CAPSTONE PROJECT – GYM LOCATION IN BARCELONA

IBM Data Science Professional Certificate

Abstract

Exploration and Segmentation of Neighbourhoods in Barcelona by their relative fitness to host a newly created gym or fitness center.

Guillermo Brugarolas Sobejano

gbrugarolas@gmail.com

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

Content

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

1. Introduction/Business Problem

Gyms/fitness centers are a type of business that is going up in popularity in recent years, as more and more people become aware of the benefits of getting and staying in shape. In many big metropolises as well as small towns there is a fierce competition between the different gyms/fitness centers, which can be an individual, small business (rare), or be a part of a brand or chain. There is competition for the best location, the best facilities, the best value for money for their users…

This project is set in the city of **Barcelona, Spain**. It is a famous Mediterranean city with a very distinctive Catalan culture. Its nice weather and beautiful landmarks attract millions of visitors from all over the world every year.

The city has an intense relationship with sport and fitness; you can see people practicing a sport at almost any time, be it swimming in the sea, jogging in one of the long avenues, playing football or, of course, weight-lifting in one of the city's many gyms.

One of the big questions to answer before deciding to start a gym/fitness center is: What are the best locations? The location of a gym or fitness center can determine its future success or failure to attract clients as much as other factors. So, this not a question with an easy, straight-forward answer. One might look at locations in terms of municipalities, or districts, or neighbourhoods, or even in terms of streets within a neighbourhood. Of course, determining what makes a location suitable or fit for hosting a new gym/fitness center is up to the analyst.

This project was done to answer the question: **Which are the best and worst neighbourhoods in Barcelona to open or start a new gym/fitness center?** This question was answered based on the following three "common-sense" facts, in order of significance:

1. Assuming that all neighbourhoods are, more or less, the same demographically speaking, neighbourhoods with a smaller gym per population density ratio should be more interesting. Neighbourhoods that are packed with these types of businesses are not recommended because of the high competition that is already there.

2. A well-connected neighbourhood with many public transport stations, bike rentals, etc. nearby should be more interesting, as the potential clients of the gym/fitness center need to have easy access to it; if they do not, they stop going there, and therefore stop paying their membership.

3. A neighbourhood with many different services nearby, such as medical centers, food & drink stores, police stations, etc. should also be more interesting. The potential clients of the gym/fitness center might need quick medical assistance if they injury themselves, might want to grab a snack right before or after their workout session, or might need to report a theft/robbery immediately. The assumption here is that a client who has all these services right next or very close to the gym/fitness center will be a little more likely to stay/pay.

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

2. Data

2.1. Foursquare Location Data

The first group of datasets is obtained from the Foursquare location API service. In particular, the API endpoint ***explore*** was used to obtain venues data for every neighbourhood in the city. This data is used in the project to obtain the ratios and parameters that are later used in the clustering of the neighbourhoods.

2.2. Barcelona City Council Open Data

The Barcelona City Council has an open web service from which one can obtain different types of datasets: https://opendata-ajuntament.barcelona.cat/

This service offers a wide range of datasets related to the city, from demographical data to geographical data. The data of most if not all datasets found is up to date.

To work with in this project, two datasets were downloaded, cleaned, and pre-processed:

these are snippets of the two datasets

1. Barcelona_Neighbourhoods_GeographicalData:

Year	District_Code	District_Name	Neighbourhood_Code	Neighbourhood_Name	Area (ha)	Latitude	Longitude
2019	1	Ciutat Vella	1	el Raval	110	41.38	216.861
2019	1	Ciutat Vella	2	el Barri Gotic	81.6	41.382.778	2.176.944
2019	1	Ciutat Vella	3	la Barceloneta	109.5	4.137.944	218.917
2019	1	Ciutat Vella	4	Sant Pere Santa Caterina i la Ribera	111	413.847	21.826
2019	2	Eixample	5	el Fort Pienc	92.9	41.395.675	2.183.703
2019	2	Eixample	6	la Sagrada Familia	104.2	41.403.561	2.174.347
2019	2	Eixample	7	la Dreta de l'Eixample	212	41.395.278	2.166.667
2019	2	Eixample	8	l'Antiga Esquerra de l'Eixample	122.8	41.390.061	2.155.061
2019	2	Eixample	9	la Nova Esquerra de l'Eixample	134.1	41.383.389	2.149
2019	2	Eixample	10	Sant Antoni	80.4	4.137.801	215.949
2019	3	Sants-Montjuic	11	el Poble Sec	458.5	4.137.278	21.625
2019	3	Sants-Montjuic	12	la Marina del Prat Vermell	1403.4	41.352.103	2.139.436

2. Barcelona_Neighbourhoods_PopulationData:

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

Year	District_Code	District_Name	Neighbourhood_Code	Neighbourhood_Name	Sex	Year_registered	Number
2020	10	Sant Marti	71	Provensals del Poblenou	Dona	Menys d'1 any	637
2020	10	Sant Marti	72	Sant Marti de Provensals	Dona	Menys d'1 any	757
2020	10	Sant Marti	73	la Verneda i la Pau	Dona	Menys d'1 any	792
2020	1	Ciutat Vella	1	el Raval	Dona	D'1 a 5 anys	7395
2020	1	Ciutat Vella	2	el Barri Gotic	Dona	D'1 a 5 anys	3229
2020	1	Ciutat Vella	3	la Barceloneta	Dona	D'1 a 5 anys	2277
2020	1	Ciutat Vella	4	Sant Pere Santa Caterina i la Ribera	Dona	D'1 a 5 anys	3890
2020	2	Eixample	5	el Fort Pienc	Dona	D'1 a 5 anys	3893
2020	2	Eixample	6	la Sagrada Familia	Dona	D'1 a 5 anys	6276
2020	2	Eixample	7	la Dreta de l'Eixample	Dona	D'1 a 5 anys	5510
2020	2	Eixample	8	l'Antiga Esquerra de l'Eixample	Dona	D'1 a 5 anys	5426
2020	2	Eixample	9	la Nova Esquerra de l'Eixample	Dona	D'1 a 5 anys	6711
2020	2	Eixample	10	Sant Antoni	Dona	D'1 a 5 anys	4516
2020	3	Sants-Montjuïc	11	el Poble Sec	Dona	D'1 a 5 anys	5112

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

3. Methodology

In the development of this project, a certain methodology was followed to guarantee the consistency and the accuracy of the techniques applied and of the results.

The methodology is divided in the following five sections, that were developed and completed in a sequential manner:

3.1. Download, Pre-process, and Explore Datasets

In this first section, the two datasets explained in the previous section are downloaded and converted into pandas DataFrames.

Pre-processing tasks were performed to put the data of both datasets in the format needed for the development of this project.

High-level exploratory analysis was performed, to visualize the contents of both datasets.

A final dataset was obtained from the merge of these two original datasets:

	Neighbourhood_Name	Number	Year	District_Code	District_Name	Neighbourhood_Code	Area (ha)	Latitude	Longitude	Population Density
0	Baro de Viver	2625	2019	9	Sant Andreu	58	23.0	41.447906	2.200742	114.130435
1	Can Baro	9331	2019	7	Horta-Guinardo	34	38.4	41.416384	2.162356	242.994792
2	Can Peguera	2234	2019	8	Nou Barris	47	11.9	41.434900	2.166188	187.731092
3	Canyelles	6869	2019	8	Nou Barris	49	79.0	41.442684	2.166015	86.949367
4	Ciutat Meridiana	11091	2019	8	Nou Barris	55	37.7	41.460914	2.174433	294.190981
5	Diagonal Mar i el Front Maritim del Poblenou	13526	2019	10	Sant Marti	69	120.3	41.409600	2.216306	112.435578
6	Horta	28363	2019	7	Horta-Guinardo	43	307.0	41.429503	2.160100	92.387622
7	Hostafrancs	16203	2019	3	Sants-Montjuic	15	41.0	41.375556	2.143056	395.195122
8	Montbau	5225	2019	7	Horta-Guinardo	40	205.5	41.430925	2.142967	25.425791
9	Navas	22457	2019	9	Sant Andreu	63	42.4	41.415744	2.186900	529.646226
10	Pedralbes	11936	2019	4	Les Corts	21	268.5	41.394722	2.113056	44.454376

3.2. Explore Gyms/Fitness Centers, Communications and Services by Neighborhoods in Barcelona

In this section, the API endpoint function ***explore*** is used to get the list of venues for every neighbourhood in the city.

Specifically, different category IDs were passed as parameters to the function, to get the Gyms/Fitness centers, the communication & transports and the different services located in every neighbourhood. These are the three main categories of venues that are included in the cluster analysis.

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

3.3. Standardization and Feature Weighting

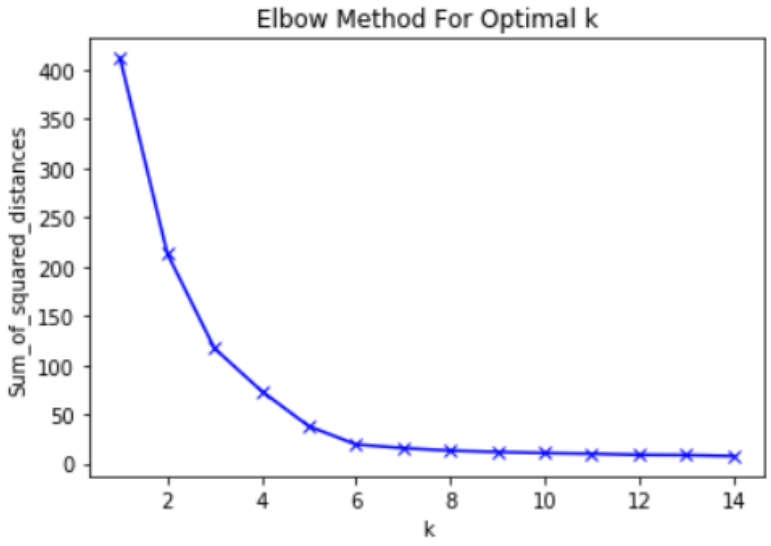
In this short section, the features for the subsequent cluster analysis are extracted and standardized.

One of the features, which is the Gyms/Population Density parameter, is assigned double the weight of the other two features, which are Communications Density and Services Density, respectively. This was done to represent computationally the fact that the current density of gyms/fitness centers in a particular neighbourhood is a much more important factor to consider (because of the competition) than the other two factors.

3.4. Cluster Neighborhoods

In this section, the 3D K-Means clustering analysis is performed.

Firstly, the elbow method is used to determine the best initial number of clusters (centroids):



A value of **k = 4** is selected.

The K-Means algorithm is run several times and the results are averaged.

A visualization map with Folium is shown, where the different neighbourhoods, represented by a circle located at their center (approximately), are assigned the colour of the cluster they fall into.

3.5. Examine Clusters

In this last section, the 4 resulting clusters are examined, and a name is given to every one of them, according to the common characteristics of its members.

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

4. Results

The results were shown with a Folium map.

Finally, let's visualize the resulting clusters

In [ ]:
```
# create map
map_clusters = folium.Map(location = [latitude, longitude], zoom_start = 11)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(df_nhs_bcn['Latitude'], df_nhs_bcn['Longitude'], df_nhs_bcn['Neighbourhood_Name'], df_nhs_bcn['Cluster Labels']):
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

{"meta":{"code":429,"errorType":"quota_exceeded","errorDetail":"Quota exceeded","requestId":"6016c13c1d350c33c6932c6a"},"response":{}}

Four clusters of neighbourhoods were generated, based on to the values of the mentioned three features.

CLUSTER NUM.	CLUSTER NAME	CLUSTER POPUL.	GYMS/POP. DENSITY	COMMS DENSITY	SERVICES DENSITY
1	Outlier	1	Very High	High	High
2	Risky	40	High	High	Low
3	Normal Interesting	20	Low	Low	High
4	Opportunity	12	Very Low	High	High

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

5. Observations

According to the results shown above, there are at least twelve neighbourhoods (approximately) in Barcelona that could be recommended to be the location of a new gym/fitness center, based on the density of those different types of venues in their area.

But these results come from one of many different approaches that can be taken to answer the original question: Which are the best neighbourhoods to open/start a new gym or fitness center?

Other types of clustering algorithms could have been used in this project, such as density-based algorithms (DBSCAN) or other.

There were limitations as for the use of the Foursquare API. Apart from the Premium calls not being available, the daily regular call limitations prevented the project to expand because of the related time restrictions.

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

6. Conclusions

As was already explained in the previous section, according to the results shown above, there are at least twelve neighbourhoods (approximately) in Barcelona that could be recommended to be the location of a new gym/fitness center, based on the density of those different types of venues in their area.

The insights provided by the results of this project are considered valuable to hypothetical individual entrepreneurs or to companies in the fitness industry looking for locations for their new gyms or centers.

A more accurate description of the city neighbourhoods could have been included in this project, to be able to use other powerful parameters, such as median household income or wealth per capita, in the clustering analysis.

IMPORTANT NOTE: The notebook could not be executed completely due to the limitation of daily Regular Foursquare API calls. The visualizations and maps could therefore not be generated at the time of delivery. This will all be updated with the definitive version as soon as the Foursquare API calls are available again. Thank you for your understanding.

7. References

Coursera IBM Data Science Professional Certificate

https://es.foursquare.com/developers

http://en.wikipedia.org/