Capstone Project - The Battle of the Neighborhoods in Warsaw

Contents

Introduction: Business Problem

Warsaw is the capital of Poland and the largest city in the country. The city's population is 1.8 million with a density of 3,469 people per square kilometer. Warsaw is divided into 18 districts with different population densities and areas. As a person who lived there for 5 years, I decided to choose this place for my project.

There are numerous companies and universities in the city attracting a huge number of interested people. Everyone is interested in various attractions in the city, depending on their needs. A family with children will have different needs than a student. The need to find an interesting and cheap point for yourself is a very important aspect. Additional needs are the knowledge of attractions available in the selected area.

Data on average housing prices will be presented on the map for individual districts. Additionally, we will have clustered information about venue density.

Data

Based on the definition of our problem, the factors that will influence our decision are:

* number of points of interest broken down by type;

* Average prices of apartments in the area.

The following sources will be used to obtain the necessary information:

* Wikipedia - Administrative divisions of Warsaw;

* Forsquare API to obtain information about the points of interest;

* Public data on average prices per square meter in individual districts

Neighborhood Candidates

First, let's download the list of Warsaw districts from Wikipedia. These data will help us in the next steps of the analysis.

Wikipedia data was retrieved using the Python Beautiful Soup library. This library allows you to download data from websites for use in data analysis.

	Neighborhood	Count_of_citizens	Density_of_citizens	Area
0	Mokotów	218 265	6162	35,42
1	Praga-Południe	180 789	8078	22,38
2	Ursynów	151 304	3455	43,79
3	Wola	141 407	7342	19,26
4	Bielany	131 592	4069	32,34

Data retrieved from Wikipedia

	Neighborhood	Price	Currency	Latitude	Longitude
0	Praga-Południe	9840	PLN	52.241663	21.080788
1	Bemowo	9593	PLN	52.240780	20.912336
2	Targówek	8874	PLN	52.282402	21.065551
3	Ursus	8824	PLN	52.194752	20.883897
4	Wawer	8740	PLN	52.195783	21.180152

Spatial data for neighborhoods

	Neighborhood	Count_of_citizens	Density_of_citizens	Area	Price	Currency	Latitude	Longitude
0	Mokotów	218 265	6162	35,42	11500	PLN	52.194494	21.031829
1	Praga-Południe	180 789	8078	22,38	9840	PLN	52.241663	21.080788
2	Ursynów	151 304	3455	43,79	10105	PLN	52.138363	21.032758
3	Wola	141 407	7342	19,26	11830	PLN	52.233072	20.958001
4	Bielany	131 592	4069	32,34	10014	PLN	52.293450	20.929596

Merged data from Wikipedia and spatial data.
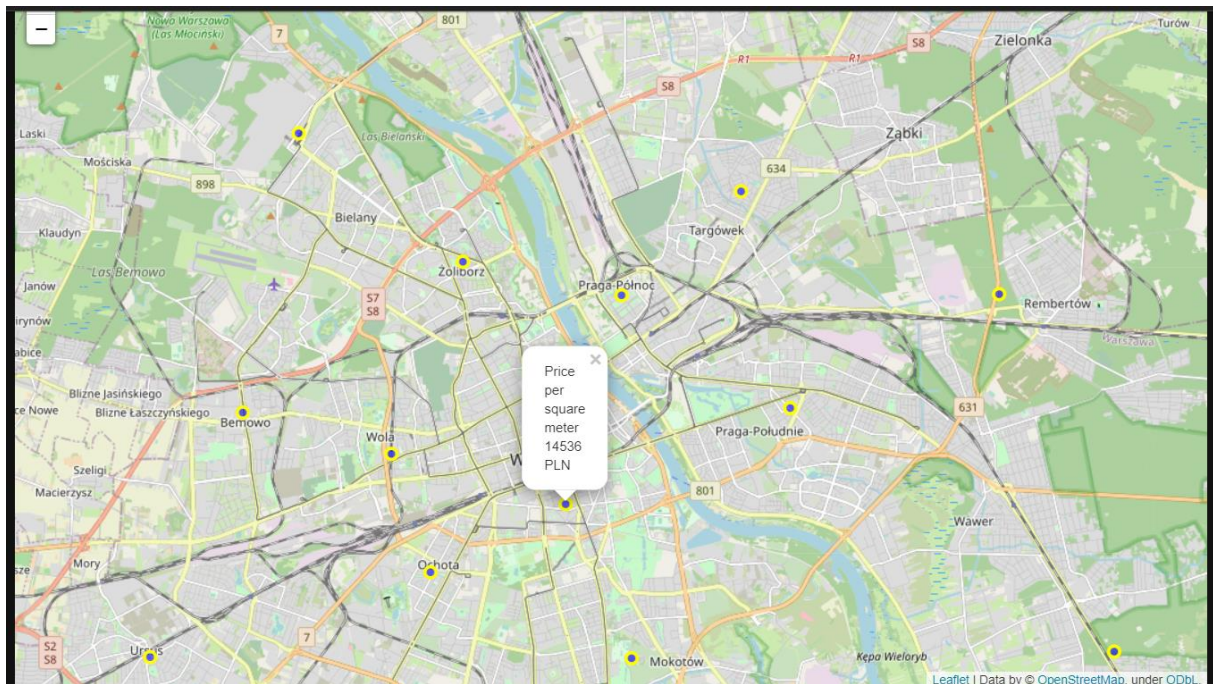
The most expensive district is Śródmieście. Price per square meter is araound 14 536 PLN. Let's check how lool like density of venues in districts.

Collection of data from Forsquare API will give us information about density of venues. Belowe is example if frequency per district for top 5 venues.

```
----Wola----
                   venue  freq
0                   Café  0.18
1            Bus Station  0.18
2              Rock Club  0.09
3              Gastropub  0.09
4    Gym / Fitness Center  0.09


----Włochy----
                   venue  freq
0           Tram Station  0.22
1      Accessories Store  0.11
2      Turkish Restaurant  0.11
3            Pizza Place  0.11
4       Electronics Store  0.11


----Śródmieście----
                          venue  freq
0                          Café  0.11
1   Vegetarian / Vegan Restaurant  0.05
2                        Bakery  0.05
3                    Coffee Shop  0.05
4                         Plaza  0.04
```

	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	Bemowo	Bus Station	Food & Drink Shop	Diner	Ice Cream Shop	Japanese Restaurant	Furniture / Home Store	Park	Pet Store	Pizza Place	Donut Shop
1	Bielany	Pizza Place	Doner Restaurant	Shoe Store	Grocery Store	Coffee Shop	Bike Rental / Bike Share	Bus Station	Paintball Field	Ramen Restaurant	Pub
2	Mokotów	Park	Gym Pool	Tennis Court	Sports Club	Restaurant	Gym / Fitness Center	Gym	Gun Range	Paintball Field	Climbing Gym
3	Ochota	Hotel	Electronics Store	Dessert Shop	Park	Tennis Court	Pub	Bus Station	Flea Market	Department Store	Supermarket
4	Praga-Południe	Coffee Shop	Hostel	Café	Italian Restaurant	Kebab Restaurant	Fast Food Restaurant	Eastern European Restaurant	Restaurant	Clothing Store	Supermarket

Data needed for clustering – grouped neighborhood with top 10 venues
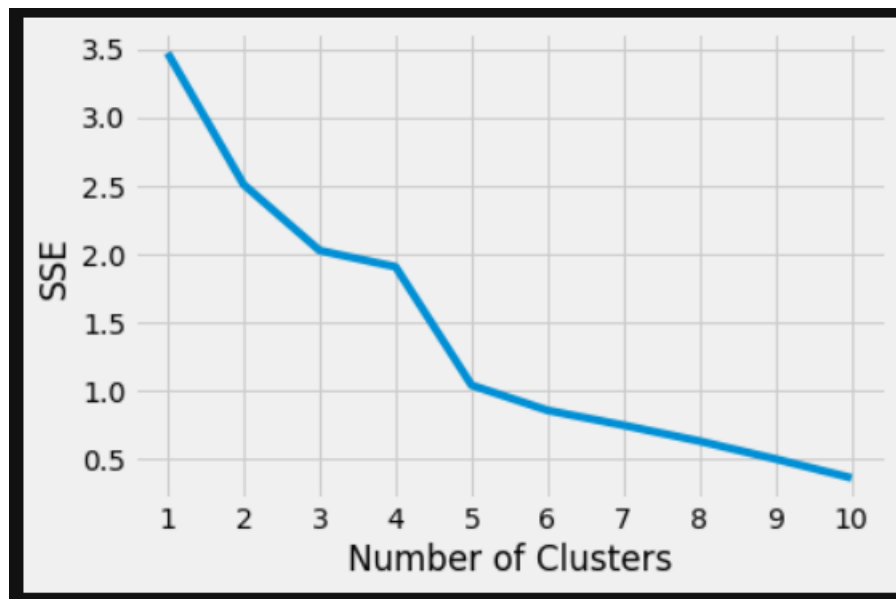
To select correct value of k I used elbow method. It gave me result of k=5. This k gave us split categories by 5 clusters visible on below map.
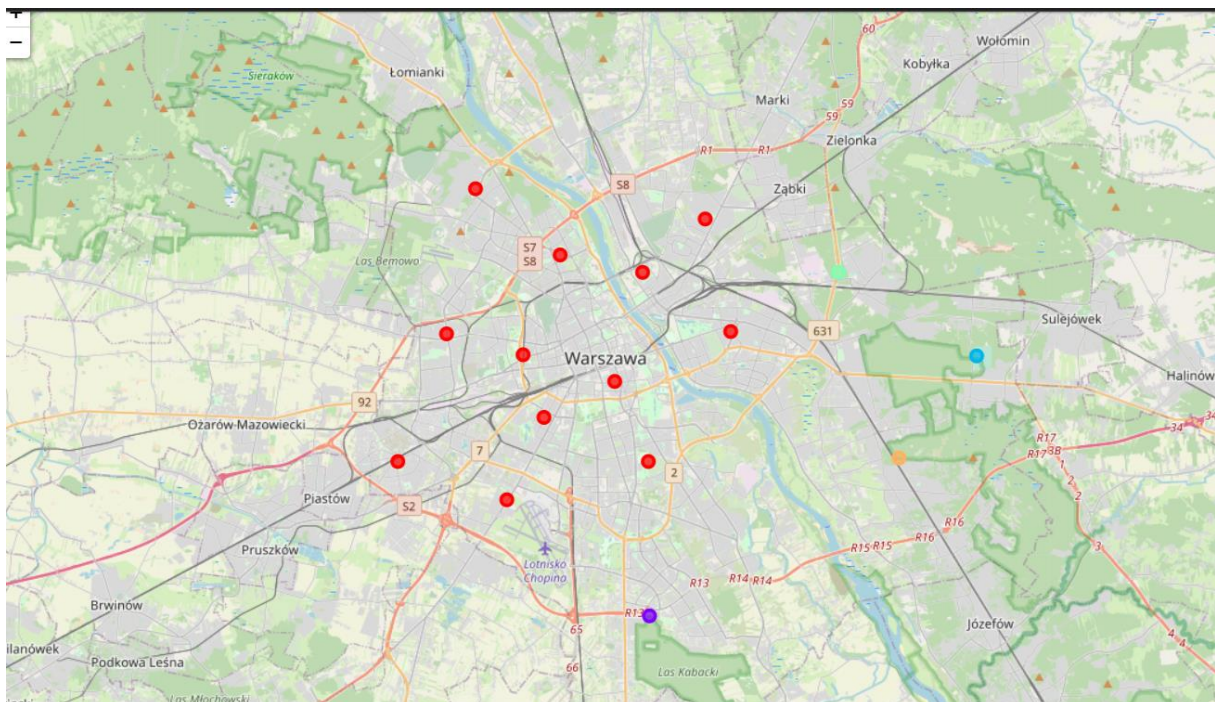
Methodology

The aim of the analysis is to find the optimal district for the resident in accordance with his needs.

The first analyzed case is a family with children, and the second is a student.

As we all know, both families with children and students will be guided, among other things, by the price and attractions available nearby.

In the section on data, we have gathered the necessary information to decide. Data on home prices come from the Sonarhome.pl website, which collects data on prices.

Data on available attractions in each district was downloaded from the Forsquare API website.
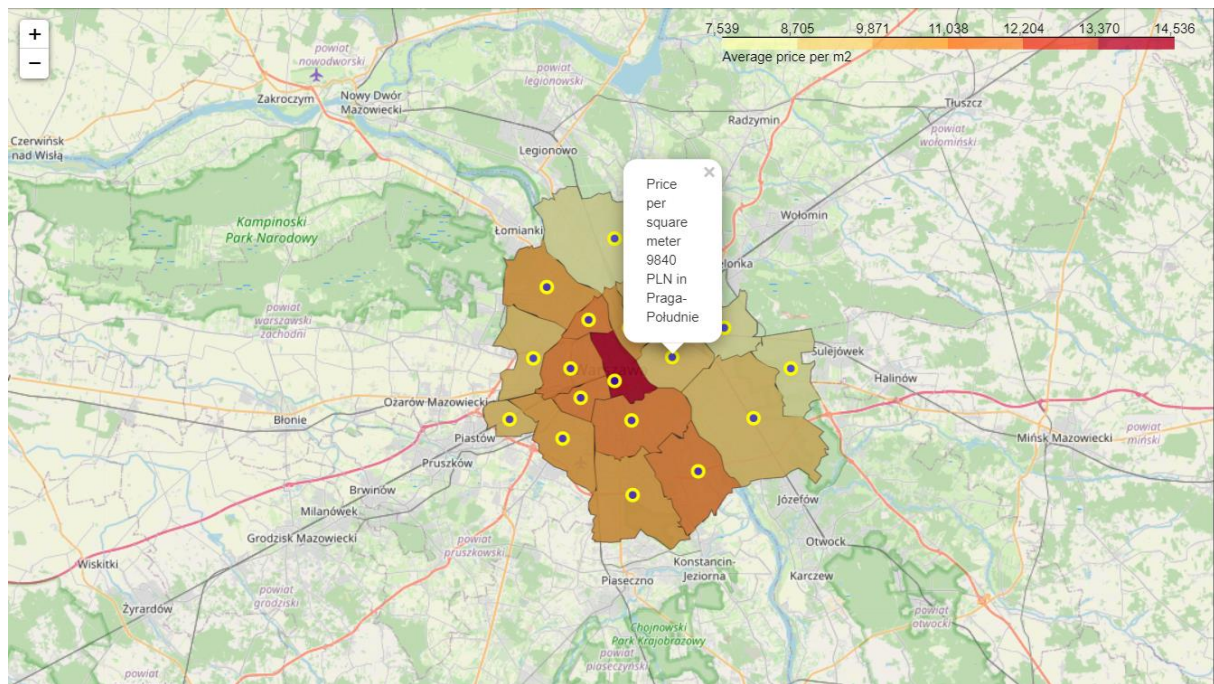
GeoJson file with Warsaw district was downloaded from git repository and edited to fit data from Wikipedia.

Analysis

There are 104 unique types of service categories in Warsaw. When analyzing the Top 1 for individual districts, we can see a large differentiation. The dominant types are cafe, hotel and park.
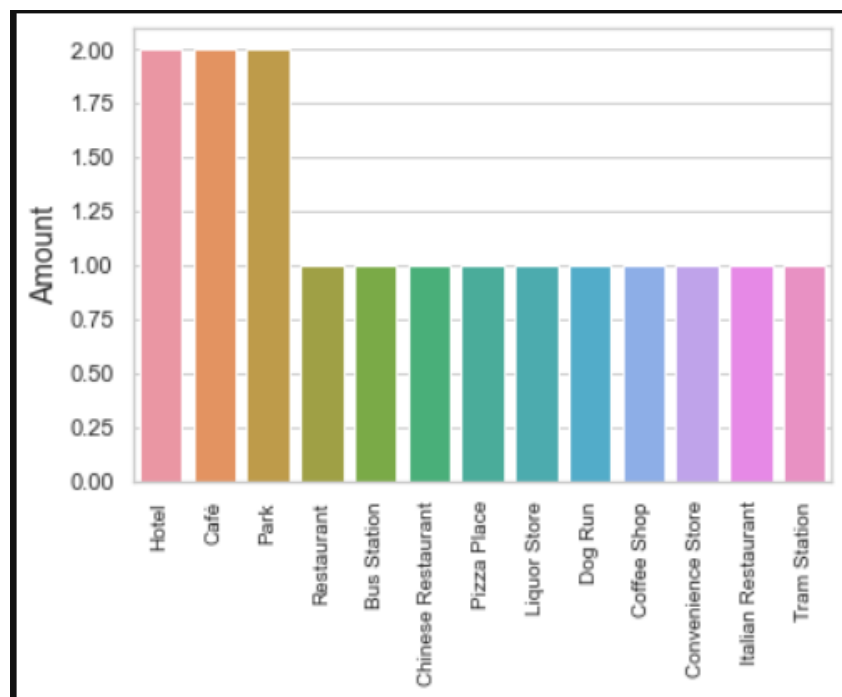
Let's check choropleth map to review prices for flats.

Results and Discussion

As mentioned at the beginning of the analysis, Warsaw is a large multicultural city. People come to Warsaw to start their adventure with a big city, studies or a better job. Analyzes like this can help them choose the right place to buy a home when they don't know the area.

As you can see on the last map, the highest prices are in the city center. This is due to the proximity of the subway, train station and the city's highest concentration of attractions. Well-connected neighborhoods will always be the most expensive.

When selecting the ka number during clustering, I was guided by the elbow method. The selected number of clusters is 5. 16 out of 18 districts of Warsaw were used in the analysis, as there was no data available in the Forsquare API for two of them.

Conclusion

Such an analysis may be useful for visitors to choose the optimal location for their new apartment, taking into account its price and the surrounding area.

Future investors can use such data to spot the gap for their business, and city managers can better manage it.

Spatial analyzes can be used in everyday life by each of us, especially when we do not know the city.

Magdalena Siedlecka