Relationship between Housing prices & maturity index of different postal areas of Helsinki 2018

Capstone Project - The Battle of the Neighborhoods (Week 2)

Yun Xiao

1. Introduction

1.1 Description of the problem & Discussion of the background

Helsinki is the capital city of Finland. Located on the shore of the Gulf of Finland. Together with the cities of Espoo, Vantaa, and Kauniainen, and surrounding commuter towns, Helsinki forms the Greater Helsinki metropolitan area, which has a population of nearly 1.5 million. The region's land area accounts for 1.2% of the area of the entire country, but it accommodates 27% of the Finnish population.

Helsinki metropolitan area is the dominant region in the Finnish property market. The area is well recognised among foreign investors, and many large global players only invest in the capital region. Positive supply developments have, in turn, bolstered housing investment while at the same time mitigating price growth significantly.

In this project, we are aiming to analyse the relationship between the housing price & maturity index of various postal districts in Helsinki.

1.2 description of the data and how it will be used to solve the problem

To solve the above problem, we first need to define maturity index. As residents enjoy various amenities that have been developed over the years, which include transport networks, shopping malls, schools, and parks. We can use the diversity and abundance of the district to indicate the maturity index.

Foursquare defines a three-level hierarchical structure of categories for venues. With nine top level categories: Arts & Entertainment, College & University, Food, Professional & Other Places, Nightlife Spot, Outdoors & Recreation, Shop & Service, Travel & Transport and Residence.

Here is the thought: we can use Foursquare to get the information of different categories or venue one district has, to stand for its diversity and abundance of the district, in another word, the maturity index.

Then using Machine Learning method to analyse the relationship between housing price & maturity index of various postal districts in Helsinki.

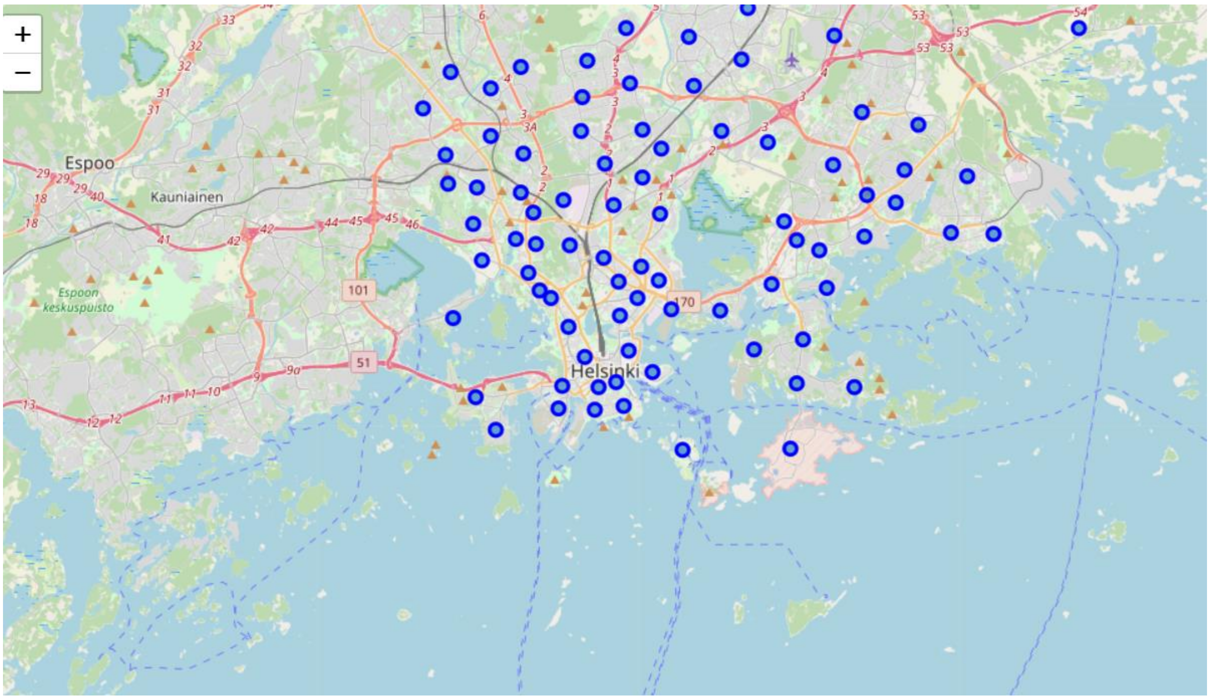
2. Data

Giving the problems, we list the data that will be used as below:

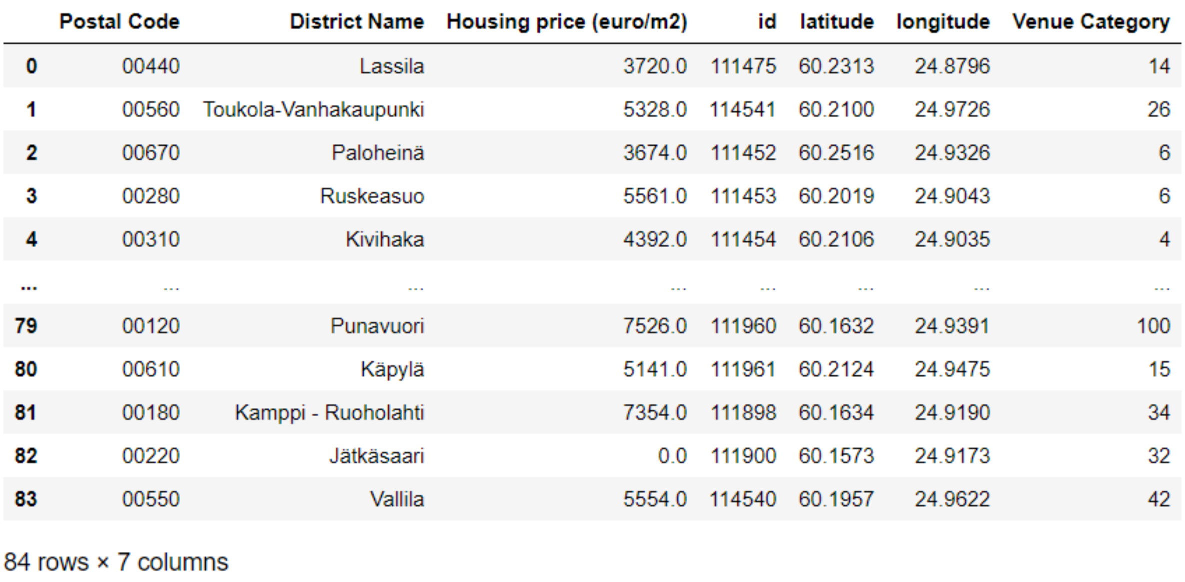
- The “Statistical yearbook of Helsinki 2019”, which includes the average housing sale price of 2018 for various postal districts in Helsinki, has been collected from Helsinki Region infoshare. The .xlsx table file has been downloaded. Data has been cleaned and reduced. We will used this data later to analyse.

- Geocoder Python package will be used to obtain the coordinates of different postal code area of Helsinki.
- Helsinki districts boundaries by postal code .geojson file is obtained from Helsinki City Survey Division. This will be used to show the districts map of Helsinki.

- Forsquare API will be used to get the number of different venues of given districts of Helsinki.



After pre-processed, we have all our data settled. Let’s have a look here



	Postal Code	District Name	Housing price (euro/m2)	id	latitude	longitude	Venue Category
0	00440	Lassila	3720.0	111475	60.2313	24.8796	14
1	00560	Toukola-Vanhakaupunki	5328.0	114541	60.2100	24.9726	26
2	00670	Paloheinä	3674.0	111452	60.2516	24.9326	6
3	00280	Ruskeasuo	5561.0	111453	60.2019	24.9043	6
4	00310	Kivihaka	4392.0	111454	60.2106	24.9035	4
...	...	...	...	...	...	...	...
79	00120	Punavuori	7526.0	111960	60.1632	24.9391	100
80	00610	Käpylä	5141.0	111961	60.2124	24.9475	15
81	00180	Kamppi - Ruoholahti	7354.0	111898	60.1634	24.9190	34
82	00220	Jätkäsaari	0.0	111900	60.1573	24.9173	32
83	00550	Vallila	5554.0	114540	60.1957	24.9622	42

84 rows × 7 columns

We will use this dataset to analyse in the below.

3. Methodology

In this project we will direct our efforts on detecting the relationship between Housing prices & maturity index of different postal districts in Helsinki.

In first step we have collected the required data: the housing price based on different postal areas in Helsinki of 2018. Then we obtained the location(coordinates) of every postal code area in Helsinki.

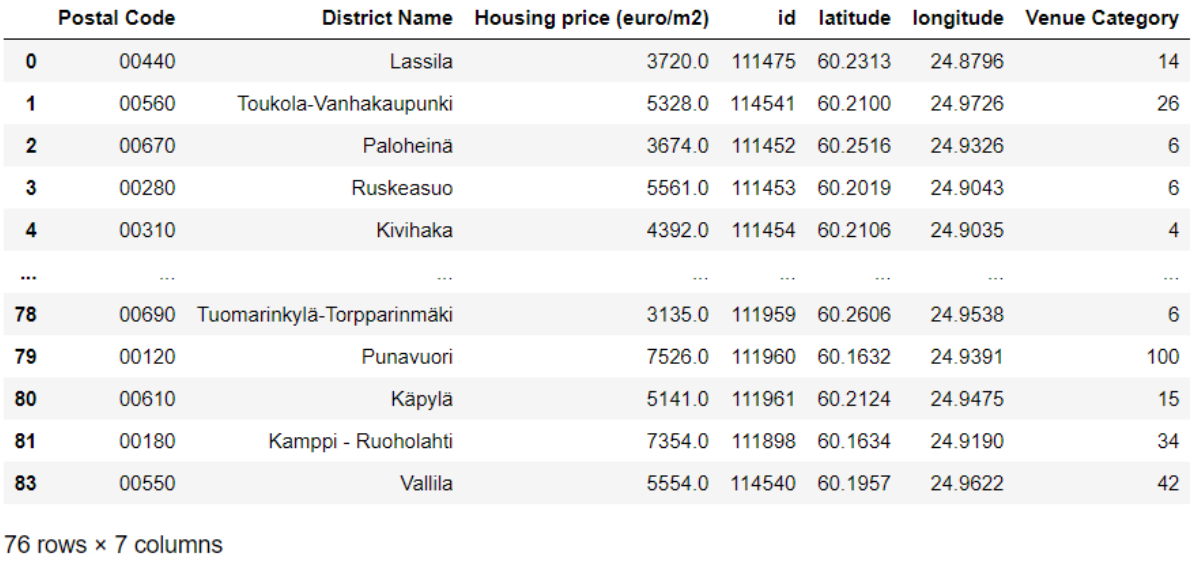
After that, we calculated the Venue Category numbers via foursquare (according to Foursquare categorization).

In third and final step we will focus on exploration of the relationship between Housing prices & maturity index of different postal districts in Helsinki. We will use Machine Learning Method - Regression or Non-regression. our features will be Venue Category and our label will be housing price.

4. Analysis

As mentioned above, we can see there are some 0.0 values for housing price, which means there are no sales in that areas due to different reasons. In order to make the results more reliable, we will drop the areas without housing sales. And here is our new dataset, which has a shape of 76 x 7. In our dataset, we have following information: postal code of this area, coordinates(both latitude and longitude), the corresponding district name, venue category and also housing price as well.

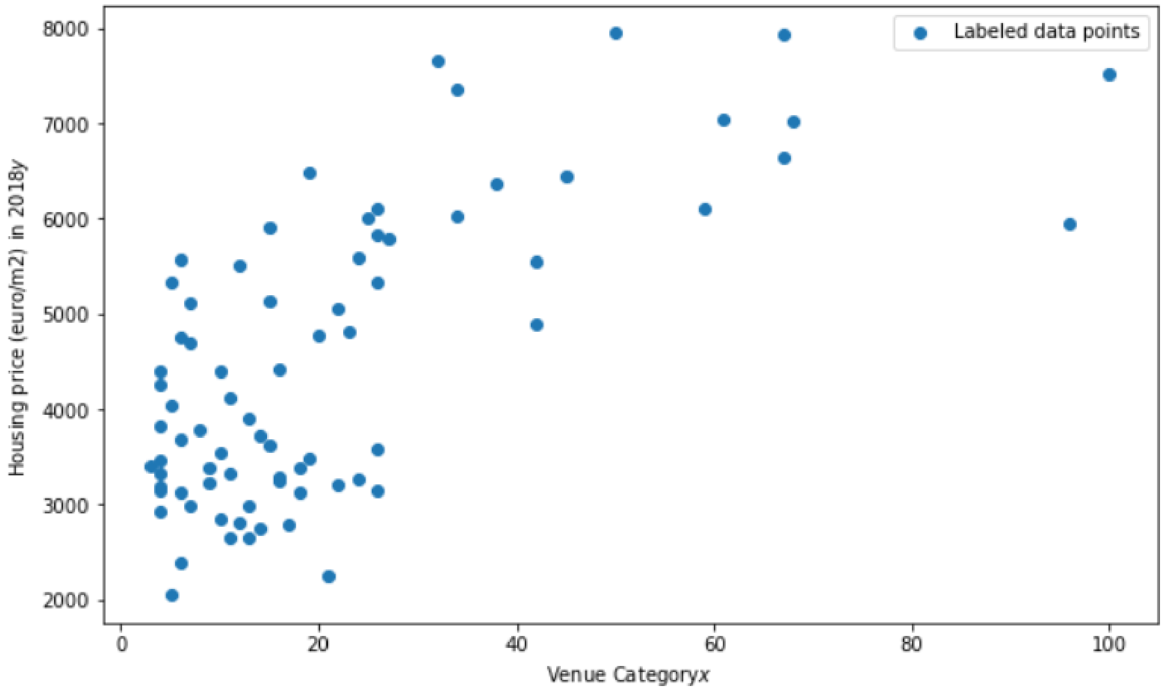
As planned, venue category will be our feature data, and housing price would be the label data.



	Postal Code	District Name	Housing price (euro/m2)	id	latitude	longitude	Venue Category
0	00440	Lassila	3720.0	111475	60.2313	24.8796	14
1	00560	Toukola-Vanhakaupunki	5328.0	114541	60.2100	24.9726	26
2	00670	Paloheinä	3674.0	111452	60.2516	24.9326	6
3	00280	Ruskeasuo	5561.0	111453	60.2019	24.9043	6
4	00310	Kivihaka	4392.0	111454	60.2106	24.9035	4
...	...	...	...	...	...	...	...
78	00690	Tuomarinkylä-Torpparinmäki	3135.0	111959	60.2606	24.9538	6
79	00120	Punavuori	7526.0	111960	60.1632	24.9391	100
80	00610	Käpylä	5141.0	111961	60.2124	24.9475	15
81	00180	Kamppi - Ruoholahti	7354.0	111898	60.1634	24.9190	34
83	00550	Vallila	5554.0	114540	60.1957	24.9622	42

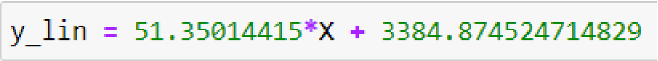
76 rows × 7 columns

We first scatter the data to have the initial ideas about the possible models. Here is the plot of our data, we choose the 'Housing price ' as ylabel, and the total 'venues category' as xlable, since we hypothsis that the housing price is somehow related to the maturity index in this area.

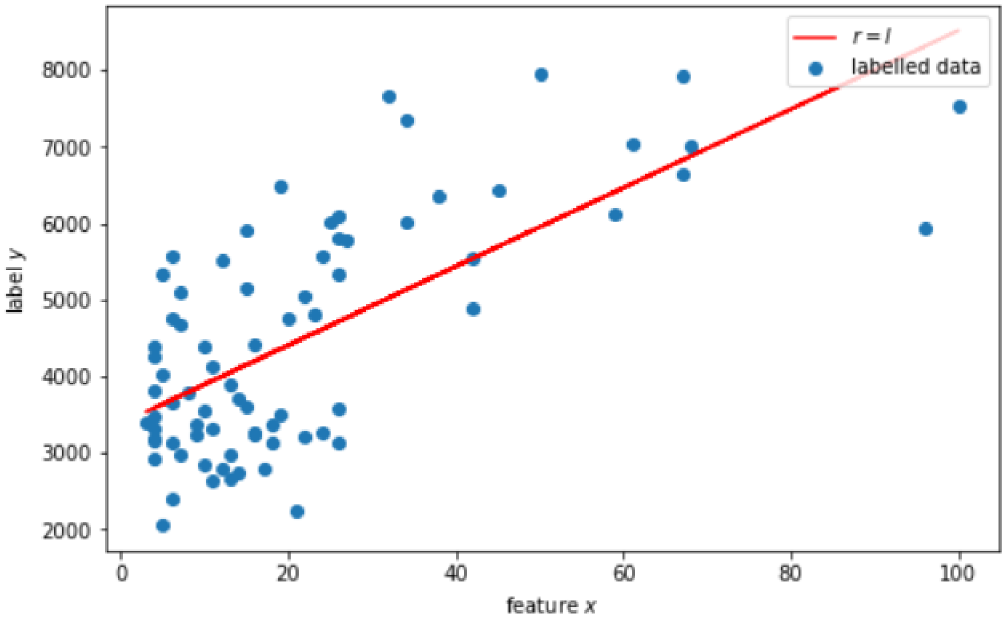


From an initial look at the plot, we find out somehow liner Regression trend or Logarithmic function trend. In this study, we will examine both models.

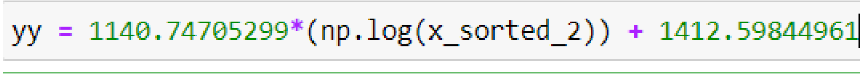
We fit the data to linear Regression model, and here is the output we got:



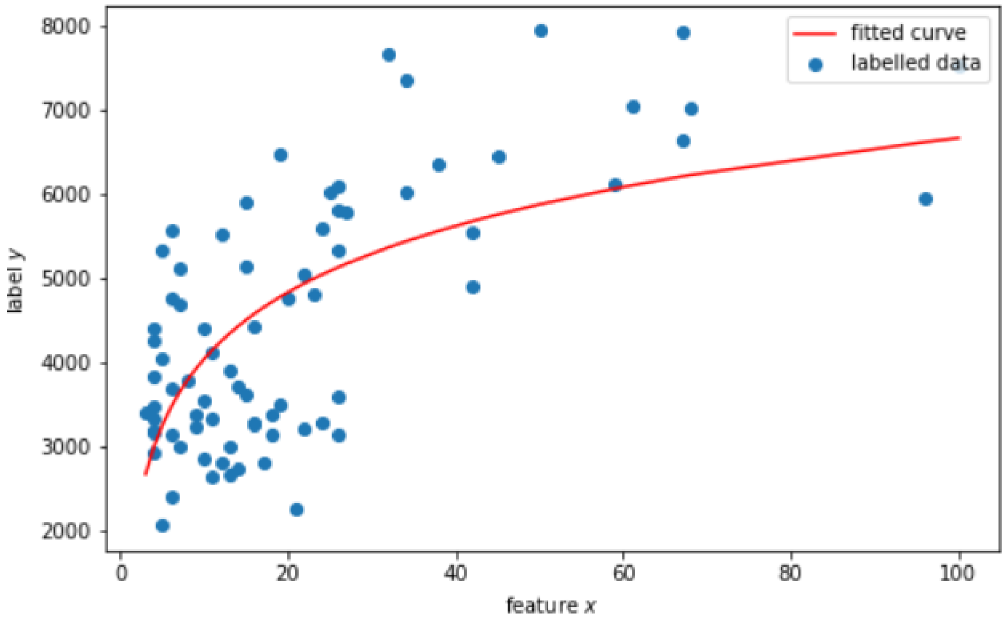
```
y_lin = 51.35014415*X + 3384.874524714829
```



Then, we fit a Logarithmic function, which got the relationship as follows:



```
yy = 1140.74705299*(np.log(x_sorted_2)) + 1412.59844961
```



In order to evaluate both above models to choose the better one. We calculate the R2 scores for both models, **Logarithmic function: R2-score: -0.18, while, Linear Regrssion: R2-score: 0.47**.

It seems that both models are not very suitable for the prediction. But linear Regression is obvisouly has much better prediction than Logarithmic function. Further studies can be done using PolynomialFeatures.

5. Results and Discussion

This project shows the initial relationship between the housing price and maturity index of different districts. Generall speaking, if more different venues in this aera, the housing price in this area would be higher than others.

Linear Regression give us a R2 score about 0.47, it is not strong, but still related. The possible reasons could be as follows:

- When obtained the venues number, we have limited the number to 100, and radius to 500, but the area radius could be larger than 500 meters. and some areas have obviously more than 100 venues.
- We sum up all different categories together to indicate the maturity index. We could in the further study, separately calcaute the numbers of different categories(says Foursquare has 9 root categories, we can count the number for each category, then we have 9 different features), with more acurate features, the 'weights' that contribute to the housing price can be also obtained.

Thus, we can improve this model in the furture from those 2 aspects.

But generally, we can say that the maturity index from one district has relationship with its housing price.

6. Conclusion

Purpose of this project was to check the relationship between housing price and maturity index of different postal districts in Helsinki. Final decision on relationship is around 0.5 related, but the analyse can be improved and experiment can be better organized by the discussion above. And of course increase the dataset number to the whole Finland.