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.