Opening an International Restaurant in Bratislava: A prediction based on Neighborhood Clustering

Carlos Andres Pizarroso Troncoso

July 16th, 2021

1. Introduction

1.1. Background

Our current world became that globalized that the fact of going abroad for studies or work became more accessible for people. During the time a person spends abroad, one is able to try local food, which could not only be new and exotic, but also, very different from one's traditional cuisine. As a result, sometimes, people can miss their own kind of food. As a person who is living abroad, I believe that this could be of interest for anyone who wonders where it is recommendable to open an international restaurant in another country.

Based on the previous premise, I initially thought on working and developing a predictive model for my own country, Bolivia. However, due to the fact that Bolivia does not have a postal code (zip code) system, I decided to work with a country of personal interest which uses this postal code system: Slovakia.

1.2. Business Problem

One of the most important questions, if not the first one, for a future restaurant owner will definitely be "where should I open my restaurant?" In business terms, it is possible to reformulate this question as "where is the most convenient location for opening my restaurant?" By convenient I mean, to know the location where this restaurant will have a better impact, or, to know the location where the influx of people will be greater considering the different venues around. Certainly, this question can be answered after doing an extensive research in the desired area by observing and getting information about the types of restaurants there, the influx of people daily, and considering many other factors. Nonetheless, this method would probably take and consume big quantities of time, economic and human resources.

In the following report I will introduce a potential alternative to this. Instead of spending resources, why not using available data from trusted sources on the internet? By applying Data Analysis with Machine Learning techniques, it is possible to process and use this information to develop a potential solution for our problem, considering the previous points and finally, answering the business problem question for this case: **"where is the best place for opening an international restaurant in Bratislava?"**

2. Data Processing

2.1. Sources

I used three different datasets for building this model:

- **Slovak Streets and Postal Codes.**

This dataset is the most important one, I obtained it from the official Slovak Post website[1] and it contains detailed information (in Slovak language) of the streets in all Slovakia. Each street is associated with a Postal Code, a Postal Office location and a City where it belongs. The document is an Excel file which can be read by Jupyter Notebook.

- **Slovak Postal Codes Geographical Coordinates**

This data set contains information of the geographical coordinates (Latitude and Longitude) for each Postal Code. The document is an Excel file which I built after getting information from two different sources[2, 3].

- **Slovak Neighborhoods Details**

This data set was also built by myself and it is a subset from the first dataset. The Data Wrangling section will indicate more details about the reasoning behind the naming of Slovak neighborhoods. I must state this **is not** an official nomenclature and it was done only for simplifying the analysis of this work.

2.2. Data Wrangling

The first dataset, which was denominated as “postalcodes_sk.xlsx” was first read by the program and then displayed.

	DULICA	ULICA	PSC	DPOSTA	POSTA	POZNAMKA	OBCE
0	Banská	Banská	976 32	Badín	Badín	NaN	Badín
1	Borovicová	Borovicová	976 32	Badín	Badín	NaN	Badín
2	Družstevná	Družstevná	976 32	Badín	Badín	NaN	Badín
3	Hliny	Hliny	976 32	Badín	Badín	NaN	Badín
4	Krčméryho	Krčméryho	976 32	Badín	Badín	NaN	Badín

Fig. 1. First 5 rows displayed from postalcodes_sk.xlsx.

In Fig. 1, it is possible to see that there are 7 columns. From these 7 columns only “ULICA”, “PSC”, “POSTA”, and “OBCE” were used for this analysis. The columns “DULICA” and “DPOSTA” repeated the values of the columns “ULICA” and “POSTA” for security reasons, thus I decided to delete them. The column “POZNAMKA” means “Notes” and it was the column that contained the biggest number of empty elements. Because these notes about some streets were not relevant for the analysis, I also decided to delete this column.

“ULICA” means “street”, however, for the present work, I classified streets according to their respective Postal Code, those streets became Neighborhoods, thus “ULICA” column was treated as “Neighborhood”. “PSC” stands for “Postal Code”. “POSTA” means “Post Office”, every region in Slovakia has a post office, therefore I treated every Post Office as a Borough. Finally, “OBCE” means “City” and this column contains all names of cities and towns in Slovakia. After changing the names to what was described before, the following table was obtained:

	Neighborhood	PostalCode	Borough	City
0	Banská	976 32	Badín	Badín
1	Borovicová	976 32	Badín	Badín
2	Družstevná	976 32	Badín	Badín
3	Hliny	976 32	Badín	Badín
4	Krčméryho	976 32	Badín	Badín

Fig. 2. Dataset after renaming the columns and eliminating the irrelevant ones.

The next step was identifying the missing values. I built some functions which helped on check if the dataset contained any missing value and identify it. There were 123 rows out of 10547 which contained missing values and, from the 123 mentioned rows, 122 belong to small streets without their own Postal Code, because 122 barely represents 1.16% of the total data and because they lacked an important feature for analysis, I decided to exclude them from the dataset. Lastly, the remaining row had a missing value for the City column. After some research, I found out the street belongs to Bratislava city. Consequently, instead of eliminating this row, I decided to manually add the missing value to its respective column.

Until this point, the dataset contained information from all Slovak cities and towns. Because of the geographical information available for postal codes and for simplifying the analysis, I decided to work with Bratislava city uniquely. After selecting the streets which were located in Bratislava, the dataset was reduced to 2142 rows, however, many streets shared the same Postal Code. As a result, I grouped streets by postal codes, reducing the dataset to 56 rows. At this point, I decided to denominate each Neighborhood by the first street that appeared in its alphabetical order list. Doing this was possible thanks to the third dataset, named as “bratislava_neighborhoods.xlsx”. Fig. 3. shows what the dataset looked like at this stage.

	PostalCode	Neighborhood	Borough
0	811 01	Neighborhood 1 (Beblavého)	Bratislava 1
1	811 02	Neighborhood 2 (Bartókova)	Bratislava 1
2	811 03	Neighborhood 3 (Bartoňova)	Bratislava 1
3	811 04	Neighborhood 4 (Banícka)	Bratislava 1
4	811 05	Neighborhood 5 (Anenská)	Bratislava 1

Fig. 3. Reduced dataframe showing the representative street for each neighborhood.

Afterwards, I merged this dataset with the information obtained from the second dataset, “bratislava_pcc.xlsx”. I re-checked for any missing value and fixed two instances which had typing-style mistakes. The resulting dataset, ready for analysis can be seen in Fig. 4.

	PostalCode	Neighborhood	Borough	Latitude	Longitude
0	811 01	Neighborhood 1 (Beblavého)	Bratislava 1	48.14527	17.10989
1	811 02	Neighborhood 2 (Bartókova)	Bratislava 1	48.14687	17.09470
2	811 03	Neighborhood 3 (Bartoňova)	Bratislava 1	48.14808	17.10469
3	811 04	Neighborhood 4 (Banícka)	Bratislava 1	48.16131	17.09738
4	811 05	Neighborhood 5 (Anenská)	Bratislava 1	48.15561	17.10973

Fig. 4. Dataframe that was used for the Data Analysis section.

This section can be visualized with more details in the Jupyter Notebook.

3. Methodology

3.1. Exploratory Data Analysis – Visualization

Rather than a statistical analysis, the first analysis I made was visual. For achieving this, I decided to plot the geographical points in a map by using Geopy and Folium Maps. By creating a function which could plot points of a dataframe in a map based on their latitude and longitude, I obtained a map of neighborhoods in Bratislava.

As it is possible to see in Fig. 5, all 56 points, which represented the 56 Postal Codes of Bratislava city, were plotted as blue dots. Some neighborhoods are close to each other while others are separated. The map and neighborhood points might not be perfect but it seems to be a fair approach.

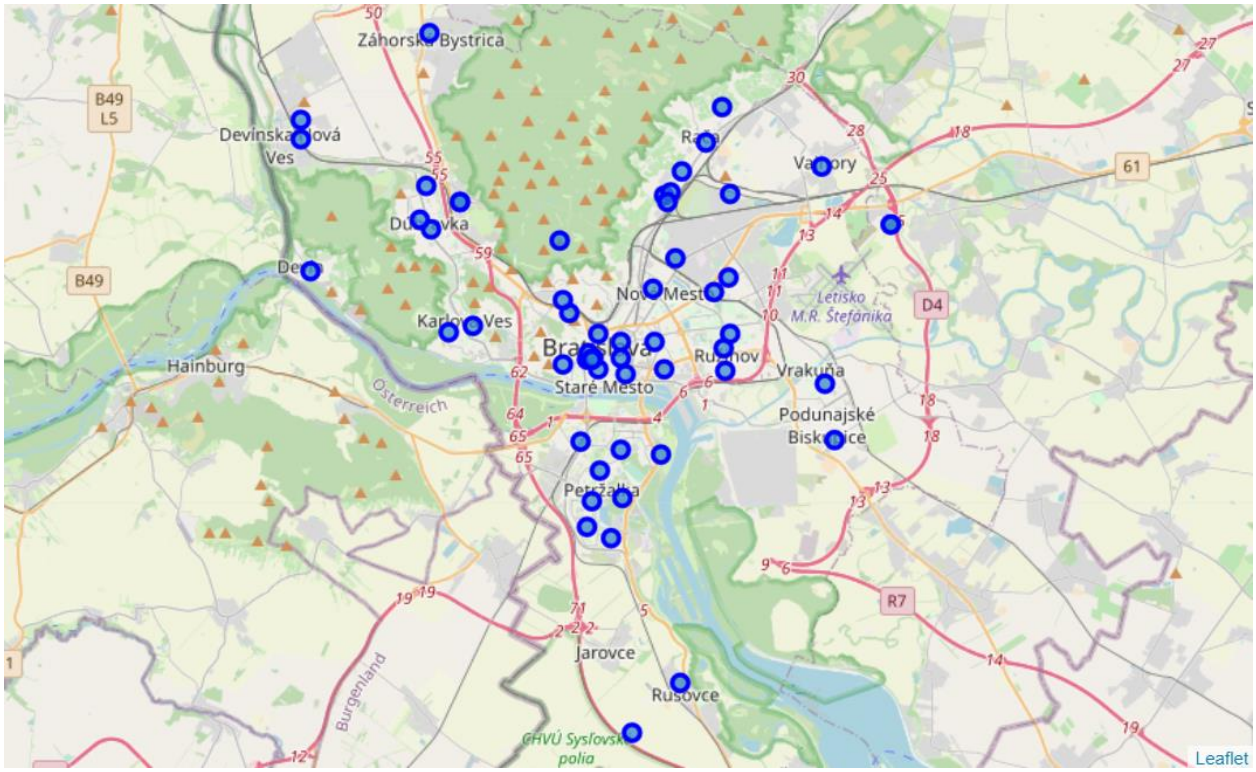

Fig. 5. Folium map which shows 56 Slovak neighborhoods.

3.2. Exploratory Analysis – Retrieving More Data

Through Foursquare API, I first retrieved information about venues in the first neighborhood, Neighborhood 1 (Beblavého). Fig. 6 shows the first 5 venues, out of 100, obtained.

	name	categories	lat	lng
0	Soho	Thai Restaurant	48.144105	17.111880
1	Stará tržnica	Event Space	48.144719	17.111225
2	Vespa Caffeteria	Café	48.143714	17.111012
3	Martinus	Bookstore	48.147157	17.110099
4	Viecha malých vinárov	Wine Bar	48.144707	17.111148

Fig. 6. First 5 venues retrieved for Neighborhood 1 (Beblavého).

After this, I used a function which would automatically retrieve the same information for each of the remaining 55 neighborhoods. However, not all neighborhoods reached the limit (100 venues), for instance, there were neighborhoods which only had less than 5 venues. I allocated all the retrieved data in a different dataframe, which would be used for the Machine Learning analysis.

	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	Neighborhood 1 (Beblavého)	Café	Bar	Wine Bar	Coffee Shop	Italian Restaurant	Pub	Brewery	Cocktail Bar	Hotel	Bistro
1	Neighborhood 10 (Astrová)	Café	Restaurant	Pizza Place	Japanese Restaurant	Gym	Drugstore	Eastern European Restaurant	Rental Car Location	Salad Place	Pharmacy
2	Neighborhood 11 (Astronomická)	Café	Italian Restaurant	Belgian Restaurant	Buffet	Poke Place	Burrito Place	Bus Stop	Furniture / Home Store	Paella Restaurant	Coffee Shop
3	Neighborhood 12 (Albrechtova)	Pizza Place	Italian Restaurant	Eastern European Restaurant	Pub	Bus Stop	Hotel	Tram Station	Gastropub	Cafeteria	Playground
4	Neighborhood 13 (Ambrušova)	Dessert Shop	Hotel	Bus Stop	Café	Grocery Store	Tennis Court	Eastern European Restaurant	Playground	Park	Italian Restaurant

Fig. 7. All Bratislava neighborhoods with their 10 most common venues.

Before continuing, I decided to check if there was any row with missing values. Eventually, the API could not retrieve information for three neighborhoods, thus, they were discarded.

3.3. Machine Learning: Clustering Neighborhoods

In order to cluster Bratislava neighborhoods, I used K – Means algorithm and the Elbow Method for obtaining the optimal value for K. Fig. 8 shows behavior of K respect to the squared error.

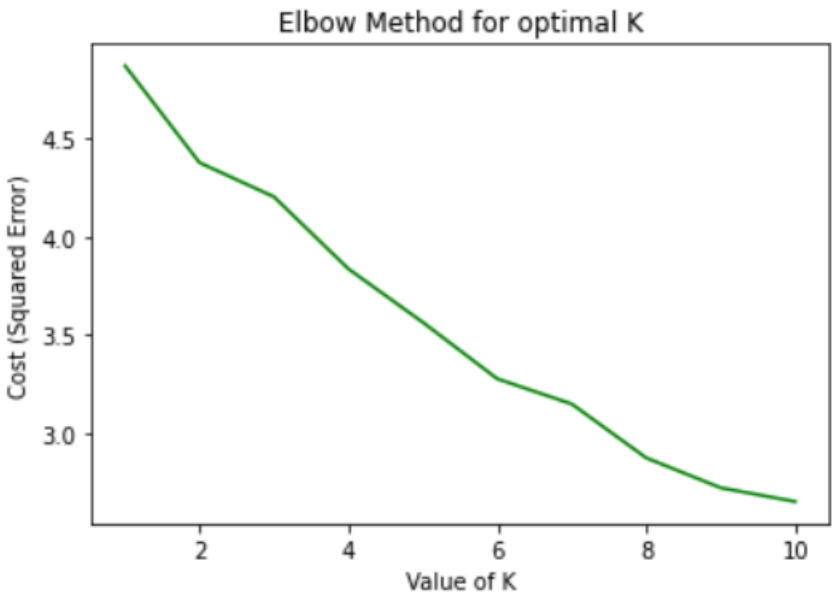

Fig. 8. Values of K vs Cost (Squared Error).

It is possible to see that there are two small “elbows” formed at K = 2 and K = 6. I decided to work with 6 since 2 is a small number for K. Consequently, the quantity of clusters obtained from K – Means algorithm was 6. These 6 clusters were plotted in a map again, for easy visualization.

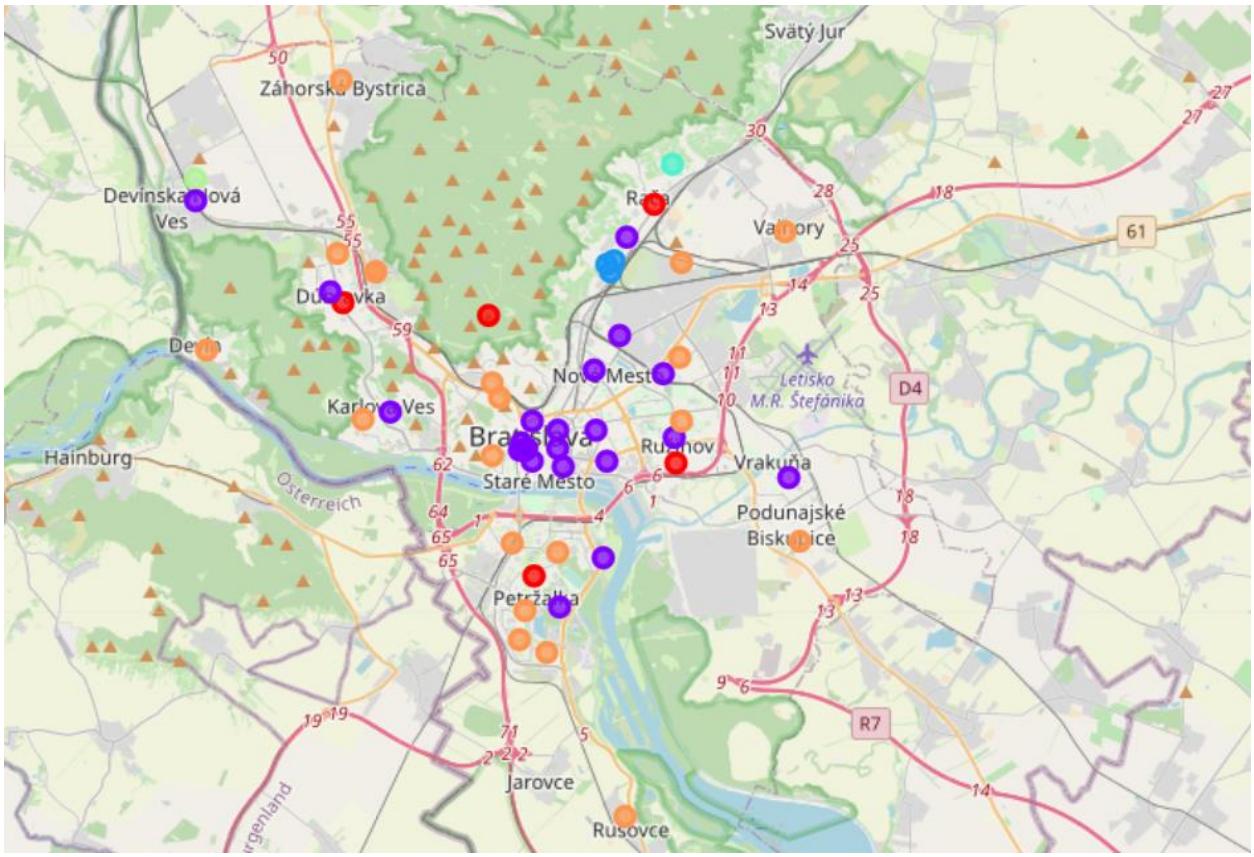

Fig. 9. The 6 clusters obtained from K – Means algorithm.

4. Results

Now that it is possible to see Bratislava's 6 neighborhood clusters, I will point out some important visualization results. First, I consider important to learn the differences between cluster and cluster, or, in other words, the reason why the model built the clusters in this precise way. For that, I think it is fundamental to consider how many and what kind of venues each neighborhood has. As a first step, Fig. 10. shows the total number of venues for each neighborhood.

There were 5 neighborhoods which had more than 60 venues retrieved by the Foursquare API. I built a function which would locate the cluster where these neighborhoods belonged to. Through this function, I found that these 5 neighborhoods belonged to Cluster 2. This made me think that the neighborhood with more venues would have a bigger influx of people, therefore, Cluster 2 could hold a good candidate neighborhood for opening a restaurant. In order to confirm this, some further observations in other results are needed.

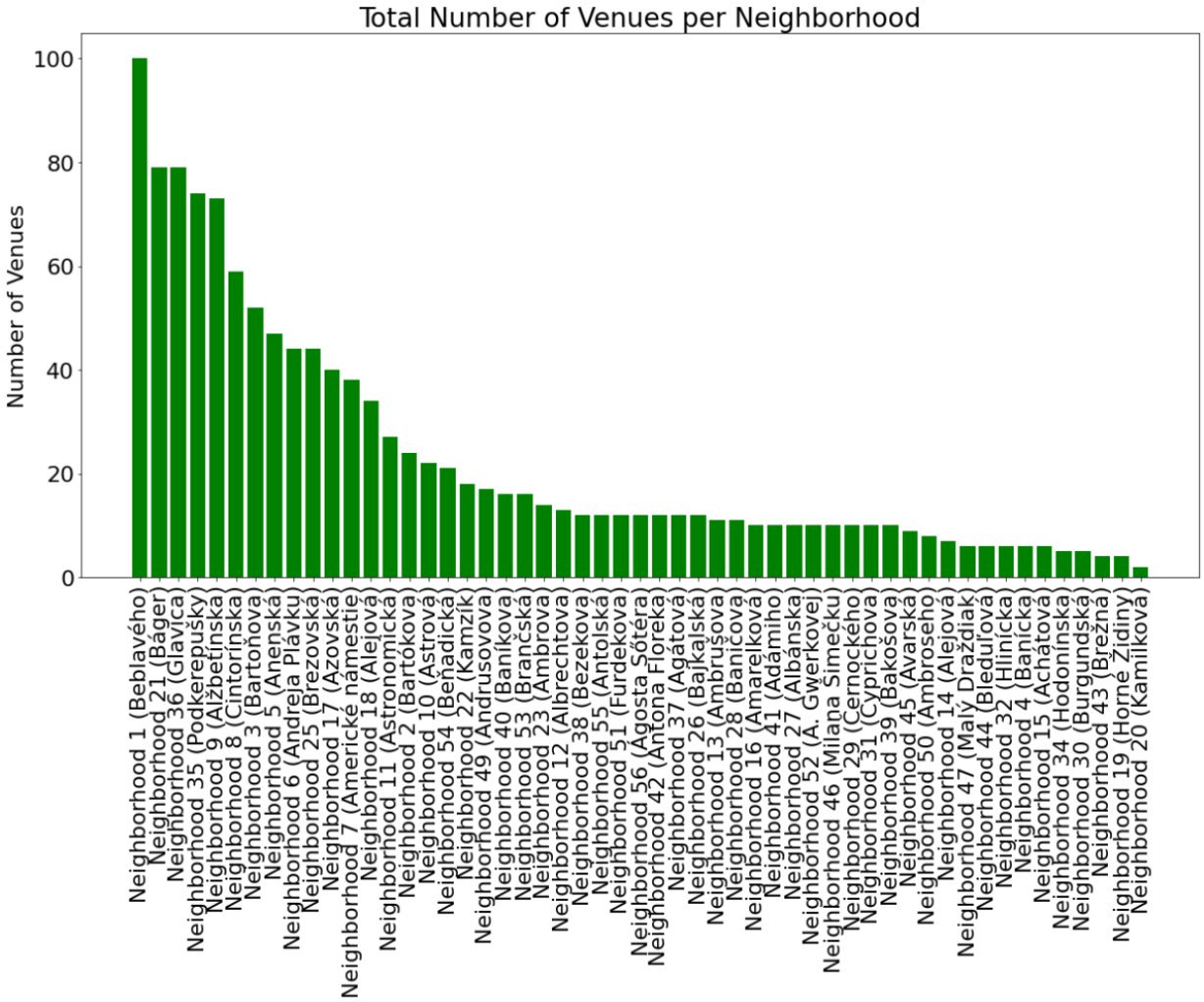

Fig. 10. Descending Ordered Chart of Total Number of Venues per Neighborhood.

As a next step, I built a function which would show the quantity of venues per cluster in a pie chart (Fig. 11.). As the pie shows, Cluster 2 is the cluster that has most of the total venues retrieved (230 venues calculated by the sum function), followed by Cluster 6, which has 200 venues. The remaining clusters have significantly lower number of venues compared to the first two ones (Cluster 1 with 50, Cluster 3 with 30, Cluster 4 with 10 and Cluster 5 with 10 venues as well). Because of the big difference between the first two clusters with the rest, I decided to focus only in Cluster 2 and Cluster 6.

- **Cluster 2**

In order to understand why Cluster 2 has the biggest number of venues, it is important to consider which neighborhoods are part of this cluster. For that, I reused the function for drawing maps and introduced geographical data of the neighborhoods which belonged to Cluster 2 exclusively, this result can be seen in Fig. 12.

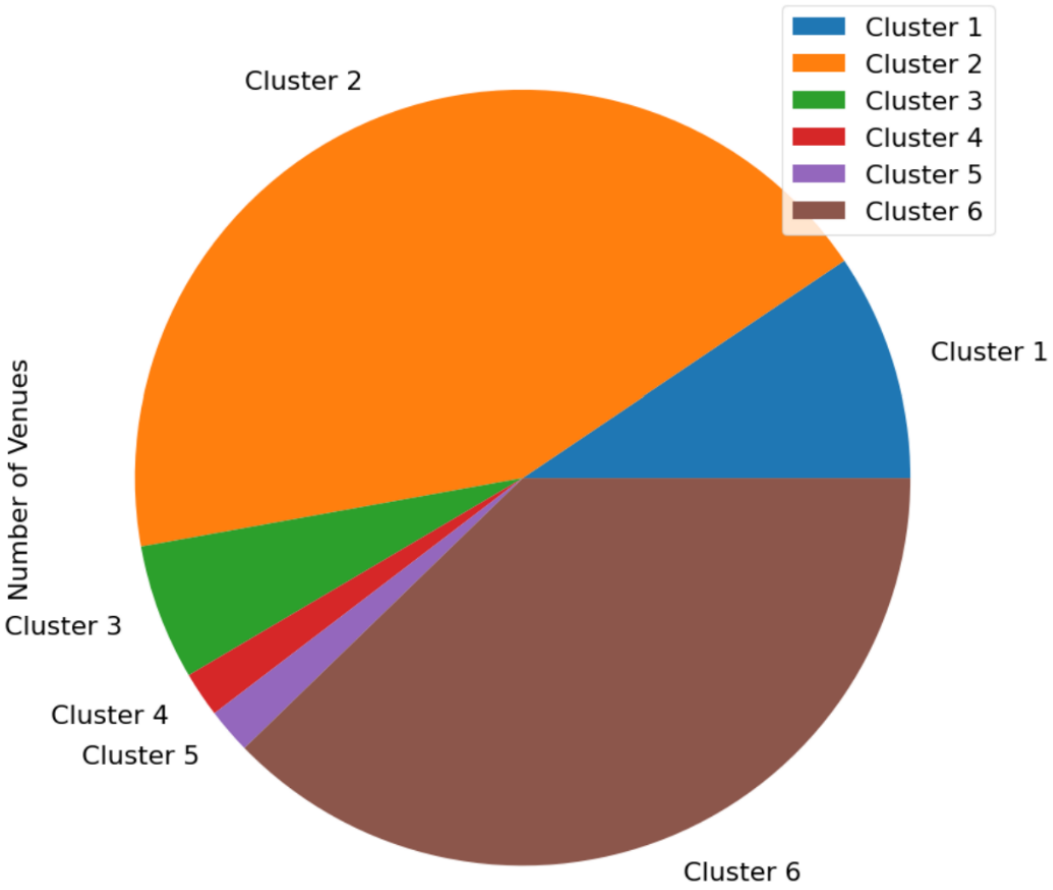

Fig. 11 Proportion of Venues in Bratislava.

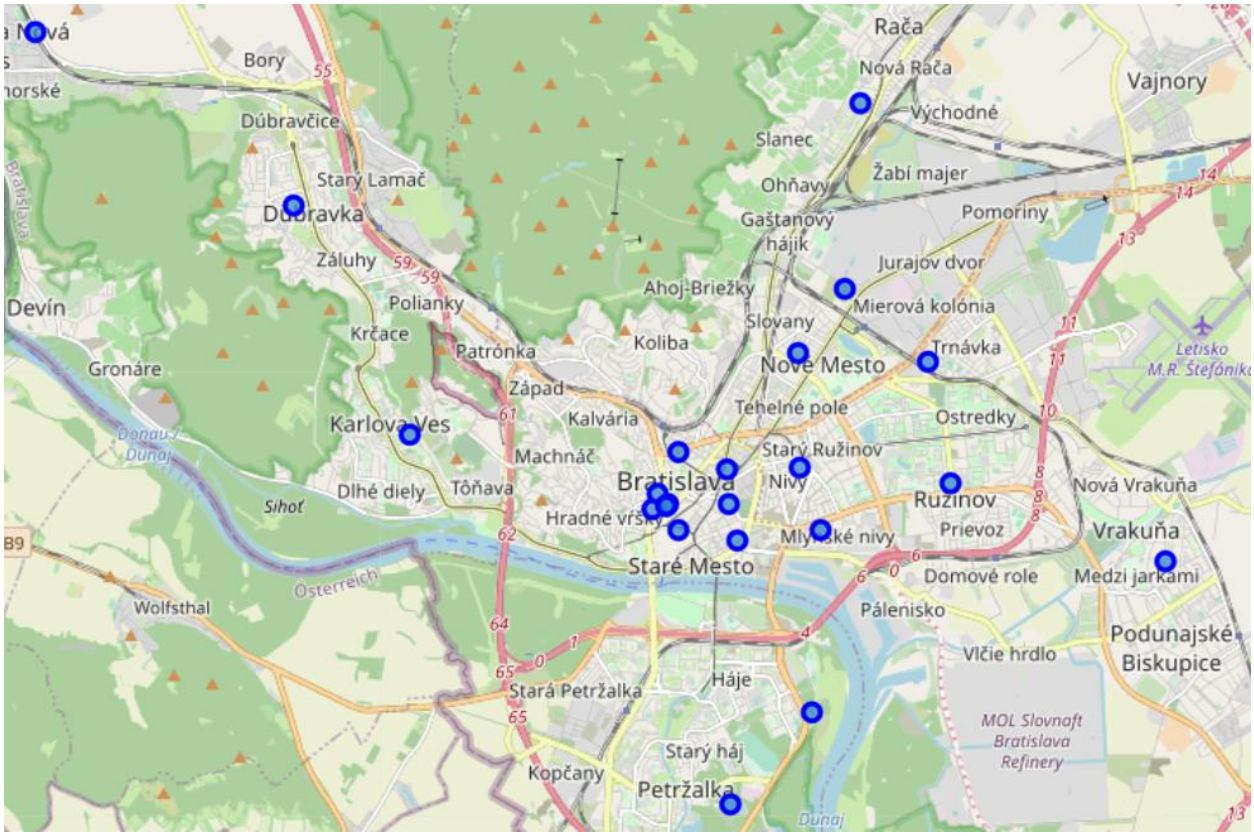

Fig. 12. Cluster 2 Neighborhoods.

In addition, I consider important to visualize what kind of venues are the most frequented in this cluster. Fig. 13. displays a bar chart with the 10 most common types of venues for Cluster 2.

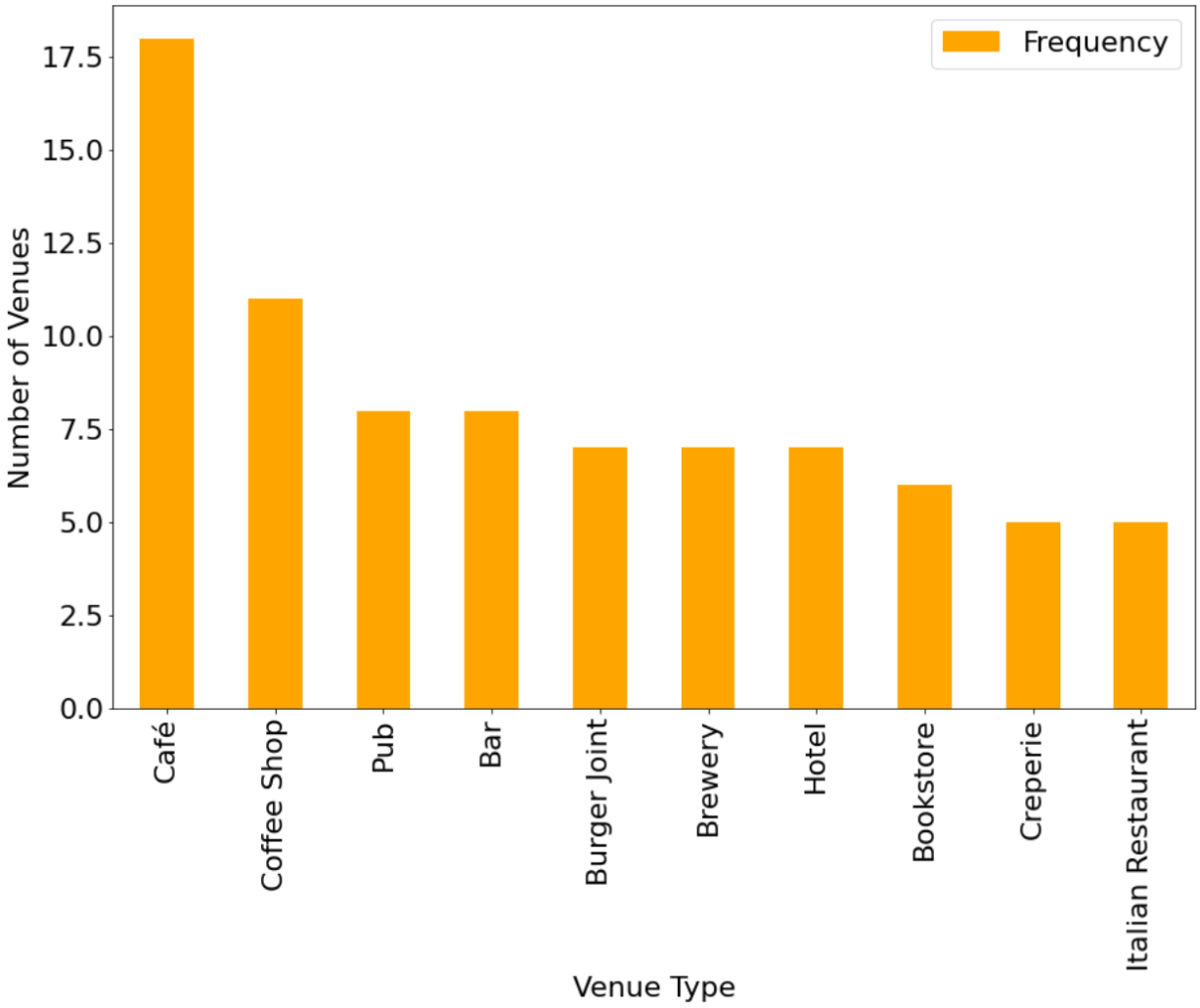

Fig. 13. Most common types of venues for Cluster 2.

As Fig. 12 shows, most of the neighborhoods in Cluster 2 are located in what is known as the center of Bratislava, and, from Fig. 13 it is possible to see that most of the venues are coffee shop style or casual dining ones like pubs, bars or burger joints. Moreover, Fig. 13 also shows the number of venues for each kind and, it is noticeable that Cluster 2 is composed mostly of coffee shop venues (29 in total) and casual dining (23 venues in total). This is something to consider and it will be properly discussed in the next section.

- **Cluster 6**

For Cluster 6, I repeated the process used for Cluster 2. First, I gathered the geographical data for all the neighborhoods which belonged to Cluster 6 and drew a map through the function already created. This result is shown in Fig. 14.

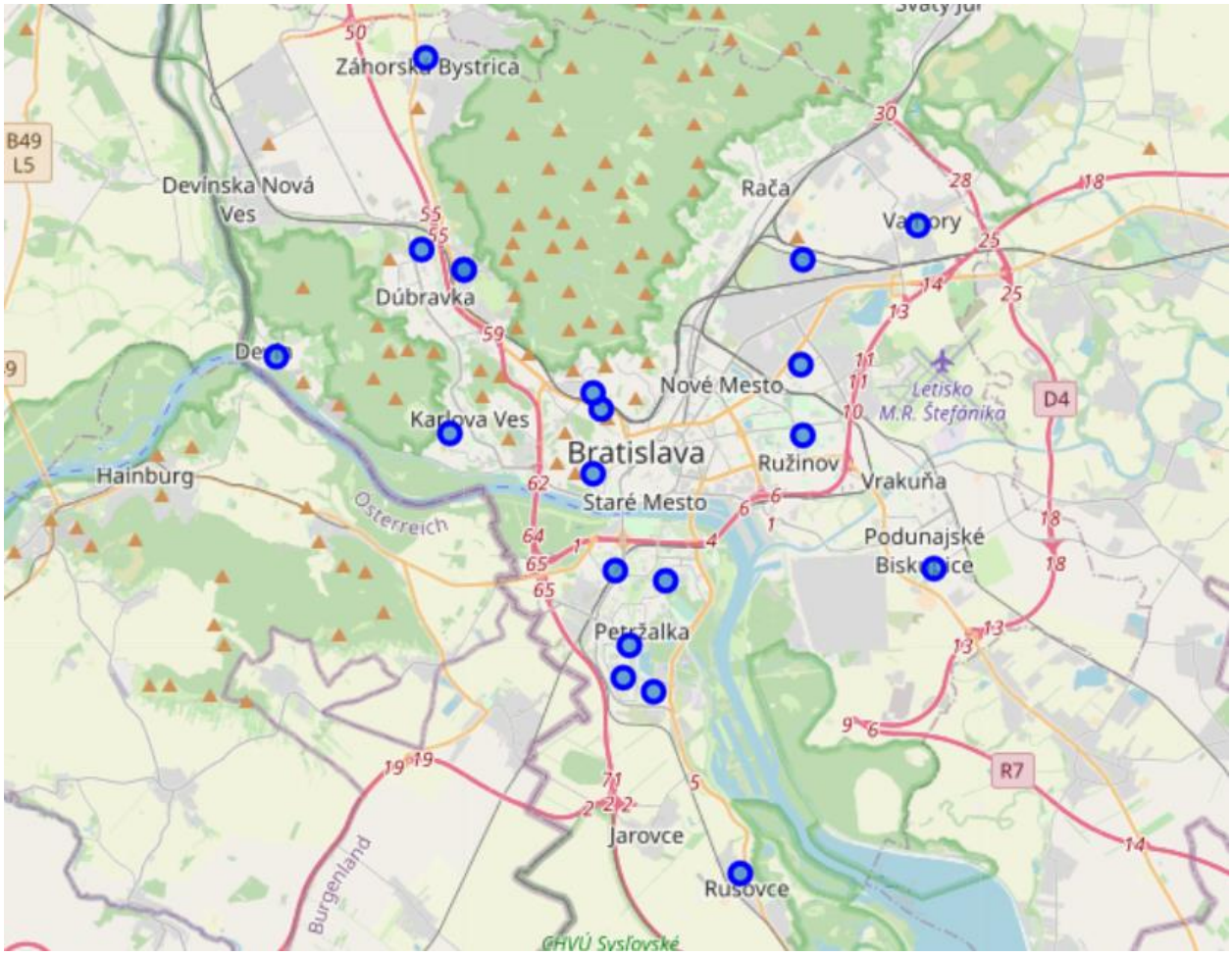

Fig. 14. Cluster 6 Neighborhoods.

The neighborhoods in Cluster 6 are more spread across the Bratislava region compared to the neighborhoods in Cluster 2, however, there is another important point to bear in mind. The neighborhoods of Cluster 6 are mostly residential areas in Bratislava; thus, the most representative venues should be different compared to the venues shown for neighborhoods in Cluster 2. In order to verify this prediction, I plotted Cluster 6 different types of venues in Fig. 15.

As the bar chart shows, the biggest number of venues correspond to bus stops. This could be because the neighborhoods in Cluster 6 are located further away from each other, therefore, the model found this kind of venue relevant. There is another type of venue which is important to point out: (local) restaurants. A big difference compared to the neighborhoods in Cluster 2 is that Cluster 6 ones have a big number of restaurants (22 in total, counting regular Slovak restaurants, Eastern European restaurants and Italian restaurants). Moreover, there are venues like Supermarkets, Pharmacies and Plazas which are proper from residential areas.

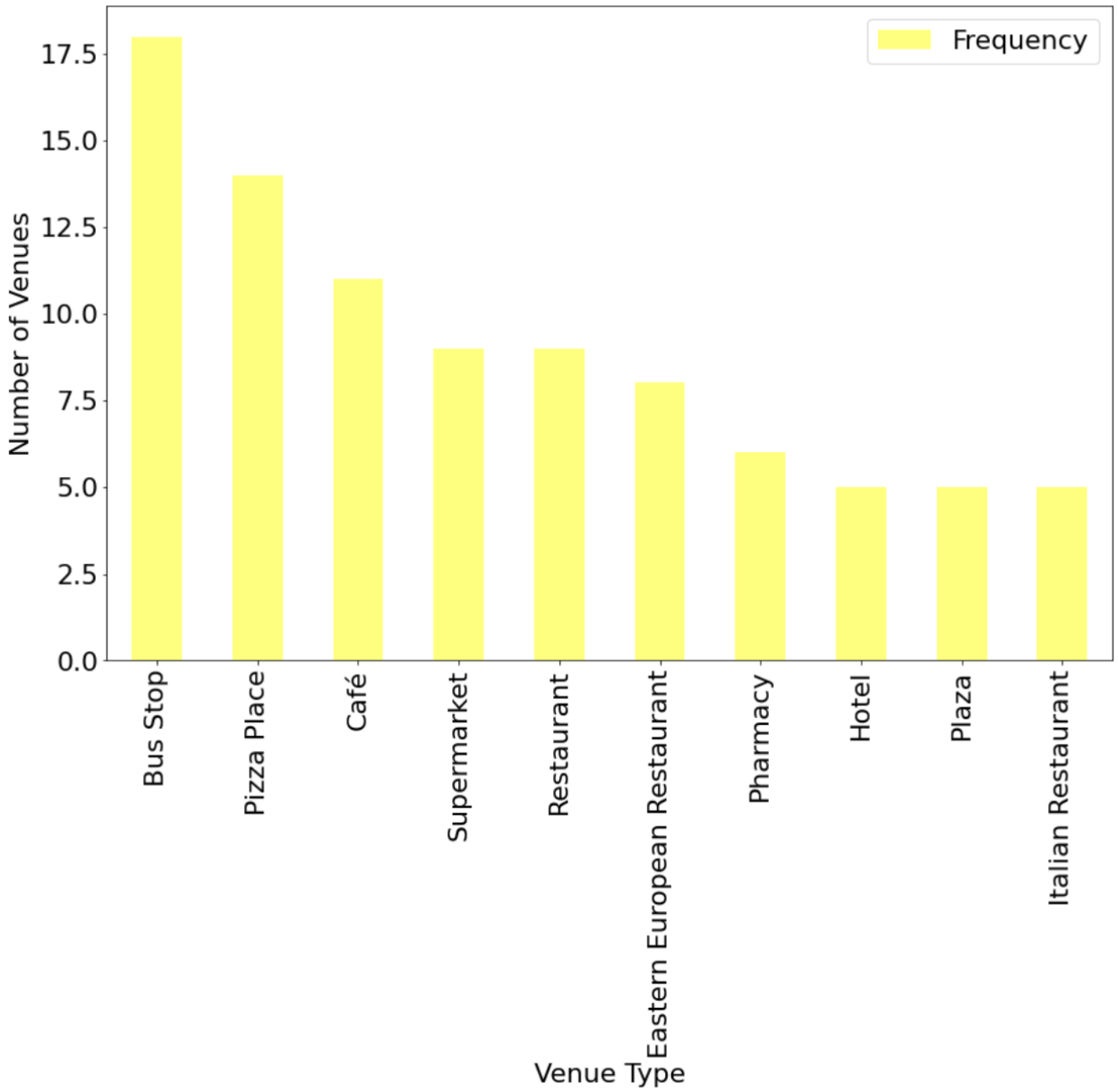

Fig. 15. Most common types of venues for Cluster 6.

5. Discussion

So, in order to answer the question “where is the best location for opening my restaurant?”, it is important to pay attention to the differences between Cluster 2 and Cluster 6.

First, the location of the neighborhoods in each cluster. As it was mentioned in the results section, these neighborhoods are in different locations, Cluster 2 neighborhoods are mostly located in the center of the city (Downtown and touristic spots) while neighborhoods from Cluster 6 are located mostly in residential areas.
But how does the location of neighborhood affect in choosing where to open the restaurant? From my point of view, this is almost determinant since, because of the location of neighborhoods, each cluster has different style of venues. Fig. 13 and Fig. 15 show how different these venues are: Cluster 2 venues show more Coffee Shop

style venues (29). Also, casual dining venues like Pubs, Bars and Burger Joints are popular and characteristic of a downtown area. I am pointing out this because these kinds of venues attract more people, in other words, the area has a bigger influx of people. On the other hand, representative venues from Cluster 6 are local restaurants and places where people can shop, get groceries and medicines. As it was already mentioned, these kinds of venues are proper from residential areas, however, these does not guarantee a bigger influx of people compared to Cluster 2 style of venues.

6. Conclusion

In this work, I looked for potential places where an interested person could open an international restaurant based on neighborhood clustering in Bratislava, Slovakia. For this, I considered groups of streets which share the same Postal Code as neighborhoods. These neighborhoods were plotted in a map and through K – Means algorithm, I successfully obtained clusters of neighborhoods.

Following the discussion in the previous section, I could conclude that a potential candidate for opening an international restaurant in Bratislava is any neighborhood located in the center of the city, known as the Downtown. This conclusion is based on the fact that the kind of venues like coffee shops and casual dining attract more people. Therefore, the influx is bigger and this would mean that the chances of getting clients will be bigger than in any other neighborhood in the city. Another point to support this conclusion is that, usually, people tend to have lunch in a restaurant and then go for a coffee, or, people meet for having some dinner in a restaurant and then go to a bar or pub. Thus, the area seems to be ideal. Certainly, economical factor was not considered in this analysis because of the scope of the project.

As for the future, I plan to improve and implement this analysis by adding more and different kinds of data (like housing – renting prices in different areas in Bratislava), so more details could be taken into account. Also, I intend to include more Machine Learning techniques that could provide a variety of results. Lastly, I plan to include accuracy for models in order to analyze the reliability of the results.

7. References

[1] https://www.posta.sk/sluzby/postove-smerovacie-cisla
[2] https://www.postcodesdb.com/AlphabeticSearch.aspx?country=Slovakia&city=Bratislava
[3] https://github.com/OpenDataSk/datasets/blob/master/postalcodecoordinates_sk.json