Find the optimal Restaurant Location in Ulm using Machine Learning

Introduction and Business Case

Setting

Ulm is a city in the South of Germany. Its population is ca. 120.000 (EU notation).
The city is divided up into 18 districts, which can further be split up into 55 neighborhoods.
Just like in every other city, each district and neighborhood has its own characteristics regarding attractiveness, venues and population.
This notebook tries to find locations for new restaurants in Ulm that promise high revenues.

Objective

This analysis tries to find the perfect neighborhood location for a new restaurant. Two questions are asked:

- In which neighborhoods is the average revenue per restaurant especially high?
- Are these neighborhoods similar in their characteristics (i.e. venues and therefore attractiveness they offer that lead to people visiting these neighborhoods)?

In case similar neighborhoods have similar restaurant revenues, a good location for a new restaurant can be found in those neighborhoods that are similar and have a high average restaurant revenue. The user can choose from the best neighborhoods.

Audience

The location of a restaurant is crucial for its revenue.
This analysis addresses people who want to open a restaurant in Ulm and try to find a location that boosts their revenue.

Tools

- Folium maps (including choropleth)
- Foursquare API calls to obtain restaurant and venue information
- Geopy for geocoding
- K-Means Clustering from sklearn

Data

Geospatial Information

Geospatial information, i.e. the 55 neighborhoods of Ulm as well as their borders are obtained from a **GeoJSON**-file found on this website http://daten.ulm.de/datenkatalog/offene_daten/31.
This file is provided by the city of Ulm.
It will be loaded into the notebook with the help of the "json" package contained in python.
See the image below:

Neighborhood Information

Information on the venues of each neighborhood identified is obtained from the Places-Endpoint of Foursquare, a company that accumulates gigantic spatial datasets, see https://enterprise.foursquare.com/products/places
Foursquare powers Apple Maps, amongst others.
The notebook will communicate with the Places-Endpoint of Foursquare via RESTful API calls handled by python package "requests".

Restaurant Revenue Information

The average restaurant revenue per neighborhood in Ulm is obtained from a csv-file found in a data catalogue provided by the city of Ulm. See the website

http://daten.ulm.de/datenkatalog/offene_daten/40.
The average revenue is an estimate based on statistical inference, since not every restaurant reports its revenue.
The dataset contains revenue values for 36 from 55 neighborhoods. Missing values will be filled with the average reported revenue in my code.
The dataset matches to the GeoJSON-file based on the "Neighborhood ID".
See the image below:

	ID	Revenue
0	ul-stv160	406370.0
1	ul-stv180	269584.0
2	ul-stv110	334637.0
3	ul-stv156	615312.0
4	ul-stv141	113757.0
5	ul-stv134	143617.0
6	ul-stv133	799629.0
7	ul-stv131	275716.0
8	ul-stv190	397276.0
9	ul-stv144	400177.0
10	ul-stv124	105885.0
11	ul-stv130	507082.0
12	ul-stv151	168826.0
13	ul-stv121	241691.0

Methodology

Pre-Processing

The data from the data section is pre-processed. This comprises two parts:

First, making spatial features computational by converting occurrences of venues to venue probabilities.

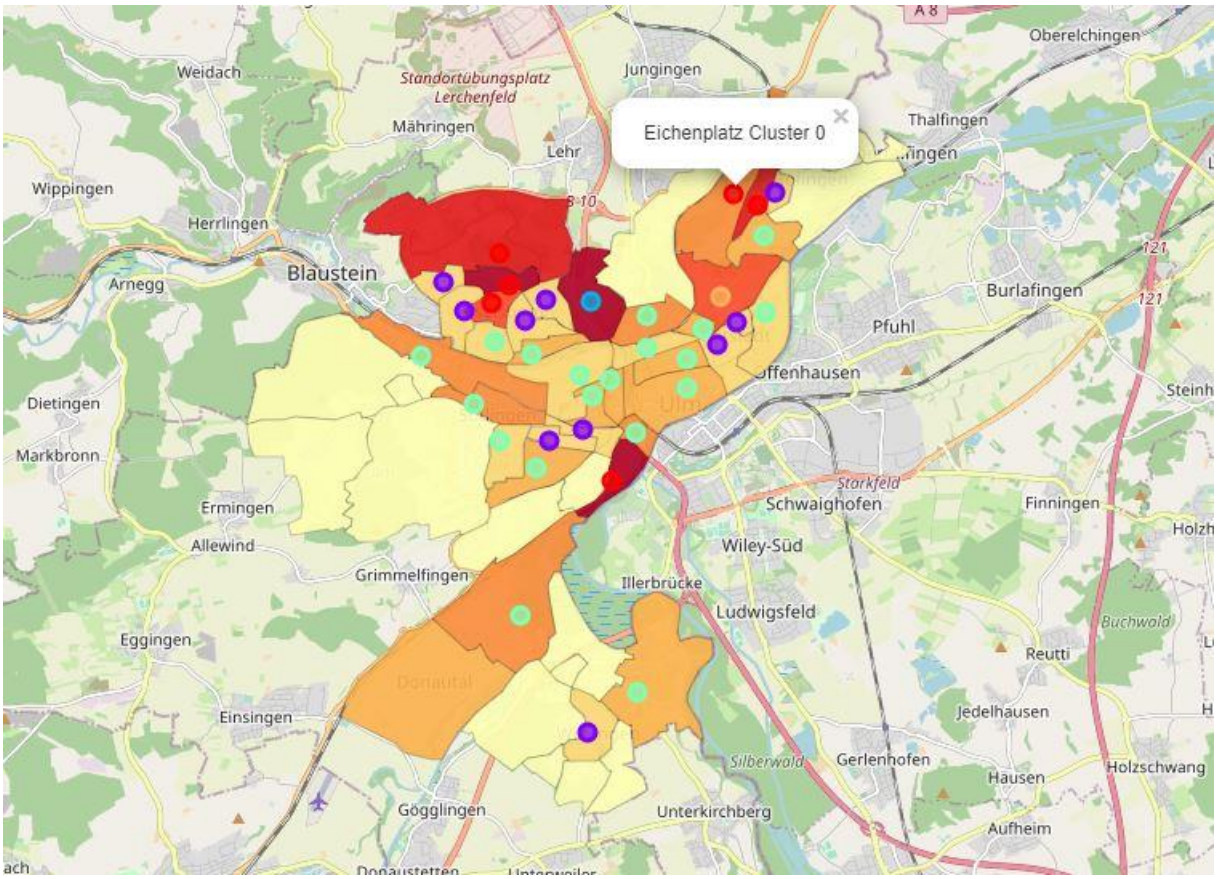
Second, maps are plotted which give an indication of the neighborhoods. Several features are added to these maps to conduct visual analyses. This comprises markers and choropleth visualizations based on the GeoJSON-file introduced above.

Machine Learning: K-Means Clustering

***k*-means clustering** is a method of vector quantization, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. *k*-means clustering minimizes within-cluster variances (squared Euclidean distances).

In general, k-means clusters the neighborhoods based on venue similarity. Venues have been pre-processed and are made computational by assigning occurrence probabilities. Each neighborhood has a specific occurrence probability for a feature/venue. Thus, neighborhoods can be compared according to their similarity.

See the image below for a picture:



Results

We can see that **neighborhoods of cluster number 0** (red circles on the map above) **on average have the highest restaurant revenue** (dark red choropleth color).

What does this mean? Well, first of all: neighborhoods of the same circle color are similar with respect to the venues they offer. This is why k-Means has placed them in the same cluster.

Neighborhoods of cluster number 0 (red circle) seem to offer similar venues that attract people to visit these neighborhoods. Once people visit these neighborhoods, they become hungry and end up in a restaurant. **This behavior boosts the revenues of the restaurants of these neighborhoods.**

What venues are the ones boosting revenues?

Each neighborhood contained in cluster number 0 possesses venues that are connected to **“nature", i.e. parks, rivers, forest, etc..**

People seem to enjoy restaurants that are placed in "green" neighborhoods.

Discussion

The outcome of this analysis is interesting. I would recommend people to open up a restaurant in one of the neighborhoods of cluster number 0.

Nevertheless, this analysis should not be the only indicator for choosing a proper restaurant location. Not taken into account are property prices. I assume that the identified locations have high rental prices, which is a significant cost factor that has to be set into relation to the expected high revenue.

In addition, I do not know how accurate the average restaurant revenue per neighborhood data is. It is the best I could find and should be workable.

In total, I think that the analysis helps to find good restaurant locations and should be taken into account as one of several factors that lead to a decision.

Conclusion

We have found a way to compare neighborhoods based on their similarity regarding venues. The neighborhoods are clustered and set into relation to expected restaurant revenues. This was done using k-Means and has created interesting results.

In order to deepen the understanding in the next step, further analyses should apply **additional machine learning algorithms**. Since I have assigned numerical values to venue features, one could run regression models that identify the numerical impact of features on expected revenues.

Thanks for taking the time and I hope you found this analysis educational.