Capstone Project Report:

Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Table of Contents

1. Introduction / Business Problem

Target of this project is finding the most promising neighborhoods in Munich, Germany for opening new fitness centers. Stakeholders are fitness trainers who want to start their own business or big fitness companies who want to expand their business within Munich. The final result of this project is a single key indicator for the 25 districts of Munich which allows the stakeholders to identify the best location for their new business activities. This indicator does not only reflect the total number of existing gyms but also take into account different factors from the structure of the population in Munich (e.g. Population density, Age structure, unemployment rate, etc.).

2. Data

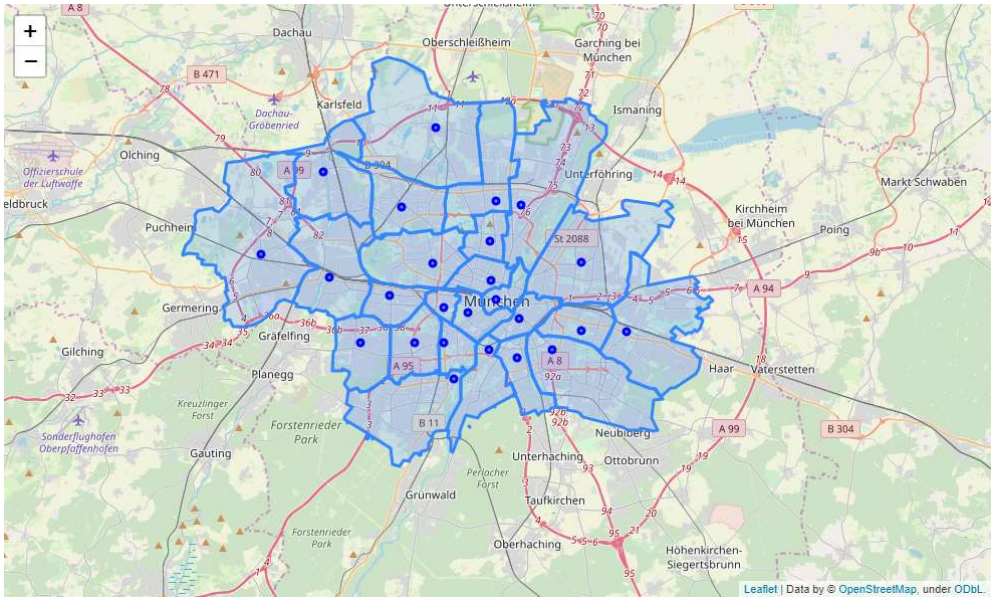
2.1 Geo-Data for the neighborhoods in Munich

The geojson file for the Neighborhoods in Munich can be found here:
https://gist.github.com/webtobesocial/f8c1ea64b2862c6eda8771daba4f297b

In the next step the neighborhood names are read from the geojson file and coordinates for each neighborhood are located with geopy geolocator function. The result is a dataframe with all Neighborhoods in Munich including coordinates:

NeighborhoodID	Neighborhood	Latitude	Longitude
01	Altstadt-Lehel	48.1378	11.5746
02	Ludwigsvorstadt-Isarvorstadt	48.1318	11.5558
03	Maxvorstadt	48.1466	11.5714
04	Schwabing-West	48.1644	11.5704
05	Au-Haidhausen	48.1288	11.5905
06	Sendling	48.118	11.5391
07	Sendling-Westpark	48.118	11.5193
08	Schwanthalerhöhe	48.1342	11.539
09	Neuhausen-Nymphenburg	48.1542	11.5315
10	Moosach	48.1799	11.5106
11	Milbertshofen-Am Hart	48.1824	11.575
12	Schwabing-Freimann	48.1806	11.5918
13	Bogenhausen	48.1548	11.6335
14	Berg am Laim	48.1235	11.6335
15	Trudering-Riem	48.1232	11.6641
16	Ramersdorf-Perlach	48.1147	11.6132
17	Obergiesing-Fasangarten	48.1112	11.5889
18	Untergiesing-Harlaching	48.115	11.5702
19	Thalkirchen-Obersendling-Forstenried-Fürstenried-Solln	48.1013	11.5462
20	Hadern	48.1181	11.4818
21	Pasing-Obermenzing	48.1478	11.4607
22	Aubing-Lochhausen-Langwied	48.1584	11.4141
23	Allach-Untermenzing	48.196	11.457
24	Feldmoching-Hasenbergl	48.2159	11.5333
25	Laim	48.1396	11.5022

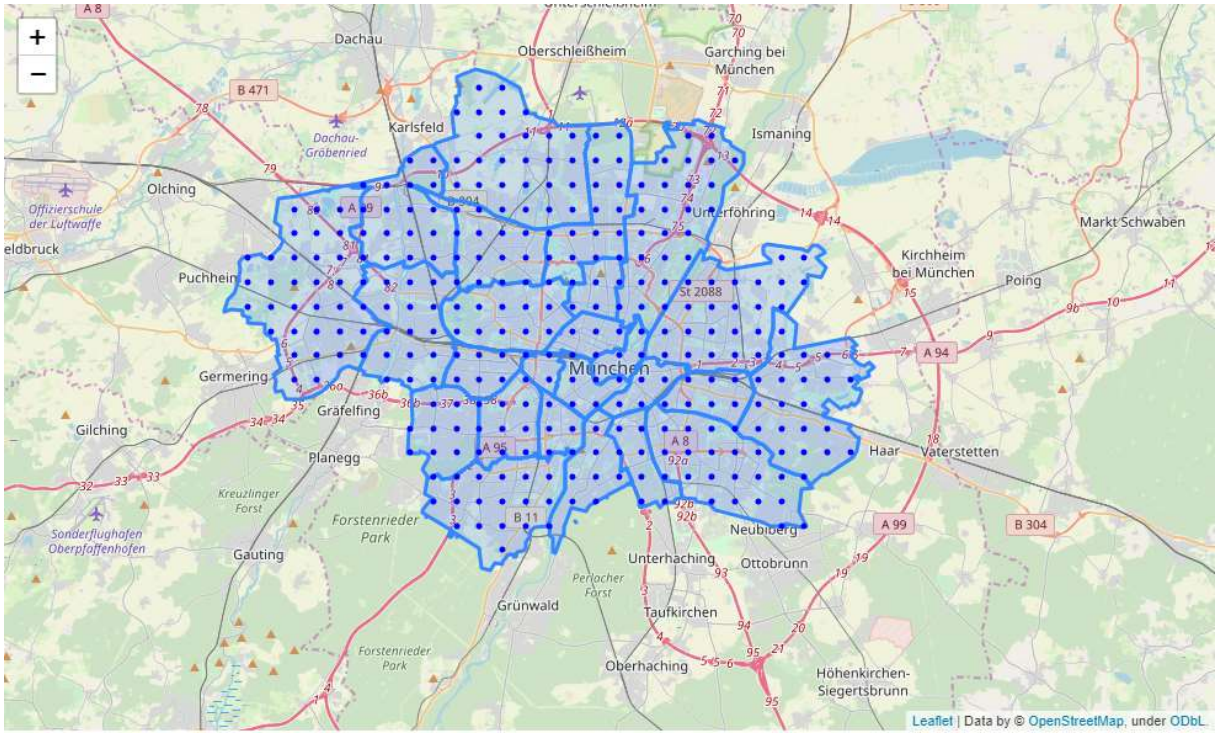
Map of Neighborhoods in Munich:



2.2 FOURSQUARE Data of fitness centers in Munich

For this project a non-commercial free FOURSQUARE Account is used. Under these conditions the search result is limited to 50 venues per search. So, the search radius has to be reduced and the number of searches have to be increased to find all Fitness Centers in Munich. Therefore, a search grid with a mesh of about 1 km is created:

Search grid for the FOURSQUARE venue search:



The FOURSQUARE search Endpoint is executed with a variable radius around 1.5 km depending on the distance to the city center. This overlapping search ensures to find all fitness centers. Duplicate venues must be dropped afterwards.
The search is filtered to the category group *Fitness Centers*.
This category group also contains several subcategories which are out of scope (e.g. Yoga Studio, Martial Arts, etc.). These subcategories will be excluded from the result venues afterwards.

In the next step the neighborhood is located for all venues and added to the DataFrame. Fitness Centers outside of the city borders are dropped.

Data cleaning:

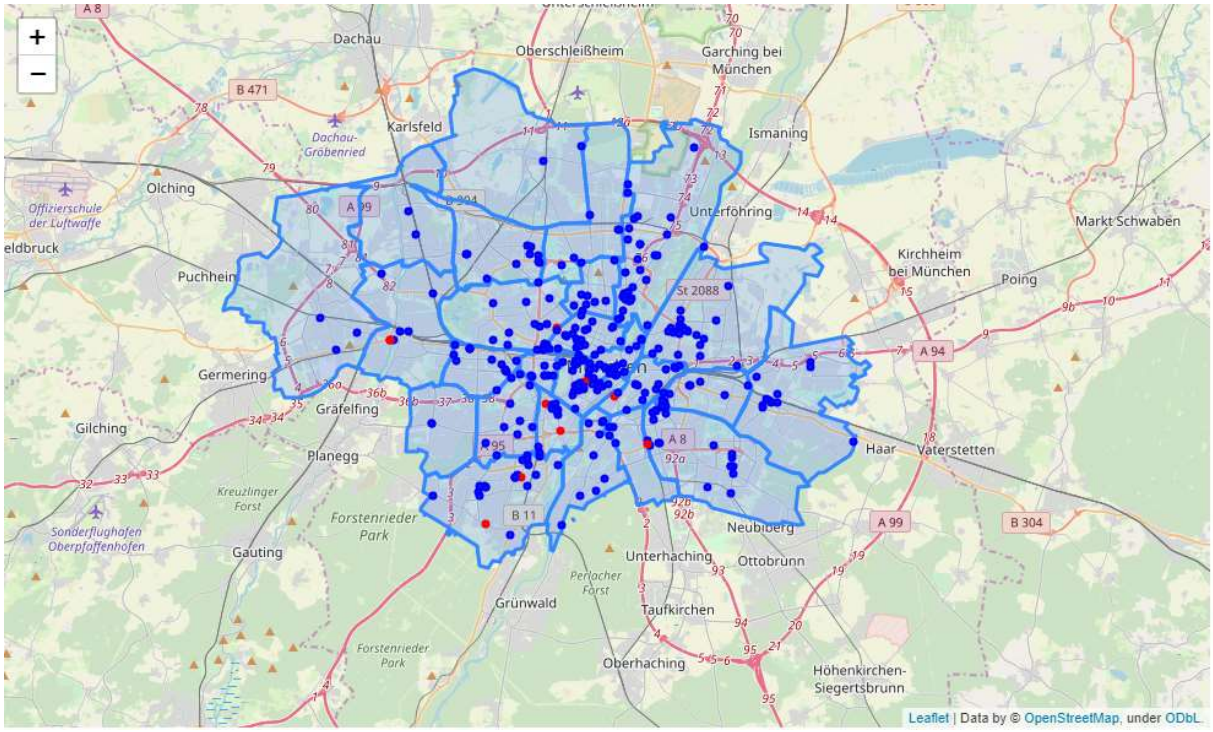
- Gyms which belong to public schools, societies or hotels are dropped from the data by keyword search

Fitness Centers which are for women only are also identified by keyword search but kept in the dataset.

Example Data after Cleaning including Neighborhood information:

	id	name	location.city	location.lat	location.lng	category.id	category.name	Neighborhood
0	4ade0d28f964a5201b6b21e3	Kieser Training	München	48.161105	11.568546	4bf58dd8d48988d175941735	Gym / Fitness Center	Schwabing-West
1	4b476e72f964a520d83126e3	körperLOUNGE	München	48.146972	11.560949	4bf58dd8d48988d175941735	Gym / Fitness Center	Maxvorstadt
2	4b4a5253f964a520c98326e3	McFIT	München	48.108272	11.580093	4bf58dd8d48988d175941735	Gym / Fitness Center	Obergiesing-Fasangarten
3	4b4b6361f964a520c79926e3	Fitness First Black Label Club	München	48.137731	11.571460	4bf58dd8d48988d175941735	Gym / Fitness Center	Altstadt-Lehel
4	4b57390ef964a520d72b28e3	McFIT	München	48.139260	11.524564	4bf58dd8d48988d175941735	Gym / Fitness Center	Laim

Map of Fitness Centers in Munich:



('Women only' marked in red)

2.3 Munich Population Statistics - OpenData

Several population statistics from https://www.opengov-muenchen.de/dataset are used for the later clustering of Neighborhoods. The aim is to get clusters of neighborhoods with similar living conditions.

The different datasets are provided by the muncipality in csv format. The following datasets are loaded:

1. Population and Population density
2. Average age
3. Single-Person Households
4. Women percentage
5. Birthrate
6. Aging index
7. Youth index
8. Unemployment rate
9. Welfare recipients rate
10. Migration background percentage

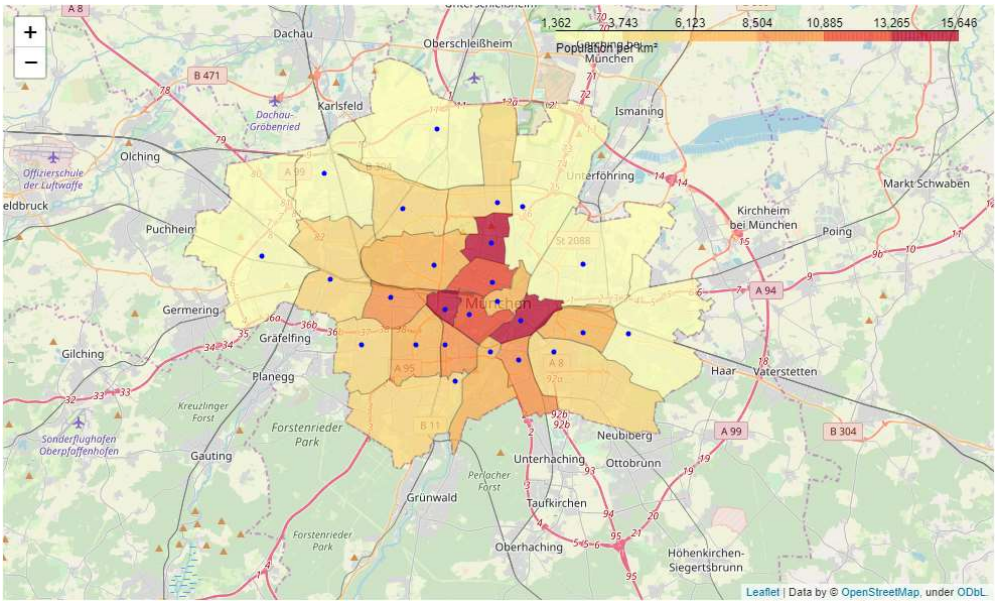
Always the latest available year of data is used (2017 or 2016).

The different files from OpenData Munich are loaded into Pandas DataFrames and finally merged together in one big DataFrame including also the fitness center data by neighborhood. The neighborhoods in the OpenData Files are named slightly different, therefore a mapping table has to be created in advance.

Example of the resulting DataFrame:

NeighborhoodID	Neighborhood	properties.name	Latitude	Longitude	Population	km2	Population_per_km2	Age_Average	single_share	women_percentage
01	01 Altstadt - Lehel	Altstadt-Lehel	48.1378	11.5746	20926	3.1	6652	41.7	64.8	49.7
02	02 Ludwigsvorstadt - Isarvorstadt	Ludwigsvorstadt-Isarvorstadt	48.1318	11.5558	51632	4.4	11731	39	64.7	48.5
03	03 Maxvorstadt	Maxvorstadt	48.1466	11.5714	51311	4.3	11939	38.6	68.9	49.9
04	04 Schwabing - West	Schwabing-West	48.1644	11.5704	68265	4.4	15646	40.8	61.6	52.6
05	05 Au - Haidhausen	Au-Haidhausen	48.1288	11.5905	60937	4.2	14441	40.1	60.3	51.3

Example for Open Data - Population Density:



3. Methodology

Overview:

1. For the fitness center data 3 key figures per neighborhood are calculated:
   a. Number of Fitness Center (gym_count)
   b. Number of Fitness Center per 1,000 inhabitants (gyms_per1000)
   c. Number of Fitness Center per km² (gyms_per_km2)

2. Based on the 10 Population key figures the 25 Neighborhoods are clustered into 5 Clusters of similar living conditions.
   Therefor skicit-learn k-means clustering is used.

3. In the next step the fitness center key figures are averaged per cluster

4. Then for each neighborhood the deviation of all 3 Fitness center key figures against the cluster average is calculated

5. Finally, the arithmetic mean of the 3 deviation figures gives the final '*Gym index*' key figure for each Neighborhood. This final figure is scaled to a base value of zero, which means no deviation to the cluster average. Values less than zero indicate Neighborhoods with less gym density than average and values greater zero indicate Neighborhoods with higher gym density than average.

3.1 Calculation of the Fitness Center key figures

Population: from OpenData Munich

Area km²: from OpenData Munich

Population Density: $\frac{Population}{Area\ [km^2]}$

Gym count: from FOURSQUARE venue search

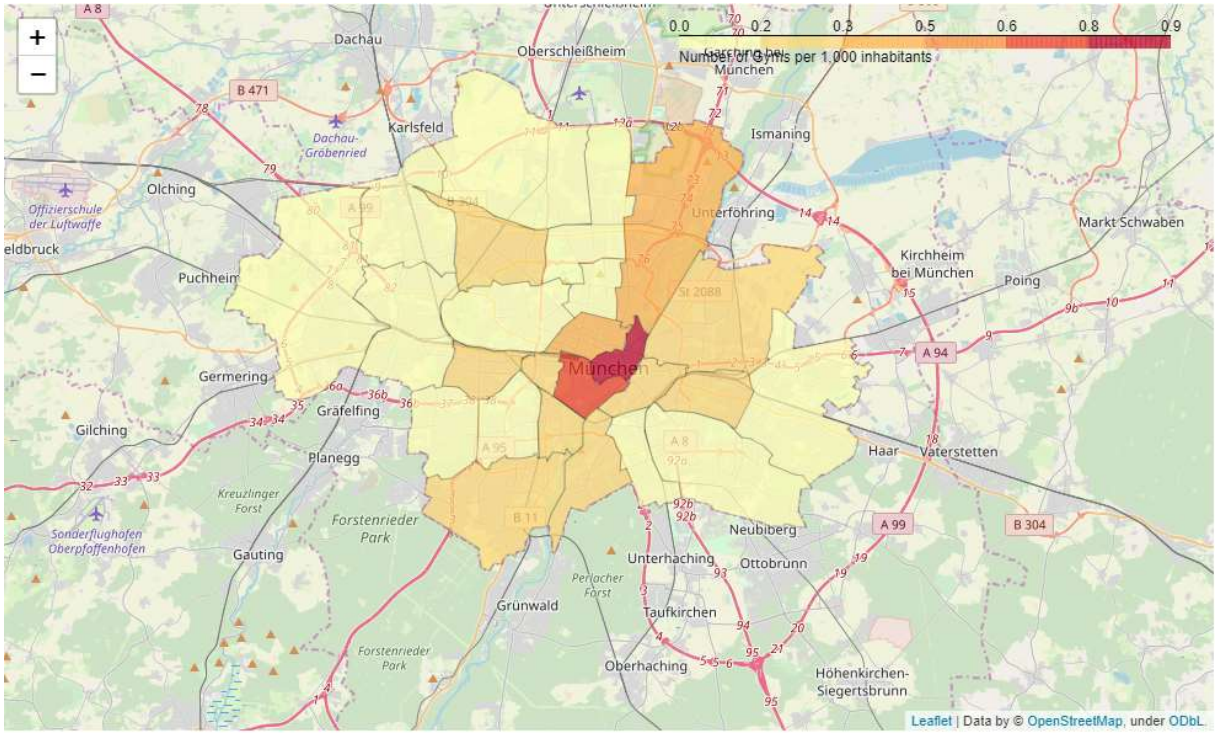
Gmys per 1,000 inhabitants: $\frac{Gym\ count}{Population} \times 1000$

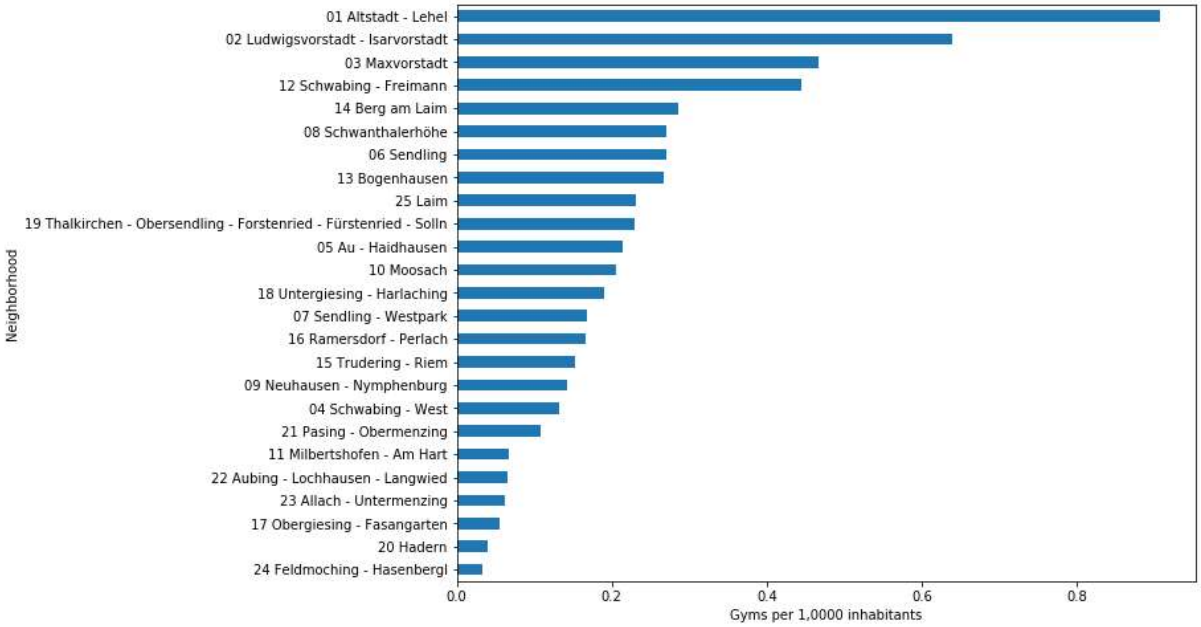
Gmys per Area km²: $\frac{Gym\ count}{Area\ km^2}$

Example values:

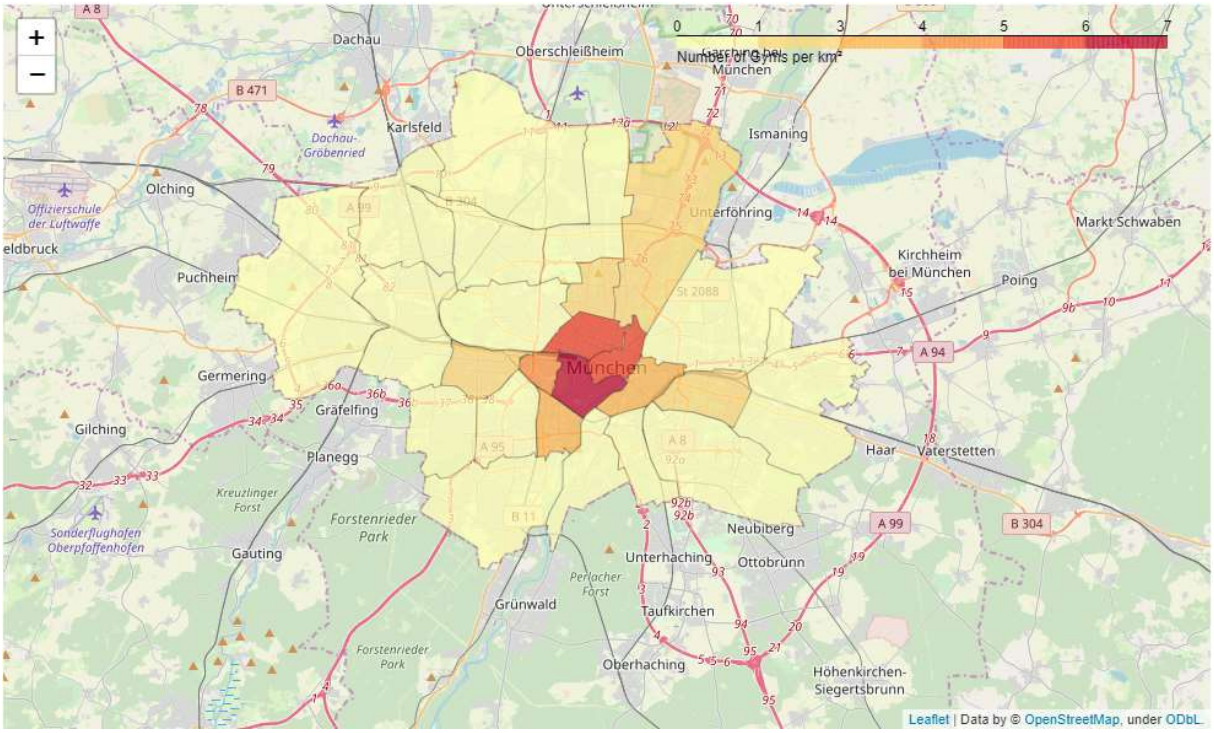
Neighborhood	Population	km2	Population per_km2	Gym count	Gyms per_1000	Gyms per_km2
01 Altstadt - Lehel	20926	3.1	6652	19	0.907961	6.129032
02 Ludwigsvorstadt - Isarvorstadt	51632	4.4	11731	33	0.639139	7.500000
03 Maxvorstadt	51311	4.3	11939	24	0.467736	5.581395
04 Schwabing - West	68265	4.4	15646	9	0.131839	2.045455
05 Au - Haidhausen	60937	4.2	14441	13	0.213335	3.095238

Number of Gyms per 1,000 inhabitants:



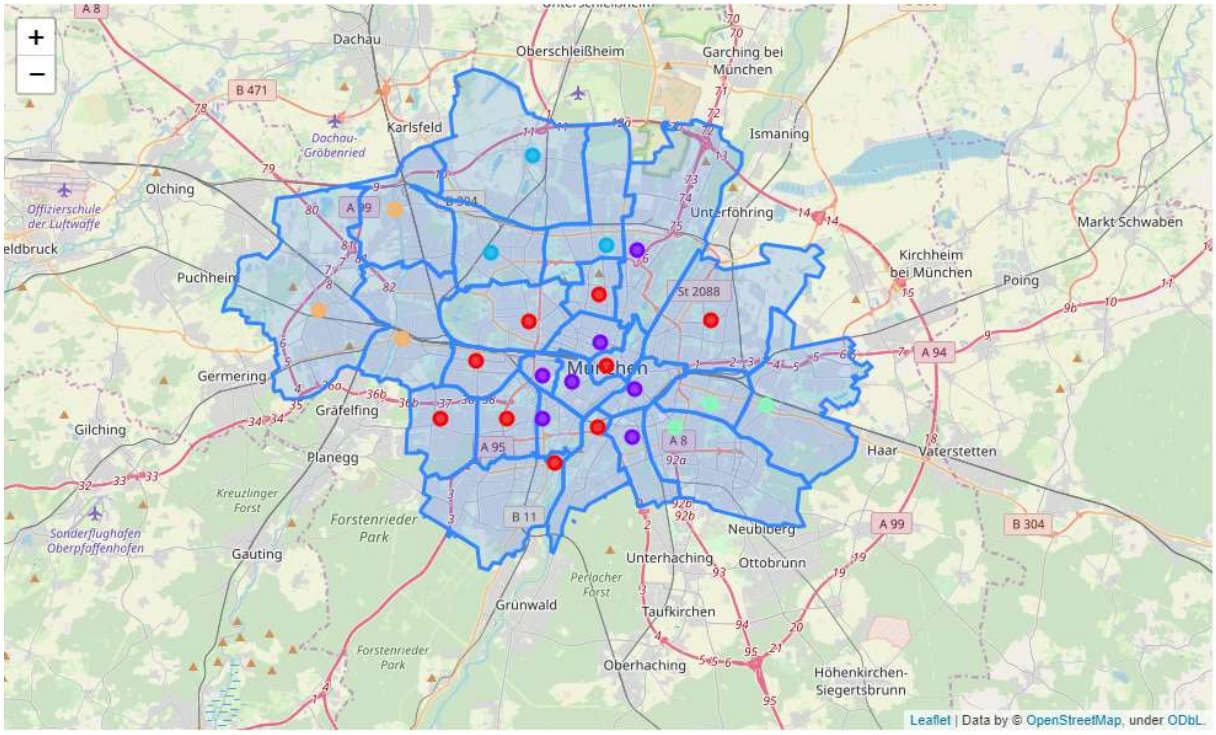


Number of Gyms per km²:



3.2 Clustering

Based on the 10 Population key figures (e.g. birthrate, average age, welfare recipient's rate, etc.) the 25 Neighborhoods are clustered into 5 Clusters of similar living conditions:



3.3 Calculate the average key figures per cluster

For the 3 key figures an average per cluster is calculated:

Label	avg_gym_count	avg_gyms_per_1000	avg_gyms_per_km2
0	13.555556	0.256607	1.850335
1	18.000000	0.337494	3.522278
2	6.000000	0.101572	0.477776
3	14.333333	0.201312	1.169052
4	4.333333	0.077949	0.233952

3.4 Calculate the key figure deviation for each Neighborhood

Now for each Neighborhood the deviation between own value and its cluster average for all 3 key figures are calculated:

deviation_gym_count = $\frac{gym\_count}{avg\_gym\_count}$

deviation_gyms_per_1000 = $\frac{gyms\_per\_1000}{avg\_gyms\_per\_1000}$

deviation_gyms_per_km2 = $\frac{gyms\_per\_km2}{avg\_gyms\_per\_km2}$

3.5 Calculate the final gym index as mean average of the deviations

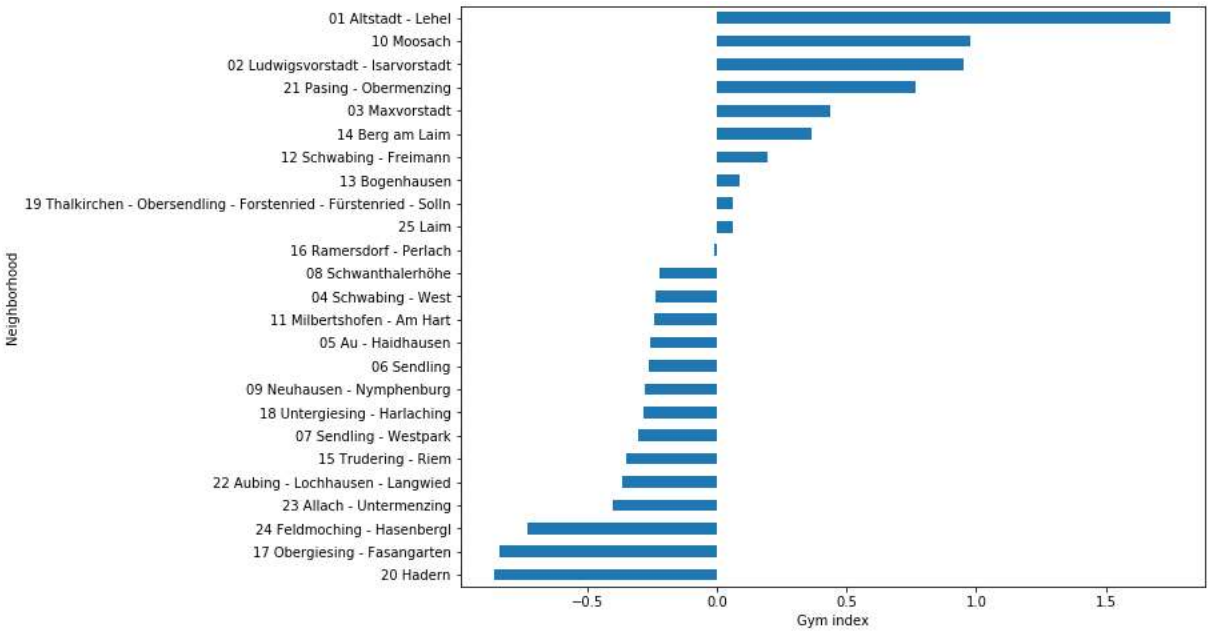
The final key figure ‘Gym index’ is the mean average of the 3 deviations scaled to zero:

gym_index = $\left(\frac{deviation\_gym\_count + deviatio \_gyms\_per\_1000 + deviation\_gyms\_per\_km2}{3}\right) - 1$

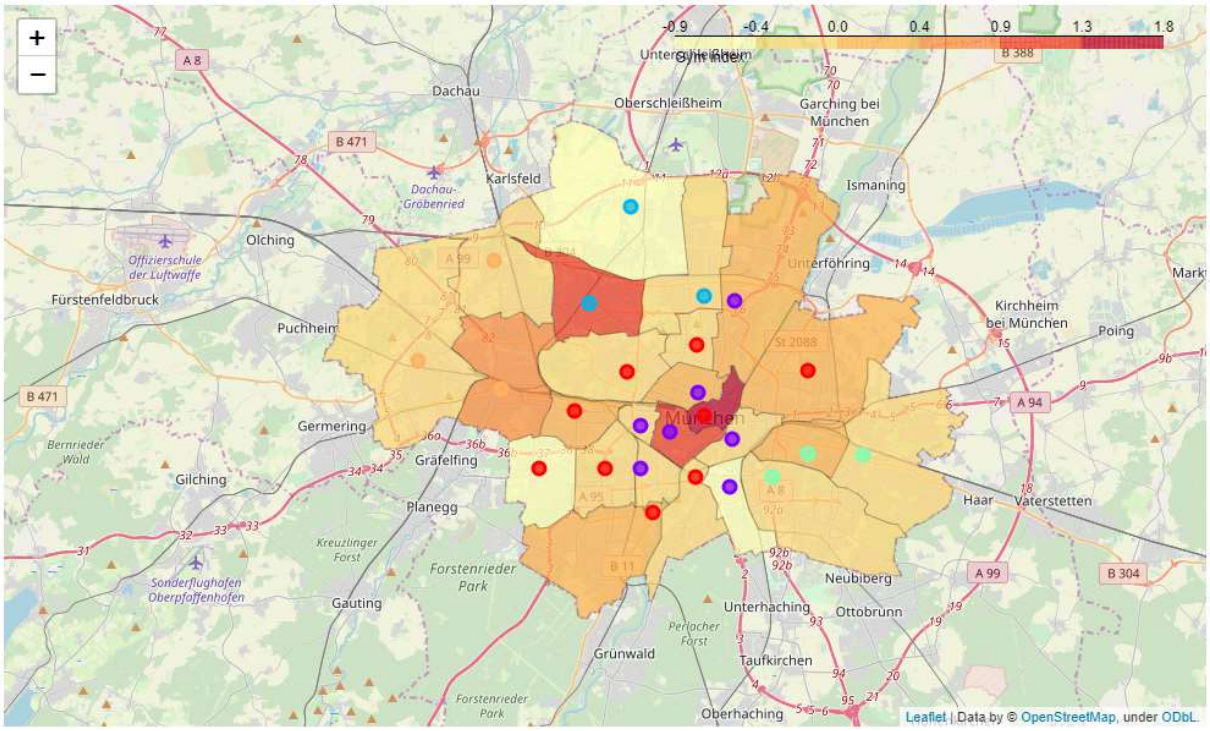
Example Data:

Neighborhood	Label	gym count	gyms per_1000	gyms per_km2	avg gym count	avg gyms per_1000	avg gyms per_km2	gym_index
01 Altstadt - Lehel	0	19	0.907961	6.129032	13.555556	0.256607	1.850335	1.750789
02 Ludwigsvorstadt - Isarvorstadt	1	33	0.639139	7.500000	18.000000	0.337494	3.522278	0.952139
03 Maxvorstadt	1	24	0.467736	5.581395	18.000000	0.337494	3.522278	0.434614
04 Schwabing - West	0	9	0.131839	2.045455	13.555556	0.256607	1.850335	-0.238945
05 Au - Haidhausen	1	13	0.213335	3.095238	18.000000	0.337494	3.522278	-0.255634

A gym index of zero means no deviation to the cluster average. Values less than zero indicate Neighborhoods with less gym density than average and values greater zero indicate Neighborhoods with higher gym density than average. So negative values showing high market potentials for fitness center:



This can also be show in a choropleth map:

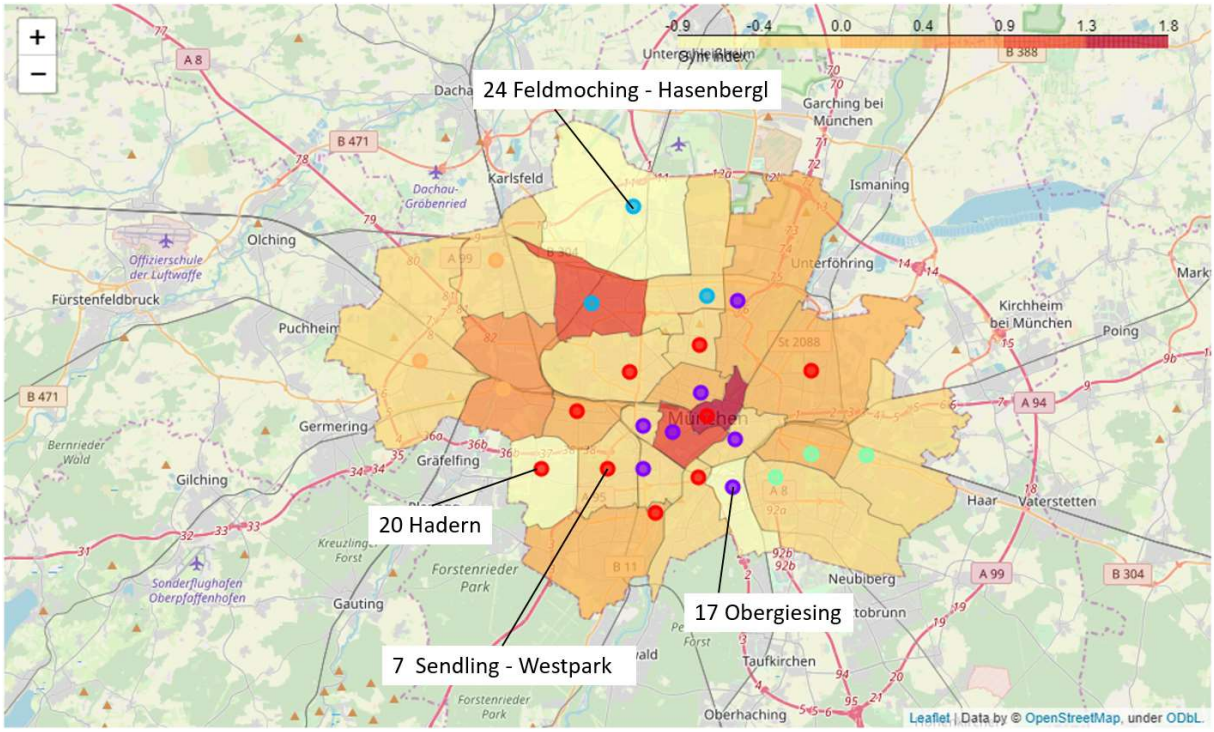


4. Results and Discussion

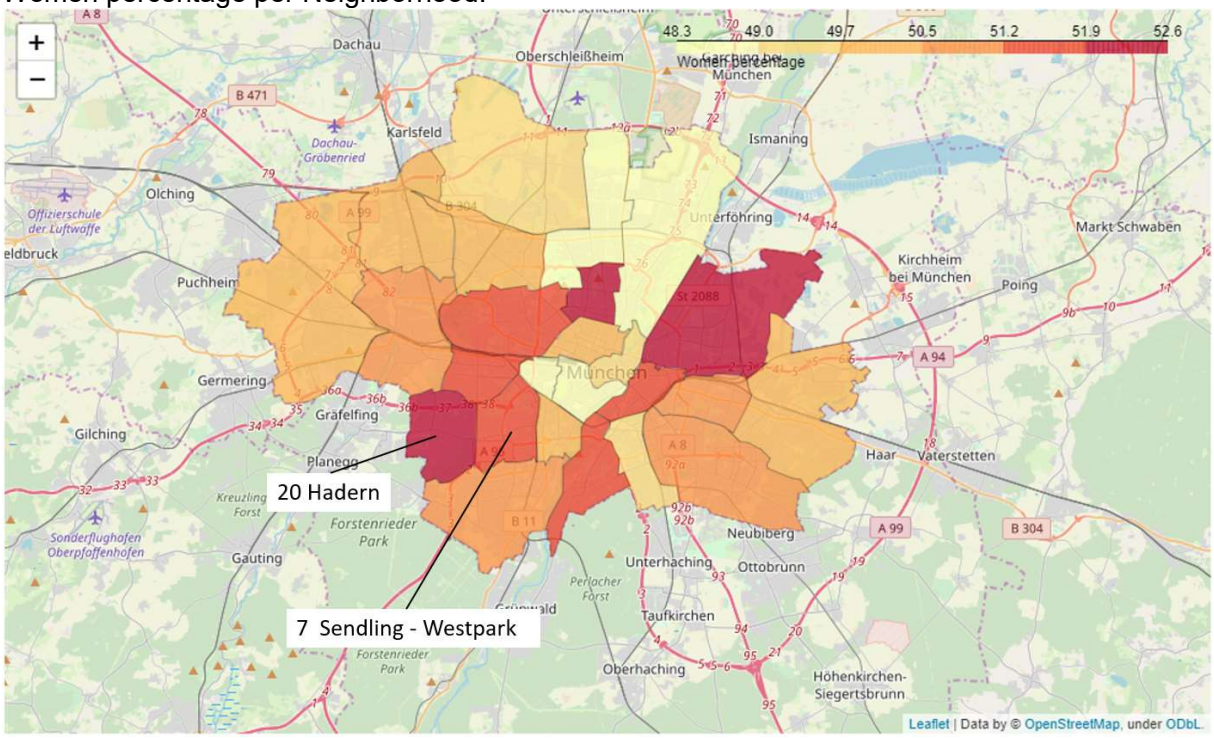
Based on the final gym Index Neighborhoods with high market potential can be identified easily. Also interesting is the location of such Neighborhoods given by the map. The first 6 Neighborhoods with the highest potential are at the outskirts. Let's have a look at the top 10 most interesting Neighborhoods with all their key figures including the distance to the city center. Distance to the city center could also be an important additional decision criteria.

Neighborhood	Popu-lation	Pop. per_km2	Gym count	Women gym	Gyms per1000	Gyms per km2	Women %	Gym index	Distance (km)
20 Hadern	49626	5380	2	0	0.0403	0.2173	52.5	-0.8593	7.2
17 Obergiesing - Fasangarten	53937	9429	3	0	0.0556	0.5263	49.7	-0.8397	3.0
24 Feldmoching - Hasenbergl	60933	2106	2	0	0.0328	0.0692	49.6	-0.7328	9.3
23 Allach - Untermenzing	32677	2115	2	0	0.0612	0.1290	49.8	-0.4005	10.9
22 Aubing - Lochhausen - Langwied	46385	1362	3	0	0.0646	0.0879	50.1	-0.3673	12.2
15 Trudering - Riem	72006	3207	11	0	0.1527	0.4888	49.9	-0.3518	6.7
07 Sendling - Westpark	59386	7599	10	0	0.1683	1.2820	51.3	-0.3044	4.6
18 Untergiesing - Harlaching	52600	6529	10	0	0.1901	1.2345	51.2	-0.2847	2.4
09 Neuhausen - Nymphenburg	98520	7629	14	1	0.1421	1.0852	51.8	-0.2756	3.7
06 Sendling	40682	10329	11	2	0.2703	2.8205	49.9	-0.2623	3.4

District ***20 Hadern*** has the highest market potential according to the gym index.
Neighborhood ***17 Obergiesing - Fasangarten*** has almost the same market potential but is only about 3 kilometers away from city center what makes this Neighborhood also very interesting.
7 Sendling - Westpark is the first inner city Neighborhood with a good market potential and would be also a choice for a new opening under such restrictions.



Women percentage per Neighborhood:



When we have a look at women percentage from the OpenData we can see that Neighborhood ***20 Hadern*** has almost the highest share of women but no Women only Fitness Center. ***07 Sendling-Westpark*** has also a quite high women percentage and also no women only gym. Both neighborhoods have a high market potential and would be very interesting for a women only fitness center opening.

4. Conclusion

The calculated *gym index* together with the OpenData about the Munich Population gives a very good basis for decision making of the stakeholders. Within the selected Neighborhood a detailed further analysis on possible exact locations on street level is necessary. Therefore the location map of all gyms in Munich from Part *2.2. FOURSQUARE search* is also a good starting point to find a building for the opening of a new fitness center. If we would concentrate on ***17 Obergiesing - Fasangarten*** the map for decision making would look like this: