The Battle of Neighborhoods – the case of Boston

1. Introduction

I choose the city of Boston. Here I want to discuss the issue of housing prices.

Particular I'm interested to understand what makes a house cheap or expensive. Also are there areas where houses more expansive/ cheap than elsewhere. This is an actual topic as rents are rising everywhere and the costs of living are increasing.

But besides external economic factors like increasing building costs, inflation there are other reasons as well which are often neglected. There are obvious reasons how much a house costs like the size of property or the equipment. But besides that what are factors which are also important? Factors that are related to the neighborhood?

2. Dataset

By looking for data to answer these questions I found a dataset within the scit-learn libary. It contains around 500 house prices for Boston in the 1970s. This is also the biggest limitation of the data as this is already old.

On the other side the dataset contains interesting variables that can give further insights to what external factors can have an influence on the house price.

The variables are:

CRIM:	Per capita crime rate by town
ZN:	Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS:	Proportion of non-retail business acres per town
CHAS:	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX:	Nitric oxide concentration (parts per 10 million)
RM:	Average number of rooms per dwelling
AGE:	Proportion of owner-occupied units built prior to 1940
DIS:	Weighted distances to five Boston employment centers
RAD:	Index of accessibility to radial highways
PTRATIO:	Pupil-Teacher-Ratio by town
TAX:	Full-value property tax rate per $10,000
B: $1000(Bk - 0.63)^2$,	where Bk is the proportion of [people of African American descent] by town
LSTAT:	Percentage of lower status of the population
MDEV:	median value of the house in 10000 $

3. Approaches

To learn more about housing prices I chose two approaches.

First I will try to buildt a model to predict the price based on the given features. Therefore I will use a multiple linear regression. This will include a check for missing values that need to be excluded, a graphical analysis to check if every feature is useful for a linear regression and the equation of that model. Afterwards the model needs to be verified based on data which was not used to create the model. This model will than enable users to predict housing prices.

The second part will deal with a cluster analysis. It is the aim of this area to understand if certain attributes are clustered around certain neighborhoods. For this it is necessary to check again for all variables and do not exclude variables which were not used for the linear regression.

Also to determine the amount of clusters the elbow-criteria via the sum of squared distances will be used. In the end the clusters will be described and named.

4. Exploratory Analysis

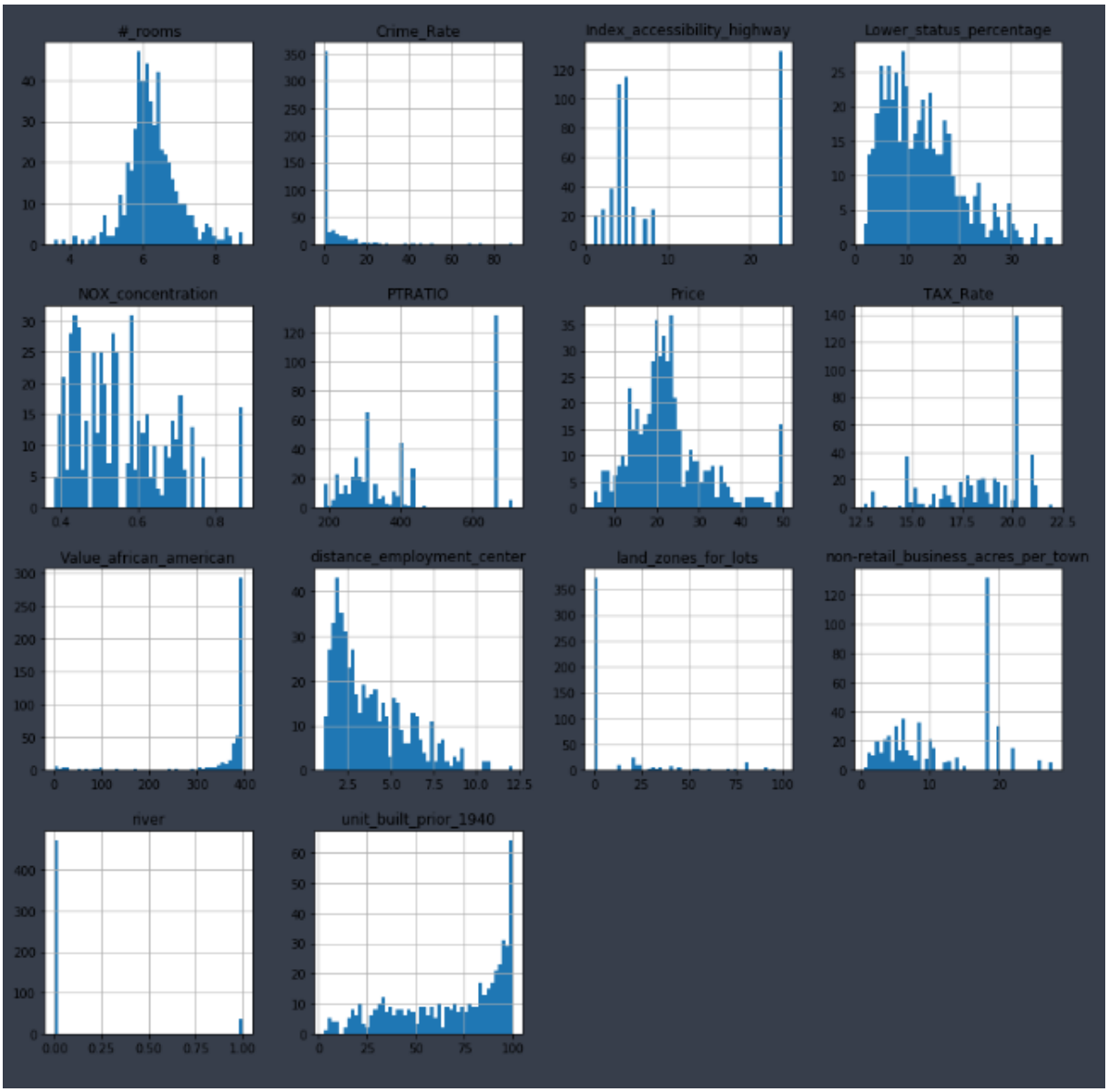
First step of the exploratory analysis was the creation of the dataset itself via import from sklearn. After features and targets where connected into one dataframe a look at the data itself proofed very valuable. For instance the variable B (Black people in town) was changed via a formula $(100*(Bk-0.63)^2$. This makes it impossible to derive the original value. Also a quick look at the formula itself makes it clear that the value is non-linear (power of 2). Therefore the changed value diguises the real percentage and has to be used with care in the further analysis.

The rest of the variables are pretty straightforward.

Next I checked for missing values in each of these variables. There were non so all entries could be used for further investigation.

As a last step an overview of all variables via describe and via graphical analysis was conducted.

This again shows the non-linear character of african-american variable and it also shows that the mean for housing prices (Price) seems to be stopped at 50 (in 10000 $). This can be deducted by the increase of homes with an exact price of 50.

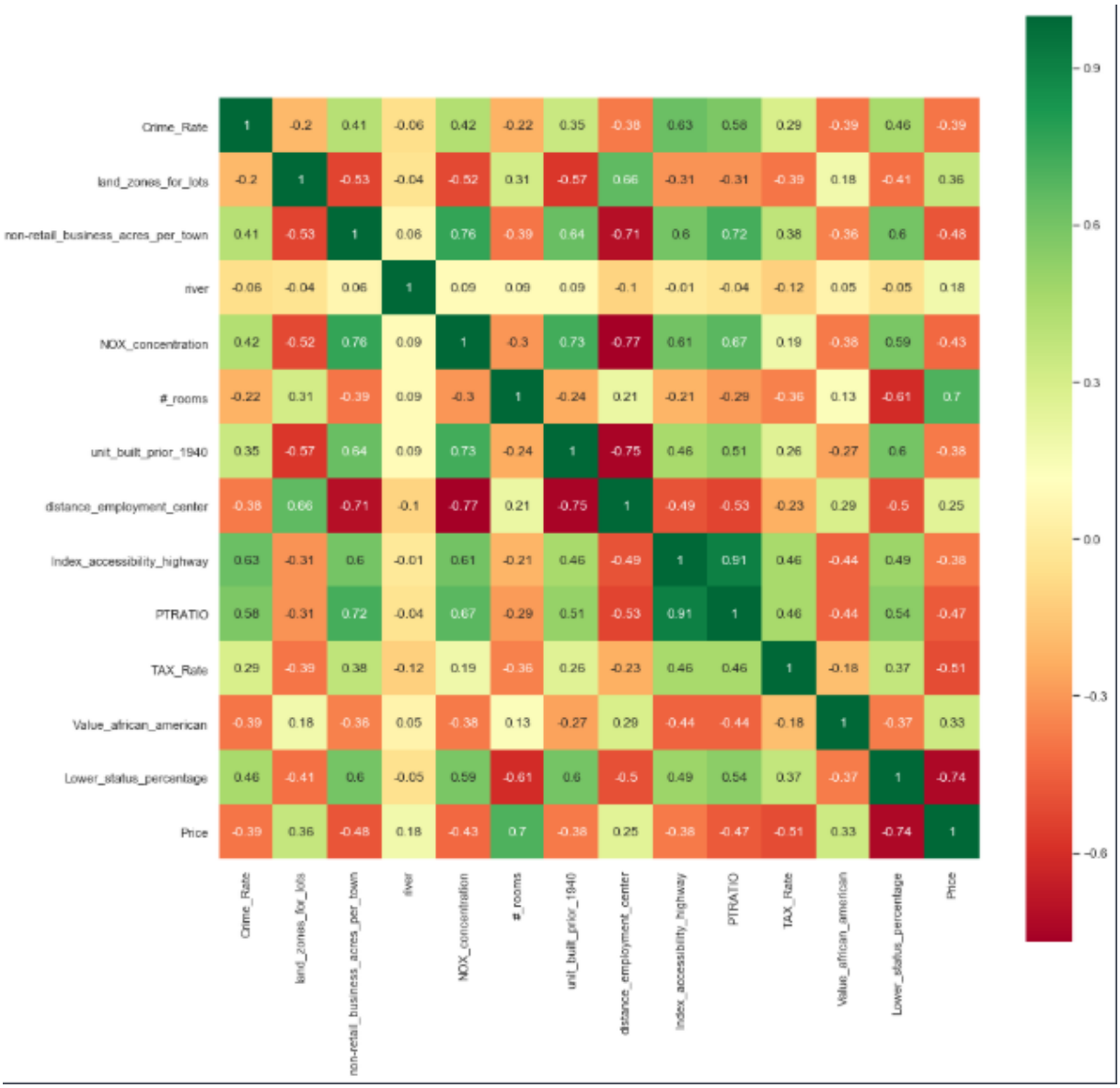


5. Results

5.1 Linear Regression

For the linear regression the first step was to check which variables have a high correlation with price as the target variable. Only these variables should be taken into consideration.

Therefore a correlation-heatmap was created.



As a dark red or a dark green implies high correlation only these values were inspected.

The following variables will be used in the regression model:

- lower status percentage (correlation with price: -0,74): poorer people means cheaper homes
- tax rate (- 0,51): rich people with expensive houses can settle at places with low taxes
- *#rooms (0,7=: more rooms means bigger house means higher prices*
- *non-retail*business_acres_per_town: (-0,48): no retail probably means industry and no one wants to live there
- PT-Ratio (-0,47): expansive houses are in neighborhoods with good schools which have a low PT Ratio

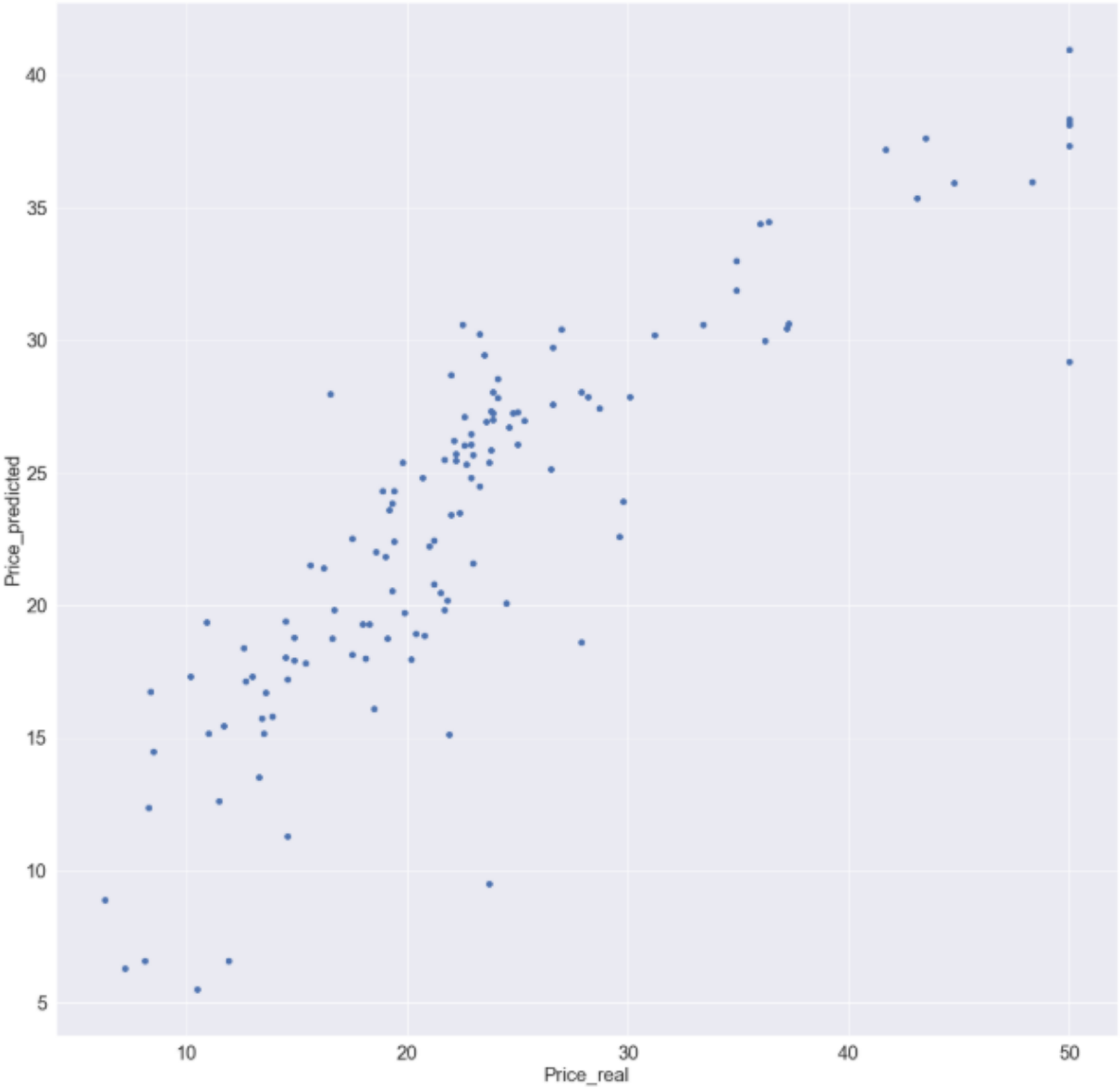
With a correlation of -0,38 the index_accessibility_highway correlates also with the price. But it also has a very high correlation with PRRatio (0,91) To avoid the danger of multicollinearity meaning two features measuring the same thing we are only using the PR-Ratio.

As now the interesting features are chosen, the actual model can be built.

Here I used machine-learning and trained a model with a 75%/25% train/ test split.

To evaluate how good the existing model is I did a graphical and statistical validation.

For the graphical validation I printed the real prices vs. the predicted prices. A perfect model would create a straight line.



As the graph shows, the the line is visible but the model is far from perfect. A statistical evaluation can give a more accurate answer how valid the model is.

For this I used R^2 which gives us the percentage of variation between predicted and real values which can be explained by our model. In our case this is 72%. This means only 28% of the variation is caused not by the model but by external factors.

The equation to predict the price is:

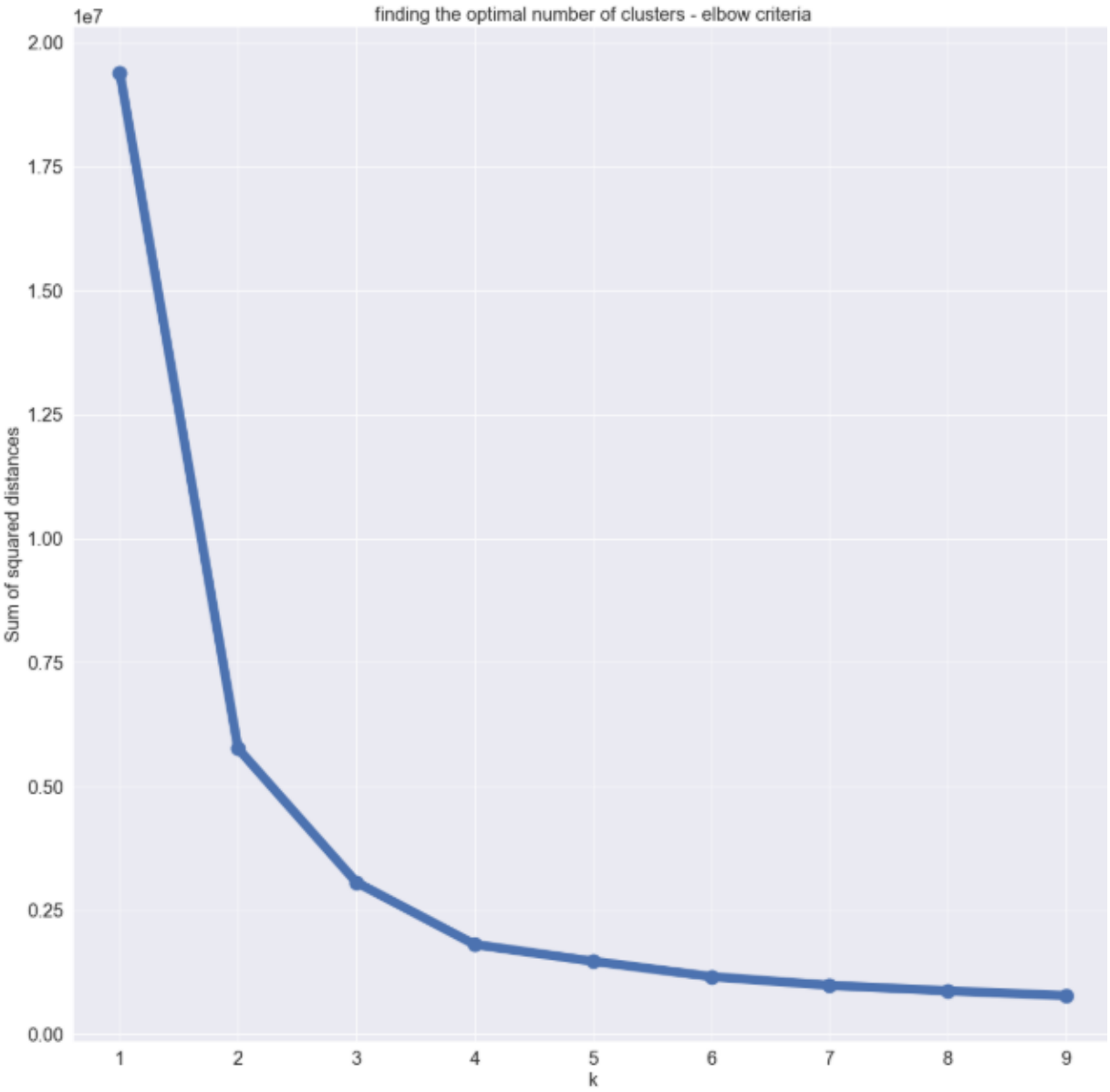
Price = 24,4 -0.59380*'Lower_status_percentage'* *-0.91042*'TAX_Rate' + 3.655074*'#_rooms'* *+0.027139*'non-retail_business_acres_per_town' -0.001779*'PTRATIO'

5.2 Cluster Analysis

Another way to approach the data and to check how the house price is influenced is via a cluster analysis. Herefore the correct amount of clusters needs to be calculated. We use the elbow-criteria

for this. As a cluster means that datapoints within that cluster are similiar to each other the distance is used as a perimeter to calculate the similarity. To avoid a bigger influence of larger numbers we standardize all variables with the standard scaler.

Then we use the sum of the squared distances for clusters between 1 – 10 to be able to print a graph which shows the elbow criteria.



As the elbow is starting at a clusteramount of 3 we therefore go with 3 clusters.

As mentioned before we use all variables in the cluster analysis because the limitations of the multiple linear regression do not nessecarily als apply for the cluster analysis

The result shows us three clusters:

Cluster	count	Percentage
0	366	72
1	102	20
2	38	8

Here is an overview of all clusters by the mean of every variable.

	Cluster 1	Cluster 2	Cluster 3
Clusterlabels	0.000000	1.000000	2.000000
Crime_Rate	0.374993	10.910511	15.219038
land_zones_for_lots	15.710383	0.000000	0.000000
non-retail_business_acres_per_town	8.359536	18.572549	17.926842
river	0.071038	0.078431	0.026316
NOX_concentration	0.509863	0.671225	0.673711
#_rooms	6.391653	5.982265	6.065500
unit_built_prior_1940	60.413388	89.913725	89.905263
distance_employment_center	4.460745	2.077164	1.994429
Index_accessibility_highway	4.450820	23.019608	22.500000
PTRATIO	311.232240	668.205882	644.736842
TAX_Rate	17.817760	20.195098	19.928947
Value_african_american	383.489809	371.803039	57.786316
Lower_status_percentage	10.388661	17.874020	20.448684
Price	24.931694	17.429412	13.126316

The three clusters can be described as:

- Cluster 0: rich suburban people because of
    - lowest crime rate,
    - most expensive houses,
    - most teacher for pupils,
    - far away from
      employment centres,
    - only a few poor people,
    - cleanest evironment,
    - no industry,
    - lowest taxes,
      lowest share of old buildings,
    - biggest houses
- Cluster 1: urban middle class people, because of
    - significant crime rate,
    - expensive houses,
    - not that many teacher fors pupils,
    - close to employment centres,
    - some poor people,
    - pollution,
    - some industry,
    - highest taxes,
      mainly old buildings
- Cluster 2: urban poor people, because of
    - highest crime rate,
    - cheap houses,
    - not that many teacher fors pupils,
    - close to employment centres,
    - poor people,
    - pollution,
    - a lot of industry,

- high taxes,
- mainly old buildings

6. Conclusion

The previous analysis has shown two ways on how to approach a dataset. First with a machine learning approach (multiple linear regression) then a cluster analysis.

The linear regression gives us an easy way to create a model to describe – based on core features – how much a house will cost on average.

It shows us the importance of the house size to predict the price as an internal factor. But it also shows us the imporance of external factors (percentage of lower status people, taxes, schools).

The cluster analysis teaches us how similiar houses can be grouped based on different features.

Nevertheless shows the clusteranalysis the typical divide of a city and confirms therefore what was to be expected. As the data is from the 1970‘ s it would be interesting to see where these different groups are now. E.g. are the most expensive houses still far away from the centre? Or did the go back towards the centre? Or did they go back to the centre and than moved again away?

So both techniques are ways to describe the data. While the regression gives us a model for the individuell datapoint, the cluster analysis helps us to compress the huge amount of information into small easy to understand chuncks.

Next step could be to do a regression analysis based on the cluster analysis so that in the end we have three regressions, one for each cluster. This could help to get a more exact modell for each cluster.