The Battle of the Neighborhoods - Madrid

Applied Data Science Capstone - December 2020

Problem Description

- The target of the project is to assess an investor to decide where would be the best location to place a Gym.

- The objective is to cluster the neighborhoods based on the ratio gym/ population to support the decision.

Data

- The information needed about districts, neighborhoods and population can be found in the website: “https://datos.madrid.es/portal/site/egob"

- An Excel sheet can be downloaded named “Panel de Indicadores”. It contains population density. Two CSVs can be also downloaded from the same web with District and Neighborhood codes

- The neighborhood coordinates will be extracted from the geocoder library

- Finally data about venues will be extracted from Foursquare API

Methodology

- The first step is to obtain a dataframe with codes and descriptions about districts and neighborhoods of Madrid city

- Once we have identified all neighborhoods and its location coordinates, we can use the Foursquare API to find the venues existing in each neighborhood

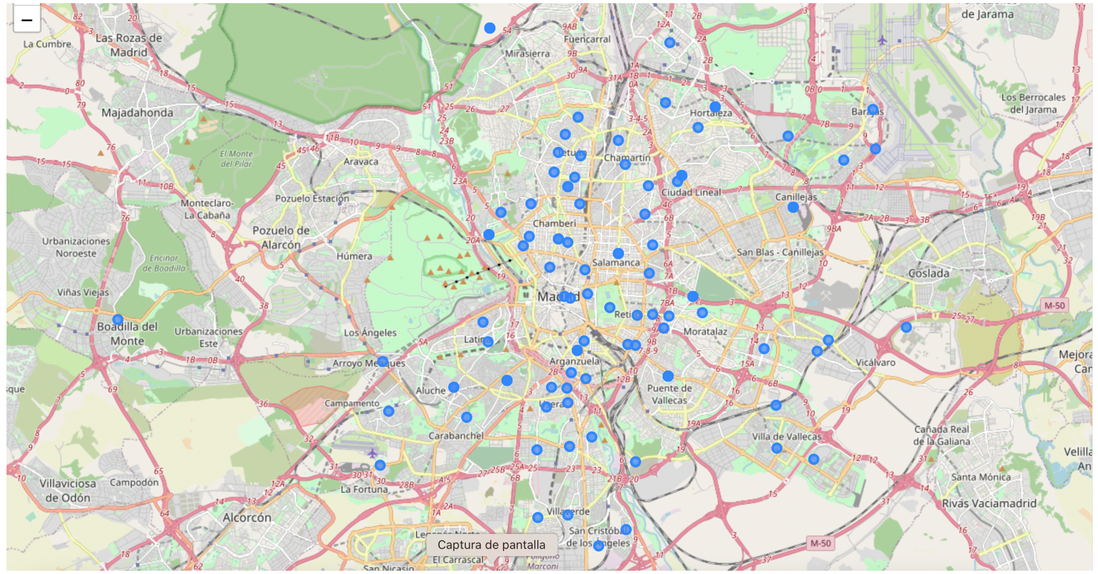
- With this new data frame we will group the venues to get the top 14 most common venues in each neighborhood based in the number of repetitions

	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

Methodology

- 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.

- Through the folium library show a map with the clusters to make their identification easier and more visual



Results

- 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

Discussion

- 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.