# Lost on Purpose in Budapest

IBM Data Science Capstone Project
Gabor Pacsirszky
2019-October-31

**Introduction where you discuss the business problem and who would be interested in this project.**

Being in control of where we are and where we are heading is the norm in most part of our life. I am offering a bit of excitement, some room for spontaneity and the element of surprise - get lost in a Budapest neighborhood of your choice.

I will create choropleth maps that will help with the initial oriantation. So if e.g. you are a fan of Cafes, you will have a map of Budapest neighborhoods, colored based on the distribution of Cafes. If you want to have some nightlife experience, a map will be available about the distribution of bars/nightclubs, but exact spots will not appear. What exactly I choose will be based on data availability and quality.

My target group: Tourists and locals who enjoy exploring cities on foot, and want to open up for an experience that is not planned/controlled in great detail before it happens. Forget targeted places and reviews this time, just take a look at the map and off you go. Feel the excitement of soon bumping into an unknown representative of the theme you selected.

**Data where you describe the data that will be used to solve the problem and the source of the data.**

I have already explored data availability and realized that this is probably the most difficult part of this project. Budapest neighborhoods are a sub-territory of its districts and even district level data is hard to find in the required format, as this project rightly says:
https://medium.com/starschema-blog/draw-a-map-of-the-districts-of-budapest-using-the-overpass-api-of-openstreetmap-and-python-bd0417469935

I will use osm (openstreetmap.org) data. First I learn their query language a bit to be able to fetch what I want. It is possible to create a connection with their API from within jupyter notebook but it is not possible to get GEOJSON format this way, therefore I will just export it from overpass-turbo.eu which provides such functionality. Once I drew the bounds of the neighborhoods in a folium map I will query above described feature elements belonging to the neighborhoods. I will first try this in foursquare (will check foursquare for bounds too again), if this is not possible I will rely on OSM.

**Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.**

As expected, the main challenge in this project was to find the data that I needed, and to structure it in a way that is a valid input for folium choropleth maps.

Folium choropleth map needs **two data sets**, one defining the **areas**, and another that has the **quantities** for the desired feature in the areas.

The areas are called multipolygons which are defined by lat-lon coordinates, ordered specifically so that when dots are connected the desired shape appears. As much as I checked, it was not possible to request this data from Foursquare API with my account. I knew about OpenStreetMaps (OSM) and was rightly hoping that this information can be retrieved from them. Took a long time to find out how exactly, had to familiarize with their own query language: https://wiki.openstreetmap.org/wiki/Overpass_API/Language_Guide

The basic components of OSM are: nodes, ways, relations. Nodes are the atomic part, ways are built from nodes and relations are usually a group of ways and nodes: https://wiki.openstreetmap.org/wiki/Elements

A great help for OSM testing and playing around is overopass turbo where any query can be run and result is shown.

I needed a GeoJson file for the areas. GeoJson file format is not available in Overpass API so I ended up downloading it from overpass turbo and uploading it in Jupyter Notebook.

Luckily, quantities data was possible to get in Json format from within Jupyter Notebook through the API.

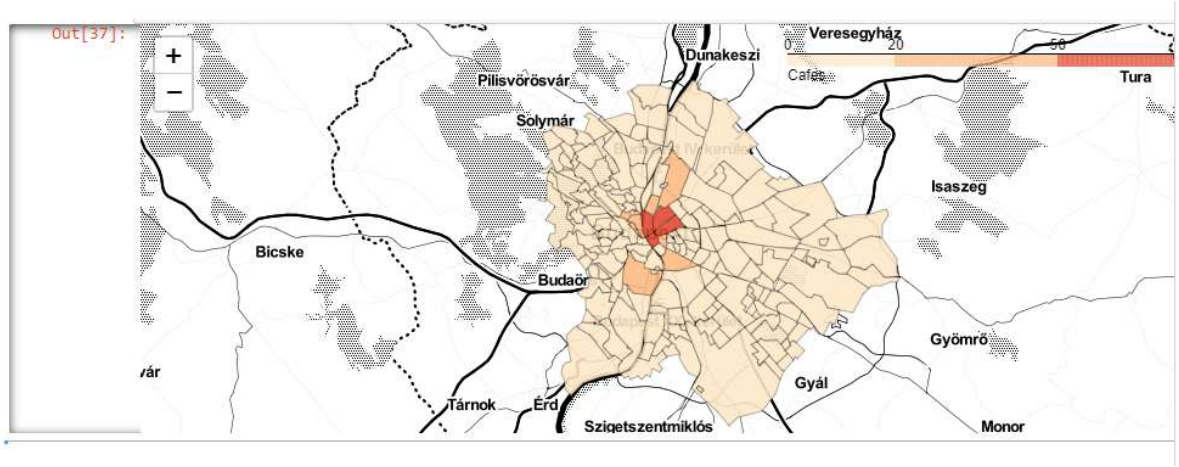
I haven't used machine learning or any statistical testing since data set was small enough to understand just by looking at it. A potential next step of this project could be to bring in other features and analyse their relationship, e.g.:

- tourist traffic and cafes
- pace of life and cafes (more outdoor cafes ?=?  more relaxing environment)
- and an always applicable factor: **weather** and café consumption.

This time though the outcome of my project is just a simple choropleth map.

**Results section where you discuss the results.**

The result is this map:




**Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.**

I created three main buckets of neighborhoods in the map:

- light-yellow
- orange
- reddish

The concentration in the city center with 2-3 neighborhoods having more than 50 Cafes each is not a surprise really (reddish).
More surprising is the very small number of Cafes in most of Budapest Neighborhoods (below 20, often 0). This group is becoming irrelevant from the project's perspective, people will not visit these places because they would not bump into any Cafe there (light-yellow).

**Conclusion section where you conclude the report.**

I was hoping to offer some areas which are currently less well-known for tourists but worth to explore. There are a few areas like that, colored in orange, meaning 20-50 cafes per neighbourhood. I would warm-heartedly recommend these places for my target group, there is a different vibe there, more authentic, not specialized on tourists (orange).
Unfortunately, Github is not rendering folium maps, so I can show it only as a picture, making it lose a few functionalities such as zooming.