The Battle of Neighborhoods — Coursera IBM Capstone Project

1) Problem Description

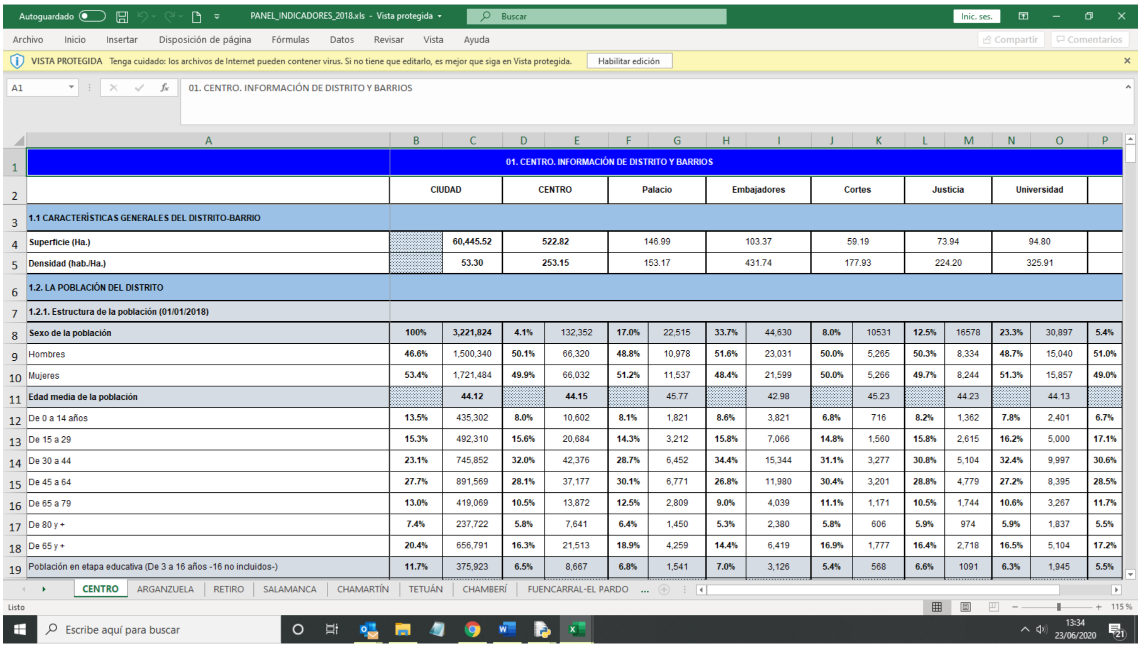
The basis of this study is to help an investor to open a new business in Madrid. The investor needs to know the most common venues in the city and the areas where the businesses are located.

The objective is to identify the top ten common venues for each Neighborhood in Madrid classified by category and locate them in clusters in a map.

2) Data

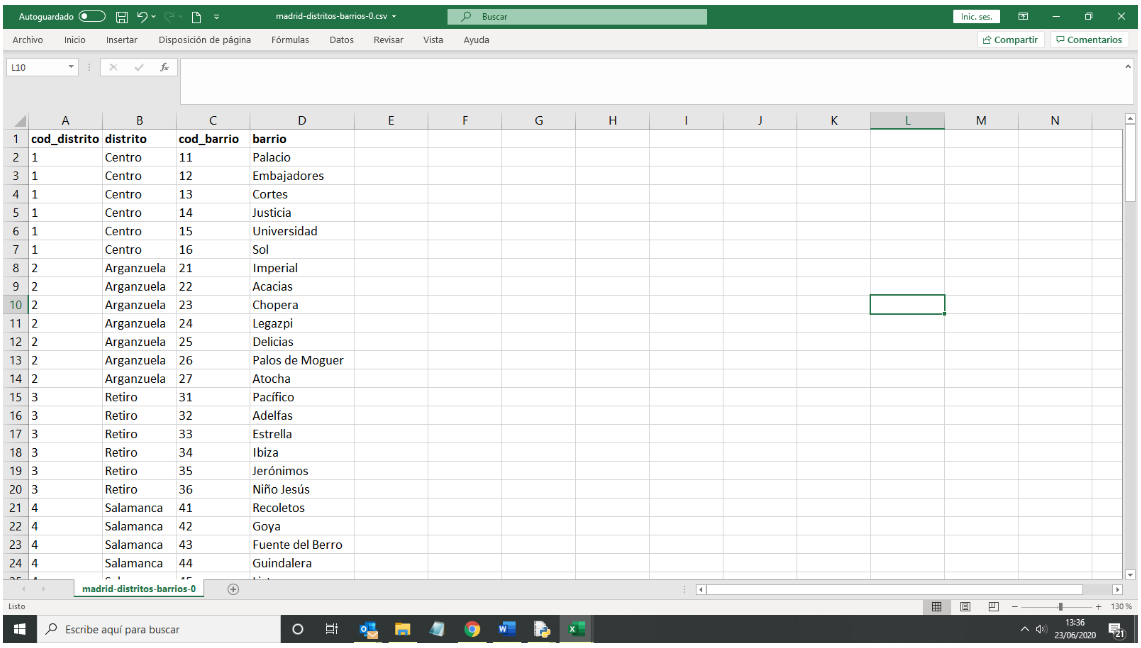
The information needed about districts, neighborhoods and population can be found in the website "Portal de datos abiertos del Ayuntamiento de Madrid": https://datos.madrid.es/portal/site/egob

From this website can be obtained a excel file ("PANEL_INDICADORES") with several indicators about the population classified by neighborhood from which the necessary data will be extracted, such as district and neighborhood codes and their names that will be used for it.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		01. CENTRO. INFORMACIÓN DE DISTRITO Y BARRIOS														
2		CIUDAD		CENTRO		Palacio		Embajadores		Cortes		Justicia		Universidad		
3	1.1 CARACTERÍSTICAS GENERALES DEL DISTRITO-BARRIO															
4	Superficie (Ha.)		60,445.52		522.82		146.99		103.37		59.19		73.94		94.80	
5	Densidad (hab./Ha.)		53.30		253.15		153.17		431.74		177.93		224.20		325.91	
6	1.2. LA POBLACIÓN DEL DISTRITO															
7	1.2.1. Estructura de la población (01/01/2018)															
8	Sexo de la población	100%	3,221,824	4.1%	132,352	17.0%	22,515	33.7%	44,630	8.0%	10531	12.5%	16578	23.3%	30,897	5.4%
9	Hombres	46.6%	1,500,340	50.1%	66,320	48.8%	10,978	51.6%	23,031	50.0%	5,265	50.3%	8,334	48.7%	15,040	51.0%
10	Mujeres	53.4%	1,721,484	49.9%	66,032	51.2%	11,537	48.4%	21,599	50.0%	5,266	49.7%	8,244	51.3%	15,857	49.0%
11	Edad media de la población		44.12		44.15		45.77		42.98		45.23		44.23		44.13	
12	De 0 a 14 años	13.5%	435,302	8.0%	10,602	8.1%	1,821	8.6%	3,821	6.8%	716	8.2%	1,362	7.8%	2,401	6.7%
13	De 15 a 29	15.3%	492,310	15.6%	20,684	14.3%	3,212	15.8%	7,066	14.8%	1,560	15.8%	2,615	16.2%	5,000	17.1%
14	De 30 a 44	23.1%	745,852	32.0%	42,376	28.7%	6,452	34.4%	15,344	31.1%	3,277	30.8%	5,104	32.4%	9,997	30.6%
15	De 45 a 64	27.7%	891,569	28.1%	37,177	30.1%	6,771	26.8%	11,980	30.4%	3,201	28.8%	4,779	27.2%	8,395	28.5%
16	De 65 a 79	13.0%	419,069	10.5%	13,872	12.5%	2,809	9.0%	4,039	11.1%	1,171	10.5%	1,744	10.6%	3,267	11.7%
17	De 80 y +	7.4%	237,722	5.8%	7,641	6.4%	1,450	5.3%	2,380	5.8%	606	5.9%	974	5.9%	1,837	5.5%
18	De 65 y +	20.4%	656,791	16.3%	21,513	18.9%	4,259	14.4%	6,419	16.9%	1,777	16.4%	2,718	16.5%	5,104	17.2%
19	Población en etapa educativa (De 3 a 16 años -16 no incluidos-)	11.7%	375,923	6.5%	8,667	6.8%	1,541	7.0%	3,126	5.4%	568	6.6%	1091	6.3%	1,945	5.5%

The information about districts and neighborhoods will be used to obtain de geographical coordinates (latitude and longitude) where these are located by using the geocoder library.



cod_distrito	distrito	cod_barrio	barrio
1	Centro	11	Palacio
1	Centro	12	Embajadores
1	Centro	13	Cortes
1	Centro	14	Justicia
1	Centro	15	Universidad
1	Centro	16	Sol
2	Arganzuela	21	Imperial
2	Arganzuela	22	Acacias
2	Arganzuela	23	Chopera
2	Arganzuela	24	Legazpi
2	Arganzuela	25	Delicias
2	Arganzuela	26	Palos de Moguer
2	Arganzuela	27	Atocha
3	Retiro	31	Pacífico
3	Retiro	32	Adelfas
3	Retiro	33	Estrella
3	Retiro	34	Ibiza
3	Retiro	35	Jerónimos
3	Retiro	36	Niño Jesús
4	Salamanca	41	Recoletos
4	Salamanca	42	Goya
4	Salamanca	43	Fuente del Berro
4	Salamanca	44	Guindalera

The Foursquare API will be used to collect information about the venues and possible competitors in the neighborhoods of Madrid.

With the venues obtained from Foursquare it will be possible to classify them by category and finally stablish a clustering for neighborhood by means of KMeans algorithm.

3) Methodology

These are the sequential steps necessary to identify the top ten common venues for each neighborhood in Madrid classified by category and located in clusters in a map.

- The first step is to obtain a dataframe with codes and descriptions about districts and neighborhoods of Madrid city.
- Next, it is necessary to clean the dataframe, define columns and drop duplicates.
- Then, call argcis from geocoder library to obtain the latitude and longitude coordinates for each neighborhood in the dataframe. This will be necessary to find the Madrid venues by means of the Foursquare API.

Code	Borough	Neighborhood	Latitude	Longitude
111	Centro	Palacio	40.409630	-3.879790

- After that, by calling the Foursquare API for each neighborhood, the venues are obtained within a radius of 500 meters.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Palacio	40.40963	-3.87979	Proverbium	40.408192	-3.877232	Italian Restaurant

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Palacio	40.40963	-3.87979	The London Walk	40.408197	-3.880682	Irish Pub
2	Palacio	40.40963	-3.87979	Go!Sushing	40.408424	-3.880492	Japanese Restaurant
3	Palacio	40.40963	-3.87979	VIPS Boadilla	40.408106	-3.880623	Burger Joint
4	Palacio	40.40963	-3.87979	El Rincón Del Bierzo	40.409606	-3.880125	Mediterranean Restaurant
...	...	...	...	...	...	...	...
482	Alameda de Osuna	40.45818	-3.58953	Dia %	40.455036	-3.586630	Grocery Store
483	Alameda de Osuna	40.45818	-3.58953	El Kiosko de Pepe	40.454098	-3.589498	Bookstore
484	Alameda de Osuna	40.45818	-3.58953	Hiper Bazar Padre Nuestro	40.455040	-3.585681	Shop & Service
485	Alameda de Osuna	40.45818	-3.58953	Gin Terrace Hilton Madrid	40.454964	-3.585670	Cocktail Bar
486	Alameda de Osuna	40.45818	-3.58953	Plaza Del Navio	40.454987	-3.585642	Plaza

487 rows × 7 columns

- By using Pandas, obtain the places by categories organizing them in columns, grouping them by Neighborhood, and classifying them according to the number of repetitions.

- With this new dataframe, get the top ten common venues for each Neighborhood based on the number of repetitions.

	Neighborhood	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
0	Alameda de Osuna	Restaurant	Tapas Restaurant	Plaza	Bookstore	Hobby Shop	Fried Chicken Joint	Italian Restaurant	Metro Station	Cocktail Bar	Pizza Place
1	Bellas Vistas	Spanish Restaurant	Bar	Grocery Store	Bakery	Supermarket	Tapas Restaurant	Pizza Place	Coffee Shop	Seafood Restaurant	Farmers Market
2	Casa de Campo	Pub	Stadium	Spanish Restaurant	Pool	Gym	Grocery Store	Fast Food Restaurant	Convenience Store	Cuban Restaurant	Deli / Bodega
3	Casco Histórico de Vallecas	Pizza Place	Scenic Lookout	Bakery	Yoga Studio	Food	Department Store	Dessert Shop	Diner	Discount Store	Donut Shop
4	Comillas	Coffee Shop	Fast Food Restaurant	Grocery Store	Supermarket	Flea Market	Plaza	Farmers Market	Dumpling Restaurant	Falafel Restaurant	Electronics Store

- Now, through K-Means Clustering unsupervised algorithm, it divides the data into K non-overlapping clusters grouping similar venues. It will be used for K=5.

- Next, merge the madrid data containing neighborhoods and coordinates, with the neighborhoods venues clustered.
- Finally, through the folium library show a map with the clusters.

4) Results

With the data now ready, we run k-means to cluster the neighborhoods into five clusters. The cluster number was established after multiple samplings and iterations. With our clusters established, this dataframe is merged with the total scores data to provide us with our final pieces of criteria in selecting the appropriate neighborhood(s).

The 1st Most Common Venues

Neighborhood	1st Most Common Venue	Cluster Labels	Population
Palacio	Spanish Restaurant	2.0	22515
Imperial	Tapas Restaurant	2.0	22719
Pacífico	Spanish Restaurant	2.0	33601
Recoletos	Spanish Restaurant	2.0	15756
El Viso	Spanish Restaurant	2.0	17145
Bellas Vistas	Spanish Restaurant	2.0	28750
Gaztambide	Spanish Restaurant	2.0	22666
El Pardo	NaN	NaN	3456
Casa de Campo	Pub	2.0	12900
Los Cármenes	Restaurant	2.0	17192
Comillas	Coffee Shop	2.0	22248
Orcasitas	Sporting Goods Shop	2.0	22555

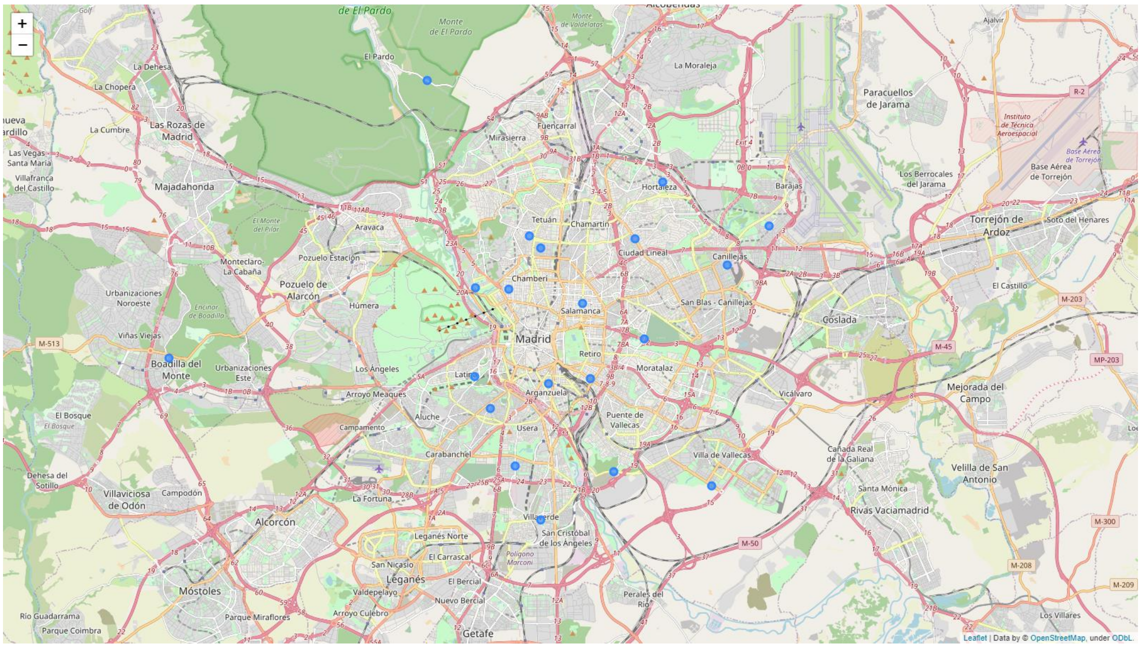
Entrevías	NaN	NaN	33986
Ventas	Chinese Restaurant	2.0	47744
Palomas	Spanish Restaurant	2.0	6738
Villaverde Alto	Mediterranean Restaurant	2.0	44299
Casco Histórico de Vallecas	Pizza Place	1.0	39740
Simancas	Restaurant	2.0	27242
Alameda de Osuna	Restaurant	2.0	19446

The top ten most common venues

	Cluster Labels	Neighborhood	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
0	0	Alameda de Osuna	Restaurant	Tapas Restaurant	Plaza	Bookstore	Hobby Shop	Fried Chicken Joint	Italian Restaurant	Metro Station	Cocktail Bar	Pizza Place
1	0	Bellas Vistas	Spanish Restaurant	Bar	Grocery Store	Bakery	Supermarket	Tapas Restaurant	Pizza Place	Coffee Shop	Seafood Restaurant	Farmers Market
2	0	Casa de Campo	Pub	Stadium	Spanish Restaurant	Pool	Gym	Grocery Store	Fast Food Restaurant	Convenience Store	Cuban Restaurant	Deli / Bodega
3	1	Casco Histórico de Vallecas	Pizza Place	Scenic Lookout	Bakery	Yoga Studio	Food	Department Store	Dessert Shop	Diner	Discount Store	Donut Shop

	Cluster Labels	Neighborhood	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
4	0	Comillas	Coffee Shop	Fast Food Restaurant	Grocery Store	Supermarket	Flea Market	Plaza	Farmers Market	Dumpling Restaurant	Falafel Restaurant	Electronics Store

1. The clusters are visualized via folium map:



5) Discussion

From the results discovered and presented, the following observations and recommendations can be made:

- Based on the criteria given by the investor group and the cluster data, the main recommendation for a new business would be a ***Chinese Restaurant*** in neighborhood ***Ventas*** due to the largest population.
- A secondary recommendation is made for the neighborhood of *Villaverde Alto* for a *Mediterranean Restaurant* with the second largest population.
- In general terms it can be seen that in all the neighborhoods the main business is related to the restaurant business.

6) Conclusion

In conclusion, the scope of this of the analysis is somewhat limited. The Venues to do business is ever changing, and the information afforded us may be dated due to relying on user information via Foursquare. Overall though, the model created can easily be replicated again and again with monitored data via the Foursquare API and the data from the forthcoming census in 2021. With the data analyzed and scoring system established by the investor group, we stand by the recommendations made.