Capstone Project - Dance Studio in Budapest

By Gabor Nemeth

May, 2020

Business Problem and Background

This project intends to identify the optimal place for a dance studio in Budapest. The topic "dance" comes from my brother being a dance teacher, who used to be on the hunt for a place to open his own studio some years back. At that time, I was not even close to data science, but moving the situation to present days in theory, I wonder if/how I and data could help him make the close to best decision.

So, my stakeholder is my onetime brother, but the results of my analysis could be interesting to anyone having similar plans in the capital of Hungary nowadays (well, after the current COVID situation is over, as dancing is a very contact-intensive hobby).

According to my brother's information, the competition has always been "strong" in the dancing business. He claims there were and still are many such studios in Budapest compared to the size of the city (population). I am going to try and check if data confirms his statement, and will be searching for areas that have no or just a few dance studios nearby, but which are easily accessible for a number of people without the usage of cars, as parking is almost impossible in Budapest.

Finding the best spot is not in my intention, as I understand there are many other factors to be taken into account, some of them being less data driven, but I will attempt to locate the top 3 neighborhoods to help narrow down the list of potential areas to consider for a new dance studio.

Data

As per the problem setting above, I will focus on the below aspects:

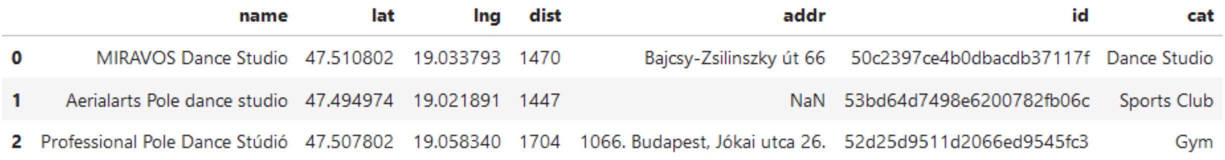
What data would be useful/needed?

1. The number and location of existing dance studios in Budapest as of today.
2. The distance of the studios from the center of the city.
3. The average ratings the studios received from the users.
4. Crime data of the neighborhoods/districts.
5. Entry fees or hourly rates of the studios.

What data are available?

1. With a bit of probing I found that the Foursquare page has various but limited information about the places, like location (address, geo coordinates), tips, reviews and ratings. Useful for my purpose from these are the geo ones and the ratings, so I will concentrate on retrieving those primarily, via the Foursquare API.
2. Understanding the crime heatmap about the districts would be an asset to find a rather secure location. There can be venues, like night clubs that may to a certain degree magnetize crime to a specific area. Though the open hours of the dance studios (mainly evenings) and that of the night clubs (late night) have no intersection, this would be worth looking into. Unfortunately, I did not find any available dataset about the criminal records nor on Foursquare neither on any other website, so I put this view aside for the current analysis.
3. As for the prices, I have no information about the fees and rates for those studios the Foursquare site lists. Hence, I postpone this aspect too for a later study.

So as above mentioned, the source of the data to be used will be Foursquare. I will first capture the latitude and longitude coordinates of Budapest via an API call, then pull the list of venues in the city, transform the json file into a dataframe and clean it for the purpose (filter on the dance studios, let go of the unnecessary columns, etc.) Feature selection part I would skip for this project as I would be applying the clustering method, so I would rather concentrate on the pre-processing in that direction (slice the venue part from the JSON file and transform it into a pandas dataframe, tidy up the category so that it displays the relevant part, the name only, filter out the needed columns and rename them, etc.) After these steps the head of the dataframe takes the below shape:



	name	lat	lng	dist	addr	id	cat
0	MIRAVOS Dance Studio	47.510802	19.033793	1470	Bajcsy-Zsilinszky út 66	50c2397ce4b0dbacdb37117f	Dance Studio
1	Aerialarts Pole dance studio	47.494974	19.021891	1447	NaN	53bd64d7498e6200782fb06c	Sports Club
2	Professional Pole Dance Stúdió	47.507802	19.058340	1704	1066. Budapest, Jókai utca 26.	52d25d9511d2066ed9545fc3	Gym

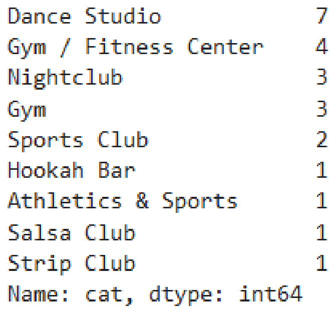
Unfortunately, I am not going to be able to use ratings as there are none for the selected/filtered studios.

Methodology and analysis

This project focuses on analyzing the dance studio density. I will use folium to visualize their spread with a heatmap first, looking for the low numbered areas and then use k-means clustering machine learning technique (as we are talking about location data) to group the studios based on their location and see if further patterns, insights might become visible enabling better exploration for the optimal location.

Exploratory data analysis

Going through the venues I see “dance” being spelled with both capital and lower case "d" in their names, so I filter on both version carefully when looking for the “dance studios” from the retrieved list of venues of various kinds. Then join the two dataframes together to have my complete list of dance studios. Another round of sanity check is performed on the categories to see if there is any to close out¶



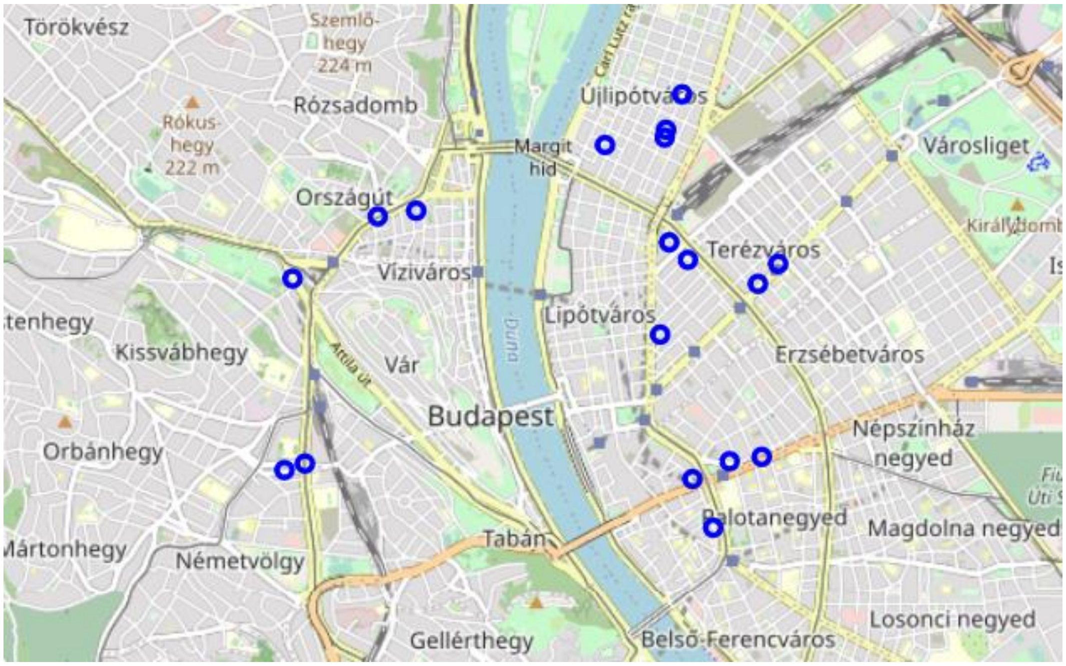
```
Dance Studio            7
Gym / Fitness Center    4
Nightclub               3
Gym                     3
Sports Club             2
Hookah Bar              1
Athletics & Sports      1
Salsa Club              1
Strip Club              1
Name: cat, dtype: int64
```

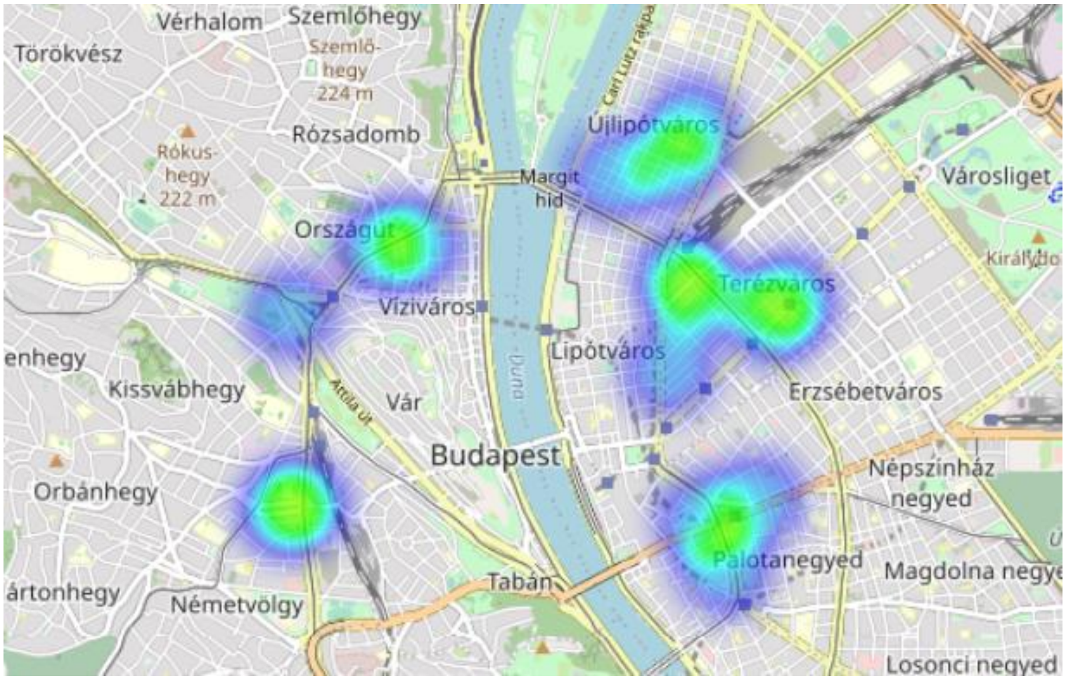
Categories seems to vary, yet the names of places suggest similarity in scope, so many of those not under dance studio might represent competition. Even pole dance venues could be "dangerous" as they can be an option for the female target audience, so I remove the night club-like places only. Those are the: Nightclub (3), Strip Club (1), Hookah Bar (1).

The *describe* gives me some main stats about the distance. There is an average of 1776 metres of distance from city center and a standard deviation of 361 meters. This indicates the dance studios are not straight in the so-called center, but a bit further away. I'll use the folium map to visualize them and see if there is any noticeable pattern about their location.

Seeing the max 2534 meters gives me a hint on what radius I could have chosen earlier on but seem with the 5000 I made no mistake of closing out any places.

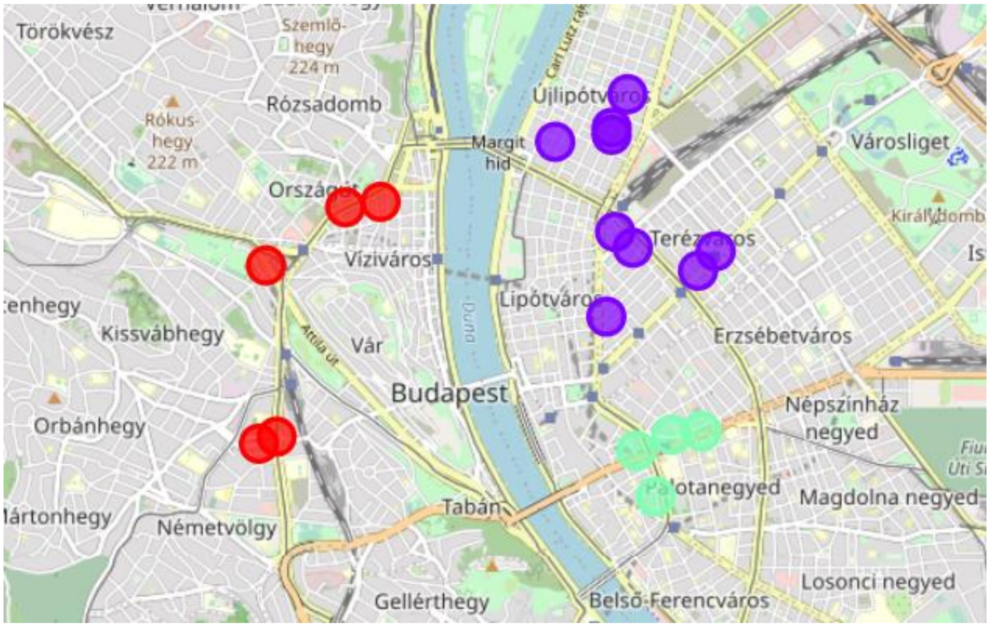
Then I put the studios on a map and have the first look. Then try with a heatmap to better visualize the density of the studios.


Törökvész
Szemlő-
hegy
224 m
Rózsadomb
Rókus-
hegy
222 m
Carl Lutz ra
Újlipótváros
Margit
híd
Városliget
Országút
Víziváros
Terézváros
Királydomb
Is
Lipótváros
tenhegy
Kissvábhegy
Attila út
Vár
Duna
Erzsébetváros
Budapest
Orbánhegy
Népszínház
negyed
Fiu
Úti S
Palotanegyed
Magdolna negyed
Mártonhegy
Németvölgy
Tabán
Losonci negyed
Gellérthegy
Belső-Ferencváros

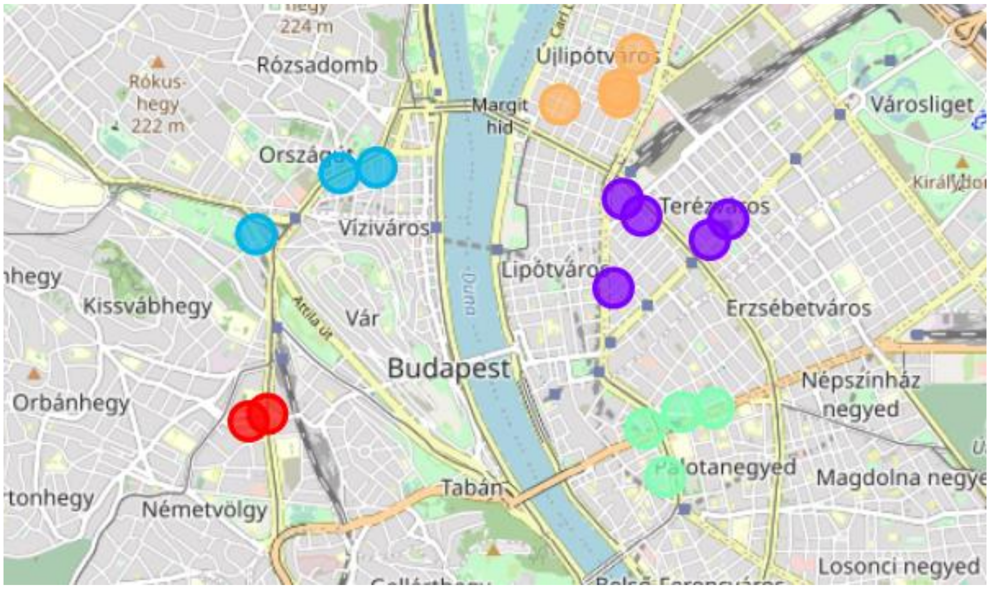

Vérhalom
Szemlőhegy
Törökvész
Szemlő-
hegy
224 m
Carl Lutz rakpa
Újlipótváros
Rókus-
hegy
222 m
Rózsadomb
Margit
híd
Városliget
Országút
Királydo
Víziváros
Terézváros
Lipótváros
enhegy
Kissvábhegy
Attila út
Vár
Duna
Erzsébetváros
Budapest
Népszínház
negyed
Orbánhegy
Ú
Palotanegyed
Magdolna negye
Tabán
ártonhegy
Németvölgy
Losonci negyed

Clustering dance studios in Budapest

As the number of studios is not that high, I'll try with 2 iterations. First with 3 clusters and then seeing how it works, maybe increase the number a bit.



Though the groups look ok for the first sight, but for anyone knowing the city a bit it is evident that these clusters cover a large area, the studios in them are quite far from each other, so probably were grouped together because of the low number of clusters. To find out, I increase the number of clusters a bit, say to 5.



As to be seen, the groups separate nicely breaking into smaller ones with studios a lot closer to each other.

Results and Discussion

As per the above map, there are a number of insights to draw.

First is that, though it is relative what we call "many" and my biased brother would certainly doubt my coming statement, but I find the number of the dance studios in Budapest (at least those that are listed by Foursquare, 18 after the filtering) quite reasonable, definitely not way too many.

The slightly large distance from the city center seems confirmed, as it is around the letter of "a" in the word "Budapest" in the middle of the Map. So that means, studios are situated not straight in deepest downtown. This makes sense bearing in mind those areas are either mainly very narrow streets, for mostly tourists, and/or with limited and difficult public transportation for a mass of people. The green area just below the Budapest word (called Tabán) is a park, above it there is a castle where no locals go unless they live there. On the Pest side, same situation, there is the Parliament there, and the shopping streets, which we do like to avoid going.

It is eye-catching also, that the existing studios tend to be around, or very close to the inner and middle boulevard of the city. As to be seen, they are more or less on a circle-shaped curve. These are the main streets of the city with decent public transport (trams, buses, underground stations and still considered downtown-ish.) Interesting to see that further out there are again no studios.

The Buda side (left to river Danube) is less "crowded" of dance schools. Though this is not part of this analysis due to lack of dataset, but the real estate prices, rent fees, etc. are way higher there than on Pest side, so anyone planning to open any kind of place there must be making decent revenue in order to make the investment profitable. This makes me concentrate on the Pest side (right to the river).

Orange and purple ones (and even reds on Buda) are relatively close to railway stations. These are areas of the city there the commuting people arrive from and leave to their agglomeration homes.

These above observations help me narrow down the list of potential areas to the following:

1. Rákóczi út between East Railway station and Blaha Lujza square,
2. Károly boulevard and Bajcsi Zsilinszki út between Astoria and West Railway station,
3. Elisabeth boulevard.

Conclusion

This project had the purpose to find an ideal location of a new dance studio in Budapest. Under ideal I understood, being in an attracting neighborhood but also being away from the rest of the schools. I pulled and wrangled the Foursquare data about the currently existing (and listed) dance studios, their names, location, categories, etc. Unfortunately, there were no ratings available for me to retrieve for these studios, so I could not analyze what the averages of ratings per neighborhoods are like.

Considering the nature of the data at hand (locations) I used the clustering technique and complemented my analysis with a heatmap for density visualization. The results showed me not only where the studios are and so the areas where there are not that many or none, but also highlighted some key features/indicators to take into account from the location perspective (railway-proximity, etc.).

As indicated in the beginning, my goal was not to find the "best" place. There are many other factors (rent price, crime rates, class fees, socializing potential, etc. that, if/once sufficient data is available, could be subject to further analysis) influencing what the best is, so I wanted to offer a top 3 options about the location based on other schools locations.

Final decision is to be made by the stakeholder(s) based on specific nature of the neighborhoods and locations for each suggested street (intentionally not talking about a specific address of a given street).