

The Battle of Neighborhoods: Capstone Project

Cluster analysis to choose where to open a gym in Palermo

Rosario Portera

16/01/2021

Introduction

Today people are more and more attentive to their physical appearance, and spend many hours a week training.

The main place to do physical activity are the gyms, in the gym thanks to the guidance of an expert coach you can achieve results faster, moreover the gyms are a great place to meet other people.

The purpose of this project is to analyze and select the best locations in the city of Palermo, in Italy, to open a new gym.

Using data science methodology and machine learning techniques like clustering, we will carry out the geospatial analysis of the city of Palermo to understand what would be the best place to open a new gym. The population of Palermo urban area is estimated by Eurostat to be 855,285, while its metropolitan area is the fifth most populated in Italy with around 1.2 million people. In the central area, the city has a population of around 676,000 people.

Data

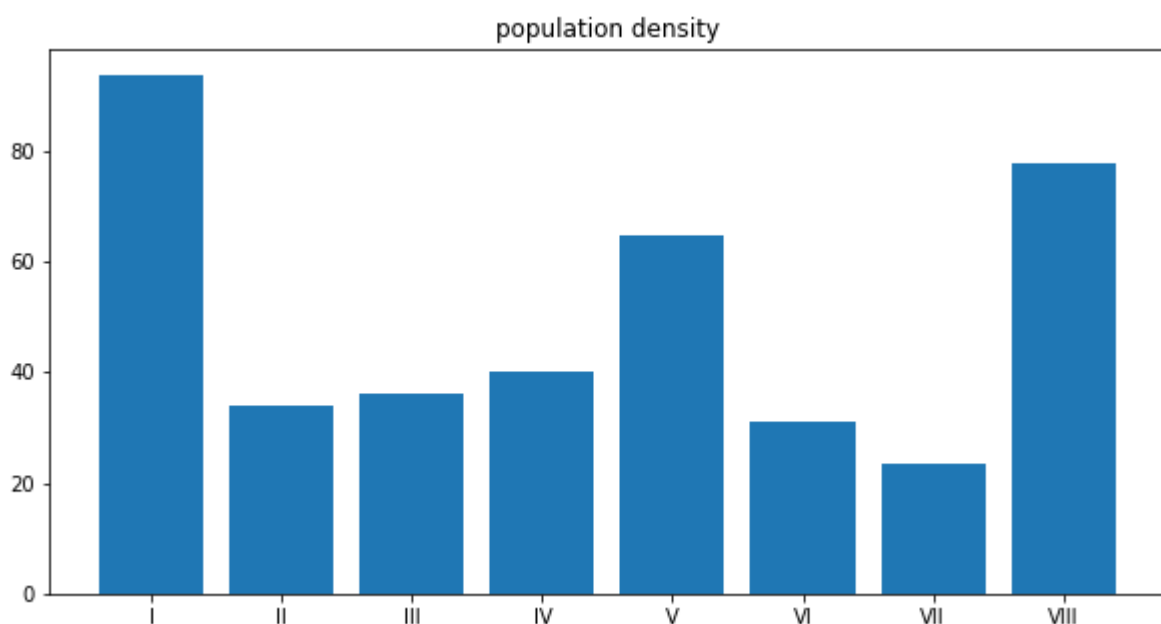
To solve the problem, we will need the following data:

1. List of neighbourhoods in Palermo.
2. Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data
3. Venue data, particularly data related to gyms. We will use this data to perform clustering on the neighbourhoods.

From Wikipedia https://it.wikipedia.org/wiki/Circoscrizioni_di_Palermo, we collect data relating to the subdivision of the city, the surface and the population of each subdivision,

	borough	population	area [ettari]	population density
0	I	23384	249.7	93.648378
1	II	72888	2139.0	34.075736
2	III	73509	2034.7	36.127685
3	IV	104448	2616.3	39.922027
4	V	113425	1753.1	64.699675
5	VI	73895	2390.0	30.918410
6	VII	76671	3295.5	23.265362
7	VIII	119341	1532.7	77.863248

We select only the districts where the population density exceeds the average



	borough	population	area [ettari]	population density
0	I	23384	249.7	93.648378
4	V	113425	1753.1	64.699675
7	VIII	119341	1532.7	77.863248

we use the geocoder library to derive the latitude and longitude of each Neighborhood and of Palermo:

	Borough	Neighborhood	Latitude	Longitude
0	I	Tribunali	38.118578	13.351957
1	I	Palazzo Reale	38.108921	13.354074
2	V	Zisa	38.116774	13.341055
3	V	Uditore-Passo di Rigano	38.127320	13.318294
4	V	Borgo Nuovo	38.129106	13.294677
5	VIII	Politeama	38.124659	13.355521
6	VIII	Libertà	38.139227	13.346306
7	VIII	Montepellegrino	38.139013	13.358323
8	VIII	Malaspina	38.127542	13.342255

The geograpical coordinate of Palermo, Italy are 38.1112268, 13.3524434.

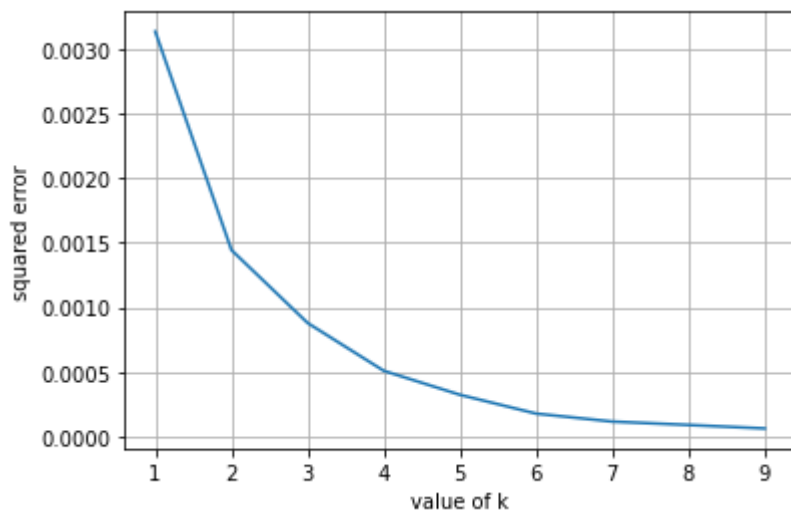
Methodology

1. Using FourSquare API we will find all venues for each neighborhood.
2. We will filter out all Palermo Gyms.
3. we clean up the data
4. we will Cluster Neighborhoods with KMeans

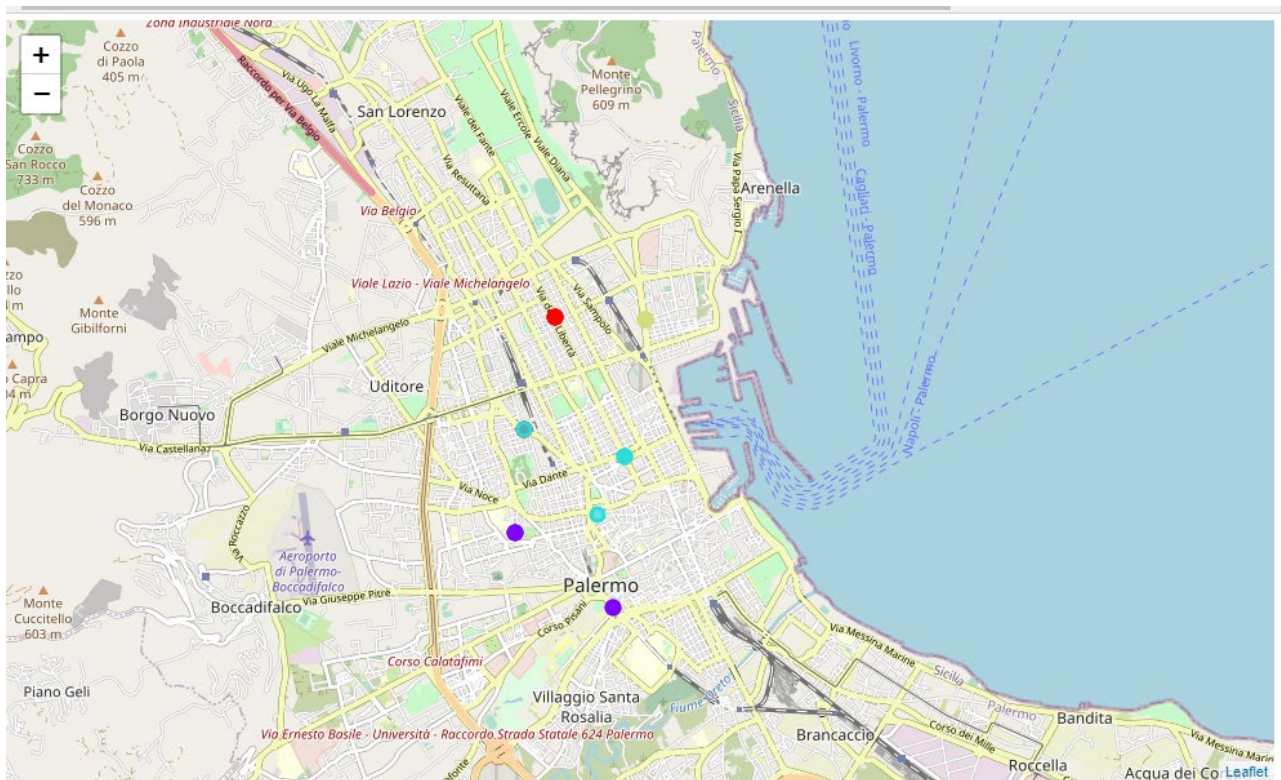
Now We use Foursquare for derive the Gym in Palermo and deryve all the gyms in Neighborhood of Palermo

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Tribunali	38.118578	13.351957	Paradisea	38.122446	13.351343	Yoga Studio
1	Palazzo Reale	38.108921	13.354074	Marte Club	38.105303	13.353556	Gym
2	Palazzo Reale	38.108921	13.354074	Gay	38.112378	13.352096	Track
3	Zisa	38.116774	13.341055	metabolica	38.113922	13.340920	Gym
4	Zisa	38.116774	13.341055	Palestra Metabolica	38.113749	13.341366	Gym
5	Zisa	38.116774	13.341055	A.S.D. Jogging	38.113158	13.341767	Gym
6	Politeama	38.124659	13.355521	Body Studio City Center	38.123629	13.354908	Gym / Fitness Center
7	Politeama	38.124659	13.355521	Dakar fitness club	38.125699	13.357197	Gym
8	Politeama	38.124659	13.355521	Paradisea	38.122446	13.351343	Yoga Studio
9	Libertà	38.139227	13.346306	Omnia Fitness	38.136677	13.344625	Gym / Fitness Center
10	Libertà	38.139227	13.346306	Fontana Health Club	38.136639	13.343261	Gym / Fitness Center
11	Libertà	38.139227	13.346306	Attiva	38.140974	13.349706	Gym
12	Libertà	38.139227	13.346306	Centro di Cultura Rishi	38.142202	13.350374	Yoga Studio
13	Montepellegrino	38.139013	13.358323	Astoria Palace Gym & Spa	38.139297	13.357835	Gym
14	Montepellegrino	38.139013	13.358323	L'accademia	38.135684	13.358600	Gym
15	Montepellegrino	38.139013	13.358323	Atlantis Fitness	38.135642	13.360993	Gym
16	Montepellegrino	38.139013	13.358323	Scalart	38.135124	13.356839	Climbing Gym
17	Malaspina	38.127542	13.342255	Palestra Body Mind	38.129240	13.340362	Gym
18	Malaspina	38.127542	13.342255	Orso Bianco	38.129232	13.346064	Gym
19	Malaspina	38.127542	13.342255	Palermo Yoga - Associazione Salvatore Sanfilippo	38.131575	13.344676	Yoga Studio
20	Malaspina	38.127542	13.342255	Aikya Training Center	38.126767	13.346951	Martial Arts School

finding the best value of K for Kmean



From the above image, we see that the best value of K will be 4 according to the Elbow method.



Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster:

Cluster 0

	Cluster Labels	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
9	0	Libertà	38.139227	13.346306	Omnia Fitness	38.136677	13.344625	Gym / Fitness Center
10	0	Libertà	38.139227	13.346306	Fontana Health Club	38.136639	13.343261	Gym / Fitness Center
11	0	Libertà	38.139227	13.346306	Attiva	38.140974	13.349706	Gym
12	0	Libertà	38.139227	13.346306	Centro di Cultura Rishi	38.142202	13.350374	Yoga Studio
17	0	Malaspina	38.127542	13.342255	Palestra Body Mind	38.129240	13.340362	Gym
19	0	Malaspina	38.127542	13.342255	Palermo Yoga - Associazione Salvatore Sanfilippo	38.131575	13.344676	Yoga Studio

Cluster 1

	Cluster Labels	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	1	Palazzo Reale	38.108921	13.354074	Marte Club	38.105303	13.353556	Gym
2	1	Palazzo Reale	38.108921	13.354074	Gay	38.112378	13.352096	Track
3	1	Zisa	38.116774	13.341055	metabolica	38.113922	13.340920	Gym
4	1	Zisa	38.116774	13.341055	Palestra Metabolica	38.113749	13.341366	Gym
5	1	Zisa	38.116774	13.341055	A.S.D. Jogging	38.113158	13.341767	Gym

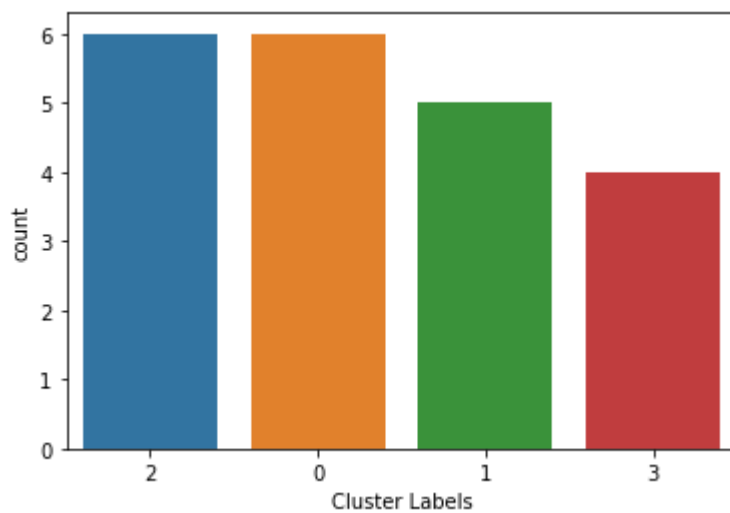
Cluster 2

	Cluster Labels	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	2	Tribunali	38.118578	13.351957	Paradisea	38.122446	13.351343	Yoga Studio
6	2	Politeama	38.124659	13.355521	Body Studio City Center	38.123629	13.354908	Gym / Fitness Center
7	2	Politeama	38.124659	13.355521	Dakar fitness club	38.125699	13.357197	Gym
8	2	Politeama	38.124659	13.355521	Paradisea	38.122446	13.351343	Yoga Studio
18	2	Malaspina	38.127542	13.342255	Orso Bianco	38.129232	13.346064	Gym
20	2	Malaspina	38.127542	13.342255	Aikya Training Center	38.126767	13.346951	Martial Arts School

Cluster 3

	Cluster Labels	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
13	3	Montepellegrino	38.139013	13.358323	Astoria Palace Gym & Spa	38.139297	13.357835	Gym
14	3	Montepellegrino	38.139013	13.358323	L'accademia	38.135684	13.358600	Gym
15	3	Montepellegrino	38.139013	13.358323	Atlantis Fitness	38.135642	13.360993	Gym
16	3	Montepellegrino	38.139013	13.358323	Scalart	38.135124	13.356839	Climbing Gym

plot gyms for any clusters



Result

The results of clustering analysis is summarized below :

- in clusters 2 and 0 there is the largest number of gyms
- cluster 0 has the least number of gyms

Discussion

According to the analysis, cluster 3 will provide the least competition for an upcoming gyms in Palermo. Some drawbacks of analysis are: the clustering is completely based on the data provided by Foursquare API. Furthermore, these results also could potentially vary if we use some other clustering techniques like DBSCAN.

Conclusion

Finally, to conclude this project, I have used some frequently used python libraries to plotting graphs, and other exploratory data analysis. Use Foursquare API to major boroughs of Palermo and their neighborhoods. All of the above analyses is depended on the adequacy and accuracy of Four Square data. A more comprehensive analysis and future work would need to incorporate data from other external databases.