Yoga classes in Munich

Taras Slipets

March, 2020

1. Introduction

1.1. Background

Nowadays working at an office or home office is one of the most frequent setups. In most cases such type of work requires seating for multiple hours a day, which brings very unhealthy static tension to the body.
There are multiple types of sports activities to eliminate and compensate for such bad influences, one of very popular is yoga. Yoga implies a combination of physical exercises with mental relaxing which also helps to decrease overall stress. Moreover most yoga exercises do not require extreme physical pressure and are quite safe from injury prospective.
All that said, yoga is quite a nice choice for office workers who live in big urban areas.

1.2. Problem

Thus, building a recommendation system for finding the best suitable yoga class for office workers based on certain criteria is a valuable analytical problem that perfectly fits into Clustering type of Data Science problems which could be solved by unsupervised learning algorithms.

1.3. Interest

Described problem and its analysis will be the most interest of 2 categories of people:

- people who would like to attend yoga classes and interested in most convenient and cost-efficient offers on the market
- people who would like to start or extend their yoga classes business by opening new locations or adjusting their existing services according to changing market demands

2. Data acquisition and cleaning

2.1. Data sources

Geo-data about Munich boroughs has been taken from Wikipedia and surface and population resources. Afterwards manually aggregated into CSV-file.
To obtain information about Yoga Studios in each borough Foursquare API has been used. Specially useful are following endpoints:

- https://developer.foursquare.com/docs/api/venues/search
- https://developer.foursquare.com/docs/api/venues/details

To narrow search results to Yoga classes Venues only we use Yoga Studio (categoryId = 4bf58dd8d48988d102941735) from available API categories values.
Total number of venues in all boroughs after collection is truncated to no more than 100 rows.

2.2. Data cleaning and Feature selection

Raw JSON data about Venues retrieved from Foursquare API should be filtered to the following structure:

- Foursquare ID
- Name
- Geo-location:
- Latitude
- Longitude
- Contacts:
- Phone
- Website
- Facebook
- Twitter
- Instagram
- Opening hours
- Rating

Mentioned structure is then populated with prices information manually to the best of researcher's effort. Populated data is then flattened and one-hot encoded to generate feature-file for K-Means Clustering algorithm to determine main types of offered Yoga classes in Munich (e.g. far from city center, but cheap; popular in the city center, etc.)
Most records do not contain information about working time, and no information about prices.
Thus, this information is collected manually and added to the table with Yoga studios data.
Working hours are collected manually and with one-hot-encoding right away:

- if studio works at 08:00-10:00 AM, Mon-Fri, than feature is 1, otherwise 0
- if studio works at 18:00-22:00 PM, Mon-Fri, than feature is 1, otherwise 0
- if studio works on weekend Sat-Sun, than feature is 1, otherwise 0

Price information is divided into 3 columns:

- “trial lesson” is 1 if free one-time trial lesson is available
- “single lesson price” - cost of one single lesson in EUR
- “month price” - cost of month abonnement in EUR

Another important detail - only collected info from websites, so Yoga studios which do not have websites are excluded from the resulting list of studios for analysis.
After some more detailed data analysis, filtering and preparation for k-means clustering, the decision has been made to retain only following features:

- distance to center
- morning work
- evening work
- weekend work
- month price

Beside that magnitude normalisation for numeric features has been performed to avoid weight bias in the model of certain features.

3. Exploratory Data Analysis

Here is the map of borough centers locations (Figure 1) by their latitude and longitude:

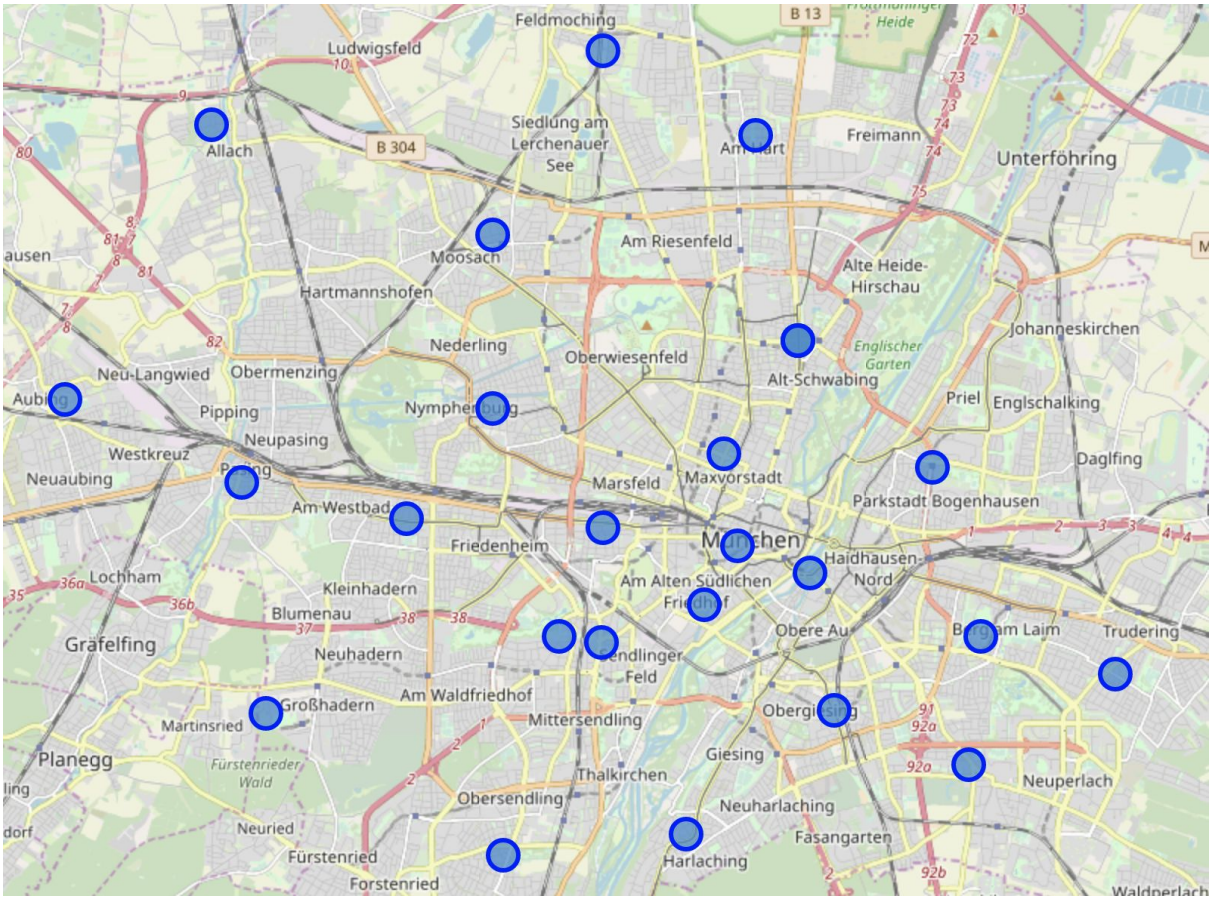

Figure 1. Munich boroughs centers locations

This information has been used to calculate approximate distance from city center to each of yoga classes based on their location.
For actual distance calculation Haversine formula has been used (1):

$$d = 2r \arcsin\left(\sqrt{\operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)}\right)$$
$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

(1)

After several rounds of running K-means clustering algorithm and semantical interpretation of obtained groups the final number of 10 clusters has been chosen.
Clustered groups has been labeled with values from 0 to 9 and here are their semantic meaning:

- 0 - most expensive classes in the city center
- 1 - quite expensive classes close to the city center, work in mornings and on weekends
- 2 - average price classes that work over the weekend
- 3 - yoga classes in far-far away galaxy :)
- 4 - affordable classes in the city center only during work time
- 5 - classes without price information only during work time
- 6 - classes far from city center, but working in mornings and on weekends
- 7 - affordable classes in the city center working in mornings and on weekends
- 8 - classes without price information but working on weekends
- 9 - classes quite distanciated from center, but with average price

Contribution of each cluster to total number of classes (Figure 2):

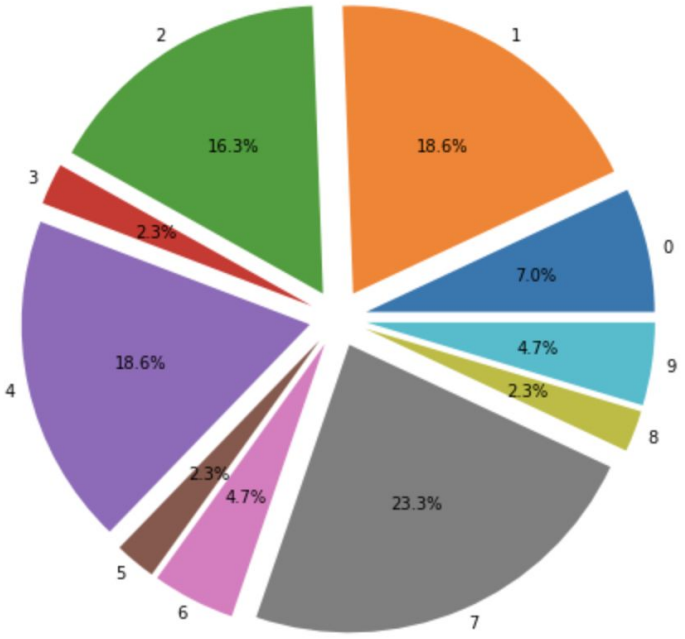

Figure 2. Clusters percentage

Percentage distribution of classes per borough (Figure 3):

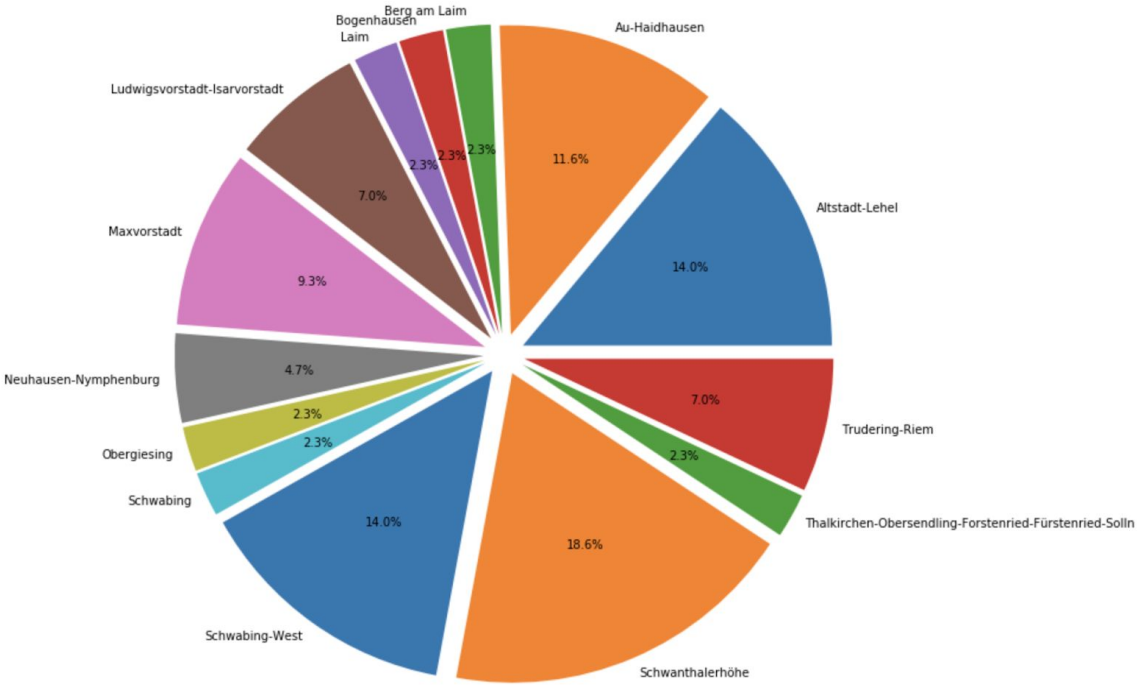

Figure 3. Distribution of yoga classes per borough in Munich

Based on information about population in each borough, people per yoga classes each borough average has been calculated and depicted on Figure 4:

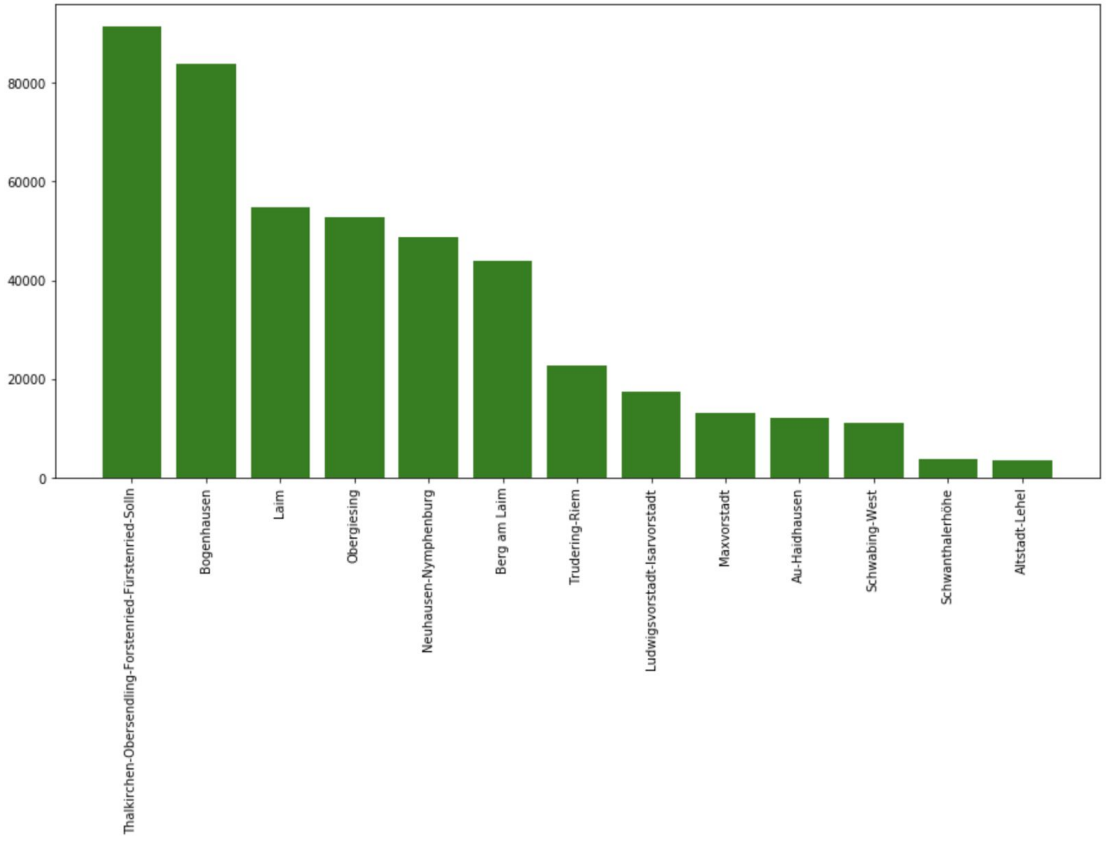

Figure 4. Average number of people per yoga classes in each borough

For city's boroughs with no online information about yoga classes let's depict overall population (Figure 5):

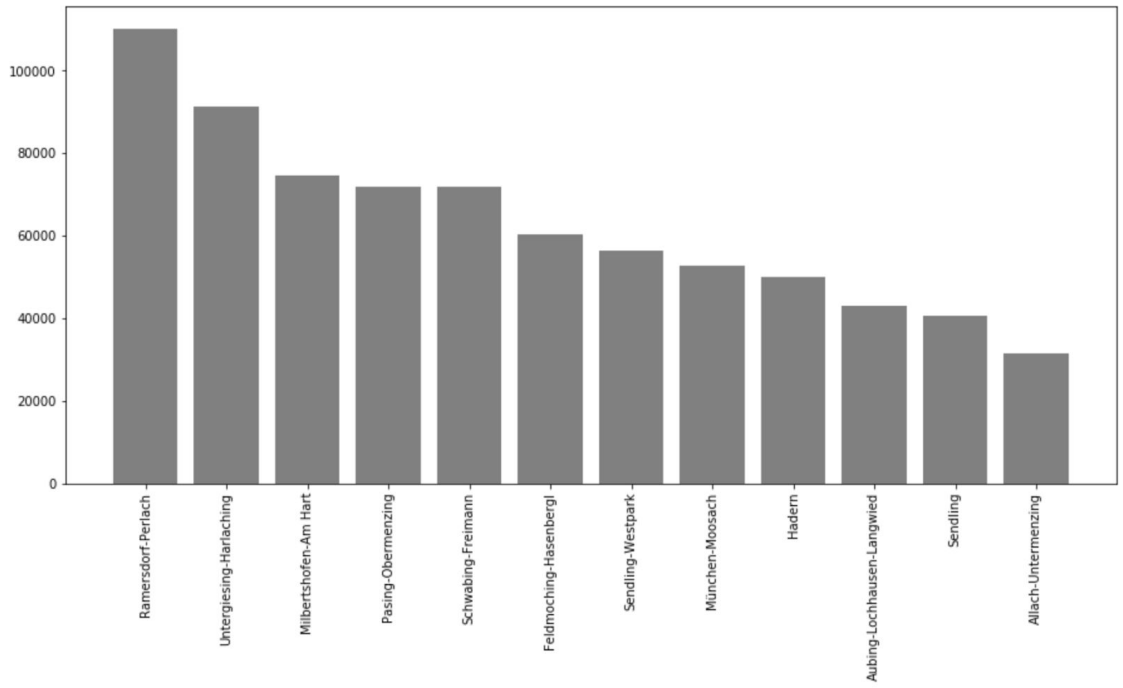

Figure 5. Per borough population for areas with no online information about yoga classes

4. Conclusions

Based on clusters interpretation, described in details in "Results" section, we can conclude, that 3 most significant numbers are:

- affordable classes in the city center working in mornings and on weekends (23.3%)
- quite expensive classes close to the city center, work in mornings and on weekends (18.6%)
- affordable classes in the city center only during work time (18.6%)

Thus, top-3 clusters unambiguously confirm, that there are 2 most important aspects: *distance to the city-center* and *working in mornings and on weekends.*

Next, let's summarise per-boroughs analysis results. Most crowded boroughs are:

- Thalkirchen-Obersendling-Forstenried-Fürstenried-Solln
- Bogenhausen
- Laim

Good boroughs to someone, who consider conducting yoga classes, because demand there is highest.

On contrary, from attendees prospective, 3 least crowded boroughs are:

- Altstadt-Lehel
- Schwanthalerhöhe
- Schwabing-West

where you have least visitors per class, so attendance should be comfortable.
For boroughs, which do not have online information about yoga classes we can only recommend it to someone, who plans to *conduct* yoga classes.
Top-3 candidates to consider are:

- Ramersdorf-Perlach
- Untergiesing-Harlaching
- Milbertshofen-Am Hart

Beside that, consider that there is very high demand for *morning and weekend working hours*.

5. Future directions

With no doubt one of the main challenges of this project is collecting information about actual yoga classes and their locations. Thus, increasing the size of the input sample would be one of the main follow up steps to improve statistical quality of analysed dataset.
Another reasonable variation would be comparison of conducted research with other clustering algorithms applied to collected dataset.