**Introduction/Business Problem**

Budapest is the capital of Hungary and the tenth-largest city in the European Union by population within city limits. The city has an estimated population of 1,7 MM distributed over a land area of about 525 square kilometres (203 square miles). Budapest consists of 23 districts, 6 in Buda, 16 in Pest and 1 on Csepel Island between them. I've been living in Budapest for a couple of years now so that's why it's my choice for a project. Budapest is one of the most popular Central Europe's attractions both for tourists and for business owners. Millions of tourists visit the city every year, international students come to study and a lot of major corporations has branches in Budapest - considering all above the real estate market is very competitive and it's hard to find a perfect place to rent. To approach this problem I'm going to cluster districts according to the various social factors and create a map showing the results.

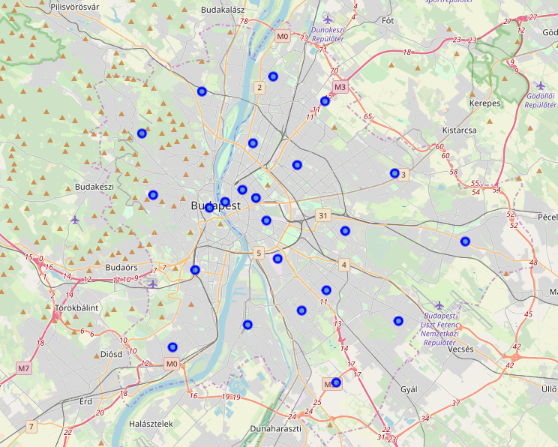
**Data**

To collect data for my research I used the following sources:

* Foursquare API to gather venues data in Budapest
* I created a dataset in GitHub repository that contains center points coordinates of each District obtained from Openstreetmap Nominatim
* json file with Budapest boundaries created from the polygon shapefile containing the first-level administrative divisions of Hungary

**Methodology**

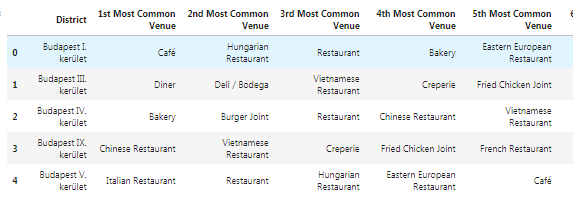
1. Budapest districts were visualized using Folium library showing boundaries of each district;

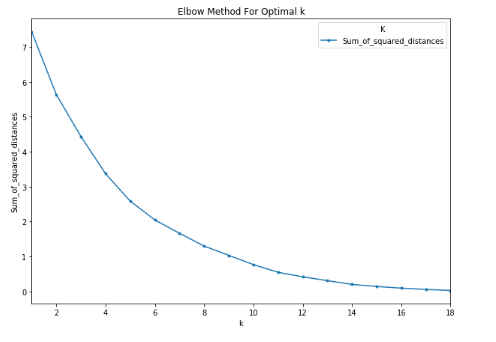


1. I explored districts using Fourquare API. To filter my selection I requested venues under "Food" category only. As I was using Foursquare API for learning purposes my search was limited by 500 venues and 500 m. radius. The returned data consisted of 297 venues in 51 categories. No venues were returned for some of the districts and those Districts were removed from the further analysis. This didn't necessary mean that there were no venues in the given districts as my search was a) based on the center point coordinates b) limited by 500 venues c) limited by 500 m radius;



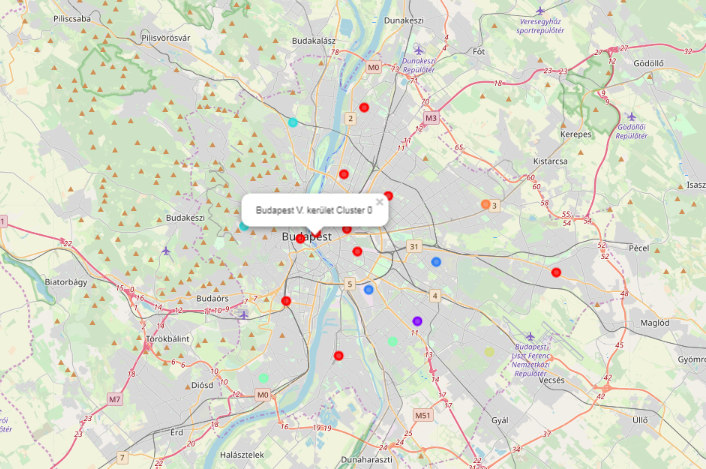
1. Based on Foursquare API results I created a dataset showing top 10 venue categories for each district. With the elbow method I identified the right number of clusters and then using a simple unsupervised machine learning algorithm K-mean all districts were divided in 7 clusters.



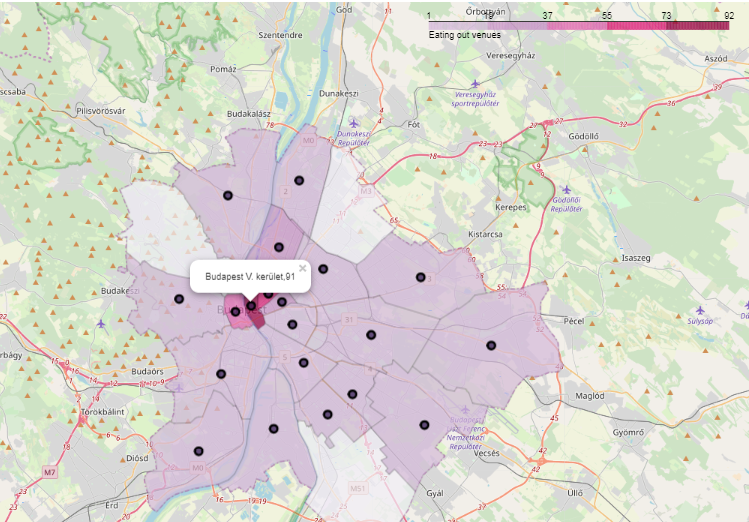


**Results**

1. I created a map to visualize clustered districts. According to the clustering results the city center is in the same cluster with more or less similar selection of venues and there are 6 small clusters of 1-2 districts with "abnormal" selection of venues;



1. To add even more clarity to my search I created a choropleth map showing a total number of venues and a name for each district. It's visible that the number of venues decrease as we move away from the city center.



**Discussion**

Even taking into consideration that my analysis was limited by Foursquare API capacity, center point coordinates usage and only one method to determine the optimal number of clusters for k-means clustering the results I got are still very useful both for personal and for business purposes. It is possible to expand original search or combine it with other factors (e.g. rent prices).

**Conclusion**

This type of analysis is beneficial for potential investors, city newcomers, tourists and business owners. The possibilities of usage are endless. What is more important is that the data used for this research is not static making possible trend analysis and various monitorings.