Estate prices in Helsinki

Discussion of the background

My client is doing his business as estate agent. So far he is not quite acquainted with data analysis and he is wondering whether he can obtain any benefit from analysis. Thus he has asked me to do some common analysis in his business. His agency is located in Helsinki and mainly his customers live in central areas of Helsinki. Mostly assignments are located in that area too. Therefore the basis of analysis will be neighborhoods of Helsinki. There are 60 neighborhoods in Helsinki. Some of them are not taken into account, because these areas are not populated. This is due to being sea, harbor area or industrial area.

Description of the problem

In this analysis I will provide my client common characteristics of neighborhoods of Helsinki. I will focus on characteristics related to pricing process of estates. The market is always right but in the the case I find bias of prices in certain neighborhood, my client will profit from known bias. Using this knowledge he is able to use more precise prices, when price estates. It is not quite probable to find this kind of bias because, as mentioned, the market is mostly right. In all cases I will provide pricing model to my client. He will use this model in accordance with his knowledge and experience and this way my analysis give him gain.

Data

All data will be per neighborhood and data is originated from years 2017-2018. In rare cases the data is from 2016. This kind of inconsistency need to be included in the final report. The following data will be used: Prices of estates in closed deals, Population, Population cohorts, housing stock, amount of square meters per house, amount of square meters per inhabitant, share of owned houses, share of rental houses, share of municipally subsidized houses, education, healthiness, number of workplaces and amount of labour force.
In addition to these, I will also use FourSquare to find relevant location data, venues of Helsinki. Interesting venues in area may raise the value of estates.

Source of Data

The study is based on neighborhoods of Helsinki. In order to use maps while reporting to my client, I need geological locations of neighborhoods. Fortunately this information is provided here. https://hri.fi/data/fi/dataset//helsingin-kaupunginosat. The information in in KML format, that is actually XML datatype.
The needed statistical information is publicly available as open data from here: https://hri.fi/fi/.

Data preparation

All the datasets are in excel-format. Those datasets are produced manually and therefore lot of cleaning work is needed. I will import each dataset into Pandas data frame, end remove unneeded rows and features using Python as well as I will trim names of neighborhoods, remove language variants of names, rename columns etc. This will be quite laborious stage.

Another cleaning task is to transfer KML formatted location data of areas into JSON format. The reason for this is Folium library. Folium uses JSON. Fortunately lxml-library provides easy means to parse KML. Therefore it is easy to transfer locations from KML to JSON.

Data consistency

Consistency is problem, because the found data utilizes various area division principles. For example some data uses postal code areas and some other uses boroughs. The analysis under construction utilizes neighborhoods and therefore all data will be transferred manually to neighborhood division. That process is dependent of additional information. The needed information is located here: https://fi.wikipedia.org/wiki/Helsingin_alueellinen_jako. Using this info it is possible to do the work.
Also the location of venues from FourSquare may need some preparation too.
After consistency check, all relevant data will be combined into single data frame and the data is ready for analysis.

Methods

The methods of analysis are the same as in courses. The first I will try clustering and the I will continue to classification. With these studies I may find areas of Helsinki, that share same characteristics. After identification of these areas, I will try several regression models. My intention is to find optimal model to determine prices of estates.