COURSERA IBM APPLIED DATA SCIENCE CAPSTONE

Investigation of Zurich neighbourhoods and apartment renting prices

Nico Biagioli
May 2020

Business problem

- As major international, financial and high-tech industrial centre, Zurich continuously attracts people willing to relocate and work there.
- Housing is critical and the research of rented apartments at reasonable price can be problematic. The aim of the present work is to give insights on the city neighbourhoods so that a flat seeker, without prior knowledge of the city, can gain some general understanding.
- First intention is to create a model relating the venues to the district average renting prices, that are known, and with that to predict the neighbourhood prices. This will allow a more granular look at the market.
- Second objective is to cluster the neighbourhoods based on venues in order to differentiate them and understanding better what they offer.

Data sources

The following are the used data sources:

- Zurich neighbourhoods: https://en.wikipedia.org/wiki/Subdivisions_of_Z%C3%BCrich
- GPS coordinates of Zurich and its neighbourhoods from geopy, ArcGis geocoding and Google Maps.
- Zurich maps from Folium.
- Average apartment rent prices (per square meter) per district: https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bauen-wohnen/mietpreise/mietpreise-strukturerhebung.html
- Venues from Foursquare API

Methodology

1. Collect district and neighbourhood data from Wikipedia
2. Get the neighbourhood GPS coordinates from ArcGis geocoding. Verify locations on Folium map and revise wrongly located neighbourhoods by manually replacing corrected coordinates obtained from Google Maps.
3. Find the venues for each neighbourhood with Foursquare API.
4. Group the venues results for districts selecting, filtering and merging the most relevant venue categories.
5. Try to apply multilinear regression to model district price based on venue categories. Attempt with regressions failed and therefore K-Neighbors classifier was used to model 3 different levels of prices.
6. Use the K-Neighbors model above to predict neighbourhood prices and compare with relative district prices.
7. Cluster neighbourhood based on full venue categories mean frequencies with K-Means algorithm and 3 clusters. Optimal number of clusters determined with the elbow method.
8. Identify the neighbourhoods whose prices are below the average of the district they belong to and hence can be defined favourable
9. Show predicted neighbourhood clusters and price levels on Folium maps.

Results: predicted neighbourhood prices

By comparing predicted neighbourhood prices with their average district prices we can see the following cases:

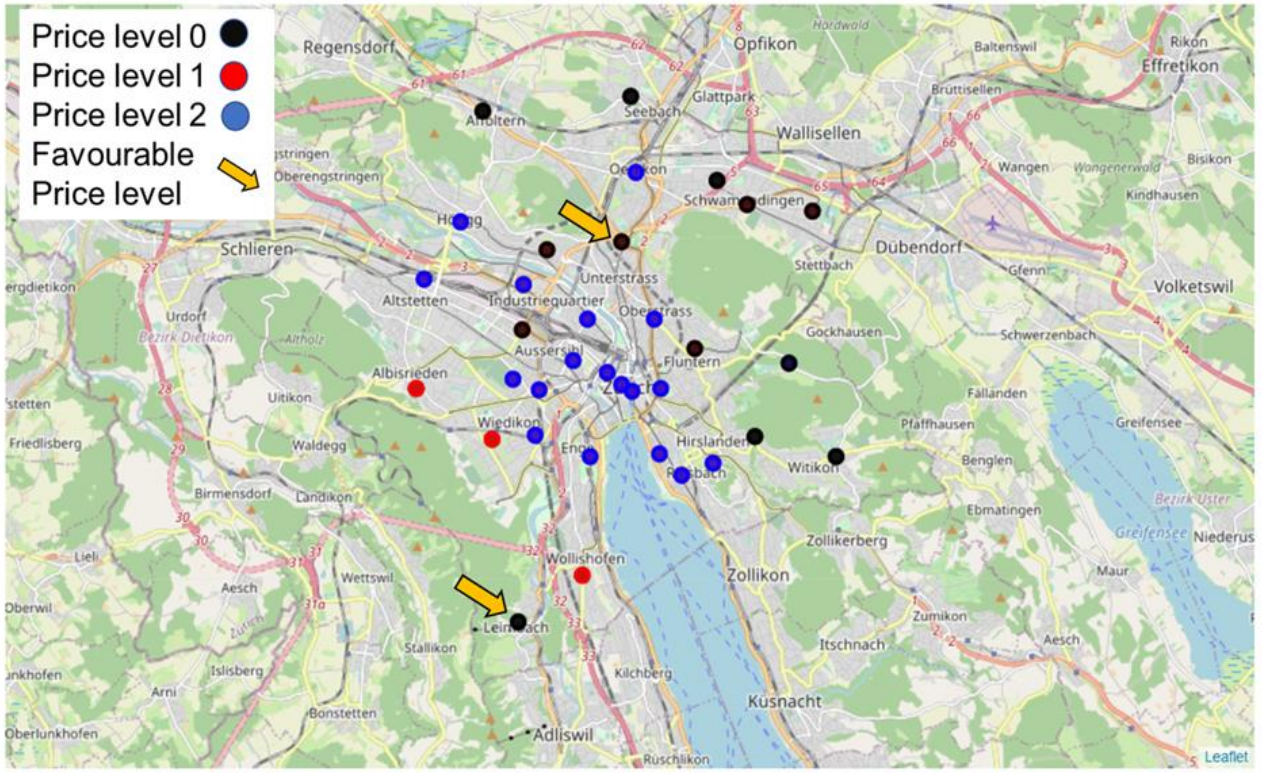
A. Neighbourhood prices being all identical to district prices: this can be interpreted as uniform price distribution within the district.

B. Neighbourhood prices being all higher or lower than district prices: this is clearly unrealistic and can be due either to model limitations or actual prices not being much related to venues.

C. Neighbourhood mean price identical to district price: this is a potentially realistic and most interesting scenario, where the non-homogeneous price distribution within the district is captured.

D. Neighbourhood mean price different from district price: less realistic case, it could give some indications of price distribution within the district but needs more analysis.

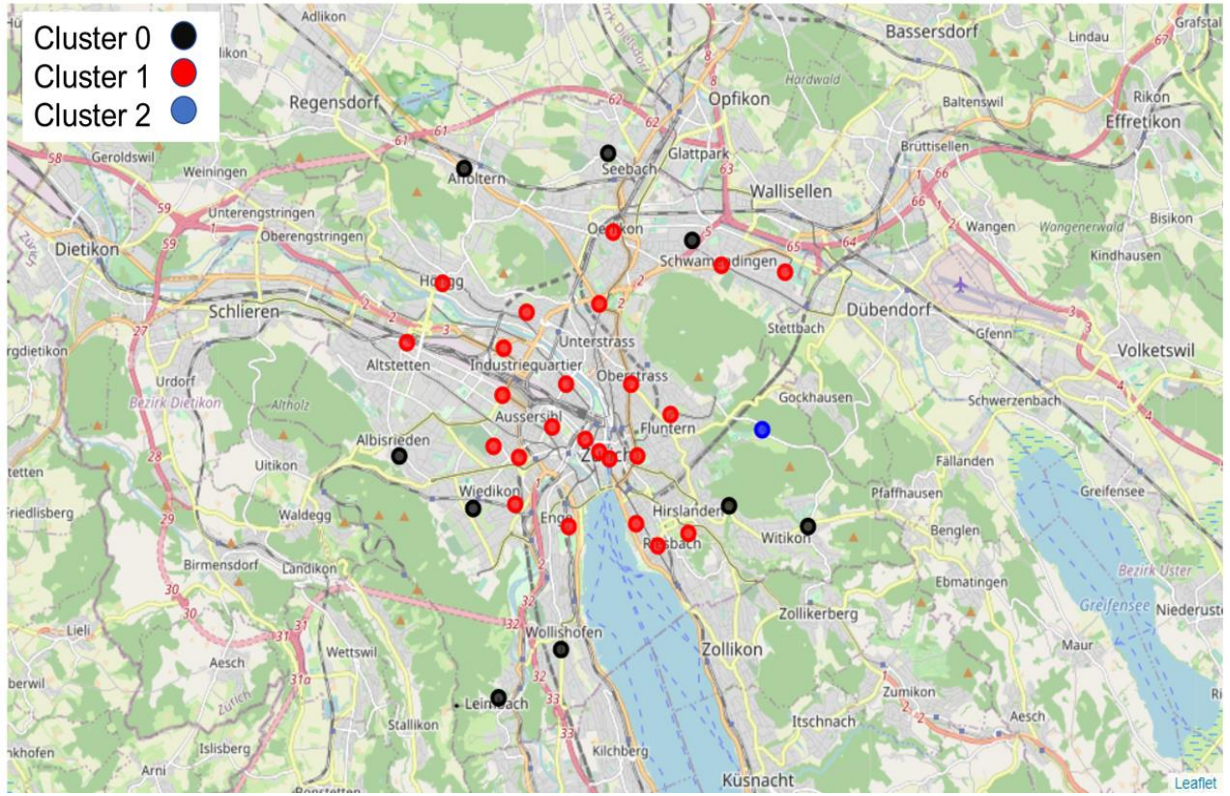
Two neighbourhoods satisfying case C are identified and considered favourable.



Results: neighbourhood clustering

Based on the results of the K-Means we can describe the 3 clusters as follows:

- Cluster 0: neighbourhoods that mostly include bus stations, supermarkets, few restaurants and very few bars. Clearly it identifies peripherical neighbourhoods.
- Cluster 1: neighbourhoods with higher prevalence of cafés, restaurants and bars. This is the inner part of the city.
- Cluster 2: only one neighbourhood falls under this category, being characterized by many restaurants in combination with high number of sport venues and markets.



Discussion

- Difficulties in predicting neighbourhood prices based on venues are related to two main issues. First: market depending also on many other factors. Second is the selection and filtering of venue categories to model prices, more solutions should be explored. Alternative classifier algorithms should also be investigated.
- As expected, the higher price levels can be found on the inner neighbourhoods. Looking the favourable neighbourhoods, one is located not far from the city center and the second one is more far but close to the lake which is a general valuable feature.
- The clustering provides useful information to someone interested in looking in detail at different venues and facilities that each neighbourhood can offer based on the whether is located more centrally or peripherically.
- By comparing the neighbourhood price prediction with clustering, results are consistent.

Conclusion

- Aim of the work was to provide insights on Zurich neighbourhoods to an apartment seeker with no previous knowledge of the city.
- A model, although rough and improvable, that predicts neighbourhood renting price levels based on venues was proposed and allows to see if prices are uniform within a district or vary based on different neighbourhoods.
- Neighbourhoods were clustered into 3 categories so that, by looking into details of their most common venues, it is possible to understand what different areas of the city offer.
- At least 2 favourable neighbourhoods that have price levels lower than their average district prices and are located in interesting areas were identified.
- Limitations of the results and further investigations and improvements were suggested.