CLUSTERING MADRID

MIGUEL RODRÍGUEZ MORALES

1. PROBLEM DESCRIPTION

In this project, the problem attempted to solve will be to find the best possible location or the most optimal, for a new restaurant in the city of Madrid, Spain. To achieve this task, an analytical approach will be used, based on advanced machine learning techniques and data analysis, concretely clustering and perhaps some data visualization techniques.

During the process of analysis, several data transformations will be performed, in order the find the best possible data format for the machine learning model to ingest. Once the data is set up and prepared, a modeling process will be carried out, and this statistical analysis will provide the best possible places to locate a new bussines in the city of Madrid.

2. DATA

The data be used to develop this project were based on two sites:

1. The Foursquare Api: This data will be accesed via Python, and used to obtain the most common venues per neighborhood in the city of Madrid. This way, it is possible to have a taste of how the city's venues are distributed, what are the most common places for leisure, and in general, it will provide an idea of what people's likes are.
2. The Madrid City Hall's Web Portal: This site provides several data sources of great utility to solve this problem. The files are provided in Excel format, and they are built over a statstical exploitation and use basis. In this case we will use the dsitricts location data and boundaries. You can access the data by clicking this link: 'https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Distritos-en-cifras/Distritos-en-cifras-Informacion-de-Distritos-/?vgnextfmt=default&vgnextoid=74b33ece5284c310VgnVCM1000000b205a0aRCRD&vgnextchannel=27002d05cb71b310VgnVCM1000000b205a0aRCRD'

3. METODOLOGY

The metodology used to aproach this problem includes statistical exploration, data visualization and machine learning techniques. In this case, clustering, in concrete K-Means algorithm was used, implemented with Python.

3.1. Data acquisition on districts

From The Madrid City Hall's Web Portal, and using Pandas library to read, clean and refine the data, the following dataframe is obtained:

	cod	District	Area_km2	Latitude	Longitude
0	1	Centro	5.23	40.415347	-3.707371
1	2	Arganzuela	6.46	40.3960833	-3.6938472
2	3	Retiro	5.47	40.408072	-3.676729
3	4	Salamanca	5.39	40.43	-3.677778
4	5	Chamartin	9.18	40.453333	-3.6775
5	6	Tetuan	5.37	40.460556	-3.7
6	7	Chamberi	4.68	40.432792	-3.697186
7	8	Fuencarral - El Pardo	237.84	40.478611	-3.709722
8	9	Moncloa - Aravaca	46.53	40.4565887	-3.7544059
9	10	Latina	25.43	40.402461	-3.741294
10	11	Carabanchel	14.05	40.383669	-3.727989
11	12	Usera	7.78	40.381336	-3.706856
12	13	Puente de Vallecas	14.97	40.386548	-3.6635396
13	14	Moratalaz	6.10	40.409869	-3.644436
14	15	Ciudad Lineal	11.43	40.4373649	-3.6499612
15	16	Hortaleza	27.42	40.469457	-3.640482
16	17	Villaverde	20.19	40.345925	-3.709356
17	18	Villa de Vallecas	51.47	40.3717661	-3.620269
18	19	Vicalvaro	35.27	40.4042	-3.60806
19	20	San Blas - Canillejas	22.29	40.426001	-3.612764
20	21	Barajas	41.92	40.470196	-3.58489

Table 1. Madrid Districts and coordinates

With this dataframe we plotted a map, including the boundaries of each district.

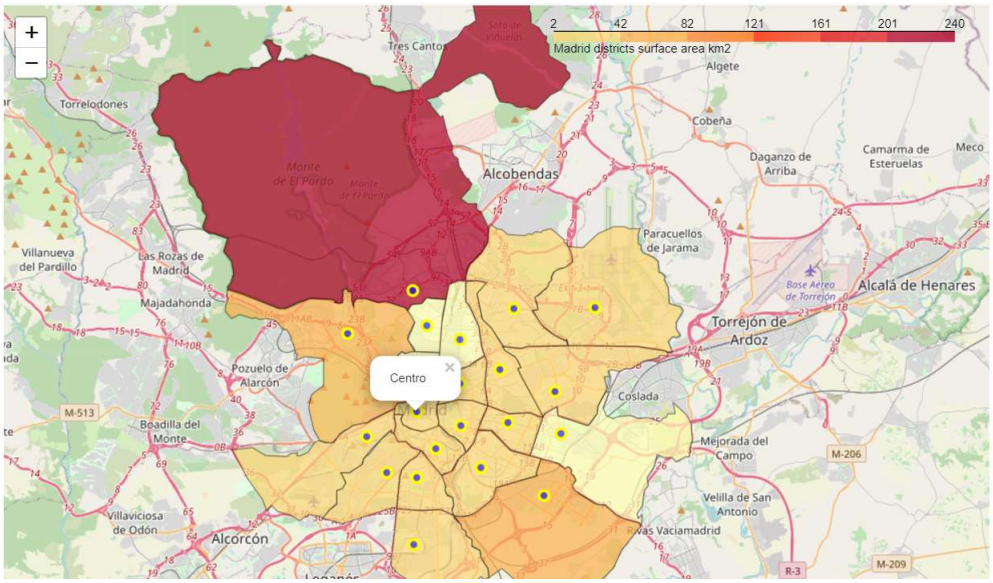

Figure 1. Map showing the Madrid Districts.

3.2. Data acquisition and wrangling on top venues

The Foursquare API is used to get the top 100 venues in each Planning Area, with search radius individualised to each Planning Area. API calls are made to Foursquare by passing the coordinates and search radius of each Planning Area in a Python loop. Foursquare returns the venue data in JSON format. Using the json() function, the venue names, venue latitude, venue longitudes, and venue categories, were extracted and appended into a list. The list is then converted into a pandas dataframe using the pandas.DataFrame() method. The first 5 rows of the dataframe are shown below, which contains 987 rows

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Centro	40.415347	-3.707371	The Hat Madrid	40.414343	-3.707120	Hotel
1	Centro	40.415347	-3.707371	La Taberna de Mister Pinkleton	40.414536	-3.708108	Other Nightlife
2	Centro	40.415347	-3.707371	Plaza Mayor	40.415527	-3.707506	Plaza
3	Centro	40.415347	-3.707371	Gyoza Go!	40.416179	-3.708612	Dumpling Restaurant
4	Centro	40.415347	-3.707371	Bodegas Ricla	40.414266	-3.708077	Wine Bar

Table 2. First five venues from all districts.

This dataframe contains 176 unique venue categories. The pandas.drop_duplicates() method is used to remove duplicates due to overlapping search results from the Foursquare API request. The number of duplicated venues removed is 75.

4. EXPLORATORY ANALYSIS OF THE DATA

After removing duplicates, the data frame is grouped by "Venue Category", and values for each row were counted using the pandas.count() method. The values are sorted by descending order. Table 3 shows the top 10 Venue Categories in all Planning Areas.

	Venue Category	Count
0	Spanish Restaurant	115
1	Restaurant	64
2	Tapas Restaurant	44
3	Bar	40
4	Plaza	33
5	Hotel	28
6	Coffee Shop	21
7	Café	21
8	Bakery	21
9	Gym	16

Table 3. Top 10 venues from all districts.

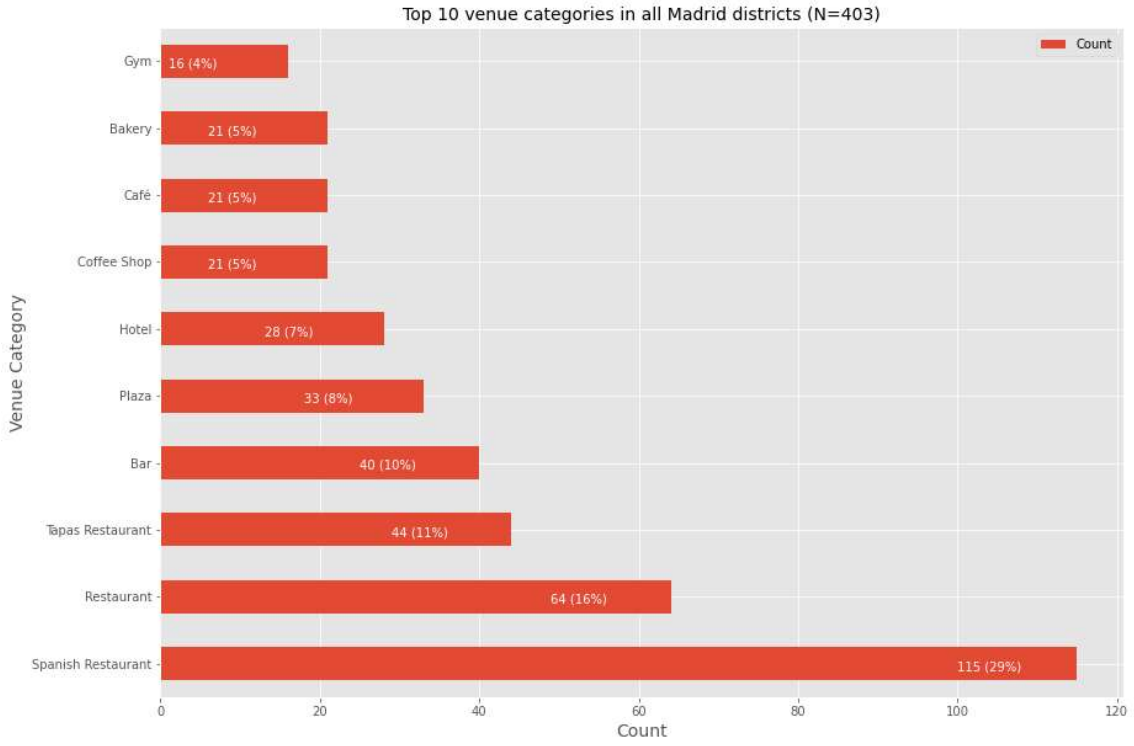

Figure 2. Bar chart of top 10 Venues.

5. K- MEANS CLUSTERING

K-Means clustering is a type of partition clustering that divides data into K non-overlapping subsets or clusters without any cluster internal structure or labels. It is an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

The values were fitted using the KMeans() function from sklearn.cluster library, with number of clusters = 5.

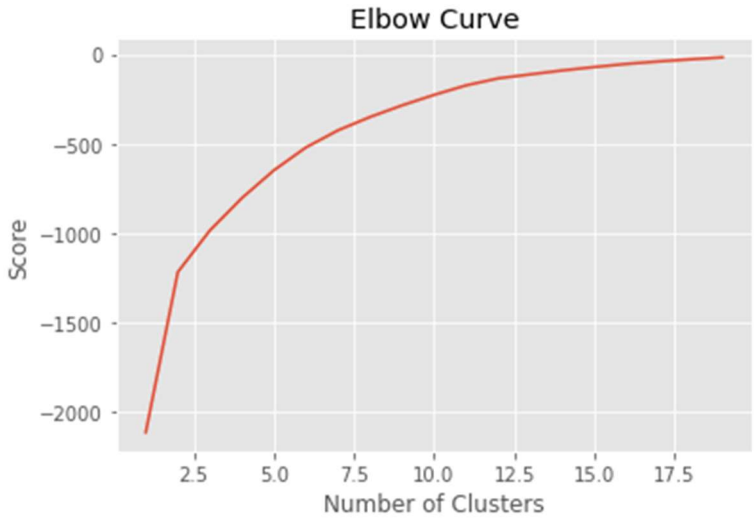

Figure 3. Optimal k value

6. RESULTS

Using the Map() function from folium library, all Planning Areas are plotted on the map, with different coloured map markers according to their cluster labels.

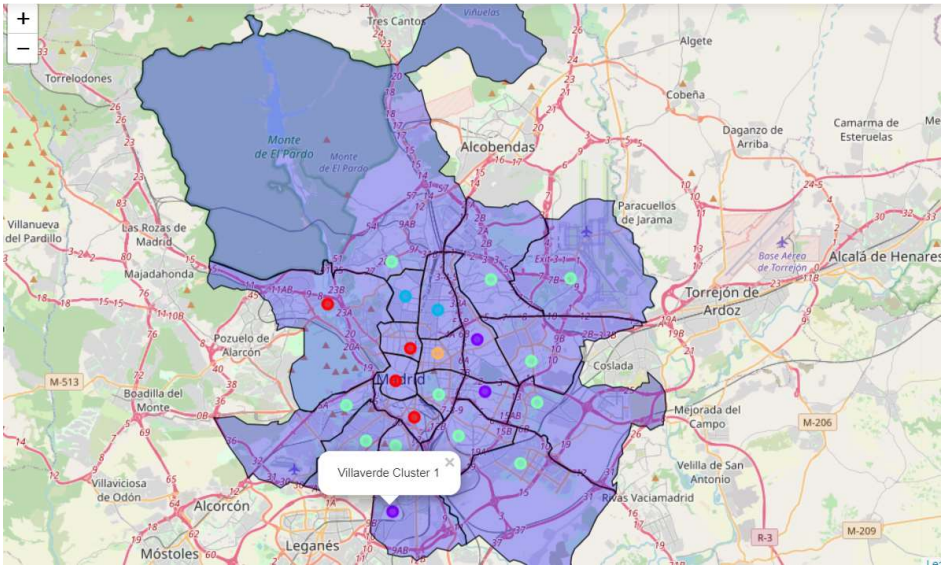

Figure 5. Map of Madrid Districts with coloured-coded clusters

The merged dataframe can be grouped by clusters, and a summary of each cluster is obtained for further analysis. Also we plot a plot a WordCloud to see the most common venues in each cluster.

Cluster 0

	District	1st Most Common Venues	2nd Most Common Venues	3rd Most Common Venues	4th Most Common Venues	5th Most Common Venues	6th Most Common Venues	7th Most Common Venues	8th Most Common Venues	9th Most Common Venues	10th Most Common Venues
0	Centro	Spanish Restaurant	Tapas Restaurant	Plaza	Hotel	Hostel	Ice Cream Shop	Mexican Restaurant	Cocktail Bar	Pastry Shop	Gourmet Shop
1	Arganzuela	Spanish Restaurant	Restaurant	Tapas Restaurant	Bakery	Beer Garden	Plaza	Mediterranean Restaurant	Gym / Fitness Center	Dessert Shop	Café
6	Chamberi	Restaurant	Spanish Restaurant	Plaza	Bar	Tapas Restaurant	Café	Brewery	Mediterranean Restaurant	Japanese Restaurant	Hotel
8	Moncloa - Aravaca	Golf Course	Fast Food Restaurant	Café	Tennis Court	American Restaurant	Noodle House	Opera House	Optical Shop	Other Nightlife	Paella Restaurant

Table 4. Cluster 0 districts.

Figure 6. Wordcloud of cluster 0

Cluster 1

	District	1st Most Common Venues	2nd Most Common Venues	3rd Most Common Venues	4th Most Common Venues	5th Most Common Venues	6th Most Common Venues	7th Most Common Venues	8th Most Common Venues	9th Most Common Venues	10th Most Common Venues
13	Moratalaz	Bar	Park	Ice Cream Shop	Plaza	Coffee Shop	Café	Nightclub	Brewery	Breakfast Spot	Pub
14	Ciudad Lineal	Spanish Restaurant	Tapas Restaurant	Restaurant	Bar	Park	Bakery	Plaza	Fast Food Restaurant	Sporting Goods Shop	Clothing Store
16	Villaverde	Train Station	Spanish Restaurant	Mediterranean Restaurant	Café	Diner	Gastropub	Mobile Phone Shop	Metro Station	Design Studio	Train

Table 5. Cluster 1 districts.

Figure 7. Wordcloud of cluster 1

Cluster 2

	District	1st Most Common Venues	2nd Most Common Venues	3rd Most Common Venues	4th Most Common Venues	5th Most Common Venues	6th Most Common Venues	7th Most Common Venues	8th Most Common Venues	9th Most Common Venues	10th Most Common Venues
4	Chamartin	Spanish Restaurant	Restaurant	Tapas Restaurant	Gastropub	Pizza Place	Plaza	Grocery Store	Café	Mexican Restaurant	Bar
5	Tetuan	Spanish Restaurant	Restaurant	Coffee Shop	Tapas Restaurant	Bar	Asian Restaurant	Hotel	Breakfast Spot	Gym	Grocery Store

Table 6. Cluster 2 districts.

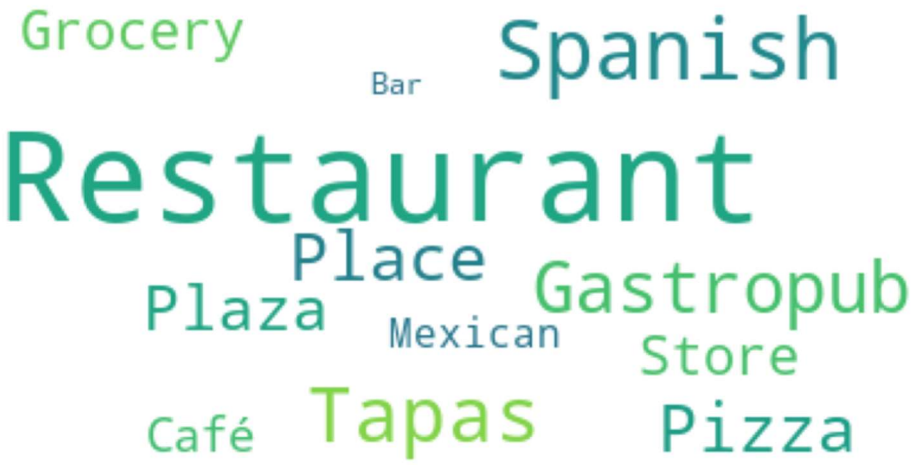

Figure 8. Wordcloud of cluster 2

Cluster 3

	District	1st Most Common Venues	2nd Most Common Venues	3rd Most Common Venues	4th Most Common Venues	5th Most Common Venues	6th Most Common Venues	7th Most Common Venues	8th Most Common Venues	9th Most Common Venues	10th Most Common Venues
2	Retiro	Spanish Restaurant	Bar	Café	Hotel	Brewery	Pizza Place	Museum	Grocery Store	Gym	Plaza
7	Fuencarral - El Pardo	Restaurant	Bar	Tapas Restaurant	Fast Food Restaurant	Café	Park	Spanish Restaurant	Clothing Store	Gym / Fitness Center	Video Game Store
9	Latina	Pizza Place	Lake	Student Center	Scenic Lookout	Tapas Restaurant	Metro Station	Sandwich Place	Gym	Park	Fast Food Restaurant
10	Carabanchel	Plaza	Metro Station	Bakery	Grocery Store	Colombian Restaurant	Soccer Field	Café	Nightclub	Spanish Restaurant	Mobile Phone Shop
11	Usera	Seafood Restaurant	Chinese Restaurant	Theater	Nightclub	Restaurant	Bakery	Spanish Restaurant	Asian Restaurant	Bubble Tea Shop	Noodle House
12	Puente de Vallecas	Spanish Restaurant	Pub	Tapas Restaurant	Gym	Music Venue	Food & Drink Shop	Park	Concert Hall	Restaurant	Coffee Shop
15	Hortaleza	Spanish Restaurant	Plaza	Restaurant	Gym	Pizza Place	Pharmacy	Pub	Cosmetics Shop	Shopping Mall	Snack Place
17	Villa de Vallecas	Restaurant	Soccer Field	Food Truck	Gym	Park	Basketball Court	Tapas Restaurant	Bakery	Cupcake Shop	Asian Restaurant
18	Vicalvaro	Park	Spanish Restaurant	Ice Cream Shop	Grocery Store	Pizza Place	Restaurant	Food & Drink Shop	Café	Plaza	Breakfast Spot
19	San Blas - Canillejas	Gym	Beer Garden	Bar	Snack Place	Chinese Restaurant	Breakfast Spot	Grocery Store	Metro Station	Spanish Restaurant	Pizza Place
20	Barajas	Hotel	Spanish Restaurant	Coffee Shop	Restaurant	Argentinian Restaurant	Tapas Restaurant	Wine Bar	Deli / Bodega	Grocery Store	Supermarket

Table 6. Cluster 3 districts.

Figure 9. Wordcloud of cluster 3

Cluster 4

	District	1st Most Common Venues	2nd Most Common Venues	3rd Most Common Venues	4th Most Common Venues	5th Most Common Venues	6th Most Common Venues	7th Most Common Venues	8th Most Common Venues	9th Most Common Venues	10th Most Common Venues
3	Salamanca	Spanish Restaurant	Restaurant	Seafood Restaurant	Mediterranean Restaurant	Boutique	Burger Joint	Tapas Restaurant	Coffee Shop	Bakery	Bar

Table 5. Cluster 4 districts.

Figure 10. Wordcloud of cluster 4

7. DISCUSSION

As you can see, most of the venues are food related: restaurants, bar, tapas bar, etc. Each cluster has its own characteristics, but also common spots with other clusters.

As a recommendation, it must be said in a study of this size, to make good predictions about where to open a certain business or shop, more data is needed. For example, socio-demographic data about the population, like their income level, if they have children or not, the education level, what kind of job do they make a living from, etc.... Also, one of the most important data to examine carefully are the data related to the people's likes and tastes about how they prefer to spend their leisure time, what kinds of food do they like, or what are their hobbies. With all these data gathered, a more in depth analysis could be performed, and the segmentations would be more accurate. For this project, these data weren't available, and was also out of the project's scope.

8. CONCLUSSIONS

As said before, most of the venues in Madrid districts are food related. There are some other bussines like in the case of Salamanca district where boutiques and exclusive shops are in the top ten list.

So, as an easy conclussion, we can say that is not a good idea to open a restaurant in Madrid. There are tons of them. But in any case, a deeper study is needed.