Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Fitness Center / Gym in Munich, Germany

By: Michael Gruber

Oktober 2019

Introduction

Today, more and more people are trying to live healthier lives by eating healthier foods or doing more sports. In particular, the phenomenon of "fitness" has become a trend in modern times and attracts more and more people to the fitness center or gym. According to Statista, a leading company for statistical analysis, in 2018 more than 11+ Mio people in Germany are registered in fitness centers / gyms. Due to the high demand for such facilities, it is even more important for studio operators to choose the right location to open a new studio. Opening Fitness Centers allows studio owners to earn consistent monthly income through membership fees. Of course, as with any business decision, opening a new Fitness Center requires serious consideration and is a lot more complicated than it seems. The location plays a crucial role for the success of the fitness center / gym. Therefore it is interesting to investigate what the situation is in a metropole such as Munich, Germany.

Business Problem

The objective of this project is to analyse and select the best locations in the city of Munich, Germany to open up a new Fitness Center / Gym. Using Data Science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: If a property developer is looking to open a new Fitness Center / Gym in Munich, where would you recommend it?

Target Audience of this project

This project is particularly useful to fitness studio owner looking to open a new fitness Center / gym in the cities of Germany, eg. Munich. This project is timely, as the city has already a wide range of fitness centers or gyms. However, regarding the statistics from Statista, the fitness hype still persists and the amount of fitness memberships is increasing every year. In fact there are more fitness memberships than eg. memberships for football clubs. But also the number of studios increased by 3.9% to more than 9300 fitness centers / gyms in Germany. These facts show that the fitness industry is still booming and owner of fitness centers could profit from opening a new studio.

Data

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

1. A list of boroughs in Munich. This defines the scope of this project which is confined to the city of Munich.

2. Latitude and Longitude coordinates of those boroughs. This is required in order to plot the map and also get the venue data.

3. Venue data, particularly data related to Fitness Centers and gyms. We will use this data to perform clustering on the boroughs.

Sources of data and methods to extract them

This Wikipedia page (https://de.wikipedia.org/wiki/Liste_der_Stadtteile_M%C3%BCnchens) contains a list of boroughs in Munich, with a total of 56 boroughs. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the boroughs using Python Geocoder package which will give us the latitude and longitude coordinates of the boroughs.

After that, we will use Foursquare API to get the venue data for those boroughs. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Fitness Center / Gym category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of boroughs in the city of Munich. Fortunately, the list is available on Wikipedia (https://de.wikipedia.org/wiki/Liste_der_Stadtteile_M%C3%BCnchens)
We will do web scraping using Python requests and beautifulsoup packages to extract the list of boroughs data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the boroughs in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Munich.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the boroughs in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each borough and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each borough by grouping the rows by borough and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Fitness Center / Gym" data, we will filter the "Gym / Fitness Center" and "Gym" as venue categories for the boroughs. Due to the two categories we then merge the results by adding the value of each frequency of occurrence and form an extra column called "Fitness".

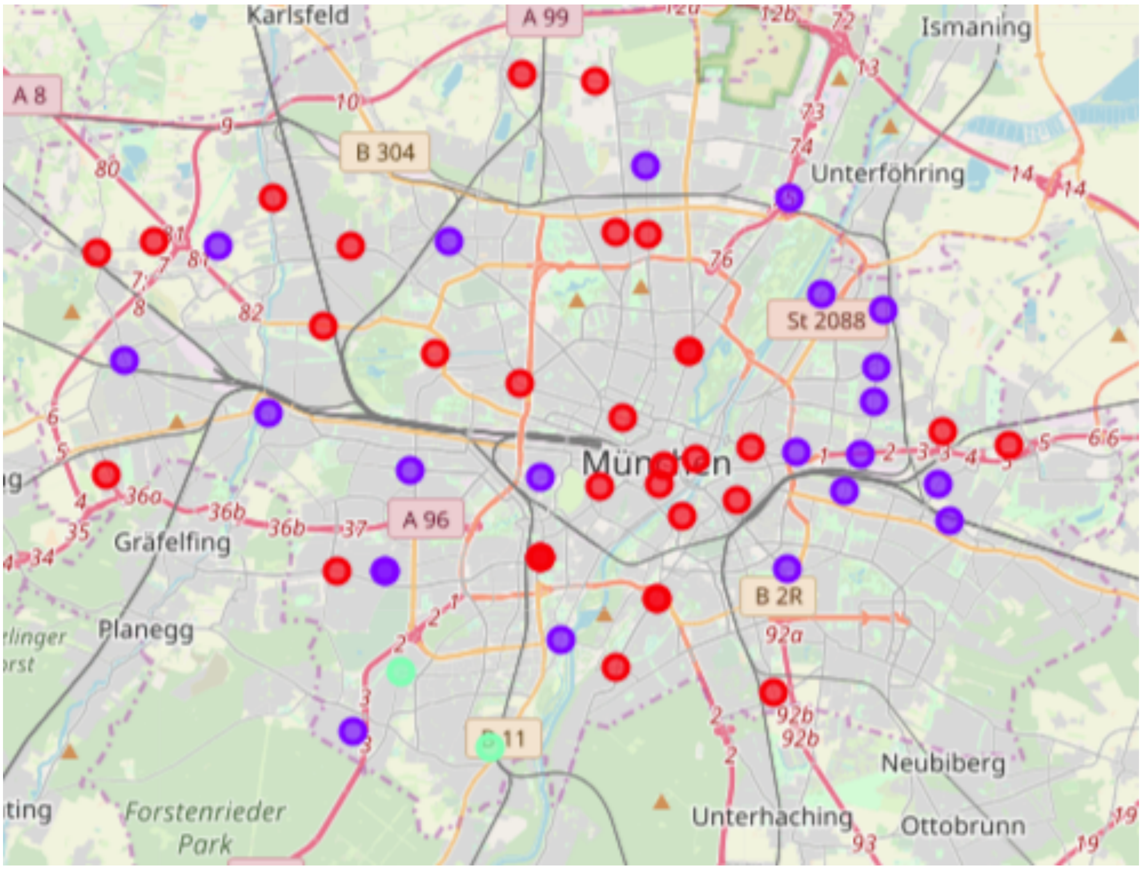
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the boroughs into 3 clusters based on their frequency of occurrence for "Fitness". The results will allow us to identify which boroughs have higher concentration of fitness centers / gyms while which boroughs have fewer number of fitness centers / gyms. Based on the occurrence of fitness centers / gyms in different boroughss, it will help us to answer the question as to which boroughs are most suitable to open new fitness centers / gyms.

Results

The results from the k-means clustering show that we can categorize the borougs into 3 clusters based on the frequency of occurrence for “Fitness”:

- Cluster 0: Boroughs with low to no existence of number of fitness centers / gyms
- Cluster 1: Boroughs with moderate number fitness centers / gyms
- Cluster 2: Boroughs with high concentration of fitness centers / gyms

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

As observations noted from the map in the results section, most of the fitness centers / gyms are concentrated in the surrounding area of Munich city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no fitness centers in the boroughs closed to the city center. This represents a great opportunity and high potential areas to open new gyms as there is very little to no competition from existing fitness centers. Meanwhile, fitness centers in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of gyms. From another perspective, the results also show that the oversupply of gyms mostly happened in the surrounding area of the city, with the center area still have very few gyms. Therefore, this project recommends studio operators to capitalize on these findings to open new fitness centers in boroughs in cluster 0 with little to no competition. Studio operators with unique selling propositions to stand out from the competition can also open new gyms in boroughs in cluster 1 with moderate competition. Lastly, studio operators are advised to avoid boroughs in cluster 2 which already have high concentration of gyms and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of fitness centers / gyms, there are other factors such as population and income of residents that could influence the location decision of a new fitness center / gym. However, to the best knowledge of this researcher such data are not available to the borough level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new fitness center / gym. In addition, this project made use ot he free Sandbox Tier Account of Foursquare API that came with limitations as ot he number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. studio operators regarding the best locations to open a new fitness center / gym. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The boroughs in cluster 0 are the most preferred locations to open a new fitness center / gym. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new fitness center.

References

Wikipedia (2019, October 20): Liste der Stadtteile Münchens. Retrieved from https://de.wikipedia.org/wiki/Liste_der_Stadtteile_M%C3%BCnchens

Foursquare (2019, October 20): Foursquare Developers Documentation. Retrieved from https://developer.foursquare.com/docs

Statista (2019, October 20): Mitgliederzahl der Fitnessstudios in Deutschland von 2003 bis 2018 (in Millionen). Retrieved from https://de.statista.com/statistik/daten/studie/5966/umfrage/mitglieder-der-deutschen-fitnessclubs/

Pressportal (2019, October 20): Eckdaten der deutschen Fitness-Wirtschaft 2019 Über 11 Millionen Mitglieder in Fitness- und Gesundheits-Anlagen. Retrieved from https://www.presseportal.de/pm/70906/4221602

Appendix

Cluster 0

Allach
Altstadt
Am Riesenfeld
Au
Bogenhausen
Daglfing
Fasangarten
Feldmoching
Freiham
Ludwigsvorstadt
Maxvorstadt
Milbertshofen
Neuhausen
Nymphenburg
Obermenzing
Riem
Schwabing
Sendling
Untermenzing
Giesing
Hadern
Haidhausen
Harlaching
Hasenbergl
Isarvorstadt
Langwied
Lehel
Lochhausen

Cluster 1

Am Hart
Am Moosfeld
Aubing
Berg am Laim
Denning
Englschalking
Forstenried
Freimann
Holzapfelkreuth
Johanneskirchen
Laim
Moosach
Oberföhring
Pasing
Perlach
Ramersdorf
Schwanthalerhöhe
Steinhausen
Thalkirchen
Trudering
Zamdorf

Cluster 2

Fürstenried
Solln