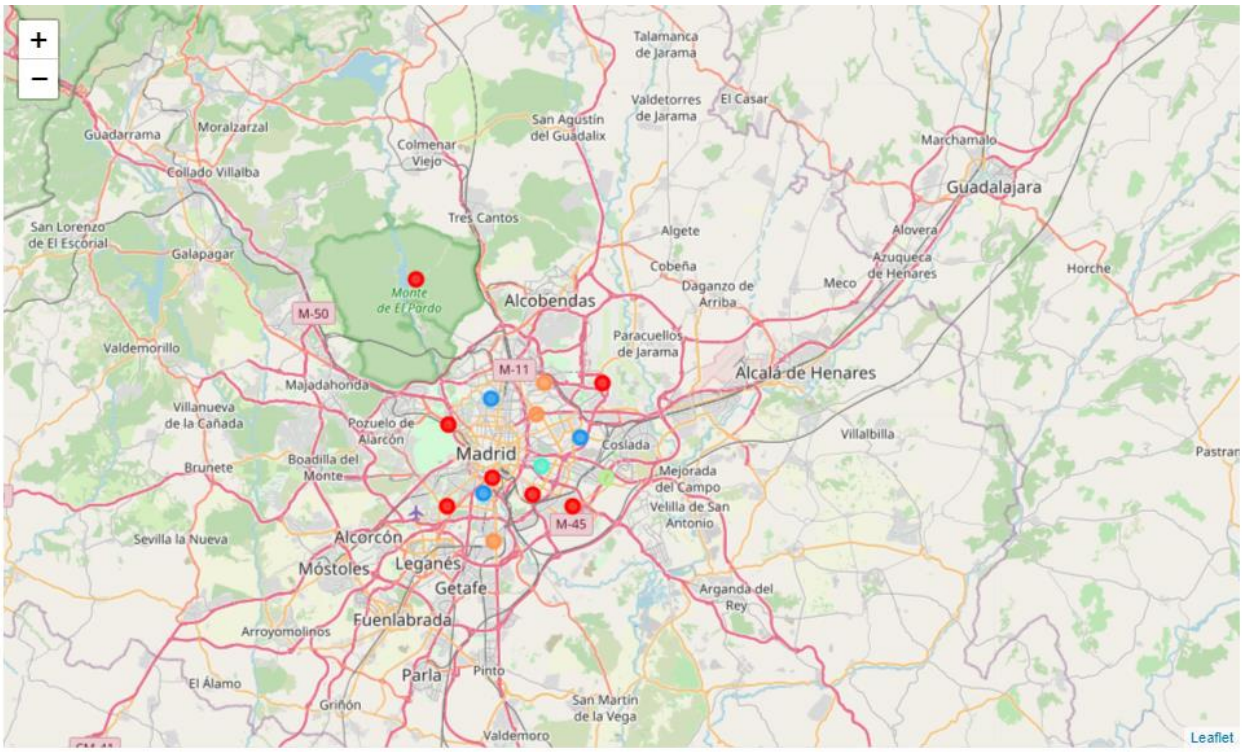
Exploratory Data Analysis visualization below shows that the Top 10 types of restaurants are Spanish and Tapas, followed by Restaurant which are not very specific as to what type of restaurant it is. The other types of restaurants are Chinese, Asian, Seafood, Argentinian, Mediterranean, Brazilian, and Middle Eastern restaurants.



Elbow method plot showed the optimal number of clusters for grouping similar restaurants. In this case, optimal k-value number of clusters was 6. This was used in K-Means to generate the clusters.



We get following clusters using K-Means, we see different colors for 6 clusters. K-Means is a clustering algorithm used in unsupervised machine learning It helps to find what types of groups exits or to identify unknown groups in a data set. This way we can segregate groups with similar traits and assign them into clusters.



For Cluster 1 we get these 10 most common restaurants

	Neighborhood	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10t Co Res
0	Centro	1.732406	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	Retiro	-75.510642	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
4	Chamartín	-4.957008	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
7	Fuencarral-El Pardo	-3.778591	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
8	Moncloa-Aravaca	-3.744204	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9	Latina	13.012591	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
18	Vicálvaro	-3.576622	0	Mediterranean Restaurant	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fast Food Restaurant	Res

Cluster 2 is the largest of all the clusters and has the following restaurants.

	Neighborhood	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10t Co Rest
1	Arganzuela	-3.697289	1	Tapas Restaurant	Spanish Restaurant	Restaurant	Chinese Restaurant	Mediterranean Restaurant	Korean Restaurant	Italian Restaurant	Falafel Restaurant	Turkish Restaurant	Fas Res
3	Salamanca	-5.664018	1	Tapas Restaurant	Spanish Restaurant	Restaurant	Middle Eastern Restaurant	Italian Restaurant	Turkish Restaurant	French Restaurant	Asian Restaurant	Brazilian Restaurant	C Res
5	Tetuán	-3.698281	1	Spanish Restaurant	Restaurant	Brazilian Restaurant	Chinese Restaurant	Japanese Restaurant	Sushi Restaurant	Seafood Restaurant	Asian Restaurant	French Restaurant	C Res
10	Carabanchel	-3.744676	1	Spanish Restaurant	Restaurant	Tapas Restaurant	Turkish Restaurant	Himalayan Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Res
11	Usera	-3.706446	1	Spanish Restaurant	Seafood Restaurant	Chinese Restaurant	Asian Restaurant	Restaurant	Fast Food Restaurant	Turkish Restaurant	Himalayan Restaurant	Brazilian Restaurant	C Res
12	Puente de Vallecas	-3.654535	1	Asian Restaurant	Spanish Restaurant	Tapas Restaurant	Turkish Restaurant	Himalayan Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fas Res
14	Ciudad Lineal	-3.650495	1	Spanish Restaurant	Argentinian Restaurant	Restaurant	Chinese Restaurant	Comfort Food Restaurant	Tapas Restaurant	Himalayan Restaurant	Asian Restaurant	Brazilian Restaurant	Res
19	San Blas-Canillejas	-3.604002	1	Seafood Restaurant	Restaurant	Tapas Restaurant	Turkish Restaurant	Himalayan Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Res
20	Barajas	-3.579845	1	Spanish Restaurant	Restaurant	Argentinian Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Fast Food Restaurant	Himalayan Restaurant	Turkish Restaurant	M Res

Cluster 3 has the following restaurants

	Neighborhood	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th M Comm Restau
13	Moratalaz	-3.644874	2	Sushi Restaurant	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fast Food Restaurant	Fre Restau

Cluster 4 has the following restaurants

	Neighborhood	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th M Comm Restaur
6	Chamberí	5.920364	3	French Restaurant	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fast Food Restaurant	Himalay Restaur

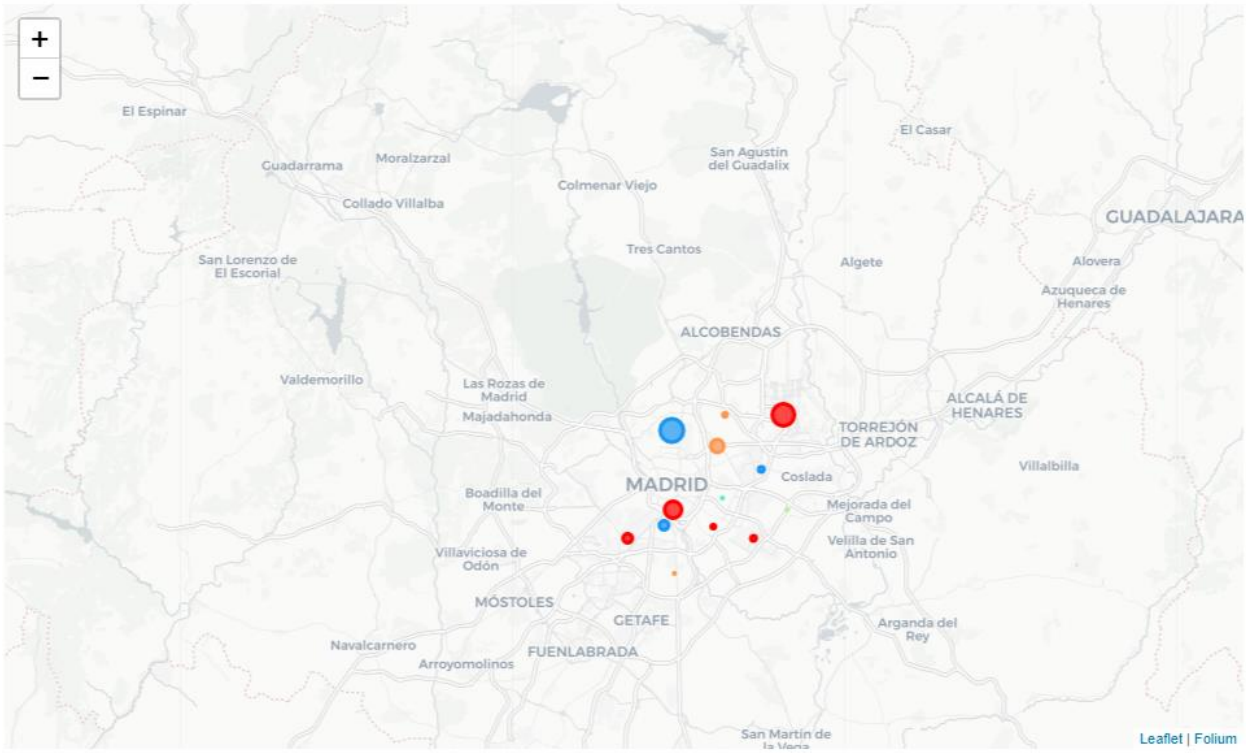
Cluster 5 has the following restaurants

	Neighborhood	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th M Comm Restau
15	Hortaleza	-3.642552	4	Spanish Restaurant	Tapas Restaurant	Turkish Restaurant	Himalayan Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fast F Restau
16	Villaverde	-3.695956	4	Spanish Restaurant	Turkish Restaurant	Italian Restaurant	Asian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fast Food Restaurant	Fre Restau

Cluster 6 has the following restaurants

	Neighborhood	Longitude	Cluster Labels	1st Most Common Restaurant	2nd Most Common Restaurant	3rd Most Common Restaurant	4th Most Common Restaurant	5th Most Common Restaurant	6th Most Common Restaurant	7th Most Common Restaurant	8th Most Common Restaurant	9th Most Common Restaurant	10th M Comm Restau
17	Villa de Vallecas	-3.612163	5	Tapas Restaurant	Asian Restaurant	Turkish Restaurant	Italian Restaurant	Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Falafel Restaurant	Fast Food Restaurant	Fre Restau

We can also see which district has higher number of restaurants when we compare the districts.



6. Discussion

Based on the location of the district you may find different types of restaurants which may be dependent on the population density, customers wanting to try different types of cuisine and the foot traffic to allow for success of the restaurant.

Foursquare API returns venues and categories, but it has limitations since in Cluster 1 there are 6 districts with null values which needed to be dropped so we could visualize which district had more restaurants compared to others. The data in Foursquare may be missing or not identified for those districts or may not be updated. Also, in our exploratory data analysis visualization, the third most frequent restaurant type are not identified correctly, it is only listed as Restaurant. This can possibly skew our analysis of the data since many restaurants are not identified as what type.

In Cluster 1, the most common restaurant is Mediterranean restaurant in Vicalvaro district. In Cluster 2, the most common restaurant is Spanish, Tapas, Asian and Seafood depending on which district. In Cluster 3, the most common restaurant is Sushi restaurant in Moratalaz district. In Cluster 4, the most common restaurant is a French restaurant in Chamberi district. In Cluster 5, the most common restaurant is Spanish restaurant in Hortaleza and Villaverde district. In Cluster 6, the most common restaurant is Tapas restaurant in Villa de Vallecas district.

When we compare the districts, we find that certain districts have many more restaurants than others. Barajas district has the highest number of restaurants, followed by Salamanca, Tetuan, and Arganzuela. Folium map confirms this by showing a larger cluster circle in these districts compared to other districts.

7. Conclusion

This project successfully segmented 20 districts or neighborhoods into 6 clusters applying K-Means technique for machine learning.

Based on these clusters if I wanted to have French cuisine, I would look for it in Chamberi district area. Similarly, if I was in the mood for some Tapas cuisine then I would most likely find it in Arganzuela, Salamanca and Villa de Vallecas districts. If you are in mood for Sushi during your visit to Madrid, try looking for it in Moratalaz district area.