COURSERA IBM APPLIED DATA SCIENCE CAPSTONE

Investigation of Zurich neighbourhoods and apartment renting prices

Nico Biagioli

May 2020

1. INTRODUCTION

Zurich is the largest city and economical capital of Switzerland, with more than 400,000 inhabitants (1.1 millions including the suburban area). It is a world major financial and banking center and hosts important research centres and global high-tech corporations. Zurich is also a major railway hub and it has an international airport with connections to every continent. Its high standard and quality of life are renowned and continuously attract foreign young professionals as well as families willing to relocate there. With more than 30% of non-Swiss citizens, Zurich offers a truly international and multicultural environment.
In this context, not surprisingly, housing is a major issue and there is high competition for getting rented apartments in the city. The market is tough for flat seekers and prices are generally high.

2. BUSINESS PROBLEM

Zurich municipality is divided into twelve districts (locally called: Kreis 1..12) and it includes 34 neighbourhoods. An official list of average apartment renting prices based on districts is available from the city of Zurich, but not for each individual neighbourhood.
The present work aims at helping an apartment seeker, one in particular that has no or little knowledge of the city, by providing more insights on neighbourhood differentiations based on venues and neighbourhood renting prices. In particular the following two issues will be explored:

1. Based on the city venues distribution in the 12 districts, is it possible to derive a model for a rough estimation of the renting price of each neighbourhood? This will allow to see if, within a district, prices are uniform or there are more favourable neighbourhoods.
2. Understanding the differences among the neighbourhoods based on their venues density and distribution and classify them. Compare the classifications in relation to district and estimated neighbourhood prices.

3. DATA

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.

The neighbourhood names will be used to research for the gps coordinates of each neighbourhood. The coordinates will allow map plots and also to search for the venues within each neighbourhood with Foursquare API. District average rent prices will be employed to create a model correlating the venues to the prices for the districts and using it for prediction of neighbourhood prices.

4. METHODOLOGY

First step was to extract the neighbourhoods and district data from Wikipedia. Due to the multi-table layout in Wikipedia and the low amount data, the most time-efficient way was to copy paste data into an Excel sheet, save it as .csv and import it into Jupyter notebook.

The following tasks included research of the neighbourhood GPS data. The search was performed with ArcGis geocoding library, based on neighbourhood names. The results are added to the main dataframe and the locations are shown on the map of Zurich using Folium. By examining the map, I could see that several neighbourhoods were not correctly located, because the coordinates provided by geocoding were inexact. I then created a list of those wrongly located neighbourhoods and manually researched their coordinates on Goggle Map. The revised coordinates were added to the dataframe and the neighbourhoods again shown on the map, this time they all resulted correctly placed.

Foursquare API was used to search for the venues within a range of 500 m radius of each neighbourhood. Because our analysis started by creating a model that relates venues to district renting prices (available information), the dataframe including venues and their categories was grouped by districts neglecting neighbourhoods (for the moment). In order to have a manageable understanding of the most common and relevant categories of venues for each district, the categories were filtered, reduced and merged based on major type (i.e.: restaurants, bars, sport etc.). This also allowed a more efficient and effective way of modelling, that otherwise would have to deal with way too many independent variables.

A linear regression was used to model the dependency of district renting prices on venue categories. The three categories having the strongest correlation to price were selected and a train and test sets used. The resulting regression can accurately predict the train set but largely fails on the test set. Previous attempts of using different combinations or single independent variables and polynomial regressions failed as well and are not included in the attached notebook code.

Due to the results above, it was decided to use a classification algorithm instead of a regression. K-Neighbors was selected as it provides a simple and easy to understand model with no need of many adjustments. For the present work scope of rough prediction, it is appropriate. To use a classifier, price data needed to be transformed into discrete price levels. After a few attempts, the most accurate results were produced with 3 price levels, see picture 1.

	Average square m price (CHF)	Price level
Distrct 1	17.9	2
Distrct 8	16.1	
Distrct 10	15.9	
Distrct 11	15.7	
Distrct 2	15.6	1
Distrct 6	15.4	
Distrct 7	15.4	
Distrct 9	15.4	
Distrct 12	14.9	0
Distrct 3	14.9	
Distrct 5	14.9	
Distrct 4	14.7	

Picture 1

The best accuracy was obtained with K=3. That provides 100% accuracy on the test dataset and 75% on the train set.

The next steps included to use the classifier trained on the district venues data to predict neighbourhood prices. A dataframe with the same venue macro categories was created, this time grouped by neighbourhoods and the neighbourhoods predicted prices calculated.

To obtain more detailed insights on the neighbourhoods an unsupervised K-Means clustering algorithm was run. Since this time we want to have a look at the details of the different venues and perform a clustering, all the categories provided by Foursquare were used going back to the original initial complete dataframe. The frequencies of occurrence of the categories for each neighbourhood were used to run the K-Means. The optimal cluster number resulted to be 3 according to elbow method. For each resulting cluster, a dataframe with the 10 most common venue categories for each neighbourhood is shown and also includes district, district price and predicted neighbourhood price.

Lastly, the neighbourhoods having predicted prices lower than the average of the district they belong to are shown along with their most common venues. This selection can help an apartment seeker to identify predicted convenient neighbourhoods and understand their facilities.

5. RESULTS

5.1 Prediction of the neighbourhood renting prices

As described above, the attempts of using linear (and polynomial) regression failed. The classifier based on 3 price levels only provides a rough estimation. By observing the complete dataframe (see attached notebook) that includes both neighbourhood predicted price levels and district price levels, it possible to observe 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.

Case C, occurring for districts 2 and 6, should be particularly looked at in order to identify more favourable neighbourhoods within their districts.

5.2 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.

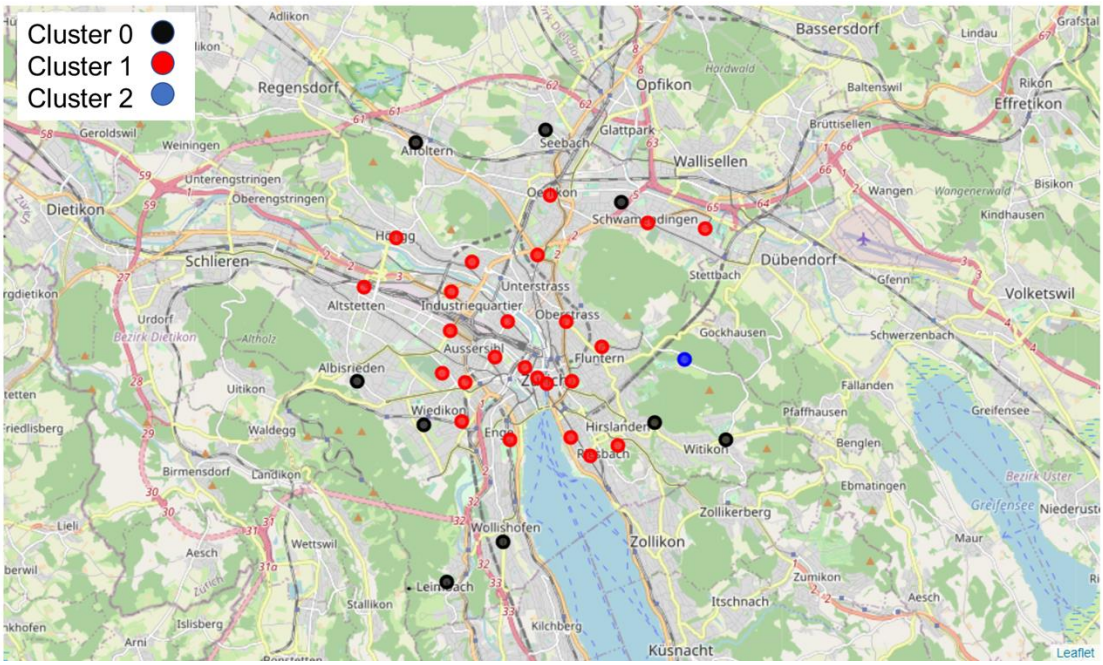

Figure 2: Zurich neighbourhood clusters

5.3 Favourable neighbourhoods

The prediction of neighbourhood prices allows identification of the neighbourhoods having prices lower than the average of the districts they belong to. Thus providing an interesting option to explore for an apartment seeker, because they are located within district whose average value is higher than the neighbourhood. Considering what said above (paragraph 5.1) when discussing the price prediction results, although 9 neighbourhoods fall into this category (see notebook), only 2 can be realistically taken into consideration. These are the ones that satisfy case C (paragraph 5.1), namely Leimbach and Unterstrass. All neighbourhood price levels and the favourable ones are shown on the map in figure 3.

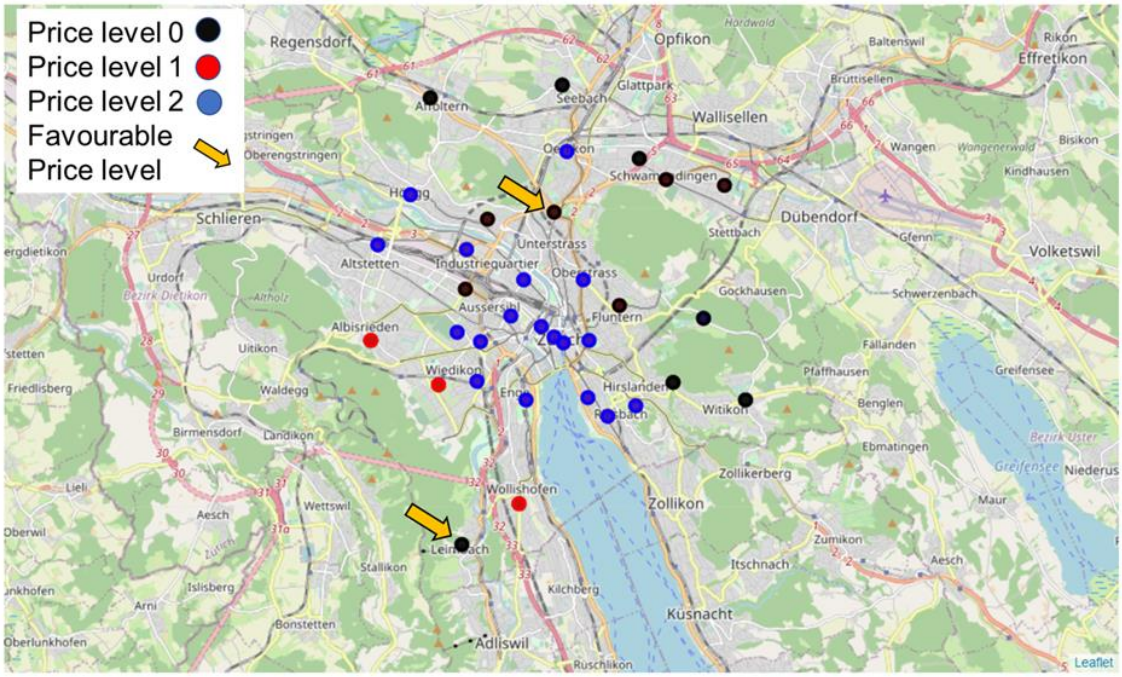

Figure 3: Zurich neighbourhood price levels

6. DISCUSION

The prediction of neighbourhood prices based on venues resulted to be challenging. A regression model could not be found and the adopted classifier only provides very course indication and prediction of 3 price levels. The are two possible main reasons. First, renting prices can only be partially correlated to venues. They clearly depend on other market factors including, among many, neighbourhood/district demography, real estate characteristics etc. Thus a model only considering venues cannot capture price trends effectively. Second problem is the selection of the venue categories to model the prices. Foursquare API provides more than 170 categories for the city of Zurich. In this work 8 macro category types were identified and included the most popular categories and those that would better characterize a neighbourhood/district.
Selection of different macro categories, different aggregation of categories or related strategies could be investigated to asses if they can provide a better and more accurate neighbourhood price prediction. Another point of possible future investigation can be the selection of the classifier algorithm. The reason of choosing K-Neighbors was mentioned in section 4, but it's certainly worth to explore the use of other approaches. I would particularly have a look at Naïve Bayes classifiers, due to the ability of dealing with low amount of data.
Considering the results of predicted neighbourhood prices, looking at the map, as clearly expected, the higher price levels can be found on the more central 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 characteristic.
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 it is located more centrally or peripherically.
By comparing the neighbourhood price prediction with the clustering we see that almost all neighbourhoods belonging to Cluster 1, the central ones, have highest price level (2) with few exceptions. The outside neighbourhood, Cluster 0, have all low price levels, 0 or 1. This shows at least consistency between the two parts of the investigation even though different venue categories were considered, as we have seen.

7. CONCLUSION

The present work aimed at investigating the neighbourhoods of Zurich, through the analysis of the venue distribution, to provide a rough estimation of the neighbourhood renting prices and helping differentiating and clustering the areas based on the most common venues. The analysis is particularly suitable for an apartment seeker, with low or no knowledge of the city, who would like to have a first understanding before a deep dive into the market search.
A classifier model trained on known average district prices was used to roughly predict neighbourhood prices. By comparing district and neighbourhood price levels, at least two neighbourhoods resulted particularly favourable because offering lower prices than their average district ones and in combination with interesting locations.
Limitations of the results, with particular regard to the price prediction model, were described and possible future ways of improvements proposed.