IBM Data Science Professional Certificate
Applied Data Science Capstone
Capstone Project - The Battle of Neighborhoods

Where should you install your startup headquarter to guarantee a pleasant working and rest environment to your collaborators?

Antony Borel

2020

1. Introduction: background and problem

An Hungarian startup is quickly growing. Its current headquarter is a small room in the countryside, located 4 hours from the capital city, Budapest. **The company needs to relocate its headquarters in Budapest** in a larger space as the collaborators are getting more numerous and they have to go abroad and host partners from abroad to maintain and increase the international growth of the company.

The two founders of the company pays particular attention to the **well-being of their collaborators and their partners visiting the company**. Therefore, the choice of their new headquarter will be highly related to the accessibility of facilities allowing relaxing and/or facilitating the daily harmony between professional and personal activities.

The startup founders are thus **asking for suggestion about where to settle their new headquarter**. The location of the new headquarters must meet the following criteria:

- one of the famous **official (thermal) bath** from Budapest should be no more than **1km** (as the crow flies) from the headquarter
- a **conference center** should be no more than **1km** (as the crow flies) from the headquarter
- a **library** should be no more than **1km** (as the crow flies) from the headquarter
- at least one restaurant providing **vegetarian or vegan food** and one restaurant providing **hungarian food** should be no more than **500m** (as the crow flies) from the headquarter
- a **fitness center** should be no more than **500m** (as the crow flies) from the headquarter
- location from the Liszt Ferenc **International Airport** of Budapest is also important. Going to the airport should take **less than 45 minutes** by car/taxi.

In case of more than one location responding to all the criteria, the duration of the trip from/to the airport could be used to hierarchize these locations.

2. Data acquisition and cleaning

2.1 Data sources

Based on the request from the startup we had to gather **geospatial data** as well as **name and categories of venues in Budapest** in order to find a solution which meets the given criteria.

Geopy was used to obtain **latitude and longitude of Budapest city and Budapest international airport**. It was also used to get coordinates of the suggested place for the headquarter.

The given distance for the first 5 criteria are "as the crow flies" so we used **Foursquare API** (https://developer.foursquare.com/) to get **latitude and longitude associated with name and categories of each of the venues**. However, a maximum duration for the trip by car/taxi to the airport is given as criterion, this information cannot be obtain through Foursquare. So, to obtain **distance and trip duration to the airport** we used **MapQuest API** (https://developer.mapquest.com/).

2.2 Data cleaning

The startup founders indicated that they wish to be close to an "official (thermal) bath" of Budapest. So we **filtered the Foursquare results** in order to consider only **official baths**. Based on the official website of Budapest baths (http://www.budapestgyogyfurdoi.hu/gyogyfurdok-es-strandok) we knew that there are 12 official baths: Széchenyi, Gellért, Rudas, Lukács, Király, Dandár, Paskál, Palatinus, Csillaghegy, Pesterzsébet, Római and Pünkösdfürdő.

Based on the details given on this webpage we could evaluate that a radius of 12km was necessary in the Foursqaure queries in order to obtain the data of all the official thermal baths. For each query, we also gave a limit number of results of 100 as, even if many thermal baths not considered as official city bath exist in Budapest, this limit was enough to get the 12 baths we were looking for. We performed several tests in order to get better, more direct, results from Foursquare, using different types of queries and combined queries, but none of them was giving the 12 baths at once. Indeed, these baths are classified in different categories ("Spa", "Water park" and "Pool"), with no string in common in their name, etc., and it seems that the API in not very stable. Best results were obtained first searching, within the names, for "Gyógyfürdő" (hungarian term for "thermal bath" or "spa") and then searching for "Strand" (hungarian term for "pool" or "water park"). After each query we filtered the data on the category to keep only the categories of interest. We finally get the data of the 12 baths splitted into 2 dataset. We concatenated them into one dataset named "WaterPoints".

For the other venues (i.e. Vegetarian /vegan restaurants, Hungarian food restaurants, fitness centers, conference rooms and libraries), results from Foursquare were **filtered to keep only relevant category** and make sure that all venues were associated with geolocation data.

We kept the same radius and limit for the queries concerning the other venues in order to target the same area.

After each dataset was cleaned we run a **quick visualization with folium library in order to verify the consistency of the returned venues locations.**

This allowed us figuring out that the **coordinates of Szent Lukas thermal bath were wrong**. We informed Foursquare about this mistake but we corrected it directly in our dataset. To do so we searched for the real coordinates in Google map as geopy did not find this address. Google map returns: 47.517898, 19.036682 (Fig. 1).

We finally **concatenated the 6 datasets** into one, called DataFull (Fig. 2). This dataset contained **103 venues**: 12 baths, 23 vegetarian/vegan restaurants, 10 Hungarian food restaurants, 28 fitness centers, 11 conference rooms and 19 libraries.

Distances to the airport obtained with the Mapquest API were **given in miles** so we add to **convert them into metric system**. We stored distance and duration data into a dataset called distAirport (Fig. 3).

2.3 Features selection

In the DataFull dataset we used the features **"name", "category", "lat" (latitude) and "lng" (longitude)**.

From the distAirport dataframe we used the attributes **"from"** (which corresponds to the venue from which the distance and duration is calculated), **"distance to airport in km" and "duration to airport"**.

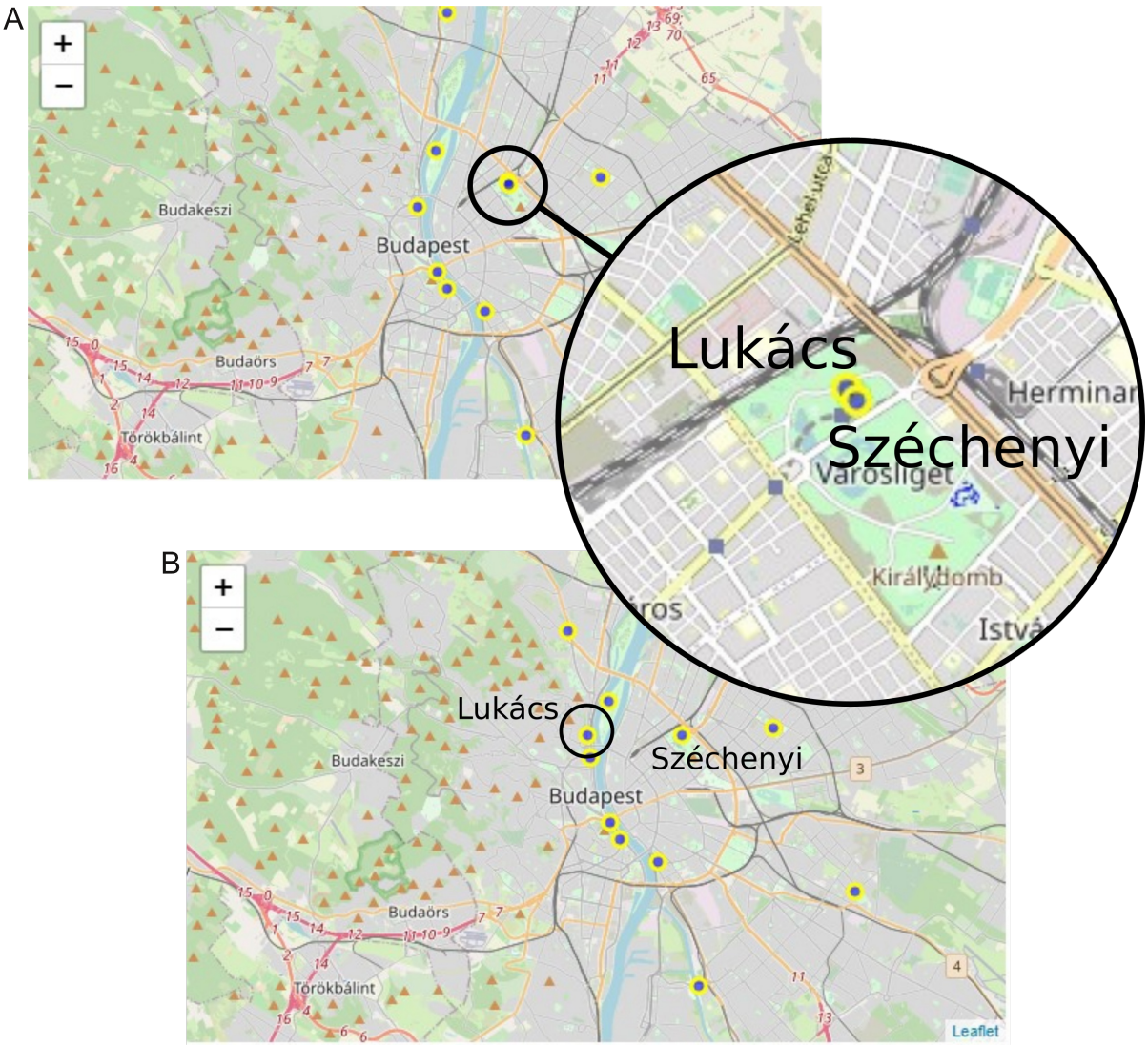

Figure 1: Location error (A) of the Lukács bath and its correction (B).

	name	categories	lat	lng
0	Rudas Gyogyfurdo es Uszoda	Spa	47.489188	19.047761
1	Szent Gellert Gyogyfurdo es Uszoda	Spa	47.483917	19.052256
2	Szent Lukacs Gyogyfurdo es Uszoda	Spa	47.517898	19.036682
3	Dandar Gyogyfurdo	Spa	47.476337	19.071061
4	Szechenyi Gyogyfurdo es Uszoda	Spa	47.518302	19.082394
5	Kiraly Gyogyfurdo	Spa	47.510608	19.038185
6	Palatinus Strandfurdo	Water Park	47.529170	19.046946
7	Paskal Gyogy- es Strandfurdo	Water Park	47.520571	19.127469
8	Romai Strandfurdo	Pool	47.574811	19.052087
9	Punkosdfurdoi Strand	Pool	47.594627	19.067900
10	Pesterzsebeti Jodos-Sos Gyogy- Es Strandfurdo	Pool	47.435448	19.090965

Figure 2: I0 first rows of the DataFull dataset.

	from	Distance_to_airport_in_km	Duration_to_airport
0	Rudas Gyogyfurdo es Uszoda	21.684301	00:24:08
1	Szent Gellert Gyogyfurdo es Uszoda	20.799162	00:23:35
2	Szent Lukacs Gyogyfurdo es Uszoda	24.338109	00:29:37
3	Dandar Gyogyfurdo	18.787482	00:20:08
4	Szechenyi Gyogyfurdo es Uszoda	22.585534	00:25:20
5	Kiraly Gyogyfurdo	25.413151	00:27:43
6	Palatinus Strandfurdo	26.441522	00:29:48
7	Paskal Gyogy- es Strandfurdo	22.505066	00:26:37
8	Romai Strandfurdo	47.091015	00:36:15
9	Punkosdfurdoi Strand	44.725279	00:32:55
10	Pesterzsebeti Jodos-Sos Gyogy- Es Strandfurdo	19.347534	00:20:56

Figure 3: I0 first rows of the distAirport dataset.

3. Methodology

We first displayed all the venues of interest using the **folium library** in order to **explore geospatial data** and observe their organization in Budapest. We ploted circles of a radius of Ikm around each official bath in order to gain a better idea about the potential areas of interest for headquarter. Indeed, the headquarter should not be further than Ikm from a bath. As official baths are spread at quite large distance from each other, as they are not numerous and as the startup founders gave it as the first criteria, we decided to base primarily our observations on this category of venue.

As the headquarter should be surrounded by few (at least 6 of different given categories) venues, we performed a **cluster analysis** to identify clusters of venues where the HQ would be close enough to everything. We use **DBSCAN algorithm** as this algorithm does not require to set a number of clusters a priori and do not constrain the shape of the clusters. DBSCAN was performed with the **Scikit-learn library**. Clusters were built using the distance of each venue to the airport and the corresponding trip duration as well as using the location (latitude and longitude) of each venue.

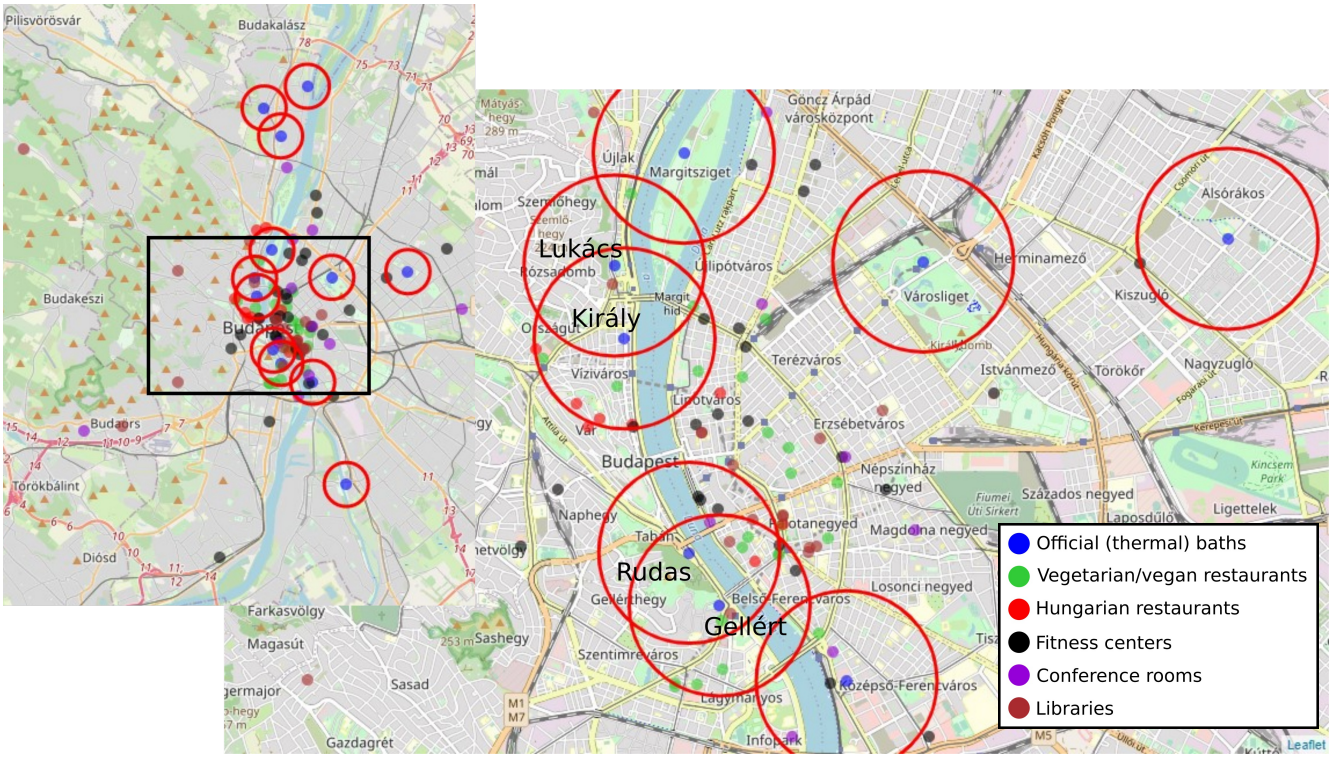

Figure 4: Map of the venues of interest in Budapest. Red circles represent an area of a radius of lkm around each official bath (blue points).

4. Results

4.1 Exploration of geospatial data

We first explored the organisation of the venues using folium. From this map, we saw that **most baths are on Buda side** (West of the Danube) of Budapest while **most venues of interest are on Pest side** (East of the Danube). Also, **half of the baths do not have any venue of interest in their surrounding** (i.e. at less than lkm). This showed that, in order to meet the criteria of the startup, **the best place for the headquarter is very likely to be in the city center**. The surroundings of the Gellért and Rudas baths seem to be good candidates. Lukács and Király baths areas may also be acceptable locations (Fig. 4).

4.2 Clustering of geospatial, distance and duration to airport data

In order to reduce the number of possibilities and to include all the criteria given by the startup founders, we carried out a cluster analysis using **DBSCAN on the latitude and longitude of each venue and their distance and trip duration to the airport**.

Two clusters are identified (numbered 0 in blue and I in purple) (Fig. 5). Numerous venues have been classified as -I (red points), showing that numerous venues could not be associated to a cluster. These venues are two far from each other and/or to the airport by car to be considered as being of interest in our case. The

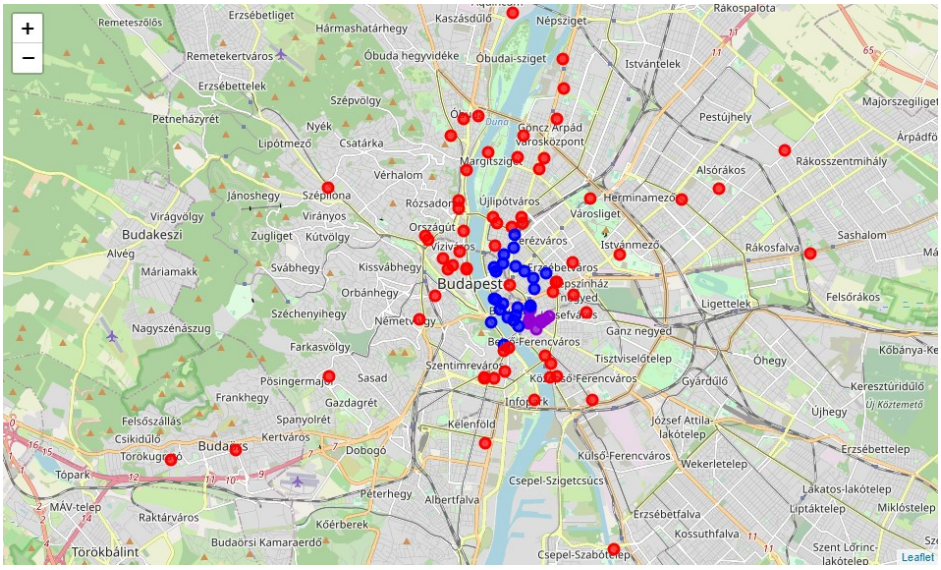

Figure 5: Results of the cluster analysis using DBSCAN algorithm. Red points are outliers. Two clusters of venues are identified in the city center.

cluster I is spread close to Kálvin tér, while **cluster 0 is spread on Belváros, Lipótváros and the surrounding of the Opera**.

To better apprehend the potential areas of interest, we remove the venues of cluster -I which can be considered as outliers. They are venues which does not have 6 neighbors in their close surrounding. Thus they are unlikely to be in an area meeting the criteria for the HQ. Then, we plotted the 35 remaining venues (Fig. 6).

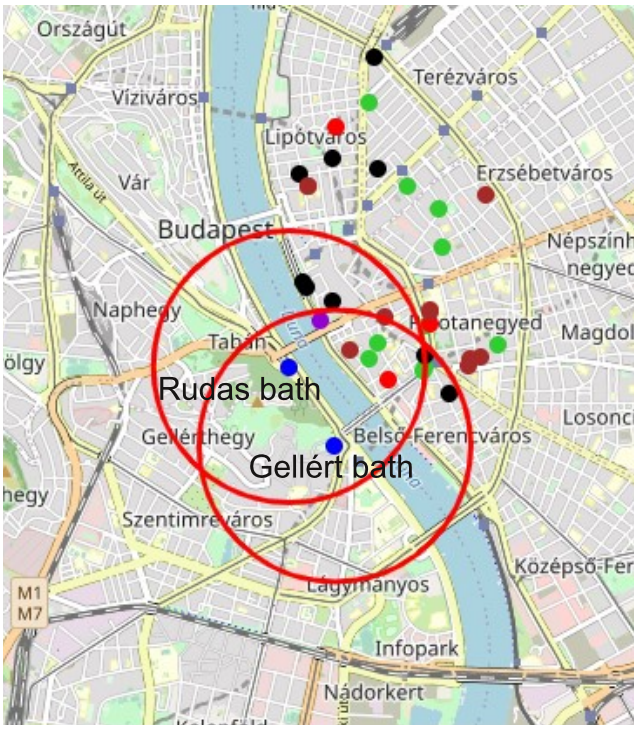

Figure 6: Map showing the venues which were associated to cluster 0 or I in the cluster analysis.

We observe that **both Gellért and Rudas baths have all required facilities** in a radius of Ikm, but all in the other side of the Danube. On the side of most facilities (Pest side), both circle of Ikm radius crossed each other close to Kálvin tér. **Kálvin tér** is also an area where venues are clustered (see cluster I). If the headquarter would be close to this area, collaborators would have **access to both baths and to all other categories of venue**. Therefore, this seems to be the good place for the headquarter.

4.3 Visualization of the suggested area and description of venues accessibilities

We used geopy to get the coordinates of Kálvin tér and we used these coordinates with the Mapquest API to compute the distance and duration of a trip by car from Kálvin tér to the airport.

Kálvin tér is located **around 20km from the airport** and it takes **less than 23 minutes to reach the airport by car**.

We observed the accessible venues within a distance of 500m and Ikm (as the crow flies) from Kálvin tér. We also displayed a circle showing an area of 700m and I.2km around Kálvin tér in order to take into consideration that the headquarter may be several meters from the exact position of Kálvin tér. We suggest to search for a real estate in an area of 200m of radius around Kálvin tér (Fig. 7).

If a real estate can be found at the direct vicinity of Kálvin tér:

- **Gellért bath**, one of the famous official (thermal) bath from Budapest, will be less than Ikm from the headquarter
- **Quince conference room** will be less than Ikm from the headquarter
- **7 libraries** will be less than Ikm from the headquarter, 6 of them will be less than 500m away
- **5 restaurants providing vegetarian or vegan food** and **2 restaurants providing Hungarian food** will be less than 500m from the headquarter
- **2 fitness centers** will be less than 500m from the headquarter

Kálvin tér is therefore the best candidate, as it meets all the given criteria and even overpass them.

5. Discussion

Our analysis shows that even if there are I2 **official (thermal) baths** in Budapest, they are very **far from the airport and/or isolated** without many venues in their surroundings. Both Gellért and Rudas baths could have been good candidates to be selected as the bath close to the headquarter. Howerever, they are both on the West side of the Danube while most of the required venues are located on the East side. Therefore, **this criterion in particular reduced drastically the candidate areas to host the new headquarter**.

Combining exploration of the geospatial data and clustering using DBSCAN algorithm, we could eliminate venues which were too far or isolated to have a chance to satisfy the criteria. Based on the results of the maps description and of the clustering, the surrounding of **Kálvin tér have been identified as the best candidate area to host the new headquarter**.

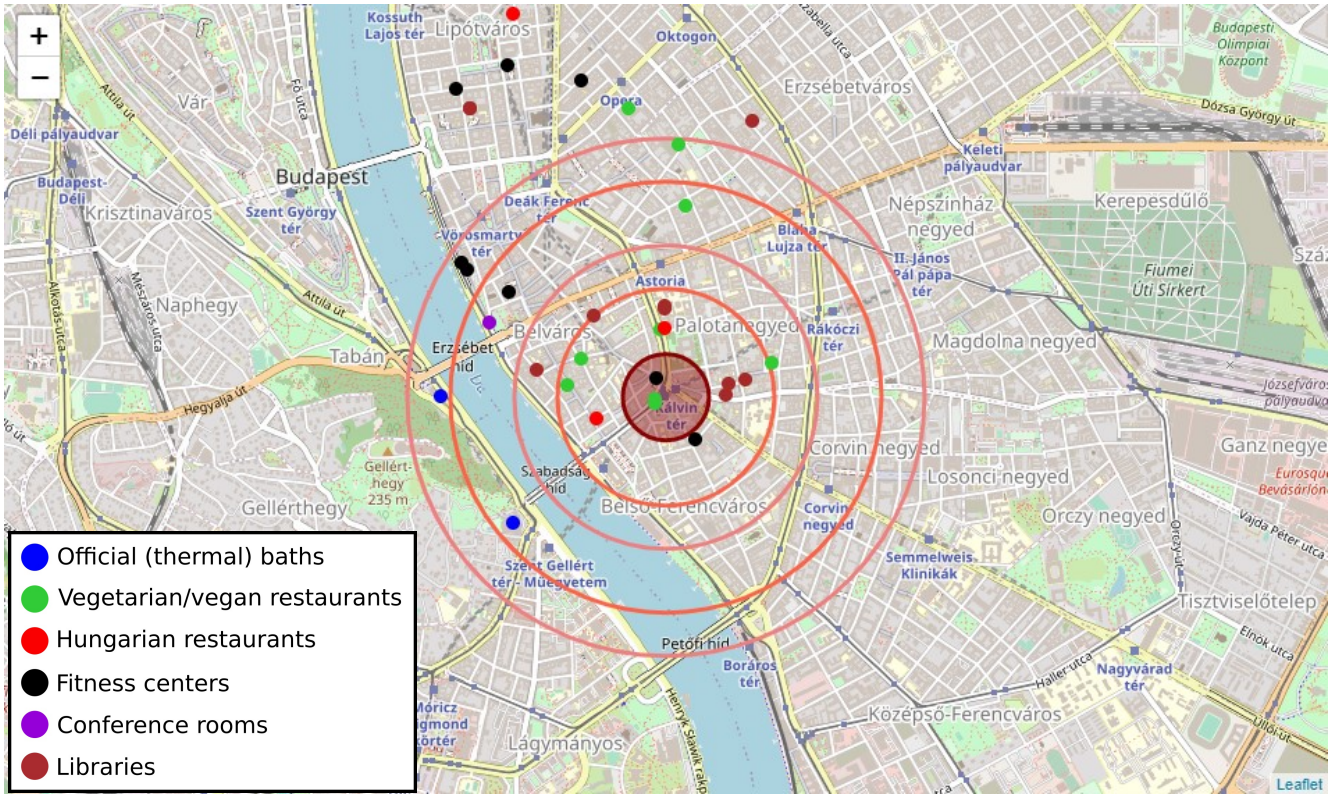

Figure 7: Map showing the accessibility of each category of venues from Kálvin tér, which is represented by the filled circle of 200m of radius. From the center towards the exterior, the first empty circle represents the limit of 500m from Kálvin tér, the second is at 700m, the third at 1000m and the last one at 1200m.

Kálvin tér is in **straight connection with the Gellért bath**, which would be located less than 1km from the HQ. Even if they are on different side of the Danube, they are connected with the Szabadság (Liberty) bridge. Going to the bath will be easy either walking or by public transportation. **One conference room and 7 libraries** are located less than 1km from Kálvin tér. Also, the collaborators will have the opportunity to choose between **5 vegetarian/vegan restaurants and 2 Hungarian restaurants**, located less than 500m from Kálvin tér. Fitness is also available with **2 fitness centers** located at less than 500m from Kálvin tér. Finally, Kálvin tér is one of the best location to go to the airport as it is on the main road joining the city center and the airport. Therefore, it will take **less than 23 minutes by car to reach the airport terminal**. Even if not requested, we can add that we found that Kalvin Tér has metro station for 2 lines and is connected by this way to 3 train stations and to the bus station which goes to the airport by public transportation.

As finding a real estate in the direct vicinity of Kálvin tér may be challenging, we also provided visualization for the venues accessibility considering that the HQ may be in an area of 200m of radius around Kálvin tér. In such case we recommend to locate the HQ rather in the west of Kálvin tér if it is desired to strictly keep a distance of less than 1km with an official bath.

6. Conclusion

Using Foursquare and Mapquest API, it has been possible to gather the necessary data to response to the given problem. Mapping and description as well as clustering of geospatial and itinerary data allow to propose a solution to the startup founders. We could point the best area in Budapest which meet all the criteria they had to settle their new headquarter. In continuation, if they would like suggestion to optimize the organization of the time of their collaborators, we could search for the best venue of each category considering the duration to reach each of them from the HQ by walking and by public transportation.