**Capstone Project Report:**

Battle of the Neighborhoods

**Applied Data Science Capstone by IBM/Coursera**

[**Capstone Project Report:** 1](#_Toc15216360)

[Battle of the Neighborhoods 1](#_Toc15216361)

[**Applied Data Science Capstone by IBM/Coursera** 1](#_Toc15216362)

[**1. Introduction / Business Problem** 1](#_Toc15216363)

[**2. Data** 2](#_Toc15216364)

[**2.1 Geo-Data for the neighborhoods in Munich** 2](#_Toc15216365)

[**2.2 FOURSQUARE Data of fitness centers in Munich** 3](#_Toc15216366)

[**2.3 Munich Population Statistics - OpenData** 5](#_Toc15216367)

[**3. Methodology** 6](#_Toc15216368)

**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**

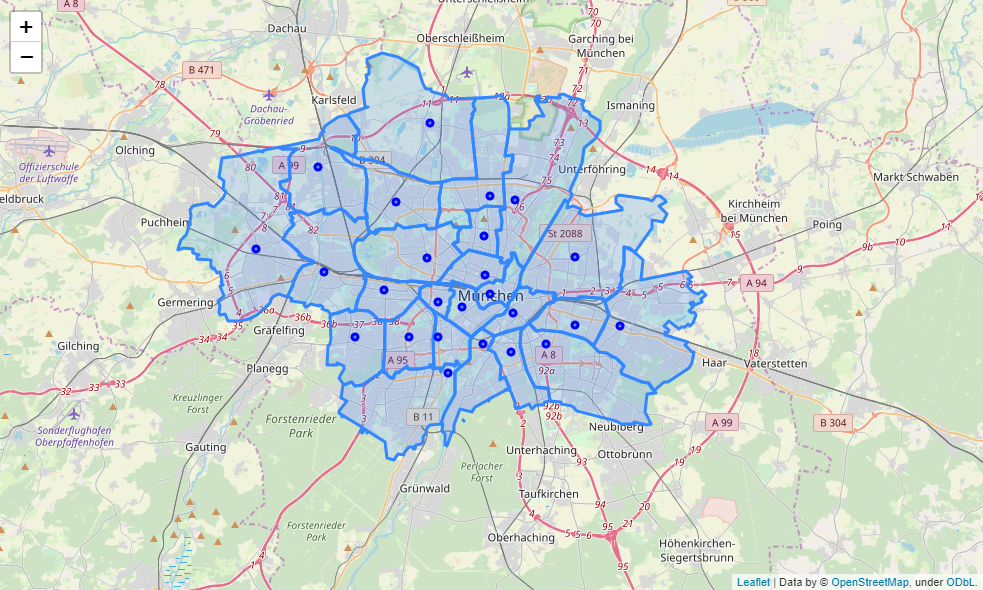
**2.1 Geo-Data for the neighborhoods in Munich**

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:



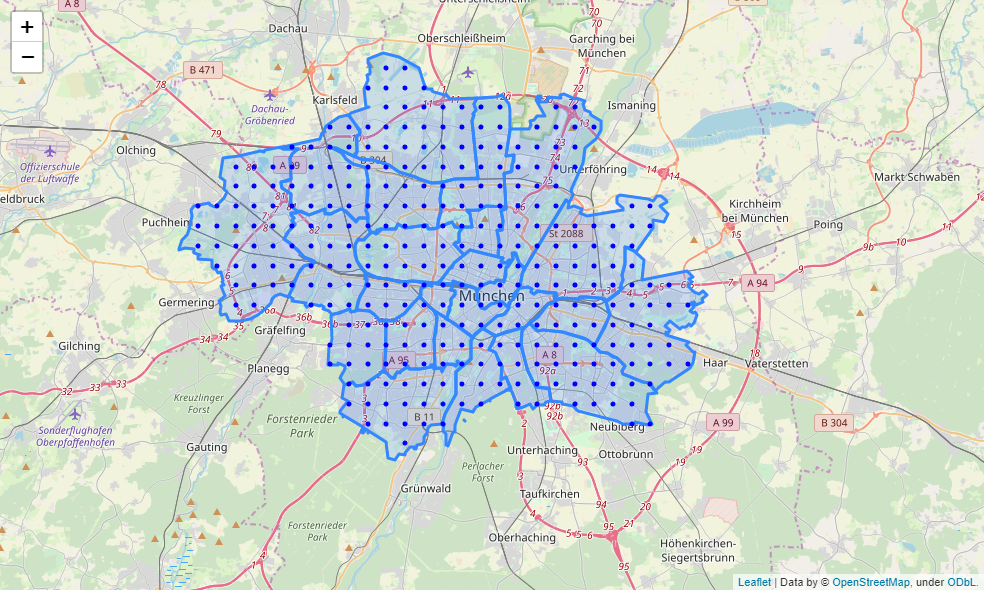
Map of Neighborhoods in Munich:



**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.

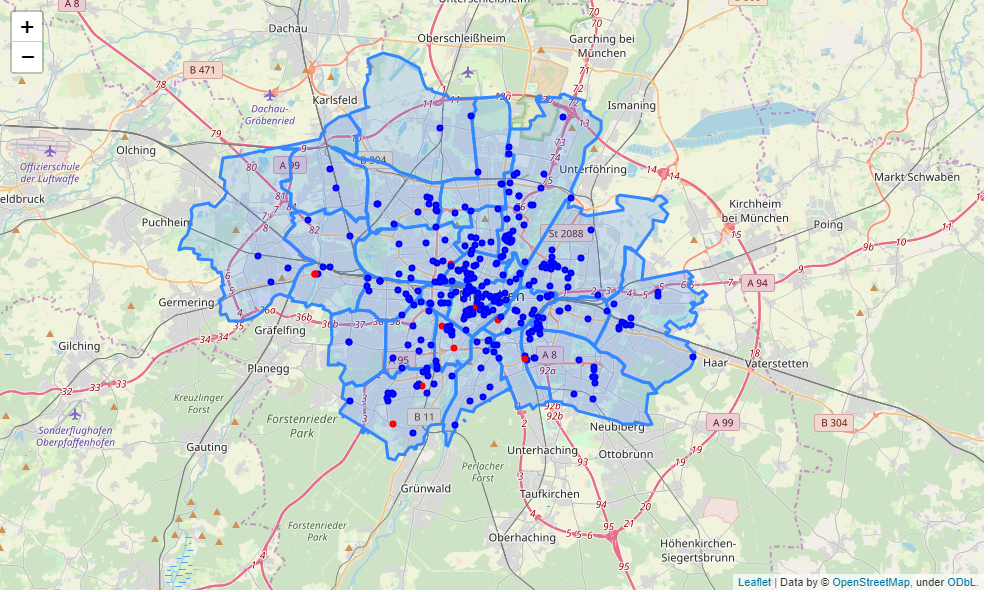
**Data cleaning:**

* 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:



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.

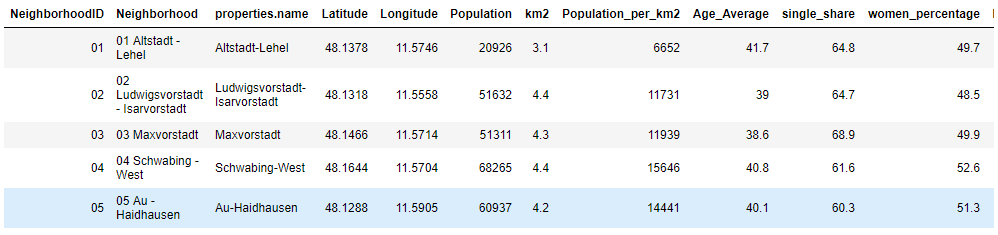
The different datasets are provided by the muncipality in csv format. The following datasets are loaded:

1. [Population and Population density](https://www.opengov-muenchen.de/dataset/0be6dc92-9ca5-4ae9-8a08-ba4039f2a225/resource/3f4aea4c-a79a-4f5b-ab01-a6ad540449f0/download/indikatorenatlas1812bevoelkerungeinwohnerdichte.csv)
2. [Average age](https://www.opengov-muenchen.de/dataset/9b072087-6098-4558-b221-01e7d119b512/resource/9177f886-7341-49d2-9bfa-15d4d8631327/download/indikatorenatlas1812bevoelkerungaltersdurchschnitt.csv)
3. [Single-Person Households](https://www.opengov-muenchen.de/dataset/86612b72-4312-4905-a0c5-581baac650c5/resource/bfedaa2b-e096-40b4-a48a-51faff42adf3/download/indikatorenatlas1812bevoelkerungeinpersonenhaushalte.csv)
4. [Women percentage](https://www.opengov-muenchen.de/dataset/c0ef5f11-fa60-4f5f-95cb-90ae18c8f9a6/resource/f67b2cf6-a617-4ccb-8d56-65fe14a6426b/download/indikatorenatlas1812bevoelkerungfrauenanteil.csv)
5. [Birthrate](https://www.opengov-muenchen.de/dataset/bbeafa23-2c03-4fca-9b2e-c4b78c4ff850/resource/1ec077d4-93f1-4b64-b2a4-663a55cd7c57/download/indikatorenatlas1812bevoelkerungrohegeburtenrate.csv)
6. [Aging index](https://www.opengov-muenchen.de/dataset/e46c1590-6000-4556-b66c-62faf5a61048/resource/fe9aeef9-3479-407c-b38f-db73a4dadf9f/download/indikatorenatlas1812bevoelkerungueberalterungsquotient.csv)
7. [Youth index](https://www.opengov-muenchen.de/dataset/754aefa4-2e36-4e0c-b768-fa9b7d003f45/resource/81f2cede-f395-4327-aeb5-24edcdbb2bee/download/indikatorenatlas1812bevoelkerungjugendquotient.csv)
8. [Unemployment rate](https://www.opengov-muenchen.de/dataset/0771b22f-b1e7-4480-8fe5-d641c2586f3e/resource/55a1adb0-6c06-403b-ab7f-7e6c41d18e32/download/indikatorenatlas1812arbeitsmarktarbeitslosendichte.csv)
9. [Welfare recipients rate](https://www.opengov-muenchen.de/dataset/103f72fa-f48a-4b7d-9402-e83fc291685e/resource/921072cb-3ed2-4624-aa59-5e13a3825033/download/indikatorenatlas1812arbeitsmarktalgii-empfaengerdichteerwerbsfaehigeregelleistungsberechtigtendi.csv)
10. [Migration background percentage](https://www.opengov-muenchen.de/dataset/cc5e9a41-6f29-458a-b8e1-103da957a333/resource/f98bf568-c84d-4f9f-b3d3-ebc985933f7b/download/indikatorenatlas1812bevoelkerungmigrationshintergrundquote.csv)

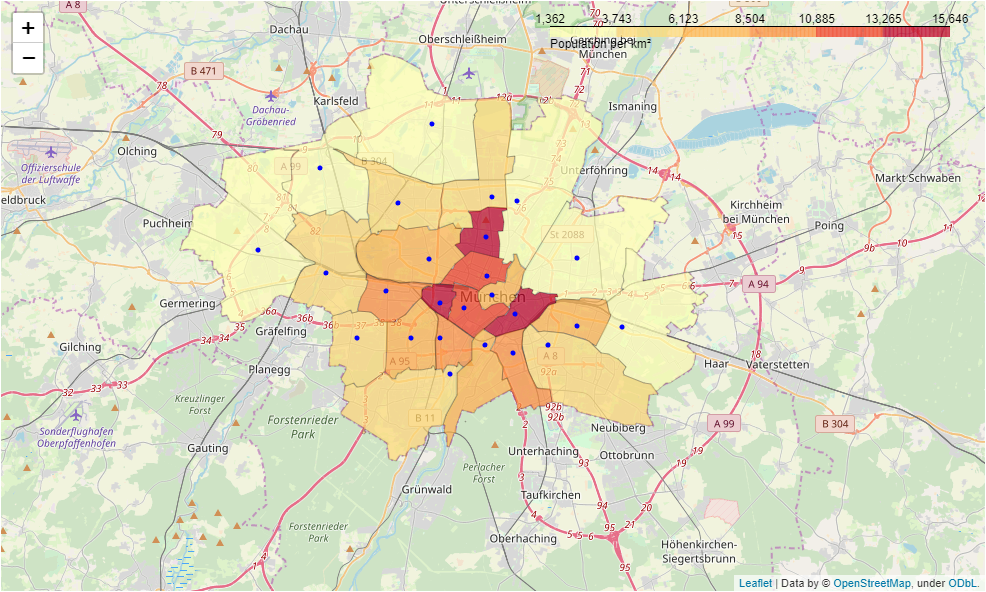
Always the latest available year of data is used (2017 or 2016).

The different files from [OpenData Munich](https://www.opengov-muenchen.de/dataset" \t "_blank) 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:



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:

Gym count: from FOURSQUARE venue search

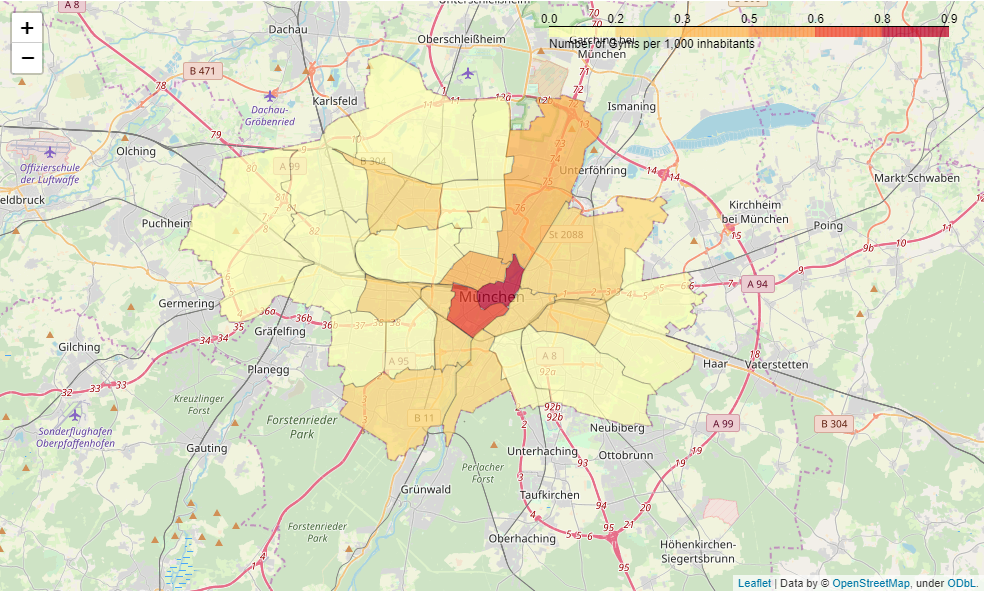
Gmys per 1,000 inhabitants:

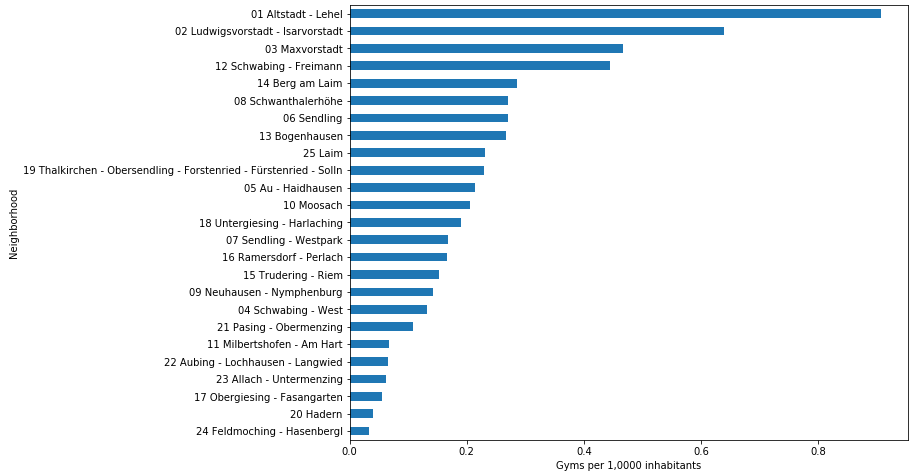
Gmys per Area km²:

Example values:

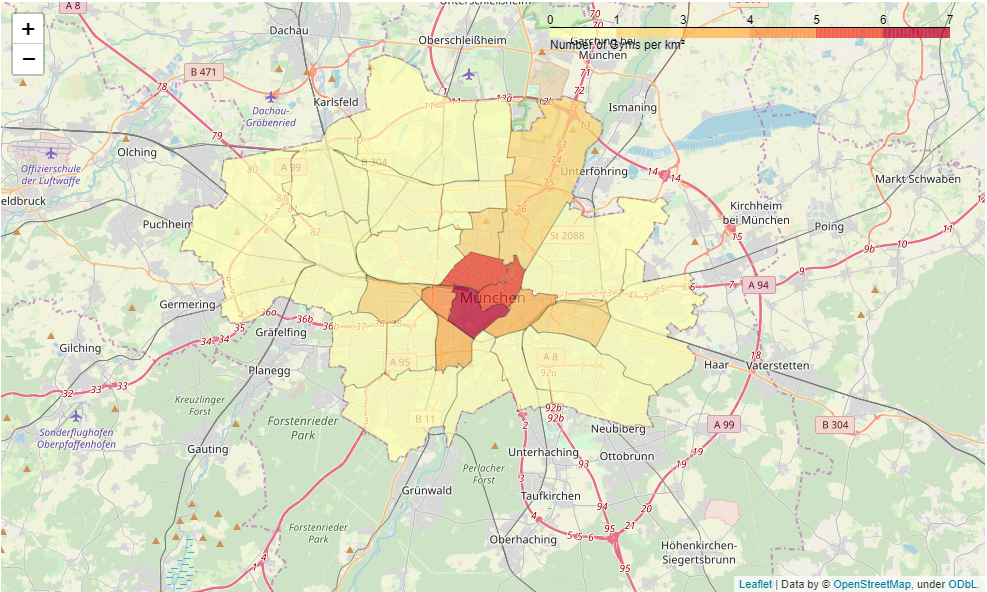
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Neighborhood | Pop ulation | km2 | Population \_per\_km2 | Gym count | Gyms per\_1000 | Gyms per\_km2 |
| 01 Altstadt - Lehel | 20926 | 3.1 | 6652 | 19 | 0.907961 | 6.129032 |
| 02 Ludwigsvorstadt - Isarvorstadt | 51632 | 4.4 | 11731 | 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:



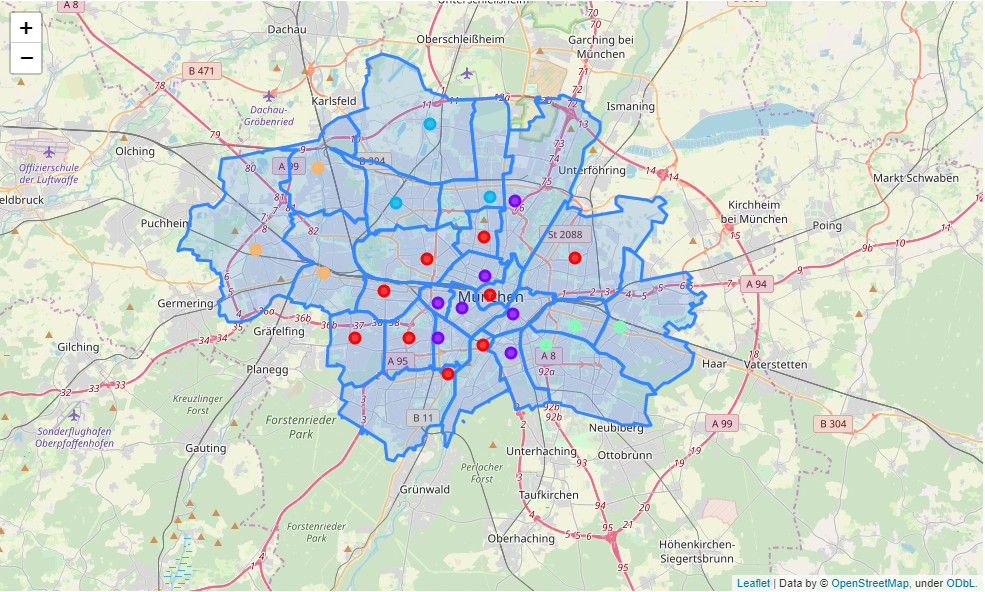


Number of Gyms per km²:



**3.2 Clustering**

Based on the 10 Population key figures (e.g. birthrate, average age, welfare recipients 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 =

deviation\_gyms\_per\_1000 =

deviation\_gyms\_per\_km2 =

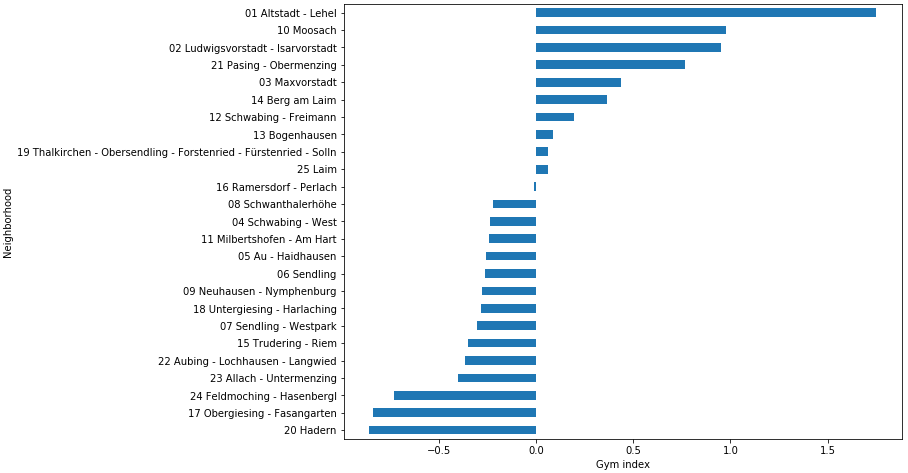
**3.5 Calculate the final gym index as mean average of the deviations**

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

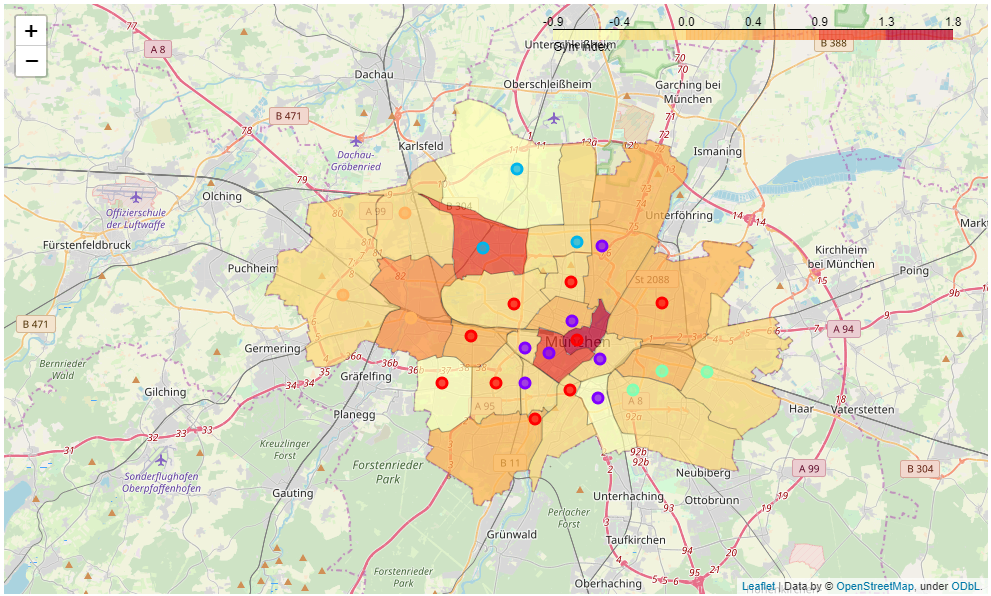
Example Data:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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.