IBM Data Science Capstone Project

Picking the right location for a new restaurant in Budapest

Gábor László Kőrös

March 7, 2020

1. Introduction

The IBM Data Science Professional certificate course on Coursera concludes with a Capstone Project. This project is about using data science toolset on a real-life problem and demonstrating the creation of value by applying the learned skills. This report presents this capstone project. The analysis was performed in Python.

2. Problem Definition

a. The Problem

For this project, I chose a hypothetical business problem. The question that we are trying to answer is the following.

A successful owner of multiple mid to high-end restaurants decided to open a new restaurant in Budapest, Hungary. Having visited the city many times in recent years, he couldn't disregard the big boom in gastronomy. He is keen on opening a new unit, which will focus on the European and Asian fusion kitchen.

Taking into account the price level at which the restaurant will operate, the intent is to find an optimal location in an area, where gastronomy is booming and which is easily accessible for tourists and for wealthier local citizens as well.

b. Assumptions and business logic

The assumption behind the analysis is that we can use unsupervised machine learning to create clusters of districts that will provide us with a list of areas for consideration for the restaurant. The intent is that the restaurant to be situated close to one of the gastronomical centres and touristic hotspots.

c. Audience

While here we are assuming a concrete business owner to whom we are addressing this report, but actually this restaurant owner can be treated as a persona and thus this analysis could be useful for a group of market players (restaurant owners).

3. Data

To perform this analysis, we will need the following data:

1. List of the districts of Budapest
2. Geo-coordinates of the districts in Budapest
3. Top venues of districts

List of districts will be obtained from Wikipedia.
(https://en.wikipedia.org/wiki/List_of_districts_in_Budapest)

Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.

Top venues data will be obtained from Foursquare through an API.

4. Methodology

a. Use of data and a high-level roadmap

After tidying up and exploring the data, we will apply the K-means machine learning technique for creating clusters of districts. We will use the silhouette score for choosing the optimal number of clusters.

b. Analysis

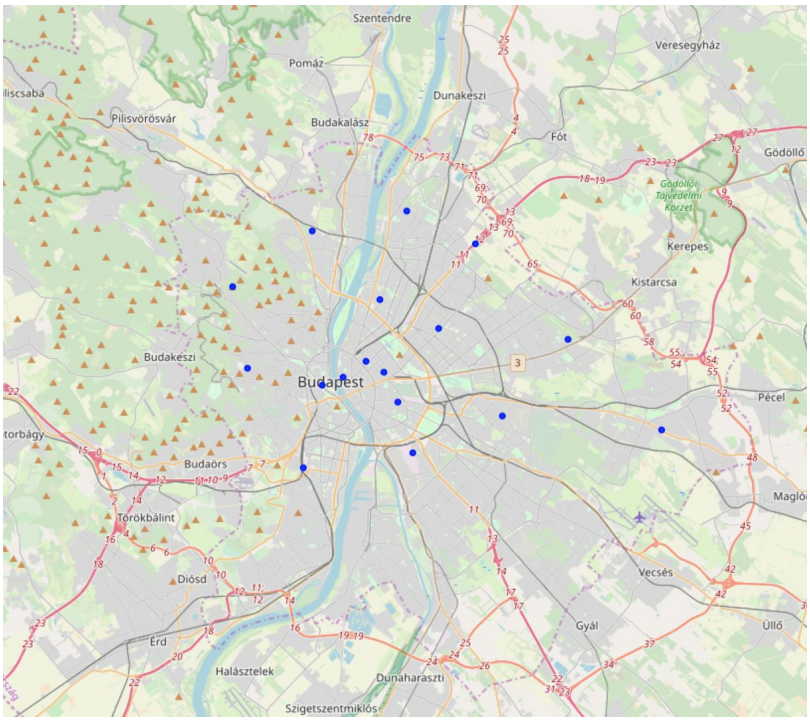
i. Data Preparation and exploration

As part of preparing the data, we start by creating a list of districts in Budapest and add the geo-coordinates of each district to this table. That is done by first importing a list of districts and then using this list and geocode python library, we add the latitude and longitude coordinates to each district. After performing this task, we get the following table that we use in pandas dataframe format.

	District	Name	Latitude	Longitude
0	I. kerület	Várkerület ("Castle District")	47.496994	19.034263
1	II. kerület	-	47.542471	18.972903
2	III. kerület	Óbuda-Békásmegyer ("Old Buda-Békásmegyer")	47.568691	19.027668
3	IV. kerület	Újpest ("New Pest")	47.577779	19.093164
4	V. kerület	Belváros-Lipótváros ("Inner City - Leopold Town")	47.500336	19.048971
5	VI. kerület	Terézváros ("Theresa Town")	47.508077	19.064426
6	VII. kerület	Erzsébetváros ("Elisabeth Town")	47.502627	19.077243
7	VIII. kerület	Józsefváros ("Joseph Town")	47.488755	19.086433
8	IX. kerület	Ferencváros ("Francis Town")	47.465070	19.096752
9	X. kerület	Kőbánya ("Quarry")	47.482405	19.158975
10	XI. kerület	Újbuda ("New Buda")	47.458334	19.021351
11	XII. kerület	Hegyvidék ("Highlands")	47.504800	18.982815
12	XIII. kerület	Angyalföld-Újlipótváros-Vizafogó ("Angel's Fie...	47.536804	19.074199
13	XIV. kerület	Zugló	47.523004	19.114513
14	XV. kerület	Rákospalota-Pestújhely-Újpalota	47.562714	19.140218
15	XVI. kerület	-	47.518266	19.204295
16	XVII. kerület	Rákosmente	47.475693	19.268780

There are 23 districts in Budapest, but due to technical issues with the geocode tool, it failed to provide an output for the whole list, but only for 17 districts. Thus, going forward this list of 17 districts will be used in the analysis. This decision is also very unlikely to have a material effect on the outcome of the analysis as districts from 18-23 are typically outskirts and thus they are not candidates in the race for choosing the target location.

In the next step, we create a visual representation of how the districts are situated in Budapest. For this, the folium library was used.



In the next step of the analysis, the districts were explored in greater detail. It means venues were collected for each district via Foursquare API. The data from Foursquare is received in json format. After arranging the data, we have up to 100 venues for each district. Venues are collected within a radius of 1000 meters from the point of district coordinates. The collected and arranged data looks like this. The following table shows some venues from the first district.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	I. kerület	47.496994	19.034263	Stand25 Bisztró	47.497673	19.032679	Bistro
1	I. kerület	47.496994	19.034263	Tabán Kinotéka	47.495818	19.034303	Indie Movie Theater
2	I. kerület	47.496994	19.034263	Budavári Mikve	47.498546	19.035846	Historic Site
3	I. kerület	47.496994	19.034263	Szelence Café	47.497767	19.031901	Café
4	I. kerület	47.496994	19.034263	Dísz tér	47.499100	19.036163	Plaza

We can check how many venues have been collected for each district. The following table gives that summary.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
I. kerület	100	100	100	100	100	100
II. kerület	43	43	43	43	43	43
III. kerület	28	28	28	28	28	28
IV. kerület	39	39	39	39	39	39
IX. kerület	14	14	14	14	14	14
V. kerület	100	100	100	100	100	100
VI. kerület	100	100	100	100	100	100
VII. kerület	100	100	100	100	100	100
VIII. kerület	85	85	85	85	85	85
X. kerület	28	28	28	28	28	28
XI. kerület	48	48	48	48	48	48
XII. kerület	31	31	31	31	31	31
XIII. kerület	90	90	90	90	90	90
XIV. kerület	70	70	70	70	70	70
XV. kerület	16	16	16	16	16	16
XVI. kerület	24	24	24	24	24	24
XVII. kerület	12	12	12	12	12	12

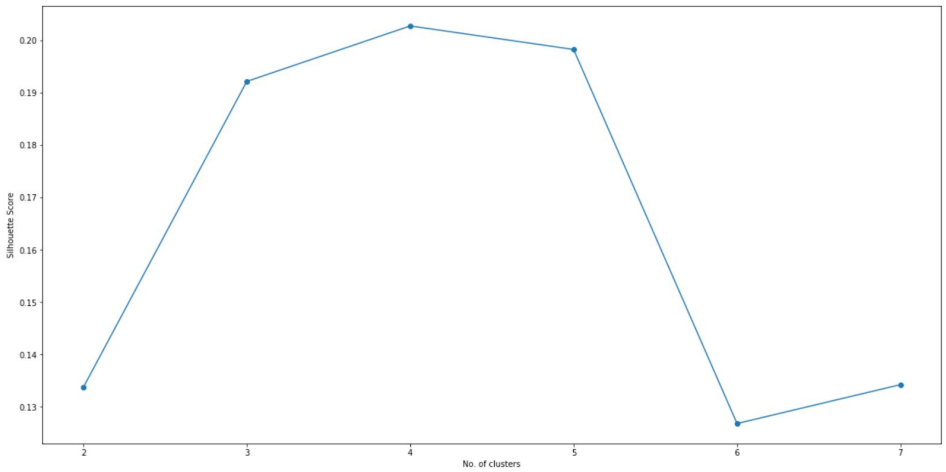
For analysing the districts, we focus on venue categories. For that purpose, we use the one-hot encoding. This creates dummy variables for categories so the data set could be used for machine learning.

After performing manipulations with the dataset, we get the following table, which shows the top ten most common venues for each district (first four shown in the table).

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	I. kerület	Café	Pub	Park	Coffee Shop	Plaza	Historic Site	Hotel	Hungarian Restaurant	Scenic Lookout	Playground
1	II. kerület	Grocery Store	Pizza Place	Tram Station	Park	Smoke Shop	Bus Stop	Gym	Forest	History Museum	Shopping Mall
2	III. kerület	Bus Stop	Grocery Store	Train Station	Dessert Shop	Eastern European Restaurant	School	Clothing Store	Deli / Bodega	Department Store	Yoga Studio
3	IV. kerület	Bus Stop	Soccer Field	Park	Hotel	Food & Drink Shop	Burger Joint	Bus Station	Pharmacy	Café	Soccer Stadium

ii. Clustering

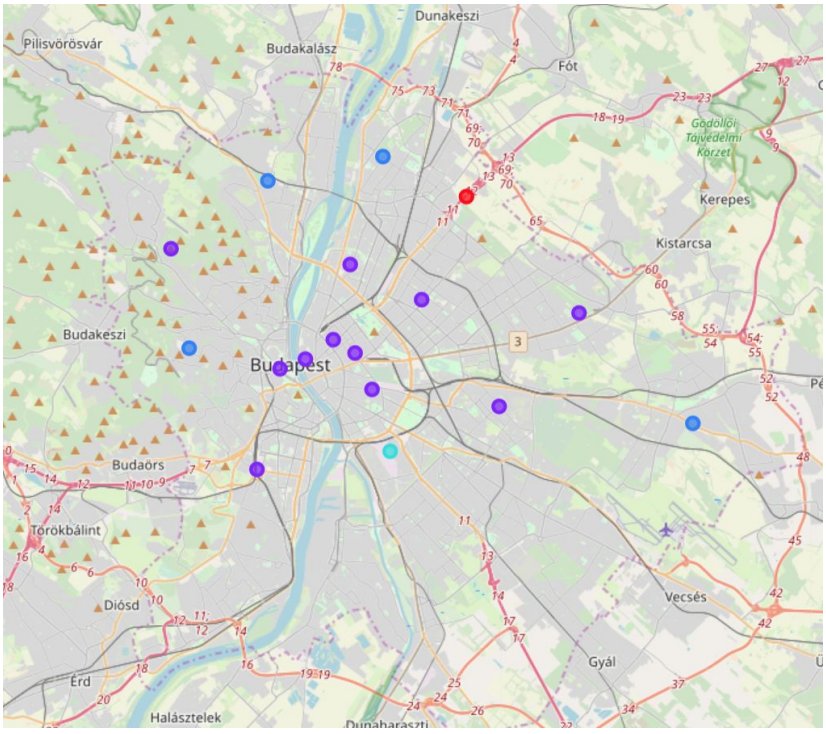
Now that we have the dataset ready, we perform clustering. For this, unsupervised machine learning technique will be used based on K-means. For K-means clustering, we need to decide on the number of clusters that we want to use. To avoid the trial and error approach, the silhouette score was used. The following graph shows the silhouette scores for a range of clusters variations.



From the graph, we can read that the optimal number of clusters to use is 4 (where the score is the highest). In the next step, we run the K-means clustering algorithm with the parameter of 4 as the number of clusters. When done, we add the cluster labels to the dataset. We get the following table.

	Neighborhood	Name	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	I. kerület	Várkerület ("Castle District")	47.496994	19.034263	1	Café	Pub	Park	Coffee Shop	Plaza	Historic Site	Hotel	Hungarian Restaurant	Scenic Lookout	Playground
1	II. kerület	-	47.542471	18.972903	1	Grocery Store	Pizza Place	Tram Station	Park	Smoke Shop	Bus Stop	Gym	Forest	History Museum	Shopping Mall
2	III. kerület	Óbuda-Békásmegyer ("Old Buda-Békásmegyer")	47.568691	19.027668	2	Bus Stop	Grocery Store	Train Station	Dessert Shop	Eastern European Restaurant	School	Clothing Store	Deli / Bodega	Department Store	Yoga Studio
3	IV. kerület	Újpest ("New Pest")	47.577779	19.093164	2	Bus Stop	Soccer Field	Park	Hotel	Food & Drink Shop	Burger Joint	Bus Station	Pharmacy	Café	Soccer Stadium
4	V. kerület	Belváros-Lipótváros ("Inner City - Leopold Town")	47.500336	19.048971	1	Hotel	Hungarian Restaurant	Restaurant	Coffee Shop	Italian Restaurant	Plaza	Modern European Restaurant	Salad Place	Dessert Shop	Sandwich Place
5	VI. kerület	Terézváros ("Theresa Town")	47.508077	19.064426	1	Coffee Shop	Pizza Place	Italian Restaurant	Thai Restaurant	Bar	Beer Bar	Theater	Indian Restaurant	Hungarian Restaurant	Bakery
6	VII. kerület	Erzsébetváros ("Elisabeth Town")	47.502627	19.077243	1	Hotel	Coffee Shop	Bar	Restaurant	Hungarian Restaurant	Beer Bar	Pizza Place	Gastropub	Burger Joint	Dessert Shop
7	VIII. kerület	Józsefváros ("Joseph Town")	47.488755	19.086433	1	Hotel	Park	Coffee Shop	Burger Joint	Chinese Restaurant	Bakery	Pub	Vietnamese Restaurant	Comedy Club	Bistro
8	IX. kerület	Ferencváros ("Francis Town")	47.465070	19.096752	3	Restaurant	Tram Station	Train Station	Bus Station	Electronics Store	Fast Food Restaurant	Soccer Field	Office	Furniture / Home Store	Department Store
9	X. kerület	Kőbánya ("Quarry")	47.482405	19.158975	1	Tram Station	Bus Stop	Arts & Entertainment	Sporting Goods Shop	Brewery	Market	Fast Food Restaurant	Supermarket	Grocery Store	Gym
10	XI. kerület	Újbuda ("New Buda")	47.458334	19.021351	1	Bakery	Bus Stop	Platform	Bus Station	Dog Run	Gym	Smoke Shop	Pharmacy	Pub	Hungarian Restaurant
11	XII. kerület	Hegyvidék ("Highlands")	47.504800	18.982815	2	Bus Stop	Park	Playground	Trail	Platform	Bakery	Bus Station	Mountain	Food	Grocery Store
12	XIII. kerület	Angyalföld-Újlipótváros-Vizafogó ("Angel's Fie...	47.536804	19.074199	1	Coffee Shop	Pub	Park	Gym / Fitness Center	Grocery Store	Indian Restaurant	Chinese Restaurant	Electronics Store	Restaurant	Café
13	XIV. kerület	Zugló	47.523004	19.114513	1	Bus Stop	Gym / Fitness Center	Gym	Grocery Store	Chinese Restaurant	Bakery	Café	Spa	Pharmacy	Pizza Place
14	XV. kerület	Rákospalota-Pestújhely-Újpalota	47.562714	19.140218	0	Supermarket	Fast Food Restaurant	Toy / Game Store	Bus Stop	Gym	Rest Area	Clothing Store	Furniture / Home Store	Eastern European Restaurant	Food & Drink Shop
15	XVI. kerület	-	47.518266	19.204295	1	Park	Light Rail Station	Cupcake Shop	Bus Stop	Dessert Shop	Soccer Field	Smoke Shop	Shop & Service	Mexican Restaurant	Paintball Field
16	XVII. kerület	Rákosmente	47.475693	19.268780	2	Bus Stop	Cosmetics Shop	Carpet Store	Supermarket	Grocery Store	Gym	Bakery	Pet Store	Dessert Shop	Restaurant

Also, we can visualise the clusters on the map that we created earlier.



c. Limitations

The analysis has some limitations that should be taken into account.

1. The analysis is performed on 17 of the 23 districts in Budapest. That is due to technical limitations with geocoder.
2. The analysis is performed on a district level.
3. When collecting venues a 1000 meter radius is used around the centre coordinates of the districts. The number of collected venues is limited to 100 per districts.

5. Results

Understanding the Clusters

By looking at the cluster data, we can see that cluster 2 is the one that we are the most interested in.

1. Cluster 1

The first cluster (Cluster label 0) is an outer district where top gastronomy is not really represented (supermarket and fast food are in the top).

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
14	XV. kerület	Supermarket	Fast Food Restaurant	Toy / Game Store	Bus Stop	Gym	Rest Area	Clothing Store	Furniture / Home Store	Eastern European Restaurant	Food & Drink Shop

2. Cluster 2

Cluster 2 (Cluster label 1) is the biggest cluster, but this is where we see lots of gastronomy related venues (coffee shop, pizza place, Thai restaurant, beer bar, pub, modern European restaurant, etc..).

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	I. kerület	Café	Pub	Park	Coffee Shop	Plaza	Historic Site	Hotel	Hungarian Restaurant	Scenic Lookout	Playground
1	II. kerület	Grocery Store	Pizza Place	Tram Station	Park	Smoke Shop	Bus Stop	Gym	Forest	History Museum	Shopping Mall
4	V. kerület	Hotel	Hungarian Restaurant	Restaurant	Coffee Shop	Italian Restaurant	Plaza	Modern European Restaurant	Salad Place	Dessert Shop	Sandwich Place
5	VI. kerület	Coffee Shop	Pizza Place	Italian Restaurant	Thai Restaurant	Bar	Beer Bar	Theater	Indian Restaurant	Hungarian Restaurant	Bakery
6	VII. kerület	Hotel	Coffee Shop	Bar	Restaurant	Hungarian Restaurant	Beer Bar	Pizza Place	Gastropub	Burger Joint	Dessert Shop
7	VIII. kerület	Hotel	Park	Coffee Shop	Burger Joint	Chinese Restaurant	Bakery	Pub	Vietnamese Restaurant	Comedy Club	Bistro
9	X. kerület	Tram Station	Bus Stop	Arts & Entertainment	Sporting Goods Shop	Brewery	Market	Fast Food Restaurant	Supermarket	Grocery Store	Gym
10	XI. kerület	Bakery	Bus Stop	Platform	Bus Station	Dog Run	Gym	Smoke Shop	Pharmacy	Pub	Hungarian Restaurant
12	XIII. kerület	Coffee Shop	Pub	Park	Gym / Fitness Center	Grocery Store	Indian Restaurant	Chinese Restaurant	Electronics Store	Restaurant	Café
13	XIV. kerület	Bus Stop	Gym / Fitness Center	Gym	Grocery Store	Chinese Restaurant	Bakery	Café	Spa	Pharmacy	Pizza Place
15	XVI. kerület	Park	Light Rail Station	Cupcake Shop	Bus Stop	Dessert Shop	Soccer Field	Smoke Shop	Shop & Service	Mexican Restaurant	Paintball Field

3. Cluster 3

Cluster 3 (Cluster label 2) contains districts where public travel rated at the top, but behind that, parks and playgrounds are also present. These are mainly areas with family houses where people live, but not really the vibrant, lively part of the city.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	III. kerület	Bus Stop	Grocery Store	Train Station	Dessert Shop	Eastern European Restaurant	School	Clothing Store	Deli / Bodega	Department Store	Yoga Studio
3	IV. kerület	Bus Stop	Soccer Field	Park	Hotel	Food & Drink Shop	Burger Joint	Bus Station	Pharmacy	Café	Soccer Stadium
11	XII. kerület	Bus Stop	Park	Playground	Trail	Platform	Bakery	Bus Station	Mountain	Food	Grocery Store
16	XVII. kerület	Bus Stop	Cosmetics Shop	Carpet Store	Supermarket	Grocery Store	Gym	Bakery	Pet Store	Dessert Shop	Restaurant

4. Cluster 4

Cluster 4 (Cluster label 3) contains only one district. Here we see the restaurant category at the top, but behind that, it is about public transport.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	IX. kerület	Restaurant	Tram Station	Train Station	Bus Station	Electronics Store	Fast Food Restaurant	Soccer Field	Office	Furniture / Home Store	Department Store

6. Discussion and Recommendations

Based on what we learned about the clusters, we can advise the restaurant owner to consider the districts from cluster 2 as a potential location for the new restaurant. These are the districts where gastronomy is well represented and also hotels are frequent. These satisfy the two original criteria that the location should be in a gastronomical centre and in a location that is easily accessible for tourists.

7. Conclusion

This paper discussed the process of coming up with an answer for a hypothetical though real-life like business problem. The analysis was performed based on the toolset of data science and relied heavily on the use of Python and Python libraries such as Pandas, Scikit, Folium to name a few. Data was collected from a different type of sources and in different formats. For analysis, machine learning technique was used. The output of the analysis provided a thorough base for the recommendation for the business problem in question.

8. References

The Jupyter notebook of the analysis can be found on GitHub.
https://github.com/gaborkoros/Coursera_Capstone/blob/master/IBM%20Capstone%20Project%20%20final.ipynb