The Pizza in Prague

IBM Data Science Capstone Project

Tomasz Biały

August, 4th 2021

I. Problem understanding

Prague is capital of Czech Republic with population of 1.3M and it is one of the most visited cities in Europe by tourists. It has huge opportunities and prospects for opening a restaurant business. Our stakeholder is willing to open the pizza restaurant in Prague but choosing a location is not an easy task, since there are a lot of criteria that should be satisfied in order to achieve the highest revenue.
In this project I am going to find the most optimal place to open a pizza restaurant considering factors like the density of other restaurants and the density of specifically pizza restaurants in a city of Prague.

II. Data collection

In order to solve the business problem described previously, I am going to collect and analyze the following data:

- Names of districts in Prague. The data is collected by parsing the *Wikipedia* page of Prague using the *Beautiful Soup* package.

- Geographic location of districts. The data is collected using the *Nominatim* library to find the geographic coordinates by name and address (geocoding).

- Name, category and geographic location of the existing food venues in Prague. The data can be retrieved with *Foursquare* - a location data provider with information about venues and events within an area of interest that can be obtained through the API. The Foursquare uses a defined search radius to sweep and discover the restaurants within its range, however the response is limited to 100 venues per call, therefore the search location need to be spitted in smaller portions.

I am going to start with collection of names of the districts in Prague, then I can find out the geographical location of their centers using *Nominatim.* Having those, I can use *Foursquare* to explore the food venues at given coordinates within 1.5km search radius. If the response limit is reached, the search location is divided into 8 smaller ones with 0.75km search radius. Finally, the duplicate data that comes from overlapping search regions is dropped.

III. Exploratory data analysis

The total of 3286 food venues information has been obtained and stored in dataframe during the data collection (Figure 1). It would be useful to visualize the location of the obtained venues on the map of Prague. It is possible to observe that venues are more densely packed in the city center (Figure 2).

Venue_category	Venue_name	Venue_latitude	Venue_longitude
Vietnamese Restaurant	DuHa	50.098531	14.399085
Café	Kavárna Alibi	50.097803	14.396178
Indian Restaurant	Indian by Nature II	50.098503	14.402786
Gastropub	U Veverky	50.098991	14.402154
Pelmeni House	Bistro Váleček	50.099056	14.402783
...	...	...	...
Café	Trdlocafé	50.108025	14.581490
Café	Café Coffee Day Emporio	50.108480	14.584162
Restaurant	Sconto restaurace	50.111445	14.582952
Pizza Place	Pizzeria Gattino	50.008910	14.427877
Turkish Restaurant	Kebab House Modřany	50.008784	14.427084

Figure 1. Dataframe with venues information

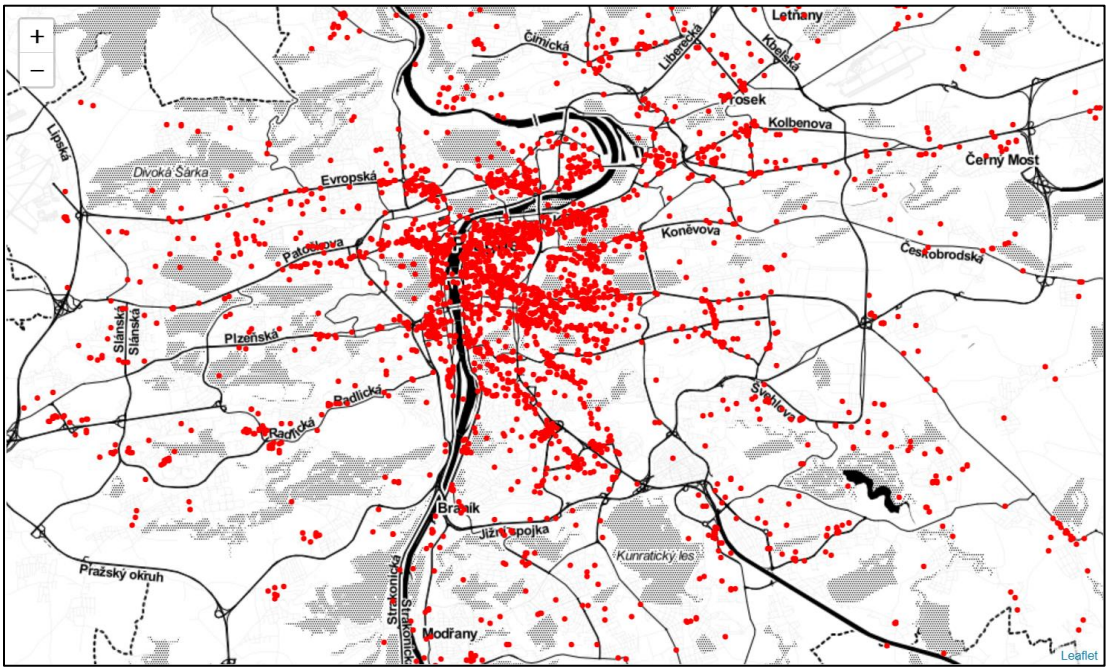

Figure 2. Food venues distribution on the Prague map.

Another thing that is worth checking is the popularity of specific types of the food venues (Figure 3). The Café is the most numerous type of venue in the city. The pizza restaurants are also quite popular and they are 6% of all food venues. Vietnamese restaurant are also popular venues since the Vietnamese people are the third-largest ethnic minority in Czech Republic and their cuisine is very unique.

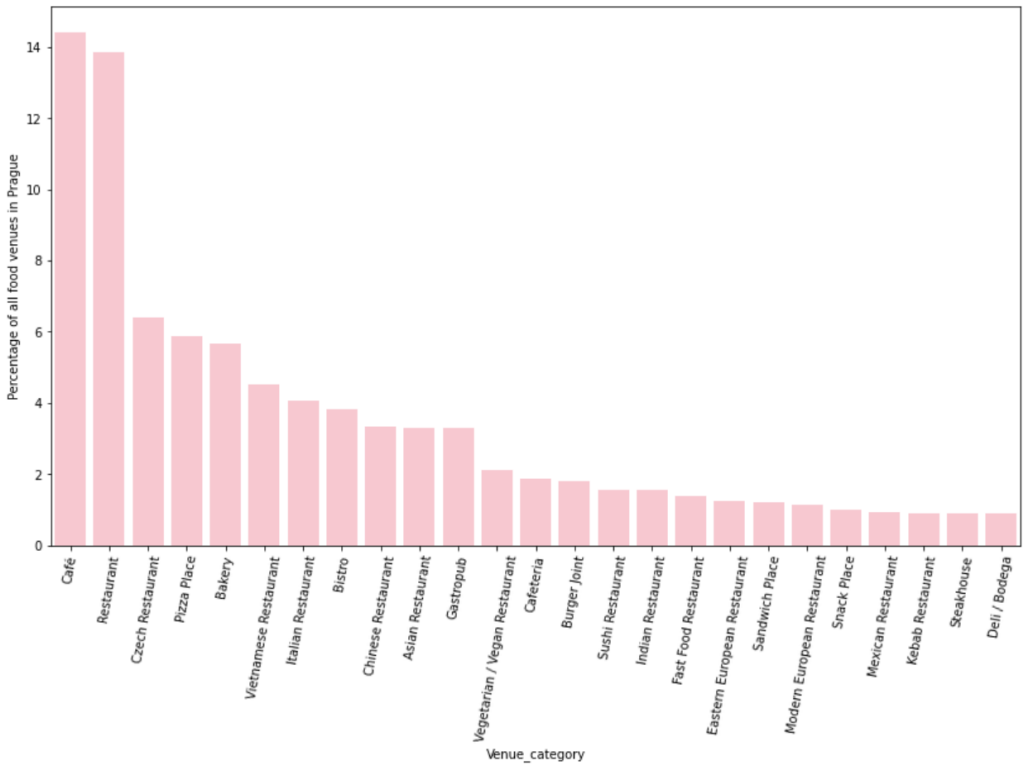

Figure 3. Popularity of most numerous food venues types

IV. Analytical method

Based on the collected data, I have sufficient information to try to solve the business case. I will cluster the locations of existing restaurants and food venues using DBSCAN clustering algorithm to find out where are the popular spots for eating out. The parameters that will be used for model are:

- Minimum number of samples = 25
- Epsilon = 200 meters

That mean that the popular spot will be considered if there are at least 25 restaurants within 200 meter range from each place.

In the next step, I am going to find out in which clusters the pizza restaurants are underrepresented. That will allow me to choose a popular spots that is not occupied by the competition. Using this results, our stakeholder can take the necessary decision about opening the pizza restaurant.

IV. Results and discussion

The DBSCAN model returned 18 clusters that are considered as popular spots for eating out. The each cluster was assigned an unique color and has been plotted on the map (Figure 4 and Figure 5). The outliers – restaurants that do not belong to any cluster were colored in grey.

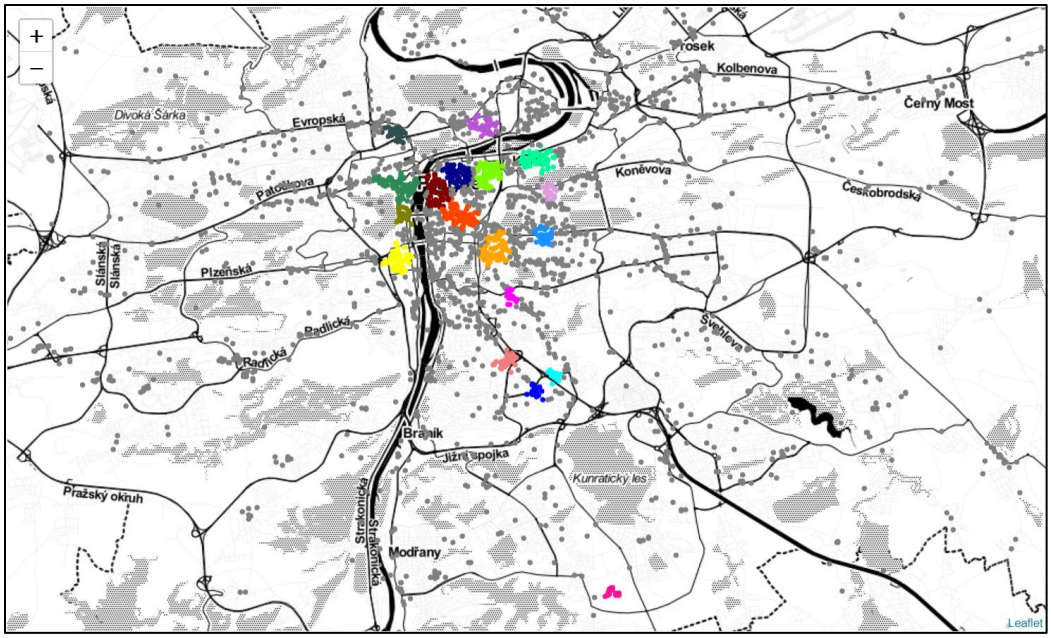

Figure 4. The cluster of food venues – popular spots for eating out

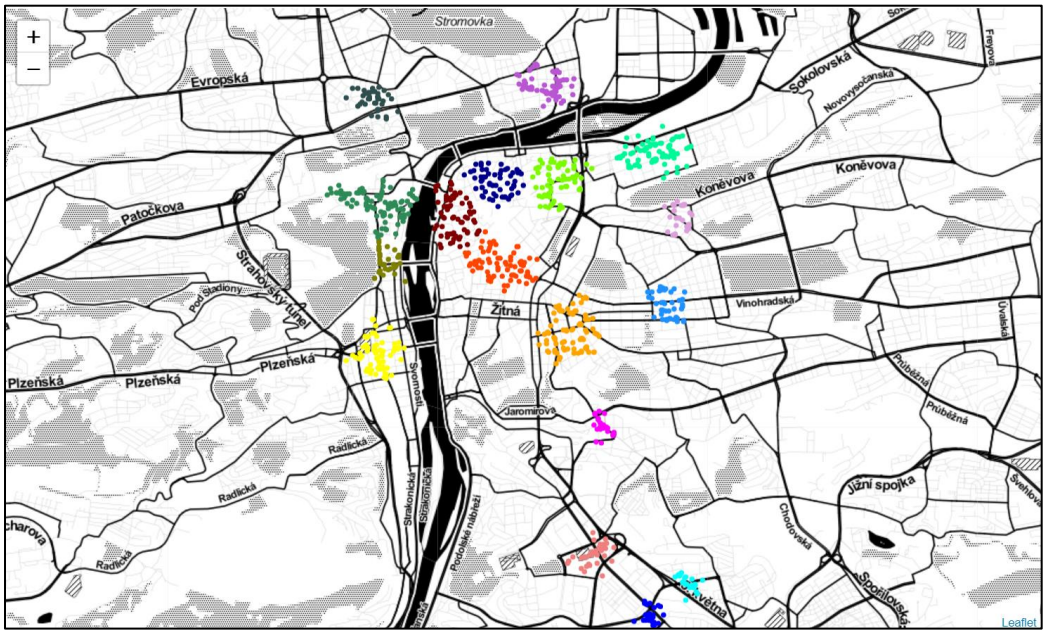

Figure 5. The cluster of food venues – popular spots for eating out (zoomed-in and outliers removed)

In the next step, I will check how many pizza restaurants are in each cluster in order to see how big is the competition there (Figure 6).

```
Cluster #1      Total places: 104      Pizza restaurants: 1.0%
Cluster #6      Total places: 84       Pizza restaurants: 4.8%
Cluster #10     Total places: 77       Pizza restaurants: 2.6%
Cluster #8      Total places: 75       Pizza restaurants: 0.0%
Cluster #5      Total places: 74       Pizza restaurants: 5.4%
Cluster #7      Total places: 72       Pizza restaurants: 4.2%
Cluster #2      Total places: 68       Pizza restaurants: 1.5%
Cluster #9      Total places: 64       Pizza restaurants: 4.7%
Cluster #4      Total places: 59       Pizza restaurants: 5.1%
Cluster #0      Total places: 42       Pizza restaurants: 2.4%
Cluster #13     Total places: 42       Pizza restaurants: 0.0%
Cluster #15     Total places: 41       Pizza restaurants: 2.4%
Cluster #3      Total places: 37       Pizza restaurants: 2.7%
Cluster #17     Total places: 36       Pizza restaurants: 2.8%
Cluster #12     Total places: 31       Pizza restaurants: 3.2%
Cluster #14     Total places: 27       Pizza restaurants: 22.2%
Cluster #11     Total places: 25       Pizza restaurants: 0.0%
Cluster #16     Total places: 25       Pizza restaurants: 4.0%
```

Figure 6. Amount of pizza restaurants in identified clusters

In order to avoid competition from businesses that serve same type of food, the recommended spots will be the ones that have the least pizza restaurants. Therefore, the clusters that contain less than 1% of pizza venues are selected:

- Cluster #1 with 1.0% pizza restaurants,
- Cluster #8 with no pizza restaurants,
- Cluster #13 with no pizza restaurants,
- Cluster #11 with no pizza restaurants.

Finally, the recommended spots for opening a pizza restaurant are shown in the map below:

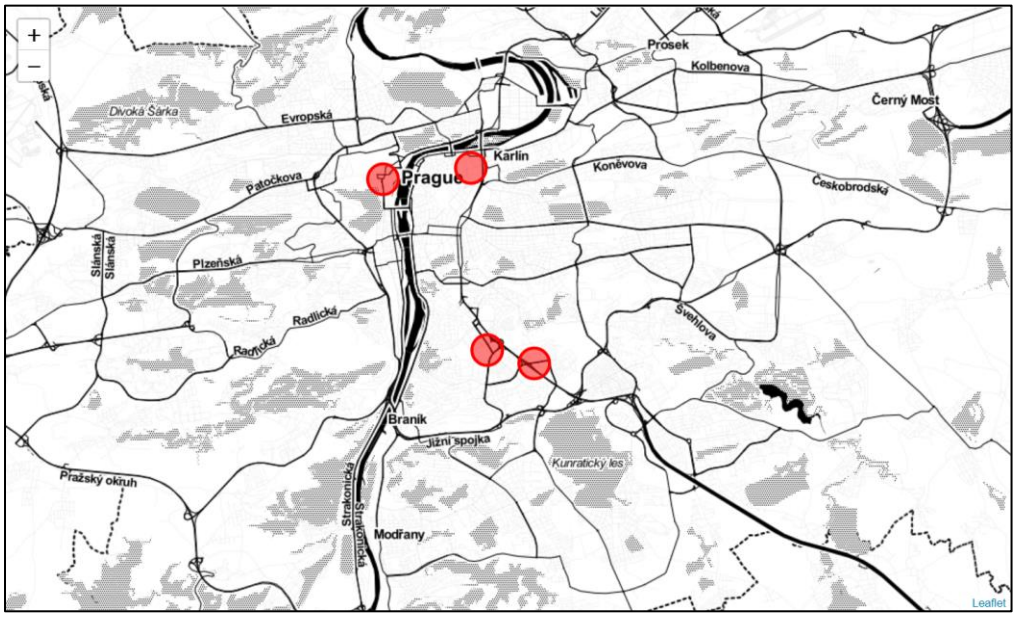

Figure 7. The recommended spots for opening the pizza restaurant

V. Conclusions

The basic data analysis was performed to identify the most optimal places to open the pizza restaurant in Prague. The clustering helped to highlight popular places where people go to eat out and categorizing the venues discarded places with the competition. Finally, the Malostranské náměstí, Náměstí Republiky, Pankrác and Michle areas were chosen as the most attractive options for setting a business.