The Battle of Neighborhoods

Finding the best location to open a sports shop in Madrid, Spain

IBM COURSERA DATA SCIENCE CAPSTONE

Authored by: Mario Gasco Durán

Index

1. Introduction - Business Problem

1.1. Problem Background

With more than six million of population, Madrid is city with a friendly and lively atmosphere. The capital of Spain stands out for its leisure and sightseeing offers and also for offering business and personal development opportunities. For that reason, Madrid is a city in constant growth.

Nevertheless, as a well-developed city, Madrid has one of the highest cost of business of the country, so that any investment might be very thoughtful and studied.

1.2. Problem Description

MADSport, a recently created company, wants to set up its first business in Madrid, and are planning to opening a large surface dedicated to the sale of sport equipment of all kinds. As a startup company they need the choose very carefully the starting location in order to attract possible customers and obtain benefits as much as possible for recover the investment done.

1.3. Location requirements

With the purpose of having an early succes, the company is looking for an area to establish its shop close to a large amount of sports facilities and if it's possible with a majority of young and adult population. Other characteristics of the desired area is the proximity to shopping places and popular places.

2. Data description

In order to carry out the proposed study, data has been collected from different databases. Hereunder, the features of the data and its source of origin will be described.

2.1. Madrid districts

This dataset called “Divisiones administrativas: distritos, barrios y divisiones históricas” is provided by the Madrid city government. It contains a list of the city’s districts and neighborhoods .

This dataset is available in the following URL: https://es.wikipedia.org/wiki/Anexo:Distritos_de_Madrid

Additionally, a dataset called “Divisiones administrativas: distritos, barrios y divisiones históricas” contains the geographical coordinates of the diferents districts.

This dataset is available in the following URL: https://datos.madrid.es/egob/catalogo/200078-10-distritos-barrios.zip

2.2. Madrid census populations

This dataset called “Características de la población” is provided by the Madrid city government. It contains a list of the city’s population by district, neighborhoods and age ranges. Through it you can determine the districts that concentrate the most potential buyers. In that project the selected age range was: 16-64 years. People in this range are considered potential buyers.

This dataset is available in the following URL: http://www-2.munimadrid.es/CSE6/control/seleccionDatos?numSerie=3010102262

2.3. Madrid sports facilities

This dataset called “Instalaciones básicas deportivas municipales” is also provided by the Madrid city government and it contains a list of all the city sports facilities; both sports centres and outdoor basic facilities, and its geographical coordinates.

This dataset is available in the following URL: https://datos.madrid.es/egob/catalogo/200215-0-instalaciones-deportivas.json

2.4. Foursquare API data

The Foursquare API will be used for providing a list of venues within a specific location, based on the latitude and longitude coordinates and a radius. Acquiring the location of different places, it will be possible to associate each place its nearest neighborhood.
In the following link the Foiursquare API from Madrid is available: https://es.foursquare.com/v/madrid/4d683074d4c288bf50da7065

3. Methodology

3.1. Data obtention and cleaning

The first step of this project was obtaining and preparing the data later analysis. The data comes from different locations specified in the previous section and presents differents formats: csv, json, txt and xls.

Madrid districts

This dataframe was created by the union of two different dataset.

The first dataset was obtained directly from the web, it contains information such us the population, the surface, the neighborhoods of each district from Madrid. It presents the following aspect:

	Number	District	Surface (ha.)	Population	Density (Hab./ha.)	Image	Neighborhoods
0	1	Centro	&&&&&&&&&&&0522.820000522,82	&&&&&&&&&0131928.&&&&&0131 928	&&&&&&&&&&&0252.340000252,34		Palacio (11)Embajadores (12)Cortes (13)Justici...
1	2	Arganzuela	&&&&&&&&&&&0646.220000646,22	&&&&&&&&&0151965.&&&&&0151 965	&&&&&&&&&&&0235.160000235,16		Imperial (21)Acacias (22)Chopera (23)Legazpi (...
2	3	Retiro	&&&&&&&&&&&0546.620000546,62	&&&&&&&&&0118516.&&&&&0118 516	&&&&&&&&&&&0216.820000216,82		Pacífico (31)Adelfas (32)Estrella (33)Ibiza (3...
3	4	Salamanca	&&&&&&&&&&&0539.240000539,24	&&&&&&&&&0143800.&&&&&0143 800	&&&&&&&&&&&0266.670000266,67		Recoletos (41)Goya (42)Fuente del Berro (43)Gu...
4	5	Chamartín	&&&&&&&&&&&0917.550000917,55	&&&&&&&&&0143424.&&&&&0143 424	&&&&&&&&&&&0156.310000156,31		El Viso (51)Prosperidad (52)Ciudad Jardín (53)...

Figure 1: Districts dataset

Then, due to having other data sources that provided a more valious information about the population, it was decided to drop some columns.

	Number	District
0	1	Centro
1	2	Arganzuela
2	3	Retiro
3	4	Salamanca
4	5	Chamartín

Figure 2: Districts dataset

Having the dataset containing the districts and its associate number, the geographical coordinates were concatenated. The resulting dataframe was called Madrid Districts.

	Number	District	Latitude	Longitude
0	1	Centro	40.418308	-3.70275
1	2	Arganzuela	40.400021	-3.69618
2	3	Retiro	40.413170	-3.68307
3	4	Salamanca	40.429722	-3.67975
4	5	Chamartín	40.451000	-3.67500

Figure 3: Madrid Districts dataframe

Madrid census population

This data source contains the population of each district divided by different range of ages. For this project, people in the 16-64 years range are considered potential buyers. After making this decision, the data was concatenated to de Madrid Districts dataframe.

	Number	District	Latitude	Longitude	Population16 to 64 ages
0	1	Centro	40.418308	-3.70275	102.065
1	2	Arganzuela	40.400021	-3.69618	104.784
2	3	Retiro	40.413170	-3.68307	73.652
3	4	Salamanca	40.429722	-3.67975	94.649
4	5	Chamartín	40.451000	-3.67500	91.757

Figure 4: Madrid Districts dataframe

Madrid sports facilities

Surfing on the internet a data source containing all the municipal sports facilities was found. This dataset contains information about the type of facilities and its geographical coordinates. After removing rows with missing values, the Sports dataset was created.

	Name	Facilities	Latitude	Longitude
0	Instalación Deportiva Básica Jardines de José ...	Circuito de bicicletas	40.433861	-3.710817
1	Instalación Deportiva Básica Jardines del Teni...	Pista polideportivaPista de hockeyÁrea multide...	40.439356	-3.704372
2	Instalación Deportiva Básica Parque de Enrique...	Pista de baloncesto	40.439356	-3.704372
3	Instalación Deportiva Básica Sala Municipal de...	Taller de reparación y mantenimiento16 pistas ...	40.440631	-3.709030
4	Instalación Deportiva Municipal Básica Abrante...	1 Campo de fútbol	40.374804	-3.734643

Figure 5: Sports dataset

Foursquare API

Using the Foursquare API and a custom function, a dataset containing a maximum of 500 venues 2000 meters around the center of each district was created. The dataset called Madrid Venues contained information about the geographical coordinates and the venue category.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Centro	40.418308	-3.70275	Puerta del Sol	40.417027	-3.703443	Plaza
1	Centro	40.418308	-3.70275	La Pulpería de Victoria	40.416506	-3.701709	Seafood Restaurant
2	Centro	40.418308	-3.70275	LUSH	40.419012	-3.704898	Cosmetics Shop
3	Centro	40.418308	-3.70275	Club del Gourmet Corte Ingles	40.417497	-3.704686	Gourmet Shop
4	Centro	40.418308	-3.70275	Rosi La Loca	40.415821	-3.702955	Tapas Restaurant

Figure 6:Venues dataset

Then, a little dataset containing the number of venues per district was created.

	District	N_venues
0	Arganzuela	100
1	Barajas	79
2	Carabanchel	100

Figure 7: N_venues dataset

3.2. Exploratory data analysis

The city of Madrid has 21 different districts. Among the more than 6 million of inhabitants of the city, the population between 16 and 64 years were determined as potential buyers. The next step was to observe which districts had the highest concentration of that group. The districts which a highest 16-64 years population were Carabanchel and Puente de Vallecas, and the ones with least population were Moratalaz and Barajas.

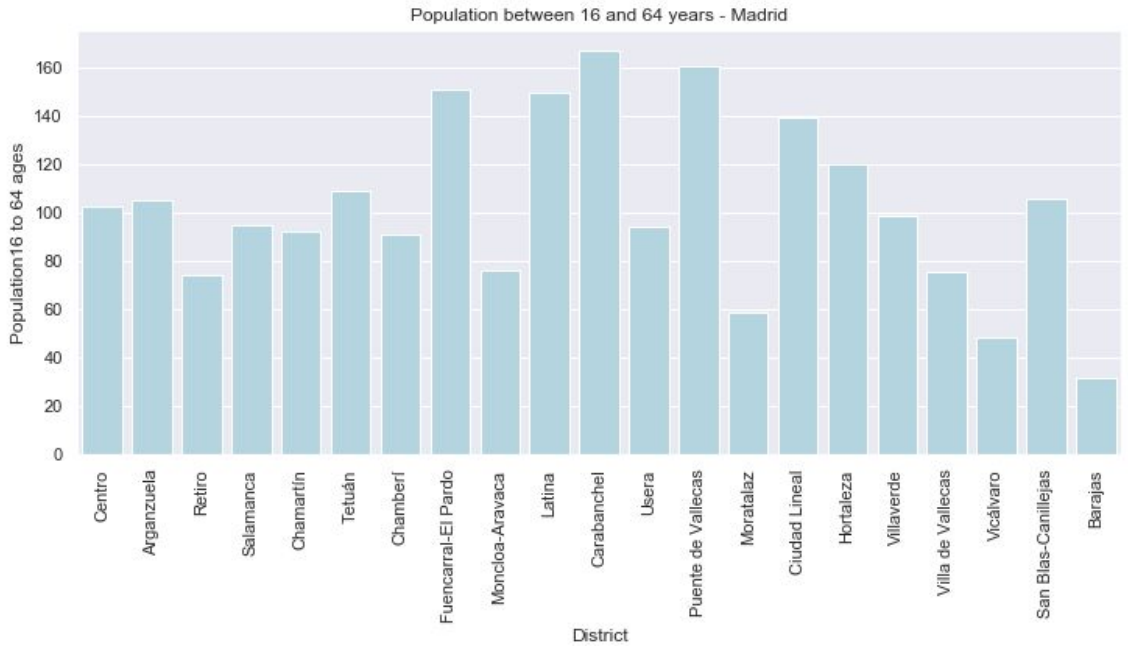

Figure 8: Population between 16 and 64 years

In order to observe the location of this districts and be able to observe if there is any connection between them, a map of Madrid containing all the districts was created.

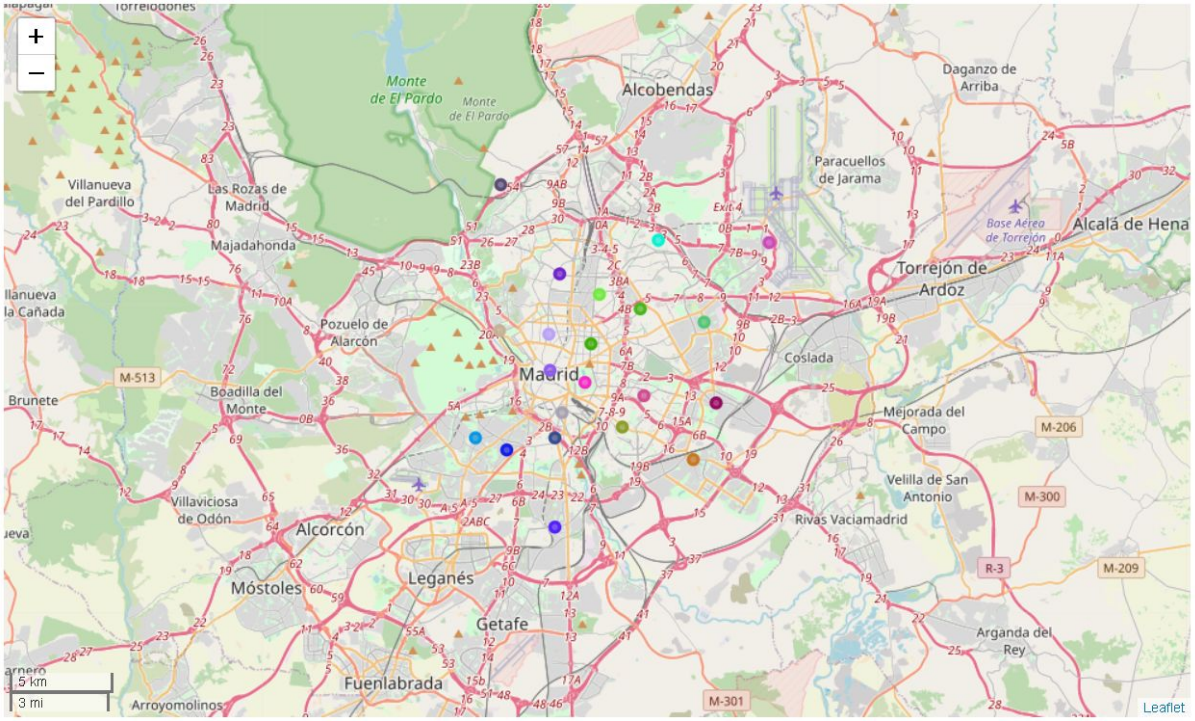

Figure 9: Madrid Map districts

Additionally, a map with all the sports facilities of the city was represented.

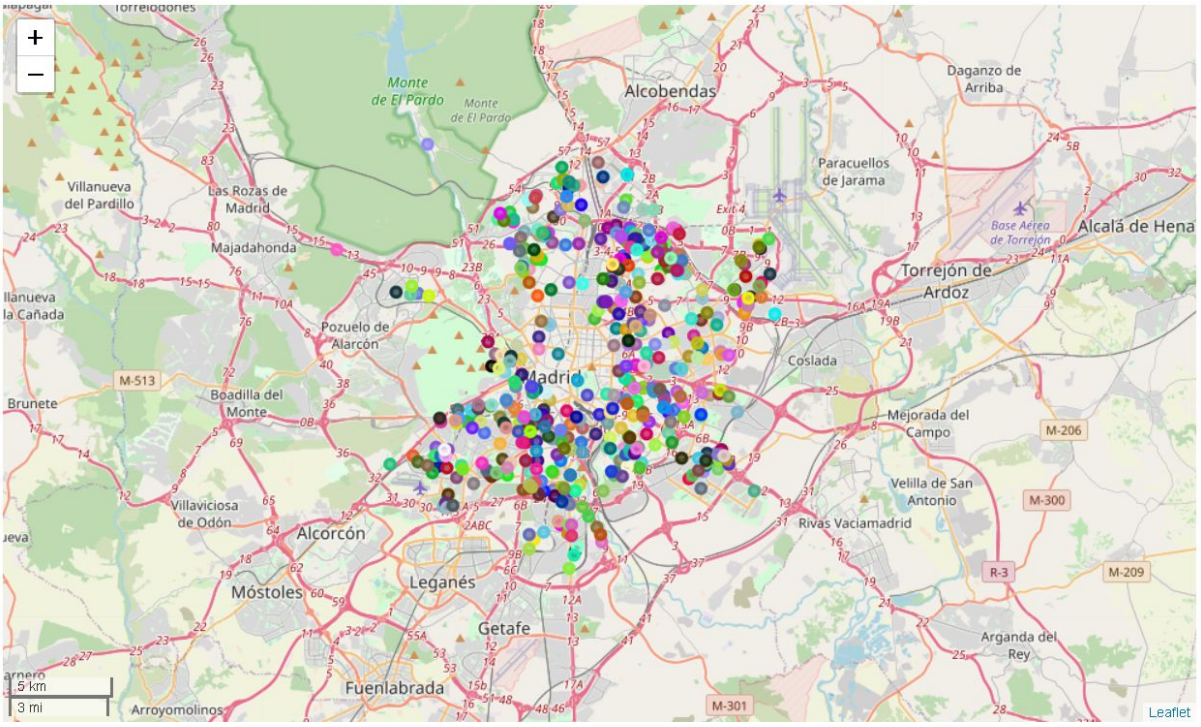

Figure 10: Madrid Map sports facilities

Lastly, using the Madrid Venues dataframe, it was determined that 15 districts of Madrid contained more than 100 venues. Instead, the districts of Villaverde, Hortaleza and Fuencarral-El Pardo do not exceed 50 venues.

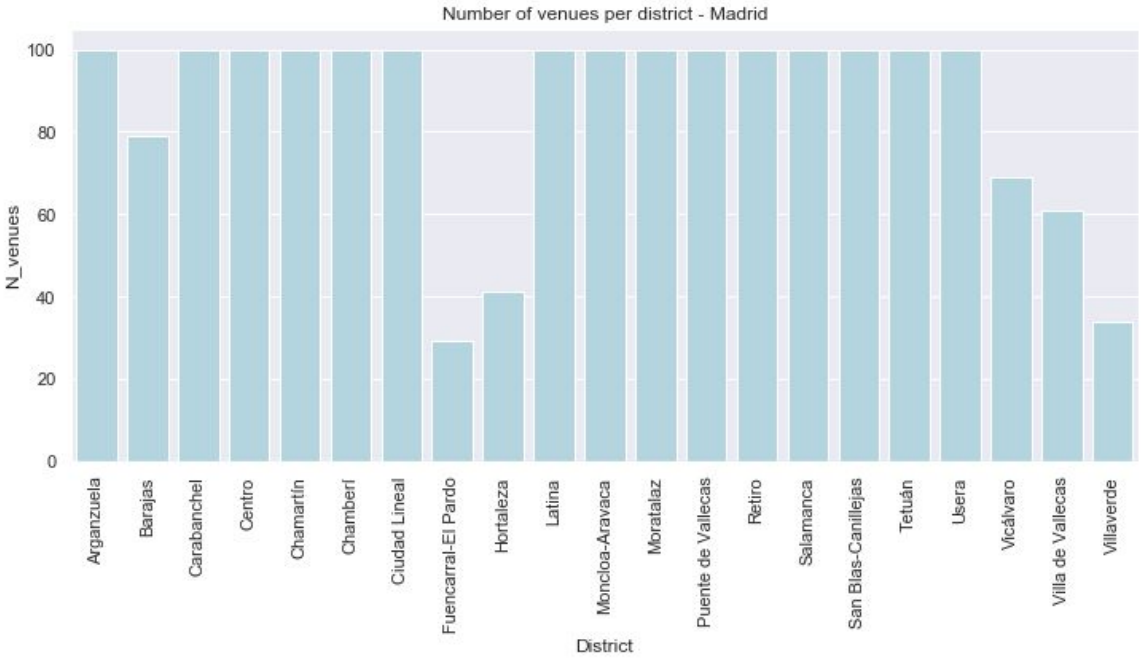

Figure 11: Number of venues per district

3.3. Clustering

Having done the exploratory analysis of the data, the next step was to find out which district was closest to each sport facility.
First of all, a total of 490 sports facilities were represented in the city of Madrid.

The selected technique for forming the clusters has been KMeans. For this, a the Sports Clustering containing the name and the geographical coordinates of each facility has been created. The starting points selected were the coordinates of the different districts of the city, 21 in total.

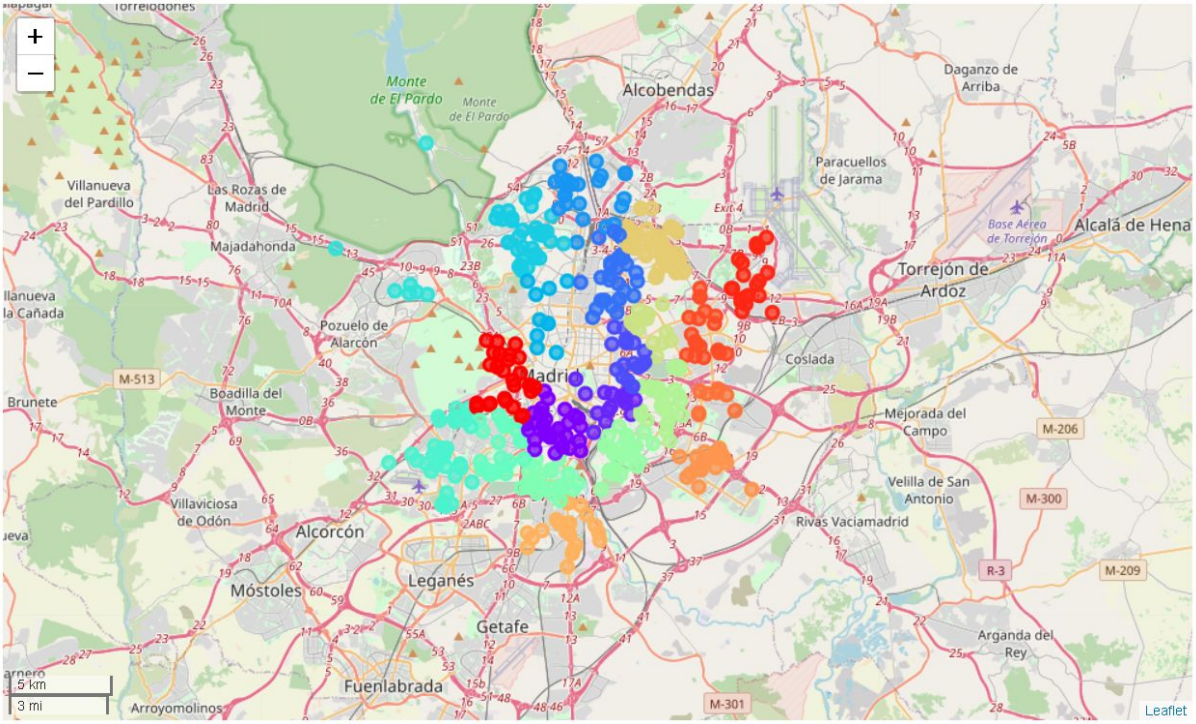

Figure 12:Sports facilities cluster

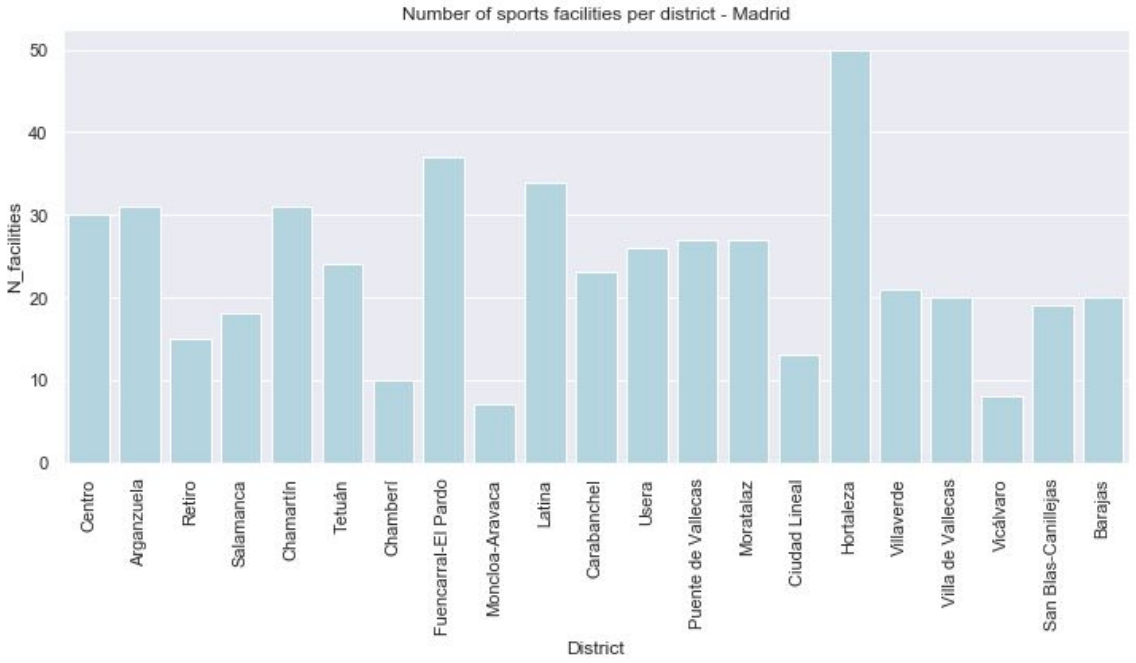

Figure 13: Number of sports facilities per district

The district of Hortaleza is the one with more municipal sport facilities, followed by Fuencarral el Pardo. On the other side, Moncloa-Aravaca and Vicálvaro are the less equipped parts of the city.

3.4. Candidats districts: Ranking

Having explored the features of the 21 districts of the city of Madrid, a score was given to each one in orden to establish a ranking and determinate the best 3 candidates.

The established criteria used was given a score between 0 and 1 to each district. To generate these scores, a weight has been assigned to the different features studied.They are specified below:

-Population between 16 and 64 years. Weight:25%

-Number of sports facilities per district. Weight:50%

-Number of venues per district. Weight:25%

First of all, the Madrid Districts dataset was completed by adding the columns N_facilites and N_venus.

	Number	District	Latitude	Longitude	Population16 to 64 ages	N_facilities	N_venues
0	17	Villaverde	40.349998	-3.70000	98.273	21	34
1	18	Villa de Vallecas	40.379600	-3.62135	75.338	20	61
2	19	Vicálvaro	40.404200	-3.60806	48.226	8	69
3	12	Usera	40.388660	-3.70035	93.873	26	100
4	6	Tetuán	40.459751	-3.69750	108.901	24	100

Figure 14: Madrid Districts dataset

The next step was normalizing the columns containing the population, the number of facilities and venues. The technique used was been: $Val_{nomalized} = Val/Val_{max}$.

39]:

	Number	District	Latitude	Longitude	Population16 to 64 ages	N_facilities	N_venues	N_facilities_Normalized	Population_Normalized	N_venues_Normalized
0	17	Villaverde	40.349998	-3.70000	98.273	21	34	0.42	0.588510	0.34
1	18	Villa de Vallecas	40.379600	-3.62135	75.338	20	61	0.40	0.451164	0.61
2	19	Vicálvaro	40.404200	-3.60806	48.226	8	69	0.16	0.288803	0.69
3	12	Usera	40.388660	-3.70035	93.873	26	100	0.52	0.562161	1.00
4	6	Tetuán	40.459751	-3.69750	108.901	24	100	0.48	0.652156	1.00

Figure 15: Madrid Districts dataset

The ranking dataset is obtained by making the sum of the three columns with their corresponding weight.

	Number	District	Latitude	Longitude	Population16 to 64 ages	N_facilities	N_venues	N_facilities_Normalized	Population_Normalized	N_venues_Normalized	Score
11	10	Latina	40.388969	-3.745690	149.500	34	100	0.68	0.895285	1.0	0.813821
8	13	Puente de Vallecas	40.393540	-3.662000	160.115	27	100	0.54	0.958853	1.0	0.759713
18	11	Carabanchel	40.383669	-3.727989	166.986	23	100	0.46	1.000000	1.0	0.730000
20	2	Arganzuela	40.400021	-3.696180	104.784	31	100	0.62	0.627502	1.0	0.716875
17	1	Centro	40.418308	-3.702750	102.065	30	100	0.60	0.611219	1.0	0.702805

Figure 16: Madrid Districts dataset

4. Results

By sorting the ranking dataset, these are the TOP 10 districts to open the new Sports Shop.

	Number	District	Latitude	Longitude	Population16 to 64 ages	N_facilities	N_venues	Population_Normalized	N_facilities_Normalized	N_venues_Normalized	Score
0	10	Latina	40.388969	-3.745690	149.500	34	100	0.895285	0.68	1.00	0.813821
1	16	Hortaleza	40.474441	-3.641100	119.751	50	41	0.717132	1.00	0.41	0.781783
2	13	Puente de Vallecas	40.393540	-3.662000	160.115	27	100	0.958853	0.54	1.00	0.759713
3	11	Carabanchel	40.383669	-3.727989	166.986	23	100	1.000000	0.46	1.00	0.730000
4	2	Arganzuela	40.400021	-3.696180	104.784	31	100	0.627502	0.62	1.00	0.716875
5	1	Centro	40.418308	-3.702750	102.065	30	100	0.611219	0.60	1.00	0.702805
6	5	Chamartín	40.451000	-3.675000	91.757	31	100	0.549489	0.62	1.00	0.697372
7	8	Fuencarral-El Pardo	40.498402	-3.731400	150.648	37	29	0.902159	0.74	0.29	0.668040
8	6	Tetuán	40.459751	-3.697500	108.901	24	100	0.652156	0.48	1.00	0.653039
9	12	Usera	40.388660	-3.700350	93.873	26	100	0.562161	0.52	1.00	0.650540

Figure 17: Top 10 districts

The districts belonging to the top 5 have something in common, they are all located on the outskirts of Madrid. We can see this in the figures 9 and 10, which show the few facilities in the center and the large number on the outskirts. In addition, it also coincides that the selected districts have a large population.

5. Conclusions

In this project, all the stages described in IBM Data Science have been carried out and all the concepts referring to data search, treatment and exploration have been used. In addition, one of the machine learning techniques explained has been applied: KMeans.

The study has allowed determining the most favorable locations to open a sports store with the data consulted. But this does not end here, the study could be extended by also adding private facilities to the study and comparing the prices of commercial places in those districts belonging to the top 5.