Analysis of House Sales Prices in Érd, Hungary
(report)

1. Introduction/Business Problem

I live in Hungary (a small country in Eastern Europe with population of 9.5 million), in a small city near the capital, called Érd. In the last 10 years, this city has developed a lot. Thanks to it more and more people are moving here, who want more peaceful and clearer environment around them. The population of Érd has grown by 12% in 10 years, and the population density is 1140 people per square kilometer (the area of Érd is 60 square kilometer).

The city is just 20km from the capital, that's why 60% of the inhabitants works in the capital (commuter). The rapid growth of the population has led to the construction of housing in large numbers. In the last 10 years the number of dwellings increased by 9%. This development attracts investors and entrepreneurs, so more and more shops, stores, restaurants are opening.

Who want to move to Érd, try to find a place for their new home at the best price, but they also want easy access to transport and social places. I live in Érd for 2 years, and I remember it was very hard to gather all of these information and find a good place near to railway or bus station, near to school and so on. Unfortunately there is not such a comparison web page, therefore **in my project I will show a map about Érd and its districts, the real estate prices in these districts and make an analysis of the connection to the venue density. Finally, I cluster the city based on the venue density and visualize it.**

2. Data

To solve the problem, I needed the following data:

- the boroughs of Érd and their coordinates
- the average sales price for houses in each borough
- the venues in each district

I used the following applications and sources:

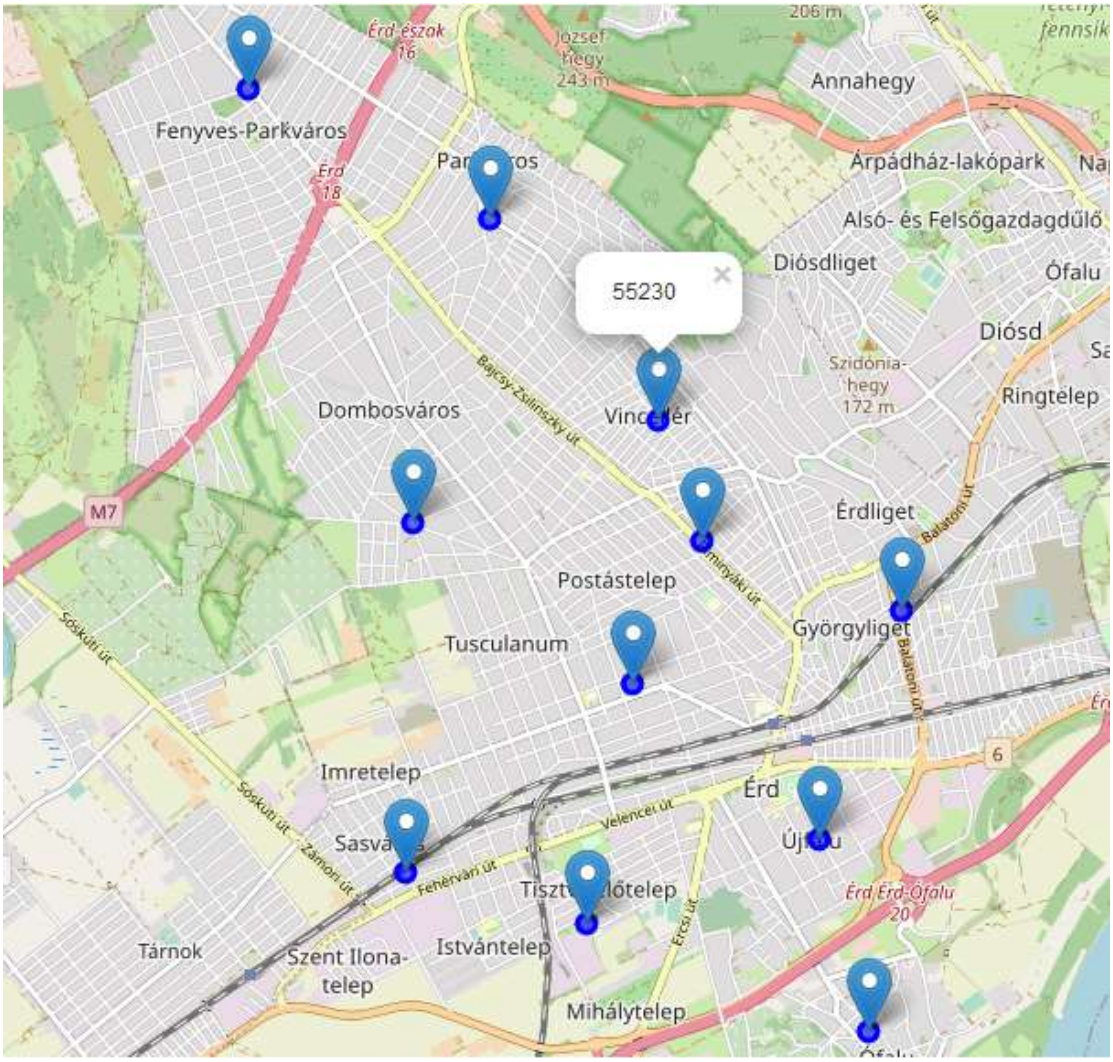
- I got the average real estate sales price from the page of National Statistical Office, I cleaned the data and reduced it to city of Érd.
- I used Google Map to get the center coordinates of each district.
- I used Forsquare API to get the most common venues of given district.
- I used Folium to visualize the average price in each district on map.
- I used unsupervised learning K-means algorithm to cluster the boroughs.

3. Methodology

Érd is a small town in Hungary near to the capital Budapest. It has a border with Budapest, so it is like a suburb.

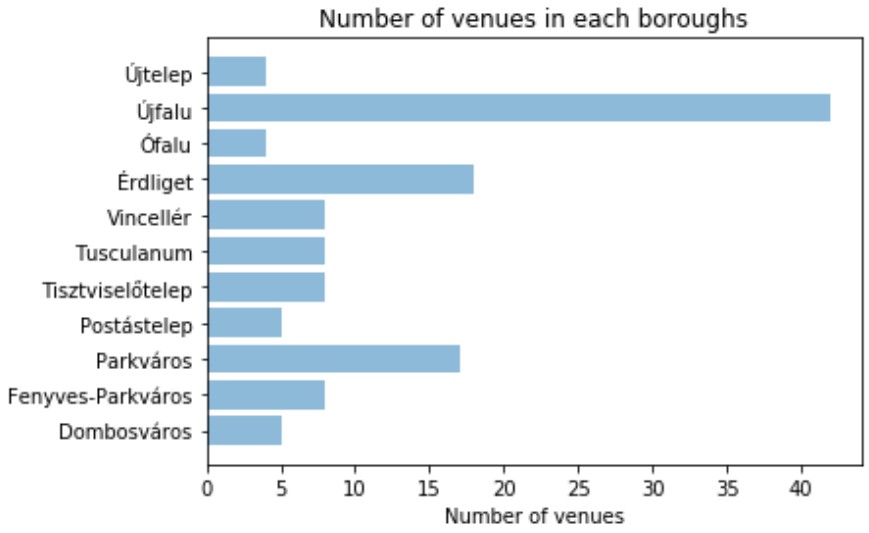
Érd has 11 districts, and I found their latitude and longitude coordinates with the help of Google Maps Nearby function. I wrote the values into an excel file, then I made a csv file out of it, that pandas can handle. I downloaded the average real estate sales prices for Budapest and its surrounding cities from web page of National Statistical Office in Hungary, I dropped the unnecessary columns and rows, so the data show the average price only for the boroughs of Érd. There is a big difference among the average prices, the lowest prices are in Ófalu, Újfalu and Újtelep - these are the oldest parts of the town near to the river Danube.

Thereafter I merged the two table into one, because I want to show the boroughs and their average prices in a map.



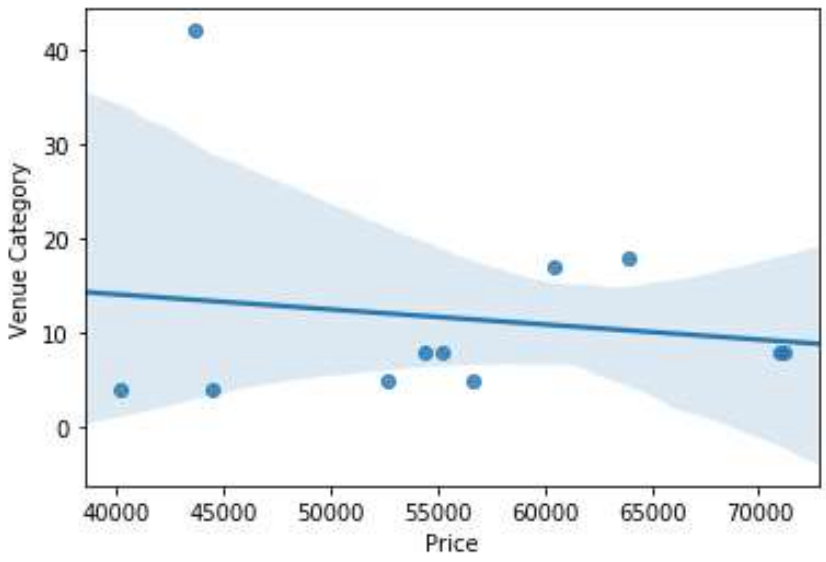
3.1 Exploring venues

I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as 50 venues and the radius 1000 meter for each borough from their given latitude and longitude information. There are 127 venues in 11 boroughs in 61 unique categories.



The most venue is in Újfalu, Érdliget and Parkváros - these districts are near to the railway station and motorway, and these are the fastest and best developing areas in Érd. The other districts have almost the same number of venues, but under 10.

Then I merged my price table with venue category to see the connection between the prices and venues. I try to find connection between the prices and the number of venues in each district so I use a correlation analysis for that (linear regression).

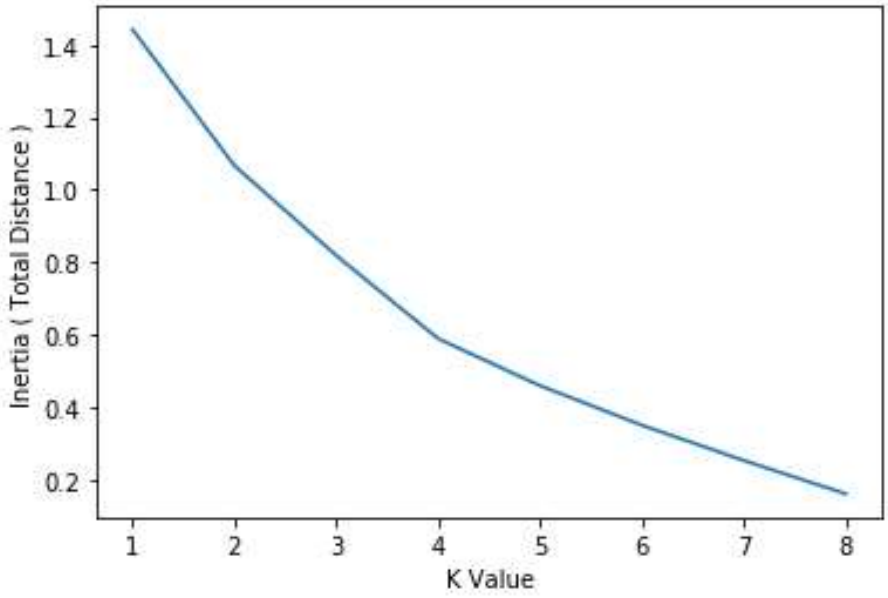


Unfortunately, the regression plot shows a weak correlation between the price and number of venues in my town. That means that for people who move into this town lots of stores or venues are not so important, this feature does not increase the housing price.

3.2 Clustering

Then I searched for the 10 most common venues in each neighborhood, that is needed for clustering. Among the most common venues we can find transport hubs, social venues as well as stores and eating places.

To cluster the town, I used unsupervised machine learning technique called K-means algorithm. First, I tried to find the best k value for clustering with elbow method. The plot shows me that there is no sharp shift in the line, but there are two breakings at 2 and 4 K value. 2 is two small value for clustering, so I chose the value 4 for K-means clustering.



Finally, I merged the origin boroughs-price table with the table contains the most common venues. The new table also contains the cluster labels, so we can see which neighborhood belongs to which cluster.

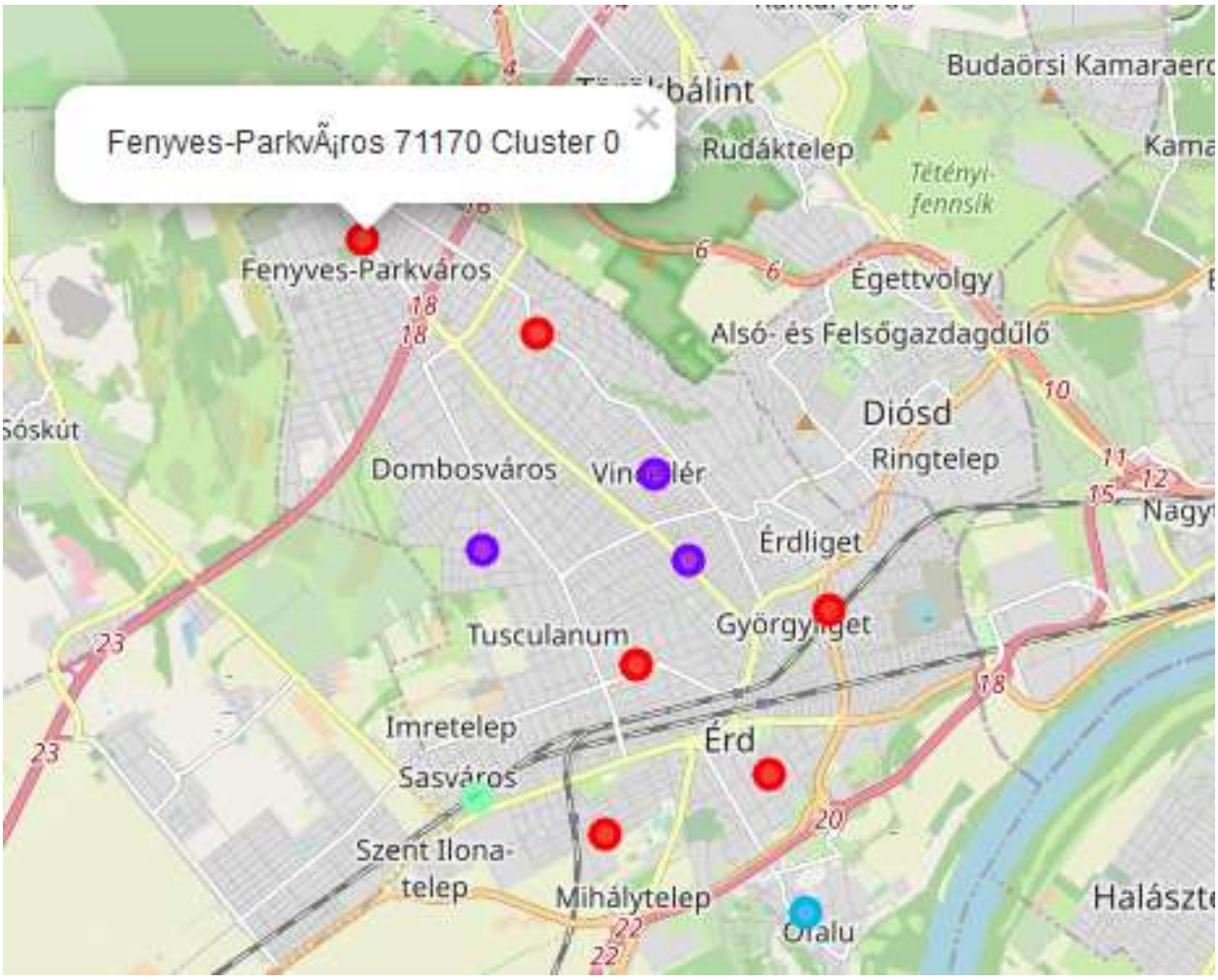
As it seems from the table, most neighborhood got to cluster 0 and 1, and only one-one boroughs belong to the cluster 2 and 3. I can label the clusters according to their main content as follows:

- Cluster 0 = mainly supermarkets, shops and restaurants
- Cluster 1 = mainly traffic stations and social venues
- Cluster 2 = sport field and nature
- Cluster 3 = small shops

4. Result

Finally, I wanted to visualize the results of clustering, I used geolocator package for that.

The markers shown that the town is divided into well-separated groups (cluster). Cluster 0 marked by red, is located on the north of Érd and on the both side of railway, Cluster 1 is between the red points, and the remaining two clusters are the furthest parts of the town. The label shows the name of borough, the average price and the name of cluster.



5. Discussion

In this project I examined the average housing sales prices in a small town in Hungary, Érd and I tried to find connection between the prices and the venues located in the districts and its neighborhoods.

I used linear regression to find correlation between prices and the number of venues in each borough, but the analysis showed the opposite: the number of venues does not influence the price. It is an interesting task to figure out the reasons why people move to Érd in such a big number, but I have no enough data and information to that, and it is out of scope of this study.

I used the K-means algorithm for clustering the districts. When I tested the Elbow method, I set the optimum k value to 4. All the 11 district coordinates were used in the sample. For more detailed and accurate result, the data set and the radius for searching venues can be expanded, maybe bigger k value can be used.

Finally, I ended my study by visualizing the price and clustering information on the map.

6. Conclusion

The population of Érd grows year by year, at the same time the number of stores, restaurants and supermarkets also increase. In the last 4 years the housing sales price has increased by almost 50%. To understand the real causes of this increasing is important for the entrepreneurs and the potential habitants, to find the right place for a new home or a new successful business. This kind of data analysis can help to bring the supplier and the customer together.