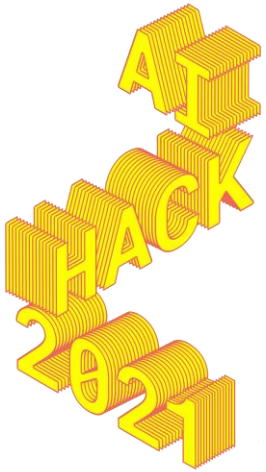
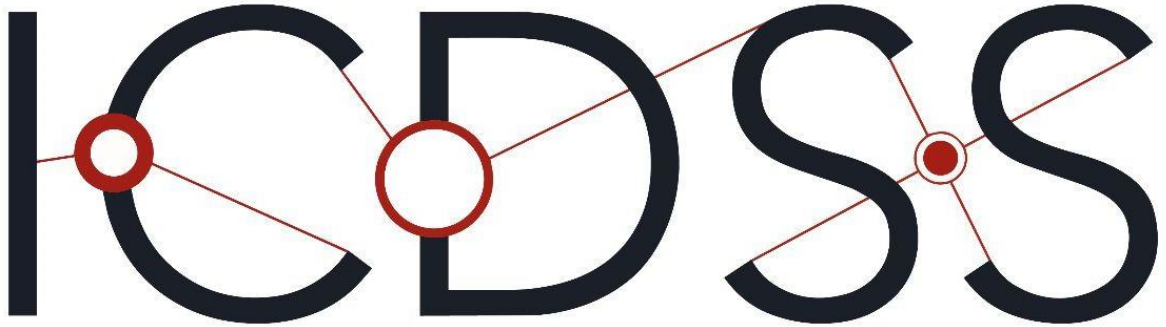


Imperial College London

Boston Housing Challenge

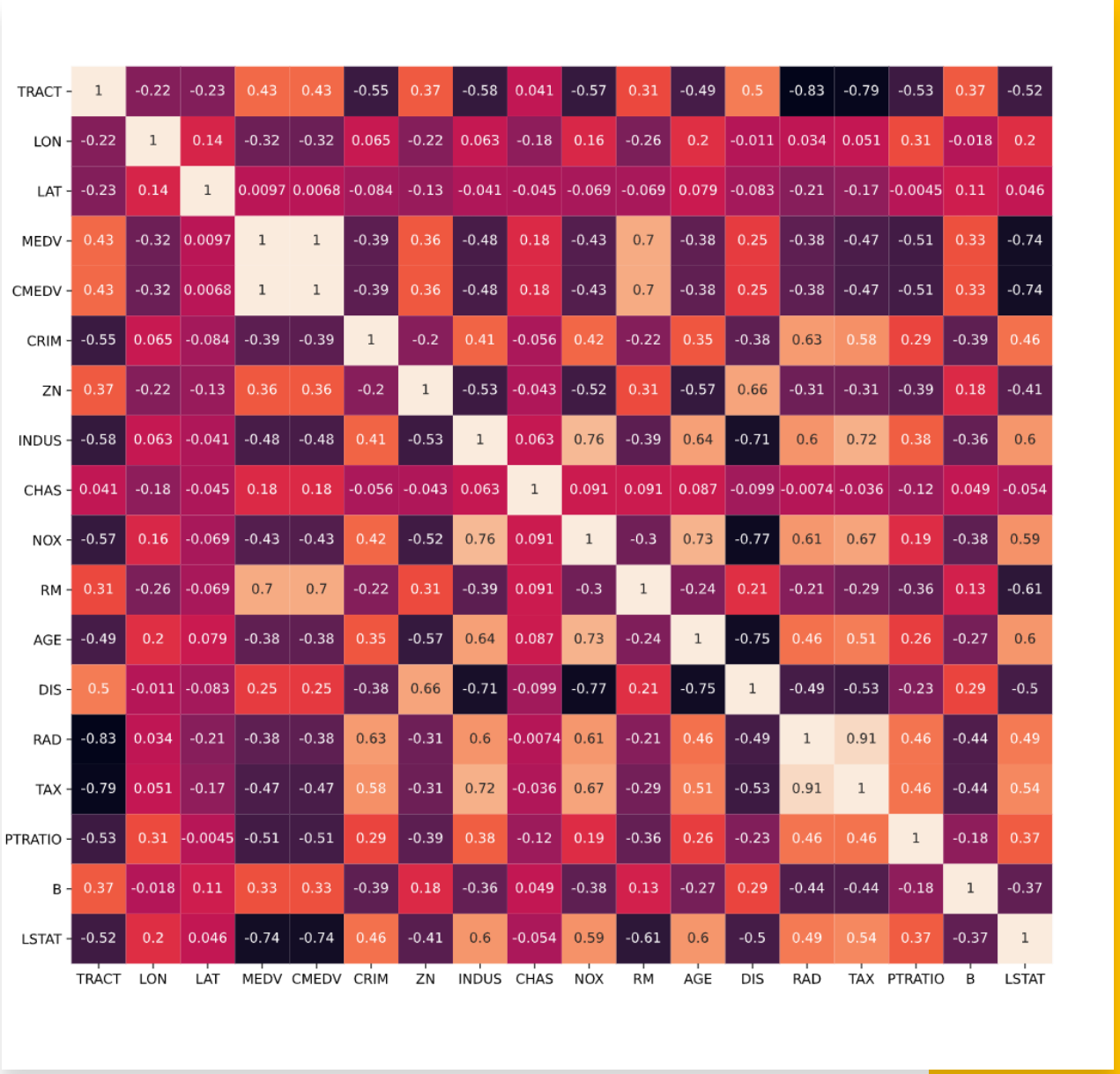


Team Name: The Outliers

gh: https://github.com/SamtheSaint/AIHACK21

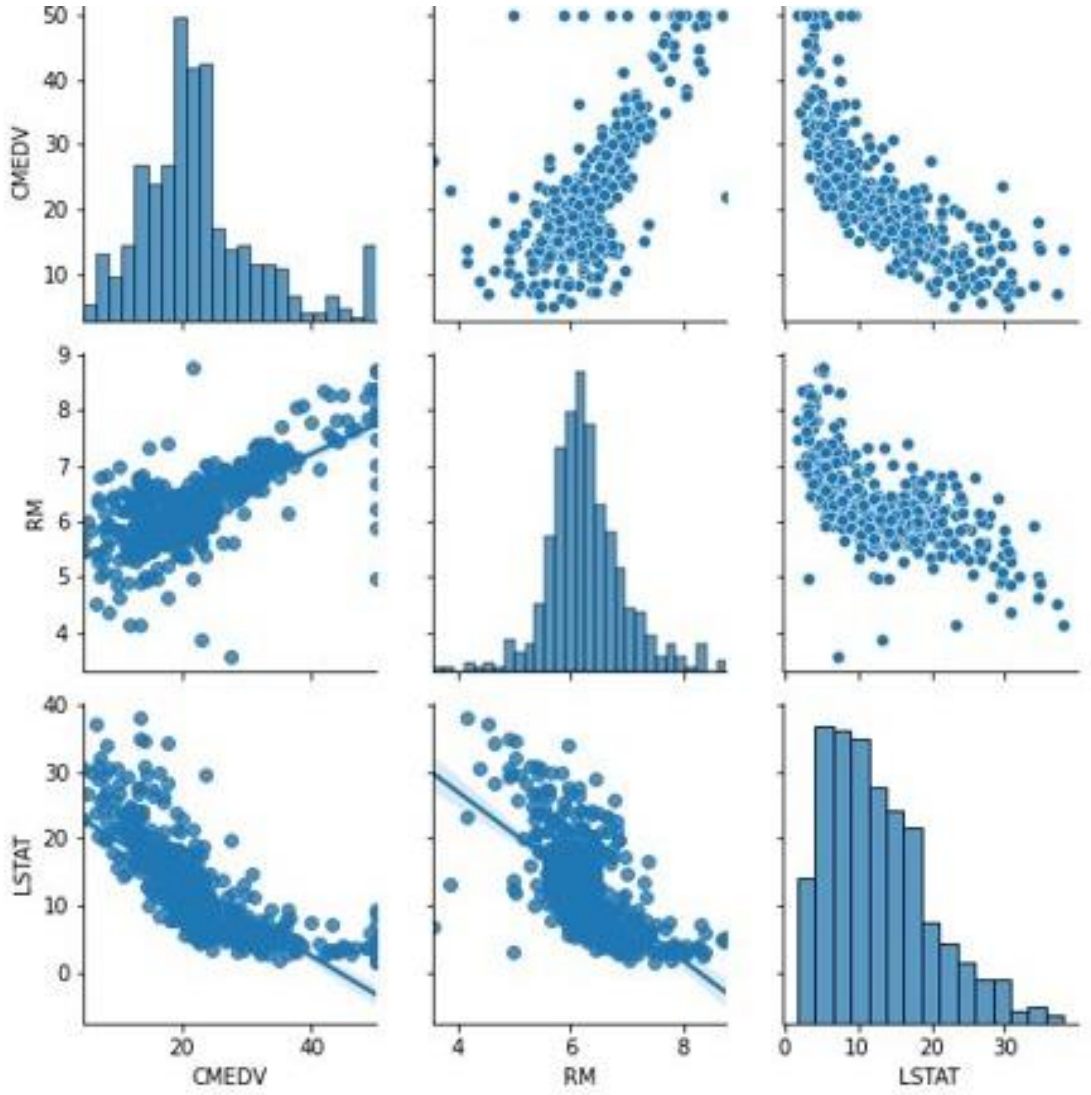
Initial Exploration of Dataset

- Computed correlation coefficients
- With regards to price the number of rooms and percentage of lower status population in the area are the most strongly correlated



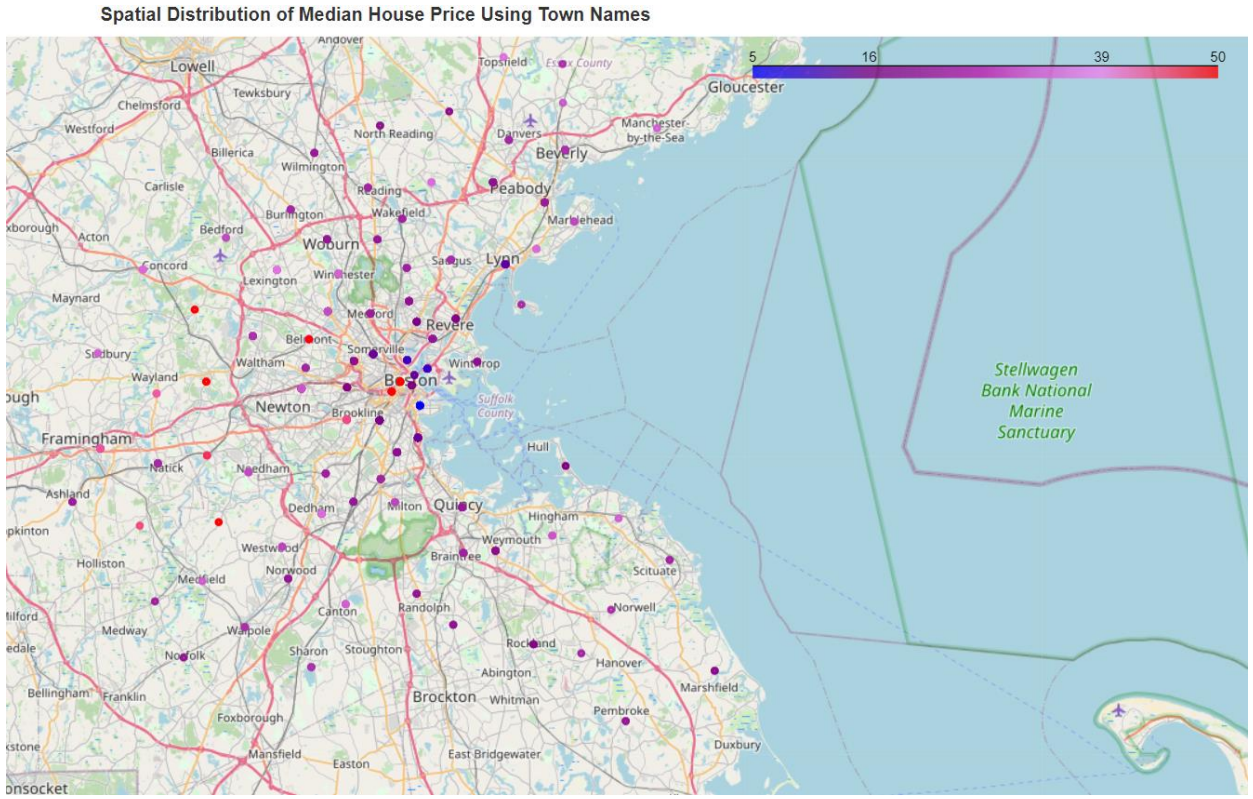
Initial Exploration of Dataset

- More evidence for the strong correlation between price and the number of rooms and the percentage of lower status population
- Notes on distribution of price and related factors

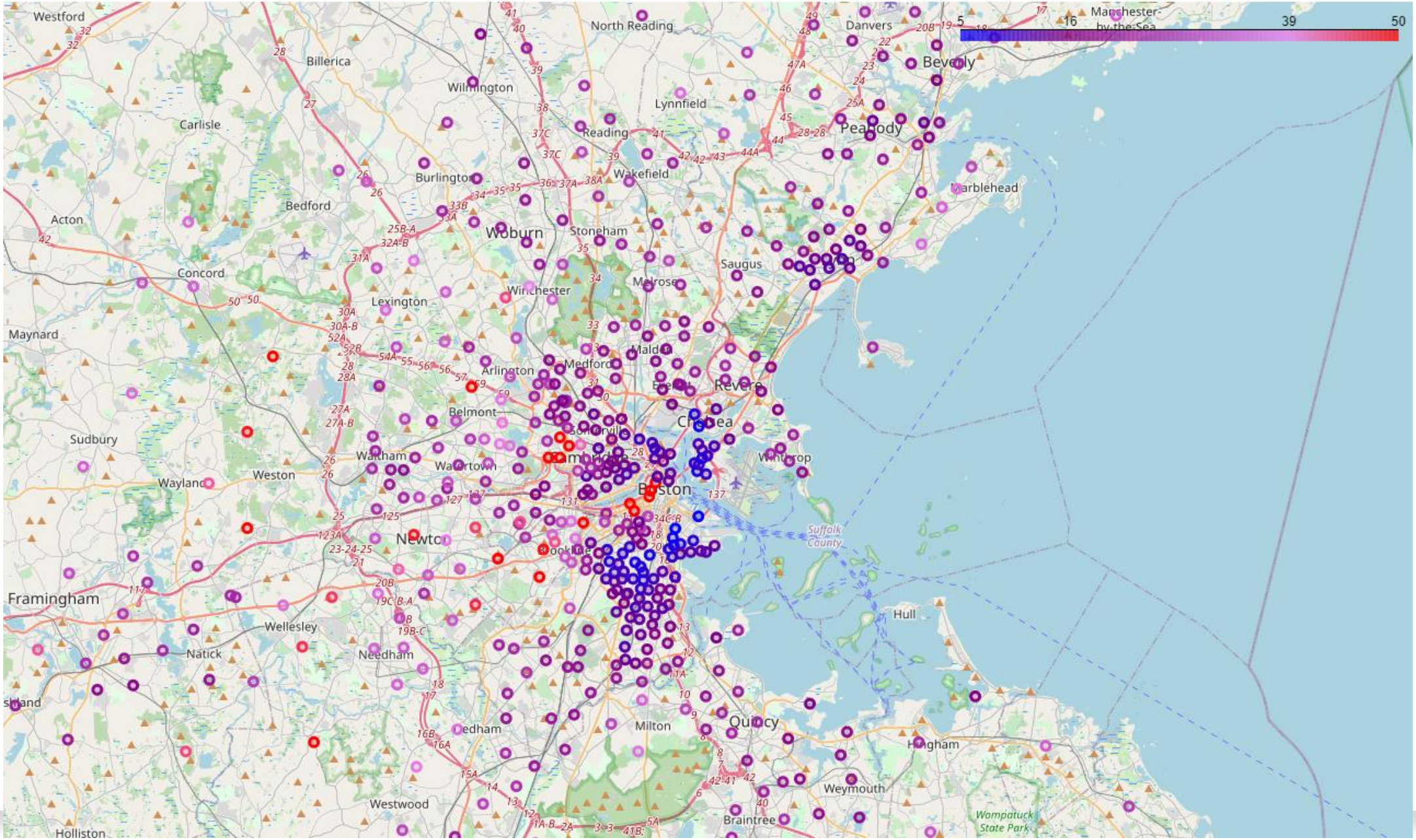


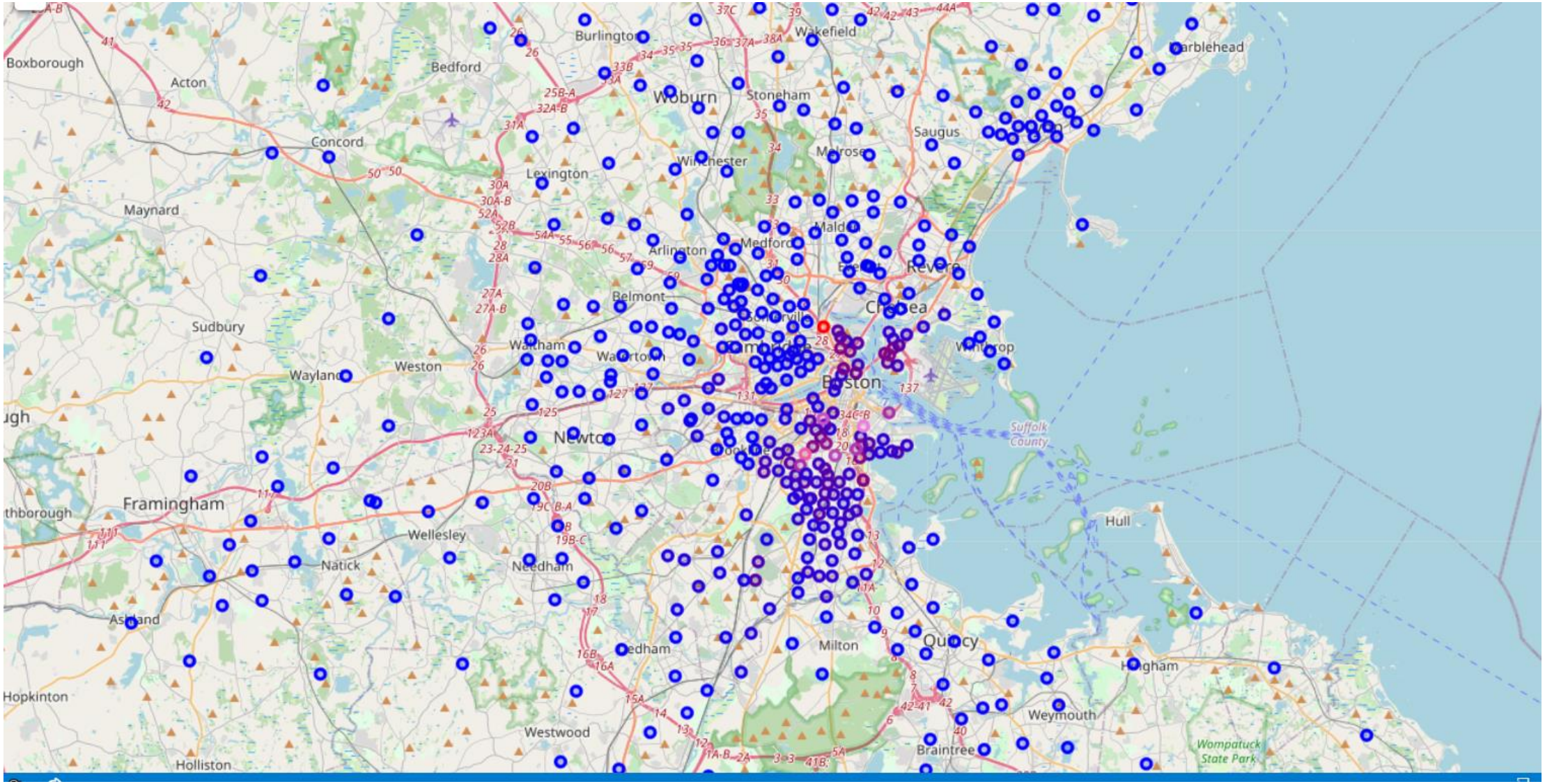
Spatial Analysis

- Initially provided longitude and latitude data was invalid, town names were used to plot distribution of metrics
- With Longitude/Latitude data provided the following day, the spatial distribution map was updated as shown in the next slide
- The effect of ‘House Price Spatial Spillover could be seen, and this was explored through clustering’



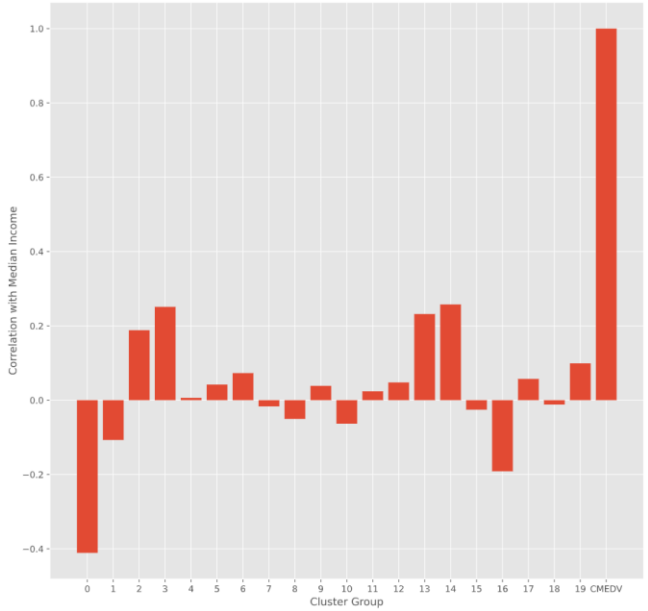
Spatial Distribution of Median House Price Using Lat/Lon Data

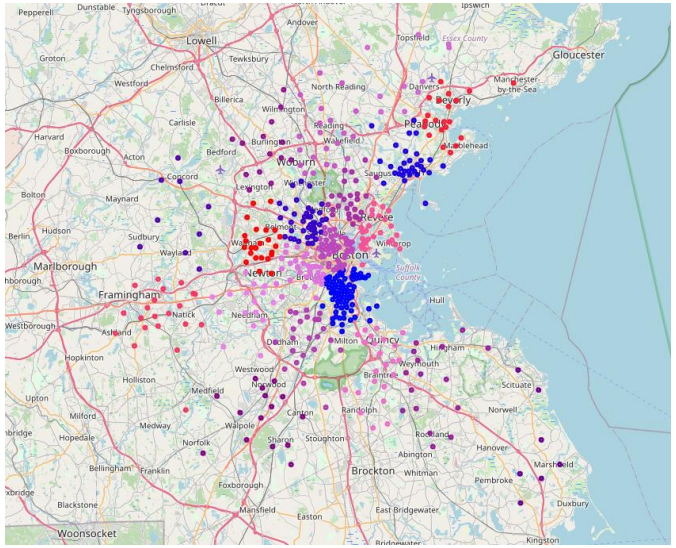



Boxborough
Acton
Concord
Maynard
Sudbury
Weston
Wayland
Framingham
Natick
Ashland
Hopkinton
Holliston
Wellesley
Needham
Westwood
Bedford
Burlington
Woburn
Stoneham
Wakefield
Lexington
Winchester
Medford
Arlington
Belmont
Waltham
Boston
Chelsea
Revere
Saugus
Malden
Milton
Quincy
Hull
Weymouth
Braintree
Suffolk County
Wompatuck State Park

Clustering

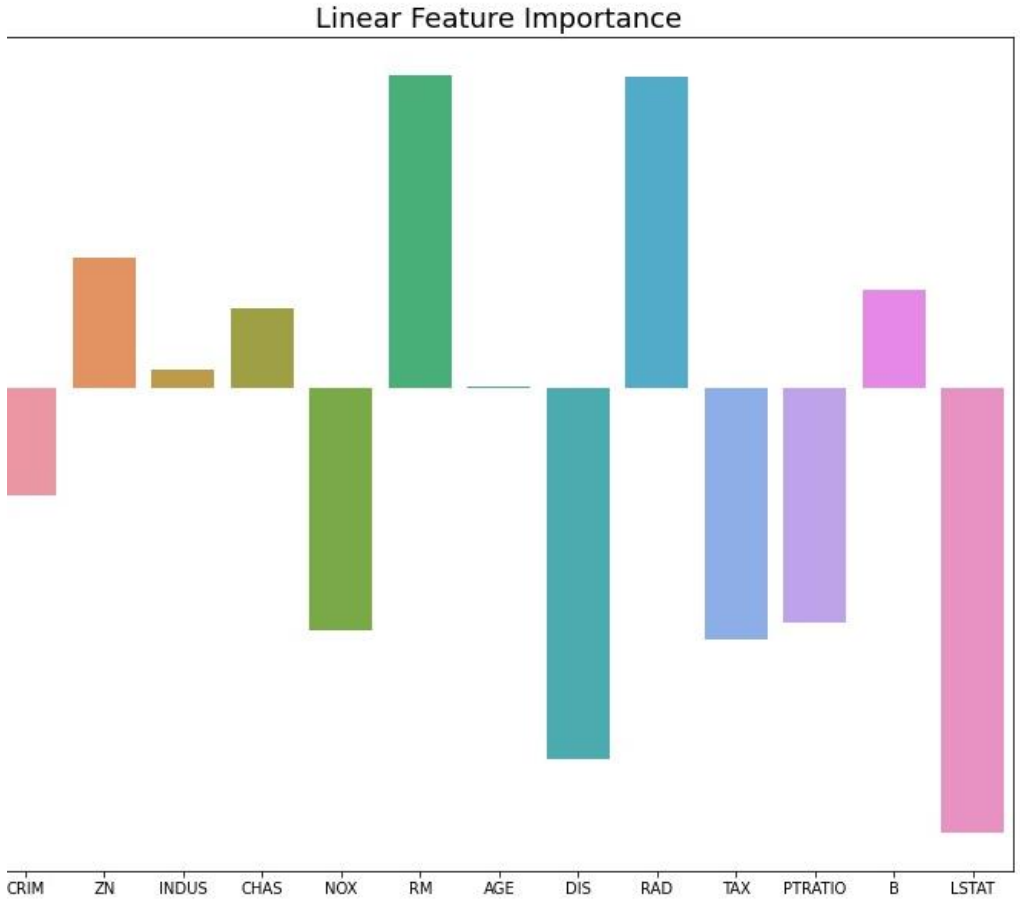
- Using K-Means clustering to organize the spatial data.
- r^2 value of 41 %





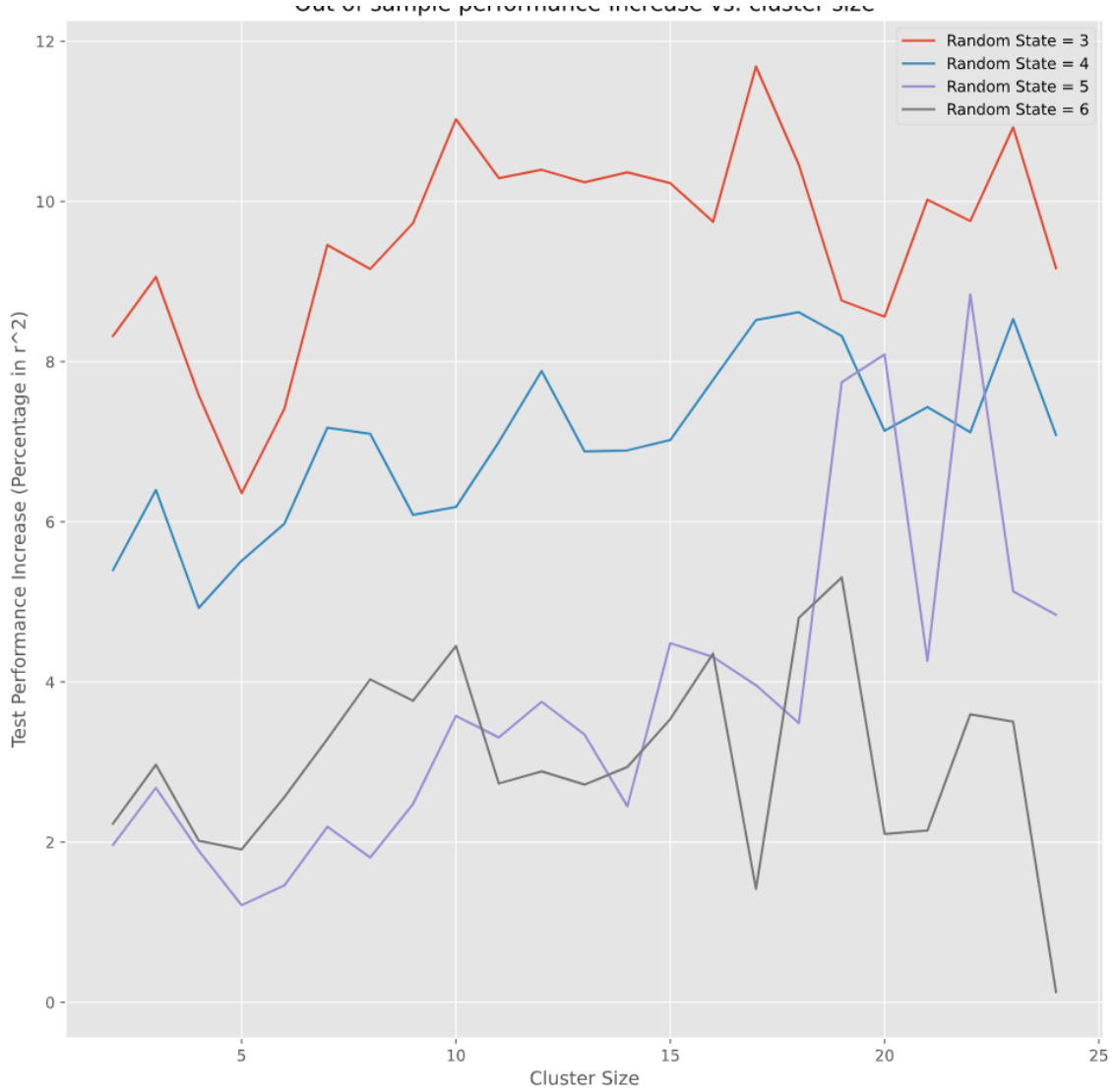
Regression on Price – Linear Regression

- Uses correlations between the features and the price as weights in the model.
- Results:
  - MSE: 22.95671894828107
  - R2: 0.7178446686547839



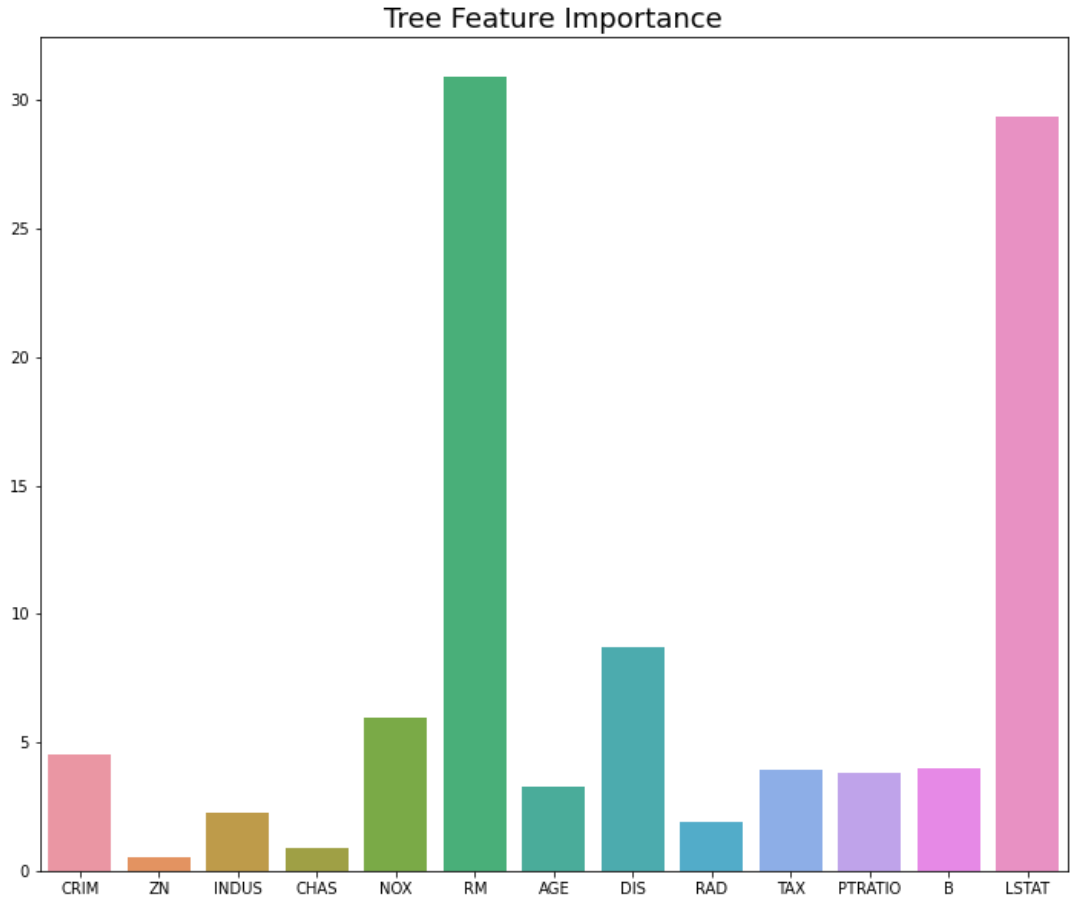
Using Cluster Group In Linear Regression Model

- With a Linear Regression model, from a base r^2 value of 68% using an 80-20% test- train split, by adding the cluster groups, performances of up to 11% could be achieved on the test set. The performance increase was measured for different cluster sizes and random states of the test train split to assess robustness



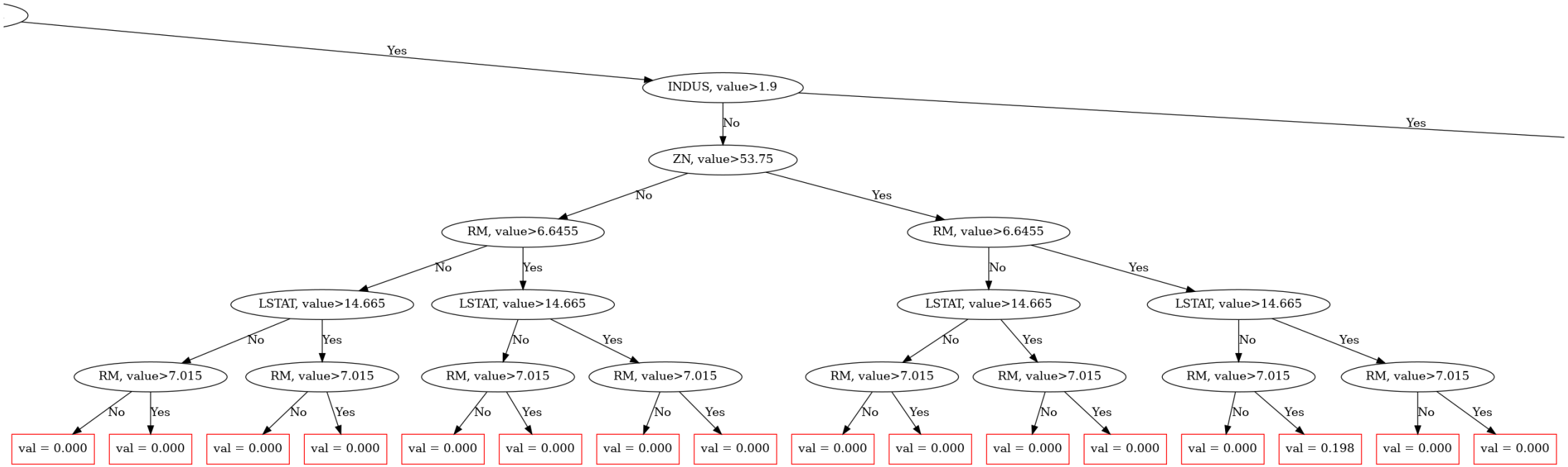
Regression on Price – Ensemble Methods

- Improved on linear regression using different Gradient Boosting trees: AdaBoost, XGBoost and CatBoost
- CatBoost had the best performance:
  - MSE: 9.112841668593683
  - R2: 0.890024039152204
- Optimal hyper parameters are exposed in the notebook



Hyper parameter search - Optuna

- We used optuna to search for optimal hyper parameters to use in the regression model coming up.
- Led to a x% increase in model performance



Regression on Price – Ensemble Methods

- Extract from one of the trees used in the CatBoost Regressor
- Splits are made using most important features

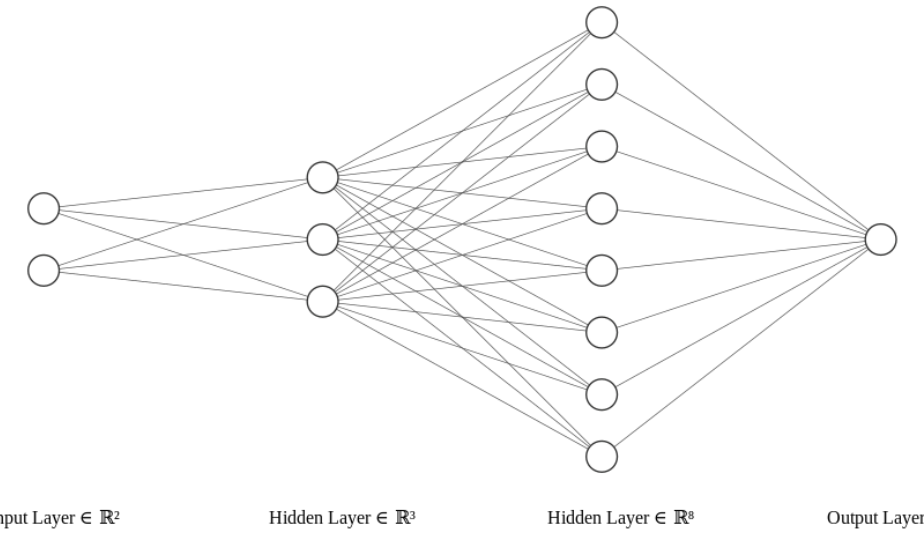
Adding the engineered cluster feature to the CatBoost model resulted in a performance boost

Regression on Price – Clustering Feature

CatBoost Model	R2	MSE
Without engineered spatial feature	0.8900240391522004	9.112841668593683
With engineered spatial feature	0.8934898067515067	8.725613355102025
% Improvement	0.389	4.25

Regression on Price – Neural Networks

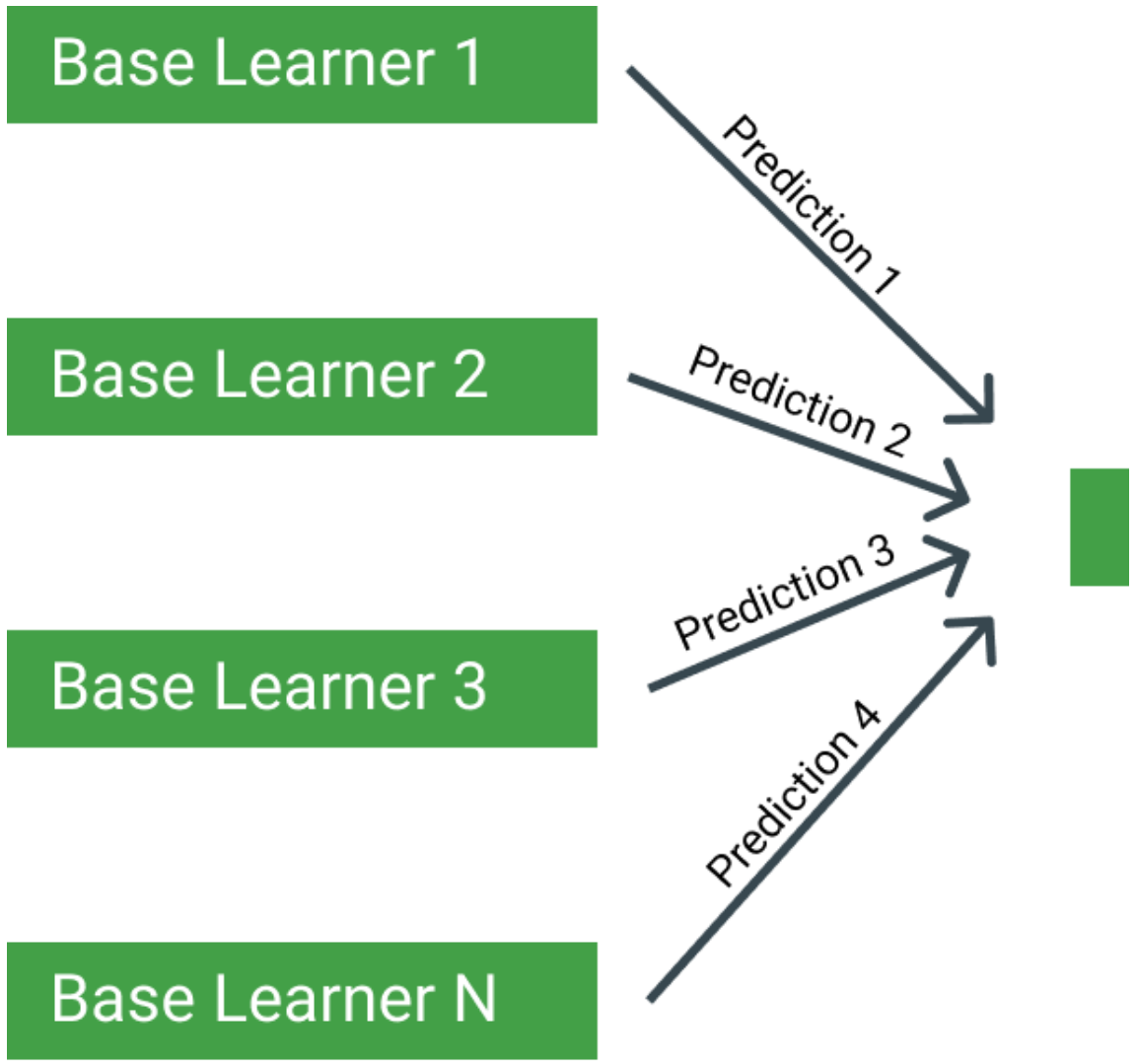
- Used simple artificial neural network structure
- Results:
  - MSE: 25.921140647451026
  - R2: 0.677052182685883
- Not enough data for deep learning methods
- Optimal hyper parameters are exposed in the notebook



Scaled down version of neural network architecture

Regression on Price – Stacking Regression

- Used results from XGBoost and CatBoost as base learners then a simpler Decision Tree model as a meta learner.
- Results:
  - MSE: 15.515438358099788
  - R2: 0.814410486439004
- Not too far off performance but very computationally expensive



Regression on Price - Summary

Model	R2 score	MSE score	Hyperparameter search time (s)	Fit time (s)
Linear Regressor	0.71784	22.957	n/a	0.00457
AdaBoost	0.83157	13.760	82	0.10824
XGBoost	0.88116	9.8720	82	0.27811
CatBoost	0.89002	9.1128	n/a	1.43539
Feedforward Neural Network (MLP)	0.67705	25.921	472	0.65697
Stacking Regressor	0.81441	15.515	n/a	20.229

Policy Based Implications

- Based on the correlation between NOx pollution, crime rates, % of lower status individuals, and house prices, there are very worrying trends shown in the inner city, particularly towards the south side of the city.
- The government should work towards developing these areas – as gentrification takes place, these areas of higher crime will only move elsewhere and wont be solved.

Intepretation and Conclusions from results

- Percentage of lower status population is strongly negatively correlated with median house price
- Existence of geographic spill-over in house prices shown through spatial regression
- Strong predictive performance of CatBoost model when combined with spatial data!