Project Report

Introduction

Currently, in most rural areas of Southern Bavaria, Germany, young people are mostly not staying at the smaller cities, but want to move to the larger cities like Munich, Ingolstadt or Augsburg. This is also the case for young families. (I define young families as families, where recently a child was born) Unfortunately, the decision where to move to in larger cities is dependent of a number of personal preferences.

Especially for young families, particular preferences matter. Among them are the crime rate of a certrain district of a city, the number of schools in a certain district and especially the real estate prices. The latter point is important due to the fact that often, in young families, only one parent is working while the other is taking care of the child or children.

In the upcoming project, I will refer to the city of Munich and analyze its 25 main districts. In particular, I will search for clusters of districts based on the three aforementioned preferences: number of schools in a district, real estate prices of a district and the crime rate of a district.

With my analysis, I provide a basis for young families, which want to move to Munich, to decide to which district to move to based on several clusters.

Data

To reach my goal, I will collect publicly available data from various sources.

- First of all, to visualize Munich, I will use the GeoJSON file provided by https://www.suche-postleitzahl.org/downloads/. It contains all postal codes from Germany. It has to be filtered to the postal codes of Munich.
- The postal code per district can be extracted from https://www.muenchen.de/leben/service/postleitzahlen.html. I have to collect both, the postal codes AND the districts of Munich, as some of the following data is available per postal code and some per district.
- I will collect data about the crime rates in Munich. Those are available for the year 2017 at the website of the city council www.muenchen.de for each district
- Regarding the real estate prices, I refer to https://suedbayerische-immobilien.de/Immobilienpreise-Muenchen, which list the price per square meter in Euro per district.
- Finally, to collect data about the schools, I use the foursquare API. Foursquare provides the names of, e.g., elementary schools (“Grundschulen”). Additionally, Foursquare provides the postal codes of each school that can be linked to the GeoJSON file.

Methodology

From a methodological perspective, we first have to link the data described above to one consistent dataframe, which can afterwards also be visualized.

To give young families an opportunity to choose from the districts with low crime rates, a larger number of schools and affordable real estate prices, I will cluster the districts of Munich using the k-nearest-neighbor (KNN) algorithm. We will see that there is not “the best” district of Munich.

Clustering the districts gives young families the opportunity to choose the district according the individual preferences. It may be that a family has much more money and the other two characteristics are, therefore, more important.

In this regard, I will in particular cluster on the three characteristics mentioned already. During the data preparation I will create one dataframe that contains all data necessary to apply the KNN algorithm.

- Column “DISTRICT” provides the name of the district
- Column “CRIME” contains the number of criminal activities in 2017
- Column “EUROPERSQM“ contains the price in Euro per square meter
- Column „COUNT_SCHOOLS“ contains the number of schools in a district

While the characteristics CRIME and EUROPERSQM are directly available on the websites presented in the previous section, the data for column COUNT_SCHOOLS is collected via the Foursquare API. However, the JSON file from Foursquare does only partly contain the name of the districts, while the postal code is available for most schools. Thus, due to data quality reasons, I created an additional table that links the postal codes to the districts and then matches the postal codes in that table with those from the Foursquare JSON file. Afterwards I grouped (groupby) the district using the count() method. This led to the final column “COUNT_SCHOOLS” which could afterwards be merged to the table previously described. See the exemplary table below:

	DISTRICT	CRIME	EUROPERSQM	COUNT_SCHOOLS
0	Allach-Untermenzing	889	5699	5
1	Altstadt-Lehel	7868	9208	6
2	Au-Haidhausen	3407	7872	7
3	Aubing-Lochhausen-Langwied	1533	5396	2

To afterwards visualize the data and the clusters, an additional table contains the previously mentioned tables, and in additional the following columns:

- Column “PLZ”, which contains the postal codes
- Column “PEOPLE”, which contains the number of people living in a district
- Column “CRIME_PER_PERSON, which contains the criminal activities per person in a district (the column was calculated simply by “CRIME/PEOPLE”)
- Column “LATITUDE”, which contains the latitude geo data of a district
- Column “LONGITUDE”, which contains the longitude geo data of a district
- Column “Cluster Labels”, which contains the cluster of the district using the KNN algorithm

Results

Applying the KNN to the dataset, I chose the value K = 4 (4 cluster) as the value to be the one clearly building distinguishable clusters. Thus, as a result, after normalizing the data, I will present four distinguishable clusters in the following.

In total, Munich contains 25 districts near the “city”, which are the inner districts, where one may want to live.

Cluster 1 contains the following 6 districts:

	DISTRICT	CRIME	EUROPERSQM	COUNT_SCHOOLS	Cluster Labels
2	Au-Haidhausen	3407	7872	7	0
5	Bogenhausen	2243	8399	4	0
10	Maxvorstadt	4321	7968	2	0
13	Neuhausen-Nymphenburg	3731	7419	4	0
17	Schwabing-Freimann	5273	8673	5	0
18	Schwabing-West	2122	7628	3	0

Cluster 2 contains following 10 districts:

	DISTRICT	CRIME	EUROPERSQM	COUNT_SCHOOLS	Cluster Labels
0	Allach-Untermenzing	889	5699	5	1
3	Aubing-Lochhausen-Langwied	1533	5396	2	1
6	Feldmoching-Hasenbergl	2072	4824	2	1
7	Hadern	1555	5991	3	1
8	Laim	2053	5489	2	1
14	Obergiesing-Fasangarten	2298	5364	2	1
19	Schwanthalerhoehe	1643	6964	1	1
20	Sendling	1784	6426	3	1
21	Sendling-Westpark	2165	6008	1	1
24	Untergiesing-Harlaching	1640	6095	1	1

Cluster 3 contains the following 2 districts:

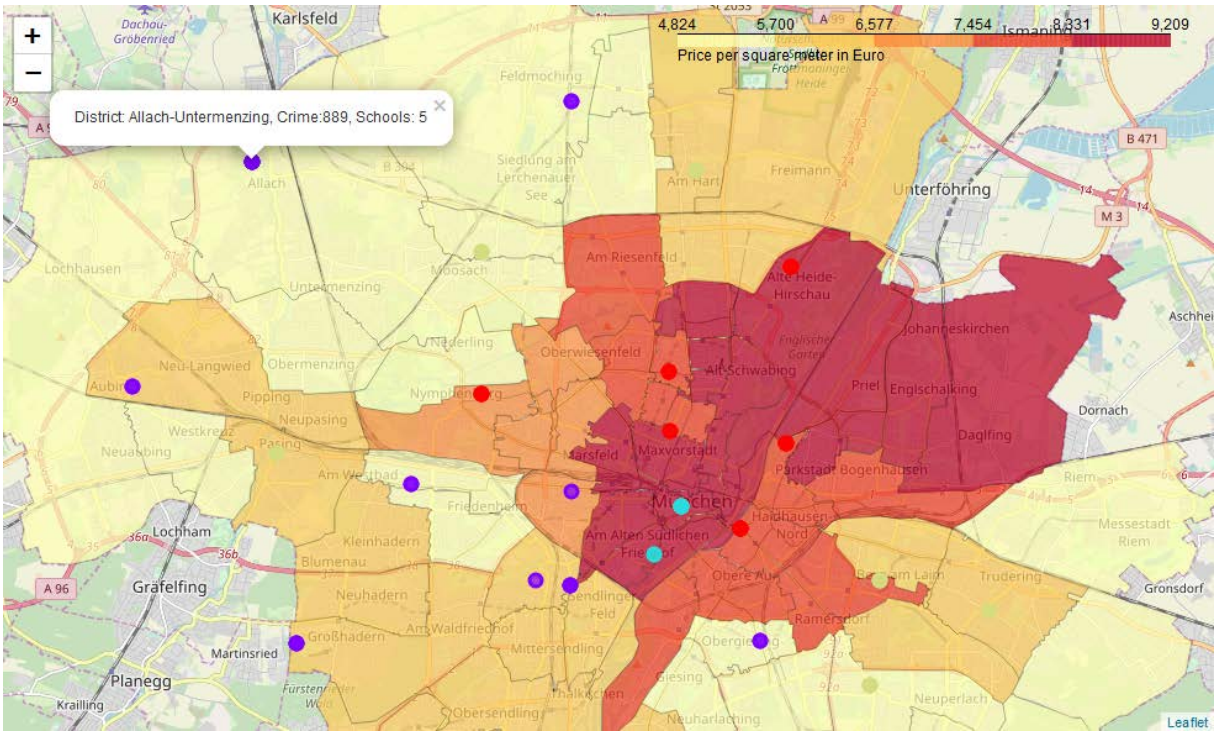
	DISTRICT	CRIME	EUROPERSQM	COUNT_SCHOOLS	Cluster Labels
1	Altstadt-Lehel	7868	9208	6	2
9	Ludwigsvorstadt-Isarvorstadt	11818	8464	2	2

Cluster 4 contains the remaining 7 district:

	DISTRICT	CRIME	EUROPERSQM	COUNT_SCHOOLS	Cluster Labels
4	Berg am Laim	2579	5921	3	3
11	Milbertshofen-Am Hart	3827	5886	3	3
12	Moosach	2776	5643	3	3
15	Pasing-Obermenzing	3240	6061	2	3
16	Ramersdorf-Perlach	4591	5590	3	3
22	Thalkirchen-Obersendling-Fuerstenried-Forstenr...	2952	5852	4	3
23	Trudering-Riem	2907	5549	2	3

Cluster 1 and 3 contain higher real estate prices and crime rates, while clusters 2 and 4 contain moderate real estate prices and crime rates. Cluster 2 seem to have a bit lower numbers of schools as compared to the others.

To get a better overview of the data, I visualized the districts in a choroplepth map by the real estate prices and show the center of each district in different colours (clusters):



Cluster 1: red Cluster 2: violet Cluster 3: turquoise Cluster 4: beige

Discussion

Based on these result, one has a starting point to choose where you would want to live in Munich. In a first case, I set the scenario of a family with much money. Thus, the price per square meter is not of importance and according to the previous tables, one may choose the district based on crime rate and number of schools. Thus, the family may choose cluster 1 as the list of districts. In another case, looking at a family with less money, they may have to look at districts of cluster 2 or 4. As it seems, cluster 3 is out of scope for families anyway, as it is very expensive with a high crime rate. Still, it is interesting to note that looking at district Altstadt-Lehel, it is a district with the highest crime rate, highest real estate prices, but the second highest number of schools. This is maybe due to the reason that at Altstadt-Lehel all subways meet, meaning it has – as the city center – the best connection to all outer regions of Munich.

With these scenarios, it seems like the clusters serve as a good starting point as to choose the district of Munich to live in.

Conclusion

In this project, I clustered the districts of Munich based on three characteristics: the crime rate, the price in Euro per square meter and the number of schools. I performed a KNN clustering algorithm to cluster the districts 4 groups. Looking at the groups in detail, we see they are distinguishable based on the three characteristics.

For young families, this analysis serves as a basis to choose a district to live in based on the three characteristics.

In further research, I will put additional characteristics into consideration, e.g., the number parks in a district, where the kids can go out and play, and also other venues kids (and also parents) are interested in, e.g., zoo, museums, etc.

Attachements

Link to the analysis on GitHub:

https://github.com/BWehner1988/coursera_capstone_project/blob/master/Capstone_Notebook_Project_Munich.ipynb

Link to the presentation on GitHub:

https://github.com/BWehner1988/coursera_capstone_project/blob/master/Capstone_Presentation.pdf