The Czech Capital vs The Second Biggest City

Introduction

In this project two biggest Czech cities will be compared. Prague, the capital with over one million inhabitants, is supposed to be an international metropolitan city. Venues there should be numerous, often large with global appeal.

Brno is the second biggest city in Czechia with over 375 000 inhabitants. It is in the east-south region in the middle if winery region. In nature, it should be calmer and friendlier, big student community has a noticeable impact on Brno's culture. Venues there are usually smaller and more alternative.

The main reason to open a place in Prague is its size and international environment. On the other hand, experimental bars and other places tend to open often in Brno, because it is big enough to find and audience but small enough to spread the word.

With such different advantages and specifics to both cities, it is interesting to see if clustering finds differences between those two cities. If there indeed is a different structure of venues, an entrepreneur might consider the cities' spirits and preferences when opening a new establishment.

This project has two audience groups in mind. Firstly, there is a little ongoing internal dispute between the capital and Brno about which city is better. In this analysis, I would like to address differences found. Second audience are investors. If there are notable differences between those two cities, one should consider if their plan fits into the vibe of the city or if it is the project the neighbourhood was missing all along.

Data

This project will be based on previous geolocation project. First, a list of Prague's and Brno's neighbourhoods with postal codes will be obtained. To ensure uniqueness of locations, each postal code will be represented once.

Next, longitude and latitude will be established based on those postal codes. A list of postal codes with geolocation figures was also found in case there are any troubles during establishing centres of chosen neighbourhoods.

With that geolocation information, Foursquare will be utilised to find the closest venues to the neighbourhoods' centres such as bars, parks or bus stops. Based on those data, cities will be put together based on their similarities and divided on their dissimilarities. In the end of the report, each cluster will be described by its top mentioned values and I will attempt to determine, which city has better opportunities or offers better social life.

Data description

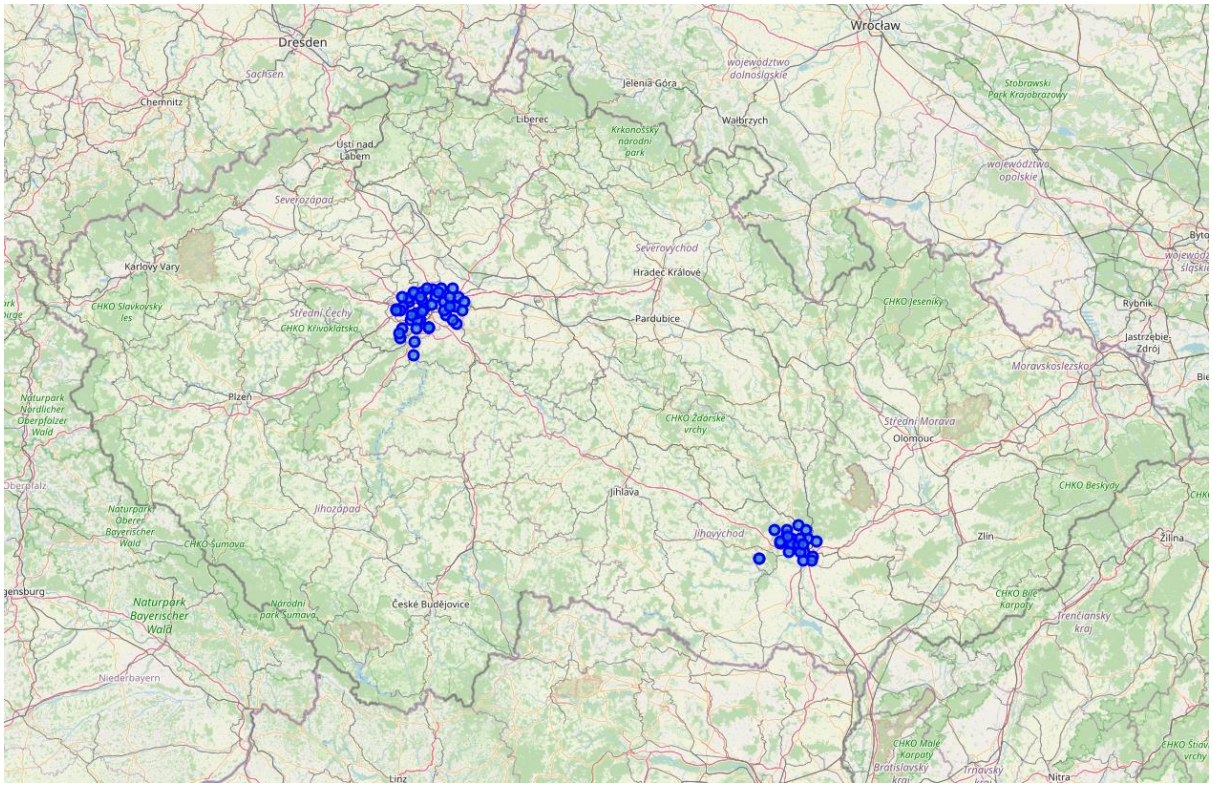
Cleaned dataset with latitude and longitude information which does not include duplicates and consisted of 90 observations, 30 postal codes from Brno and 60 postal codes came from Prague.

After initial visualisation, it was clear that some of the observations were too far of from both Prague and Brno. Alternative data sources did not improve the situation, it seems that data quality about Czech cities' administrative parts is not as high as it was in Canada or the USA.

Observations which were too far away from the cities were therefore deleted from the dataset. In case of Prague, dots seemed to be reasonably spaced around the city. This, however, was not the

case of Brno which had close to no data points in the city centre itself. Therefore, seven data points were randomly manually added to main parts of the city.

This data set looked reasonably and could serve as a basic indicator about neighbour situation in both cities. It then consisted of 72 observations – 24 from Brno region and 48 Prague and its surroundings. The final visualization of the data set can be found below.



Picture 1: Data points

Those data points were then used to find according information from Foursquare. Radius of 500 metres from the data centres was considered for this search as it was feasible to find as many venues in the area but create not many duplicates. Based on 72 data points, 585 venues were found in 165 unique categories.

Methodology

After locating the venues, data was transformed to desired encoding. Each row represented one city part with its top 15 values. First few rows of the table can be seen below. Because of readability of the picture, top 10 most common venues were displayed instead of all 15, however the whole analysis was conducted on those top 15 venues.

	Name	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	Benice	Trail	Train Station	Restaurant	Athletics & Sports	Bus Stop	Basketball Court	ATM	Monument / Landmark	Motel	Motorcycle Shop
1	Bohnice (část)	Bus Stop	Restaurant	Supermarket	Playground	Movie Theater	Hospital	Chinese Restaurant	Park	Soccer Field	Brewery
2	Bohunice	Food Court	Climbing Gym	Boat or Ferry	Bus Station	Bike Rental / Bike Share	Electronics Store	Baseball Stadium	Golf Course	Baseball Field	Pool
3	Bosonohy	Beer Garden	ATM	Nature Preserve	Middle Eastern Restaurant	Monument / Landmark	Motel	Motorcycle Shop	Movie Theater	Museum	Music Venue
4	Braník (část)	Bus Stop	Hotel	Gym / Fitness Center	Soccer Field	Italian Restaurant	Motorcycle Shop	Electronics Store	Dog Run	Czech Restaurant	Pet Store

Picture 2: Top venues

The first method used was k-means which divides data into given number of clusters. Those clusters are not overlapping. To ensure that the global maximum was found, the process was tun multiple times.

The number of clusters for the analysis was chosen to be 7. As the data had originally no labels, it was not possible to base this number of any exact metric. Multiple analysis versions were conducted, and this amount seemed visually as the most feasible.

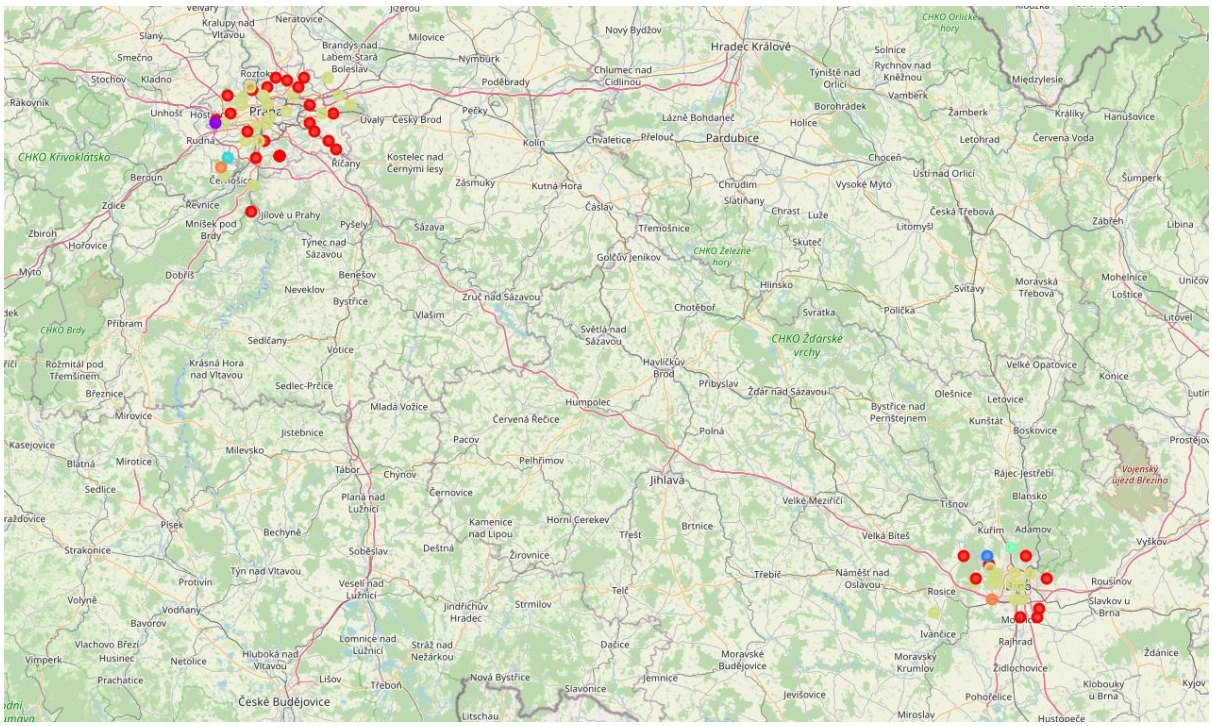
The second method was hierarchical clustering, where all observations begin as single clusters and then the closest two observations are connected into one new cluster. This process iterates until all observations belong to one cluster. Euclidean distances and Ward method were used to create clusters. Based on dendrogram, 4 clusters were chosen as the optimal number of clusters.

Results

In this section, results from both methods will be presented.

K-means

Based on this analysis, 7 clusters were found. It can be seen that similar structure of venues can be found in both Brno and Prague. Based on top venues in the most prominent clusters (yellow and red), it can be found that the main difference was found in the distance from the centre rather than between the cities.

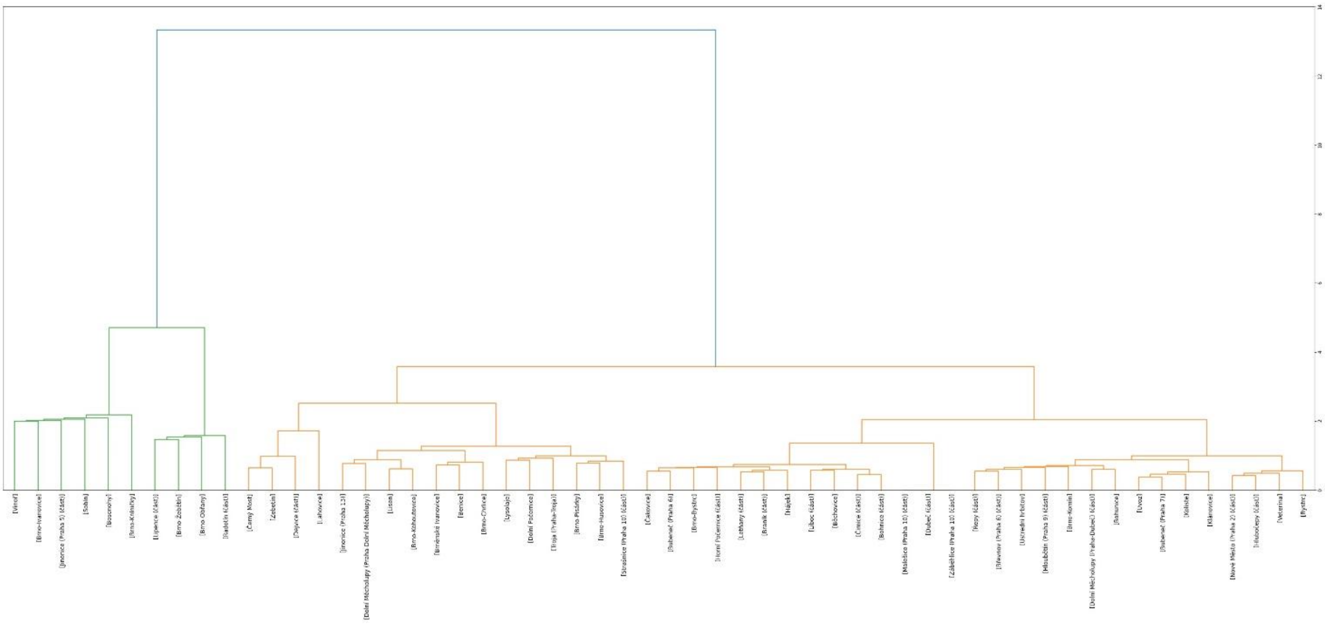


Picture 3: K-means map

Yellow cluster have the highest amount of restaurants, coffees and gyms in their top positions. Red clusters on the other hand have the highest amount of bus stops, hotels and restaurants.

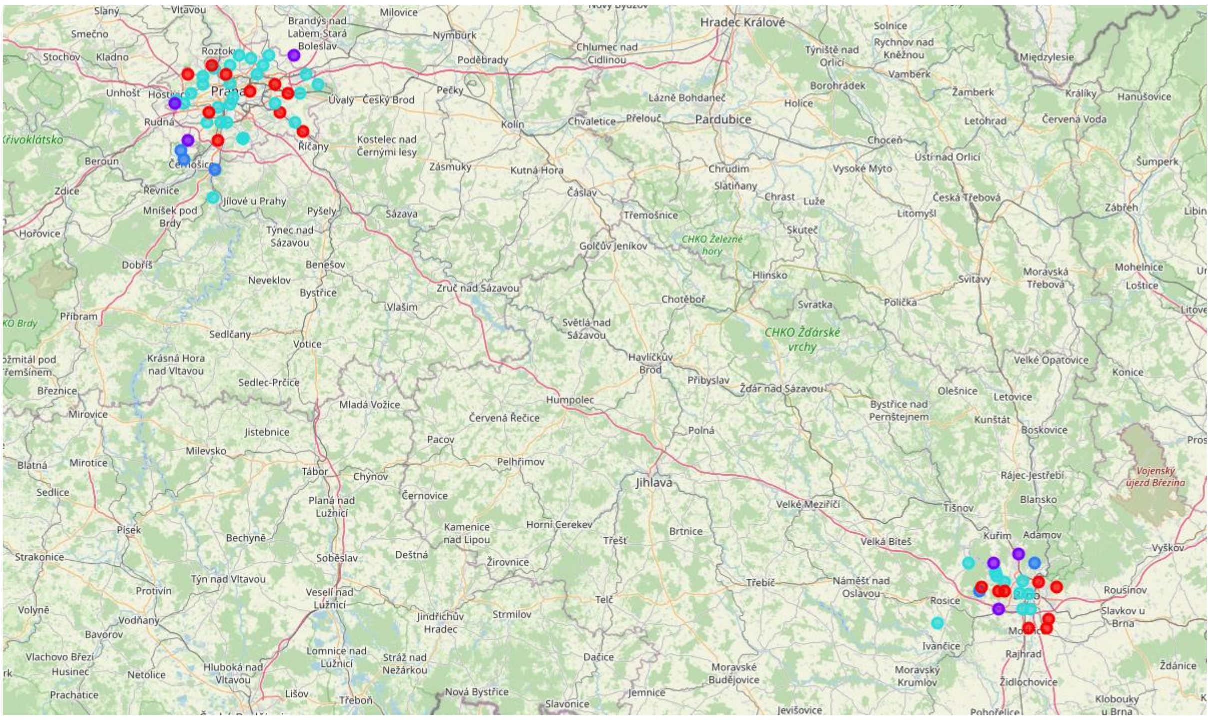
Hierarchical clustering

Based on dendrogram presented below, it was found out that the best amount of clusters is 4 as then the horizontal distances were rather small.



Picture 4: Dendrogram

Based on clusters presented in picture 5, we can again observe, that clusters tend to be more dependent on the location in the city rather than the city itself. The teal cluster's most common venues are bus stops, hotels, restaurants and supermarkets which mostly represents the busy inner city. The red cluster is then greatly represented by bust stops and stations, followed by offices and restaurants.



Picture 5: Hierarchical clustering map

Discussion

As can be seen in two approaches mentioned in results, the city itself did not have big effect on clusters created. The locations were assigned to clusters more based on the distance from the city center where we could observe different types of distinction into busy center and calmer resident and office areas.

Better results might be obtained by limiting venues based on proposed establishment type, such as Italian restaurant or coffee shop. The aim of this assignment was however to try to capture general feeling differences between those two cities.

Another factor that could cause issues was imperfect geolocation of both cities, especially of Brno. More feasible approach to capture the are better could be to omit postal codes and create a web of points with fixed distance among them on both cities.

Conclusion

This assignment tried to compare the capital and largest city of Czechia, Prague, and the second largest city, Brno. The aim was to discover venue differences in both cities to discover potentials for investors and to forever end disputes between those two cities.

The underlying theory was that both cities have different feeling to them. While Prague is the huge metropolitan, busy and international city, Brno is smaller with a large student and hipster communities and more experimental places.

Based on k-means and hierarchical clustering, it was found out that provided data were not sufficient to capture differences in venues in both cities. The prominent clusters in both cases were based on location withing the city and looked similarly in both Brno and Prague. Therefore, it was conducted that both cities are similar enough, all disputes should be abandoned, and investors need to use more specific models which limit venues to their desired industries. The competition in both cities is fiercer close to city center and a niche establishment could find big enough audience in both, while resident and office areas could offer space to expansion in more classical types.