The Rules Behind Housing Prices in the Neighborhoods of Munich City

- An applied data science project with machining learning by Nan Chen

Description: this project demo shows how to apply simple approaches in data science to obtain insights into some data as common as the housing prices in Munich. The methods adopted by the project include python libraries `numpy`, `pandas` and `lxml` to extract and process data from online source, and `geopy` to transfer address into geographical coordinates. The data visualization part is performed by `matplotlib` for plotting charts, `seaborn` to reveal variable correlations, as well as drawing maps by `folium` plus detailed features by `requests` and `json`. As next step, the location provider `Foursquare` is used to search for the nearby venues. After making a better generalization based on the original venue categories provided by `Foursquare API`, they can be processed by `KMeans` for neighborhood clustering. In this way, all neighborhoods are clustered by the machine by their shared interesting characters. As the final step, the counting numbers of venue categories are fed into `sklearn` to formulate an insightful price model. The entire data set is divided into a train group for developing model and a test group for checking result. Different models (*Simple Linear Regression* vs. *Multiple Linear Regression* vs. *Multiple-variable Polynomial Fit*) are attempted to select the best. Finally, the new price models can be proved by feeding it with some new data, in which the estimated prices are found to be meaningful with tolerable errors compared to the reality.

1. Introduction

The data description about the Munich city neighborhoods (districts or “Bezirk” in German) can be found from wiki [1]. The tabular information can be extracted by `pd.read_html` and processed into a district data frame. From the area and population numbers we can calculate the population density. The housing prices of Munich by districts are listed on a local real-estate website [2]. Their neighborhood names are mostly consistent, yet not exactly the same. Thus, a treatment has to be made to align the price data to the district data. I dropped accidentally three rows of “unofficial” districts, but they will be still made use of later in the project to testify my price model.

The population density and the prices can be combined together as the following plot:

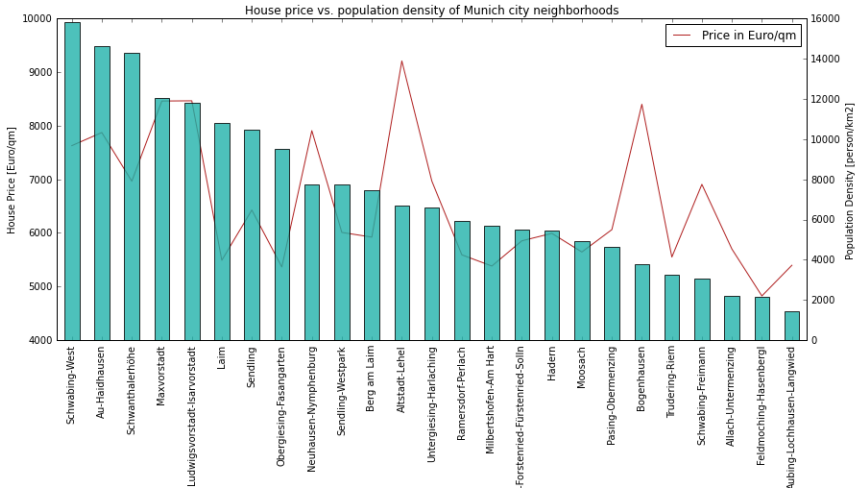

Fig.1: Population density and housing prices per neighborhood of Munich city

This data set will be used as the starting point of analysis throughout the whole project.