Setting up Gym in Munich - Report

Own Capstone Project – as part of Applied Data Science Capstone Course

Introduction

The purpose of this project is to find a suitable neighborhood for opening up a new gym in the city of Munich (München), Germany, which is also a city where I live. With 1.5 mil inhabitants, Munich is the 3rd largest city in Germany. It contains 25 neighbhorhoods ("Stadtbezirke" in German).

To solve this problem, I will combine three different data sources:

- the data from Foursquare API
- demographical data
- the average private property rent accross neighborhoods

The subject interested in this analysis would be a subject wanting to open up a gym.

Data sources and extraction

Foursquare API

I carried out the data extraction and used the codes as shown in the course *Applied Data Science with Python*. Later on I only filtered out the gym venues to see how many there are in each neighborhood.

Demographical Data

The demographical data is extracted from the statistical website of Munich city coucil.

Website link: http://www.mstatistik-muenchen.de/indikatorenatlas/export/export.htm

File link: http://www.mstatistik-muenchen.de/indikatorenatlas/export/mv_export_csv_ab2010.xlsx

It contains all sorts of demographic data from the years 2010 - 2017. From there I also extracted the neighborhood names. The data is in German but I renamed (translated into English) the headings of the columns that I keep.

Rental Data

I decided to add private property rental prices to my analysis as it reflects the affluence of the neighborhood. Areas with high rent are more expensive to run a gym in but also have greater purchase power. Depending of what we want to offer, more expensive neigbhorhoods can afford luxury gyms with a variety of services, whether cheaper areas are more likely to accomodate a simple gym with lower membership fees.

The data comes from a website of a real estate agent, where it is shown in the second table: https://suedbayerische-immobilien.de/Mietpreise-Muenchen-Stadtteile . I extracted the data using Beautiful Soup.

Methodology

Obtaining neighborhood names and coordinates

First I extracted the neighborhood names from the Munich council statistical dataset. Here the names are shown with their official name. Other sources might use different spelling or area separation, but here the official administrative way is shown so I will adhere to it when I review and rename my other data sources.

I turned the names data into a list and used it to obtain the coordinates using the Geopy Nominatim library. Although this library is not optimal for bulk work, as there are only 25 neighborhoods in Munich, this way worked perfectly. I prefer this libary over the Geocoder library, which seems to be very slow and sometimes fail to obtain the coordinates at all.

After obtaining the coordinates in form of a list, I added all three lists together to form a new data frame.

I visualised the neighborhoods using this data frame and the Folium library.

Exploring neighborhoods in Munich using Foursquare API

I used my Foursquare creadentials and applied the fuction *getNearbyVenues* as mentioned in the Coursera course (in the New York Clustering assignment) to download the json file with all venues in each neighbhorhood and transforme it into a data frame. A street name of each venue has also been added. From the list of all unique venue categories I selected those related to a fitness center/gym and filtered them out in the data frame. This way I obtained a data frame with all gyms across neighborhoods.

When comparing the number of lines in this ‚gym' data frame with the number of unique gyms, I could see that many lines have been duplicated. That is quite presumably so, as the area size of each neighborhood vary. As the radius has been set to 2500 meters to cover even the largest neighborhoods, some venues are then captured more times in the data frame, each time in a different neighborhooring area. To amend this, I applied the the Geopy Nominatim library again to double-check the name of the neighborhood name based on the venue street as taken from the Foursquare file. I then added these corrected names to my data frame and worked with them to obtain a grouped by data: number of gyms in each neighborhood.

Reading demographical data

From the demographical data file I used the sheet named BEVÖLKERUNG (Population) and gathered the number of inhabitants of each neighbhorhood. I concentrated on the age group of 15 – 65 years which is presumably the age group where most customers fall into.

Then I calculated the number of inhabitants per gym (under the presumtion an additional gym would be created in the neighborhood).

Collecting rental data

I extracted the table from the website as mentioned in the Data Sources section. Using Beautiful Soup I obtained a list of the table rows and them split them by '\n' into a further level of lists. Then I created an empty dictionary and added a list of neighborhoods as the first column and a list of rent prices as the second column. From the dictionary I created a data frame and adjusted the neighborhoods names using Regex.

In the end, all the data have been merged into one data frame to be handy for the clustering.

K-Means Clustering

I performed the segmentation of the neighborhoods employing K-means clustering algorithm. I chose this method as from the clusters we will be able to see which areas would most likely welcome a new gym and what kind of gym that would be.

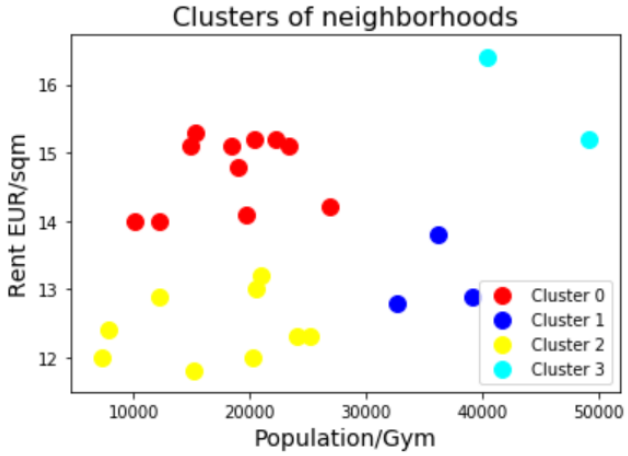
The inputs were the population per gym and the average rent in each neighborhood. I standardized the data to compare the measures better as they were spread over different scales (8k to 50k resp 11.8 to 16.4).

To establish the optimal number of clusters, I used a graphical method called the Elbow method. It is based on Withing Cluster Sum of Squares (WCSS). This is a sum of squared distances between points and the centroid. WCSS is calculated for each option of different number of clusters and then plotted as a line graph. With growing number of clusters the WCSS comes closer to zero. However, at certain point the marginal gain will drop - the „elbow“ point on the graph, which shows the optimal number of clusters.

I ran the K-means algorithm using the scikit learn library. Then I added the cluster labels to the final data frame and plotted it as a scatter plot.

Results

After running the K-Means algortihm, the neighborhoods have been divided into four clusters. They are distinguished by the number of potential customers the new gym could receive and the wealthiness of the area represented by the average rent.



Discussion

The clusters can be characterised as follows:

Cluster 0 - depicts areas with higher rent, so obviously wealthy neighborhoods that cater for luxury gyms with high membership fees, the density is 10 to 30 k potential customers per gym.

Cluster 1 - the neighborhoods here are of medium to lower affluence, with 33 - 40k population per gym density. This seems to be an ideal cluster to target if we want to open up a rather simple gym with average membership fees.

Cluster 2 - shows the neighborhoods with lower rent with the population to gym density ranging from 8 to 27 k. Simple, no-frill gyms with lower membership fees can be found here.

Cluster 3 - the smallest cluster with the highest proportion of potential customers per gym (40 to 50k). Also an area with the top private property rent. This is the cluster to target, as there could be a high demand that has not been saturated yet. These are the high income areas, perfect for setting up an upscale gym.

Conclusion

The data analysis and the K-Means clustering process have shown that are still areas in Munich with unsaturated demand for a gym.

If somebody was to invest into opening a new gym, I would recommend to target the areas in Cluster 1 and 3, as they are quite populous and there is no gym yet. The clusters contain the following 5 neighborhoods:

Area Name	Population 15 - 65 Yrs	Number of Gyms	Potential Customers	Rent EUR/sqm	Cluster Label
Ludwigsvorstadt - Isarvorstadt	40440.0	0.0	40440	16.4	3
Schwabing - West	49129.0	0.0	49129	15.2	3
Obergiesing - Fasangarten	39182.0	0.0	39182	12.9	1
Untergiesing - Harlaching	36244.0	0.0	36244	13.8	1
Hadern	32602.0	0.0	32602	12.8	1

The areas **Ludwigsvorstadt - Isarvorstadt** and **Schwabing - West** are the most affluent with high rent so we can expect that people living there have a high income but running a gym there would be therefore expensive. It would be an area of interest for luxury gym operators who can offer spacious gym sites with additional services such as sauna, pool, massage and child minding facilities.

Untergiesing - Harlaching is an area with an average rent so here I would recommend to open up a moderately expensive gym that could cater for middle-class customers.

The remaining two areas (**Obergiesing - Fasangarten, Hadern**) would represent lower income population. Opening a convenient simple gym with reasonable membership fees would be the best solution here.