Venues' relation to apartment prices in Helsinki, Finland

Introduction to problem

Housing investment is popular business amongst wide range of investors from small private investors to big capital funds. It is well understood that location has big impact on the apartment price. What is less clear, is the exact venue types that might affect to the apartment prices. In this project I will find what kind of venues there is to be found in higher valued living areas in Helsinki. This will be useful information for any real estate investor.

I will not make prediction about cause and effect relations: whether the venues themselves cause the value increase, or if the real cause is something else, like population density. I will simply look into the correlation between certain venue types and apartments' square meter price.

Introduction to Data usage

I have gotten the apartment prices for public sources supplies by city of Helsinki. This data comes with Postcode, apartment square meter price on that postcode and Borough name for that postcode. I combined this data with another source, that has all the streets in Finland, their location data, and their postcode. I calculated the average built area location of every postcode location of Helsinki.

Foursquare was used for getting the data about the venues. Amount of each venue type was grouped by postcodes and multiple linear regression model with backward elimination was applied.

Data Sources

Apartment prices: https://www.asuntojenhinnat.fi/myytyjen-asuntojen-tilastot/kunta/helsinki
Foursquare.com for venues data
Another csv file with all the streets, postcodes and location data of them.

Methodology

The data was gathered and combined from different sources. I first created one dataframe with postcodes and their habited center (based on average location of streets in that postcode area), combined with average square meter price of apartments in that area. I took the nearest venues of these postcode areas with foursquare. I combined everything to one dataframe.

When getting the venues data, I was able to look the general view of how many of each type of venue there is in Helsinki.

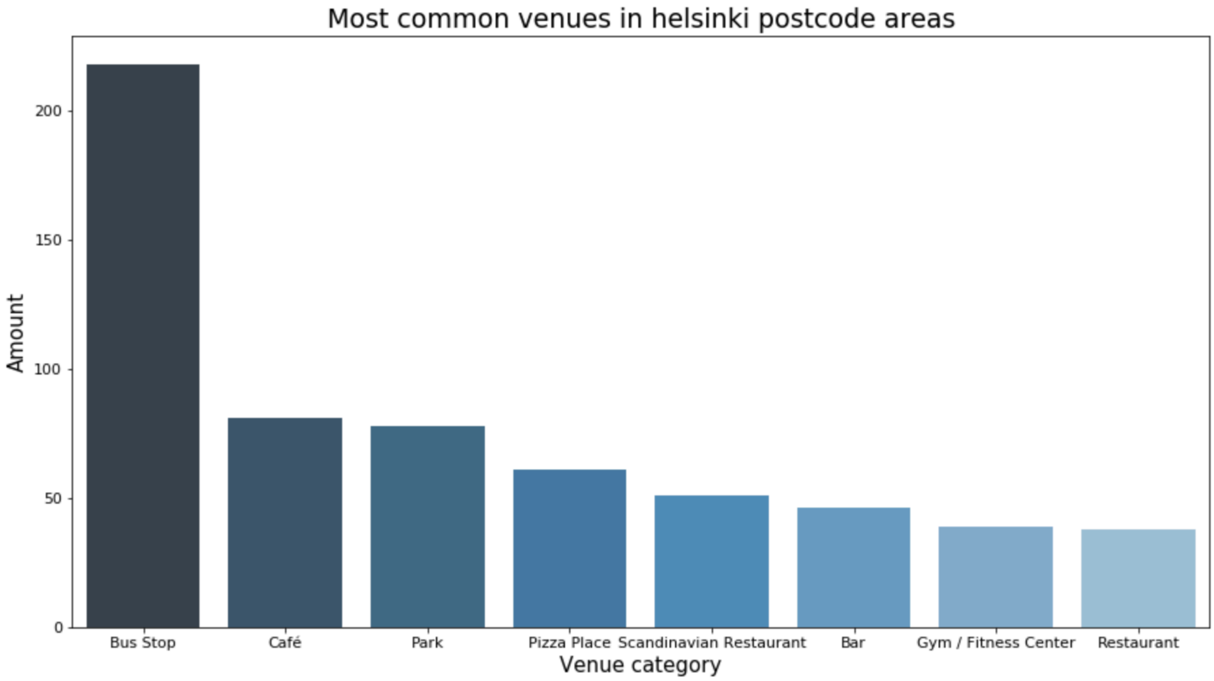
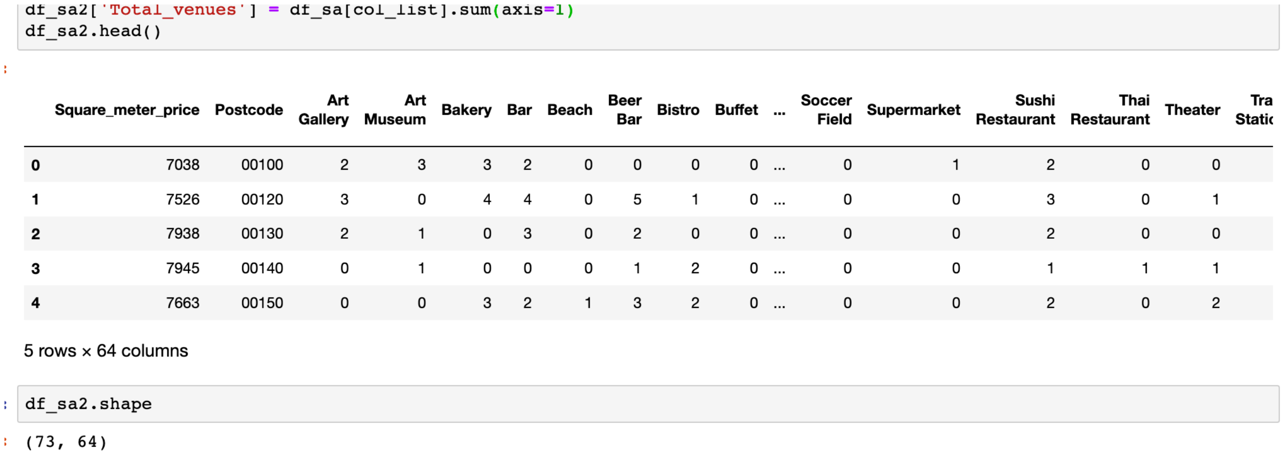


Figure 1. Most common venues

I chose multiple linear regression as the regression model for this project. Since I wanted to predict the square meter price based on multiple variables, multiple linear regression with backward elimination is just what the doctor ordered to do the job with precision and quality. This way I can remove the less relevant variables, that does not correlate with square meter price and will have the relevant variables (venue types) left.

Since some venues were very rare in Helsinki, I dropped the most uncommon ones away and was left with dataframe with 62 variables (venue types) and n=73 (postcodes).



```
df_sa2['Total_venues'] = df_sa[col_list].sum(axis=1)
df_sa2.head()
```

	Square_meter_price	Postcode	Art Gallery	Art Museum	Bakery	Bar	Beach	Beer Bar	Bistro	Buffet	...	Soccer Field	Supermarket	Sushi Restaurant	Thai Restaurant	Theater	Tra Static
0	7038	00100	2	3	3	2	0	0	0	0	...	0	1	2	0	0	
1	7526	00120	3	0	4	4	0	5	1	0	...	0	0	3	0	1	
2	7938	00130	2	1	0	3	0	2	0	0	...	0	0	2	0	0	
3	7945	00140	0	1	0	0	0	1	2	0	...	0	0	1	1	1	
4	7663	00150	0	0	3	2	1	3	2	0	...	0	0	2	0	2	

5 rows × 64 columns

```
df_sa2.shape
```

```
(73, 64)
```

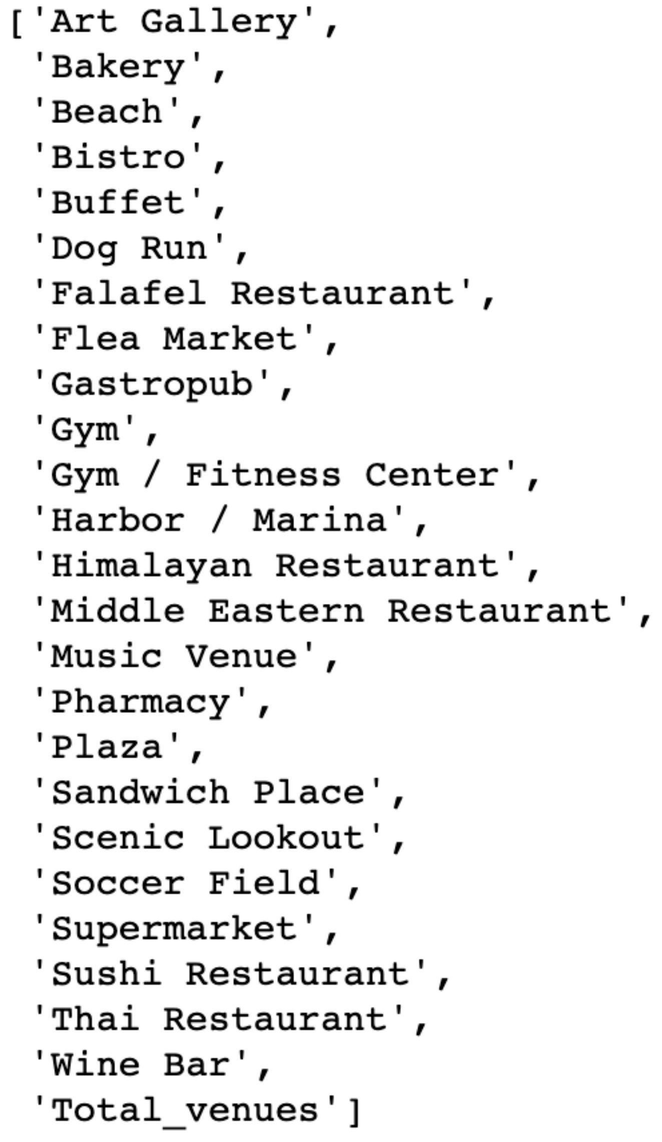
Figure 2. Final dataframe before machine learning

I divided the sample to training and test set and fitted it into the regression model. Backward elimination was done by choosing the variables with biggest P-value. When all the variables had P-value < 0.05, the elimination process was stopped.

Lastly I made a dataframe to show the accuracy of my regression model, and to find out whether or not the actual model could be used for estimating the exact price of an apartment.

Results

After backward elimination process I was left with 25 venue types showing a very strong correlation to apartment prices. They are the following:



```
['Art Gallery',
 'Bakery',
 'Beach',
 'Bistro',
 'Buffet',
 'Dog Run',
 'Falafel Restaurant',
 'Flea Market',
 'Gastropub',
 'Gym',
 'Gym / Fitness Center',
 'Harbor / Marina',
 'Himalayan Restaurant',
 'Middle Eastern Restaurant',
 'Music Venue',
 'Pharmacy',
 'Plaza',
 'Sandwich Place',
 'Scenic Lookout',
 'Soccer Field',
 'Supermarket',
 'Sushi Restaurant',
 'Thai Restaurant',
 'Wine Bar',
 'Total_venues']
```

Figure 3. The relevant venues

The last item, “Total_venues”, is the total number of venues in postcode area. When looking into this list and understanding the geography of Helsinki, some assumptions can be made. Apartment prices are most affected of the sea side places, activity venues and cultural locations. Public transportation does not seem to be an important factor, which might be because Helsinki is so well covered with public transportation. Therefore it does not matter, if you have metro, tram, bus or train station around you, since connections are always at least good.

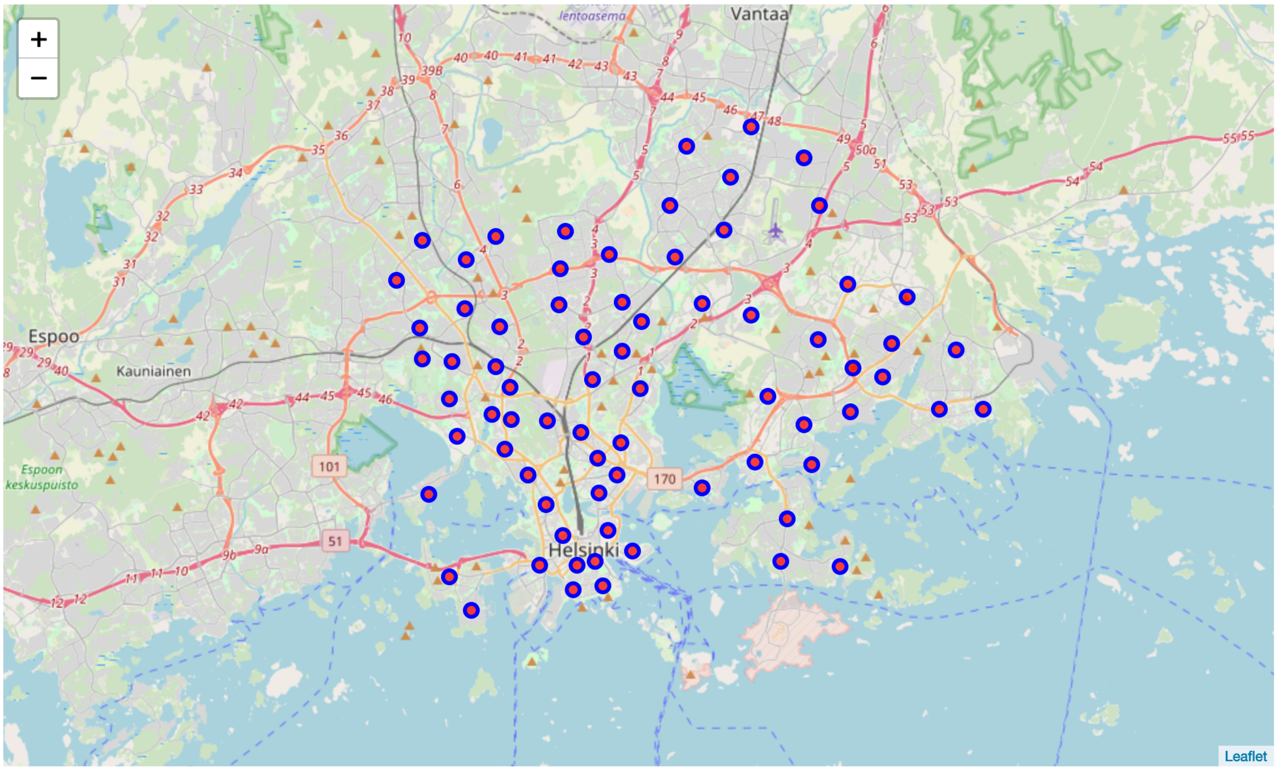
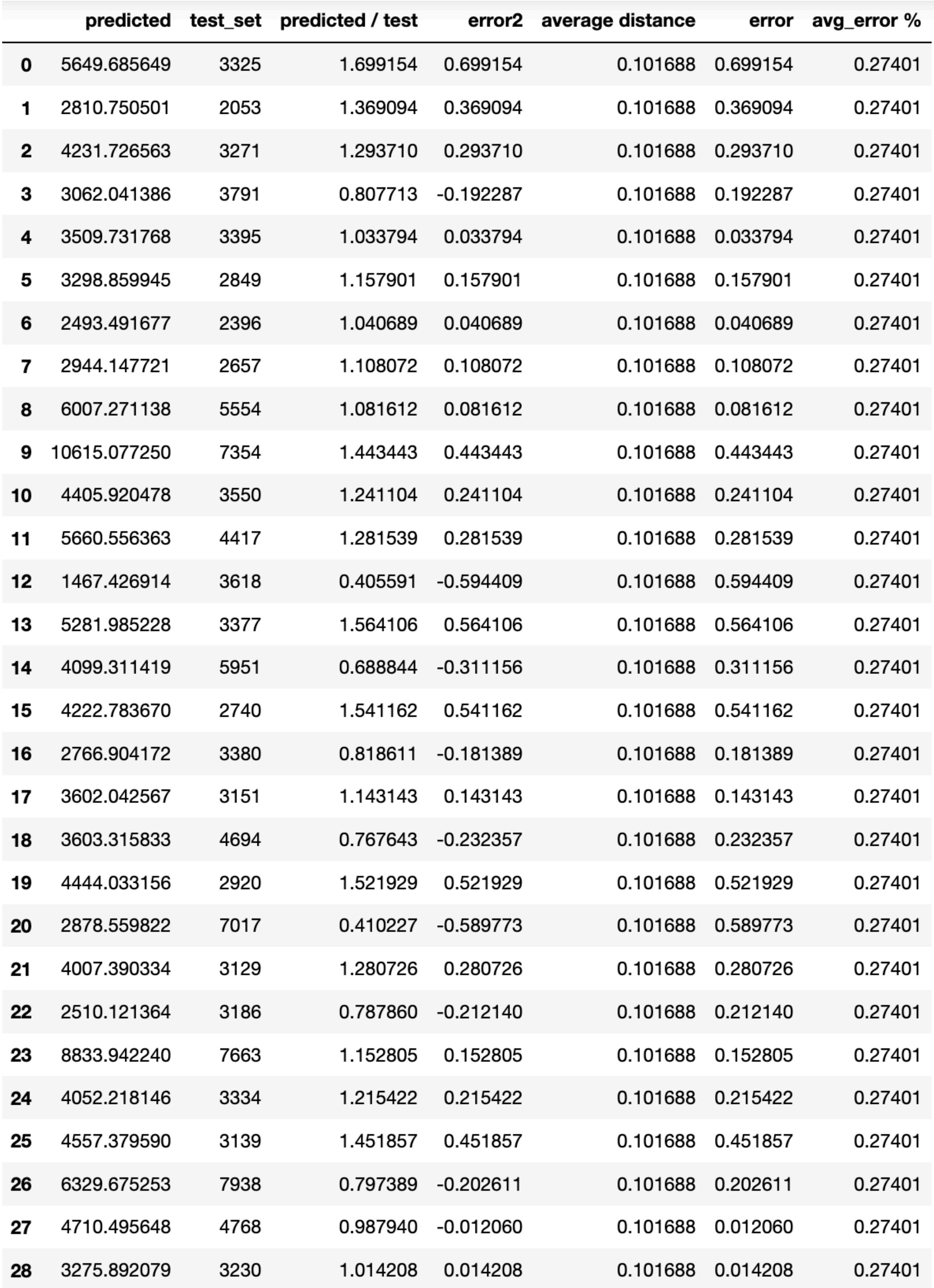


Figure 4. Postcodes on the map



	predicted	test_set	predicted / test	error2	average distance	error	avg_error %
0	5649.685649	3325	1.699154	0.699154	0.101688	0.699154	0.27401
1	2810.750501	2053	1.369094	0.369094	0.101688	0.369094	0.27401
2	4231.726563	3271	1.293710	0.293710	0.101688	0.293710	0.27401
3	3062.041386	3791	0.807713	-0.192287	0.101688	0.192287	0.27401
4	3509.731768	3395	1.033794	0.033794	0.101688	0.033794	0.27401
5	3298.859945	2849	1.157901	0.157901	0.101688	0.157901	0.27401
6	2493.491677	2396	1.040689	0.040689	0.101688	0.040689	0.27401
7	2944.147721	2657	1.108072	0.108072	0.101688	0.108072	0.27401
8	6007.271138	5554	1.081612	0.081612	0.101688	0.081612	0.27401
9	10615.077250	7354	1.443443	0.443443	0.101688	0.443443	0.27401
10	4405.920478	3550	1.241104	0.241104	0.101688	0.241104	0.27401
11	5660.556363	4417	1.281539	0.281539	0.101688	0.281539	0.27401
12	1467.426914	3618	0.405591	-0.594409	0.101688	0.594409	0.27401
13	5281.985228	3377	1.564106	0.564106	0.101688	0.564106	0.27401
14	4099.311419	5951	0.688844	-0.311156	0.101688	0.311156	0.27401
15	4222.783670	2740	1.541162	0.541162	0.101688	0.541162	0.27401
16	2766.904172	3380	0.818611	-0.181389	0.101688	0.181389	0.27401
17	3602.042567	3151	1.143143	0.143143	0.101688	0.143143	0.27401
18	3603.315833	4694	0.767643	-0.232357	0.101688	0.232357	0.27401
19	4444.033156	2920	1.521929	0.521929	0.101688	0.521929	0.27401
20	2878.559822	7017	0.410227	-0.589773	0.101688	0.589773	0.27401
21	4007.390334	3129	1.280726	0.280726	0.101688	0.280726	0.27401
22	2510.121364	3186	0.787860	-0.212140	0.101688	0.212140	0.27401
23	8833.942240	7663	1.152805	0.152805	0.101688	0.152805	0.27401
24	4052.218146	3334	1.215422	0.215422	0.101688	0.215422	0.27401
25	4557.379590	3139	1.451857	0.451857	0.101688	0.451857	0.27401
26	6329.675253	7938	0.797389	-0.202611	0.101688	0.202611	0.27401
27	4710.495648	4768	0.987940	-0.012060	0.101688	0.012060	0.27401
28	3275.892079	3230	1.014208	0.014208	0.101688	0.014208	0.27401

Figure 5. Model’s predictions on test set

In this table above we can see the regression model fitted to the test set. The first column shows the predicted square meter price and the second column shows the actual price. Third column tells y-hat size to y_test size, and error2 column tells the difference. What is most important value is the last column avg_error %. It tells that the average difference between y-hat and y_test. The average difference is 27%, which is very high for almost any kind of analytics situation, especially with real estate. Therefore this model cannot not be used as stand alone model to evaluate housing prices, but it could be potentially used as one tool in the toolbox, or to be merged with another machine learning model to improve accuracy.

Discussion

Like told in Results section, the final model is not exactly accurate, but it could be helpful and useful. It would be interesting to have the actual data from all postcode areas and not only what foursquare offers. Then it could be possible to make much better model with much higher accuracy.

If continuing with this project, it could be interesting to make detailed analysis with many regression models to see how precise prediction can be gotten. Also the aforementioned idea of listing all the venues from the postcode areas could be interesting, but extremely time consuming and not practical.

It would be very interesting to use this model with another machine learning model that is used for apartment price predictions and to see how venue data combined with the data from the apartment work out. Since this study showed clear correlation between venues and square meter prices, I believe this method could make significant difference some other prediction algorithm that does not use venue data yet.

Conclusion

The goal of this study was to find what venue types has most correlation to apartments' square meter price. Another goal was to build a machine learning model for estimating the square meter price. Both goals were matched and clear valuable information was gained. On this basis, I could say the project has been success.

Since this was rather small project, more accurate model could be built with more testing, i.e. having bigger sample with postcodes of near by cities and finding more accurate venue data than what only foursquare can offer.