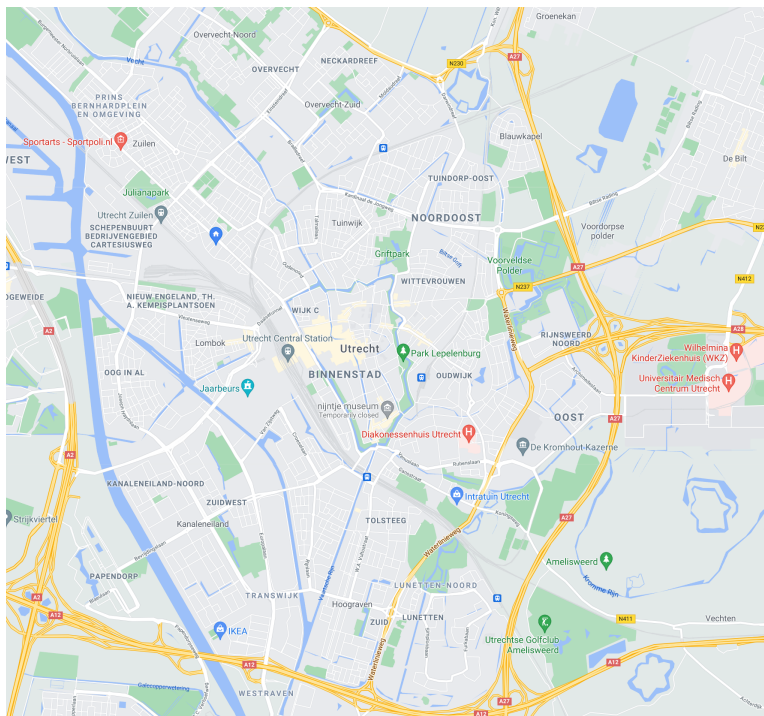


Introduction:

Housing prices in Utrecht, Netherlands are high, this causes trouble for many people. A lack of transparency in pricing only makes this worse. People need to have the right information available when making these purchase decisions. However, for many it is unclear why prices are much higher in some areas than in others. Should this affect your purchase decision?



The beautiful city of Utrecht, Netherlands

Breaking down the price of housing in Utrecht allows people to understand this phenomena and act accordingly. Because neighbourhood data is used in this analysis, it could help when looking for new and upcoming areas to live or invest in. In this assignment I will combine Utrecht housing price data, Utrecht neighbourhood data, and Foursquare API data to compare housing prices in neighbourhoods in Utrecht.

Apart from the benefits to the Utrecht community, this information can help the government understand various factors influencing housing pricing on a local level which can lead to improvements in housing policies to combat the rise of unaffordable housing.

Data:

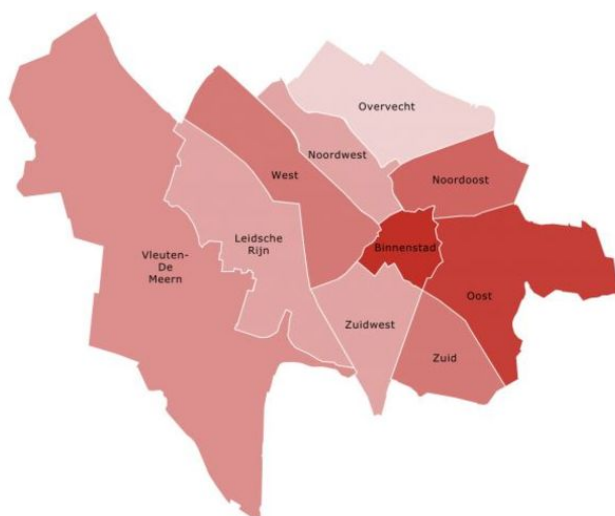
The data used in this analysis is from 3 sources. Utrecht neighbourhood data, Utrecht housing price data, and Foursquare API.

Utrecht neighbourhood data. The Dutch Central Bureau for Statistics (CBS) has a vast amount of data available on neighbourhoods. I have used the data set on neighbourhoods to find information on the number of inhabitants, children, addresses, neighbourhood size, and more. This data is available on various government levels. Initially, the idea was to use this data on 'Buurt' level (which is essentially a sub-neighbourhood), because utrecht has more than a 100 of those. However because of the lack of housing price data on such a small level of granularity, I had to make the choice to go with 'Wijk' level (large/parent neighbourhood). In the methodology section I will discuss the implications of that choice (significantly less data points).

	ID	WijkenEnBuurten	Gemeentenaam_1	SoortRegio_2	Codering_3	IndelingswijzigingWijkenEnBuurten_4	AantalInwoners_5	Mannen_6	Vrouwen_7	k_0T
0	14564	Wijk 01 West	Utrecht	Wijk	WK034401	1	29270	14210	15055	
1	14576	Wijk 02 Noordwest	Utrecht	Wijk	WK034402	1	45255	22060	23195	
2	14590	Wijk 03 Overvecht	Utrecht	Wijk	WK034403	1	34295	17200	17095	
3	14601	Wijk 04 Noordoost	Utrecht	Wijk	WK034404	1	39680	18780	20895	
4	14613	Wijk 05 Oost	Utrecht	Wijk	WK034405	1	32080	14905	17170	
5	14627	Wijk 06 Binnenstad	Utrecht	Wijk	WK034406	1	19165	9815	9345	
6	14639	Wijk 07 Zuid	Utrecht	Wijk	WK034407	1	27895	13685	14210	
7	14648	Wijk 08 Zuidwest	Utrecht	Wijk	WK034408	1	38620	19430	19190	
8	14656	Wijk 09 Leidsche Rijn	Utrecht	Wijk	WK034409	1	41290	20685	20605	
9	14672	Wijk 10 Vleuten-De Meern	Utrecht	Wijk	WK034410	1	49795	24465	25330	

10 rows × 118 columns

General Neighbourhood data



Location and shape of each neighbourhood

The next dataset is on housing prices in Utrecht's neighbourhoods. This data is from 2019, as the data from 2020 was not available at the time of this analysis. The data only includes transactions from 2019. The dataset was scraped from the web using the standard pandas scraping functionality (see notebook). This dataframe is merged with the already existing dataframe containing general neighbourhood data.

	Neighbourhood	Price_2019Q3	Price_per_m2
0	West	341.181	4.219
1	Noordwest	309.690	4.015
2	Overvecht	274.954	2.908
3	Noordoost	445.016	4.788
4	Oost	507.770	4.887
5	Binnenstad	434.747	5.091
6	Zuid	327.887	3.802
7	Zuidwest	313.324	3.661
8	Leidsche rijn	415.235	3.673
9	Vleuten-De Meern	428.002	3.562

House prices table in euros

The last dataset is coming from the Foursquare API. A central point was picked in each neighbourhood, and the coordinates of those points were added into the already existing dataset. Next, using the Foursquare API, all venues in a radius of 700m were identified. The 700 meter radius was chosen because 1) neighbourhood size is roughly 1.4km diameter, and 2) city centre reached the API return limit of 100 venues, so a bigger area would require multiple API calls. The returning values are transformed so that it represents the number of venues in that neighbourhood/area and appended to the existing dataset.

Latitude	Longitude	Venue
52.1091	5.0668	4
52.1108	5.0902	24
52.1180	5.1080	24
52.1035	5.1351	17
52.0840	5.1547	7
52.0889	5.1175	100
52.0641	5.1245	15
52.0758	5.1003	20
52.0954	5.0457	19
52.1002	5.0021	6

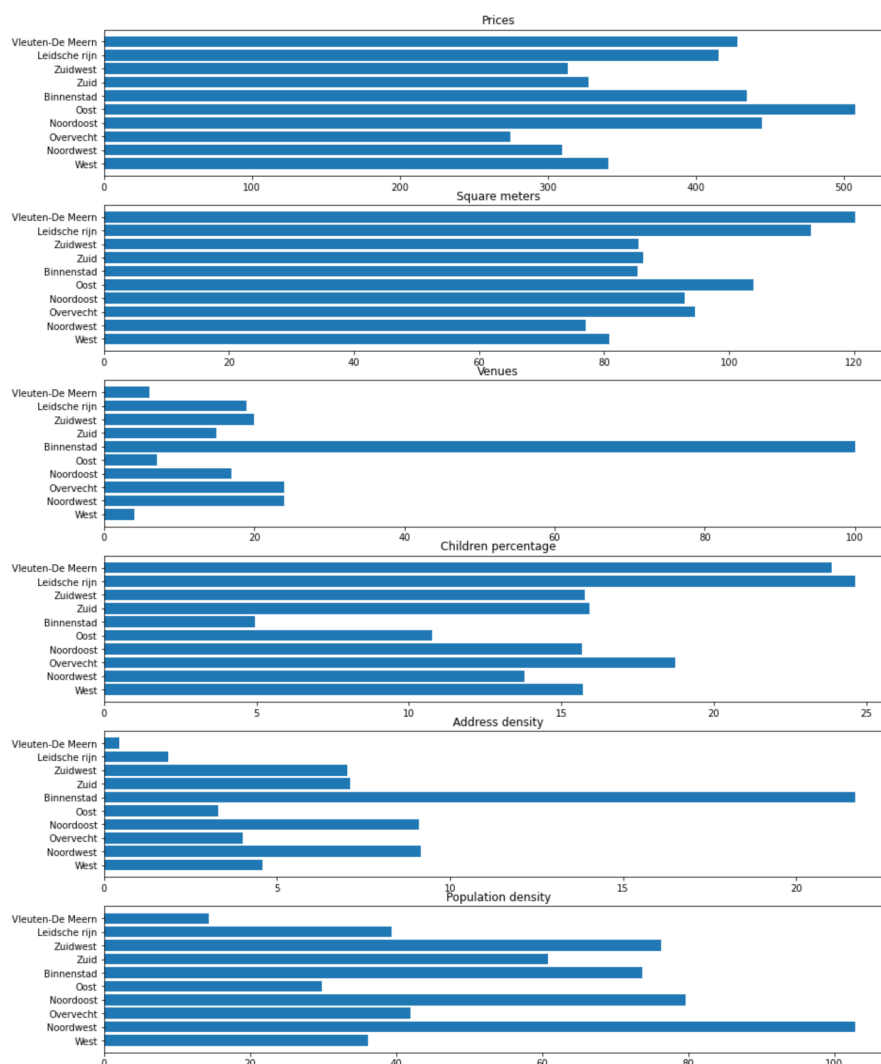
Neighbourhoods coordinates and number of Foursquare venues

Methodology

Using these datasets described in the Data section, I created variables that are believed to be more meaningful than the variables already in the dataset. Variables like Inhabitants and Addresses, we cannot use on face value because these are inherently linked to the landmass of a neighbourhood. To accommodate for that the variables were transformed to density measures, instead of absolute numbers. New variables:

Venues	= Count API results
Children_percentage	= Children / Inhabitants
Address_density	= Addresses / Landsurface
Population_density	= Inhabitants / Landsurface
Square_meters_2019	= Price / Price per square meter

Exploratory analysis consisted of comparing the variables above among neighbourhoods. The graph below shows some of the output:



The difference among neighbourhoods in the key variables.

Just from looking at these numbers, some similarities and differences can be discovered. For example, Vleuten-De Meern & Leidsche rijn are less populated, have more children, bigger surface area, medium price. This makes a lot of sense for neighbourhoods that are on the edge of Utrecht. To investigate this further, a cluster analysis will be discussed in the results section using K-means.

The main analysis on price, will be performed using multiple linear regression. This choice has been made because I do not just want to predict or estimate the average housing price, I also want to understand why the price is what it is. Many machine learning algorithms work as a black box, but this is not the case for regressions.

The variables are not standardized at all, therefore the choice has been made to normalize them using min-max feature scaling. This allows an apples to apples comparison because all variables are scaled to be in the 0 to 1 range.

Results

Clustering

The result of the k-means cluster analysis is 4 clusters:

	Neighbourhood	Cluster
0	West	1
1	Noordwest	1
2	Overvecht	1
3	Noordoost	0
4	Oost	0
5	Binnenstad	2
6	Zuid	1
7	Zuidwest	1
8	Leidsche rijn	3
9	Vleuten-De Meern	3

Result of cluster analysis

Based on the the data and the knowledge I have on Utrecht, this makes each cluster unique:

- Cluster 0: High prices, few children
- Cluster 1: Low prices, small surface area
- Cluster 2: High prices, few children, high address density, many venues
- Cluster 3: Large surface area, low address & population density, many children

Regression

Results:

	0	Coefs
0	Venue	-410.461378
1	children_percentage	-234.167489
2	address_density	474.366364
3	population_density	-59.148807
4	square_meters_2019	302.084812

Coefficient results of regression analysis

Because the variables have been normalized using min-max scaling prior to the analysis, we can compare them on face value. The results seem to suggest the amount of venues has a negative effect on housing prices, and address density a positive effect. Important to note is that these variables are negatively correlated with each other (collinearity), therefore the actual effect is smaller because they cancel each other out.

Discussion

So why are the results what they are? The density of children is negatively related to housing prices. This could have various reasons. Children are expensive, and so are houses, so people without children could perhaps afford more housing. Additionally, in general people in the city centre have less children, this is also where prices have risen the most in recent years. Population density is also slightly negatively correlated to price. This is a bit harder to explain, perhaps because population density is negatively correlated to square meters (surface area) of dwellings? This is definitely an area where more research is needed.

This brings me to the last point in this discussion. Obviously the sample size of this study is too small to make any conclusive statements. This is mainly due to data limitations and time limitations. However, this work could be used as a starting point, or a framework to expand on by including other cities, or digging up the data on a lower level of granularity in Utrecht. Sadly it seems that this data is only available through kadaster (paid).

Conclusion

The results show that there the neighbourhoods can be clustered in four distinct groups based on the variables used in this study. Additionally, children have a large negative effect on housing prices, surface area obviously has a big positive effect. Population density has a slight negative effect.

