The Battle of Neighborhoods

Gym in Madrid

The target of this project is to assess an investor to decide where would be the best location to place a Gym in Madrid. Using Foursquare data we want to cluster the neighborhoods based on their similarity in on the “Outdoors and Recreation” category so we can understand which are the sport offers each cluster offers to its population.

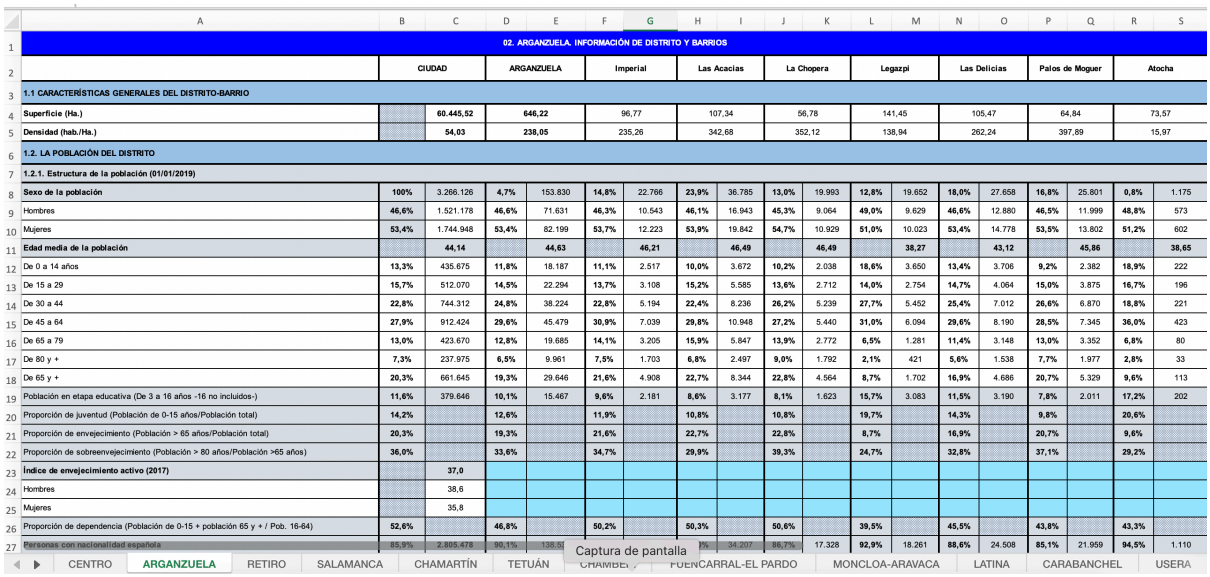
New gym in Madrid is about to open

Based on this premise, we will be able to find the neighborhoods where a Gym / Fitness Center has the better change to get more potential clients.

Data

The information he 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 we can download an excel file with several indicators about (“panel_indicadores_distritos_barrios_2019.xls”) 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	Q	R	S
1	02. ARGANZUELA. INFORMACIÓN DE DISTRITO Y BARRIOS																		
2		CIUDAD		ARGANZUELA		Imperial		Las Acacias		La Chopera		Legazpi		Las Delicias		Palos de Moguer		Atocha	
3	1.1 CARACTERÍSTICAS GENERALES DEL DISTRITO-BARRIO																		
4	Superficie (Ha.)		60.445,52		646,22		96,77		107,34		56,78		141,45		105,47		64,84		73,57
5	Densidad (hab./Ha.)		54,03		238,05		235,26		342,68		352,12		138,94		282,24		397,89		15,97
6	1.2. LA POBLACIÓN DEL DISTRITO																		
7	1.2.1. Estructura de la población (01/01/2019)																		
8	Sexo de la población	100%	3.266.126	4,7%	153.830	14,8%	22.766	23,9%	36.785	13,0%	19.993	12,8%	19.652	18,0%	27.658	16,8%	25.801	0,8%	1.175
9	Hombres	46,6%	1.521.178	46,6%	71.631	46,3%	10.543	46,1%	16.943	45,3%	9.064	49,0%	9.629	46,6%	12.880	46,5%	11.999	48,8%	573
10	Mujeres	53,4%	1.744.948	53,4%	82.199	53,7%	12.223	53,9%	19.842	54,7%	10.929	51,0%	10.023	53,4%	14.778	53,5%	13.802	51,2%	602
11	Edad media de la población		44,14		44,63		46,21		46,49		46,49		38,27		43,12		45,86		38,65
12	De 0 a 14 años	13,3%	435.675	11,8%	18.187	11,1%	2.517	10,0%	3.672	10,2%	2.038	18,6%	3.650	13,4%	3.706	9,2%	2.382	18,9%	222
13	De 15 a 29	15,7%	512.070	14,5%	22.294	13,7%	3.108	15,2%	5.585	13,6%	2.712	14,0%	2.754	14,7%	4.064	15,0%	3.875	16,7%	196
14	De 30 a 44	22,8%	744.312	24,8%	38.224	22,8%	5.194	22,4%	8.236	26,2%	5.239	27,7%	5.452	25,4%	7.012	26,6%	6.870	18,8%	221
15	De 45 a 64	27,9%	912.424	29,6%	45.479	30,9%	7.039	29,8%	10.948	27,2%	5.440	31,0%	6.094	29,6%	8.190	28,5%	7.345	36,0%	423
16	De 65 a 79	13,0%	423.670	12,8%	19.685	14,1%	3.205	15,9%	5.847	13,9%	2.772	6,5%	1.281	11,4%	3.148	13,0%	3.352	6,8%	80
17	De 80 y +	7,3%	237.975	6,5%	9.961	7,5%	1.703	6,8%	2.497	9,0%	1.792	2,1%	421	5,6%	1.538	7,7%	1.977	2,8%	33
18	De 65 y +	20,3%	661.645	19,3%	29.646	21,6%	4.908	22,7%	8.344	22,8%	4.564	8,7%	1.702	16,9%	4.686	20,7%	5.329	9,6%	113
19	Población en etapa educativa (De 3 a 16 años -16 no incluidos-)	11,6%	379.646	10,1%	15.467	9,6%	2.181	8,6%	3.177	8,1%	1.623	15,7%	3.083	11,5%	3.190	7,8%	2.011	17,2%	202
20	Proporción de juventud (Población de 0-15 años/Población total)	14,2%		12,6%		11,9%		10,8%		10,8%		19,7%		14,3%		9,8%		20,6%	
21	Proporción de envejecimiento (Población > 65 años/Población total)	20,3%		19,3%		21,6%		22,7%		22,8%		8,7%		16,9%		20,7%		9,6%	
22	Proporción de sobreenvejecimiento (Población > 80 años/Población >65 años)	36,0%		33,6%		34,7%		29,9%		39,3%		24,7%		32,8%		37,1%		29,2%	
23	Índice de envejecimiento activo (2017)		37,0																
24	Hombres		38,6																
25	Mujeres		35,8																
26	Proporción de dependencia (Población de 0-15 + población 65 y + / Pob. 16-64)	52,6%		46,8%		50,2%		50,3%		50,6%		39,5%		45,5%		43,8%		43,3%	
27	Personas con nacionalidad española	85,9%	2.805.478	90,1%	138,5				34.207	86,7%	17.328	92,9%	18.261	88,6%	24.508	85,1%	21.959	94,5%	1.110

Captura de pantalla

CENTRO | ARGANZUELA | RETIRO | SALAMANCA | CHAMARTÍN | TETUÁN | CHAMBERÍ | FUENCARRAL-EL PARDO | MONCLOA-ARAVACA | LATINA | CARABANCHEL | USERA

We will additionally use two .CSV files downloaded from the same webpage containing information about which neighborhood belongs to each borough both of them you can find them in the project repository.

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.

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 establish a clustering for neighborhood by means of KMeans algorithm.

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. I used two CSV files described in the “Data” section:

	borough code	borough	neighborhood
0	1	CENTRO	PALACIO
1	1	CENTRO	EMBAJADORES
2	1	CENTRO	CORTES
3	1	CENTRO	JUSTICIA
4	1	CENTRO	UNIVERSIDAD
...	...	...	...
126	21	BARAJAS	ALAMEDA DE OSUNA
127	21	BARAJAS	AEROPUERTO
128	21	BARAJAS	CASCO H.BARAJAS
129	21	BARAJAS	TIMON
130	21	BARAJAS	CORRALEJOS

- Next, it is necessary to clean the dataframe, define columns, drop duplicates and unnecessary information like “district surface”

- 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.

	borough code	borough	neighborhood	Latitude	Longitude
0	1	CENTRO	PALACIO	40.40958	-3.88009
1	1	CENTRO	EMBAJADORES	40.39152	-3.69289
2	1	CENTRO	CORTES	40.41607	-3.69893
3	1	CENTRO	JUSTICIA	40.42479	-3.69308
4	1	CENTRO	UNIVERSIDAD	40.42565	-3.70726
...	...	...	...	...	...

- Once we have identified all neighborhoods and its location coordinates, we can use the Foursquare API to find the venues existing in each neighborhood. We will search for venues within 1km from the neighborhood center. As we are looking for Gyms, we will filter the venues to obtain only those ones related to the “Outdoors and Recreation” category.

	neighborhood	neighborhood Latitude	neighborhood Longitude	venue	venue Latitude	venue Longitude	venue Category
0	PALACIO	40.40958	-3.88009	BeOne	40.404435	-3.885560	Gym / Fitness Center
1	PALACIO	40.40958	-3.88009	Skatepark de Boadilla	40.410489	-3.888021	Skate Park
2	PALACIO	40.40958	-3.88009	Fitness19	40.409288	-3.880009	Gym / Fitness Center
3	PALACIO	40.40958	-3.88009	Parque Sofía de Grecia II	40.409460	-3.882442	Plaza
4	PALACIO	40.40958	-3.88009	Parque Santillana del Mar	40.407508	-3.882004	Park
...	...	...	...	...	...	...	...
2930	CORRALEJOS	40.46540	-3.61164	Recinto Para Perros Parque JCI	40.462782	-3.608200	Dog Run
2931	CORRALEJOS	40.46540	-3.61164	Piscina de Novotel	40.462584	-3.614951	Pool
2932	CORRALEJOS	40.46540	-3.61164	Fitness Pullman Hotel	40.462311	-3.614936	Gym / Fitness Center
2933	CORRALEJOS	40.46540	-3.61164	Lago del Parque Juan Carlos I	40.460311	-3.609649	Lake
2934	CORRALEJOS	40.46540	-3.61164	Donut Juan Carlos I	40.459320	-3.606842	Sculpture Garden

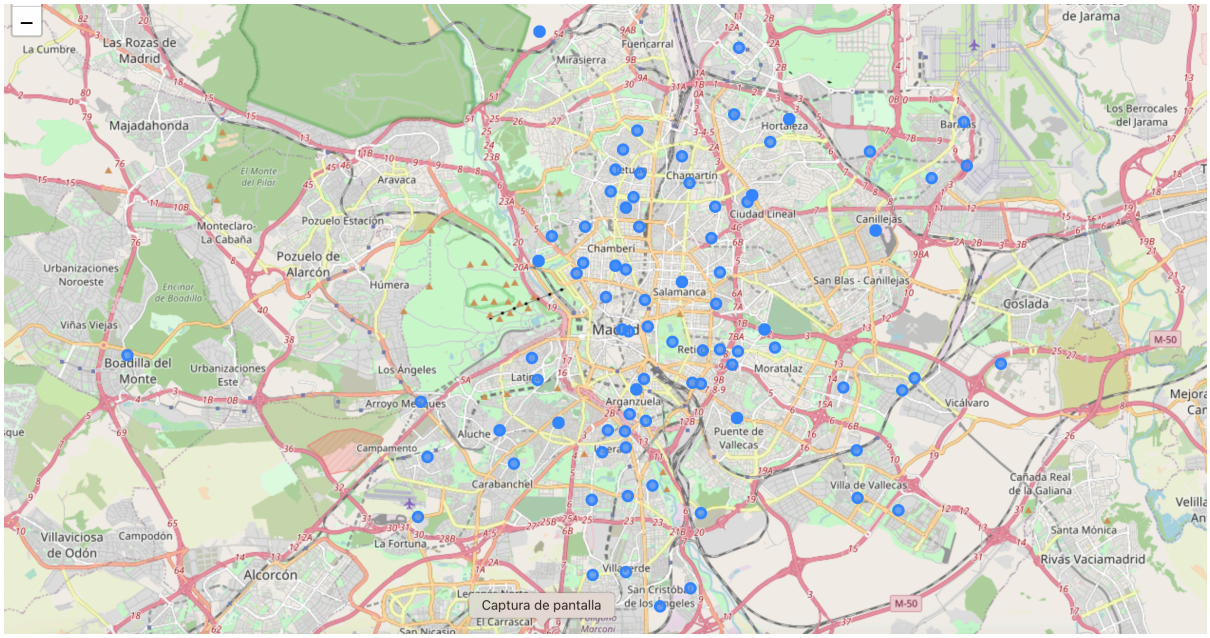
- With this new data frame we will group the venues to get the top 14 most common venues in each neighborhood based in the number of repetitions

	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue
0	ABRANTES	Plaza	Gym / Fitness Center	Athletics & Sports	Park	Playground	Roof Deck	Mountain	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gym	Gun Range
1	ACACIAS	Plaza	Park	Gym	Gym / Fitness Center	Yoga Studio	Garden	Playground	Pool	Gym Pool	Gymnastics Gym	Lake	Outdoors & Recreation	Pedestrian Plaza	Athletics & Sports
2	ADELFAS	Gym	Plaza	Gym / Fitness Center	Athletics & Sports	Park	Harbor / Marina	Martial Arts School	Skate Park	Soccer Field	Gymnastics Gym	Gym Pool	Lake	Basketball Court	Indoor Play Area
3	AEROPUERTO	Sculpture Garden	Playground	Plaza	Soccer Field	Garden	Gym	Yoga Studio	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gym / Fitness Center	Gun Range
4	ALAMEDA DE OSUNA	Park	Plaza	Gym	Gym / Fitness Center	Tennis Court	Garden	Scenic Lookout	Yoga Studio	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gun Range

- 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. We will then merge Madrid data containing neighborhoods and coordinates, with the neighborhoods venues clustered.

	borough code	borough	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	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue
0	1	CENTRO	PALACIO	40.40958	-3.88009	0.0	Gym / Fitness Center	Plaza	Athletics & Sports	Park	Pool	Skate Park	Cycle Studio	Dog Run	Lake	Basketball Court	Indoor Play Area	Harbor / Marina
1	1	CENTRO	EMBAJADORES	40.39152	-3.69289	3.0	Park	Plaza	Gym	Playground	Gym / Fitness Center	Soccer Field	Yoga Studio	Garden	Pool	Skating Rink	Skate Park	Fountain
2	1	CENTRO	CORTES	40.41607	-3.69893	0.0	Plaza	Gym	Gym / Fitness Center	Park	Athletics & Sports	Garden	Roof Deck	Fountain	Yoga Studio	Sculpture Garden	Pedestrian Plaza	Lake
3	1	CENTRO	JUSTICIA	40.42479	-3.69308	0.0	Plaza	Park	Yoga Studio	Gym	Gym / Fitness Center	Athletics & Sports	Gymnastics Gym	Fountain	Playground	Garden	Gym Pool	Outdoors & Recreation
4	1	CENTRO	UNIVERSIDAD	40.42565	-3.70726	0.0	Plaza	Gym	Gym / Fitness	Park	Athletics & Sports	Garden	Gym Pool	Gymnastics	Outdoors &	Roof Deck	Pedestrian Plaza	Playground

- Through the folium library show a map with the clusters.



- Now is time to analyze the different clusters to identify the best cluster to invest in a gym based on how frequently you can find them

Cluster 1

```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 0, madrid_merged.columns[[1] + list(range(5, madrid_merged.shape[1]))]]
```

	borough	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	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue
0	CENTRO	0.0	Gym / Fitness Center	Plaza	Athletics & Sports	Park	Pool	Skate Park	Cycle Studio	Dog Run	Lake	Basketball Court	Indoor Play Area	Harbor / Marina	Botanical Garden	Boxing Gym
2	CENTRO	0.0	Plaza	Gym	Gym / Fitness Center	Park	Athletics & Sports	Garden	Roof Deck	Fountain	Yoga Studio	Sculpture Garden	Pedestrian Plaza	Lake	Playground	Tree

Cluster 2

```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 1, madrid_merged.columns[[1] + list(range(5, madrid_merged.shape[1]))]]
```

	borough	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	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue
16	RETIRO	1.0	Park	Plaza	Garden	Fountain	Gym	Gym / Fitness Center	Lake	Yoga Studio	Harbor / Marina	Athletics & Sports	Playground	Sports Club	Scenic Lookout	Outdoors & Recreation
60	LATINA	1.0	Park	Athletics & Sports	Plaza	Gym	Playground	Soccer Field	Gym / Fitness Center	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gun Range	Mountain

Cluster 3

```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 2, madrid_merged.columns[[1] + list(range(5, madrid_merged.shape[1]))]]
```

	borough	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	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue
43	FUENCARRAL-EL PARDO	2.0	Scenic Lookout	Yoga Studio	Outdoors & Recreation	Martial Arts School	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gym / Fitness Center	Gym	Gun Range	Golf Course	Garden
45	FUENCARRAL-EL PARDO	2.0	Scenic Lookout	Yoga Studio	Outdoors & Recreation	Martial Arts School	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gym / Fitness Center	Gym	Gun Range	Golf Course	Garden

Cluster 4

```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 3, madrid_merged.columns[[1] + list(range(5, madrid_merged.shape[1]))]]
```

	borough	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	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue
1	CENTRO	3.0	Park	Plaza	Gym	Playground	Gym / Fitness Center	Soccer Field	Yoga Studio	Garden	Pool	Skating Rink	Skate Park	Fountain	Lake	Golf Course
13	RETIRO	3.0	Gym	Plaza	Park	Athletics & Sports	Martial Arts School	Harbor / Marina	Gym / Fitness Center	Playground	Gymnastics Gym	Skate Park	Soccer Field	Dog Run	Garden	Golf Course

Cluster 5

```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 4, madrid_merged.columns[[1] + list(range(5, madrid_merged.shape[1]))]]
```

	borough	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	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue
44	FUENCARRAL-EL PARDO	4.0	Park	Mountain	Dog Run	Gym	Martial Arts School	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gym / Fitness Center	Yoga Studio	Outdoors & Recreation	Golf Course
46	FUENCARRAL-EL PARDO	4.0	Park	Mountain	Dog Run	Gym	Martial Arts School	Lake	Indoor Play Area	Harbor / Marina	Gymnastics Gym	Gym Pool	Gym / Fitness Center	Yoga Studio	Outdoors & Recreation	Golf Course

Results

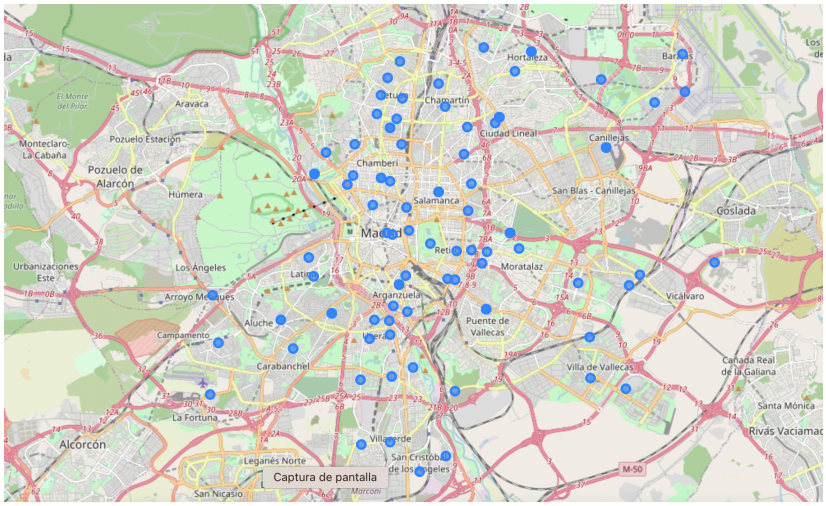
With the data now ready, we run k-means to cluster the neighborhoods into five clusters. The cluster number was established after multiple samplings and iterations. With our clusters established, this dataframe is merged with the total scores data to provide us with our final pieces of criteria in selecting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

Based on the analysis, the cluster where gym is lower in the “top 14 most common venues” list is the 2nd cluster, where we can observe the average position is about 10th position.

From the cluster 2 neighborhoods, the selection must be done based on the population and the lowest possible average age to maximize the number of potential clients

	borough	Population	Average age
16	RETIRO	119379	47,13
60	LATINA	238154	46,55
74	USERA	139501	42,34
79	PUENTE DE VALLECAS	234770	43,3
85	MORATALAZ	94609	47,61
91	CIUDAD LINEAL	216270	45,87
107	VILLAVERDE	148883	42,15
113	VILLA DE VALLECAS	110436	38,9
114	VICALVARO	72126	40,83

On the map:



Discussion

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

- Based on the criteria given by the client and the cluster data, the main recommendation for a new gym would be Villa de Vallecas, followed by Puente de Vallecas (as both has a good potential and they are nearby).
- Villaverde and Usera are also good locations based on the population/age ratio
- Latina, even if it is the biggest in population, the average age of its population is quite high, reducing the amount of potential clients as the younger ones are more interested in high intensity activities, so my recommendation would be not to choose it at first instance.

Conclusion

In conclusion, the scope of this of the analysis is somewhat limited. The venues to do business is ever changing, and the information available may be outdated as it relays 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.