THE BATTLE OF NEIGHBORHOODS

This report contains the study performed for the Capstone and is structured as follows: Section 1 presents the business problem. Section 2 describes the data that is going to be analyzed. In Section 3 the methodology used is presented. Section 4 shows the results obtained, while Section 5 presents the discussion. Finally, Section 6 contains the conclusions of the project.

1. INTRODUCTION / BUSINESS PROBLEM

An important gym group is willing to open a new gym in Madrid and they want to find the best place to open the club. Currently more and more people finds specialized places to practice sports and some neighborhoods can be crowded, so opening a new gym without analyzing the current venues in the city and the potential customers can be risky.

Additionally, our customer has transmitted to us that they're age target group is people aged between 20 and 39 years. Moreover, we will rely on income data per district to refine the final decision.

Consequently, the company came to us to provide them with useful information about Madrid districts and neighborhoods so they can base the decision of the new gym location in reliable data.

2. DATA

The data used to solve this problem came from four sources:

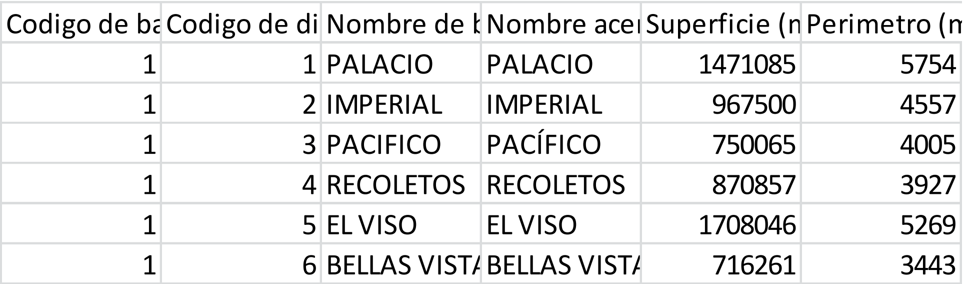
1. **District and neighborhood names from Madrid**

   Source: https://www.madrid.es/

   The columns extracted for this study are :

   - Hood number : Unique number of the hood
   - Hood: Name of the hood
   - Neighborhood: Name of the neighborhood

Here is an extract of the first rows of this dataset:



Codigo de ba	Codigo de di	Nombre de b	Nombre acei	Superficie (n	Perimetro (n
1	1	PALACIO	PALACIO	1471085	5754
1	2	IMPERIAL	IMPERIAL	967500	4557
1	3	PACIFICO	PACÍFICO	750065	4005
1	4	RECOLETOS	RECOLETOS	870857	3927
1	5	EL VISO	EL VISO	1708046	5269
1	6	BELLAS VIST/	BELLAS VIST/	716261	3443

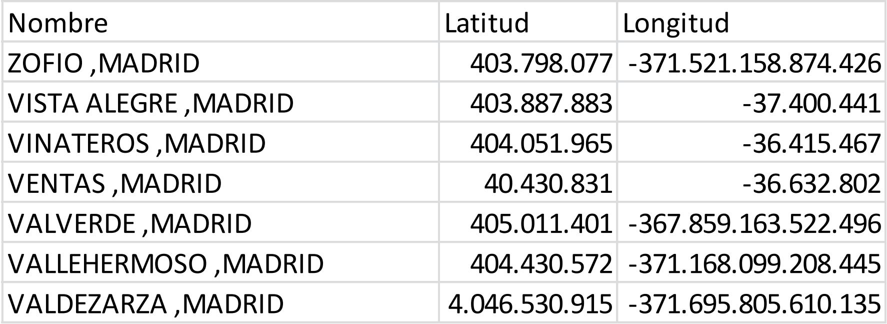
2. **Geographical data**

   Source: https://www.123coordenadas.com

This web was used to calculate coordinates for each neighborhood and obtain a CSV file with the following columns:

- Longitude : Geographical longitude
- Latitude : Geographical latitude

Here is an extract of the first rows of this dataset:



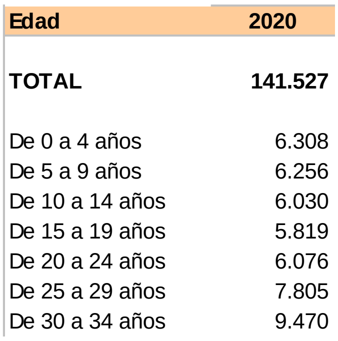
Nombre	Latitud	Longitud
ZOFIO ,MADRID	403.798.077	-371.521.158.874.426
VISTA ALEGRE ,MADRID	403.887.883	-37.400.441
VINATEROS ,MADRID	404.051.965	-36.415.467
VENTAS ,MADRID	40.430.831	-36.632.802
VALVERDE ,MADRID	405.011.401	-367.859.163.522.496
VALLEHERMOSO ,MADRID	404.430.572	-371.168.099.208.445
VALDEZARZA ,MADRID	4.046.530.915	-371.695.805.610.135

3. **Population data**

Source : https://www.madrid.es/

An Excel file with the population of Madrid classified in districts and age ranges was obtained and used to determine the number of persons with ages between 20 and 39 years per district. This data is stored in a column called Population.

Here is an extract of the first rows of this dataset:



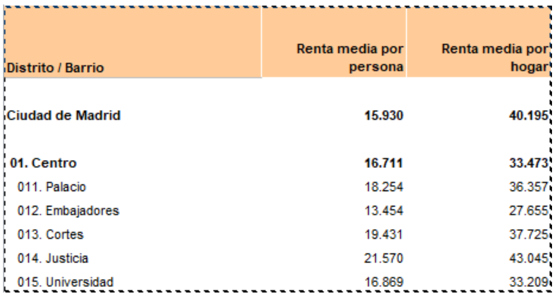
Edad	2020
TOTAL	**141.527**
De 0 a 4 años	6.308
De 5 a 9 años	6.256
De 10 a 14 años	6.030
De 15 a 19 años	5.819
De 20 a 24 años	6.076
De 25 a 29 años	7.805
De 30 a 34 años	9.470

4. **Income data**

Source : https://www.madrid.es/

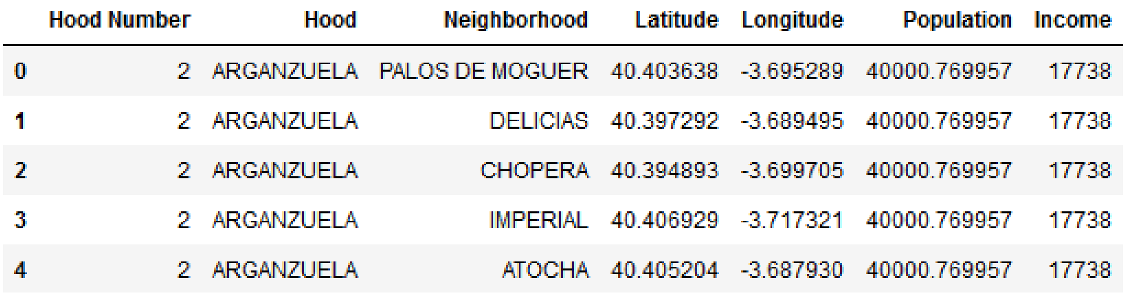
An Excel file with the income per person of Madrid classified in districts was obtained and used to determine the average income for each district. This data is stored in a column called Income.

Here is an extract of the first rows of this dataset:



Distrito / Barrio	Renta media por persona	Renta media por hogar
Ciudad de Madrid	15.930	40.195
01. Centro	16.711	33.473
011. Palacio	18.254	36.357
012. Embajadores	13.454	27.655
013. Cortes	19.431	37.725
014. Justicia	21.570	43.045
015. Universidad	16.869	33.209

Using these four sources a dataframe is created. The first 5 rows of this dataframe is depicted in Figure 1.



	Hood Number	Hood	Neighborhood	Latitude	Longitude	Population	Income
0	2	ARGANZUELA	PALOS DE MOGUER	40.403638	-3.695289	40000.769957	17738
1	2	ARGANZUELA	DELICIAS	40.397292	-3.689495	40000.769957	17738
2	2	ARGANZUELA	CHOPERA	40.394893	-3.699705	40000.769957	17738
3	2	ARGANZUELA	IMPERIAL	40.406929	-3.717321	40000.769957	17738
4	2	ARGANZUELA	ATOCHA	40.405204	-3.687930	40000.769957	17738

Figure 1. Dataframe head

3. METHODOLOGY

The methodology followed to perform the study is described in this section.

1. **Obtain the data:** The data was obtained from the sources described in Section 2.

2. **Join data in a single data frame:** Since the data was obtained from different sources, a join process was needed. Luckily, Madrid districts and neighborhoods are identified by an unique number, so this process was robust and avoided possible mismatches due to different ways of writing the district and neighborhoods names.
   The location data was added later, after obtaining a CSV file with longitudes and latitudes for each neighborhood.

3. **Get venues in Madrid neighborhoods:** Using Foursquare API, a list of all the nearest venues to each neighborhood was obtained and grouped for analysis.

4. **Obtain gyms:** The list obtained in Section 3 was filtered to obtain only venues related to gyms and related places.

5. **Analyze type of gyms:** Since several venues contains the word "Gym", an analysis was made to determine which are the different categories of gyms and how many of them were in each neighborhood.

6. **Group gyms by neighborhood:** After analyzing gym types, it was determined that all of them were similar for the purpose of the study, so they are summed in a single column called "Gyms" and grouped by neighborhood.

7. **Prepare dataframe for classification:** The original dataframe is merged with the previously obtained one, and the NaN values are converted to 0. This NaN values appeared after the merge because some neighborhoods doesn't have gyms.

8. **Normalize data:** The data is normalized to give zero mean and unit variance, for KNN algorithm to work properly.

9. **Cluster neighborhoods:** KNN alhorithm with K = 4 was used to classify neighborhoods.

10. **Visualize resulting clusters:** The resulting clusters were depicted in a map of Madrid.

11. **Analyze clusters:** The clusters were analyzed using statistics indicators to determine their characteristics and find the best solution for our problem.

4. RESULTS

After the execution of the classification algorithm, four clusters were obtained. The geographical representation of these clusters can be seen in Figure 2.

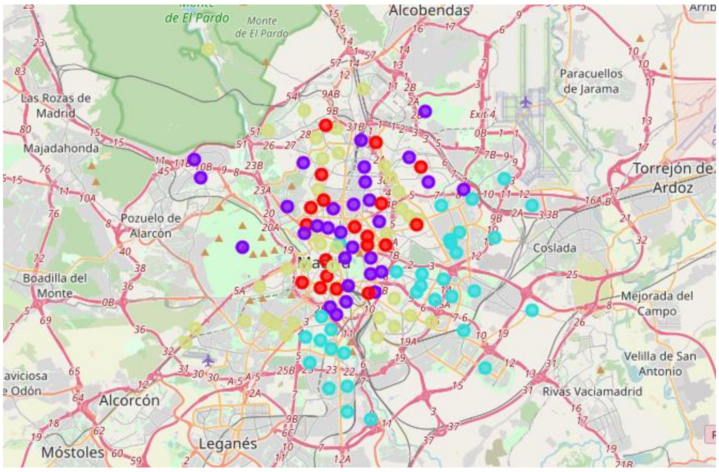


Figure 2. Madrid map with clusters

The clusters were analyzed in terms of statistical indicators to determine the features of the neighborhoods classified in each of them. Table 1 summarizes the main indicators:

- **Gym count**: Number of gym clubs in the cluster.
- **Income mean**: Mean of the income per person in the cluster.
- **Population mean**: Mean number of people living in the cluster.
- **Population/gyms rate**: Result of dividing the number of people between the available gyms per cluster.

The last column contains the mean of each indicator for all the clusters.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Mean
Gym count	19	30	32	38	29,75
Income mean	19.834,00 €	22.082,00 €	12.353,00 €	14.073,00 €	17.085,5 €
Population mean	40.389	33.905	27.967	52.296	38.639,25
Population/gyms rate	2.125,73	1.130,16	873,96	1.376,21	1.376,52

Table 1. Cluster indicators summary

5. DISCUSSION

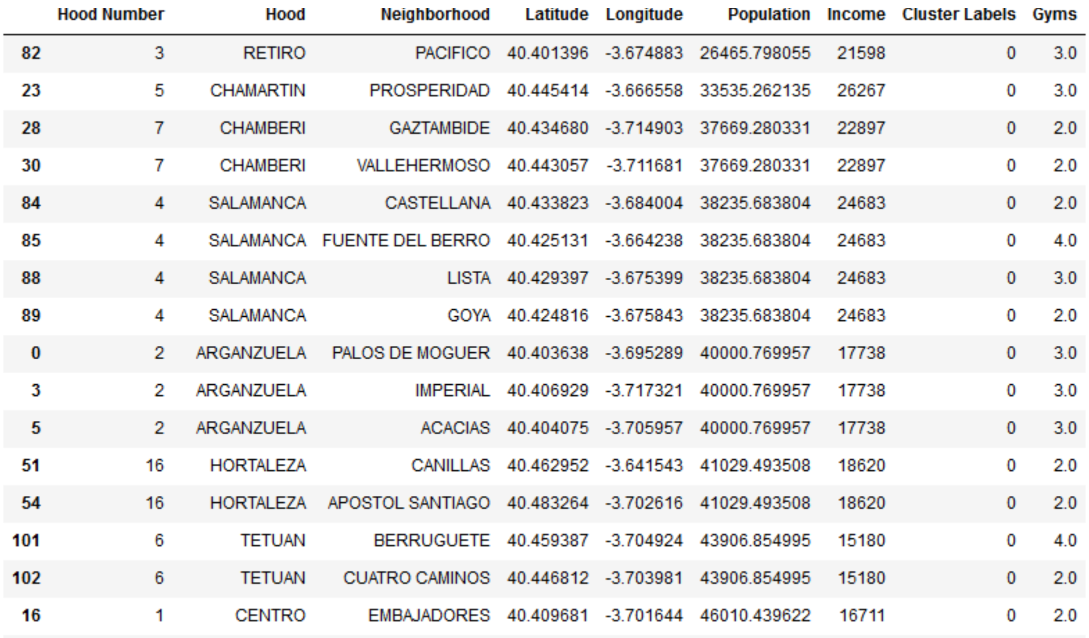
Having a look at the results, it seems that Cluster 1 is the most promising result taking into account business requirements for the following reasons:

1. The Population/gyms rate is the highest of all clusters, which means that there are few gyms for the volume of people in those neighborhoods.

2. The average income is above Madrid mean, which means that there could be more potential customers with good salaries in these areas.

3. The population mean is above Madrid mean, which means that these neighborhoods are well populated.

Once we decided which cluster is the most suitable, we will have a close look to it to determine which neighborhoods are the best options to open the new gym club. Figure 3 presents the detailed data of cluster 1.



	Hood Number	Hood	Neighborhood	Latitude	Longitude	Population	Income	Cluster Labels	Gyms
82	3	RETIRO	PACIFICO	40.401396	-3.674883	26465.798055	21598	0	3.0
23	5	CHAMARTIN	PROSPERIDAD	40.445414	-3.666558	33535.262135	26267	0	3.0
28	7	CHAMBERI	GAZTAMBIDE	40.434680	-3.714903	37669.280331	22897	0	2.0
30	7	CHAMBERI	VALLEHERMOSO	40.443057	-3.711681	37669.280331	22897	0	2.0
84	4	SALAMANCA	CASTELLANA	40.433823	-3.684004	38235.683804	24683	0	2.0
85	4	SALAMANCA	FUENTE DEL BERRO	40.425131	-3.664238	38235.683804	24683	0	4.0
88	4	SALAMANCA	LISTA	40.429397	-3.675399	38235.683804	24683	0	3.0
89	4	SALAMANCA	GOYA	40.424816	-3.675843	38235.683804	24683	0	2.0
0	2	ARGANZUELA	PALOS DE MOGUER	40.403638	-3.695289	40000.769957	17738	0	3.0
3	2	ARGANZUELA	IMPERIAL	40.406929	-3.717321	40000.769957	17738	0	3.0
5	2	ARGANZUELA	ACACIAS	40.404075	-3.705957	40000.769957	17738	0	3.0
51	16	HORTALEZA	CANILLAS	40.462952	-3.641543	41029.493508	18620	0	2.0
54	16	HORTALEZA	APOSTOL SANTIAGO	40.483264	-3.702616	41029.493508	18620	0	2.0
101	6	TETUAN	BERRUGUETE	40.459387	-3.704924	43906.854995	15180	0	4.0
102	6	TETUAN	CUATRO CAMINOS	40.446812	-3.703981	43906.854995	15180	0	2.0
16	1	CENTRO	EMBAJADORES	40.409681	-3.701644	46010.439622	16711	0	2.0

Figure 3. Cluster 1 data

From the results it can be extracted the best locations that are described in Table 2. They all share a big population, high income and the lowest number of gyms.

Hood Number	Hood	Neighborhood	Population	Income	Gyms
4	SALAMANCA	CASTELLANA	38.235,68	24683	2.0
4	SALAMANCA	GOYA	38.235,68	24683	2.0
7	CHAMBERI	GAZTAMBIDE	37.669,28	22897	2.0
3	RETIRO	ADELFAS	26.465,80	21598	2.0
16	HORTALEZA	CANILLAS	41.029,49	18620	2.0
16	HORTALEZA	APOSTOL SANTIAGO	41.029,49	18620	2.0

Table 2. Best locations

6. CONCLUSION

In this project, a business problem related to a gym club willing to opening new venues in Madrid was introduced. Data was obtained from several sources and prepared to represent Madrid districts, neighborhoods, income per habitant, population and geo data.

Then a methodology was used to process the data and extract valuable knowledge regarding gym clubs distribution in Madrid. A classification algorithm was used to segment neighborhoods attending to the features selected in the data stage.

After the results were obtained, they were analyzed regarding the business problem questions, and a final proposal for the customer was elaborated.