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

Introduction - Business Problem

- *Problem Background:* high cost of business of the city of Madrid
- *Problem Description:* set up a new sports shop called MADSport
- *Location requirements:*
  - Sports facilities nearer
  - Popular venues
  - High number of population

Data description

- *Madrid districts:* districts and geographical coordinates.
- *Madrid census populations:* population between 16-64 years per district.
- *Madrid sports facilities:* name, type of facility and geographical coordinates.

Methodology

Data obtention and cleaning

Madrid districts

	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

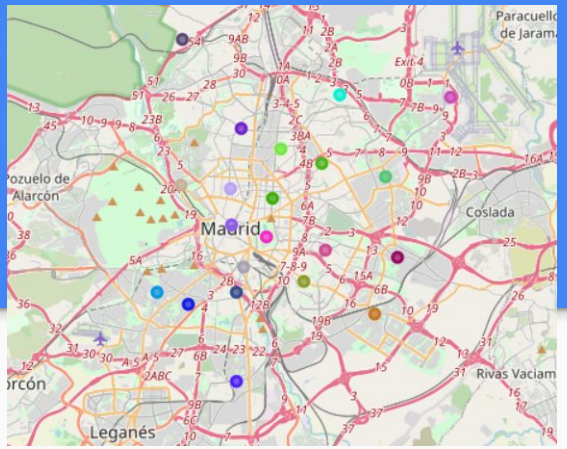
Madrid sports facilities

	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

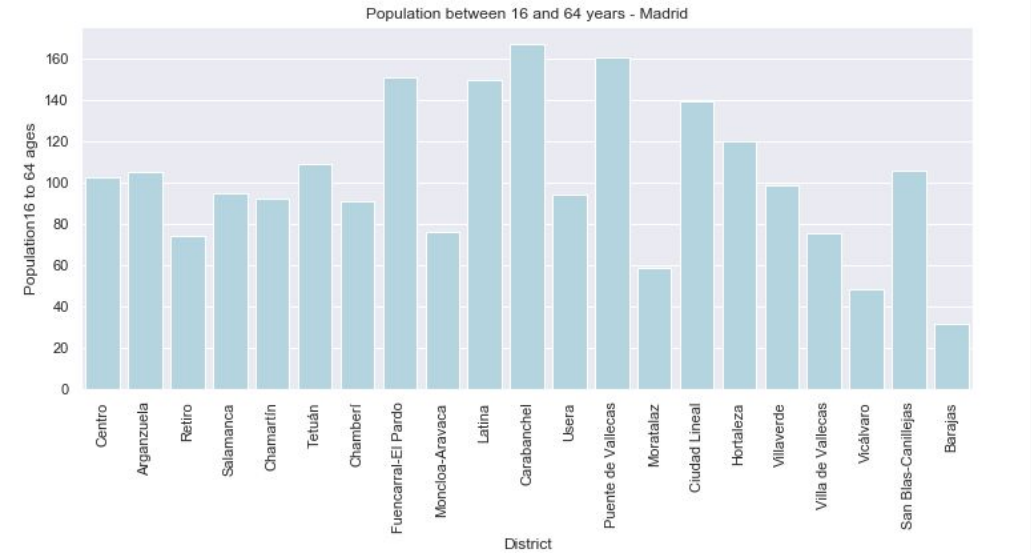
Foursquare API

	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

Exploratory data analysis: districts

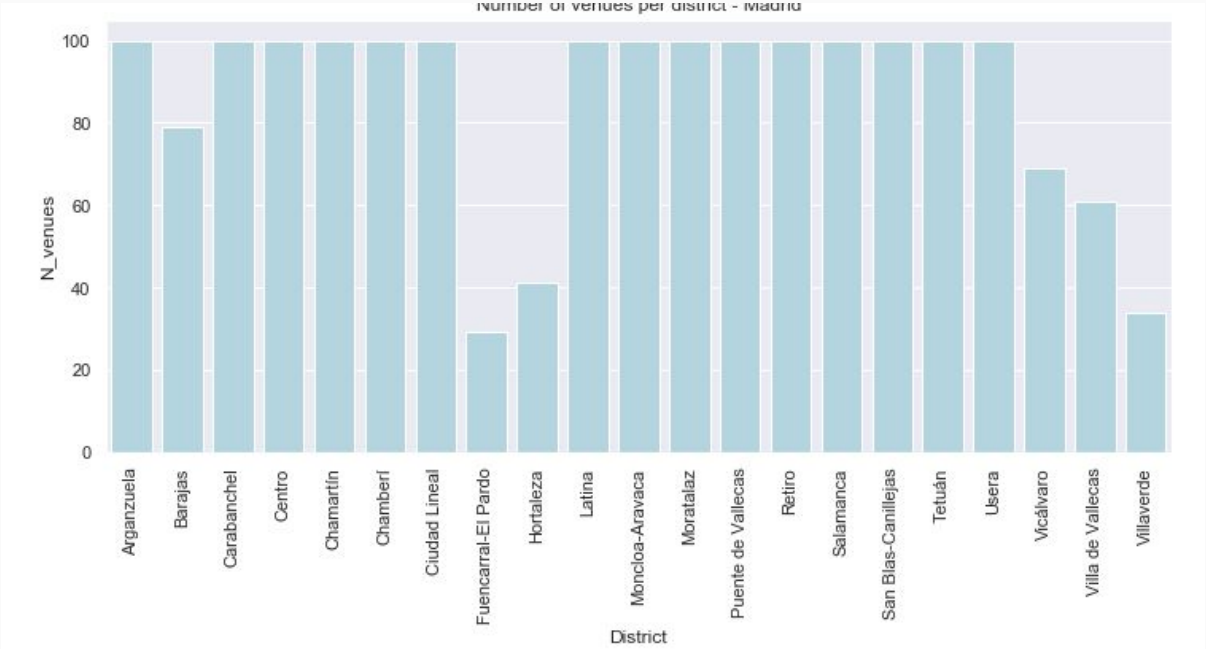


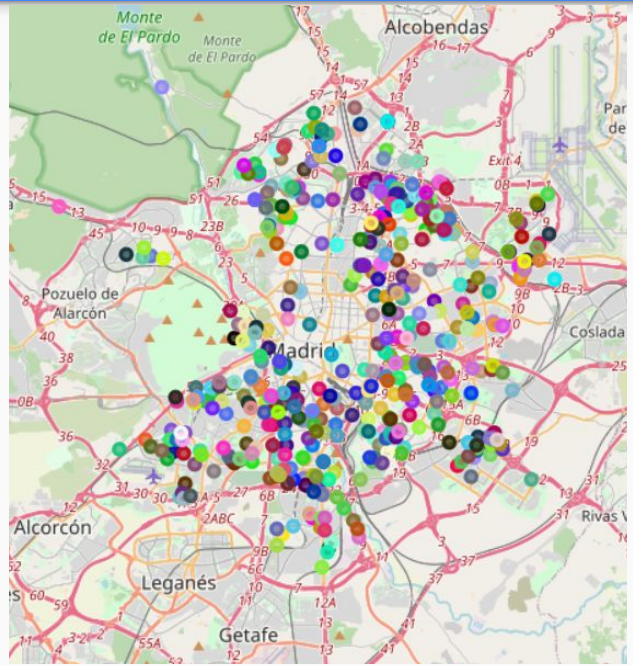
- 21 districts
- Most populated districts:
  - Carabanchel
  - Puente de Vallecas
- Less populated districts:
  - Moratalaz
  - Barajas



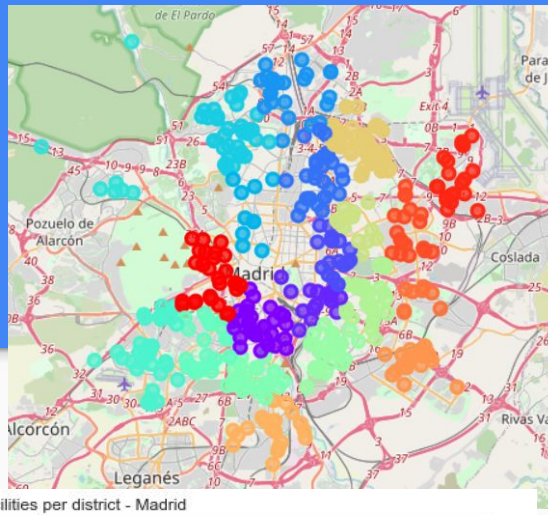
Exploratory data analysis: sport facilities

- 490 sport facilities in Madrid





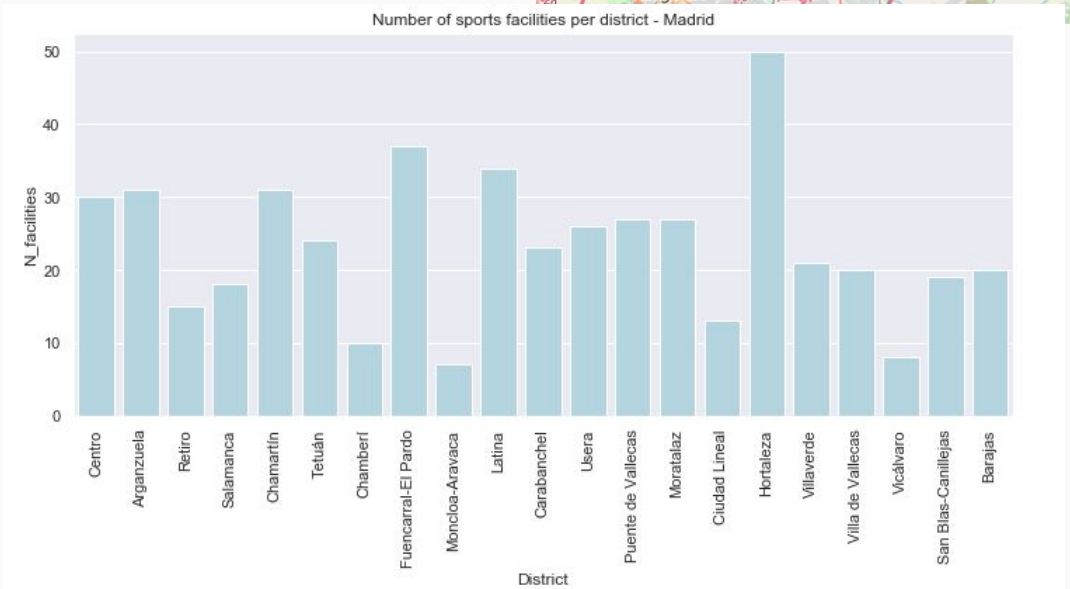
Clustering



- KMeans
- Geographical coordinates of

  each district as starting points

- 21 clusters
- Districts with most facilities:
  - Hortaleza
  - Fuencarral-El Pardo



Candidates districts: Ranking

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%

Candidates districts: Ranking

	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

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

Candidates districts: Ranking

	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

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

Candidates districts: Ranking

	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

Results

- Top 5 districts located in the outskirts
- Top 5 locations between the ones with most population

	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

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.