

# Coursera Capstone Project Report:

## Predicting House Prices in Zagreb, Croatia

### Introduction

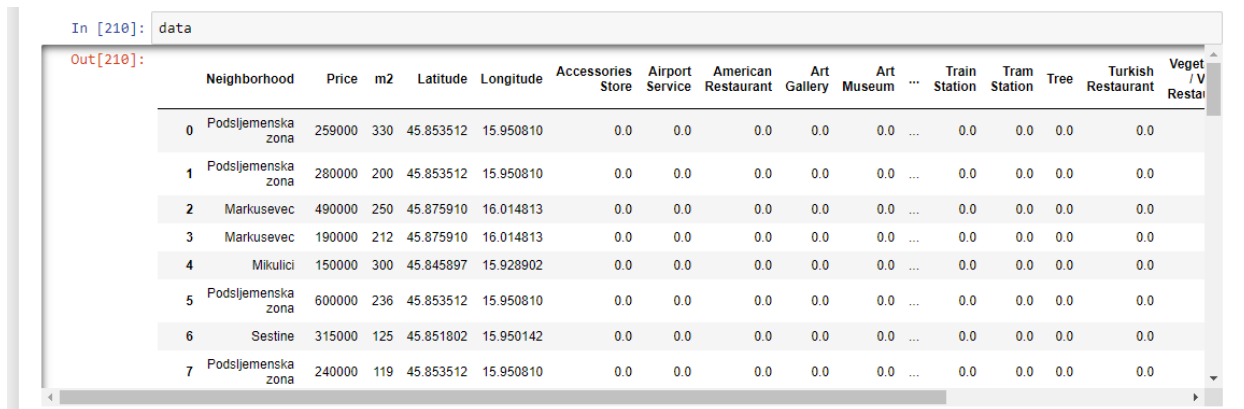
The problem to be investigated is to what extent do nearby venues determine house prices. The location chosen for this project is the city of Zagreb, Croatia. The people interested in solving this sort of problem would be as follows: people looking to buy houses in Zagreb, real-estate agents, city planners, investors etc.



Figure 1. Zagreb, Croatia

## The data

In this project the Foursquare location data is leveraged to explore and compare neighborhoods, and the house prices (as well their size in meters squared) is parsed from [www.gohome.hr](http://www.gohome.hr) website using BeautifulSoup4, html2text and urllib python libraries and modules. By manually creating and filling the corresponding .csv file with neighborhoods' latitudes and longitudes, and assigning those latitudes and longitudes to all the houses in the same neighborhood, the first part of the dataset is obtained. As already mentioned, using the Foursquare location data to each house a vector of the number of nearby venues is joined. That's how the whole dataset is obtained. The data will be used to find correlations between the selected features (venue categories and size of the house) and the house prices in EUR.



In [210]: data

Out[210]:

	Neighborhood	Price	m2	Latitude	Longitude	Accessories Store	Airport Service	American Restaurant	Art Gallery	Art Museum	...	Train Station	Tram Station	Tree	Turkish Restaurant	Veget / V Restai
0	Podsljemenska zona	259000	330	45.853512	15.950810	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
1	Podsljemenska zona	280000	200	45.853512	15.950810	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
2	Markusevec	490000	250	45.875910	16.014813	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
3	Markusevec	190000	212	45.875910	16.014813	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
4	Mikulici	150000	300	45.845897	15.928902	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
5	Podsljemenska zona	600000	236	45.853512	15.950810	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
6	Sestine	315000	125	45.851802	15.950142	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
7	Podsljemenska zona	240000	119	45.853512	15.950810	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	

Figure 2. The dataset

## Methodology

Using Nominatim module from geopy python library, Zagreb's geographical coordinates were extracted. Then, by utilizing folium python library's mapping capabilities, a visualization was created to better portray the various neighborhood distribution within the city of Zagreb, Croatia.

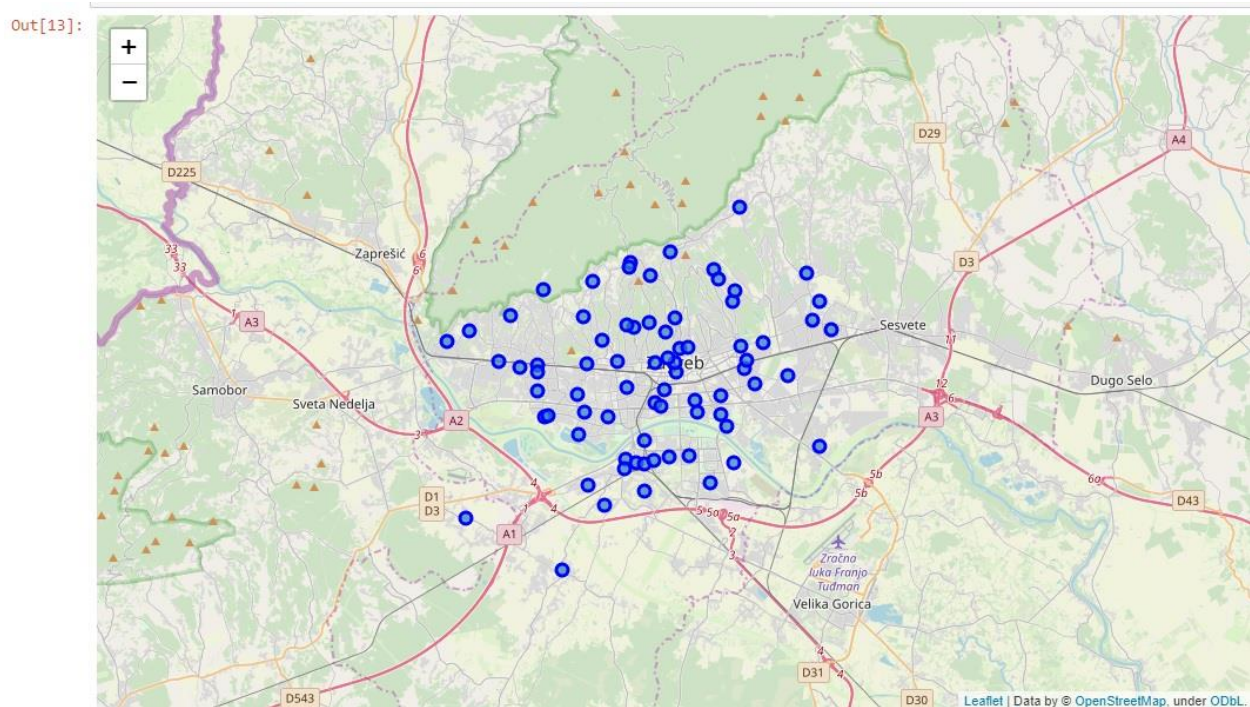


Figure 3. Neighborhood distribution in Zagreb, Croatia

Having the Foursquare developer account level credentials, nearby venues (1 kilometer) for neighborhood are pulled from their servers, subsequently aggregated by type so that they could be assigned to all the houses in the same neighborhood. Preliminary exploratory data analysis was performed where the top correlation values between venues categories and house price were compared and plotted in a bar chart. According to the data, these are the venue categories that affect house prices the most.

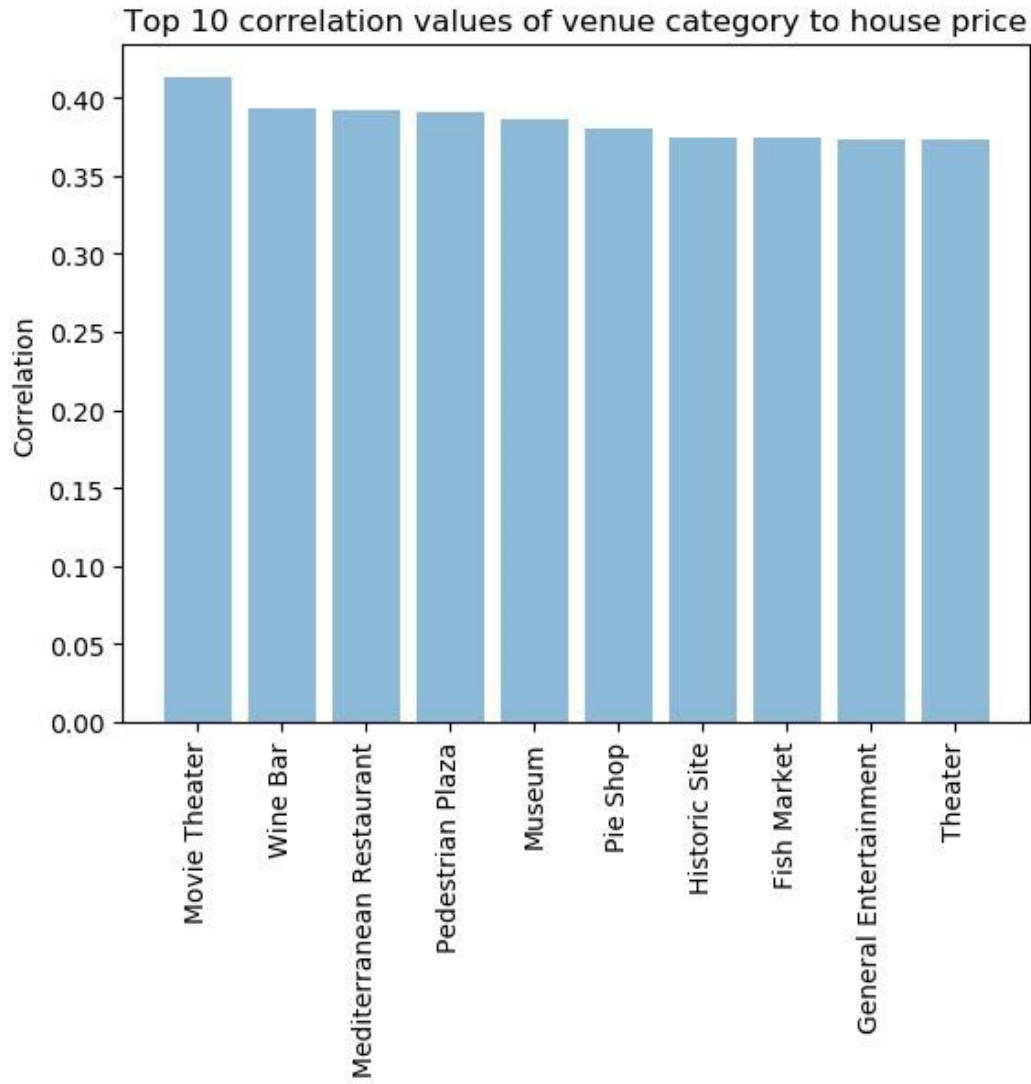


Figure 4. Top 10 correlation values of venue category to house price

Another analysis that was done was to identify which neighborhoods in Zagreb have the highest mean house price to  $m^2$  ratio and how far are these neighborhoods geographically distanced to the center of the city which is empirically determined the priciest of them all.

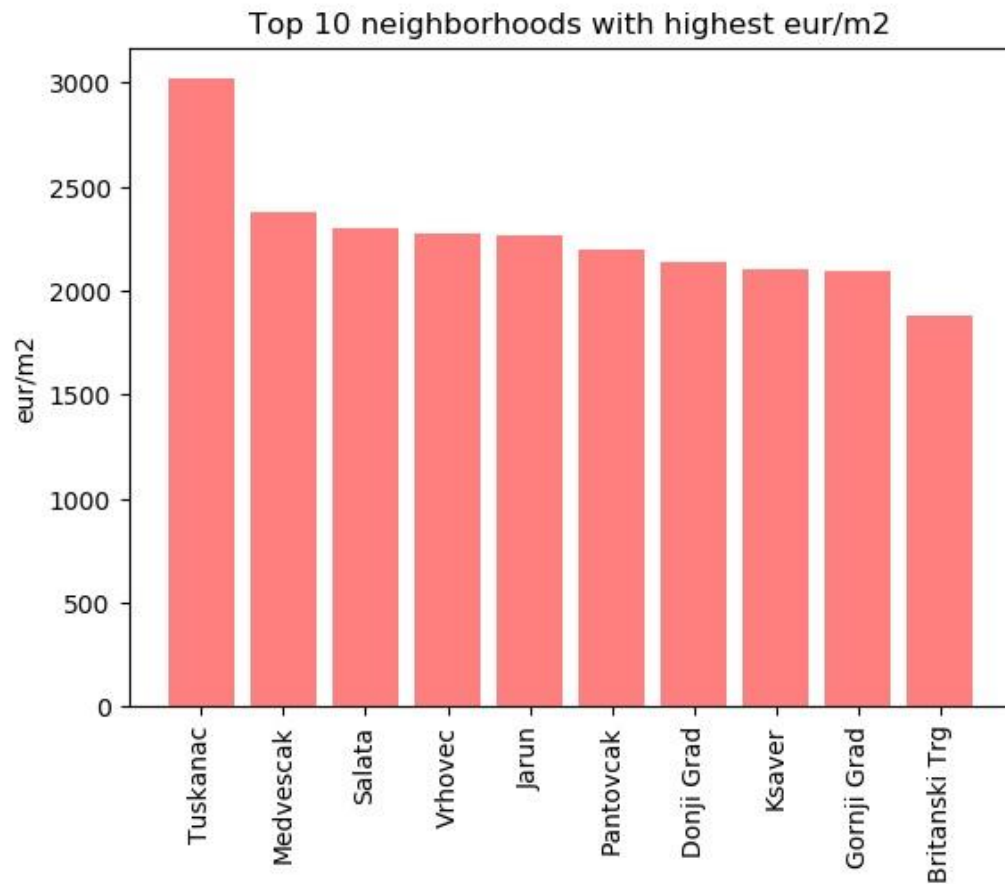


Figure 5. Top 10 neighborhoods with highest mean price to m<sup>2</sup> ratio

Of the machine learning algorithms, polynomial linear regression of degree 2 was used where the features included only the top 40 venue categories with the highest correlation to house price. Of other features, meters squared was also used. scikit-learn's (python library for machine learning) StandardScaler transformed the features into data with zero mean and variance of one. All the data was split into 80/20 training and test sets.

## Results and discussion

The lowest residual sum of squares that was achieved was 47278303818.68 and the highest variance score (fit) of 0.65. The best result was achieved by trying to combine the most optimal values of hyperparameters (number of venue categories, degree of polynomial features, type of scaling etc.). It is assumed that better results could be achieved by the following means: more house samples in the dataset, more venues in the Foursquare database for the city of Zagreb, Croatia, more computer memory for higher degree polynomial features, using a neural network to fit the data, and others.

## Conclusion

This Coursera capstone project allowed for practicing with the skills and the tools to use location data to explore a geographical location, as well as applying all the knowledge acquired during previous courses in the IBM Data Science Professional Certificate specialization. This course also enabled the student to deal with real-world data that is often times messy and incomplete. In conclusion, a big part of the data science profession is to also be able to deal with these kind of problems.