A reference source for house buyers in the Capital Region of Finland

Capstone Project - The Battle of the Neighbourhoods

Applied Data Science Capstone by IBM/Coursera

Hang Nguyen - April 2021

Table of contents

1. Introduction

1.1. Background of housing in Finland[1]

Finland is one of the safest countries in the world, and all residential districts in Finland are safe to live in. Homes are well equipped and warm throughout the year, and even in cities, nature is nearby. Public transport works well in cities.

Income differences in Finland are lower than in most other countries. This also shows in housing; differences between residential areas are not as great as in countries that have high income differences. Finland has a wide range of housing options. A resident can rent or buy a home, and there are also other options that fall between the two. About two-thirds of Finns live in owner-occupied homes. About one-half of Finns live in a detached or semi-detached house. Wooden houses are common in Finland. About one-third of Finns live in a block of flats, but in cities it is more common. On average, about two people live in the same home. Over 40% of Finns live alone.

In the long term, buying a home is often cheaper than renting. Housing prices vary widely across Finland. The price is affected by the location and age of the home. In large cities, housing costs more than elsewhere in Finland. However, a home in a large city is usually a safe investment; housing prices will probably not fall. In rural areas and small towns, homes are much cheaper, but it is more difficult to sell the home later.

Finnish cities have many housing estates where buildings were built in the 1950s, 1960s or 1970s. There are also many new buildings in large cities. Homes in new buildings are usually more expensive than in old buildings. The average size of Finnish homes is about 40 m² per person, but the differences are great. The largest amount of space is available to those who live alone.

1.2. Description of the problem

Currently, I am living with my family in Finland. Our family is planning to buy our own house to settle down. We are a middle-class family with a little child and buying a house is a huge investment for us. That's why we would like to have more information of a certain living area before deciding. Our criteria contain average house price and the availabilities of social services of the area especially services for little children. We do not care about pubs, bars or tourist attractions. Besides, during over a year of pandemic, mostly we work remotely from home, this means we can choose a house in countryside or quiet area but near to public transportation. Therefore, although we prefer the urban area to live, we also open to see options from quiet area in the capital area of Finland

The Finnish Capital Region consists of four municipalities with city title, Helsinki, Vantaa, Espoo and Kauniainen and forms the core part of the Helsinki-Uusimaa Region. Total population is about 1.2 million (2020). Most of the inhabitants live in the urban areas of the cities, but within the boundaries of these cities there are also suburban and rural

[1] Source: Housing problem - https://www.infofinland.fi/en/living-in-finland/housing/housing-problems

areas. The region is both a major commercial and a cultural center. Helsinki, the University City, is a modern city with high-tech offices for international companies. Helsinki is also the location of many academic and governmental institutions, as well as historical buildings[2]

1.3. Who are interested in?

As I mentioned before, buying a house is a huge investment and in long term, owning a house is cheaper than renting. People who want to buy a house will try to look for as much as information they can get. Therefore, every house buyer in Finland will be interested in the analysis of this project, especially family with children.

2. Data acquisition and wrangling

Based on the description of given problem, factors which will effect the acquisition of data are:

- Housing sale price.
- Social services for basic wellbeing especially for children
- Public transportation

Following data sources will be needed to extract/generate the required information:

- The list of areas in the capital region of Finland. This information can be found in National Statistical Service of Finland
- Geographic information of each areas such as latitude and longitude can be calculated by using Pgeocode library
- House price or price per square meter in each area, this information also can be found and obtained from National Statistical Service of Finland
- Social services (grocery store, education services, hospital, etc.) and public transportation (train/metro station, etc.) will be obtained by using Foursquare API

From National Statistical Service of Finland, I found the excel file of all postal codes for each district in Finland. There are 3027 rows in total, however, because I focus on finding house in the captial region of Finland. The Finnish Capital Region consists of four municipalities with city title, Helsinki, Vantaa, Espoo and Kauniainen. So, I extract data from 4 municipalities only.

[2] Helsinki – Uusimaa Regional Council - https://www.uudenmaanliitto.fi/en/helsinki-uusimaa_region/finnish_capital_region

municipality	postalcode	area
Espoo	46	46
Helsinki	84	84
Kauniainen	1	1
Vantaa	37	37

I need the data of latitude and longitude of each area for later use in FourSquare API to obtain more detail of venues. To get this information, I use Pgeocode library to get latitude and longitude based on the postal code

Also, from National Statistical Service of Finland, I found statistical data of housing sale price for each area or district in Finland. The statistics were reported in 2020 so the data is quite up to date. In this data, I need to do some cleaning because the postal code, name of areas and municipality are mixed. So, I extract first character of the name column and put them into new column named 'postalcode'. Postal code is unique, so I used it as a key to merge data together. There are quite a lot of missing values of housing sale price (no recorded) in data file. I decided to delete all rows contain missing house sale price. Here is my data set after cleaning and merging with postal code and geographic data.

	postalcode	area	municipality	latitude	longitude	price_per_m2
0	00100	Helsinki Keskusta - Etu-Töölö	Helsinki	60.1714	24.9316	7575
1	00120	Punavuori	Helsinki	60.1632	24.9391	8160
2	00130	Kaartinkaupunki	Helsinki	60.1645	24.9487	7825
3	00140	Kaivopuisto - Ullanlinna	Helsinki	60.1578	24.9525	8713
4	00150	Eira - Hernesaari	Helsinki	60.1570	24.9369	8367

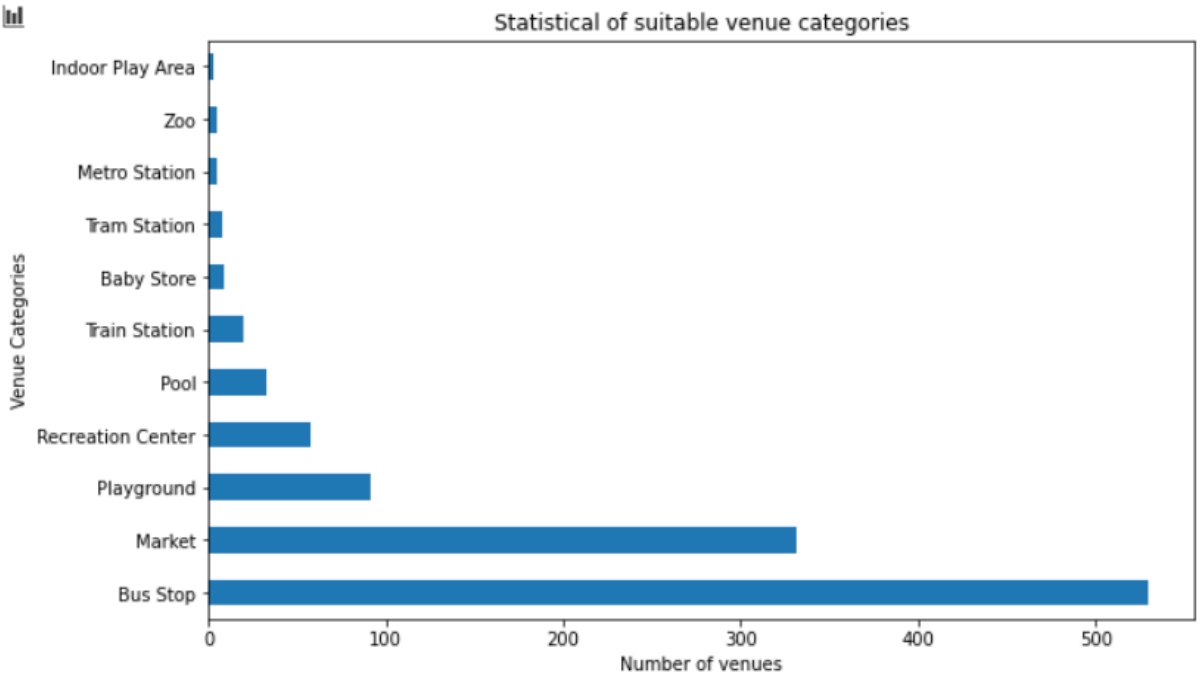
The next step is gathering data of social services in each area. This information was collected by Four Square API. Total 4.842 venues in 148 areas are returned from Four Square API service.

As I mentioned before in the introduction section, the give problem of this project was raised by a family who would know more about living location with some key factors such as children friendly places and public transportation nearby. They do not care about other services in an area like pub, cafe, restaurant or tourist attractions. Thus, after checking all returned venue categories, I sort out some categories which meet the requirement of problem:

- Grocery Store, Supermarket, baby store

- Service for entertainment (especially for kids): Playground, Recreation Center, Library, Pool, Indoor play area
- Public transportation: Train Station, Metro Station, Bus Stop, Tram station

I filtered the set of data from Four Square by suitable venue categories, did some cleaning because some categories are the same but named differently. For example: baby store and kids' store; or grocery store, supermarket and market.



The dataset of all suitable venues contains 1093 records, I use this dataset to cluster and segment all areas in the capital region of Finland.

3. Methodology

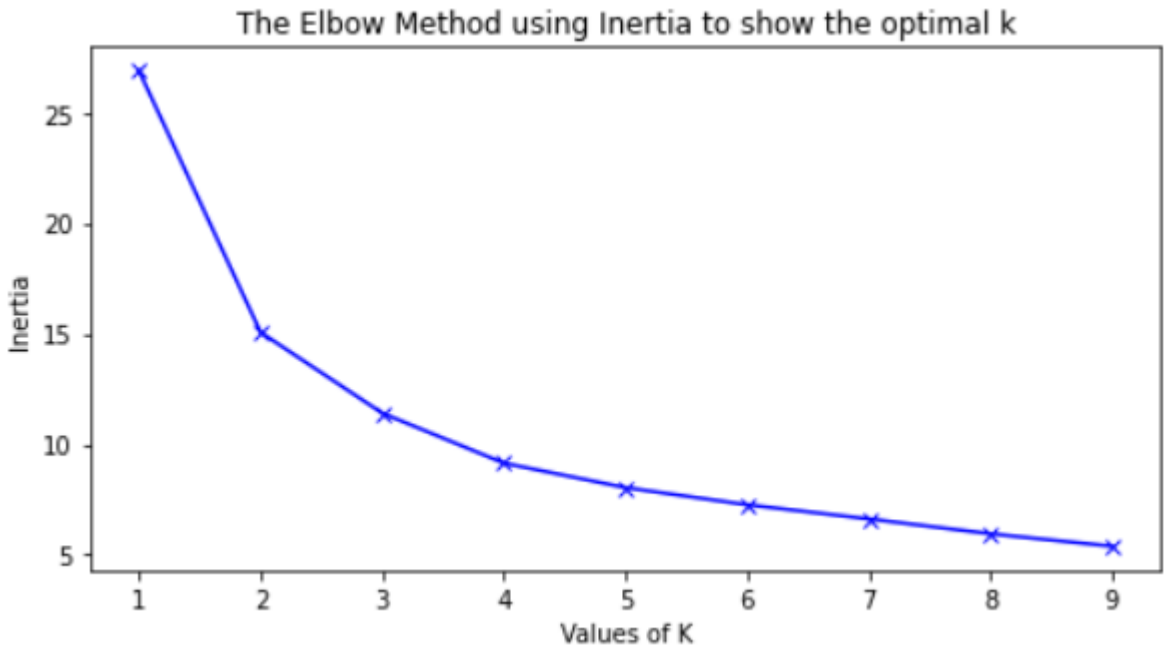
I used K-means algorithm for the clustering and segmentation of target areas into different clusters. Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data. The clustering factors are venue categories (public transportation, services for children). I will use Elbow method to find an optimal K for K-means algorithm

The results of clustering and segment are labelled friendly and later they are displayed on popup text of each area on Folium map

In term of housing sale price, I divided range of price into 5 categories (using function linspace() of numpy library): low price, medium price, high price, expensive price and very expensive price. Those labels of housing sale price are also displayed together with cluster's labels on the Folium map.

3.1.Analysing all areas of the capital region

With the dataset of venues in each area, I use one-hot coding to encode categorical features (venue categories) as a one-hot numeric array. After that I use Elbow method to find optimal K in unsupervised learning K- means algorithm.



From the result of Elbow method (see the graph above), I choose K = 4 to run K-means algorithm. Here is the result of K-means clustering algorithm for my dataset.

```
array([0, 0, 0, 1, 0, 0, 0, 3, 0, 0, 1, 3, 0, 2, 0, 0, 3, 1, 3, 1, 3, 3,
       0, 0, 2, 1, 3, 3, 0, 2, 1, 2, 0, 3, 2, 2, 0, 0, 3, 0, 0, 2, 0, 3,
       1, 2, 3, 0, 1, 3, 3, 3, 0, 1, 0, 0, 0, 1, 0, 1, 3, 3, 1, 0, 0, 0,
       1, 3, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 2, 3, 0, 0, 0, 0, 0, 1,
       2, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 3, 0, 2, 2, 0, 3, 0,
       1, 3, 1, 3, 3, 2, 0, 0, 1, 0, 3, 3, 1, 0, 0, 1, 1, 3, 0, 1, 0, 1,
       1, 0, 1, 2, 0, 0, 1, 3, 3, 1, 1, 1, 0, 1])
```

3.2.Adding labels for clusters and housing sale price

Adding friendly labels for clusters

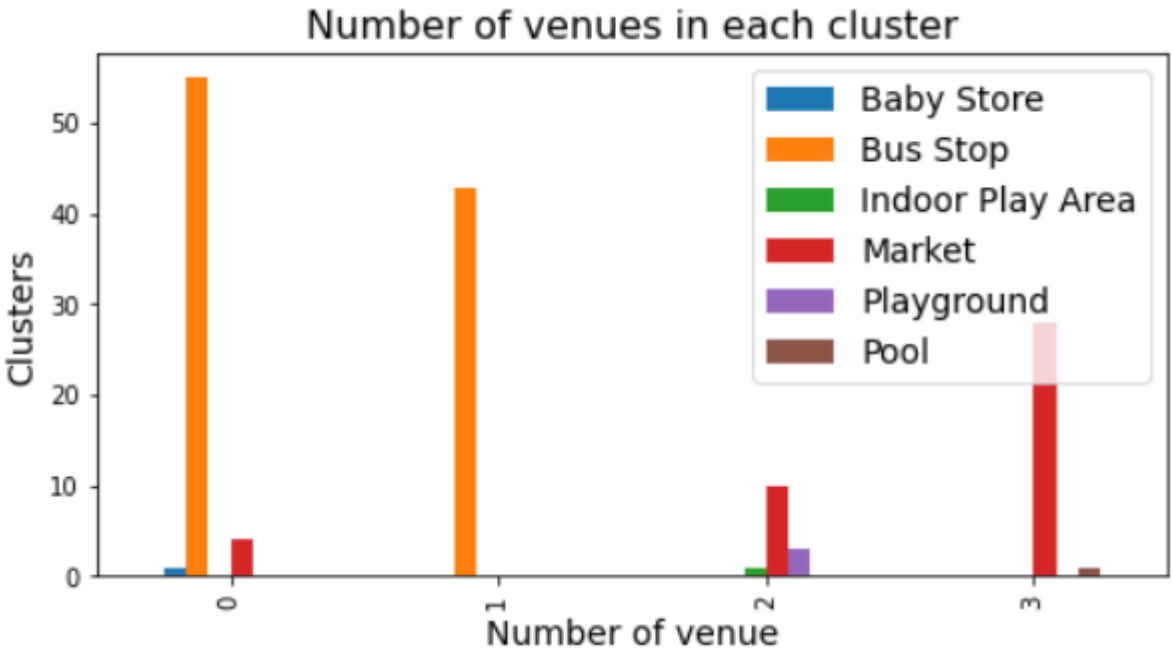
Firstly, I created the new data frame and added the top 5 venues for each area.

	postalcode	Area	municipality	latitude	longitude	price_per_m2	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	00100	Helsinki Keskusta - Etu-Töölö	Helsinki	60.1714	24.9316	7575	2	Playground	Market	Pool	Baby Store	Bus Stop
1	00120	Punavuori	Helsinki	60.1632	24.9391	8160	2	Market	Playground	Pool	Baby Store	Bus Stop
2	00130	Kaartinkaupunki	Helsinki	60.1645	24.9487	7825	3	Pool	Market	Baby Store	Bus Stop	Indoor Play Area
3	00140	Kaivopuisto - Ullanlinna	Helsinki	60.1578	24.9525	8713	2	Market	Playground	Baby Store	Bus Stop	Indoor Play Area
4	00160	Katajanokka	Helsinki	60.1671	24.9684	7208	3	Market	Pool	Recreation Center	Baby Store	Bus Stop

Secondly, I checked the most common venue in each cluster. I visualized the result of the most common venues in each cluster to get an idea of labeling for each cluster.

	Cluster Labels	1st Most Common Venue	Counts
0	0	Baby Store	1
1	0	Bus Stop	55
2	0	Market	4
3	1	Bus Stop	43
4	2	Indoor Play Area	1
5	2	Market	10
6	2	Playground	3
7	3	Market	28
8	3	Pool	1

The bar chart shows the most common venues in each cluster



From the bar chart above, I named 4 clusters as below:

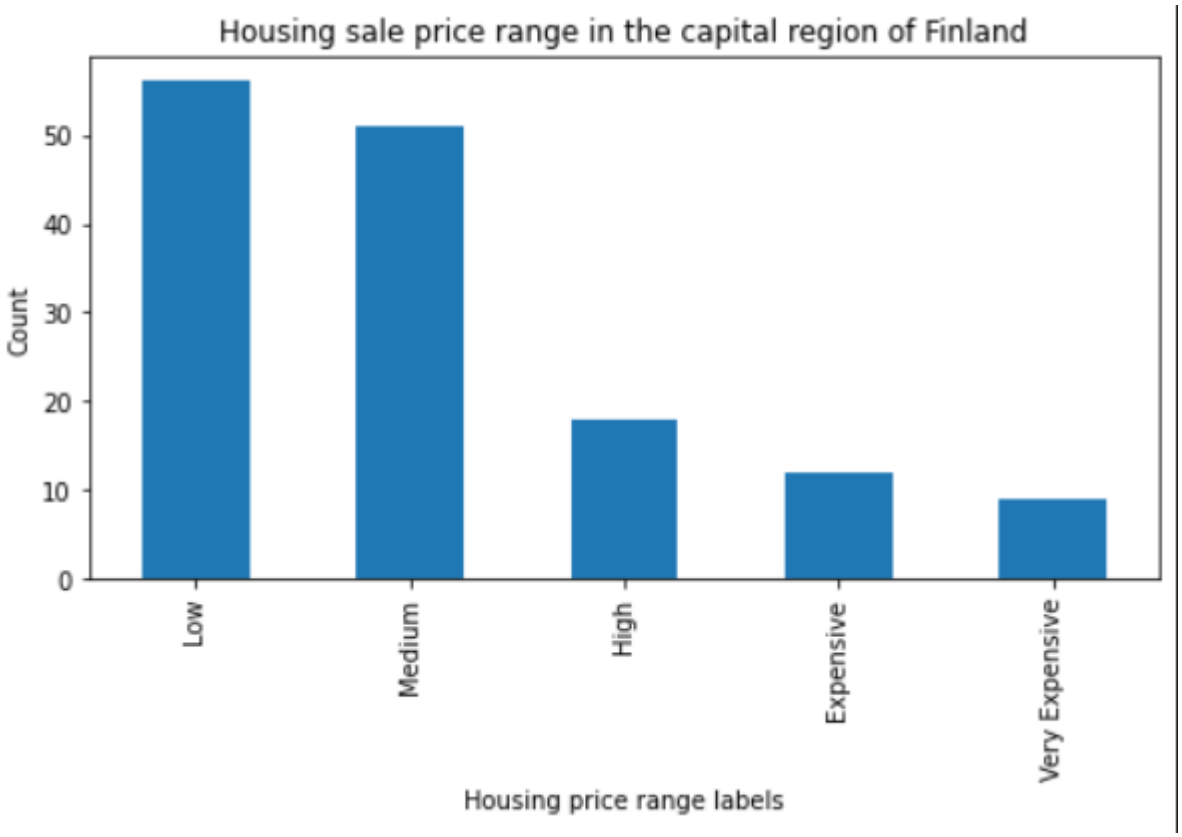
- Cluster 0: Easy access by bus, various markets and baby stores
- Cluster 1: Easy access by bus
- Cluster 2: Various markets & children friendly places
- Cluster 3: Various markets and pools

Finally, I added those friendly labels in the dataset

Adding labels for housing sale price

Pulling out the data of housing sale price from my dataset, I visualized it with histogram chart to clearly see the range of price

I defined housing sale price into 5 labels: 'Low', 'Medium', 'High', 'Expensive', 'Very Expensive'. So, I would like to have 5 bins with equal size, I use linspace() function of numpy

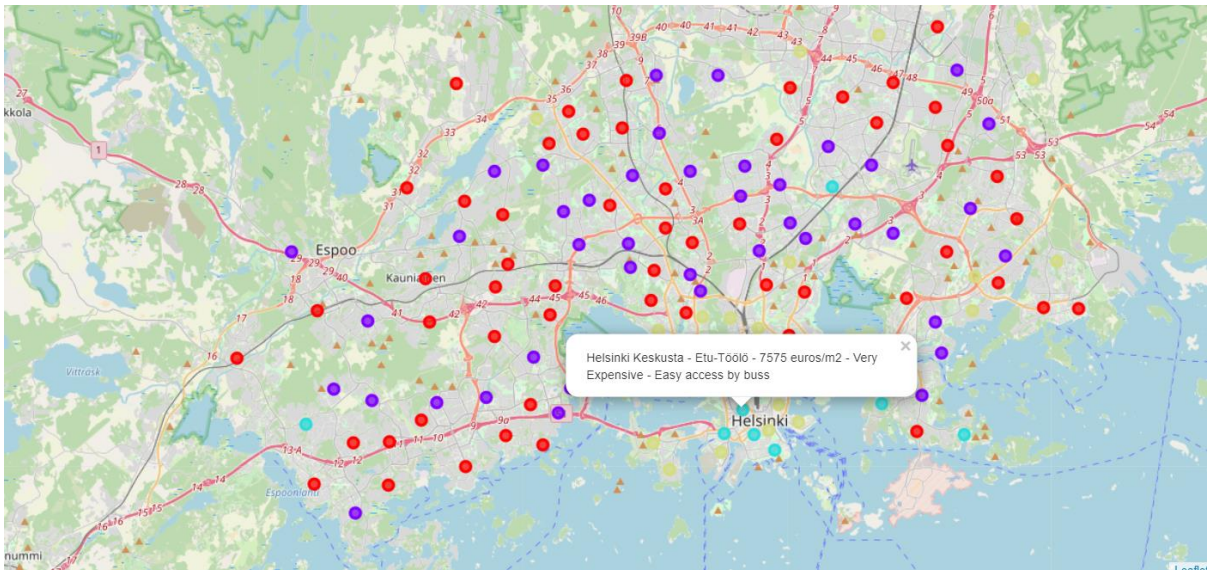


Finally, I also added those labels into my main dataset. Here is how it looks like

	postalcode	Area	municipality	latitude	longitude	price_per_m2	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	[illegible] Most Common Venue	[illegible] Most Common Venue	Friendly_cluster_labels	price_lables
0	00100	Helsinki Keskusta - Etu-Töölö	Helsinki	60.1714	24.9316	7575	2	Playground	Market	Pool	Baby Store	Bus Stop	Easy access by buss	Very Expensive
1	00120	Punavuori	Helsinki	60.1632	24.9391	8160	2	Market	Playground	Pool	Baby Store	Bus Stop	Easy access by buss	Very Expensive
2	00130	Kaartinkaupunki	Helsinki	60.1645	24.9487	7825	3	Pool	Market	Baby Store	Bus Stop	Indoor Play Area	Various markets & children friendly places	Very Expensive
3	00140	Kaivopuisto - Ullanlinna	Helsinki	60.1578	24.9525	8713	2	Market	Playground	Baby Store	Bus Stop	Indoor Play Area	Easy access by buss	Very Expensive
4	00160	Katajanokka	Helsinki	60.1671	24.9684	7208	3	Market	Pool	Recreation Center	Baby Store	Bus Stop	Various markets & children friendly places	Expensive

4. Result

The result of this project is a map of each target area with information of clustering and housing sale price. I used Folium to visualize.



5. Discussion

This project aims to support the decision of house buyers with some requirements. The problem was raised by a family who would know more about living location with some key factors such as children friendly places and public transportation nearby. They do not care about other services in an area like pub, cafe, restaurant or tourist attractions. Therefore, considering the requirements, I limited the categories of venue in dataset of every areas. The result of this analysis can be considered as an informative source to support the decision of a family in term of buying a house

As I mentioned before, the features of areas are limited based on requirements of buyer (children's friendly areas but near public transportation), in the future development, the features can be extended more by using all data of returned venue categories.

Limitation

- I planned to visualize the housing sale price with choropleth map to give a stunning and clear visualization of housing sale price. However, I cannot find the detail geometry data of the capital region. I tried many sources (Spatial Data Repository of NYU, Paikkatietoikkuna, Paavo postal code area statistics, etc.) however, all I got was geometry of municipalities and cannot drill deeper. In the future, I will continue to find deeper geometry data of the capital region in order to make this analysis more informative to users.
- Four Square data also has limitation of its own. After examining the returned data from Four Square, I found out that there is almost none of venue related to education available (only one Day-care returned). With a common knowledge of living in Finland, I know that in each area, there are several daycares, kindergartens and schools around.

6. Conclusion

Purchasing a house or apartment is always a huge decision and investment in a family. Purpose of this project is to provide more informative data on certain location of living for house buyers as a trustful reference when they decide to invest their financial resource on an accommodation.