Location Recommendation System for New Businesses in Budapest, Hungary

1. Introduction

This is the capstone project report for the IBM Data Science course at Coursera. In the project, I try to help people who want to found a new business or expand their existing one in Budapest, Hungary.

One important element of starting a successful business is finding the best location for it. It is well known that similar businesses tend to gravitate to each other, creating clusters of them all around the city. If you open your new shop close to your competitors, you have a higher chance that people will find and try it just by chance.

The user of the recommendation system can enter a category as an input, e.g. "Coffee Shop", "Restaurant", "Bookstore", etc. and the system will determine the best location of a new venue based on the distribution of the already existing ones.

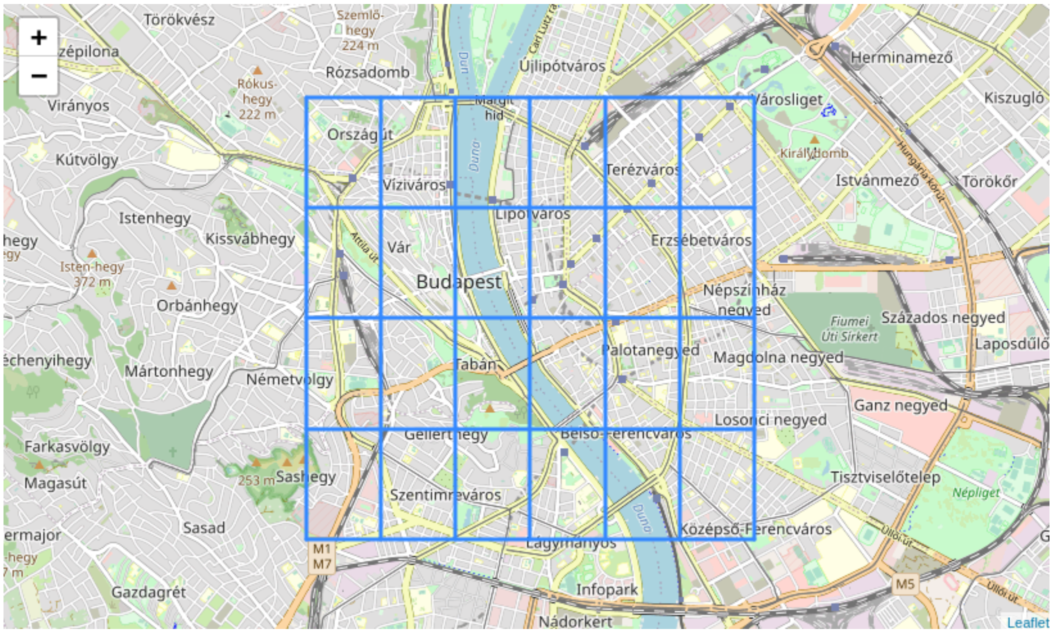
In this report, I explain in detail how the necessary data is acquired and processed, how the location recommendation is generated, and discuss the results of some example inputs.

2. Data

The data is taken from Foursquare, using its so-called Places API. The user enters a category and the system searches for the kind of business in the central part of Budapest, and it collects their location information. What we will need for the analysis is only the geographical coordinates of the venues.

The list of categories, that the user can use, is also taken by using the same API. Currently, there are 459 different categories available in Foursquare. The category list is also available on the following link:
https://developer.foursquare.com/docs/build-with-foursquare/categories/

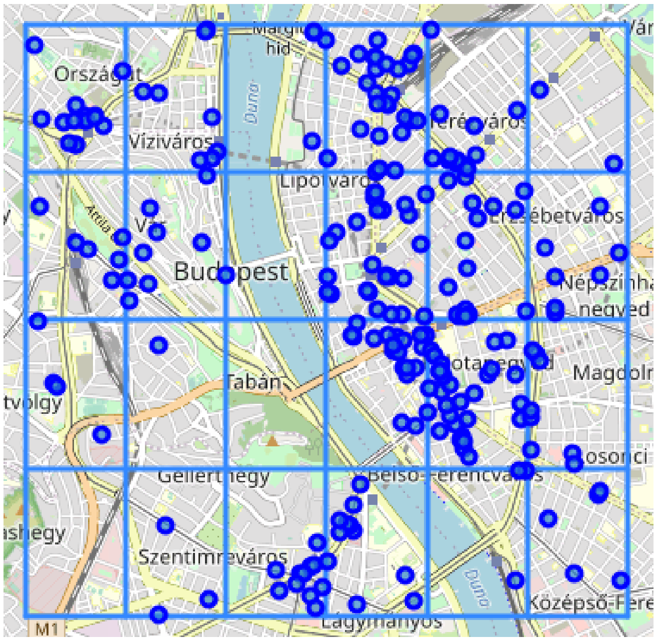
In order to overcome the problem that Foursquare returns only 50 venues at a time, I divided the area of interest into 24 equal-sized, rectangular regions as shown it the picture below. The size of a region is 0.01 x 0.01 degrees in terms of geographical coordinates. As one longitude degree and one latitude degree are getting more and more different in as we move away from the equator, these regions have a rectangular shape instead of a square.



Once the regions are specified and we know the bordering coordinates of all of them, we can fetch the necessary information from Foursquare. With the “sw” and “ne” parameters of the Places API, we can search for a rectangular area, and by this means we can search for venues in the whole area of interest without overlapping.

First, I generate 24 URL-s for the 24 regions, then I call the API with all of them and store the resulted coordinates in a list. I do not store any other information, like name, address, or rating of a given venue. For our purposes, the coordinates are enough.

Here is an example dataset as we fetch all bookstores in the area of interest.



3. Methodology: Clustering with Mean Shift algorithm

Once the location information of a given venue category is available, we will cluster the results with the Mean Shift algorithm. An important feature of Mean Shift is that it is not necessary to specify the number of clusters in advance. Instead, we specify the so-called bandwidth. The smaller the bandwidth, the more clusters will be identified with fewer members that are closer to each other.

We try to find the clusters in which the venues are the closest to each other. In the examples below we set the bandwidth to a relatively small value, which will result in numerous clusters, especially, if there are many results for the given category. By experience, a bandwidth of 0.002 provides good results for most of the categories. As the data we are clustering is geographical coordinates, the unit of the bandwidth parameter is also degrees of geographical coordinates. The bandwidth of 0.002 will result in clusters whose size is roughly one-fifth of the above-specified rectangular regions. In practice, the size of a typical cluster will be just a few blocks.

Once the clustering is done, we choose the 5 clusters with the most members, and that gives us the recommended locations of our new venue. The algorithm provides the coordinates of the cluster centers, but maybe it is more important to visualize the members of these best clusters. The size and shape of the resultant clusters are well visible on the map and it gives us a good intuition about the best location of a new business, which is of course not only a single point on the map but sometimes a section of a street, sometimes a few blocks around a given position, and many times it is a shopping mall.

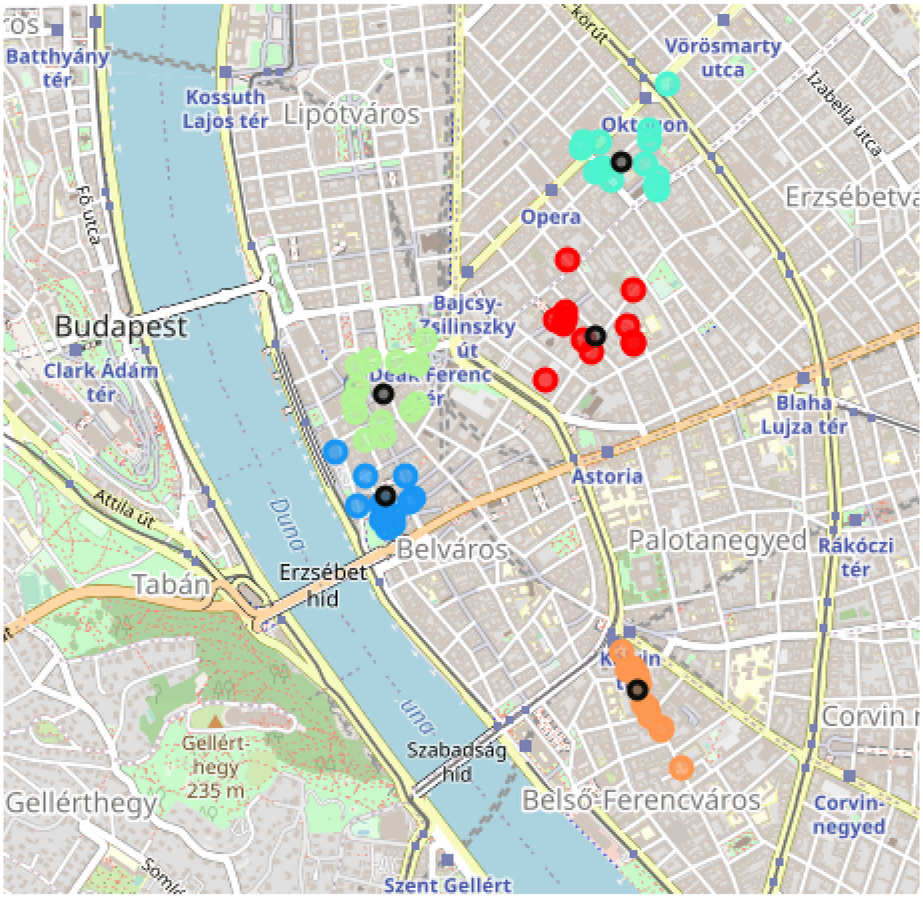
The number of recommendations, i.e. the number of how many clusters we choose from the results, is a parameter, which can be modified by the user.

4. Results and Discussion

In the section below, we will check out the results of the location recommendations of some selected categories and we will do a quick sanity check. Black markers are the cluster centers, colored ones are the already existing venues from the category in question. Let’s see.

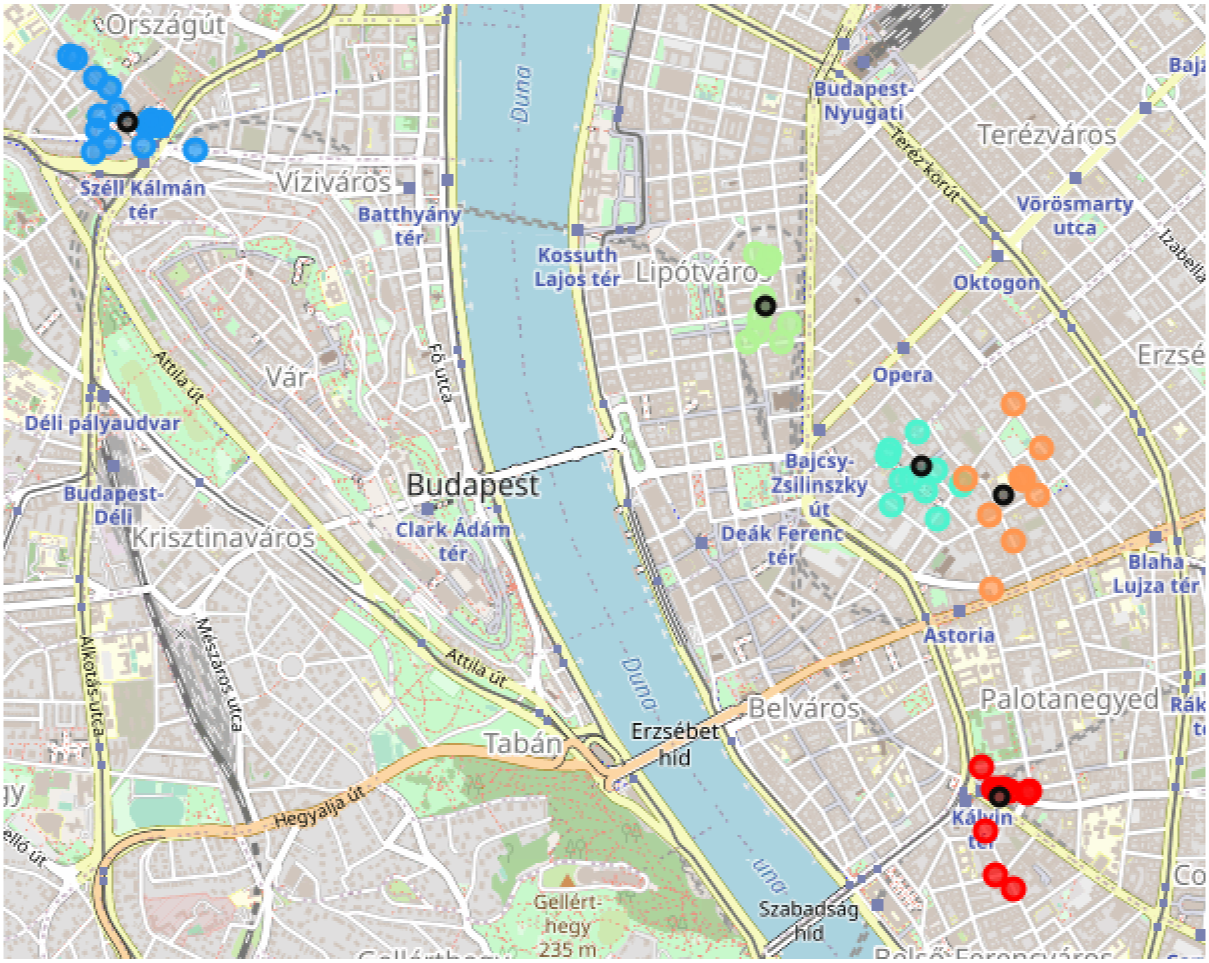
Restaurants

The system has found 405 restaurants, which are grouped into 114 different clusters. Not surprisingly, all of the 5 best ones are in downtown. The green cluster is simply a downtown location with plenty of tourists and plenty of restaurants. Blue is the neighborhood of Váci street, also focusing mostly on tourists. Red is the party zone of Budapest, cyan is Ferenc Liszt square, a well-known location of good restaurants, and orange is the famous Ráday street, with many good places to eat. The results make perfect sense to me, let’s continue with something else.



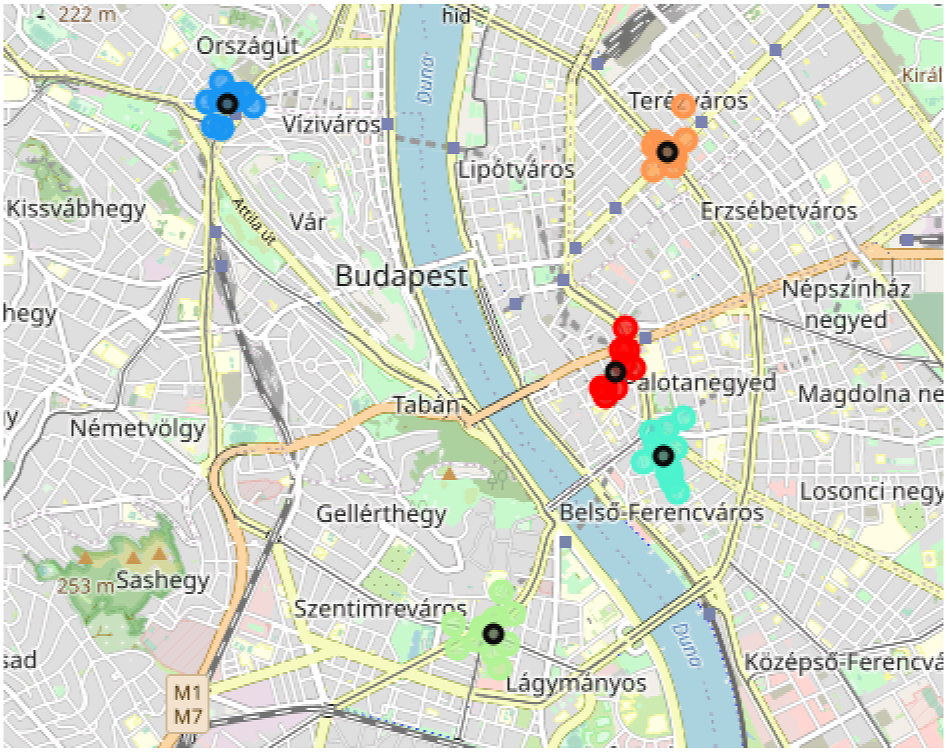
Coffee Shops

There are 345 coffee shops, which are grouped into 92 different clusters. Cyan and orange is the party zone, red is again the Ráday street. Blue is a large shopping mall called Mammut. I do not know about the features of green, but seemingly it is a good spot to open a coffee shop in that area.



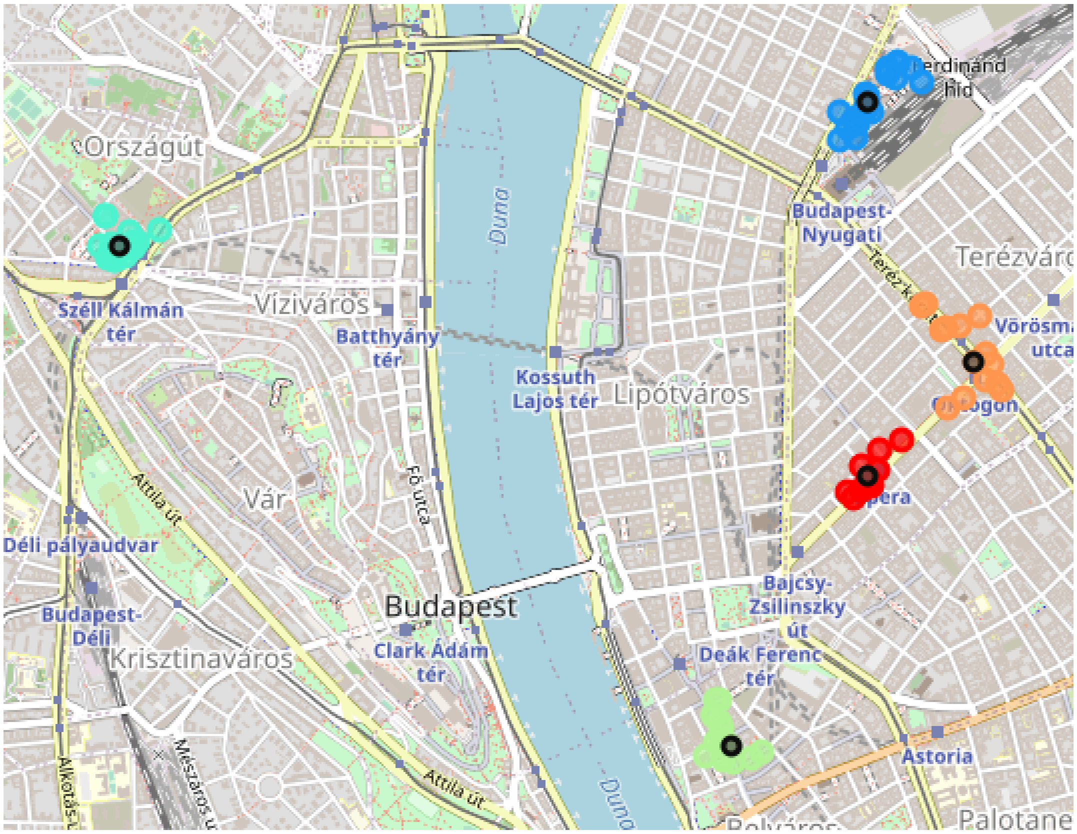
Bookstores

There are 228 bookstores, which are grouped into 77 different clusters. Interesting results. Green and red are close to the biggest universities, blue is the shopping mall “Mammut” and its vicinity. Interestingly, Ráday street (cyan) is also a good spot for bookstores however it is known mostly from its restaurants. Orange is a busy square called Oktogon. I have no clue what brings so many bookstores here, except of course the tens of thousands of people moving around here each day.



Jewelry Stores

There are 213 jewelry stores, which are grouped into 64 different clusters. Cyan is the shopping mall “Mammut” that we already know. Blue is another one called “Westend”. Green is the tourist focusing Váci street. Red indicates another, maybe the most famous street in Budapest, Andrássy Avenue. Especially around the opera building, it is also a good spot for a jewelry store. Orange is the vicinity of Oktogon, a busy area in general.



Notaries

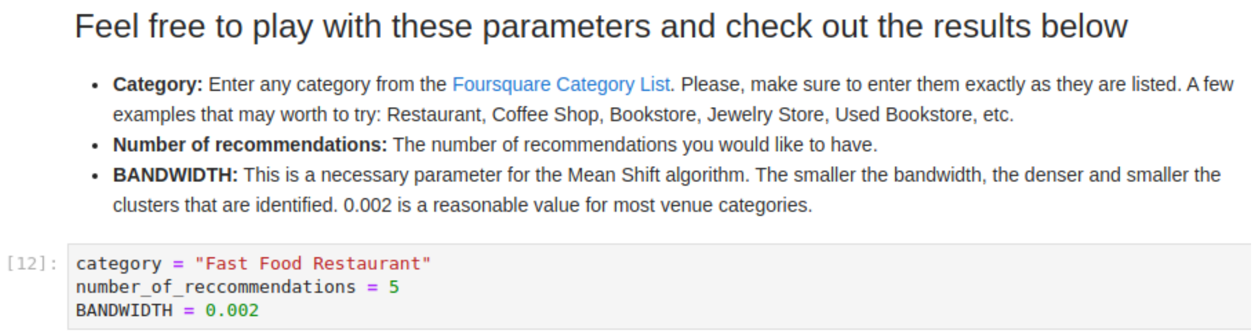
My system found only 2 notaries in Budapest, grouped into 2 different clusters, and we are not able to draw a meaningful conclusion from this result. Well, I am pretty sure that there are more than 2 notaries in the downtown of Budapest. This is just an example showing the limitations of Foursquare and my location recommendation tool. Unfortunately, it is not good for everything.

5. Conclusion

In the capstone project presented above, we have built a location recommendation tool helping people to find the best location of a new venue in the city of Budapest. The user can enter the category of their interest, then the tool, using Mean Shift clustering over the location data of the existing venues in the same category, shows the 5 best locations on the map.

It is tested for a couple of categories and the results are discussed. It turned out that the tool provides meaningful results for many categories, especially if there are a lot of results provided by Foursquare's Places API. However, some kind of businesses like notaries is seriously underrepresented in Foursquare's database. In this case, it is not possible to draw a meaningful conclusion and to recommend a good location.

Feel free to try it. The tool is available on GitHub and it is compressed in only 3 cells of Jupyter Notebook. You can just enter the 3 available parameters in the second cell, and check the results, executing the 3rd cell.



Feel free to play with these parameters and check out the results below

- **Category:** Enter any category from the Foursquare Category List. Please, make sure to enter them exactly as they are listed. A few examples that may worth to try: Restaurant, Coffee Shop, Bookstore, Jewelry Store, Used Bookstore, etc.
- **Number of recommendations:** The number of recommendations you would like to have.
- **BANDWIDTH:** This is a necessary parameter for the Mean Shift algorithm. The smaller the bandwidth, the denser and smaller the clusters that are identified. 0.002 is a reasonable value for most venue categories.

[12]:

```
category = "Fast Food Restaurant"
number_of_reccommendations = 5
BANDWIDTH = 0.002
```